

launch your AI mission with data science

edition 1.9



Intellerts

We recommend this paper if:

- You want to know more about Data Science
- You think Data Science is just about programming in R or Python
- Or you're a manager, advisor or consultant that wants to understand Data Science

In fact, this paper is for anyone who wants to improve their business BUT...

This is not an exhaustive course or a course in statistics or ML

But it will help you understand Data Science better and, ultimately, help you make better business decisions.

Index

PROLOGUE

Data-Driven Organizations	4
UNDERSTANDING DATA	4
- WHERE DO YOU STAND?	
- 5 STEPS WHEN EMBARKING ON YOUR DATA SCIENCE JOURNEY	
- 7 ADVANTAGES OF A DATA DRIVEN ORGANIZATION	
EVERYONE IS A DATA SCIENTIST	7
OUR 8-STEP DATA SCIENCE MODEL	10

CHAPTER ONE

The Art Of Asking Questions	13
5 AREAS TO ADDRESS	13
5 KEY SUCCESS FACTORS	15

TWO

Data Landscaping	16
8 CONSIDERATIONS WHEN EXPLORING DATA SOURCES	18
3 REASONS TO USE EXTERNAL DATA	19
TYPES OF METADATA	20
PURPOSES OF METADATA	20
5 DIMENSIONS OF BIAS	20
1. DATA CONDITIONS	
2. MODEL CONDITIONS	
3. DATA SCIENTIST CONDITIONS	
4. IMPACT ON STAKEHOLDERS	
5. IMPACT ON COMMUNITY	
INTEGRATED APPROACH OF DATA SCIENCE AND ETHICS	

THREE

Data Understanding	24
6 THINGS ABOUT DATA QUALITY	24
10 THINGS ABOUT DATA MANAGEMENT	25
6 ELEMENTS OF GDPR	26
5 THINGS ABOUT ANONYMIZATION	28

FOUR

Mining & Combining	30
3 DATA SCALES	30
4 TYPES OF JOINES	31
5 LEVELS OF DATA VALIDATION	33
EXPLORATORY DATA ANALYSIS (EDA)	34

FIVE

Deep Into Domain	35
3 TYPES OF DOMAIN KNOWLEDGE	36

SIX

Modeling	37
ASPECTS OF KEY PERFORMANCE INDICATORS (KPI)	38
3 CONDITIONS FOR AN ALGORITHM	39
TYPES OF ANALYTICS	39
TYPES OF AI	40
3 TYPES OF MACHINE LEARNING	41

SEVEN

Feed The Eyes	43
DATA VISUALIZATION ORIGINS	44
VISUAL VARIABLES	45
RANKING VISUAL VARIABLES	
PRINCIPLES FOR DESIGN	47
MISTAKES TO AVOID	47

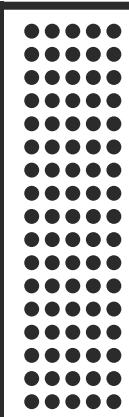
EIGHT

Tell The Story	48
EVALUATION FIRST	48
DIFFERENT AUDIENCES	48
STORYTELLING FORMATS	49
3 WAYS TO MAKE A DASHBOARD	50

EPILOGUE

Final Thoughts	52
WHY DATA SCIENCE PROJECTS FAIL	52
DATA SCIENCE IS A TEAM SPORT	54

Data-Driven Organizations



UNDERSTANDING DATA

What is a data-driven organization? Are all companies not data driven, to some extent? Why are we now labelling this as something new? Is it because we have now labeled data as a precious substance?

When it comes to AI, there's a lot of confusion. Everyone is jumping on the AI bandwagon. Companies want algorithms. They may not understand why but they are certainly afraid of missing out on the latest tech trend.

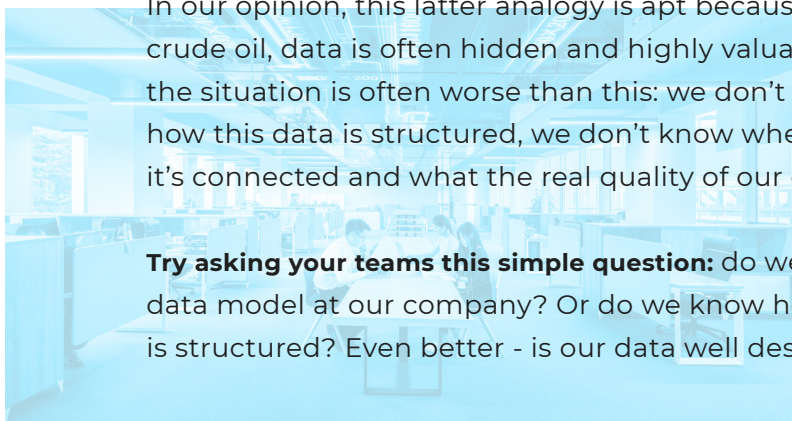
AI is something of a misnomer and companies often just need some business intelligence, those data analysis strategies and technologies to extract insights from their corporate information.

In other words, today's companies need information to optimize their business. That seems like a simple request, but is, in reality, a continuous challenge.

Let's take a moment to understand the data. Data is regularly called the new corporate currency or the new corporate oil.

In our opinion, this latter analogy is apt because, just like crude oil, data is often hidden and highly valuable. Actually, the situation is often worse than this: we don't even know how this data is structured, we don't know where it is, how it's connected and what the real quality of our data is.

Try asking your teams this simple question: do we have a data model at our company? Or do we know how our data is structured? Even better - is our data well described?



So many individuals labor over such questions, trying to identify the right path to become a data-driven organization with grand visions and intuitive technologies. But first, you must get to grips with your data.

Let's look at an example. You may remember mail-order companies. One commendable quality of these companies was that they continuously invested in maintaining a high-quality customer base. When these businesses were sold, their value was based on the number of customer records. This is because the buyer was, effectively, paying for a valuable customer database.

Thanks to advancing digitization, things have changed. Data has moved away from physical, paper-based repositories and into online, cloud-based databases. To truly be a data-driven company, **you must now understand that an individual customer record requires investment to acquire and to maintain, and that value is only created by accurately capturing your relationship with those customers.**

In other words, you must start treating your data as a critical company asset. This provides a sound basis for your decision making with regards to your AI, BI and Big Data technology investments.

WHERE DO YOU STAND?

Let's start by assessing your data maturity. Where do you think you sit on this scale?

DATA AVERSE - "We don't need data to tell us what we already know."

DATA UNAWARE - "Data has never been a priority for us. What's the benefit?"

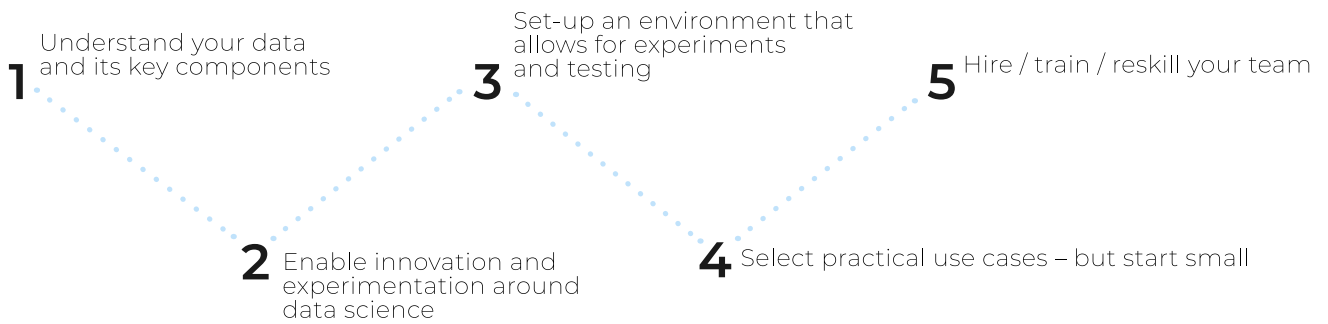
DATA AWARE - "We need to make better use of our data, but we don't know how."

DATA GUIDED - "Data confirms every decision we make and the actions we take."

To become data driven, the textbooks often advise companies to 'create a data culture'. While that's a very commendable recommendation, cultures do not change overnight. They may not even change in a couple of years.



5 STEPS WHEN EMBARKING ON YOUR DATA SCIENCE JOURNEY



7 ADVANTAGES OF A DATA-DRIVEN ORGANIZATION

As you become more data-driven, you will realize many benefits for your business. These benefits provide you with a competitive advantage, having a positive impact on your bottom line. In fact, research from Forbes reveals data-driven organizations realize 20% to 30% improvements in their EBITDA. Advantages include:

ONE STRATEGY OR POLICY DEVELOPMENT

Decisions about your strategy and policy are now targeted because data-driven insights give a broad, objective and complete view of your customers and business.

FOUR EFFECTIVENESS

Your information is approved and available across your whole organization.

SEVEN QUALITY

The consequences of your actions are measurable, providing your business with the ability to adjust and adapt appropriately.

TWO AGILITY

Fast and accurate responses provide quick insights, giving your business the right information, at the right time.

FIVE INNOVATION

Advanced analytics enables the development of new products and services.

THREE TRANSPARENCY

Your choices are now auditable and, consequently, better informed with the right analysis.

SIX EFFICIENCY

Gains are provided because your decisions are based on facts, the easy identification of bottlenecks and seamless automation.

EVERYONE IS A DATA SCIENTIST

The shortage of Data Science talent is well documented. Organizations face an uphill struggle to recruit (and retain) Data Scientists and, for now, we just must deal with this fact.

But this situation is changing. Coding is regularly touted as an essential skill and is now a popular classroom activity with many children now taking courses in R or Python.

This seems like sound advice. Data Science is on the rise as a profession and the volume of data is skyrocketing. This has a knock-on effect as our technologies continue to change our inherent need for AI-fueled knowledge.

However, data exploration is not the exclusive domain of the Data Scientist.

Many jobs are similar in nature to the role of a Data Scientist. This fact is often overlooked. But everyone regularly gathers data, cleans data, performs analyses, pivots in Excel, models problems, interprets results, extrapolates, and optimizes.

Data Science is now a diverse domain and no one Data Scientist can cover all of the available domains.

THE ROLE OF TODAY'S DATA SCIENTISTS

Looking at their day-to-day lives, Data Scientists spend almost 70% of their time cleaning, preparing and organizing data. This is a crucial part of the job, but you could argue that some of these tasks do not require a PhD in mathematics or statistics.



What's more, many young and talented Data Scientists are not particularly happy in their job. They regularly move between companies, with an average tenure of less than two years per job.

This is because these individuals need support and mentoring to find their feet and connect with the key business challenges of their work. But, in a real-world setting, there is often no time to focus on their professional development. Data is demanding. It needs cleaning, reports must be made, and so on. When Data Scientists do not get the right support, technical tools or the opportunity to apply latest Machine Learning techniques, disillusionment soon kicks in.

To overcome this, organizations must help Data Scientists understand their business functions and how they link to the world of AI and Data Science.

In the end, all business professionals must apply analytical thinking. So, what next?

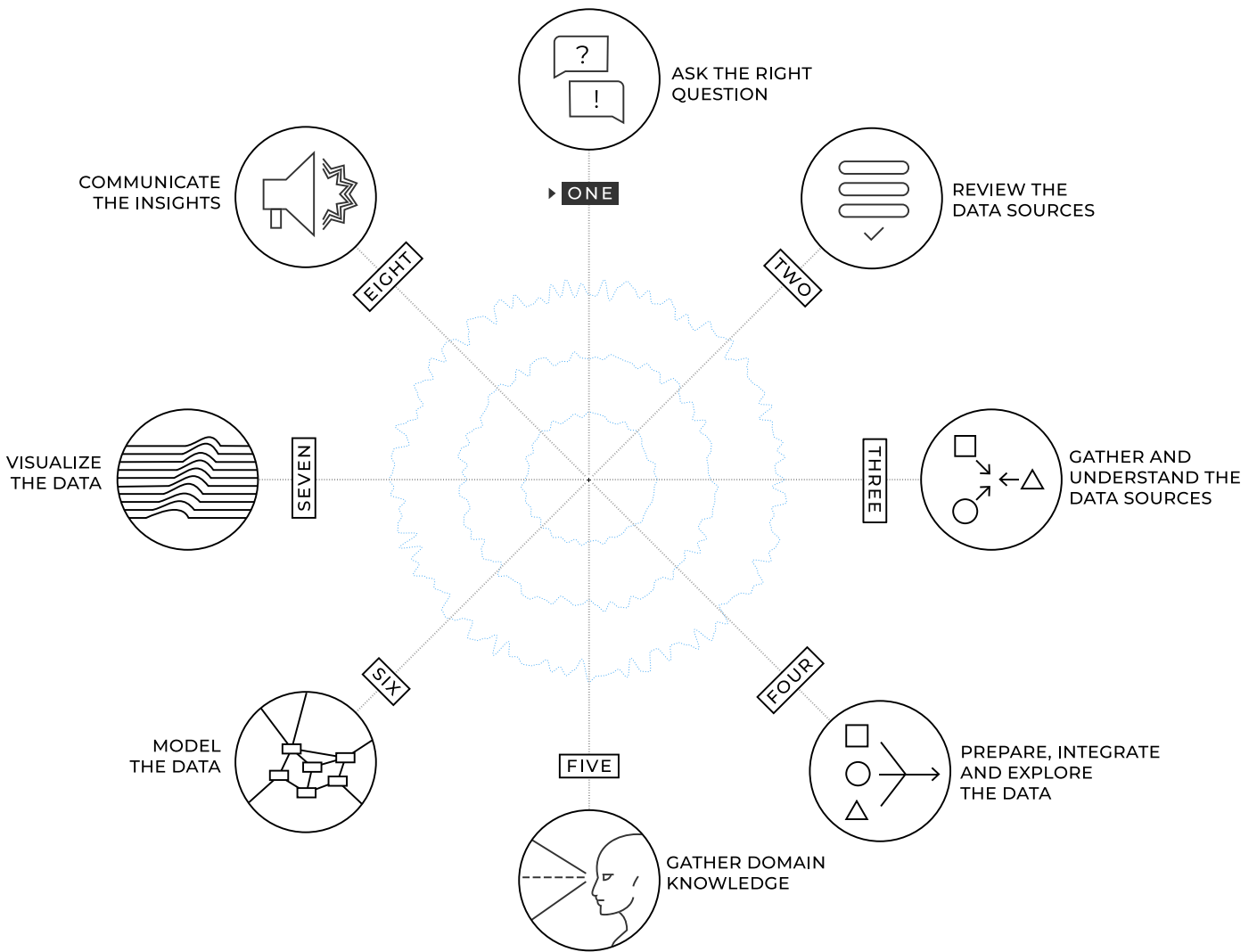
First, we must make fundamental changes both from a business and Data Science perspective.

Everyone who is **not** considered a Data Scientist must start to embrace data, whether they like it or not, understanding how data is organized and why. After all, data is the new corporate currency, the new corporate oil. Everyone must understand the basics of the technologies involved, and the inner workings of AI, traditional statistical modeling and optimization methods.

Today's Data Scientists must also embrace the world of business and gain a better understanding of how real-world innovation really works. When working with data and algorithms, building solutions, they are the domain experts. But that does not mean that their efforts will be thoroughly embraced by 'the business' (either internally or by the customer).

An integral competence framework highlights how the quality of a Data Scientist should not be judged purely on the quality of their R or Python skills. Yes, it is important to have these skills, but they represent just one of the many areas of knowledge now required by today's Data Scientists.





OUR 8-STEP DATA SCIENCE MODEL

We have defined an 8-step Data Science model, providing a framework for managing your data science projects.

In this section, we explain how you can organize your work to help you use data analytics to streamline your Data Science initiatives.

This 8-step model works at every stage of your Data Science project cycle. *It brings the right people into your organization to optimize your processes and technologies.*

As a result, your organization achieves the agility and transparency to make the right data-driven decisions, at the right time. You can meet your data management and data governance requirements. The 8-step Data Science model also stimulates innovation across your enterprise, helping you develop new data-based products and services.

