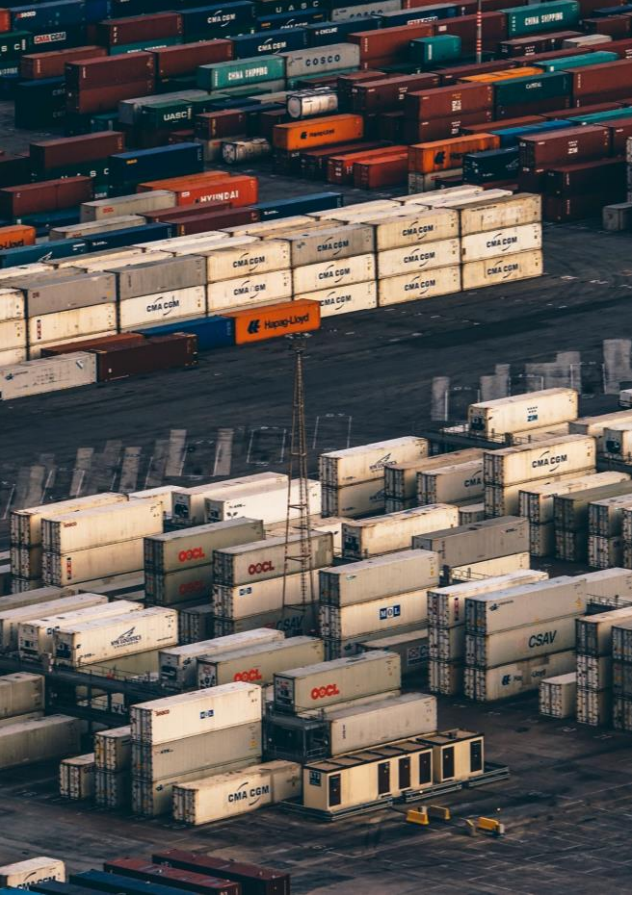
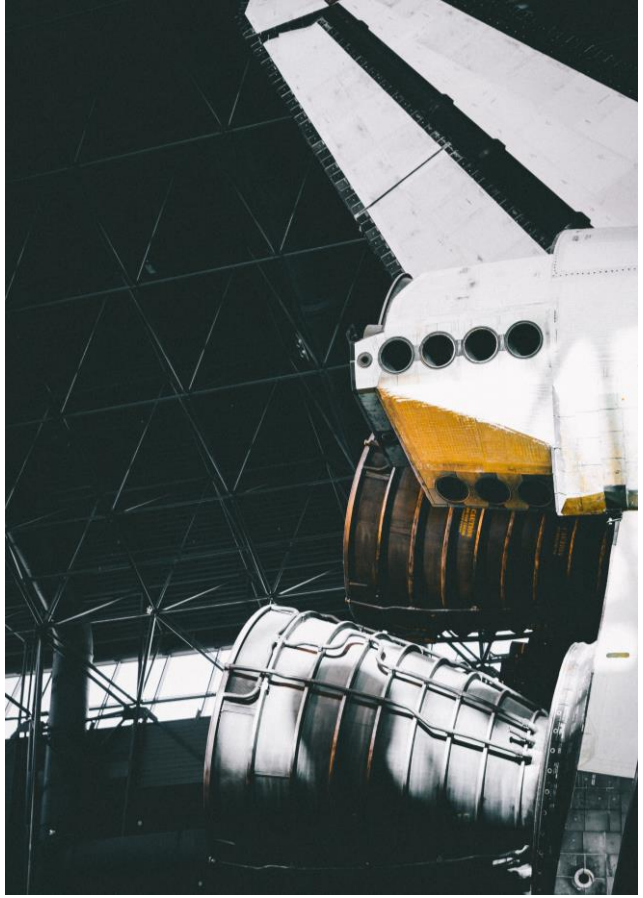
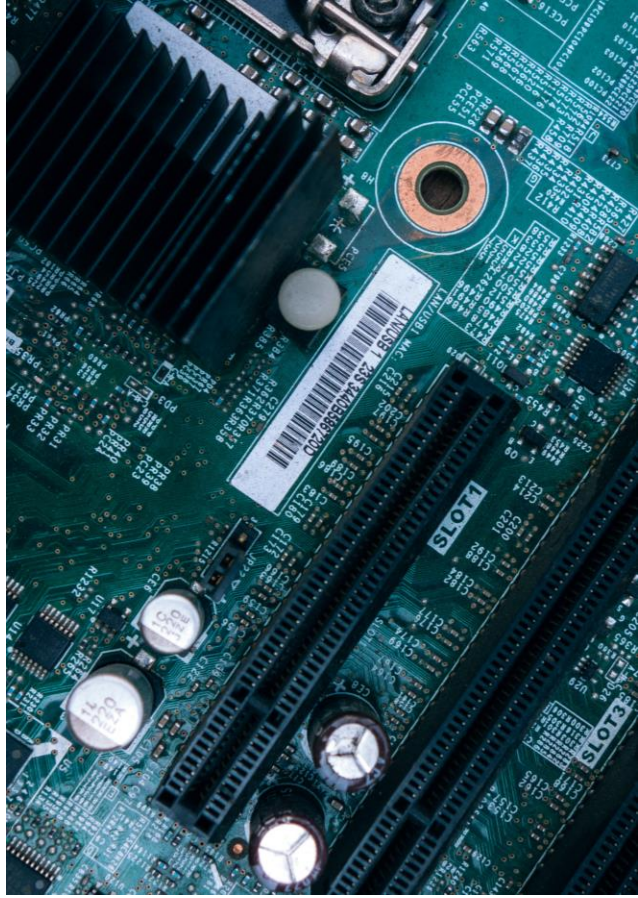




THE DATA SCIENCE BOOKLET

365 DataScience

The problem:

WHAT IS DATA SCIENCE?

Data Science is a term that escapes any single complete definition, which makes it difficult to use, especially if the goal is to use it correctly. Most articles and publications use the phrase freely, with the assumption that it is universally understood. However, data science – its methods, goals, and applications – evolve with time, and technology. Data science 25 years ago referred to gathering and cleaning datasets and then applying statistical methods to that data. In 2018, data science has grown to a field that encompasses data analysis, predictive analytics, data mining, business intelligence, machine learning, deep learning and so much more.

The solution:

What this booklet is designed to do is gather all those terms, everything you've heard about data science, and around data science, and order them neatly on a canvas.

We will show you how each term fits into the data science big picture, what's the timeline of data science activities, and how some of the core pillars of data science work.

Why? Because we want you to be best prepared when making the decision to walk into the data science club. And because data science is our area of expertise and we believe in sharing quality knowledge.

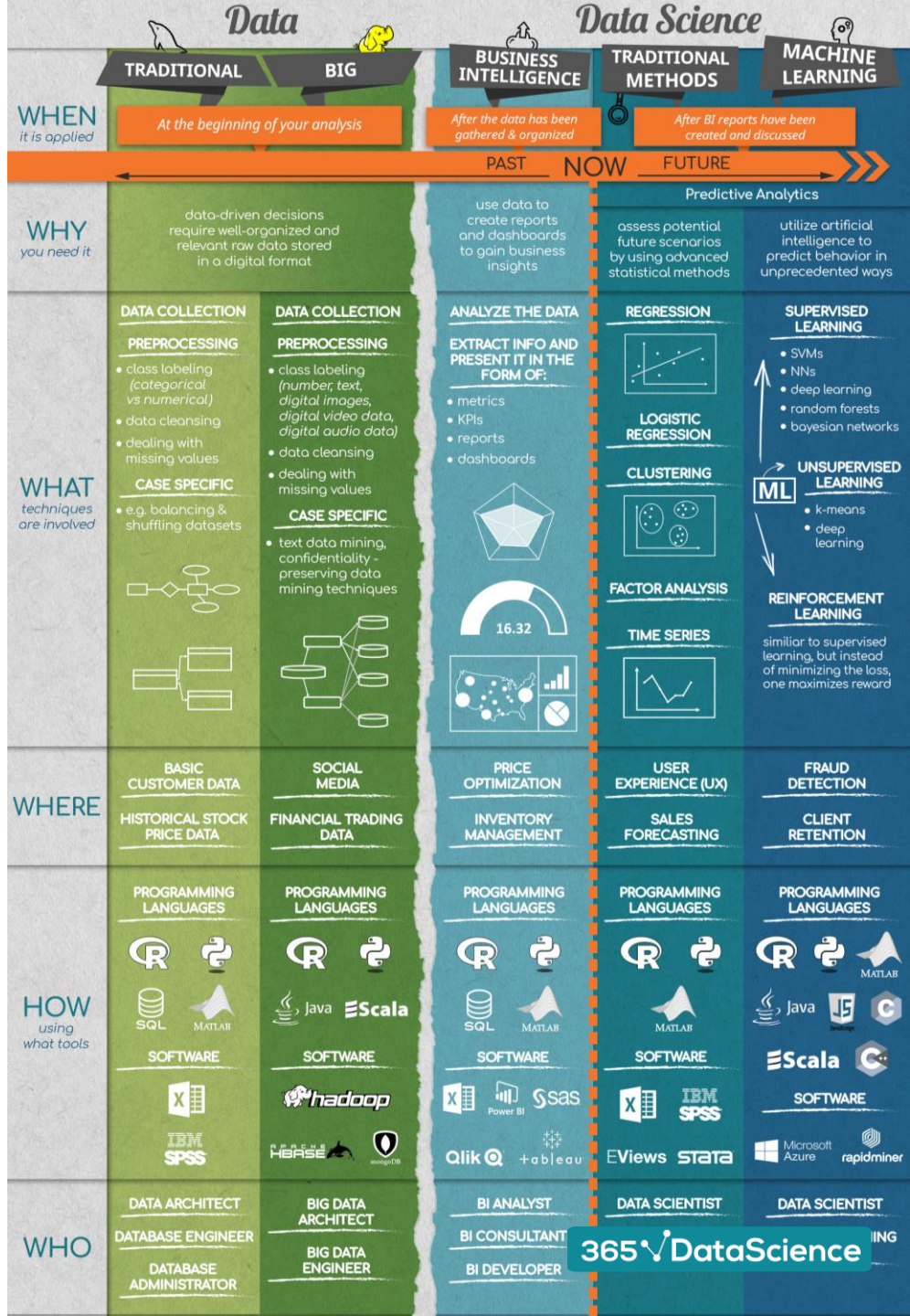
CONTENTS

LET'S BREAK THINGS DOWN

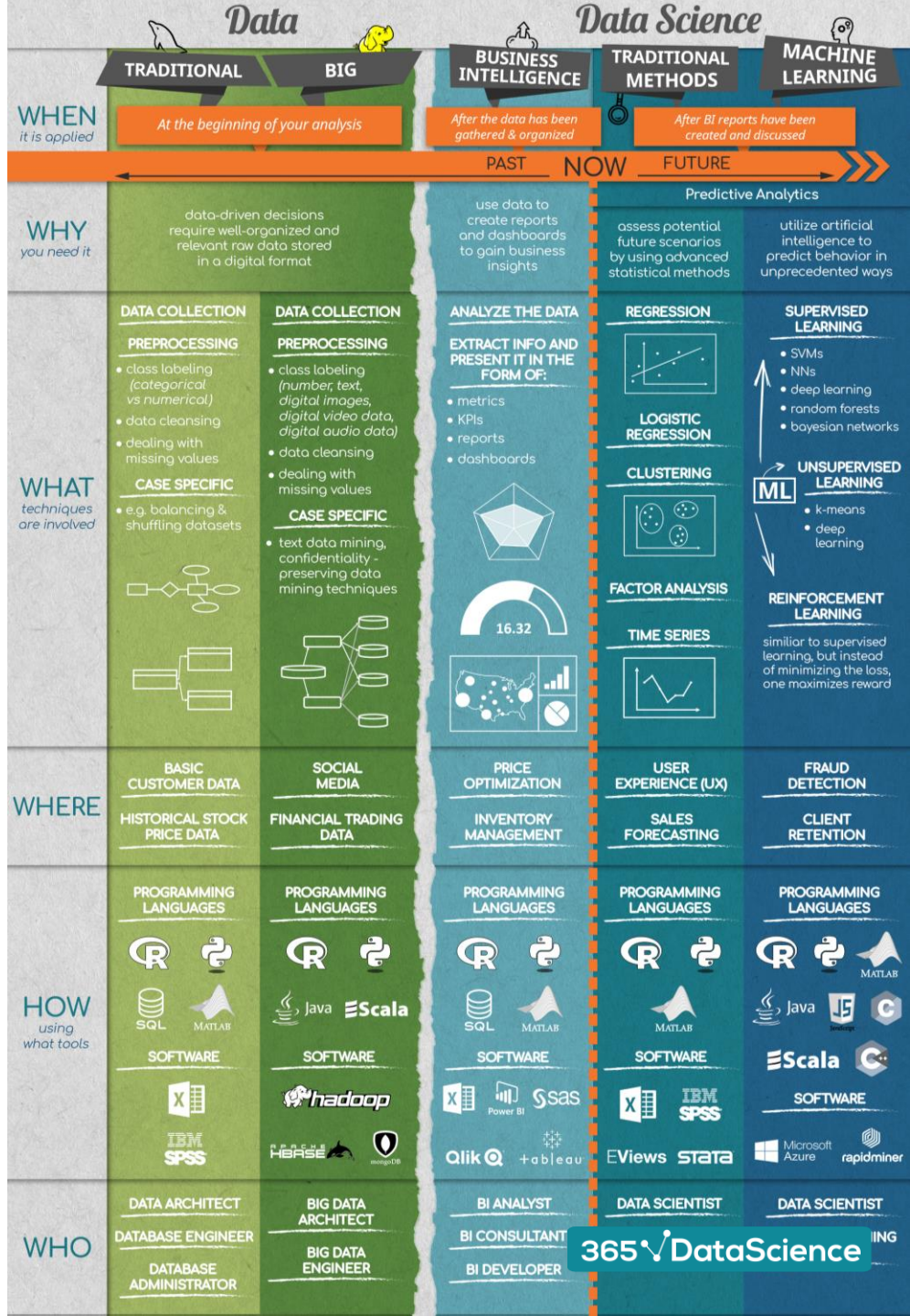
1. The data in data science
   - What do you do to data (big and traditional)
   - Where does data come from
   - Who handles the data

