LEARNING MADE EASY

Data Science Strategy

for dummies®
A Wiley Brand

Adopt a data-driven mindset for business success

Keep your data science program focused on generating value

Nurture a top-quality data science team

Ulrika Jägare

Foreword by Lillian Pierson

Data Science Strategy

by Ulrika Jägare

FOREWORD BY **Lillian Pierson**
CEO of Data-Mania

for dummies®
A Wiley Brand

Data Science Strategy For Dummies®

Published by: **John Wiley & Sons, Inc.,** 111 River Street, Hoboken, NJ 07030-5774, www.wiley.com

Published simultaneously in Canada

Library of Congress Control Number: 2019942827

ISBN: 978-1-119-56625-0; 978-1-119-56626-7 (ebk); 978-1-119-56627-4 (ebk)

Manufactured in the United States of America

10 9 8 7 6 5 4 3 2 1

Contents at a Glance

Table of Contents

Foreword

We're living in a make-or-break era; the ability to generate business value from enterprise data will either make or break your organization. We didn't get here overnight. For years, experts have been professing how vital it is that business reframe itself to become more data-driven.

Some listened, some did not.

Organizations that took their business by its big data helm (like Netflix, Facebook, and Walmart) set the precedent. You better believe they have extremely robust data strategies in place governing those operations. The ones that did not? This book was written for you.

Sadly, over the last decade, some organizations got caught up in the media buzz. They've spent a huge amount of time and money working to hire data scientists, but haven't seen the ROI they'd expected.

Part of the problem is that it's both expensive and difficult to hire data scientists. In 2018, the median salaries for data scientists in USA ranged between $95,000 and $165,000 (see the 2018 *Burtch Works' Data Science Strategy Report*). Making matters worse, the demand for analytics-savvy workers is twice the supply (see *The Quant Crunch,* prepared for IBM by Burning Glass Technologies). No surprise that it's exceedingly difficult to recruit and retain these type of professionals.

But a bigger part of the problem is just this — contrary to what most advocates will tell you, **just sourcing and hiring a team of "Data Scientists" isn't going to get your organization where it needs to be.** You'll also need to secure a robust set of big data skill sets, technologies, and data resources. More importantly, you'll need a comprehensive big data strategic plan in place, to help you steer your data ship.

It takes a lot more than just implementation folks dealing with all the details of your data initiatives; you also need an expert to manage them. You need someone who can communicate with and manage your data team, can communicate effectively with organizational leaders, can build relationships with business stakeholders, and who can perform exhaustive evaluations of both your business and your data assets in order to form the data strategy your business will need to survive in the digital era. Read this book for details on how to get these elements in place.

All around the world, I've been on the frontlines supporting organizations that know their data's value and are ready to make big changes to start extracting that value. At Data-Mania, we provide results-driven data strategy services to optimize our client's data operations. We are also leading the change by training our client's staff with the data strategy and data science skills they need to succeed. Through our partnerships with LinkedIn and Wiley, over the last five years we've educated about a million technical professionals globally. Across both of these functions and with each project we engage, one message strongly resounds — The people and organizations who are committed to taking necessary actions to transform enterprise data to business value are the ones that will prevail in the digital era.

I want to be the first to congratulate you! Just by picking up this book and making the effort to educate yourself on the problems and solutions related to data strategy, you've already taken the first step. Whether you're a C-suite executive that's looking for guidance on next steps for your organization, or if you're a data professional looking to move forward in your career, *Data Science Strategy For Dummies* will provide you a solid framework around which to proceed.

It's an exciting time to be alive. Never before have businesses had access to such a powerful upper hand. Those of us who recognize this in our business data are the ones who are primed to blaze the trail and build a true legacy with the work we do in our careers. Some of us have been on this path for a while now, while others are new. Welcome aboard!

Lillian Pierson, P.E.

Data Strategist & CEO of Data-Mania

Introduction

A revolutionary change is taking place in society. Everybody, from small local companies to global enterprises, is starting to realize the potential in digitizing their data assets and becoming data driven. Regardless of industry, companies have embarked on a similar journey to explore how to drive new business value by utilizing analytics, machine learning (ML), and artificial intelligence (AI) techniques and introducing data science as a new discipline.

However, although utilizing these new technologies will help companies simplify their operations and drive down costs, nothing is simple about getting the strategic approach right for your data science investment. And, the later you join the ML/AI game, the more important it will be to get the strategy right from the start for your particular area of business. Hiring a couple of data scientists to play around with your data is easy enough to do — if you can find some of the few that are available — but the real heavy lifting comes when you try to understand how to utilize data science to create value throughout your business and put that understanding into an executable data science strategy. If you can do that, you are on the right path for success.

A recent survey by Deloitte of "aggressive adopters" of cognitive technologies found that 76 percent believe that they will "substantially transform" their companies within the next three years by using data and AI. IDC, a global marketing intelligence firm, predicts that by 2021, 75 percent of commercial enterprise apps will use AI, over 90 percent of consumers will interact with customer support bots; and over 50 percent of new industrial robots will leverage AI.

However, at the same time, there remains a very large gap between aspiration and reality. Gartner, yet another research and advisory company, claimed in 2017 that 85 percent of all big data projects fail; not only that, there still seems to be confusion around what the true key success factors are to succeed when it comes to data and AI investments. This book argues that a main key success factor is a great data science strategy.

The target audience for this book is anyone interested in making well-balanced strategic choices in the field of data science, no matter which aspect you're focusing on and at what level — from upper management all the way down to the individual members of a data science team. Strategic choices matter! And, this book is based on actual experiences arising from building this up from scratch in a global enterprise, incorporating learnings from successful choices as well as mistakes and miscalculations along the way.

So far, there seems to be little in-depth research or analysis on the topic of data science and AI strategies and little practical guidance as well. In fact, when researching for this book, I couldn't find another single book on the topic of data science strategy. However, several interesting articles and reports are available, like TDWI's report, "Seven Steps for Executing a Successful Data Science Strategy" (https://tdwi.org/research/2015/01/checklist-seven-steps-successful-data-science-strategy.aspx?tc=page0&m=1) or The Startup's "How To Create A Successful Artificial Intelligence Strategy" https://medium.com/swlh/how-to-create-a-successful-artificial-intelligence-strategy-44705c588e62). However, these articles primarily focus on easily consumable tips and tricks, while bringing up a few aspects of the challenges and considerations needed. There is an obvious lack of in-depth guidance which is not really accessible in an article format.

At the same time, the main reasons companies fail with their data science or AI investment is that either there was no data science strategy in place or the complexity of executing on the strategy wasn't understood. Although this enormous transformation is happening right here, right now, all around us, it seems that few people have grasped how data science will impose a fundamental shift in society — and therefore don't understand how to approach it. This book is based on more than ten years of experience spent driving different levels of strategic and practical transformation assignments in a global enterprise. As such, it will help you understand what is fundamentally important to consider and what you should avoid. (Trust me: There are many pitfalls and areas to get stuck in.) But if you want to be in the forefront with your business, you have neither the time nor the money to make mistakes. You really want a solid, end-to-end data science strategy that works for you at the level you need in order to bring your organization forward. The time is now! This is the book that everyone in data science should read.

About This Book

This book will help guide you through the different areas that need to be considered as part of your data science strategy. This includes managing the complexity in data science and avoiding common data challenges, making strategic choices related to the data itself (including how to capture it, transfer it, compute it, and keep it secure and legally compliant), but also how to build up efficient and successful data science teams.

Furthermore, it includes guidance on strategic infrastructure choices to enable a productive and innovative environment for the data science teams as well as how to acquire and balance data science competence and enable productive ways of working. It also includes how you can turn data into enhanced or new business

opportunities, including data-driven business models for new data products and services, while also addressing ethical aspects related to data usage and commercialization.

My goal here is to give you relevant and concrete guidance in those areas that require strategic thinking as well as give some advice on what to include when making choices for both your data and AI investment as well as how best to come up with a useful and applicable data science strategy. Based on my own experience in this field, I'll argue for certain techniques or technology choices or even preferred ways of working, but I won't come down on one side or the other when it comes to any specific products or services. The most I'll do in that regard is point out that certain methods or technology choices are more appropriate for certain types of users rather than others.

Foolish Assumptions

Because this book assumes a basic level of understanding of what data science actually is, don't think of it as an introduction to data science, but rather as a tool for optimizing your analytics and/or ML/AI investment, regardless of whether that investment is for a small company or a global enterprise. It covers everything from practical advice to deep insights into how to define, focus, and make the right strategic choices in data science throughout. So, if you're looking to find a broad understanding of what data science is, which techniques and ML tools come recommended, and how to get started as a data scientist professional, I instead warmly recommend the book *Data Science For Dummies*, by Lillian Pierson (Wiley).

How This Book Is Organized

This book has six main parts. Part 1 outlines the major challenges that companies (small as well as large) face when investing in data science. Whereas Part 2 aims to create an understanding of the strategic choices in data science that you need to make, Part 3 guides you in successfully setting up and shaping your data science teams. In Part 4, you find out about important infrastructure considerations, managing models in development and production and how to relate to open source. In Part 5 you learn all about commercializing your data business and monetizing your data. And, and is the case with all *For Dummies* books, this book ends with The Part of Tens, with some practical tips, including what not to do when building your data science strategy and spelling out why you need to create a data science strategy to begin with.

Icons Used In This Book

I'll occasionally use a few special icons to focus attention on important items. Here's what you'll find:

This icon with the proverbial string around the finger reminds you about information that's worth recalling.

Expect to find something useful or helpful by way of suggestions, advice, or observations.

The Warning icon is meant to grab your attention so that can you steer clear of potholes, money pits, and other hazards.

This icon may be taken in one of two ways: Techies will zero in on the juicy and significant details that follow; others will happily skip ahead to the next paragraph.

Beyond The Book

This book is designed to help you explore different strategic options for your data science investment. It will guide you in your choices for your business, from data-driven business models to data choices and from team setup to infrastructure choices and a lot more. It will help you navigate the most common challenges and steer you toward the success factors.

However, this book is aimed at covering a very broad range of areas in data science strategy development, and is therefore not able to deep-dive into specific theories or techniques to the level you might be looking for after reading parts of this book.

In addition to what you're reading right now, this product comes with a free access-anywhere Cheat Sheet that offers a number of data-science-related tips, techniques, and resources. To get this Cheat Sheet, visit `www.dummies.com` and type **data science strategy for dummies cheat sheet** in the Search box.

Where To Go From Here

You can start reading this book anywhere you like. You don't have to read in chapter order, but my suggestion is to start by studying how data science is framed in this book, which is outlined in Chapter 1. In that chapter, you can also learn about the complexity and challenges you will encounter, before diving into subsequent chapters, where I explain how to tackle the challenges most enterprises encounter when strategically investing in data science.

1 Optimizing Your Data Science Investment

IN THIS PART . . .

Defining a data science strategy

Grasping the complexity in data science

Tackling major challenges in the field of data science

Addressing change in a data-driven organization

IN THIS CHAPTER

- Clarifying the concept of data science
- Understanding the fundamentals of a data-driven organization
- Putting machine learning in context of data science
- Clarifying the components of an effective data science strategy

Chapter 1
Framing Data Science Strategy

In this chapter, I aim to sort out the basics of what data science is all about, but I have to warn you that data science is a term that escapes any single complete definition — which, of course, makes data science difficult to understand and apply in an organization. Many articles and publications use the term quite freely, with the assumption that it's universally understood. Yet, data science — including its methods, goals, and applications — evolves with time and technology and is now far different from what it might have been 25 years ago.

Despite all that, I'm willing to put forward a tentative definition: *Data science* is the study of where data comes from, what it represents, and how it can be turned into a valuable resource in the creation of business strategies. Data science can be said to be a multidisciplinary field that uses scientific methods, processes, algorithms, and systems to extract insights from data in various forms, both structured and unstructured. Mining large amounts of structured and unstructured data to identify patterns and deviations that can help an organization rein in costs, increase efficiencies, recognize new market opportunities, and increase the organization's competitive advantage.

Data science is a concept that can be used to unify statistics, analytics, machine learning, and their related methods and techniques in order to understand and analyze actual phenomena with data. It employs techniques and theories drawn

from many fields within the context of mathematics, statistics, information science, and computer science.

Behind that type of definition though, lies the definition of how data science is approached and performed. And because the ambition of this part of the book is to frame data science strategy, I need to first frame this multidisciplinary area of data science and its life cycle more properly.

Establishing the Data Science Narrative

It never hurts to have an image when explaining a complicated process, so do take a look at Figure 1-1, where you can see the main steps or phases in the data science life-cycle. Keep in mind, however, that the model visualized in Figure 1-1 assumes that you've already identified a high-level business problem or business opportunity as a starting point. This early ambition is usually derived from a business perspective, but it needs to be analyzed and framed in detail together with the data science team. This dialogue is vital in terms of understanding which data is available and what is possible to do with that data so you can set the focus of the work going forward. It isn't a good idea to just start capturing any and all data that looks interesting enough to analyze. Therefore, the first stage of the data science life cycle, *capture,* is to frame the data you need by translating the business need into a concrete and well-defined problem or business opportunity.

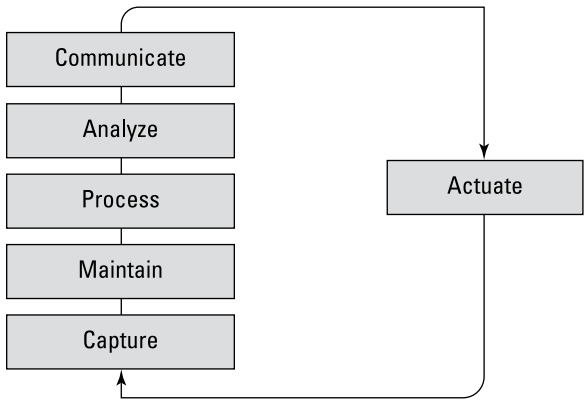

FIGURE 1-1: The different stages of the data science life cycle.

TIP

The initial business problem and/or opportunity isn't static and will change over time as your data-driven understanding matures. Staying flexible in terms of which data is captured as well as which problem and/or opportunity is most important at any given point in time, is therefore a vital in order to achieve your business objectives.

The model shown in Figure 1-1 aims to represent a view of the different stages of the data science life cycle, from capturing the business and data need through preparing, exploring, and analyzing the data to reaching insights and acting on them.

The output of each full cycle produces new data, which provides the result of the previous cycle. This includes not only new data or results, which you can use to optimize your model, but can also generate new business needs, problems, or even a new understanding of what the business priority should be.


REMEMBER

These stages of the data science life cycle can also be seen as not only steps describing the scope of data science but also layers in an architecture. More on that later; let me start by explaining the different stages.

Capture

There are two different parts of the first stage in the life-cycle, since *capture* refers to both the capture of the business need as well as the extraction and acquisition of data. This stage is vital to the rest of the process. I'll start by explaining what it means to capture the business need.

The starting point for detailing the business need is a high-level business request or business problem expressed by management or similar entities and should include tasks such as

- **Translating ambiguous business requests** into concrete, well-defined problems or opportunities
- **Deep-diving into the context of the requests** to better understand what a potential solution could look like, including which data will be needed
- **Outlining (if possible) strategic business priorities** set by the company that might impact the data science work

Now that I've made clear the importance of capturing and understanding the business requests and initial scoping of data needed, I want to move on to describing aspects of the data capture process itself. It's the main interface to the data source that you need to tap into and includes areas such as

- Managing data ownership and securing legal rights to data capture and usage
- Handling of personal information and securing data privacy through different anonymization techniques
- Using hardware and software for acquiring the data through batch uploads or the real-time streaming of data

- Determining how frequently data will need to be acquired, because the frequency usually varies between data types and categories
- Mandating that the preprocessing of data occurs at the point of collection, or even before collection (at the edge of an IoT device, for example). This includes basic processing, like cleaning and aggregating data, but it can also include more advanced activities, such as anonymizing the data to remove sensitive information. (*Anonymizing* refers to removing sensitive information such as a person's name, phone number, address and so on from a data set.)


REMEMBER

In most cases, data must be anonymized before being transferred from the data source. Usually a procedure is also in place to validate data sets in terms of completeness. If the data isn't complete, the collection may need to be repeated several times to achieve the desired data scope. Performing this type of validation early on has a positive impact on both process speed and cost.

- Managing the data transfer process to the needed storage point (local and/or global). As part of the data transfer, you may have to transform the data — aggregating it to make it smaller, for example. You may need to do this if you're facing limits on the bandwidth capacity of the transfer links you use.

Maintain

Data *maintenance* activities includes both storing and maintaining the data. Note that data is usually processed in many different steps throughout its life cycle.


WARNING

The need to protect data integrity during the life cycle of a data element is especially important during data processing activities. It's easy to accidentally corrupt a dataset through human error when manually processing data, causing the data set to be useless for analysis in the next step. The best way to protect data integrity is to automate as many steps as possible of the data management activities leading up to the point of data analysis.


REMEMBER

Keeping business trust in the data foundation is vital in order for business users to trust and make use of the derived insights.

When it comes to maintaining data, two important aspects are

- **Data storage:** Think of this as everything associated with what's happening in the data lake. Data storage activities include managing the different retention periods for different types of data, as well as cataloging data properly to ensure that data is easy to access and use.

- **Data preparation:** In the context of maintaining data, data preparation includes basic processing tasks such as second-level data cleansing, data staging, and data aggregation, all of which usually involve applying a filter directly when the data is put into storage. You don't want to put data with poor quality into your data lake.

REMEMBER

Data retention periods can be different for the same data type, depending on its level of aggregation. For example, raw data might be interesting to save for only a short time because it's usually very large in volume and therefore costly to store. Aggregated data on the other hand, is often smaller in size and cheaper and easier to store and can therefore be saved for longer periods, depending on the targeted use cases.

Process

Processing of data is the main data processing layer focused on preparing data for analysis, and it refers to using more advanced data engineering methodologies, such as

- **Data classification:** This refers to the process of organizing data into categories for even more effective and efficient use, including activities such as the labeling and tagging of data. A well-planned data classification system makes essential data easy to find and retrieve. This can also be of particular importance for areas such as legal and compliance.
- **Data modeling:** This helps with the visual representation of data and enforces established business rules regarding data. You would also build data models to enforce policies on how you should correlate different data types in a consistent manner. Data models also ensure consistency in naming conventions, default values, semantics, and security procedures, thus ensuring quality of data.
- **Data summarization:** Here your aim is to use different ways to summarize data, like using different clustering techniques.
- **Data mining:** This is the process of analyzing large data sets to identify patterns or deviations as well as to establish relationships in order to enable problems to be solved through data analysis further down the road. Data mining is a sort of data analysis, focused on enhanced understanding of data, also referred to as *data literacy.* Building data literacy in the data science teams is a key component of data science success.

  With low data literacy, and without truly understanding the data you're preparing, analyzing, and deriving insights from, you run a high risk of failing when it comes to your data science investment.

WARNING

Analyze

Data *analysis* is the stage where the data comes to life and you're finally able to derive insights from the application of different analytical techniques.

REMEMBER

Insights can be focused on understanding and explaining what has happened, which means that the analysis is descriptive and more reactive in nature. This is also the case with real-time analysis: It's still reactive even when it happens in the here-and-now.

Then there are data analysis methods that aim to explain not only *why* something happened but also *what* happened. These types of data analysis are usually referred to as *diagnostic* analyses.

Both descriptive and diagnostic methods are usually grouped into the area of reporting, or business intelligence (BI).

To be able to predict what will happen, you need to use a different set of analytical techniques and methods. Predictions about the future can be done strategically or in real-time settings. For a real-time prediction you need to develop, train and validate a model before deploying it on real-time data. The model could then search for certain data patterns and conditions that you have trained the model to find, to help you predict a problem before it happens.

Figure 1-2 shows the difference between reporting techniques about what has happened (in black) and analytics techniques about what is likely to happen, using statistical models and predictive models (in white).

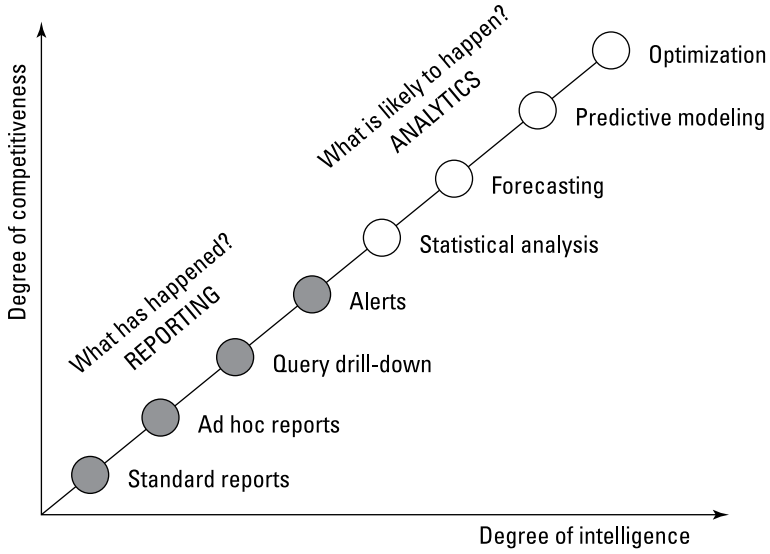

FIGURE 1-2: The difference between reporting and analytics.

This list gives you examples of the kinds of questions you can ask using different reporting and BI techniques:

- **Standard reports:** What was the customer churn rate?
- **Ad hoc reports:** How did the code fix carried out on a certain date impact product performance?
- **Query drill-down:** Are similar product-quality issues reported in all geographical locations?
- **Alerts:** Customer churn has increased. What action is recommended?

And this list gives you examples of the kinds questions you can ask using different analytics techniques:

- **Statistical analysis:** Which factors contribute most to the product quality issues?
- **Forecasting:** What will bandwidth demand be in 6 months?
- **Predictive modeling:** Which customer segment is most likely to respond to this marketing campaign?
- **Optimization.** What is the optimal mix of customer, offering, price, and sales channel?

Analytics can also be separated into two categories: basic analytics and advanced analytics. *Basic* analytics uses rudimentary techniques and statistical methods to derive value from data, usually in a manual manner, whereas in advanced analytics, the objective is to gain deeper insights, make predictions, or generate recommendations by way of an autonomous or semiautonomous examination of data or content using more advanced and sophisticated statistical methods and techniques.

Some examples of the differences are described in this list:

- **Exploratory data analytics** is a statistical approach to analyzing data sets in order to summarize their main characteristics, often with visual methods. You can choose to use a statistical model or not, but if used, such a model is primarily for visualizing what the data can tell you beyond the formal modeling or hypothesis testing task. This is categorized as basic analytics.
- **Predictive analytics** is the use of data, statistical algorithms, and machine learning techniques to identify the likelihood of future outcomes based on historical data. This is categorized as advanced analytics.

- **Regression analysis** is a way of mathematically sorting out which variables have an impact. It answers these questions: Which factors matter most? Which can be ignored? How do those factors interact with each other? And, perhaps most importantly, how certain am I about all these factors? This is categorized as advanced analytics.
- **Text mining or text analytics** is the process of exploring and analyzing large amounts of unstructured text aided by software that can identify concepts, patterns, topics, keywords, and other attributes in the data. The overarching goal of text mining is, to turn text into data for analysis via application of natural language processing (NLP) and various analytical methods. Text mining can be done from a more basic perspective as well as from a more advanced perspective, depending on the use case.

Communicate

The *communication* stage of data science is about making sure insights and learnings from the data analysis are understood and communicated by way of different means in order to come to efficient use. It includes areas such as

- **Data reporting:** The process of collecting and submitting data in order to enable an accurate analysis of the facts on the ground. It's a vital part of communication because inaccurate data reporting can lead to vastly uninformed decision-making based on inaccurate evidence.
- **Data visualization:** This can also be seen as *visual communication* because it involves the creation and study of the visual representation of data and insights. To help you communicate the result of the data analysis clearly and efficiently, data visualization uses statistical graphics, plots, information graphics, and other tools. Effective visualization helps users analyze and reason about data and evidence because it makes complex data more accessible, understandable, and usable.

Users may have been assigned particular analytical tasks, such as making comparisons or understanding causality, and the design principle of the graphical visualization (showing comparisons or showing causality, in this example) follows the task. Tables are generally used where users can look up a specific measurement, and charts of various types are used to show patterns or relationships in the data for one or more variables.

Figure 1-3 below exemplifies how data exploration could work using a table format. In this specific case, the data being explored regards cars, and the hypothesis being tested is which car attribute impacts fuel consumption the most. Is it, for example, the car brand, engine size, horse power or perhaps the weight of the car?

As you can see, exploring the data using tables has its limitation, and does not give an immediate overview. It requires you to go through the data in detail to discover relationships and patterns. Compare this with the graph shown in Figure 1-4 below, where the same data is being visualized in a completely different way.

Make	Model	Origin	DriveTrain ▲	Type	Cylinders	Engine Size (L)	Frequency	Fuel Consumtion (L/10 km)	Horsepower	Invoice	Length (IN)	MSRP
Mercedes-B...	E500	Eur...	All	Wa...	8	5	1	1,17605	302	56 47...	190	60 67...
Isuzu	Ascender S	Asia	All	SUV	6	4,2	1	1,3440571429	275	29 97...	208	31 84...
Toyota	Tundra Access Cab V6 SR5	Asia	All	Truck	6	3,4	1	1,517483871	190	23 52...	218	25 93...
Audi	A6 3.0 Quattro 4dr	Eur...	All	Sedan	6	3	1	1,094	220	35 99...	192	39 64...
Volvo	S60 R 4dr	Eur...	All	Sedan	5	2,5	1	1,094	300	35 38...	181	37 56...
Acura	MDX	Asia	All	SUV	6	3,5	1	1,17605	265	33 33...	189	36 94...
Nissan	Titan King Cab XE	Asia	All	Truck	8	5,6	1	1,4700625	305	24 92...	224	26 65...
Toyota	Sequoia SR5	Asia	All	SUV	8	4,7	1	1,517483871	240	31 82...	204	35 69...
Audi	S4 Quattro 4dr	Eur...	All	Sedan	8	4,2	1	1,3835882353	340	43 55...	179	48 04...
Audi	A8 L Quattro 4dr	Eur...	All	Sedan	8	4,2	1	1,1473658537	330	64 74...	204	69 19...
Subaru	Impreza WRX STi 4dr	Asia	All	Sports	4	2,5	1	1,120047619	300	29 13...	174	31 54...
BMW	330xi 4dr	Eur...	All	Sedan	6	3	1	0,9600408163	225	34 11...	176	37 24...
Infiniti	FX45	Asia	All	Wa...	8	4,5	1	1,3835882353	315	33 12...	189	36 39...
Toyota	Land Cruiser	Asia	All	SUV	8	4,7	1	1,5680666667	325	47 98...	193	54 76...
Subaru	Forester X	Asia	All	Wa...	4	2,5	1	0,9600408163	165	19 64...	175	21 44...
GMC	Sierra HD 2500	USA	All	Truck	8	6	1	1,517483871	300	25 75...	222	29 32...
Mercedes-B...	C240 4dr	Eur...	All	Sedan	6	2,6	1	1,0691363636	168	31 18...	178	33 48...
Jaguar	X-Type 3.0 4dr	Eur...	All	Sedan	6	3	1	1,094	227	30 99...	184	33 99...
Volvo	S80 2.5T 4dr	Eur...	All	Sedan	5	2,5	1	1,000893617	194	35 68...	190	37 88...
Volvo	XC70	Eur...	All	Wa...	5	2,5	1	1,000893617	208	33 11...	186	35 14...
Audi	A6 2.7 Turbo Quattro 4dr	Eur...	All	Sedan	6	2,7	1	1,094	250	38 84...	192	42 84...
Dodge	Grand Caravan SXT	USA	All	Sedan	6	3,8	1	1,094	215	29 81...	201	32 66...

FIGURE 1-3: Example of data exploration using a table.

Figure 1-3 is based on a screenshot generated using SAS® Visual Analytics software. Copyright © 2019 SAS Institute Inc., Cary, NC, USA. SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. All Rights Reserved. Used with permission.

In Figure 1-4, a visualization in the shape of a linear regression graph has been generated for each car attribute, together with text explaining the strength of each relationship to fuel consumption. (Linear regression involves fitting a straight line to a dataset while trying to minimize the error between the points and the fitted line.) The graph in Figure 1-4 shows a very strong positive relationship between the weight of the car and fuel consumption. By studying the relationship between the other attributes and fuel consumption using the graph generated for each tab, it will be quite easy to find the strongest relationship compared to using the table in Figure 1-3.

However, in data exploration the key is to stay flexible in terms of which exploration methods to use. In this case, it was easier and quicker to find the relationship by using linear regression, but in another case a table might be enough, or none of the just mentioned approaches works. If you have geographical data, for example, the best way to explore it might be by using a geo map, where the data is distributed based on geographical location. But more about that later on.

Actuate

The final stage in the data science life cycle is to *actuate* the insights derived from all previous stages. This stage has not always been seen as part of the data science life cycle, but the more that society moves toward embracing automation, the more the interest in this area grows.

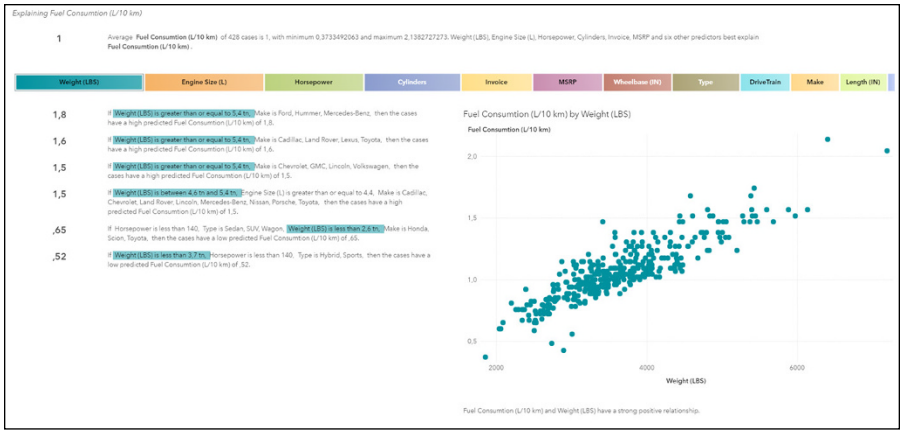

FIGURE 1-4: Visualizing your data.

Figure 1-4 is based on a screenshot generated using SAS® Visual Analytics software. Copyright © 2019 SAS Institute Inc., Cary, NC, USA. SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. All Rights Reserved. Used with permission.

Decision-making for actuation refers to connecting an insight derived from data analysis to trigger a human- or machine-driven decision-making process of identifying and deciding alternatives for the right action based on the values, policies, preferences, or beliefs related to the business or scope of the task.

TECHNICAL STUFF

What actually occurs is that a human or machine compares the insight with a predefined set of policies for what needs to be done when a certain set of criteria is fulfilled. If the criteria are fulfilled, this triggers a decision or an action. The actuation trigger can be directed toward a human (for example, a manager) for further decisions to be made in a larger context, or toward a machine when the insight falls within the scope of the predefined policies for actuation.

REMEMBER

Automation of tasks or decisions increases speed and reduces cost, and if set up properly, also produces continuous and reliable data on the outcome of the implemented action.

The stage where decisions are actuated — by either human hand or a machine — is one of the most important areas of data science. It's fundamental because it will provide data science professionals (also known as *data scientists*) with new data based on the results of the action (resolution or prevention of a problem, for example), which tells the data scientists whether their models and algorithms are performing as expected after deployment or whether they need to be corrected or improved. The follow-up regarding model and algorithm performance also supports the concept of continuous improvement.

PUTTING AUTOMATION IN THE CONTEXT OF DATA SCIENCE

What is actually the relationship between data science and automation? And, can automation accelerate data science production and efficiency? Well, assuming that the technology evolution in society moves more and more toward automation, not only for simple process steps previously performed by humans but also for complex actions identified and decided by intelligent machines powered by machine-learning-developed algorithms, the relationship will be a strong one, and data science production and efficiency will accelerate considerably due to automation.

The decisions will, of course, not really be decided by the machines, but will be based on human-preapproved policies that the machine then acts on. *Machine learning* doesn't mean that the machine can learn unfettered, but rather that it always encounters boundaries for the learning set up by the data scientist — boundaries regulated by established policies. However, within these policy boundaries, the machine can learn to optimize the analysis and execution of tasks assigned to it.

Despite the boundaries imposed on it, automation powered by machines will become more and more important in data science, not only as a means to increase speed (from detection to correction or prevention) but also to lower cost and secure quality and consistency of data management, actuation of insights, and data generation based on the outcome.

When applying data science in your business, remember that data science is transformative. For it to fully empower your business, it isn't a question of just going out and hiring a couple of data scientists (if you can find them) and put them into a traditional software development department and expect miracles. For data science to thrive and generate full value, you need to be prepared to first transform your business into a data-driven organization.

Sorting Out the Concept of a Data-driven Organization

Data is the new black! Or the new oil! Or the new gold! Whatever you compare data to, it's probably true from a conceptual value perspective. As a society, we have now entered a new era of data and intelligent machines. And it isn't a passing trend or something that you can or should avoid. Instead, you should embrace it and ask yourself whether you understand enough about it to leverage it in your

business. Be open-minded and curious! Dare to ask yourself whether you truly understand what being data-driven means.


TIP

The concept of being data-driven is a cornerstone that you need to understand in order to correctly carry out any strategic work in data science, and it's addressed in several parts of this book. In this chapter, I try to give you a big-picture view of how to think and reason around the idea of being data-driven.

If you start by putting the ongoing changes happening in society into a wider context, it's a common understanding that we humans are now experiencing a fourth industrial revolution, driven by access to data and advanced technology. It's also referred to as the *digital revolution.* But be aware! Digitizing or digitalizing your business isn't the same as being data-driven.


REMEMBER

Digitization is a widely-used concept that basically refers to transitioning from analog to digital, like the conversion of data to a digital format. In relation to that, digitalization refers to making the digitized information work in your business.

The concept of digitalizing a business is sometimes mixed up with being data-driven. However, it's vital to remember that digitalizing the data isn't just a good thing to do — it's the foundation for enabling a data-driven enterprise. Without digitalization, you simply cannot become data-driven.

Approaching data-driven

In a data-driven organization, the starting point is data. It's truly the foundation of everything. But what does that actually mean? Well, being data-driven means that you need to be ready to take data seriously. And what does *that* mean? Well, in practice, it means that data is the starting point and you use data to analyze and understand what type of business you should be doing. You must take the outcome of the analysis seriously enough to be prepared to change your business models accordingly. You must be ready to trust and use the data to drive your business forward. It should be your main concern in the company. You need to become "data-obsessed."


TIP

Before I explain what it means to be data obsessed, consider how you're doing things today in your company. Is it somewhat data-driven? Or perhaps not at all? Where is the starting point in different business areas?

Figure 1-5 shows a model (with examples) for comparing a more traditional approach to a data-driven approach related to approaching different business aspects.

Business Question	Traditional business (starting point)	Data-driven business (starting point)
• What business shall we do?	Enhance current portfolio Build on what we have!	What is the data telling us? Data is the starting point for everything!
• Base for business decisions?	Experience based decisions What we know, is what we trust!	Data-driven insights/predictions We trust and use the insights for decisions!
• How to get things started?	Ownership and structure A defined setup is the way to get started!	Enablement (data, teams, infrastructure) Focus is to enable teams to find new solutions!
• Focus during execution?	Tools an system interaction It is about efficient tools and systems!	Data, models and algorithms Efficient data utilization, regardless of tools!
• How to track progress?	Reports Standard reports gives us what we need!	Data and measurements Study patterns and deviations in data and measurements!
• Increasing productivity?	Improve processes and tools We know how to be more efficient!	Utilize automation and AI Drive efficiency beyond human capabilities!

FIGURE 1-5: The difference between a traditional business and a data-driven business.

Comparing the approaches in a traditional business versus a data-driven organization is worthwhile. Many companies' leaders actually think that their companies are data-driven just because they collect and analyze data. But it's all about how data drives (or doesn't drive) the business priorities, decisions, and execution that tells you how data-driven your business really is. Understanding what the starting point is will help you define your ground zero and identify which areas need more attention in order to change.

Being data obsessed

So, what does the term *data-obsessed* actually mean? It's really quite simple: It means that you should always assume that the access and usage of data can improve your business – in *all* aspects. Use the following list of questions to determine how data-obsessed your organization actually is:

- Which data do you need to use as a company, based on your strategic objectives? Do you collect that data already? If not, how do you get it?
- Do you own all the data you need? If not, how can you secure legal rights to use it for your needs (internal efficiency or business opportunities)?
- Is the data geographically distributed across countries? If yes, what needs to happen to your infrastructure in order to enable you to use it efficiently?

- Is the data sensitive? That is, does it contain personal information? If yes, what are the applicable laws and regulations related to the data? (Be sure to note whether those laws and regulations change, depending on which country houses a specific data storage facility.) How do you intend to use sensitive data?
- Do you need access to the data in real-time to analyze and realize your use cases? If yes, what type of data architecture do you need?
- What data retention periods do you need to establish for the different types of data used by your organization? What will you use the selected data types for? Are you in control when it comes to expected data volumes and data storage costs per data type?
- Can you automate most of the data acquisition and data management activities? If yes, what is the best data architectural solution for that?
- Do you need to account for an exploratory development environment as well as an efficient and highly automated production environment in the same architecture? If yes, how will you realize that?
- Are employees ready to become data-driven? Have the potential, value, and scope of the change been clearly stated and communicated? If so, are employees ready for that change?
- Are managers and leaders on board with what it means to become data-driven? Do they fully understand what needs to change fundamentally? If so, are managers and leaders ready to start taking vital decisions based on data?

REMEMBER

The questions I post here don't comprise an exhaustive list, but they cover some of the main areas to address from a data-driven perspective. Notice that these questions don't cover anything related to using machine learning or artificial intelligence techniques. The reason that isn't covered is because, in practice, a company can be data-driven based only on data, analytics, and automation. However, companies that also effectively integrate the use of technologies like machine learning and artificial intelligence have a better foundation for responding to the machine-driven evolution in society.

Sorting Out the Concept of Machine Learning

People often ask me to explain the difference between advanced analytics and machine learning and to say when it is advisable to go for one approach or the other. I always start out by defining machine learning. *Machine learning* (ML) is

the scientific study of algorithms and statistical models that computer systems use to progressively improve their performance on a specific task. Machine learning algorithms build a mathematical model based on sample data, known as *training data*, in order to make predictions or decisions without being explicitly programmed to perform the task.

So, here's how advanced analytics and ML have some characteristics in common:

- Both advanced analytics and machine learning techniques are used for building and executing advanced mathematical and statistical models as well as building optimized models that can be used to predict events before they happen.
- Both methods use data to develop the models, and both require defined model policies.
- Automation can be used to run both analytics models and machine learning models after they're put into production.

What about the differences between advanced analytics and machine learning?

- There is a difference in who the actor is when creating your model. In an advanced analytics model, the actor is human; in a machine learning model, the actor is (obviously) a machine.
- There is also a difference in the model format. Analytics models are developed and deployed with the human-defined design, whereas ML models are dynamic and change design and approach as they're being trained by the data, optimizing the design along the way. Machine learning models can also be deployed as *dynamic,* which means that they continue to train, learn and optimize the design when exposed to real-life data and its live context.
- Another difference between analytical models and machine learning models regards the difference in how models are tested using data (for analytics) and trained using data (for machine learning). In analytics data is used to test that the defined outcome is achieved as expected, while in machine learning, the data is used to train the model to optimize its design depending on the nature of the data.
- Finally, the techniques and tools used to develop advanced analytics models and ML models differ. Machine learning modeling techniques are much more advanced and are built on other principles related to how the machine will learn to optimize the model performance.

Figure 1-6 shows how the different models can be developed, tested, or trained and then deployed. As you can see, analytics models are always developed and tested in a *static* manner, where the human actor decides which statistical methods to use and how to test the model using the defined sample data set in order to reach the optimal model performance. And, regardless of how much data (or which data) you push through an analytical model, it stays the same until the human actor decides to correct or evolve the model.

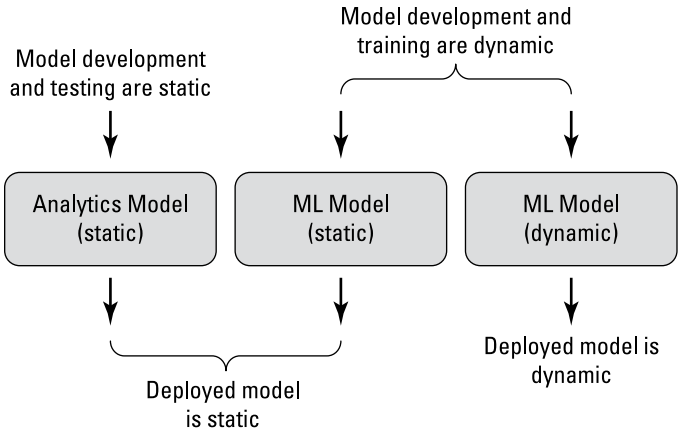

FIGURE 1-6: The difference in how development, training, and deployment are done for an analytical model versus a machine learning model.

TECHNICAL STUFF

In ML development, a human actor also decides which technique or method to be used. Training methods in ML differ depending on which technique is used — you can use supervised learning, for example, or unsupervised learning, semi-supervised learning, reinforcement learning, or even deep learning, which is a more complex method. It's even possible to combine two methods, like combining reinforcement learning with deep learning to what is referred to as deep reinforcement learning.

Instead of the static approach used in traditional model testing, with ML models you first train a model using a selected training data set that should represent the target environment where you intend to deploy the ML model. During the training, the model performance is tested to monitor the learning progress as well as measure the model accuracy. Within the scope of the chosen ML method, you then let the algorithm (machine actor) train itself on the training data set to reach the target that has been set. The machine then continues to train the ML model to evolve and find the most optimized model performance as long as you let it. The time will come when the model accuracy cannot be improved on using the training set. At that stage, you have to evaluate whether the model accuracy is good enough for deployment.

TECHNICAL STUFF

If you decide that a sufficient level of training has been reached by the machine actor, you need to decide how to deploy the model in the target environment, — deploy to production, in other words. You have two options at this point. You can decide that the model is sufficiently trained to achieve its purpose and that you can deploy it as a static model — meaning that it will no longer learn and optimize performance based on data, regardless of what changes occur in the target environment. Or, you can decide to deploy the ML model into production as a dynamic model, meaning that it will continue to evolve and optimize its performance driven by the data and behaviors that populate the model in the production environment. This is sometimes also referred to as *online training.*

So, when should you go for what type of model and deployment approach? Well, it depends on many factors. As a guiding rule, you should never use ML if you can get the job done using an analytics approach. Why? For the same reason you don't use a sledgehammer to drive a nail. You might perhaps succeed, but you can just as easily destroy the nail and hurt yourself, causing loss of time and money.

When it comes to a static or dynamic deployment, it depends on the business model and whether the target environment is static (changes happen seldom and are usually minor) or dynamic (changes occur often and on a large scale). If you're developing an algorithm to make online recommendations based on previous user behavior, for example, it's necessary to deploy a dynamic ML model; otherwise, you cannot fulfil your objective.

If, on the other hand, the purpose of the ML model is to let the machine find the optimal way to automate a set of complex tasks that you expect to stay the same over time, it is advisable to deploy the ML model as a static model in its target environment.

WARNING

Be aware that implementing ML models in live environments requires more resources from you. Machine learning training is complex and requires a lot of processing capacity as well as more monitoring of the ML model. You need to make sure that the ML model continues to perform as expected and doesn't degrade or deviate from its objective as part of its live training. Another aspect to consider is the need to ensure that the model can interact with other dynamic ML models in the target environment without disturbing each other's purpose or act in a way that leads to models canceling each other out. (What you're doing here is often referred to as *ensuring model interoperability.*)

Defining and Scoping a Data Science Strategy

To understand the constituent parts of a data science strategy as well the strategy's current and future significance, it's worthwhile to look at some of the major components on a high level. I then address each of these different parts in detail throughout this book. But before that I need to make a short clarification of the difference between a *data science strategy* and a *data strategy.*

On a high-level a data science strategy refers to the strategy you define with regards to the entire data science investment in our company. It includes areas such as overall data science objectives and strategic choices, regulatory strategies, data need, competences and skillsets, data architecture, as well as how to measure the outcome. The data strategy on the other hand, constitutes a subset of the data science strategy, and is focused on outlining the strategic direction directly related to the data. This includes areas such as data scope, data consent, legal, regulatory and ethical considerations, storage collection frequency, data storage retention periods, data management process and principles, and last, but not least; data governance.

Both strategies are needed in order to succeed with your data science investment and should complement each other in order to work. The details of how to define the data strategy is captured in part 2 of this book.

Objectives

If I ask about the objectives of a data science strategy, I'm asking whether there are clear company objectives set and agreed on for any of the investments made in data science. Are the objectives formulated in a way that makes them possible to execute and measure success by? If not, then the objectives need to be reformulated; this is a critically important starting point that must be completed properly in order to succeed down the line.

Data science is a new field that holds amazing opportunities for companies to drive a fundamental transformation, but it is complex and often not fully understood by top management. You should consider whether the executive team's understanding of data science is sufficient to set the right targets or whether they need to be educated and then guided in setting their target.


TIP

Whether you're a manager or an employee in a small or large company, if you want your company to succeed with its data science investment, don't sit and hope that the leadership of your company will understand what needs to be done. If you're knowledgeable in the area, make your voice heard or, if you aren't, don't hesitate to accept help from those who have experience in the field.


REMEMBER

If you decide to bring in external experts to assist you in your data science strategizing, be sure to read up on the area yourself first, so that you can judge the relevance of their recommendations for your business — the place where *you* are the expert.

Approach

Taking the right initial approach is a fundamental part of your data science strategy — it will determine whether your company takes the appropriate implementation and transformation approach for the data science investment. For example, is the approach ambitious enough — or is it *too* ambitious, considering time estimates related to available competence? Is there a clear business strategy and expected value that the data science strategy can relate to? Taking the time to think through the approach is sure to pay off because, if you don't know where you're going, you are most unlikely to end up there.

Choices

The term *choices* here refers to the strategic choices necessary to drive the data science transformation forward.


REMEMBER

The strategy you create cannot be about doing everything. It's equally important to make strategic choices about what to do as it is to make decisions about what *not* to do.

Decisions can also be distributed differently over time, because the choices can be about starting with a particular business area or set of customers, learning from that experience, and then continuing to include other areas or customers. The same strategy applies to choices of data categories or types to focus on early rather than later on as the company matures and capabilities expand.

Data

Defining a data strategy is a cornerstone of the data science strategy — it includes all aspects related to the data, such as whether or not you understand the various types of data you need to access in order to achieve your business objectives. Is the data available? How will you approach data management and data storage? Have you set priorities on the data? Have you identified and set data quality targets?


REMEMBER

Another important aspect of data relates to data governance and security. Data will be one of your most valuable assets going forward; how you treat it is fundamental to your company's success.

Legal

Understanding the legal implications for the data you need in terms of access rights, ownership, and usage models is vital. If you aren't on top of this aspect early on, you might find yourself in a situation where you cannot get hold of the data you need for your business without breaking the law, or, even if you can get hold of the data, you may realize that you cannot use it in the way you need in order to fulfil your business objectives.


REMEMBER

Laws and regulations related to data privacy stretch further than many people think, and they keep changing in order to protect people's data integrity. This is good from a privacy perspective, but doesn't always work well with data innovation. Therefore, as a good investment, you should always stay informed about laws and regulations related to the data needed for your business.

Ethics

Ethics, an area of growing importance, refers to the creation of clear ethical guidelines for how data science is approached in the company. Internally, this term refers to securing a responsible approach to data usage and management when it comes to preserving the data privacy of your customers or other stakeholders. One way of protecting privacy is through anonymizing personal information in the data sets.

Externally, insisting on the ethics of data science is vital when it comes to gaining your customers' trust in how you handle data. When machine learning or artificial intelligence is introduced — especially when automation of decisions and preventive actions are involved — it touches on another ethical perspective: the "explainability" of algorithms. It refers to the idea that it must be possible to explain a decision or action taken by a machine. Machine learning or artificial intelligence cannot become an automated, black box execution by a machine. Humans must stay in control to secure the transparency of AI algorithms and ensure that ethical boundaries are kept.

Competence

Based on the objectives that are set, choices that are made, and approach that is chosen, you must ensure that you put the right competence in place to execute on your targets. Putting together an experienced and competent data science team is easier said than done. Why is that? Well, you really need three main categories of competencies, and the availability of experienced data scientists in the market is now very low, simply because few data scientists have the sufficient experience and because the demand for these types of competencies is very high.

REMEMBER

You can't get by with simply hiring only data scientists. Data engineers with a genuine understanding of the data in focus is fundamental. Without good data management, data scientists cannot perform their algorithmic magic. It's as simple as that.

Finally, you need to secure domain expertise for the targeted area, whether it's a vast business understanding or an exceptional operational understanding. It's absolutely crucial to have the domain experts working closely with the data engineers and the data scientists to achieve productive data science teams in your organization.

Infrastructure

When talking about infrastructure, it's all about understanding what is needed in terms of data architecture and applications in order to enable a productive and innovative environment for your data science teams. It includes considering both a development environment (a workspace where you innovate, develop, train, and test new capabilities) and a production environment (a runtime environment where you deploy and run your solutions).

Infrastructure includes all aspects, from how you'll set up your data collection/data ingest, anonymization, data storage, data management, and application layer with tools for the analytics and ML/AI development and production environment.

WARNING

It is impossible to identify and set up the perfect environment, especially because the technology evolution in this area is moving very fast. However, a vital part of the infrastructure setup is to avoid getting locked into a situation where you become entirely dependent on a certain infrastructure vendor (hardware, software, or cloud, for example). I don't mean that you should only go for open source products, but I do mean that you have to think carefully which building blocks you're using and then make sure that they're exchangeable in the long run, if needed.

Governance and security

Working actively with data governance and security will make sure you stay in control of data usage at all times. It isn't important only in terms of gaining your customers' trust, but it is in many cases also a necessity for following the law. Keeping track of which data is collected, stored, and used for which use cases is a minimum requirement for most types of data.

WARNING

Overworking the area of governance and security will have an impact on your data science productivity and innovation. A common mistake is to be overprotective with regard to data usage, keeping all data locked in to a degree that nobody can access what they need in order to do their job. Therefore, you should approach the

setup of data governance and security with a mindset of openness when it comes to sharing data amongst employees within the organization. Lock the gates to outsiders, but strive for an open-data approach internally, boosting collaboration, reuse, and innovation.

Commercial/business models

As part of your company's data science strategy, you need to consider whether you only want to focus your efforts internally as a means of improving operational efficiency or whether you have ambitions to utilize data science to improve your commercial business models. Improving your business using data science will absolutely expand your possibilities, both in improving current business as well as helping you find new opportunities.

WARNING

Tread carefully when commercializing data. If you haven't transformed internally first by implementing data-driven operations, you'll likely be unable to fully leverage a data science approach externally in the business perspective.

That doesn't mean you need to implement and run data-driven operations throughout the company, but such operations will be needed for the areas connected to the new data-science-based business models and commercial offerings you're aiming to realize.

Measurements

Without measuring your success, how will you ever know whether you have actually achieved your objectives? Or be able to prove that. Still, many companies fail to think of measurements early on.

REMEMBER

Measurements are needed not only from an internal operational efficiency perspective but also to measure whether you have managed to deliver on the promises made to customers. This is important regardless of whether the agreed-on customer targets have been contracted or not. It should always be a priority for you to know how your business is performing against your objectives. The feedback will give you all the information you need to determine where the business stands, what needs to improve, and what has perhaps already been achieved.

Yes, establishing measurements early on is fundamental when it comes to securing continuous learning in your company, but it also shows customers that you care about reaching your targets. However, don't forget to think through the metrics structure you plan to use. It isn't an easy task to identify and define the correct set of metrics from the start. This is also something that needs to be reevaluated over time, based on which measurements actually give you the insights and feedback needed on what is going well — and what isn't going so well.

IN THIS CHAPTER

- Understanding why data science is inherently complex
- Realizing the potential in complexity
- Avoiding common pitfalls
- Managing complexity

Chapter 2
Considering the Inherent Complexity in Data Science

Cities are complex systems, and city policies are typically made in complex environments where many factors covering a whole spectrum of social, environmental, economic, and technological factors must be taken into consideration. However, in recent years, urban complexities have been better managed by evolutions in data science. The ability to perform urban modeling and simulate different future scenarios based on actual data has opened up many new possibilities related to urban planning and investments. These evolutions in data science have enabled government agencies to better understand complex urban issues, anticipate possible scenarios, and make the best policy and investment decisions.

But what does *complexity* really mean and refer to? Well, in my view, society has the general misconception that complexity is always bad. Yes, the simplest solution is often the best one — a truism that has been around ever since the 13th century, when the Franciscan friar William of Ockham came up with the original formulation, now known universally as *Occam's razor.* But in some cases, it's the

actual complexity of a matter that makes it interesting for a certain technical solution. This is the case in data science. If the problem at hand is simple and can be solved with a simple solution (using ordinary code, for example). it makes no sense to use machine learning techniques to solve the problem and throw self-learning software at it.


REMEMBER

In fact, if your business is simple and straightforward, and you want to keep it that way and not expand beyond your current business models, you might gain very little value by adding machine learning and artificial intelligence into the mix. However, if you're interested in evolving your business as well as your product or service capabilities beyond what is currently possible, more advanced data science could be a way to make that goal achievable.


WARNING

Nevertheless, the journey will not be simple. Data science is definitely an enabler, and you can use it to start simple and grow from there. But operating with data science at the core of your business requires skilled data scientists and data architects. Data science is a craft that requires skills across several disciplinary fields, including a good architectural understanding. It isn't a competence you acquire simply by taking a course in R or Python; it's much more than that, and you should not underestimate the level of expertise required. More advanced data science where machine learning and artificial intelligence is used is a complex matter that is used to solve complex problems. Therefore, this chapter aims to help you understand the fundamentals of why data science is complex — and also why the potential lies in that same complexity.

Diagnosing Complexity in Data Science

What does it mean when people say that data science is inherently complex? Well, by its very nature, data science — and especially techniques like machine learning and artificial intelligence — are built to solve complex problems that cannot be solved even by the brightest humans. It doesn't mean that the machine will outplay humans on day one, but over time it will — at least as far as we *want* the machines to outperform us, regulated by the policies we use to constrain the machines ability to enhance its learning. At the end of the day, it isn't about how smart machines can become, but rather how smart they can make us humans.

In machine learning algorithms are built to learn how to optimize their realizations first on a training data set in a lab environment and then on real life data later on. Algorithms can be built to learn from many different data sources and parameters, many more than is possible for the human brain to incorporate and process quickly and continuously. Let's face it; as long as they are running on a flexible and scalable architecture, machines need no sleep, no rest, and basically

have no limits in terms of capacity when it comes to bringing in more data or other policies and constraints. Humans just can't measure up to that.


REMEMBER

Automating a repetitive task in real-time using many different data sources doesn't necessarily have to be solved by machine learning. Many automation tasks can be carried out using a static statistical model; if the model doesn't need to change and optimize over time, then there's no need for machine learning. It really should only come into play when the problem at hand is dynamically changing and complex. Only then is the machine-learning algorithm needed to manage a complexity that the human brain cannot cope with (in real-time or not), improving the realization fast enough and in as many dimensions as required.


REMEMBER

Since data science serves as the foundation in managing the increasing complexity of our soon fully digitalized and connected society, it's at the core of the solutions needed for our technical evolution going forward. The attractiveness of data science is very much connected to the rapidly growing access to data and higher availability of technologies like machine learning and artificial intelligence. However, managing complexity is often not merely *difficult* to manage with a simple solution — sometimes it's *impossible.*

Given this fact, it's of vital importance to understand that, although data science is the scientific discipline that will be instrumental in bringing our society forward toward a future of more automation, robotics and self-learning software, it has never claimed to be based on simple science. Instead, data science is by nature inherently complex.

Recognizing Complexity as a Potential

If we assume that data science is complex, how can we turn that into a business potential? Well, due to the complex nature of data science, it's going to require skills that are not easily acquired and therefore not something that every company has on hand. Approaching it right, within the right time frame, could therefore be turned into a competitive advantage for your company.


TIP

Because data science is complex, it also means that in order to understand what it's about, you need to invest significant time and money to enhance your understanding of where you need to start, what it means for your business, and which business outcomes you could expect. One key component in getting this right, is to spend time on building a really good and useful Data Science Strategy.

WARNING

Rushing into an investment in data science without a clear objective or understanding of how the business must fundamentally change in order to capture the business value desired could produce the opposite result. There are many ways your investment in data science could go wrong. Some of these problems or pitfalls are more easy to avoid than others. Some are impossible to avoid, but can be managed through increased awareness of how to approach them.

Enrolling in Data Science Pitfalls 101

Part of coming to terms with the complexity of a solution is realizing that — despite our best thoughts and intentions — we are still drawn to simplistic solutions for complex problems. Data science solutions are no exception to this rule. In coming up with a data science strategy, you're bound to encounter many "reasonable" assertions that are in fact far from reasonable and could potentially endanger the success of your data science initiative. (I refer to these "reasonable" assertions as "pitfalls" because if you let them establish themselves, you and your data science initiatives will fall into a pit with no strategy for getting out.) You need to work constantly against the assumption that "the simplest solution is the best one" by stressing again and again that mastering complexity is really the only way you can ensure the success of any data science strategy. Some challenges can be avoided, whereas other challenges are unavoidable and need to be managed.

In order to help you on your endeavors, I'm going to walk you through an overview of some of the common pitfalls you need to avoid (and an explanation on why) in order for your data science strategy to succeed. A lot is won if you can focus your efforts on efficiently managing the challenges you cannot avoid.

Believing that all data is needed

Data voraciousness is a fault common to quite a few companies. They spend a lot of time and money on investing in infrastructure components so they can collect and store all the data available in a certain segment relevant to its business. Data is being acquired without strategically thinking through what is actually needed and when.

WARNING

What happens when you bring in all data? Time and money gets spent on getting the data, sorting it out, and making sure the infrastructure can cope with the huge amount of data being brought in. That means there's nothing left over for investing in the task of making use of the data.

It sometimes even gets to a point where so much effort is spent on managing the data that there is barely any time left over for producing the insights the data was meant to produce. And on top of that, the insights that are derived from the data are often never put into action — the focus is elsewhere, stuck on managing the overload of data coming into the company.

Thinking that investing in a data lake will solve all your problems

Many companies have spent a considerable amount of time and money investing in *data lakes*, believing that by replacing the scattered data repositories (usually spread across various applications and traditional database systems) with a new and common data repository (usually in the cloud), all problems are solved. But you need to be careful so that you don't see the data lake as the silver bullet — that part of your infrastructure that will solve any problem. Please be aware that the data lake should be seen as a temporary storage point for your data, not a permanent one. Remember that it adds value only as long as the data stored in it is used. Of course, a company may have other reasons for storing data — regulations that require storing data for a certain time period, for example. Keep in mind however, that the data lake should primarily be seen as a layer in your infrastructure that should be focused on enabling secure and efficient data usage by the next layer.


WARNING

Avoid thinking of the data lake as a "warehouse" where you throw in all the data you've collected and lock the door for usage only by an unknown individual, for an unknown purpose, later at some unclear point in the future. (It's called a data *lake*, not a data *abyss*.) Think ahead instead; it's of vital importance that you clearly define in your data strategy which data will be stored where, for which purpose, and with which priority. You also need to think about how long you want your retention periods for each different type of data to be, based on what you're aiming to achieve with the data.

Another important aspect to regard strategically involves the costs associated with data storage. If you're collecting huge amounts of data on a regular or real-time basis, which means you anticipate a constant flow of new data coming in that will grow the total data volume over time, you can expect an exponential growth in data storage costs over the short and long term.


TIP

Before data is even put into the data lake, you also need to consider how to structure the data lake so that you'll be able to find data quickly and efficiently. You need to separate between sensitive and non-sensitive data, as well as between data you own versus data that you do not own yourself but have the rights to use. It's also very important to think through the data access rights internally from a data governance perspective. Perhaps not everyone should have access to everything? Just be careful in that regard. Don't overdo it in terms of restricting

data access within your company. Locking your data in just to be on the safe side is not a good idea, since it will decrease your data lake utilization efficiency. Restrict only what is absolutely necessary from a legal or company policy perspective with regards to data privacy, restricted customer data, financial data, or other sensitive data.


REMEMBER

Without this basic data lake structure with data categorization, tagging, defined retention periods, access rights and so on, you run the risk of having loads of data at your disposal, but cannot use it efficiently, because it is lost or locked in the lake.

Focusing on AI when analytics is enough

The ambition to stay in tune with the evolution in the industry at any cost, is another typical pitfall that can be found in a growing number of companies. This derives from the fact that companies want to stay in tune with the latest technology evolution in the market but lack a real understanding of what it actually means. When it comes to artificial intelligence (AI), there is a misplaced confidence in what AI is and what it can do, but at the same time there is an underestimation of what analytics can do without adding the complexity of AI.


REMEMBER

Saying that AI is overly hyped doesn't mean that AI cannot be relevant. On the contrary, AI can most probably enhance most businesses in many aspects. However, you should not try to solve a simple problem by using a complex technology like AI if it can, in fact, be solved using analytics.

Analytics helps you explore the data using different techniques, enabling you to find dependencies and correlations as well as build models to forecast or predict a certain outcome or behavior. Analytics falls short when one of these factors exists:

- The problems are too complex for humans to understand and design an optimized solution for.
- The data environment where the algorithm needs to run is dynamic and constantly changing.

In these situations, you need something else. An analytics model put into production has a fixed design; its behavior is static and cannot adapt or change, meaning that it will stay the same over time, even if data and conditions change.

Rather than jump immediately on the AI bandwagon, then, take the time to think through which type of environment you're targeting with your solutions and in which context you need these solutions to work. Ask yourself which ones are more static, and thus more likely to stay the same over time in terms of data and

behavior, and which are constantly changing. Understanding that to a certain level of detail, will help you to get a better overview of what approach to use for which problem.

