UNIVERSITÄT BERN

CAS Applied Data Science - Module 1

Data Acquisition and Management

PD Dr. Sigve Haug

Bern, 2020-09-20

Module 1 Overview

First day

- About data and working with data
- Infrastructures for data
- Data sources and acquisition

Second day

- Visualisation of data
- Data management planning

Third day

- Collecting data from www
- Data bases
- Project abstracts and clarifications

Project

- Written report by 2018-09-23

Module 1
Second day

09:00 Discussion session

09:30 Visualisation

- Lecture
- Notebook tutorial

10:30 Break

11:00 Notebook tutorial

-

12:30 Lunch

13:30 Databases and MySQL

~~17:00~~ End

Questions from yesterday

- ...
- ...

- How can you show all rows of a dataframe in a jupyter notebook? Set maxrows = None
- Is jupyter safe to use (from a data security standpoint)? Jupyter is as safe as any other file on your computer (assuming you don't use it in the cloud, google colab, etc.).
- Why do you not always import the full libraries but just certain modules from libraries? In order to be memory-efficient

Visualisation of data

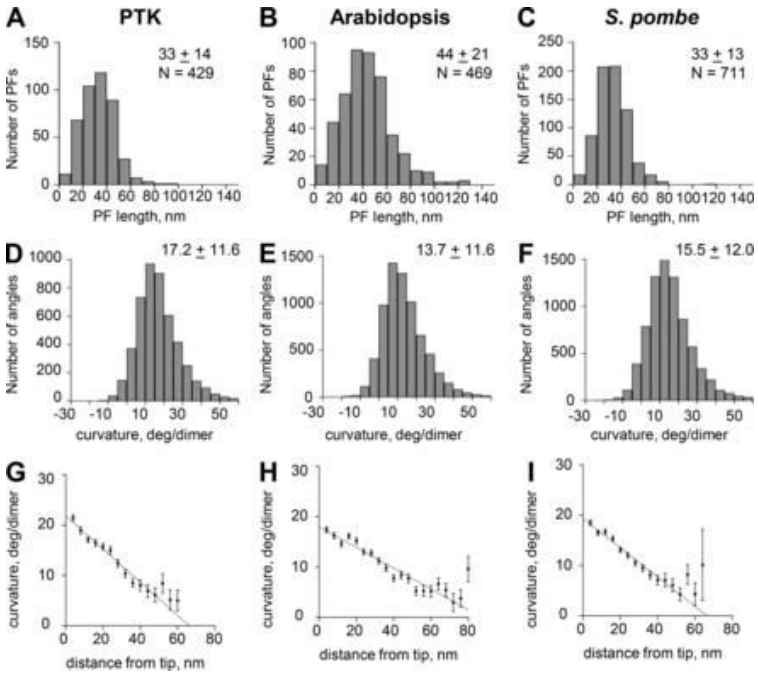
Why visualization?
Because it is more intuitive for humans (much loss for computers though).

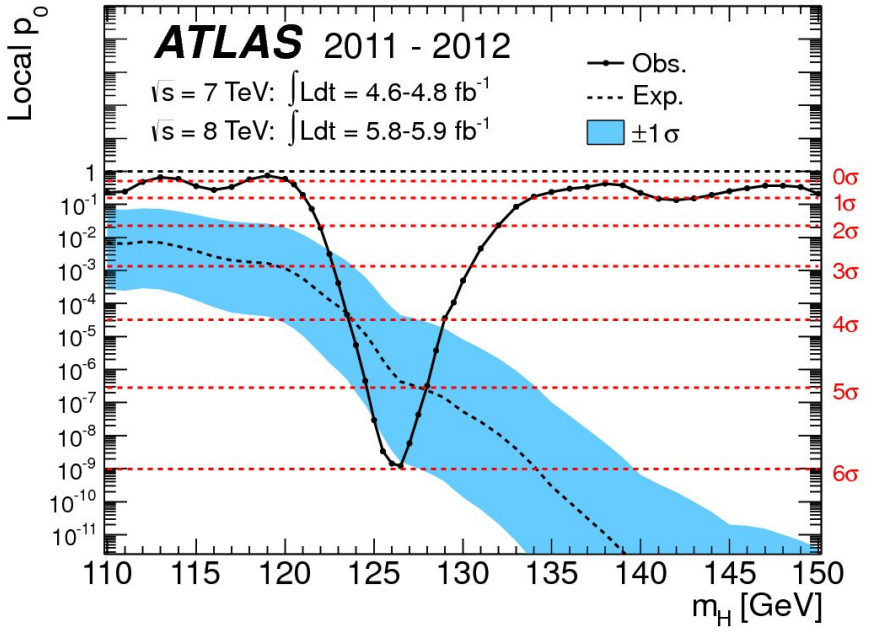
Data is any sequence of symbols.
Information is data in context.

