Business Intelligence & Process Modelling

Frank Takes

Universiteit Leiden

Lecture 4 — BI & Predictive Analytics

Recap

- **Business Intelligence**: anything that aims at providing actionable information that can be used to support business decision making
  - Business Intelligence
  - Visual Analytics
  - **Descriptive Analytics**
  - **Predictive Analytics**
- Process Modelling (April and May)

Categories of techniques

- Machine learning
  - Supervised learning: learning on labeled data
  - Semi-supervised learning: partially labeled data
  - Unsupervised learning: leaning/mining on unlabeled data
  - Reinforcement learning: agents learning to act in an environment

Universiteit Leiden

Supervised learning

Supervised learning

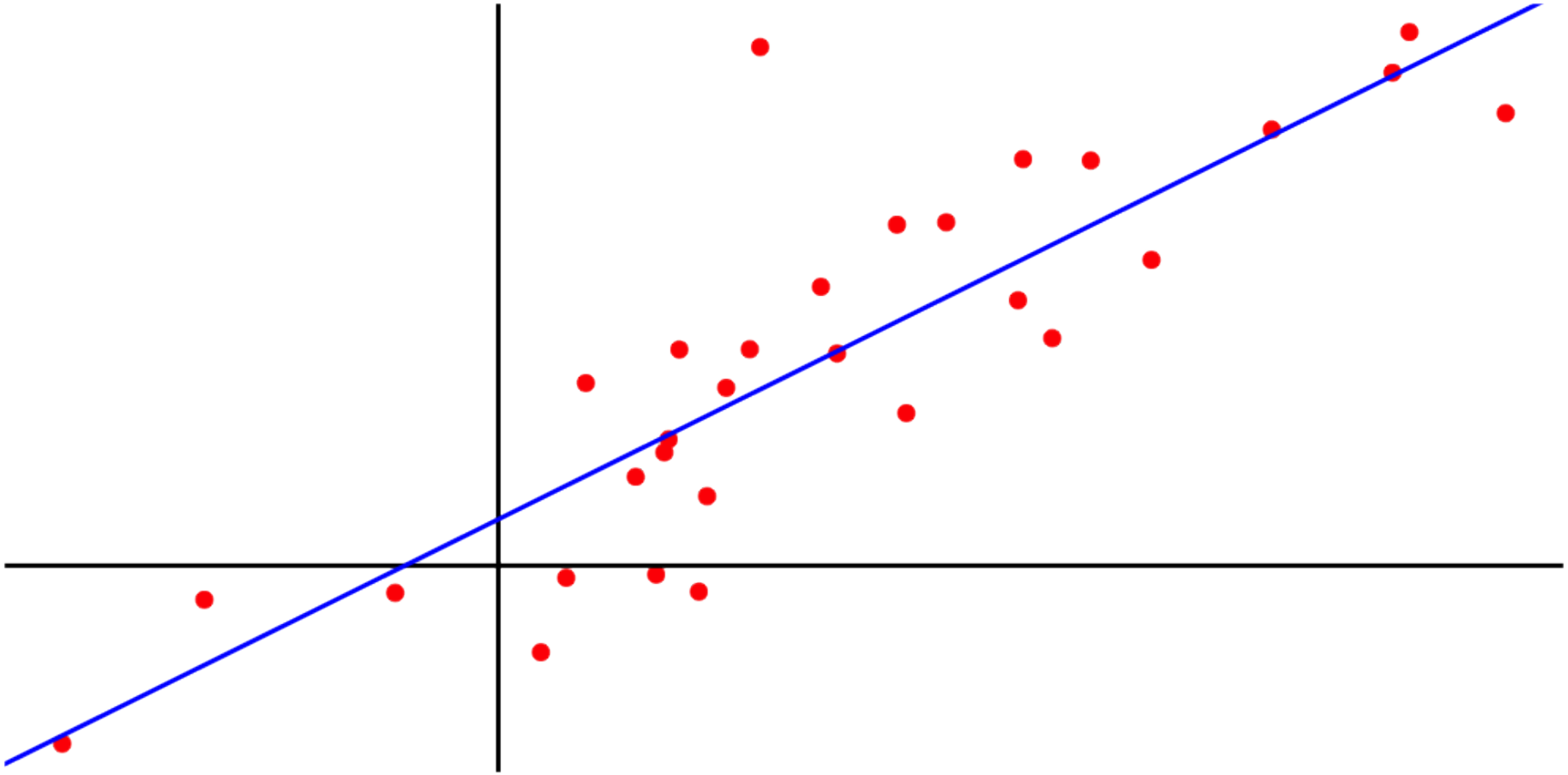
- Regression
- Classification
- Bayesian Networks
- Support Vector Machines
- Link prediction

Example dataset

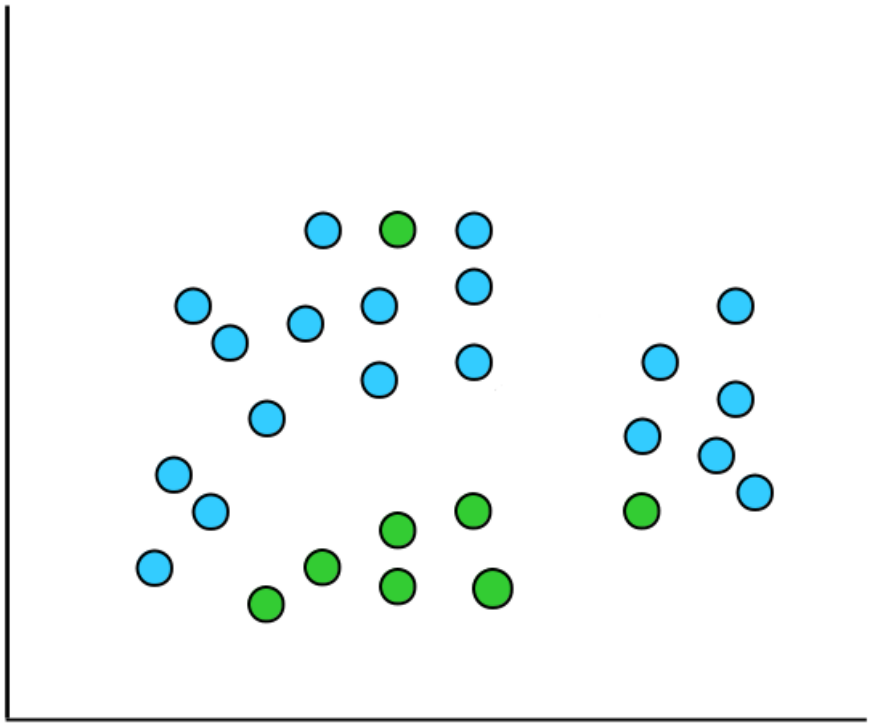
- 2 attributes and a **Class** attribute
- 50 datapoints

x	y	**Class**
2	3	Blue
3	2	Green
3	4	Blue
...	...	...

Regression as a model



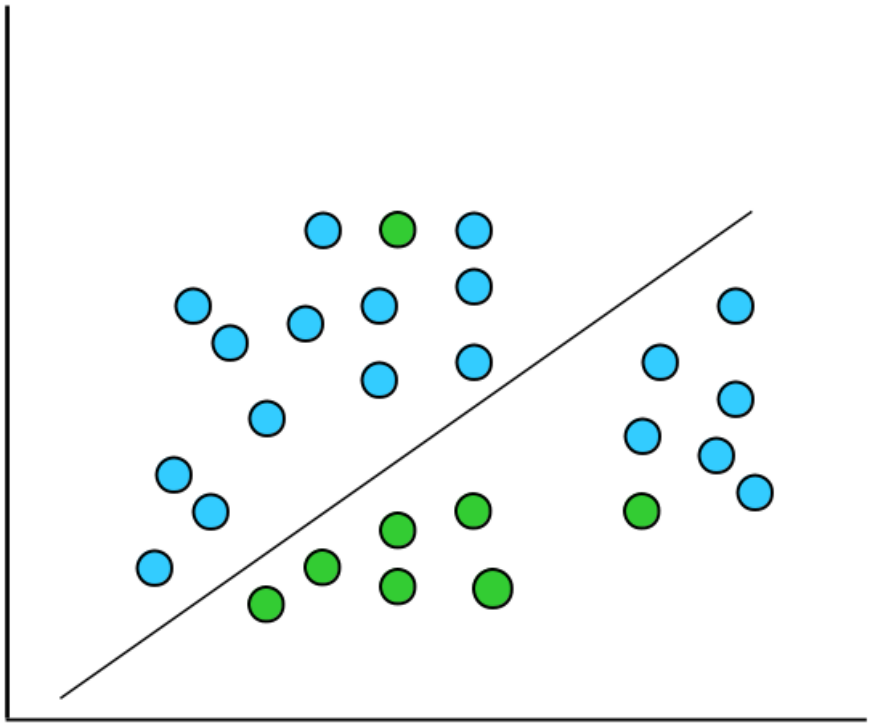
Classification: Regression



Linear Regression

- Given n variables $x_1, \ldots x_n$
- Find weights $w_0, \ldots w_n$ such that
  $w_0 + w_1 x_1 + \ldots w_n x_n \geq 0$