Believing in the 1-tool approach

Many companies are of the opinion that a harmonized tool approach offers the most efficient IT environment. And perhaps that is many times the case, especially when you want to drive toward aligned ways of working in the company, harmonizing data input, and so on. But, when it comes to data science, you must consider which parts needs to conform, and which are more efficient if they stay diverse and flexible.


REMEMBER

As a general approach, strive for as much alignment in the basic layers of capture, storage, and management of data. But when you approach the upper layers of analyzing and communicating insights and decisions, you need to allow a much higher level of freedom for data scientists, business analysts, and other interested parties. When approaching data exploration and analysis as well as algorithm development, you need different techniques and tools to be available for your teams.

The same is true when it comes to communicating or utilizing results from the analysis: You need tailored approaches that align best with the information you want to convey. Forcing this to happen in one and the same environment will hinder, rather than boost, innovation and will definitely limit the impact of data science on your overall business. For a data science strategy to succeed, it needs to be integrated into all necessary aspects of your company. And for that you require a variety of tools and applications for different purposes.


TIP

Many companies' leaders tend to view the data science setup from a cost perspective, and often believe that the main cost for the environment is connected to the application layer rather than to the enablement part of the infrastructure — capture, storage, and management of data, in other words. Aligning and optimizing the basic layers correctly in your infrastructure will not only give you cost control of your data science infrastructure investment but also allow the data science teams greater freedom in the layers on top. With this approach, you have a greater chance to enhance your data science productivity overall.

Investing only in certain areas

Leaders of larger companies often tend to think that you can select one or two areas for the data science investment rather than go for a full end-to-end implementation across the company. It's understandable because such an implementation is not only costly but also fundamentally transformative in terms of how

tasks are approached and executed. Of course, change on such a massive scale is seen as a major risk from a company-wide perspective.


REMEMBER

To truly benefit from your data science investment, you need to approach it from an end-to-end perspective. As long as you take the time to think through your investment from a long-term perspective, it's not only possible but also advisable to first start small and then grow over time, business area by business area. To achieve that, however, you need a plan to incorporate the business folks in the data science investment. All parts need to transform over time — and that transformation may reach much further than you think. If your company is large, you might even have to consider transforming your relationship with your sub-suppliers and vendors. If your business becomes data- and value-driven in all aspects, can you then really work with a sub-supplier that is cost-driven?


WARNING

When planning to take small steps toward having a company that's fully focused on data science, you cannot count on that approach being the best one from a cost perspective. Instead, it's more likely than not that until data science is implemented as a driving force throughout the company, you will see only minor benefits from a company-wide perspective. Remember that having only parts of the organization become data-driven could even increase your overall cost short term, because it means that you need to maintain two or more types of setups (infrastructure, processes, competences, and so on) in parallel.

Leveraging the infrastructure for reporting rather than exploration

The common problem of being report-focused is typically connected to a situation where top management has the wrong idea about what data science can bring to the company. The situation usually arises because some company leaders believe that the main purpose of data science is to produce a set of answers to certain predefined questions raised by management. Answering these specific questions should therefore be the main driver for the implementation of data science and should thus result in a report back to management.


WARNING

You might be asking yourself what's wrong with trying to fulfill requests coming from Corporate — is not the starting point of all analysis a set of business questions you want answers to? Well, in a sense, that is correct. However, it's equally important to be mindful that the questions you're asking might not be the right ones to ask. Why? Simply because that predetermined set of questions is based on your current understanding of your business, market, and customer base. If your company is mainly experience-based rather than data-driven, the questions might be correct — or not. You simply do not know if you're approaching a certain problem or opportunity from the wrong angle.

TIP

Approach your data science investment as an opportunity for your company to be based on data, insights, and facts that will help guide you correctly in a data centric society. Like in the company Husqvarna that is dealing with outdoor power products, they have now started to enable connectivity for their chain saws. The company is doing this in order to collect data on how they are used, or not used, when cutting trees, to be able to understand more about their own business. Simply exploring the data for patterns or anomalies that might point to new (and perhaps unexpected) questions worth asking, is a good way to start.

Underestimating the need for skilled data scientists

Becoming a data scientist is an acquired competence; it can be learned, in other words, with the help of books and training courses and workshops. Becoming an *experienced* data scientist, however, takes time and requires certain skills that are not as easy to acquire.

It's important to respect the difference between a basic data scientist and an experienced one. It's equally important to realize that the senior ones are difficult to come by, so if you have some in your company, make sure you hold on to them. A senior data scientist in the AI space is someone who has worked in the area between five to ten years, knows several programming languages, but most importantly is very skilled at using various ML/AI techniques when building algorithms. To be seen as senior, it also includes having experience from developing and deploying algorithms based on various use cases and in different types of target environments.

However, the key to creating successful data science teams does not lie in acquiring as many senior data scientists as possible. In fact, it's better to have fewer senior ones and spread them across many teams, which will allow them to function as mentors for more junior data scientists, hence contributing to the overall company data science maturity in a better way.

REMEMBER

Although senior data scientists are expensive to hire, it's worth thinking of them in terms of the contributions you can expect from them. If supported by domain experts, senior data scientists can work across any disciplinary field and, given the right preconditions in terms of data and a capable infrastructure, help you approach basically any problem or opportunity in an efficient and innovative way.

`Navigating the Complexity

Arming yourself with persuasive arguments designed to counter those advocates of the "simple is better" philosophy at your company is a good starting point. Recognizing any and all of the challenges that may arise on your company's journey to becoming fully data-driven is crucial, but just being aware of the challenges doesn't rid you of them automatically. It requires not only a constant awareness of the necessity of not thinking about things the wrong way but also a strategic mindset — and plan — to navigate around these potential problems as they arise.

Taking the time to study up on the different scenarios and the solutions to them is worthwhile. When they occur (and they are certain to occur), you will already have a level of understanding on how to deal with them. Even better, given what you know, you can act proactively to make sure you never end up in one of these less-than-favorable situations for your business.


TIP

Write up your identified risk list and proposed mitigation plan for all scenarios, and add it to the company's data science strategy so that you have an agreed-on view of what to do — and what not to do — when stuff happens.

IN THIS CHAPTER

- Handling data in motion
- Driving productivity through data consistency
- Ensuring explainability in artificial intelligence
- Elaborating on the difference between software development and machine learning
- Dealing with the rapid technology evolution

Chapter 3
Dealing with Difficult Challenges

This chapter addresses a number of complex challenges that are difficult to avoid and that will require the right set of tactics to manage successfully. More specifically, I'll show you what you need to do to make the right decisions when it comes to acquiring and managing your data efficiently and consistently, setting up your data science environment, managing the legal constraints related to the data and algorithms you need for your business, as well as preparing for rapid evolutions in the area of data science as a whole that is sure to come.

Getting Data from There to Here

When a company decides to embark on a journey to become data driven, the focus is naturally on the data itself, which inevitably leads to a greater awareness of the actual variety of data needed to gain full proactive and data-driven control of their current business. On top of that, companies soon realize that in order to expand

beyond what is possible today, the data sets need to become even more varied. At this point, many companies start to realize that the data which is fundamental to becoming truly data driven might actually belong to someone else or is located in another country, with other data regulations. This section explains how to strategically approach such practical challenges as part of your data acquisition.

Handling dependencies on data owned by others

Dealing with proprietary data is an unavoidable yet manageable challenge faced by any company striving toward becoming fully data-driven. Typically, what happens is that you have identified and carefully specified all the data you need in your data strategy and when you then start looking into how to strategically approach capturing the data, you realize that you have a data ownership problem.


REMEMBER

If you use only data generated from your internal IT environment, you have, of course, less of a problem. If that's the case, however, then your company probably isn't truly data-driven in the proper sense. A data-driven business accounts for how its products and/or services are used and how it performs in real-life settings, not merely in the lab environment. And anytime you start using data generated by life in the real world, you run into the data ownership problem.

What kind of data am I talking about? First and foremost, this involves data owned by your customers, but it can also include data owned by your customers' customers, depending on which business you're in. You have to take the time to truly understand the detailed context of the data you need. It can relate to issues of data privacy, but it doesn't have to. It can simply be the case that the data you need in order to better understand your business performance or potential belongs to someone else.


TIP

Don't get discouraged when it comes to ownership issues. Most situations can be solved from a legal perspective if you're willing to address them openly with the data owners, explaining why you need the data and how you will treat the data after it's in your possession. It's all about gaining trust with regard to how, and for what purpose, the data will be used. (It wouldn't hurt to also spell out how your work may, if possible, contribute back to the owners of the data.)

At the end of the day, you need to be absolutely certain that you understand (and are complying with) the legal constraints that apply for each different type of data you intend to use. Your use of the data must also be regulated by way of a contractual setup with the party owning the data, including what rights your company has related to data access, storage, and usage over time.

REMEMBER

Laws and regulations have a habit of changing over time. Lately, the trend is to increase restrictions even further in order to protect an individual's right to their own data. One recent example is the quite restrictive General Data Protection and Regulation (GDPR) enacted by the European Union (EU) that went into effect in May 2018. Given recent news of the misuse of data by entities such as Cambridge Analytica and Facebook, the U.S. and Canada are definitely looking into legislation similar to the EU's GDPR.

Anything that helps to protect an individual's right to privacy is all for the best, but just remember that the way you deal with privacy legislation today will most probably be quite different in the near future. Therefore, you should strategically and proactively think through your infrastructure setup and your data needs to ensure that you account for these types of constraints in your current and evolving data science environment.

Managing data transfer and computation across-country borders

If your company has divisions in a number of different countries or does business (and therefore has many customers) in many countries, one major challenge you might face is how to manage data that needs to cross international borders.

You need to carefully consider a number of different aspects of the data puzzle if your company has an international component. Here's a list of the major concerns:

- **Legality:** Legal constraints to moving data across borders is a consideration that a company must stay on top of. Laws and regulations differ from country to country, so different solutions may be possible, depending on which country you're doing business in. The restrictions are also different depending on which type of data you're moving out of a country. Data with personal information is usually much more difficult to move than non-sensitive data. Breaking laws related to data transfer can be quite costly and can severely impact the company brand if it is determined that you violated customer trust.
- **Data transfer approach:** This refers to how you actually execute the data transfer. It's typically quite costly and also differs from country to country. Depending on the volume of data you want transferred, and the data transfer frequency, you can either rent space in existing connectivity infrastructures and data links or — if you cannot get your requirements met regarding aspects such as capacity, security, or exclusivity — invest in your own links.

- **Possibilities for local computation and storage:** If you can store the data and carry out the analysis in the country where the data has been captured, you might be able to lower the cost and increase the speed of delivery. However, to get this setup to work efficiently, you need to properly think through what your distributed computational architecture will look like. What will be done where? and where will the source data be kept, for example? Will there be a central point of data storage and global analysis, or will there be only distributed setups? How you answer these questions depends a lot on what type of business is being conducted and what the setup looks like in different countries.

Managing Data Consistency Across the Data Science Environment

It might seem like a simple task to ensure data consistency across the different parts of the data science environment, but it's much more difficult than it seems. First off, this area tends to be more complex than it needs to be, eating up more time and resources than originally estimated. The need for consistency includes aspects such as data governance and data formats, but also the labeling of data consistently —using customer IDs across many different sources to enable correlation of different data types related to the same customer, for example.

The challenge is that there is a built-in contradiction in terms infrastructure between enabling usage of special tools to allow data scientists and data engineers to be innovative and productive and at the same time ensuring consistency in the data. This is because specialized tools are optimized to focus on solving certain problems but either don't keep the format consistent or don't interface well with other tools needed in the end-to-end flow. Optimized, specialized machine learning tools are simply not good at playing together with other, similar specialized tools that are addressing comparable or other adjacent problems.

REMEMBER

But is it really that bad? Well, it can lead to real problems, depending on how much freedom is allowed in the architectural implementation and among the teams. Some examples of problems that can stem from a lack of consistency across the AI environment are described in this list:

- **Ad hoc solutions:** Every case is treated as an isolated problem that needs to be solved *this instant* in order for the team to move forward. The result? No long-term solution and no learning between teams.

- **Increased cost:** When you have to duplicate tool capabilities in order to manage a lack of consistency or when you have to build capabilities into purchased tools to secure just the basic consistency, those costs add up.
- **End-to-end not working:** Inconsistencies can occur when the infrastructure is implemented across several cloud vendors, which then makes it difficult or impossible to transfer data and keep data consistent across different virtualized environments.


REMEMBER

Because corporate management cannot enforce, and may not want to enforce, data consistency across the organization as a company policy, they have to use other means to preserve data consistency end-to-end. One way is to ensure that all teams follow proper and relevant guidelines for evaluating and purchasing new tools that incorporate specific directives related to data consistency. Clearly motivating why this is key to a successful data science strategy execution.


WARNING

It's also vital to consider which limits are needed for each individual company, depending on the type of business, their objectives, and so on. Hold the line when it comes to data consistency: Otherwise, you may end up with a cumbersome and costly implementation of data science, one far removed from the productive data science environment you were hoping for.

Securing Explainability in AI

Explainable AI (XAI), also referred to as Transparent AI, involves the ability to explain how an algorithm has reached a particular insight or conclusion that results in a certain decision to take action. Though an important aspect to consider as part of the evolution of AI, it isn't easy to solve technically, especially if the AI is acting in real-time and thus using streaming data that hasn't been stored. To bring this point home, imagine, if you will, that you cannot explain to your customer why the machine made a certain decision — a decision you would not have made based on your own experience. What do you tell the customer then?


REMEMBER

Addressing explainable AI is becoming increasingly important in terms of our human ability to understand more about why and how the AI is performing in a certain way. In other words, what can be understood by studying how the machine is learning by processing these huge amounts of data from many dimensions, looking for certain patterns or deviations? What is it that the machine detects and understands that you missed or interpreted differently or simply were not capable of detecting? Which conclusions can be drawn from that?

WARNING

Ethically, AI explainability will be even more important when data scientists start building more advanced artificial intelligence, where many different algorithms are working together. It will be the key to understanding exactly *what* machines interpret as well as *how* the machine's decision-making process is carried out. Knowing this information is crucial to staying on top of the policy framework needed to set the boundaries for what the machine shall and shall not do, as well as how these policies need to be expanded, or perhaps restricted, going forward.

From a purely existential perspective on one hand and the need for humans to remain in control of the intelligent machines that are being built on the other, you cannot simply view AI as black box. (The *black box* challenge in AI refers to the need to ensure that, when an algorithm takes a decision based on the techniques that have been used to train the algorithm, that decision-making process must be transparent to humans. Algorithm transparency is possible when many of the more basic ML techniques — supervised learning, for example — are being used, but so far nobody has yet found a way to gain transparency when it comes to algorithms based on deep learning techniques. For example, there must be a way to explain why a certain decision was taken when something went wrong. A pertinent example is the self-driving car, where a bunch of algorithms are in play, working together and (hopefully) following policies predefined for how to act in certain circumstances. All works according to plan, but then a totally unknown and unexpected event occurs and the car takes an unexpected action that causes an accident. In such situations, people in general would naturally expect that there would be some way to extract information from the self-driving car on why this specific decision was made — hence, they expect *explainability* in AI.

Apart from the technical, ethical, and existential reasons for ensuring the explainability of AI, there is now also a legal reason. The EU's General Data Protection Regulation (GDPR) has a clause that requests algorithmic interpretability. Right now, these demands aren't too strict, but over time this will likely change dramatically. The GDPR request now requires the ability to explain how the algorithm functions based on the following questions:

- Which data is used?
- Which logic is used in the algorithm?
- What process is used?
- What is the impact of the decision made by the algorithm?

Dealing with the Difference between Machine Learning and Traditional Software Programming

It is quite well established and commonly agreed in the software industry on what the actual difference is between traditional programming and machine learning. However, when it comes to how this difference should be handled, there's little agreement. Given this division, I want to take the time to explain what to consider when it comes to your implementation approaches as well as how to deal with these differing viewpoints in terms of development aspects as well as the production environment. But first let me start you off by looking at what the argument's all about.

The traditional programming approach, shown in Figure 3-1, has you decide beforehand how to solve a certain problem by using the program being developed. The main target for the software developer is to build the requested functionality.

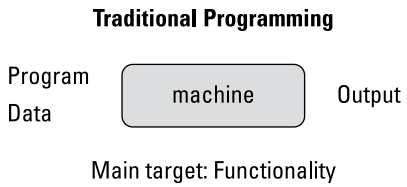

FIGURE 3-1: The traditional programming approach.

Based on the data and the program, the machine performs the analysis exactly the way you want, regardless of whether it's the most optimized way to solve the problem. The assumption is that you-the-programmer (rather than the machine) know best how to solve the problem.

On the other hand, when it comes to machine learning development, the starting point is to empower the machine to find the best solution when you set the boundaries of which data to use and which outcome to achieve — and nothing more. (See Figure 3-2.) The assumption is that, given these conditions, the machine will find the most optimized program to solve the problem.

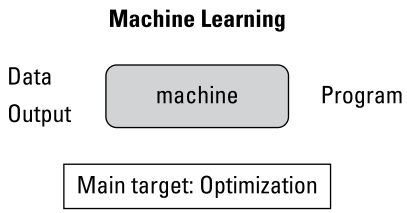

FIGURE 3-2: A machine learning approach.

So, what do these distinct approaches mean in terms of your development and production environments? One main aspect to consider is that traditional programming embraces a much stricter process. It's rule-based and follows predefined design principles. The starting point for machine learning development, on the other hand, is much more explorative and open-ended. As you might have guessed, this will have quite a significant impact on how the development environment needs to be set up.

WARNING

Some companies have a tendency to downplay the impact of the development environment setup and which impact it will have on data science productivity. If you start from this vantage point, you may well conclude that you can use the same (or similar) infrastructure setup for both your traditional software development environment and your data science environment. Nothing could be further from the truth — taking such an approach means that you're setting up major barriers toward achieving your goal of useful artificial intelligence/machine learning output.

Traditional programming is much more restrictive when it comes to which programming languages to use and which principles to apply for what task. This, of course, impacts how both the development and production environments need to be set up. Figure 3-3 gives you a graphical representation of how traditional programming happens.

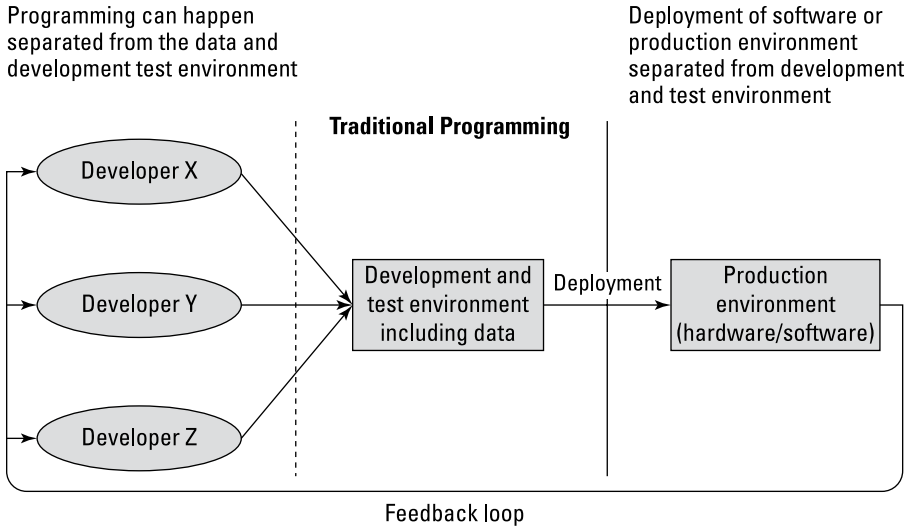

FIGURE 3-3: The traditional programming flow.

As you can see on the left side of Figure 3-3, traditional software programming can happen separately from both data and the development and test environment. It doesn't have to happen separately, but the fact is that it can be done in isolation — even on a laptop in a coffee shop — and then integrated with other code in the development and test environment. At this point, data can be added to the model in order to achieve the desired output.

Figure 3-3 also shows that the deployment of the software program is done in a separate environment (into a software/hardware product or similar production environment, for example) outside the development and test environment.

Turning once again to a machine learning approach, you need to recognize that the explorative and learning nature of machine learning development requires the setup that's available to your data scientists to be extremely flexible. Efficient data management, easy data access, and a variety of specialized machine learning tools must be easily available. Nobody walks into the process with predefined notions of exactly which machine learning technique to use, because all that needs to be explored and because the most optimized solution may become clear only after the process has started.

REMEMBER

As Figure 3-4 shows, machine learning development cannot happen in isolation and without the data. It all starts and ends with the data in a machine learning development flow because the data itself is what trains the model for an optimized design. For this to work, you obviously have to have a constant data flow, which means that you need a stable data pipeline — preferably, a virtualized one that offers more infrastructure flexibility over time.

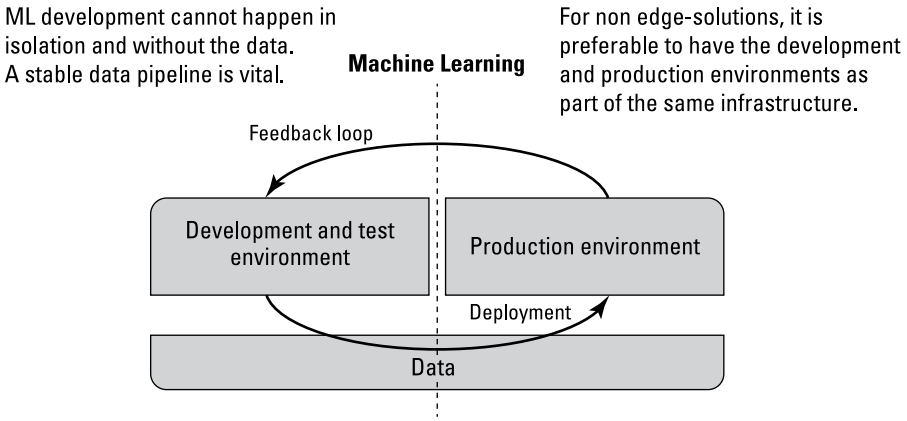

FIGURE 3-4: A machine learning flow.

TIP

For machine learning virtualized production environments not implemented on the edge (inside IoT devices like a mobile phone, a car, a watch, a fridge, or other types of devices that are connected and where a ML algorithm can run, in other words), try to keep your development and production environment close or as part of the same infrastructure setup. This facilitates machine learning productivity when moving between development and production, with faster and more efficient feedback loops as part of the benefits. You also gain a cost efficiency benefit when you don't need to duplicate the infrastructure, because both are built on the same data pipeline.

Managing the Rapid AI Technology Evolution and Lack of Standardization

AI/ML technologies are constantly evolving and becoming more and more advanced. As computational efficiency increases, such technologies can now adjust to run on a smaller hardware footprint. These advances push analytics, ML, and AI realization also to the edge, meaning that an algorithm has computational support to run inside a device rather than that the device just provides data to the algorithm running remotely in the cloud, for example. That's a good trend, because it will allow society to utilize machine intelligence to a broader extent across system environments and billions of mobile devices and other connected entities. However, one area is not keeping up with all the rapid changes: ML/AI standardization.

The lack of standardization isn't, of course, something that you or a single company can solve, but it's important to be aware of this situation as part of your data science strategy. And of course, at the end of the day, data scientists all have the responsibility to strive toward more standardization in machine learning and artificial intelligence. But just because no official, international standardization exists yet, it doesn't mean that no initiatives exist. The standardizations that are out there are mostly based on something often referred to as *de facto* standardization, derived from influential open source initiatives coordinated through universities like UC–Berkeley (AMP lab and RISE lab) and companies like Google (Google Beam) and AT&T (Acumos).

Another trend that can be detected deals with increased concerns when it comes to access to (and usage of) personal information for nontransparent or even hidden reasons. This has resulted in stricter legislation in different countries, but has also led to more ongoing discussions about the need to increase regulation and impose standardizations related to AI ethics. This positive trend will hopefully continue to push human society to better envision — as a group — what the future of AI utilization should look like.

Of course, this trend has a downside. Because so little standardization is now available to lean on for your data science investment, you need to account for the possibility that you will have to make serious adjustments to your infrastructure when the new standards finally pop up in the near future. The worst-case scenario? You may have to repeat his process several times, or even totally remodel your entire infrastructure.

My advice to you? Continuously follow trends, and be on the lookout for any indication that new laws or regulations or standardization initiatives are coming down the pike — especially open source ones. And be prepared to adjust your data science approach to match the changes that are coming.

IN THIS CHAPTER

- Exploring why change in data science is different
- Approaching ways to manage change in data science
- Listing things to avoid in change management
- Utilizing various change techniques
- Guiding steps to start your change journey

Chapter 4

Managing Change in Data Science

Investing in data science and a data-driven approach means understanding and dealing with the change that needs to happen. Although the inevitable data science transformation in society may not have fully arrived yet, organizations still need to get ready. The time for standing on the sidelines, waiting to see what other companies are doing, is over. The time to act is now.

Those companies best positioned to manage the needed change driven by data science in the next decade will be the ones that start preparing now. The day has come for companies to invest time in strategically building up an understanding of what is needed and capture the intent in a data science strategy — not just in one area or function, but throughout the company.

Understanding Change Management in Data Science

In a study done by PricewaterhouseCoopers (PwC) and Iron Mountain, 1,800 senior business leaders in North America and Europe at midsize companies and enterprise-level organizations responded to a survey which showed that only a small percentage of the companies actually considered themselves to have effective data management practices.

The study found that although 75 percent of business leaders from companies of all sizes, locations, and sectors feel that they are "making the most of their information assets," in reality, only a minor portion seem to be strategically approaching these major changes in the right way. Overall, as much as 43 percent of company leaders answered that they "obtain little tangible benefit from their information," and 23 percent "derive no benefit whatsoever," according to the study. So, what are companies doing wrong?

One lesson to draw from the survey is that investing in the technology to become data driven is only the beginning. To ensure success, companies must do much more than focus on the tools needed to manage the data. Data science transformation deals with sophisticated and interconnected data, small as well as big data sets, which impacts a whole range of business operations and has implications on people, cultures, organizations, processes, and skill sets in data science. The glue that connects and holds all these elements all together is the people. And the key is to get people motivated. This can be achieved in many ways, but using data to communicate relevant examples and proof points in combination with firm leadership is a good way to start.

REMEMBER

Strong leadership to drive the change includes not only the line management support but is also very much dependent on strong leaders and change drivers who can generate trust that the change will bring results. Without these dedicated change drivers across the company, it does not matter if you have the perfect plan — this type of totally transformative change will not happen, at least not to its full extent.

In data science, the methods and techniques used for everything from knowing how to capture and process data to building models and deriving insights continue to evolve, creating a constant need to manage change. This change is also happening in areas such as regulatory practices, security, and privacy, continually altering the base and framework for how to approach data science. For a data science strategy to succeed, organizations need to understand and accept the fact that the skill sets needed to handle different aspects of data science will continue to change. To manage this continual change, you have to have an open mind and

be willing to leverage and explore new technologies and methodologies as they become available.

In practice, this means that individuals need to adopt a data-driven mindset and a commitment to lifelong learning as an extension of their work if they ever hope to manage change. Only when you actively use data to explore new avenues and solve real problems can you justify the data science investment.


TIP

Defining a relevant and applicable process for change management should be a joint organizational effort, approached through brainstorming and idea refinement. Usually, agreeing that change is needed is easier than deciding how change should be approached.

Approaching Change in Data Science

Managing change effectively is a multistep process that requires significant investments of time and money. I can recommend a generic change management approach for you to follow, but you also have to consider some specific characteristics. Figure 4-1 graphically illustrates what has to happen, and the next few sections describe in detail the recommended steps.

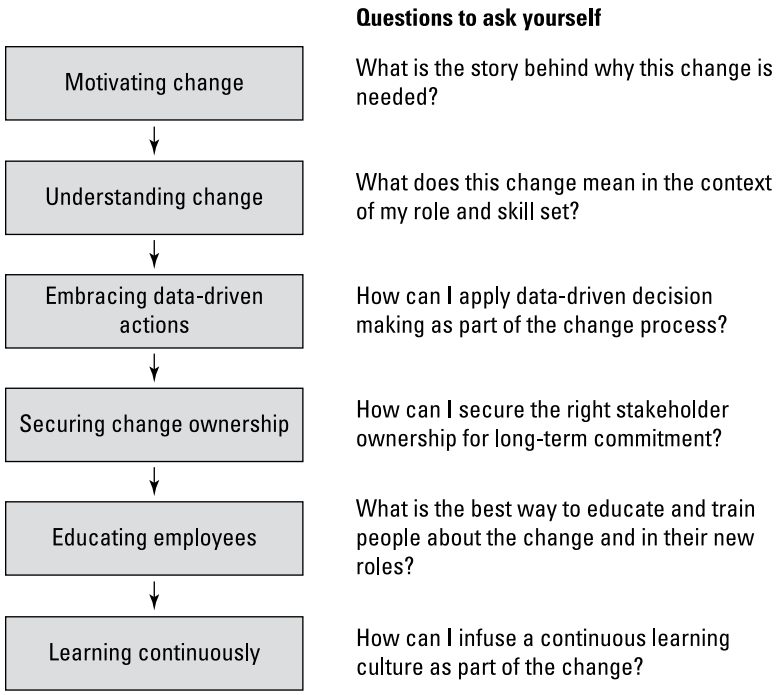

FIGURE 4-1: Driving change in data science.

Motivating change

Creating a convincing case for change is a necessary starting point. This compelling story should define what the data science investment will enable for the company and the organization in relation to not only internal policies, processes, and employees but also competitors and customers. By relying on a story-based approach, one which uses relevant business examples as part of your argument, you will be able to help your organizations understand the full impact of the changes coming its way at the very beginning of the process.

REMEMBER

To be able to clearly motivate change, an organization must have a thorough understanding of what each change will mean and where the changes will occur across the spectrum of business and IT operations.

Understanding change

The next step in data-driven readiness is to define the changes in operational terms in a way that employees can relate to. This includes aspects such as explaining the purpose of the change or how upcoming changes might impact structure, processes, skills, and performance goals.

WARNING

Change is never easy. Employees will be exposed to new roles, capabilities, competencies, and ways of working, so the way that companies prepare employees for this fundamental change is critical. First and foremost, you need to focus on educating employees with relevant role-based information and preparing them to be data evangelists in the organization. This personalized approach to making change real and meaningful drives the readiness that is needed for introducing data science successfully.

Embracing data-driven actions

Data science is creating a cultural shift — one that is most evident when it comes to how decisions are made. In data science, decision-making leverages a data-driven approach much more so than approaches relying on experience or gut feeling. It also assumes a culture of collaboration in the organization, because only by working together can people across the organization discover the full value of the insights in a relevant business context that support a permanent change towards data-driven decisions and action.

A reliance on agile methods and DevOps (*dev*elopment *op*erations) teams is quickly becoming best practice when managing data science transformations. In an agile approach, an organization empowers its people to work where, when, and how they choose, with maximum flexibility and minimum constraints in order to optimize their performance and deliver best-in-class value and customer service.

A DevOps team approach combines software development (Dev) with information technology operations (Ops). The goal of DevOps is to shorten the development life cycle while delivering frequently in close alignment with business objectives.

These team-related changes also add a layer of complexity for employees, where the traditional walls between organizational teams are displaced to form collaborative teams. Organizations therefore need their leaders to

- Embrace change willingly
- Communicate with employees about the changes that are happening
- Take some time to listen and learn from employees


REMEMBER

In the world of data science, building temporary cross-functional teams like task forces isn't enough to solve complex business problems or build innovative solutions: Organizations must be willing to foster informal groups where individuals are encouraged to seek and uncover hidden opportunities or problems they can address. For leadership, it's equally important to acknowledge the contributions made by such groups in order to empower them and sustain them for the long term.

Securing change ownership


TIP

Without question, the best way to manage the complexity of transformation is to create ownership among the stakeholders who will ultimately deliver on the promise of the new technologies and capabilities.

The idea of appointing traditional change leaders is old-school. Instead, a new innovative model recognizes that business leaders that take on a more operational role as part of the change are the most trusted sources of information and credibility within an organization, and thus should deploy the new technology and own the change as such.


TIP

Create a story that leaders can embrace. One way to do this is by effectively using targeted workshops that demonstrate how the anticipated changes will significantly improve business processes, systems, and practices across different business segments. In these workshops, you can enable early adopters among leadership to work collaboratively with other stakeholders throughout the company.

By using this expanding model for managing data science change, you can touch all the stakeholders you'll need in order to deliver on the promise of data — and also reduce the risk of employees feeling demotivated and alienated by the change.

Educating employees

I recommend blending the necessary education and training programs with elements other than the standard skills training that will (obviously) be needed. Place at the top of your list areas like psychology, gaming, and communication. Adding these elements to the mix helps to focus employee learning and development efforts and enables employees to pick up new competencies and skills beyond the technical aspects of digital or cloud-based data science solutions.


TIP

Consider the idea that it might be necessary to require learning on a broader scale to ensure that the basic principles of data science are understood by a significant subset of your employees. Google, for example, has developed a machine learning course that is mandatory for every technical employee.

Learning continuously

Data science transformation requires a new way of thinking about how change impacts people, cultures, organizations, processes, and more. Don't view the data science program as a small part of the process; rather, you must see it as a part of the entire digital transformation journey for your company.

For example, leaders should maintain ongoing blended learning-and-development programs that engage employees by describing the practical uses of data science so that understanding and familiarity build up among your workforce over time. Ongoing support helps employees embrace the agile culture and creates practitioners who learn in small increments continuously, building knowledge and expertise iteratively.


REMEMBER

Choosing a continuous learning approach incorporates the learning preferences of a multigenerational workforce and is effective where there is significant workforce turnover, regardless of whether it's planned or unplanned. In either approach (traditional or ongoing learning), managing the impact of the change should be seen as the core of a well-planned program, supported by fact-based content and by relevant and timely communication.

Recognizing what to avoid when driving change in data science

Around the globe, a number of businesses have made significant investments in data science, having realized (correctly) its revolutionary potential. Not having done their homework in terms of a proper situational analysis, however,

many of these businesses have suffered huge losses rather than the expected benefits.

Failing with the data science investment is particularly common among smaller and medium-size businesses. Why is that? Why are medium- and small-scale businesses unable to derive sufficient value by implementing data science? What obstacles stand in their way?

In an attempt to come up with some answers to these questions, Computer Associates interviewed 1,000 IT managers across companies with more than half a billion dollars in revenue in a range of different industries, from retail to financial services to pharma. Their research findings revealed that the biggest obstacle by far is an insufficient infrastructure. Many times, companies are stuck with their legacy environment due to previous costly investments that "cannot be thrown away." So, rather than creating a new, modern data architecture which puts the focus on the data, companies tend to add applications and system elements to their old environment, making data science inefficient and even more costly.

The second largest obstacle is organizational complexity. This usually becomes a problem when the company management underestimates how transformative data science is. All aspects of the company must change in order to become data-centric, meaning that all managers across the company must understand and use data and new data-driven insights for decision making related to finance, marketing, sales, product and service development, and so on. However, in reality many companies tend to treat data science like a side business by adding new roles and functions to work with data rather than transforming existing functions and roles.

The third most significant obstacle is security and other compliance concerns. This is not surprising, considering the growing awareness of the importance of handling data in a secure and ethically correct manner. New laws and regulations are becoming more and more strict in order to protect people's right to privacy, and as long as there is still very little standardization in data science, requirements will keep on changing.

A general finding in the study was that, based on the type of analytics approach that was chosen, the level of resistance varied. That's worth a closer look, so I walk you through some different types of high-level analytics projects in the following sections. Then you can get a better sense of the major factors underlying the success (or failure) of a data science transformation project.

Descriptive analytics transformation projects

Descriptive analytics projects involve tasks aimed at using data to describe what has happened or how things are right now — why, for example, we have sold x number of products this month of this specific product type. It includes activities such as developing graphs, charts, and dashboards, accompanied by no (or relatively simple) data analysis functions. The focus is on identifying the right set of metrics and presenting information in an effective manner.


REMEMBER

Descriptive analytics solutions generally face lesser resistance challenges in their implementations. The reasons are obvious — the deliverables are easily understood by stakeholders.

However, it is sometimes difficult to justify the business value of descriptive analytics projects. At the end of the day, with the limited analysis happening in descriptive analytics, what is really the value of investing in understanding what happened yesterday when what you *really* want is to be prepared for tomorrow?

Diagnostic analytics transformation projects

The objective of diagnostic analytics projects is to understand the reasons for a particular phenomenon and to conduct a root cause analysis. Diagnostic analytics projects can culminate in the development of statistical models (explanatory models, causal models, and so on) and dashboards. However, the output must include insights and recommendations designed to help stakeholders understand the reasons for what's happening and initiate appropriate actions.


REMEMBER

Organizations are usually receptive for analytical findings and insights based on diagnostic analytics outcomes, but there is a slightly higher resistance when it comes to implementing recommendations. This is mainly due to the fact that business users are aware that some recommendations aren't actionable because they require too many changes or have too many restrictions.

Predictive analytics transformation projects

Predictive analytics projects involve forecasting a certain metric or predicting a certain phenomenon. *Predictive modeling* is the process of applying a statistical model or data mining algorithm on data for the purpose of predicting new or future observations. Predictive models can be used for not just predictions but also simulation purposes. Examples include clinical research, sales prediction, production failure, and weather forecasting.

WARNING

As you might expect, predictive analytics solutions face the highest degree of resistance. Diagnostic and descriptive solutions largely deal with what has already happened, and predictive solutions relate to something that is yet to happen. Thus, business users have reservations about predictive solutions. This skepticism isn't groundless, because the cost of making wrong predictions can be astonishing.

Using Data Science Techniques to Drive Successful Change

For your data science investment to succeed, the data science strategy you adopt should include well-thought-out strategies for managing the fundamental change that data science solutions impose on an organization. One effective and efficient way to tackle these challenges is by using data-driven change management techniques to drive the transformation itself — in other words, drive the change by "practicing what you preach." I'll walk you through some examples of how to do this in practice.

Using digital engagement tools

For companies, there is a new generation of real-time employee opinion tools that are starting to replace old-fashioned employee opinion surveys. These tools can tell you far more than simply what employees are thinking about once a year. In some companies, employees are surveyed weekly using a limited number of questions. The questions and models are constructed in such a way that management can follow fluctuations in important metrics as they happen rather than the usual once or twice a year. These tools have obvious relevance for change management and can help answer questions like these:

- Is a change being equally well received across locations?
- Are certain managers better than others at delivering messages to employees?

Assume that you have a large travel-and-tourism firm that is using one of these tools for real-time employee feedback. One data-driven approach to use in such a situation is to experiment with different change management strategies within selected populations in the company. After a few changes in the organization, you can use the data collected to identify which managers prove to be more effective

in leading change than others. After that has been established, you can observe those managers to determine what they're doing differently. You can then share successful techniques with other managers.

This type of real-time feedback offers an opportunity to learn rapidly how communication events or engagement tactics have been received, thus optimizing your actions in days (rather than in weeks, which is typical of traditional approaches). The data can then feed into a predictive model, helping you determine with precision which actions will help accelerate adoption of a new practice, process, or behavior by a given employee group.


TIP

You can find some commercial tools out there — culture IQ polls, for example — that support this kind of data collection. These kinds of polls sample groups of employees daily or weekly via a smartphone app to generate real-time insights in line with whatever scope you have defined. Another tool, Waggl.com (`www.waggl.com`), has a more advanced functionality, allowing you to have an ongoing conversation with employees about a change effort as well as allowing change managers to tie this dialogue to the progress of initiatives they're undertaking.


REMEMBER

These different types of digital engagement tools can have a vast impact on change programs, but the data stream they create could be even more important. The data that's generated can be used to build predictive models of change. Using and deploying these models on real transformation projects and then sharing your findings helps to ensure a higher success rate with data-driven change initiatives in the future.

Applying social media analytics to identify stakeholder sentiment

Change managers can also look beyond the boundaries of the enterprise for insights about the impact of change programs. Customers, channel partners, suppliers, and investors are all key stakeholders when it comes to change programs. They are also more likely than employees to comment on social media about changes a company is making, thus giving potentially vital insight into how they're responding.

Ernst & Young (now known as EY) is using a tool for social media analytics called SMAART, which can interpret sentiment within consumer and influencer groups. In a project for a pharmaceutical company, EY was able to isolate the specific information sources that drove positive and negative sentiment toward the client's brand. The company is now starting to apply these techniques to understand the external impact of change efforts, and it's a simple leap to extend these

techniques within the enterprise. Advances in the linguistic analysis of texts mean that clues about behavior can now be captured from a person's word choices; even the use of articles and pronouns can help reveal how someone feels.


TIP

Applying sentiment analysis tools to data in anonymized company email or the dialogue in tools like Waggl.com can give fresh insight about your organization's change readiness and the reactions of employees to different initiatives. And, the insights gained from analyzing internal communication will be stronger when combined with external social media data.

Capturing reference data in change projects

Have you ever worked in an organization where different change programs or projects were compared to one another in terms of how efficiently they made the change happen? Or one where a standard set of measurements were used across different change initiatives? No? Me, neither. Why is it that organizations often seem obsessed with measuring fractional shifts in operational performance and in capturing data on sales, inventory turns, and manufacturing efficiency, but show no interest in tracking performance from change project to change project, beyond knowing which ones have met their goals?


REMEMBER

Some people may claim that you can't compare change projects within an organization; it would be like comparing apples to oranges. I disagree: Different projects may have unique features, but you'll find more similarities than differences between different types of projects. Capturing information about the team involved, the population engaged in the change, how long it took to implement, what tactics were used, and so on is a good idea. It enables you to build a reference data set for future learning, reuse, and efficiency benchmarking. However, remember that although it may not yield immediate benefit, as the overall data set grows, it will make it easier to build accurate predictive models of organizational change going forward.

Using data to select people for change roles

For quite a long time, companies have been using data-driven methods to select candidates for senior management positions. And today some businesses, such as retailers, are starting to use predictive analytics for hiring frontline staff. Applying these tools when building a change team can both improve project performance significantly and help to build another new data set.

TIP

If every change leader and team member would undergo testing and evaluation before a change project starts, that data could become important variables to include as you search for an underlying model on what leads to successful change projects. This can even be extended to more informal roles like change leaders, allowing organizations to optimize selection based on what they know about successful personalities for these types of roles.

Along these lines, the California start-up LEDR Technologies is pioneering techniques to predict team performance. It integrates data sources and uses them to help teams anticipate the challenges they may face with team dynamics so that the team can prevent them before they occur.

Automating change metrics

Picture a company or an organization that has a personalized dashboard it has developed in partnership with the firm's leadership team — one that reflects the company's priorities, competitive position, and future plans.

These dashboards should also be used to offer insights related to the different transformation investments you've made. Keep in mind that much of the data that can act as interesting change indicators are already available today — they're just not being collected.

REMEMBER

When a company builds a dashboard for identifying recruitment and attrition, it's teaching the executive team to use data to perform people-related decisions. However, it can take quite some time to set it up correctly and iron out the bugs. My suggestion? Don't wait. Start building these type of dashboards as fast as possible now and, where possible, automate them. Why the automation? Change dashboards are vulnerable to version control issues, human error, and internal politics. Automating data management and dashboard generation can make it more transparent and help you keep data integrity.

Getting Started

As organizations collect more data and build more accurate models, change managers will be able to confidently use them to prescribe strategies to enable organizations to meet their goals. They'll be able to answer important questions, such as these:

- W Which stakeholders are involved? What type of change approach works with groups that share these characteristics?
- What risks are associated with programs that share these features?
- What are the techniques that accelerate the delivery of business benefit, and what are their relative costs?
- What is the cause-and-effect of specific types of investment?

All these questions can be answered with data and will underpin data-driven transformation plans.

REMEMBER

Developing these sorts of metrics isn't quick or easy. They aren't one-off installations, but rather multiyear commitments to capture data, build models, and refine dashboards. Establishing stable and reliable data sets takes time. Data quality is an issue everywhere, and so is the need for a common data language that allows organizations to know that they're measuring what they intend to measure. This has been a problem for data analytics in other fields; there's no reason to think that change management will be any different.

Although it will take time, you'll eventually be able to close the causal loop and make reliable predictions for how an action or initiative in a change program will impact a given metric. This will move investment in change from being an act of faith to being a data-driven decision. Change management will move from a project-based discipline that's struggling to justify adequate investment to one that is advising on business outcomes and how to deliver them. This will lead to a decline in the one metric that is well known across change programs — the failure rate. And, as part of introducing data-driven change management, it should finally be possible to solve the great puzzle of why so many transformation efforts fail.

2 Making Strategic Choices for Your Data

IN THIS PART . . .

Sorting out the role of data

Selecting and exploring your data

Addressing ethical aspects related to data, teams, and algorithms

Transforming to a data-driven organization

Developing towards a machine-driven company

IN THIS CHAPTER

- Explaining the basic elements in data
- Identifying data value
- Describing the trends around data
- Understanding potential future options

Chapter 5

Understanding the Past, Present, and Future of Data

Business decisions force you to focus, allocate scarce resources, and think hard about exactly how to be unique compared to the competition. It is important to remember that you can't be everything to everyone. Strategically, you should think simplicity over complexity, since a clear and simple strategy is a lot easier to explain and to put into action. But what about making strategic choices for your data? Sure, data can help you understand your strategic options as well as the potential impact of various choices from a business perspective, but how do you utilize data to understand more about data itself?

Well, to make choices, you need to create choices. Real choices for your data cannot be made if you do not know what your options are and what options you are decisively rejecting. When deciding on a viable strategy, too often alternative strategic options are only considered superficially, and you need more to make the right choices. This chapter will therefore focus on sorting out the fundamental elements of data.

Sorting Out the Basics of Data

The terms *data* and *information* are often used interchangeably; there is a difference between them, however. For example, data can be described as raw, unorganized facts that need to be processed — a collection of numbers, symbols, or characters before it has been cleaned and corrected. Raw data needs to be corrected to remove flaws like outliers and data entry errors. Raw data can be generated in many different ways. *Field* data, for example, is raw data that has been collected in an uncontrolled live environment. *Experimental* data has been generated within the context of a scientific investigation by observation and recording. Data can be as simple and seemingly random and useless until it's organized, but once data is processed, organized, structured, or presented in a given context that makes it useful, it's called *information.*

Historically, the concept of data has been most closely associated with scientific research, but now data is being collected, stored, and used by an increasing number of companies, organizations, and institutions. For companies, examples of interesting data can be customer data, product data, sales data, revenue, and profits; for governments, it can include data such as crime rates and unemployment rates.

During the second half of the 1900s, there were several attempts to standardize the categorization and structure of data in order to make sense of its various forms. One well-known model for this is the DIKW (*d*ata, *i*nformation, *k*nowledge, and *w*isdom) pyramid, described in the following list; the first version of this model was drafted already in the mid-1950s, but it first appeared in its current state in the mid-1990s, as an attempt to make sense of the growing amounts of data (raw or processed) that were being generated from different computer systems:

- **Data** is raw. It simply exists and has no significance beyond its existence (in and of itself). It can exist in any form, usable or not. Data represents a fact or statement of event without relation to other factors — *it's raining,* for example.
- **Information** is data that has been given a meaning by way of some sort of relationship. This meaning can be useful, but does not have to be. The information relationship can be related to cause-and-effect — *the temperature dropped 15 degrees and then it started raining*, for example.
- **Knowledge** is the collection of information with the purpose to be useful. It represents a pattern that connects discrete elements and generally provides a high level of predictability for what is described or what will happen next: *If the humidity is very high and the temperature drops substantially, the atmosphere is often unlikely to be able to hold the moisture, and so it rains,* for example.

» **Wisdom** exemplifies more of an understanding of fundamental principles within the knowledge that essentially form the basis of the knowledge being what it is. Wisdom is essentially like a shared understanding that is not questioned; *It rains because it rains,* for example. And this encompasses an understanding of all interactions that happen between raining, evaporation, air currents, temperature gradients, changes, and rain.

The DIKW pyramid offered a new way to categorize data as it passes through different stages in its life cycle and has gained some attention over the years. However, it has also been criticized, and variants have appeared that were designed to improve on the original. One major criticism has been that, although it's easy enough to understand the step from data to information, it's much harder to draw a clear and valid line from information to knowledge and from knowledge to wisdom, making it difficult to apply in practice.


REMEMBER

Conceptual models are *heuristic* devices: They're useful only insofar as they offer a way to learn something new. One model or another may be more appealing to you, but from the perspective of a data science implementation, the most important thing for you to consider is a question like this: Will my company gain value from having the four levels of the DIKW pyramid, or will it just make implementation more difficult and complex? (Personally, I like a pyramid with just two levels: data and insights. That one has worked fine for me so far, and is far easier to explain and garner support for.)

Explaining traditional data versus big data

Traditional data is data in a volume and format that makes it easy to access, work with, and act on. Big data is a different animal, however, defined more by the volume of the data, the variety of the types of data involved, and the velocity at which it's processed. If all three of these characteristics are fulfilled in terms of being too big to handle in an ordinary processing environment, you can assume that you're dealing with big data and not traditional data (now known, after the appearance of big data on the scene, as *small* data).

Furthermore, the term *big data* is used to refer to data sets that are too large or complex to be handled by traditional data processing application software. Big data challenges include tasks such as capturing data, transferring data, storing data, cleaning and preparing data, exploring and analyzing data, searching data, sharing and reusing data, visualizing data, updating data, and managing privacy, data ownership, and governance.

Although big data was originally described by three key concepts — volume, variety, and velocity — two other concepts have lately been added: veracity (data quality, in other words) and value. These additional characteristics have been

added to describe two other important aspects of big data that you should consider when estimating the potential benefits of a big-data data set.

What's so important about veracity? If the data set that a certain company wants to explore fulfills the criteria of big data based on volume, variety, and velocity, it could still be useless for the company to invest in if the data quality is poor and cannot be corrected. Poor data quality (the data set is incomplete, corrupt, or biased, for example) directly impacts trust in the data itself, ultimately impacting the perceived value of the overall data set.

Figure 5-1 graphically represents the key concepts, know, as the "five *V*s," that define big data.

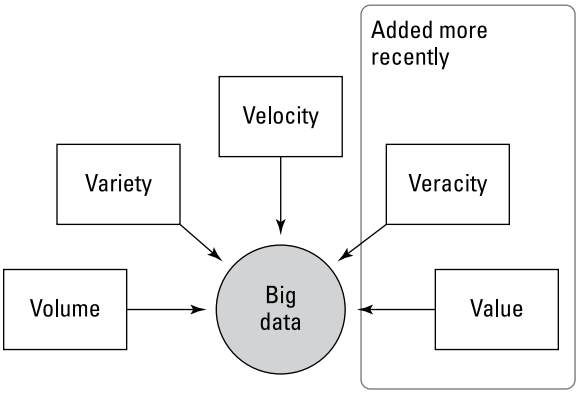

FIGURE 5-1: Defining big data.

I add my own take on these concepts in the following list:

- **Volume:** Refers to the quantity of generated and stored data. The size of the data determines the value and potential insight, and whether it can be considered big data.

  As you begin to unleash the power of big data, you soon discover that the data streams connected to your business will grow exponentially. This sharp increase in data volume can cause your organization significant difficulties if you haven't planned accordingly, because each new data set places considerable strain on your current data storage and computational setup.

WARNING

- **Velocity:** Refers to the speed at which the data is generated and processed to meet your organization's business objectives. Big data is often available in real-time. Compared to small data, big data is produced more continually. These two types of velocity are related to big data:
  - *Frequency of generation*
  - *Frequency of capture*

  
  REMEMBER

  If you find it difficult to cope with the exponentially increasing volume of data in your company, the fact that the speed will also increase probably sounds like an intimidating detail. However, to take full advantage of big data, you must focus not only on how much information you collect, but also how fast you can use the data to either make business decisions or incorporate and utilize it as part of a service or product offering from your company.

- **Variety:** Refers to the type and context of the data. Variety helps people who analyze data use the resulting insight effectively. Historically, the data was in a well-structured format at the point of collection. Now, as the use of big data becomes standard practice, it's unstructured data that is fast becoming the norm in the corporate world.

  
  REMEMBER

  If you're looking to expand your use of big data, you' have to become accustomed to the lack of structured data. But with the right analytical setup, these new varieties of data can increase your business growth, enabling the exploration of new opportunities in what was previously unknown territory.

- **Veracity:** Refers to the level of noise in the data. The quality of captured data can vary greatly, affecting the accuracy of the analysis. Yes, big data provides your business with the chance to accumulate information from places you never thought possible, but if the data isn't accurate or timely, it doesn't matter what you decide to do with it.
- **Value:** Refers to the business value gained from big data.

So there you have it — the five *V*s of big data: volume, velocity, variety, veracity, and value. Not all of these characteristics of big data are equal in importance, however. Four of them (volume, velocity, variety, and veracity) can be seen as enablers. To achieve the fifth *V* (value) requires an understanding of what the organization is trying to accomplish. That's why I emphasize the need to be perfectly clear about your overall strategic business objectives before you can leverage volume, velocity, variety, and veracity to achieve value in big data.

Knowing the value of data

The statement "Data is the new oil" is one that lots of people make, but what does it mean? In some ways, the analogy *does* fit: It's easy to draw parallels because of

the way information (data) is used to drive much of the transformative technology available today via artificial intelligence, machine learning, automation, and advanced analytics — much like oil drives the global industrial economy.

So, as a marketing approach and a high-level description, the expression does its job, but if you take it as an indication of how to strategically address the value of data, it might lead to investments that cannot be turned into value. For example, storing data has no guaranteed future value, like oil has. Storing even more data has even less value because it becomes even more difficult to find it so that you can put it to use. The value in data lies not in saving it up or storing it — it lies in putting it to use, over and over again. That's when the value in data is realized.

If you start by looking at the core of the analogy, you can see that it refers to the value aspects of data as an enabler of a fundamental transformation of society — just like oil has proven to be throughout history. From that perspective, it definitely showcases the similarities between oil and data. Another similarity is that, although inherently valuable, data needs processing — just as oil needs refining — before its true value can be unlocked.

However, data also has many other aspects that cause the analogy to fall apart when examined more closely. To see what I mean, check out some of the differences I see between these two enablers of transformation:

- **Availability:** Though oil is a finite resource, data is an endless and constantly increasing resource. This means that treating data like oil (hoarding it and storing it in siloes, for example) has little benefit and reduces its usefulness. Nevertheless, because of the misconception that data is similar to oil (scarce), this is often exactly what is done with the data, driving investments and behavior in the wrong direction.
- **Reusability:** Data becomes more useful the more it's used, which is the exact opposite of what happens with oil. When oil is used to generate energy like heat or light, or when oil is permanently converted into another form such as plastic, the oil is gone and cannot be reused. Therefore, treating data like oil — using it once and then assuming that its usefulness has been exhausted and disposing of it — is definitely a mistake.
- **Capture:** Everyone knows that as the world's oil reserves decline, extracting it becomes increasingly difficult and expensive. With data, on the other hand, it's becoming increasingly available as the digitalization of society increases.
- **Variety:** Data also has far more variety than oil. The raw oil that's drilled from the ground is processed in a variety of ways into many different products, of course, but in its raw state, it's all the same. Data in its raw format can represent words, pictures, sounds, ideas, facts, measurements, statistics, or any other characteristic that can be processed by computers.

REMEMBER

The fact nevertheless remains that the quantities of data available today comprise an entirely new commodity, though the rules for capturing, storing, treating, and using data are still being written. Let me stress, however, that data, like oil, is a vital source of power and that the companies that utilize the available data in the most optimized way (thereby controlling the market) are establishing themselves as the leaders of the world economy, just as the oil barons did a hundred years ago.

Exploring Current Trends in Data

Big data was definitely *the thing* just a couple of years ago, but now there's much more of a buzz around the idea of *data value* — more specifically, how analysis can turn data into value. The next few sections look at some of the trends related to utilizing data to capture new value.

Data monetization

Monetizing data refers to how companies can utilize their domain expertise to turn the data they own or have access to into real, tangible business value or new business opportunities. Data *monetization* can refer to the act of generating measurable economic benefits from available data sources by way of analytics, or, less commonly, it may refer to the act of monetizing data services. In the case of analytics, typically these benefits appear as revenue or cost savings, but they may also include market share or corporate market value gains.

REMEMBER

One could argue that data monetization for increased company revenue or cost savings is simply the result of being a data-driven organization. Though that argument isn't totally wrong, company leaders are taking an increasing interest in the market to explore how data monetization can drive the innovation of entirely new business models in various different business segments.

One good example of how this process can work is when telecom operators sell data on the positions of rapidly forming clusters of users (picture the conclusion of a sporting event or a concert by the latest YouTube sensation) to taxi companies. This allows taxi cars to be available proactively in the right area when a taxi will most likely be needed. This is a completely new type of business model and customer base for a traditional telecom operator, opening up new types of business and revenues based on available data.

Responsible AI

Responsible AI systems are characterized by transparency, accountability, and fairness, where users have full visibility into which data is being used and how. It also assumes that companies are communicating the possible consequences of using the data. That includes both potential positive and negative impact.

Responsible AI is also about generating customer and stakeholder trust based on following communicated policies and principles over time, including the ability to maintain control over the AI system environment itself.

Strategically designing your company's data science infrastructure and solutions with responsible AI in mind is not only wise, but could also turn out to be a real business differentiator going forward. Just look at how the opposite approach, taken by Facebook and Cambridge Analytica, turned into a scandal which ended by putting Cambridge Analytica out of business. You might remember that Cambridge Analytica gained access to the private and personal information of more than 50 million Facebook users in the US and then offered tools that could then use that data to identify the personalities of American voters and influence their behavior. Facebook, rather than being hacked, was a willing participant in allowing their users' data to be used for other purposes without explicit user consent.

THE ROLE OF OPEN SOURCE IN DATA SCIENCE

Open source data architectures are no longer analogous to research projects forever running in lab environments for trials and experimentation. Now considered mainstream in IT environments, these architectures are widely deployed in live production in several industries. In fact, it has become so common that if you're building a modern data architecture, chances are you're using an open source stack. Some companies have even found that using open source architectures provides the only cost-effective path to getting something done.

The tipping point is more or less here, which means that the time is now to decide how to strategically react to the opportunities associated with open source. It's past the point where making small incremental changes or playing it safe with traditional proprietary infrastructures was sufficient. Now, if you continue to play it safe or stick with baby steps, you will leave your company at risk of being left behind while competitors move ahead. It's also important to remember that deciding on utilization of open source software is not incremental; rather, it necessitates a full-bore disruptive architectural approach.

The data included details on users' identities, friend networks, and "likes." The idea was to map personality traits based on what people had liked on Facebook, and then use that information to target audiences with digital ads. Facebook has also been accused of spreading Russian propaganda and fake news which, together with the Cambridge Analytica incident, has severely impacted the Facebook brand the last couple of years. This type of severe privacy invasion has not only opened many people's eyes in terms of the usage of their data but also impacted the company brands.

Cloud-based data architectures

More and more companies are moving away from on-premise-based data infrastructure investments toward virtualized and cloud-based data architectures. The driving force behind this move is that traditional data environments are feeling the pressure of increasing data volumes and are unable to scale up and down to meet constantly changing demands. On-premise infrastructure simply lacks the flexibility to dynamically optimize and address the challenges of new digital business requirements.


TIP

Re-architecting these traditional, on-premise data environments for greater access and scalability provides data platform architectures that seamlessly integrate data and applications from various sources. Using cloud-based compute and storage capacity enables a flexible layer of artificial intelligence and machine learning tools to be added as a top layer in the architecture so that you can accelerate the value that can be obtained from large amounts of data.

Computation and intelligence in the edge

Edge computing describes a computing architecture in which data processing is done closer to where the data is created— Internet of Things (IoT) devices like connected luggage, drones, and connected vehicles like cars and bicycles, for example. There is a difference between pushing computation to the edge (edge compute) and pushing analytics or machine learning to the edge (edge analytics or machine learning edge). Edge compute can be executed as a separate task in the edge, allowing data to be preprocessed in a distributed manner before it's collected and transferred to a central or semi-centralized environment where analytics methods or machine learning/artificial intelligence technologies are applied to achieve insights. Just remember that running analytics and machine learning on the edge requires some form of edge compute to also be in place to allow the insight and action to happen directly at the edge.

The reason behind the trend to execute more in the edge mainly depends on factors such as connectivity limitations, low-latency use cases where millisecond response times are needed to perform an immediate analysis and make a decision (in the case of self-driving cars, for example). A final reason for executing more in the edge is bandwidth constraints on transferring data to a central point for analysis. Strategically, computing in the edge is an important aspect to consider from an infrastructure-design perspective, particularly for companies with significant IoT elements.

REMEMBER

When it comes to infrastructure design, it's also worth considering how the edge compute and intelligence solutions will work with the centralized (usually cloud-based) architecture. Many view cloud and edge as competing approaches, but cloud is a style of computing where elastically scalable technology capabilities are delivered as a service, offering a supporting environment for the edge part of the infrastructure. Not everything, however, can be solved in the edge; many use cases and needs are system- or network-wide and therefore need a higher-level aggregation in order to perform the analysis. Just performing the analysis in the edge might not give enough context to make the right decision. Those types of computational challenges and insights are best solved in a cloud-based, centralized model, as illustrated in Figure 5-2.

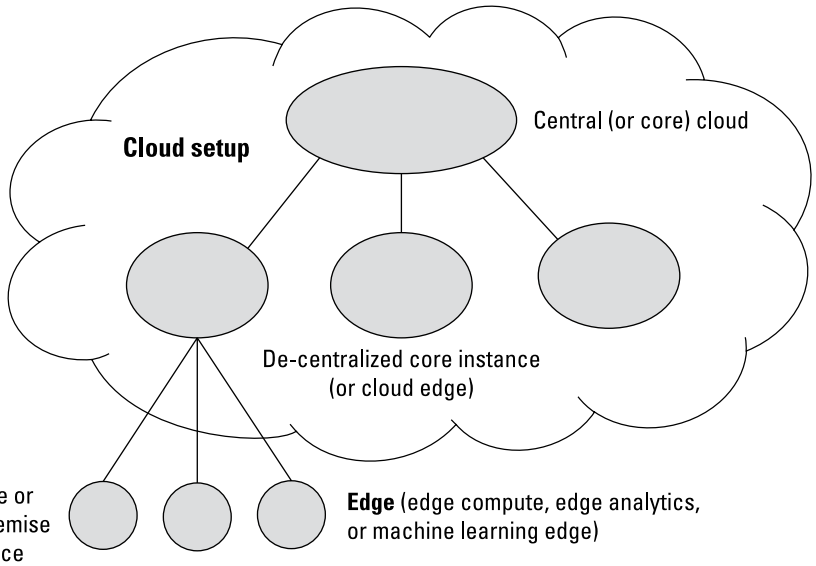

FIGURE 5-2: A model for cloud/edge computing.

