Intro to Data and Data Science

1. The different data science fields

1.1 The different data science fields

There is no denying that in today's day and age data is at the foundation of any successful company. Leading entrepreneurs are aware that looking deeper into data is what will make them tower above the competition.

Someone who qualified as a statistician 25 years ago and kept up with modern technologies could fit into a multitude of professional categories today.

Make sure you check out our infographic, which puts together an aggregated, concise and to the point structure when it comes to receiving an introduction to the world of data and data science.

The 365 Data Science All in One Infographic

1.2 Analysis vs analytics

Analysis and analytics are not two interchangeable terms. The reason for one often being used instead of the other is the lack of a transparent understanding of both.

Analysis – Dividing data into digestible components that are easier to understand and examining how different parts relate to each other. Performed on past data, explaining why the story ended in the way that it did. We want to explain ***'how'*** and ***'why'*** something happened.

Analytics – Explores the future. The application of logical and computational reasoning to the component parts obtained in an analysis. In doing this, you are looking for patterns and exploring what you can do with them in the future.

We can have:

- Qualitative analytics – using intuition and experience in conjunction with analysis to plan your next business move
- Quantitative analytics – applying formulas and algorithms to numbers that you have gathered from your analysis.

1.3 Intro to Business Analytics, Data Analytics, and Data Science

Some business activities are data-driven, while others are subjective or experience-driven.

Something, which is confusing in practice is that analytics has become a term comprising both 'analysis' and 'analytics'.

Data science is a discipline reliant on the availability of data, while business analytics does not completely rely on data. However, data science incorporates part of data analytics. Mostly the part that uses complex mathematical, statistical, and programming tools.

1.4 Adding Business Intelligence (BI), Machine Learning, and Artificial Intelligence (AI) to the picture

Business Intelligence (BI) is the process of analysing and reporting historical business data. After reports and dashboards have been prepared, they can be used to make informed strategic and business decisions by end-users such as the general manager. Concisely put, business intelligence aims to explain past events using business data.

Business Intelligence can be seen as the preliminary step of predictive analytics. First, you analyse past data and then using these inferences would allow you to create appropriate models that could predict the future of your business accurately.

Machine learning is the ability of machines to predict outcomes without being explicitly programmed to do so. It is about creating and implementing algorithms that let machines receive data and use this data to:

- Make predictions
- Analyse patterns
- Give recommendations

AI simulates human knowledge and decision making with computers. Humans have managed to reach **AI** through machine and deep learning.

Symbolic reasoning is a type of AI that makes an exception and does not use ML and deep learning. It is based on high-level human-readable representations of problems and logic. Very rarely used in practice.

Advanced analytics stands for all types of analytics processes.

1.5 An overview of the 365 Data Science infographic

From a data scientist's perspective, the solution to every task comes with having a proper dataset. This is the first thing on your to-do list. The information in the 365 Data Science infographic is split into 5 columns each detailing different stages of the process of solving a business task:

1) Working with traditional data
2) Working with big data
3) Doing business intelligence
4) Applying traditional data science techniques
5) Using ML techniques

The rows of the infographic answer several very important questions:

- When is this part of the process applied?
- Why do we need it?

- What are the techniques?
- Where and in which real-life cases can it be applied?
- How is it implemented? Using what tools?
- Who is doing this?

2. The relationship between different data science fields

Data can be defined as information stored in a digital format, which can then be used as a base for performing analysis and decision making. We can distinguish between two types of data:

- **Traditional data:** Data in the form of tables containing numeric or text values; Data that is structured and stored in databases
- **Big data:** Extremely large data; Humongous in terms of volume. It can be in various formats:

- structured

- semi-structured

- unstructured

Big data is often characterized with the letter 'V'. Under different frameworks we may have 3,5,7, and even 11 Vs of big data; The main ones are volume, variety, velocity.

Data science is a broad subject. It's an interdisciplinary field that combines statistical, mathematical, programming, problem-solving, and data-management tools.

The 365 Data Science infographic divides data science in 3 segments: business intelligence (analyse the past that you acquired), traditional methods and machine learning (forecast future performance).

Business Intelligence includes all technology-driven tools involved in the process of analysing, understanding, and reporting available past data. It allows you to make decisions, extract insights, and extract ideas.

