

Natural Language Processing  
SoSe 2014



## Machine Learning for NLP

*Dr. Mariana Neves*

*April 30th, 2014*

(based on the slides of Dr. Saeedeh Momtazi)

# Introduction

- „Field of study that gives computers the ability to learn without being explicitly programmed“
  - Arthur Samuel, 1959
- Learning Methods
  - Supervised learning
    - Active learning
  - Unsupervised learning
  - Semi-supervised learning
  - Reinforcement learning

# Outline

- Supervised Learning
- Semi-supervised learning
- Unsupervised learning

# Outline

- Supervised Learning
- Semi-supervised learning
- Unsupervised learning

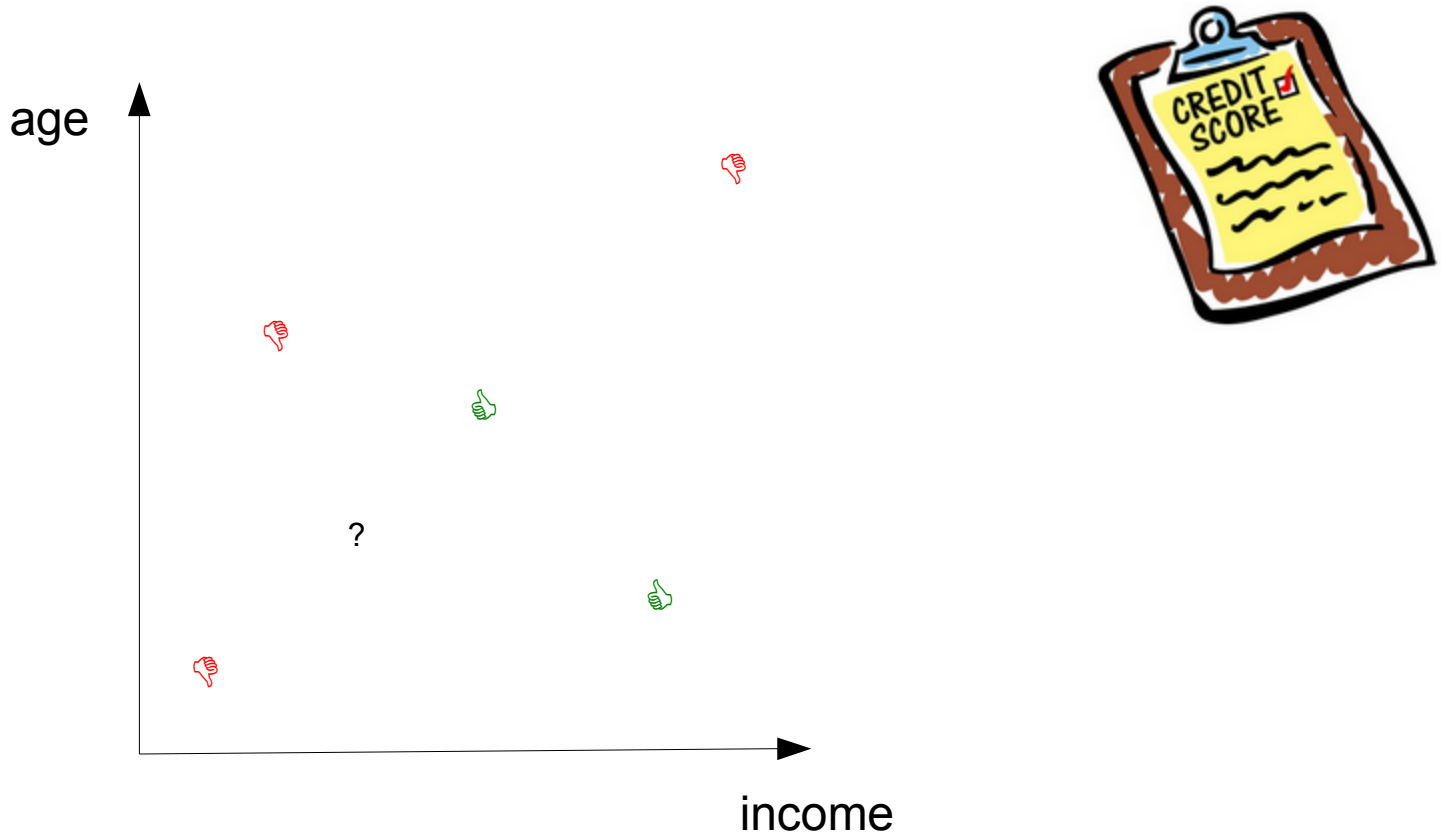
# Supervised Learning

- Example: mortgage credit decision
  - Age
  - Income



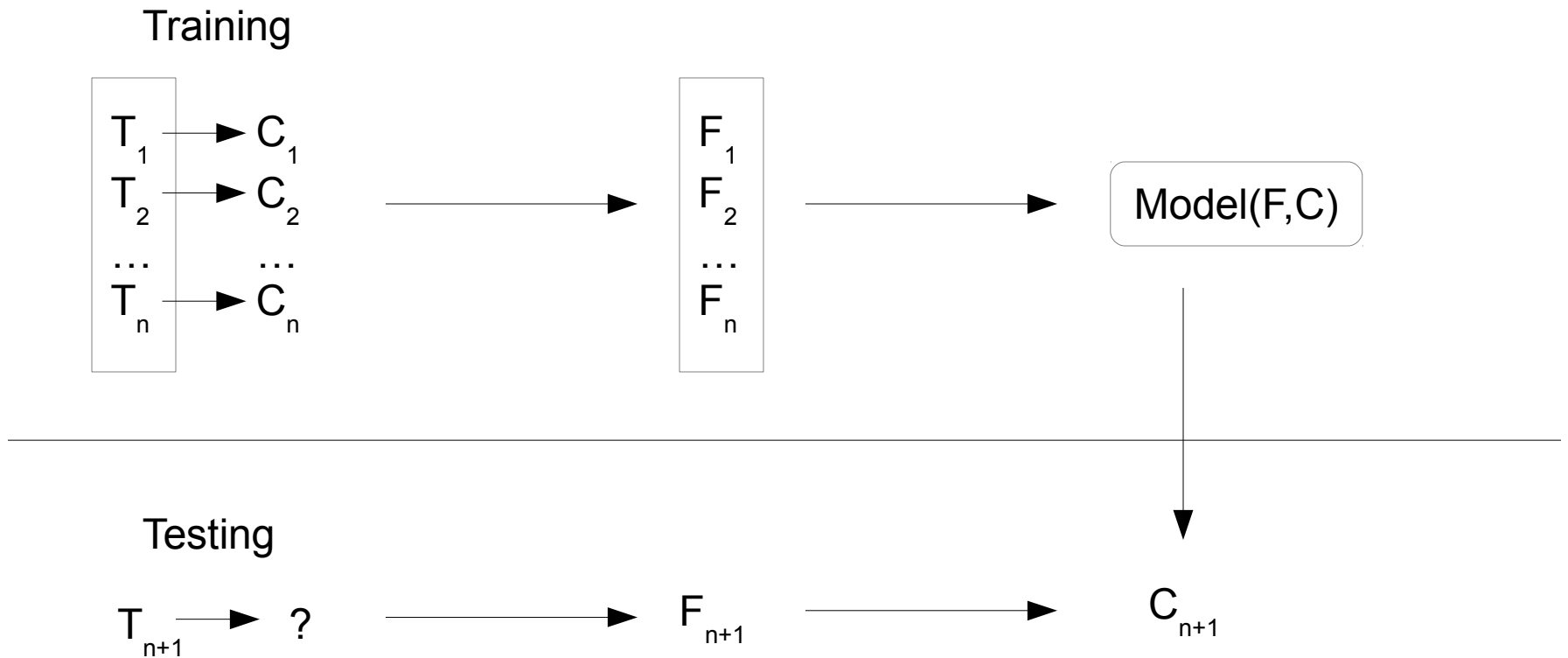
<http://nationalmortgageprofessional.com/news24271/regulatory-compliance-outlook-new-risk-based-pricing-rules>

# Supervised Learning



<http://nationalmortgageprofessional.com/news24271/regulatory-compliance-outlook-new-risk-based-pricing-rules>

# Classification



# Applications

| Problems                  | Items    | Categories            |
|---------------------------|----------|-----------------------|
| POS tagging               | Word     | POS                   |
| Named entity recognition  | Word     | Named entity          |
| Word sense disambiguation | Word     | Word's sense          |
| Spam mail detection       | Document | Spam/Not Spam         |
| Language identification   | Document | Language              |
| Text categorization       | Document | Topic                 |
| Information retrieval     | Document | Relevant/Not relevant |



