

02450 Introduction to machine learning and data modeling

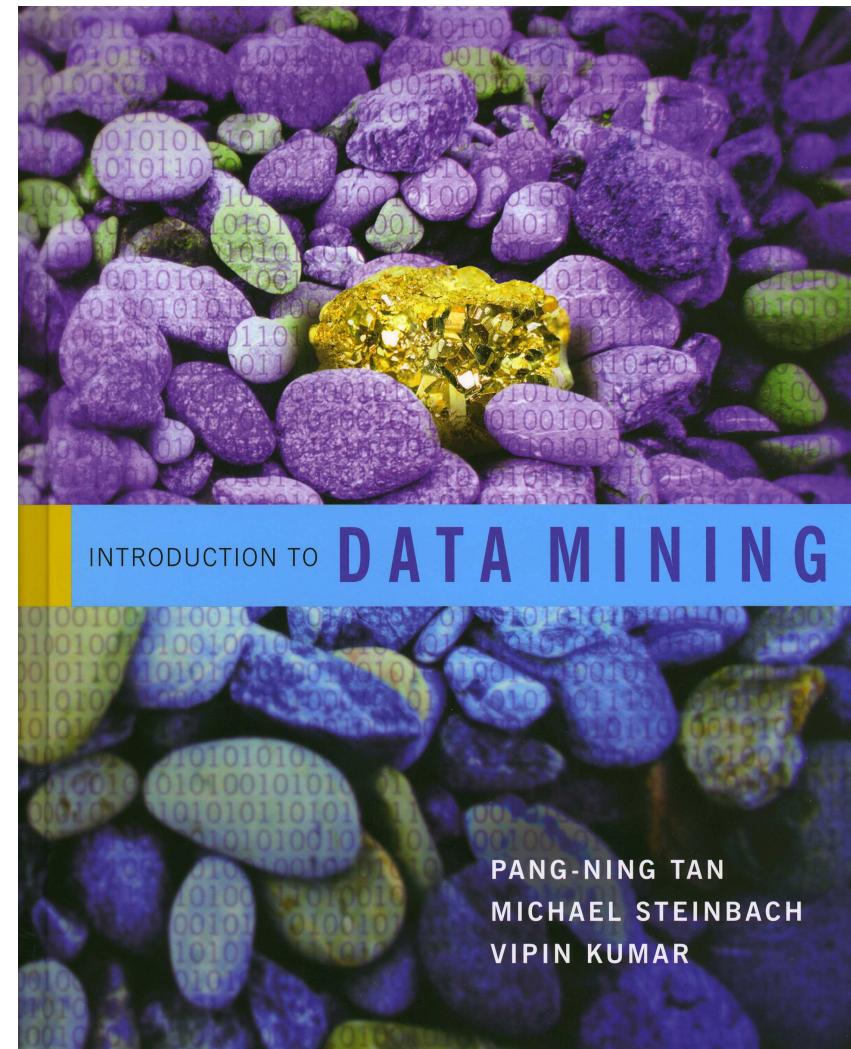
$$f(x+\Delta x) = \sum_{i=0}^{\infty} \frac{(\Delta x)^i}{i!} f^{(i)}(x)$$

$\Theta^{\sqrt{17}} + \Omega \int_0^{\infty} \delta e^{i\pi} =$
 $\Sigma! \gg \chi^2 = \{2.7182818284$

Reading material

Tan, Steinbach and Kumar
"Introduction to Data Mining"

Section 1.1-1.4



Lecture schedule

1. Introduction **(Tan 1.1-1.4)**

Data: Feature extraction and visualization

2. Data and feature extraction **(Tan 2.1-2.2 +(A)+ B.1)**

3. Measures of similarity and summary statistics **(Tan 2.4 + 3.1-3.2 + C1-C2)**

4. Data visualization **(Tan 3.3)**

Supervised learning: Classification and regression

5. Decision trees and linear regression **(Tan 4.1-4.3 + D)**

6. Overfitting and performance evaluation **(Tan 4.4-4.6)**

7. Nearest neighbor, naive Bayes, and artificial neural networks **(Tan 5.2-5.4)**

8. Ensemble methods and multi class classifiers **(Tan 5.6-5.8)**

Unsupervised learning: Clustering and density est.

9. K-means and hierarchical clustering **(Tan 8.1-8.3 + 8.5.7)**

10. Mixture models and association mining **(Tan 9.2.2 + 6.1-6.3)**

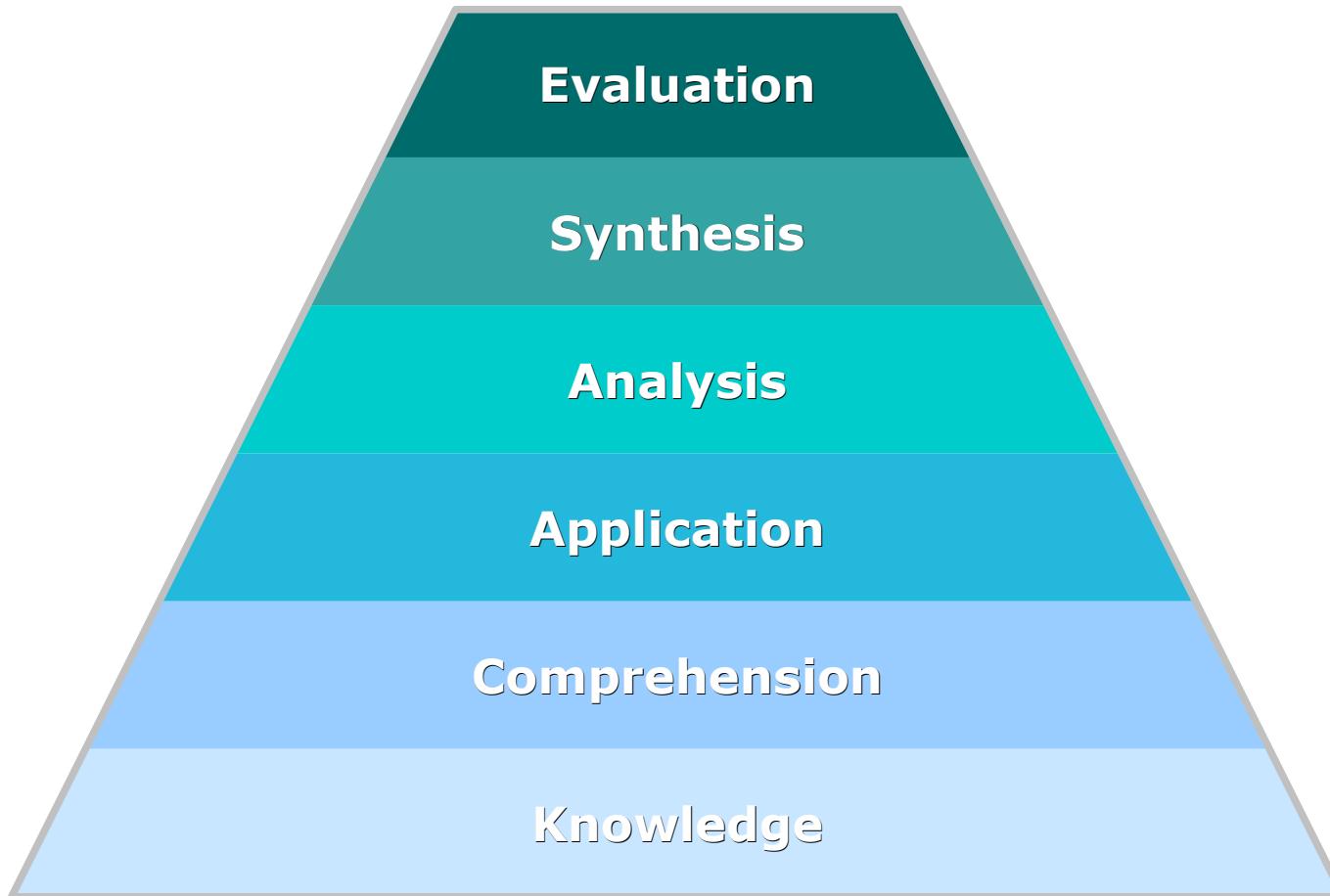
11. Density estimation and anomaly detection **(Tan 10.1-10.4)**

Machine learning and data modelling in practice

12. Putting it all together: Summary and overview

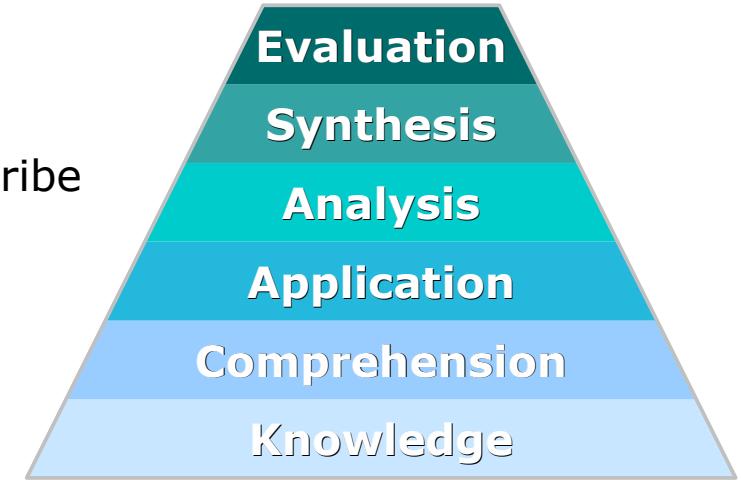
13. Mini project presentation

Blooms taxonomy



Learning objectives

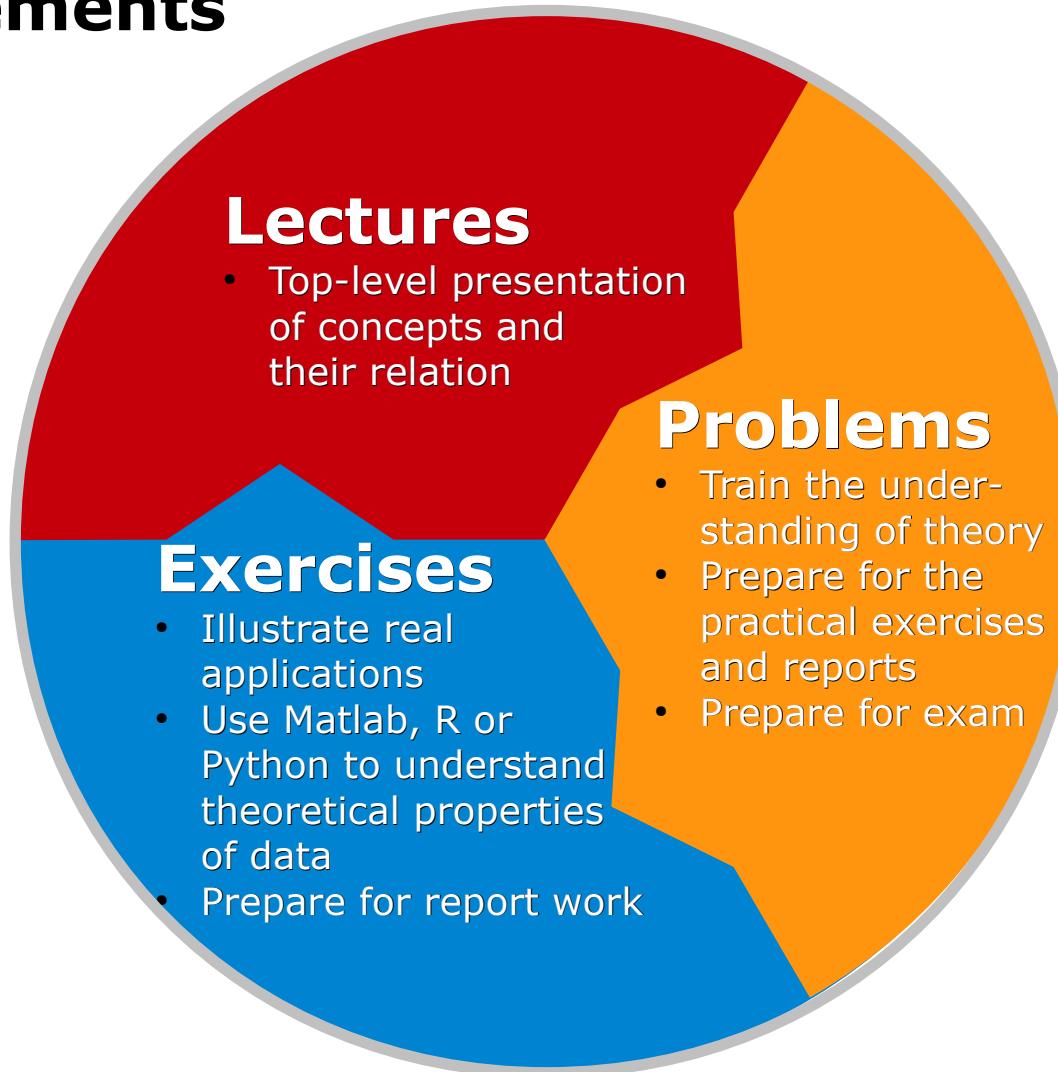
1. Describe the major steps involved in data modeling from preparing the data, modeling the data to evaluating and disseminating the results.
(Knowledge)
2. Discuss key machine learning concepts such as feature extraction, cross-validation, generalization and over-fitting, prediction and curse of dimensionality.
(Comprehension)
3. Sketch how the data modeling methods work and describe their assumptions and limitations.
(Knowledge and Comprehension)
4. Match practical problems to standard data modeling problems such as regression, classification, density estimation, clustering and association mining.
(Comprehension and Application)
5. Apply the data modeling framework to a broad range of application domains in medical engineering, bio-informatics, chemistry, electrical engineering and computer science.
(Application)
6. Compute the results of the data modeling framework by use of Matlab, R or Python.
(Application)
7. Use visualization techniques and statistics to evaluate model performance, identify patterns and data issues.
(Analysis)
8. Combine and modify data modeling tools in order to analyze a data set of their own and disseminate the results of the analysis.
(Application, Analysis, Synthesis and Evaluation)



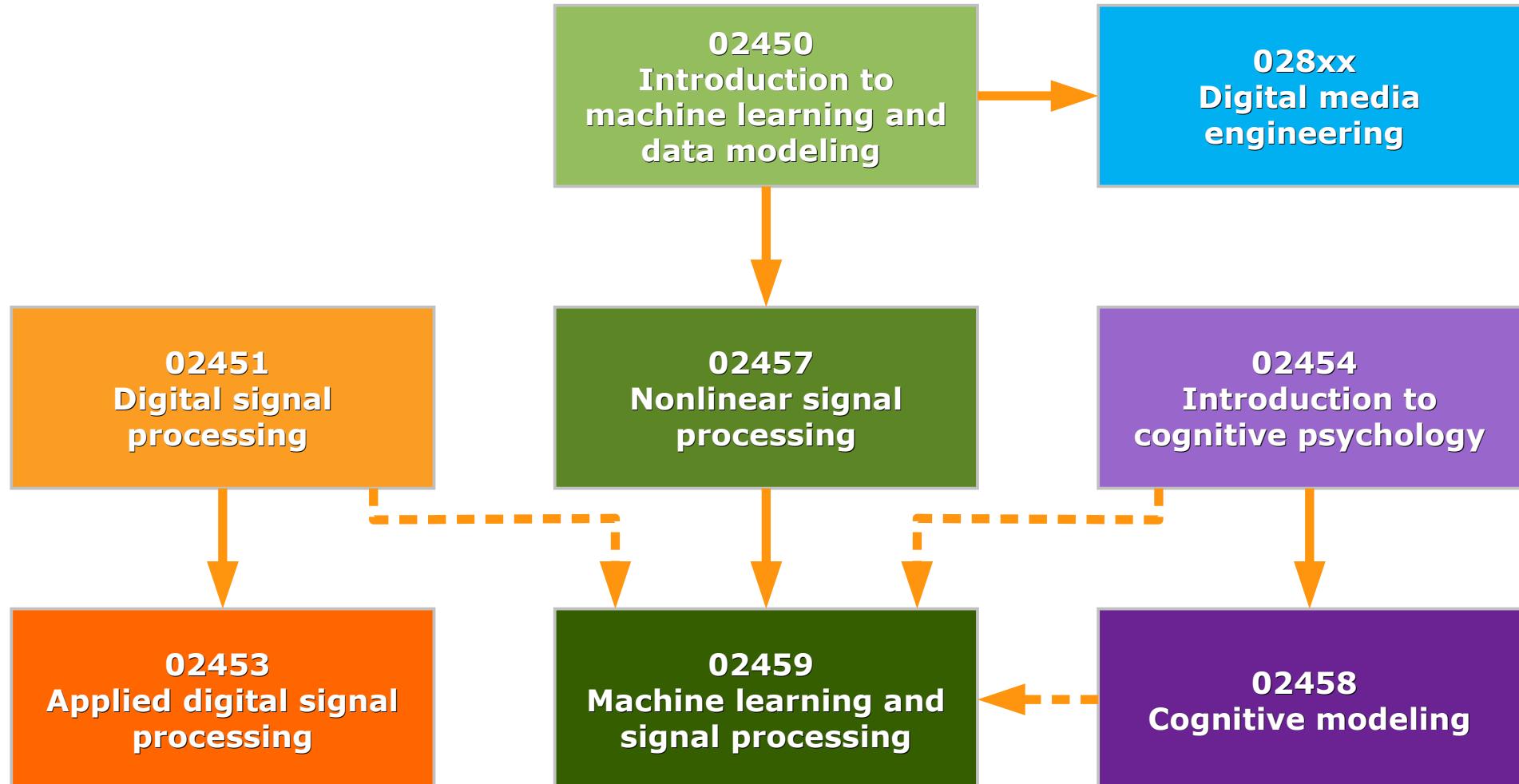
Examination

- 4 hours written examination (multiple choice)
Concentrated on the same topics as the lectures and exercises and linked to the learning objectives
- Group reports
Report Deadlines:
 - Report 1: 1st October, **Feature extraction and visualization**
 - Report 2: 5th November, **Supervised learning: Classification and regression**
 - Report 3: 3rd December, **Unsupervised learning: Clustering and density est.**
- Final grade based on an overall assessment of the reports and written exam

Course elements



Relevant courses at Section for Cognitive Systems

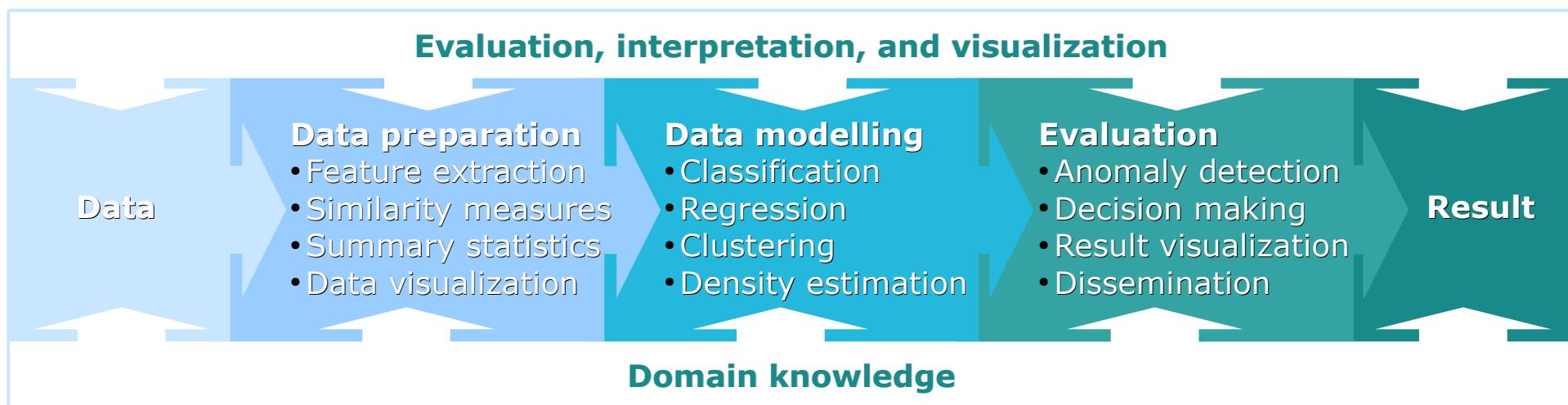


Pretest

The purpose of the pretest is to assess the students background and academic level in order to allow the teachers to adjust the presentation of the course material as well as measure student learning. This pretest will not be graded and will not influence exam results in any way. We do not expect you to be able to answer all questions.

<http://obsurvey.com/S2.aspx?id=FDCA52C2-A77D-4773-907C-4DA79793D3B2>

Data modeling framework



Every day, we create 2.5 quintillion (10^{18}) bytes of data — so much that 90% of the data in the world today has been created in the last two years alone. This data comes from everywhere: sensors used to gather climate information, posts to social media sites, digital pictures and videos, purchase transaction records, and cell phone GPS signals to name a few.

Source: <http://www-01.ibm.com/software/data/bigdata/>

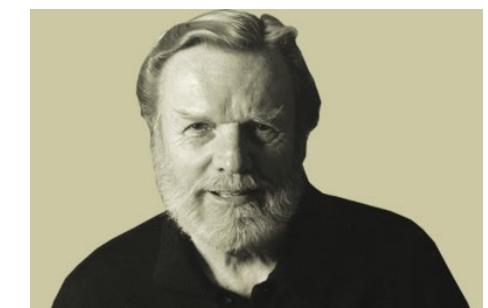
"If data had mass, the earth would be a black hole"

Stephen Marsland



"We are drowning in information and starving for knowledge"

John Naisbitt



We are entering the era of big data

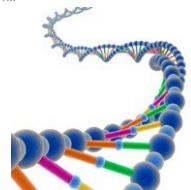


- ~1 trillion webpages

(<http://googleblog.blogspot.dk/2008/07/we-knew-web-was-big.html>)

- One hour of video is uploaded to youtube every second resulting in 10 years of content every day

(source: youtube)



- We have sequenced more than 1000 peoples genome of $3.8 \cdot 10^9$ base pairs

(source: K. P. Murphy "Machine Learning")

- Walmart handles more than 1 mio. transactions per hour and has databases containing more than $2.5 \cdot 10^{15}$ bytes of information

(source: K. P. Murphy "Machine Learning")



- Each night the worlds astronomy laboratories store high-resolution of the night sky of around a terabyte (10^{12})

(source: Stephen Marsland "Machine Learning An Algorithmic Perspective")



- In total, the four main detectors at the Large Hadron Collider (LHC) produced 13 petabytes (10^{15}) of data in 2010

(source: wikipedia "Big Data")



- Facebook handles 40 billion photos from its user base.

(source: wikipedia "Big Data")



- FICO Falcon Credit Card Fraud Detection System protects 2.1 billion active accounts world-wide

(source: wikipedia "Big Data")



The End of Theory: The Data Deluge Makes the Scientific Method Obsolete

By Chris Anderson  06.23.08



THE PETABYTE AGE:

Sensors everywhere. Infinite storage. Clouds of processors. Our ability to capture, warehouse, and understand massive amounts of data is changing science, medicine, business, and technology. As our collection of facts and figures grows, so will the opportunity to find answers to fundamental questions. Because in the

"All models are wrong, but some are useful."

So proclaimed statistician George Box 30 years ago, and he was right. But what choice did we have? Only models, from cosmological equations to theories of human behavior, seemed to be able to consistently, if imperfectly, explain the world around us. Until now. Today companies like Google, which have grown up in an era of massively abundant data, don't

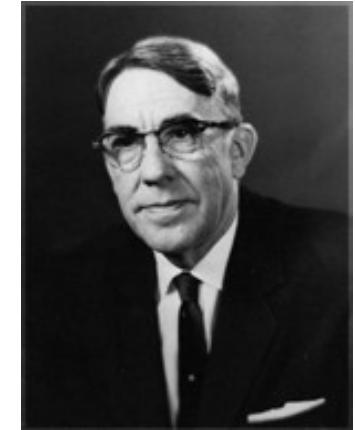


Peter Norvig
Research director of Google

Analysis of massive amounts of data will be the main driving force of all sciences in the future!!!

What is machine learning?

- Arthur Samuel (1959)
 - **Machine learning:** "Field of study that gives computers the ability to learn without being explicitly programmed"
 - Samuels wrote a checkers playing program
 - Had the program play 10000 games against itself
 - Work out which board positions were good and bad depending on wins/losses
- Tom Michell (1999)
 - **Well posed learning problem:** "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E."
 - The checkers example,
 - E = 10000 games
 - T = playing checkers
 - P = if you win or not



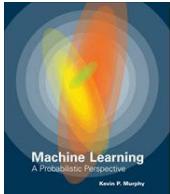
Arthur Samuel
(1901-1990)



Tom Michell

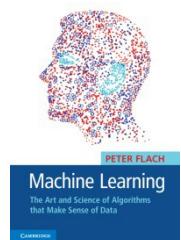
Source: http://www.holehouse.org/mlclass/08_Neural_Networks_Representation.html

What is machine learning?



"The goal of machine learning is to develop methods that can automatically detect patterns in data, and then to use the uncovered patterns to predict future data or other outcomes of interest"

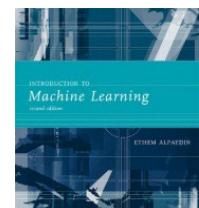
Kevin P. Murphy
"Machine Learning" 2012



"Machine learning is the systematic study of algorithms and systems that improve their knowledge or performance with experience"

Peter Flach

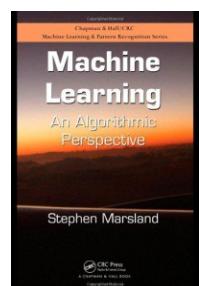
"Machine Learning The Art and Science of Algorithms that Make Sense of Data", 2012



"Machine learning is programming computers to optimize a performance criterion using example data or past experience. We have a model defined up to some parameters, and learning is the execution of a computer program to optimize the parameters of the model using the training data or past experience."

Ethem Alpaydin

"Introduction to Machine Learning", 2010



"[Machine learning] lies on the boundary of several different academic disciplines, principally computer science, statistics, mathematics, and engineering"

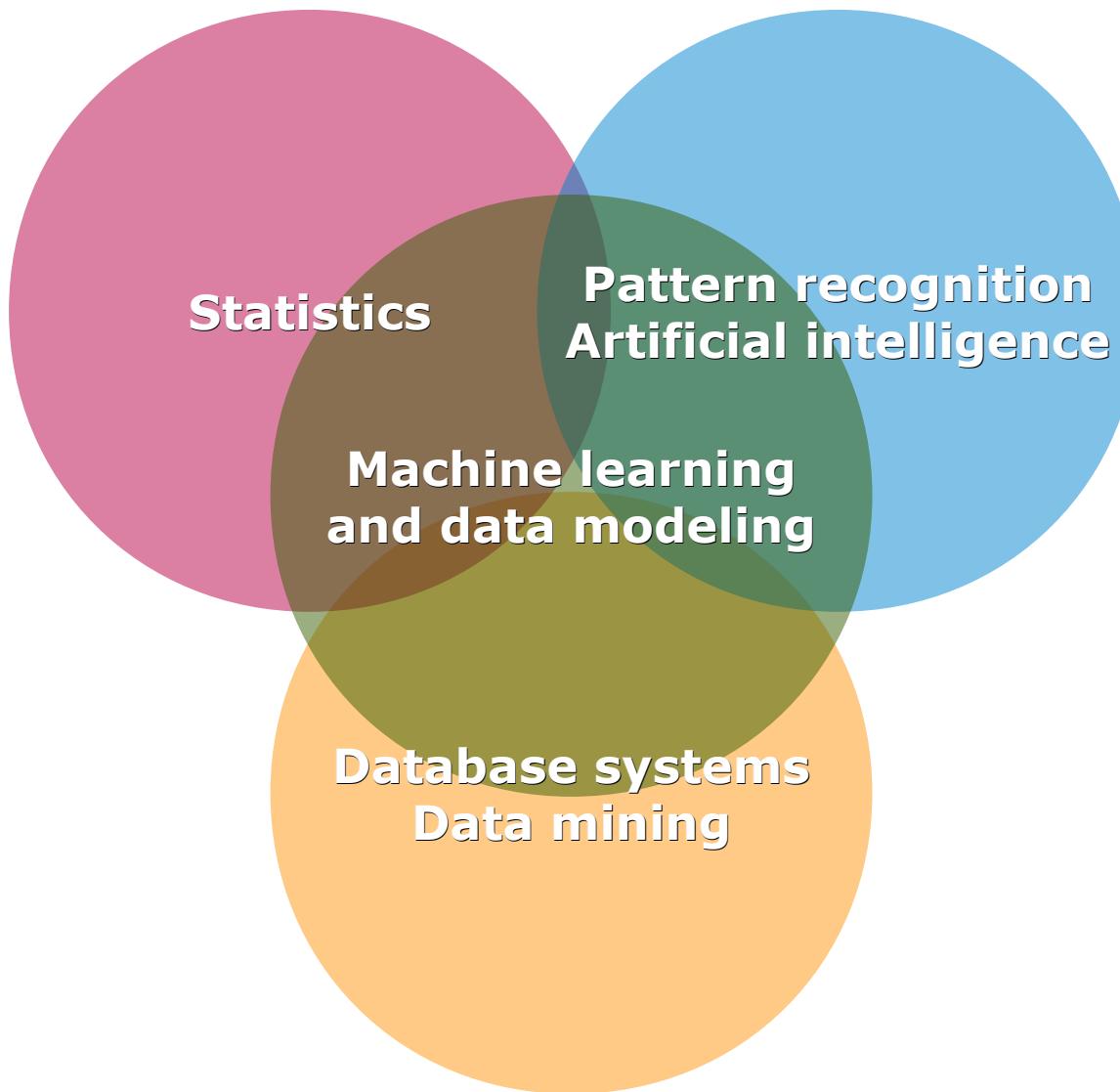
Stephen Marsland

"Machine Learning An Algorithmic Perspective", 2009



Source: http://www.holehouse.org/mlclass/08_Neural_Networks_Representation.html

Machine learning and data modeling





Motivating challenges

• Scalability

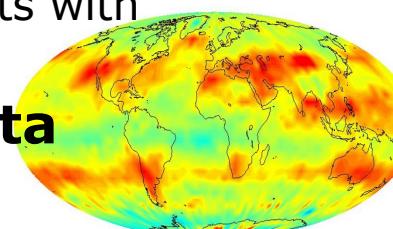
- Datasets with sizes of gigabytes, terabytes and petabytes are becoming common

• High Dimensionality

- It is now common to encounter data sets with hundreds or thousands of attributes

• Heterogenous and Complex Data

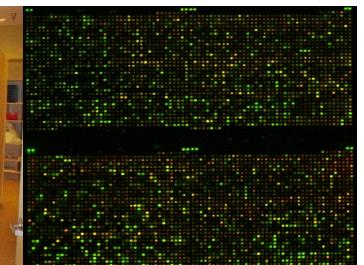
- Attributes commonly of different types
- Multiple types of datasets



Climate data



PET/fMRI



Micro-array data

• Data Ownership and Distribution

- Data distributed across many locations/organizations
- Security Issues, privacy preserving issues.



• Non-traditional Analysis

- Traditional statistical approach: Hypothesize-and-test paradigm
- Current data analysis tasks often require the generation and evaluation of thousands of hypotheses.
- Data mining can automate this process to identify interesting hypotheses for formal testing.

Applications

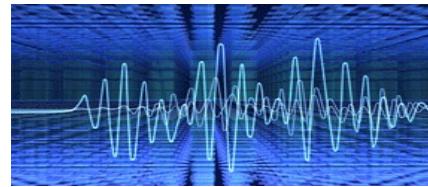
- **Chemistry**

- Spectrometry
- Chemical sensors



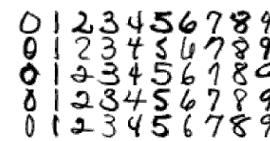
- **Audio processing**

- Spoken digit classification
- Music genre classification



- **Image processing**

- Hand-written digit recognition
- Image tagging and classification
- Number plate recognition



- **Informatics**

- Collaborative filtering
- Text corpus analysis
- Spam filters
- Computer games

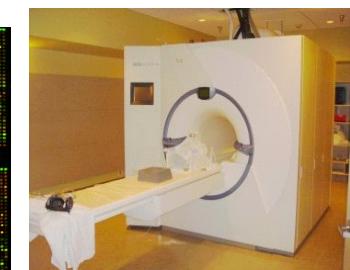
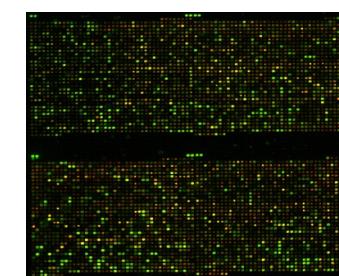
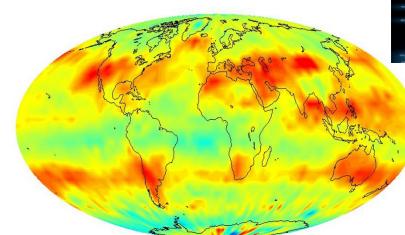
amazon.com

Google **eBay**

NETFLIX.

- **Biomedical**

- Micro-array gene analysis
- Medical Imaging



- **Financial data mining**

- Market predictions



- **Climate data**

- Weather forecast

Data Mining and Machine Learning Tasks

Predictive tasks (Supervised learning)

- Use some variables to predict unknown or future values of other variables

- **Classification**

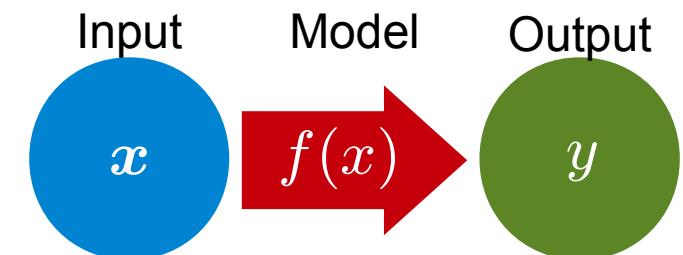
- Discrete output

(Determine which class a new data object belongs to)

- **Regression**

- Continuous output

(Determine the output value from the input variables)



Descriptive tasks (Unsupervised learning)

- Find human-interpretable patterns that describe the data

- **Clustering**

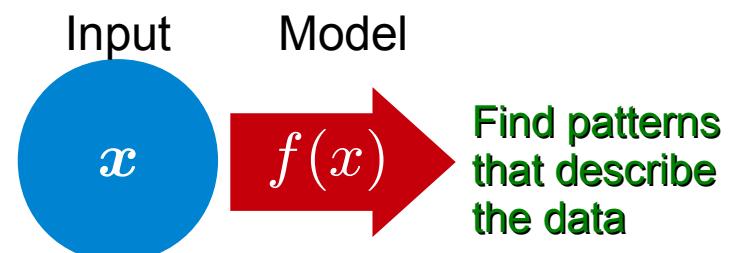
- Discover group structure in data

- **Association rule discovery**

- Discover how data objects relate to each other

- **Anomaly detection**

- Find data objects that are abnormal



Classification: Definition

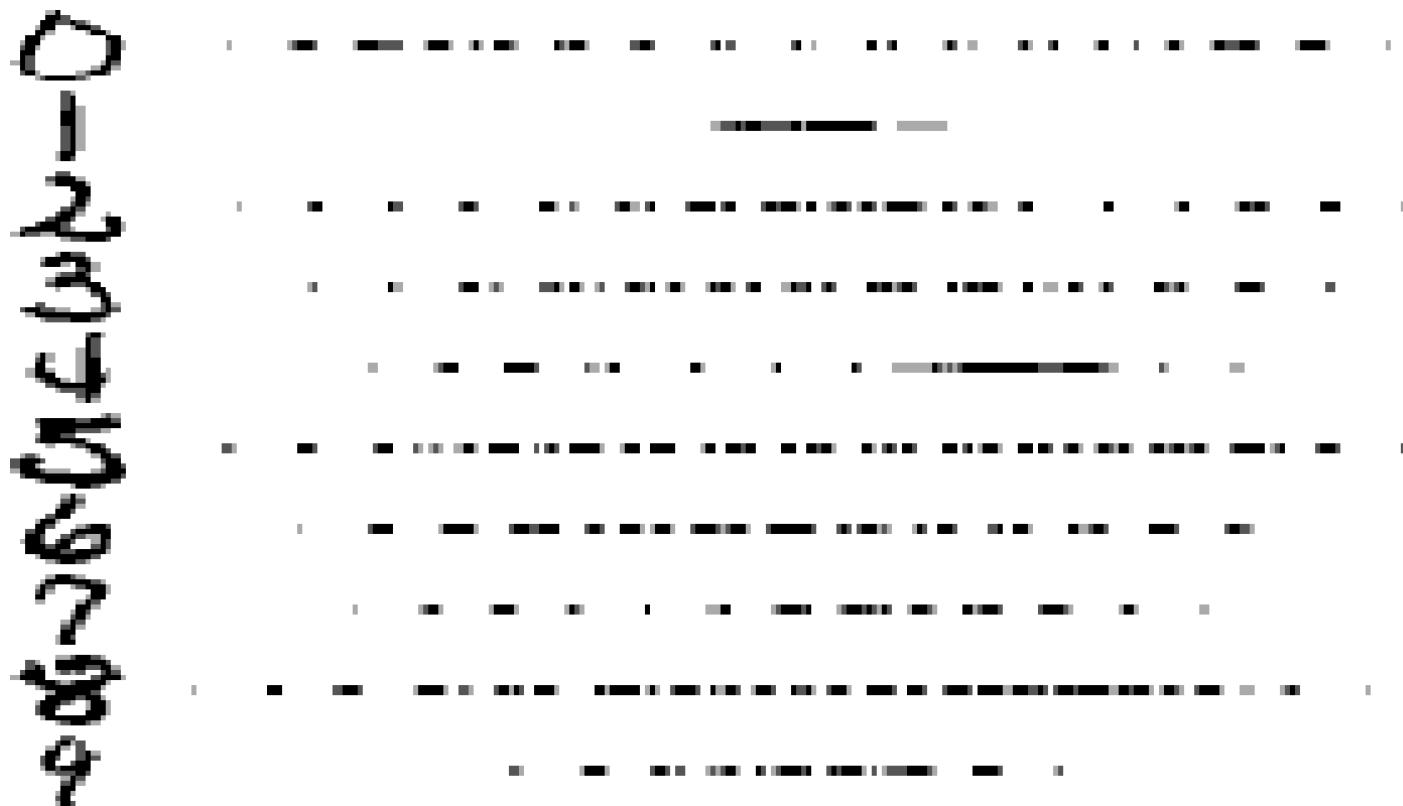
- Given a collection of data objects (**training set**)
 - Each object has associated a number of features
 - Each object belongs to a certain class
- Define a **model** for the class given the other features
- Goal: Assign a class label to a **previously unseen object**

Classification: Example

Training set										Classify		
0	1	2	3	4	5	6	7	8	9	?	?	?
0	1	2	3	4	5	6	7	8	9	5	2	4
0	1	2	3	4	5	6	7	8	9	9	9	9
0	1	4	2	3	3	6	7	8	9	9	9	9
0	1	2	3	3	4	5	6	7	8	9	9	9
0	1	2	3	4	5	6	7	8	9	9	9	9

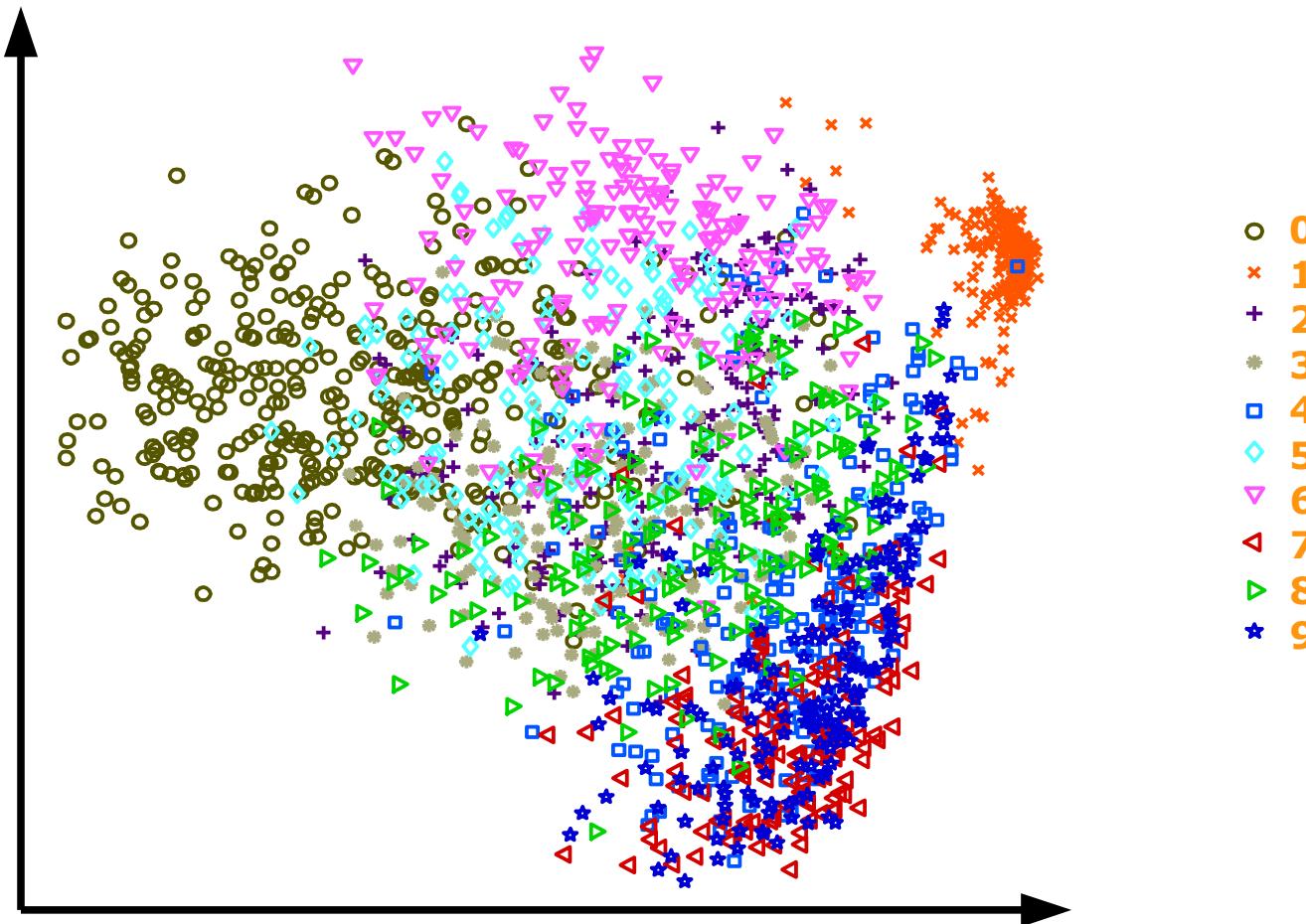
Classification: Example

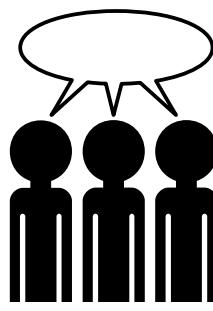
Data representation



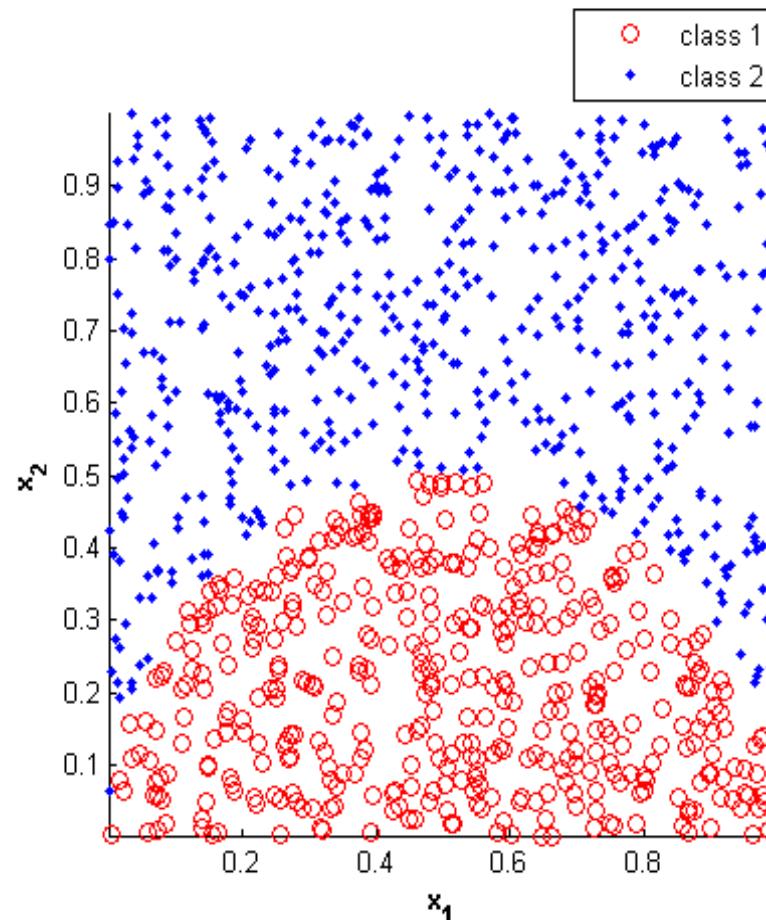
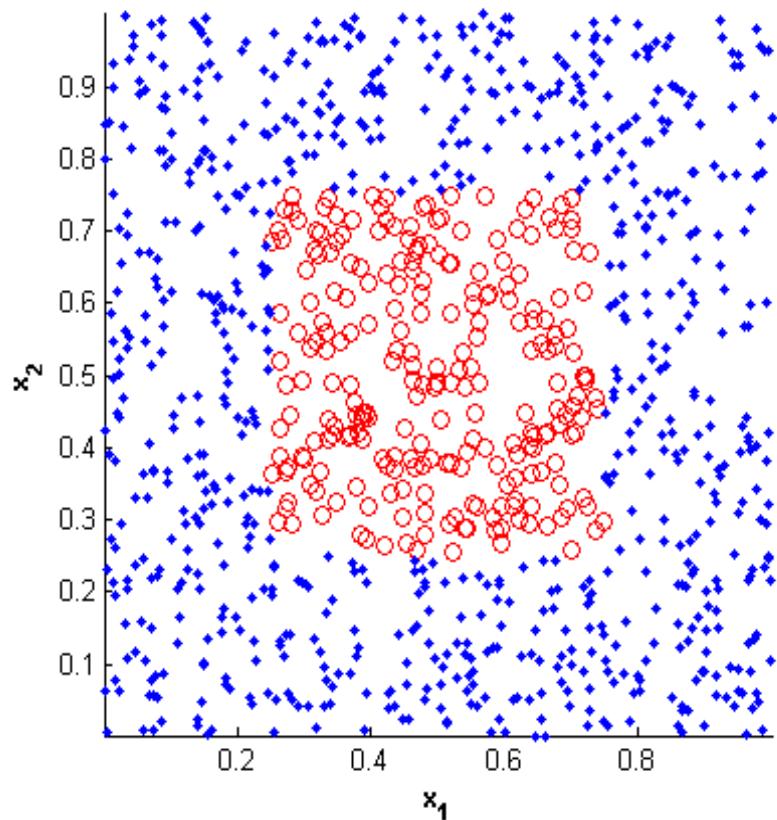
Classification: Example

Visualization



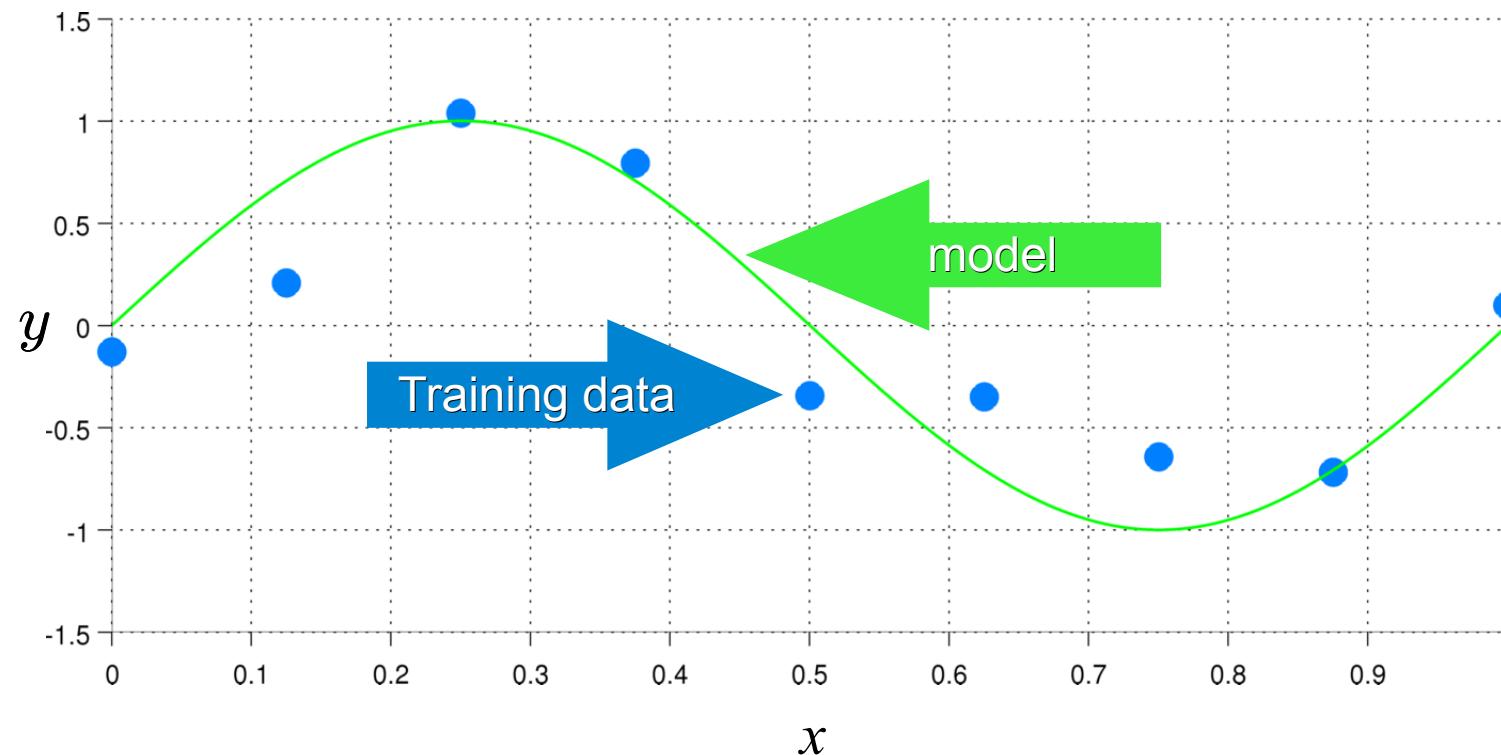


How would you characterize the two classes and how would you determine what class a new observation belongs to?



Regression: Definition

- Given a collection of data objects
 - Each object has associated a number of features
 - Each object has associated a **continuous valued variable**
- Define a **model** for the variable given the features
- Goal: Predict the value of the variable for a **previously unseen object**

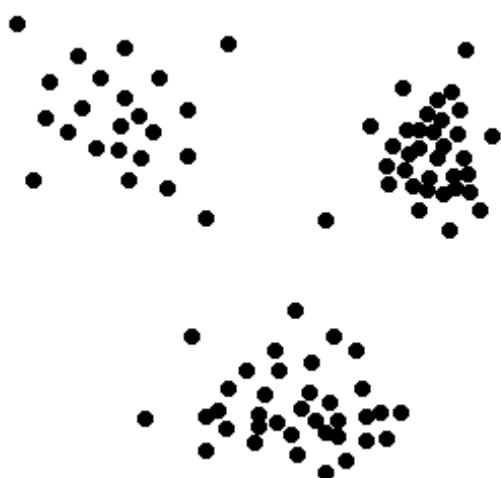


Regression: Example

- Predict **sales amounts** of new product based on
 - advertising expenditure
- Predict **wind velocity** as a function of
 - temperature, humidity, and air pressure
- Predict the value of a **stock market index** based on
 - previous index time series and market indicators

Clustering: Definition

- Given a collection of data objects
 - Each object has associated a number of features
 - A measure of **similarity** between objects is defined
- Goal: **Group the objects** into clusters such that
 - Objects within each cluster are similar
 - Objects in separate clusters are less similar



Clustering: Definition

- Given a collection of data objects
 - Each object has associated a number of features
 - A measure of **similarity** between objects is defined
- Goal: **Group the objects** into clusters such that
 - Objects within each cluster are similar
 - Objects in separate clusters are less similar



Clustering: Example

Document clustering

- Goal

- Find groups of documents that are similar to each other based on the important words appearing in them

- Approach

- Identify frequently occurring words in each document
- Define a similarity measure based on the word frequencies
- Perform clustering: Find groups of documents

- Gain

- Use the clusters to relate a new document to existing documents
- Better search algorithms: Return documents that are similar but do not have the exact search keywords



Association rule discovery: Definition

- Given a set of **records**
 - Each containing a number of **items from a set**
- Goal: Produce dependency rules
 - Predict the occurrence of an item based on occurrences of other items

Association rule discovery: Example

Market basket analysis



Training set

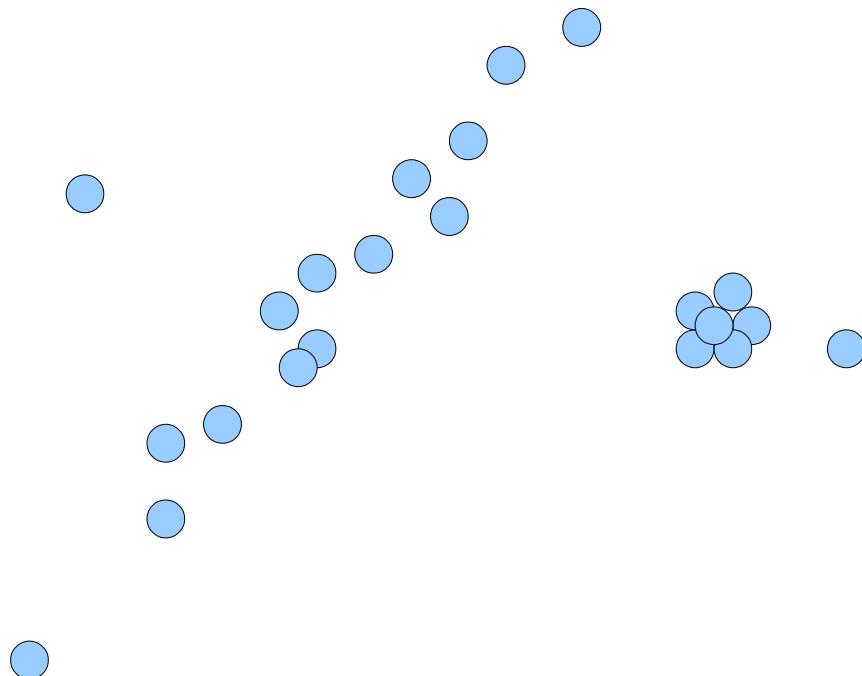
-
- 1. {Bread, Coke, Milk}
 - 2. {Beer, Bread}
 - 3. {Beer, Coke, Diaper, Milk}
 - 4. {Beer, Bread, Diaper, Milk}
 - 5. {Coke, Milk}

Rules discovered

-
- {Milk} \rightarrow {Coke}
 - {Diaper, Milk} \rightarrow {Beer}

Anomaly detection: Definition

- Given a collection of data objects
 - Each object has associated a number of features
- Detect which objects **deviate from normal** behaviour



Anomaly detection: Example

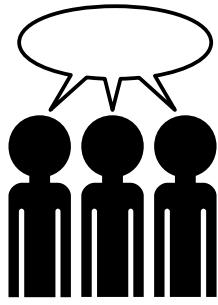
- Credit card **fraud detection**
 - Recognize dubious credit card transactions based on the transaction history of the card holder
- Detection of **outliers** in data measurements
 - Remove erroneous measurements due to misreading from an instrument
- **Fault detection** in system health monitoring
 - Detect when a wind turbine performs poorly due to ice coating on blades

Discussion

Which of these activities are machine learning / data modeling tasks?

(Tan Ex. 1.1)

- a) Dividing the customers of a company according to their gender
- b) Dividing the customers of a company according to their profitability
- c) Computing the total sales of a company
- d) Sorting a student database based on student identification number
- e) Predicting the outcomes of tossing a (fair) pair of dice
- f) Predicting the future stock price of a company using historical data
- g) Monitoring the heart rate of a patient for abnormalities
- h) Monitoring seismic waves for earthquake activities
- i) Extracting the frequencies of a sound wave



Group discussion

You all have different backgrounds and experiences

- Give an example of a data modeling problem from
 - your previous work or
 - your interests or
 - your imagination
- Which of the five modeling tasks could be applied to solve the problem?

Data modeling tasks

- **Classification**
 - Determine which class a new data object belongs to
- **Regression**
 - Learn a functional relationship from input to output
- **Clustering**
 - Discover group structure in data
- **Association rule discovery**
 - Discover how data objects relate to each other
- **Anomaly detection**
 - Find data objects that are abnormal

Exercises

All the exercises will be in Matlab, Python and R.

-Two options for running Matlab on your computer:

1)Install Matlab on your computer and run it using internet connection to a license server: <http://www.gbar.dtu.dk/downloads/#>

2) Run Matlab on the GBAR from your computer using thinlinc:
<http://gbar.dtu.dk/ThinLinc>

-Python and R is freely available

(exercise 1 today will guide you through how to install the programs)

You should form groups of 2-3 people for the exercises and for the 3 (group) reports.

Each group will give feedback to the teachers on the lectures and exercises of one of the course weeks.

Lecture schedule

1. Introduction (Tan 1.1-1.4)

Data: Feature extraction and visualization

2. Data and feature extraction (Tan 2.1-2.2 +(A) + B.1)

3. Measures of similarity and summary statistics (Tan 2.4 + 3.1-3.2 + C1-C2)

4. Data visualization (Tan 3.3)

Supervised learning: Classification and regression

5. Decision trees and linear regression (Tan 4.1-4.3 + D)

6. Overfitting and performance evaluation (Tan 4.4-4.6)

7. Nearest neighbor, naive Bayes, and artificial neural networks (Tan 5.2-5.4)

Report 1

Report 2

8. Ensemble methods and multi class classifiers (Tan 5.6-5.8)

Unsupervised learning: Clustering and density est.

9. K-means and hierarchical clustering (Tan 8.1-8.3 + 8.5.7)

10. Mixture models and association mining (Tan 9.2.2 + 6.1-6.3)

11. Density estimation and anomaly detection (Tan 10.1-10.4)

Machine learning and data modelling in practice

12. Putting it all together: Summary and overview

13. Mini project presentation

Report 3

See also the course homepage for the lecture outline and reading material for each week: www.imm.dtu.dk/courses/02450

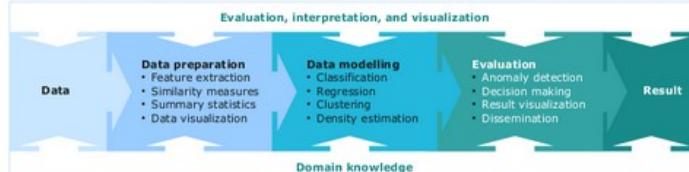


Section for Cognitive Systems
DTU Compute

02450 Introduction to Machine Learning and Data Modeling

Machine learning and data modeling

The course is designed around a data modeling framework shown in the figure. Each lecture/assignment will focus on an aspect of the data modeling framework.



We emphasize the holistic view of modeling in order to motivate and stress the relevance of individual components and building blocks, disseminate the obtained competence (see the course [learning objectives](#)), and make them applicable for a broad spectrum of engineering problems in e.g. biomedical engineering, chemistry, electrical engineering, and informatics.

Resources

Location

The lectures and computer assignments will take place in building 324 room 040 and 050. Please bring your own laptop computer.

Reading material

The course is based on the book: Pang-Ning Tan, Michael Steinbach, and Vipin Kumar, "Introduction to Data Mining". The book is available from Polyteknisk Boghandel.



Tan et al., Introduction to Data Mining



Morten Marup



Mikkel N. Schmidt



Ditte Høvenhoff Hald



Kit Melissa Larsen

Lecture slides and exercises

Lecture slides, handouts, assignment instruction etc. is available at the DTU Campusnet course page (requires formal enrolment to the course).

Course description

A description of the course can be found at the DTU Coursebase.

Teacher

- Morten Marup (MM), mmor@dtu.dk
- Mikkel N. Schmidt (MNS), mns@dtu.dk
- Ditte Høvenhoff Hald (DHH), ditha@dtu.dk
- Kit Melissa Larsen (KMELA), kmela@dtu.dk

Lecture schedule

No.	Date	Teacher	Subject	Reading material
1	3 Sep 2013	MM	Introduction	1.1-1.4
Data: Feature extraction, and visualization				
2	10 Sep 2013	MM	Data and feature extraction	2.1-2.3 + (A) + B.1
3	17 Sep 2013	MM	Measures of similarity and summary statistics	2.4 + 3.1-3.2 + C1-C2
4	24 Sep 2013	MM	Data visualization	3.3
Supervised learning: Classification and regression				
5	1 Oct 2013	MM	Decision trees and linear regression	4.1-4.3 + D
6	8 Oct 2013	MM	Oversampling and performance evaluation	4.4-4.6
7	15 Oct 2013		Autumn Holiday	
8	22 Oct 2013	MM	Nearest neighbor, naive Bayes, and artificial neural networks	5.2-5.4
9	29 Oct 2013	MM	Ensemble methods and multi-class classifiers	5.6-5.8
Unsupervised learning: Clustering and density estimation				
10	5 Nov 2013	MM	K-means and hierarchical clustering	8.1-8.3 + 8.5-7
11	12 Nov 2013	MM	Mixture models and association mining	9.2.2 + 6.1-6.3
12	19 Nov 2013	MM	Density estimation and anomaly detection	10.1-10.4
Machine learning and data modelling in practice				
13	26 Nov 2013	MM	Putting it all together: Summary and overview	
14	3 Dec 2013	MM	Project presentation	

Today's exercise:

- **Find a dataset to analyze throughout the course:**

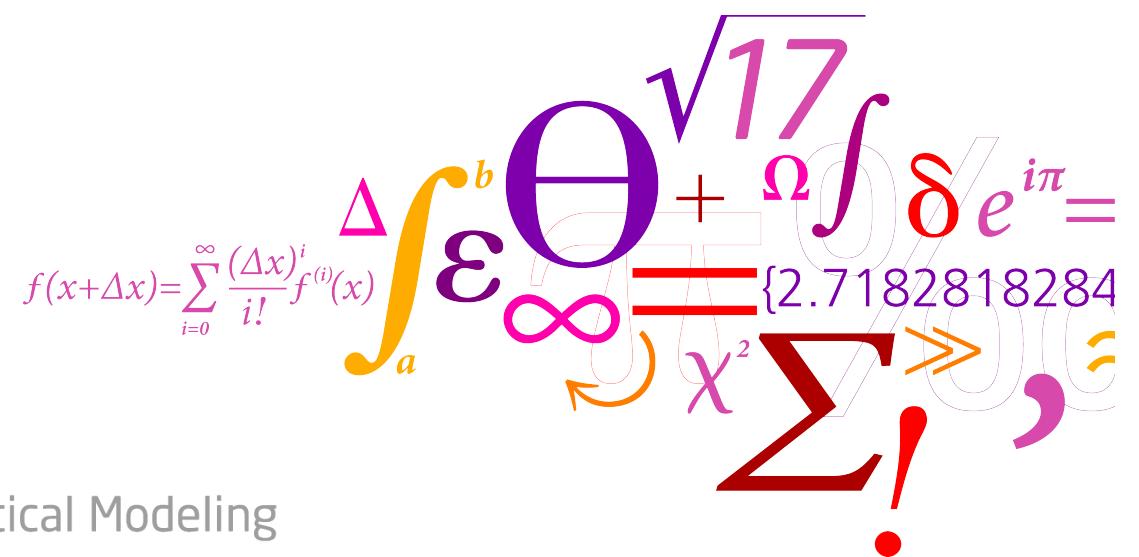
Each group will find a dataset of their own. Either your own dataset, a dataset you find yourself or for instance taken from one of the resources given in the guideline "FindingADatasetForTheReport.pdf" in the folder Project Descriptions on Campusnet. The 3 group reports will be based on the dataset that in turn will be analyzed by the various approaches taught during the course. Once you have found a dataset you need to have the dataset approved by me for the course.

- **Deadline for finding and having the dataset approved: 17th September**

- **Familiarize yourself with Matlab, Python or R:**

Todays exercise is a brush up course on either Matlab, Python or R targeted for those not familiar with these programming language. We recommend you use Matlab unless you are much more familiar with Python or R.

02450 Introduction to machine learning and data modeling

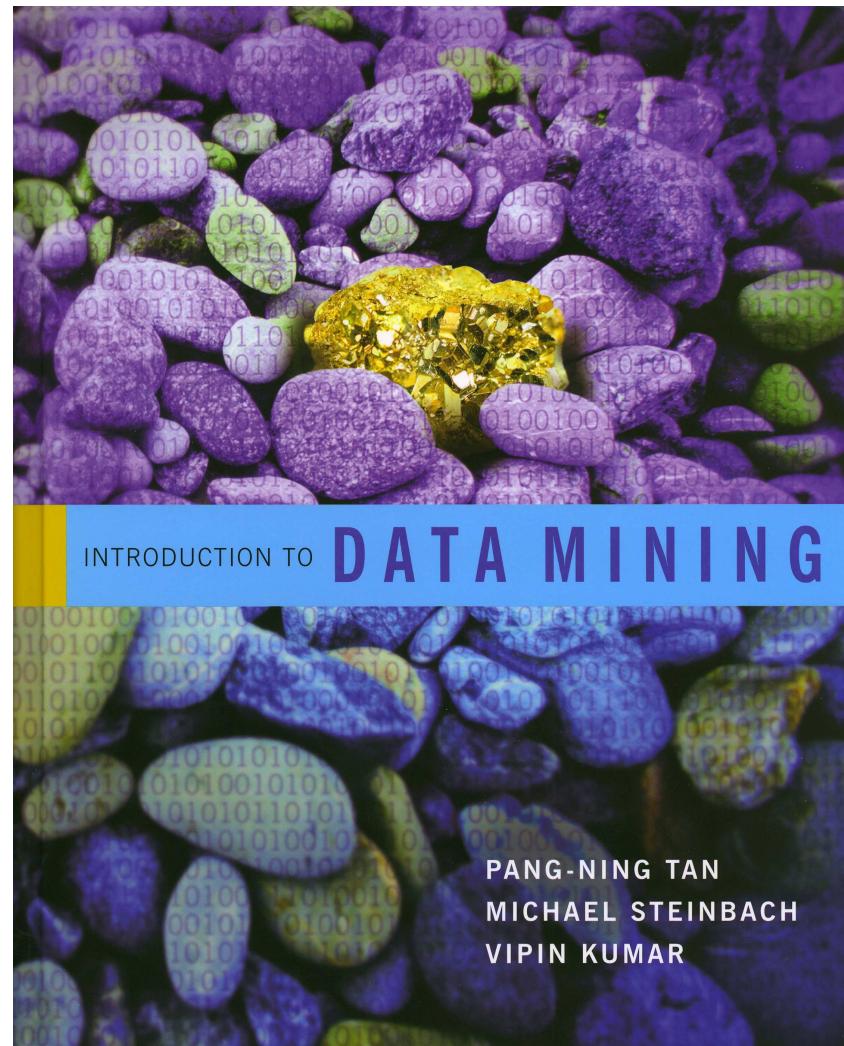
$$f(x+\Delta x) = \sum_{i=0}^{\infty} \frac{(\Delta x)^i}{i!} f^{(i)}(x)$$


The collage includes the following mathematical symbols and numbers:

- $\Theta^{\sqrt{17}}$
- $\Omega \int \delta e^{i\pi} =$
- $\{2.7182818284$
- $\Sigma \gg ,$
- χ^2
- $\infty =$
- \sum
- $\Delta \int_a^b \epsilon$

Reading material

Tan, Steinbach and Kumar
"Introduction to Data Mining"
Section 2.1-2.3 + B1 (+ A)



Group(s) of the day:
Thordis Kristjansdottir
Gudrun Svana Hilmarsdottir
Lasse Regin Nielsen
Benjamin Holm Glaas

Lecture schedule

1. Introduction
(Tan 1.1-1.4)

Data: Feature extraction and visualization

2. Data and feature extraction
(Tan 2.1-2.3 + (A) + B.1)

3. Measures of similarity and summary statistics
(Tan 2.4 + 3.1-3.2 + C1-C2)

4. Data visualization
(Tan 3.3)

Supervised learning: Classification and regression

5. Decision trees and linear regression
(Tan 4.1-4.3 + D)

6. Overfitting and performance evaluation
(Tan 4.4-4.6)

7. Nearest neighbor, naive Bayes, and artificial neural networks
(Tan 5.2-5.4)

8. Ensemble methods and multi class classifiers
(Tan 5.6-5.8)

Unsupervised learning: Clustering and density est.

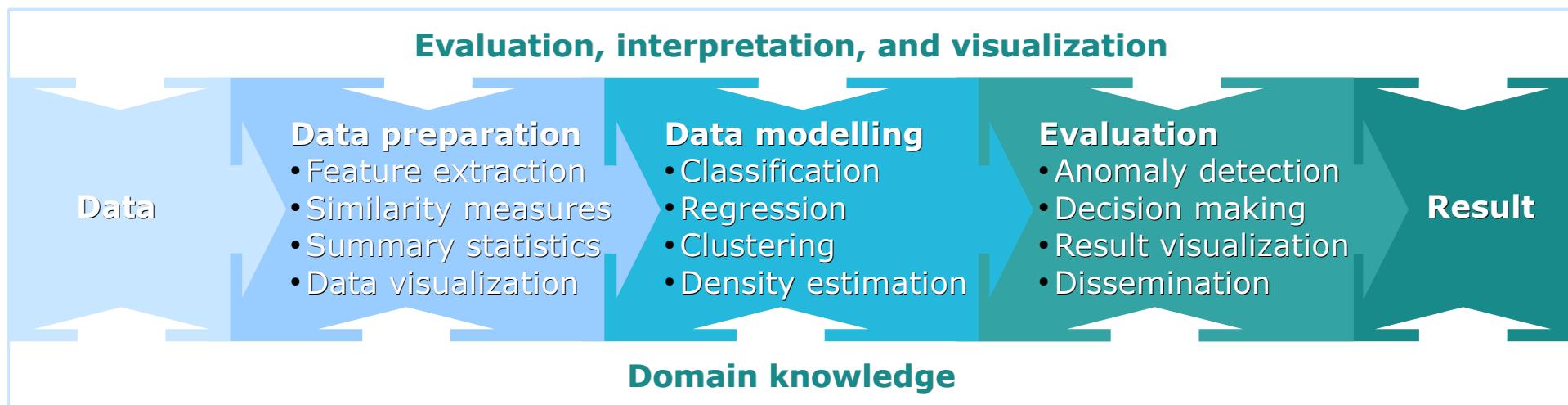
9. K-means and hierarchical clustering
(Tan 8.1-8.3)
10. Mixture models and association mining
(Tan 9.2.2 + 6.1-6.3)

11. Density estimation and anomaly detection
(Tan 10.1-10.4)

Machine learning and data modelling in practice

12. Putting it all together: Summary and overview
13. Mini project

Data modeling framework



Todays learning objectives:

Understand types of data, their attributes and data issues.

Be able to apply principal component analysis for data visualization and feature extraction.

Today we will enter the Matrix!



What is data?

- Collection of data objects and their attributes
- An attribute is a property or characteristic of an object
 - Also known as variable, field, characteristic, or feature
- Collection of attributes describe an object
 - Also known as record, point, case, sample, entity, or instance

Attributes

ID	Age	Gender	Name
1	31	F	Alex
2	24	M	Ben
3	52	F	Cindy
4	35	M	Dan
5	58	M	Eric
6	46	F	Fay
7	42	M	George

Data objects

Discrete / continuous attributes

- **Discrete**

- Finite (or countably infinite) set of values
- Examples:
 - Zip codes
 - Counts
 - Set of words in a collection of documents
- Often represented as integer variables

- **Continuous**

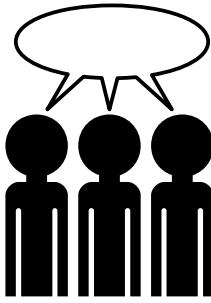
- Has real numbers as attribute values
- Examples:
 - Temperature
 - Height
 - Weight.
- Often represented as floating point variables

Types of attributes

- **Nominal:** Objects belong to a category (Equal / Not equal)
 - ID numbers
 - Eye color
 - Zip codes
- **Ordinal:** Objects can be ranked (Greater than / Less than)
 - Taste of potato chips on a scale from 1-10
 - Grades
 - Height in {short, medium, tall}
- **Interval:** Distance between objects can be measured (Addition / Subtraction)
 - Calendar dates
 - Temperature in Fahrenheit and Celcius
- **Ratio:** Zero means absence of what is measured (Multiplication / Division)
 - Length
 - Time
 - Counts
 - Temperature in Kelvin

Qualitative

Quantitative



Discussion

- **Classify the following attributes**
 - a) Military rank
 - b) Angles measured in degrees
 - c) A persons year of birth
 - d) A persons age in years
 - e) Coat check number
 - f) Distance from center of campus
 - g) Number of patients in a hospital
 - h) Sea level

- **Discrete**
 - Finite (or countably infinite) set of values
 - **Continuous**
 - Real number
-
- **Nominal** (Equal / Not equal)
 - Objects belong to a category
 - **Ordinal** (Greater than / Less than)
 - Objects can be ranked
 - **Interval** (Addition / Subtraction)
 - Distance between objects can be measured
 - **Ratio** (Multiplication / Division)
 - Zero means absence of what is measured

Types of data sets

- **Record data**

- Collection of data objects and their attributes
 - Representation: Table

- **Relational data**

- Collection of data objects and their relation
 - Representation: Graph

- **Ordered data**

- Ordered collection of data objects
 - Representation: Sequence

Record data example: Market basket data

- Transaction data table

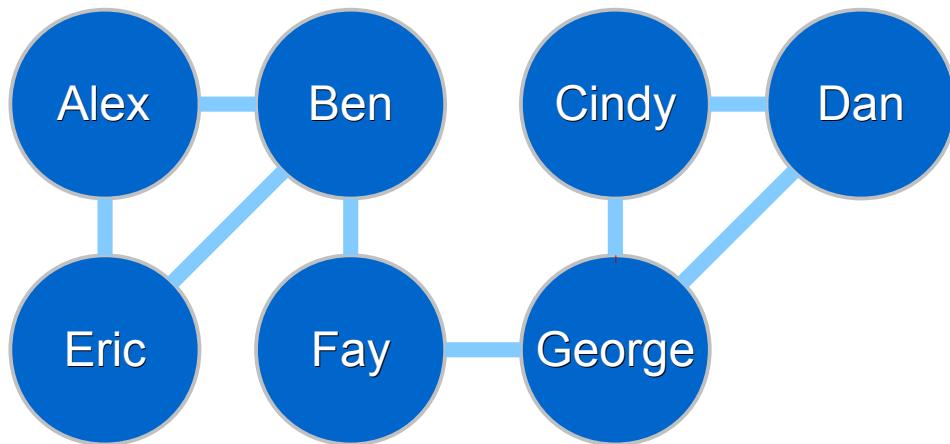
ID	Items
1	Bread, Soda, Milk
2	Beer, Bread
3	Beer, Soda, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Soda, Diaper, Milk

- Matrix

ID	Bread	Soda	Milk	Beer	Diaper
1	1	1	1	0	0
2	1	0	0	1	0
3	0	1	1	1	1
4	1	0	1	1	1
5	0	1	1	0	1

Relational data example: Who knows who?

- Graph

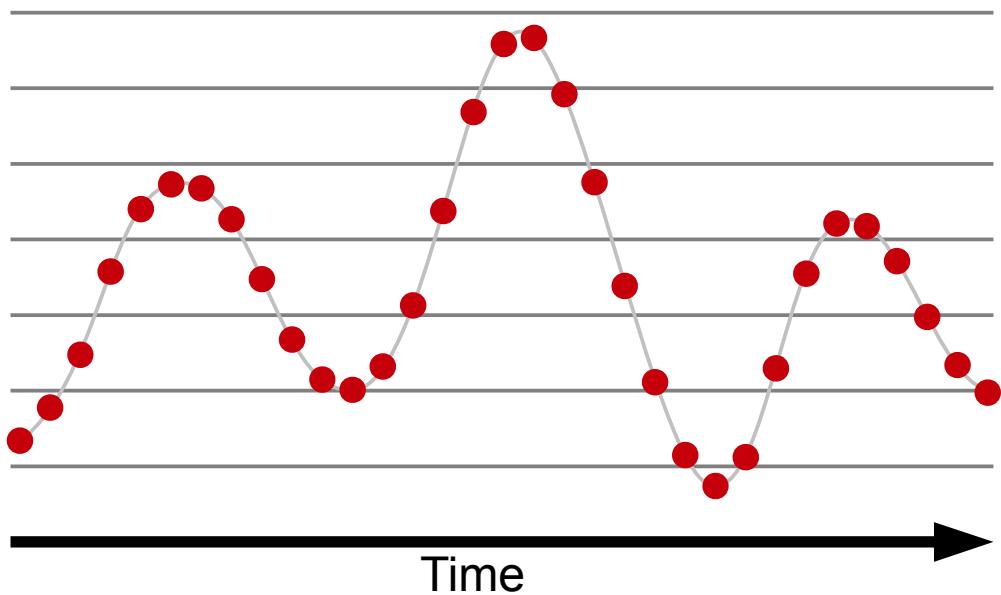


- Matrix

	A	B	C	D	E	F	G
A	0	1	0	0	1	0	0
B	1	0	0	0	1	1	0
C	0	0	0	1	0	0	1
D	0	0	1	0	0	0	1
E	1	1	0	0	0	0	0
F	0	1	0	0	0	0	1
G	0	0	1	1	0	1	0

Ordered data example: Time series

- Sequence



- Matrix

Time	Value
0	1.3
1	1.8
2	2.5
3	3.6
4	4.4
5	4.7
6	4.6
7	4.3
8	2.4
9	2.1
10	2.0
11	2.3
12	3.1

Data quality

- **Data is of high quality if they**
 - Are fit for their intended use
 - Correctly represent the phenomena they correspond to
- **Examples of quality problems**
 - Noise
 - Outliers
 - Missing values



Noise

- **Definition**

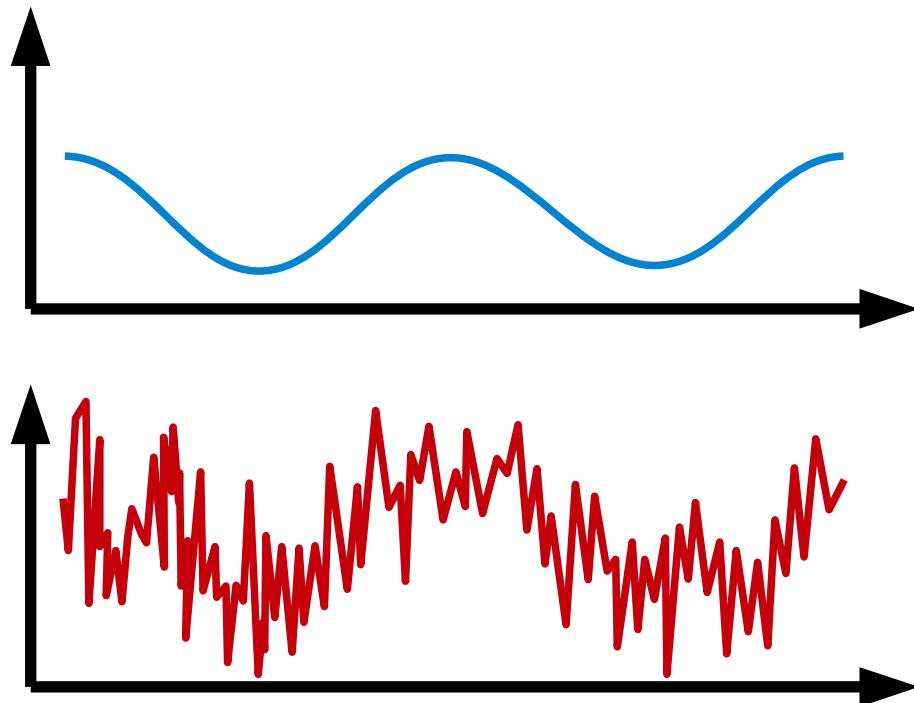
- Unwanted perturbation to a signal
- Unwanted data

- **Reasons for noise**

- Fundamental limits in measurement accuracy
- Interference from other signals
- Measurement of attributes not related to the data modeling task

- **Handling noise**

- Exclude noisy attributes
- Remove noise by filtering
- Include a model of the noise



Outliers

- **Definition**

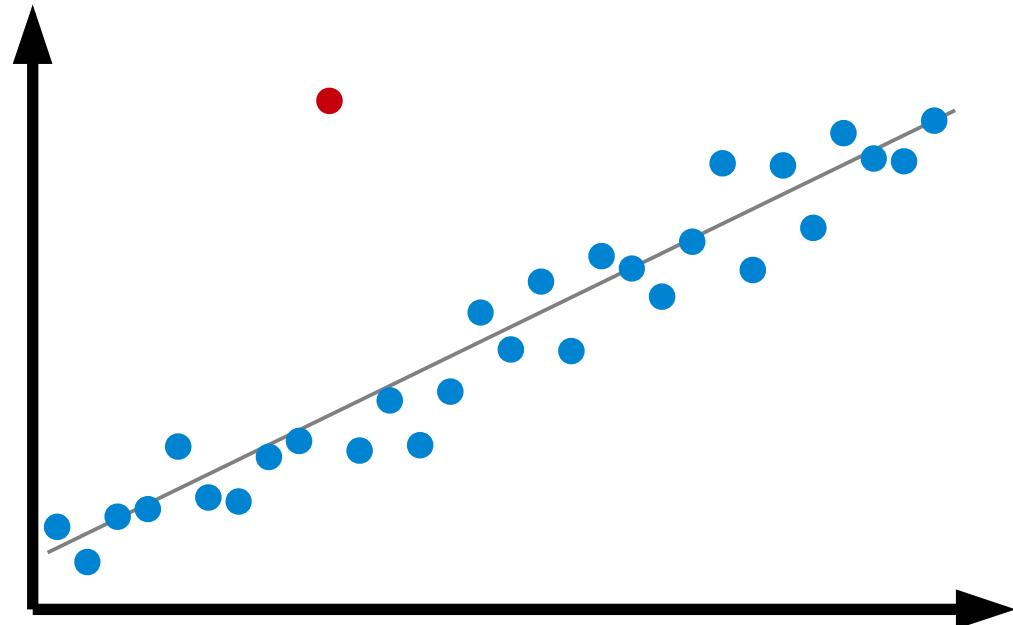
- Data objects which are significantly different from most others

- **Reasons for outliers**

- Measurement error
 - Natural property of data

- **Handling outliers**

- Identify outliers
 - Exclude anomalous outliers
 - Model the outliers



Missing values

- **Definition**

- No value is stored for an attribute in a data object

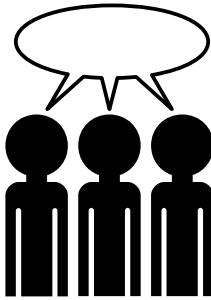
- **Reasons for missing values**

- Information is not collected
 - People decline to give their age
- Attribute is not applicable
 - Annual income is not applicable to children

- **Handling missing values**

- Eliminate data objects
- Estimate missing values
- Ignore the missing value in analysis
- Replace with an average value

ID	Age	Gender	Name
1	31	F	Alex
2	(?)	M	Ben
3	52	F	Cindy
4	35	(?)	Dan
5	(?)	M	Eric
6	(?)	F	Fay
7	42	M	(?)



Discussion

- A group of people were asked to write how many children they have
 - Their response was this

3 1 ~~NONE~~ 2 7 3 15 0 1 3 2 ~~zero~~ *

- A research assistant typed the results into a table
 - His table looked like this

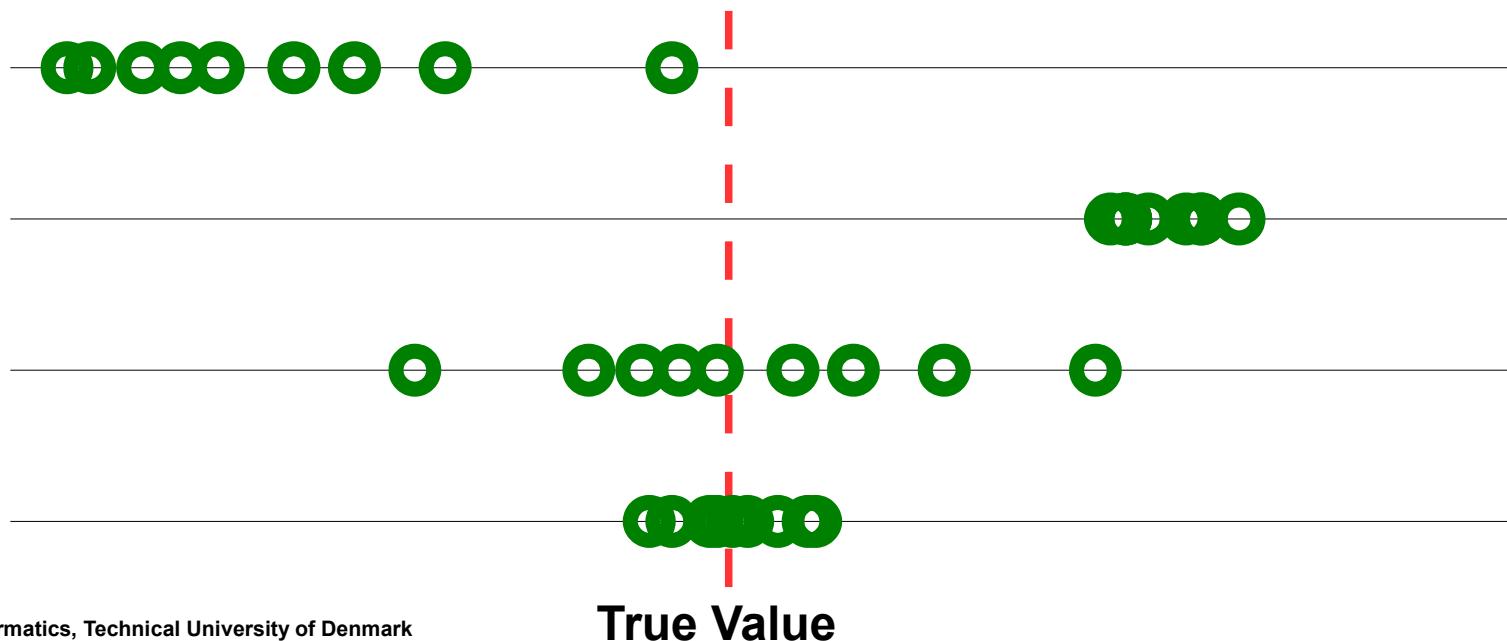
Children	3	1	0	2	7	5	15	0	1	3	-2	0	0	0	1
----------	---	---	---	---	---	---	----	---	---	---	----	---	---	---	---

- Are there any data quality issues?
 - Noise?
 - Outliers?
 - Missing values?
- Why have these issues occurred, and how should they be handled?

Precision, Bias and Accuracy

Assume we make repeated measurements of the same underlying quantity and use this set of values to calculate a mean value (average) that serves as our estimate of the true value.

- **Definition 2.3 (Precision):** The closeness of repeated measurements (of the same quantity) to one another (often measured by standard deviation)
- **Definition 2.4 (Bias):** A systematic variation of measurements from the quantity being measured.
- **Definition 2.5 (Accuracy):** The closeness of measurements to the true value of the quantity being measured.



Data preprocessing and dimensionality reduction

- **Aggregation**

- Combining several attributes into a single attribute

- **Sampling**

- Selecting a representative subset of data points

- **Dimensionality reduction**

- Project data to a low-dimensional subspace

- **Feature subset selection**

- Choose a subset of attributes

- **Feature extraction**

- Create new features from existing attributes

- **Discretization and binarization**

- Reduce continuous attributes to discrete

- **Attribute transformation**

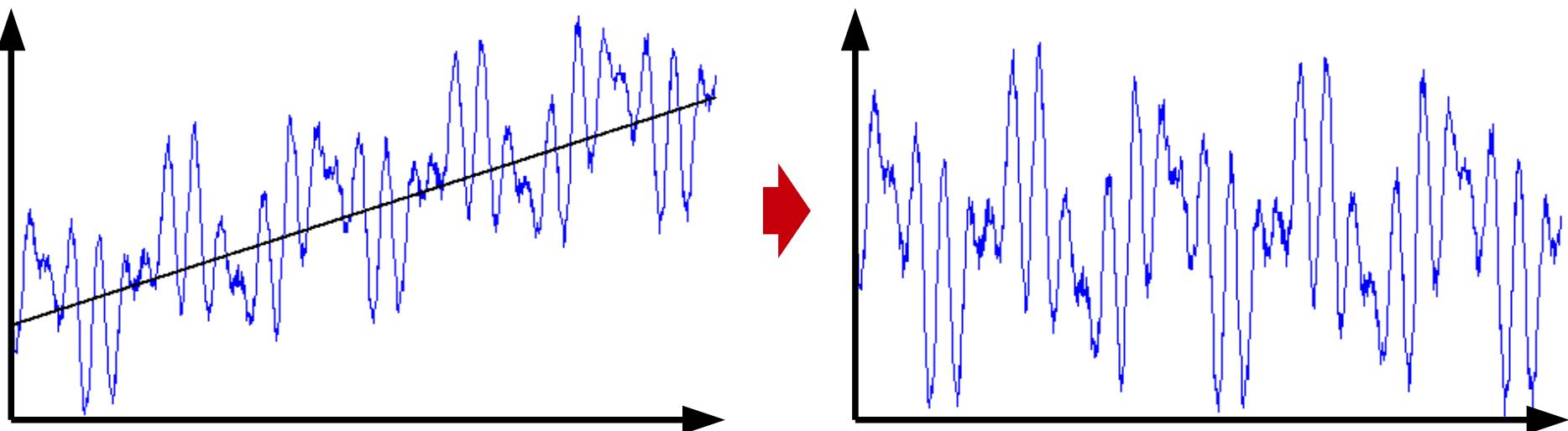
- Apply a fixed transformation to an attribute

Filtering

- Eliminating, suppressing, or attenuating certain aspects of the data
 - Noise removal in audio signals
 - Elimination of common words in text documents
 - Removal of background in images
 - Removal of examples which are corrupted
 - De-trending data (if it is not stationary)

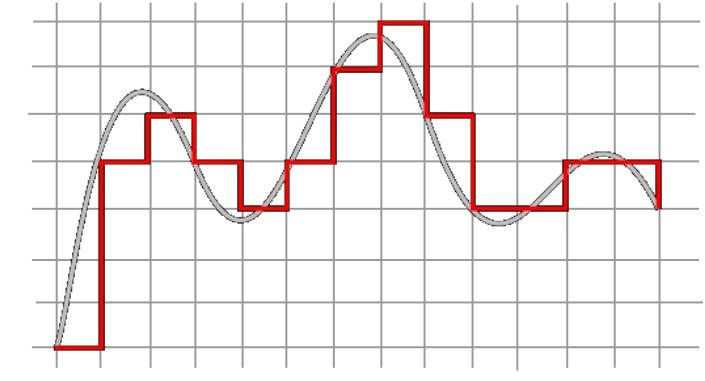
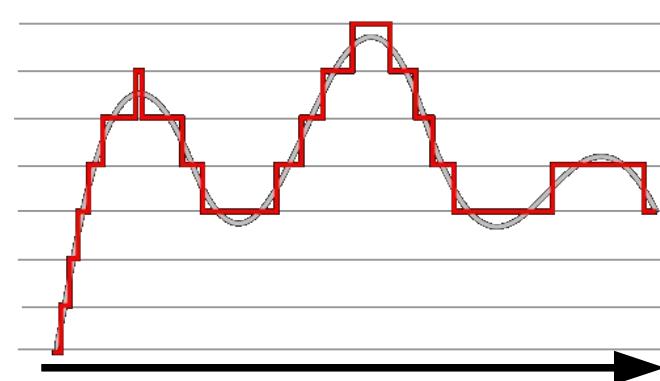
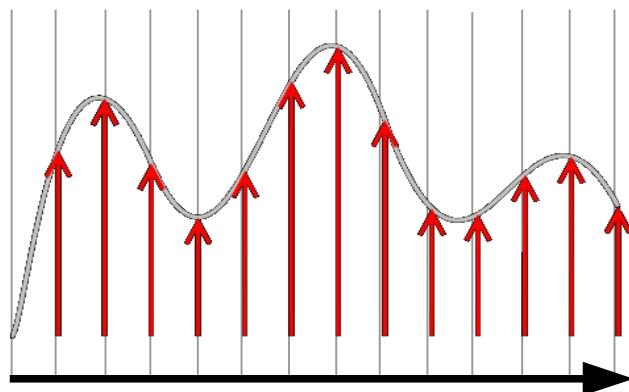


Example of de-trending data



From Analog to Digital

- Data are often **analog** and needs to be converted to a **digital** format



- Some data are “born” digital
 - On/off sensors
 - Questionnaires
 - User ratings



Chemical sensor data

- Example: The nano-nose



	A	B	C	D	E	F	G	H	I	J	K	L
1	Nanonose			A	B	C	D	E	F	G	H	
2	Sample type	Concentration										
3	Water	9200		95,5	21		6	11,94231	64,13462	21,49856	5,56784	1,174135
4	Water	9200		94,5	17		5	5,484615	63,20577	19,65856	4,968	1,883444
5	Water	9200		92	16		3	11,05769	62,58654	19,81312	5,19248	0,564835
6	Water	4600		53	7,5		2,5	3,538462	35,16346	6,876207	1,641724	0,144654
7	Water	4600		51	7,5		2,5	4,865385	34,05769	6,757241	1,613966	0,059663
8	Water	4600		50	8		2,5	3,980769	33,61538	6,773103	1,776552	0,075509
9	Water	2300		27,5	4		1,5	2,2	18,35577	2,798333	0,5635	0,030383
10	Water	2300		27,5	4,5		1,5	2,2	17,25	2,629667	0,854833	0
11	Water	2300		27	4		1,5	3,1	17,47115	2,645	0,624833	0
12	Water	1150		13,5	1,5		0	1,769231	9,730769	1,114063	0,495938	0
13	Water	1150		13	2		0,5	2,653846	10,61538	1,078125	0,452813	0
14	Water	1150		13,25	2,5		0,25	2,653846	8,846154	0,424063	0,115	0
15	Water	575		4,5	0,8		0	0,442208	2,432692	0,651667	0,529	0
16	Water	575		4	0,8		0	0,884615	3,671154	0,406333	0,348833	0
17	Water	575		3,7	0,5		0	1,326923	3,759615	0,329667	0,253	0
18	Water	288		1,5	0		0	0	1,107697	0,132324	0,144	0

Bag of words

- First three sentences on **wikipedia.org**
 - The bag-of-words model is a simplifying assumption used in natural language processing and information retrieval
 - In this model, a text (such as a sentence or a document) is represented as an unordered collection of words, disregarding grammar and even word order
 - The bag-of-words model is used in some methods of document classification



Bag of words

- First three sentences on **wikipedia.org**
 - The bag-of-words model is a simplifying assumption used in natural language processing and information retrieval
 - In this model, a text (such as a sentence or a document) is represented as an unordered collection of words, disregarding grammar and even word order
 - The bag-of-words model is used in some methods of document classification



- We will treat **this text** as a data set and create a bag-of-words model of it



Bag of words

- Elimination of common words (so-called stop words)
 - The bag-of-words model is a simplifying assumption used in natural language processing and information retrieval
 - In this model, a text (such as a sentence or a document) is represented as an unordered collection of words, disregarding grammar and even word order
 - The bag-of-words model is used in some methods of document classification



Bag of words

- Representation as matrix

Word	Sentence		
	1	2	3
bag-of-words	1		1
model	1	1	1
simplifying	1		
assumption	1		
natural	1		
language	1		
processing	1		
information	1		
retrieval	1		
text		1	
sentence		1	
document		1	1
represented		1	
unordered		1	
collection		1	
words		1	
disregarding		1	
grammar		1	
word		1	
order		1	
methods			1
classification			1

Bag of words

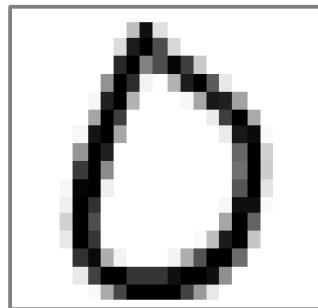
- Stemming

Word	Sentence		
	1	2	3
bag-of-word*	1		1
model*	1	1	1
simplif*	1		
assum*	1		
natural*	1		
languag*	1		
process*	1		
information*	1		
retriev*	1		
text*		1	
sentence*		1	
document*		1	1
represent*		1	
unorder*		1	
collect*		1	
word*		2	
disregard*		1	
grammar*		1	
order*		1	
method*			1
classif*			1

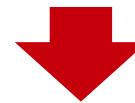
Image representation

- Example: Handwritten digits

- Preprocessing
 - Digitalization
 - Centering
 - Rotation
 - Scaling



$$M_0 = \begin{bmatrix} 0 & & & & & 0 \\ 0 & 0.3 & 1 & 0.2 & 0 & 0 \\ | & & & & & | \\ 0 & & & & & 0 \end{bmatrix}$$



- Vectorization

$$\mathbf{x}_0 = [0 \ 0.3 \ 1 \ 0.2 \ 0]^\top$$

- Matrix representation of data set

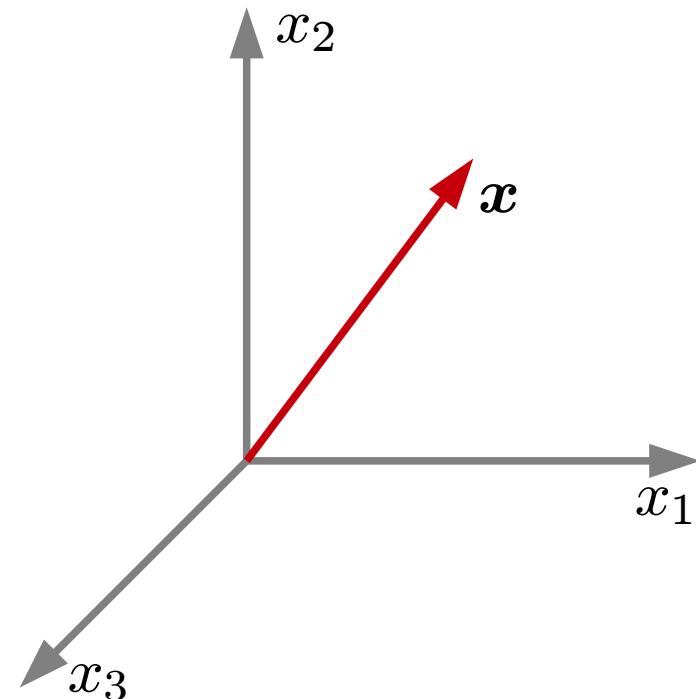
$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_0 & \mathbf{x}_1 & \dots & \mathbf{x}_N \end{bmatrix}$$



Vector space representation

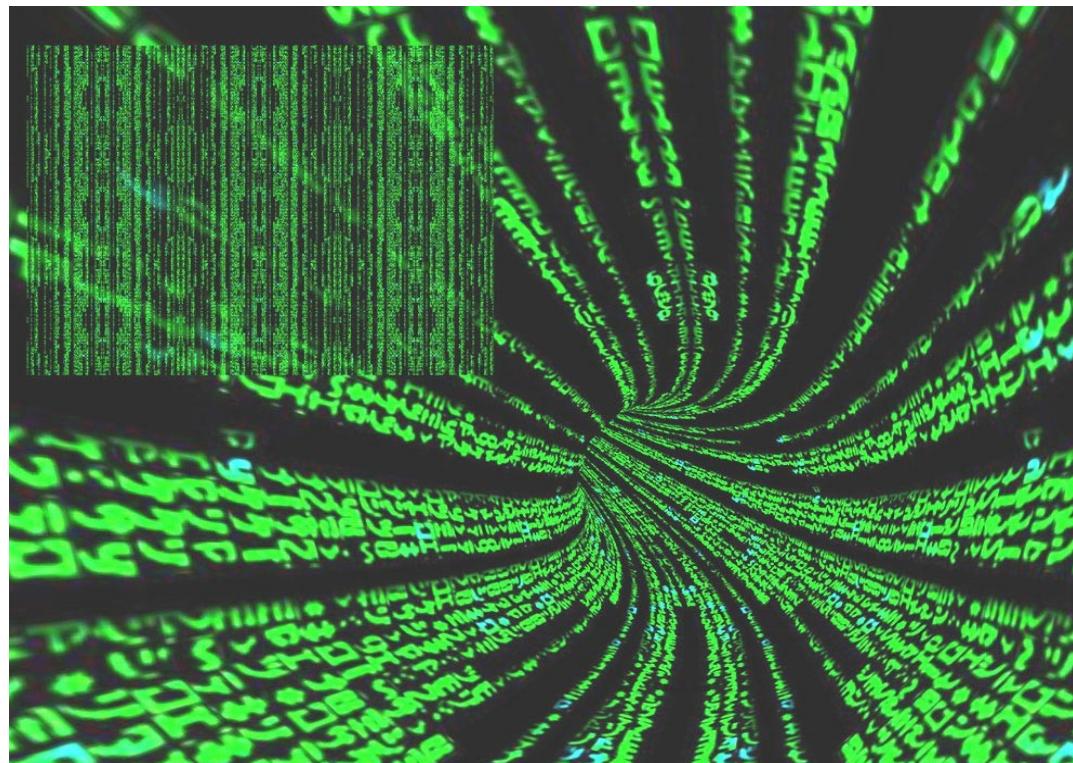
- All these data objects have a vector space representation

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_N \end{bmatrix}$$



Vectors and matrices

- Data represented as vectors and matrices
 - Linear algebra useful for manipulating and analyzing data
- We will derive the Singular Value Decomposition (SVD)
 - A highlight of linear algebra
 - Very important for data visualization



Vectors and matrices

- Common matrix notation

$A, \bar{A}, \overline{\overline{A}}$

$$A = \begin{bmatrix} a_{1,1} & \cdots & a_{1,M} \\ | & \searrow & | \\ a_{N,1} & \cdots & a_{N,M} \end{bmatrix} \in \mathbb{R}^{N \times M}$$

- Common vector notation

$x, \boldsymbol{x}, \overline{x}, \vec{x}$

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ | \\ x_N \end{bmatrix} \in \mathbb{R}^N$$

Matrix Multiplication

- Two matrices can be multiplied $\mathbf{AB} = \mathbf{C}$
 - if the number of columns in the first equals the number of rows in the second

$$\begin{array}{c}
 \textcolor{orange}{A} \times \textcolor{orange}{B} = \textcolor{orange}{C} \\
 L \times M \quad M \times N \quad L \times N
 \end{array}$$

$\begin{bmatrix} \text{3}\times 4 \text{ matrix} & \begin{bmatrix} \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ 1 & 2 & 3 & 4 \end{bmatrix} & \begin{bmatrix} \text{4}\times 5 \text{ matrix} & \begin{bmatrix} \cdot & \cdot & \cdot & a \\ \cdot & \cdot & \cdot & b \\ \cdot & \cdot & \cdot & c \\ \cdot & \cdot & \cdot & d \end{bmatrix} & \begin{bmatrix} \text{3}\times 5 \text{ matrix} & \begin{bmatrix} \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & x_{3,4} \end{bmatrix} & \cdot \end{bmatrix} = \begin{bmatrix} \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \end{bmatrix}$

$$x_{3,4} = 1 \cdot a + 2 \cdot b + 3 \cdot c + 4 \cdot d$$

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \begin{bmatrix} 5 & 6 \\ 7 & 8 \end{bmatrix} = ?$$

Matrix transpose

- The transpose of a matrix

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 3 & 4 & 6 \\ 7 & 8 & 9 \end{bmatrix} \quad A^\top = \begin{bmatrix} 1 & 3 & 7 \\ 2 & 4 & 8 \\ 3 & 6 & 9 \end{bmatrix}$$

- Transpose of a sum

$$(A + B)^\top = A^\top + B^\top$$

- Transpose of a product

$$(AB)^\top = B^\top A^\top$$

$$(Ax)^\top y = x^\top A^\top y = x^\top (A^\top y)$$

The identity matrix

- Ones on the diagonal and zeros everywhere else

$$I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \quad I^\top = I$$

- Multiplying by the identity does not change anything

$$\begin{aligned} IA &= A \\ I_2 &= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad A = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \\ I_2 A &= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} a & b \\ c & d \end{bmatrix} = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \end{aligned}$$

Matrix inverses

- For a square matrix, the inverse satisfies

$$\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$$

- Inverse of a product of square matrices

$$(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$$

- Transpose of inverse

$$(\mathbf{A}^{-1})^\top = (\mathbf{A}^\top)^{-1}$$

Norms

- The norm of a vector is usually written as

$$\|\mathbf{x}\|_p = \left(\sum_{n=1}^N |x_n|^p \right)^{1/p}$$

- Of particular interest is the 1-norm and the 2-norm

$$\|\mathbf{x}\|_1 = \sum_{n=1}^N |x_n|$$

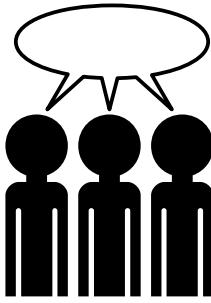
$$\|\mathbf{x}\|_2 = \sqrt{\sum_{n=1}^N |x_n|^2} = \sqrt{\mathbf{x}^\top \mathbf{x}}$$

- The Frobenius norm of a matrix

$$\|\mathbf{X}\|_F^2 = \sum_{i,j} x_{i,j}^2 = \text{trace}(\mathbf{X} \mathbf{X}^\top) = \text{trace}(\mathbf{X}^\top \mathbf{X})$$

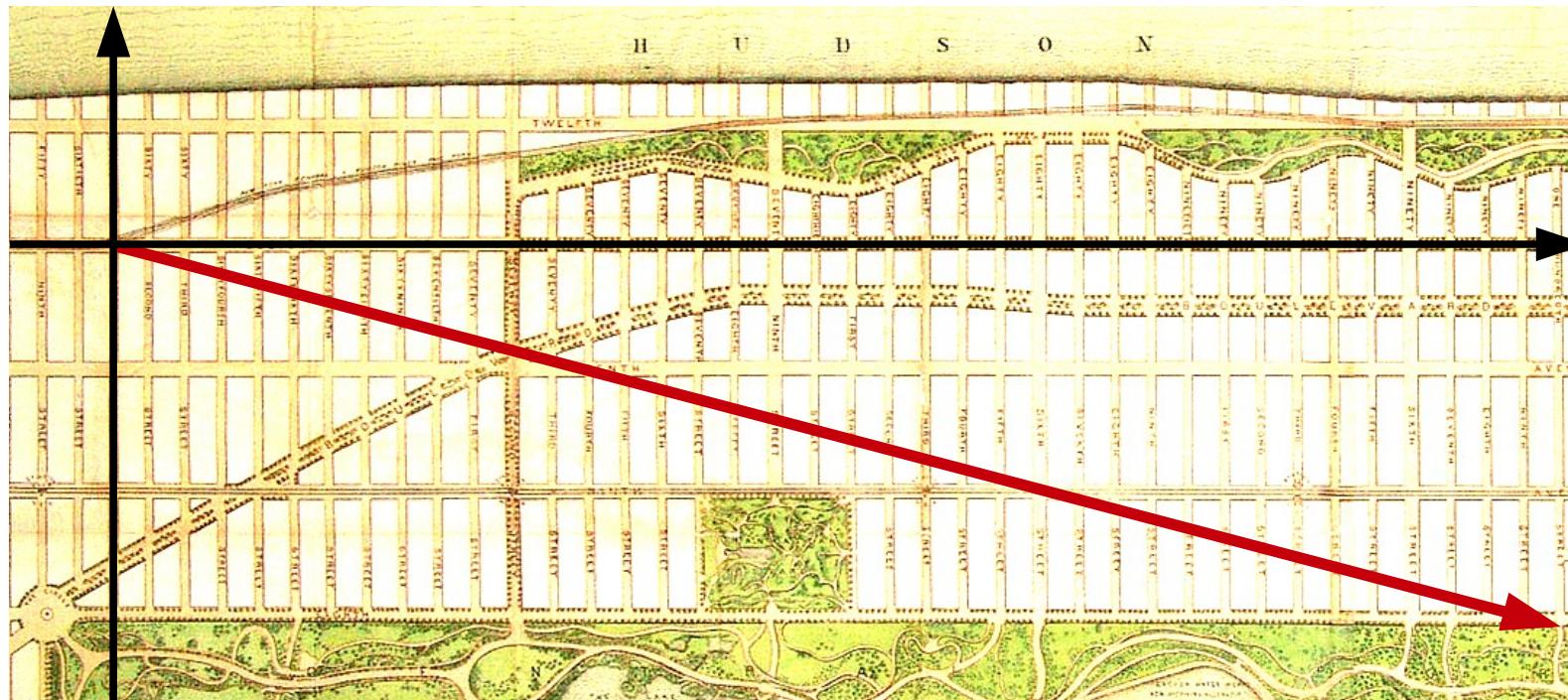
Where trace takes the sum of the diagonal elements, i.e.

$$\text{trace}(\mathbf{A}) = \sum_i a_{i,i}$$



Discussion

- I want to go from **61st and West End Ave.** to **110th and Central Park West**
- I have created a coordinate system and a vector to my destination



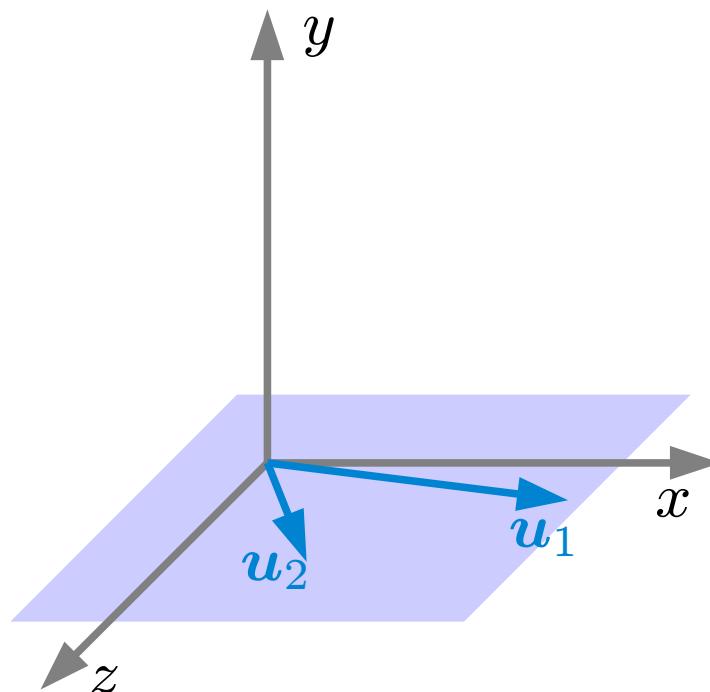
- Which vector norm should I use to measure the length of my trip?

Vector spaces

- Vector spaces can be of arbitrary size
- Typically defined using a matrix of basis vectors
- The basis vectors must be **linearly independent**

$$Ux = 0 \quad \Rightarrow \quad x = 0$$

$$U = \begin{bmatrix} | & | & | \\ u_1 & u_2 & \cdots & u_N \\ | & | & | \end{bmatrix}$$



Vector spaces

- Often the vectors are taken to be **mutually orthogonal** and of **unit length**

$$\mathbf{u}_i^\top \mathbf{u}_j = 0$$

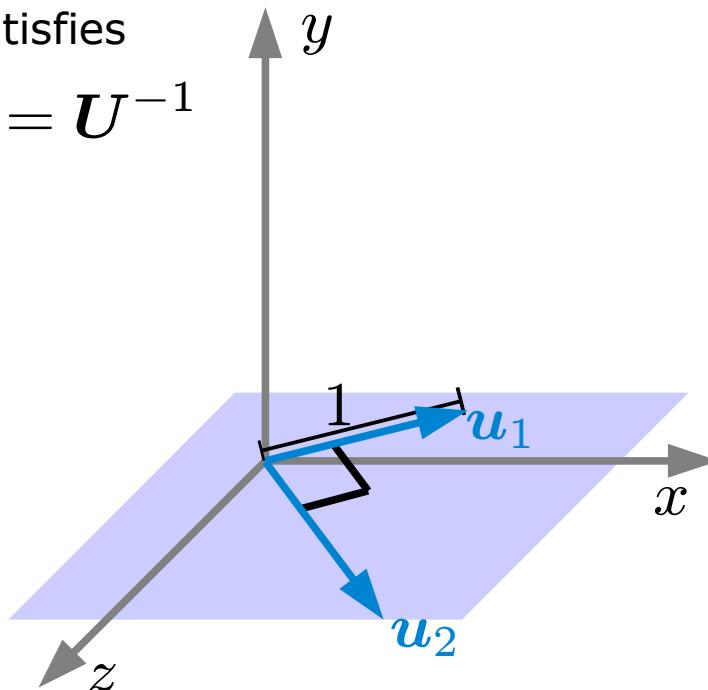
$$\|\mathbf{u}_i\|_2 = 1$$

– This defines an orthonormal basis for the vector space

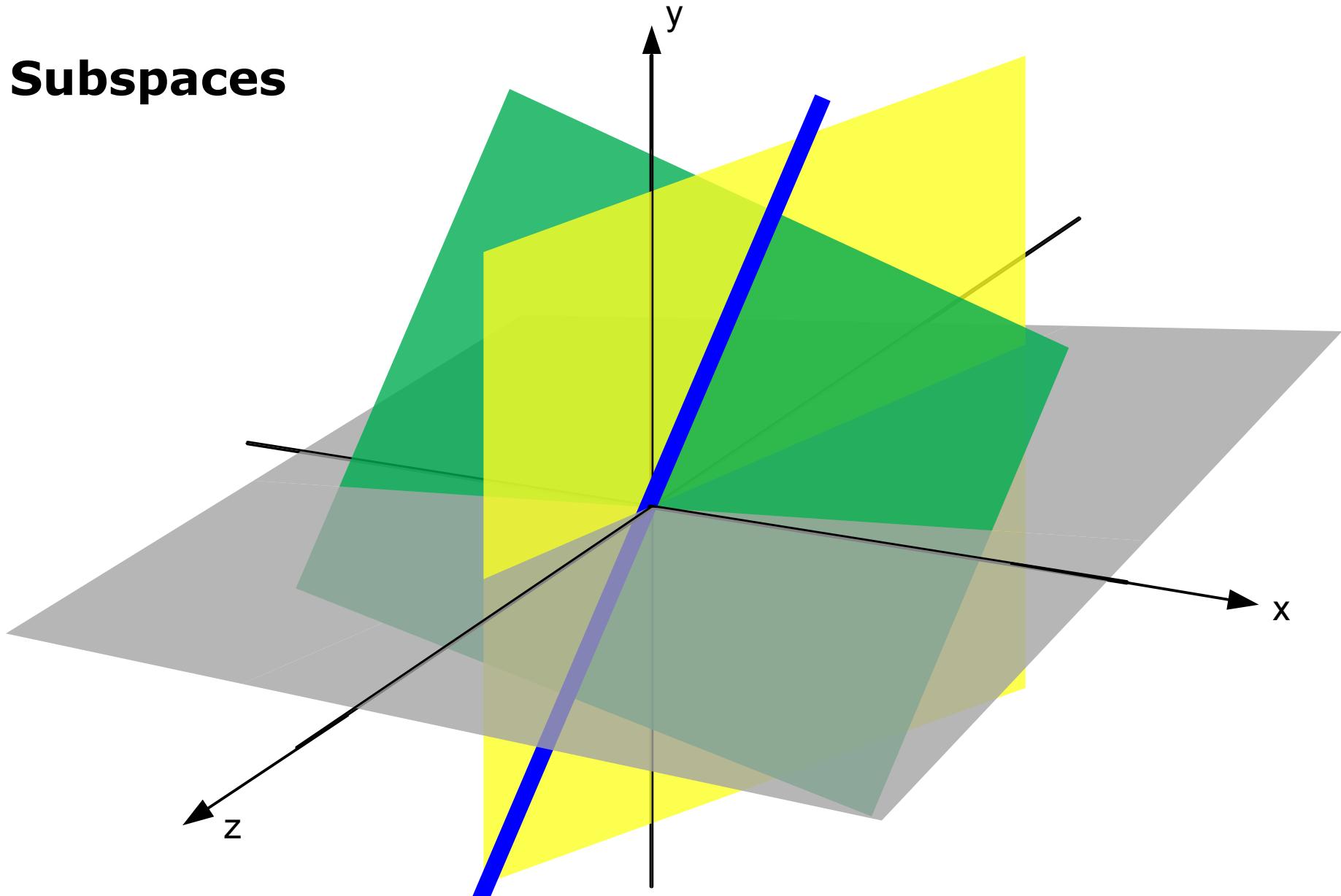
$$U = \begin{bmatrix} & & \\ \mathbf{u}_1 & \mathbf{u}_2 & \cdots & \mathbf{u}_N \\ & & \end{bmatrix}$$

- An orthonormal matrix satisfies

$$\mathbf{U}^\top \mathbf{U} = \mathbf{I}, \quad \mathbf{U}^\top = \mathbf{U}^{-1}$$

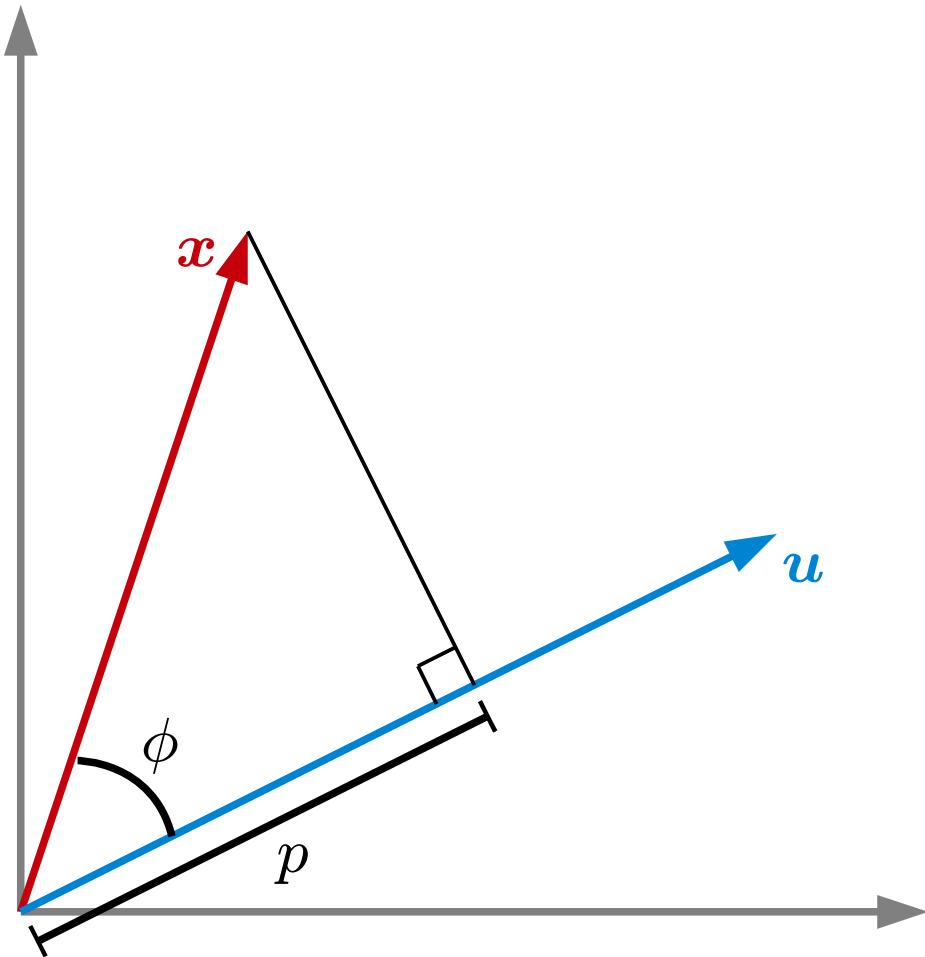


Subspaces



Projection

- Projection onto a vector



- Angle between vectors

$$\cos(\phi) = \frac{\mathbf{u}^\top \mathbf{x}}{\|\mathbf{x}\|_2 \|\mathbf{u}\|_2}$$

- Length of projection

$$p = \|\mathbf{x}\|_2 \cos(\phi) = \frac{\mathbf{u}^\top \mathbf{x}}{\|\mathbf{u}\|_2}$$

- Projection onto unit vector

$$p = \mathbf{u}^\top \mathbf{x}$$

Projection onto subspace

- **Projection onto a subspace**

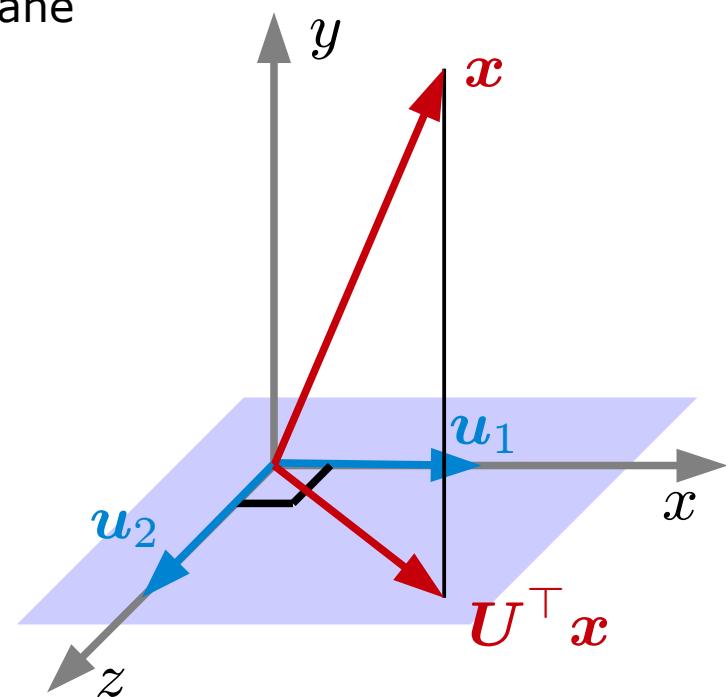
- Subspace defined by a orthonormal basis matrix
- Projection given by

$$U^\top x$$

- **Example:** Projection of 3-D vector onto the (x,z) plane

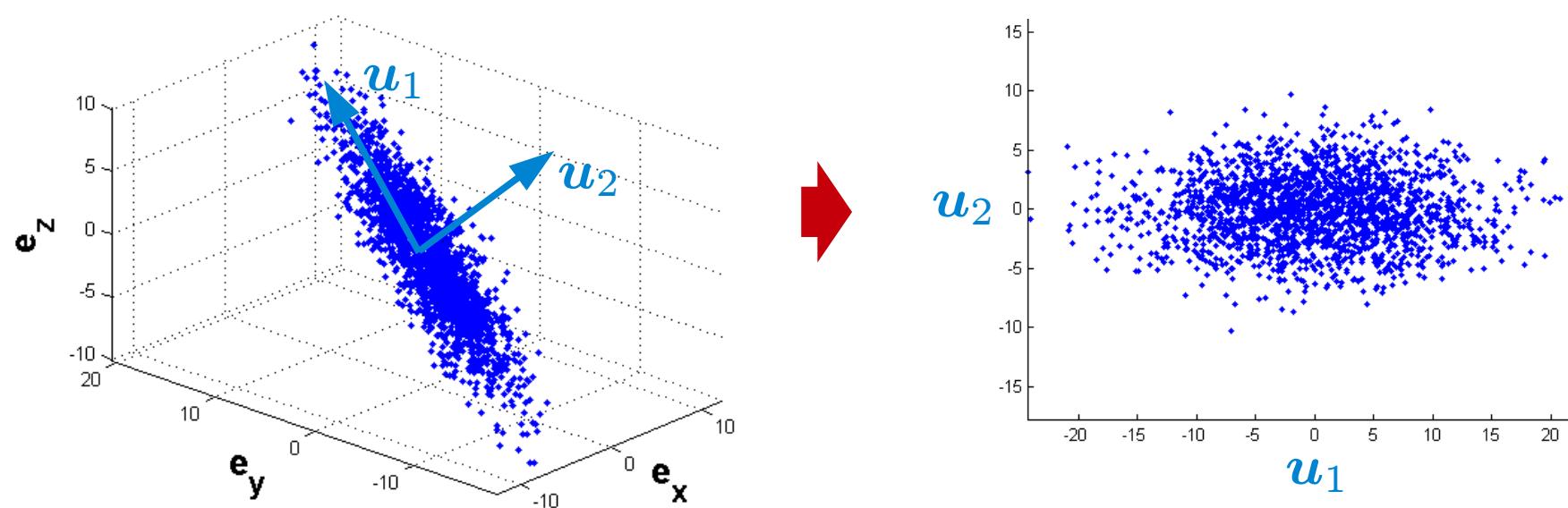
$$U = \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 1 \end{bmatrix} \quad x = \begin{bmatrix} x \\ y \\ z \end{bmatrix}$$

$$U^\top x = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} x \\ z \end{bmatrix}$$



Visualizing high dimensional data

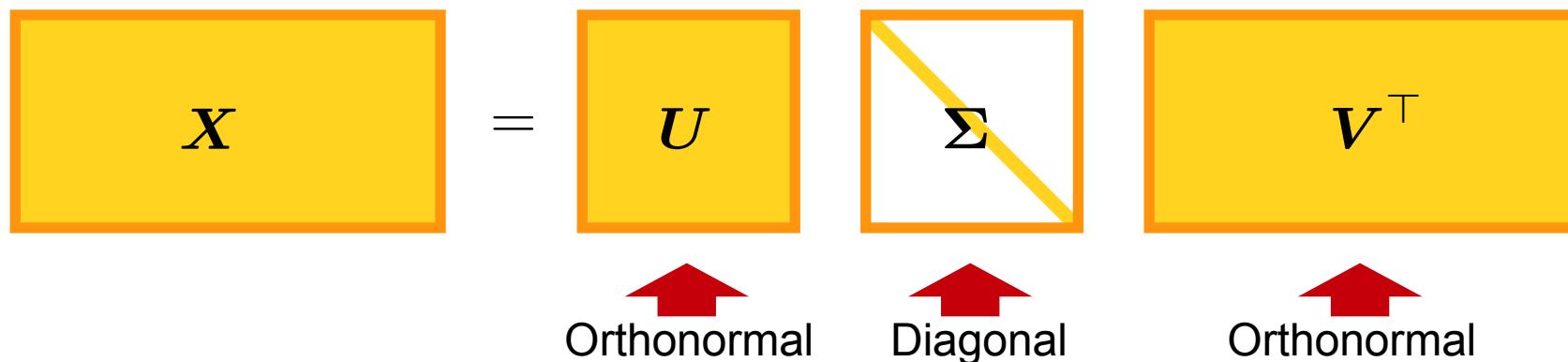
- Humans perceive well data in 2-D
 - We have a hard time visualizing high dimensional data
- We can project high dimensional data to a lower dimensional subspace
- But what is a good projection?
 - Account for as much of the variation as possible



Principal component analysis

- 1) Subtract the mean from each attribute
- 2) Apply singular value decomposition (SVD)

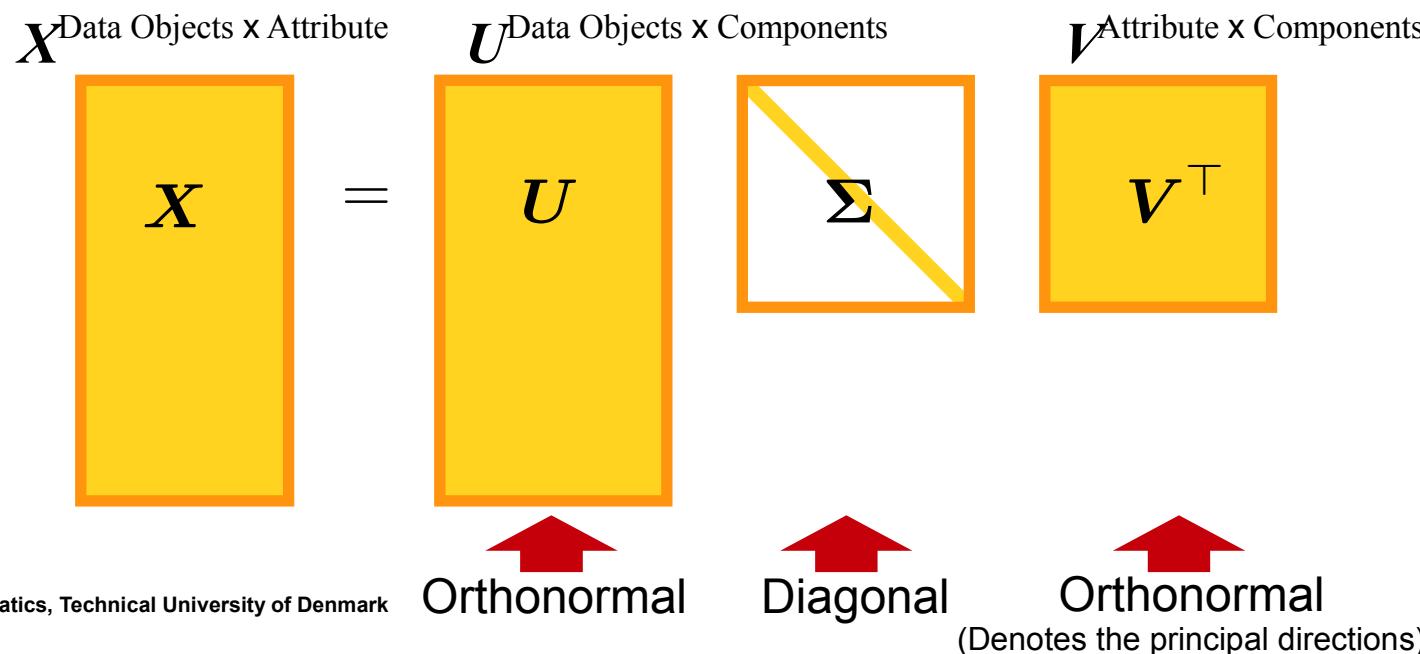
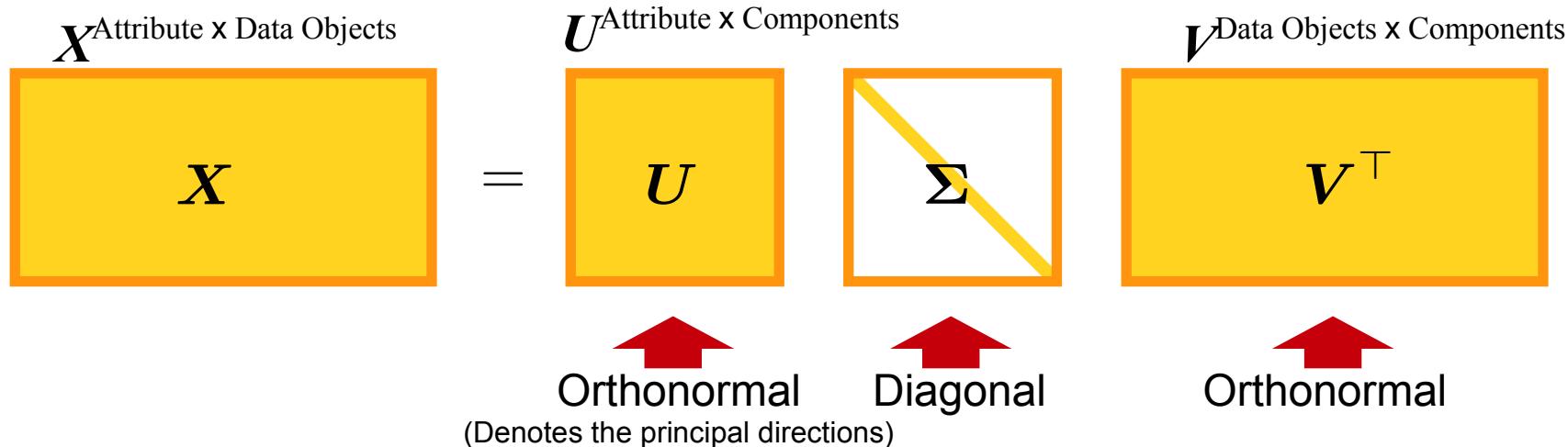
$$\mathbf{X} = \mathbf{U}\Sigma\mathbf{V}^\top$$



- Orthonormal basis

$$\mathbf{U} = \begin{bmatrix} | & | & | \\ \mathbf{u}_1 & \mathbf{u}_2 & \cdots & \mathbf{u}_N \\ | & | & | \end{bmatrix}$$

PCA of a data matrix X defined by Attribute x Data Objects vs. Data Objects x Attributes

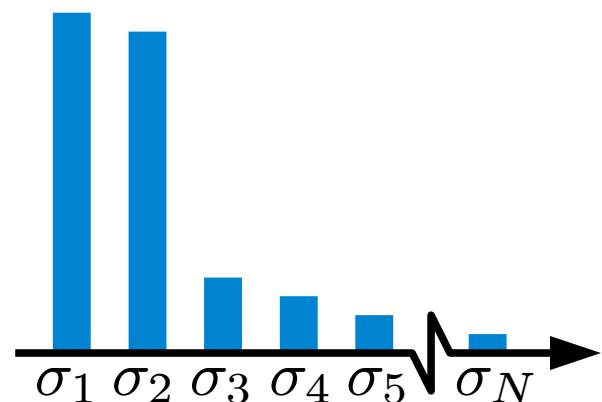


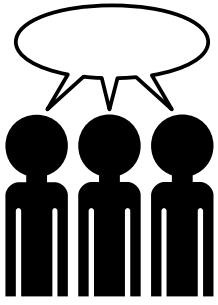
Principal component analysis

- Entries in the diagonal matrix are called **singular values**
 - They are sorted (largest first)
 - Indicate how much variability is explained by the corresponding component
 - 1st component explains most of the variability
 - 2nd component explains most of the remaining variability
 - Etc.

$$\Sigma = \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_N \end{bmatrix} \quad \sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_N$$

- Singular value spectrum





- Show that

$$\|X\|_F^2 = \sum_i \sigma_i^2$$

where

$$\sigma_i = \Sigma_{i,i}$$

Fraction of the variation in the data explained by the i^{th} principal component is given by:

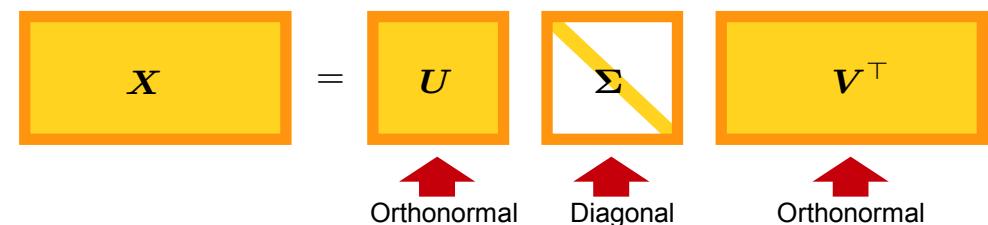
$$\frac{\sigma_i^2}{\sum_i \sigma_i^2}$$

And by the first K principal components

$$\frac{\sum_{i=1}^K \sigma_i^2}{\sum_i \sigma_i^2}$$

Hints:

$$X = U\Sigma V^\top$$



$$\|X\|_F^2 = \text{trace}(XX^\top)$$

$$\text{trace}(AB) = \text{trace}(BA)$$

Fishers Iris Data

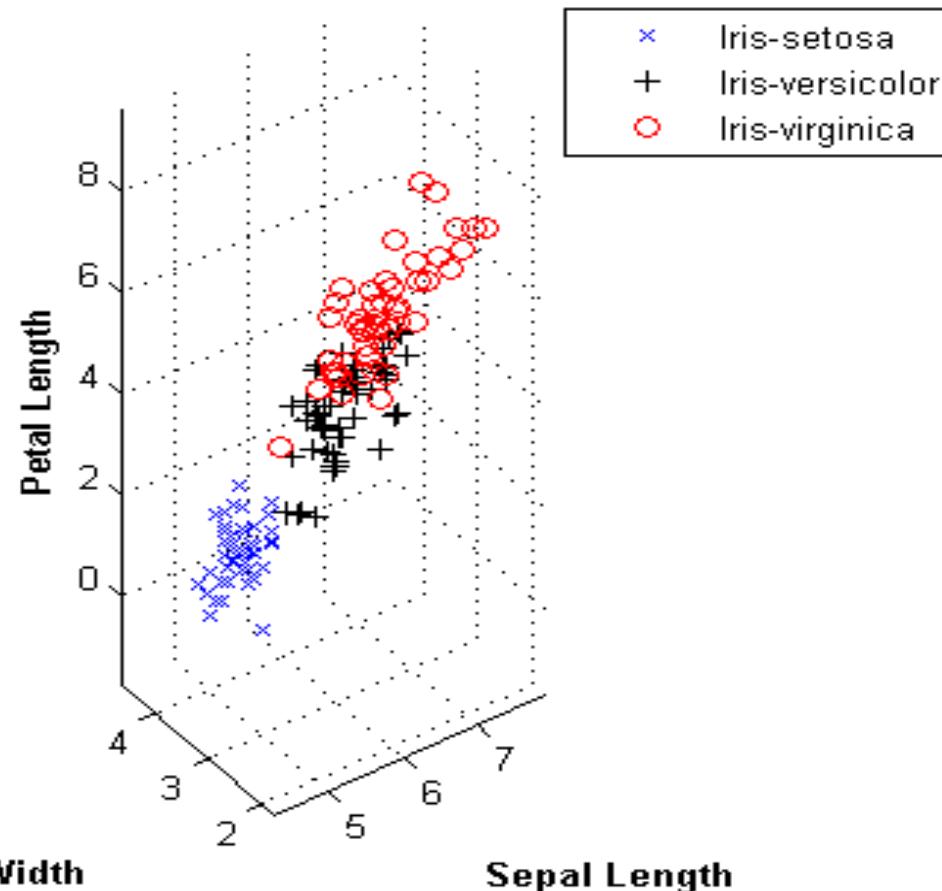


**Three types of flowers:
Iris Setosa, Iris Versicolor, Iris Virginica**

Flower ID	Attribute			
	Sepal Length	Sepal Width	Petal Length	Petal Width
1	5.1	3.5	1.4	0.2
2	4.9	3.0	1.4	0.2
3	4.7	3.2	1.3	0.2
4	4.6	3.1	1.5	0.2
.
.
150	5.9	3.0	5.1	1.8

We will presently consider the first 3 attributes, i.e. Sepal length, Sepal Width and Petal Length.

3D scatter plot of the data

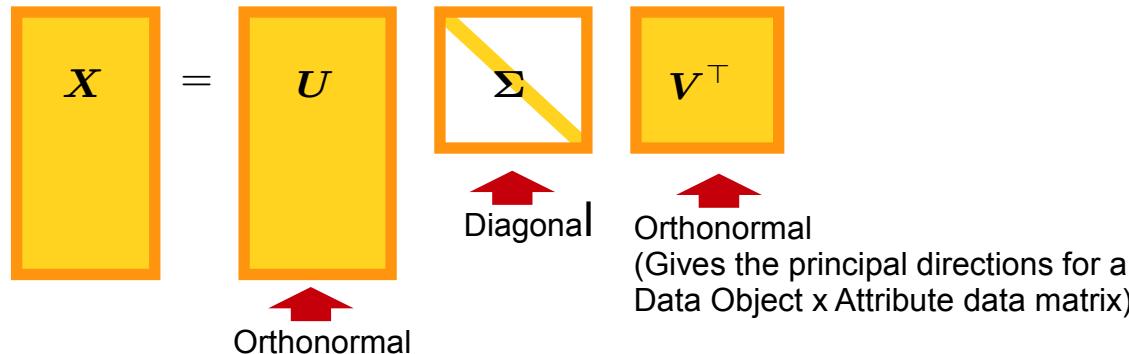


What fraction of the total variation in the data do you think the first principal component accounts for?

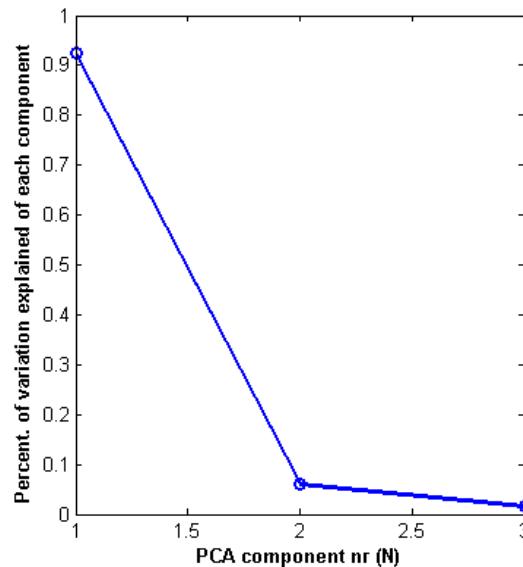
Principal Component Analysis

- 1) Subtract the mean
- 2) Apply singular value decomposition (SVD)

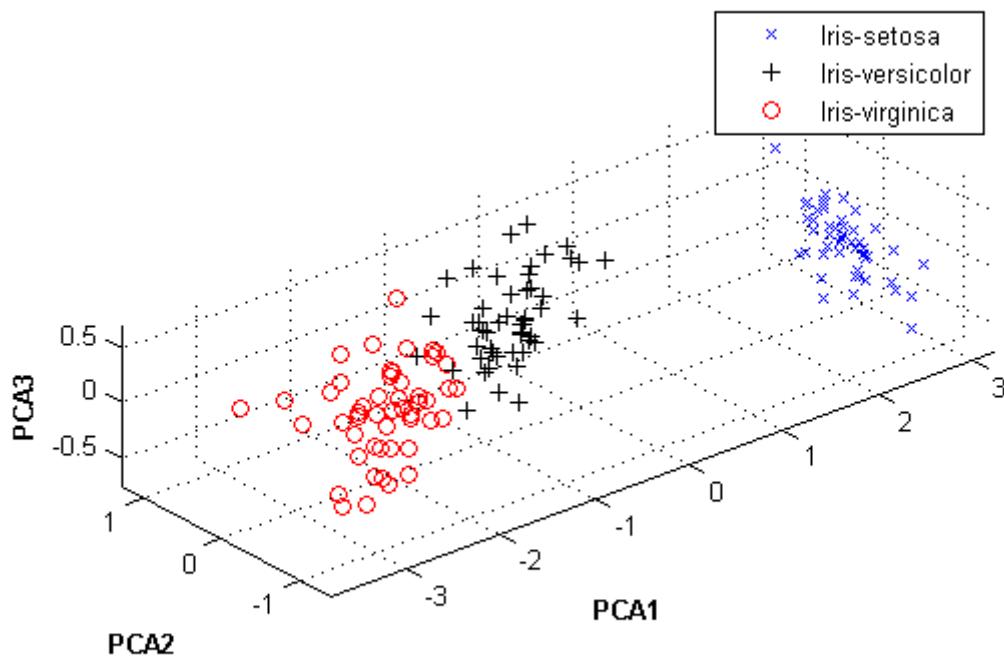
$$X = U \Sigma V^\top$$



Evaluate the singular values to determine how much of the dynamics is lost when reducing the dimensionality

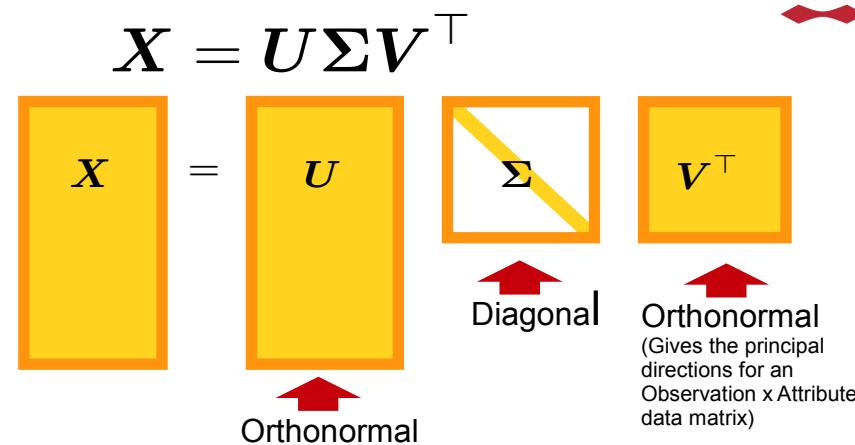


Visualizing the data projected onto the space of the principal components



The principal directions V

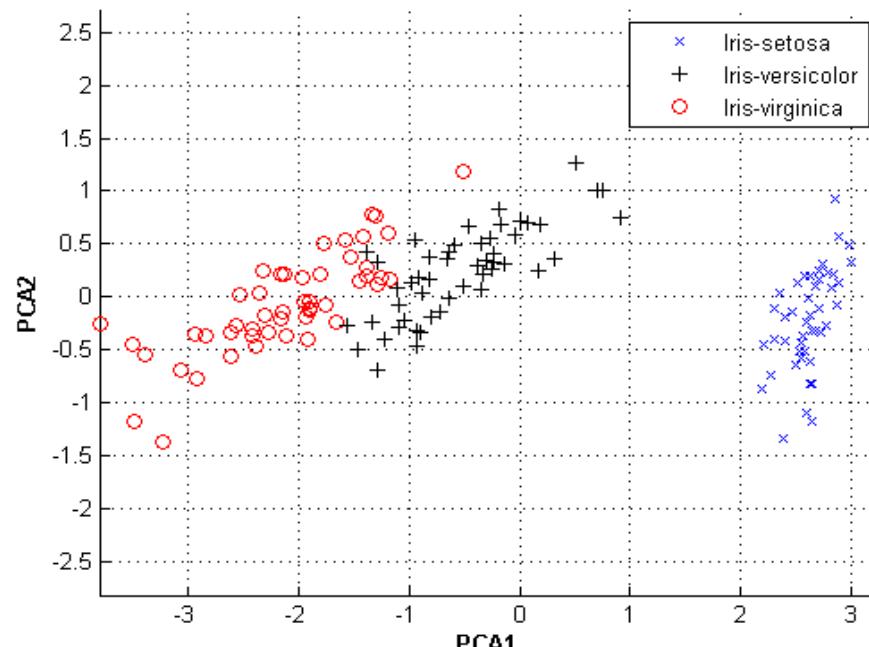
$$\begin{array}{l} \text{Sepal Length} \\ \text{Sepal Width} \\ \text{Petal Length} \end{array} \quad V = \begin{bmatrix} -0.3902 & -0.6392 & -0.6627 \\ 0.0887 & -0.7425 & 0.6640 \\ -0.9165 & 0.2003 & 0.3464 \end{bmatrix}$$



$$PCA1: p_1 = Xv_1 = u_1\sigma_1$$

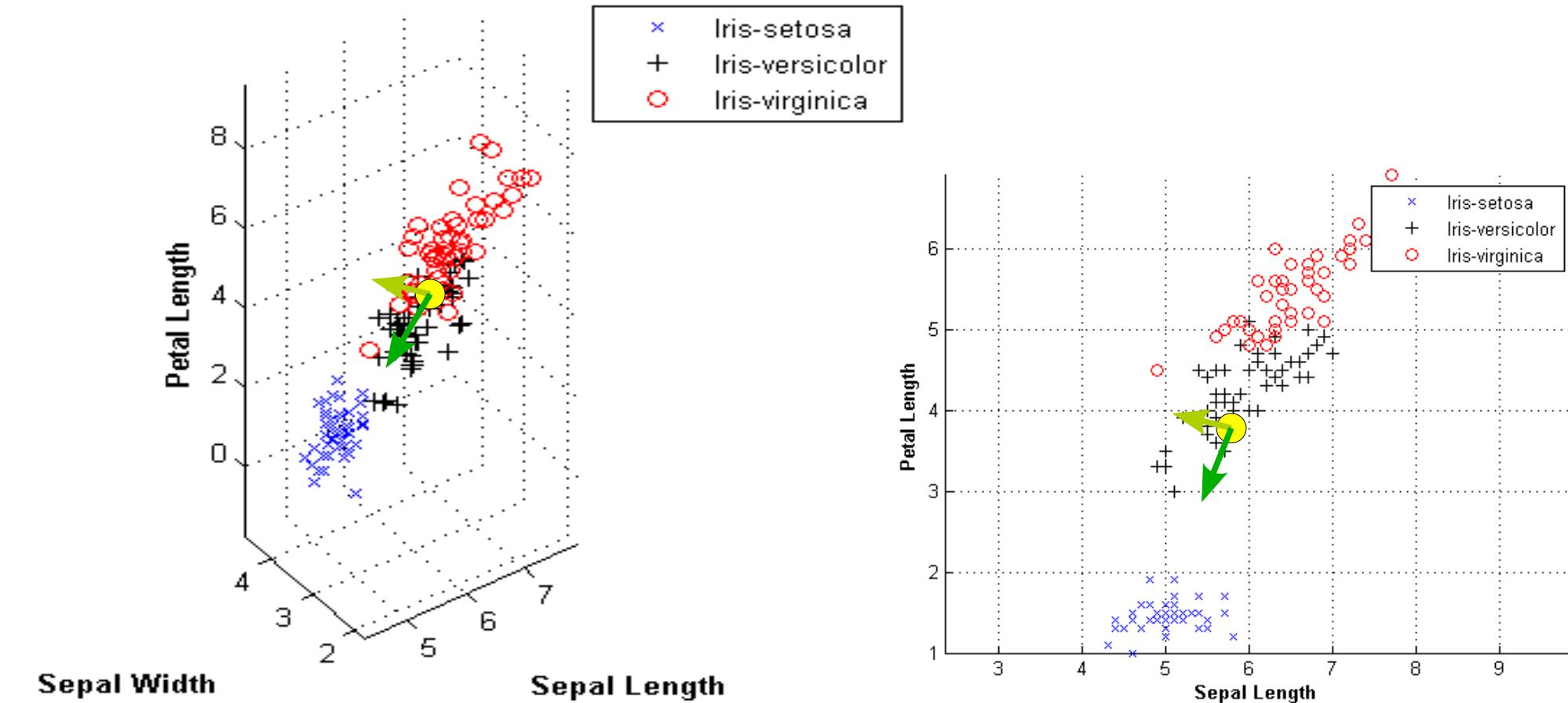
$$PCA2: p_2 = Xv_2 = u_2\sigma_2$$

$$PCA3: p_3 = Xv_3 = u_3\sigma_3$$



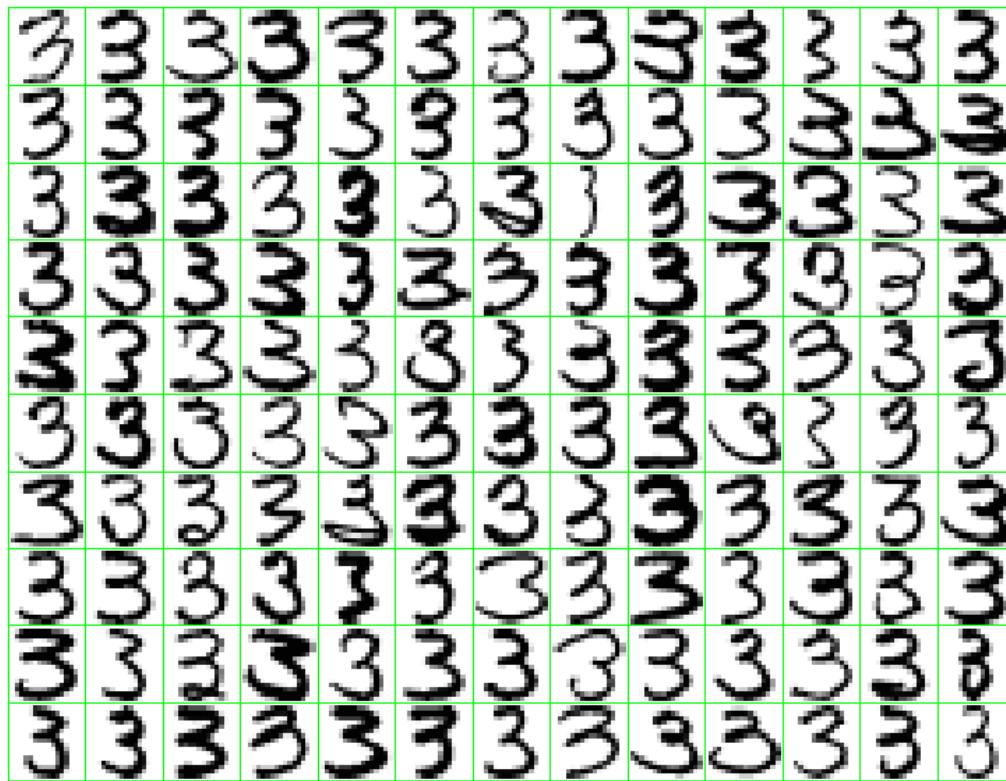
$$\mu = \begin{bmatrix} 5.8433 \\ 3.0540 \\ 3.7587 \end{bmatrix} \quad v_1 = \begin{bmatrix} -0.3902 \\ 0.0887 \\ -0.9165 \end{bmatrix} \quad v_2 = \begin{bmatrix} -0.6392 \\ -0.7425 \\ 0.2003 \end{bmatrix}$$

Sepal Length
 Sepal Width
 Petal Length

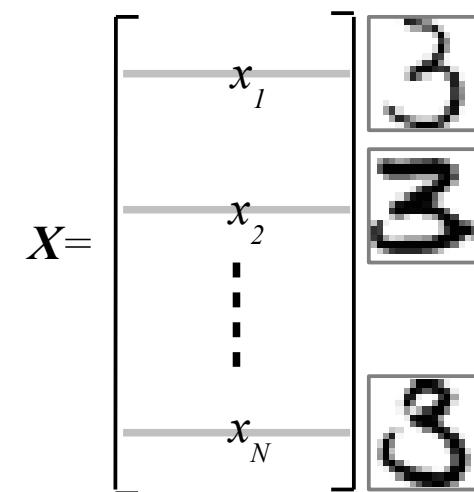


Visualization of hand written digits

- Data: Hand written 3's



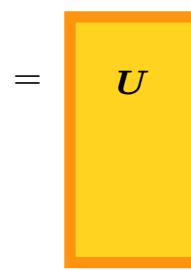
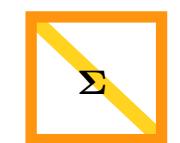
- Data matrix



Each image is 16×16 pixels forming a $M=16^2=256$ dimensional space

- Principal component analysis

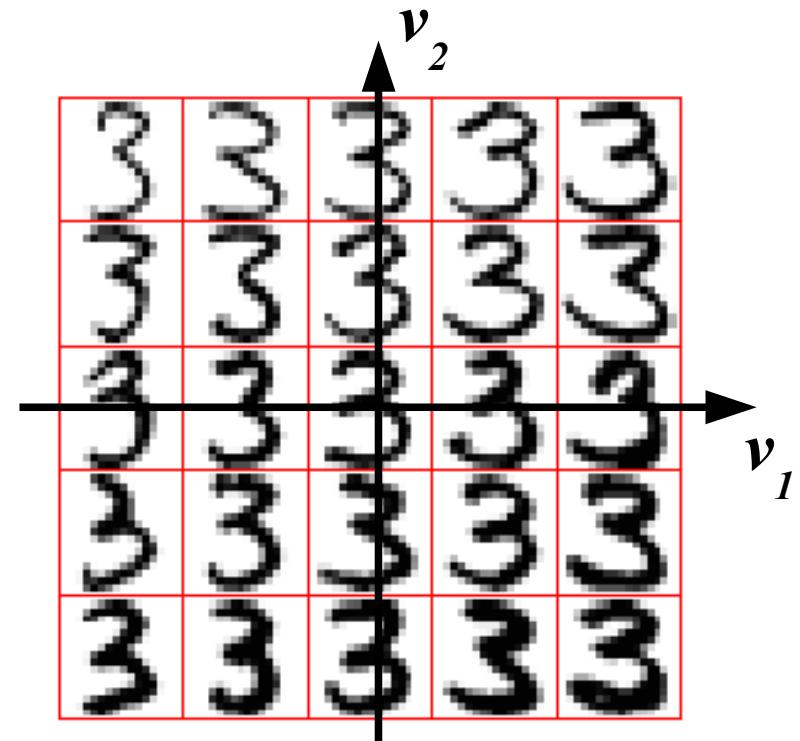
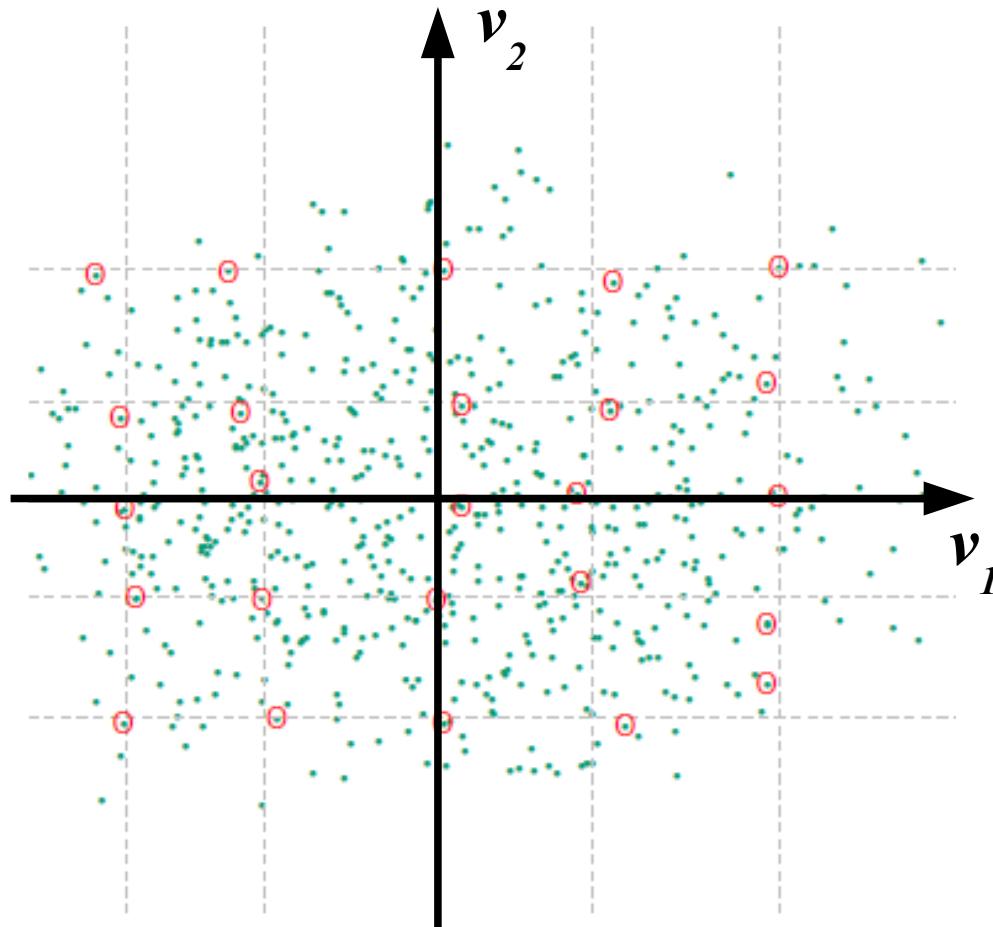
$$X = U \Sigma V^\top$$

 =
 



 ↓ Orthonormal ↓ Diagonal ↓ Orthonormal
 (Gives the principal directions for a Data Object x Attribute data matrix)

10/09/13

Visualization of hand written digits

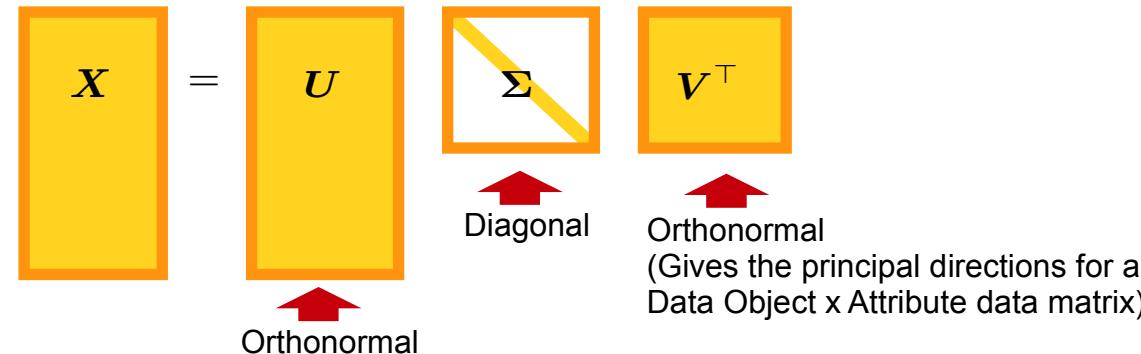


What is the dynamics captured by the first two principal components?

