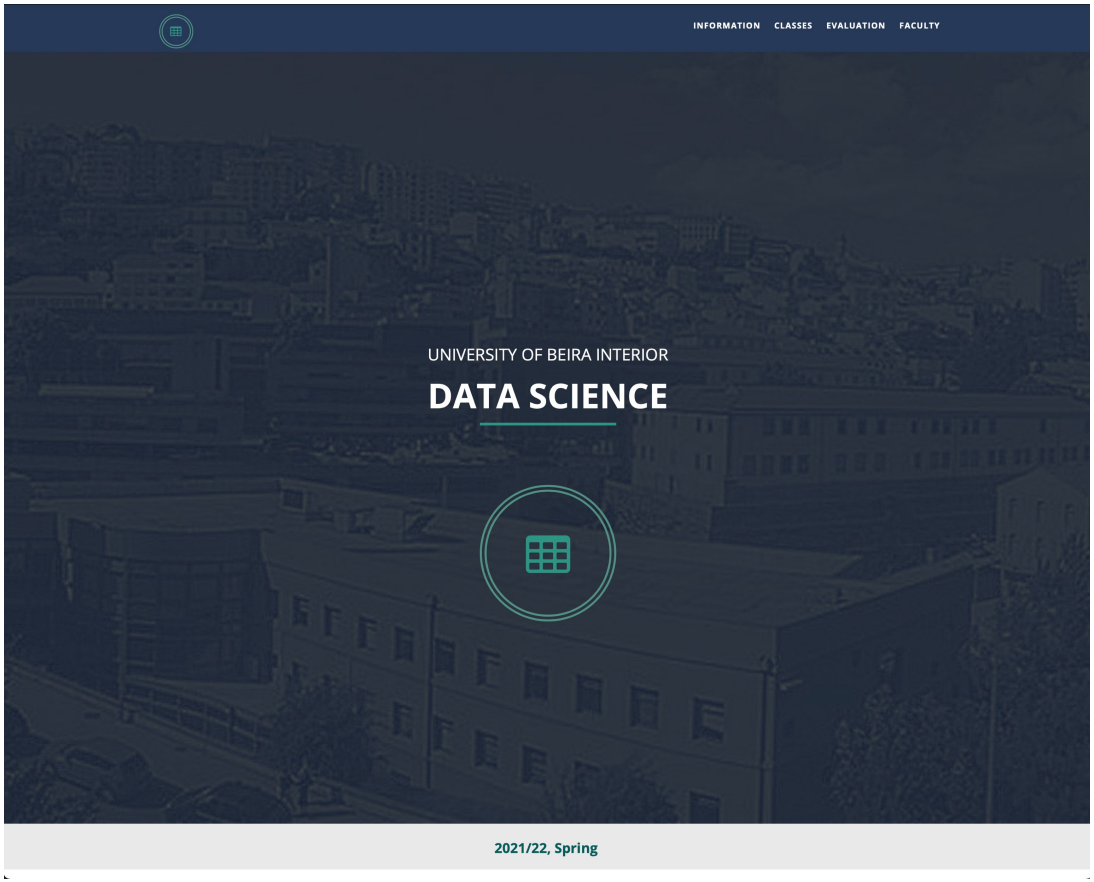
DATA SCIENCE MEI/1

University of Beira Interior
Department of Informatics

Hugo Pedro Proença,
hugomcp@di.ubi.pt, **2021/22**

Data Science @ UBI: Main Page

- **URL**: http://di.ubi.pt/~hugomcp/dataScience/
  - News
  - Program, Evaluation Criteria, Bibliography
  - Classes (Theoretical + Practical)
  - Evaluation



Data Science? What is It?