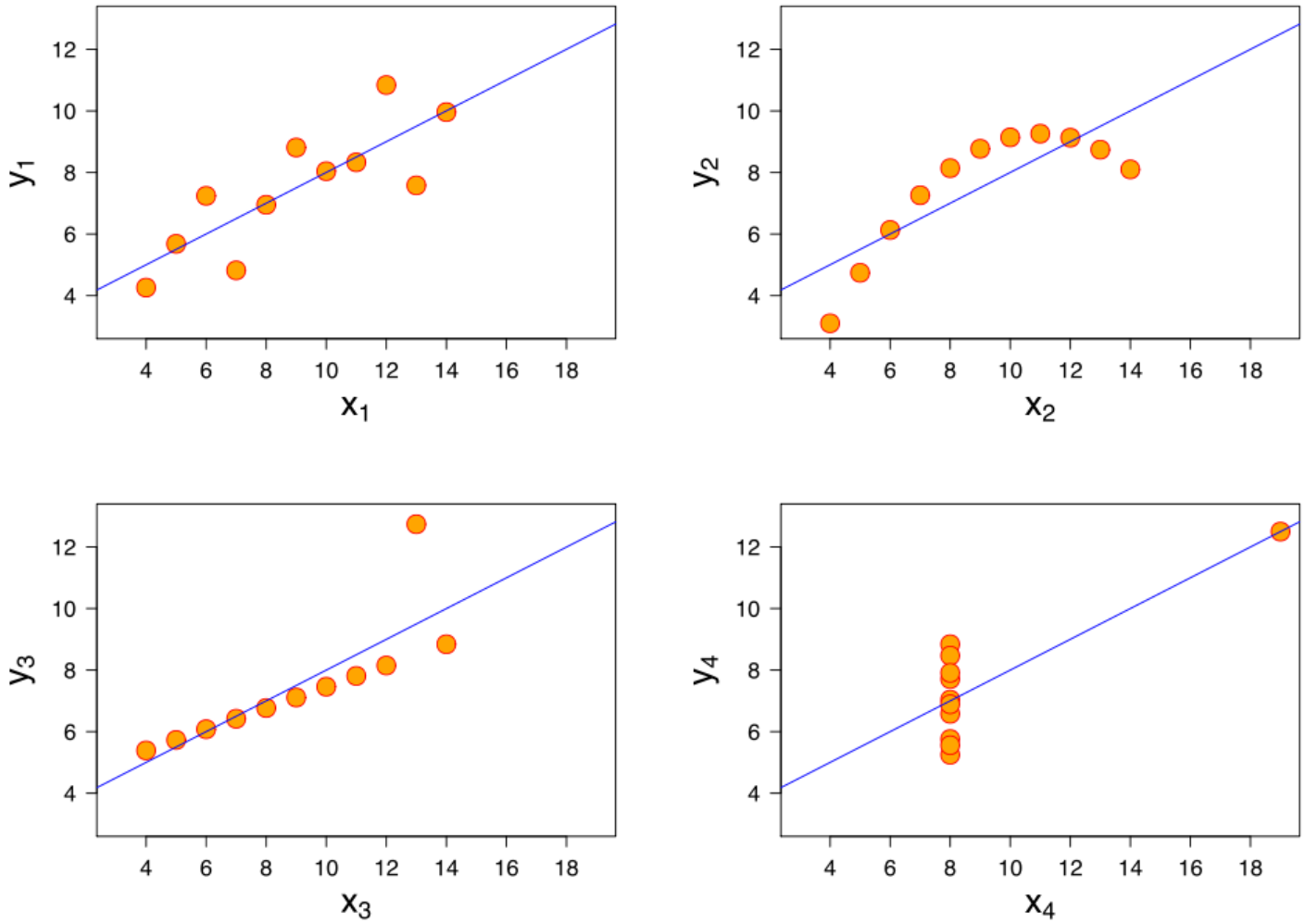
Classification: Regression



Linear Regression

- Given n variables $x_1, \ldots x_n$
- Find weights $w_0, \ldots w_n$ such that
  $w_0 + w_1 x_1 + \ldots w_n x_n \geq 0$
- Example: $n = 2$
  $w_0 + w_1 x + w_2 y \geq 0$

Regression disclaimer



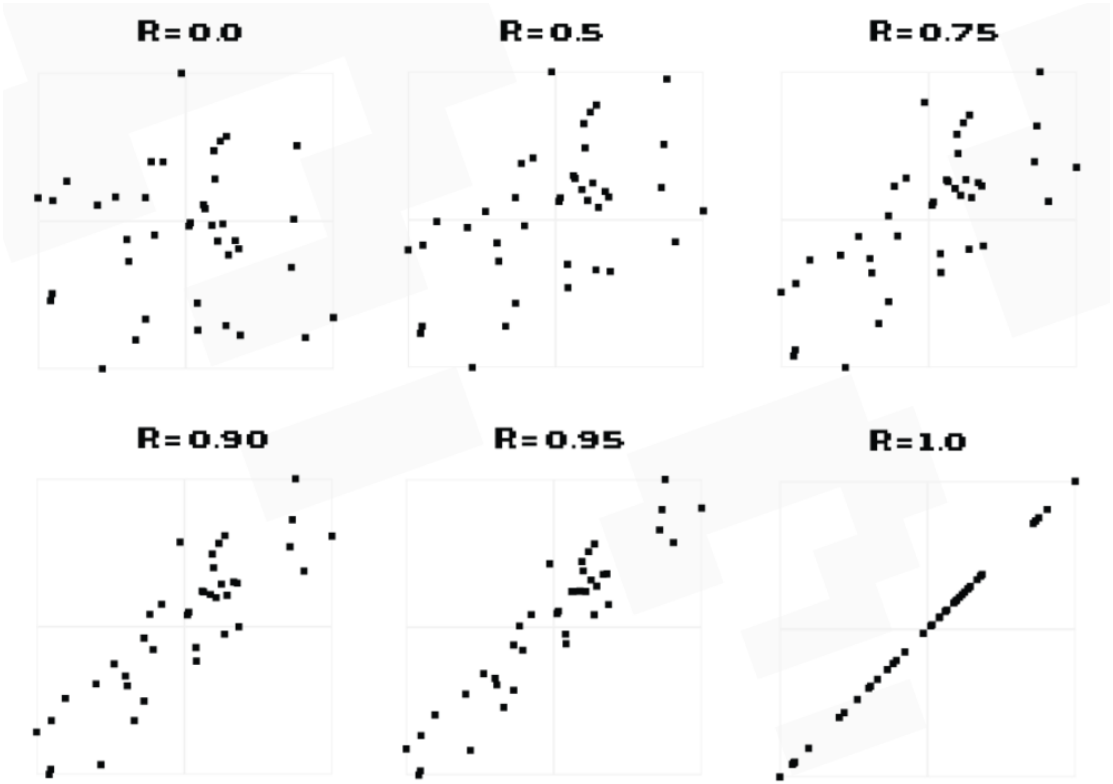
http://en.wikipedia.org/wiki/Linear_regression

Correlation

- Pearson correlation $r \in [0; 1]$ describing the extent to which the relation between variables can be described in a linear way.