Data and Domain driven feature extraction

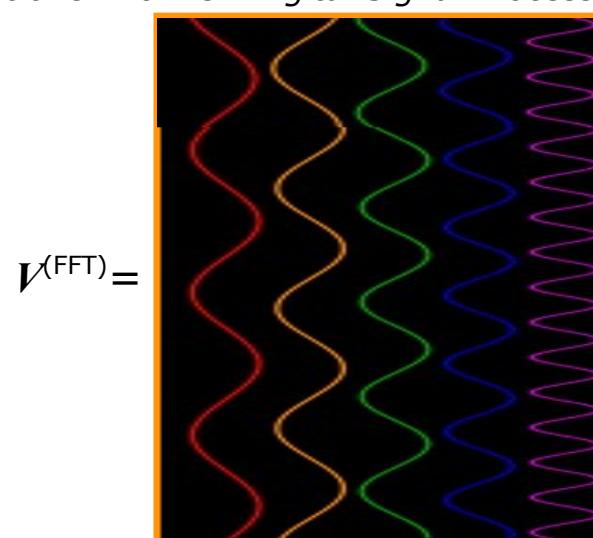
PCA is an example of a data driven approach for feature extraction

i.e., we define from data the features extracted in terms of the projections $V^{(PCA)}$ that preserve most of the variation in the data



The fourier transform is an example of a domain driven approach for feature extraction

i.e., in the analysis of sound good features are often to use spectral representations. These can be found by projecting the data using the so-called fourier transform matrix $V^{(FFT)}$ where the components are defined as specific frequencies such that the projection of the data onto these frequencies defines the extend to which these frequencies are present in the data. (you can learn much more about this in 02451 Digital Signal Processing)



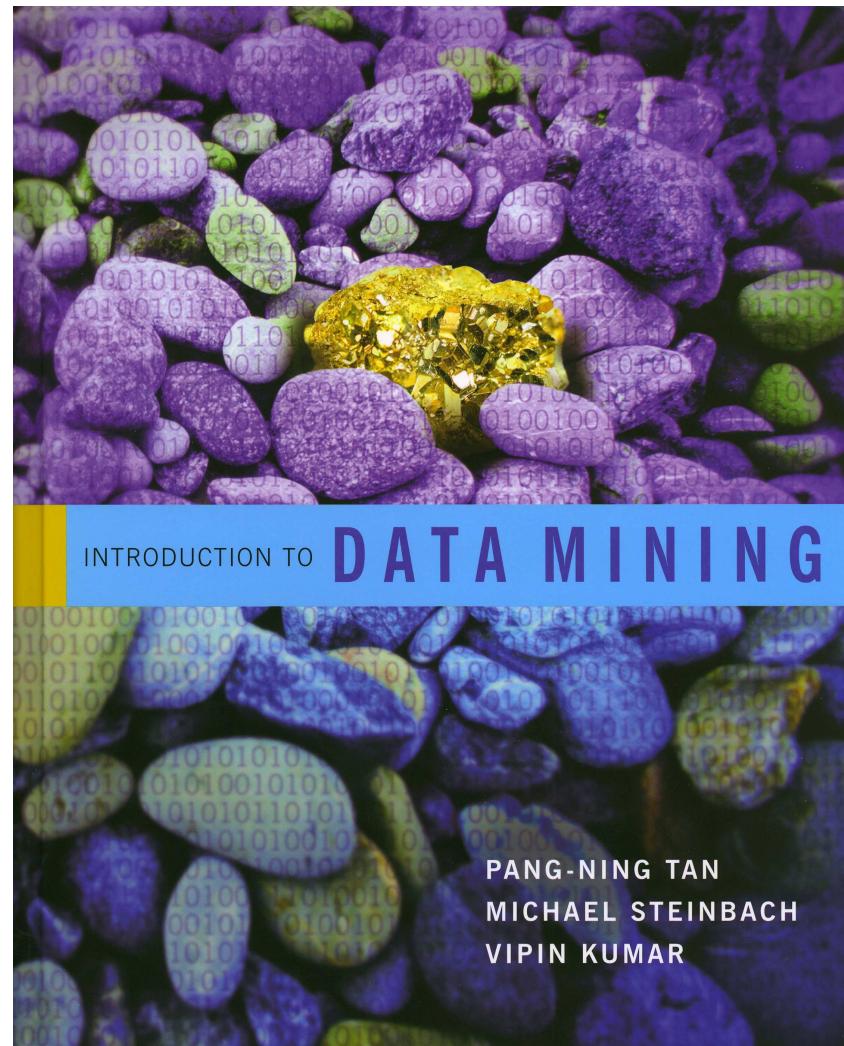
02450 Introduction to machine learning and data modeling

Reading material

Tan, Steinbach and Kumar
"Introduction to Data Mining"

Section 2.4 + 3.1-3.2 + C1-C2

Group(s) of the day:
Jan Selliah
Carsten Nilsson
Mette Vestergaard Lauridsen
Anders Vinther Olsen
Mikkel Liisborg hansen
Nanna Thorning-Schmidt



Lecture schedule

1. Introduction
(Tan 1.1-1.4)

Data: Feature extraction and visualization

2. Data and feature extraction
(Tan 2.1-2.3 + B1 (+ A))

3. Measures of similarity and summary statistics *(Tan 2.4 + 3.1-3.2 + C1-C2)*

4. Data visualization
(Tan 3.3)

Supervised learning: Classification and regression

5. Decision trees and linear regression
(Tan 4.1-4.3 + D)

6. Overfitting and performance evaluation
(Tan 4.4-4.6)

7. Nearest neighbor, naive Bayes, and artificial neural networks
(Tan 5.2-5.4)

8. Ensemble methods and multi class classifiers
(Tan 5.6-5.8)

Unsupervised learning: Clustering and density est.

9. K-means and hierarchical clustering
(Tan 8.1-8.3+8.5.7)

10. Mixture models and association mining
(Tan 9.2.2 + 6.1-6.3)

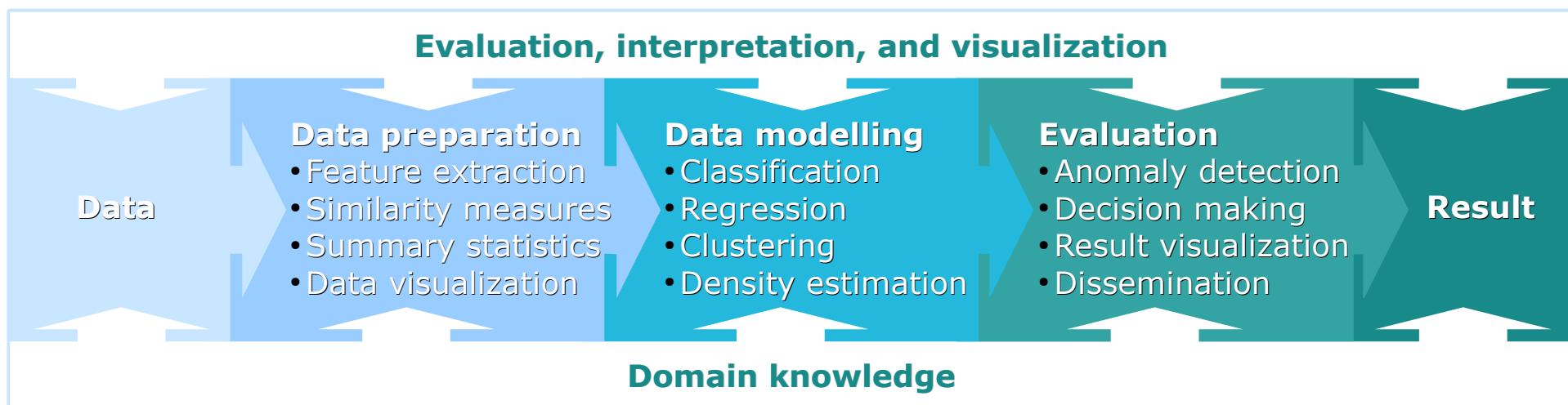
11. Density estimation and anomaly detection
(Tan 10.1-10.4)

Machine learning and data modelling in practice

12. Putting it all together: Summary and overview

13. Mini project

Data modeling framework



Todays learning objectives:

Be able to calculate various measures of similarity and dissimilarity.

Understand how various summary statistics are calculated and can be interpreted

Explain and apply Bayes theorem

Understand the normal and multi-variate normal distribution and the role of the covariance matrix

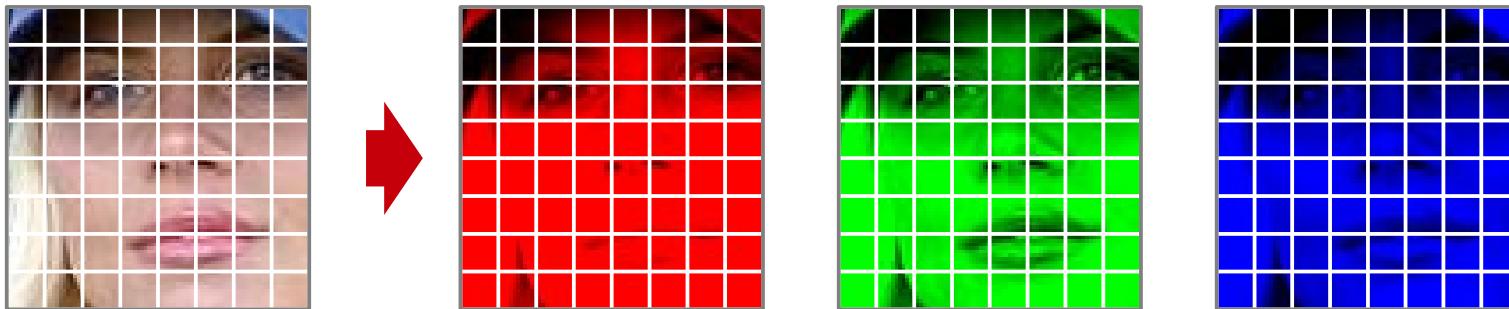
Example: Principal component analysis of images



- 1000 images, 86 x 86 pixels, 3 RGB intensities

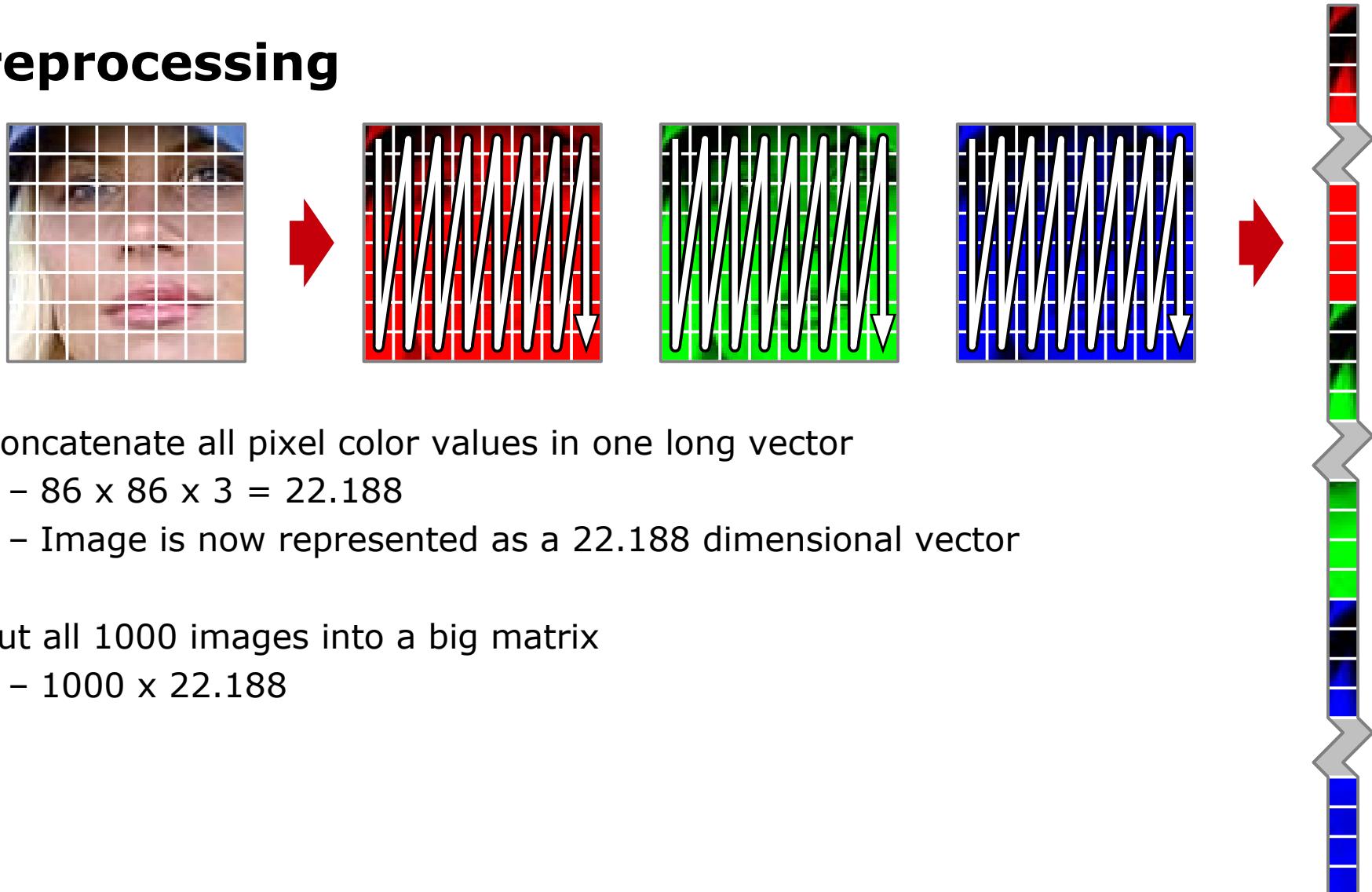
Tamara Berg "Faces in the wild"

Preprocessing



- Each image
 - 86×86 pixels
 - 3 RGB intensities
- Split image into red, green, and blue color channels

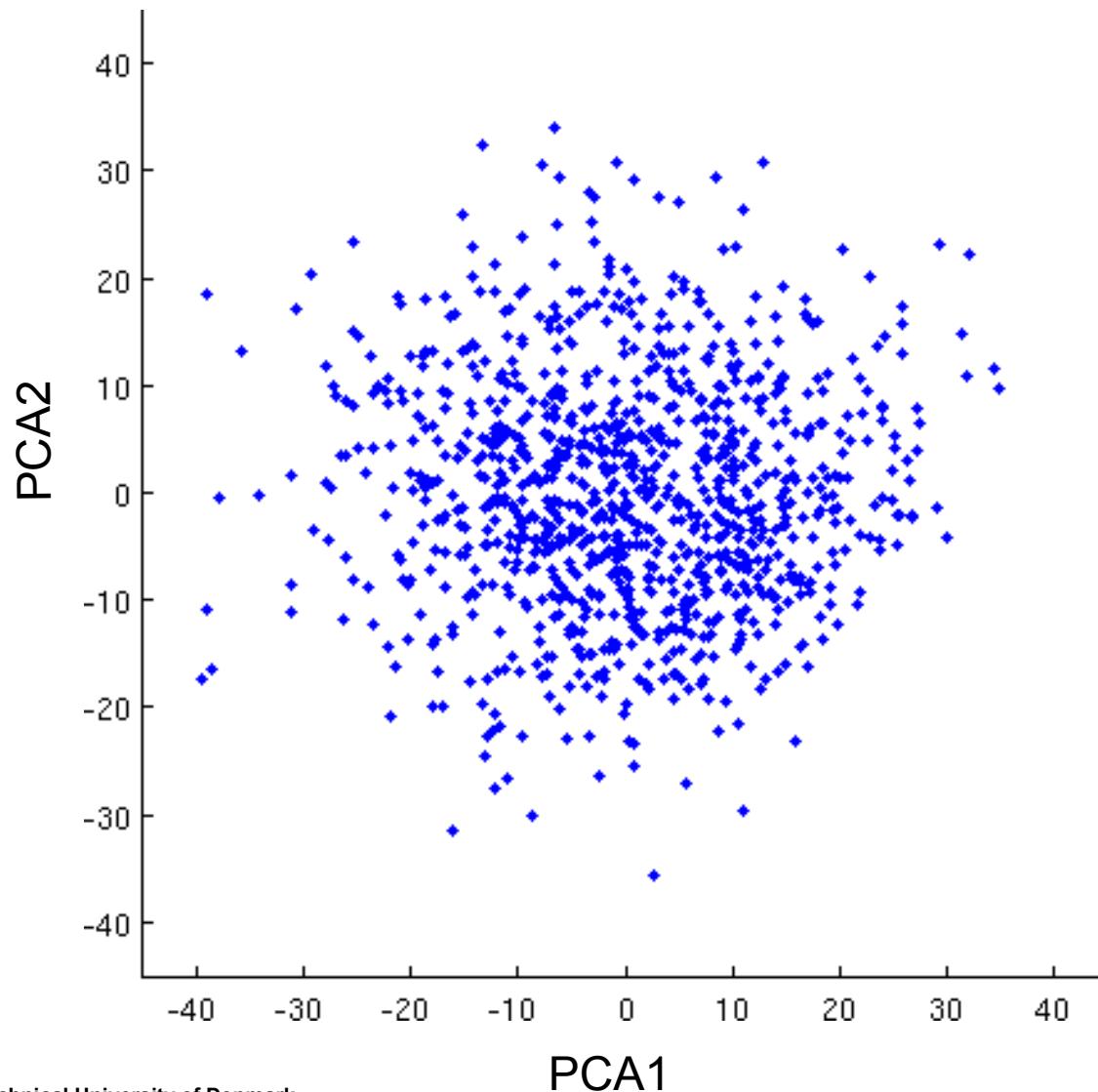
Preprocessing



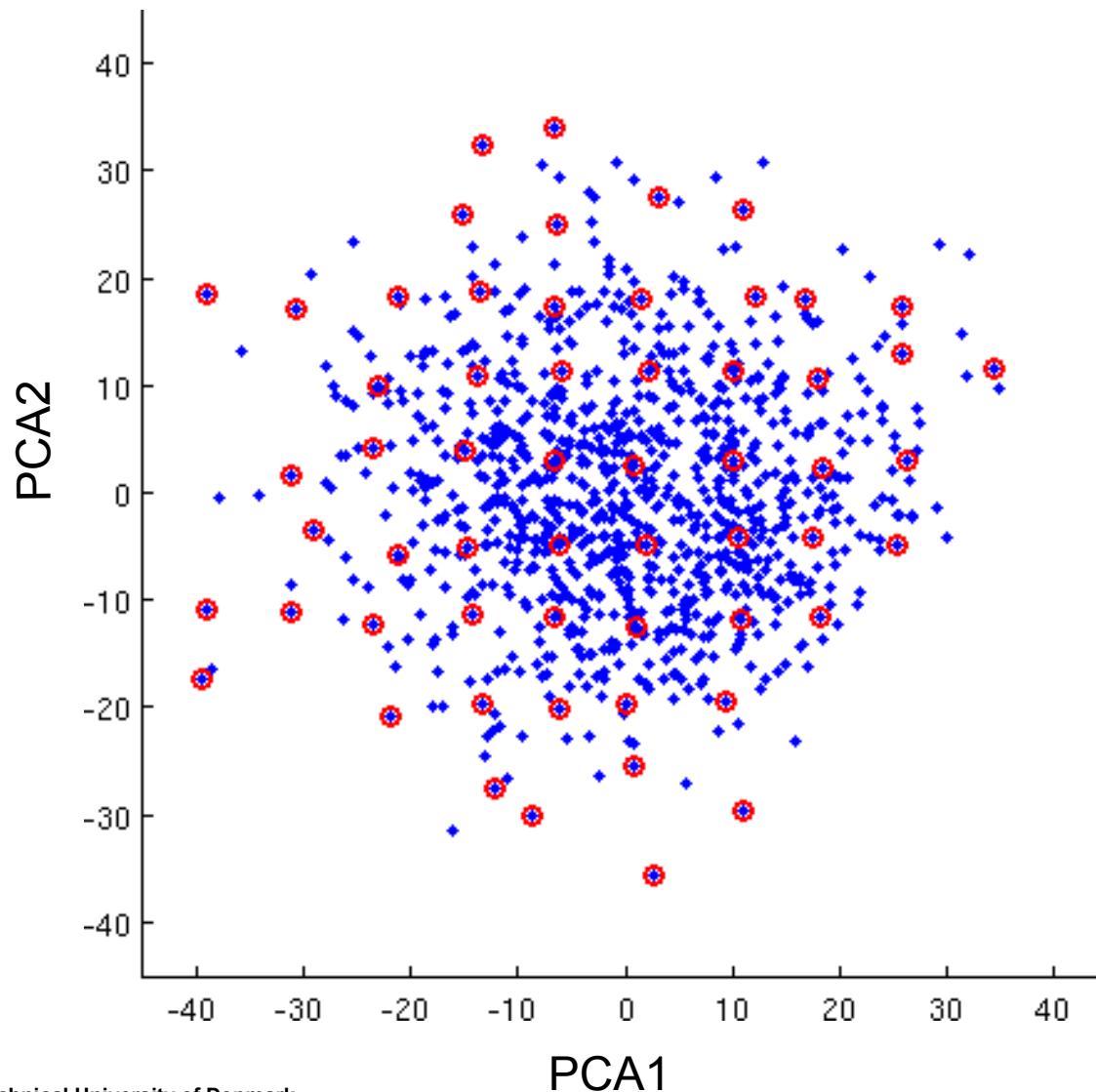
Principal component analysis (PCA)

- 1. Subtract the mean**
- 2. Compute the singular value decomposition (SVD)**
 - Orthogonal linear transformation
 - Transforms data to a new coordinate system
 - Greatest variance along the first axis
 - Second greatest variance along the second axis
 - Etc.
- **Plot data in the transformed coordinate system**
 - Corresponds to looking at data from an angle where it is most spread out

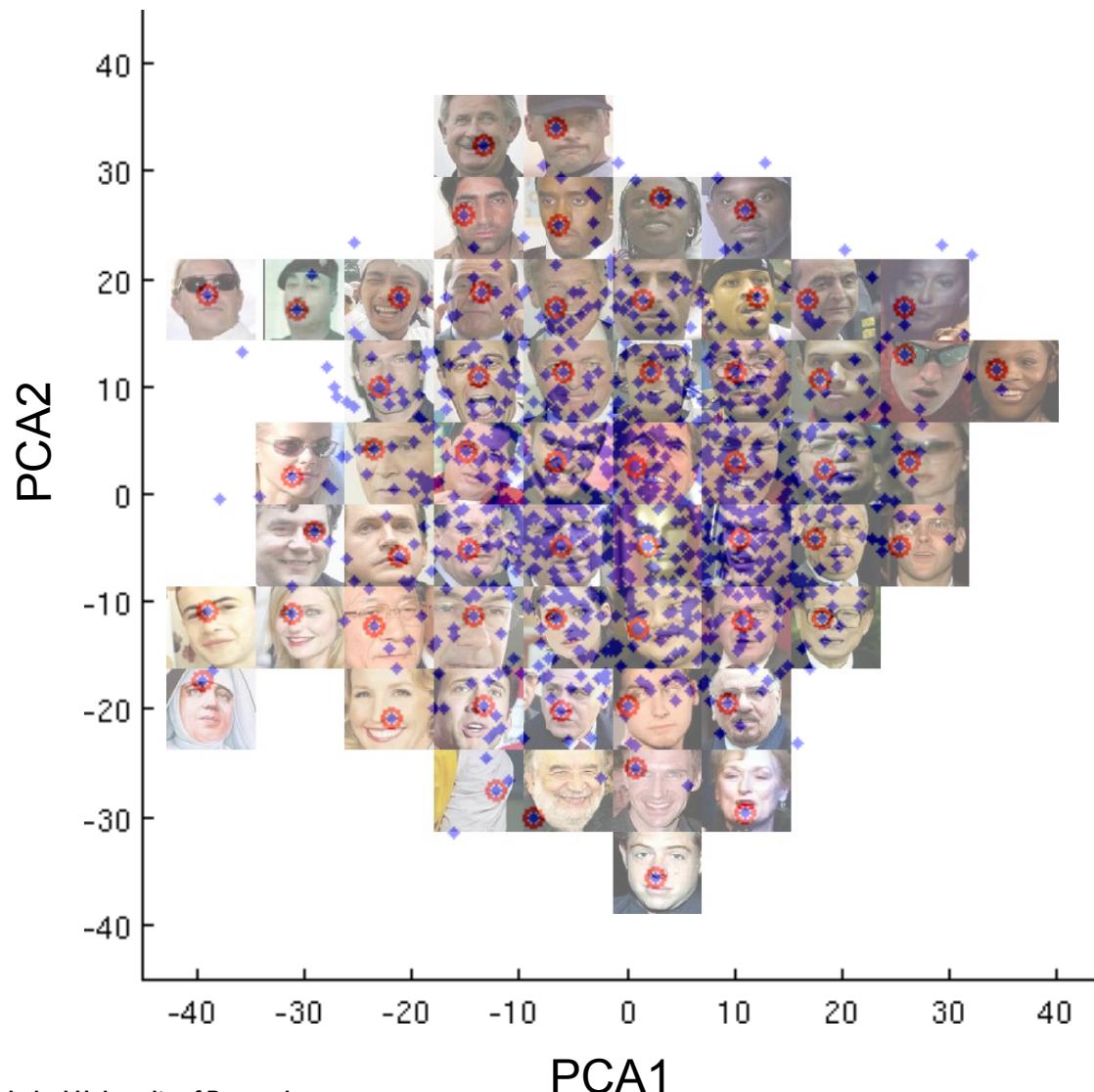
PCA of face images

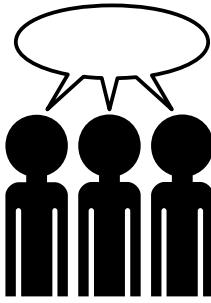


PCA of face images



PCA of face images

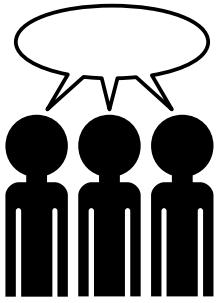




Discussion

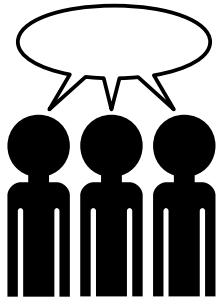
- What information do the two principal axes capture?





- What information do the two principal axes capture?





- What information do the two principal axes capture?



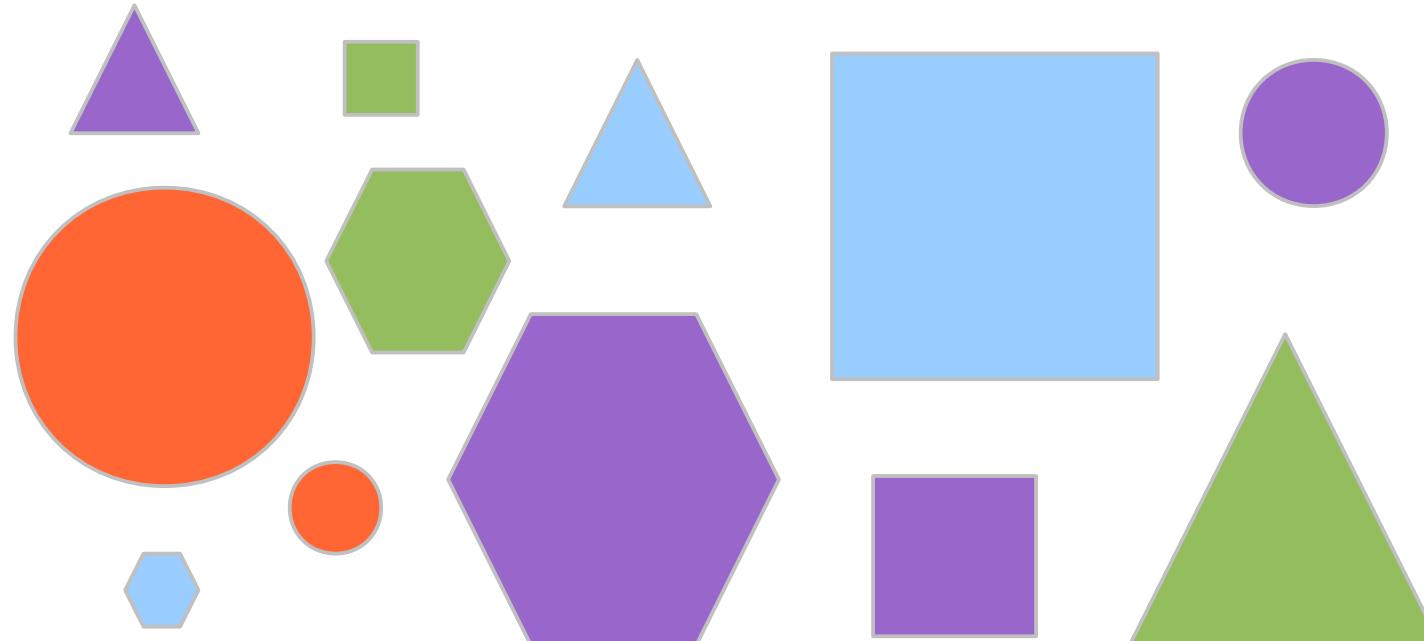
Similarity / Dissimilarity

- **Definition**

- A numerical measure of *how alike/different* two data objects are
- Often defined on the interval [0,1]

- **Similarity / dissimilarity between two data objects**

$$s(x, y), \quad d(x, y)$$



Dissimilarity measures

- Euclidean distance (2-norm)

$$d_2(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{n=1}^N (x_n - y_n)^2}$$

- Minkowski distance (p-norm)

$$d_p(\mathbf{x}, \mathbf{y}) = \left(\sum_{n=1}^N |x_n - y_n|^p \right)^{1/p}$$

Properties of dissimilarity measures

- **Positivity**

$$d(x, y) \geq 0$$

$$d(x, y) = 0 \Leftrightarrow x = y$$

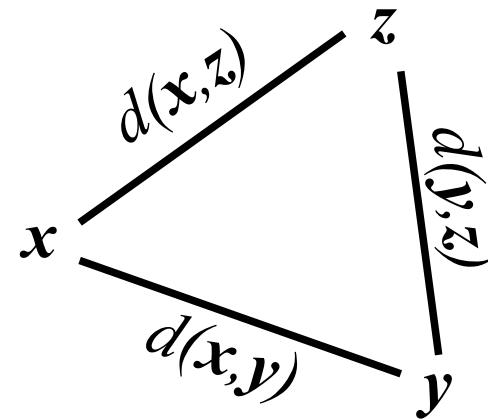
- **Symmetry**

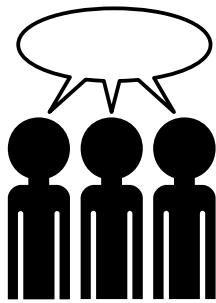
$$d(x, y) = d(y, x)$$

- **Triangle inequality**

$$d(x, z) \leq d(x, y) + d(y, z)$$

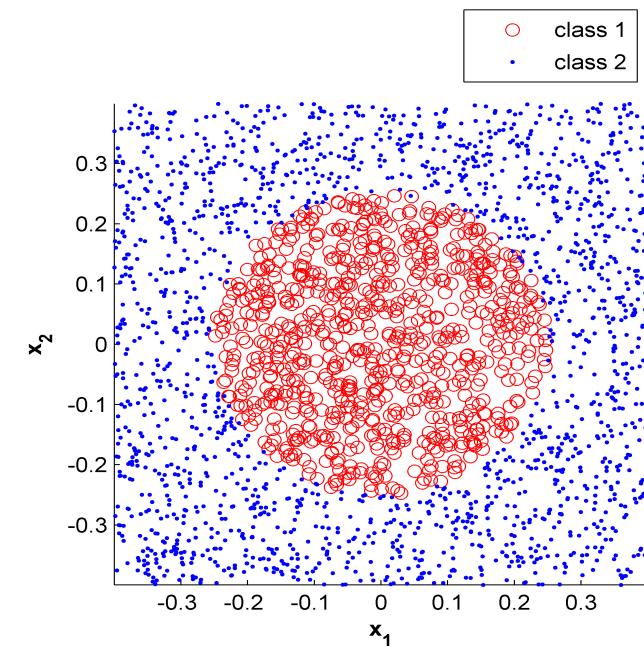
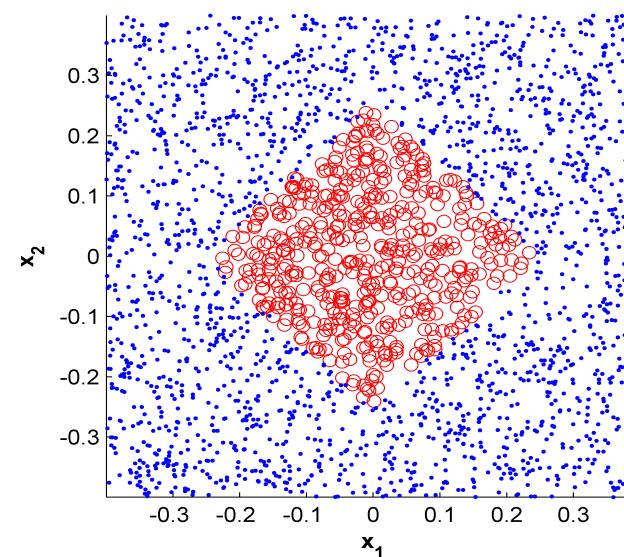
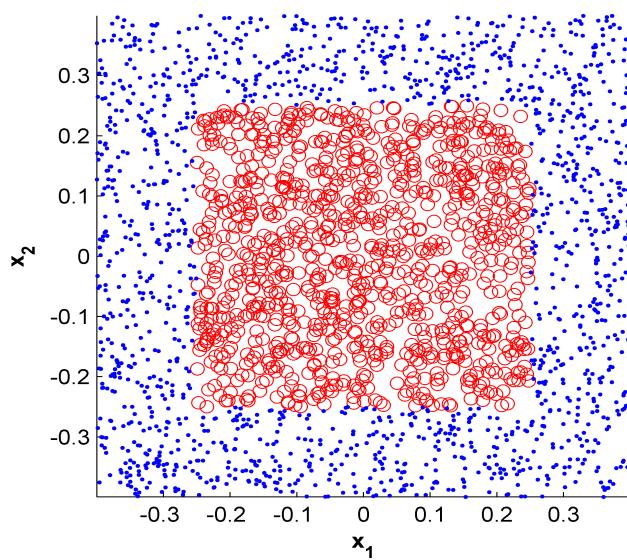
- Measures that satisfy all these properties: **Metrics**





Minkowski distance

Which Minkowski distance (p -norm) can be used to separate the two classes, by measuring the distance to the origo $(0,0)$?



$$d_p(\mathbf{x}, \mathbf{y}) = \left(\sum_{n=1}^N |x_n - y_n|^p \right)^{1/p}$$

Binary similarity measures

- **Simple matching coefficient (SMC)**

- Symmetric: Counts present and absent attributes equally

$$\text{SMC}(\mathbf{x}, \mathbf{y}) = \frac{f_{00} + f_{11}}{K}$$

- **Jaccard coefficient**

- Asymmetric: Counts only present attributes

$$J(\mathbf{x}, \mathbf{y}) = \frac{f_{11}}{K - f_{00}}$$

K : Total number of attributes

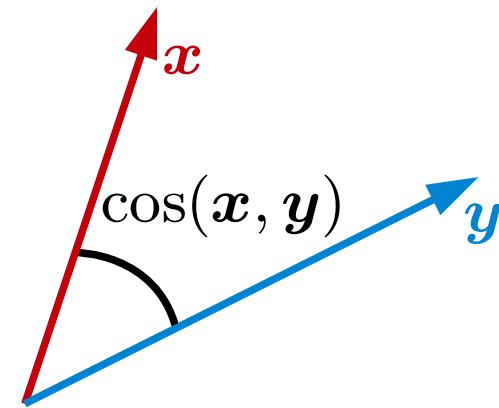
f_{00} : Number of attributes where $x_k = y_k = 0$

f_{11} : Number of attributes where $x_k = y_k = 1$

Continuous similarity measures

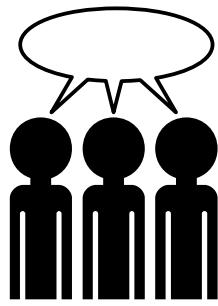
- Cosine similarity

$$\cos(x, y) = \frac{x^\top y}{\|x\|_2 \|y\|_2}$$



- Extended Jaccard coefficient

$$EJ(x, y) = \frac{x^\top y}{\|x\|_2^2 + \|y\|_2^2 - x^\top y}$$



Calculate the SMC, Jaccard, Cosine and Extended Jaccard similarity between customer 1 and customer 2 in the market basket data below.

ID	Bread	Soda	Milk	Beer	Diaper
1	1	1	1	0	0
2	0	1	1	0	1

$$\text{SMC}(\mathbf{x}, \mathbf{y}) = \frac{f_{00} + f_{11}}{K}$$

$$\text{J}(\mathbf{x}, \mathbf{y}) = \frac{f_{11}}{K - f_{00}}$$

$$\cos(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x}^\top \mathbf{y}}{\|\mathbf{x}\|_2 \|\mathbf{y}\|_2}$$

$$EJ(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x}^\top \mathbf{y}}{\|\mathbf{x}\|_2^2 + \|\mathbf{y}\|_2^2 - \mathbf{x}^\top \mathbf{y}}$$

K : Total number of attributes

f_{00} : Number of attributes where $x_k = y_k = 0$

f_{11} : Number of attributes where $x_k = y_k = 1$

Empirical statistics

- Empirical mean

$$\bar{x} = \frac{1}{N} \sum_{n=1}^N x_n$$

- Empirical covariance

$$\text{cov}(\mathbf{x}, \mathbf{y}) = \frac{1}{N-1} \sum_{n=1}^N (x_n - \bar{x})(y_n - \bar{y})$$

- Empirical variance

$$\text{var}(\mathbf{x}) = \text{cov}(\mathbf{x}, \mathbf{x}) = \frac{1}{N-1} \sum_{n=1}^N (x_n - \bar{x})^2$$

- Empirical standard deviation

$$\text{std}(\mathbf{x}) = \sqrt{\text{var}(\mathbf{x})}$$

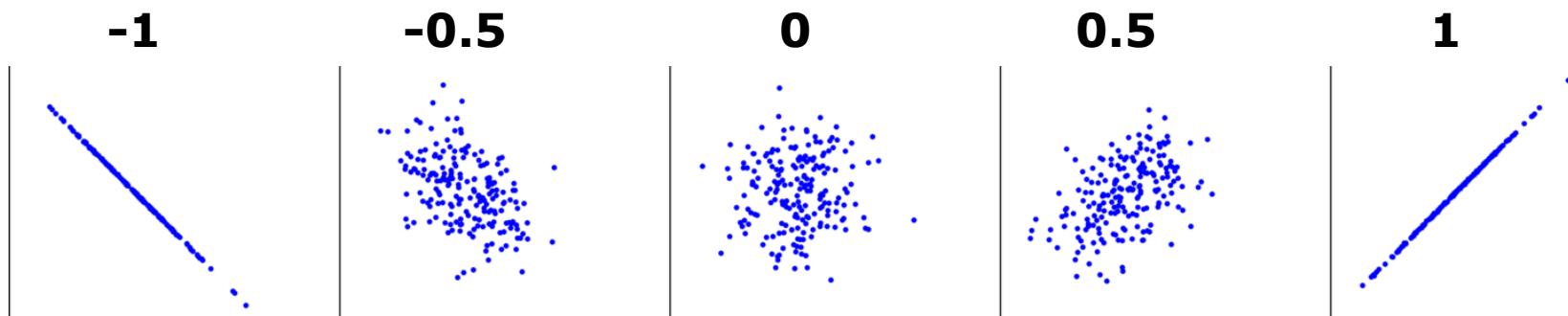
Correlation

- **Measure of linear relation**

$$\text{corr}(\mathbf{x}, \mathbf{y}) = \frac{\text{cov}(\mathbf{x}, \mathbf{y})}{\text{std}(\mathbf{x})\text{std}(\mathbf{y})}$$

- A correlation of **1** or **-1** means there is a perfect linear relation

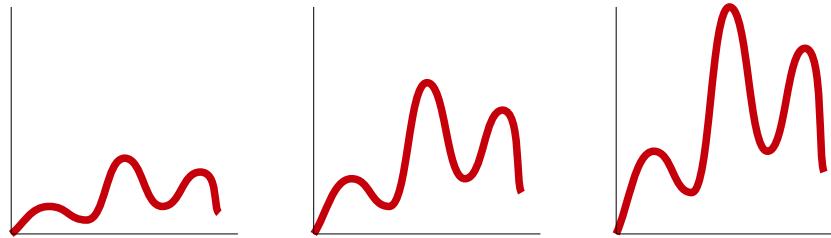
$$x_k = a y_k + b$$



Invariance

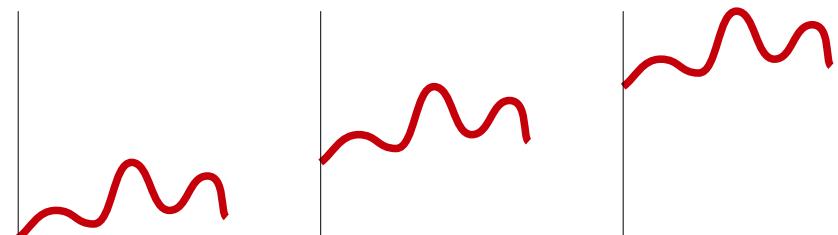
- **Scale**

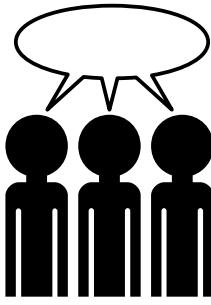
$$d(\mathbf{x}, \mathbf{y}) = d(\alpha\mathbf{x}, \mathbf{y})$$



- **Translation**

$$d(\mathbf{x}, \mathbf{y}) = d(\beta + \mathbf{x}, \mathbf{y})$$



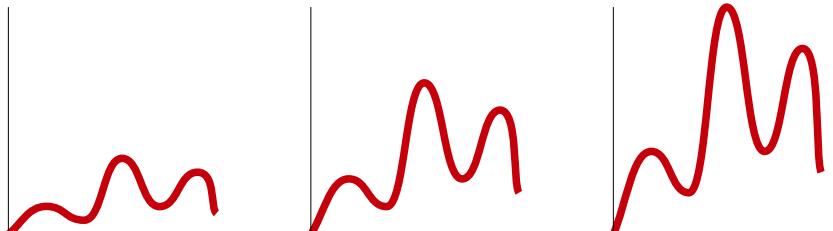


Discussion

- When would a **scale invariant** similarity measure be useful
 - Give an example
- When would a **translation invariant** similarity measure be useful
 - Give an example

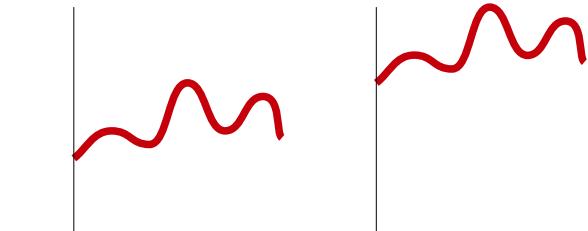
Scale invariance

$$d(x, y) = d(\alpha x, y)$$

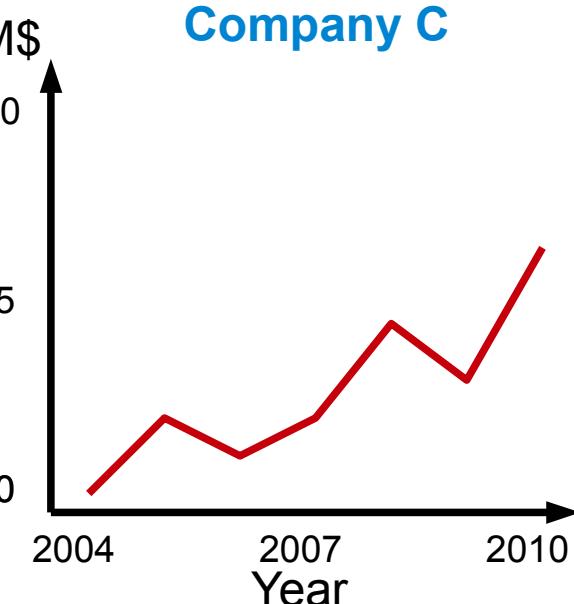
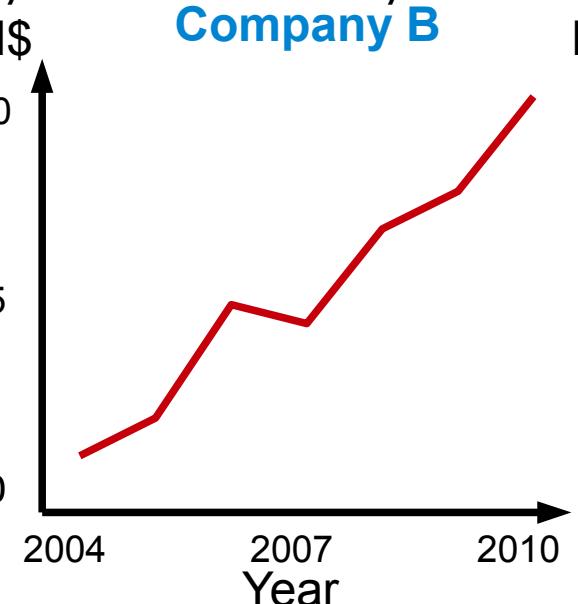
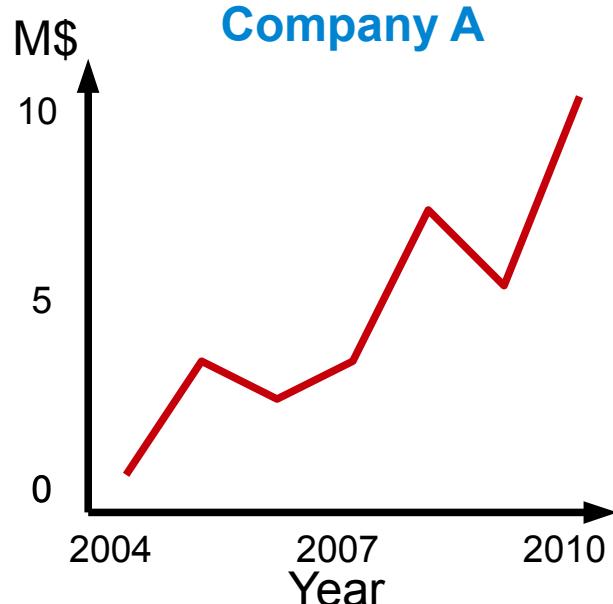


Translation invariance

$$d(x, y) = d(\beta + x, y)$$



- Which of these three companies are similar and in what way?
 - Which similarity measure would you use?



Issues in proximity calculation

- **What to do when attributes have different**
 - **Scale?**
 - Standardize attributes
 - Use scale invariant similarity measure
 - **Type?**
 - Compute similarities for each attribute and combine
 - **Importance?**
 - Compute a weighted similarity measure

Standardization

- **Attributes have different scales.**
 - Example:
 - **Number of children** ~ 0-5
 - **Age** ~ 0-100 years
 - **Annual income** ~ 0-50.000 €
- Unless we do something, **Annual income** will dominate
 - **Standardization**: Subtract mean and divide by standard deviation

$$x_k^* = \frac{x_k - \bar{x}_k}{\text{std}(x_k)}$$

Combining heterogeneous attributes

- **Attributes have different type**
 - Example:
 - **Age:** Continuous
 - **Education:** Binary
 - Primary (yes/no)
 - Secondary (yes/no)
 - Tertiary (yes/no)
- Similarity measure must handle **continuous** and **binary** features
 - **Compute similarities for each attribute and combine**

$$s_{\text{Age}} = d_1(x_{\text{Age}}, y_{\text{Age}}) \quad s_{\text{Edu.}} = \text{SMC}(x_{\text{Edu.}}, y_{\text{Edu.}})$$

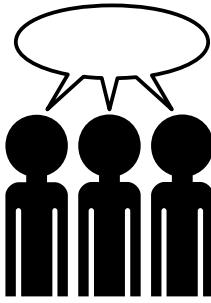
$$s(\mathbf{x}, \mathbf{y}) = \frac{1}{2}(s_{\text{Age}} + s_{\text{Edu.}})$$

Weighting attributes by importance

- **Attributes have different importance**
 - Example:
 - **Age:** Very important
 - **Education:** Less important
- Similarity measure must take **importance** into account
 - Introduce **importance weights** for each attribute

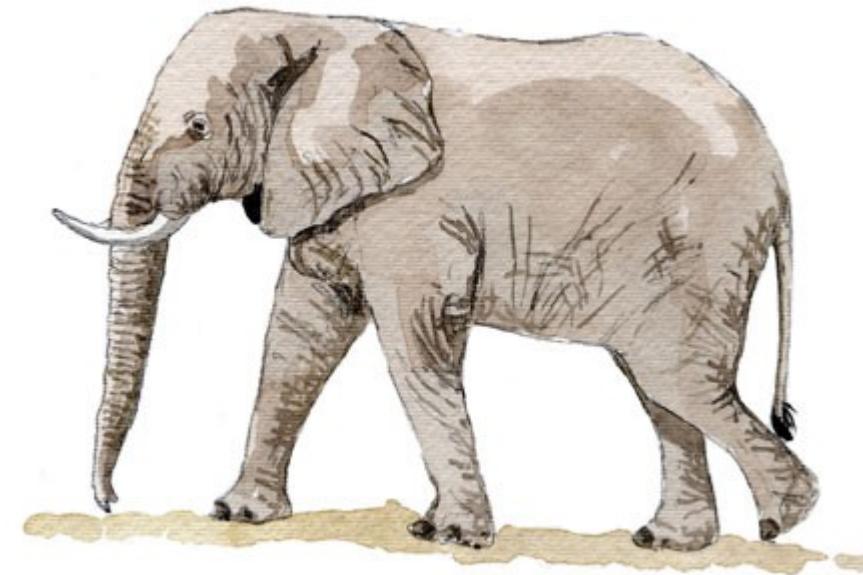
$$s_{\text{Age}} = d_1(x_{\text{Age}}, y_{\text{Age}}) \quad s_{\text{Edu.}} = \text{SMC}(x_{\text{Edu.}}, y_{\text{Edu.}})$$

$$s(\mathbf{x}, \mathbf{y}) = 0.99 \cdot s_{\text{Age}} + 0.01 \cdot s_{\text{Edu.}}$$



Discussion

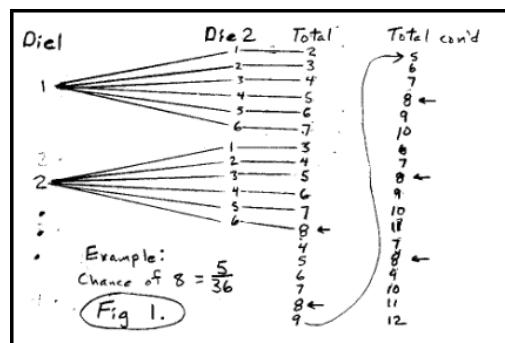
- The following **attributes** are measured for a herd of elephants
 - Weight
 - Height
 - Tusk length
 - Trunk length
 - Ear area
 - Gender
- **Based on these measurements**
 - How would you evaluate how similar elephants are?
 - Justify your answer



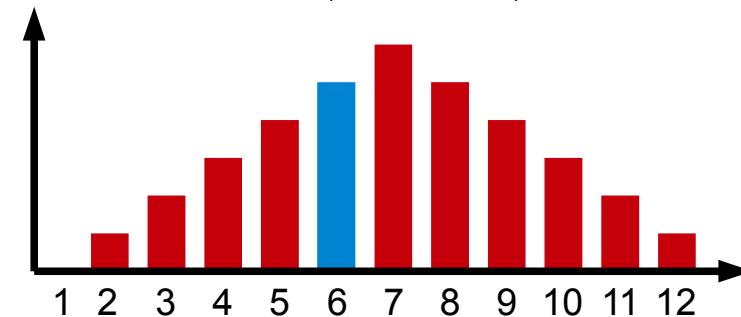
Probabilities

- **Discrete: Probability mass**

- Example: The sum of two dice



$$P(X = v)$$

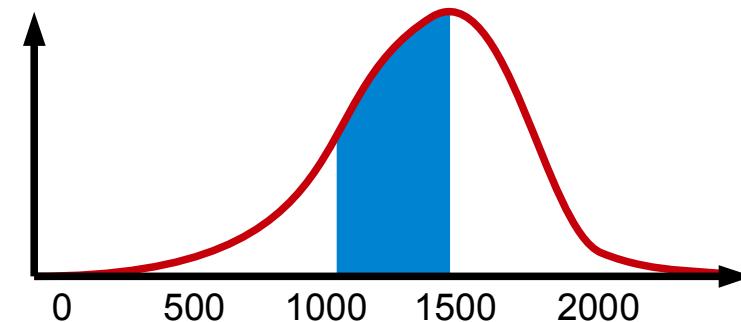


- **Continuous: Probability density**

- Example: Lifetime of a light bulb

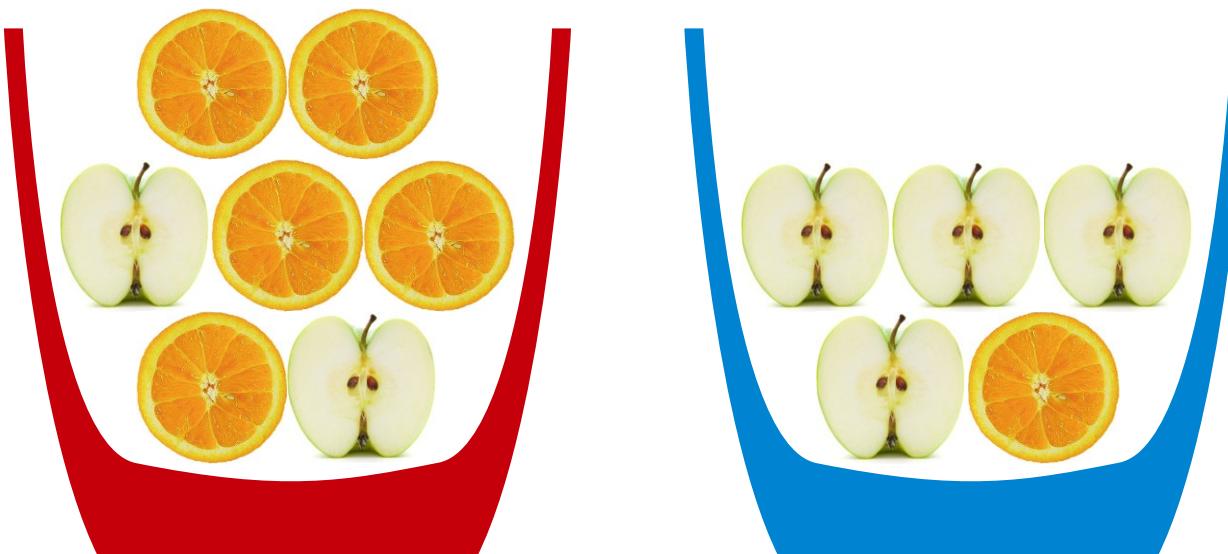


$$P(a \leq X \leq b) = \int_a^b p(X)dX$$



Probabilities

- What is the probability of an **orange** if the bowl is **red**?
- What is the probability of the **red** bowl if the fruit is **orange**?



Probabilities

- Basic rules of probability

- Sum rule

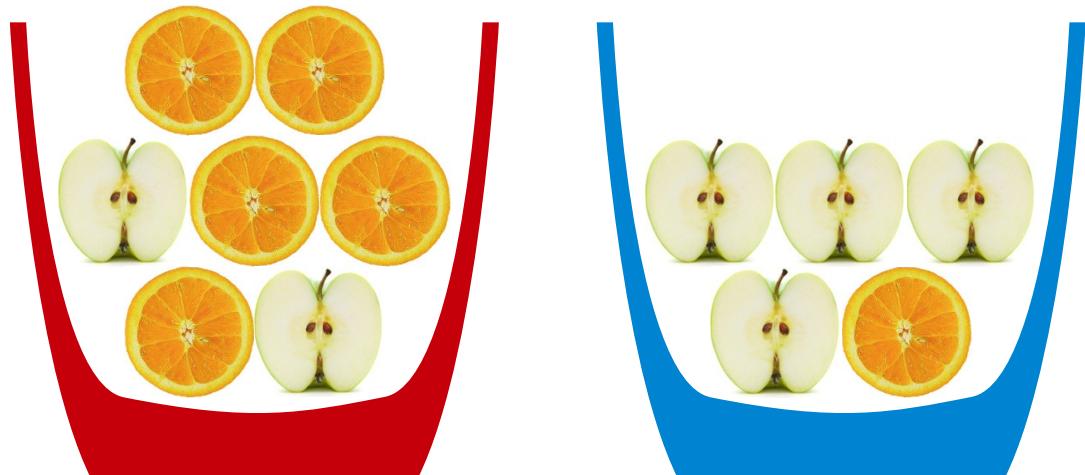
$$p(x) = \sum_y p(x, y)$$

- Product rule

$$p(x, y) = p(x|y)p(y)$$

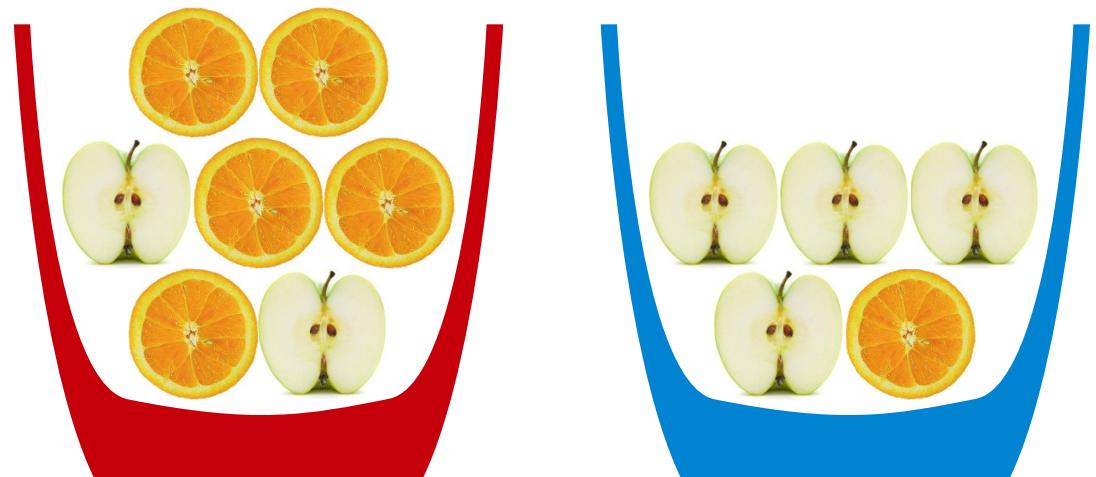
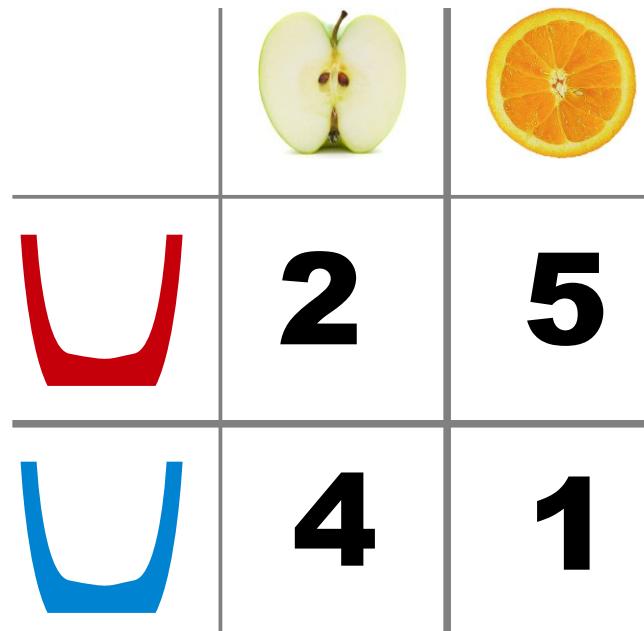
- Bayes' rule

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)}$$



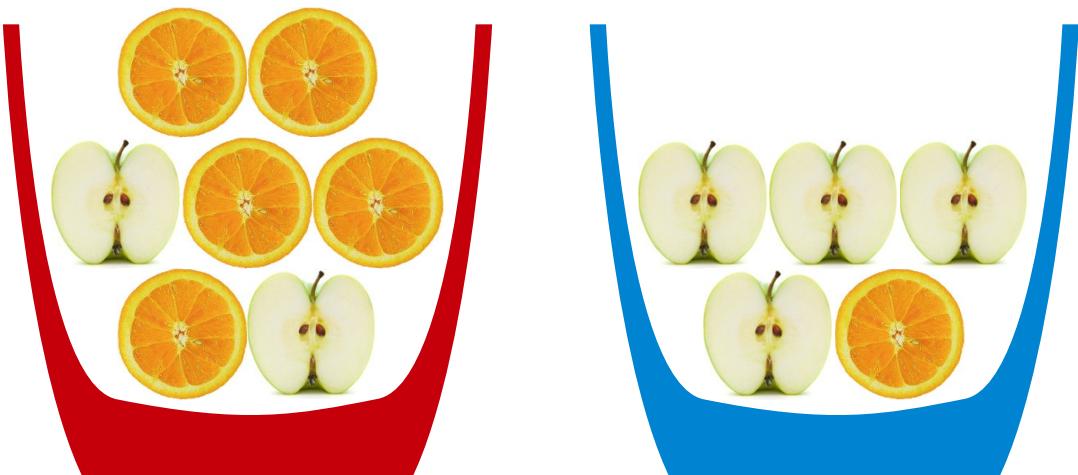
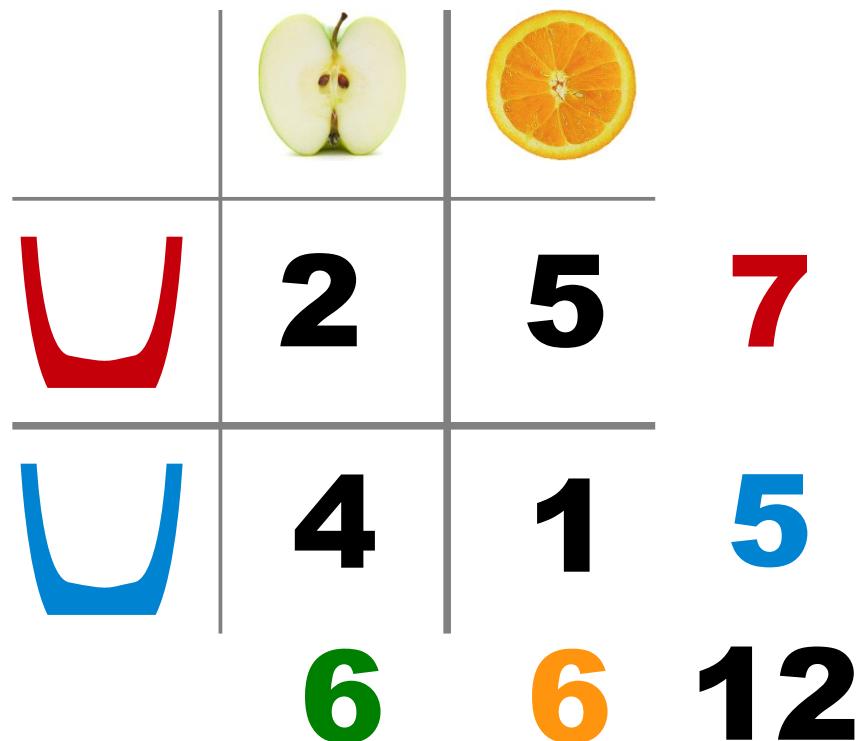
Probabilities

- What is the probability of an **orange** if the bowl is **red**?
- What is the probability of the **red** bowl if the fruit is **orange**?



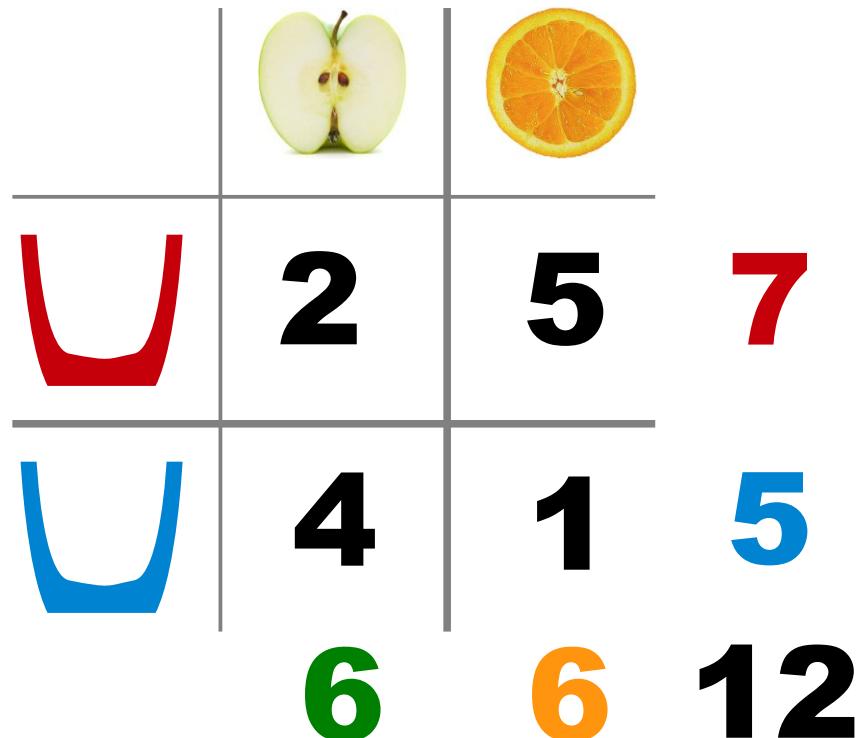
Probabilities

- What is the probability of an **orange** if the bowl is **red**?
- What is the probability of the **red** bowl if the fruit is **orange**?



Probabilities

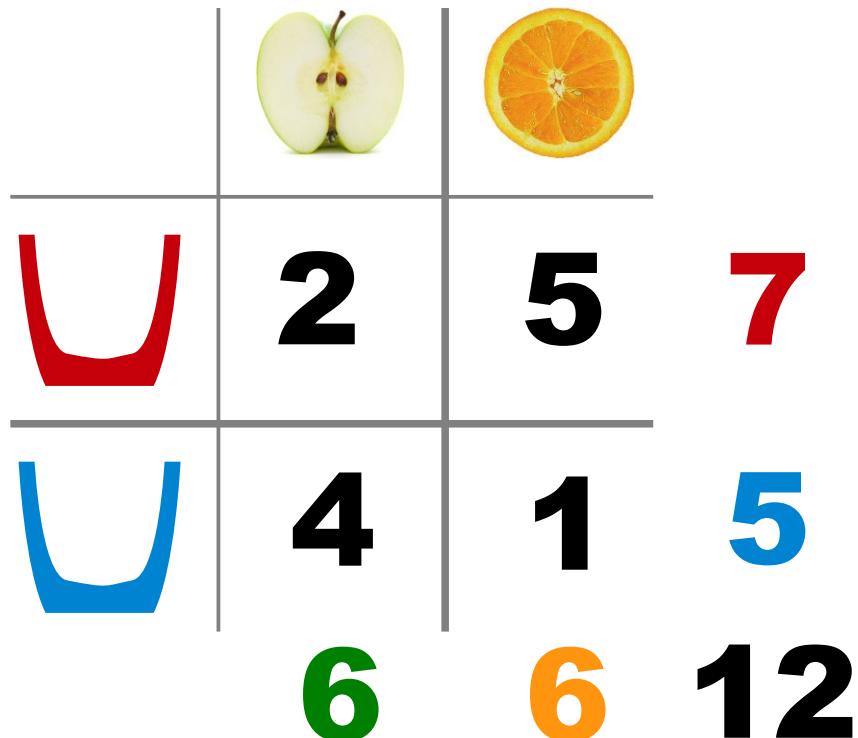
- What is the probability of an **orange** if the bowl is **red**?
- What is the probability of the **red** bowl if the fruit is **orange**?



$$p(o|r) = \frac{p(r, o)}{p(r)} = \frac{\mathbf{5/12}}{\mathbf{7/12}} = \mathbf{5/7}$$

Probabilities

- What is the probability of an **orange** if the bowl is **red**?
- What is the probability of the **red** bowl if the fruit is **orange**?

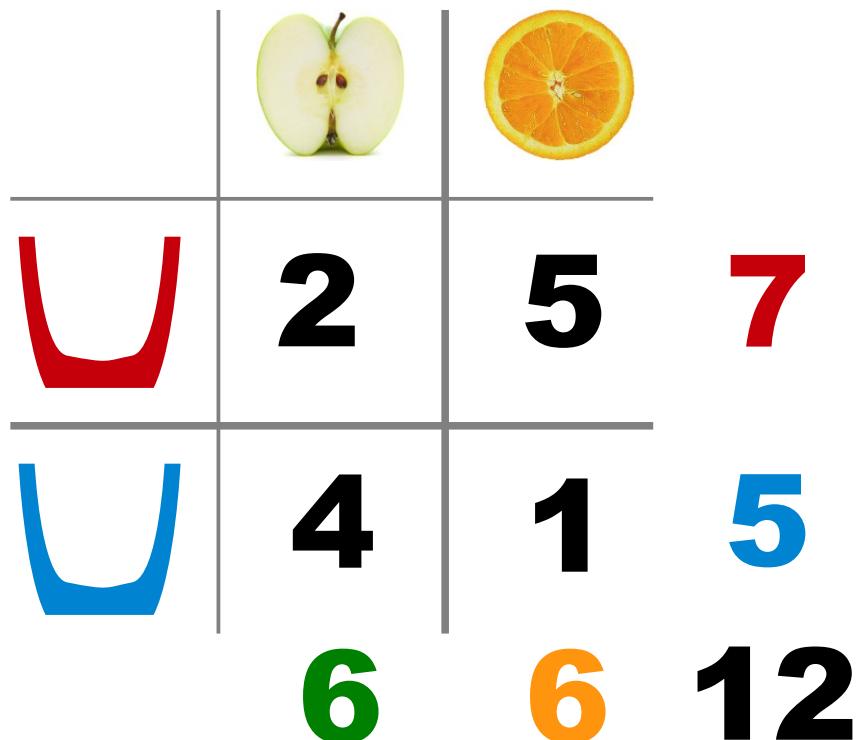


$$p(o|r) = \frac{p(r,o)}{p(r)} = \frac{\mathbf{5/12}}{\mathbf{7/12}} = \mathbf{5/7}$$

$$p(r|o) = \frac{p(r,o)}{p(o)} = \frac{\mathbf{5/12}}{\mathbf{6/12}} = \mathbf{5/6}$$

Probabilities

- What is the probability of an **orange** if the bowl is **red**?
- What is the probability of the **red** bowl if the fruit is **orange**?

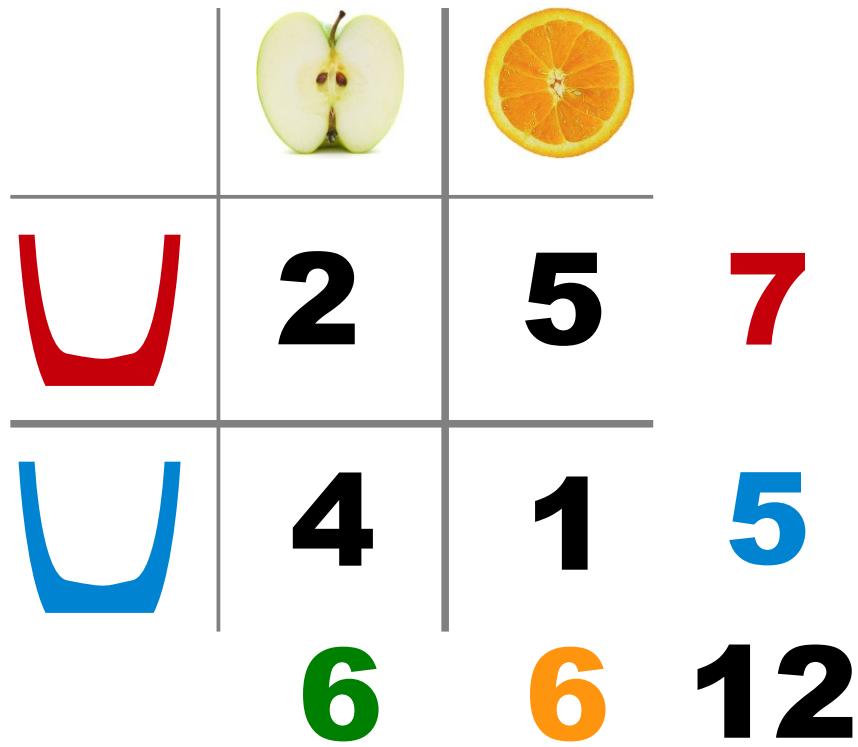


$$p(o|r) = \frac{p(r,o)}{p(r)} = \frac{\mathbf{5/12}}{\mathbf{7/12}} = \mathbf{5/7}$$

$$\begin{aligned} p(r|o) &= \frac{p(r,o)}{p(o)} = \frac{\mathbf{5/12}}{\mathbf{6/12}} = \mathbf{5/6} \\ &= \frac{p(o|r)p(r)}{p(o)} \end{aligned}$$

Probabilities

- What is the probability of an **orange** if the bowl is **red**?
- What is the probability of the **red** bowl if the fruit is **orange**?

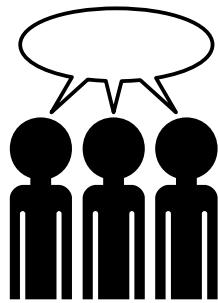


$$p(o|r) = \frac{p(r,o)}{p(r)} = \frac{\mathbf{5/12}}{\mathbf{7/12}} = \mathbf{5/7}$$

$$p(r|o) = \frac{p(r,o)}{p(o)} = \frac{\mathbf{5/12}}{\mathbf{6/12}} = \mathbf{5/6}$$

$$= \frac{p(o|r)p(r)}{p(o)}$$

$$= \frac{\mathbf{5/7} \cdot \mathbf{7/12}}{\mathbf{6/12}} = \mathbf{5/6}$$



Medical test

A medical test for a given disease

- Correctly identifies the disease 99% of the time (true positives), and
- Incorrectly turns out positive 2% of the time (false positives).

You know that

- 1% of the population suffers from the disease.

You go to the doctor to get tested, and the test turns out to be positive.

What is the probability you have the disease?

Hints:

- Identify from the text:
 $p(\text{Positive}|\text{Disease})$
 $p(\text{Positive}|\text{No Disease})$
 $p(\text{Disease})$
 $p(\text{No Disease})$
- Use the basic rules of probability given to the right to find:
 $p(\text{Disease}|\text{Positive})$

$$p(x) = \sum_y p(x, y)$$

$$p(x, y) = p(x|y)p(y)$$

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)}$$

Frequency and mode

- **Frequency:** Percentage of time a value occurs
 - Example: Given the attribute **Gender** and a representative population of people, the value **Female** occurs about 50% of the time
- **Mode:** The most frequent attribute value
 - Example: Given the attribute **Operating System** and a representative population of computers, the value **Microsoft Windows** is the mode
- The notions of frequency and mode are typically used with categorical data

Percentiles

- **Percentiles:** Given an ordinal or continuous attribute \mathbf{x} and a number \mathbf{p} between 0 and 100, the \mathbf{p} th percentile is a value \mathbf{x}_p of \mathbf{x} such that \mathbf{p} percent of the observed values of \mathbf{x} are less than \mathbf{x}_p .
 - Example: The 10th percentile of \mathbf{x} is the value $\mathbf{x}_{10\%}$ such that 10% of all values are less than $\mathbf{x}_{10\%}$.
- **Median:** The 50th percentile
 - Sort the numbers and take the middle value
(if there are an even number of values, average the two middle values)

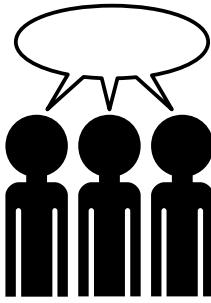
Measures of location

- **Mean:** Average

$$\bar{x} = \frac{1}{N} \sum_{n=1}^N x_n$$

- **Median:** Sort the numbers

$$\text{median}(x) = \begin{cases} x_{R+1} & \text{if } N \text{ is odd } (N = 2R + 1) \\ \frac{1}{2}(x_R + x_{R+1}) & \text{if } N \text{ is even } (N = 2R) \end{cases}$$



Discussion

- What are the frequencies and the mode of these numbers and what is the mean and median?

0, 1, 1, 3, 5, 590

- Explain also what to be careful about when using the mean and median

- **Frequency:** Percentage of time a value occurs
- **Mode:** The most frequent attribute value
- **Mean:** Average

$$\bar{x} = \frac{1}{N} \sum_{n=1}^N x_n$$

- **Median:** Sort the numbers

$$\text{median}(x) = \begin{cases} x_{R+1} & \text{if } N \text{ is odd } (N = 2R + 1) \\ \frac{1}{2}(x_R + x_{R+1}) & \text{if } N \text{ is even } (N = 2R) \end{cases}$$

Measures of spread

- **Range**

$$\text{range}(x) = \max(x) - \min(x)$$

- **Variance**

$$\text{variance}(x) = s_x^2 = \frac{1}{N-1} \sum_{n=1}^N (x_n - \bar{x})^2$$

- **Absolute average deviation (AAD)**

$$\text{AAD}(x) = \frac{1}{N} \sum_{n=1}^N |x_n - \bar{x}|$$

- **Median absolute deviation (MAD)**

$$\text{MAD}(x) = \text{median} \{|x_1 - \bar{x}|, \dots, |x_N - \bar{x}|\}$$

- **Interquartile range (IQR)**

$$\text{IQR}(x) = x_{75\%} - x_{25\%}$$

Expected values

- Discrete random variable

$$\mathbb{E} [g(X)] = \sum_i g(x_i)P(X = x_i)$$

- Continuous random variable

$$\mathbb{E} [g(X)] = \int_{-\infty}^{\infty} g(X)p(X)dX$$

Statistics

- **Mean**

$$\bar{x} = \mathbb{E}[x]$$

- **Covariance**

$$\text{cov}(x, y) = \mathbb{E}[(x - \bar{x})(y - \bar{y})]$$

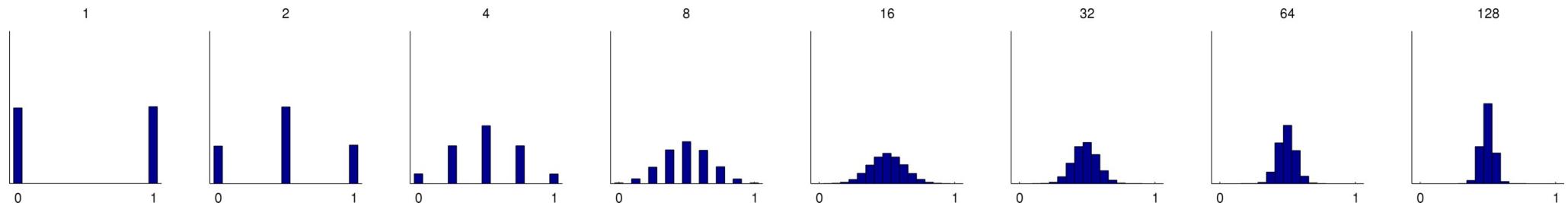
- **Variance**

$$\text{var}(x) = \text{cov}(x, x) = \mathbb{E}[(x - \bar{x})^2]$$

- **Standard deviation**

$$\text{std}(x) = \sqrt{\text{var}(x)}$$

Normal distribution

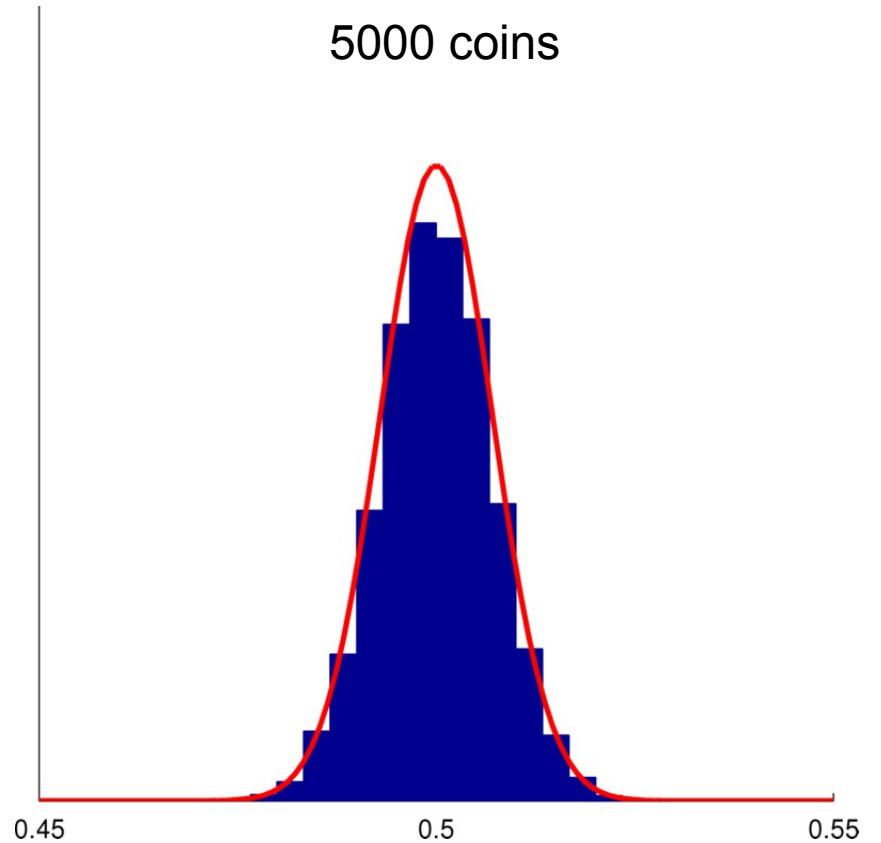


- **Central limit theorem**

- The mean of a large number of random variables will tend to a Normal distribution irrespective of the distribution of the random variables
(Under certain conditions)

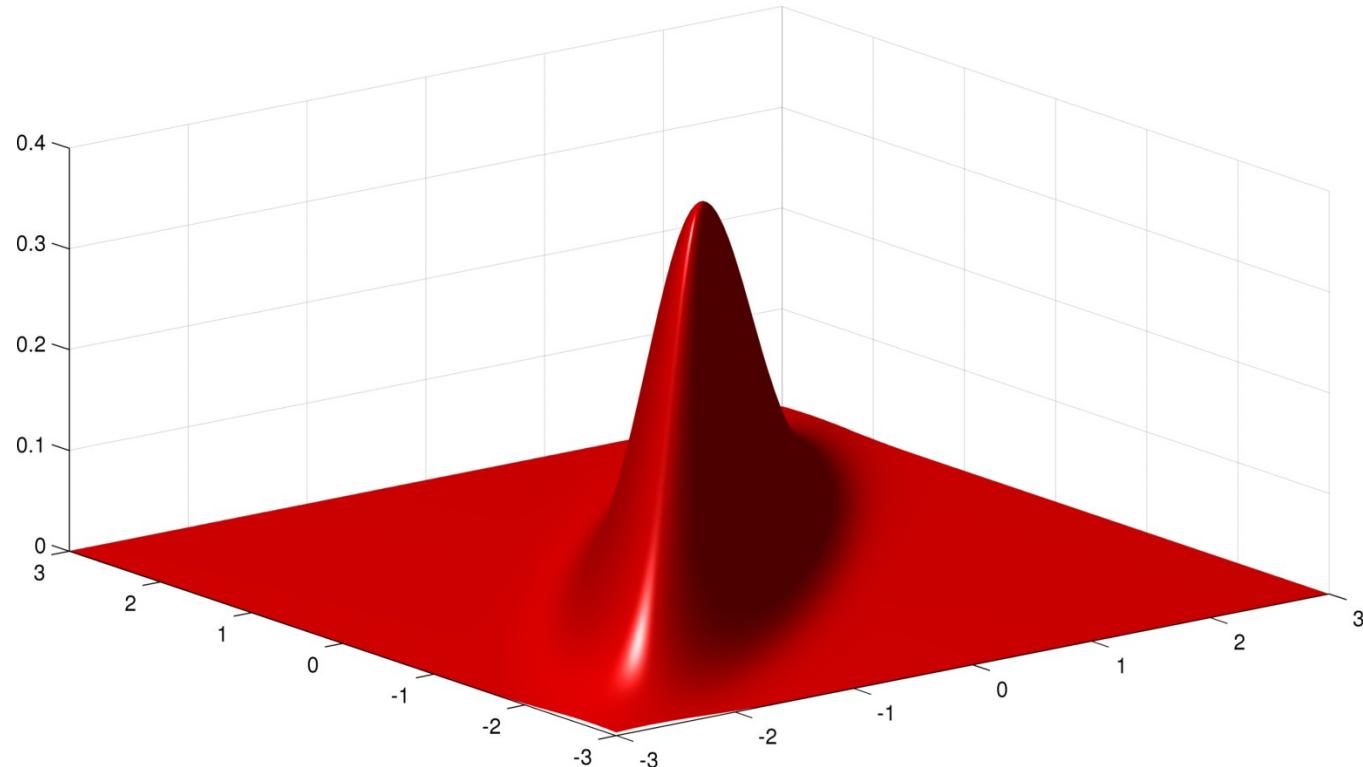
$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

- **Example:** Proportion of heads when flipping
 - 1 coin, 2 coins, 4 coins etc.



Multivariate Normal distribution

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{k/2} |\Sigma|^{1/2}} \exp \left(-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right)$$



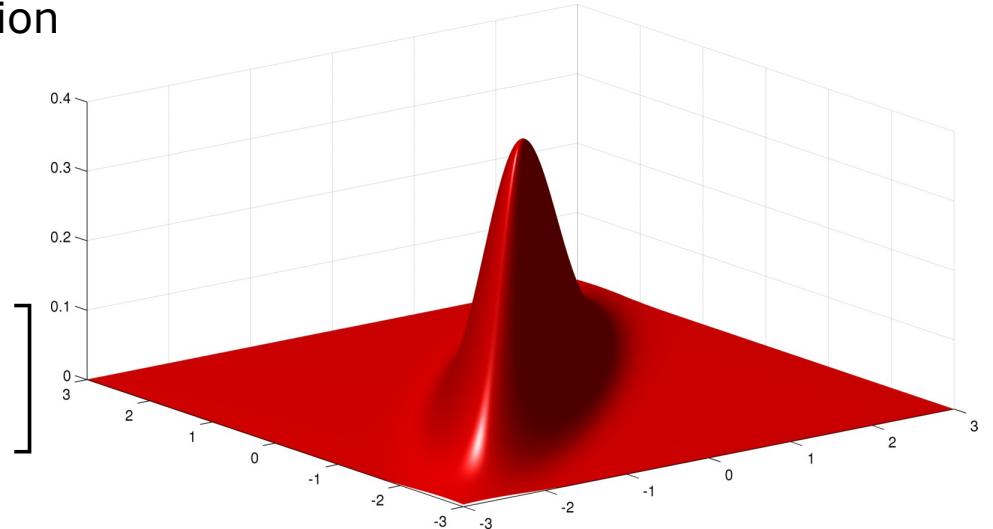
Multivariate Normal distribution

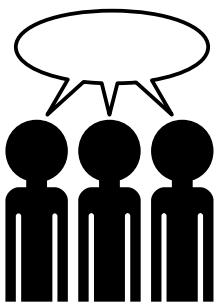
$$p(\mathbf{x}) = \frac{1}{(2\pi)^{k/2} |\Sigma|^{1/2}} \exp \left(-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right)$$

- Example: 2-dimensional Normal distribution

$$\boldsymbol{\mu} = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \end{bmatrix}$$

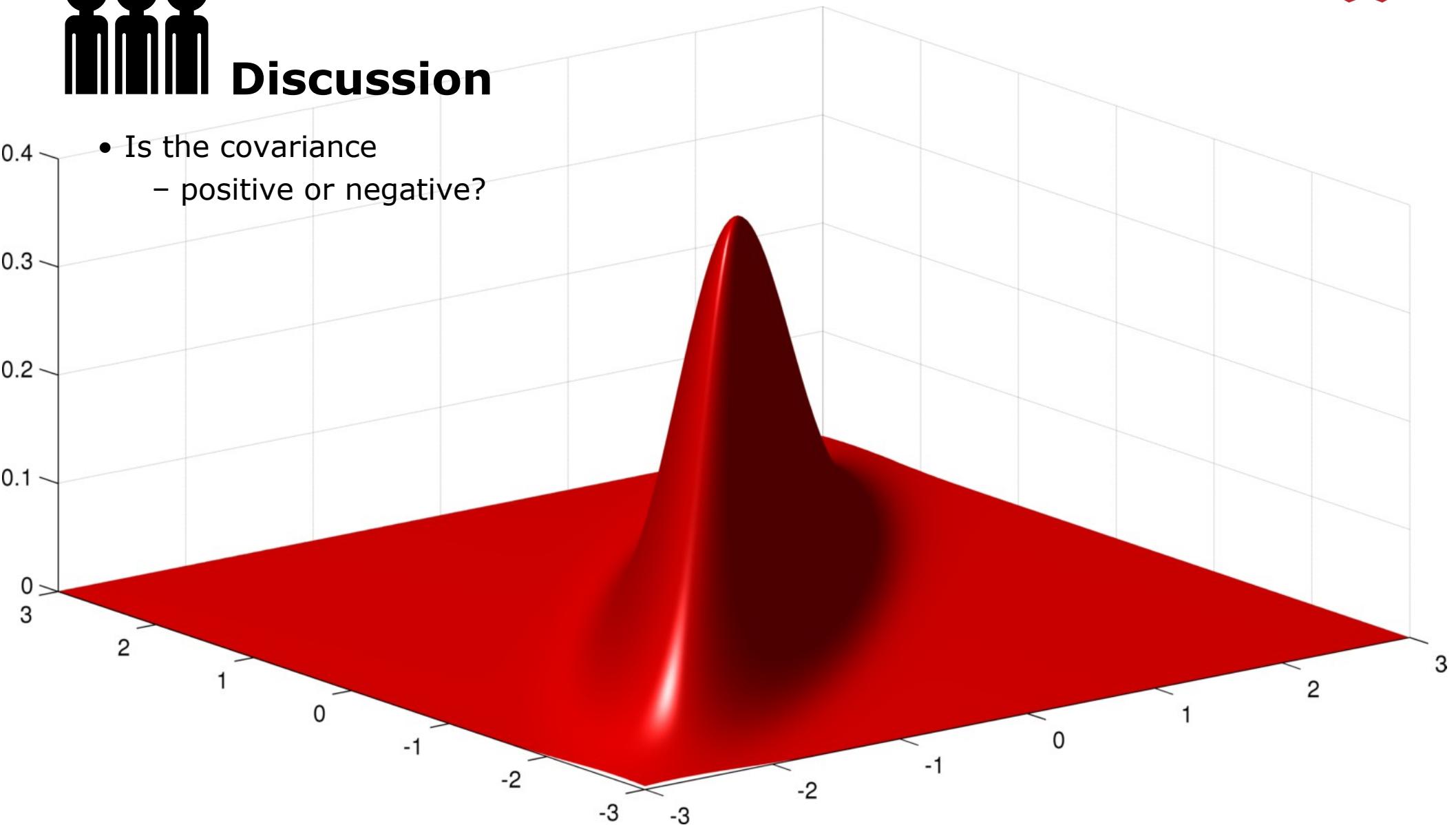
$$\boldsymbol{\Sigma} = \begin{bmatrix} \text{var}(x_1) & \text{cov}(x_1, x_2) \\ \text{cov}(x_2, x_1) & \text{var}(x_2) \end{bmatrix}$$

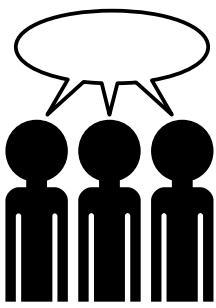




Discussion

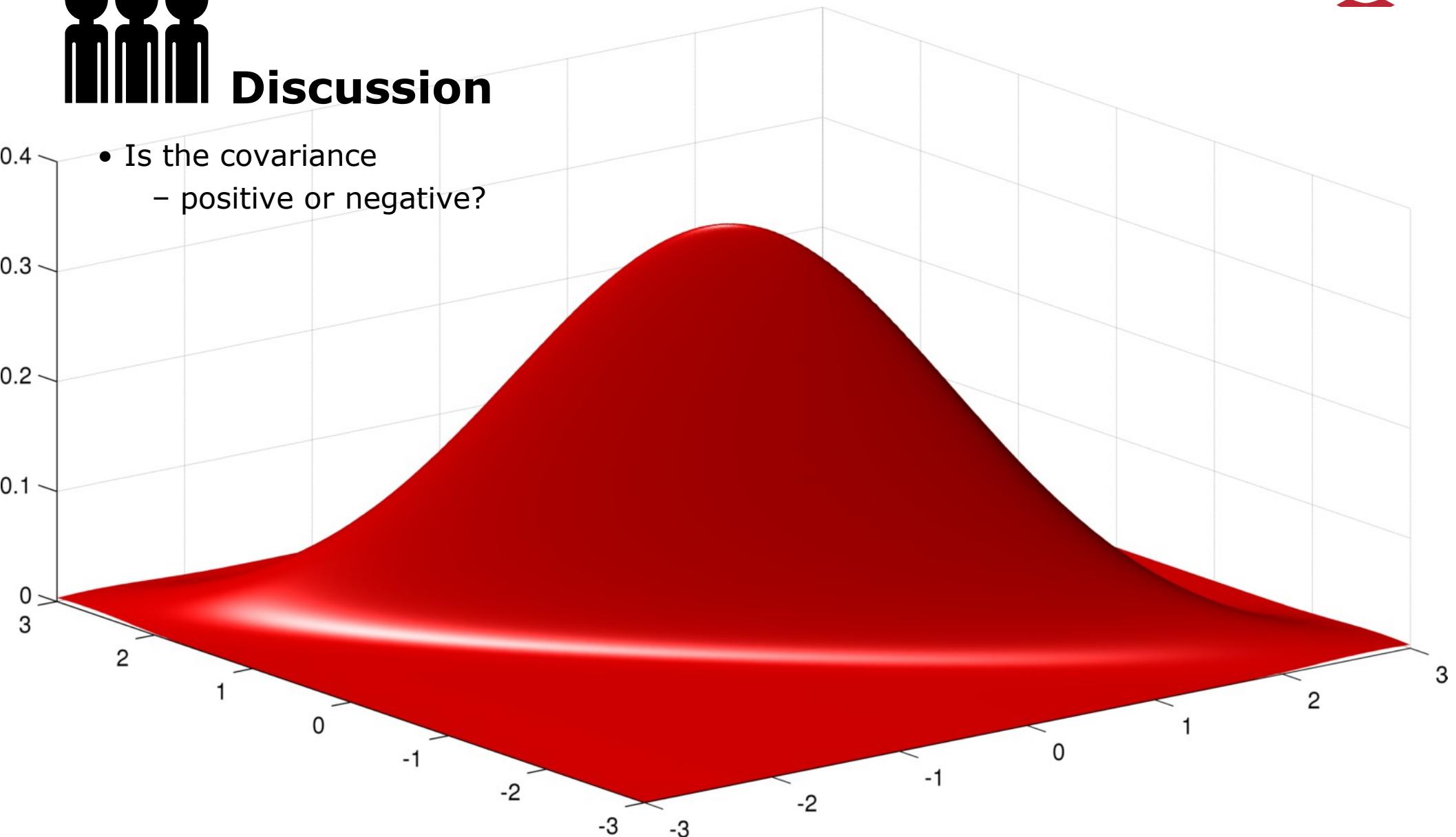
- Is the covariance
 - positive or negative?





Discussion

- Is the covariance
 - positive or negative?



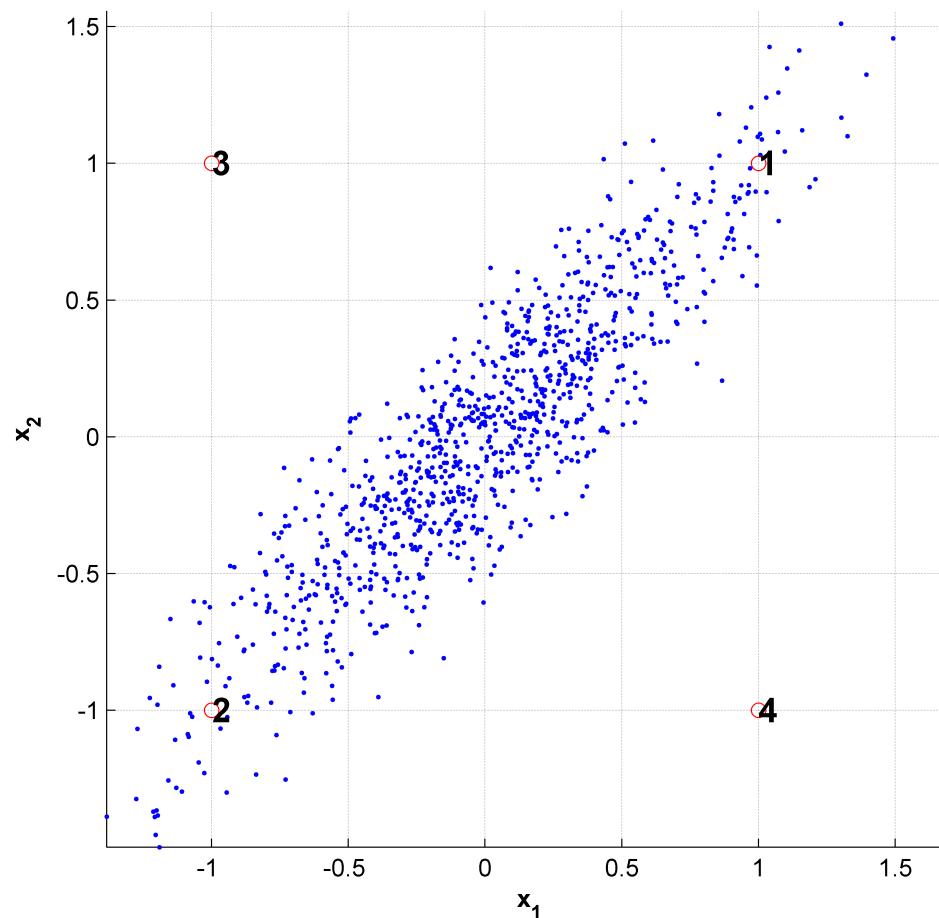
The Mahalanobis distance

How far are x_1 and x_2 apart?

- $\text{mahalanobis}(x_1, x_2) = 4.2$
- $d_{\text{Euclidean}}(x_1, x_2)^2 = 8.0$

How far are x_3 and x_4 apart?

- $\text{mahalanobis}(x_3, x_4) = 80$
- $d_{\text{Euclidean}}(x_3, x_4)^2 = 8.0$



$$\text{mahalanobis}(\mathbf{x}, \mathbf{y}) = (\mathbf{x} - \mathbf{y})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \mathbf{y})$$

$$d_{\text{Euclidean}}(\mathbf{x}, \mathbf{y})^2 = (\mathbf{x} - \mathbf{y})^\top \mathbf{I}^{-1} (\mathbf{x} - \mathbf{y})$$

02450 Introduction to machine learning and data modeling

Reading material

Tan, Steinbach and Kumar
"Introduction to Data Mining"

Section 3.3

Group(s) of the day:

Martin Christiansen

Marie Højén

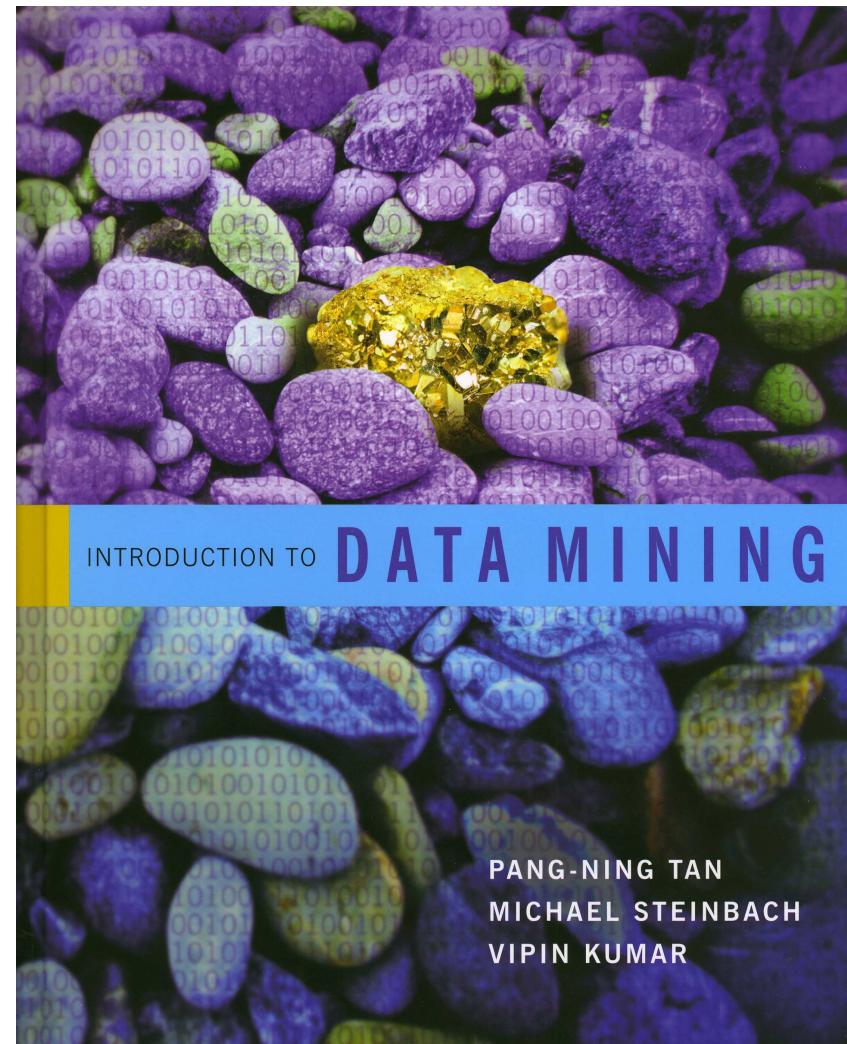
Andreas Borup Svendsen

Kåre Wedel Jacobsen

Hao Tong

Henri Nikula

Loris Dal Lago



Lecture schedule

1. Introduction
(Tan 1.1-1.4)

Data: Feature extraction and visualization

2. Data and feature extraction
(Tan 2.1-2.3 + B1 (+ A))
3. Measures of similarity and summary statistics
(Tan 2.4 + 3.1-3.2 + C1-C2)

4. Data visualization *(Tan 3.3)*

Supervised learning: Classification and regression

5. Decision trees and linear regression
(Tan 4.1-4.3 + D)
6. Overfitting and performance evaluation
(Tan 4.4-4.6)
7. Nearest neighbor, naive Bayes, and artificial neural networks
(Tan 5.2-5.4)

8. Ensemble methods and multi class classifiers
(Tan 5.6-5.8)

Unsupervised learning: Clustering and density est.

9. K-means and hierarchical clustering
(Tan 8.1-8.3, 8.5.7)
10. Mixture models and association mining
(Tan 9.2.2 + 6.1-6.3)
11. Density estimation and anomaly detection
(Tan 10.1-10.4)

Machine learning and data modelling in practice

12. Putting it all together: Summary and overview
13. Mini project

Probabilities (revisited from last week)

- Basic rules of probability

- Sum rule

$$p(x) = \sum_y p(x, y)$$

- Product rule

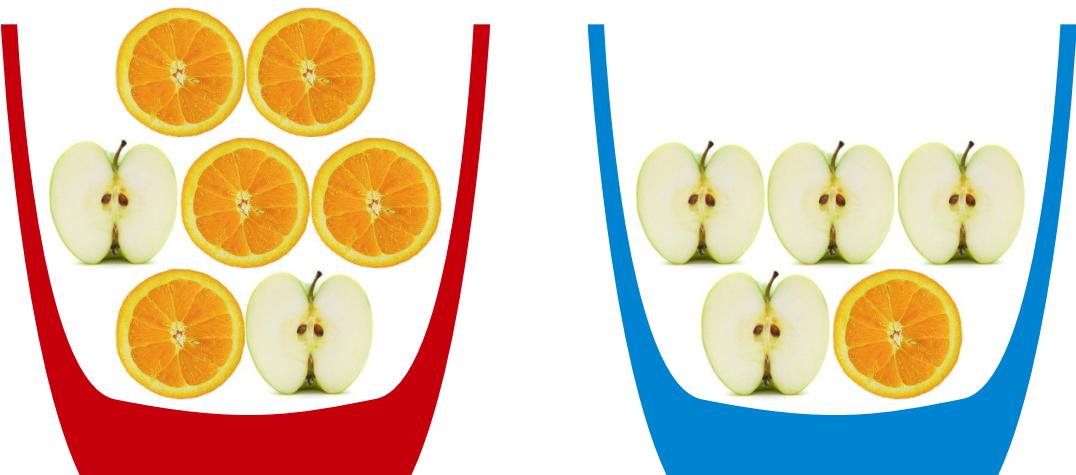
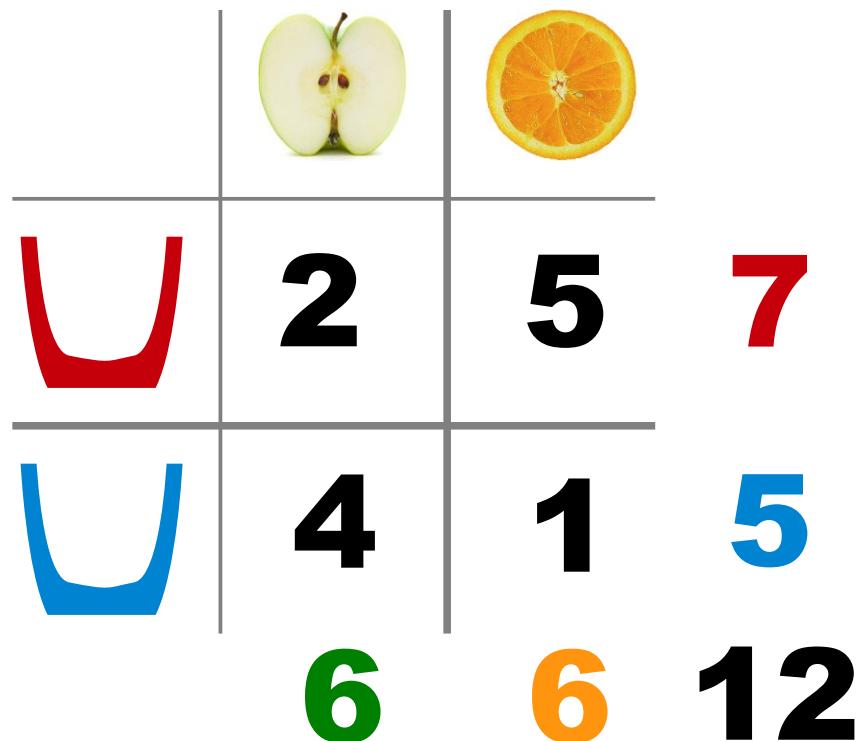
$$p(x, y) = p(x|y)p(y)$$

- Bayes' rule

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)}$$

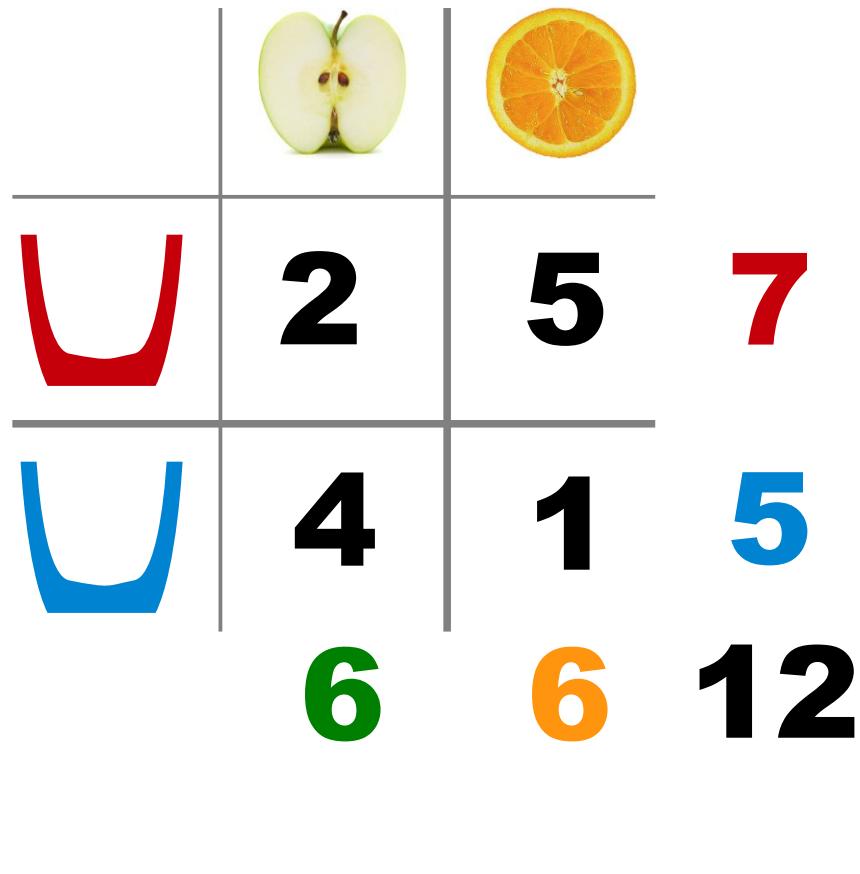
Probabilities

- What is the probability of an **orange** if the bowl is **red**?
- What is the probability of the **red** bowl if the fruit is **orange**?



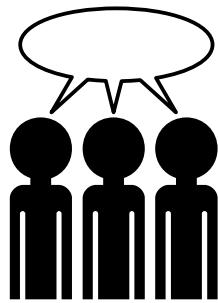
Probabilities

- What is the probability of an **orange** if the bowl is **red**?
- What is the probability of the **red** bowl if the fruit is **orange**?



$$p(o|r) = \frac{p(r,o)}{p(r)} = \frac{\mathbf{5/12}}{\mathbf{7/12}} = \mathbf{5/7}$$

$$\begin{aligned}
 p(r|o) &= \frac{p(r,o)}{p(o)} = \frac{\mathbf{5/12}}{\mathbf{6/12}} = \mathbf{5/6} \\
 &= \frac{p(o|r)p(r)}{p(o)} \\
 &= \frac{\mathbf{5/7} \cdot \mathbf{7/12}}{\mathbf{6/12}} = \mathbf{5/6}
 \end{aligned}$$



The news paper “Slam the Glam”

News media agency Reuters Bureau sends news stories to the tabloid paper “Slam the Glam”:

- 80% of the news stories from Reuters are positive and 20% of the news stories are negative.
- 90% of the negative news stories are published in “Slam the Glam” while only 5% of the positive stories are published.

What is the probability that a published story (from Reuters) is positive?

Hints:

- *Sum Rule*

$$p(x) = \sum_y p(x, y)$$

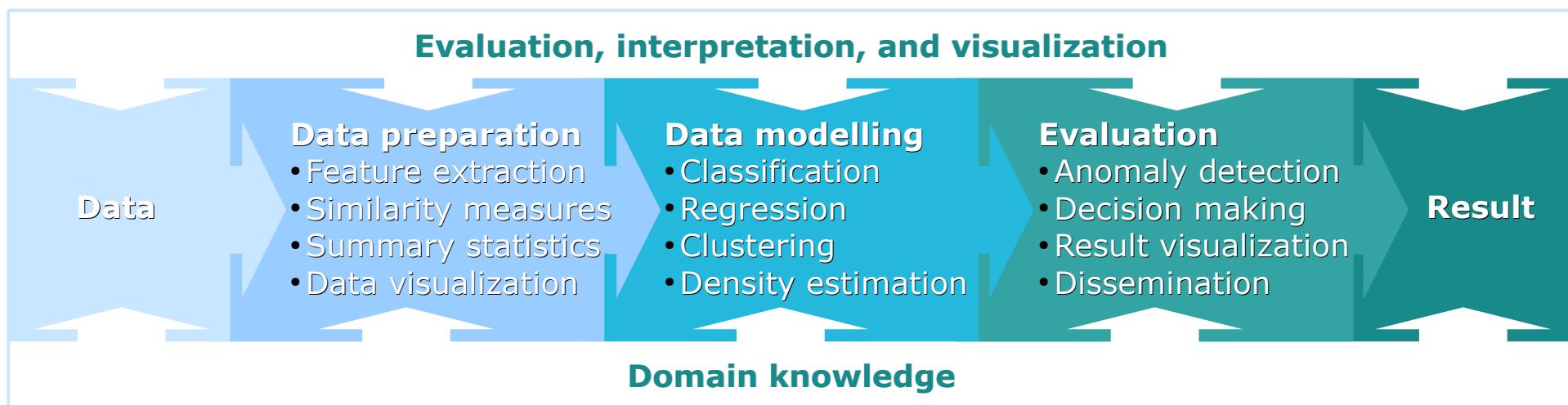
- *Product rule*

$$p(x, y) = p(x|y)p(y)$$

- *Bayes theorem*

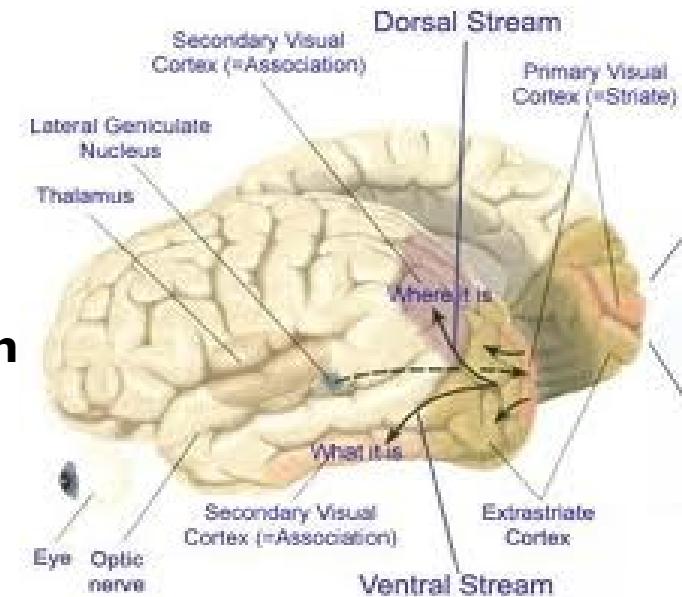
$$p(x|y) = \frac{p(y|x)p(x)}{p(y)}$$

Data modeling framework



Visualization

- **Mapping of data to a visual format**
 - Data objects, attributes, and relations
- **Humans are good at understanding visual information**
 - See patterns and trends
 - Detect outliers



The adage "**A picture is worth a thousand words**" refers to the idea that a complex idea can be conveyed with just a single still image. It also aptly characterizes one of the main goals of visualization, namely making it possible to absorb large amounts of data quickly. (http://en.wikipedia.org/wiki/A_picture_is_worth_a_thousand_words)

Information visualization presumes that "**visual representations and interaction techniques take advantage of the human eye's broad bandwidth pathway into the mind** to allow users to see, explore, and understand large amounts of information at once. Information visualization focused on the creation of approaches for conveying abstract information in intuitive ways." James J. Thomas and Kristin A. Cook (2005)

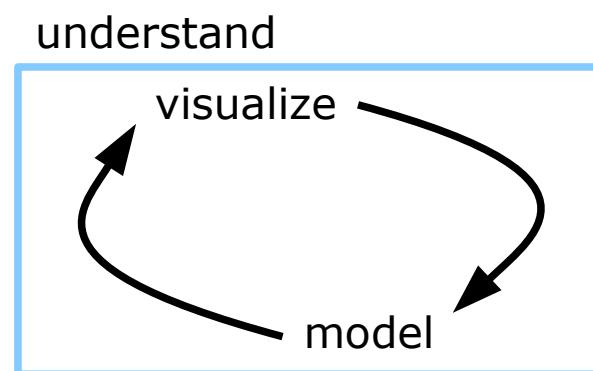
Visualization

- **Visualization**

- Uncovers the unexpected
- Cognitive bias

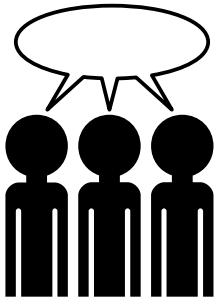
- **Modeling**

- Mathematically well founded
- Discovers what we anticipate



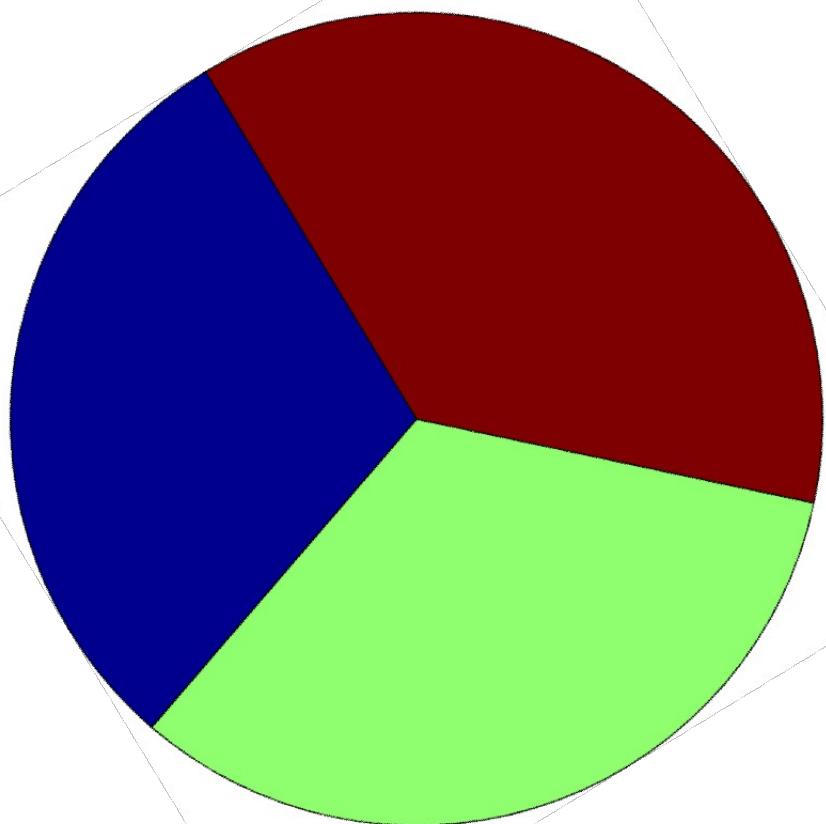
Visualization

- **Representation:** How will you map objects, attributes, and relations to visual elements?
 - Positions, lengths, colors, areas, orientation
- **Arrangement:** How will you display the visual elements?
 - Viewpoint, transparency, separation, grouping
- **Selection:** How will you handle a large number of attributes and data objects?
 - Display a subset, focus on a region of interest, show summaries



Representation

- **Area represents proportion**
 - Which is smallest, middle, and largest?
 - What are the proportions approximately?

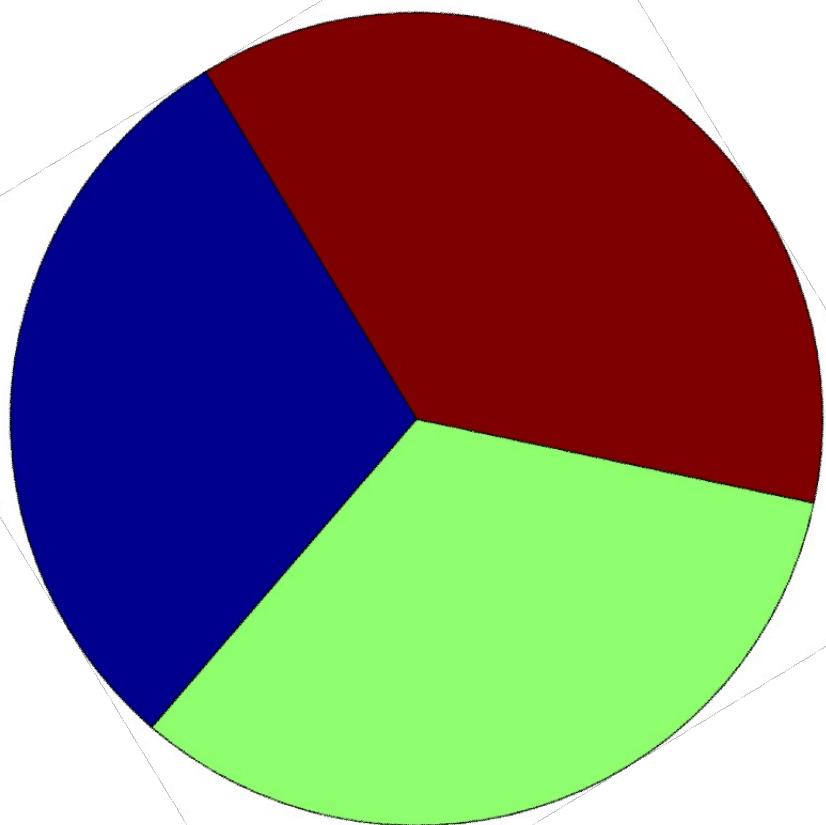




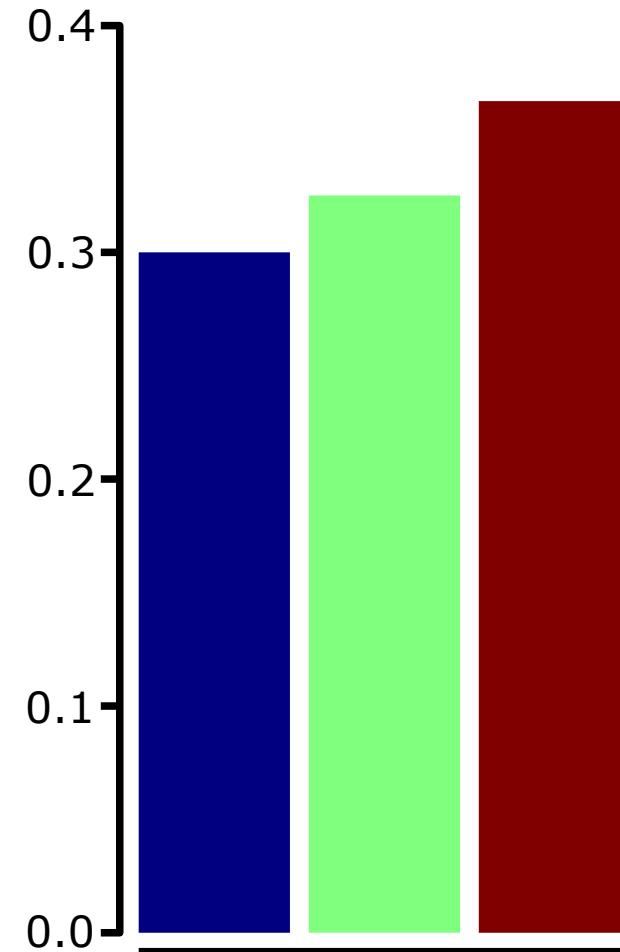
Representation

- **Area represents proportion**

- Which is smallest, middle, and largest?
- What are the proportions approximately?



- **Height represents proportion**



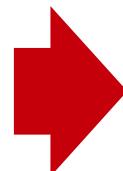
Arrangement

- **Placement of visual elements**

- Can make a great difference in how easy it is to interpret data

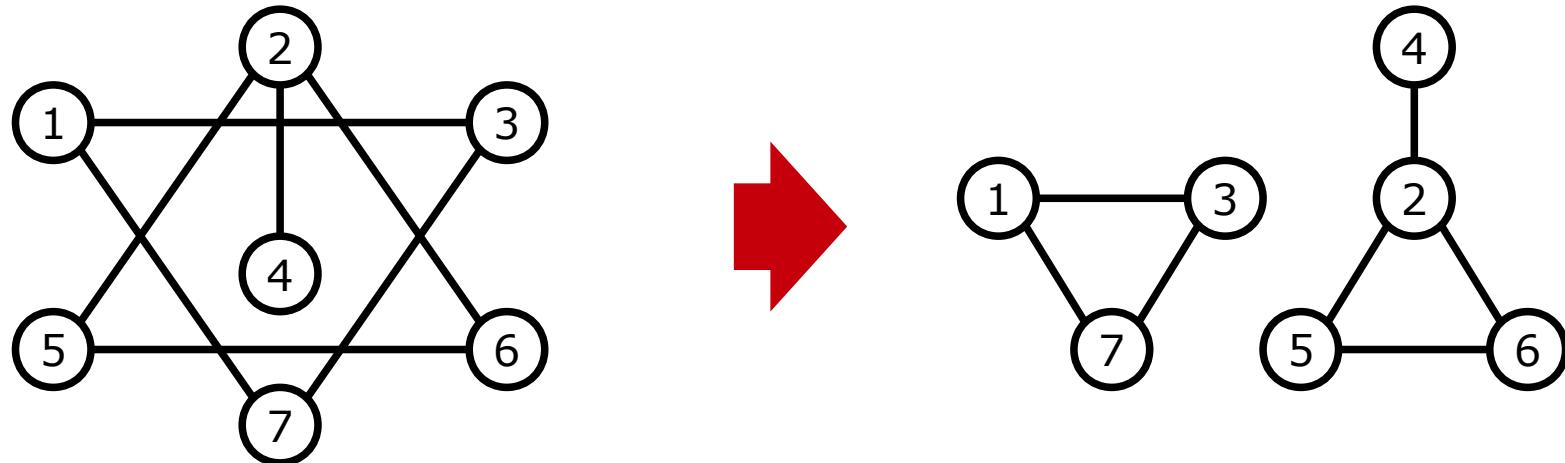
- **Example**

	1	2	3	4	5	6
1	0	1	0	1	1	0
2	1	0	1	0	0	1
3	0	1	0	1	1	0
4	1	0	1	0	0	1
5	0	1	0	1	1	0
6	1	0	1	0	0	1
7	0	1	0	1	1	0
8	1	0	1	0	0	1
9	0	1	0	1	1	0



	6	1	3	2	5	4
4	1	1	1	0	0	0
2	1	1	1	0	0	0
6	1	1	1	0	0	0
8	1	1	1	0	0	0
5	0	0	0	1	1	1
3	0	0	0	1	1	1
9	0	0	0	1	1	1
1	0	0	0	1	1	1
7	0	0	0	1	1	1

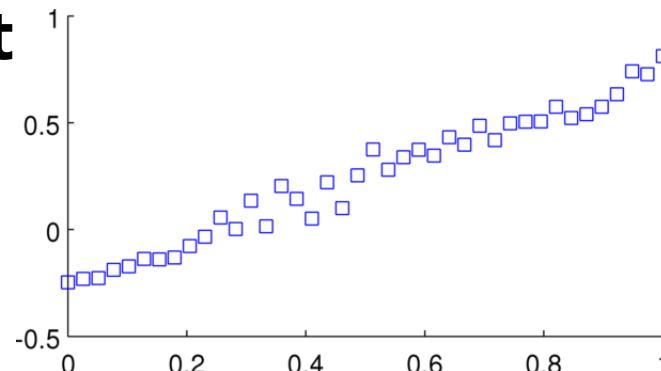
Arrangement



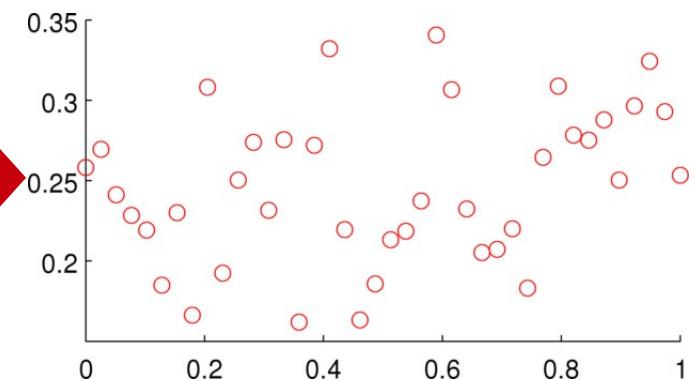
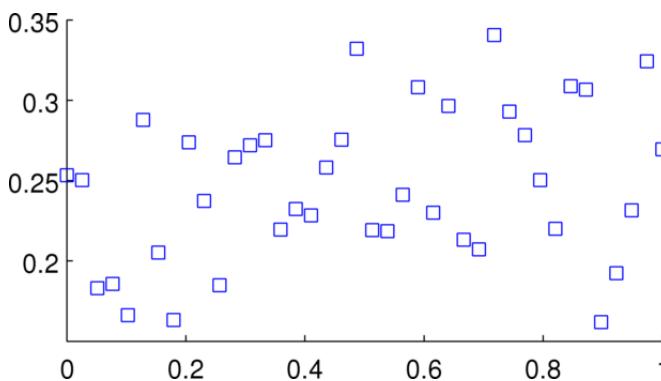
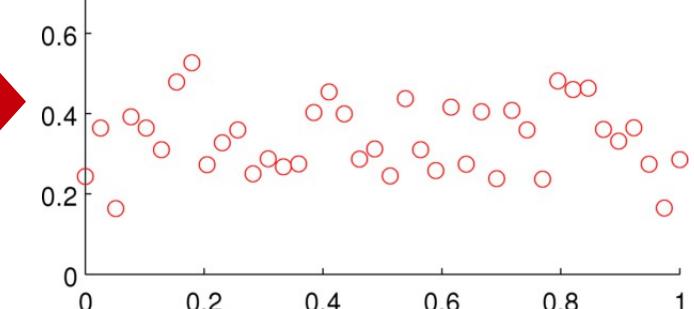
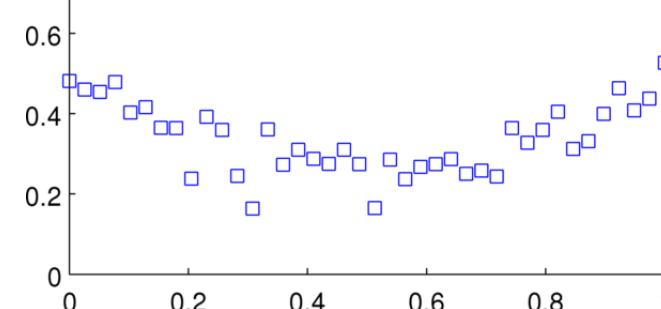
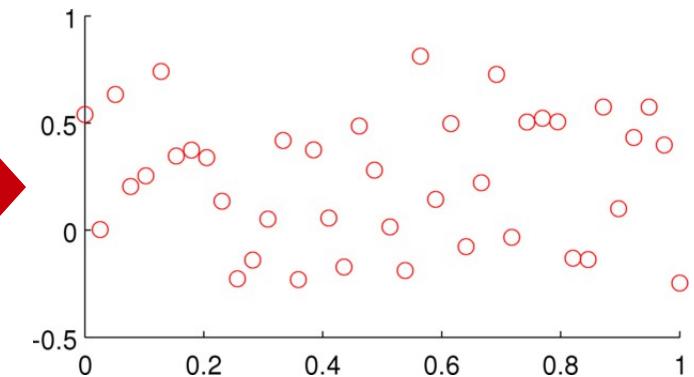
Arrangement

- Plots of permuted data can help reveal structure

Original



Permuted



```
scatter(x,y, 's')
```

```
ind=randperm(length(y));  
scatter(x,y(ind), 'r')
```

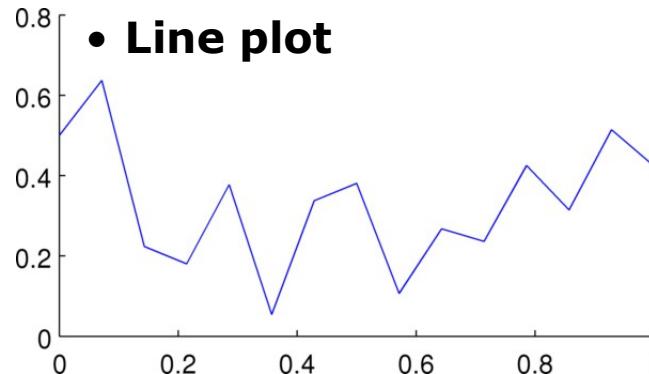
Selection

- Elimination or de-emphasis of certain objects or attributes
- A subset of **attributes**
 - **Why?** A graph can only show so many attributes – focus on the relevant
 - **How?**
 - Dimensionality reduction
 - Plot pairs of attributes
- A subset of **objects**
 - **Why?** A graph can only show so many objects – focus on the relevant
 - **How?**
 - Random sampling
 - Display of region of interest

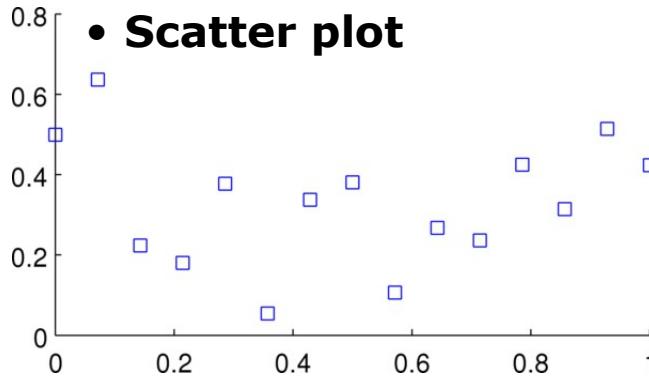
Types of plots

- **Basic plots**
- **Distribution of a single attribute**
 - Histogram
 - Empirical cumulative distribution
 - Percentile plots
 - Box plot
- **Relation between attributes**
 - 2-d histogram
 - Scatter plots
- **Visualization of high-dimensional objects**
 - Matrix plots
 - Parallel coordinates
 - Star plots

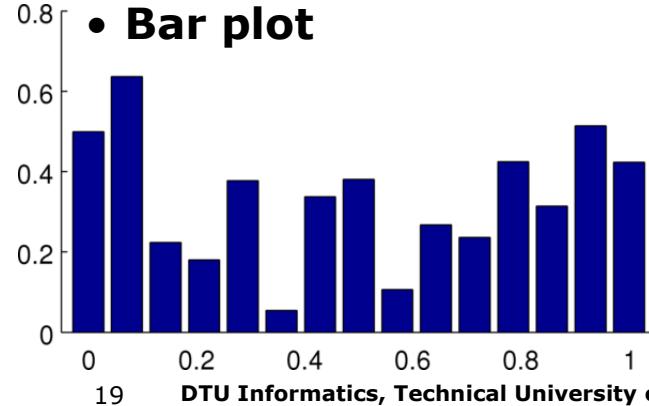
Basic plots



```
plot(x,y);
```



```
plot(x,y,'s');
scatter(x,y,'s')
```



```
bar(x,y);
```

The iris data set

- **Three flowers**
 - 50 instances of each class, 150 in total
- **Attributes**
 - Sepal (outermost leaves)
 - length in cm
 - width in cm
 - Petal (innermost leaves)
 - length in cm
 - width in cm
 - Class of flower
 - Iris Setosa
 - Iris Versicolour
 - Iris Virginica

Flower ID	Attribute			
	Sepal Length	Sepal Width	Petal Length	Petal Width
1	5.1	3.5	1.4	0.2
2	4.9	3.0	1.4	0.2
3	4.7	3.2	1.3	0.2
4	4.6	3.1	1.5	0.2
.
.
150	5.9	3.0	5.1	1.8



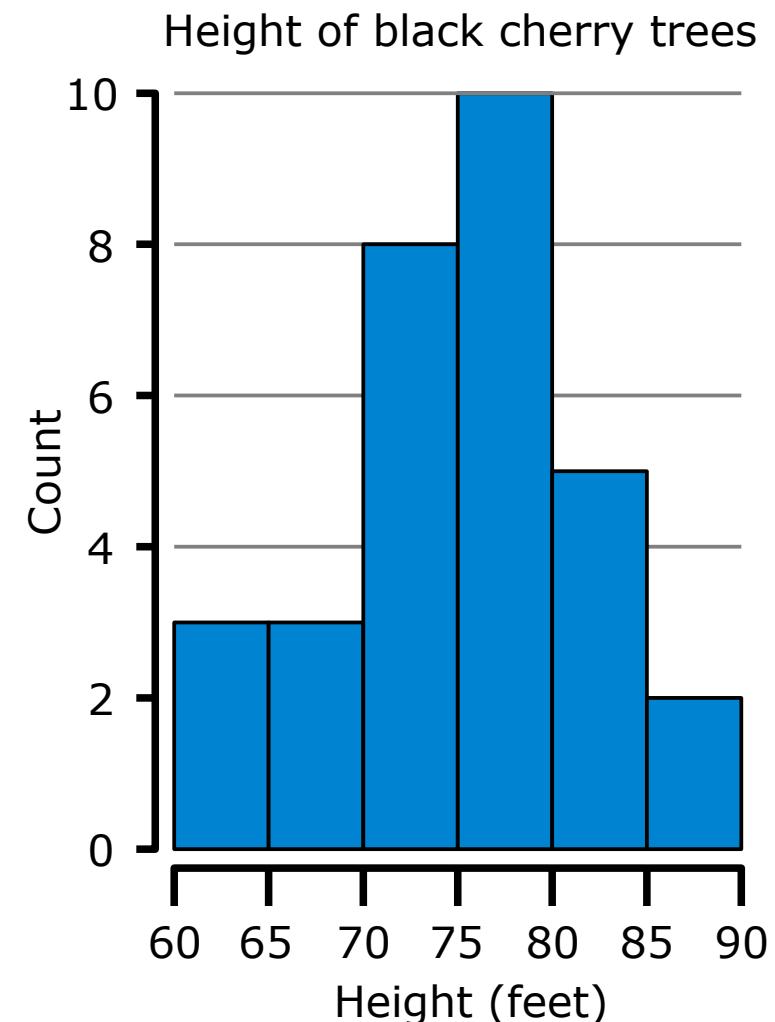
$X^{\text{Observation} \times \text{Attribute}}$

Matlab var.	Type	Size	Description
X	Numeric	$N \times M$	Data matrix: The rows correspond to N data objects, each of which contains M attributes.
y	Numeric	$N \times 1$	Class index: For each data object, y contains a class index, $y \in \{0, 1, \dots, C - 1\}$, where C is the total number of classes.
classNames	Cell array	$C \times 1$	Class names: Name (string) for each of the C classes.
attributeNames	Cell array	$M \times 1$	Attribute names: Name (string) for each of the M attributes.
N	Numeric	Scalar	Number of data objects.
M	Numeric	Scalar	Number of attributes.
C	Numeric	Scalar	Number of classes.

Distribution of a single attribute

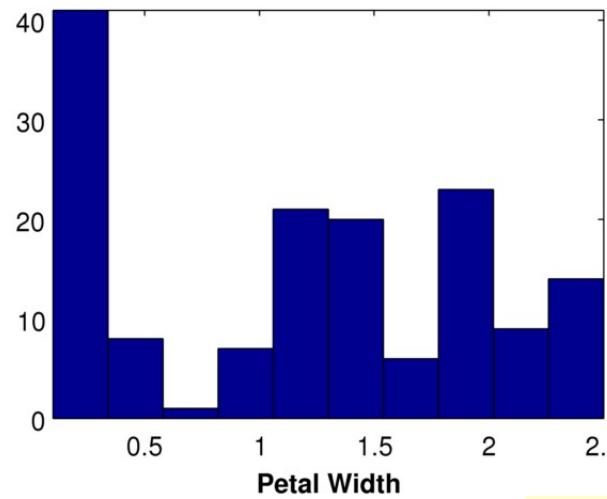
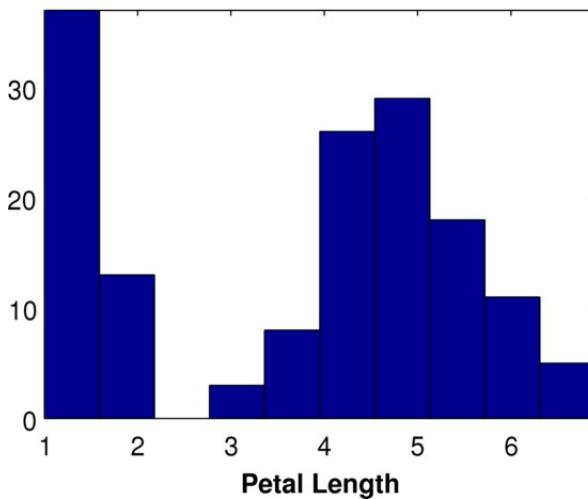
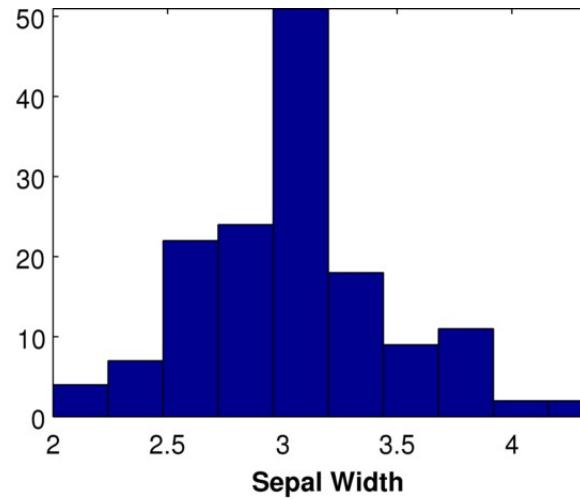
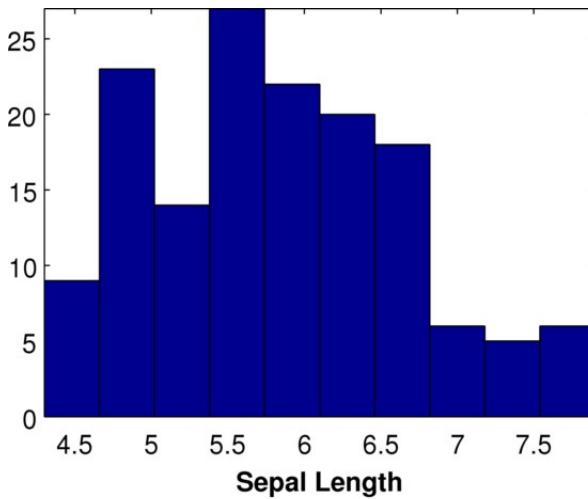
Histograms

- **Shows distribution of a single variable**
 - Divide the values into bins
 - Bar plot of the number of values in bin
 - Height indicates count of values
 - Shape determined by
 - Distribution of data
 - Number of bins / bin width



$$H = \{60, 64, 64, 66, 67, 69, 71, 72, 72, 72, 72, 73, 74, 74, 75, 75, 76, 76, 76, 77, 77, 78, 78, 79, 80, 80, 81, 82, 84, 85, 85, 89\}$$

Histograms of the Iris data attributes

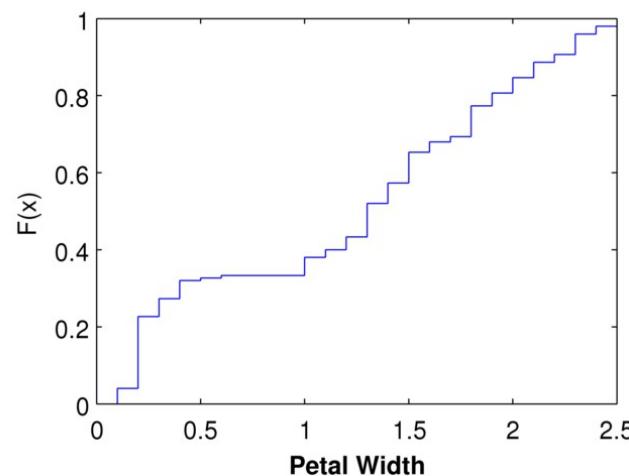
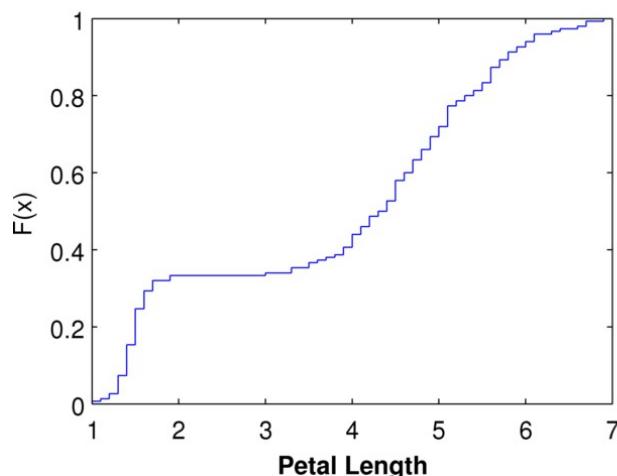
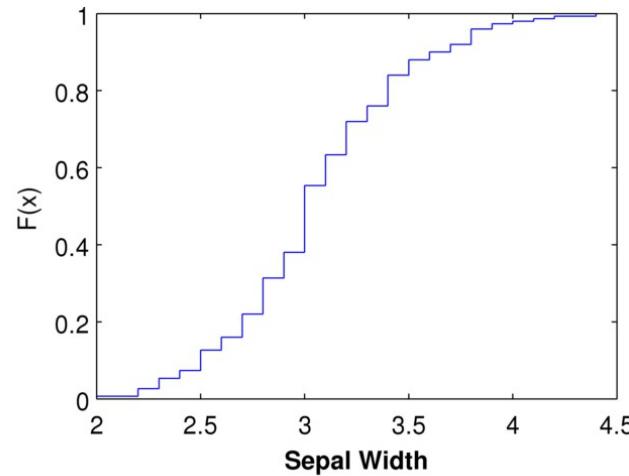
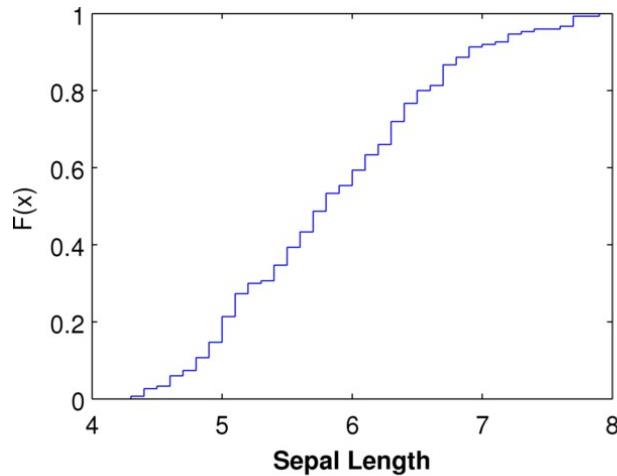


```

for m=1:M
    subplot(2,2,m)
    hist(X(:,m),10);
    axis tight;
    xlabel(attributeNames{m})
end

```

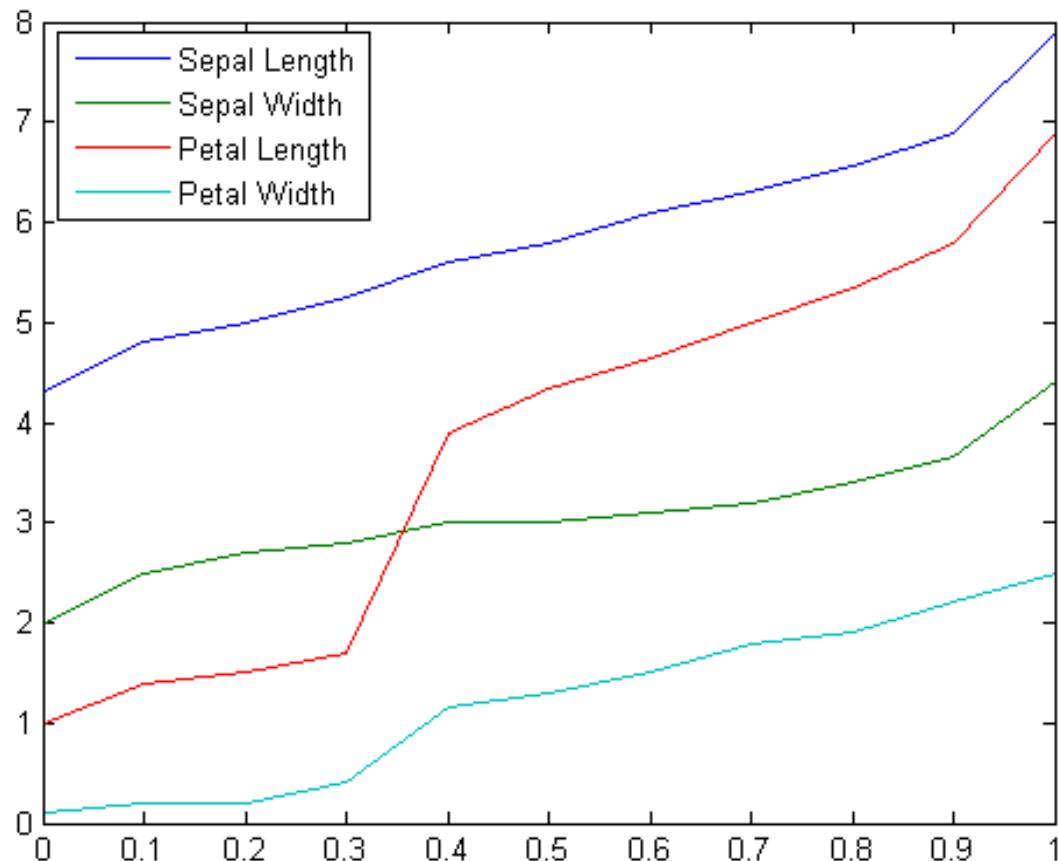
Empirical cumulative distributions



```
for m=1:M
    subplot(2,2,m);
    ecdf(X(:,m));
    xlabel(attributeNames{m});
end
```

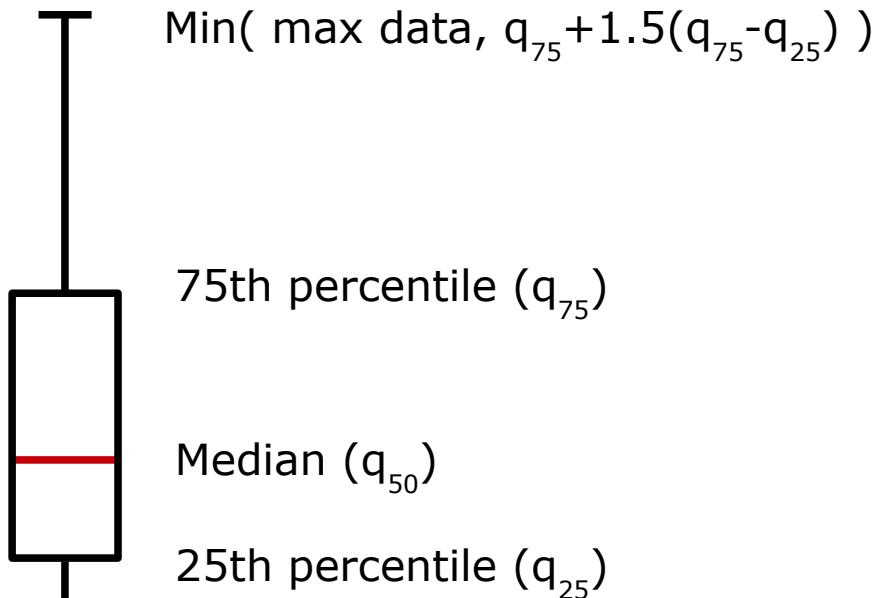
Percentile plots

Percentiles: Given an ordinal or continuous attribute \mathbf{x} and a number \mathbf{p} between 0 and 100, the \mathbf{p} th percentile is a value \mathbf{x}_p of \mathbf{x} such that \mathbf{p} percent of the observed values of \mathbf{x} are less than \mathbf{x}_p .

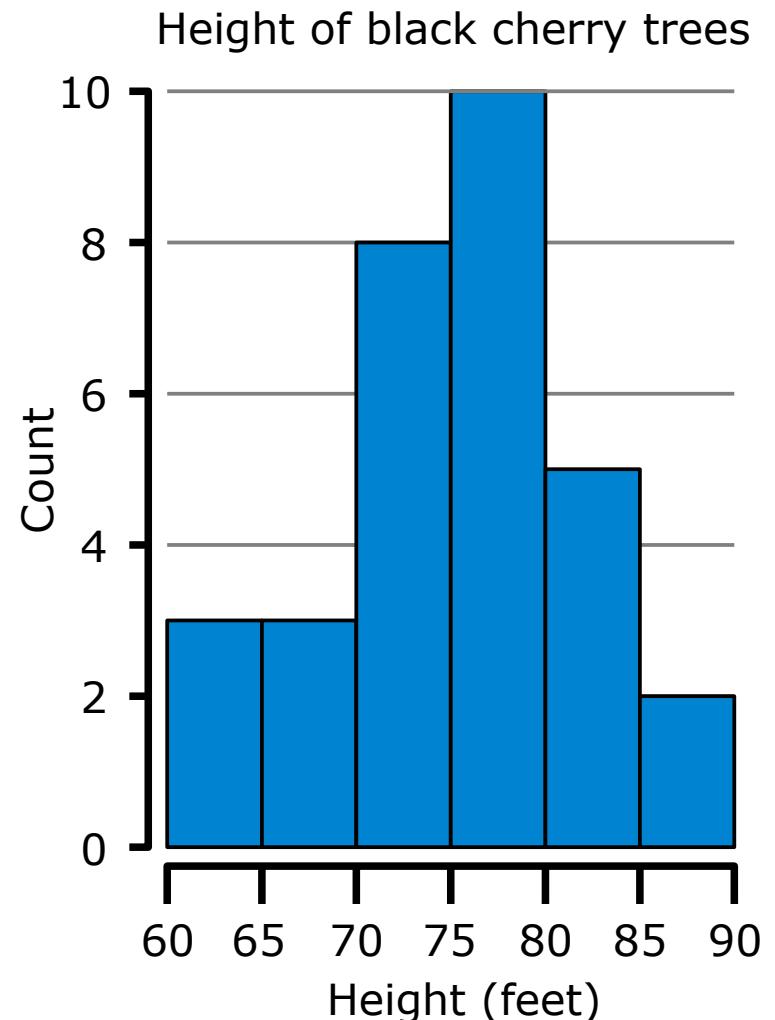


```
prctile = 0:0.1:1;
Y = quantile(X,prctile);
plot(prctile,Y);
legend(attributeNames);
```

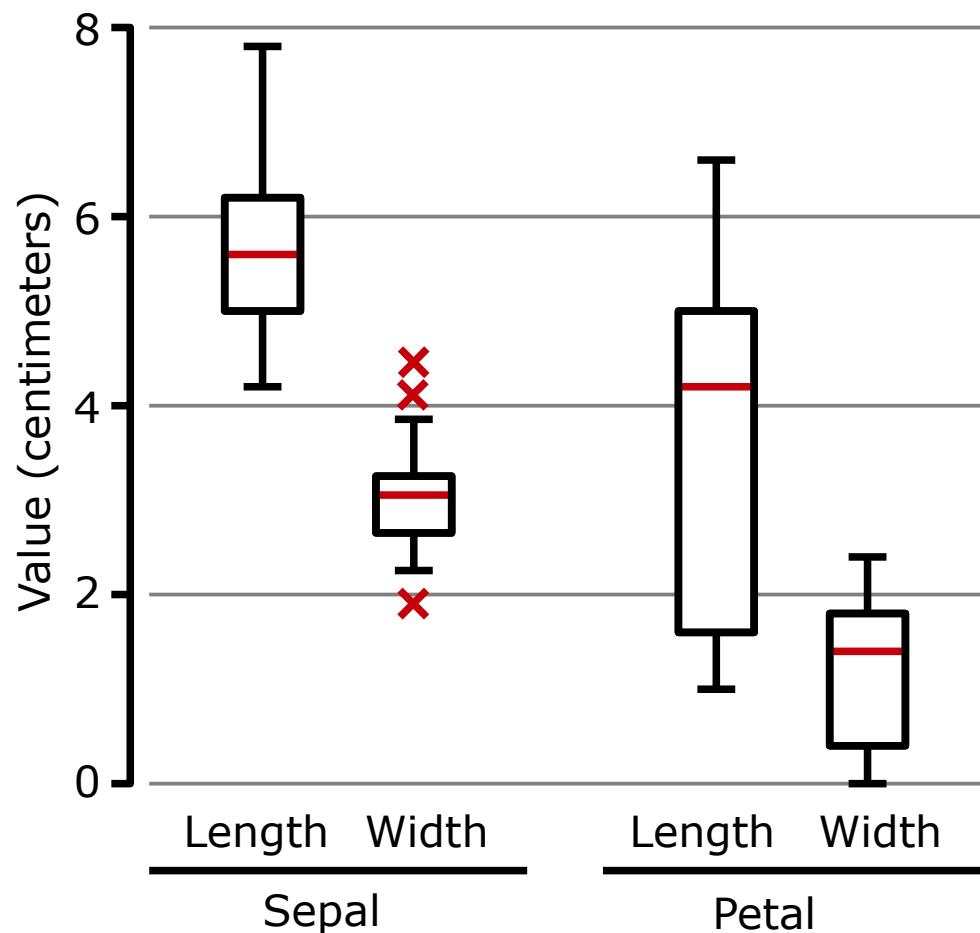
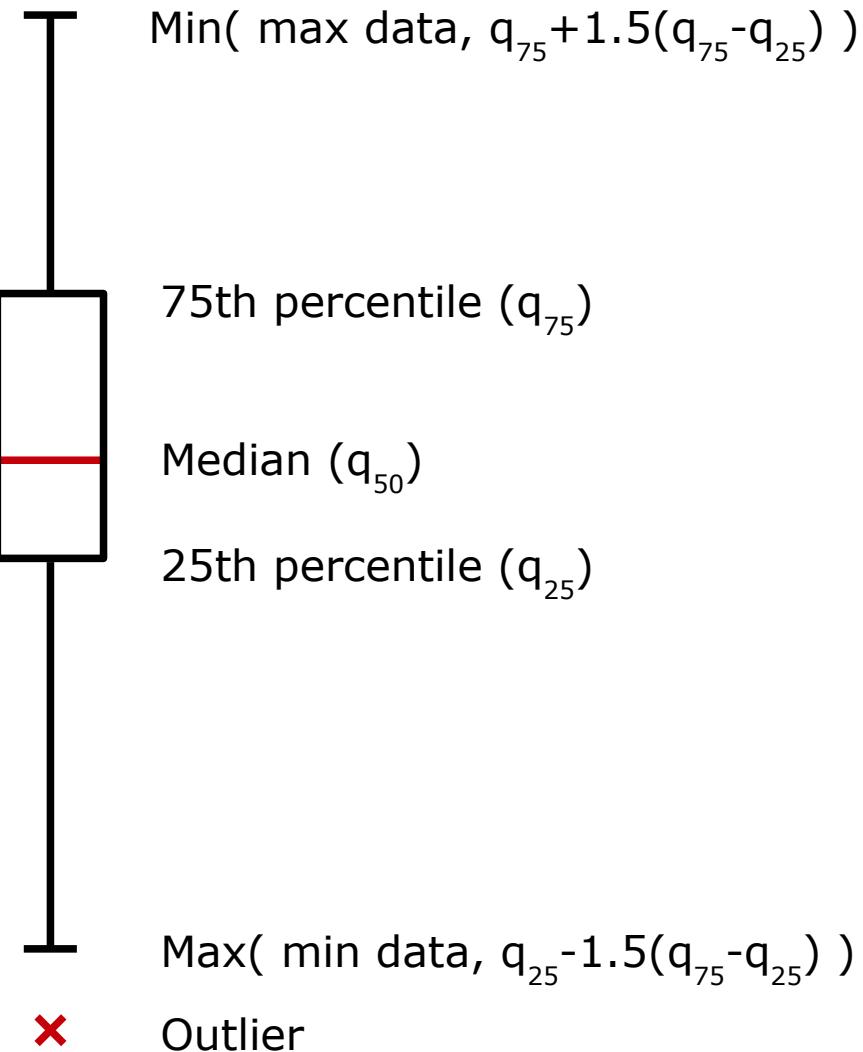
Box plots



The plotted whisker extends to the adjacent value, which is the most extreme data value that is not an outlier.

