Lecture 1: Fundamentals of Data Visualisation

4 Vs of Big Data
- Data Volume: MB, GB, TB, EB
- Data Velocity: Real Time, Static
- Data Veracity: certain, uncertain
- Data Variety: homogeneous, heterogeneous

Data Science:
- data wrangling, databases, software engineering, visualisation, domain program, statistics, machine learning

Visualisation:
- the process of transforming data into static or interactive machine representation
- provide visual representations of dataset designed to help people carry out tasks more effectively
- the process that transforms data into interactive graphical representations for the purpose of **exploration**, **confirmation**, or **presentation**

Why visualise:
- Presentation: creating visualisations to just communicate something they already know
- Confirmation: creating visualisations to verify or falsify hypothesis that you have
- Exploration: find something new in the data that you have not expected

Presentation:
- Present that the creator of the visualisation already knew

Exploration/confirmation:
- people are looking at the data and tried to find something new
- Exploratory data analysis: an approach of analysing data to summarise the main characteristics without using statistical model or having formulated a prior hypothesis

Why have a human in the loop
- Not to visualize:
  - Well-defined questions on well-defined dataset:
    - Computer can be better than human
    - use statistics/machine learning
- To visualize:
  - Can't formulate the question in advance
  - III-specified question, such as caner research

Why have a computer in the loop
- Scale
  - Drawing by hands unfeasible
  - Interaction allows to drill down into the data
  - Integration with algorithms
- Efficiency
  - Re-use charts for different datasets: apply to any data that you upload
- Quality
  - Precise data-driven rendering
- Storytelling

Why show Data in Detail

- Anscombe;s Quartett: different datasets have the same basic statistical measures, but different characteristics is they are plotted

Why use interactivity:
- When the data gets bigger and more complex, you are running into the limitations that people in this place have. Therefore, a single static visualisation or a seamless static view that only shows one of a few aspects of your data cannot cover the whole complexity of the dataset .

Why is it so hard to create effective visualisations:
- There are countless visual encodings and possible interactions that you could combine in order to create an effective visualisations

Peutinger Map:
- showing roads of Roman Empire
- distortion: to compress the content to get the road network onto these parchment pages

Minards' Map:
- Napoleons' March on Moscow
- first example of multiple coordinates view where you have two visualisations that are coordinated

Data Types and Characteristics (WHAT)

Data Types
- Fundamentals units
- Combinations make up dataset types
- Structural interpretation of data
- Items: discrete individual entities
  - machine, worker, city
- Attributes: measured, observed or logged properties of items
  - age, price, temperature
- Links: relationship between items:
  - Facebook friendship, connections between circuit elements
- Position: spatial data providing location in 2D or 3D space
  - Long/lat pair of city, pixel in photo, voxels in MRI scan
- Grids: sampling strategy for continuous data
  - Grid of weather stations in a region

Dataset Type
- Tables:
  - Items, Attributes
- Network & Trees:
  - Items (nodes), Links, Attributes
- Fields:
  - Grids
  - Positions
  - Attributes
- Geometry:
  - Items, Positions
- Clusters, Set, Lists
  - Items

Attribute Types:

- **Categorical (nominal)**:
  - Compare equality, non implicit order. e.g., fruit, gender, product category, file types
- **Order**:
  - Ordinal: great/less than defined. e.g. shirt size, rankings
  - Quantitative: arithmetic possible. e.g. length, weight, count
- **Interval (arbitrary zero)**: cannot compare directly, only differences can be compared. e.g., dates, temperature in C&F
- **Ratio (true zero):** there is nothing of the measured entity observed, can measure rates & proportions. e.g., length, mass
- Sequential: homogeneous from min to max. e.g., # of people in countries
- Diverging: two or multiple sequences that meet at common zero point. e.g., elevation dataset (above sea level & below sea level)
- Cyclic: time(hour, week, month, year). e.g., seasons of the year

Visualising graph:

- Node-link diagrams:
- Adjacency Matrix: all nodes are contained as columns and rows
- True map: visualise the content of a file system or if you want to visualise the map of the market