S31: Data Analysis

Course Introduction

Summerschool Data Science

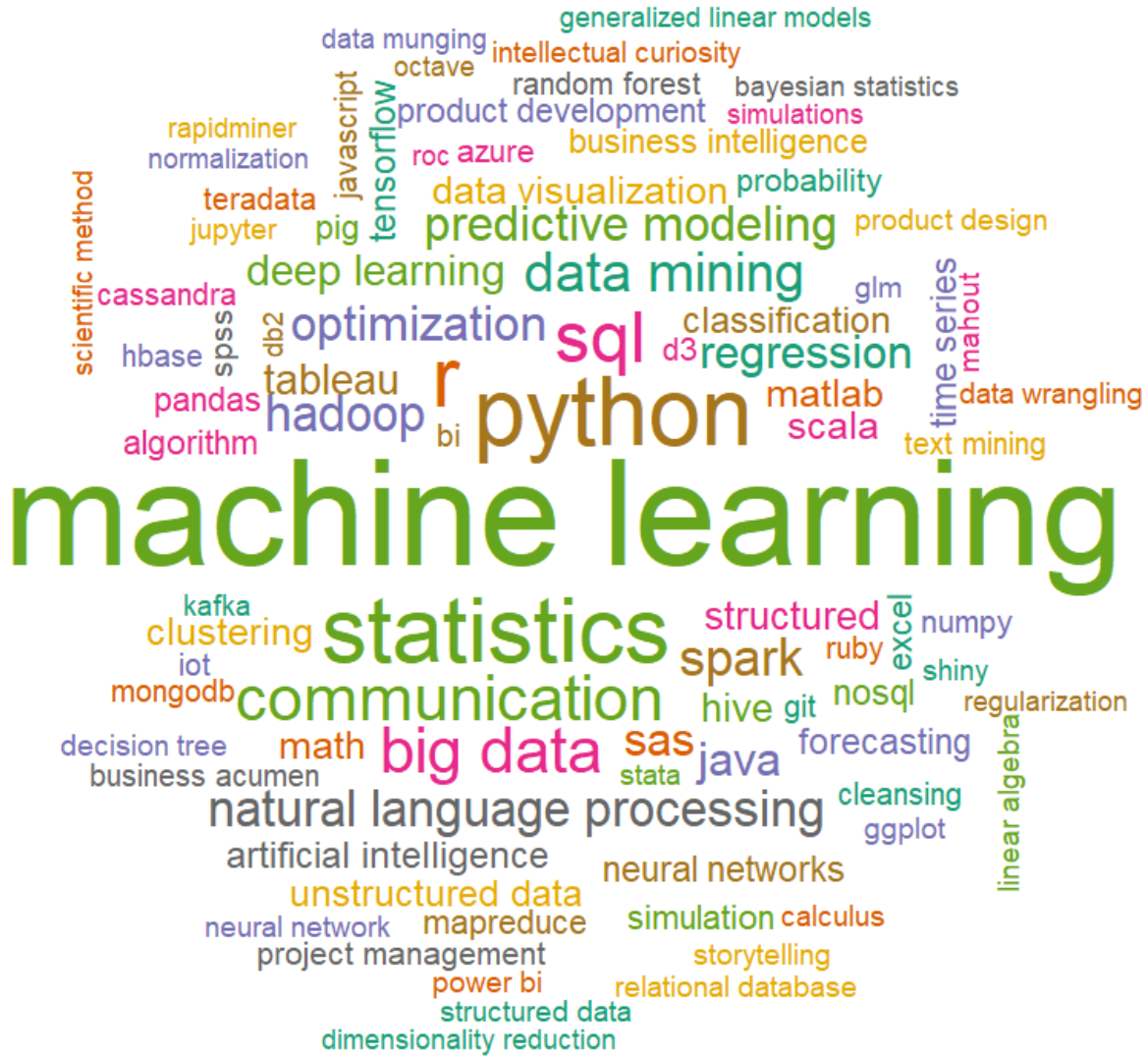
Part of Data Science program

- Statistical programming with R
- Multiple imputation in practice
- Introduction to text mining
- **Data analysis**
- Applied text mining

