Course Book

USE CASE AND EVALUATION

DLMDSUCE01

iu
INTERNATIONAL
UNIVERSITY OF
APPLIED SCIENCES

USE CASE AND EVALUATION

MASTHEAD

Publisher:
IU Internationale Hochschule GmbH
IU International University of Applied Sciences
Juri-Gagarin-Ring 152
D-99084 Erfurt

Mailing address:
Albert-Proeller-Straße 15-19
D-86675 Buchdorf
media@iu.org
www.iu.de

DLMDSUCE01
Version No.: 001-2024-0430

N. N.

TABLE OF CONTENTS

USE CASE AND EVALUATION

Appendix

INTRODUCTION

WELCOME

SIGNPOSTS THROUGHOUT THE COURSE BOOK

This course book contains the core content for this course. Additional learning materials can be found on the learning platform, but this course book should form the basis for your learning.

The content of this course book is divided into units, which are divided further into sections. Each section contains only one new key concept to allow you to quickly and efficiently add new learning material to your existing knowledge.

At the end of each section of the digital course book, you will find self-check questions. These questions are designed to help you check whether you have understood the concepts in each section.

For all modules with a final exam, you must complete the knowledge tests on the learning platform. You will pass the knowledge test for each unit when you answer at least 80% of the questions correctly.

When you have passed the knowledge tests for all the units, the course is considered finished and you will be able to register for the final assessment. Please ensure that you complete the evaluation prior to registering for the assessment.

Good luck!

BASIC READING

Gilliland, M., Tashman, L., Sglavo, U., Makridakis, S. & Petropoulos, F. (2021). *Business forecasting: The emerging role of artifical intelligence and machine-learning.* Wiley; Safari. http://search.ebscohost.com.pxz.iubh.de:8080/login.aspx?direct=true&db=cat05114a&AN=ihb.52472&site=eds-live&scope=site

Hyndman, R. (2018). *Forecasting: Principles and practices (2nd ed.).* OTexts. (Available on the Internet)

Nussbaumer Knaflic, C. (2015). *Storytelling with data: A data visualization guide for business professionals.* Wiley. http://search.ebscohost.com.pxz.iubh.de:8080/login.aspx?direct=true&db=cat05114a&AN=ihb.44883&lang=de&site=eds-live&scope=site

Osterwalder, A., & Pigneur, Y. (2010). *Business model generation: A handbook for visionaries, game changers, and challengers.* Wiley. http://search.ebscohost.com.pxz.iubh.de:8080/login.aspx?direct=true&db=cat05114a&AN=ihb.45837&site=eds-live&scope=site

Parmenter, D. (2020): *Key performance indicators : Developing, implementing, and using winning KPIs (4. ed.).* John Wiley & Sons, Incorporated. http://search.ebscohost.com.pxz.iubh.de:8080/login.aspx?direct=true&db=cat05114a&AN=ihb.49461&site=eds-live&scope=site

REQUIRED READING

UNIT 1

Osterwalder, A., & Pigneur, Y. (2010). *Business model generation: A handbook for visionaries, game changers, and challengers.* Wiley. Chapter 1. http://search.ebscohost.com.pxz.iubh.de:8080/login.aspx?direct=true&db=cat05114a&AN=ihb.45837&site=eds-live&scope=site

UNIT 2

Gilliland, M., Tashman, L., & Sglavo, U. (2016). *Business forecasting: Practical problems and solutions.* John Wiley & Sons. Chapters 3.4, 3.7 & 3.16. http://search.ebscohost.com.pxz.iubh.de:8080/login.aspx?direct=true&db=nlebk&AN=1127827&site=eds-live&scope=site

UNIT 3

Parmenter, D. (2020): *Key performance indicators : Developing, implementing, and using winning KPIs (4. ed.).* John Wiley & Sons, Incorporated. Chapter 1. http://search.ebscohost.com.pxz.iubh.de:8080/login.aspx?direct=true&db=cat05114a&AN=ihb.49461&site=eds-live&scope=site

UNIT 4

Parmenter, D. (2020): *Key performance indicators : Developing, implementing, and using winning KPIs (4. ed.).* John Wiley & Sons, Incorporated. Kapitel 10. http://search.ebscohost.com.pxz.iubh.de:8080/login.aspx?direct=true&db=cat05114a&AN=ihb.49461&site=eds-live&scope=site

UNIT 5

Parmenter, D. (2015). *Key performance indicators: Developing, implementing, and using winning KPIs.* John Wiley & Sons. Chapter 10. http://search.ebscohost.com.pxz.iubh.de:8080/login.aspx?direct=true&db=nlebk&AN=985129&site=eds-live&scope=site

FURTHER READING

UNIT 1

Kerzel, U. (2021). *Enterprise AI canvas integrating artificial intelligence into business.* Applied Artificial Intelligence, 35(1), 1–12. http://search.ebscohost.com.pxz.iubh.de:8080/login.aspx?direct=true&db=bsu&AN=147676988&site=eds-live&scope=site

UNIT 2

Hyndman, R. J., & Koehler, A. B. (2006). *Another look at measures of forecast accuracy.* International Journal of Forecasting, 22(4), pp. 679—688. http://search.ebscohost.com.pxz.iubh.de:8080/login.aspx?direct=true&db=edselp&AN=S0169207006000239&site=eds-live&scope=site

Fleming, P. J., & Wallace, J. J. (1986). *How not to lie with statistics: the correct way to summarize benchmark results.* Communications of the ACM, 29(3), pp. 218—221. (Available on the Internet).

Hyndman, R. J. (2006). *Another look at forecast accuracy metrics for intermittent demand.* International Journal of Applied Forecasting, 4, pp. 43—46. http://search.ebscohost.com.pxz.iubh.de:8080/login.aspx?direct=true&db=edsbas&AN=edsbas.C4F99B37&site=eds-live&scope=site

UNIT 3

Agresti, A., & Coull, B. (1998). *Approximate is better than "exact" for interval estimation of binomial proportions app.* The American Statistician, 52(2). Taylor & Francis, Ltd. (Available online)

Cook, J. (2005). *Exact calculation of beta inequalities.* http://search.ebscohost.com.pxz.iubh.de:8080/login.aspx?direct=true&db=edsbas&AN=edsbas.463C0D49&site=eds-live&scope=site

UNIT 4

Bean, R. (2022). *Why becoming a data-driven organization is so hard.* Harvard Business Review Digital Articles, 1–6. http://search.ebscohost.com.pxz.iubh.de:8080/login.aspx?direct=true&db=bsu&AN=155479868&lang=de&site=eds-live&scope=site

Desjarding, J. (2017). *Every single cognitive bias in one infographic.* (Available on the Internet)

Kotter, J. P. (2012). *Leading change.* Harvard Business Review Press. http://search.ebscohost.com.pxz.iubh.de:8080/login.aspx?direct=true&db=nlebk&AN=675195&site=eds-live&scope=site

Treder, M. (2019). *Becoming a data-driven organisation: Unlock the value of data.* Springer. http://search.ebscohost.com.pxz.iubh.de:8080/login.aspx?direct=true&db=edsvle&AN=edsvle.AH36997874&site=eds-live&scope=site

LEARNING OBJECTIVES

The course **“Use Case and Evaluation”** focuses on the practical applications and fundamental aspects of designing a data-driven project. Several methodologies explaining how to identify use cases suitable for data-driven or machine learning based approaches are introduced. A detailed discussion emphasizes the role of the data in these use cases, such as the types of data that exist, in which systems they are likely to be stored, and how to judge and improve the quality of the data.

Defining “success” is a critical part of any project, and the course discusses in detail how to evaluate predictive models, both from a model-centric view, which focuses on the predication made by the model itself, and from a business-centric view. The latter evaluates how the decisions, which are derived from the predictions by the model, influence the performance of the businesses.

Furthermore, the course discusses how to identify and avoid common mistakes which arise from our evolutionary heritage as humans. Certain behavioral patterns are often detrimental to the objective evaluation of predictive models.

Introducing new approaches and methods into an organization often leads to the need for change. The final part of the course discusses typical patterns and pitfalls of change management, with focus on the integration of data-driven approaches into businesses and organizations.

UNIT 1

USE CASE EVALUATION

STUDY GOALS

On completion of this unit, you will have learned …

- how to define a use case.
- how to identify use cases that can be approached with machine learning and artificial intelligence.
- how to specify the use case requirements.
- how to identify possible data sources and data handling requirements.
- how to consider typical data quality issues in the discussion of new use cases.

1. USE CASE EVALUATION

Introduction

Before starting a new project, it is important to address the question: “How will the project be executed?” The arguably more important question should also be addressed: “Why should the project be executed?” This is referred to as the evaluation of a use case where we need to determine who benefits from a particular aspect, how to evaluate whether or not a project is successful, and how it should be tackled. The business model canvas and the machine learning canvas are two useful tools that help us to ask the right questions and guide the project team in setting up the project comprehensively, or lead them to the decision that a particular project idea is not worth pursuing. The latter is also an important outcome — rather than chasing ideas which have little chance of being successful or valuable, a thorough evaluation of use cases allows the project team to focus on ideas that are most likely to result in the highest return.

When dealing with data centric projects typically found in data science and artificial intelligence (AI) applications, the types of data that are available and how to deal with any data quality issues that might be encountered should also be discussed.

1.1 Identification of Use Cases

Use Case
A scenario or project that will hopefully create value in a unique way.

The identification of suitable **use cases** is the first critical step of any new project, venture, or enterprise. It typically starts with the fundamental question: “What do we want to do?”

Developing ideas for new business models is quite challenging. The main questions to address are:

- Is it a physical or digital product, or is it a service?
- Who would buy this product or service? Who is the target market and how big is it?
- How much would customers pay for the product or service?
- Who is the competition? Is it possible to create a market niche in which there is little or no competition?
- What is the unique selling point (USP), i.e. the offering which distinguishes this product or service from others?
- Is it possible to make a profit?

Use cases for data science and AI topics always include some form of prediction model, i.e. data are measured, stored, processed, and then fed into a predictive model. The resulting predictions need to be evaluated to determine whether they are “good” or “bad” and, crucially, whether they should be transformed into operational decisions that the business can act upon. This step is very important because any prediction or insight produced by a **predictive model** can only create value for a business if this is transformed into an action.

This can either be through a decision support system for a human expert or through an automated system which may be monitored by humans while operational decisions are made automatically.

Predictive Model
This uses a machine learning algorithm trained to predict unknown or future events from data.

Since all predictive models in data science or AI use cases are based on data, access to sufficient data is a key requirement. This implies that repeatable operational decisions are good candidates for a machine learning based scenario, whereas strategic decisions are likely incompatible with a machine learning approach. For example, few companies will open new production plants or factories often enough to build up a large database covering all aspects that are important when building a new factory. This scenario would require a strategic decision best made by humans: What should be produced at the factory? Where should the new factory be located? Is the infrastructure in the area (power, roads, and railway) suitable for the requirements of the factory? What are the political and environmental implications of building a factory at a given place? On the other hand, operational decisions such as the amount of raw materials, or power required per day, as well as the subsequent ordering decisions, can often be automated using predictive models and optimization. Use cases focusing on research and development (R&D) are best considered as a separate category: Although they may contain strategic or operational elements, use cases focused on R&D typically focus on achieving longer term goals or increasing the capabilities of a company. In many situations, this type of use case is not immediately concerned with operational decisions, but rather with improving the skill-set of employees or the development of the next-to-next generation of potential products. In many cases, R&D use cases produce a design study or proof-of-concept from which the next generation product is derived rather than directly producing a marketable product. Cross-industry or academic collaboration can be very fruitful in R&D use cases.

The evaluation of use cases involves multiple steps. Firstly, ideas which could be developed into a concrete use case must be collected. For example, a blank board and a brainstorming technique can be used to collect many different ideas. Quite often, however, a number of ideas and problems already exist which can assist in the formation of new ideas for use cases. It is very important at this stage to be open to new ideas and not preemptively think about implementation, cost, and other constraints — these important details will come soon enough. Limitations on new ideas may result in missing out on “moonshots”, or valid use case ideas which are considered unusual.

Furthermore, the requirement of success should be avoided at this early stage of the use case definition. If, for example, a manager requests that employees must “develop ten ideas, of which nine must be successful,” the subsequent process of identifying promising use cases is severely constrained from the start. First of all, the definition of success is unclear. Should the new activity be profitable? If yes, after how long? Is it a new area the company wants to expand into? Then, since some form of success is required, the resulting use cases will likely be very modest. Quite often some easy and unoriginal ideas will exist which can be easily covered — this should be done, of course, but the idea behind identifying use cases is to discover new and original ideas, not something obvious. Then, if use cases must be successful by definition, ambitions or risky ideas with great potential will not be considered. It is therefore better to encourage ideas with both potential and risk. At a later stage, this enables an informed decision concerning which use case scenario to proceed with.

Defining success is a critical part of establishing a new idea for use cases. It enables a collective understanding of what makes a use case successful. Particularly when discussing data science or AI use cases, the definition of success is often coupled with the performance metrics of the underlying predictive model. While these model-centric performance metrics are, of course, important when assessing whether the model can accurately predict future behavior, they do not necessarily capture the business value created from the model as they are more technical. The model provides a prediction which is then translated into an optimized business decision. This decision is based more heavily on the outcome of the use case than the predictive model itself. Therefore, when defining success of a use case, both components must be addressed:

- Model-centric evaluation: How good is the underlying predictive model?
- Business-centric evaluation: How would the project manager or senior management view the project? Which business metrics should be used to assess the performance from a business perspective? What is an acceptable range for these metrics?

Once a project's definition of success has been established, time should be spent discussing what should happen if the identified metrics are not reached or fall out of the acceptable range: Who needs to be notified? What action needs to be taken and by whom?

Overall, the process can be summarized in the following steps:

1. Collect ideas: Start from a blank board and brainstorm, or collect existing ideas, problems, and new ideas for business models.
2. Identify the type of use case:
   a) strategic decisions
   b) operational decisions
   c) R&D
3. Identify the ideas that should be discarded at this stage. Possible reasons for this could include the following:
   a) The idea is not compatible with the existing business offering or remit.
   b) The idea would expand the business offering, but in a way that is not compatible with the overall business strategy.
   c) There is insufficient data available, the required data cannot be collected, or the collection of the data would be too expensive, e.g. because a large fraction of the IT infrastructure would have to be changed.
4. Identify how the use case would generate value for the company. What benefit would the use case have if successfully implemented? Is there a market for it? How much would customers pay for it? Does the use case solve a real problem or aid users in a concrete way? What are the implications of implementing this use case in terms of production, marketing, support, customer-service, etc.?
5. Structure the ideas:
   a) Which data are required?
   b) Is a specific domain expertise required? Are the relevant experts available?
   c) How are the resulting decisions implemented?
6. Define success: Which metrics and KPIs are used? How will this be evaluated and monitored?
7. Assess the potential and risk of the use case.

Once all aspects of the use case have been discussed, performing a "pre-mortem" can help to identify any aspect which may have been missed. Despite careful planning of all relevant details, projects may still fail. In most cases, a **"post-mortem" analysis** is then performed to identify what went wrong and why. The idea of a "pre-mortem" is to perform the same analysis, but before the project starts. After all details have been discussed, the team should think about the following scenario: "The project team is now one year into the development and implementation of the project and it has gone wrong." Trying to identify all potential steps where a project might fail can help to locate potential risks and mitigation strategies.

Post-Mortem Analysis
An investigation into why a particular process or project failed.

Approaches

- Asset driven: What is available and how can value be derived from it? E.g. collection of unique data can gain value through offered insights and a data provider.
- Capability driven: What can I do and how can I derive value from it?
- Vision driven: What do I want to do? Outline a new idea that has both potential and risk, and work towards implementing it.
- Solution driven: Which issues can be addressed? How could the use case help with a given situation?

Example of a capability driven approach: IBM Watson

Watson is an AI system that was originally developed to demonstrate that a machine based system can answer almost any question.

In 2011, the system famously won the Jeopardy challenge against Ken Jennings, 74-time winner and Brad Rutter, 20-time champion.

However, since then, IBM has struggled to find a valuable commercial application for Watson's capabilities besides playing "Jeopardy". The following quote highlights the issue:

> 'IBM Watson has great AI,' one engineer said, who asked to remain anonymous so he wouldn't lose his severance package. 'It's like having great shoes but not knowing how to walk—they have to figure out how to use it' (Strickland, 2018).

IBM Watson is a very capable AI system, but it is currently a solution to a problem that does not exist yet. Starting with the solution does not guarantee that there is a problem to solve, or, in other words, having advanced capabilities does not necessarily imply that these capabilities can be used in a meaningful way beyond the original development process. The following quotation highlights the main challenge:

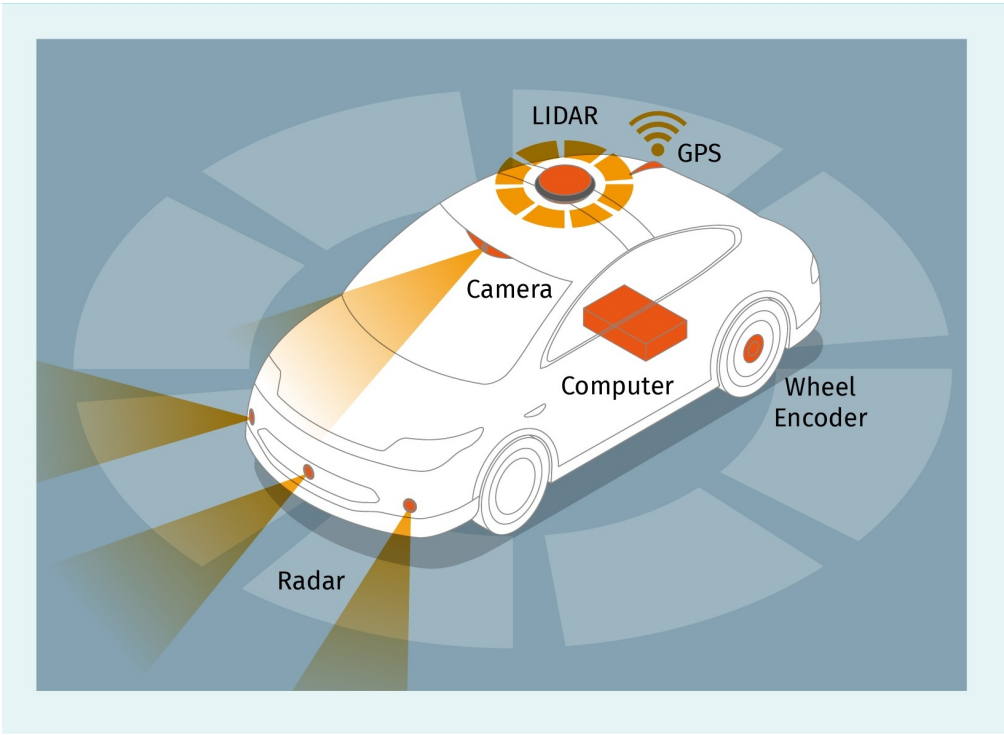
> Nothing that IBM has done in the past five years shows it has succeeded in using the core technology behind the original Watson demonstration to crack real-world problems... IBM's initial plan was to apply Watson to extremely hard problems, announcing in early press releases moonshot projects to end cancer and accelerate the development of Africa. Some of the promises evaporated almost as soon as the ink on the press releases had dried (Waters, 2016).

Apart from the technical challenges, there is a fundamental epistemological difference: “On Jeopardy! there’s a right answer to the question” (Waters, 2016).

Unlike “Jeopardy”, medical problems and illnesses do not have one fixed solution or cure that can be applied with 100% certainty. In many cases, a combination of approaches will have the highest chance of success for an individual patient. This means that the practical uses of Watson may be unable to extend to medical issues as it had originally intended.

Example of a vision driven approach: Self-driving car

Figure 1: Self-Driving Car



Source: Pixabay, 2019.

Many different people and companies are working on building a self-driving car, including start-ups and well-known tech companies. Their focus is to make the driver obsolete — but why?

Traditional car manufacturers continue to add more intelligent assistant systems that will allow the car to be driven more and more autonomously in the future. Start-ups and tech companies, however, aim to build a car that can immediately drive itself. This approach lacks focus as it does not address urban sprawl, attempt to include groups who cannot drive themselves (elderly, teenagers on a night out), or have a specific target group (taxis, public transport). It appears that they are tackling it as a technical challenge to be overcome, rather than a practical application that will be used by consumers. For example,

Google's self-driving car initiative from 2009 became Waymo (n.d.) – 10 years later, the self-driving car does not exist, and shows no sign of becoming available to the public any time soon.

Example of a solution driven approach: Fresh water in aviation

Figure 2: Fresh Water in Aviation

Source: Pixabay, 2019.

The successful operation of passenger planes requires the execution of many minor operations, for example fresh water must be provided for the lavatories. Unlike fuel, fresh water isn't important to safety and is therefore subject to a much more stringent optimization due to its weight. On one hand, sufficient water should be available to ensure normal operations, but on the other hand, limiting the amount of fresh water on board helps to reduce the overall weight of the plane, and therefore the amount of fuel required.

Being able to accurately predict how much water will be needed for each flight has a strong potential for cost optimization across an airline fleet.

1.2 Specifying Use Case Requirements

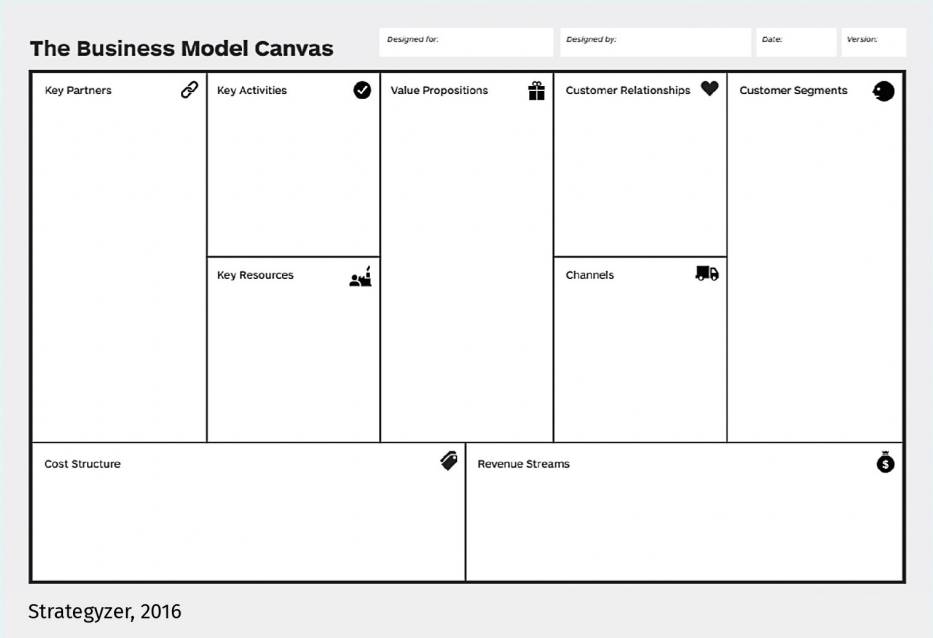
After the first brainstorming exercise for new use cases, and the indentification of which ones should be discarded at this stage, we are left with a short-list of the most promising use case ideas. The next step is to formally structure these ideas to ensure that all vital aspects have been covered. Various approaches exist to aid this process, for example, a canvas based method enables the discussion of all vital aspects while presenting the information in a compact form.

Business Model Canvas

The Business Model Canvas was developed by Osterwalder and Pigneur (2010), and started as the work of Osterwalder's PhD thesis (2004). The Business Model Canvas (BMC) is concerned with the development of new business models in general. The BMC focuses on the business aspects of use cases rather than details about data science or AI approaches. The main elements of the Business Model Canvas are:

- Customers
  - Who are the customers? Why would they buy the new product or service?
  - Who is the target market?
  - What is the relationship between customer and vendor? How is this relationship established and maintained (if necessary)?
- Value proposition
  - What is the value generated for customers?
  - Why would they buy the product or service?
  - Which need does the product or service satisfy? Is it a “real” need?
- Key partners
  - What is the value generated by the product or service?
  - How is the relationship with a partner established and maintained?
- Key activities
  - What exactly does the company do? How is the value being created?
- Key resources
  - Which resources are required to generate value for the customer?
- Channels
  - Which sales channels exist, e.g. stores, e-commerce?
  - How can customers integrate the channel into their processes?
- Cost
  - Which costs will have to be covered? E.g. manufacturing, logistics, warehousing, marketing, sales, etc.
- Revenue
  - Which revenue streams exist? E.g. single purchase, recurring revenue, licensing, consulting.
  - How much are customers willing to pay?
  - How much do customers pay for similar products?

Figure 3: Business Model Canvas

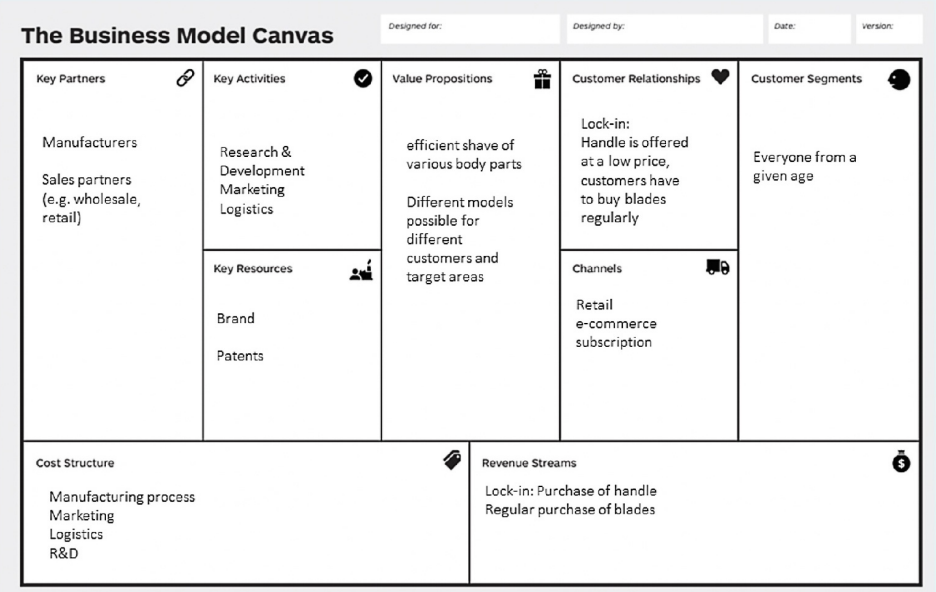


Source: Strategyzer, 2016.

Example: Business Model Canvas

The use of the Business Model Canvas can be illustrated by using the example of a company producing razors. The corresponding BMC may look like this:

Figure 4: Business Model Canvas for a Manufacturer of Razors



Author, based on Strategyzer, 2016

Source: Ulrich Kerzel, based on Strategyzer, 2016.

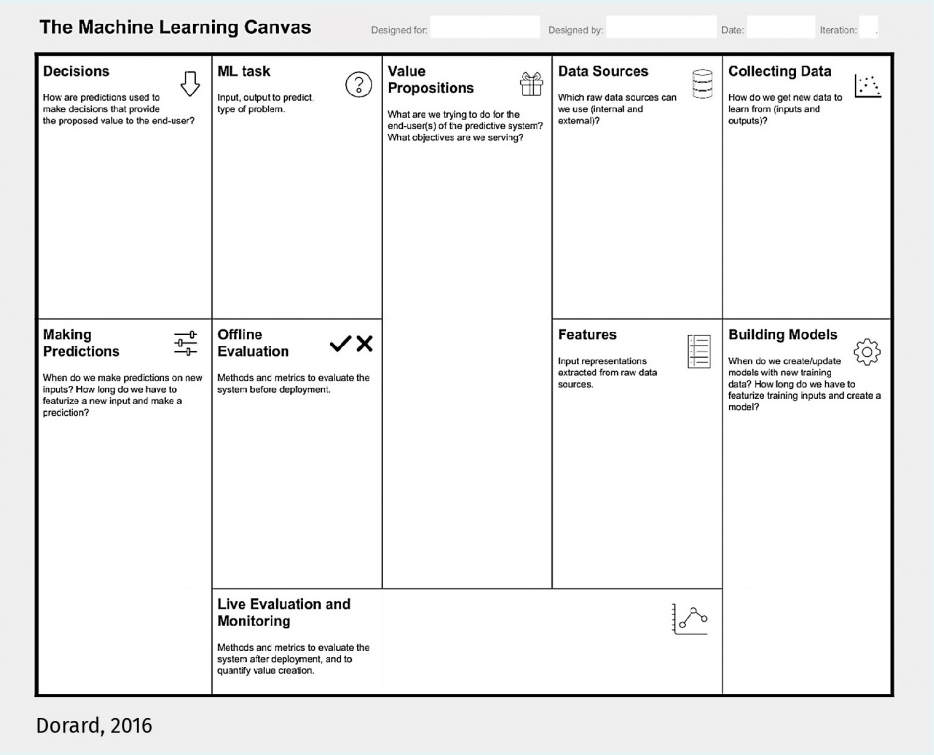
Machine Learning Canvas

The Business Model Canvas (BMC) is a helpful approach that can be used to identify business opportunities. However, it does not focus on how the value is created for customers in detail. This can be useful as it helps to avoid being preoccupied with the details too early in the process. However, in the case of data science or AI based use cases, some technical details are required at this early stage. In particular:

- Which data are required and can be obtained?
- What needs to be predicted?
- How is the prediction translated into an (operational) decision?
- How are predictions and the subsequent decisions evaluated and monitored?

The Machine Learning Canvas (MLC) was proposed by Dorard (2016) and focuses specifically on the technical aspects of implementing a machine learning based use case in a company. Both the MLC and BMC share the value proposition as a central element of the respective canvas, i.e. addressing the question the way that value is created for the customers. While the Business Model Canvas focuses on the business perspective, the Machine Learning Canvas integrates the questions that are vital for designing the machine learning model.

Figure 5: Machine Learning Canvas



Dorard, 2016

Source: Dorard, 2016.

Example: Supermarket replenishment

To make the use of the Machine Learning Canvas more tangible, the following example can be used:

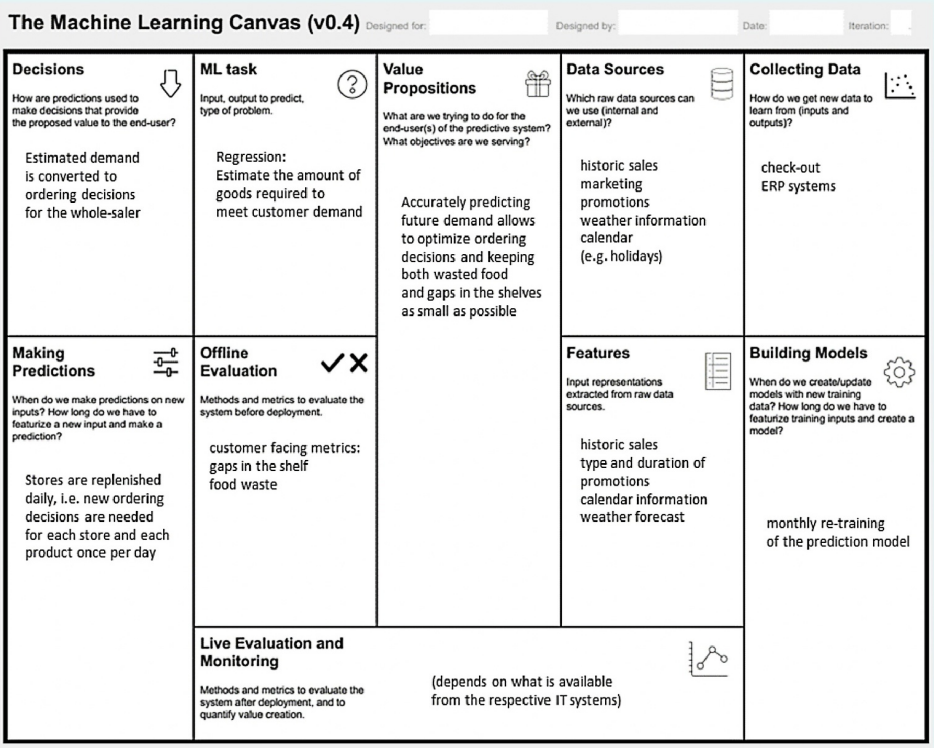
Supermarkets offer a wide range of products, ranging from fresh produce, flowers, and meat, to frozen food, drinks, and household items. Since customers do not typically pre-order, the actual daily demand for the products is unknown to the operators. Therefore, they must predict the demand so that they can order adequate amounts of the products and customer demand can be met without being left with excess produce. This is particularly challenging for perishable food such as fresh produce, dairy, or meat as they can only be kept on the shelves for a few days at most. After this time, these products must be disposed of. In order to optimize the replenishment process, estimating the expected demand is vital. Since no exact information exists concerning customer orders, the demand must be predicted using influencing factors such as previous sales, advertisements, and weather information.

This use case is ideally suited to a machine learning based approach since the relevant data can be recorded, for example, from the check-out systems. This information could help to make the thousands of ordering decisions that are required every day. In the case

of perishable food, the main customer facing metrics are the amount of food which has to be disposed of, and the number of gaps in the shelves. The challenging aspect of these two metrics is that they contradict each other. Putting too much emphasis on limiting the amount of food which has to be thrown out will lead to more gaps on the shelves which, in turn, reduces customer satisfaction, whereas reducing the gaps on the shelves will lead to increased waste. Defining success for this use case must include a strategic decision concerning the amount of waste and how many gaps on the shelves are acceptable for the specific chain or store.

The resulting Machine Learning Canvas can, for example, look like this:

Figure 6: Machine Learning Canvas for Supermarket Replenishment



Author, based on Dorard, 2016

Source: Ulrich Kerzel, based on Dorard, 2016.