As you can see, the cloud setup can be done in a decentralized manner as well, and these decentralized instances are referred to as *cloud-edge*. For a larger setup on a regional or global scale, the decentralized model can be used to support edge implementations at the IoT device level in a certain country or to support a telecom operator in its efforts to include all connected devices in the network. This is useful for keeping the response time low and not moving raw data over country borders.

Digital twins

A *digital twin* refers to a digital representation of a real-world entity or system — a digital view of a city's telecommunications network built up from real data, for example. Digital twins in the context of IoT projects is a promising area that is now leading the interest in digital twins. It's most likely an area that will grow significantly over the next three to five years. Well-designed digital twins are assets that have the potential to significantly improve enterprise control and decision-making going forward.

Digital twins integrate artificial intelligence, machine learning, and analytics with data to create living digital simulation models that update and change as their physical counterparts change. A digital twin continuously learns and updates itself from multiple sources to represent its near real-time status, working condition, or position. (See Figure 5-3 for an overview of this process.)

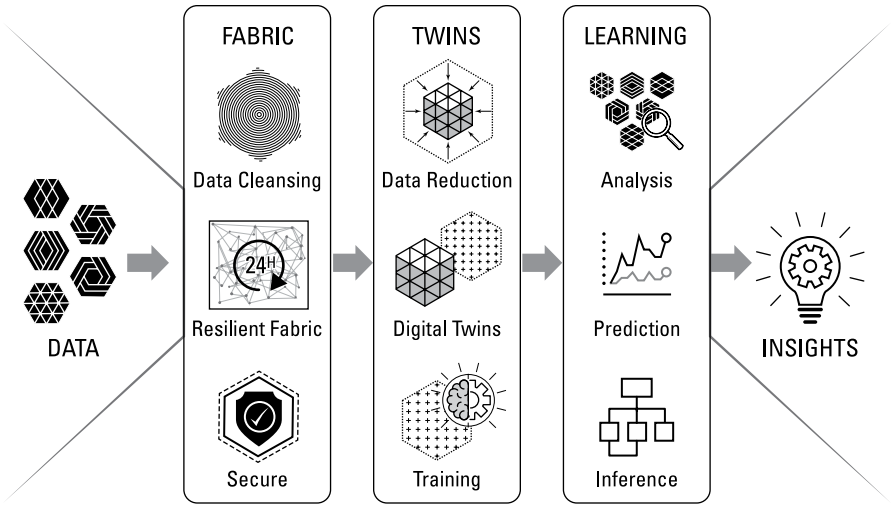

FIGURE 5-3: How digital twins produce insights.

REMEMBER

Digital twins are linked to their real-world counterparts and are used to understand the state of the system, respond to changes, improve operations, and add value. Digital twins start out as simple digital views of the real system and then evolve over time, improving their ability to collect and visualize the right data, apply the right analytics and rules, and respond in ways that further your organization's business objectives. But you can also use a digital twin to run predictive models or simulations which can be used to find certain patterns in the data building up the digital twin that might lead to problems. Those insights can then be used to prevent a problem proactively.

TIP

Adding automated abilities to make decisions based on the digital-twin concept of predefined and preapproved policies would be a great capability to add to any operational perspective — managing an IoT system such as a smart city, for example.

Blockchain

The blockchain concept has evolved from a digital currency infrastructure into a platform for digital transactions. A *blockchain* is a growing list of records (blocks) that are linked using cryptography. Each block contains a cryptographic hash of the previous block, a timestamp, and transaction data. By design, a blockchain is resistant to modification of the data. It's an open and public ledger that can record transactions between two parties efficiently and in a verifiable and permanent way. A blockchain is also a decentralized and distributed digital ledger that is used to record transactions across many computers so that any involved record cannot be altered retroactively without the alteration of all subsequent blocks. The blockchain technologies offer a significant step away from the current centralized, transaction-based mechanisms and can work as a foundation for new digital business models for both established enterprises and start-ups. Figure 5-4 shows how to use blockchain to carry out a blockchain transaction.

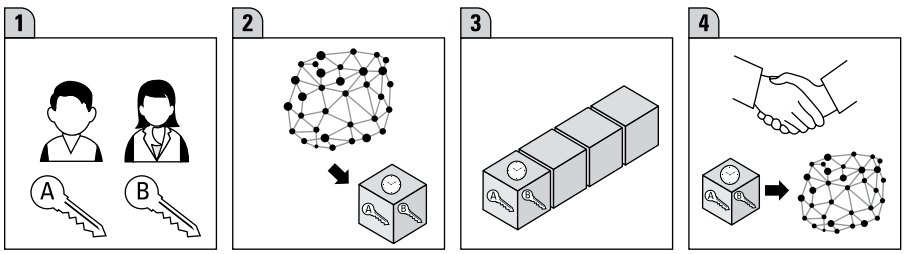

FIGURE 5-4: Creating a blockchain transaction.

WARNING

Although the hype surrounding blockchains was originally focused on the financial services industry, blockchains have many potential areas of usage, including government, healthcare, manufacturing, identity verification, and supply chain. Although blockchain holds long-term promise and will undoubtedly create disruption, its promise has yet to be proven in reality: Many of the associated technologies are too immature to use in a production environment and will remain so for the next two to three years.

Conversational platforms

Conversational AI is a form of artificial intelligence that allows people to communicate with applications, websites, and devices in everyday, humanlike natural language via voice, text, touch, or gesture input. For users, it allows fast interaction using their own words and terminology. For enterprises, it offers a way to build a closer connection with customers via personalized interaction and to receive a huge amount of vital business information in return. Figure 5-5 shows the interaction between a human and a bot.

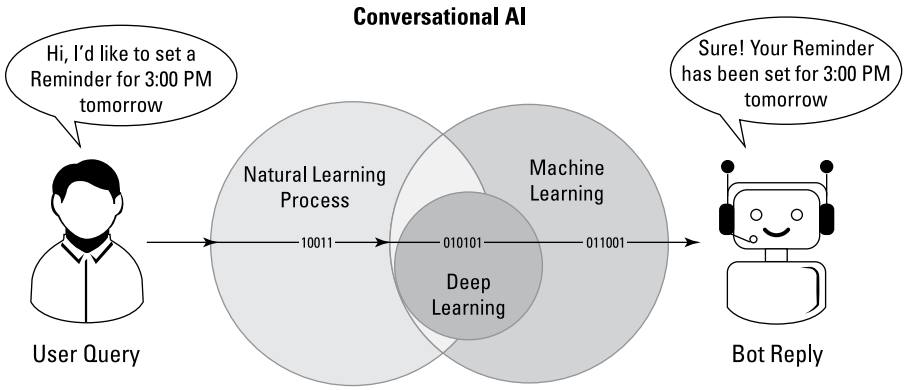

FIGURE 5-5: An example of how to use a conversational platform.

This type of platform will most likely drive the next paradigm shift in how humans interact with the digital world. The responsibility for translating intent shifts from humans to machines. The platform takes a question or command from the user and then responds by executing some function, presenting some content, or asking for additional input. Over the next few years, conversational interfaces will become a primary design goal for user interaction and will be delivered in dedicated hardware, core OS features, platforms, and applications.

Check out the following list for some potential areas where one could benefit from applying conversational platforms by way of bots:

- **Informational:** Chatbots that aid in research, informational requests, and status requests of different types
- **Productivity:** Bots that can connect customers to commerce, support, advisory, or consultative services
- **B2E (business-to-employee):** Bots that enable employees to access data, applications, resources, and activities
- **Internet of Things (IoT):** Bots that enable conversational interfaces for various device interactions, like drones, appliances, vehicles, and displays

Using these different types of conversational platforms, you can expect increased bot productivity (because they can concentrate on the most valuable interactions), a 24/7 automated workforce, increased customer loyalty and satisfaction, new insights into customer interactions, and reduced operational expenses.


TECHNICAL STUFF

Conversational platforms have now reached a tipping point in terms of understanding language and basic user intent, but they still aren't good enough to fully take off. The challenge that conversational platforms face is that users must communicate in a structured way, and this is often a frustrating experience in real life. A primary differentiator among conversational platforms is the robustness of their models and the application programming interfaces (APIs) and event models used to access, attract, and orchestrate third-party services to deliver complex outcomes.

Elaborating on Some Future Scenarios

Although the explosion of new use cases (those specific situations in which data science could potentially be used) and applications in data science is happening all around us, there are still scenarios yet to come. In this section, I describe some potential future scenarios in the data science space, including challenges and motivations.

Standardization for data science productivity

Standardization generally ensures the smooth operation of processes and builds credibility over time. Best practices ensure efficiency and reduce redundancy. Today, the amount of data generated in the world is increasing by the second.

As this data is collected and stored, it's critical to standardize and normalize it for optimal usage. Otherwise, it can get very noisy.

One of the biggest challenges in building an optimized data management solution is the lack of standardization when collecting data from all over the Internet — or even just across different parts of a large, global company. Standardization is vital when trying to avoid redundancy and increase accuracy in matching data types. As necessary as it is, it's still a difficult problem to solve. Lack of standardization proves to be a hindrance to many business systems. All data needs to be converted to a predefined format, which requires domain expertise as well as agreement (both internally and externally) on data definitions and structure needs to be reached.

So, are there any reasons why standardization in the ML/AI space is particularly important? I'm glad you asked. The following list gives you some of the reasons why it's particularly needed:

- **Model interoperability:** Interoperability standards are important for not only a global community in general but also any single company implementing machine learning and artificial intelligence. The reason is that in order to scale company development of algorithms, interoperability between models in production is needed to ensure not only that the models can work together, but that they can also maximize the model performance in a multi-model environment.
- **Process standardization:** An emphasis on process control also brings into question which standardizations are needed regarding safety, performance, latency, reliability, bias, and even privacy. By standardizing the best practices of developing, for example, complex techniques like deep learning, not only can more teams accelerate their development but more innovative solutions can also be developed independently and be plugged in to accelerate a much larger process.
- **Human-Machine compatibility:** Here, compatibility refers to standardizing the interaction between human and machine — when the human uses certain words or phrases, the machine uses a standardized set of responses or actions, for example. Many errors can occur because of different objectives and methods for the interactions. For mission-critical systems, imagine a machine-managed air traffic control tower; it could be a disaster if the interaction isn't standardized between systems or at least cumbersome to have to learn all different versions of how to achieve an objective in collaboration with the machine, depending on which implementation is in use and where.

» **Ethics:** The challenges of AI standardization cover many levels of concern, but this one is vital — any form of AI standardization should include methods for how to best drive AI for the maximum benefit of humanity. It would be a total failure if standardization would lead to more advanced autonomous weaponry or more enhanced methods to predict and manipulate human behavior.

From data monetization scenarios to a data economy

Huge advances in technology have led to an explosion in the rate at which new data is being created. The data economy is delivering what businesses and governments across the world want: Create high-quality jobs, generate economic growth, and enable organizations across all sectors to expand successfully and serve their customers.

But with the growth of data, are company leaders realizing its true potential? As the data economy emerges, changes in customer expectations and technological advancements will transform supply chains into complex ecosystems. Production strategies will shift, and collaboration across organizations and ecosystems will create a more open flow of information and ideas. Companies will need to reinvent themselves by defining their desired roles in the data economy by way of an evaluation of their engagement in these ecosystems. This will allow organizations to assess whether new business units, joint ventures, and acquisitions will be required.

An explosion of human/machine hybrid systems

Hybrid human/machine systems combine machine and human intelligence to overcome the shortcomings of existing AI systems. The need for human involvement to overcome the mistakes and limitations of AI systems is already acknowledged in critical domains such as medicine and driving. (A driver of a semiautonomous car is expected to continuously watch over the decisions of the machine and correct it when needed to prevent accidents, for example.) However, successfully integrating human and machine intelligence has its challenges. Human intelligence is a valuable resource associated with higher costs and constraints like an 8-hour workday for example. The quality and availability of human input may also vary depending on other factors, including the condition of the human, like illness or fatigue.

One way to overcome the challenges of hybrid human/machine systems is to change the way machines access human intelligence. For that to happen, AI systems would need to be equipped with reasoning capabilities that can make effective decisions about how it should access that human intelligence. Here, recent advances in human computation may provide some clues about how AI systems could accomplish this. Crowdsourcing platforms provide easy access to human intelligence on demand in a scalable and adaptable way. Simply defined, crowdsourcing happens when a company or an institution outsources a function once performed by a limited number of employees to an undefined (and generally large) network of people in the form of an open call.

This approach can take the form of peer-production (when the job is performed collaboratively), but is also often undertaken by single individuals. For AI systems in which a user isn't in the loop to provide help, the human help needed by the system may be provided by the crowd. For many research efforts, including the ones presented here, crowdsourcing platforms function as test beds for data collection and experimentation related to the challenges of accessing and working with human intelligence.

Quantum computing will solve the unsolvable problems

If you spend more than five minutes on the Internet, watching the news, and otherwise staying current with the world, you have heard the excitement surrounding recent advances in the development of quantum computer systems.

This is not an exaggeration — it really will change everything. Quantum computers have the potential to blow right through obstacles that limit the power of classical computers, solving problems in seconds that would take a classical computer the entire life of the Universe just to attempt to solve — encryption and research on new advanced medicine, for example. When chemists research new medicines, much of their work is testing hundreds of possible variables in a chemical formula in order to find the desired characteristics needed to treat a variety of illnesses. This process of experimentation and discovery often leads to a development time of more than 10 years before a new drug is brought to market — often at a cost of billions of dollars. Computation today is done on computers that have to combine and recombine elements to test the results.

Needless to say, the race is now on to make quantum computers into practical everyday tools for business, industry, and science in order to gain a competitive advantage. Quantum computing is here to stay, it is growing, and if it doesn't solve all of the world's problems, it could potentially solve quite a few

Quantum computers are qualitatively different from standard computers in how they compute data. On the one hand, you have your standard binary digital electronic computer, where the data needs to be encoded into *b*inary dig*its* (bits), each of which is always in one of two definite states (0 or 1). Quantum computation uses *qu*antum *bits* (*qubits*), which can be in *superpositions* of states — that is to say, just like Schrödinger's cat could be both alive and dead, a qubit can be both 0 and 1.

IN THIS CHAPTER

- Managing the selection and collection of data effectively
- Examining and documenting the data
- Enhancing your understanding of the data by using an explorative approach
- Analyzing and rating data quality

Chapter 6
Knowing Your Data

Approaching your data strategy in the right way is fundamental for you to secure a stable foundation for the rest of your data science investment. And it's not just about securing the integrity in the data; you also need to make sure that the data types you choose for your business objectives are the right ones and are selected for the right reasons. For that to happen, you need to understand the data you're targeting. To gain that understanding, you have to successfully work four main steps: Select data, describe data, explore data, and assess data quality.

Selecting Your Data

Data selection is the process of determining the appropriate data type and source — as well as the suitable methods — to collect data. Data selection precedes the actual task of data collection.


REMEMBER

The main objective of data selection is to determine the appropriate data type, source, and method(s) necessary to provide you with the answers you need to questions you've posed. The selection is often connected to a certain area — finance, sales, product, or consumer, for example — and is mostly driven by the type of analysis you intend to use, as well as by your ability to access the necessary data sources.

WARNING

Integrity issues can arise when the decisions to select appropriate data to collect are based primarily on cost and convenience considerations rather than on the ability of data to effectively answer the questions posed. Certainly, cost and convenience are valid factors in the decision-making process. However, you have to assess to what degree these factors might impact the integrity of the analysis.

In this first part of the data selection process (see Figure 6-1), consider your answers to these questions:

- What questions are you trying to answer?
- What is the scope of the analysis?
- Within the field you're aiming to analyze, what type of data is the industry generally targeting?
- What data format is needed to answer your business questions: quantitative, qualitative, or both?

Figure 6-2 captures the main areas of concern related to data collection.

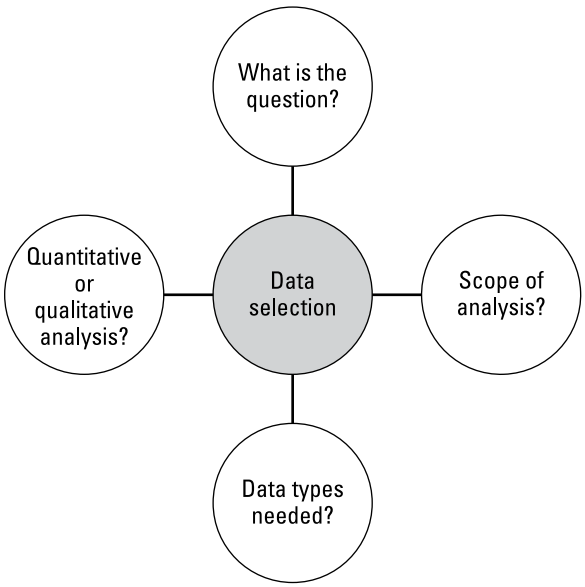

FIGURE 6-1: Aspects to consider when selecting data.

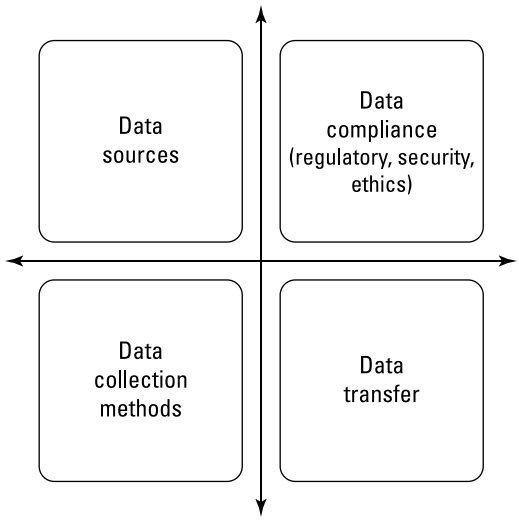

FIGURE 6-2: Data collection areas to address.

As part of the data collection process, consider your answers to these questions:

- Where are the data sources you need to tap into? Do they exist, or do you need to create new data (using a survey, for example)? If the data exists, do you already have it in your company, or do you need to acquire it?
- Are there any legal, ethical, or security restrictions related to the data you need, like data privacy, data ownership, or data retention periods?
- How do you need to collect the data? Are frequent batch uploads sufficient, or do you need to stream live data? Will your infrastructure be able to manage different ways of acquiring the data?
- Is there a need to move data across country borders? Are there legal restrictions related to the specific data you're targeting? If such restrictions exist, how will they be dealt with?

Describing Data

Describing data refers to the task of examining and documenting the collected data so that elements such as data format, quantity of data, and metadata can be noted and recorded. A *metadata* element is data type used to describe other data in order to increase the usefulness of the original data. Creating and maintaining metadata is a vital part of your data science environment.

Keep in mind that you'll encounter a number of different types of metadata when setting up a data science environment. I briefly describe the different kinds in this list:

- **Descriptive:** Used to describe the main characteristics of a data element for the purposes of discovery and identification. It can include elements such as title, abstract, author, and keywords.
- **Structural:** Deals with metadata about groups of data and indicates how multiple objects are put together — how pages are ordered to form chapters, for example. It describes data categories, versions, relationships, and other characteristics of digital materials.
- **Administrative:** Provides information to help manage a data element, such as when and how it was created, file type and other technical information, as well as who can access it.
- **Reference:** Describes the contents and quality of statistical data.
- **Statistical:** Describes processes that collect, process, or produce data statistics (also referred to as *process* data).

Describing your data is a strategically important task in order to evaluate whether the collected data satisfies your identified business requirements — and it includes several distinct steps. This list gives you an overview of what activities need to take place:

- Analyze the data volume and try to estimate the level of complexity.
- Describe the different tables needed and their relationships to one another.

  
  REMEMBER

  Data tables help you keep information organized. If you're collecting data from an experiment or scientific research, saving it in a data table will make it easier to look up later. Data tables can also help you make graphs and other charts based on your information.
- Check the availability of attributes, which helps to describe the context of each data type — the geographical location it was collected from, for example, or the date it was collected.
- Determine whether there are different types of attributes needed. Types could include the following:
  - *Nominal*: ID numbers, eye color, zip codes
  - *Ordinal:* Rankings (taste of potato chips on a scale from 1-10, for example), grades, height in categories (tall, medium, short)
  - *Interval:* Calendar dates, temperatures in Celsius or Fahrenheit
  - *Ratio*: Exact temperature in Kelvin, length, time, counts.

Consider how you intend to use the data to determine which attributes will be required.

- Describe the value range of the selected attributes (if applicable). Remember that the same attribute can be mapped to different attribute values. (Height can be measured in feet or meters, for example.)
- Analyze potential attribute correlations, such as gender and height or date and temperature.
- Understand the meaning of each attribute, and describe the value in business terms.
- For each attribute, compute basic statistics (distribution, average, maximum, minimum, standard deviation, variance, mode, and skewness, for example) and relate the results to their meaning in business terms.
- Decide attribute relevance related to the specific business objective by involving domain experts.
- Determine whether the meaning of each attribute is used consistently.
- Decide whether it's necessary to balance the data, if the data distribution is distorted.
- Analyze and document key relationships in the data.

REMEMBER

Without the data properties attached to your data, the value and usefulness of your data will be significantly reduced. The documentation of data properties is a cornerstone for further analysis of the data and must be kept relevant as part of your data management activities as long as you intend to use the data in any way.

Exploring Data

A vital step in getting to know your data better is to explore it. *Data exploration* is an approach similar to an initial data analysis, where a data analyst uses visual exploration to understand what is in a dataset as well as the characteristics of the data. These characteristics can include size or amount of data, completeness of the data, correctness of the data, or possible relationships or insights that may be hidden in the data, for example. Visual data exploration is the activity of searching and finding out more about the data using various statistical models visualized through graphical representation of the data — heat maps, geo maps, box plots and word clouds, for example. Just by looking at the same data set, using various graphical representations, it is possible to detect data correlations and dependencies, as well as new insights in the data.

Data exploration is usually conducted using a combination of these types of methods:

- **Automated:** Can include data profiling or data visualization to give the analyst an initial view of the data as well as an understanding of key characteristics.
- **Manual:** Often follows an automated action by manually drilling down or filtering the data to identify anomalies or patterns identified automatically. Data exploration usually also require manual scripting and queries into the data (using languages such as Python or R, for example) or using Excel (for smaller data sets) or similar tools to view the raw data.

REMEMBER

The actual data mining task is the semiautomatic or automatic analysis of large quantities of data to extract previously unknown and/or interesting patterns, such as groups of data records (cluster analysis), unusual records (anomaly detection), and dependencies (association rule mining, sequential pattern mining).

Figures 6-3, 6-4, 6-5, and 6-6 show different ways of exploring data. Figure 6-3 starts things off by looking at school grades in Sweden. The bar graph to the left displays the classical medium value per region; in that view, Stockholm is shown to be performing best. But the medium value in the bar graph cannot be trusted. Why? A closer look at the Stockholm region in the box-plot shows that there are problems in that area. There's a very big spread of the grades — actually, the largest spread in Sweden is in Stockholm — and when you study this in more detail you can see that there is a very large difference between the schools. This is referred to as segregation. If you compare these numbers with the Norrbotten region, you will see a much smaller spread of grades, which indicates a higher overall performance.

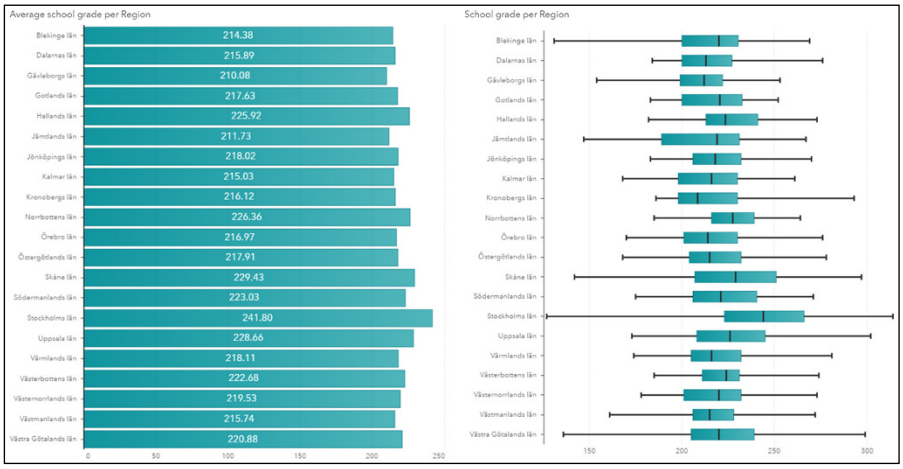

FIGURE 6-3: Data exploration on school grades in Swedish regions using a box-plot.

Figure 6-3 is based on a screenshot generated using SAS® Visual Analytics software. Copyright © 2019 SAS Institute Inc., Cary, NC, USA. SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. All Rights Reserved. Used with permission.

In Figure 6-4, you can see an example where a scatterplot is used to understand whether or not the educational level of parents actually impacts school grades. And, as you can read from the graph, there is definitely a strong correlation indicated by the distribution of values grouped along a diagonal 45-degree line from the left hand lower corner to the upper right corner.

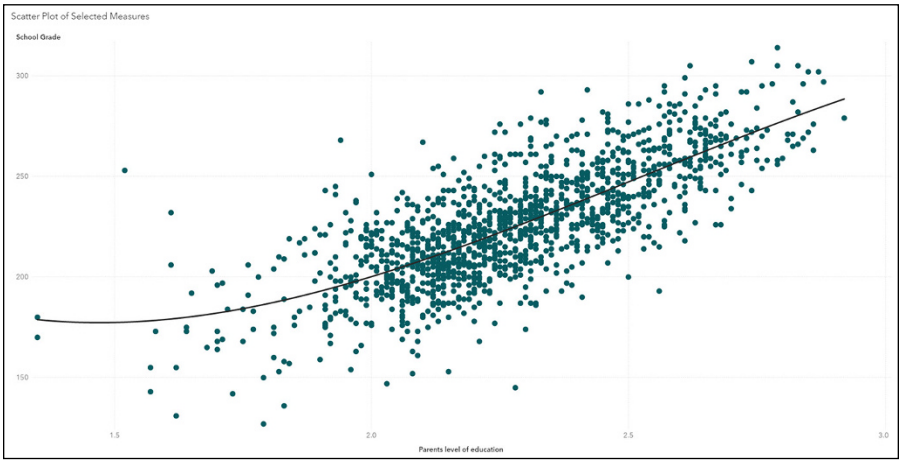

FIGURE 6-4: A scatter-plot exploring dependencies in the data.

Figure 6-4 is based on a screenshot generated using SAS® Visual Analytics software. Copyright © 2019 SAS Institute Inc., Cary, NC, USA. SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. All Rights Reserved. Used with permission.

Figure 6-5 uses data to show how visitors to a website are moving between different parts of the website. With the data, it is possible to determine where most people tend to start browsing (search engines, referral, direct links and other), and where they tend to move from their point of entry. Gaining a better understanding of these type of patterns can be helpful in terms of understanding the user experience and if there is a pattern to where they tend to "drop-out" and leave the website.

Figure 6-6 shows another way of exploring data using a *heat map* — a graphical representation of data that uses a system of color-coding to represent different values. This example is analyzing manufacturing yield data for various products to find dependencies or correlation between a certain product and the city where it's manufactured. Manufacturing yield refers to the quality of the product from a perspective of how often a product needs to be replaced (level of yield). In the heat map in Figure 6-6, you can see that the coloring reaches from lighter (which is bad) to darker (which is good). By using a heat map, you can quickly get an overview of the situation and see which cities are having manufacturing problems for certain products, based on their yield rates.

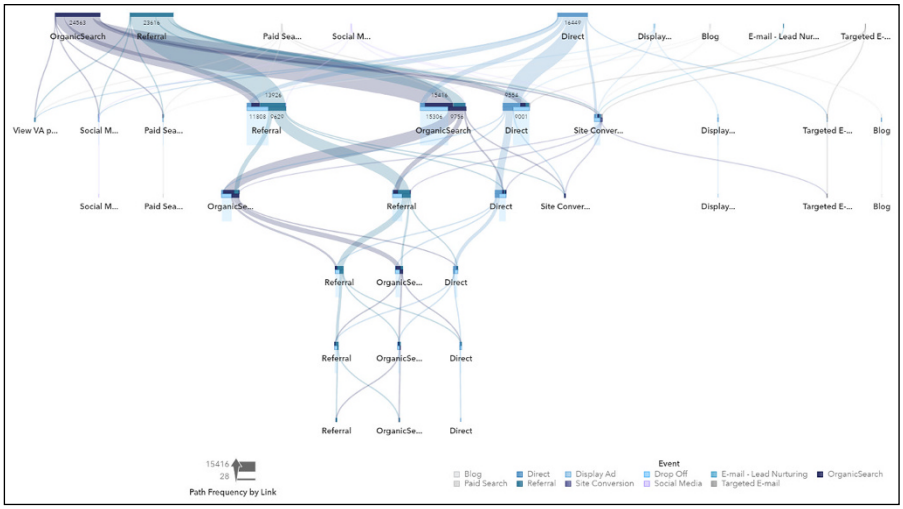

FIGURE 6-5: A path analysis chart using data to show how users enter, move and leave a specific website.

Figure 6-5 is based on a screenshot generated using SAS® Visual Analytics software. Copyright © 2019 SAS Institute Inc., Cary, NC, USA. SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. All Rights Reserved. Used with permission.

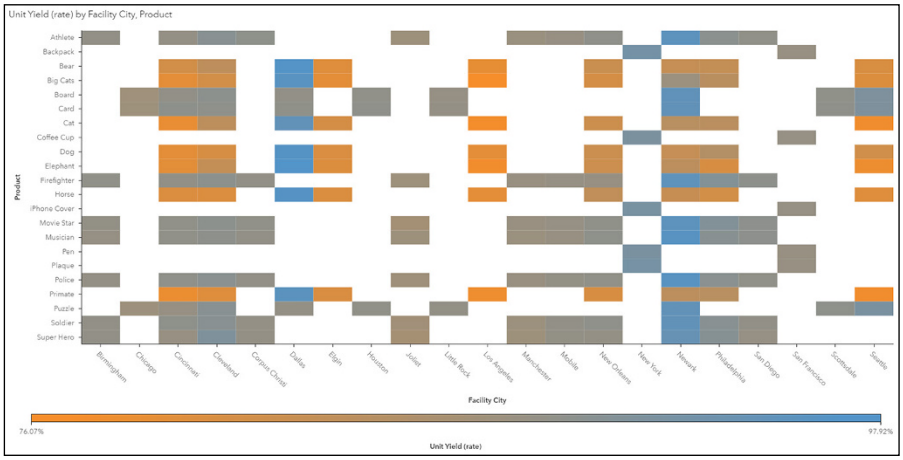

FIGURE 6-6: A heat map analyzing potential correlation between product manufacturing yield and product manufacturing city.

Figure 6-6 is based on a screenshot generated using SAS® Visual Analytics software. Copyright © 2019 SAS Institute Inc., Cary, NC, USA. SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. All Rights Reserved. Used with permission.

REMEMBER

All these exploration tasks are aimed at creating a clearer view of the data so you can get to know your data better, all in hopes of gaining some first insights based on your exploration that you might want to analyze further. This is a key first step the analyst and is just as important as defining basic metadata (statistics, structure, relationships) for the data set that can then be used in further analysis.

Assessing Data Quality

Another fundamental part of data understanding involves gaining a detailed view of the quality of the data as soon as possible. Many businesses consider data quality and its impact too late — well past the time when it could have had a significant effect on the project's success. By integrating data quality with operational applications, organizations can reconcile disparate data, remove inaccuracies, standardize on common metrics, and create a strategic, trustworthy, and valuable data asset that enhances decision making. Also, if an initial analysis suggests that the data quality is insufficient, steps can be taken to make improvements. One way to refine the data is by removing unusable parts; another way is to correct poorly formatted parts.


TIP

Start by asking yourself questions such as these: Is the data complete? Does it cover all required cases? Is it correct, or does it contain errors? If there are errors, how common are they? Does the data have missing values? If so, how are they represented, where do they occur, and how common are they?

A more structured list of data quality checkpoints includes steps such as these:

- Check data coverage (whether all possible values are represented, for example).
- Verifying that the meaning of attributes and available values fit together. For example, if you are analyzing data on geographical location for retail stores, is the value captured in latitude and longitude, rather than the name of the regional area it is placed in?
- Identifying missing attributes and blank fields.
- Classifying the meaning of missing or wrong data, and double-check attributes with different values but similar meanings.
- Checking for inconsistencies in the spelling and formatting of values (situations where the same value sometimes starts with a lowercase letter and sometimes with an uppercase letter, for example). Without consistent naming conventions and numerical format, data correlation and analysis will not be possible cross data sets.
- Reviewing deviations and deciding if any of them qualify as mere noise (outliers) or indicate an interesting phenomenon (pattern).
- Check whether there is noise and inconsistencies between data sources.

REMEMBER

If you detect data quality problems as part of the quality check, you need to define and document the possible solutions. Focus on attributes that seem to go against common sense; visualization plots, histograms, and other ways of visualizing and exploring data are great ways to reveal possible data inconsistencies. It may also be necessary to exclude low-quality or useless data entirely in order to perform the needed analysis.

Figure 6-7 shows table formatted to show an overview of a data set. Tables like these are a good way to get a first overview of your data from a quality perspective because it uses descriptive statistics to quickly detect extreme values in terms of things like minimum values, maximum values, median, medium, and standard deviation. The table also allows us to analyze the key values to make sure that they are 100% unique and do not include any duplicated or missing values. If you are studying data related to your customers, for example, you want to make sure that a customer does not occur twice due to a spelling error — or is missing from the list altogether!

Column	Unique	Primary Key ...	Null	Bla...	Pattern Count	Mean	Median	Minimum	Maximum	Standa...
Age	0,09 % (43)	No				39,29	39,00	19,00	61,00	5,59
City	0,09 % (42)	No			16			Auckland	Valencia	
Continent	0,01 % (5)	No			4			Africa	South Ame...	
Country	0,06 % (29)	No			10			Argentina	Venezuela	
Customer	100,00 % (45...	Yes			3			ARBUENO...	ZAJOHAN...	
CustomerDistance	100,00 % (45...	Yes				5,56	5,74	0,04	10,00	2,34
CustomerInDays	4,54 % (2073)	No				156,66	20,00	0,00	3 647,00	396,05
CustomerLat	100,00 % (45...	Yes				25,60	40,52	-37,92	60,07	32,85
CustomerLon	100,00 % (45...	Yes				7,20	2,15	-77,14	174,85	55,35
FirstOrderAmount	97,10 % (44...	No				199,74	161,82	4,73	1 874,32	159,25
FirstOrderCostOfSales	100,00 % (45...	Yes				165,85	140,92	4,20	1 566,99	126,92
FirstOrderCostOfSalesPerc	100,00 % (45...	Yes				0,85	0,91	0,57	0,99	0,11
FirstOrderCustomerSatisfac	100,00 % (45...	Yes				0,51	0,45	0,18	1,00	0,20
FirstOrderDeliveryTime	0,05 % (25)	No				4,30	3,00	1,00	25,00	3,38
FirstOrderDiscount	0,68 % (312)	No				14,02	6,00	0,00	767,00	27,64
FirstOrderDiscountPerc	57,0...	No				0,07	0,04	0,00	0,88	0,09
FirstOrderExp	3,74 % (1707)	No				0,41	0,37	0,01	1,00	0,20
FirstOrderListAmount	3,93 % (1794)	No				213,76	173,58	5,61	1 905,72	167,17
FirstOrderProductQuality	100,00 % (45...	Yes				0,77	0,78	0,56	0,92	0,06
FirstOrderProfit	0,94 % (430)	No				-0,83	-10,00	-234,00	270,00	41,67

FIGURE 6-7: Profiling the data to get an overview of the data quality.

Figure 6-7 is based on a screenshot generated using SAS® Visual Analytics software. Copyright © 2019 SAS Institute Inc., Cary, NC, USA. SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. All Rights Reserved. Used with permission.

Figure 6-8 shows a graphical visualization of the same data, but this graph focuses on just one column; country. By looking at the data from a country perspective, you can validate the data distribution in another way, and possibly detect inconsistencies or missing values that were difficult to detect from the overview. In this specific example, the tool actually has a functionality called *Pattern* which indicates when data values are deviating from the norm.

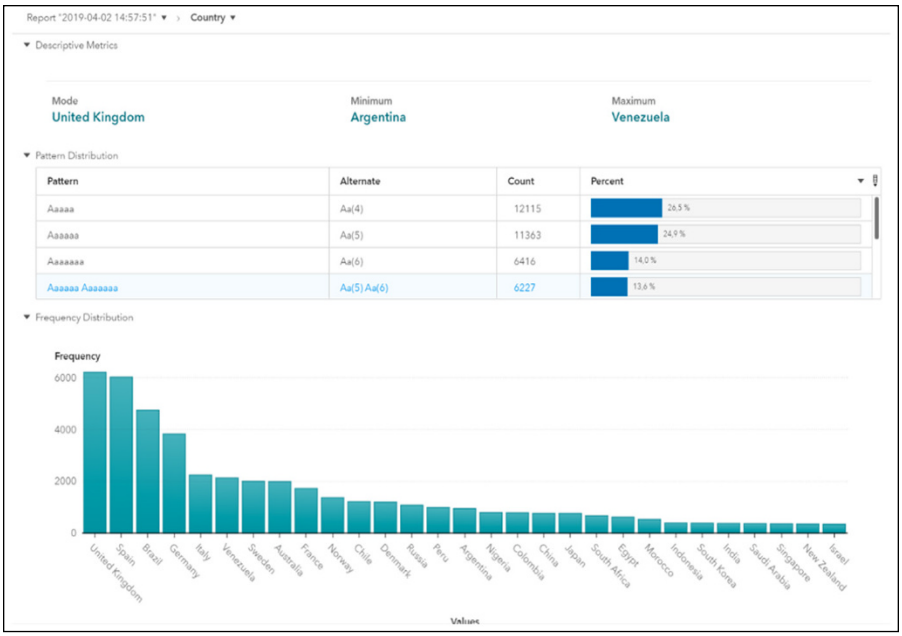

FIGURE 6-8: Data profiling and validation from a country perspective.

Figure 6-8 is based on a screenshot generated using SAS® Visual Analytics software. Copyright © 2019 SAS Institute Inc., Cary, NC, USA. SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. All Rights Reserved. Used with permission.

Ask any thriving organization the secret to success and you'll get many answers: a solid data strategy, or calculated risks in combination with careful budgeting. They're all good business practices which come from the same place: a solid foundation of high-quality data. When you have accurate, up-to-date data driving your business, you're not just breaking even — you're breaking records. A data quality assessment process is essential to ensure reliable analytical outcomes. This process depends on human supervision-driven approaches since it is impossible to determine a defect based only on data.

Improving Data Quality

So, what do you do, practically speaking, if you realize that your data quality is really bad? My recommendation is to use the four-step approach outlined below to get started on highlighting the gaps and defining a road map to implement needed improvements in data quality.

1. **Scope:** Define the data quality problem and describe the business issue related to the quality problem. Specify data details and business processes, products and/or services impacted by the data quality issues.

2. **Explore:** Conduct interviews with key stakeholders on data quality needs and expectations, as well as data quality problems. Review data quality processes and tool support (if any) across business functions and identify needed resources for the quality improvement activity.
3. **Analyze:** Assess current practices against industry best practices for data quality and align it with the findings from the exploration phase.
4. **Recommend:** Develop a road map for improving the data quality process and define how the technical architecture for the data quality process must look like, incorporating your findings of what is not working today and what is essential from a business perspective.

IN THIS CHAPTER

- Walking through ethical considerations related to using artificial intelligence
- Managing various aspects of responsible artificial intelligence
- Designing your artificial intelligence approach with ethics in mind

Chapter 7

Considering the Ethical Aspects of Data Science

According to Gartner's CIO Agenda Survey from 2018, 85% of AI projects through 2020 will deliver erroneous outcomes due to bias in data, algorithms, or development teams. This is serious figure that must be addressed. Think about it: if more and more companies and organizations are becoming data and AI-driven as well as automating their operations based on artificial intelligence technologies that cannot be trusted, that means there's trouble on the horizon is — not only for the evolution of AI, but for society as a whole.

In this context, addressing the ethical aspects of artificial intelligence is fundamental and will be my focus in this chapter. But I also want to make clear that you shouldn't start thinking about ethics only when you get around to implementing your data science strategy. An ethical perspective is, in fact, *hugely* important to consider from the start — that is to say, from the moment you start designing your business models, architecture, infrastructure, and ways of working and building up the teams themselves. This chapter explains the basics you need to consider from a strategic as well as practical perspective.

Explaining AI Ethics

So, what does AI ethics actually refer to and which areas are important to address to generate trust around your data and algorithms? Well, there are many aspects to this concept, but there are five cornerstones to rely on;

- **Unbiased data, teams, and algorithms**. This refers to the importance of managing inherent biases that can arise from the development team composition if there isn't a good representation of gender, race, and sex. Data and training methods must be clearly identified and addressed through the AI design. Gaining insights and potentially making decisions based on a model that is in some way biased (a tendency toward gender inequality or racist attitudes, for example) isn't something you want to happen.
- **Algorithm performance**. The outcomes from AI decisions shall be aligned with stakeholder expectations that the algorithm performs at a desired level of precision and consistency and doesn't deviate from the model objective. When models are subsequently deployed in their target environment in a dynamic manner and continue to train and optimize model performance, the model will adjust to the potential new data patterns and preferences and might start deviating from the original goal. Setting sufficient policies to keep the model training on target is therefore vital.
- **Resilient infrastructure**. Make sure that the data used by the AI system components and the algorithm itself are secured from unauthorized access, corruption, and/ or adversarial attack.
- **Usage transparency and user consent**. A user must be clearly notified when interacting with an AI and must be offered an opportunity to select a level of interaction or reject that interaction completely. It also refers to the importance of obtaining user consent for data captured and used. The introduction of the General Data Protection Regulation (GDPR) in the EU has prompted discussions in the US calling for similar measures, meaning that the awareness of the stakes involved in personal information as well as the need to protect that information are slowly improving. (For more on the GDPR, refer to Chapter 3.) So, even if the data is collected in an unbiased manner and models are built in an unbiased setup, you could still end up with both ethically challenging situations (or even breaking the law) if you're using personal data without the right permissions.
- **Explainable models**. This refers to the need for AI's training methods and decisions criteria to be easily understood, documented, and readily available for human assessment and validation. It refers to situations where care has been taken to ensure that an algorithm, as part of an intelligent machine, produces actions that can be trusted and easily understood by humans.

The opposite of AI explainability is when the algorithm is treated as a black box, where even the designer of the algorithm cannot explain why the AI arrived at a specific insight or decision.


TECHNICAL STUFF

An additional ethical consideration, which is more technical in nature, relates to the reproducibility of results outside of the lab environment. AI is still immature, and most research-and-development is exploratory by nature. There is still little standardization in place for machine learning/artificial intelligence. De facto rules for AI development are emerging, but slowly and they are still very much community driven. Therefore, you must ensure that any results from an algorithm are actually *reproducible*— meaning you get the same results in the real, target environment as you would not only in the lab environment but also between different target environments (between different operators within the telecommunications sector, for example.)

Addressing trustworthy artificial intelligence

If the data you need access to in order to realize your business objectives can be considered ethically incorrect, how do you manage that? It's easy enough to say that applications should not collect data about race, gender, disabilities, or other protected classes. But the fact is that if you do not gather that type of data, you'll have trouble testing whether your applications are in fact fair to minorities.


REMEMBER

Machine learning algorithms that learn from data will become only as good as the data they're running on. Unfortunately, many algorithms have proven to be quite good at figuring out their own proxies for race and other classes, in ways that run counter to what many would consider proper human ethical thinking. Your application would not be the first system that could turn out to be unfair, despite the best intentions of its developers. But, to be clear, at the end of the day your company will be held responsible for the performance of its algorithms, and (hopefully) bias-related legislation in the future will be stricter than it is today. If a company isn't following laws and regulations or ethical boundaries, the financial cost could be significant — and perhaps even worse, people could lose trust in the company altogether. That could have serious consequences, ranging from customers abandoning the brand to employees losing their jobs to folks going to jail.

To avoid these types of scenarios, you need to put ethical principles into practice, and for that to happen, employees must be allowed and encouraged to be ethical in their daily work. They should be able to have conversations about what ethics actually means in the context of the business objectives and what costs to the

company can be weathered in their name. They must also be able to at least discuss what would happen if a solution cannot be implemented in an ethically correct manner. Would such a realization be enough to terminate it?

Data scientists in general find it important to share best practices and scientific papers at conferences, writing blog posts, and developing open source technologies and algorithms. However, problems such as how to obtain informed consent aren't discussed quite as often. It's not as if the problems aren't recognized or understood; they're merely seen as less worthy of discussion. Rather than let such a mindset persist, companies should actively encourage (rather than just allow) more discussions about fairness, the proper use of data, and the harm that can be done by the inappropriate use of data.

WARNING

Recent scandals involving computer security breaches have shown the consequences of sticking your head in the sand: Many companies that never took the time to implement good security practices and safeguards are now paying for that neglect with damages to their reputations and their finances. It is important to exercise the same due diligence now accorded security matters when thinking about issues like fairness, accountability, and unintended consequences of your data use. It will never be possible to predict all unintended consequences of such usage and, yes, the ability to foresee the future is limited. But plenty of unintended consequences could easily have been foreseen. (Facebook's Year in Review feature, which seemed to go out of its way in to remind Facebook users of deaths in the family and other painful events, is a prime example.)

REMEMBER

Mark Zuckerberg's famous motto, "Move fast and break things," is unacceptable if it hasn't been thought through in terms of what is likely to break. Company leaders should insist that they be allowed to ponder such aspects — and stop the production line whenever something goes wrong. This idea dates back to Toyota's Andon manufacturing method: Any assembly line worker can stop the line if they see something going wrong. The line doesn't restart until the problem is fixed. Workers don't have to fear consequences from management for stopping the line; they are trusted, and are expected to behave responsibly.

What would it mean if you could do this with product features or AI/ML algorithms? If anyone at Facebook could have said, "Wait, we're getting complaints about Year in Review" and pulled it out of production, Facebook would now be in a much better position from an ethical perspective. Of course, it's a big, complicated company, with a big, complicated product. But so is Toyota, and it worked there.

The issue lurking behind all these concerns is, of course, corporate culture. Corporate environments can be hostile to anything other than short-term profitability. However, in a time when public distrust and disenchantment are running at an all-time high, ethics is turning into a good corporate investment. Upper-level management is only starting to see this, and changes to corporate culture won't happen quickly, but it's clear that users want to deal with companies that treat them and their data responsibly, not just as potential profit or as engagements to be maximized.


TIP

The companies that will succeed with AI ethics are the ones that create space for ethics within their organizations. This means allowing data scientists, data engineers, software developers, and other data professionals, to "do ethics" in practical terms. It isn't a question of hiring trained ethicists and assigning them to their teams; it's about living ethical values every single day, not just talking about them. That's what it means to "do good data science."

Introducing Ethics by Design

What's the best way to approach implementing AI ethics by design? Might there be a checklist available to use? Now that you mention it, there is one, and you'll find it in the United Kingdom. The government there has launched a data ethics framework, featuring the data ethics workbook. As part of the initiative, they have isolated seven distinct principles around AI ethics. The workbook they came up with is built up around a number of open-ended questions designed to probe your compliance with these principles. Admittedly, it's a lot of questions — 46, to be exact, which is rather too many for a data scientist to continuously keep track of and incorporate efficiently into a daily routine. For such questions to be truly useful then, they need to be embedded not only in the development ways of working but also as part of the data science infrastructure and systems support.


REMEMBER

It isn't merely a question of making it possible as a practical matter to follow ethical principles in daily work and to prove how the company is ethically compliant — the company must also stand behind these ambitions and embrace them as part of its code of conduct. However, when a company talks about adding AI ethics to its code of conduct, the value doesn't come from the pledge itself, but rather emerges from the process people undergo in developing it. People who work with data are now starting to have discussions on a broad scale that would never have taken place just a decade ago. But discussions alone won't get the hard work done. It is vital to not just *talk* about how to use data ethically but also to *use* data ethically. Principles must be put into practice!

Here's a shorter list of questions to consider as you and your data science teams work together to gain a common and general understanding of what is needed to address AI ethical concerns:

- **Hacking:** To what extent is an intended AI technology vulnerable to hacking, and thus potentially vulnerable to being abused?
- **Training data:** Have you tested your training data to ensure that it is fair and representative?
- **Bias:** Does your data contain possible sources of bias?
- **Team composition:** Does the team composition reflect a diversity of opinions and backgrounds?
- **Consent:** Do you need user consent to collect and use the data? Do you have a mechanism for gathering consent from users? Have you explained clearly what users are consenting to?
- **Compensation:** Do you offer reimbursement if people are harmed by the results of your AI technology?
- **Emergency brake:** Can you shut down this software in production if it's behaving badly?
- **Transparency and Fairness:** Do the data and AI algorithms used comply with corporate values for technology such as moral behavior, respect, fairness and transparency? Have you tested for fairness with respect to different user groups?
- **Error rates:** Have you tested for different error rates among diverse user groups?
- **Model performance:** Do you monitor model performance to ensure that your software remains fair over time? Can it be trusted to perform as intended, not just during the initial training or modelling but also throughout its ongoing "learning" and evolution?
- **Security**: Do you have a plan to protect and secure user data?
- **Accountability**: Is there a clear line of accountability to an individual and clarity on how the AI operates, the data that it uses, and the decision framework that is applied?
- **Design:** Did the AI design consider local and macro social impact, including its impact on the financial, physical, and mental well-being of humans and our natural environment?

IN THIS CHAPTER

- Realizing why it is a necessity to become data driven
- Moving towards a data-driven approach
- Scoping and defining a data strategy
- Establishing a data-driven mindset

Chapter 8
Becoming Data-driven

Unless a company invests big money in becoming data-driven in today's business climate, it will eventually perish. Companies that don't believe that their data is an asset (and therefore should be managed accordingly) will end up in a lot of trouble within the next five years. This chapter explains why it's necessary for your company to become data driven and offers some advice on what steps your company needs to take in order to become data-driven.

Understanding Why Data-Driven Is a Must

Companies and organizations across many business areas have begun their journey to capture, create, and use data in ways that are fundamentally changing how people work and live. And, as you're probably already aware, the starting point for any data-driven organization is simply the realization that data is at the core of everything it does. It is truly the foundation of everything, and organizations across the business spectrum are now becoming aware of the transformative power of data, analytics, and AI.

Companies are also starting to understand the real challenges that lie ahead. For many, it's tough enough to catalog and categorize all the data available; identifying and adding rules for processing and using the data in order to translate the data into tangible value seems an almost insurmountable task. But, although it's

difficult, it isn't impossible, and several companies are now starting to address this challenge more strategically *and* more practically. "Nice to know," you might say, "but where do we start?"


TIP

Rather than start by hiring an external data person (as this person is commonly known), my advice to you, based on my experience in the field, is to invest some time and effort in finding a key person in your organization who is willing to lead your data-driven initiative. This person should be someone who can see the bigger picture and help create a data strategy based on thorough insights into how the company functions as well as its future business objectives. That key person should also have the people skills and communication skills to transform an entire organization — with sufficient support, of course — and should be willing (and patient enough) to be on the frontline to move the company from a simple data integrator all the way to a market innovator.


REMEMBER

When businesses today claim to recognize that data has a value, their contentions aren't necessarily correct in terms of the true value of the data. It's easy enough to understand that transactional data can be used for reporting or data analytics, which can then lead to better decision-making. But even though the perceived value of data has increased over the past two decades, many companies still lag behind when it comes to efficiently capturing, sharing, and managing data. This is mainly because their systems and processes reflect an outdated belief that data is simply the byproduct of some other activity — rather than the key to their business success. To move beyond such an approach and actually enter the 21st century, such organizations need to invest significantly more time and effort into creating a data strategy.

That's all well and good, but what does it actually mean when I say that a company needs to create a data strategy? First and foremost, developing a data strategy means making sure that all data resources are positioned in such a way that they can be used, shared, and moved easily and efficiently. In other words, having a data strategy ensures that data is managed and used as an asset and not simply as a byproduct of another application. By establishing common methods, practices, and processes to manage, manipulate, and share data across the company in a repeatable manner, a data strategy ensures that the goals and objectives to use data effectively and efficiently are aligned in a conscious and strategic manner.

Unfortunately, just as many companies still use data as a byproduct rather than as the core value of their businesses, many don't resolve their data problems by creating and following through on their own data strategies, but rather hire data analysts tasked with the chore of "finding things in the data." The result is that instead of having someone onboard with the clear business objective of turning data into insights, you give a bunch of data analysts access to a database and have them run queries that any basic analytics tool could provide in seconds instead. What's the point of that?

WARNING

Without a data strategy, it doesn't matter whether the people you hire are called data analysts or data scientists or data engineers or machine learning/artificial intelligence engineers. If you're hiring people with fancy titles just because everybody else in the market seems to be doing it, it isn't a sign that you have suddenly understood the value of a data strategy. Without clearly defined objectives that you've committed to carrying out in a strategic way, all the hires you make will be worse than window dressing because you'll be investing time and money and receiving little in return. Essentially, you're spending an enormous amount of money to attract and retain data analysts (or scientists or engineers) who spend most of their days extracting, cleaning, and modeling data without knowing a) which problems to focus on and b) how one could create a new business opportunity that generates revenues or profit for the company.

Transitioning to a Data-Driven Model

Becoming data-driven is both an organizational ambition and an absolute necessity. It involves cultural aspects as well as technology aspects. It's about using data to take direct action, in addition to building relationships and trust around the data. But it's also about how you look at the data, and how you come to use it as part of your day-to-day business. Technologically savvy management teams understand that attempting to "boil the ocean" when adopting a data-centric strategy is a foolish thing to do. Some have already learned this the hard way, by undertaking less-than-successful "big-bang" transformation projects. Given the scope and complexity of the dynamic nature of the digital world, companies are starting to understand that change will continue to accelerate, even as they achieve or progress beyond the target state.

REMEMBER

Management must set visionary goals for data-centricity, but they also need to allow for change during implementation — and even perhaps for changes to the vision itself. Given these realities, it's imperative that management approach both data management and analytics and machine learning/artificial intelligence initiatives from a continuous improvement perspective, driving progress toward goals while following a roadmap that is adjusted as business and organizational needs evolve.

Becoming data-driven is a new way to approach your business, and it's understandable that many companies get lost in all their data and ambitions. On top of that, everything is moving very fast in terms of the constantly evolving techniques used in the areas of data science, artificial intelligence, and virtualized infrastructures. (And don't even get me started on the new and expanded policies in regulatory, security, and ethical practices.)

REMEMBER

When you introduce a data-driven approach to your organization, it isn't enough to have clear objectives on what needs to be achieved — you also need a way to measure your achievements toward those objectives. On top of tracking progress, you must be able to measure and prove the value and impact that data science and machine learning have on your business.

Securing management buy-in and assigning a chief data officer (CDO)

The most important decision that company management has to make is to make someone fully responsible for data. Doing so sends the right message across the organization not only internally but also externally, toward the market and its customers. You want everyone to know that you're taking seriously the task of becoming data-driven and that it is what will drive the company forward into the future. (For more on the role of the CDO as well as advice on how to establish the function in your organization, see Chapter 12.)

REMEMBER

A large part of the CDO's responsibility should be the company-wide management and use of data as an organizational and strategic asset. This means working together with every single department to design a common way to acquire, store, manage, share, and use data. As important as that is, it's even more important to ensure that the culture adopts a data-driven way of thinking so that the decision-making process is informed by discussions enabled by sharing and reuse of data, models, and insights. At the end of the day, it's all about making use of the data in real decisions across the company and as part of its portfolio of products and/or services.

TIP

Top management sponsorship and approval is of course essential for the success of any data strategy, but management must take on more than just a stewardship or enforcement function. Leadership must also "walk the talk," embracing fact-based decision-making, pushing for more and better data, and recognizing achievement when efforts succeed.

Management must also provide a clear vision, prioritize analytical applications, understand return on investment, allocate appropriate resources, manage talent, ensure cross-functional coordination, and remove some of the barriers that will inevitably pop up during implementation. Finally, management must insist on compliance with legal and regulatory requirements on the data in scope.

Identifying the key business value aligned with the business maturity

The starting point for any data analysis should be an understanding of the most significant business opportunities and/or problems for your company. Given that starting point, you can then focus your analysis on identifying and describing how a data-driven approach can contribute and provide value in that perspective. In either case, the whole point of a data-driven approach is to provide value where there was none before.

WARNING

Never get caught up in the potential of a particular technology; always stay focused on its application. Rather than asking, "What can this new technology do for us?" you should ask, "What problems do I need to solve?" The situation has intensified due to the rapid technology evolution, creating a skills shortage in most companies. Of course, there's room for some experimentation with regard to what the new technologies can enable, but only as long as you stay focused on the primary goal, which is driving the business forward. Successfully harnessing data, analytics, and machine learning/artificial intelligence ensures that you'll be able to accelerate the benefits of adopting a data-driven approach, which will in turn increase the drive for further deployment.

All companies go through different stages in a *business life cycle* — that progression of a business through various phases over time, most commonly divided into the five stages of launch, growth, shake-out, maturity, and decline. However, it's worth noticing that in large, global companies, different business areas within the same company may be in different stages of a business life cycle at any given time. This happens, for example, when a new business segment or area is added to a setup with a more traditional set of business areas — a new segment addressing digital business for example. This mix of new and more traditional business areas in a large enterprise can be difficult to handle when it comes to aligning business objectives or carrying out a business transformation plan.

In either case, however, it is important to align the business objectives with the business maturity, since the success of implementing a data-driven approach is very much dependent on the readiness of the company. A recently established company in its launch or growth phase might already be based on a fully digitalized business model and half way towards becoming data-driven already. Setting business objectives for such a company should be ambitious and implementation should be pretty straight forward, given that other conditions are favorable such as access to data, rights to data, infrastructure setup and management approach. For a company in the maturity phase or perhaps even entering into the decline phase, setting balanced and achievable objectives for a fundamental change will be trickier.

Given this problematic mix of old and new, a company's approach when transforming itself into a data-driven organization either needs to be aligned with the organization's primary needs for the overall company or for the needs of a certain targeted area. It is possible to transform different business areas at different speeds and different ambitions and even at different times. However, depending on the integration and dependencies between different business areas, that could cause problems related to data and infrastructure dependencies, as well as portfolio offerings and customer communication. An important priority of the assigned CDO should therefore be to determine which approach to take so that it's possible to align it with the overall data science strategy.

The following list shows a few examples of the approaches that could be adopted when becoming data-driven:

- **The data integrator:** The company should focus its data-driven implementation primarily on a modern, integrated, internal data infrastructure designed to bring onboard new and more data that it can use to achieve business objectives related to the different ways it could monetize its data across the business.
- **The business optimizer:** A company committed to optimizing its current business should focus primarily on exploiting the currently available data in order to make internal and customer-centric business processes as effective and efficient as possible.
- **The market disruptor/innovator:** For a company that has the ambition to become a market innovator, the focus should be on augmenting human capabilities using machine learning and artificial intelligence techniques. That will lay the foundation for the company to become a digital market disruptor.

Developing a Data Strategy

After your company's objectives have become clearer, your CDO, as part of an overall data science strategy, needs to create a business-driven data strategy fleshed out with a significant level of detail. In addition, that person needs to define the scope of the desired data-driven culture and mindset for your company and move to drive that culture forward. In this section, I spell out what a CDO needs to keep in mind in order to accomplish these tasks, as well as an example of a data strategy scope.

Caring for your data

One key aspect in any data strategy involves caring for your data as if it were your lifeblood — because it is. You need to address data quality and integration issues as key factors of your data strategy, and you need to align your data governance programs with your organizational goals, making sure you define all strategies, policies, processes, and standards in support of those goals.

REMEMBER

Organizations should assess their current state and develop plans to achieve an appropriate level of maturity in terms of data governance over a specific period. It's important to recognize that data governance is never complete; by necessity, it evolves, just as corporate needs and goals, technology, and legal and regulatory aspects do.

Governance programs can range from establishing company-level, business-driven data and information programs for data integrators, to establishing customized, segment-based programs for the business optimizers and market disruptors/innovators. However, even the best strategy can falter if the business culture isn't willing to change. Data integrators flourish in an evidence-based operational environment where data and research is used to establish a data-driven culture, whereas business optimizers and market disruptors/innovators need to adopt a "fail-fast" agile software development culture in order to increase speed-to-market and innovation.

Democratizing the data

As important as it is to understand the value of the data your company has access to, it's equally important to make sure that the data is easily available to those who need to work with it. That's what *democratizing* your data really means. Given its importance, you should strive to make sure that this democratization occurs throughout your organization. The fact of the matter is, everyone in your company makes business decisions every single day, and those decisions need to be grounded in a thorough understanding of all available data. We know that data-driven decisions are better decisions, so why wouldn't you choose to provide people with access to the data they need in order to make better decisions?

WARNING

Although most people can understand the need for data democratization, it isn't at all uncommon for a company's data strategy to instead focus on locking up the data — just to be on the safe side. Nothing, however, could be more devastating for the value realization of the data for your business than adopting a bunker mentality about data. The way to start generating internal and external value on your data is to use it, not lock it up. Even adopting a radical approach of a totally open data environment internally is better than being too restrictive in terms of how data is made available and shared in the company.

Driving data standardization

A third key component in any data strategy is to standardize to scale quickly and efficiently. Data standardization is an important component for success — one that should not be underestimated. A company cannot hope to achieve goals that assume a 360-degree view of all customers underpinned by the correct data without a common set of data definitions and structures across the company and the customers.