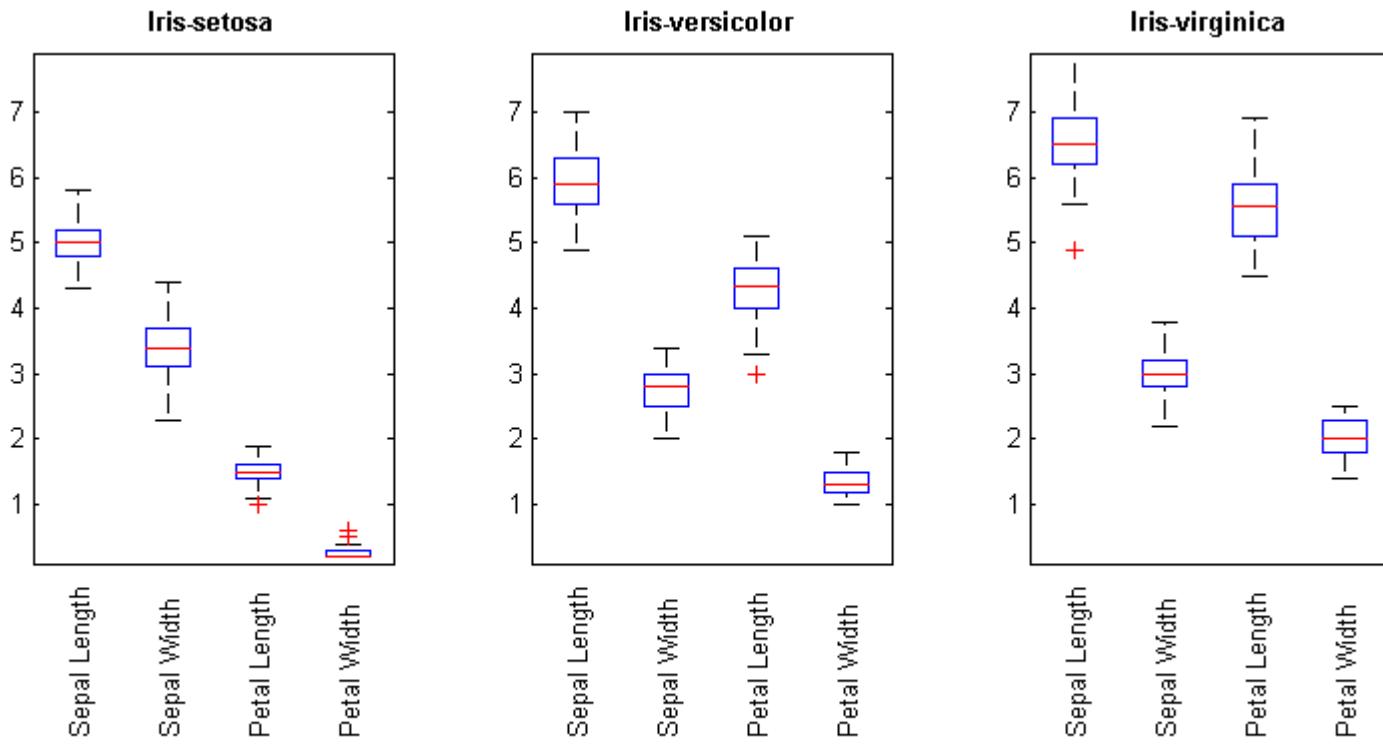


Box plots



The plotted whisker extends to the adjacent value, which is the most extreme data value that is not an outlier.

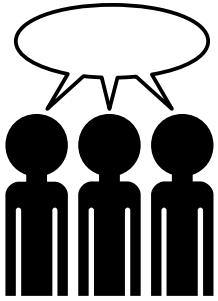
Box plots



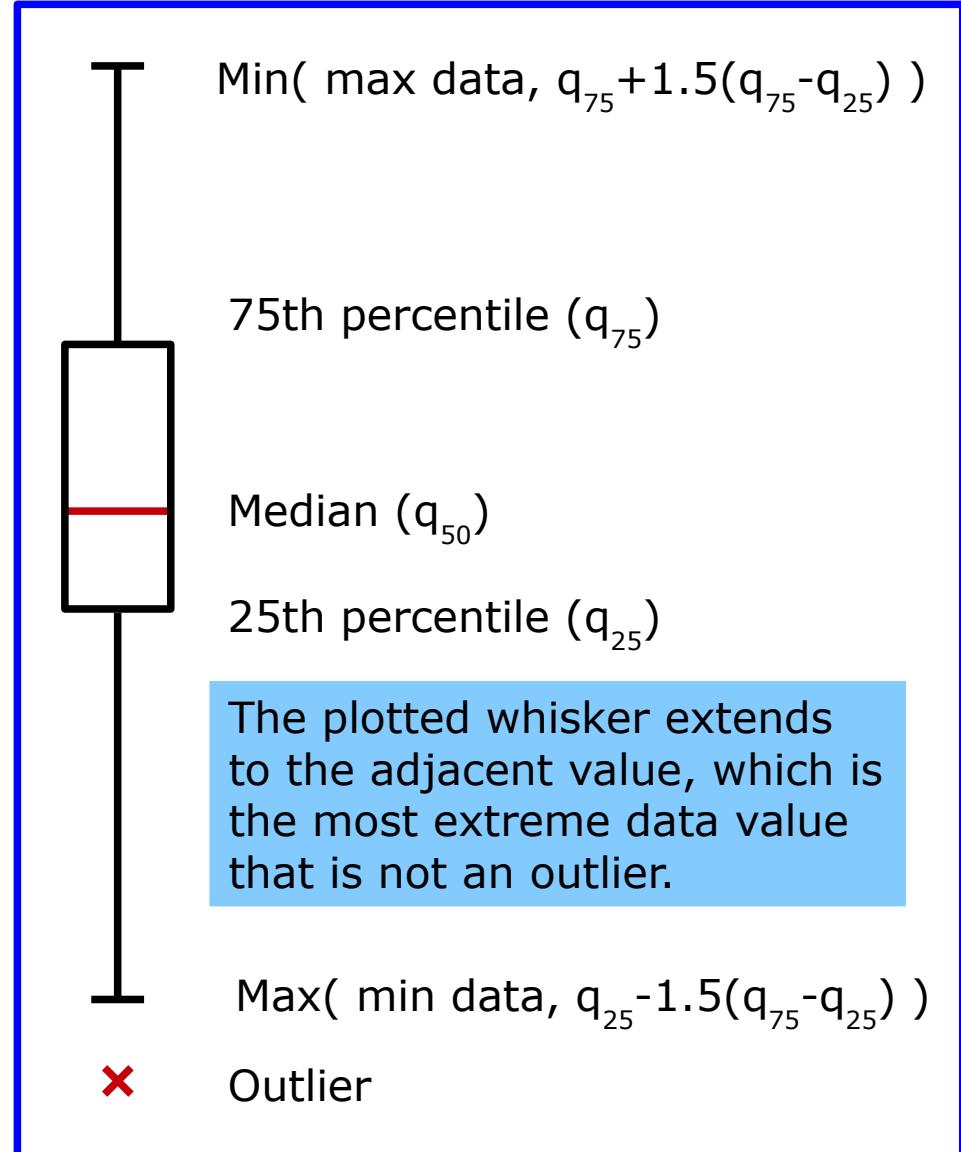
```

y_lim=[min(min(X)) max(max(X)) ];
for c = 1:C
    subplot(1,C,c)
    boxplot(X(y==c-1,:),'labels',attributeNames,'labelorientation','inline');
    axis([0.5 M+0.5,y_lim])
    title(classNames{c}, 'FontWeight','bold');
end

```



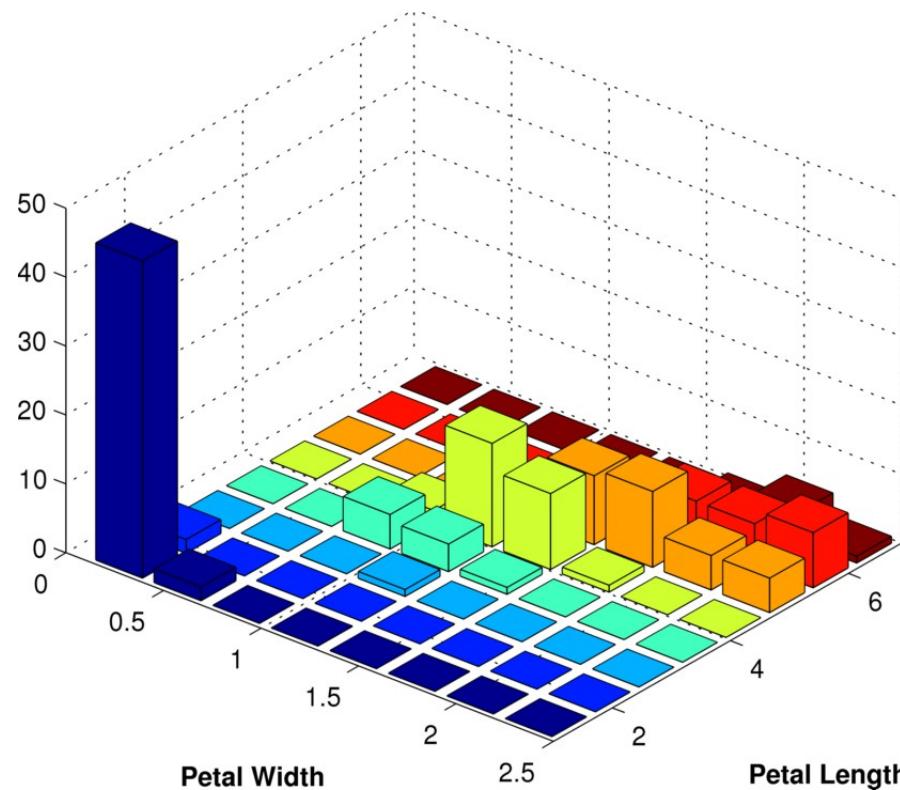
Draw the Box-plot for the dataset $D=\{0, 1, 1, 3, 3, 4, 4, 10\}$



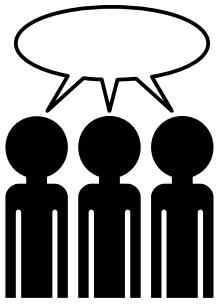
Relation between attributes

Two-dimensional histograms

- Shows joint distribution of two variables

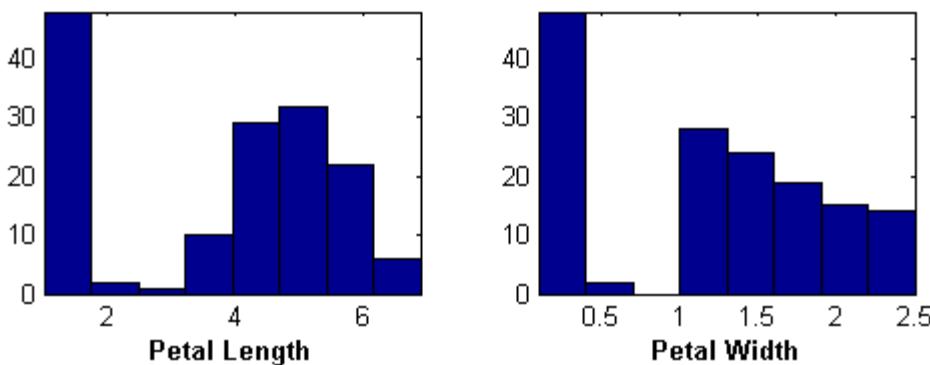


```
nBins = 8;
ind = [3 4]; % Indices of the two variables to create the 2D histogram from
[n,x,data] = hist2d(X(:,ind)',nBins);
bar3xy(x(1,:),x(2,:),n);
axis tight;
xlabel(attributeNames{ind(1)},'FontWeight','bold')
ylabel(attributeNames{ind(2)},'FontWeight','bold')
```

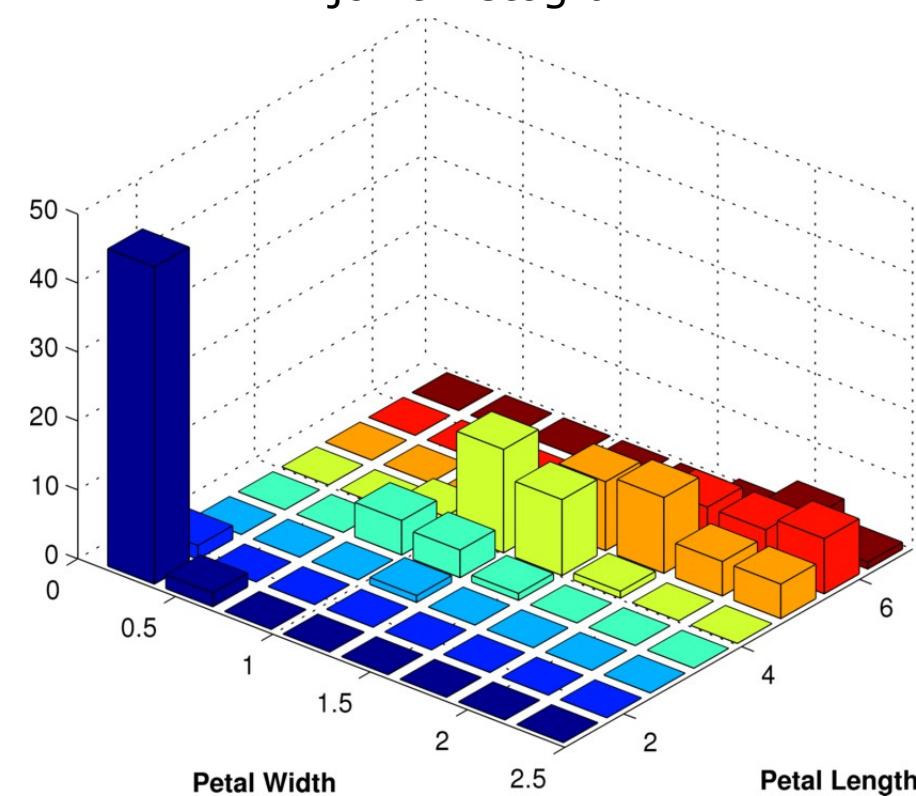


Let B be the number of bins for each mode in the 1D and 2D histogram (B=8 in the figures below). How many values are there to be estimated in the 1D and 2D histogram as a function of B? What happen to the number of estimated values if we estimated a 3D histogram of the joint distribution between three attributes? What would happen if we estimated a MD histogram of the joint distribution of M attributes?

1D histogram of each attribute

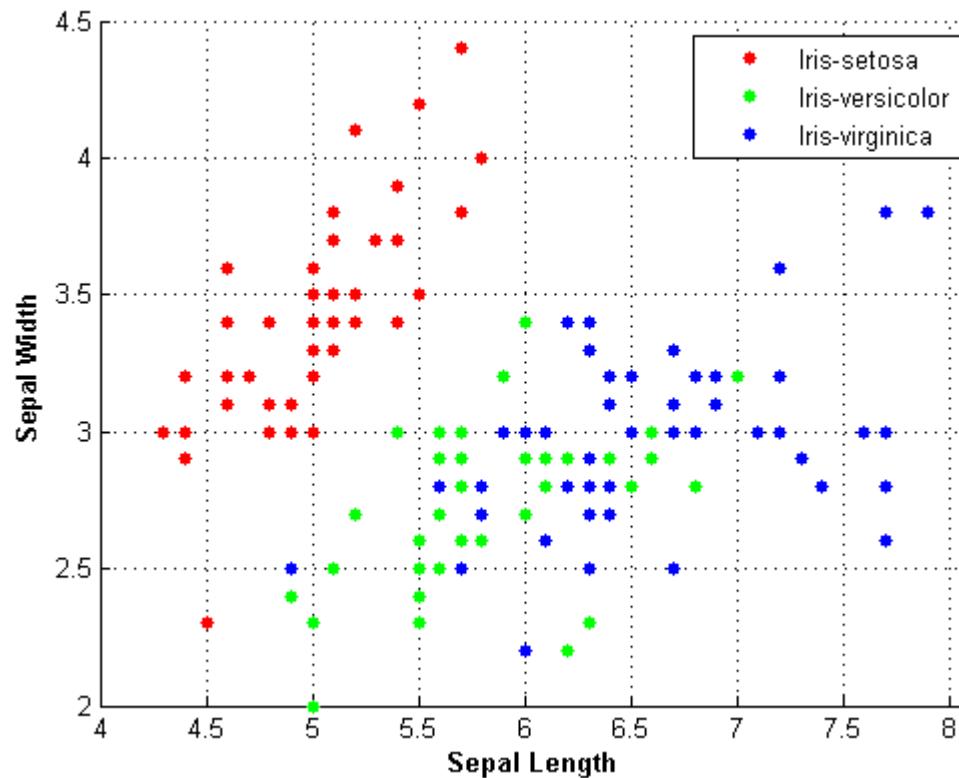


2D joint histogram



Scatter plots

- Shows **relation** between attributes
 - Assess dependence between attributes
 - Used with classes to assess separability



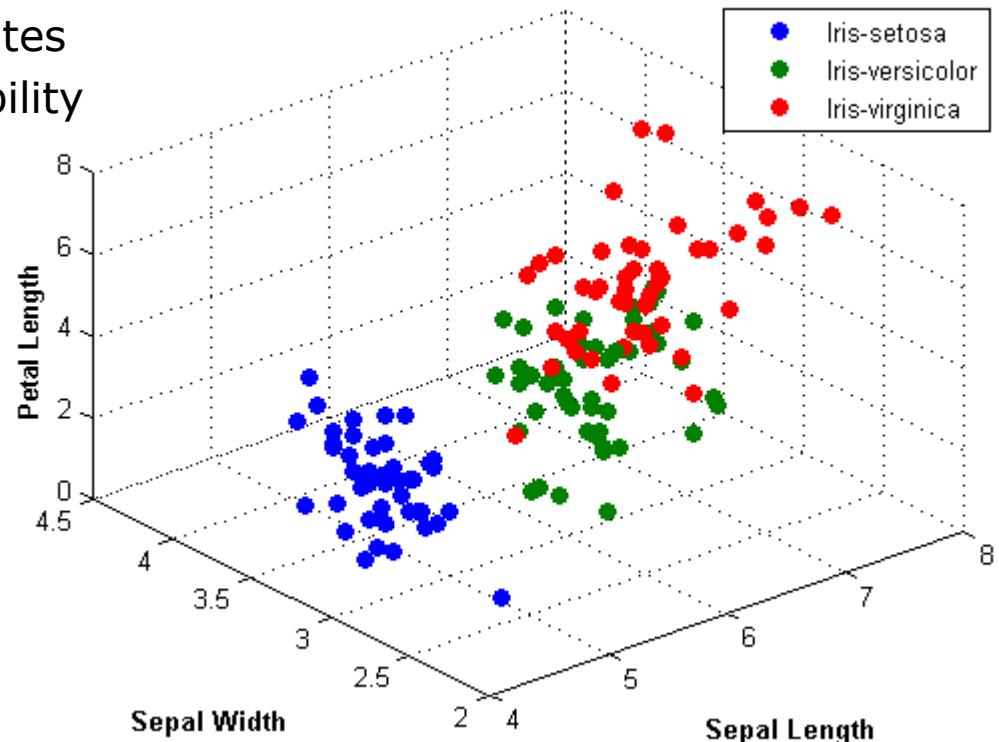
```

gscatter(X(:,1),X(:,2),classNames(y+1));
grid on;
xlabel(measurementType(1), 'fontweight', 'bold');
ylabel(measurementType(2), 'fontweight', 'bold');

```

Scatter plots

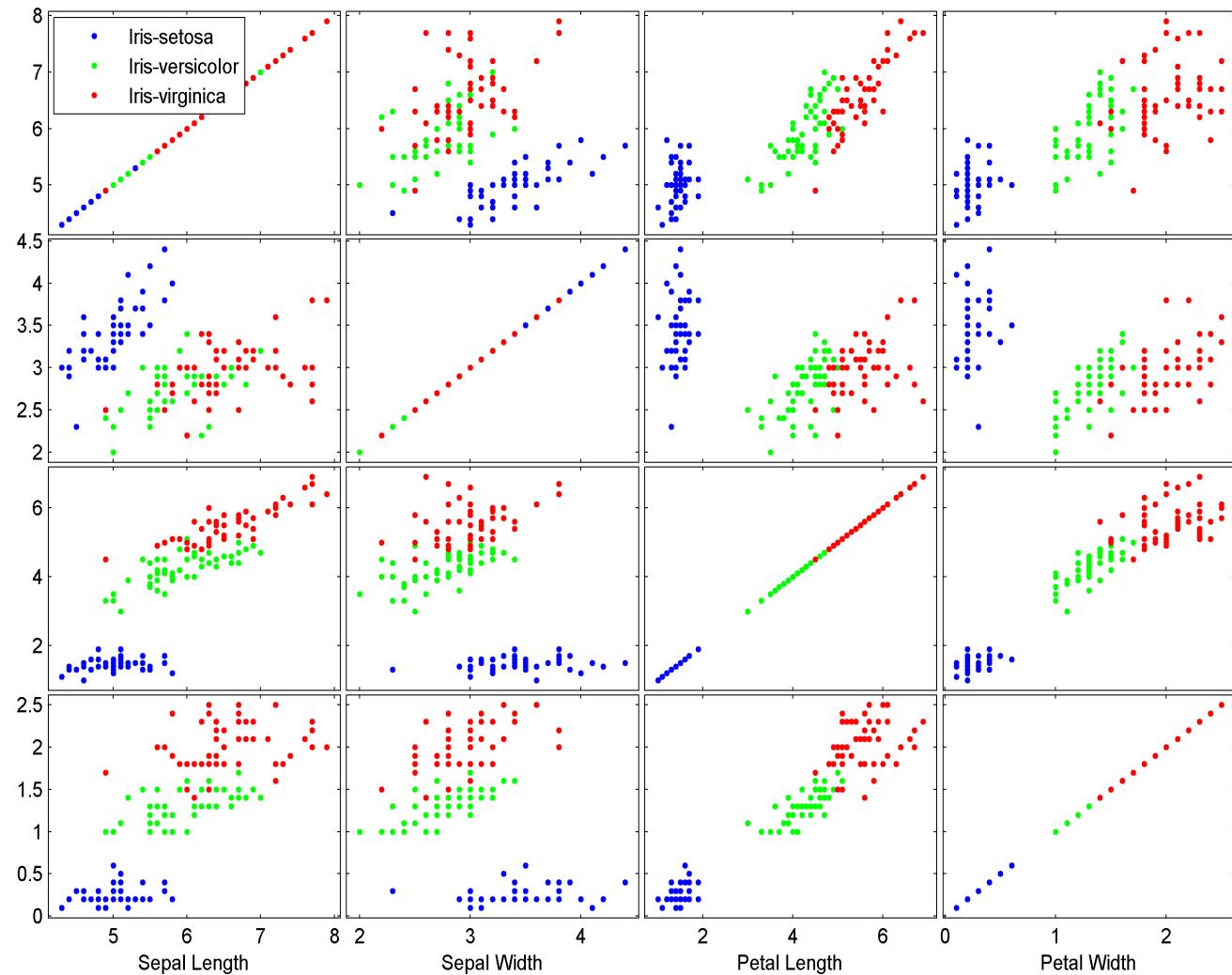
- Shows **relation** between attributes
 - Assess dependence between attributes
 - Used with classes to assess separability



```
hold all;
for c = 1:C
    scatter3(X(y==c-1,1),X(y==c-1,2),X(y==c-1,3),'filled');
end
grid on;
xlabel(attributeNames{1}, 'FontWeight', 'bold');
ylabel(attributeNames{2}, 'FontWeight', 'bold');
zlabel(attributeNames{3}, 'FontWeight', 'bold');
legend(classNames)
```

Scatter plots

- Scatter plot matrix
 - All pairs of attributes



```
gplotmatrix(X,X,classNames(y+1),[],[],[],'on',' ',attributeNames);
```

Visualization of high-dimensional objects

Matrix plots

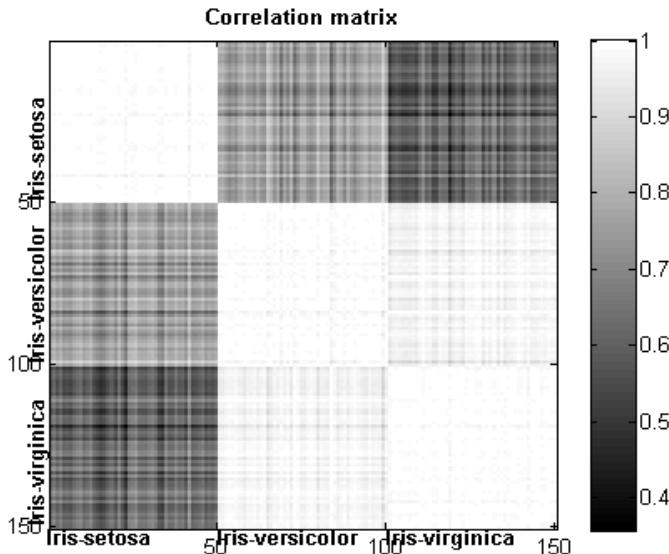
- **Plot of raw data matrix**

- Useful when objects are sorted according to class
- Typically, attributes are normalized

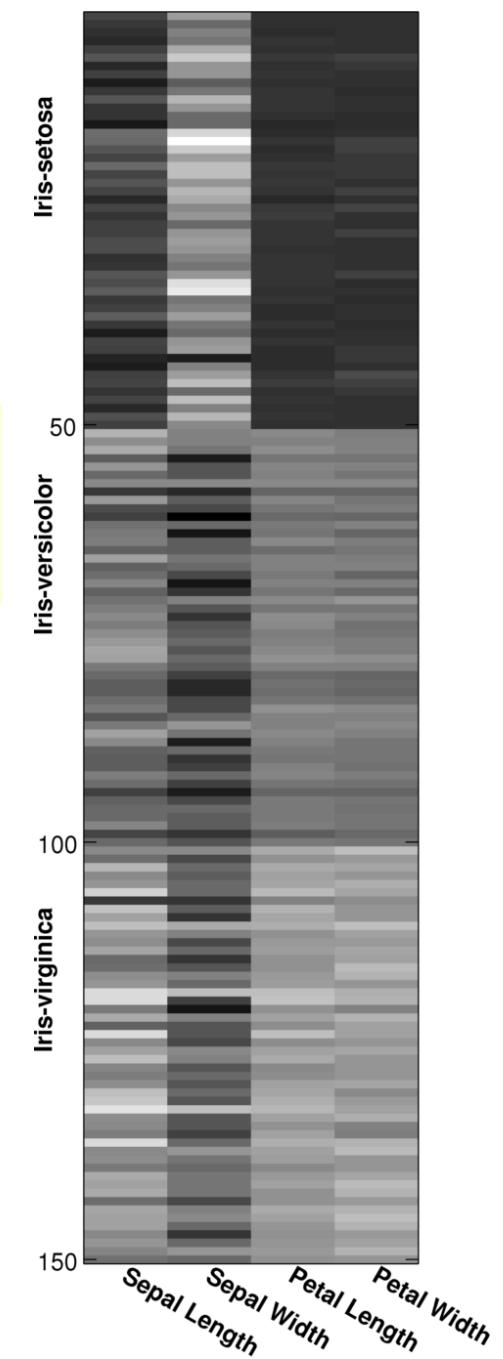
```
Z = zscore(X); % Standardize data
imagesc(Z);
colormap(gray)
title('Normalized data matrix');
```

- **Plots of similarity matrices**

- Useful for visualizing the relation between objects

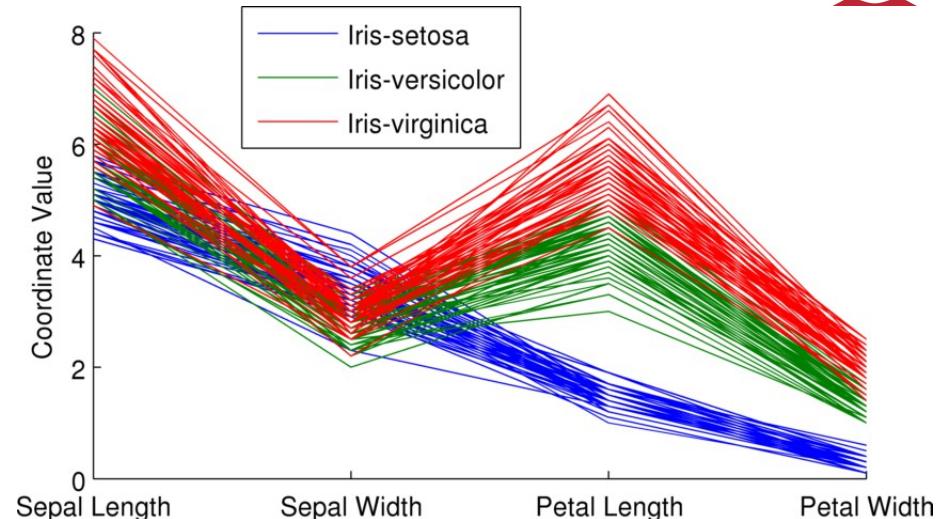


```
C = corrcoef(X');
h2 = imagesc(C);
colormap(gray);
axis equal;
axis tight;
colorbar;
title('Correlation matrix');
```



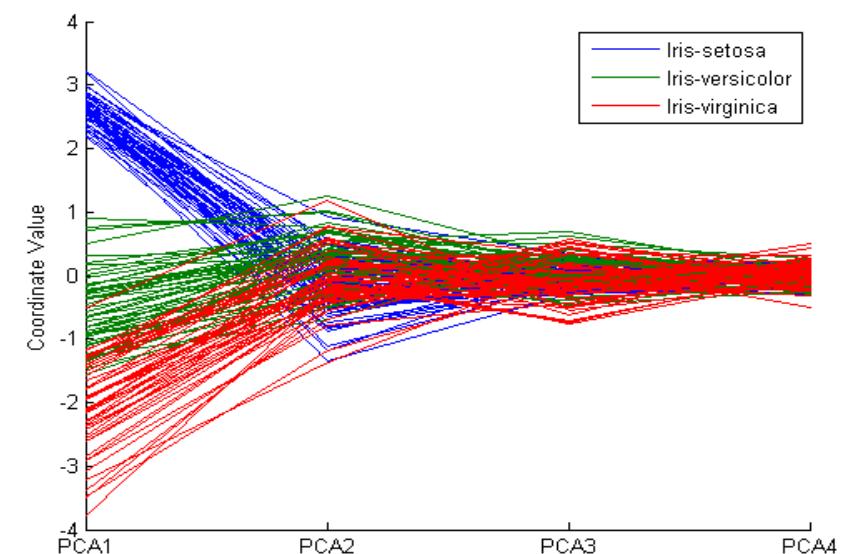
Parallel coordinates

- Plot high-dimensional data
- Instead of perpendicular axes
 - Use parallel axes
- Attribute values are plotted as a point
 - and the points are connected by a line
- Each object is represented as a line
- Lines representing a group of objects
 - Are similar in some sense
 - Ordering of attributes is important in seeing such groupings



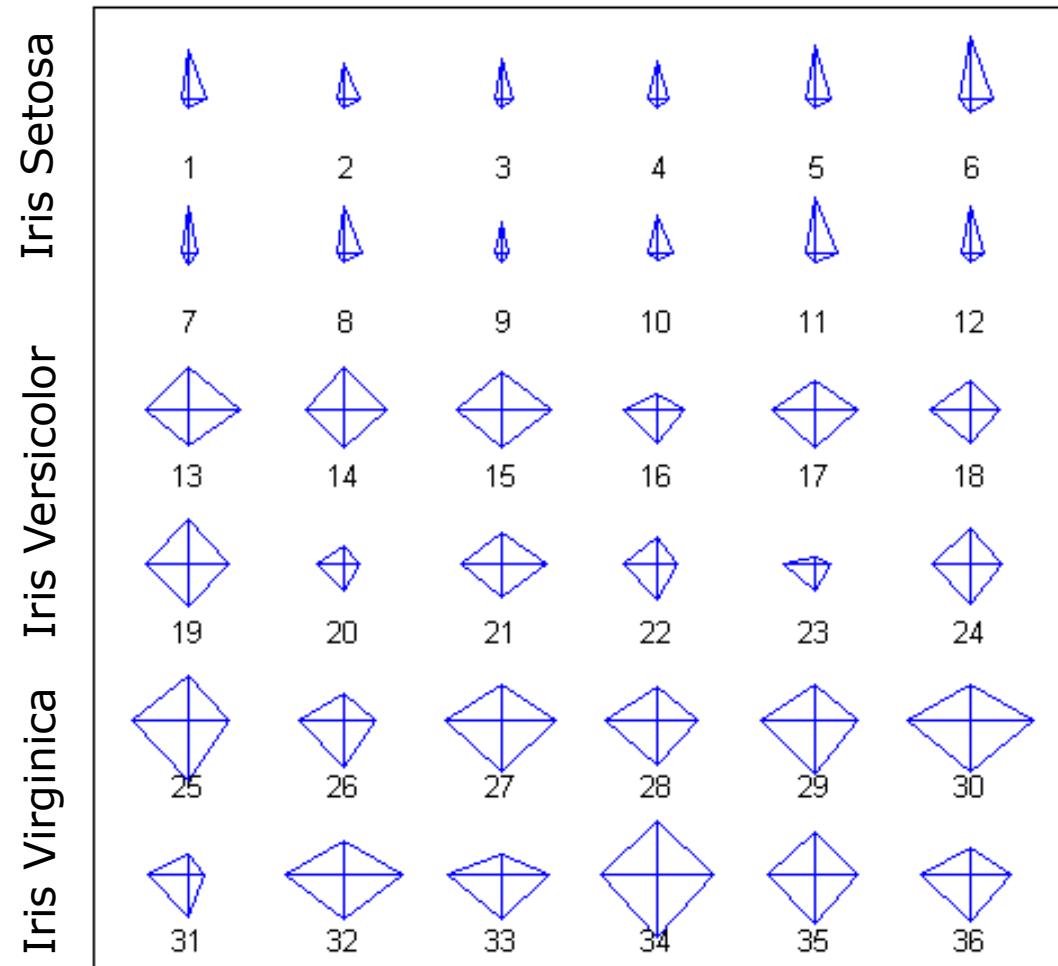
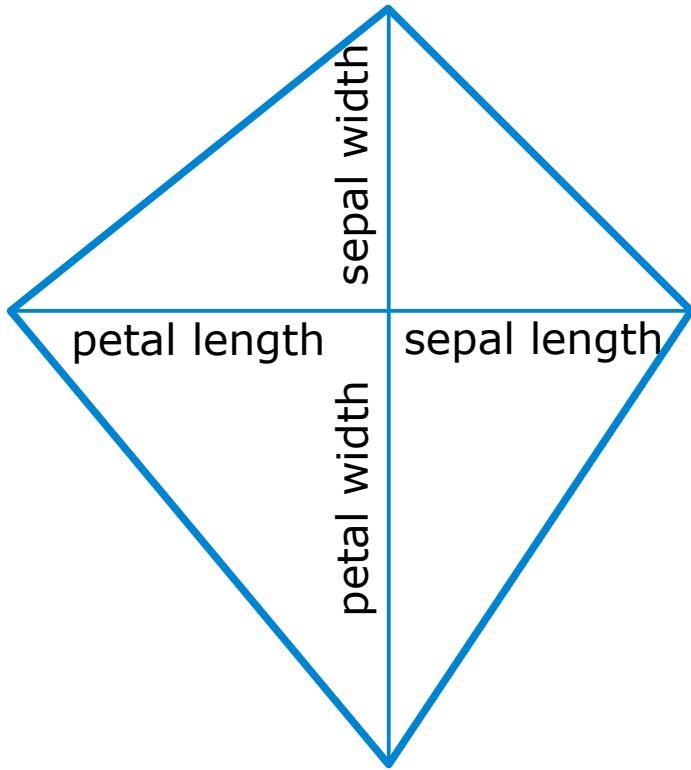
```
parallelcoords(X, 'group', classNames(y+1), ...
    'labels', attributeNames);
```

```
% Project data by PCA
Y = bsxfun(@minus, X, mean(X));
[U, S, V] = svd(Y);
Z = U*S;
% Plot data in space spanned by PCA
parallelcoords(Z, 'group', classNames(y+1), ...
    'labels', {'PCA1', 'PCA2', 'PCA3', 'PCA4'});
```



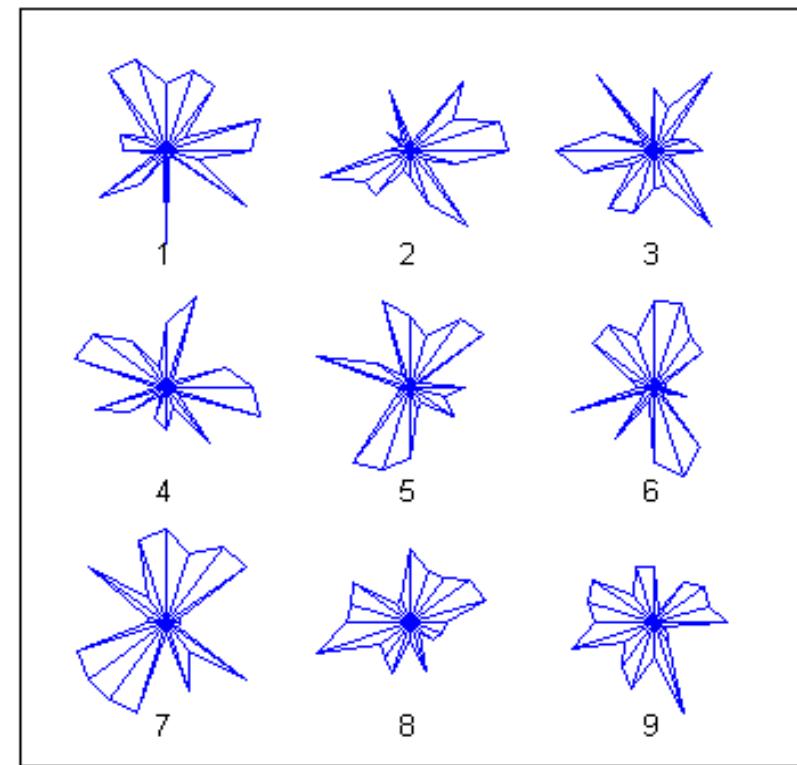
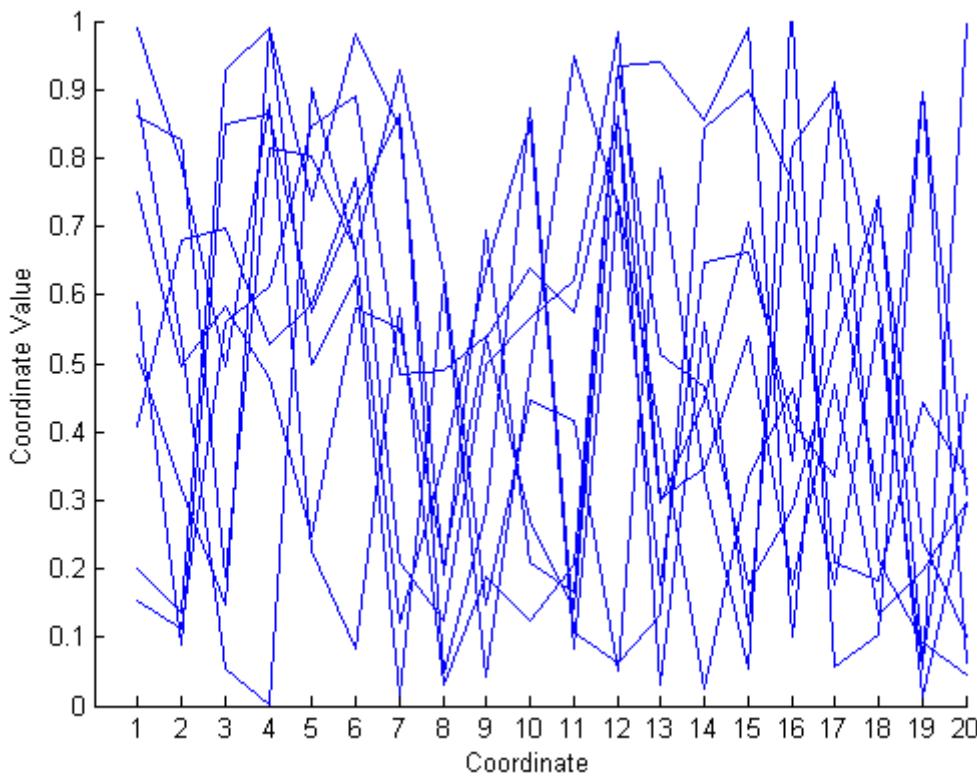
Star plots

- **Plot high-dimensional data**
- Similar to parallel coordinates
 - Axes radiate from center
 - Connecting line is a polygon



```
indGlyph=[1:12 50+(1:12) 100+(1:12)];  
glyphplot(X(indGlyph,:), 'glyph', 'star');
```

Example of Parallel Coordinates and star plots for 20 attributes



ACCENT

- **Apprehension**

- Is it easy to see what is important in the graph?

- **Clarity**

- Are the most important elements visually most prominent?

- **Consistency**

- Have you used the same colors, shapes, etc. as in other graphs?

- **Efficiency**

- Does it convey its information in the most simple and efficient way?

- **Necessity**

- Are all elements of the graph necessary to represent data?

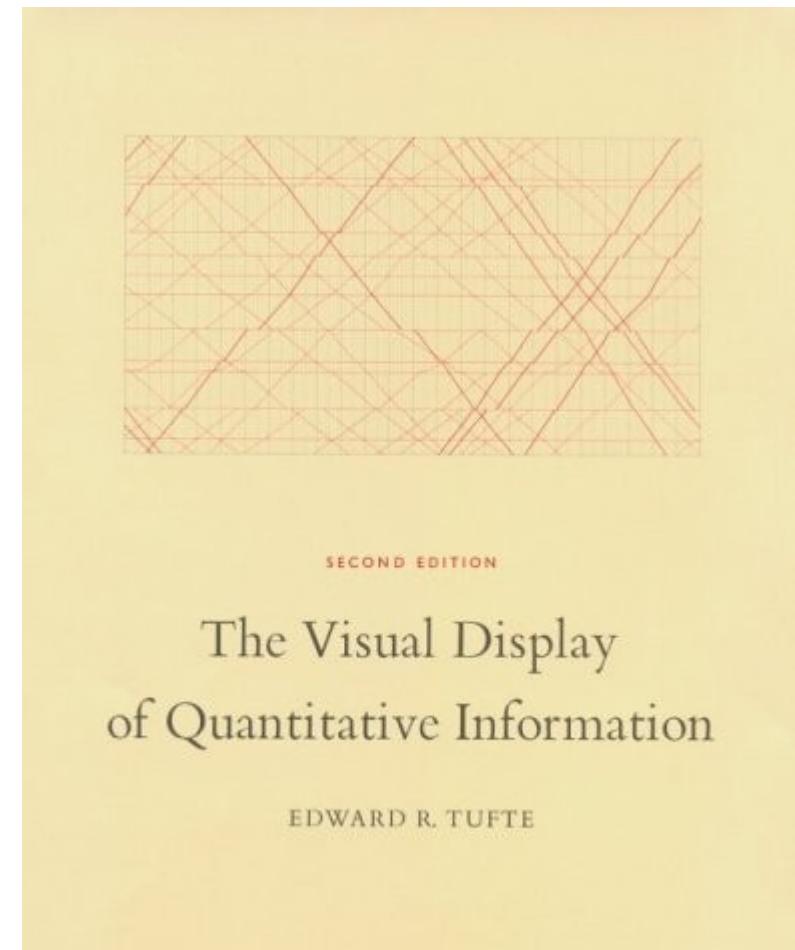
- **Truthfulness**

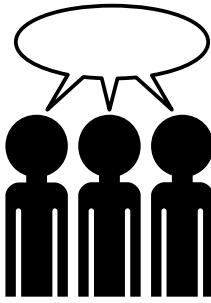
- Does the graph represent the data correctly?

Tufte's guidelines

- **Graphical excellence**

- Well-designed presentation of interesting data – a matter of
 - substance, statistics, and design
- Complex ideas communicated with
 - clarity, precision, and efficiency
- Gives the viewer
 - the greatest number of ideas
 - in the shortest time
 - with the least ink
 - in the smallest place.
- Nearly always multivariate
- Requires telling the truth about the data





Apprehension: Is it easy to see what is important in the graph?

Clarity: Are the most important elements visually most prominent?

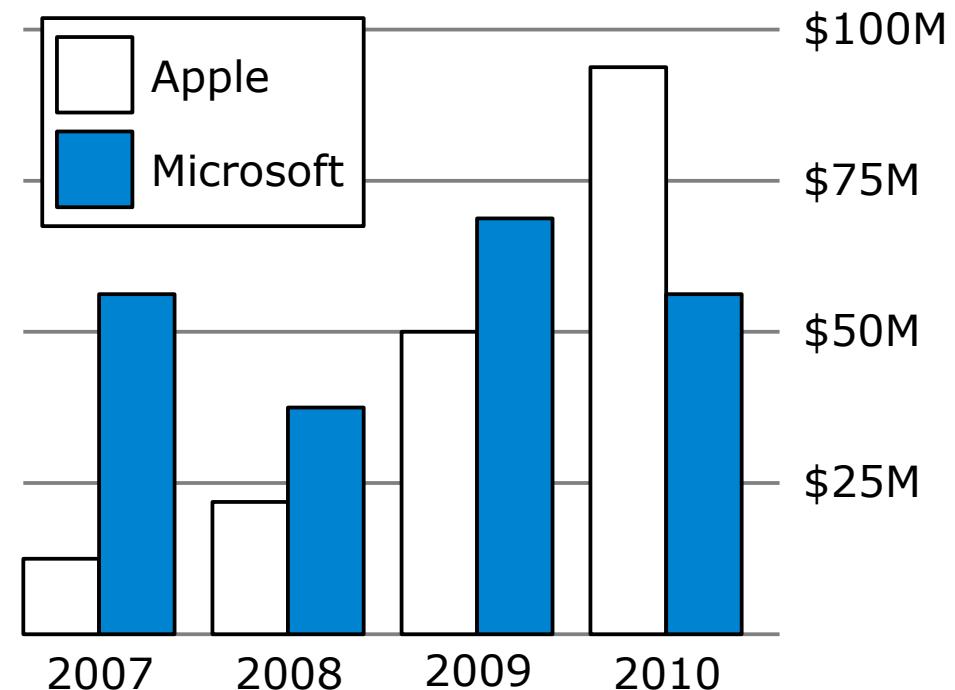
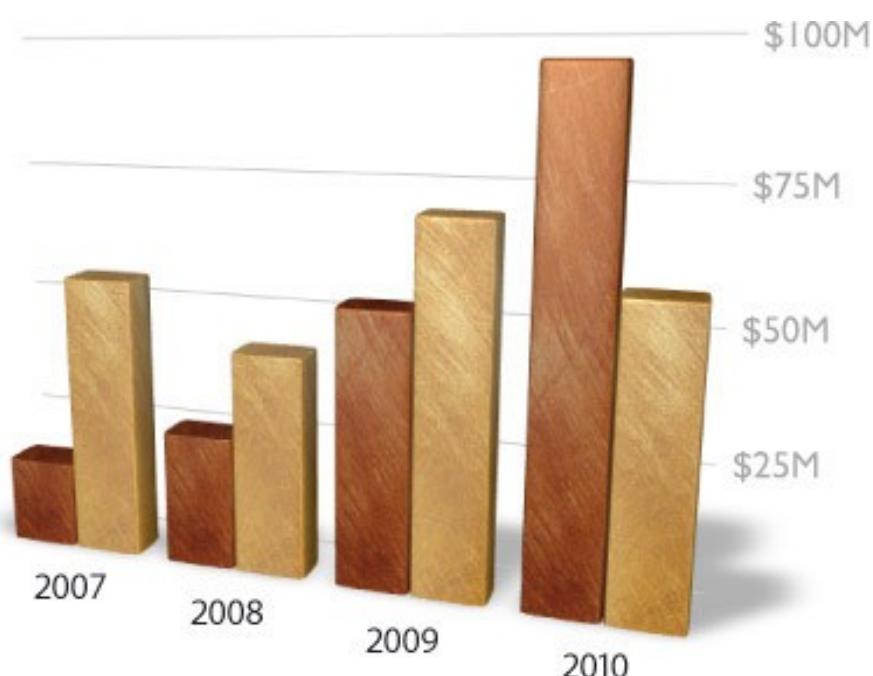
Consistency: Have you used the same colors, shapes, etc. as in other graphs?

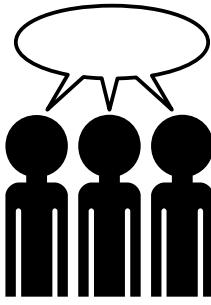
Efficiency: Does it convey its information in the most simple and efficient way?

Necessity: Are all elements of the graph necessary to represent data?

Truthfulness: Does the graph represent the data correctly?

- Compare the graphs using ACCENT





Apprehension: Is it easy to see what is important in the graph?

Clarity: Are the most important elements visually most prominent?

Consistency: Have you used the same colors, shapes, etc. as in other graphs?

Efficiency: Does it convey its information in the most simple and efficient way?

Necessity: Are all elements of the graph necessary to represent data?

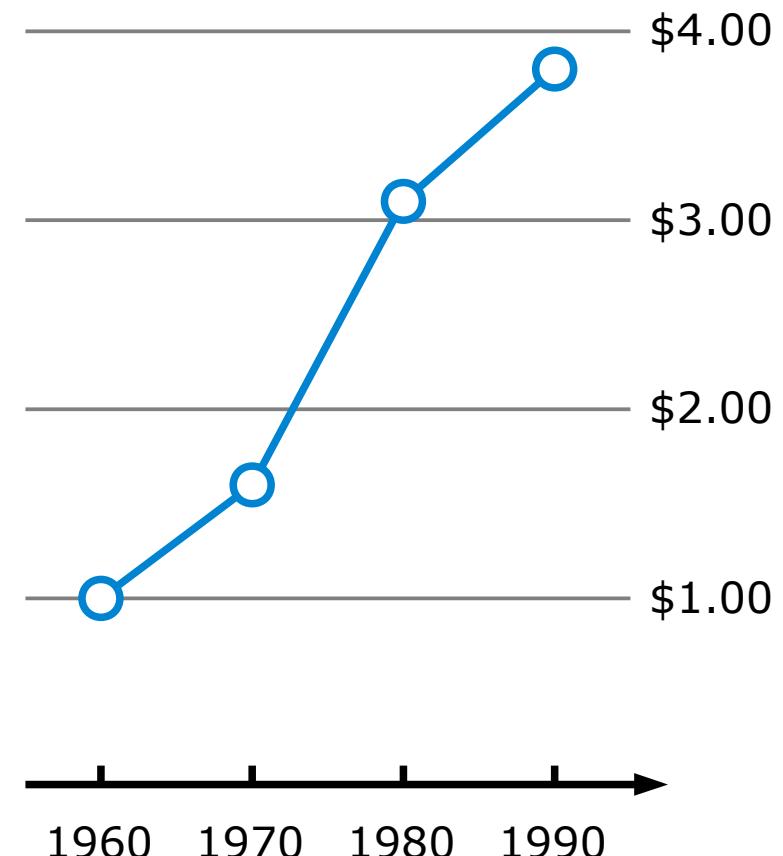
Truthfulness: Does the graph represent the data correctly?

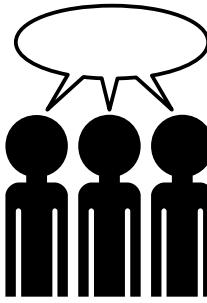
- Compare the graphs using ACCENT

Minimum wage

1960		\$1.00
1970		\$1.60
1980		\$3.10
1990		\$3.80

Minimum wage





Apprehension: Is it easy to see what is important in the graph?

Clarity: Are the most important elements visually most prominent?

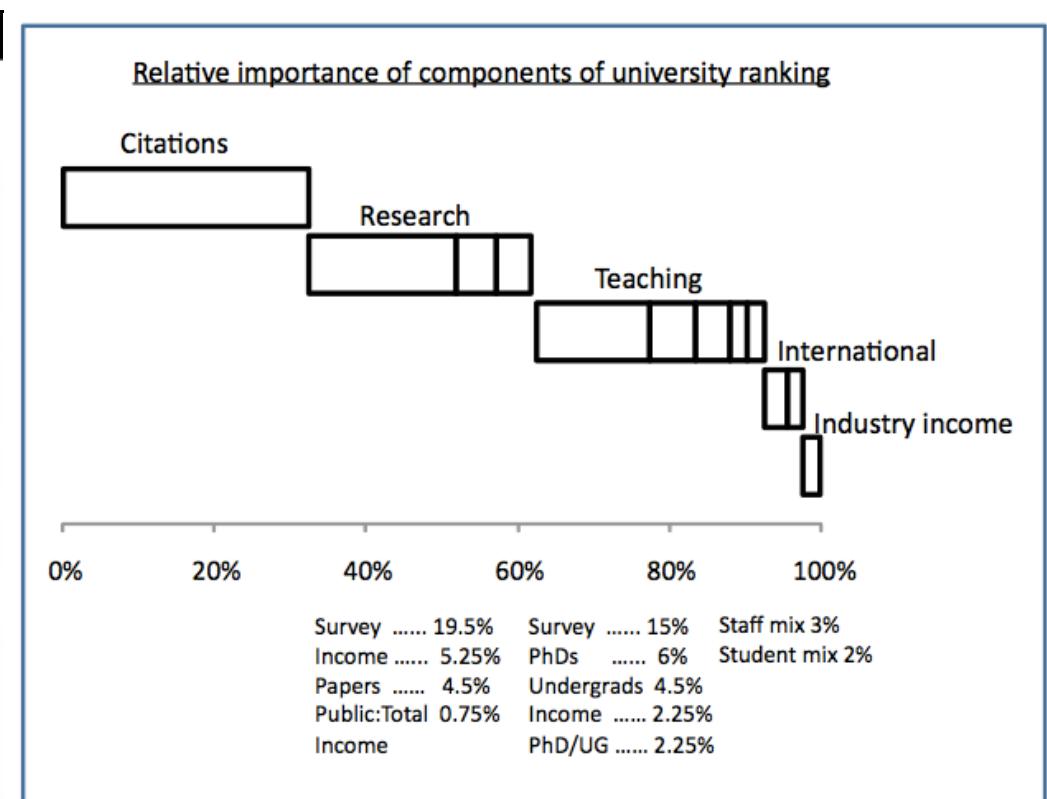
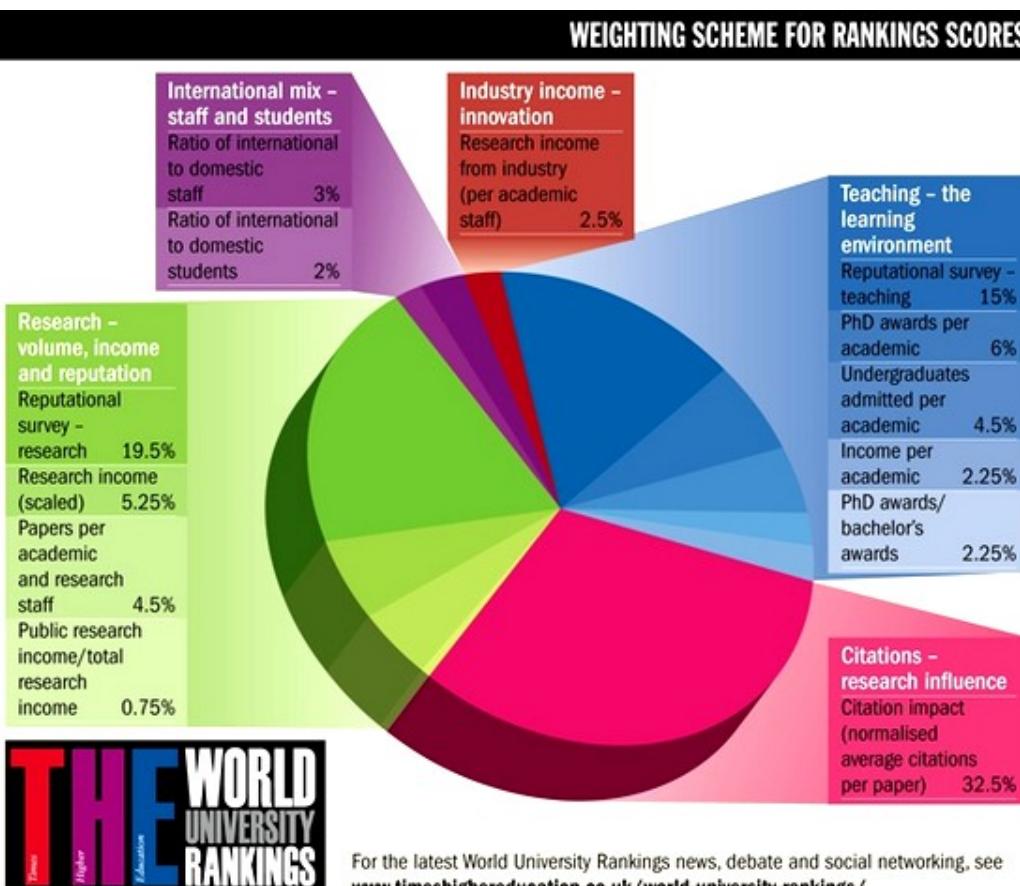
Consistency: Have you used the same colors, shapes, etc. as in other graphs?

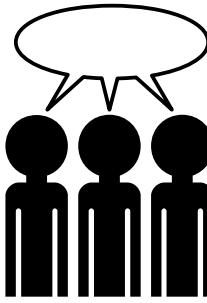
Efficiency: Does it convey its information in the most simple and efficient way?

Necessity: Are all elements of the graph necessary to represent data?

Truthfulness: Does the graph represent the data correctly?

- Compare the graphs using ACCENT





Apprehension: Is it easy to see what is important in the graph?

Clarity: Are the most important elements visually most prominent?

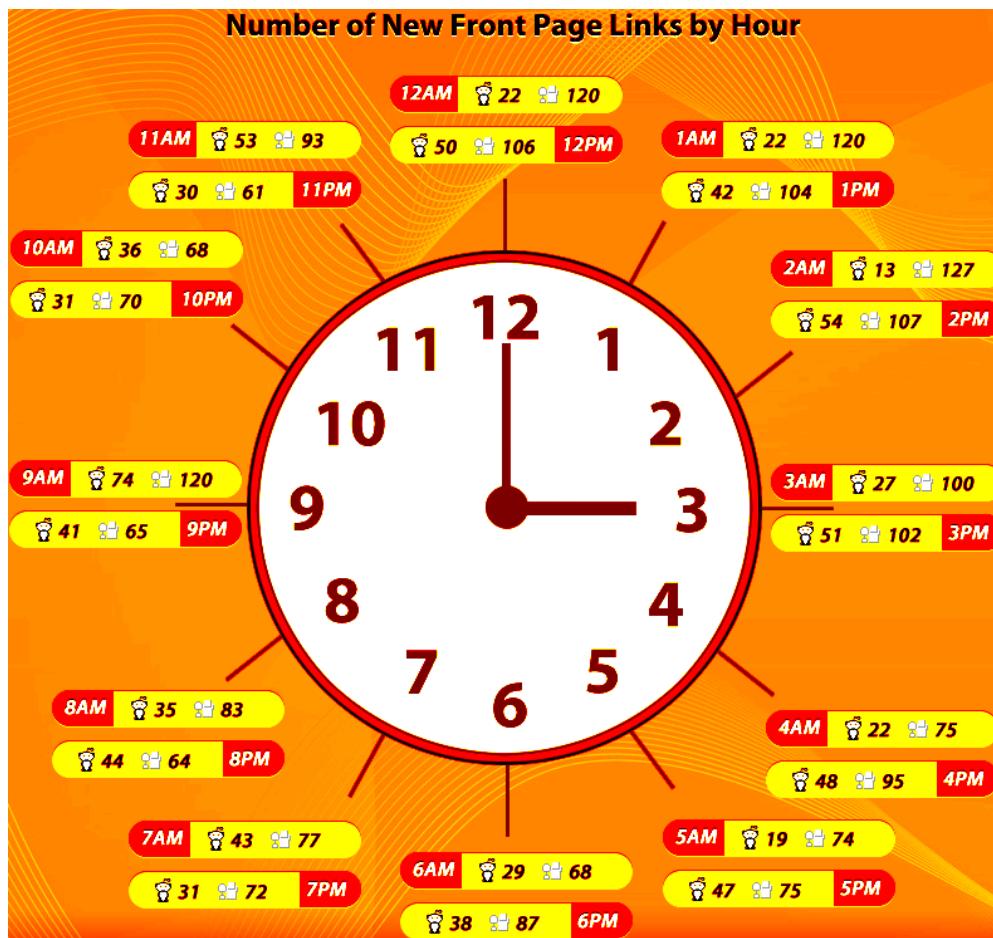
Consistency: Have you used the same colors, shapes, etc. as in other graphs?

Efficiency: Does it convey its information in the most simple and efficient way?

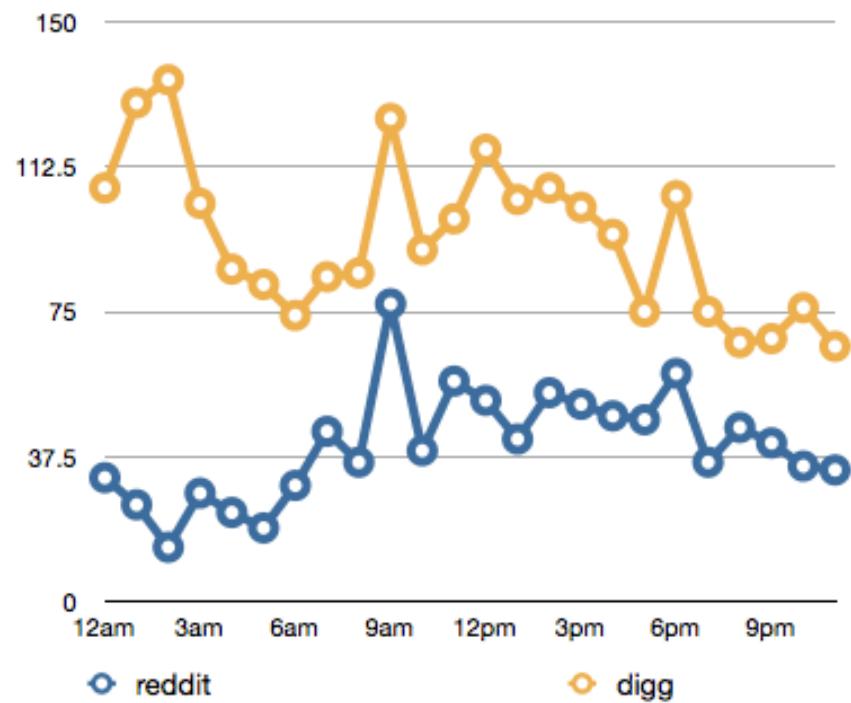
Necessity: Are all elements of the graph necessary to represent data?

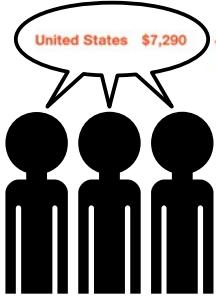
Truthfulness: Does the graph represent the data correctly?

- Compare the graphs using ACCENT



Number of New Front Page Links by Hour





Apprehension: Is it easy to see what is important in the graph?

Clarity: Are the most important elements visually most prominent?

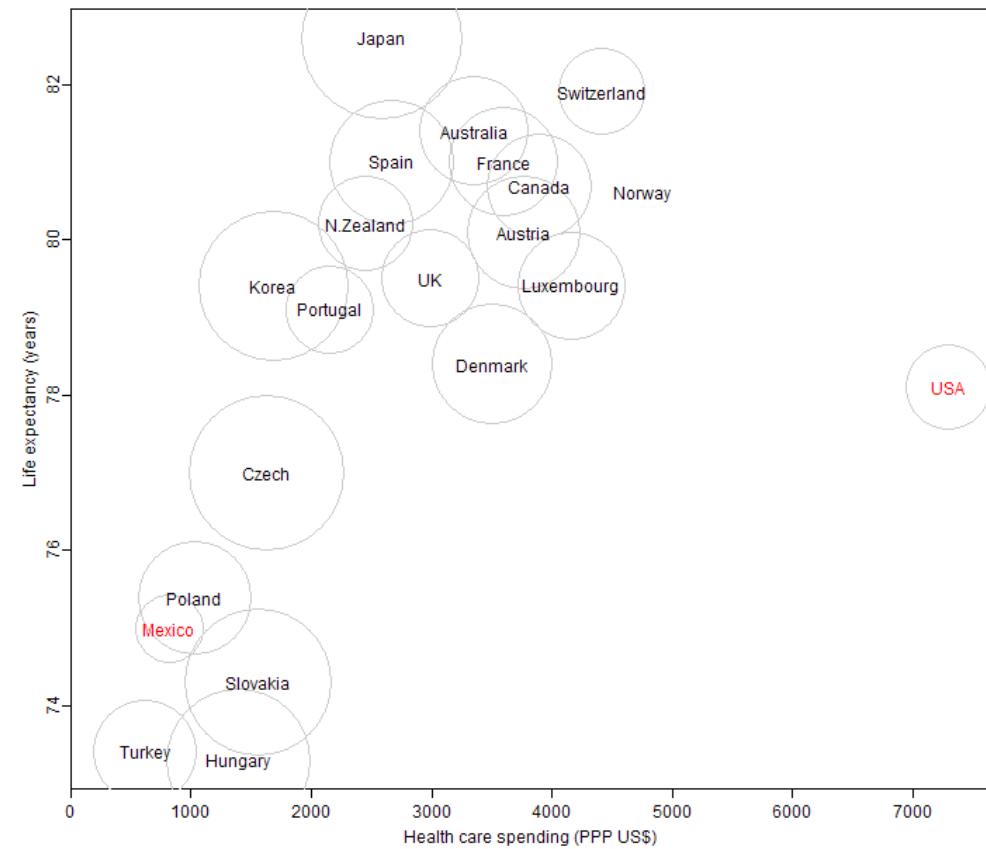
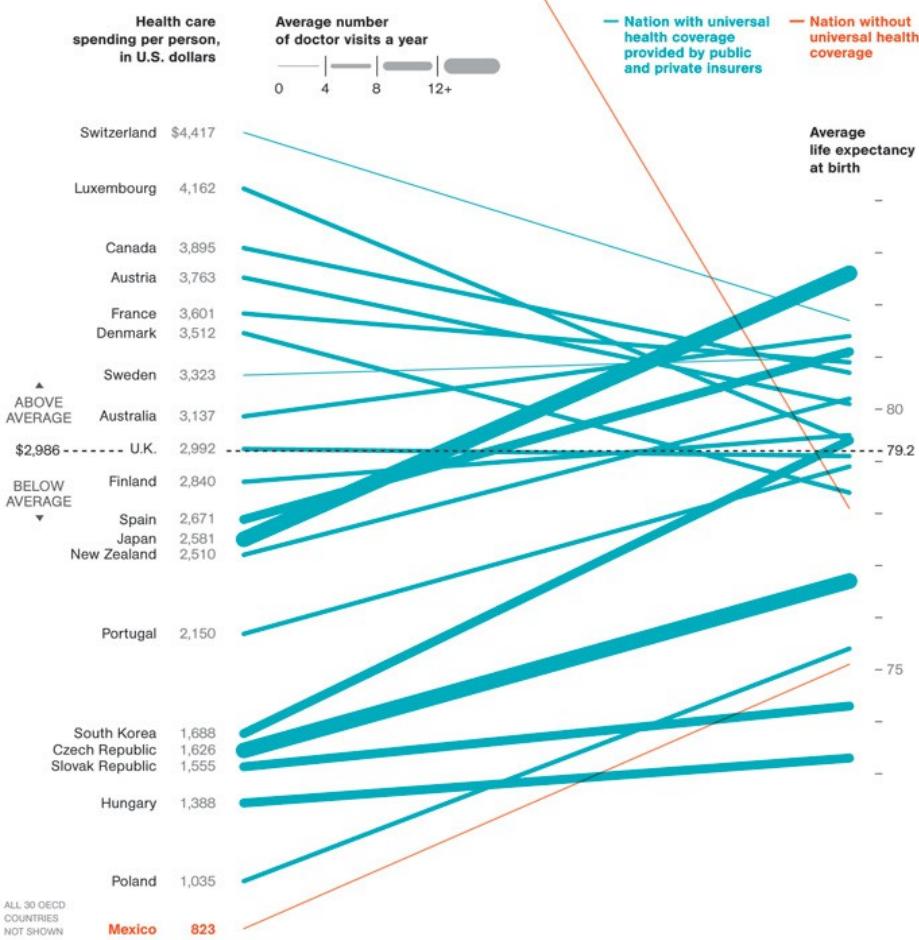
Consistency: Have you used the same colors, shapes, etc. as in other graphs?

Efficiency: Does it convey its information in the most simple and efficient way?

Necessity: Are all elements of the graph necessary to represent data?

Truthfulness: Does the graph represent the data correctly?

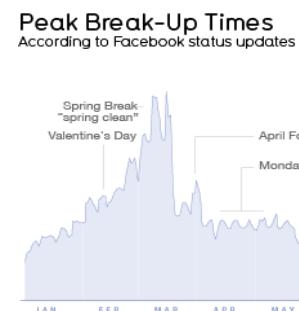
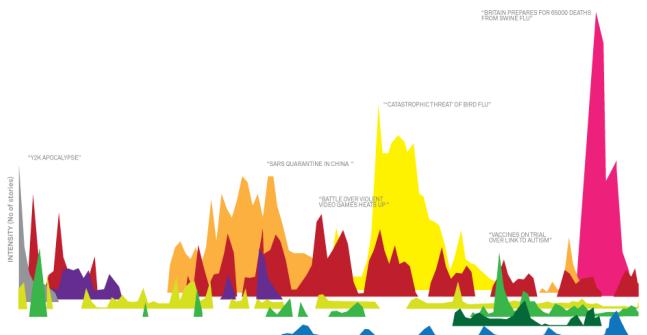
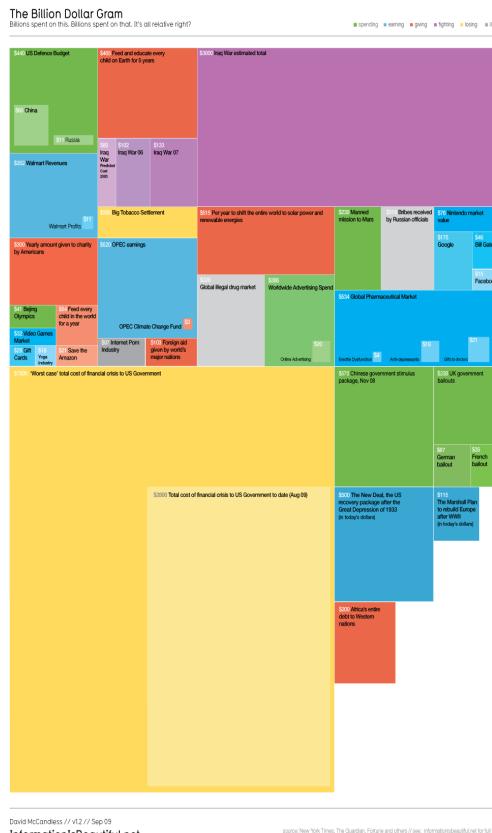
- Compare these graphs using ACCENT



Making good data visualizations is an art

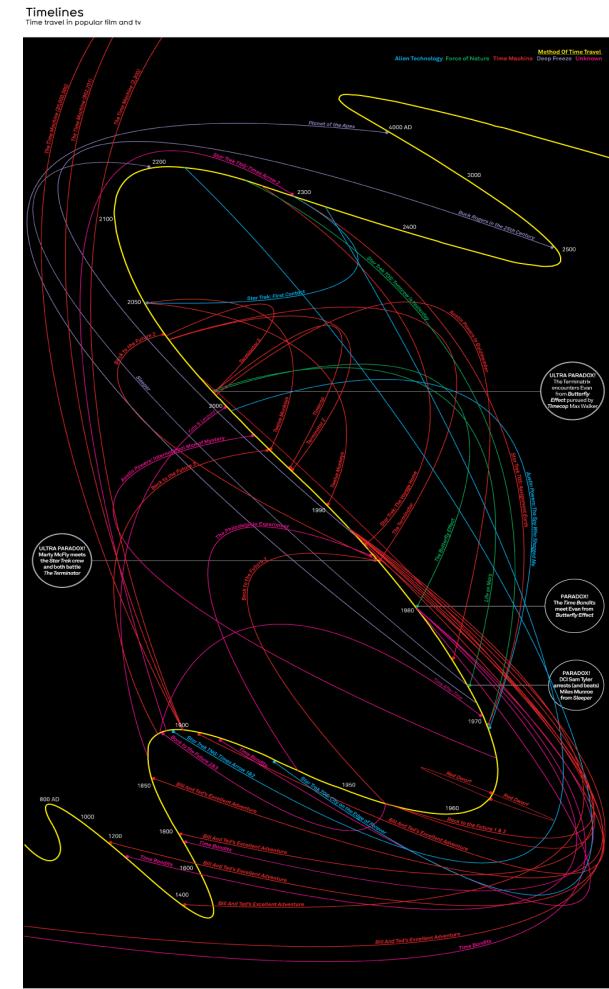
For some interesting data visualizations see also

http://www.ted.com/talks/david_mccandless_the_beauty_of_data_visualization.html
<http://www.informationisbeautiful.net/>



David McCandless & Lee Byron
 InformationIsBeautiful.net / LeeByron.com

source: searches for "we broke up because"
 taken from the infographic ultrabook
 The Visual Miscellaneum





Imagine you have a dataset where some of the attributes are numeric given in the matrix X but you also have a categorical attribute given by TXT (see below). You would like to carry out a PCA on the data taking both the numeric and categoric attributes into account. How would you proceed?

Age Height Weight
(Standardized)

```
X= [-0.2248 -0.4762 -0.2097
-0.5890 0.8620 0.6252
-0.2938 -1.3617 0.1832
-0.8479 0.4550 -1.0298
-1.1201 -0.8487 0.9492
2.5260 -0.3349 0.3071
1.6555 0.5528 0.1352
0.3075 1.0391 0.5152
-1.2571 -1.1176 0.2614
-0.8655 1.2607 -0.9415
-0.1765 0.6601 -0.1623
0.7914 -0.0679 -0.1461
-1.3320 -0.1952 -0.5320
-2.3299 -0.2176 1.6821
-1.4491 -0.3031 -0.8757
0.3335 0.0230 -0.4838
0.3914 0.0513 -0.7120
0.4517 0.8261 -1.1742
-0.1303 1.5270 -0.1922
0.1837 0.4669 -0.2741]
```

Nationality

```
TXT= ['Sweden'
'Sweden'
'Sweden'
'Sweden'
'Norway'
'Norway'
'Norway'
'Norway'
'Sweden'
'Norway'
'Denmark'
'Denmark'
'Sweden'
'Sweden'
'Sweden'
'Sweden'
'Denmark'
'Norway'
'Denmark']
```

	Denmark	Norway	Sweden
'Sweden'	0	0	1
'Sweden'	0	0	1
'Sweden'	0	0	1
'Sweden'	0	0	1
'Norway'	0	1	0
'Norway'	0	1	0
'Norway'	0	1	0
'Norway'	0	1	0
'Sweden'	0	0	1
'Norway'	0	1	0
'Denmark'	1	0	0
'Denmark'	1	0	0
'Sweden'	0	0	1
'Sweden'	0	0	1
'Sweden'	0	0	1
'Sweden'	0	0	1
'Denmark'	1	0	0
'Norway'	0	1	0
'Denmark'	1	0	0

X_tmp=

	Age	Height	Weight	Denmark	Norway	Sweden
-0.2248	-0.4762	-0.2097	0	0	1.0000	
-0.5890	0.8620	0.6252	0	0	1.0000	
-0.2938	-1.3617	0.1832	0	0	1.0000	
-0.8479	0.4550	-1.0298	0	0	1.0000	
-1.1201	-0.8487	0.9492	0	1.0000	0	
2.5260	-0.3349	0.3071	0	1.0000	0	
1.6555	0.5528	0.1352	0	1.0000	0	
0.3075	1.0391	0.5152	0	1.0000	0	
-1.2571	-1.1176	0.2614	0	1.0000	0	
-0.8655	1.2607	-0.9415	0	0	1.0000	
-0.1765	0.6601	-0.1623	0	1.0000	0	
0.7914	-0.0679	-0.1461	1.0000	0	0	
-1.3320	-0.1952	-0.5320	1.0000	0	0	
-2.3299	-0.2176	1.6821	0	0	1.0000	
-1.4491	-0.3031	-0.8757	0	0	1.0000	
0.3335	0.0230	-0.4838	0	0	1.0000	
0.3914	0.0513	-0.7120	1.0000	0	0	
0.4517	0.8261	-1.1742	0	0	1.0000	
-0.1303	1.5270	-0.1922	0	1.0000	0	
0.1837	0.4669	-0.2741	1.0000	0	0	

```
[X_tmp, attributeNames_tmp]=categoric2numeric(TXT);
X=[X X_tmp];
attributeNames=[attributeNames; attributeNames_tmp];
```

	Age	Height	Weight	Denmark	Norway	Sweden
-0.2248	-0.4762	-0.2097	0	0	1.0000	
-0.5890	0.8620	0.6252	0	0	1.0000	
-0.2938	-1.3617	0.1832	0	0	1.0000	
-0.8479	0.4550	-1.0298	0	0	1.0000	
-1.1201	-0.8487	0.9492	0	1.0000	0	
2.5260	-0.3349	0.3071	0	1.0000	0	
1.6555	0.5528	0.1352	0	1.0000	0	
0.3075	1.0391	0.5152	0	1.0000	0	
-1.2571	-1.1176	0.2614	0	1.0000	0	
-0.8655	1.2607	-0.9415	0	0	1.0000	
-0.1765	0.6601	-0.1623	0	1.0000	0	
0.7914	-0.0679	-0.1461	1.0000	0	0	
-1.3320	-0.1952	-0.5320	1.0000	0	0	
-2.3299	-0.2176	1.6821	0	0	1.0000	
-1.4491	-0.3031	-0.8757	0	0	1.0000	
0.3335	0.0230	-0.4838	0	0	1.0000	
0.3914	0.0513	-0.7120	1.0000	0	0	
0.4517	0.8261	-1.1742	0	0	1.0000	
-0.1303	1.5270	-0.1922	0	1.0000	0	
0.1837	0.4669	-0.2741	1.0000	0	0	

One-out-of-K coding

02450 Introduction to machine learning and data modeling

A collage of mathematical symbols including integrals, summation, infinity, and various Greek letters like theta, epsilon, and chi.

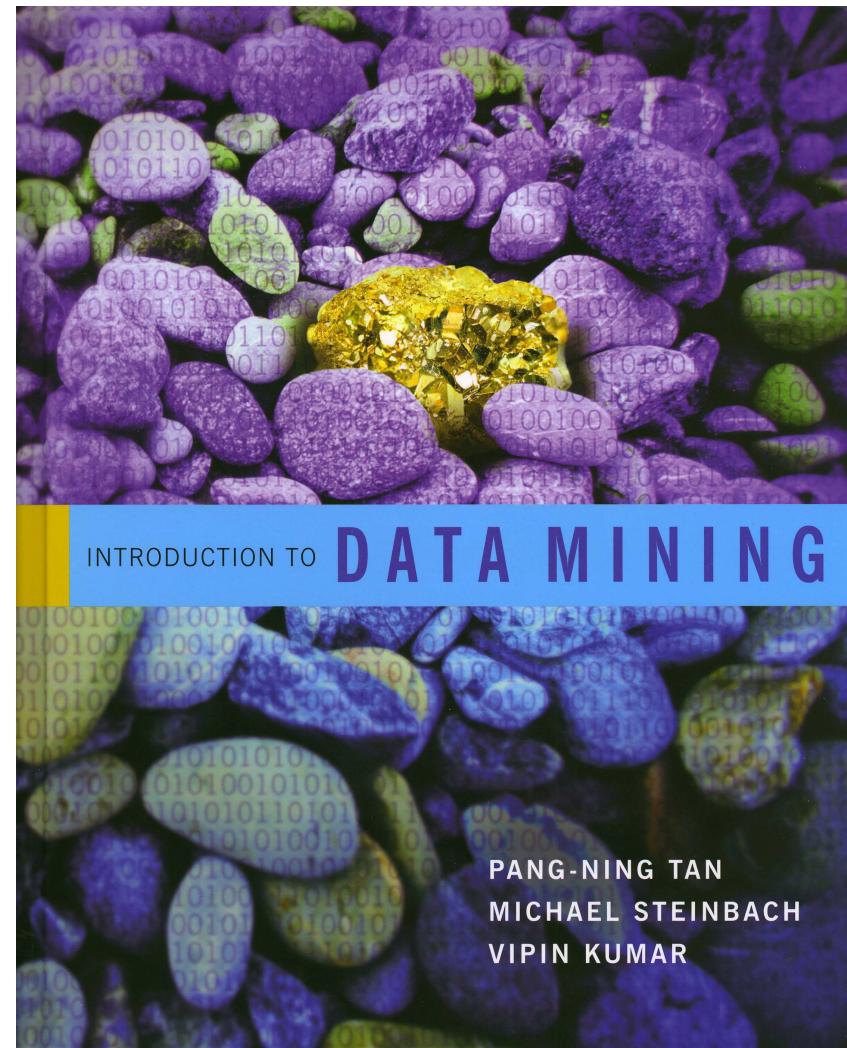
Reading material

Tan, Steinbach and Kumar
"Introduction to Data Mining"

Section 4.1-4.3 + Appendix D

Groups of the day

Esben Folger Thomas
Anne Frahm
Andreas Madsen
Frederik Wolgast Rørbech
Tomasz Maciazek
Florin Maticu
Steffen Angstmann



Lecture schedule

1. Introduction
(Tan 1.1-1.4)

Data: Feature extraction and visualization

2. Data and feature extraction
(Tan 2.1-2.2 +(A) + B.1)

3. Measures of similarity and summary statistics
(Tan 2.4 + 3.1-3.2 + C1-C2)

4. Data visualization
(Tan 3.3)

Supervised learning: Classification and regression

5. **Decision trees and linear regression**
(Tan 4.1-4.3 + D)

6. Overfitting and performance evaluation
(Tan 4.4-4.6)

7. Nearest neighbor, naive Bayes, and artificial neural networks
(Tan 5.2-5.4)

8. Ensemble methods and multi class classifiers
(Tan 5.6-5.8)

Unsupervised learning: Clustering and density est.

9. K-means and hierarchical clustering
(Tan 8.1-8.3 + 8.5.7)

10. Mixture models and association mining
(Tan 9.2.2 + 6.1-6.3)

11. Density estimation and anomaly detection
(Tan 10.1-10.4)

Machine learning and data modelling in practice

12. Putting it all together: Summary and overview

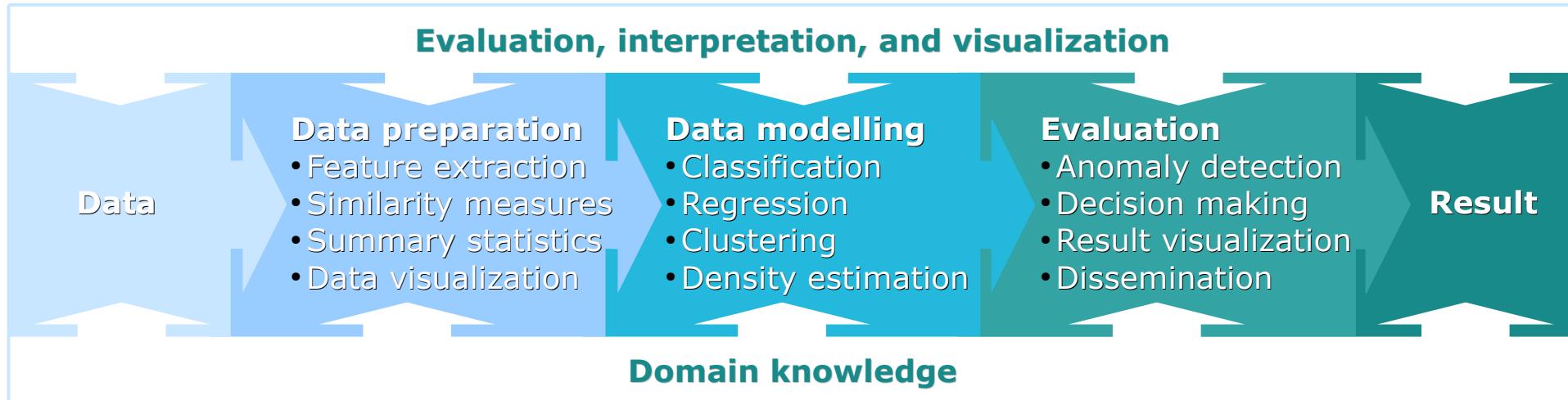
13. Mini project presentation

Report 1

Report 2

Report 3

Data modeling framework



After today you should be able to:

Explain what supervised learning is

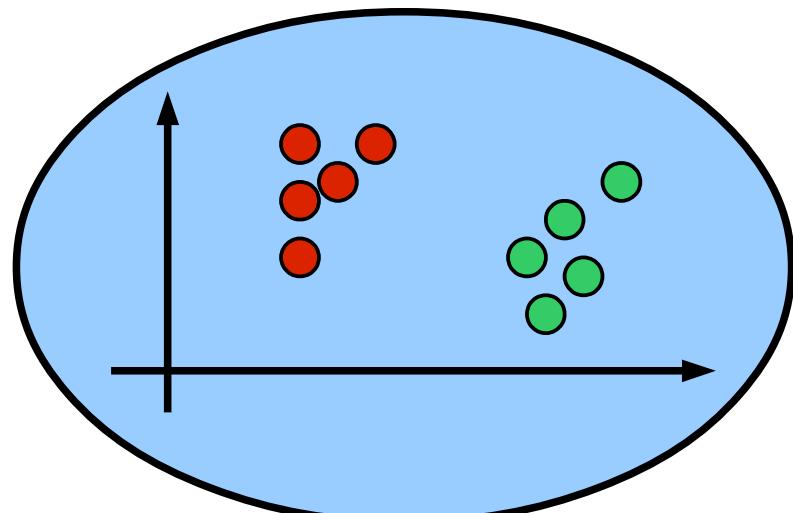
Explain the difference between classification and regression

Be able to evaluate classifiers in terms of the confusion matrix, error rate and accuracy

Understand the principles behind decision trees and Hunt's algorithm

Apply and interpret decision trees, linear regression and logistic regression

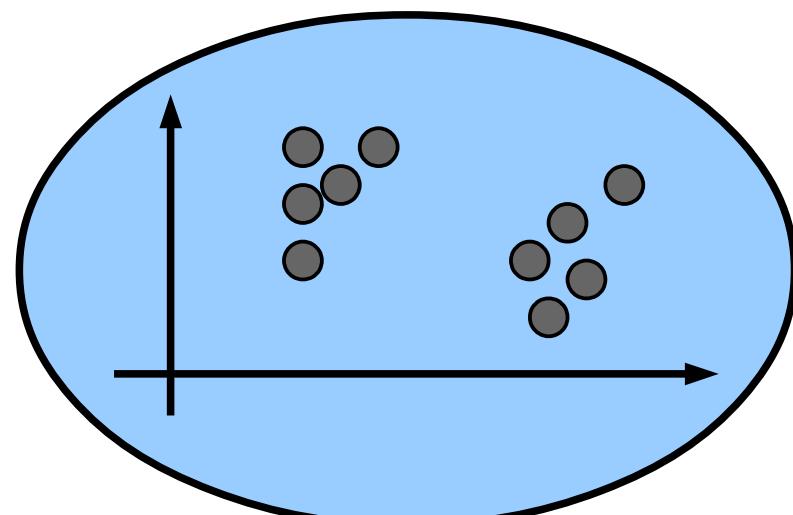
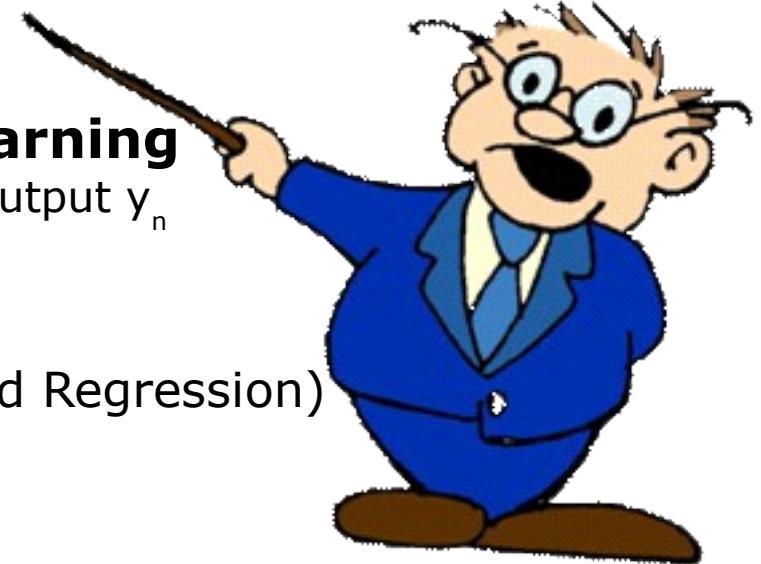
Supervised and Unsupervised learning



Supervised Learning

Input data \mathbf{x}_n and output y_n

(Classification and Regression)



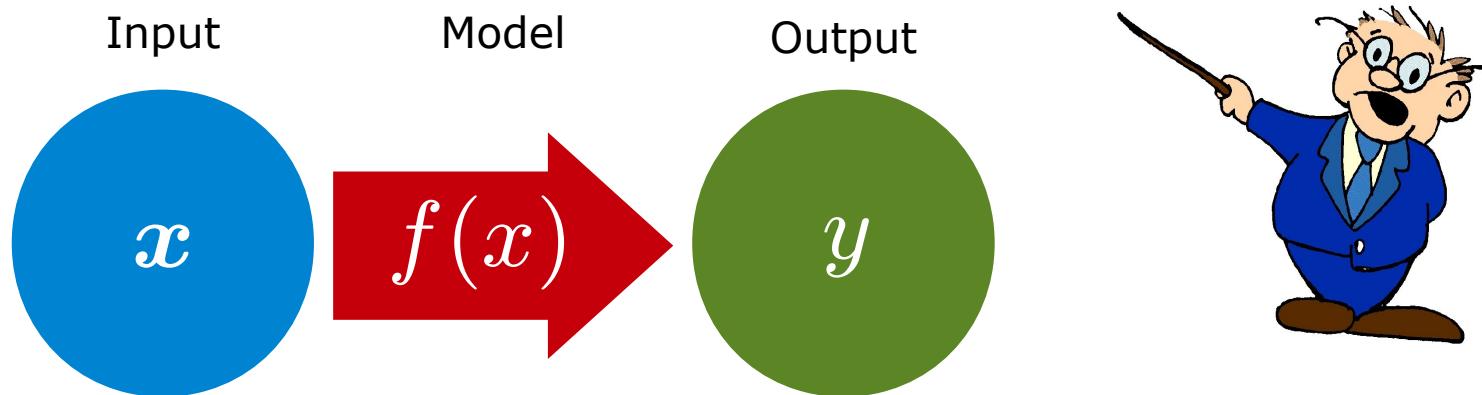
Unsupervised Learning

Input data \mathbf{x}_n alone

(Exploratory analysis)



Supervised learning



- **Mapping between domains**
 - Classification: Discrete (nominal) output
 - Regression: Continuous output

Supervised learning

- **Data**

- Inputs and outputs

$$\{\mathbf{x}_n, y_n\}_{n=1}^N$$

- **Model**

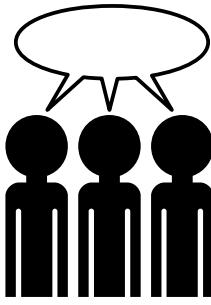
- Function that maps inputs to outputs

$$f(\mathbf{x})$$

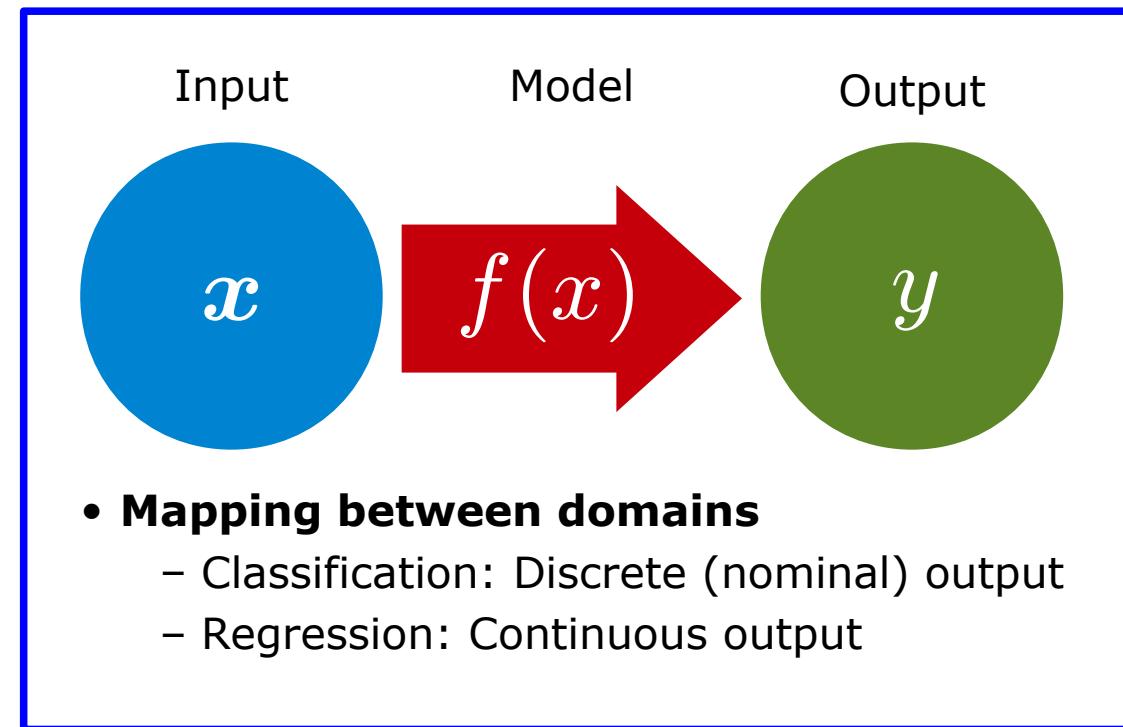
- **Cost function**

- Dissimilarity measure between data and model

$$d(y, f(\mathbf{x}))$$



Give an example of a classification and a regression problem and explain what the model $f(x)$ can be used for.



Classification

- **Definition:** Learning a function that maps a data object to a discrete class
- **Why classify?**
 - Descriptive modeling
 - Explain / understand the relation between attributes and class
 - Predictive modeling
 - Predict the class of a new data object

Confusion matrix

- Visualization of actual versus predicted class labels

- **Accuracy**

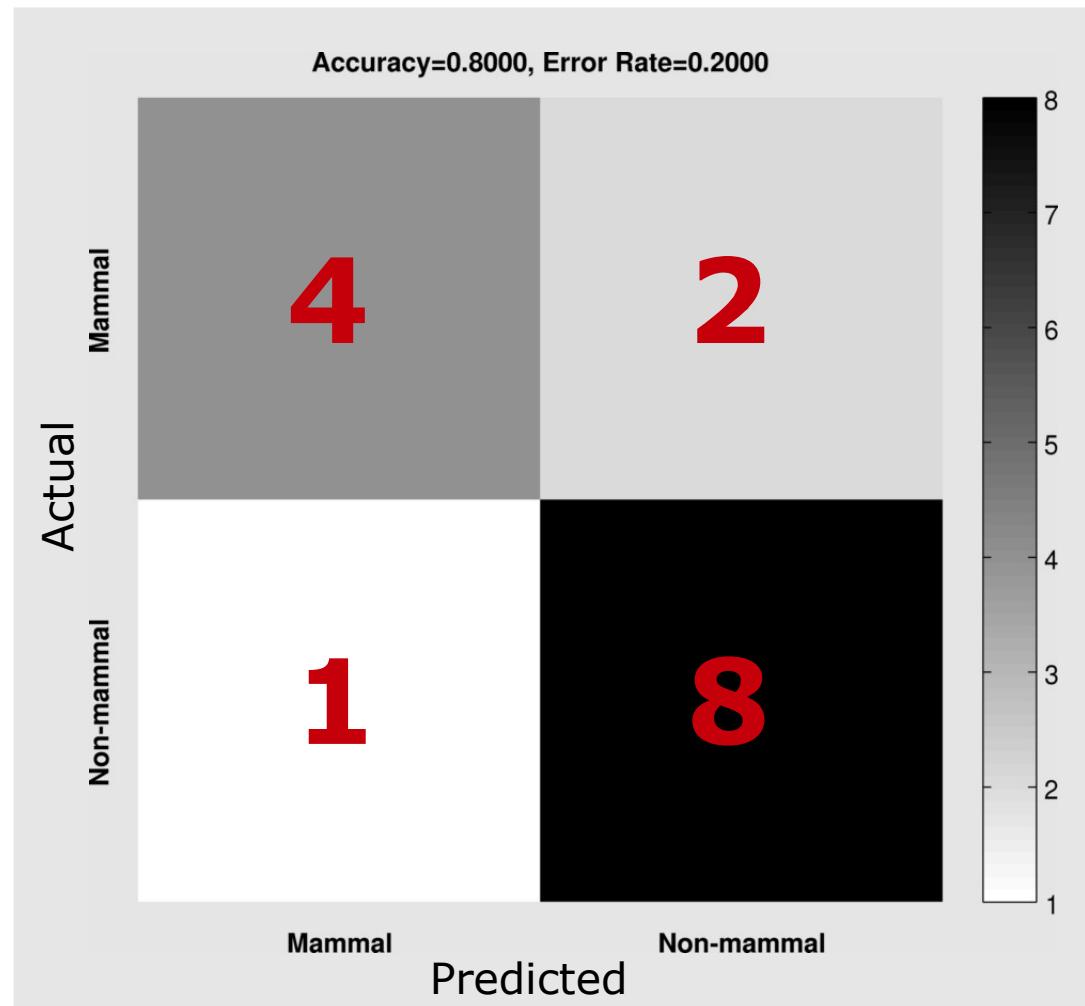
(Number of correctly predicted observations divided by the total number of observations)

$$\frac{4 + 8}{4 + 2 + 1 + 8} = 80\%$$

- **Error rate**

(Number of in-correctly predicted observations divided by the total number of observations)

$$\frac{2 + 1}{4 + 2 + 1 + 8} = 20\%$$



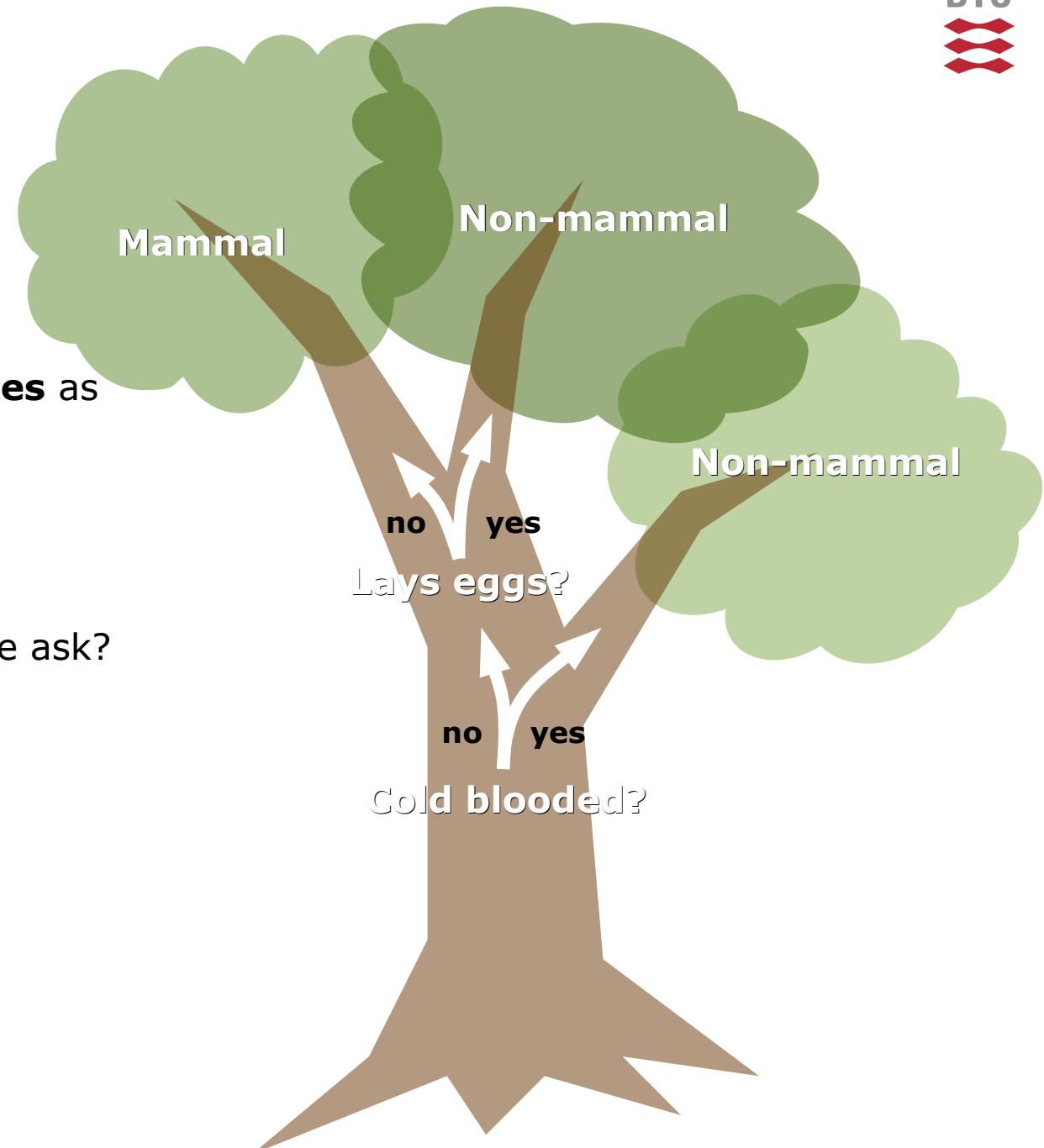
Decision trees

- Remember the game “20 questions to the professor”? (see also www.20q.net)

- Q1. Is it an Animal? Yes.
- Q2. Can you hold it? No.
- Q3. Does it live in groups (gregarious)? Yes.
- Q4. Are there many different sorts of it? No.
- Q5. Can it jump? Yes.
- Q6. Does it eat seeds? No.
- Q7. Is it white? Sometimes.
- Q8. Is it black and white? No.
- Q9. Does it have paws? Yes.
- Q10. Can you see it in a zoo? Yes.
- Q11. Does it roar? Yes.
- Q12. Is it worth a lot of money? Yes.
- Q13. Does it have spots? Yes.
- Q14. Is it multicoloured? Yes.
- Q15. Can you make money by selling it? Yes.
- Q16. Does it live in the jungle? Yes.
- Q17. I guessed that it was a leopard? Wrong.
- Q18. Does it like to play? Yes.
- Q19. I guessed that it was a cheetah? Wrong.
- Q20. I am guessing that it is a siberian tiger? Correct.

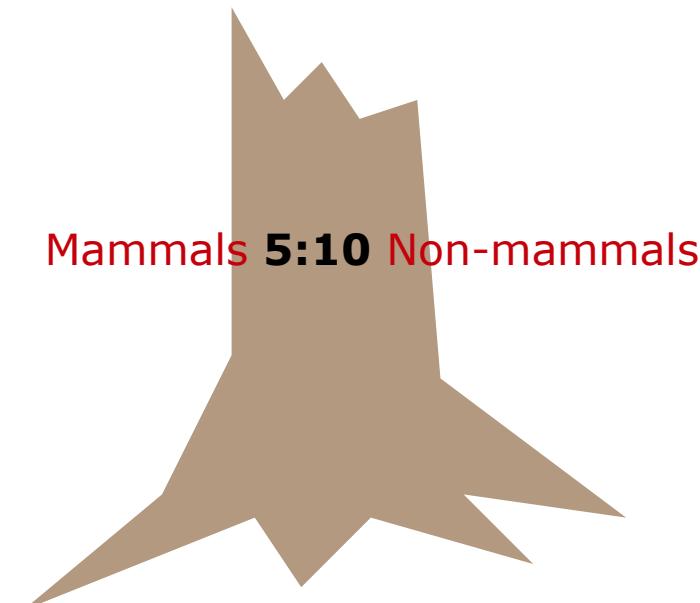
Decision trees

- Ask a series of questions until a conclusion is reached
- **Example:** Classify **vertebrates** as
 - **Mammal** or
 - **Non-mammal**
- **Learning task**
 - Which questions should we ask?



Hunts algorithm

- Assign all data objects to the root



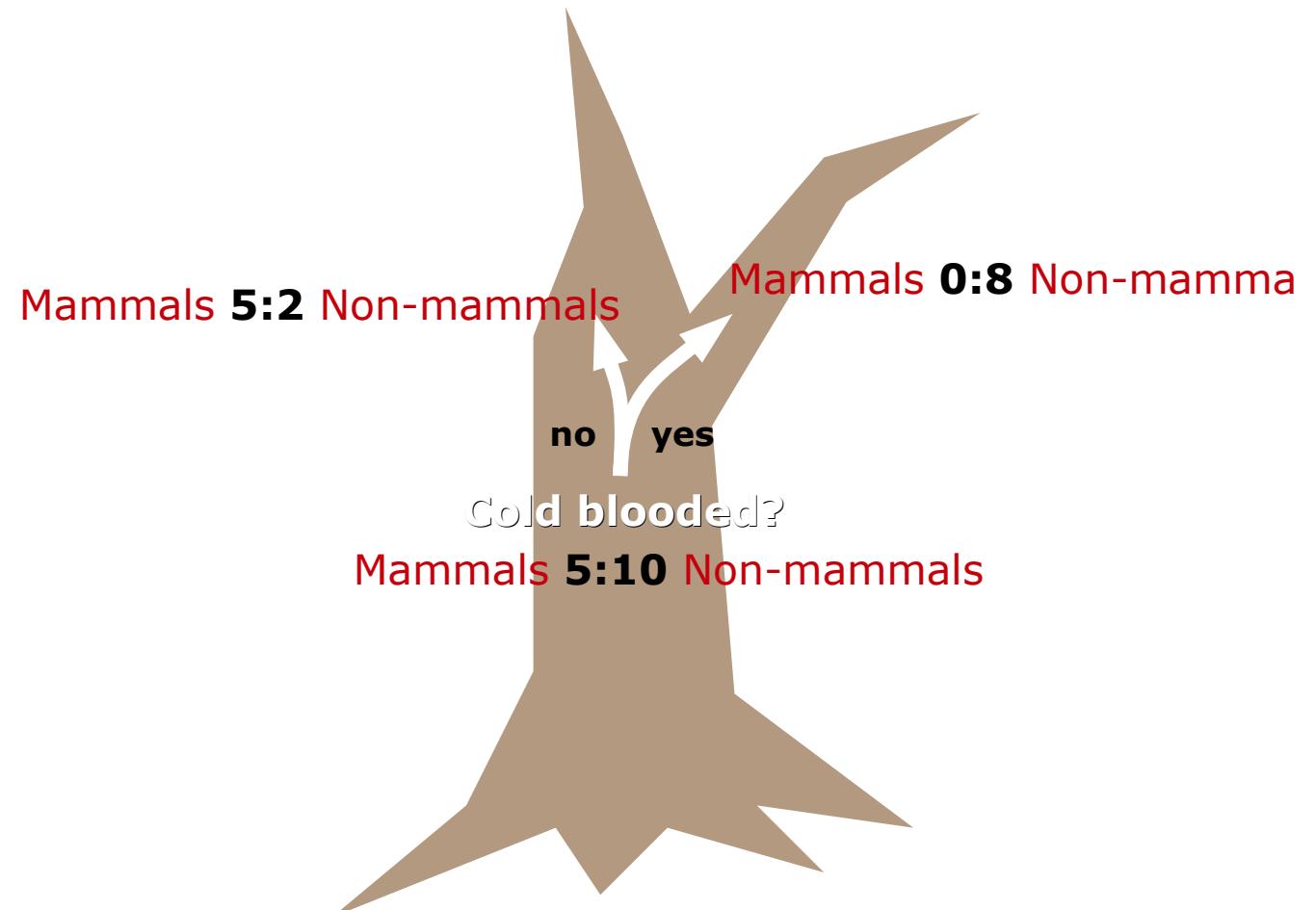
Hunts algorithm

- Select an attribute test condition
 - Find a good question to ask



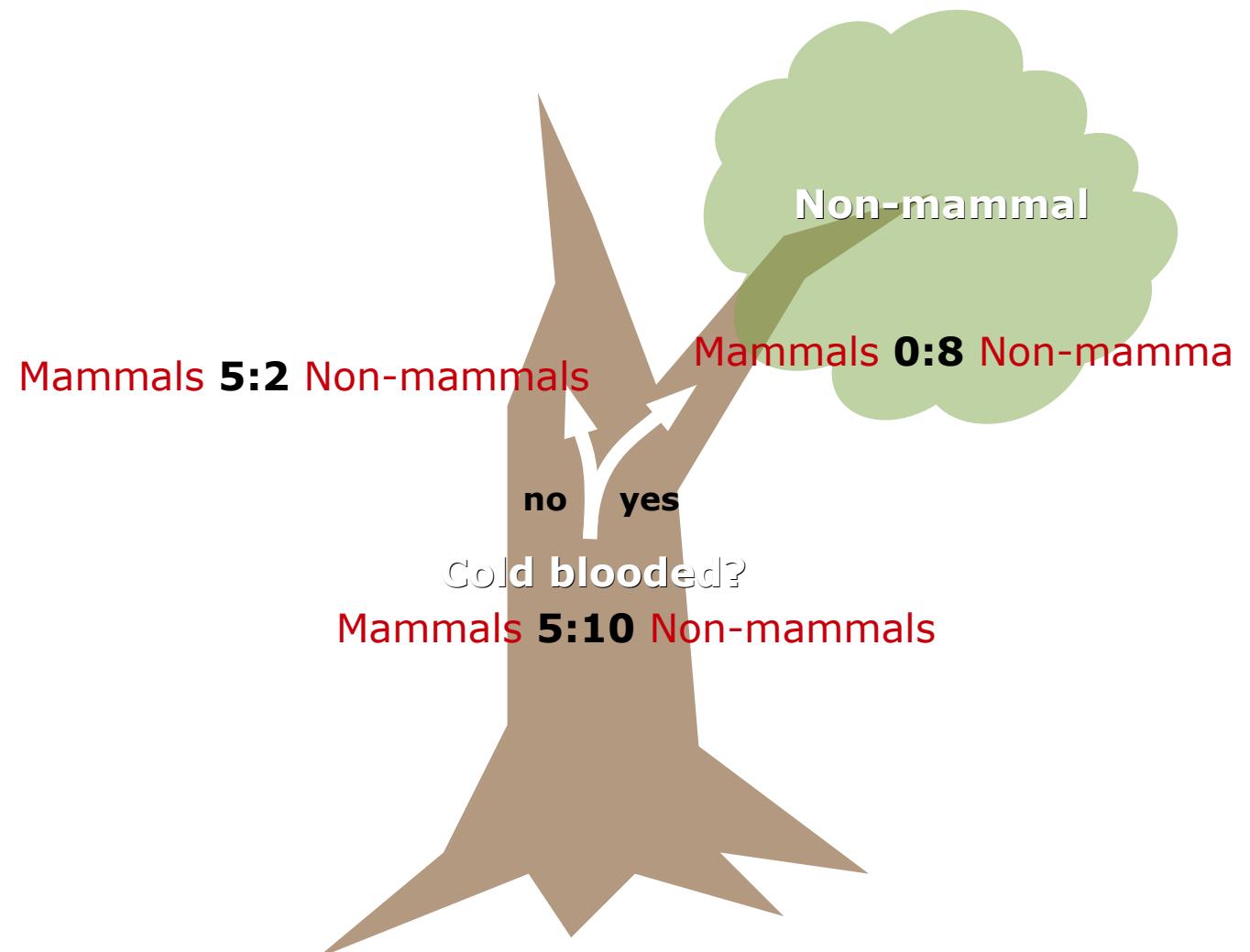
Hunt's Algorithm

- Partition the data objects into subsets according to the test condition



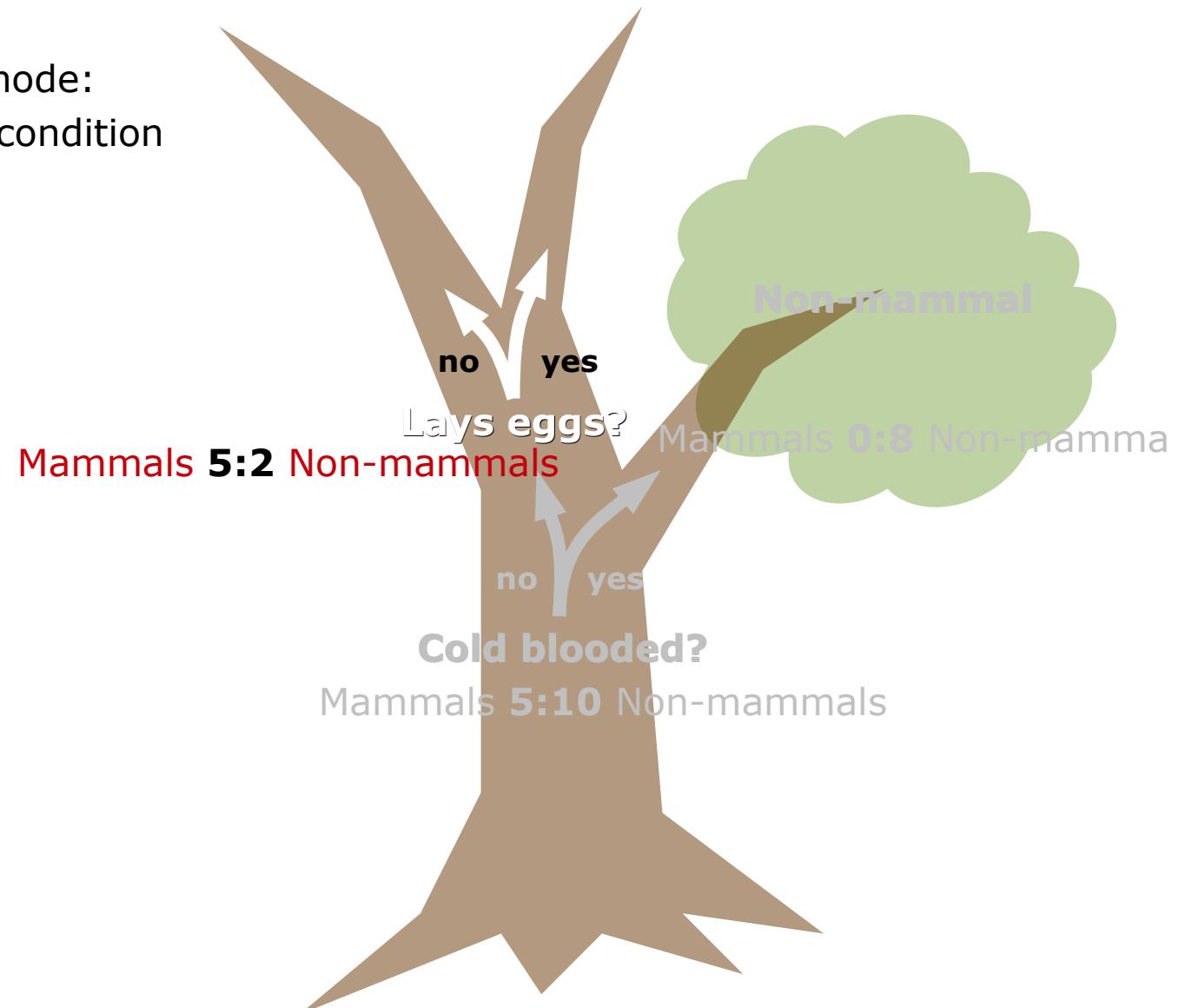
Hunts algorithm

- If all data objects belong to the same class
 - Create a leaf node



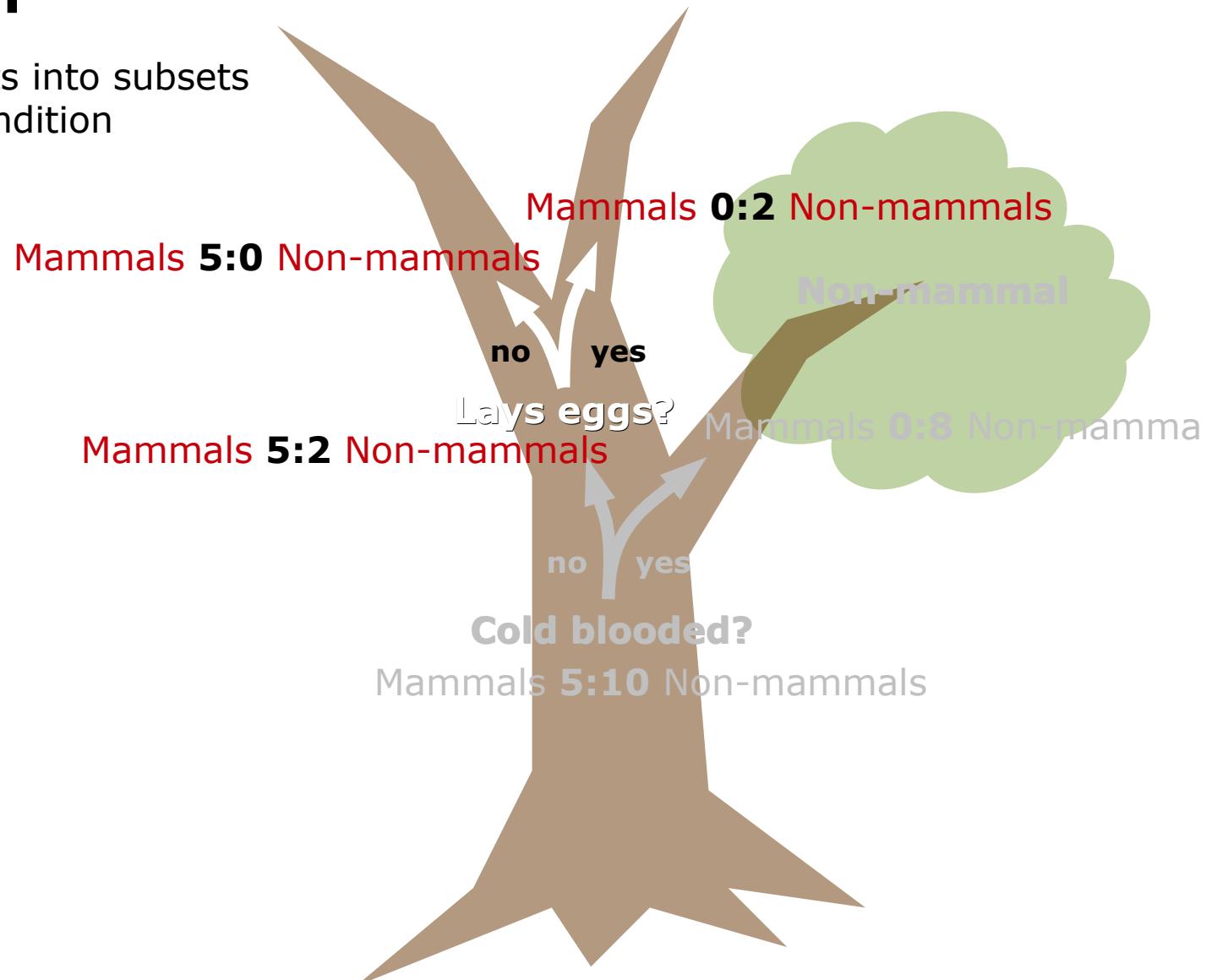
Hunts algorithm

- Repeat for each non-leave node:
 - Select an attribute test condition



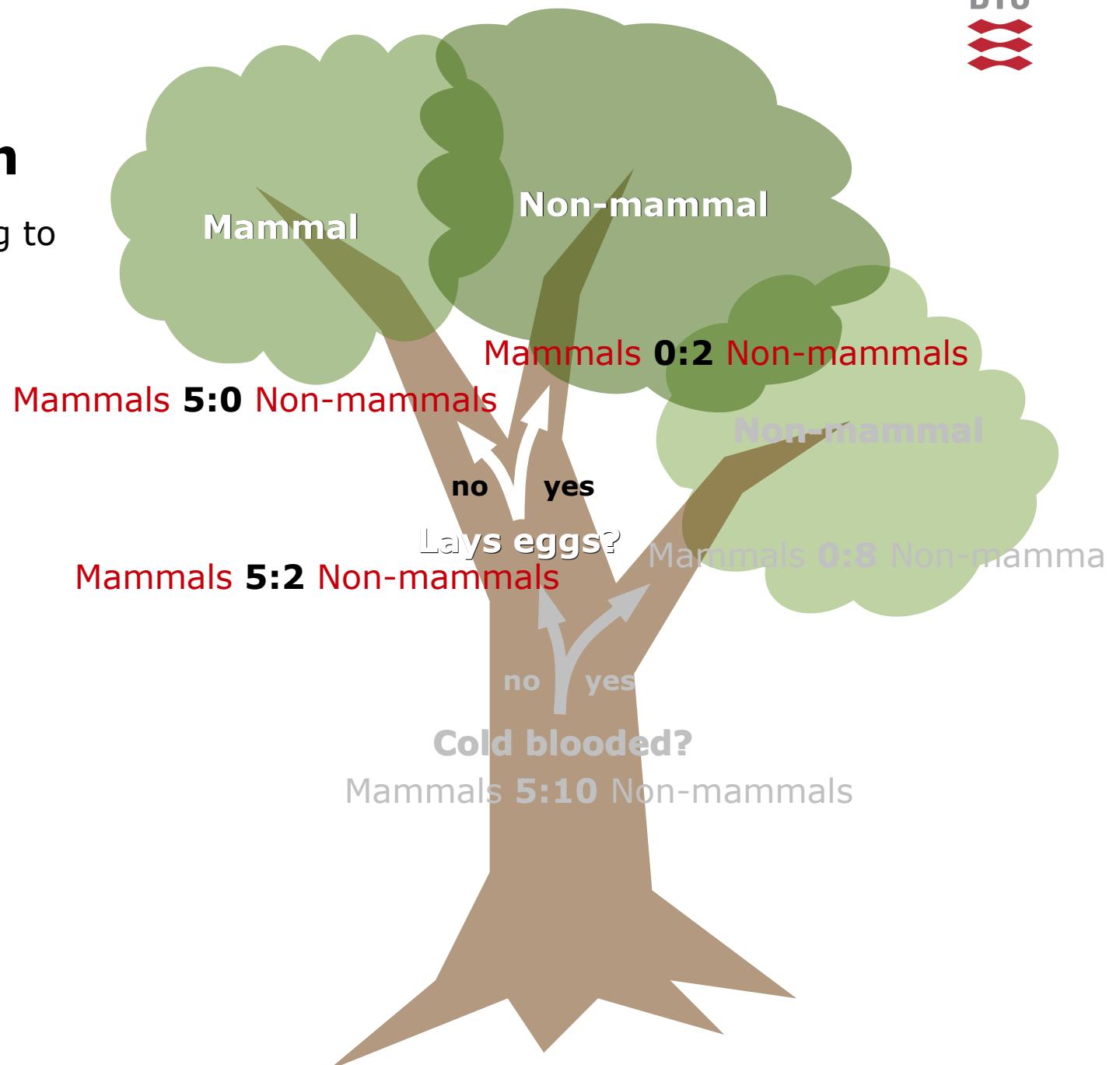
Hunts algorithm

- Partition the data objects into subsets according to the test condition



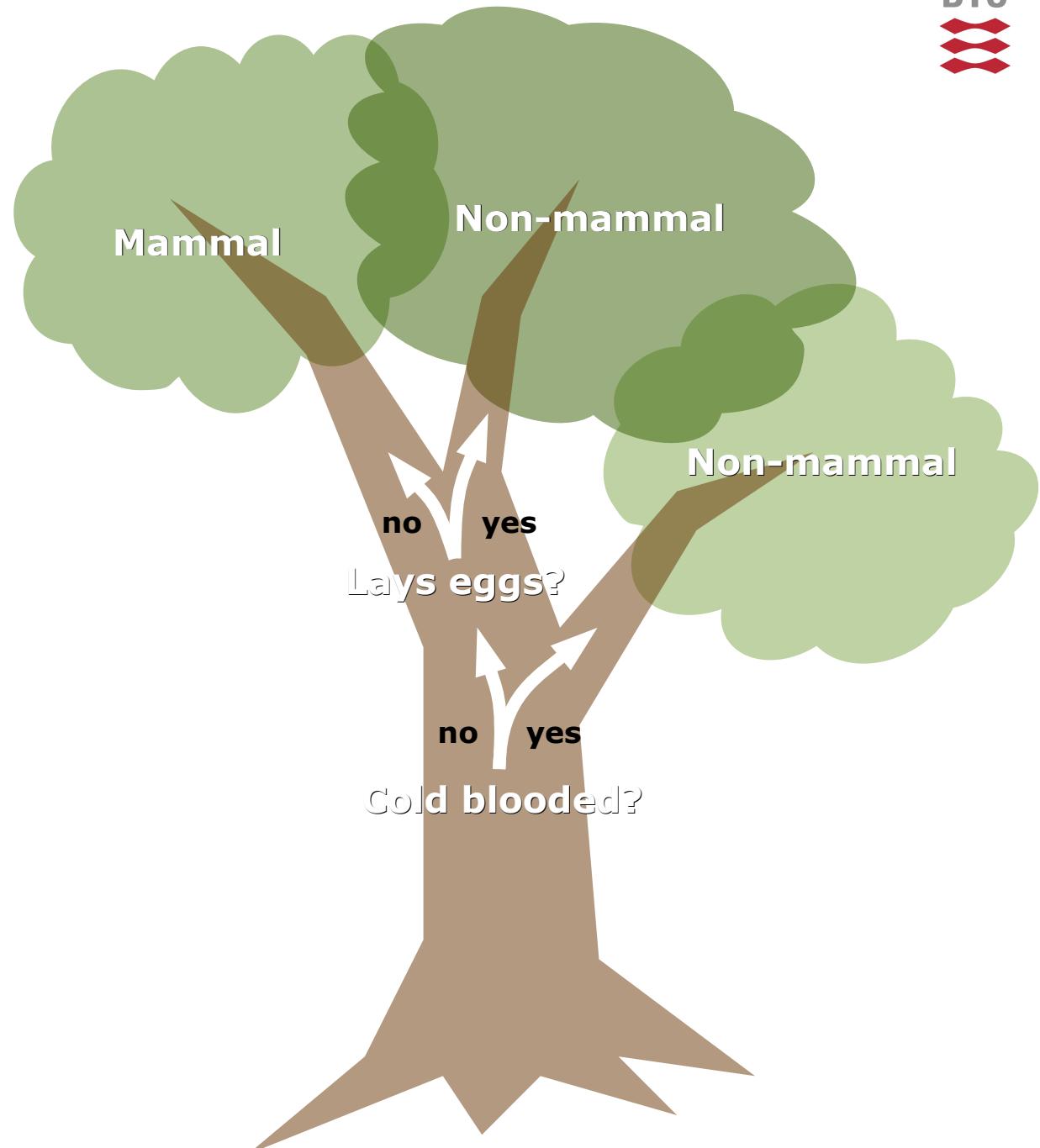
Hunts algorithm

- If all data objects belong to the same class
 - Create a leaf node



Hunts algorithm

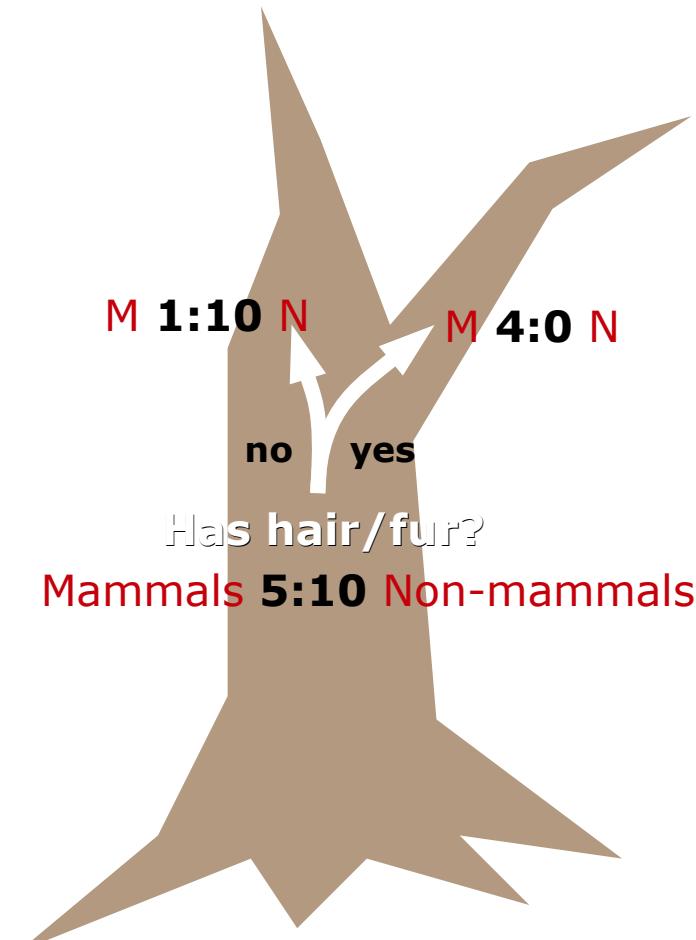
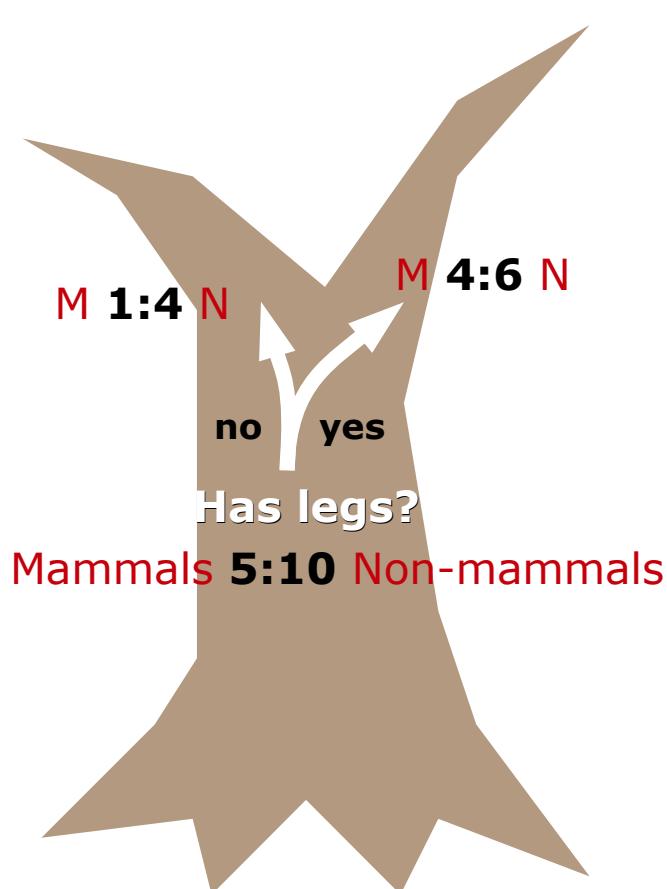
- But how do we find the **best question** at each step?





Selecting the best split

- Which of these two questions is best and why?



Selecting the best split

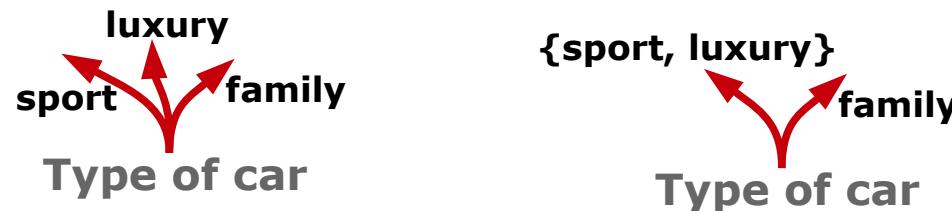
- Consider a large number of possible splits
- Compute a measure of impurity after the proposed split
 - For each new branch of the tree
 - Compute weighted average impurity
- Choose split that reduces impurity most

Which splits to consider

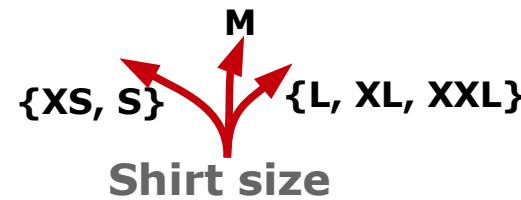
- Binary



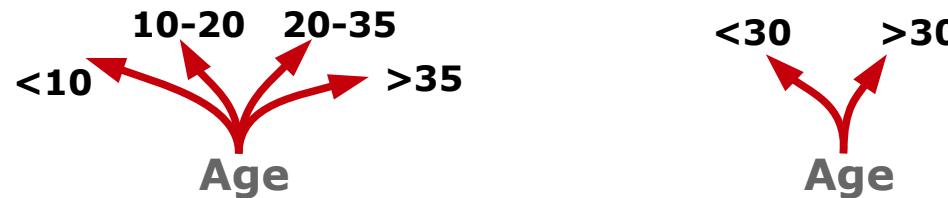
- Nominal

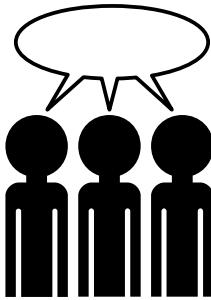


- Ordinal



- Continuous





Selecting the best split: Impurity measures

- Compute the purity gain, Δ

$$\text{Entropy}(t) = - \sum_{i=0}^{c-1} p(i|t) \log_2 p(i|t)$$

$$\text{Gini}(t) = 1 - \sum_{i=0}^{c-1} (p(i|t))^2$$

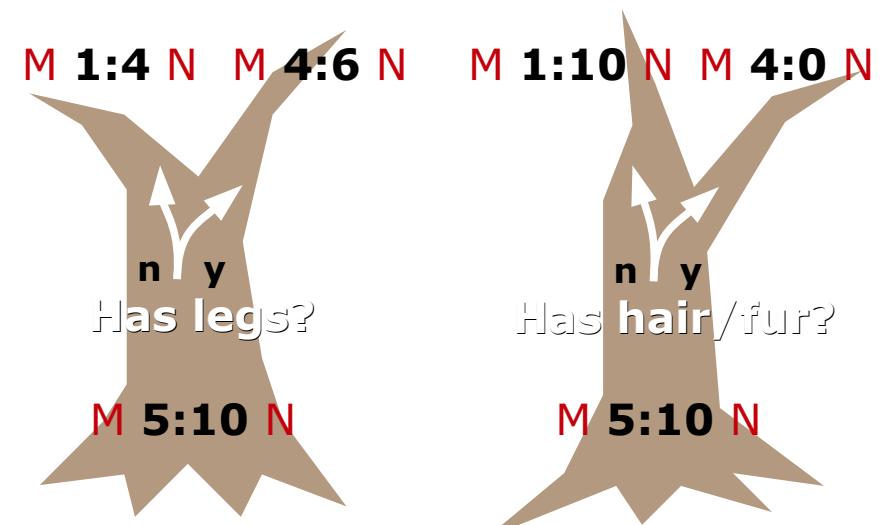
$$\text{Class. error}(t) = 1 - \max_i p(i|t)$$

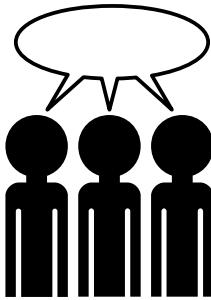
 $\Delta = ?$ $\Delta = ?$

$p(i|t)$ Fraction of objects that belong to class i

$$\Delta = I(\text{parent}) - \sum_{j=1}^k \frac{N(v_j)}{N} I(v_j)$$

Hint: Look at page 158-160 in Tan et al.





Selecting the best split: Impurity measures

- Compute the purity gain, Δ

$$\text{Entropy}(t) = - \sum_{i=0}^{c-1} p(i|t) \log_2 p(i|t)$$

$$\text{Gini}(t) = 1 - \sum_{i=0}^{c-1} (p(i|t))^2$$

$$\text{Class. error}(t) = 1 - \max_i p(i|t)$$

$p(i|t)$ Fraction of objects that belong to class i

$$\Delta = I(\text{parent}) - \sum_{j=1}^k \frac{N(v_j)}{N} I(v_j)$$

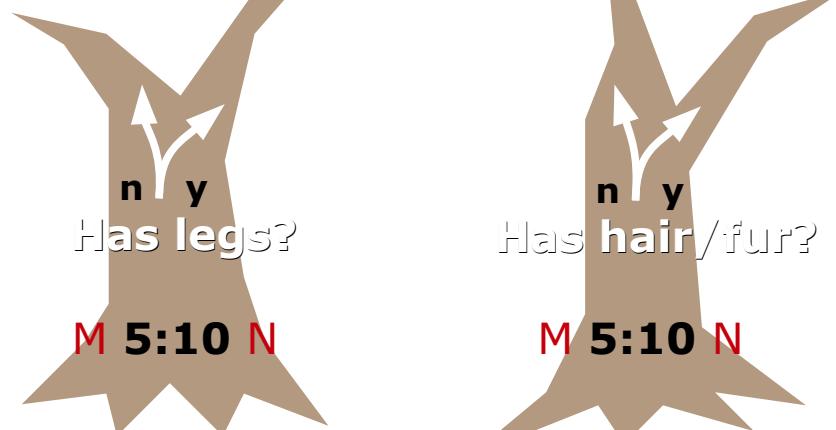
Hint: Look at page 158-160 in Tan et al.

$I(\text{Parent}) = -5/15 \cdot \log(5/15) - 10/15 \cdot \log(10/15)$ $= 0.9183$ $I(\text{left}) = -1/5 \cdot \log(1/5) - 4/5 \cdot \log(4/5)$ $= 0.7219$ $I(\text{right}) = -4/10 \cdot \log(4/10) - 6/10 \cdot \log(6/10)$ $= 0.9710$ $\Delta = 0.9183 - 5/15 \cdot 0.7219 - 10/15 \cdot 0.9710$ $= 0.0303$	$I(\text{Parent}) = -5/15 \cdot \log(5/15) - 10/15 \cdot \log(10/15)$ $= 0.9183$ $I(\text{left}) = -1/11 \cdot \log(1/11) - 10/11 \cdot \log(10/11)$ $= 0.4395$ $I(\text{right}) = -4/4 \cdot \log(4/4) - 0/4 \cdot \log(0/4)$ $= 0$ $\Delta = 0.9183 - 11/15 \cdot 0.4395 - 4/15 \cdot 0$ $= 0.5960$
--	--

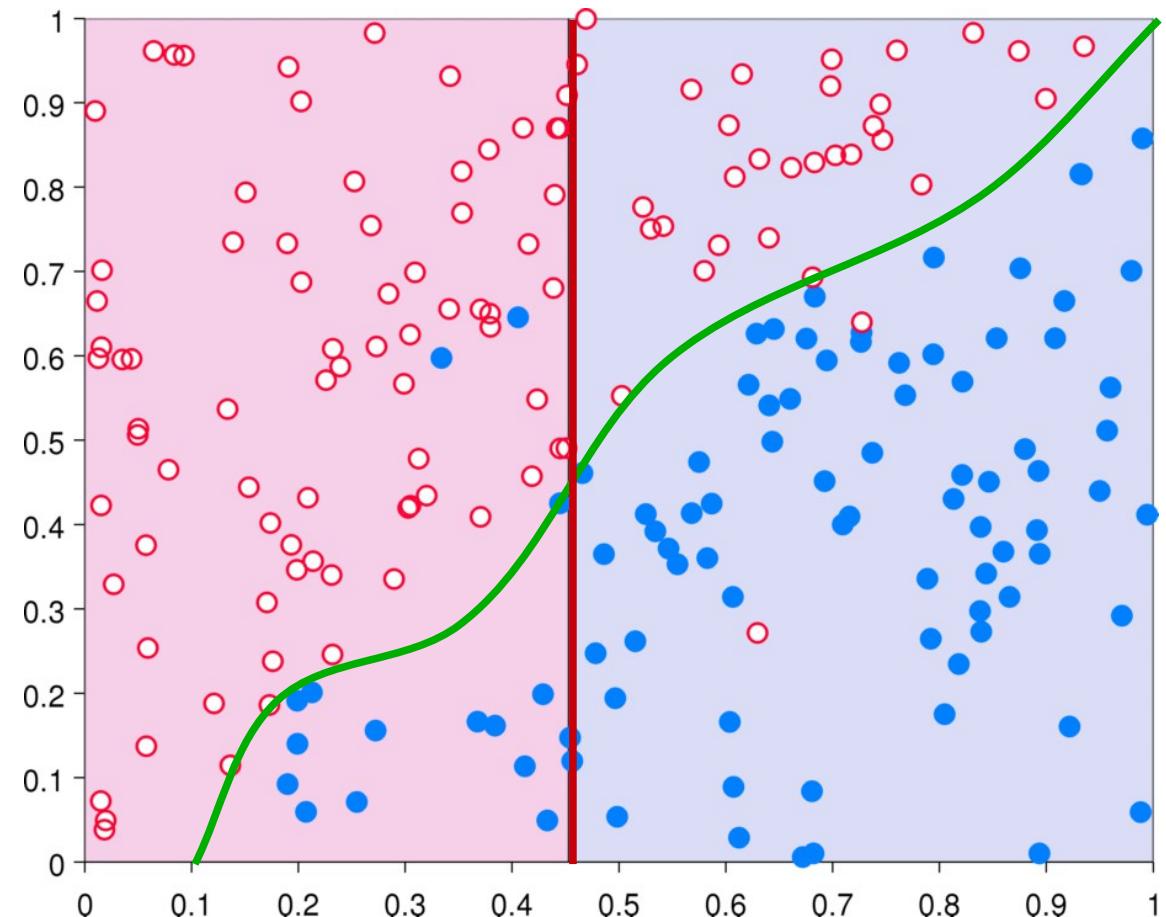
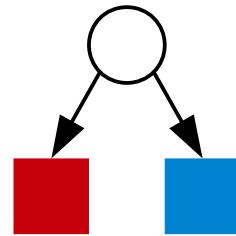
$I(\text{Parent}) = 1 - (5/15)^2 - (10/15)^2$ $= 0.4444$ $I(\text{left}) = 1 - (1/5)^2 - (4/5)^2$ $= 0.3200$ $I(\text{right}) = 1 - (4/10)^2 - (6/10)^2$ $= 0.4800$ $\Delta = 0.4444 - 5/15 \cdot 0.3200 - 10/15 \cdot 0.4800$ $= 0.0177$	$I(\text{Parent}) = 1 - (5/15)^2 - (10/15)^2$ $= 0.4444$ $I(\text{left}) = 1 - (1/11)^2 - (10/11)^2$ $= 0.1653$ $I(\text{right}) = 1 - (4/4)^2 - (0/4)^2$ $= 0$ $\Delta = 0.4444 - 11/15 \cdot 0.1653 - 4/15 \cdot 0$ $= 0.3232$
--	---

$I(\text{Parent}) = 1 - 10/15$ $= 5/15$ $I(\text{left}) = 1 - 4/5$ $= 1/5$ $I(\text{right}) = 1 - 6/10$ $= 4/10$ $\Delta = 5/15 - 5/15 \cdot 1/5 - 10/15 \cdot 4/10$ $= 0$	$I(\text{Parent}) = 1 - 10/15$ $= 5/15$ $I(\text{left}) = 1 - 10/11$ $= 1/11$ $I(\text{right}) = 1 - 4/4$ $= 0$ $\Delta = 5/15 - 11/15 \cdot 1/11 - 4/15 \cdot 0$ $= 0.2667$
---	---

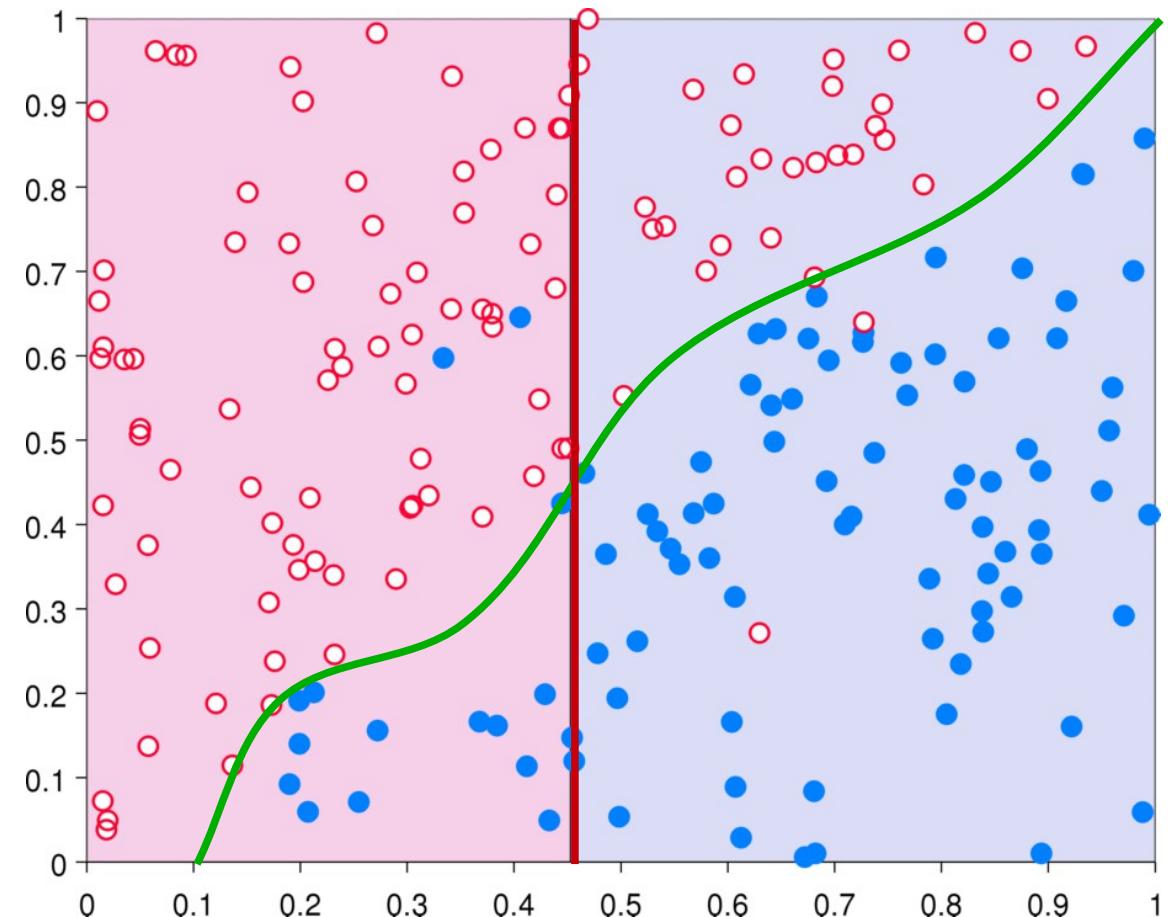
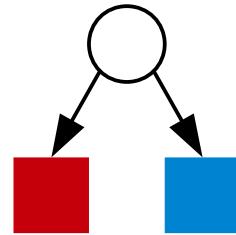
M 1:4 N M 4:6 N M 1:10 N M 4:0 N



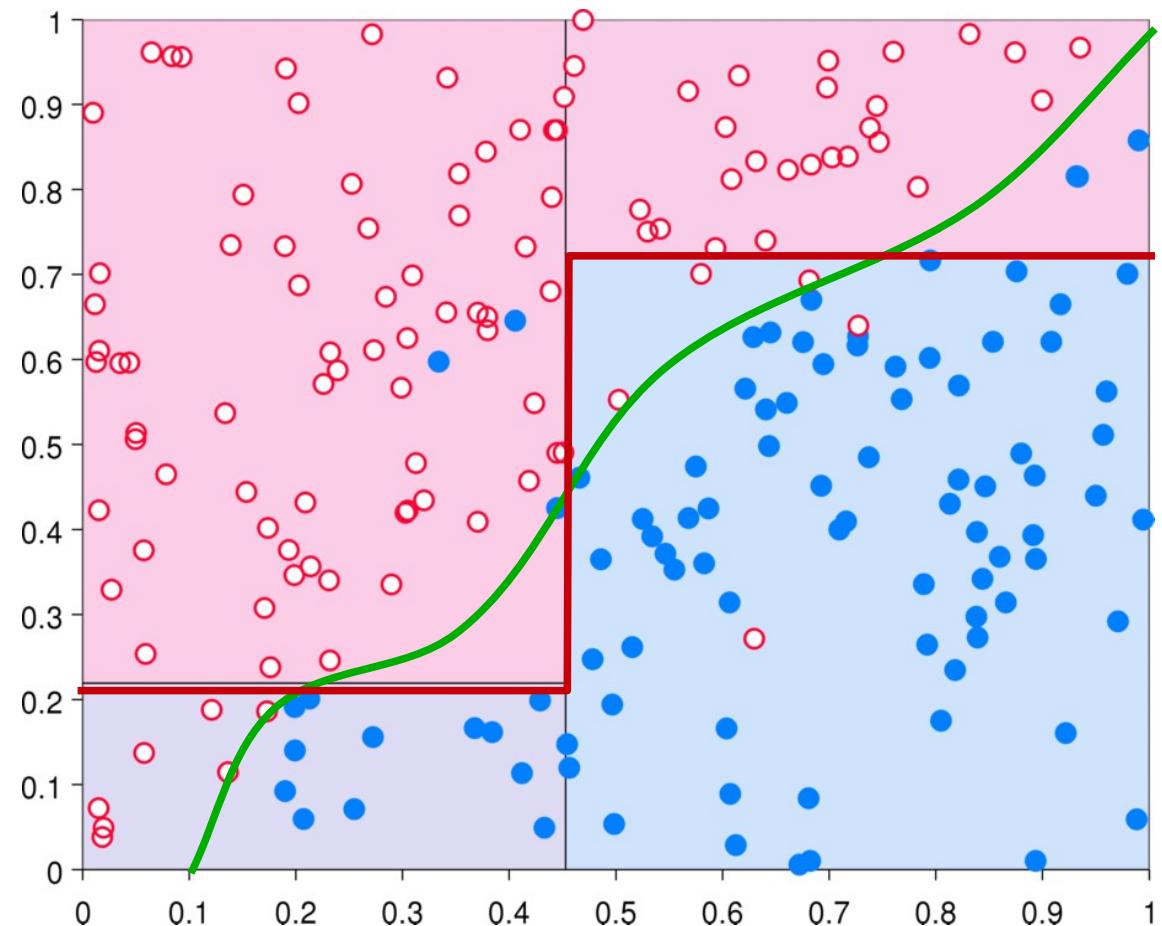
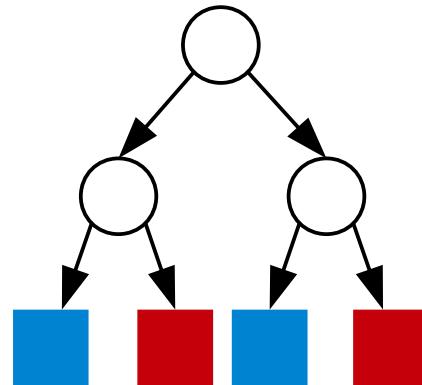
Classification Trees



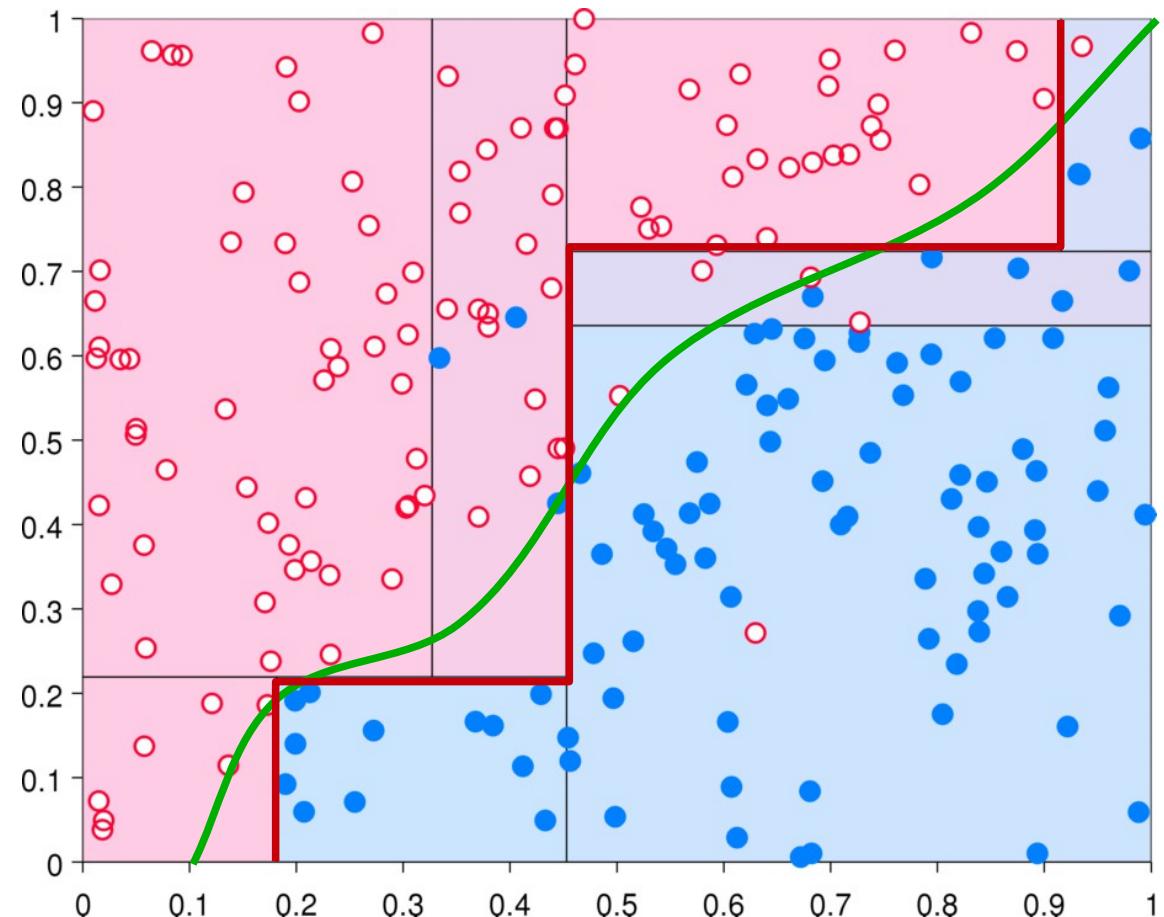
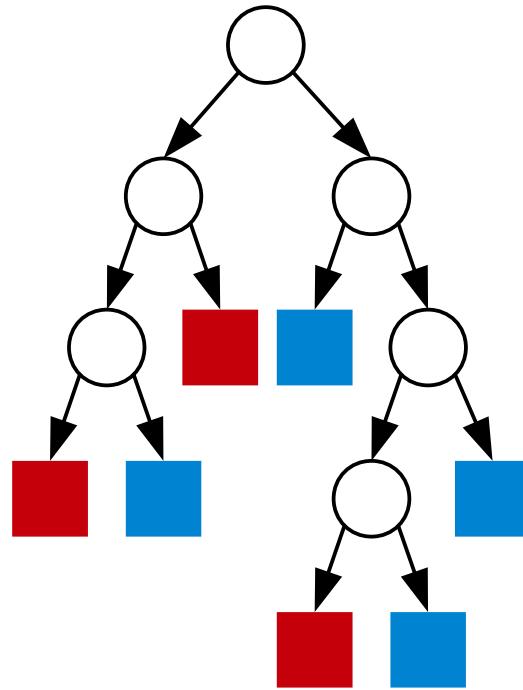
Classification Trees



Classification trees



Classification trees

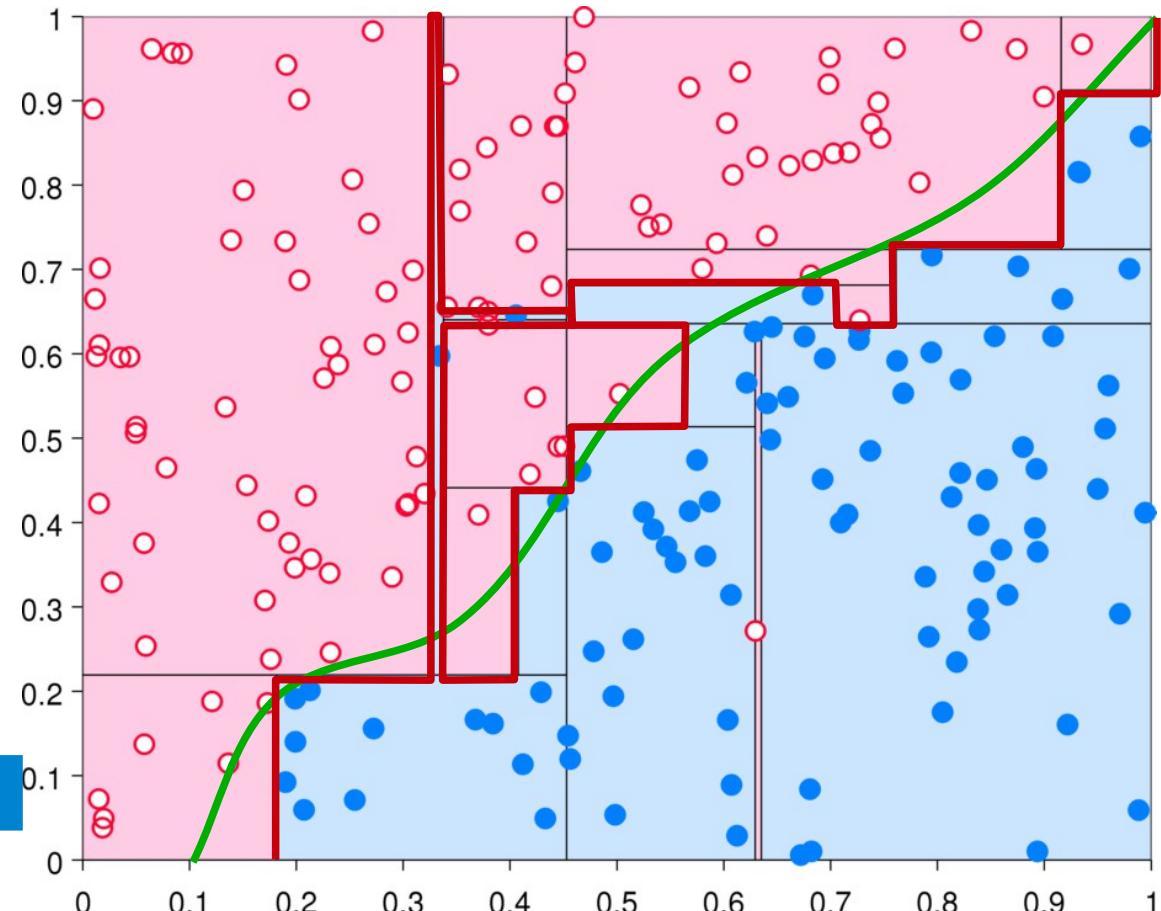
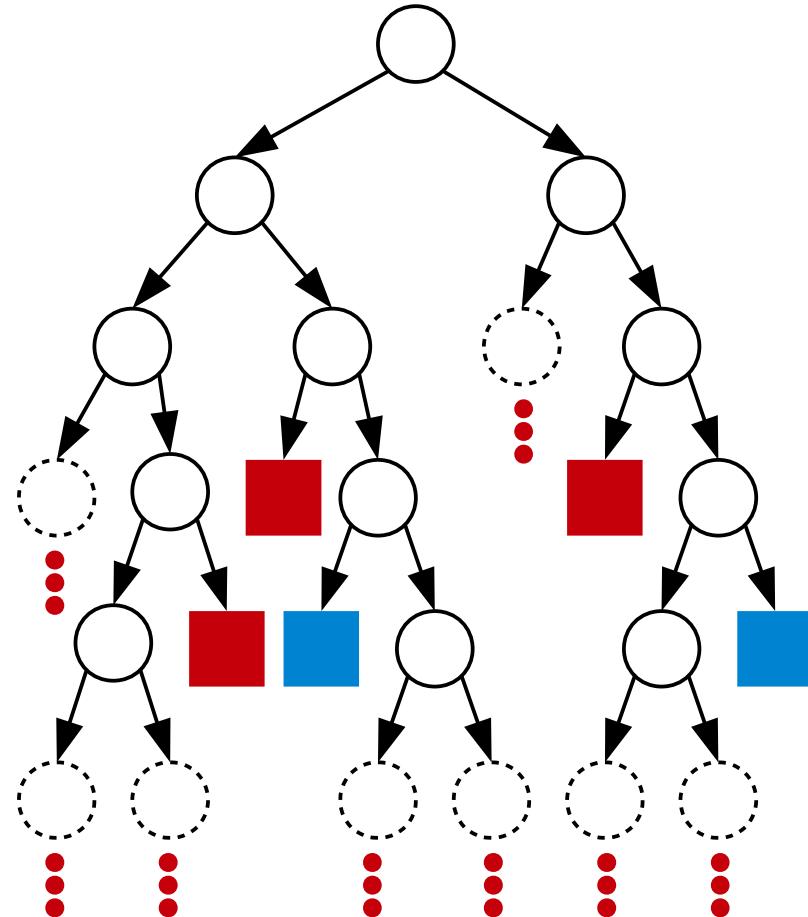


Classification trees

Common stopping criteria:

All records have the same class label

The number of observations have fallen below some minimum threshold



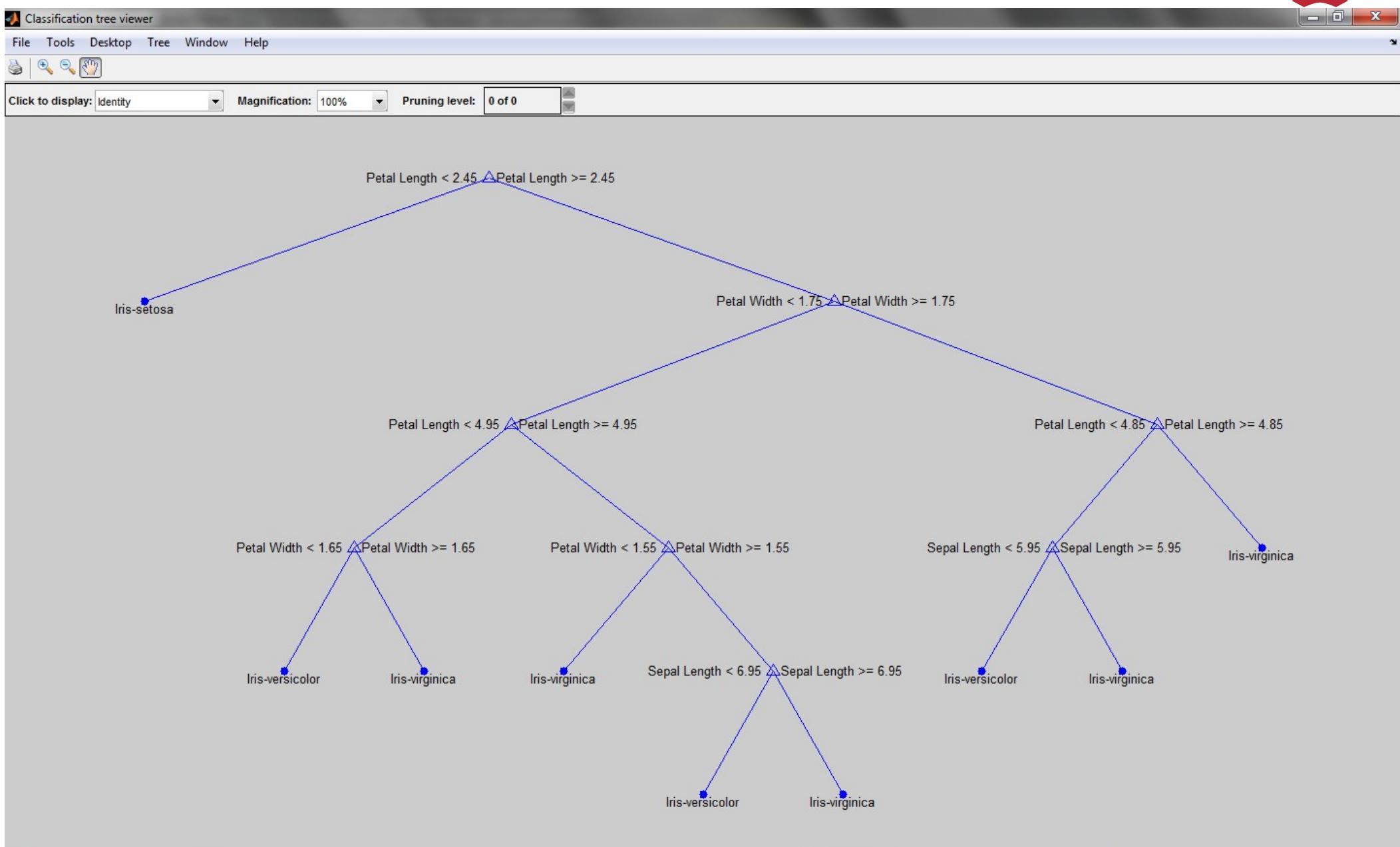
The iris data set

- **Three flowers**
 - 50 instances of each class, 150 in total
- **Attributes**
 - Sepal (outermost leaves)
 - length in cm
 - width in cm
 - Petal (innermost leaves)
 - length in cm
 - width in cm
 - Class of flower
 - Iris Setosa
 - Iris Versicolour
 - Iris Virginica

Flower ID	Attribute			
	Sepal Length	Sepal Width	Petal Length	Petal Width
1	5.1	3.5	1.4	0.2
2	4.9	3.0	1.4	0.2
3	4.7	3.2	1.3	0.2
4	4.6	3.1	1.5	0.2
.
.
150	5.9	3.0	5.1	1.8



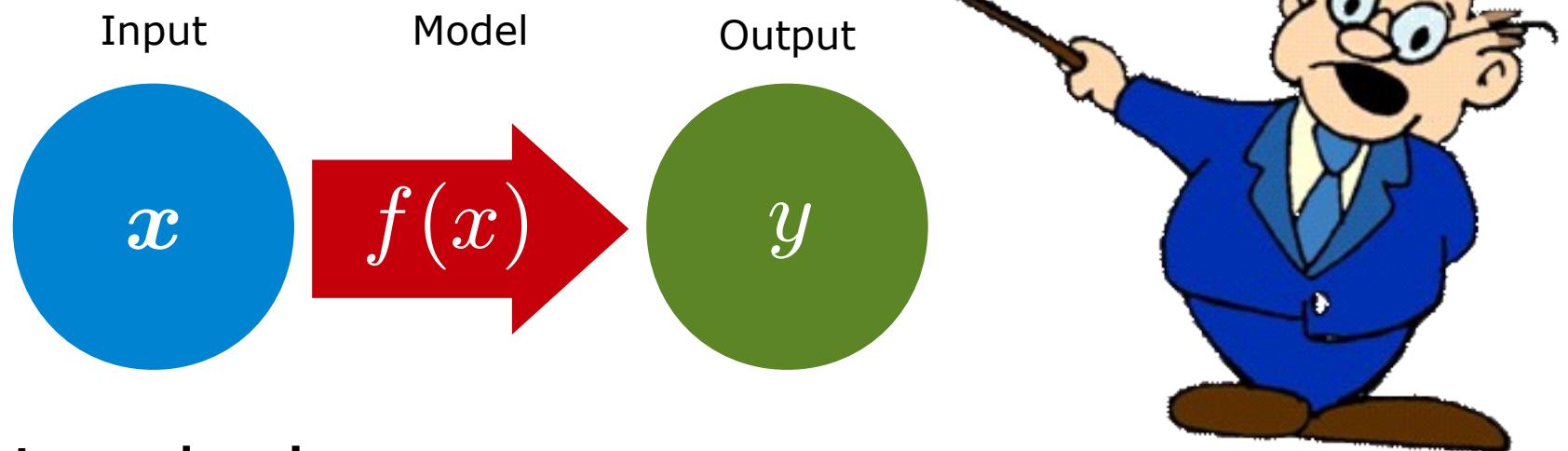
$X^{\text{Observation} \times \text{Attribute}}$



What would the following iris flower be classified as?