1.3 Data Sources and Data Handling

Data are the foundation upon which all machine learning models in data science or AI use cases are built. The analysis of data allows conclusions to be drawn about correlations or causal relationships observed in the use case setting, and the sophisticated machine learning model can use this data to help to predict future events.

Types of Data

Data are classified into three general categories: structured data, unstructured data, and streaming data.

Structured data

Structured data are data which can be represented as a table in a fixed format, such as sales records of a shop, order books, student records, or marketing campaigns.

For example, structured data could look like this:

Table 1: Example for Structured Data (Sales Records)

Date	Product ID	Store ID	Promotion Flag	Sales Amount
2019-01-02	123	456	false	5
2019-01-02	125	456	false	3
2019-01-04	345	789	true	10
2019-01-07	378	978	true	7

Source: Ulrich Kerzel, 2020.

This example shows the (fictitious) sales of various products in several stores of a retail chain. Sales records are kept for each product in the store on a daily basis and, in addition to the number of products sold, it is indicated whether the product was promoted or advertised during the measured period.

This simple example shows that data are often recorded with a specific granularity. Instead of recording each individual sale with a more detailed time-stamp (e.g. minutes and seconds), sales records in this example are aggregated to the level of days. This choice is typically made when the IT systems are designed and results in any finer granularity being inaccessible. For example, if a particular use case definition requires the availability of each individual sales record, this use case cannot be pursued unless the IT infrastructure is changed.

An important type of data is time-series based data such as sensor readings during a manufacturing process. While each individual measurement can be represented as a structured data point (e.g. time-stamp, sensor ID, sensor value), the individual records are generally not independent from one another as they refer to the same underlying process. For example, if a temperature sensor is attached to a kettle, the sensor will record the rising water temperature when the kettle is switched on and the water is brought to a boil. Depending on how often the sensor is read, the expected temperature variation is quite small. If the kettle heats the water slowly and the temperature is measured, for example, once per second, the difference between each measurement should be very small. These correlations between data-points are generally higher if the sensor readings occur more

often than the change of the physical process they monitor. While time-series data can be stored in the same way as other structured data, it is worth considering whether dedicated data-storage solutions are a better alternative.

In general, it would be preferable to keep data at the lowest granularity possible as it is always possible to aggregate data at a later stage, however, once data are aggregated, the finer underlying structure is lost. On the other hand, storing data at a fine granularity also requires a large data storage and processing capacity. The expense of this type of storage may prevent it completely.

Unstructured data

Unstructured data are data that cannot be represented by a fixed format or table. Examples of unstructured data are images, audio recordings, text, etc.

In many applications, unstructured data cannot be used directly, but specific features such as the topic or the mood of a text, the number of speakers in a dialogue, etc. can be created automatically using deep learning approaches.

Unstructured data is often complemented with meta-data, which is data about data. The meta-data contains a range of details about what the data is about, or how it was recorded. For example, the recording of a telephone conversation is an example of unstructured data, whereas details such as the duration, telephone numbers of the participants, time of the call, etc. are meta-data which describe the details of the telephone recording. However, this kind of meta-data does not contain any information about the content of the recording.

Streaming data

Streaming data are data which are processed and recorded as a continuous stream, unlike batch data which are processed as blocks of data. Strictly speaking, streaming data is not a separate category as both structured and unstructured data can be “streamed” and processed or stored continuously. Examples include audio and video streams, sensor readings in a factory, or other data which are transmitted immediately unlike batch data which must wait until a block of data is ready to be sent.

However, streaming data typically requires a dedicated setup and an IT infrastructure that puts different constraints on the data processing infrastructure than a batch oriented architecture such as low-latency networks, high-speed storage systems, load balancing systems, etc. These requirements are amplified if the streaming data needs to be processed within a short time-span such as a few milliseconds.

Data Sources

Data used in data science and AI use cases can come from a wide variety of systems. Ideally, all data are stored electronically, contained in some catalogue, and easily accessible. However, in most real-world scenarios, data need to be traced, and relevant experts need to be found who know how to access the data or to export them to more modern systems if historic data has been archived in legacy systems.

Data are often found in:

- Databases for both structured and unstructured data: These databases typically contain operational data from various business processes and business units. Examples include sales records, marketing campaigns, sensor data and other measurements, images, videos, texts, or audio samples.
- Data warehouses: Data warehouses typically consolidate operational data in a central data storage system. Operational data are transferred to a staging area and then moved into the data warehouse. In order to limit the amount of data stored in the data warehouse, data are typically aggregated to a higher level after a certain amount of time. For example, sales records may be kept for a limited amount of time at the granularity of the operational systems, e.g. daily records for retail outlets and stores are kept for three years. Next, the data are aggregated to weekly, and then monthly numbers, so the original data in the operational systems and the data warehouse are deleted. If the use case requires data at a given granularity, the available historic data is often limited by the aggregation level in the data warehouse. Ideally, all data are kept at the most granular level, however, this is often impossible due to cost considerations and practicality.
- ERP (enterprise resource planning) systems: ERP systems typically contain all electronic records and data that are relevant to business processes such as contracts, the placement and scheduling of orders, invoices, employee records, etc. ERP systems have a significant overlap with operational storage systems and data warehouses, however, ERP systems generally focus more on the business process rather than individual data points.
- CRM (customer relationship management) and customer service systems: These systems contain all relevant information about the customers of a given business. Information includes who interacted with whom in relation to which issue, which people are involved in particular business processes, which support issues have arisen, how they have been resolved, etc.
- Filesystems: These exist for a wide range of data files such as CSV or Excel files for data, and document files such as PDF or Word. Even though many companies have an extensive setup comprising of various databases and data warehouses, a surprising amount of data is stored in PDF or Word files, as well as in hand-written notes, e.g. order forms or maintenance reports.
- Paper archives: Although it is tempting to focus on data which are already available in electronic form or in databases, paper documents and PDFs often hold critical information that is not contained in other systems, such as construction plans, maintenance logs, or other data that predate the change to electronic systems. In many use cases, these types of data are vital and an important part of accessing the data is evaluating ways to transform these data into electronic records.

- Data streams: Real-time data are often transmitted via a streaming infrastructure. In order to access these data, a client that subscribes to the relevant data stream needs to be registered.

When looking at potential data sources, it is often helpful to create a grid that lists the relevant parts of the value chain or departments on one side, and data sources which are expected to be vital for the project on the other.

Table 2: Example of Table Used for Potential Data Sources

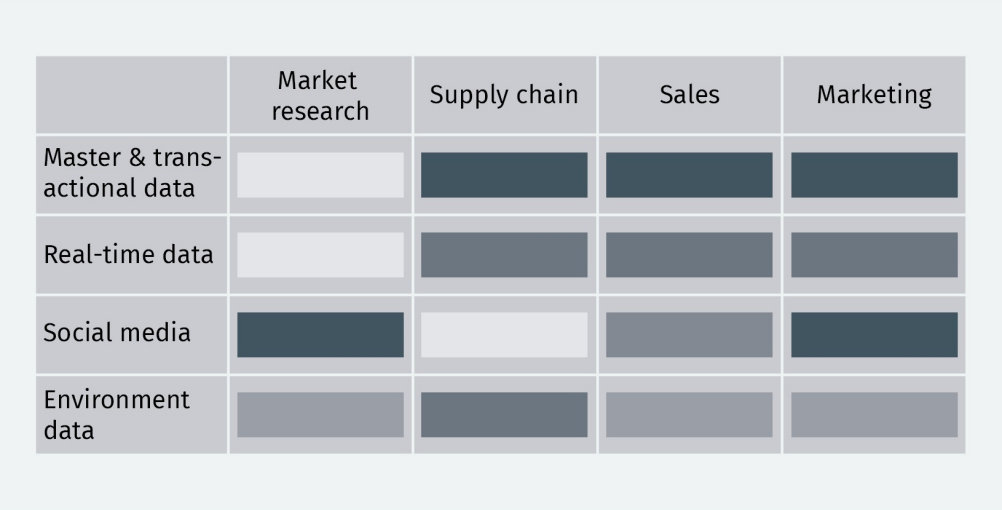
	Department 1	Department 2	Department 3	Department 4
Data Source 1				
Data Source 2				
Data Source 3				
Data Source 4				

Source: Ulrich Kerzel, 2020.

This makes it easier to determine which department has access to which data assets.

For example, the grid may look like this for a fashion retailer, where the shades indicate how important a given source of data is for the relevant business unit. In this case, the darkest shading indicates the most important sources.

Figure 7: Example of Table Used for Fashion Retailers



Source: Ulrich Kerzel, 2020.

In this example, the unit “market research” scouts new trends for the upcoming season as well as investigating the ways that consumers respond to products that are currently offered. The “supply chain” unit takes responsibility from the production of the items until they reach the stores, where the “sales” unit takes over and is responsible for all aspects of the sales process including returns. The “marketing” department is responsible for all catalogues, advertisements, marketing campaigns, social media interaction, and the determination of the price for each item.

On the data side, the “master & transactional data” refers to all data that relate to the articles such as article description, internal IDs, times that the article was on sale at a given price, etc. as well as the quantity of each item sold per store per day.

“Real-time data” refers to any data which are updated in real-time, for example, stock information. “Social media” data refers to all data recorded from social media sites such as Twitter, Instagram, Facebook, etc. as well as all data from the retailer's webpage itself. Details include, for example, how many visitors were there on a given day at any specific time, what they looked at, and where they came from.

The “environment data” refers to any data captured in or around the stores, e.g. number of visitors or customers, parking situation, weather information.

Each department will have different data assets as their operations have different requirements. Filling out this grid helps to identify which data sources are available where. As accessing each data asset typically takes a significant amount of time and effort to connect to the systems, a color coded grid can be used to establish an order in which to acquire the assets.

Data Quality

One of the most important aspects when dealing with data is its quality, i.e. if the data are correct, complete, and represent the underlying physical system. Inevitably, when dealing with data, some errors are to be expected. In particular, when beginning a new project, significant time and effort must be spent cleaning the data as any later machine learning model will only be trained using clean data. However, data quality management is an ongoing effort throughout each project as each new data delivery and system which is connected may contain new and previously unseen errors.

The examples below illustrate some typical issues when dealing with structured data. Consider the following table which contains the (fictitious) sales record of a retailer:

Table 3: Structured Data (Sales Record) – No Data Quality Issue

Date	Product ID	Store ID	Promotion flag	Sales amount
2019-01-02	123	456	false	5
2019-01-02	125	456	false	3
2019-01-04	345	789	true	10

Date	Product ID	Store ID	Promotion flag	Sales amount
2019-01-07	378	978	false	7

Source: Ulrich Kerzel, 2020.

The table consists of a date, identifiers for the store in which a specific product was sold, a Boolean flag indicating whether or not a promotion was applied, and the number of products sold. In the above example, no data quality issue can be found.

One typical issue is missing records, e.g.

Table 4: Sales Record Missing

Date	Product ID	Store ID	Promotion flag	Sales amount
2019-01-02	123	456	false	5
2019-01-02	125	456	false	
2019-01-04	345	789	true	10
2019-01-07	378	978	false	7

Source: Ulrich Kerzel, 2020.

In the above example, the sales amount for product 125 is missing. In order to be able to process this entry, this record must be cleaned using one of the following methods:

- removing the record,
- imputing the missing value, e.g. the mean of all records, or
- flagging the record, e.g. adding **NaN**.

NaN
This means "not a number" and is used to represent an undefined data value.

The best option depends on the specific problem, as well as the relevant domain and background knowledge. In general, handling missing data should not introduce an additional bias into the dataset. Depending on the algorithms used later in the analysis, flagging the record with NaN might be best, but other algorithms might require different approaches.

Table 5: Missing Sales Record Replaced By NaN

Date	Product ID	Store ID	Promotion flag	Sales amount
2019-01-02	123	456	false	5
2019-01-02	125	456	false	NaN
2019-01-04	345	789	true	10
2019-01-07	378	978	false	7

Source: Ulrich Kerzel, 2020.

This could include using an imputed value but adding a flag to indicate that the data record was altered. Using a flag in this way is advantageous because a machine learning algorithm used later in the modelling processes can, at least in principle, learn that the original data was modified. Furthermore, since this final preprocessing of the data happens only just prior to building a machine learning model, the individual capabilities of the chosen algorithms can be taken into account. Any decision about the way to handle this type of quality data issue should be considered thoroughly.

Another common data quality issue is an outlier, i.e. a value that seems irregular compared to the majority of the data:

Table 6: Outlier in Sales Amount

Date	Product ID	Store ID	Promotion flag	Sales amount
2019-01-02	123	456	false	5
2019-01-02	125	456	false	25
2019-01-04	345	789	true	10
2019-01-07	378	978	false	7

Source: Ulrich Kerzel, 2020.

In the above example, the sales record for product 125 is recorded as 25 units — much higher than any other value, especially considering that no promotion has been applied. However, deciding what to do in this case is not straightforward and needs careful consideration. First, it should be established whether or not there is a technical fault which has led to a wrong number being recorded. If the possibility of such a fault can be excluded, the value may be correct, despite being unlikely. A possible explanation in this case is that a customer bought all available stock. In this case, the data record should not be changed since it is a valid record. In other cases, e.g. if a technical fault is likely, the value should be replaced by NaN.

An example for this type of data could be wind speeds in moderate climate regions such as central Europe. Most of the time, wind speeds are quite low, rising to an occasional storm. In rare events, a strong storm or tornado may hit the area — since this happens very rarely, these wind speeds are outliers compared to the bulk of the data — but perfectly valid.

Yet another example of a data quality issue is the appearance of numbers outside the expected range, e.g.

Table 7: Outlier – Negative Sales Amount

Date	Product ID	Store ID	Promotion flag	Sales amount
2019-01-02	123	456	false	5
2019-01-02	125	456	false	–1

Date	Product ID	Store ID	Promotion flag	Sales amount
2019-01-04	345	789	true	10
2019-01-07	378	978	false	7

Source: Ulrich Kerzel, 2020.

In this case, the sales amount for product 125 is negative. At first sight this looks like an obvious data quality issue. Expected sales are zero or positive as no "negative items" can be sold. However, negative entries are possible if returned products are tracked as part of the sales volume. A good example is the power consumption of households with solar panels on the roof. In some regions, utility providers are required to buy back excess solar power if the production exceeds the demand of the household. This case is often modeled as negative power consumption by that household for billing purposes.

A further typical example of data quality issues is the change in the usage patterns of data entries, e.g.

Table 8: Unexpected Value in Promotion Flag

Date	Product ID	Store ID	Promotion flag	Sales amount
2019-01-02	123	456	false	5
2019-01-02	125	456	false	3
2019-01-04	345	789	true	10
2019-01-07	378	978	false	7
2019-03-01	123	456	1	8

Source: Ulrich Kerzel, 2020.

In this example, the numerical value "1" was entered in the column "promotion flag." Again, this looks like an obvious data quality issue at first sight. Instead of "1," the Boolean value "true" should have been entered. The system processing the data should have noticed the change because the schema definition of a database should not allow for numerical values.

However, in practice, these schema definitions are not always rigorously defined, enforced, or changed. In this specific example, there is a large gap between the first few records and the last row. A plausible explanation is that the column "Promotion Flag" was extended to accomodate different types of promotions, each represented by a numerical identifier. Instead of changing the format of the whole database, the table definition was changed. Although this does not follow best practices, it is often encountered in real-world scenarios and emphasizes the need to investigate any potential data quality issue carefully.

Further data quality issues which are often found in practical applications include:

- Interval-of-validity: Data are typically only valid for a limited amount of time, however, this valid range is often not explicitly specified. For example, a product is manufactured from a given list of raw materials. Even if the overall specifications of the final product don't change, the product itself is different if the raw materials are changed or the vendor providing the product is changed. As this may lead to issues when the product is used, any change made should be tracked.
- Tracking of interventions: Issues outside of the remit of the data are often not tracked. For example, a range of sensors record values in a manufacturing process. During maintenance, some of the sensors could be changed to a newer model which may behave differently when exposed to some ranges of the measured values. If these interventions are not tracked, the later model-building process will not be able to comprehend why some ranges of the data behave differently from others.
- Incomplete data history: Data storage systems are often designed so that the data are correct at any given moment in time, but they do not track each change that is made to the data. For example, a big company is restructured, and different parts of the company are now responsible for the manufacturing of products or different target markets, etc. Typically, great care is taken to make sure that all data storage systems and databases reflect the new responsibilities after the restructuring so that important aspects such as financial metrics can be tracked accurately. However, if this migration does not take in to account the old responsibilities, the historic data are practically useless since no one will be able to see which metric was valid for which business unit at a given time in the past.

In the previous paragraphs, we have discussed the importance of the quality of available data for building a sophisticated machine learning model. As the data are used to train the machine learning algorithm, any data quality issues, inaccuracies, or impurities can, at least in principle, be picked up by the machine learning algorithm and lead to sub-optimal or wrong predictions. It is therefore paramount to improve the quality of the training data as much as possible, as well as making sure that the data that trains the machine learning algorithms are representative of the use case that they are used in.

Ideally, a good algorithm is able to generalize from the training data, i.e. some level of noise can be tolerated. However, the following scenarios are particularly problematic:

- Systematic trends or shifts: If the data contain a bias which systematically distorts parts of the data, it is unlikely that the algorithm will be able to identify and correct for this, and it will assume that it is a “real” aspect of the data instead.
- Outliers: The treatment of outliers generally requires detailed knowledge of the systems that the data are taken from. Removing rare but unusual values introduces a bias in the dataset.
- Imputation of missing values: Naively, imputing missing values “cures” data quality issues, however, this can introduce a new bias into the data as the information is lost, which would have indicated that part of the data is missing. Furthermore, missing values may occur in specific scenarios — a missing value could contain further information. For example, if a sensor is faulty or operated outside its specifications, it might report non-sensical values leading to missing data in the storage system. If the missing value is

imputed without further consideration, this information is lost. This biases the data and removes the chance to fix these cases or to treat them differently to normal operating conditions.

SUMMARY

Identifying potential use cases and specifying requirements involves multiple steps, either starting from the beginning, or from previously identified opportunities. Special care should be taken not to limit use case ideas to projects which are likely to succeed, as this typically leads to use cases which are not ambitious. A critical factor in the definition of a use case is the definition of success, i.e. which measurable quantity needs to be fulfilled for the use case to be considered as a successful project.

Use cases can be classified into strategic use cases and operational use cases. In general, strategic use cases are better suited for human expertise and decision making, whereas operational use cases typically involve many repeatable decisions, meaning that they are ideal for data science and AI based approaches.

Use cases for data science and AI projects rely on high quality data. In most cases, data need to be accessed from a wide range of systems. Since these data are then used to train machine learning models, great care needs to be taken to assess and improve the quality of the data. When dealing with data quality issues, it is important to avoid introducing an additional bias into the cleaned dataset as a side-effect of removing the issue.

UNIT 2

MODEL CENTRIC EVALUATION

STUDY GOALS

On completion of this unit, you will have learned …

- which commonly used metrics exist to evaluate predictions.
- the advantages and disadvantages of commonly used metrics.
- how visual aides can support evaluating a prediction model.

2. MODEL CENTRIC EVALUATION

Introduction

A key point in the determination of use cases is the definition of success, i.e. how to decide whether the project is successful or needs to be improved. The model centric evaluation focuses on the predictive model itself. It aims to evaluate whether or not the model, which is used to predict unknown or future events based on new data, is reliable and accurate. In particular, the evaluation of the model focuses on the following aspects:

- Is the prediction accurate, and are the prediction errors as low as possible?
- Is the variance of the predictions as low as possible?
- Does the prediction show any bias?

Ex-post
Short for Latin "Ex post facto", meaning an evaluation of the facts or observations after they have occurred.

These questions are typically addressed in an **ex-post** test on an independent data set, which is used to evaluate the performance of the predictive model. As the data are already recorded, each event in the test is associated with the true (observed) outcome. Denoting a prediction for event i with $\mathrm{p_i}$ and the true observed event with $\mathrm{t_i}$, the prediction error can be defined as:

$$\mathrm{e_i} = \mathrm{p_i} - \mathrm{t_i}$$

To address the above questions, the value and spread of the prediction error $\mathrm{e_i}$ should be as small as possible and should not show any systematic trends. A wide range of metrics are used to determine whether the predictions made by a given model are of sufficient quality to be used in the wider context of a project. It is important to perform these validations on an independent dataset which shows the same characteristics as the data, the algorithm, or the classifier was trained with but has not been used in the training process. This dataset is also called the **test data** as it enables the performance of the prediction model to be tested on data where the true outcome is known, but the data haven't been seen by the algorithms before. In many cases, the test data is obtained by randomly choosing a fraction of the available data and separating it from the data used in the training process. As a rule of thumb, approximately 20 percent of the available data should be retained for testing.

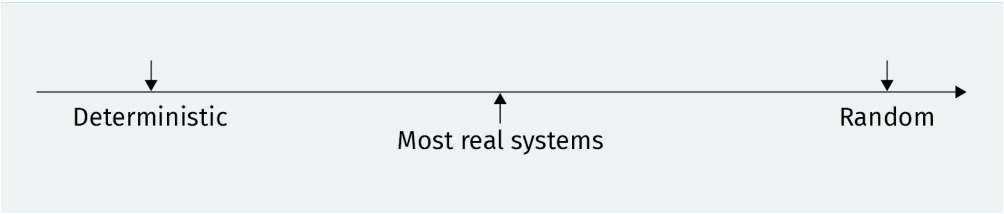
Test Dataset
This is an independent dataset used to verify the predictions and check for overtraining, which has the same characteristics as the data used to train the predictive model.

A key point in understanding the evaluation of predictions is that the predictions are essentially a **random variable** with an intrinsic uncertainty.

Random Variable
A variable where the value depends on the outcome of a random process. The values of a random variable might, for example, represent the outcome of a future experiment or process which is determined by a particular probability distribution.

The predictability of a model can generally be placed in a continuum between the extremes of a fully deterministic system and a system for which no prediction can be made:

Figure 8: Predictability Between Deterministic and Random Cases



Source: Ulrich Kerzel, 2020.

In a fully deterministic system, all outcomes of the system can be modelled precisely and the system can generally be described by a set of (complex) rules or equations. An example of a deterministic system is a grandfather clock where the movements of a pendulum are used to track the progress of time. The draw of lottery numbers, on the other hand, is an example of a random system. Even though the apparatus used to draw the lottery numbers looks like a simple mechanical system, the sequence of the numbers cannot be predicted. Most realistic use cases are somewhere in between: They have both a deterministic and a random component. The deterministic component allows the prediction of new or future events and the random component ultimately leads to an uncertainty associated with the prediction. This means that even if a perfect prediction model could be developed, the predictions will fluctuate around the true value due to the stochastic nature of the process and, therefore, the prediction error $\mathrm{e_i}$ cannot be zero for all predicted events.

The predictive model is based on a finite set of training data, meaning that the model is optimized using the events contained in the training data. If the model is then applied to new events outside the range of the training data, whether or not the prediction is useful depends on the generalization abilities of the model. This also implies that all data-driven models are inherently **Bayesian** when the training data is considered to be the prior knowledge available about the system that will be modelled.

Bayes
Reverend Thomas Bayes (18th century) created a school of statistics that uses conditional probabilities and evidence to calculate the probability of unknown events.

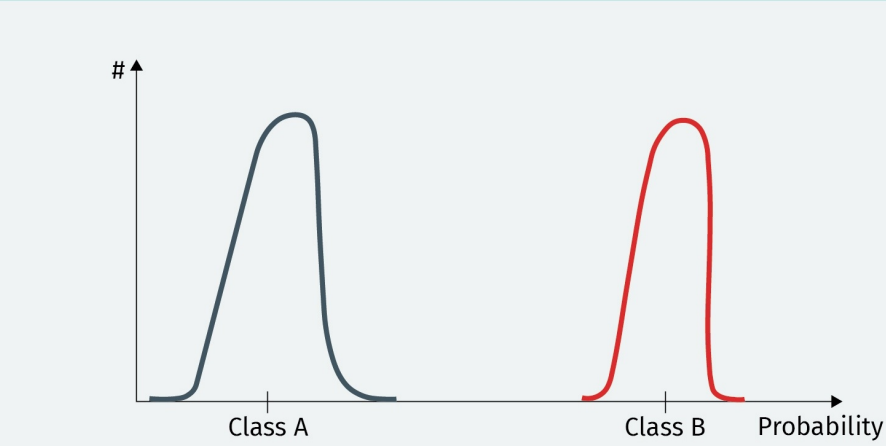
2.1 Common Metrics for Classification and Regression

Metrics for Classification

Type I and type II errors

In classification tasks, elements are assigned a specific class based on the threshold defined in the use case. For simplicity, we assume that there are only two relevant classes: A and B. The machine learning classifier assigns a probability between 0—100 percent to each classified event and the item is assigned to class A or class B depending on whether the predicted probability for this specific event is above or below the threshold. No machine learning algorithm can produce results which are perfect in any circumstance, but ideally, the distribution for the different classes are separated.

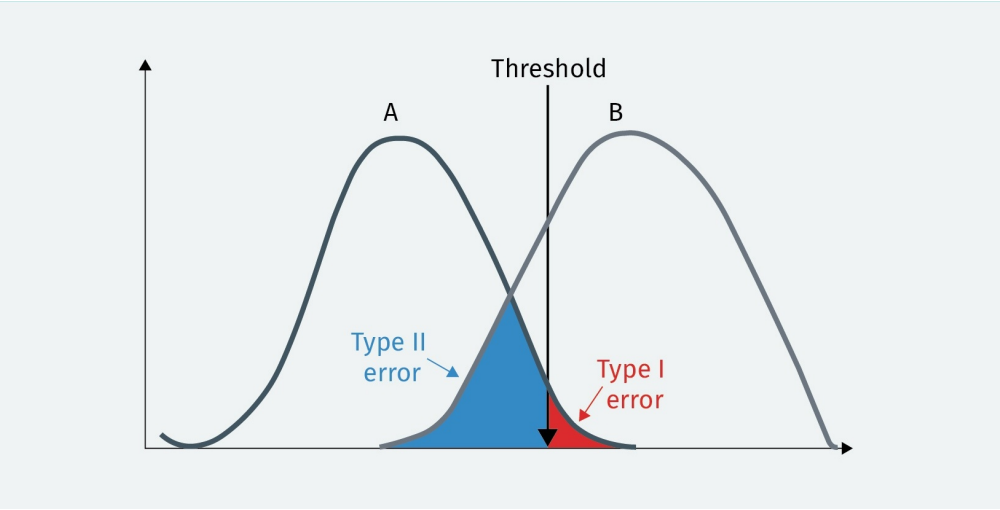
Figure 9: Example of Two Well Separated Probability Distributions



Source: Ulrich Kerzel, 2020.

Since, in this case, the two classes are well separated, any threshold set between the classes will lead to perfect results. However, in most realistic cases, there will be some events which can't be easily distinguished:

Figure 10: Type | and Type II Errors



Source: Ulrich Kerzel, 2020.

This implies that, depending on the exact value of the threshold, events (in the overlap region) can be misclassified as class A even though they belong to class B or vice versa. These are known as type I and II errors (Pearson & Neyman, 1930):

- Type I error: The hypothesis "event belongs to class A" is rejected, even though it is true. This is also called a "false positive."

- Type II error: The hypothesis “event belongs to class A” is wrong, i.e. the event belongs to class B, however, the event lies in the acceptance region for class A. This is called a “false negative.”

Any sample that passes the threshold will contain events which are classified correctly, i.e. events identified correctly as members of the class (“true positives”), as well as events which are misclassified, i.e. events which pass the threshold but do not belong to this class.

Positive and negative rates

If an event belongs to class A it is said to be a “positive”, otherwise a “negative”. For example, in fraud analysis, a known fraudulent event is a positive, a known non-fraudulent event a negative. The number of positive and negative events in an ensemble are denoted as condition positive (P) and condition negative (N), respectively. If an event is identified correctly as a positive event (e.g. a fraud case), it is denoted as a “true positive” (TP), if the event is correctly rejected as a negative it is a “true negative” (TN). If an event is wrongly identified to be positive (i.e. false alarm, type I error), it is known as a “false positive” (FN), if an event is rejected although it is a positive (i.e. missed event, type II error), it is a “false negative” (FN).

Then the following quantities can be defined:

- True Positive Rate (TPR) = TP/P = TP/(TP + FN) = 1 – FNR
- True Negative Rate (TNR) = TN/N = TN/(TN + FP) = 1 – FPR
- False Negative Rate (FNR) = FN/P = FN/(FN + TP) = 1 – TPR
- False Positive Rate (FPR) = FP/(FP+TN) = 1 – TNR
- False Discovery Rate (FDR) = FP/(FP + TP)
- Accuracy Acc = (TP + TN)/(P + N)

Precision, sensitivity, and recall

The precision or positive predictive value (PPV) of a classifier is then defined as (Perry, Kent, & Berry, 1955):

$$\text{Precision} = \frac{\text{selected and true (true positives)}}{\text{selected}}$$

i.e. the number of selected events which truly belong to the class divided by the number of events which have been selected by the classifier, irrespective of whether they belong to the class or not. This means that a classifier with a precision of 1 only returns correctly identified events.

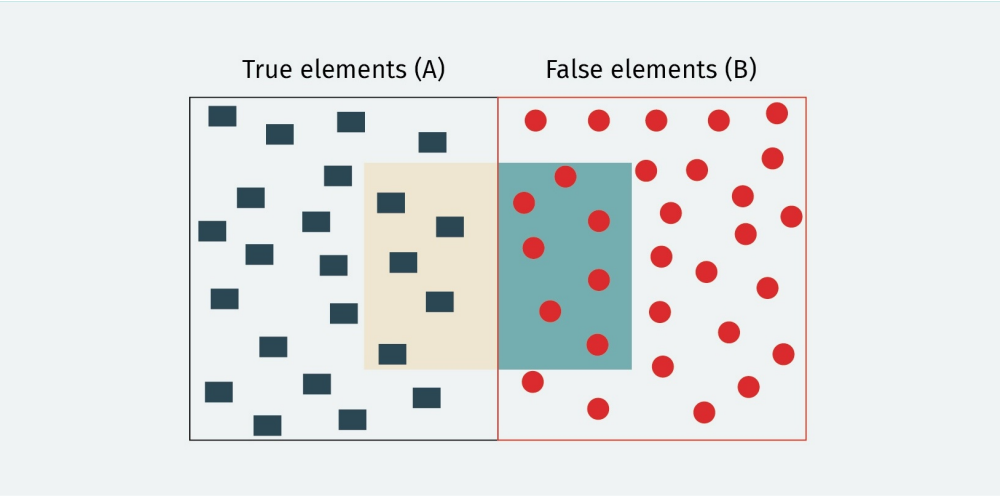
The recall or sensitivity (also: true positive rate) of a classifier is defined as (Perry, Kent, & Berry, 1955):

$$\text{Sensitivity} = \frac{\text{selected and true (true positives)}}{\text{true elements of relevant category}}$$

The sensitivity measures how many members of a given class are identified correctly. For example, if a (test) dataset contains 100 known fraud cases, the sensitivity measures how many of these cases are identified correctly when a particular threshold is applied to determine whether the predicted probability of a given event should be treated as fraud or not.

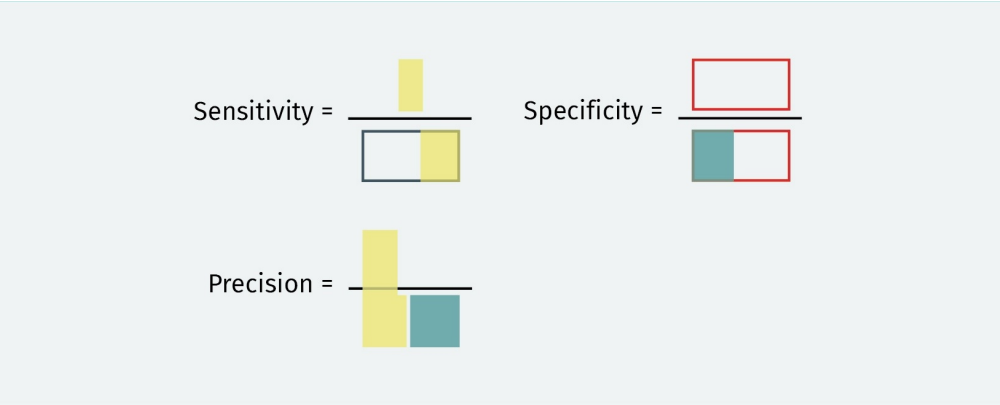
The specificity (also: true negative rate) complements the sensitivity and measures how many true negative events (i.e. events which do not belong to the class) are identified correctly.

Figure 11: Union of All True and False Elements Used in Predictions



Source: Ulrich Kerzel, 2020.

Figure 12: Definition of Sensitivity, Specificity, and Precision



Source: Ulrich Kerzel, 2020.