TM Forum, a nonprofit industry association for service providers and their suppliers in the telecommunications industry, developed something they call the Information Framework (SID) in concert with professionals from the communications and information industries working collaboratively to provide a universal information and data model. (The SID part of the name comes from *S*hared *I*nformation *D*ata model.) The benefits of this common model come from its ability to significantly support increased standardization around data in the telecommunications space and include aspects such as;

- Faster time to market for new products and services
- Cheaper data and systems integration
- Less data management time
- Reduced cost and support when implementing multiple technologies

REMEMBER

Organizations have long recognized the need to seek standardization in their transactional data structures, but they need to realize the importance of seeking standardization in their analytical data structures as well. Traditional analytics and business intelligence setups continue to use data warehouses and data marts as their primary data repositories, and yes, they are still highly valuable to data-driven organizations, but enabling dynamic big data analytics and machine learning/artificial intelligence solutions requires a different structure in order to be effective.

Structuring the data strategy

The act of creating a data strategy is a chance to generate data conversations, educate executives, and identify exciting new data-enabled opportunities for the organization. In fact, the process of creating a data strategy may generate political support, changes in culture and mindset, and new business objectives and priorities that are even more valuable than the data strategy itself. But what should the data strategy actually include? The list below gives you an idea.

- Data centric vision and business objectives including user scenarios
- Strategic data principles, including treating data as an asset
- Guidelines for data security, data rights, and ethical considerations
- Data management principles, including data governance and data quality
- Data infrastructure principles regarding data architecture, data acquisition, data storage, and data processing
- Data scope, including priorities over time

WARNING

Don't mix-up the data strategy with the data science strategy. The main difference is that the data strategy is focused on the strategic direction and principles for the data and is a subset of the data science strategy. The data science strategy includes the data strategy, but also aspects such as organization, people, culture and mindset, data science competence and roles, managing change, measurements, and business commercial implications on the company portfolio.

Establishing a Data-Driven Culture and Mindset

An obvious but vital step in becoming data-driven is to take the time to get the employees onboard in terms of what this fundamental change really means in their day-to-day business activities. It will take some time and effort to get there, but not only is it worth the time invested, it's also the main prerequisite for change to happen and to last over time.

TIP

In the early stages of introducing a data-driven mindset, focus on explaining what's happening with the help of examples close to what is already being done. Start with concrete examples of what the changes will actually mean in their daily setup. For example, if the current ways of working in a company are strongly *reactive* — meaning that a process starts with a customer complaint — what would the new starting point be? How would a data-driven and proactive approach impact current workflows, practically speaking, when the flow starts instead with predictive analysis of the data and the ambition to prevent complaints?

On a day-to-day basis as part of ordinary decision-making, leaders must actively encourage employees to a) establish the habit of asking for the data input they need and b) make use of the data at their disposal. A practical and clear request by management to actually use the available data will ripple through the organization and will have a much greater impact than you might think. It's truly an

important step to establish data-first thinking in the company. Furthermore, this effort could be underscored in even stronger terms by establishing a system of rewards for those employees who promote and drive the culture around using data as the primary input for the decision-making.

Recently, word has been going around that "analytics are getting easier," and in some respects, this is true. Some quite capable off-the-shelf analytics tools are available, and certain new self-service analytical applications have removed some of the complexity, making data analytics available to more people with different types of roles — and hence supporting a data-driven culture.

However, although analytics may be getting easier, data management is becoming harder. This is mainly because of the growing variety of sources and structures and the ever-increasing velocity at which it is generated. Therefore, it's vital to consider having the right talents and skills supporting data engineering aspects, as well as data science as a whole, and it's also important to keep in mind that this state of affairs will remain so for the foreseeable future.

IN THIS CHAPTER

- Embarking on the road of digitization
- Executing the data-driven approach
- Making operations more efficient through automation
- Realizing the full value of being machine driven

Chapter 9

Evolving from Data-driven to Machine-driven

To be driven by data and to be driven by a machine are not the same things. You can call them *related* states in data science, if you want, but they definitely exist at different stages of maturity. Being *data-driven* above all else refers to the idea that any progress in an activity is bound by data rather than by intuition or personal experience. Being *machine driven,* on the other hand, refers not to the boundaries of an activity, but rather to how that activity is carried out — more specifically, that it is an activity connected, automated, and controlled by a machine and its implementation of a certain model or algorithm. Machine driven is the final stage of industrializing and automating data science from end to end, driven by the designed intelligence of the machine.

Today, few companies and organizations are at a stage where they can be said to be fully data-driven or, definitely, fully machine-driven. However, several companies are now starting to explore how to make parts of their businesses machine-driven as a first step toward full transformation at a later stage. This approach is not only possible but also advisable for early experimentation on the potential efforts, benefits, and impact.

When you're starting to experiment with consolidating parts of your business into a more machine-driven approach, you need to be aware of four major steps — all with interdependencies, but with each and every one of them requiring the same starting point: digitizing the data.

Figure 9-1 drives that point home — always digitize the data first. Without that step, none of the other steps can be taken. However, after you've digitized the data and made it available, you can try any of the following steps in any order (although I do recommend following the prescribed sequence in Figure 9-1 for both a full, companywide scope and selected parts of the business). Figure 9-1 gives you the roadmap, and the rest of this chapter describes in detail the steps along the way.

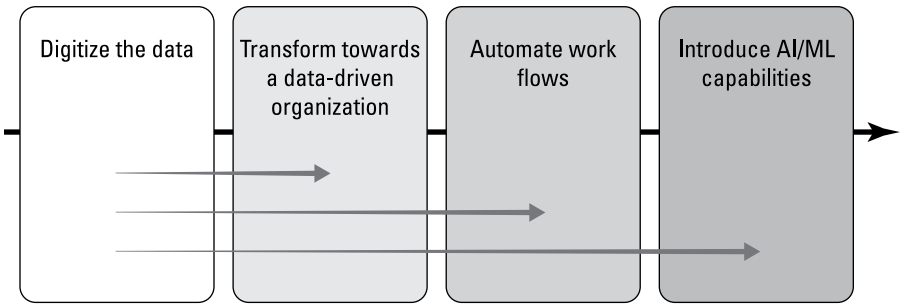

FIGURE 9-1: On the road to a machine driven approach.

Digitizing the Data

When you digitize data, you're transforming data assets into digital, machine-readable content. Simply spoken, without digitalizing the data in your business, you can't take the next step toward becoming machine driven.

TECHNICAL STUFF

Digitization of data (sometimes also referred to as *digitalization* which I discuss in Chapter 1 in relation to a data-driven organization) is the process of converting information into a digital (computer-readable) format. The output is a digital representation of an object, an image, a sound, a document, or a signal (usually, an analog signal), and it includes the change of analog source material into a numerical format. The result is the digitized data in the form of binary numbers, which facilitate computer processing and other machine led operations.

Digitizing data is of crucial importance to data processing, storage, and transmission because it allows information of all kinds and in all formats to be carried with the same efficiency. Digitized data, unlike analog data (which typically loses quality each time it's copied or transmitted), can (in theory) be circulated an indefinite number of times with absolutely no degradation in quality.

So, which assets am I referring to that need to be digitized? How about *everything?* That means everything from the company's customer data, internal routines, and processes and workflows to employee information, product information, and other types of information. Digitizing all this data opens up the possibility of using the available data as input for data analysis driving internal efficiency improvements and new business opportunities as part of a data-driven approach.

Applying a Data-driven Approach

After the data is transformed into a digital and machine readable format, you can start taking steps toward making your business data-driven from end to end.


REMEMBER

As described in Chapter 8, introducing a data-driven approach is not a minor tweak to current operations; for it to be successful, you need to carry out a fundamental transformation of your organization from end to end, starting with top management buy-in. After the decision has been made to carry out this transformation and it is generally understood what it will mean in terms of not only the expected business benefits but also the required efforts, the first step is to create a solid and well-thought-out data science strategy. This strategy will help you plan and execute all needed activities in order to meet your overall business objectives. If the long term objective includes becoming machine driven, then automation, machine learning, and artificial intelligence aspects need to already be included in the data science strategy at this stage.

When a data science strategy has been defined and agreed on, the hard work to transform the company to a data-driven one then begins in earnest, including aspects such as enabling and using data, establishing a data-driven mindset and culture across the company, and making sure that data is requested and used in decisions and measurements across the entire business.


TIP

To prepare the organization for not only an enhanced data understanding but also a necessary change in mindset and culture, spend some time investigating and experimenting with different machine driven solutions for selected areas. Doing so helps prepare the organization for what is coming and also helps exemplify what it means in terms of competence needed, impact on current ways of working, architecture and infrastructure impact — and also benefits gained. Yes, the starting point for a data-driven organization is that it all starts and ends with data, but for data driven to deliver value, remember that it needs to be understood and implemented in the real context of the company.

Automating Workflows

The step to automate either a part of or all of your processes or workflows can be taken directly after you have digitized your content; it doesn't require that you first turn your business into a data-driven one, even if it's advisable in order to achieve your long term objective of becoming machine driven.

Automating workflows is the first concrete step in transferring control from humans to machines. Remember that, by automating workflows, humans are still deciding on the approach and the steps that the machine needs to take; you're just moving the responsibility of executing the steps to the machine. This method (also called *process automation*) involves using computer technology and software engineering to help systems and processes work better. (The prime examples here are helping plants and factories operate more efficiently, more safely, and at a lower cost in industries as diverse as paper, mining, and cement.)


REMEMBER

Process automation can also be applied at the business level, where it's then referred to as business process automation (BPA). When applied here, it refers to the technology enabled automation of complex business processes. BPA can be used for streamlining many aspects of a business, including enhancing cost efficiencies, achieving digital transformation, increasing service quality, and improving service delivery. A full implementation of BPA usually includes activities such as enabling data, integrating applications, restructuring employees, and applying software applications (machines) to automate the tasks throughout the organization. Robotic process automation (RPA) is an emerging field within BPA, which is taking the next step toward becoming machine driven and can (in more advanced versions) add artificial intelligence capabilities to the machine scope.

Introducing AI/ML capabilities

When you introduce artificial intelligence / machine learning capabilities to your company, you take a significant step toward becoming machine driven. This step includes focusing on what the business objective is and letting go of some human control of when or how certain tasks should be done.


TECHNICAL STUFF

Adding intelligence capabilities to the machine means that models and algorithms are designed to use the data to find and achieve the best possible way to execute the defined objective. Practically speaking, you must set up and industrialize an infrastructure in the company that is both data- and machine-oriented. There must be room for exploratory development environments for data scientists to identify new opportunities and create new models, but you must also provide a stable production environment to enable the algorithms to run as part of the operational reality of the company. At the end of the day, being machine driven is

all about letting the algorithms run and then find and execute the best way to solve or predict and prevent a problem.

THE ROLE OF HUMANS IN AI

Even though being machine driven is all about the machines, don't forget about the human involvement outside of algorithm development. Your main objective here is not to create the smartest machines in the world, but rather to use intelligent machines to augment the human capabilities of your organization so that you can achieve more. Therefore, although the role that humans play in a machine driven business may be different, they're still quite important. The role includes tasks you'd expect, such as monitoring model and algorithm performance to make sure the machine is doing what it needs to do with an expected level of quality, but it also includes directly interacting with the machine.

The point here is that early machine-led activities are usually focused on serving new insights and recommending actions, not on running fully closed-loop automation where decisions are both taken and acted on entirely by the machine. As the technology matures within an organization, however, more and more of these closed-loop, machine driven scenarios will start to come alive across different businesses, especially where you can limit the algorithm scope and complexity. Examples of these types of machine led activities include closed systems designed for particular outcomes in industries such as mining, manufacturing, and smaller IoT systems.

Another important aspect from the human perspective is the level of competence needed and the diverse skill sets required to not only implement a machine driven approach but also design the data science strategy to account for all necessary aspects and dimensions. Unfortunately, the access to experienced data science expertise in the area of full end-to-end machine driven business automation is limited, to say the least, especially when it comes to more complex AI techniques in the cognitive reasoning space.

However, when it comes to the availability of data scientists in a general sense, it is increasing slowly. The interest in various universities across the world is growing, and so is the availability of data science university programs and other types of data science educational programs.

However, because few up-and-running, end-to-end machine driven businesses are in play today, it's even more important to explore the area of machine driven business in small and manageable steps. This will allow competence to slowly grow and spread in the company over time — and allow curiosity to lead the change instead of imposing it from the top down. Enforcing data science in a company without fully understanding either the impact of the efforts needed or the benefits to be expected is not a good way to get started.

3 Building a Successful Data Science Organization

IN THIS PART . . .

Explaining what comprises an efficient data science team

Exploring different options for an organizational setup

Motivating the role of the chief data officer

Identifying and hiring needed resources and competence

IN THIS CHAPTER

- Defining the data science leadership
- Understanding the team preconditions for success
- Forming a team
- Establishing the business purpose

Chapter 10
Building Successful Data Science Teams

Data scientist has become the most attractive role in today's competitive job market. Entry level salaries can range into six figures, and roughly 700,000 job openings are projected by 2020. What's driving this demand? Simple — business value. The job of the data scientist is to extract insights hidden inside mountains of data — insights that can then be used to achieve diverse business goals, ranging from fraud detection to facial recognition. Acknowledging this diversity at the core of data science is the key to building efficient data science teams, which must be composed of individuals with highly specialized and complementary skill sets in order to be successful. But before you can embark on the journey to get the perfect data science team in place, you need to make sure you have the right leadership in place.

Starting with the Data Science Team Leader

When hiring a data science team leader, keep in mind that you're not hiring a data scientist — you're hiring a *leader* in data science. There's a difference. First, remember that the leadership skills are more important than the expert skills in

data science. That doesn't mean that the manager can't be a former data scientist, but it does mean that the person needs to be ready to take on a leadership role.


WARNING

If you appoint a leader who doesn't understand the basic area of data science and how it differs from ordinary software development, building an efficient and successful data science team will be slow going, in both the short term and long term.

Remember that leadership is a skill in and of itself. Just because someone has proven to be a successful team contributor in the past doesn't mean that he has the skills necessary to retain and develop great talent while delivering valuable insights, products, and outcomes to your customers and back to the organization. Great data scientists have loads of career options and won't tolerate bad managers for long. If you want to retain great data scientists, a good starting point is to commit to having great managers.


REMEMBER

If the leader is a former software development manager, it is fundamental to understand that being a leader in data science is different. A minimum requirement in that situation is that the manager of the data science team knows that he or she is not knowledgeable enough in data science and thus needs to keep an open mind and wants to learn more. That situation also benefits from having experienced data scientists on the team that can support the leader in the beginning.

Adopting different leadership approaches

Understanding how to lead data science teams compared to pure software development teams is fundamental for becoming a successful leader in this domain. In this section, I present a list of areas and aspects pinpointing the main differences between pure software development teams and data science teams — differences that must be taken into account as part of leading these teams:

- **Team working methods:** In data science, the methodology differs from what you'd find in traditional software development. Data science requires much longer development cycles to get to a result and validate that result, because of its dependency on acquiring and preparing the data to a sufficient level of quality before any type of analysis can be done. Because data science is also still in an early phase of its evolution, it's still in a state of flux. That means techniques and methodologies are constantly evolving, which calls for a high degree of experimentation in the approach.
- **Technology and techniques used:** In data science, you need to use technologies not often used in software development (statistics, machine learning, and artificial intelligence, for example) in order to create dynamic, self-learning algorithms and dynamic implementation environments. Areas such as

robotics are also becoming more and more important in the area of data science. These technologies require other techniques, programming languages, and tool sets on top of what is used in traditional software engineering, depending on the data science strategy and focus. More ambitious and complex machine learning/artificial intelligence modeling and solutions require even more advanced tool sets.

- **Competencies required:** Data science is its own area of competency, with its own roles and tool sets. A former software developer can become a good data scientist, but it requires a different skill set —one related to the statistical models, with more advanced math competencies, for example. There are also other important roles to consider, such as data architect, data engineer, domain expertise, and automation engineer.
- **Infrastructure needed:** When dealing with data science, the infrastructure focus is also different because everything is centered around the data. It's all about having a capable infrastructure that can support efficient data capture, anonymization, transfer, legal compliance, data security and governance, ethical considerations, storage, and processing before you even start working on analysis and model development. As you can imagine, the number and variety of data science needs place high demands on the infrastructure capacity and its capability to enable the analysis and exploration of huge data sets while maintaining data quality and integrity.

  As a comparison, a pure software development infrastructure generally more self-contained with fewer external dependencies. It has also less need for scalability in terms of compute power and storage. Another aspect is that the software development area is also much more mature from a software ecosystem perspective, which means that it's far more streamlined and standardized than data science, which is still in an experimental and exploratory mode.
- **Other important considerations:** In data science, the ethical-, privacy-, and security-related aspects of data usage and model performance are of the utmost importance when it comes to understanding and managing a data science team. If due diligence isn't carried out and any ethical or regulatory requirements were to be violated, all data science activities would have to come to a screeching halt. In addition to the costs of downtime, you could also be facing huge fines associated with the breaking of such laws and regulations. As a leader, it's therefore vital to make sure company policies around such matters are built into strategies, guidelines, systems, and control mechanisms during the entire life cycle of the data as well as the models.

Approaching data science leadership

A fundamental aspect of becoming a great data science leader is to think through how to generate trust, authenticity, and loyalty within the team and toward the leader. This may well be true for all team/leader relationships, but it's nevertheless especially true for data science, where companies are still very much confused about its role in the organization. This means that the data science leader is responsible for protecting team members from unreasonable requests and for explaining the team's role to the rest of the organization. Your team needs to trust that you will "have their back."

WARNING

Having your employees' backs doesn't mean blindly defending them at all costs. It *does* mean making sure that they know you value their contributions. The best way to do that is to make sure your team members have interesting projects to work on and that they aren't overburdened by projects with vague requirements, strange use cases, or unrealistic timelines, accompanied by insufficient data and an inadequate data science environment.

To build trust over time, the data science leader/manager should invest in openness and honesty. Data scientists are competent people who are trained to analyze and handle information. Be transparent during the entire data science process, including recruiting, onboarding, ensuring daily operations, and discussing the team's performance and focus — as well as the overall company strategy. Being sincere and open in all aspects can be painful, but is vital for team success.

TIP

Make feedback consistent and bidirectional. Great data scientists excel at spotting whether you mean what you say. If you ask for feedback but do not intend to act on the feedback, you will soon discover that your best employees may want to leave.

WARNING

Valuable employees seldom leave a company because they're unhappy with the company itself. They leave because of poor leadership.

Finding the right data science leader or manager

So, how do you approach finding the right data science manager to build a successful team? One approach is to use the same test for the manager candidates as you do with data scientists. It should include the same challenges, but you should expect different outcomes or results from the test.


TIP

Know what a good result looks like. Thoroughly think through the actual level of proficiency you require from the data science leader. For instance, if you're hiring a manager for a machine learning team, let the manager solve a specific data science task — one that deals with image similarity, for example. A good outcome of that exercise is for the manager to be able to explain the different techniques needed to solve the task, starting with deep learning and moving on to other machine-learning strategies. The idea here is to gauge the breadth of the candidate's knowledge, even if you're not asking them to produce actual models or code a solution.


TIP

I also recommend implementing a two-tiered leadership for data science teams, where one person will be the line manager demonstrating an understanding of the overall business objectives and accepting the responsibility for the team's success, while another person will be the data science expert who reports to the overall line manager. This type of setup creates a dynamic environment that combines a deep technical data science leadership with a good business orientation.

The two levels of leadership will not always agree, but that is the purpose of the setup. The business-oriented manager will challenge the technology manager, and the technology manager will challenge cooperate decisions. It is of the uttermost importance that these two leaders are able to cooperate closely and openly with each other, with a great amount of trust. This needs to be tested as early as possible in the hiring process.

Defining the Prerequisites for a Successful Team

You can often come up with quite specific technical requirements for each role within the data science organization, but there needs to be a common understanding of what is required for a data science *team* to be successful. While technical skill sets are vital for a successful data science team, there is a far more critical set of success factors for a data science team that you need to be aware of. The next few sections walk you through them.

Developing a team structure

How the data science team is structured is vital to the effectiveness and efficiency of the team. The importance of close cooperation between data engineers and data scientists shouldn't be underestimated. Usually, these two different roles are not on the same team, which means it's even more crucial to secure an efficient

cross-team collaborative environment. Such a collaborative environment should be able to handle all analytical processes end-to-end across different systems and organizations, and ensure that productivity is not lost along the way.


REMEMBER

As you might have guessed, maintaining these lines of communication is never an easy task. You should nevertheless strive to minimize the inevitable hand-offs between data engineers and data scientists, in the process creating a seamless workflow from data capture to the deployment of models into production. It also drives efficiency if data engineers and data scientists are able to share insights with the business users from the same data system environment.

Establishing an infrastructure

The data science infrastructure needs to be highly scalable — that is, it should enable the data scientist to focus on analyzing data and building models, not struggle with acquiring data or computing data. Implementing a secure-but-enabling infrastructure is the key to success with all your artificial intelligence investments — including the investments you've made in building a successful data science team.


REMEMBER

When approaching artificial intelligence infrastructure work, many companies tend to forget to define a clear data science strategy that is agreed upon across the company before getting started. Remember that a poorly-thought-out infrastructure strategy will inevitably increase complexity and cost in development and operations (DevOps). If you aren't following an infrastructure plan, approaching it in a more build-as-you-go fashion, the complexity can be more difficult to handle in terms of setting up and maintaining the infrastructure over time, managing upgrades and fixes, scaling the infrastructure with growing data volumes as well as in providing high-performance infrastructure for large teams of data scientists, sometimes spread out over a large geographical area.


WARNING

When it comes to increasing data processing efficiency, distributed computing is often presented as the best solution to ensure performance at scale. But be aware that the complexity of managing distributed computing, from cloud-edge to edge on device or component level, requires special skills, which can be difficult to get hold of.

Ensuring data availability

Enabling access to the data that's needed may seem like an obvious task to prioritize, but you might be amazed at how often this is a major concern for data scientists. Sometimes access issues are caused by external factors that are more

difficult to manage, such as new legal constraints on data usage, but many times the issues arise because the company has no common and agreed-on strategy for enabling access at a consistent level.


WARNING

It might seem a simple task to ensure that the data is available for data scientists to use in performing their analyses and building their models — much simpler, for example, than chasing data and mending broken data environments. The task is actually far from a simple one. In fact, unreliable data availability is a common source of frustration among data scientists, making them leave for other companies that are better at meeting the data needs. This is really an unnecessary risk to put on your company, so make sure you don't end up in that situation. If you manage to get hold of one of the few senior data scientists who are available on the market, do whatever is needed to retain that person!


TIP

One way to minimize the risk is to lower your ambitions and start with data that is already available to you. You can then work in parallel to ensure that you have all the necessary rights to the data, that the data is secure and governed correctly, and that you have the means to collect and prepare the data for the data scientists.

Insisting on interesting projects

One important aspect to take into account when trying to ensure the success of a data science team is to offer your data scientists interesting challenges and difficult problems to solve. If your problems are too simple, senior data scientists will become bored. If the problems are impossible to solve because of lack of data or a malfunctioning environment, the data scientists might stay a bit longer and try to fix it, but eventually they will leave if they aren't supported.


REMEMBER

Data scientists have the most sought after competence in the market right now. They know their value, and so should you. They won't waste their value on a company that cannot offer the right opportunities for them to be part of a successful data science team, working on interesting and challenging projects. So, if you're after the experts, you must be ready to have interesting ideas and complex problems to offer them.

Promoting continuous learning

Because data science is an area under constant change, data scientists want to (and should) stay updated on new techniques, methods, and trends appearing in the market. There must be opportunities available to all roles in data science to continue to learn as part of their job.

TIP

One way to enable continuous learning is to let data science team members participate in different technical data science conferences or open source projects or venues. Formal training doesn't necessarily keep up with the pace that's needed, and data science is still very much driven from an open source perspective, with regard to technology as well as methodology.

Encouraging research studies

TIP

Data science team members should also be allowed to spend time on research, including writing white papers and participating in white paper reading sessions based on work done by colleagues. Because the data science area is evolving *fast*, it's crucial to stay in tune with new methods, research, and technology developments as well as contribute to standardization discussions and open source initiatives.

Building the Team

When building a team of data scientists, you should always focus the scope and purpose around questions that reflect your company's strategic business objectives. This focus could mean, for example, attracting new customers, automating processes, or introducing new, innovative data products in the portfolio.

You want to be able to get your stakeholders and decision-makers on board with your data science ambitions as early as possible and be able to argue for the return on investment (ROI). You should, for example, consider these questions:

- What will drive an optimal outcome, and what are the incentives?
- How will the data science team work with stakeholders?
- How are investments in infrastructure approached with regard to priority, approval process, funding, and management?
- How will costs be allocated?
- How will business, legal, IT, and data teams operate without creating unacceptable risk?

REMEMBER

An effective communication strategy, clearly defined priorities, and the ability to manage expectations are vital, regardless of which approach you decide to take for structuring and developing your data science capabilities.

Developing smart hiring processes

Many professionals are trying to break into the "sexiest profession of the 21st century," so, as a data science manager, you're sure to get lots of applications, and you'll have to be selective. Take advantage of that and be picky in the right way. Make sure you care about your hiring process.

WARNING

One common area where companies fail is in the trade-off between the short- and long-term perspectives. For instance, it's easy to start thinking that you are late to the data science game and therefore there isn't enough time to recruit all the people you need. That approach is a huge mistake. If you believe that there isn't enough time to find the right talent and to scrutinize your interview and onboarding processes, you probably don't have the time to manage a new employee, either. Creating a great hiring process pays off in the long term.

So, what does a great hiring process look like? For one thing, it doesn't focus only on technical skills. Social skills, like empathy and communication, are undervalued in data science and the disciplines from which data scientists usually emerge, but they're critical for a team. Make this a part of your hiring process.

TIP

Rather than focus on whether you can get along with a candidate, ask yourself whether there is a lens through which this person sees the world that expands the team's knowledge and value. That dimension should be valued as highly as you value other attributes, such as technical ability and domain expertise. This is why it's important to prioritize diversity. That includes diversity of academic discipline and professional experience but also of lived experience and perspective.

When it comes to diversity, a few areas stand out as especially important in data science. First, you should not be focusing on hiring senior people. Not only are they expensive and difficult to get hold of, but less experienced employees tend to not be so influenced by history and can therefore more easily ask questions about why things are done in a certain way. The questions asked are more free of the usual assumptions that more experienced professionals at some point stop being aware of having. It isn't difficult to become obsessed with a particular way of doing things and to forget to question whether a favored approach is still the best solution to a new task.

REMEMBER

Data scientists come from a variety of academic backgrounds, including computer science, math, physics, statistics, and many others. What matters most is having a creative mind coupled with first-rate critical thinking skills.

Another important consideration is to hire individuals whose strengths complement one another, rather than build a team whose members all excel in the same area. Having a person who always sees the big picture, someone who can articulate stories with data, and a visualization wizard working together can collaborate

to produce outcomes that none could do independently. It's all about taking advantage of these complementary skills to the greatest extent possible in order to create great solutions that nobody had thought of before.


TIP

One way to make sure that the team actually works as a team and collaborates is to ask team members to regularly read each other's code and check each other's models. It's a great way to foster team collaboration centered around technical discussions. Making sure your team engages in these types of planned collaborative activities helps ensure that you get the most out of this sort of team diversity.

Finally, it's important to build a team that reflects the people whose data you're analyzing: For example, if you're analyzing social media data from an application used only by women, the data science team cannot only consist of men. This is the only way to ensure that you have a resilient team that will ask better questions and have a wider perspective from which to ask these questions. This way, each individual's blind spots are covered by another team member's past experiences and skill set.

Letting your teams evolve organically

When you start building your data science teams, you should begin by having small, versatile teams, where each team encourages team members to "wear many hats" and do lots of different kinds of data science. As the teams mature and prove their value in different ways, roles will become more defined and some activities will likely move to other teams. An example of that is that once data science is more established and understood in the company, activities related to infrastructure, operations, security, and so on are usually handled separately, enabling the data science teams to be more specialized.

And yet, I'd warn against specializing teams too soon. Team specialization works only when team responsibilities are clearly defined and efficient ways of working have been put in place to balance the lack of speed and additional costs associated with multiple teams working together. Because accomplished data scientists are hard to find, let the teams organically evolve into more specialization. (My mantra is "Specialization will come; no need to rush it.") And don't forget to pair up experienced data scientists with less experienced ones. It's usually easy to find smart and driven data scientists who (with a little dedicated coaching) are eager to learn more. This includes learning how to appropriately frame a problem, manage a small project, develop and train a model, integrate with APIs, and put a project into production. And, if you eventually manage to acquire a couple of accomplished data scientists, you already have the structure in place to integrate them into these learning teams to make sure you can get the most out of their vast experience.

Connecting the Team to the Business Purpose

Data scientists are generally purpose driven, so to get the most benefit from their time, they need to have a clear understanding of the task at hand and believe in the business objective behind the projects they're assigned to. There must also be room for prioritizing the ideas coming from the data science team itself. Anchoring your team's work in the context of the data science strategy and the overall business objectives are among the most important jobs a leader of a data science team has to accomplish. Unfortunately, it isn't always an easy job to do.

A data science project often starts with a question from someone outside the team. Often, however, the question that the person asks isn't exactly what the data science team feels should be investigated. Therefore, managing a data science team usually involves a lot of discussion and fine-tuning of questions from stakeholders to better understand the information they actually want and how it will be used.


TIP

Don't let questions or requests become projects for your team until you know exactly what the stakeholder wants to understand and how it will be used. Having clear objectives for the data related questions that come your way is one of the most important things you can provide for your team.

At the same time, stakeholders won't always be able to answer the questions from the data science team, even if they want to. They might not know the full context or long-term objective of the question they're asking, what a finished data science product would look like, or even how they would apply it. To overcome this hurdle, try some of the activities described in this list:

- **Understand business value**. Ensure that product managers understand the business value of data science. They need to understand it in terms of how it drives business innovation and value and be ready to prioritize it even when they might not understand all aspects of it.
- **Participate in strategy meetings.** Make sure members of the data science team are regularly invited to product and strategy meetings. This way, they can be part of the creative process rather than merely responding to requests. (It also doesn't hurt to ask a lot of questions.) Integrating data scientists into the business dialogue also contributes to an increased company understanding of the data science approach to business opportunities.

- **Collaborate across organizational borders.** Clearly, data science teams aren't the only ones seeking collaboration with business stakeholders, but it is nevertheless the case that they should be given priority when decisions are made concerning who collaborates with which teams across business divisions. It significantly increases motivation for the data scientists when they know not only that they are contributing to something that they understand but also that their contribution is valued and prioritized by the organization as a whole.
- **Prove the value.** Data science teams need to be able to prove their value. If the team comes up with a new idea, the organization and the data science manager must challenge the team with questions such as "How can we prove that this is contributing to our business?" and "How do we know that this is the best solution?"

IN THIS CHAPTER

- Describing different ways to organize data science
- Assessing a center of excellence
- Implementing a common function for data science

Chapter 11

Approaching a Data Science Organizational Setup

The power of the data revolution remains strong, and companies of various sizes are actively building and expanding their data science teams in a variety of ways. Companies realize that they must be able to use data to improve decision-making and operational efficiency, but they also see that they must have the capacity to create new products and processes based on data-driven insights. In order to do so, companies must embed at the corporate level the necessary organizational and cultural changes it will take to succeed.

The organizational structure needed for the data science team varies based on the size of the company, the number of different business functions there are, the geographical distribution, the company culture, and other similar aspects. However, there are some common factors to consider when integrating data scientists into a larger data-driven organization. This chapter takes a look at these factors.

Finding the Right Organizational Design

To figure out the best team setup for the data science function from an organizational perspective, consider these five main tasks:

- **Decide where in the organization you want to place the data science function, and then figure out the optimal setup.** For example, do you want a centralized model or a decentralized one?
- **Align the organizational design with its business functions.** As part of that process, you define and implement an efficient governance structure based on openness and transparency toward the business, and you have to build in decision-making structures that allow for the business to impact which problems and opportunities the data science teams should prioritize.
- **Ensure that the organizational setup fits with your overall business strategy, including partnership setups.** The structure should also enable data science teams to easily connect to needed data science ecosystems related to data, tools, and models.

  
  REMEMBER

  In data science, being part of the ecosystem is essential. The evolution is moving too fast and is too complex and costly to address all by yourself. Moving fast requires companies to share as well as reuse components and capabilities as part of a data science ecosystem. This involves an organizational approach that allows participating in — and contributing to — open source communities and other open frameworks.
- **Identify which roles you need on the team and how many team members for each role.** Consider how many of the roles can be filled via internal recruitment. This stage includes identifying the needed level of training for the team and ensuring that the company executives engage in the training.
- **Scale the team over the long term.** Consider how that will be approached and what the timelines are. Will external recruitment be part of the scale-up, or will it be based on growing organically, for example. Also consider standardizing processes and governance structures *before* major scale-up in order to better manage growth.

That's the big picture — the bare outline of what needs to happen. The next few sections fill in the details.

Designing the data science function

Time for a closer look at how to approach the first of the main tasks to consider: Determine the setup of the data science function in your organization. The setup options range from a centralized model to a highly distributed model.

A centralized model is sometimes called a *shared model* or a *center of excellence*, where all the data scientists are working together in the same organizational unit. This model encourages collaboration and cohesiveness of the data science team, allowing team members to bounce ideas off each other and get quick help for questions with the data science work. This setup also allows for on-the-job-training for less experienced data scientists.

In large companies, where the business units are financially strong and independent or where the business units conduct different types of business, there's a tendency to instead choose a distributed or decentralized model. In the distributed model, the business units (BUs) themselves hire their own set of data scientists. This allows data scientists to work closely and more continuously with managers, engineers, and stakeholders. In this setup, data scientists have the opportunity to gain valuable domain expertise as well as insights into the real problems that their colleagues face every day. This setup tends to better empower data science teams, but you also run the risk of duplicating some of the work the teams are doing, because coordination and collaboration across teams in different business units tend to be less pronounced.

WARNING

When small data science teams or even single data scientists are embedded into different business units, it can have the side effect of leaving data scientists overly isolated. If data scientists lack data science peers to discuss and elaborate with, the productivity goes down and motivation drops. There is also a tendency to question isolated data scientists more, since data science is not generally understood in traditional software organizations. The best way to mitigate this is to avoid spreading data scientists too thin in the organization.

The *hybrid* organizational model, consisting of one central unit with one or several data science teams combined with multiple specialized teams embedded in business units across several projects, is starting to emerge as the best proven strategy for many companies. From a data scientist's perspective, the main benefit of this model is the ability to work closer to the real business functions, which can mean gaining more knowledge about how things work in practice. Since the hybrid model adopts aspects of both a centralized and decentralized approach, it balances the benefits and drawbacks of those models. In the hybrid model, the central unit serves as a hub to promote sharing and reuse of best practices and propagates those to each of the distributed data science teams.

Another, less commonly used way of setting up a hybrid model is to have all data scientists assigned to specific business units but report into a common centralized data science unit.

Figure 11-1 shows a graphical illustration of the three different ways to set up your data science function in the company. Which one you should decide to go for depends on aspects such as the maturity of your company, but it is also depends

on your data science objective and availability of various data science competencies. If you are setting up a data science function which is more focused on analytics than machine learning and artificial intelligence, for example, the access to required competence is much less of a problem than in the machine learning space. When you can count on having access to all the competence you need, you can design the organization as needed, directly from start. The best model can then be chosen depending on the company size, line of business, and ambition around analytics, rather than being limited by a lack of required data science competence.

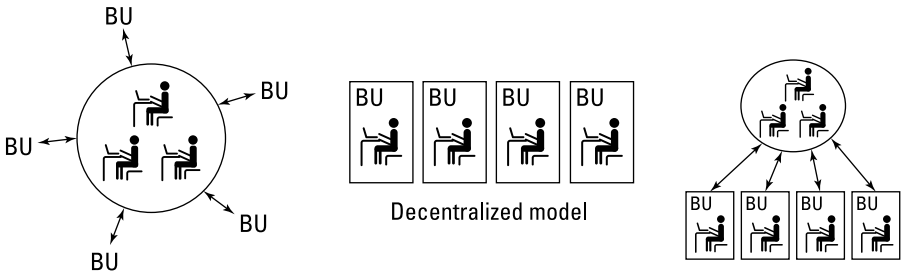

FIGURE 11-1: The different organizational models for data science teams.

REMEMBER

In the machine learning and artificial intelligence space, where the availability of data scientists is scarce and the general understanding of how ML/AI techniques shall be utilized in the company is much less understood, the situation is different. You might come to the conclusion that a decentralized model is what you want, but you realize that with so few data scientists available to recruit, you will spread yourself too thin in the company and not be able to achieve your objectives using that approach. Therefore, starting with a centralized model which offers an adequate number of data science teams working together to make a difference might be the only way to get started. However, over time a centralized model might transform into a hybrid model, as your pool of data scientists grows organically, but also through recruitment. Also, once the area of machine learning becomes more of a commodity in the industry, data scientists can become an integral part of any development unit, and there is no longer any need for the centralized unit keeping it all together.

Evaluating the benefits of a center of excellence for data science

Going for the centralized model for your data science function as your preference or as a necessary starting point really means setting up a central unit in the company, also known as a data science *center of excellence* (CoE). Although there are

different opinions about the effectiveness and efficiency of CoE's in the industry, there are still several proven benefits of using a centralized approach when building up a data science function in the company, as listed below.

- **Speed:** A data science CoE is essential to accelerate the data-driven approach across the business at scale. It reduces implementation times drastically and therefore the time-to-market span needed to deploy new data-driven products and services.
- **Reuse:** A CoE facilitates sharing best practices and methodologies across different teams in the organization.
- **Evolution:** A CoE makes it easier to allocate time for the team to stay updated on market trends and the technology evolution in data science.
- **Skill sets:** The CoE equips the business with the needed set of data science skills when it is needed.
- **Terminology:** A centralized function helps ensure that the entire organization uses a common terminology and "speaks the same language" by developing a common set of standards while deploying data science methods and techniques.
- **Culture:** A CoE can serve as the driver for cultural change to become a data-obsessed and action-driven organization when it comes to using data science techniques.


REMEMBER

It isn't necessary to set up the centralized CoE function strictly according to the centralized model. You can use a light version of the hybrid model and still gain the benefits of a central CoE. Having a 80-20 weighted setup might still work with the data science resource availability — meaning that you have 80% of the COE centrally located and 20% embedded in the business units and then you can let it grow over time towards a 50-50 or even 20-80 approach.

Identifying success factors for a data science center of excellence

If you want to ensure that your CoE delivers real business value to your company, you have to make sure that it succeeds in achieving these three distinct goals:

- **The CoE needs to be seen as an enabler of business value:** In other words, the CoE should be recognized as capable of enabling a deep cultural change around leveraging automation, analytics, machine learning, and artificial intelligence. That should include having the ability to attract the best talent possible and be generally viewed as a driver of value.

- **The CoE needs to be seen as an autonomous entity yet fully supported by management.** The CoE has to function as an independent unit that owns the tools, standards, and methodologies in data science. It cannot be up to anyone in the organization to suboptimize infrastructures and tool sets for specific product or service benefits in various areas rather than for the overall company value. The CoE should also continuously engage supportive senior leadership to further embed data science into the organization as part of its day-to-day business.
- **The CoE needs to be impact-oriented.** The unit should prioritize work on use cases aligned with strategic priorities and utilize a value-driven use case roadmap with a quantifiable impact.

Making sure that your CoE achieves all three goals significantly increases the chance of establishing an efficient and successful data science center-of-excellence function.

Applying a Common Data Science Function

You have several important decisions to make when approaching the establishment of a data science function common across the company. For larger enterprises, geographical location is usually one of the major decisions to address. And, if you start with the important aspect of where the common data science function should be placed, it might seem that co-locating it with the company headquarters could be a good idea. However, because this is a strategically important decision, make sure that it's based on actual data as well as on well-thought-out selection criteria. Let's look into this a bit further.

Selecting a location

In a traditional center of excellence focused on analytics rather than machine intelligence, the actual location usually isn't an important factor. And yet, since a common data science function usually requires not only niche skills, which are difficult to acquire, but also talent availability for future potential as well as industry ecosystem presence to secure alignment and influence on market standards as they are evolving, placing the data science function at the same location as the headquarters might not make the most sense. However, this of course depends on how well the headquarters fulfills the chosen location selection criteria.

You also have to consider other aspects related to location. The following list describes the four main selection criteria for location, listed in a more structured way, including an estimation of their importance in the overall decision:

- **Talent:** Refers to the availability of a new and existing talent pool in actual figures at or near the location in question. This criterion is vital — let it carry about half the total weight of the decision. It includes aspects such as tech-related graduates, the existing talent pool in needed roles and competence areas, and the quality of education in the area.
- **Ecosystem:** Refers to a set of factors that allows for smoother running of data science teams, thanks to qualitative characteristics of locations, such as capacity to retain and attract talent, availability of data scientists and other key competencies, availability of the latest technologies, capacity for innovation, number of start-ups, and potential for venture capital funding. This important criterion should carry about a third of the total weight of the decision.
- **Infrastructure:** This is all about the practical realization of the data science environment, including good connectivity, data links, the availability of the data itself, and the digital channels you can utilize. (I'd weight it around 10 percent.) It's about making sure that the infrastructure is making the data science environment easy to operate and use for employees as well, including aspects like real estate availability, attractive business and living environments, and airports and other transportation facilities supporting good location accessibility.
- **Cost:** This refers to cost indications for mainly three areas — labour and real estate costs, as well as salary inflation. However, this criterion is much less important than availability of the right talent and competence and therefore carries only around 10 percent of the weight of the decision.


REMEMBER

Even if you're planning for a centralized setup for your data science function, you can still distribute that function to several geographical locations. Doing so may even be advisable, especially if it means accessing more and better talent as well as improving the ability to start tapping into existing data science ecosystems. (The fact of the matter is, there's little chance of fulfilling all the criteria in a single location, so you might be forced to branch out anyway.)

Approaching ways of working

After you agree on the location(s) of your data science function, it's time to start thinking of ways to make the data science function work for you. Important aspects to consider are, of course, how to make sure the new function is both effective (doing the right things) and efficient (doing things right), as well as keeping the rest of the organization feeling that it's proving its value.

REMEMBER

If set up correctly, the new data science function will play an important role in your organization by bringing together the different business units and finding clear and agreed-on responsibilities related to the IT function.

Figure 11-2 explains how some of the interactions between the business units and the common data science function could look like.

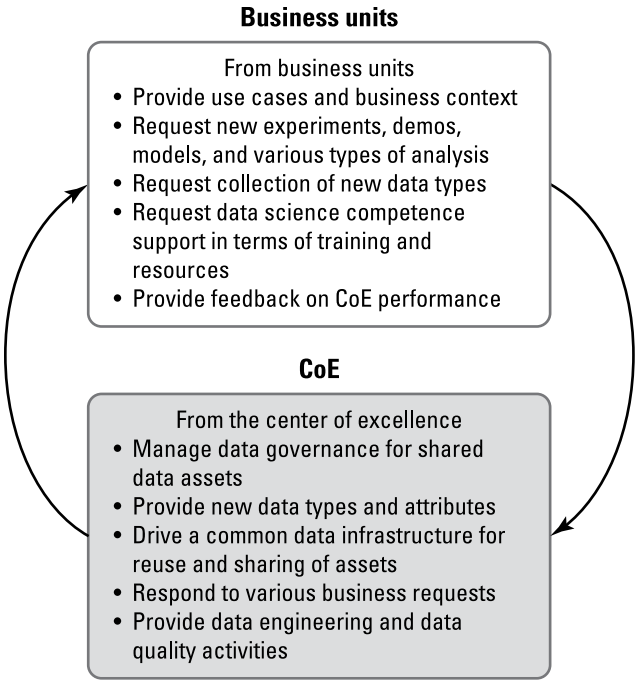

FIGURE 11-2: Example of dividing responsibilities between business units and the CoE.

Figure 11-2 exemplifies how the interaction can work, but keep in mind that you have many aspects to consider when dividing up roles and responsibilities. Don't be too strict when it comes to what's being done by which unit. If senior data scientists are embedded in the business units and are capable of driving certain strategic areas, let them. Empower people to take charge and make a difference using data science. (This is the long-term goal, anyway.) Just make sure that these distributed groups don't become too isolated and disconnected from the common data science function. (One way to ensure the necessary degree of integration is to insist that such groups share models and insights through the common team, including following agreed-on company standards and principles.)


TIP

Encourage employees to enhance their skills in data science by attending recommended training programs and exploring data and use cases in various areas. As the pool of data science employees grows, working together to carry out the data science function should naturally lead to different teams and groups of people finding ways to network and share insights and learnings across business segments.


REMEMBER

Though it's important to make sure that the IT function has a role to play in the new data-driven organization being built, keep in mind that there is always a balance when it comes to involving the internal IT functions. It varies, of course, from company to company, but my personal experience is that the IT function usually wants to do more than it's capable of, often leading to situations where the IT function makes promises they can never deliver on — not because they don't want to, but because they are driven mainly by cost and they lack the business context needed to understand priorities and make the necessary strategic decisions.

Given this reality, clearly separate the tasks that IT is responsible for in the area of data science and the ones the data science team is responsible for. Perhaps you'll have IT focus on the storage infrastructure or certain aspects of data management. Or perhaps you'll want them to focus on operating the data governance model from a system perspective. Yet no matter how you divide up responsibilities, make clear to IT that taking care of their responsibilities in a data science context is not merely a minor IT issue, but rather a crucial part of the overall business operations in the company.


TIP

IT departments tend to be better at operating solutions than at defining and developing them. So, whatever area you make IT responsible for, do not allocate to them the overall responsibility for data science. They simply lack the business competence and data science competence to manage it successfully.

Managing expectations

Another important aspect to consider is to clearly communicate the common data science function's purpose. Everyone needs to understand that the idea is not to have all the data science work be done in that common function. Far from it. For a company to become data driven and focus on data science throughout, it is vital that this is everyone's business in the company.

So, what is the role of this common data science function? First and foremost, it is to secure cooperation across the company, facilitate the sharing of data and algorithms, support the organization with highly skilled data scientists, and provide other relevant competencies to achieve company objectives. All this needs to

happen regardless of whether or not the activities are driven from a centralized function or if competence is injected into the business units to enable and accelerate desired results.


WARNING

You want to make sure that the organization does not perceive this new common data science function as "responsible" for making data science happen in the company. That is not a desirable situation, since it puts the employees in a state of mind where it is no longer a company ambition or everyone's responsibility to make data science happen. The perception then becomes that it is the common data science function's role to make it happen.

Therefore, to clearly state the role of the common data science function is very important. It is a supporting organization, designed to enable the company to succeed with the data science investment. For some companies, this function is of more importance when starting up than in the longer perspective, once data science is more known and the competence is more spread. But for other companies, the common data science function becomes the hub of the data-driven approach, vital for the survival of data science in the company even from a long-term perspective.

Selecting an execution approach

After you've made all the strategic decisions in terms of where the common data science function will be located and everyone agrees on how the new data science function will operate in relation to the rest of the organization, it's time to get started. But where do you start? In my mind, you have only two logical ways to go about it: the big bang approach or the use-case-driven, scale-up approach. The next couple of sections describe exactly what's involved in the two approaches and list their benefits and drawbacks.

The big-bang approach

One way that you can establish the new common data science function is by using the big bang approach, where you directly define and implement the long-term vision and target organization with the desired head count and the full-blown data science infrastructure all at once. This includes the full data scope, data architecture, data governance framework, competence programs and other aspects, supported by a clear rollout plan that's consistent with the company's strategic priorities.

The benefits of this approach are that it communicates a strong company commitment to data science, both internally toward employees and externally toward the market, customers, vendors, and even competitors. This approach is likely to generate motivation and interest among employees, where they feel empowered

and inspired to pursue new opportunities in the area. It's also likely that the investment actually happens, because company management would lose too much credibility if they chose not to honor a clearly communicated commitment.

WARNING

Potential drawbacks of the big bang approach revolve around the idea that the data science organizational setup might be perceived as being forced onto the organization from the top down, without trying it out first or anchoring it in the organization. It's also a huge investment to make up front, before any value-added of the data science investment has been proven. On top of that, the big bang approach makes it difficult for the new common data science function to find time to prove its value and gain support from the business functions when it's stuck on defining a data science strategy, hiring the necessary data science specialists, building up a solid infrastructure, securing data, and identifying and prioritizing first cases of interest.

The use-case-driven scale-up approach

If the big bang approach doesn't fit your company's needs, perhaps you can introduce the new common data science function by using the use-case-driven, scale-up approach. This more cautious approach starts with a detailed design of the strategic steps to take, in which order, and focused on the type of value. This approach enables you to strategically select cases that will prove the value both internally (for the employees, securing commitment and empowerment) and externally (generating external commitment by proving how real value can be achieved using data science in areas that had previously proven to be intractable).

The benefits from the use case-driven approach stem from the fact that the upfront investment is much less and you also get to prove the use value of a data science function on a case-by-case basis *before* scaling up for the next step. You can pick and choose between short-term benefits or high-value use cases. Finally, you gain room to breathe by setting a slower pace, which means that you have time to anchor the setup and approach among employees, which means that they can become more involved in defining requirements and strategic priorities. You can also stop at a less-advanced level of the setup without wasting time or money, if that turns out to be better for the company.

WARNING

In the use-case-driven approach, you run the risk of never proving the worth of a data science function sufficiently enough to stakeholders to actually get the common data science function you need. You get stuck in proving case after case and never gain the final approval, which means that the company never gains the benefits of scale, where a unit can drive a common data science strategy that promotes the sharing and reuse of data, models and insights across the company.

IN THIS CHAPTER

- Defining the CDO role
- Justifying the need for a CDO
- Getting started with a CDO role for your company
- Predicting future CDO scenarios

Chapter 12

Positioning the Role of the Chief Data Officer (CDO)

Assuming that there is strategic agreement on setting up a common data science function in your company, who will then be the strategic spokesperson for all data science efforts? Who will ensure that data science is understood and included on corporate leadership's agenda? It's more than giving a boardroom presentation now and then in order to make sure that the top brass has an idea of what's going on; it's about making sure that data science becomes a vital part of the everyday agenda. Here is where the CDO role becomes extremely important.

The *chief data officer (CDO)* is the corporate officer responsible for overseeing a range of data related functions to ensure that your organization gets the most benefit from what could be its most valuable asset — its data. The position's scope includes enterprise-wide governance and the utilization of data and information as an asset, including all aspects related to data architecture, data management and governance, data utilization, and data commercialization realized through the data science function or functions, regardless of whether it's through a centralized, decentralized, or hybrid setup.

However, the actual placement of the CDO in the corporate structure is not a given, and many examples exist. Although the significance of the role is rising with increased understanding of the value of data, it's still rarely the case that the CDO reports directly to the chief executive officer (CEO). Usually, the CDO is linked to the functions of the CIO (chief information officer), CTO (chief technology officer), or even CSO (chief strategy officer) or CMO (chief marketing officer).

Scoping the Role of the Chief Data Officer (CDO)

To describe the scope of a CDO, you first need to determine how the position relates to that of a chief analytics officer (CAO). Although the CDO and the CAO are two distinct roles, these two positions are customarily held by the same person or else only one role, the CDO role, is used, but when these roles are combined into one, it is sometimes also referred to as a CDAO role. However, in situations where these two roles are separate and held by two different functions, the main difference can be summarized by the title itself: data versus analytics. The main difference in the area of responsibility is captured in Figure 12-1.

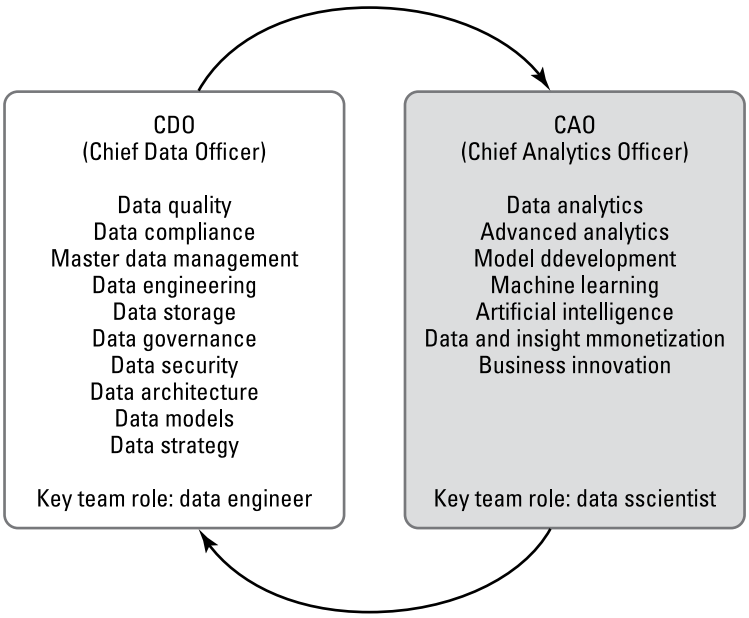

FIGURE 12-1: Comparing CDOs and CAOs.

If the CDO is about data enablement, the CAO role is about how you drive insights from that data — in other words, how you make the data actionable. The CAO is much more likely to have a data science background, and the CDO, a data engineering one.

Let me clarify that both the CDO and CAO positions are essentially carve-outs from the traditional CIO job in the IT domain. In the case of the CDO role, the CIO may well have welcomed eliminating some of these responsibilities. However, when it comes to the portion of the CIO role that is about IT cost for new data assets, the CIO can be deeply challenged by the new realities of big data. Both the CDO and CAO would need to argue for initially storing huge amounts of data, even if its value isn't immediately evident. These aspects pose a significant but important change in mindset for the CIO role, one that probably would not have been recognized the same way without the introduction of the CDO and CAO roles.

REMEMBER

When it comes down to the practical implementation of this role, it's all about securing an efficient end-to-end setup and execution of the overall data science strategy across the company. Which solution can function as the most optimized setup for your company will depend on your line of business and how you're organized. Just remember to keep these two roles working closely together, including the teams that are attached to the roles. Separation between data engineering teams and data science teams is not advisable, especially since there is a need for a strong common foundation based on these two parts in data science. The teams may have a different focus, but they need to work closely together in an iterative way to achieve the speed, flexibility, and results expected by the business stakeholders.

In cases where the CDO role is the only role in a company — where CDO and CAO responsibilities have been merged, in other words — the mandate of the CDO role is usually described in terms of Figure 12-2. The area referring to the business mandate refers mainly to driving areas such as:

- Establishing a company-wide data science strategy
- Ensuring the adoption of a dominant data culture within the company
- Building trust and legitimizing the usage of data.
- Driving data usage for competitive advantage
- Enabling data-driven business opportunities
- Ensuring that principles for legal, security, and ethical compliance are upheld

When it comes to the technology mandate, the following aspects are usually included:

- Establishing a data architecture
- Securing efficient data governance
- Building an infrastructure that enables explorative and experimental data science
- Promoting the continuous evolution of data science methods and techniques
- Designing principles for legal, security, and ethical aspects
- Securing efficient data and model life-cycle management

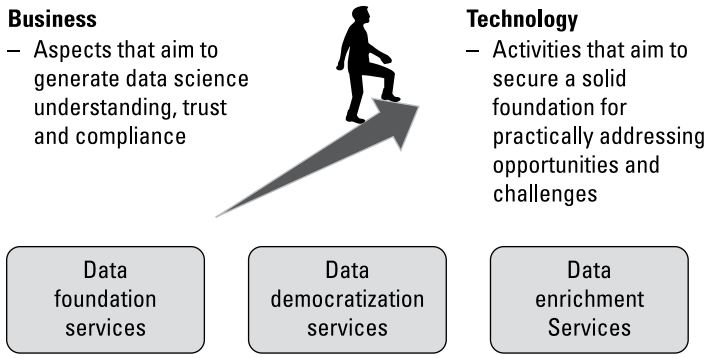

FIGURE 12-2: The mandate of the CDO role, when it includes CAO responsibilities.

Notice that, in Figure 12-2, I highlight three distinct services a merged CDO/CAO must manage. This list gives a sense of what each service entails:

- **Data foundation services** includes areas such as managing data provenance and data stewardship, data architecture definition, data standards, and data governance as well as risk management and various types of compliance.
- **Data democratization services** refers to areas such as establishing a data-driven organizational culture through the business validation of data initiatives, making non-sensitive data available to all employees (data democratization) as well as proper evaluation of available data.
- **Data enrichment services** includes areas such as deriving and creating value from data through applying various analytical and machine based methods and techniques, exploring and experimenting with data as well as ensuring a smart and efficient data lake/data pipeline setup supporting a value realization of the data science investment.

Explaining Why a Chief Data Officer Is Needed

In addition to exploring revenue opportunities, developing acquisition strategies, and formulating customer data policies, the chief data officer is charged with explaining the strategic value of data and its important role as a business asset and revenue driver to executives, employees, and customers. Chief data officers are successful when they establish authority, secure budget and resources, and monetize their organization's information assets.

REMEMBER

The role of the CDO is relatively new and evolving quickly, but one convenient way of looking at this role is to regard this person as the main defender and chief steward of an organization's data assets. Organizations have a growing stake in aggregating data and using it to make better decisions. As such, the CDO is tasked with using data to automate business processes, better understand customers, develop better relationships with partners, and, ultimately, sell more products and services faster.

A number of recent analyses of market trends claim that by 2020, 50 percent of leading organizations will have CDOs with similar levels of strategy influence and authority as CIOs. CDOs can establish a leadership role by aligning their priorities with those of their organizations. To a great extent, the role is about change management. CDOs first need to define the role and manage expectations by considering the resources made available to them.

WARNING

Despite the recent buzz around the concept of CDOs, in practice it has proven to be difficult for them to secure anything other than moderate budgets and limited resources when reporting into existing business units, like IT. Moreover, with usually only a handful of personnel, the CDO group must operate virtually by tagging onto, and inserting themselves into, existing projects and initiatives throughout the organization. This, of course, isn't an optimal setup when it comes to proving the value of the CDO function.

TIP

For the CDO function to truly pay off, you need to break up the silos and optimize the company structures around the data. It's all about splitting up the scope of responsibility for your IT department so that you can separate out the data assets from the technology assets and let the CDO take ownership of the data and information part, as well as the full data science cycle when there is no CAO role appointed.

Establishing the CDO Role

The main task of a chief data officer is to transform the company culture to one that embraces an insight- and data-driven approach. The value of this should not be underestimated. Establishing a data-first mentality pushes managers at all levels to treat data as an asset. When managers start asking for data in new ways and view data science competence as a core skill set, they will drive a new focus and priority across all levels of the company. Changing a company's cultural mindset is no easy task — it takes more than just a few workshops and a series of earnest directives from above to get the job done. The idea that one must treat data as an asset needs to be firmly anchored in the upper management layer — hence the importance of the CDO role.

Let this list of common CDO mandates across various industries serve as an inspiration for what can fit in your company. A CDO can

- Establish a data-driven culture with effective data governance. As part of that project, it is also vital to gain trust the trust of the various business units so that a company-wide sense of data ownership can be established. The idea here is to *foster,* not hinder, the efficient use of data.
- Drive data stewardship by implementing useful data management principles and standards according to an agreed-on data strategy. It is also important to industrialize efficient data-quality management, since ensuring data quality throughout the data life-cycle requires substantial system support.
- Influence decision-making throughout the company, supported by quality data that allows for analytics and insights that can be trusted.
- Influence return on investment (ROI) through data enrichment and an improved understanding of customer needs. The idea here is to assist the business in delivering superior customer experiences by using data in all applicable ways.
- Encourage continuous data-driven innovation through experimentation and exploration of data, including making sure that the data infrastructure enables this to be done effectively and efficiently.

As in most roles, you always face a set of challenges impacting the level of success that can be achieved. Just by being aware of these challenges, you'll be better able to avoid them or at least have strategies in place to deal with them if, or when, they arise. The following list summarizes some of the most common challenges related to the CDO role:

- **Assigning business meaning to data:** A CDO must make sure that data is prioritized, processed, and analyzed in the right business context in order to generate valuable and actionable insights. One aspect of this could be the timing of the *insight generation:* If the time it takes to generate the insight is too slow from a business usefulness perspective, the insight, rather than steering the business, would merely confirm what just happened.
- **Establishing and improving data governance:** The area of data governance is crucial for keeping data integrity during the life cycle of the data. It's not just about managing access rights to the data, but very much about managing data quality and trustworthiness. As soon as manual tasks are part of data processing activities, you run the risk of introducing errors or bias into the data sets, making analysis and insights derived from the data less reliable. Automation-driven data processing is therefore a vital part of improving data governance.
- **Promoting a culture of data sharing:** In practice, it is the common data science function that will drive the data science activities across the company as it promotes data sharing from day to day. However, it's also significant to have a strong spokesperson in management who enforces an understanding and acceptance of data sharing. The main focus should be to establish that you won't be able to derive value from data unless the data is used and shared. Locking in data by limiting access and usage is the wrong way to go — an open data policy within the company should be the starting point. With that in place, you can then limit access on sensitive data and still ensure that such limitations are well motivated and cannot be solved by using anonymization or other means.
- **Building new revenue streams, enriching and leveraging data-as-a-service:** A person in the CDO role also promotes and supports innovation related to data monetization. This is an inspiring task, but not always an easy one. Driving new business solutions that require data-driven business models and potentially completely new delivery models might inspire a lot of fear and resistance in the company and with management in general. Remember that new data monetization ideas might challenge existing business models and be seen as a threat rather than as new and promising business potential. "Be mindful and move slowly" is a good approach. Using examples from other companies or other lines of business can also prove effective in gaining trust and support from management for new data monetization ideas.
- **Delivering Know Your Customer (KYC) in a real and tangible fashion:** Utilizing data in such a way that it can enable data-driven sales is a proactive and efficient way to strengthen customer relationships and prove how knowledgeable the company is. However, there should be a balance here in how data is obtained and used: The last thing you want is for your customers to feel intruded on. You want the company to be perceived as proactive and

forward-leaning with an innovative drive that is looking out for its customers, not as a company that invades its customers' private sphere, using their data to turn it into an advantage in negotiations. The CDO must master that balance and find a way to strategically toe that line — a line that can be quite different, depending on your business objectives and line of business.