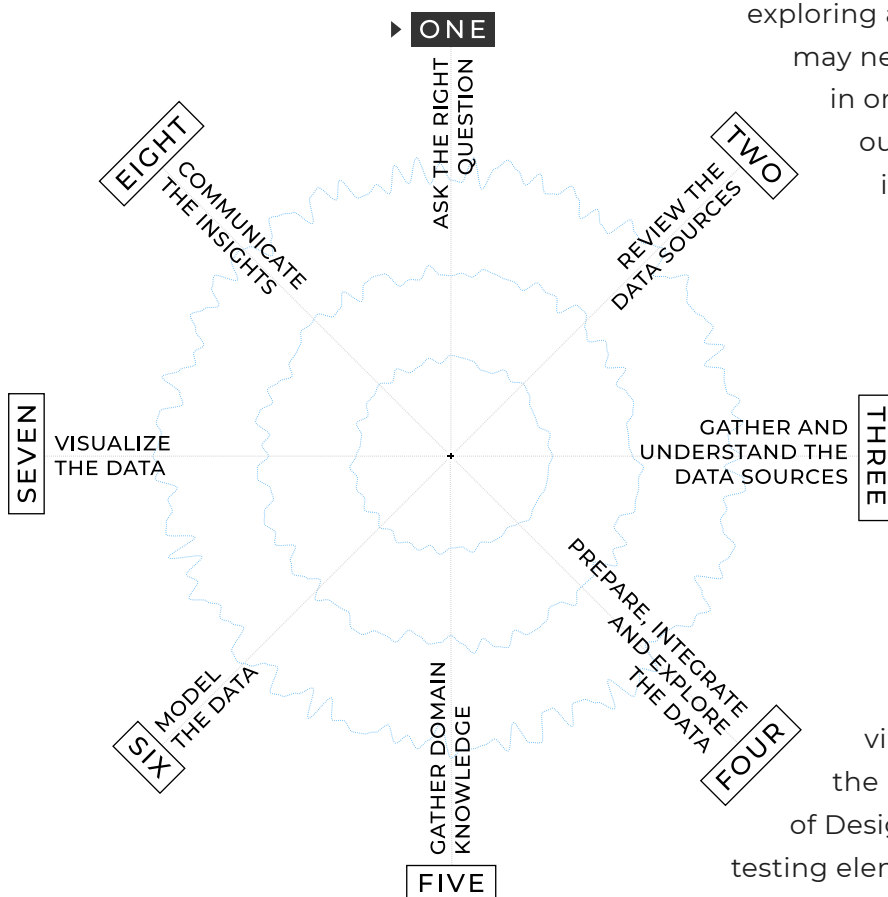
Let's examine how it works. The 8-step Data Science model is cyclical in nature. Although sequential steps are defined, you must often iterate back to previous steps, for example, when you realize certain questions require clarification.

Some activities can also happen in parallel. For example, you should start gathering domain knowledge when you start qualifying the problem area. But, after exploring and understanding the data, you may need to move forward to the next step in order to interpret the intermediate outcomes and, possibly, adjust your initial questions.

It's a harsh fact, but many data projects fail. To increase your chance of success, it is important to follow a proven methodology.

Our 8-step Data Science framework is a combination of existing methodologies and elements from the Design Thinking (problem solving) and Lean Startup (business model viability) methods. We have taken the breakthrough thinking element of Design Thinking and used the product testing element from Lean Start-up.


We are introducing a scientific approach to the world of innovation. We are combining an explorative approach to existing problems and, when we reveal new insights and information, your product ideas can further tested and validated.





These concepts are often criticized for oversimplifying the design process and trivializing the role of technical knowledge and skills. By combining this with a more scientific and technical approach from the existing BoK methodology, we are bringing together the robustness of technology, with the benefits of creativity.


As the volume of data continues to skyrocket, this is a much-needed capability. New business models and entire industries are now pivoting. As a result, creativity and technology must go hand-in-hand.


The 8-step Data Science model provides a practical and comprehensible framework for any data project, whether it is a BI-related project or an advanced-AI one. This framework is split into eight distinct steps:


 **ONE ASK THE RIGHT QUESTION** Here the problem or challenge is identified and qualified. The required resources and support for project success are also arranged.


 **TWO REVIEW THE DATA SOURCE** Once you have a clear understanding of the problem, you can start to explore the relevant and available data sources, including internal and external sources. The estimated quality and (direct and indirect) costs are also taken into account.


 **THREE GATHER AND UNDERSTAND THE DATA** The previously selected data sources are reviewed to gain a clear understanding of the content, as well as the data quality and structure.

 **FOUR INTEGRATE AND EXPLORE** The goal here is to gain insights from the data by looking for patterns and correlations. Before a Data Scientist can start their exploration work, the data must be cleaned and prepared, often by integrating multiple data sources and by creating new variables.

 **FIVE GATHER DOMAIN KNOWLEDGE** Once the Data Scientist has identified patterns and correlations, it is time to reach out to domain experts to validate or explain those results.

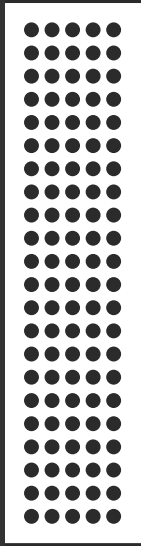
 **SIX BUILD A MODEL** This step is all about building and validating a data model. For a BI project, you may want to create KPIs. For an AI project, you may need to create a machine learning model, as an example.

 **SEVEN VISUALIZE THE DATA** Visualization is an important step for any data project. For a BI-project, it is an integral deliverable. For AI-projects, visualization plays an important role to explain the function of the machine learning models.

 **EIGHT COMMUNICATE THE INSIGHTS** In order to achieve true communication, collaboration is needed to fully understand and exchange insights created.

CHAPTER ONE

The Art Of Asking Questions



GOAL

Understand the problem and define the goals ("pain and gain").

TOPICS

Clarify the requirements of the question, define and sharpen the scope, manage expectations, determine the ethical considerations for go/no-go.

PITFALLS

Scope is too wide or unclear, accountability and responsibilities are not clear, no ethical checks.

"THE IMPORTANT THING IS NOT TO STOP QUESTIONING. CURIOSITY HAS ITS OWN REASON FOR EXISTING." Albert Einstein

The failure of many data science projects is often fixed from the start. This is because the problem is not clearly defined. Albert Einstein fully recognized this issue and there are a few reasons why this is so important.

FIRST, everyone in the organization must have the same understanding of the problem. If you can achieve this, it much easier to get a shared understanding of the complexity of the problem. **NEXT**, you must understand and agree on the impact of the problem and the solution. Then, you can assess whether the solution is aligned with your business strategy. Once this happens, your organization can determine the resources and required level of support.

5 AREAS TO ADDRESS

ONE To gain a shared understanding of the problem across the organization.

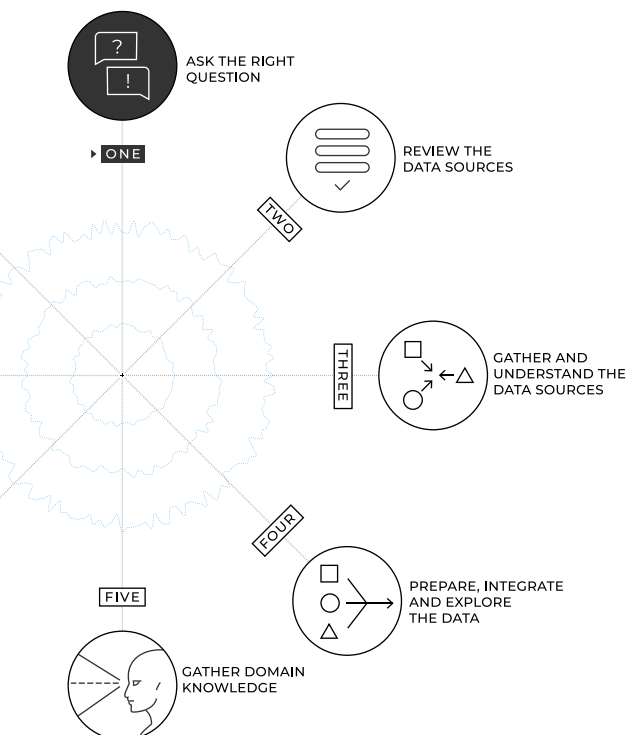
TWO To understand the impact of the problem and solution.

THREE To understand whether the solution is aligned with the business strategy.

FOUR To arrange resources and support within the organization.

FIVE To understand the complexity of the problem.

Defining a clear problem statement is not a task. It is a process. The first step is to explore the current situation to get a high-level understanding of the problem and its impact. This is then followed by a deeper examination to explain the problem.



Before thinking about a solution, it is important to assess whether there is a deeper, underlying problem. Only then can you start thinking about writing a problem statement.



Remember, the solution must be aligned to the business strategy, clearly articulating the desired benefits for the organization. Before writing the problem statement, it is important to contextualize the problem. This step has a backward- and forward-looking perspective. You must look back at the lessons learned from your previous approaches. You must also look forward at any possible constraints that might exist when implementing a solution.

By asking the right questions, you can arrive at a clear and agreed problem statement and your desired outcome. In real-life, organizations often make mistakes during this process, adversely affecting your chance of success. The most common mistakes are linked to a lack of common understanding of the problem. This can be caused by thinking about the solution too early, not identifying the deep-lying problems, or not understanding or describing the problem adequately.

Problem statement issues can also arise if the relevant domain experts are not involved in the process. It is also important to link this work to the wider business, aligning the problem with your business strategy and ensuring buy-in from stakeholders.



5 KEY SUCCESS FACTORS

With a good understanding of the particular business challenge, we can prepare for a data science project. Here are 5 success factors to consider::

ONE
A CLEAR UNDERSTANDING OF THE SCOPE

A clear view of the size and complexity of the project is required. When the project scales, it is better to adjust the scope to ensure the project's manageability and make sure the short-term results can be achieved. This makes it easier to manage expectations and communicate the project scope across the different levels of the organization.

TWO
UNDERSTANDING THE IMPACT OF THE PROJECT

To start, you must identify the people who will be impacted by the project. They must be involved at an early stage of the project. Moreover, Data Science projects typically lead to changes in your tools, processes and teams. So, it is important that these changes are properly addressed, and change management is applied to help your organization during this transformation.

THREE
GETTING THE RIGHT RESOURCES

You must set up a team with the right skills and sufficient bandwidth. These skills must go beyond technical data science skills, and also include project management, change management and communication skills. The project must also be properly funded with a sufficient budget to invest in the right tools, infrastructure, training and, potentially, recruitment.

FOUR
PREPARING FOR CHANGE READINESS

This critical component is often overlooked. When starting a data project, you need a clear picture of how change-ready your organization is. This is often a formality for those organizations with experience in this area. For those entering unknown territory, it is important to ensure you have enough support and don't overstretch. The project must be supported by change management to avoid opposition to your proposed project. The "leading coalition" has a major role to play here. This is a group of multi-talented people who can support the organization.

FIVE
ROLES AND RESPONSIBILITIES

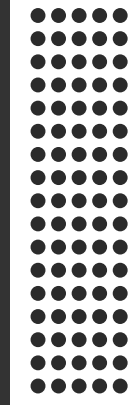
The tasks and roles in a Data Science projects require a broad set of competencies and demand strong collaboration. Data Science is a team sport and in order to be successful as a team it is essential to have roles and responsibilities.

In our next coming chapter, we will get into exploring the data landscape.



CHAPTER TWO

Data Landscaping



GOAL

Gain a detailed insight into the available internal and external data sources.

TOPICS

Exploring relevant sources. For each eligible source, understand how the data is collected and whether there is possible bias, data privacy check.

PITFALLS

High data costs, data is not available or is of poor quality, where good descriptions or metadata are lacking.

Once you establish a clear understanding of the problem, you can start to explore the relevant data sources.

Exploring the data landscape is often an underestimated task. This is mainly because companies lack a good understanding of their data structure.

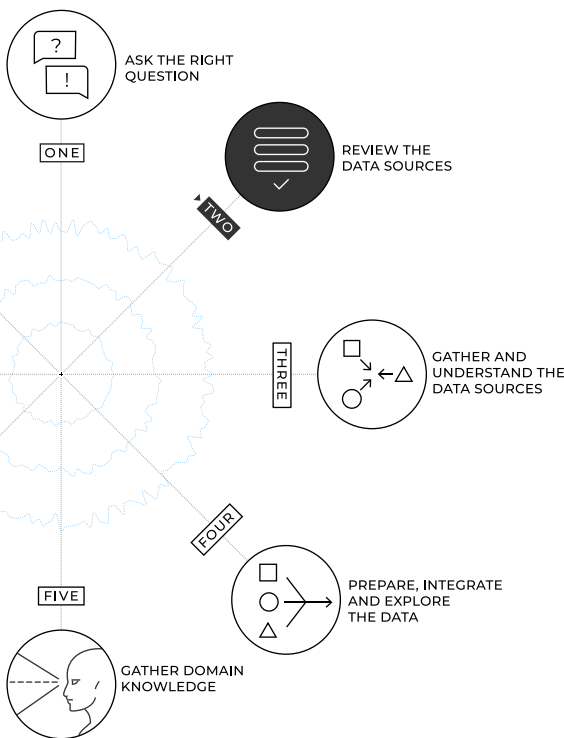
In our current age of digitization, data is growing at an astonishing rate and AI is predicted to have a dramatic impact on our society and workplaces. Well, AI is already impacting a range of real-world scenarios. But these scenarios are relatively narrow, limiting the impact of AI to a selected set of use cases.

When we scale up and start to use AI in more comprehensive applications, AI implementations often fail. For example, you may start to use AI to support your everyday decisions and assess *complex* data.

This is not a novel scenario, it's very similar to the issues other IT projects face like ERP and CRM implementations, for example. When you start to scale up the size and *complexity* of the problem you're dealing with, your problems also grow exponentially.

Data is key to resolve these issues. Yet, the term "data is the new corporate currency" is thrown around regularly nowadays. So, how are you going to capitalize on this data? The obvious answer is to "mine" your data.

However, we would say you need to "mind" your data to get the best data models (and results) possible.



In fact, the data model is the one element consistently underestimated by today's enterprises. If you ever want to truly benefit from AI in your organization, you must understand the data models you use.

This issue crops up time and time again. When embarking on any data initiative, you must understand the data, and ask your information experts the following question: Do we have a data model for our key systems?

The definition of a data model, according to Wikipedia, is:

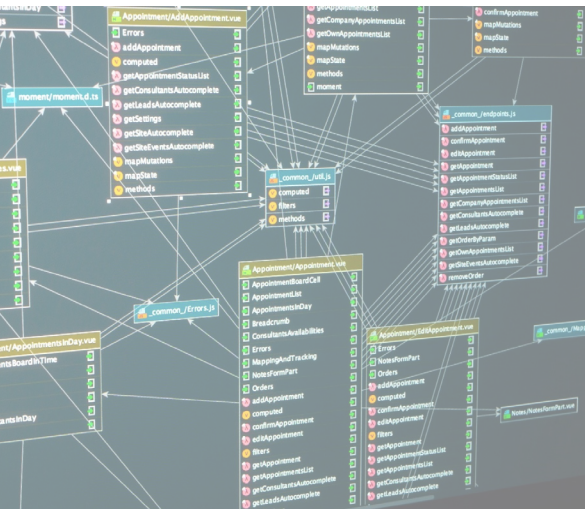
“A data model (or datamodel) is an abstract model that organizes elements of data and standardizes how they relate to one another and to properties of the real-world entities.”

OK, in defense of every IT professional out there, creating a data model is a challenging task. The size and complexity of your systems, both on and off the cloud, are all conspiring against you.

For any type of analytics project, basic or advanced, understanding your data structures is just the starting point. To create accurate decision support processes and systems, you need a deep understanding of these data structures and how your business processes are reflected in the data.