F-measure

The F_1 score (Chinchor, 1992) is defined as the harmony of precision and sensitivity, and combines both quantities into a single measure.

$$\mathrm{F}_1 = 2\frac{\mathrm{PPV} \cdot \mathrm{TPR}}{\mathrm{PPV} + \mathrm{TPR}}$$

Including a parameter β, a more general definition is (Chinchor, 1992):

$$\mathrm{F}_\beta = \left(1 + \beta^2\right)\frac{\mathrm{PPV} \cdot \mathrm{TPR}}{\beta^2\mathrm{PPV} + \mathrm{TPR}}$$

Confusion matrix

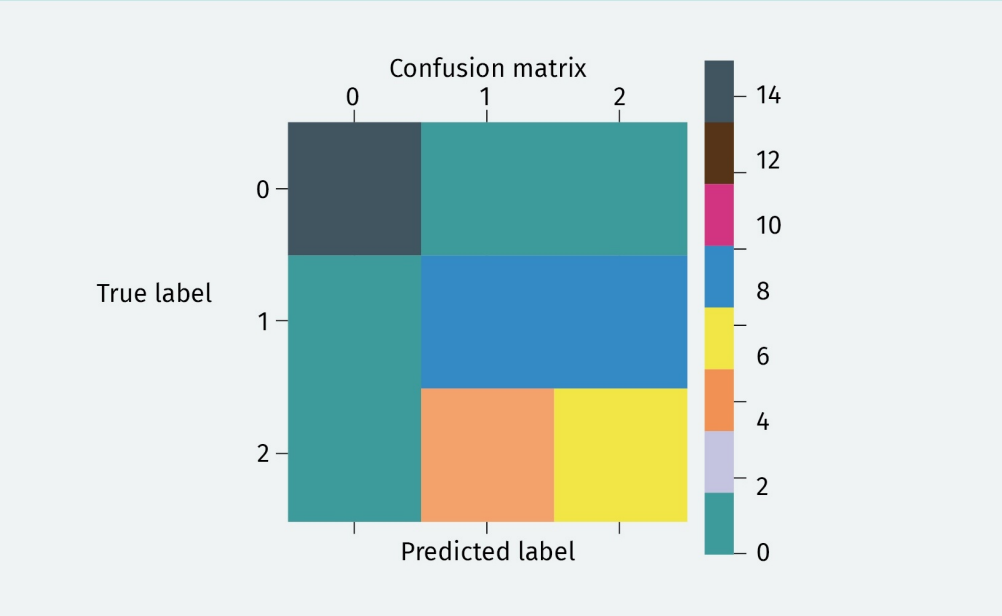
The confusion matrix (Stehman, 1997) is a tabular representation of the performance of a classifier. The rows and columns represent the true classes (or labels) in the test data, and the predictions made by the classifier. For example, the rows of the confusion matrix can represent the true labels, i.e. the true assignment of a given data point. The columns of the confusion matrix then contain the prediction made by the classifier. The individual cells in the table contain the respective combination of prediction and true label.

Ideally, the confusion matrix should have a diagonal form, i.e. all predicted labels agree with the true labels and the prediction algorithm works perfectly. In most realistic cases, the prediction algorithm will not be able to classify each data sample correctly and the confusion matrix helps to identify any classes that the algorithm cannot separate easily.

The data in the examples below belong to three different classes (denoted by 0, 1, 2). The color of the fields indicates the number of entries in each segment of the confusion matrix, evaluated using an independent test dataset. In the first example, the classifier is able to separate the first class (0) from the other classes (1, 2). However, the second row (true label = 1) shows that the classifier has difficulties distinguishing between classes 1 and 2.

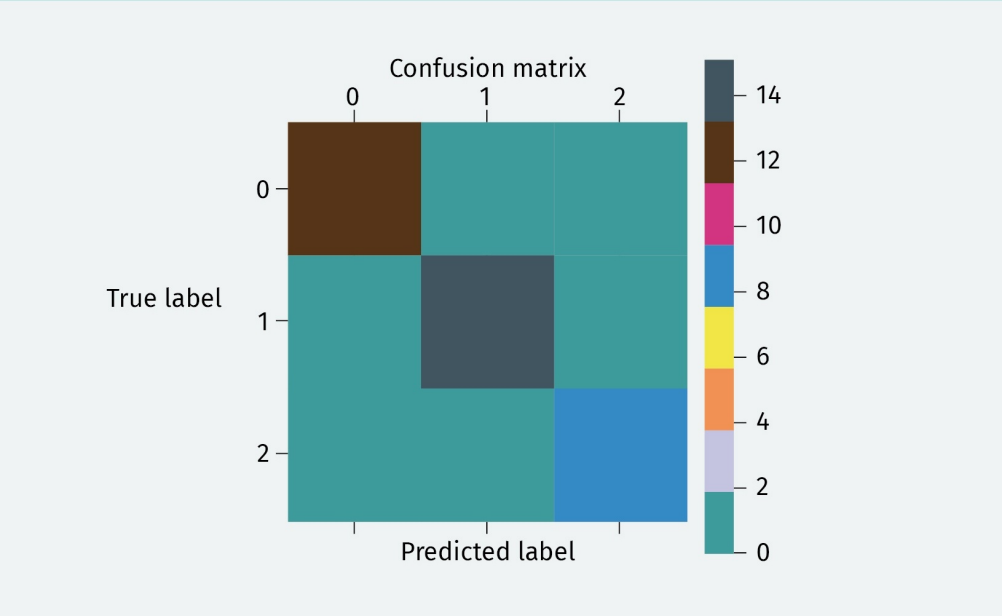
The confusion matrix in the second classifier looks much better: The main entries are on the diagonal line and all three classes are identified correctly.

Figure 13: Confusion Matrix for Three Different Cases — Mediocre Classifier



Source: Ulrich Kerzel, 2020.

Figure 14: Confusion Matrix for Three Different Cases — Well Performing Classifier



Source: Ulrich Kerzel, 2020.

Metrics for Regression

Mean absolute deviation (MAD)

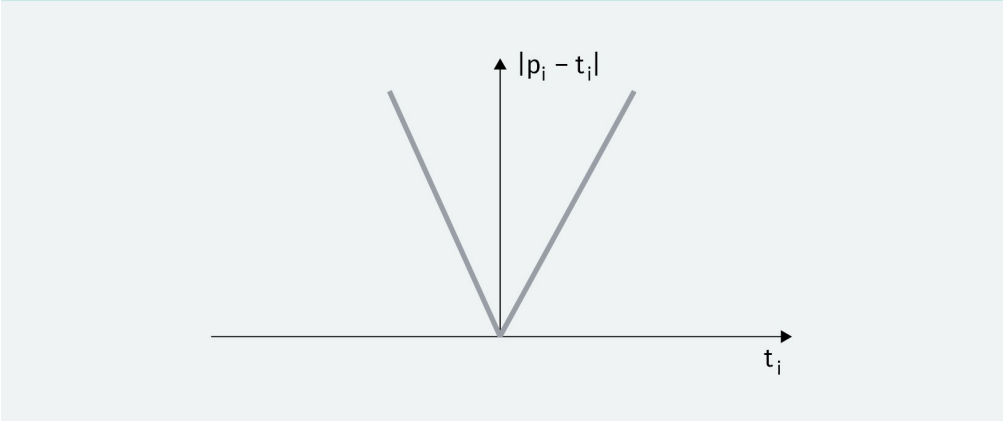
The MAD is defined as:

$$\mathrm{MAD} = \frac{1}{\mathrm{N}} \sum_{\mathrm{i}=1}^{\mathrm{N}} |\mathrm{p_i} - \mathrm{t_i}|$$

It measures the mean of the absolute deviation between the observed event $\mathrm{t_i}$ and the predicted point estimator $\mathrm{p_i}$. The MAD can deal with outliers more easily than other metrics since all points are entered linearly with the deviation between observed and true event into the equation.

It can be shown that the optimal point estimator for the MAD is the median of the predicted probability distribution. In the simple case that the costs are the same for predictions which are too low or too high, the cost function has the following structure:

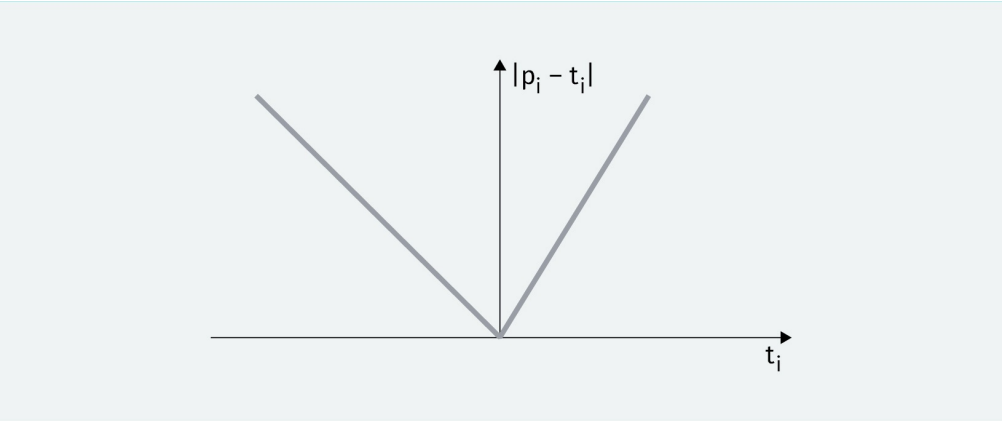
Figure 15: Absolute Cost Function C(|pi – ti|) – Same Slope for Under and Overage Costs



Source: Ulrich Kerzel, 2020.

The slope indicates the penalty associated with a wrong classification. The fact that the MAD is well equipped to handle outliers, and the fact that it is associated with the median of a predicted probability distribution makes the MAD the preferred choice in most practical applications. The MAD can be easily extended by adding different penalties depending on whether the prediction is too low or to high:

Figure 16: Absolute Cost Function C(|pi – ti|) – Different Slopes for Under and Overage Costs



Source: Ulrich Kerzel, 2020.

Using different penalties allows the modelling of simple constraints of the system for which the predictions are made.

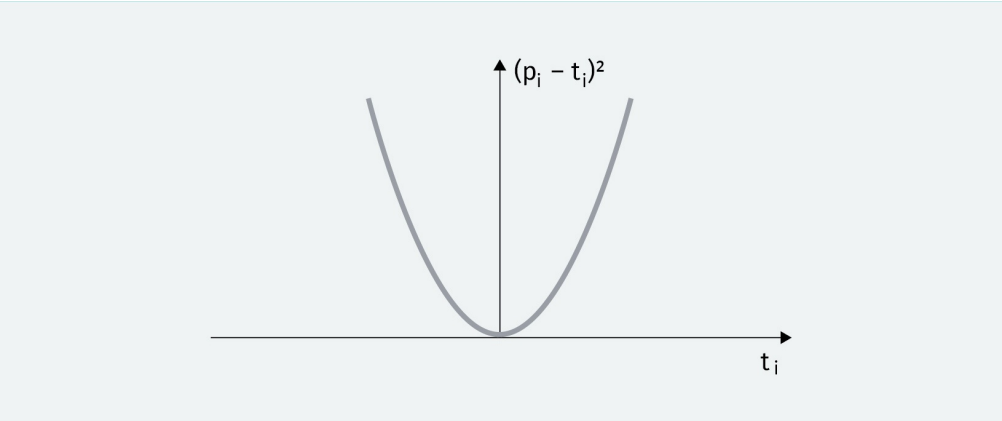
Mean squared error (MSE)

The MSE is defined as

$$\mathrm{MSE} = \frac{1}{\mathrm{N}} \sum_{\mathrm{i}=1}^{\mathrm{N}} \left(\mathrm{p_i} - \mathrm{t_i}\right)^2$$

and averages over the squared deviations of the observed event $\mathrm{t_i}$ and the predicted point estimator $\mathrm{p_i}$:

Figure 17: Quadratic Cost Function C((pi – ti)2)



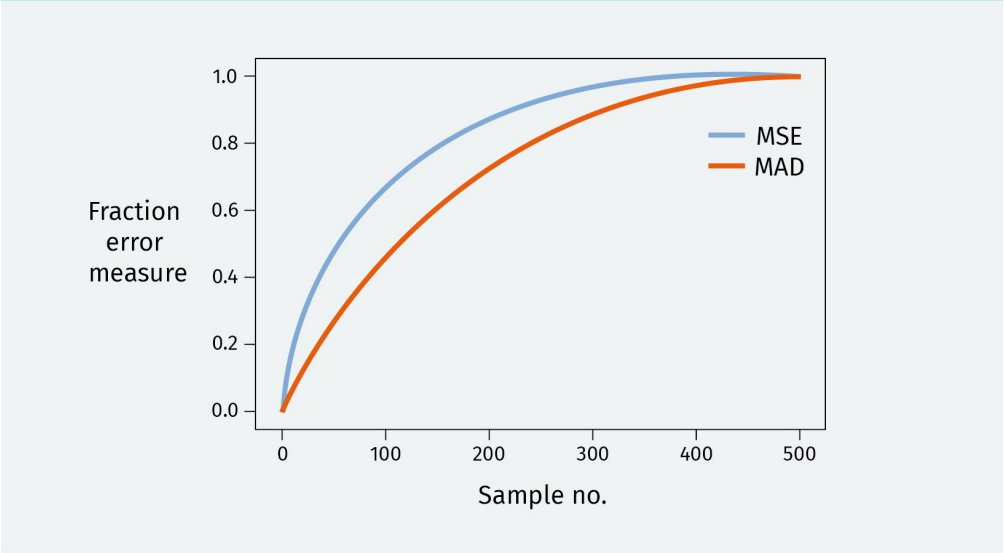
Source: Ulrich Kerzel, 2020.

The main advantage of the MSE is that this metric is associated with the mean of predicted probability distributions. Since many forecasting methods only provide a single number, this is typically assumed to be the mean of a probability distribution. Unlike the MAD, the MSE is very sensitive to outliers due to the quadratic term (Berger, 1985). While this may be beneficial in very specific use cases, this behavior distorts the influence that outliers have on the overall result.

Comparing the behavior of the MSE and MAD

The different behaviors of MAD and MSE are illustrated in the graph below. The "prediction" is taken as the "true" event to which a Gaussian noise term has been added. The resulting errors $e_i = p_i - t_i$ are then put in to descending order. The MAD and MSE are calculated, taking the errors into account one after the other, and the overall value is normalized to the respective value of the MAD and MSE after incorporating all elements:

Figure 18: Comparison Between Mean Absolute Deviation and Mean Squared Error



Source: Ulrich Kerzel, 2020.

As illustrated by the figure, outliers and large deviations contribute to the total value of the MSE much more severely than the MAD. This effectively implies that a significantly smaller sample of test events contributes to the overall value of the evaluation metric in case of the MSE compared to the MAD.

Mean absolute percentage error (MAPE)

The MAPE is defined as

$$\mathrm{MAPE} = \frac{1}{N}\sum_{i=1}^{N}\frac{|p_i - t_i|}{t_i}$$

Except for the denominator, the MAPE is defined in the same way as the MAD. However, each evaluated event is divided by the value of the true observed event t_i. One motivation behind this is that both the MSE and MAD are absolute error metrics, hence their value increases with the number of events observed. From a practical point of view, a relative metric that relates the deviation between the predicted point estimator and observed value to the true event would be preferable. However, dividing by the true observed event, the definition is undefined if $t_i = 0$. For example, a typical supermarket will have many fast moving items which are sold several times a day, hence $t_i >> 0$. However, many items will be slow moving goods which are sold only a few times a day, such as luxury goods. In these cases, $t_i = 0$ and these events cannot be analyzed in a meaningful way. This must then be mitigated by choosing a different metric or aggregating, in which case detailed information is lost.

Root mean squared error (RMSE)

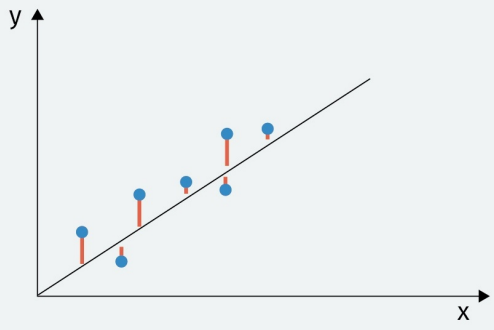
The RMSE is defined as

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(p_i - t_i)^2}$$

which is essentially the same as the MSE, but the square root is taken to compute the final value. Due to their similarities, the RMSE inherits the disadvantages of the MSE, in particular the susceptibility to the influence of outliers.

The behavior of the RMSE can be visualized as the deviation of the residuals from a simple linear regression line in case of an unbiased estimator with Gaussian residuals:

Figure 19: Residuals in Linear Regression



Source: Ulrich Kerzel, 2020.

Mean absolute scaled error (MASE)

The MASE was proposed by Hyndman and Koehler (2006) to overcome the issue that absolute metrics, such as the MAD and the MSE, depend on the scale of the problem, i.e. on the number of samples evaluated. Instead, they suggest normalizing the error $\mathrm{e_i} = \mathrm{p_i} - \mathrm{t_i}$ to the MAD calculated from a naïve forecast, i.e.

$$\mathrm{q_i} = \frac{\mathrm{e_i}}{\mathrm{MAD_i^b}}$$

the MASE is then calculated as the simple arithmetic mean of the above quantity:

$$\mathrm{MASE} = \mathrm{mean}(|\mathrm{q_i}|)$$

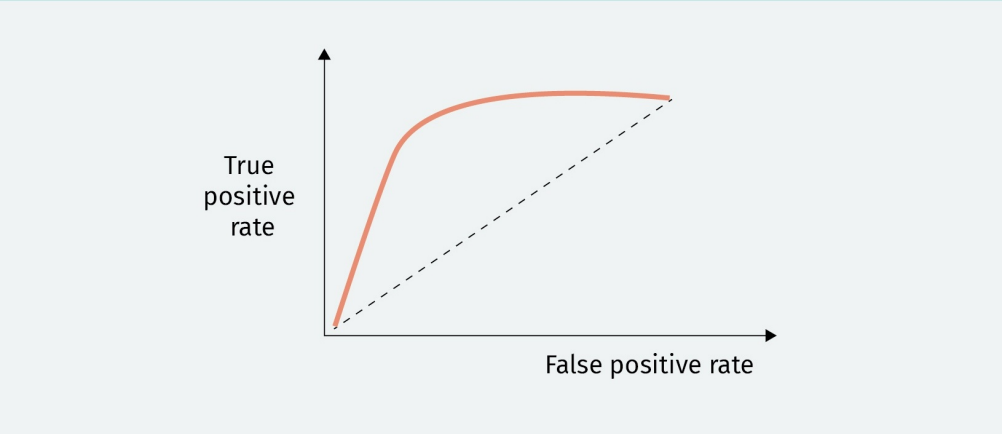
However, there are several issues with this proposed metric. As Fleming & Wallace (1986) point out, using the arithmetic mean for normalized quantities can lead to confusing or incorrect results. Furthermore, the new prediction model is evaluated relative to a naïve forecast. By definition, the naïve forecast is not suitable to predict future events, otherwise it would be used as a regular operation. Hence, the MASE is effectively biased towards the performance and features of the naïve forecasting model, which may distort the evaluation of the performance of a more advanced model under consideration. Finally, the scaled errors $\mathrm{q_i}$ depend on the average deviation computed from the naïve model, hence favoring large deviations in $\mathrm{MAE^b}$ which in turn reduces the scaled errors $\mathrm{q_i}$.

2.2 Visual Aides

Receiver operating Ccurve (ROC)

The receiver operating curve (ROC) (Egan, 1975) is a graphical tool used to illustrate the performance of a prediction model for classification. The value of the x-axis is determined by the false positive rate, i.e. the ratio of wrongly accepted events (false positives) to the number of negative events in the entire sample. The value of the y-axis is determined by the true positive rate. The ROC is obtained by evaluating the false positive rate and the true positive rate for various threshold settings of the prediction model. The diagonal line in the plot is associated with random predictions, i.e. when the model has no predictive power to separate between two classes. The ideal point is on the upper left corner of the plot, (0,1), which means that all events are classified correctly. The closer the graph comes to the ideal point (0,1), the better the prediction model is.

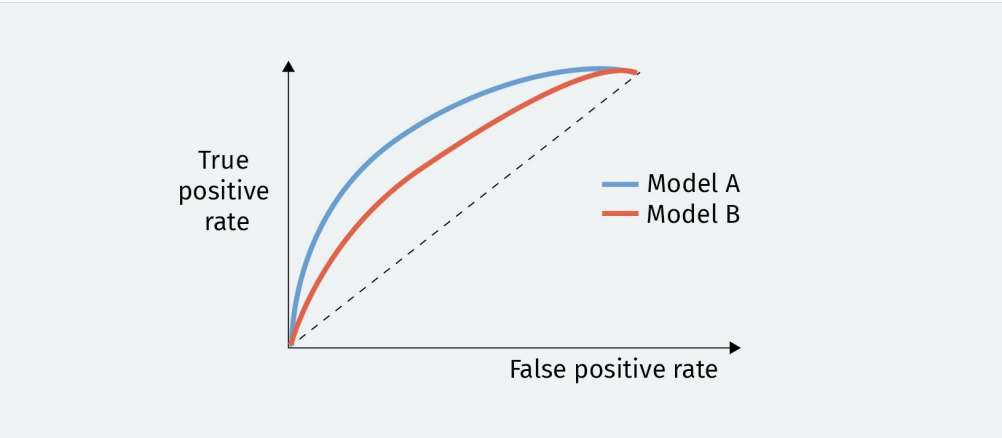
Figure 20: ROC Curve



Source: Ulrich Kerzel, 2020.

The ROC allows the performance of several prediction models to be compared — each model is represented by a single line in the graph as shown in the comparison below. In this case, the blue model is generally better than the red model as it is further away from the diagonal line, indicating random assignment.

Figure 21: ROC Curve Used to Compare Prediction Models



Source: Ulrich Kerzel, 2020.

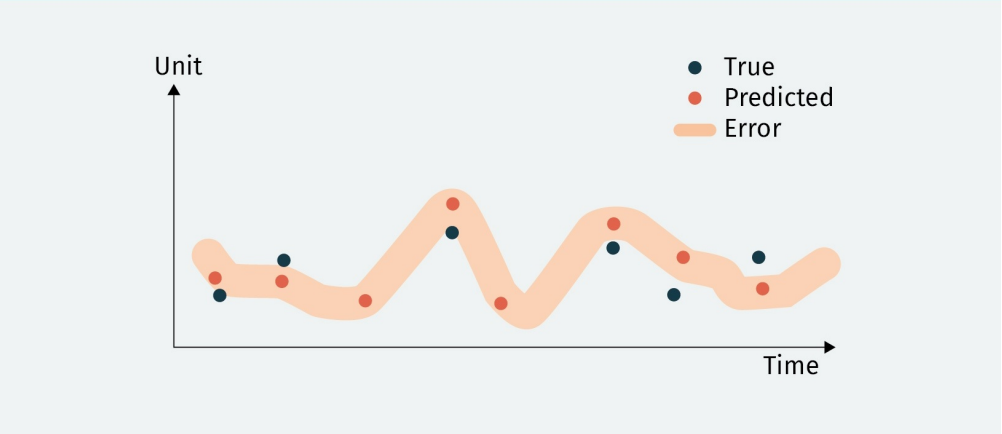
Time series

Time Series
A sequence of numbers or events which are auto-correlated with each other.

In many applications and use cases, predicted and observed values are given as a **time-series**. This means that a given data point depends on what has previously happened. Examples are the price of stocks and shares, the sales of goods, the number of cars on the road, etc.

A graphical representation of the time-series data is therefore a helpful supplement to judge the performance of a predictive model as indicated in the figure below. It combines the predicted point estimator, a measure of the uncertainty associated with each prediction, and the observed true event if available. While this illustration is not sufficient to judge the model quality by itself, it does allow us to judge the overall performance and spot any systematic trends or difficulties with periodic aspects of the data, etc.

Figure 22: Visualization of a Time Series with Prediction, True Values, and Predicted Uncertainties



Source: Ulrich Kerzel, 2020.

Diagnonal plot

In an ideal case, a predictive model should be an unbiased **estimator** of the future or the unknown true event. Since all predictions are essentially random variables described by an underlying probability density function, the predicted value of a particular point estimator is not expected to match the observed event in each case. However, in general the predicted values should reproduce the true events if they are considered on a sufficiently large statistical scale. For example, if we observe that a supermarket sells on average 100 apples, we expect that the predictions will describe this accordingly, subject to some uncertainty. In each individual case, the observed value may be different from the predicted value due to the particularities of the underlying probability density distribution, but, given a large enough ensemble, the prediction should faithfully reproduce the observed trends.

Estimator
A model or method used to calculate the value of unknown or future events based on observed or measured data.

In many cases, the range of values from the predicted and observed events is quite large. For example, an insurance company may receive claims between a few hundred and several hundred thousand dollars. Using a single value such as the average insurance claim does not allow us to judge whether or not the predictive model faithfully reproduces the observed values. Instead, one should use a profile plot. Using the predicted values as the x-axis, the values are binned appropriately. For example, one could choose an equidistant binning or choose the bin borders so that each bin contains the same number of events. The number of events in each bin should be sufficiently high so that an appropriate ensemble can be evaluated. Then, in each bin of the predicted values, the mean of the

true observed values is calculated along with a measure of the volatility, such as the standard deviation of the observed values, the root-means-squared (RMS), or the **full width half maximum(FWHM)**. Different choices are possible, but the volatility should be estimated from the observed data. In case of highly asymmetric distributions, it may be beneficial to introduce asymmetric measures for the volatility as well. The diagonal plot is then constructed so that the mean and spread of the true observed values are shown in each bin of the prediction. Ideally, the resulting graph should feature the mean of the true observed values scattered around the diagonal line, but they should be statistically compatible with the diagonal line, taking the measures of the volatility into account. Any systematic deviations in one or more regions of the diagonal plot indicate potential issues with the predictive model and should be investigated further.

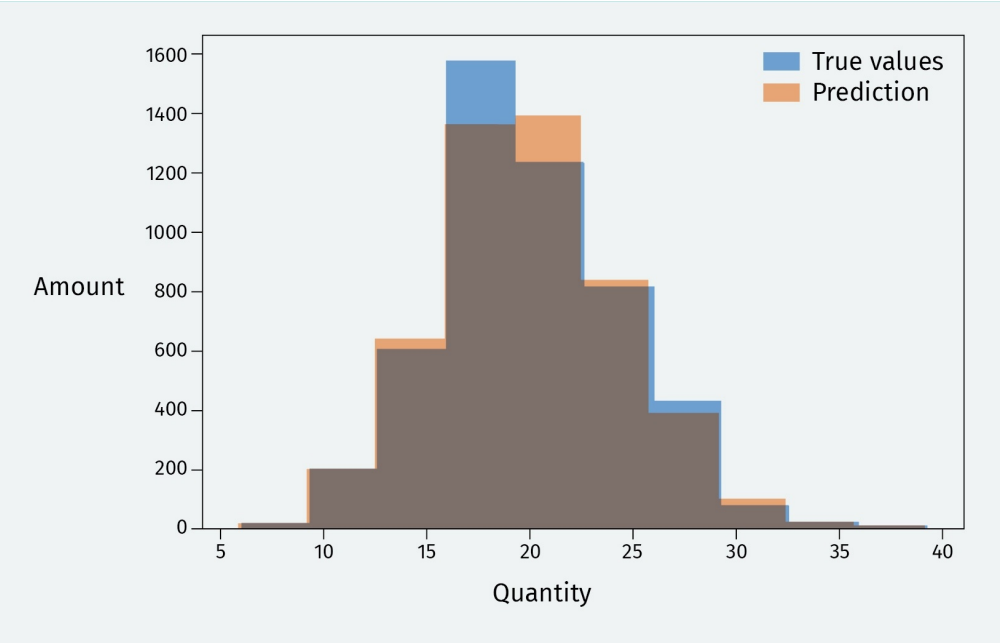
Full Width Half Maximum (FWHM)
Starting from the highest, the points of the distribution where the value is half of the peak value ("half maximum") are identified. The full distance between these points are taken as a measure of the uncertainty.

Example

The following example uses numbers generated from a Poisson probability density function. This function can be used to model discrete events such as the sales of products in a supermarket, etc. The corresponding "predictions" are taken as the value of the true event with some random noise added, modelled by a Gaussian distribution added, i.e. the true events are given by: $y = \mathrm{Poisson}(\mu)$ and the predictions $\widehat{y} = \mathrm{G}(0, \sigma)$, where μ is the mean of the Poisson distribution and σ the standard deviation of the Gaussian distribution centered around zero.

In the first case, a mean $\mu{=}20$ is taken with minimal noise $\sigma{=}0.2$:

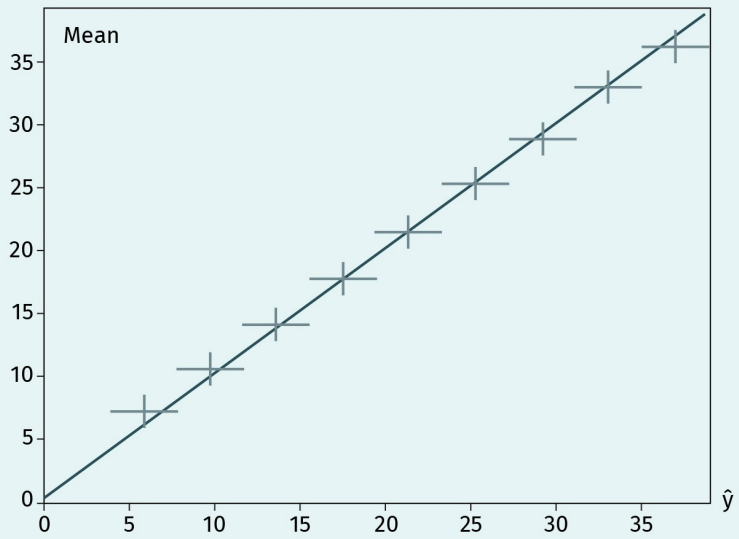
Figure 23: Prediction vs. True Value Using a Poisson Distribution



Source: Ulrich Kerzel, 2020.

The resulting profile plot of predictions vs. the true values is on a perfect diagonal line which has been added to guide the eye.

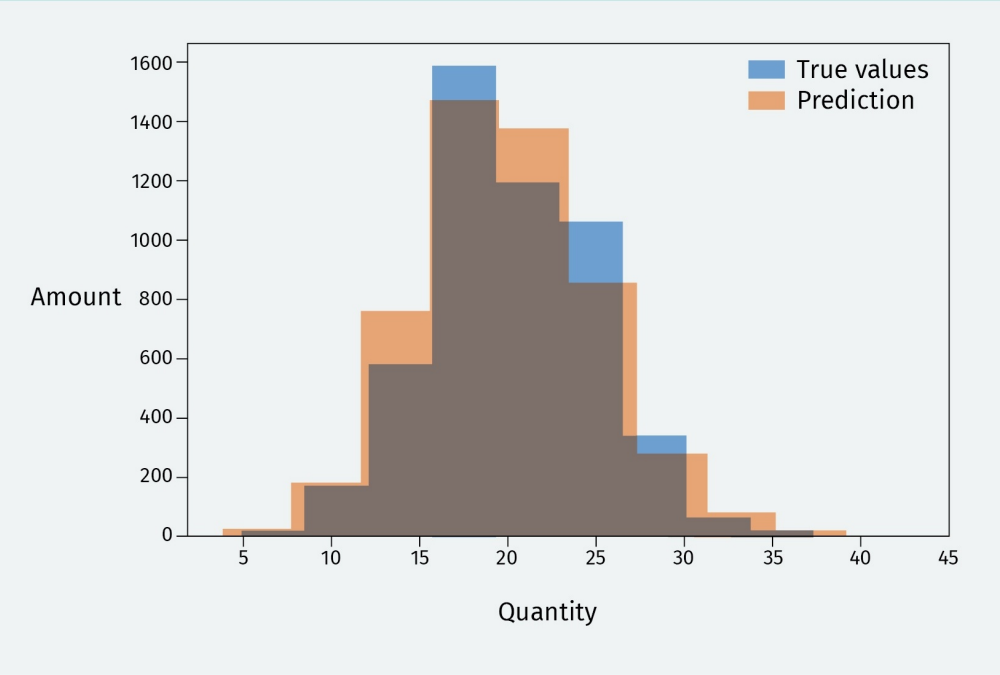
Figure 24: Diagonal Plot for the First Case



Source: Ulrich Kerzel, 2020.

In the second example, the noise term is increased to $\sigma=2.0$:

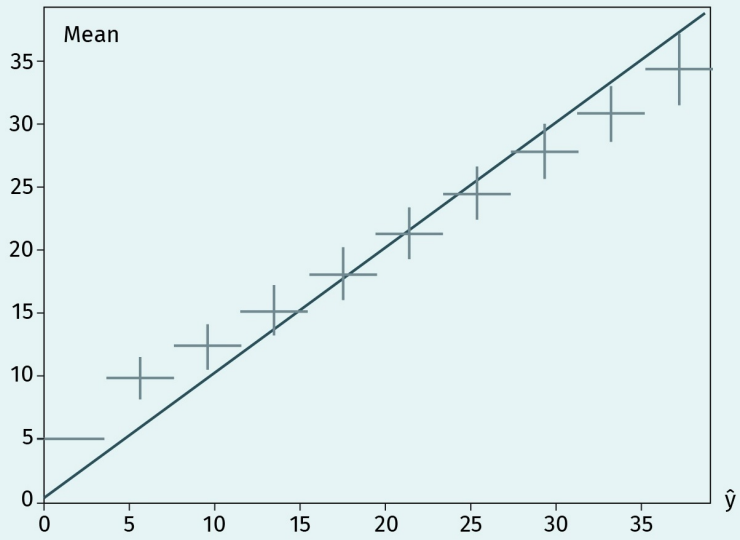
Figure 25: Increasing the Noise



Source: Ulrich Kerzel, 2020.

In this case, the central part of the resulting diagonal plot is still compatible with the diagonal line, however, the bins with few statistics in the tails of the distribution start to deviate. Note that this is purely a statistical effect that comes from adding a large amount of noise.