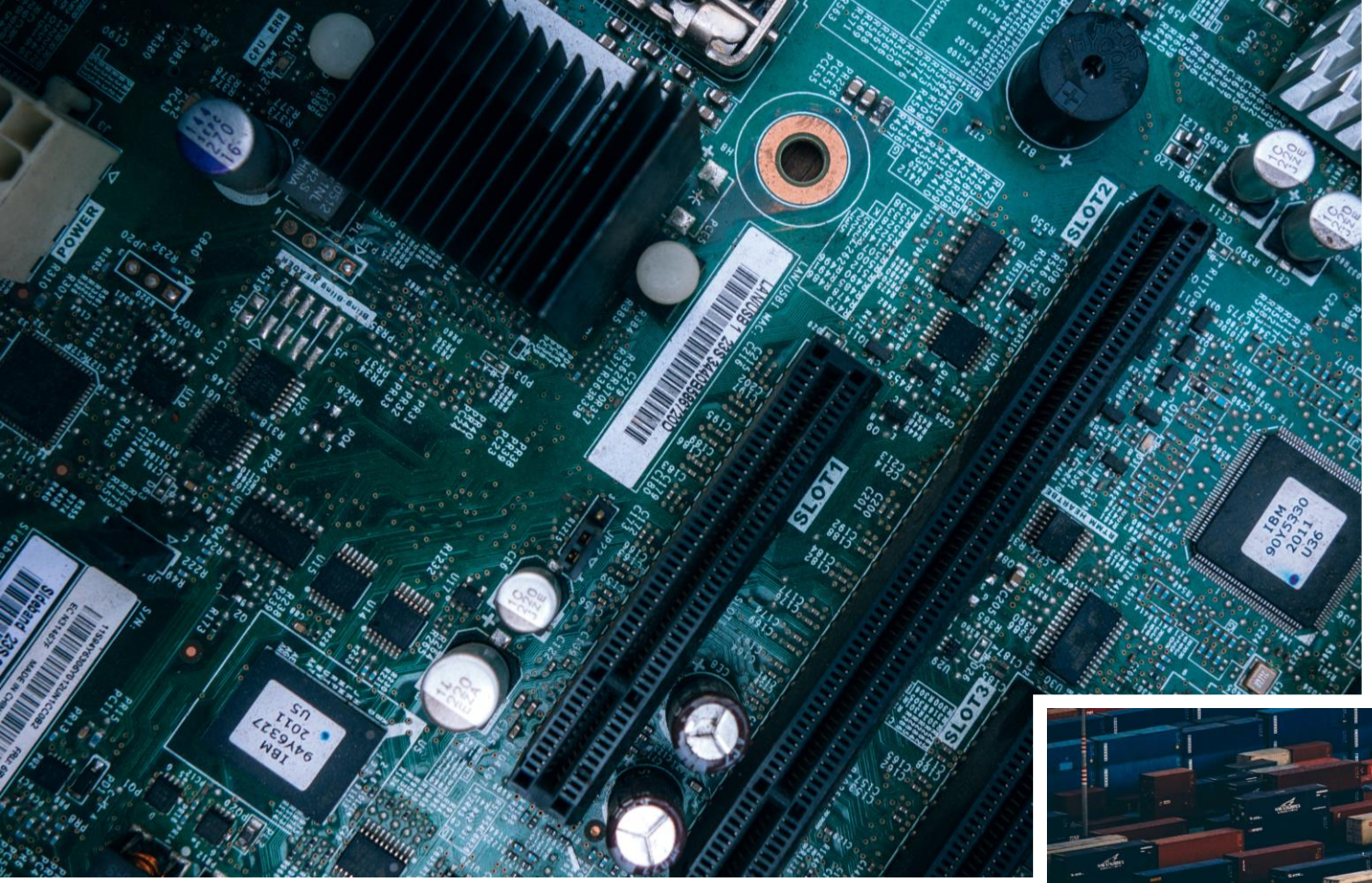
2. Data science explaining the past
   - What does Business Intelligence do
   - Where is Business Intelligence used
   - Who does the BI branch of data science

3. Data science predicting the future
   TRADITIONAL METHODS
   - Traditional forecasting methods
   - Where are they used
   - Who uses traditional forecasting methods

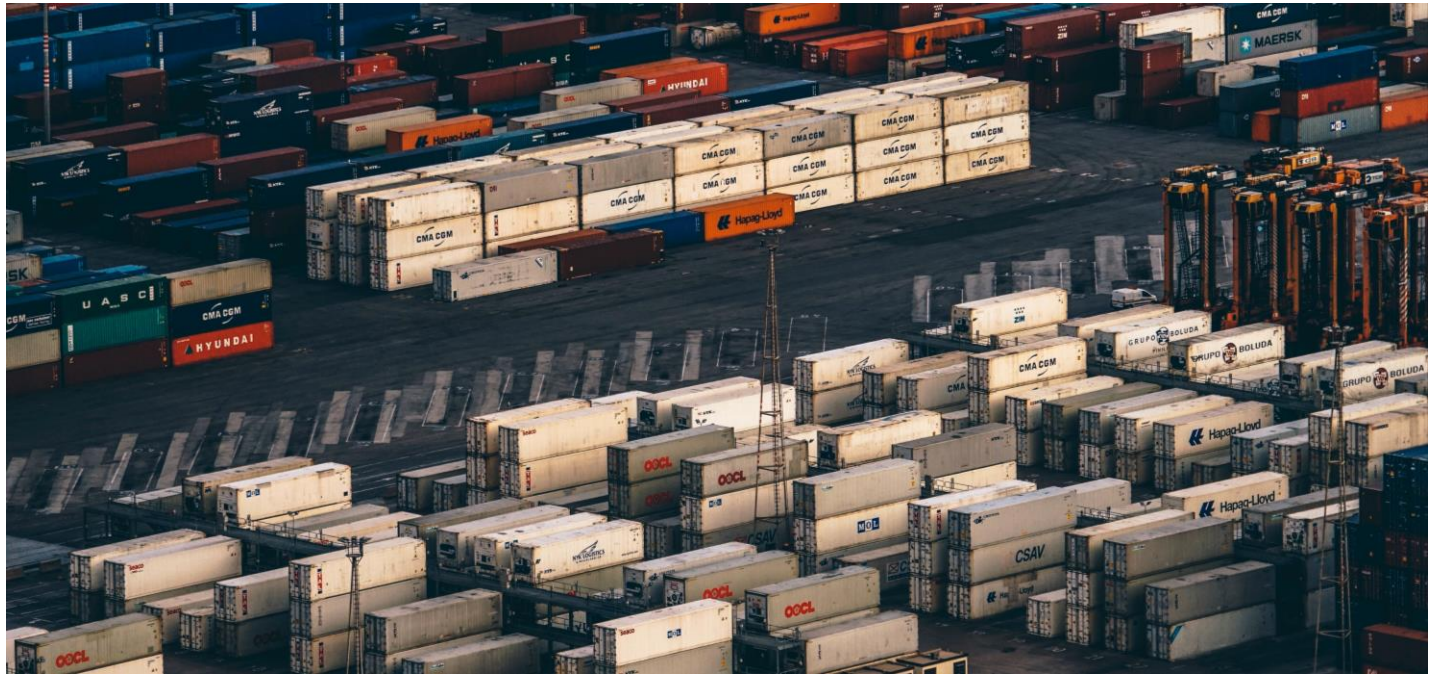


CONTENTS
(CONTINUED)




POWER
SLOT1
SLOT2
SLOT3
IBM
90Y5330
2011
U36


img
IX


MAERSK
HYUNDAI
CMA CGM
Hapag-Lloyd
OOCL
GRUPO BOLUDA

DATA

TRADITIONAL DATA

Data is the foundation of data science; it is the material on which all the analyses are based. In the context of data science, there are two types of data: traditional and big data.

Traditional data is data that is structured and stored in databases which can be managed from one computer; it is in table format, containing numeric or text values.

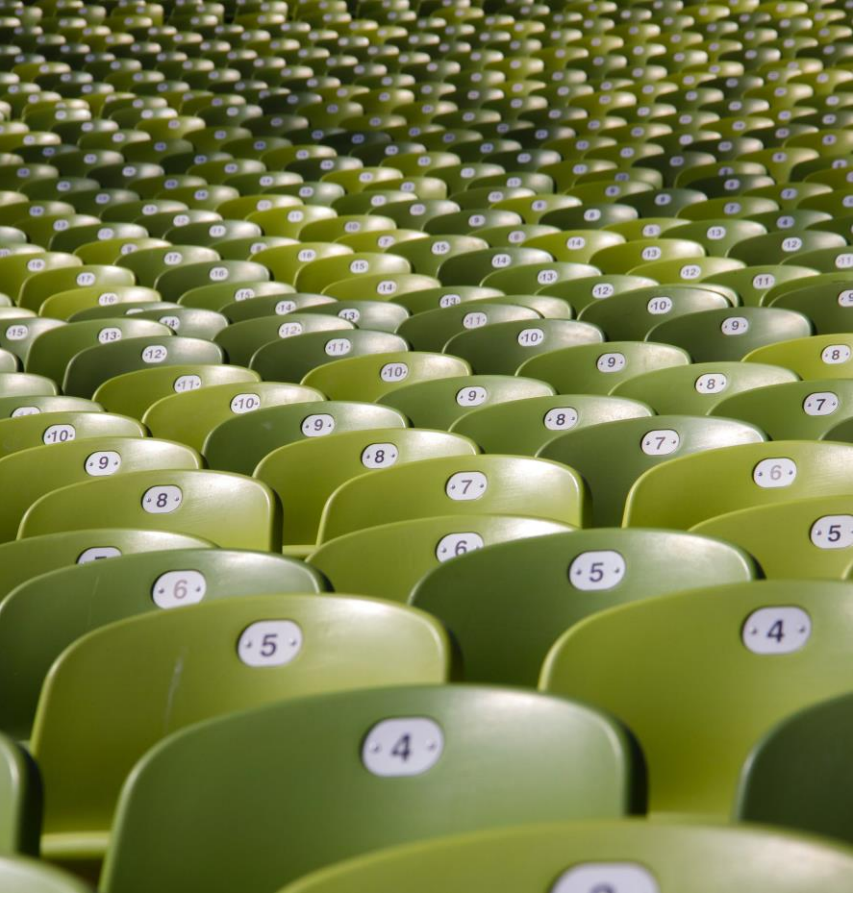
The term "*traditional data*" is not part of the official vernacular. It is something we are introducing for clarity. We believe it helps emphasize the distinction between **big data** and… non-big data.

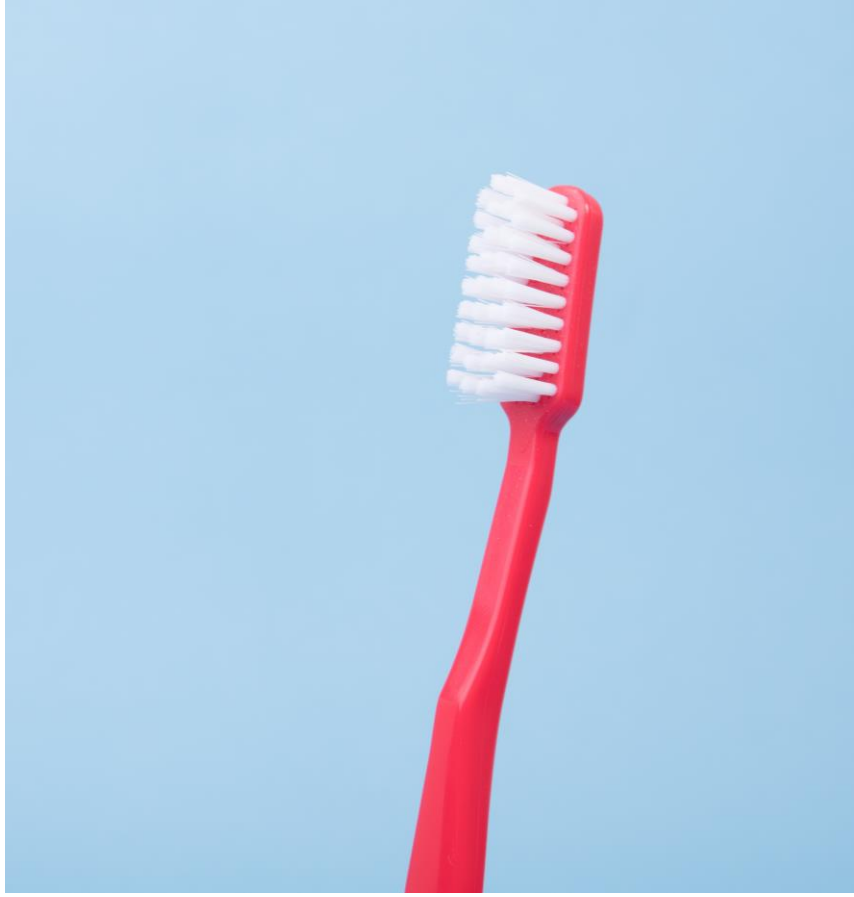
BIG DATA

Big data is bigger than traditional data, but not in the trivial sense. This is extremely large data, distributed across a network of computers, but it is not *just* characterized by its volume. This data can be in various formats; it can be structured, semi-structured or unstructured.

You will often see big data characterized by the letter "V". This stems from "**the 3Vs of big data**":

- Variety – numbers, text, but also images, audio, mobile data, etc. Big data can be in various formats
- Velocity – it is retrieved and computed in real time
- Volume – big data is measured in tera-, peta-, exa-bytes (that's 1 million terabytes).





WHAT DO YOU DO TO DATA

Before the answer...