Overview [edit]

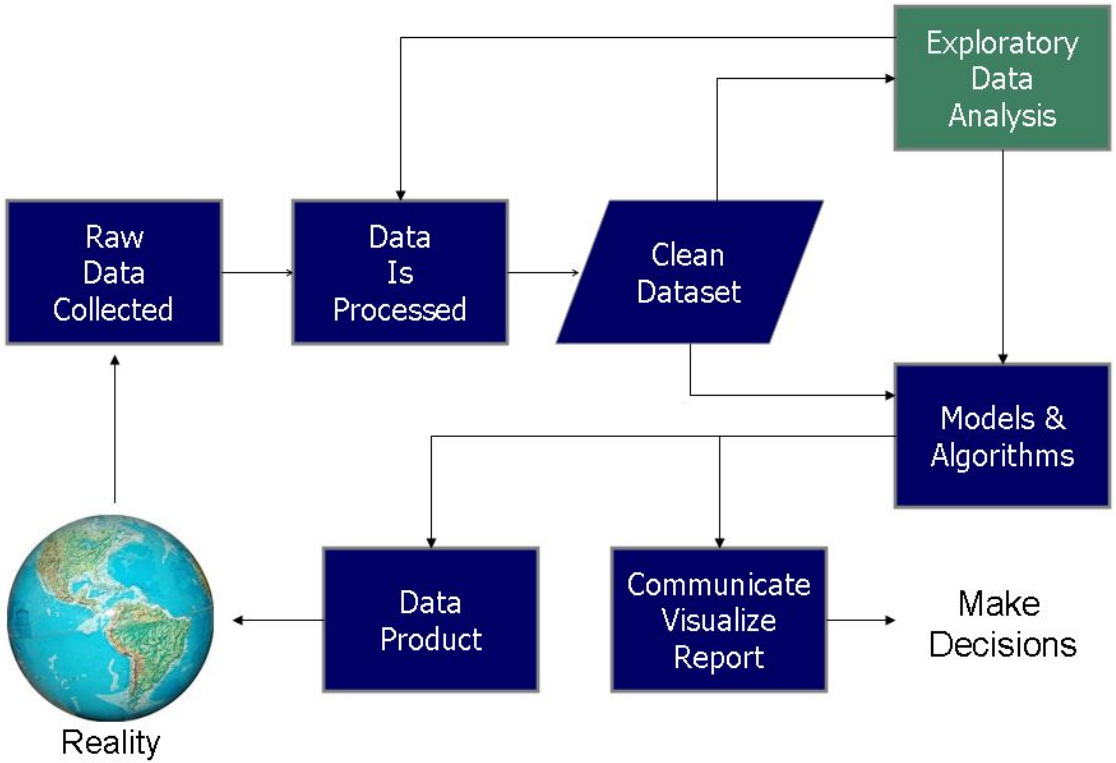
Data visualization refers to the techniques used to communicate data or information by encoding it as visual objects (e.g., points, lines or bars) contained in graphics. The goal is to communicate information clearly and efficiently to users. It is one of the steps in data analysis or data science. According to Friedman (2008) the

Visualisation examples





Data Science Process



Example of visualisation in the data science process (actually it is all over)

Visualisation of data - a division

Visualisation is basically about communication (sender / receiver). Your type of visualization depends on your target group (do they have domain knowledge / statistical knowledge?, etc.)