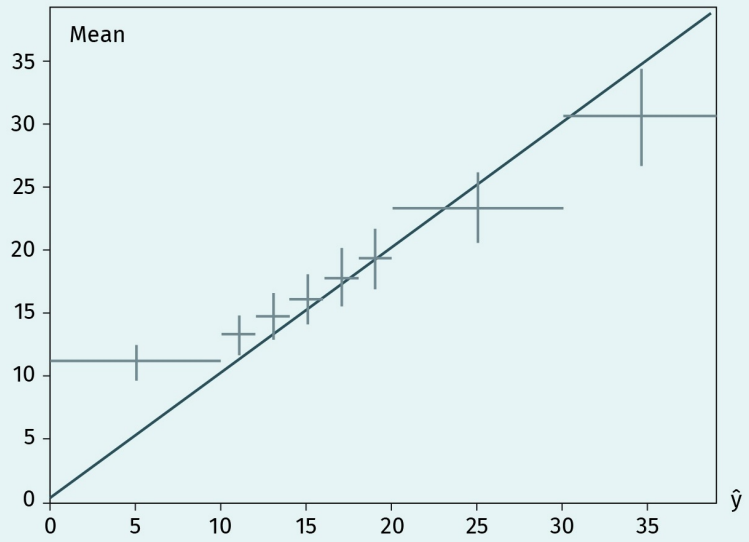
Figure 26: Diagonal Plot with Increased Noise



Source: Ulrich Kerzel, 2020.

The diagonal plot seems to show a systematic bias towards low numbers, however, the effect is less pronounced when you choose a more appropriate non-equidistant binning.

Figure 27: Diagonal Plot with Non-Equidistant Binning



Source: Ulrich Kerzel, 2020.

SUMMARY

Evaluating the performance predictive model is a core part of implementing data science and AI use cases in practical applications. Defining success is an integral part of establishing how data-driven approaches create value in a particular application.

The model centric evaluation focuses on the output of the predictive model and quantifies how well the predictions agree with the "true" outcomes. A wide range of metrics exist for classification, where an event is attributed to a specific category, and for regression, which focuses on the prediction of a quantity.

For classification, the most relevant metrics are precision, sensitivity, and recall, as well as the understanding of type I and type II errors. In the case of classification, the mean absolute error (MAD) is a suitable metric in most cases, although the mean squared error (MSE) can be used if supported by the use case.

Graphs like the ROC curve or confusion matrix aid the understanding of the performance of the predictive model visually. The diagonal plot is a crucial tool to evaluate whether the prediction is unbiased across the range of expected values.

UNIT 3

BUSINESS CENTRIC EVALUATION

STUDY GOALS

On completion of this unit, you will have learned …

- the definition of a cost function.
- how cost functions can be used to evaluate the performance of a prediction.
- the definition of KPIs.
- how KPIs can be used to evaluate the performance of a predictive model from a business perspective.
- how to use an A/B test to determine the performance of a model.

3. BUSINESS CENTRIC EVALUATION

Introduction

The business centric evaluation focuses on the wider implications that a predictive model has on the part of the project that is business related. Of course, project managers would not want to rely on inaccurate predictive models, however, it is not the predictions themselves that are typically used in the business setting, but the operations that are derived from them.

For example, a core operational task for a supermarket is to replenish the goods for customers to buy. Typically, supermarkets offer a range of products which can be stored on the shelf ranging from a few days (e.g. fresh meat, fish, fruits, and vegetables) to many weeks or months (canned or frozen food, drinks, etc.). Each day, an inventory is taken, perished food is removed, and new products are ordered. To accurately estimate the amount of product that needs to be ordered, supermarket managers typically rely on a sophisticated prediction to estimate the demand for a particular item for the next delivery period. The demand is predicted using a complex model, taking in to account a variety of factors such as the past sales of the product, location of the supermarket, promotion and advertisement, day of the week, etc. The predictive model can be evaluated based on historical data, however, for an ordering manager, this is not sufficient — or even relevant. Items cannot be ordered in any quantity, they are typically grouped into larger units with a given lot-size, so not every item can be delivered in each delivery cycle. Hence, the demand prediction needs to be transformed into an operational order. This order is based on the demand prediction, but must take in to account all operational constraints of the supermarket supply chain. Furthermore, as a business, the supermarket managers focus less on the technical quality of the predictive model, but rather on what this implies for their operational excellence.

Ideally, managers aim to maximize profit, so it would be ideal to optimize directly for profit. However, in practical situations, the overall profit is influenced by many factors which are difficult to disentangle, identify, or measure. Hence, in many practical applications, companies typically rely on an agreed set of metrics called key performance indicators (KPIs) which represent how well a particular aspect of the business operates. In most practical applications, a set of these KPIs are used, however, they can contradict each other as optimizing one KPI can deteriorate the other. Coming back to the example of the supermarket, typical KPIs in these settings are the waste rate (i.e. how many products have to be disposed of as they can no longer be sold), inventory (i.e. how many products should be kept in the store) or the stockout rate (i.e. how often customers face an empty shelf). An ordering manager then has to decide how many products to order so that they can improve their on-shelf availability (and hence reduce the stockout risk) and minimize the waste of perished products. Therefore, the demand prediction should not only be evaluated from a technical perspective but also from this business perspective to determine how the operational aspects are related to a “good” or “bad” performance of the predictive model.

Figure 28: Fresh Produce in a Supermarket

Source: Ali, 2018.

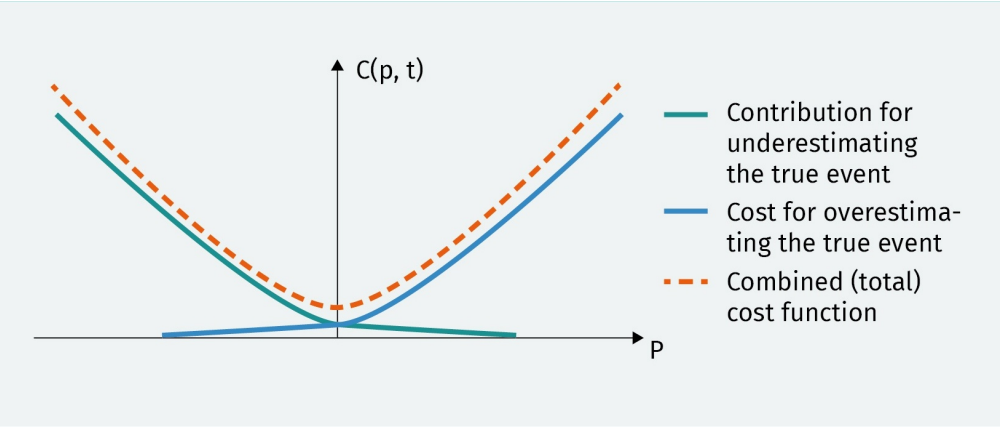
3.1 Cost Function and Optimal Point Estimators

Businesses focus on a number of objectives and implement a variety of strategies to reach them, such as expanding their market share, or developing new products. Above all, companies aim to increase their profit, hence minimizing the operational costs is equivalent to maximizing the profits. In a data-centric enterprise, the **cost function** $\mathrm{C}\ (\mathrm{p_i},\ \mathrm{t_i})$ for an individual product or service depends on the prediction $\mathrm{p_i}$ as well as the corresponding true (observed) event $\mathrm{t_i}$, and summarizes the monetary impact of a prediction which is below or above the true value. For example, if the prediction is too low, emergency orders, which are more expensive, might have to be placed to satisfy customer demand. In some cases, customers may leave, and missed revenue has to be taken into account — in some cases, customers that are angry about the empty shelves may not return, meaning that the loss of future sales must also be considered. On the other hand, if too many products are ordered, this increases inventory and logistics costs. These costs may differ from product to product, however they will generally show a linear dependency as the cost can be attributed to single units of the product or service. For example, warehouse costs rise linearly (within the capacity of the building), as doubling the number of units doubles the space that they occupy inside the warehouse. In general, other costs such as marketing,

Cost Function
A mathematical function which depends on predicted and true (observed) values. The cost function defines a value or penalty based on the deviation of the prediction from the observation.

promotions, etc. should also be included in the cost function. However, many of the elements contributing to the cost structure of a given product are hard to determine in practice. A typical cost function is shown below:

Figure 29: Typical Cost Function



Source: Ulrich Kerzel, 2020.

Evaluating the predictive model in terms of the cost function is beneficial because the predictions are directly related to the operational structures of the business, and the impact of the performance can be directly evaluated in the relevant context.

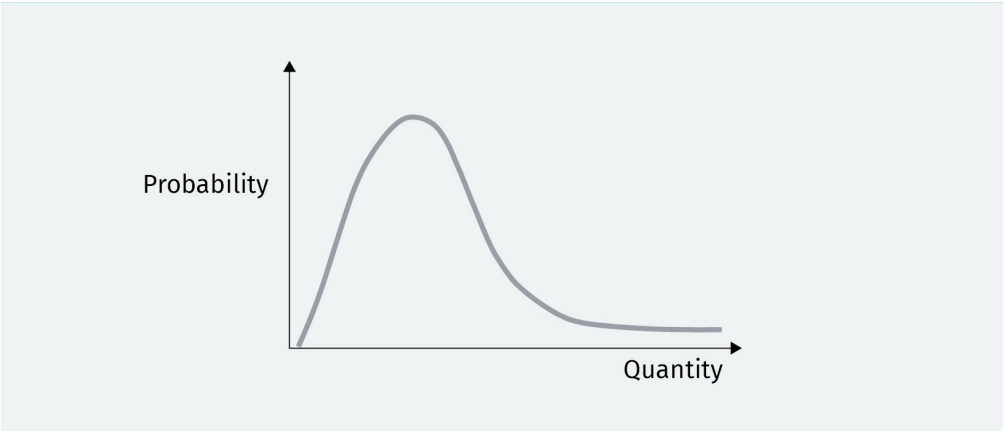
Optimal point estimators

Working with a suitable cost function is also beneficial because the value of the predicted future behavior, which is used for later operational decisions, can be chosen so that it is optimally suited for the individual case.

Recall that predictions are essentially random variables, i.e. predictions made by a predictive model are not a single fixed number, but effectively a probability. For example, if a predictive model must determine whether or not a transaction is fraudulent, the predictive model should not just return a "yes/no" decision, but a probability between 0 and 100 percent indicating how likely it is that a given transaction is fraudulent. Then, depending on the use case and cost structure, a suitable threshold can be set which separates fraudulent and normal cases. The exact level of this threshold depends on the operational requirements. In some cases it may be better to flag events and investigate them manually, whereas in other cases, it would be easier to focus on the cases which can be identified with high probability. In the case of regression, i.e. the prediction of a quantity, the prediction itself should not just be a number (e.g. "tomorrow 3 apples will be sold") or a number with an uncertainty measure (e.g. "tomorrow between 3 ± 1 apples will be sold"), ideally it should be a **probability density distribution** (PDD).

Probability Density Distribution
This reflects the continuous probability for each occurrence of a random variable.

Figure 30: Prediction of Future Sales as a Probability Distribution



Source: Ulrich Kerzel, 2020.

The prediction of a full probability density distribution for each individual event contains all available information and covers all possible outcomes. For each value of the predicted quantity, the associated probability shows how likely this outcome is. Predicting a single number does not take in to account any information about the expected uncertainty or volatility. The shape of the true underlying probability density distribution is, in general, asymmetric. This can be caused, for example, by physical boundaries, e.g. one cannot sell less than zero apples whereas there is no fundamental upper limit — in principle, one can sell as many apples as possible. If a predictive model predicts a single number and its uncertainty (e.g. "3 ± 1 apples"), this asymmetric shape can no longer be captured, and one effectively assumes a specific probability distribution such as the normal distribution.

Predicting a probability density distribution may capture all available information, but, like in the example of fraud detection, we must translate this prediction into an actionable number. The ordering manager does not want to know about the probability distribution for the demand for individual products, but rather about the best estimate to base the order on. This estimate, the optimal point estimator, can be obtained from the cost function. The best single number p that can be used in operational decisions is the one which minimizes the cost function:

$$\frac{\partial \mathrm{E}[\mathrm{C}[\mathrm{p},\mathrm{t}]]}{\partial \mathrm{p}} = 0 \quad \text{and} \quad \frac{\partial^2 \mathrm{E}[\mathrm{C}[\mathrm{p},\mathrm{t}]]}{\partial \mathrm{p}^2} > 0$$

For any given cost function C, the optimal point estimator p can then be derived. Ideally, all relevant contributions are contained in a realistic estimate of the cost function. In many cases, this will not be possible in practice as the individual contributions are too difficult to obtain. However, in this case, one can build a model or simplified cost function based on reasonable assumptions of the general cost structure.

Example: Linear cost function

In many cases, the cost function can generally be expected to show a linear behavior. The simplest cost function within this assumption is then given by $\mathrm{C(p,\ t)} = |\mathrm{p} - \mathrm{t}|$, i.e. the absolute deviation of prediction p and observed true event t. Note that this is the same as the mean absolute deviation (MAD).

The optimal point estimator is then given by:

$$\begin{aligned} 0 &= \partial_{\mathrm{p}} \mathrm{E[C(p,t)]} \\ &= \partial_{\mathrm{p}} \int_{-\infty}^{\infty} \mathrm{f(t)|p-t|dt} \\ &= \partial_{\mathrm{p}} \left(\int_{-\infty}^{\mathrm{x_q}} \mathrm{f(t)|p-t|dt} + \int_{\mathrm{x_q}}^{\infty} \mathrm{f(t)|p-t|dt} \right) \\ &= \partial_{\mathrm{p}} \left(\int_{-\infty}^{\mathrm{x_q}} \mathrm{f(t)(p-t)dt} - \int_{\mathrm{x_q}}^{\infty} \mathrm{f(t)(p-t)dt} \right) \end{aligned}$$

Here, $\mathrm{f(t)}$ is the probability density distribution that the true value follows, and the expectation value E is given by the convolution $\mathrm{E[g[x]]} = \int \mathrm{f(x)g(x)dx}$. In the last step above, the first integral t is smaller than p (and hence the absolute value is positive) and in the second integral t is larger than p which yields an additional factor, –1. Executing the partial derivation with regard to p then gives:

$$\begin{aligned} 0 &= -\int_{-\infty}^{\mathrm{x_q}} \mathrm{f(t)dt} + \int_{\mathrm{x_q}}^{\infty} \mathrm{f(t)dt} \\ \int_{-\infty}^{\mathrm{x_q}} \mathrm{f(t)dt} &= \int_{\mathrm{x_q}}^{\infty} \mathrm{f(t)dt} \end{aligned}$$

This means the number $\mathrm{x_q}$ splits the integral in half, i.e. the value of the integral from $-\infty$ to $\mathrm{x_q}$ is the same as from $\mathrm{x_q}$ to ∞, which is the definition of the median, i.e. the median is the optimal quantile of the predicted probability density function or optimal point estimator in case of a linear cost function.

3.2 Evaluation Using KPIs

Critical Success Factors (CSF)

For companies to be successful, they need to excel in one or more areas and also offer a unique value proposition which cannot be easily replicated by their competition, or execute a particular offering or business model exceptionally well so that customers will use this company as their preferred provider. These critical success factors (CSFs) are at the very core of a business. The Australian government's specialist business program delivery division (AusIndustry) defines the CSFs as a "list of issues or aspects of organizational performance that determine ongoing health, vitality, and well-being" (AusIndustry, 1999). For example, the CSFs for a supermarket could be the availability of goods on the shelf (avoiding stockout situations) or the freshness of produce, dairy, and meat products. Timely arrival and departure of trains are central to the operation of a train station. A news agency needs to provide accurate and timely information about current events, etc.

Parmenter distinguishes between operational CSFs, which are focused on the internal processes of a company, and external outcomes which are concerned with what the company provides to its customers (Parmenter, 2015, p. 162).

It is important that the CSFs are specific, quantifiable, and don't describe general statements. Although platitudes are tempting because everyone can agree with them, they don't offer any guidance on how to achieve set goals. For example, the sentence "we need to become (more) profitable" is true for most companies but it is unclear how to do this. Similarly, "we need to increase customer satisfaction" doesn't address the way that this goal should be achieved or measured. Parmenter (2015, p. 171) follows the "SMART" approach (Doran, 1981) and points out that CSF should be:

- Specific: Avoid empty phrases and focus on specific aspects.
- Measurable: The CSFs can be used to derive a measurable quantity. For example, instead of "improving customer satisfaction," a company could focus on reducing the number of complaints they receive.
- Achievable: Although this sounds obvious, it is important to note that CSFs need to be achievable in practice otherwise they will be ignored.
- Relevant: The CSFs must be relevant to the operational and daily tasks of the employees of the company. The management of the company cannot expect the employees to align their actions to the CSFs if it is not clear who is responsible for the overall goal during day-to-day work.
- Time sensitive: CSFs should be focused on the immediate and near future. Although strategic long-term goals may play a vital role in the definition of the CSFs, the success factors themselves should be relevant to the operational tasks.

What are key performance indicators (KPIs)?

What Are Key Performance Indicators (KPIs)?

In a nutshell, key performance indicators (KPIs) are the measurable quantities which translate the critical success factors (CSFs) into objective metrics. Once the CSFs are correctly determined, the KPIs should be easily identified.

The main aim of the KPIs is to measure on an ongoing basis how well the company fulfills the CSFs. As the KPIs are highly relevant to the well-being and operational excellence of the company, the progress made towards achieving them should be directly reported to the **senior management team**, as well as to the CEO on a daily basis. Since the KPIs are reported directly to the highest echelon of a company, they should fulfill the following characteristics (Parmenter, 2015, p. 206):

Senior Management Team
High level managers who are responsible for the well-being of a company. Includes C-level suite (CEO, CTO, CFO, etc.) and heads of business units.

- No more than ten organization-wide KPIs: Although it is tempting to report everything to the CEO, this can result in information overload, meaning that the important information is buried in a mass of details. The KPIs should be used to summarize the current status and highlight the status-quo of the company. Often, companies measure a wide range of factors and report them to the senior management team and CEO. In order to be able to focus on the most relevant metrics, the senior management team may need to abolish old metrics and reports in order to start from a clean slate.

- Measured frequently: The KPIs need to be measurable and should be measured as often as the team or (automated) mechanism can react to the change. For example, it is not helpful to measure the value of a KPI in real-time if the KPI is only impacted once every full hour, etc. On the other hand, if it could be impacted at any time, measuring the value of the KPI only once per hour wastes precious time in which the team or algorithm could have intervened.
- Actionable: Measurements of KPIs need to clearly show whether or not the values are sufficient. This implies that nominal ranges of KPI values must be defined beforehand. They should indicate whether the current measurement of the KPI is acceptable or if action must be taken. This also means that it must be clear which deviation from the allowed range of KPI values triggers which action. In particular, employees of a company must be able to understand the KPI and how their actions can influence its value.
- Responsibility and accountability: Ultimately, a team or business unit needs to be responsible for a KPI. The CEO needs to be able to call a manager or team-lead who can describe the current measurement of the KPI, and explain who can do what to improve the situation if the KPIs are outside their nominal range. A team, division, or business unit must be able to take ownership for a specific KPI or set of KPIs and ensure that the value stays in the agreed acceptable ranges.

For example, two CSFs for a supermarket chain could be the amount of food that must be thrown away, and the on-shelf availability of goods in the store. The associated KPIs could then, for example, be the waste rate, i.e. the amount of food which has to be disposed of, and the number of gaps on the shelf during opening hours. As both KPIs affect each other (e.g. increasing the amount of goods on the shelf will improve the availability but lead to a higher waste rate), the possible values of both quantities and their dependency on each-other have to be established first. Then, allowed ranges should be defined, e.g. a waste-rate between ten and 30 percent and an on-shelf availability between 80 and 100 percent. In most practical scenarios, these KPIs can be measured daily and reported to the senior management team, the operational team responsible for the procurement of new goods, and the operational staff in the shops.

The above discussion uses the phrase “key performance indicators” when referring to measurements reported to the senior management team exclusively. In other interpretations, the “key” part of KPI is understood to be a core metric of the project or responsibility of the team. Hence, KPIs can be introduced at any level of hierarchy.

Types of indicators

As discussed above, an organization should not have more than 10 key performance indicators (KPIs) which are regularly reported to the senior management team.

However, individual business units, divisions, and project teams may need a wider range of indicators to measure their own performance. Since the overall number of KPIs is limited, it is unfeasible to run an entire business unit on one KPI. Also, some departments such as HR or accounting may not have KPIs at all but are still critical to the success of the company.

In order to address this situation and allow for a wider range of reporting option, Parmenter (2015, p. 3) suggests to use the following:

- key performance indicators,
- performance indicators,
- key result indicators, and
- result indicators.

Hence, Parmenter (2015) discriminates between result and performance indicators, as well as “key” and “normal” indicators. As discussed previously, the “key” indicators are immediately relevant for the senior management team and the CEO, and should be directly reported to them. The “normal” indicators are still relevant, but may be more detailed and may focus on specific aspects of the operations of a company. Hence, they should be reported to the team-leader or manager of a particular business unit. Parmenter makes a clear distinction between metrics intended for the senior management team and metrics for all other levels of the hierarchy. Following the previous discussion, performance indicators are immediately connected to the critical success factors of a company and are, in a simplified way, the measurable quantities associated with a specific CSF. Result indicators, on the other hand, summarize the efforts of several teams or business units, for example, employee satisfaction, EBIT, etc. They are not directly connected to the critical success factors of a company but are relevant to its well-being and status-quo. For example, net profit or EBIT are very important to the overall performance of a company and the senior management team needs to be aware of them at all times. However, these, and similar measures, are not tied to specific teams but are the responsibility of all divisions. For example, the CEO cannot simply call the team-lead of team A and say, for example: “The profitability of our company is low, fix it,” — each team needs to contribute to the overall goal.

When can KPIs work?

KPIs are a vital instrument used to monitor the well-being and status-quo of a company, and to indicate if and when action needs to be taken. However, in order for KPIs to work as intended, the company must put a corresponding framework in place.

First of all, KPIs should be derived from the critical success factors, i.e. the KPIs should measure things that are directly relevant to the overall well-being of the company. Only the metrics associated with the CSFs should be measured and reported. Often, a wide range of metrics are being measured without a clear connection to the way that they can help the company. Rather than adding more and more metrics over time, the senior management team should decide on a few critical KPIs and abolish all other metrics, measurements, and processes that are ineffective. This does not mean that the CEO should abolish everything which does not directly contribute to a given metric, but they should critically question why any given quantity is measured and reported, and how it is connected to the overall critical success factors or the goals of the company.

One of the most effective aspects of a good KPI is that employees are empowered to act on it. As a team or division takes ownership of a KPI, they must also be given the opportunity to execute various actions to ensure that the KPIs are within the approved ranges.

This in turn requires the senior management team and CEO to give control, within a given boundary, to the team or division. For example, if the KPIs for a supermarket are defined as the fraction of food that must be disposed of and the number of gaps on the shelf, the procurement team needs to be able to order the right amount of goods and the teams responsible for the shop floor need to be able to transfer the goods from the delivery bay to the shop floor.

Gaming the system — Unintended consequences of KPIs

KPIs are defined in order to help a company or business to focus on what makes them successful. This can be summarized with the slogan "what gets measured gets done." It is important to know that KPIs are a tool to enable the realization of a goal (meeting the requirements of the Critical Success Factors), but that in the day-to-day work, employees will be responsible for keeping the value of the KPIs within the allowed range. If the KPIs are not carefully designed and implemented, it will lead to unintended consequences. Parmenter uses the example of hospital staff in the emergency department: The KPI was defined as the "average time to treat patients" (Parmenter, 2015, p. 44). As the patients arriving by ambulance had not been waiting long, they were kept waiting while other patients in the waiting room were treated with a higher priority. However, in general, patients arriving by ambulance have a more severe condition, and by delaying their acceptance into the hospital, the severity of the condition was not taken into account — even though the KPI improved, people's health was put at risk. Before a new KPI is introduced (or an existing one kept), managers should critically reflect on the possible consequences of the KPI, and how to mitigate any adverse effects.

Common fallacies and best practices

The following aspects should be considered when using Key Performance Indicators:

- Too many KPIs make it difficult to focus on the crucial aspects. The more KPIs are introduced, the higher the risk that they may contradict each other.
- Abolish what hasn't been successful. Keeping KPIs because they have always been there is not a good starting point for a successful evaluation. Instead, critically consider why any KPI is there, what it measures, and which actions are associated with it.
- An allowed range needs to be defined for any KPI to indicate when an action needs to be taken.
- Teams or divisions need to take ownership of one or more KPIs and should be held accountable for them. This also implies that these teams and divisions need to be given the power and budget to act accordingly.
- "What gets measured gets done" — KPIs are both a blessing and a curse. If KPIs are not carefully designed, unintended consequences may jeopardize the project or company.

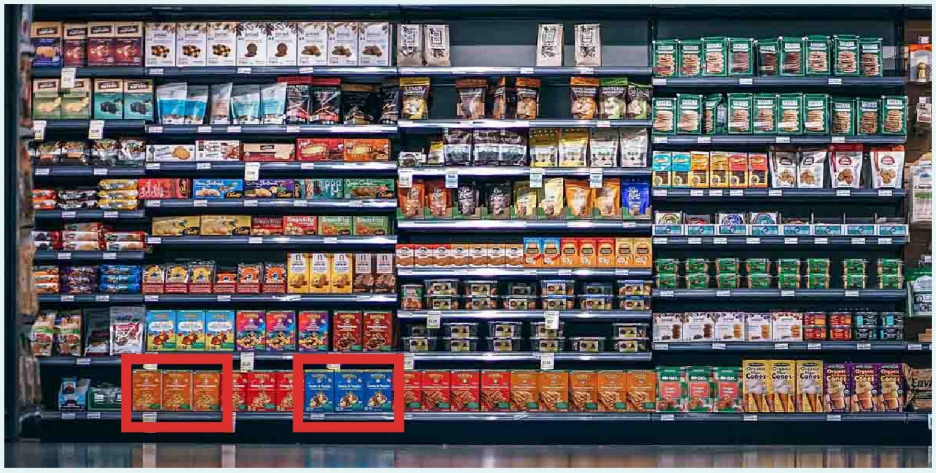
3.3 A/B Test

A/B tests are used to determine which of two variants perform better. A/B tests can be used in a variety of scenarios. For example, when designing a new webpage, one variant could be shown to a set of randomly chosen visitors and a different variant to another set of visitors. Using a suitable metric, it can then be decided which design is more "successful" based on e.g. the average time a user spends on the webpage, how many sub-pages are opened, if a specific link is clicked, or what is bought in the online shop, etc.

A/B tests can be used whenever the outcome of two alternative approaches should be evaluated. For example, A/B testing can also be used to determine the performance of a new algorithm which automatically optimizes the prices of goods as indicated in the picture below. The "old" way of determining the optimal price is applied to the red product and the "new" algorithm to the blue product. Assuming the products are comparable, the sales record of the two products can be compared to determine which pricing approach is more successful.

It should be emphasised that an A/B test compares the outcomes of two valid and functional campaigns, it is not a useful approach when attempting to identify issues in the business model, operational issues, or general management issues.

Figure 31: Products in a Supermarket Used for an A/B Test



Source: Ulrich Kerzel, 2020.

Prerequisites for an A/B Test

In order to perform an A/B test, two independent approaches are executed simultaneously. In some cases, this is relatively easy to implement in practice. Coming back to the example of a web-page or online shop, one half of the visitors gets to see one version, the other half, the other version. In case of physical products or objects, this is often not as trivial. In order to be able to evaluate both approaches at the same time, two comparable

objects have to be found. This could be similar products with very similar characteristics for which customers do not show a strong brand loyalty, or two different sales locations which are very similar with regard to neighborhood, size, etc. Finding two comparable groups is often quite challenging in practice as the two objects must be directly comparable otherwise the A/B test will be biased. Furthermore, the overall frequency at which the objects are presented to customers or visitors has to be sufficiently high. As the A/B test will only run for a limited time, a large enough sample has to be collected in order to be able to draw any conclusions from the test.

Performing an A/B test

Once the two objects or samples have been chosen, the A/B test can be executed. Approach A is applied to object A and approach B is applied to object B. After a specified amount of time, the samples are swapped, i.e. approach A is used for object B and approach B for object A to eliminate any remaining bias. However, the times that the first and second sample were taken can still play a major role. For example, imagine that the objects A and B are Valentine's chocolates — if the first testing period is just before Valentine's day and the second just after, the two samples will not be comparable.

In general, great care should be taken to avoid any confounding variables entering the A/B test and distorting the result, such as:

- launching or stopping marketing campaigns or promotions,
- special events (e.g. Easter, Christmas, Valentine's Day),
- major news about the company (e.g. new products are being launched or old ones discontinued, positive or negative reporting in the media and press, etc.),
- redesign of the website, shop, or marketing material, or
- prolonged periods of unavailability (e.g. products cannot be shipped, the webpage or e-commerce store is not accessible, etc.).

In addition, external factors may have to be taken into consideration, such as the weather or seasonal changes for seasonal products.

A key aspect of performing an A/B test (or any experiment) is to run the test until the predefined time-window ends. Ending the test early introduces an arbitrary cut-off and likely leads to a bias once the "desired" or "obvious" result is seen. An exception are medical tests where the well-being of patients may be at risk if a test is not stopped when severe side-effects are noticed.

Disadvantages of an A/B test

Although A/B tests can be a powerful way to evaluate two alternative approaches, they come at a significant cost. First of all, suitable products or objects which can be used in the test must be identified. Then it has to be accepted that revenue will be lost because either A or B will perform worse than the other. From an operational aspect, further costs arise from these tests as both approaches have to be executed, maintained, and monitored. Finally, all data must be recorded and analyzed once the tests are concluded.

Furthermore, for an A/B test to be conclusive, a substantial amount of data must be recorded, implying that the system being tested generates lots of data (e.g. a website with high traffic or fast-selling products) or has to run for a long time.

Evaluating an A/B test

The easiest evaluation of an A/B test is to calculate the uplift, which is defined as the relative change in success rate between the two approaches:

$$\text{uplift} = \frac{\mathrm{R_a} - \mathrm{R_b}}{\mathrm{R_a}}$$

where $\mathrm{R_a}$ is the success ratio of method A and $\mathrm{R_b}$ the success ratio of method B.

Example:

An e-commerce owner wants to improve the advertisement for their web shop so they start a new ad-campaign with a marketing agency. To prove that the new campaign works, the owner of the web shop suggests an A/B test comparing the current ads on the web with the new one. Each ad gets shown to potential customers (“impressions”) who can then click on the ad to get to the web shop (“conversion”) — or not. The rate of impressions and conversions for both approaches is monitored for a few weeks to decide whether the new ad campaign is better than the old one.

For example, after 6 weeks we could have the following results:

- current campaign: 15700 impressions, 30 conversions
- new campaign: 16000 impressions, 50 conversions

Is the new campaign better? We can calculate the uplift based on the conversion rate, i.e. the ratio of conversions over impressions:

$$\text{uplift} = \frac{\mathrm{CR_{new}} - \mathrm{CR_{current}}}{\mathrm{CR_{current}}}$$

where CR is the conversion rate, i.e. the number of times the ad was clicked (“conversion”) divided by the number of times the ad was shown (“impression”).

In the above example, the uplift is 0.63 or 63 percent, i.e. the new campaign works better than the current approach.

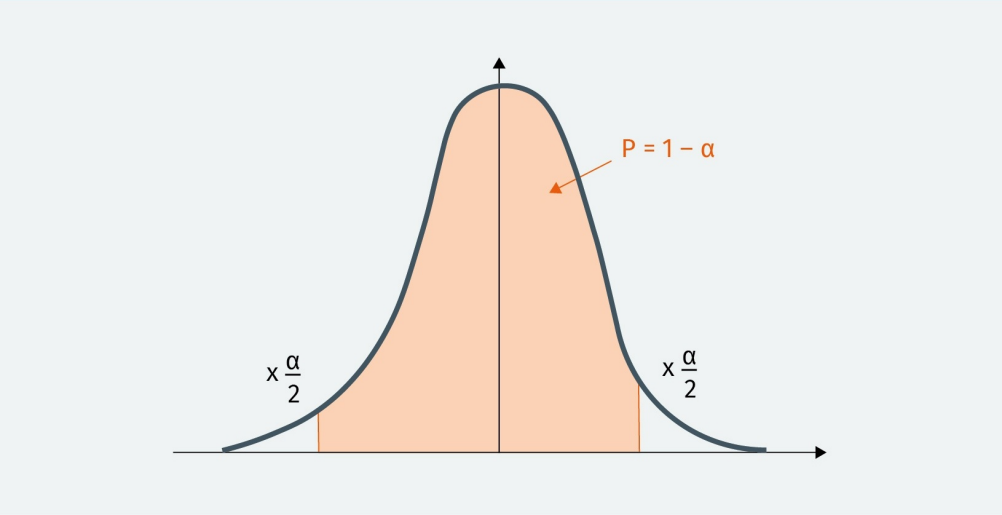
Evaluating the significance of the A/B test in counting experiments

After the A/B test has been concluded, the subsequent analysis aims to determine whether approach A or approach B works better.

