

Data Analysis and Visualization

Course Introduction

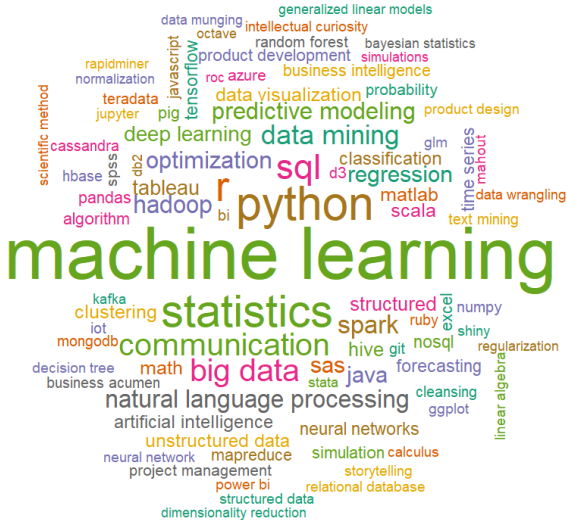
Part of Data Science program

1. Statistical programming with R
2. Multiple imputation in practice
3. **Data analysis and visualization (DAV)**

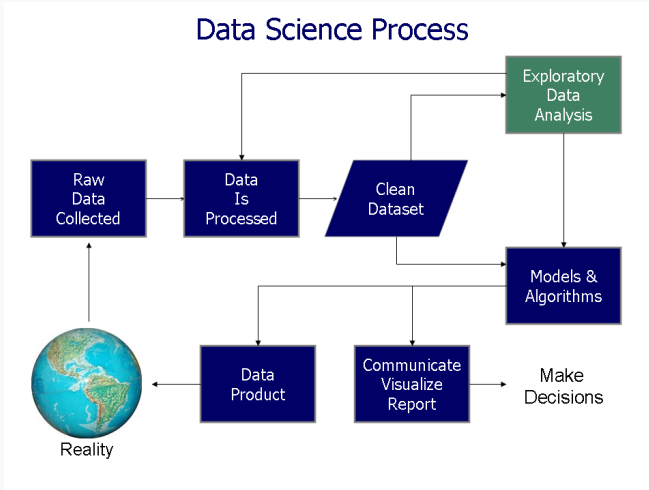
DAV course materials:

<https://github.com/MaartenCruyff/S31>

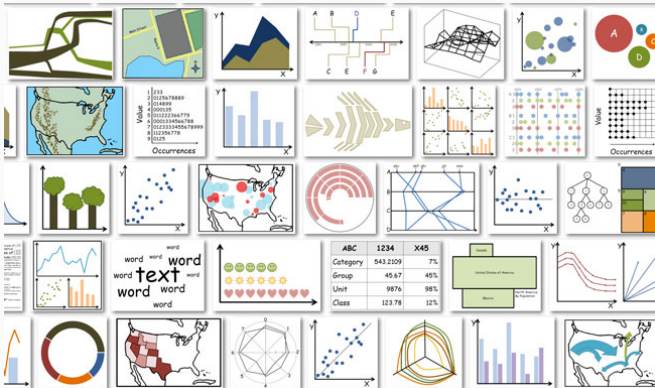
Data science word cloud



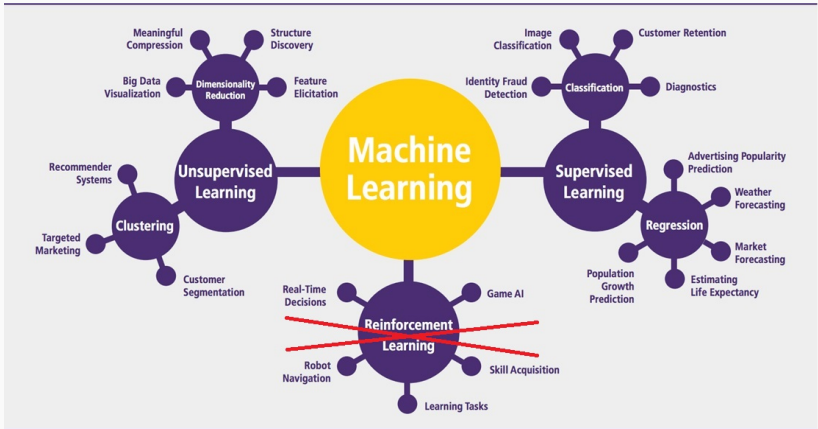
Data science process



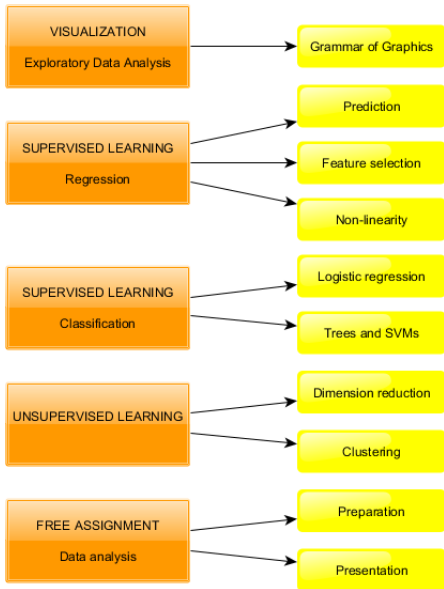
Visualizations



Models



Program



Visualization

1. Basics of grammar of graphics
2. Working with the R packages `dplyr` and `ggplot2`

Data analysis

1. Overview of models/techniques for statistical learning
2. Basic understanding the underlying algorithms
3. Ability to fit the models in R

Course is non-technical, emphasis on applications

Structure of the course

Day 1 to 4: Mornings and afternoon session

- Q&A previous lab (15 min.)
- introduction new topic (45 min.)
- lab session (2 hr)

Day 5: Presentations

- Groups of 3-6 students
- Perform data analysis (morning)
- Present results slideshow (afternoon)

Work environment

- R/RStudio
- work in scripts, preferably with Rmd files (template provided)
- render a HTML document with text and R in- and output
- no experience? Learn on the job, ask for assistance

Code folding

- R code for labs is made visible by clicking on the CODE symbol
- first try it yourself before peeking
- experiment with the code, try out other options

Hastie, Tibshirani and Friedman (2002). *The Elements of Statistical Learning*. Springer.

James, Witten, Hastie and Tibshirani (2009). *An Introduction to Statistical Learning with Applications in R*. Springer.

Wickham and Groleman (2017). *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. O'Reilly.