But data is now a big and difficult beast to tame. **The activities of any Data Scientist are split roughly into 90% data and 10% science tasks.** Extracting, cleaning, combining, and transforming data is the most resource-intensive challenge in many AI (and BI) initiatives.

This is one of the big paradoxes of data science. On the one hand, we evangelize the gospel of AI and, on the other hand, we seem to forget the fundamentals of good data management.



For any kind of advisor or consultant, your data is the starting point - that is whether you are handing out a simple piece of advice or planning a major digital transformation.

So, when embarking on any data science initiative, the first question to ask is: "Can you show me the data model?"

8 CONSIDERATIONS WHEN EXPLORING DATA SOURCES

Data is everywhere. Sometimes, it is structured and placed in ERP and CRM systems. But a lot of data is unstructured. This includes data from sensors, audio, video, or pictures, for example. Unstructured data must be converted into structured data so it can be used in your AI and machine learning models.

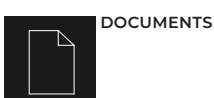
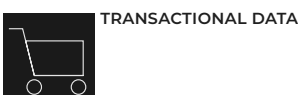
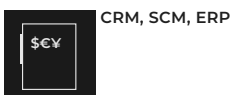
These are elements to consider when you start exploring the data landscape:

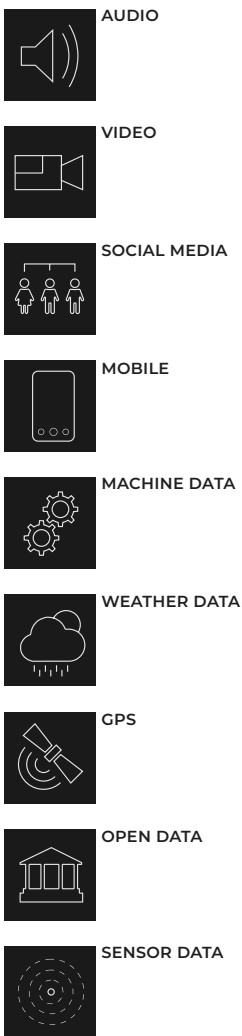
QUALITY Data quality is obviously an important parameter to decide which sources are usable. It is more difficult to determine the data quality when the source is external and not used. A good starting point is to gain an understanding of how the data has been gathered. You can also request a small sample of the data to assess its quality. Data quality is easier to determine when the source has been used. A new assessment of the quality might sometimes then still be required. This is because data quality is not one dimensional: the quality might be sufficient for one purpose, but insufficient for another.

BIAS Bias is a special case of data quality. Here, the data may not represent the target population. Models that are based on biased data lead to inaccuracies.

LINKING MULTIPLE SOURCES When different sources are combined, they must contain a common key or way to derive this key.

DATA TYPE EXAMPLES





COSTS Costs are incurred for both internal and external data sources. For external sources, there is the cost of purchasing the data, but there are also other (indirect) costs. These include the cost of extracting and preparing the data, for example, and long-term costs of maintaining this data.

STRUCTURED VS. UNSTRUCTURED To convert the unstructured data into structured data, advanced techniques are required, such as NLP. Such processes take time and effort, incurring extra costs and additional expertise.

DATA PRIVACY Data sources must be evaluated for personal data and other sensitive information. This requires additional work (e.g. pseudonymization and anonymization) and incurs extra costs.

METADATA Metadata is often described as “data about data”. Metadata can refer to the description of technical data such as tables, fields, data types. It can also be a description of the content to help you understand it or find the data source.

REAL TIME VS. HISTORICAL

The frequency of your data points, as well as the amount of historical data, are key to consider when you are investigating data sources. Sometimes real-time data is essential, often it is enough to have a good quality range of historical data points.

3 REASONS TO USE EXTERNAL DATA

With many data science projects, internal data will suffice. In some cases, it is necessary to acquire external data. This is because:

ONE The quality of the internal data is insufficient and must be corrected.

TWO You need to enrich internal data with information, which is not available in the internal data sources. For example, the internal data is enriched with external data where the original external data can be structured or unstructured.

THREE The organization requires information about non-customers or the wider competitive landscape. This information can facilitate targeted sales or marketing efforts or perform detailed market analysis.

ONE TECHNICAL METADATA

Technical metadata provides a foundation. It defines the objects and processes that make up the data warehouse or DataMart from a technical perspective.

TWO BUSINESS METADATA

Business metadata gives business context by describing the content of the data warehouse or DataMart in a user-accessible way.

THREE OPERATIONAL METADATA

Operational metadata describes the results of processing and accessing data (e.g. ETL). This data is valuable to improve your data processing step.

TYPES OF METADATA

PURPOSES OF METADATA

INFORMATION RETRIEVAL AND DISSEMINATION

Metadata makes information and resources discoverable.

OWNERSHIP AND RIGHTS MANAGEMENT

Metadata gives insights into the data ownership and the intellectual property rights attached to content.

MANAGING USERS

Metadata allows better management of users and improves data security.

PRESERVATION AND RETENTION

Metadata assists in securing important assets and resources.

RESOURCE DESCRIPTION

Metadata helps us understand the content of resources.

5 DIMENSIONS OF BIAS

“People generally see what they look for, and hear what they listen for.”

- To Kill a Mockingbird, Harper Lee

AI and Data Science adoption rates are soaring as more organizations pursue a data-driven agenda. *But have you stopped to consider the ethics of AI?* It is a complex undertaking, with many businesses struggling to apply ethical considerations in their day-to-day work.

‘**Bias**’ is a term that often gets thrown around, stalling data-driven initiatives, complicating project implementations and confusing stakeholders.

Bias is an important consideration when reviewing potential data sources. Bias is best described as “when data is not representative of the population of interest”. Bias can be introduced when the data is gathered. For example, the data may not have been randomly collected and certain groups will be over- or under-represented, compared to the population of interest. **This is selection bias.**

In the case of **surviving bias**, one group (the failures) are not taken into account at all. Models using biased data *result in inaccurate predictions*.

Wikipedia defines bias as:

“Bias is a disproportionate weight in favor of or against an idea or thing, usually in a way that is closed-minded, prejudicial, or unfair. Biases can be innate or learned. People may develop biases for or against an individual, a group, or a belief.[1] In science and engineering, a bias is a systematic error. Statistical bias results from an unfair sampling of a population, or from an estimation process that does not give accurate results on average.”

To estimate the extent to which a data source is biased, it is important to understand how the data has been collected. Bias can also be caused by prejudice or human bias. This type of bias is more difficult to detect.

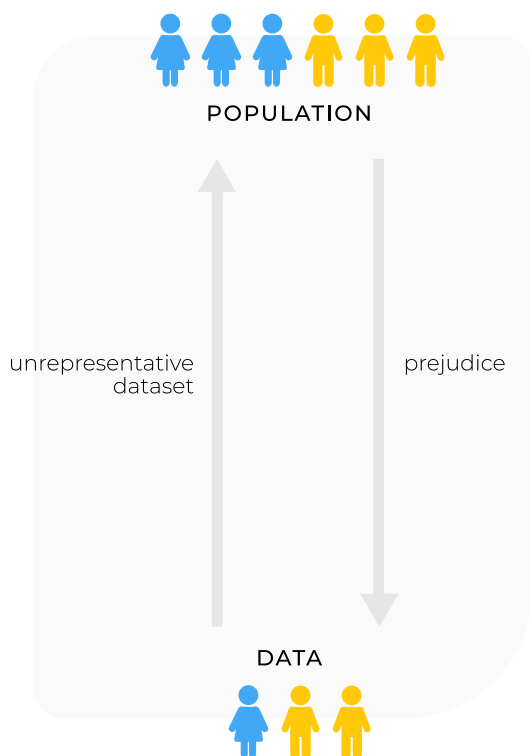
So, how can your organization achieve the right balance between the ethics of AI and achieving your business objectives? In this section, we'll focus on a few important elements of bias, explaining how your business can embrace AI and Data Science in an ethical manner for digital success.

Let's unpack the basics of that statement and explain why they are important from a Data Science context.

ONE - DATA CONDITIONS

Good quality data is not important for good AI, right? Wrong. Ask any experienced data scientist and they'll tell you the same thing: to make accurate (and therefore ethical) decisions based on your data, **the quality of your data is essential**.

Another misconception is that data is objective. Bias within data, however, can lead to incorrect conclusions or reinforce existing prejudices within your data. As such, the state of your data and your data management efforts are incredibly important. **Data privacy and data security are, therefore, vital boundary conditions for ethical data usage.**



From another perspective: you may think all databases are biased since, by their very nature, they are a selection of datasets (and cannot include everything). However, it is more important to understand the basics of your data sample, including how your selection of data (i.e. your database), and/or its sub-selections relate to one another.

TWO - MODEL CONDITIONS

You must take data bias and quality into account at the modeling stage. Bias can show up in the data and it can also be introduced when you select attributes for an AI model.

The **transparency of your model matters.** You must have justifiable reasons to opt for a more powerful but less transparent model. The good news is that transparency is not impossible to achieve. You can increase the transparency of, for example, a complicated neural network model by analyzing its operation or function, or by introducing human supervision.

Either way, **an AI model must be auditable** to ensure the output of the model or to ensure the steps leading to the model are replicable. To achieve this, an external company or your internal teams can conduct an audit.

THREE - DATA SCIENTIST CONDITIONS

Whatever project you are working on, it is unethical to act against your existing policies, rules or regulations.

This tenet also applies to data science. **But you must have a clear accountability agreement in place** to provide a consistent approach to ethics across your team. Your data scientists must also **work in a proportional and transparent manner**, adopting the least intrusive data strategies and clearly documenting your policies, rules and regulations.

FOUR - IMPACT ON STAKEHOLDERS

Your AI and Data Science project has a people impact on both your employees and the data owners.



You should **allow employees to provide feedback** across the project lifecycle, including after deployment. You should also **allow data owners to report any suspected issues**. You may also need to make special considerations around the impact of your data project on vulnerable groups.

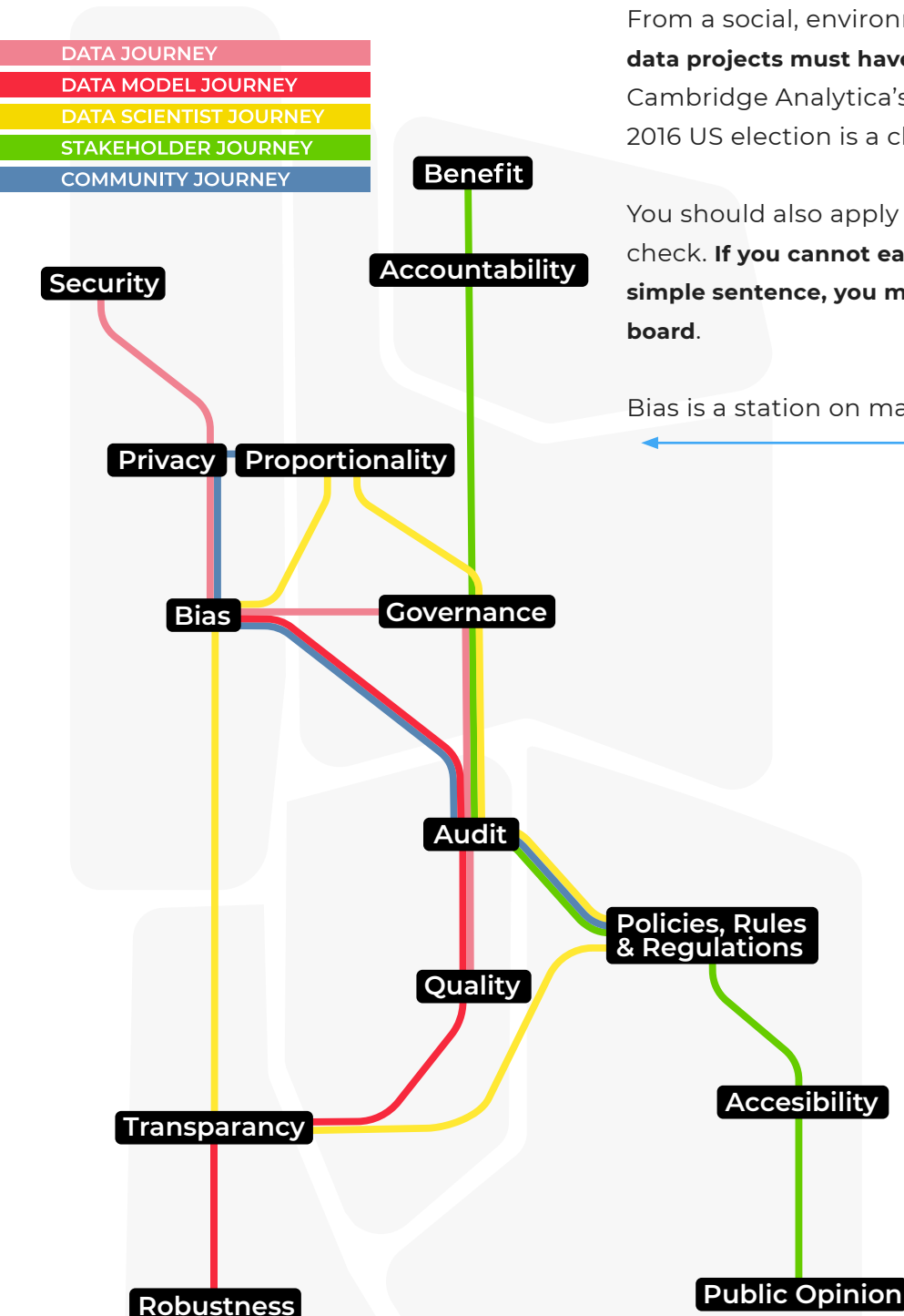
Accessibility is another consideration where people should have access to your AI products and services. This will safeguard certain groups within society, ensuring they are not discriminated against when your AI-based technologies are used in the wider world.

FIVE - IMPACT ON COMMUNITY

From a social, environmental, and democratic perspective, **data projects must have a positive impact on our community**. Cambridge Analytica's use of Facebook data during the 2016 US election is a clear example here of what not to do.

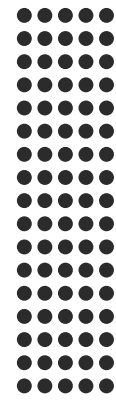
You should also apply one final consideration: the headline check. **If you cannot easily justify your data project in one simple sentence, you may want to leave it on the drawing board.**

Bias is a station on many different journeys.



CHAPTER THREE

Data Understanding



GOAL

To understand the data quality, verify how the data has been captured, prioritize sources, and anticipate possible data privacy challenges.

TOPICS

Data understanding (structure, fields, granularity), anonymization, restructuring of the data (aggregation, pivot), correcting bad quality data.

PITFALLS

Insufficient data quality checks, too much work to make necessary corrections to the data quality, where good descriptions or metadata are lacking.

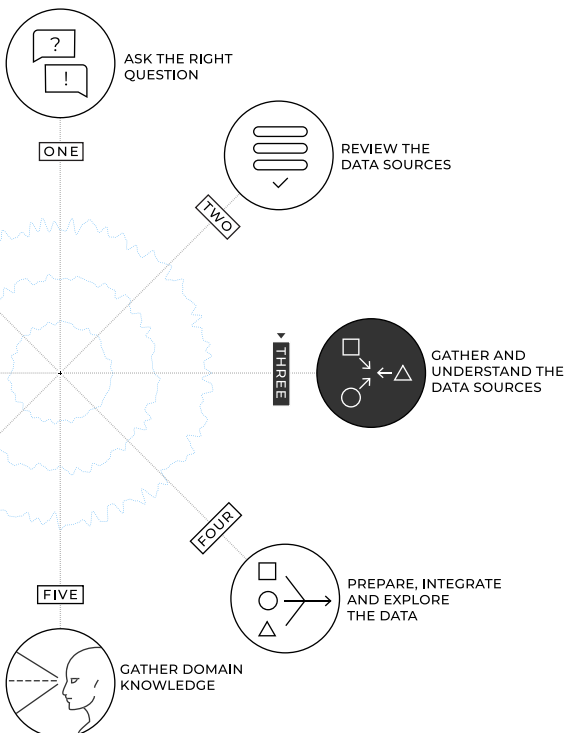
Once the data sources have been reviewed, you must select the best-qualified data sources. To achieve this, you must gather and assess these sources to understand the exact content, structure, and quality of the data. When the data sources contain personal data, the data must be anonymized.