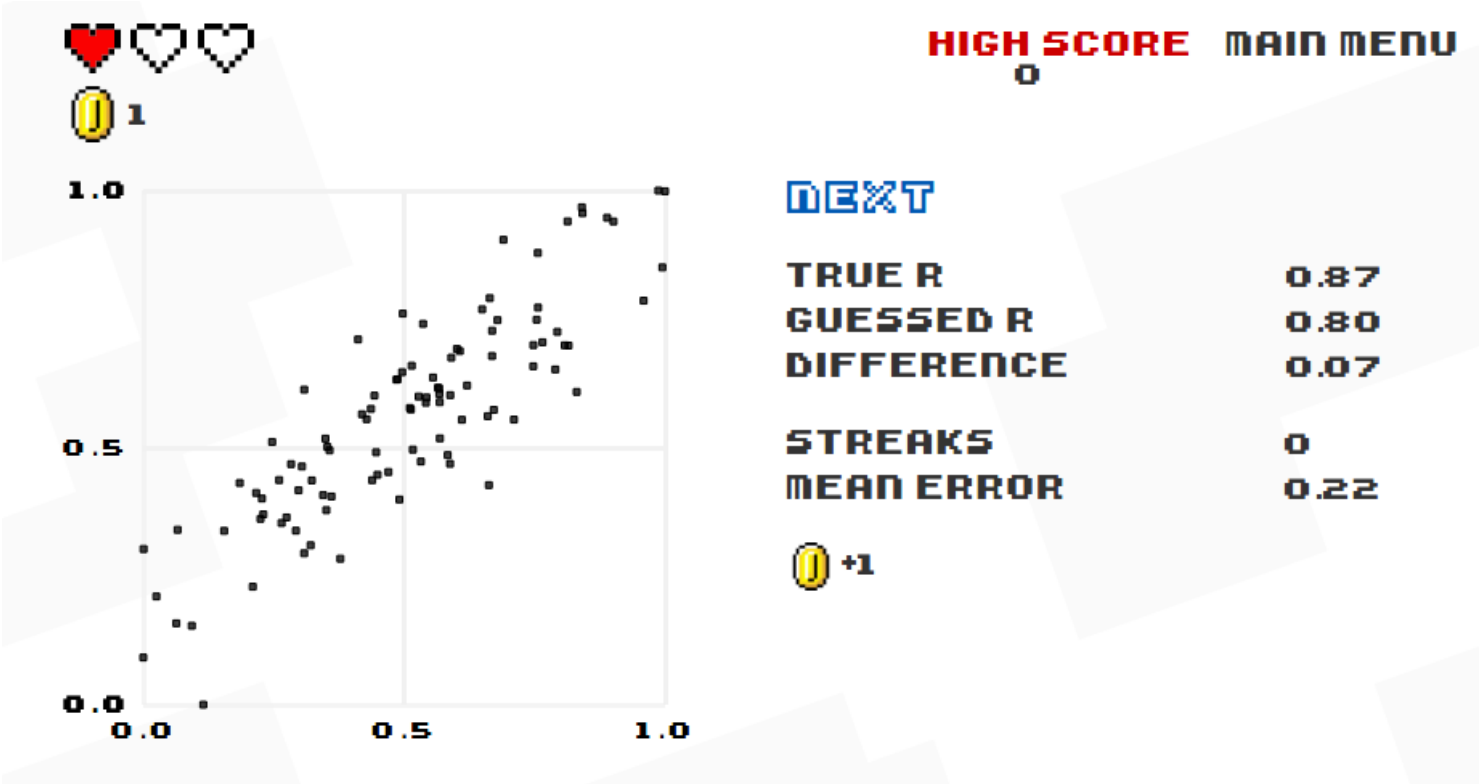
Correlation

- Pearson correlation $r \in [0; 1]$ describing the extent to which the relation between variables can be described in a linear way.

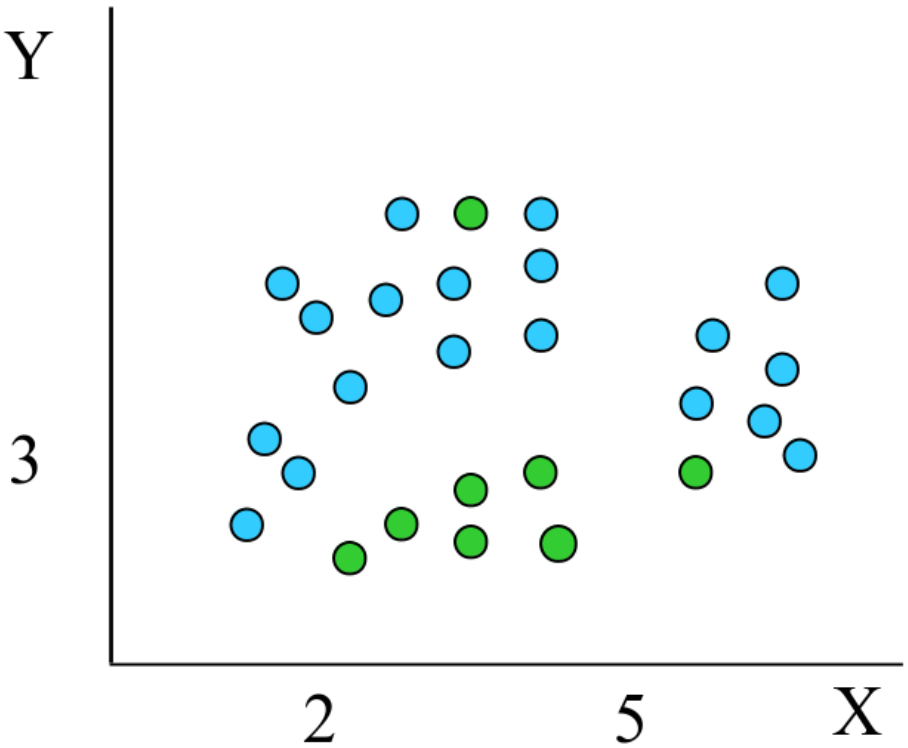


Correlation

- How do we perceive correlations?
- Study by University of Cambridge — Gamification
- `http://guessthecorrelation.com`