CONSIDER THIS

Imagine the following scenario: you are an analyst for a large company and your manager approaches you with the following request:

"We need to consider client satisfaction in the next quarter, so we can predict churn rate. Oversee the process and come up with some numbers."

OR

"We have an enormous amount of customer data from the previous quarter. Can you oversee the analysis and deliver an approximation of churn rates for the next quarter?"

Give yourself a minute to name the difference between the two scenarios.

Before the answer...

CONSIDER THIS

Imagine the following scenario: you are an analyst for a large company and your manager approaches you with the following request:

NO DATA

"We need to consider client satisfaction in the next quarter, so we can predict churn rate. Oversee the process and come up with some numbers."

RAW DATA

"We have an enormous amount of customer data from the previous quarter. Can you oversee the analysis and deliver an approximation of churn rates for the next quarter?"

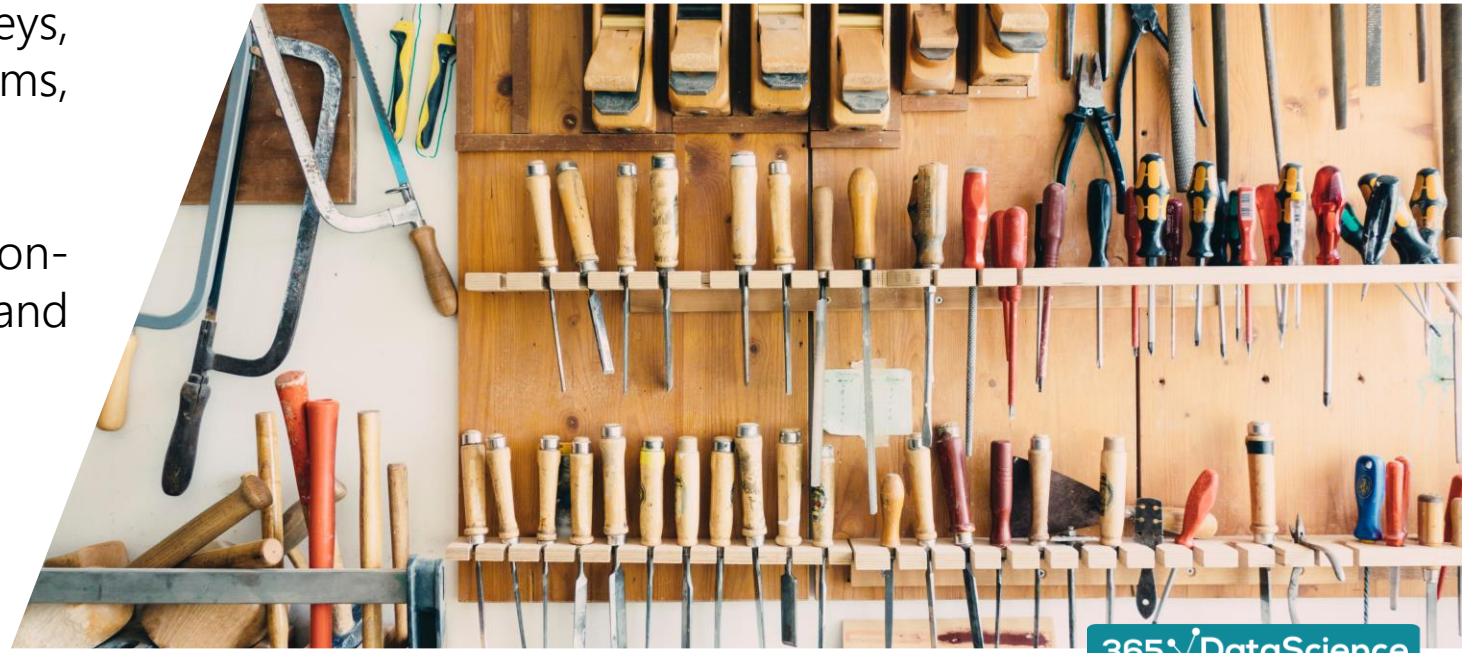
PREPROCESSING

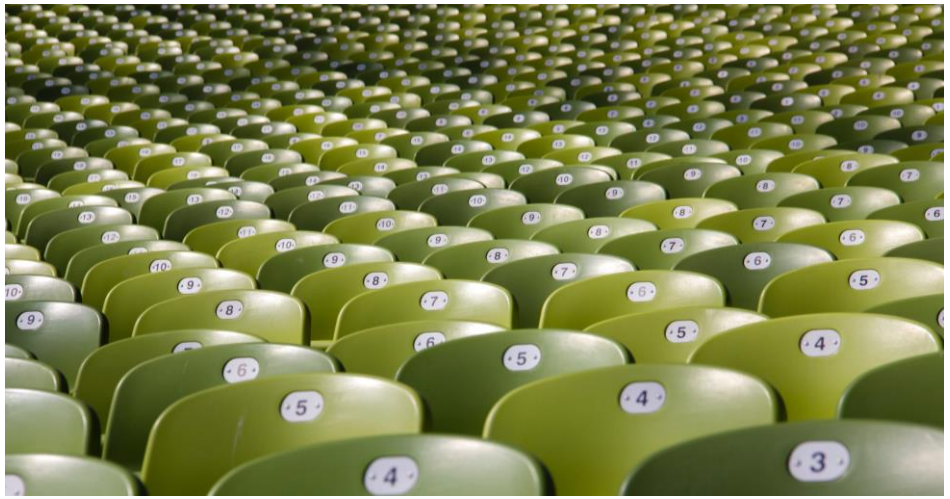
PRELIMINARY DATA SCIENCE

Regardless of whether a data scientist is given just-collected data or they have to go through the gathering process themselves, this data is going to be in **raw format**. This data can come from surveys, or through automatic data collection paradigms, like cookies on a website.

Raw data is untouched data that needs to be converted into a form that is more understandable and useful for further processing.

The group of operations that do this are called **preprocessing**.





PREPROCESSING

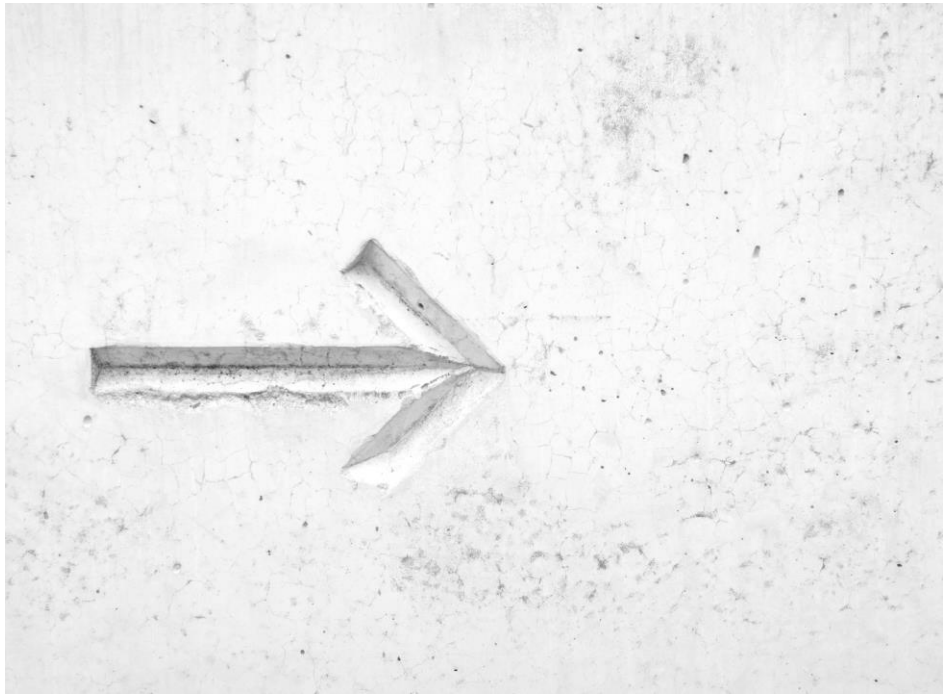
COMMON PROCESSES

Class-labeling the observations

This consists of arranging data by category, or labelling data points to the correct data type (e.g., for traditional data this can be numerical / categorical; for big data – txt, digital image, digital audio).

Data cleansing / data scrubbing

Dealing with inconsistent data, such as misspelled categories & missing values.



PREPROCESSING

COMMON PROCESSES

Data balancing

If the categories in the data contain an unequal number of observations, they may not be representative of the population.

Balancing methods, like extracting an equal number of observations for each category, and preparing *that* for processing, fix the issue.

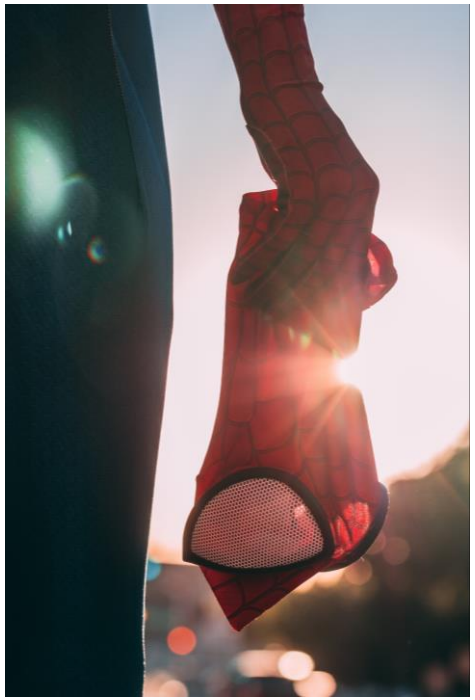
PREPROCESSING

COMMON PROCESSES

Data shuffling

Re-arranging data points to eliminate unwanted patterns and improve predictive performance further on.

This is applied, for example, if the first 100 observations in the data are from the first 100 people who have used a website; the data isn't randomized, and patterns due to sampling emerge.

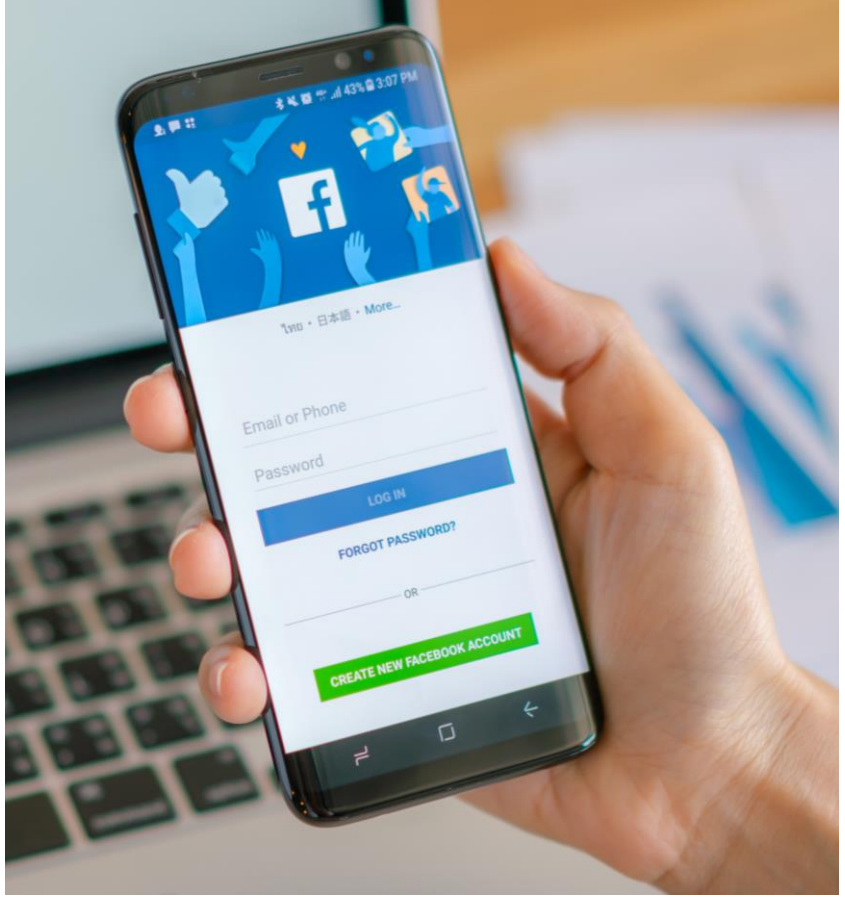


PREPROCESSING

COMMON PROCESSES

Data masking (big data)

Aims to ensure that any confidential information in the data remains private, without hindering the analysis and extraction of insight. Masking involves concealing the original data with random and false data, allowing the scientist to conduct their analyses without compromising private details. This can be done to traditional data too, but big data information can be much more sensitive.