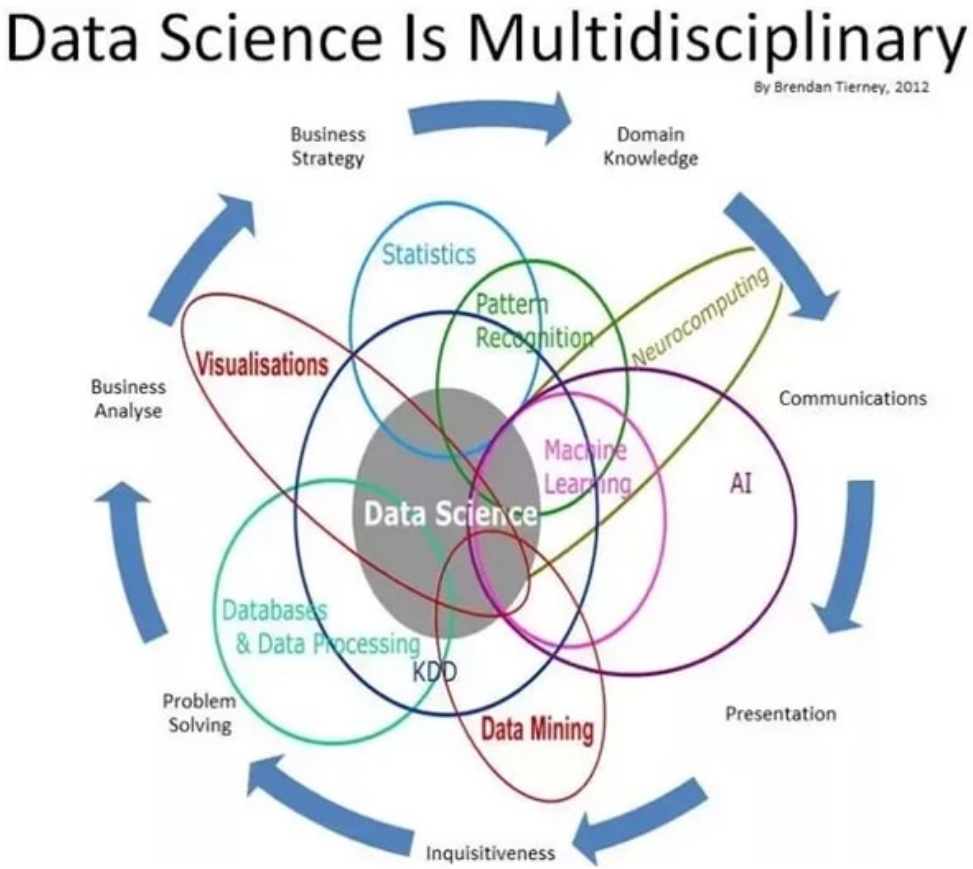
- At its core, **Data Science** is "*a field of study that aims to use a scientific approach to extract meaning and insights from data*".
  - But isn't this the same definition of **Pattern Recognition**?
  - **Pattern Recognition** is "*the automated recognition of patterns and regularities in data*".
  - In the **Machine Learning** context, Pattern Recognition refers to the assignment of a label to a given input value.
    - However, in Statistics, the term "Discriminant Analysis" was introduced for this same purpose in 1936.
- Though it may sound obvious, Data Science relies on data.
- The massive growth of Data Science was spurred by the availability of massive datasets and cheap computing power.
  - Only with these incredible resources Data Science becomes effective.
  - Small datasets, messy data, and incorrect data can waste a lot of time, creating models that produce meaningless or misleading results.
  - If the data doesn't capture the actual "*cause of variation*", Data Science will fail.

Data Science? What is It?

- **Data Science** is a broad term for multiple disciplines
  - It can be seen as an umbrella that encompasses Data Analytics, Data Mining and several other disciplines.
  - At the end, a **Data Scientist** is expected to forecast the future based on past patterns.
- In this setting, Machine Learning fits within data science.
  - Machine learning uses various techniques, such as regression and supervised clustering.
  - On the other hand, the data in **Data Science** may or may not evolve from a machine or a mechanical process.
- The main difference between the two terms is that Data Science is a broader term, that not only focuses on algorithms and statistics but also **takes care of the entire data processing methodology**.

Data Science? What is It?

☐ **Data Science** intercepts a broad range of disciplines**:**



Source: https://www.simplilearn.com/data-science-vs-data-analytics-vs-machine-learning-article

Data Science Topics

- **Programming Skills**
  - It is expected that Data Scientists know how to use the tools of the trade. This means a statistical programming language, like **R** or **Python** and a database querying language like **SQL**.
- **Statistics**
  - A good understanding of statistics is very important. Data Scientists should be familiar with statistical tests, distributions, maximum likelihood estimators, etc.
- **Machine Learning**
  - It is important to be familiar with the most popular machine learning methods.
    - Things like k-nearest neighbours, random forests, ensemble methods, and more.
    - It's true that a lot of these techniques are provided *per se* using R or Python libraries.
    - However, the most important is to understand the broad strokes and really understand when it is appropriate to use each different technique.
- **Multivariate Calculus and Linear Algebra**
  - Understanding these concepts is particular important for cases where small improvements in the predictive performance can lead to huge wins for the company.

Data Science Topics

- **Data Wrangling**
  - It's important to know how to deal with imperfections in data.
  - Some examples of data imperfections include missing values, inconsistent string formatting (e.g., 'New York' versus 'new york' versus 'ny'), and date formatting ('2017-01-01' vs. '01/01/2017', unix time vs. timestamps, etc.).
- **Data Visualization and Communication**
  - Visualizing and communicating data is particularly important.
    - Communicating, this means describing your findings, or the way techniques work to audiences, both technical and non-technical.
    - Visualization-wise, it can be helpful to be familiar with data visualization tools like "matplotlib", "ggplot", or "d3.js".
      - Tableau has become a popular data visualization and dashboarding tool as well. It is important to not just be familiar with the tools necessary to visualize data, but also the principles behind visually encoding data and communicating information.
- **Software Engineering**
  - Typically, the Data Scientist will be responsible for handling a lot of data logging, and potentially the development of data-driven products.