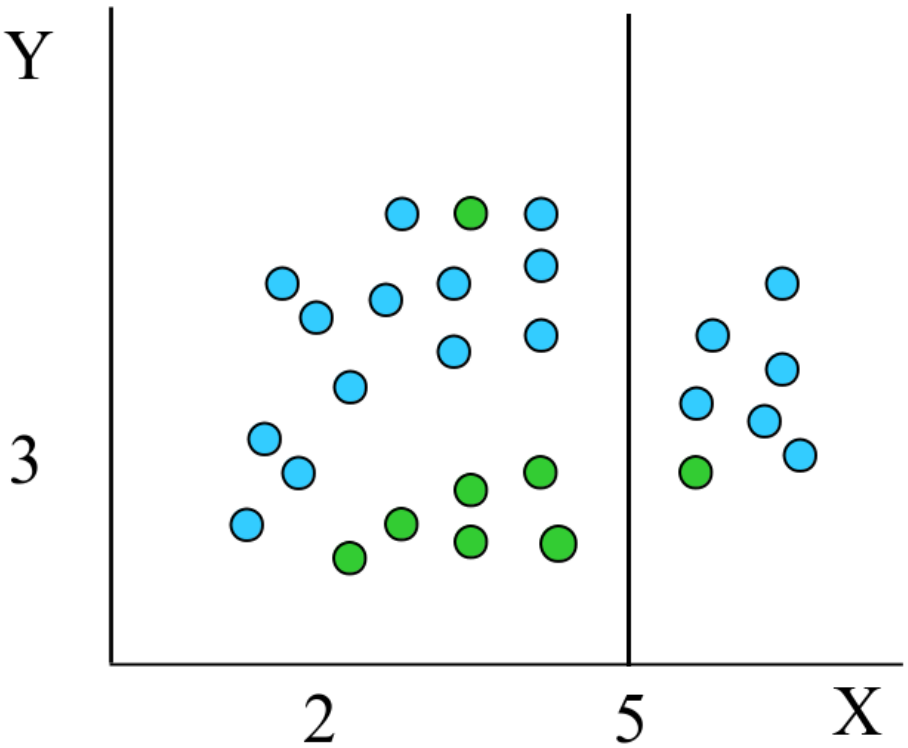
Classification: Decision trees



Decision Tree ($d = 0$)

```
return MAJORITY-CLASS();
```

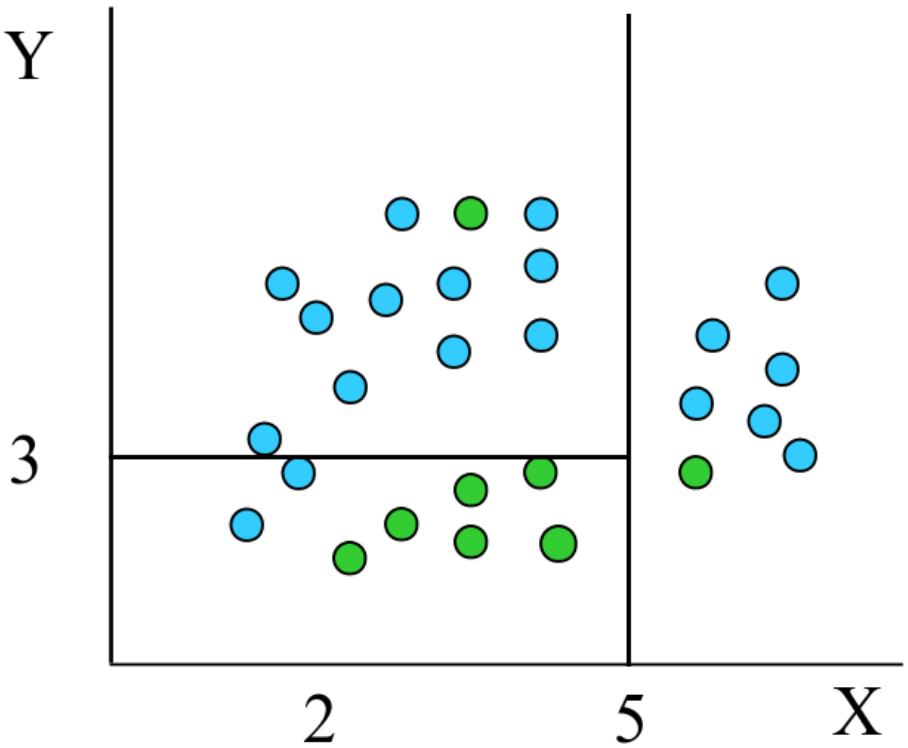
Classification: Decision trees



Decision Tree ($d = 1$)

```
if(X > 5) return BLUE;
else return GREEN; // oops!
```

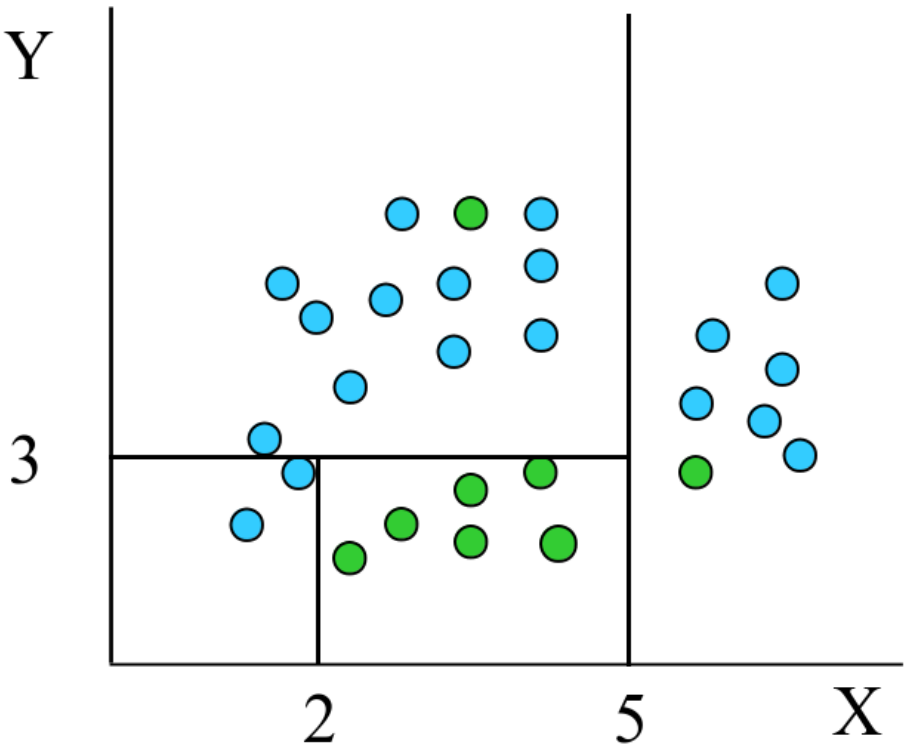
Classification: Decision trees



Decision Tree ($d = 2$)

```
if(X > 5) return BLUE;
elseif(Y > 3) return BLUE;
else return GREEN;
```

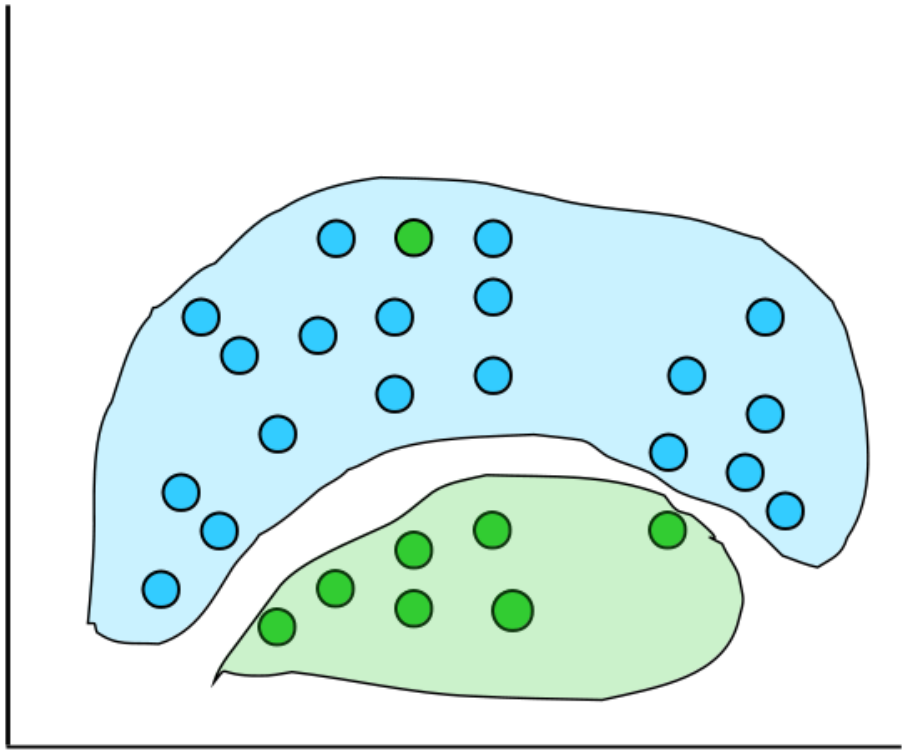
Classification: Decision trees



Decision Tree ($d = 3$)

```
if(X > 5) return BLUE;
elseif(Y > 3) return BLUE;
elseif(X > 2) return GREEN;
else return BLUE;
```

Classification: Neural networks



Neural Networks

- Perceptron
- Multi-level
- Backpropagation
- **Deep learning**

Universiteit Leiden

Data mining results validation

Data mining results validation

- Domain experts
- Ground truth
- Correlation vs. causation
- Outlier or data error
- Manual inspection vs. numeric measures

Evaluating results

- 200 student test results are predicted
- grade $\in$ {failed, passed, cum laude}

		predicted class		
		failed	passed	cum laude
actual class	failed	178	22	0
	passed	21	175	2
	cum laude	1	3	18

