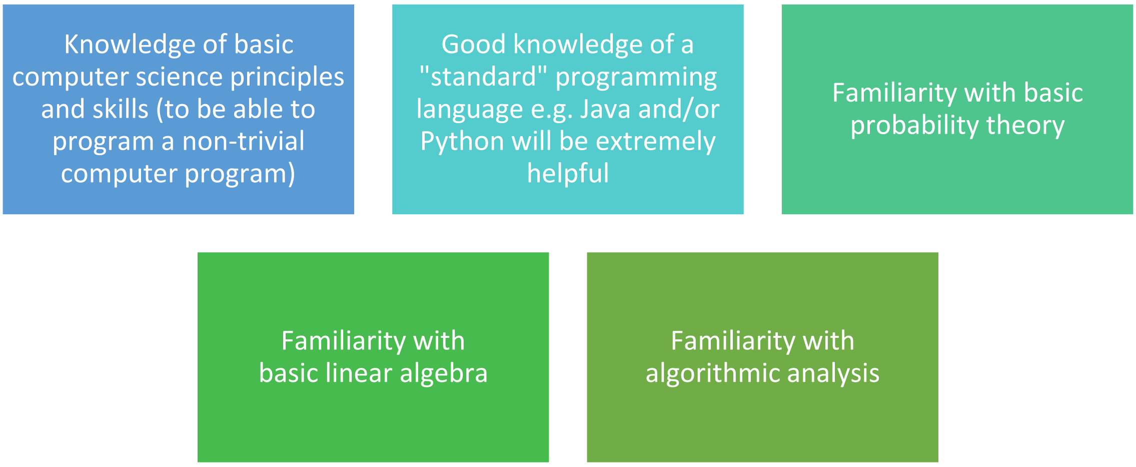
Data Analytics and Big Data

Description

The course discusses data mining and machine learning algorithms with the emphasis on analysing large amounts of data (Big Data). Starting by describing the data mining pipeline and basic machine learning algorithms, this course will later elaborate on tools for creating parallel algorithms that can process very large amounts of data.

Topics include: CRISP-DM standard process model for data miming, data acquisition and preprocessing techniques, model evaluation techniques, decision trees, frequent item sets and association rules, nearest neighbour search (for high dimensional data), classification rules, locality sensitive hashing (LSH), various dimensionality reduction strategies, recommender systems, clustering, link analysis, data streams, web and text mining.

Prerequisites



Knowledge of basic computer science principles and skills (to be able to program a non-trivial computer program)

Good knowledge of a "standard" programming language e.g. Java and/or Python will be extremely helpful

Familiarity with basic probability theory

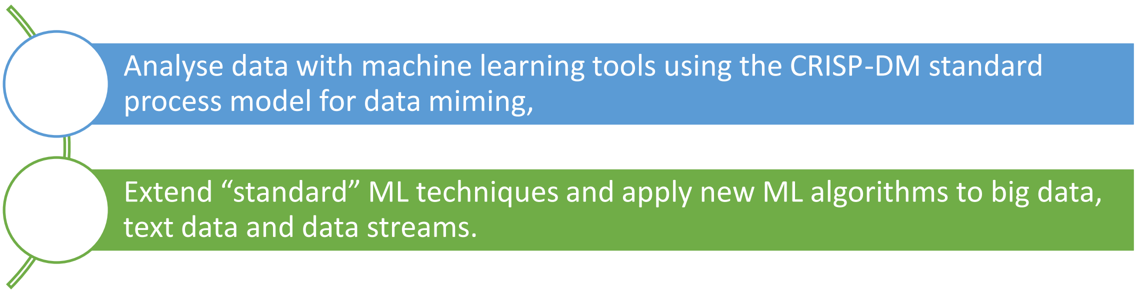
Familiarity with basic linear algebra

Familiarity with algorithmic analysis

Contents

The 6 phases of the CRISP-DM standard process model for data miming (problem understanding, data understanding, data preprocessing, modelling, evaluation, deployment),
Data acquisition, understanding, preprocessing and the principles of ETL (Extract, Transform, Load),
Understanding and using the “standard” machine learning (ML) algorithms (majority classifier, naïve Bayes, decision trees and rules, k-NN, clustering, frequent item sets and association rules, (non)linear and logistic regression, regression trees, random forests, support vector machines (SVMs), ...),
Evaluating the outcomes of learning (model error and accuracy, confusion matrix, ROC curve and AUC, train, test and validation sets, random data splits, cross-validation, the bootstrap principle, cost-sensitive evaluation),
Extending classical ML to handle large amounts of data (Big Data) (sampling techniques, locality sensitive hashing (LSH), dimensionality reduction, recommender systems, clustering big data, link analysis, data streams, web and text mining).

Expected learning outcomes



Analyse data with machine learning tools using the CRISP-DM standard process model for data miming,

Extend “standard” ML techniques and apply new ML algorithms to big data, text data and data streams.