The following approach is valid for A/B tests that are also counting experiments, e.g. how often has a link been clicked during the test, a product been bought, etc. This means that there are only two possible outcomes: the action was taken or not. Assuming that the probability for success is the same for each testing period and the trials are statistically independent, this can be treated as a Bernoulli experiment for a random variable X with outcome $\mathrm{X} = 1$ (e.g. link clicked) or $\mathrm{X} = 0$ (link not clicked). The distribution of X can be described by the binomial distribution which depends on only one parameter p defining the probability of success. Hence $\mathrm{P}(\mathrm{X} = 1) = \mathrm{p}$ and $\mathrm{P}(\mathrm{X} = 0) = 1 - \mathrm{p}$. The expectation value of the binomial distribution is given by $\mathrm{E}[\mathrm{x}]=\mathrm{p}$ and the variance by $\mathrm{var}[\mathrm{x}] = \sigma^2 = \mathrm{p}(1 - \mathrm{p})$. If the sample size is sufficiently large so that the central limit theorem holds, the confidence interval can be defined in the following way: The estimator $\widehat{\mathrm{p}}$ for the sample proportion is $\widehat{\mathrm{p}} = \frac{\mathrm{X}}{\mathrm{n}}$, i.e. the number of successes observed in the sample. Hence $\mathrm{E}[\widehat{\mathrm{p}}] = \mathrm{E}[\mathrm{X}/\mathrm{n}] = \mathrm{np}/\mathrm{n} = \mathrm{p}$ and $\mathrm{var}[\widehat{\mathrm{p}}] = \mathrm{var}[\mathrm{X}/\mathrm{n}] = 1/\mathrm{n}^2\mathrm{var}[\mathrm{x}] = 1/\mathrm{np}(1-\mathrm{p})$. Subtracting the mean and scaling by the standard deviation of $\widehat{\mathrm{p}}$ results in a normal distribution if the sample size is sufficiently large and the probabilities are not too close to the end values. The confidence level α can then be calculated as the two-sided probability mass

$$\mathrm{P}\left(-\mathrm{z}_{\alpha/2} < \frac{\widehat{\mathrm{p}} - \mathrm{p}}{\sqrt{\mathrm{p}(1-\mathrm{p})/\mathrm{n}}} < \mathrm{z}_{\alpha/2}\right) = 1 - \alpha$$

Figure 32: Confidence Level for a Two Sided Probability Density Distribution



Source: Ulrich Kerzel, 2020.

$$\frac{\widehat{\mathrm{p}} - \mathrm{p}}{\sqrt{\mathrm{p}(1-\mathrm{p})/\mathrm{n}}} = \pm\, \mathrm{z}_{\alpha/2}$$

And hence:

$$\mathrm{p} = \widehat{\mathrm{p}} \pm \mathrm{z}_{\alpha/2}\sqrt{\mathrm{p}(1-\mathrm{p})/\mathrm{n}}$$

Approximating $\mathrm{p} = \widehat{\mathrm{p}}$ on the right hand side (Wald approximation) gives

$$\mathrm{p} = \widehat{\mathrm{p}} \pm \mathrm{z}_{\alpha/2}\sqrt{\widehat{\mathrm{p}}(1-\widehat{\mathrm{p}})/\mathrm{n}}$$

for a 95 percent confidence interval (i.e. α=5), $\mathrm{z} = 1.96$ in the Gaussian approximation.

Bayesian A/B testing

The approaches to A/B testing that have been discussed so far are based on the frequentist approach to statistics which treats the results as numbers obtained from a repeatable experiment, and the resulting metric, such as the uplift, is a single number. In a **Bayesian** approach, everything is based on probability density distributions. While this may seem complicated at first, it has the benefit of computing probabilities instead of simple numbers.

Bayes
Reverend Thomas Bayes (18th century) created a school of statistics that uses conditional probabilities and evidence to calculate the probability of unknown events.

One of the most important cases for A/B testing is the analysis of tests with two potential outcomes, for example, if a user clicked on an advertisement on a webpage or not — or more generally, "success" vs. "failure" or "yes" vs. "no", etc.

These kinds of scenarios are called "Bernoulli trials," and represent a random experiment with only two different outcomes. If these experiments are repeated n times, they can be described by the binomial distribution, assuming that each experiment is statistically independent from the previous ones. For example, tossing a coin n times results in n statistically independent data-samples (head or tail) as the coin does not retain some sort of "memory" between the throws.

The binomial distribution is given by:

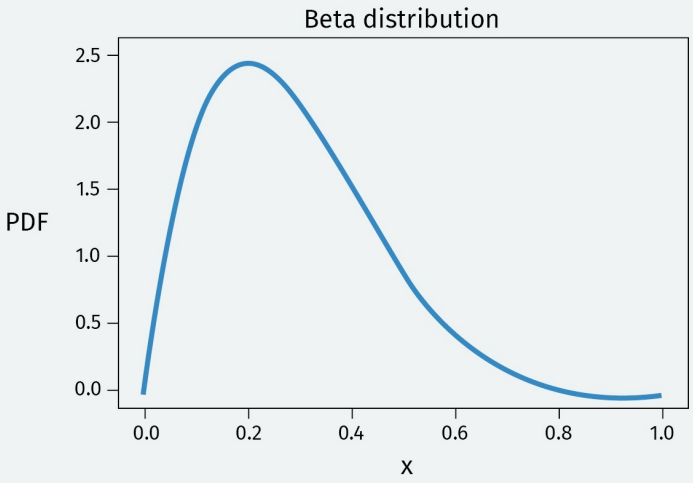
$$\mathrm{P}(\mathrm{k}) = \binom{\mathrm{n}}{\mathrm{k}}\mathrm{p}^{\mathrm{k}}\mathrm{q}^{\mathrm{n}-\mathrm{k}}$$

where p is the probability of a true result, q=1–p is the probability of a false result, and $\mathrm{P}(\mathrm{k})$ is the probability of observing k "successes", e.g. the coin comes up head or the visitor clicked on the advertisement.

In order to work in the Bayesian framework, the prior needs to be included. The **conjugate prior** for the Bernoulli, binomial, negative-binomial, and geometric distribution is given by the beta distribution. The beta distribution is a continuous probability density distribution defined in the interval [0, 1] and has two positive parameters controlling its shape:

Conjugate Prior
In Bayesian statistics, prior and posterior distributions are called conjugate distributions if they belong to the same family of distributions. In this case, the prior is the conjugate prior of the likelihood function.

Figure 33: Beta Distribution Where α = 2 and β = 5



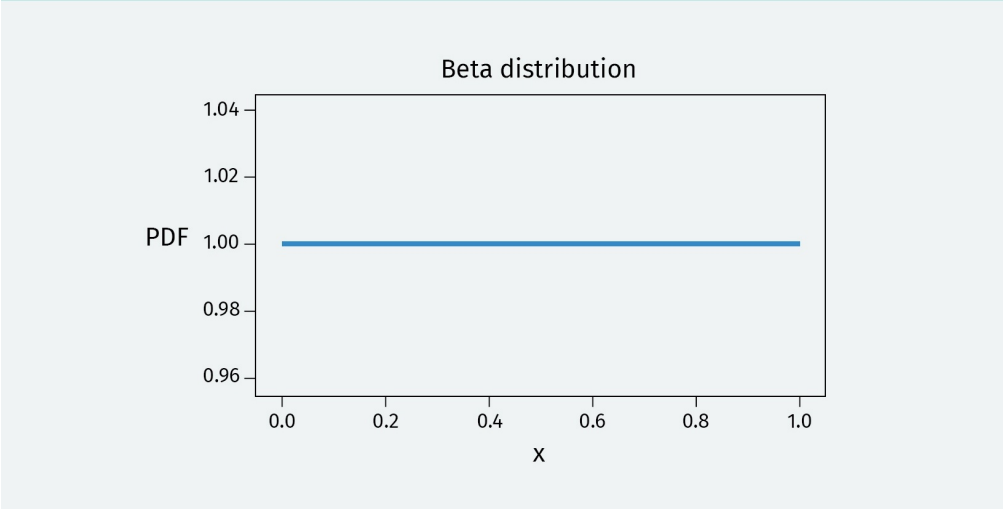
Source: Ulrich Kerzel, 2020.

Since the beta distribution is the conjugate distribution to the Bernoulli and binomial distribution, it can be used to model any prior knowledge for Bernoulli or binomial trials, e.g. in A/B tests.

When used for modelling the prior for Bernoulli or binomial trials, the parameter α represents the number of successes, and β the number of failures. For example, if in $n=10$ trials, the experiment gave 4 successes, $\alpha=4$ and $\beta=6$.

The values $\alpha = \beta = 1$ result in a flat or uniform distribution of the beta distribution:

Figure 34: Beta Distribution Where α = β = 1



Source: Ulrich Kerzel, 2020.

In terms of Bayesian statistics, we can use this to model an uninformative prior, i.e. if we don't know anything about the outcomes of the A/B test, we assume that all outcomes are equally likely. Note that this assumption is not always valid and a flat distribution is not always a good choice for an uninformative prior.

Using the values $\alpha = \beta = 1$ resulting in a flat beta distribution as an uninformative prior, the prior for the Bernoulli trials is then modelled on top of this flat distribution:

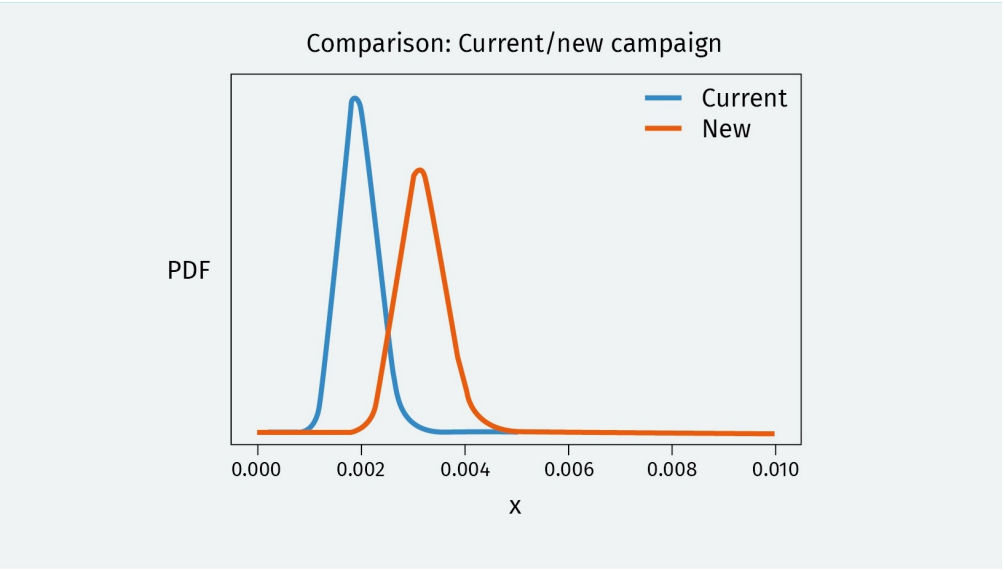
- $\alpha = 1 + \text{number of successes}$
- $\beta = 1 + \text{number of failures} = 1 + \text{number of trials} - \text{number of successes}$

Example:

Continuing with the e-commerce example, where the following result was obtained: The current campaign resulted in 15700 impressions and 30 conversions, whereas the new campaign had 16000 impressions leading to 50 conversions. The parameters of the beta distributions for these cases are given by

For the current campaign: $\alpha_c = 1 + 30; \beta_c = 1 + 15700 - 30$ and for the new campaign: $\alpha_n = 1 + 50; \beta_n = 1 + 16000 - 50$

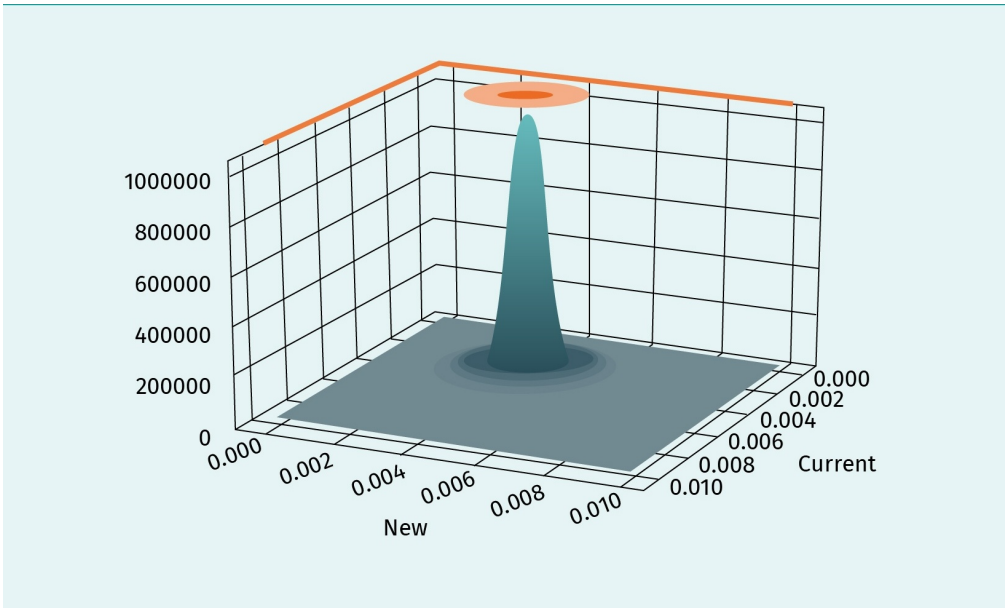
Figure 35: Beta Distributions Representing the Current and New Campaign



Source: Ulrich Kerzel, 2020.

The two experiments (running the current and new campaign as part of the A/B test) are now represented as two probability density distributions, one for each experiment. The joint probability density distribution of the two beta distributions of the new and current approach can be visualized like this:

Figure 36: Joint PDF of the Beta Distributions Representing the Current and New Approach



Source: Ulrich Kerzel, 2020.

The crucial aspect in the evaluation of an A/B test is to determine whether the new campaign is more successful than the old one, i.e. the volume of the joint probability density distribution where the marginal distribution representing the new approach is “bigger” than the one for the current approach. In general, this has to be evaluated numerically. However, for the special case of the beta distribution, an analytic solution exists. Let the function $\mathrm{g(a,b,c,d)}$ represent the probability that a sample taken from a beta function with parameters $\mathrm{(a,b)}$ is larger than an independent sample taken from a beta function with parameters $\mathrm{(c,d)}$, where $\mathrm{a,b,c,d>0}$:

$$\mathrm{g(a,\ b,\ c,\ d)} = \int_0^1 \frac{\mathrm{x^{a-1}(1-x)^{b-1}}}{\mathrm{B(a,\ b)}} \mathrm{I_x(c,\ d)dx}$$

where $\mathrm{I_x(c,\ d)}$ is the incomplete beta function. For the special case $\mathrm{d=1}$, this formula can be evaluated as:

$$\mathrm{g(a,\ b,\ c,1)} = \frac{\Gamma(\mathrm{a+b})\Gamma(\mathrm{a+c})}{\Gamma(\mathrm{a+b+c})\Gamma(\mathrm{a})}$$

and it can be shown that all cases where d is a positive integer can be reduced to this case (Cook, 2005). Hence, for the special case of A/B tests where all numbers are positive integers, no numerical integration is needed and the above formulae can be used.

Example (continued):

Continuing with the above example, we can now evaluate the probability that the first sample (the new approach) is larger than the second: $\mathrm{g(a_n,\ b_n,\ a_c,\ b_c)} = 0.98$.

Hence, the uplift from the original sample was 63 percent and the new approach has a probability of 98 percent that it will be better than the current campaign.

SUMMARY

Defining success is one of the critical aspects in establishing a use case for a new data-driven project. The predictive model itself is typically evaluated in terms of previously discussed metrics. However, the predictions are typically far removed from the actual business operations as the predictions first need to be transformed into optimized business decisions. These then have a direct impact on the success of the project.

The best way to evaluate the performance in a business context is to define a suitable cost function which contains and quantifies all relevant aspects related to the project. The optimal point estimator of the predicted probability density distribution (in case of regression) can then be identified by finding the minimum of the cost function. The optimal decisions can be based on this estimator.

In many cases, however, obtaining the full cost function is difficult or prohibitively expensive. In these cases, a set of suitable key performance indicators (KPIs) that measure the impact and benefit of the project should be identified. The optimal decisions derived from the predictions are then evaluated in terms of these KPIs.

Often, a new approach needs to be verified against an existing method. In these cases, an A/B test is performed where the new and the current approach are executed side-by-side to measure the uplift of the new method compared to the existing approach.

UNIT 4

MONITORING

STUDY GOALS

On completion of this unit, you will have learned …

- why monitoring is important to the success of data-driven projects.
- how dashboards can be used for monitoring.
- how to design good dashboards for different target audiences.
- how to use automated reporting and alerts as systematic checks of the project performance.

4. MONITORING

Introduction

Defining the success of a project, either from a model or from a business centric approach, is only the first step. Once relevant metrics are defined, they need to be measured frequently to determine whether or not they are in the allowed range. Action must be taken once the measures are outside the range which has been deemed acceptable for a given project.

Two complementary approaches are discussed in this unit:

Monitoring
Continuous assessment of one or more quantities.

- Visual **monitoring** using dashboards: Dashboards are a form of visual communication. They take the raw numbers from performance metrics, KPIs, etc. and present them in so that it is immediately obvious whether or not the values are in the allowed range, which aspects are considered "good", and where action is required.
- Automated monitoring and reporting: In many cases, specific actions are required when certain conditions are met. Automating the measurement of relevant metrics, as well as the required actions, frees more time to think about appropriate responses for cases in which pre-defined actions are not possible, and more deliberation and discussion are required. Furthermore, many data are acquired by a wide range of automated systems and sensors. Automating checks and alerts can help to improve the quality of the recorded data by ensuring that each system is operating within its nominal parameters and alerting the appropriate operator if a condition is met.

4.1 Visual Monitoring Using Dashboards

Dashboard
A collection of graphical representations of critical measurements, designed to give a quick overview of the overall situation.

Dashboards are a powerful means of visual communication which can be used to convey critical information about the current status quo and whether action needs to be taken. Dashboards aggregate and visualize information that has been stored and processed in the IT infrastructure of an organization, i.e. dashboards are an interface for information that is already present elsewhere. This implies that all relevant information is already present, and part of designing a dashboard is to make sure that all required information is recorded and can be accessed from the IT infrastructure.

Dashboards are used to convey relevant information to a wide range of audiences for different purposes. Below are a few typical examples.

C-Level and Senior Management

A dashboard designed for the senior management team should convey high-level information on one single screen. It should summarize the current status quo of the whole organization and indicate the status of the KPIs. Since screen size is limited, designing a single dashboard is extremely challenging as it is very tempting to put lots of details on the dash-

board. However, this quickly leads to visual clutter and unreadable charts which means that the benefit of the dashboard is lost. When designing the dashboard, it is important to keep in mind which operational data the senior management team actually needs on a day-to-day basis, and to remember that KPIs are tied to specific teams or projects. In case questions arise, these teams can be held accountable for their respective KPIs and should be able to quickly answer any question. In addition, it is helpful to design more detailed dashboards for each member of the senior management team which focus on their respective area of responsibility. The chief financial officer (CFO) will need a more detailed overview of the financial situation, the chief sales or revenue officer (CRO) will be more interested in the current status of the sales pipeline, number and **total contract value** (TCV), and the most promising current opportunities. The chief operating officer (COO) or chief technology officer (CTO) will typically require more internal information about the current status of projects, infrastructure, and resources. When designing the dashboard, it should be kept in mind that most of the information on the dashboard is confidential. While the dashboards should be easily accessible for the intended audience, external visitors and most internal employees are not supposed to see these details.

Total Contract Value
A business metric which establishes how much money a deal brings to a company, including one-time and recurring fees.

Department summary

Similar to dashboards for the heads of particular departments or business units, these are intended for the management team of the respective unit. They are typically more detailed than the high-level summary for the senior and executive management, as they are intended to give a quick overview of the current status quo and serve as a basis for further discussion about upcoming issues.

Project or team summary

Once projects have left the main development phase and become operational, they have a number of critical factors which need to be monitored continuously. For example, operators might need to monitor the performance of the casting process in a steel mill, the throughput in an assembly line, etc. A dedicated dashboard for the project or team manager provides a detailed overview of the current situation and can also be integrated into the daily meeting of the team to discuss the current performance, any issues which have arisen in the last day, and which actions may be necessary for the following day.

Dashboards should be designed for their intended target audience. Each audience is unique and will need access to certain information at a specific level of granularity. When composing the dashboard, the relevant information should be clear, and sufficient context should be provided so that it is easy to understand what is shown on the dashboard without having to refer to a manual. Visualizations should be clear of “graphical clutter,” i.e. elements which are only added to decorate the chart. KPIs and other metrics should be visualized using appropriate graphs with visual cues that indicate whether the current value is acceptable or if action needs to be taken. Adding tables or raw numbers should be avoided as they are harder to read and their interpretation requires more time than carefully chosen charts.

Example dashboard

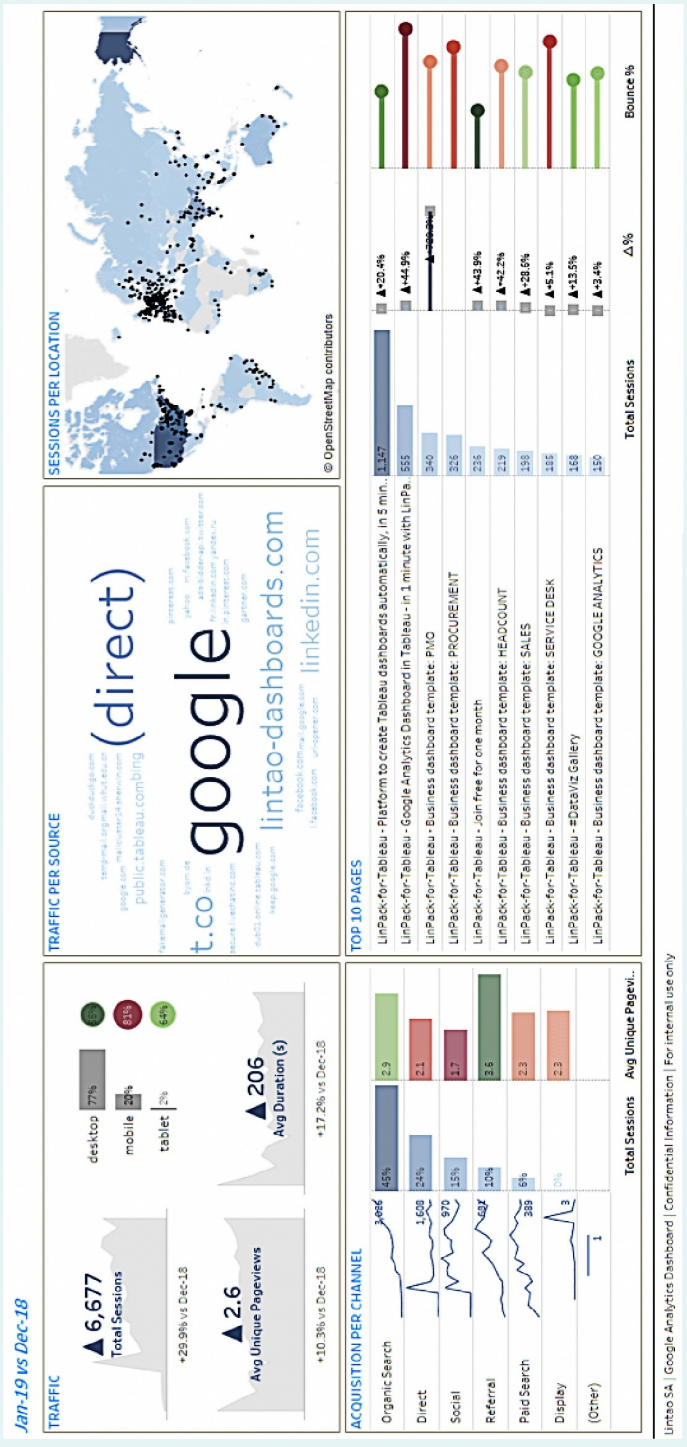
The following example shows a dashboard focussing on high level aspects which are relevant to managers at a food retailer. The dashboard shows a few relevant KPIs as well as some graphs. However, graphical clutter is avoided as the graphs contain no values on the x- or y-axis. If further details are required, these should be shown in a separate dashboard or a detailed report.

Figure 37: Management Dashboard for a Food Retailer

Source: Linpack, 2019a.

The next example focuses on the marketing effort of the same food retailer, specifically, a visualization of Google Analytics which analyzes the online marketing efforts. This dashboard is designed specifically for marketing managers of a company and includes relevant technical terms which are too specific for a general management discussion.

Figure 38: Marketing Dashboard for a Food Retailer



Source: Linpack, 2019b.

Design Elements

Visual representations of data such as graphs, dashboards, or reports use visual design elements to convey the data. These visual elements can be used to represent numbers, highlight different aspects, guide the eye of the viewer to relevant points, or lead them to conclusions. Creating graphs and visual representations has been discussed for many years; one of the earliest books on the topic is the seminal work by A. C. Haskell in 1919 which came just after the work by the Joint Committee on Standards for Graphic Presentation in 1915.

Line style:

In a graph, lines can be used to connect points and guide the eye. The importance of a connection can be emphasized using the thickness of a line:

Figure 39: Line Thickness

Source: Ulrich Kerzel, 2020.

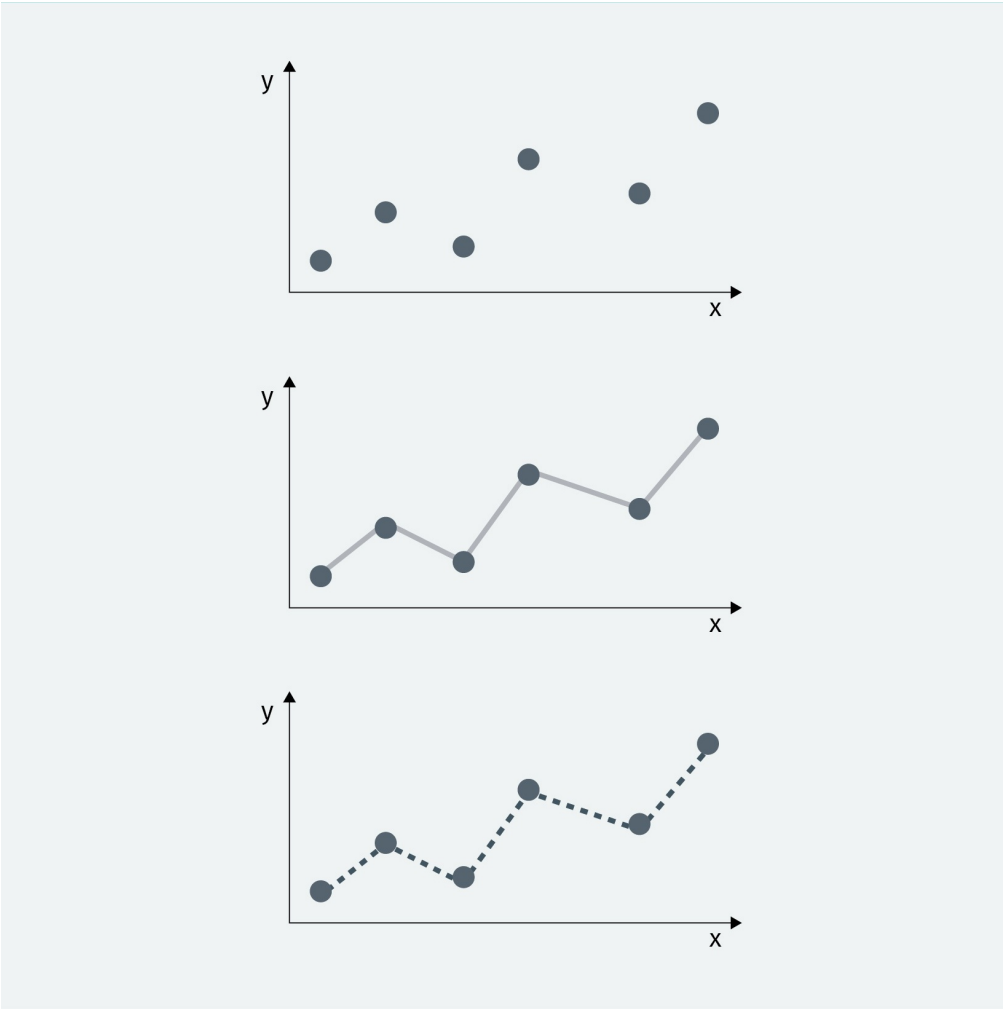
Or its drawing style:

Figure 40: Line Drawing Style

Source: Ulrich Kerzel, 2020.

The different styles can be used to represent different connections. A solid line will generally indicate a stronger connection than a faint dotted line.

Figure 41: Faint Dotted Line vs. Strong Solid Line

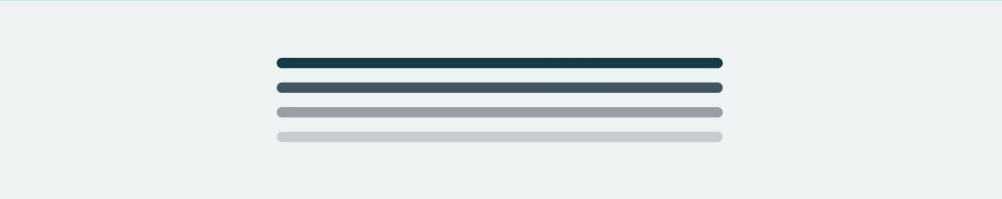


Source: Ulrich Kerzel, 2020.

Note that in the case of the strong solid line, the viewer will jump to the conclusion that there is a linear relationship between the data points and the values will simply rise or fall from one point to the next whereas, in reality, each point represents a single measurement and nothing is known in between. Although a simple line might be expected in many cases, the behavior might be more complex in reality.

Different shades can also highlight individual lines.

Figure 42: Different Shades

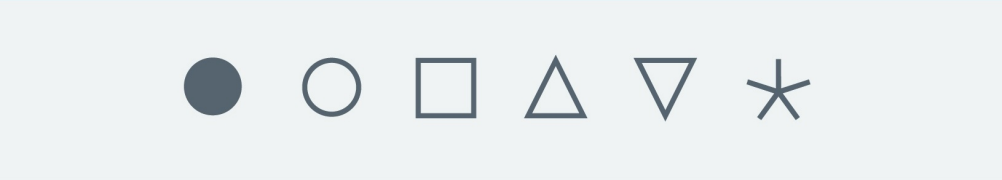


Source: Ulrich Kerzel, 2020.

Symbols:

A variety of symbols can be used to differentiate between data-points.

Figure 43: Different Symbols

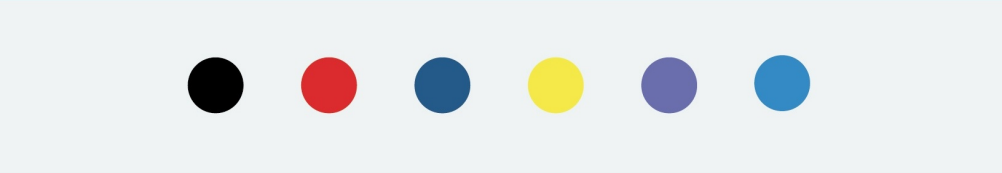


Source: Ulrich Kerzel, 2020.

Color:

Colors are a further means to differentiate data points from one another.

Figure 44: Different Colors



Source: Ulrich Kerzel, 2020.

Colors can also be used to highlight special values. For example, try to find every “3” in the following random number:

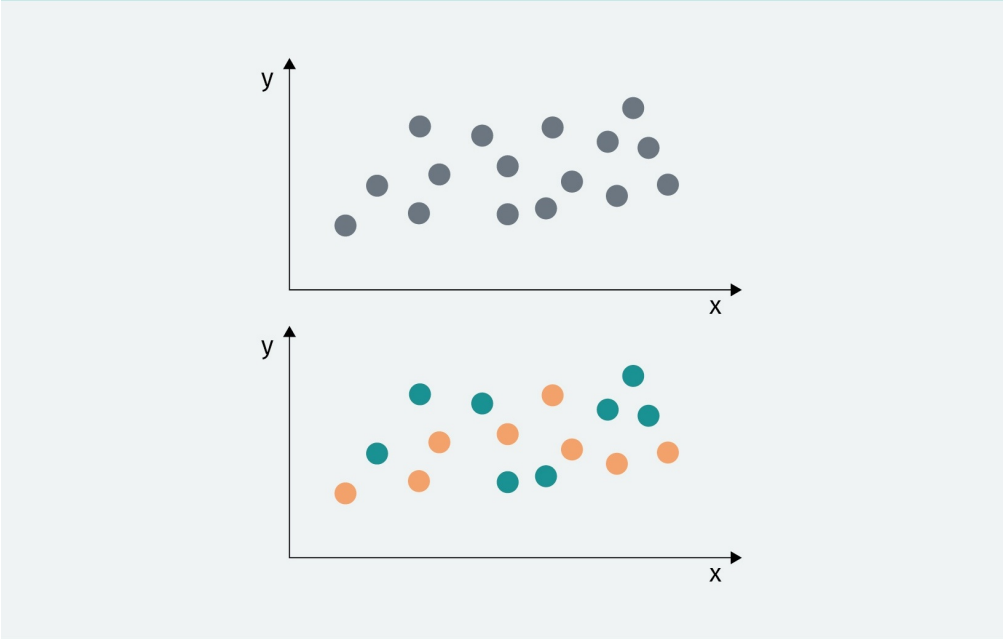
19012966820397955975113129729782206165717225863361773210616561995815285719029278497764
33878609426925052135886989460175050482412461629034083412784181855331872370495749535491183212

This is much easier when the 3s are bolded.

190129668203979559751131297297822061657172258633617732106165619958152857190292784977643387860942692505213588698946017505048241246162903408341278418185533318723704957495354911832 12

Colors can also be used to group different data points and emphasise their relationship with each other.

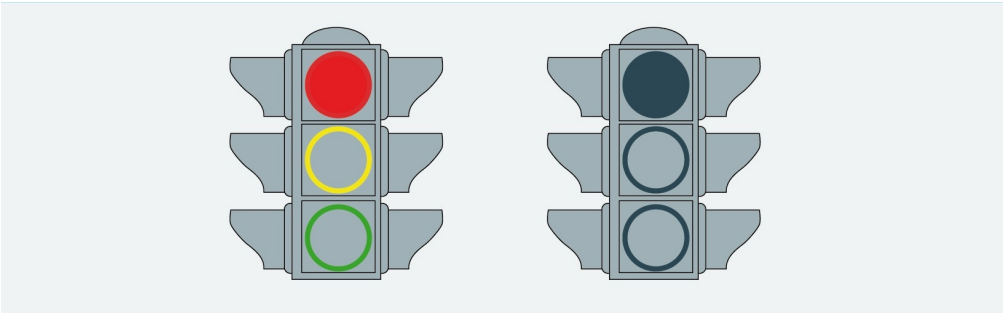
Figure 45: Use of Colors to Group Data Points



Source: Ulrich Kerzel, 2020.