Confusion matrix

		predicted class: +	predicted class: -	
actual class	+	*tp*	*fn*	*p*
	-	*fp*	*tn*	*n*
		p'	*n'*	*N*

name	*formula*
error	*(fp+fn)/N*
accuracy	*(tp+tn)/N*
tp-rate	*tp/p*
fp-rate	*fp/n*
precision	*tp/p'*
recall	*tp/p*

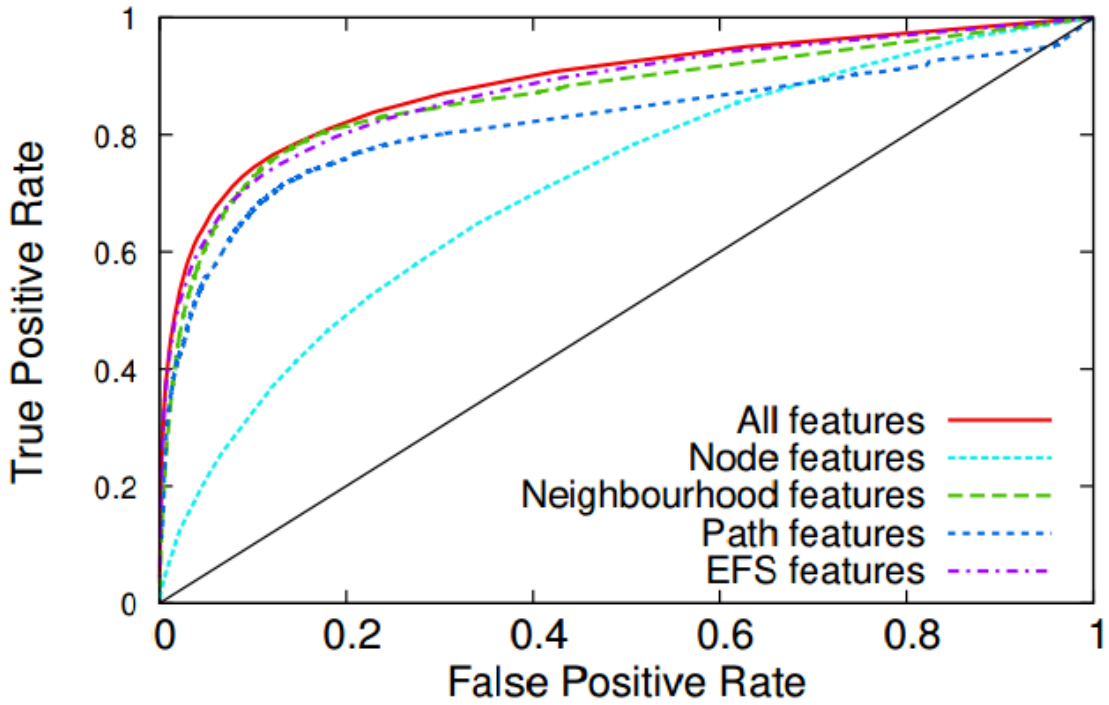
Data mining validation techniques

- Confusion matrix measures
- F-measure to balance precision and recall

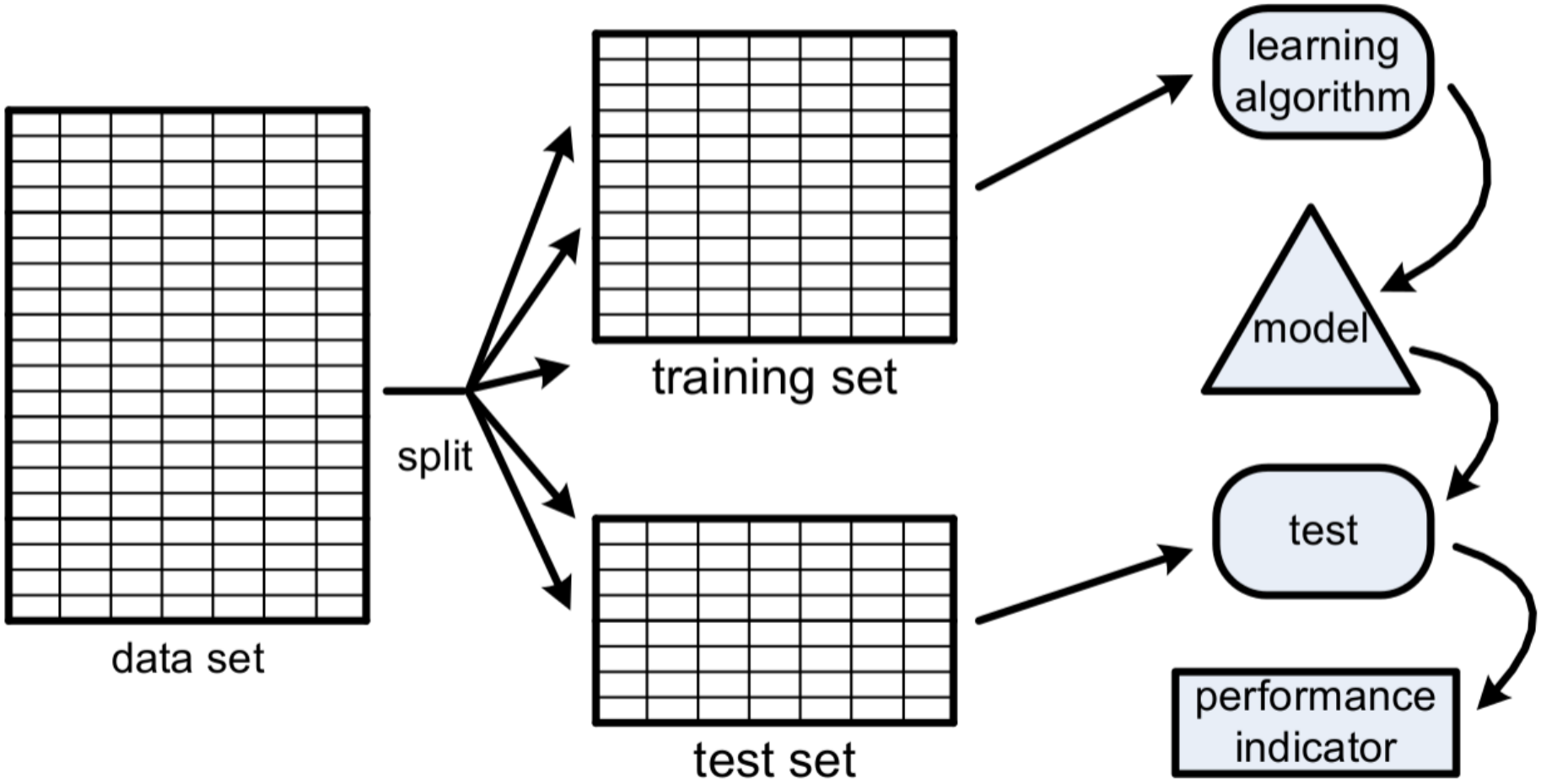
$$\text{F-measure} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

- ROC curves
- AUROC
- Separate training, testing and validation sets — **cross-validation**
- ***k*-fold cross-validation**