- **Fixing legacy data infrastructure issues while investing in the future of data science:** This challenge is tricky to handle. You can't just switch from old legacy infrastructures (often focused on data transactions and reporting) to the new, often cloud based infrastructures focused on handling data enablement and monetization in completely new and different ways. There has to be a transition period, and during that period you have to deal with maintaining the legacy infrastructure, even when it's costly and feels like an unnecessary burden. At the same time, management is expecting fast and tangible results, based on the major investments needed. But be aware that the longer you have the two infrastructures working in parallel, the harder it is to truly get people to change their mindset and behavior toward the new data-driven approach realized through the new infrastructure investments. Neither will you see any real savings, because you need to cost-manage both the legacy environments and the new environments. Even if it proves difficult, try to drive this swap with an ambitious timeline, keeping in mind that there's no going back.

The Future of the CDO Role

The appointment of chief data officers in large organizations has ramped up in recent years as companies realize the importance of data as a fundamental business asset, with nine out of ten enterprises expected to fill this role by 2020.

Businesses are increasingly placing the chief data officer (CDO) at the hub of their operations, with a responsibility linked to all functions, as the dependence on information and data-driven decision-making increases. At the same time, the CDO role is becoming broader and less technical. In many firms, smarter analytics tools are rendering moot some of the complex data science formerly required. Only a few years ago, a CDO would typically have been expected to have a highly technical background, but now the role is emerging from various parts of the business.

TIP

CDOs now need to know far more about business contexts, strategies, and risks than what was required just a decade ago. The data in focus today isn't just "customer data" — it's everything in a business, and that's why the role is changing. The ability to act on real-time data is central to business strategies, which is why elevating the role of chief data officer as an all-encompassing responsibility is such a big deal. At the senior level, the CDO needs to be someone who understands the business, is able to equip teams with the right infrastructure and tool set, and knows how to make data accessibility both efficient and simple.

WARNING

So the role of the CDO is obviously evolving, in terms of both its placement in the organizational hierarchy and its increasing scope and mandate. As the CDO role matures, it seems to be following a certain path, especially in those companies that established the role early on. You can see a high-level view of these steps on the way to maturity in Figure 12-3.

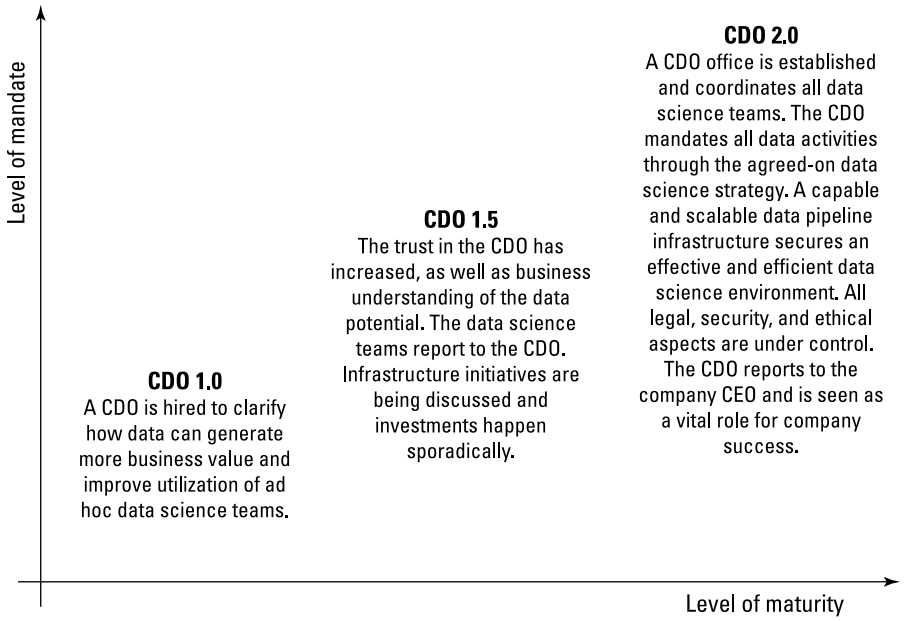

FIGURE 12-3: The evolution of the CDO role.

When you're talking about a CDO 1.0 company context (refer to Figure 12-3), you probably start out by testing some initial initiatives, which usually means hiring data scientists sporadically across the company. After you determine that the piecemeal approach isn't generating much value, you usually establish a few data science teams as supporting roles around the data scientists. Then, finally, a CDO is hired to try to bring some order to the chaos, making sure that the data science teams are utilized efficiently. The person in the CDO role usually reports to a C-level executive in the company management team or one management level below.

In the CDO 1.5 context, some recognition and trust is starting to happen around the role as such, and common investments are being made. There is nevertheless still a lack of company-wide strategic alignment, and a lot of ad hoc activities continue to pop up. However, some alignment between different data science teams is usually ongoing, and management is starting to express its expectations when it comes to results. At this maturity level, the CDO isn't yet part of the company management team.

Finally, in the CDO 2.0 maturity level, the company truly recognizes the importance of the CDO role for company success. Usually, a CDO office is established, the CDO reports directly to the CEO, and all activities and investments are driven from an agreed-on and approved data science strategy. Data science organization setup is agreed on and coordinated throughout the company by way of the CDO office, which also ensures standards, principles, and infrastructure alignment.

So, where can CDOs turn for assistance to drive up the recognition and maturity of the role? Many problems for the CDO relate to proving the stability of the function when everything is evolving *fast.* When members of company management have problems keeping up with how the area of data science continuously transforms, it is seen as an area to be treated differently — more like a start-up or an innovation unit rather than as a business function like any other. This sidelining, which is a major problem, is hindering the CDO role in becoming a vital and integrated function in the company.

WARNING

At the day-to-day level, the real problems facing the CDO come from having to wade through the vast number of services offered by the data industry: transformation agencies, cloud services, data cleaners, and algorithm designers, for example. How can a CDO find the right services among all of this? In this case, success is hard to judge, especially in a role that has yet to be well defined by industry, whereas failure can be pretty obvious. If your company is front page news because of a major data breach or privacy violation, it's a bad day to be the CDO. To find out what makes a good day — well, that requires more companies to dare to trust, and invest in, a CDO.

IN THIS CHAPTER

- Determining the skill sets needed for an effective data science team
- Exploring the characteristics of a data scientist
- Applying structure to a data science team
- Keeping the talent you acquire

Chapter 13

Acquiring Resources and Competencies

Nearly every company now has the ability to collect data, and the amount of data is growing larger and larger. This has led to a higher demand for employees with specialized skills who can effectively organize and analyze this data to glean business insights. Unfortunately, not only does the demand for data scientists surpass the available supply, many of the aspiring data scientists in the market don't have the skillset or experience needed for available positions.

WARNING

The specialized, complex nature of data science work poses a significant problem for hiring. In fact, there's still genuine confusion in the job market about what the term *data scientist* actually means. There are often specific technical requirements that different roles within the data science organization demand, but there needs to be a common understanding of what is required for a data science team to be successful.

Identifying the Roles in a Data Science Team

In the past couple of years, an avalanche of different data science roles has overwhelmed the market, and for someone who has little or no experience in the field, it's hard to get a general understanding of how these roles differ and which core skills are actually required. The fact is that these different roles are often given different titles, but tend to refer to the same or similar jobs — admittedly, sometimes with overlapping responsibilities. This crazy-quilt of job titles and job responsibilities is yet another area in data science that is in need of more standardization.

Before attempting some hard-and-fast role definitions, then, let me start by sketching out the different task sets you'd typically find on a data science team. (See Figure 13-1.) The idea here is to scope the high-level competence areas that need to be covered on a data science team, regardless of who actually carries out which task. The three main areas are mathematics/statistics, computer science, and business domain knowledge.

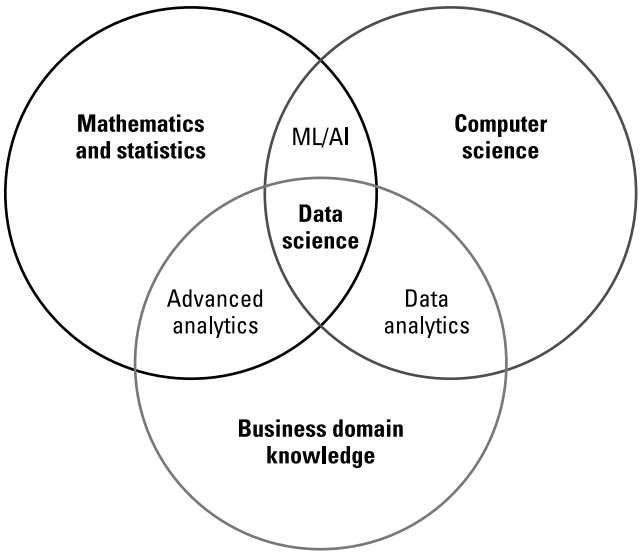

FIGURE 13-1: Competence areas needed on a data science team.

Figure 13-1 shows the easy part because there's general agreement on which competencies are required for a successful and efficient data science team — though you still need to define the roles and areas of responsibility for each team member. The definitions in this section aim to give you a general understanding of the most important roles you'll need on your data science team. Just remember that variants may apply, depending on your own specific setup and strategic focus.

Data scientist

In general terms, a *data scientist* produces mathematical models for the purposes of prediction. And, because the development and interpretation of mathematical models requires deep technical knowledge, most data scientists have graduate level training in computer science, mathematics, or statistics. Data scientists also need strong programming skills in order to effectively leverage the range of available software tools. Aside from being technically savvy, data scientists need critical thinking skills, based on common sense as well as on a thorough understanding of a company's business objectives in order to produce high quality models.

REMEMBER

Sometimes, a role referred to as *data analyst* is set apart from the data scientist role. In such cases, the data analyst role is like the Sherlock Holmes of the data science team in that they focus on collecting and interpreting data as well as analyzing patterns and trends in the data which they draw conclusions from in a business context. The data analyst must master languages like R, Python, SQL, and C, and, just like the data scientist, the skills and talents that are needed for this role are diverse and span the entire spectrum of tasks in the data science process. And, to top it all off, a data analyst must demonstrate a healthy I-can-figure-it-out attitude. It's really up to you to decide whether you want to have all your company's data scientists take up the tasks associated with a data analyst or if you want to set up a data analyst as a separate role.

Within the role of the data scientist, you'll find another, more traditional role hidden away: the statistician. In historical terms, the statistician was the leader when it came to data and the insights it could provide. Although often forgotten or replaced by fancier-sounding job titles, the statistician role represents what the data science field stands for: getting useful insights from data. With their strong background in statistical theories and methodologies, and a logically oriented mindset, statisticians harvest the data and turn it into information and knowledge. They can handle all sorts of data. What's more, thanks to the quantitative background, modern statisticians are often able to quickly master new technologies and use these to boost their intellectual capacities. A statistician brings to the table the magic of mathematics with insights that have the ability to radically transform businesses.

Data engineer

The role of the *data engineer* is fundamental for data science. Without data, there cannot be any data science, and the job of data scientists is a) quite impossible if the requisite data isn't available and b) definitely daunting if the data is available but only on an inconsistent basis. The problem of inconsistency is frequently faced by data scientists, who often complain that too much of their time is spent on data acquisition and cleaning. That's where the data engineer comes in: This

person's role is to create consistent and easily accessible data pipelines for consumption by data scientists. In other words, data engineers are responsible for the mechanics of data ingestion, processing, and storage, all of which should be invisible to the data scientists.

If you're dealing with small data sets, data engineering essentially consists of entering some numbers into a spreadsheet. When you operate at a more impressive scale, data engineering becomes a sophisticated discipline in its own right. Someone on your team will need to take responsibility for dealing with the tricky engineering aspects of delivering data that the rest of your staff can work with.


REMEMBER

Data engineers don't need to know anything about machine learning (ML) or statistics to be successful. They don't even need to be inside the core data science team, but could be part of a larger, separate data engineering team that supplies data to all data science teams. Based on my experience, however, you should never place your data engineers and data scientists too far apart from one other organizationally. If these roles are separated into different organizations, with potentially different priorities, this could heavily impact the data science team productivity. Data science methods are quite experimental and iterative in nature, which means that it must be possible to continuously modify data sets as the analysis and algorithm development progress. For that to happen, data scientists need to be able to rely on a prompt response from the data engineers if trouble arises. Without that rapid response, you run the risk of slowing down a data science team's productivity.

Machine learning engineer

Data scientists build mathematical models, and data engineers make data available to data scientists as the "raw material" from which mathematical models are derived. To complete the picture, these models must first be deployed (put into operation, in other words), and, second, they must be able to act on the insights gained from data analysis in order to produce business value. This task is the purview of the *machine learning engineer.*

The machine learning engineer role is a software engineering role, with the difference that the ML engineer has considerable expertise in data science. This expertise is required because ML engineers bridge the gap between the data scientists and the broader software engineering organization. With ML engineers dedicated to model deployment, the data scientists are free to continually develop and refine their models.

REMEMBER

Variants are always a possibility when setting up a data science team. For example, the ML engineer deployment responsibilities are often also handled by the data scientist role. Depending on the importance of the operational environment for your specific business, it can make more or less sense to separate this role from the data scientist responsibilities. It is, again, up to you to implement this responsibility within the team.

Data architect

A *data architecture* is a set of rules, policies, standards and models that govern and define the type of data collected and how it is used, stored, managed and integrated within an organization and its data systems. The person charged with designing, creating, deploying, and managing an organization's data architecture is called a *data architect,* and they definitely need to be accounted for on the data science team. (For more details on data architecture, see Chapter 14.)

Data architects define how the data will be stored, consumed, protected, integrated, and managed by different data entities and IT systems, as well as any applications using or processing that data in some way. A data architect usually isn't a permanent member of a single data science team, but rather serves several data science teams, working closely with each team to ensure efficiency and high productivity.

Business analyst

The *business analyst* often comes from a different background when compared to the rest of the team. Though often less technically oriented, business analysts make up for it with their deep knowledge of the different business processes running through the company — operational processes (the sales process), management processes (the budget process) and supporting processes (the hiring process). The business analyst masters the skill of linking data insights to actionable business insights and can use storytelling techniques to spread the message across the entire organization. This person often acts as the intermediary between the "business guys" and the "techies."

Software engineer

The main role of a *software engineer* on a data science team is to secure more structure in the data science work so that it becomes more applied and less experimental in nature. The software engineer has an important role in terms of collaborating with the data scientists, data architects, and business analysts to ensure alignment between the business objectives and the actual solution. You

could say that a software engineer is responsible for bringing a software engineering culture into the data science process. That is a massive undertaking, and it involves tasks such as automating the data science team infrastructure, ensuring continuous integration and version control, automating testing, and developing APIs to help integrate data products into various applications.

Domain expert

It takes a lot of conversations to make data science work. Data scientists can't do it on their own. Success in data science requires a multiskilled project team with data scientists and domain experts working closely together. The *domain expert* brings the technical understanding of her area of expertise, sometimes combined with a thorough business understanding of that area as well. It usually includes familiarity with the basics of data analysis, which means that domain experts can support many roles on the data science team. However, the domain expert usually isn't a permanent member of a data science team; more often than not, that person is brought in for specific tasks, like validating data or providing analysis or insight from an expert perspective. Sometimes the domain expert is allocated for longer periods to a certain team, depending on the task and focus. Sometimes one or several domain experts are assigned to support multiple teams at the same time.

Seeing What Makes a Great Data Scientist

There's a lot of promise connected with the data scientist role. The problem is not only that the perfect data scientist doesn't exist, but also that the few truly skilled ones are too few and too difficult to get hold of in the current marketplace. So, what should you be doing instead of searching for the perfect data scientist?


TIP

The focus should be on finding someone with the ability to solve the specific problems your company is focusing on — or, to be even more specific, what your own data science team is focusing on. It's not about hiring the perfect data scientist and hoping that they're going to do all the things that you need done, now and in the future. Instead, it's better to hire someone with the specific skills needed to meet the clearly defined organizational objectives you know of today.

For instance, think about whether your need is more related to ad hoc data analysis or product development. Companies that have a greater need for ad hoc data insights should look for data scientists with a flexible and experimental approach and an ability to communicate well with the business side of the organization. On the other hand, if product development is more important in relation to the problems you're trying to solve, you should look for strong software engineering skills, with a firm base in the engineering process in combination with their analytical skills.

REMEMBER

If you're hoping to find a handy checklist of all the critical skills that you should be looking for when hiring a data scientist, you'll be sorely disappointed. The fact is, not even a basic description of important traits the role should possess is agreed on across the industry. There are many opinions and ideas about it, but again the lack of standardization is troublesome.

So, what makes coming up with a simple checklist of the needed tool sets, competencies, and technical skills required so difficult? For one, the area is still evolving fast, and tools and techniques that were important to master last year might be less important this year. Therefore, staying in tune with the evolution of the field and continually learning new methods, tools, and techniques is the key in this space. Another reason it's difficult to specify a concrete checklist of skills is because the critical skill sets needed are actually outside the data science area — they qualify more as soft skills, like interpersonal communication and projecting the right attitude. Just look at the data scientist Venn diagram of skills, traits, and attitude needed, shown in Figure 13-2. The variety of skill sets and mindset traits that a perfect data scientist must master is almost ridiculous.

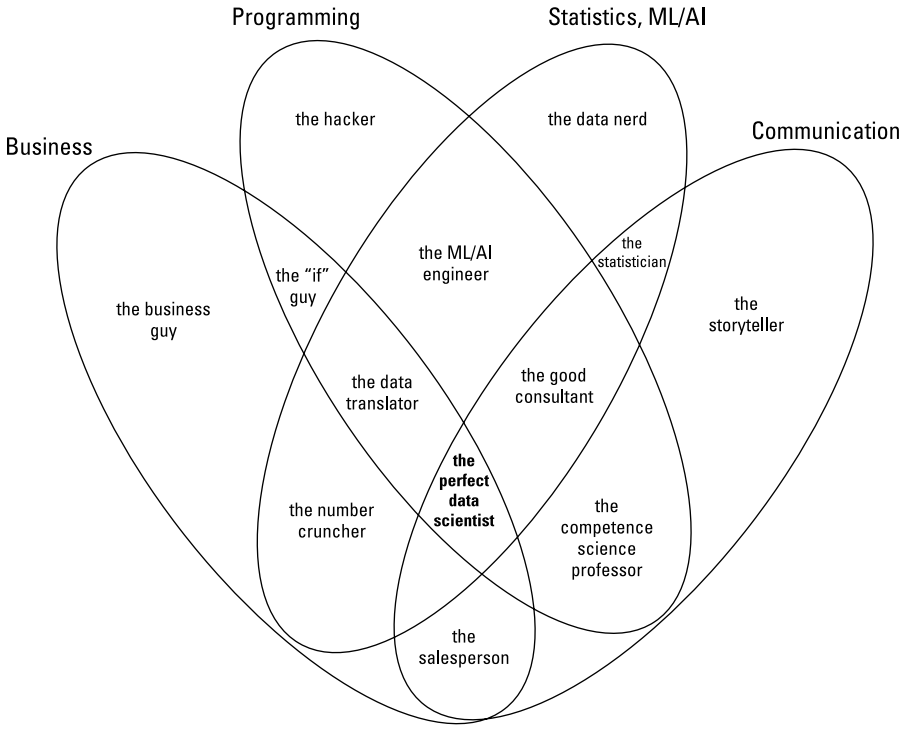

FIGURE 13-2: A data scientist Venn diagram of skills, traits, and attitude needed.

So, bearing in mind that specifying competencies needed for a data scientist is more a question of attitude and mindset in combination with a certain skill set, I have still compiled this list:

- **Business understanding:** Having the ability to translate a problem from business language into a hypothesis is important and refers to how a data scientist should be able to understand what the business person describes, and then be able to translate that into technical terms and present a potential solution in that context.
- **Impactful versus interesting:** Data scientists must be able to resist the temptation to always prioritize the interesting problems when there might be problems that are more important to solve because of the major business impact such solutions would have.
- **Curiosity:** Having an intellectual curiosity and the ability to detail a problem into a clear set of hypotheses that can be tested is a major plus.
- **Attention to detail:** As a data scientist, pay attention to details from a technical perspective. A model cannot be nearly right. Building an advanced technical algorithm takes time and dedication to detail.
- **Easy learner:** The data scientist must have an ability to learn quickly, because the rapidly changing nature of the data science space includes technologies and methodologies but also new tools and open-source models that are made available and become ready to build on.
- **Agile mindset:** Stay flexible and agile in terms of what is possible, how problems are approached, how solutions are investigated, and how problems are solved.
- **Experimentation mindset:** The data scientist must not fear to fail or try assumptions that might be wrong in order to find the most successful way forward.
- **Communication:** A data scientist must be able to tell a story and describe the problem in focus or the opportunity that he's aiming for, as well as describe how great the models are once they are finished and what they actually enable.

Of course, there are additional skills of interest, such as in statistics, machine learning, and programming, but remember that you do not need one person to fit all categories here. First of all, you should be looking for data scientists who possess the most important skills that meet your needs. However, in the search for that top notch data scientist, remember that the list above could also be used for hiring a complementary team of data scientists which together possess the skills and mindset needed.

TIP

After your team of data scientists is in place, encourage their professional development and lifelong learning. Many data scientists have an academic mindset and a willingness to experiment, but in the pursuit of a perfect solution, they sometimes get lost among all the data and the problems they're trying to solve. Therefore, it's important that they stay connected with the team, though you should allow enough independence so that they can continue to publish white papers, contribute to open source, or pursue other meaningful activities in their field.

Structuring a Data Science Team

When building a data science team with the right type of skill sets, it's all about finding that optimized set of team members. What are the key drivers for different types of roles, and how should you combine them into one team?

Let's be honest: There's no formula you can apply that will solve this equation for you. It's a little more complicated than that. How the team structure needs to look and be balanced is very much related to what your objectives are, what your processes look like, how the intended target environment is defined, and so on. However, you can always start with a simple standard setup based on the role descriptions in the previous section and work from there. You can see this simplified setup in Figure 13-3.

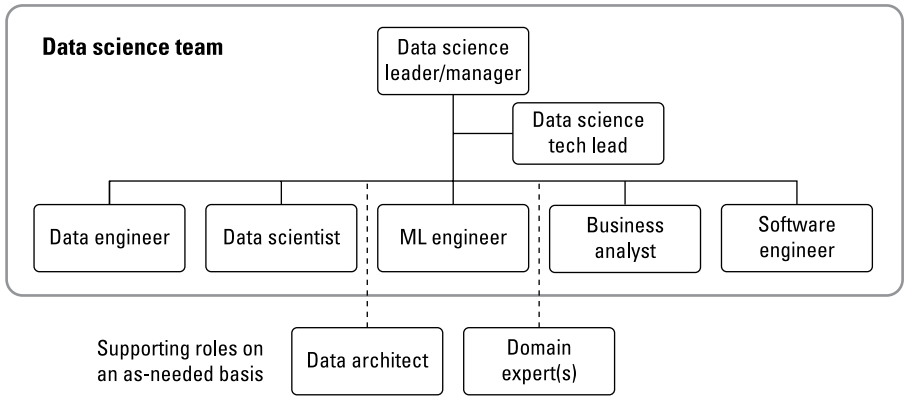

FIGURE 13-3: A typical data science team structure.

When balancing needed resources per role in the data science team, you need to consider aspects such as the ones described in this list:

- **Data scope and complexity:** What is the scope of your data science challenge? Are you targeting internal business efficiency gains, or are you going for a commercial data business? The scope of the data needed and the

complexity of the data acquisition involved will, for example, impact the need for data architectural support and experienced data engineers.

- **Data product or service type:** If you're going for a commercial data product, are you aiming for something sold off-the-shelf or are you aiming to build an operational model from a data service perspective? How do you aim to deliver your data product or service? Depending on the offering type as well as the delivery model, it's certain to impact the number of software engineers you will need for the end-to-end solution development and delivery platform creation.
- **Level of machine learning/artificial intelligence techniques used:** How complex are your use cases and the targeted solution? Is there a need for a highly technical solution with a lot of self-learning algorithms, or is a simpler model sufficient? The level of complexity involved will drive the need for different data scientist skill sets, ranging from analytics to advanced analytics and from machine learning to artificial intelligence competence.
- **Data science environment setup (development and production):** What will the data science infrastructure look like? Is it cloud based or on-premise based? Is it globally distributed, or is there a single local instance? Or do you have a global setup with a centralized instance and cloud-edge instances in different countries, or even edge instances on device level? The infrastructure setup can vary a lot between companies, depending on the size of company, if it's local or global, the line of business it is focused on, whether the data is owned or you need rights to use it, and so on. However, from a resource balancing aspect, you will soon realize that the data science environment setup impacts the number of data science teams you will be needing and also which competences you need more of — as well as where in the world the teams need to be placed.
- **Data science organizational model:** This refers to the organizational setup you have decided on in terms of having a centralized team, a decentralized team, or a hybrid. (See Chapter 11 for more on these models.) The resource balancing depends on which role a centralized function will have in your company. For example, if you have a common centralized data science function, does that mean all data scientists should work there, serving the whole company? Or does it mean that the centralized function works only on common parts, which is relevant across different business units, meaning that the rest of the organization is allowed to acquire and build their own data science competence? Those are important questions to be clear about in order to better understand your data science resource balancing parameters.

Hiring and evaluating the data science talent you need

When making hiring decisions in data science, your goal is to have a well-functioning team, not just a set of skilled individuals. Equally important is the need to create a diverse team, where individuals with different backgrounds and different life experiences can work comfortably together. The trick is to start your search by looking for individuals who represent the different disciplines — data scientists, data engineers, and software engineers, for example — but then to always make the final decision based on a candidate's ability to function well within a data science context.

So, how can you tell whether someone will function well in such a context? I always look at three main areas when evaluating whether a candidate has the right skill set and personality traits to succeed (by the way, I always apply these criteria to *all* individuals On the team, not just a select few):

- Cultural fit
- Engineering skills
- Data science competence

In the area of engineering skills, look for competencies in system design, formal coding, and algorithm development. As for data science skills, you should insist on competencies in instance modeling and algorithms, ML framework and tools (like TensorFlow), and data processing.

When it comes to a cultural fit, start out by looking at people who clearly share the values of the company as well as the other team members. From there, move on to evaluating how the candidate works as part of a team and then gauge their personal motivation and drive. However, remember the rule of not hiring people who represent the same background, gender, education, and age. A diverse team secures diverse results and actively works against bias in the data and insights.


TIP

To evaluate the candidates from all these aspects, it's important to have a clear and mutually agreed-on perception of what good looks like. This should be in the form of generic evaluation criteria that everyone who is participating in the hiring process is in agreement about — and you should definitely have it in writing. This seems like an easy thing to do, but it's often hard to do in practice because what "good" looks like is very much a matter of personal opinion. But as hard as it is, defining evaluation criteria is really important to do. All candidates need to be evaluated in as unbiased a fashion as possible against the same criteria.

To get the data on your candidates you need in order to evaluate the different areas, you should use a combination of assignments and tests with interviews. Test results, outcomes, and any insights gained from personal interviews then need to be mapped to what good looks like. For instance, you might do a screening interview, an engineering and system design interview, a data science assignment, followed by a data science fundamentals interview. If the candidate progresses through all these stages, the data science leader then needs to evaluate the cultural fit to make sure that the right *person* (and not merely the right *skill set*) is hired.

You might be wondering why you should be evaluating a data engineer by looking at the same areas and using the same criteria you'd use for a data scientist. I'd argue that it pays to be able to gauge the breadth of a candidate's knowledge, especially in neighboring fields of expertise. Regardless of the findings, it could be useful to know whether a candidate is a skilled data scientist but a pretty bad engineer, even if that individual might never be asked to do any coding. And, when you come right down to it, there really are some important cross-functional skills that are essential for a data science team. For example, it would be best to have a data engineer who is skilled with coding and knows about system design and DevOps, yet knows enough about the fundamentals of data science to know to what extent and in what way their talents can be put to use in a data science context. No, that person doesn't need to have the same level of data science understanding that a machine learning engineer with a background as a mathematician would have, but they need to be at a level where they know how they can contribute.

To cater to these needed cross-functional skill sets, you should map what skills and responsibilities are either less important or more important for the different roles on your team. Yes, a data scientist should take the system design test, but if he scores poorly, it should not be an immediate black mark against him. But then you should require this candidate to score high on his understanding of topics like naive Bayes methods and logistic regression techniques as a counterpoint. Figure 13-4 shows a high level mapping of skill sets to certain roles in order to define what good looks like.

After mapping the relative importance of various skill sets, be sure to map out which areas or skill sets you expect the interviews, assignments, and tests to provide data on in order to make sure that you cover all needed data points for evaluating the candidate properly. Keep in mind that it's really quite hard to build successful teams, and it also costs a pretty penny. Take this work seriously and come well prepared for the interviews. This is also important because the access to experienced competence in data science is scarce, and it's not only you who is evaluating the candidate — the candidate is also evaluating you and your level of competence in the area, as well as the maturity of the company. Being well prepared and having a well-thought-out interview structure is a good starting point.

Area	Skill Set	Data Scientist	Data Engineer
Engineering	System design	Relevant	Important
Engineering	Formal coding	Relevant	Important
Engineering	Code-based problem solving	Important	Important
Engineering	Data architecture and data models	Relevant	Important
Data science	Data processing	Important	Important
Data science	Modeling and algorithms	Important	Relevant
Data science	Analysis and evaluation	Important	Relevant
Data science	Frameworks and tools	Important	Relevant
Culture fit	Company fit	Relevant	Relevant
Culture fit	Personal motivation	Important	Important
Culture fit	Career goals	Relevant	Relevant
Culture fit	Team fit and diversity aspects	Important	Important

FIGURE 13-4: An example of mapping the importance of skill set to certain roles.

Retaining Competence in Data Science

What can companies do to get the most out of their data science teams and to motivate them to make a more robust contribution to the business? One important part is giving data scientists the time they need to invent. Remember that you're dealing with people who, on one hand, want to push the boundaries and, on the other hand, get bored easily if they're asked to do the same thing over and over again.

REMEMBER

These scientists are scarce talents who want to work on the company's most important functions. If they're asked to spend their time performing repetitive tasks such as data acquisition, data management, and extensive massaging of results forecasting, they often feel underutilized. Tasking data scientists with forward looking projects gives them the opportunity to invent the way the company can benefit from big data.

Also make sure that company management is involved at the right stage of the data projects. Without access to senior management, data science teams may focus on the wrong problems. This should preferably be managed by way of the CDO role, if that role is in place within the company. (For more on the CDO role, see Chapter 12.) In general, it's crucial for data science teams to engage senior management at three stages in any project: early on, to help define the problem the company wants to solve; after the first results start rolling in; and when it comes time for the resulting insights to be implemented or acted on.

If handled correctly, your data scientists can develop a tremendous reputation for knowledge inside the organization. Ensuring that the dialogue between the data science teams and senior management occurs early and often also increases the likelihood that data scientists' suggestions are actually implemented. Again, this is a vital role to play for the CDO.

TIP

When it comes to motivating strategic data science talent to stay in the company, one other strategy is to let data scientists out of the data box. Data scientists are natural learners who are positioned to see all aspects of the business as informed by data, rather than through a traditional software development or marketing lens. Because of this perspective, they can make connections others can't to broader conversations and innovative ideas through their observation of the overall business.

I'd also recommend that you consider cross-training your company's data scientists. Whether or not data scientists have the sexiest job of the 21st century, as *Harvard Business Review* declared, is debatable, but what is not in dispute is that they're hard to identify, hard to recruit, and in short supply. *Cross-training* data scientists means moving people from a data science organization into operations management, digital marketing, or customer relationship management, which are all analytically grounded disciplines and can open up new opportunities for personal and career development, not only for the data scientists but also for the company as a whole when their competence and knowledge are spread out in a more practical manner, driving data-driven thinking and applying statistical models in practice outside traditional domains. Such cross-training can also act as a motivator for more people to want to learn more about data science and pursue a career in that area through formal and on-the-job training.

REMEMBER

When business leaders confuse data reporting for analysis, a company can have trouble addressing problems effectively. By the same token, data scientists need to learn how to address senior management on senior management's terms. Data scientists tend to want to explain everything they've done, describe how hard they've worked, and emphasize what an accomplishment it was. Senior management, on the other hand, has three rules: Be clear, be quick, and be gone.

TIP

Developing the business acumen of data scientists helps them contribute more holistically to conversations within the company, allowing them to initiate analyses and experiments rather than simply react to requests. That is a long-term benefit that costs companies little to implement, but it's a crucial competence over the long term.

Understanding what makes a data scientist leave

Unfortunately, whatever your ambitions are in terms of the new data scientists you're bringing into the company, many data scientists tend to move on, often within their first year. Why is that? To complement the earlier section on what you should do to retain your valuable data science resources, this section aims to pinpoint four main concerns that drive data scientist dissatisfaction. Here's my list:

- **Expectation doesn't match reality.** Many companies hire data scientists without really understanding what data science is all about. For example, without a suitable infrastructure in place to start getting value out of their data science investment, coupled with the fact that these companies fail to hire senior or experienced data practitioners before hiring juniors, you now have a recipe for a disillusioned and unhappy relationship for both parties. The data scientist likely enters the company with the ambition to write smart machine learning algorithms to generate insights, but soon discovers that they can't do this because their first job is to sort out the data infrastructure and/or create reports on demand. In contrast, many times, the level of data science maturity is so low in the company that all they want is a chart they can present in their board meeting each day. Leaders at such companies then get frustrated because they don't see value being generated quickly enough and all of this of course leads to the data scientist being unhappy in the role and eventually leaving.
- **Company politics is more important than data science skills.** A data scientist often assumes that knowing lots of machine learning algorithms will make him the most valuable person in the company. However, the data scientist soon discovers that those expectations do not match reality. The truth is, the people in the business with the most influence need to see the worth of any employee they're thinking of entrusting with greater responsibilities, regardless of whether they are data scientists or not. From a data scientist perspective, that means first making yourself available and then working to make yourself irreplaceable. For that to happen, you need to be ready to handle a constant flow of ad hoc work, such as getting numbers from a database to give to the right people at the right time, doing simple projects just so that the right people get the right perception of you, the data scientist,

as someone who is trustworthy, reliable, and innovative. As frustrating as it may sound, putting yourself out there is a necessary part of the job that any data scientist must accept if they hope to get to the point where they can achieve something more interesting and impactful.

- **The data scientist role isn't understood.** Following on from doing anything to please the right people, those same people with all that power often don't understand what is meant by the term *data scientist.* This means that data scientists are expected to be the analytics experts as well as the go-to reporting folks, and let's not forget the database experts, too. It isn't just nontechnical executives who make too many assumptions about data scientist skills: Other colleagues in technology assume that the data scientist knows everything that is data related. The conventional wisdom states that the data scientist should know her way around Spark, Hadoop, Hive, Pig, SQL, Neo4J, MySQL, Python, R, Scala, TensorFlow, A/B testing, NLP, anything related to machine learning, and anything else you can think of that is related to data. But it doesn't stop there. Because the data scientist supposedly knows all of this and obviously has access to all the data, the expectation is that the data scientist has the answers to all the questions within minutes. Trying to tell everyone what you actually know and have control of can be both difficult and frustrating.

- **Working on an isolated team limits productivity.** When you see successful commercial data products, you often see expertly designed user interfaces with intelligent capabilities and, most importantly, a useful output, which, at the very least, is perceived by the users to solve a relevant problem. Now, if a data scientist spends his time only learning how to write and execute machine learning algorithms, he can only be a small (although necessary) part of a team that leads to the success of a long effort that ends up producing a valuable data product.

  That's one scenario — yes, part of a larger team, but that often means being a small cog in a much larger machine. Still, this is probably preferable to being shunted to the side and asked to work in isolation on something "data science-y." When cut off from those processes that actually create products to sell, data science teams end up struggling to provide value. Despite this, many companies still ask data science teams to come up with their own projects and write code to solve a problem they've defined. In some cases, this can be sufficient. For example, if all that's needed is a static spreadsheet that is produced once a quarter, the team can provide some value. On the other hand, if the goal instead is to optimize how to provide intelligent suggestions in an adjustable website, this will involve many different skills that shouldn't be expected of the vast majority of data scientists. (Only the true data science

unicorn can solve this one.) So, if you task an isolated data science team with this project, cut off from all other resources, it's most likely to fail (or take a very long time to solve, because organizing isolated teams to work on collaborative projects in large enterprises isn't easy).

The time-proven wisdom about managing teams bears repeating: Data scientist teams, like others, flourish best when there is effective leadership, a strong mandate from the company executive team, and clear objectives based on a solid and agreed strategy in place. Remember that keeping your valued data scientists in your company requires not only a path for data science teams to take key initiatives in a collaborative and agile manner from design through implementation enabled by a fit-for-purpose data infrastructure, but is also very much about managing expectations. In both directions.

4 Investing in the Right Infrastructure

IN THIS PART . . .

Designing and building the data architecture

Governing data to secure integrity, value, and reliability

Managing models in different phases of the data science life cycle

Understanding open source in the context of data science

Addressing infrastructure realization

IN THIS CHAPTER

- Understanding what a data architecture is
- Identifying key characteristics of a data architecture
- Understanding the architectural layers
- Defining key technologies for a data architecture
- Building a modern data architecture

Chapter 14
Developing a Data Architecture

Building a data architecture is similar to what happens when a traditional architect designs a home or a building: First create a blueprint that aligns with the short- and long-term objectives of an organization, and then make sure that the blueprint becomes a reality.

A general view is that a data architecture defines a standard set of products and tools that an organization uses to manage data. But it's much more than that. Any truly effective data architecture must take into account the unique cultural and contextual requirements of an organization, like the company size, setup, and line of business as well as potential technical, legal, security, or other constraints. In addition, a data architecture needs to define the processes to capture, transform, and deliver usable data to business users. Most importantly, it identifies the people who will consume that data and their unique business requirements. I cover all of this (and much more) over the course of this chapter.

Defining What Makes Up a Data Architecture

Within the area of information technology, a data architecture consists of models, policies, rules, and standards that govern which data is collected as well as how it's stored, arranged, integrated, and put to use in data systems and in organizations. Data is usually one of several architectural domains that form the pillars of an enterprise architecture or solution architecture for business operations internally or for a commercial data product or service portfolio offering externally.

REMEMBER

A data architecture should set data standards for all the data systems as a vision or a model of the interactions between an organization's various data systems. Data integration, for example, is dependent on data architecture standards and structures used by the various business units and the selected system applications and defines how the data interaction must work. These standards and structures address data in storage and data in motion and include descriptions of data storage solutions, data categories, and data types, including mappings of those data entities to data quality levels, relevant applications, usage or storage locations, and so on.

One key cornerstone in how a data architecture realizes a company's business objectives is how the data architecture describes how data is processed, stored, and utilized in an industrialized setting or system at work. It has to provide criteria for data processing operations to make it possible to design data flows and also control the flow of data in the data science life cycle.

When it comes to defining the overall data architecture, the responsible party here is, of course, the data architect. However, the data architect is also typically the key person charged with making sure that the data architecture blueprint is followed and understood as part of the realization and build-up of the actual data science infrastructure. This could, of course, also include modifications of the data architecture itself, because of the real-life adjustments that need to happen based on potential legal, security, ethical, geographical, cultural, or technical limitations occurring when the data architecture blueprint is put into practice.

Describing traditional architectural approaches

A data architecture includes a complete analysis of the relationships among an organization's functions, available technologies, and data types. When defining a data architecture for your company, you should approach your task with these three perspectives in mind:

- **Conceptual:** A conceptual data architecture, also sometimes referred to as the *semantic data model*, represents all relevant business entities from a data perspective.
- **Logical:** A logical data architecture, also called a *system data model,* represents the logic of how the included data entities are related and linked to each other from a data flow perspective.
- **Physical:** A physical data architecture represents the actual realization of the architecture in its physical environment — in other words, how the actual data architecture is implemented as part of the technology infrastructure.

The data architecture should be defined during the planning phase of the new data infrastructure setup. As part of that process, your data strategy needs to capture — in a manner that is complete, consistent, and understandable — all the major data categories and data types, as well as the sources of data necessary to support the enterprise's strategic ambitions.


REMEMBER

The primary requirement at this early planning stage is to define all relevant data categories and data types in relation to your organization's business needs and objectives, not to specify which tools or applications should be used to deal with them.

Elements of a data architecture

When it comes to data architecture, it's crucial that certain elements already be defined during the design phase. For example, you need to define the administrative structure and related methodologies and processes required for managing the data during the different stages of its life cycle. Not paying enough attention to the importance of administrating both the data and the data architecture could result in chaos, corrupted data, or a serious blow to your data integrity — any of which could seriously impact the value and usefulness of the data for your company.


REMEMBER

A vital part of your data architecture includes a description of the technology choices. Will your architecture be realized through a virtualized and cloud based environment, or through an on-premises solution, for example? Or will the realization include a local, single site and instance, or will it be deployed in a larger, multisite setup? Will it perhaps even be globally distributed? All these questions need to be understood and answered early on, in order for your data architecture to be designed in a way that supports your business objectives.

TIP

Consider the kinds of interfaces your other systems will need to access your data, as well as the kind of infrastructure design necessary for supporting common data operations (emergency procedures, data imports, data backups, and external transfers of data, for example.)

WARNING

Without the guidance of a properly implemented data architecture design, you might have common data operations implemented in wildly different ways, depending on where you are in the organization. Such a crazy quilt approach makes it extremely difficult to understand and control the flow of data within your organization. This sort of fragmentation is highly undesirable, not only due to its potentially increased cost but also due to the data disconnects that it could involve. These sorts of difficulties are not uncommon in rapidly growing enterprises or in enterprises that have a broad product and service portfolio serving different lines of business.

TIP

Properly executed, the data architecture design phase forces an organization to precisely specify and describe both internal and external information flows. These are patterns that the organization may not have previously taken the time to conceptualize and think through properly. It is therefore possible at this stage to identify costly information shortfalls, disconnects between departments, and disconnects between organizational systems and data that may not have been evident before the data architecture analysis.

Exploring the Characteristics of a Modern Data Architecture

Still waiting on a concrete definition of what a data architecture actually is? Start by looking at these characteristics a data architecture simply must include:

- **A business orientation:** Rather than focus on the data or the technology during the definition phase, a modern data architecture starts with the business users and the overall business objective and flows backward. Customers can be internal or external to an organization, and their needs may vary by role, by department, and over time. A good data architecture therefore continuously evolves to meet new and changing business and customer data needs.
- **Adaptability:** In a modern data architecture, data flows easily from source systems to business users. The purpose of the architecture is to manage that flow by creating a series of interconnected and bidirectional data pipelines that serve various constantly changing business needs.

- **Automation:** To create an easily adaptable architecture in which data flows continuously and data integrity is protected, the architecture must be as automated as possible. The architecture must ensure the profiling and tagging of data at the point of data ingestion and map it to existing data sets and attributes — a key function of creating data catalogs as well, by the way. In the same manner, the data architecture must also enable the detection of changes in the data sources as well as quantify the impact of changes on any architectural component at any time. In a real-time production environment, it must be able to detect anomalies on the fly and either notify the appropriate instances (human and/or machine) or trigger alerts if needed.
- **Intelligence:** The ideal data architecture has more going for it than just automation; it uses machine learning and artificial intelligence to actually build the data objects, tables, views, and models that keep data flowing. In other words, it uses intelligence not only for analyzing the data but also as part of managing and processing the data. Machine learning and artificial intelligence can be applied to identify data types, find common keys and join paths, identify and fix data quality errors, map tables, identify relationships, recommend related data sets and analytics, and so on. A modern data architecture uses intelligence to learn, adjust, alert, and recommend, making people who manage and use the environment more efficient and effective in their jobs.
- **Flexibility:** A modern data architecture needs to be flexible enough to support a variety of business needs. That means it needs to support multiple types of business users, load operations and refresh rates (batch, mini-batch, and stream), query operations (create, read, update, delete), deployments (on-premise, public cloud, private cloud, hybrid), data processing engines (relational, OLAP, MapReduce, SQL, graphing, mapping, programmatic), and pipelines (data warehouse, data mart, OLAP cubes, visual discovery, real-time operational applications.) A modern data architecture has to be all things to all people in the company at any given time.
- **A collaborative spirit:** Unlike in the past, where the IT department built everything, a modern data architecture usually splits the responsibility for acquiring and transforming data between IT and the business units. The IT department may still do the heavy lifting of ingesting data from internal operational systems and create generic reusable building blocks. However, data from external data sources like social media data, customer data, product performance data from live environments and so on, is usually collected by the business. The reason is that the business units already owns that interface, like IT owns the interface to the internal systems. Letting IT focus on the infrastructure backbone of data storage setup and management, as well as on data transfer, is usually a good split — once the data is acquired and ingested, data engineers in the business units are ready to apply data preparation and data catalog tools to create custom data sets to power the business units' analytical and machine learning activities run by data scientists

and business analysts. This collaboration between the data engineers and the data scientists means that IT doesn't have to be involved in business related details around the data.

- **Ease of governance:** A modern data architecture defines access points for each type of user to meet their data need. From the bird's-eye view, you generally have four types of business users — data consumers, data explorers, data analysts, and data scientists — and each type needs different access points to the data. Ensuring that access is what governance is all about, which means that the governance, surprisingly enough, is really the key to a good self-service environment.
- **Simplicity:** Your first assumption should always be that the simplest architecture is usually the best architecture. Ensuring such simplicity can be quite challenging, however, given the diversity of the data needs and the complexity of components in a present-day data architecture. To apply the simplicity rule, an organization with small data sets should seriously consider an out-of-the box analytics tool with a built-in data management environment. To reduce complexity in a big data context and avoid creating a rigid environment, organizations should strive to limit data movement and data duplication by promoting a unified data structure, a data integration framework, and a harmonized analytical and machine learning environment supporting innovation and experimentation, without adding infrastructure complexity. Exactly how to approach this is described later in this chapter.
- **Scalability:** In the age of big data and variable workloads, organizations need a scalable, elastic architecture that adapts to changing data processing requirements on demand. Many companies are now gathering around cloud platforms (both public and private) to obtain on-demand scalability at affordable prices. Elastic architectures free administrators from having to calibrate capacity exactly, control usage if necessary, and constantly overbuy hardware. Scalability also spawns many types of applications and use cases, such as on-demand development and test environments, analytical sandboxes, and prototyping playgrounds.
- **Security:** A modern data architecture must manage to be a collaborative and innovative workspace while at the same time being secure, reliable, and trustworthy. It must manage to provide authorized users ready access to data while keeping hackers and intruders at bay. It must do all that while still complying with privacy regulations, including governmental statutes like the Health Insurance Portability and Accountability Act (HIPPA) in the US as well as regulations like the EU's General Data Protection Regulation (GDPR). The data architecture shall encrypt data upon its ingestion into the data storage, masking personally identifiable information, and track all data elements in a data catalog, including their lineage, usage, and audit trail. Life cycle management ensures that each data object has an owner, a location, and a defined retention period.

» **Resiliency:** A data architecture must also be resilient, with high availability, a robust disaster recovery capability, and a stable infrastructure for backup and restoring data. This is especially true in a modern data architecture that often runs on huge server farms in the cloud, where outages are common. The good news is that many cloud providers offer built-in redundancy and failover with good service level agreements (SLAs) and allow companies to set up mirror images for disaster recovery in geographically distributed data centers at low cost.

Explaining Data Architecture Layers

You need to be aware of various constraints and influences when deciding on a data architecture, because it might impact the architecture's design. These include aspects you'd expect, such as business requirements, key technology choices, financial considerations, different types of business policies, and data processing needs. But it's also important to understand the main architectural layers that make up the basis of any data architecture.

In the past, organizations built fairly static IT-driven data architectures, where systems were complex, difficult, time-consuming to design, and where damage to the database affected virtually all applications running in the environments. These were called *data warehouses.* Because of the underlying technology and design patterns, most data warehouses take an army of people to build and manage — so they provide a minimal return on your investment. Most are overvalued corporate data dumps, where an organization stores all data collected without a defined purpose and structure in the belief that just collecting and dumping the data in a data warehouse adds value in itself. However, there are some existing examples of well-designed and successful implementations that provide a well-functioning environment for data analysis.

REMEMBER

A modern data architecture may still act in part like a data warehouse, but ideally it should be a data warehouse that is flexible, scalable, and agile. Just remember that the storage aspects of a data warehouse comprise just one potential component of a modern data architecture. The new data environment should be approached like a living, breathing organism that detects and responds to changes, continuously learns and adapts, and provides governed, tailored access for every individual.

Figure 14-1 uses the data science flow I define in Chapter 1 to illustrate how you need to build up your data architecture in different layers. Each layer has a defined purpose in the data architecture, based on the specific business context of your

company. This means that a data architectural realization setup and components you decide to use might look quite different, depending on whether you're focusing on internal business analytics or developing an architecture to support a commercial data product or service on a global scale.

Using data science ways of working to drive data architecture design is an excellent way of making sure you get a data architecture that supports the way your business needs to work. Once you defined this, you can then apply the needed systems and tools to realize your architecture in an end-to-end data science infrastructure, as exemplified in Chapter 18 of this book. But let's start by looking at each layer in more detail.

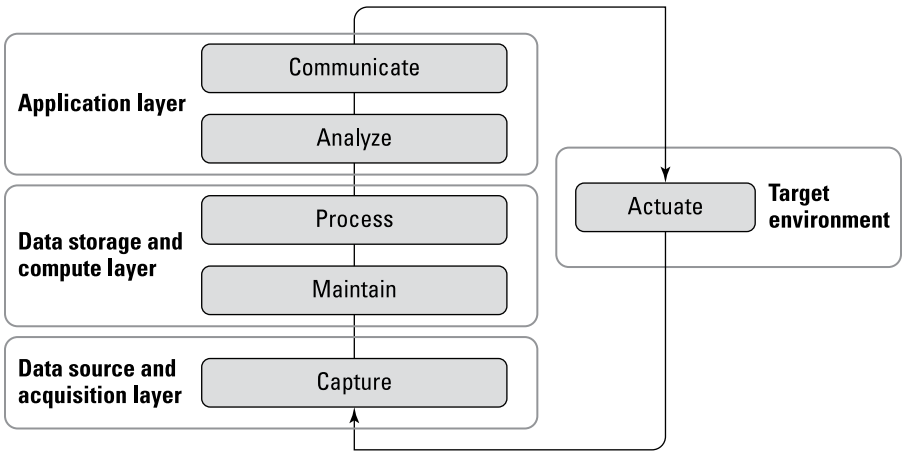

FIGURE 14-1: Using the data science flow to define your data architecture.

- **Data source and acquisition layer:** This layer is all about making sure you understand your data needs based on your business objective and what this means in terms of where the data resides, who owns it, how large it is, whether it's sensitive (and therefore needs to be anonymized), how you should collect it, collection frequency, and so on. It's also advisable to perform the first data processing activities already at the point of collection, because you don't want to spend time and money collecting and storing dirty data. Examples of early processing activities prior to storage include validating completeness of the data collected, identifying and removing duplicated data records, data aggregation to minimize transfer capacity impact, data anonymization for personal data, data encryption, and so on.
- **Data storage and compute layer:** In this layer of the data architecture, you need to consider aspects such as how you want to store the data (retention periods for different types of data, for example). You also need to consider the next level of data processing (data cleansing, mapping, labeling, and so

on), and how this can be done with as little manual intervention as possible. (Automating data management tasks helps to protect data integrity by not unconsciously adding bias to the data.) In this layer, you also need to decide whether you want to store and process the data using an on-premises solution or a cloud solution. What you select could depend on the size of the data you're working with, but also on whether you expect the data to grow rapidly and unexpectedly, which would suggest that you need a cloud based solution for fast and easy scale-up of the environment. However, a public cloud solution — from Amazon, Azure, or Google for example — fits perfectly (even for small data environments) since it removes the upfront cost of investing in your own infrastructure as you only pay for capacity used. Another factor to consider is that, if you need to collect and compute data from many different countries in your environment, you might not be able to transfer the data out of all countries due to strict laws and regulations. To solve that problem, you should consider a distributed cloud setup, where data can be processed in a data center within country borders but where insights and models can be transferred to a central instance for reuse and sharing across the distributed setups.

- **Application layer:** The application layer is straightforward. As its name implies, it's where you implement the applications and tools you want to run on top of your data. It can be a mix of different applications, like open source tools and frameworks such as TensorFlow, Scikit-learn, or Keras, but also out-of-the-box applications from established analytics vendors such as SAS, IBM, or Tableau. Here it's important to think through what type of users you will have in your environment as well as their level of competence and interest. Maybe all you have to do is ensure that insights are easily communicated through various predesigned dashboards and visualizations for decision support; but the best approach is usually to make sure you build the data architecture in such a way that you can swap applications in and out, depending on what becomes available in the industry, and also to support the changing expectations and needs in the organization. Users might start out by wanting things to be provided to them, but as maturity grows, users might want to do more themselves. This is especially true for companies investing in commercial data products and services.

REMEMBER

Flexibility in the application layer isn't the major cost in the architecture. If the bottom layers in the architecture are common to all, a lot is won in terms of keeping costs down and data integrity and reliability in the data up. Flexibility in the application layer ensures that users stay happy and that various needs are met. It also minimizes the risk of dissatisfied users branching off and building their own environment, in the process increasing total company cost as well as creating siloed data and insights.

- **Target environment:** This refers to the environment where you intend to implement your insights and models. Again, what this really ends up being varies a lot from company to company. If the data architecture is built for internal analytics needs only, the target environment might be various internal systems, but it could also refer to how insights flow into, and are used in, various organizational structures and decision forums. In a live operational setting or for a commercial data product or service, the target environment could refer to the live production environment. The output from these target environments then feed new data back to the data source and acquisition layer, providing feedback data through the change it has implemented as part of the live production environment, and the data science cycle starts again.

Listing the Essential Technologies for a Modern Data Architecture

The drive today is to refactor the enterprise technology platform to enable faster, easier, more flexible access to large volumes of precious data. This refactoring is no small undertaking and is usually sparked by a shifting set of key business drivers. Simply put, the platforms that have dominated enterprise IT for nearly 30 years can no longer handle the workloads needed to drive data-driven businesses forward.

Organizations have long been constrained in their use of data by incompatible formats, limitations of traditional databases, and the inability to flexibly combine data from multiple sources. New technologies are now starting to deliver on the promise to change all that. Improving the deployment model of software is one major step to removing barriers to data usage. Greater data agility also requires more flexible databases and more scalable real-time streaming platforms. In fact, no fewer than seven foundational technologies are needed to deliver a flexible, real-time modern data architecture to the enterprise. These seven key technologies are described in the following sections.

NoSQL databases

The relational database management system (RDBMS) has dominated the database market for nearly 30 years, yet the traditional relational database has been shown to be less than adequate in handling the ever-growing data volumes and the accelerated pace at which data must be handled. NoSQL databases — "no SQL" because it's decidedly nonrelational — have been taking over because of their speed and ability to scale. They provide a mechanism for storage and retrieval of

data that is modeled in means other than the tabular relations used in relational databases. Because of their speed, NoSQL databases are increasingly used in big data and real-time web applications.

NoSQL databases offer a simplicity of design, simpler horizontal scaling to clusters of machines (a real problem for relational databases), and finer control over availability. The data structures used by NoSQL databases (key-value, wide column, graph, or document, for example) are different from those used by default in relational databases, making some operations faster in NoSQL. The particular suitability of a given NoSQL database depends on the problem it must solve. Sometimes the data structures used by NoSQL databases are also viewed as more flexible than relational database tables.

Real-time streaming platforms

Responding to customers in real-time is critical to the customer experience. It's no mystery why consumer-facing industries —Business-to-Consumer (B2C) setups, in other words — have experienced massive disruption in the past ten years. It has everything to do with the ability of companies to react to the user in real-time. Telling a customer that you will have an offer ready in 24 hours is no good because they will have already executed the decision they made 23 hours ago. Moving to a real-time model requires event streaming.

Message-driven applications have been around for years, but today's streaming platforms scale far better and at far lower cost than their predecessors. The recent advancement in streaming technologies opens the door to many new ways to optimize a business. Reacting to a customer in real-time is one benefit. Another aspect to consider is the benefits to development. By providing a real-time feedback loop to the development teams, event streams can also help companies improve product quality and get new software out the door faster.

Docker and containers

Docker is a computer program that performs operating-system-level virtualization, also known as *containerization.* First released in 2013 by Docker, Inc., Docker is used to run software packages called *containers,* a method of virtualization that packages an application's code, configurations, and dependencies into building blocks for consistency, efficiency, productivity, and version control. Containers are isolated from each other and bundle their own application, tools, libraries, and configuration files and can communicate with each other by way of well-defined channels.

All containers are run by a single operating system kernel and are thus more lightweight than virtual machines. Containers are created from images that specify their precise content. A container image is a self-contained piece of software that includes everything that it needs in order to run, like code, tools, and resources.

Containers hold significant benefits for both developers and operators as well as for the organization itself. The traditional approach to infrastructure isolation was that of static partitioning, the allocation of a separate, fixed slice of resources, like a physical server or a virtual machine, to each workload. Static partitions made it easier to troubleshoot issues, but at the significant cost of delivering substantially underutilized hardware. Web servers, for example, would consume on average only about 10 percent of the total computational power available.

The great benefit of container technology is its ability to create a new type of isolation. Those who least understand containers might believe they can achieve the same benefits by using automation tools like Ansible, Puppet, or Chef, but in fact these technologies are missing vital capabilities. No matter how hard you try, those automation tools cannot create the isolation required to move workloads freely between different infrastructure and hardware setups. The same container can run on bare-metal hardware in an on-premises data center or in a virtual machine in the public cloud. No changes are necessary. That is what true workload mobility is all about.

Container repositories

A *container image repository* is a collection of related container images, usually providing different versions of the same application or service. It's critical to maintaining agility in your infrastructure. Without a DevOps process with continuous deliveries for building container images and a repository for storing them, each container would have to be built on every machine in which that container could run. With the repository, container images can be launched on any machine configured to read from that repository. Where this gets even more complicated is when dealing with multiple data centers. If a container image is built in one data center, how do you move the image to another data center? Ideally, by leveraging a converged data platform, you will have the ability to mirror the repository between data centers. A critical detail here is that mirroring capabilities between on-premises and the cloud might be vastly different than between your on-premises data centers. A converged data platform will solve this problem for you by offering those capabilities regardless of the physical or cloud infrastructure you use in your organization.

Container orchestration

Instead of static hardware partitions, each container appears to be entirely its own private operating system. Unlike virtual machines, containers don't require a static partition of data computation and memory. This enables administrators to launch large numbers of containers on servers without having to worry so much about exact amounts of memory. With container orchestration tools like Kubernetes, it becomes easy to launch containers, kill them, move them, and relaunch them elsewhere in an environment.

Assuming that you have the new infrastructure components in place (a document database such as MapR-DB or MongoDB, for example) and an event streaming platform (maybe MapR-ES or Apache Kafka) with an orchestration tool (perhaps Kubernetes) in place, what is the next step? You'll certainly have to implement a DevOps process for coming up with continuous software builds that can then be deployed as Docker containers. The bigger question, however, is what you should actually deploy in those containers you've created. This brings us to microservices.

Microservices

Microservices are a software development technique that structures an application as a collection of services that

- Are easy to maintain and test
- Are loosely coupled
- Are organized around business capabilities
- Can be deployed independently

As such, microservices come together to form a microservice architecture, one that enables the continuous delivery/deployment of large, complex applications and also enables an organization to evolve its *technology stack* — the set of software that provides the infrastructure for a computer or a server. The benefit of breaking down an application into different, smaller services is that it improves modularity, which then makes the application easier to understand, develop, and test and to become more resilient to *architecture erosion* — the violations of a system's architecture that lead to significant problems in the system and contribute to its increasing fragility. With a microservices architecture, small autonomous teams can run in parallel to develop, deploy, and scale their respective services independently. It also allows the architecture of an individual service to emerge through

continuous *refactoring* — a disciplined technique for restructuring an existing body of code, altering its internal structure without changing its external behavior (thus ensuring that it continues to fit within the architectural setting).


REMEMBER

The concept of microservices is nothing new. The difference today is that the enabling technologies like NoSQL databases, event streaming, and container orchestration can scale with the creation of thousands of microservices. Without these new approaches to data storage, event streaming, and infrastructure orchestration, large-scale microservices deployments would not be possible. The infrastructure needed to manage the vast quantities of data, events, and container instances would not be able to scale to the required levels.


TECHNICAL STUFF

Microservices are all about delivering agility. A service that is micro in nature generally consists of either a single function or a small group of related functions. The smaller and more focused the functional unit of the work, the easier it will be to create, test, and deploy the service. These services must be *decoupled*, meaning you can make changes to any one service without having an effect on any other service. If this is not the case, you lose the agility promised by the microservices concept. Admittedly, the decoupling must not be absolute — microservices can, of course, rely on other services — but the reliance should be based on either balanced REST APIs or event streams. (Using event streams allows you to leverage request-and-response topics so that you can easily keep track of the history of events; this approach is a major plus when it comes to troubleshooting, because the entire request flow and all the data in the requests can be replayed at any time.)

Function as a service

Just as the microservices idea has attracted a lot of interest in the software industry, so has the rise of server-less computing — perhaps more accurately referred to as Function as a Service (FaaS). Amazon Lambda is an example of a FaaS framework, where it lets you run code without provisioning or managing servers, and you pay only for the computing time you consume.


TECHNICAL STUFF

FaaS enables the creation of microservices in such a way that the code can be wrapped in a lightweight framework built into a container, executed on demand based on some trigger, and then automatically load-balanced, thanks to the aforementioned lightweight framework. The main benefit of FaaS is that it allows the developer to focus almost exclusively on the function itself, making FaaS the logical conclusion of the microservices approach.

REMEMBER

The triggering event is a critical component of FaaS. Without it, there's no way for the functions to be invoked (and resources consumed) on demand. This ability to automatically request functions when needed is what makes FaaS truly valuable. Imagine, for a moment, that someone reading a user's profile triggers an audit event, a function that must run to notify a security team. More specifically, maybe it filters out only certain types of records that are to be marked as prompting a trigger. It can be selective, in other words, which plays up the fact that, as a business function, it is completely customizable. (I'd note that putting a workflow like this in place is tremendously simple with a deployment model such as FaaS.)