Many colors have a special meaning to us as we associate them with an everyday experience. For example the following are much more prominent with the use of color:

Figure 46: Colors and Their Meanings in Everyday Life

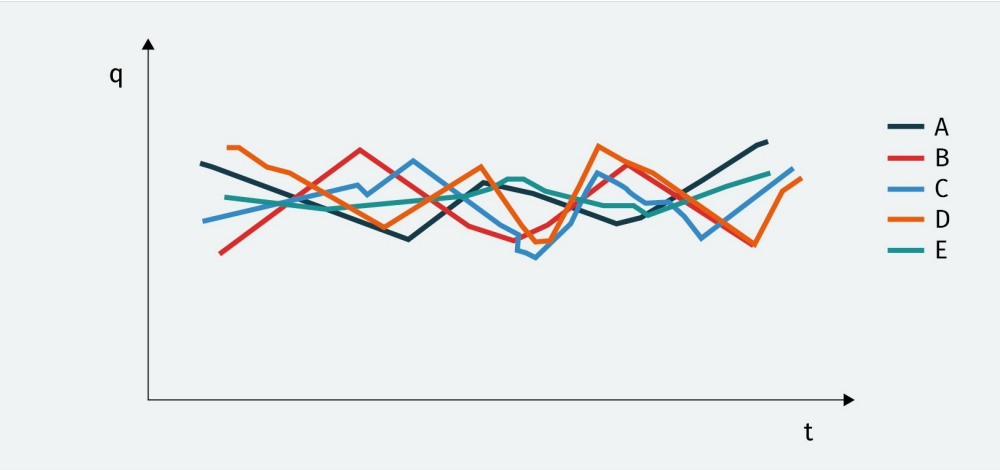


Source: Ulrich Kerzel, 2020.

Building a traffic light without color does not convey the same meaning and sense of urgency as the one with color.

However, using color to squeeze too much information into a graph quickly results in information overload and it becomes impossible to follow what is going on:

Figure 47: Color Overload



Source: Ulrich Kerzel, 2020.

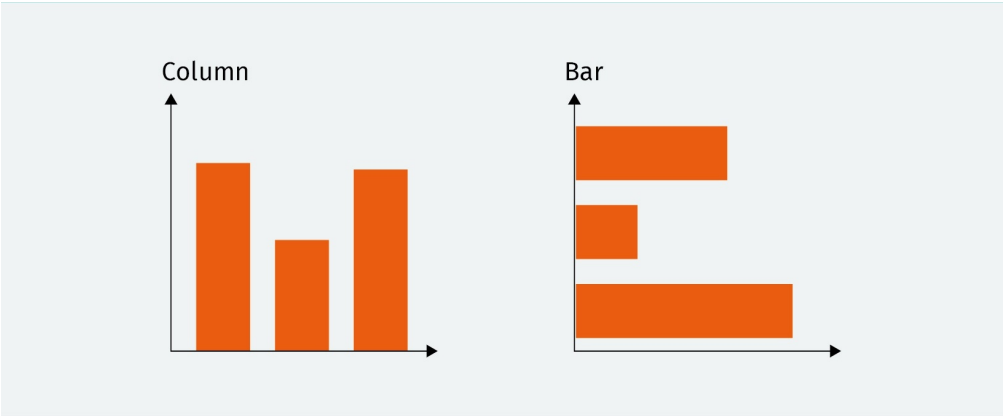
Chart types

A wide range of chart types are available to illustrate information graphically. Different charts have different strengths and weaknesses, so the appropriate method should be chosen for each visualization. As with any other form of communication, considering the needs of the intended target audience is critical when designing a graphical representation. A presentation to the CEO or a member of the senior management team will need a different approach when summarizing the current status quo than a presentation to the project manager.

Column/Bar Graphs:

The column or bar graph introduced by Playfair in 1801 is a simple visualization to illustrate different values of a given variable or category:

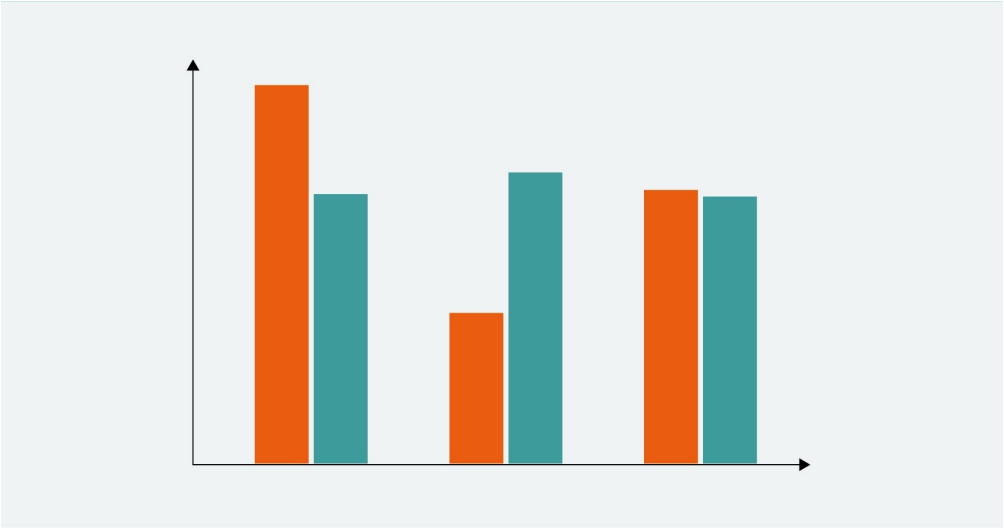
Figure 48: Column and Bar Graphs



Source: Ulrich Kerzel, 2020.

Both types of graph can be used to effectively compare the differences between several categories as shown below.

Figure 49: Column Graph Used to Compare Categories

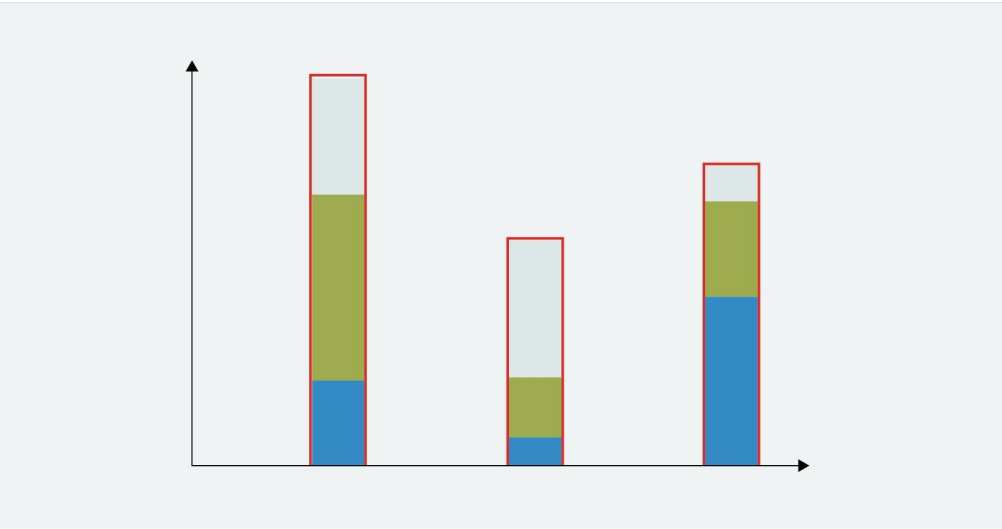


Source: Ulrich Kerzel, 2020.

However, as more categories are added, this graph type quickly becomes overloaded and confusing, and adding appropriate labels becomes difficult.

Stacked graphs allow us to judge the behavior of categories compared to the whole range, thus allowing an aggregated and more detailed view at the same time. However, this type should be limited to three categories to avoid making the graph unreadable.

Figure 50: Stacked Graph



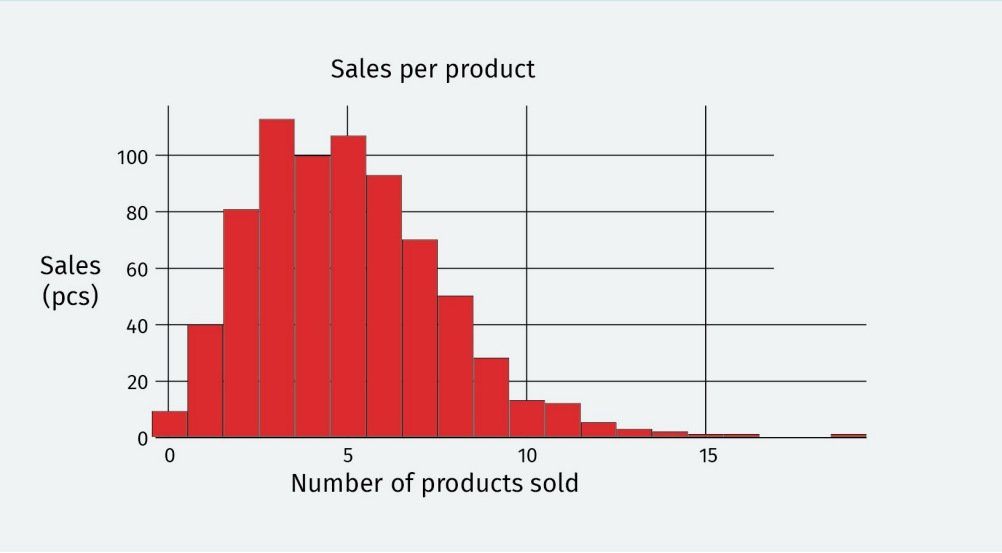
Source: Ulrich Kerzel, 2020.

Histogram:

(Histogram) Bins
Discrete interval into which a continuous variable is split.

The histogram (Pearson, 1895) is related to the bar chart, however, a continuous variable is visualized in **bins**. This allows us to visualize large amounts of data effectively. The number and width of each interval should be chosen carefully. Which choice is more appropriate depends on the context of the variable shown.

Figure 51: Histogram

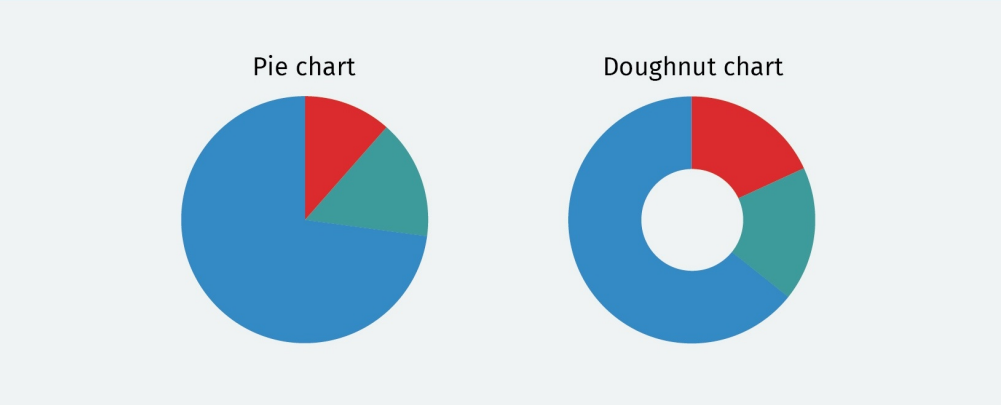


Source: Ulrich Kerzel, 2020.

Circular Charts:

Circular charts compare fractions to a whole. They work best when considering only a few fractions as these chart types can become confusing if too many fractions are represented. Furthermore, it should be kept in mind that humans are not very good at judging angles, hence these visualizations are best suited to give a general overview and should be avoided if a high level of detail is required. The pie chart (Playfair, 1801) can be described as a filled circle where sections are used to illustrate the respective fractions. The doughnut chart (Harris, 1996) is a variation where the middle of the chart is left blank.

Figure 52: Pie Chart and Doughnut Chart

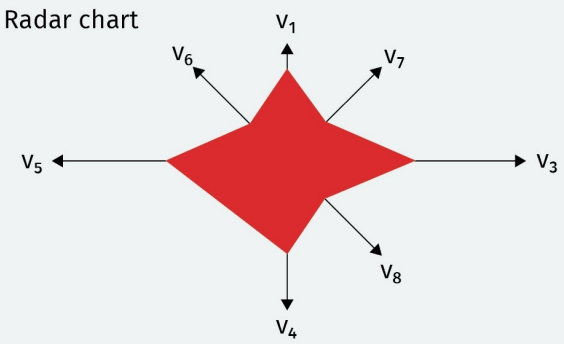


Source: Ulrich Kerzel, 2020.

Radar Charts:

The radar chart (Porter & Niksiar, 2018) allows us to visualize a set of variables in a 2-dimensional representation. The variables are arranged on a circle with a common origin in the middle of the circle. Typically, the variables are normalized relative to each other to have a common scale, e.g. 0 percent in the middle and 100 percent at the furthest point. The relative position of the variables around the outer circle is usually not used to convey any particular information.

Figure 53: Radar Chart

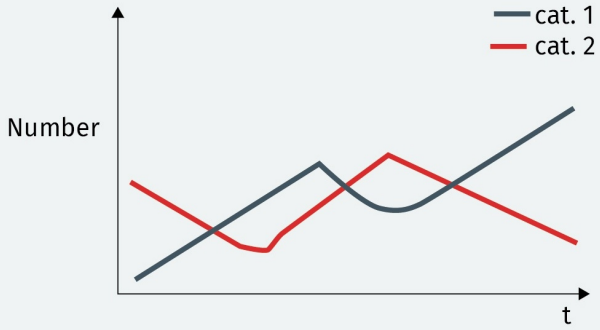


Source: Ulrich Kerzel, 2020.

Line Charts:

Line charts (Spear, 1952) are typically used to illustrate time-dependent data, for example, a time-series such as the sales of a product over time. The line chart can be used to visualize different categories at the same time to enable judgement concerning the way that these categories evolve and how they relate to each other. However, the graph can become too crowded and confusing if too many categories are used. Using lighter and more intense variants of the respective colors can help to highlight individual categories.

Figure 54: Line Chart

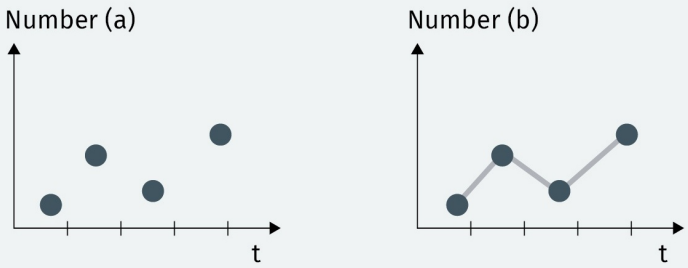


Source: Ulrich Kerzel, 2020.

It should also be noted that a solid line is very suggestive to the human observer. For example, the following graphs contain the same information: At certain points in time, a number is recorded and added to the line graph. In the left graph a), the four observed

data points are added but no further indication is given with regard to how they are connected to each other. In the right graph, each point is connected, indicating a linear relationship between them:

Figure 55: Connection of Data Points

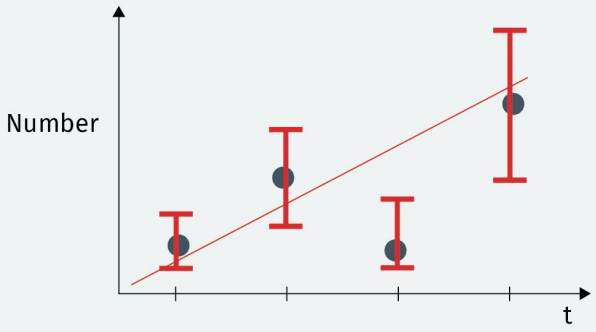


Source: Ulrich Kerzel, 2020.

Unless such a relationship is known, this can mislead the observer of the chart.

For example, measured data points are usually associated with uncertainty, e.g. due to a measurement error or the intrinsic resolution of a sensor. Taking further information into account, the connection between the points could be very different, for example:

Figure 56: Different Connection of Data Points



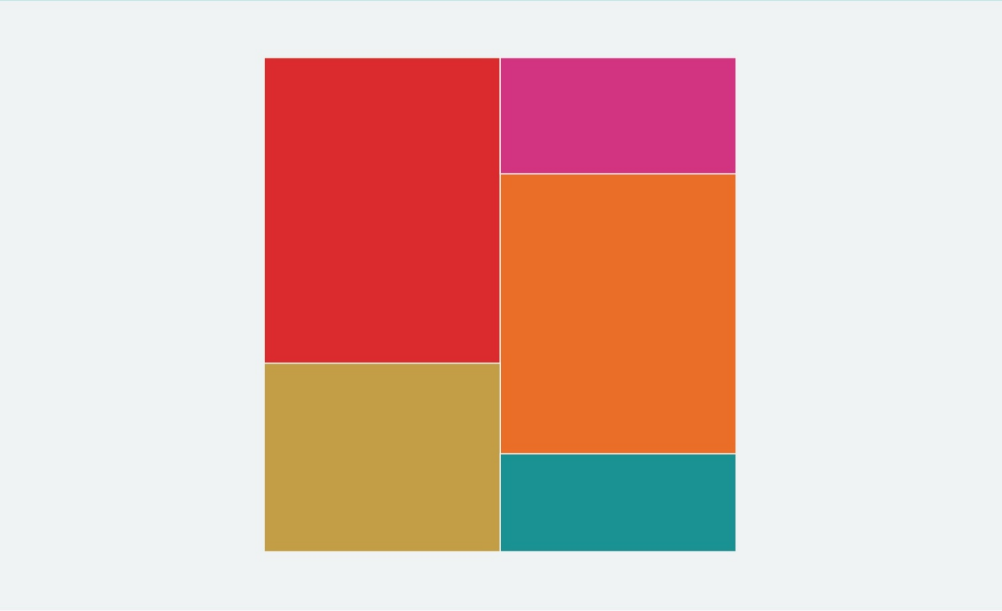
Source: Ulrich Kerzel, 2020.

In this example, each point is now associated with an asymmetric uncertainty and the resulting trend has been added.

Tree Maps:

Tree maps (Shneiderman, 1992) allow us to visualize proportional values and hierarchical relationships at the same time, and are ideally suited to visualize many hierarchical values. Tree maps can be used to give a general overview and some level of detail at the same time. However, labelling becomes challenging as the size of the boxes gets smaller.

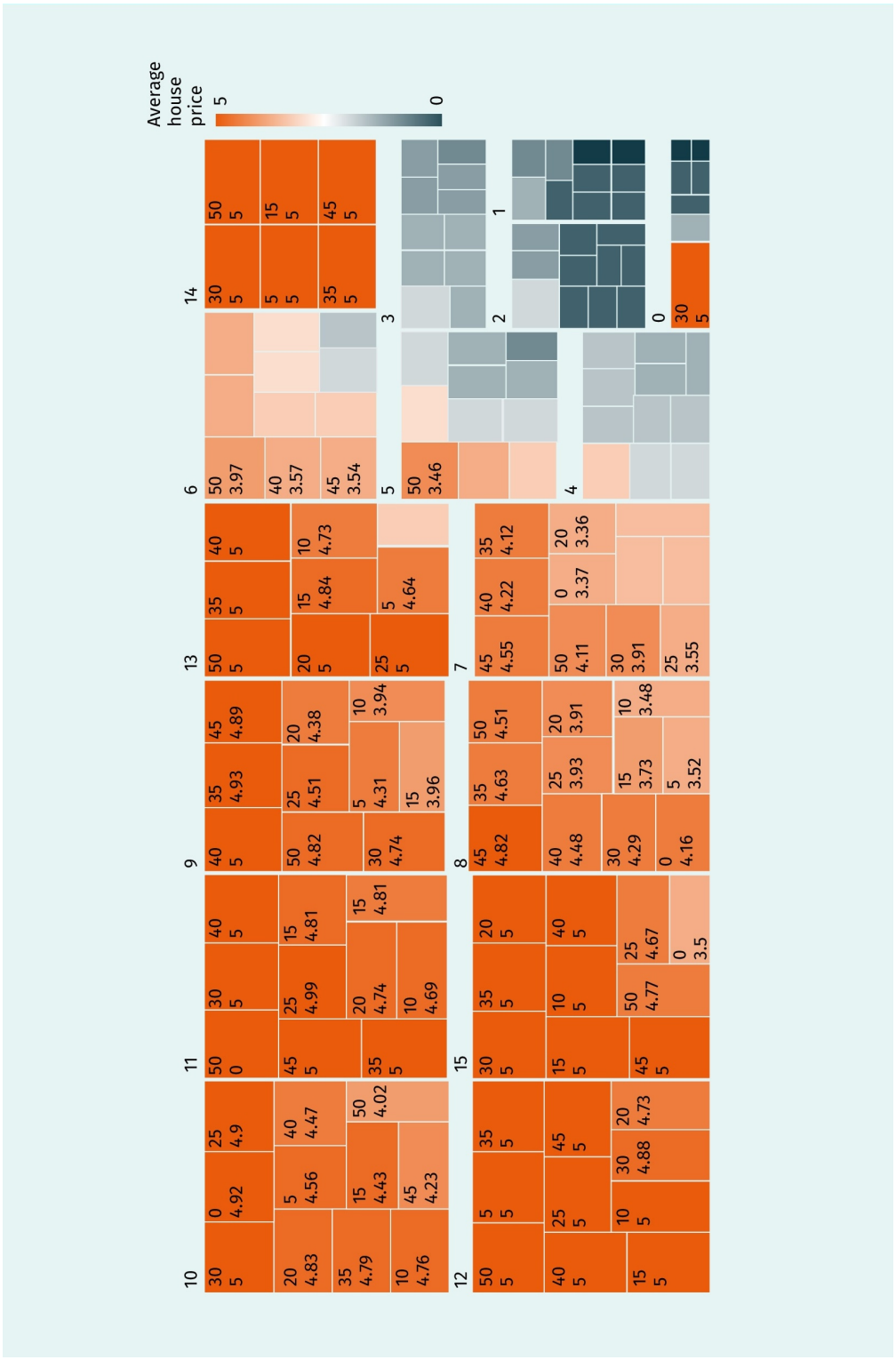
Figure 57: Tree Map



Source: Ulrich Kerzel, 2020.

The following example shows a tree map of the average house price (on a logarithmic scale) in California depending on the median income in the neighborhood and the age of the house. The color is used to visualize the price of the house (on a logarithmic scale between 0 and 5), and the nested boxes represent median income and age:

Figure 58: Tree Map of the Average House Price in California

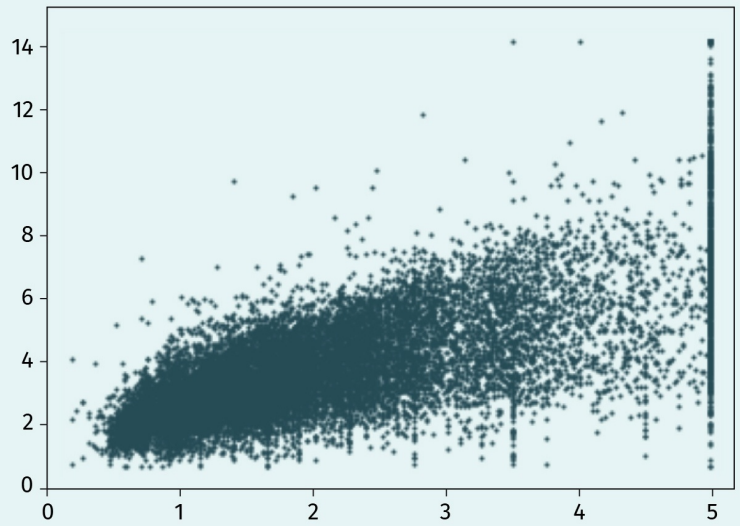


Source: Ulrich Kerzel (2020), based on Pace & Barry, 1999.

Scatter Plot and Scatter Plot Matrix:

The scatter plot is used to visualize the behavior of two variables: The x-axis is used for one variable, the y-axis for the other. A historical account leading to the invention of this type of plot is given by Friendly & Denis (2005). Each measured data-point is added to this two-dimensional chart as shown in the example below:

Figure 59: Scatter Plot

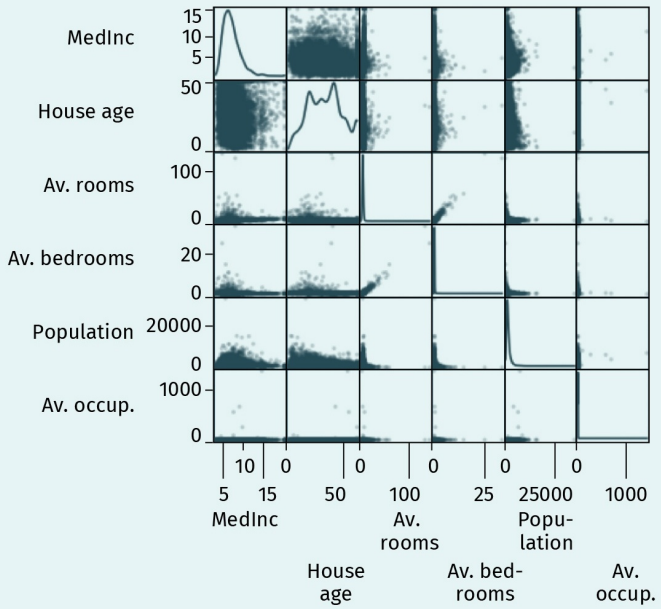


Source: Ulrich Kerzel, 2020.

However, as the number of data points increases or if variables with only a few categories are used, the scatter plot becomes very difficult to interpret as the individual data points are drawn too close to each other to convey much information.

The scatter plot matrix is made of scatter plots for each combination of two variables. The diagonal is often used for a histogram of the respective variable. Again, this visualization type becomes difficult to interpret for large datasets or variables.

Figure 60: Scatter Plot Matrix



Source: Ulrich Kerzel, 2020.

Bubble Plot:

The bubble plot is a variant of the scatter plot where each data point is replaced by a bubble of variable size. This additional parameter can be used to add information from a third variable. However, this chart type is only useful for a few data points.

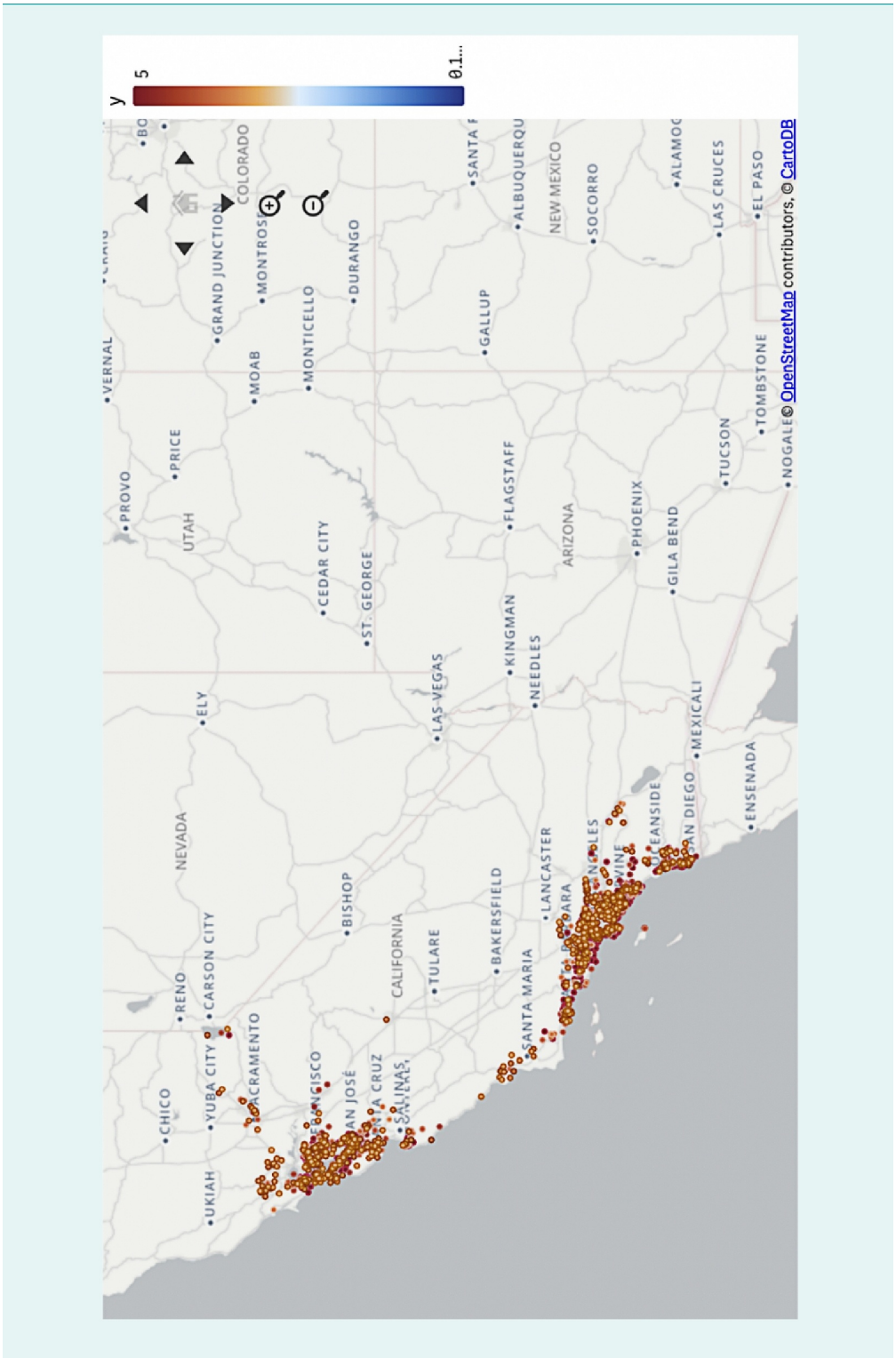
Figure 61: Bubble Plot

Source: Ulrich Kerzel, 2020.

Maps:

Maps are ideal to visualize geo-spatial information. Data points for each location can be highlighted using size, color, and shape to convey further information as shown in the example below, highlighting house prices in California on a logarithmic scale.

Figure 62: Map Highlighting House Prices in California



Source: Ulrich Kerzel (2020), based on Pace & Barry, 1999.

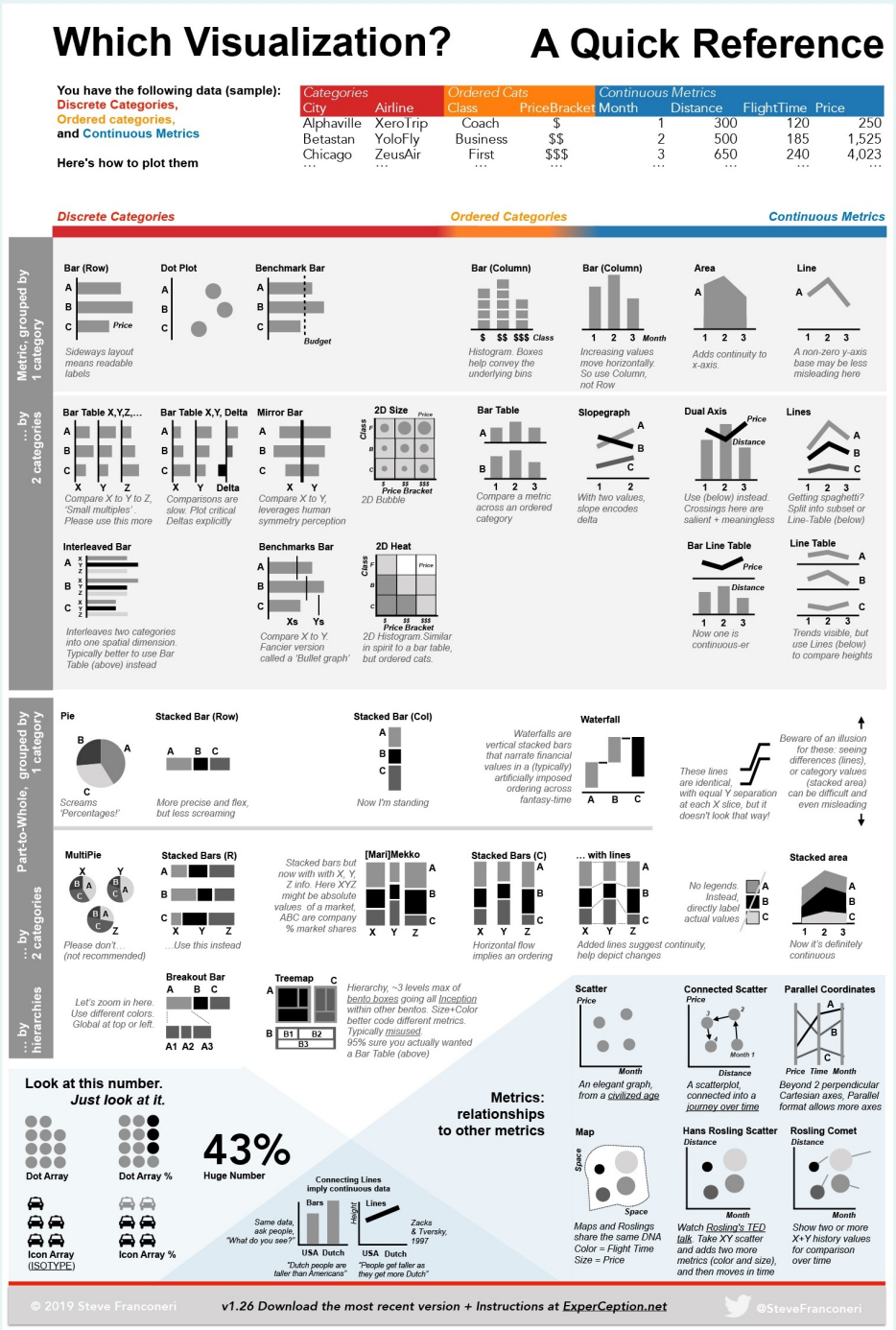
Best Practices

Parmenter (2015, p. 222) summarizes a few key concepts developed by St. Few to bear in mind when designing visualizations:

- Keep within the boundaries of a single screen: Instead of providing many options of unfolding detailed charts and data representations, think carefully about how the information should be presented to the intended audience.
- Provide sufficient context: Indicate whether the numbers are within a "good" or "bad" range. This range has to be decided beforehand by taking in to account the advice of experts.
- Provide adequate level of detail or precision: The level of detail in charts or the precision of the numbers shown should reflect the overall message of the visualization. For example, $25 million is much easier to understand than $25,000,000.
- Start scales at zero: It is often tempting to start the scale of graphs at some other point than zero. However, this introduces a cognitive bias and distorts the magnitude of differences.
- Keep a consistent color scheme: All visualizations should have the same color scheme, for example, low numbers indicated by blue, high numbers by red. In addition, the color scheme should use as few colors as possible but as many as necessary to avoid clutter and keep the visualizations simple.
- Avoid decorations: Additional decorations and graphical elements without functionality can clutter the visualization and should be avoided unless they provide additional context.

