Data acquisition and management

CAS Applied Data Science - Module 1

Dr. Anja Mühlemann

Module 1 – Purpose and Format

Purpose

- Think about data
- Get used to the tools for working with data
- Establish skills needed for the upcoming modules

Format

- Presentations
- Discussions
- Work on Notebooks

Schedule

Wednesday

- About data
- Data Management

Thursday

- Data visualisation
- Web scraping and APIs
- Project clarifications

Friday

- Web scraping and APIs
- Databases

Project

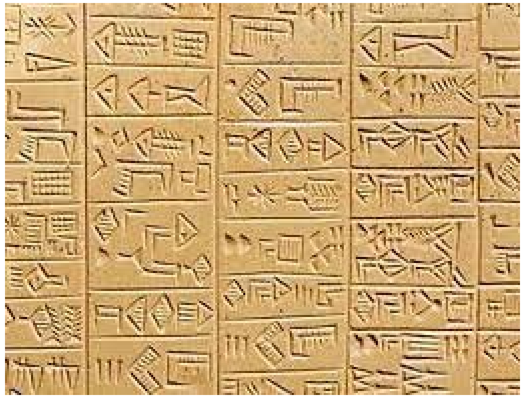
- Produce a Conceptual Design Report for a Data Science Project (deadline 2025-10-05?)

What is data?

- plural of *datum*, "(thing) given“
- observable, measurable or statistically collectable values. For example, in the form of symbols or numbers.
- can be digital or analog
- Needs processing and interpretation to become information.

Examples

- Survey responses
- Prices for same product in different shops



Data and Metadata

Data example

- year of birth,
- gender,
- weight,
- heigth,
- and serum iron levels

of participants of a study.

Metadata (data about the data) example

- Units
- Author
- Date
- Location
- …

Data Representations

- Often data is represented by numbers, words or symbols.

Common data types

- Integer (natural numbers)
- Float (decimal numbers)
- Boolean (TRUE/FALSE)
- Character (a,b,c,…)
- String (sequence of characters)
- Array (list of elements)
- Dataframe (combination of the aforementioned)

Declaration

- In most programming languages the data types must be specified.
  eg: *int counter = 2*
- In Python and R the data types don't need to be specified.
  eg: *counter = 2*

Storing Data on Computers

- Computers work based on electrical currents. Thus, there are only two states *current* (1) or *no current* (0) for transmission, or *presence of an electrical charge* (1) or *absence* for storage.
- Therefore, any number or character is saved as a binary number.

Example of binary representation

The number 13 using decimal numbers equals the binary number
$13_{10} = 8 + 4 + 1 = 1101_2$

$2^4 = 16$	$2^3 = 8$	$2^2 = 4$	$2^1 = 2$	$2^0 = 1$
	1	1	0	1

$10^4 = 10000$	$10^3 = 1000$	$10^2 = 100$	$10^1 = 10$	$10^0 = 1$
			1	3

Storing Data on Computers

- The space needed to save one binary digit is called *bit*
- 8 bits = 1 Byte (space needed for one letter using extended ASCII)
- 1000 Bytes = 1 kB
- 10^6 Bytes = 1 MB
- 10^9 Bytes = 1 GB
- This is where the terminology for storage of computers, USB-sticks, etc. comes from.

Example: The text of the Lord of the Rings trilogy uses approximately 2.5 MB of storage. An average hard drive could hold about 200’000 copies. In comparison, a single compressed photo uses about 5MB.

Formats

- Moderately sized data sets are often recorded in CSV or XLSX.
- When the data exceeds Excel's capacity or requires extra safeguards, databases are used.
- If the data set includes images, sounds, or similar content, the appropriate formats are used, with the database storing the paths to their respective locations.

Challenges

Working with data often presents several challenges

- Data entry errors
- Data from different sources and formats
- Missing data
- Large amounts of data

→ Today we look at how to import, handle and join data sets in Python.