

- » 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.

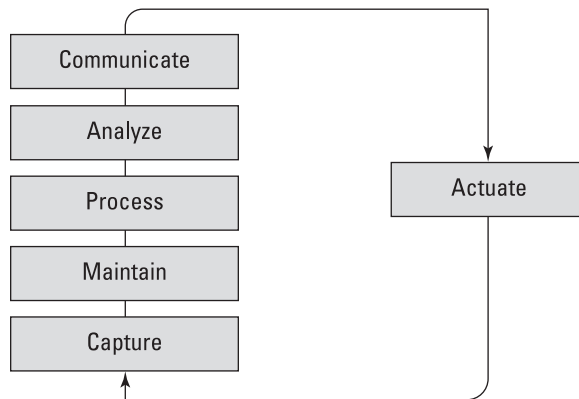


FIGURE 1-1:
The different
stages of the data
science life cycle.



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.)

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.



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.



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.



WARNING

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.

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).

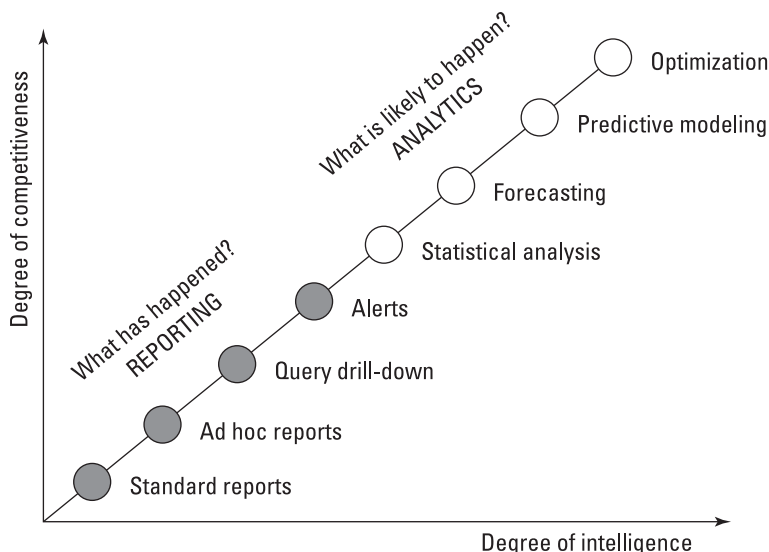


FIGURE 1-2:
The difference
between
reporting and
analytics.

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)	Frequenc	Fuel Consumption (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	Tran 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	330i 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	SUV	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	XType 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	8	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.

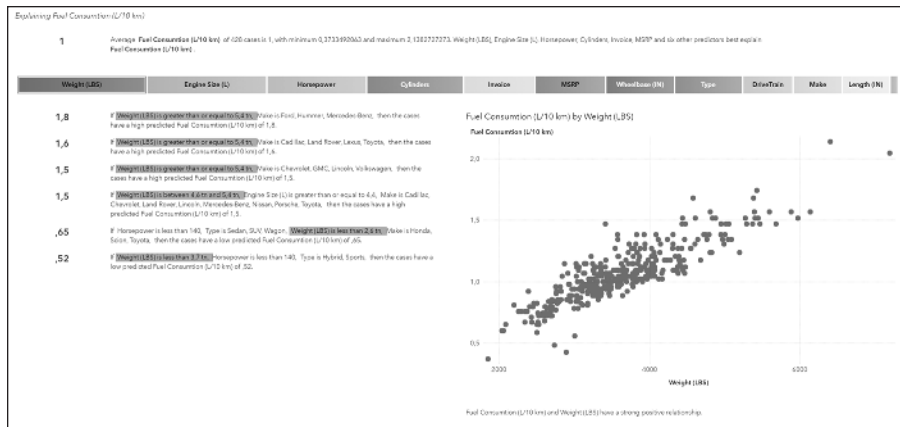


FIGURE 1-4:
Visualizing
your data.

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.

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

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 and 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!

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.

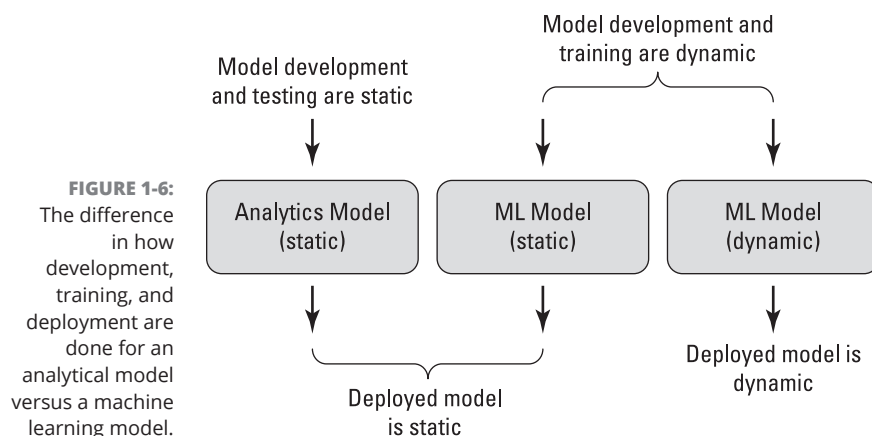
So, here's how advanced analytics and ML have some characteristics in common:

- » 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.



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.