Data Analysis and Visualization

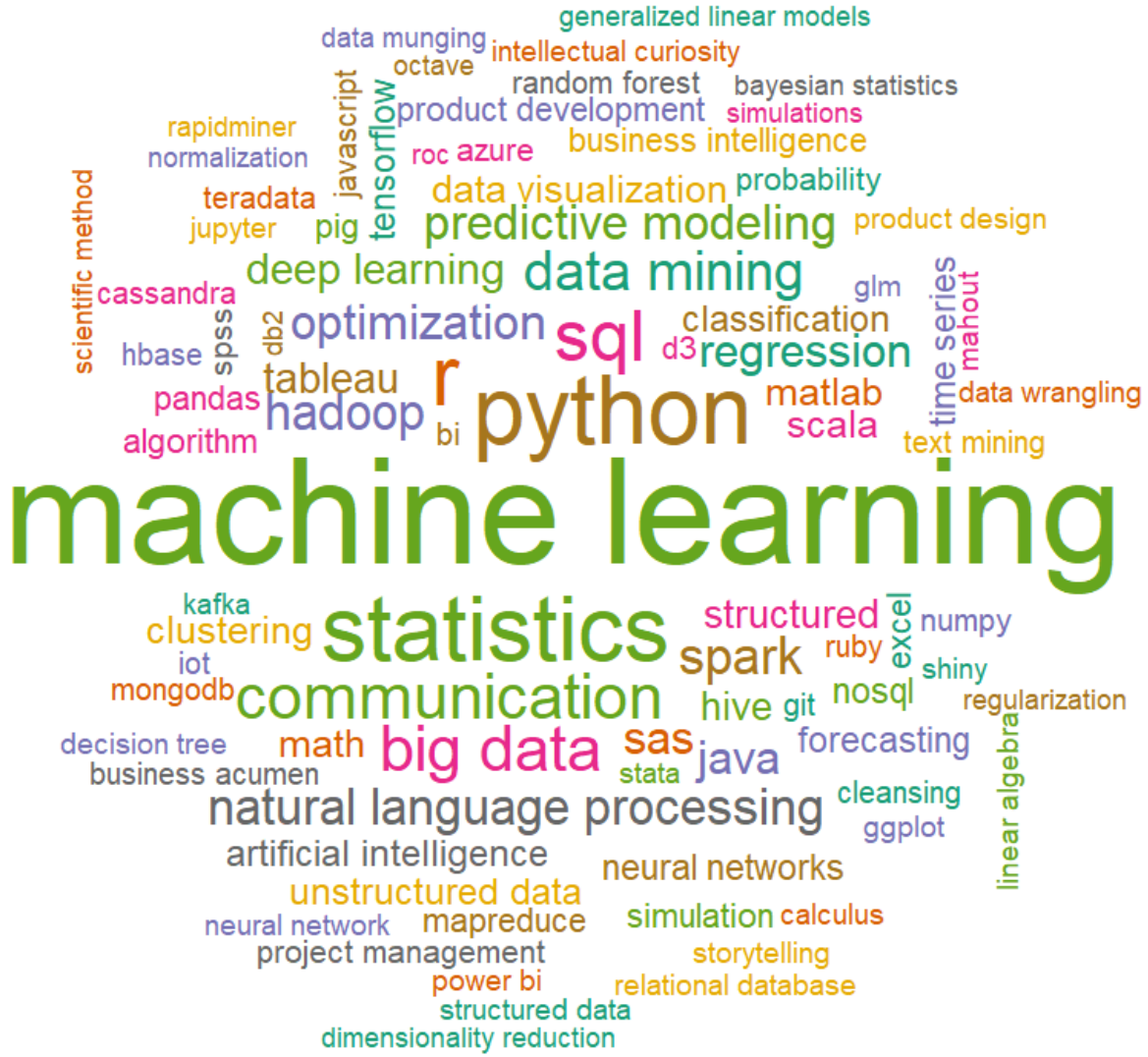
Course Introduction

Summerschool Data Science

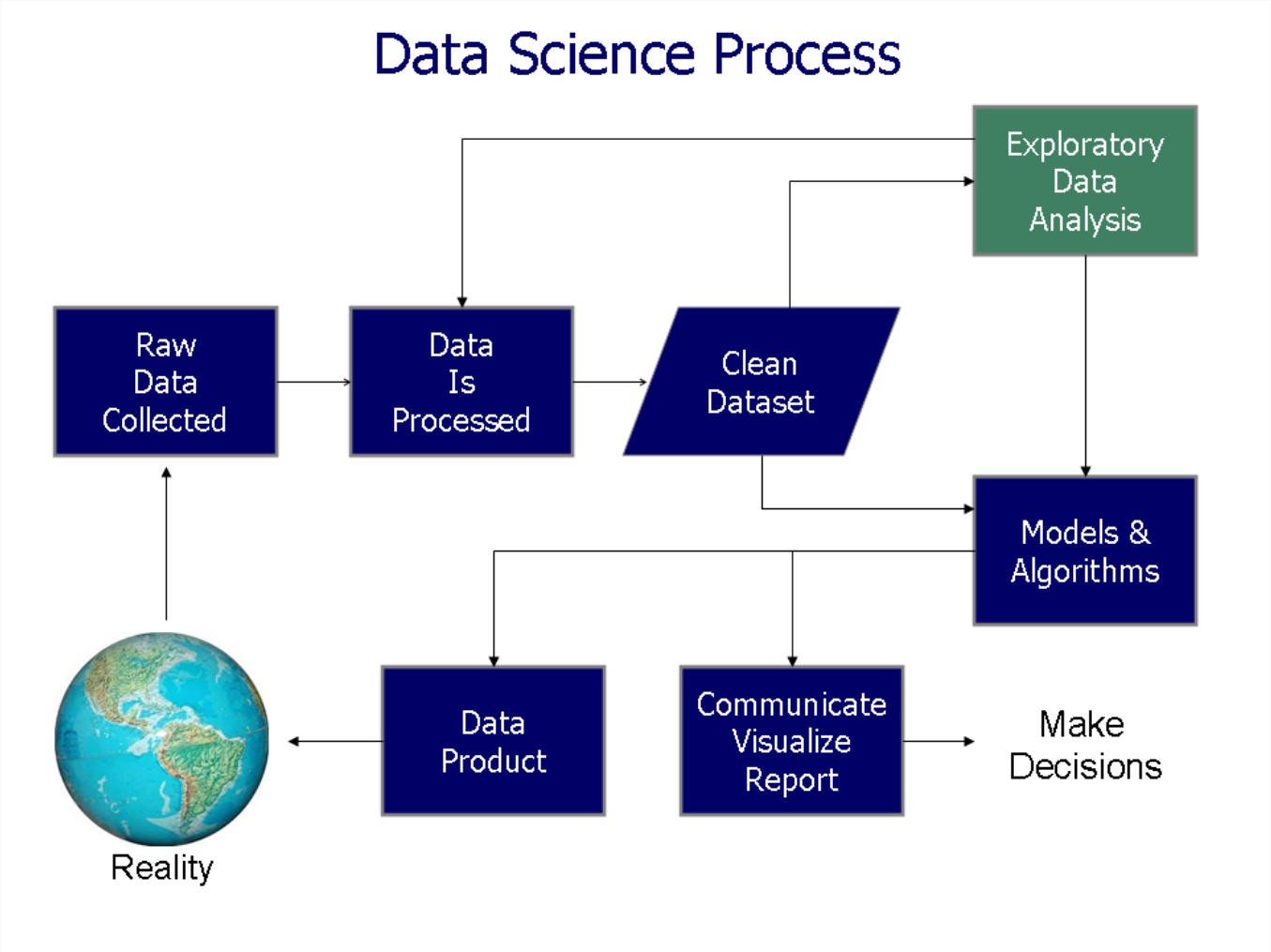
Part of Data Science program

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

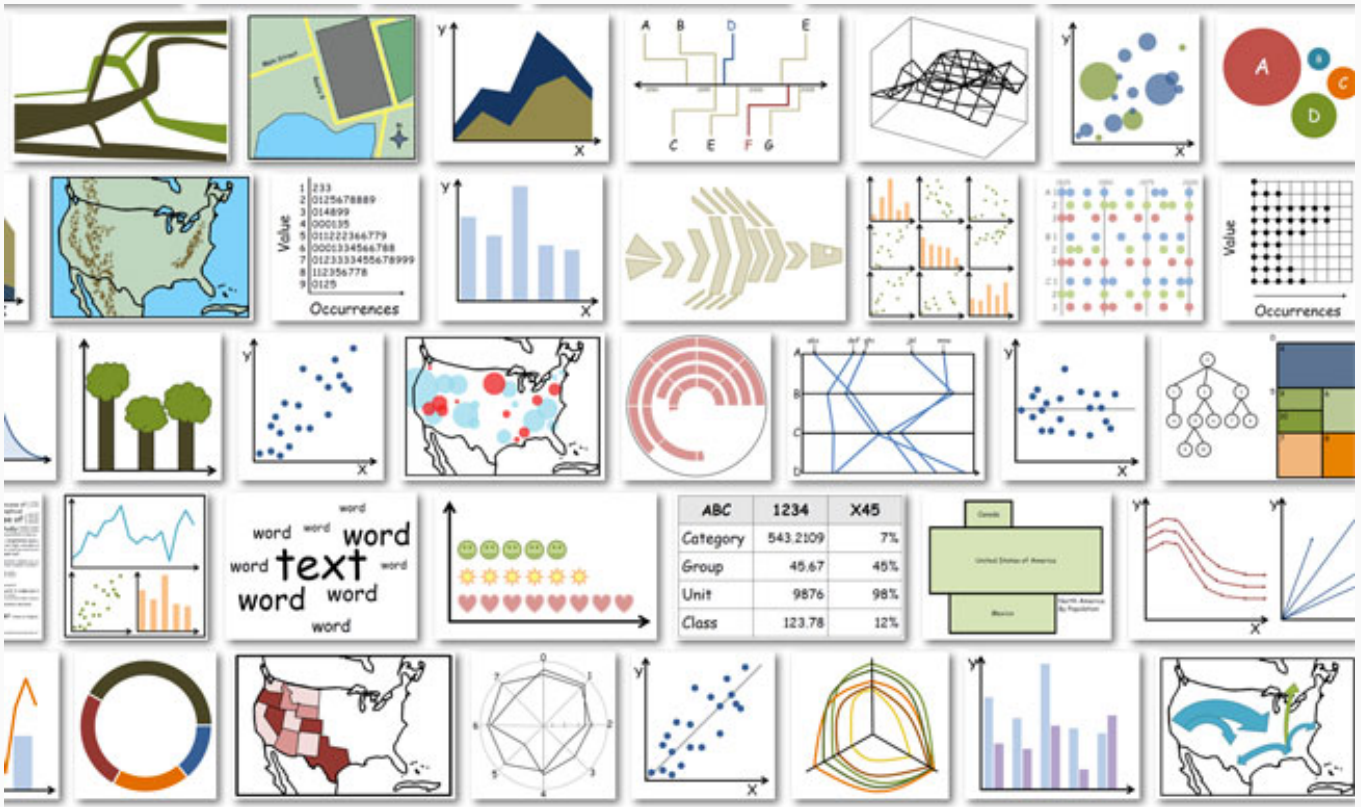
Data science word cloud



