Basics of Data Science

Dr. Anthony Mile
CUK

What is Data Science?

Data science is the study of data to extract meaningful insights for business. It is a multidisciplinary approach that combines principles and practices from the fields of mathematics, statistics, artificial intelligence, and computer engineering to analyze large amounts of data.

In Wikipedia, **Data Science** is defined as *a scientific field that uses scientific methods to extract knowledge and insights from structured and unstructured data, and apply knowledge and actionable insights from data across a broad range of application domains*.

Why Data Science?

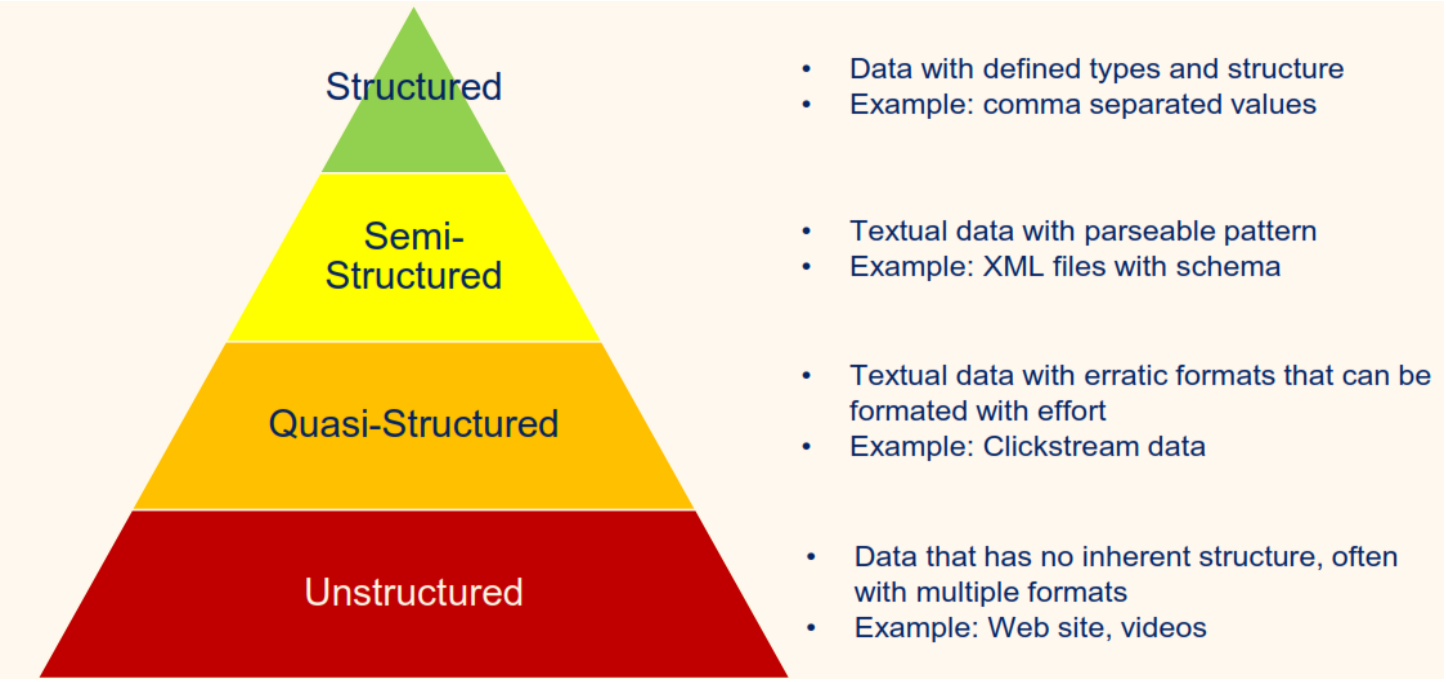
- Data is the oil for today's world. With the right tools, technologies, algorithms, we can use data and convert it into a distinct business advantage
- Data Science can help us to detect fraud using advanced machine learning algorithms
- It helps us to prevent any significant monetary losses
- Allows to build intelligence ability in machines
- It enables us to take better and faster decisions
- It helps us to recommend the right product to the right customer to enhance your business

Other Related Fields

- Databases
- Big Data
- Machine Learning
- Artificial Intelligence
- Visualization