WHERE DOES DATA COME FROM

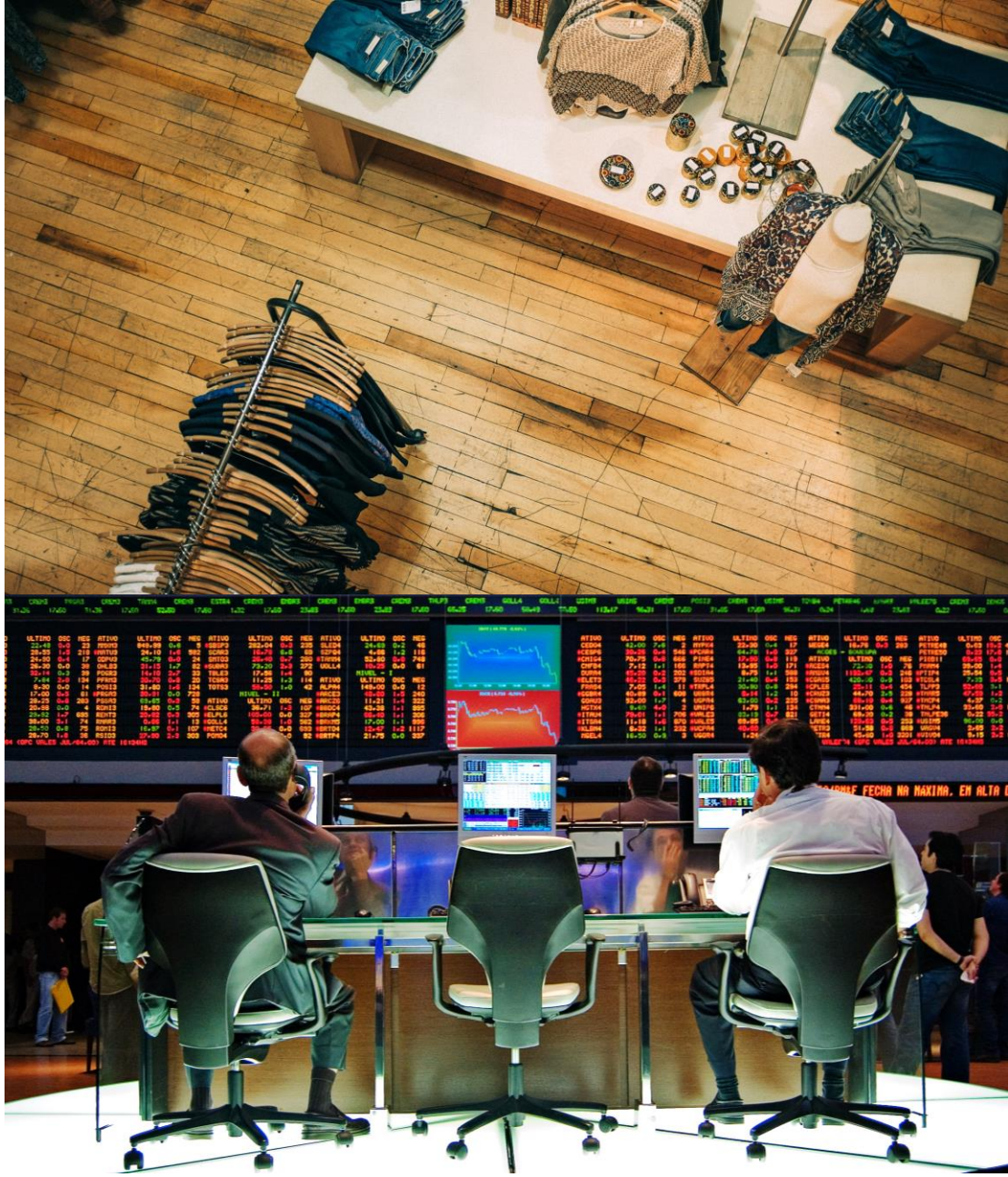
WHERE DOES DATA COME FROM?

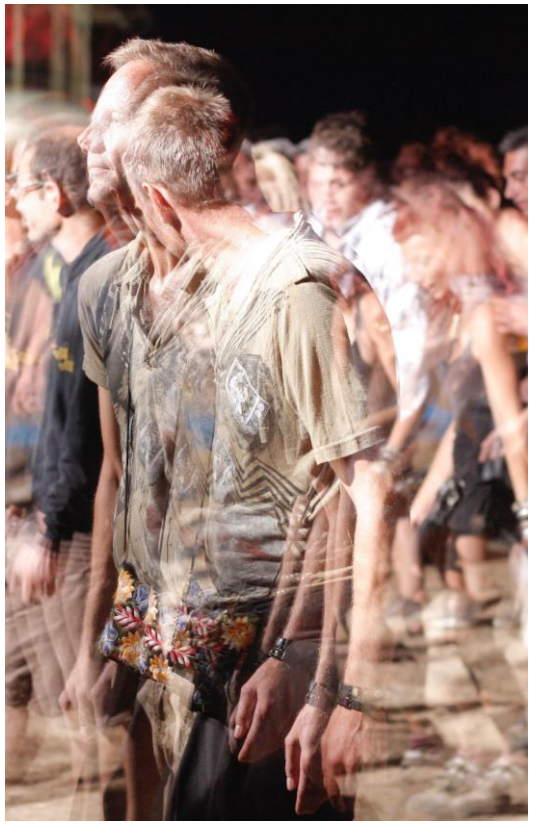
TRADITIONAL DATA

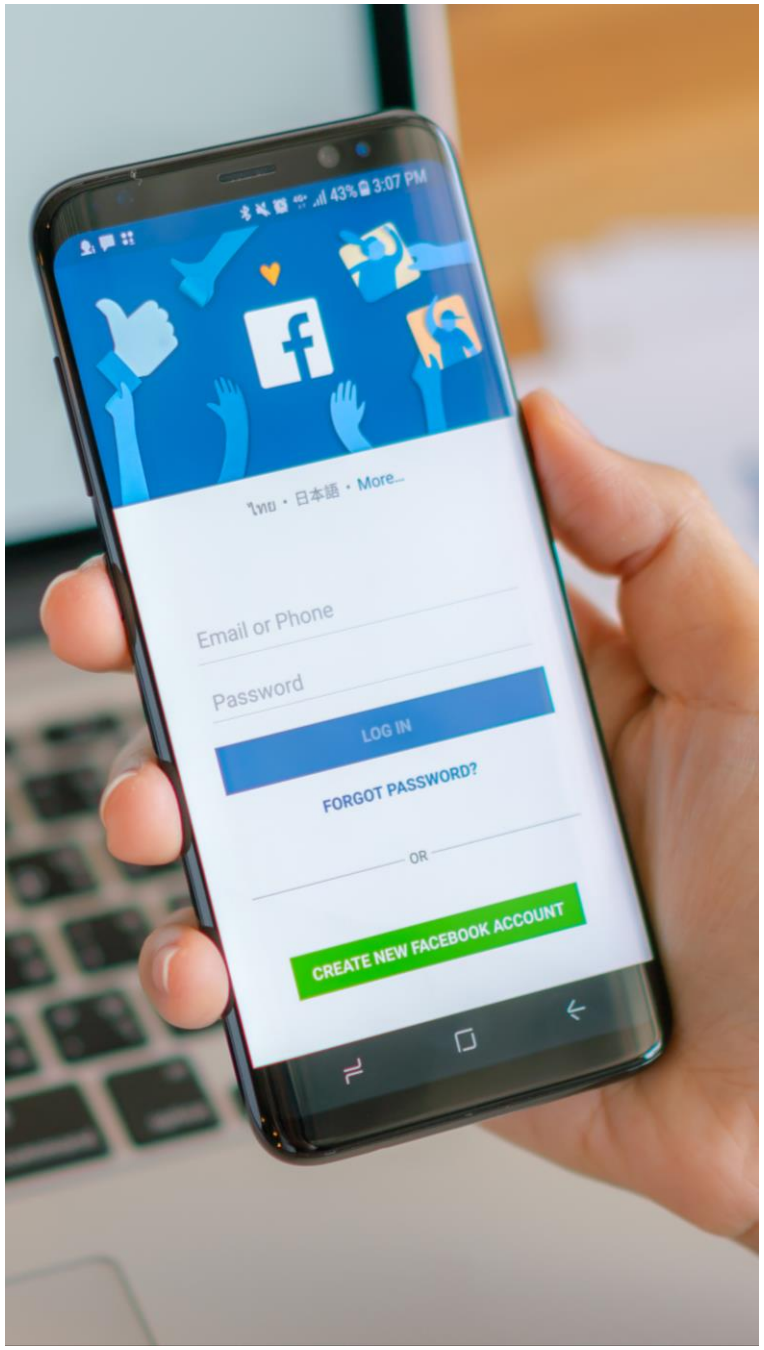
Traditional data may come from basic customer records, or historical stock price information.

Basic customer records can contain information like the customer ID, how many times they have filed a complaint, the amount of money spent in a single purchase, if they are part of a members scheme, their address, contact information and so on.

Historical stock price data, on the other hand, can contain dates of the observations, the names of the stocks, and their prices.





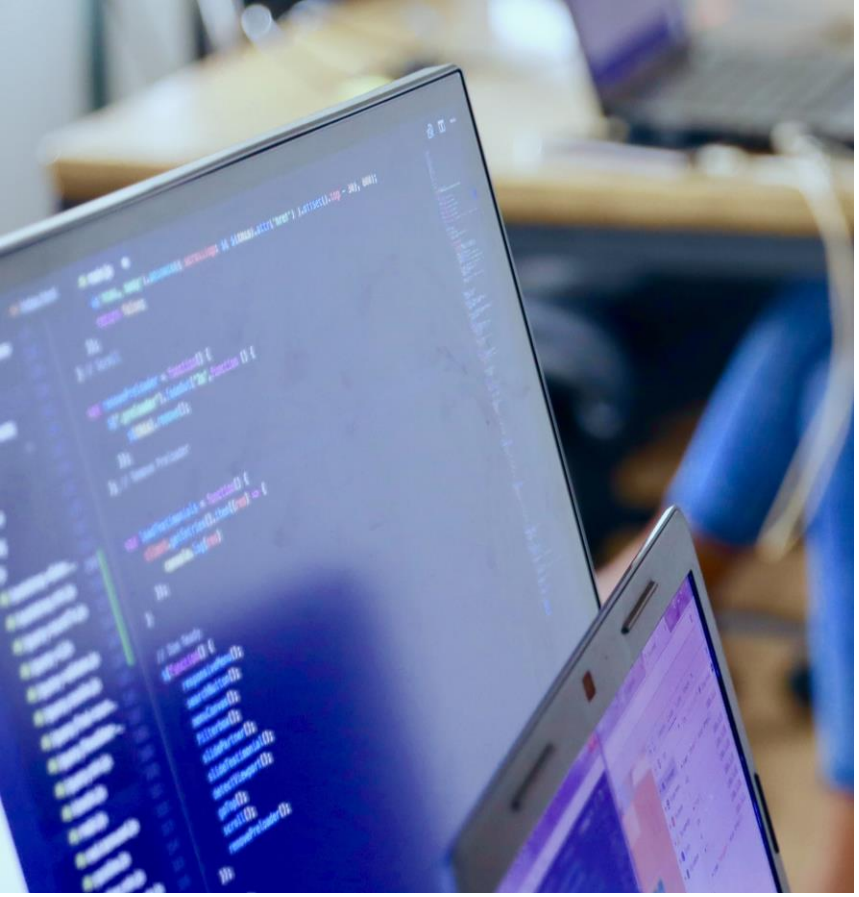
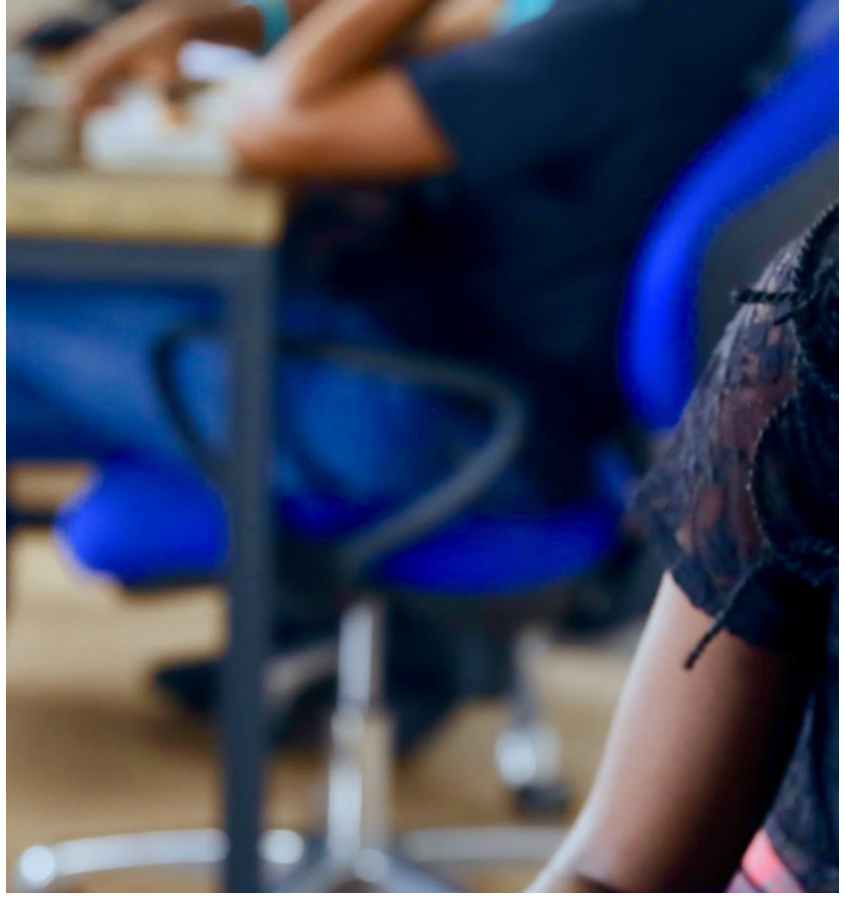


WHERE DOES DATA COME FROM?

BIG DATA

A consistently growing number of companies and industries use and generate big data. Consider online communities, for example, Facebook, Google, and LinkedIn; or financial trading data. Machine data from sensors in industrial equipment also amounts to big data. And, of course, wearable tech.

Currently, the volume of digital data amounts to 3.2 zettabytes, and 90% of this data has been gathered in the last 2 years only.


WHO HANDLES THE DATA

WHO HANDLES DATA

DATA ARCHITECTS AND ENGINEERS

The data specialists who deal with raw data and preprocessing, with creating databases, and maintaining them can go by a different name. But although their titles are similar sounding, there are palpable differences in the roles they occupy. Consider the following.

Data Architects and **Data Engineers** (and Big Data Architects, and Big Data Engineers, respectively) are crucial in the data science market. The former creates the database from scratch; they design the way data will be retrieved, processed, and consumed. Consequently, the data

WHO HANDLES DATA

DATABASE ADMINISTRATORS

engineer uses the data architects' work as a stepping stone and then processes (preprocesses) the available data. They are the people who ensure the data is clean and organized and ready for the analysts to take over.