The following "cheat sheet" is also helpful when deciding on which visualization to use:

Figure 63: Visualization Quick Reference



Franconeri, 2019 (used with permission)

Source: Franconeri, 2019.

4.2 Automated Reporting and Alerting

Dashboards and visual representations of KPIs are ideally suited to convey relevant information and to indicate whether key aspects and numbers are within nominal parameters. They provide high level information at a glance and indicate whether action needs to be taken.

However, a company, a single project, or a system typically requires much more detailed data and result indicators than are used for a dashboard. Firstly, all relevant detailed data need to be stored and processed in appropriate IT systems to be able to construct the aggregated measures which are displayed in the dashboard. Although this sounds obvious, this step alone often poses significant challenges in practice as IT systems have grown over the decades and relevant information is scattered across multiple, often incompatible, systems. Automated checks to determine whether the data in the system are sensible are the only way to monitor and improve data quality as the overall volume, as well as the rate at which new data are added, are far too high to be able to monitor them manually.

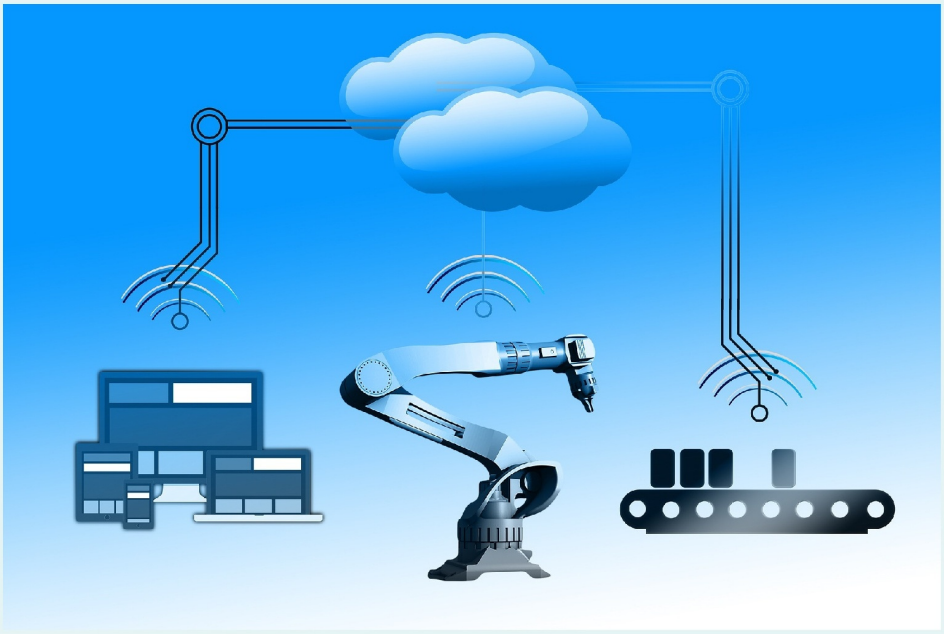
Similarly, the overall status or performance of most systems, projects, or divisions are monitored via a wide range of metrics, results, or performance indicators. Many of these will be too detailed to be used in higher level monitoring or visualizations, however, they provide valuable insights into the individual performance of a particular aspect of a project or system. Including automated alerts also helps to draw attention to any particular aspect which may be hard to find if only aggregated views, such as high-level dashboards, are used. As there are typically too many metrics to be monitored manually on a regular basis, a valid range of acceptable values should be defined for each one so that automated checks can report whenever these boundaries are violated. If possible, specific actions can be linked to the alerts beforehand, allowing for an immediate response. For example, if a specific part of a machine is sending irregular data, this might indicate that the part is broken and the maintenance crew can be alerted automatically so that they can inspect the machine.

In addition, any sudden change should be reported, even if the change occurs within the allowed range, as this may indicate a change in operating conditions or external influences.

A good example is low-level data such as individual sensor readings of a machine or production pipeline. Each sensor has a valid range of operating conditions and output values. Both aspects of the operational conditions of each individual sensor have to be monitored. Can the sensor operate normally and within design specifications? Is the output of the sensor sensible and within the expected range? The first question addresses the issue of whether the sensor can, in principle, work as expected. For example, a temperature sensor may only be sensitive between 0 and 500 degrees Celsius, but would provide nonsensical values if placed in an environment colder than 0°C. The second question aims to monitor whether the sensor itself is working normally or if it is broken. Smart sensors may include many self-check options and can report any failure themselves, however, simple sensors often do not provide such information and have to be monitored externally. Any deviation from expected values needs to be recorded and an automated alert highlighting a potential issue must be sent to the operator. As a typical production plant or processing

pipeline is fitted with tens to hundreds of thousands of sensors, automating data quality measures and alerts are the only way to ensure high-quality data and identify broken sensors.

Figure 64: Connected Factories



Source: Daily News Hungary, 2019.

SUMMARY

Continuous monitoring is a key aspect of running a data-driven project. Visual aides such as dashboards are ideal when conveying relevant information at a glance. These dashboards should be tailored to different audiences such as senior management, project management, or operational staff.

When designing dashboards, focus on what is important to the target audience and take care to avoid overloading the dashboard with too many details.

Automated reports and alerts allow for the continuous monitoring of a project's performance and minimizes the requirement for regular manual evaluation.

UNIT 5

AVOIDING COMMON FALLACIES

STUDY GOALS

On completion of this unit, you will have learned …

- what cognitive biases are.
- the most important cognitive biases and how to avoid them.
- how statistical effects can influence the evaluation of a project.
- the key aspects of change management transforming companies into data-driven enterprises.

5. AVOIDING COMMON FALLACIES

Introduction

Using data to make decisions is very hard — despite best efforts and advances in technology, many decisions are still made based on gut feeling or the experience of individual experts instead of data. This is unavoidable for urgent decisions that must be made when insufficient data are available. An example are strategic decisions such as whether to open a new factory or expand into a new market. In these situations, and in similar cases, one can perform market research and evaluate different options, but unless a company opens a new factory every few days, there will not be sufficient data available to predict whether not not it will be a success. Instead, the minimal data available can be used to develop an informed opinion to assist in the evaluation of different options. However, in many other cases, making data-driven decisions will be difficult but might be possible. This unit focuses on some of the most common fallacies and pit-falls. **Cognitive biases** and behavioral aspects make data-driven decision making difficult, as the way we make decisions is wired into our brains due to millennia of evolution. This is beneficial if we need to survive in nature, but impedes our ability to make deliberate decisions. Furthermore, we don't have an intuitive understanding of statistical effects. Our everyday world is mostly deterministic, and dealing with stochastic effects requires a lot of experience. Finally, changing an organization to enable data-driven decisions in the first place is a challenge on its own. Many organizations have strict hierarchies which often rely on managers "doing the right thing," therefore preventing the use of the best decision possible given the current data. This makes them vulnerable to cognitive or statistical biases.

Cognitive Bias
The way in which our behavior and decisions are influenced, often subconsciously.

5.1 Cognitive Biases

Humans have evolved over millennia and, for almost all of this time, our existence has mainly focused on survival, agriculture, hunting, or warfare. Although science and philosophy are thousands of years old, mass education is a relatively new concept. The computerization of our everyday life started less than 100 years ago and, within a few decades, data and data-driven approaches have evolved so much that they are prominent in all aspects of our daily lives. This implies that our brain structures and thought processes will not have adapted in such a short amount of time, and are therefore still wired for comparatively simple lives. Although this has guided our survival throughout history, it makes it difficult for humans to deal with data-driven decisions.

Kahneman distinguishes between two systems which are generally used in human decision making, these are System 1 and System 2 (Kahneman, 2012, p. 20):

- "System 1 operates automatically and quickly, with little or no effort and no sense of voluntary control."

- “System 2 allocates attention to the effortful mental activities that demand it, including complex computations. The operations of System 2 are often associated with the subjective experience of agency, choice and concentration.”

When we think about the decisions that we make, we mainly think of System 2. However, most of our daily actions are governed by System 1 which continually assesses our environment and establishes a sense of normality. From an evolutionary perspective, System 1 decides whether there are any threats and answers the following key questions: “Are things going well? Should my attention be redirected? Is more effort needed for this task?” (Kahneman, 2012, p. 59).

Millennia ago, this was required to spot any dangers, for example, when roaming the savannah, whereas in our modern life, we encounter this in less dangerous situations, such as when driving a car etc. An experienced driver can typically drive longer distances and familiar routes without needing to put in much effort. Quite often, we don’t even remember driving most of the way as System 1 handles many of the actions required to operate the car automatically so a journey only demands our attention if an unusual situation requires a decision from System 2. The main characteristic of System 2 is that it handles all tasks which require our full attention and careful decision making. Examples include making complex calculations, filling out a tax declaration, parking a car in a narrow spot, focusing attention on the voice of a particular person in a crowd, etc. (Kahneman, 2012, p. 22).

Figure 65: Danger in the Savannah

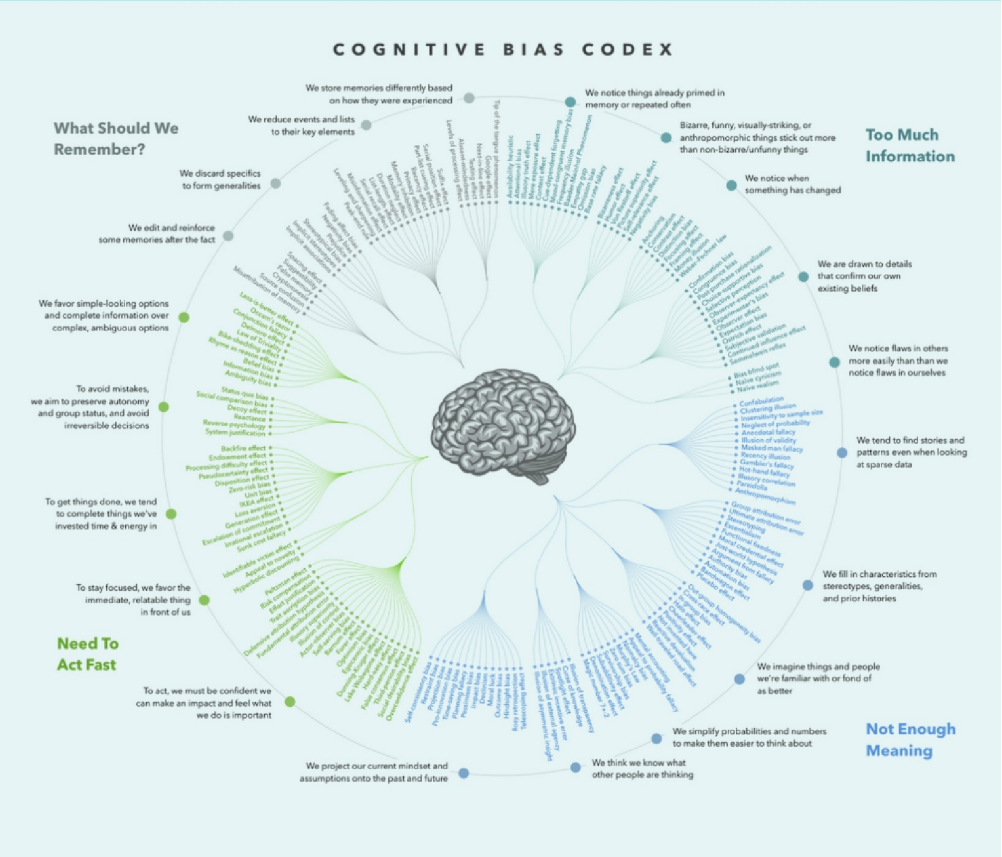
Source: Lindsay, 2016.

One of the most important aspects of System 1 regarding decision making, analyzing data, or evaluating a machine learning model is that System 1 is always trying to find causal connections and construct a plausible story (Kahneman, 2012, p. 75), even if there is no causal relationship in the data. In fact, one of the measures of success for System 1 is how coherent the newly created story is (Kahneman, 2012, p. 85). While creating a causal relationship or story, System 1 is largely insensitive to the amount and quality of the data that it uses to construct this (Kahneman, 2012, p. 86), meaning that the fewer data points are available, the more plausible a story appears as fewer pieces need to be fitted in. Crucially, System 1 does not keep track of alternative causal relationships or stories it constructs — or even the fact that there were alternatives (Kahneman, 2012, p. 80). The creation of a causal relationship is an automatic function of System 1 and we cannot undo it. System 2 is capable of overruling this, but in most cases, it readily accepts the story or connection invented by System 1 (Kahneman, 2012, p. 76).

Many of the errors of judgement we encounter are due to the interaction between System 1 and 2, in particular the automatic function of System 1 that constructs a causal story which is readily accepted by System 2.

The human mind and, in particular, the interplay of System 1 and System 2 gives rise to a large number of cognitive biases. There are currently more than 180 different biases known (Jm3, 2016), some of the most important ones are discussed below in more detail.

Figure 66: Cognitive Bias Codex



Source: Jm3, 2016

Pattern Spotting

Because System 1 automatically tries to construct a relationship or story from the available data, we tend to look for patterns even when we know that none exist. We tend to associate clusters of data points or similar occurrences with an underlying pattern even if they are only due to random fluctuations. For example, when rolling a single die each number has the same probability to occur, provided the dice is manufactured evenly and has no deliberate or accidental imperfections. As each throw of the die is independent of the next, we expect that, with enough attempts, each number will occur with approximately the same frequency, $1/6^{th}$ of the total number of tries. If one observes the sequence of numbers, it is quite often noticed that a given number appears several times. For example, the probability that only unique numbers will appear in a sequence of five throws is quite low. For the first throw, all numbers are valid as no number can be repeated so far. For the second throw, only five numbers are now valid to avoid multiple occurrences of the same number, for the third throw only 4 numbers and so on: . Hence, the probability that a sequence of five rolls of a die results in unique numbers is only about 10 percent. Any sequence is possible, though we would naturally assume that the sequence 52461 is more random than 33444.

Figure 67: Clouds in the Sky

Source: Pixabay, 2019.

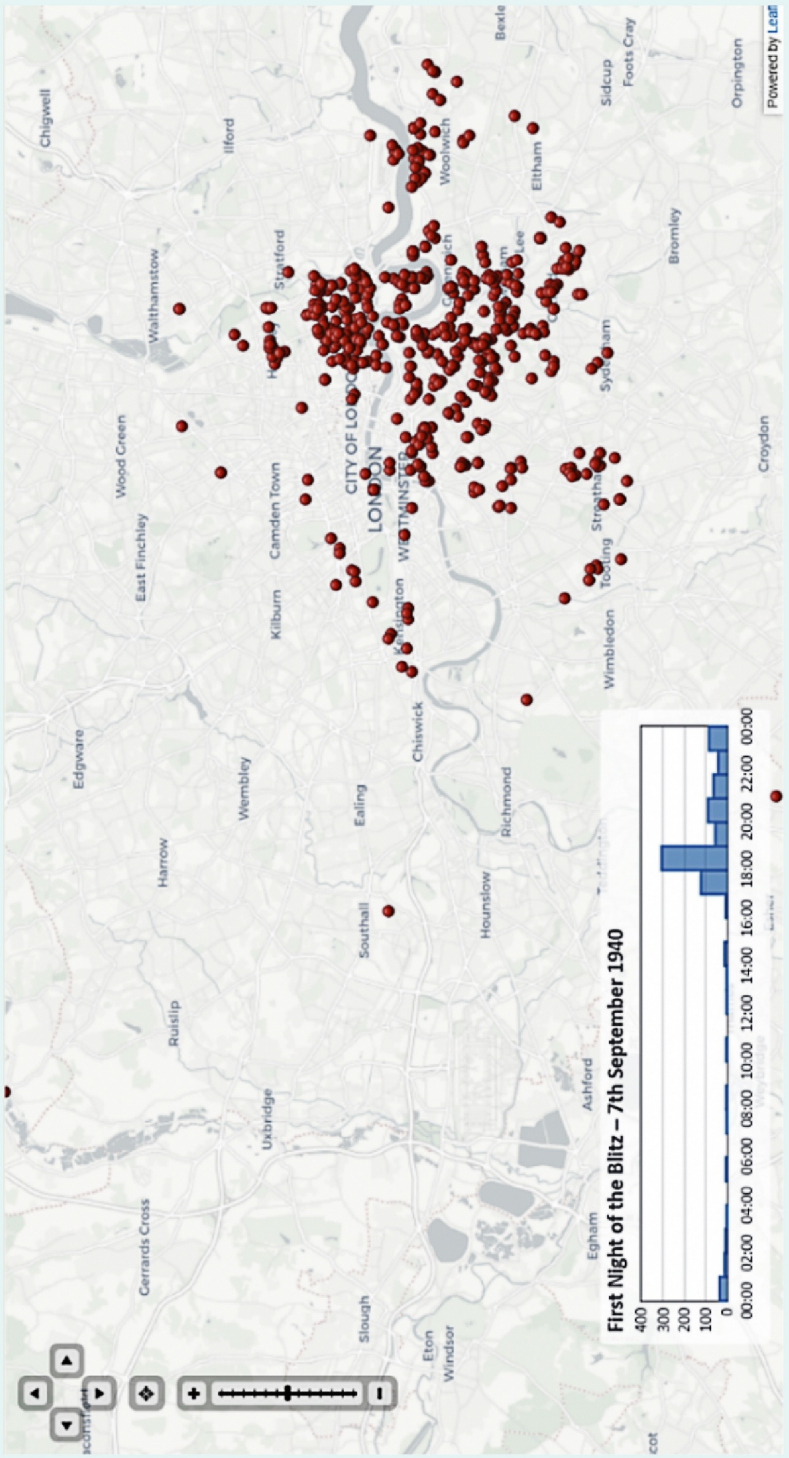
An activity often enjoyed by young children is spotting animals and shapes in the clouds. At night, we often look at the stars and assign names to apparent constellations of stars like Taurus, Libra, etc. From our vantage point at earth, the stars seem to be aligned in a certain pattern — though, of course, in the vast expanse of space the stars are very far away from each other and unrelated.

Figure 68: Star Signs

Source: Pixabay, 2019.

A further example is the placement of German bombs that hit London during the second world war (Kahneman, 2012, p. 115). When added to a map, the apparent pattern first suggested that the Germans must have had access to detailed intelligence information about potential targets. However, a careful statistical analysis showed that the pattern was compatible with a random distribution.

Figure 69: Map of German Bombs That Hit London in World War Two



Bombsight, 2011

Source: Bombsight, 2011.

Just as we tend to spot patterns in random events, we are equally likely to look for patterns and explanations in data even if we have no way of knowing if a pattern might be present. Our desire (of System 1) to find relationships and coherent stories is so strong that we tend to look for anything in the data which would result in a plausible story or a causal relationship.

Priming

Our thought processes are mostly influenced by events, words, sounds, smells, visual input, or even thoughts. For example, if you have recently read the word “eat” you are much more likely to read the following word “so_p” as “soup” than “soap”. However, if you have instead read “washing” recently, “soap” will come to your mind easier than “soup” (Kahneman, 2012, p. 52).

This effect is called “priming” and means that a previous observation or thought makes it more likely that you will think of certain things or act in a certain way. We are not even aware of the fact that we have been “primed”, and primed ideas can lead to further primed ideas. In the famous “Florida effect” experiment, students were asked to build four-word sentences from a set of five words. One group was given a list of words which were associated with elderly. After having completed their assignment, the students were asked to walk across the corridor to another room. Those students who were given a list of words associated with old age tended to walk slower and more deliberately than the students in the control group. Even though the world “old” was never mentioned, the students were primed to old age and even changed their movements unconsciously (Kahneman, 2012, p. 53).

Since ideas can prime further ideas, we have often no explanation why we arrive at the wrong intuition. For example, if you sat on a wobbly chair during the lunch break, you might be primed to “unstable”, “shaky”, “toppling over.” If your first task upon returning to work after lunch is to evaluate the predications of a machine learning model, you might be primed to “unstable predictions” and may look for supporting patterns even if there is no evidence to support such a hypothesis.

Priming has also far-reaching consequences when designing e.g. a dashboard or monitoring system. Depending on what is shown in which way at first, the audience of the dashboard will likely be primed for the whole time they analyze the dashboard. For example, if the focus is put on few but problematic issues first, the target audience will likely be primed to look for inconsistencies and abnormal events. Even if the metrics and charts show an overwhelmingly positive picture, if the audience are primed for negative issues, they will focus on them.

Halo Effect

The halo effect describes that we tend to associate our liking or dislike of a particular aspect with the whole system or person. For example, if we find a person easy to talk to, we tend to assume that we like the person as a whole with all their traits even if we don’t know anything about them (Kahneman, 2012, p. 82). Kahneman employs the example of describing the personality traits of two people:

- Alan: intelligent – industrious – impulsive – critical – stubborn – envious
- Ben: envious – stubborn – critical – impulsive – industrious – intelligent

Both Alan and Ben are described with the same words, yet most will view Alan more favorably than Ben (Kahneman, 2012, p. 82). This is because the sequence of the words matters. Our first impression of Alan and Ben is formed with the first one or two words which are much more significant than the last ones. Even though both Alan and Ben are described as "envious", the halo of the prominent position of this description in Ben's case gives rise to a much more negative description of Ben's personality compared to Alan's.

One way to reduce the halo effect is to separate out individual tasks or descriptions and focus on each aspect individually or derive information from many independent sources. This approach of decorrelating the influence is also one explanation why the "wisdom of the crowds" works. For example, while individuals may perform very poorly at estimating a quantity, e.g. the number of pennies in a jar, but the average over a large number of people tends to be fairly accurate. This is because each individual estimate is independent from the next. One person may wildly over- or underestimate the quantity, but the average tends to be close to the true value, provided that all estimates are indeed independent from another and don't share a common bias (Kahneman, 2012, p. 84).

Hindsight Bias

A very common mistake when evaluating an event or the predictions of a predictive model is to confuse which information was available at which time. Often, we tend to take all of the available information into account when we evaluate the outcome, including information which has become available only after the event has occurred. Consider the example of a professional football player who kicks the ball aimed at the goal and misses. Many will jump up in their seats and shout something like "I knew that player would miss!"

Figure 70: Footballer Aiming at the Goal

Source: Montgomery, 2017.

The fact that the player missed the goal is obvious once the event has happened, but the spectators had no way of knowing that the ball would miss the goal before it happened. The trajectory of the ball once kicked may have indicated that it would miss the goal — but, in this case, any chance for intervention had already passed. The last possible moment that the player could influence the event was before their foot hit the ball, and the last possible moment before someone external could intervene was a long time before that. For example, the coach could have swapped the player, but, at the time, all relevant information pointed to a favorable outcome.

Figure 71: Clothes on a Rack in a Retail Store

Source: Modern Retail, 2018.

Hindsight bias can often distort the evaluation of predictions in many scenarios. For example, a retailer has a wide range of clothes that they offer in their store. In order to always have optimal stock levels, future demand is predicted to optimize the order quantities from the wholesaler. In many cases, the quality of the prediction is then evaluated by comparing the sales of the top 10 percent of the most sold articles to their predicted demand, and, in many cases, these predictions will be far from the observed sales once they have occurred. This example has two aspects which are related to each other. Firstly, each prediction is, ideally, a full probability density distribution from which an optimal point estimator is derived to calculate the best order quantity. However, reducing the distribution to a single number also removes all information about the shape of the distribution, the volatility or variance, etc. If the predicted distribution has a high volatility, the optimal point estimator will have a large uncertainly which is often neglected. Then, in the final evaluation, an arbitrary cut-off is introduced based on information which was not available when the prediction was made. With hindsight knowledge, it becomes obvious which articles have sold best, but at the time the order had to be placed, this information was not available. All information available at the time was used to make the best possible prediction.

Kahneman summarizes this discrepancy as: “[Hindsight bias] leads observers to assess the quality of a decision, not by whether the process was sound, but by whether the outcome was good or bad” (Kahneman, 2012, p. 203).

Overconfidence

Figure 72: Overconfidence

Source: Pixabay, 2019.

Overconfidence is defined as a person's inability to critically judge their own performance. This bias comes in three variants:

- Overestimating their own performance
- Overestimating their own performance with regard to others
- Overestimating the precision of their own performance

In each case, a person assumes that their own performance is much better or much more precise than their peers or a machine. According to Kahneman, this is another consequence of System 1's tendency to build a coherent story in which their own estimate makes sense (Kahneman, 2012, p. 261). This way, even if System 2 was engaged and mental effort was spent on the task, System 2 will mostly accept the causal story presented by System 1.

In certain professions, showing or dealing with uncertainty would be considered a weakness. Medical practitioners, for example, often state a firm diagnosis or treatment plan even if the outcome is uncertain. However, discussing outcomes of treatments with patients in terms of probabilities would not be acceptable to the wider population.

Confirmation Bias

One of the fundamental principles of science is to test hypotheses by trying to find evidence against them. However, our own belief system tends to work the other way round. System 1 constructs a plausible and causal story from whatever data is available and passes this on to System 2 if it is engaged. Contrary to scientific practice, we tend to search for evidence which agrees with our beliefs (Kahneman, 2012, p. 81). This often leads to ignoring evidence that is incompatible with our beliefs, so we are more likely to look at data that confirms what we think we already know.

Anchoring Bias

The anchoring bias describes the way that people tend to stay close to a number that they have just encountered, even if that number has nothing to do with the task at hand. Hence, this number becomes the anchor around which the further development of events will pivot. In a way, the anchoring bias is related to priming. Being exposed to a certain number as an anchor primes us to pivot around this number, even if that number is random or has nothing to do with the current task (Kahneman, 2012, p. 119).

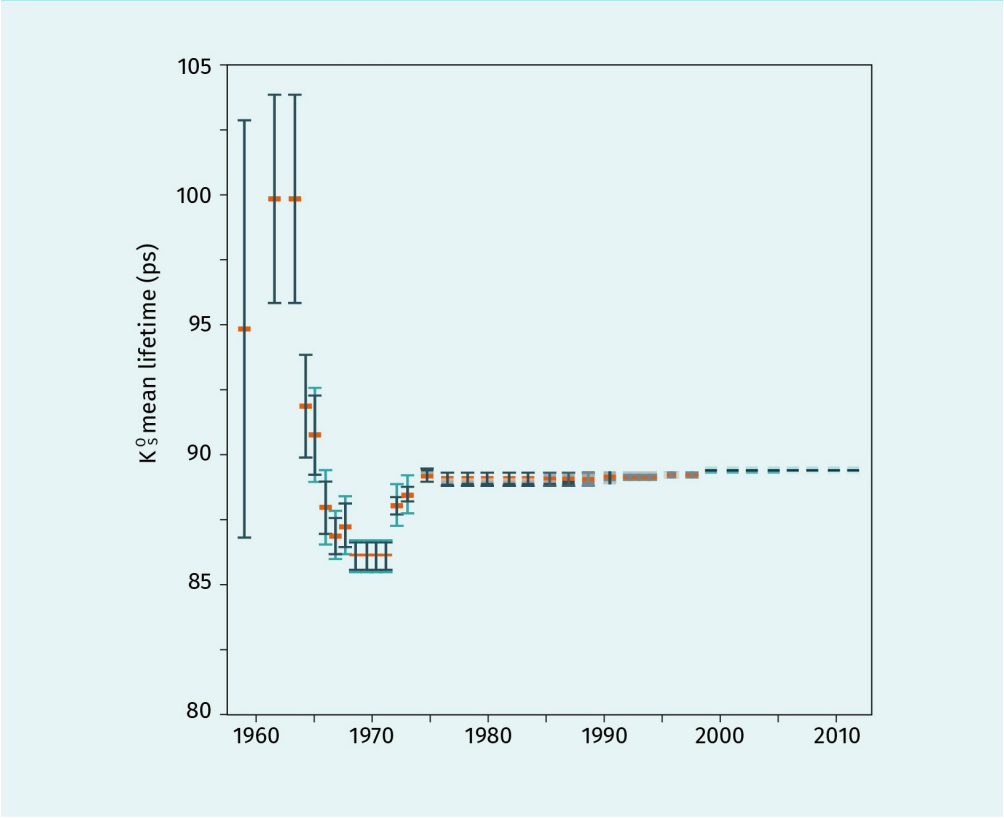
Figure 73: Haggling at a Market Stall

Source: Architects, 2019.

This effect can often be observed when haggling over prices at a market. Both the potential seller and buyer try to influence their counterpart with unrealistic high and low numbers. In a study, the authors found that the anchoring effect is also prevalent in professional settings. Judges who took part in the study were first asked to roll dice to simulate the prosecutor's demand for a sentence. The dice were loaded such that one group of judges would always roll low numbers whereas the other group would always roll high numbers. The study found that the numbers obtained from rolling the dice served as anchors for the final sentence, i.e. those judges who rolled higher numbers tended to hand out harsher sentences than those who rolled smaller numbers (Englich, Mussweiler, & Strack, 2006).

Even the best scientists are prone to the anchoring bias which can be illustrated by visualizing the result of scientific measurements over time. The following illustration shows the lifetime of a particular particle created in high-energy collisions (Particle Data Group, 2018):

Figure 74: Measured Lifetime of Kaons as a Function of Time



Source: Ulrich Kerzel, 2020.

The graph shows that measurements tend to cluster around certain values. The final (correct) measured value is around 90ps, however, measurements over time do not show a tendency to quickly converge to this value but stay close to the previously observed values: Even though they were not correct, scientists performing a new measurement are biased towards the number reported in previous publications.

Substituting Questions

If an answer to a question cannot be found immediately, System 1 is prone to finding a related question that is easier to answer (Kahneman, 2012, p. 97). In particular, if asked about probabilities, people tend to judge something else and believe that they have answered the question about probability. For example, when asked whether sales predictions are reasonable, an easier heuristic to answer would be, for example, to look at a time series of historic sales events and corresponding predictions, then judge whether the curves match. Although this may be a good indicator of the quality of the prediction, the evaluation is much more complex than a cursory glance. Also, the design of the visualization or dashboard plays a large role in the substituted heuristic: Does it prime for potential issues, even if these are rare, or does it give a more balanced overview?

5.2 Statistical Effects

Most people find it extremely difficult to judge events objectively without looking for causal explanations. On one hand, this is hard-wired into our brains in the form of System 1 which always tries to find causal relationships and construct a coherent story. Kahneman summarizes this as: “People are prone to applying causal thinking inappropriately, to situations that require statistical reasoning. Statistical thinking derives conclusions about individual cases from properties of categories and ensembles. Unfortunately, System 1 does not have the capability for this mode of reasoning; System 2 can learn to think statistically, but few people receive the necessary training” (Kahneman, 2012, p. 97).

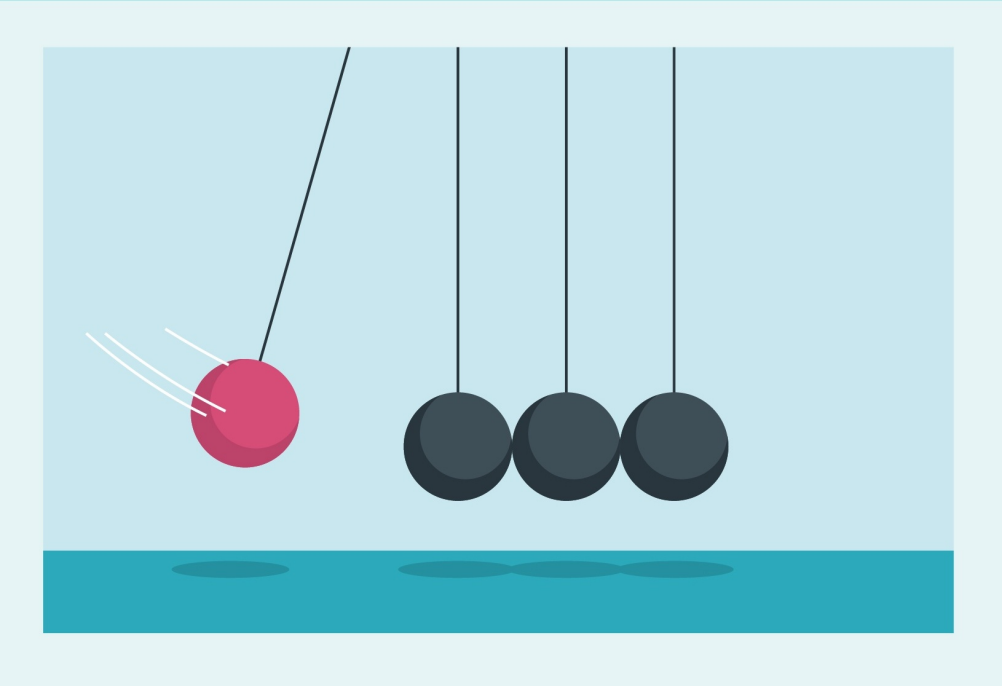
Figure 75: Curious Child

Source: White, 2019.

