

S31: Data Analysis

Course Introduction

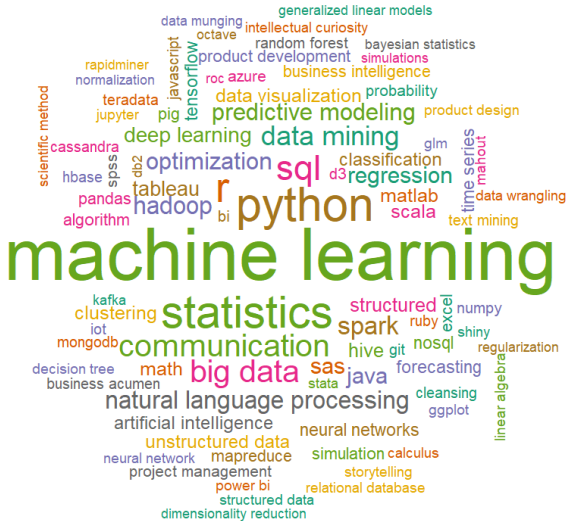
Part of Data Science program

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

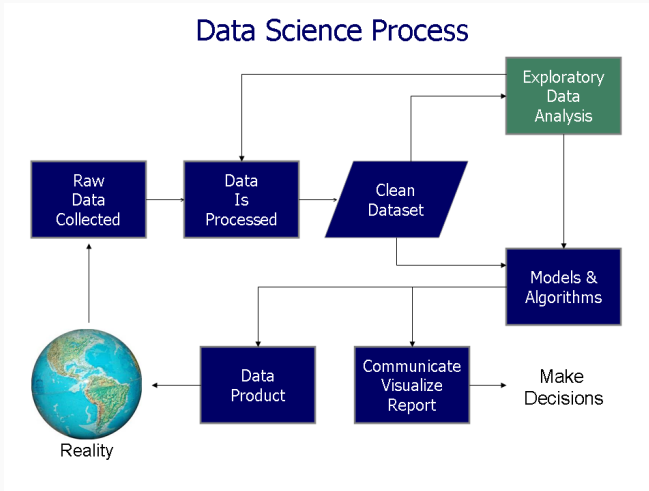
DAV course materials:

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

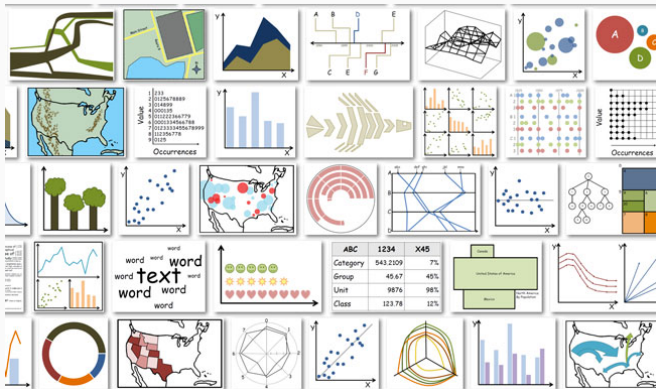
Data science word cloud



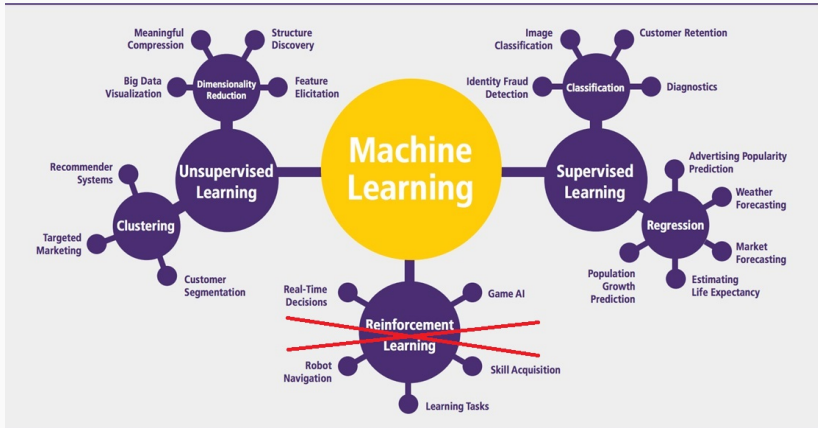
Data science process



Data visualizations



Models & Algorithms



S31: Data Analysis	Course description	Day 1	Day 2	Day 3	Day 4	Day 5
Day 1: Data science <ul style="list-style-type: none">a. Data visualizationb. Bias-variance trade-off Day 2: Regression <ul style="list-style-type: none">a. Feature space expansionsb. Feature selection Day 3: Classification <ul style="list-style-type: none">a. Logistic regression and LDAb. Trees and SVMs Day 4: Unsupervised learning <ul style="list-style-type: none">a. PCAb. Clustering Day 5: Presentations <ul style="list-style-type: none">a. Prepare a brief slide showb. Presentations	<p>Summerschool Utrecht Data Science course S31: Data Analysis</p> <p>Content</p> <p>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.</p> <p>Structure</p> <p>Morning session (9:00-12:15) and afternoon sessions (13:45-17:00) consisting of:</p> <ul style="list-style-type: none">▪ slide presentation of new topic▪ R lab to practice with new topic▪ Q&A on the R lab <p>Software</p> <p>The software needed for the R labs is freely available on the internet, and includes:</p> <ul style="list-style-type: none">▪ R▪ RStudio▪ R packages, including the <code>tidyverse</code> package <p>Make sure to you have the latest versions of R and RStudio installed, and that you regularly update your packages!</p> <p>Course materials</p> <p>The following course materials are available via the links in this manual:</p> <ul style="list-style-type: none">▪ slides▪ R labs (HTML)▪ R labs (R Markdown templates) <p>The R labs include the exercises, and the R code to do these exercises. The R code is hidden, but can be made visible by clicking the <code>code</code> button. The recommended way to perform the analyses is to first try to write the code yourself, and only when you get stuck to use the <code>code</code> button.</p> <p>The R Markdown templates include the text of the R labs and empty <code>R</code> chunks that can be used to do the exercises. When finished, clicking the <code>knit</code> button renders the original HTML, including all R code and output. It is highly recommended to use the <code>Rmd</code> lab templates to make the lab exercises in.</p> <p>References</p> <ul style="list-style-type: none">▪ James, Witten, Hastie & Tibshirani. (2013). <i>An Introduction to Statistical Learning with Applications in R (ISLR)</i>, 1st ed. New York: Springer▪ Golemund and Wickham (2016). <i>R for Data Science</i>					

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 (30 min.)
- lunch from 12:15 to 1:45 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 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  Knit button

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