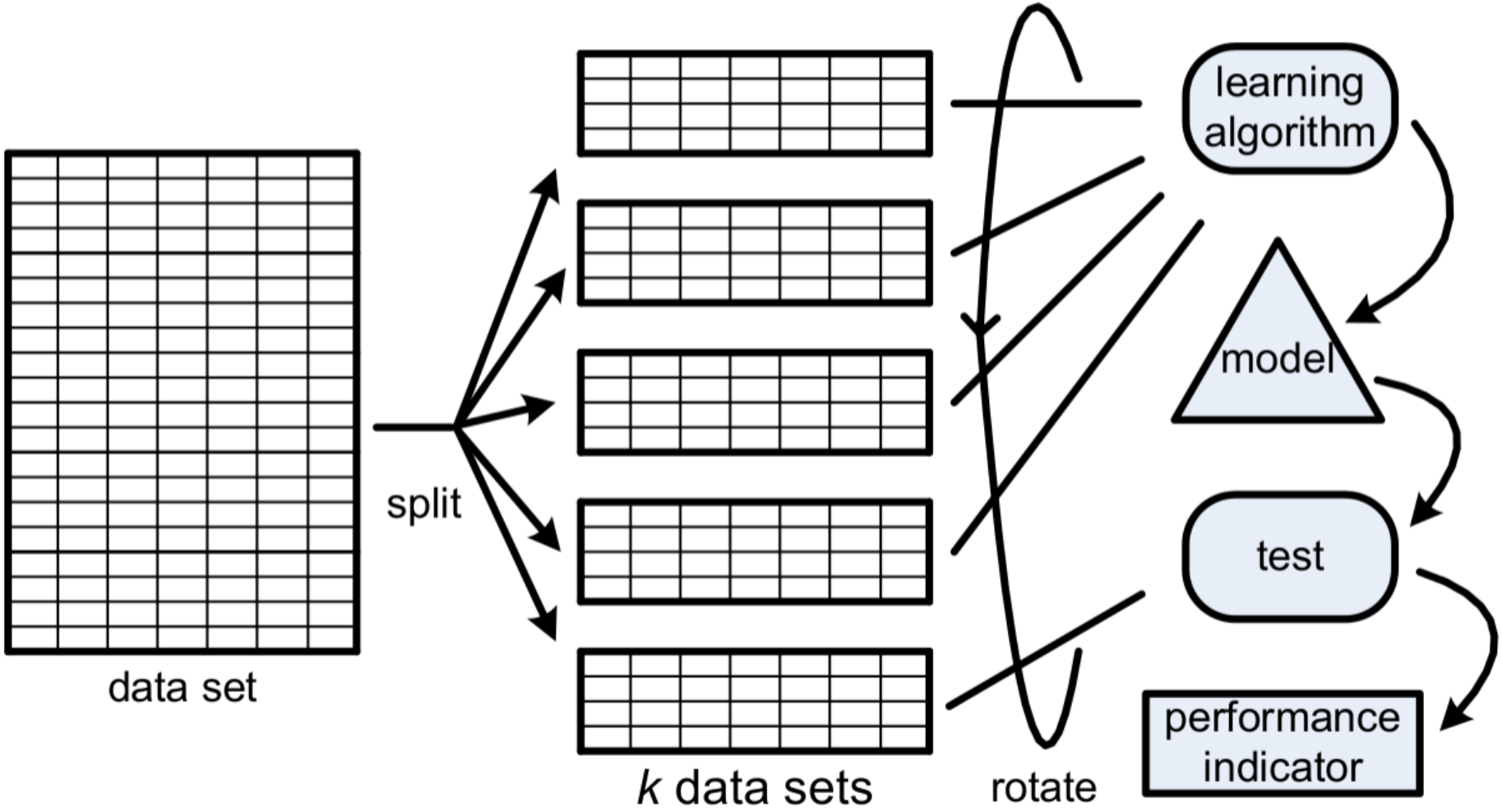
ROC


True Positive Rate
False Positive Rate
0
0.2
0.4
0.6
0.8
1
All features
Node features
Neighbourhood features
Path features
EFS features

Cross-validation



k-fold Cross-validation



Occam's razor

Occam's razor

- William of Ockham
- 14th century logician
- "One should not increase, beyond what is necessary, the number of entities required to explain anything"
- **Minimal Description Length (MDL)** principle: best model uses minimal number of bits to describe the data
- Watch out for **overfitting**

Curse of dimensionality

- Too many attributes

Curse of dimensionality

- Too many attributes
- Or: too many values per categorical attribute

Curse of dimensionality

- Too many attributes
- Or: too many values per categorical attribute
- Problematic:
  - Algorithms may have exponential runtime in the number of attributes
  - Sparse data on combinations of attributes makes training difficult
  - To get a statistically relevant results, required number of data records grows exponentially

Dimensionality reduction

- Solve problem of many attributes
- Possible solutions:
  - Principal Component Analysis (PCA)
  - Backward Feature Elimination
  - Forward Feature Construction

