Fundamentals of Data Science

Introduction: Core Concepts and Technologies

Southampton
Data Science
Academy

The data in Data Science

Data is everywhere

Data is not all the same

- Qualitative vs quantitative
- Discrete vs continuous
- Structured vs unstructured (vs semi-structured)
- Open vs closed

Quantitative vs qualitative data

Quantitative

- numerical data
- discrete or continuous
- could be ordered

Qualitative

- non-numerical data
- could be ordered
- expressed in natural language
- types eg. binomial; nominal & ordinal

Discrete vs continuous data

- Discrete data takes specific values
  - E.g., integers, number of cars owned, days of the week, age, etc.
- Continuous data may take any value
  - E.g., real numbers, weight, height, GDP etc.
- Distinction sometimes subject to conventions/application scenarios

Examples

- Number of cars owned – quant, discrete
- Favourite football team – qual, nominal
- Day of the week – qual, ordinal
- Height (mm) – quant, continuous
- Height (short/medium/tall) – qual, ordinal

Structured vs unstructured data

Structured

- stored or organised according to a pre-defined data model
- has typically been cleaned to remove erroneous or irrelevant elements
- often stored in a (relational) database eg sensor data; call records; web server logs

Unstructured

- no pre-defined data model, and data that is not organised according to any particular structure eg text

Unstructured data

- Requires new types of storage
- Relational databases aren’t suitable when the data structure could change at any time, or new types of data could be incorporated.
- Solutions such as NoSQL help to overcome these restrictions.

Source: http://www.couchbase.com/cn/why-nosql/nosql-database

Semi-structured data

- A form of structured data, but without the formal structure of a relational database.
- Contains tags or markers to provide structure.
- Can therefore be transformed easily into relational data.
- E.g., XML, JSON

Open Data

Open data is: (various definitions)

- “data that anyone can access, use and share.” (The ODI, 2015)
- Most definitions focus on the possibility to freely use, re-use and redistribute licensed data.

Governments and businesses alike are publishing open data to increase transparency and drive innovation.

https://theodi.org/guides/what-open-data