Descriptive

- Describes the data
- Helps understand the data
- Do as many as possible at every stage
- Do different spaces/representations
- Look for patterns, similarities differences, significant features, correlations ..

Inferential

- Communicates information and knowledge inferred from the data
- Can be complex / compact
- Normally your “goal” - whole data science process is about improving the final inferential graph

How do you read a publication (paper, book, report)?

Typically

- Quickly read the abstract
- Scan introduction and conclusion (for important numbers)
- Study **figures and graphs**
- Study tables
- Check if there are known references
- Dig into the text

So visualisation is important

- People with power don’t have time
- Normally your space and time for communication are limited
- Need to pass your message in an elevator (20 seconds)
- Good visualisation communicates trust, results and interpretations
- Also helps you understand your data

Visualisation of data - general considerations

Communication

- Sender - Message - Receiver
- The sender should have a clear motivation and be trustworthy
- Choose the right medium for message
- The message should be clear and decodable and interpretable for the receiver

Human cognition

- Most graphics (still) target humans
- Should therefore take
  - Cognition
  - Pre-attentive attributes
- into account

Visualisation of data - general guidelines

Graphs should reveal data

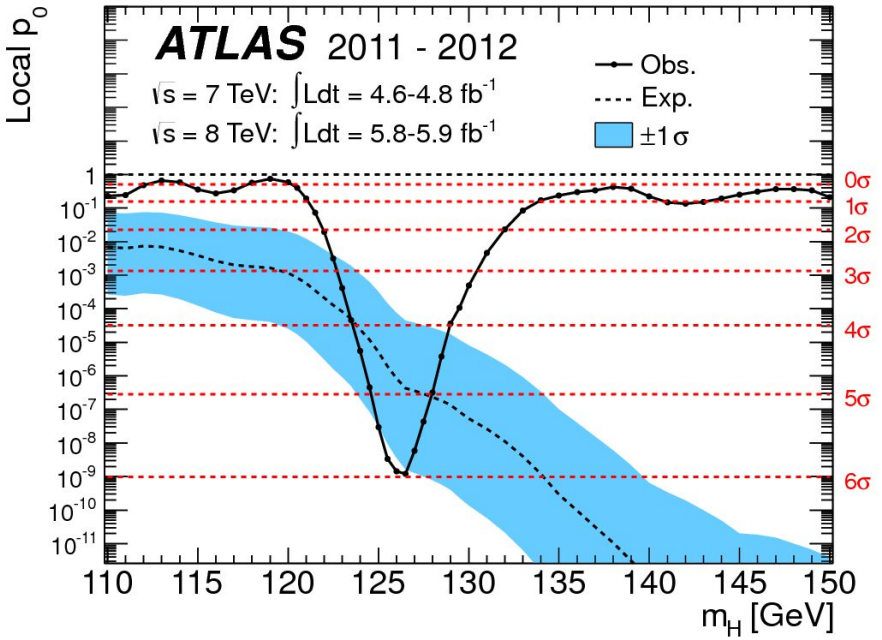
- Show the data
- Make the viewer think about the message/data
- Avoid distorting from the message/data
- Present many number in a small space
- Encourage comparison of different pieces of data
- Show several levels of detail (from overview to fine structures)
- Serve a clear purpose
- Be closely integrated with other description of the the data (text, tables etc)

Edward Tufte, The Visual Display of Quantitative Information, 1983

Visualisation of data - concrete guidelines

Important points for graphs

- **Axes labeling with units** most important!
- Sufficient but not redundant informatic for understanding
- Readability and visibility
- In publications **figure legend and reference in text**
- In science very often **uncertainties** should be included see graphic (the blue band stands for the standard deviation)



Visualisation of data

8 message types and graphs

- Time series
- Ranking
- Part to whole
- Deviation
- Frequency distributions
- Correlation
- Nominal comparison
- Geospatial and geographic

Stephen Few-Perceptual Edge-Selecting the Right Graph for Your Message-2004

For sure there are more ...