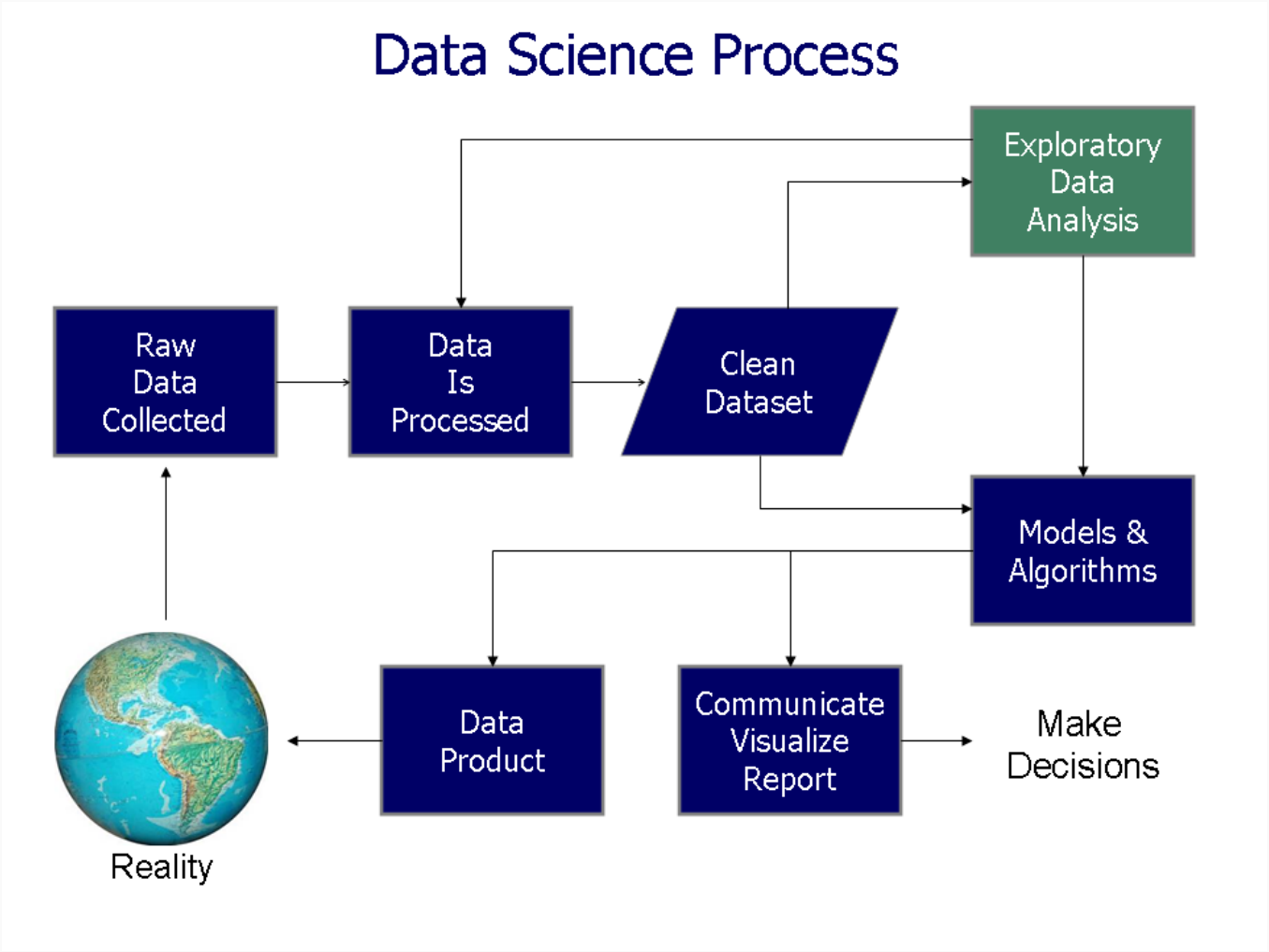
DAV course materials:

- links to slides, labs and literature in the course manual

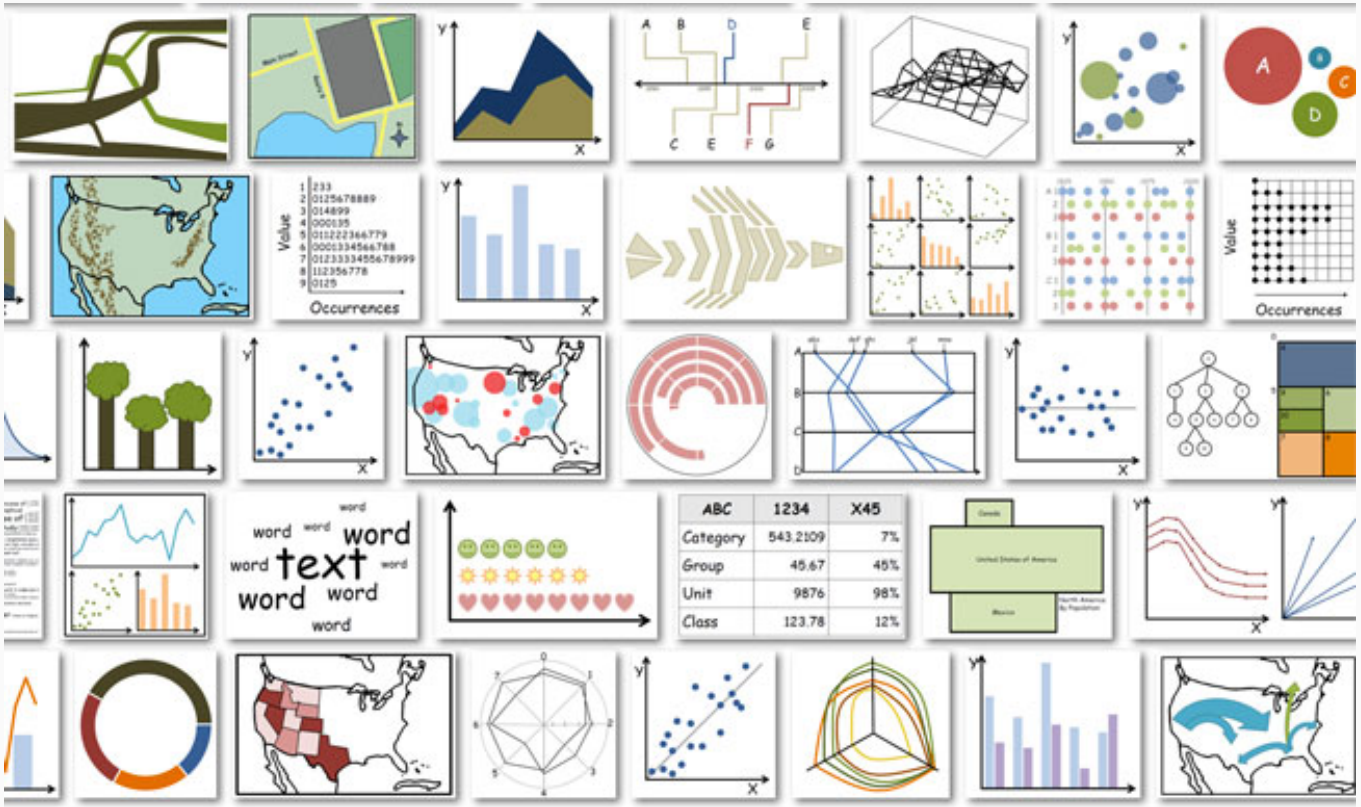
Data science word cloud



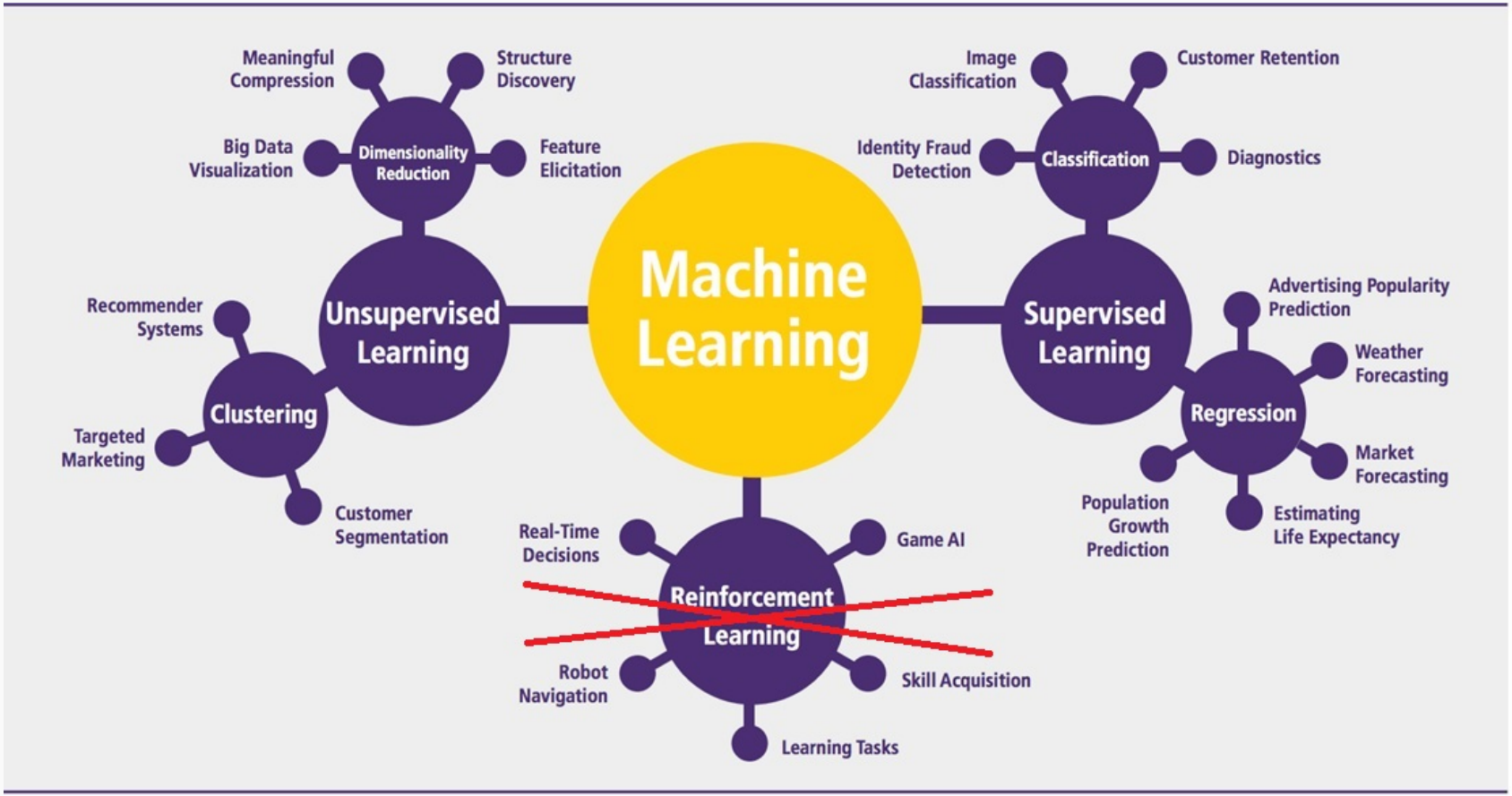
Data science process



Data visualizations



Models & Algorithms



Manual

S31: Data Analysis | Course description | Day 1 | Day 2 | Day 3 | Day 4 | Day 4

Day 1: Data science
- a. Data visualization
- b. Bias-variance trade-off

Day 2: Regression
- a. Feature space expansions
- b. Feature selection

Day 3: Classification
- a. Logistic regression
- b. Trees and SVMs

Day 4: Unsupervised learning
- a. PCA
- b. Clustering

Day 5: Presentations
- a. Prepare a brief slide show
- b. Presentations

Summerschool Utrecht Data Science course S31: Data Analysis

Content

The course Data Analysis is part of a series of Data Science courses offered by Summerschool Utrecht. The course offers a range of statistical techniques and algorithms from statistics, machine learning and data mining to make predictions about future events and to uncover hidden structures in data. The course has a strong practical focus; participants actively learn how to apply these techniques to real data and how to interpret their results. The course covers both classical and modern techniques of data analysis.

Structure

Morning session (9-12:30) and afternoon sessions (1:30-5) consisting of:

- slide presentation of new topic
- R lab to practice with new topic
- Q&A on the R lab

Software

The software needed for the R labs is freely available on the internet, and includes:

- R
- RStudio
- R packages, including the `tidyverse` package

Course materials

The following course materials are available via the links in this manual:

- slides
- R labs (HTML)
- R labs (R Markdown templates)

Aims

Data analysis

1. Basics of data visualization with `ggplot2`
2. Overview of models/techniques for statistical learning