The magic behind a triggering service is really nothing more than working with events in an event stream. Certain types of events are used as triggers more often than others, but any event you want can be made into a trigger. The event could be a document update, or maybe running an OCR process over the new document and then adding the text from the OCR process to a NoSQL database. The possibilities here are endless.

TIP

FaaS is also an excellent area for creative uses of machine learning — perhaps machine learning as a service or, more specifically, "a machine learning function aaS." Consider that whenever an image is uploaded, it could be run through a machine learning framework for image identification and scoring. There's no fundamental limitation here. A trigger event is defined, something happens, the event triggers the function, and the function does its job.

WARNING

FaaS is already an important part of microservices adoption, but you must consider one major factor when approaching FaaS: vendor lock-in. The idea behind FaaS is that it's designed to hide the specific storage mechanisms, the specific hardware infrastructure, and the software component orchestration — all great features, if you're a software developer. But because of this abstraction, a hosted FaaS offering is one of the greatest vendor lock-in opportunities the software industry has ever seen. Because the APIs aren't standardized, migrating from one FaaS offering in the public cloud to another is difficult without throwing away a substantial part of the work that has been performed. If FaaS is approached in a more methodical way — by leveraging events from a converged data platform, for example — it becomes easier to move between cloud providers.

Creating a Modern Data Architecture

In many larger companies, the IT function is usually tasked with defining and building data systems, especially for data generated by internal IT systems. It is many times the case, however, that data coming from external sources — customers, products, or suppliers —are stored and managed separately by the

responsible business units. When that's the case, you're faced with the challenge of making sure that all share a common data architecture approach, one that enables all these different data types and user needs to come together by means of an efficient and enabling data pipeline. This data pipeline is all about ensuring an end-to-end flow of data, where applied data management and governance principles focus on a balance between user efficiency and ensuring compliance to relevant laws and regulations.

In smaller companies or modern data-driven enterprises, the IT function is usually highly integrated with the various business functions, which includes working closely with data engineers in the business units in order to minimize the gap between IT and the business functions. This approach has proven very efficient.

So, after you decide which function will set up and drive which part of the data architecture, it's time to get started. Using the step-by-step guide provided in this list, you'll be on your way in no time:

1. **Identify your use cases as well as the necessary data for those use cases.**

   The first step to take when starting to build your data architecture is to work with business users to identify the use cases and type of data that is either the most relevant or simply the most prioritized at that time. Remember that the purpose of a good data architecture is to bring together the business and technology sides of the company to ensure that they're working toward a common purpose. To find the most valuable data for your company, you should look for the data that could generate insights with high business impact. This data may reside within enterprise data environments and might have been there for some time, but perhaps the means and technologies to unearth such data and draw insights from it have been too expensive or insufficient. The availability of today's open source technologies and cloud offerings enable enterprises to pull out such data and work with it in a much more cost-effective and simplified way.

2. **Set up data governance.**

   It is of the utmost importance that you make data governance activities a priority. The process of identifying and ingesting data as well as building models for your data needs to ensure quality and relevance from a business perspective is important and should also include efficient control mechanisms as part of the system support. Responsibility for data must also be established, whether it concerns individual data owners or different data science functions. (For more on data governance, see Chapter 15.)

3. **Build for flexibility.**

   The rule here is that you should build data systems designed to change, not ones designed to last. A key rule for any data architecture these days is to not

build in dependency to a particular technology or solution. If a new key solution or technology becomes available on the market, the architecture should be able to accommodate it. The types of data coming into enterprises can change, as do the tools and platforms that are put into place to handle them. The key is therefore to design a data environment that can accommodate such change.

4. **Decide on techniques for capturing data.**

   You need to consider your techniques for acquiring data, and you especially need to make sure that your data architecture can at some point handle real-time data streaming, even if it isn't an absolute requirement from the start. A modern data architecture needs to be built to support the movement and analysis of data to decision makers when and where it's needed.

REMEMBER

   Focus on real-time data uploads from two perspectives: the need to facilitate real-time access to data (data that could be historical) as well as the requirement to support data from events as they're occurring. For the first category, existing infrastructure such as data warehouses have a critical role to play. For the second, new approaches such as streaming analytics and machine learning are critical. Data may be coming from anywhere — transactional applications, devices and sensors across various connected devices, mobile devices and, telecommunications equipment, and who-knows-where-else. A modern data architecture needs to support data movement at all speeds, whether it's sub-second speeds or with 24-hour latency.

5. **Apply the appropriate data security measures.**

   Do not forget to build security into the architecture. A modern data architecture recognizes that threats to data security are continually emerging, both externally and internally. These threats are constantly evolving and may be coming through email one month and through flash drives the next. Data managers and data architects are usually the most knowledgeable when it comes to understanding what is required for data security in today's environments, so be sure to utilize their expertise.

6. **Integrate master data management.**

   Make sure that you address *master data management,* the method used to define and manage the critical data of an organization to provide, with the help of data integration, a single point of reference. With an agreed-on and built-in master data management (MDM) strategy, your enterprise is able to have a single version of the truth that synchronizes data to applications accessing that data. The need for an MDM-based architecture is critical because organizations are consistently going through changes, including growth, realignments, mergers, and acquisitions. Often, enterprises end up with data systems running in parallel, and often, critical records and information may be duplicated and overlap across these silos. MDM ensures that applications and systems across the enterprise have the same view of important data.

7. **Offer data as a service (aaS).**

   This particular step is a relatively new approach, but it has turned out to be quite a successful component — make sure that your data architecture is able to position data as a service (aaS). Many enterprises have a range of databases and legacy environments, making it challenging to pull information from various sources. With the aaS approach, access is enabled through a virtualized data services layer that standardizes all data sources, regardless of device, applicator, or system. Data as a service is by definition a form of internal company cloud service, where data — along with different data management platforms, tools, and applications — are made available to the enterprise as reusable, standardized services. The potential advantage of data as a service is that processes and assets can be prepackaged based on corporate or compliance standards and made readily available within the enterprise cloud.

8. **Enable self-service capabilities.**

   As the final step in building your data architecture, you should definitely invest in self-service environments. With self-service, business users can configure their own queries and get the data or analyses they want, or they can conduct their own data discovery without having to wait for their IT or data management departments to deliver the data. The route to self-service is providing front-end interfaces that are simply laid out and easy to use for your target audience. In the process, a logical service layer can be developed that can be reused across various projects, departments, and business units. IT could still have an important role to play in a self-service-enabled architecture, including aspects such as data pipeline operations (hardware, software, and cloud) and data governance control mechanisms, but it would have to spend less and less of its time and resources on fulfilling user requests that could be better formulated and addressed by the user themselves.

IN THIS CHAPTER

- Describing what data governance is all about
- Making the case for the value of data governance
- Introducing data stewardship
- Setting up a data governance structure

Chapter 15
Focusing Data Governance on the Right Aspects

Now more than ever, the ability to handle vast amounts of data in a way that manages to balance risks and business opportunities is critical to a company's success. Yet even with the emergence of data management functions and chief data officers (CDOs), most companies aren't performing at their best when it comes to managing and monetizing their data. Cross-industry studies show that, on average, less than half of an organization's structured data is actively used in making decisions and less than 1 percent of its unstructured data is analyzed or used at all. Many times, companies lock in the data, just to be on the safe side, and employees are forced to spend a lot of time explaining why they need a certain data set. At the same time, data breaches are more and more common, as data sets are spread in silos and their value is many times not understood or handled accordingly. On top of that, a common problem is that the company data infrastructure and applications don't live up to the expectations placed on them. In this chapter, I walk you through the key elements of data governance and guide you on the best way to approach the topic.

Sorting Out Data Governance

The concept of data governance refers to the people, processes, and system support required to establish a consistent and correct handling of an organization's data across the company. It supports data management with the necessary foundation, strategy, and structure needed to ensure that data is managed as an asset and transformed into meaningful and actionable insights.

Figure 15-1 shows that data governance is a capability that enables an organization to ensure the high data quality of your data throughout its complete life cycle. The key focus areas of data governance include ensuring availability, usability, consistency, integrity, and security of your data at all times and includes establishing processes to ensure effective data governance throughout the organization. The different areas managed through data governance can be explained according to this list:

- **Data availability:** This term is used to describe to what extent a data element can be easily accessed at any level of performance. The level of data availability can be measured by factors like how easy the data is to manage and maintain, the ability to restore or recover any services or data in case of any error or failure, the ability to deliver a service, and the ability to understand problems with the data, diagnose their root cause, and repair them as soon as possible.
- **Data usability:** This refers to the state of the data you currently have (in its raw format) and how it fits its purpose. How do you know whether your data is usable? Here are some questions you can ask: Is the raw data value correct or incorrect? How granular and precise are the data attributes? How integrated is the data with other data sources and data objects?
- **Data consistency:** This term is used to describe how useful and reliable the data is from a trustworthiness perspective. Consistency is usually checked from three main perspectives: point-in-time consistency (that data stay consistent over time), transaction consistency (that data stays consistent during a transaction), and application consistency (that data stays consistent between different applications).
- **Data integrity:** This refers to the maintenance and assurance of the accuracy and consistency of data over its entire life cycle. It's a critical aspect of the design, implementation, and usage of any system that stores, processes, or retrieves data. (Data integrity is the opposite of data corruption.)
- **Data security:** This refers to protecting digital data, such as those in a database, from destructive forces and from the unwanted actions of unauthorized users — users who might initiate a cyberattack or a data breach, for example.

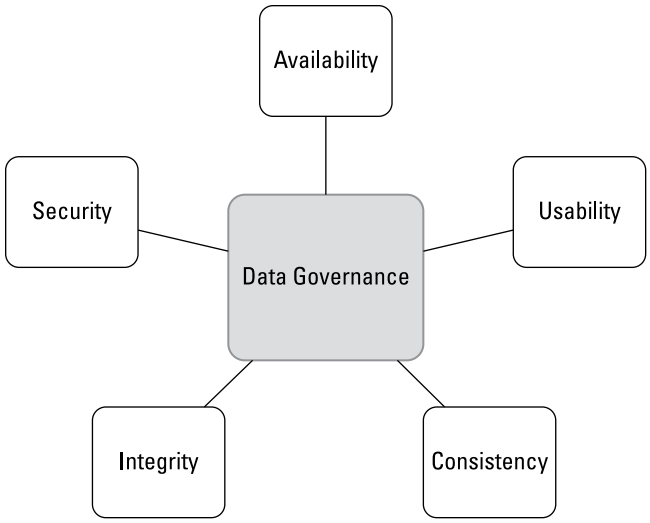

FIGURE 15-1: The data aspects managed by data governance.

REMEMBER

As a starting point, understand and seek company agreement on what is expected from the data governance activities and frameworks. Is it something you engage in just to make sure you're toeing the line when it comes to laws and regulations, or is the ambition to leverage data governance as an enabler for a greater, much more reliable business? The starting point is important because it will help you determine your priorities and approach when establishing and implementing data governance.

Data governance for defense or offense

When it comes to data governance, you can play either defense or offense. Data defense is all about minimizing risk, and includes aspects such as ensuring compliance with regulations, using analytics or machine learning to detect and limit fraud, and building systems to prevent theft. Defensive efforts also include measures designed to ensure the integrity of data flowing through a company's internal systems. This happens by way of identifying, standardizing, and governing main data sources (customer, product, or sales data, for example) so that the company can rely on a single source of truth for its most important data.

Data offense, on the other hand, focuses on supporting business objectives such as increasing revenue, profitability, and customer satisfaction. It typically includes activities like data analysis and modeling that are designed to generate customer insights in order to support management decision-making. However, data offense also includes activities related to pursuing business opportunities with data products or services based on research or development work.

TIP

Every company needs both offense and defense to succeed, but getting the balance right is complicated and may differ a lot between different types of companies. These two approaches usually compete for limited resources, funding, and people. However, please note that while putting equal emphasis on the two is optimal for some companies, for others it might be a lot wiser to favor one or the other. This is highly impacted by the company context related to which industry it's part of and how competitive the environment is, as well as regulatory limitations.

Hospitals, for example, exist in a low competitive market with a highly regulated context where requirements on data quality and privacy protection are extremely high. They must therefore prioritize data defense over offense. Companies in the retail business, on the other hand, are much less regulated and are therefore able to work with sensitive personal data as part of a strategy for beating the competition and responding quickly to market changes. These types of companies typically prioritize offence over defense.

Objectives for data governance

You (or perhaps your bosses) might be asking why data governance is so important. Right off the bat I can tell you that ineffective data governance within a company inevitably leads to one thing: poor data. This poor data is visible through inconsistent definitions, duplicates, missing fields, and other classic data problems. These are clear issues that should be avoided.

Goals for data governance may be defined at all levels of the company, but by encouraging stakeholders to contribute to the goal setting, you may be able to ensure that they recognize the importance of data governance. In the following list, you can find some useful examples of data governance objectives, divided into the defensive and offensive camps:

Defensive:

- Increasing consistency and confidence in decision-making
- Decreasing the risk of regulatory fines
- Improving data security as well as defining and verifying the requirements for data distribution policies
- Designating accountability for data quality
- Enabling better planning by supervisory staff
- Reducing operational friction
- Protecting the needs of data stakeholders
- Training management and staff to adopt common approaches to data issues

Offensive:

- Maximizing the income generation potential of commercial data products
- Encouraging high degrees of sharing and reuse of data and insights
- Using data derived insights to inform business investment decisions
- Seeing that research and development activities are driven by data, analytics, and machine learning/artificial intelligence and are focused on exploring new business opportunities
- Optimizing staff effectiveness
- Minimizing or eliminating "re-work"
- Establishing process performance baselines to enable improvement efforts

REMEMBER

You can realize any of these goals by implementing data governance programs or launching initiatives that use change management techniques.

Explaining Why Data Governance is Needed

Imagine that you've submitted a budget designed to beef up your data governance practices. You hear back that not a penny will be granted unless you can convince upper management that it's a good investment. What are the best arguments to use to save your initiatives as well as your budget? Read on to find out.

Data governance saves money

First and foremost, data governance increases efficiency in the organization. Duplicated accounts lead to duplicate efforts or, at the very least, to time wasted tracking down duplicate accounts in your marketing, sales, finance, development, or analytical efforts. Data governance reduces errors in the source data, giving your business a solid base to work from, and saves precious time that would otherwise be used correcting your existing data. Time saved is money saved.

TIP

Data governance forces your company to define its core data as well as the rules governing that core data. The initiation of a data governance project could be a golden opportunity to get everyone on the same page about core data definitions. The enforcement of these definitions ensures greater operational efficiency over time.

Bad data governance is dangerous

Lack of effective data governance is a security concern for two reasons:

- There are outside security risks associated with dirty, unstructured data.
- Bad data governance can result in regulatory compliance issues.

WARNING

Badly structured data poses a security risk for the simple reason that if you have dirty, unstructured data clogging your data pipeline, you can't quickly tell when something is about to go wrong and you can't efficiently monitor which data is at risk.

Regulatory compliance and data governance is becoming a hotter topic with each passing day. As people are becoming more aware of the importance of their personal data, governments are beginning to take extremely seriously how companies store, protect, and use their customers' data. With a messy, ungoverned data swamp, it may prove impossible for a company to guarantee that all data regarding a particular individual is deleted when requested. This opens up your company to great risk and potentially enormous fines.

Good data governance provides clarity

Effective data governance provides the peace of mind that your company's data is generally clean, standardized, and accurate. The effects of this reassurance resonate throughout a company and provide important benefits. An obvious but important benefit is assurance that the integrity in your data is kept over its life cycle, meaning that data is trusted and used as a base for important business decisions as well as for research and development of new insights or data products and services.

Establishing Data Stewardship to Enforce Data Governance Rules

Within an organization, a data steward is responsible for utilizing an organization's data governance processes to ensure the fitness of its various data elements. As such, data stewards carry out a specialist role that incorporates processes, policies, guidelines, and responsibilities for administering a company's entire data scope in compliance with policy and/or regulatory requirements.

Data stewardship is concerned with taking care of data assets that don't belong to the stewards themselves and may represent the needs of the entire organization. Others may be tasked with representing a smaller data scope related to a particular business unit or department or even a certain type of data. In some organizations, data stewards are senior representatives of appointed stakeholder groups — a structure designed to ensure sufficient engagement in — and decisions about — the treatment of certain data assets. However, in other organizations, data stewards operate independently, ensuring that the general rules and controls are applied to data appropriately throughout the organization.


REMEMBER

The overall objective of a data steward is ensuring data quality for the main data elements that have been decided on. This includes capturing metadata for each data element, such as definitions, related rules/governance, physical manifestations, and related data models. Data stewards begin the stewarding process with the identification of the elements that they will steward, with the ultimate result of standards, controls, and data entry.


TIP

The data steward needs to work closely with stakeholders involved with data standardization in order to drive alignment on data standards; with data architects in order to understand and secure adherence to data dependencies; and with system support experts in order to secure automated and built-in control mechanisms for data quality checks. Data controls can be preventive, detective, and corrective and be executed manually, aided by technology, or completely automated.

Implementing a Structured Approach to Data Governance

All organizations need to be able to make decisions about how to manage data, realize value from it, minimize cost and complexity, manage risk, and ensure compliance with ever-growing legal, regulatory, and other requirements. A company needs to create rules, ensure that the rules are being followed, and be able to deal with noncompliance, ambiguities, and any data-related issues that may arise.


REMEMBER

In short, a company needs to do more than just manage data. There's a need for a governance system that sets the rules of engagement for management activities across the organization. Small organizations or ones with simple data environments may be able to succeed in these goals through an informal system of governance. They may not even be aware of when they're switching between making management decisions and broader governance decisions. On the other hand, larger organizations or ones with more complex data or compliance environments generally find that they need to step back and agree on a more formal system of governance.

Defining and establishing a framework or a structured approach to data governance includes activities such as these:

- **Define your objectives.** Ask yourself whether a defensive or offensive approach is most important for your line of business, market situation, and regulatory situation.
- **Decide on the focus area.** You have to start somewhere, so ask yourself where your data governance project starts — full scope across the company or only for a selected unit or department?
- **Set data definitions and rules.** What, exactly, do you need to govern? Avoid casting your net too widely, because it might hinder innovation and efficiency.
- **Specify decision rights.** Who will be able to decide on the governance rules? What is the data governance framework needed for your company?
- **Define and implement control mechanisms.** How will you ensure that rules are followed? Avoid manual (human) control as much as possible by building in control mechanisms to your data systems with as high a level of automation as possible.
- **Identify data stakeholders.** Determine who'll be using the data and how. Take the time to really understand the business need and the stakeholders for the selected focus area.
- **Set up a data governance board (DGB).** Data governors make up the DGB. They are ultimately accountable for business data use, data quality, and prioritization of data-related issues. They make decisions that impact data based on recommendations from the data stewards. The board has the authority to decide how the budget for data management improvements related to data governance shall be spent. This step is applicable mainly for larger companies or larger implementations when complexity and dependencies are significant.
- **Assigning and training data stewards.** Who will be working with data governance on a daily basis? How will you capture the need across the organization and make sure that the framework stays relevant over time? Establishing the role of data stewards to manage data governance according to your line of business, is an important step.
- **Designing and implementing needed processes**. The final step is to make sure you have processes in place that are sufficient yet clear and simple for your company to execute satisfactory data governance.

REMEMBER

One of the most important factors when it comes to data governance is ensuring that there's a common, agreed-on set of principles and best practices common to all teams and individuals in charge of collecting, governing, and consuming the data. Ensure that everyone is on board and that there are clear goals, clearly defined processes, and clear permission levels to make everything run smoothly.

TIP

The key to data governance is effective collaboration. The right data governance tools should go hand in hand with these principles. Make sure that whichever tools you're evaluating for adoption are easy to use for business and IT users alike, enable seamless collaboration across teams, and are flexible enough to evolve with your changing business needs.

IN THIS CHAPTER

- Describing the basics of managing models
- Sorting out why model management is important
- Considering key steps for implementing model management
- Identifying and handling model risks

Chapter 16

Managing Models During Development and Production

Although managing data is essential in order to succeed with your data science investment, understanding why model management is a key part is equally important. In this chapter, I briefly explore what model management is all about, as well as list some of the important aspects to consider when it comes to model development and deployment.

Unfolding the Fundamentals of Model Management

An *algorithm* is a step-by-step method of solving a problem, commonly used for data processing, calculation, and other related computer and mathematical operations. An algorithm is also used to manipulate data in various ways, such as inserting a new data item, searching for a particular item, or sorting an item.

Technically, computers use algorithms to list the detailed instructions for carrying out a task. For example, to compute an employee's paycheck, the computer uses an algorithm. To accomplish this task, appropriate data must be entered into the system. In terms of efficiency, various algorithms are able to accomplish operations or problem solving easily and quickly.

So, an algorithm is the general approach you will take. The *model* is what you get when you run the algorithm over your training data and subsequently use to make predictions on new data. You can generate a new model with the same algorithm but with different data, or you can get a new model from the same data but with a different algorithm.

Working with many models

A common misunderstanding around machine learning models when compared to the software development space is the notion that the objective of a machine learning modeling session is to build one successful model, deploy it, and then walk away, patting yourself on the back for a job well done. That's a fantasy. In reality, working with machine learning models involves working with many models over an extended period, even after the model has been deployed in production. It's quite common to have several models in production at the same time and have new models ready to replace older models in production when conditions change. It's also important to be able to manage these model replacements in a smooth manner, without disturbing the ongoing service. In model development, you also work with more than one model as you experiment with multiple tools and compare model performance in order to find the best-performing model.

That's the general lay of the land, but it doesn't quite explain what model management is really all about. To get a better handle on that, consider a situation where an organization has hundreds of models embedded in various production systems to support decision-making in marketing, pricing, credit risk, operational risk, fraud, and finance functions. In this example, which is a common one across the software industry, data scientists across different business units are free to develop their models with no formalized or standardized processes for storing, deploying, and managing them. Some models don't have the necessary documentation describing the model's owner, business purpose, usage guidelines, or other information necessary for managing the model or explaining it to regulators, because the units were told to prioritize speed over proper documentation. Furthermore, after the model results are achieved in this imaginary scenario, they're subjected to limited controls and requirements as they make their way to decision-makers. Not surprisingly, because different data sets and variables were used to create the models, the results turn out to be inconsistent. There's little validation or back-testing for accuracy, and decisions are made on the model results as is — and then everyone just hopes for the best.

The scenario just described, with total modeling confusion, may look all too familiar to many organizations. In a diverse and loosely managed modeling environment, it can become quite difficult to answer critical questions about the predictive analytical or machine learning models that your organization relies on for not only the day-to-day business operations but also strategic decision-making. It's simply vital for your organization to build a solid foundation for managing models in a reliable, transparent, scalable, and reusable manner. But how should that be done?

A good starting point when reviewing your current model management situation or trying to determine how your architectural team conceives of it is to ask the following questions:

- Are we tracking who created the models and why?
- Do we know which input variables are used to make predictions or, in the machine learning case, which training data was used to train the model?
- Are we keeping track of how the models are used?
- Are we measuring model performance and do we know when these models were last updated?
- Is there enough supporting documentation in place to enable model reuse by other data science teams?
- Is it taking a long time to put new or updated models into production?

Faced with this list of questions, companies tend to respond in one of two ways: They're either able to answer these questions positively, but come to the conclusion that more could be done to increase efficiency and value from the models, or they're not able to answer any of these questions affirmatively. Why the latter? Because they haven't realized the importance of good model management in a company driven by data and machine learning/artificial intelligence. In a data- and model-driven enterprise, models are at the heart of critical business decisions. Models can identify new opportunities, help you forge new or better relationships with customers, and enable you to manage uncertainty and risks. For these reasons and many more, they should be created and treated as high-value organizational assets.


TIP

Model management isn't just about applying new guidelines or a new governance structure; you need to have software on hand that can wrangle your data into shape and quickly create many accurate models you can rely on. On top of that, it takes efficient and repeatable processes that are fully supported by a reliable infrastructure and your various architectural elements in order to manage and trace your models for optimal performance throughout their entire life cycle.

Making the case for efficient model management

The notion that efficient model management is crucial for the success of an organization is gaining more and more ground as the importance of data science is getting recognized in a broader sense. Some even say that, going forward, it will be a stronger competitive advantage to be model-driven, as opposed to being only data-driven. It also looks like the evidence is piling up that companies that are successful in getting value out of their data science investment are the ones that treat models as a new type of business asset.

Companies most successful in data science today treat models very differently from how they treat data and software. The way that they build models, develop models, deploy models, manage and have governance around models, as well as how they create the technology infrastructure systems to support models, are different from what they've done in the past when setting up systems for data or software. Why is that?

First of all, the raw materials data scientists use to create models are different from other business assets because models require computationally intensive algorithms. That requirement, of course, drives the growing need for elastic, scalable computational power as well as for specialized hardware, like graphics processing units (GPUs). Those are architectural components that software engineering teams don't normally need.

TIP

Another critical raw material for model development is the open source ecosystem. There are new tools, new packages, or updated packages coming out every day, especially around Python and R. If a company is trying to compete and have the best, most innovative models, they need ways to give data scientists quick access to the very rapidly evolving ecosystem without stifling their flexibility.

The third property that differentiates models from other types of software development is the process. Models by their very nature emerge from a research process, and such processes are inherently experimental, emergent, and exploratory. That's quite different from how software development works, and it's quite different from how systems acquire data. A data science team developing models might try hundreds of ideas before finding one that works; that's just fine, but it does create different requirements on the underlying infrastructure. Teams developing models need different capabilities to facilitate rapid experimentation and rapid exploration so that they can drive breakthroughs. In software, it's about de-risking and driving to clarity of requirements. In model development, it's about rapid experimentation where you try as many ideas as you can as fast as possible.

The fourth critical property of models to keep in mind is how they behave. In software engineering, there is typically a specification that developers aim for, and tests that can confirm whether the spec has been met. There's nothing of the sort when building predictive models. Instead, data science models are probabilistic. They don't have a correct answer. They just have better or worse answers when they're alive in the real world. What that means is that organizations need new ways for quality control, monitoring, governing, and reviewing models to ensure model reliability and anticipated behavior by the algorithm, meaning that it's performing as expected.

The unique requirements for succeeding with model management in data science suggests that model management should be treated as a dedicated discipline. Data infrastructures should not limit model management to a software application but should instead incorporate much more of a model driven approach, not only data driven.

Implementing Model Management

The next breakthrough in data science will probably not be new revolutionary algorithms (those will keep on coming, no matter what), but rather the ability to rapidly combine, deploy, and maintain existing algorithms in rapidly changing live environments.

Many corporations have now realized the need for a centralized repository for storing predictive models along with detailed metadata for efficient workgroup collaboration and version control of various models. Successful model management involves a collaborative team of modelers, architects, model scoring managers, model auditors, and validation testers. However, many companies are struggling with the process of signing off on the development, validation, deployment, and retirement life cycle management milestones for their models.

REMEMBER

You must readily know exactly where each model is in the life cycle, how old the model is, who developed the model, and who is using the model for what application. The ability to version-control the model over time is another critical business need that includes event logging and tracking changes to understand how the model form and usage is evolving over time.

Model degradation, where the model is no longer performing as expected and model accuracy is declining, is another serious challenge faced by many organizations. Standardizing the metrics used for measuring model performance is urgently needed. Currently, it's up to each company — or even each data scientist — to define and determine when a model needs to be retrained or

replaced. On top of that, there's always a need for managing retired models because they need to be archived and not just thrown away. Finally, a more reliable process for managing your model scoring is a must because it's a key requirement to ensure that you can evaluate model performance and profitability over time.

Successful organizations recognize that models are essential corporate assets that produce and deliver answers to production systems for improved customer relationships, improved operations, increased revenues, and reduced risks. Few companies, however, are capable of fully managing all the complexities of the complete model life cycle, just because it's such a multifaceted task. So, if you can't do it all, what should you focus on? In other words, what does it take, from a model management perspective, to be able to make a lot of good, fast operational decisions that consistently reflect overall organizational strategy and at the same time keep your organization faster and better than anyone else's? Well, as in most cases, there's no silver bullet, but there are some main aspects to focus on:

- Data-driven systems**:** The entire operational setup in the company *must* use data to produce answers for people or systems, depending on your level of automation, so that the right actions are initiated.
- Reliable and updated models**:** Having relevant and up-to-date models that the business can rely on for optimal decisions and actions at the right time is key. These decisions can be made by machine intelligence driven systems that are using your models for automated decision-making in an operational setting.
- Integrated business rules**:** The integration of business rules and a predictive analytical approach into operational decision flows has to occur if you want to provide the instructional insight needed for vetted, trusted decisions.
- Model monitoring**:** Nothing will work for you in the long run if you don't find a way to manage and monitor your analytical models to ensure that they're performing well and continue to deliver the right answers.
- A modern **data architecture:** Make sure you have a modern data architecture that addresses your needs and is supported by efficient and relevant processes that can grow to address new needs, like streaming data and building more detailed predictive models faster than ever.

Pinpointing implementation challenges

Although it might seem easy to move from a well-thought-out data science strategy into the implementation phase, it isn't always the case. There are still many problems that can arise during implementation. Being aware of these common challenges from a model management perspective helps you to navigate around them, or at least gives you some tips on how to handle them once they occur:

- **Getting models into production too slowly:** Because processes are often manual and ad hoc, it can take months to get a model implemented into the production environment. And, because it can take so long to move models through the development and testing phases, they can be stale by the time they reach production — or they never get deployed at all. Internal and external compliance issues can make the process even more challenging, especially as the regulatory situation in data science is evolving at such a rapid pace.
- **Difficulty interpreting model recommendations:** The step of translating model results into business actions for operational decisions requires clear, agreed-on business rules that need to become part of the governed environment because these are the rules that define how you'll use the model results. For example, a fraud detection model might return a fraud risk score as a number between 100 and 1,000 (similar to a FICO credit score). It's up to the business to decide what level of risk requires action. If the trigger for a fraud alert is set too high, fraud might go unnoticed. If the trigger for fraud is set too low, the alerts create too many false positives. Both outcomes will decrease the value these models create and also reduce trust in the results.
- **Models not performing as expected:** Too often, poorly performing models remain in production even though they're producing inaccurate results that lead to bad business decisions. Model results change as the data changes, a reflection of new conditions and behaviors that the model might not be able to adapt to, even if it's a machine learning model. This would not be a problem if the inaccuracies were caught quickly enough, but that's often not the case. The main reasons for this situation are a lack of a central repository for models, no consistent metrics to monitor model performance, and insufficient guidance or control mechanisms to determine when a model needs to be retrained or replaced.
- **Processes for model management not working in practice:** Organizations often find themselves in a reactive mode when put under pressure and are responding in a rush to meet deadlines. (This is especially true for data science teams in their early stages, when their processes are first being established and they feel they have much to prove to management.) It might cause situations where each group has a different approach for handling and validating a model, which can result in a wide array of reports with different levels of detail for review or models that are inconsistently described, making interpretation difficult. No one is sure how the highest scoring model (the champion model) was selected, how a particular model score was calculated, or what governs the business rules that trigger the model.

- **Lack of transparency:** If you don't actively address transparency in model management, you're not going to gain much visibility into the different stages of model development or much knowledge of who touches the model as it goes through its life cycle. In a small company, that may be okay, but in a larger enterprise you'll soon find out that such a situation can be quite cumbersome. Conflicting assumptions may arise and cause additional confusion, and when, as a last resort, unbiased reviewers are called in to validate the models as they pass through each group, you're facing a big resource drain and an additional hit to the development lead-time.
- **Loss of important model knowledge:** With inadequate documentation of models, important intellectual property will stay in the mind of the model creator, severely impacting the ability for model reuse. An approach that heavily depends on key individuals also increases the risk of losing vital information entirely — when that person leaves the company, the knowledge is gone.
- **Insufficient skill sets:** Even with increasing numbers of data scientists entering the marketplace, the shortage of analytical skills needed for model creation and deployment is still a big challenge for many organizations. Without sufficient skillsets in the company, progress can be slow and results poor.

Managing model risk

A vital part of your approach for implementing model management should be focused on understanding and measuring the risk of using and trusting an artificial intelligence model for strategic decision-making and operational setups. Risk goes with the territory because machine learning/artificial intelligence models are *probabilistic:* — they give you the best answer possible, but that answer might still be wrong. It isn't an absolute truth.

Another important risk consideration stems from the fact that machine learning/artificial intelligence models are designed to learn dynamically, which means that, if they're deployed in a dynamic manner, they'll evolve in a live production environment. This also means that the decision framework for the model may change over time, moving away from the principles it was originally trained on in the lab environment when the model is exposed to new data which triggers the model to learn and respond to new behaviors. Therefore, it's important to implement sufficient policy constraints to make sure model learning stays under control.

A more obvious risk you need to manage when it comes to your machine learning/artificial intelligence models lies in the fact that poor training, bias, and bad or

corrupt data can affect your model outcomes. (Garbage in, garbage out.) Securing diverse, and representative training data of sufficient quality is a good start when it comes to lowering this risk.

Measuring the risk level

The purpose of measuring the risk of a machine learning/artificial intelligence model is very much related to understanding and defining the risk profile for a certain model. Simply put, if you're considering using the findings and recommendations derived from a specific model for a very important (and costly) investment decision or for an important customer recommendation, wouldn't you like to know the risk involved with trusting the model for that?

So, what would such a risk assessment look like? You'd need to start by fully understanding the technical risks involved in trusting the model in relation to the impact such misplaced trust would have if the model failed. The technical risks associated with a model include aspects such as these:

- The model's objectives
- Its functional capabilities
- The model's learning approach
- The environmental conditions
- The level of human oversight

The impact is first and foremost measured by the potential financial, emotional, and physical impact a model failure might have to external and internal users, but it also includes estimating the impact from a reputational, regulatory, and legal perspective.


TIP

Once you understand the technical risks and the impact of failure, you need to consider how to establish the right type of control mechanisms. They key here is finding the right balance in your data science infrastructure between a workspace which is innovative and productive while at the same time remaining accurate and reliable.

Identifying suitable control mechanisms

Understanding which control mechanisms to put in place for which risk isn't an easy task, but Table 16-1 shows a couple of examples of how this could be addressed.

TABLE 16-1 Examples of Model Risks and Possible Control Mechanisms

Model Risk	Control Mechanism
Insufficient model supervision	Procedures are in place to monitor the model performance and respond to deviations from the expected performance. Strict safety and control measures are employed to prevent uncontrolled evolution of the model.
Lack of model explainabiity	A data protection impact assessment has been performed and results have been communicated to relevant stakeholders.
Biased results	Collection of a diverse set of training data across all relevant classes to avoid latent bias Use of regularization techniques to penalize for imbalances in selection across targeted data types A diverse team is used to test the model outcomes for latent bias
Poor model performance	Procedures are in place to test the model under varied live conditions to help ensure required performance under deployment Cross-validation and assembling techniques (combining predictions from a few different models) are used to help prevent over-fitting of the model which occurs if the model performance is too tightly connected to a particular data set.

REMEMBER

Although risk awareness and control are definitely good things, data science by its very nature is all about using an experimental- and machine-driven approach to augment human behavior to reach beyond what is possible today. Controlling and constraining that approach too much from an infrastructure and process perspective will only hamper data science model innovation and productivity. The key is to find a balanced approach in terms of how to manage model risk and control versus model innovation and business creativity.

IN THIS CHAPTER

- Understanding the importance of open source in data science
- Explaining open source programming languages
- Exploring open source frameworks and tools
- Deciding what to select

Chapter 17
Exploring the Importance of Open Source

The biggest names in data science are open source, with many of them even part of the same (open source) Apache family: Spark, Hadoop, Kafka, and Cassandra. Though closed source databases are still incredibly popular, open source alternatives are growing at a rapid pace. It is clear that, if they keep growing, those closed source databases won't be that popular for much longer. This chapter focuses on explaining why open source is important in data science, as well as giving you an overview of popular tools and frameworks.

Exploring the Role of Open Source

The popularity of open source systems in data science is growing, for a number of reasons. First, open source principles are based on the sharing of assets, an approach that allows different people in different areas to effectively work together. When companies share their work and allow others to contribute, it

allows for more people to find both new issues and new possibilities. Techniques like deep learning, for example, owe a lot to big players like Google and Facebook, which actively give their data and resources back to the community.


REMEMBER

It always looks as though technology is developing very quickly, but the process itself isn't a rapid one. If companies were to attempt to tackle big data software on their own, with no input or help from various open source software, it would be a painfully slow process. There is a serious need to keep up with the times, and data science is a rapidly growing field with a constant shortage of the right competence and skill sets. This not only affects small businesses looking to keep up but also major investors who could change the course of a business at large. Companies are looking to rapidly expand their data science departments and usages, but the talent pool and technology aren't yet there. Open sourcing that data and technology at least eases the burden and allows companies to move forward at an even pace.


TIP

The community approach also means that users have the chance to ask questions and get helpful answers. Rather than go into a tailspin whenever a problem arises, a user will likely find several others in the community who have the answer or, more likely, who know how to find it. Creative open source users also tend to look for ways to work economically and save money. They are likely to find or tweak inexpensive hardware, whereas a major software company with a monopoly may push users to buy very specific and expensive gear.

Understanding the importance of open source in smaller companies

Once companies decide to put their data to use, they often find themselves absorbed in activities focused on implementing a data lake. The overarching objective suddenly becomes gathering and storing data but without the proper resources or, for smaller companies, the funds to harness them, data is absolutely useless. If a small company were to pay for every bit of software and training required to use data, there would be a much smaller incentive to try to integrate big data into the workplace.

Open source, however, has that try-before-you-buy mentality. For companies that offer products based on open source software, potential customers are often familiar and comfortable with the open source aspects of products. New users can take a chance on data with little risk, and experts can move between different solutions with relative ease.

Understanding the trend

If current trends continue, the entire next-generation data platform will be open source, meaning gains for open source companies and those who build on them. In data science, open source is the norm. Even training in the area supports the open source community by often remaining free. Though university degrees will certainly prove useful in the future, many businesses and programmers are simply looking for further training on big data topics to add to their arsenal. Free online courses in data are plentiful, and programs from Udacity (`www.udacity.com`), IBM's Cognitive Class (`https://cognitiveclass.ai`, formerly Big Data University), and others are trying to fill the gap between data science wannabes and users. Even Google has held free courses on how to use data.

The incredible growth experienced by open source programs and communities is the real proof that it is the future of data. The companies that are taking great steps with data are the ones also pushing open source. This not only proves the effectiveness of open source but also shows where finances in data are headed and just where companies are placing their bets. SAS Institute, an American giant in data and analytics, is putting a lot of its investment in open source compatibility, understanding that the idea isn't meant to compete with open source but rather to figure out how to complement it. For example, in the new SAS cloud-based solution, a data scientist can continue to work in an open source environment using favorite open source tools or programming languages during development, but once it's time to put the algorithm into production, that person can deploy it into the virtualized SAS environment to ensure greater reliability, performance, and monitoring.

The past, present, and future of big data is strongly rooted in open source tech, and that will be one of its greatest strengths. With the shortage of data scientists and skilled workers, it will be paramount that companies and individuals have easy access to powerful and up-to-date solutions without fear of paying every last penny to stay in the game. Especially as companies like Google and Facebook share their knowledge, the future of data will only get better and more powerful.

Describing the Context of Data Science Programming Languages

The landscape of data science is evolving quickly, and tools used for extracting value from data science have also increased in number. To fully utilize the potential with open source tools and frameworks, it's important to strategically make sure that your company builds and acquires skills in open source programming

languages in the data science space. Although there is no specific order to this list of popular languages for data science, Python and R are fighting for the top spot. However, having data scientists with more than one language skill gives your organization more flexibility.

How many open source programming languages are in the running? Let me count the ways:

- **Python:** Python is an extremely popular, general purpose, dynamic, and widely used language within the data science community. It is commonly referred to as the easiest programming language to read and learn. Because it combines quick improvement with the capacity to interface with high performance algorithms written in Fortran or C, it has become the leading programming language for open source data science. With the advancement of technologies such as artificial intelligence, machine learning, and predictive analytics, the demand for experts with Python skills is rising significantly.

  A weakness with Python is that it executes with the help of an interpreter instead of a compiler, which makes it slightly slower than for example C or C++. Python also has quite high memory consumption due to its flexibility in managing various data types.

- **R:** R is an open source language and software environment for statistical computing and graphics, supported by the R Foundation for Statistical Computing. This skill set has high demand across recruiters in machine learning and data science.

  R provides many statistical models, and numerous analysts have composed their applications in R. It's the favorite language for open statistical analysis, and there's a clear focus on statistical models that have been composed utilizing R. The public R package archive contains more than 8,000 contributed packages. In R, the unit of shareable code is the package. Microsoft, R Studio, and a number of other organizations give business support to R-based computing.

  One disadvantage of R is that it's harder to maintain when the code grows bigger. Another issue is that because R is extremely flexible, you may find yourself in many situations when you can do something well in a hundred ways. For maintainability and for working in teams, this may not be what you want.

  To support the usage of Python and R, you should also consider using Anaconda (applicable for both Python and R) or RStudio (only for R). Anaconda offers an easy way to perform Python and R machine learning on Linux, Windows, and Mac OS X and has over 11 million users worldwide, it is the industry standard for developing, testing, and training on a single machine, enabling individual data scientists to download Python and R data science

packages, manage libraries, dependencies, and environments. It offers support for developing and training machine learning and deep learning models with Scikit-learn, TensorFlow, and Theano as well as analyze and visualize data using other specialized open source software.

- **Java:** Java is a popular, general purpose language that runs on the Java Virtual Machine (JVM). Many organizations, particularly multinational corporations, use this language to create back-end systems and desktop/mobile/web applications. It is an Oracle-supported computing system that empowers portability between platforms.

  One big advantage of Java lies in its huge user base in enterprise software, which means there's quite a large community out there, with a lot of skilled developers available to you. It's been around for a long time, and most software engineers are packing Java skills. However, even Java has its drawbacks. For example, Java is comparatively slower and takes more memory space than the other native programming languages, like C and C++.

- **SQL:** SQL (Structured Query Language) is another popular programming language in the data science field that has been around for a while. It's great for querying and editing the information stored in a relational databases and has been used for decades for storing and retrieving data. It has proven to be especially useful for managing particularly large databases, reducing the turnaround time for online requests by its fast processing time. Having SQL skills can be an important asset for machine learning and data science professionals, as SQL is a preferred skill set for many organizations.

- **Julia:** Julia is a high-level dynamic programming language designed to address the needs of high-performance numerical analysis and scientific computing. As such, it is rapidly gaining popularity among data scientists. Because of its faster execution, Julia has become a perfect choice for dealing with complex projects containing high-volume data sets. For many basic benchmarks, it runs 30 times faster than Python and regularly runs somewhat faster than C code. If you like Python's syntax yet have to deal with a massive amount of data, Julia is the next programming language to learn.

- **Scala:** Scala (short for *sca*lable *la*nguage), is now the go-to language for functional programming. This general-purpose, open source programming language runs on the JVM. It's an ideal choice for those working with high -volume data sets and has full support for functional programming.

  Because it was developed to run on the JVM, it allows interoperability with Java itself, making Scala a great general-purpose language while also being a perfect option for data science. (It just so happens that the cluster computing framework Apache Spark is written in Scala, so if you want to juggle your data in a thousand-processor cluster and have a pile of legacy Java code, Scala is a good open source solution.)

REMEMBER

Most programming languages have drawbacks, and Scala is no exception. Scala is definitely hard to learn and therefore difficult to adopt. Moreover, it doesn't have much of community presence and is hampered by limited backward compatibility. If those minuses outweigh the pluses in your eyes, Scala isn't for you.

Unfolding Open Source Frameworks for AI/ML Models

A machine learning framework is an interface, library, or tool that allows developers to more easily and quickly build machine learning models without getting into the nitty-gritty of the underlying algorithms. It provides a clear, concise way of defining machine learning models using a collection of prebuilt, optimized components. Overall, an efficient machine learning framework reduces the complexity of machine learning, making it accessible to more developers.

Some of the key features of a good machine learning framework are that it

- Is optimized for runtime performance
- Is developer friendly and utilizes traditional ways of building models
- Is easy to understand and code on
- Provides parallelization to distribute the computation process, to make it faster

The dramatic rise of artificial intelligence in the past decade has spurred a huge demand for artificial intelligence and machine learning skills in today's job market. Machine-learning-based technology is now used in almost every industry, from finance to healthcare. In the following subsections I describe a selection of popular machine learning frameworks and libraries, pointing out their strengths and weaknesses when it comes to building machine learning models.

REMEMBER

You're going to be making some weighty decisions when it comes to choosing a framework, but don't forget that open source allows you to try them out first. Just keep in mind that not all machine learning frameworks are optimized for all types of machine learning techniques. Though some are good for natural language processing (NLP), others have been built to focus on deep learning (DL), and though some are more suitable for different types of hardware, others are tailored for the cloud. It's important to consider what your focus area is and, because these

frameworks are constantly evolving and also complement each other, allow for some freedom of choice for your users in the application layer.

TensorFlow

Developed by Google, TensorFlow is an open source software library built for deep learning or artificial neural networks. With TensorFlow, you can create neural networks and computation models using flow graphs. It is one of the most well-maintained and popular open source libraries available for deep learning. The TensorFlow framework is available in both C++ and Python formats. Other similar deep learning frameworks that are based on Python include Theano, Torch, Lasagne, Blocks, MXNet, PyTorch, and Caffe. You can use TensorBoard for easy visualization so that you can see the computation pipeline. Its flexible architecture allows you to deploy easily on different kinds of devices. On the negative side, TensorFlow doesn't have symbolic loops (symbolic-driven and dynamic programs for finding bugs) and doesn't support distributed learning, where machine learning algorithms are run in a distributed processing setup spread over several sites or target environments. Furthermore, it doesn't support Windows.

Theano

Theano is a Python library designed for deep learning. With the help of this tool, you can define and evaluate mathematical expressions, including multidimensional arrays. Optimized for GPU, the tool comes with a number of handy features, including integration with NumPy, dynamic C code generation, and symbolic differentiation. However, to get a higher-level, more intuitive view that make it easier to develop deep learning models regardless of the computational backend used, the tool will have to be used with other libraries such as Keras, Lasagne, and Blocks. The tool is great for cross-platform work because it's compatible with the Linux, Mac OS X, and Windows operating systems.

Torch

Torch is an easy-to-use open source computing framework for machine learning algorithms. The tool offers efficient GPU support, N-dimensional arrays, numeric optimization routines, linear algebra routines, and routines for indexing, slicing, and transposing. Based on a scripting language called Lua, the tool comes with an ample number of pretrained models. This flexible and efficient machine learning research tool supports a broad array of major platforms, including Linux, Android, Mac OS X, iOS, and Windows.

Caffe and Caffe2

Caffe is a popular deep learning tool designed for building apps, and it just so happens to have a good Matlab/C++/ Python interface. The tool allows you to quickly apply neural networks to the problem using text without writing code. The tool supports a variety of operating systems such as Ubuntu, Mac OS X, and Windows.


TIP

As new computation patterns have emerged — distributed computation, mobile computation, reduced precision computation, and more nonvision use cases, meaning when the use case has no image representation — the Caffe design has shown some limitations. The introduction of Caffe2 improves Caffe 1.0 in a number of ways, including first-class support for large-scale distributed training, mobile deployment, new hardware support (in addition to CPU and CUDA), and flexibility for future directions such as quantized computation.

The Microsoft Cognitive Toolkit (previously known as Microsoft CNTK)

Microsoft Cognitive Toolkit empowers developers to harness the intelligence within massive data sets through deep learning and by providing scaling, speed, and accuracy with commercial-grade quality and compatibility with a number of different programming languages and algorithms. It is one of the fastest deep learning frameworks with C#/C++/Python interface support. The open source framework comes with a powerful C++ API and is faster and more accurate than TensorFlow. The tool also supports distributed learning with built-in data readers providing a very efficient way to access data. It supports algorithms such as feed-forward, CNN, RNN, LSTM, and sequence-to-sequence. The tool's platform support is a tad limited because it works only with Windows and Linux.

Keras

Written in Python, Keras is an open source library designed to make the creation of new deep learning models easy. This high-level and intuitive neural network API makes it easier to develop deep learning models regardless of computational backend and can be run on top of deep learning frameworks like TensorFlow, Microsoft CNTK, and so on. Keras is known for its user-friendliness and modularity, making it the ideal tool for fast prototyping. The tool is optimized for both CPU and GPU.

Scikit-learn

Scikit-learn (formerly scikits.learn) is a free software machine learning library for the Python programming language. It features various classification, regression, and clustering algorithms, including support vector machines, random forests, gradient boosting, k-means, and DBSCAN, and it's designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy. The tool supports operating systems like Windows and Linux. On the downside, it isn't very efficient with GPU.

Spark MLlib

Apache Spark MLlib is a scalable machine learning library that includes clustering, dimensionality, regressing, collaborative filtering, decision trees, and higher-level pipeline APIs. It is a distributed machine learning framework that can be used in Java, Scala, Python, and R. Designed for processing large-scale data, it has been developed on top of Apache Spark Core and is widely used and focused on making machine learning easy. The tool interoperates with NumPy in Python and R libraries.

Azure ML Studio

Azure ML Studio is a modern cloud platform that data scientists can use to develop machine learning models in the cloud. With a wide range of modeling options and algorithms, Azure is good for building larger machine learning models. The service provides a variety of storage space per account and applies a "pay-as-you-go" model and can be used with R and Python programs.

Amazon Machine Learning

Amazon Machine Learning (Amazon ML) is a robust, cloud based service that makes it easy for developers of all skill levels to use machine learning technology. Amazon ML provides visualization tools and wizards that guide you through the process of creating machine learning models without having to learn complex machine learning algorithms and technology, including other data science services like frameworks and data security services.

Choosing Open Source or Not?

Obviously, you have endless choices of tools and frameworks in the open source space in data science, and because it's the open source space that drives data science evolution, does that mean that open source is the only way to do successful data science? No, of course not.

However, it must be understood that open source is a powerful factor in the data science space. It isn't something to disregard or marginalize. Although your data science investment might seem like a small fish in the ocean, it is still part of the ocean, meaning that it needs to function within the vast data science ecosystem. Few data science investments made today can be seen as isolated environments not dependent on anything external, like data, regulatory demands, vendors, customers, and so on. Considering that, and the fact that most standardization today is also driven from a de facto standardization approach in open source communities, you need to understand and closely monitor what is happening in the open source space, even if you choose to invest in a fully commercial "ready-to-go" solution.

IN THIS CHAPTER

- Understanding what it means to realize a data infrastructure
- Identifying the key infrastructure elements for success
- Utilizing automation to drive speed and efficiency
- Creating a collaborative workspace to increase data science productivity

Chapter 18
Realizing the Infrastructure

Data plays a key role in every use case of data science, although the type of data used can vary. For example, innovation can be fueled by having machine learning models find insights in the large amounts of data being generated by businesses. In fact, it's possible for a business to cultivate an entirely new way of thinking inside the organization, based on data science alone, if management pushes in that direction. The key is understanding the role that data plays at every step in the data science workflow and how the infrastructure must be designed and operated to maximize utilization of the data as well as enable high data science productivity. In this chapter, I help you focus on how all the pieces need to come together to realize a productive data infrastructure supporting your data science setup.

Approaching Infrastructure Realization

A *data infrastructure* is a decidedly *digital* infrastructure promoting data sharing and consumption. As is the case with other infrastructures, it is the structure that is needed for a system to function. How to realize and set up the data

infrastructure depends heavily on company objectives, the size of the company, its line of business, and other factors. But by using a reference model for your data infrastructure, you'll be able to get an overview of areas that need to be covered and what you need to consider in order to make strategic choices in each area.

REMEMBER

A *reference model* for a data infrastructure is an abstract framework for understanding significant entities and relationships between them. The purpose is to facilitate understanding of existing data infrastructures when comparing them in terms of functionality, services, and boundary conditions related to the scope of what is included or not.

Figure 18-1 shows a bird's-eye view of a data infrastructure framework, without the included components for each layer and area. Each area covered in this framework needs to be analyzed thoroughly and components selected carefully in relation to the overall company strategy and ambition, as well as industry context, line of business, and legacy infrastructure setup.

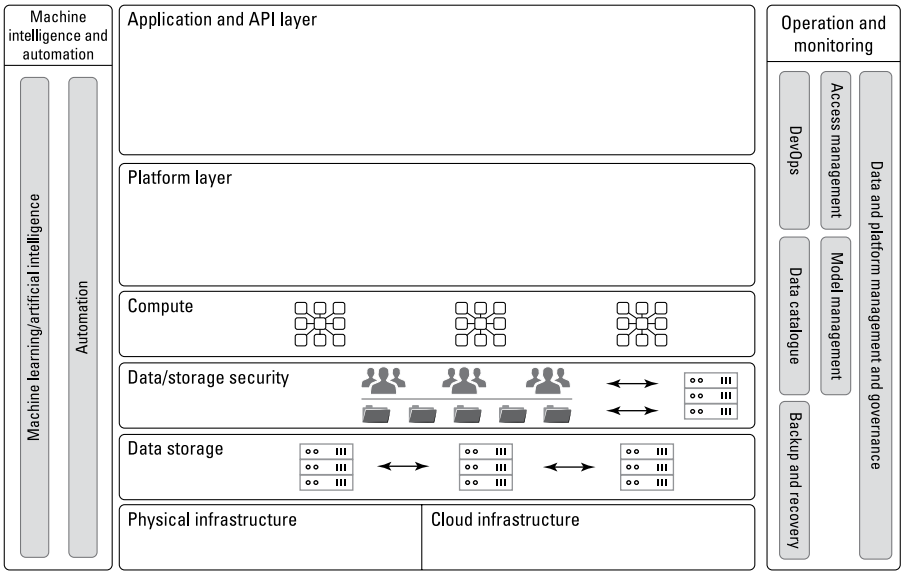

FIGURE 18-1: An example of a data infrastructure framework.

The different areas labeled in Figure 18-1 — physical infrastructure and data security, for example — represent the significant entities in a data infrastructure. The work on your data infrastructure realization should start from these important entities, where you need to determine how they'll serve the overall data science objective in terms of optimizing your data usage, and data science productivity. You know your data, customer, and market best, and by applying that knowledge and understanding of how the infrastructure needs to support your ways of

achieving your targets, you'll be able to figure out how to realize that infrastructure in terms of which components should be placed in which layer.

Turning back to Figure 18-1, here's a quick look (starting at the bottom) at the entities and layers you'll need to work with in developing your data infrastructure:

- **Physical infrastructure:** The hardware assets necessary for implementation, including the communication networks (cables) and data centers.
- **Cloud infrastructure:** Hardware and software components — servers, storage, a network, and virtualization software — needed to support the computing requirements of a cloud computing model.
- **Data storage:** Concerns archiving of data in various forms. However, different types of data storage play different roles in a computing environment. In addition to forms of hard data storage, there are remote data storage, such as cloud computing, significantly improving the ways that users access data.
- **Data storage security:** A specialty area of security concerned with securing data storage systems and ecosystems and the data that resides on these systems.
- **Compute:** Resources that are used to process data are called compute resources, and in cloud computing these are usually provided by central processing units (CPUs) working together in clusters. To enable fast and capable big data computation there are also other computational resources like accelerated processing units (APUs), and graphics processing units (GPUs).
- **Platform:** The environment in which a piece of software is executed. It may be the hardware or the operating system (OS) or even a web browser and associated application programming interfaces or other underlying software, as long as the program code is executed with it. The platform layer is the stage on which computer programs can run.
- **Applications and API's:** A layer that specifies the shared communications protocols and interface methods used by hosts in a communications network. It consists of protocols that focus on process-to-process communication across an IP network and provides a firm communication interface and end-user services. The application layer is used in both of the standard models of computer networking: the Internet Protocol Suite (TCP/IP) and the Open Systems Interconnection (OSI) model.

When realizing your data infrastructure, you'll also need to consider how automation and machine learning techniques will be used across different layers in the infrastructure, because using these techniques as part of the infrastructure processing activities is essential from a speed, cost, and data integrity perspective.

Another aspect is, of course, how you intend to operate the infrastructure and its data systems in practice. This includes aspects such as determining the principles for data management and governance, regulatory aspects, maintenance, and monitoring of the end-to-end infrastructure.

REMEMBER

It is one thing to define and set up your infrastructure, but in order to achieve a successful realization, you need to properly think through both operations and life cycle management of the end-to-end environment.

Listing Key Infrastructure Considerations for AI and ML Support

Historically, the infrastructure for artificial intelligence and machine learning projects have been set up and run by the data science teams themselves, but in larger companies these tasks are now slowly being transitioned to IT infrastructure professionals as these technologies start to move into the mainstream infrastructure. As this transition happens and artificial intelligence initiatives become more widespread, IT organizations need to start carefully considering what type of infrastructure best enables artificial intelligence productivity in this constantly evolving data science space.

TIP

Rather than purchase servers, network infrastructure, and other components for specific projects, the goal should be to think more broadly about the business's needs both today and tomorrow, similar to the way data centers are run today. How can the infrastructure be built in a way that it consists of both a stable foundation in the bottom layers for speed, cost efficiency, and reuse but with user flexibility and a self-service approach on top?

Location

Artificial intelligence and machine learning initiatives are not solely conducted in the cloud, nor are they handled only on premises. These initiatives should be executed in the location that makes the most sense, given the expected output. For example, a facial recognition system at an airport might be forced to conduct the analysis locally because of data privacy and security reasons, and in some cases perhaps depending on latency requirements (response time to and from the cloud). It's critical to ensure that the infrastructure can be deployed in various ways; in the cloud, in an on-premises data center, or at the edge (on a device) so that the performance of AI initiatives is optimized, depending on its requirements and context.

Location is also more than infrastructure location; it could also refer to geographical location. Whether your company is a global one or a local one will impact the geographical spread needed for your infrastructure setup.

One final location aspect to consider has to do with whether you're aiming for a data science environment for internal business efficiency gains or whether the infrastructure should also (or only) support a commercial business offering of data products and services. Location becomes a key component here if there is a data regulatory, system latency, or other need that forces you to be as close as possible to your customers.

Capacity

Artificial intelligence performance is highly dependent on the underlying infrastructure. For example, graphical processing units (GPUs) can accelerate deep learning by 100 times compared to traditional central processing units (CPUs). An underpowered server will cause delays in the process, whereas an overpowered server wastes money. Whether the strategy is end-to-end or best-of-breed, ensure that the computational hardware has the right mix of processing capabilities and high-speed storage. This requires choosing a vendor that has a broad portfolio that can address any phase in the artificial intelligence life cycle.

Data center setup

In terms of data centers, a data infrastructure doesn't live in isolation — it's always considered an extension of the current setup, by either expanding the number of data centers on hand or transforming parts of the current infrastructure to one driven by data science. Ideally, companies should look for a solution that can be managed with their existing tools (or at least as part of a configuration that complements what you already have) instead of scrapping everything and starting over. In some cases, however, this might not be possible, even as a stopgap measure. If your current infrastructure is hopelessly outdated, with costly and low-performing legacy tools that cannot be virtualized or containerized, it isn't worth the effort to keep using it. If your company's ambition is to transform itself into a company fully driven by data and artificial intelligence, one that intends to build a virtualized business on a fully cloud-based infrastructure, the current setup just has to go.

End-to-end management

There's no single "AI in a box" that can be dropped in and turned on to kick off the AI process. It's composed of several moving parts, including servers, storage, networks, and software, with multiple choices at each position. The best solution

is a holistic one that includes all (or at least most) of the components that could be managed through a single interface. Although this is complex, try to think "simplicity" in terms of how it needs to be managed.

REMEMBER

Utilize automation in as many aspects as possible in the management and operational aspects of your data environment.

Network infrastructure

When deploying artificial intelligence solutions, emphasize GPU-enabled servers, flash storage, and other computational infrastructure elements. This makes sense because artificial intelligence is heavily processor- and storage-intensive. Don't forget, however, that your data has to get to your storage systems and servers somehow, which means that you have to pay attention to your network capabilities. Think of infrastructure for artificial intelligence as a 3-legged stool, where one leg consists of the servers; another, of the storage system; and the third, the network. Each must be equally fast to keep up with each other and not cause an imbalance in the infrastructure, and it can never be stronger or faster than the weakest or slowest part. A lag in any one of these components can impair performance. The same level of due diligence given to servers and storage should be given to the network — checking link capacity for data transfer between point of collection to point of computing, for example. Remember that the network infrastructure setup could span over more than one country, at least for global companies or companies dependent on big data volumes from other countries.

Security and ethics

Artificial intelligence often involves extremely sensitive data, such as patient records, financial information, and personal data. Having this data breached could be disastrous for the organization. Also, the infusion of bad or biased data could cause the AI system to make incorrect inferences, leading to flawed decisions. The AI infrastructure must be secured from end to end with state-of-the-art technology, including both security aspects and control mechanisms for AI ethics. And although security aspects are seldom forgotten, ethical ones are. More details on control mechanisms for managing ethical aspects in AI models can be found in Chapter 16.

REMEMBER

As regulatory restrictions become stricter, they're starting to include ethical infrastructure principles — using representative and unbiased data, for example, or forming diverse and unbiased data science teams — that must be fulfilled in order to avoid situations where poorly handled data could cause AI systems to become discriminatory, intrusive, or even hazardous.

Advisory and supporting services

Although services like data science training and various consultancy assignments aren't technically considered part of the infrastructure, they need to be part of the infrastructure decision. Most organizations, particularly inexperienced ones, don't have the necessary skills in-house to make AI successful and productive. A services partner can deliver the necessary training, advisory, implementation, and optimization services across the data science life cycle and should be considered a core component of the deployment.

Ecosystem fit

Data ecosystems can be described as consisting of a number of actors that interact with each other to exchange, produce, and consume data. Such ecosystems provide various vital components for creating, managing, and sustaining data as part of a data infrastructure.

Some say that data science, especially with regards to machine learning and artificial intelligence, is at a maturity level comparable to where the software business was in the 70s or 80s. Therefore, staying ahead of the competition in terms of managing efficient machine intelligence ecosystems can become a major competitive advantage going forward.

REMEMBER

No single AI vendor can provide all technology everywhere. You must select vendors that provide broad ecosystem support and can bring together all, or a lot, of the components of AI to deliver a fully capable, end-to-end solution. Having to put together the components yourself usually leads to unnecessary delays and even failures. Choosing a vendor with a strong ecosystem might instead provide a fast path to success.

