Introduction – Data Science

What is Data Science?

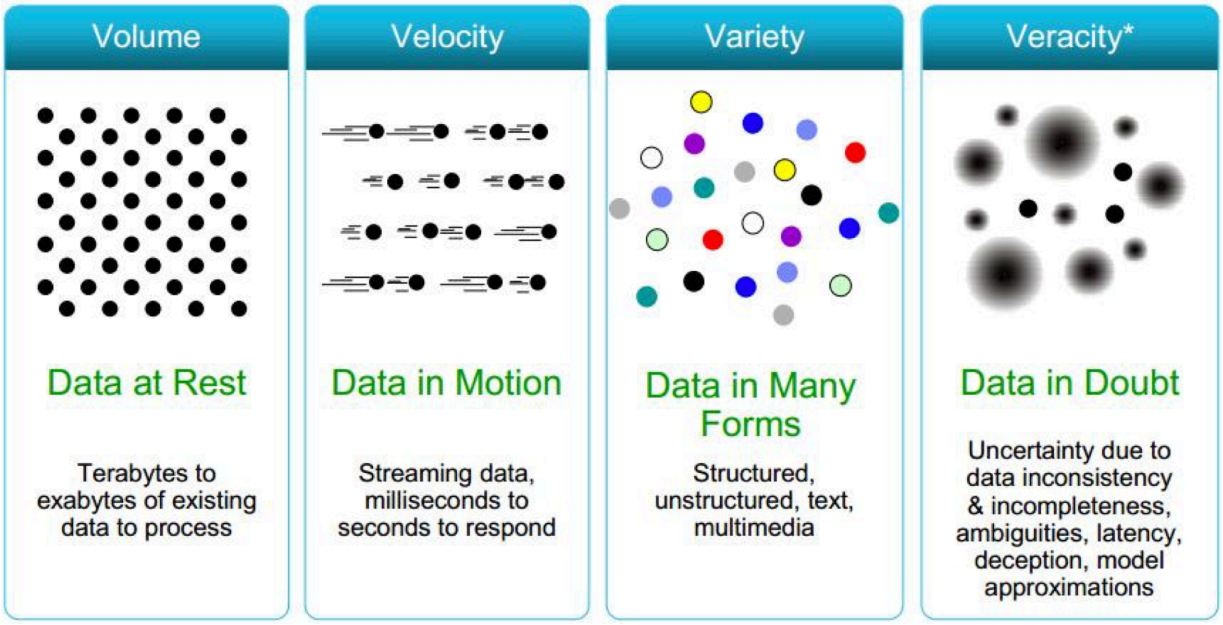
- The extraction of knowledge from large volumes of data that are structured or unstructured, which is a continuation of the field of Data Mining and Predictive Analytics.

When is Data Science useful?

- Data Science is only useful when the data are used to answer a specific concrete question.

The structure of Data Science

- The process of formulating a question
- Collecting and cleaning the data
- Analysing the data
- Communicating the answer



Why is volume important?

- We have data we did not have before: GPS coordinates from photos & cars, large genome databanks.
- This data allows us to answer questions we could not answer before.

- Large sets of (human generated) data can replace complex algorithms
    - People leave traces of 'intelligence' in the data with their intelligent behaviour on everyday tasks
    - You need a lot of data (specially to cover rare events)

Why is uncertainty important?

- If data is factual, answering is straightforward
- However, data must always contain uncertainty
- Therefore, we need statistics to estimate possible solutions to questions.

Wat is Machine Learning:

- Computer program that modifies or adapts its actions so that these actions get more accurate/optimal

Manual engineering becomes infeasible when:

- Data volume grows
- Data complexity grows (many features)
- There is less time to develop