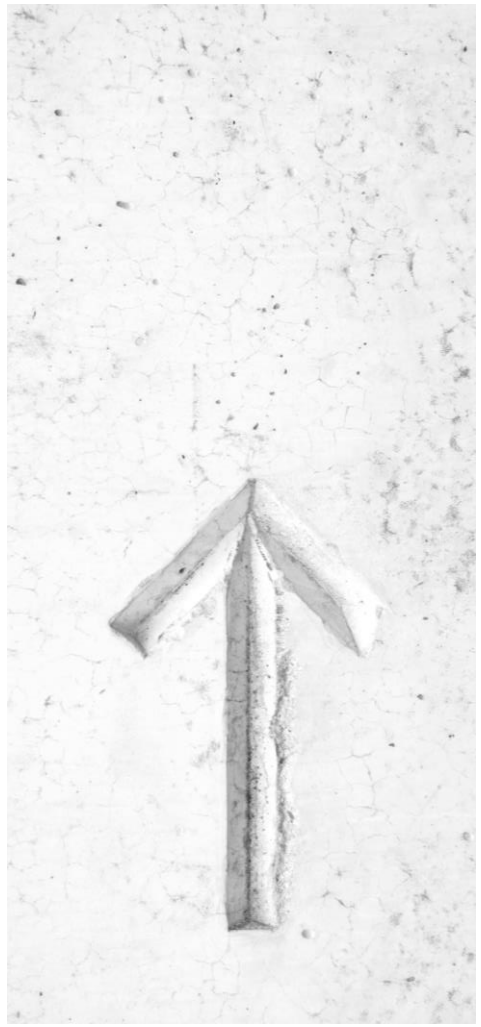
The **Database Administrator**, on the other hand, is the person who controls the flow of data into and from the database. Of course, with Big Data almost the entirety of this process is automated, so there is no real need for a human administrator. The Database Administrator deals mostly with traditional data.

DATA SCIENCE EXPLAINING THE PAST

BUSINESS INTELLIGENCE

There are also two ways of looking at data: with the intent to explain behavior that has already occurred, and you have gathered data for it; or to use the data you already have in order to predict future behavior that has not yet happened.

Before data science jumps into predictive analytics, it must look at the patterns of behavior the past provides, analyze them to draw insight and inform the path for forecasting. Business intelligence focuses precisely on this: providing data-driven answers to questions like…



BUSINESS INTELLIGENCE

- How many units were sold?
- In which region were the most goods sold?
- Which type of goods sold where?
- How did the email marketing per-form last quarter in terms of click-through rates & revenue gene-rated?
- How does that compare to the per-formance in the same quarter of last year?

Although Business Intelligence does not have “data science” in its title, it is part of data science, and not in any trivial sense.

365 DataScience



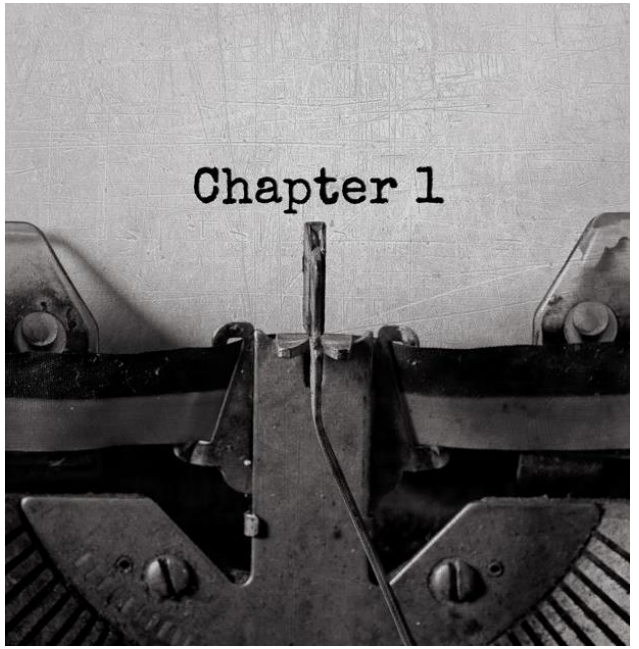
WHAT DOES A BI ANALYST DO

WHAT DOES A BI ANALYST DO

Of course, data science can be applied to **measure business performance**. But in order for the Business Intelligence Analyst to achieve that, they must employ specific data handling techniques.

The starting point of all data science is data. Once the relevant data is in the hands of the BI Analyst (monthly revenue, customer, sales volume, etc.), they must:

- quantify the observations
- calculate KPIs
- examine the measures to extract insights from their data.

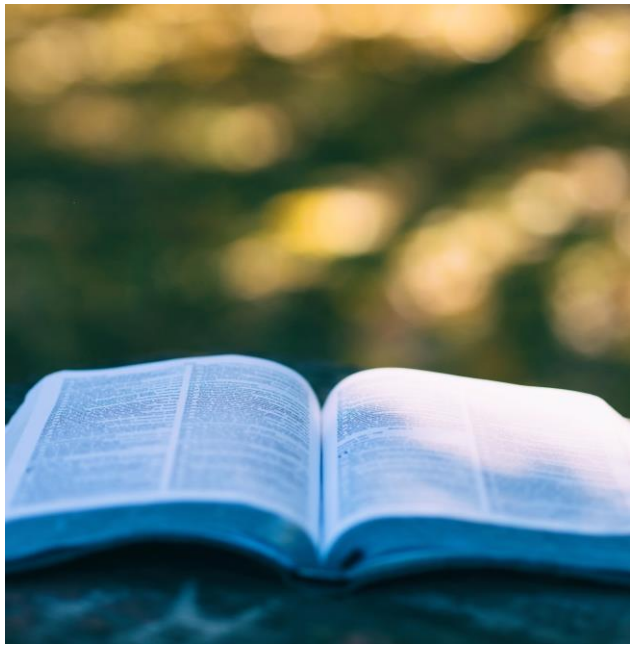


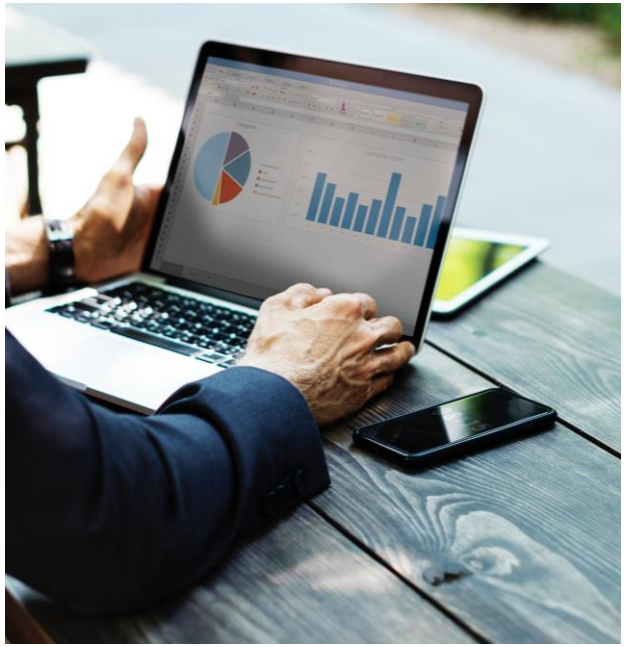
WHAT DOES A BI ANALYST DO

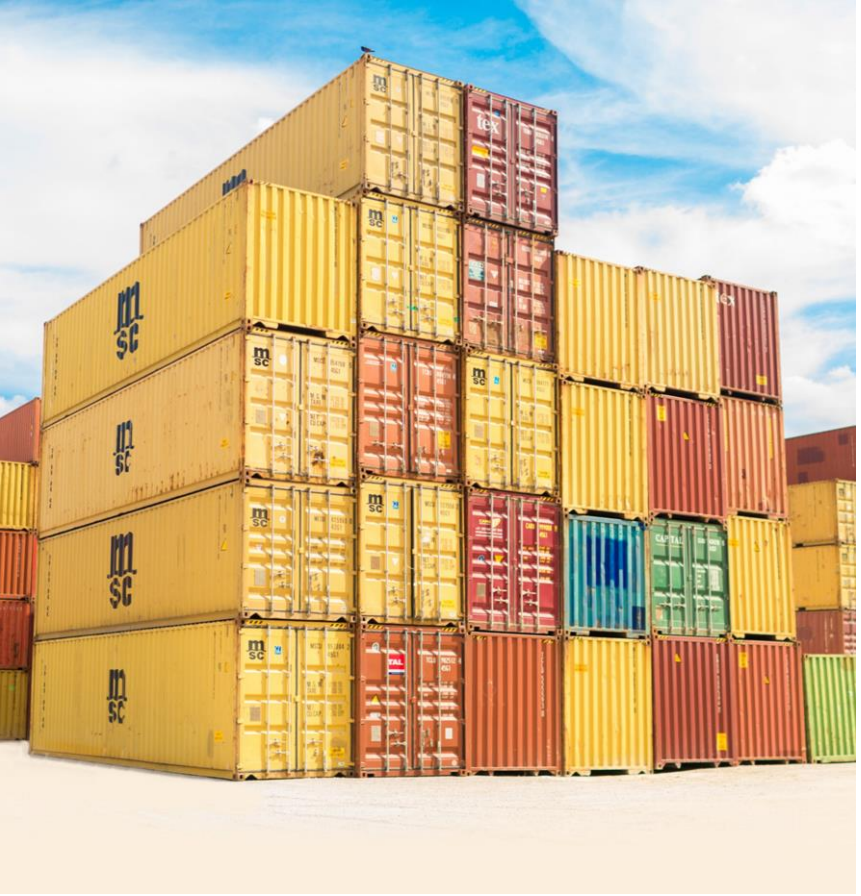
Data Science is about telling a story.

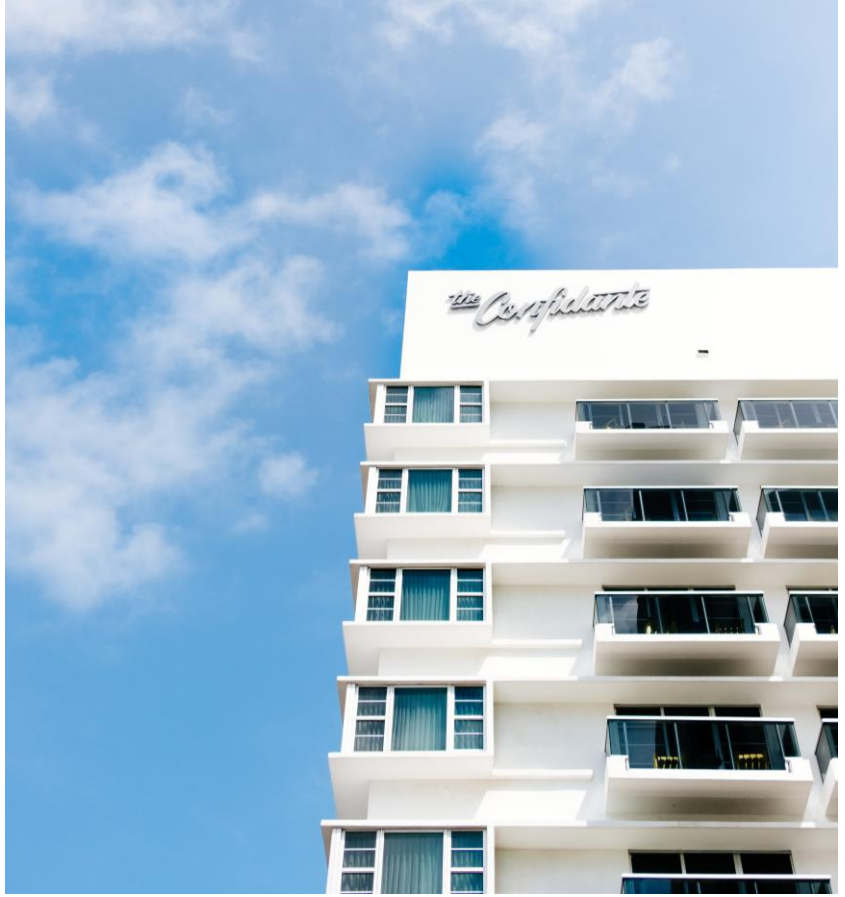
Apart from handling strictly numerical information, data science, and specifically BI, is about visualizing the findings, and creating easily digestible images supported only by the most relevant numbers. All levels of management should be able to understand the insights from the data and inform their decision-making.

BI analysts create dashboards & reports, accompanied by graphs, diagrams, maps, and other comparable visualizations, to present the findings relevant to the current business objectives.









WHERE IS BI USED

PRICE OPTIMISATION

Notably, data science is applied to inform things like price optimization techniques.

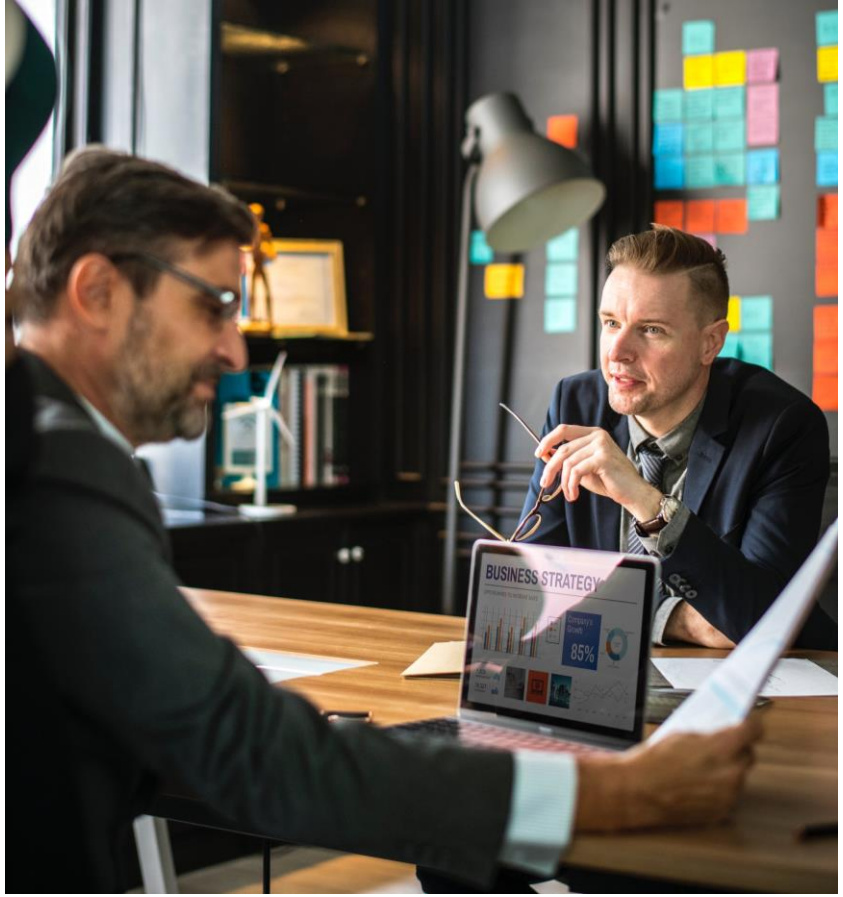
How does that work? With BI! The relevant information is extracted in real time, it is compared with historicals, and actions are taken accordingly.