Automating Workflows in Your Data Infrastructure

In a data-driven organization, data drives every business process, but you can't fully accelerate and optimize its processes if you don't automate data management workloads every step of the way.

Manual touchpoints and workflows throughout the data management lifecycle impede many companies' ability to take in, aggregate, store, process, analyze, consume, and otherwise make the most of their data resources. Automating more

data pipeline processes can help your company execute transactions, make decisions, rethink strategies, and seize competitive opportunities better and faster.

More and more data professionals are adopting an approach focused on creating an automation-driven architecture in which repeatable tasks, such as data integration scripts and machine learning models, can be deployed rapidly into production environments.

However, automating large parts of the data pipeline demands integration of DevOps practices into the working lives of all functions and roles dependent on the pipeline. This includes roles such as data scientists, data engineers, business analysts, and data administrators. Even IT operations and other stakeholders might be impacted, depending on your company setup and whether the pipeline is set up and operated by IT or not.

Automation also requires that DevOps practices span across your entire infrastructure, including your diverse data centers, mainframes, and private clouds, as well as any externally sourced "as-a-service" offering that is being used by the business. Ideally, from an operational perspective, you should have a single visual interface with which to develop repeatable scripts, run scheduled jobs, develop nuanced rules and orchestrations, and otherwise automate the scheduling, consumption, and administration of resources throughout your distributed data environment to fully utilize the time and cost efficiency gains of automated workflows.

Enabling an Efficient Workspace for Data Engineers and Data Scientists

Over the past five years, I've heard many stories from data science teams about their successes and challenges when building, deploying, and monitoring models. Unfortunately, I've also heard about the misconception that data science should be treated just like software development.

This misconception is completely understandable. Yes, data science does involve code and data, yet people leverage data science to discover answers to previously unsolvable questions. As a result, data science work is more experimental, iterative, and exploratory than software development. Data science work involves computationally intensive algorithms that benefit from scalable computational resources and sometimes requires specialized hardware like GPUs. Data science work also requires data — a lot more data than typical software products require. All these needs, and more, highlight how data science work differs from software

development. These needs also highlight the vital importance of collaboration between data science and engineering for innovative, model driven companies seeking to maintain or grow their competitive advantage.

As the amount of data in an organization grows, so do the numbers of engineers, analysts, and data scientists needed to analyze this data. Today, IT teams constantly struggle to find a way to allocate big data infrastructure budgets among different users in order to optimize performance. Data users such as data scientists and analysts also spend enormous amounts of time tuning their big data infrastructure, which might not be their core expertise, or at least not what they are assigned to work on.

The importance of an efficient cross-team collaborative workspace shouldn't be underestimated. Such a collaborative workspace should be able to handle all analytical processes from end to end, across different systems and organizations, and ensure that productivity isn't lost along the way. As you might imagine, this isn't an easy task, but with a collaborative workspace approach for the entire data science team, you minimize the inevitable hand-offs between data engineers and data scientists, creating a seamless workflow from data capture to deployment of models into production.

The common, collaborative workspace should also have sufficient ecosystem support for the most popular languages and tools, which allows practitioners to use their preferred toolkit. A good cross-team collaborative workspace should also foster teamwork between data engineers and data scientists via interactive notebooks (which works as an interactive coding environment), APIs, or their favorite integrated development environments (IDEs), all backed with version control and change management support.

Practitioners must also be able to access all needed data in one place and automate the most complex data pipelines with job scheduling, monitoring, in predefined workflows. That access gives the data science teams full flexibility to run and maintain data pipelines at scale and at every part of the data science life cycle.

5 Data as a Business

IN THIS PART . . .

Exploring commercial opportunities in data science

Deciding on commercial approach

Using a data-driven approach to customer engagements

Working with data-driven business models

Considering new delivery models for data products and services

IN THIS CHAPTER

- Sorting out the concept of data monetization
- Explaining how to start treating data as an asset
- Elaborating on the data economy

Chapter 19
Investing in Data as a Business

The "data is an asset" idea is not new. However, despite the large number of people who have made the claim, there's still a huge difference between talking about making data an asset and actually *doing* it. And, when it comes to pushing the slightly more expanded message "data as a business," few are willing to take that ambitious step, even though (as I would argue) the possibilities are endless for those who dare to actually go on and do it.

So, yes, there's been a bit of a buzz out there about data for a while, but the interesting question right now is why an increasing amount of attention is being given to data as an asset and why does it seem as if some companies are discovering it for the first time? And, why do many businesses still undervalue data and information — or are unable to leverage it — although analysts, vendors, and others are repeating the message of the importance of data over and over again? This chapter tries to provide the answers to these questions.

Exploring How to Monetize Data

By now, almost everyone realizes that there's money to be made in data, but not everyone has a good grasp of *how* that money gets made. The following bullet list reveals the secrets by displaying examples of some of the different types of data monetization opportunities out there.

- **Digital advertising:** Right content, right audience, right time
- **Financial services:** Cross and upsell and detect fraud
- **Managing traffic:** Alleviate congestion and optimize delivery routes
- **Optimized billboard ads:** Understand the traffic and tailor the message
- **Public transportation:** Passenger satisfaction, operational efficiency, revenue opportunities
- **Retail:** Optimize store placements and staffing, monitor competitors
- **Entertainment and Events:** Manage traffic, target promotions
- **IoT (Internet of Things):** Add value through location data and more

REMEMBER

Data isn't only the main driver for connecting people — it is the lifeblood of so many company perspectives, like understanding and improving customer experience improvement and customer service. And it's not just about the data — it's also about getting access to it, whether the reason is to build new business models, do data-driven marketing, or simply gain access to the right data that enables better decision-making.

It isn't that companies fail to understand the importance of data, information, and actionable intelligence. (Well, some do.) It's mainly that many businesses often don't fully grasp how much of a business asset data really is or how the data is distinguished from the technology through which it flows. However, the emergence of a chief data officer (CDO) in many organizations and across industries indicates a growing recognition of data as a strategic business asset that stands on its own.

TIP

In most organizations with a CDO, that role will either participate in or effectively lead data monetization efforts as a way of demonstrating the CDO's own value. This leadership role can do much to enhance the growing influence of the CDO, but to truly realize true economic benefits of data, companies must champion a CDO's initiatives and start treating data as a real business asset.

Approaching data monetization is about treating data as an asset

Following the seven steps in this section will help your organization or company take a structured approach to monetizing data so that you can start treating data as an asset:

1. **Assign a data product manager.**

   Give your data the same product management attention that you give other valuable assets and competencies. Organizations typically have a defined approach for managing and marketing products. Likewise, if you're considering licensing data in any form, you need someone whose job is to define and develop the market for the data asset and to turn it into a real data product.

2. **Get to know your data and make an inventory of available data assets.**

   It is essential that you evaluate what data you have access to and what you need going forward. And, if you don't become fully data literate and truly understand the details of your data, you won't be able to leverage it as an asset. Make sure you identify all types of data — including operational, commercial, public, social media, and Internet content — that you can mine for new forms of value. Next, help business leaders understand the range of data available and use various data management and data mining techniques to refine the raw data into more consumable and communicable forms. (For more on such techniques, see Chapter 6.)

3. **Evaluate direct and indirect methods for monetizing data.**

   Indirect monetization requires using data internally to improve a process or product in a way that results in measurable outcomes, such as income growth or cost savings. Direct monetization involves a transaction of some sort, or incorporating data into a new data product or service. It can also mean actually selling the data itself (usually, by licensing it) in one form or another.

4. **Observe others.**

   Borrow ideas from other industries. Check out what other industries are doing in order to jump-start your own monetization efforts. It's becoming increasingly important to look beyond a certain industry, not just to find good ideas but also as an early warning sign about how other companies are evolving their information monetization initiatives that could intrude on your market in unexpected ways.

5. **Test ideas for feasibility.**

   Put ideas to the test by asking a series of feasibility questions regarding whether your ideas are practical, marketable, scalable, legal, ethical, economical, and so on.

6. **Prepare the data.**

   Think through how you're going to gather the data, and from which data sources. Then you need to work with the data to enhance its analytical and potential economic value. Again, think of how physical production processes use raw materials to eventually create finished goods.

7. **Decide on a marketing strategy.**

   Finally, for data products that you want to bring to a commercial market, you should focus on the marketing aspect. An important aspect to start with is to package the data product and determine how it will be positioned, priced, and sold. You also need to consider which terms and conditions will be applicable for the data product's specific usage. The appointed data product manager will play an important role in this process.


REMEMBER

Don't perceive monetizing data as an extraordinary task or one intended only for cutting-edge applications of digital business. Rather, you can view it as a core competency for every organization today — one, in other words, that has the potential to generate significant economic value from the varying range of data assets at each company's disposal.


TIP

ROI from data products will be achieved when data products are treated as real products and nothing else. An important step on this journey is to apply traditional product management methods to the data. You need to demystify the notion that a commercial data product is something different that needs to be treated with specific rules and methods.

Data monetization in a data economy

When treating your data products as any other product in your company's product portfolio, you need to continually remind yourself that data products are part of a global market phenomenon referred to as the *data economy.* Understanding the ecosystem of this global data economy is vital for you to succeed when it comes to introducing a new data product or service into that economy. Simply put, the data economy is an economy based on data, data technologies, and data products and services. It has its origin in the new global economy — the one based on the transition from a manufacturing based economy to a service and information based economy.


REMEMBER

The data economy is a digital ecosystem and a network of different players, such as data suppliers and data users. The term *data economy* refers to the ability of organizations and people to leverage data as an asset. Data is utilized to make strategic decisions, improve operational efficiencies, and drive sustainable growth, well-being, and innovations. The value and impact of data is increased by situational, contextual, historical, and time based factors.

WARNING

Integrating, refining, and sharing data increases its value and impact in a data economy. The effective use of data can lead to company growth, improvements in quality of life, and the creation of efficient societies. However, the effective use of data can be hindered by national or regional laws and regulations restricting data usage and the efficient exchange of data, which is explained in more detail in chapter 24.

Figure 19-1 shows the different technology areas enabled in the data economy and can be used to characterize companies and their roles, capabilities, and overall trends in how they currently act in the data economy as well as in the future market place. Companies grow their presence and potential value in the data economy in many ways and it is not necessary for a company to stay within one of these layers. Leading companies usually expand within a layer, or across multiple layers, or in the entire technology stack.

Architectural layers in the data economy	**Examples of data enabled technology areas**
Visualization provider	applications, user experience, user interface
Insight and execution generator	applications, methods/techniques, analytics and machine learning models and algorithms, open source, programming languages
Platform provider	sensors, devices, servers, cloud, lab, open source, networks
Aggregator	capture, transfer, transformation, quality, correlation, aggregation
Controller	API's, control, quality, anonymization, security
Enabler/Producer	sensors, applications, connectivity, cloud, social media, websites, open source, financial transactions, surveys, digitized hard copies, embedded chips, attached wearables, mobile phones

FIGURE 19-1: Different technology areas in the data economy.

WHERE MYDATA WANTS TO TAKE US

MyData is a human centered approach to personal data management that combines the industry need for data with digital human rights. The mission with the MyData movement mission is to empower individuals by improving their right to self-determination regarding their personal data. The human-centric paradigm is striving for a fair, sustainable, and prosperous digital society, where the sharing of personal data is based on trust. The aim is to build a balanced relationship between individuals and organizations. The 'MyData Movement' came off the ground in September 2016 as part of the MyData2016 conference in Helsinki in Finland.

The data economy can also be viewed in different ways in terms of how data is perceived in the perspective of the world's economy.

- **The big data economy:** Can be defined as algorithm-based analysis of large-scale digital data for the purpose of predicting, measuring, and governing data assets.
- **The human-driven data economy:** This is a fair and functioning data economy in which data is controlled and used fairly and ethically in a human-oriented manner. The human-driven data economy is linked to the MyData movement (see sidebar) and a human-centered approach to personal data management.
- **The personal data economy:** This is enabled by individuals focused on using the personal data that all people generate and provide directly or indirectly. Consumers of personal data become the suppliers and controllers, like when Facebook uses our personal data to provide similar topics of interest to us in the application. Similarly, Uber declared recently that algorithms will analyze the personal data in real time and will charge customers what the algorithm predicts that you're willing to pay, rather than a flat rate.
- **The algorithm economy:** This is where companies and individuals can buy, sell, trade, or donate individual algorithms or app components.

Looking to the Future of the Data Economy

The data-driven economy is increasing competitiveness, innovation, and business opportunities on a world-wide scale. Recent estimates report that rising global data flows have boosted world GDP by more than 10%. This can be compared to

figures for Europe only, where the new policy regulations, legislative conditions, and investments in ICT are expected to increase the value of the European data economy to 739 billion euros by 2020, representing 4 % of the overall EU GDP. Key sectors in the data economy either are already data-driven or are on the way to becoming so in areas such as manufacturing, agriculture, automotive, telecommunications, and smart living environments. Healthcare and pharma are also at the core of the data economy.

The world is also moving toward a fairer data economy that benefits everyone. Managing personal information in a responsible manner makes everyday life easier and adds to the well-being of the many. A unified procedure opens up opportunities for user-oriented innovations and business activities.

Individuals are now starting to have more control over (and transparency into) the data concerning themselves. Individuals can actively define the conditions under which their personal information is used. Those service providers who are worthy of customer trust can also gain access to significantly more extensive and varied data e-services.

WARNING

All is not sweetness and light on the data economy horizon — some true challenges are definitely coming down the pike. It won't be easy to constantly come up with ever new approaches to dealing with data breaches as hacker techniques adapt to new security measures. The values to be gained through hacking also increase as the data economy grows. Other challenging issues include determining compensation to victims of data product malfunctions (accidents with self-driving cars, for example) and coming up with sufficient incentives for enterprises so that they take the necessary steps to invest in data security. Add to that the uncertainties for companies about data regulatory burdens and litigation risks and you can see that much work awaits companies and societies as we go forward.

REMEMBER

The regulation of the data economy is closely linked with data privacy. The present approach is flexibility, finding a balance between protecting privacy, and allowing citizens to decide for themselves. The European Union's GDPR regulation is one cornerstone of this new regulatory framework. A new paradigm for data governance is needed, with data ethics as a central component in all regulatory reforms.

IN THIS CHAPTER

- Deciding what to concentrate your data science investment on
- Recognizing what drives internal business insights
- Investing in commercial opportunities enabled by data
- Finding a balance in your strategic objectives

Chapter 20
Using Data for Insights or Commercial Opportunities

If you're planning to use data science mainly to enable data-driven operations and fact-based decisions and to drive internal efficiencies, you must understand your main strategic areas of concern from a business optimizer perspective. If, on the other hand, your main ambition is to underpin your commercial offerings using data science, there are some other strategic considerations to be aware of as a market innovator or disruptor. This chapter explains the strategic aspects you need to consider depending on what your business objectives are.

Focusing Your Data Science Investment

If you and your company have just recently embarked on the data science and machine intelligence journey, it isn't an easy task to strategically decide how you want to focus your investments going forward. Yet choosing the right starting

point, is vital, knowing all along that such a decision matters tremendously! On top of that, the data science and machine intelligence areas are quite complex, extremely transformative, and continually evolving with new techniques and methodologies (including new technical solutions for faster and more efficient computation and analysis) coming seemingly every day.

Considering the level of investment needed in terms of money and the effort required to bring about change, how can you be sure that a choice you make today will still be valid a couple of years from now? The simple answer: You can't. That doesn't mean, however, that you should then wait until the data science field has stabilized and matured. If you do that, you can be sure that your competition has passed you by. Instead, focus your investment on the most flexible data architecture setup possible — one that will allow you to change direction if needed in terms of business focus and data scope. The idea here is that any setup you choose should allow you to swap out old applications and machine learning/artificial intelligence tools so that you can incorporate new ones.


REMEMBER

You should make one basic decision early on — whether to focus your data science efforts internally (on business effectiveness and efficiency) or externally (on commercial offerings). Just keep in mind that the main purpose of this (crucial) decision is to guide and focus your efforts rather than to decide the future direction of your business once and for all.


WARNING

If your strategy is to "go big" and spread your data science investment across both internal efficiencies and commercial opportunities from the start, that is of course an option, too. But be aware that focusing equally on both aspects as the first thing you do isn't an easy task — even when your company already has some basic understanding of data science. If your company is new to the data science area, I strongly recommend abandoning that approach: Start instead with an internal focus and then move outward. That will enable your company to leverage a stable data-driven foundation to underpin your commercial offering.

Determining the Drivers for Internal Business Insights

Data science and machine intelligence are having a fundamental impact on businesses and are rapidly becoming critical resources for market differentiation and sometimes even for company survival. It's all about ensuring that your company is focused on doing the right things (effectiveness) and doing them the right way (efficiency).

Recognizing data science categories for practical implementation

Although you might have the best intentions when it comes to your data science investment in terms of establishing a data-driven approach in your company, it isn't an easy task to manage a data-driven business on a day-to-day basis. Even if you have built up the best infrastructure support possible and have a great data science team driving your efforts, it is still hard work to get all aspects right from a practical perspective. There are so many aspects to consider and so many parts of the company that need to fundamentally change. At the same time, you're bedeviled by questions of not only where you should start but also how then to avoid getting lost in all the changes.

Well, being able to quickly categorize every potential impact into one of five categories and then being able to communicate the potential of each one is an efficient way to help leaders drive better results using data science. Luckily for you, I've done the spade work and can now describe these five categories for you:

- **Innovation:** This category is about fostering new thinking and identifying potential business and market disruptions based on data science.

  Data scientists hold the ability to frame complex business problems as machine learning or operations research problems, which is the key to finding better, more optimized solutions to old problems. They may even reveal new problems and approaches that were previously unknown.

- **Exploration:** This category refers to how to explore unknown transformative patterns in data, thus identifying unknown business potential by thinking outside the box.

  Data scientists should be encouraged to make data discovery expeditions, where there are no clear objectives other than to explore the data for previously undiscovered value. Being data driven is all about challenging old ways of addressing problems by getting rid of conscious or unconscious preconceptions of how things work and therefore must be handled. Data exploration enables you to let the data lead the way to new, more optimized solutions, based on facts.

- **Experimentation:** Let free experimentation and prototyping take place by challenging the status quo with radically new ideas and solutions, not just data insights. Experimentation usually happens in live settings, not in the lab.

  With the availability of an ever increasing amount of data and constantly changing customer needs and expectations, human decision-making is becoming increasingly inadequate. Data science, and especially machine learning, excels in solving the kind of highly complex, data-rich problems that overwhelm even the smartest person.

The list of business or government challenges that data science can tackle is potentially endless. Taking just one example, imagine finding the most optimized engine for recommending items that a customer might be interested in based on previous behavior, purchases, preferences, and profile.

How would such a recommendation engine work? With experimentation, you'd use different machine learning algorithms in parallel in a live setting to generate recommendations. All the algorithms are given the same purpose and objective, which is to maximize *conversion* — in other words, maximizing the likelihood of customers buying something on the site.

Different potential buyers are exposed to different algorithms, and after either a predefined period has passed or a certain number of results have been achieved to secure the necessary statistical significance, the outcome is analyzed and one of the competing algorithms in the experimentation is appointed the winner.

- **Improvement:** This category is all about continuously improving existing business processes and current portfolio offerings.

  Improvement is perhaps the most common application of data science since data scientists many times have established models for refining internal processes and methodologies related to the data their organization collects. Common examples are marketing firms using customer segmentation for marketing campaigns, retailers tweaking dynamic pricing models, and banks adjusting their financial risk models. In the product development dimension, improvement could include enhancing development and distribution efficiencies in terms of time and cost, but it could also mean enhancing a service offering with the support of machine intelligence capabilities and machine task automation.

- **Firefighting:** Firefighting regards how to identify the drivers of *reactive* behavior — the bad stuff that happens. (Clearly, you'd prefer to focus your efforts on predictive, proactive, and preventive behavior, but if you don't put out fires, your business may burn to the ground.)

  Let's face it: Firefighting is sometimes a must. When something has gone wrong in your system — business profitability is decreasing, for example, or a customer has an urgent complaint — you need to react and respond as quickly and efficiently as possible. Data scientists can not only find the best solution to the problem fast but also help identify why this problem occurred in the first place and try to prevent it from ever happening again, by implementing algorithms to predict and prevent it going forward.

Applying data-science-driven internal business insights

Categorizing challenges is one thing, but how do you determine which data science activities are the most important? And how do you then apply those activities to gain real business value out of your data science investment? Last but not least, what role does the data scientist play in all this? I'll answer the last question first by showing how data scientists can apply practical internal business values in your company. So, here are the values you want to promote as well as advice on how data scientists can promote them:

- **Empower management to make better decisions**. If approached correctly, an experienced data scientist is likely to be seen as a trusted advisor and strategic partner to the organization's upper management. A data scientist can communicate and demonstrate the value of the company's data to facilitate improved decision-making processes across the entire organization, not only as stand-alone activities or predefined dashboards for management but also by integrating a need for data and insights into the operational model in the company. The data scientist has the ability to set up the model in such a way that the data truly becomes the fuel for the entire company operations, underpinning every decision, action, and evaluation.
- **Explore opportunities and how to apply insights**. The role of the data scientist is also to examine and explore the organization's data, after which recommendations can be made prescribing certain actions that will help improve company performance, better customer engagement, and ultimately increase profitability.
- **Introduce employees to the usefulness of the data science environment.** Another responsibility of a data scientist is to ensure that employees are familiar with (and informed about) the organization's data science development and production environment for analyzing and identifying value. Data scientists prepare the organization for success by demonstrating an effective use of the system to extract insights and drive action. Once employees understand the capabilities of the data science environment, their focus can shift to addressing key business challenges.
- **Identify new opportunities**. A vital part of that role is to question the existing processes and assumptions for the purpose of developing additional methods as well as analytical models and algorithms to continuously and constantly improve the value that is derived from the organization's data.
- **Promote decision-making based on quantifiable, data-driven evidence.** With the arrival of data scientists, the ability to gather and analyze data from various channels has ruled out the need to take high stake risks. Data scientists can now create models using existing data that simulate a variety of potential

actions; in this way, an organization can determine which path enables the best possible business outcomes.

- **Test decisions.** At the end of the day, it's all about making certain decisions (and not others) and then implementing the changes. But it doesn't end there. It's crucial to know the impact of the decisions made and deployed in terms of how they actually affect the organization. Data scientists can help the organization identify the key metrics related to important changes and quantify their success.
- **Identify and refine the view of the customer.** Most companies have at least one source of customer data to work with, but if it isn't being used well, the data is pretty much worthless. One important aspect of data science is the ability to combine existing data that isn't necessarily useful on its own with other data points to generate insights an organization can use to learn more about its customers and other target audiences. A data scientist can help with the identification of the key groups with precision through an analysis of disparate sources of data. With this in-depth knowledge, organizations can tailor services and products to customer groups and help profit margins flourish.
- **Recruit the right talent.** Reading through résumés all day is a daily, repetitive task for a recruiter, but that is now starting to change due to the possibility of using data science for these types of tasks as well. By mining the vast amount of data that is already available —in-house résumés and applications, for example, or even sophisticated data-driven skill tests and games — can help your recruitment team make faster and more accurate hiring decisions.

Using Data for Commercial Opportunities

When you want to make a strategic business move, you need to have proper reasoning and motivation behind your actions. Also, if you really want to seize opportunities, you can't afford to wait months for regular business evaluations. Data science gives business owners a way to make decisions quickly and efficiently while avoiding risks.

REMEMBER

To be clear, using data for identifying and seizing new commercial opportunities has little to do with making better business decisions regarding your company's current endeavors. Instead, it refers to how you can use data to identify, scope, and invest in new commercial business initiatives based on data and data-related products. In other words, it means investing in completely new products and services that can either complement your current line of business or disrupt your existing business model altogether. It really depends on the extent of your ambition and on which opportunities you find and are willing to invest in.

Defining a data product

Here's what seems like an easy question: What is a product based on data? A so-called data product, in other words? It turns out that the question isn't an easy one to answer. As in many areas of data science, there's no clear definition of *data product*, though — if forced to come up with a working definition — I'd say it's a product that facilitates an end goal through the use of data.

Perhaps this definition doesn't actually help your understanding of the concept because it might appear quite broad at first glance — referring to almost anything. Though it's definitely true that many different types of data products are available, you can nevertheless divide them into only two major categories:

- **Data *enabled:*** A functionally oriented product that needs data to fulfill its goal — it runs on data input to fulfill its functional objective, in other words. One example here is when data is used to derive a predictive insight about an automated system action that's needed in order to prevent a problem from occurring (automated decision-making). Without feeding the system with data, the system cannot analyze and act in advance to prevent the problem.
- ***Pure* data:** This type of product consists of data and has a data centered purpose (not a functionality-oriented one). In other words, it generates an insight as the end result, rather than any ability to perform a functional task. A pure data product can also refer to a situation where what you're selling is either the raw data or processed data itself or other related insights derived from data compiled into a report or set of recommendations.


REMEMBER

It pays to reflect a bit on the differences in definition between data products and other technology products. The different types are generally defined by diverse characteristics and should therefore be addressed differently from a strategic perspective. Though many of the standard product development rules applyto both technology products and data products — solving a customer need, learning from feedback, or prioritizing requirements, for example — there are still plenty of areas where the two types of products differ significantly.

This list uses popular products to spell out some of these differences:

- **Gmail**: Is Gmail a data product? No, Gmail is an email service with the primary objective of enabling written, digital communication between individuals. However, Gmail's sorting of our emails as Important or Not Important is a data product because the primary objective is to sort emails based on their data content and relevance, not on their functionality.

- **Google Analytics**: Is Google Analytics a data product? Yes, it is. The main objective is to explain a quantitative understanding of online behavior to the user. Here, data is central to the interaction with the user and, unlike the other technology products focused on a functional outcome, its key objective is to derive insights from that data.
- **Instagram**: Is Instagram a data product? No, but if you divide Instagram into different products, some parts (data tagging, search, and discovery, for example) can be considered data products.

Distinguishing between categories of data products

I talk about the distinction between data enabled products and pure data products earlier in this chapter, but there are other, more granular ways of dividing up the data product pie. One way is to sort the products into five main categories: raw data, processed data, algorithms/insights, decision support, and automated decision-making. Figure 20-1 gives a graphical representation of the range, and the following list describes some of the details for each product type:

- **Raw data:** This term refers to collecting data and making it available just as it is (or perhaps with some small processing or cleansing steps). The user can then choose to use the data as appropriate, though most of the work will be done on the customer or user side.
- **Processed data:** Processed data is one step up from raw data, meaning that some sort of data cleaning and transformation has occurred in converting the raw data into a format that can then be analyzed and visualized to provide insights to the user of the data product or the intended customer. In the case of customer data, additional attributes can be added for additional value — assigning a customer segment to each customer, for example, or calculating the likelihood of a customer clicking on an ad or buying a product from a certain category.
- **Algorithms/insights:** Data products related to models and algorithms or algorithms-as-a-service are the newest types of digital data product offerings. Here, an algorithm acts on some data — sometimes in a machine learning context, sometimes not — and the result is new information or insights. An example is the algorithm used in Google Images. When the user uploads a picture, it receives a set of images that are the same or similar to the one that's uploaded. Behind the scenes, the product extracts an image's salient

features, classifies it, and matches it to stored images, returning the ones that are most similar. Insights are added to this same category because the data product can sometimes be the insight itself, not the algorithm that generated the insights. A typical buyer of an insight (via a report or an insight-as-a-service model) is nontechnical. A typical insight related to Google Images could be derived through using machine learning to compare hundreds of images of your product to detect certain patterns of customer preferences when using your products, for example. The insight into how customers prefer to use your product could then be used in future marketing campaigns or as input to model future product enhancements.

- **Decision support system:** This category provides information to the user in order to support decision-making, though the final decisions are still made by the user. Analytics products such as Google Analytics, Flurry, and SAS Visual Analytics are examples that fall into this category. A lot of effort is needed in order to create a decision support system, and it's expected to do most of the job with the intention to give the user relevant information in an easy-to-digest format — dashboards to allow users to take better decisions, for example. When using these analytics tools, the insights gained could lead to changes in the editorial strategy, plans for addressing leaks in the conversion funnel, or a doubling down on a given product strategy. The important thing to remember with this type of data product is that, although the product has collected the data, compiled the data, and displayed the data, the user is still expected to interpret the data. Users are in control of the decision to act (or not act) on that data.
- **Intelligent and automated decision support:** In this type of data product, all intelligence within a given domain is included, meaning that the product can stand on its own and both provide and act upon insights within the specific data product. One example is Netflix product recommendations, where data from previous user preferences on series and films are used by the algorithm to recommend new selections. Since the result of the decision on what to view is captured in the same environment (the Netflix application) the model can capture each choice and learn to give even more precise and intelligent recommendations in the future. Other examples of more complex data products in this category would include closed loop automation with examples such as self-driving cars and automated drones. This type of data product allows the algorithm to do the job, learn from data and actions, and then present the user with the final output. The outcome sometimes comes with an explanation of why the AI chose that option; at other times, the decision-making process is completely hidden.

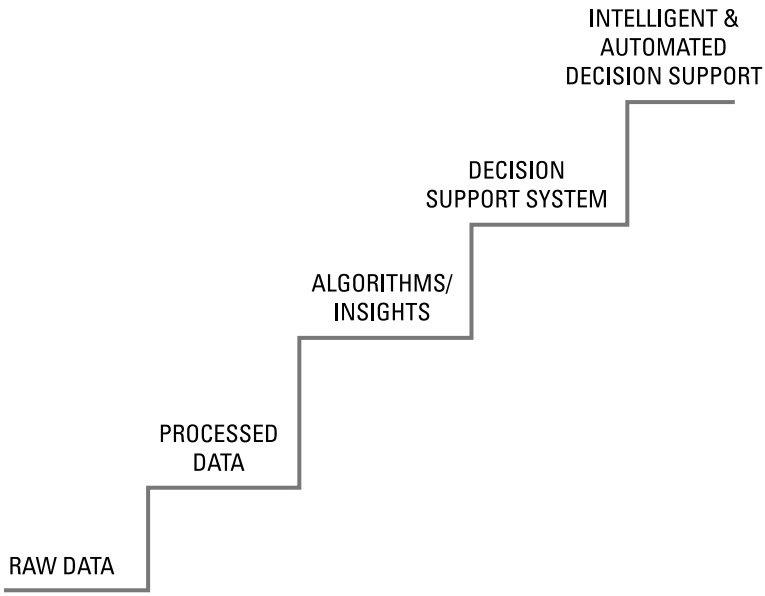

FIGURE 20-1: Different categories of data products.

REMEMBER

A main difference between these categories is the built-in level of complexity. More specifically, the categories in Figure 20-1 are classified in terms of their increasing internal complexity and (should have) less complexity on the user side. For example, though raw data has little built-in complexity to start with, it requires complex techniques and skills to develop a product that generates value out of the raw data. On the other hand, with a data product that is built on a complex machine learning algorithm, you get a simple user interface for the customer with less thinking required. The data product manages the complexity internally through the machine learning algorithm.

REMEMBER

Typically, (but not exclusively), raw data, processed data, and algorithms are focused on technical users. Insights, decision support, and automated decision-making products tend to have a more balanced mix of technical and nontechnical users.

Balancing Strategic Objectives

Suppose that you own a shoe company. Wouldn't it be useful if you found out that a segment of your target market prefers to buy running shoes in the month of June? Wouldn't it be more useful if you found out that the segment belongs to the 16–21 age group, that they prefer road running shoes to trail running shoes, that they can afford to spend $100 on a pair of shoes, and that they love the colors blue and red?

You don't always need to sell data to make money from it. Thanks to the ever increasing popularity of the Internet, affixing a small chip to every product (for example, shoes) and tracking the usage and other details is now simpler than ever. This doesn't mean that your company needs to listen to your customers' private conversations or keep track of everything your customers are up to. You only need to track the data that you think is beneficial for your business. That data may, however, range from nonsensitive to sensitive data (product usage, interests and preferences, or Internet activities, for example) to even your customers' friends' interests and SMS and call logs. Therefore, you must have a good strategy for handling data ethically and in a legally correct manner.

Hence, the applicability of data as input to or output from your business is an important strategic decision that needs some consideration. Spending time to understand the current trends in the market for your line of business is important, but so is making sure that your ambitions are feasible. If your line of business is more traditional and isn't yet digitalized and is remains far from being data driven, perhaps the best way to start is with an internal focus on turning things around, before you try disrupting the market with new data products.

IN THIS CHAPTER

- Getting to know your customers better
- Ensuring satisfied customers
- Improving efficiency in customer services

Chapter 21
Engaging Differently with Your Customers

Because humans everywhere now live in the age of the customer, it's time to clarify what the term customer experience management (CEM) really means. It may help, however, to first see what it is not. CEM is not about collecting feedback, responding to feedback, or tracking that tried-and-true metric of customer loyalty, your Net Promoter Score. None of these actions individually represents CEM. Instead, you could say that CEM refers to the complete philosophy and methodology that makes your business delightful to work with for your customers. In this chapter, I want to show you how an effective data strategy can guide you to a better and more insightful approach to your customers.

Understanding Your Customers

Optimizing the customer experience is a great way to attract new customers, but it's also one of the best ways to foster customer loyalty to retain the ones you already have.

Despite this benefit, marketers and other organizational leaders alike often neglect the customer before and after the sale. The biggest barrier to even beginning to

turn around this counterproductive practice is usually the lack of a deep understanding of the customer in the first place.

REMEMBER

Having a comprehensive understanding of your customers is the key to achieving core business goals, whether you're trying to build (or optimize) the customer experience, create more engaging content, or increase sales.

To see how you can come to a better understanding of your customers, I recommend that you look at some key activities you need to perform in order to really get to know your customers. The next few sections show you the way.

Step 1: Engage your customers

An optimized customer experience is, of course, valuable for revenue and retention, but if you get it right, it can also be a great source of customer insight. Engaging with your customers in real-time has become more easily accessible, thanks to a variety of new tools. Messenger is becoming an ever more popular customer service channel, and tools like Drift allow you to talk with your customers as they browse your website. Drift is especially exciting because it acts as an entirely new way to approach the customer experience in real-time in a conversational format compared to traditional customer engagement. So, speed here is definitely a plus, but there are a number of other pluses, as Figure 21-1 makes clear.

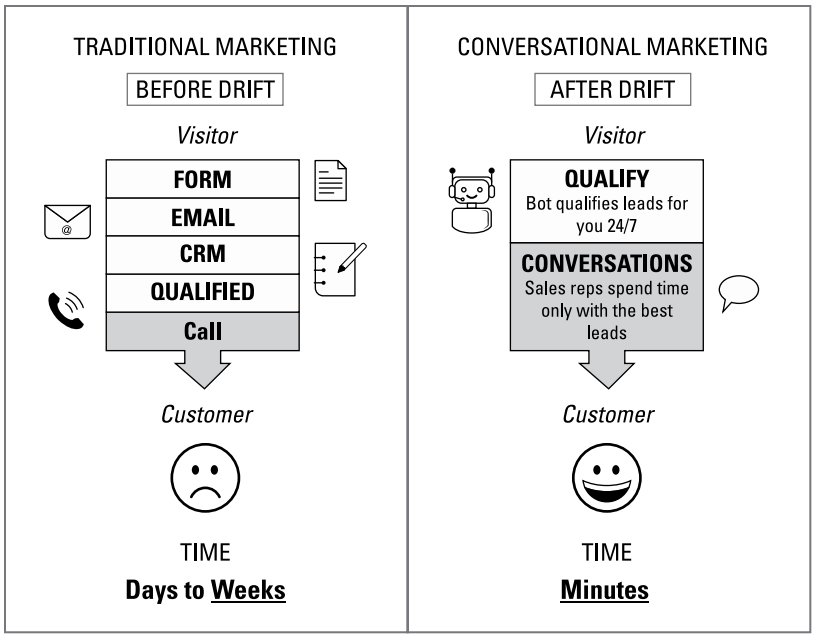

FIGURE 21-1: The old-versus-new ways of performing customer marketing.

Though channels like Messenger and Drift are clearly great ways to collect customer insight, they're not often used in the most effective manner. If your engagement is ad hoc or piecemeal, you're not putting the true power of these channels to use; such engagement needs to be part of a bigger plan. That means companies and organizations must have the foresight to invest the time and money it takes to understand the entire customer journey. It's not enough to take a single point along the way to survey and understand your customer; without a broader context, these spot-checks could be worse than useless. You wouldn't be able to answer basic questions, such as how did the customer get to this point or what were they looking for or where are they heading in the overall journey, because you don't have the information you need in order to come up with an informed answer.

If you invest the time needed in understanding the entire customer journey, you'll be able to flesh out how your customers experience your brand over the course of their relationship with that brand. That context will let you ask your customers the right questions at the right time, thereby building brand engagement as well as the kind of customer trust needed to help guide the journey to the point of purchase.


TIP

As you work to keep your customers engaged during the first stages of the customer journey, think of your relationship as a 2-way street. Encourage customers to share their thoughts and opinions by including a customer satisfaction survey on a regular basis in your ordinary emails.

Follow these three principles when designing a survey:

- **Remove bias.** Ask customers for their opinions without projecting your own. Get their uninfluenced, impartial opinions. You want genuine insights, even if they're negative. An example might be something as simple as this: "What do you think we could do better?"
- **Be concrete.** Use simple language that asks for feedback on a specific topic. For example, the question "How have you improved your marketing effectiveness using our machine learning algorithm?" will help to determine the value your customers are getting from you.
- **Focus.** Your surveys should address just one area of the customer experience. The aim is to get insights that you can then act on.

Keep these principles in mind as you personalize your customer survey with questions relating to your brand and product or service.

Step 2: Identify what drives your customers

Many marketers make the mistake of using generic demographics — like age, profession, and location — to develop a sense of the range of their customer base.

These data points simply don't provide enough information to create messaging that resonates with your customers on an emotional level.

One way to dig deeper into customer preferences is to use the Acquisitions tab on Google Analytics to see which social media outlets, industry blogs, and professional forums your site traffic comes from. Then apply this information to your identities so that you can find out where and when to reach them more effectively.


TIP

Acquiring keyword data is another helpful way to discover the terms and descriptions that certain buyer identities use to describe your services. To segment customers based on keyword searches, for example, first use Google Webmaster Tools to create a list of common keywords that drive people to your site, group the keywords into overarching themes, and then assign them to different customer categories based on the data you have available. To put this effort into action, incorporate these keywords across your website and then map content marketing efforts and other online interactions towards these new customer categories based on what attracts different type of buyers. Being attentive to customer preferences and speaking the same language as your customers is a subtle way to make your current audience feel more welcomed.

Step 3: Apply analytics and machine learning to customer actions

From clicking on a link to reading a web page, every customer action offers valuable insight into customer behavior. To determine how customers interact with your website, you can try user behavior tracking tools such as Google Analytics and Inspectlet. They're great tools for gathering insights such as time-on-page and bounce rate. Inspectlet can even provide short videos of users on your page in real-time.

The behavioral data you collect should lead you to conclusions about what your audience doesn't understand, what they like and don't like, and how you can create a stronger website experience. If people had trouble navigating to a certain sales page, for example, you can adjust the interface to allow for a more user-friendly experience.


TIP

If people spend more time on one page than others, analyze that page's content to see what may be needing extra attention. For example, if people are spending too much time on the checkout page, perhaps it's time to improve the customer payment experience on your site. Most importantly, though, if you have a page with a high bounce rate, try to see what is making people leave.

Recommender systems first became popular in the retail industry, mainly in online retail or e-commerce for personalized product recommendations. One most common usage is for Amazon's section on "Customer who bought this item also bought . . .". Recommender systems could be seen as an intelligent and sophisticated salesman who knows the customer's taste and style and can thus make more intelligent decisions about what recommendations would benefit the customer. Though it started off in e-commerce, it is now gaining popularity in other sectors, especially in media. Some of the examples are YouTube "recommended videos" or Netflix "other movies you may enjoy". Other industries are now also realizing the value of using recommender systems. (The transportation industry is one example.)

Step 4: Predict and prepare for the next step

Creating a plan for future customer engagement is just as important as creating a plan for the present. This puts customer experience teams in the right frame of mind to respond to customers during stressful or challenging situations.

Predictive modeling software mines existing customer data to identify cyclical patterns and trends that can inform decision-making. Two great tools for these tasks are the custom analytics programs from RapidMiner and Angoss, both of which create realistic future models. To see how predictive modeling informs customer strategy, imagine that you work for a SaaS company that wants to adjust its product road map to anticipate customer needs. By looking at the historical behavioral data, you can see which features customers have found most valuable over time and which features they didn't use. Understanding your most popular and most visited pages can also influence your content strategy, focusing on topics and formats that will best solve your audience's challenges.


TIP

Identify similarities across the most commonly used features to determine why your customers liked them. Additionally, looking at market trends and analysis gives you a good idea of what other companies in your space have already accomplished so that you can devise new features that explore these areas.


REMEMBER

Many companies turn to market research firms only as a form of insurance — a means of reducing business risk related to investing in the wrong product or service, in other words. But market research can be used in product development not only as insurance but also as a tool to establish market needs and to obtain a better understanding of market potential. Continuous market research throughout the product road map naturally leads to more sales. The more you understand your market, the better product/market fit you have.

Step 5: Imagine your customer's future

The only way to understand the unique and dynamic customer buying journey is to put yourself in your customer's shoes. This is made possible by an advanced technique called *customer journey mapping*, a method where companies create a detailed, graphical representation of the customer journey based on critical touch points. These touch points are interactions between a customer and your brand before, during, or after purchase.

Modern-day customer interactions encompass a wide range of touch points: mobile, web, social, interactive voice response (IVR), in-store, chatbots, and more. This brings me to the concept of *omnichannel*, a cross-channel content strategy that organizations use to improve their customer experience. Customers today switch between many different channels frequently — say, in the middle of a purchase or even during discovery. Analyses of the trends in customer behavior across industries show that omnichannel is only going to get bigger with the growth and variety of channels; therefore, customer journey mapping has to include every touch point and channel where your customers have a presence.

Figure 21-2 illustrates the broad variety of possible touch points you can use to reach your customers as well as how these points vary depending on whether it's pre-purchase, during purchase, or post-purchase. Capturing data from all three phases enhances your overall understanding of what is driving your customer to buy — and, hopefully, buy again.

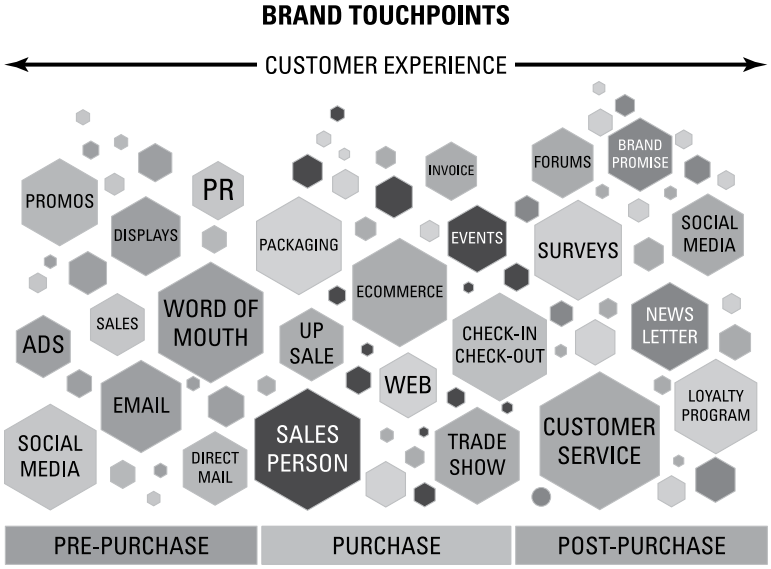

FIGURE 21-2: A word cloud for brand touch points.

REMEMBER

The first condition to being omnichannel is that you need to be *multichannel:* You need to be available wherever your customers are. However, having these channels up and running is one thing; making them work together seamlessly as part of the overall journey is another.

Check out Uber as an example of how to define touch points and how to apply them to customer journey mapping. Minor touch points include activities like downloading the app or just following the app on social media. Major touch points include things like actually requesting a ride or completing driver training.

Once you define the touchpoints, you need to explore the circumstances affecting each of them. For example, a marketer at Uber might ask, "What influenced the rider to download the app for the first time? Was it related to Uber's customer referral program?" The internal team should be engaged with these issues to get a well-rounded perspective and promote collaborative problem solving.

When failed touch points are identified (when a customer fails to use the downloaded Uber app, for example), you need to establish a plan for contacting these customers. Creating milestones might be a good idea, such as when an app user hasn't logged in to the account in three months or when an avid customer suddenly stops using the product. It's best if the customer experience team can call, write, or meet with customers directly to understand why they're disengaged. If these resources aren't available, you could create an email marketing message specifically focused on reengaging your customers based on certain milestones.

Keeping Your Customers Happy

The first thing you need to address when aiming to keep your customers happy by improving customer satisfaction is to get a better understanding of what the current customer attrition rate actually is. In that way, you can focus your initial efforts on the group of customers most likely to leave or on their way to leave already.

TIP

To reduce customer attrition (also referred to as *churn*), you should use historical customer data to map snapshots of customers taken at a given point in time. Such snapshots would, for example, record who they are, what they bought, and how they interacted with the products and/or services sold to them. You should map that information to whether they later churned (ceased being a customer, in other words). You could then study each current customer you've determined is likely to churn and then rate how valuable that customer is. Finally, you could determine what action needs to be taken in order to prevent the most valuable customers from churning.

Here are some actions you can take to prevent your customers from churning:

- **Enhance marketing campaigns by cross-selling products.** Make sure your marketing campaigns are really using the right message to target the right group of customers. There is nothing as annoying as being the subject of a marketing campaign that offers you something you have absolutely no interest in. It really gets you wondering whether the company knows its customers at all. Avoiding this embarrassment is best addressed by the cross-selling of products. Start by mapping customer/product pairs to purchase indicators as recorded in historical data. By doing so, you'll know whom to target when launching a new product or when promoting an existing one.
- **Optimize products and pricing.** Even if you're able to offer the right product or service to the right customer, it's also important to offer it at the right price. To find out which price is right for a certain product, you need to map product characterizations and price to numbers of sales. Then you can change the price and other characteristics to see how they impact revenue (price × number of sales). This gives you a good understanding of the most optimized price level.
- **Increase customer engagement.** Finally, it's important to know more about how you can increase your engagement with the customer. What is the customer really interested in? You can learn more about this by observing customer behavior when customers are presented with different products or services. This is needed in order to map customer/item pairs to indicators of customer interest. This enables you to predict needs and interests and take them into account when evaluating the service provided to the customer.

So, how do you know when a customer is served well? Simply put, serving customers well implies that you need to either offer your customers products that they are interested in and can afford or provide services they engage with.


TIP

To keep your customers happy and fully satisfied, it's vital that each and every part of your company collaborates with one another. Keeping your customers happy depends on not only the quality of your customer service crew but also other departments (those responsible for production, for example.) Only when all the gears in your company are well oiled and tightly connected can you expect the best results.


REMEMBER

Even if you implement all measures you can think of to make your company customer-centric, you can always do better. The same rule applies to customer service. Setting up some business goals, linked with customer service, as well as key performance indicators (KPIs), will help you stay on track with all your efforts to make customer satisfaction grow.

WARNING

If your customer comes with an issue in a communication channel that is less preferred — Facebook, for example — don't force the customer to use your chosen channel in its place. If the customer reached out to you on social media, it was because it was the most comfortable way for her to communicate. Instead, you should offer the customer some different choices, not only the one solution. Also remember to keep your customer informed about when the problem could be solved, instead of keeping the customer waiting. It's all about offering your customers the same services that you would demand from others.

Serving Customers More Efficiently

Serving customers more efficiently mainly refers to improving operations in order to reduce costs. But of course it also means serving customers well — as in anticipating issues before they occur or improving the handling of customer support requests when they arise.

TIP

Data science can increase efficiency through the use of supervised machine learning techniques. The idea is to map situations to outcomes so that you can predict outcomes in new situations. One example is when a customer has been exposed to a new product: The outcome here is defined by whether the customer will show interest in the new product. In such situations, machine learning techniques need examples to work with and train on. That means you need data on situations (characterized as finely as possible, along with any contextual information) and outcomes observed in these situations. An analysis of the sample data allows you to first find patterns and then the relationships between situations and outcomes. Predictions on outcomes are made automatically by using these relationships.

The following activities (presented in their business contexts) can help you improve efficiency in terms of your customer management through a predictive and preventive approach.

Predicting demand

Predicting demand is important to businesses that observe high variability in demand for their services and/or products — businesses that sell fresh goods and need to avoid having too much or not enough in stock, for instance. This enables benefits in terms of the ability to measure demand and the context in which it happened so you can map context to demand. You can also use the insights gained to determine how much staff to hire in anticipation of how busy the business is going to be.

Automating tasks

You can save time by having machines perform certain repetitive or intelligent tasks automatically. Sometimes you might already be performing these tasks with hand-crafted rules, but when you introduce machine learning abilities to run the activity instead, you can tap into a new potential of optimizing how the activity is done. Using machine learning to automate how the task is done means that the machine automatically learns rules from sample data, and, over time, it might optimize how the task is done, depending on which techniques you're using. In this context, one obvious example is the scoring of credit applications or insurance claims, where you're expected to either approve or reject something. Another example is automatically performing risk analyses from historical data using machine learning. This one gives you a better base for making decisions before investing time and money into new projects.

Making company applications predictive

You have much to gain by making applications used by employees related to customer relationship management, enterprise resource planning, human resources predictive. By adding predictiveness capabilities to these applications, people can do their jobs more efficiently. Here are some of the benefits of using predictive applications:

- **You can prioritize things.** Predictive company applications enable you to direct user focus to what is most important. This can be email (such as Google Priority Inbox), customer support requests (so that you can reply more quickly to the most important ones), or other external requests toward your company that you need to reply to urgently, for example.
- **You can better adapt the workflows.** A proactive approach enables you to use adaptive workflows based on predictions rather than on predefined manual rules. You could, for example, route customer support requests to those best equipped to handle them, in which case the outcome is a customer support team or person.
- **You can adjust the user interface.** You can easily increase user efficiency by adapting the interface to show just what users need at the time they use the app. All you need to do is map a context to an action that will need to be performed to trigger the adjustment.
- **You can automate user settings.** Predictive applications let you set configurations and preferences automatically by analyzing application usage data, and thereby speeding up user efficiency.

IN THIS CHAPTER

- Describing what a business model is
- Sorting out the fundamentals of a data-driven business model
- Using a framework for data-driven business models

Chapter 22
Introducing Data-driven Business Models

The increased use of data is transforming the way companies do business. With advanced analytics, machine learning, and access to new data sources, companies in one sector can play a role in the products and services of others — even those far removed from their traditional line of business. This blurs the boundaries between industries and changes competitive dynamics. Companies that embrace the full range of opportunities and transform their business models in parallel with these shifts will find new opportunities for revenue streams, customers, products, and services. In this chapter, I describe how you can approach the area of data-driven business models.

Defining Business Models

First you need a working definition of a business model. It's generally described as the foundation of how an organization creates, delivers, and captures value in economic, social, cultural, or other contexts. The process of business model construction and modification, also called *business model innovation,* forms a part of ordinary business strategy development.

The fact is, though, that the term *business model* is used for a broad range of informal and formal descriptions to explain core aspects of a business, including purpose, business process, target customers, offerings, strategies, infrastructure, organizational structures, sourcing, trading practices, and operational processes and policies including the company culture.


TIP

Given the wide use of the term "business model," I recommend defining it as broadly as possible. For me, that means defining business models simply as the design of organizational structures to endorse a commercial opportunity. Business models are used to describe and classify businesses, especially in an entrepreneurial setting, but they're also used by managers inside companies to explore possibilities for future development.


REMEMBER

Today, the type of business model that's needed for a certain company might actually depend on how the underlying technology is used. For example, entrepreneurs on the Internet have also created entirely new models that depend completely on existing or emergent technology. Using technology, businesses can reach a large number of customers with minimal costs. In addition, the rise of outsourcing and globalization has meant that business models must also account for strategic sourcing, complex supply chains, and moves to collaborative, relational contracting structures.

As you'd expect, business model design generally refers to activities that focus on designing a company's business model. It's part of the business development and business strategy process and involves design methods. However, there's a big difference between defining an entirely new business model when none is in place and changing an existing business model.


REMEMBER

In the case of designing a new business model, a common challenge is usually to understand and allocate needed resources in time. When changing an existing model to a new business model, however, the challenge is rather to manage resistance or lack of interest from employees as well as adapt organizational and product structures to new ways of developing and selling. And, depending on the size and distribution of employees, this can be a challenging task.

Technology-centric communities sometimes have specific frameworks for business modeling that attempt to define what can often be a difficult approach to defining business value streams. (At some point, tech start-ups have to start making money, right?) Business model frameworks represent the core aspect of any company, striving to represent the total picture of how a company selects its customers, but they also include how a company defines and differentiates its offerings, defines the tasks it will perform itself and those it will outsource, configures its resources, goes to market and creates usefulness for customers, and captures profits.

There's one final prism to use when looking at a business modeling framework: Is the focus on internal factors, such as market analysis, product/services promotion, development of trust, social influence, and knowledge sharing, or is the concentration more on external factors, like competitors and technological aspects?

It would seem as if we're asking business modeling frameworks to do too much, and it's true that the scope can be very broad — at times too broad. And yet, when used correctly, business modeling frameworks can be incredibly useful tools. In the context of data science, however, new frameworks for business modelling have emerged — data-driven business model frameworks. These will be explained and exemplified in more detail later in this chapter, but first I want to explain what a data-driven business model actually is.

Exploring Data-driven Business Models

The increased utilization of data in any modern business of today is challenging traditional ways of adding business value and present significant risks to companies that don't respond accordingly. And of course it offers opportunities to those that do. Companies that transform their business models in parallel with these shifts will find new doors opening for them.

For example, in the home thermostat market, which is a traditionally a relatively stable sector with a small, settled list of competitors, a start-up called Nest has been able to challenge the established companies by introducing a thermostat that uses analytics to learn customers' preferences by analyzing data patterns — patterns that are then used for building a model for how the thermostat should adjust itself accordingly. The example of how Nest's novel, data-driven business model enabled it to enter a market long closed to outsiders is a good example of how data-driven business models can totally disrupt any traditional market segment.

However, the payoff isn't just for new players. For established companies, new data-driven business models can help keep and expand their share in an existing market. One recent example in the automobile insurance sector is the Snapshot app offered by major player Progressive. With Snapshot, data is collected from a small device that customers plug into their car's diagnostic port to help calculate premiums based on actual driving habits. Among the data analyzed is when and how far the customer drives and the number of hard brakes he makes. Good drivers are rewarded with lower premiums. On average, it could mean a savings of 10 to 15 percent, which can be a compelling value proposition to many drivers.

Creating data-centric businesses

The large volume of data that companies generate and the insights that data generates may well have value to other companies and organizations, both within and outside the industry it belongs to. Social media sites, for example, often capture data related to users' preferences and opinions, which could be information of interest to manufacturers that want to better focus their product development efforts and marketing campaigns. Mobile network operators routinely collect subscriber location data, which could be of value to retailers that want to know where consumers are shopping. By making this information available (for a price, of course), companies can develop new revenue streams through data monetization.


TIP

Though the sale of personal information traceable to specific individuals can raise privacy concerns, companies can greatly reduce sensitivity by aggregating and ensuring the anonymity of data through segmentation, for example. This means that individuals are first put into a group or segment based on their consumption habits, neighborhood, age, and so on. When the grouping is done, all personal data (name, address, and phone number, for example) is then removed so that it becomes anonymized.


REMEMBER

Identifying relevant applications is just the first step in deriving value from big data. New capabilities, new organizational structures (and mindsets), and significant internal change will also be required. But you should not underestimate the importance of zooming in on the right opportunities. You need to think outside the box, embrace new models, and even reimagine how and where you want to do business. A culture that encourages innovation and experimentation — and even some radical thinking — will serve this undertaking well, but so will calling in outside help when needed to assess, prioritize, and develop the different routes to value.

Data and machine intelligence isn't just changing the competitive environment; it's fundamentally transforming it. And your business needs to change along with it. Seeing where the opportunities lie and creating strategies to seize them will help your company turn the data promise into a reality. And that new reality will enable you and your company to gain new customers, new revenue, and even new markets along the way.

Investigating different types of data-driven business models

An important first step in realizing the potential benefits of data in your business is deciding what the business model(s) will be. The data economy supports an entire ecosystem of businesses and other organizations. These are often dependent on each other's products and services, so the strength of the sector as a

whole is crucial. For example, companies and organizations may share or sell data, models, algorithms, and insights, which is incorporated into new or enhanced solutions by other companies.

It's also worth considering that data products require a business model to determine how users will benefit from the service provided and how the value from data products and services will be generated. Many models are available for how to capitalize on the value of data and the services based on data. Which one you and your company should go for really depends on factors such as the type of service provided, whether it's related to a platform or a product, and how the customer will benefit from it. (One common monetization example is the freemium model, where users are offered part of a service for free but are charged for upgrading to the full service or are charged a premium for additional data services with an existing product.)

Figure 22-1 shows different high-level categories of data-driven business models and examples of areas within each category. The next few sections look at the categories in greater detail.

Data-driven business model types	Examples of offerings
Differentiation through data	• Data-driven and predictive business • Business contextual relevance • Create new offerings
Data and insight brokering	• Sell data • Sell insights • Sell analytical models • Sell ML models • Provide benchmarking
Infrastructure brokering	• Analytics and statistics tools • Cloud and data center services (IaaS) • Platform services (PaaS) • Data and analytics consultancy services
Data delivery networks	• Cross-licensing • 2-sided business models • Targeted advertisements
ML/AI functionality enablement	• Intelligent automation • Technology evolution • Robotics • New intelligent systems

FIGURE 22-1: Different categories of data-driven business models.

Differentiation through data

The differentiation-through-data category of data-driven business models refers mainly to how you use data to *differentiate* your current business — taking steps to strengthen it and make it more competitive. This can be done by using data to better understand your market and your customers, using data to drive decisions throughout the company or becoming more predictive, proactive, and preventative in business operations and toward your customers.

This category can also include areas such as expanding your current business by developing new types of services based on data related to your current business. In this sense, differentiation also creates new experiences. For a decade or so now, the world has seen technology and data add new levels of personalization and relevance to advertisements and location-based services as two examples. Google's AdSense delivers advertising that is actually related to topics users are looking for. Online retailers are able to offer — via FedEx, UPS, and even the US Postal Service — up-to-the minute tracking of where your packages are. Map services from Google, Microsoft, Yahoo!, and now also Apple provide information linked to where you are.

Data and insight brokering

Another business model category relates to how you can become a broker of data and insights. This includes selling raw, aggregated, or processed data (cleansed, labeled, or even correlated data, for example) of which you are the owner. It could also include selling data of which you are not the original owner, but then you need to make sure you have the rights to sell the data to a third party.

Another business model in this category includes selling specific analytical models for stand-alone purposes or to integrate into another solution or other applicable usages. Data-driven business models offer opportunities for many more service offerings that will improve customer satisfaction and provide contextual relevance. Imagine a map-based service that links your fuel supply to the availability of fueling stations. If you were low on fuel and your car spoke to your maps app, it could not only provide you with routes to the nearest open gas stations within a 10-mile radius but also receive the price per gallon. Who wouldn't pay a few dollars a month for a contextual service that delivers the peace of mind of never running out of fuel on the road?


TIP

Here's another example. In this scenario, as part of a business, you manage to own millions of pictures of items, including descriptions of what these pictures depict. On top of using this data for enhancing your own business, you could sell access to this data set for the training of machine learning models, — Deep Learning models, for example — as an additional revenue source.

Data and insight brokering can also include selling generic or specific machine-learning-based models designed either to run as a stand-alone products or to be integrated into existing software for enhancing its output. An example of the latter is a web based application aimed at providing a meeting place or an online marketplace for people who want to sell and buy used goods — the Swedish company Blocket or the American counterpart eBay, for example. None of these firms had machine learning models enhancing the applications in the beginning; these kinds of functionality — the ones supporting search, recommendations, and automated classification of images when publishing a new ad — were added later on.

Finally, such brokering could also include benchmarking services, which in this context refers to using data from several companies in a certain market or business segment to compare aspects such as market penetration, customer rating, or adherence to a certain applicable standard.

Infrastructure brokering

Infrastructure brokering takes a slightly different tack from the previous two categories. It aims to sell products needed to enable the first two categories. For example, it could offer infrastructure solutions for acquiring and collecting data and/or storing and processing data offered via cloud services. It could also refer to various types of reporting or analytics tools, offered on-premise or in the cloud. The solutions could be used for many purposes, including data exploration, data visualization, deriving insights for internal usage, or even commercializing the outcome by selling the derived insights. Finally, this category could also include consultancy services connected to infrastructure setup and usage.

Data delivery networks

The data-delivery-networks category of business models refers to those areas where profit comes out of bringing different businesses together in various marketplaces for the sharing and selling of data products — in other words, a convenient place to meet and share, even for competitors. In such a scenario, retailers like Amazon could sell raw information on the hottest purchase categories, and additional data on weather patterns and payment volumes from other partners could help suppliers pinpoint demand signals even more closely. These new analysis and insight streams could be created and maintained by information brokers who could sort by age, location, interest, and other categories. With endless variations, brokers' business models would align by industry, geography, and user role.

Cross-licensing of data is one type of offering sometimes used among competitors. Here, both parties agree to give the other party a license to collect and use data owned and locked in by the other party. By using the model of cross-licensing, each party gains (or loses) equal insight into the competitors' data. You often see such cross-licensing used in the telecommunications equipment-vendor business.

That's because many telecom networks are serviced by multiple vendors that have delivered telecommunications equipment to an operator's network. Given this multivendor telecommunications environment, cross-licensing of data comes in handy when there's a need to know more about the installed base of a competitor's equipment — which hardware, software, or set of configurations are used, for example — or where there could be a need to gain access to performance data from a competitor's equipment in order to gain full insight into the performance of the entire network. The data and insights derived could then be used to sell insights or other types of data-related services.