`http://www.kdnuggets.com/2015/05/7-methods-data-dimensionality-reduction.html`

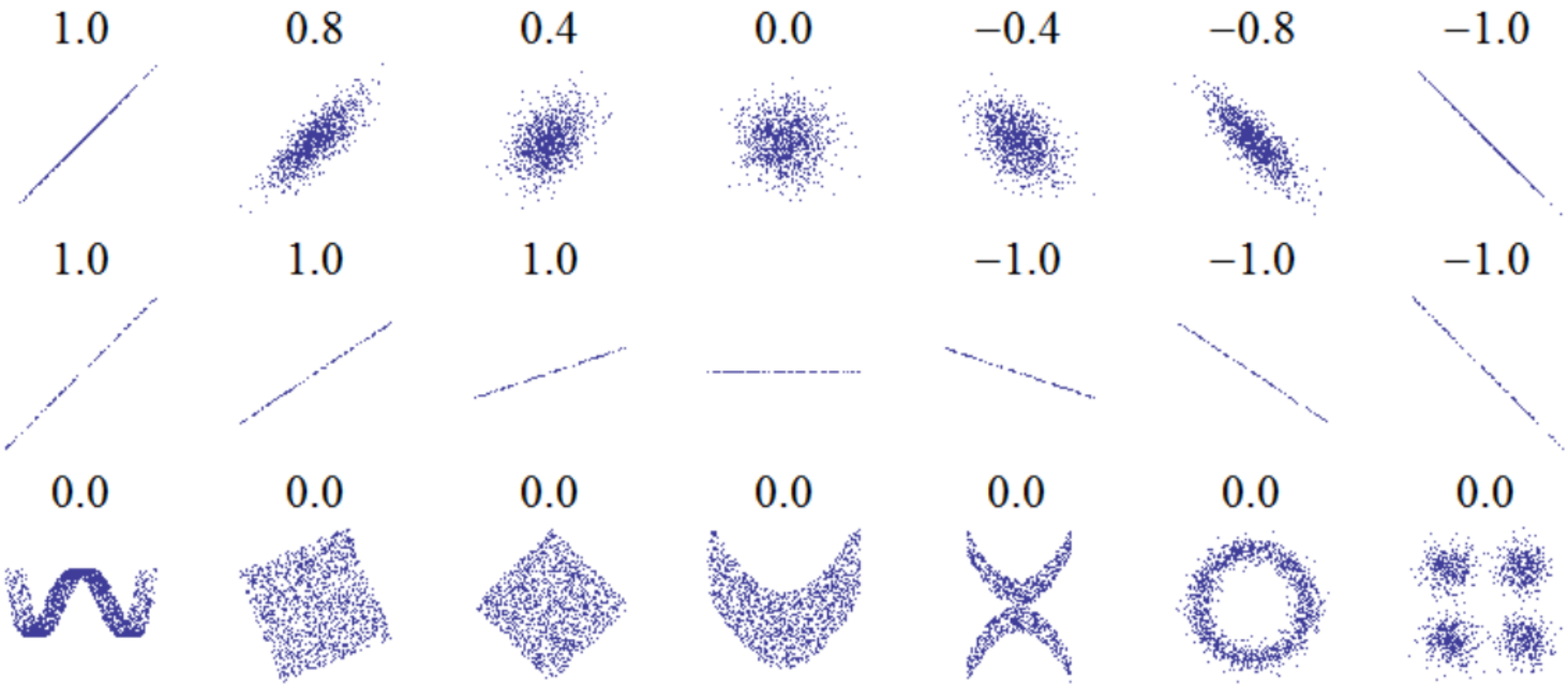
Feature extraction

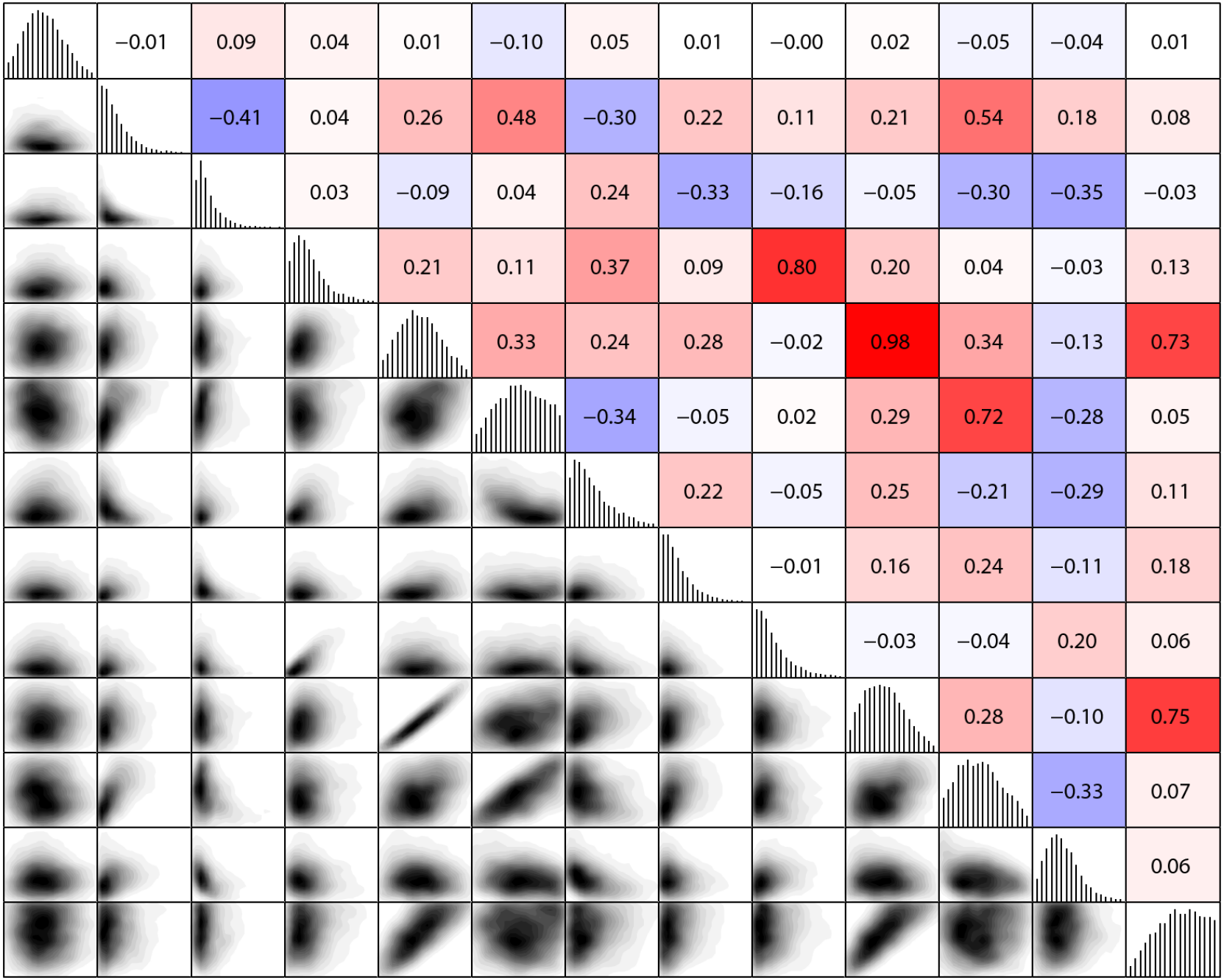
- Process of dimensionality reduction
- Derive features from original data that are
  - Informative
  - Non-redundant
  - Facilitating for the mining activity
- Manual: Correlation matrix

Correlation matrix

- Correlation: extent to which two variables are
  - related (positive value between 0 and 1) or
  - inversely related (negative value between -1 and 0)
- **Correlation matrix**: matrix in which each cell represents the correlation of the variables represented by the respective row and column

Correlation





https://www.bgc-jena.mpg.de/bgi/uploads/People/MaartenBraakhekke/correlationMatrix.png

Case: Churn prediction

- Master project by P. Kusuma
- Dutch telecom provider
- 700 million call records
- Churn: customer switching to competitor
- Use (data mining) techniques to predict churn
- Problem: class imbalance

P.D. Kusuma et al., Combining Customer Attributes and Social Network Mining for Prepaid Mobile Churn Prediction, in Proceedings of the 22th Belgian Dutch Conference on Machine Learning, pp. 50-58, 2013.

Churn

http://www.exacaster.com/

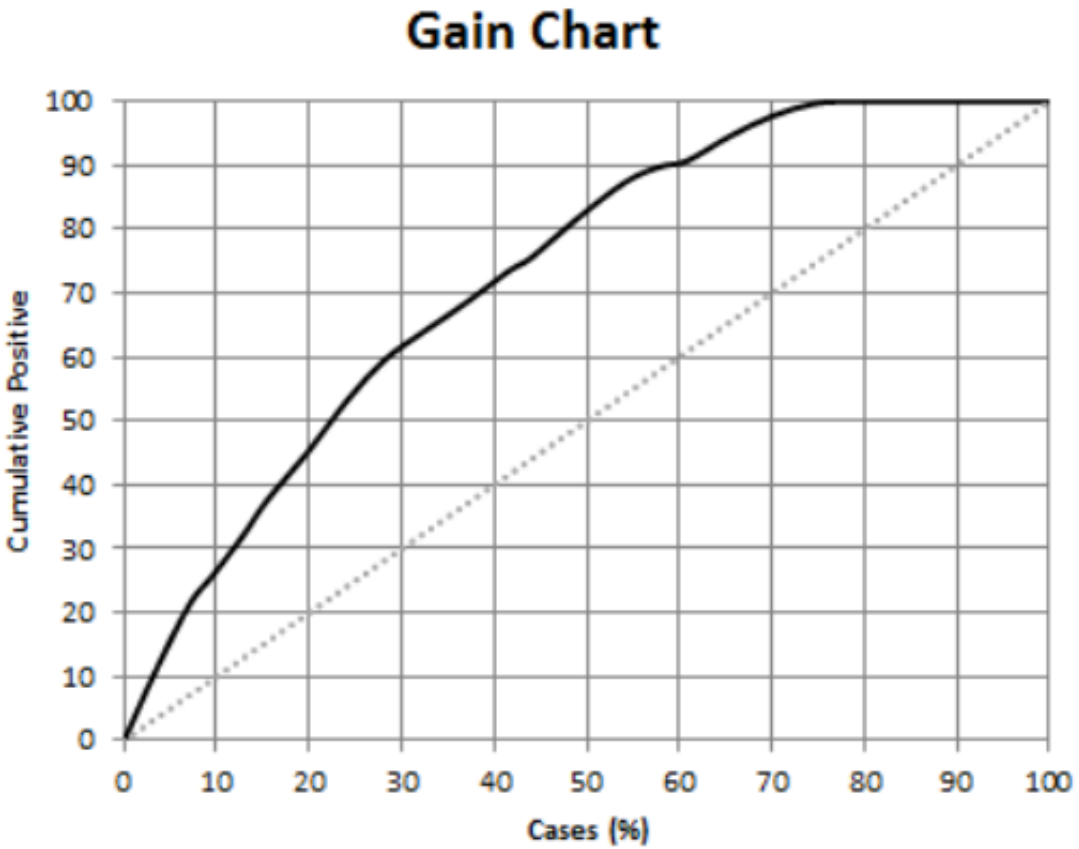
Customer attributes

- Demographic characteristics (age)
- Contractual information (type of subscription and package plan)
- Handset information (handset model and manufacturer)
- Service usage (voice call duration, SMS count and data usage)
- Churn identification (churner or non-churner)
- Network attributes (churners in call neighborhood, text, neighborhood, etc.)