Data quality and data management are not sexy topics. In almost all data science projects, this area is a hard nut to crack.

Gartner | “Poor data quality is responsible for an average of \$15 million per year in losses.”

FORRESTER | “One thirds of analysts spend more than 40 percent of their vetting and validating their analytics data before it can be used for decision-making.”

experian | “75% of businesses are wasting 14% of revenue due to poor quality”



6 THINGS ABOUT DATA QUALITY

The importance of good quality data is often overlooked. Good quality data requires thorough assessment, validation, and correction by expert analysts and Data Scientists. This is an important step because the quality of your data has a direct impact on the accuracy of your resulting business decisions. **Simply put, if your data quality is poor, your business will make the wrong decisions and, ultimately, see revenue losses.**

Excessive data correction also takes time. As a result, your Data Scientists are focused on data correction, instead of value-added tasks.



Many people take a one-dimensional view of data quality. But data quality has six different dimensions:

ONE - COMPLETENESS: are there any missing values?

TWO - VALIDITY: do the values correspond to the expected format?

THREE - CORRECTNESS: are the entered values correct?

FOUR - CONSISTENCY: is the data the same across all systems?

FIVE - UNIQUENESS: does the database contain duplicates?

SIX - TIMELINESS: is the data refreshed in time?

Some data quality issues have more of an impact than others. Validity issues can be corrected by converting the data into the appropriate format. Uniqueness issues are also easy to resolve, although it is advisable to find the underlying cause for the duplication. The other issues are more complicated and time-consuming to resolve.

Data quality is also context dependent. The quality of your data might be sufficient for one purpose but inadequate for another.

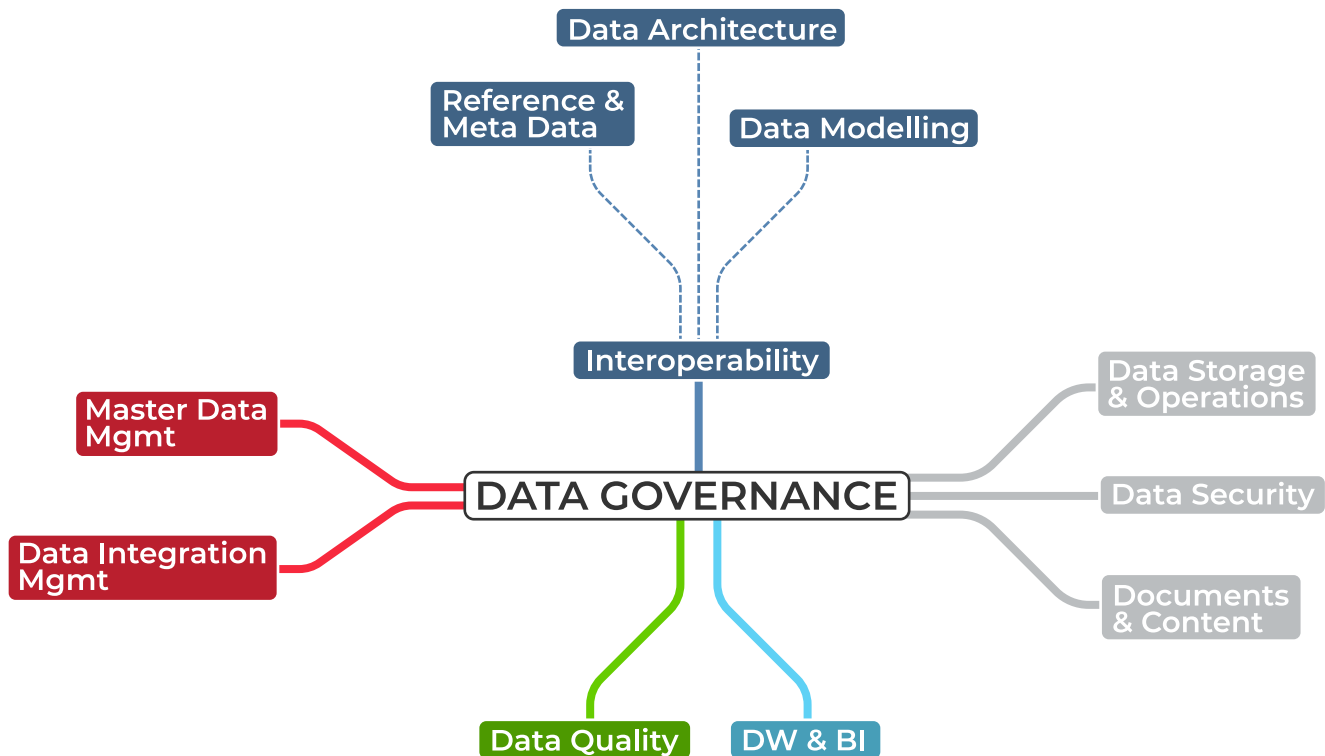
You can improve data quality when the source is used or refreshed. This is not a comprehensive approach - **it is better to address the problems in a structural way, using a combination of data management and data governance.** It's a route not often taken, since it digresses from the objective of the exercise and doesn't seem to deliver immediate value. However, in the long run, it pays to feedback the kind of data issues you have encountered into the operational processes and how adjustments can improve your data quality.

10 THINGS ABOUT DATA MANAGEMENT

Data management and data governance are intrinsically linked. *For data management*, you manage your data to achieve specific goals. These goals may include improving the usability of the data, making the data available and managing data security.



For data governance, you must ensure the data is managed appropriately across your people, processes and technologies. To achieve this, you must include your senior management, who can act as an executive sponsor and data owner. Senior management can give the mandate to staff to operationally manage data governance, while the responsibility remains at the management level. The Data Management of Body of Knowledge (DMBOK) from the Data Management Association (DAMA) is a popular data management and governance framework.



6 ELEMENTS OF GDPR

DMBOK covers data privacy from a data management perspective. The legal side of data protection and privacy is in the EU regulated through the General Data Protection Regulation (GDPR). *The primary aim of the GDPR is to give individuals control over their personal data and to simplify the regulatory environment for international business by unifying the regulation within the EU.* According to the GDPR, there are six different legal grounds where it is lawful to process personal data:

ONE - CONSENT Consent means that a person has given consent for a personal data processing activity for one or more specific purposes. It is important that the purpose for which consent is given is fully clear to the person.

TWO - LEGITIMATE INTEREST Where data processing is necessary for the purposes of the legitimate interests pursued by the controller or by a third party. This could occur, for example, when a company stores customer details in their system.

THREE - PUBLIC INTEREST In the GDPR, public interest is described as “processing is necessary for the performance of a task carried out in the public interest or in the exercise of official authority vested in the controller”.

FOUR - CONTRACTUAL NECESSITY To enter in any contractual relationship, personal data must be provided and processed by the individual.

FIVE - LEGAL OBLIGATIONS If a controller has a legal duty where particular personal data requires processing, then that processing is permitted.

SIX - VITAL INTEREST Life-threatening circumstances also provide legal ground for processing personal data.

It is important to note that there is no lawful basis if you can reasonably achieve the same purpose without the processing of **personal data**.

“Personal data’ means any information relating to an identified or identifiable natural person (“data subject”); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person.”

- GDPR Article 4

What is personal data? Article 4 of the GDPR provides the official definition (see above). In summary, all data that identifies a natural person or makes a person identifiable is considered as personal data.

PERSONAL DATA

YES

Social Security Number

Name

Email address

Data about the only female person in a postcode area

IP-adress

Photos

Pseudonymised Data

NO

Anonymised personal data (not relatable)

Data about companies or organizations

Aggregated personal data

Deceased persons

Generic email address

So, data including social security numbers, names, email addresses are all personal data. Data not considered to be personal data includes company data, aggregated personal data, a generic email address or data about deceased persons.

The GDPR takes the protection of personal data very seriously. However, you can still analyze and store data sources that originally contain personal data. A simple technique is to store the data on a higher aggregation level. You could aggregate data about consumers or households at a postcode level, for example. GDPR does not consider aggregated data as personal, as long as individuals cannot be traced.

5 THINGS ABOUT ANONYMIZATION

When aggregation is not a viable option, anonymization is another option. With anonymization, all personally identifiable information is stripped from the data source. There are different anonymization techniques:

GENERALIZATION Replace exact values with a more general value. For example, replacing age with an age category or address information with the province.

SUPPRESSION (OR MASKING) Personal data is deleted or altered in such a way that the individual can no longer be identified.

DATA SWAPPING (OR PERMUTATION) A technique used to rearrange the dataset attribute values so they do not correspond with the original records.

PERTURBATION This technique modifies the original dataset slightly by applying techniques that round numbers and adding random noise. The range of values must be proportional to the perturbation.

SYNTHETIC DATA Synthetic data is algorithmically manufactured data. This synthetic data is used to create artificial datasets instead of altering the original dataset. The process involves creating statistical models based on patterns found in the original dataset.



According to GDPR, data is only truly anonymized when the anonymization is irreversible. Anonymization has also some disadvantages. It not only limits your ability to derive value and insights from your data, but you can no longer enrich the data by linking it through a personal identifier.

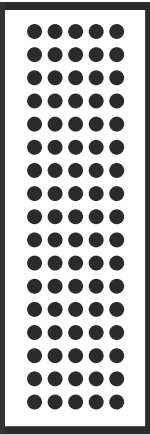
These disadvantages do not exist with pseudonymization. Pseudonymization is a method that replaces private identifiers with fake identifiers or pseudonyms. Pseudonymization preserves statistical accuracy and data integrity and allows for linking with other databases that have used the same algorithm for pseudonymization.

Although data privacy is protected with pseudonymization, GDPR still considers pseudonymized data as personal data since it is possible to re-identify a person.



CHAPTER FOUR

Mining & Combining



GOAL

To capture patterns, correlations, and anomalies.

TOPICS

Enrichment and integration of data, data validation (micro/macro), bias check, data privacy after integration, explorative analyses, patterns and correlations.

PITFALLS

Wrong joins, databases on wrong level, no bias or data privacy check, additional DataMarts required to answer all questions within scope.

Once the relevant data sources are gathered, the preparation, integration and exploration of the data can begin. This is the most time-consuming step in the framework. According to Forrester, one-third of analysts spend more than 40% of their time vetting and validating data.

When preparing your data, the first step is to assess the content of your data sources. Namely, do the files contain only quantitative data or qualitative data as well? If the data is quantitative, is the data discrete or continuous?

3 DATA SCALES

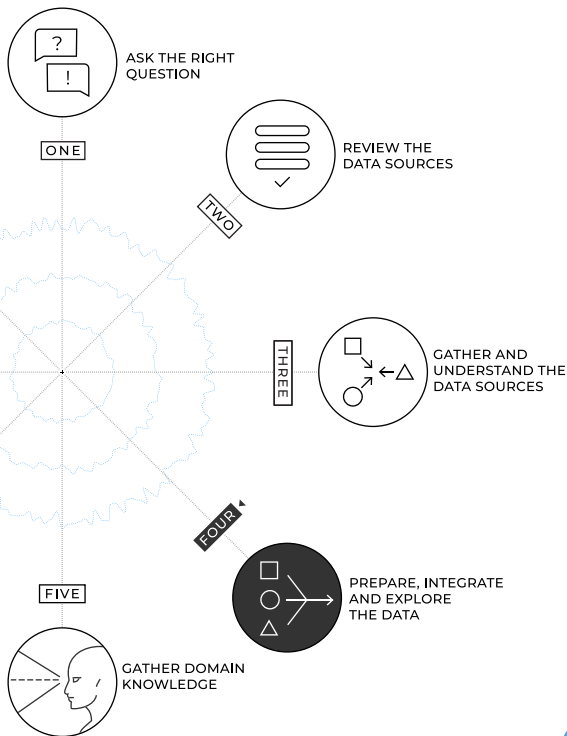
Quantitative data is scaled in one of three ways, depending on whether that data is discrete or continuous:

ONE - INTERVAL With an interval scale, the data has a standardized order. So, the difference between each level on the scale is the same. There is no zero point.

TWO - RATIO An absolute and meaningful zero is a unique feature of the ratio scale. Ratio scale data can be multiplied or divided.

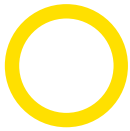
THREE - CIRCULAR Whether it is time, days of the week, months of years, data linked to duration has a circular scale. Circular scale data is a special type of interval scale data.

After preparing the data, the next step is **integration**. A common way to integrate multiple sources is by **linking** them. To link the files, the sources need at least one field in common. *This field is called a key*. The presence of one key is often sufficient but sometimes, when the linking of two files, multiple keys are required.



4 TYPES OF JOINS – AND UNIONS

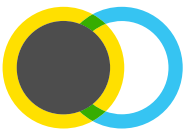
In database terms, the link between two files is called a “join”. There are four types of joins:



File1		
Name	Car	Fuel
Olivia	Opel	Diesel
George	Mercedes	Diesel
WilliamP	BMW	atrol
MiaP	Kia	atrol



File2	
Name	Commuting
Olivia	10 km
George	15 km
Jane	20 km



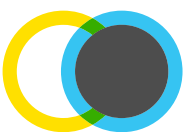
Left Join			
Name	Car	Fuel	Commuting
Olivia	Opel	Diesel	10 km
George	Mercedes	Diesel	15 km
William	BMW	Patrol	null
Mia	Kia	Patrol	null

ONE - LEFT JOIN A left join returns all records from the left table (the master) and the matched records from the right table (the child). The result is NULL from the right side if there is no match.



Inner Join			
Name	Car	Fuel	Commuting
Olivia	Opel	Diesel	10 km
George	Mercedes	Diesel	15 km

TWO - RIGHT JOIN This is the same as a left join, but the master and child are reversed.



Right Join			
Name	Commuting	Car	Fuel
Olivia	10 km	Opel	Diesel
George	15 km	Mercedes	Diesel
Jane	20 km	null	null

THREE - INNER JOIN All records from both tables are returned where the key from one table matches the key from the other table.



Outer Join			
Name	Car	Fuel	Commuting
Olivia	Opel	Diesel	10 km
George	Mercedes	Diesel	15 km
William	BMW	Patrol	null
Mia	Kia	Patrol	null
Jane	null	null	20 km

FOUR - OUTER JOIN An outer join returns all records in both tables, whether they are matched or not. If there is no match, the result is NULL from either the right or left side.

Join Example

Key	Column 1	Column 2	Column 3	Column 4
Row 1	Data	Data	Data	Data
Row 2	Data	Data	Data	Data
Row 3	Data	Data	Data	Data
Row 4	Data	Data	Data	Data

new columns

Union Example

Key	Column 1	Column 2
Row 1	Data	Data
Row 2	Data	Data
Row 3	Data	Data
Row 4	Data	Data
Row 5	Data	Data
Row 6	Data	Data
Row 7	Data	Data

new rows

Another way to integrate data is to use a “**union**”. A union is a method for combining data by appending rows of one table onto another table.

If two tables are “unioned” together, then the data from the first table is in one set of rows, and the data from the second table in another set. The rows contain the same results.

So, what’s the difference between a union and join? *With a union, you are adding records. With a join, you are adding columns.*

When tables have the same number of fields and the same field names, then a union generates an integrated table, which matches the layout of the original tables. When one or more of the tables contain extra fields or field names that are not identical, then a union produces a table containing a column for each unique field name. The field name is “null” for records belonging to the original table that do not contain that field name.

In most cases, the choice between a union or join is self-evident. You can also use these three factors as guidance:

ONE A join adds new fields, whereas a union adds additional records.