# Part-of-speech tagging

|   |   |
|---|---|
| 1 | <p>Confidence in the pound is widely expected to take another sharp dive if trade figures for September, due for release tomorrow, fail to show a substantial improvement from July and August 's near-record deficits.</p> |
| 2 | <p>Chancellor of the Exchequer Nigel Lawson 's restated commitment to a firm monetary policy has helped to prevent a freefall in sterling over the past week.</p>   |
| 3 | <p>But analysts reckon underlying support for sterling has been eroded by the chancellor 's failure to announce any new policy measures in his Mansion House speech last Thursday.</p>                                      |
| 4 | <p>This has increased the risk of the government being forced to increase base rates to 16% from their current 15% level to defend the pound, economists and foreign exchange market analysts say.</p>                      |
| 5 | <p>"The risks for sterling of a bad trade figure are very heavily on the down side," said Chris Dillow, senior U.K. economist at Nomura Research Institute.</p>   |
| 6 | <p>"If there is another bad trade number, there could be an awful lot of pressure," noted Simon Briscoe, U.K. economist for Midland Montagu, a unit of Midland Bank PLC.</p>  |

<http://weaver.nlpplab.org/~brat/demo/latest/#/not-editable/CoNLL-00-Chunking/train.txt-doc-1>

# Named entity recognition

|    |   |           |           |           |
|----|---|-----------|-----------|-----------|
| 1  | Characterization of undifferentiated <b>human</b> ES cells and differentiated <b>EBs</b> by antibodies  | cell type | cell type | anat      |
| 2  | All monoclonal antibodies were initially selected for their abilities to recognize recombinant proteins in direct ELISAs.   |           |           |           |
| 3  | A subset were also tested by Western Blot analysis using recombinant proteins and cell lysate to confirm binding to a single epitope.   |           |           |           |
| 4  | The best clone was later screened for its applications for immunocytochemistry and flow cytometry using various cell lines.   |           |           |           |
| 4  | Human peripheral blood platelets were used for screening mouse anti-human CD9 antibody.   | spc       | anatomy   | component |
| 5  | MCF-7 cells were used for screening mouse anti-human E-Cadherin and PODXL (podocalyxin-like) antibodies.  | spc       | spc       | gene      |
| 6  | MG-63 cells were used for screening mouse anti-human GATA1 (GATA binding protein 1) antibody.   | spc       | spc       | gene      |
| 7  | Beta-TC6 cells were used for screening for mouse anti-human/mouse PDX-1 (pancreatic duodenal homeobox-1) antibody.  | spc       | spc       | spc       |
| 8  | NTERA-2 cells were used for screening mouse anti-human Oct3/4 and SOX2 (sex-determining region Y-box 2) antibodies.   | spc       | spc       | gene      |
| 9  | All polyclonal antibodies were affinity-purified using recombinant proteins and validated by direct ELISAs and Western.   |           |           |           |
| 10 | Caco-2 cells were used for validation of goat anti-human GATA6 antibody and NTERA-2 cells were used for validation of goat anti-human Nanog and anti-human Oct3/4 antibodies (Summarized in Table 1). | spc       | gene      | spc       |
|    | Table 1: Summary list of antibody verification by western blot.   | spc       | gene      | spc       |
|    | AntibodySample used for analysisMol.  | spc       | gene      | spc       |

[http://corpora.informatik.hu-berlin.de/index.xhtmll#/cellfinder/version1\\_sections/16316465\\_03\\_results](http://corpora.informatik.hu-berlin.de/index.xhtmll#/cellfinder/version1_sections/16316465_03_results)

# Word sense disambiguation

**Michael Jordan** (born 1963) is an American basketball player.

**Michael Jordan** may also refer to:

- [Michael Jordan \(mycologist\)](#), English mycologist
- [Michael Jordan \(footballer\)](#) (born 1986), English goalkeeper (Arsenal, Chesterfield, Lewes)
- [Michael Jordan \(insolvency baron\)](#) (born 1931), English businessman
- [Mike Jordan](#) (born 1958), English racing driver
- [Mike Jordan \(baseball\)](#) (1863–1940), baseball player
- [Michael Jordan \(Irish politician\)](#), Irish Farmers' Party TD from Wexford, 1927–1932
- [Michael B. Jordan](#) (born 1987), American actor
- [Michael I. Jordan](#) (born 1957), American researcher in machine learning and artificial intelligence
- [Michael H. Jordan](#) (1936–2010), American executive for CBS, PepsiCo, Westinghouse
- [Michael-Hakim Jordan](#) (born 1977), American professional basketball player
- [Michal Jordan](#) (born 1990), Czech ice hockey player
- "Michael Jordan", a song by Kendrick Lamar featuring Schoolboy Q on the album [Overly Dedicated](#)

# Spam mail detection

## Neue Nachricht

Peter Schmidt [noreply@comment.am]

**Sent:** Tuesday, April 29, 2014 10:32 AM

**To:** [Forschungskolleg](#)

Guten Tag,

Sie nutzen derzeit einen Krankenkassen Tarif, der durch einen g?nstigeren ersetzt werden kann.

Damit Sie erfahren welcher Tarif g?nstiger ist und bessere Leistungen bietet, m?ssten Sie einfach nur kurz einen kostenlosen Vergleich auf unserer Internetseite durchf?hren. Dieses dauert weniger als 1 Minute.

Durch einen Wechsel in einen privaten Krankenkassentarif k?nnen Sie derzeit enorm viel sparen. Darum r?t unsere Gesellschaft unbedingt zum Vergleich. Oft sind es ?ber 2.500 Euro die gespart werden k?nnen. Dazu erhalten Sie dann auch noch andere und bessere Leistungen als in Ihrem alten Tarif.

Besuchen Sie unsere Webseite unter:

<http://www.pkv-check2014.com>

Ich hoffe ich konnte Ihnen helfen

Aus Newsletter austragen unter:

<http://www.pkv-check2014.com/unsubscribe>

# Language identification

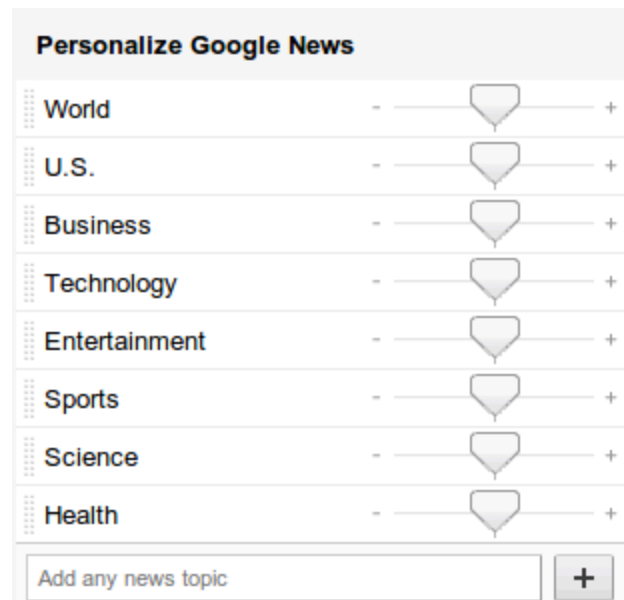
Deutsch Spanish Portugiesisch Sprache erkennen ▾

Va guanyar sis anells de campió de l'NBA amb els Chicago Bulls, on va aconseguir una mitjana de 30,1 punts per partit, la mitjana més gran de la història de la lliga. A més, també va guanyar 10 títols com a màxim anotador, va ser escollit 5 vegades com el MVP de la temporada, 6 com el MVP de les finals, en deu ocasions va formar part del millor quintet de l'NBA, i nou vegades en el millor quintet defensiu; durant tres temporades va ser líder en robatoris de pilotes, i un cop va rebre el premi al millor defensor de la temporada.