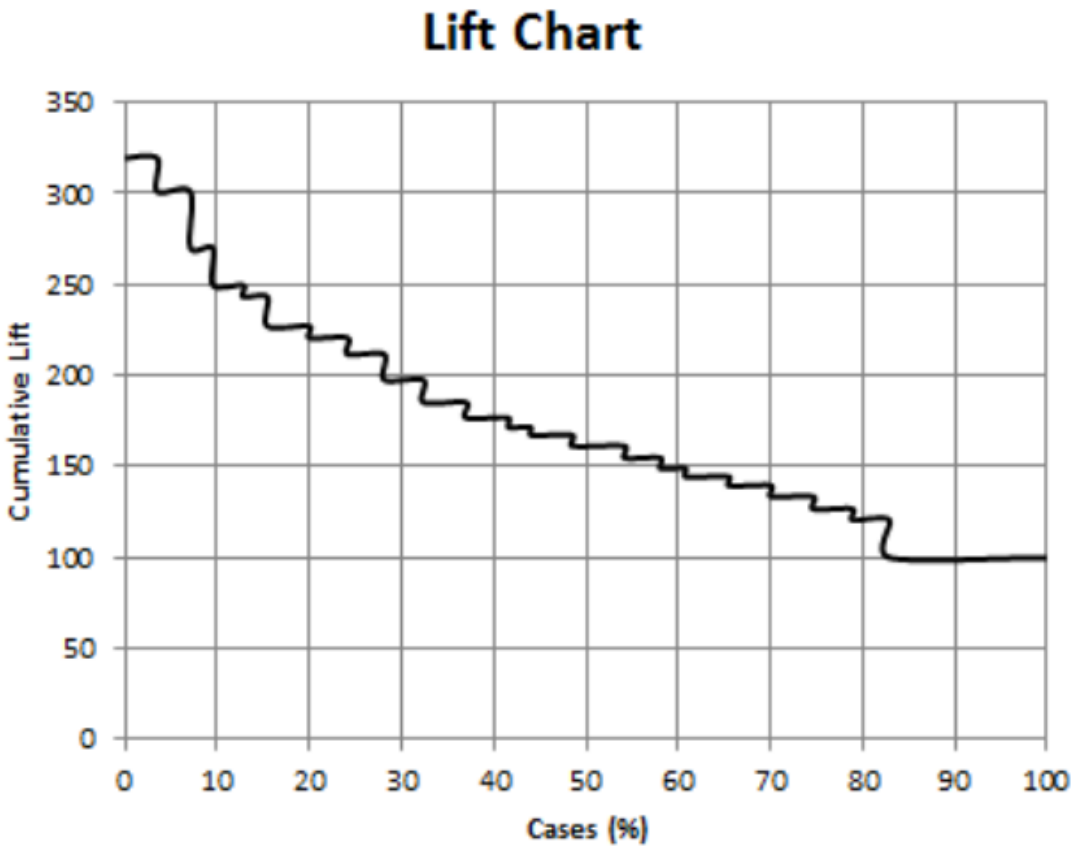
Techniques

- Decision trees
- Social Network models
- Separate training, testing and validation sets
- Rank customers based on their probability of being a churner
- **Gain chart**: percentage of customers contacted (horizontal axis) vs. percentage of churning customers found (vertical axis)
- **Lift chart**: percentage of customers contacted (horizontal axis) vs. gain improvement over random selection

Gain chart


Gain Chart
Cumulative Positive
Cases (%)
0
10
20
30
40
50
60
70
80
90
100

Lift chart


Lift Chart
Cumulative Lift
350
300
250
200
150
100
50
0
0
10
20
30
40
50
60
70
80
90
100
Cases (%)

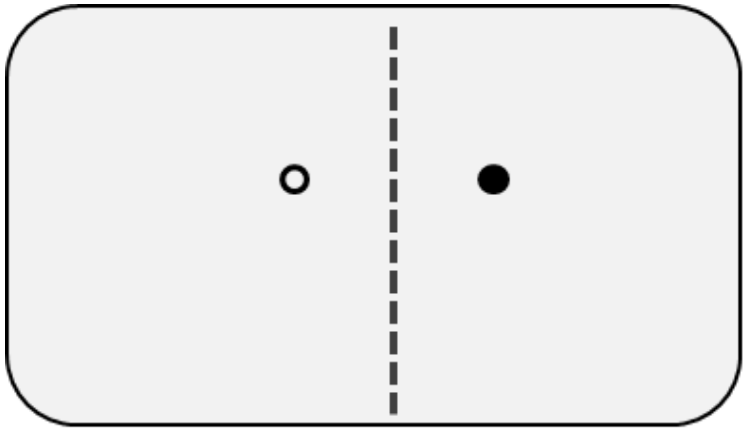
Categories of techniques

- Supervised learning: learning on labeled data
- **Semi-supervised learning**: partially labeled data
- Unsupervised learning: leaning/mining on unlabeled data
- Reinforcement learning: agents learning to act in an environment

Semi-supervised learning

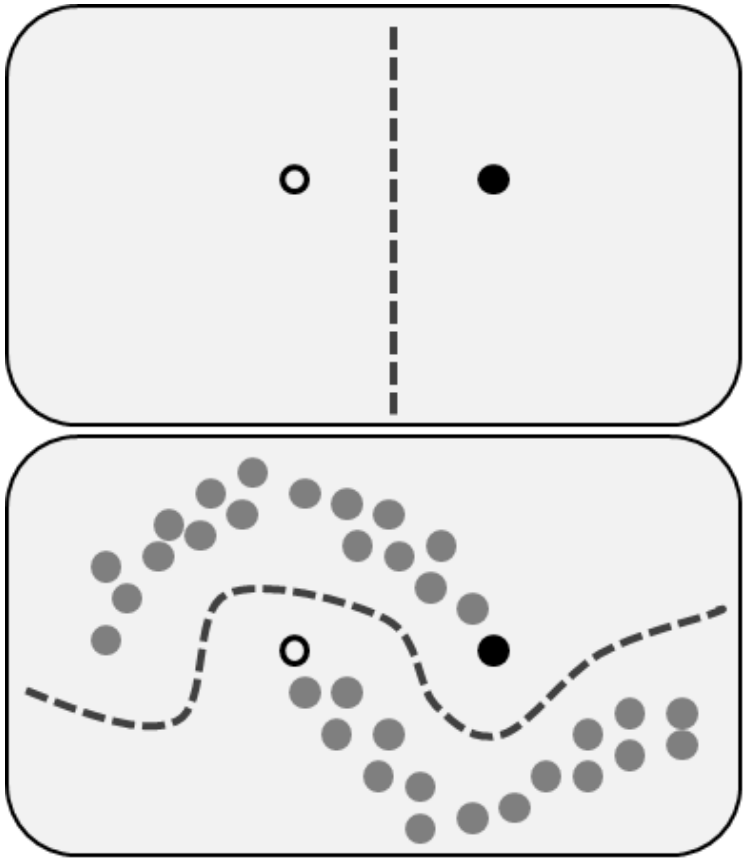
- **Semi-supervised learning**: learning from both labeled and unlabeled data
- Smoothness assumption: data points close to each other, are more likely to share the same label
- Cluster assumption: data tends to form discrete clusters, and points in the same cluster are more likely to share a label
- Lower dimensionality assumption: probably, the effective dimensionality of the data is much lower than the number of input attributes

Semi-supervised learning



http://en.wikipedia.org/wiki/Semi-supervised_learning

Semi-supervised learning



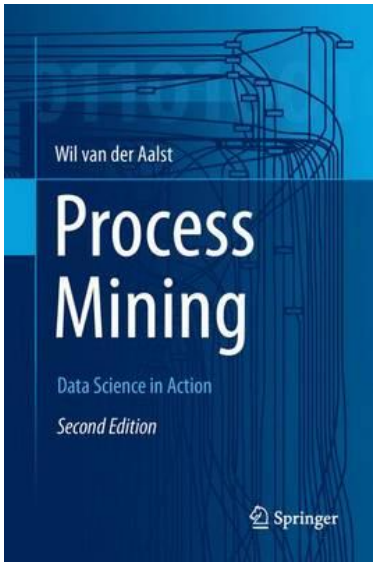
http://en.wikipedia.org/wiki/Semi-supervised_learning

Lab session March 2

- Finalize Assignment 1
- Ask your final questions

Credits

Lecture partially based on (slides of the (previous edition of the)) course book: W. van der Aalst, *Process Mining: Data Science in Action*, 2nd edition, Springer, 2016.



Slides partially based on "From Data Mining to Knowledge Discovery: An Introduction" by Gregory Piatetsky-Shapiro (`KDnuggets.com`)