TWO When the file structure is different, a join is the logical choice. With a union, the structure of the files is similar or the same.

THREE A join can only be performed when a key is present in the table, which needs to be integrated. With a union, this is not required.

There are cases, however, when the choice is still not immediately obvious. Here, you must consider the pros and cons of both options. A factor to take into account is the impact on your KPIs.

5 LEVELS OF DATA VALIDATION

When the preparation and integration of the data are complete, the data is explored. However, before the exploration can begin, there are different kinds of validation to perform. Some of these validations typically occur during the integration and transformation of the data, but some are done after the completion of these two steps. There are five different validations:

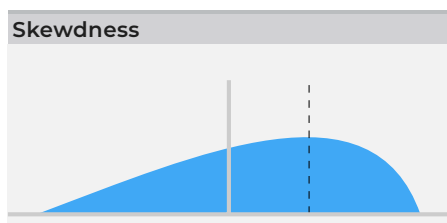


EXPLORATORY DATA ANALYSES (EDA)

Once the data is prepared, it is ready to be explored. *The aim is to gain insights into the distribution of the available fields and to detect trends in order to discover relationships between fields.* The **univariate analysis** provides many insights into the distribution of fields: measurements like median, mode, skewness, and the existence of outliers. Outliers are important and can have a major impact on the analysis. So, it is vital to correct or delete them.

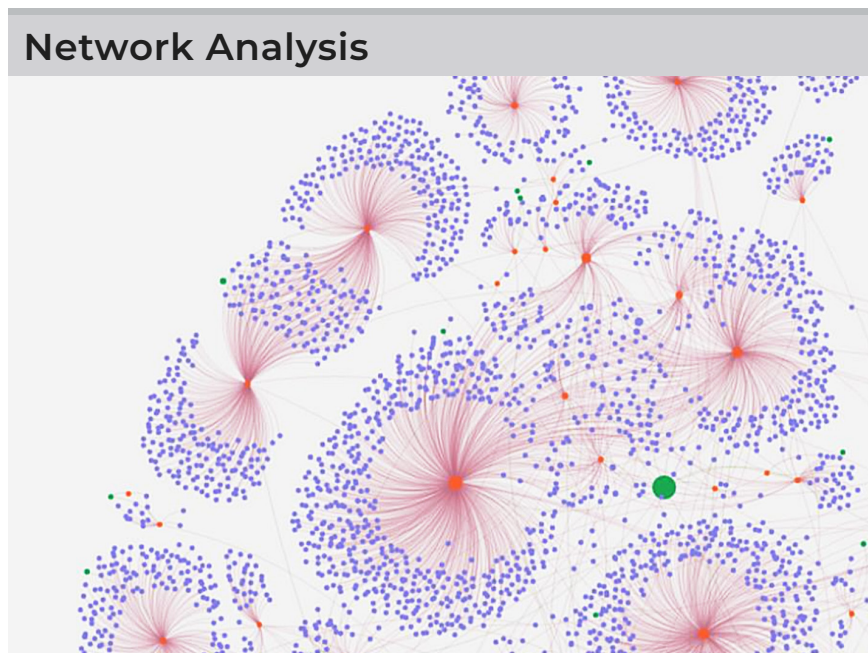
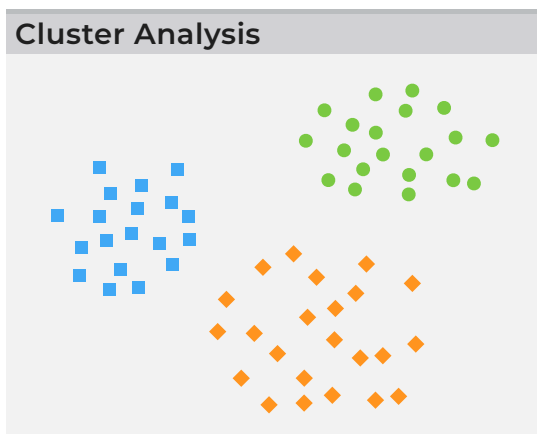
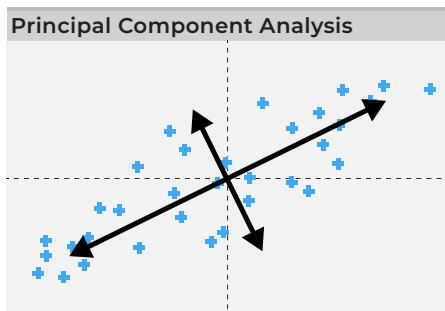
The **multivariate analysis** provides insights into the relationships between two or more fields. In the case of the scatter plot, it can also be helpful to detect outliers. The **correlation matrix** is typically a good first step to find which fields are correlated with the targeted variable. However, correlation does not automatically indicate causality between two fields.

There is also a more advanced exploratory analysis. The two most common types are **principal component analysis (PCA)** and **cluster analysis**. PCA identifies the most important fields for a model. This is especially helpful when there are many fields available. Cluster analysis detects which clusters or distinct groups within the data are distinguishable. It builds an understanding of the hidden relationships in the data and is sometimes the goal of a data science project.



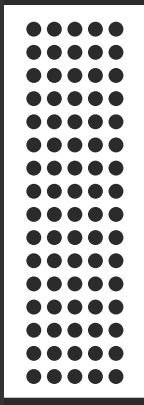
Correlation Matrix

	1	2	3	4
1	1.0	-0.4	0.2	-0.1
2	-0.4	1.0	0.1	0.0
3	0.2	0.1	1.0	0.1
4	-0.1	0.0	0.1	1.0



CHAPTER FIVE

Deep Into Domain



GOAL

Obtain solid knowledge of the business domain from domain experts.

TOPICS

Validation or explanation of results from explorative phase, investigate further the possibility of bias or existing prejudice, understand whether certain topics or insights are sensitive.

PITFALLS

A step is skipped, or not enough time is spent. Limited access to domain experts.

While domain experts interact throughout a data project, this collaboration is essential at this stage. During the exploration of the data, the Data Scientist can familiarize themselves with the data and detect trends, abnormalities, patterns and correlations. *These insights now need to be explained or put into context by domain experts.*

Domain expertise is also required during other stages of the 8-step process, including step 1 (asking questions), step 2 (data landscaping), and step 6 (modeling).

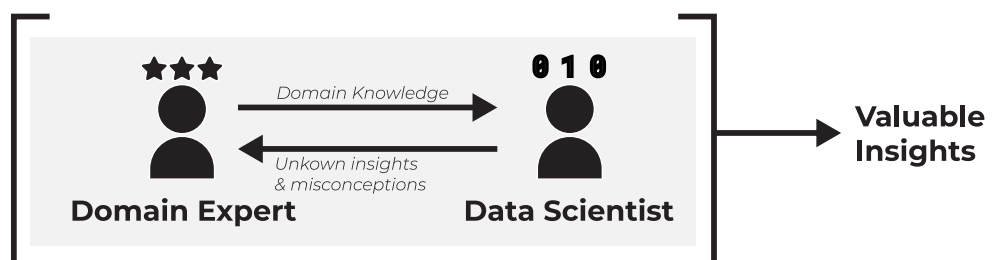
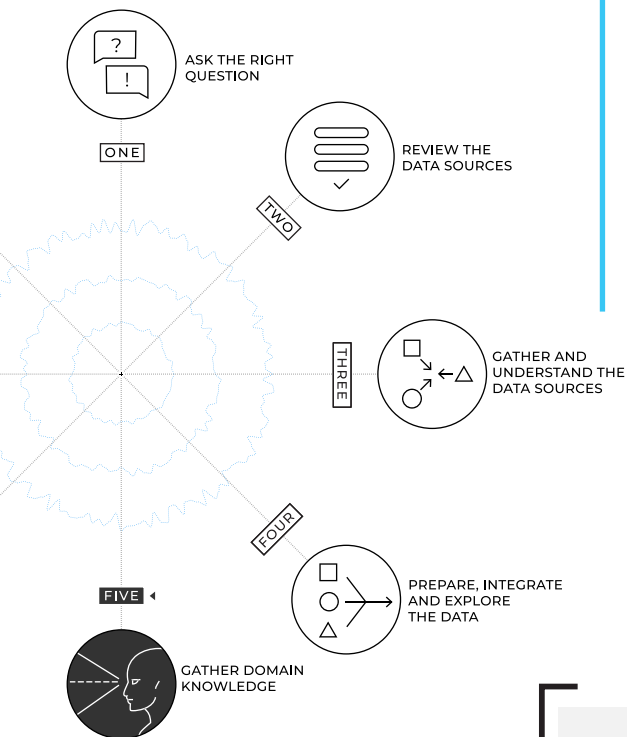
“What is domain knowledge?”

Is knowledge of a specific, specialized discipline or field, in contrast to general knowledge, or domain-independent knowledge.

The term is often used in reference to a more general discipline, as, for example, in describing a software engineer whoe has general knowledge of programming, as well as domain knowledge about the pharmaceutical industry. People whoe have domain knowledge, are often considered specialists or experts in the field.”

- Wikipedia

Thus, domain knowledge is knowledge about a specific, specialized discipline or field related to a certain (business) domain that somebody has acquired through years of experience.



3 TYPES OF DOMAIN KNOWLEDGE

There are three types of domain knowledge:

ONE - CONTEXT OF THE PROBLEM. Domain knowledge builds a further understanding of the problem. The context of the problem is discussed during the first step of the 8-step model (ask the right questions). But, as more insights into the problem are gained, domain experts can provide further insights. Previous attempts to solve the problem can also provide further insights into the context of the problem.

TWO - SPECIALIZED INFORMATION OR EXPERTISE. After the exploration of the data, a domain expert can interpret the resulting insights or validate the (tentative) conclusions for the Data Scientist. The Data Scientist can also review the proposed features with the domain expert. Once a model is created, the domain expert can help interpret the results or finetune the model and then validate the model to assess its effect on vulnerable groups.

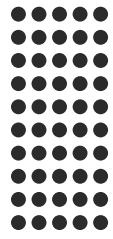
THREE - KNOWLEDGE OF DATA COLLECTION. Domain experts are also a useful resource to identify potential data sources. As a result, you can get a better understanding of how data was collected or identify data quality issues.

The interaction between the Data Scientist and the domain expert is ideally a two-way street. The Data Scientist can tap into the knowledge of the domain expert. But the Data Scientist can also reveal their data-based insights to the domain expert.



CHAPTER SIX

Modeling



GOAL

To build and validate a data model.

TOPICS

Creation of KPIs, selection of AI model based on performance and ethical factors (e.g. law and regulation, transparency).

PITFALLS

Ethical considerations not taken in account during model selection, the performance of the model is not regularly tested, no bias check of the model.

'Model the data' is the main aim of any Data Science project. Now, we can finally put all the work in the previous steps to good use. You may assume this step is the most time-consuming but, for most projects, this is not the case. *That is because Data Science is 90% data preparation and 10% science.*

There are many different ways to model the data. The most basic form is reporting. **Reporting** is a form of diagnostic analytics, where we observe the past and try to explain it. In other words, we evaluate past decisions.

Monitoring is another method, bordering both diagnostic and predictive analytics. It is more in line with predictive analytics, detecting any possible deviations at an early stage to provide decision support. By understanding better what is likely to happen, a company can mitigate the consequences of the forecasted event or profit from them.

Forecasting and **predictive** models are two other forms of predictive analytics. Predictive models sit between predictive and prescriptive analytics since a predictive model implicitly prescribes what to do. With prescriptive analytics, it is possible to automate your decisions, but this is also a complex method to implement. However, such complex analytics typically provide additional business benefits. As such, the prescriptive model is widely regarded as the ultimate form of advanced analytics.

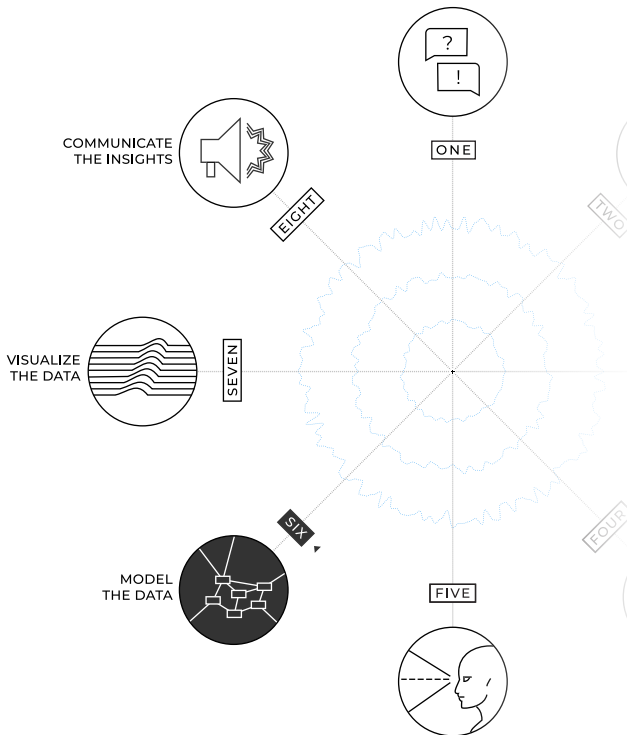
ASPECTS OF KEY PERFORMANCE INDICATORS (KPI)

ONE Variable to analyze the performance of organizations/projects.

TWO Quantitative and measuring an objective or a Critical Success Factor.

THREE Internal and external effects to perform better and to deploy effective means/policy.

FOUR Factors which can be influenced by the organization.



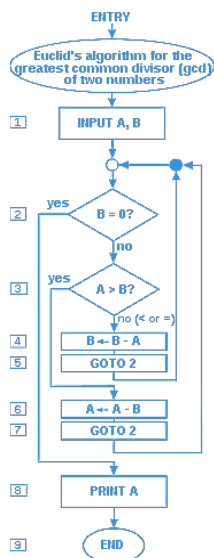
The most rudimentary form of a data model is a key performance indicator (KPI). KPIs are used to analyze the performance of organizations or projects, by measuring an objective or a critical success factor. A KPI measures an internal or external factor that can be influenced by the organization. In other words, a KPI is actionable

WHY ARE KPI'S USEFUL?

To make objectives measurable,
To make targeted data analyses,
Draw conclusions based on data,
And used as a basis for making better choices for policy/ action/prevention.

KPIs make an objective (or the factors to reach the objective) measurable. They enable organizations to perform root cause analysis, allowing you to understand why a certain target was (or was not) reached. As such, they provide an excellent base to define and track policies and actions.

3 CONDITIONS FOR AN ALGORITHM



Informally, an algorithm is any well-defined computational procedure that takes some value, or set of values, as input and produces some value, or set of values, as output. An algorithm is thus a sequence of computational steps that transform the input into the output.

- Thomas H. Cormen, Chales E. Leiserson (2009), **Introduction to Algorithms 3rd edition.**

Advanced models, including predictive and prescriptive models, make use of algorithms. These algorithms vary between a simple “if... then... else” statement and incredibly complex sequences. But, in all cases, the operation of an algorithm is always the same.

As stated by Thomas H. Cormen and Charles E. Leieron, an algorithm is a sequence of computational steps that transform the input into an output. This also means there is nothing “smart” about algorithms. They just follow certain steps without any real intelligence. But an



algorithm can be so complex that its operation is totally incomprehensible to a human being.

An algorithm must satisfy the following conditions. It should:

BE FINITE An algorithm that never ends is useless, as it will never solve the problem.

HAVE WELL-DEFINED INSTRUCTIONS Each step of the algorithm must be precisely defined.

BE EFFECTIVE The algorithm should provide the desired output.

3 TYPES OF ANALYTICS

Different techniques that use algorithms. These include:

TRANSFORMING ANALYTICS These techniques are used in step 4 of the 8-step model, namely - prepare, integrate, and explore the data. Examples include *data aggregation*, *enrichment*, and *processing techniques* such as data cleaning, preparation, and separation.