Traditional methods a set of methods that are derived mainly from statistics and are adapted for business.

Machine learning is all about creating algorithms that let machines receive data, perform calculations and apply statistical analysis to make predictions with unprecedented accuracy.

3. What is the purpose of each data science field

Data-driven decisions require well-organized and relevant new data stored in a digital format.

Data is the foundation. It is the material on which you base your analysis. Without data, a decision maker would not be able to test their decisions and ensure they have taken the right course of action.

While the goal of 'traditional methods' and 'machine learning' are essentially the same, and techniques can overlap, there is a difference between the two. Traditional methods relate to traditional data. They were designed prior to the existence of big data, where the technology simply wasn't as advanced as it is today. They involve applying statistical approaches to create predictive models.

4. Common data science techniques

4.1 Traditional data: Techniques

The term data can refer to 'raw facts', 'processed data', or 'information'.

Raw data, also called 'primary data' is data which cannot be analysed straight away. It is untouched data you have accumulated and stored on the server. The gathering of raw data is referred to as **data collection.**

Data can be collected in a number of ways. One example would be the use of surveys, asking people to rate how much they like or dislike a product or experience on the scale of 1-10. Alternatively, gathering data could be automatic (for example cookies).

Data preprocessing needs to be performed on raw data to obtain meaningful information. This is a group of operations that will basically convert your raw data into a format that is more understandable

Class labelling: Labelling the data point to the correct data type (or arranging data by category).

Data cleansing: ('data cleaning', 'data scrubbing'): deal with inconsistent data. For example, working on a dataset containing US states and finding that some of the names are misspelled.

Data balancing: Ensuring that the sample gives equal priority to each class. For example, if you work with a dataset that contains 80% male and 20% female data, and you know that the population contains approximately 50% men and 50% women, then you need apply a balancing technique to counteract this problem (using an equal number of data from each group).

Data shuffling: Shuffling the observations from your dataset just like shuffling a deck of cards. This will ensure your dataset is free from unwanted patterns caused by problematic data collection.

4.2 Traditional data: Real-life examples

Numerical variable: Numbers that are easily manipulated (for ex. Added), which gives us useful information

Categorical variable: Numbers that hold no numerical value can be considered categorical data. Dates are also considered categorical data.

4.3 Big data: Techniques

Examples of big data: text data, digital image data, digital video data, digital audio data, etc.

With a wide variety of data types comes a wider range of data cleansing methods.

Text data mining: The process of deriving valuable, unstructured data from text.

Data masking: As a business, when you work with user private data, you must be able to preserve confidential information. However, this doesn't mean that the data can't be touched or used for analysis. Instead you must apply some data masking techniques to utilise the information without compromising private details. In essence, data masking conceals the original data with random and false data, allowing you to conduct analysis and keep confidential information in a secure place.

4.4 Big data: Real-life examples

We find big data in increasingly more industries and companies.

Probably the most notable example of a company leveraging the true potential of big data is Facebook. The company keeps track of its users' names, personal data, photos, videos, recorded messages and so on. This means their data has a lot of variety. And with 2 billion users worldwide, the volume of data stored on their servers is tremendous.

4.5 Business Intelligence: Techniques

Business intelligence requires the combination of data skills and business knowledge in an effort to explain the past performance of your company. It answers the questions "What happened?", "When did it happen?", "How many units did we sell?", "In which region did we sell the most goods?" etc.

The job of a business intelligence analyst requires her to understand the essence of a business and strengthen that business through the power of data.

Metric: refers to a value that derives from the measures you have obtained and aims at gauging business performance or progress. Has a business meaning attached to it.

Measure: simple descriptive statistics of past performance

Metric = Measure + Business meaning

KPIs: It doesn't make sense to keep track of all metrics. So, companies choose to focus on the most important ones.

KPIs = metrics + Business objective

Filtering out the boring metrics and turning the interesting and informative KPIs into easily understood and comparable visualizations is an important part of the business intelligence analyst job.

4.6 Business Intelligence: Real-life examples

BI allows you to adjust your strategy to past data as soon as it is available. If done right, Business Intelligence will help to efficiently manage your shipment logistics and, in turn, reduce costs and increase profit.