Sepal Length	Sepal Width	Petal Length	Petal Width
4.0	3.5	3.0	2.0

Supervised learning



- **Mapping between domains**
 - Classification: Discrete (nominal) output
 - Regression: Continuous output

Supervised learning

- **Data**

- Inputs and outputs

$$\{\mathbf{x}_n, y_n\}_{n=1}^N$$

- **Model**

- Function that maps inputs to outputs

$$f(\mathbf{x})$$

- **Cost function**

- Dissimilarity measure between data and model

$$d(y, f(\mathbf{x}))$$

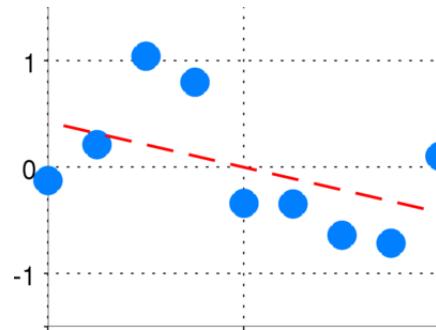
Regression

- **Definition:** Learning a function that maps a data object to a continuous-valued output
- **Why Regression?**
 - Descriptive modeling
 - Explain / understand the relation between attributes and continuous-valued output
 - Predictive modeling
 - Predict the output value of a new data object

Linear regression

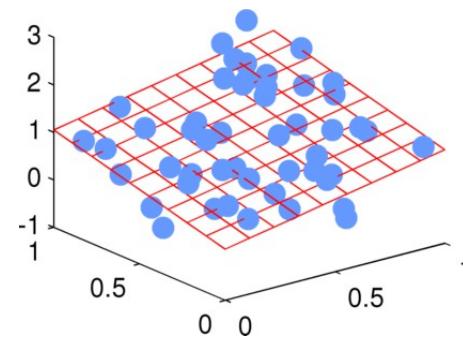
- 1-dimensional inputs

$$f(x) = w_0 + w_1 x$$



- 2-dimensional inputs

$$f(\mathbf{x}) = w_0 + w_1 x_1 + w_2 x_2$$



- K-dimensional inputs

$$f(\mathbf{x}) = w_0 + w_1 x_1 + w_2 x_2 + \cdots + w_K x_K$$

Linear regression

- K-dimensional inputs

$$f(\mathbf{x}) = w_0 + w_1x_1 + w_2x_2 + \cdots + w_Kx_K$$

- Non-linearly transformed inputs

$$f(x) = w_0 + w_1x + w_2x^2 + \cdots + w_Kx^K$$

$$f(x) = w_0 + w_1\sin(x) + w_2\cos(x)$$

Linear regression

- K-dimensional inputs

$$f(\mathbf{x}) = w_0 + w_1 x_1 + w_2 x_2 + \cdots + w_K x_K$$

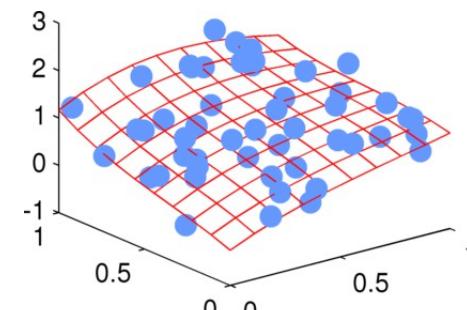
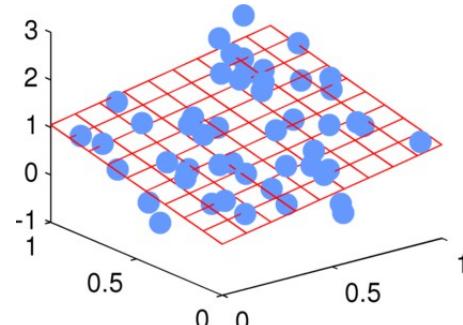
- Non-linearly transformed inputs

$$f(x) = w_0 + w_1 x + w_2 x^2 + \cdots + w_K x^K$$

$$f(x) = w_0 + w_1 \sin(x) + w_2 \cos(x)$$

- Example

$$f(\mathbf{x}) = w_0 + w_1 x_1 + w_2 x_2$$



$$\begin{aligned} f(\mathbf{x}) = & w_0 + w_1 x_1 + w_2 x_2 \\ & + w_3 x_1^2 + w_4 x_2^2 + w_5 x_1 x_2 \\ & + w_6 x_1^3 + w_7 x_1^2 x_2 + w_8 x_1 x_2^2 + w_9 x_2^3 \end{aligned}$$

Vector notation

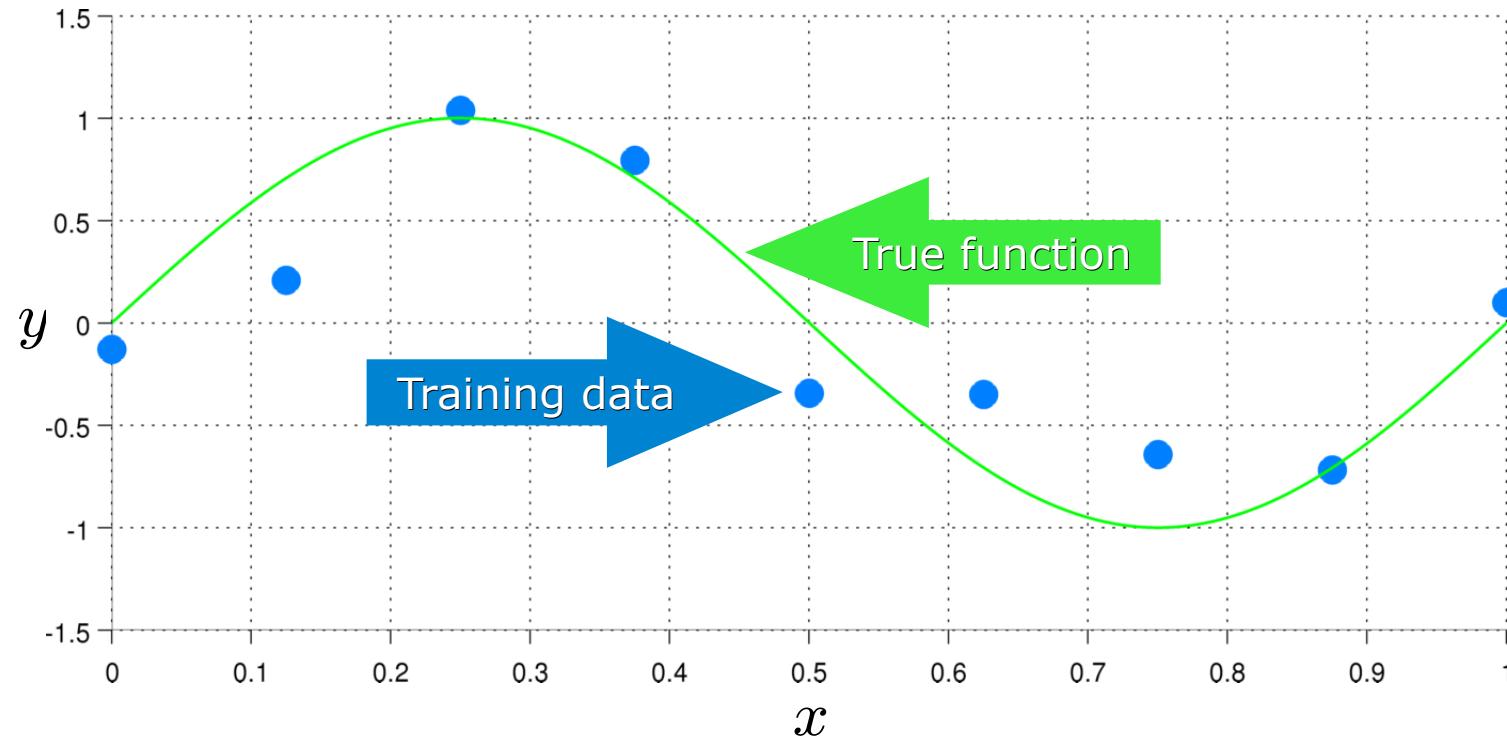
- The linear model can be written compactly using vector notation

$$f(\mathbf{x}) = w_0 + w_1 x_1 + w_2 x_2 + \cdots + w_K x_K$$

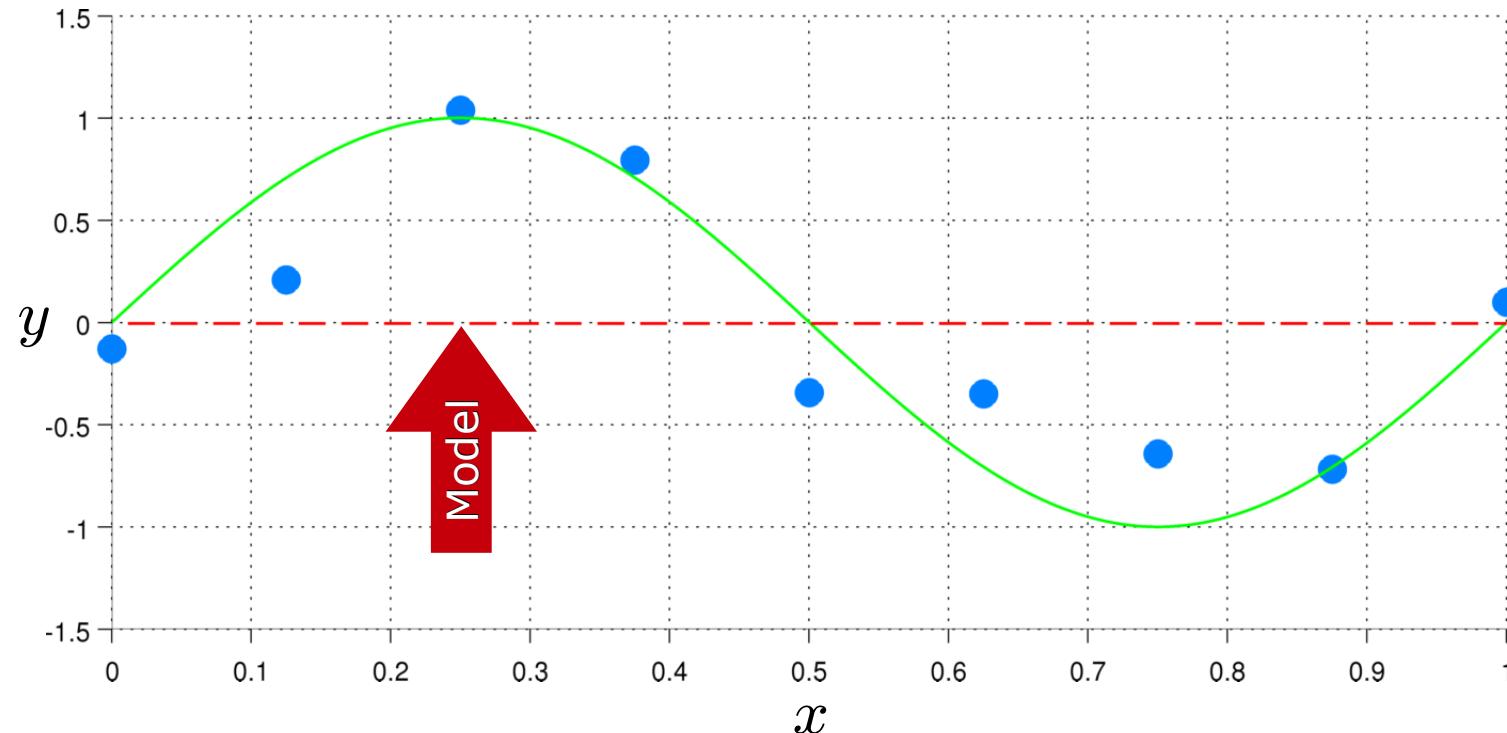
$$= \sum_{k=0}^K w_k x_k = \boxed{\mathbf{x}^\top \mathbf{w}}$$

- where $x_0 = 1$

Linear regression



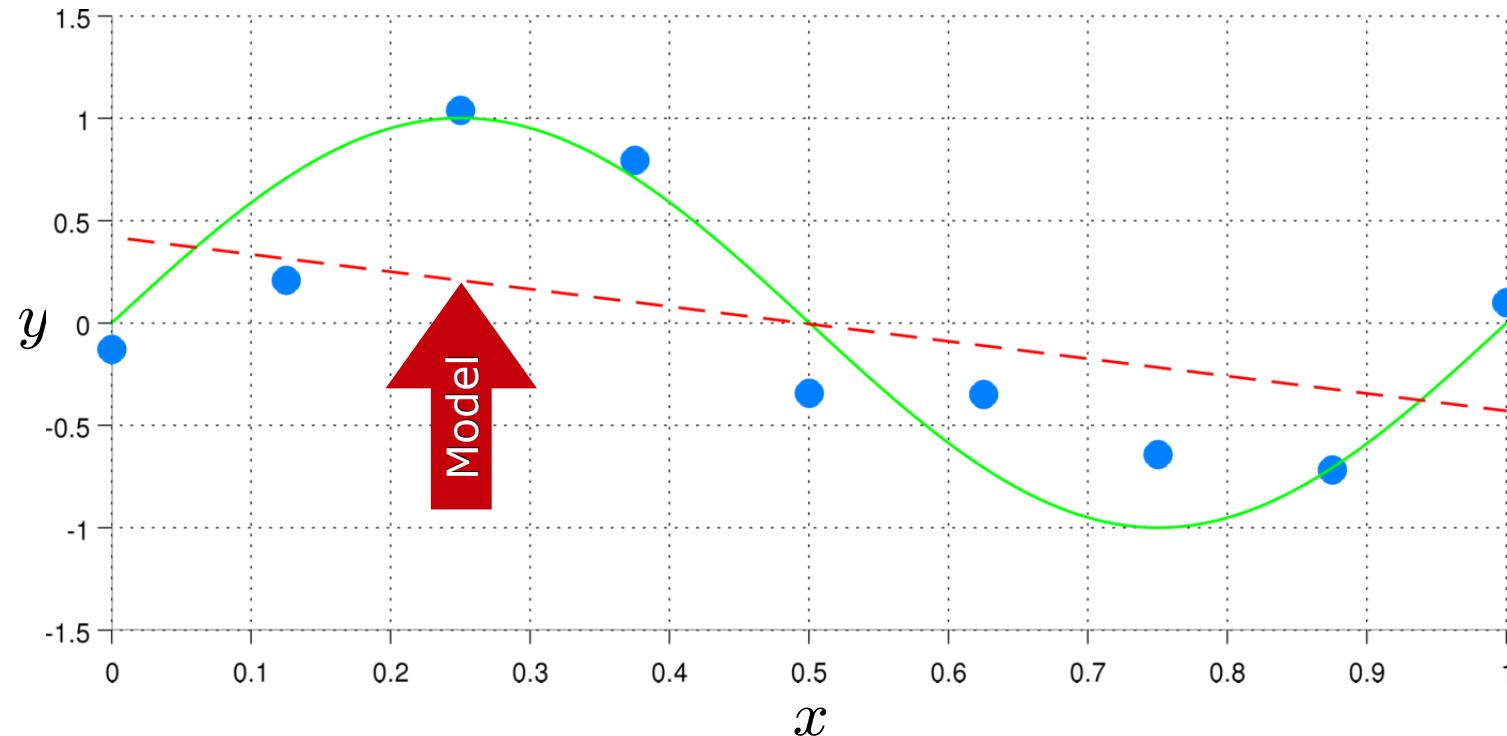
Linear regression



Model

$$f(x) = w_0$$

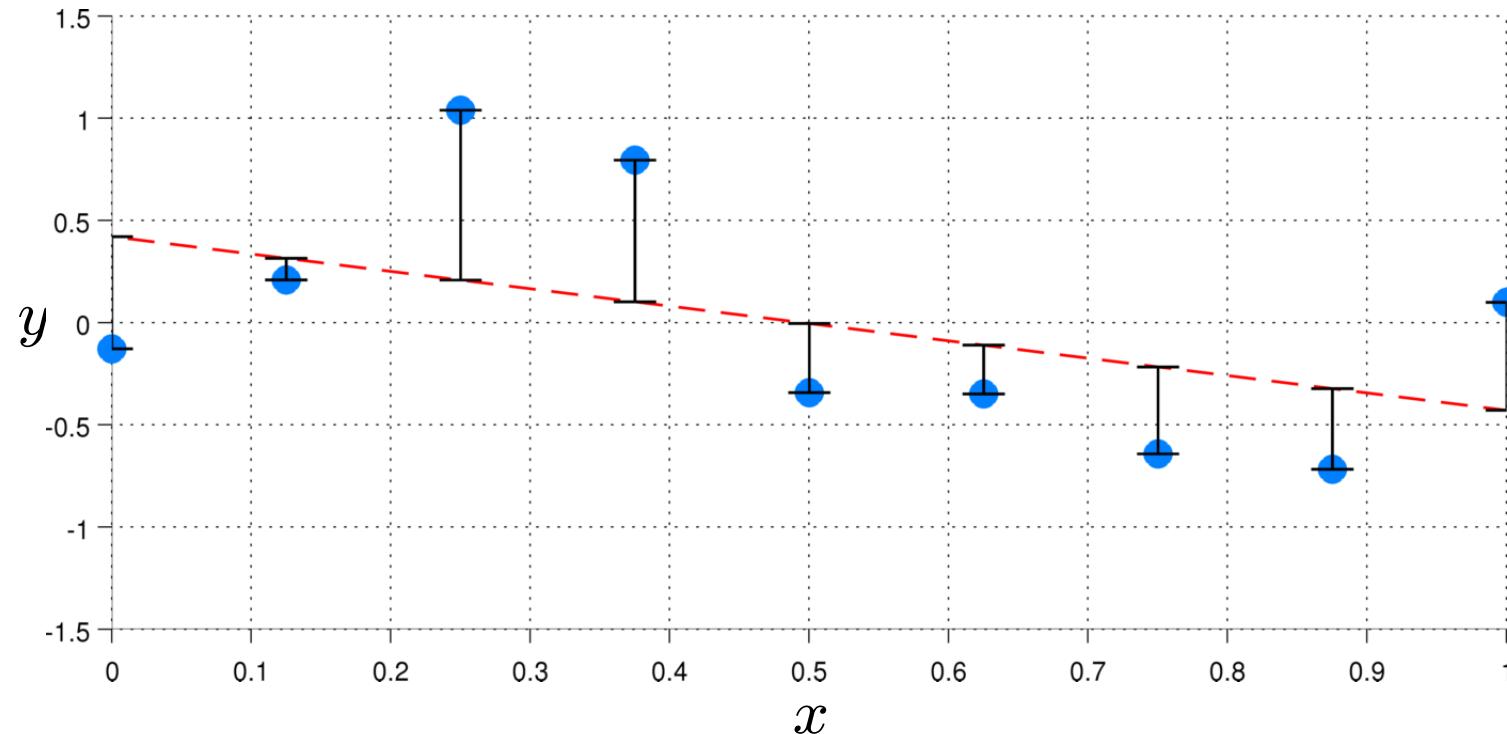
Linear regression



Model

$$f(x) = w_0 + w_1 x$$

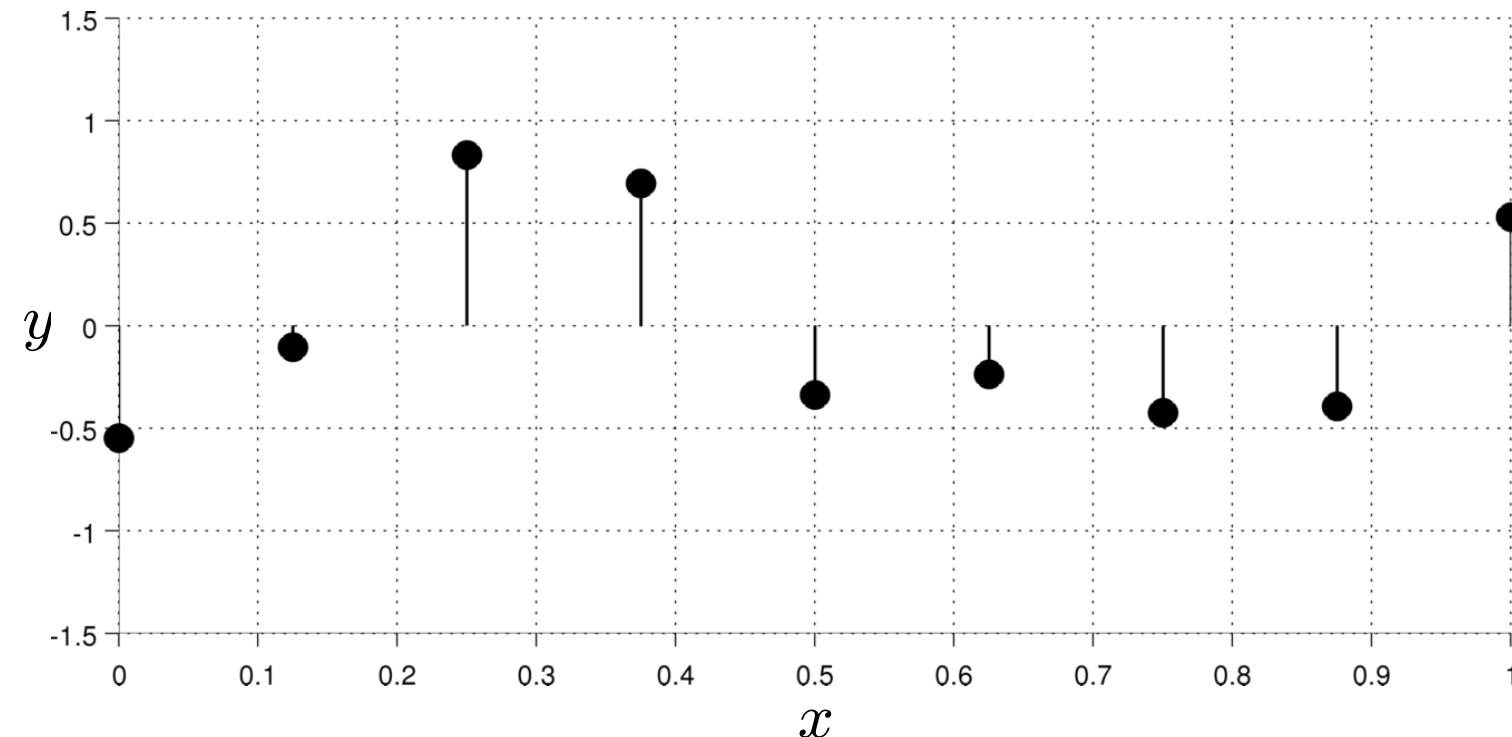
Residual error



Model

$$f(x) = w_0 + w_1 x$$

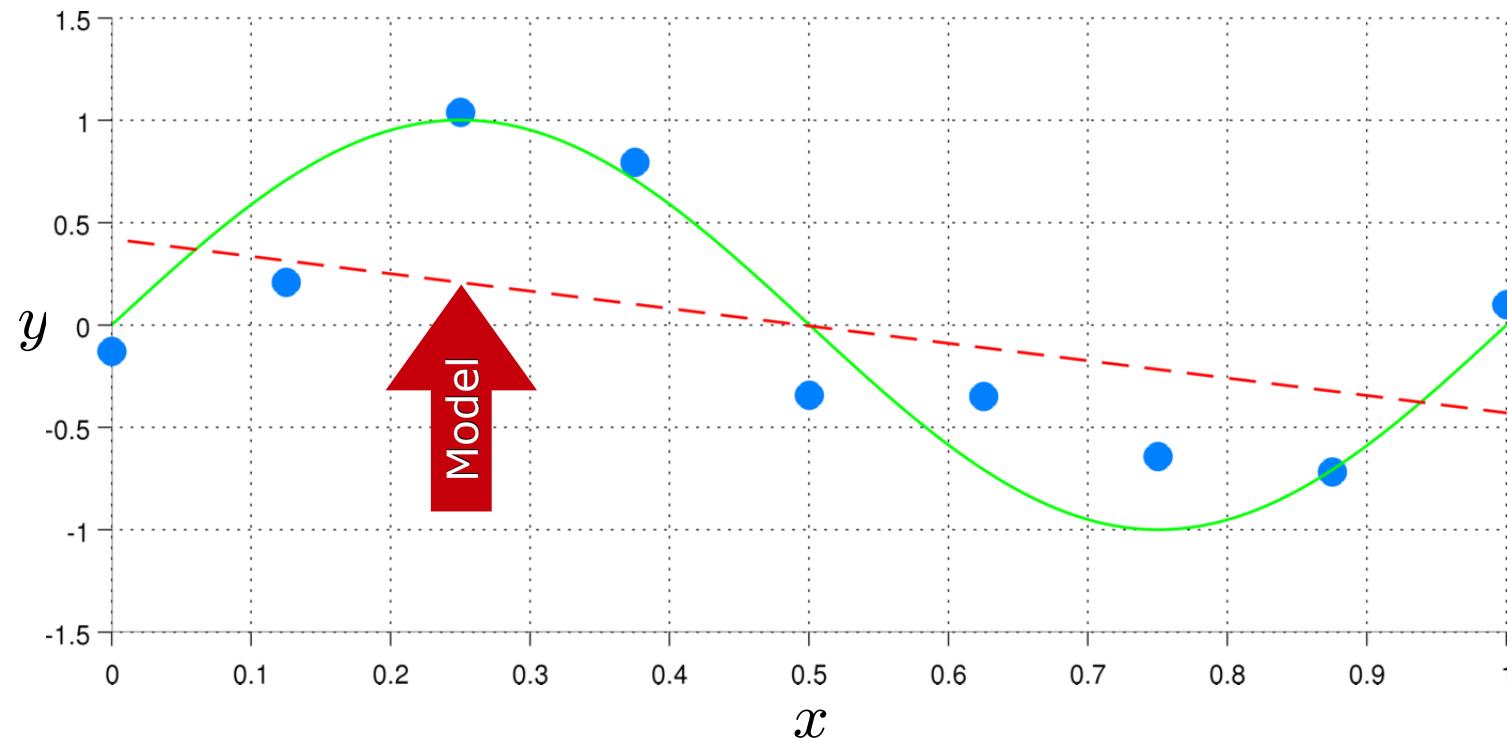
Residual error



Model

$$f(x) = w_0 + w_1 x$$

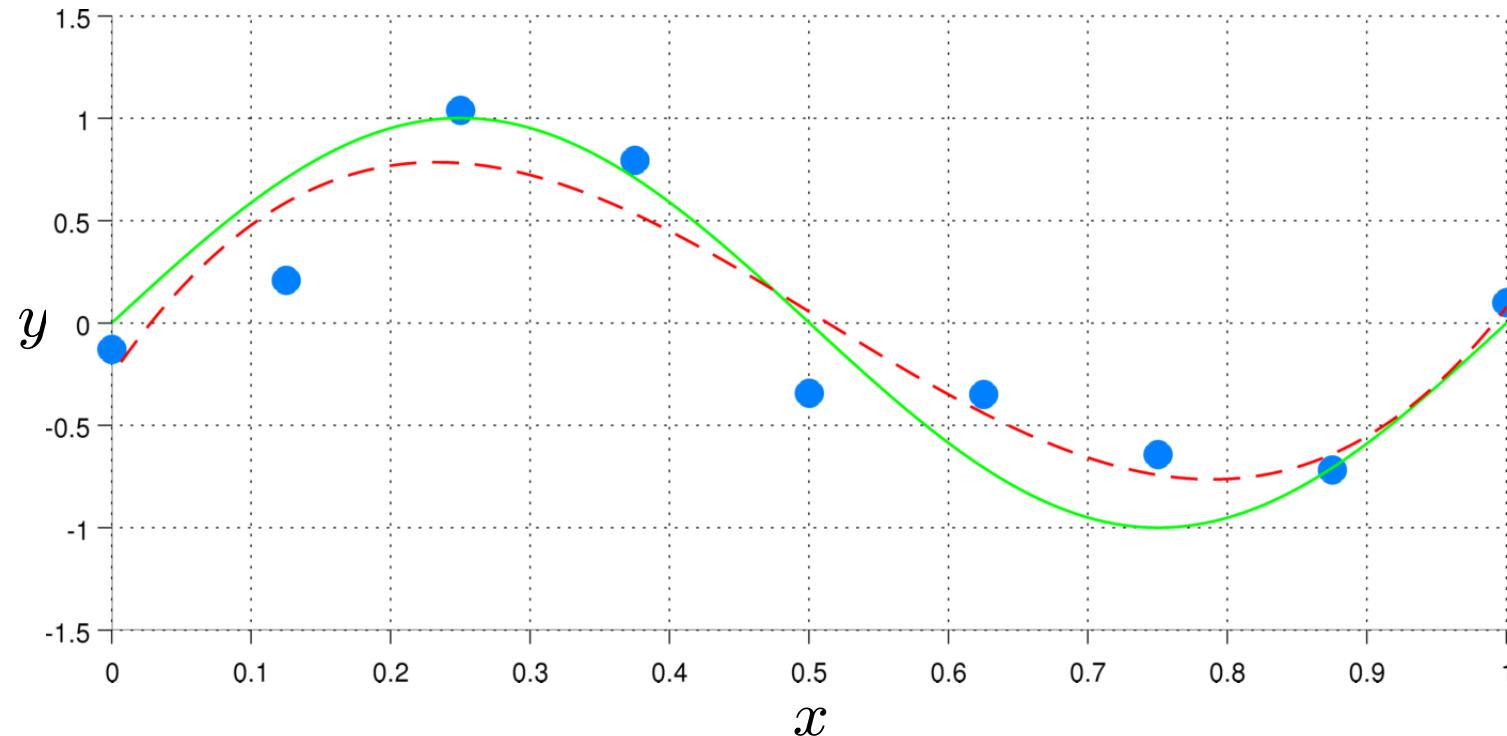
Linear regression



Model

$$f(x) = w_0 + w_1 x$$

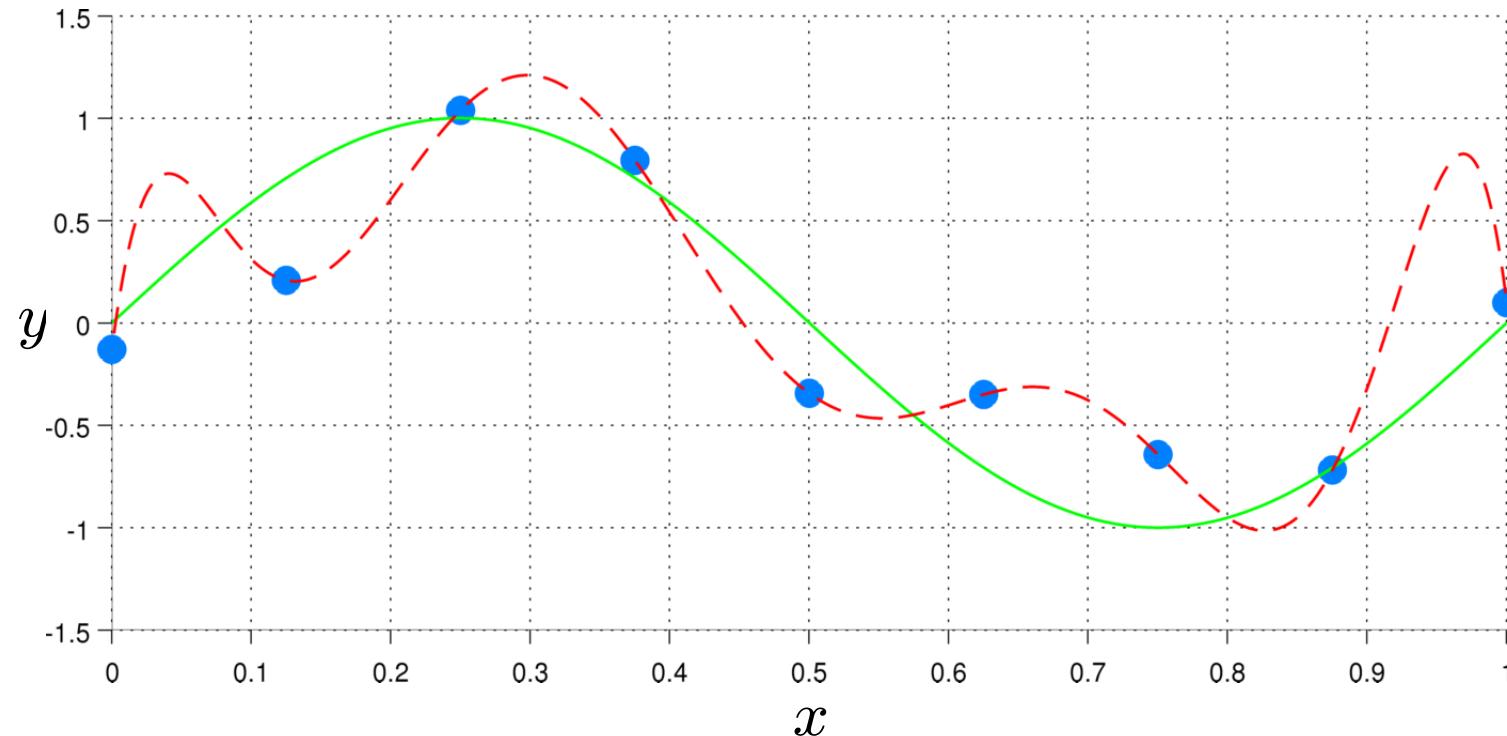
Linear regression



Model

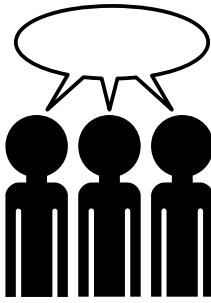
$$f(x) = w_0 + w_1x + w_2x^2 + w_3x^3$$

Linear regression



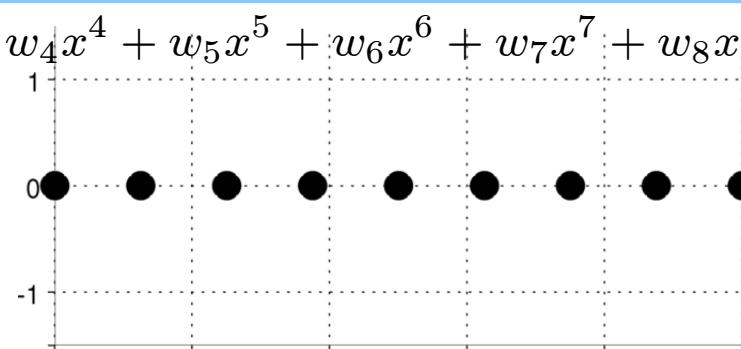
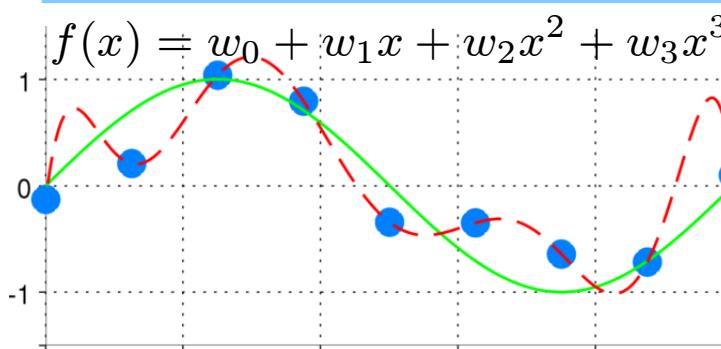
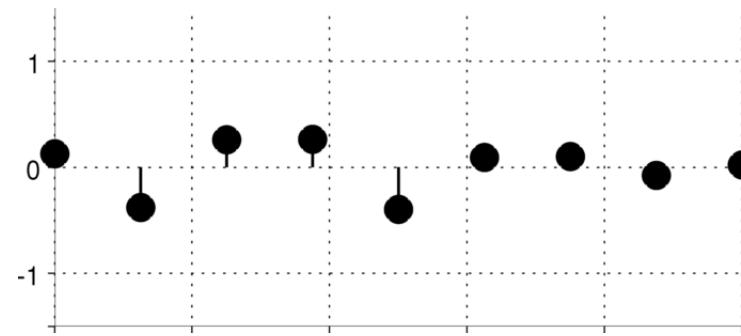
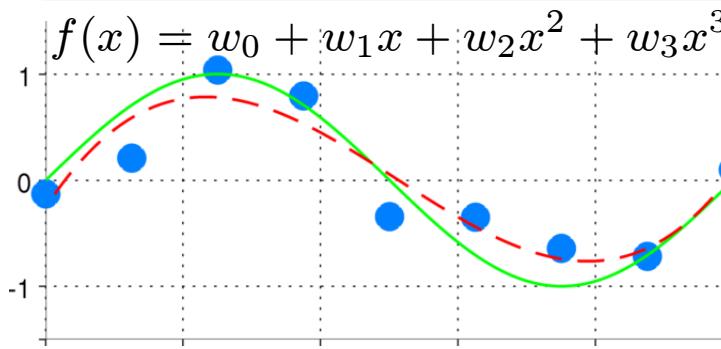
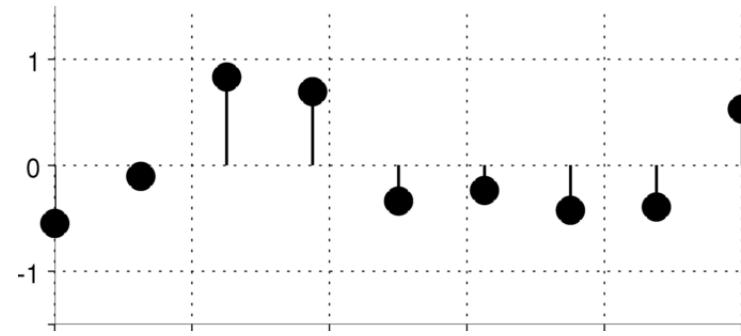
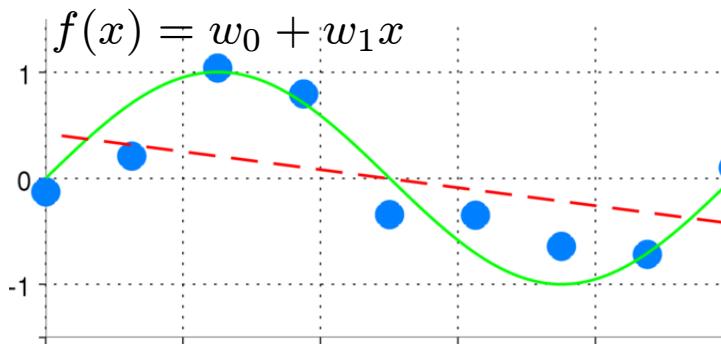
Model

$$f(x) = w_0 + w_1x + w_2x^2 + w_3x^3 + w_4x^4 + w_5x^5 + w_6x^6 + w_7x^7 + w_8x^8$$



Model order

- Which model order
 - Gives the best fit?
 - Do you think is most "correct"?



Estimating parameters

- How do we compute the parameters?
 - Most simple approach: **Minimize cost function over data set**

– **Data** $\{\mathbf{x}_n, y_n\}_{n=1}^N$

– **Model** $f(\mathbf{x}) = \mathbf{x}^\top \mathbf{w}$

– **Cost function** $d(y, f(\mathbf{x}))$

– **Parameters** $\mathbf{w} = \arg \min_{\mathbf{w}} \sum_{n=1}^N d(y_n, f(\mathbf{x}_n))$

Least Squares Regression

- **Cost function:** Squared error

$$d(y, f(\mathbf{x})) = (y - f(\mathbf{x}))^2$$

Model: Linear regression

$$f(\mathbf{x}) = \mathbf{x}^\top \mathbf{w}$$

- **Parameters**

$$\mathbf{w} = \arg \min_{\mathbf{w}} \sum_{n=1}^N d(y_n, f(\mathbf{x}_n)) = \arg \min_{\mathbf{w}} \sum_{n=1}^N (y_n - \mathbf{x}_n^\top \mathbf{w})^2$$

$$\frac{\partial E}{\partial \mathbf{w}} = 2(\mathbf{y} - \mathbf{X}\mathbf{w})^\top \mathbf{X} = 0$$

$$\Rightarrow 2\mathbf{y}^\top \mathbf{X} = 2\mathbf{w}^\top \mathbf{X}^\top \mathbf{X}$$

$$\Rightarrow \mathbf{w} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$

Logistic Regression

(for binary classification, $y \in \{0,1\}$)

- **Cost function:**

negative log of the Binomial distribution

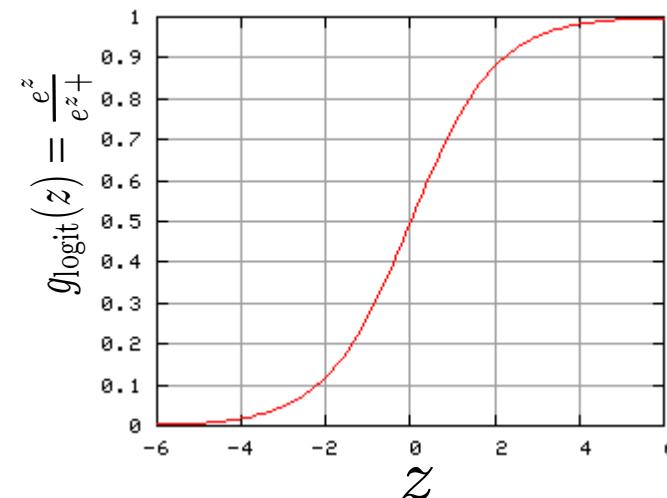
$$d(y, f(\mathbf{x})) = -y \log(f(\mathbf{x})) - (1 - y) \log(1 - f(\mathbf{x})) \quad f(\mathbf{x}) = g_{\text{logit}}(\mathbf{x}^\top \mathbf{w})$$

Model: Logit link function

- **Parameters**

$$\mathbf{w} = \arg \min_{\mathbf{w}} \sum_{n=1}^N d(y_n, f(\mathbf{x}_n))$$

$$g_{\text{logit}}(z) = \frac{e^z}{e^z + 1} = \frac{1}{1 + e^{-z}}$$



Interpretation of $f(x)$: The probability that the observation belongs to class 1

Generalized linear model

- **Cost function:** Choose one

$$d(y, f(\mathbf{x}))$$

Model: Linear + non-linear link

$$f(\mathbf{x}) = g_{\text{link}}(\mathbf{x}^\top \mathbf{w})$$

- **Parameters:** Optimize using numerical optimization methods

$$\mathbf{w} = \arg \min_{\mathbf{w}} \sum_{n=1}^N d(y_n, f(\mathbf{x}_n))$$

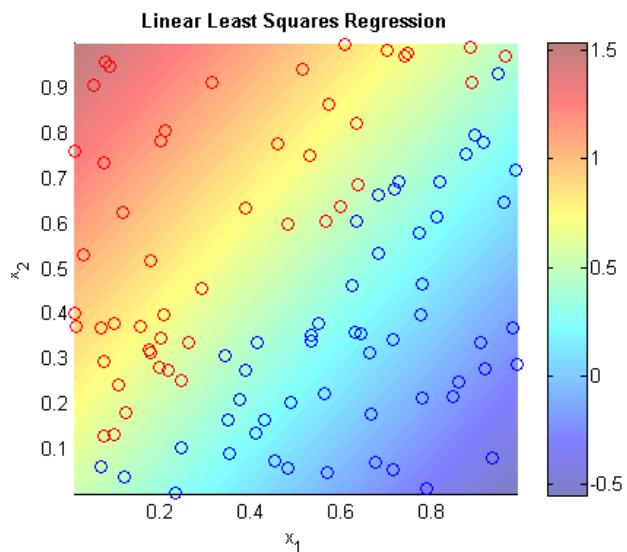
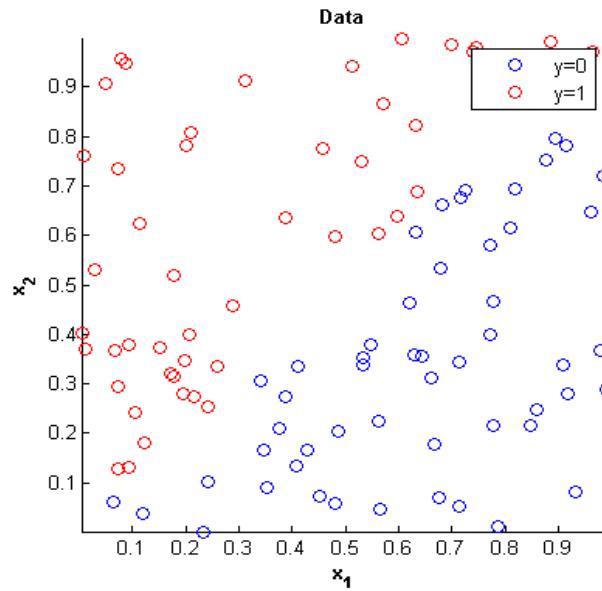


Matlab: `glmfit`
Python: `sklearn.linear_model`
R: `glm`

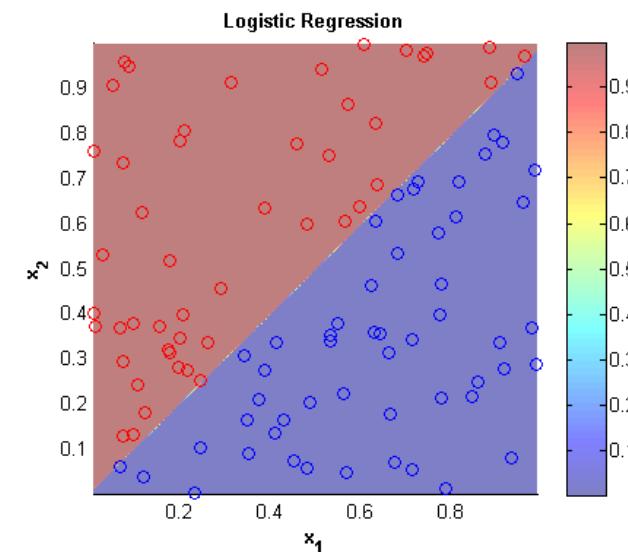
$$g_{\text{identity}}(z) = z$$

$$g_{\text{logit}}(z) = \frac{e^z}{e^z + 1} = \frac{1}{1 + e^{-z}}$$

Linear vs. Logistic Regression for a classification problem



$$f(x) = 0.5463 - 1.1081x_1 + 1.0009x_2$$



$$f(x) = 1/[1+\exp(-(14.5-3316x_1+3352x_2))]$$

02450 Introduction to machine learning and data modeling

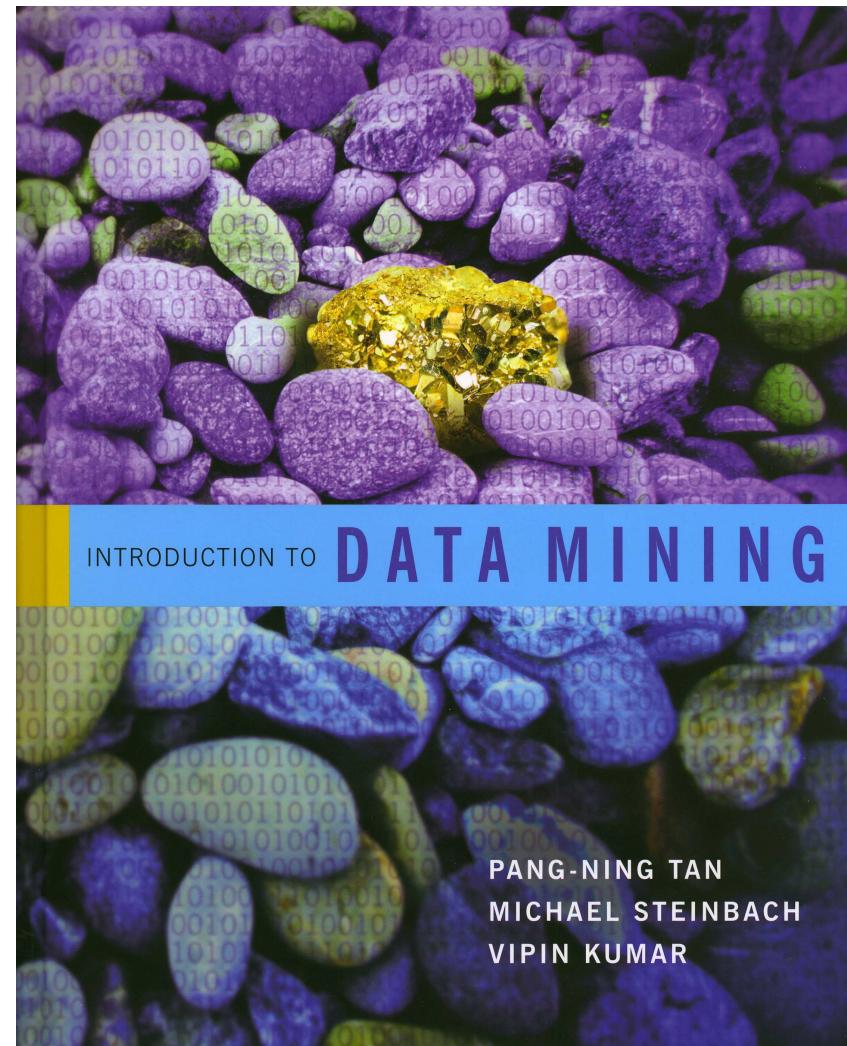
Reading material

Tan, Steinbach and Kumar
"Introduction to Data Mining"

Section 4.4-4.6

Group of the day

Jesper Plantener
Jacob Bøgelund Hansen
Tobias Brasch
Erik W. Rasmussen
Steffen Karlsson
Christian Graver Larsen
Erik Derner
Jeppe Kristensen Bloch
Patrick Bach Andersen



Lecture schedule

1. Introduction
(Tan 1.1-1.4)

Data: Feature extraction and visualization

2. Data and feature extraction
(Tan 2.1-2.3 + B1 (+ A))
3. Measures of similarity and summary statistics
(Tan 2.4 + 3.1-3.2 + C1-C2)
4. Data visualization
(Tan 3.3)

Supervised learning: Classification and regression

5. Decision trees and linear regression
(Tan 4.1-4.3 + D)
6. **Overfitting and performance evaluation**
(Tan 4.4-4.6)
7. Nearest neighbor, naive Bayes, and artificial neural networks
(Tan 5.2-5.4)

8. Ensemble methods and multi class classifiers
(Tan 5.6-5.8)

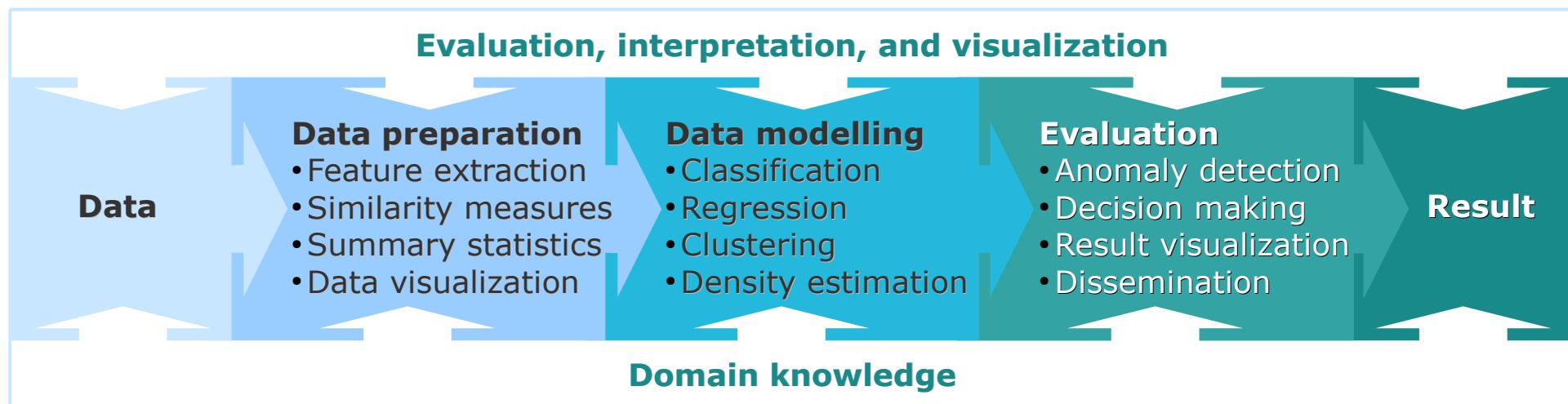
Unsupervised learning: Clustering and density est.

9. K-means and hierarchical clustering
(Tan 8.1-8.3+8.5.7)
10. Mixture models and association mining
(Tan 9.2.2 + 6.1-6.3)
11. Density estimation and anomaly detection
(Tan 10.1-10.4)

Machine learning and data modelling in practice

12. Putting it all together: Summary and overview
13. Mini project

Data modeling framework



After today you should be able to:

Explain the difference between training and test (generalization) error

Explain how cross-validation can be used for model selection

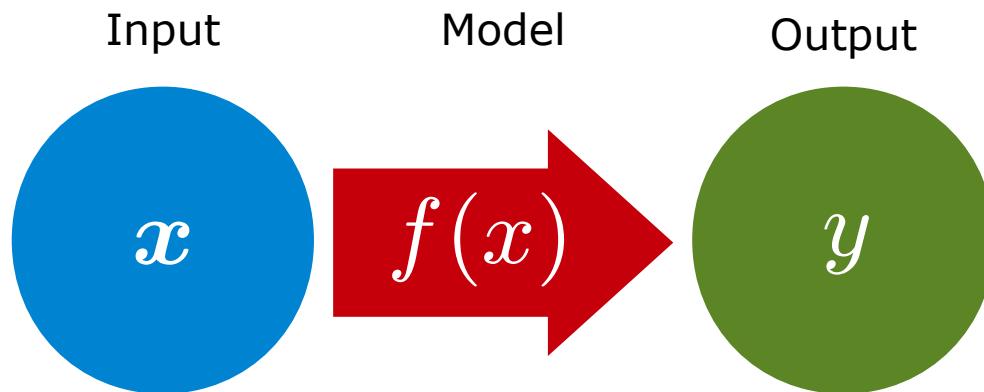
Apply forward and backward selection

Prune decision trees

Understand the Bias-Variance tradeoff as illustrated for regularized least squares estimation

Test the significance of classifiers

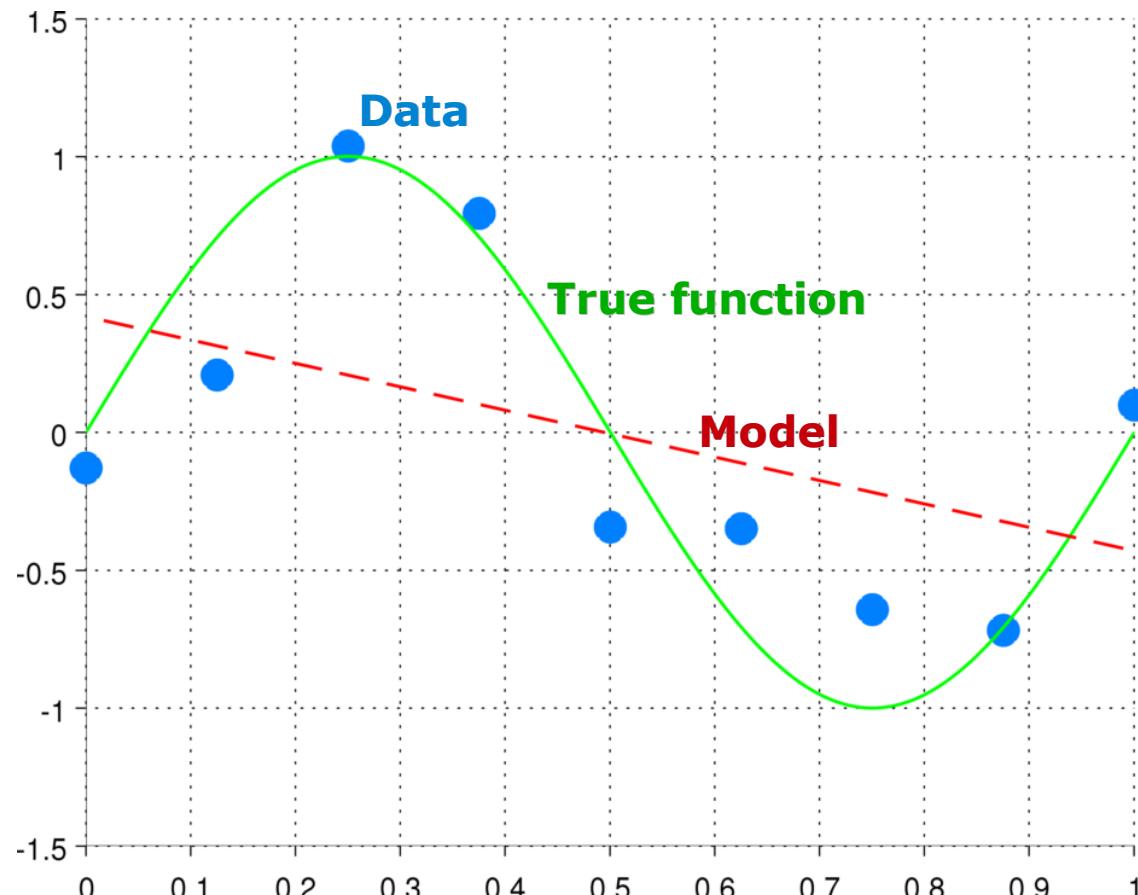
Supervised learning



- **Mapping between domains**
 - Classification: Discrete output
 - Regression: Continuous output

Linear regression

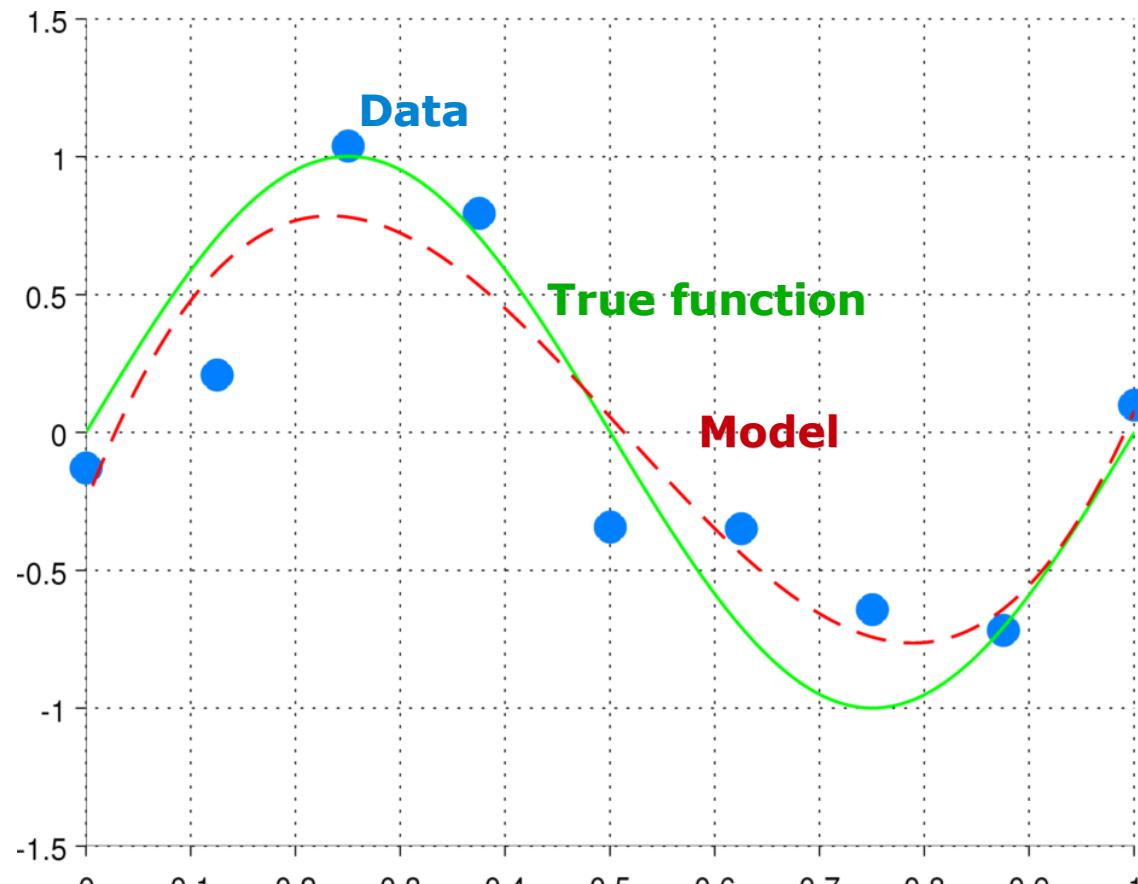
- Bad fit
- **Too simple model**



$$f(x) = w_0 + w_1 x$$

Linear regression

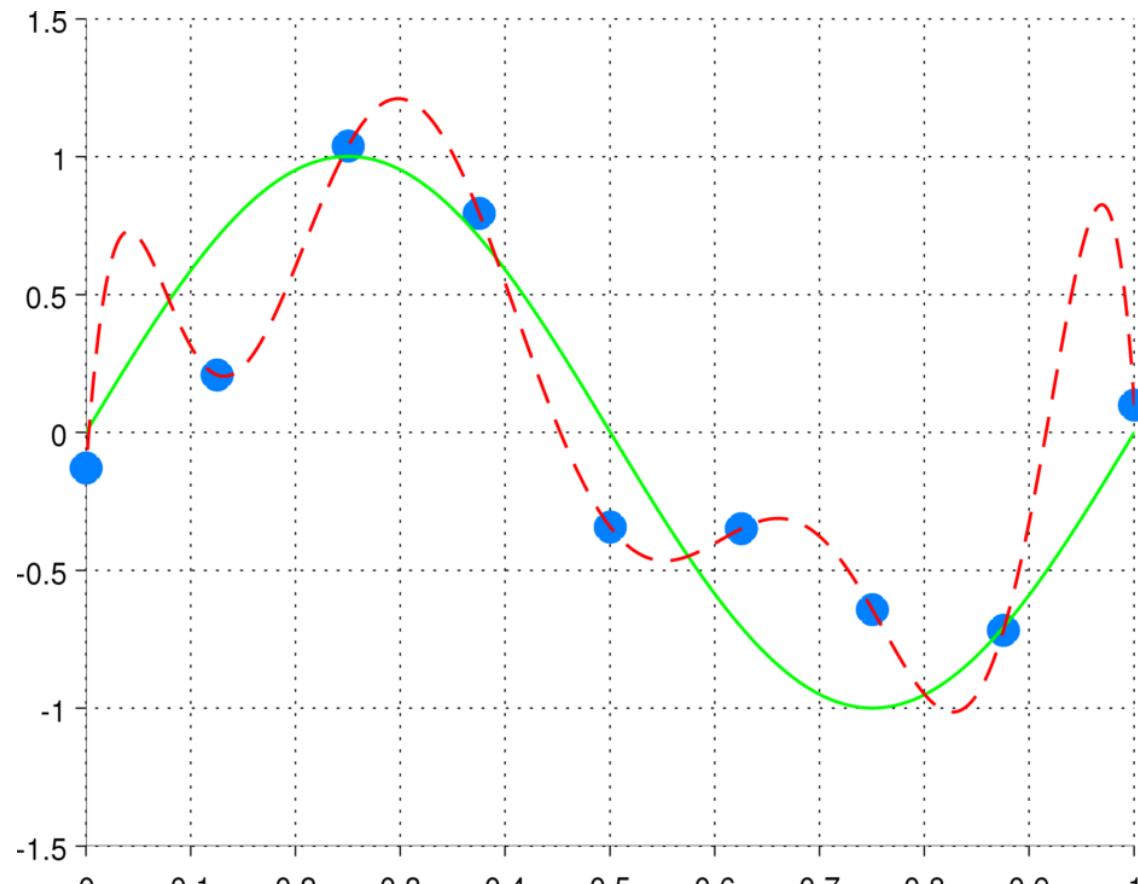
- Reasonable fit
- **Reasonable model**



$$f(x) = w_0 + w_1x + w_2x^2 + w_3x^3$$

Linear regression

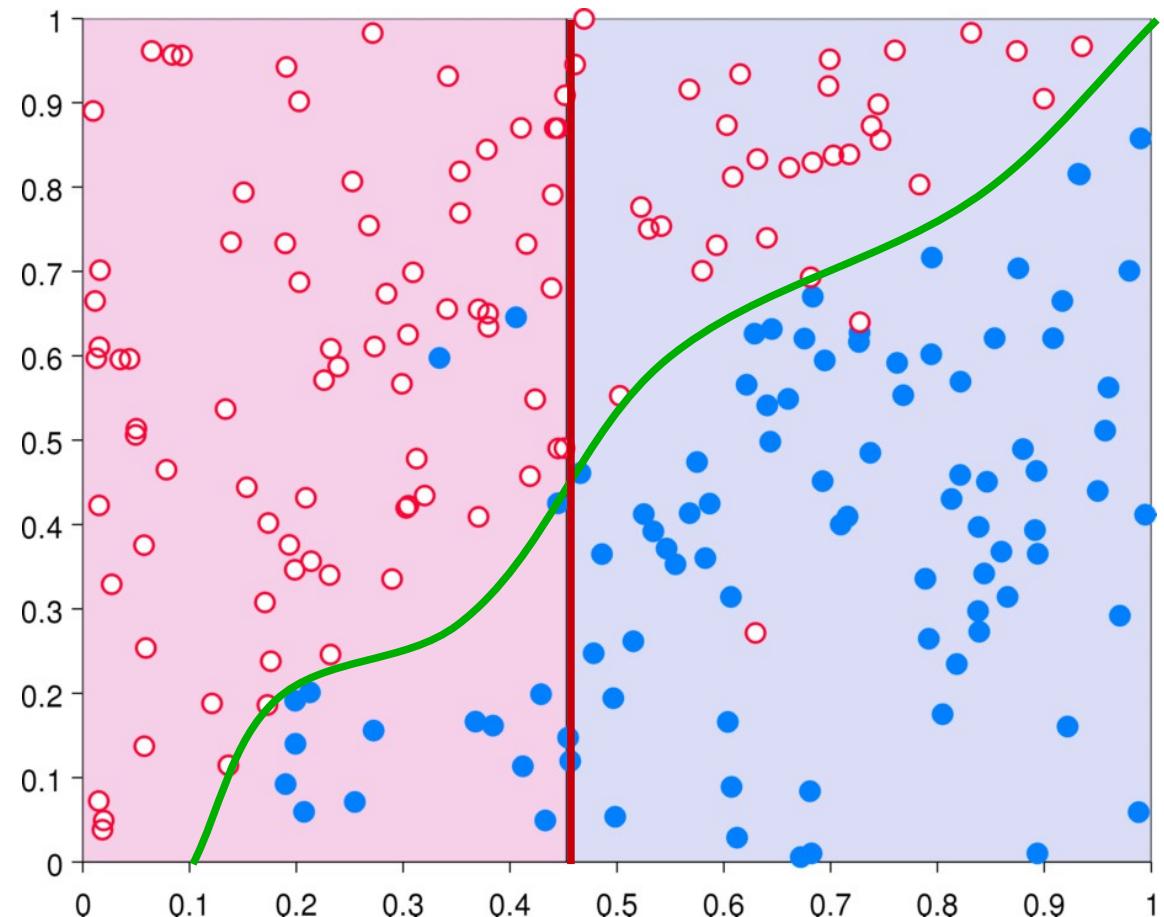
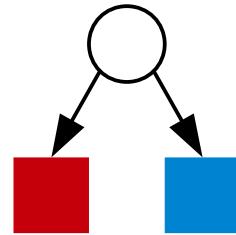
- Perfect fit
- **Too complex model**



$$f(x) = w_0 + w_1x + \cdots + w_8x^8$$

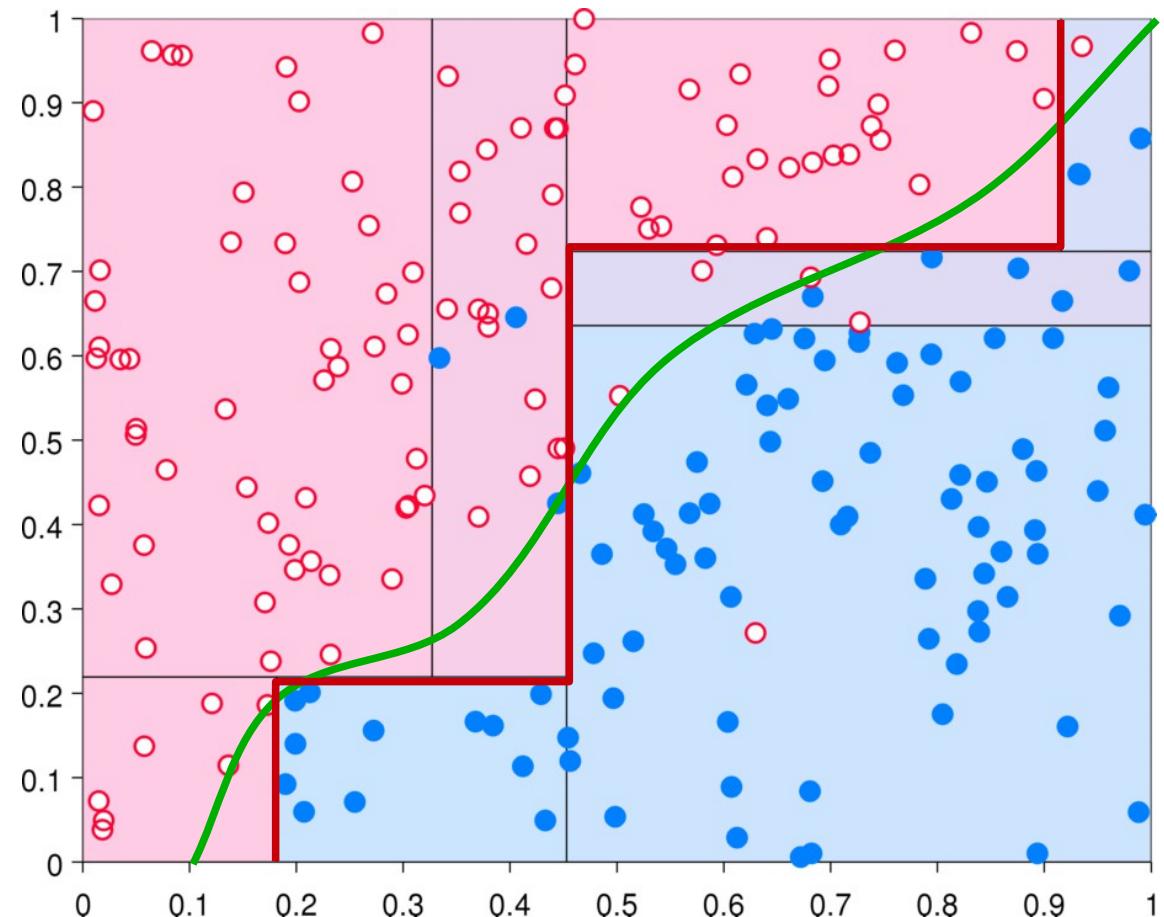
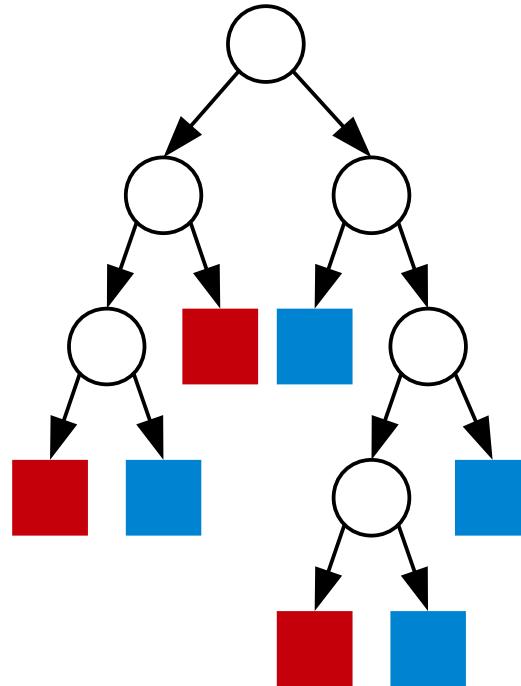
Regression trees

- Bad fit
- **Too simple model**



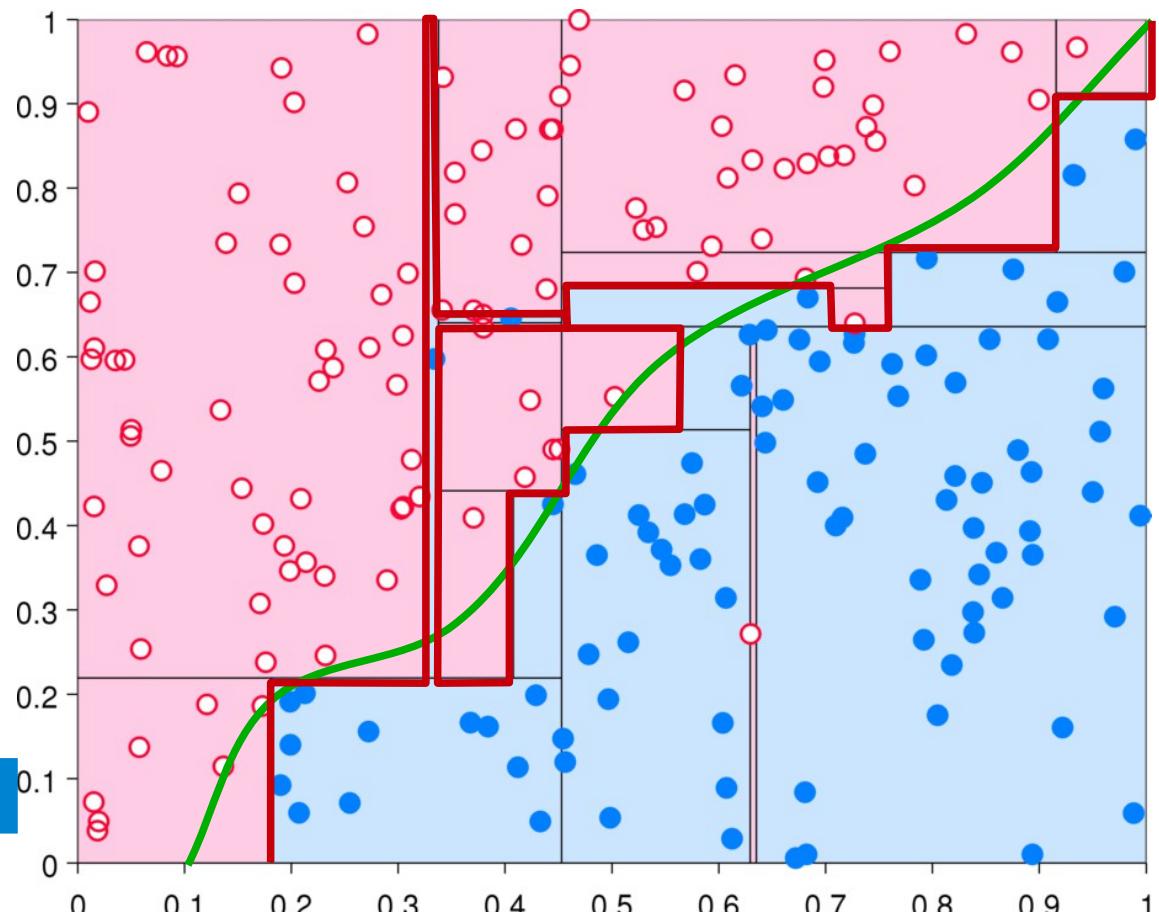
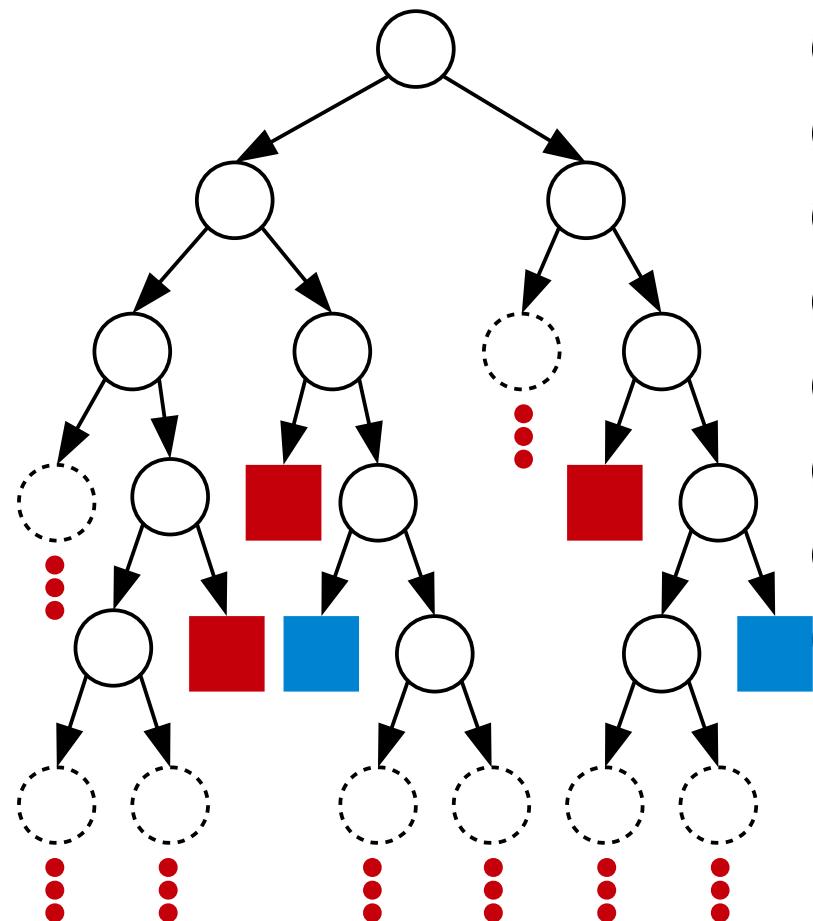
Regression trees

- Reasonable fit
- **Reasonable model**

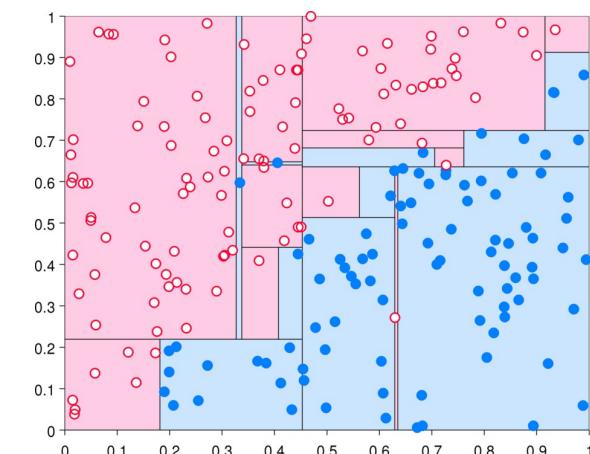
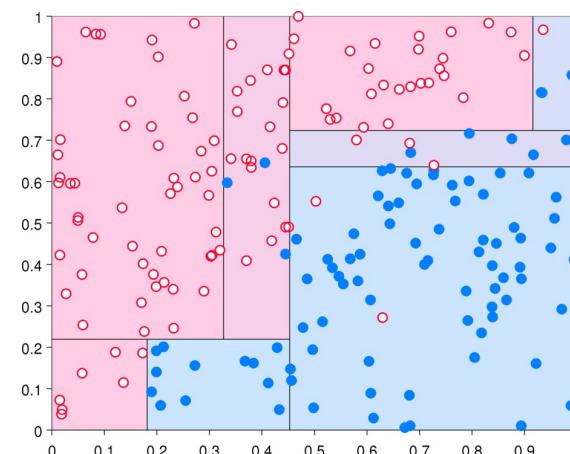
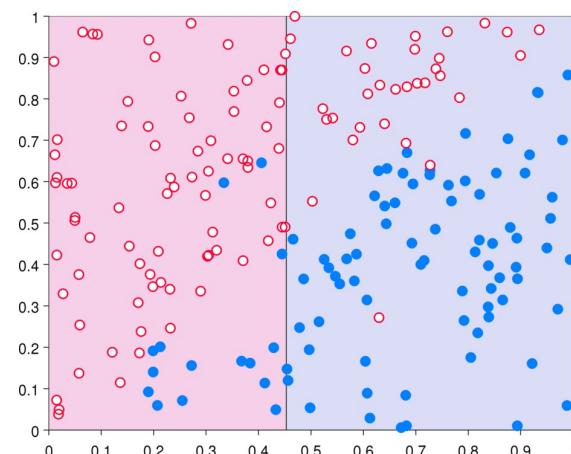
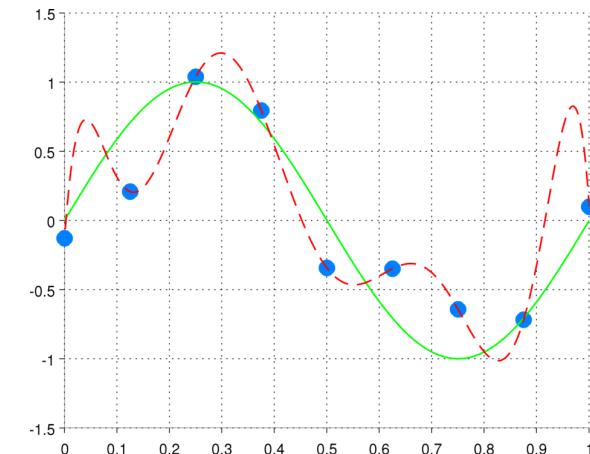
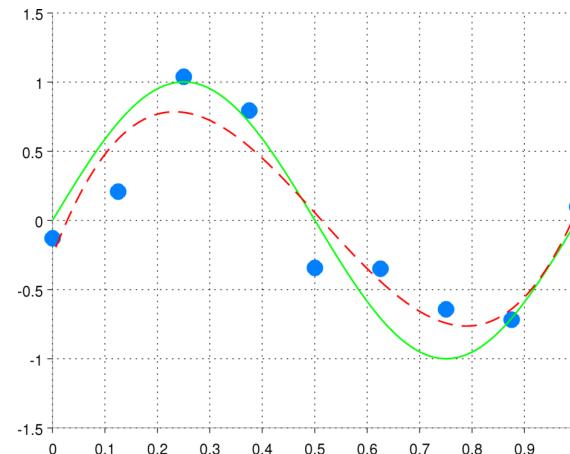
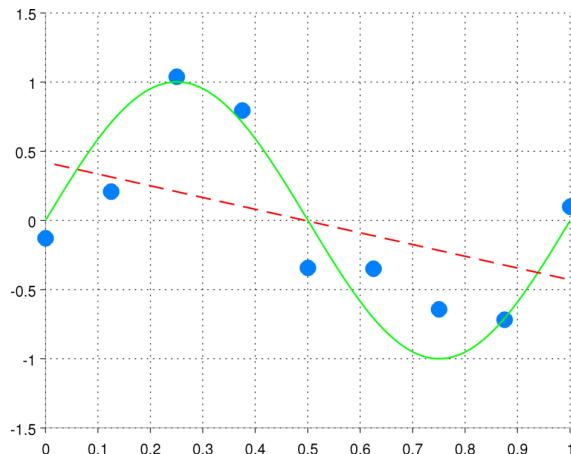


Regression trees

- Perfect fit
 - **Too complex model**

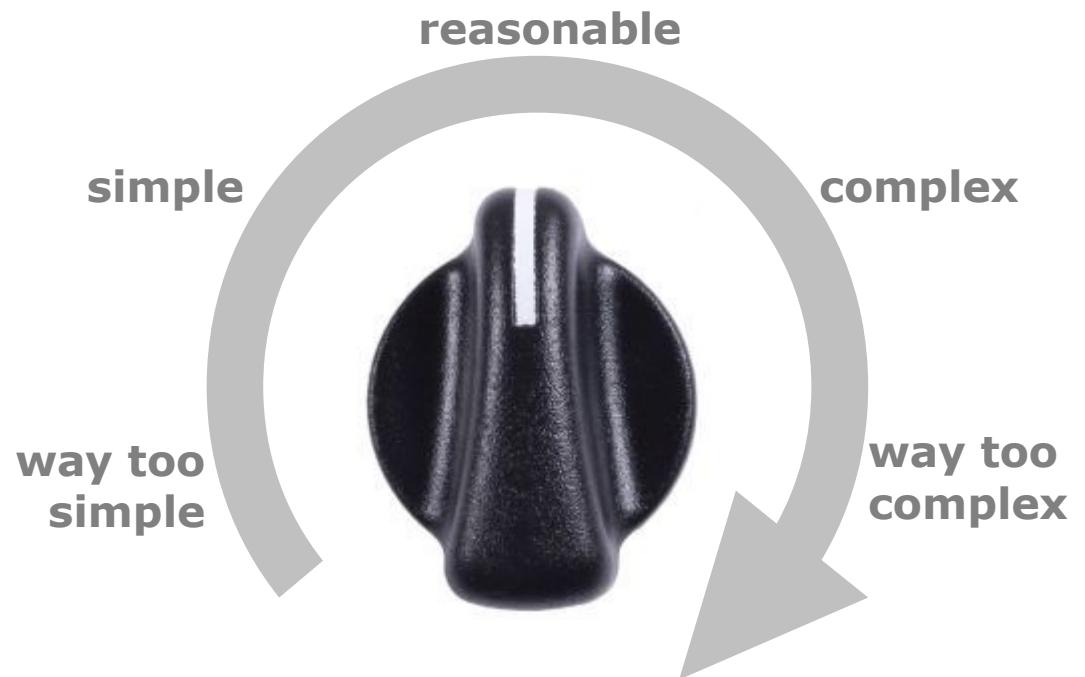


Model overfitting



Control the model complexity

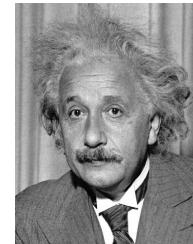
- Find **parameter** or **mechanism** in model that controls complexity



Lex Parsimoniae, Law of parsimony



"Given two models with same predictive performance, the simpler model is preferred over the more complex model" - William of Occam



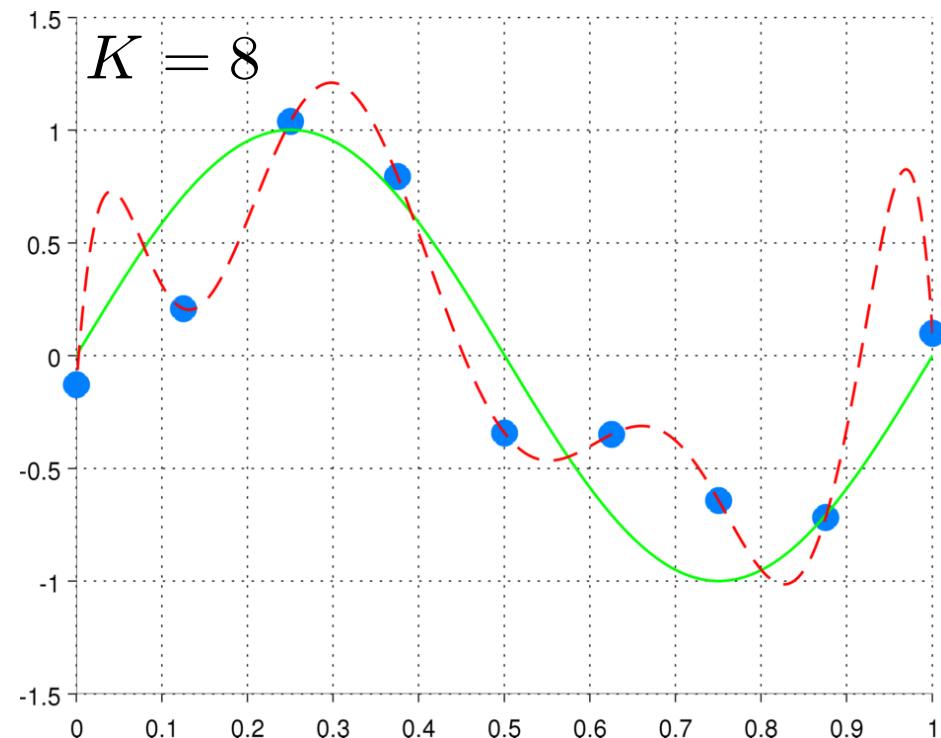
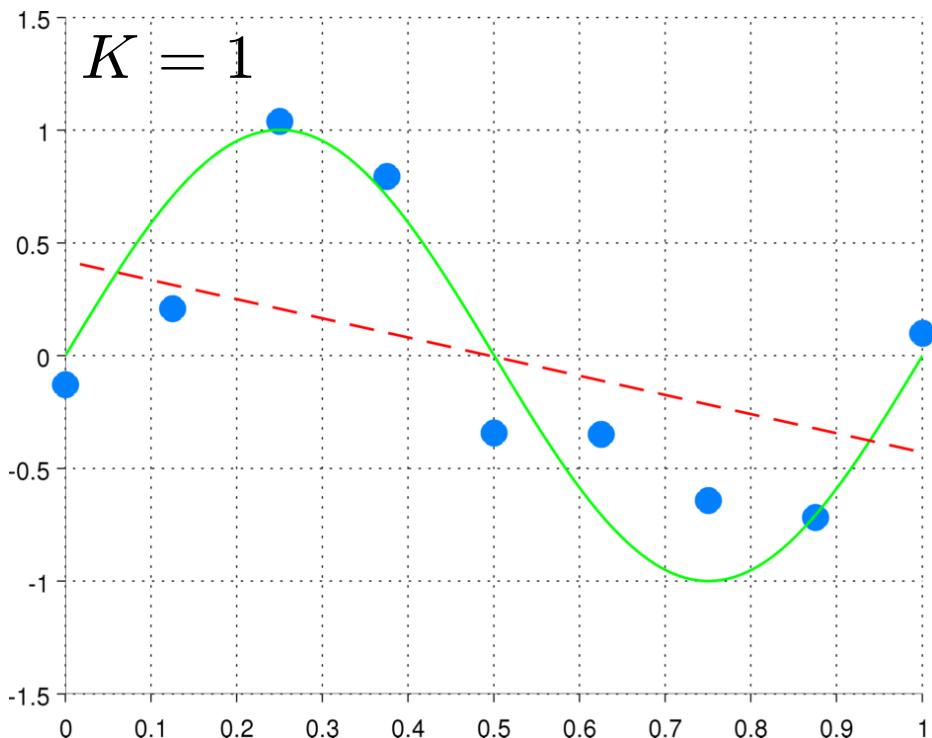
"Everything should be made as simple as possible, but not simpler" - Einstein

Linear regression

- Linear regression on non-linearly transformed inputs (polynomials)

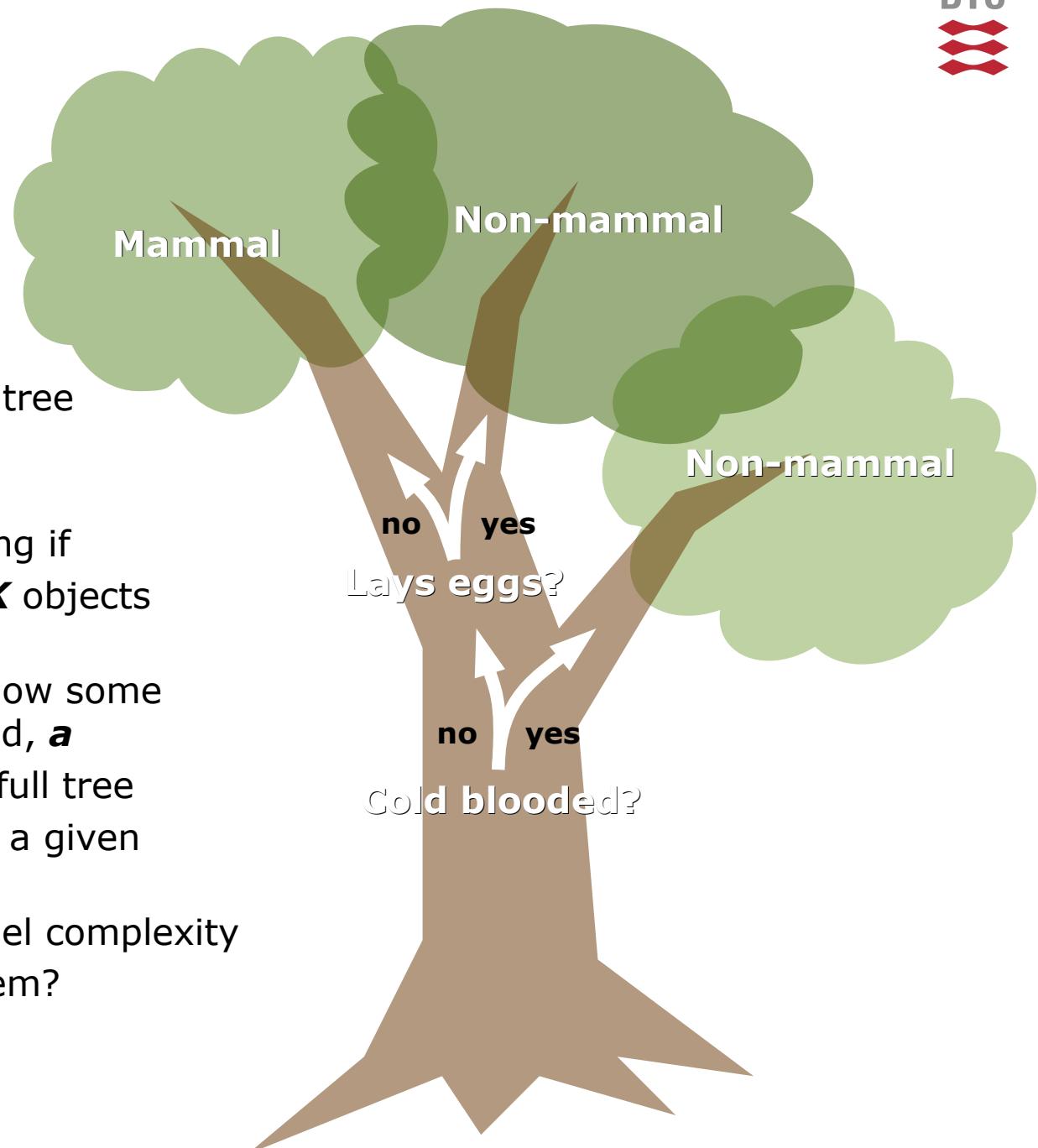
$$f(x) = w_0 + w_1x + w_2x^2 + \cdots + w_Kx^K$$

– **Control complexity:** Choose a suitable value for K



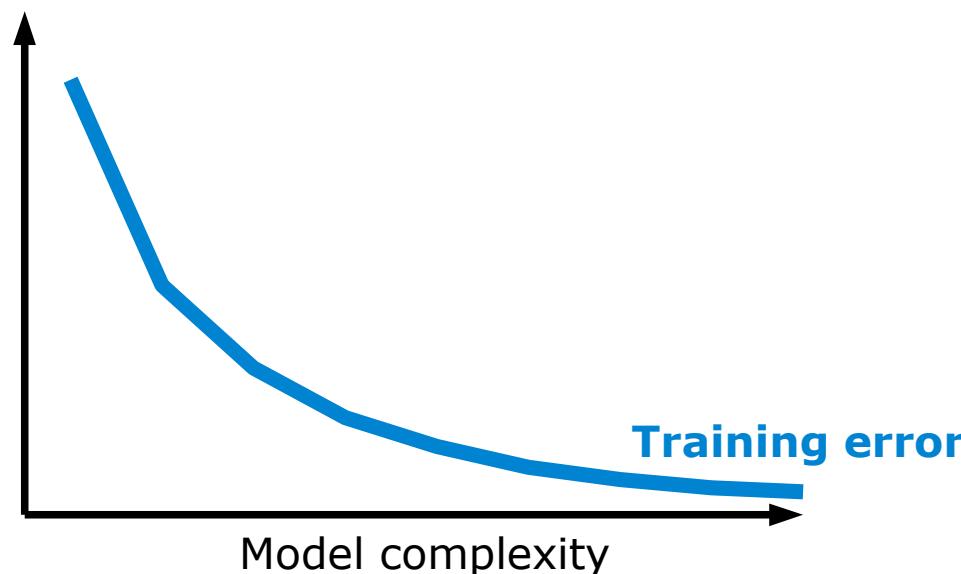
Decision trees

- Hunt's algorithm
 - Continue splitting until each node is pure
 - Results in a very complex tree (overfitting)
- **Control complexity**
 - **Pre-pruning:** Stop splitting if
 - There is less than K objects on the branch
 - Impurity gain is below some predefined threshold, a
 - **Post-pruning:** Generate full tree
 - Cut off branches to a given pruning level, c
- K , a , and/or c determine model complexity
 - How should we choose them?



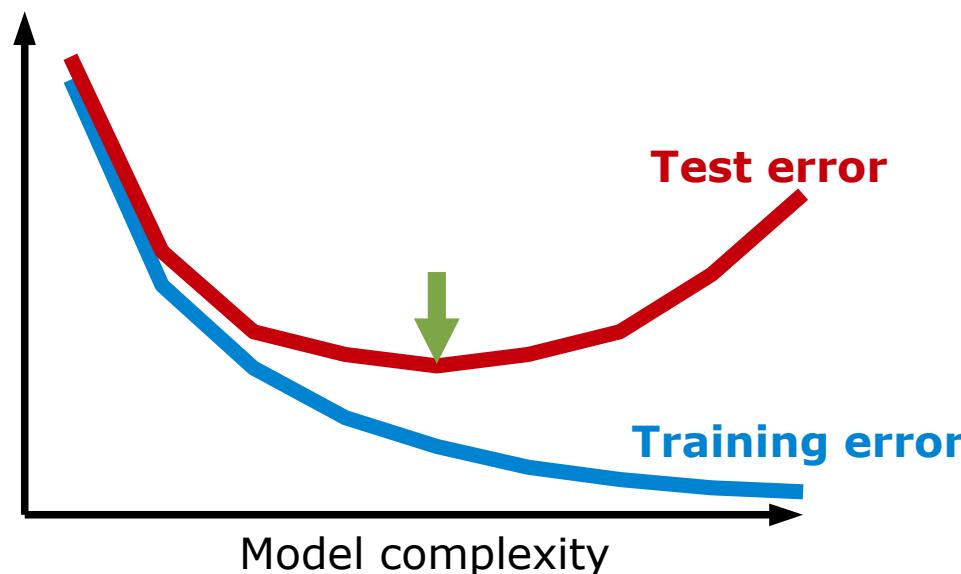
Cross-validation

- Partition the data into
 - **Training** and **test** set
- Using the **training set**
 - Train the model for a range of different complexities
 - From too simple to too complex
 - Compute the **training error**
 - It will always be lowest for the most complex model



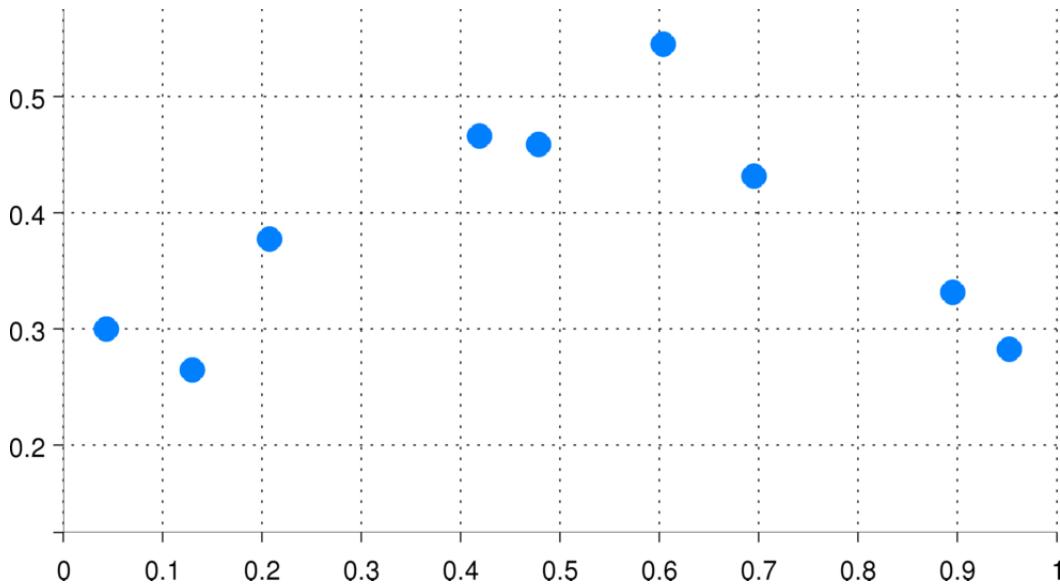
Cross-validation

- Using the **test set**
 - For each model complexity, compute the **test error**
 - Choose the model complexity that gives the lowest test error
(The **generalization error** is the expected value of the the test error.)



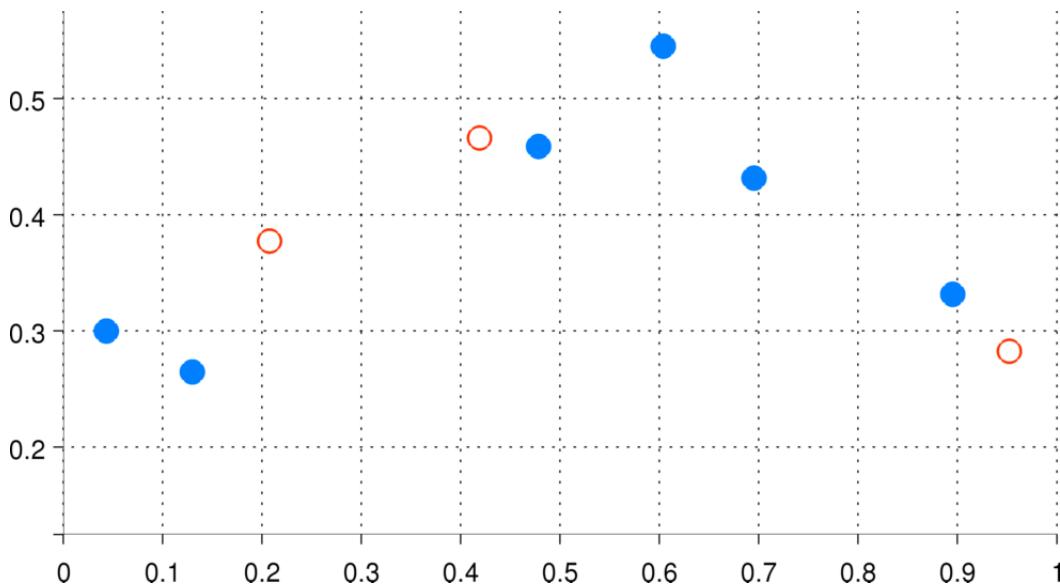
Holdout method

- Randomly choose a subset of data points to be in a **test set**
 - For example choose 1/3 of the points
- The rest is the **training set**



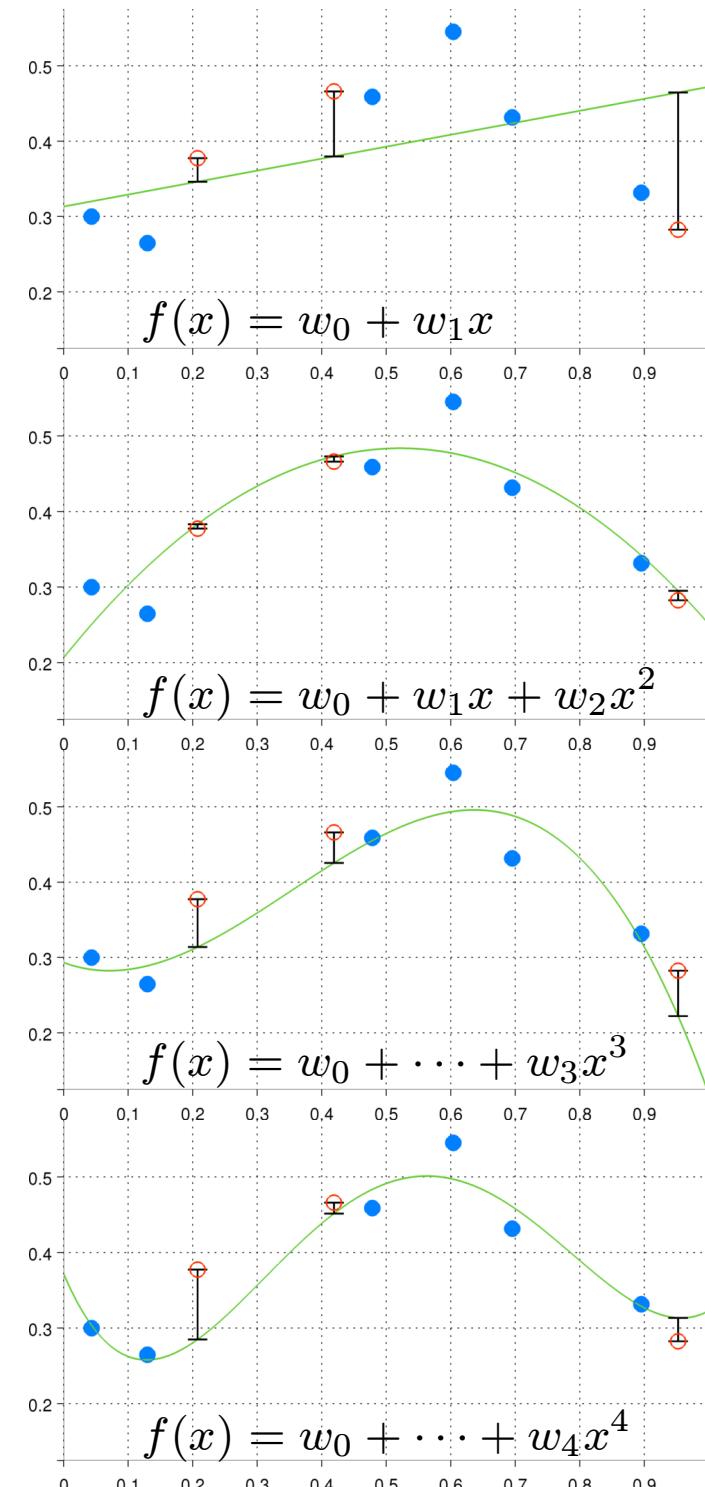
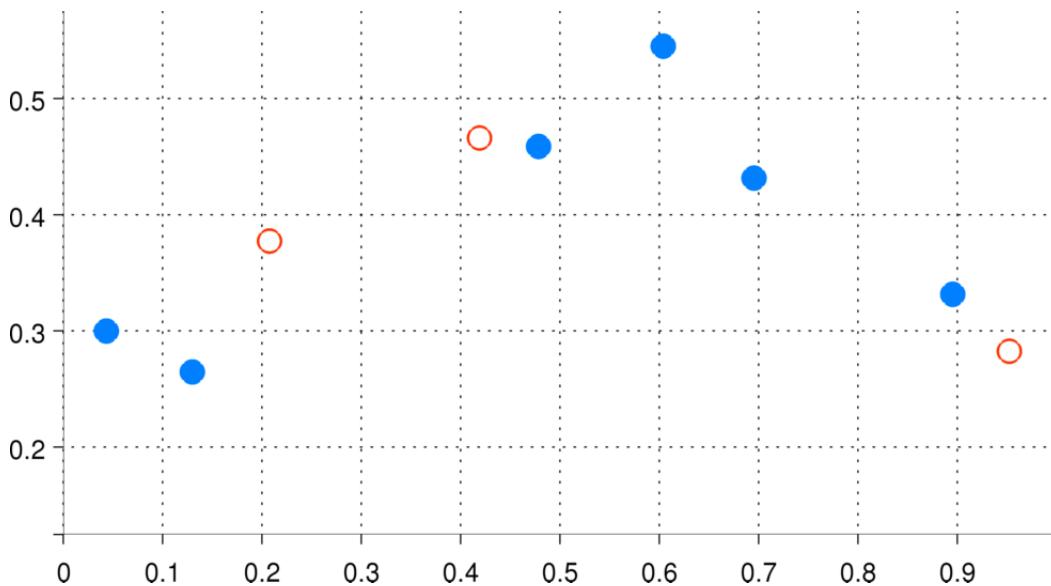
Holdout method

- Randomly choose a subset of data point to be in a **test set**
 - For example choose 1/3 of the points
- The rest is the **training set**



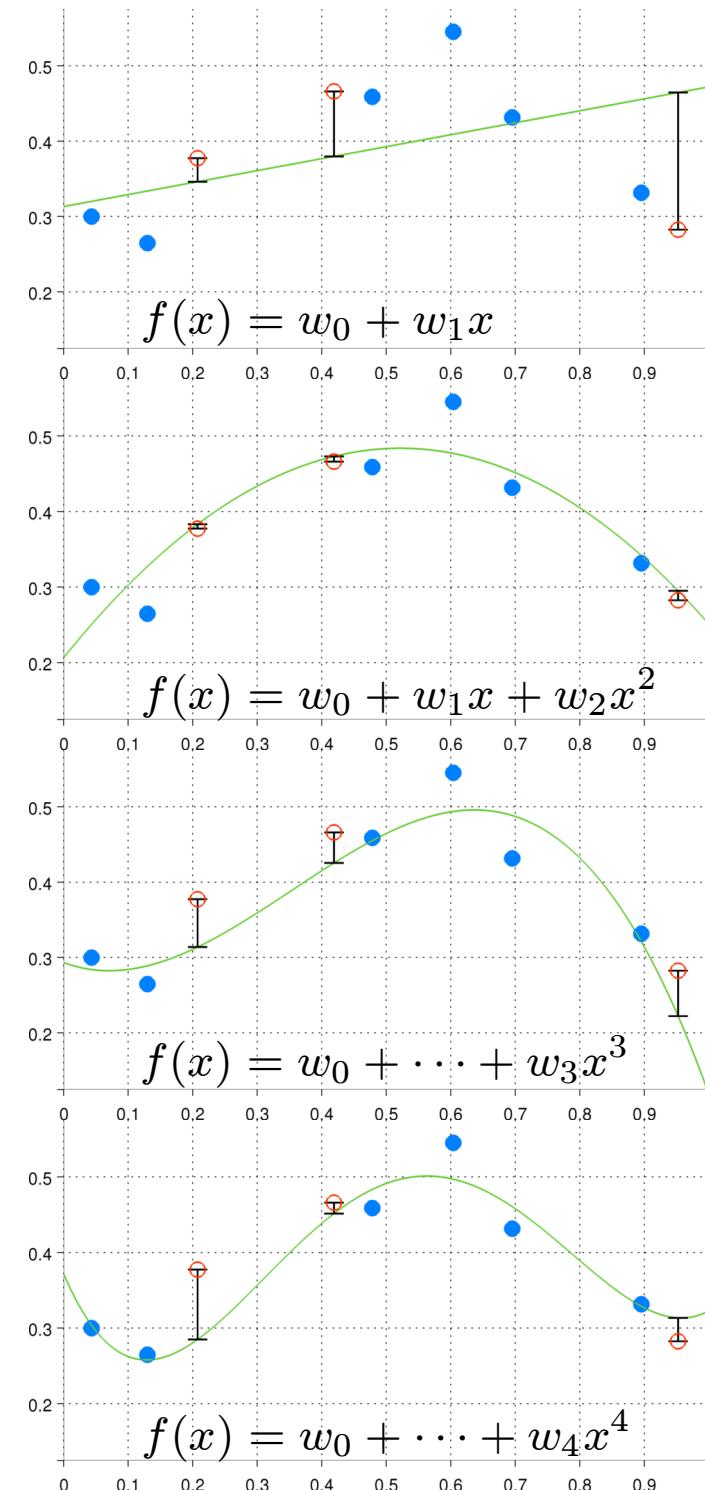
Holdout method

- Using the **training set**
 - Train the model for different complexities
- Using the **test set**
 - Compute the test error
- Choose the model with lowest **test error**



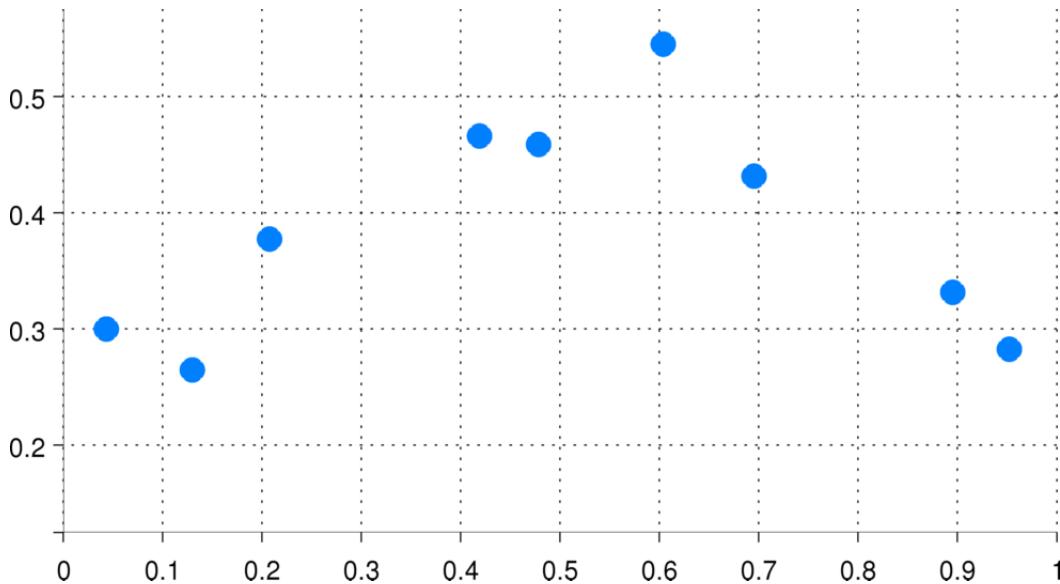
Holdout method

- Using the **training set**
 - Train the model for different complexities
- Using the **test set**
 - Compute the test error
- Choose the model with lowest **test error**



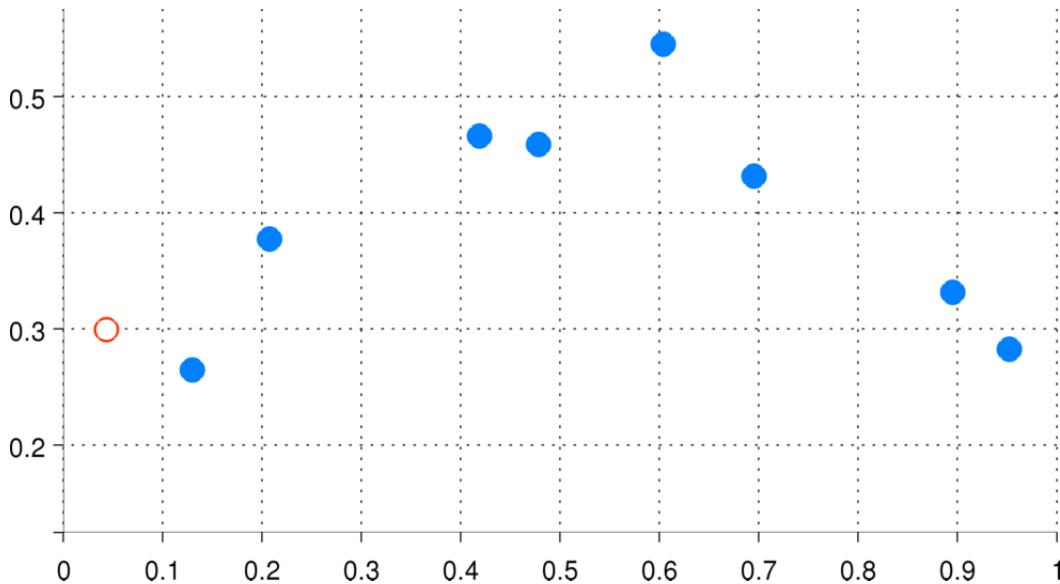
Leave-one-out

- Choose the first data point as a **test set**
- The rest is the **training set**



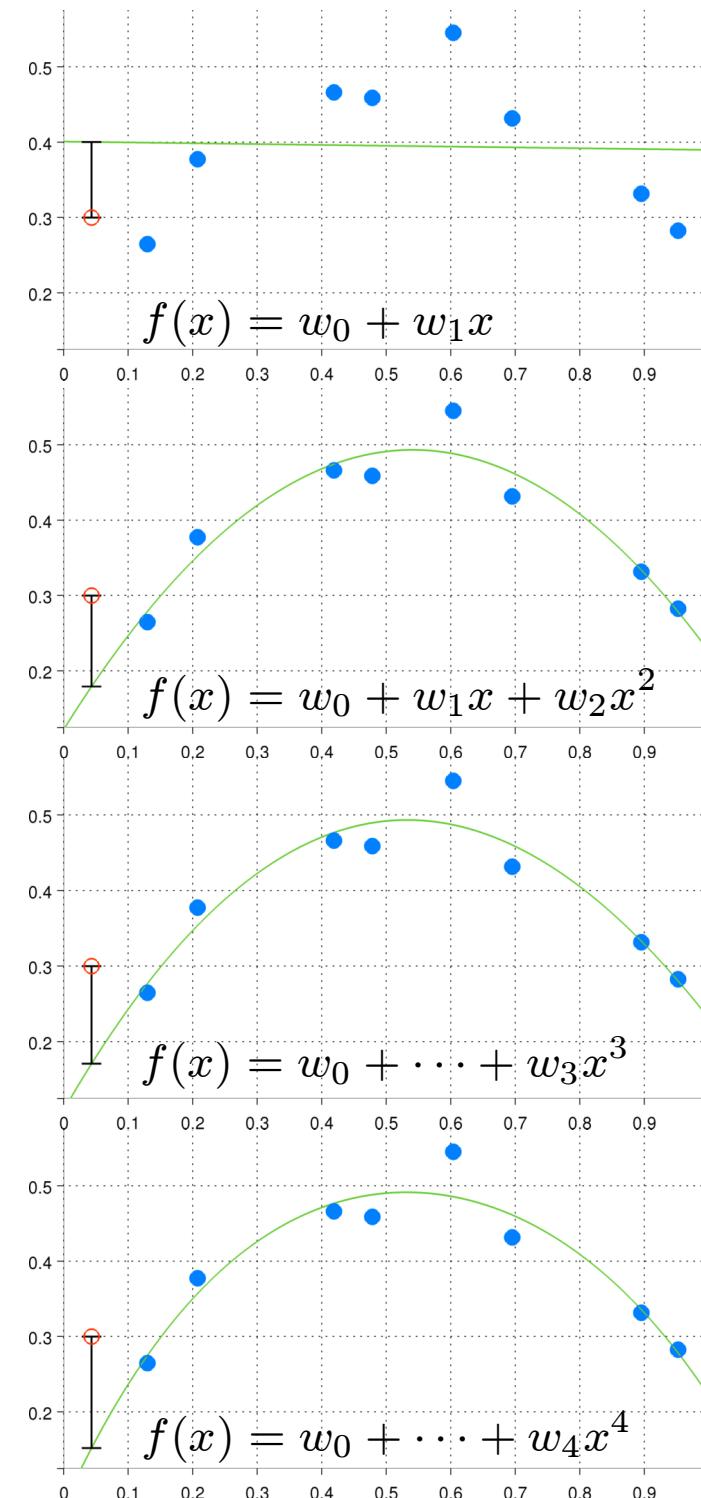
Leave-one-out

- Choose the first data point as a **test set**
- The rest is the **training set**

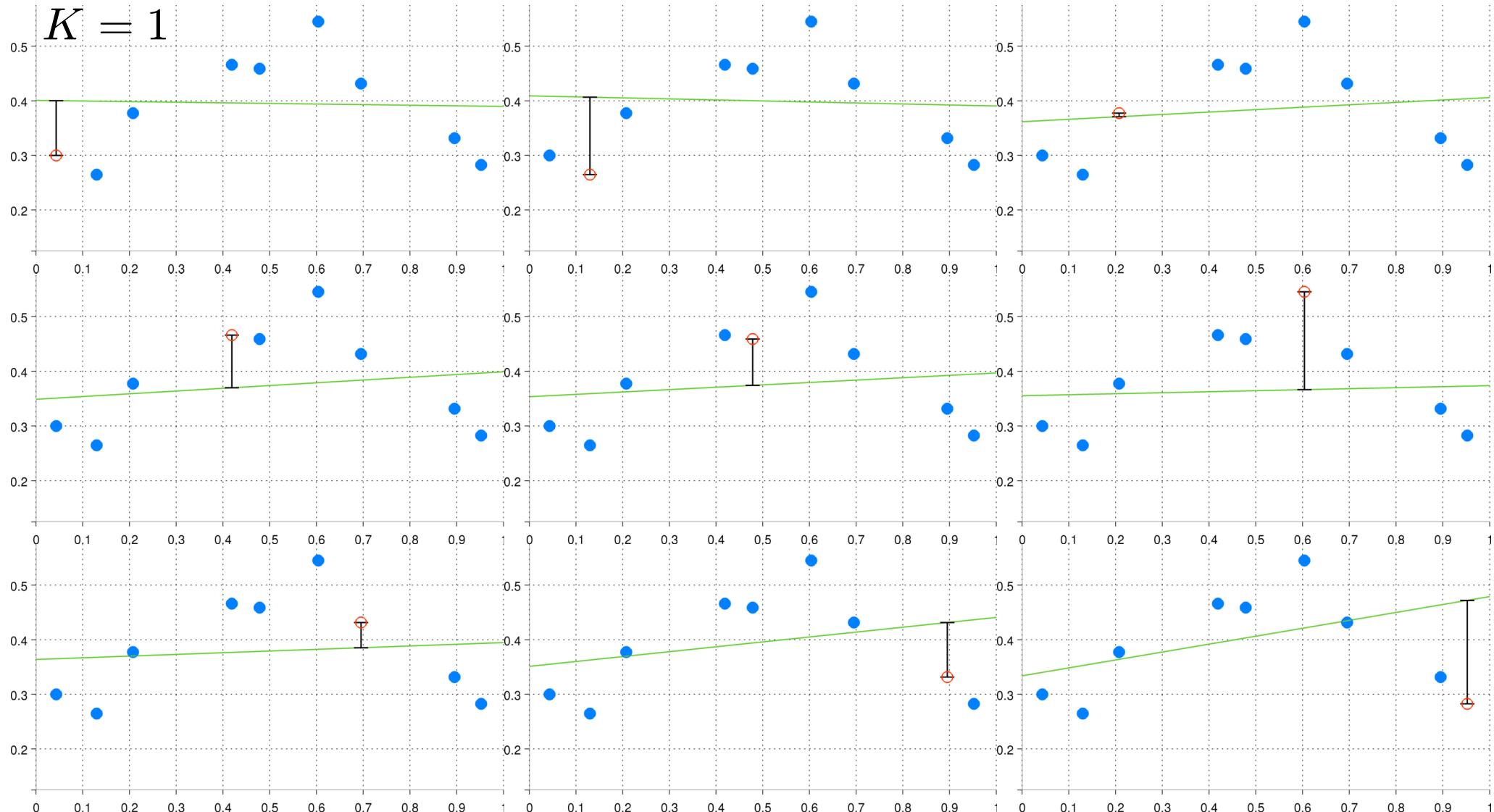


Leave-one-out

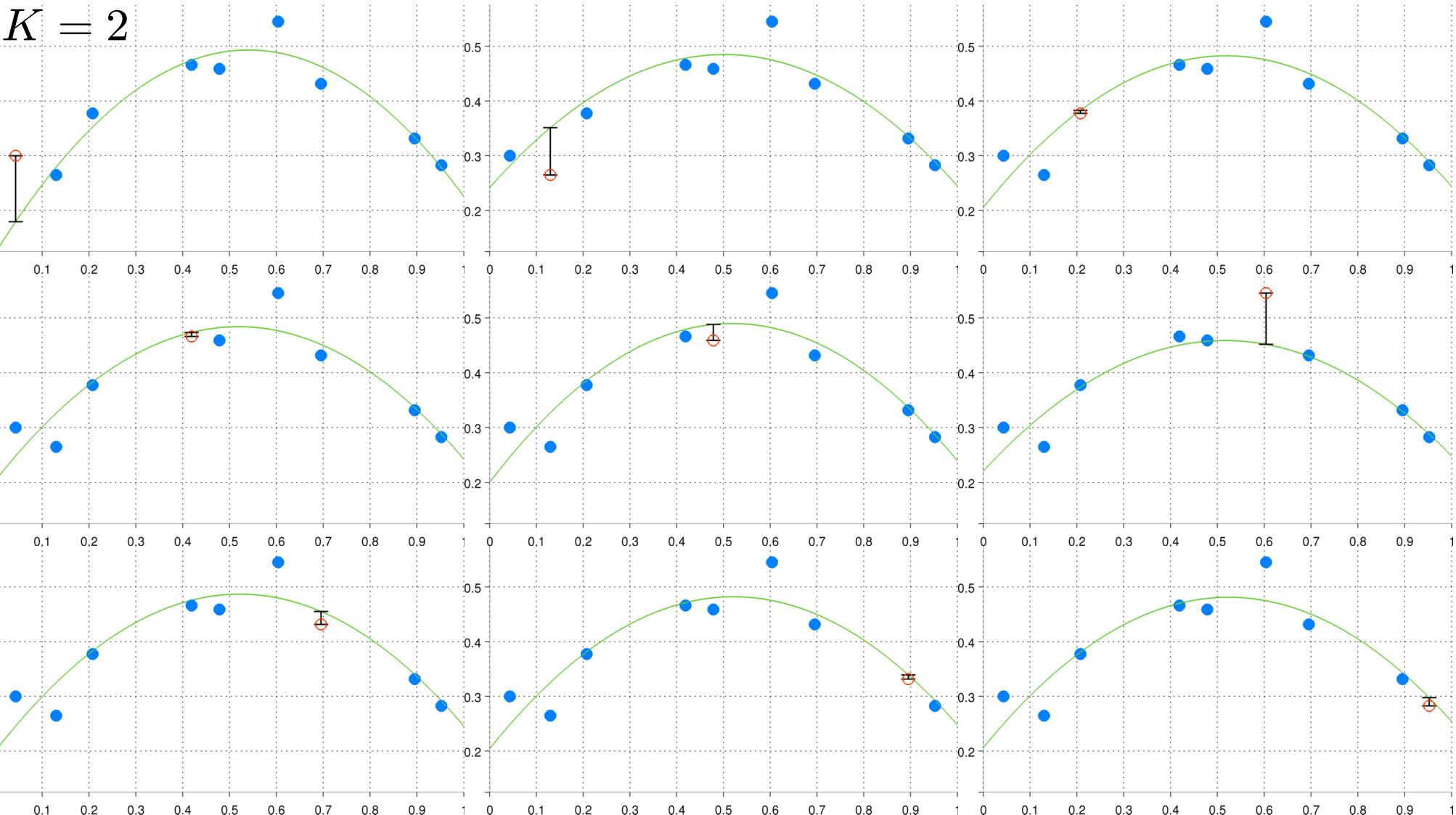
- Using the **training set**
 - Train the model for different complexities
- Using the **test set**
 - Compute the test error
- **Repeat for all data points**
 - All data points get to be test set
 - Compute **average test error**



Leave-one-out

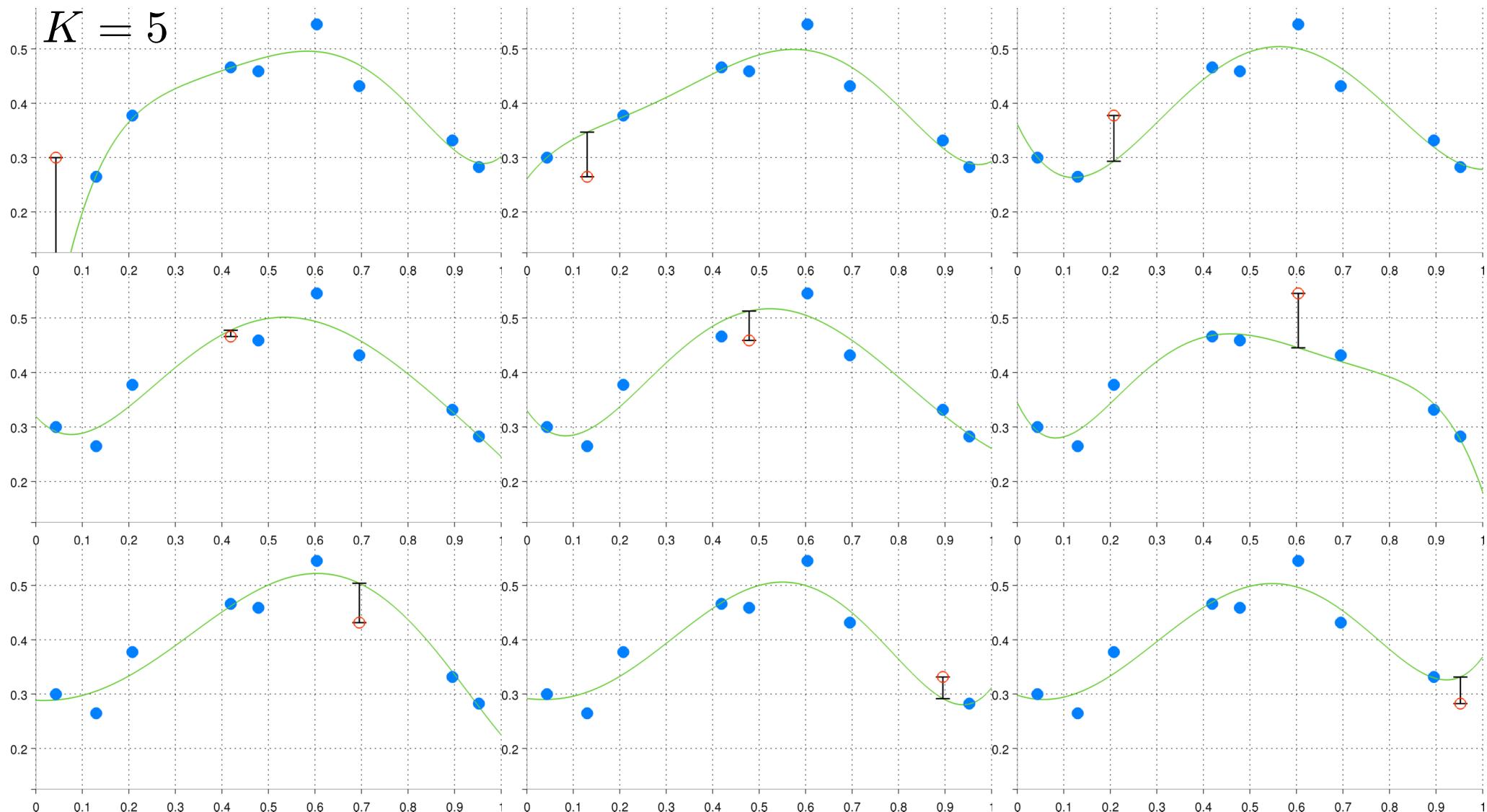


Leave-one-out



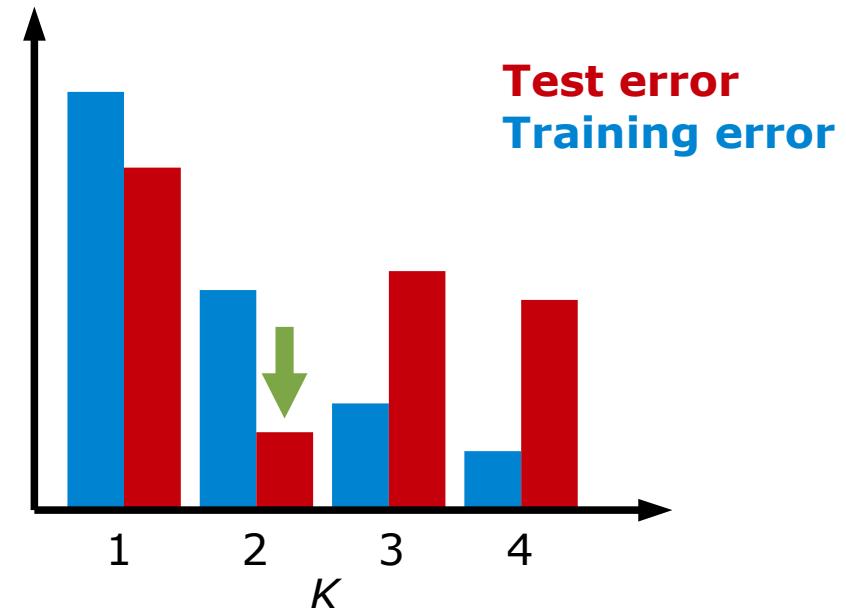
Leave-one-out cross-validation

$$K = 5$$



Leave-one-out

- Using the **training set**
 - Train the model for different complexities
- Using the **test set**
 - Compute the test error
- **Repeat for all data points**
 - All data points get to be test set
 - Compute **average test error**

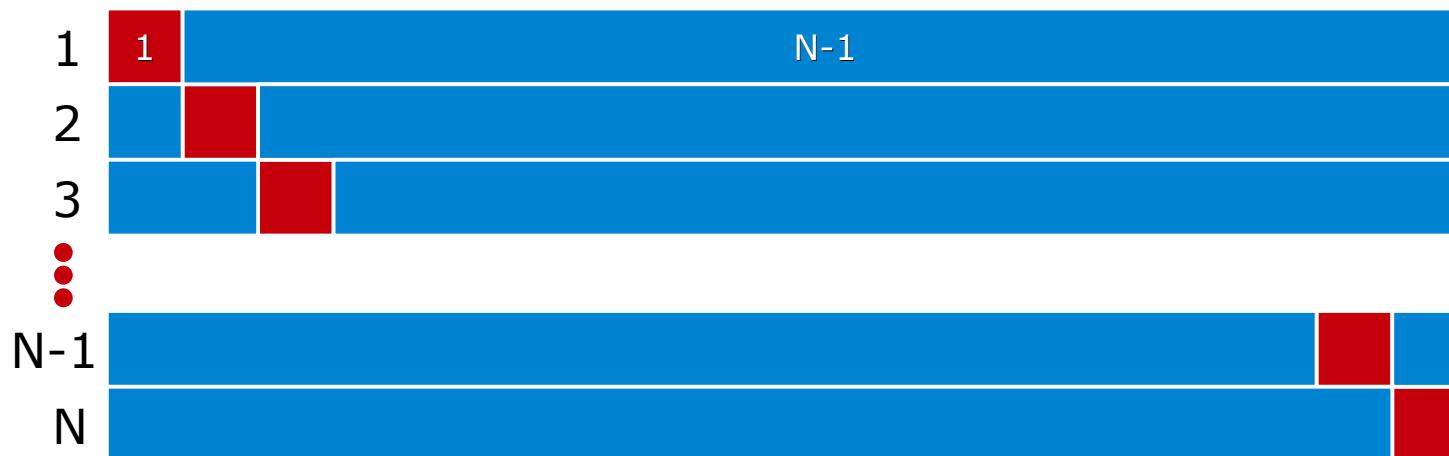


Cross-validation methods

Holdout method

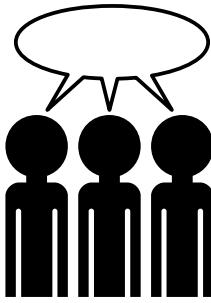


Leave-one-out



K-fold cross-validation (3-fold)





Cross-validation methods

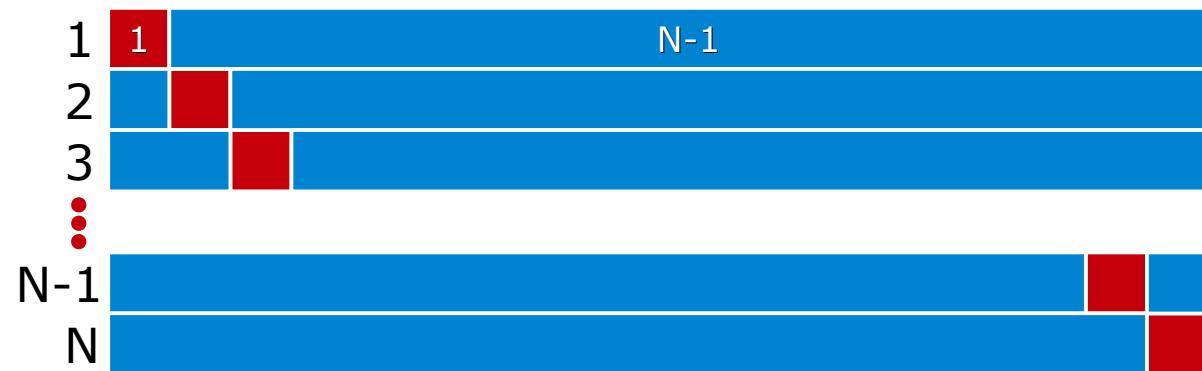
- Compare these three methods
 - What are their pros and cons?
- 10-fold cross-validation is very often used in practice
 - Why do you think?

Holdout method

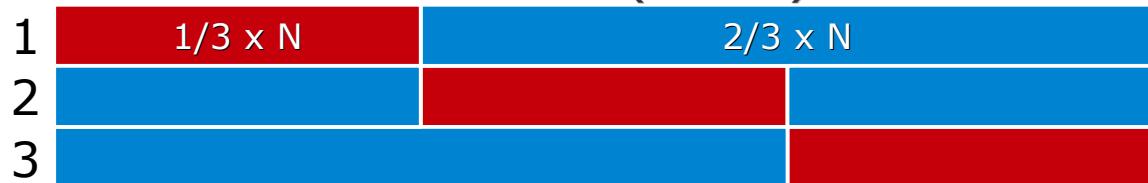


**Test
Training**

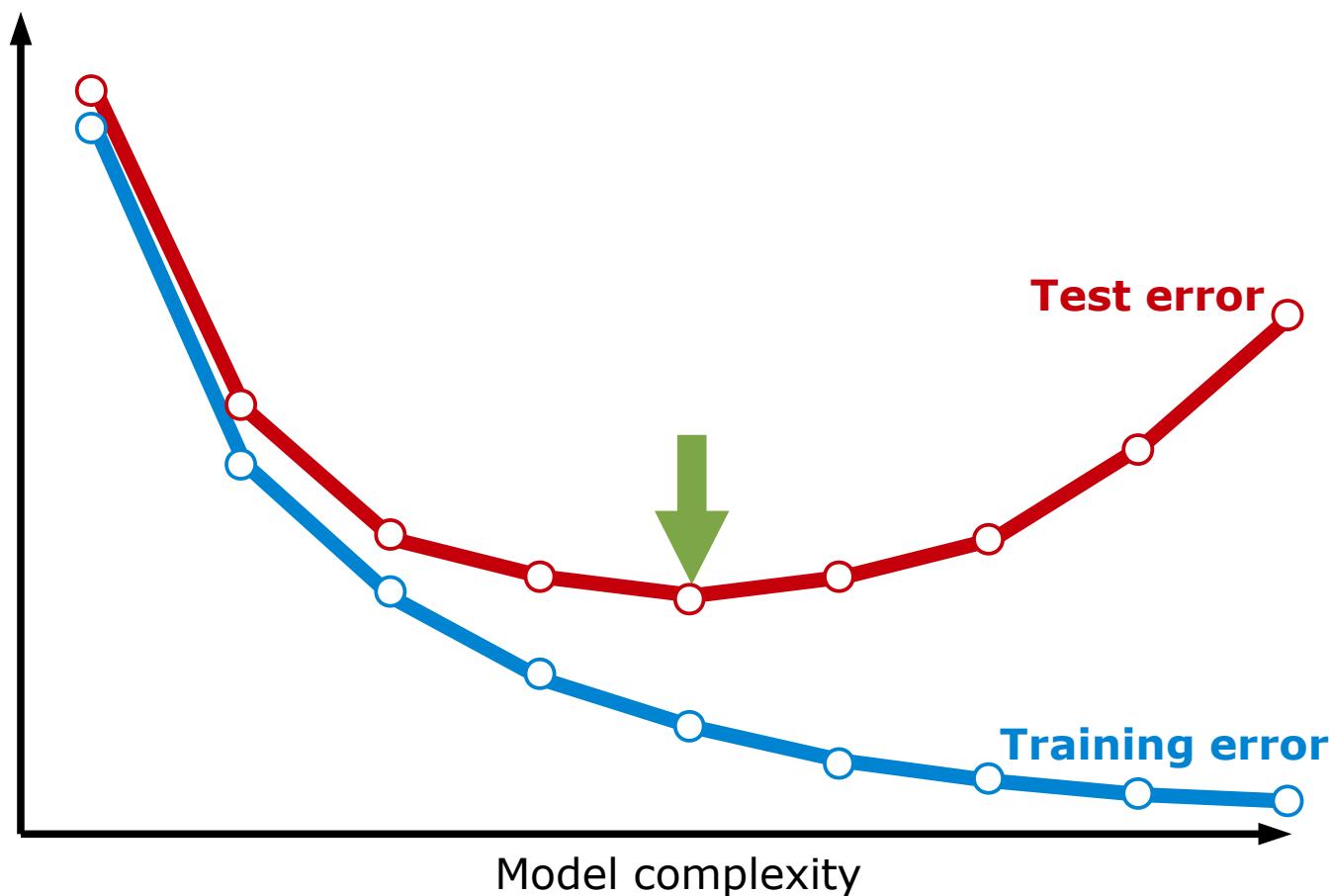
Leave-one-out



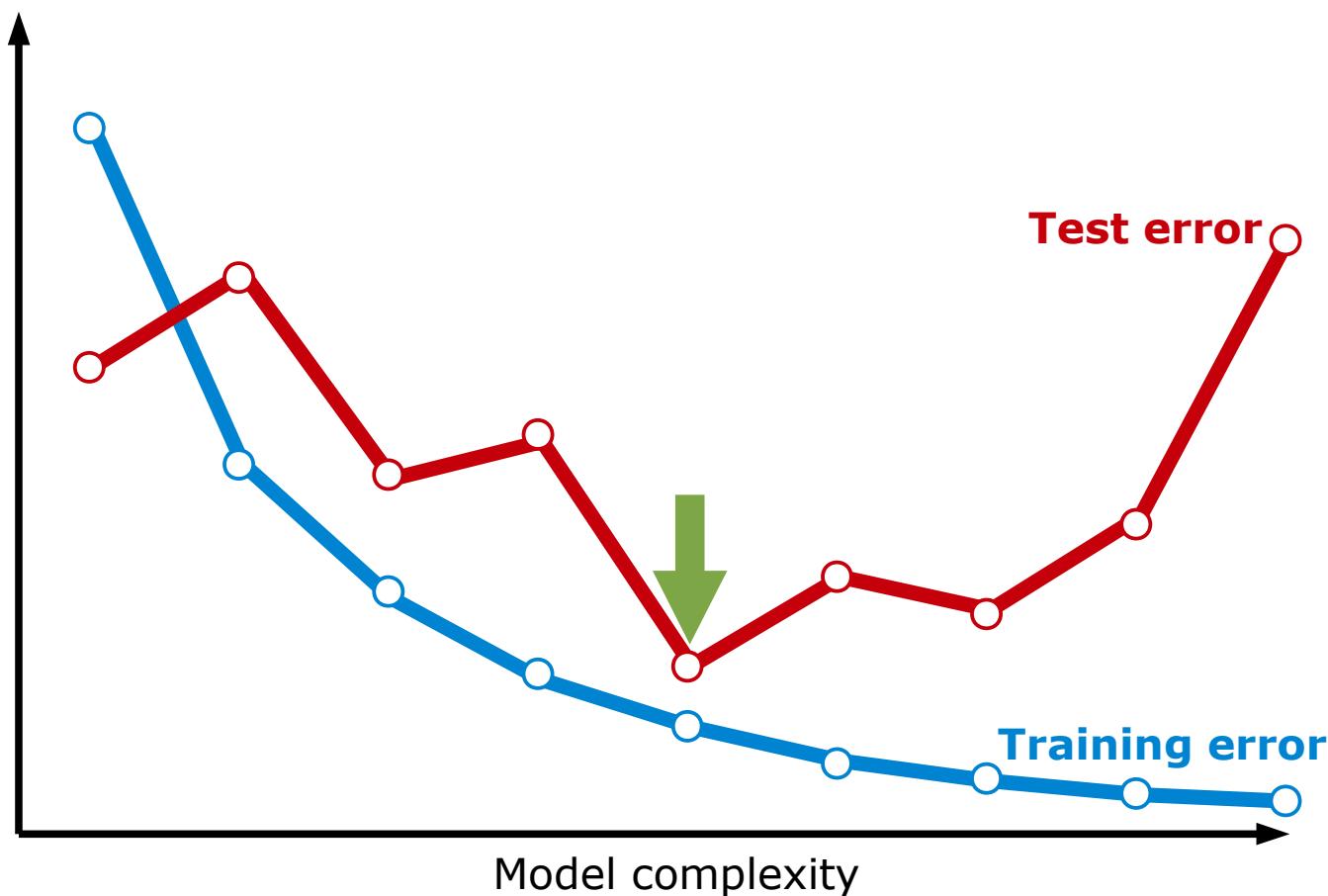
K-fold cross-validation (3-fold)

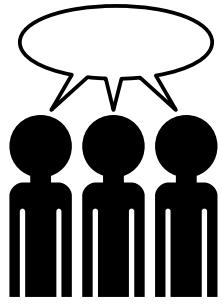


Training and test error



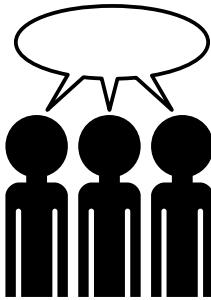
Training and test error





Cross-validation

- What if we want to try out a huge range of complexities, say K from 1 to 10,000
 - Can you come up with a strategy?