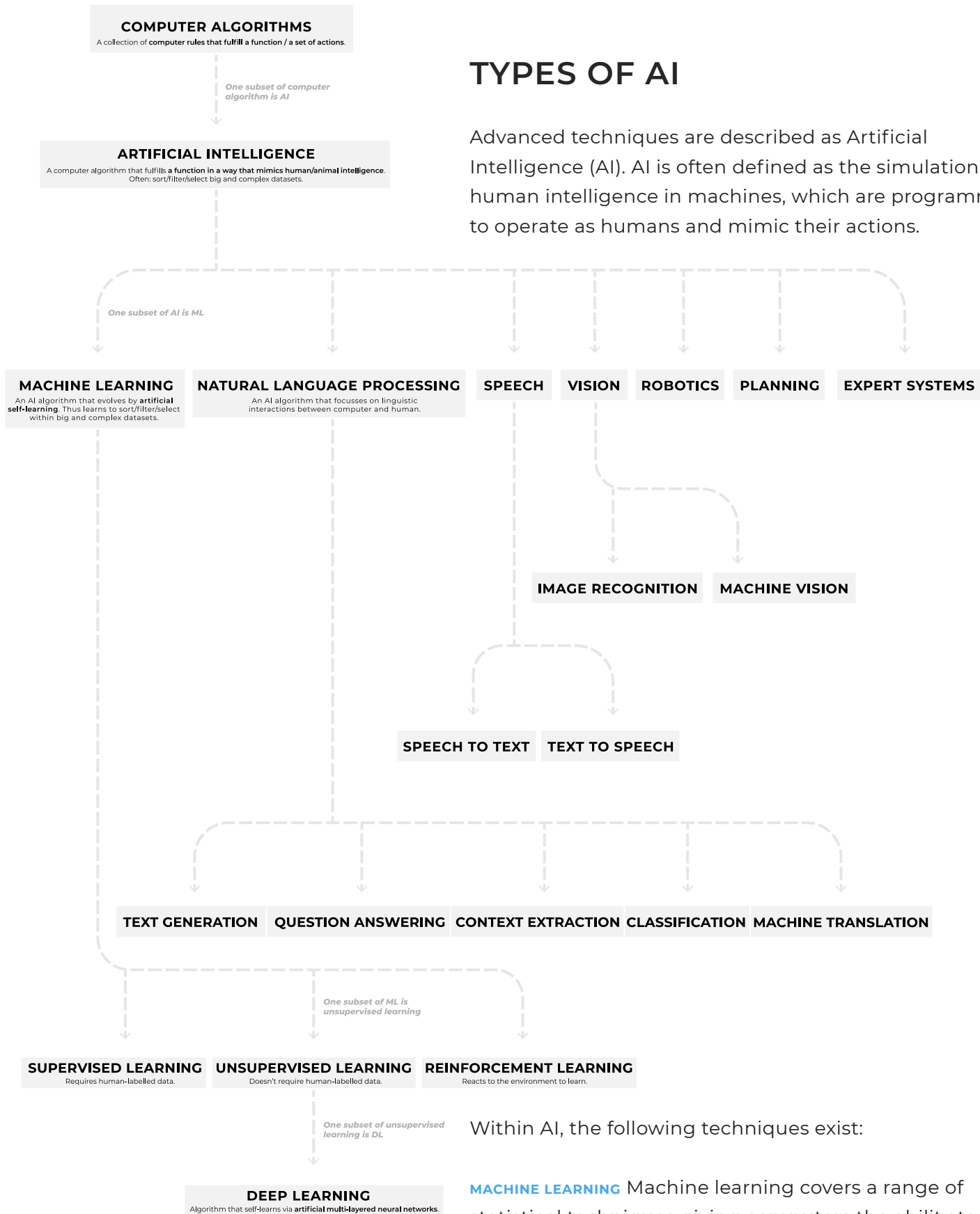
LEARNING ANALYTICS These analytics provide insights into relationships and can also classify objects into groups. *Regression*, *clustering*, *classification*, and *recommendation* are all examples of learning analytics techniques.

PREDICTIVE ANALYTICS These include *simulation* or *optimization* techniques. With simulation, a simplified representation of reality is created. This may be a process or a system, for example. The model used is either predictive or prescriptive in nature. But both try to predict the future using the patterns in the data. Optimization simply refers to operation research techniques.



TYPES OF AI

Advanced techniques are described as Artificial Intelligence (AI). AI is often defined as the simulation of human intelligence in machines, which are programmed to operate as humans and mimic their actions.



Within AI, the following techniques exist:

MACHINE LEARNING Machine learning covers a range of statistical techniques giving computers the ability to learn. In other words, they progressively improve their capacity to execute a task. Machine learning can also be split into Deep Learning, unsupervised and supervised learning.

NATURAL LANGUAGE PROCESSING Natural Language Processing (NLP) is an area of AI concerned with the interactions between computers and human (natural) languages. The field of NLP includes text generation (e.g. a machine writing a book), question answering (e.g. chatbots), context extraction (e.g. anonymization and summarization), classification (e.g. sentiment analysis) and machine translation (e.g. Google Translate).

EXPERT SYSTEMS These represent a simple AI system, typically consisting of lists of “if... then” statements and other such associations, which are written in a human-like language.

SPEECH Speech is linked to NLP, but it is a unique technique within AI. Compared to NLP, the words are converted from speech to text or from text to speech.

VISION This technique deals with the way computers see and understand digital images and videos including, for example, facial recognition.

PLANNING Planning is a branch of AI concerning strategies and action sequences. Self-driving cars are an example of planning.

ROBOTICS Robotics is also known as Robotic Process Automation (RPA). Robotics automates the manual tasks usually done by a human.

3 TYPES OF MACHINE LEARNING

There are three types of machine learning algorithms:

SUPERVISED In supervised learning, the algorithm uses a labeled dataset. This label is the outcome the algorithm needs to predict. A part of the labeled data is then used to train a model in an iterative way; every step the model compares the label from the database with the outcome of the model and then readjusts it until the model has been optimized. The model can then be applied to the other part of the labeled dataset to validate and measure the accuracy of the model. When the accuracy is sufficient the model can then be used to predict the label for the



unlabeled data. Classification and regression are examples of supervised learning.

UNSUPERVISED In unsupervised learning, the algorithm uses an unlabeled dataset. The algorithm is searching for unknown patterns and relationship in the data. It is important for the Data Scientist to research whether the outcome of the model is useful and/or actionable. There are different types of unsupervised learning:

- **Clustering** Creating groups in such a way that objects in the same group (called a cluster) are more similar to each other than to those in other groups.
- **Anomaly Detection** Identification of rare events or observations that raise suspicions by differing significantly from the majority of the data. Detecting bank fraud is an example of this type of unsupervised learning.
- **Association** Discovering interesting relations between variables in large databases. The aim is to predict what other attributes are commonly associated with a couple of key variables. Recommendations to add products to a shopping basket is one example of an association model.
- **Autoencoders** The aim of autoencoders is to remove “noise” from visual data like images, video, or medical scans. This is done by learning a representation (encoding) for the data set and then generating the reduced encoding, which is a close representation to its original input.

REINFORCEMENT

With reinforcement learning, the agent relies both on learning from past feedback and exploration of new tactics that may present a larger payoff. This involves a long-term strategy where the agent tries to maximize the cumulative reward. This is an iterative process where, the more rounds of feedback, the better the agent’s strategy becomes. This technique is especially useful for training robots, which make a series of decisions during tasks like steering an autonomous vehicle or managing a warehouse inventory.

CHAPTER SEVEN

Feed The Eyes



GOAL

To visualize business dashboards.

TOPICS

Importance of strong visualization, the Minard diagram, variables to consider, principles for design.

PITFALLS

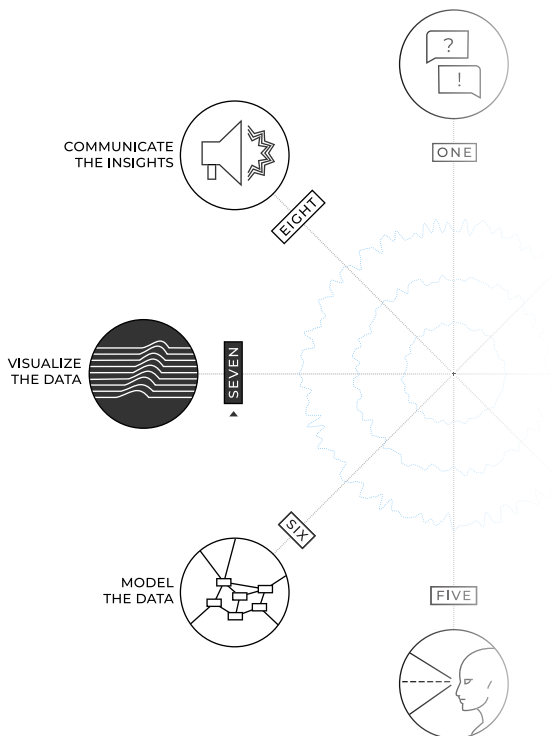
Understanding audiences, 9 mistakes to avoid.

Why do we visualize data? The most basic answer is that it's very hard to read data encoded in a database or dataset. Visualization makes use of the basics of human perception to intuitively present the data.

Even if you never made a graph, you've probably already created a data visualization. A mere table is a data visualization; it's data visually placed in rows and columns, helping the viewer read the data.

We also use data visualizations to find different points-of-view that help us interpret the data. A table is one example; although the way in which it shows the data often makes it difficult to identify trends and make comparisons. You need to look at the values in each cell, store them in your short-term memory, execute an analysis, and correctly form a conclusion. Many people would need to turn to pen and paper to complete these tasks. Data visualization can offer many different points-of-view. Some of these views could be tables and others could be charts and plots. All of them have different up- and down-sides and could help the user with their needs.

Reading and understanding data is one of the most important parts of data visualizations, but there are more. For instance, visualizations must also be memorable, convincing, entertaining, sleek, or whatever else is important to you and your objectives.



Turin Papyrus Map. Fragment 1 of 5



DATA VISUALIZATION ORIGINS

There are various theories about the origin of the first data visualization. But because these origins exist before humanity began to document their activities, it is impossible to pin-point an exact year. The **Turin Papyrus Map** is an ancient Egyptian map dating back to 1150 BC. It is considered to be the oldest found map and the first documentation of data visualization. But the **Tally Chart** would probably win as the “oldest data visualization” as it dates back to between 33000 BC and 23000 BC. This just goes to show how naturally this craft came to humanity.

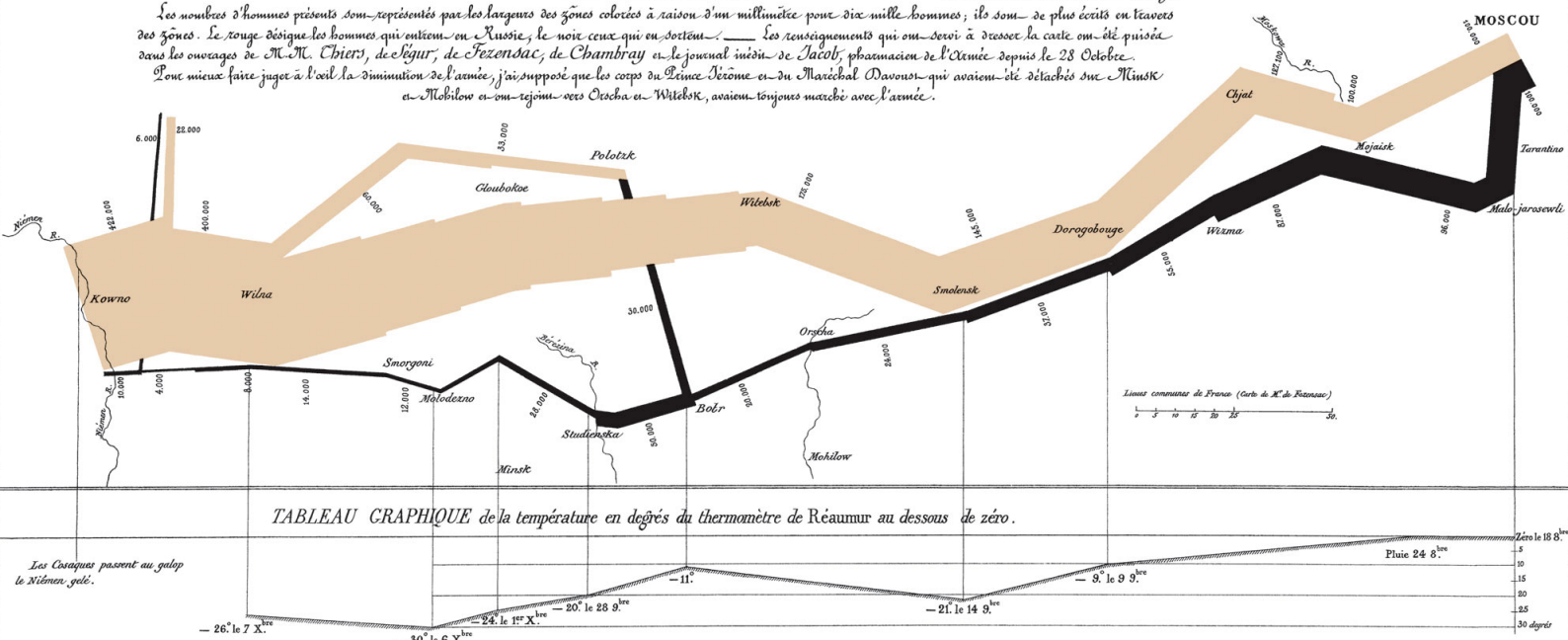
Data Visualization has experienced a few big growth spurts throughout history. One of these was between 1860 and 1890 when statisticians, governments, and municipal authorities were eager to discover the possibilities and problems of graphic representation. Visualization was widely adopted, and graphics were officially recognized by government agencies, becoming a feature of official publications.

One famous example is the **Minard diagram** created by Charles Joseph Minard, which was published in 1869. It is a precursor of the Sankey diagram, which depicts the course of Napoleon’s Russian campaign in 1812. It combines several aspects of the campaign into one diagram, combining the size and the direction of the army’s movements with information including the geography.

Carte Figurative des pertes successives en hommes de l'Armée Française dans la campagne de Russie 1812-1813.

Dessinée par M. Minard, Inspecteur Général des Ponts et Chaussées en retraite. Paris, le 20 Novembre 1869.

Les nombres d'hommes présents sont représentés par les largeurs des zones colorées à raison d'un millimètre pour dix mille hommes; ils sont de plus écrits en lettres des zones. Le rouge désigne les hommes qui entrent en Russie, le noir ceux qui en sortent. Les renseignements qui ont servi à dresser la carte ont été puisés dans les ouvrages de M. M. Chiers, de Léger, de Fezensac, de Chambray et le journal inédit de Jacob, pharmacien de l'Armée depuis le 28 Octobre. Pour mieux faire juger à l'œil la diminution de l'armée, j'ai supposé que les corps du Prince Jérôme et du Maréchal Davout qui avaient été détachés sur Minsk et Mohilow et ont rejoint vers Orscha et Witebsk, avaient toujours marché avec l'armée.

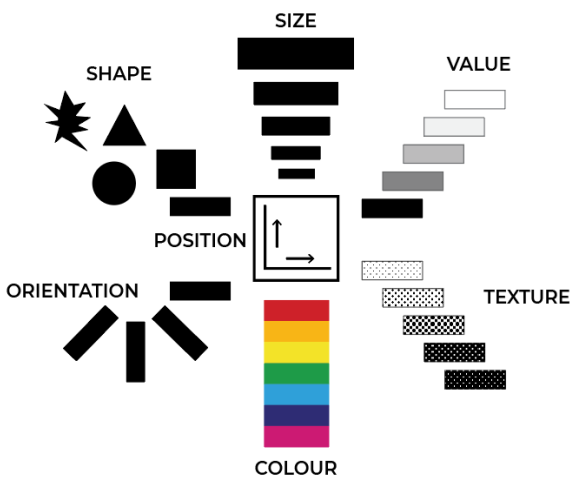


VISUAL VARIABLES

Visual variables are what a data visualization consists of. They are like the flavors of a dish, the words of a sentence, or the musical components of a song (tempo, notes, rhythm, ...). In 1967, the variables were first systematized by the French cartographer Jacques Bertin. In his book *Sémiologie Graphique* he teases out different components: position, size, shape, value, hue, orientation, and texture.

