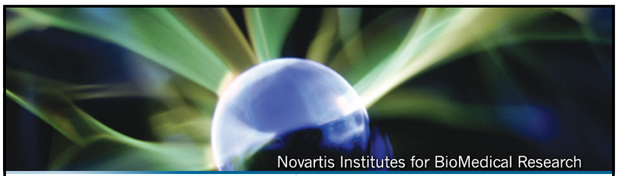
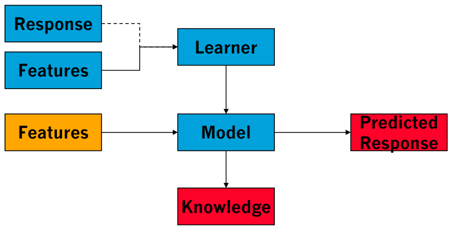


Data mining / Statistical Learning

Peter Gedeck
CADD/GDC
Novartis Institutes for BioMedical Research
NHRC, Horsham, UK

Data mining / Statistical Learning

- The nontrivial extraction of implicit, previously unknown, and potentially useful information from data
- It's about learning from data, understanding data, extracting knowledge and applying the knowledge


Response
Features
Learner
Features
Model
Predicted Response
Knowledge

Data mining / Statistical Learning

- Types of models
  - Regression
  - Classification
  - Dimensionality reduction
  - Clustering
- Type of learning
  - Supervised Learning
  - Unsupervised Learning

Regression

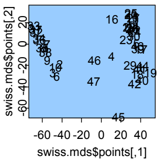
- A typical scenario:
  - You've made a couple of chemical compounds and measure a physicochemical property, e.g. solubility.
  - The measurement is very time-consuming, however you can describe the compounds using more easily accessible *features*
  - You would like to be able to predict the solubility of new compounds without doing the measurement
  - Luckily we have the set of compounds for which we know the solubility and the features
  - Using this dataset we can build a *prediction model*
- This is a regression problem
  - The response is a continuous variable
  - In this course, we will focus on regression

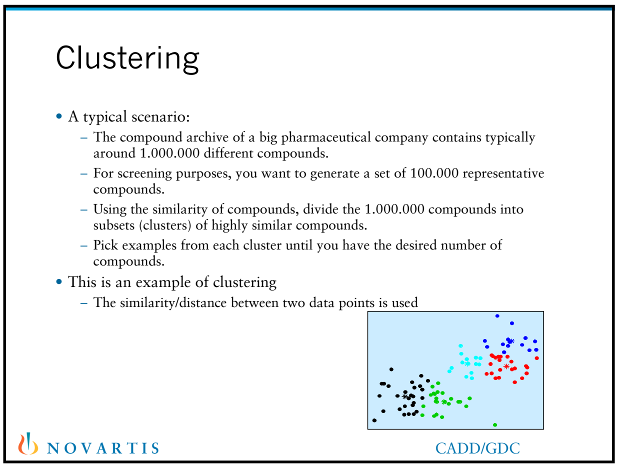
Classification

- A typical scenario:
  - For High-throughput screening, a pharmaceutical company wants to purchase 100.000 compounds.
  - It is known that drugs require specific properties that not all compounds have.
  - It is however not trivial to identify simple rules.
  - Generate a dataset of know drugs and a dataset of non-drugs (e.g. building blocks).
  - By comparing the properties of drugs and non-drugs, it is possible to build a statistical model.
  - Apply the model prior to purchasing new compounds to increase the chance of purchasing drug-like compounds.
- This is a classification problem
  - The response is a class membership