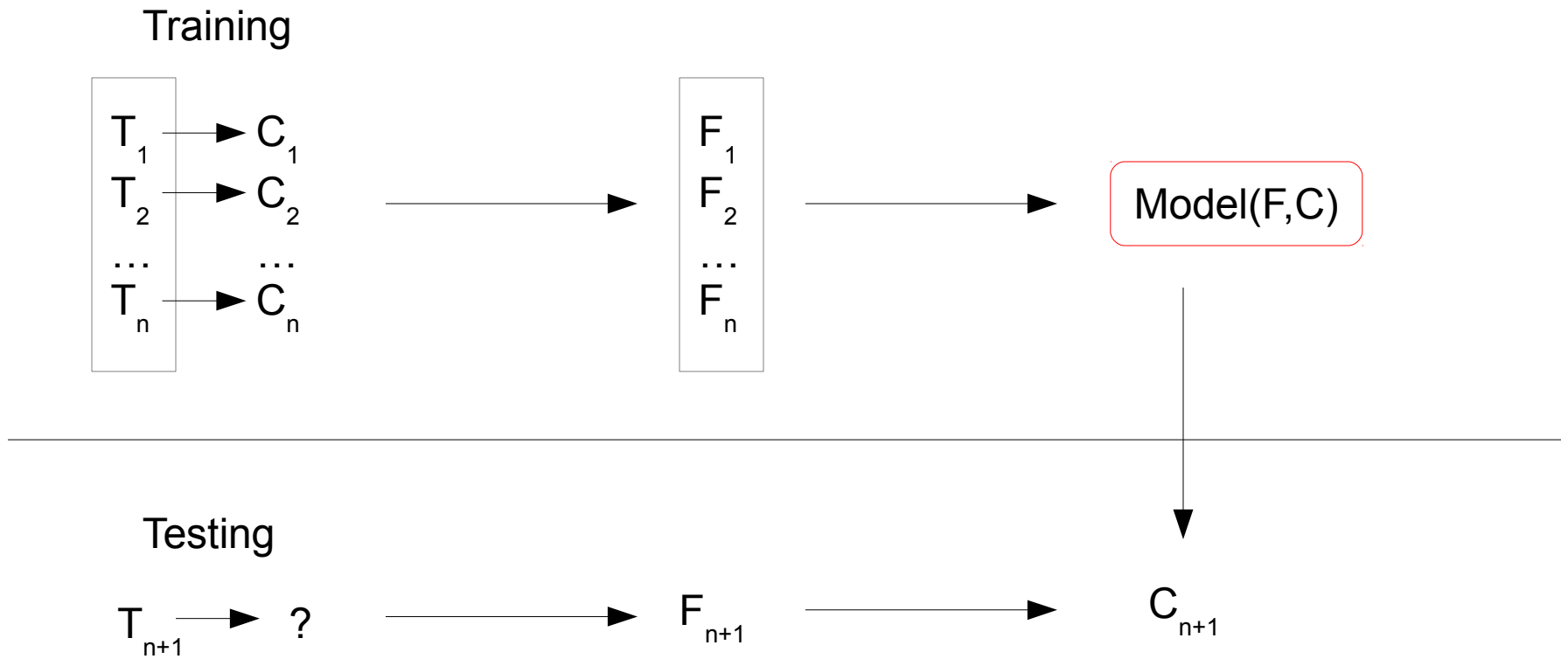
🔊 ⌨ ▾ 🔊 🗨

Ausgangssprache: [Katalanisch](#)

# Text categorization



# Classification



# Classification algorithms

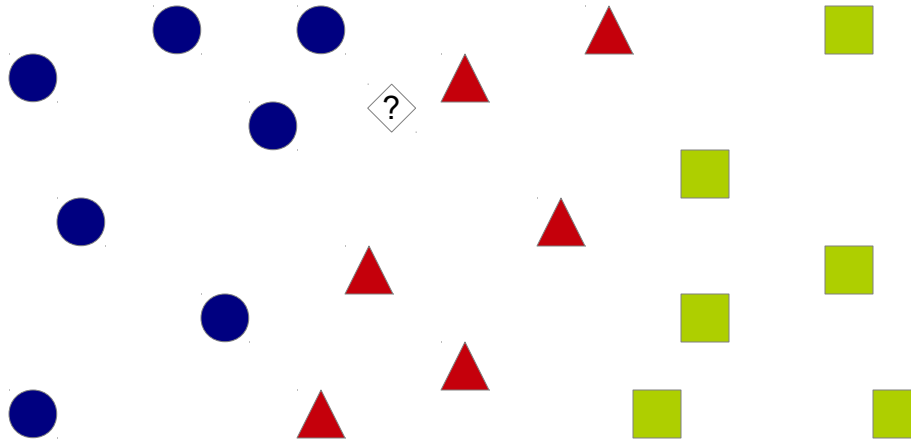
- K Nearest Neighbor
- Support Vector Machines
- Naïve Bayes
- Maximum Entropy
- Linear Regression
- Logistic Regression
- Neural Networks
- Decision Trees
- Boosting
- ...



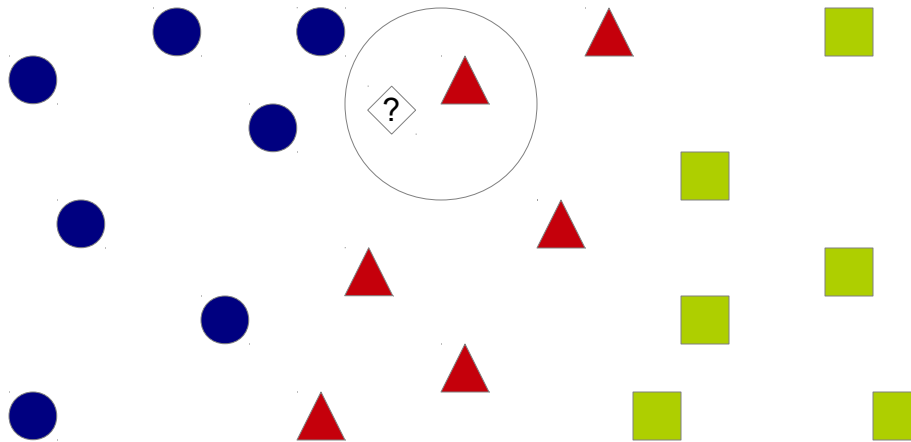
# Classification algorithms

- **K Nearest Neighbor**
- Support Vector Machines
- Naïve Bayes
- Maximum Entropy
- Linear Regression
- Logistic Regression
- Neural Networks
- Decision Trees
- Boosting
- ...