Visualisation tools

Spreadsheets

- Easy plotting by mouse clicking
- Limited customisation possibilities
- In (hard) sciences often below standard

R, Python etc

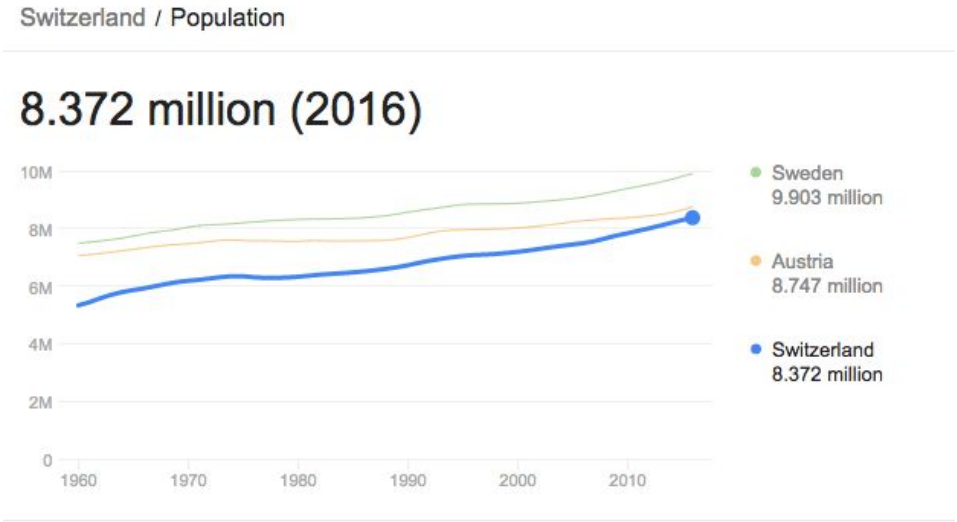
- Programmable plotting
- Highly customisable and automisable
- With effort almost everything can be done (scripted/programmed)

We don’t cover drawings, diagrams, sketches etc (see for example gimp)

Visualisation of data - graphs

Line charts

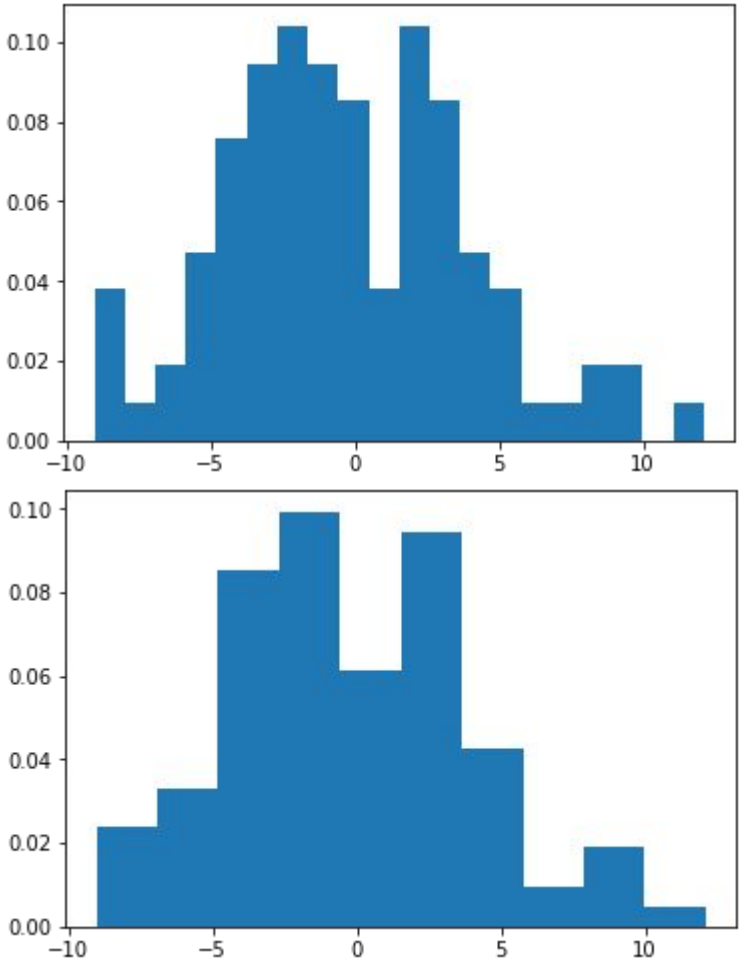
- Time series
- How a variable varies with time
- Example - CH population
-