This can also be understood from the way we experience the world, starting when we are small children. Many everyday systems are deterministic and don't change their properties **stochastically**. This allows us to explore the world and learn from our actions, as repeated actions always lead to the same outcome. On the other hand, many systems have both a deterministic part and a stochastic part.

Stochastic
Random processes where the individual outcome cannot be predicted. At most, probabilities can be assigned to each possible outcome.

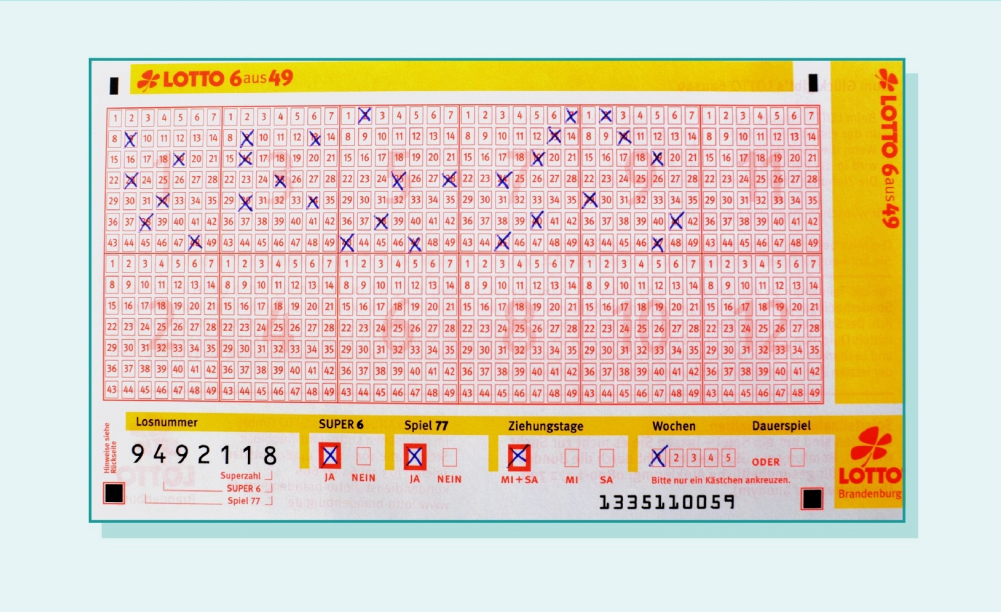
Figure 76: Newton's Cradle



Source: Pixabay, 2019.

The familiar Newton's cradle is fully deterministic, ignoring effects from friction and other losses. Once one ball has been lifted, the effect of the other balls can be calculated as the dynamics of the system and the initial conditions are known.

Figure 77: Lottery Ticket



Source: Storms, 2019.

Lottery, on the other hand, is not deterministic at all. Trying to predict the winning numbers of next week's lottery is futile and will never be successful. Most realistic systems are somewhere in between. The deterministic part allows predictions about future events to be made, and the stochastic part leads to uncertainty or volatility. This implies that the prediction of future events can only be made in terms of probabilities or probability density functions.

Discretization Effect

In many cases the observable events are discrete. For example, most goods can only be sold in fixed quantities, such as bottles, prepacked goods, or clothes. However, a prediction for the (average) sales per time-period will mostly be a continuous number. Considering the example of the sales of an expensive bottle of champagne, the manager of a store might expect that, on average, 1.2 bottles will be sold in any given week. The observed sales events are, of course, discrete. Some weeks, no bottles will be sold, in some other weeks one bottle, in a different week maybe two or three bottles. Even in the best case, the prediction and observed events cannot agree in this scenario. If one bottle was sold in a given week, a prediction of 1.2 bottles would imply a prediction error of 20 percent. Worse, if two bottles were sold, the prediction would be off by 80 percent. This interpretation is, of course, nonsense, as the effect is only due to the fact that the average predicted sales is a continuous number, but the goods can only be sold in discrete quantities. Since the average is slightly larger than one, it is to be expected that in some weeks two or more bottles will be sold.

Figure 78: Champagne Bottles That May Be Sold

Source: Pixabay, 2019.

Law of Small Numbers

Figure 79: Medical Devices

Source: LeSage, 2019.

The number of cases considered in an analysis can lead to biases or apparent contradictions. Wainer and Zwerling (2006) use the example of age-adjusted death rates due to kidney cancer in America. Highlighting the counties with the lowest 10 percent of cases, the resulting map shows mainly rural, sparsely inhabited counties in the Midwest, South, and West America. An immediate explanation springs to mind: Living in rural areas implies less pollution and easier access to healthy food and produce. However, highlighting counties with the top 10 percent of cases, the map mainly shows rural, sparsely inhabited countries in the Midwest, South, and West America. Although an intuitive explanation might be that rural areas have less access to health-care, the comparison shows that the lowest and highest decile of age-adjusted death-rates are in counties with a very similar structure.

Both intuitive explanations make sense at first. Living in rural areas implies an overall healthier lifestyle, less pollution, better food — but also less access to high-quality health-care. Hence, our System 1 and System 2 immediately jump to conclusions and construct a plausible story (Kahneman, 2012, p. 109). However, in this case, there is no underlying causal relationship or reason. Rural counties are generally sparsely populated, hence, the death-rate per county, normalized to the number of inhabitants, have a very large variance or volatility. Densely populated counties with many inhabitants have a smaller var-

iance as the cases are normalized to a much larger sample. Although it is tempting to look for a plausible causal story — and our cognitive biases tend to do this automatically — the main effect is statistical due to the large variance found in small samples.

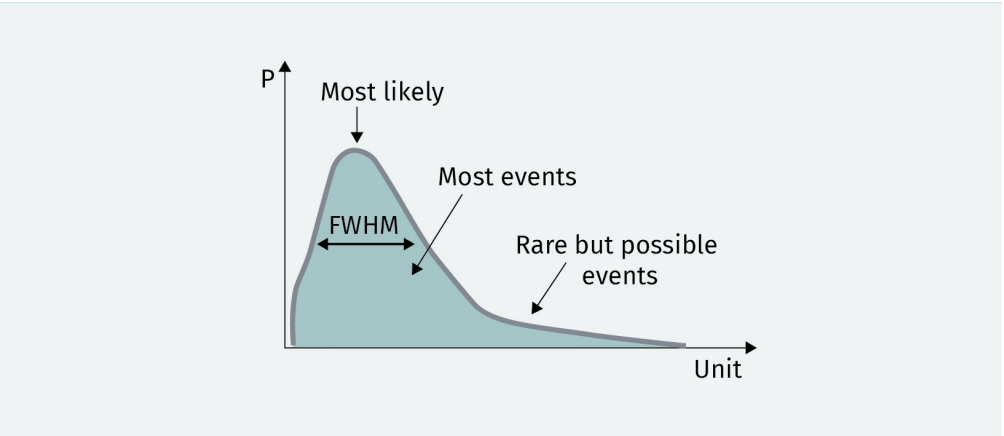
Outliers and black swan events

"Black Swan Events" Rare and unpredictable outliers which may have a big impact on the system considered.

The terminology of **black swan events** was shaped by Taleb (2010), going back to the Latin phrase "rara avis in terris nigroque simillima cygno" ("a good person is as rare as a black swan") (Puhvel, 1984). A black swan event is possible as future events can only be predicted as probabilities or probability density distributions — but have a very low probability of occurring. Hence, if they do occur, they are surprising and — similarly to the cognitive bias of the law of small numbers — we tend to look for simple explanations and a plausible causal story, rather than treating them as a purely statistical occurrence. These rare events are often given significant attention, looking for a causal relationship which may not be there.

Consider the following (predicted) probability density distribution: The shape is very asymmetric, the value zero represents a limit, i.e. all events have to be greater than zero. Most of the distribution is concentrated in a narrow region close to zero. However, the distribution has very long tails to high numbers with a very low probability. Hence, most of the events, maybe more than 90 percent, will occur in a small region close to zero where the bulk of the distribution is located. Events in the long tails will occur rarely — however, given the shape of the distribution, these events can occur as their probability is small but non-zero. As the observer is familiar with the "typical" occurrence, any event which originates from the tails will be considered as an outlier or Black Swan event, even if there is nothing special about this particular event apart from the fact that it rarely occurs.

Figure 80: Skewed Distribution With Long Tails



Source: Ulrich Kerzel, 2020.

5.3 Change Management: Transformation to a Data-Driven Company

Integrating Data Driven Decisions

Integrating artificial intelligence and data driven decisions into an organization or company poses significant challenges. One main fear is that the increased use of machine learning to drive decision-making will lead to a widespread loss of jobs. A PWC (Barriman & Hawksworth, 2017) study estimates that between 30—40 percent of jobs could be at risk due to automation. However, automation has defined our lives since humans started to create it. Automation ranges from mechanical aids to plough fields through the eons, to the first three industrial revolutions in which more and more mechanical tasks were automated and taken over by machines. As Bylund (2018) points out: “Nobody works for the sake of work—people strive to create value, which helps pay our salaries and feed our families.” Each transition time and industrial revolution was painful to many, where individuals were losing their jobs and struggling to find new ones. Roles disappeared, but new ones were created and society, as well as industry, changed as a whole.

This is why **change management** on the level of a business unit or organization, as well as on the wider scale of the society or economy, is crucial when man and machine start to work together, combining human and artificial intelligence. Many new jobs will emerge which require a higher level of training than before, and more traditional jobs will also change. For example, Ghafourifar (2017) highlights that the role of the manager will focus more on decision making, mentoring, and innovation as AI based systems take over more administrative tasks. This means that managers will be able to “manage” more and exercise their core qualities. This is certainly a plus for strong leaders, but it also means that managers can't hide in administrative tasks, thus exposing the need for change and further education in this role.

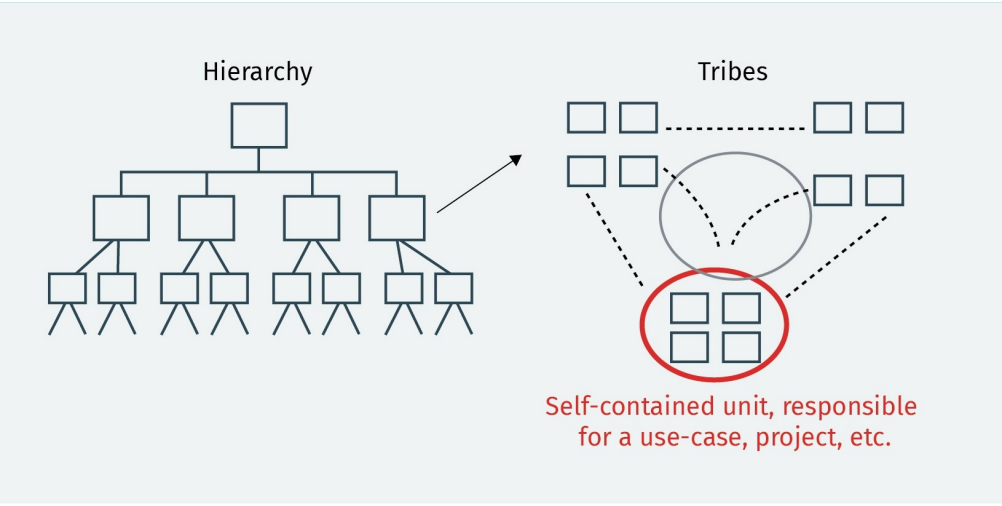
Change Management
A managerial process which aims to control the way that an organization changes to a new structure, or how employees can be prepared for large changes in the company.

This change is particularly visible if future events are predicted as probabilities or probability distributions. In order to derive operational decisions, thresholds have to be set. For example, at which probability would a transaction be considered fraudulent? Using a high threshold would focus on relatively few events which are almost certainly fraudulent. However, a number of fraudulent transactions will be missed as they look reasonably “normal”. Alternatively, one could choose a low threshold which would result in many false alarms which have to be verified manually. Any threshold is possible but the specific value needs to be determined in accordance to the business model and overall strategy. This implies that the managers responsible for such a task can make more powerful decisions than in the past, but they need a deeper understanding to be able to make such a decision.

A recent Deloitte study investigates the impact on organizational structures and concludes that the organizational structure will evolve from the traditional hierarchy towards tribes where teams focus on particular aspects or tasks and are loosely coupled with other teams within the organization. This transformation is a drastic change from traditional structures and poses significant challenges for all levels of the management. In the past, senior and C-level management were used to lead large units. Depending on the individ-

ual manager, this often led to silos within the organization with little contact between units. Furthermore, integrating new ideas was often difficult due to the long chain of bottom-up and top-down decision chains. Working in “tribes” has the immediate benefit that the teams can focus on particular topics or aspects and take ownership of them. Managers need to transform into supporters of the tribes instead of viewing themselves as supervisors of direct reports (Bersin et. al., 2017).

Figure 81: Organizational Charts: Hierarchy vs. Tribes



Source: Ulrich Kerzel, 2020.

As Parmenter (2015, p. 110) points out, this includes “transfer of power to the front line.” Employees should be able to take action to improve or rectify a situation instead of using long decision chains involving many layers of management.

McKinsey points out a number of pitfalls that explain why digital strategies fail. They found that many business leaders don’t have a clear understanding of how “digital” will impact businesses, but tend to see it as an upgraded IT system or isolated use case rather than a disruptive transformation which can potentially impact the whole business model (Bughin et. al., 2018). Furthermore, most managers learned the principles of economics a long time ago and find it challenging to imagine that digital products and services may change the way products are developed or used. Digital products also put the customer more often in direct contact with the manufacturer or service provider who can change or decimate the distribution chain and make products more interchangeable for the customer. Train or plane tickets are mainly bought via a smartphone app instead via travel agents, cloud offerings eliminate or reduce the need to buy and operate computer hardware at scale, etc. Not being ready for the change that digital business models bring endangers the operation of many companies. Another pitfall the authors of the McKinsey study identified is to overcompensate for the threat of a digital business model: “If I’m going to be disrupted, then I need to create something completely new” while forgetting that existing business models, customers, and market shares need to be readied for the digital age.

Davenport and Bean (2018) summarize that many companies have identified artificial intelligence and data-driven approaches as vital for their future strategies and have started to create new management roles such as the CDO (Chief Data Officer). However, they found that there is no clear understanding of the scope or responsibilities of these new management positions and how they are related to the other management roles in a company. The authors point out that a clear understanding of how to establish a data-driven culture throughout the company is vital to transform companies into data-driven businesses.

Similarly, DJ Patil, a former advisor to the White House, and Mason (2015) stress that building a successful data-driven enterprise is not about technical solutions or people with "crazy math skills," but rather a cultural change where data is accessible throughout the organization and forms the basis of discussions and decisions.

Leading change

Changing a project team, organization, or business is a difficult task, although change is generally inevitable. While many might agree that an organization needs to change, few would initially volunteer to change themselves or lead the change in a team. Parmenter stresses that the first part of the change process is selling the change to the management and, as with any sales process, this is not done by presenting logical arguments but by focusing on emotional drivers (Parmenter, 2015, p. 147).

In his seminal work "Leading Change", Kotter (2012) outlines an eight-step process:

1. Establish a sense of urgency: Convince at least 75 percent of the management team that the current status quo is more dangerous than change. Managers need to be convinced that change is the only way forward and that the current system could threaten their organization.
2. Form a powerful guiding coalition: In the second stage, an initial team with shared commitments should be formed. They should have enough power to lead the change effort and act outside normal hierarchies. This team should come from all relevant levels of the organization to make sure that all voices can be heard.
3. Create a vision: The successful change effort needs a vision so that everyone in the organization can understand why the change is necessary. The vision needs to be simple enough to be understood by every member of the organization within a few minutes.
4. Communicate the vision: Each member of the organization needs to be aware of the change effort and where the change will lead. The vision is the key to bringing everyone on board.
5. Empower others to act on the vision: New structures and systems are needed so that the vision can be acted upon. Managers should encourage risk-taking, and a "just-do-it" approach to non-traditional ideas, activities, and actions to help abandon the old structures.
6. Plan for and create short-term wins: Quick wins and improvements should be engineered as a motivating factor into the change effort. This highlights the success of the new approach and encourages everyone to participate. Praise should come from the highest levels such as the CEO to support the change effort.

7. Consolidate improvements and produce more change: The positive effect from the quick wins and improvements should fuel more change. Managers should support and promote those who participate in the change process and drive it forward.
8. Institutionalize new approaches: As new approaches driven by the change process start to take hold, managers need to establish the new approaches as the norm, for example by highlighting the connection between the new approach and success.

The key to establishing new approaches and structures is the employees. Each member of the organization needs to understand why change is necessary and, in particular, which change is required, which old structures need to be left behind, and which new best practices should be followed. Effective communication must also address employees' resistance and hesitations. Management should lead by example and show employees that they practice what they preach (Parmenter, 2015, p. 156) (Troyani, 2014).

Beckhard (1987) summarizes the success of the change process in the formula: D*V*F > R, where the combination of the driving forces dissatisfaction, vision, and first steps needs to be larger than the resistance to change.

Stereotypes

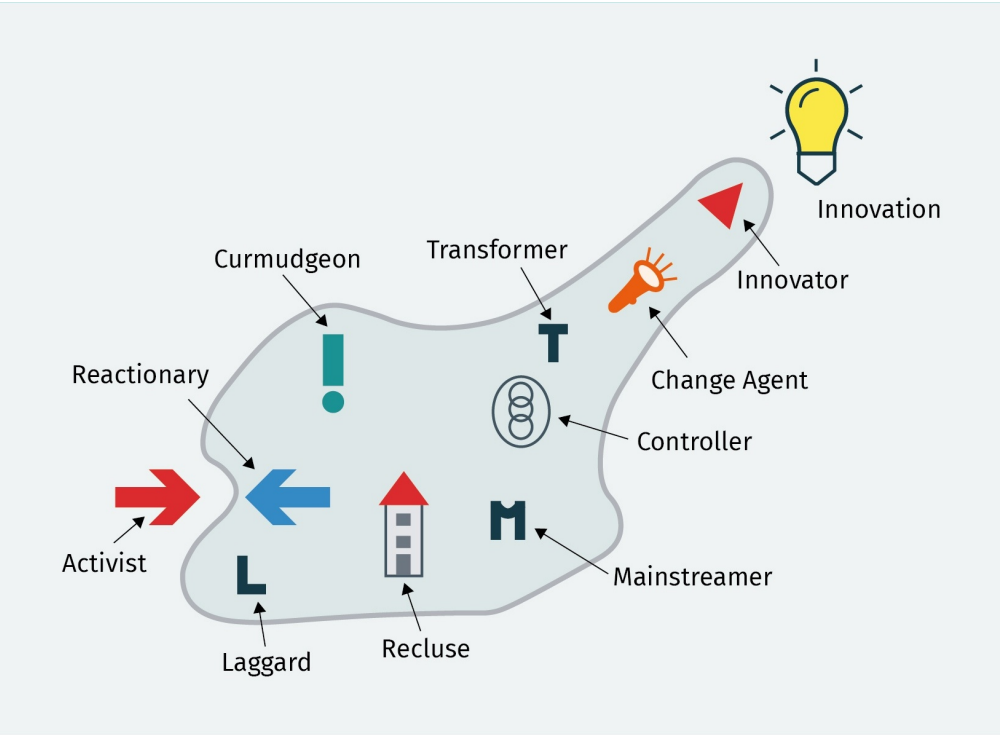
As noted earlier, organizational change is driven by people. Since people are all different, they tend to take a different attitude towards the change process. AtKisson (2016) has identified a number of typical roles that people take. He uses the image of an amoeba. This single-cell organism has the ability to alter its shape. In relation to the change process, the amoeba symbolizes the entire organization which stretches and changes its shape in order to reach a goal. The different constituents symbolize the typical roles people within the organization may take:

- Innovator: These are the drivers of the change process who try to bring new ideas in to the organization. They identify with the new ideas and are often very attached to them.
- Change agents: They are able to understand the new ideas and the benefit they would bring to the organization, but they also understand that the other people in the organization need to understand why the change is necessary and what needs to change in the first place. Change agents help others through the change process.
- Transformers: They are typically early adopters and promote the use of innovative approaches. Transformers are often well respected in their teams, though they are not necessarily a team leader or part of the management.
- Mainstreamers: They “go with the flow” and do what others do — they are not the first to change and see what others do. Once the change movement is sufficiently large, they join in.
- Laggards: They are almost the opposite of mainstreamers. They like the way things are and are reluctant to change. They are not opposed to change per se, but it is very difficult to push them out of their comfort zone.
- Reactionaries: They actively oppose the change and act against it. Their behavior can be motivated by a range of different reasons: They might genuinely think that the proposed change is bad for the organization, or they have something to lose, for example a (management) position or project.

- Curmudgeons: They are not only against the initiated change, but against everything. Curmudgeons always find a reason to complain.
- Recluses: Every organization has them. They tend to work differently compared to the normal processes as they have other interests or priorities. They might be well respected, such as researchers who work quietly in their area of expertise, but they keep themselves out of the everyday hustle and bustle. Seeking them out and integrating them in the change process can be very valuable as recluses might have considerable knowledge.
- Controllers: They are symbolized by the DNA of the amoeba. They make the final decisions about what will happen within the organization for example, the CEO or board of directors.
- Activists: They are outside of the organization and try to push the organization into a new direction. Activists see and vocalize what they think is wrong. Often, non-government organizations or journalists point to organizations which need to change. Aligning the change process with them can be helpful, e.g. by pointing them against the reactionaries so that they have less time to fight the change process within the organization.

When starting a change process, the picture of the amoeba is helpful to structure the change process. Before the actual change is initiated, the group driving the change should identify which people in the organization are likely to take which role. This should then be developed further into a strategy, including how to communicate with each role or person, how to anticipate their likely behavior, and how they can be integrated into the change process.

Figure 82: Amoeba of Change



Source: Ulrich Kerzel (2020), based on AtKisson, 2016.

SUMMARY

The most critical aspect in any project are the humans who design, implement, monitor, and evaluate it. Despite best intentions and careful consideration taken when designing the predictive model and its evaluation in the model and business centric ways, humans are still prone to confusion due to cognitive biases and statistical effects.

Much of this is rooted in our evolutionary origins. Behavioral patterns which have suited humans well across the millennia lure us into wrong interpretations, sometimes without us realizing what is happening. Being aware of these effects gives us the chance to compensate for them as much as possible.

Introducing a new approach or method into an organization is difficult and many efforts fail in practice because they cannot be implemented across the organization. Change management helps to guide the change process and establish new methods or technologies in a company. The image of the amoeba identifies typical roles often taken by people

within an organization when confronted with change. This forms the basis of designing a strategy to deal with the anticipated obstacles when implementing a change project within an organization.

BACKMATTER

LIST OF REFERENCES

Agresti, A., & Coull, B. (1998, May). Approximate is better than exact for interval estimation of binomial proportions app. *The American Statistician, 52(2),* 119—126.

Ali, M. (2018, May 18). DC, DPO visit fruit, vegetable market [article]. Retrieved from https://www.urdupoint.com/en/pakistan/dc-dpo-visit-fruit-vegetable-market-348008.html

American Statistical Association. (1915, December). Joint committee on standards for graphic presentation. *Publications of the American Statistical Association,14*(112), 790—797.

Architects, J. (2019, March 25). How landscape architecture and urban design can reduce crime [article]. Retrieved from https://land8.com/page/3/?cat=0

AtKisson, A. (2016, August 6). Where do you fit on the amoeba map? [article]. Retrieved from https://www.greenbiz.com/article/where-do-you-fit-amoeba-map

AusIndustry. (1999). *Key performance indicators manual: A practical guide for the best practice development, implementation and use of KPIs.* South Melbourne: Pitman Publishing.

Barriman, R., & Hawksworth, J. (2017, March). Will robots steal our jobs? The potential impact of automation on the UK and other major economies [article]. Retrieved from https://www.pwc.co.uk/economic-services/ukeo/pwcukeo-section-4-automation-march-2017-v2.pdf

Beckhard, R. (1987). *Organizational transitions: managing complex change* (2nd ed.). Reading: Addison-Wesley.

Berger, J. O. (1985). *Statistical decision theory and Bayesian analysis* (2nd ed.). New York, NY: Springer.

Bersin, J., McDowell, T., Rahnema, A., & van Dume, Y. (2017). The organization of the future: Arriving now [article]. Retrieved from https://www2.deloitte.com/content/dam/insights/us/articles/HCTrends_2017/DUP_Global-Human-capital-trends_2017.pdf

Bomb Sight (n.d.). Mapping the WW2 bomb census [website]. Retrieved from http://www.bombsight.org

Bughin, J., Catlin, T., Hirt, M., & Willmott, P. (2018, January). Why digital strategies fail [article]. Retrieved from https://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/why-digital-strategies-fail

Bylund, P. (2018, February 22). Automating jobs is how society makes progress [article]. Retrieved from https://qz.com/work/1212722/automating-jobs-is-how-society-makes-progress/

Chinchor, N. (1992). MUC-4 evaluation metrics: Association for computational linguistics. Retrieved from http://dl.acm.org/ft_gateway.cfm?id=1072067&type=pdf

Cook, J. (2005, November 2). Exact calculation of beta inequalities [report]. Retrieved from https://www.johndcook.com/UTMDABTR-005-05.pdf

Daily News Hungary. (2019, July 5). Industrial applications of optical fiber [article]. Retrieved from https://dailynewshungary.com/industrial-applications-of-optical-fiber/

Davenport, T. H., & Bean, R. (2018, February 15). Big companies are embracing analytics, but most still don't have a data-driven culture [article]. Retrieved from https://hbr.org/2018/02/big-companies-are-embracing-analytics-but-most-still-dont-have-a-data-driven-culture

Doran, G. T. (1981). There's a S.M.A.R.T. way to write management's goals and objectives. *Management Review*, *70*, 35—36.

Dorard, L. (2016). My methodology: The machine learning canvas [website]. Retrieved from https://www.louisdorard.com/machine-learning-canvas

Egan, J. P. (1975). *Signal detection theory and ROC analysis*. New York, NY: Academic Press.

Englich, B., Mussweiler, T., & Strack, F. (2006). Playing dice with criminal sentences: The influence of irrelevant anchors on experts' judicial decision making. *Personality and Social Psychology Bulletin, 32*(2), 188—200.

Fleming, P. J., & Wallace, J. J. (1986). How not to lie with statistics: The correct way to summarize benchmark results. *Communications of the ACM, 29*(3), 218—221.

Franconeri, S. (2019). Which visualization?: A quick reference [image]. Retrieved from http://experception.net/Franconeri_ExperCeptionDotNet_DataVisQuickRef.pdf

Friendly, M., & Denis, D. (2005). The early origins and development of the scatterplot. *Journal of the History of the Behavioral Sciences*, *41*(2), 103—130.

Ghafourifar, B. (2017, September 22). AI will re-define the role of the manager [article]. Retrieved from https://venturebeat.com/2017/09/22/ai-will-redefine-the-role-of-manager/

Harris, R. L. (1996). *Information graphics: A comprehensive illustrated reference*. Oxford: Oxford University Press.

Haskell, A. C. (1919). *How to make and use graphic charts*. Charleston, SC: Nabu Press.

Hyndman, R. J. (2006). Another look at forecast accuracy metrics for intermittent demand. *International Journal of Applied Forecasting, 4*, 43—46.

Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting, 22*(4), 679—688.

Jm3 (2016, September 5). Cognitive Bias Codex: 180+ biases [image]. Retrieved from https://commons.wikimedia.org/wiki/File:Cognitive_Bias_Codex_-_180%2B_biases,_designed_by_John_Manoogian_III_(jm3).jpg

Kahneman, D. (2012). *Thinking, fast and slow*. New York, NY: Penguin Books.

Kotter, J. P. (2012). *Leading change.* Boston, MA: Harvard Business Review Press.

LeSage, K. (2019, February 27). Natural ways to lower blood pressure [blog]. Retrieved from https://blogs.sas.com/content/efs/2019/02/27/natural-ways-to-lower-blood-pressure/

Lindsay, D. (2016, December 23). This video of the “planet earth II” animals screaming like humans is genius [article]. Retrieved from https://sg.news.yahoo.com/video-planet-earth-ii-animals-203709281.html?guccounter=1&guce_referrer=aHR0cHM6Ly93d3cuZ29vZ2xlLmRlLw&guce_referrer_sig=AQAAAAvB1c-RfiLOc1EiQkXZHzXLl_hlShOQQBm47Cqa2ELK80FVukwEfXXDPMXamXbwbe9uGXZzTprpJfI-3dF8v_3CEUGd_oeyw8L8nc7Eyk9exGff0F82qhyEyzGkbztkdOzSGFok4bgFQ1HV06ltW8UOJEqU86NbIhgQvFdNPUJw9

Linpack (2019a). Google analytics dashboard [image]. Retrieved from https://linpack-for-tableau.com/tableau-dashboard/google-analytics-dashboard/

Linpack (2019b). Sales dashboard cockpit [image]. Retrieved from https://linpack-for-tableau.com/tableau-dashboard/sales-dashboard/

Mason, H. & Patil, D. (2015). *Data driven*. Sebastopol, CA: O'Reilly Media.

Modern Retail. (2018, 22 October). Bricks and clicks lead the way for the UK's high streets [article]. Retrieved from https://www.modernretail.co.uk/bricks-and-clicks-lead-the-way-for-the-uks-high-streets/

Montgomery, C. (2017, August 19). 5 exercices fondamentaux pour développer son explosivité au football [article]. Retrieved from http://jygoal.fr/5-exercices-fondamentaux-pour-developper-son-explosivite-au-football/

Osterwalder, A. (2004). *The business model ontology: A proposition in a design science approach* (Unpublished doctoral dissertation). Université de Lausanne, Switzerland.

Osterwalder, A., & Pigneur, Y. (2010). *Business model generation*. Hoboken, NJ: Wiley.

Parmenter, D. (2015). *Key performance indicators: Developing, implementing, and using winning KPIs*. Hoboken, NJ: John Wiley & Sons.

Particle Data Group (2018). History plots [image]. Retrieved from http://pdg.lbl.gov/2018/reviews/rpp2018-rev-history-plots.pdf

Pearson, E. S., & Neyman, J. (1930). On the problem of two samples. *Bulletin of the Academy of Polish Sciences,* 73—96.

Pearson, K. (1895). Contributions to the mathematical theory of evolution. II. Skew variation in homogeneous material. *Philosophical Transactions of the Royal Society A: Mathematical, Physical, and Engineering Sciences*, *186*(0), 343—414.

Perry, J. W., Kent, A., & Berry, M. M. (1955). Machine literature searching X. Machine language; factors underlying its design and development. *American Documentation*, *6*(4), 242—254.

Porter, M. M., & Niksiar, P. (2018). Multidimensional mechanics: Performance mapping of natural biological systems using permutated radar charts. *PloS One*, *13*(9).

Puhvel, J. (1984). The origin of Etruscan tusna swan. *The American Journal of Philology, 105*(2), 209.

Shneiderman, B. (1992). Tree visualization with tree-maps: 2-d space-filling approach. *ACM Transactions on Graphics*, *11*(1), 92—99.

Spear, M. E. (1952). *Charting statistics*. New York: McGraw-Hill.

Stehman, S. V. (1997). Selecting and interpreting measures of thematic classification accuracy. *Remote Sensing of Environment*, *62*(1), 77—89.

Storms, H. (2019, March 11). $273 million lottery winner to reward clerk who kept his lost winning ticket safe [article]. Retrieved from https://www.inquisitr.com/5336508/273-million-lottery-winner-to-reward-clerk-who-kept-his-winning-ticket-safe-for-him-when-he-lost-it/

Strategyzer. (2016). Why use the business model canvas? [website]. Retrieved from https://www.strategyzer.com/canvas/business-model-canvas

Strickland, E. (2018, June 25). Layoffs at Watson Health reveal IBM‘s problem with AI [article]. Retrieved from https://spectrum.ieee.org/the-human-os/robotics/artificial-intelligence/layoffs-at-watson-health-reveal-ibms-problem-with-ai

Taleb, N. N. (2010). *The black swan: The impact of the highly improbable*. New York, NY: Random House.

Troyani, L. (2014, December 9). 10 quick and easy ways to begin organizational change [article]. Retrieved from https://www.tinypulse.com/blog/10-quick-and-easy-ways-to-begin-organizational-change

Wainer, H., & Zwerling, H. L. (2006). Evidence that smaller schools do not improve student achievement. *Phi Delta Kappan, 88*(4), 300—303.

Waters, R. (2016, January 5). Artificial intelligence: Can Watson save IBM? [article]. Retrieved from https://www.ft.com/content/dced8150-b300-11e5-8358-9a82b43f6b2f

Waymo. (n.d.). We're building the world's most experienced driver [website]. Retrieved from https://waymo.com/

White, N. (2019, January 21). MacDonald's being delivered to schools amid childhood diabetes crisis [article]. Retrieved from https://iwantastandingdesk.com/blogs/standing-desk-height-adjustable-desk-news/macdonald-s-being-delivered-to-schools-amid-childhood-diabetes-outbreak

LIST OF TABLES AND FIGURES

IU Internationale Hochschule GmbH
IU International University of Applied Sciences
Juri-Gagarin-Ring 152
D-99084 Erfurt

Mailing Address
Albert-Proeller-Straße 15-19
D-86675 Buchdorf

media@iu.org
www.iu.org

Help & Contacts (FAQ)
On myCampus you can always find answers to questions concerning your studies.