The data delivery networks enable the monetization of data on a larger scale. To be truly valuable, all this data has to be delivered into the hands of those who can use it, when they can use it, through different types of marketplaces. The data delivery networks take the data and aggregate it, exchange it, and reconstitute it into newer and cleaner insight streams — kind of like what cable TV does in terms of content delivery. These data delivery networks will be the essential funnel through which information based offerings find their markets and are monetized. Although their primary function is to be a marketplace for making business, they also function as a hybrid between a new type of offering and a delivery model.


REMEMBER

Few organizations have the capital to create end-to-end content delivery networks that can go from cloud to device. Today, only a few giants — such as Amazon, Apple, Bloomberg, Google, and Microsoft — show such potential, because they own the distribution chain from cloud to device.

Cross-licensing relies on an open marketplace that acts as a platform for data and model providers to meet the users. The 2-sided business model is similar in that it's built on the concept of bringing together the data and the different parties, but there's one significant difference — the 2-sided business model isn't open to everyone. It's a restricted setup, brought into being for only a specific purpose and outfitted with an active middleman that connects the different parties, securing both model delivery and monetization.

The basic concept of the 2-sided business model is that it has (at a minimum) three types of parties involved, although it could include many more. The main party acts as the middleman — the one offering a service to a client who is in need of insights from data but is unable to do that analysis by itself. There are many reasons why the client might not be able to do it — perhaps the company doesn't have access rights to the data or it lacks the right infrastructure for managing it or it lacks the domain expertise to analyze and draw conclusions from the data — but that doesn't hide the fact that the need is still there. Here's where the middleman comes to the rescue, buying the needed data from other data vendors and then performing the analysis on behalf of the customer, which then in turn pays for the insights.

Here's an example: Imagine that a global coffee vendor wants to understand how well its marketing campaigns for coffee are working in a certain area in the United States. The vendor of this particular data delivery network service is a global telecommunications vendor, which has launched an entirely new service built on a disruptive business model. The way it works is that the vendor of the service buys data from two operators in the US and uses segmented location data for a group of people living in a certain area. (No individuals can be identified, because they're anonymized as part of a group living in a certain geographical area.) Patterns of movement to and from the closest coffee shops are then studied, using the data from the operators gathered from peoples' mobile phones. With that data, it would be possible to determine how well the marketing campaign has succeeded for people living in a certain area. And this can be analyzed without violating an individual's privacy.

Figure 22-2 extrapolates from my coffee vendor example, showing the 2-sided business model in action. I cover the steps in greater detail later in this chapter.

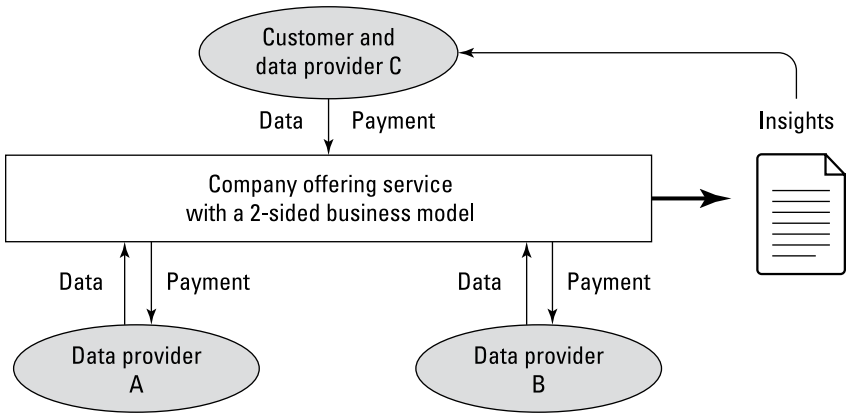

FIGURE 22-2: The 2-sided business model (data driven).

Note that Figure 22-2 shows that the customer is also providing data (as Data provider C), including data such as geographical area, time range, and location of coffee shops, but lacks the needed data from data providers A and B (the operator data with location information of coffee customers). Without that data, the global coffee shop vendor can't carry out the analysis.

Data providers A and B are competitors in the same market and therefore would refuse to just hand over their data to a competitor, even if they were to be paid, because it could reveal sensitive information about their businesses.

This opens up a possible use of the 2-sided business model, where a third party offering both data and business understanding of the telecommunication business

could act as a neutral player or middleman connecting the business of coffee shops with telecom operators.

Machine learning/artificial intelligence functionality enablement

All business models based on machine learning and artificial intelligence need data to exist and carry out their purpose (functional or other) and are therefore, per definition, data-driven business models. You can also use machine learning/artificial intelligence technologies to derive insights from the data, and those insights could be sold just as other insights can be sold. But despite the similarities with other models, business models based on machine learning and artificial intelligence are slightly different.

The main reason for investing in machine learning/artificial intelligence technology-based business models is usually to expand current business and technology with new and advanced techniques and functionality. This enhanced functionality can, for example, be used to evolve automation to a new level with intelligent automation, mainly focused on optimizing how to perform a certain automated task. For example, if the automation steps performed by any machine today are the same steps that were previously performed by a human in order to perform a certain service, with machine learning the machine can identify the best way to solve the task, regardless of which steps were previously performed. The machine isn't bound by a preconception of "the right way to do something," (assuming that the data, team, and algorithm are unbiased) but rather focuses on solving the task in the most optimized manner.

You could also use a machine learning/artificial-intelligence-focused business model to evolve the functionality of an already existing solution with the help of the dynamic and adjustable techniques that such a model offers. To see what I mean, here's an example of a telecommunications network where previously only nonmachine learning models were used in the software for thousands of base stations spread out across an entire country.

That's pretty much how things were until 2017, Now, however, you have a few machine language models being brought online inside the base stations, making it possible to dynamically adjust and better serve customers in real-time as the need for bandwidth changes with day-of-the-week, time-of-day, preferences for certain apps, geographical location, and so on. The machine learning models in the base stations are, of course, trained on real data before they're deployed, but they can then continue to learn the patterns for the different geographical areas they cover, which means they can then predict and prepare to serve customers as their needs arise. Instead of a one-model-serves-all (or nobody), the network dynamically adjusts proactively and in real-time to the constantly changing needs in the connected society.

Another way that machine learning/artificial intelligence can empower your various business models is to use it for completely new and disruptive business models, such as robotics. This is an expanding area that is now moving beyond the repetitive automation you find in factories, where the robots work alone, to scenarios where they become dynamic and intelligent assistants to humans in lab environments (co-bots), in our cars (self driving cars), in our gardens (robot lawn mowers), and even in our houses (robot vacuum cleaners).


REMEMBER

Many possible paths are open to you when it comes to monetizing the data revolution. What is crucial is to have an idea of which one you want to follow in your company. Only by understanding which business model (or models) best suits your organization can you make smart decisions on how to build, partner, or acquire your way onto the next evolutionary wave.

Using a Framework for Data-driven Business Models

As difficult as it seems to argue against the business value of data, leveraging the potential of data isn't as easy as it appears at first glance. The reasons are multilayered: Many companies lack expertise in the areas of data cleaning and storage, and data often only becomes valuable when aggregated with data from competitors or players in another industry, which might be either difficult or even impossible to achieve. This has given rise to several initiatives aimed at defining frameworks to support companies and organizations and offer a more structured approach to introducing data-driven business models.


REMEMBER

An important aspect to start with is to actively question your company's readiness and willingness to change and invest in data as a core part of the business, not just something that the company does on the side. Data-driven business models demand full dedication throughout the company, and it's dealing directly with your customers, so it's vital that you do it properly, once you decide to do it.

This short list of questions can be used as examples for how to approach your own company's self-assessment for readiness to introduce data-driven business models:

- Is my company ready to discuss data-driven business models?
- Which strategic business goals are my company pursuing? Are they strategically in line with a data-driven business model ambition — or in conflict with them?

- Is the ambition business-to-business (B2B) or business-to-customer (B2C) monetization?
- Is there organizational support (processes, governance, frameworks) for an idea to be evaluated in a structured way?
- Could the market potential be improved through cross-industry initiatives like an open source project or a standardization effort — and how would you approach that in that case?

After you have performed your company assessment and (hopefully) have decided to move ahead and introduce a data-driven business model, you have some key areas to consider:

- **Understand** what a data-driven business model means in your line of business.
- **Realize** the potential of a data-driven business model, and try to specify and quantify the relevance for your company's future.
- **Recognize** that data-driven business models are no mystery. They're already in use in various companies across different industry segments, even if most are still relatively new.
- **Use** a simple framework to guide the thinking and approach regarding data-driven business models.
- **Discover** existing and relevant patterns in the company, and get an overview of where you can get started and what to leverage early on.

Creating a data-driven business model using a framework

There's nothing simple about creating a data-driven business model for your company, so in order to offer you a practical approach to this task, I describe in this section the data-driven business model (DDBM) framework. DDBM can be used as a high-level innovation blueprint to identify benefits and challenges associated with leveraging data-driven thinking to construct data-driven business models.

The data-driven business model (DDBM) framework consists of six dimensions: key resources, key activities, value proposition, customer segment, revenue model, and cost structure. Figure 22-3 shows an overview of this model.

I describe the dimensions in the following sections.

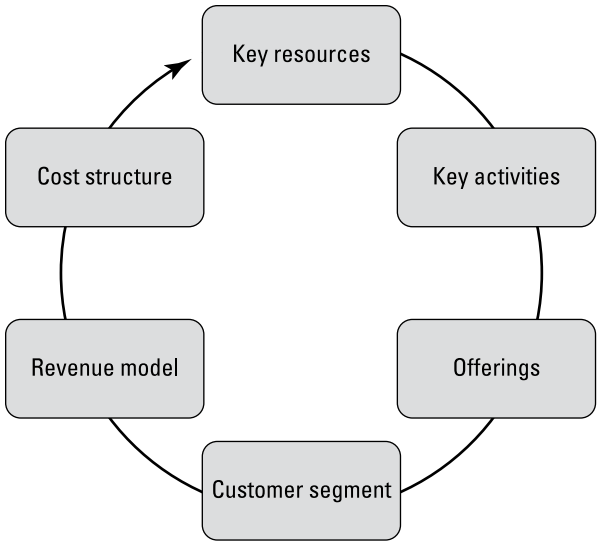

FIGURE 22-3: Data-driven business model (DDBM) dimensions.

Key resources

Companies need resources to develop their products or services as well as to create value. By definition, a DDBM has data as a key resource, but this does not imply that data is the only key resource of the respective business model. Your company might need other key resources to enable your business model — key competence and infrastructure, for example. But the main focus of the DDBM key resource is to explore and define what kind of data sources and data types are needed to fulfill the data-driven business model objective.

When identifying needed data sources, you should distinguish between external and internal ones:

- *External* data sources can refer to data types like customer or market preferences, open statistical data records, benchmarking data, or various social media data, like blogs.
- *Internal* data sources can refer to different databases with historical data, internal transactional system data, product and service data related to performance, quality, customer rating, as well as company financial data and so on.

Key activities

Like any other company, your company needs to perform different activities to develop, produce, and deliver its offering. In traditional product-centric business models, the key value-creating activities can be described using a traditional

value chain — a high-level model used to describe the process by which businesses receive raw materials, add value to the raw materials through various processes to create a finished product, and then sell the finished product to customers. However, because the traditional value chain concept is primarily focused on the physical world and treats data as a supporting element rather than as a source of value itself, it's of limited use in the context of data-driven business models.

Instead, you need to identify all key activities related to your specific data-driven business idea, including all data science life-cycle-related activities you carry out on data — capturing it, maintaining it, processing it, analyzing it, communicating it, and actuating it. The data-science-related activities will be of varying degrees of importance, depending on what your business model looks like, but it's worth spending some time thinking through the needed steps and how to secure a stable data science foundation in your DDBM.

Offering/value proposition

A *value proposition* can be defined as an expression of the experience that a customer will receive from a supplier and that is measurable by its value creation. That means the value proposition is the value created for customers through the offering. However, because it's difficult to formalize and categorize the perceived value by a customer in any industry, the DDBM framework focuses on the offering instead.


REMEMBER

What your company will offer as the core of the data-driven business model comes down to what is described earlier in this chapter as different types of business models: differentiation through data, data and insight brokering, infrastructure brokering, data delivery networks, and machine learning/artificial intelligence functionality enablement. It's important to have a solid business idea that you then take the time to specify properly and use the categories to define.

Customer segment

The customer segment dimension deals with the target group for the offering. Though there are several ways to segment customers, the most generic classification divides target customers into businesses or business-to-business (B2B) and individual consumers or business-to-customer (B2C).


REMEMBER

B2B is a situation where one business makes a commercial transaction with another business. This typically occurs when a business is sourcing materials for its production process for output (a food manufacturer purchasing salt as a raw material to produce an enhanced output, for example). B2C, on the other hand, refers to a business that sells products or provides services to end-user consumers directly.

In many cases, companies can target businesses as well as individual consumers. In B2B trade, it's often the case that the parties to the relationship have comparable negotiating power and, even when they do not, each party typically involves professional staff and legal counsel in the negotiation of terms. B2C is instead shaped to a far greater degree by an unequal balance between the parties: the company offering the product or service and the end user or consumer. In that relationship, the company is in a superior position when it comes to the end user in terms of the economic implications and access to relevant information.

Revenue model

To survive long-term, every company has to have at least one revenue stream. Several different revenue models can be distinguished using classifications such as these:

- **Asset sale:** Giving away the ownership rights of a product (a data product such as a data set, insights, or models/algorithms, for example) or a service in exchange for money
- **Lending/Renting/Leasing:** Temporarily granting someone the exclusive rights to use an asset (a data set, for example) for a defined period
- **Licensing:** Granting permission to use a protected intellectual property, like a patent (a model/algorithm, for example) or copyright, in exchange for a licensing fee
- **Usage fee:** Charging a fee for the use of a particular service (a defined scope for Insights as a Service, for example)
- **Subscription:** Charging for the use of a service or software product (machine learning enhanced software, for example) during a limited and agreed-on period
- **Brokerage:** Charging for an intermediate service where the business model works as the middleman, connecting data and insights with other parties (sometimes by establishing new marketplaces for these parties to meet and make business)
- **Advertising:** Providing advertisements on your site or in relation to a service. Doing so can supply an extra source of income or be your main income stream. For it to be effective for your advertisers, you need a good understanding of your target audience, in order to direct the right ads to the right target group, using the data at hand to explore and analyze the market and pricing potential.

TIP

Take the time to properly think through the type of revenue model or models your company is aiming for, both short-term and long-term. This will help you identify and drive other aspects in your data-driven business model when using this framework.

Cost structure

To be able to create and deliver value to customers, a company generates costs for labor, technology, purchased products, and so on. So, as part of this process, will the use of data enable a specific cost advantage? Well, typically a company would have a specific cost advantage if the data used in its product or service were created independently of the specific offering.

An example of this is a car manufacturer using data that is automatically created and stored by the electronics in the car. Then there are other companies, like Automatic, a start-up providing analytics for car owners like parking tracking, maintenance reminders, engine diagnostics, driving history and insights. To capture the same data, Automatic needs to install a specific hardware device connected to the car, which most likely would require a specific consent from each car owner as well.

Another example is Twitter, which could use its own data without additional costs to provide various analytical services like trends in opinions on a certain topic discussed in tweets, while companies like Gnip, a start-up company providing social media analytics, would have to pay Twitter for the same data, directly impacting its cost structure.

Putting it all together

After you have a proper understanding of your business idea and how you want to realize it through a data-driven business model, just go ahead and, using the DDBM framework outlined in this chapter, put together your six dimensions and the respective features per dimension. Sitting down and doing that work will help you define and develop a fully data-driven business model.

REMEMBER

For each dimension, at least one feature has to be selected; however, a company can have more than one feature for any dimension.

IN THIS CHAPTER

- Explaining the concept of delivery models for data products and services
- Exploring new delivery models
- Listing ways of delivering data products and services

Chapter 23
Handling New Delivery Models

Maybe you already have a pretty good idea about what data science means in your line of business, regarding both its challenges and potential. You might even have started to work through the different aspects of your own data science strategy, based on your business idea. Or, perhaps you're already constructing a full-scale data-driven business model and getting close to execution. No matter how far you have come, or if you have not even started, if you haven't thought through how you intend to deliver your new data products and services, you have left out a vital strategic aspect from your plans.

WARNING

Delivery models might seem like something that you can solve later, thinking that once you get started, you'll figure it out. But make no mistake, it's important to think through this aspect of your business model early on. How you intend to deliver — or might be forced to deliver, depending on customer demand or user expectations — can impose huge transformative changes on your company or organization.

Defining Delivery Models for Data Products and Services

A delivery model describes the way you intend to deliver the product or service offering you're planning to sell to a customer. For physical products, this means resolving how you plan to ship the product from the factory to a store where it will be made available for purchase by a customer. Depending on your business model, it could also be shipped directly from the factory to the end-customer — when it's been sold through online stores, for example. Things you need to consider in that context are mostly related to aspects like the number of factories and the location of factories (selecting countries for global enterprises), your need to have, depending on customer demand, expected delivery time, and frequency and desired ways to consume your products.

When it comes to data products and services, there are other considerations to take into account. For data products, it's mostly a case of digital and virtualized products and services, which require other types of delivery models and platforms, such as cloud-based services with different types of *as a Service* (aaS) models. Or, it could be unlicensed open source software or various types of data and machine learning/artificial intelligence marketplaces where some parts are open for everybody and other parts are locked unless you acquire a license.

WARNING

For data products and services, there are also legal, ethical, and security aspects to consider that differ from requirements on traditional hardware and software delivery models. Depending on legal restrictions in different countries, you might have to consider a delivery model where some countries in which you have a presence may have stricter legislation on data usage, especially related to using data containing personal information.

REMEMBER

If the data cannot legally leave the country, you cannot perform data processing, develop your models and insights, or deliver the outcome from another country than the one the data originated from.

Understanding and Adapting to New Delivery Models

In IT, the term *alternative delivery models* refers to replacing the traditional delivery models for software products and services with new kinds of strategies and processes that are intended to enhance the way technology is used. The rather

broad term *delivery model* is often carefully applied to new service models that have been made possible by advances in technology, such as those that support web delivered services.

Some of the alternative delivery models that experts most commonly talk about involve cloud services and Software as a Service (SaaS) models. Here, instead of selling software in a box on a physical CD or another storage media, the software is delivered over the Internet or another type of network connection. With these new types of alternative delivery models, users can choose to purchase services with subscription fees or buy an entire package while still getting its implementation over the Internet. Thus, alternative delivery models have actually become an important term to talk about, representing a rapid shift in the business world and in the ways that people buy and use software applications.

REMEMBER

Delivering as-a-Service is also a suitable delivery model for data products. Data products as-a-Service means to provide them on demand — scalable and secure. A user interface is often implemented via an app or a web interface, and the whole service is often made available through a cloud based infrastructure, including various platforms and applications.

TIP

Organizations that are typically not in the software industry actually need to start acting like software companies when delivering data products. There's a lot to learn from software delivery models, but be aware that the traditional ones are also constantly changing. New technology advancements, cost efficiency requirements, and user demands are some driving factors behind the constant need to find better and more appealing ways to deliver software as well as data products and services.

From the point of view of a data science strategy, it's easy to forget or underestimate the importance for your business to choose the right delivery model. This task is very important in order to reach your customers in the way they expect and need and that suits your line of business. And because the customer expectation will change over time, your delivery model must be flexible, scalable, and built in a way that you can respond to shifting demands over time. Once you understand the way you need to deliver your offerings, you will discover that the selected delivery model will influence much more than you think. It usually impacts areas such as the product and service development lifecycle, geographical presence of development sites, competence strategies, organizational and support structures, and even the actual operationalization of the data products and services.

Introducing New Ways to Deliver Data Products

The area of recommending efficient delivery models for different types of data products and services is still being researched and discovered as this book is being written. However, in this section you'll find a high-level overview of some examples of different models being used for different data product and service categories. In Figure 23-1, a new column has been added to the table from Chapter 22 (Figure 22-1), where I map business models and offerings. (Note that Figure 23-1 adds just a few examples of the different possible delivery models to my previous list and should not be considered an exhaustive.

Data Driven Business Model types	Examples of offerings	Examples of delivery models
Differentiation through Data	• Data driven & Predictive Business • Business Contextual Relevance • Create new Offerings	☐ Self-Service Analytics environments ☐ Apps, websites, product/service interfaces ☐ Products & Services
Data & Insight Brokering	• Sell data • Sell insights • Sell analytical models • Sell ML models • Provide Benchmarking	☐ Downloadable files ☐ Websites ☐ API's ☐ Cloud-services ☐ Market places
Infrastructure Brokering	• Analytics & Statistics tools • Cloud and Data center services (IaaS) • Platform Services (PaaS) • Data & Analytics consultancy services	☐ Downloadable license ☐ API's ☐ Online services ☐ On-site services ☐ Cloud-services
Data Delivery Networks	• Cross-licensing • 2-sided Business models • Targeted advertisements	☐ Market places ☐ API's ☐ Cloud-services
ML/AI Functionality Enablement	• Intelligent Automation • Technology Evolution • Robotics • New Intelligent Systems	☐ Products & Services ☐ Online Services ☐ On-site Services ☐ API's

FIGURE 23-1: Examples of delivery models for different data-driven business models and data product and service offerings.

As you can see, there are many different types of delivery models, and sometimes the same delivery model can be used for different offerings and different data-driven business models. The following sections describe these different delivery models in more detail and include some contextual examples.

Self-service analytics environments as a delivery model

When investing in a data-driven business model aimed at differentiating through data, one example of a delivery model is to utilize a self-service analytics environment. Most of these ready-to-use analytics tools are easy to use and are usually available as both on-premise installations and cloud-based solutions.

By using an off-the-shelf product for exploring and generating business insights from your data to drive better decisions, you focus your efforts on the data preparation and data analysis rather than on investing in building your own tool or platform from scratch. This is especially suitable for companies that are new to data science (and thus with little or no analytics or machine learning/artificial intelligence competence in-house), but also for internal business analytics purposes in any type of company, regardless of their analytics maturity level.


REMEMBER

It's also possible to use the output from an off-the-shelf exploration or analysis tool to generate a suitable dashboard or another visualization that could be used externally as well — toward your customers, for example — thus saving you the time it would take to design your own. Another benefit of using ready-to-use analytics tools is that they come with interactive visualization views, which is seldom the case when you create a visualization in Python or R, the two most common programming languages for data scientists. *Interactive* visualizations means that you can easily click on different parts of a visualization to zoom in or out, select an area for further analysis, or even change the scope of what you're looking at and analyzing.

Figures 23-2, 23-3, and 23-4 show how easily you can use a single application for data exploration, analytics and visualizing insights. Start with one visualization for your data set, and then easily change to another when you want to expand the analysis. Or simply combine various visualizations in to one view (like the examples below) and connect the graphs so when you change the scope of one view, the other graphs adjust to that scope too. The case below shows how you can enhance your understanding by easily adding location data to the traditional data set you usually look at.

By adding geographical context to your analysis and visualizations by combining traditional data with location data, location analysis brings the "where" dimension to the forefront so that you can analyze data in new ways to get the full picture before making decisions, while identifying location-specific opportunities.

Figure 23-2 below shows how you can get an overview of your customer data using a set of variables like where they live, how old they are, and their level of customer satisfaction. These different data variables can then be used for further

analysis and to search for possible dependencies. For example, the correlation matrix that has been automatically generated shows a strong correlation between "first order product quality" and "returning customer". Location data \ shown on the map indicates that the average customer loyalty probability across all the customers is 54 percent.

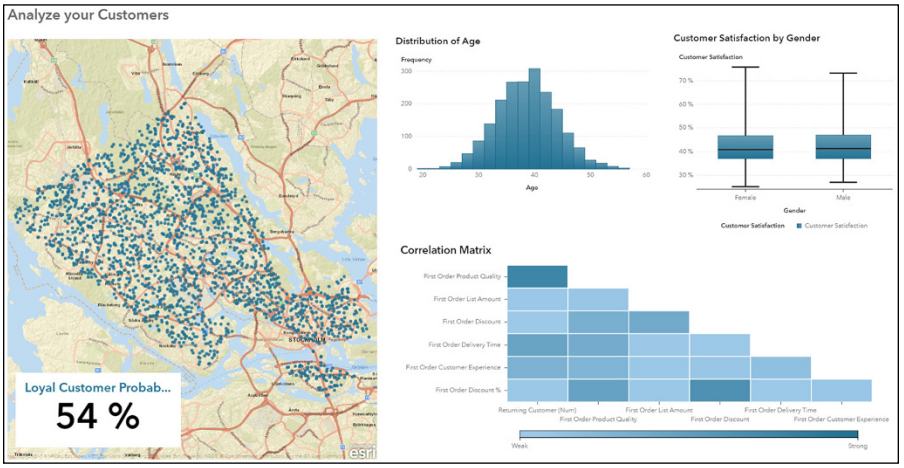

FIGURE 23-2: Using an analytics tool to explore your customer data and possible correlations related to where they live, their age and their level of satisfaction.

Figure 23-2 is based on a screenshot generated using SAS® Visual Analytics software. Copyright © 2019 SAS Institute Inc., Cary, NC, USA. SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. All Rights Reserved. Used with permission.

The graph in Figure 23-3 zooms in on one specific geographical area of the customer base. You can see the round circle in the upper corner. The graph then automatically adjusts the other connected graphs (the bar chart showing the distribution of age, the box plot showing the gender distribution, and the correlation matrix). As you can see, the customer loyalty probability in this area is significantly higher than for the average customer — 82 percent.

The decision to use easy-to-consume interactive analytics tools, however, is not the right one for all companies. Analytics and machine intelligence mature companies tend to lean toward doing everything by themselves, especially when it comes to analysis as part of commercial data products and services interfacing the customers. For example, there is a common perception that without a unique design on the user interface visualizing the data and insights, it will not differentiate towards competition. But remember to think through what to focus your development work on; do you want to develop insights or develop a tool for visualizing your insights? Differentiating is important, but focus your company efforts on the right tasks.

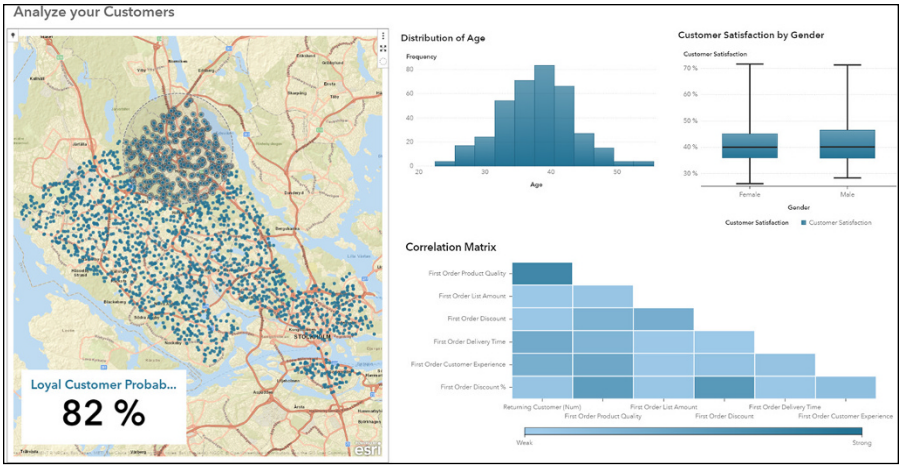

FIGURE 23-3: Graph focused on a certain geographical area selected in the map using drag and drop.

Figure 23-3 is based on a screenshot generated using SAS® Visual Analytics software. Copyright © 2019 SAS Institute Inc., Cary, NC, USA. SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. All Rights Reserved. Used with permission.

REMEMBER

The analytics tool companies are investing a lot of money, competence, and time into making these tools user friendly and state of the art in most dimensions. That is what they are specialized to do. Most of the tools are also quite adaptable for different needs, so be sure to calculate the alternative cost and time before you embark on building your totally in-house solution.

TIP

For purely internal purposes, out-of-the-box tools for analytics tend to be more cost efficient and faster from idea to insight; offer a more stable production environment; and have the ability to make analytics accessible to more and different types of employees, supporting the implementation of a data-driven organization across different company segments.

Applications, websites, and product/service interfaces as delivery models

When your ambition is to use apps, websites, or existing product and service interfaces as delivery models for differentiating through data, it's all about sharing data and findings with your customers.

This can be done by making your users' own data available to them through either the communication channel they're using or the communication channel your company is offering. For example, a mobile operator can make data available to its subscribers about their own costs and usage, the best subscription offer based on usage pattern, location based services, and more.

This empowers subscribers and gives them contextual understanding of how they're actually using their mobile phones, which means that they can be more in control of their current and future usage, including their costs. At the same time, it empowers the company that shares the data, because it sends a signal of transparency to its customers and generates trust, which can strengthen brand perception.

Another example of how a company shares users' own data to empower their brand is the Nordic company Skistar, which has facilities for downhill skiing in the Nordic countries. The company has an app in which you can create an account and upload the ID number from your digital ski pass. The ski pass is reusable for as long as it does not break, and you just reactivate it by paying the applicable fee when you need it for a new period.

The ski pass automatically connects to the ski-system during your day on the slopes every time you use a ski lift. The app provides you with data on how many rides you have taken, distance you have traveled, vertical meters achieved, calories burned, and so on. All of it nicely is aggregated per day, week, month, or year. It also allows you to connect with your friends so that you can compare your results.

Figure 23-4 shows two views available to skiers using the Skistar ski resorts.

FIGURE 23-4: Tracking your day on the slopes.

A dummy example of how Skistar gives back the data to the skiers by connecting the digital ski pass with their app to generate simple, but fun, insights.

Another example is an online business that uses the data generated on its home page to offer its customers other recommended products or services based on previous purchase patterns through the website, or offering reduced prizes to users who show a particular interest in specific items.

Websites can also be useful as delivery models for data products such as insights that you sell. The insight result could be in the shape of a dashboard or some similar visualization. When you have published the dashboard, you can give the customer a secure link to their own website where they can consume the insights through an interactive view and also download a .pdf file with a static view.


TIP

For an existing product or service that you offer, you could use new or existing data that wasn't previously made available for customers. The data can then be added to the same system interface as before — a financial system, for example. The purpose here would be to enhance the experience by adding new data, which could improve the user perception of the same product or service without functionally improving it.

Existing products and services

By integrating data as input or a key resource into an existing product or service and making your offerings data driven, you'll be able to differentiate through data. Here, the original product or service wouldn't be a data product, but could be using different forms of data for contextual purposes rather than as its main driving force. This would be the case regardless of whether the products would be hardware, software, or other, and regardless of whether the products were on-premise or were virtualized and deployed through a cloud service. The delivery model involves enhancing existing products or services, or identifying new data product opportunities through data utilization.

One example of such an approach is when you start using a data-driven and predictive approach to an existing service offering. This is especially interesting for real-time operational types of service offerings, like the extensive and complex operations of a telecommunications network. Without a data-driven approach and without the help of techniques such as predictive analytics and machine learning/artificial intelligence, services tend to stay reactive to faults or different types of alarms. But when data and models are used proactively to identify patterns in the data, it helps to understand what is causing certain problems, allowing you to predict and prevent problems from ever happening in the future. This will in turn improve service performance and network quality, as well as satisfaction of service operations by the customer, the network operator, and even the mobile network end-users, like you and me.

Downloadable files

For offerings related to data and insight brokering, the delivery model of choice is often downloadable files. If you're offering a data set small enough to put into a file and download from a website, this is an excellent delivery model. This could, for example, be the case for a file with test data for model training. Downloadable files also work when you're selling stand-alone analytical or machine learning models.


TIP

Another example when this is applicable is when your offering is an Insight-as-a-Service or some sort of report meant to be delivered. Usually, the size of a compiled report with insights, recommendations, and different statistical visualizations suits the downloadable format quite well.

However, it's worth considering from where the customer downloads the files. Constructing an easy-to-use but secure website where they can access the files is a good idea. On this site, you can also take the opportunity to inform your customer about other current and coming products that you're offering, or even open up the site for other companies to buy spots for advertising for other related data products to your customer clientele. Just make sure you pick companies that you do not perceive as current or potential future competitors, and keep the website simple and clean, with the main focus on the files you're delivering. Make it simple to access what they're after.

APIs

API stands for application programming interface, a set of clearly defined methods of communication among various components such as web-based systems, operating systems, database systems, computer hardware, or software libraries. Using an API as a delivery method is useful when customers want direct access to the data product or service (data, model, or insight) in order to integrate the product or results directly into their system environment.


TIP

APIs can also be useful when you're selling a certain machine learning capability as a service and are including the infrastructure necessary to run the model. This means you're not selling the model itself, but only the ability to use the machine learning model. The delivery model approach is that the customer uploads the data to you, using the API, and then runs the model in your environment. This is machine learning as-a-Service as well as a form of infrastructure brokering.

A concrete example of this is Amazon, which is offering this type of service using its cloud environment and its machine learning algorithm for image recognition for various purposes, like facial recognition and analysis, object and activity detection, unsafe content detection, celebrity detection, and even analysis of text

in images. Amazons image recognition offering enables you to search your image collection for similar faces by storing face metadata, using the `IndexFaces` API function. You can then use the `SearchFaces` function to return high confidence matches.

Cloud services

A *cloud* service is any service made available to users on demand via the Internet from a cloud provider's servers, as opposed to being provided from a company's own on-premises servers. Cloud services are designed to provide easy, scalable access to applications, resources, and services, and are fully managed by a cloud services provider. Examples of some well-known cloud providers are Amazon, Microsoft, and Google.

When you say you're using cloud services as a delivery platform, you could be offering a cloud service as an infrastructure or platform service for various services, like data storage, computation of data, or applications. But it could also refer to utilizing a cloud service as a delivery platform for data, insight, and model brokering, or a data delivery network for a marketplace, for example.

Because a cloud service can dynamically scale to meet the needs of its users, and because the service provider supplies the hardware and software necessary for the service, there's no need for a company to provision or deploy its own resources or to allocate IT staff to manage the service. This makes a cloud service an interesting delivery model for many different types of data products and services.

Online market places

Online marketplaces are sometimes also referred to as online e-commerce marketplaces. The marketplace is a type of e-commerce site where products or services are provided by multiple third parties and where transactions are processed by the marketplace operator.


REMEMBER

Online marketplaces are like platforms for multiple players and are well-suited fostering marketplaces for data products and data service offerings. They are already an important media and has the potential to become the main data delivery model in the future, driving deal-making in data, analytics, and artificial intelligence. It also enables advertising opportunities across multiple channels and industries.

A *data marketplace,* or *data market,* is an online store where people can buy data. Data marketplaces typically offer various types of data for different markets and from different sources. Common types of data sold include business intelligence, advertising, demographics, personal information, research, and market. Data types can be mixed and structured in a variety of ways. Data vendors may offer data in specific formats for individual clients.

Data sold in these marketplaces is used by businesses of all kinds, governments, business and market intelligence agencies, and many types of analysts. Data marketplaces have proliferated with the growth of big data, as the amount of data collected by governments, businesses, websites, and services has increased and all that data has become increasingly recognized as an asset. Data marketplaces are often integrated with cloud services. Figure 23-5 shows how the interaction between different marketplace actors may look.

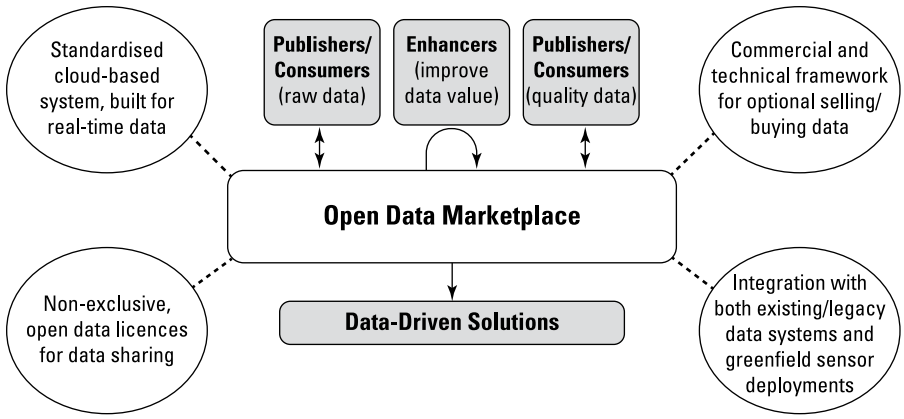

FIGURE 23-5: A model showing workflows on an open data marketplace.

Downloadable licenses

A software license is a kind of license that is used to set rules about how a piece of software can or cannot be used. After either downloading or buying the software, you need to agree with the license in order to use it.

For a business model like infrastructure brokering, downloadable licenses for different types of analytics software are common and useful. The downloadable software licenses are usually equipped with time constraints that make them impossible to use after the expiration date has occurred, unless the license is renewed.

Online services

An *online service* is a generic term that refers to any information and services provided over the Internet. These services not only allow subscribers to communicate with each other but also provide unlimited access to information. Online services can range from simple to complex. A basic online service may help subscribers gain needed data through a search engine, and a complex one might be an online mortgage application from a bank. Online services may be free or paid.

An online service is suitable for offerings like data and analytics consultancy services that don't need to be provided onsite. Many data scientists have found a lucrative business by selling their expertise online, delivering analytics and machine learning/artificial intelligence expertise in terms of data-driven recommendations and strategies, as well as machine learning/artificial intelligence models and solutions.

Onsite services

Onsite services is a delivery model that refers to services that take place on the same premises, or at the same location, as the customer. This type of service is usually needed when offsite services cannot be given or the customer is a complete novice. Compared to offsite services or online services, onsite support takes a longer time to set up and is usually more expensive.

6 The Part of Tens

IN THIS PART . . .

Grasping why you need a solid data science strategy

Learning about ten common mistakes to stay away from

IN THIS CHAPTER

- Creating an excellent base for all levels of communication
- Understanding why it's all about making choices
- Realizing what needs to be considered early on
- Aligning views and realizing objectives

Chapter 24

Ten Reasons to Develop a Data Science Strategy

This book spells out many of the challenges you'll face when embarking on a data science journey within your company. It emphasizes what is fundamental and what not to forget, but it also points out areas of specific interest and choices of specific importance. One thing it hasn't done (yet) is make the argument for why it's vital for you to develop and document all your strategic ambitions into a data science strategy. That is what this chapter is all about. Enjoy!

Expanding Your View on Data Science

Taking the time to develop a data science strategy is crucial. It forces you to learn more about what data science really is before you start investing and making important choices. Having a strategy in place lowers the risk of missing vital steps and considerations along the way.

REMEMBER

Though data science is a blend of different disciplines — like mathematics, statistics, and computer science — do not be mistaken: It is a discipline of its own. Understanding the key concepts and considerations driving the area of data science is vital but often not even done.

My view is that when you truly understand what data science is all about, you'll look at your company in another light, from another perspective. It will be obvious to you what needs to be done differently, and you'll be able to explain why this is the case. Then you can motivate those around you to make the necessary changes, because in data science, it all starts and ends with data. Perhaps many companies that have been around for a while don't think of themselves as structured around data that way, but they need to be if they are to succeed in the new data and artificial intelligence age.


REMEMBER

Google uses data as the starting point for everything. By using artificial intelligence and machine learning techniques to detect patterns and deviations in the data, Google can decide in a truly data-driven manner which business to go for and which areas to prioritize and take action on. At Google, data drives organizational change, new innovation, and business priorities. And its overall leading slogan is, as you can imagine, *AI first.*

Aligning the Company View

If you drive your data science strategy the right way, you'll have an opportunity to bring people together around the business opportunities sure to result from your data science investment. It's important to formulate that vision and mission and capture it in a data science strategy that is agreed on by all stakeholders. By doing so, you ensure that everyone is committed to the stated objectives and that they're anchored in the organizational structure early on. That gives you a strong and solid foundation for the vast and challenging work ahead.

However, it's easier said than done to align an organization around data science. Why is that? Well, for starters, people's views about the insights into what data science is and how it will be transformative for different businesses are quite varied. That means you won't be starting at the same level of understanding of what it means to introduce data science into the company. If some enter into the undertaking assuming that data science can be added into a corner of the company as some kind of add-on and be expected to generate value, you'll have issues further down the line.


TIP

To get the full potential from your data science investment, you must treat it as the dominant discipline. If you are able to align your company around such a perception, and capture the details of how this will be approached as part of a data science strategy, you have positioned your company for success.

Creating a Solid Base for Execution

By actually writing down your data science approach and priorities, you're establishing the foundation for the plans needed to execute the strategy. It helps steer the business in the right direction and provides a baseline to rely on when challenges appear and new opportunities arise.


TIP

A vital component of such a solid base is an architectural drawing, in which your infrastructure can be realized and implemented. This requires quite a lot of detailed thinking in a cross-domain team setup, not only to detail the setup and execution approach in different domains but also to think through how this will be executed in a data- and machine-driven setup across the whole company in a fluid manner.

Of course, the strategy might need to change over time, due to changing needs or priorities in the company, or even to an evolving data science technology. But, regardless, it gives you a solid foundation to stand on and start with when considering a new direction or modifying plans regarding execution.

Realizing Priorities Early

To be honest, for a medium- to large-size company, making an end-to-end investment in data science is expensive. But because the improved business potential is much greater, companies are realizing the necessity of investing in a future driven by data science.


WARNING

In order not to get lost in all the unavoidable complexity on the way toward that future, it's crucial to understand and clearly set your priorities early and then try to stick with them when things get more difficult. Doing so will help guide you right during challenging periods.

What, then, are the typical priorities you need to consider early? Well, an important part of your work in coming up with a strategy should be to look at your current business setup. In doing so, some of the more important questions to consider early on are listed here:

- What is it that you really want to change?
- What is the data science potential in your line of business?
- What are your true expectations?
- How will you realize your expectations in practice?

Putting the Objective into Perspective


REMEMBER

Creating an end-to-end data science strategy forces you to not only set clear objectives but also consider them from many perspectives. It's not only about the business potential; it's also about the legal rights, privacy concerns, or other ethical considerations related to the data you're using.

When considering the context of the fundamental change that needs to occur in your business environment when introducing data science, you might need to consider to what extent your company depends on the following factors:

- Data you might not own yourself, which could mean limitations on usage
- The need to digitalize all parts of your business in order to become data driven from end to end
- The necessity for new roles, competencies, and skill sets in data science among employees and managers
- New laws and regulations that were previously not applicable
- Working with new vendors and partners related to data and infrastructure support
- Potentially addressing an entirely new customer base

Creating an Excellent Base for Communication

Yes, it's a lot of work to put a data science strategy in place, aligned and anchored with main stakeholders, but once it's in place, it will provide you with an excellent base of communication. You can easily use the written strategy and work its defined objectives and challenges into presentation material and company targets.

The strategy can be used to build up a communication plan for the different target groups identified, including the different priorities and considerations agreed on regarding the new mindset and culture you want to enforce.


TIP

You can turn the content from your data science strategy into more consumable formats, like an FAQ, and then turn selected parts into external material for communication with your customers, partners, and vendors.

Understanding Why Choices Matter

TIP

A strategy is all about making choices — choices for what is important to pursue and what is not worth going after. If your strategy is to try to do everything, you're lost. That is worse than a bad strategy; it's no strategy at all.

REMEMBER

Because the choices you make in the strategy will be the guiding star for coming choices and priorities during the challenge of fully introducing data science, you absolutely must make the right choices at the beginning. Making the wrong choices at that point is sure to have a severe impact on the overall success of your investment in data science.

So, what can you do to make sure you make the right choices? I recommend the following:

- Take the time you need in order to get your strategy right, and be sure to iterate. Don't rush!
- Provide your main stakeholders (and you yourself) with a basic level of understanding in data science.
- Involve internal *and* external data science experts in the area to make sure you gain a broad and varied perspective on the market situation.
- Take advice from your data scientists internally (if there are any), and let them contribute actively to the strategy.
- Engage main stakeholders in decisions regarding cross-domain challenges and priorities, even if it's difficult and cumbersome.
- Make tough choices, but be ready to adjust along the way depending on enhanced understanding or changing conditions.

Identifying the Risks Early

By actually taking the time to consider risks as part of your data science strategy, you can not only detect risks early but also potentially prevent them from ever becoming a reality.

TIP

Finding a good structure to use when identifying potential risks is the key, and it will help you through this not-so-fun exercise. I know it's much more appealing to think about all the new possibilities the future may bring rather than what can potentially go wrong with your investment. However, it is time well invested to do this.

Some main risk areas to consider include the ones described in this list:

- **Data:** Are you in ownership control of all the data you will need in order to realize your internal efficiency ambitions or for realizing external business opportunities? If not, have you secured the necessary legal rights to the data for what you want to do today and potentially in the future?
- **Competence:** Do you have the necessary skill set to execute on your strategy? If not, have you set the right business ambition in relation to the availability of such competence — considering the time it takes to internally build experience and/or attracting and retaining data scientists among the scarce number available on the market, for example?
- **Infrastructure:** Have you thoroughly examined the risks related to your architectural ambitions? (Examples here include going open source or not, virtualized and cloud based environment or not, and distributed and local setups or centralized setup.) There are many risks associated with the infrastructure architectural choice as well as the implementation challenges (setups for global companies moving data and distributing computation and automation cross borders, for example.)

Thoroughly Considering Your Data Need


REMEMBER

Performing a thorough inventory of your data need is vital when it comes to your data strategy work. It provides a practical understanding of the business priorities, infrastructure need, legal and ethical considerations, data governance aspects, as well as business potential of that strategy. It all starts with the data. It's as simple as that.

The data inventory should include aspects such as these:

- Classification of data in terms of type, format, degree of sensitivity, collection point(s), and ownership
- Grouping of data types into data categories with similar attributes (lowering the number of individual types to handle)
- Usage need in terms of needed level of data granularity, collection frequency, and data retention periods

TIP

Once you have the inventory in place, you can use it to create a data model to help your understanding of (and preparation for) data interoperability (which data needs to be combined with which, and how can it be analyzed together?); how the data need impacts the infrastructure setup (what requirements may be derived from the collective data need?); and what must be protected from a legal and security perspective (which laws and regulations are applicable for which data types, and what does that mean?).

Understanding the Change Impact

The data strategy is a good way to get a firm grip on the total scope of the change that is needed in order to achieve your objectives. That means you can start planning early for the necessary cultural shift in mindset and behaviors that is needed. It enables the transition to be well planned and proactive so that it isn't rushed but is introduced step by step. Allowing employees to mature in their perception of what data science is and what this will enable is the next step.

REMEMBER

One aspect of data science that is significant to acknowledge has to do with people's fear of how the introduction of automation will impact ordinary jobs. (This is closely linked to the further fear that machine learning algorithms will replace the need for humans in the workplace.) Such fears need to be taken seriously, and it's important to communicate clearly what the intent is with the change, how it will impact your employees' tasks and roles, what benefits it will bring, and what opportunities are opening up as a result of the change.

TIP

Becoming data driven and investing in data science also means that you should apply the same thinking when managing change. Approach the transformation program from a data-driven angle, and use analytics and machine learning techniques to measure and understand the change efficiency and impact. Using a method like sentiment analytics, for example, enables you to understand how the change is perceived among stakeholders. Other aspects you want to cover include to what degree the change is actually happening and whether there are specific change roles that are more efficient than others. What are they doing that others are not doing?

REMEMBER

By defining a data science strategy, you get an opportunity to secure a broader management understanding of what data science really is, the opportunities it enables, and the fundamental change impact that data science imposes.

IN THIS CHAPTER

- Underestimating the fundamental shift that data science imposes
- Perceiving AI as a magic solution to any problem
- Forgetting about ethical aspects
- Neglecting to measure the change

Chapter 25

Ten Mistakes to Avoid When Investing in Data Science

Although you must focus on your data science strategy objectives in order to succeed with them, it doesn't hurt to also learn from others' mistakes. This chapter gives you a list of ten challenges that many companies tackle in the wrong way. Each section not only describes what you should aim to avoid but also points you in the direction of the right approach to address the situation.

Don't Tolerate Top Management's Ignorance of Data Science

A fundamental misunderstanding occurs in the area of data science regarding the target group for data science training. The common view is that as long as the skill set for the data scientists themselves is improved, or for the software engineers who are training to become data scientists, you are spot-on. However, by

adopting that approach, the company runs the significant risk of alienating the data science team from the rest of the organization. Managers and leaders are often forgotten.

REMEMBER

If managers don't understand or trust the work done by the data scientists, the outcome won't be utilized in the organization and insights won't be put into action. So, the main question to ask is how to secure full utilization of the data science investment if the results cannot be interpreted by management.

This is one of the most common mistakes committed by companies today, and the fact is that there's also little training and coaching available for line management and for leaders. But without some level of understanding of data science at the management level, how can the right strategy be put in place, and how can you expect management to dare to use the statistical results to make substantive decisions?

WARNING

Without management understanding of data science, it's not only difficult to capture the full business opportunity for the company, but it might also lead to further alienation of the data science team or to termination of the team altogether.

Don't Believe That AI Is Magic

Data science is all about data, statistics, and algorithms. There's nothing magic about it — the machine does what it's told to do. However, the notion that the machine can learn causes some to think that it has the full ability to learn by itself. To some extent, that is correct — the machine *can* learn — but it's correct only within the boundaries you set up for it. (No magic, in other words!) A machine cannot solve problems by itself, unless a machine is allowed to develop such a design. But that's advanced technology and not today's reality.

WARNING

Overestimating what artificial intelligence can do for your company can really set you off on the wrong track, building up expectations that can never be met. This could lead to severe consequences both within the company and externally, with impacts not just in terms of trust and reliability but also in terms of financial performance. As important as it is not to underestimate the potential in artificial intelligence, one should also avoid the opposite extreme, where its potential is overestimated. I repeat: Artificial intelligence isn't magic. Yes, it's called artificial intelligence, but a more correct definition is actually *algorithmic* intelligence. Why? Because at the end of the day, very advanced mathematics are applied to huge amounts of data, with the ability to dynamically interact with a defined environment in real-time.

Don't Approach Data Science as a Race to the Death between Man and Machine

Some people tend to believe that task automation, driven by machine learning predictions, truly means the end of humans in the workplace. That prediction isn't one that I believe in. However, it does mean a significant change in competence and skill sets as well as a change in which job roles will be relevant and which types of responsibilities will be the focus in the workplace.

Like the introduction of the Internet in the workplace, introducing artificial intelligence in a more mainstream format will change what jobs are and how they're performed. There will be a lot less "hands-on" work, even in the software business. And yes, machines will most probably do a lot of the basic software development going forward, which means that people in the hardware-related industry will not be the only ones replaced. At the end of the day, basically all humans will be impacted as machine learning/artificial intelligence and automation capabilities and capacity expand and evolve beyond what is possible to do today.

However, this also means that humans can move on to perform other tasks that are different from the ones we do today — managing and monitoring models and algorithms and their performance, for example, or setting priorities and acting as a human fallback solution in cooperation with the machine. Other typical human tasks might be managing legal concerns related to data, evaluating ethical aspects of algorithm-based decision-making, or driving standardization in data science. You could say that the new human tasks will be focused on managing the machines that manage the original tasks — tasks that were previously perceived to be either boring and repetitive or too complex to execute at all.

WARNING

This "putting man against machine" business isn't the way to approach your data science implementation. Allowing the narrative to be framed that way may scare your employees and even prompt them to leave the company, which isn't what you want. Your employees are valuable assets that you need in the next stages as well, but perhaps in new roles and with new acquired skill sets.

TIP

Embrace what the machine learning/artificial intelligence technology can do for a specific line of business. Company leaders who understand how to utilize these techniques in a balanced approach between man and machine to augment the total performance and let the company evolve beyond its current business are the leaders whose companies will succeed.

Don’t Underestimate the Potential of AI

As strange as it may seem, some companies just don’t understand how transformative artificial intelligence really is. They refuse to see the fundamental shift that is already starting to transform society, and cannot see artificial intelligence as anything other than just another software technique or a set of new programming languages.

TIP

The key here is to a) take the time to truly understand what data science is really all about and to b) not be afraid to accept help from experts to identify and explain the strategic potential for your specific business. Because the area of data science is complex, it requires domain expertise and experience in terms of both the development of a strategy and its implementation. It also requires the ability to read and interpret where the market is moving in this area.

WARNING

By underestimating the impact that artificial intelligence can have on your business, you run the risk of significantly limiting the future expansion of your company. Later, once the true potential is really understood, you will find yourself entering the game too late and being equipped with the wrong skill set. You may finally be put out of business by competitors that had seen the potential much earlier and therefore invested earlier and smarter in artificial intelligence.

Don’t Underestimate the Needed Data Science Skill Set

A typical sign of companies underinvesting in data science is when you find small, isolated islands of data science competence spread out in different parts of a large company. In smaller companies, you see a similar symptom when a small-but-competent data science team is working on the most important project in the company but the only one outside the team that realizes its importance is an outsider like yourself.

Both of these examples are signs that top management in the company has not understood the potential of data science. They have simply realized that something is happening in this area in the market and are just following a trend to make sure that data science doesn’t pass them by.

WARNING

If the awareness and competency level of management doesn’t improve, the area will continue to be underinvested, distributed in a way that it cannot reach critical mass, and therefore rendered incapable of being scaled up at a later stage.

Don't Think That a Dashboard Is the End Objective

It may sound strange for, someone knowledgeable in data science, to say that anyone can think that the main outcome of data science is a dashboard. I can assure you, however, that this is a common misunderstanding. This isn't only wrong — it's also one of the main reasons that many companies fail with their data science investment.

At many companies, management tends to think that the main purpose of analytics and artificial intelligence is to use all that big data that has been pumped into the expensive data lake, to automate tasks and report on progress. Given such a mindset, it should come as no surprise that the main focus of management would be to use these techniques to answer *their* questions with statistically proven methods that could produce results that could be visualized in a nice-looking dashboard. For someone new to the field of data science, that might actually seem like a good approach. Unfortunately, they would be wrong.

REMEMBER

To be absolutely clear, the main objective of analytics and machine learning/artificial intelligence isn't simply to do what you've always done but using more machines. The idea is to be able to move beyond what you're able to do today and tackle new frontiers.

If the only end goal was to create a dashboard in order to answer some questions posed by a manager, there would be no need to create a data-driven organization. The idea here is that, in a data-driven organization, it all starts with the data and not with the manager and the dashboard. The starting point is what the data is indicating that you need to look at, analyze, understand, and act on. Analysis should be predictive, in order for the organization to be proactive and for its actions to be preventive.

The role of the dashboard should be to surprise you with new insights and make you discover new questions you should be asking — not to answer the questions you've already come up with. It should enable teams to monitor and learn from ongoing preventive actions. The dashboard should also support human or machine discovery of potential trends and forecasts in order to make long-term strategic decisions.

WARNING

In the real world, the steps needed to design a dashboard tend to end up being the most important tasks to discuss and focus on. Often, dashboards end up driving everything that is done in the data science implementation program, totally missing the point about keeping an open and exploratory approach to the data. This tends to happen because the dashboard is the simplest and most concrete

deliverable to understand and hold on to in this new, complex, and constantly changing environment. In this sense, it acts like a crutch for those unwilling or unable to grasp the full potential of a data-driven business.


REMEMBER

You run the great risk of missing the whole point of being data driven when your starting point is all about designing the dashboard and laying down all the questions from the start. By doing so, you assume that you already know which questions are important. But how can you be sure of that? In a society and a market now undergoing huge transformations, if you don't look at the data first and let the algorithms do the work of finding the patterns and deviations hiding there, you might end up looking at the entirely wrong problem for your business.

Don't Forget about the Ethical Aspects of AI

What does artificial intelligence ethics actually refer to, and why do you think it's of the utmost importance? Well, there are many aspects surrounding the idea of ethics in AI, many of which can have a severe impact on the artificial intelligence results. One obvious but important ethical consideration is the need to avoid machine bias in the algorithms — biases where human preconceptions of race, gender, class, or other discriminatory aspects are unconsciously built into the models and algorithms.

Usually, people tend to believe that they don't have biased opinions, but the truth is that we all have them, more or less. People tend to lean in one direction, subconsciously or not. Modeling that tendency into self-learning algorithms can have severe consequences on the performance of the company's algorithms.

One example that comes to mind involves an innovative, online, and artificial-intelligence-driven beauty contest. The algorithm had learned to search for the ten most beautiful women in the US, using only digital photos of women. But when studying the result from the contest, it became clear that something must have gone wrong: All of the ten most beautiful women selected by the algorithm were white, blonde, and blue-eyed. So, when studying the algorithm again, it turned out that the training set used for the algorithm had a majority of white, blonde, and blue-eyed women in it, which taught the machine that this was the desired look.

Other aspects in addition to machine bias include areas such as the use of personal information, the reproducibility of results outside the lab environment, and the explainability of AI insights or decisions. It's also worth noting that this last aspect is now a law within the GDPR (General Data Protection Regulation) in the EU.


REMEMBER

Ethical considerations are for our own, human protection as machine intelligence evolves over time. You must think about such aspects early on. It's not only a fundamental aspect to consider as part of your data science investment, but it's actually also hugely important to consider already from the start, when designing your business models, architecture, infrastructure, ways of working, and the teams themselves. Not wanting to break the law is of course important, but securing a sustainable and trustworthy evolution of artificial intelligence in your business is far more important.

Don't Forget to Consider the Legal Rights to the Data

When becoming data driven, one of the most common mistakes is to forget to make a proper analysis of which data is needed. Even if your main ambition with your data science investment is focused on internal efficiency and data-driven operations, this is still a fundamental area to address.

Once the data need is analyzed, it's not unusual to discover that you need other types of data than you originally thought. It might be data other than just internally generated data, owned by you. An example might be faults found in your products or services, or perhaps performance related data. It could even be the more sensitive type of data, which falls under the category of privacy data, related to how your products or services are being used by your customers.

Data privacy is an area that's getting more and more attention, in society with consumers' enhanced awareness of how their data is being used and also in terms of new laws and regulations on data. One concrete example is the General Data Protection and Regulation law (GDPR), introduced in 2018 within the EU with significant penalties for violators.

Although you might not have any plans for monetizing your data or to build new products based on the data, the whole rights issue is still central — even when all you want to do is analyze the data in order to better understand your business, enhance and innovate the current portfolio, or just improve the efficiency of your operations.

WARNING

No matter what your reasons are for using the data, you still need legal rights in place in order to use it! It's absolutely vital to address this early on as part of the development of your data strategy. If you don't, you might end up either violating the law regulating data usage and ownership or being stuck in terms of not being able to sell your new fantastic product or service because it's using data you aren't entitled to use.

Don't Ignore the Scale of Change Needed

If you don't take the time to properly sketch out the different change scenarios for your business when introducing a data science strategy, you most likely will fail. The fundamental shift needed in the company to become truly data, analytics, and machine driven is significant and should not be underestimated.

The most common mistakes in data science related to managing change are listed here:

- Underestimating the scope of the change and not taking seriously enough what has to happen
- Failing to recognize that business models are sure to be impacted when introducing data science
- Approaching customers with a value argumentation based on introducing data science techniques without explicitly explaining what the customer value is
- Pricing models to stay the same or not reflect the increased value, only the lowered cost
- Focusing single-mindedly on cost efficiency when it comes to business operational changes
- Neither measuring nor understanding operational improvements
- Carrying out organizational changes on so small a scale that everything stays the same in practice, ensuring that the actual change never occurs
- Building the cost and dimensioning model on old and outdated criteria, therefore ensuring that the model won't capture the new values
- Failing to see the change that data science imposes on the company and not understanding that change from an ecosystem perspective
- Underestimating the need for communication related to the change

Don't Forget the Measurements Needed to Prove Value

REMEMBER

A common mistake is to forget to introduce baseline measurements before the data science investment is made and implemented. Most of the focus in these cases tends to be on the future measurements and the results targeted with the investment. This is usually because of a resistance toward investing in new measurements in the current situation, because it's being abandoned for the new strategy. Unfortunately, this means that the company will lack the ability to statistically prove the value of the investment in the next step. Don't fall into that trap! It could truly backfire on the entire strategic ambition, when top management or even the board of directors asks what the value was of this major investment.

WARNING

Financially, you could, of course, be able to motivate the investment on a high level; however, it would be difficult to prove individual parts. Efficiency gains such as speed, agility, automation level, and process reactiveness versus proactiveness are values that are more difficult to prove and put a number on if you haven't secured a measurement baseline before executing your data science strategy.

Index

Numbers

A

B

C

D

E

F

G

H

I

N

O

P

Q

R

S

T

U

V

W

X

About the Author

Ulrika Jägare is an M.Sc. Director in Technology and Architecture at Ericsson AB. With a decade of experience in analytics and machine intelligence as well as 19 years in telecommunications, she has held numerous leadership positions in both R&D and product management. Ulrika was key to the launch of Ericsson's machine intelligence strategy and commercial approach as well as the recent Ericsson Operations Engine — a new data and AI driven operational model for network operations in telecommunications.

In addition to this book, she is the author of two highly referenced technical books by Wiley: *Unified Analytics For Dummies (Databricks Special Edition)* and *Embedded Machine Learning Design For Dummies (Arm Special Edition)*.

Dedication

I dedicate this book to my patient and supporting family — Emil, Rasmus, and Fredrik. I love you all very much!

Author's Acknowledgments

I would like to express my gratitude to everyone who has helped me write and produce this book. Firstly, I would like to thank Lillian Pearson, who put me in contact with Wiley and for also writing the foreword to this book.

Additionally, I give a deep thanks to my innovative and competent colleagues at Ericsson for sharing all the ups and downs in our joint data science journey.

Then I would like to extend a huge thanks my husband Fredrik, who is also working in the data science area, for all the great discussions and elaborations.

Furthermore, I would like to thank the supportive editorial team at Wiley for the relevant and sometimes humorous feedback and suggestions; Katie Mohr, Paul Levesque, Becky Whitney, and other editorial staff.

Publisher's Acknowledgments

Acquisitions Editor: Katie Mohr
Senior Project Editor: Paul Levesque
Copy Editor: Becky Whitney
Editorial Assistant: Matthew Lowe
Sr. Editorial Assistant: Cherie Case
Production Editor: Mohammed Zafar Ali
Cover Image: © gleitfrosch/Getty Images

WILEY END USER LICENSE AGREEMENT

Go to www.wiley.com/go/eula to access Wiley’s ebook EULA.