3. Basic understanding the underlying algorithms
4. Ability to apply data analysis techniques on data

Course is non-technical, emphasis on applications

Structure of the course

Day 1 to 4: Morning and afternoon sessions

- introduction new topic (45 min.)
- R lab session (2 hr)
- Q&A R lab (45 min.)
- lunch from 12:30 to 1:30 pm

Day 5: Presentations

- Groups of 3-6 students
- Prepare slides of a data analysis (morning)
- Present results slideshow (afternoon)

Code folding

- Labs are HTML files
- R code can be made visible by clicking the `CODE` button
- try before peeking, and experiment with the code (try out other options)

The data and aesthetics arguments tell `ggplot()` where to to find the data, and where to map the variables. Together they specify the axes of the plot array, but they do not make any plot yet.

a. The first step in making plots the data specification with `ggplot(data = txhousing)`. This creates an empty plot surface.

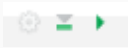
HIDE

```
ggplot(data = txhousing)
```

b. The next step is to add one or more aesthetics. We start by mapping the variable `volume` to the x-axis. Check the result.

CODE

Rmd templates

- Open the template in RStudio
- write your code in the R chunks
- test it by clicking the  button
- renders HTML document by clicking button

```
65
66  ## Data and aesthetics
67
68  The data and aesthetics arguments tell `ggplot()` where to to find the data, and where to map the
    variables. Together they specify the axes of the plot array, but they do not make any plot yet.
69
70  a. The first step in making plots the data specification with `ggplot(data = txhousing)`. This creates
    an empty plot surface.
71
72
73  ```{r}
74
75  ```
76
77  b. The next step is to add one or more aesthetics. We start by mapping the variable `volume` to the
    x-axis. Check the result.
78
79
80  ```{r}
81
82  ```
83
```