# K Nearest Neighbor

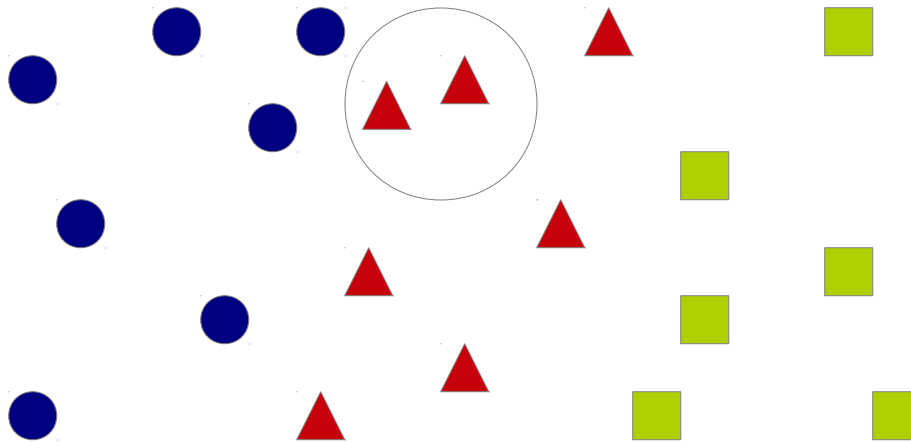


# K Nearest Neighbor



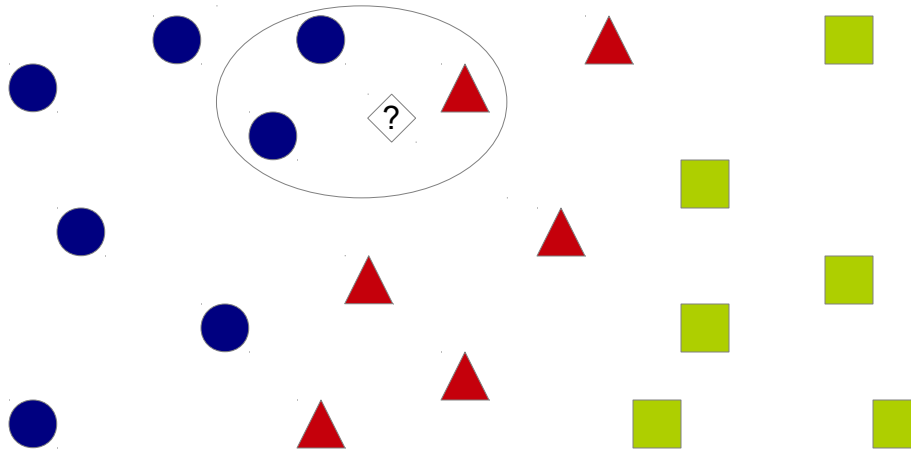
## K Nearest Neighbor

- 1-nearest neighbor



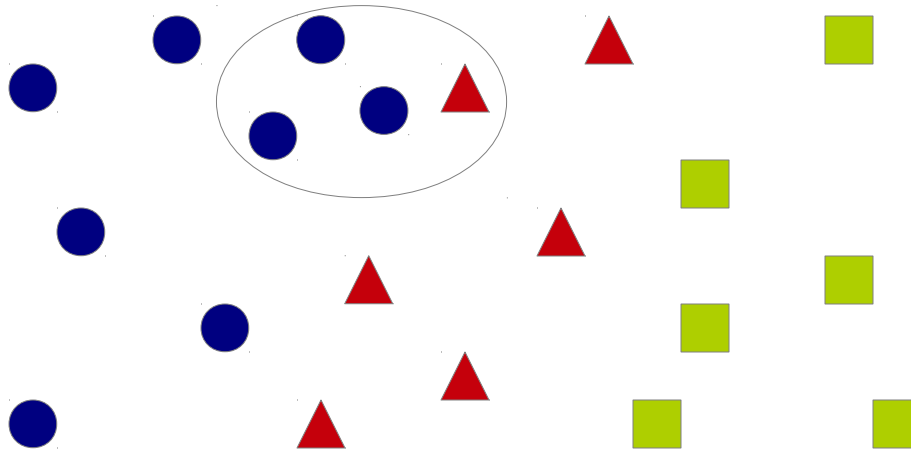
# K Nearest Neighbor

- 3-nearest neighbors



# K Nearest Neighbor

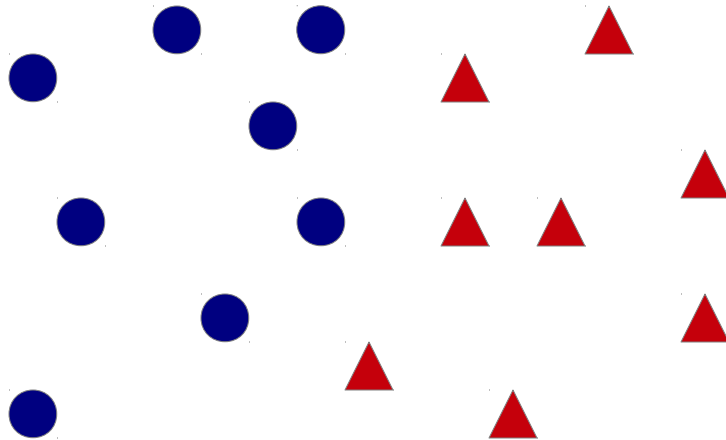
- 3-nearest neighbors



# Classification algorithms

- K Nearest Neighbor
- Support Vector Machines
- Naïve Bayes
- Maximum Entropy
- Linear Regression
- Logistic Regression
- Neural Networks
- Decision Trees
- Boosting
- ...

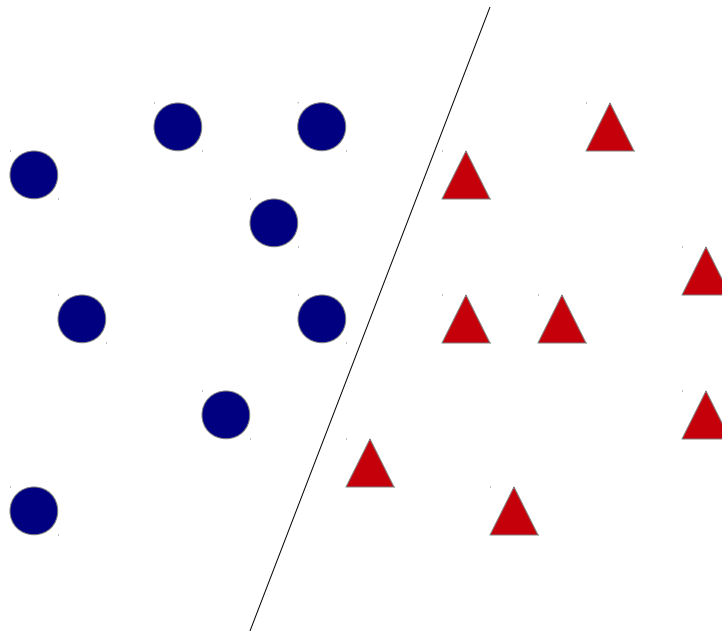
# Support vector machines





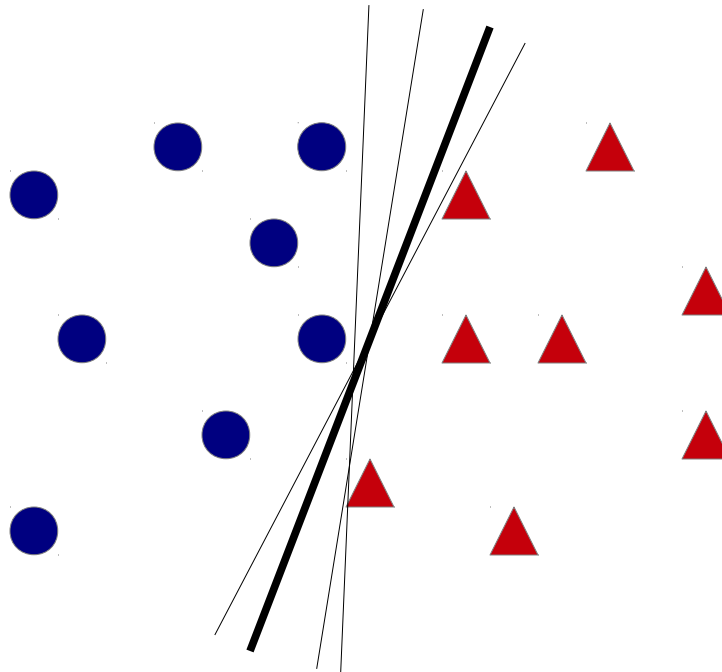
# Support vector machines

- Find a hyperplane in the vector space that separates the items of the two categories



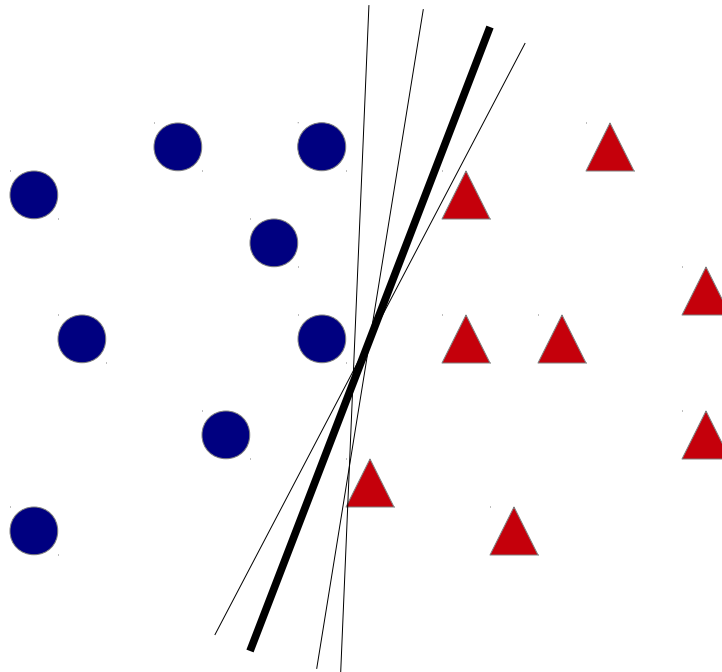
# Support vector machines

- There might be more than one possible separating hyperplane



# Support vector machines

- Find the hyperplane with maximum margin
- Vectors at the margins are called support vectors



# Classification algorithms

- K Nearest Neighbor
- Support Vector Machines
- Naïve Bayes
- Maximum Entropy
- Linear Regression
- Logistic Regression
- Neural Networks
- Decision Trees
- Boosting
- ...