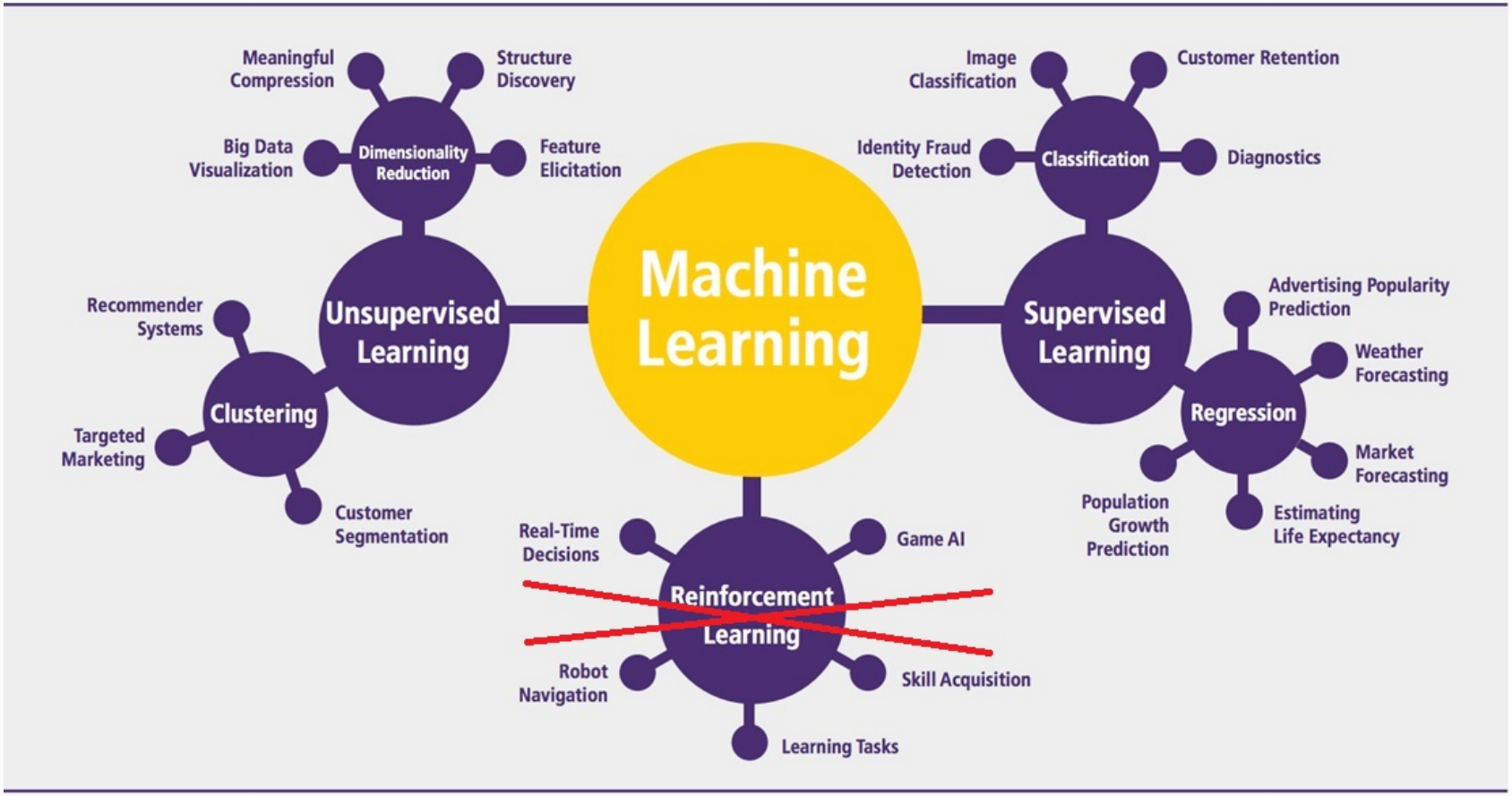
Data science process



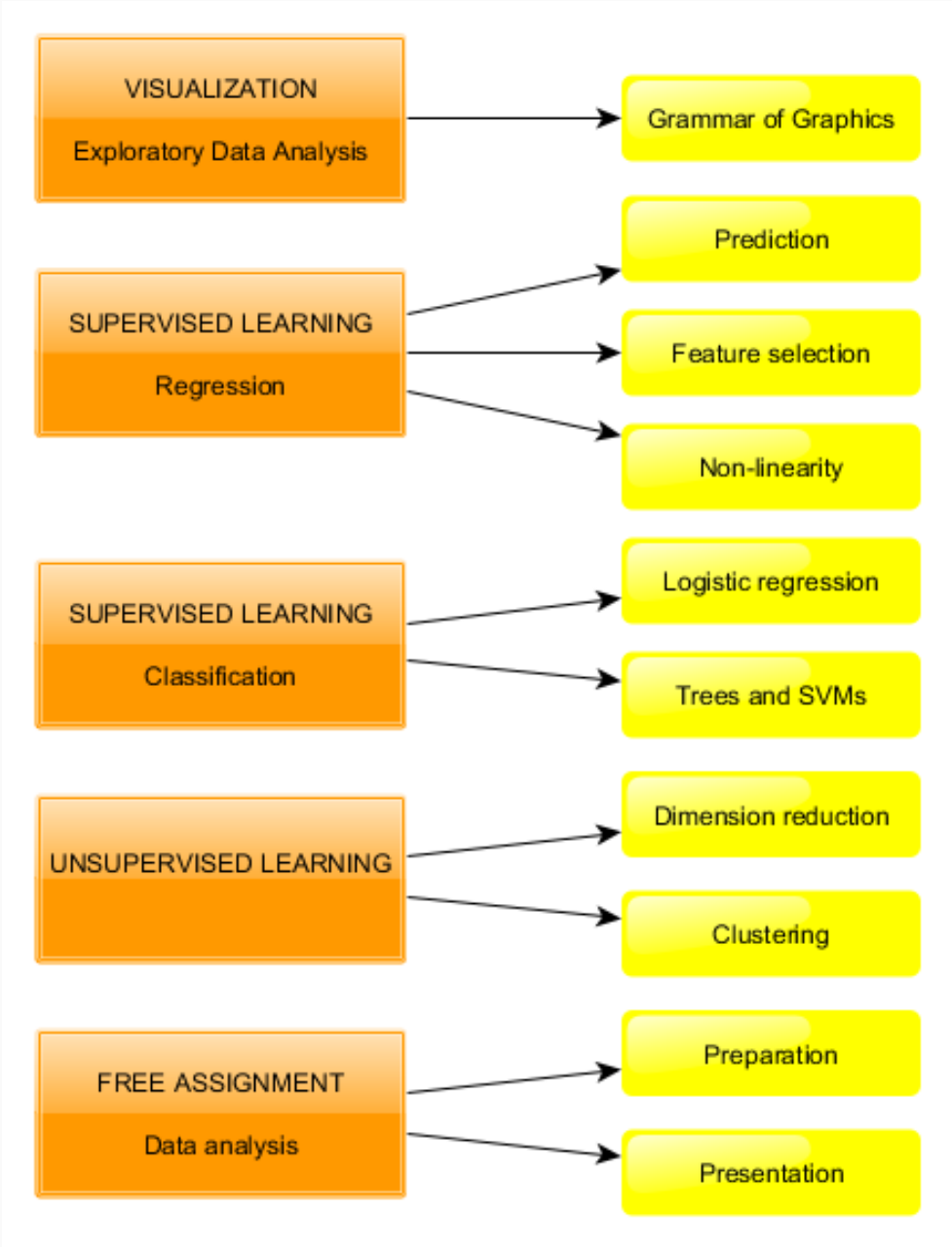
Visualizations



Models


Machine Learning
Unsupervised Learning
Dimensionality Reduction
Meaningful Compression
Structure Discovery
Big Data Visualization
Feature Elicitation
Clustering
Recommender Systems
Targeted Marketing
Customer Segmentation
Supervised Learning
Classification
Image Classification
Customer Retention
Identity Fraud Detection
Diagnostics
Regression
Advertising Popularity Prediction
Weather Forecasting
Market Forecasting
Estimating Life Expectancy
Population Growth Prediction
Reinforcement Learning
Real-Time Decisions
Game AI
Robot Navigation
Skill Acquisition
Learning Tasks

Program


VISUALIZATION
Exploratory Data Analysis
Grammar of Graphics
SUPERVISED LEARNING
Regression
Prediction
Feature selection
Non-linearity
SUPERVISED LEARNING
Classification
Logistic regression
Trees and SVMs
UNSUPERVISED LEARNING
Dimension reduction
Clustering
FREE ASSIGNMENT
Data analysis
Preparation
Presentation

Aims DAV

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)

Labs

Work environment

- R/RStudio
- work in scripts, prefebly 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

Literature

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.