Consider hotel management behavior: prices of rooms are raised in periods when many people want to visit the hotel and reduced when the goal is to attract visitors in periods with low demand.

INVENTORY MANAGEMENT

Data science, and business intelligence, are invaluable for handling over and undersupply.

How does it work? In-depth analyses of past sales transactions identify seasonality patterns and the times of the year with the highest sales, which results in the implementation of effective inventory management techniques that meet demands at minimum cost.



WHO DOES BUSINESS INTELLIGENCE

BUSINESS INTELLIGENCE ROLES

- Business Intelligence Analyst

A BI analyst focuses primarily on analyses and reporting of past historical data.

- Business Intelligence Developer

The BI developer is the person who handles more advanced programming tools, such as Python and SQL, to create analyses specifically designed for the company. It is the third most frequently encountered job position in the BI team.

BUSINESS INTELLIGENCE ROLES

- Business Intelligence Consultant

The BI consultant is often just an 'external BI analyst'. Many companies outsource their data science departments as they don't need or want to maintain one.
BI consultants would be BI analysts had they been employed, however, their job is more varied as they hop on and off different projects. The dynamic nature of their role provides the BI consultant with a different perspective, and whereas the BI Analyst has highly specialized knowledge ("depth"), the BI consultant contributes to the breadth of the data science team.

DATA SCIENCE PREDICTING THE FUTURE

DATA SCIENCE PREDICTING *the future*

Predictive analytics in data science rest on the shoulders of explanatory data analysis.

In fact, everything is connected. Once the BI reports and dashboards have been prepared and insights – extracted from them – this information becomes the basis for predicting future values. And the accuracy of these predictions lies in the methods used.

Recall the distinction between traditional data and big data in data science. A similar distinction can be made regarding predictive analytics and their methods: traditional data science methods vs. Machine Learning. One deals primarily with traditional data, and the other – with big data.

TRADITIONAL METHODS

Traditional forecasting methods in this booklet comprise the **classical statistical methods** for forecasting – linear regression analysis, logistic regression analysis, clustering, factor analysis, and time series. The output of each of these feeds into the more sophisticated machine learning analytics, but let's first review them individually.

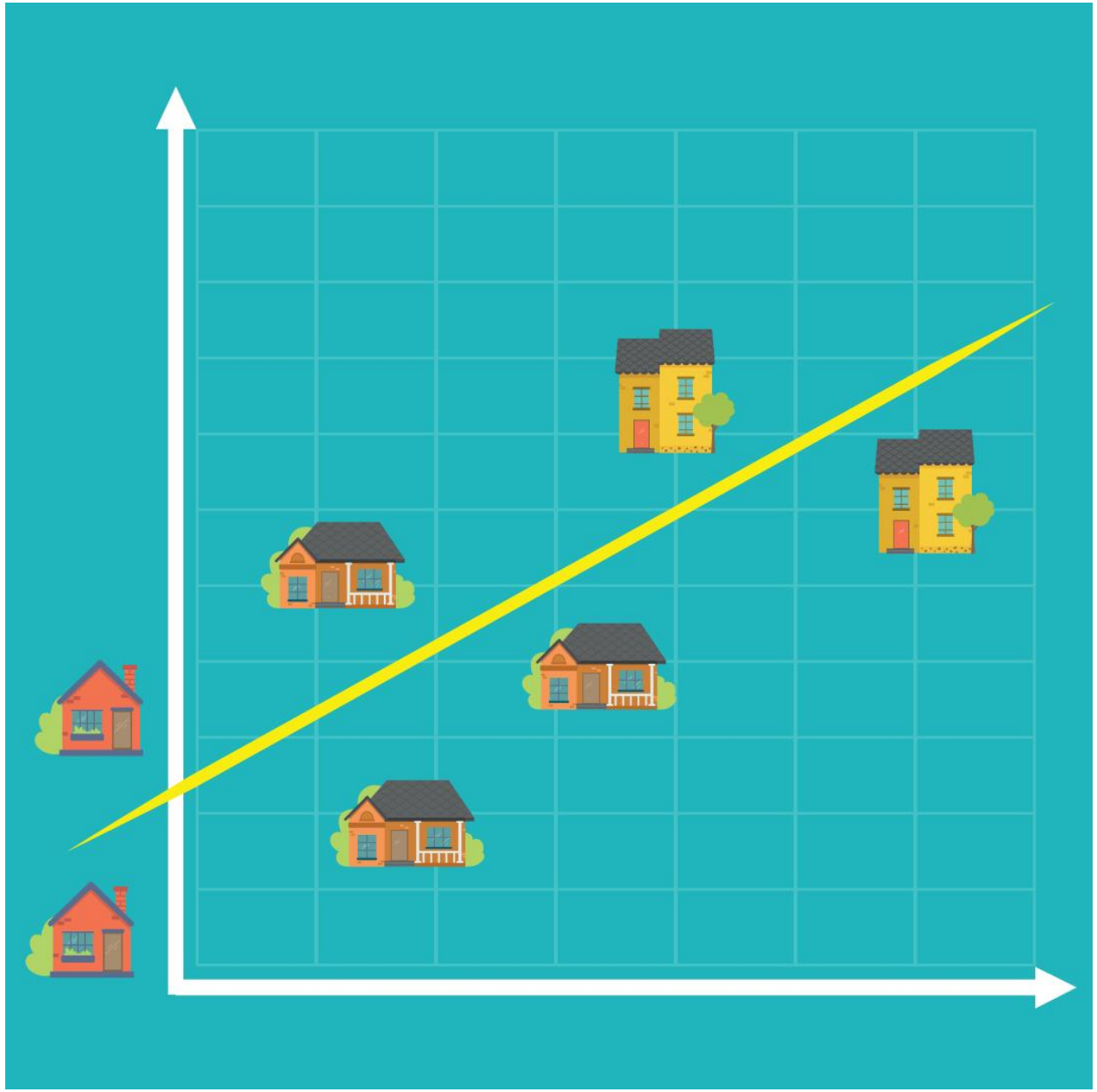
A quick side-note. Some in the data science industry refer to several of these methods as machine learning too, but in this article machine learning refers to newer, smarter, better methods, such as deep learning.

TRADITIONAL METHODS

LINEAR REGRESSION

In data science, the linear regression model is used for quantifying causal relationships among the different variables included in the analysis.

Like the relationship between house prices, the size of the house, the neighborhood, and the year built. The model calculates coefficients with which you can predict the price of a new house, if you have the relevant information available.

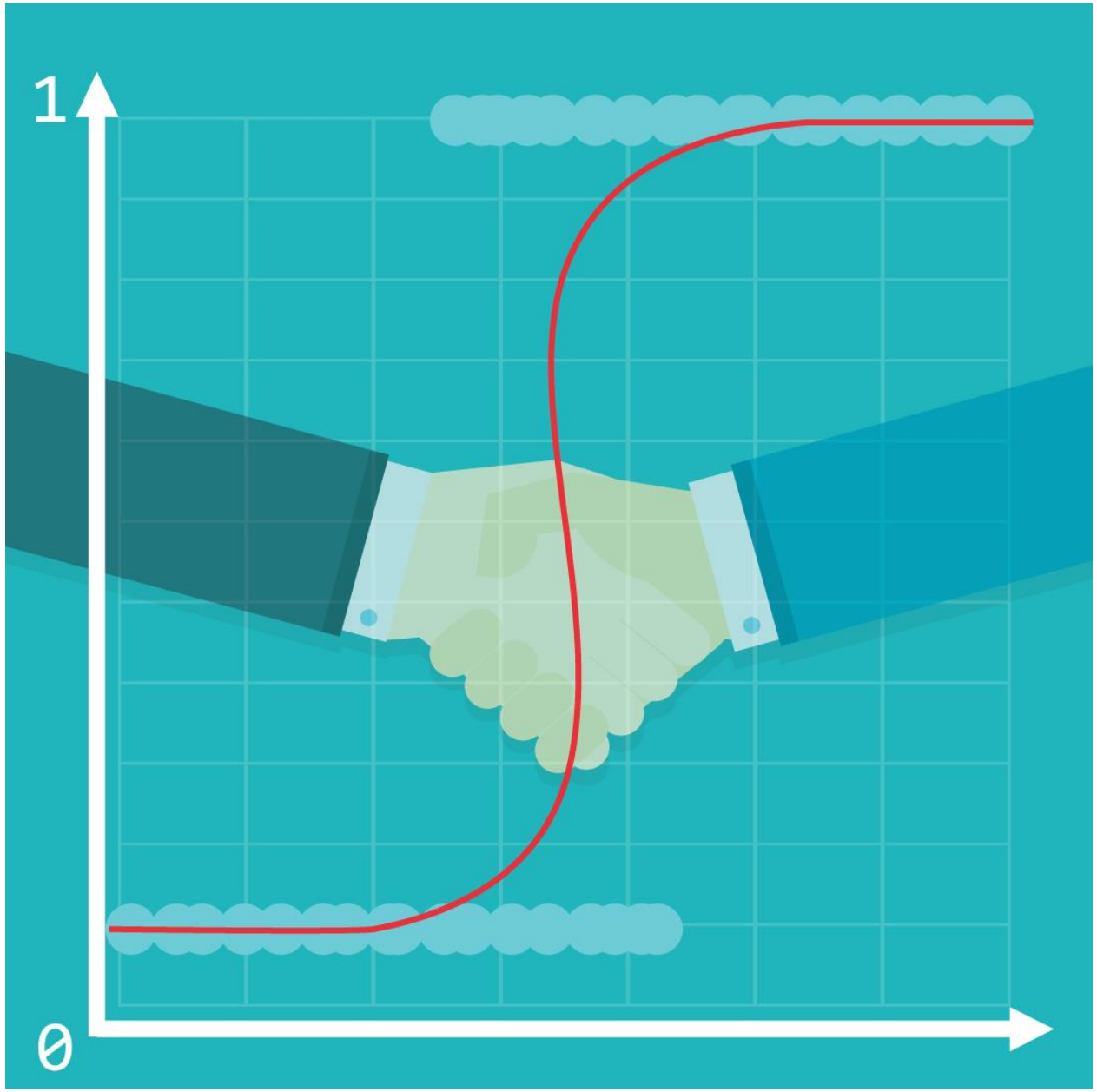


TRADITIONAL METHODS

LOGISTIC REGRESSION

Since not all relationships between variables can be expressed as linear, data science makes use of methods like the logistic regression to create non-linear models. Logistic regression operates with 0s and 1s.

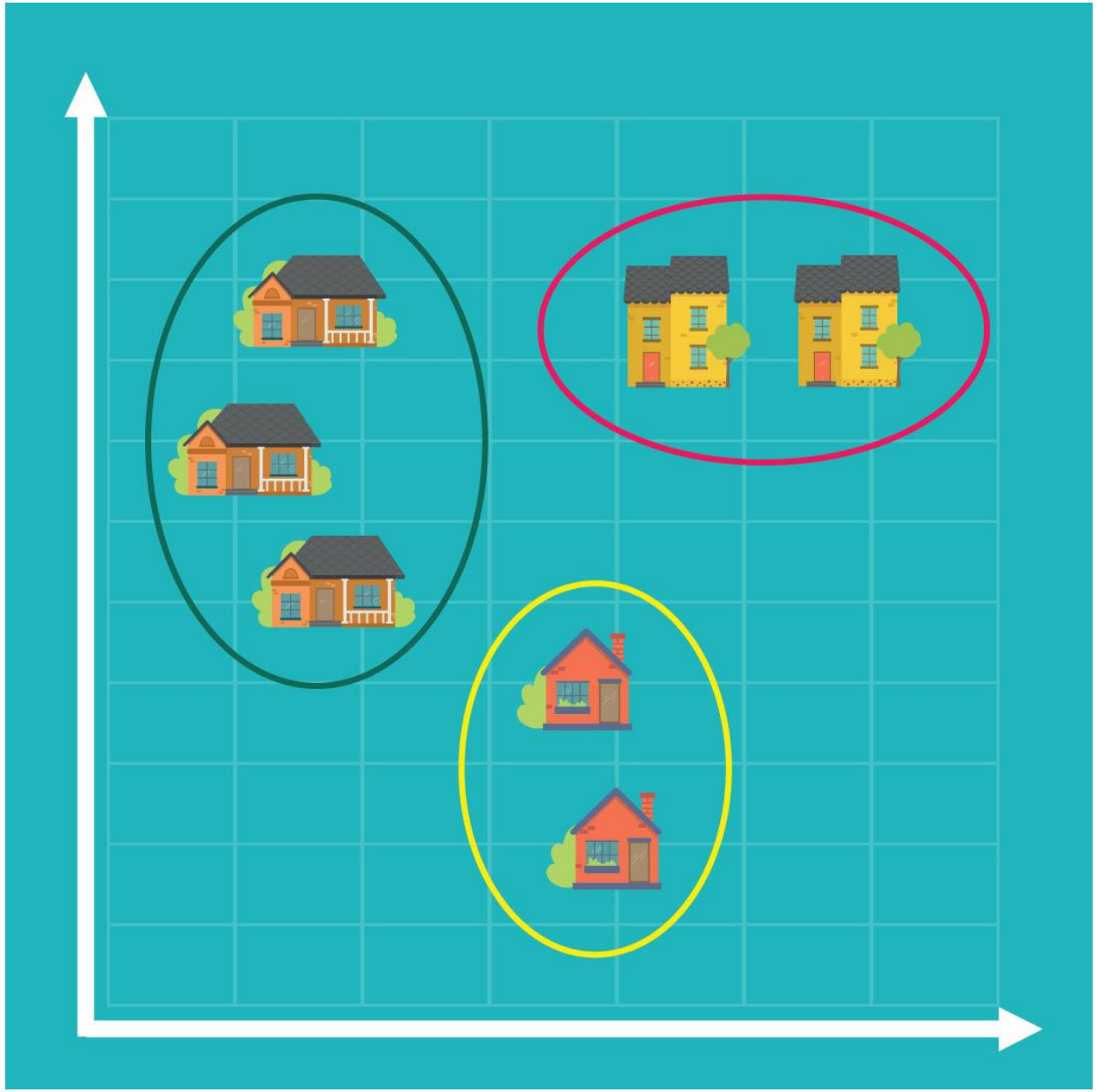
Companies apply logistic regression algorithms to filter job candidates during their screening process. If the algorithm estimates that the probability that a prospective candidate will perform well in the company within a year is above 50%, it would predict 1, or a successful application. Otherwise, it will predict 0.