Feature subset selection

- Let's say we want to do linear regression
 - We have a large number of attributes

$$x_1, x_2, \dots, x_M$$

- Using all attributes results in a too complex model
 - **Control complexity:** Choose a subset of attributes
 - Small subset = Simple model
 - Large subset = Complex model

- **How many different ways can we choose a subset?**
 - How many models must be compared for
 - M=4
 - M=10
 - M=100

$$f(x) = w_0$$

$$f(x) = w_0 + w_1x_1 + w_2x_{27} + w_3x_{88}$$

$$f(x) = w_0 + w_1x_{19} + w_2x_{76}$$

$$f(x) = w_0 + w_1x_{19} + w_2x_{76} + w_3x_{88}$$

$$f(x) = w_0 + w_1x_1 + w_2x_{27} + w_3x_{19}$$

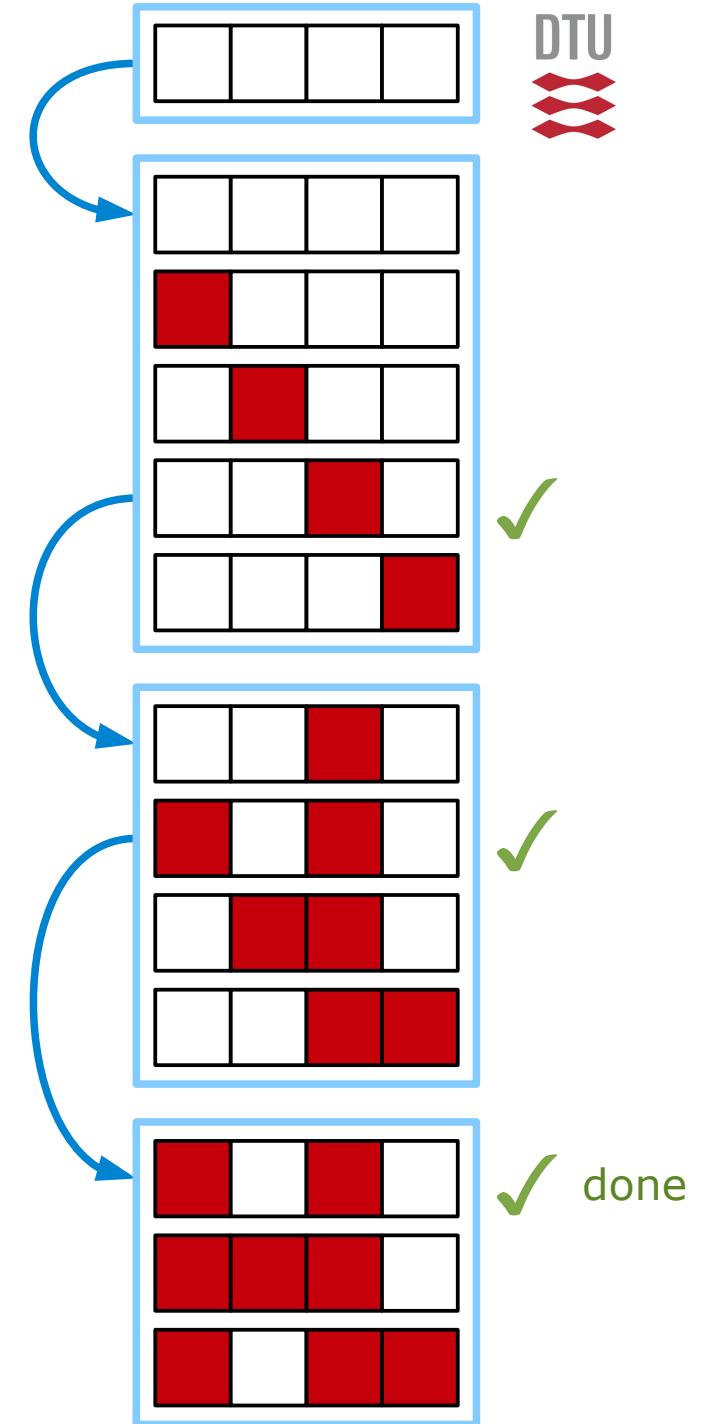
$$f(x) = w_0 + w_1x_{27} + w_2x_{88}$$

⋮

Sequential feature selection

Forward selection

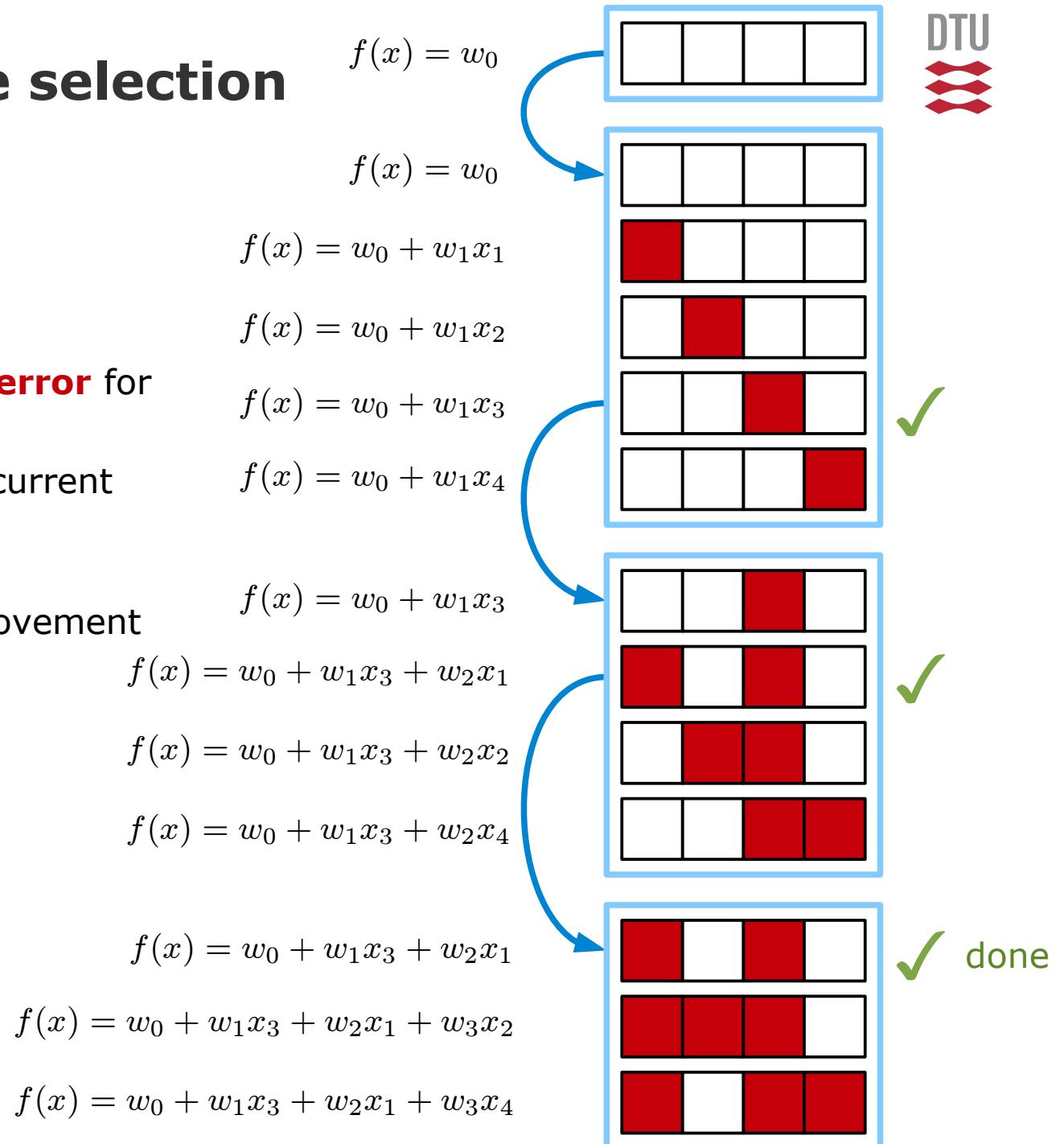
- Start with no features
 - Compute **cross-validation error** for
 - Current feature subset
 - All subsets equal to the current
 - + one added feature
 - Choose best subset
 - Repeat until no further improvement



Sequential feature selection

Forward selection

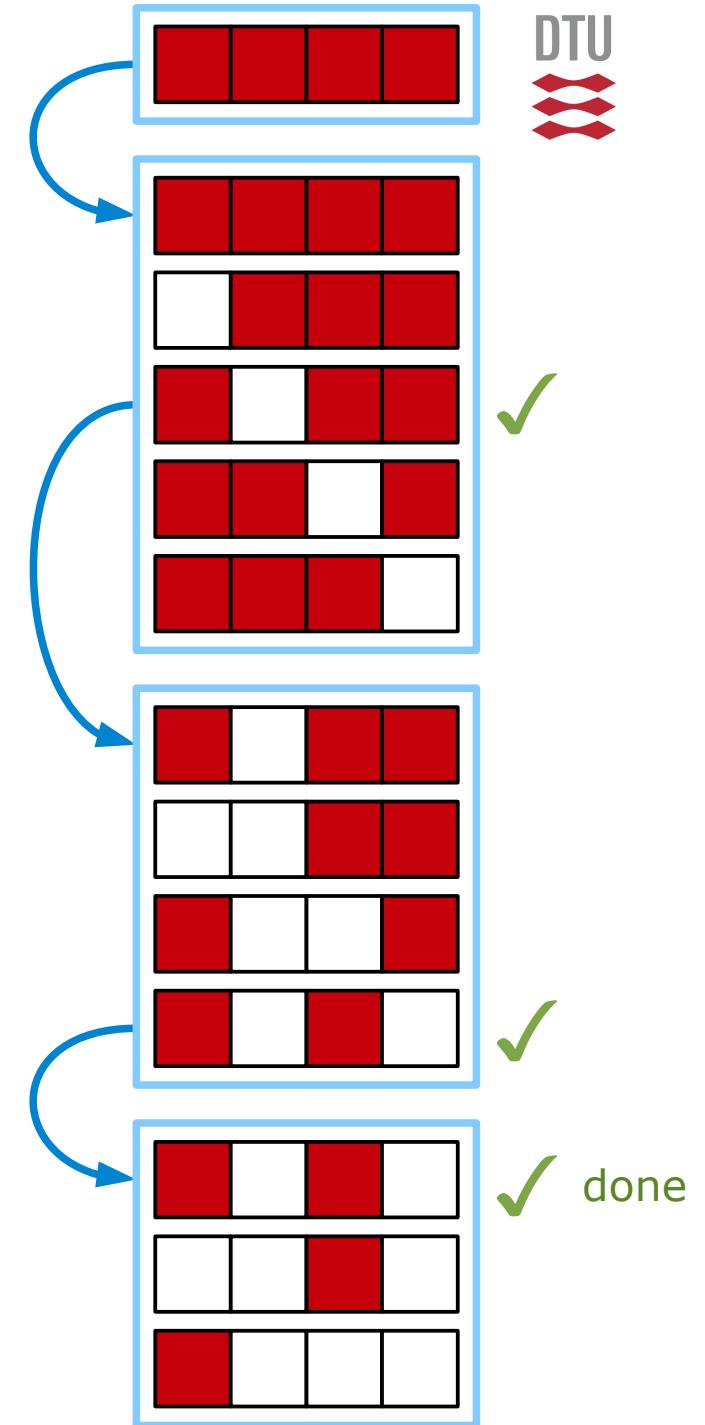
- Start with no features
- Compute **cross-validation error** for
 - Current feature subset
 - All subsets equal to the current + one added feature
- Choose best subset
- Repeat until no further improvement

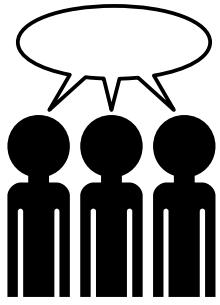


Sequential feature selection

Backward selection

- Start with all features
- Compute **cross-validation error** for
 - Current feature subset
 - All subsets equal to the current
 - one removed feature
- Choose best subset
- Repeat until no further improvement





Feature subset selection

- **How many models do we maximally have to evaluate by forward or backward selection?**

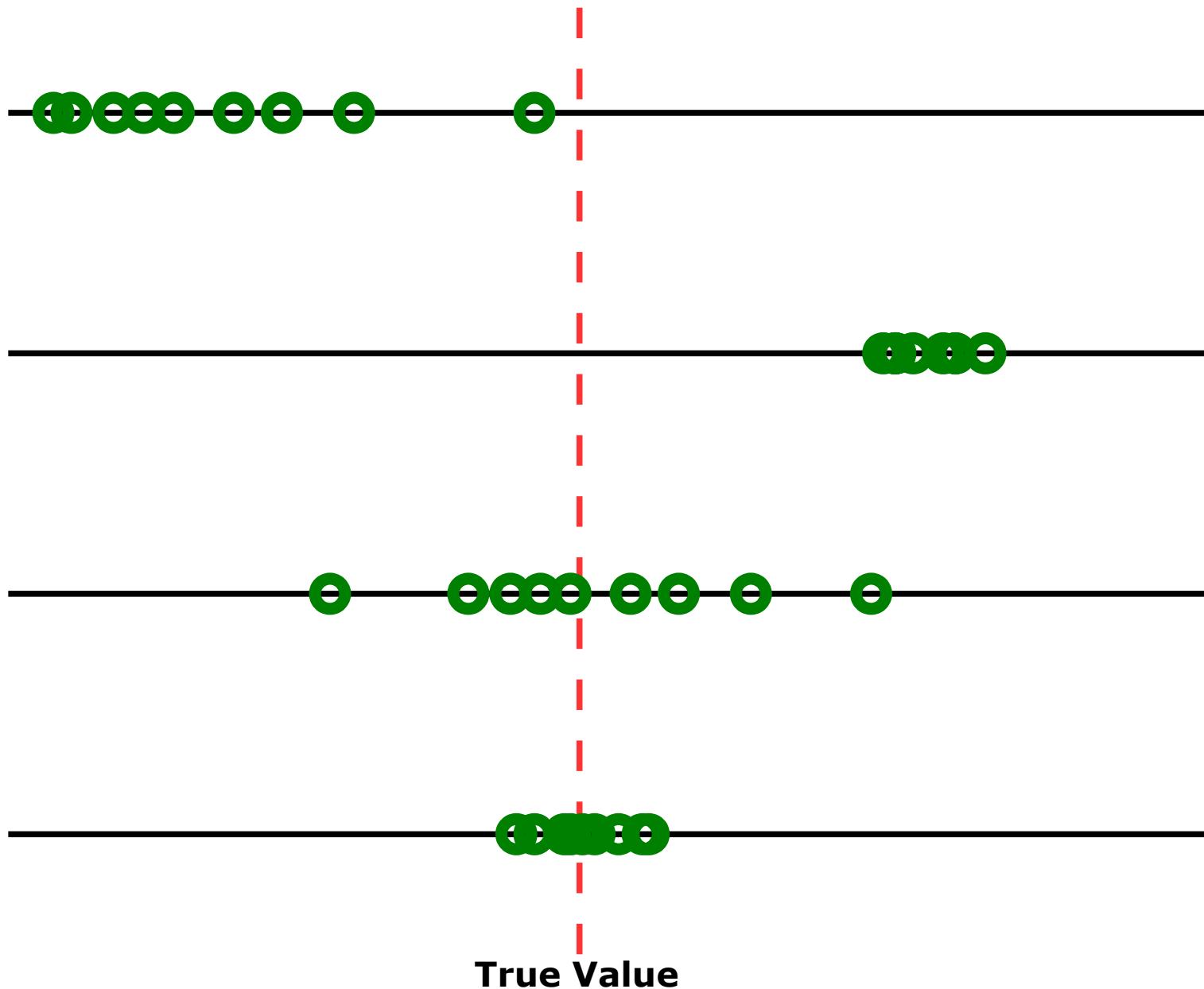
$$x_1, x_2, \dots, x_M$$

- M=4
- M=10
- M=100

Bias-Variance tradeoff

Consider learning the parameter \mathbf{w} for each of N cross-validations, denoted $\mathbf{w}^{(i)}$ and consider the squared error between the estimated and true parameter

$$\begin{aligned}
 \bar{\mathbf{w}} &= \frac{1}{N} \sum_{i=1}^N \mathbf{w}^{(i)} \\
 E[(\mathbf{w}^{true} - \mathbf{w}^{(i)})^2] &= \frac{1}{N} \sum_{i=1}^N (\mathbf{w}^{true} - \mathbf{w}^{(i)})^2 \\
 &= \frac{1}{N} \sum_{i=1}^N (\mathbf{w}^{true} - \bar{\mathbf{w}} + \bar{\mathbf{w}} - \mathbf{w}^{(i)})^2 \\
 &= \frac{1}{N} \sum_{i=1}^N (\mathbf{w}^{true} - \bar{\mathbf{w}})^2 + \frac{1}{N} \sum_{i=1}^N (\bar{\mathbf{w}} - \mathbf{w}^{(i)})^2 + \frac{1}{N} \sum_{i=1}^N 2(\mathbf{w}^{true} - \bar{\mathbf{w}})(\bar{\mathbf{w}} - \mathbf{w}^{(i)}) \\
 &= \underbrace{(\bar{\mathbf{w}} - \mathbf{w}^{true})^2}_{Bias} + \underbrace{\frac{1}{N} \sum_{i=1}^N (\mathbf{w}^{(i)} - \bar{\mathbf{w}})^2}_{Variance}
 \end{aligned}$$



Regularized Least Squares Regression

- **Cost function:** Squared error+ridge penalty

$$E = \sum_n (y_n - f(x_n))^2 + \lambda w^\top w$$

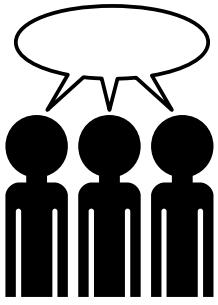
- Model:** Linear regression

$$f(x) = x^\top w$$

- **Parameters**

$$w = \arg \min_w \underbrace{\sum_{n=1}^N (y_n - x_n^\top w)^2}_{E} + \lambda w^\top w$$

$$\begin{aligned} \frac{\partial E}{\partial w} &= 2(y - Xw)^\top X + 2\lambda w^\top = 0 \\ \Rightarrow 2y^\top X &= 2w^\top (X^\top X + \lambda I) \\ \Rightarrow w &= (X^\top X + \lambda I)^{-1} X^\top y \end{aligned}$$



What do you think happens when $\lambda \rightarrow \infty$?
And how can we select the optimal value of λ ?

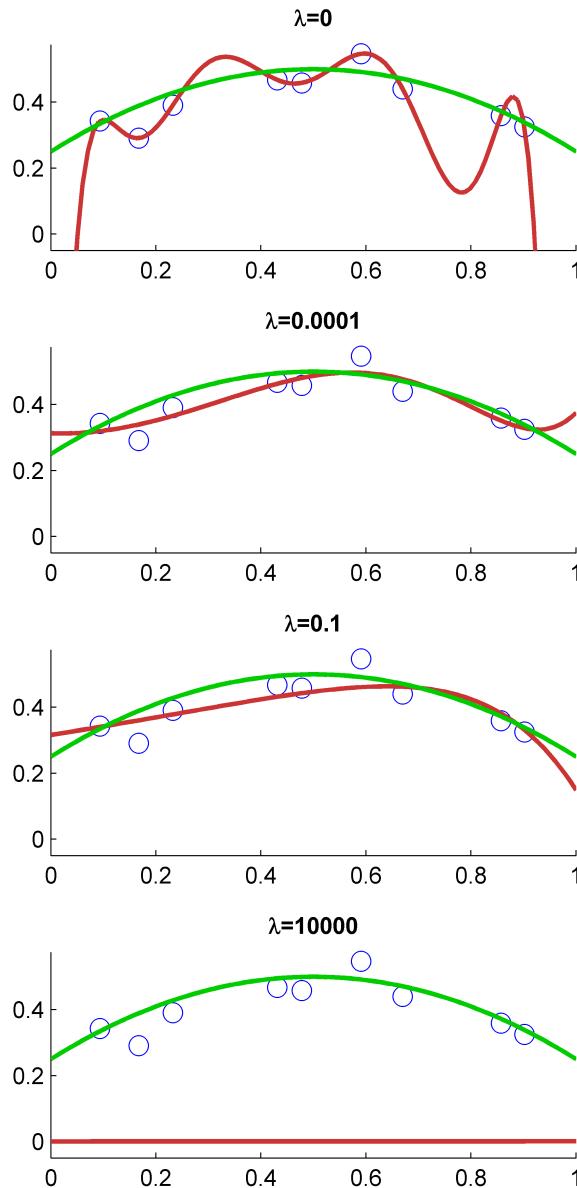
$$\mathbf{w} = \arg \min_{\mathbf{w}} \sum_{n=1}^N (y_n - \mathbf{x}_n^\top \mathbf{w})^2 + \lambda \mathbf{w}^\top \mathbf{w}$$

$$\frac{\partial E}{\partial \mathbf{w}} = 2(\mathbf{y} - \mathbf{X}\mathbf{w})^\top \mathbf{X} + 2\lambda \mathbf{w}^\top = \mathbf{0}$$

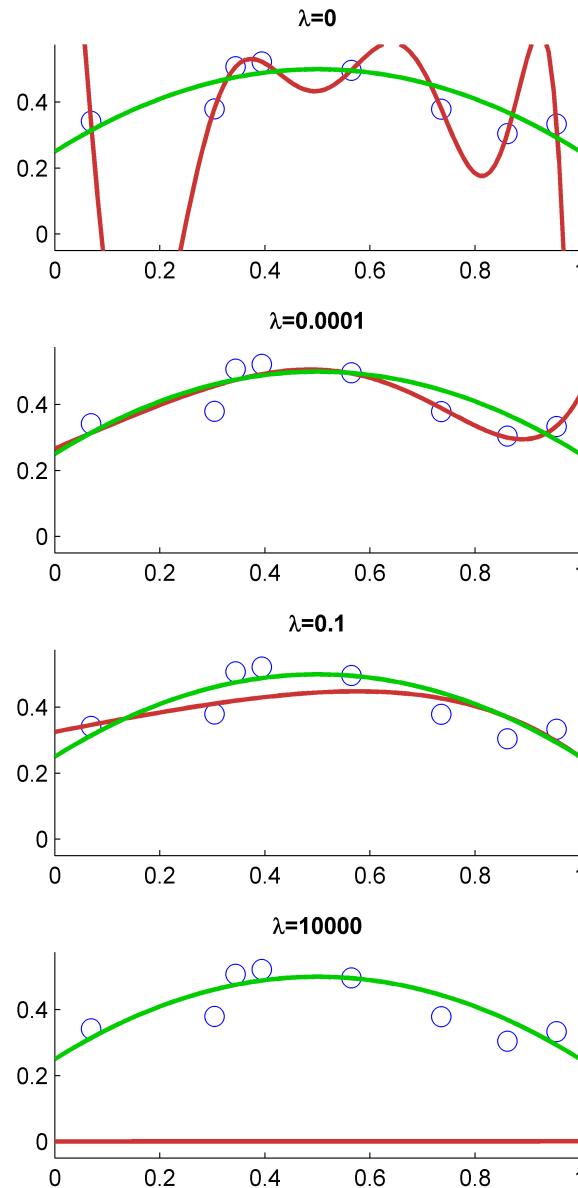
$$\Rightarrow 2\mathbf{y}^\top \mathbf{X} = 2\mathbf{w}^\top (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})$$

$$\Rightarrow \mathbf{w} = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^\top \mathbf{y}$$

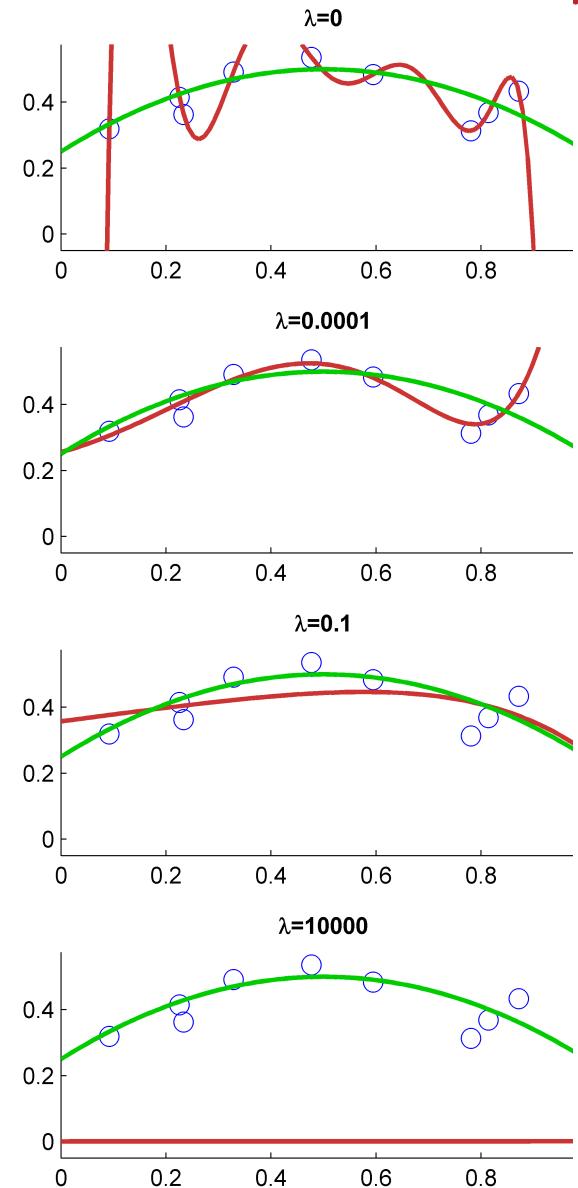
Dataset 1



Dataset 2

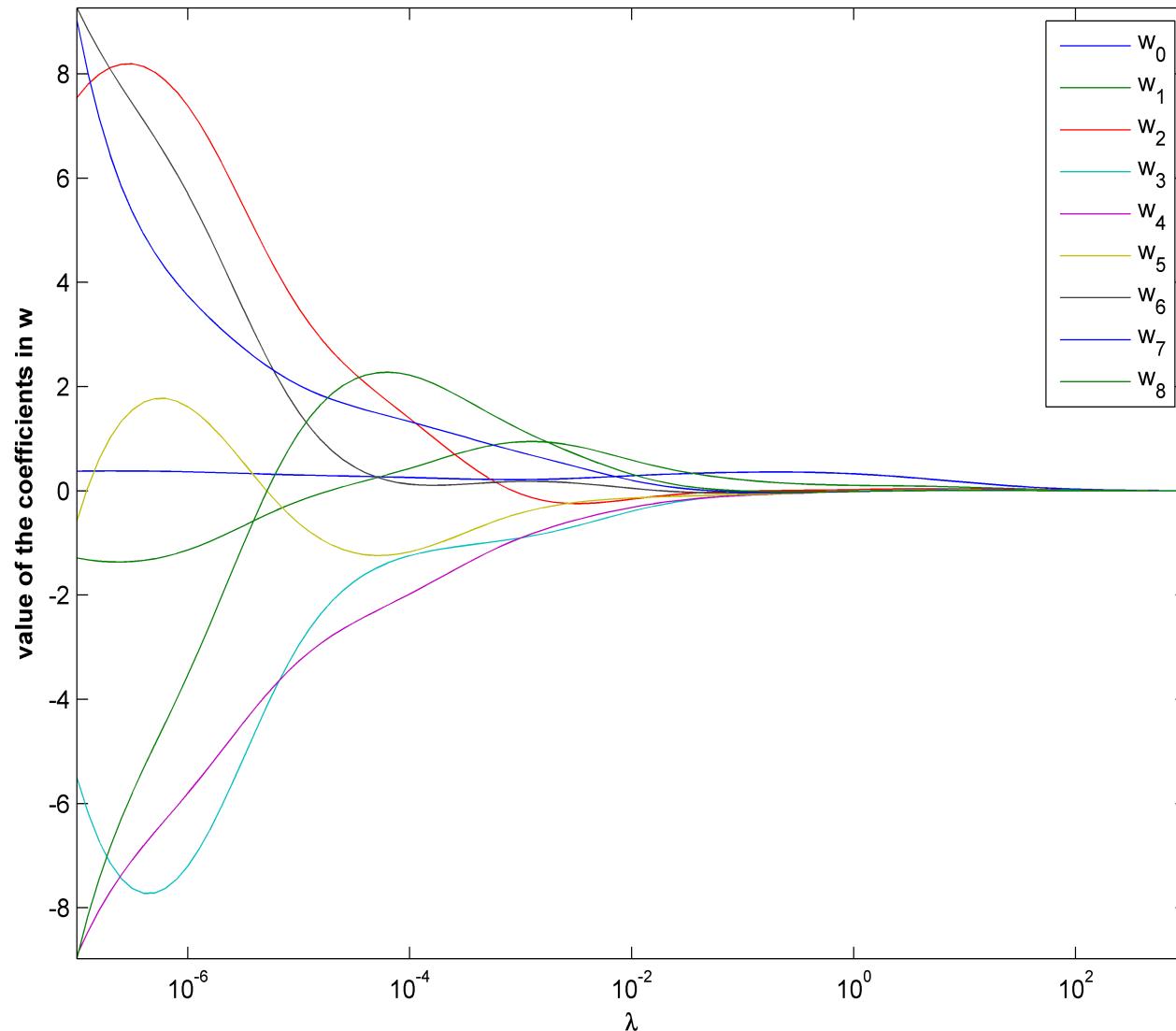


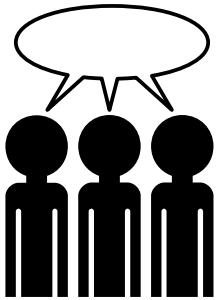
Dataset 3



By regularization we can tradeoff bias and variance, in particular we can hope to substantially reduce variance without introducing too much bias!

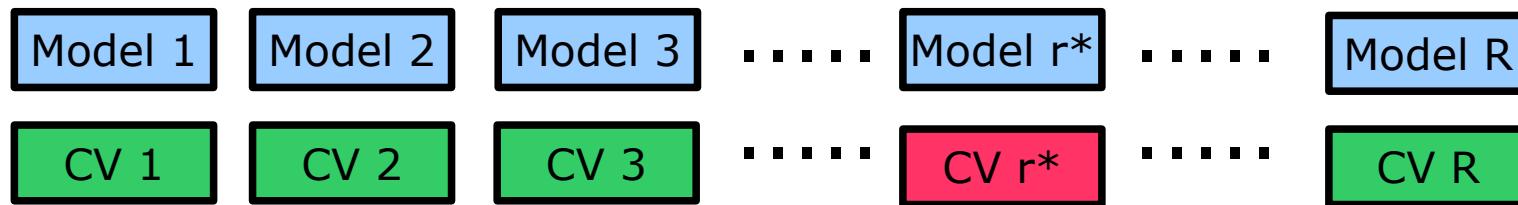
$$\begin{aligned}
 f(x) &= w_0 + w_1x^1 + w_2x^2 + w_3x^3 + w_4x^4 + w_5x^5 + w_6x^6 + w_7x^7 + w_8x^8 \\
 &= \sum_{k=0}^8 w_k x^k
 \end{aligned}$$



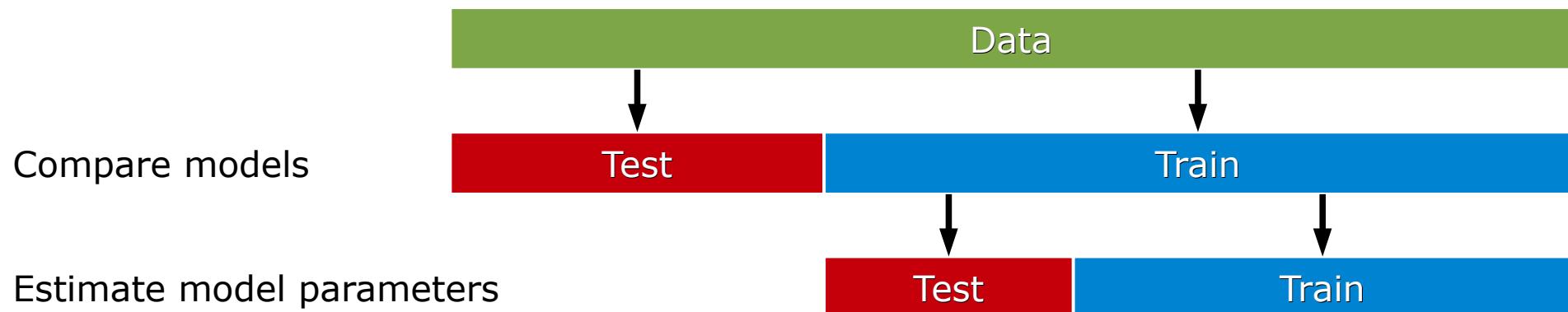


Imagine we estimate R different models and select the model r^* with the lowest cross-validation error as the best model. Will the estimated cross-validation error for this model be a correct estimate of how well the model generalizes to new data?

(i.e., is the obtained cross-validation error an un-biased estimator of the generalization error?)



Multi-level cross-validation



Evaluating the performance of classifiers

- Compare classifier performance to random guessing
 - i.e., evaluate how significantly the classifier performs relative to random guessing
- Derive confidence interval for accuracy of classifiers
 - i.e., confidence intervals are used to indicate the reliability of an estimate.
- Compare the performance of two classifiers
 - i.e., is one classifier significantly better than another classifier.

Evaluating the significance of a classifier

Imagine we train a classifier on N observations and correctly classify L of the observations. We want to know if our classifier is better than random guessing. I.e. We would like to reject the null hypothesis:

H_0 : Our classifier is guessing at random

If we were random guessing the distribution of classifying L observations correctly would follow a binomial with $p=1/2$.

$$p(L|N,p) = \frac{N!}{L!(N-L)!} p^L (1-p)^{N-L}$$

The probability by random of classifying L or more observations correctly by random is then given by

$$\begin{aligned} p(l \geq L|N,p) &= \sum_{l=L}^N \frac{N!}{l!(N-l)!} 0.5^l (1-0.5)^{N-l} \\ &= 1 - \sum_{l=0}^{L-1} \frac{N!}{l!(N-l)!} 0.5^l (1-0.5)^{N-l} \end{aligned}$$

Consider two classification problems.

Problem 1: $N=10, L=8$

$$p(l \geq 8 | 10, 0.5) = 0.0547$$

Problem 2: $N=100, L=60$

$$p(l \geq 60 | 100, 0.5) = 0.0284$$

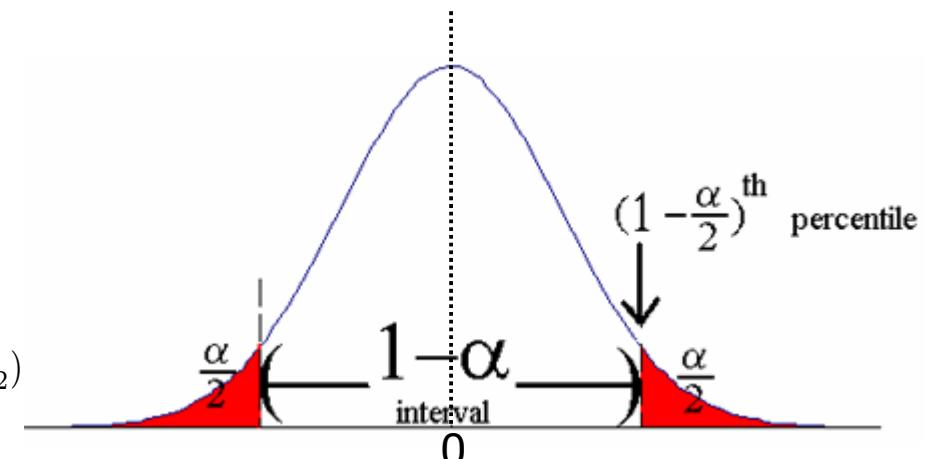
Confidence interval for the accuracy of a classifier

The empirical accuracy has mean $p=L/N$ and variance $p(1-p)/N$. For large N we can approximate the binomial distributed by the normal distribution to obtain confidence intervals for the accuracy, i.e.

$$p(Z_{\alpha/2} \leq \frac{acc - p}{\sqrt{p(1-p)/N}} \leq Z_{1-\alpha/2}) = 1 - \alpha$$

$$Z_{1-\alpha/2} = -Z_{\alpha/2}$$

$$p(Z_{\alpha/2} \leq \frac{acc - p}{\sqrt{p(1-p)/N}} \leq -Z_{\alpha/2}) = p\left(\left(\frac{acc - p}{\sqrt{p(1-p)/N}}\right)^2 \leq Z_{\alpha/2}^2\right)$$



$$\left(\frac{acc - p}{\sqrt{p(1-p)/N}}\right)^2 = Z_{\alpha/2}^2 \Rightarrow$$

$$0 = (N + Z_{\alpha/2}^2)p^2 - (2N \cdot acc + Z_{\alpha/2}^2)p + N \cdot acc^2 \Rightarrow$$

$$p \in \left[\frac{2N \cdot acc + Z_{\alpha/2}^2 - Z_{\alpha/2} \sqrt{Z_{\alpha/2}^2 + 4N \cdot acc - 4N \cdot acc^2}}{2(N + Z_{\alpha/2}^2)}; \frac{2N \cdot acc + Z_{\alpha/2}^2 - Z_{\alpha/2} \sqrt{Z_{\alpha/2}^2 + 4N \cdot acc - 4N \cdot acc^2}}{2(N + Z_{\alpha/2}^2)} \right]$$

Consider a model with accuracy 80%, the confidence interval as a function of number of observations N are

N	20	50	100	500	1000	5000
$p \in$	[0.584; 0.919]	[0.670; 0.888]	[0.711; 0.867]	[0.763; 0.833]	[0.774; 0.824]	[0.789; 0.811]

Comparing the performance of two classifiers based on k-fold cross-validation using the same splits for the two models



Let e_{1j} and e_{2j} be the error rates at cross validation split j for model 1 and model 2 respectively and $d_j = e_{1j} - e_{2j}$

$$\begin{aligned}\bar{d} &= \frac{1}{k} \sum_{j=1}^k d_j && \text{Variance of the mean value} \\ \bar{\sigma}^2 &= \frac{\sum_{j=1}^k (d_j - \bar{d})^2}{k(k-1)} \\ d^{cv} &\in [\bar{d} - t_{1-\alpha/2, k-1} \hat{\sigma}; \bar{d} + t_{1-\alpha/2, k-1} \hat{\sigma}]\end{aligned}$$

Consider two classifiers with the following error rates across 5-cross validation splits.

Classifier 1: $e_1 = [0.2 \ 0.1 \ 0.3 \ 0.2 \ 0.4]$

$d = e_1 - e_2 = [0.1 \ 0 \ 0.1 \ 0 \ 0.3]$

Classifier 2: $e_2 = [0.1 \ 0.1 \ 0.2 \ 0.2 \ 0.1]$

$\bar{d} = 0.1, \ \hat{\sigma}^2 = 0.003,$

$$\begin{aligned}d^{cv} &\in [0.1 - 2.7764 \cdot \sqrt{0.003}; 0.1 + 2.7764 \cdot \sqrt{0.003}] \\ &= [-0.0521; 0.2521]\end{aligned}$$

Comments to the report

- Consider including more summary statistics than displaying the simple mean and standard deviation, i.e. Median, range, etc. Especially if no box plots are included.
- Remember labels on your plots and explain what is on your plots (and why) + think ACCENT
- Outliers in a boxplot do not necessarily mean that it is not a feasible value in the attribute.
- Use correlation matrix/plots to describe if variables are correlated
- PCA analysis:
 - Explain what you do to your data prior to PCA (1-out-of-K, subtract mean/normalize)
 - Interpret PCA directions (plot V or a subset of directions)!
 - Explain what it is you are displaying in your plots of the PCA (E.g. the projections of data onto PC1 and PC2)
 - Do not include your output y in the PCA, but color code for instance the observations according to y if you have a classification problem.
- In general: Discuss your findings!! Document your understanding.
- If you have a regression problem: try to plot output vs. Inputs.
- Use the report description as a checklist for you report content

02450 Introduction to machine learning and data modeling

A collage of mathematical symbols including integrals, summation, infinity, and various Greek letters like theta, omega, and chi.

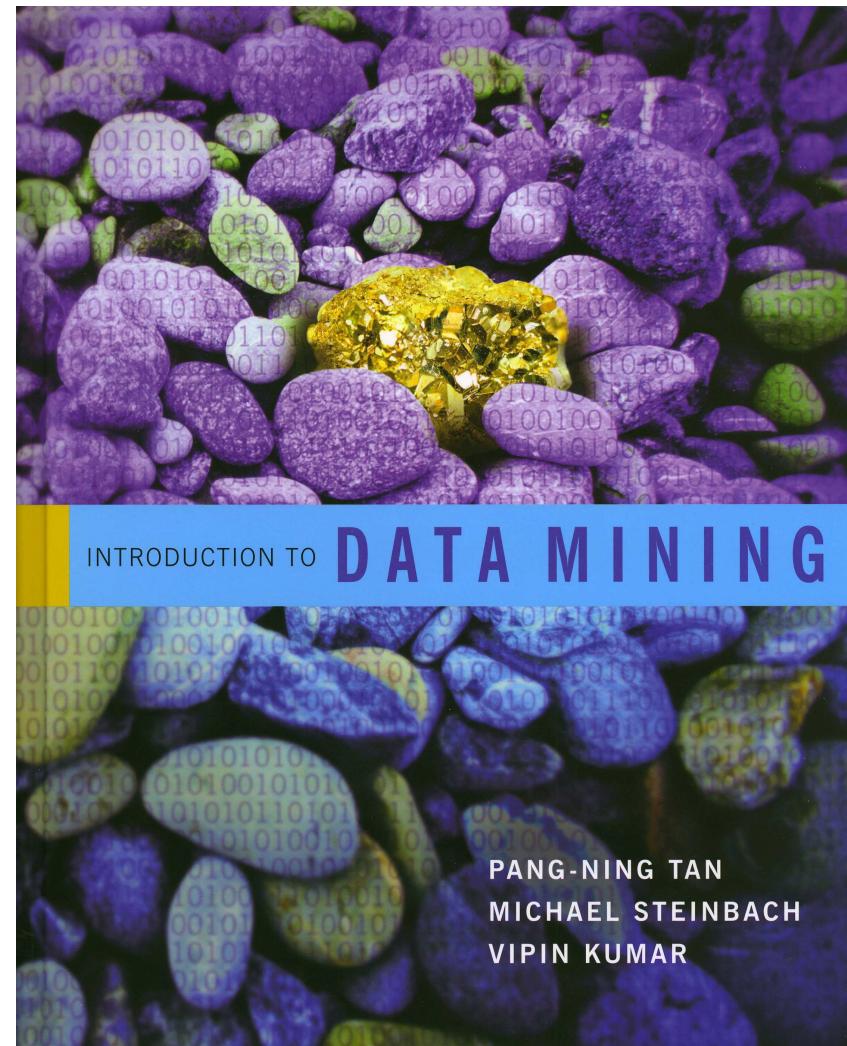
Reading material

Tan, Steinbach and Kumar
"Introduction to Data Mining"

Section 5.2-5.4

Group of the day

Alexander Bschorr
Christian Bøgh
Thorbjørn Kruse Olesen
Lasse Pedersen
Nicolai Sonne
Lasse Herløw
Jacopo Fabiani
Andrea Cuttone
Ioannis Petridis
Paul Damade
Fredrik Bror Gustav Kock
Matias Laura Rasmussen
Farzana Nasry



Lecture schedule

1. Introduction
(Tan 1.1-1.4)

Data: Feature extraction and visualization

2. Data and feature extraction
(Tan 2.1-2.3 + B1 (+ A))
3. Measures of similarity and summary statistics
(Tan 2.4 + 3.1-3.2 + C1-C2)
4. Data visualization
(Tan 3.3)

Supervised learning: Classification and regression

5. Decision trees and linear regression
(Tan 4.1-4.3 + D)
6. Overfitting and performance evaluation
(Tan 4.4-4.6)
7. **Nearest neighbor, naive Bayes, and artificial neural networks**
(Tan 5.2-5.4)

8. Ensemble methods and multi class classifiers
(Tan 5.6-5.8)

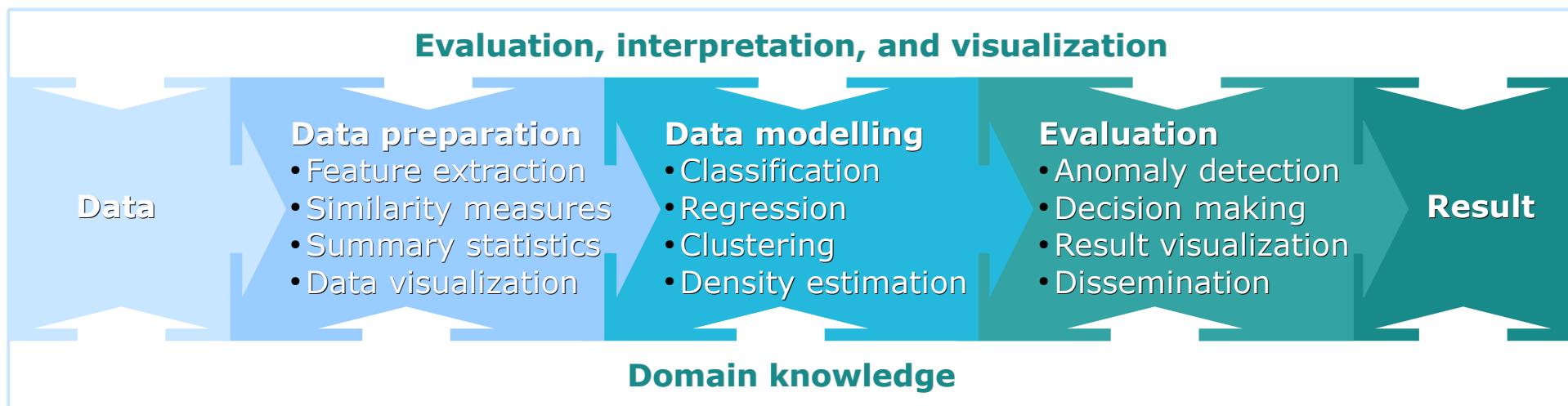
Unsupervised learning: Clustering and density est.

9. K-means and hierarchical clustering
(Tan 8.1-8.3)
10. Mixture models and association mining
(Tan 9.2.2 + 6.1-6.3)
11. Density estimation and anomaly detection
(Tan 10.1-10.4)

Machine learning and data modelling in practice

12. Putting it all together: Summary and overview
13. Mini project

Data modeling framework



After today you should be able to:

Explain how K-Nearest Neighbors can be used to classify data

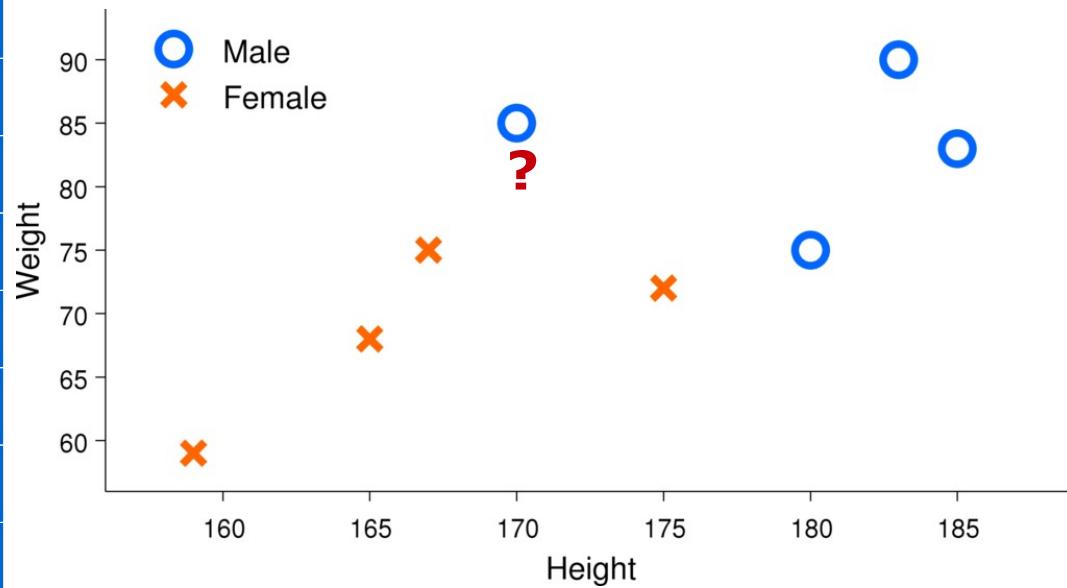
Account for the assumptions made in Naïve Bayes

Apply Bayes theorem to obtain the class posterior likelihood

Understand the principle behind artificial neural networks (ANN) and how ANN can be used for classification and regression.

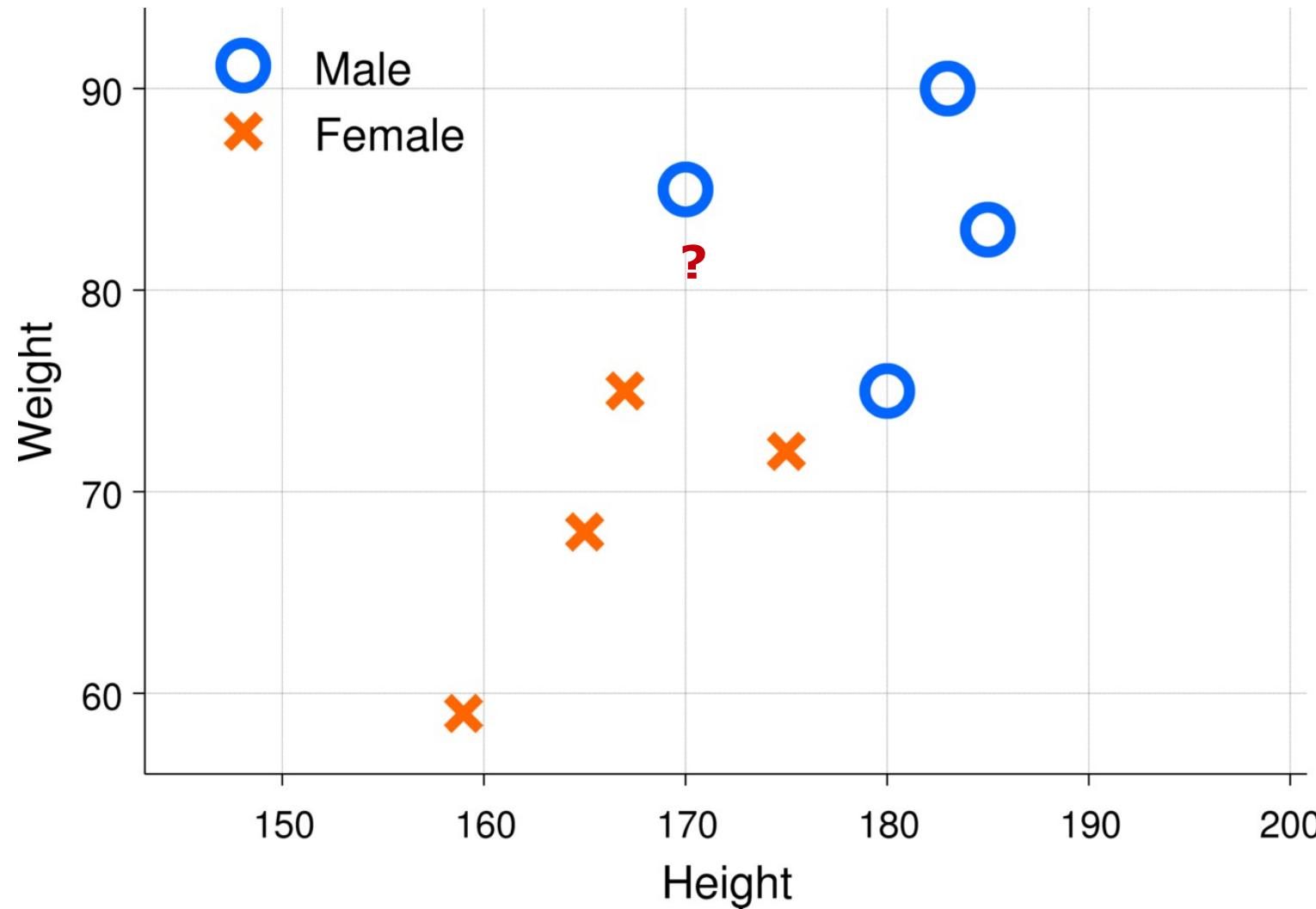
Classify gender based on height and weight

	Height	Weight	Gender
1	183	90	Male
2	180	75	Male
3	170	85	Male
4	185	83	Male
5	159	59	Female
6	167	75	Female
7	165	68	Female
8	175	72	Female
9	171	82	?



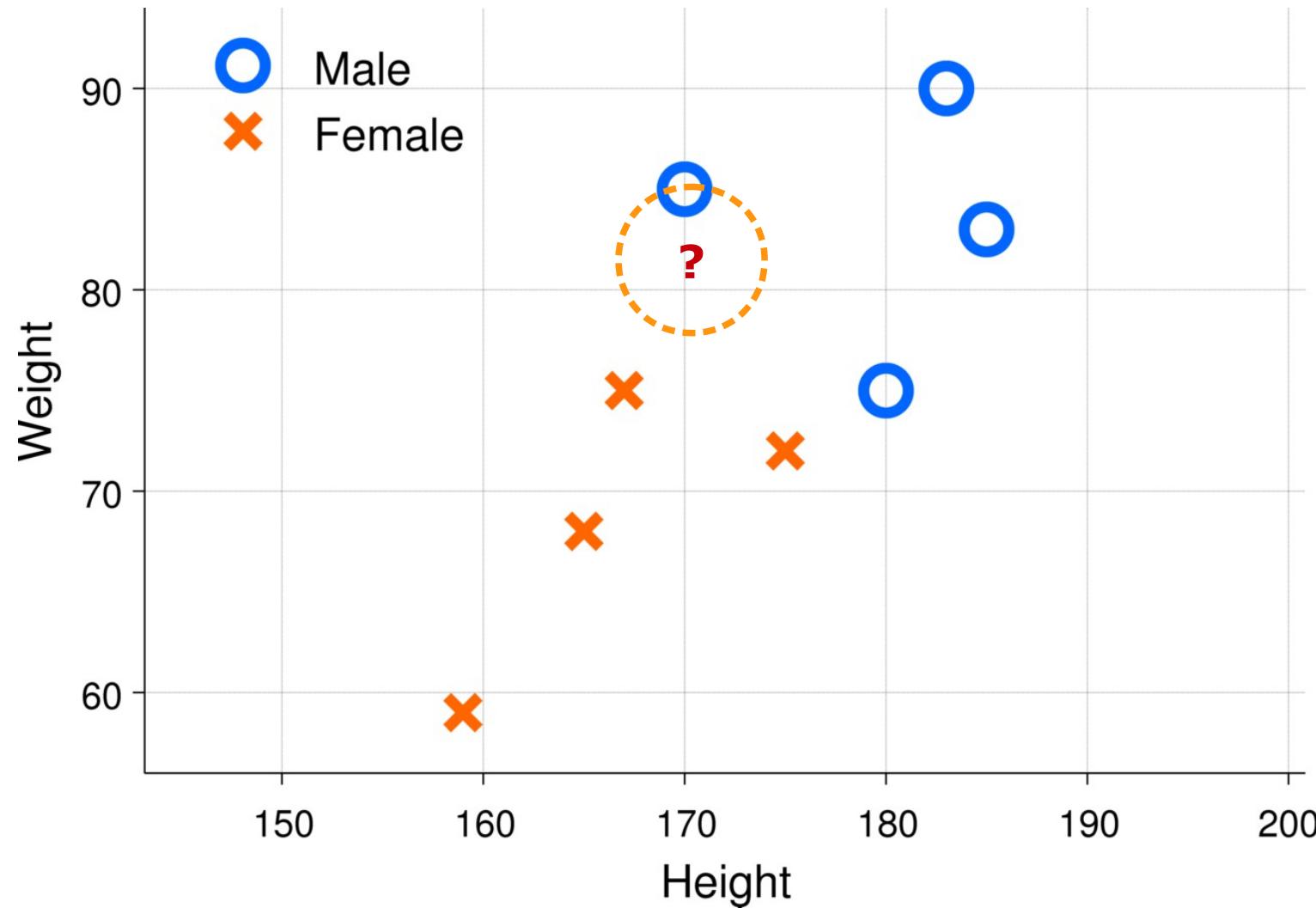
Nearest neighbor classifier

- 1 nearest neighbor



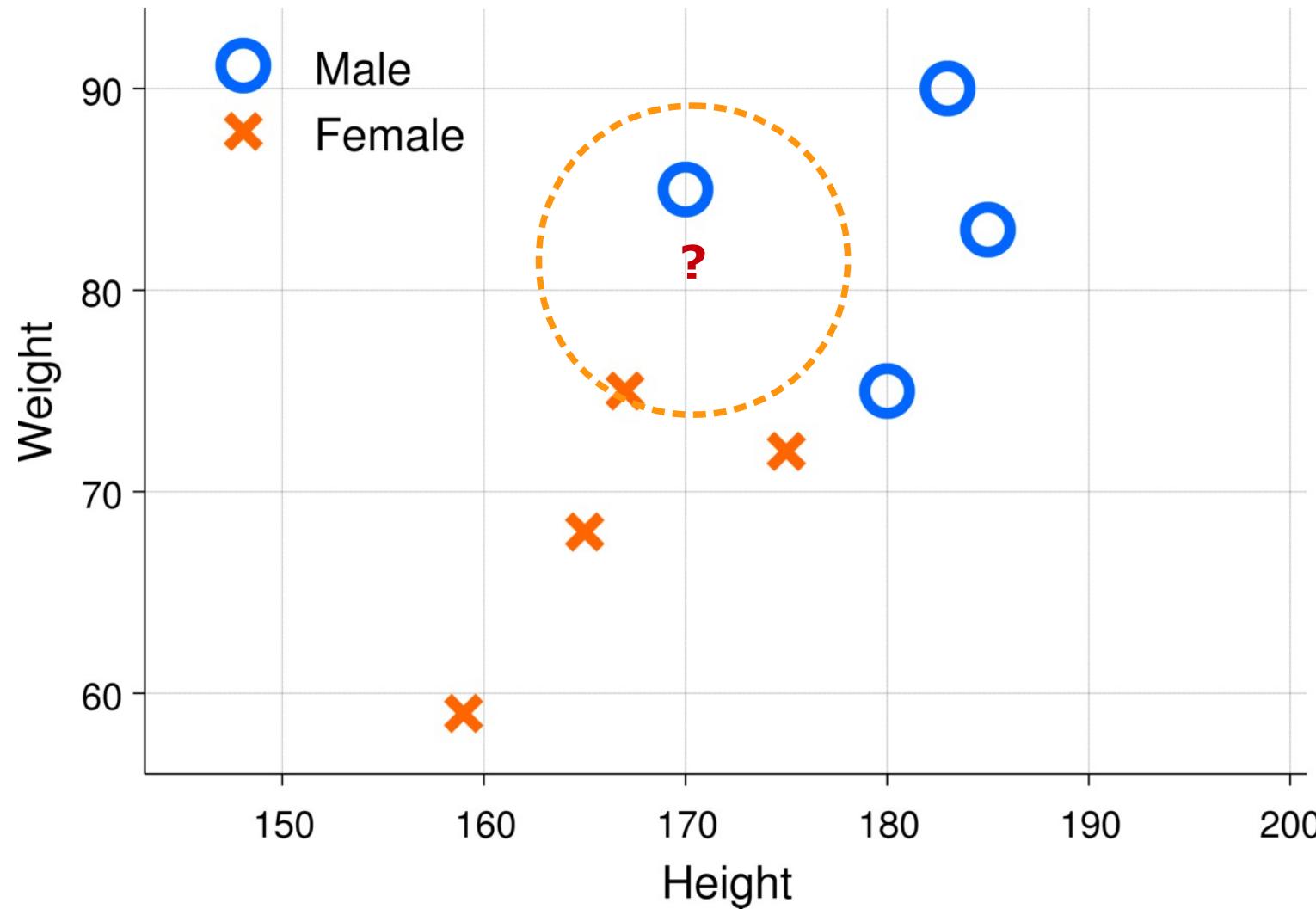
Nearest neighbor classifier

- 1 nearest neighbor



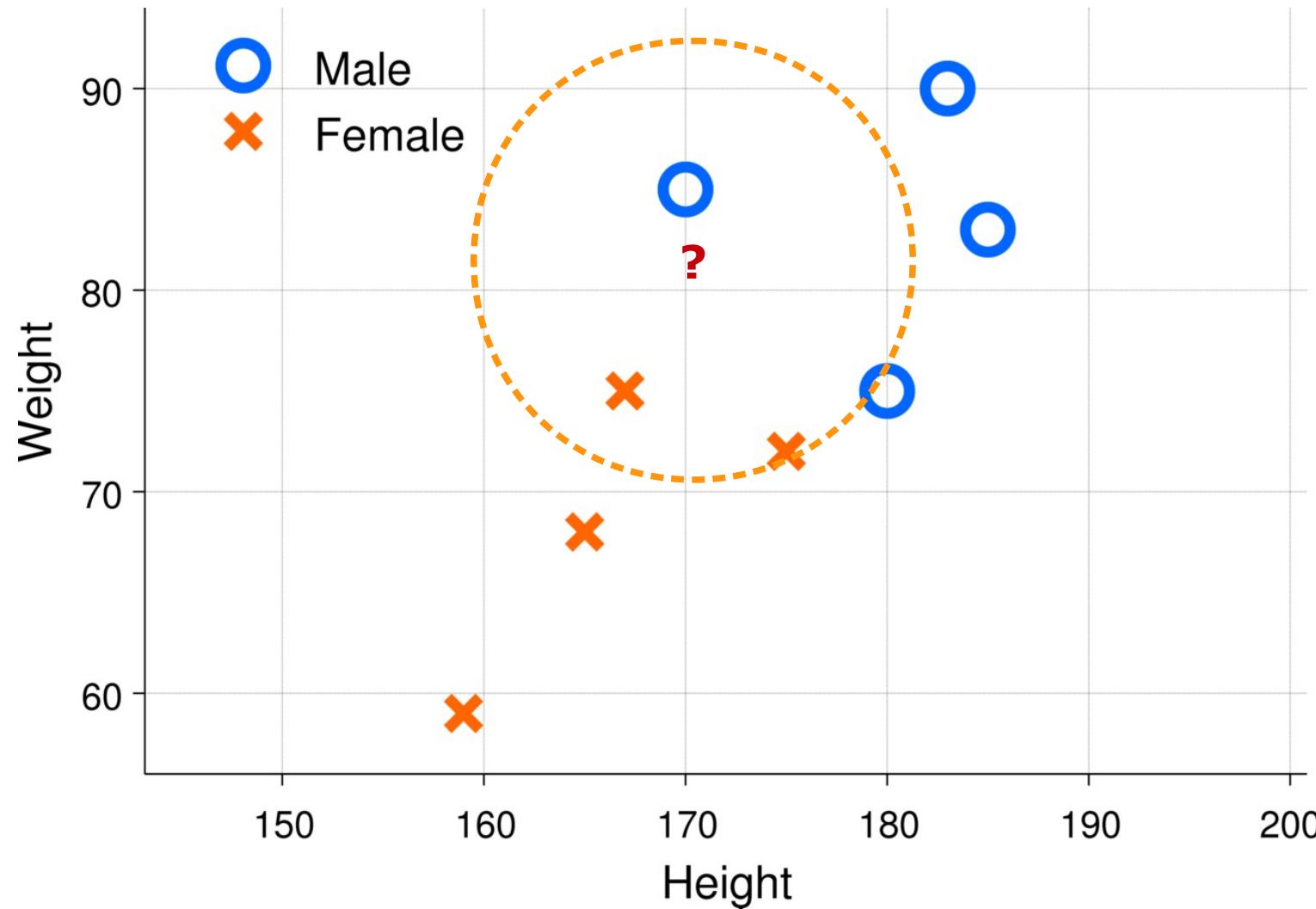
Nearest neighbor classifier

- 2 nearest neighbors



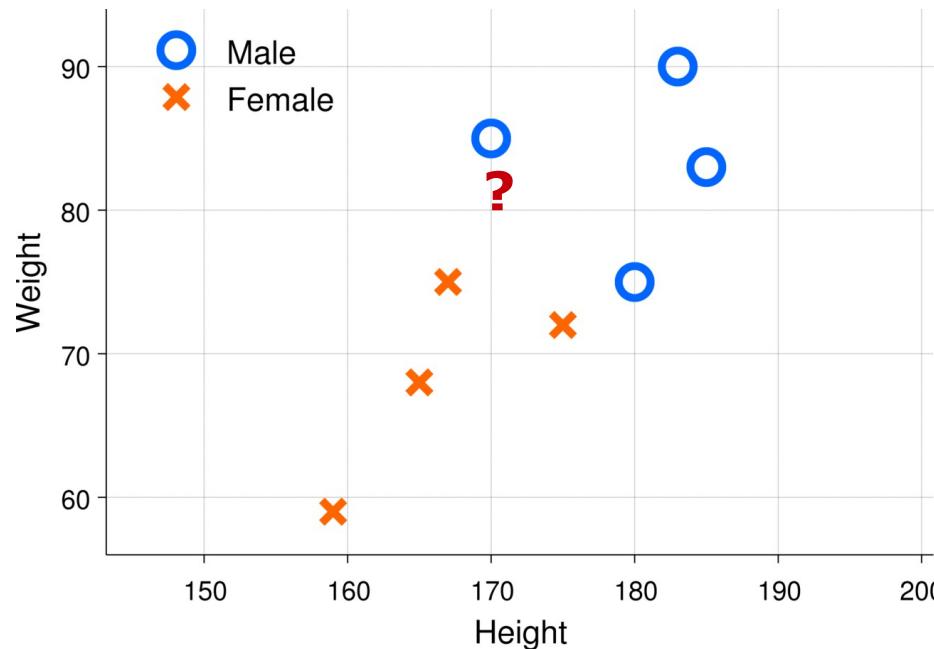
Nearest neighbor classifier

- 3 nearest neighbors

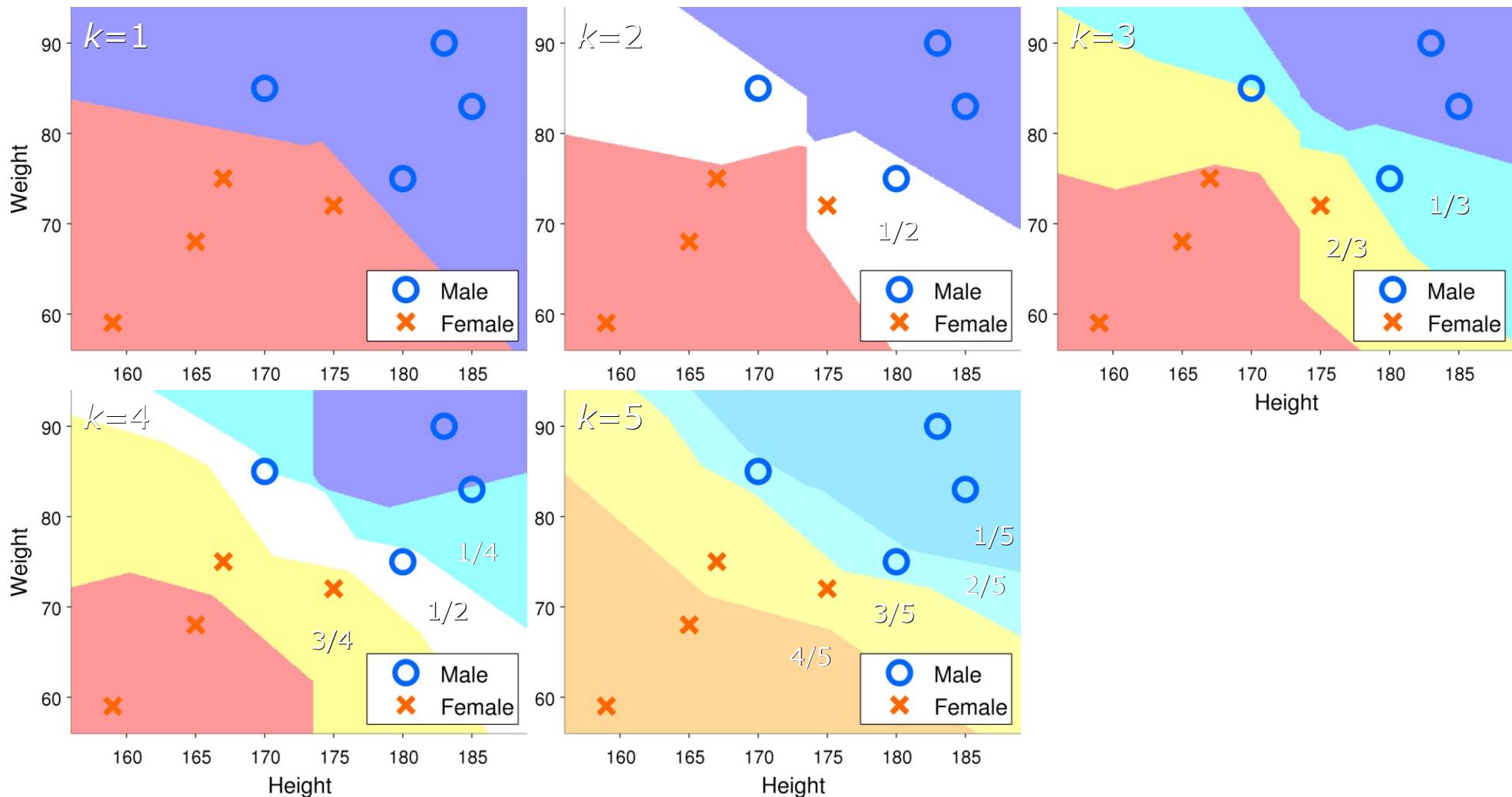


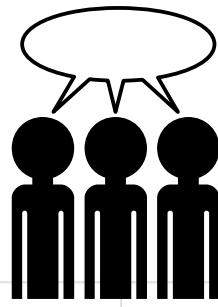
Nearest neighbor classifier

- Choose
 - The number of neighbors, k
 - A distance measure
1. Compute distance to all other data objects
 2. Find the k nearest data objects
 3. Classify according to majority of neighbors



Nearest neighbor decision surface





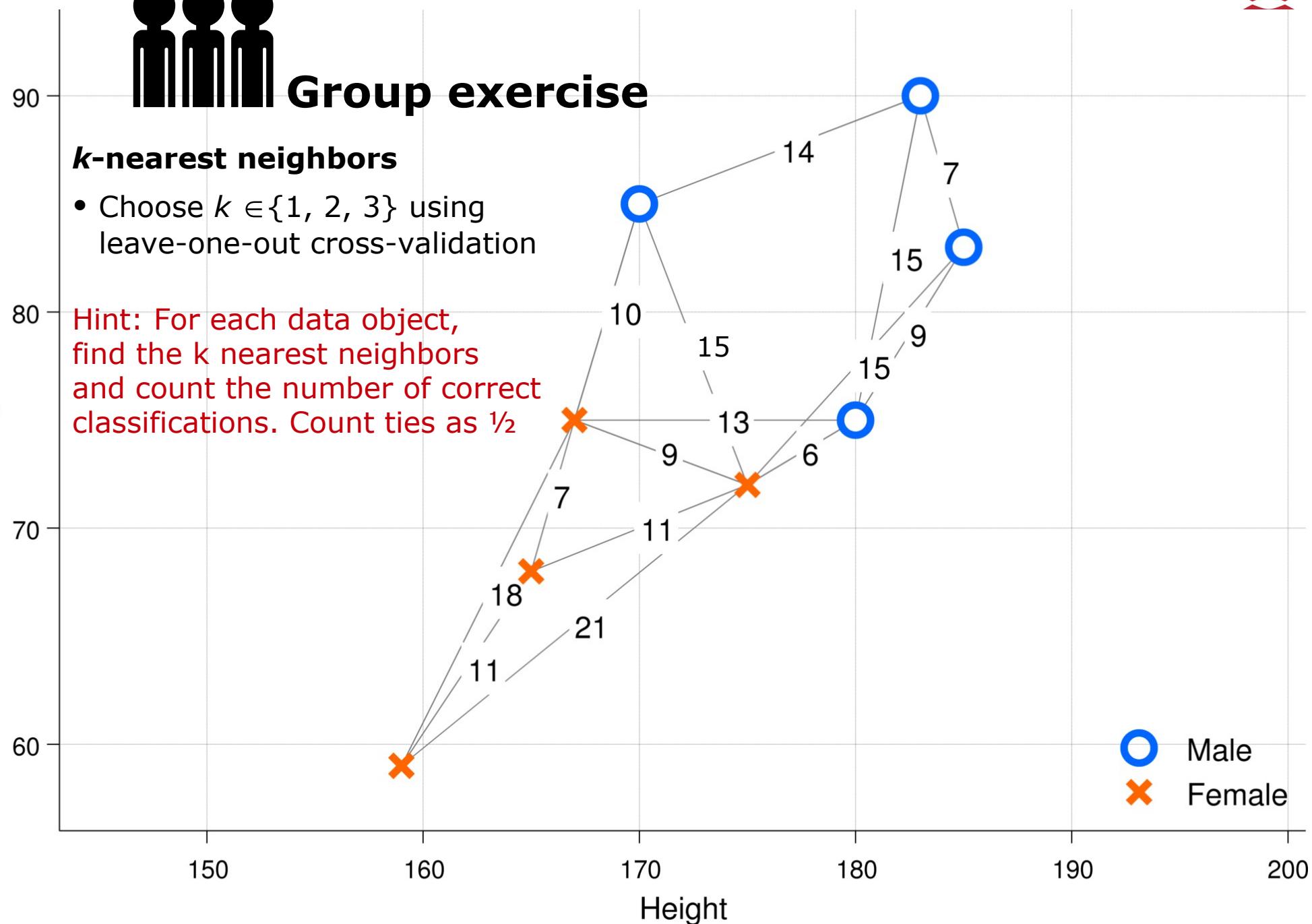
Group exercise

***k*-nearest neighbors**

- Choose $k \in \{1, 2, 3\}$ using leave-one-out cross-validation

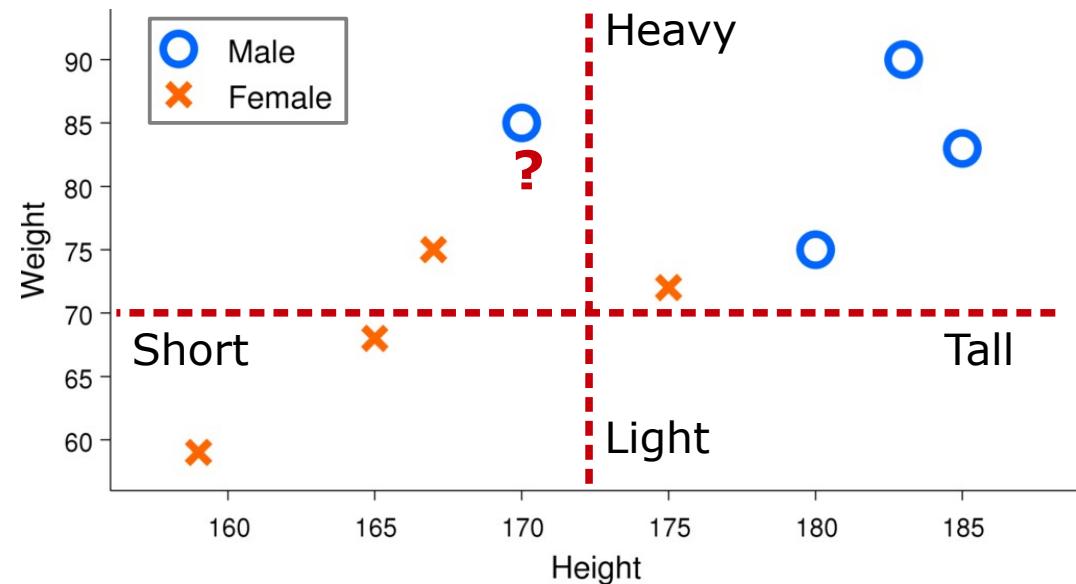
Hint: For each data object, find the k nearest neighbors and count the number of correct classifications. Count ties as $\frac{1}{2}$

Weight



Bayesian classifiers

	Height	Weight	Gender
1	Tall	Heavy	Male
2	Tall	Heavy	Male
3	Short	Heavy	Male
4	Tall	Heavy	Male
5	Short	Light	Female
6	Short	Heavy	Female
7	Short	Light	Female
8	Tall	Light	Female
9	Short	Heavy	Light



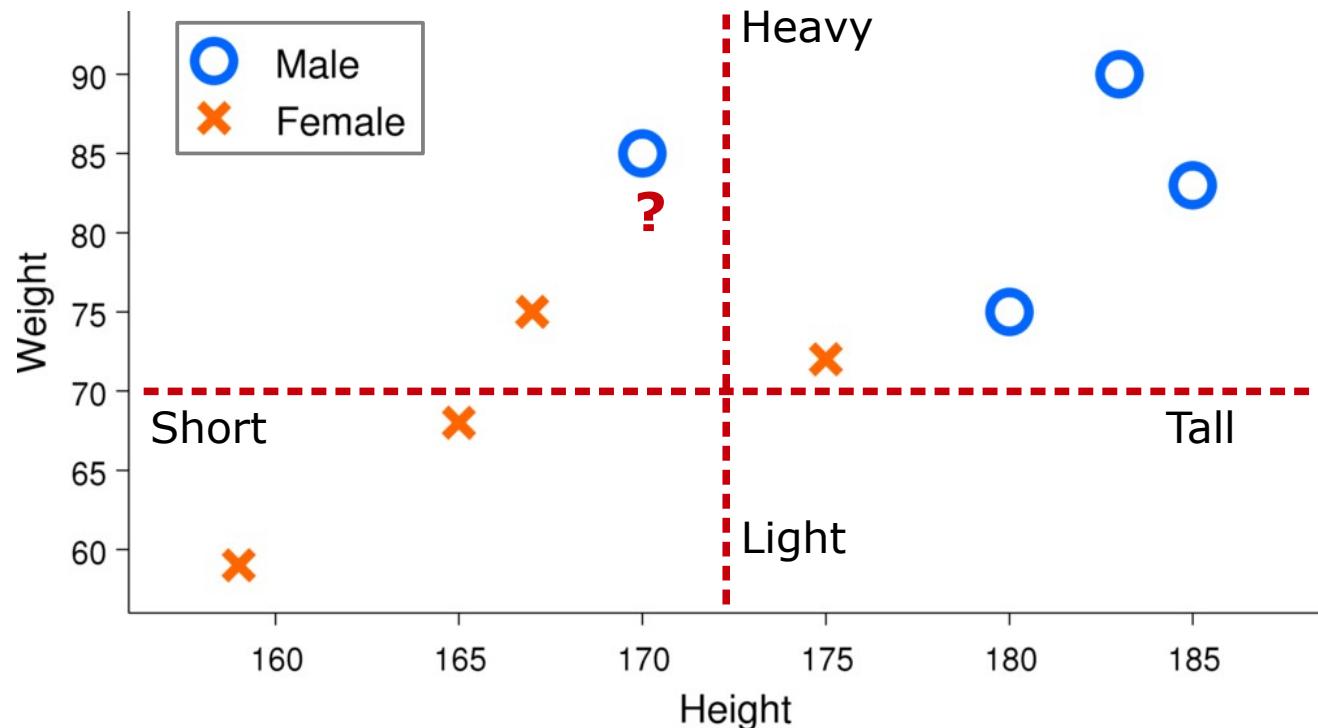
Bayesian classifiers

- What is the probability that ? is male

$$p(\text{Gender} = \text{Male} | \text{Height} = \text{Short}, \text{Weight} = \text{Heavy})$$

- Shorthand notation:

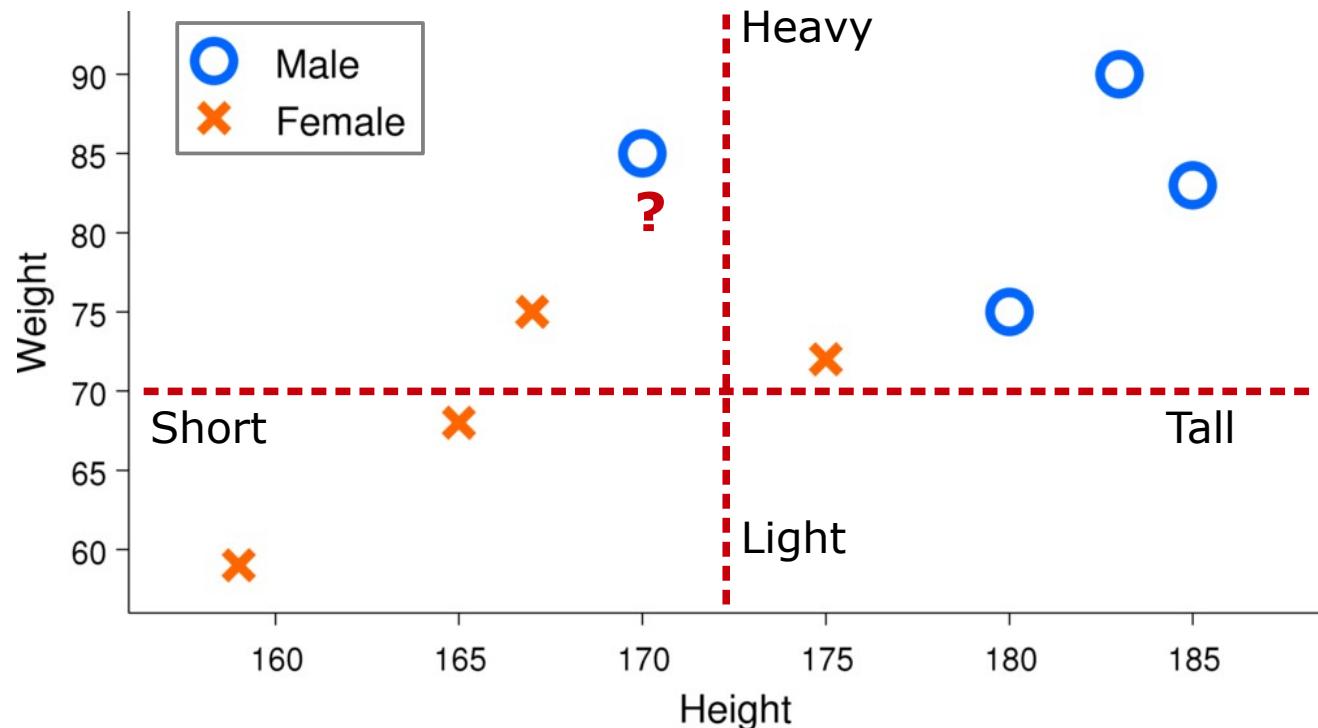
$$p(G = m | H = s, W = h) = p(m | s, h)$$



Bayesian classifiers

- Bayes rule

$$p(m|s, h) = \frac{p(s, h|m)p(m)}{\sum_{G \in \{m, f\}} p(s, h|G)p(G)} = \frac{\frac{1}{4} \cdot \frac{4}{8}}{\frac{1}{4} \cdot \frac{4}{8} + \frac{1}{4} \cdot \frac{4}{8}} = \frac{1}{2}$$



Bayesian classifiers

- **Contingency table**

- All combinations of attribute values
- Huge table



	Height	Weight	Gender
1	Tall	Heavy	Male
2	Tall	Heavy	Male
3	Short	Heavy	Male
4	Tall	Heavy	Male
5	Short	Light	Female
6	Short	Heavy	Female
7	Short	Light	Female
8	Tall	Light	Female

	Gender	Height	Weight	Fraction
Male	Short	Light	0/4	
		Heavy	1/4	
	Tall	Light	0/4	
		Heavy	3/4	
Female	Short	Light	2/4	
		Heavy	1/4	
	Tall	Light	0/4	
		Heavy	1/4	

Bayesian classifiers

- Naïve Bayes assumption
 - Conditional probabilities of attributes are independent

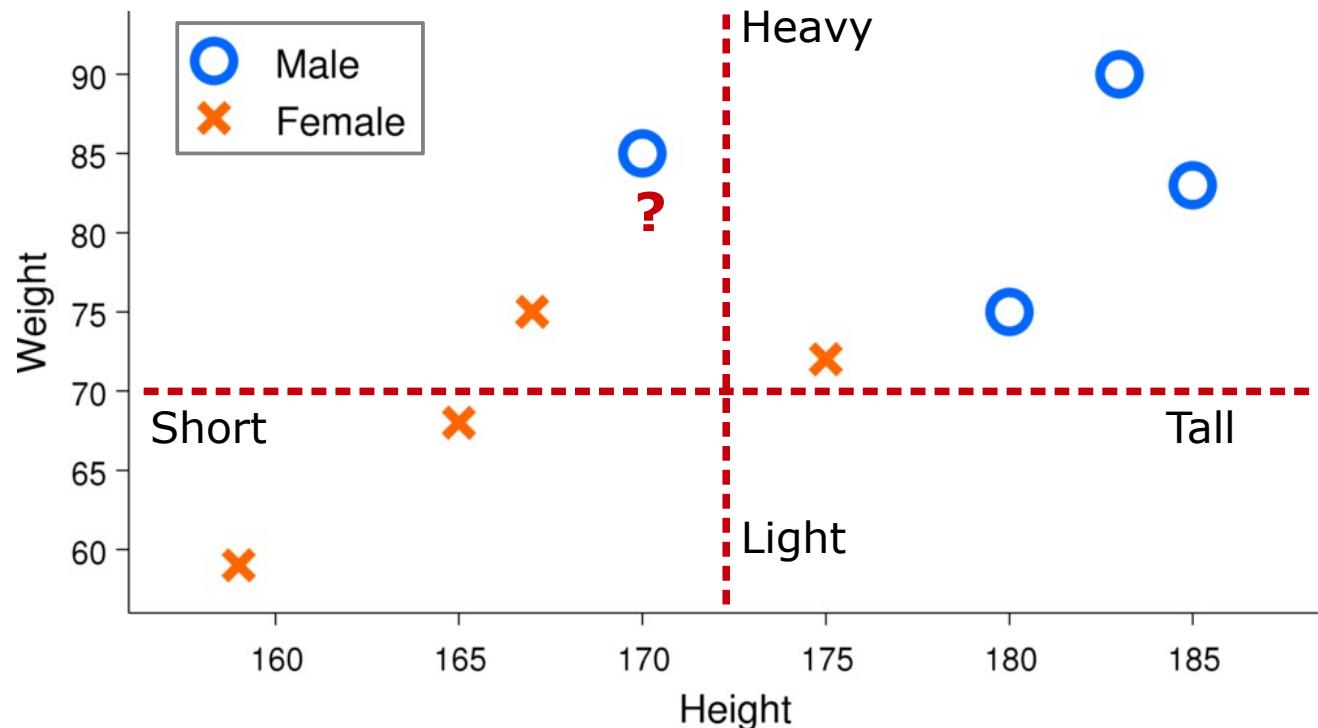
$$p(\text{Height, Weight}|\text{Gender}) = p(\text{Height}|\text{Gender}) \times p(\text{Weight}|\text{Gender})$$

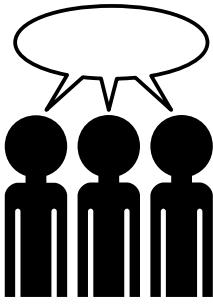
Bayesian classifiers



- Naïve Bayes classifier

$$p(m|s, h) = \frac{p(s|m)p(h|m)p(m)}{\sum_{G \in \{m, f\}} p(s|G)p(h|G)p(G)} = \frac{\frac{1}{4} \cdot \frac{4}{4} \cdot \frac{4}{8}}{\frac{1}{4} \cdot \frac{4}{4} \cdot \frac{4}{8} + \frac{3}{4} \cdot \frac{2}{4} \cdot \frac{4}{8}} = \frac{2}{5}$$





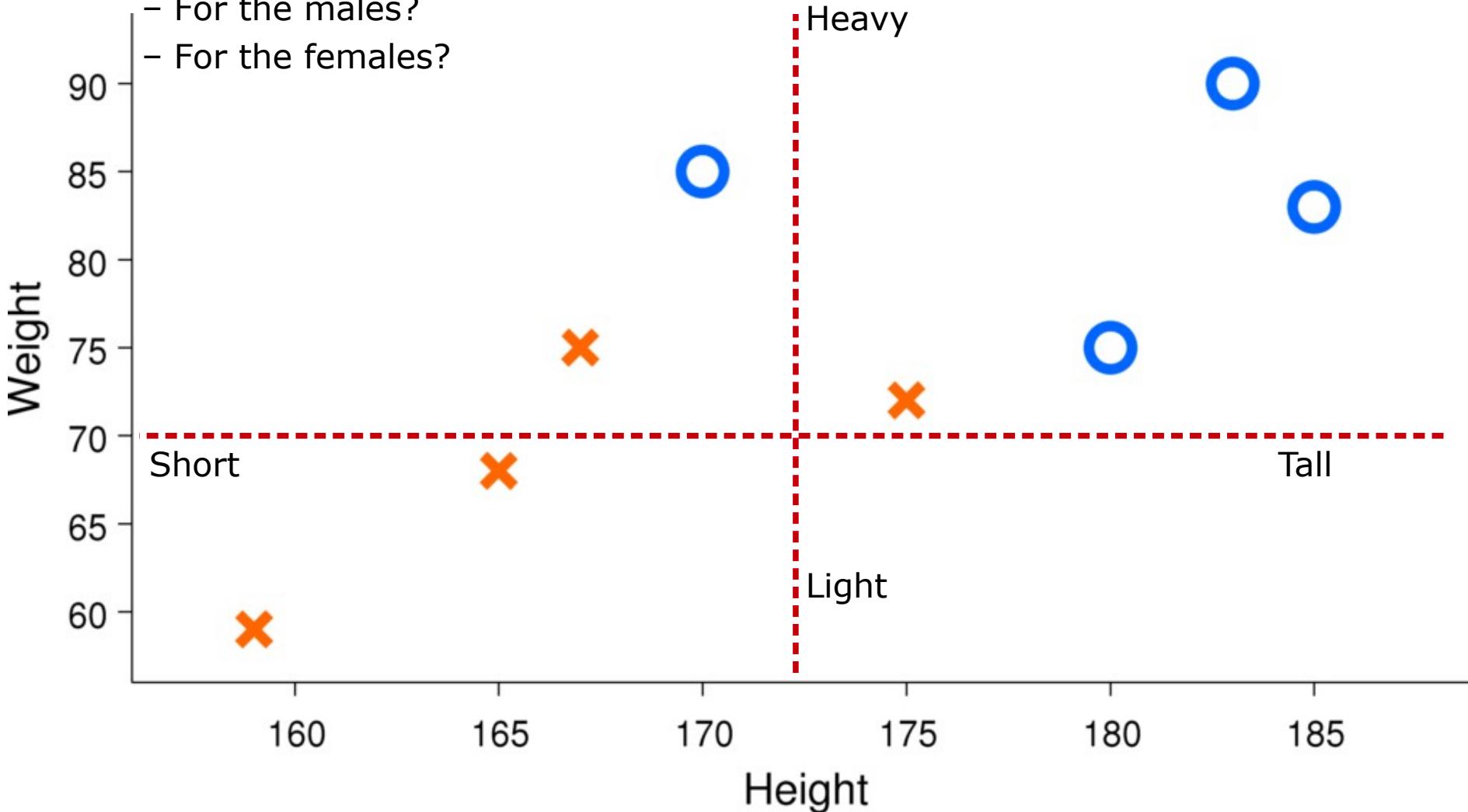
Naïve Bayes assumption

$$p(\text{Height}, \text{Weight}|\text{Gender}) = p(\text{Height}|\text{Gender}) \times p(\text{Weight}|\text{Gender})$$



Group exercise

- Does the naïve Bayes assumption hold empirically
 - For the males?
 - For the females?



Bayesian classifiers

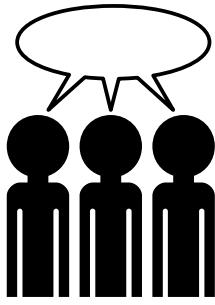
- **Naïve Bayes contingency table**

- Only counts for each attribute
- Small table

	Height	Weight	Gender
1	Tall	Heavy	Male
2	Tall	Heavy	Male
3	Short	Heavy	Male
4	Tall	Heavy	Male
5	Short	Light	Female
6	Short	Heavy	Female
7	Short	Light	Female
8	Tall	Light	Female



Gender	Attribute	Fraction
Male	Height=Short	1/4
	Weight=Light	0/4
Female	Height=Short	3/4
	Weight=Light	2/4



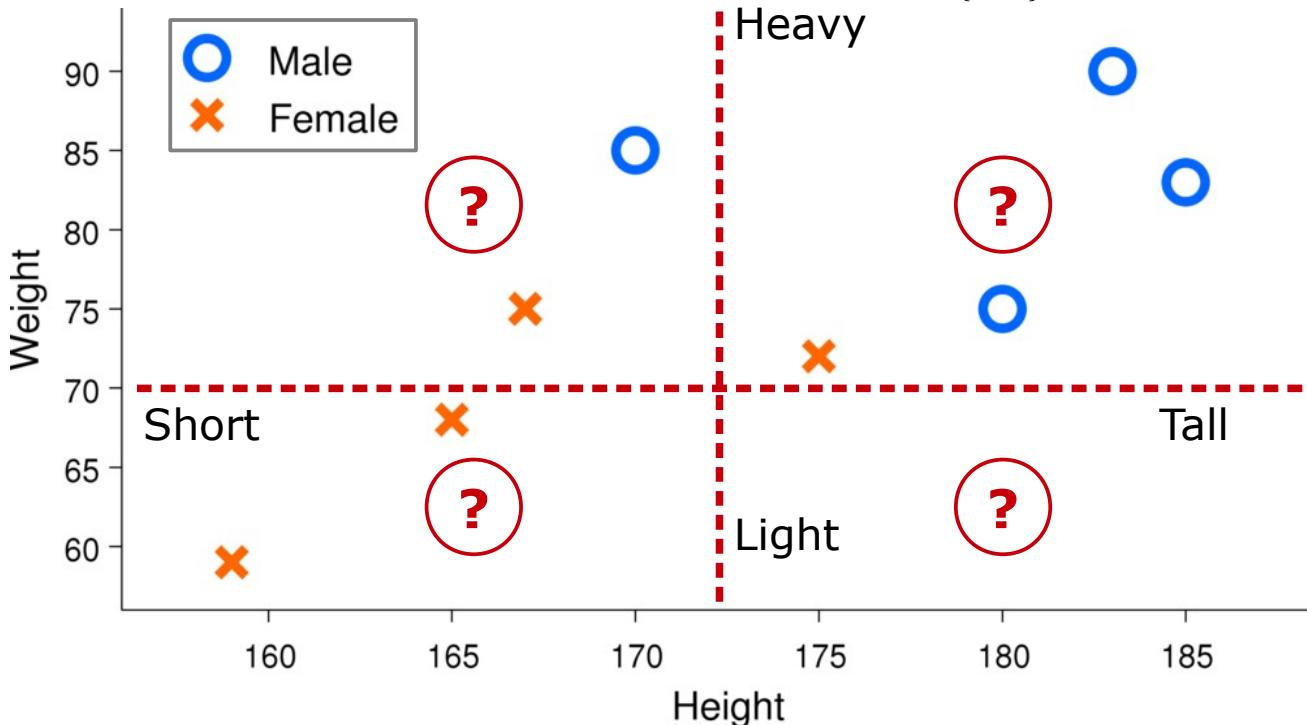
Group exercise

Bayes classifiers

- Classify (compute the posterior probability of G=m) for the four ? using
 - Bayes classifier

$$p(m|s, h) = \frac{p(s, h|m)p(m)}{\sum_{G \in \{m, f\}} p(s, h|G)p(G)}$$

$$p(m|s, h) = \frac{p(s|m)p(h|m)p(m)}{\sum_{G \in \{m, f\}} p(s|G)p(h|G)p(G)}$$

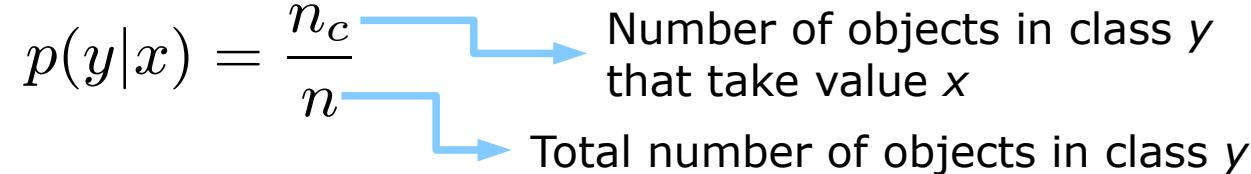


Gender	Attribute	Fraction
Male	Height = Short	1/4
	Weight = Light	0/4
Female	Height = Short	3/4
	Weight = Light	2/4

Gender	Height	Weight	Fraction
Male	Short	Light	0/4
	Short	Heavy	1/4
	Tall	Light	0/4
	Tall	Heavy	3/4
Female	Short	Light	2/4
	Short	Heavy	1/4
	Tall	Light	0/4
	Tall	Heavy	1/4

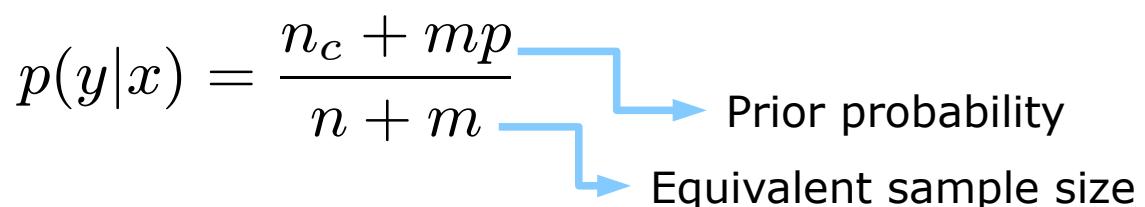
M-estimate

- Probability of class given attribute

$$p(y|x) = \frac{n_c}{n}$$


Number of objects in class y
that take value x

Total number of objects in class y

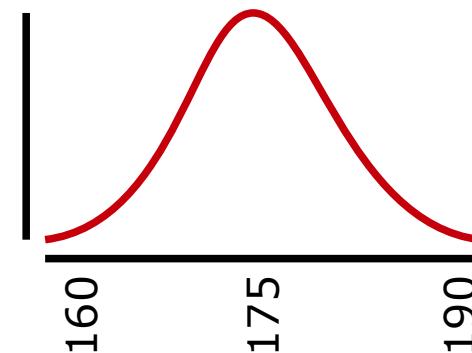
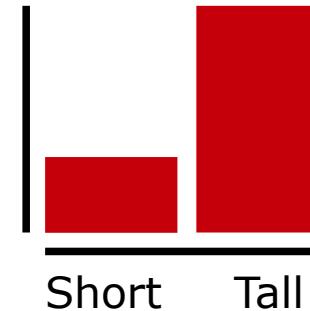
- Not defined when $n=0$
 - M-estimate
- $$p(y|x) = \frac{n_c + mp}{n + m}$$
- 
- Prior probability
- Equivalent sample size
- If no objects in class y take value x the probability will be p
 - Corresponds to putting m extra objects into the data set

Bayesian classifiers



- Handling continuous attributes
 - Two way split ($x < a$)
 - Converts into binary attribute
(We have used this in the previous example)
 - Discretize into a number of bins
 - Converts into discrete ordinal attribute
 - Probability density estimation
 - Assume attribute follows a Normal distribution
 - Use data to compute parameters
(mean and variance)

$$p(\text{Height} | \text{Gender} = \text{Male})$$

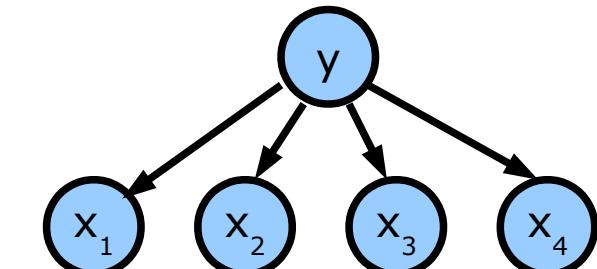


Baysian Belief Networks (BBN)

- Independence assumption may not hold for some attributes (use BBN)

Naïve Bayes

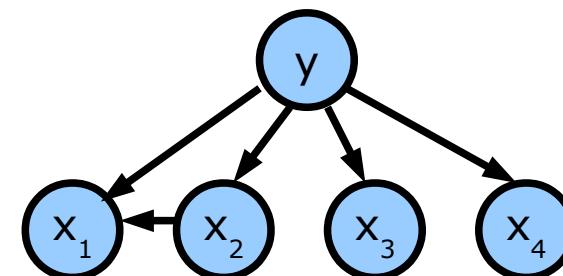
$$p(\mathbf{X}|y) = p(x_1|y)p(x_2|y)p(x_3|y)p(x_4|y)$$



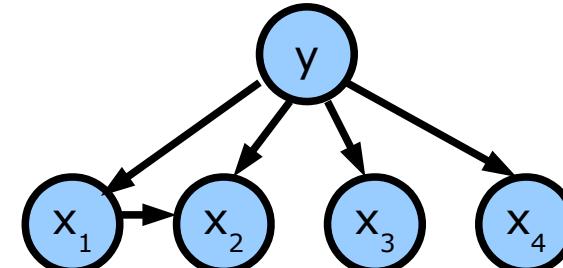
$$p(x_1|y) \quad p(x_2|y) \quad p(x_3|y) \quad p(x_4|y)$$

When x_1 and x_2 are not independent given y

$$\begin{aligned} p(\mathbf{X}|y) &= p(x_1, x_2|y)p(x_3|y)p(x_4|y) \\ &= p(x_1|x_2, y)p(x_2|y)p(x_3|y)p(x_4|y) \end{aligned}$$



$$= p(x_2|x_1, y)p(x_1|y)p(x_3|y)p(x_4|y)$$



Remember basic rules of probability

- Sum rule

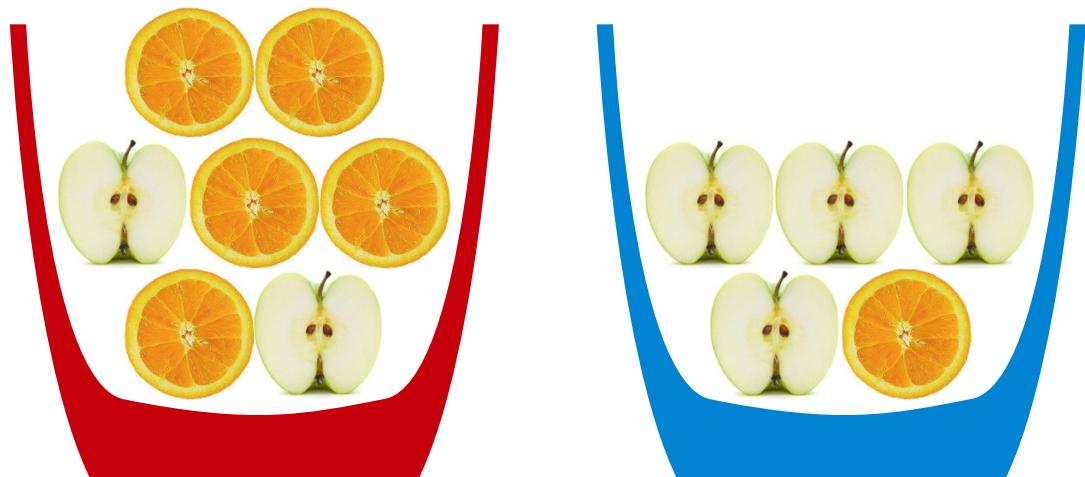
$$p(x) = \sum_y p(x, y)$$

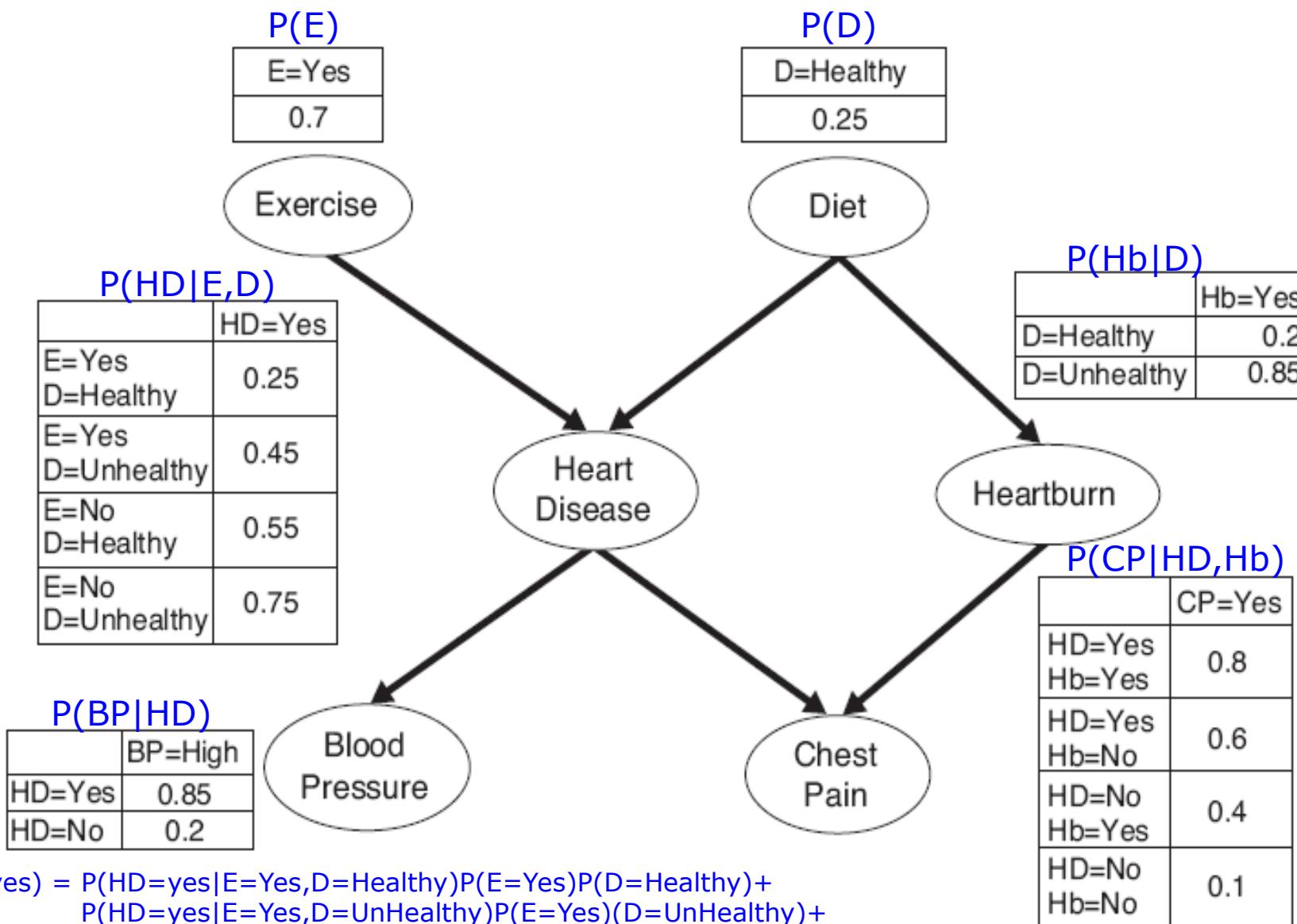
- Product rule

$$p(x, y) = p(x|y)p(y)$$

- Bayes' rule

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)}$$





$$\begin{aligned}
 P(HD=yes) &= P(HD=yes|E=Yes, D=Healthy)P(E=Yes)P(D=Healthy) + \\
 &\quad P(HD=yes|E=Yes, D=UnHealthy)P(E=Yes)(D=UnHealthy) + \\
 &\quad P(HD=yes|E=No, D=Healthy)P(E=No)P(D=Healthy) + \\
 &\quad P(HD=yes|E=No, D=UnHealthy)p(E=No)P(D=UnHealthy) = 0.49
 \end{aligned}$$

$$\begin{aligned}
 p(BP=High) &= P(BP=High|HD=Yes)P(HD=yes) + P(BP=High|HD=No)P(HD=No) \\
 &= 0.85 \cdot 0.49 + 0.2 \cdot (1 - 0.49) = 0.5185
 \end{aligned}$$

$$\begin{aligned}
 P(HD=yes|BP=High) &= P(BP=High|HD=Yes)P(HD=Yes)/P(BP=High) \\
 &= 0.85 \cdot 0.49 / 0.5185 = 0.8033
 \end{aligned}$$

$$\begin{aligned}
 P(HD=yes|BP=High, D=Healthy, E=yes) &= P(BP=High|HD=Yes)P(HD=Yes|D=Healthy, E=yes)/P(BP=High|D=Healthy, E=yes) \\
 &= 0.85 \cdot 0.25 / (0.85 \cdot 0.25 + 0.2 \cdot (1 - 0.25)) = 0.5862
 \end{aligned}$$

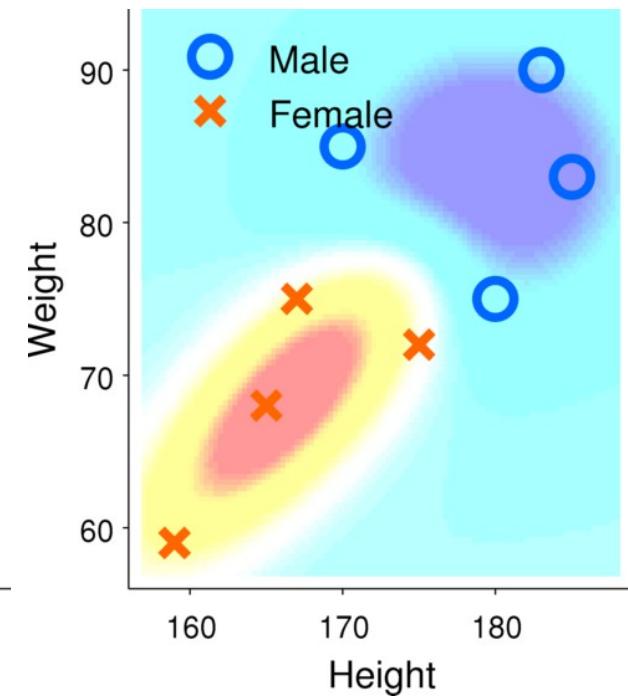
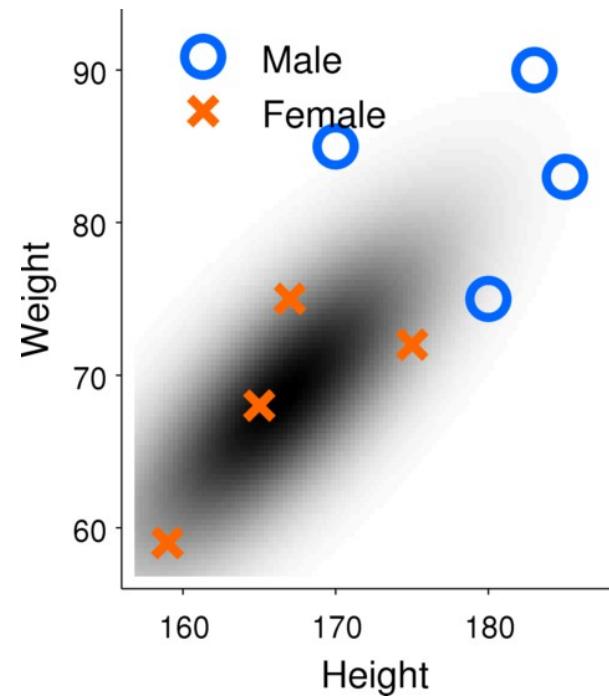
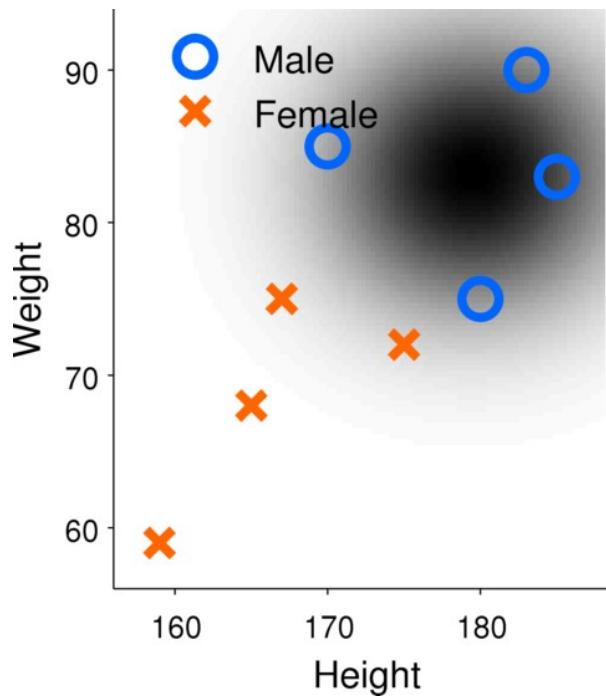
Bayesian classification by the multivariate normal distribution

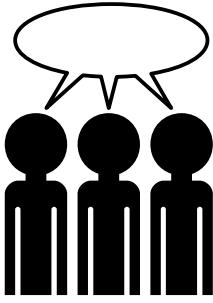
Continuous density estimation

- Fit a Normal distribution to each class
 - Compute class mean and covariance
- Classify using Bayes rule as before

$$P(\mathbf{x}|y = c) = \frac{1}{(2\pi)^{M/2} \det(\Sigma_c)^{\frac{1}{2}}} \exp \left(-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_c)^\top \boldsymbol{\Sigma}_c (\mathbf{x} - \boldsymbol{\mu}_c) \right)$$

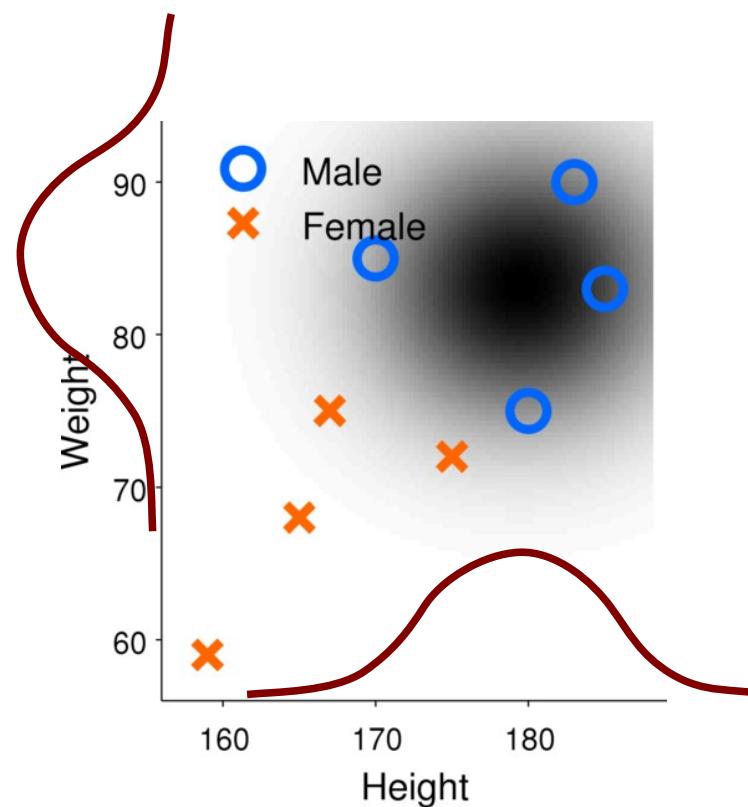
$$P(y = c|\mathbf{x}) = \frac{P(\mathbf{x}|y = c)P(y = c)}{\sum_{c'} P(\mathbf{x}|y = c')P(y = c')}$$





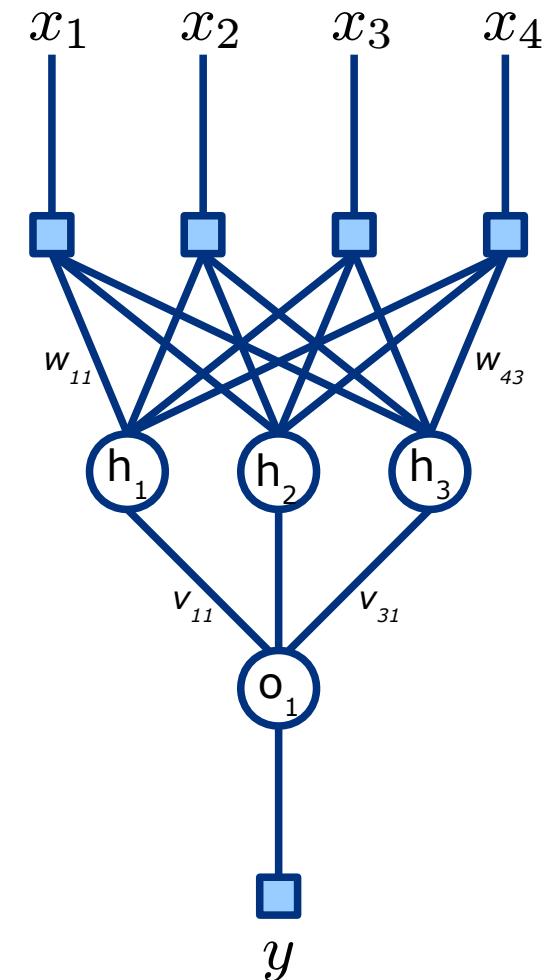
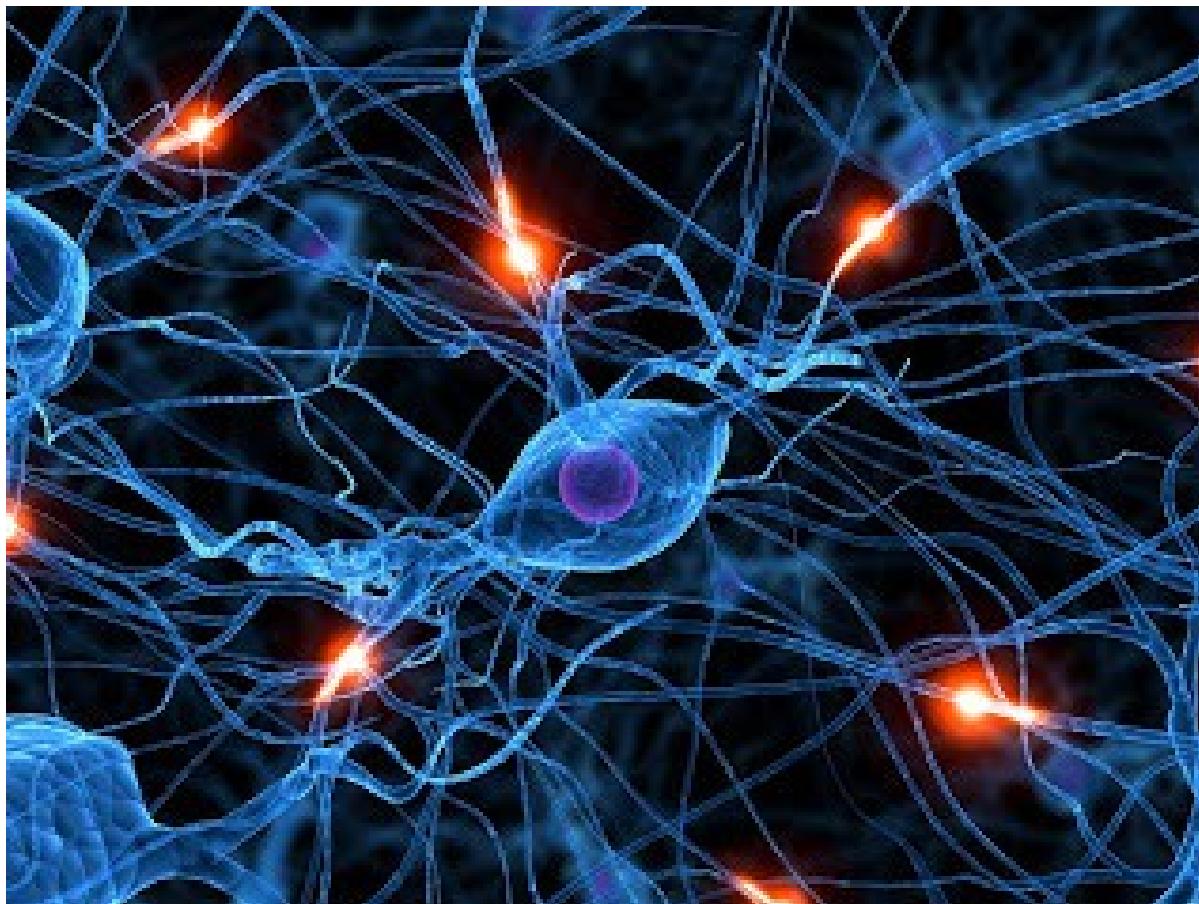
Group exercise

- What does the Naive Bayes assumption correspond to when assuming that each attribute in the multivariate normal distribution is independent of each other?



Artificial neural networks (ANN)

- The human brain contain in the order of 10^{11} neurons with 10^{15} connections
- Artificial Neural Networks are inspired by the architecture of the human brain



Artificial neural networks (ANN)

- Remember the generalized linear model?

– Data

$$\{\mathbf{x}_n, y_n\}_{n=1}^N$$

– Model

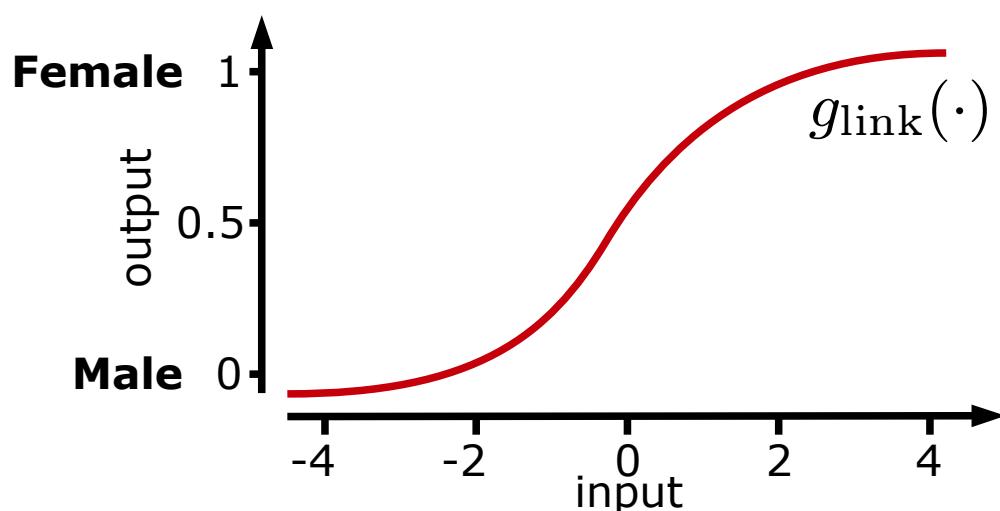
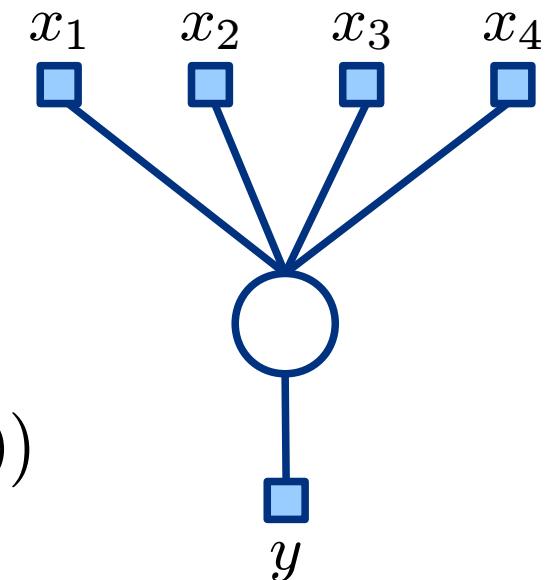
$$f(\mathbf{x}) = g_{\text{link}}(\mathbf{x}^\top \mathbf{w})$$

– Cost function

$$d(y, f(\mathbf{x}))$$

– Parameters

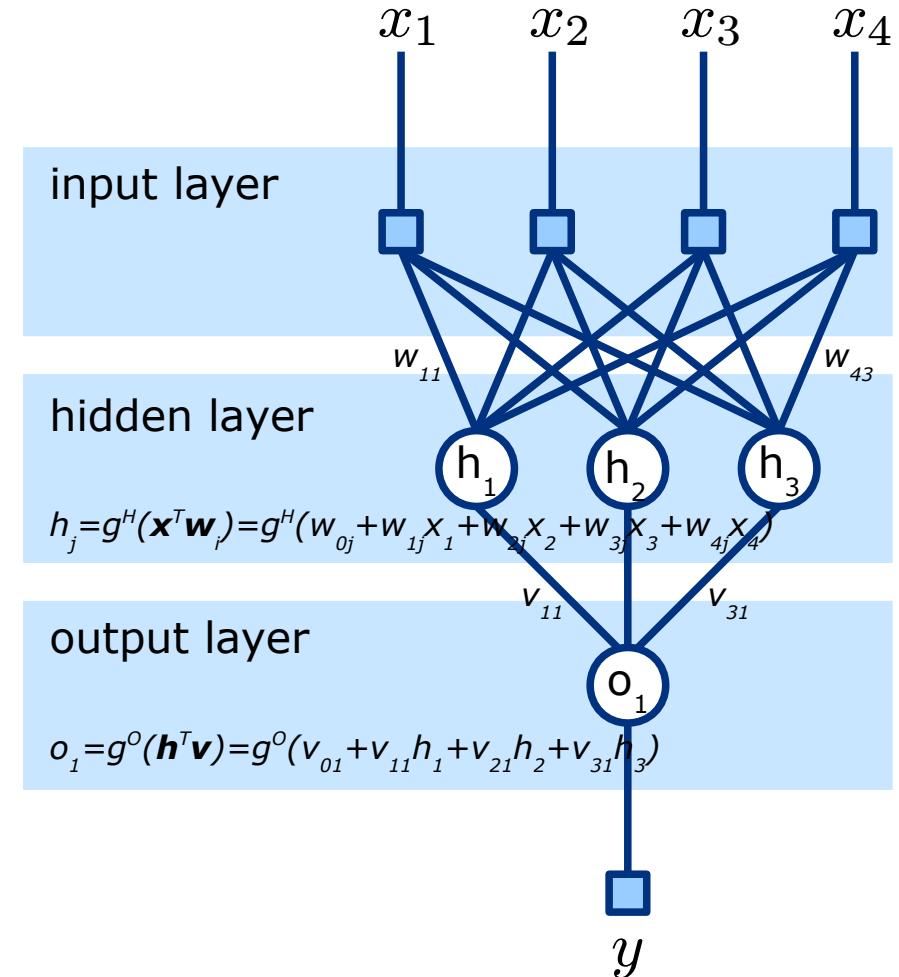
$$\mathbf{w} = \arg \min_{\mathbf{w}} \sum_{n=1}^N d(y_n, f(\mathbf{x}_n))$$



Artificial neural networks

Feed forward network

- Each “neuron”
 - Computes a non-linear function of the sum of its inputs
 - Is just like a generalized linear model
 - Has its own set of parameters
- Modeling choices
 - Cost function
 - Non-linearities
 - Number of neurons and hidden layers
 - Selection of inputs
- Parameter estimation using numerical optimization methods
- Very flexible model: Can easily overfit



Artificial Neural Networks

- The ANN we will consider in the exercises:

- Data

$$\{\mathbf{x}_n, y_n\}_{n=1}^N$$

- Model

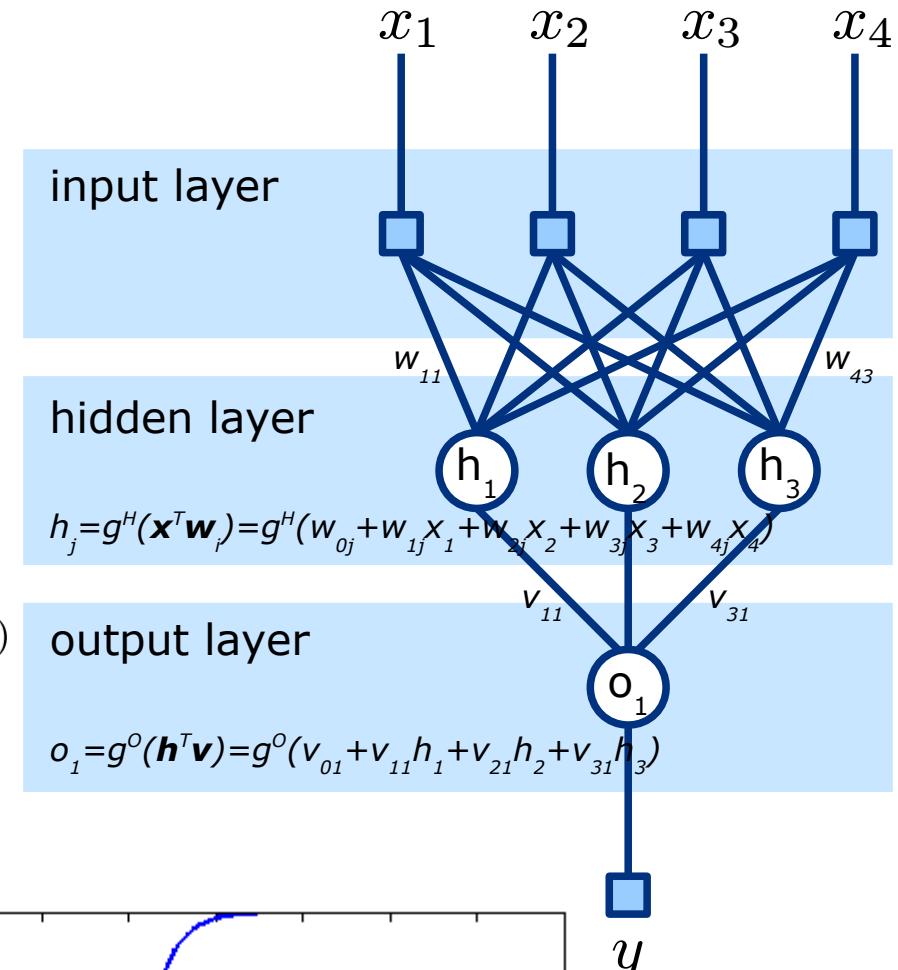
$$f(\mathbf{x}_n) = g^O(v_{10} + \sum_i g^H(\mathbf{x}^\top \mathbf{w}_i))$$

- Cost function

$$d(y, f(\mathbf{x}))$$

- Parameters

$$\mathbf{W}, \mathbf{v} = \arg \min_{\mathbf{W}, \mathbf{v}} \sum_{n=1}^N d(y_n, f(\mathbf{x}_n))$$

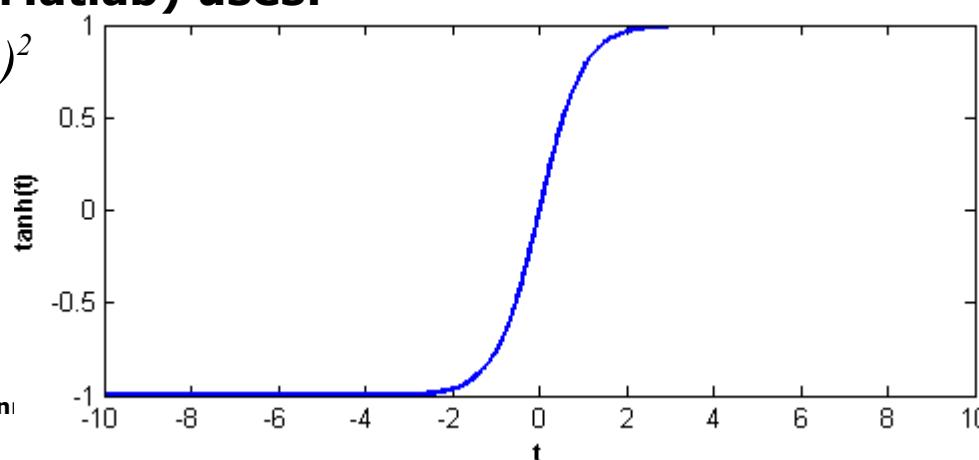


- The implementation (in Matlab) uses:

$$d(y, f(\mathbf{x})) = (y - f(\mathbf{x}))^2$$

$$g^O(t) = t$$

$$g^H(t) = \tanh(t)$$



Midterm practice test

The midterm practice test is used solely for you to test your knowledge and for me to see how well you have understood the covered material so far.

The test **does not** count towards your grade for this course.

- <http://obsurvey.com/S2.aspx?id=E100944D-B068-4B3A-A5E4-52C5F3F631E1>

02450 Introduction to machine learning and data modeling

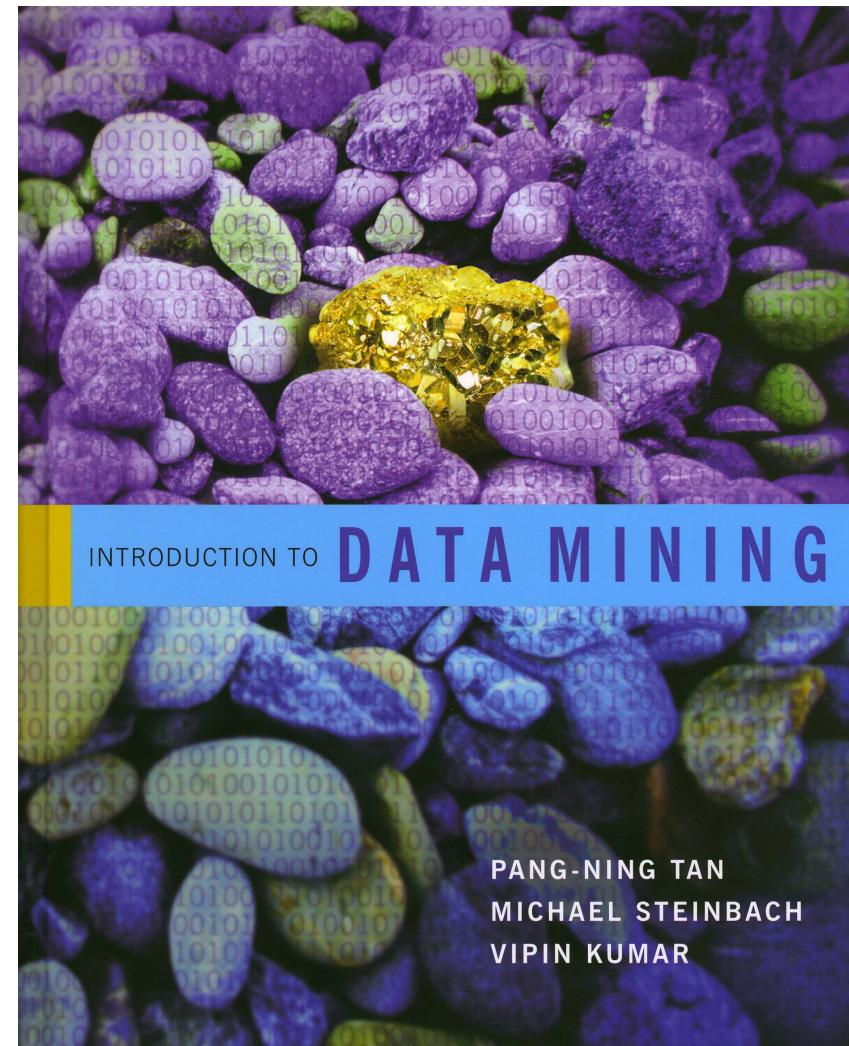
Reading material

Tan, Steinbach and Kumar
"Introduction to Data Mining"

Section 5.6-5.8

Group of the day

Mark Thomas Østerlund
Thor Bech Johannesen
Michael Thygesen Jensen
Mads Friis Hansen
Ferdinando Papale
Albert Fernandez
David Koza
Christian Moesgaard
Sebastian Schock



Lecture schedule

1. Introduction
(Tan 1.1-1.4)

Data: Feature extraction and visualization

2. Data and feature extraction
(Tan 2.1-2.3 + B1 (+ A))
3. Measures of similarity and summary statistics
(Tan 2.4 + 3.1-3.2 + C1-C2)
4. Data visualization
(Tan 3.3)

Supervised learning: Classification and regression

5. Decision trees and linear regression
(Tan 4.1-4.3 + D)
6. Overfitting and performance evaluation
(Tan 4.4-4.6)
7. Nearest neighbor, naive Bayes, and artificial neural networks
(Tan 5.2-5.4)

8. Ensemble methods and multi class classifiers
(Tan 5.6-5.8)

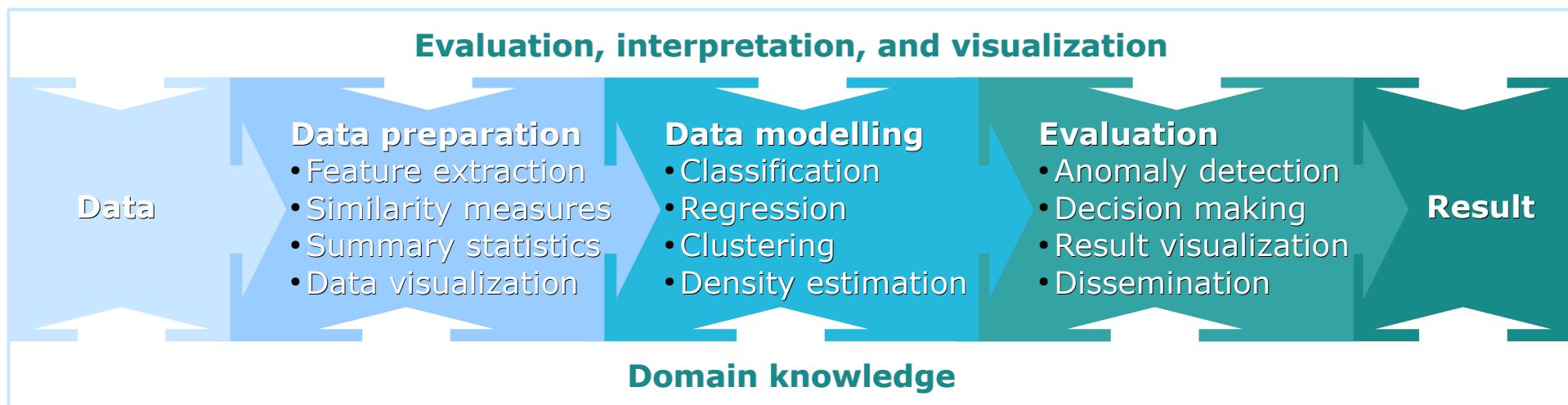
Unsupervised learning: Clustering and density est.

9. K-means and hierarchical clustering
(Tan 8.1-8.3)
10. Mixture models and association mining
(Tan 9.2.2 + 6.1-6.3)
11. Density estimation and anomaly detection
(Tan 10.1-10.4)

Machine learning and data modelling in practice

12. Putting it all together: Summary and overview
13. Mini project

Data modeling framework

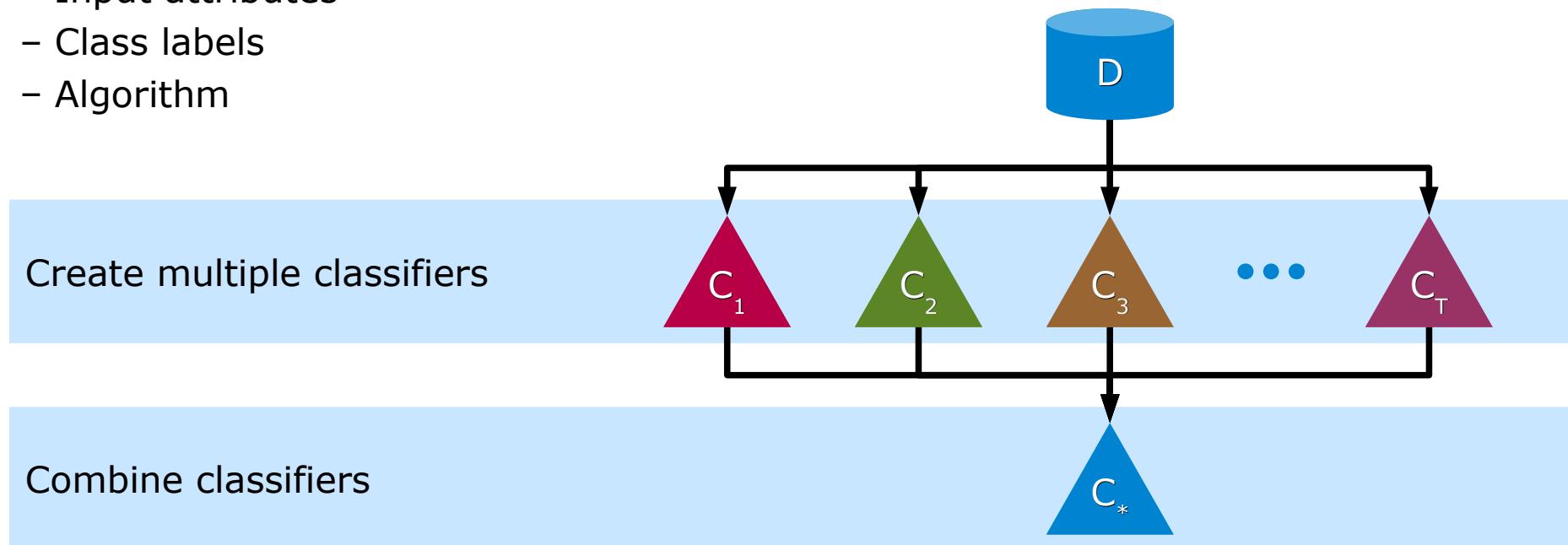


After today you should be able to:

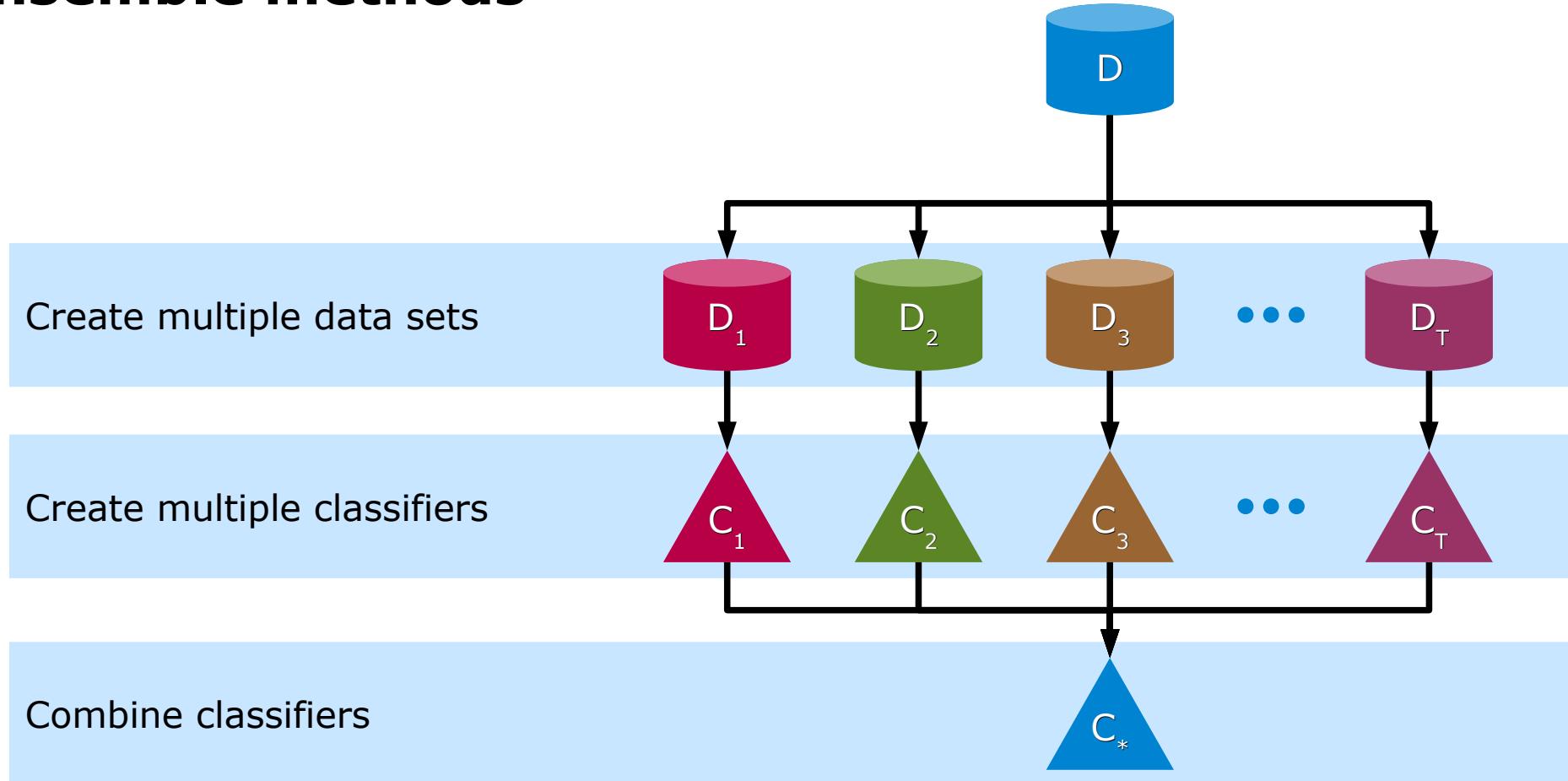
- Explain the principle behind boosting and bagging and apply it to improve classifiers
- Be able to address issues of class-imbalances by resampling
- Understand the definition of Precision, Recall, ROC and AUC
- Be able to extend binary classifiers to multi-class classification

Ensemble methods

- Combine multiple (weak) classifiers into one (strong) classifier
- Each classifier trained using different variations of
 - Data set
 - Input attributes
 - Class labels
 - Algorithm



Ensemble methods



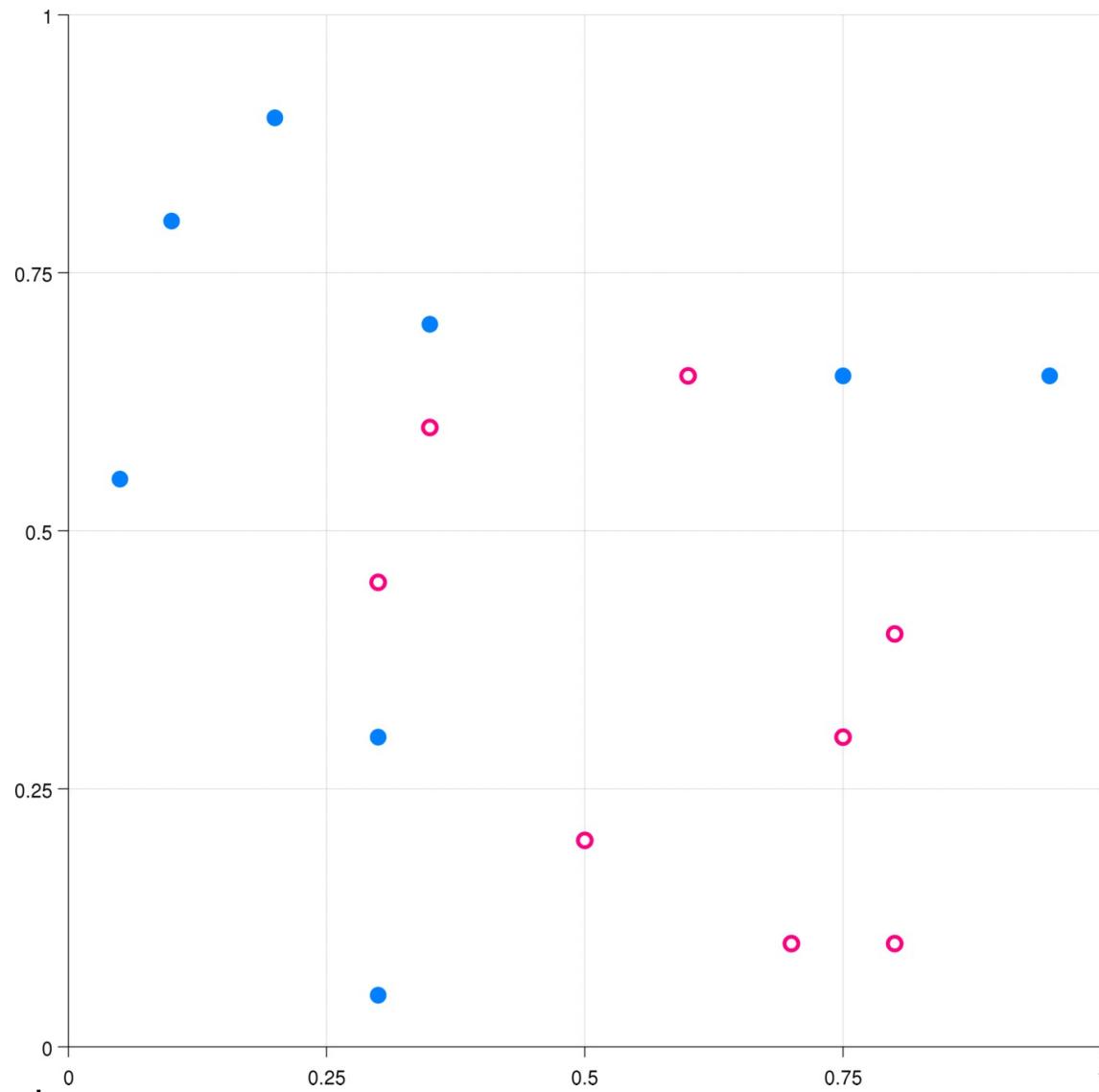
Why ensemble methods?