# Naïve Bayes

- Selecting the class with highest probability
  - Minimizing the number of items with wrong labels

$$\hat{c} = \operatorname{argmax}_{c_i} P(c_i)$$

- Probability should depend on the to be classified data (d)

$$P(c_i|d)$$

# Naïve Bayes

$$\hat{c} = \operatorname{argmax}_{c_i} P(c_i)$$

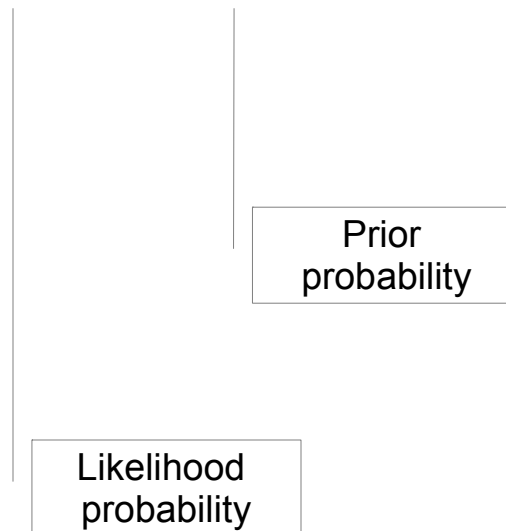
$$\hat{c} = \operatorname{argmax}_{c_i} P(c_i|d)$$

$$\hat{c} = \operatorname{argmax}_{c_i} \frac{P(d|c_i) \cdot P(c_i)}{P(d)}$$

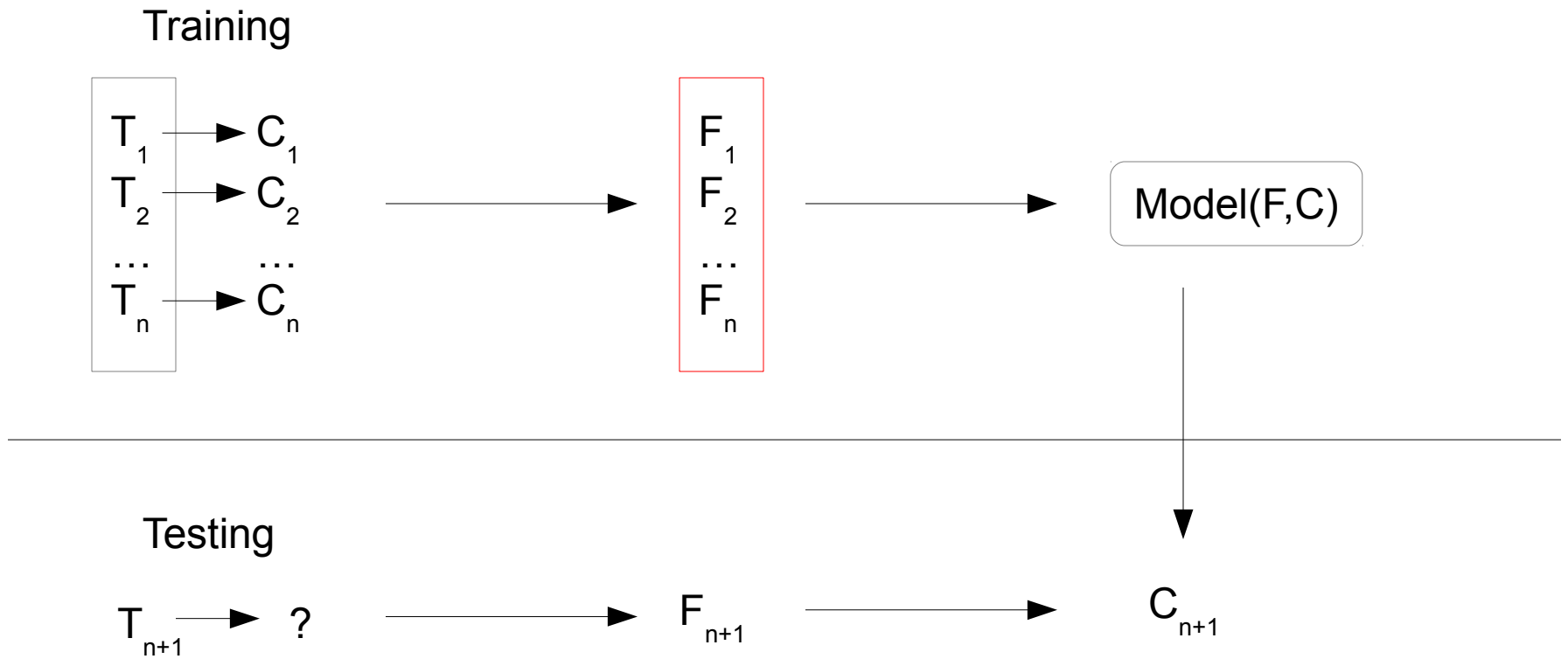
$$\hat{c} = \operatorname{argmax}_{c_i} P(d|c_i) \cdot P(c_i)$$

# Naïve Bayes

$$\hat{c} = \operatorname{argmax}_{c_i} P(d|c_i) \cdot P(c_i)$$



# Classification





# Spam mail detection

**Neue Nachricht**

Peter Schmidt [noreply@comment.am]

Sent: Tuesday, April 29, 2014 10:32 AM

To: Forschungskolleg

Guten Tag,

Sie nutzen derzeit einen Krankenkassen Tarif, der durch einen g?nstigeren ersetzt werden kann.

Damit Sie erfahren welcher Tarif g?nstiger ist und bessere Leistungen bietet, m?sstten Sie einfach nur kurz einen kostenlosen Vergleich auf unserer Internetseite durchf?hren. Dieses dauert weniger als 1 Minute.

Durch einen Wechsel in einen privaten Krankenkassentarif k?nnen Sie derzeit enorm viel sparen. Darum r?t unsere Gesellschaft unbedingt zum Vergleich. Oft sind es ?ber 2.500 Euro die gespart werden k?nnen. Dazu erhalten Sie dann auch noch andere und bessere Leistungen als in Ihrem alten Tarif.

Besuchen Sie unsere Webseite unter:

<http://www.pkv-check2014.com>

Ich hoffe ich konnte Ihnen helfen

Aus Newsletter austragen unter:

<http://www.pkv-check2014.com/unsubscribe>

**Features:**

- words
- sender's email
- contains links
- contains attachments
- contains money amounts
- ...

## Feature selection

- Bag-of-words:
  - Each document can be represented by the set of words that appear in the document
  - Result is a high dimensional feature space
  - The process is computationally expensive
- Solution
  - Using a feature selection method to select informative words

## Feature selection methods

- Information gain
- Mutual information
- $\chi$ -Square

## Information gain

- Measuring the number of bits required for category prediction w.r.t. the presence or absence of a term in the document
- Removing words whose information gain is less than a predefined threshold

$$\begin{aligned} IG(w) = & \sum_{i=1}^K P(c_i) \cdot \log P(c_i) \\ & + P(w) \cdot \sum_{i=1}^K P(c_i|w) \cdot \log P(c_i|w) \\ & + P(\bar{w}) \cdot \sum_{i=1}^K P(c_i|\bar{w}) \cdot \log P(c_i|\bar{w}) \end{aligned}$$

## Information gain

- $N$  = # docs
- $N_i$  = # docs in category  $c_i$
- $N_w$  = # docs containing  $w$
- $N_{\bar{w}}$  = # docs not containing  $w$
- $N_{iw}$  = # docs in category  $c_i$  containing  $w$
- $N_{i\bar{w}}$  = # docs in category  $c_i$  not containing  $w$

$$P(c_i) = \frac{N_i}{N}$$

$$P(w) = \frac{N_w}{N}$$

$$P(c_i|w) = \frac{N_{iw}}{N_i}$$

$$P(\bar{w}) = \frac{N_{\bar{w}}}{N}$$

$$P(c_i|\bar{w}) = \frac{N_{i\bar{w}}}{N_i}$$

## Mutual information

- Measuring the effect of each word in predicting the category
  - How much does its presence or absence in a document contribute to category prediction?

$$MI(w, c_i) = \log \frac{P(w, c_i)}{P(w) \cdot P(c_i)}$$

- Removing words whose mutual information is less than a predefined threshold

$$MI(w) = \max_i MI(w, c_i)$$

$$MI(w) = \sum_i P(c_i) \cdot MI(w, c_i)$$

## $\chi$ -Square

- Measuring the dependencies between words and categories

$$\chi^2(w, c_i) = \frac{N \cdot (N_{iw} N_{i\bar{w}} - N_{i\bar{w}} N_{\bar{i}w})^2}{(N_{iw} + N_{i\bar{w}}) \cdot (N_{\bar{i}w} + N_{i\bar{w}}) \cdot (N_{iw} + N_{\bar{i}w}) \cdot (N_{i\bar{w}} + N_{i\bar{w}})}$$

- Ranking words based on their  $\chi$ -square measure

$$\chi^2(w) = \sum_{i=1}^K P(c_i) \cdot \chi^2(w, c_i)$$

- Selecting the top words as features

## Feature selection

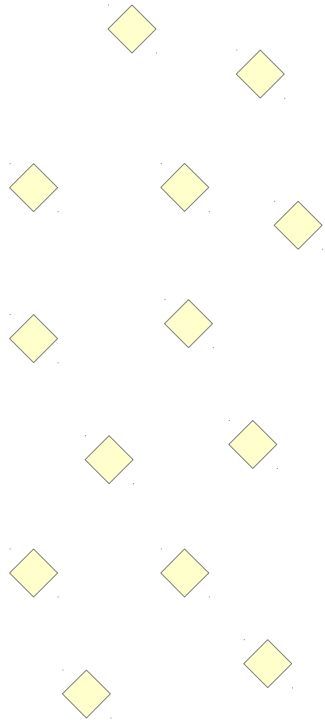
- These models perform well for document-level classification
  - Spam Mail Detection
  - Language Identification
  - Text Categorization
- Word-level Classification might need another types of features
  - Part-of-speech tagging
  - Named Entity Recognition