Dimensionality reduction

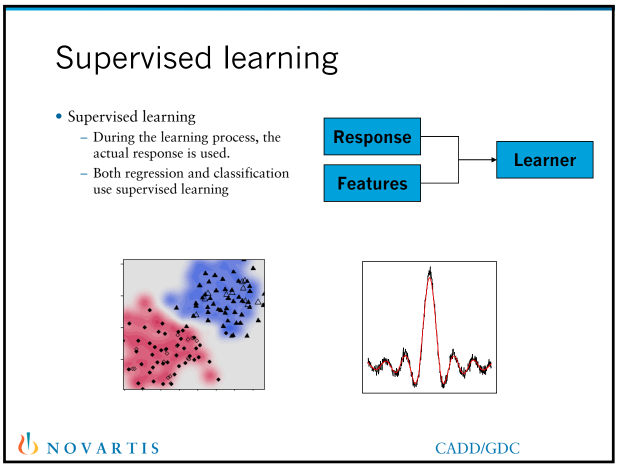
- A typical scenario:
  - You have a set of compounds and you want to visualise the compounds in a *simple* graph.
  - There is no obvious two-dimensional description of the dataset.
  - It is however possible to determine a similarity between two structures.
  - Using a multi-dimensional scaling, generate a two-dimensional model of your data.
- This is an example of dimensionality reduction
  - The similarity/distance between two data points is used





Clustering

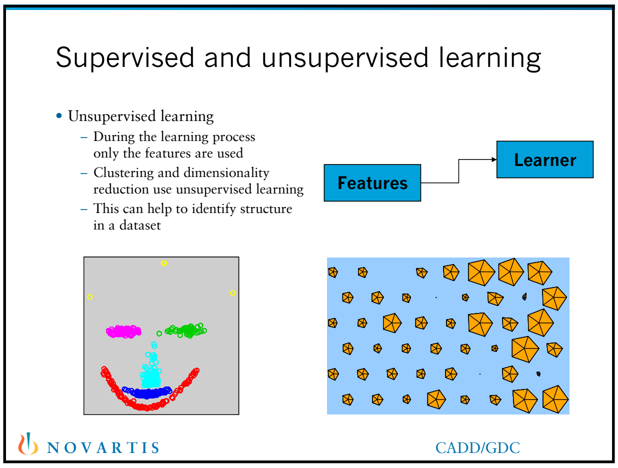
- A typical scenario:
  - The compound archive of a big pharmaceutical company contains typically around 1.000.000 different compounds.
  - For screening purposes, you want to generate a set of 100.000 representative compounds.
  - Using the similarity of compounds, divide the 1.000.000 compounds into subsets (clusters) of highly similar compounds.
  - Pick examples from each cluster until you have the desired number of compounds.
- This is an example of clustering
  - The similarity/distance between two data points is used



Supervised learning

- Supervised learning
  - During the learning process, the actual response is used.
  - Both regression and classification use supervised learning

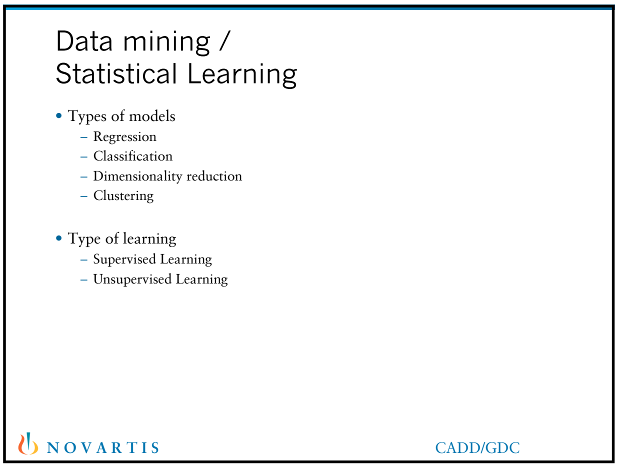
Response
Features
Learner



Supervised and unsupervised learning

- Unsupervised learning
  - During the learning process only the features are used
  - Clustering and dimensionality reduction use unsupervised learning
  - This can help to identify structure in a dataset

Features
Learner



Data mining / Statistical Learning

- Types of models
  - Regression
  - Classification
  - Dimensionality reduction
  - Clustering
- Type of learning
  - Supervised Learning
  - Unsupervised Learning