Let's say you have a set of these numbers: [3, 10, 14, 25, 30, 50, 87, 95, 100]. You could create visualizations with each visual variable.

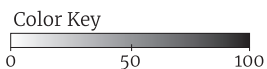
Since 1967, researchers have added variables, like angle, area, slope, volume, connection, movement (when data is animated), transparency, and interactivity. To be fair to Bertin, some of these, like movement and interactivity, were not even possible when he first thought of the system.



Shapes

Values: 3, 10, 14, 25, 30, 50, 87, 95, 100

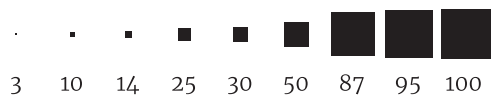
Color Value



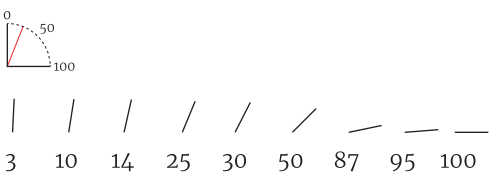
Numbers represented in color:



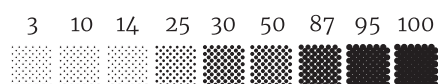
Size



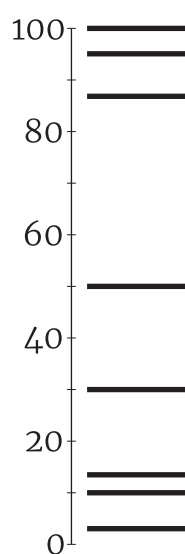
Orientation



Texture



Positions

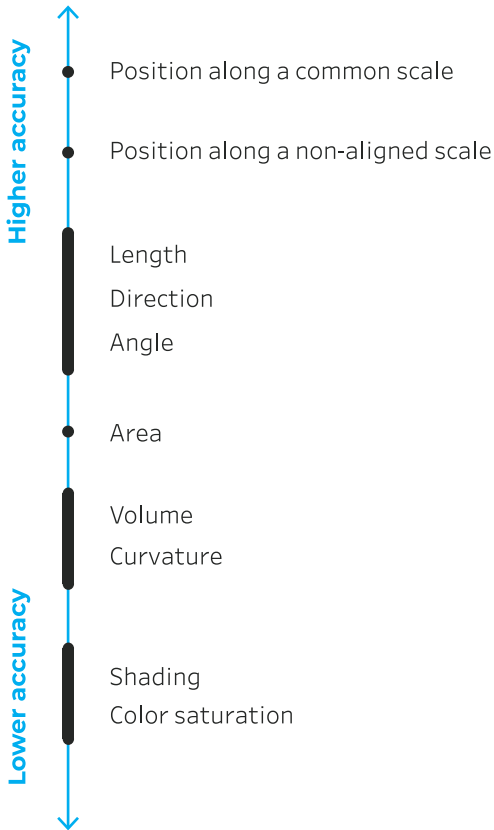


Color Hue

Red = 3
 Orange = 10
 Yellow = 14
 Green = 25
 Light blue = 30
 Dark blue = 50
 Purple = 87
 Dark grey = 95
 Light grey = 100



Enable accurate estimates



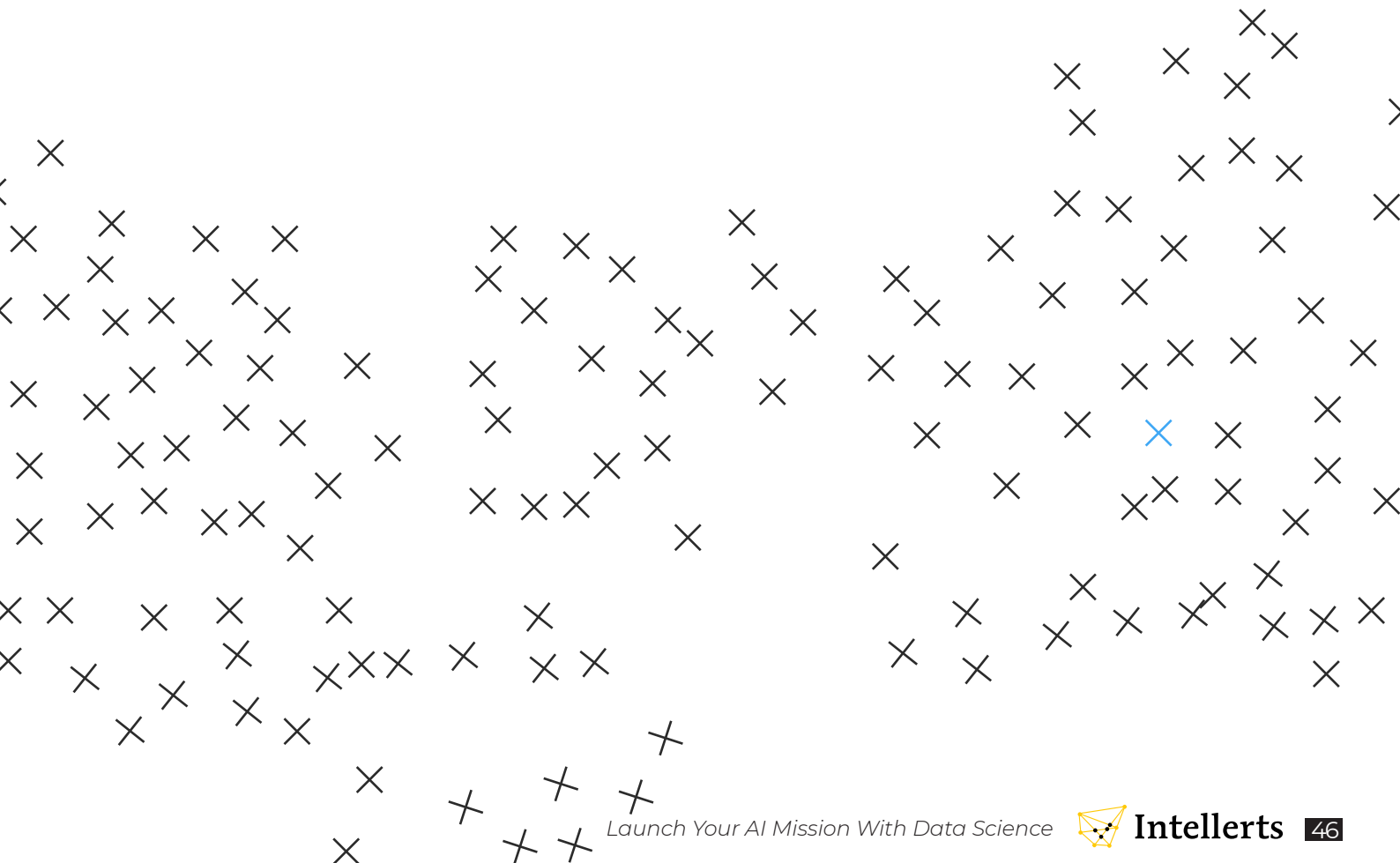
RANKING VISUAL VARIABLES

Some researchers rank the visual variables, like the ranking on the left. In this plot, you will notice *positions along a common axis*, like a bar or line chart, are the most effective at informing the user about accurate estimates.

But don't start using bar charts for every dataset. Remember, data visualizations can have different purposes. Maybe the visualization needs to turn heads and be remembered; something a bar chart isn't great at.

The picture on the left also states that color is the least accurate. This is not always true. Try to find the blue X below. Easy right? Did you use the *position along a common axis*? Not in the slightest, you used color. Color may not be the best at representing exact values, which the earlier ranking was concerned about, but it's very good at steering the eye to important details.

You need to know what your audience needs. This cannot be overlooked as there is no one-size-fits-all solution in data visualization.



PRINCIPLES FOR DESIGN

→ GOOD DATA VISUALIZATION IS TRUSTWORTHY

→ **GOOD DATA VISUALIZATION IS ACCESSIBLE** The designer must have a good understanding of how consumers consume a data visualization and change it accordingly. Additionally, the designer must understand for whom they are designing and try to be as inclusive as possible. For instance, 1 in 12 men (1 in 200 women) are color blind; picking the wrong color combinations could make a graph unreadable for 8% of men. Also, different people want different things from the data visualizations. For instance, many people want a chart accompanied by a table with the same data to gain an in-depth understanding.

→ **GOOD DATA VISUALIZATION IS ELEGANT** An elegant design attracts the audience and stimulates them to understand it. But elegance should never be prioritized ahead of trustworthiness.

MISTAKES TO AVOID

X DATA VISUALIZATION MISTAKES

- Use of 3D
- Wrong chart type
- Chart too complex
- Improper use of color
- Cherry picking
- Not following conventions
- Omitting key information
- Bubble size based on radius
- Too much clutter

Creating well-crafted and effective data visualizations is a real art. Due to the developments in and the gaining popularity of data science, there are now also people who only focus on this one aspect of data science. Even if you are not a visualization expert, you can easily improve the quality of your visualizations by avoiding the above nine data visualization mistakes.

These mistakes do not follow the three principles of good visualization design. Many of these mistakes are made unconsciously due to a lack of knowledge, but sometimes these “mistakes” are made on purpose to mislead the audience.

For example, if cherry-picking is used, or where the full picture is not revealed but the facts matching the opinion of the creator are used to create the impression of a change, even where there is relatively little change. These two examples both violate the principles of trustworthiness.

Not following common design conventions is another type of violation of the accessibility principle. Or too much clutter on your data visualization violates the elegance principle. Some mistakes have a negative impact on some principles. Improper use of color, for example, can lead to a chart not being accessible and elegant.



CHAPTER EIGHT

Tell The Story



GOAL

Create compelling storyline in which insights are communicated in a formal process (business and data scientists).

TOPICS

Explaining about the analyzed problem results, conclusions and recommendations. "Lessons learned"

PITFALLS

Presentation and visualization not adapted to the targeted audience, insufficient awareness for the sensitivity of topics, fragmented data sets and ownership

The final step of the 8-step model is to bring together and communicate all of your insights. This is when compelling presentations and dashboards are created.

EVALUATION FIRST

Whether a project is following a short-cycled lean approach or a more traditional project methodology, you must evaluate your work before you tell your story to others; only then you can tell the complete story.

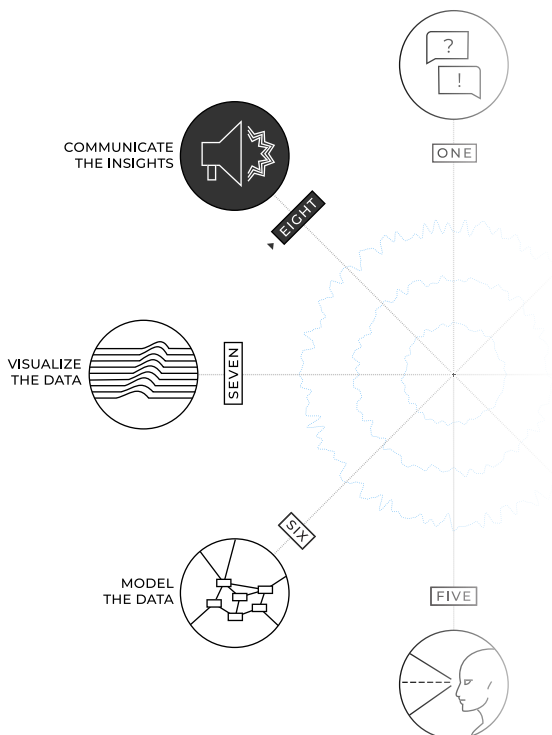
Has the problem that was defined in step one fully been solved, for example? And what are the lessons and the desired next steps?

Whether a project is successful is measured by its impact on the business, not by the model's performance. It is possible to build a highly accurate model with little or no business impact.

That's why it is so important to define a clear problem statement in step one of the 8-step model. This ensures the real problem and the desired solution have been identified at the start of the project, along with the expected benefits, and whether the solution is in line with the business strategy. All this needs to be included in your evaluation presentations.

DIFFERENT AUDIENCES

After evaluation, it is time to communicate the outcomes. Often, you must communicate across different audiences. This can be within your team, external stakeholders, or management, for example.



The content, design, and style of the presentation should be tailored to the audience. To achieve this, try asking yourself the following questions:

TEAM
STAKEHOLDERS
MANAGEMENT

- Who is the audience and what are they like?
- What is the technical knowledge of the audience?
- Should the technical details be kept to a minimum or is there a need to share them?
- What is the purpose of the presentation? Is it to inform, or to seek funding, permission, or something else?
- What are the pain points for the people in the audience and how will extending this project provide a solution?
- What are the actions for the audience?
- Can you anticipate resistance? For example, are people afraid that the project might negatively impact their job?

The answers to these questions will also help to determine which visuals should be included in the presentation. Sometimes, it is necessary to go back to step 7 and adjust your visuals or create new ones.

STORYTELLING FORMATS

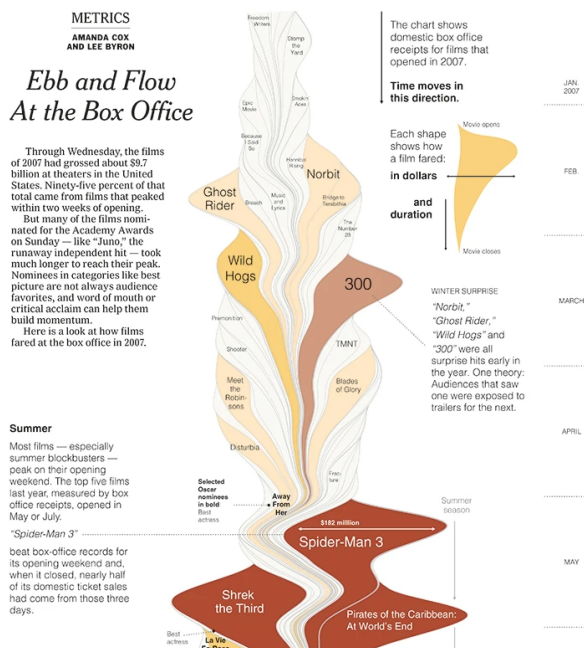
All presentations should fulfill a task the audience seeks to complete. The following three global formats can help you design with a purpose in mind.

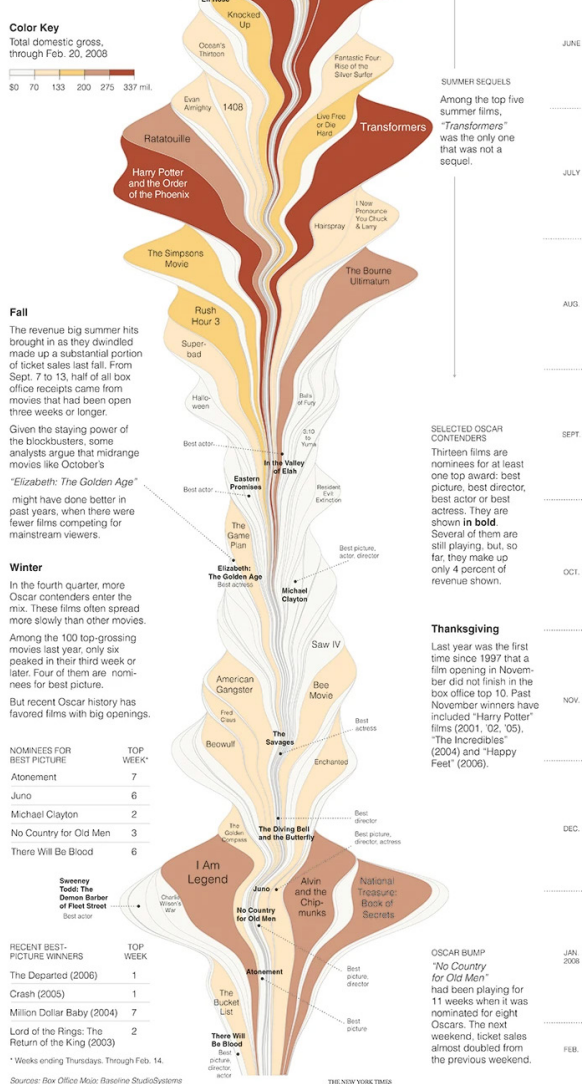
REPORT This style of report is the most 1-to-1 translation of the selected formats. In a report, facts are narrated just as they are: **A happened and then B happened, we now know fact C, D, and E, which we found doing study F.**

The core audience here are managers that quickly need to know if everything will go as planned and if there are any anomalies. KPI-focused reports (or dashboards) are a good example of this.