TRADITIONAL METHODS

CLUSTER ANALYSIS

This exploratory data science technique is applied when the observations in the data form groups according to some criteria. Cluster analysis takes into account that some observations exhibit similarities, and facilitates the discovery of new significant predictors, ones that were not part of the original conceptualization of the data. For instance, looking at housing data, you could find the following clusters: small houses, with a high price (typical for houses in the city center); big houses that cost less (houses far away from the center); and big houses that cost a lot (houses in nice suburban neighbor-hoods).

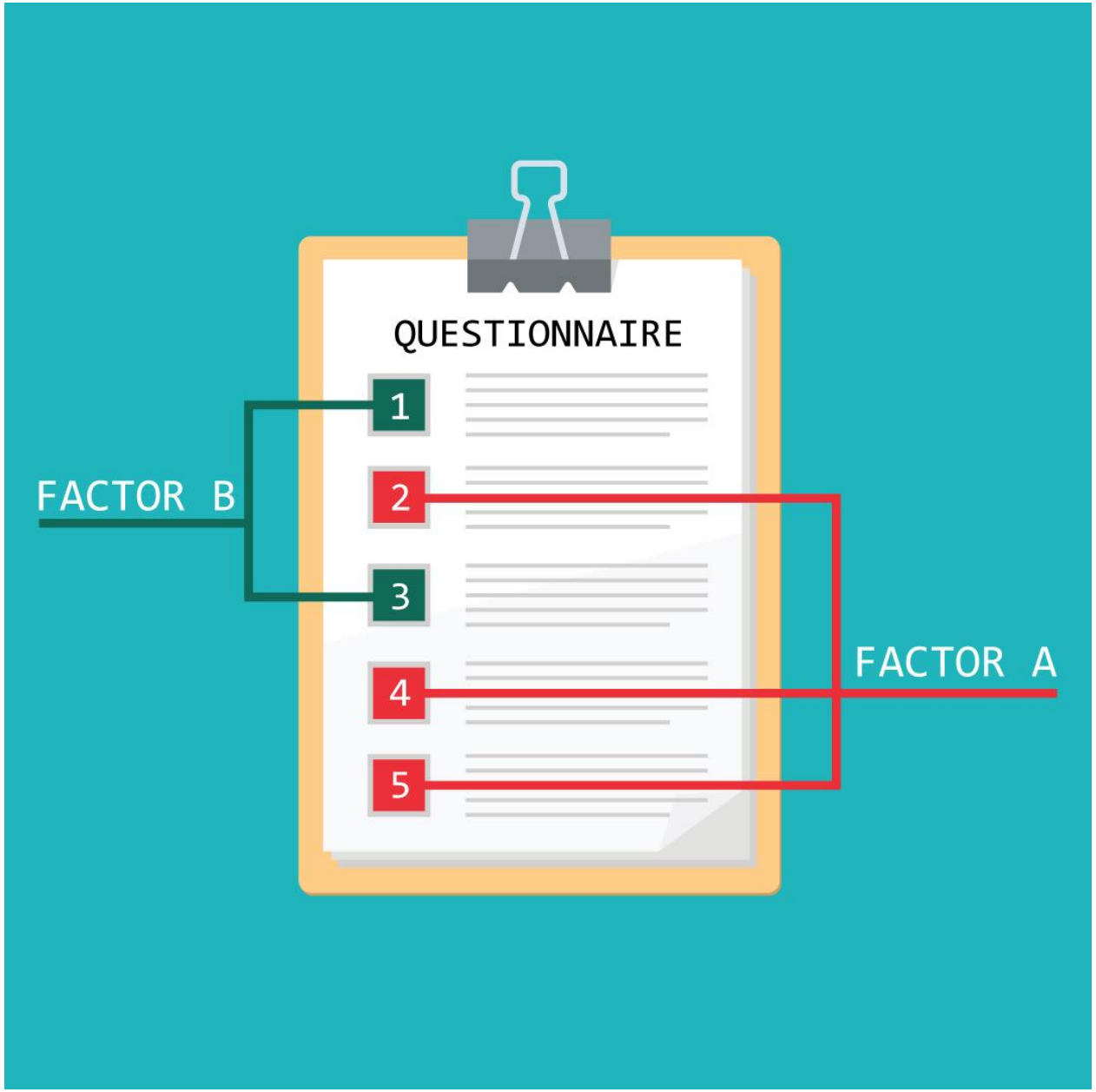


TRADITIONAL METHODS

FACTOR ANALYSIS

If clustering is about grouping observations together, factor analysis is about grouping features together. Data science resorts to using factor analysis to reduce the dimensionality of a problem.

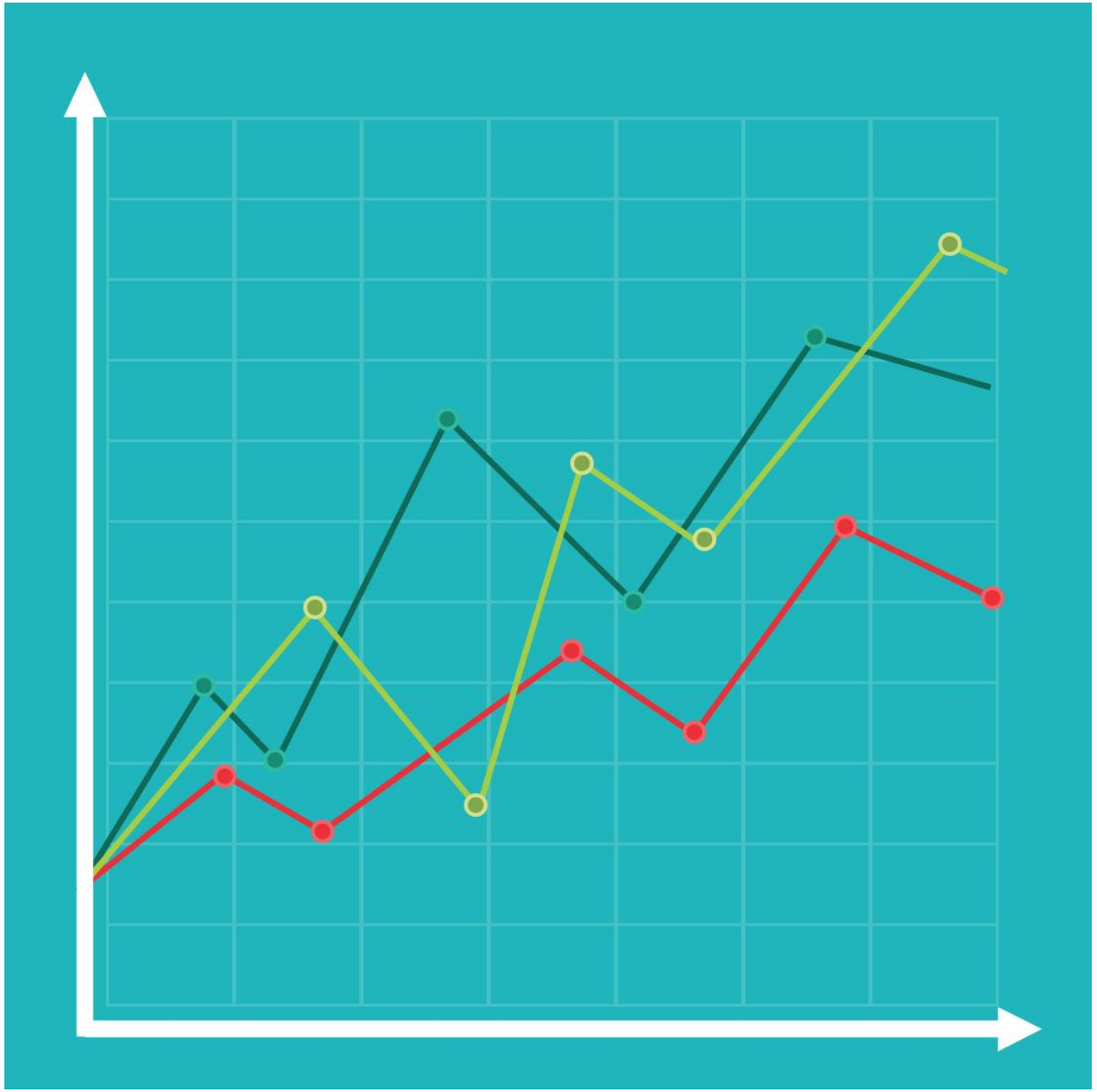
For example, if in a 100-item questionnaire each 10 questions pertain to a single general attitude, factor analysis will identify these 10 factors; they can then be used for a regression that will deliver a more interpretable prediction. Many techniques in data science are integrated like this.



TRADITIONAL METHODS

TIME SERIES ANALYSIS

Time series is a popular method for following the development of specific values over time. It is widely used in economics and finance because their subject matter is stock prices and sales volume – variables that are typically plotted against time. Time will always be on the horizontal line, as time is independent of any other variable. Therefore, such a graph can end up depicting a few lines that illustrate the behavior of your stocks over time. So, when you study the visualization, you can spot which stock performed well and which did not.



TRADITIONAL FORECASTING METHODS

REAL-LIFE APPLICATIONS

365 DataScience

TRADITIONAL FORECASTING USES

User experience:

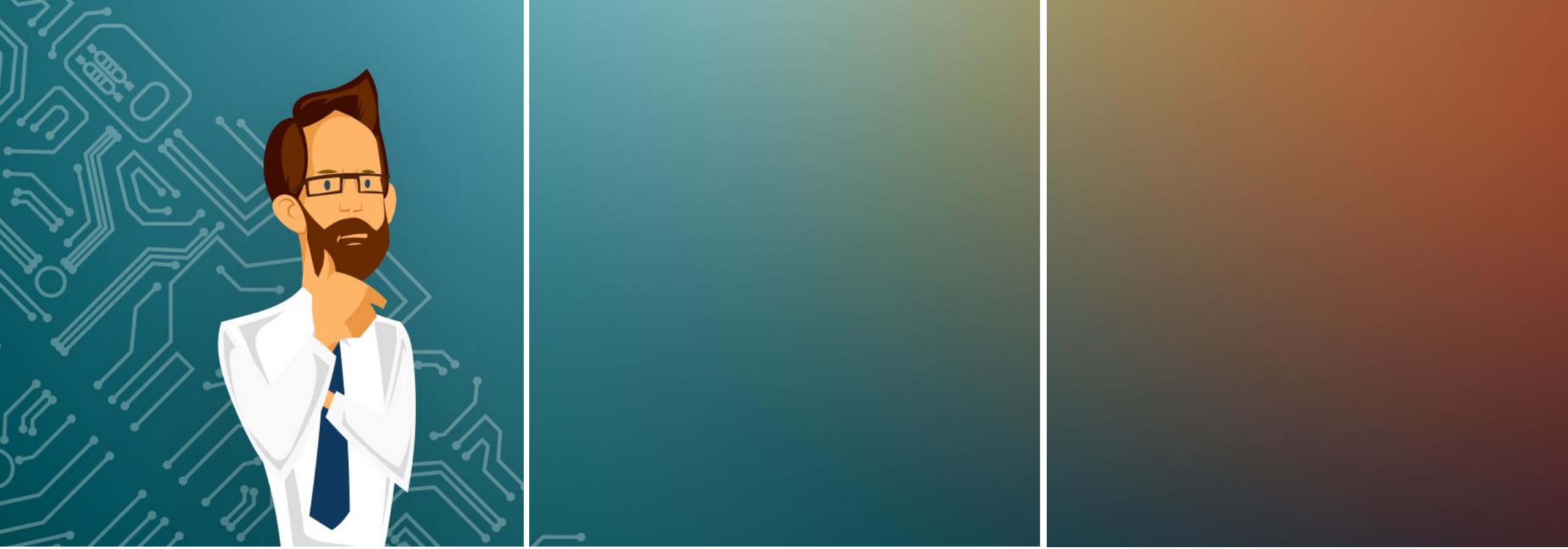
When companies launch a new product, they often design surveys that measure the attitudes of customers towards that product. Analyzing the results after the BI team has generated their dashboards includes grouping the observations by segments (e.g. regions), and then analyzing each segment separately to extract meaningful predictive coefficients. The results of these operations often corroborate the conclusion that the product needs slight but significantly different adjustments in each segment in order to maximize customer satisfaction.