What Data Science **is Not**

- The "**predictive**" feature is the main requirement in Data Science projects.
  - Being able to predict an event is a powerful thing. However, a data scientist not only builds predictive models all day at work.
- Instead, there are multiple layers in a Data Science project.
  - The model building part is just a component in the overall Data Dcience lifecycle.
  - There are many others:
    - Understanding the problem statement
    - Hypothesis building
    - Data collection
    - Verifying the data
    - Data cleaning
    - Exploratory analysis
    - Designing the model
    - Testing/Verifying the model
    - If an error is found, head back to the verification or cleaning stage
    - Putting it into production (deploying the model)

Typical Mistakes in Learning Data Science

- "*Seeing and seeing without practicioning*"
  - Data science needs practice. Everything you learn, you should also practice and try it.
  - This is fundamental to really comprehend things and when you are working in the field you will be doing a lot of different practical stuff.
  - A good knowledge on statistics, math and python won't make you a successful data scientist. You need more, you need to master your craft. Be able to use these tools to solve business problems. So if you are learning something new, and you want to understand it for real, find a scenario where you can apply it or play with it.
- "*Creating Models in an Automated Way*"
  - Before creating any model, we should understand (interpret) the problem and the data available.
  - When we are understanding we are decoding the parts that forms this complex thing, and transforming the raw data we got in the beginning into something useful and simple.
  - We do this by modelling. This is the process of understanding the "reality", the world around us, but creating a higher level prototype that will describe the things we are seeing, hearing and feeling, but it's a representative thing, not the "actual" or "real" thing.

Data Science Curiosities

- Less than **0.5%** of all data we create is ever analysed and used.
- A **10%** increase in data accessibility will result in more than **$65 million** additional net income for the typical Fortune 1000 company.
- The US leads the data science market, requiring **190,000** data scientists by next year.
- Google uses about **1,000** computers to answer every single search query.
- In **2020**, there are more than 50 billion smart connected devices in the world, collecting, analysing and sharing data.
- By **2025**, the Data Science analytics sector in India is estimated to grow eightfold, reaching **$16 billion**.
- Last year, an estimated **1 trillion** photos were taken and billions of them will be shared online.
- Nearly **80%** of photos are taken on smart phones and most will become searchable data online.
- Every second **60,000** search queries are performed on Google and **1.2 trillion** searches per year.
- In **2020**, about 1.7 megabytes of new information is created every second for every human on the planet.
- **Data scientists** earn more than the average IT employees.
- **1 billion** pieces of content are shared via Facebook's Open Graph every day.
- Bad data costs US businesses alone **$600 billion** annually.
- **70%** of data is created by individuals, but enterprises are responsible for storing and managing **80%** of that.
- There are nearly as many pieces of digital information as there are stars in the universe.

Data Science Applications

- **Health Care. LYNA: Google Machine Learning for Metastasis**
  - Google developed a new tool for identifying breast cancer tumors that metastasize to nearby lymph nodes. That can be difficult for the human eye to see, especially when the new cancer growth is small.
  - In one trial, LYNA — short for Lymph Node Assistant —**accurately identified metastatic cancer 99 percent of the time** using its algorithm.

Data Science Applications

- **Health Care. ONCORA Medical: Cancer Care Recommendations**
  - Oncora's software uses Data Science tools to create personalized recommendations for current cancer patients based on data from past ones.
  - Health care facilities using the company's platform include New York's Northwell Health. Their radiology team collaborated with Oncora data scientists to mine 15 years' worth of data on diagnoses, treatment plans, outcomes and side effects from more than 50,000 cancer records.
  - Based on this data, Oncora's algorithm learned to suggest personalized chemotherapy and radiation regimens.

Data Science Applications

☐ Logistics. UPS: Optimizing Package Routing

- ☐ **UPS** uses data science to optimize package transport from drop-off to delivery. Its latest platform for doing so, Network Planning Tools (NPT), incorporates machine-learning and AI to crack challenging logistics puzzles, such as how packages should be rerouted around bad weather or service bottlenecks.
- ☐ NPT lets engineers simulate a variety of workarounds and pick the best ones; AI also suggests routes on its own. According to a company forecast, the platform could save UPS $100 to $200 million by 2020.

Data Science Applications

- **Commerce. Zara**
  - Zara uses intensive data analytics to manage a tight supply chain and give customers exactly “what they want”.
    - Sometimes even before the customers themselves know that they want it.
  - Zara has turned the industry on its head by using data and analytics to track demand on a real-time, localized basis and push new inventory in response to customer pull. This enables them to manage one of the most efficient supply chains in the fashion industry, and to create the fast fashion category as a market leader.