- Can improve classification algorithms in terms of
 - Better classification accuracy
 - Increased stability
 - Reduced variance
 - Less overfitting
- Consider $N=25$ independent classifiers for binary classification, each with error rate $\epsilon=0.35$

$$\begin{aligned}\epsilon_* &= \sum_{n=\lceil \frac{N}{2} \rceil}^N \binom{N}{n} \epsilon^n (1-\epsilon)^{N-n} \\ &= \sum_{n=13}^{25} \binom{25}{n} 0.35^n (1-0.35)^{25-n} = 0.06\end{aligned}$$

Data example

- Classification using logistic regression

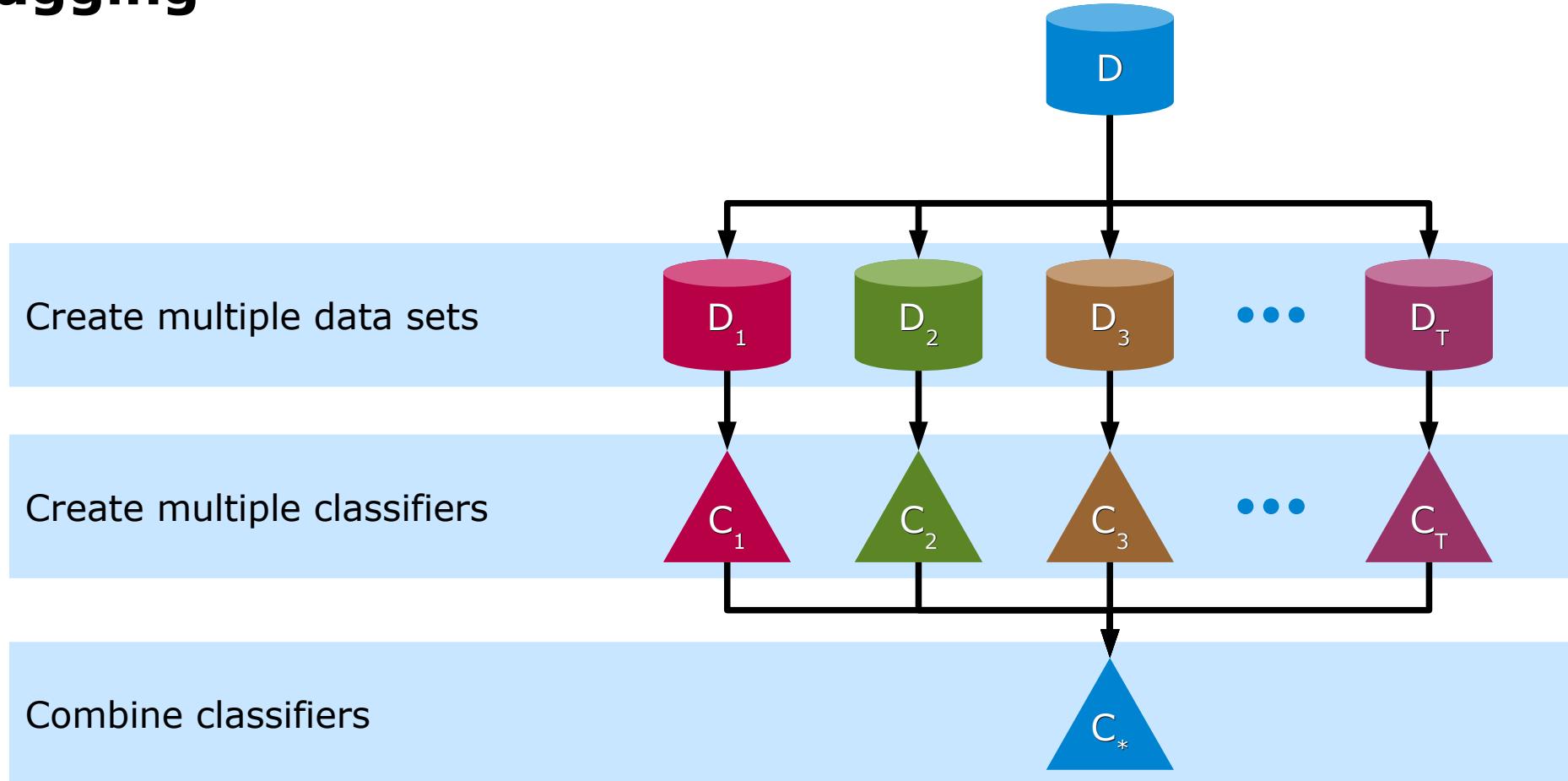


Bagging

- New training data sets drawn randomly from pool with replacement

Pool of training data	1	2	3	4	5	6	7	8	9	10
	3	5	4	3	9	7	9	5	1	1
	5	8	2	6	2	3	8	3	5	1
New training data sets	1	7	4	1	10	6	10	8	8	7
	⋮									
	4	3	8	5	2	4	7	10	10	8

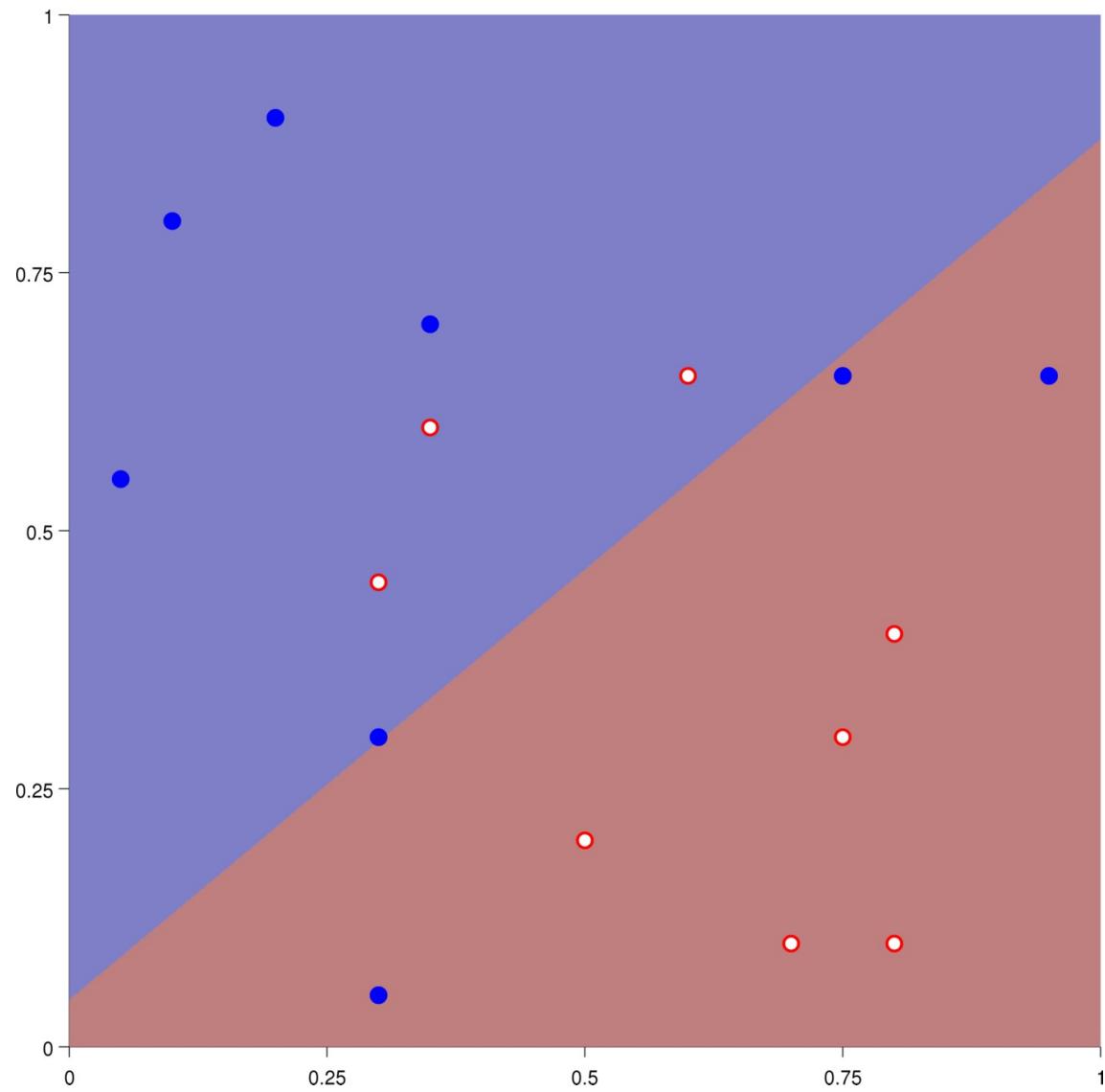
Bagging



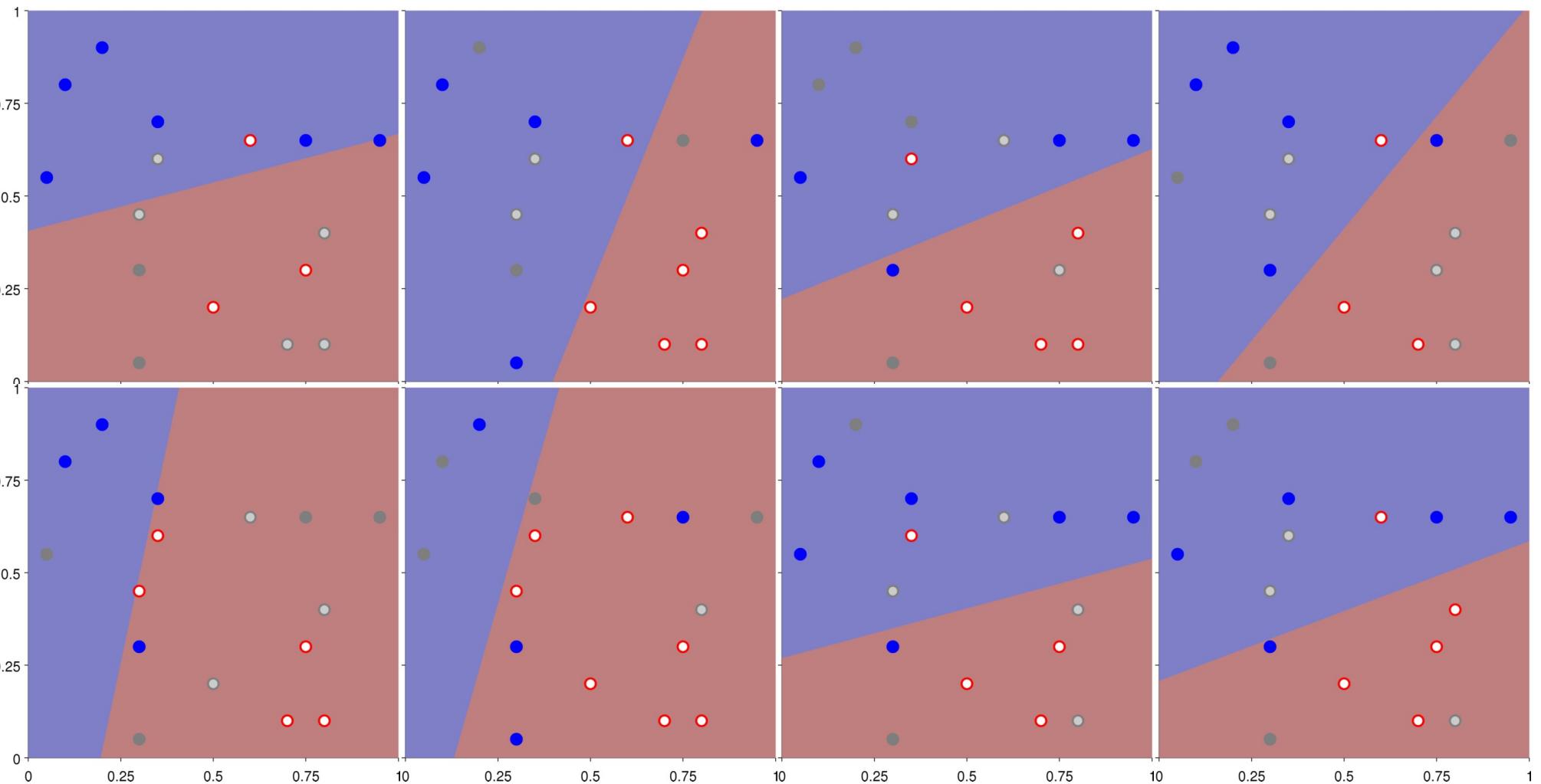
Bagging

- **Single classifier**

- Logistic regression
- Two features, (x, y)



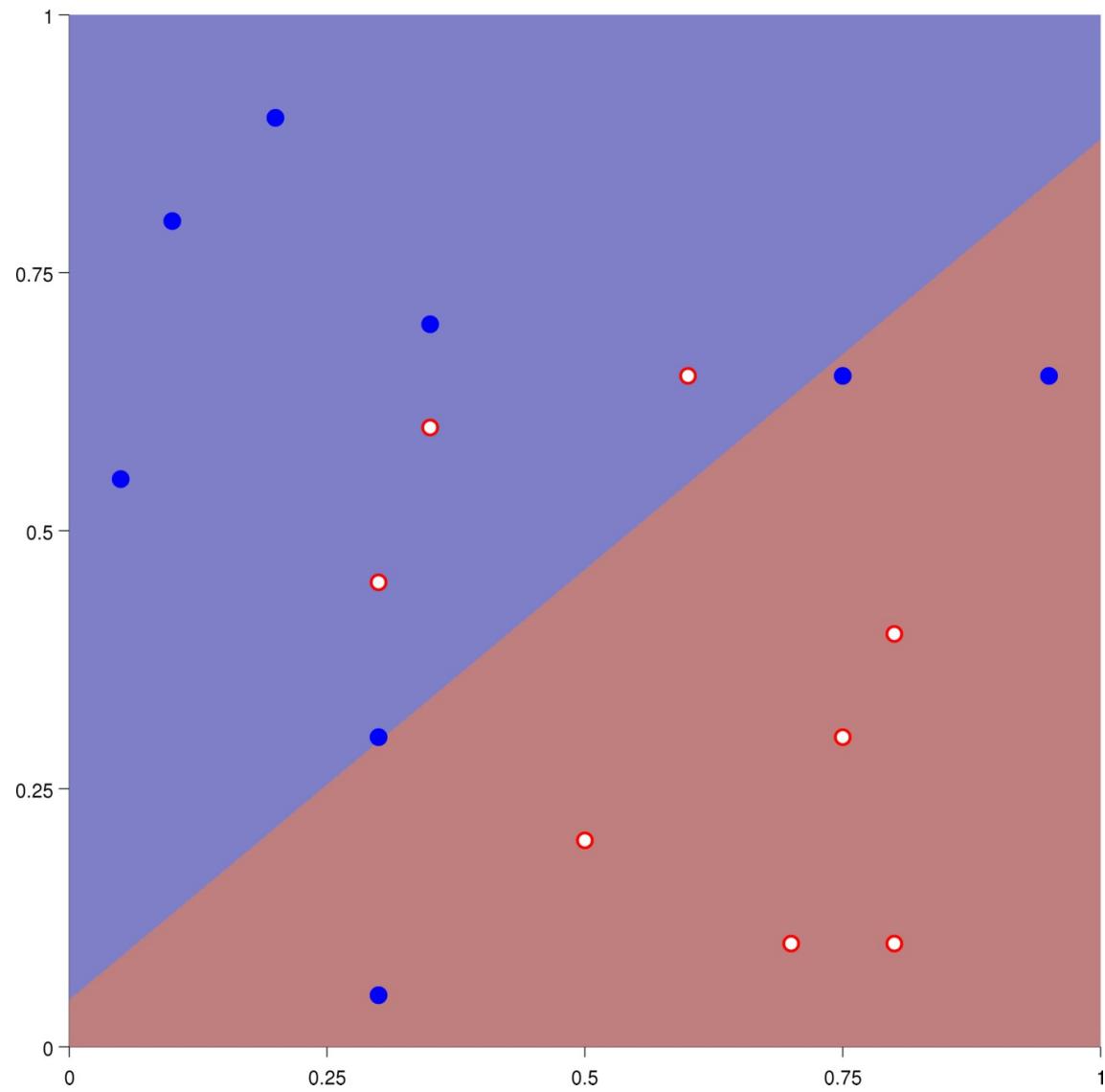
Bagging



Notice, gray dots are observations not included in bagging round

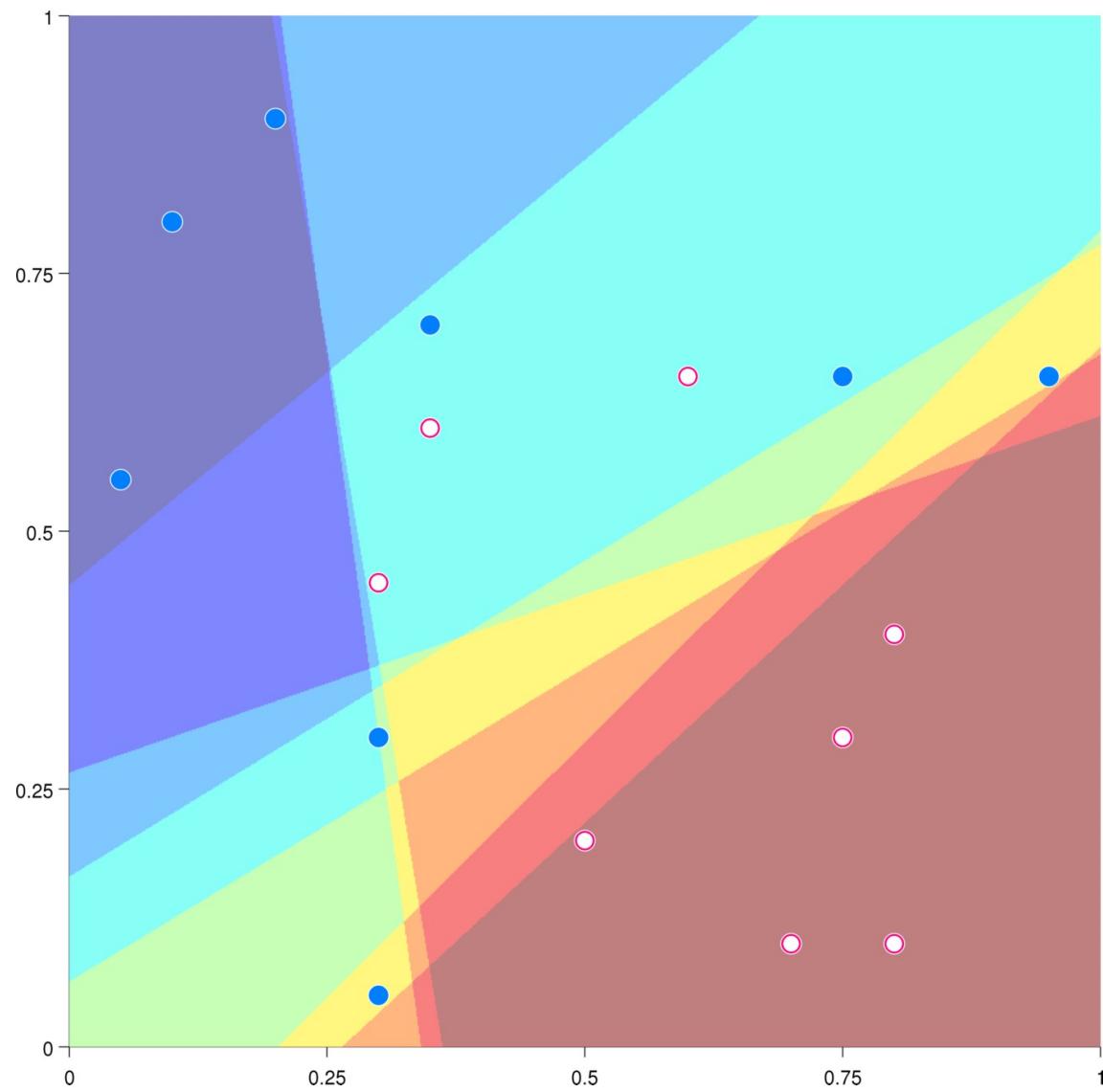
Bagging

- Single classifier



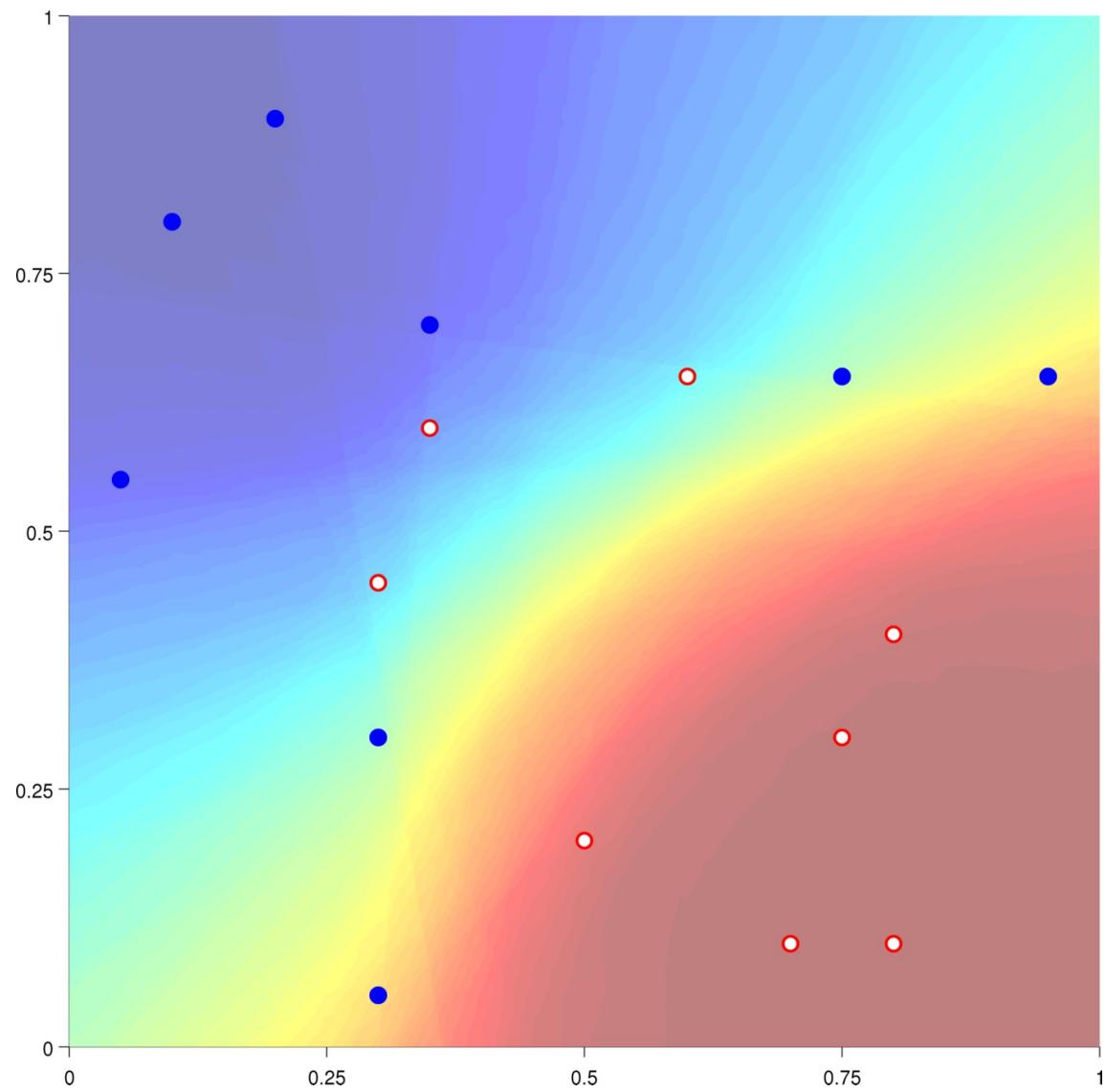
Bagging

- Majority voting of 10 bagged classifiers



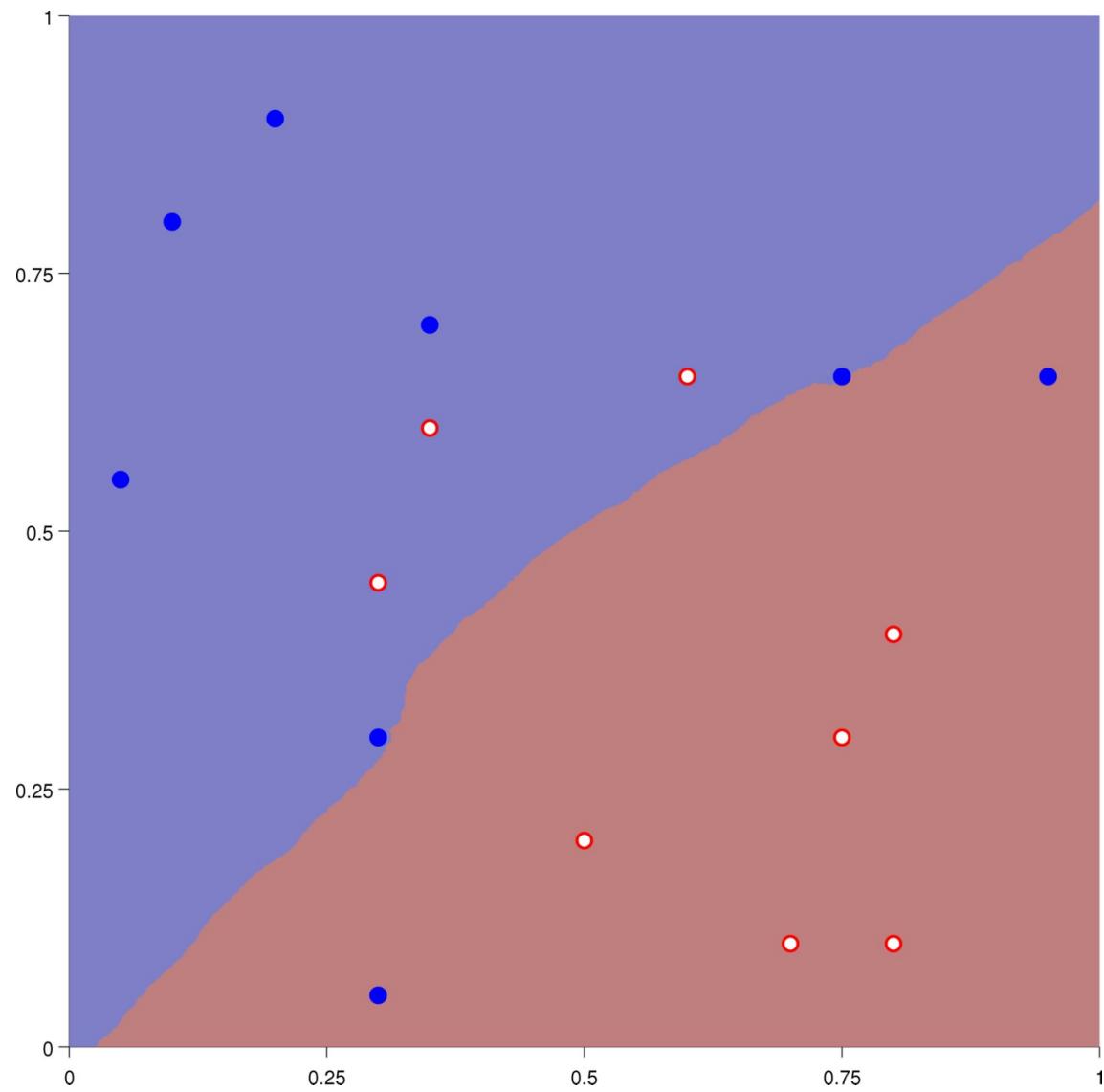
Bagging

- Majority voting of 100 bagged classifiers



Bagging

- Decision boundary of 100 bagged classifier ensemble



Boosting

Pool of training data

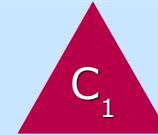
1	2	3	4	5	6	7	8	9	10
.1	.1	.1	.1	.1	.1	.1	.1	.1	.1

Weights

New training data set

3	5	4	3	9	7	9	5	1	1
---	---	---	---	---	---	---	---	---	---

Train classifier



Boosting

Pool of training data	1	2	3	4	5	6	7	8	9	10
Weights	.1	.1	.1	.1	.1	.1	.1	.1	.1	.1
New training data set										
3 5 4 3 9 7 9 5 1 1										
Train classifier	C_1									
Classify all data objects	1✓	2✗	3✓	4✗	5✓	6✗	7✓	8✓	9✓	10✓

Boosting

Pool of training data	1	2	3	4	5	6	7	8	9	10
Weights	.1	.1	.1	.1	.1	.1	.1	.1	.1	.1
New training data set	3 5 4 3 9 7 9 5 1 1									
Train classifier	 C ₁									
Classify all data objects	1✓	2✗	3✓	4✗	5✓	6✗	7✓	8✓	9✓	10✓
Update weights	.07	.17	.07	.17	.07	.17	.07	.07	.07	.07

Boosting

Pool of training data	1	2	3	4	5	6	7	8	9	10
Weights	.1	.1	.1	.1	.1	.1	.1	.1	.1	.1

New training data set	3	5	4	3	9	7	9	5	1	1
-----------------------	---	---	---	---	---	---	---	---	---	---

Train classifier										
------------------	---	--	--	--	--	--	--	--	--	--

Classify all data objects	1✓	2✗	3✓	4✗	5✓	6✗	7✓	8✓	9✓	10✓
Update weights	.07	.17	.07	.17	.07	.17	.07	.07	.07	.07

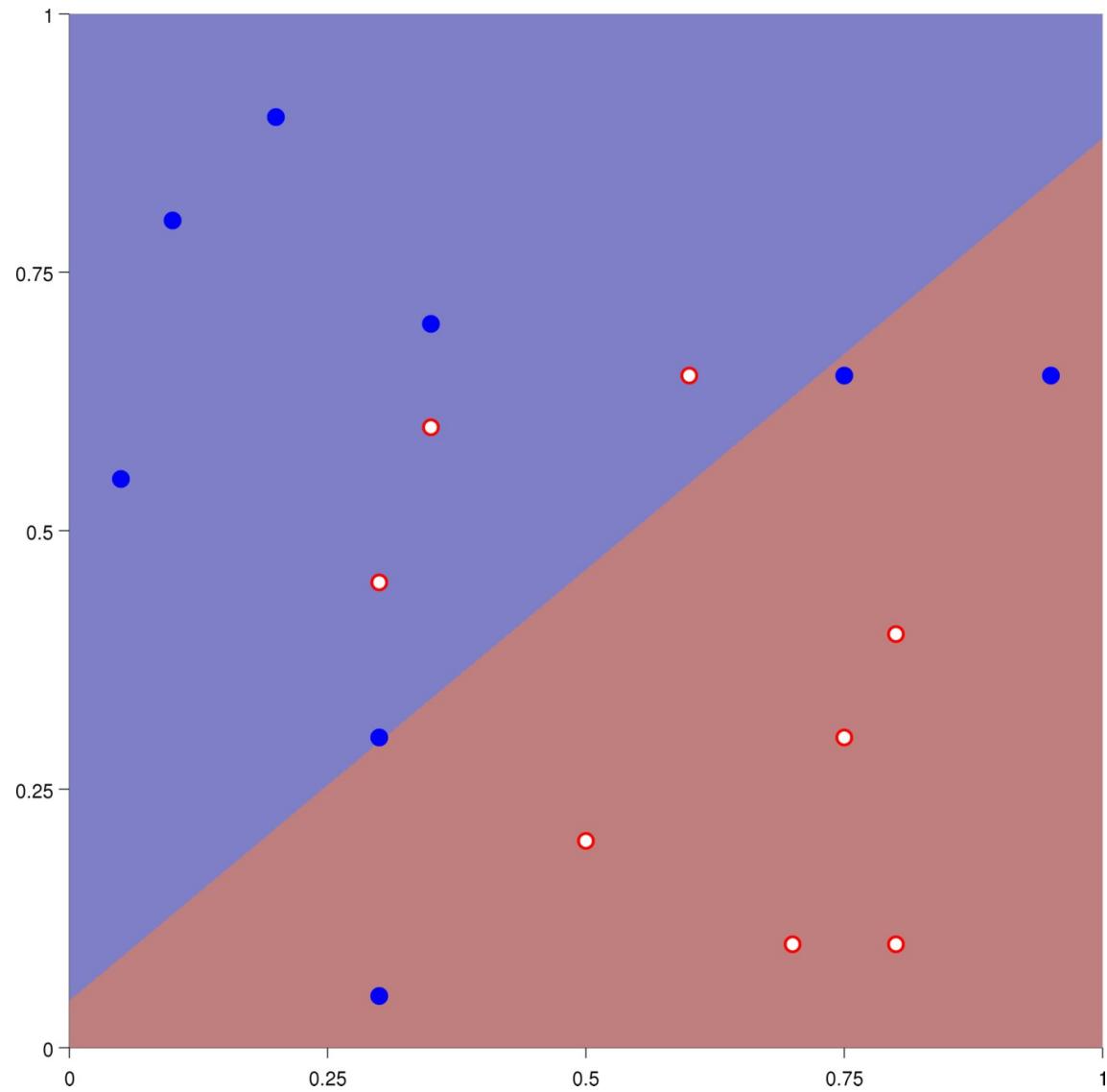
New training data set	6	4	7	3	2	4	10	2	5	6
-----------------------	---	---	---	---	---	---	----	---	---	---

Train classifier										
------------------	---	--	--	--	--	--	--	--	--	--

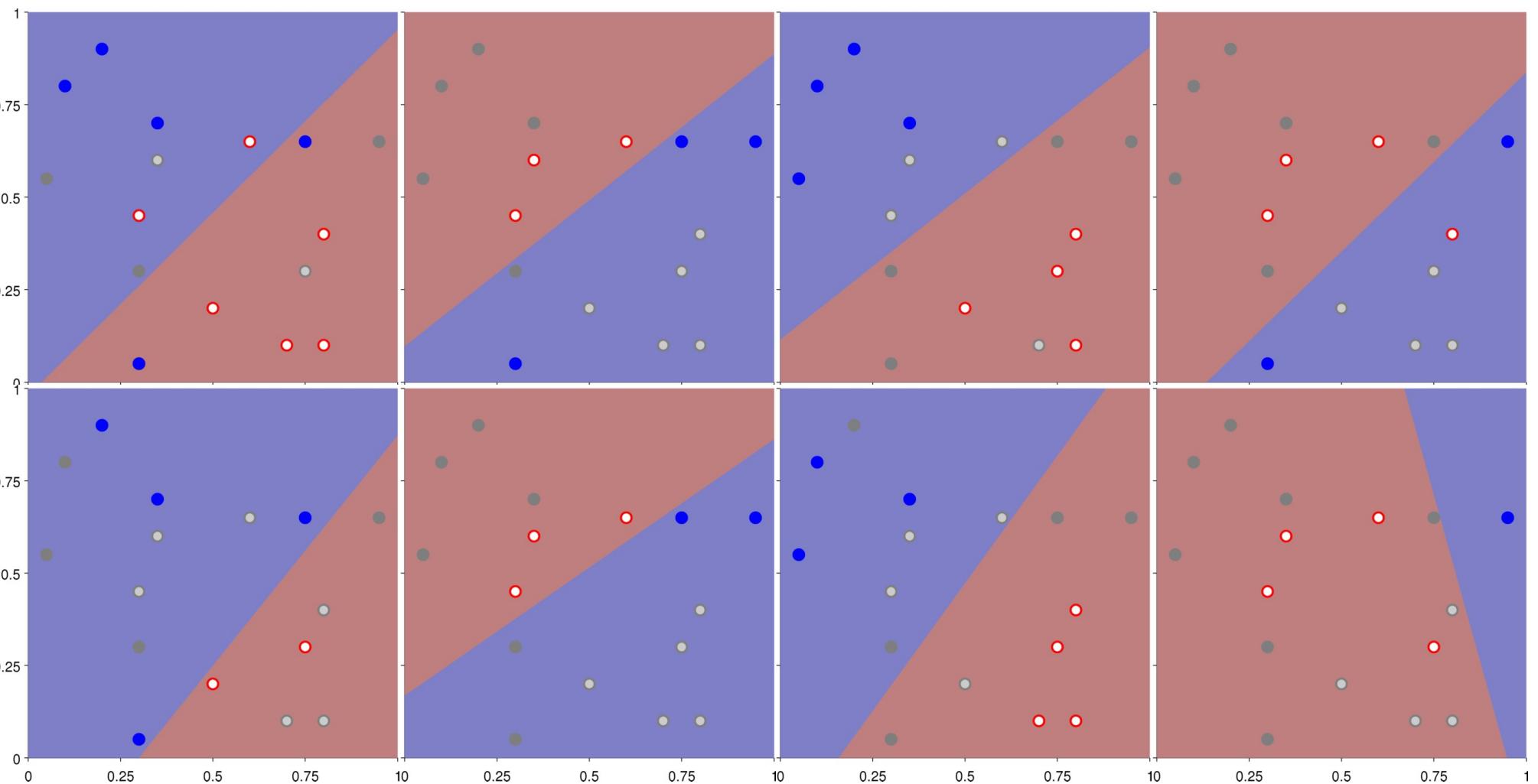


Boosting

- Single classifier



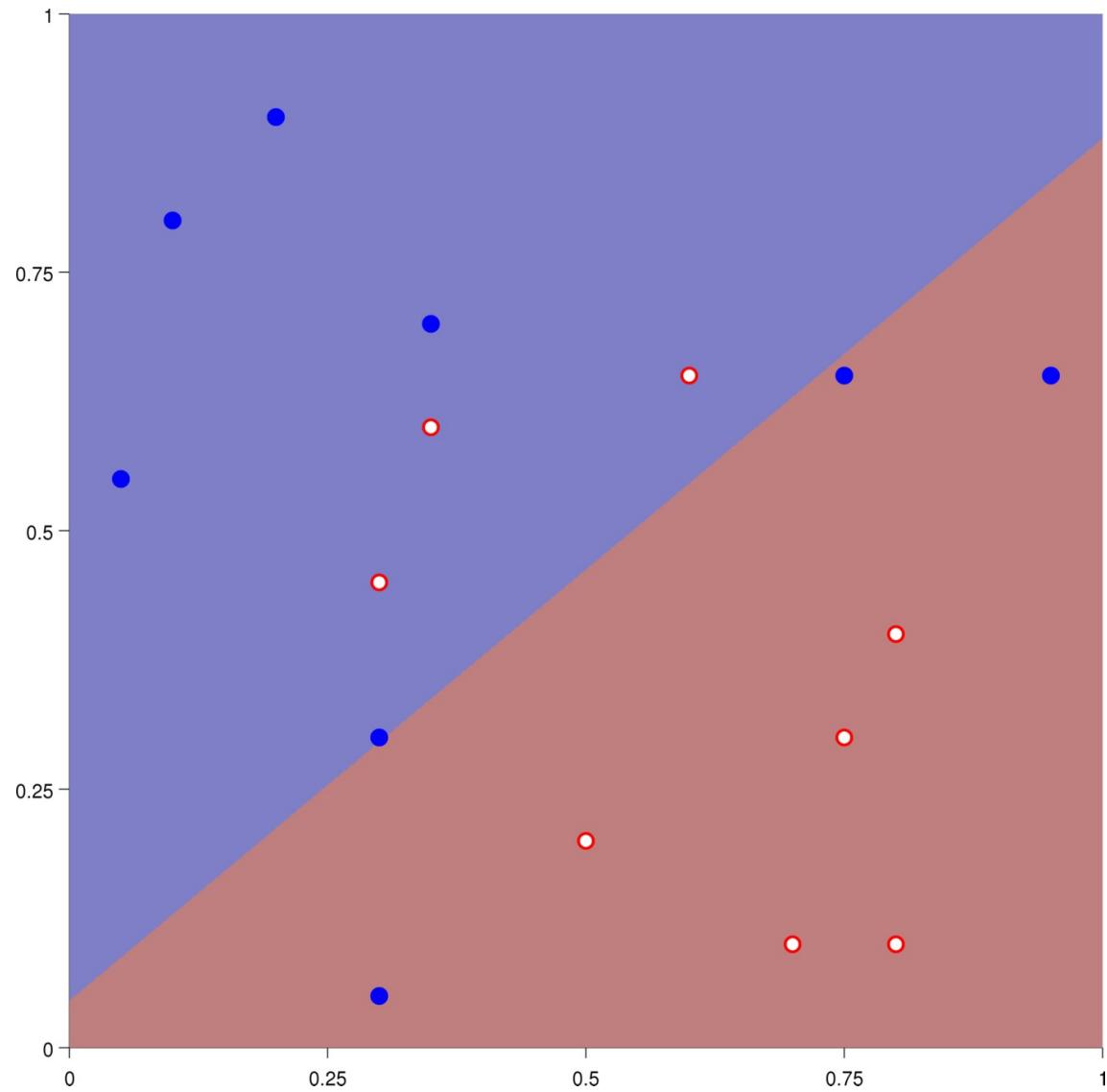
Boosting



Notice, gray dots are observations not included in boosting round

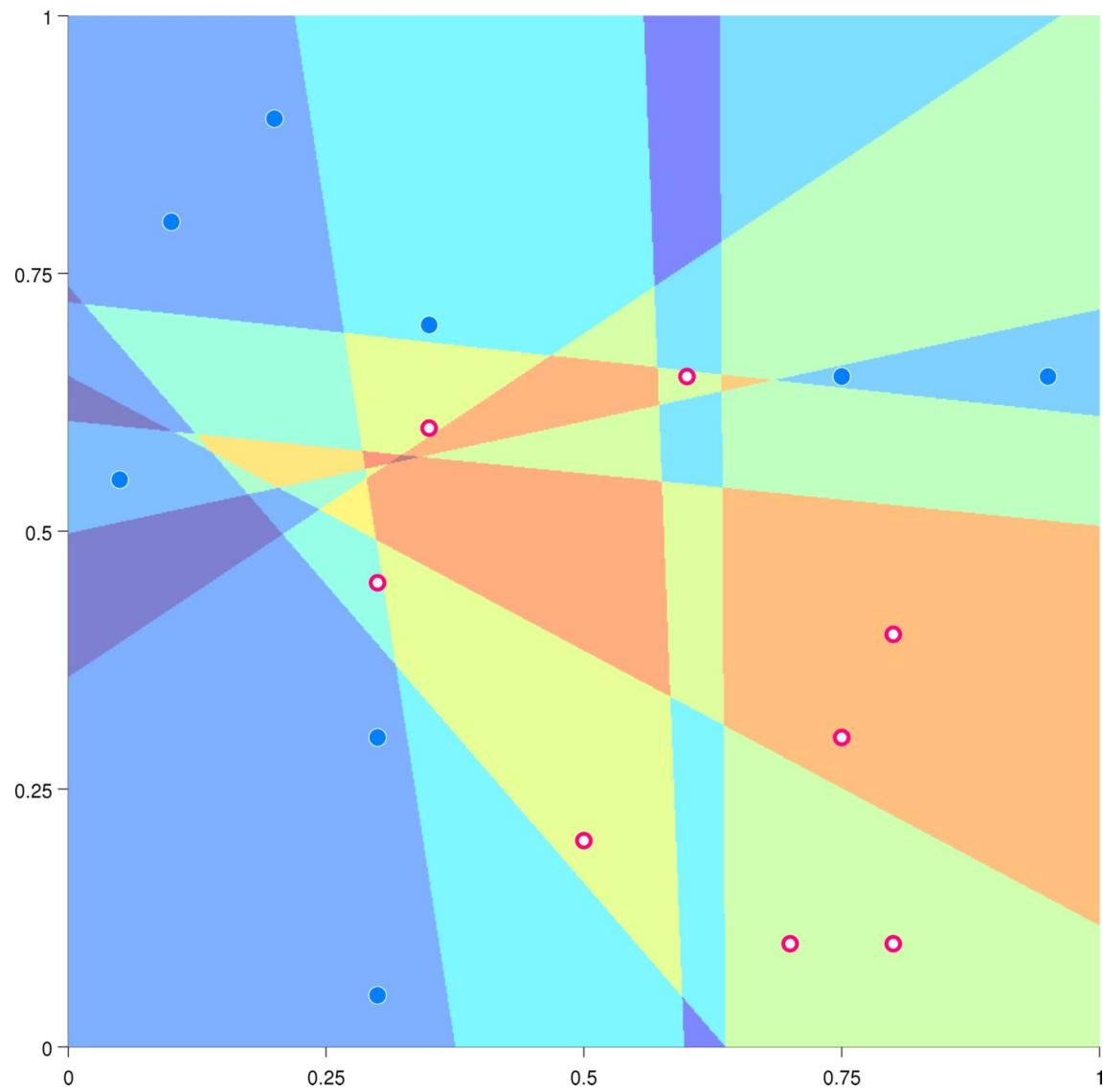
Boosting

- Single classifier



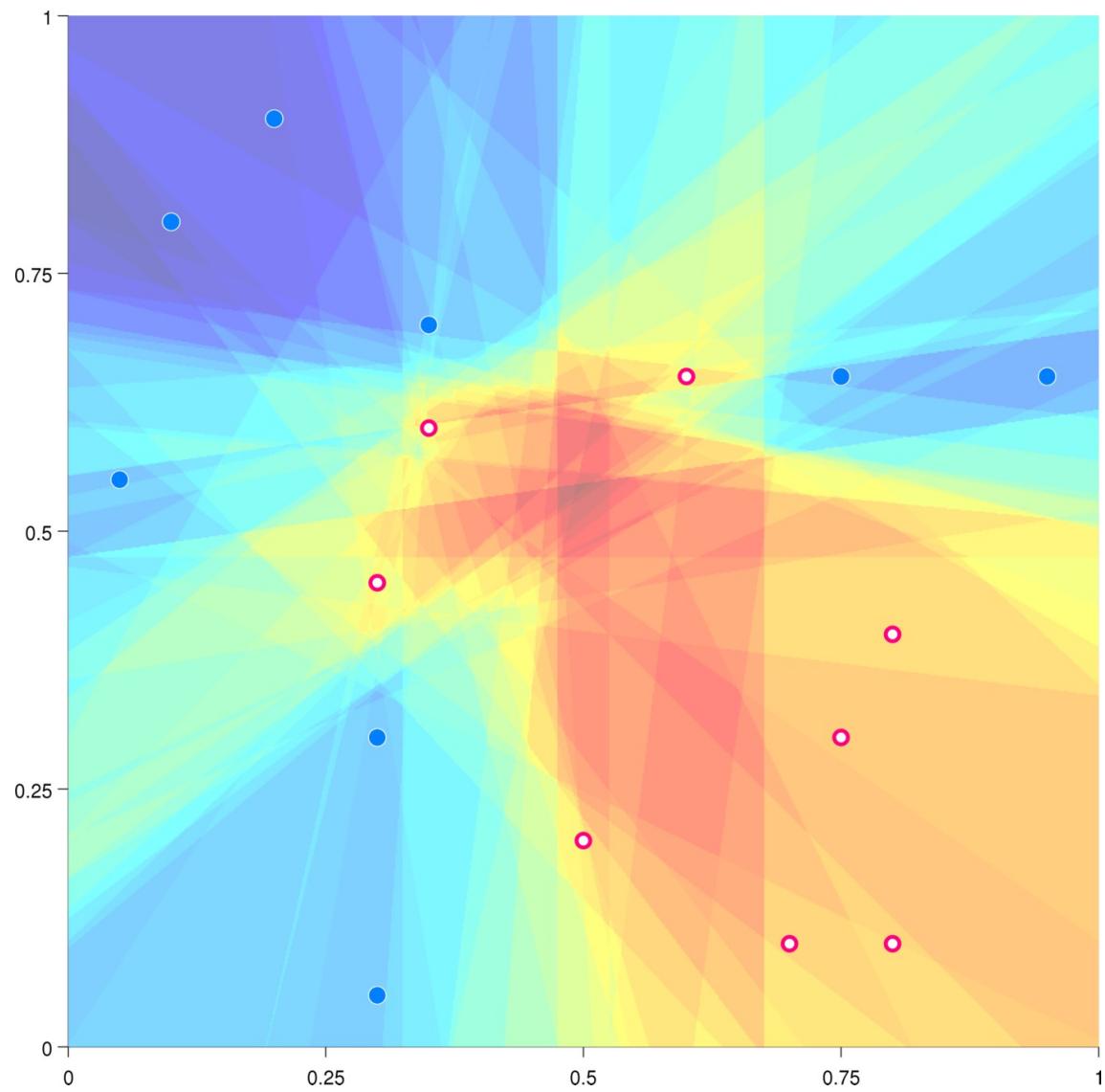
Boosting

- Majority voting of 10 boosted classifiers



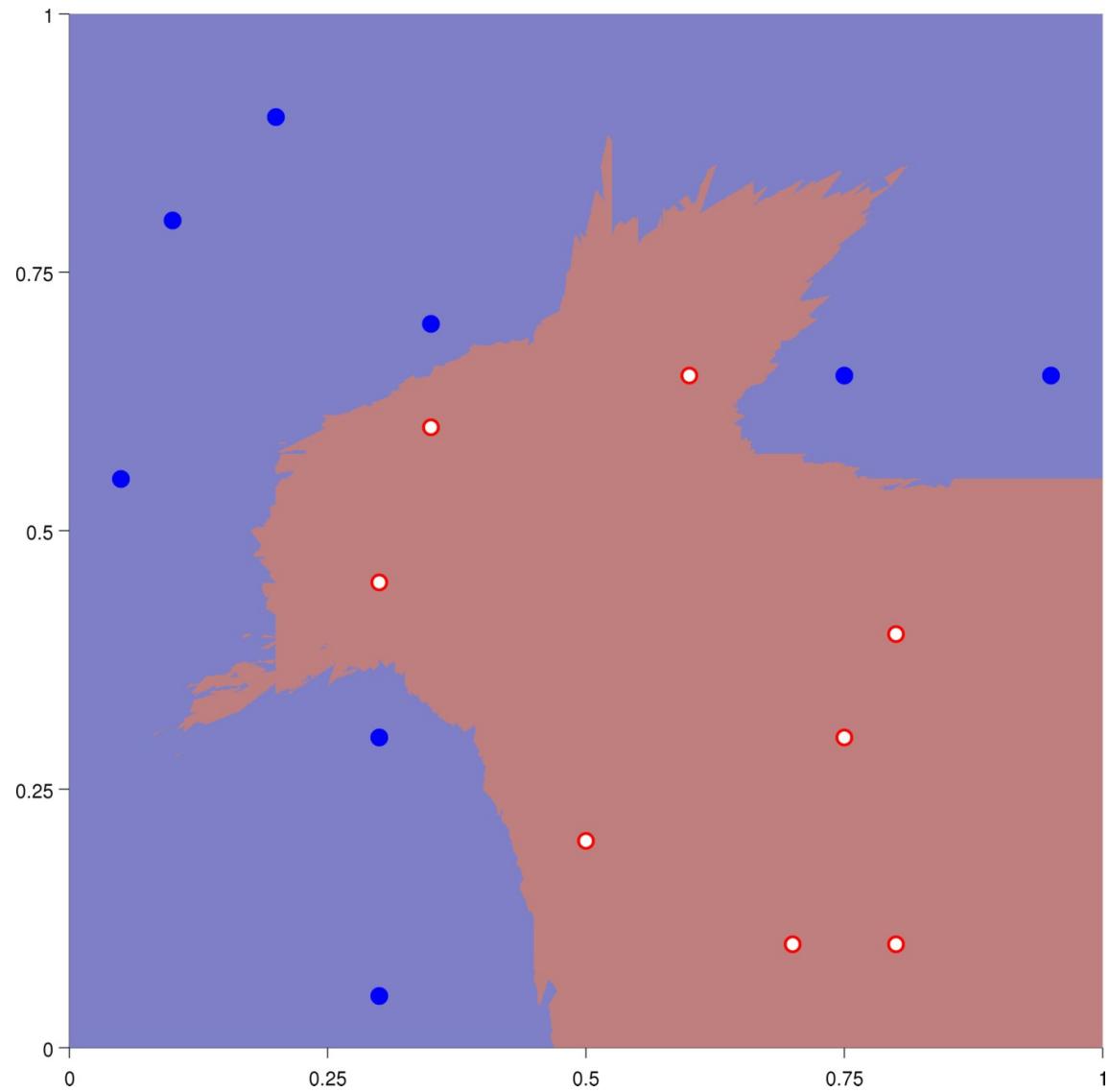
Boosting

- Majority voting of 100 boosted classifiers



Boosting

- Decision boundary of 100 boosted classifier ensemble



Class imbalance problem

- Many data sets have **imbalanced class distributions**
 - Example: Detection of defects that only occur rarely (e.g. 1/1,000,000)
 - Danger: Algorithm that says nothing is defect will be 99.999% correct
- **Solution approaches**
 - Resample to balance data sets
 - Modify existing classification algorithms
 - Measure performance in a way that takes balance into account

Resampling balanced data

- New sample has equal number of data objects from each class

- **Approaches**

- **Undersampling** majority class: Throws out potentially useful data
- **Oversampling** minority class: Increase data size and computational burden
- **Somewhere in between...**

Imbalanced training data

1	2	3	4	5	6	7	8	9	10
---	---	---	---	---	---	---	---	---	----

Oversampling

1	2	3	4	5	1	2	3	6	6
6	6	8	8	8	8				

Undersampling

3	5	6	8
---	---	---	---

Somewhere in between

3	5	4	3	9	6	6	8	8	8
---	---	---	---	---	---	---	---	---	---

Confusion matrix

		<i>Predicted</i>	
		<i>positive</i>	<i>negative</i>
<i>Actual</i>	<i>positive</i>	TP True Positive	FN False Negative
	<i>negative</i>	FP False Positive	TN True Negative

Precision and recall

- **Precision**

- Fraction of true positive among objects predicted to be positive

$$p = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

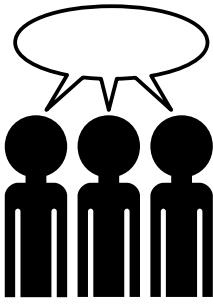
- **Recall**

- Fraction of objects predicted to be positive among all positive objects

$$r = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

		<i>Predicted</i>	
		<i>positive</i>	<i>negative</i>
<i>Actual</i>	<i>positive</i>	TP True Positive	FN False Negative
	<i>negative</i>	FP False Positive	TN True Negative





Group exercise

- You consider two different classifiers, on a test set with 20 positive objects
 - **Classifier 1** detects 54 positives of which 18 are actually positive
 - **Classifier 2** detects 16 positives of which 14 are actually positive
- Compute the **precision** and **recall** for the two classifiers
- Which classifier (if any) is the best?
- Which would you use if the objective is to detect credit card fraud
*(consider what is most costly – **missing** or **falsely detecting** a positive)*

- **Precision**

- Fraction of true positive among objects predicted to be positive

$$p = \frac{TP}{TP + FP}$$

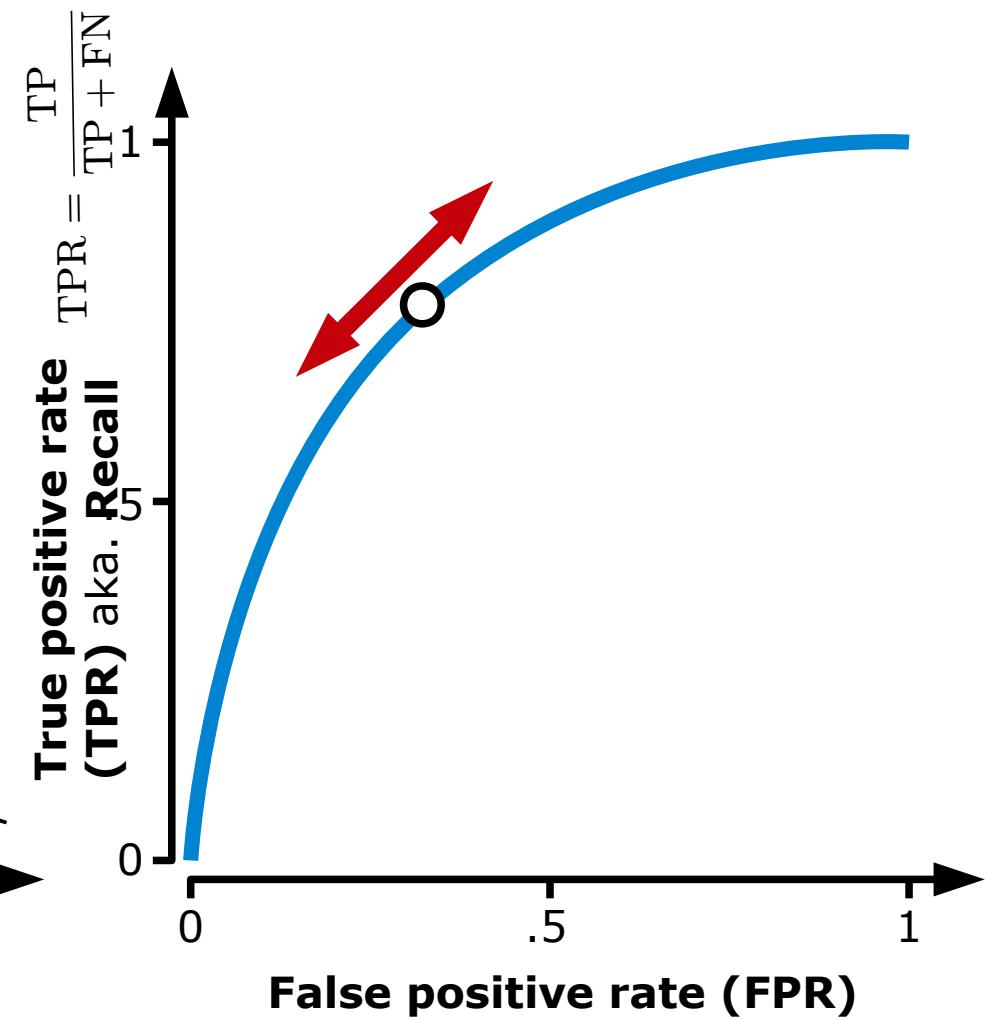
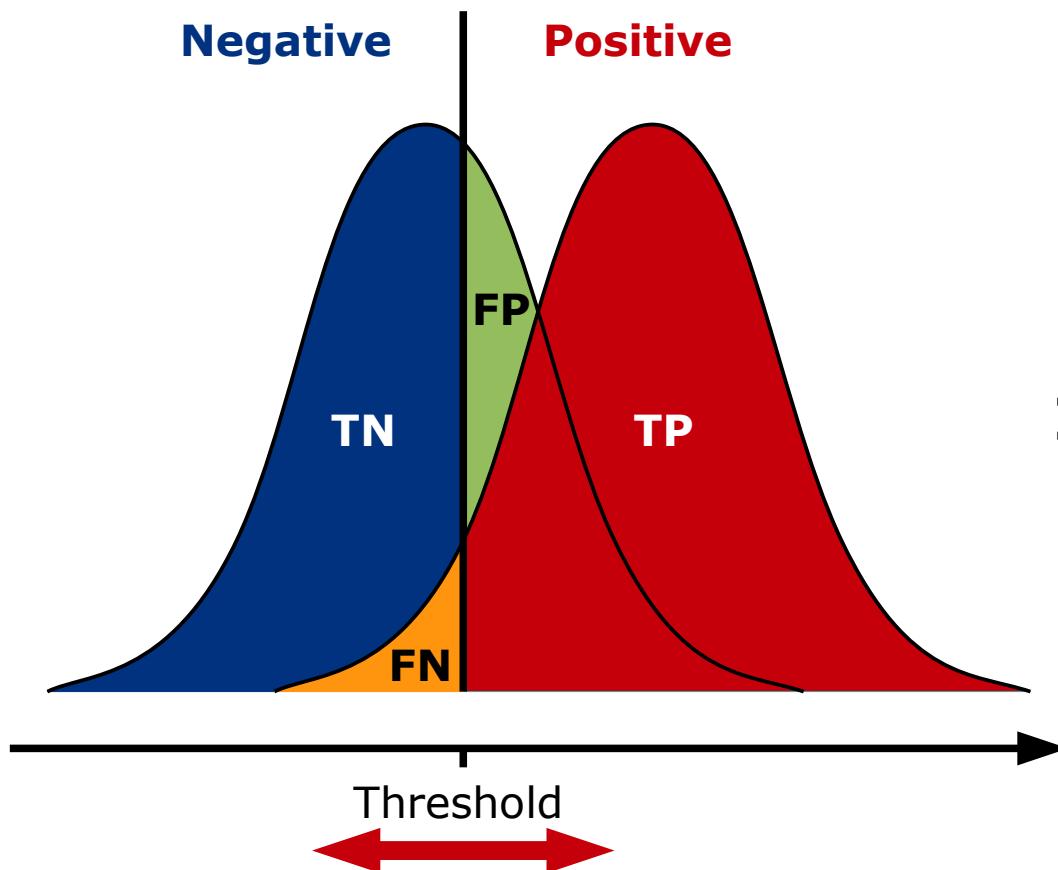
- **Recall**

- Fraction of objects predicted to be positive among all positive objects

$$r = \frac{TP}{TP + FN}$$

		<i>Predicted</i>	
		<i>positive</i>	<i>negative</i>
<i>Actual</i>	<i>positive</i>	TP	FN
	True Positive	False Negative	
	<i>negative</i>	FP	TN
	False Positive	True Negative	

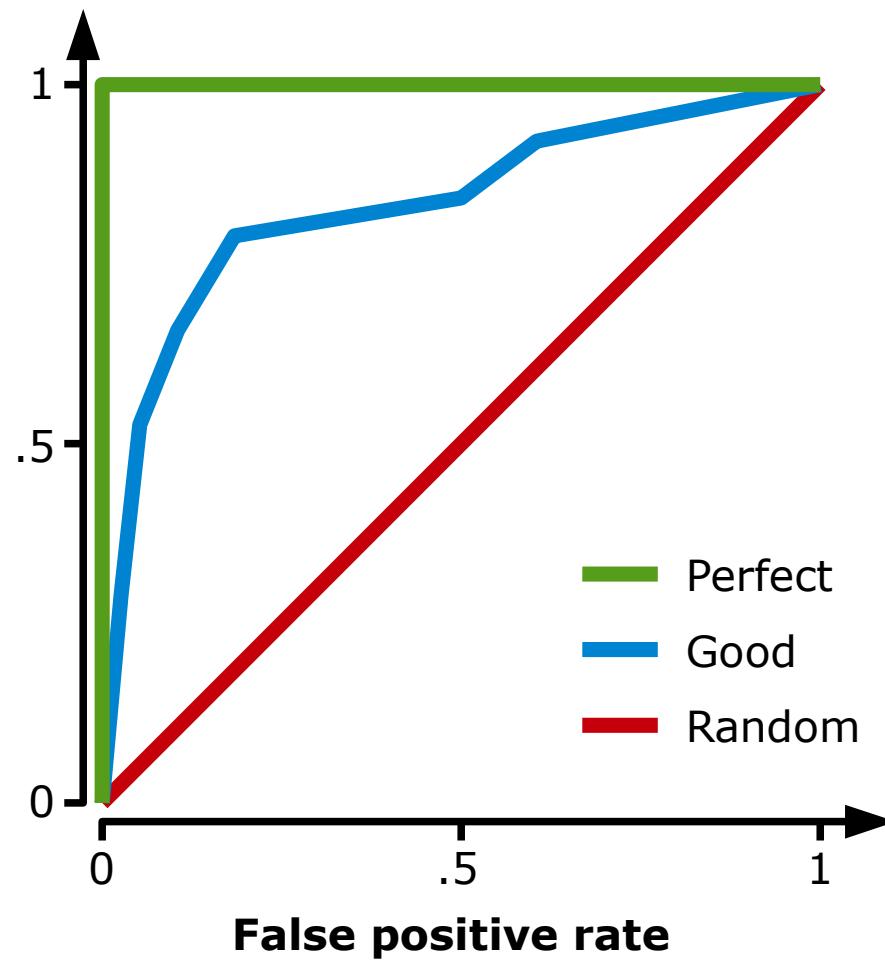
Receiver operating characteristic



Receiver operating characteristic

True positive rate
aka. **Recall**

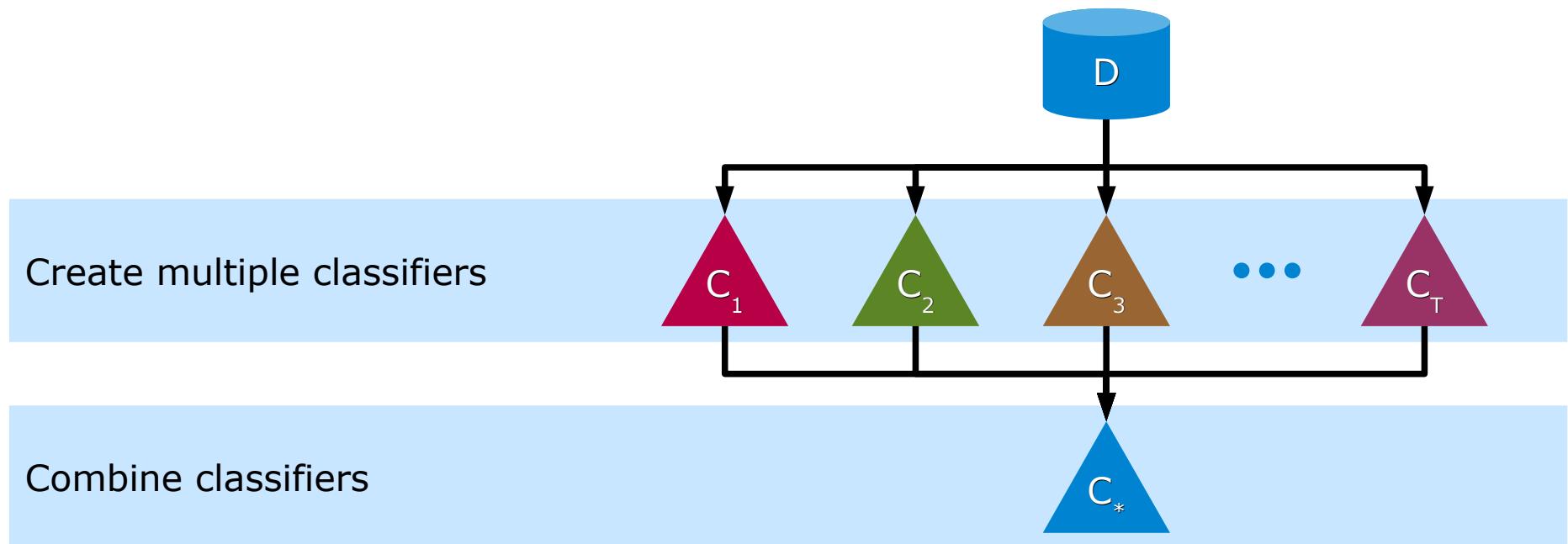
$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$



$$\text{FPR} = \frac{\text{FP}}{\text{TN} + \text{FP}}$$

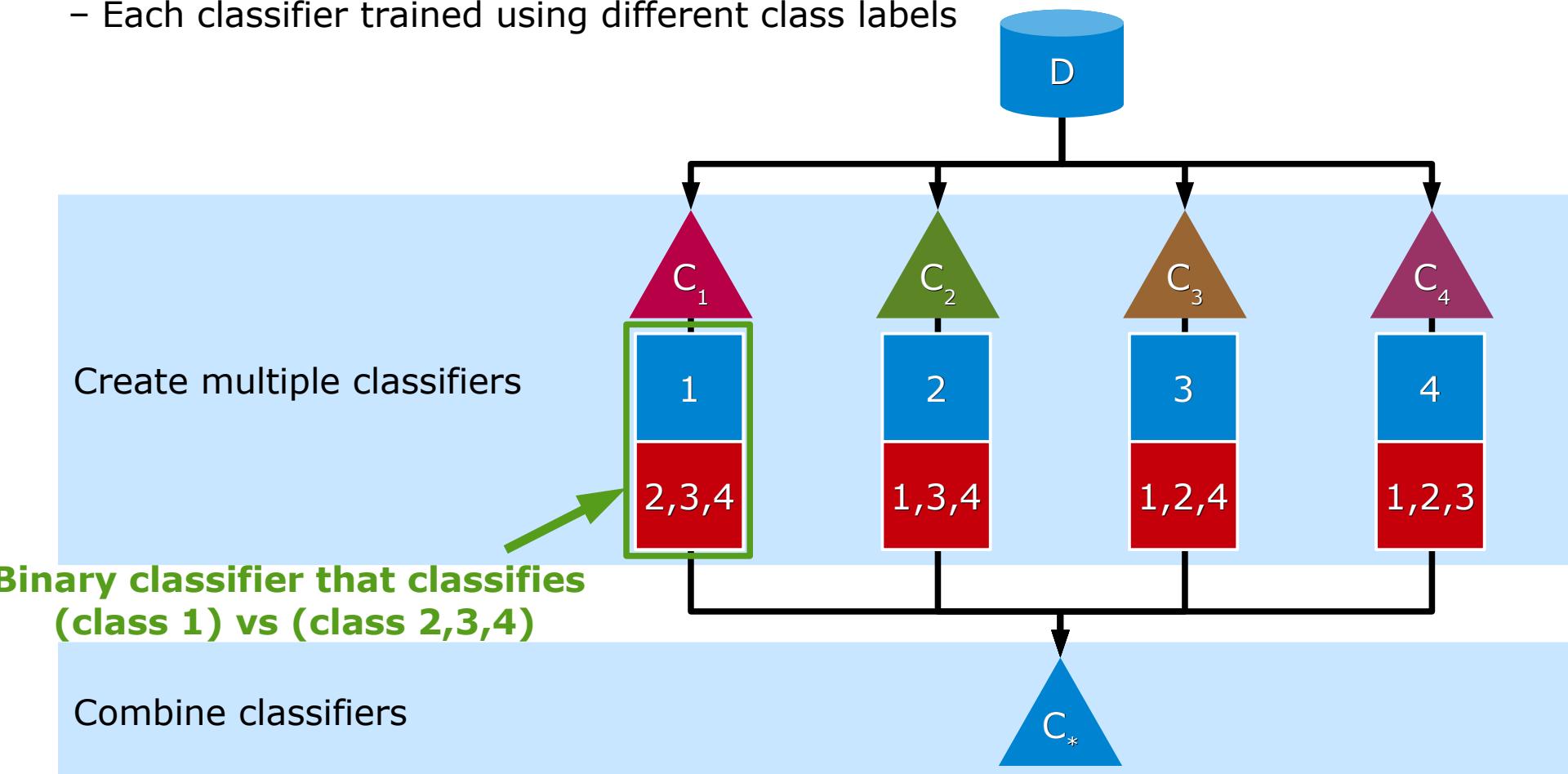
Multiclass problems

- Combine multiple **binary** classifiers into one **multiclass** classifier
 - Each classifier trained using different class labels



Multiclass problems

- Combine multiple **binary** classifiers into one **multiclass** classifier
 - Each classifier trained using different class labels



Multiclass problems

- Classification algorithms designed for **binary classification** can also be applied to **multiclass** problems
 - Train a number of classifiers and make final classification by **majority voting**

- 1-against-rest**

- K classifiers

1	2	3	4
2,3,4	1,3,4	1,2,4	1,2,3

- 1-against-1**

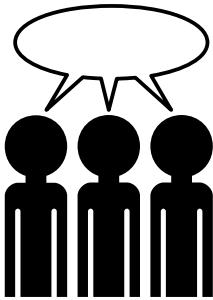
- $K \cdot (K-1)/2$ classifiers

1	1	1	2	2	3
2	3	4	3	4	4

- Error correcting output coding**

- Robustness against errors

1	1,4	1,3	1,3,4	1,2	1,2,4	1,2,3
2,3,4	2,3	2,4	2	3,4	3	4



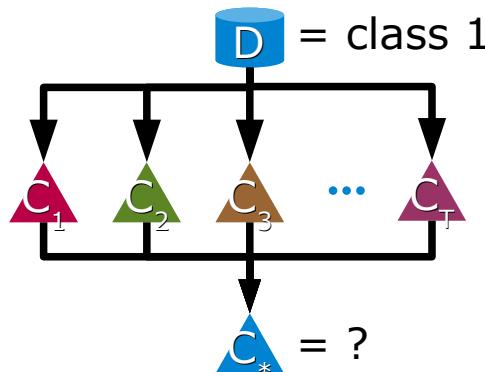
Possibly broken classifier

Group exercise

- Assuming that
all classifiers are perfect
 - Use the three multiclass methods to classify a data object belonging to class 1

Hint: Go through the classifiers and add up the votes, giving one vote to each of the winning classes. (In 1-against-1 give 0.5 to each class when class 1 is not included.)

- Assuming that
the first classifier is broken
and makes a wrong classification
 - Repeat the exercise



Binary classifier that classifies (class 4) vs (class 1,2,3)



	1-against-rest	1-against-1	Error correcting
1	1 2,3,4	1 2	1 2,3,4
2	2 1,3,4	3 1	4 1,2
3	3 1,2,4	4 1	2 1,2,4
4	4 1,2,3	1 2,3 4 3 2 1,2,3	3 4 1 2 3 4

Remember one-out-of-K coding

Nationality

		Denmark	Norway	Sweden
TXT=	X_tmp=	0	0	1
'Sweden'		0	0	1
'Sweden'		0	0	1
'Sweden'		0	0	1
'Sweden'		0	0	1
'Norway'		0	1	0
'Norway'		0	1	0
'Norway'		0	1	0
'Norway'		0	1	0
'Sweden'		0	0	1
'Norway'		0	1	0
'Denmark'		1	0	0
'Denmark'		1	0	0
'Sweden'		0	0	1
'Sweden'		0	0	1
'Sweden'		0	0	1
'Denmark'		1	0	0
'Sweden'		0	0	1
'Norway'		0	1	0
'Denmark'		1	0	0

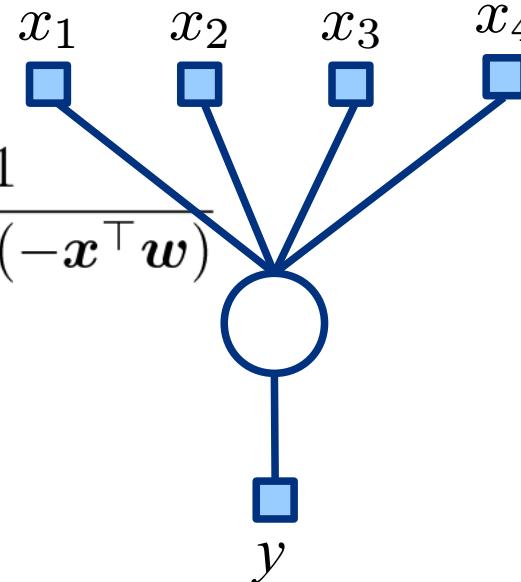
One-out-of-K coding

Extending classifiers to handle multi-class problems

- Logistic regression

$$\{\mathbf{x}_n, y_n\}_{n=1}^N$$

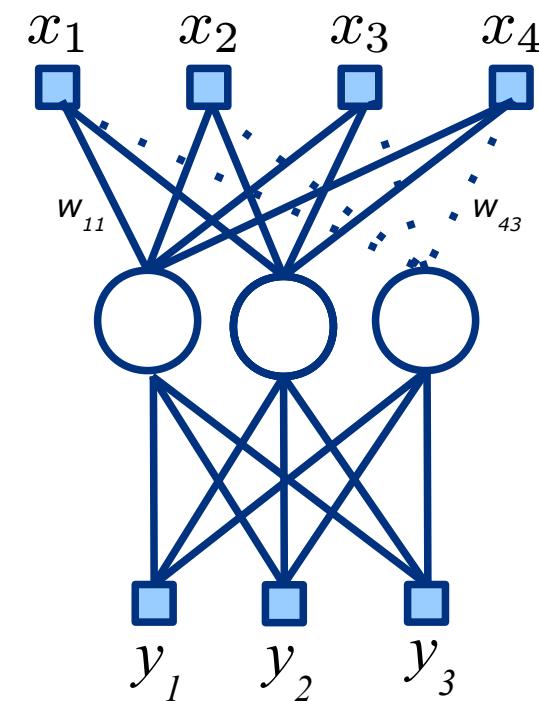
$$f(\mathbf{x}) = \text{logit}(\mathbf{x}^\top \mathbf{w}) = \frac{1}{1 + \exp(-\mathbf{x}^\top \mathbf{w})}$$

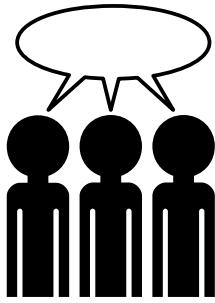


- Multinomial Regression

$$\{\mathbf{x}_n, y_n\}_{n=1}^N$$

$$f_c(\mathbf{x}) = \text{softmax}(\mathbf{x}^\top \mathbf{W}) = \frac{\exp(\mathbf{x}^\top \mathbf{w}_c)}{\sum_{c'} \exp(\mathbf{x}^\top \mathbf{w}_{c'})}$$
$$\mathbf{w}_C = \mathbf{0}$$





Group exercise

- Show that for a binary (i.e, two class problem) multinomial regression is equivalent to logistic regression if $\mathbf{w}_0 = \mathbf{0}$.

Hint: consider the softmax function for two classes:

$$f_c(\mathbf{x}) = softmax_c(\mathbf{x}^\top \mathbf{W}) = \frac{\exp(\mathbf{x}^\top \mathbf{w}_c)}{\sum_{c'} \exp(\mathbf{x}^\top \mathbf{w}_{c'})}$$

and show that this is equivalent to the logit function:

$$f(\mathbf{x}) = logit(\mathbf{x}^\top \mathbf{w}) = \frac{1}{1 + \exp(-\mathbf{x}^\top \mathbf{w})}$$

Artificial Neural Networks for multiclass classification

input layer

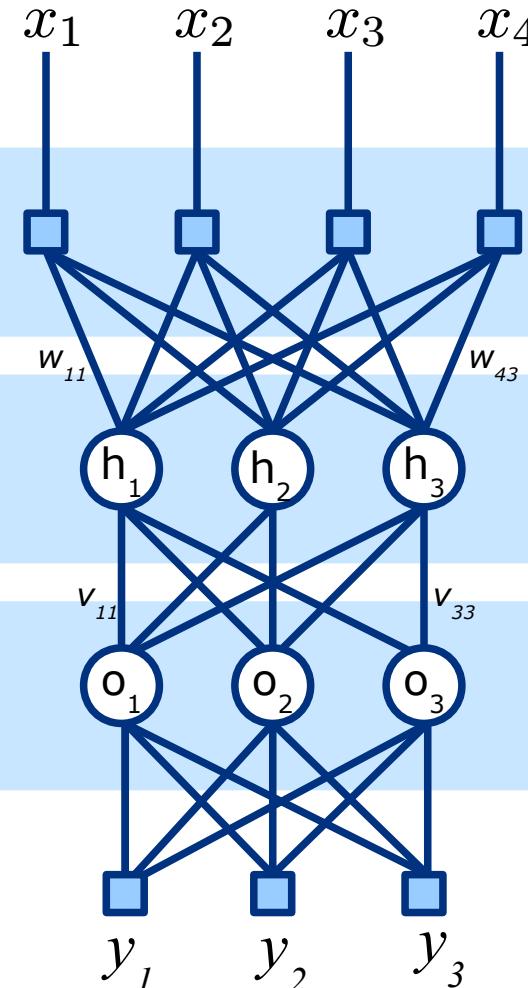
hidden layer

$$h_j = g^H(\mathbf{x}^T \mathbf{w}_j) = g^H(w_{0j} + w_{1j}x_1 + w_{2j}x_2 + w_{3j}x_3 + w_{4j}x_4)$$

output layer

$$o_c = g^O(\mathbf{h}^T \mathbf{v}) = g^O(v_{0c} + v_{c1}h_1 + v_{c2}h_2 + v_{c3}h_3)$$

$$f_c(\mathbf{o}) = softmax_c(\mathbf{o}) = \frac{\exp(o_c)}{\sum_{c'} \exp(o_{c'})}$$



02450 Introduction to machine learning and data modeling

$$f(x+\Delta x) = \sum_{i=0}^{\infty} \frac{(\Delta x)^i}{i!} f^{(i)}(x)$$

A complex mathematical expression featuring various symbols and numbers:

- Θ (purple)
- Ω (pink)
- δ (yellow)
- $e^{i\pi}$ (purple)
- $\sqrt{17}$ (pink)
- \int_a^b (yellow)
- Σ (red)
- χ^2 (orange)
- ∞ (pink)
- \gg (yellow)
- $=$ (red)
- $\{2.7182818284$ (pink)
- $,$ (pink)

Reading material

Tan, Steinbach and Kumar
"Introduction to Data Mining"

Section 8.1-8.3+8.5.7

Groups of the day

Jacob Elbæk

Alexandra Fretoft

Anders Michael Nielsen

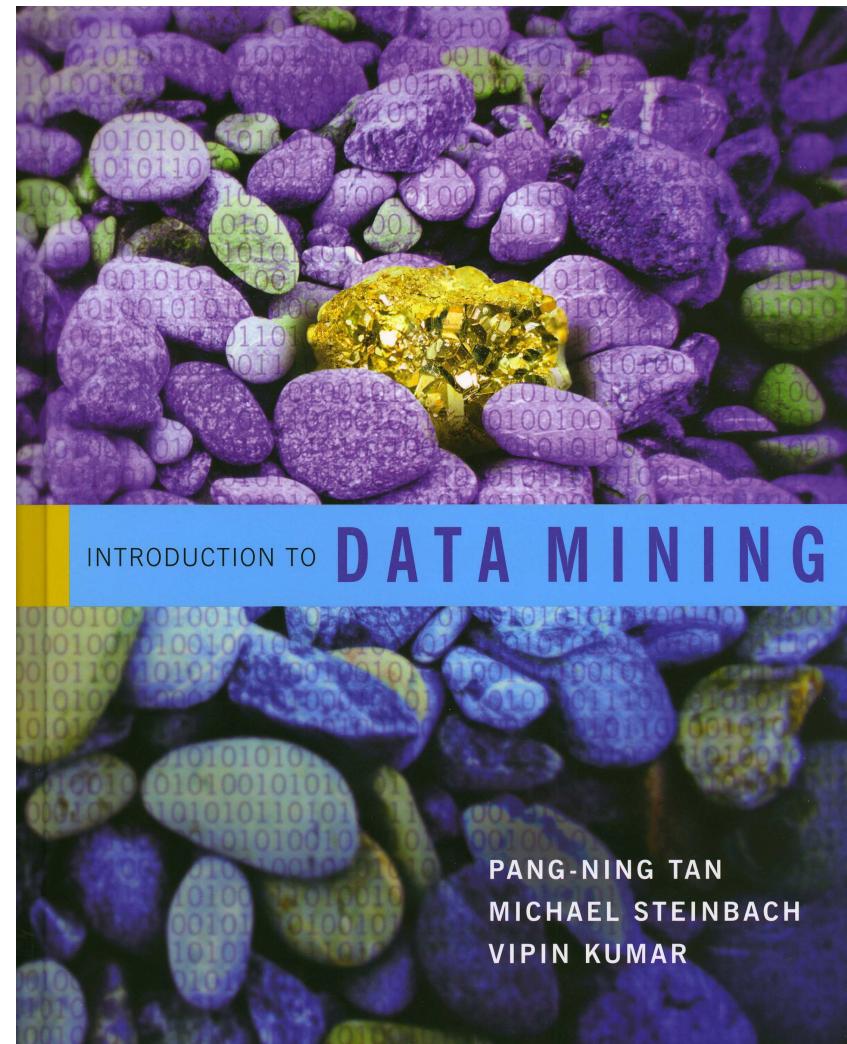
Simon Bøge Hemmingsen

Marc Marer Thomsen

Alex Pellegrini

Monica Emerson

Woody Rousseau



Lecture schedule

1. Introduction
(Tan 1.1-1.4)

Data: Feature extraction and visualization

2. Data and feature extraction
(Tan 2.1-2.3 + B1 (+ A))
3. Measures of similarity and summary statistics
(Tan 2.4 + 3.1-3.2 + C1-C2)

4. Data visualization
(Tan 3.3)

Supervised learning: Classification and regression

5. Decision trees and linear regression
(Tan 4.1-4.3 + D)
6. Overfitting and performance evaluation
(Tan 4.4-4.6)
7. Nearest neighbor, naive Bayes, and artificial neural networks
(Tan 5.2-5.4)

8. Ensemble methods and multi class classifiers
(Tan 5.6-5.8)

Unsupervised learning: Clustering and density est.

9. K-means and hierarchical clustering
(Tan 8.1-8.3+8.5.7)

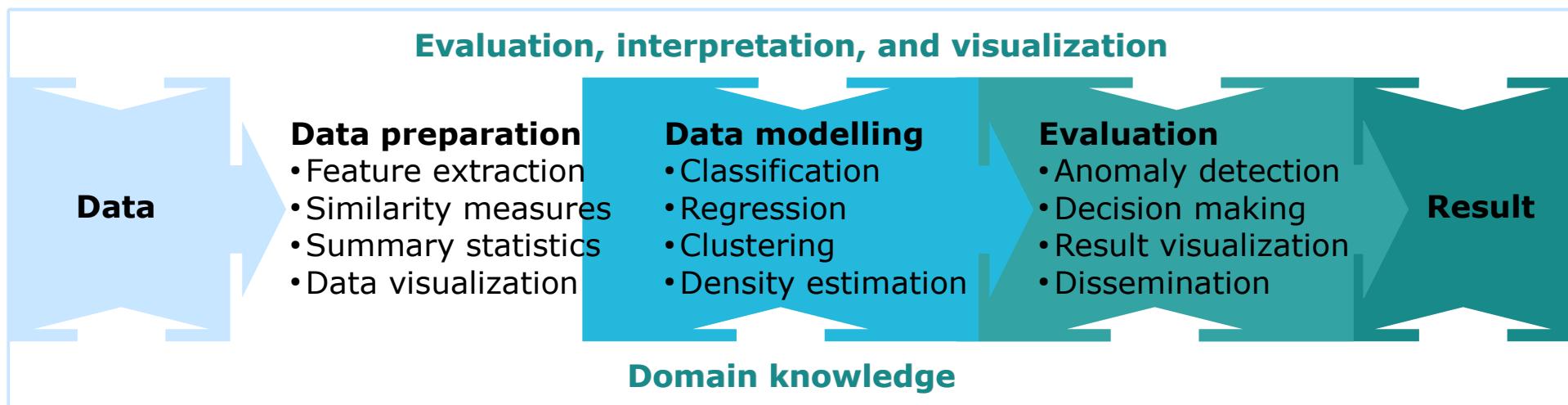
10. Mixture models and association mining
(Tan 9.2.2 + 6.1-6.3)

11. Density estimation and anomaly detection
(Tan 10.1-10.4)

Machine learning and data modelling in practice

12. Putting it all together: Summary and overview
13. Mini project

Data modeling framework



After today you should be able to:

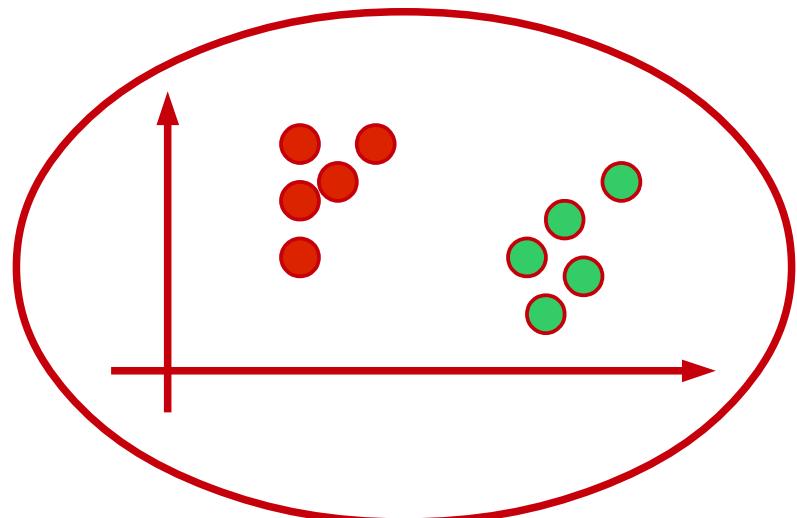
Discuss the aims of unsupervised learning

Explain the principles behind hierarchical clustering and k-means clustering

Apply hierarchical clustering and k-means clustering

Evaluate the quality of the clustering using supervised measures of cluster validity

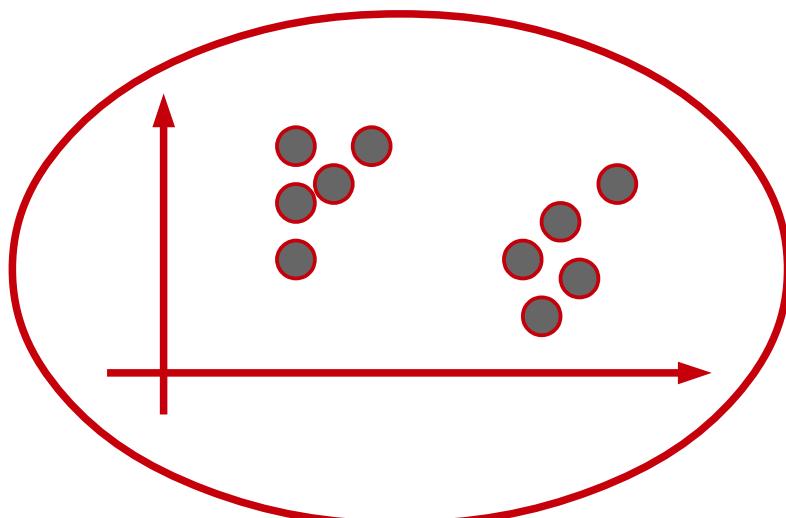
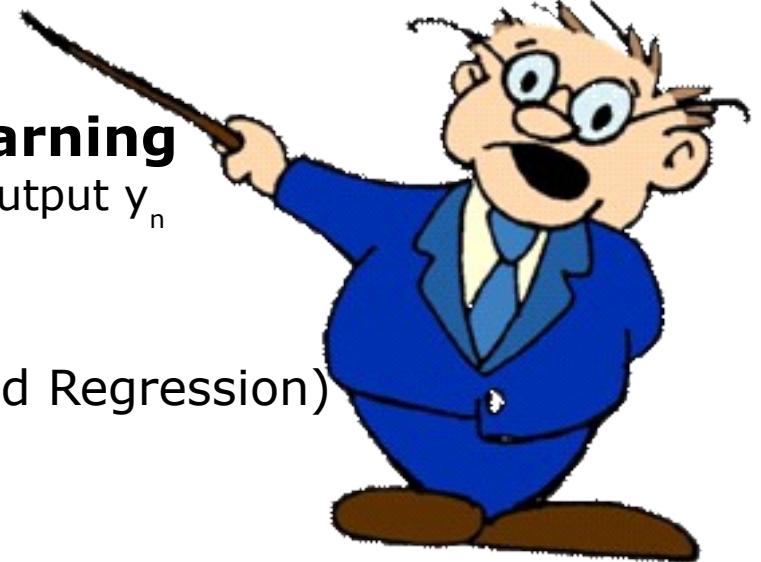
Supervised and Unsupervised learning



Supervised Learning

Input data \mathbf{x}_n and output y_n

(Classification and Regression)



Unsupervised Learning

Input data \mathbf{x}_n alone

(Exploratory analysis)



Imagine you observe the world for the first time!



<http://www.clipartlord.com/category/baby-clip-art/>



We humans are skilled at dividing objects into groups (clustering), but how do we make computers do the same?



http://commons.wikimedia.org/wiki/File:Abessinier_sorrel.jpg
http://commons.wikimedia.org/wiki/File:Cat_Eyes.jpg
http://commons.wikimedia.org/wiki/File:Black_white_cat_on_fence.jpg
<http://commons.wikimedia.org/wiki/File:Cat002.jpg>
<http://commons.wikimedia.org/wiki/File:Saurier2.jpg>
http://commons.wikimedia.org/wiki/File:Golden_Retriever_Dukedestiny01.jpg
<http://commons.wikimedia.org/wiki/File:Maspiri-Astro-SVE.jpg>
http://commons.wikimedia.org/wiki/File:GroC3%9Fer_Schweizer_Sennenhund.jpg
http://commons.wikimedia.org/wiki/File:GermanShorthPtr_wb.jpg

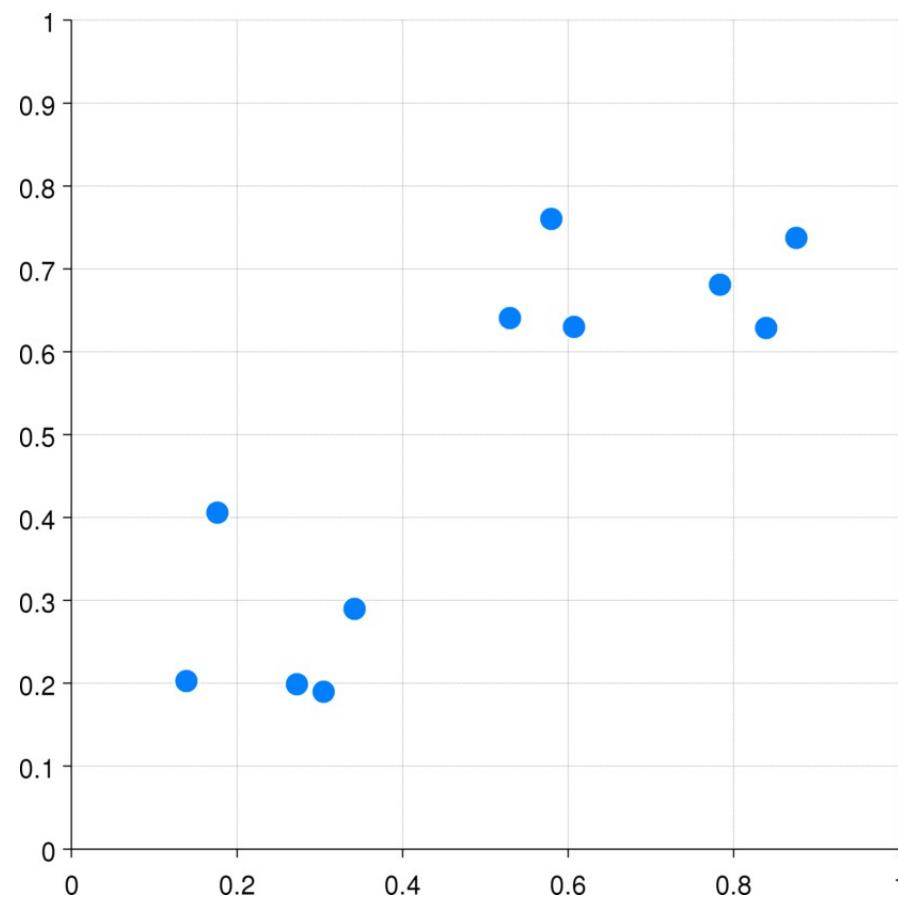
Unsupervised learning

- **Supervised learning**
 - Use the data to learn the output values
- **Unsupervised learning**
 - No output variables available
 - Use the data to learn from the data
 - Sometimes called exploratory analysis
 - What to find in the data?
 - Structure
 - Regularities
 - Hidden information
 - Etc.

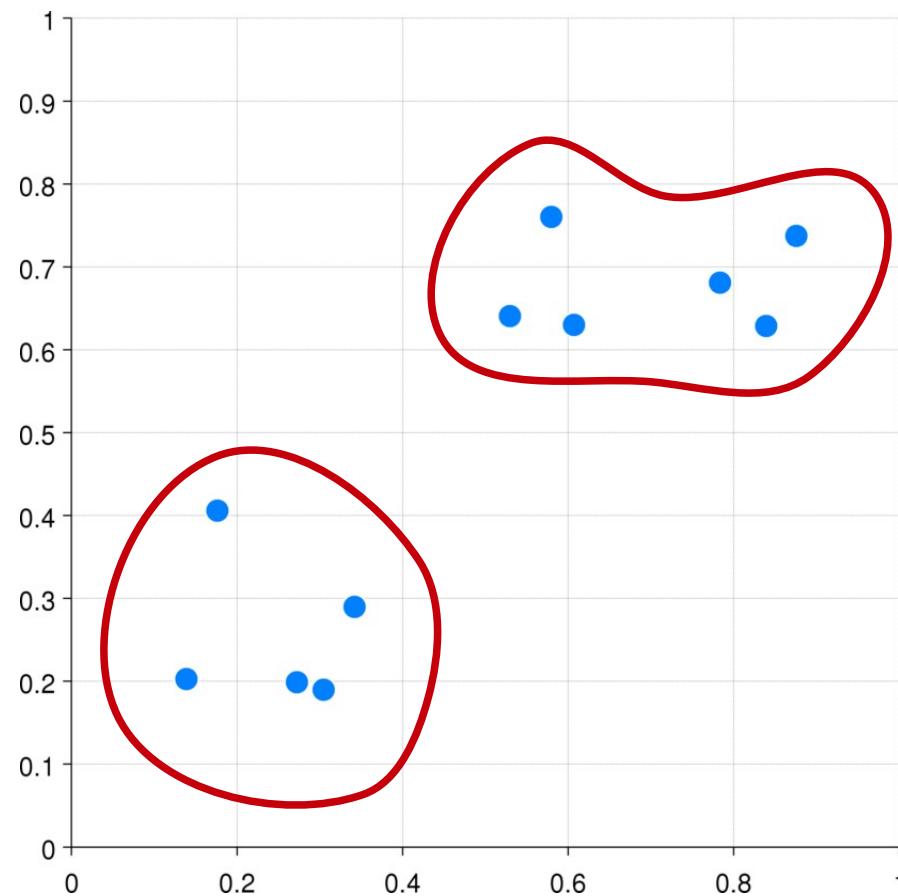
Clustering

- Divide data into groups (subsets/clusters) that are
 - **Meaningful:** Capture the natural structure of the data
 - **Useful:** Depends on purpose
- Observations in the same cluster are **similar in some sense**
- Unsupervised classification

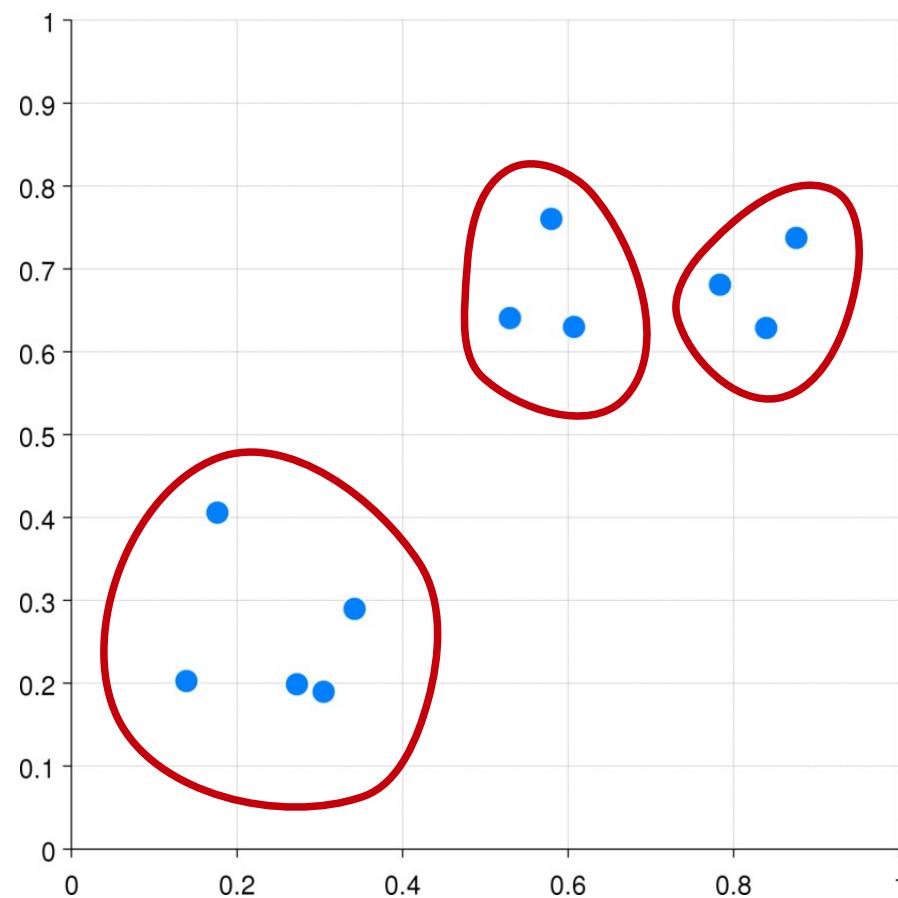
Clustering



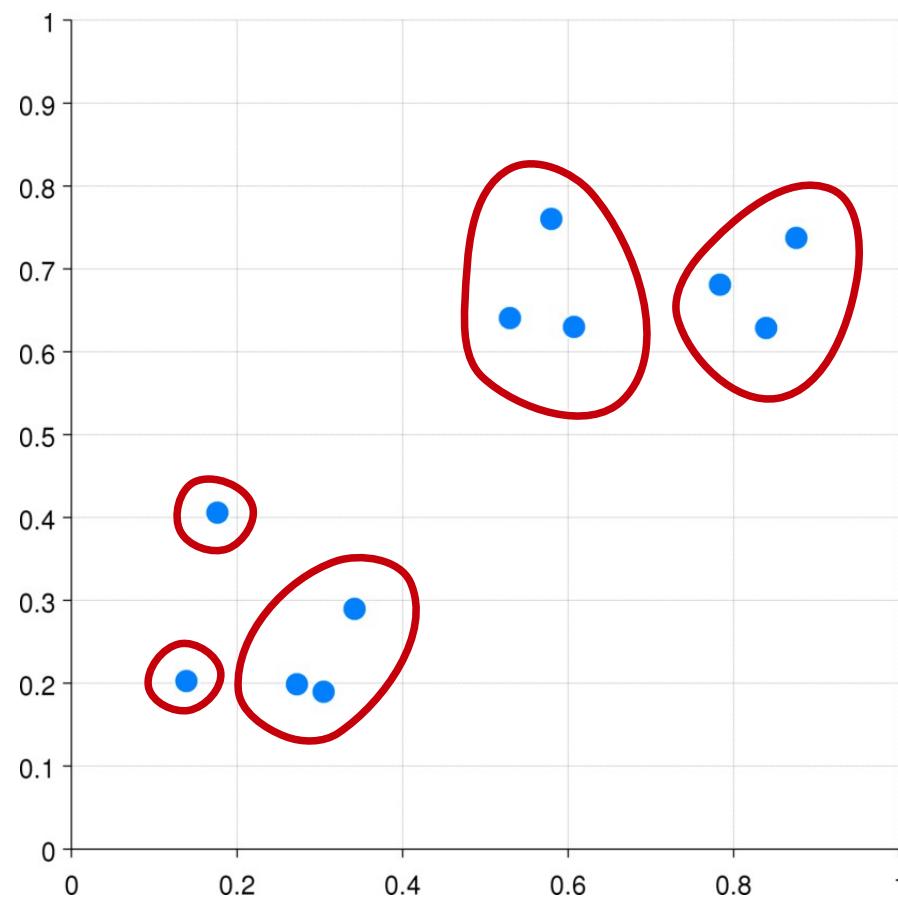
Clustering



Clustering

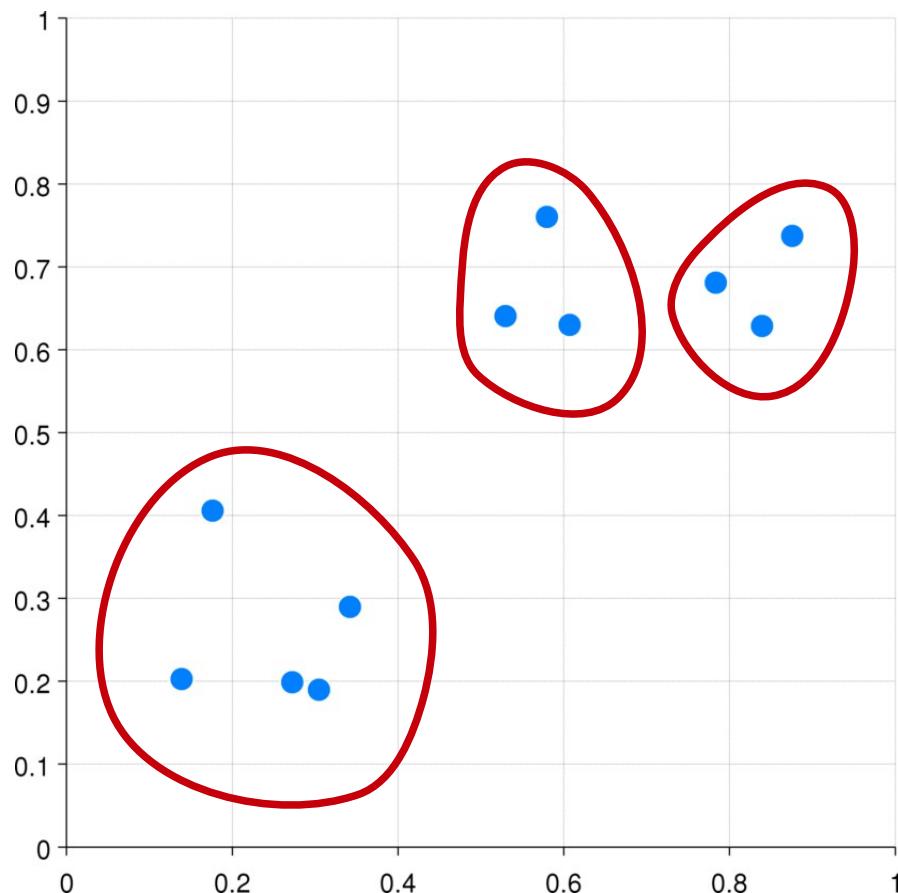


Clustering

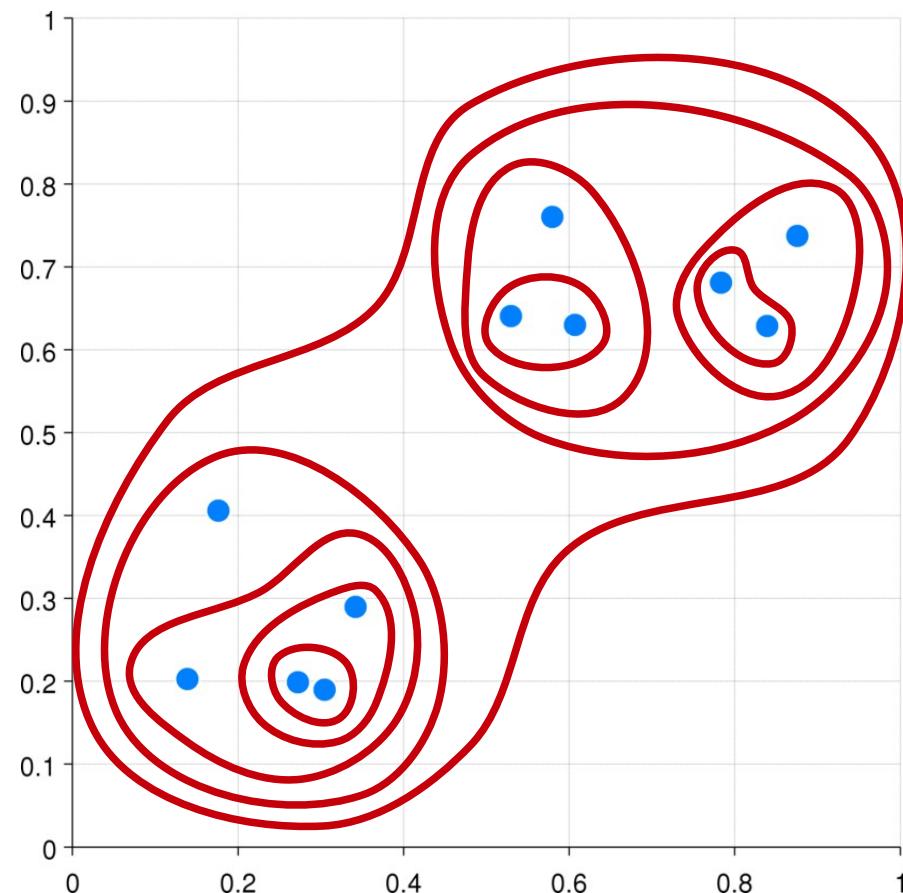


Partitional / hierarchical clustering

Partitional



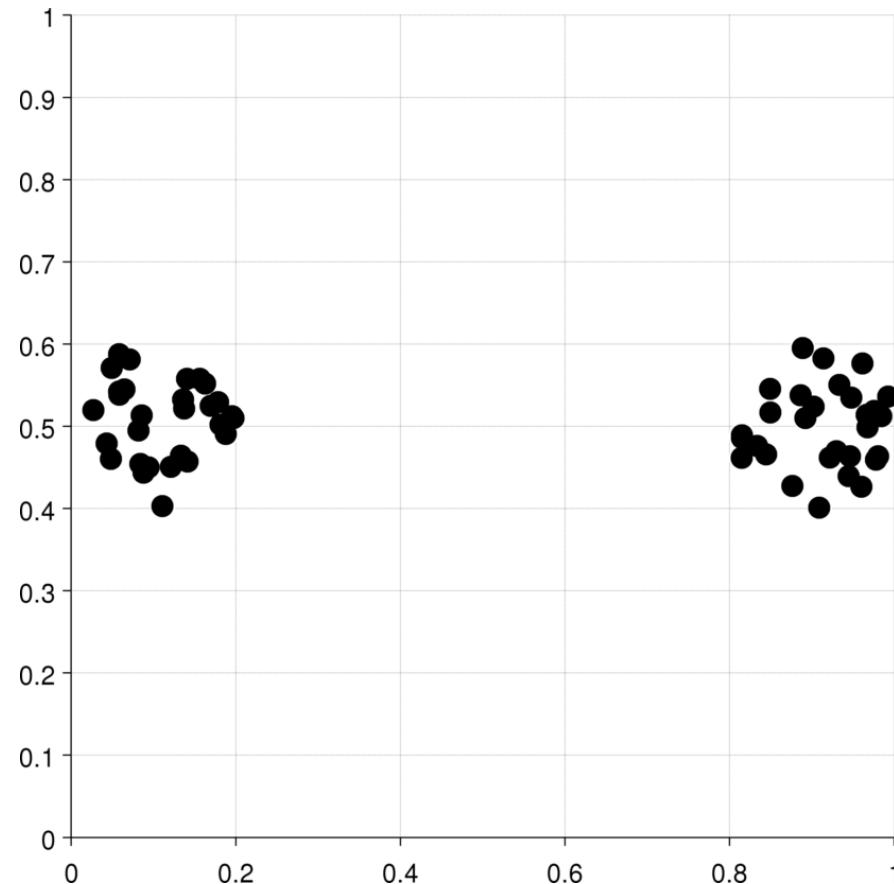
Hierarchical



Types of clustering

Well-separated

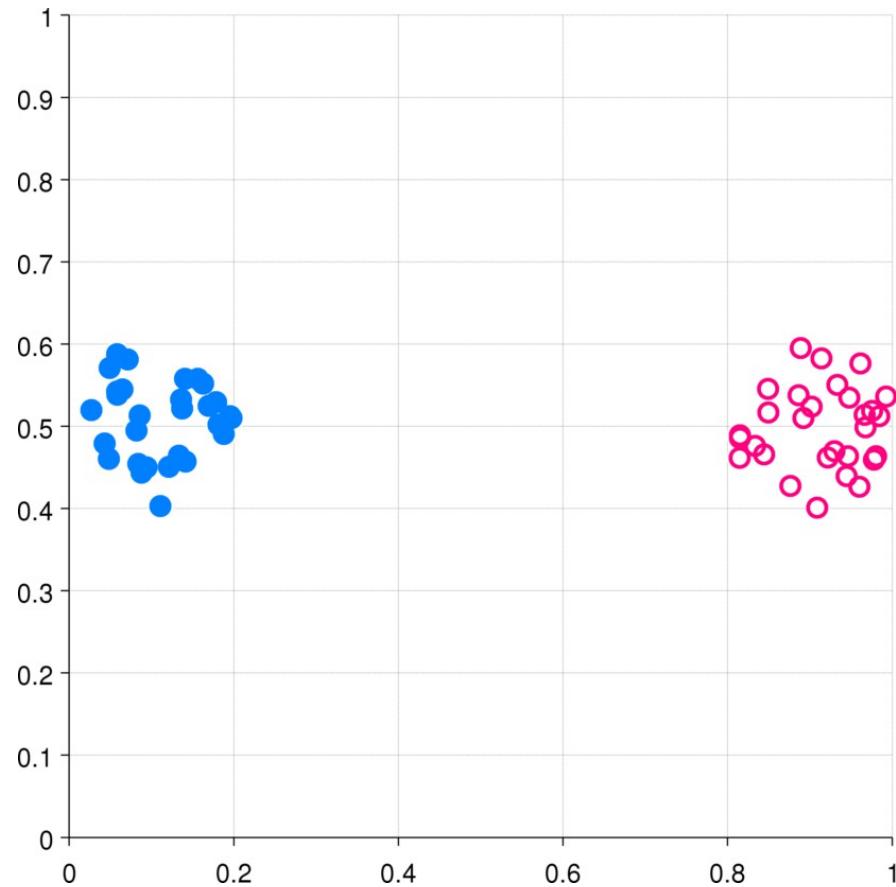
- Each point is closer to all points in its cluster than any point in another cluster



Types of clustering

Well-separated

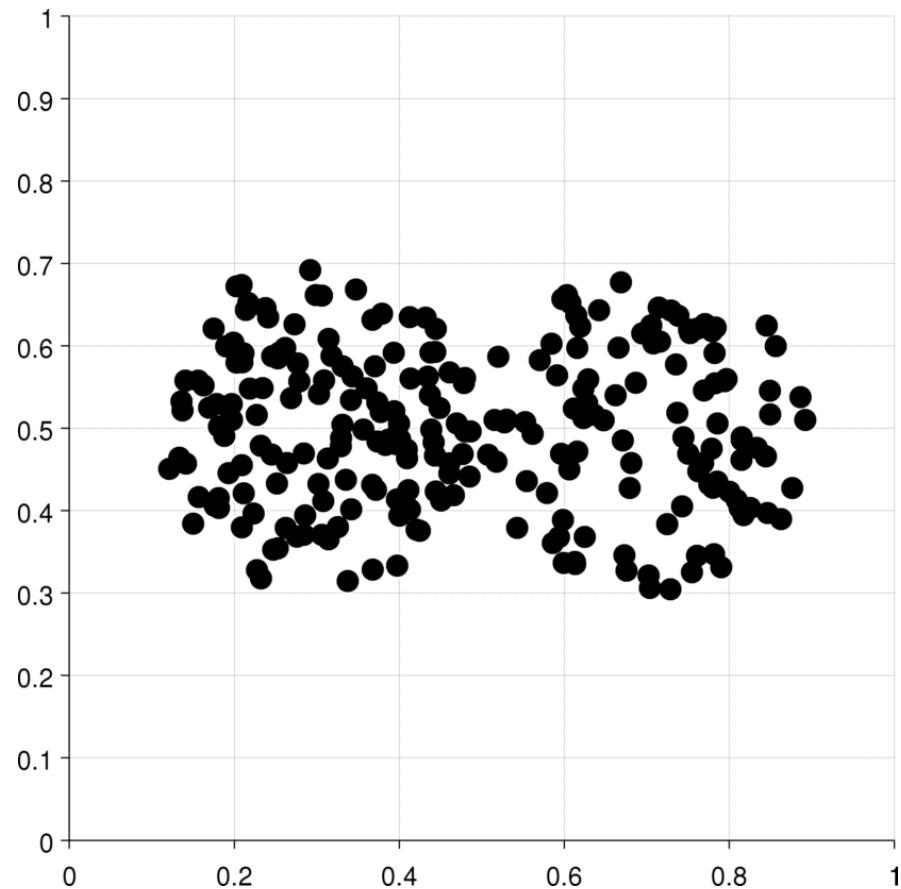
- Each point is closer to all points in its cluster than any point in another cluster



Types of clustering

Center-based

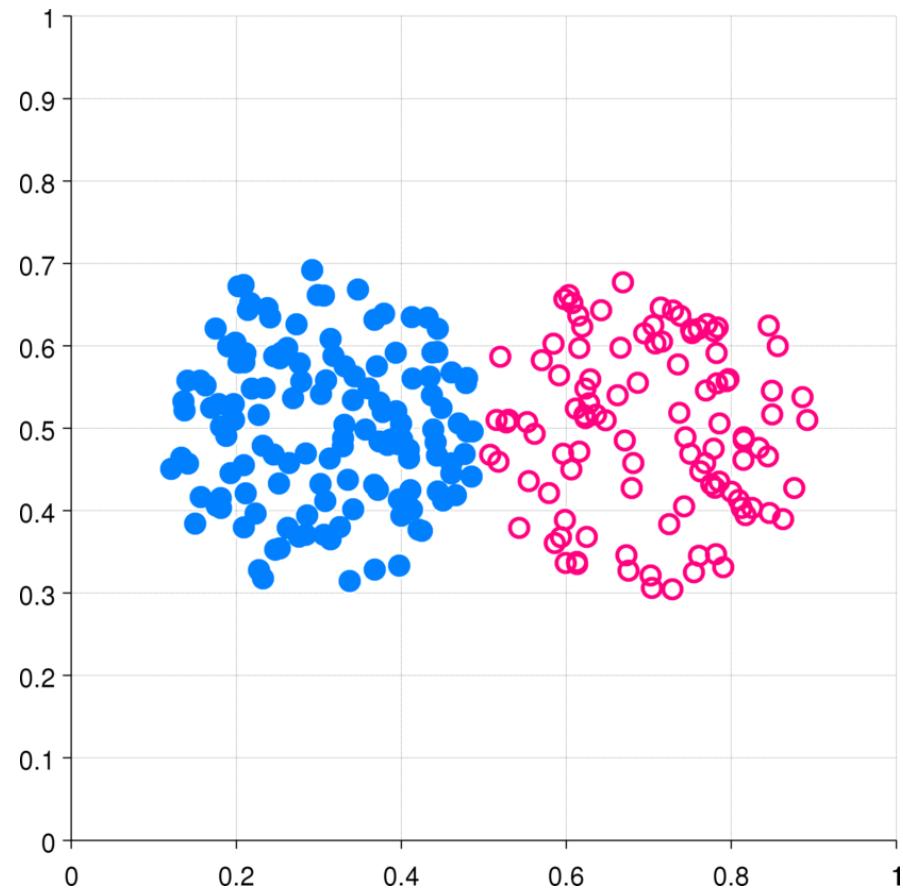
- Each point is closer to the center of its cluster than to the center of any other cluster



Types of clustering

Center-based

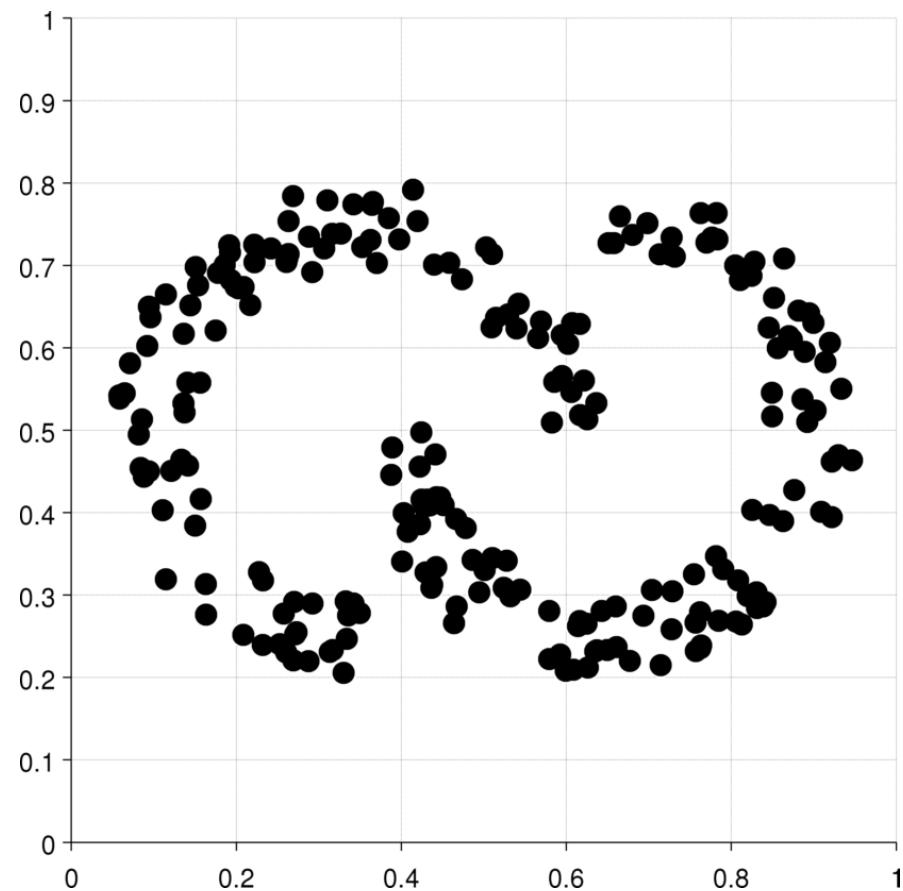
- Each point is closer to the center of its cluster than to the center of any other cluster



Types of clustering

Contiguity-based

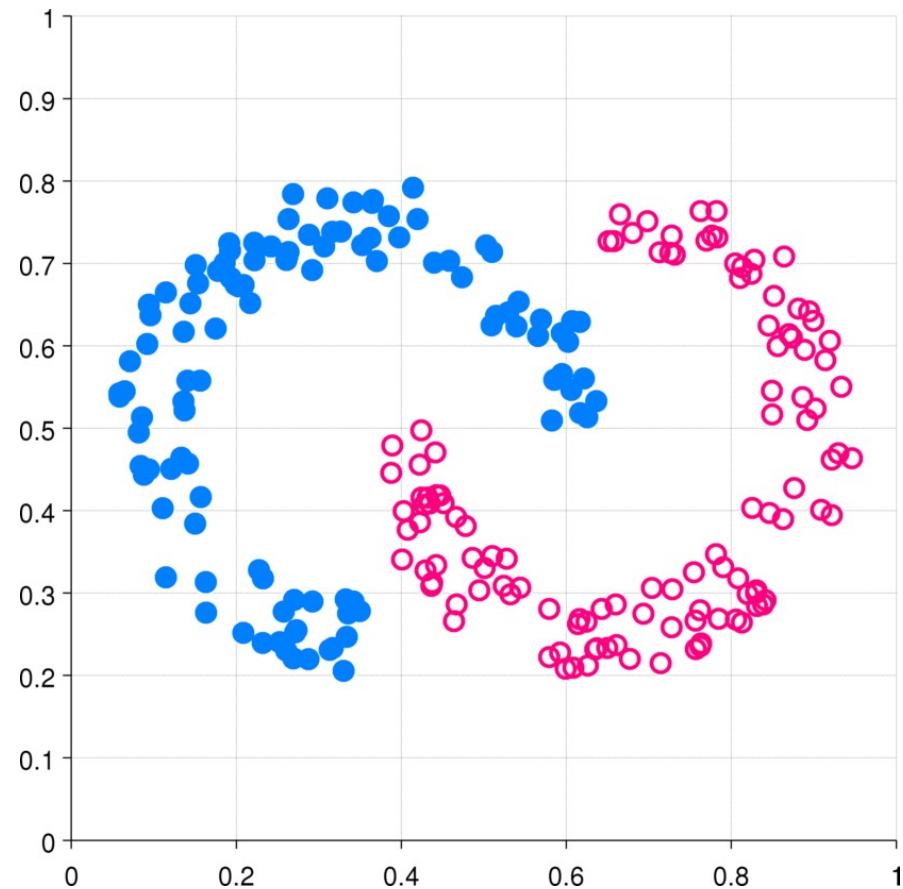
- Each point is closer to at least one point in its cluster than to any point in another cluster



Types of clustering

Contiguity-based

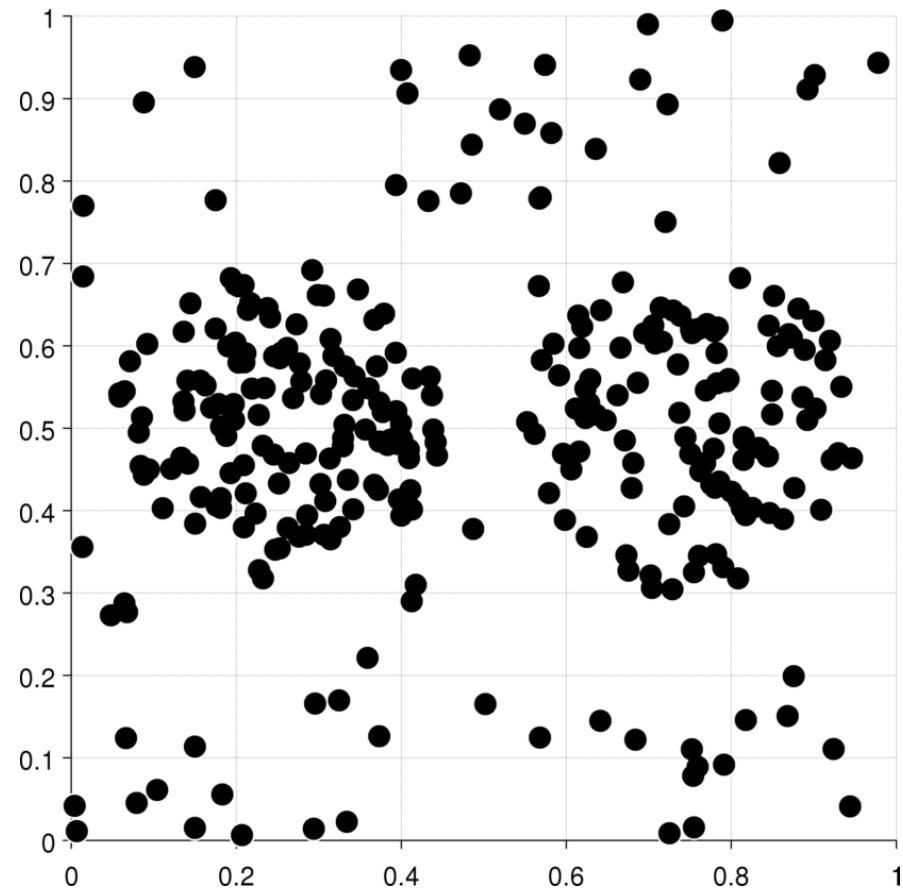
- Each point is closer to at least one point in its cluster than to any point in another cluster



Types of clustering

Density-based

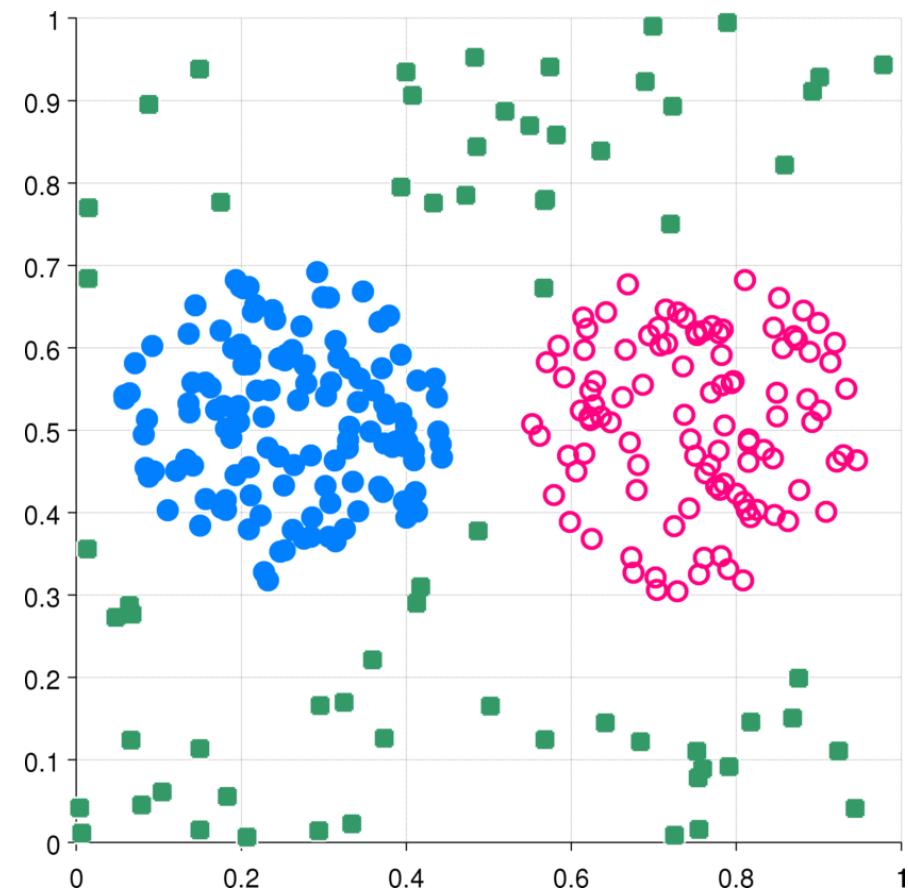
- Clusters are regions of high density separated by regions of low density



Types of clustering

Density-based

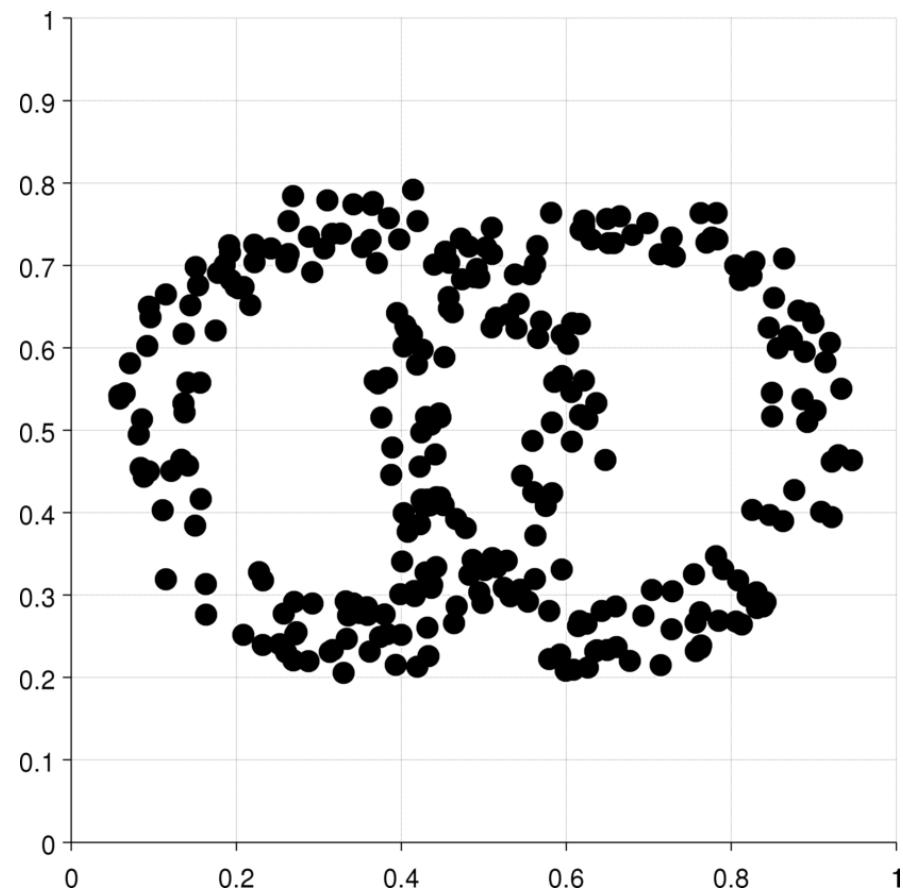
- Clusters are regions of high density separated by regions of low density



Types of clustering

Conceptual clusters

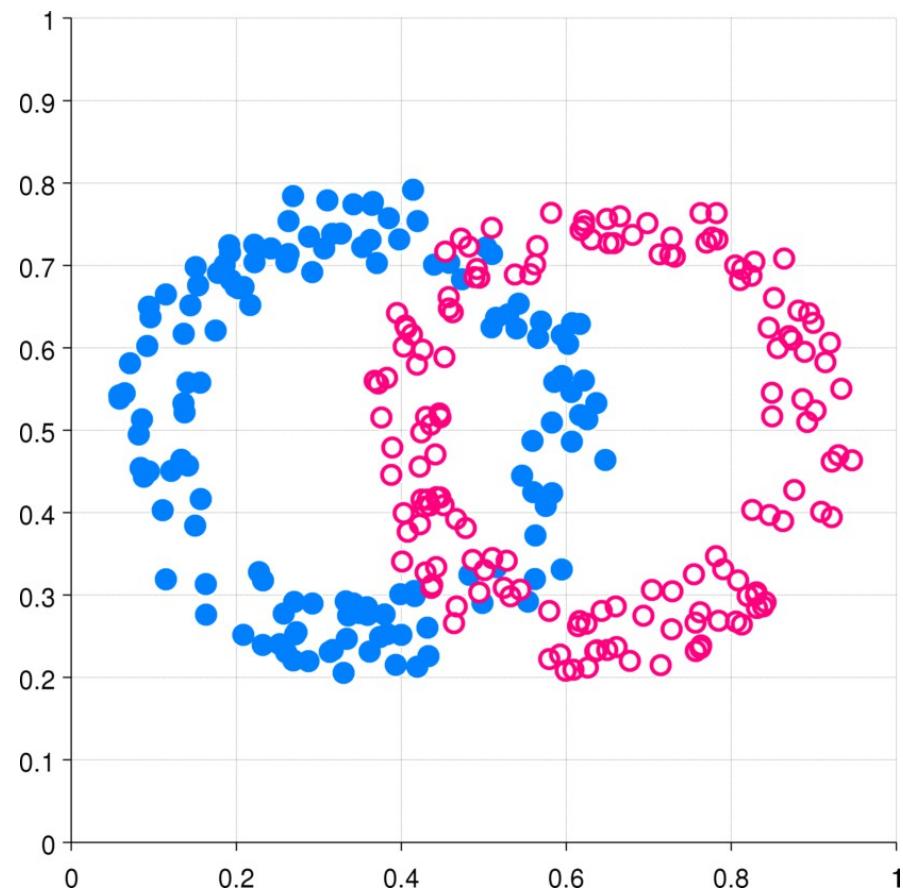
- Points in a cluster share some general property that derives from the entire set of points



Types of clustering

Conceptual clusters

- Points in a cluster share some general property that derives from the entire set of points

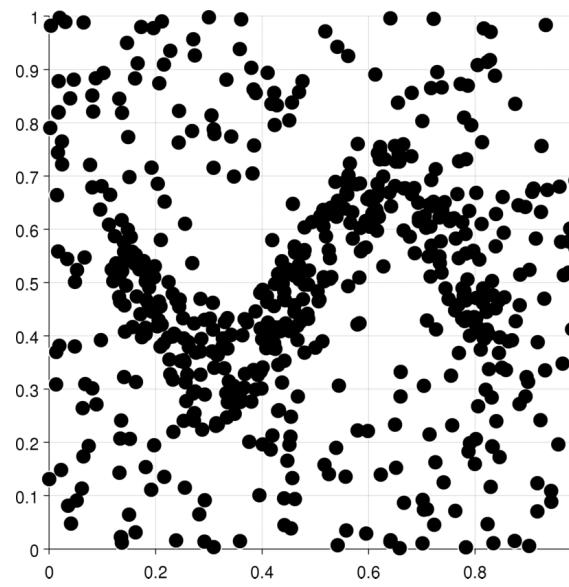
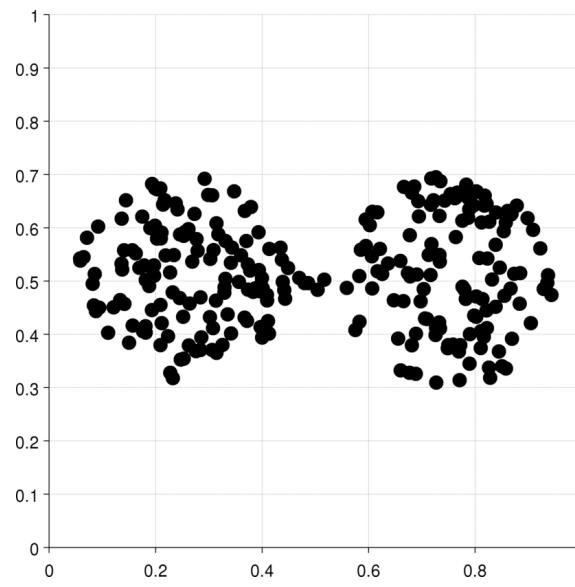
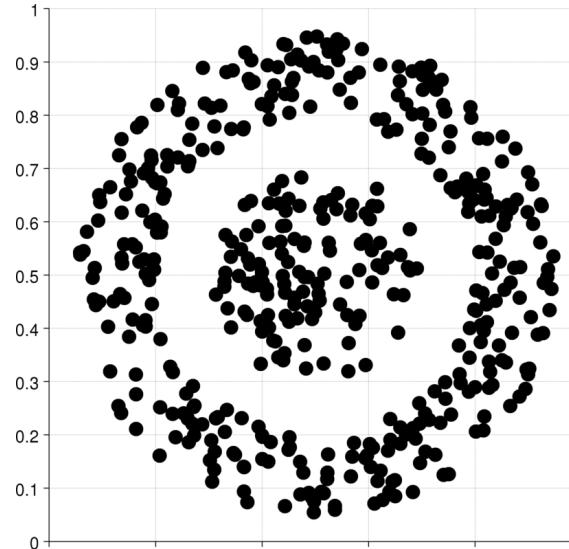




Group exercise

Using the five criteria

- How will these points be clustered?
- How many clusters?



Well-separated

- Each point is closer to all points in its cluster than any point in another cluster

Center-based

- Each point is closer to the center of its cluster than to the center of any other cluster

Contiguity-based

- Each point is closer to at least one point in its cluster than to any point in another cluster

Density-based

- Clusters are regions of high density separated by regions of low density

Conceptual clusters

- Points in a cluster share some general property that derives from the entire set of points

K-means clustering

Select K points as initial centroids

Repeat

- Form K clusters by assigning each point to its closest centroid
- Recompute the centroids of each cluster

Until centroids do not change

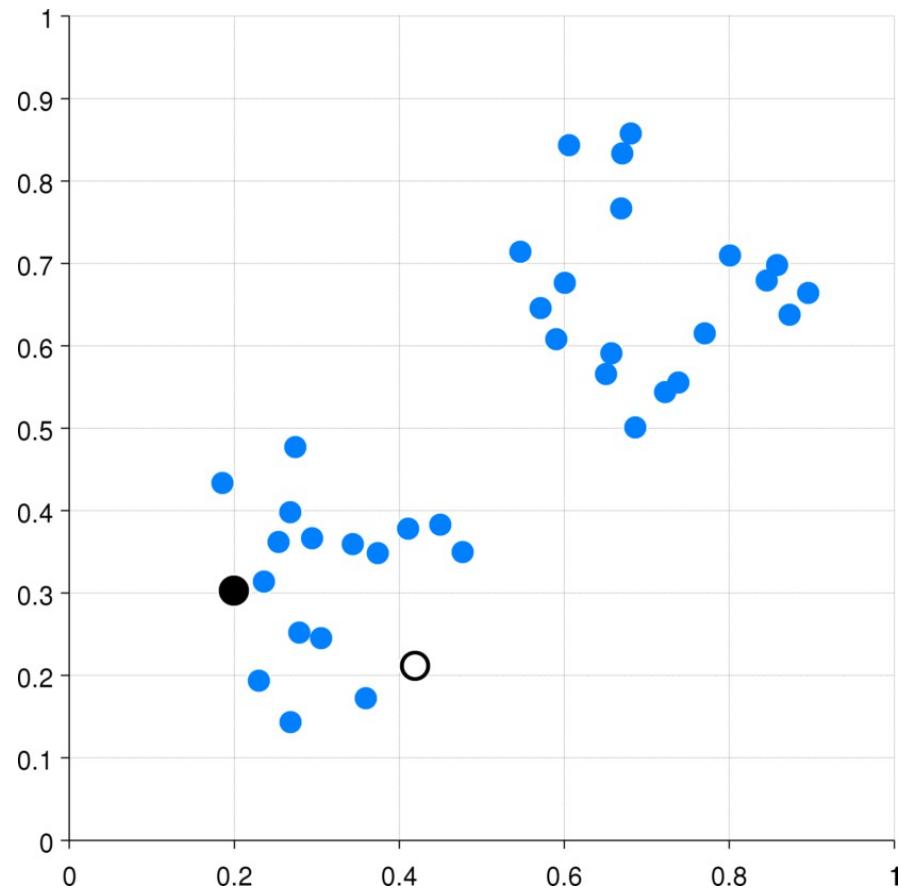
K-means clustering

Select K points as initial centroids

Repeat

- Form K clusters by assigning each point to its closest centroid
- Recompute the centroids of each cluster

Until centroids do not change



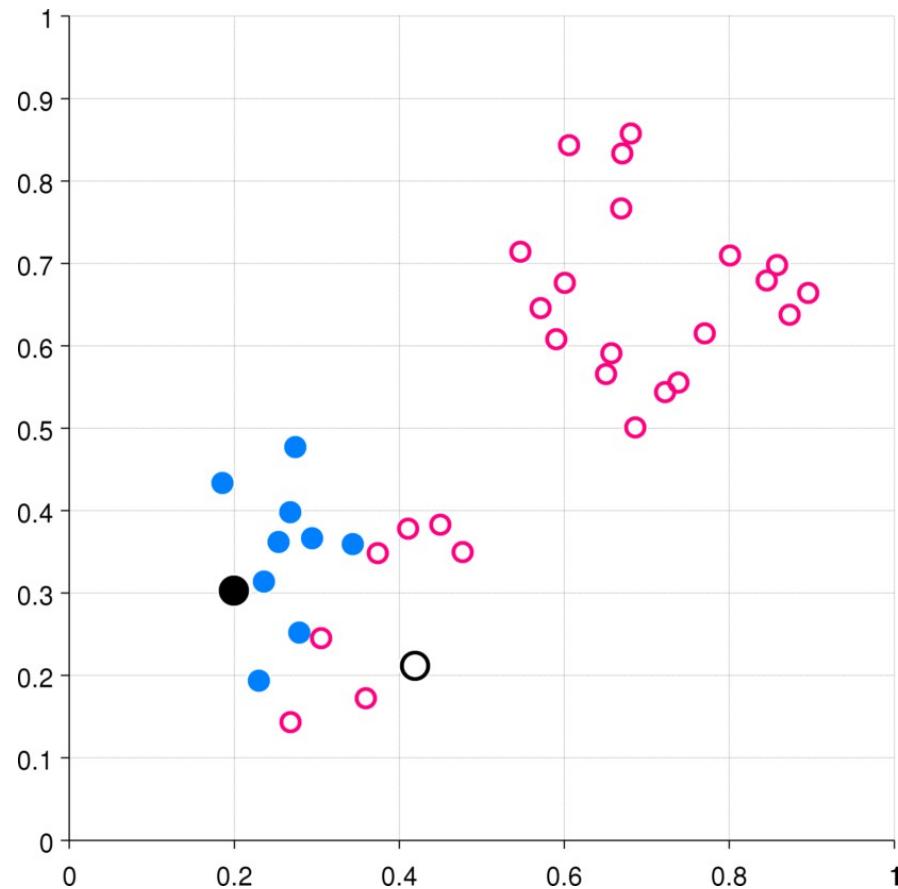
K-means clustering

Select K points as initial centroids

Repeat

- Form K clusters by assigning each point to its closest centroid
- Recompute the centroids of each cluster

Until centroids do not change



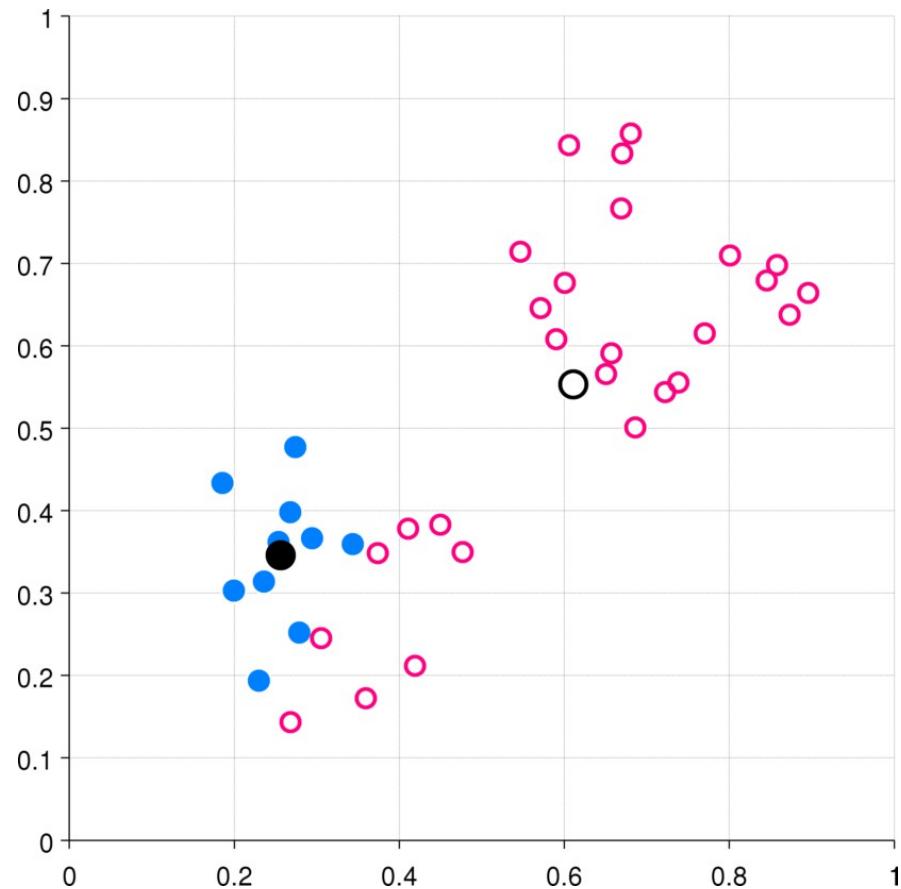
K-means clustering

Select K points as initial centroids

Repeat

- Form K clusters by assigning each point to its closest centroid
- Recompute the centroids of each cluster

Until centroids do not change



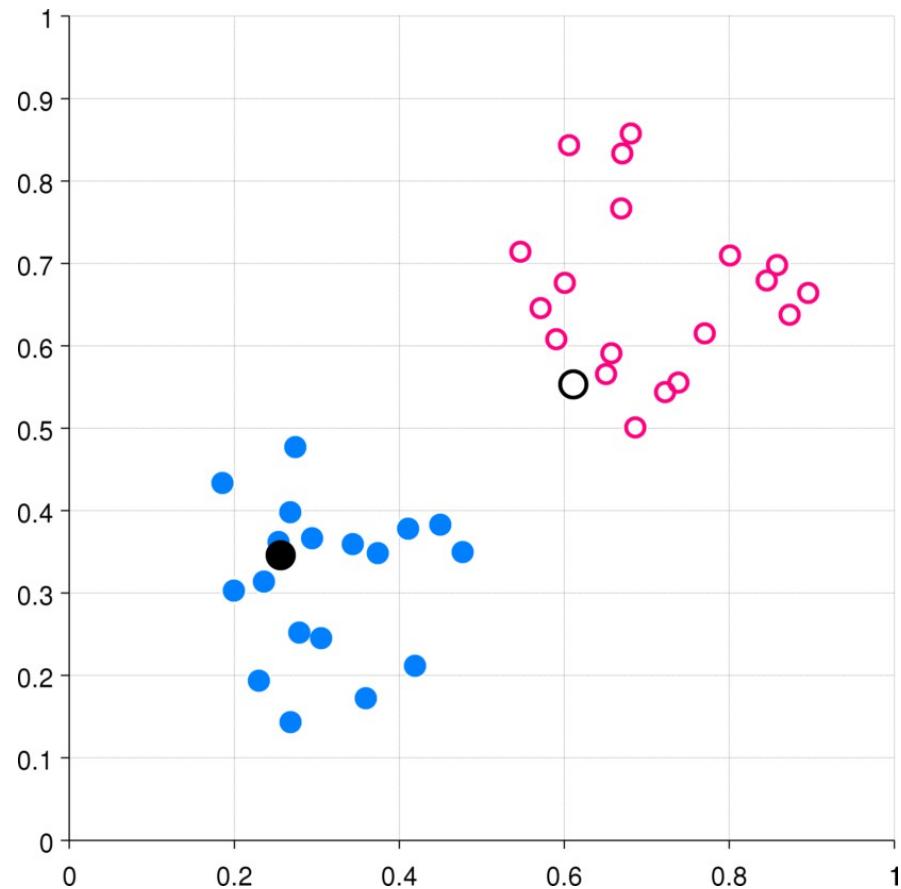
K-means clustering

Select K points as initial centroids

Repeat

- Form K clusters by assigning each point to its closest centroid
- Recompute the centroids of each cluster

Until centroids do not change



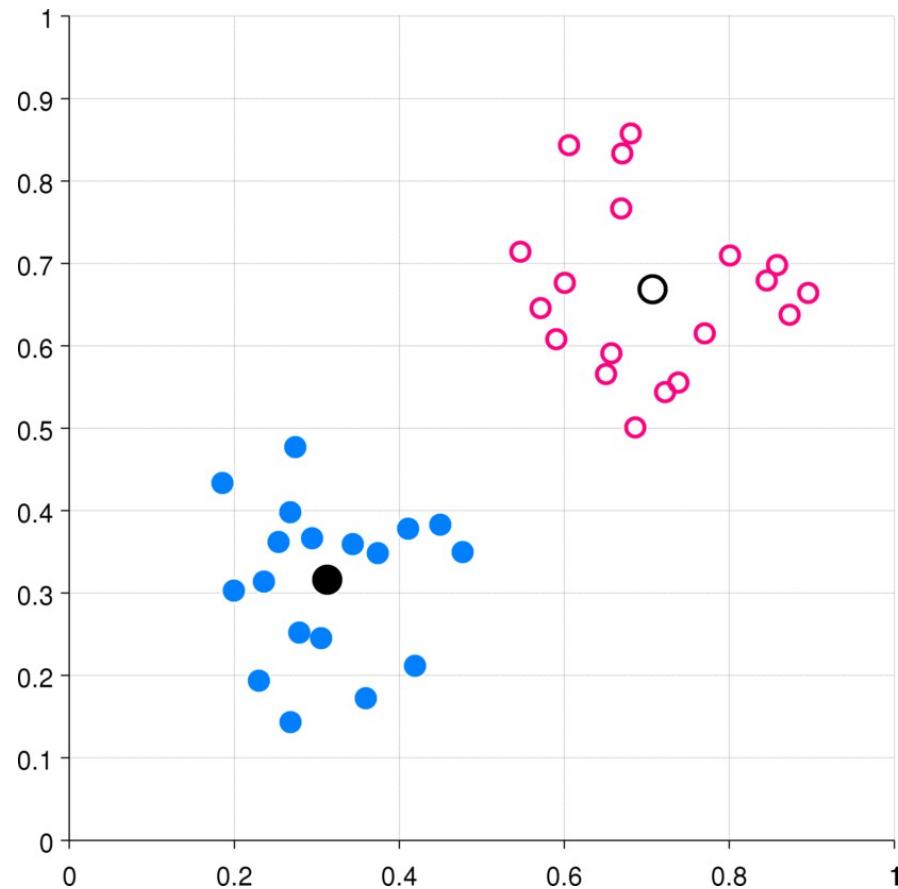
K-means clustering

Select K points as initial centroids

Repeat

- Form K clusters by assigning each point to its closest centroid
- Recompute the centroids of each cluster

Until centroids do not change



K-means clustering

How do I

- Find the closest centroid?
 - Use a suitable **dissimilarity/similarity measure**
- Compute the cluster centroids
 - Depends on dissimilarity/similarity measure
 - For example, for Euclidean distance the mean is optimal
(See Section 8.2.6 in Tan et al.)



Group exercise

**Using pen-and-paper k-means,
cluster the following data objects**

- Number of clusters
 - K=2
 - Distance measure
 - Euclidean
 - Computation of centroid
 - Mean of cluster members
 - Initial centroids
 - For example the first two data objects
 - In case of any ties, flip a coin to decide
-
- **Data objects**

$$x = \{42, 60, 17, 48, 12\}$$

Select K points as initial centroids

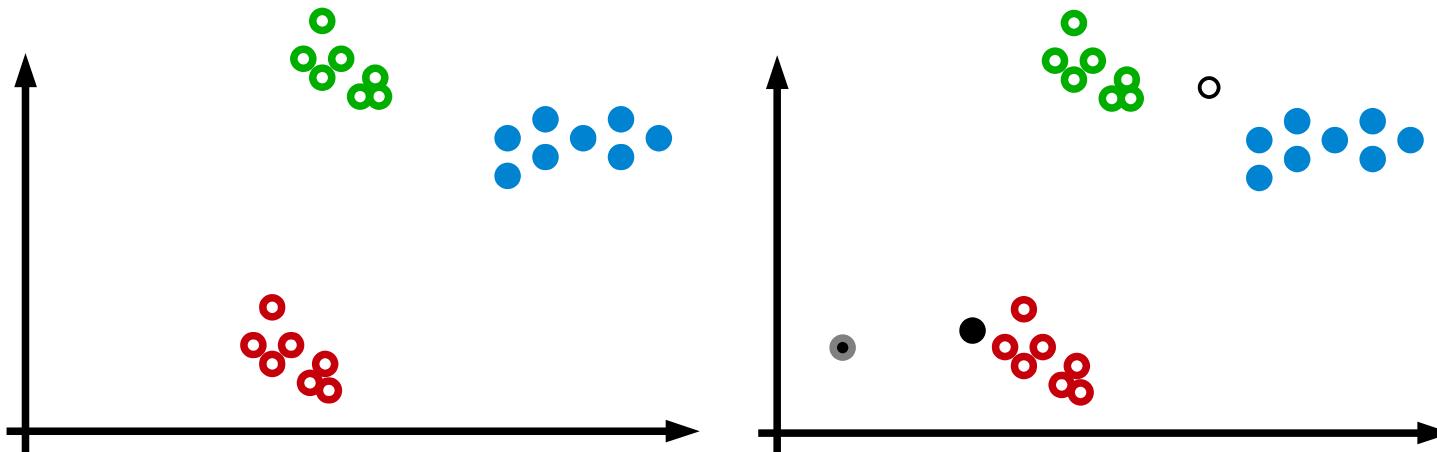
Repeat

- Form K clusters by assigning each point to its closest centroid
- Recompute the centroids of each cluster

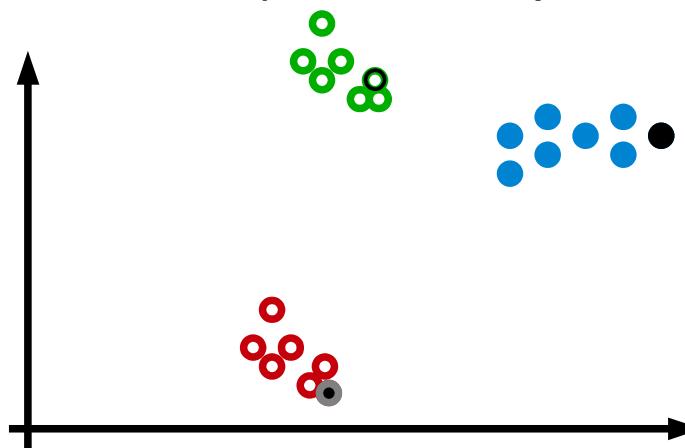
Until centroids do not change



How will the data to the left be clustered given the initialization of the three centroids below to the right?