4.7 Traditional methods: Techniques

There are two branches of predictive analytics – traditional methods (classical statistical methods for forecasting) and machine learning.

In business and statistics, a regression is a model used for quantifying causal relationships among the different variables includes in your analysis.

A logistic regression is a common example of a non-linear model. The values on the vertical line will be 1s and 0s only.

Clustering: grouping the data in neighbourhoods to analyse meaningful patterns

Time series: used in economics and finance, showing the development of certain values over time, such as stock prices or sales volume.

4.8 Traditional methods: Real-life examples

The application of traditional methods is extremely broad. Two relevant examples:

Forecasting sales data: using time series data to predict a firm's future expected sales

UX: plot customer satisfaction and customer revenue to find that each cluster represents a different geographical location

4.9 Machine Learning: Techniques

Machine learning: Creating an algorithm, which the computer then uses to find a model that fits the data as best as possible to make very accurate predictions. In most situations, 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.

There are four ingredients for machine learning: data, model, objective function, optimization algorithm

Model: the computer uses an algorithm to recognize certain types of patterns

Objective function: specification of the machine learning problem; a function to be maximized or minimized depending on the task at hand

Optimization algorithm: A process in which previous solution of the problem are compared until reaching an optimal solution

4.10 Machine Learning: Types

Three main types of machine learning:

- Supervised learning

Training an algorithm resembles a teacher supervising her students. Provides feedback every step of the way. Telling students whether they did 'good' or whether they need to improve their performance.

When using supervised learning you use labelled data (every data point is categorized as 'good' performance or as 'performance that needs improvement' in our example).

- Unsupervised learning

In this case, the algorithm trains itself. There isn't a teacher who provides feedback. The algorithm uses unlabelled data that is not categorized as 'good' or as 'performance that needs improvement'. The unsupervised ML model simply uses the data and sorts in different groups. In our example, it will be able to show us two groups – 'good performing' and 'performance that needs to be improved', however the ML model would not be able to tell us which one is which.

- Reinforcement learning

A reward system is introduced. Every time a student does a task better than it used to in the past they will receive a reward (and nothing if the task is not performed better). Instead of minimizing an error, we maximize a reward, or in other words, maximizing the objective function.

Deep learning – the modern state-of-the-art approach to machine learning – leverages the power of neural networks and can be placed in both categories – supervised and unsupervised learning.

5. Common data science tools

There are two main types of tools one can use in data science – programming languages and software. Programming languages enable you to devise programs that can execute specific operations. Moreover, you can reuse these programs whenever you need to execute the same action.

Our annual research on 1,001 data scientist profiles shows that the most popular programming language for data science is Python followed by R. These languages are not just suitable for mathematical and statistical computations. They are general purpose programming languages.

Python and R have their limitations. They are not able to address problems specific to some domains. One example is 'relational database management systems'. In these instances, SQL works best.

In terms of software, Excel plays an important role. It is able to do relatively complex computations and good visualizations quickly. SPSS is another popular tool for working with traditional data and applying statistical analysis.

There is a significant amount of software designed for working with big data – Apache Hadoop, Apache Hbase, and Mongo.

PowerBI, Qlik, Tableau are top-notch examples of software designed for business intelligence visualizations.

6. Data science job positions

Data architect – designs the way data will be retrieved processed and consumed

Data engineer – process the obtained data so that it is ready for analysis

Database administrator – handles this control of data; works with traditional data

BI analyst – performs analyses and reporting of past historical data

BI consultant – 'external BI analyst'

BI developer – performs analyses specifically designed for the company

Data scientist – employs traditional statistical methods or unconventional machine learning techniques for making predictions

Data analyst – prepare advanced analyses

Machine learning engineer – applies state-of-the-art ML techniques

7. Dispelling common misconceptions

1. 200,000 lines of data constitute big data -It is not just volume that defines a data set as 'big' – variety, variability, velocity, veracity and other characteristics play an important role as well
2. Qualitative analysis such as SWOT are not used for quantitative analysis. Hence, they are not part of business intelligence
3. Software like Excel, SPSS, and Stata can be successfully used by data science teams in many companies
4. In deep learning, there is still a debate on WHY the algorithms used outperform all conventional methods.