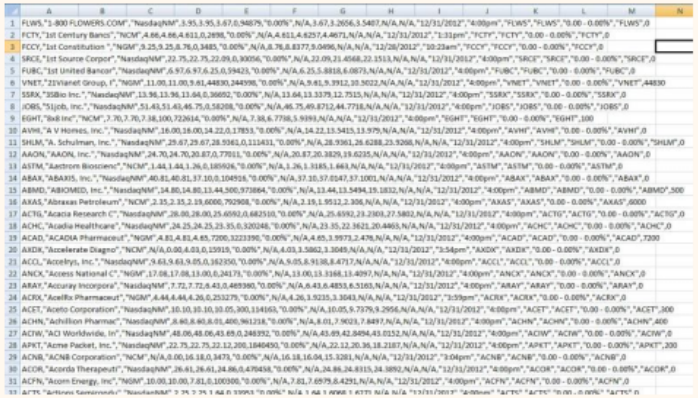
Types of Data

- Structured
- Unstructured
- Semi Structured



Semi-Structured: Doesn't follow the tabular structure associated with relational databases or other forms of data tables. However, it does contain tags and metadata to separate semantic elements and establish hierarchies of records and fields

Structured



Quasi-Structured

Semi-Structured

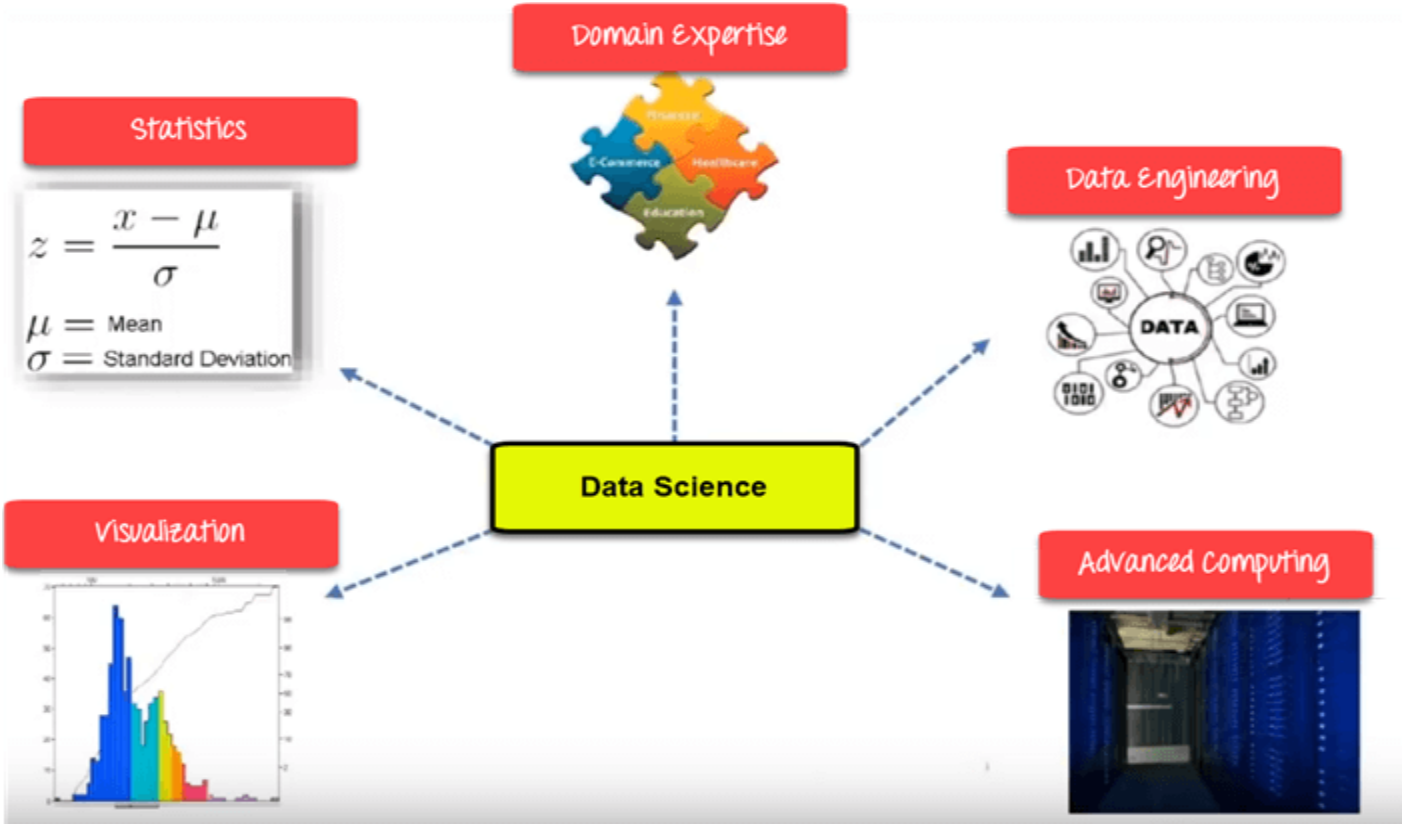
```
<?xml version="1.0" encoding="iso-8859-8" standalone="yes" ?>
<CURRENCIES>
  <LAST_UPDATE>2004-07-29</LAST_UPDATE>
  <CURRENCY>
    <NAME>dollar</NAME>
    <UNIT>1</UNIT>
    <CURRENCYCODE>USD</CURRENCYCODE>
    <COUNTRY>USA</COUNTRY>
    <RATE>4.527</RATE>
    <CHANGE>0.044</CHANGE>
  </CURRENCY>
  <CURRENCY>
    <NAME>euro</NAME>
    <UNIT>1</UNIT>
    <CURRENCYCODE>EUR</CURRENCYCODE>
    <COUNTRY>European Monetary Union</COUNTRY>
    <RATE>5.4417</RATE>
    <CHANGE>-0.013</CHANGE>
  </CURRENCY>
</CURRENCIES>
```