- What could we do if we have an empty cluster?
- What could be a good initialization procedure? (Farthest First)



Agglomerative hierarchical clustering

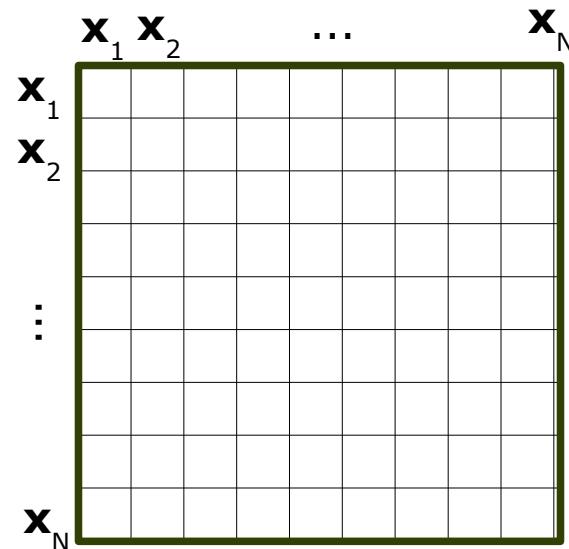
Compute the proximity matrix

Repeat

- Merge the two closest clusters
- Update the proximity matrix to reflect the proximity between the new cluster and the original clusters

Until only one cluster remains

$$D_{ij} = \text{distance}(x_i, x_j)$$



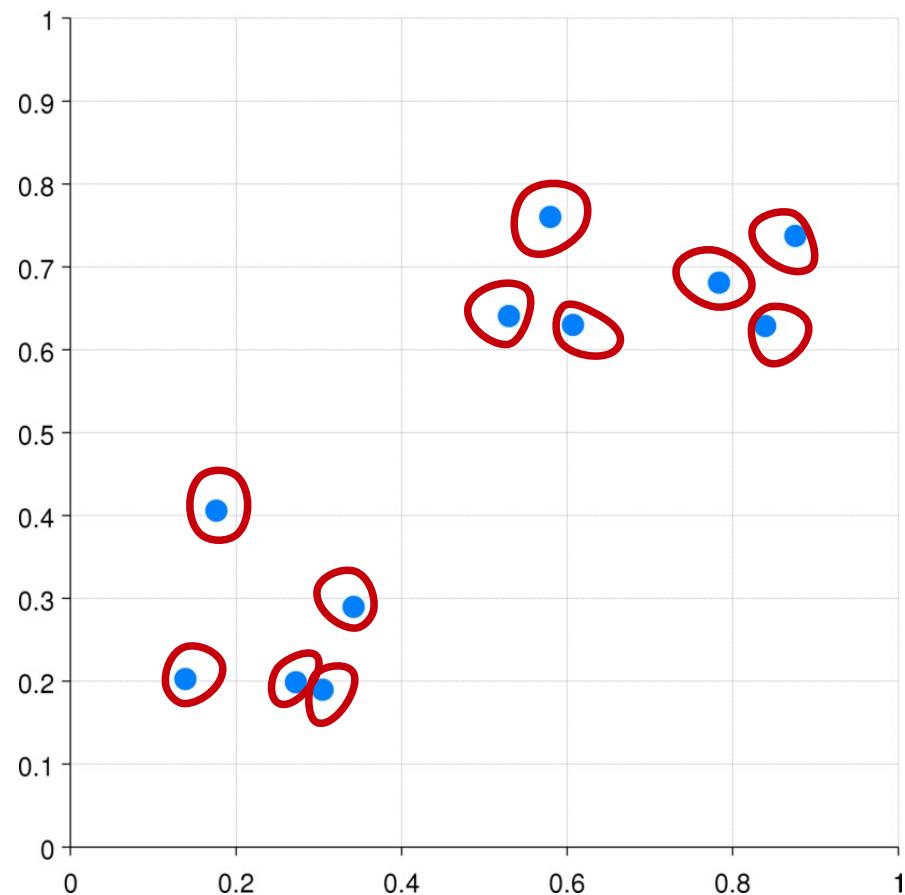
Agglomerative hierarchical clustering

Compute the proximity matrix

Repeat

- Merge the two closest clusters
- Update the proximity matrix to reflect the proximity between the new cluster and the original clusters

Until only one cluster remains



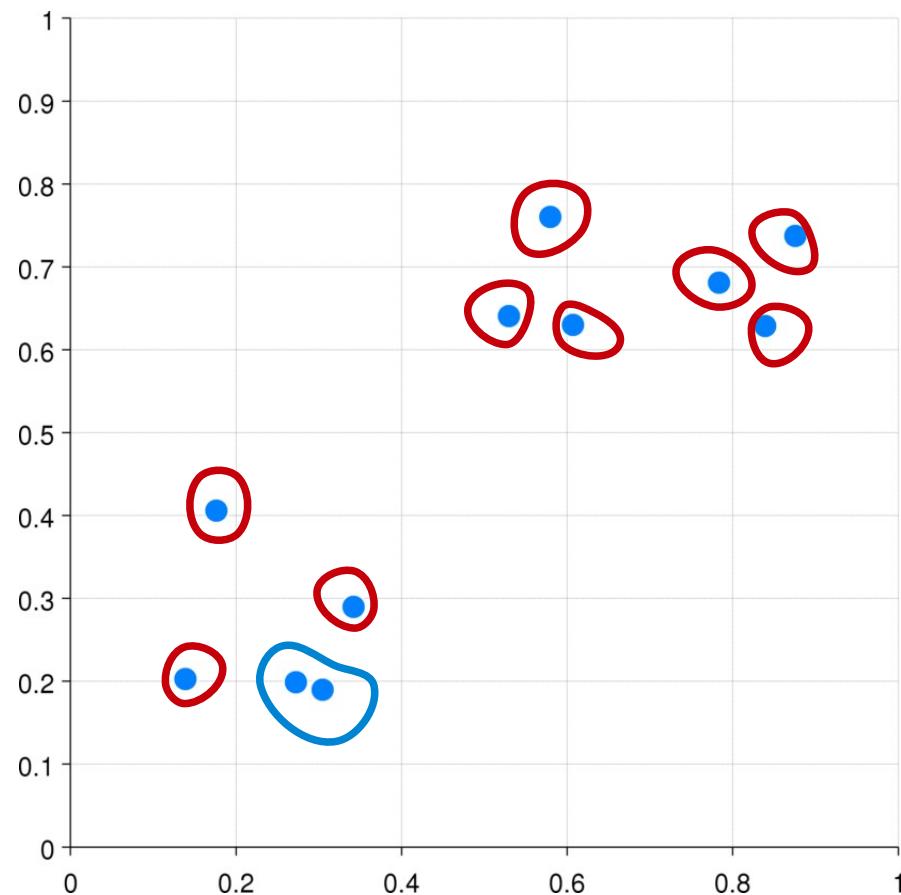
Agglomerative hierarchical clustering

Compute the proximity matrix

Repeat

- Merge the two closest clusters
- Update the proximity matrix to reflect the proximity between the new cluster and the original clusters

Until only one cluster remains



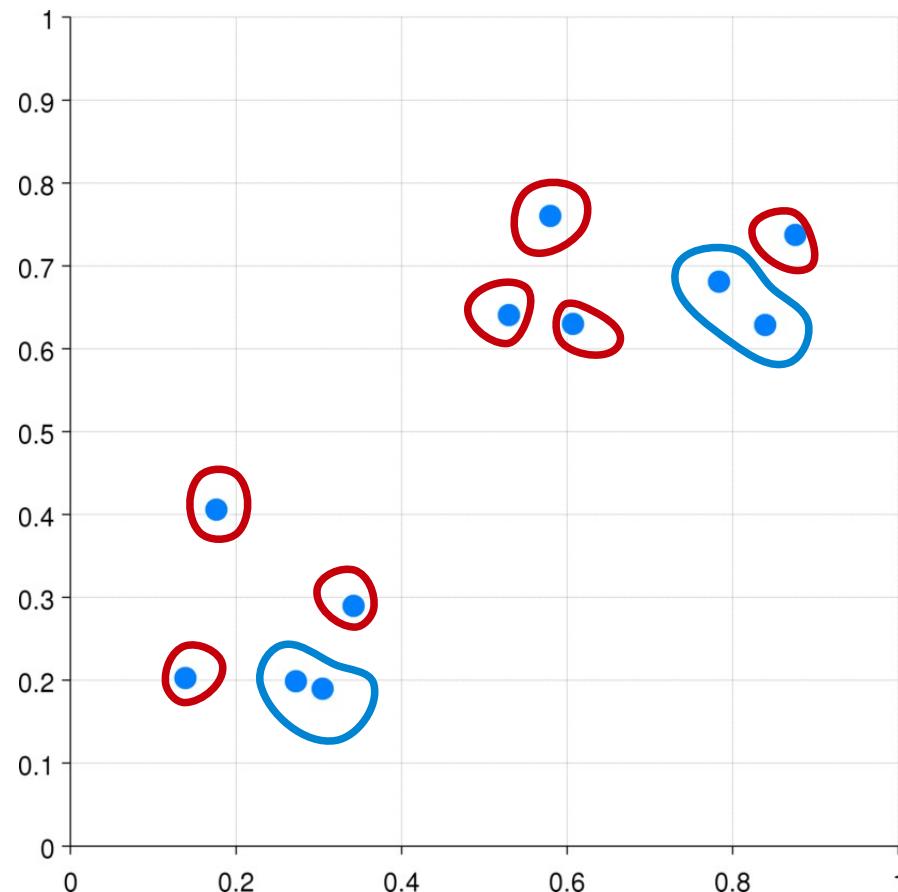
Agglomerative hierarchical clustering

Compute the proximity matrix

Repeat

- Merge the two closest clusters
- Update the proximity matrix to reflect the proximity between the new cluster and the original clusters

Until only one cluster remains



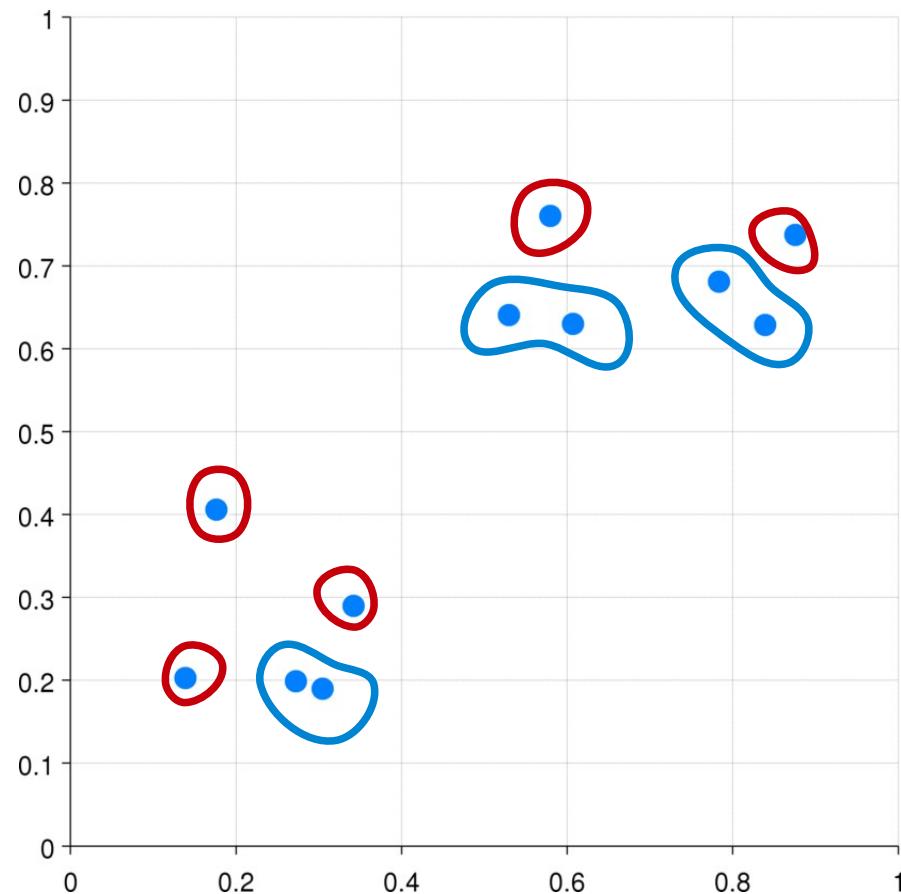
Agglomerative hierarchical clustering

Compute the proximity matrix

Repeat

- Merge the two closest clusters
- Update the proximity matrix to reflect the proximity between the new cluster and the original clusters

Until only one cluster remains



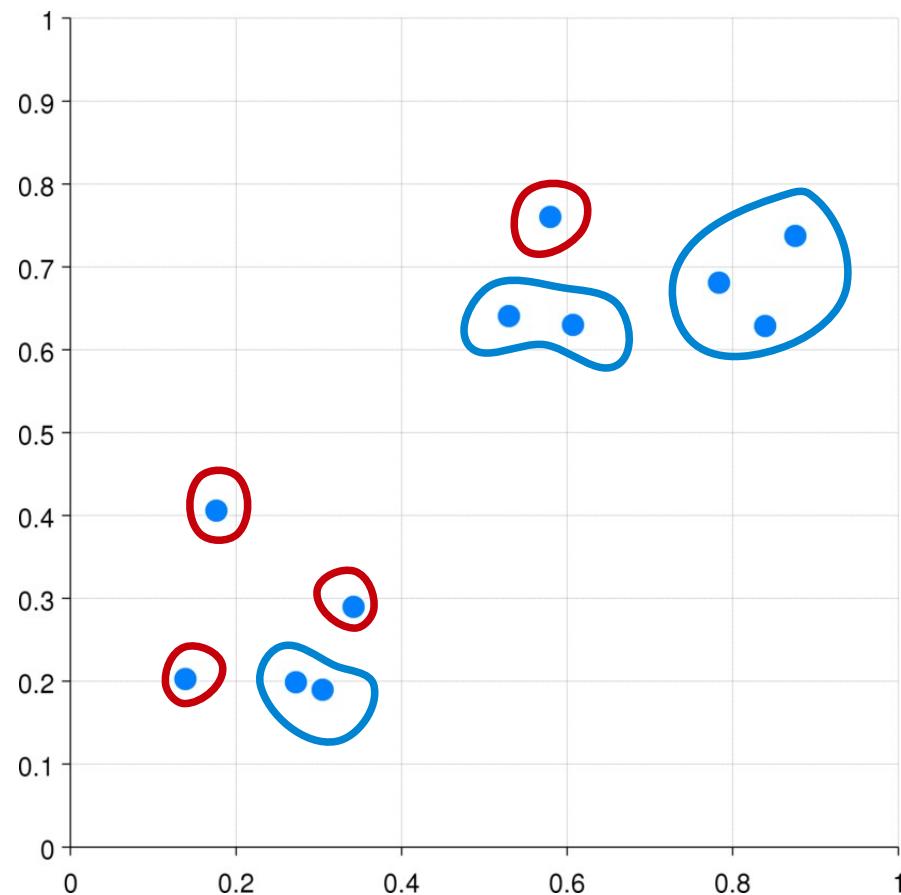
Agglomerative hierarchical clustering

Compute the proximity matrix

Repeat

- Merge the two closest clusters
- Update the proximity matrix to reflect the proximity between the new cluster and the original clusters

Until only one cluster remains



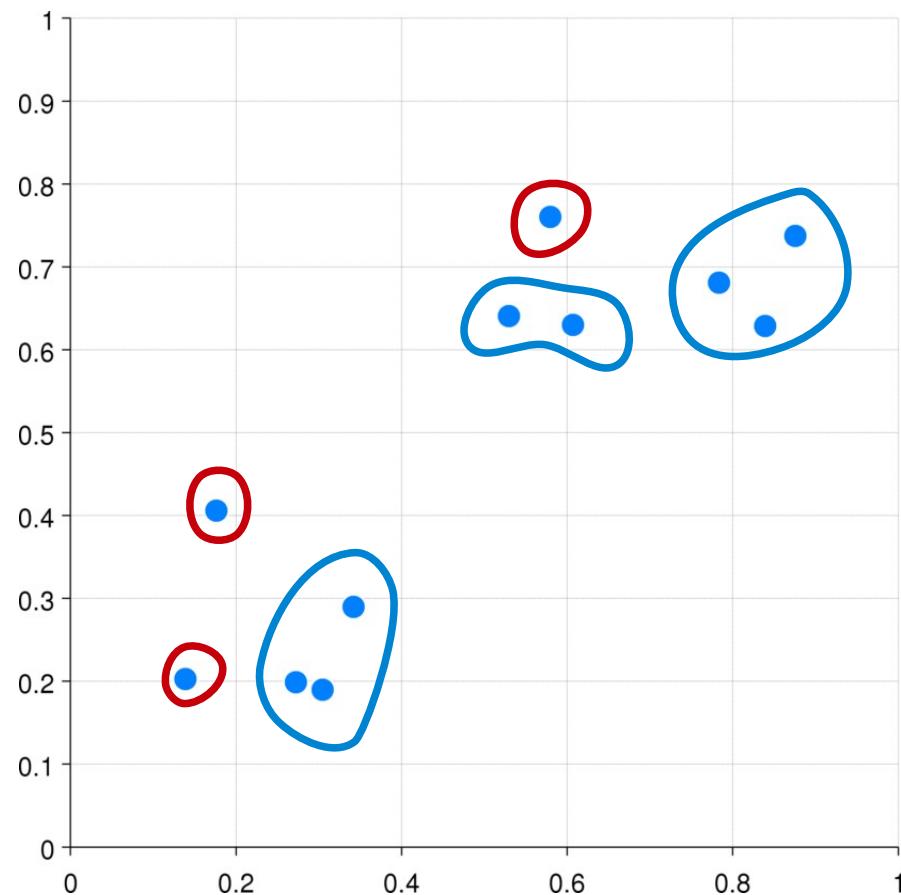
Agglomerative hierarchical clustering

Compute the proximity matrix

Repeat

- Merge the two closest clusters
- Update the proximity matrix to reflect the proximity between the new cluster and the original clusters

Until only one cluster remains



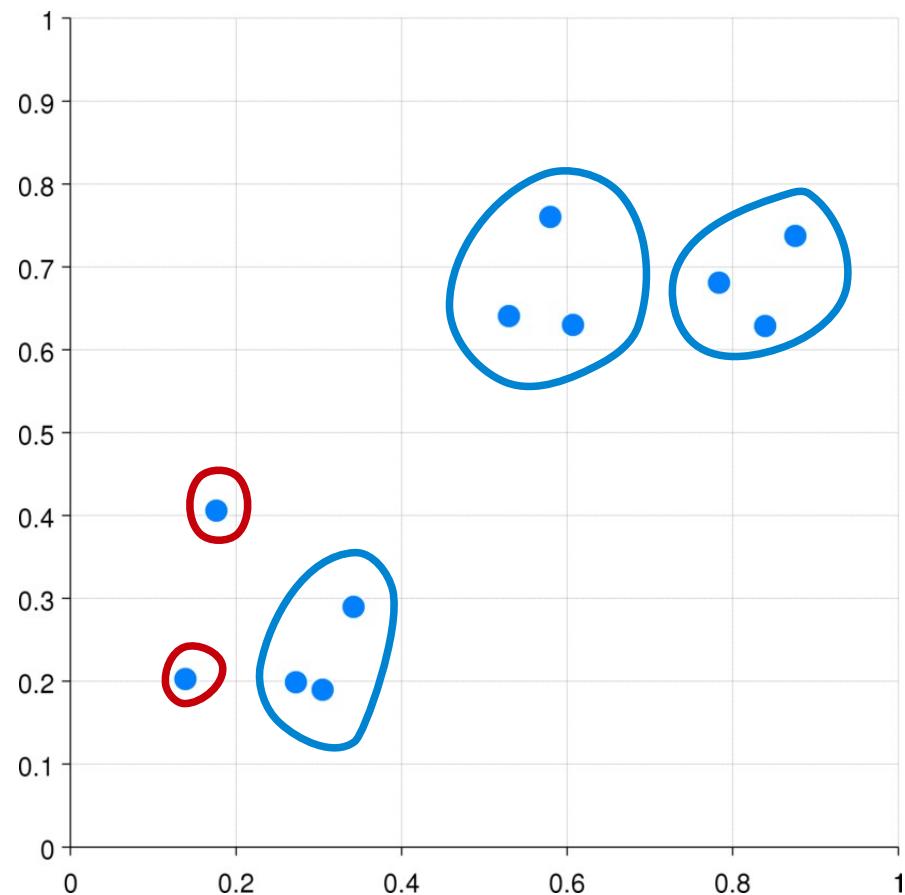
Agglomerative hierarchical clustering

Compute the proximity matrix

Repeat

- Merge the two closest clusters
- Update the proximity matrix to reflect the proximity between the new cluster and the original clusters

Until only one cluster remains



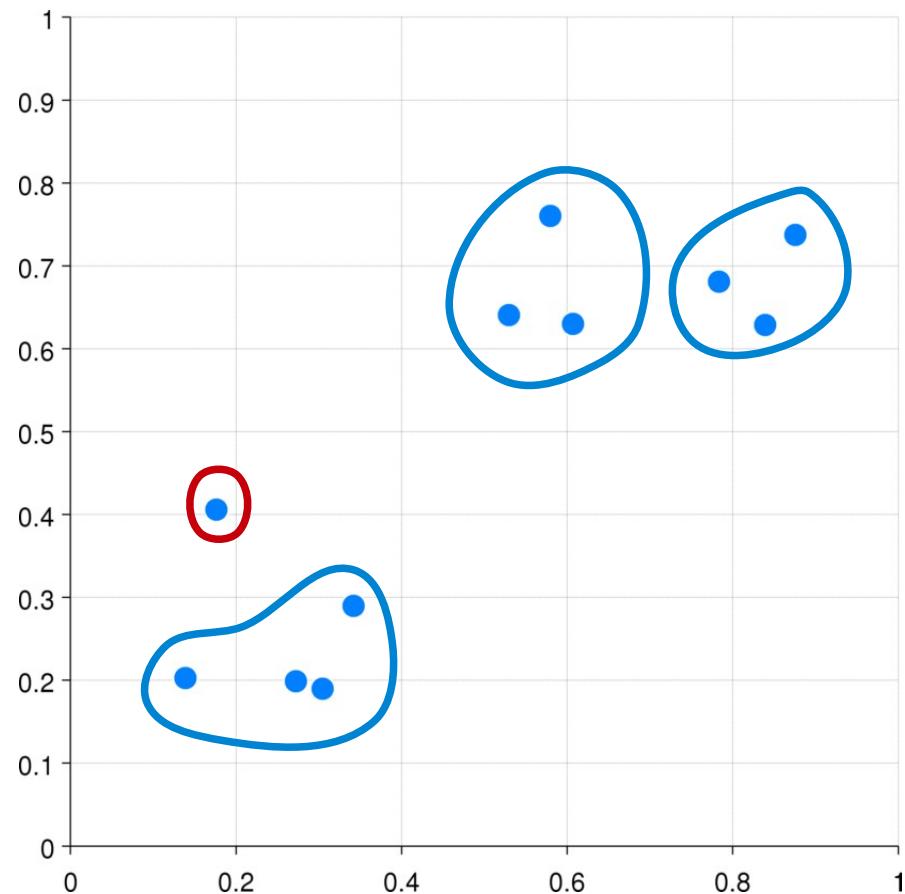
Agglomerative hierarchical clustering

Compute the proximity matrix

Repeat

- Merge the two closest clusters
- Update the proximity matrix to reflect the proximity between the new cluster and the original clusters

Until only one cluster remains



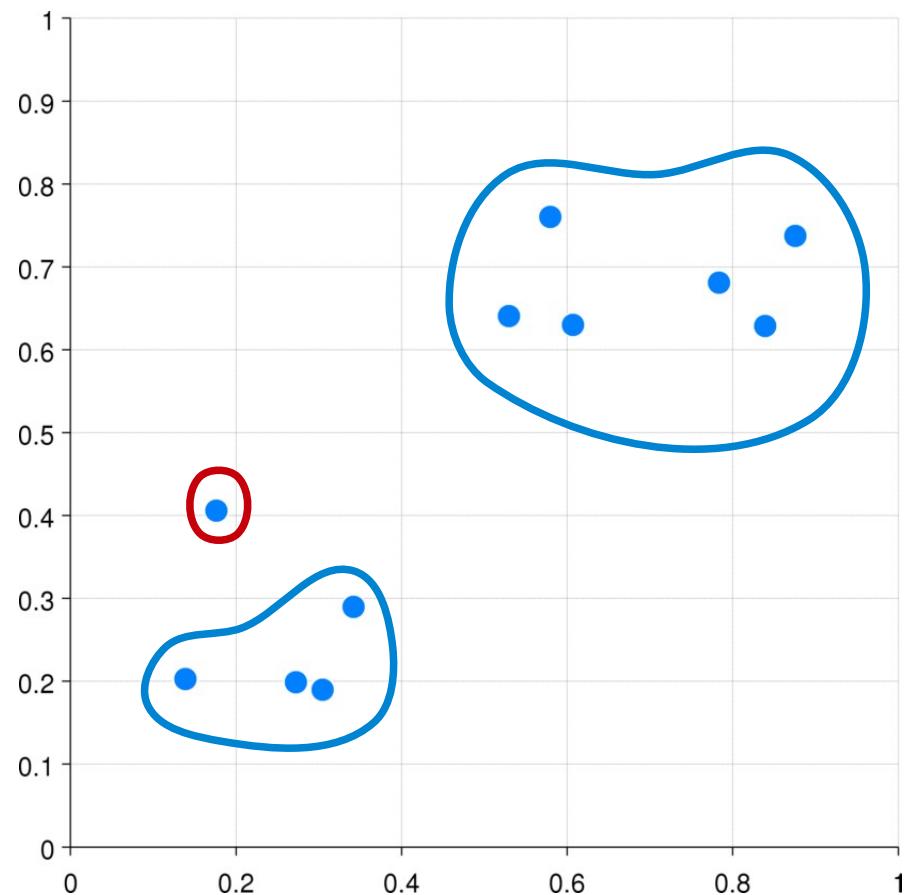
Agglomerative hierarchical clustering

Compute the proximity matrix

Repeat

- Merge the two closest clusters
- Update the proximity matrix to reflect the proximity between the new cluster and the original clusters

Until only one cluster remains



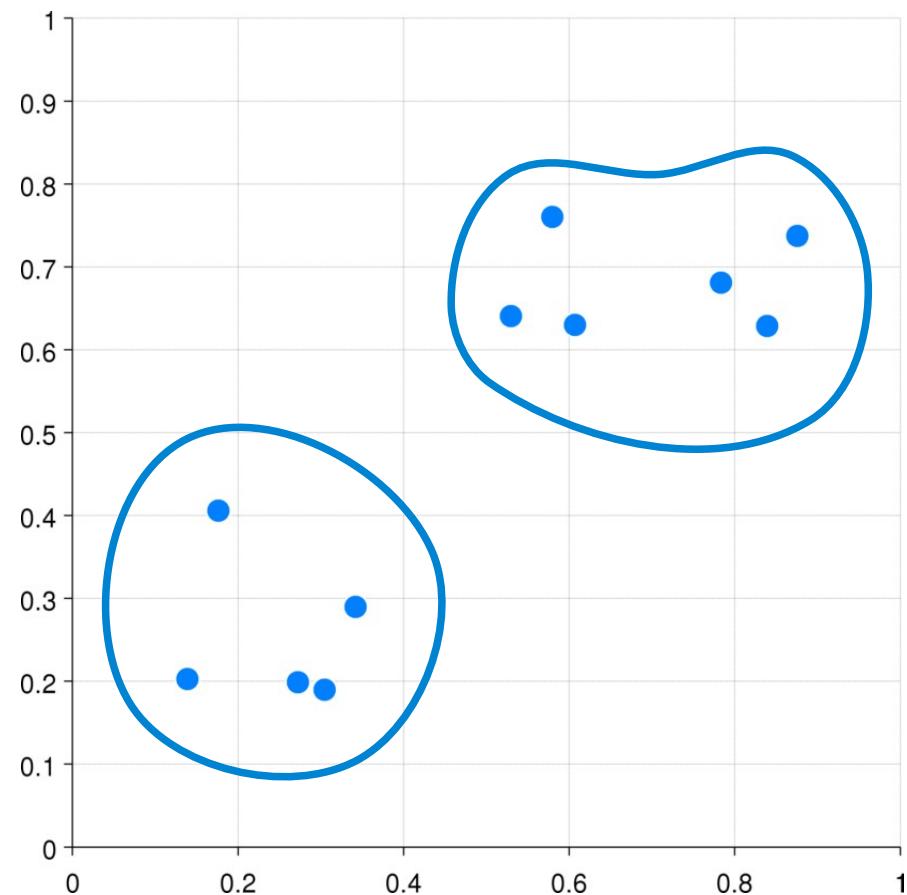
Agglomerative hierarchical clustering

Compute the proximity matrix

Repeat

- Merge the two closest clusters
- Update the proximity matrix to reflect the proximity between the new cluster and the original clusters

Until only one cluster remains



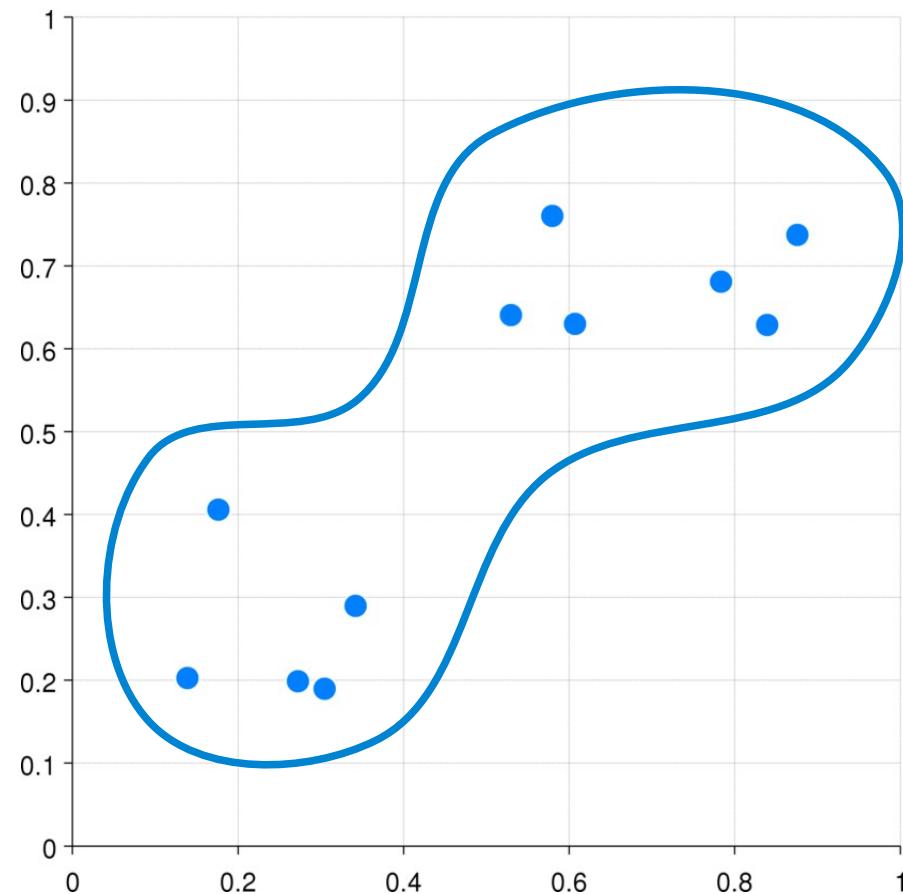
Agglomerative hierarchical clustering

Compute the proximity matrix

Repeat

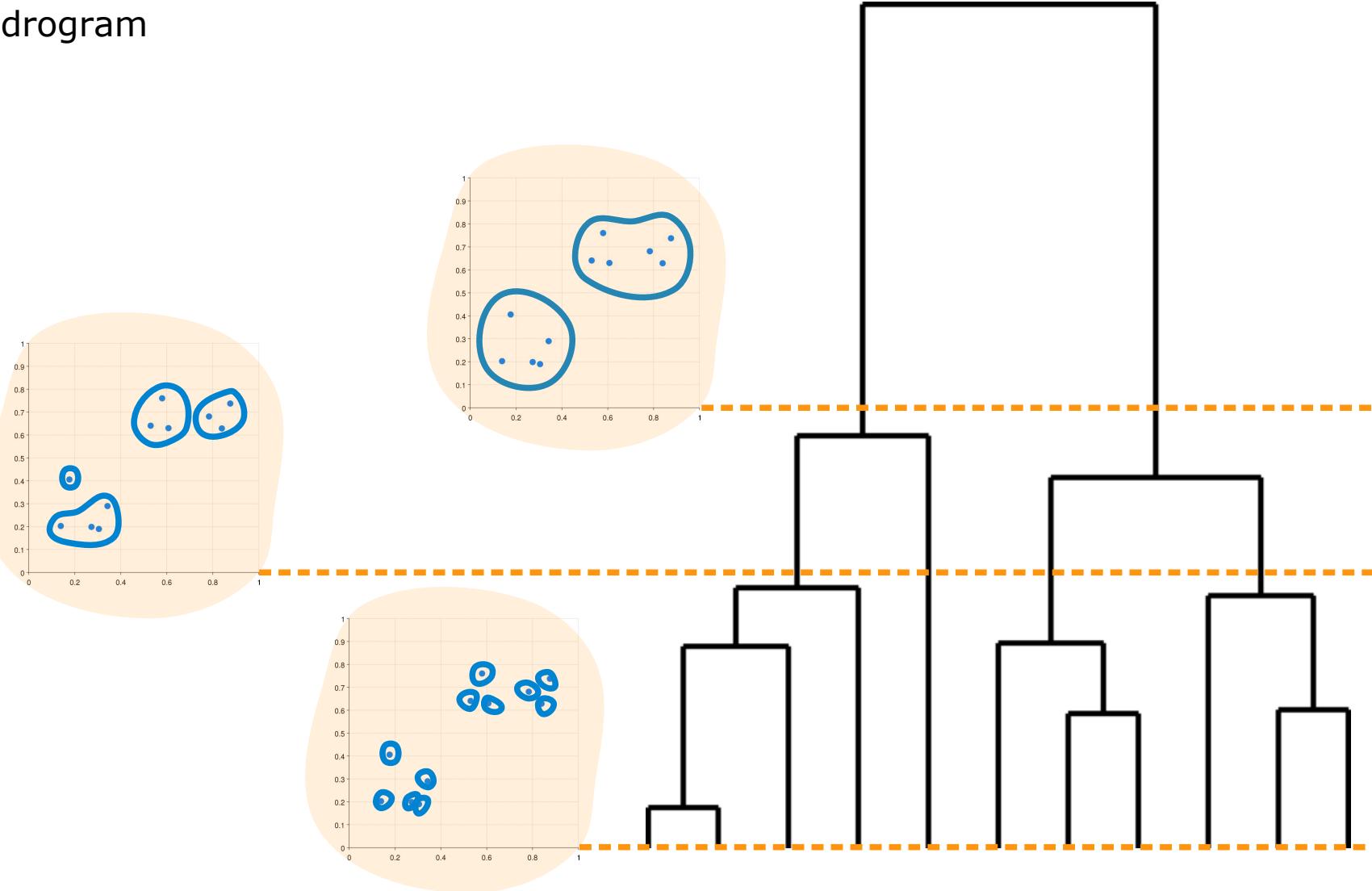
- Merge the two closest clusters
- Update the proximity matrix to reflect the proximity between the new cluster and the original clusters

Until only one cluster remains



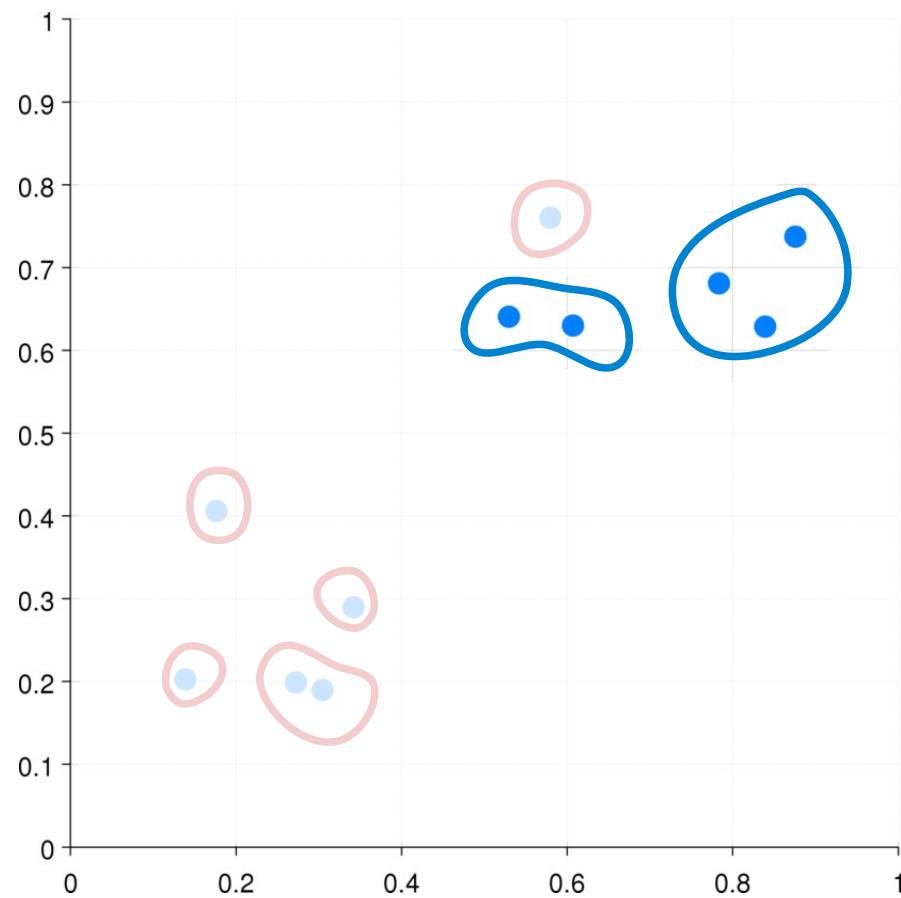
Agglomerative hierarchical clustering

- Dendrogram



Similarity between clusters

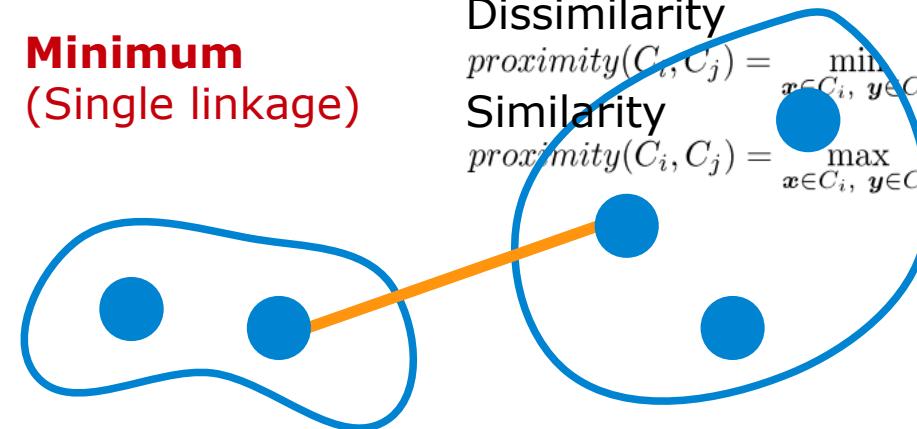
- The **key operation** in agglomerative hierarchical clustering is measuring **similarity between clusters**



Similarity between clusters

- Can be computed using **similarities between objects**

Minimum
(Single linkage)



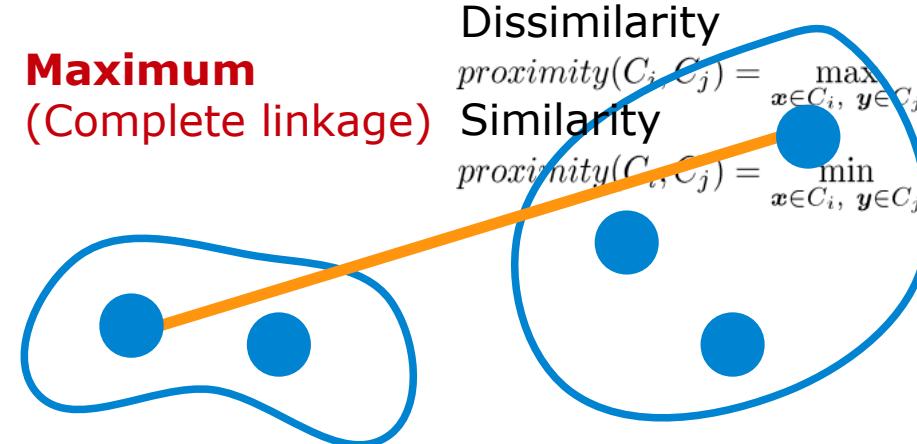
Dissimilarity

$$\text{proximity}(C_i, C_j) = \min_{x \in C_i, y \in C_j} \text{proximity}(x, y)$$

Similarity

$$\text{proximity}(C_i, C_j) = \max_{x \in C_i, y \in C_j} \text{proximity}(x, y)$$

Maximum
(Complete linkage)



Dissimilarity

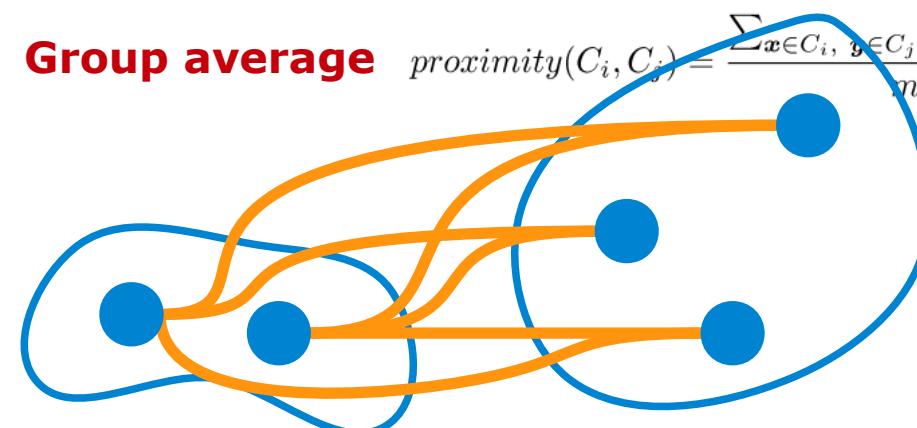
$$\text{proximity}(C_i, C_j) = \max_{x \in C_i, y \in C_j} \text{proximity}(x, y)$$

Similarity

$$\text{proximity}(C_i, C_j) = \min_{x \in C_i, y \in C_j} \text{proximity}(x, y)$$

Group average

$$\text{proximity}(C_i, C_j) = \frac{\sum_{x \in C_i, y \in C_j} \text{proximity}(x, y)}{m_i \cdot m_j}$$



C_i : Observations in cluster i

C_j : Observations in cluster j

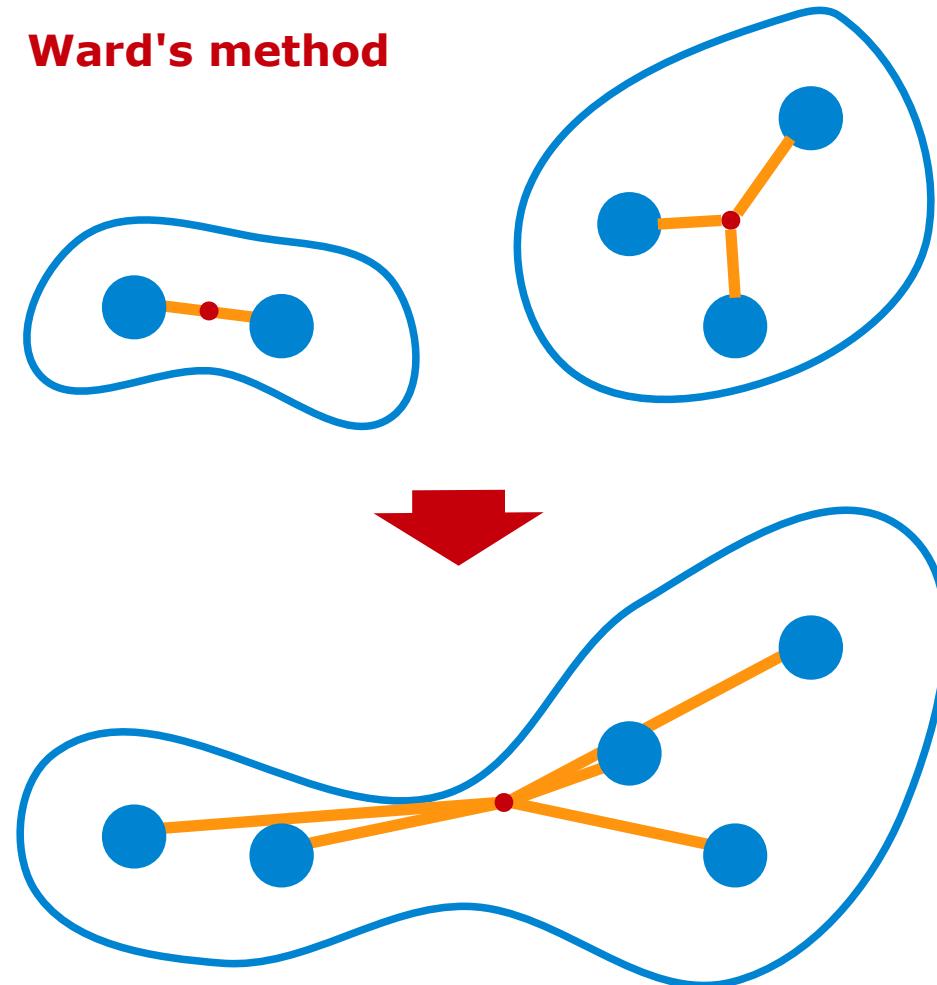
m_i : Number of observations in cluster i

m_j : Number of observations in cluster j

Similarity between clusters

- Increase in sum of squared error after merging the two clusters

Ward's method





Group exercise

Using pen-and-paper agglomerative hierarchical clustering, **cluster** the following data objects and draw the **dendrogram**

- Distance measure
 - Euclidean
 - Similarity between clusters
 - Minimum (Single linkage)

- **Data objects**

$$x = \{42, 60, 17, 48, 12\}$$

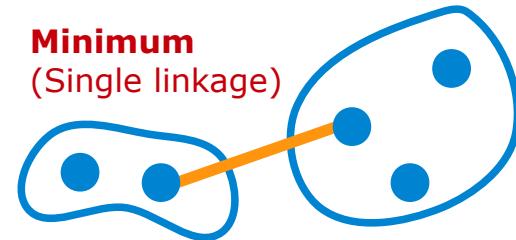
Compute the proximity matrix

Repeat

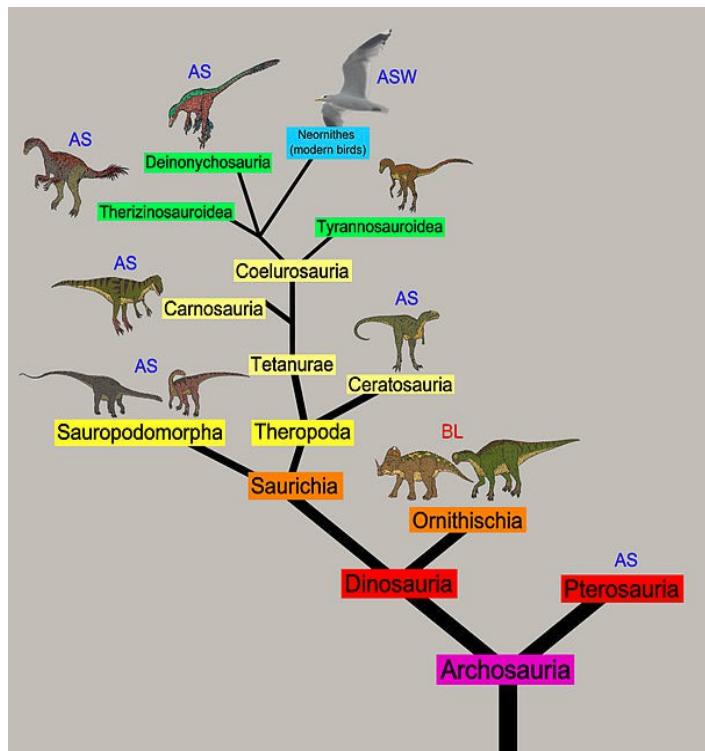
- Merge the two closest clusters
 - Update the proximity matrix to reflect the proximity between the new cluster and the original clusters

Until only one cluster remains

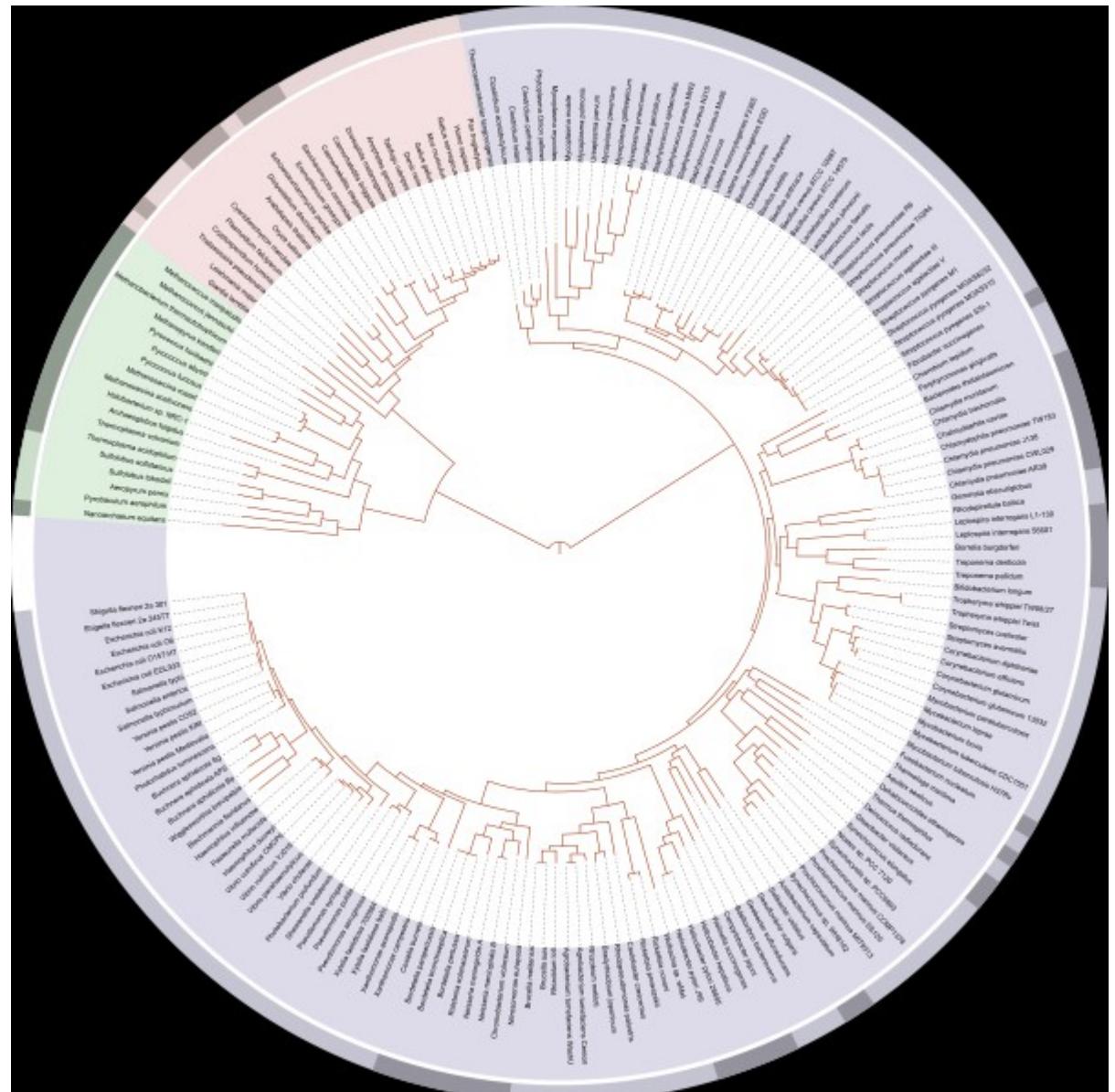
Minimum (Single linkage)



Phylogenetic trees



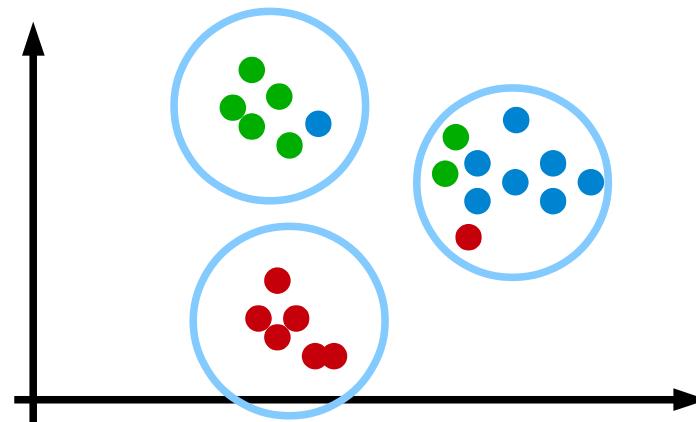
http://en.wikipedia.org/wiki/Phylogenetic_tree



http://en.wikipedia.org/wiki/File:Tree_of_life_SVG.svg

Cluster Purity measures when class labels are available

Motivation: Evaluate the extent to which manual classification process can be automatically produced by cluster analysis by comparing clustering to “ground truth”



Supervised measures of cluster validity

If we have (supervised) class labels, they can be used to measure the quality of the clustering

- **Cluster purity measures**

- Entropy
- Gini
- Class error

- **Class label accuracy measures**

- Precision
- Recall
- F-measure

- **Binary similarity measures**

- Simple matching coefficient
- Jaccard coefficient

Supervised measures of cluster validity

Cluster purity measures

- **Entropy**

$$Entropy(i) = - \sum_j p_{ij} \log_2 p_{ij}$$

- **Gini**

$$Gini(i) = 1 - \sum_j p_{ij}^2$$

- **Class error**

$$ClassError(i) = 1 - \max_j p_{ij}$$

- **Purity**

$$Purity(i) = \max_j p_{ij}$$

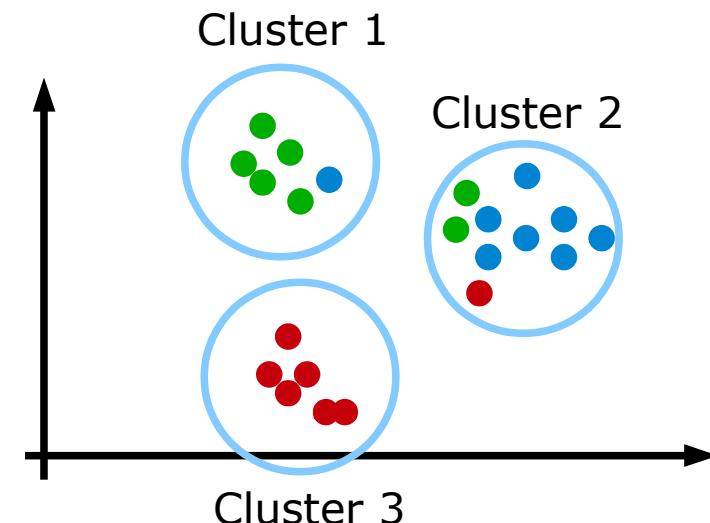
p_{ij} : Probability that member of cluster i belongs to class j $p_{ij} = \frac{m_{ij}}{m_i}$

m_{ij} : Number of objects of class j in cluster i

m_i : Total number of objects in cluster i



What is the Entropy, Gini, Class. Error and Purity for each of the three clusters given above?



	Entropy	Gini	Class. Error	Purity
Cluster 1	0.65	5/18	1/6	5/6
Cluster 2	1.16	23/50	3/10	7/10
Cluster 3	0	0	0	1

Supervised measures of cluster validity

Class label accuracy measures

- **Precision**

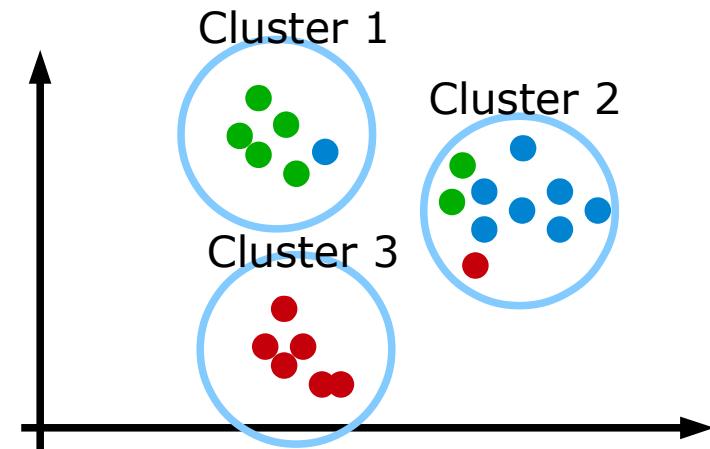
$$\text{Precision}(i, j) = \frac{m_{ij}}{m_i} = p_{ij}$$

- **Recall**

$$\text{Recall}(i, j) = \frac{m_{ij}}{m_j}$$

- **F-measure**

$$F(i, j) = \frac{2 \cdot \text{Precision}(i, j) \cdot \text{Recall}(i, j)}{\text{Precision}(i, j) + \text{Recall}(i, j)} = \frac{2m_{ij}}{m_i + m_j}$$

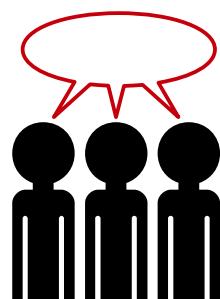


m_{ij} : Number of objects of class j in cluster i

m_i : Total number of objects in cluster i

m_j : Total number of objects in class j

p_{ij} : Fraction of objects of class j in cluster i



What is the Precision, recall and F-measure of the blue class in cluster 2?

$$\text{Precision}(\text{blue}, \text{cluster 2}) = 7/10$$

$$\text{Recall}(\text{blue}, \text{cluster 2}) = 7/8$$

$$F(\text{blue}, \text{cluster 2}) = 2 \cdot 7 / (8 + 10) = 7/9$$

Supervised measures of cluster validity

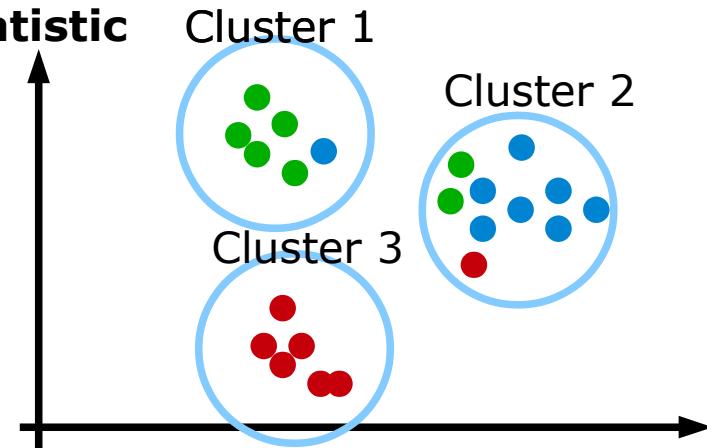
Binary similarity measures

- Simple matching coefficient (SMC)/Rand statistic

$$\text{SMC}(\mathbf{x}, \mathbf{y}) = \frac{f_{00} + f_{11}}{K}$$

- Jaccard coefficient

$$J(\mathbf{x}, \mathbf{y}) = \frac{f_{11}}{K - f_{00}}$$



K : Total number of **pairs of objects**, $N \cdot (N-1)/2$

f_{00} : Number of object pairs in **different class** assigned to **different clusters**

f_{11} : Number of objects pairs in **same class** assigned to **same cluster**

In our example we find:

$$K=22 \cdot (22-1)/2=231$$

$$f_{11}=(5 \cdot (5-1)/2+1 \cdot (1-1)/2)_{c1}+(7 \cdot (7-1)/2+2 \cdot (2-1)/2+1 \cdot (1-1)/2)_{c2}+(6 \cdot (6-1)/2)_{c3}=10+22+15=47$$

$$f_{00}=(5 \cdot (7+1)+1 \cdot (2+1)+0 \cdot (2+7))_{c1 \rightarrow c2}+(5 \cdot 6+1 \cdot 6+0 \cdot 0)_{c1 \rightarrow c3}+(2 \cdot 6+7 \cdot 6+1 \cdot 0)_{c2 \rightarrow c3}=43+36+54=133$$

$$\text{SMC}=(47+133)/231=180/231$$

$$\text{Jaccard}=47/(231-133)=47/98$$

Exam question examples

QUESTION I:

We have a one dimensional data set of size $N = 5$ with data examples

$$x_1 = 1, x_2 = 3, x_3 = 6, x_4 = 7 \text{ and } x_5 = 12.$$

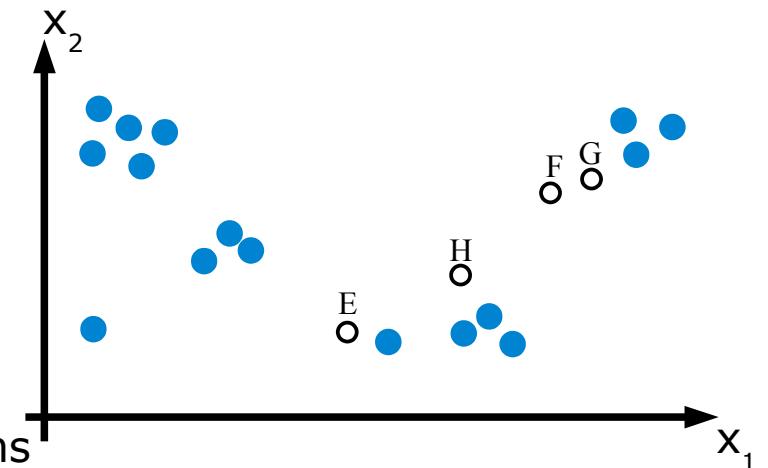
We run hierarchical clustering with a Euclidean dissimilarity between data points using group average linkage. We will use the following notation to summarize the dendrogram: $(x y)$ means that x and y are joined in the binary tree. x and y can themselves be binary trees. What is the order we build the tree?

- A. $12(34)5 \rightarrow (12)(34)5 \rightarrow ((12)(34))5 \rightarrow (((12)(34))5).$
- B. $12(34)5 \rightarrow 12((34)5) \rightarrow (12)((34)5) \rightarrow (12((34)5)).$
- C. $12(34)5 \rightarrow (12)(34)5 \rightarrow (12)((34)5) \rightarrow ((12)((34)5)).$
- D. $12(34)5 \rightarrow 12((34)5) \rightarrow 1(2((34)5)) \rightarrow (1(2((34)5))).$

QUESTION II:

Consider the clustering problem given to the right where blue dots are observations and black circles are the initial position of four centroids denoted E, F, G and H used to cluster the data by k-means using Euclidean distances as dissimilarity. Upon convergence of the k-means algorithm which one of the following statements is wrong?

- A. Cluster formed by centroid F will be empty
- B. Cluster formed by centroid E will contain 10 observations
- C. Clusters formed by centroid H will contain 4 observations
- D. Cluster formed by centroid G will contain 3 observations



(Solution: QI: A, QII: B)

02450 Introduction to machine learning and data modeling

$$f(x+\Delta x) = \sum_{i=0}^{\infty} \frac{(\Delta x)^i}{i!} f^{(i)}(x)$$

A complex mathematical expression featuring various symbols and numbers:

- Θ (purple)
- Ω (pink)
- δ (blue)
- $e^{i\pi}$ (purple)
- $\sqrt{17}$ (pink)
- \int_a^b (yellow)
- Σ (red)
- χ^2 (orange)
- ∞ (pink)
- \gg (yellow)
- $=$ (red)
- $\{2.7182818284$ (pink)
- $,$ (pink)

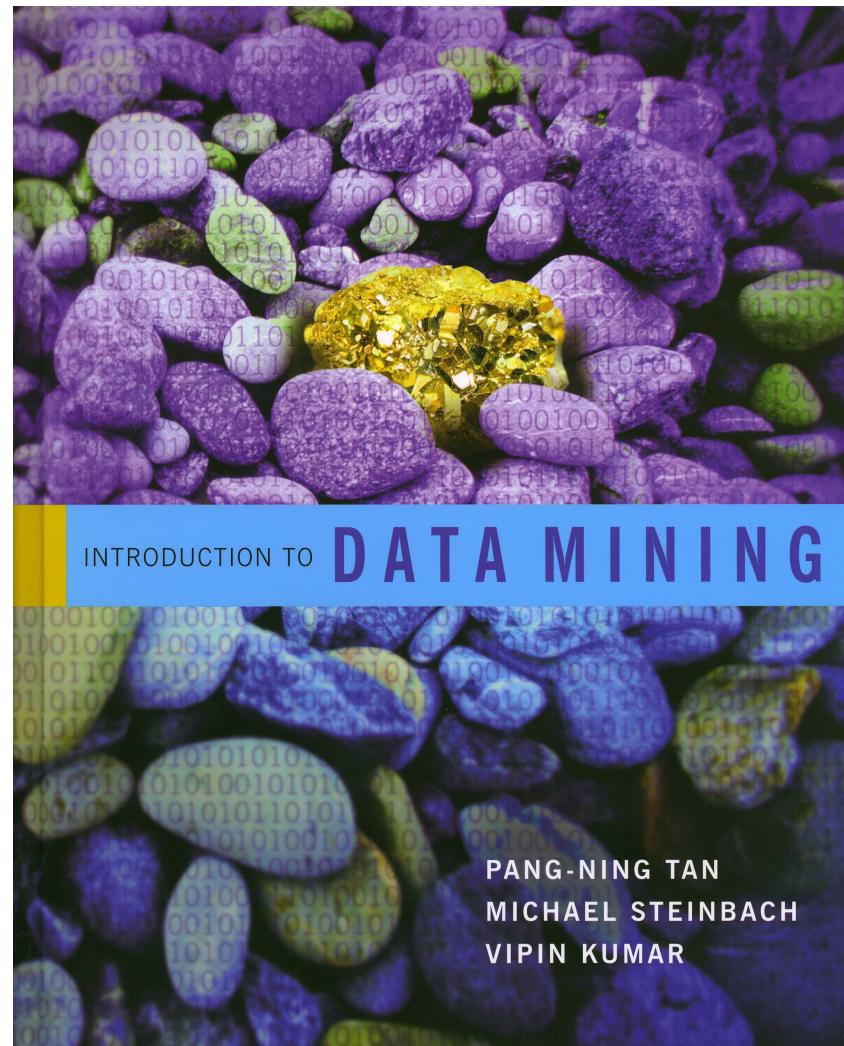
Reading material

Tan, Steinbach and Kumar
"Introduction to Data Mining"

Section 9.2.2 + 6.1-6.3

Groups of the day

Anders Bech Bruntse
Kristian Jensen
Patrick Gadd
Lenn Bloch
Hong Hao
Camille Fisichella
Aske Kargaard Scharling
Istvan Szonyi
Umaer Rashid Hanif
Caspar Aleksander Bang Jespersen
Martin Nørgaard



Lecture schedule

1. Introduction
(Tan 1.1-1.4)

Data: Feature extraction and visualization

2. Data and feature extraction
(Tan 2.1-2.3 + B1 (+ A))
3. Measures of similarity and summary statistics
(Tan 2.4 + 3.1-3.2 + C1-C2)
4. Data visualization
(Tan 3.3)

Supervised learning: Classification and regression

5. Decision trees and linear regression
(Tan 4.1-4.3 + D)
6. Overfitting and performance evaluation
(Tan 4.4-4.6)
7. Nearest neighbor, naive Bayes, and artificial neural networks
(Tan 5.2-5.4)

8. Ensemble methods and multi class classifiers
(Tan 5.6-5.8)

Unsupervised learning: Clustering and density est.

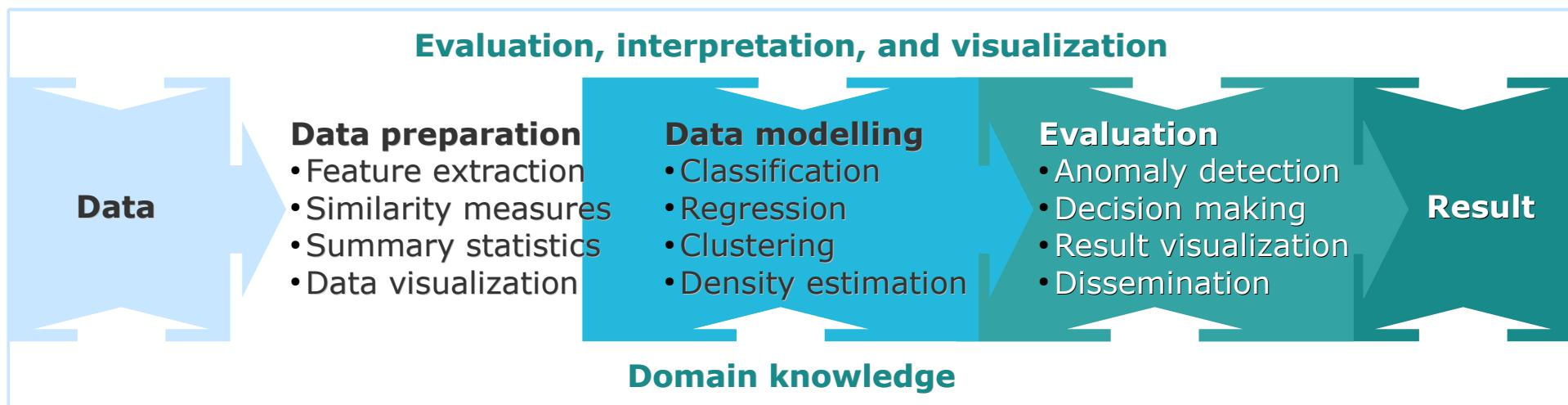
9. K-means and hierarchical clustering
(Tan 8.1-8.3+8.5.7)
10. **Mixture models and association mining**
(Tan 9.2.2 + 6.1-6.3)

11. Density estimation and anomaly detection
(Tan 10.1-10.4)

Machine learning and data modelling in practice

12. Putting it all together: Summary and overview
13. Mini project

Data modeling framework



After today you should be able to:

Explain the role of the parameters in the Gaussian Mixture Model (GMM)
and how the parameters are updated using the EM-algorithm

Explain why cross-validation can be used for GMM

Describe the Apriori principle in association mining and explain how this can be used for efficient estimation of association rules.

Calculate support and confidence of association rules

Imagine (again) you observe the world for the first time!



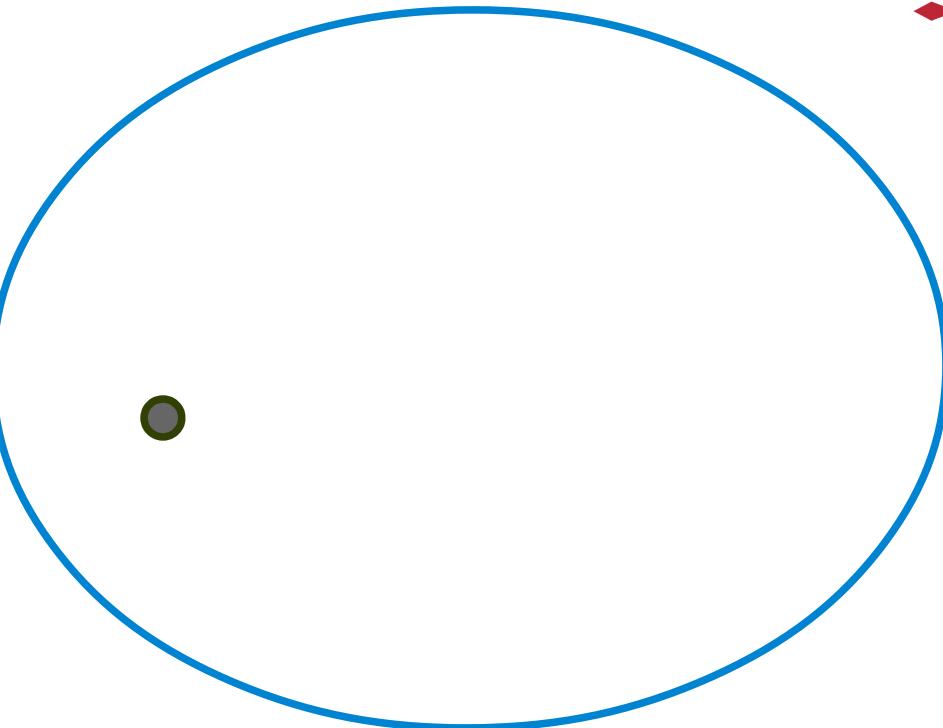
<http://www.clipartlord.com/category/baby-clip-art/>

http://commons.wikimedia.org/wiki/File:Abessinier_sorrel.jpg
http://commons.wikimedia.org/wiki/File:Cat_Eyes.jpg
http://commons.wikimedia.org/wiki/File:Black_white_cat_on_fence.jpg
<http://commons.wikimedia.org/wiki/File:Cat002.jpg>
<http://commons.wikimedia.org/wiki/File:Saurier2.jpg>
http://commons.wikimedia.org/wiki/File:Golden_Retriever_Dukedestiny01.jpg
<http://commons.wikimedia.org/wiki/File:Maspiri-Astro-SVE.jpg>
http://commons.wikimedia.org/wiki/File:GroC3%9Fer_Schweizer_Sennenhund.jpg
http://commons.wikimedia.org/wiki/File:GermanShorthPtr_wb.jpg

Imagine (again) you observe the world for the first time!



<http://www.clipartlord.com/category/baby-clip-art/>

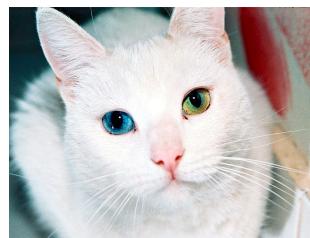
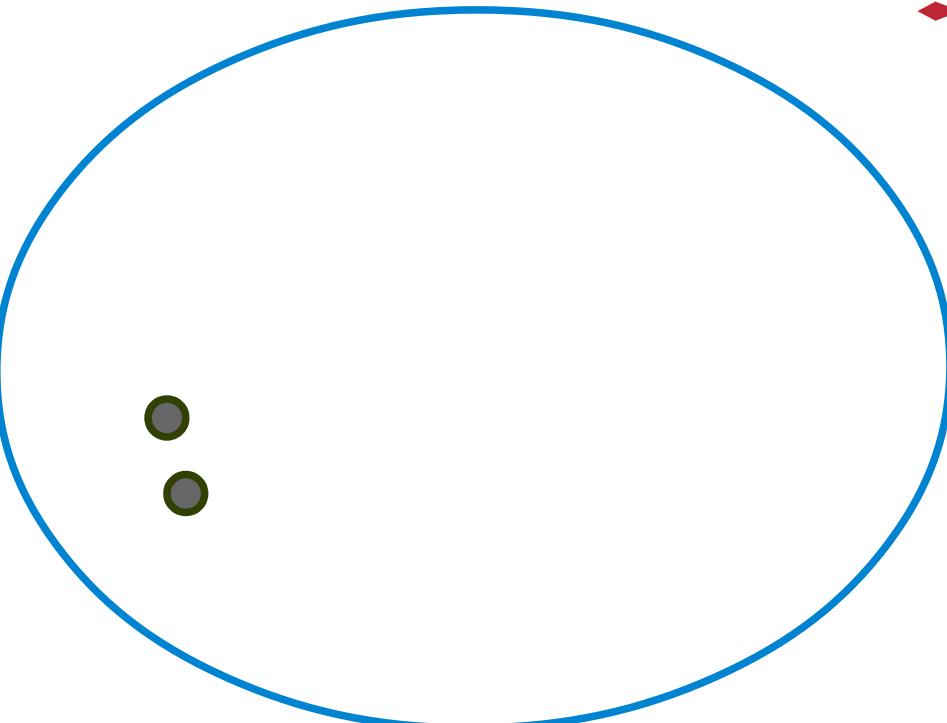


http://commons.wikimedia.org/wiki/File:Abessinier_sorrel.jpg
http://commons.wikimedia.org/wiki/File:Cat_Eyes.jpg
http://commons.wikimedia.org/wiki/File:Black_white_cat_on_fence.jpg
<http://commons.wikimedia.org/wiki/File:Cat002.jpg>
<http://commons.wikimedia.org/wiki/File:Saurier2.jpg>
http://commons.wikimedia.org/wiki/File:Golden_Retriever_Dukedestiny01.jpg
<http://commons.wikimedia.org/wiki/File:Maspiri-Astro-SVE.jpg>
http://commons.wikimedia.org/wiki/File:GroC3%9Fer_Schweizer_Sennenhund.jpg
http://commons.wikimedia.org/wiki/File:GermanShorthPtr_wb.jpg

Imagine (again) you observe the world for the first time!



<http://www.clipartlord.com/category/baby-clip-art/>

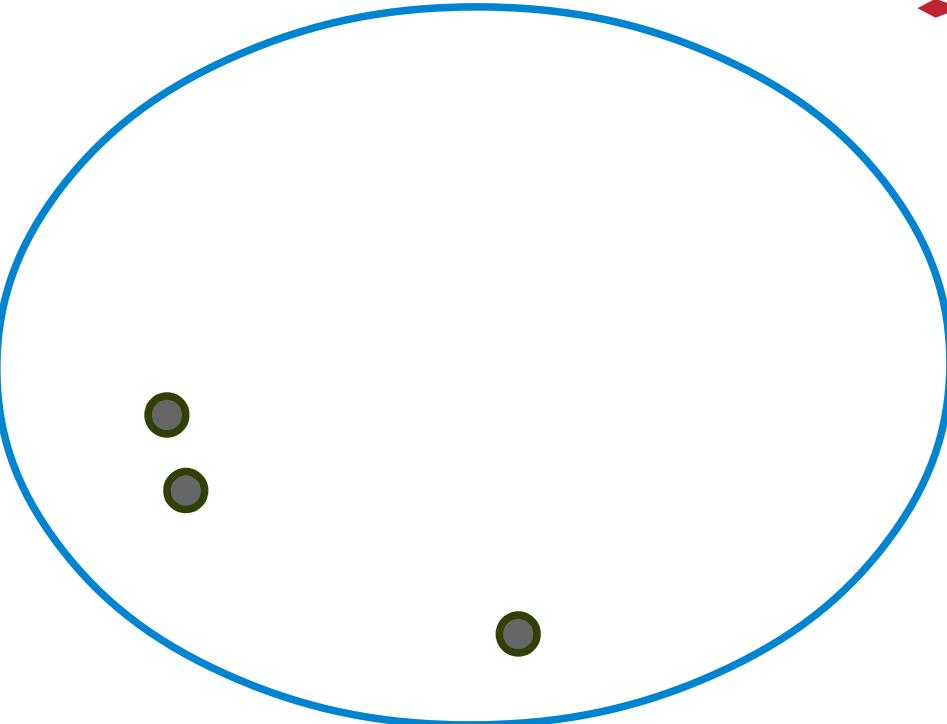


http://commons.wikimedia.org/wiki/File:Abessinier_sorrel.jpg
http://commons.wikimedia.org/wiki/File:Cat_Eyes.jpg
http://commons.wikimedia.org/wiki/File:Black_white_cat_on_fence.jpg
<http://commons.wikimedia.org/wiki/File:Cat002.jpg>
<http://commons.wikimedia.org/wiki/File:Saurier2.jpg>
http://commons.wikimedia.org/wiki/File:Golden_Retriever_Dukedestiny01.jpg
<http://commons.wikimedia.org/wiki/File:Maspiri-Astro-SVE.jpg>
http://commons.wikimedia.org/wiki/File:GroC3%9Fer_Schweizer_Sennenhund.jpg
http://commons.wikimedia.org/wiki/File:GermanShorthPtr_wb.jpg

Imagine (again) you observe the world for the first time!



<http://www.clipartlord.com/category/baby-clip-art/>

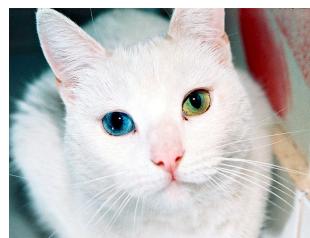
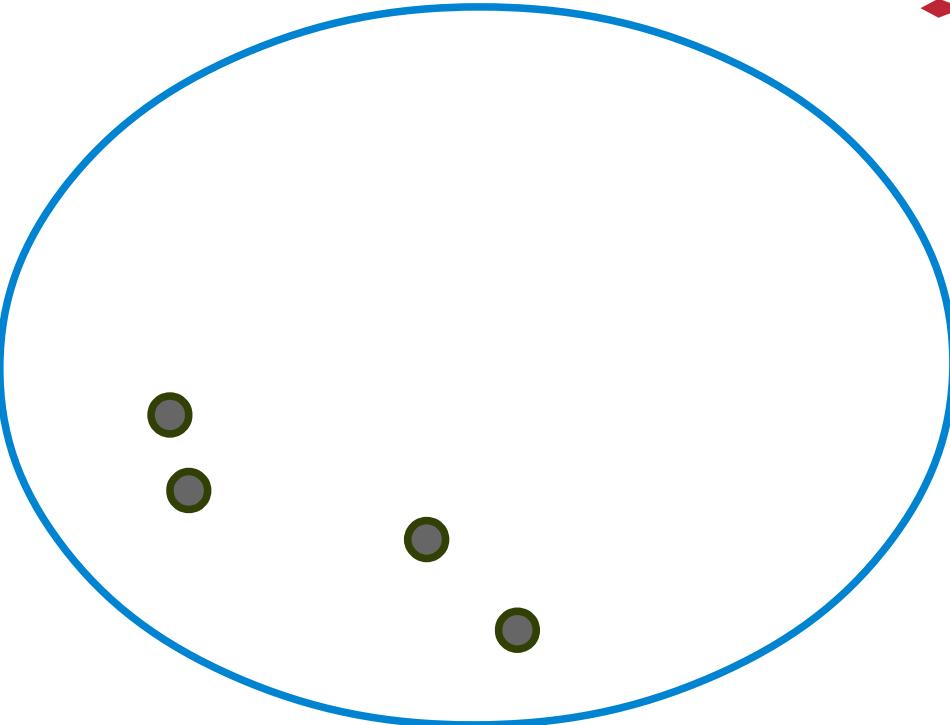


http://commons.wikimedia.org/wiki/File:Abessinier_sorrel.jpg
http://commons.wikimedia.org/wiki/File:Cat_Eyes.jpg
http://commons.wikimedia.org/wiki/File:Black_white_cat_on_fence.jpg
<http://commons.wikimedia.org/wiki/File:Cat002.jpg>
<http://commons.wikimedia.org/wiki/File:Saurier2.jpg>
http://commons.wikimedia.org/wiki/File:Golden_Retriever_Dukedestiny01.jpg
<http://commons.wikimedia.org/wiki/File:Maspiri-Astro-SVE.jpg>
http://commons.wikimedia.org/wiki/File:GroC3%9Fer_Schweizer_Sennenhund.jpg
http://commons.wikimedia.org/wiki/File:GermanShorthPtr_wb.jpg

Imagine (again) you observe the world for the first time!



<http://www.clipartlord.com/category/baby-clip-art/>

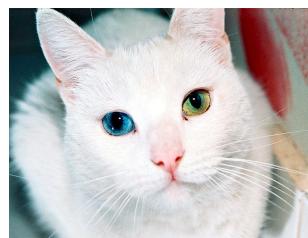
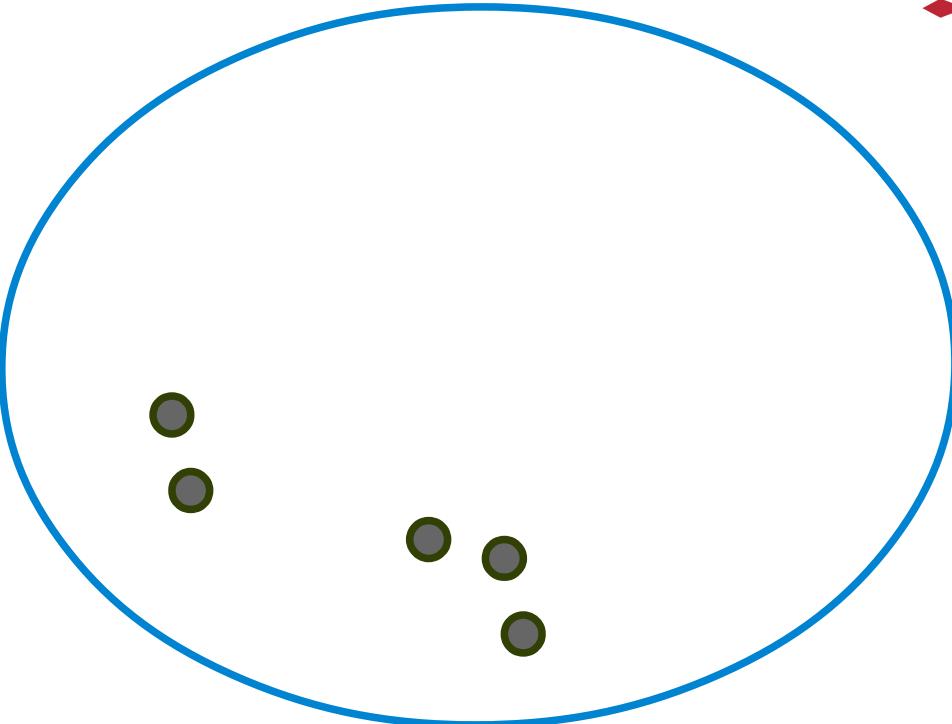


http://commons.wikimedia.org/wiki/File:Abessinier_sorrel.jpg
http://commons.wikimedia.org/wiki/File:Cat_Eyes.jpg
http://commons.wikimedia.org/wiki/File:Black_white_cat_on_fence.jpg
<http://commons.wikimedia.org/wiki/File:Cat002.jpg>
<http://commons.wikimedia.org/wiki/File:Saurier2.jpg>
http://commons.wikimedia.org/wiki/File:Golden_Retriever_Dukedestiny01.jpg
<http://commons.wikimedia.org/wiki/File:Maspiri-Astro-SVE.jpg>
http://commons.wikimedia.org/wiki/File:GroC3%9Fer_Schweizer_Sennenhund.jpg
http://commons.wikimedia.org/wiki/File:GermanShorthPtr_wb.jpg

Imagine (again) you observe the world for the first time!



<http://www.clipartlord.com/category/baby-clip-art/>

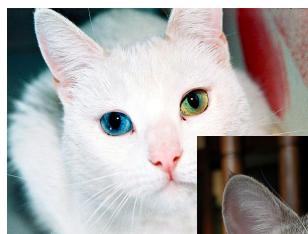
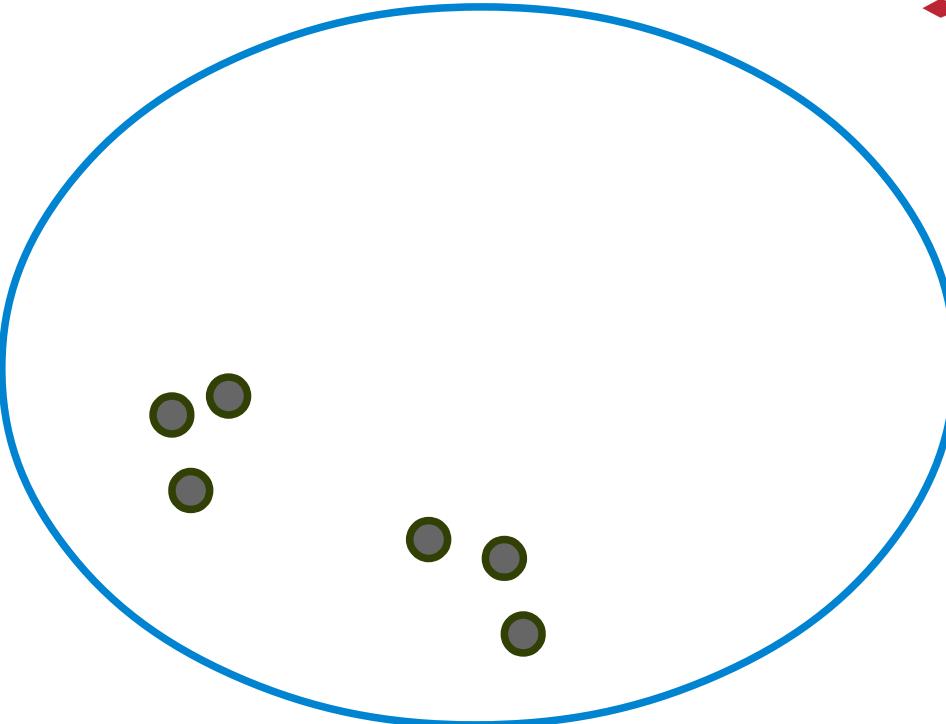


http://commons.wikimedia.org/wiki/File:Abessinier_sorrel.jpg
http://commons.wikimedia.org/wiki/File:Cat_Eyes.jpg
http://commons.wikimedia.org/wiki/File:Black_white_cat_on_fence.jpg
<http://commons.wikimedia.org/wiki/File:Cat002.jpg>
<http://commons.wikimedia.org/wiki/File:Saurier2.jpg>
http://commons.wikimedia.org/wiki/File:Golden_Retriever_Dukedestiny01.jpg
<http://commons.wikimedia.org/wiki/File:Maspiri-Astro-SVE.jpg>
http://commons.wikimedia.org/wiki/File:GroC3%9Fer_Schweizer_Sennenhund.jpg
http://commons.wikimedia.org/wiki/File:GermanShorthPtr_wb.jpg

Imagine (again) you observe the world for the first time!



<http://www.clipartlord.com/category/baby-clip-art/>

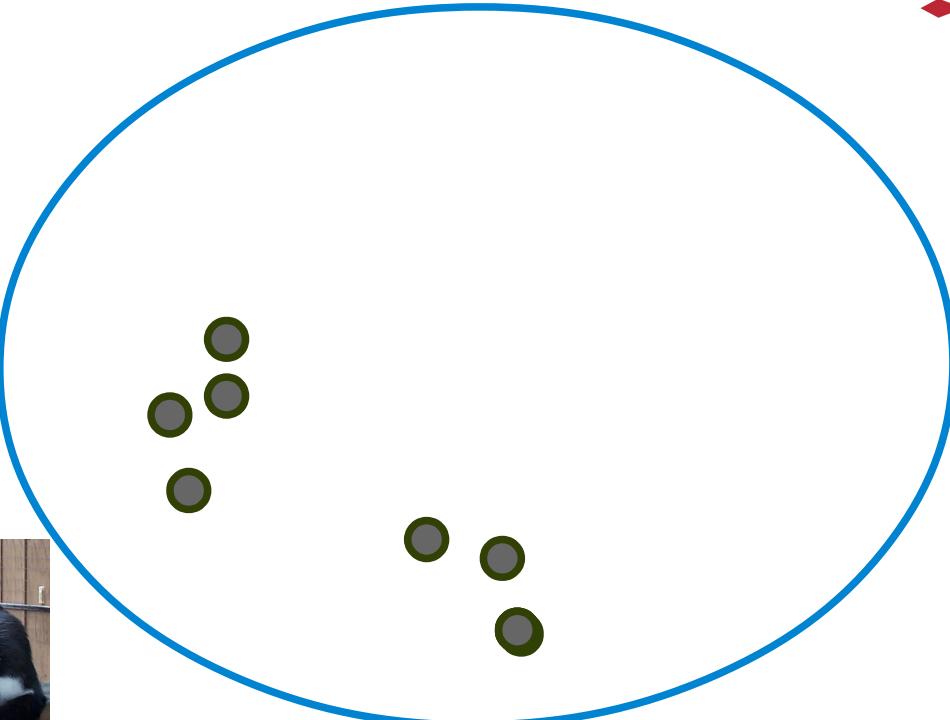
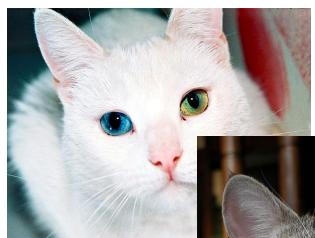


http://commons.wikimedia.org/wiki/File:Abessinier_sorrel.jpg
http://commons.wikimedia.org/wiki/File:Cat_Eyes.jpg
http://commons.wikimedia.org/wiki/File:Black_white_cat_on_fence.jpg
<http://commons.wikimedia.org/wiki/File:Cat002.jpg>
<http://commons.wikimedia.org/wiki/File:Saurier2.jpg>
http://commons.wikimedia.org/wiki/File:Golden_Retriever_Dukedestiny01.jpg
<http://commons.wikimedia.org/wiki/File:Maspiri-Astro-SVE.jpg>
http://commons.wikimedia.org/wiki/File:GroC3%9Fer_Schweizer_Sennenhund.jpg
http://commons.wikimedia.org/wiki/File:GermanShorthPtr_wb.jpg

Imagine (again) you observe the world for the first time!



<http://www.clipartlord.com/category/baby-clip-art/>

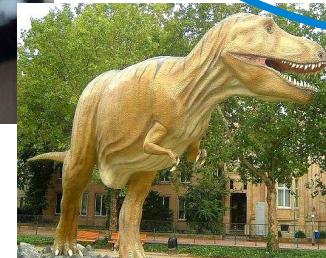
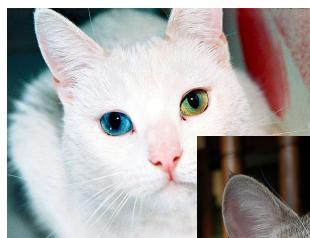


http://commons.wikimedia.org/wiki/File:Abessinier_sorrel.jpg
http://commons.wikimedia.org/wiki/File:Cat_Eyes.jpg
http://commons.wikimedia.org/wiki/File:Black_white_cat_on_fence.jpg
<http://commons.wikimedia.org/wiki/File:Cat002.jpg>
<http://commons.wikimedia.org/wiki/File:Saurier2.jpg>
http://commons.wikimedia.org/wiki/File:Golden_Retriever_Dukedestiny01.jpg
<http://commons.wikimedia.org/wiki/File:Maspiri-Astro-SVE.jpg>
http://commons.wikimedia.org/wiki/File:GroC3%9Fer_Schweizer_Sennenhund.jpg
http://commons.wikimedia.org/wiki/File:GermanShorthPtr_wb.jpg

Imagine (again) you observe the world for the first time!



<http://www.clipartlord.com/category/baby-clip-art/>

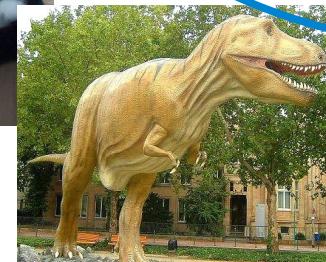


http://commons.wikimedia.org/wiki/File:Abessinier_sorrel.jpg
http://commons.wikimedia.org/wiki/File:Cat_Eyes.jpg
http://commons.wikimedia.org/wiki/File:Black_white_cat_on_fence.jpg
<http://commons.wikimedia.org/wiki/File:Cat002.jpg>
<http://commons.wikimedia.org/wiki/File:Saurier2.jpg>
http://commons.wikimedia.org/wiki/File:Golden_Retriever_Dukedestiny01.jpg
<http://commons.wikimedia.org/wiki/File:Maspiri-Astro-SVE.jpg>
http://commons.wikimedia.org/wiki/File:GroC3%9Fer_Schweizer_Sennenhund.jpg
http://commons.wikimedia.org/wiki/File:GermanShorthPtr_wb.jpg

Imagine (again) you observe the world for the first time!



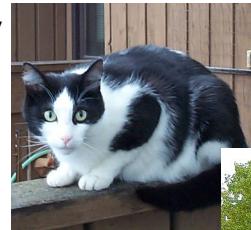
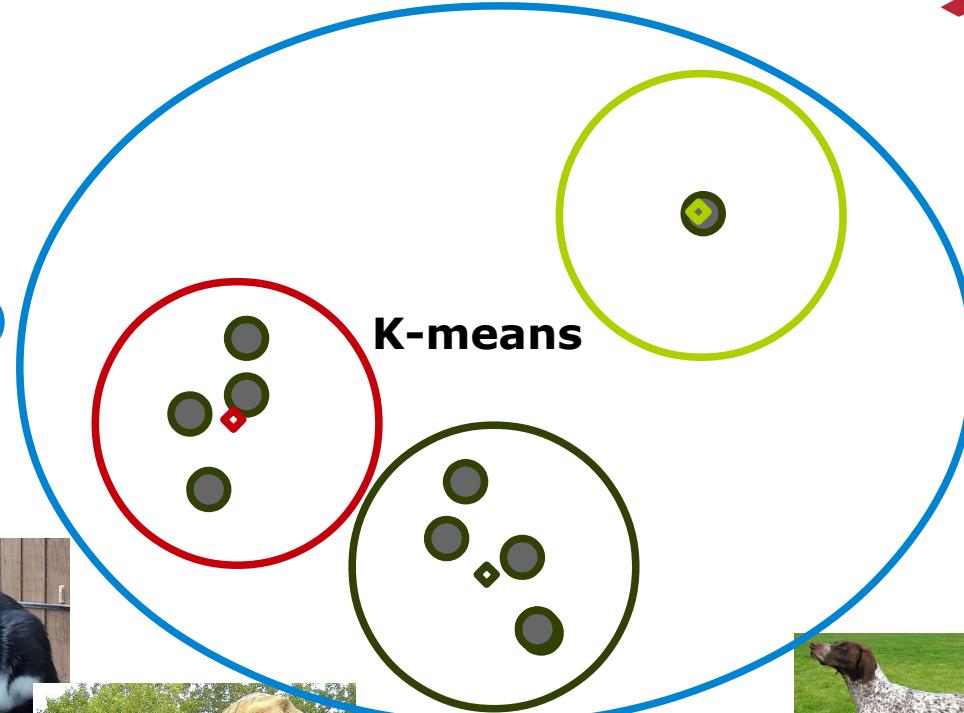
<http://www.clipartlord.com/category/baby-clip-art/>



We humans are skilled at dividing objects into groups (clustering), but how do we make computers do the same?

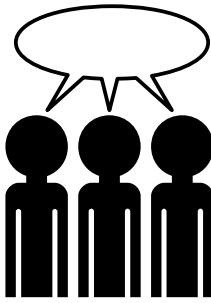
http://commons.wikimedia.org/wiki/File:Abessinier_sorrel.jpg
http://commons.wikimedia.org/wiki/File:Cat_Eyes.jpg
http://commons.wikimedia.org/wiki/File:Black_white_cat_on_fence.jpg
<http://commons.wikimedia.org/wiki/File:Cat002.jpg>
<http://commons.wikimedia.org/wiki/File:Saurier2.jpg>
http://commons.wikimedia.org/wiki/File:Golden_Retriever_Dukedestiny01.jpg
<http://commons.wikimedia.org/wiki/File:Maspiri-Astro-SVE.jpg>
http://commons.wikimedia.org/wiki/File:GroC3%9Fer_Schweizer_Sennenhund.jpg
http://commons.wikimedia.org/wiki/File:GermanShorthPtr_wb.jpg

Imagine (again) you observe the world for the first time!



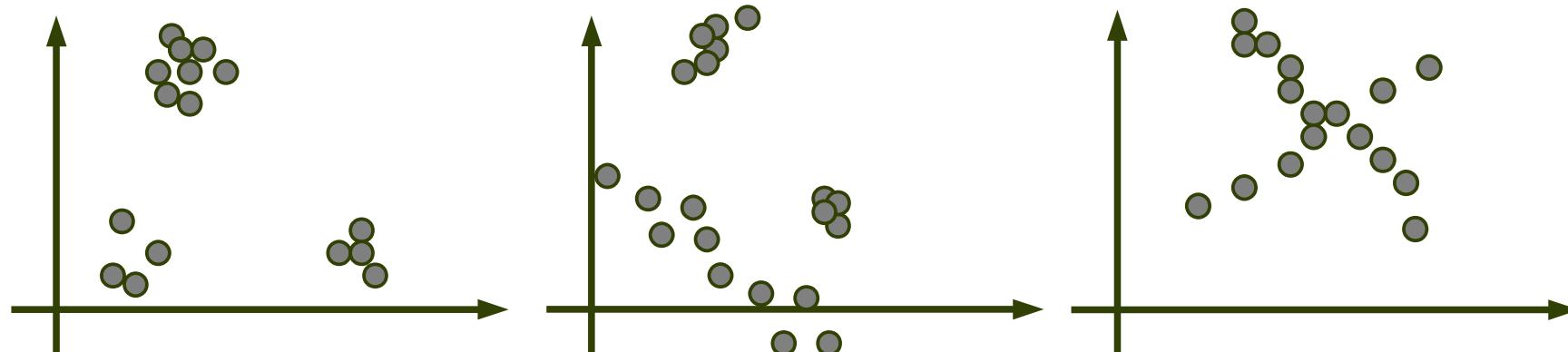
We humans are skilled at dividing objects into groups (clustering), but how do we make computers do the same?

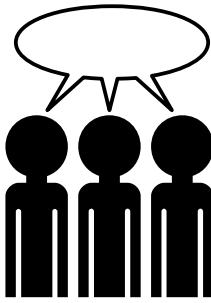
http://commons.wikimedia.org/wiki/File:Abessinier_sorrel.jpg
http://commons.wikimedia.org/wiki/File:Cat_Eyes.jpg
http://commons.wikimedia.org/wiki/File:Black_white_cat_on_fence.jpg
<http://commons.wikimedia.org/wiki/File:Cat002.jpg>
<http://commons.wikimedia.org/wiki/File:Saurier2.jpg>
http://commons.wikimedia.org/wiki/File:Golden_Retriever_Dukedestiny01.jpg
<http://commons.wikimedia.org/wiki/File:Maspiri-Astro-SVE.jpg>
http://commons.wikimedia.org/wiki/File:GroC3%9Fer_Schweizer_Sennenhund.jpg
http://commons.wikimedia.org/wiki/File:GermanShorthPtr_wb.jpg



Group exercise

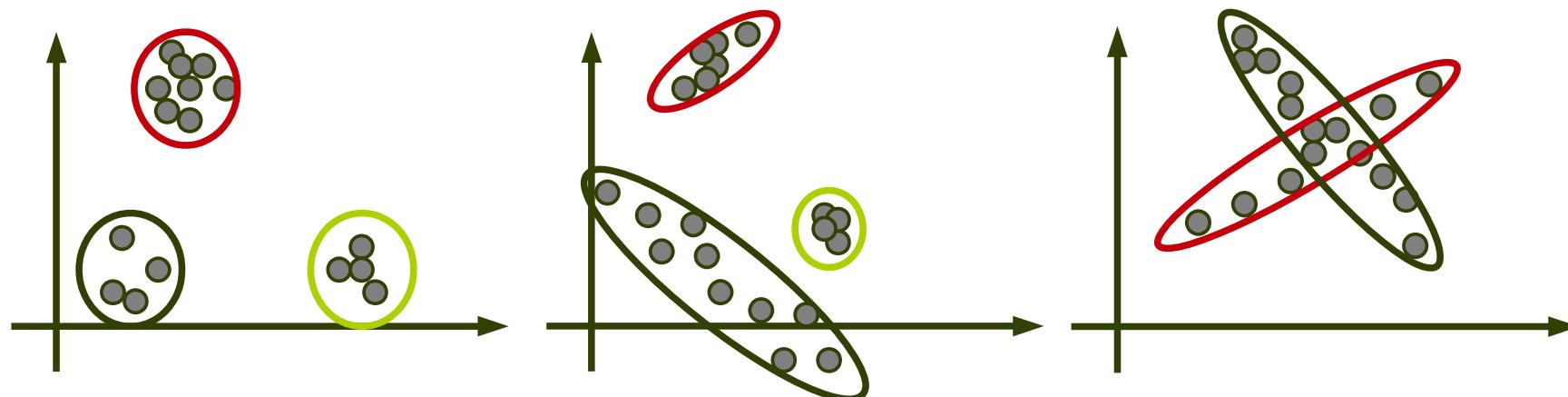
- What are the clusters below and what characterize each cluster?
- Is **k-means** well suited for modeling the clusters below?
 - Will it always find the optimum solution?
 - Can it model the sizes of the clusters?
 - Can it model the shape of the clusters?
 - How can we determine the number of clusters?





Group exercise

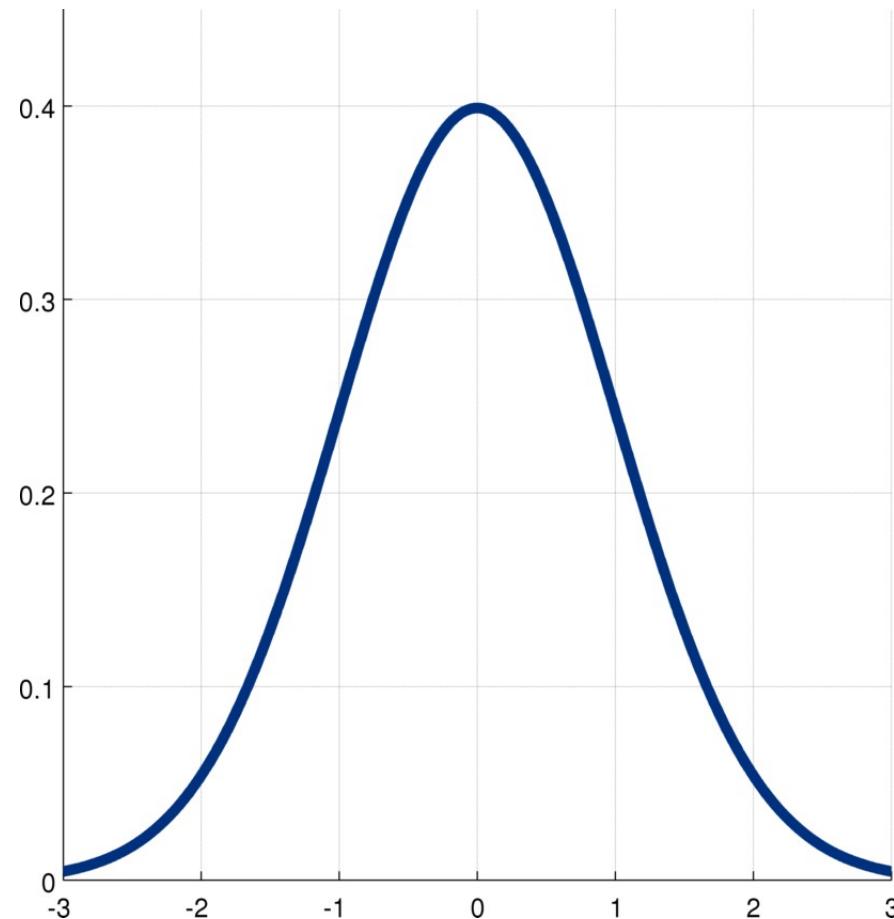
- What are the clusters below and what characterize each cluster?
- Is **k-means** well suited for modeling the clusters below?
 - Will it always find the optimum solution?
 - Can it model the sizes of the clusters?
 - Can it model the shape of the clusters?
 - How can we determine the number of clusters?



Normal distribution

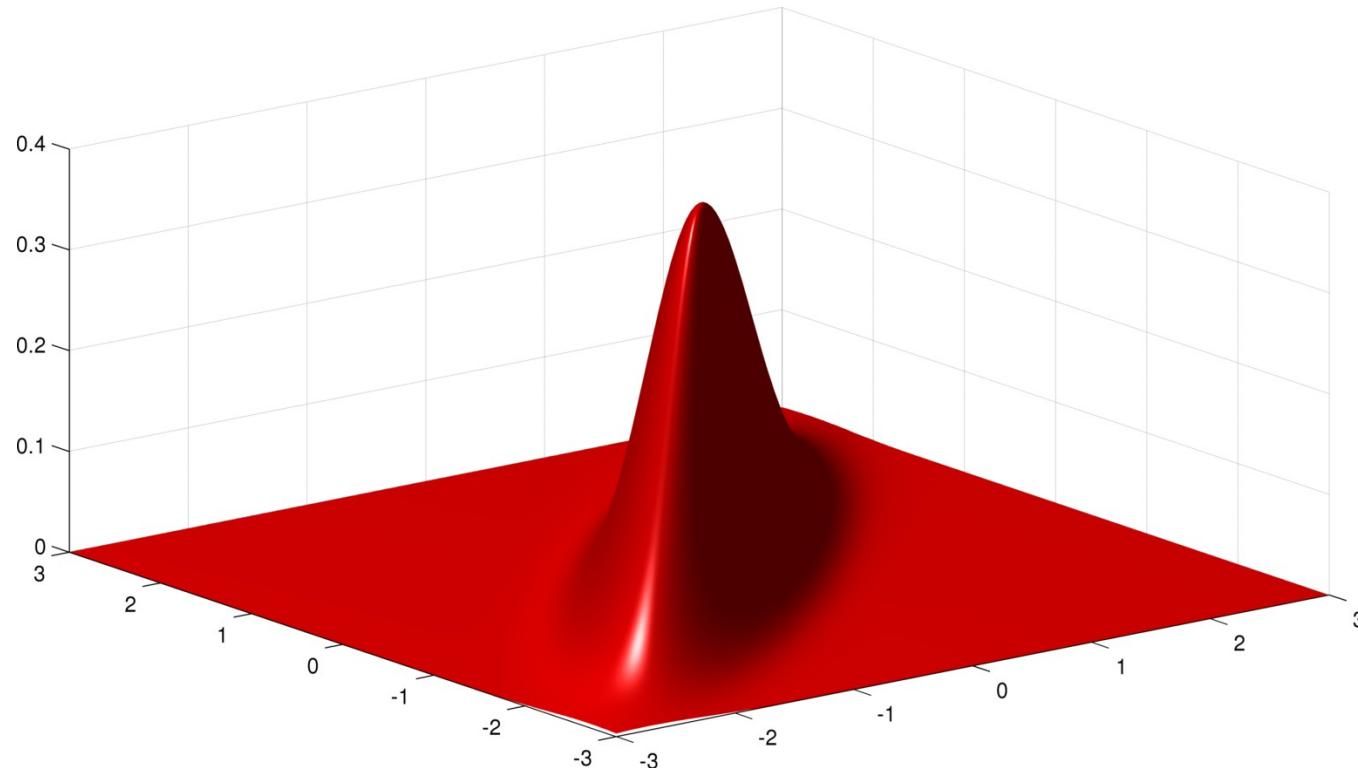
- Probability density function describes the relative chance of a given value to occur
- Normal distribution characterized by
 - Mean
 - Variance

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right)$$



Multivariate Normal distribution

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{k/2} |\Sigma|^{1/2}} \exp \left(-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right)$$



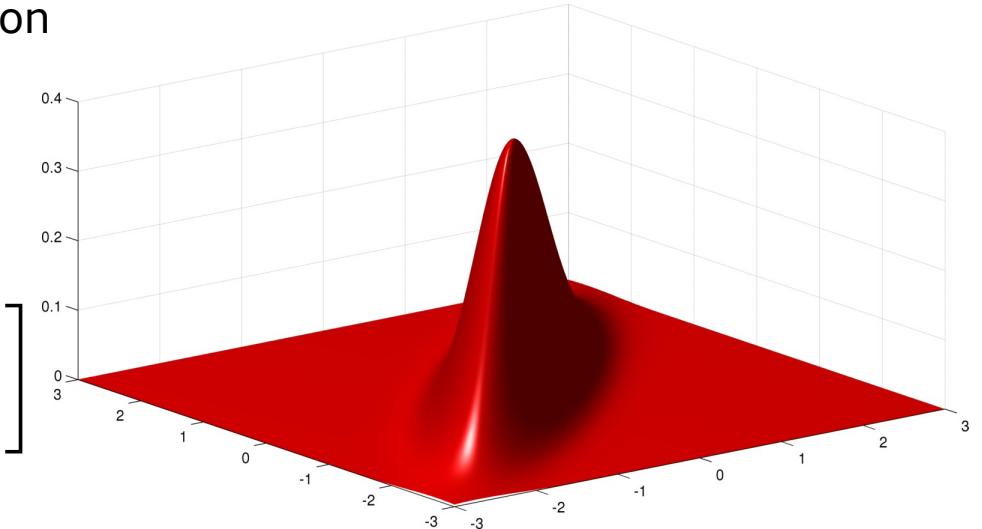
Multivariate Normal distribution

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{k/2} |\Sigma|^{1/2}} \exp \left(-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right)$$

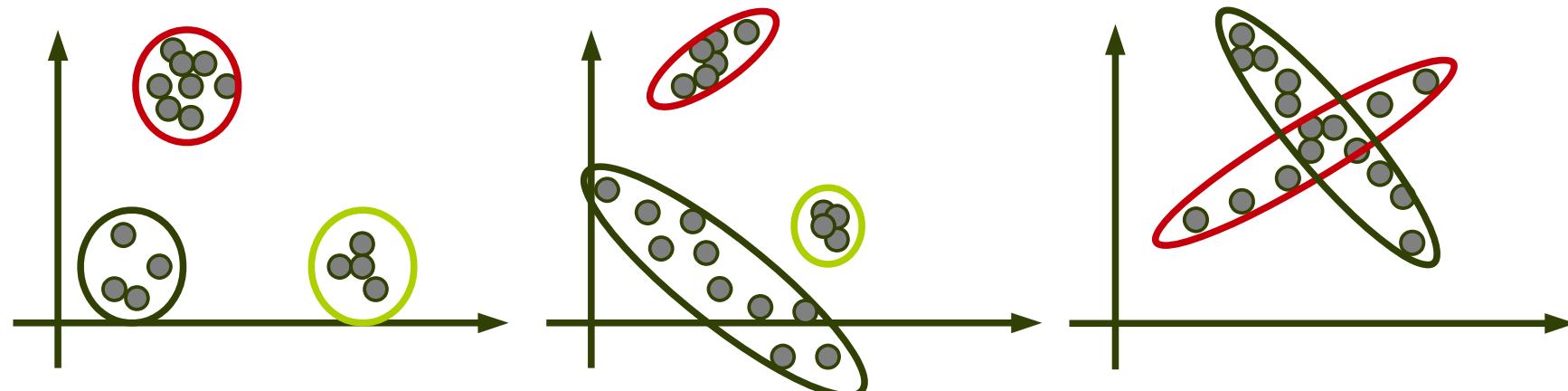
- Example: 2-dimensional Normal distribution

$$\boldsymbol{\mu} = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \end{bmatrix}$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} \text{var}(x_1) & \text{cov}(x_1, x_2) \\ \text{cov}(x_2, x_1) & \text{var}(x_2) \end{bmatrix}$$



The Gaussian Mixture Model (GMM)



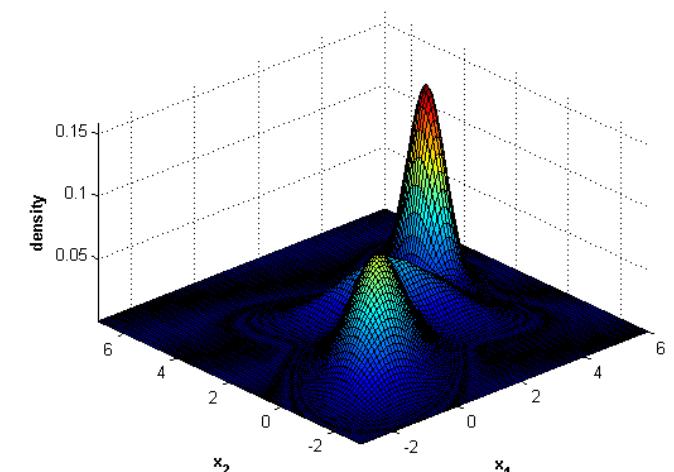
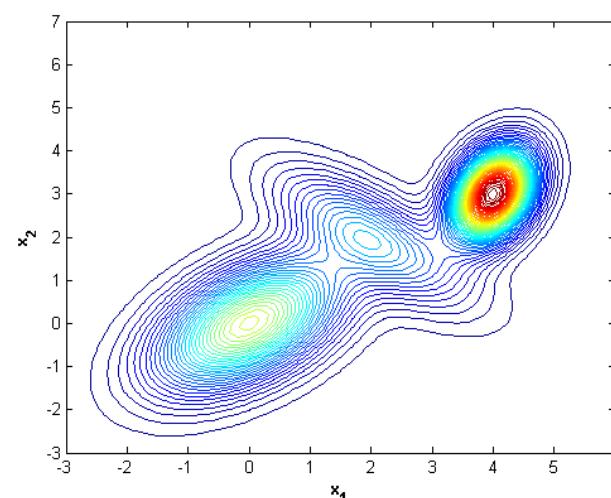
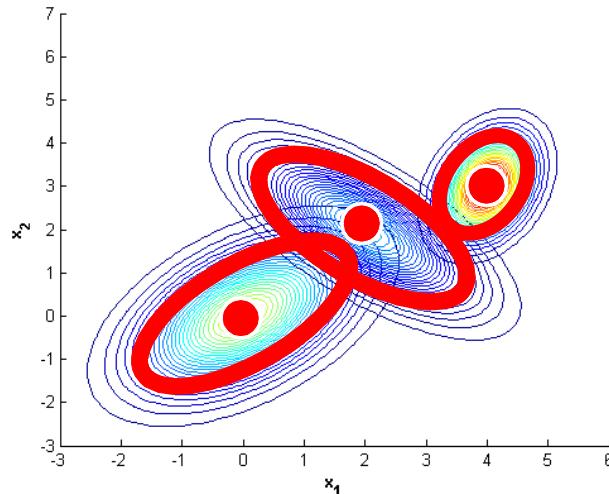
- Different locations
- Different shape
- Different sizes

Data density **Sum of cluster specific densities assumed normal distributed**

$$\mu_{(k)} \quad \downarrow \quad p(\mathbf{x}) = \sum_{k=1}^K w_k \mathcal{N}(\mathbf{x}, \mu_{(k)}, \Sigma_{(k)})$$
$$\Sigma_{(k)}$$
$$w_k$$
$$\text{s.t. } \sum_{k=1}^K w_k = 1, \quad w_k \geq 0$$

GMM example

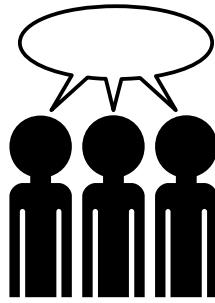
$$p(\mathbf{x}) = 0.5\mathcal{N}(\mathbf{x} | \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}) + 0.2\mathcal{N}(\mathbf{x} | \begin{bmatrix} 2 \\ 2 \end{bmatrix}, \begin{bmatrix} 1 & -0.7 \\ -0.7 & 1 \end{bmatrix}) + 0.3\mathcal{N}(\mathbf{x} | \begin{bmatrix} 4 \\ 3 \end{bmatrix}, \begin{bmatrix} 0.2 & 0.1 \\ 0.1 & 0.5 \end{bmatrix})$$



$\mu_{(k)}$: Cluster center (prototypical example in cluster)

$\Sigma_{(k)}$: Shape of the cluster

w_k : Relative size/density of the cluster



Group exercise

- Consider the Gaussian mixture model (GMM)

$$p(\mathbf{x}) = \sum_{k=1}^K w_k \mathcal{N}(\mathbf{x}, \boldsymbol{\mu}_{(k)}, \boldsymbol{\Sigma}_{(k)}) \quad \text{s.t. } \sum_{k=1}^K w_k = 1, \quad w_k \geq 0$$

- What is the value of the integral?

$$\int p(\mathbf{x}) d\mathbf{x}$$

Gaussian mixture models, EM algorithm

Select an initial set of model parameters
(mean and covariance for each cluster)

Repeat

- **Expectation**

- For each object, calculate the probability of belonging to each distribution

- **Maximization**

- For each probability distribution, estimate parameters by maximum likelihood

Until the parameters do not change

E-step

$$p(z_n = k | \mathbf{x}_n) = \frac{w_k \mathcal{N}(\mathbf{x} | \boldsymbol{\mu}_{(k)}, \boldsymbol{\Sigma}_{(k)})}{\sum_{K=1}^K w_k \mathcal{N}(\mathbf{x} | \boldsymbol{\mu}_{(k)}, \boldsymbol{\Sigma}_{(k)})}$$

M-step

$$N_k = \sum_{n=1}^N p(z_n = k | \mathbf{x}_n)$$

$$\boldsymbol{\mu}_{(k)} = \frac{1}{N_k} \sum_{n=1}^N \mathbf{x}_n p(z_n = k | \mathbf{x}_n)$$

$$w_k = \frac{N_k}{N}$$

$$\boldsymbol{\Sigma}_{(k)} = \frac{1}{N_k} \sum_{n=1}^N (\mathbf{x}_n - \boldsymbol{\mu}_{(k)}) (\mathbf{x}_n - \boldsymbol{\mu}_{(k)})^\top p(z_n = k | \mathbf{x}_n)$$

Gaussian mixture models, EM algorithm

Select an initial set of model parameters
(mean and covariance for each cluster)

Repeat

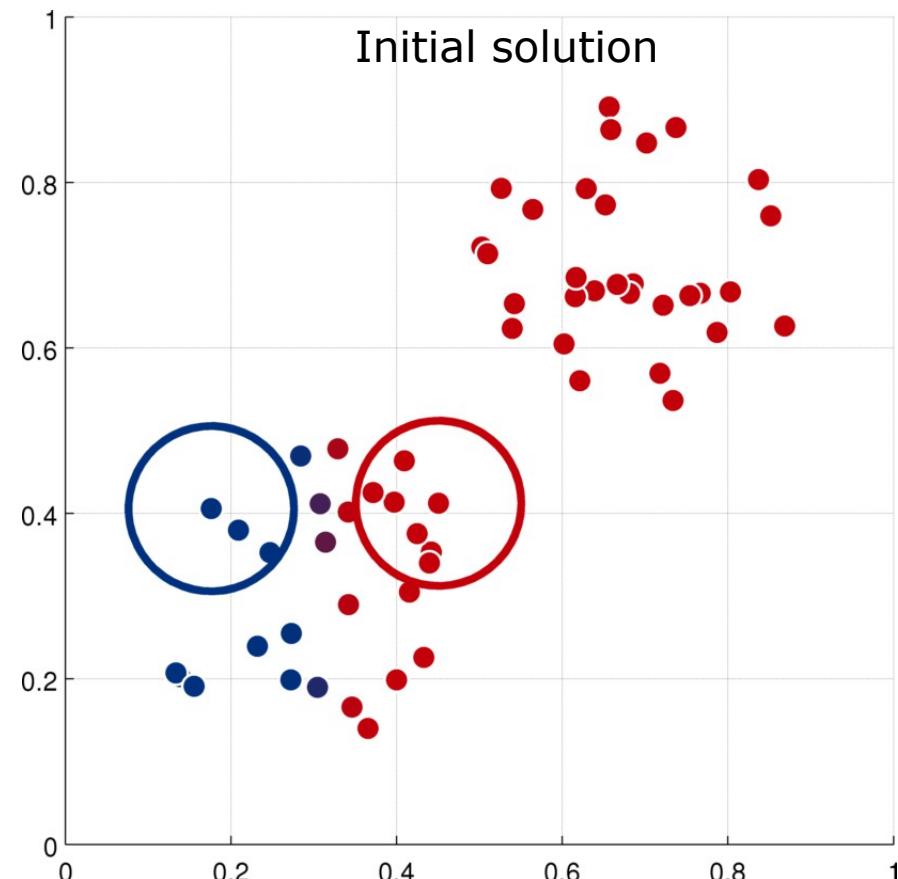
- **Expectation**

- For each object, calculate the probability of belonging to each distribution

- **Maximization**

- For each probability distribution, estimate parameters by maximum likelihood

Until the parameters do not change



Gaussian mixture models, EM algorithm

Select an initial set of model parameters
(mean and covariance for each cluster)

Repeat

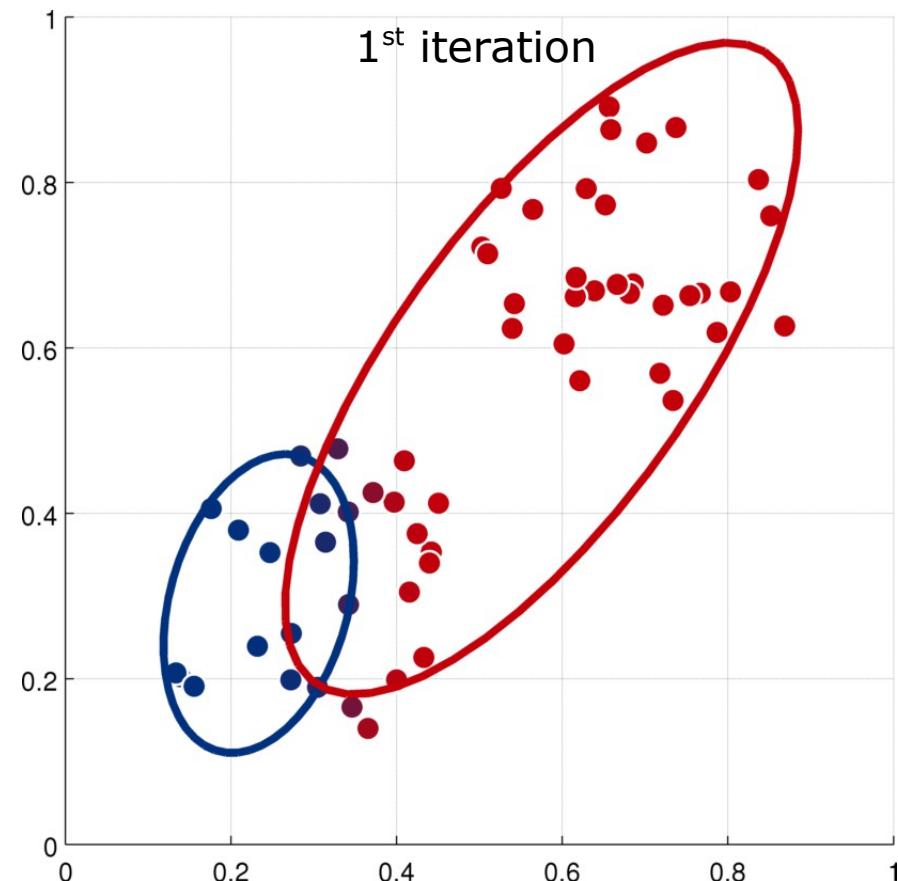
- **Expectation**

- For each object, calculate the probability of belonging to each distribution

- **Maximization**

- For each probability distribution, estimate parameters by maximum likelihood

Until the parameters do not change



Gaussian mixture models, EM algorithm

Select an initial set of model parameters
(mean and covariance for each cluster)

Repeat

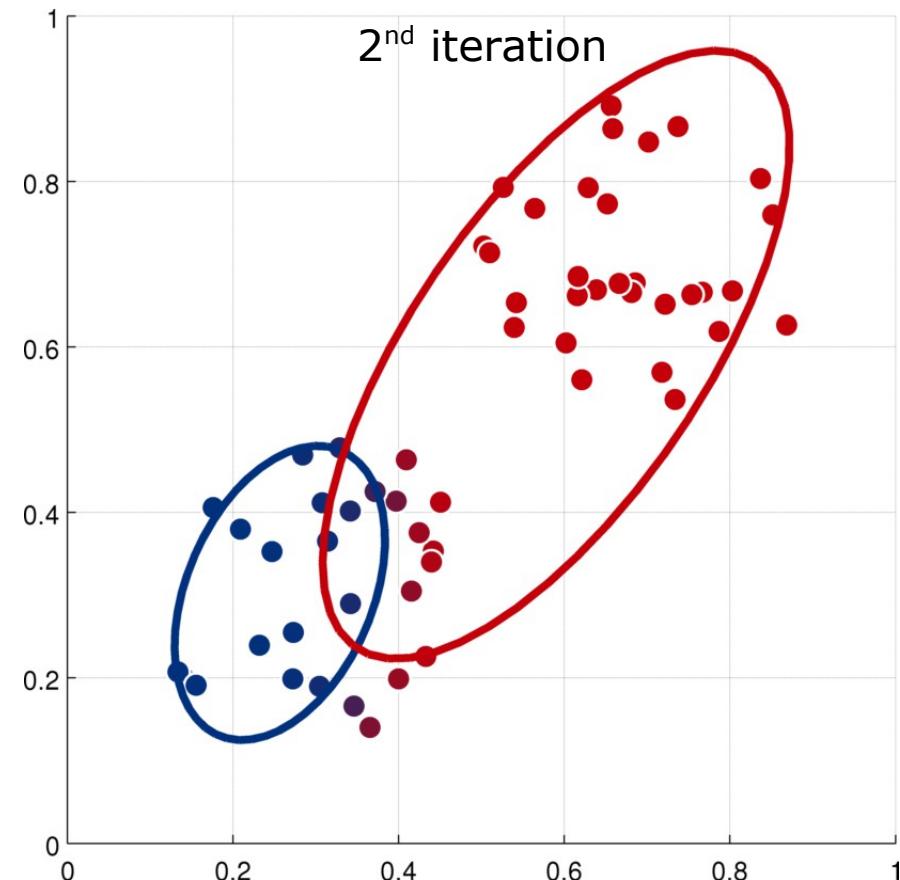
- **Expectation**

- For each object, calculate the probability of belonging to each distribution

- **Maximization**

- For each probability distribution, estimate parameters by maximum likelihood

Until the parameters do not change



Gaussian mixture models, EM algorithm

Select an initial set of model parameters
(mean and covariance for each cluster)

Repeat

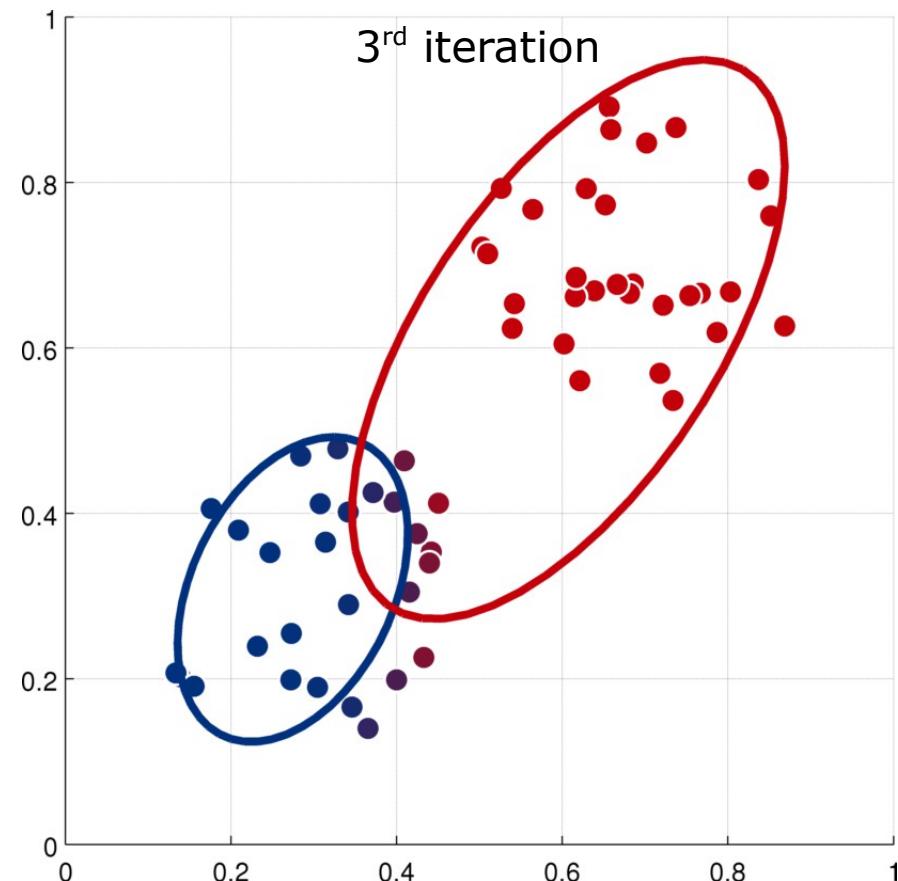
- **Expectation**

- For each object, calculate the probability of belonging to each distribution

- **Maximization**

- For each probability distribution, estimate parameters by maximum likelihood

Until the parameters do not change



Gaussian mixture models, EM algorithm

Select an initial set of model parameters
(mean and covariance for each cluster)

Repeat

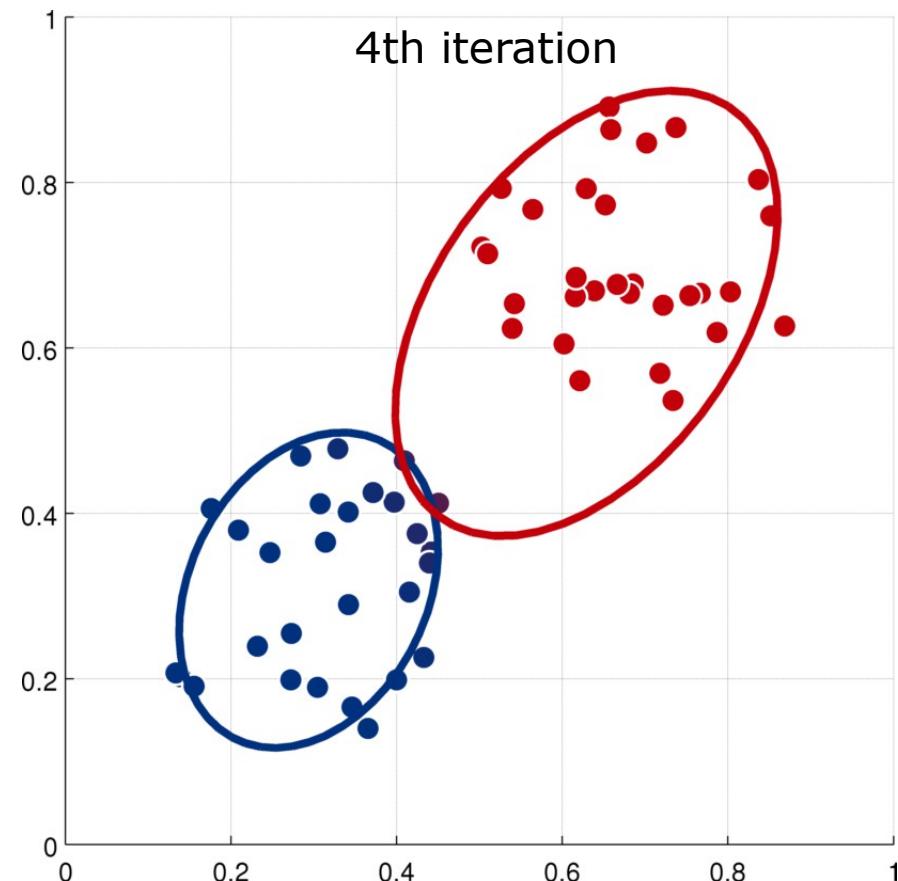
- **Expectation**

- For each object, calculate the probability of belonging to each distribution

- **Maximization**

- For each probability distribution, estimate parameters by maximum likelihood

Until the parameters do not change



Gaussian mixture models, EM algorithm

Select an initial set of model parameters
(mean and covariance for each cluster)

Repeat

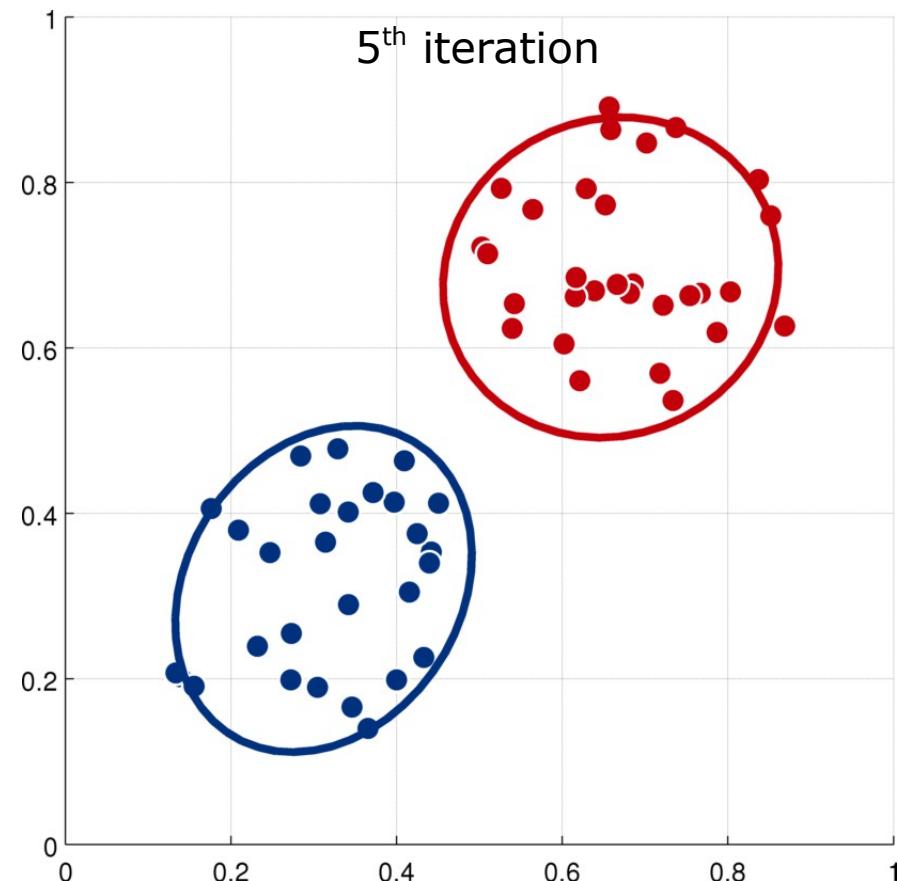
- **Expectation**

- For each object, calculate the probability of belonging to each distribution

- **Maximization**

- For each probability distribution, estimate parameters by maximum likelihood

Until the parameters do not change



Gaussian mixture models, EM algorithm

Select an initial set of model parameters
(mean and covariance for each cluster)

Repeat

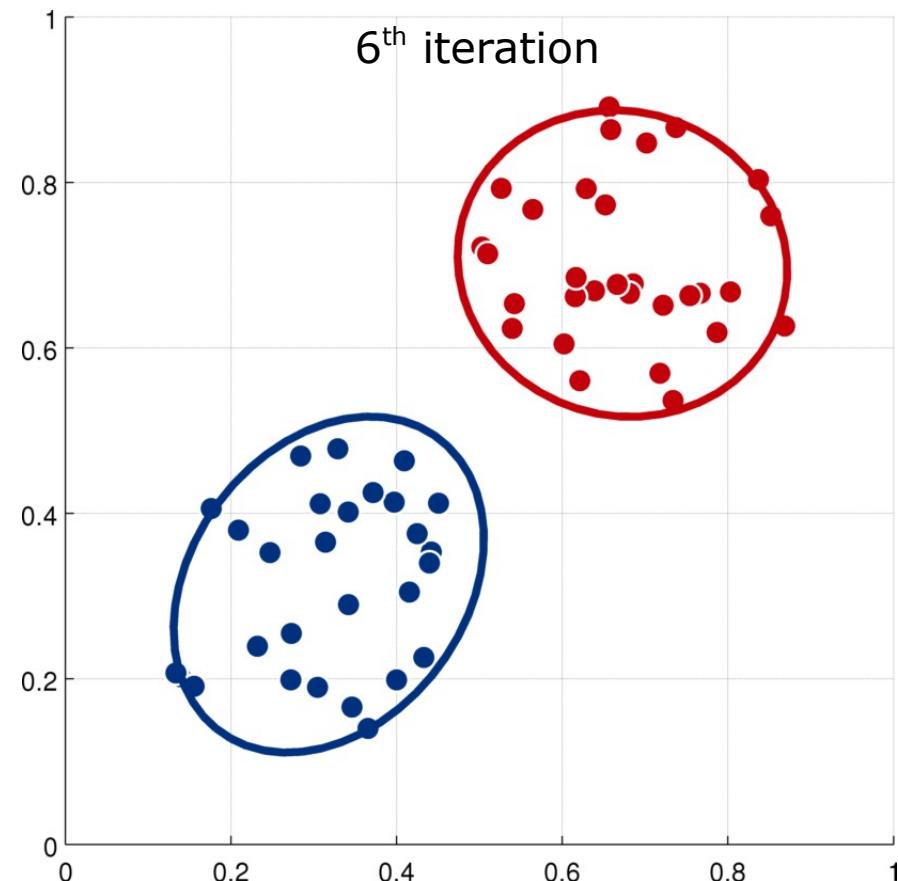
- **Expectation**

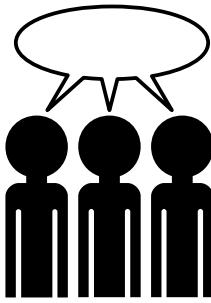
- For each object, calculate the probability of belonging to each distribution

- **Maximization**

- For each probability distribution, estimate parameters by maximum likelihood

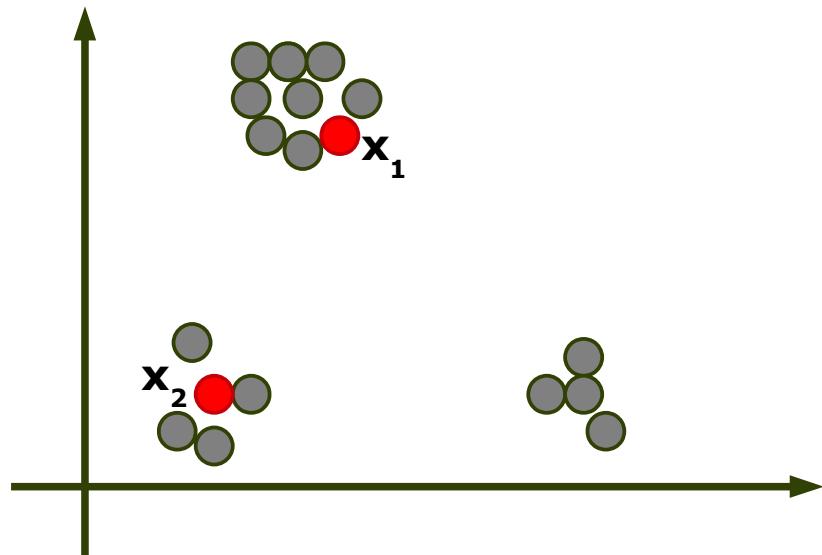
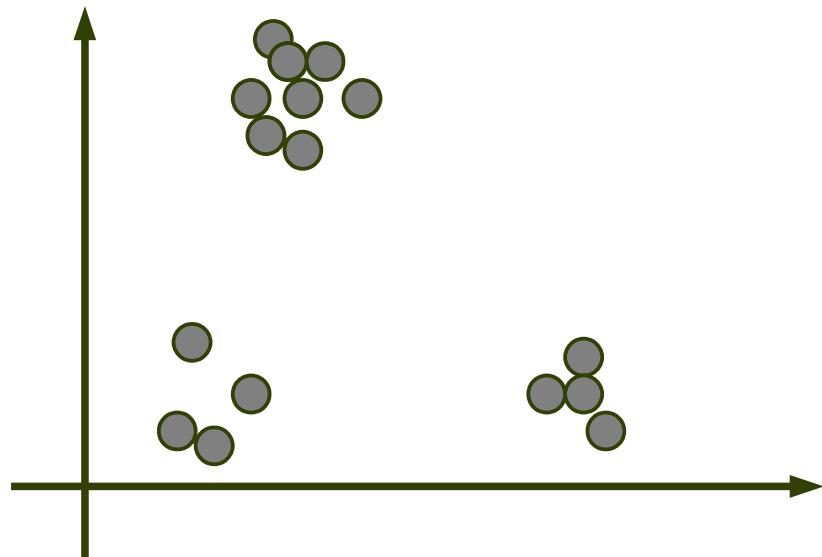
Until the parameters do not change





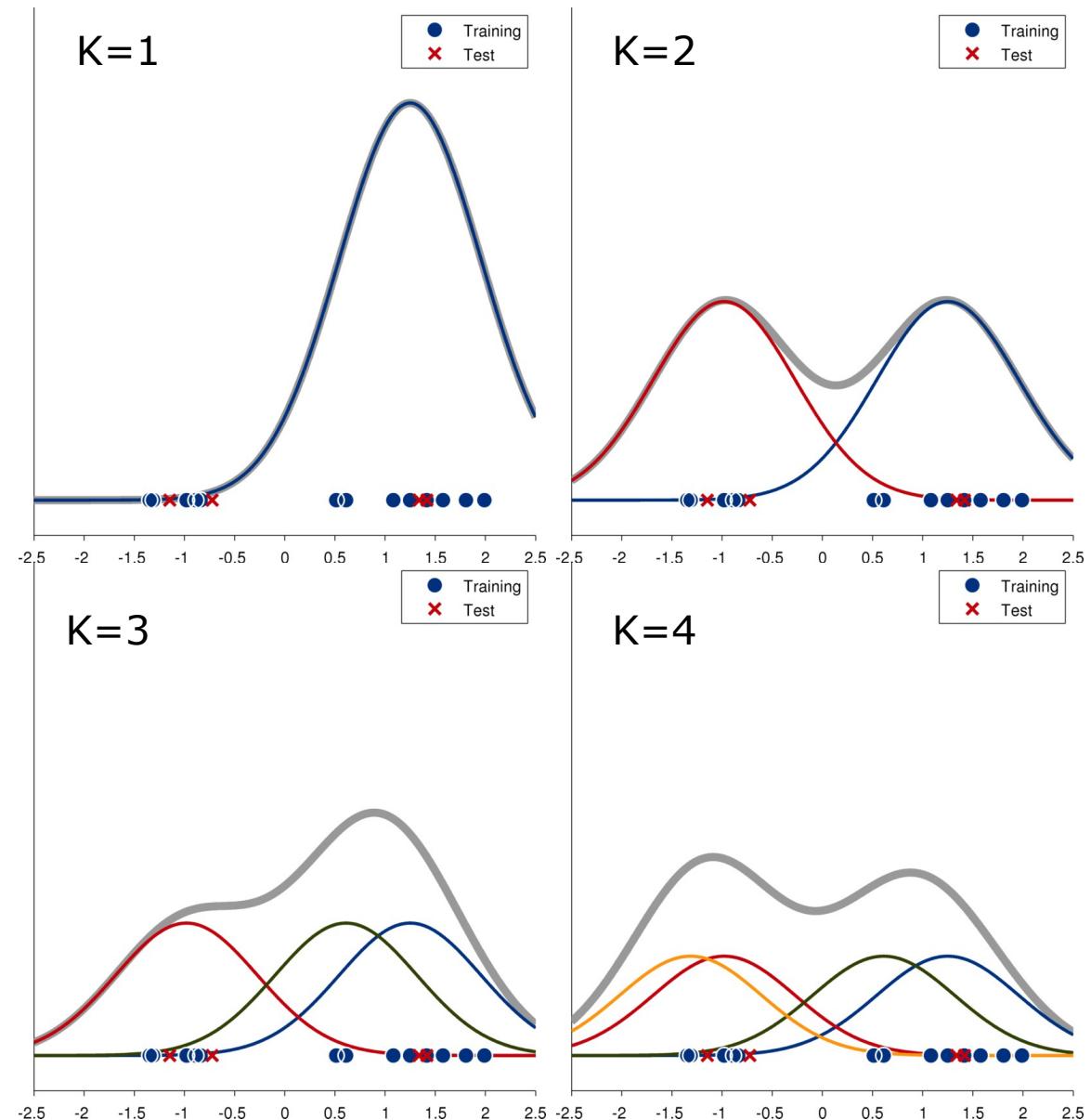
Group exercise

- Consider the data to the right with 16 observations.
 - What would ideally happen if we used a GMM with K=16 clusters to model the data?
- Imagine we have two **test observations** denoted \mathbf{x}_1 and \mathbf{x}_2 (red points) that are not used for training.
 - What happens to $p(\mathbf{x}_1)$ and $p(\mathbf{x}_2)$ if we use K=3 and K=16 clusters?



Mixture models

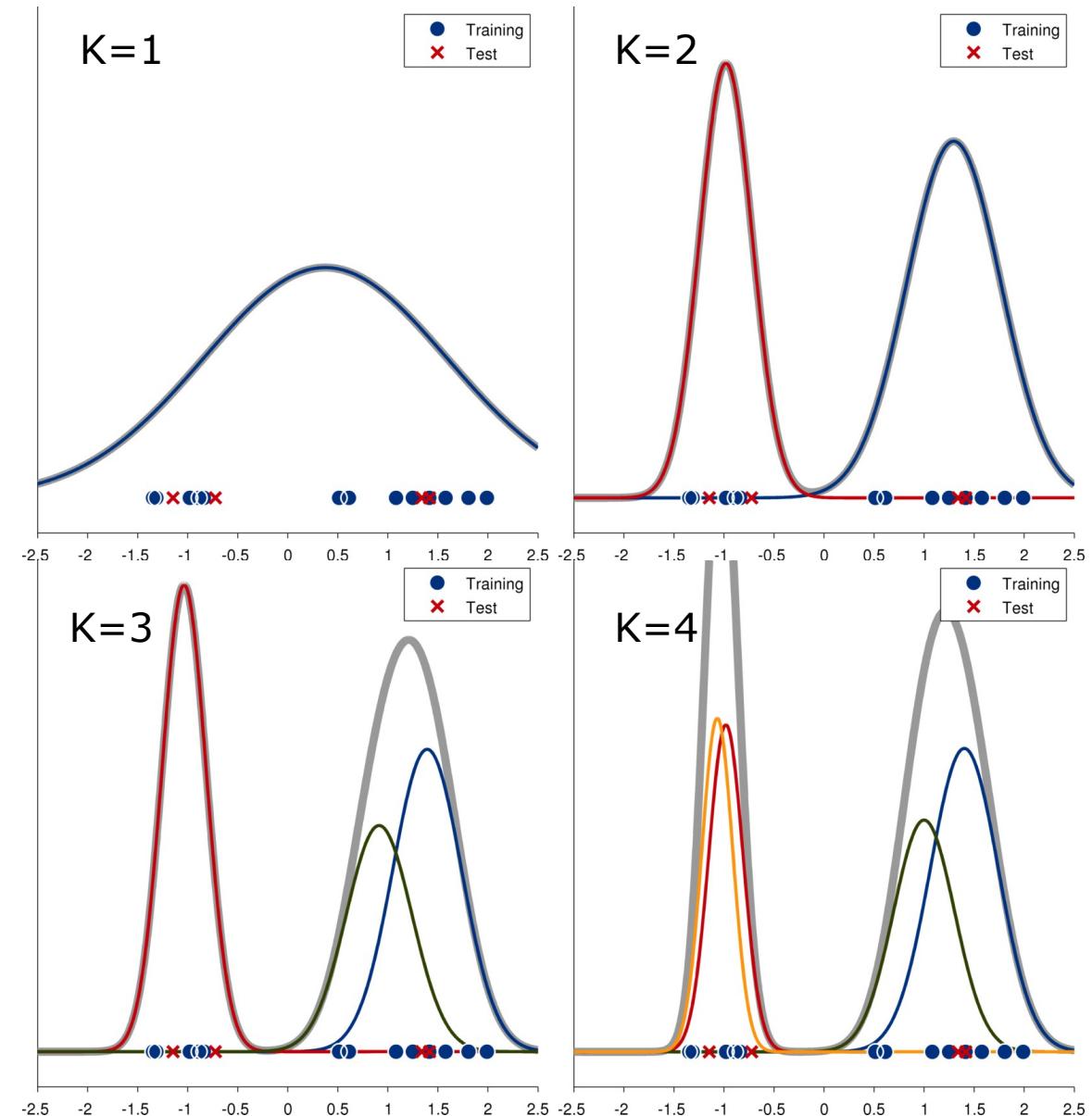
- Selecting complexity using crossvalidation



EM 1st iteration

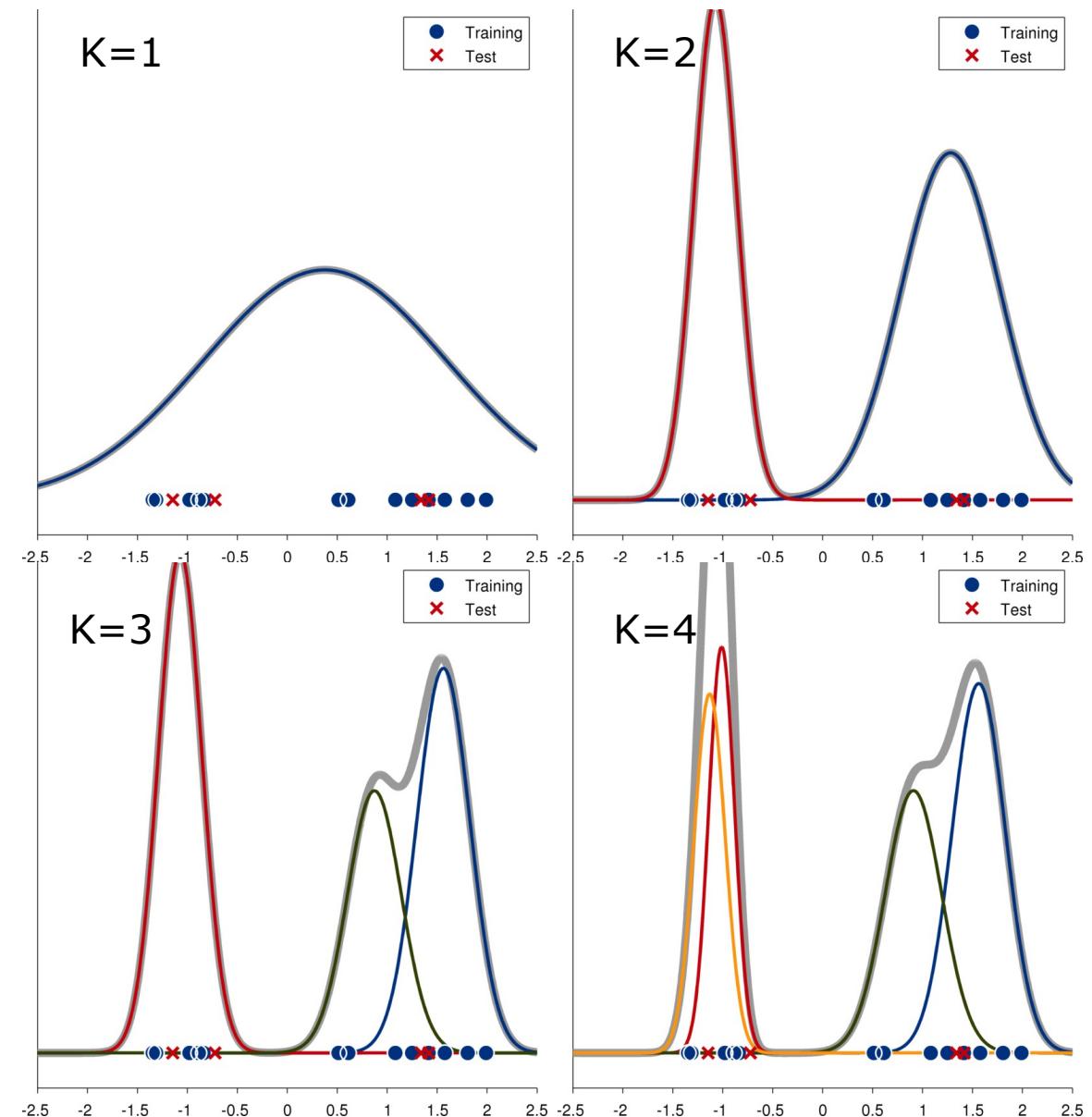
Mixture models

- Selecting complexity using crossvalidation



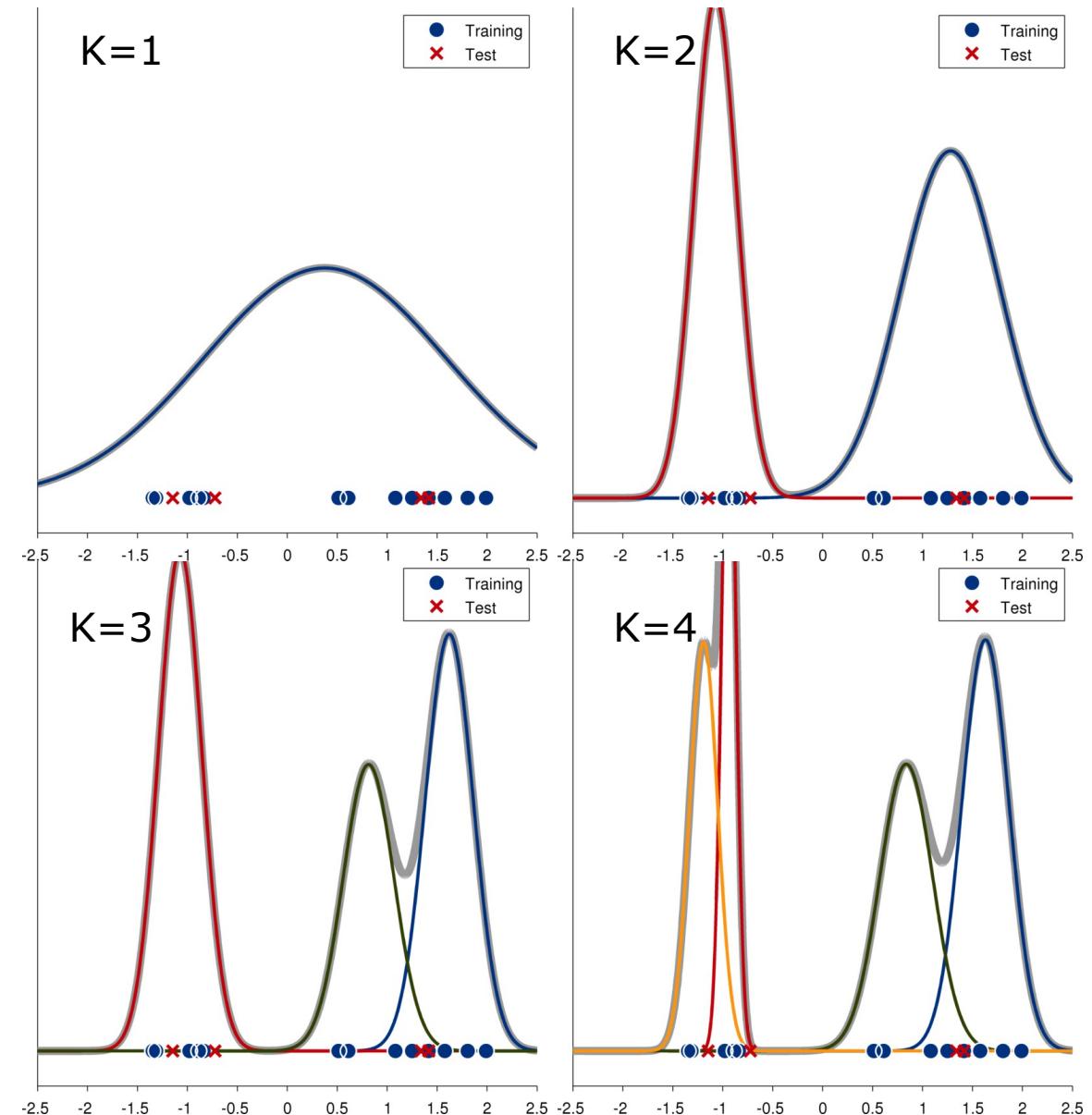
Mixture models

- Selecting complexity using crossvalidation



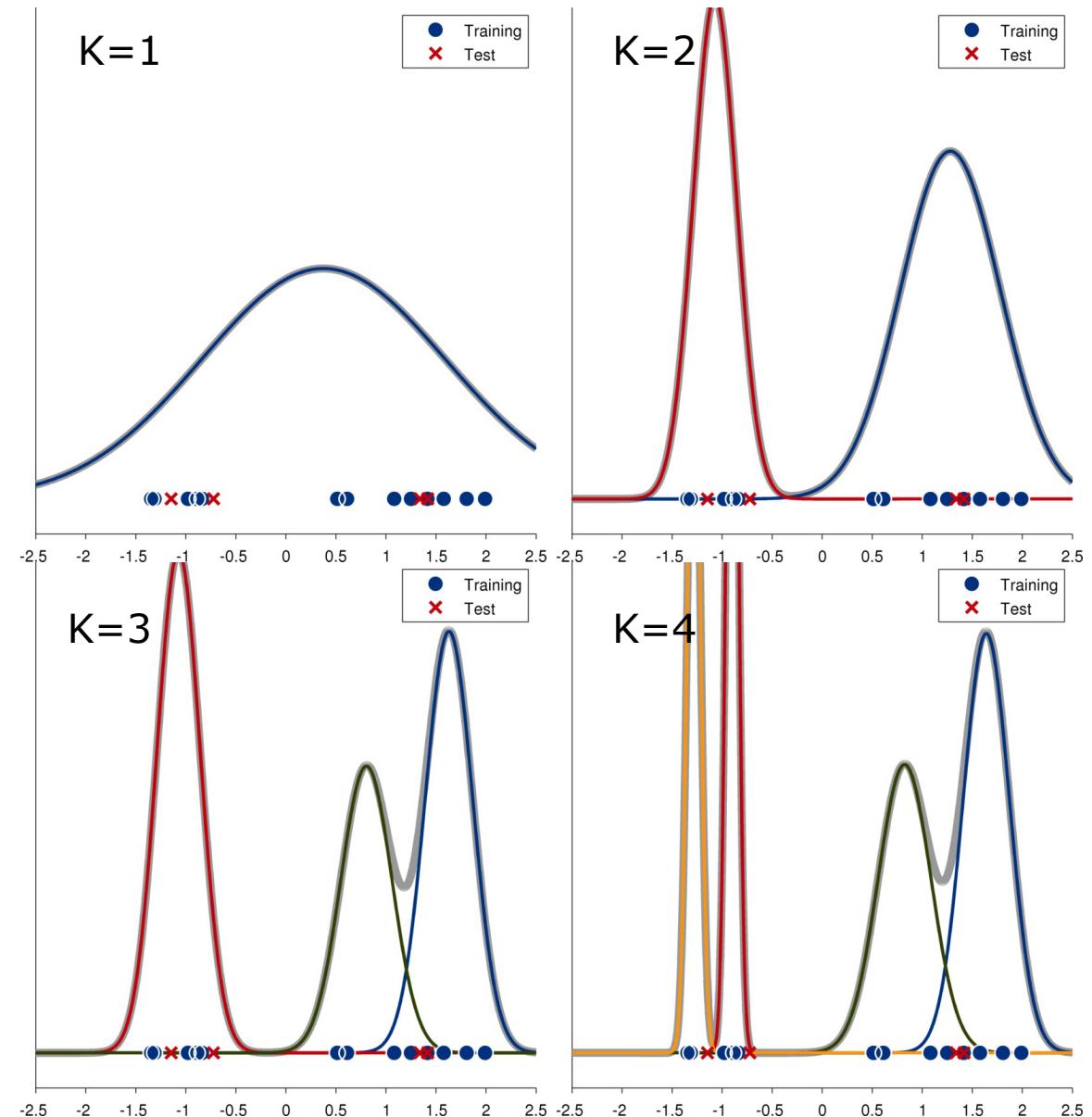
Mixture models

- Selecting complexity using crossvalidation



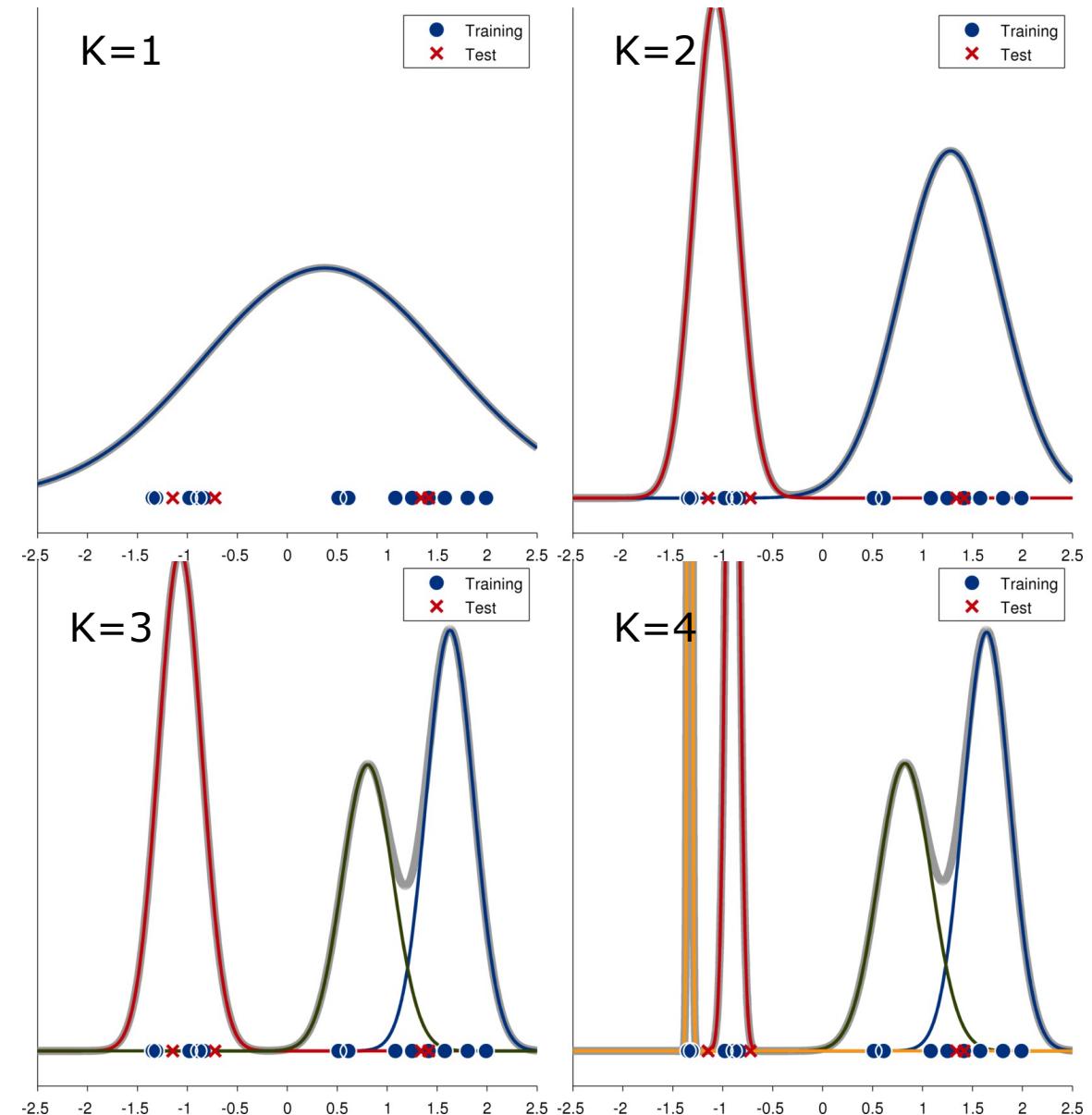
Mixture models

- Selecting complexity using crossvalidation



Mixture models

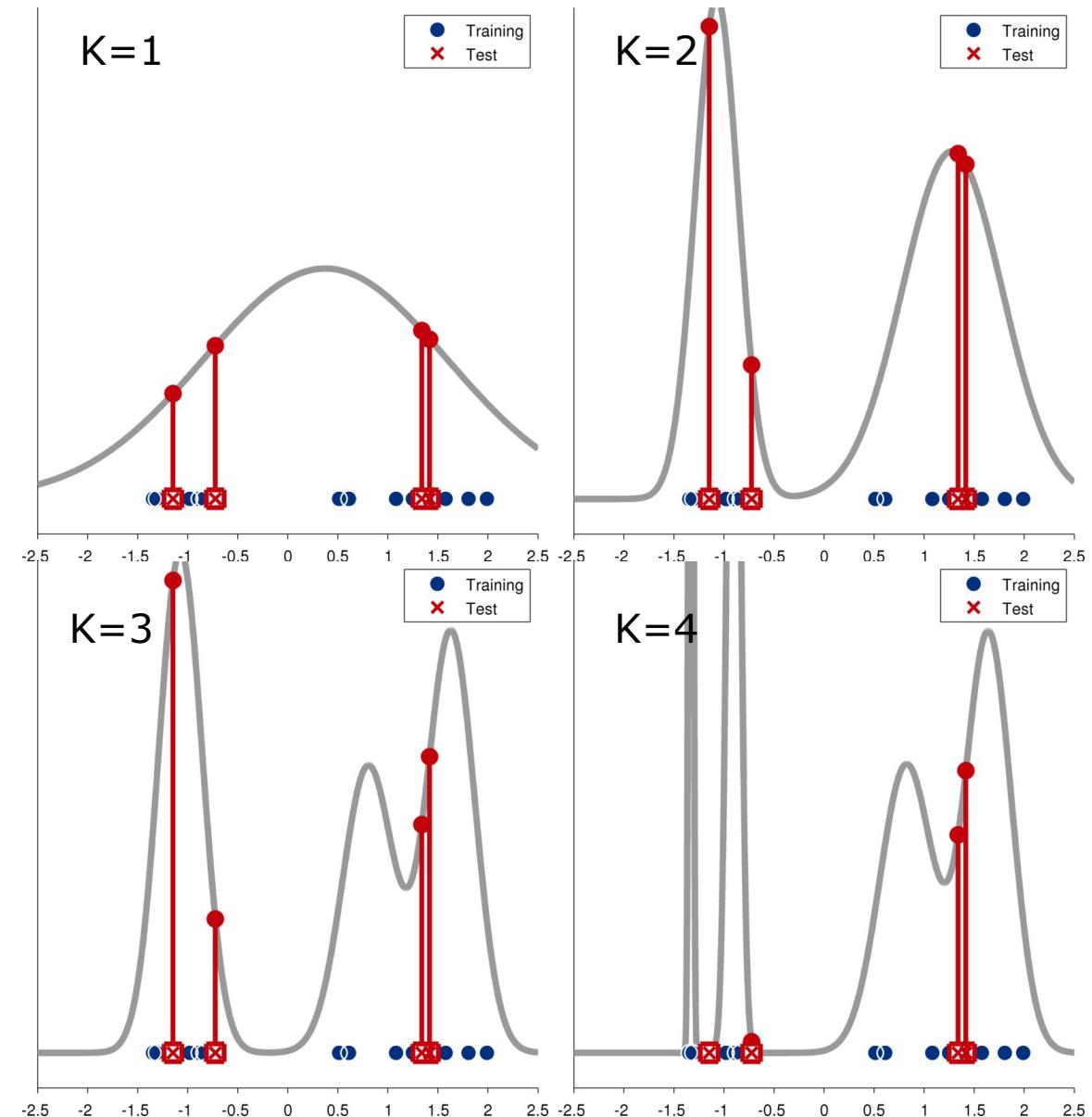
- Selecting complexity using crossvalidation



Test data evaluation

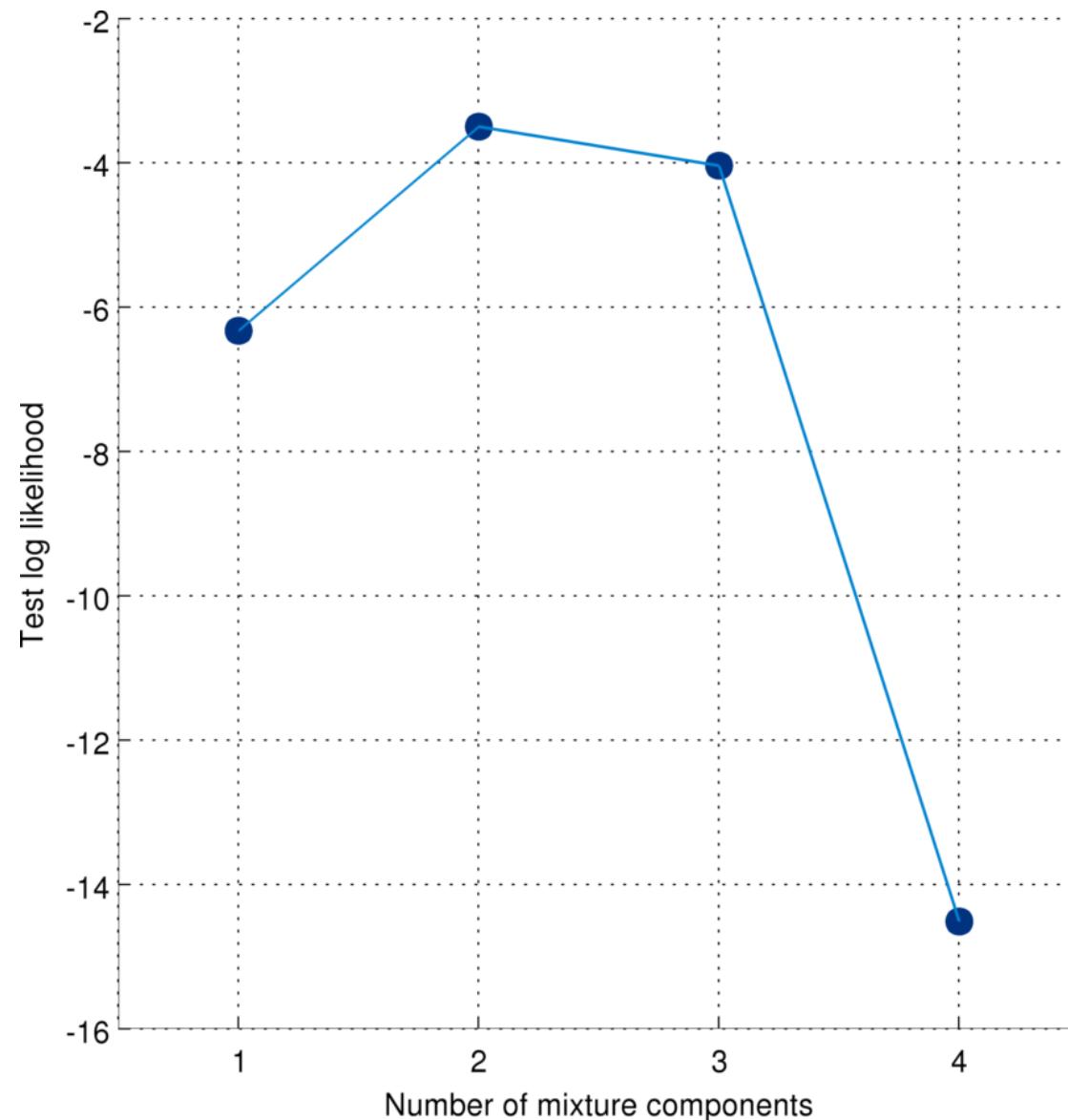
Mixture models

- Selecting complexity using crossvalidation



Mixture models

- Selecting complexity using crossvalidation



K-means versus GMM

K-means

- No guarantee of optimal solution
- Does not model shape of clusters
- Does not model the size of clusters
- Difficult to assess the number of clusters to use particularly when there is no ground truth

Gaussian mixture model (GMM)

- No guarantee of optimal solution (even more local minima issues due to the additional model parameters)
- Models shape of cluster as ellipsoid
- Models the size of clusters
- Possible to estimate the number of components by cross-validation



<http://siguealconejoblanco.es/comics/merchandising/escultura-de-superman-vs-muhammad-ali/>

K-means versus GMM

K-means

- No guarantee of optimal solution
- Does not model shape of clusters
- Does not model the size of clusters
- Difficult to assess the number of clusters to use particularly when there is no ground truth

Gaussian mixture model (GMM)

- No guarantee of optimal solution (even more local minima issues due to the additional model parameters)
- Models shape of cluster as ellipsoid
- Models the size of clusters
- Possible to estimate the number of components by cross-validation



<http://siguealconejoblanco.es/comics/merchandising/escultura-de-superman-vs-muhammad-ali/>

Association Mining

Association mining Agarwal

Ca. 24.900 resultater (0,07 sek.)