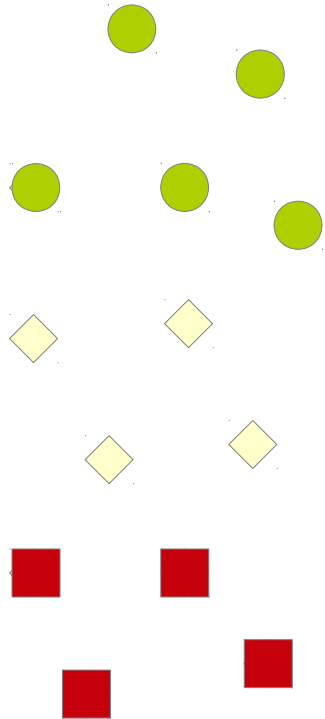
# Supervised learning

- Shortcoming
  - Relies heavily on annotated data
  - Time consuming and expensive task
- Solution
  - **Active learning**
    - Using a minimum amount of annotated data
    - Annotating further data by human, if they are very informative

# Active learning

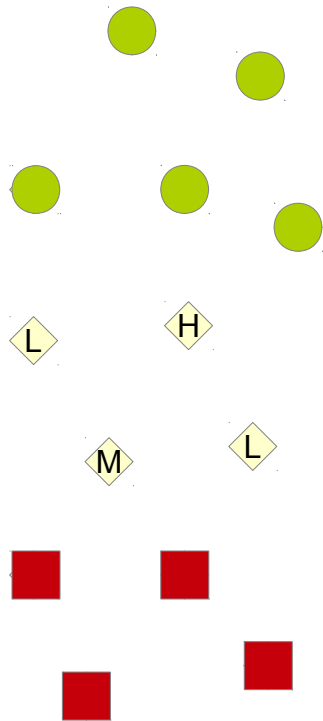


# Active learning



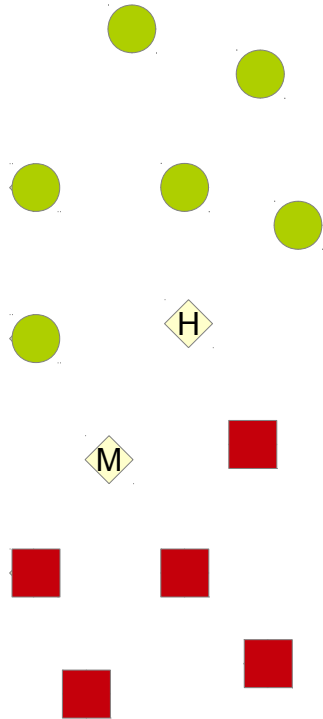
- Annotating a small amount of data

# Active learning



- Calculating the confidence score of the classifier on unlabeled data

# Active learning



- Finding the informative unlabeled data (data with lowest confidence)

- manually annotating the informative data

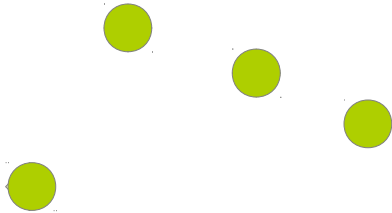
# Outline

- Supervised Learning
- Semi-supervised learning
- Unsupervised learning

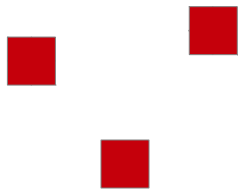
## Semi-supervised learning

- Annotating data is a time consuming and expensive task
- Solution
  - Using a minimum amount of annotated data
  - Annotating further data automatically