EXPLANATION Explanation formats are used to make the user **understand**. The New York Times, with their data-journalism pieces, could be the ultimate example of this format. Their text and data visualizations work together to create an explanative powerhouse.

NYTIMES ARTICLE
AMANDA COX
LEE BYRON





When explaining difficult subjects, it is important to start with the context. **Understanding** (*neurologically speaking*) works faster when we can place new knowledge in a bigger framework. This makes this knowledge easier to access and combine with other pieces of knowledge in the same framework (or context). Often, people like to dive deep into a topic. This works for an audience that is already knowledgeable about the topic. However, because no one can be an expert about everything, a lot of audiences need to take a step back first to really understand what is happening.

Also build on top of what people already know. *Imagine, you're describing what a car is, but not to someone in the 21st century but rather to a knight in medieval times. You wouldn't start with the shift-poke or the texture of the seats. You start with things the knight already knows, like horse riding, carriages, wheels, and travelling from Nottingham to Edinburgh.* A lot of audiences that you are going to explain to will be greatly helped if you use concepts they know and can relate to.

PITCH & DRAMA Pitches and dramatic pieces (e.g., TED talks) are crucial formats to change the viewers' attitude towards the presented topic. They often follow integrated and creative storylines to hold the attention of the audience. In this case, the audience seeks to invest or to be inspired by the topic.

3 WAYS TO MAKE DASHBOARD

Not all dashboards are created equal, and a big reason is that different users need to use them for different purposes. At Intellerts, we discriminate between three purposes and change the dashboard layout and design accordingly.

FAST REPORTING; KPI FOCUSED Managers are more concerned with steering the ship than deep analysis. They quickly want to find out how their companies (or other companies) are doing. KPIs are a great way of showing the key information. However, you should not only show the KPIs because much-needed context from data visualizations is missing and this could lead the audience to draw wildly wrong conclusions.

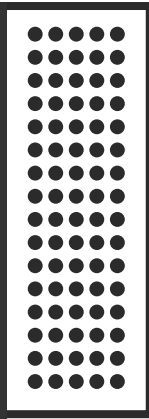
DATA EXPLORATION; ANALYSIS FOCUSED For deep analysis, you don't want to only show the analyst the conclusions (like KPIs). You want to give them the tools to delve deep into the data. Luckily, nowadays, many dashboard tools give a range of functionality to facilitate interactive data exploration.

An example of interactivity is brushing, where the user can highlight certain data points. Linking is when you link two data visualizations with each other. When you combine brushing and linking, you could highlight data points in one chart and highlight the linked data points in another visualization. This allows for strong analysis.

DATA STORYTELLING; EXPLANATION FOCUSED This is useful when the user is not acquainted with the domain or doesn't know the kind of analysis that is being conducted. The creator of the dashboard takes more authorship over the dashboard to create a better storyline.



Final Thoughts



WHY DATA SCIENCE PROJECTS FAIL

More than 85% of Data Science projects fail, according to Gartner. Every step in the Intellerts' 8-step model has pitfalls. Avoiding them will dramatically increase the chance of project success.

Data Science projects do not often fail because of operational issues. Typically, tactical and strategic aspects dictate the success of a Data Science project. These aspects are usually outlined at the start of a project. So, it is important to scope the project correctly. What's more, many critical success factors are not in place at the start of the project. This is often due to a lack of support from key stakeholders, uneducated business leaders or a lack of teamwork or good data science team.

If the change management side is neglected, this can also lead to project failure. Other, less common reasons are data quality issues and problems during the modeling step.

X COMMON MISTAKES

Thinking about the solution at too early a stage.

Understanding of the problem is too vague and lacks detail (only high level understanding).

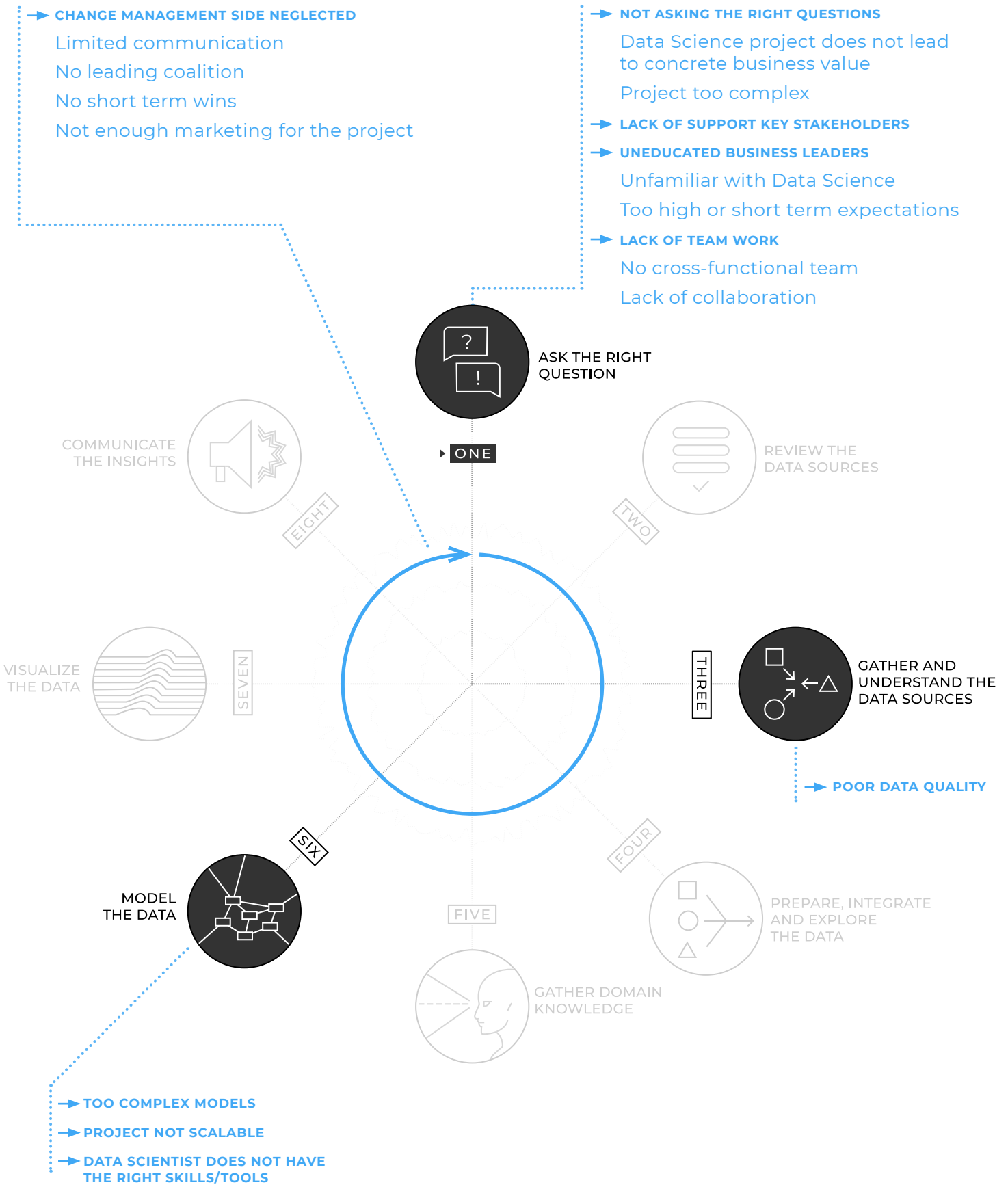
Problem is described with jargon where the meaning is not fully understood.

Problem is not aligned to the business strategy. The real, deep lying problem is not defined.

Domain knowledge not properly taken into account.

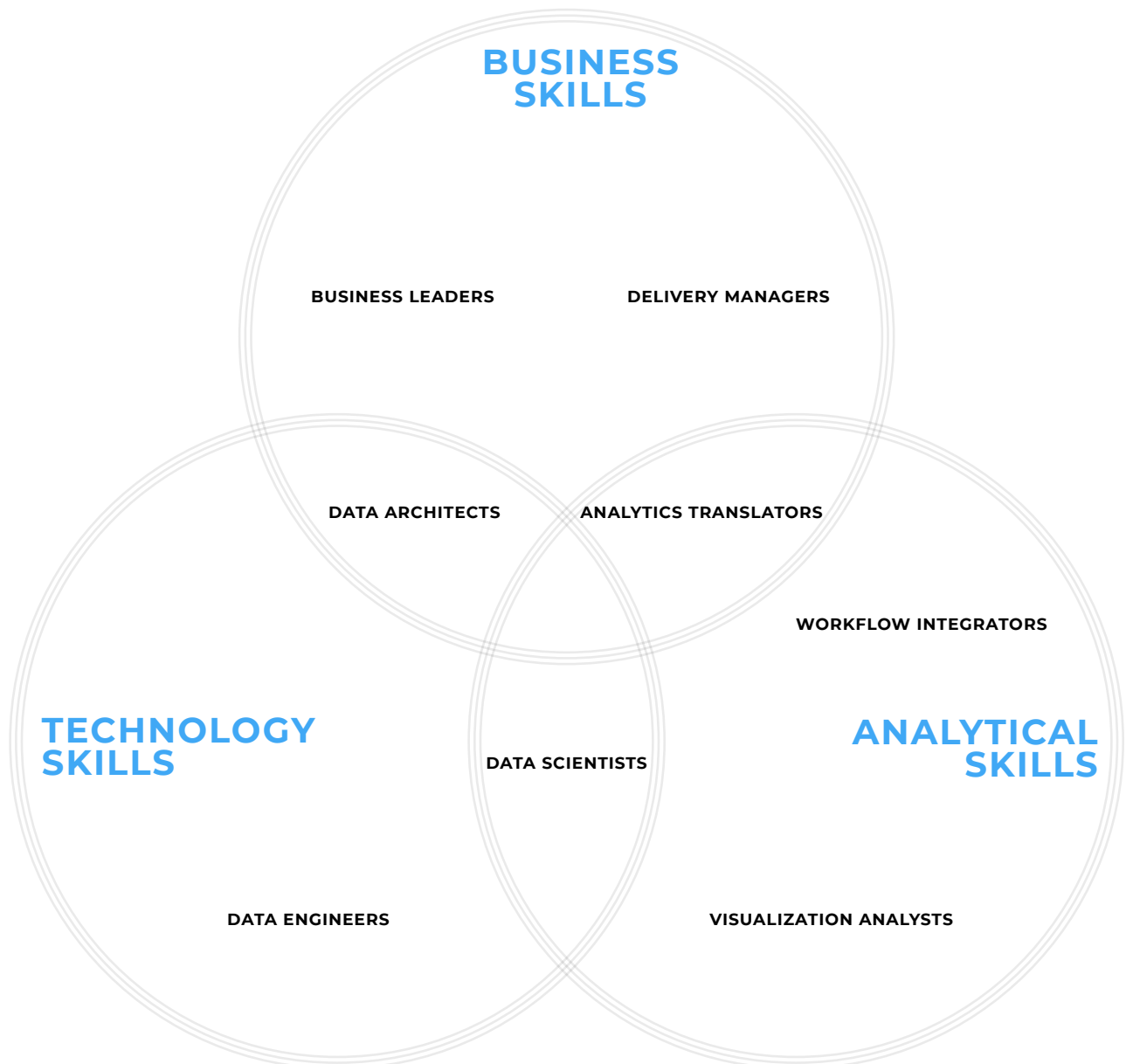
Not enough support from stakeholders.





DATA SCIENCE IS A TEAM SPORT

Data Science has developed fast in the last few years due to a range of technological advancements and the ubiquitous nature of structured and unstructured data. Now, Data Science has impacted every industry. Demand for specialized knowledge has increased. As a consequence, it is not possible for one individual to master every aspect of Data Science. It is now more important than ever to gather the required level of knowledge through teamwork.



A variety of roles and skills are now required to ensure the success of a Data Science project. The main roles include:

- ▶ **BUSINESS LEADERS:** Lead analytics transformation across the organization
- ▶ **DELIVERY MANAGERS:** Deliver data- and analytics-driven insights and interface with end-users
- ▶ **DATA ARCHITECTS:** Ensure quality and consistency of present and future data flows
- ▶ **ANALYTICS TRANSLATORS:** Ensure analytics solve critical business problems
- ▶ **DATA ENGINEERS:** Collect, structure, and analyze data
- ▶ **DATA SCIENTISTS:** Develop statistical models and algorithms
- ▶ **VISUALIZATION ANALYSTS:** Visualize data and build reports and dashboards
- ▶ **WORKFLOW INTEGRATORS:** Build interactive decision-support tools and implement solutions

Some roles only require one type of skill set. This may be analytical, technology or business skills, for example. Other roles rely on the intersection of two skill sets. These roles include data architects, data scientists and analytical translators. All of these roles play an important role, connecting two different worlds.

According to McKinsey, *the analytical translator is the new must-have role*. Translators typically have a very versatile profile with skills such as business and domain knowledge, project management as well as strong acumen in quantitative analytics and structured problem-solving. Their unique skill set can help businesses increase the return on investment from their analytics initiatives. They are instrumental in identifying the right opportunities to pursue, helping to ensure that all team members work in harmony.

It is important to bridge the technical expertise of your data engineers and data scientists with the operational expertise of marketing, supply chain, manufacturing, risk, and other frontline managers.



**Want to learn more about
Data Science?**

**We would love to help you
go a step further.**

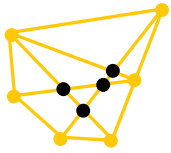
Feel free to reach out.



Martin Haagoort
CEO
M.Haagoort@Intellerts.com
+31 6 116 414 13



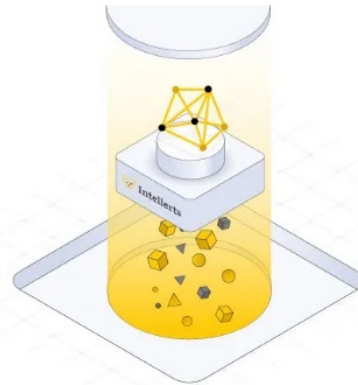
Niels van Rees
CSO
N.vanRees@Intellerts.com
+31 6 154 661 05
[Schedule a meeting](#)



Intellerts

PROFESSIONAL SERVICES

Intellerts provides a range of professional services, tailored to your needs. This includes BI/AI modelling, the development of specific AI/ML models, and recommendations tailored to your data science environment. We support your in-house AI capabilities with our unique Data Science “center-of-excellence” approach.



SOLUTIONS

We have developed a range of data science and AI solutions, across multiple industries and use cases. With our agile approach, we can maximize the value of your data, providing your business with faster insights, better decisions and improved outcomes.



PLATFORM

Our AIFA® Data Science Platform streamlines your BI, data processing and AI tasks. From onboarding data to sharing powerful and actionable insights, our intuitive workflow streamlines your end2end data process in an efficient and secure manner.



OUR CLIENTS



ISO CERTIFICATES



OUR PARTNERS