Graphs - histograms

Frequency distribution

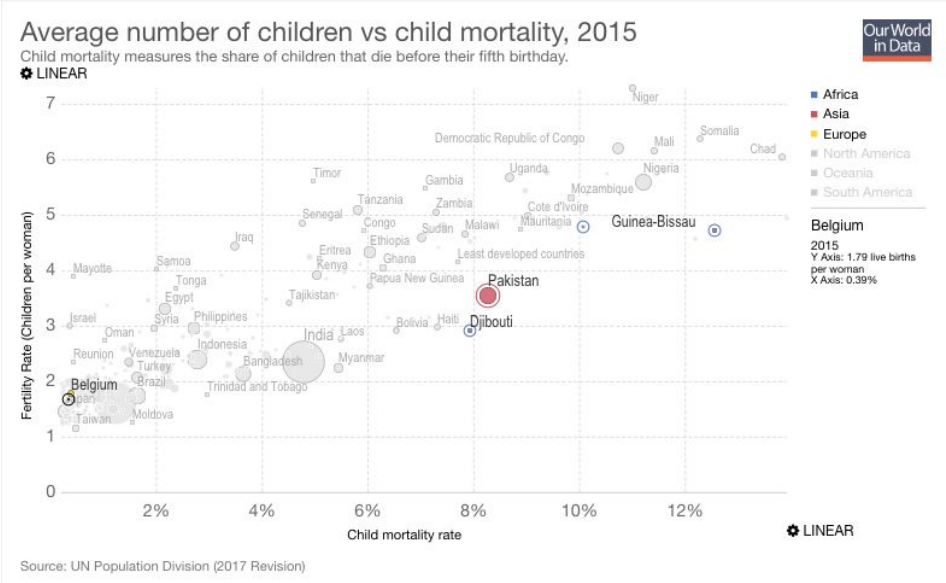
- Samples the data into bins
- Shows the amount of data in each bin
- Many bins increase the visual fluctuation
- Few bins may hide structures



Graphs - scatter plots

Shows correlations

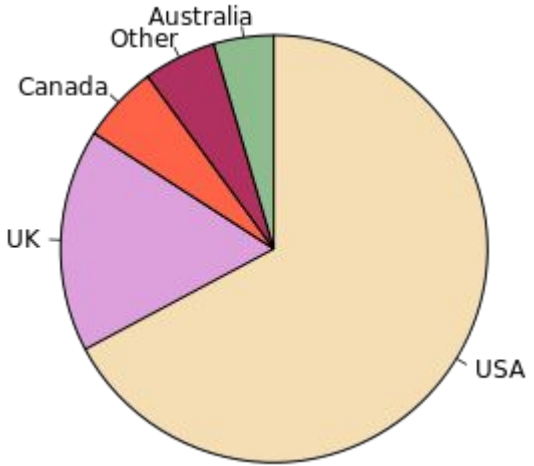
- Comparison between observations represented by two variables (X,Y) to determine if they tend to move in the same or opposite directions
- Example - human fertility versus child mortality
- Scatter plots are often used
- Can be 2 or 3 dimensional
- Box plots may indicate frequency too



Graphs - pie charts

Parts-to-whole

- Categorical subdivisions are measured as a ratio to the whole (i.e., a percentage out of 100%)
- Example - countries with their parts of the total english speaking world population
- Pie charts can be used



UNIVERSITÄT BERN

CAS Applied Data Science - Module 1

Data Acquisition and Management