365 DataScience

TRADITIONAL FORECASTING USES

Forecasting sales:

Forecasting sales volume is the type of analysis where time series comes into play. Sales data has been gathered until a certain date, and the data scientist wants to know what is likely to happen in the next sales period, or a year ahead. How does it work? They apply mathematical and statistical models and run multiple simulations; these simulations provide the analyst with future scenarios. This is at *the core* of data science, because, based on these scenarios, the company can make better predictions and implement adequate strategies.



WHO USES TRADITIONAL FORECASTING

WHO USES TRADITIONAL FORECASTING

The **data scientist**. But bear in mind that this title also applies to the person who employs machine learning techniques for analytics, too. A lot of the work spills from one methodology to the other.

The **data analyst**, on the other hand, is the person who prepares advanced types of analyses that explain the patterns in the data that have already emerged and overlooks the basic part of the predictive analytics.

MACHINE LEARNING

Machine learning is the state-of-the-art approach to data science. And rightly so.

The main advantage machine learning has over any of the traditional data science techniques is the fact that at its core resides **the algorithm**. These are the directions a computer uses to find a model that fits the data as well as possible.

The difference between machine learning and traditional data science methods is that we do not give the computer instructions on how to find the sought dependence; it takes the algorithm and uses its directions to learn on its own how to find that model.

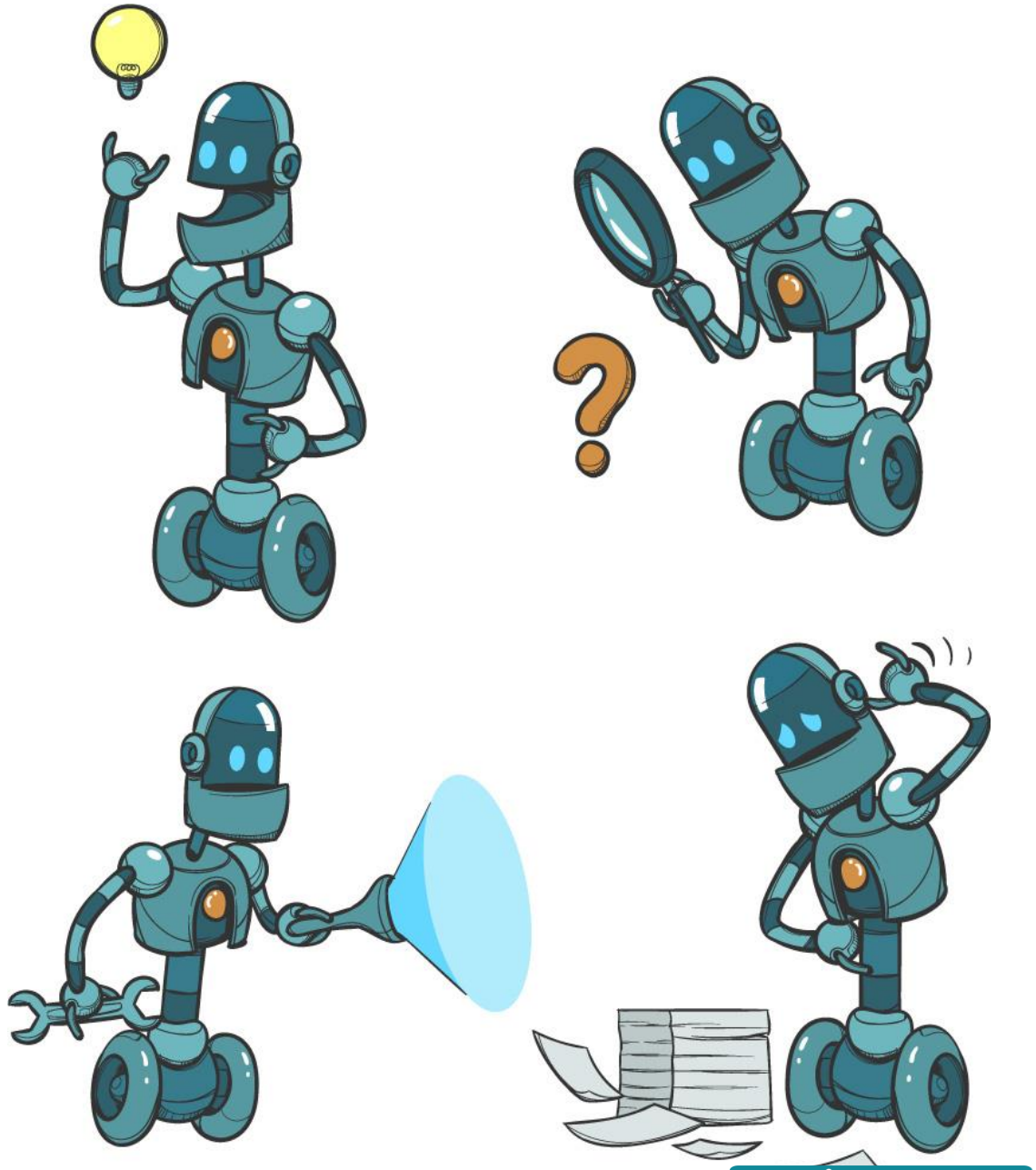
Unlike in traditional data science, human involvement is minimized. In fact, machine learning, especially deep learning algorithms are so complicated, that humans cannot genuinely understand what is happening “inside” the model.

MACHINE LEARNING
THE ALGORITHM

A machine learning algorithm is like a trial-and-error process, but the special thing about it is that each consecutive trial is at least as good as the previous one. But bear in mind that in order to learn well, the machine has to go through hundreds of thousands of trial-and-errors, with the frequency of errors decreasing throughout.

Once the training is complete, the machine will be able to apply the complex computational model it has learned to novel data still to the result of highly reliable predictions.

There are three major types of machine learning: supervised, unsupervised, and reinforcement learning.



365 DataScience

SUPERVISED LEARNING

Supervised learning rests on using labeled data. Imagine having big data consisting of video files and images, labelled as “cats”, “dogs”, & “other”.

The machine gets data that is associated with a correct answer; if the machine’s performance does not get that correct answer, an optimization algorithm adjusts the computational process, and the computer does another trial. Typically, the machine does this on hundreds of data points at once.

Support vector machines, neural networks, deep learning, random forest models, and Bayesian networks are all instances of supervised learning.

UNSUPERVISED LEARNING

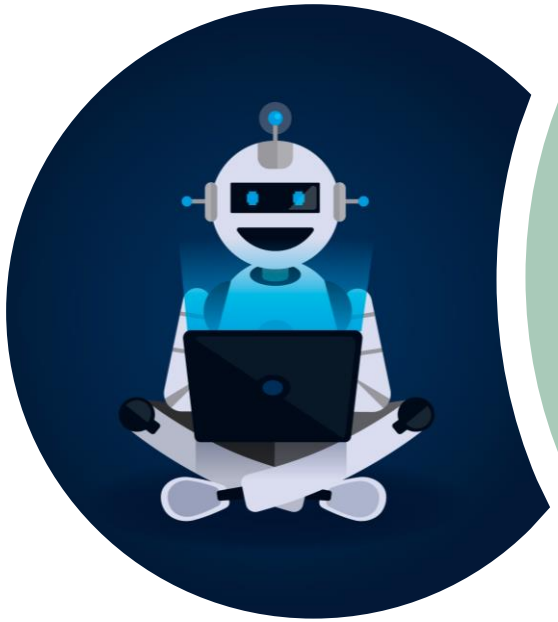
When the data is too big, or the data scientist is pressured for resources to label the data, or they do not know what the labels are at all, data science resorts to using unsupervised learning. This consists of giving the machine unlabeled data and asking it to extract insights from it. This often results in the data being divided in a certain way according to its properties. In other words, it is clustered.

Unsupervised learning is extremely effective for discovering patterns in data, especially things that humans using traditional analysis techniques would miss.

REINFORCEMENT LEARNING

This is a type of machine learning where the focus is on performance (to walk, to see, to read), instead of accuracy. Whenever the machine performs better than it has before, it receives a reward, but if it performs suboptimally, the optimization algorithms does not adjust the computation.

Think of a puppy learning commands. If it follows the command, it gets a treat; if it doesn't follow the command, the treat doesn't come. Because treats are tasty, the dog will gradually improve in following commands. That said, instead of minimizing an error, reinforcement learning maximizes a reward.



MACHINE LEARNING

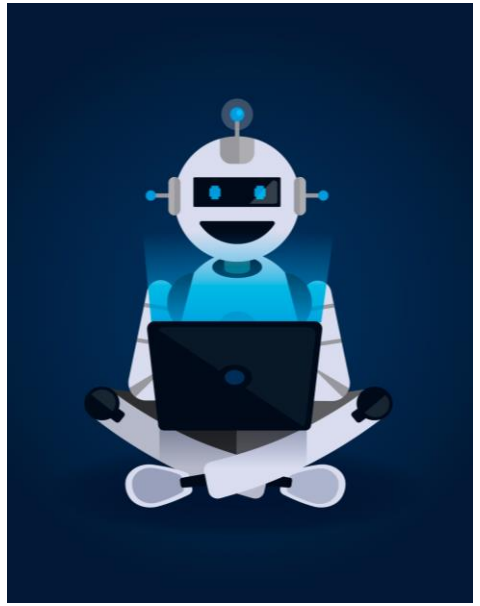
REAL-LIFE APPLICATIONS

MACHINE LEARNING USES

Fraud detection:

With machine learning and supervised learning in particular, banks can take past data, label the transactions as legitimate, or fraudulent, and train models to detect fraudulent activity. When these models detect even the slightest probability of theft, they flag the transactions, and prevent the fraud in real time.

MACHINE LEARNING USES



Client retention:

With machine learning algorithms, corporate organizations can know which customers may purchase goods from them. This means the store can offer discounts and a 'personal touch' in an efficient way, minimizing marketing costs and maximizing profits. A couple of prominent names come to mind: Google, and Amazon.

WHO USES MACHINE LEARNING

WHO USES MACHINE LEARNING

As we mentioned already, the data scientist is deeply involved in designing machine learning algorithms, but there is another star on this stage.

The **machine learning engineer**. This is the specialist who is looking for ways to apply state-of-the-art computational models developed in the field of machine learning into solving complex problems such as business tasks, data science tasks, computer vision, self-driving cars, robotics, and so on.

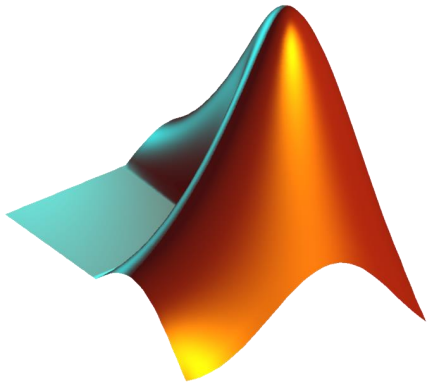
PROGRAMMING LANGUAGES & SOFTWARE

PROGRAMMING LANGUAGES

Knowing a programming language enables the data scientist to devise programs that can execute specific operations. The biggest advantage programming languages have is *flexibility*.

R, Python, and MATLAB, combined with SQL, cover most of the tools used when working with traditional data, BI, and conventional data science.

R and Python are the two most popular tools across all data science sub-disciplines. Their biggest advantage is that they can manipulate data and are integrated within multiple data and data science software platforms. They are not just suitable for mathematical and statistical computations; they are adaptable.



PROGRAMMING LANGUAGES

SQL is king, however, when it comes to working with relational database management systems, because it was specifically created for that purpose. SQL is at its most advantageous when working with traditional, historical data, for example when doing a BI analysis.

MATLAB is the fourth most indispensable tool for data science. It is ideal for working with mathematical functions or matrix manipulations.

PROGRAMMING LANGUAGES

Big data in data science is handled with the help of R and Python, of course, but people working in this area are often proficient in other languages like Java or Scala. These two are very useful when combining data from multiple sources.

JavaScript, C, and C++, in addition to the ones already mentioned, are often employed when the branch of data science the specialist is working in involves machine learning. They are faster that R and Python, and provide greater freedom.

SOFTWARE AND FRAMEWORKS

In data science, the software or, software solutions, are tools adjusted for specific business needs.

Excel is a tool applicable to more than one category – traditional data, BI, and Data Science. Similarly, SPSS is a very famous tool for working with traditional data and applying statistical analysis.

TensorFlow, on the other hand, is a software library and framework designed for working with big data and designing Machine Learning algorithms. It was developed by Google for internal use, became public in 2015, and is generally the leader for working with and deploying neural networks.

SOFTWARE AND FRAMEWORKS

Apache Hadoop, Apache Hbase, and Mongo DB, on the other hand, are software designed for working with big data.

Power BI, SaS, Qlik, and especially Tableau are top-notch examples of software designed for business intelligence visualizations.

In terms of predictive analytics, EViews is mostly used for working with econometric time-series models, and Stata – for academic statistical and econometric research, where techniques like regression, cluster, and factor analysis are constantly applied.

INTERVIEW FAQ

FAQ AT INTERVIEWS

1. What does data science mean?
2. What are the assumptions of a linear regression?
3. What is the difference between factor analysis and cluster analysis?
4. What is an iterator generator?
5. Write down an SQL script to return data from two tables.
6. Draw graphs relevant to pay-per-click adverts and ticket purchases.
7. How would you explain Random Forest to a non-technical person?
8. How can you prove an improvement you introduced to a model is actually working?
9. What is root cause analysis?

FAQ AT INTERVIEWS

10. Explain K-means.
11. What kind of RDBMS software do you have experience with? What about non-relational databases?
12. Supervised learning vs unsupervised learning.
13. What is overfitting and how to fix it?
14. What is the difference between SQL, MySQL and SQL Server?
15. How would you start cleaning a big dataset?
16. Give examples where a false negative is more important than a false positive, and vice versa.
17. State some biases that you are likely to encounter when cleaning a database.
18. What is a logistic regression?

365 DataScience

Good luck!