Tip: Søg efter resultater på Dansk alene. Du kan ændre dine sprogindstillinger i Indstillinger for Scholar.

[Mining association rules between sets of items in large databases](#)
 R Agrawal, T Imielinski, A Swami - ACM SIGMOD Record, 1993 - dl.acm.org
 Abstract We are given a large database of customer transactions. Each transaction consists of items purchased by a customer in a visit. We present an efficient algorithm that generates all significant **association** rules between items in the database. The algorithm incorporates ...
 Citeret af 12601 Relaterede artikler Alle 94 versioner Importer til BibTeX Flere▼

[\[PDF\] Fast algorithms for mining association rules](#)
 R Agrawal, R Srikant - Proc. 20th Int. Conf. Very Large Data Bases, VLDB, 1994 - it.uu.se
 Abstract We consider the problem of discovering **association** rules between items in a large database of sales transactions. We present two new algorithms for solving this problem that are fundamentally different from the known algorithms. Experiments with synthetic as well ...
 Citeret af 14210 Relaterede artikler Alle 309 versioner Importer til BibTeX Flere▼

[Mining quantitative association rules in large relational tables](#)
 R Srikant, R Agrawal - ACM SIGMOD Record, 1996 - dl.acm.org
 Abstract We introduce the problem of **mining association** rules in large relational tables containing both quantitative and categorical attributes. An example of such an **association** might be " 10% of married people between age 50 and 60 have at least 2 cars". We deal ...
 Citeret af 1717 Relaterede artikler Alle 62 versioner Importer til BibTeX Flere▼

[\[BOC\] Mining generalized association rules](#)
 R Srikant, R Agrawal - 1995 - wwwqbic.almaden.ibm.com
 ABSTRACT: We introduce the problem of **mining generalized association** rules. Given a large database of transactions, where each transaction consists of a set of items, and a taxonomy is-a hierarchy on the items, we find associations between items at any level of ...
 Citeret af 1680 Relaterede artikler Alle 82 versioner Importer til BibTeX Flere▼

Association rule discovery: Definition

- Given a set of **records**
 - Each containing a number of **items from a set**
- **Goal:** Produce dependency rules
 - Predict the occurrence of an item based on occurrences of other items

Association rule discovery: Example

Market basket analysis

Training set	Rules discovered
<ol style="list-style-type: none">1. {Bread, Soda, Milk}2. {Beer, Bread}3. {Beer, Soda, Diaper, Milk}4. {Beer, Bread, Diaper, Milk}5. {Soda, Diaper, Milk}	$\{\text{Milk}\} \rightarrow \{\text{Soda}\}$ $\{\text{Diaper}, \text{Milk}\} \rightarrow \{\text{Beer}\}$

Market basket data

- Representation as

Transaction table

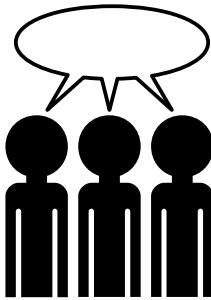
ID	Items
1	Bread, Soda, Milk
2	Beer, Bread
3	Beer, Soda, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Soda, Diaper, Milk

Data matrix

ID	Bread	Soda	Milk	Beer	Diaper
1	1	1	1	0	0
2	1	0	0	1	0
3	0	1	1	1	1
4	1	0	1	1	1
5	0	1	1	0	1

Association analysis

- **Itemset**
 - For example {Bread, Soda, Milk}, {Milk, Diaper}, {}
- **Support** for an itemset **X**
 - Percentage of transactions that contain **X**
- **Association rule**
 - Expression of the form: **X** \rightarrow **Y**
where **X** and **Y** are disjoint item sets
- **Support** for an association rule **X** \rightarrow **Y**
 - Percentage of transactions that contain **X** \cup **Y**
$$s(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{N} = P(X, Y)$$
- **Confidence** for an association rule **X** \rightarrow **Y**
 - Percentage of transactions containing **X** that also contain **Y**
$$c(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{\sigma(X)} = \frac{P(Y, X)}{P(X)} = P(Y|X)$$



Group exercise

- What is the **support** for
 - {Bread}
 - {Milk, Diaper}
- What is the **support** and **confidence** for
 - {Bread} \rightarrow {Milk}
 - {} \rightarrow {Milk}
- Find an **itemset** with
 - 0% support
 - 100% support
- Find an **association rule** with
 - 0% confidence
 - 100% confidence

ID	Bread	Soda	Milk	Beer	Diaper
1	1	1	1	0	0
2	1	0	0	1	0
3	0	1	1	1	1
4	1	0	1	1	1
5	0	1	1	0	1

Itemset

- For example
 {Bread, Soda, Milk}, {Milk, Diaper}, {}

Support for an itemset X

- Percentage of transactions that contain X

Association rule

- Expression of the form: $X \rightarrow Y$
 where X and Y are disjoint item sets

Support for an association rule $X \rightarrow Y$

- Percentage of transactions that contain $X \cup Y$

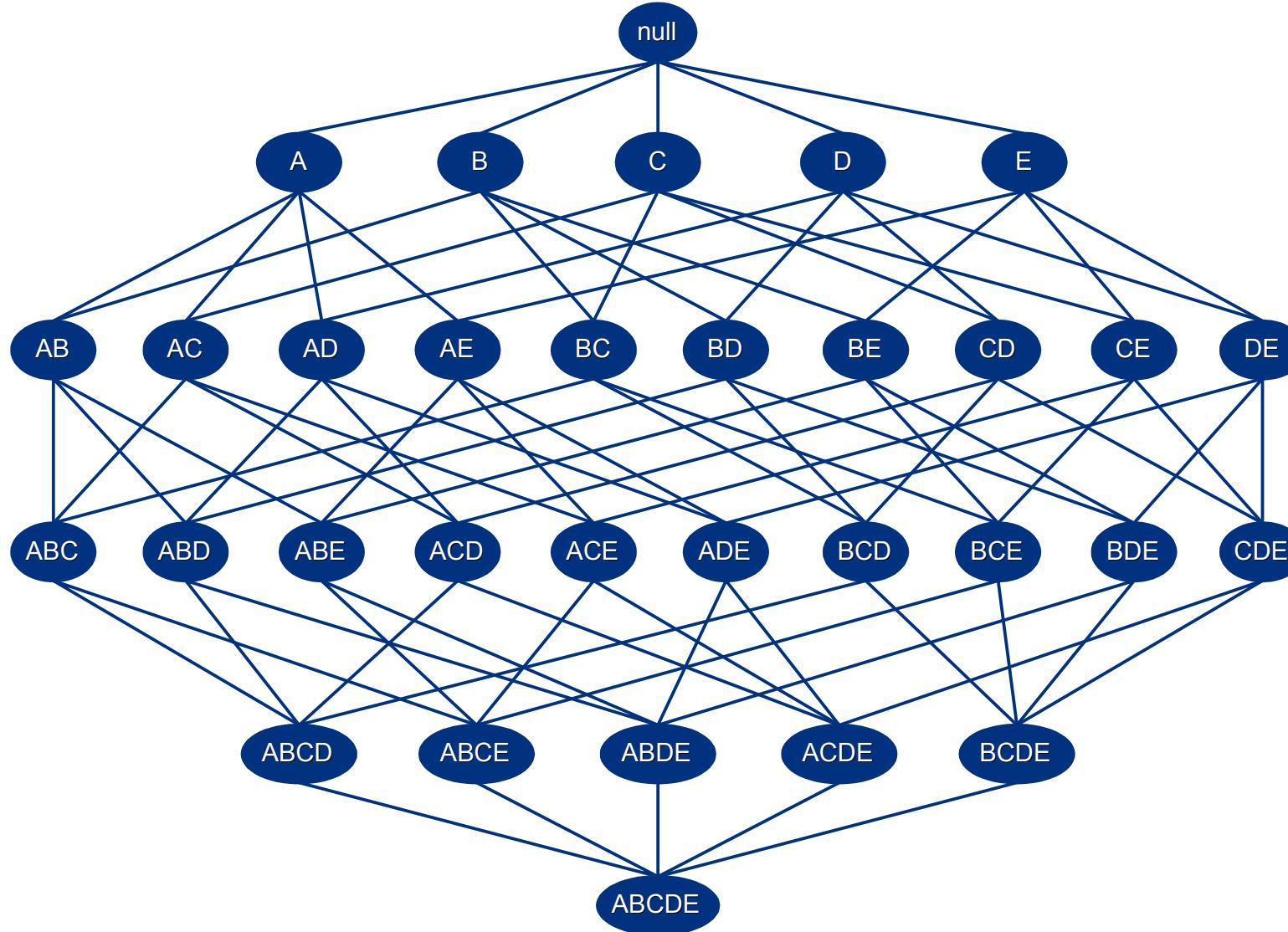
Confidence for an association rule $X \rightarrow Y$

- Percentage of transactions containing X that also contain Y

Association rule mining

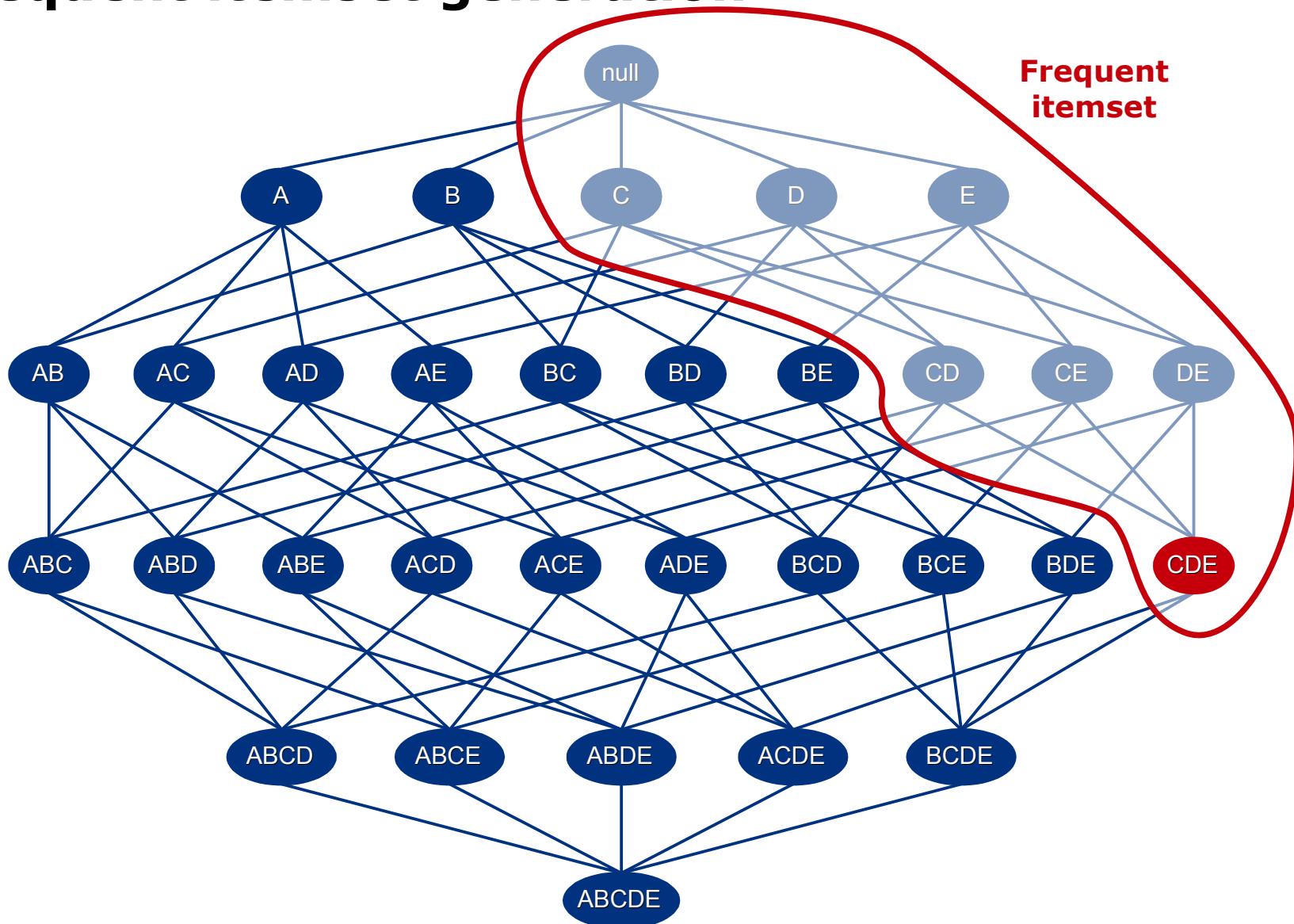
- Find all association rules that have
 - **Support** $\geq \text{minsup}$
 - **Confidence** $\geq \text{minconf}$
- Approach
 - **Frequent itemset generation**
 - Generate a list of all **itemsets** with **Support** $\geq \text{minsup}$
 - **Association rule generation**
 - Generate all **association rules** with **Confidence** $\geq \text{minconf}$

Frequent itemset generation



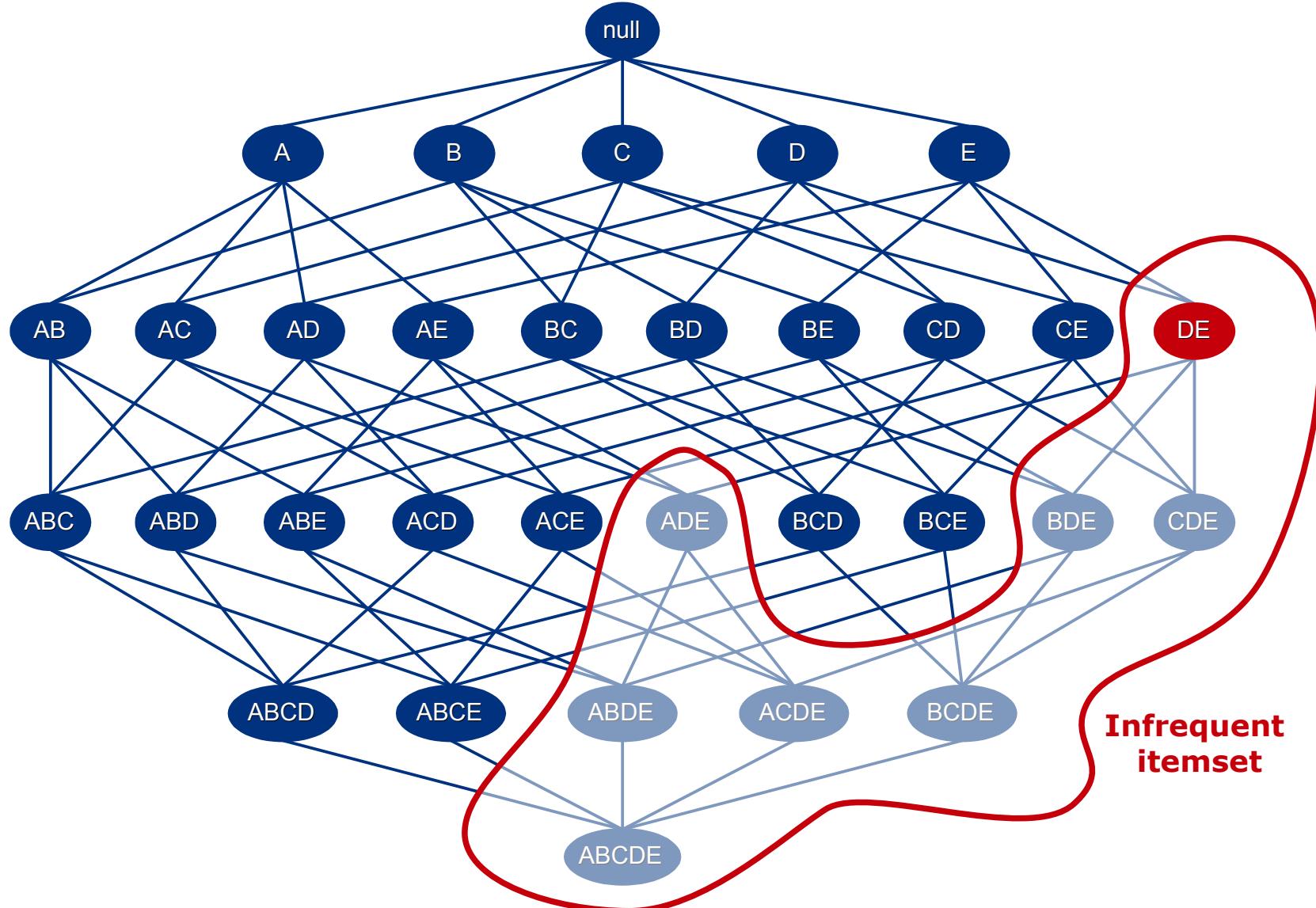
How many different itemsets can be created for a problem with a total of D items?

Frequent itemset generation

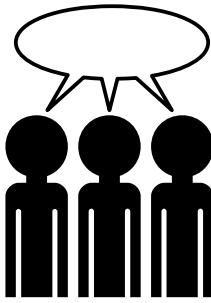


If an itemset is frequent, then all of its subsets must also be frequent

Frequent itemset generation



If an itemset is infrequent, then all of its supersets must also be infrequent

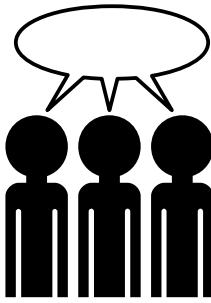


Group exercise

How many possible itemsets are there in the market basket below?

What are all the itemsets with support $\geq 35\%$?

	Juice	Milk	Beer	Cheese	Chocolate	Yoghurt	Sugar	Flour	Eeg	Wine
Customer 1	0	0	1	0	0	0	1	0	1	0
Customer 2	1	1	0	0	0	1	0	1	1	1
Customer 3	0	1	0	1	1	0	0	0	0	1
Customer 4	1	1	0	0	0	1	0	0	0	1
Customer 5	1	0	1	0	0	0	0	0	1	0
Customer 6	1	0	0	0	0	0	0	0	1	0
Customer 7	1	1	0	0	1	1	0	0	0	1
Customer 8	0	1	0	1	0	0	1	1	0	1
Customer 9	1	1	0	0	1	1	0	0	0	0
Customer 10	0	0	1	0	0	1	0	0	1	1



Group exercise

How many possible itemsets are there in the market basket below?

What are all the itemsets with support $\geq 35\%$?

	Juice	Milk	Beer	Cheese	Chocolate	Yoghurt	Sugar	Flour	Egg	Wine
Customer 1	0	0	1	0	0	0	1	0	1	0
Customer 2	1	1	0	0	0	1	0	1	1	1
Customer 3	0	1	0	1	1	0	0	0	0	1
Customer 4	1	1	0	0	0	1	0	0	0	1
Customer 5	1	0	1	0	0	0	0	0	1	0
Customer 6	1	0	0	0	0	0	0	0	1	0
Customer 7	1	1	0	0	1	1	0	0	0	1
Customer 8	0	1	0	1	0	0	1	1	0	1
Customer 9	1	1	0	0	1	1	0	0	0	0
Customer 10	0	0	1	0	0	1	0	0	1	1

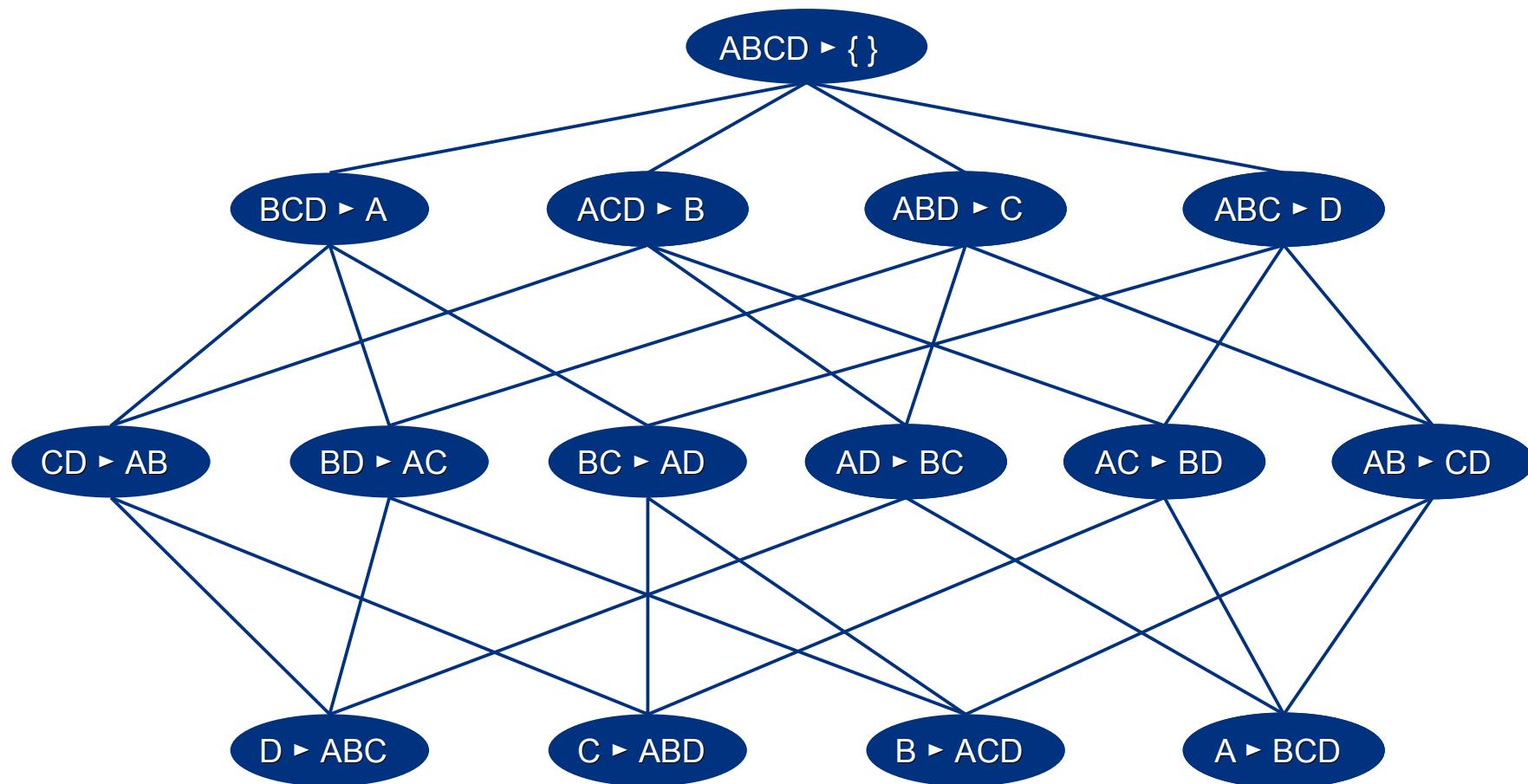
$2^{10} = 1024$ itemsets, itemsets with support $\geq 35\%$ are:

{Juice}, {Milk}, {Yoghurt}, {Egg}, {Wine}

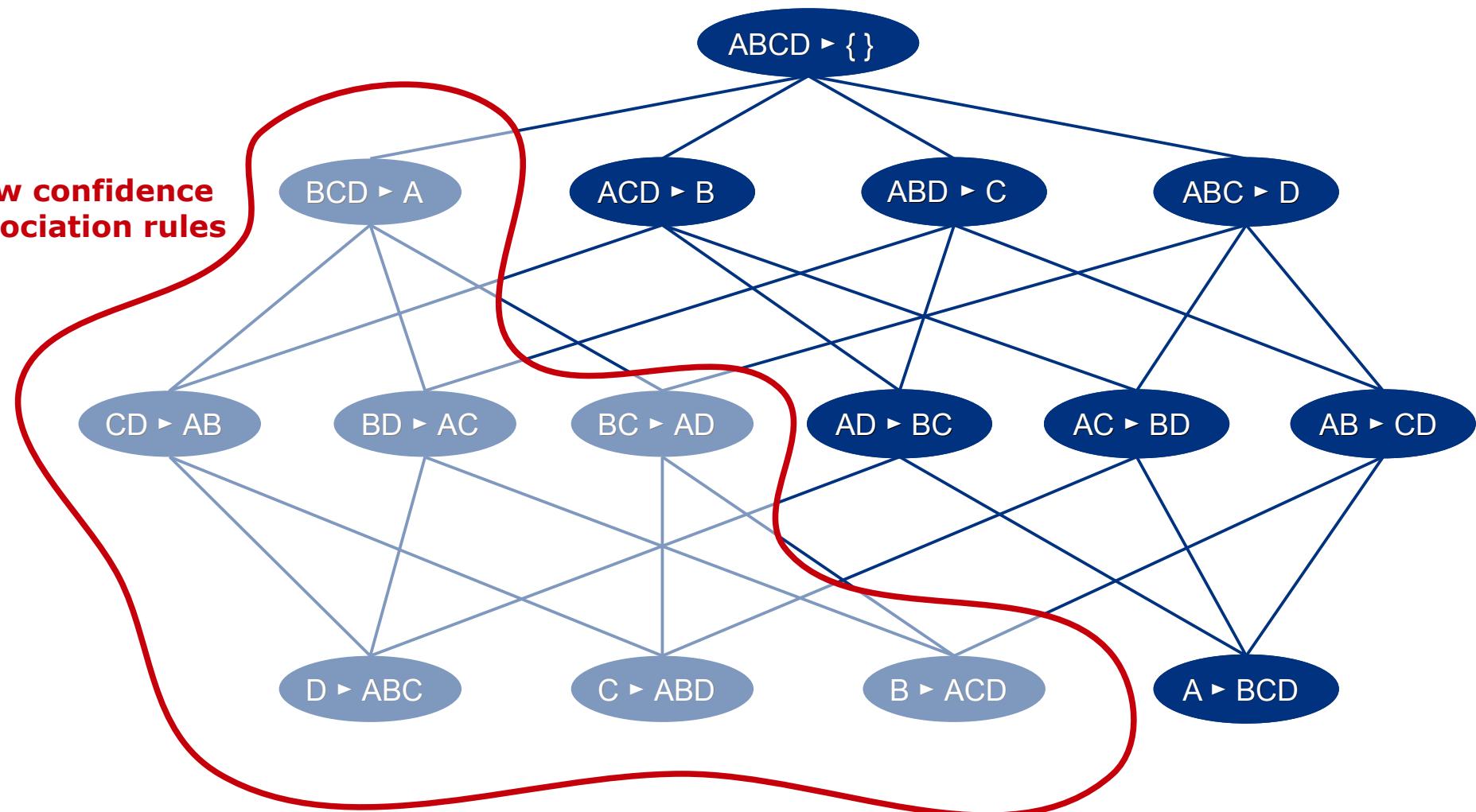
{Juice, Milk}, {Juice, Yoghurt}, {Milk, Yoghurt}, {Wine, Milk}, {Wine Yoghurt}

{Juice, Milk, Yoghurt}

Association rule generation



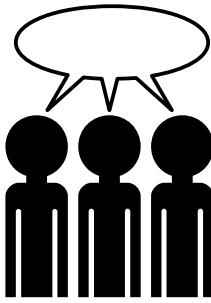
Association rule generation



Results for market basket example

Itemset	Support	Association rule	Support	Confidence
Milk	80%	{ } \rightarrow Milk	80%	80%
Bread	60%	Soda \rightarrow Milk	60%	100%
Soda	60%	Diaper \rightarrow Milk	60%	100%
Beer	60%	Soda, Diaper \rightarrow Milk	40%	100%
Diaper	60%	Beer, Diaper \rightarrow Milk	40%	100%
Diaper Milk	60%	Beer, Milk \rightarrow Diaper	40%	100%
Soda Milk	60%			
Bread Beer	40%			
Bread Milk	40%			
Soda Diaper	40%			
Beer Diaper	40%			
Beer Milk	40%			
Soda Diaper Milk	40%			
Beer Diaper Milk	40%			

ID	Bread	Soda	Milk	Beer	Diaper
1	1	1	1	0	0
2	1	0	0	1	0
3	0	1	1	1	1
4	1	0	1	1	1
5	0	1	1	0	1



Group exercise

- How can we do association mining for continuous data?

	Attribute 1	Attribute 2	Attribute 3
	0.3689	0.9827	0.6999
	0.4607	0.7302	0.6385
	0.9816	0.3439	0.0336
	0.1564	0.5841	0.0688
	0.8555	0.1078	0.3196
	0.6448	0.9063	0.5309
	0.3763	0.8797	0.6544
	0.1909	0.8178	0.4076
X=	0.4283	0.2607	0.8200
	0.4820	0.5944	0.7184
	0.1206	0.0225	0.9686
	0.5895	0.4253	0.5313
	0.2262	0.3127	0.3251
	0.3846	0.1615	0.1056
	0.5830	0.1788	0.6110
	0.2518	0.4229	0.7788
	0.2904	0.0942	0.4235
	0.6171	0.5985	0.0908
	0.2653	0.4709	0.2665
	0.8244	0.6959	0.1537

Binarize data according to percentiles

AttributeNames=	Attribute 1	Attribute 2	Attribute 3	AttributeNamesBin=	Attribute 1 0-50 %	Attribute 1 50-100 %	Attribute 2 0-33.3 %	Attribute 2 33.3-66.7 %	Attribute 2 66.7-100 %	Attribute 3 0-50%	Attribute 3 50-100%
X=	0.3689	0.9827	0.6999		1	0	0	0	1	0	1
	0.4607	0.7302	0.6385		0	1	0	0	1	0	1
	0.9816	0.3439	0.0336		0	1	0	1	0	1	0
	0.1564	0.5841	0.0688		1	0	0	1	0	1	0
	0.8555	0.1078	0.3196		0	1	1	0	0	1	0
	0.6448	0.9063	0.5309		0	1	0	0	1	0	1
	0.3763	0.8797	0.6544		1	0	0	0	1	0	1
	0.1909	0.8178	0.4076		1	0	0	0	1	1	0
	0.4283	0.2607	0.8200		0	1	1	0	0	0	1
	0.4820	0.5944	0.7184	Xbinary=	0	1	0	1	0	0	1
	0.1206	0.0225	0.9686		1	0	1	0	0	0	1
	0.5895	0.4253	0.5313		0	1	0	1	0	0	1
	0.2262	0.3127	0.3251		1	0	1	0	0	1	0
	0.3846	0.1615	0.1056		1	0	1	0	0	1	0
	0.5830	0.1788	0.6110		0	1	1	0	0	0	1
	0.2518	0.4229	0.7788		1	0	0	1	0	0	1
	0.2904	0.0942	0.4235		1	0	1	0	0	1	0
	0.6171	0.5985	0.0908		0	1	0	1	0	1	0
	0.2653	0.4709	0.2665		1	0	0	1	0	1	0
	0.8244	0.6959	0.1537		0	1	0	0	1	1	0

Further reading (not required)

Association Mining

- Rakesh Agrawal and Ramakrishnan Srikan “Fast Algorithms for Mining Association Rules”, Proc. 20th Int. Conf. Very Large Data Bases, 1994
- Rakesh Agrawal, Tomasz Imieliński and Arun Swami “Mining association rules between sets of items in large databases” Proceedings of the 1993 ACM SIGMOD international conference on Management of data

Gaussian Mixture Model

- Christopher M. Bishop “Pattern Recognition and Machine Learning”, Chapter 9, Springer 2006



http://commons.wikimedia.org/wiki/File:Old_book_bindings.jpg

Exam question examples

QI: Consider the market basket given to the right where 10 customers have purchased various types of fruits in a grocery store. Disregarding the empty set, what are all frequent itemsets with support $\geq 40\%$

A: {Orange}, {Apple}, {Banana}, {Pineapple}

B: {Orange}, {Apple}, {Banana}, {Pineapple}, {Orange, Apple}, {Orange, Pineapple}

C: {Orange}, {Apple}, {Banana}, {Pineapple}, {Orange, Apple}, {Orange, Pineapple}, {Apple, Pineapple}

D: {Orange}, {Apple}, {Banana}, {Pineapple}, {Orange, Apple}, {Orange, Pineapple}, {Apple, Pineapple}, {Orange, Apple, Pineapple}

	Orange	Apple	Banana	Kiwi	Pear	Pineapple
Customer 1	1	0	1	0	0	1
Customer 2	1	1	0	0	1	1
Customer 3	0	1	0	1	1	0
Customer 4	1	1	0	0	0	0
Customer 5	0	0	1	0	0	0
Customer 6	1	0	1	0	0	0
Customer 7	1	1	0	0	1	1
Customer 8	0	1	0	1	0	0
Customer 9	1	1	0	0	0	1
Customer 10	0	0	1	0	0	1

QII: What is the confidence of the decision rules $\{\text{Orange, Apple}\} \rightarrow \{\text{Pear}\}$

A: 25%

B: 50%

C: 75%

D: 100%

Correct answer QI: B, QII: B

02450 Introduction to machine learning and data modeling

$$f(x+\Delta x) = \sum_{i=0}^{\infty} \frac{(\Delta x)^i}{i!} f^{(i)}(x)$$

$\Theta^{\sqrt{17}} + \Omega \int_0^{\infty} \delta e^{i\pi} =$
 $\Sigma! \gg \chi^2 = \{2.7182818284$

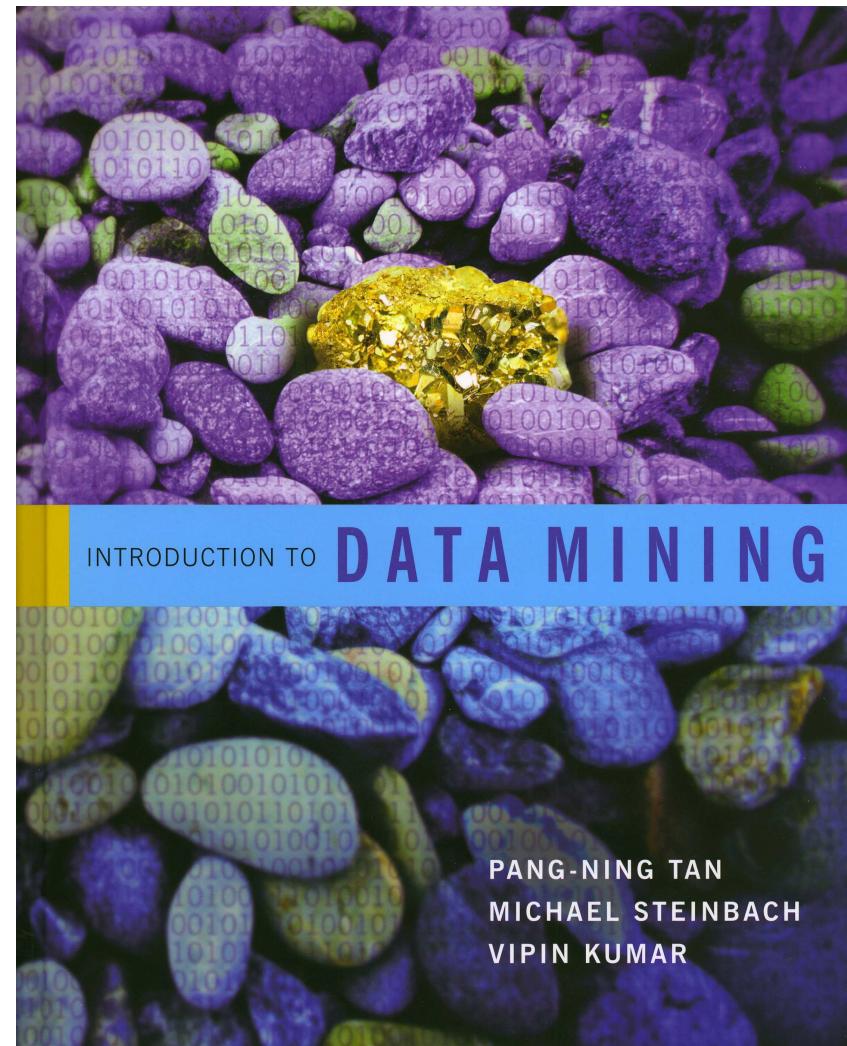
Reading material

Tan, Steinbach and Kumar
"Introduction to Data Mining"

Section 10.1-10.4

Groups of the day

Zhang Zheng
Simon Päusch
Hans Peter Halkier Nicolajsen
Nicolas Tiaki Otsu
Po-hao Huang
Zhen Li
Lotte Trap
Steen Schneidermann Lillelund
Konstantinos Blatsoukas
Mads Frøding Engels
Simon Benfeldt Jørgensen



Lecture schedule

1. Introduction
(Tan 1.1-1.4)

Data: Feature extraction and visualization

2. Data and feature extraction
(Tan 2.1-2.3 + B1 (+ A))
3. Measures of similarity and summary statistics
(Tan 2.4 + 3.1-3.2 + C1-C2)
4. Data visualization
(Tan 3.3)

Supervised learning: Classification and regression

5. Decision trees and linear regression
(Tan 4.1-4.3 + D)
6. Overfitting and performance evaluation
(Tan 4.4-4.6)
7. Nearest neighbor, naive Bayes, and artificial neural networks
(Tan 5.2-5.4)

8. Ensemble methods and multi class classifiers
(Tan 5.6-5.8)

Unsupervised learning: Clustering and density est.

9. K-means and hierarchical clustering
(Tan 8.1-8.3+8.5.7)
10. Mixture models and association mining
(Tan 9.2.2 + 6.1-6.3)

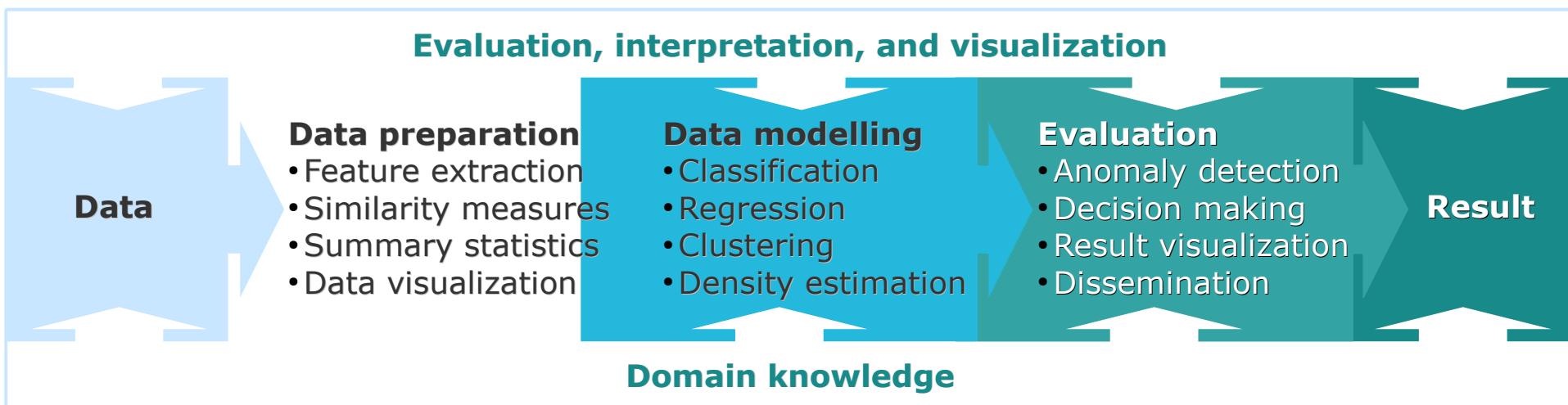
11. Density estimation and anomaly detection

(Tan 10.1-10.4)

Machine learning and data modelling in practice

12. Putting it all together: Summary and overview
13. Mini project

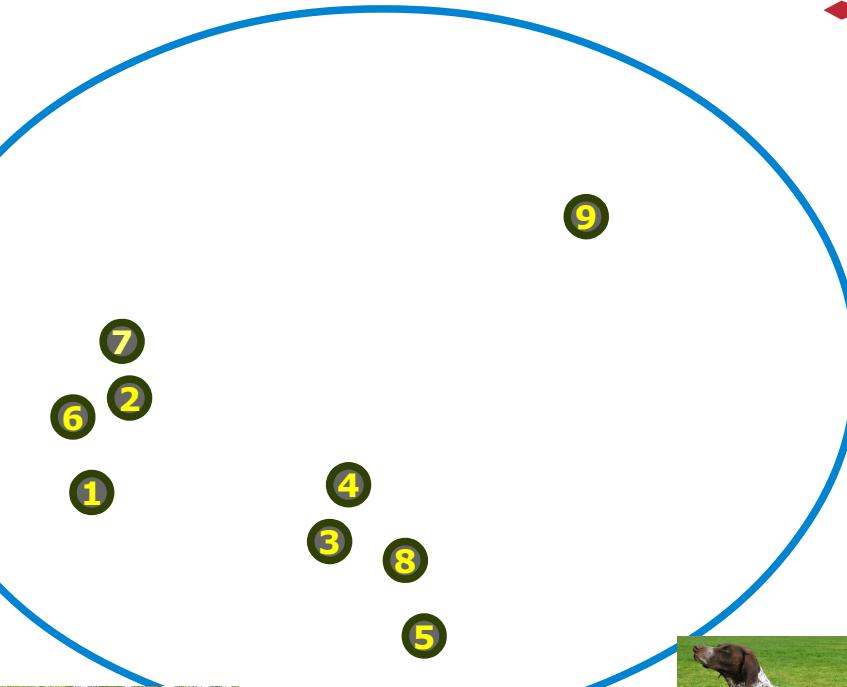
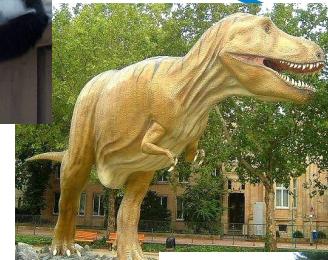
Data modeling framework



Imagine (yet again) you observe the world for the first time!



<http://www.clipartlord.com/category/baby-clip-art/>



http://commons.wikimedia.org/wiki/File:Abessinier_sorrel.jpg

http://commons.wikimedia.org/wiki/File:Cat_Eyes.jpg

http://commons.wikimedia.org/wiki/File:Black_white_cat_on_fence.jpg

<http://commons.wikimedia.org/wiki/File:Cat002.jpg>

<http://commons.wikimedia.org/wiki/File:Saurier2.jpg>

http://commons.wikimedia.org/wiki/File:Golden_Retriever_Dukedestiny01.jpg

<http://commons.wikimedia.org/wiki/File:Maspiri-Astro-SVE.jpg>

http://commons.wikimedia.org/wiki/File:GroC3%9Fer_Schweizer_Sennenhund.jpg

http://commons.wikimedia.org/wiki/File:GermanShorthPtr_wb.jpg

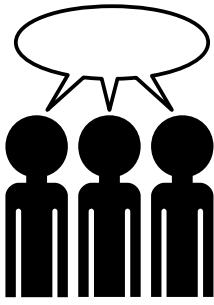
How do we detect anomalous objects (i.e., the dinosaur in the world of cats and dogs)

Anomaly detection: Definition

- Given a collection of data objects
 - Each object has associated a number of features
- Detect which objects **deviate from normal** behaviour

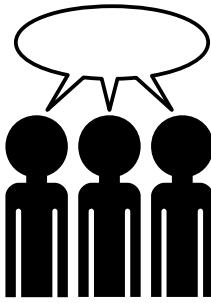
Anomaly detection: Example

- Credit card **fraud detection**
 - Recognize dubious credit card transactions based on the transaction history of the card holder
- Network **intrusion detection**
 - Detect hacker attacks, web crawlers etc.
- **Ecosystem disturbances**
 - Detect hurricanes, floods droughts, heat waves and fires
- **Health and medicine monitoring**
 - Detect abnormal behaviour in populations and patients
- **Fault detection in industry systems**
 - Detect when a wind turbine performs poorly due to ice coating on blades
- Detection of **outliers** in data measurements
 - Remove erroneous measurements due to misreading from an instrument



Group exercise

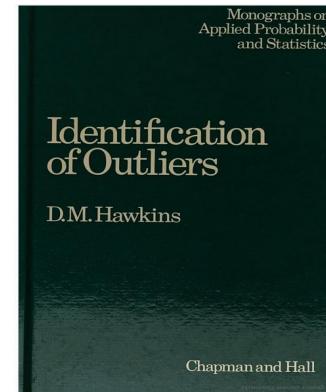
- Come up with **your own definition** of an outlier / anomaly
- How can we detect outliers using some of the methods you have already learned in the course?



Group exercise

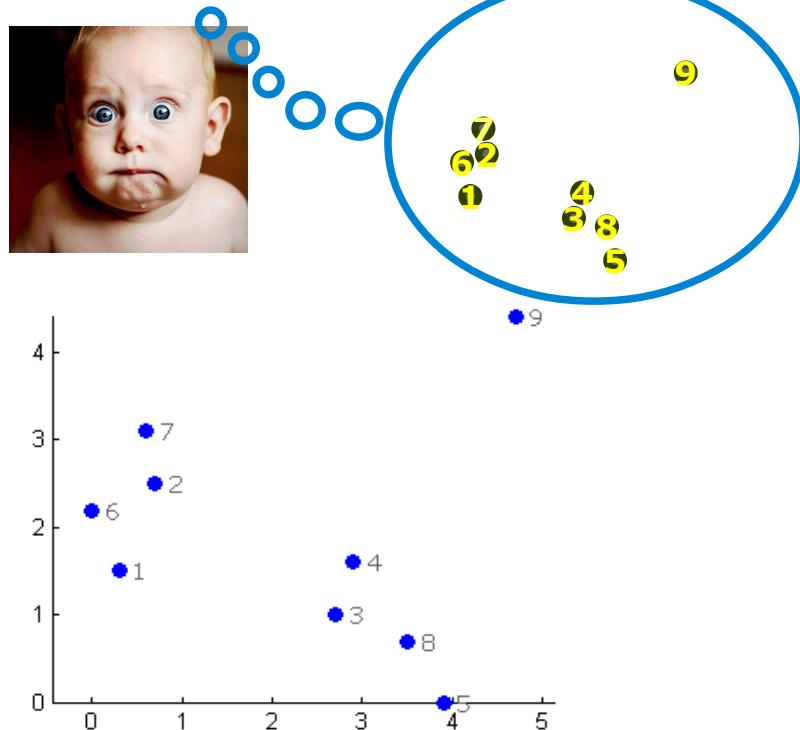
- Come up with **your own definition** of an outlier / anomaly
- How can we detect outliers using some of the methods you have already learned in the course?

Hawkins' definition of an outlier: *An outlier is an observation that differs so much from other observations as to arouse suspicion that it was generated by a different mechanism*



Probabilistic definition of an outlier: *An outlier is an object that has a low probability with respect to a probability distribution model of the data.*

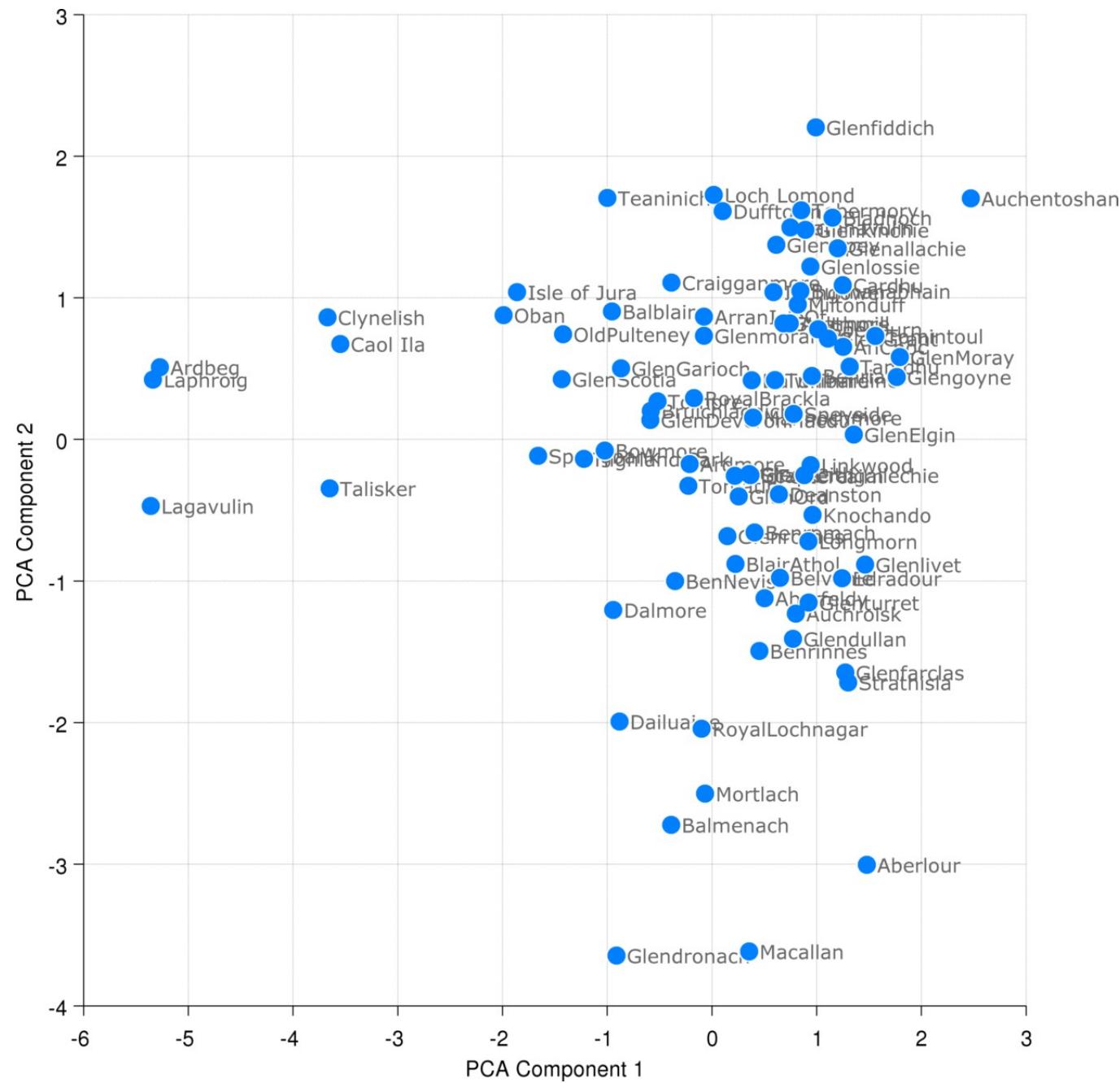
Data example I: Cats, Dogs and Dinosaurs



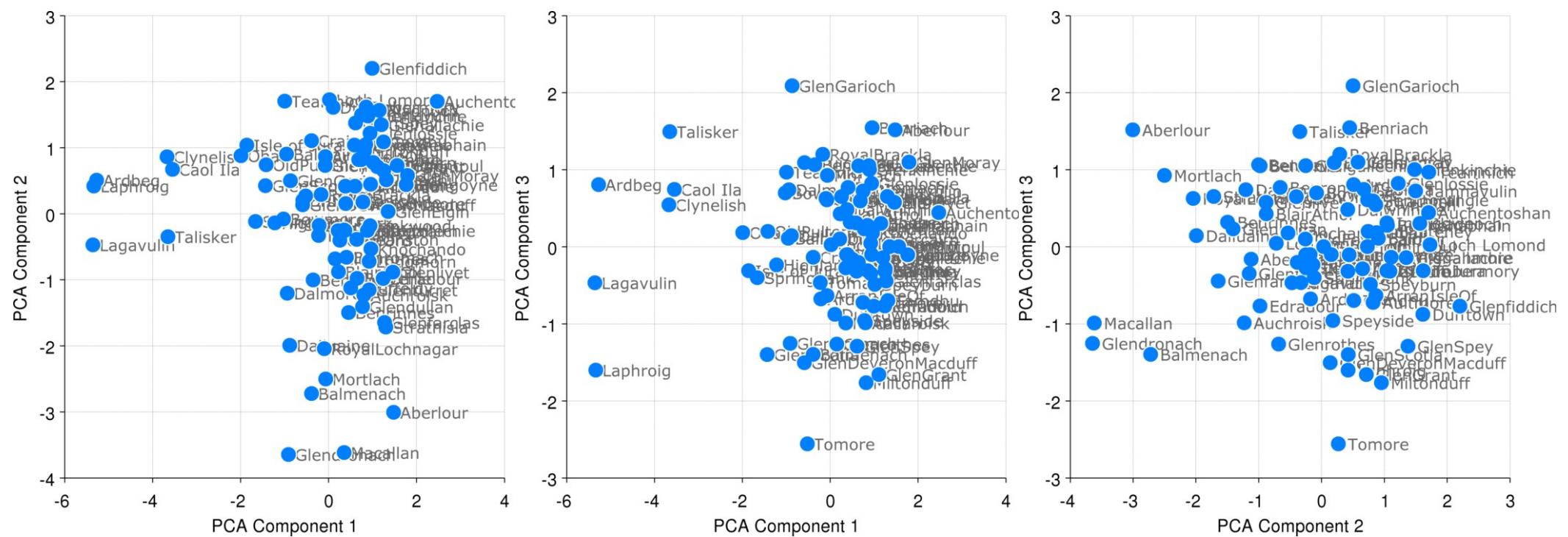
Data example II: Whisky

- 86 types of Scotch whisky
- Human ratings 1-5
- 12 taste categories
 - body, sweetness, smoky, medicinal, tobacco, honey, spicy, winey, nutty, malty, fruity, floral

PCA plot



PCA plot

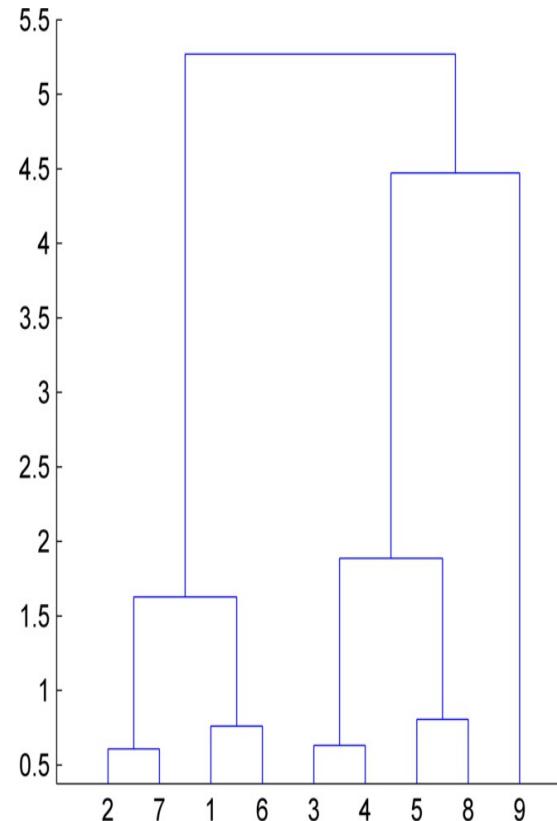


Dendrogram

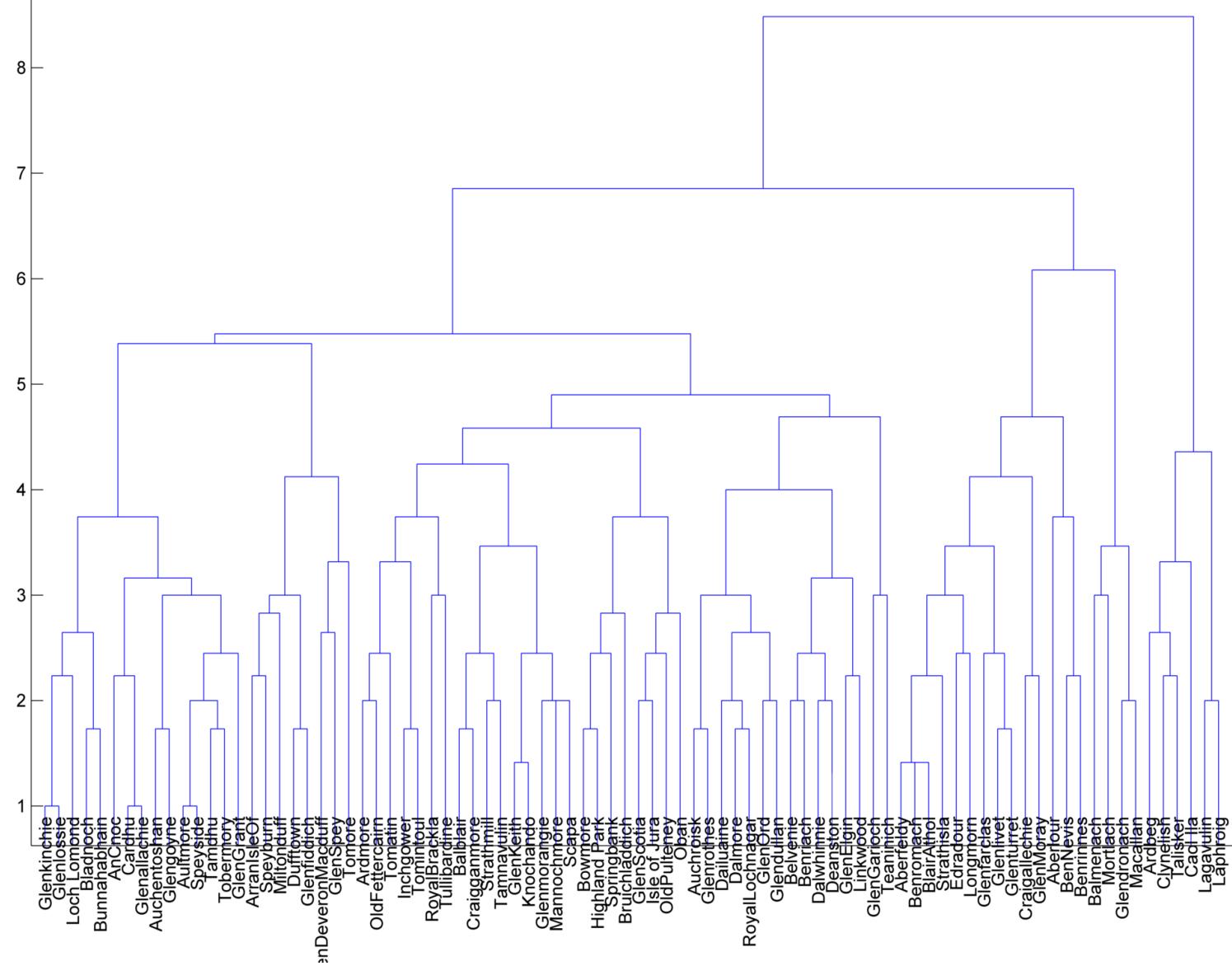


- Dendograms can be used to visualize relative distances between the observations

Data I: Cats, Dogs and Dinosaurs



Data II: Whisky



Approaches to anomaly detection

- **Density-based techniques**

- Estimate the density of data objects
- Outliers are:
 - Data objects in low density area

Approaches we will consider:

- Univariate normal distribution
- Kernel density estimation

- **Proximity-based techniques**

- Measure the distance between data objects
- Outliers are
 - Data objects far from the other data objects

Approaches we will consider:

- Mahanalobis distance to center of data
- Distance to K^{th} nearest neighbour
- Inverse average distance to K nearest neighbours (KNN density)
- Average relative KNN density

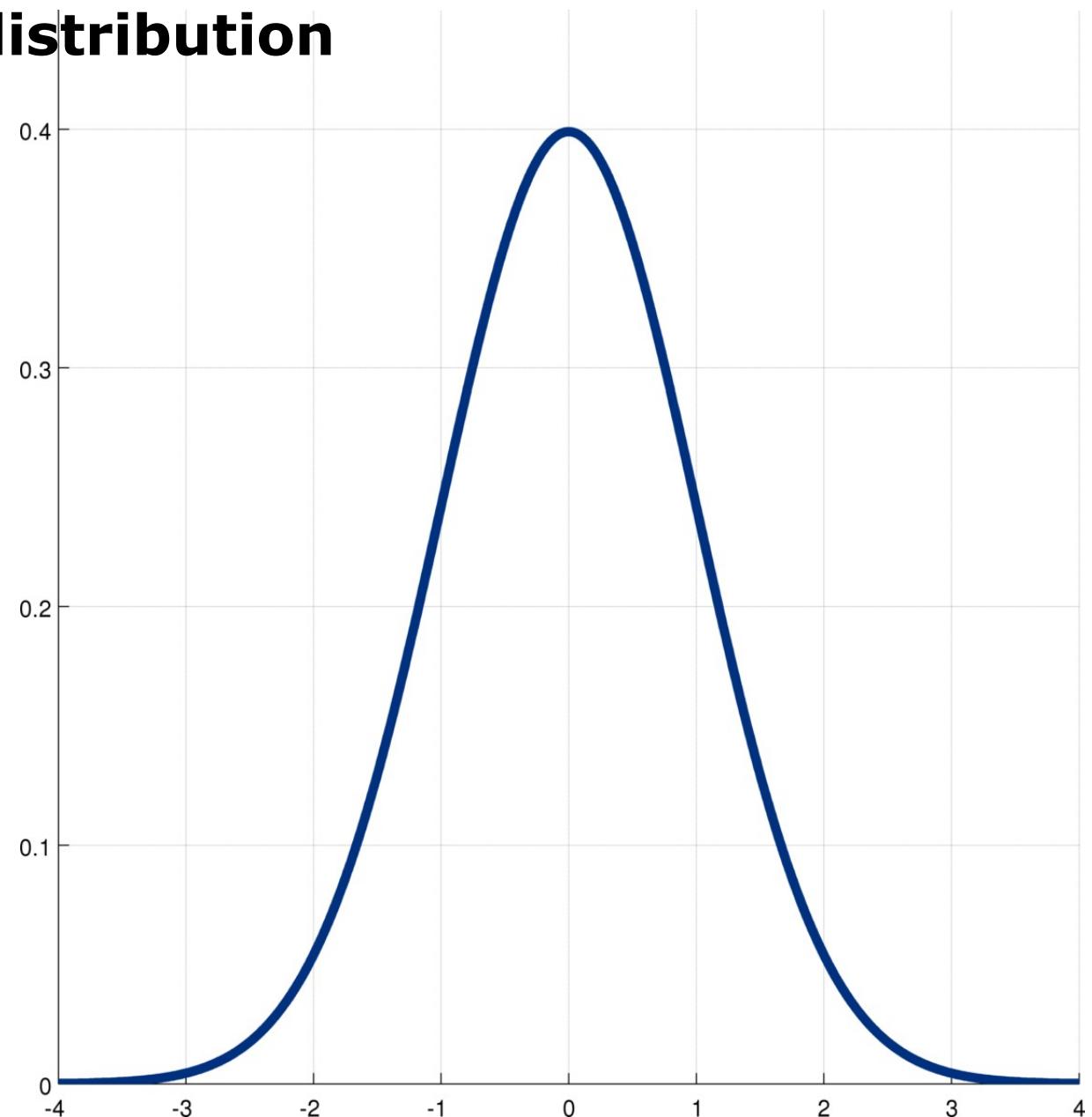
Density based techniques: Univariate normal distribution

- Map attribute to standard Normal variable

$$z = \frac{x - \mu}{\sigma}$$

- Choose a threshold

c	p(z >c)
1.0	0.3173
1.5	0.1336
2.0	0.0455
2.5	0.0124
3.0	0.0027
3.5	0.0005
4.0	0.0001



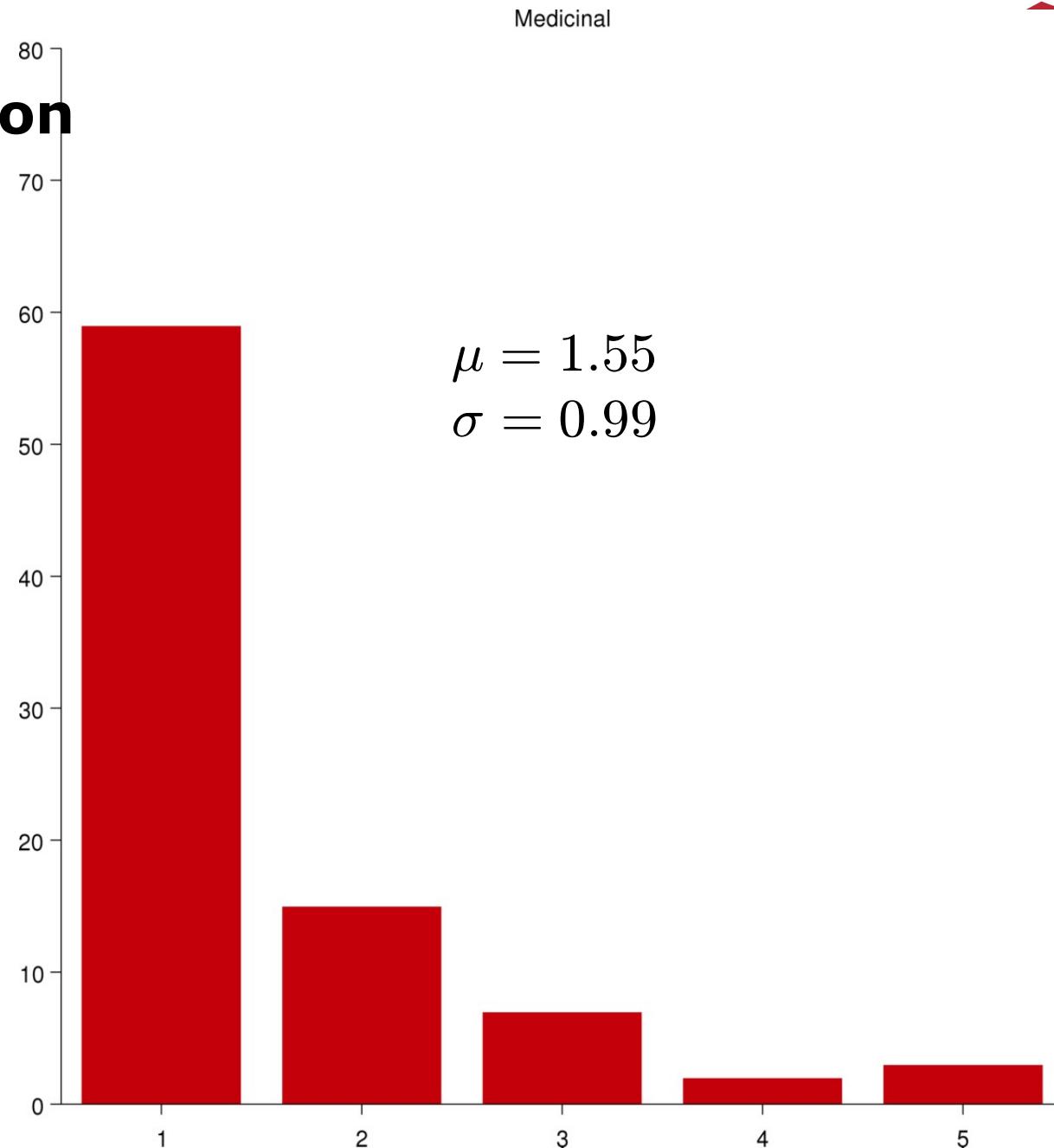
Normal distribution

- Map attribute to standard Normal variable

$$z = \frac{x - \mu}{\sigma}$$

- Choose a threshold

c	p(z >c)
1.0	0.3173
1.5	0.1336
2.0	0.0455
2.5	0.0124
3.0	0.0027
3.5	0.0005
4.0	0.0001



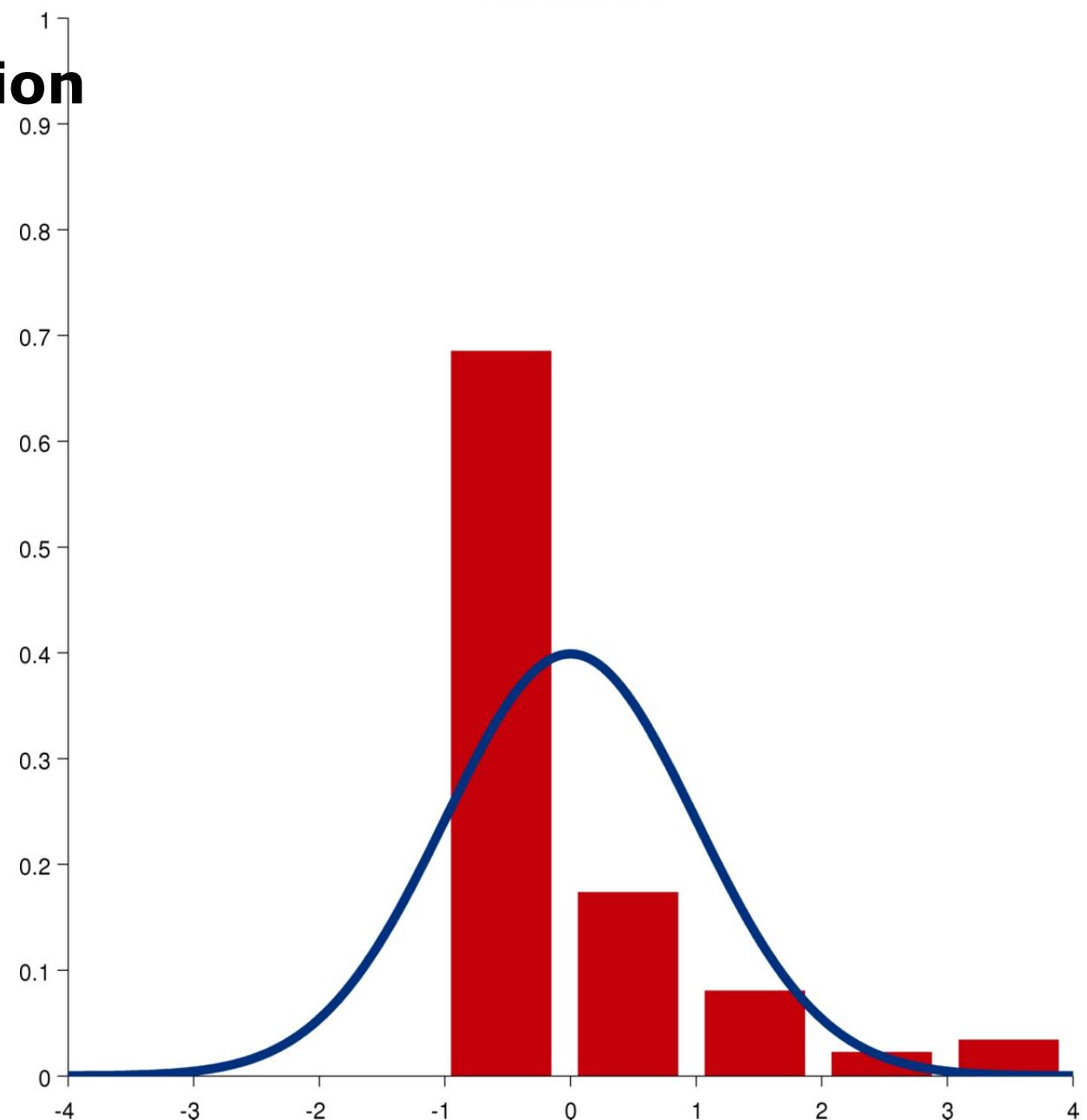
Normal distribution

- Map attribute to standard Normal variable

$$z = \frac{x - \mu}{\sigma}$$

- Choose a threshold

c	p(z >c)
1.0	0.3173
1.5	0.1336
2.0	0.0455
2.5	0.0124
3.0	0.0027
3.5	0.0005
4.0	0.0001



Normal distribution

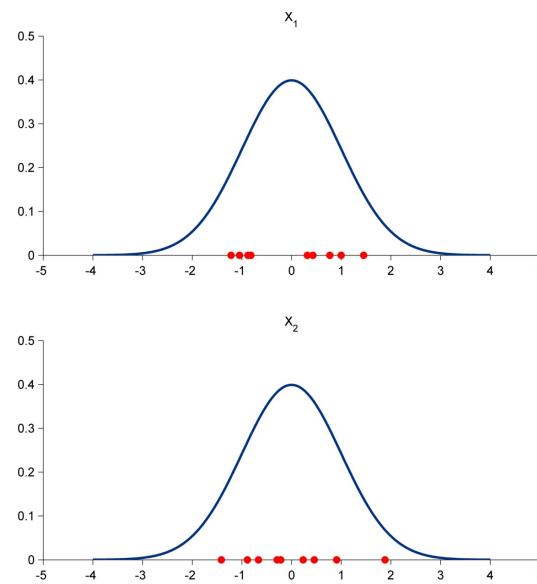
- Map attribute to standard Normal variable
- Choose a threshold

$$z = \frac{x - \mu}{\sigma}$$

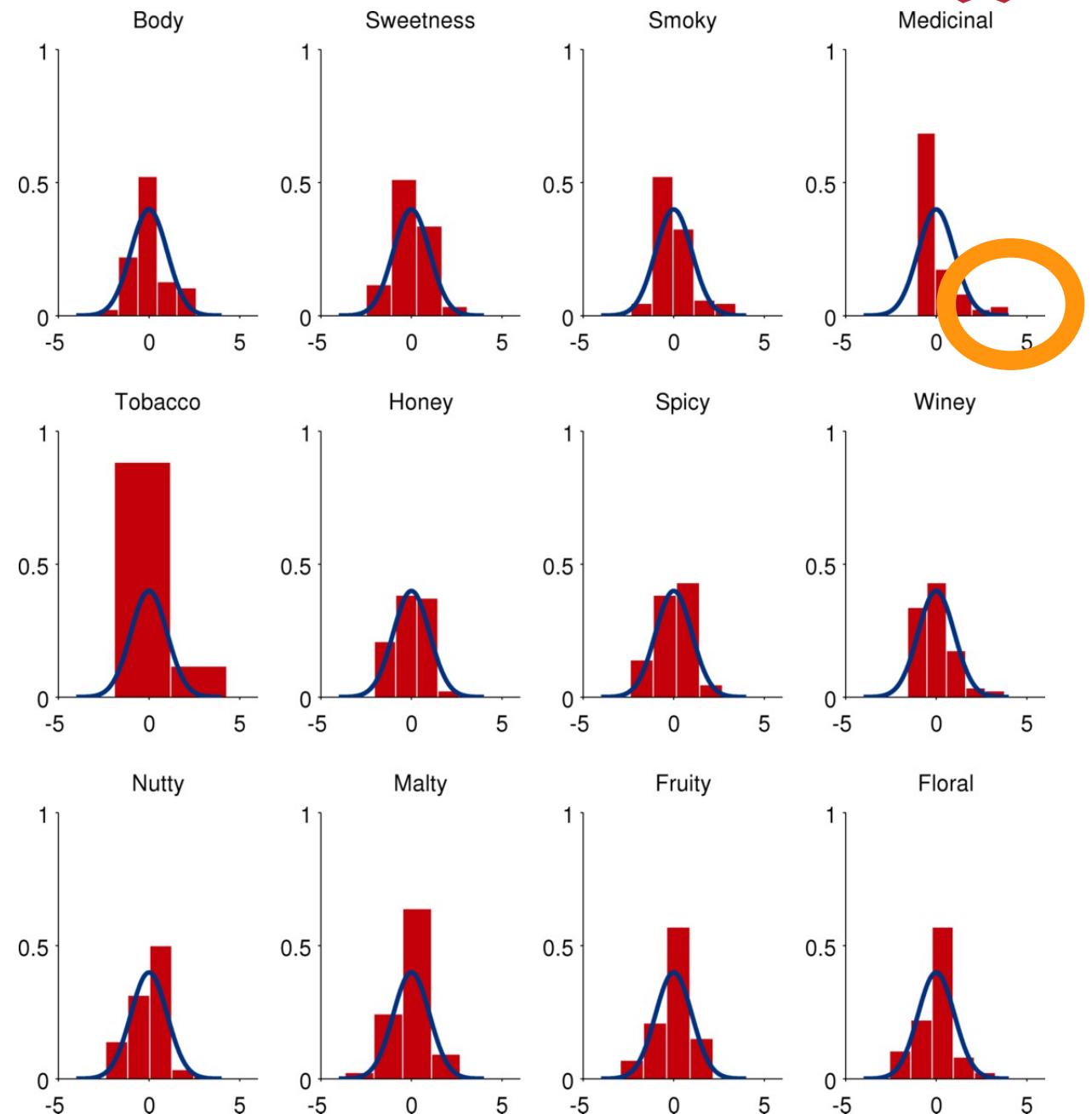
$$p(|z| > c) = 0.001$$

$$c = 3.2905$$

Data I: Cats, Dogs and Dinosaurs



Data II: Whisky



Medicinal: z-score

Normal distribution

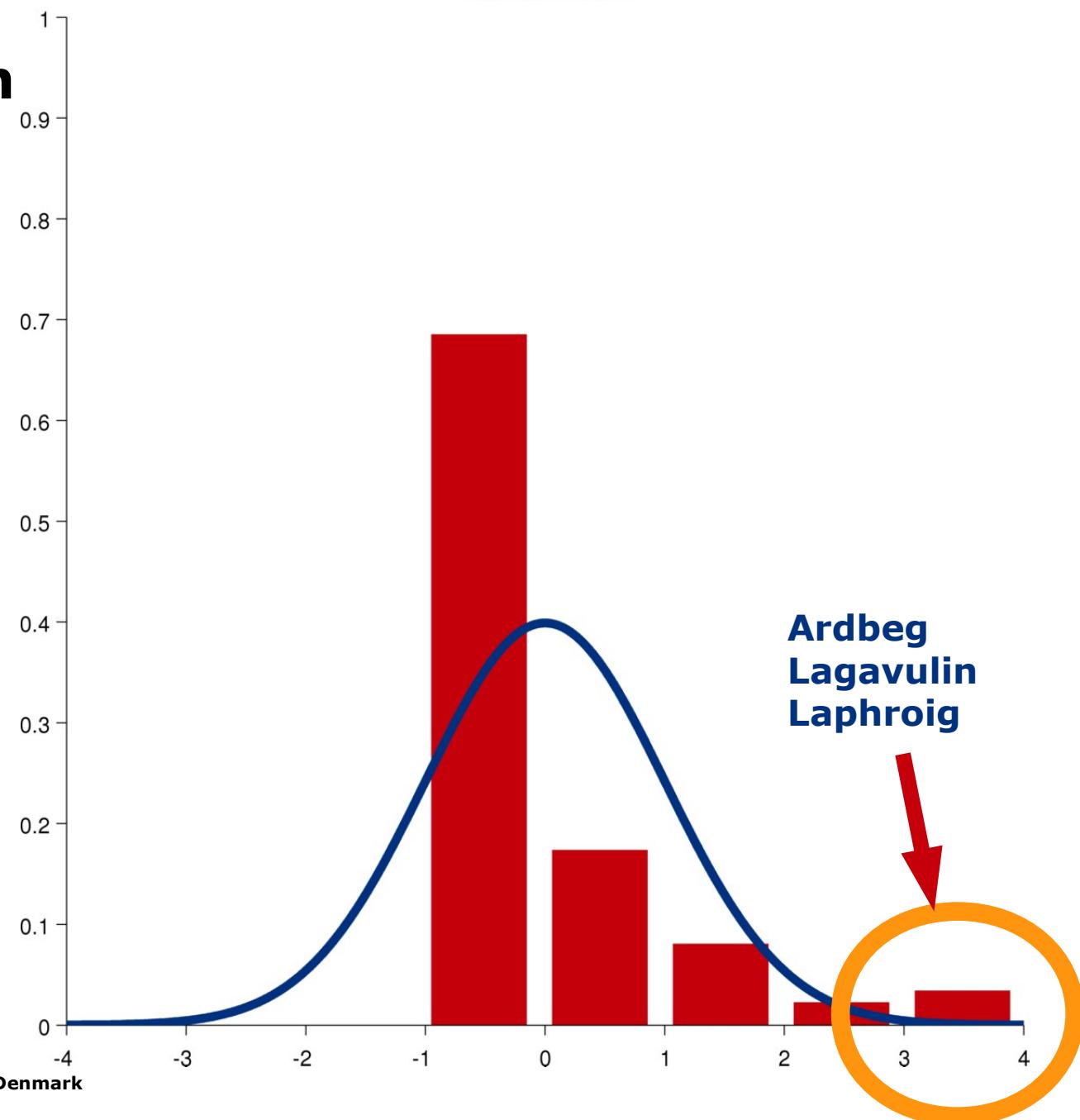
- Map attribute to standard Normal variable

$$z = \frac{x - \mu}{\sigma}$$

- Choose a threshold

$$p(|z| > c) = 0.001$$

$$c = 3.2905$$



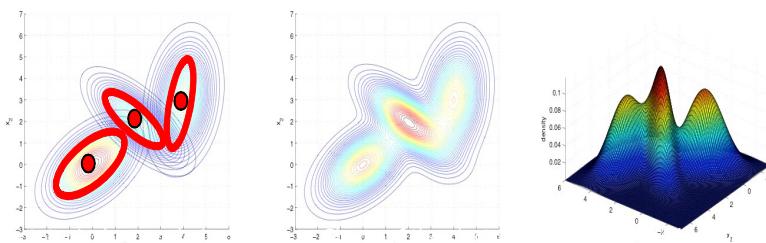
Density based techniques: Kernel Density Estimator

Remember from last week:
Gaussian Mixture Model (GMM)

Data density **Sum of cluster specific densities
assumed normal distributed**

$$p(\mathbf{x}) = \sum_{k=1}^K w_k \mathcal{N}(\mathbf{x} | \boldsymbol{\mu}_{(k)}, \boldsymbol{\Sigma}_{(k)})$$

$$(s.t. \sum_{k=1}^K w_k = 1, \quad w_k \geq 0)$$



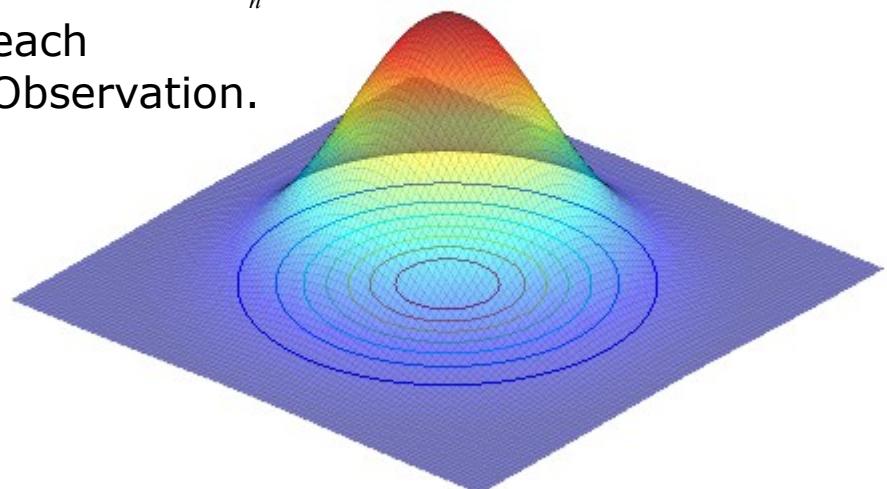
$\boldsymbol{\mu}_{(k)}$: Cluster center (prototypical example in cluster)

$\boldsymbol{\Sigma}_{(k)}$: Shape of the cluster

w_k : Relative density of the cluster

Kernel Density estimation based on Gaussian Kernel:

Consider the GMM and define a Gaussian with mean \mathbf{x}_n and co-variance $\sigma^2 \mathbf{I}$ around each Observation.



Let all observation weight the same, i.e.

$$w_n = 1/N$$

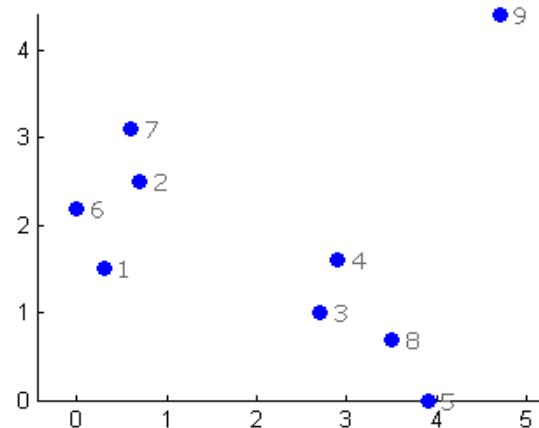
$$p(\mathbf{x}) = \sum_{n=1}^N \frac{1}{N} \mathcal{N}(\mathbf{x} | \mathbf{x}_n, \sigma^2 \mathbf{I})$$

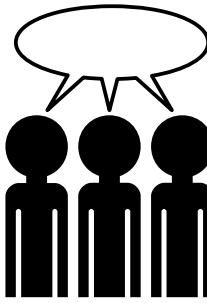
Only free parameter σ !



How do we determine σ ?

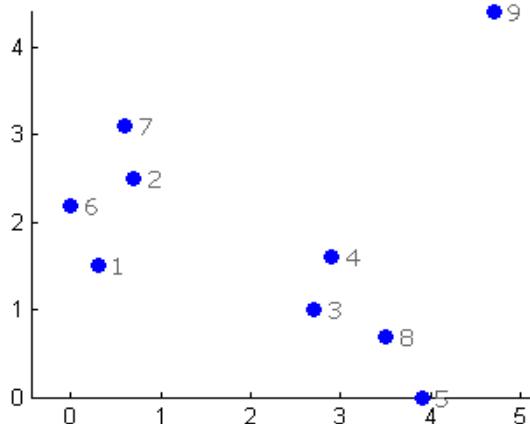
Data I: Cats, Dogs and Dinosaurs



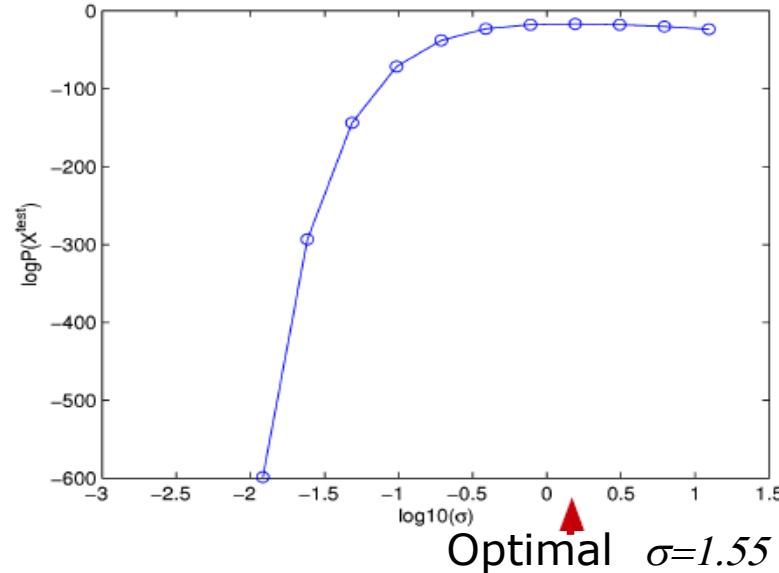


How do we determine σ ?

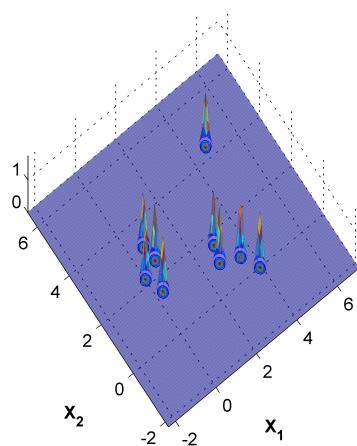
Data I: Cats, Dogs and Dinosaurs



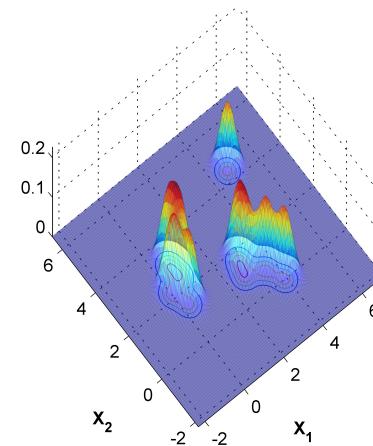
Density of test set based on leave-one-out cross validation



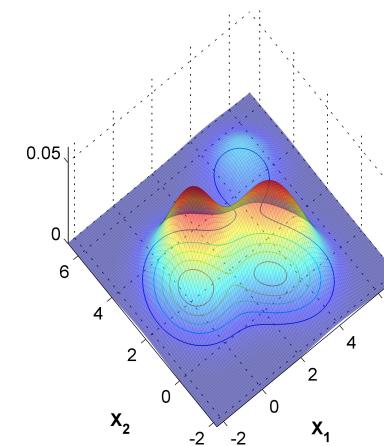
$\sigma=0.01$
Kernel Density width=0.01



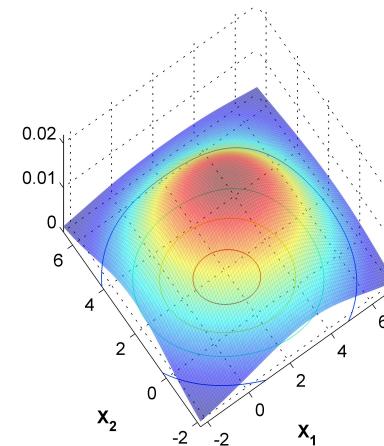
$\sigma=0.1$
Kernel Density width=0.1



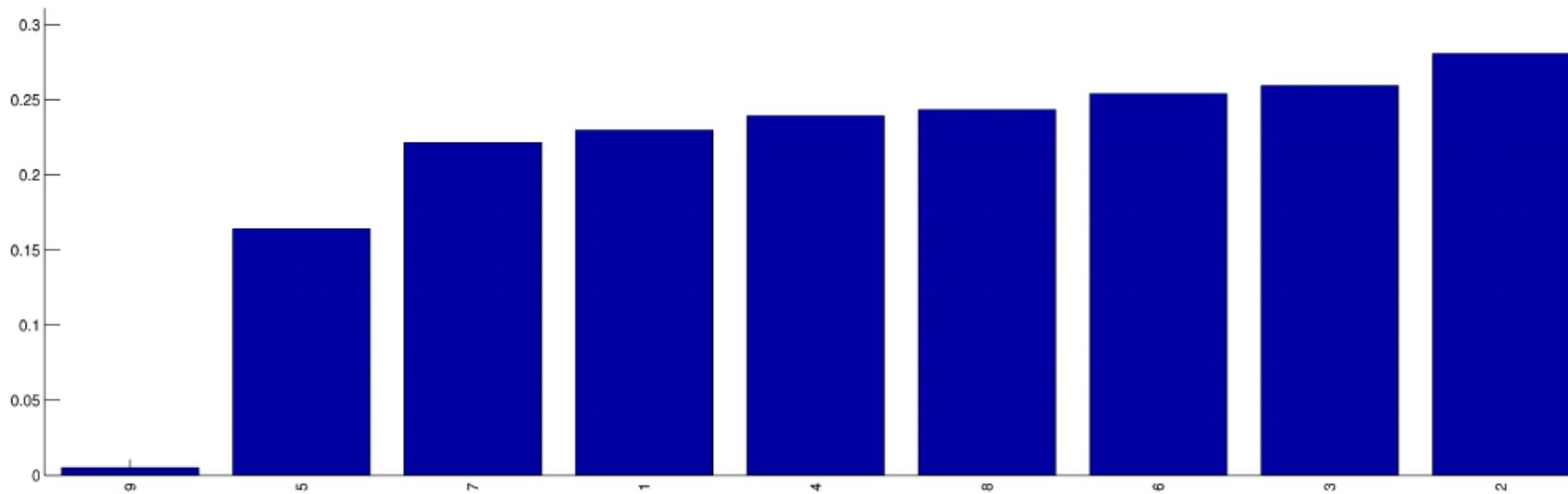
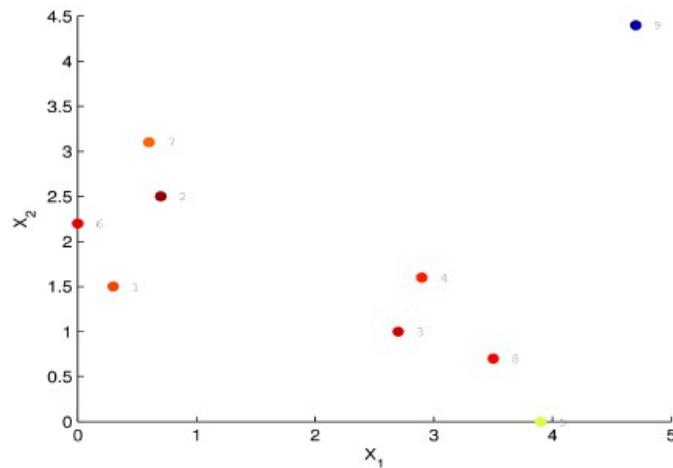
$\sigma=1$
Kernel Density width=1



$\sigma=5$
Kernel Density width=5



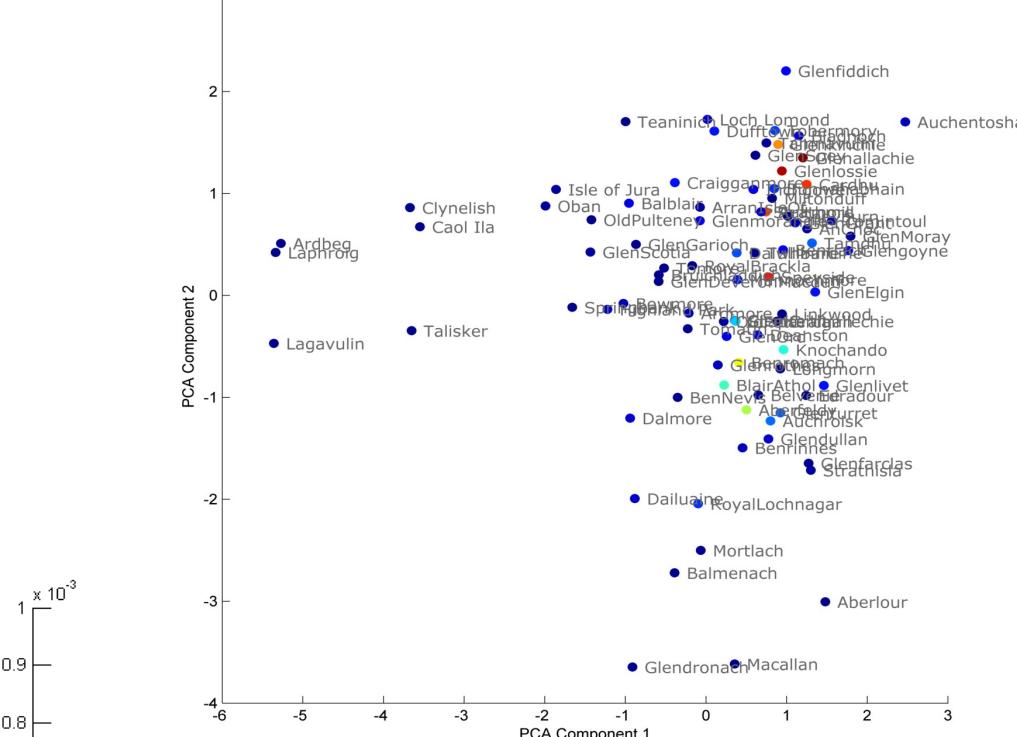
Estimated leave-one-out density evaluated at each observation



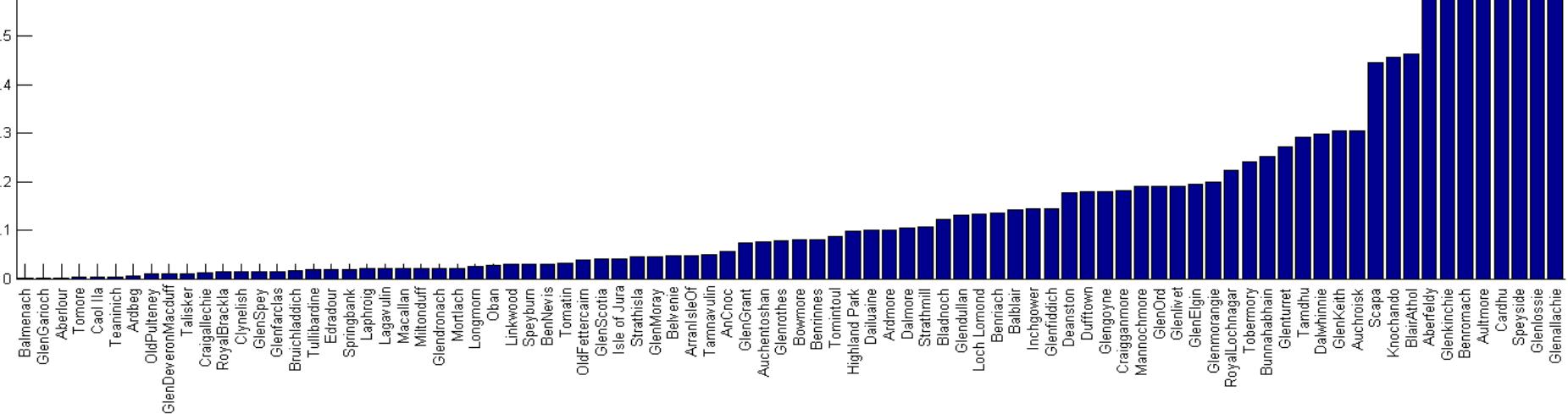
Estimated density evaluated at each observation



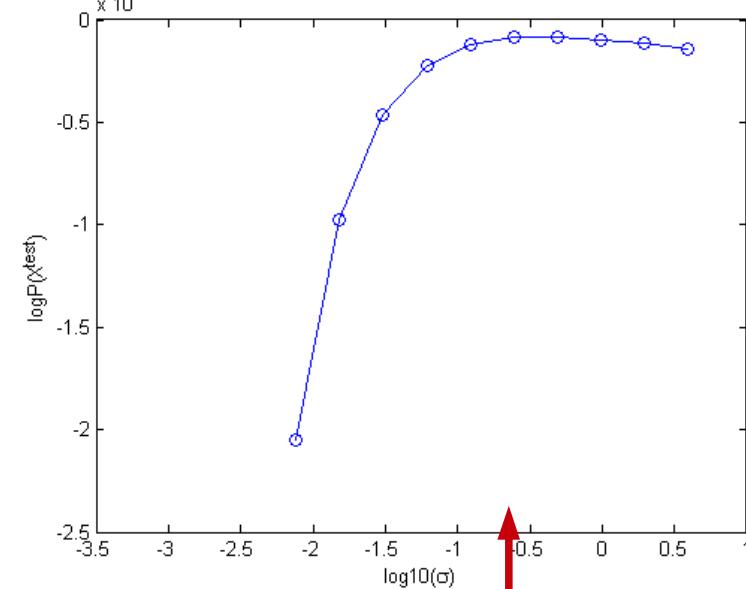
Data II: Whisky



leave-one-out densities



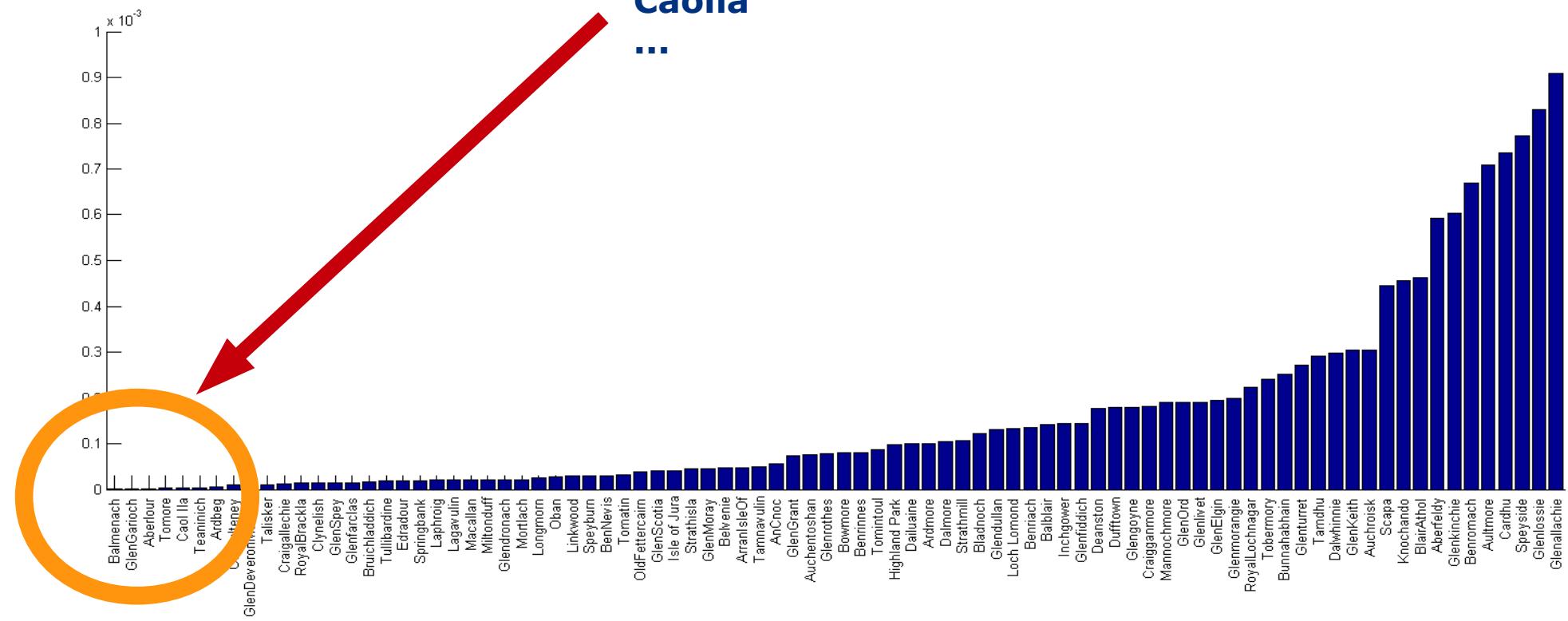
Density of test set based on
leave-one-out cross validation



Optimal $\sigma=0.49$

Data II: Whisky

Balmenach
Glen Garioch
Aberlour
Tomore
Caolla
...



Proximity-based techniques

- Mahanalobis distance to center of data
- Distance to K^{th} nearest neighbour
- Inverse average distance to KNN

Mahalanobis distance

- Distance to the mean of data, taking covariance into account

$$d_{\text{mahalanobis}}(\mathbf{x}, \bar{\mathbf{x}}) = (\mathbf{x} - \bar{\mathbf{x}})^{\top} \mathbf{S}^{-1} (\mathbf{x} - \bar{\mathbf{x}})$$

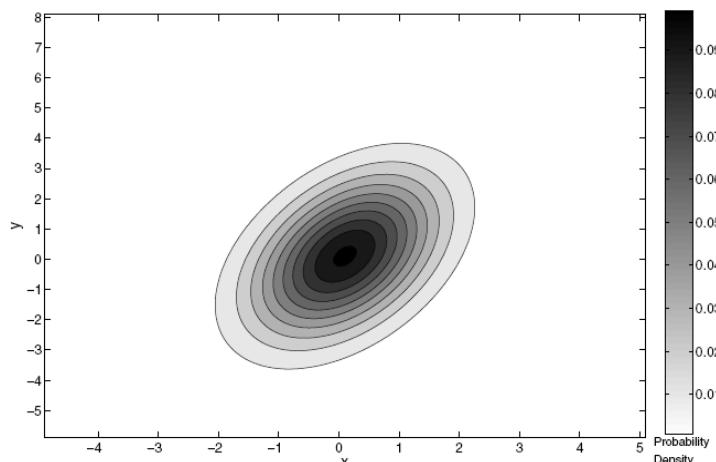


Figure 10.2. Probability density of points for the Gaussian distribution used to generate the points of Figure 10.3.

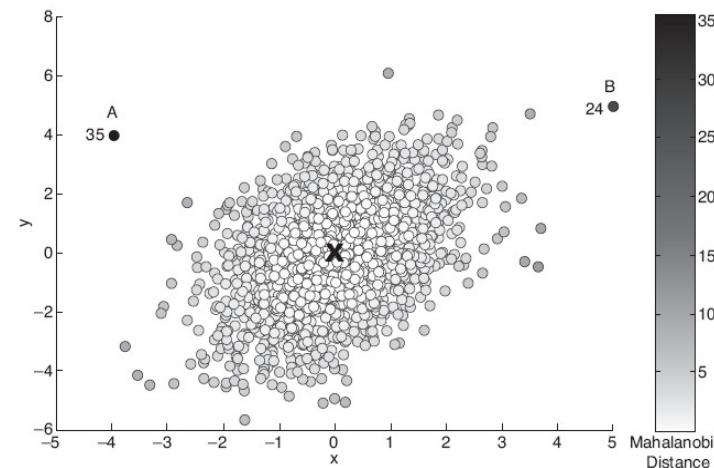


Figure 10.3. Mahalanobis distance of points from the center of a two-dimensional set of 2002 points.

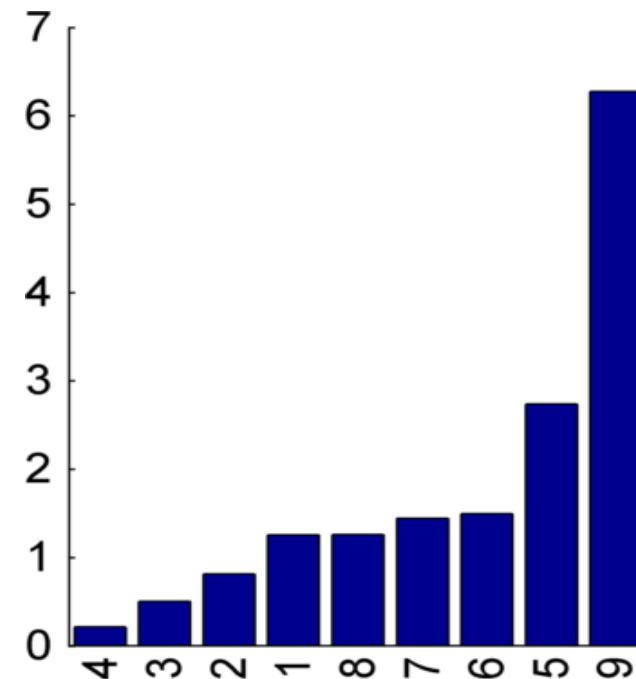
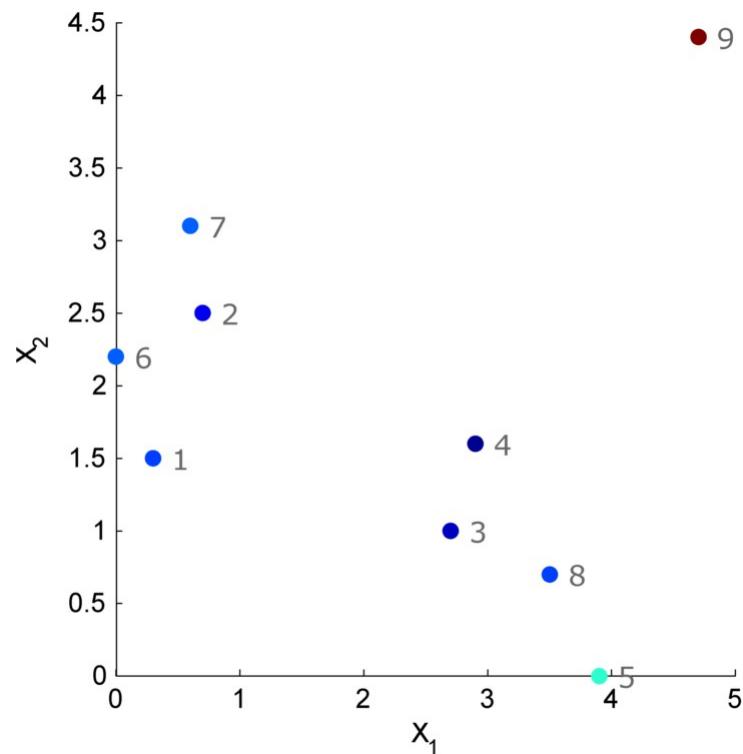
$\bar{\mathbf{x}}$ Mean vector

\mathbf{S} Covariance matrix

Mahalanobis distance

- Distance to the mean of data, taking covariance into account

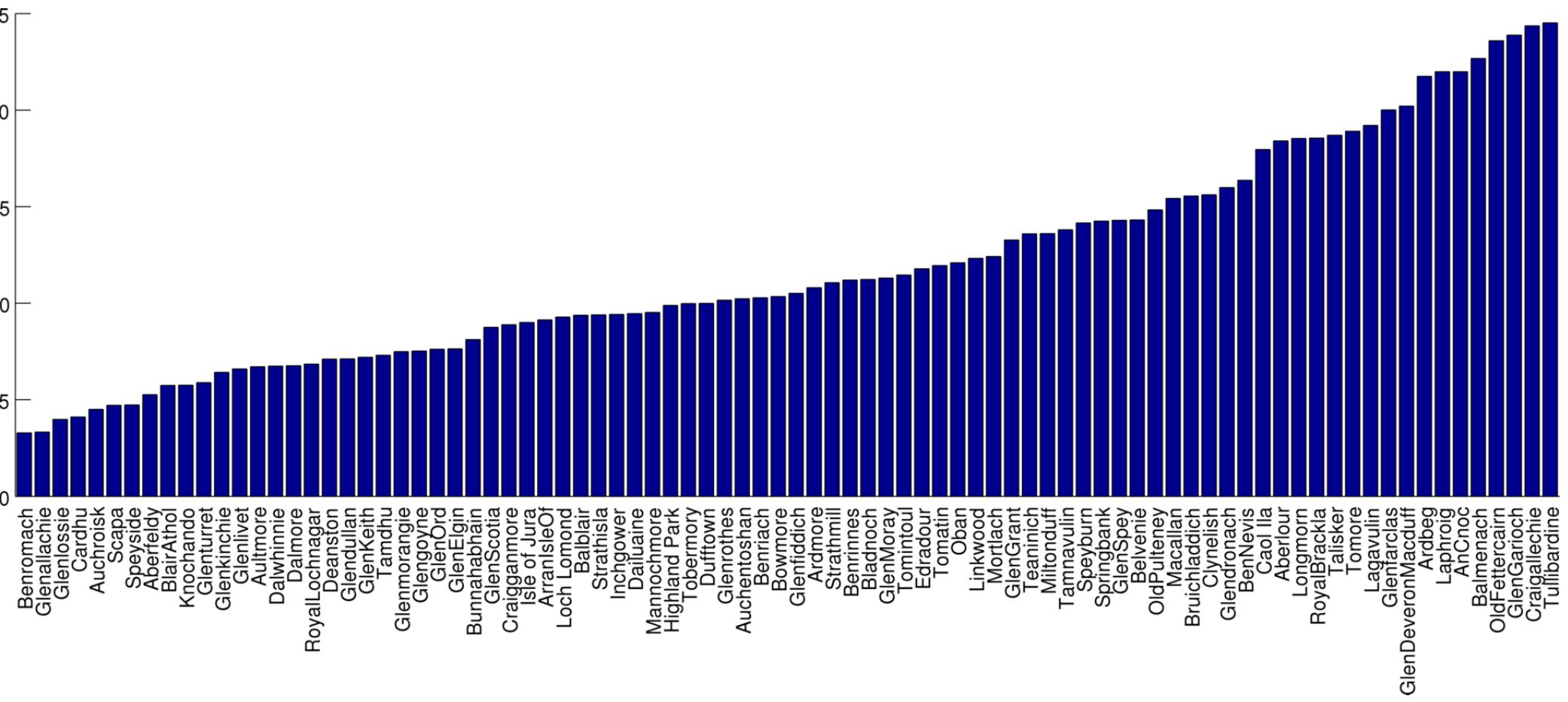
Data I: Cats , dogs and dinosaurs



Mahalanobis distance

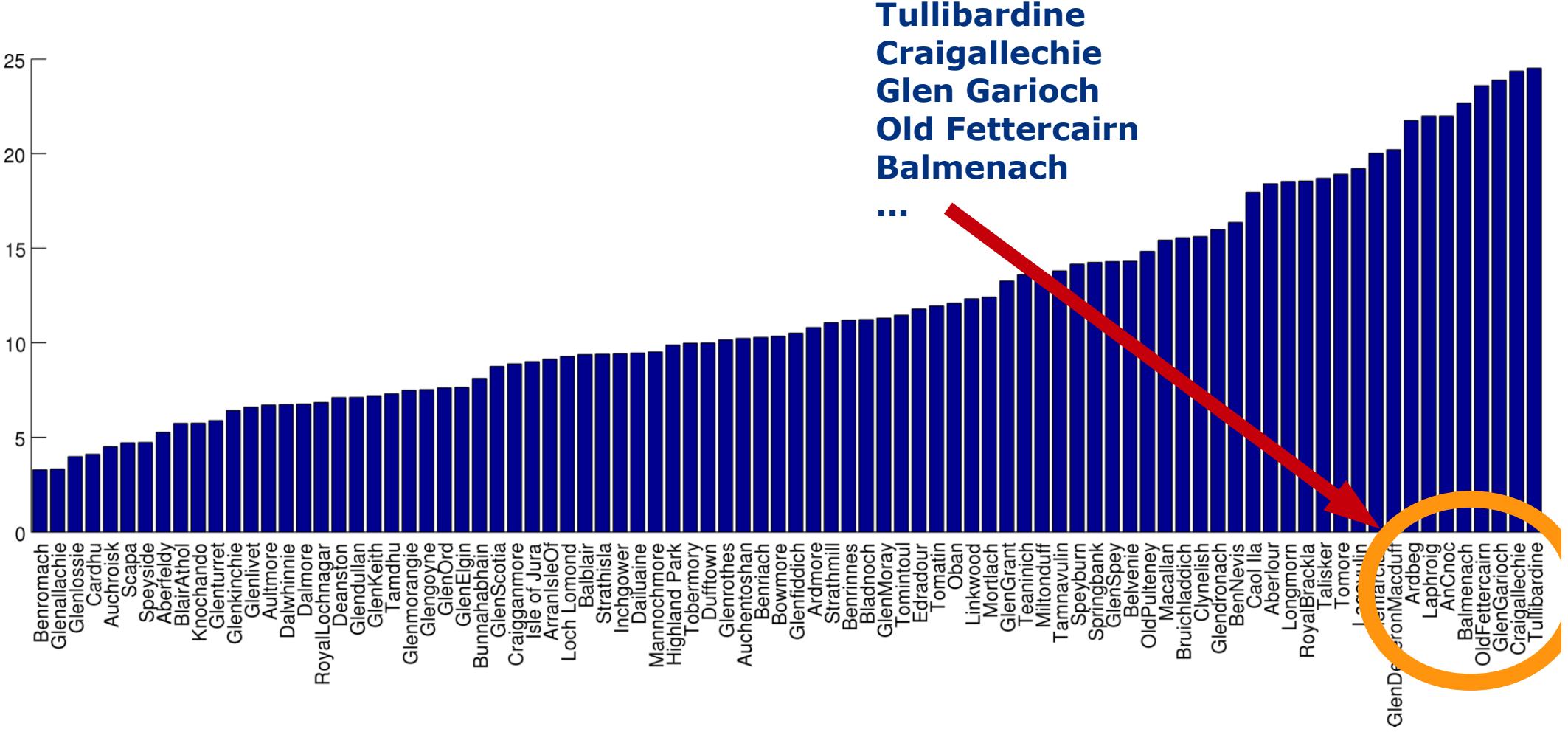
- Distance to the mean of data, taking covariance into account

Data II: Whisky



Mahalanobis distance

- Distance to the mean of data, taking covariance into account



Distance to k-nearest neighbor

- Measure the distance to the k'th nearest neighbor

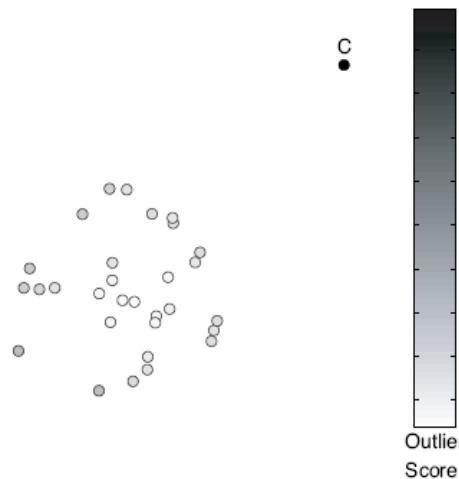


Figure 10.4. Outlier score based on the distance to fifth nearest neighbor.



Figure 10.5. Outlier score based on the distance to the first nearest neighbor. Nearby outliers have low outlier scores.

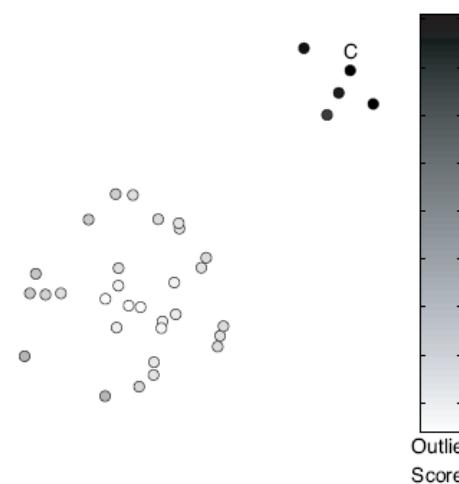


Figure 10.6. Outlier score based on distance to the fifth nearest neighbor. A small cluster becomes an outlier.

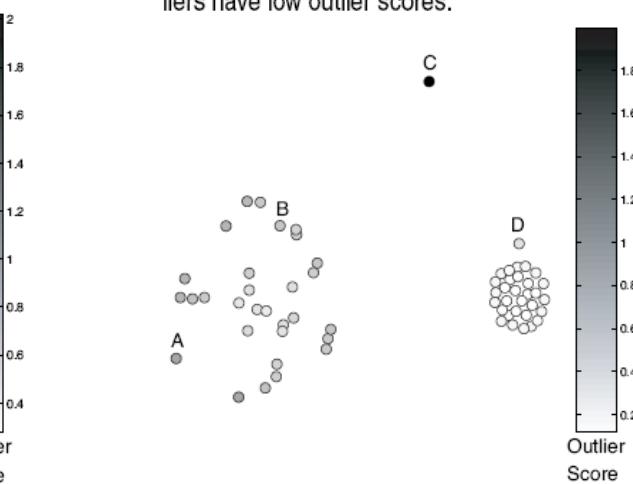
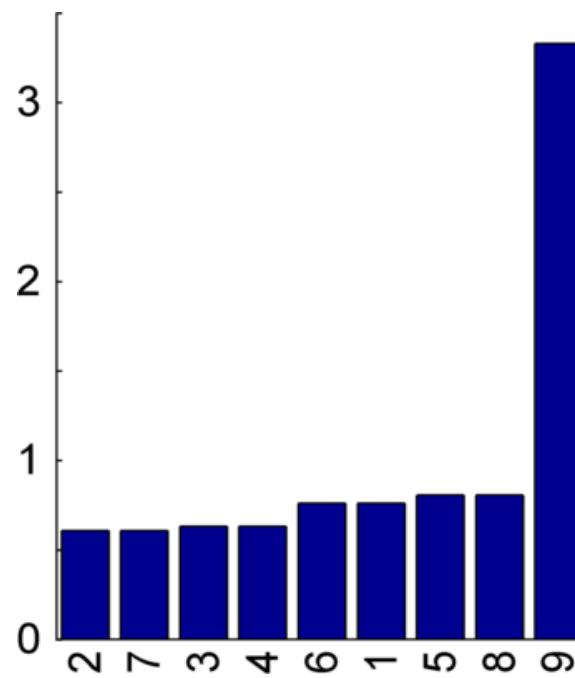
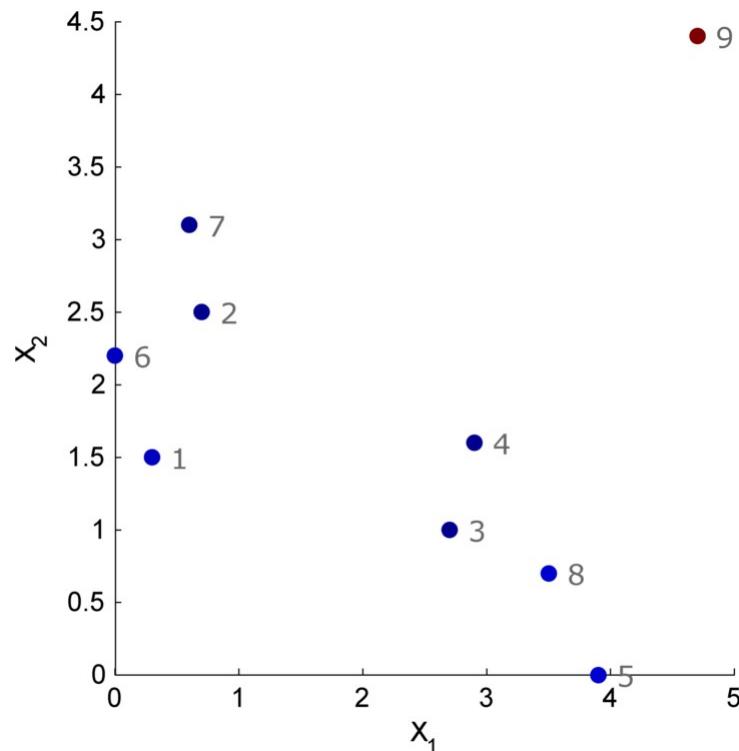


Figure 10.7. Outlier score based on the distance to the fifth nearest neighbor. Clusters of differing density.

Distance to k-nearest neighbor

- Measure the distance to the 1st nearest neighbor

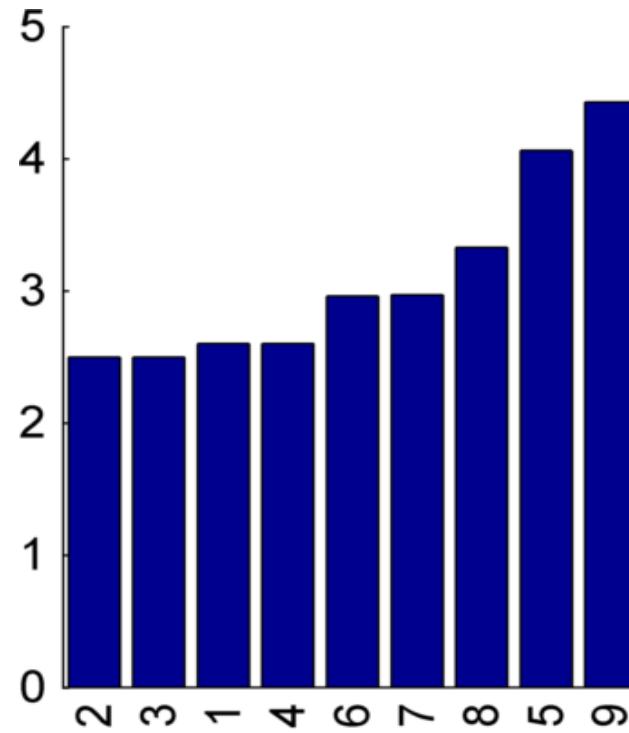
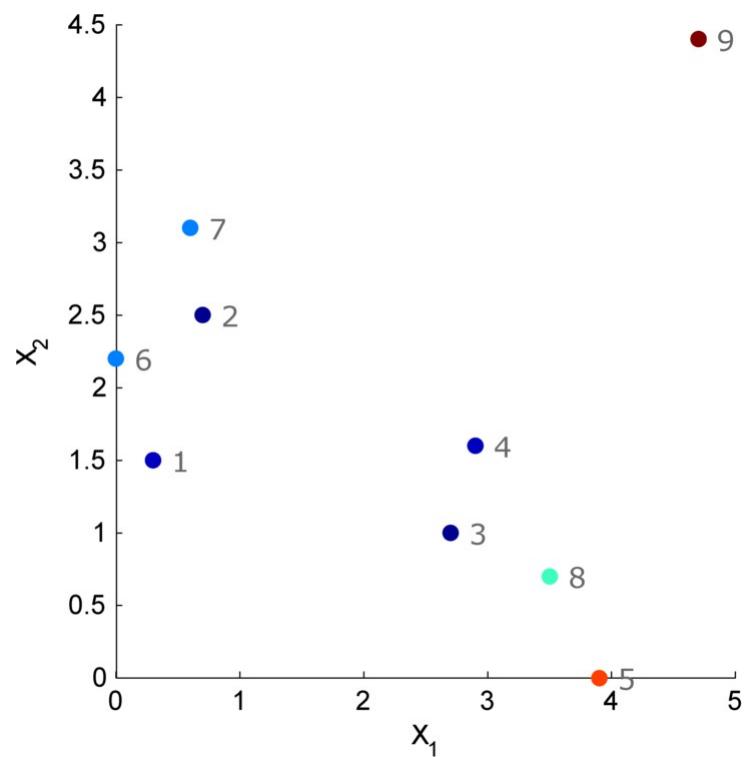
Data I: Cats , dogs and dinosaurs



Distance to kth Nearest neighbour

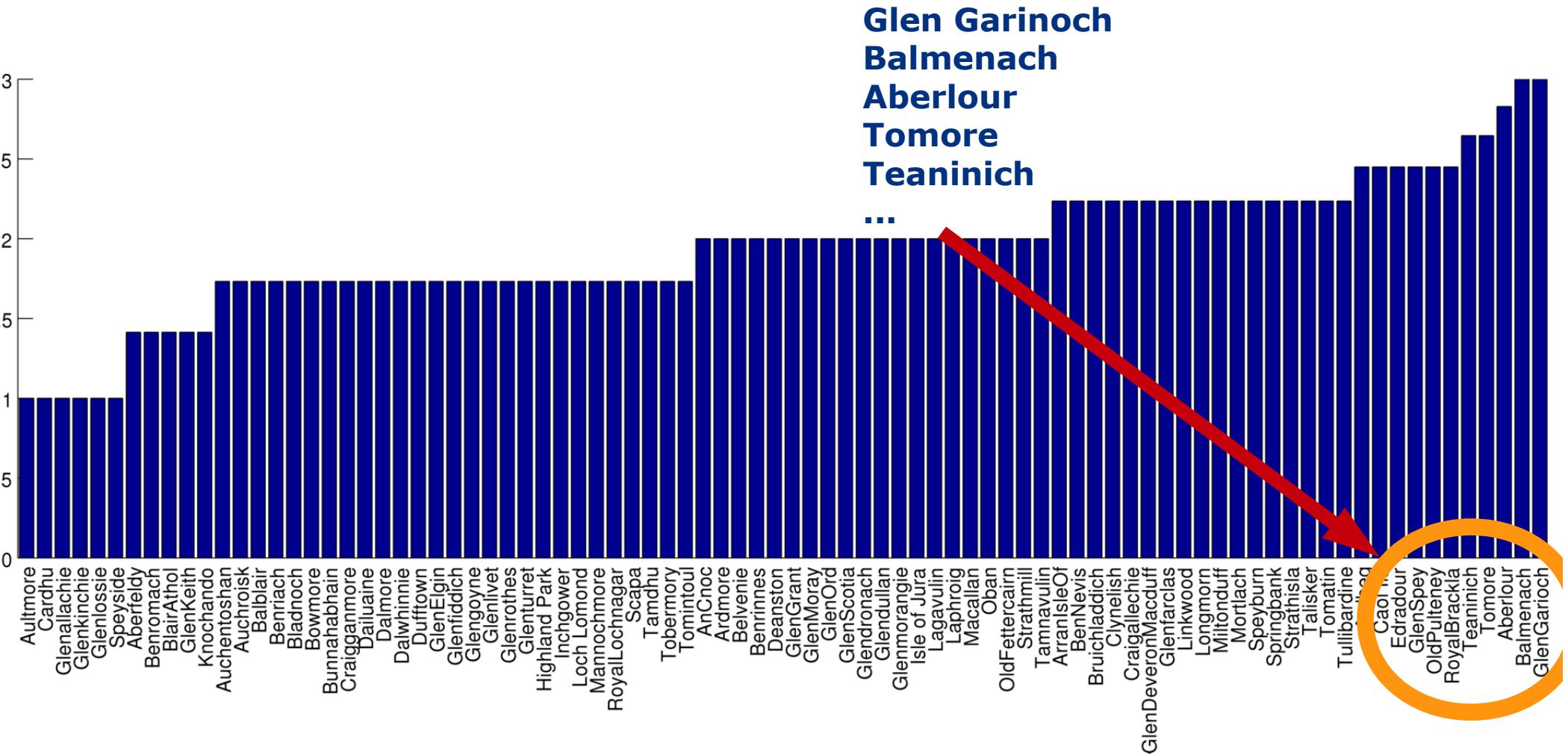
- Measure the distance to the 5th nearest neighbor

Data I: Cats , dogs and dinosaurs



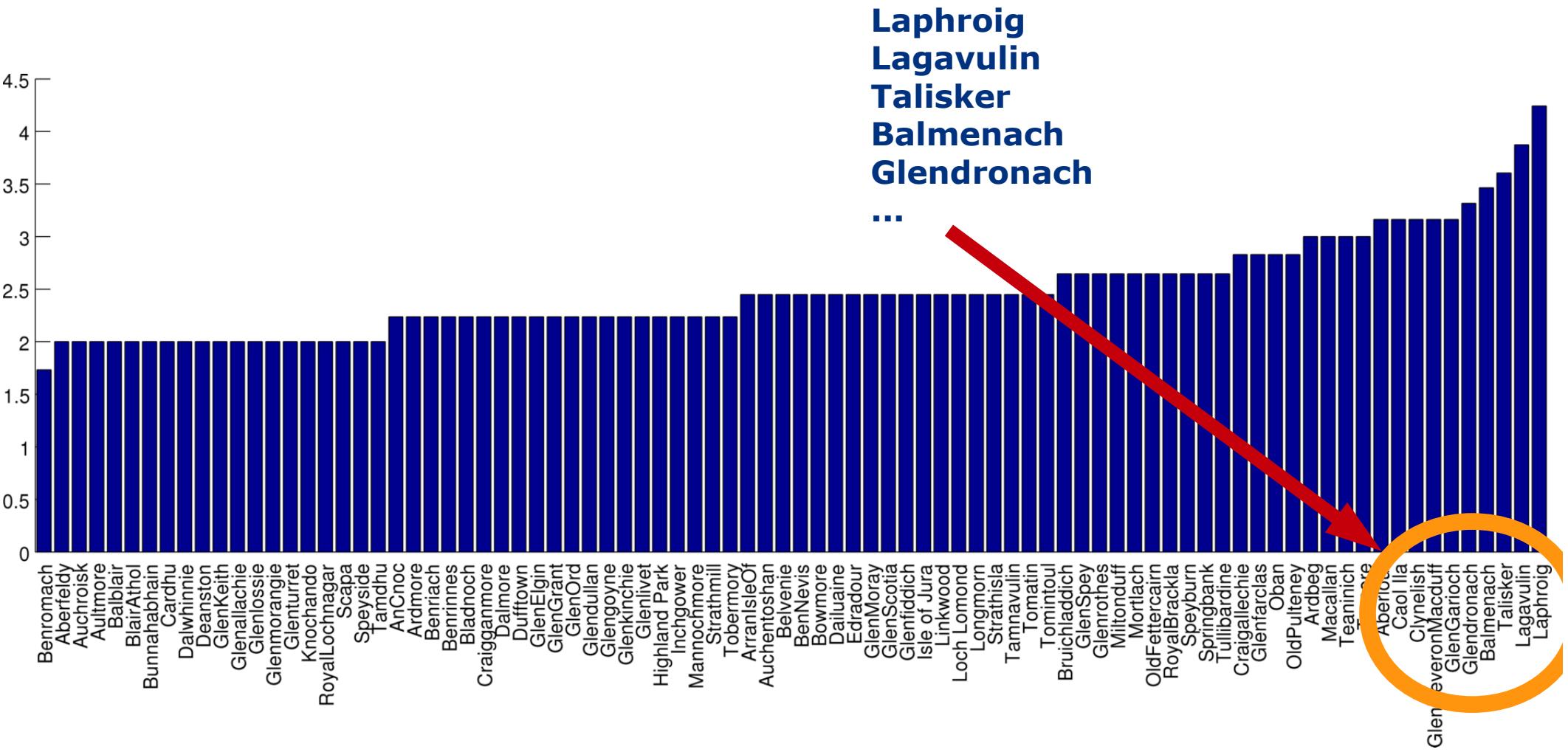
Distance to k-nearest neighbor

- Measure the distance to the 1st nearest neighbor



Distance to k-nearest neighbor

- Measure the distance to the 5th nearest neighbor



Inverse distance density estimation

- **Distance based measure of density**

- Density is inverse proportional to average distance to k nearest neighbors
- Density is low if nearest neighbors are far away

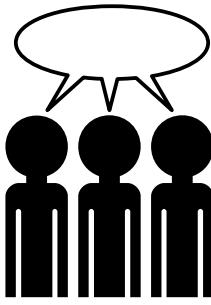
$$\text{density}(\mathbf{x}, k) = \left(\frac{1}{k} \sum_{\mathbf{y} \in N(\mathbf{x}, k)} \text{distance}(\mathbf{x}, \mathbf{y}) \right)^{-1}$$

- **Relative density**

- Density compared to density at nearest neighbors

$$\text{average relative density}(\mathbf{x}, k) = \frac{\text{density}(\mathbf{x}, k)}{\frac{1}{k} \sum_{\mathbf{y} \in N(\mathbf{x}, k)} \text{density}(\mathbf{y}, k)}$$

$N(\mathbf{x}, k)$ The set of k nearest neighbors

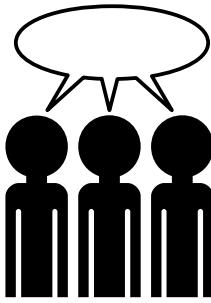


Consider the pairwise distance matrix given to the left. What is the density and average relative density of the first observation for $k=3$?

$d(\mathbf{x}_i, \mathbf{x}_j)$	\mathbf{x}_1	\mathbf{x}_2	\mathbf{x}_3	\mathbf{x}_4	\mathbf{x}_5	\mathbf{x}_6	\mathbf{x}_7
\mathbf{x}_1	0	2.5	2.4	4.0	0.8	0.6	3.3
\mathbf{x}_2	2.5	0	0.6	1.6	2.9	3.0	1.1
\mathbf{x}_3	2.4	0.6	0	1.9	3.0	2.7	1.0
\mathbf{x}_4	4.0	1.6	1.9	0	4.5	4.6	3.8
\mathbf{x}_5	0.8	2.9	3.0	4.5	0	1.1	3.9
\mathbf{x}_6	0.6	3.0	2.7	4.6	1.1	0	3.8
\mathbf{x}_7	3.3	1.1	1.0	3.8	3.9	3.8	0

$$\text{density}(\mathbf{x}, k) = \left(\frac{1}{k} \sum_{\mathbf{y} \in N(\mathbf{x}, k)} \text{distance}(\mathbf{x}, \mathbf{y}) \right)^{-1}$$

$$\text{average relative density}(\mathbf{x}, k) = \frac{\text{density}(\mathbf{x}, k)}{\frac{1}{k} \sum_{\mathbf{y} \in N(\mathbf{x}, k)} \text{density}(\mathbf{y}, k)}$$



Consider the pairwise distance matrix given to the left. What is the density and average relative density of the first observation for k=3?



$d(\mathbf{x}_i, \mathbf{x}_j)$	\mathbf{x}_1	\mathbf{x}_2	\mathbf{x}_3	\mathbf{x}_4	\mathbf{x}_5	\mathbf{x}_6	\mathbf{x}_7
\mathbf{x}_1	0	2.5	2.4	4.0	0.8	0.6	3.3
\mathbf{x}_2	2.5	0	0.6	1.6	2.9	3.0	1.1
\mathbf{x}_3	2.4	0.6	0	1.9	3.0	2.7	1.0
\mathbf{x}_4	4.0	1.6	1.9	0	4.5	4.6	3.8
\mathbf{x}_5	0.8	2.9	3.0	4.5	0	1.1	3.9
\mathbf{x}_6	0.6	3.0	2.7	4.6	1.1	0	3.8
\mathbf{x}_7	3.3	1.1	1.0	3.8	3.9	3.8	0

$$\text{density}(\mathbf{x}, k) = \left(\frac{1}{k} \sum_{\mathbf{y} \in N(\mathbf{x}, k)} \text{distance}(\mathbf{x}, \mathbf{y}) \right)^{-1}$$

$$\text{average relative density}(\mathbf{x}, k) = \frac{\text{density}(\mathbf{x}, k)}{\frac{1}{k} \sum_{\mathbf{y} \in N(\mathbf{x}, k)} \text{density}(\mathbf{y}, k)}$$

$$\text{density}(\mathbf{x}_1, 3) = [1/3(0.6+0.8+2.4)]^{-1} = 3/3.8$$

$$\text{density}(\mathbf{x}_6, 3) = [1/3(0.6+1.1+2.7)]^{-1} = 3/4.4$$

$$\text{density}(\mathbf{x}_5, 3) = [1/3(0.8+1.1+2.9)]^{-1} = 3/4.8$$

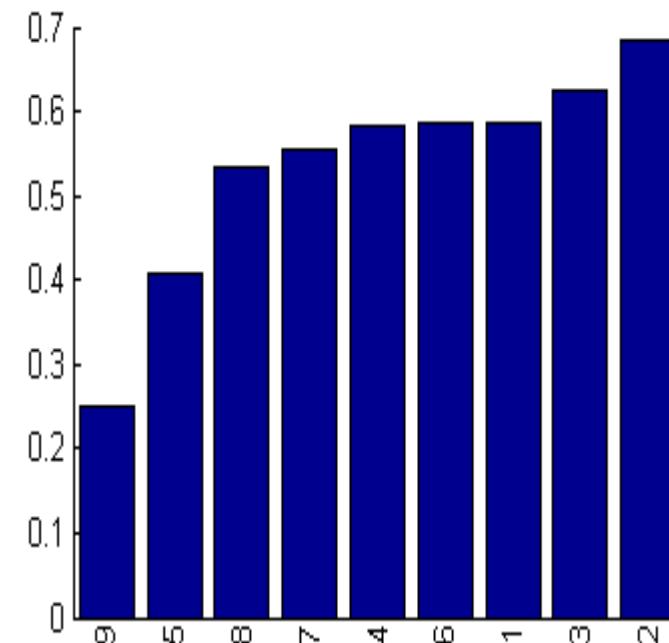
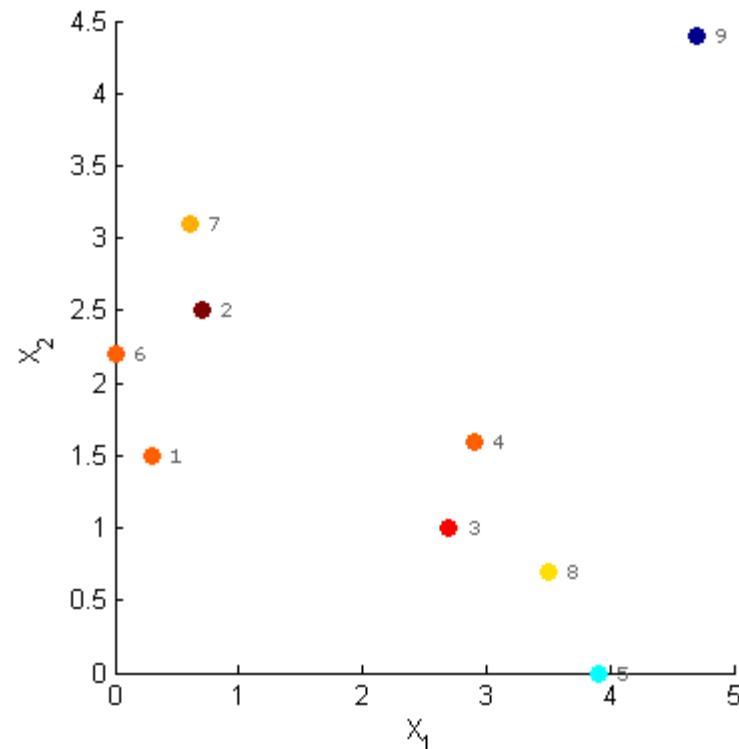
$$\text{density}(\mathbf{x}_3, 3) = [1/3(0.6+1.0+1.9)]^{-1} = 3/3.5$$

$$\text{Av. Rel. Density}(\mathbf{x}_1, 3) = [3/3.8]/[1/3 (3/4.4+3/4.8+3/3.5)] = 1.0945$$

Inverse distance density estimation

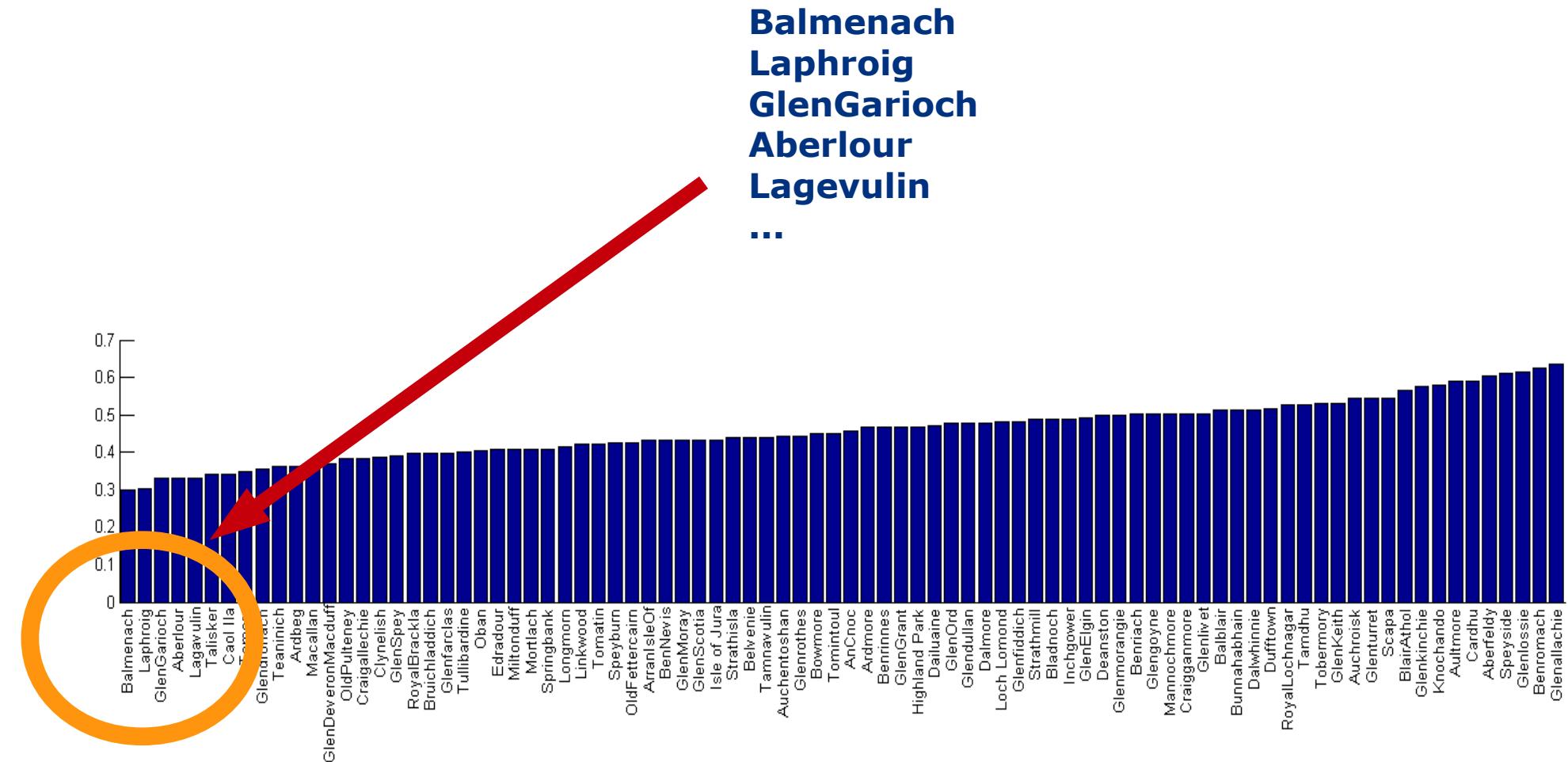
- KNN density (5 nearest neighbors)

Data I: Cats , dogs and dinosaurs



Inverse distance density estimation

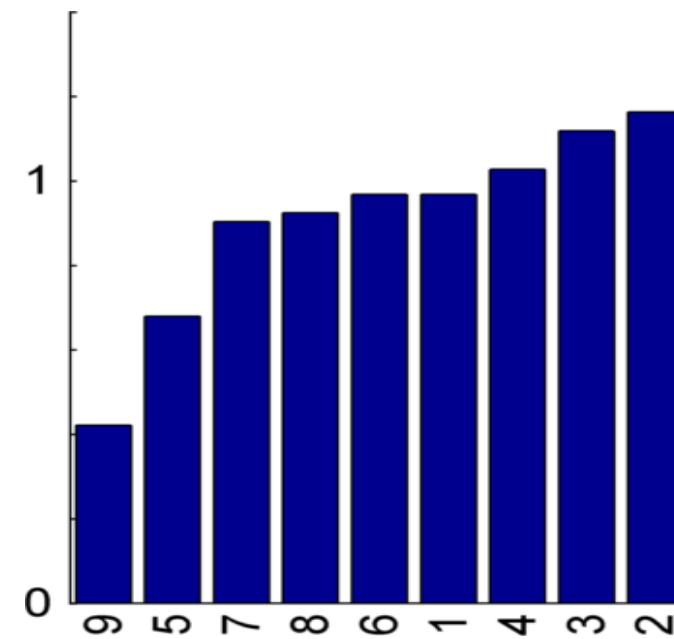
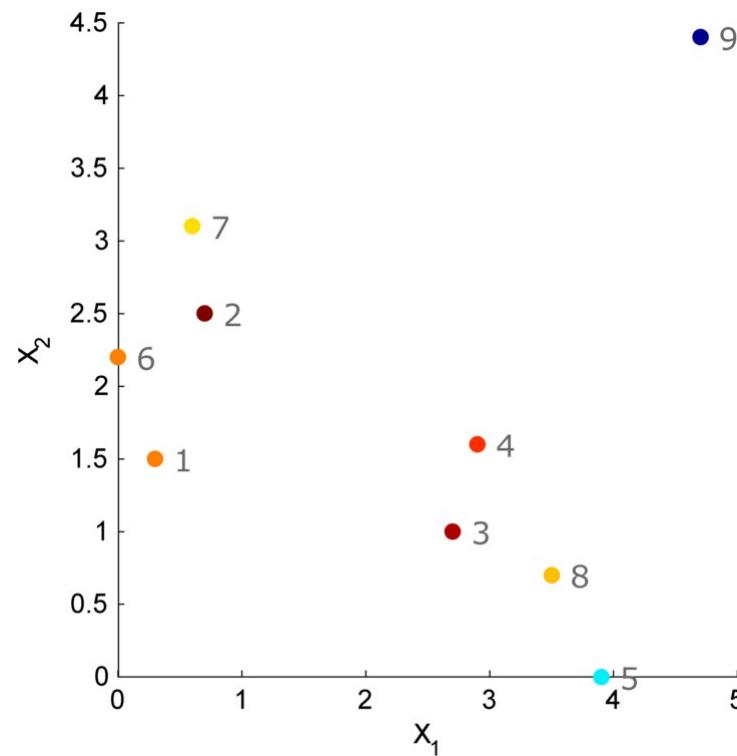
- KNN density (5 nearest neighbors)



Average Relative density

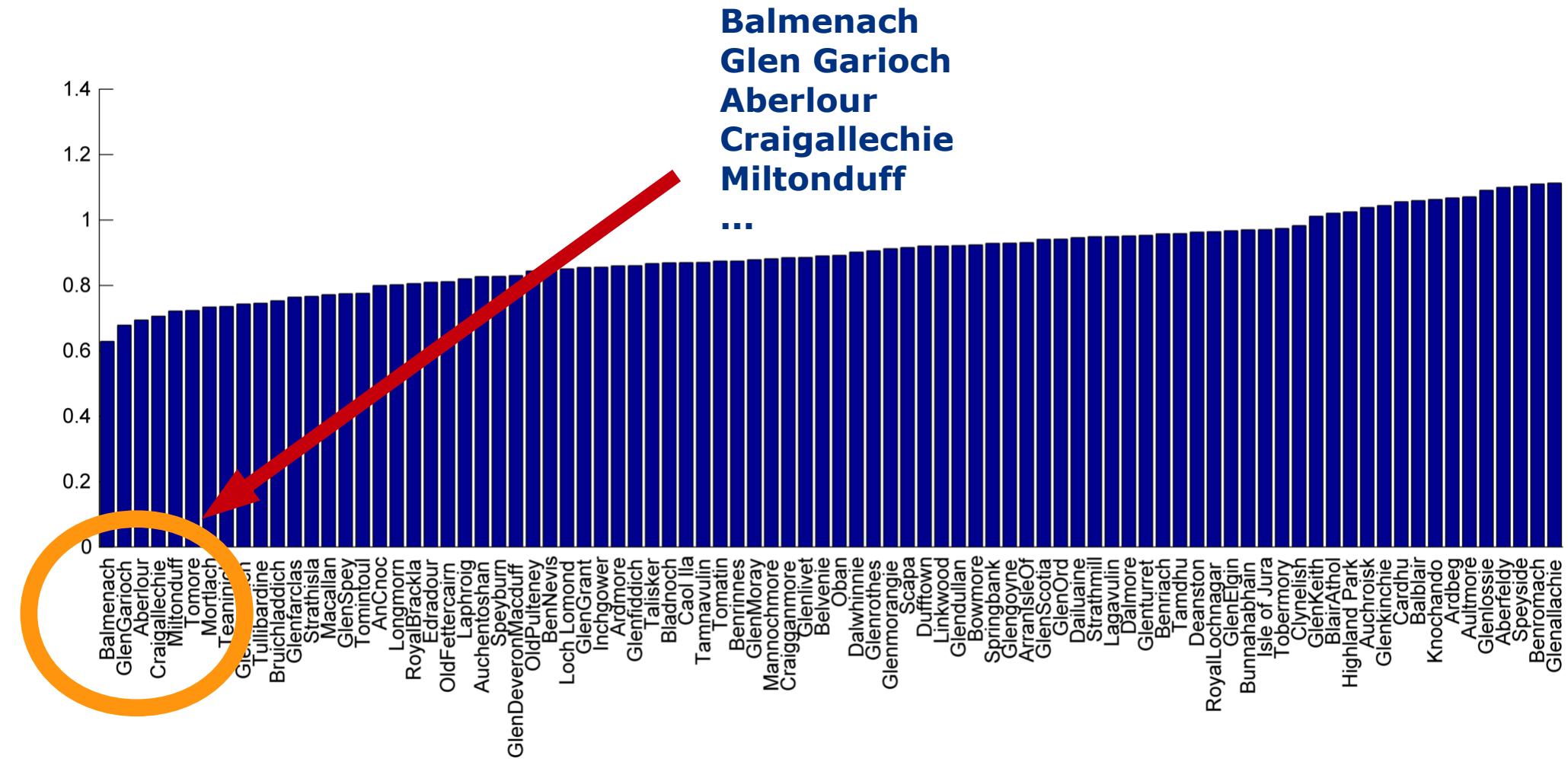
- Average Relative KNN density (5 nearest neighbors)

Data I: Cats , dogs and dinosaurs



Average relative density

- Average relative KNN density (5 nearest neighbors)



Results using different methods

- **Univariate Normal distribution**

- Ardbeg
- Lagavulin
- Laphroig

- **Kernel Density Estimation**

- Balmenach
- Glen Garioch
- Aberlour
- Tomore
- Caolla

- **Mahalanobis distance**

- Tullibardine
- Craigallechie
- Glen Garioch
- Old Fettercairn
- Balmenach

- **Distance to nearest neighbor**

- Glen Garioch
- Balmenach
- Aberlour
- Tomore
- Teaninich

- **Distance to 5th nearest neighbor**

- Laphroig
- Lagavulin
- Talisker
- Balmenach
- Glendronach

- **KNN density**

- Balmenach
- Laphroig
- Glen Garioch
- Aberlour
- Lagavulin

- **KNN average relative density**

- Balmenach
- Glen Garioch
- Aberlour
- Craigallechie
- Miltonduff

Common: Balmenach, Glen Garioch, Laphroig, Aberlour, Tomore, Lagavulin, Craigallechie

Pre-test revisited now as a post test

<http://obsurvey.com/S2.aspx?id=fdca52c2-a77d-4773-907c-4da79793d3b2>

Example of exam questions

$$\text{density}(\mathbf{x}, k) = \left(\frac{1}{k} \sum_{\mathbf{y} \in N(\mathbf{x}, k)} \text{distance}(\mathbf{x}, \mathbf{y}) \right)^{-1}$$

$$\text{average relative density}(\mathbf{x}, k) = \frac{\text{density}(\mathbf{x}, k)}{\frac{1}{k} \sum_{\mathbf{y} \in N(\mathbf{x}, k)} \text{density}(\mathbf{y}, k)}$$

QI: What is the average relative density for observation 2 (i.e. \mathbf{x}_2) for $k=2$ nearest neighbours?

- A: 1/5
- B: 3/10
- C: 7/10
- D: 1

QII: We consider a data set with 10 data objects denoted A–J, that theoretically should lie in one cluster; however, we suspect there might be one or more outliers in the data set. To examine this, we plot the distance to the k 'th nearest neighbor for $k \in \{1, 2, 3\}$ (see Figure to the right). Based on the plot, which statement is *false*?

- A: Observation F and J are both outliers
- B: Observation I is closer to A than to G
- C: Observation G is closer to I than to F
- D: The data has a main cluster of 6 observations and four observations that are far from this cluster.

$d(\mathbf{x}_i, \mathbf{x}_j)$	\mathbf{x}_1	\mathbf{x}_2	\mathbf{x}_3	\mathbf{x}_4	\mathbf{x}_5
\mathbf{x}_1	0	2.0	0.2	0.9	0.2
\mathbf{x}_2	2.0	0	1.5	0.5	2.0
\mathbf{x}_3	0.2	1.5	0	1.2	1.4
\mathbf{x}_4	0.9	0.5	1.2	0	1.0
\mathbf{x}_5	0.2	2.0	1.4	1.0	0