# Semi-supervised learning

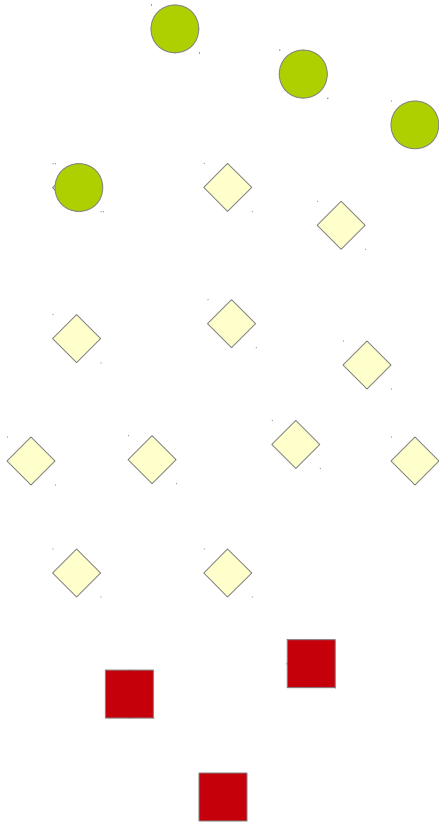


- A small amount of labeled data



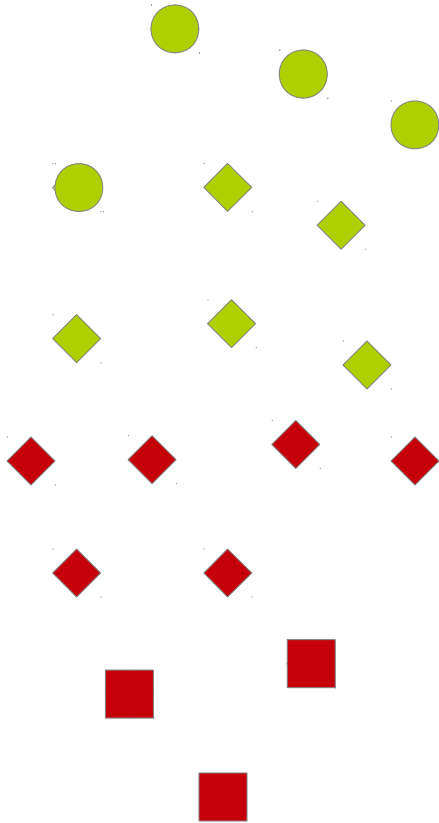


# Semi-supervised learning



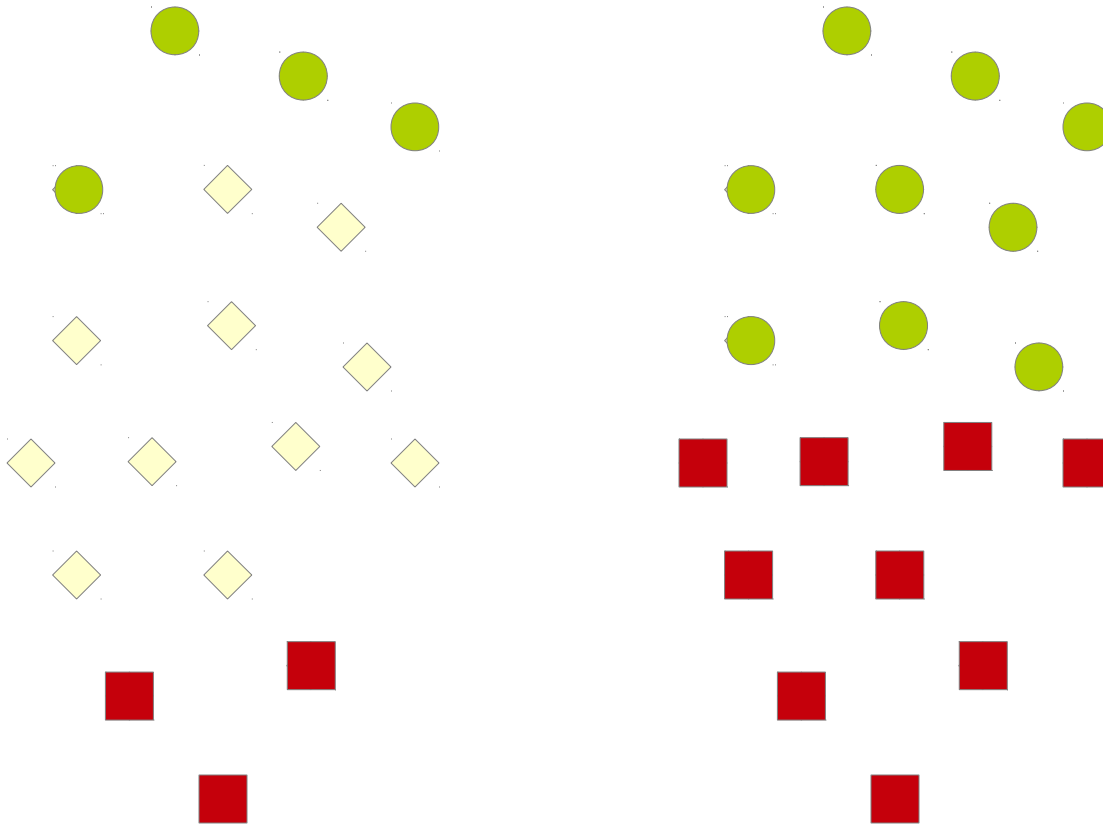
- A large amount of unlabeled data

# Semi-supervised learning



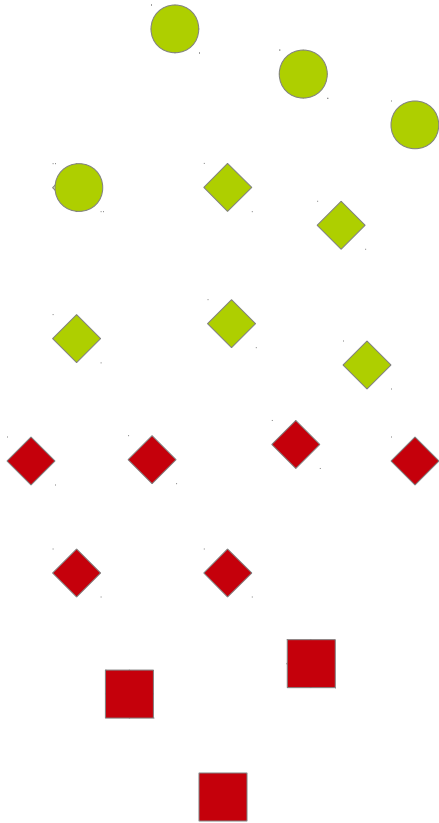
- Finding the similarity between the labeled and unlabeled data
- Predicting the labels of the unlabeled data

# Semi-supervised learning



- Training the classifier using labeled data and predicted labels of unlabeled data

# Semi-supervised learning

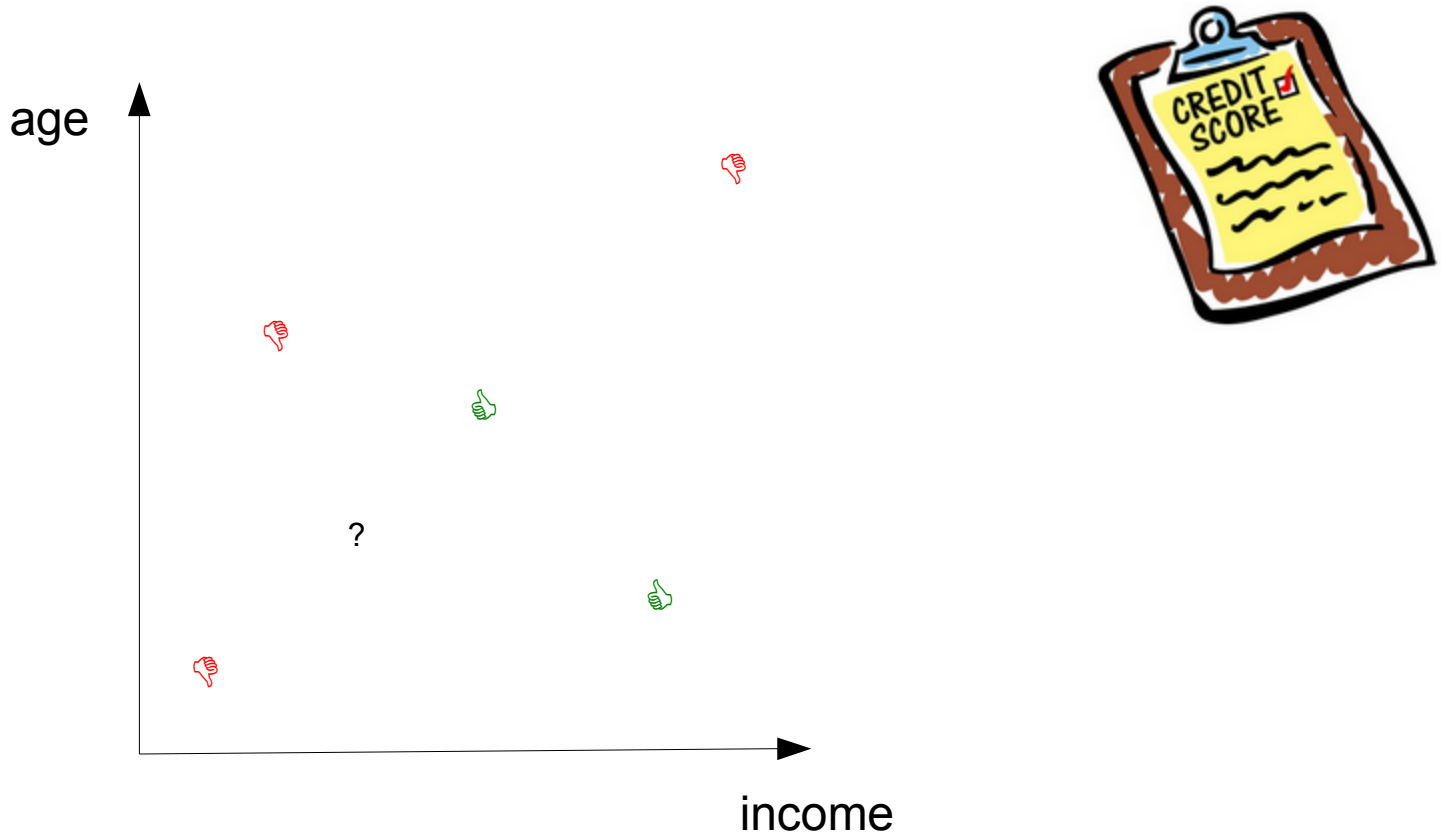


- Introducing a lot of noisy data to the system
- Adding unlabeled data to the training set, if the predicted label has a high confidence

# Outline

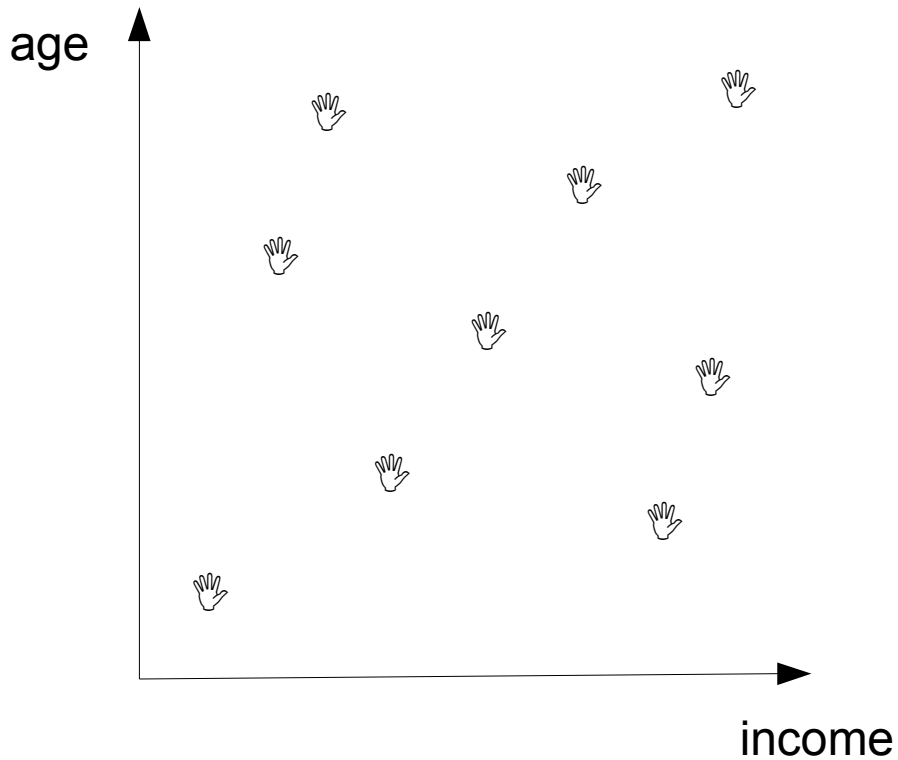
- Supervised Learning
- Semi-supervised learning
- Unsupervised learning

