Data Analysis

Introduction to Data Analysis in Pandas

Jens Hahn

Humboldt-Universität zu Berlin
Group of Theoretical Biophysics

November 23rd 2018

This lecture

1. Why should I care about data?
2. What is data?
3. What do I have to do?
4. Pandas
5. Assignments

Why should I care?

Data is everywhere

Experimental BP

- Own measurements
- Data comparison
- Storage

Theoretical BP

- Cooperation partners
- Parameter estimation
- Model analysis

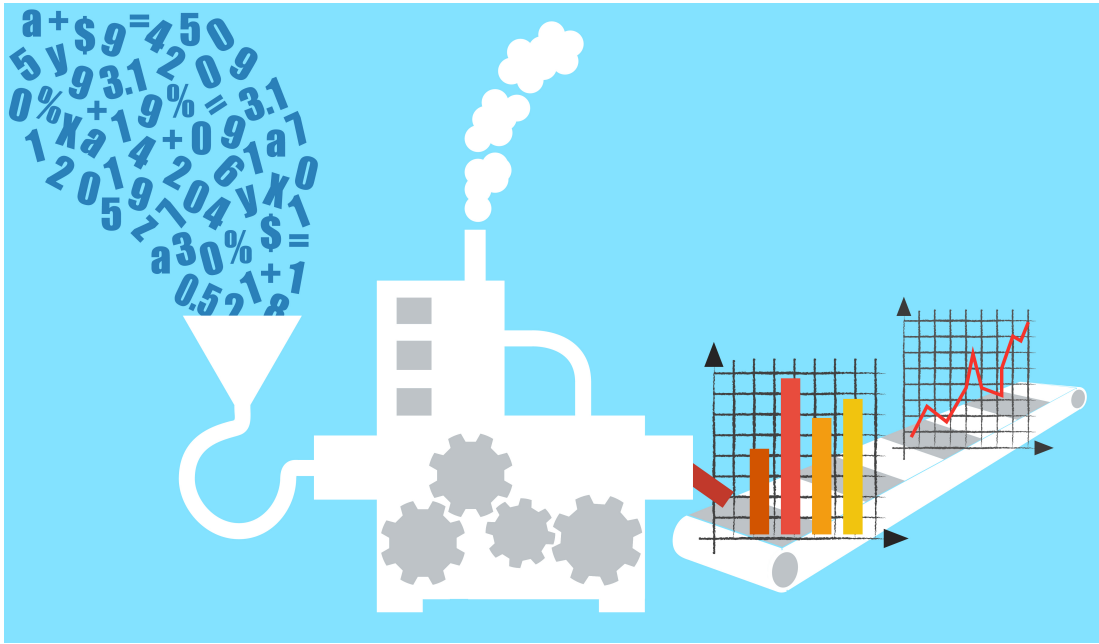
What is data?

Data never looks the same

- Pictures/Videos - e.g. Microscopy
- Ascii files - e.g. enrichment, gene sequences
- Curves - e.g. EPR, absorption spectra
- CSV

What do I have to do?

Processing



What do I have to do?

The usual steps

Understand the file

- Separators
- Symbols
- Missing values
- Decimal separator
- Anything else terrible?

Have a plan

- What do you want?
- How can you get it?
- Restructuring?
- Plotting?
- Storing and documentation

The Pandas package

Do you know R?

The most important commands

- `pd.read_csv` - load csv file
- `df.columns` - show column names
- `df.index` - show indices
- `df.loc['index', 'column']` - access via name
- `df.iloc[0, 1]` - access via index
- `df.T` - Transpose DataFrame

Assignments

CSV example

1. Calculate the mean, median, mode of the csv file (A,B,C)
2. Calculate the mean, median, mode grouped by classifiers
3. How many values are missing in the groups A,B, and C?
4. How many different values are in A,B, and C?

Interactive assignment

Elutriation data - cell sizes

- Experiment
- Data file
- Replicates
- Analysis
- Plotting....