Unstructured

Digitalization and Digital Transformation

In the last decade, many businesses started to understand the importance of data when making business decisions. To apply data science principles to running a business, one first needs to collect some data, i.e. translate business processes into digital form. This is known as **digitalization**. Applying data science techniques to this data to guide decisions can lead to significant increases in productivity called **digital transformation**.

Data Science Components



Statistics & Visualization

- Statistics is the most critical unit of Data Science basics, and it is the method or science of collecting and analyzing numerical data in large quantities to get useful insights.
- Visualization technique helps us access huge amounts of data in easy to understand and digestible visuals.

Machine Learning & Deep Learning

- AI enables the machine to think, that is without any human intervention the machine will be able to take its own decision.
- Machine Learning is a subset of Artificial Intelligence that uses statistical learning algorithms to build systems that have the ability to automatically learn and improve from experiences without being explicitly programmed.
- Deep learning is a machine learning technique that is inspired by the way a human brain filters information, it is basically learning from examples. It helps a computer model to filter the input data through layers to predict and classify information.

Data Engineering

- Data engineering refers to the building of systems to enable the collection and usage of data. This data is usually used to enable subsequent analysis and data science; which often involves machine learning.

Advanced Computing

- **It** refers to systems with the ability to process data and perform calculations at high speeds, such as supercomputers.

Data Science Process

Discovery

Discovery step involves acquiring data from all the identified internal & external sources, which helps us answer the business question.

The data can be:

- Logs from webservers
- Data gathered from social media
- Census datasets
- Data streamed from online sources using APIs

Preparation

- Data can have many inconsistencies like missing values, blank columns, an incorrect data format, which needs to be cleaned. We need to process, explore, and condition data before modelling. The cleaner our data, the better are our predictions.

Model Planning

- In this stage, we need to determine the method and technique to draw the relation between input variables. Planning for a model is performed by using different statistical formulas and visualization tools. SQL analysis services, R, and SAS/access are some of the tools used for this purpose.

Model Building

- In this step, the actual model building process starts. Here, Data scientist distributes datasets for training and testing. Techniques like association, classification, and clustering are applied to the training data set. The model, once prepared, is tested against the "testing" dataset.

Operation

- We deliver the final baselined model with reports, code, and technical documents in this stage. Model is deployed into a real-time production environment after thorough testing.

Communicate Results

- In this stage, the key findings are communicated to all stakeholders. This helps us decide if the project results are success or failure based on the inputs from the model.

Tools for Data Science

Java

- Java can be used for many of the processes:
- Data import and export.
- Cleaning data.
- Statistical analysis.
- Machine learning and Deep learning.
- Deep learning.
- Text analytics (also known as Natural Language Processing or NLP).
- Data visualization.

R

- R is a popular programming language used for statistical computing and graphical presentation.
- Its most common use is to analyze and visualize data.

Why Use R?

- It is a great resource for data analysis, data visualization, data science and machine learning
- It provides many statistical techniques (such as statistical tests, classification, clustering and data reduction)
- It is easy to draw graphs in R, like pie charts, histograms, box plot, scatter plot, etc..
- It works on different platforms (Windows, Mac, Linux)
- It is open-source and free
- It has a large community support
- It has many packages (libraries of functions) that can be used to solve different problems

Python

- Python is open source, interpreted, high level language and provides great approach for object-oriented programming. It is one of the best language used by data scientist for various data science projects/application. Python provide great functionality to deal with mathematics, statistics and scientific function. It provides great libraries to deals with data science application.

SAS

SAS stands for **Statistical Analysis Software**. It was created in the year 1960 by the SAS Institute. From 1st January 1960, SAS was used for data management, and business intelligence, Since then, many new statistical procedures and components were introduced in the software.

Why we use SAS

- Data Management
- Statistical Analysis
- Report formation with perfect graphics
- Business Planning
- Application Development
- Data extraction
- Data transformation
- Data updation and modification

MATLAB

- MATLAB offers a notebook environment, toolboxes, and apps for developing analytic models.
- Using MATLAB we can combine statistics and machine learning with application specific techniques such as signal processing, image processing, text analytics, optimization and controls

Thank u

?