# Supervised Learning



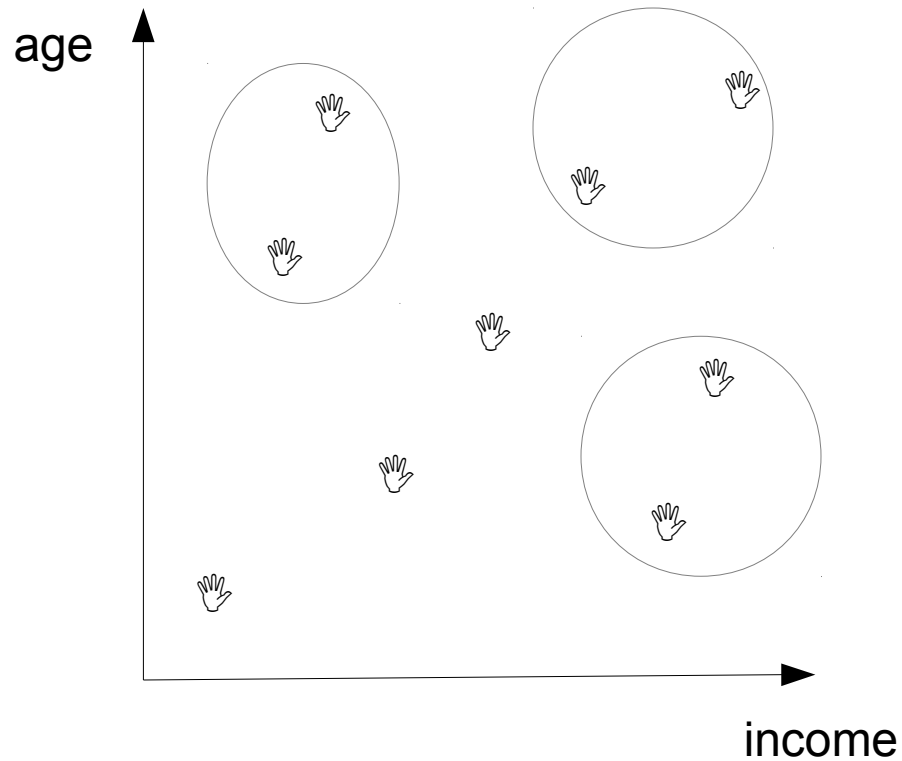
<http://nationalmortgageprofessional.com/news24271/regulatory-compliance-outlook-new-risk-based-pricing-rules>

# Unsupervised Learning



<http://nationalmortgageprofessional.com/news24271/regulatory-compliance-outlook-new-risk-based-pricing-rules>

# Unsupervised Learning



<http://nationalmortgageprofessional.com/news24271/regulatory-compliance-outlook-new-risk-based-pricing-rules>



# Clustering

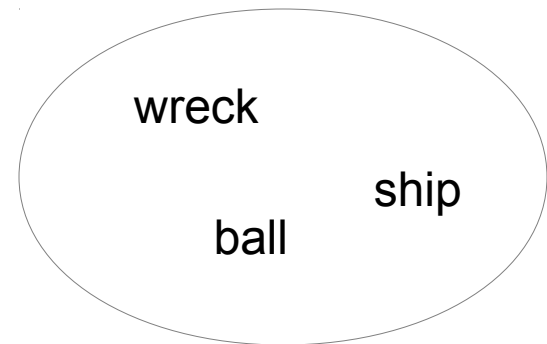
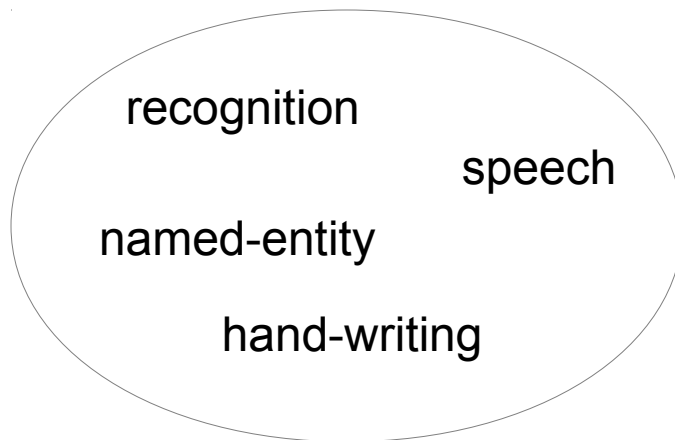
- Calculating similarities between the data items
- Assigning similar data items to the same cluster

# Applications

- Word clustering
  - Speech recognition
  - Machine translation
  - Named entity recognition
  - Information retrieval
  - ...
  
- Document clustering
  - Text classification
  - Information retrieval
  - ...

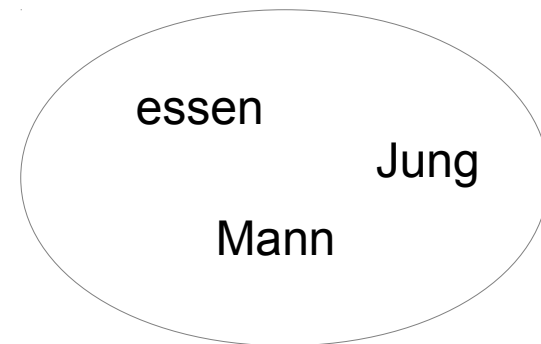
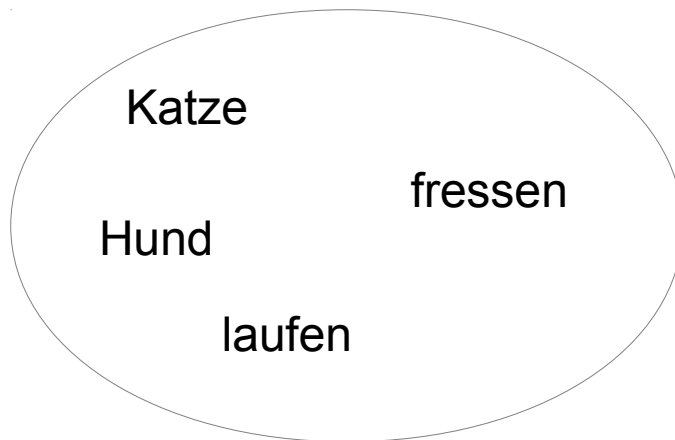
## Speech recognition

- „Computers can recognize a speech.“
- „Computers can wreck a nice peach.“



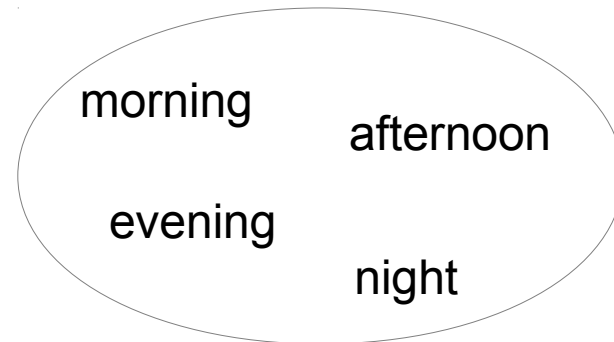
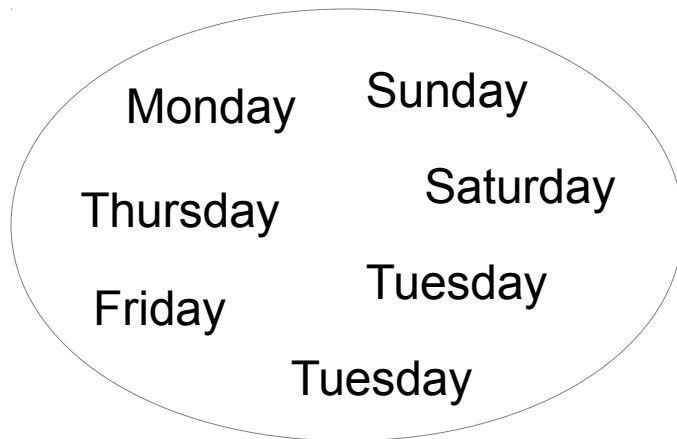
## Machine translation

- „The cat eats...”
  - „Die Katze frisst...”
  - „Die Katze isst...”



## Language modelling

- „I have a meeting on Monday evening.“
- „You should work on Wednesday afternoon.“
- „The next session is on Thursday morning.“
- „The talk is on Monday morning.“
- „The talk is on Monday molding.“



# Clustering algorithms

- Flat
  - K-means
- Hierarchical
  - Top-Down (Divisive)
  - Bottom-Up (Agglomerative)
    - Single-link
    - Complete-link
    - Average-link

## K-means

- The best known clustering algorithm
- Works well for many cases
- Used as default/baseline for clustering documents
- Defining each cluster center as the mean or centroid of the items in the cluster

$$\vec{\mu} = \frac{1}{|c|} \sum_{\vec{x} \in c} \vec{x}$$

- Minimizing the average squared Euclidean distance of the items from their cluster centers

## K-means

**Initialization:** Randomly choose k items as initial centroids

**while** stopping criterion has not been met **do**

**for** each item **do**

    Find the nearest centroid

    Assign the item to the cluster associated with the nearest centroid

**end for**

**for** each cluster **do**

    Update the centroid of the cluster based on the average of all items in the cluster

**end for**

**end while**

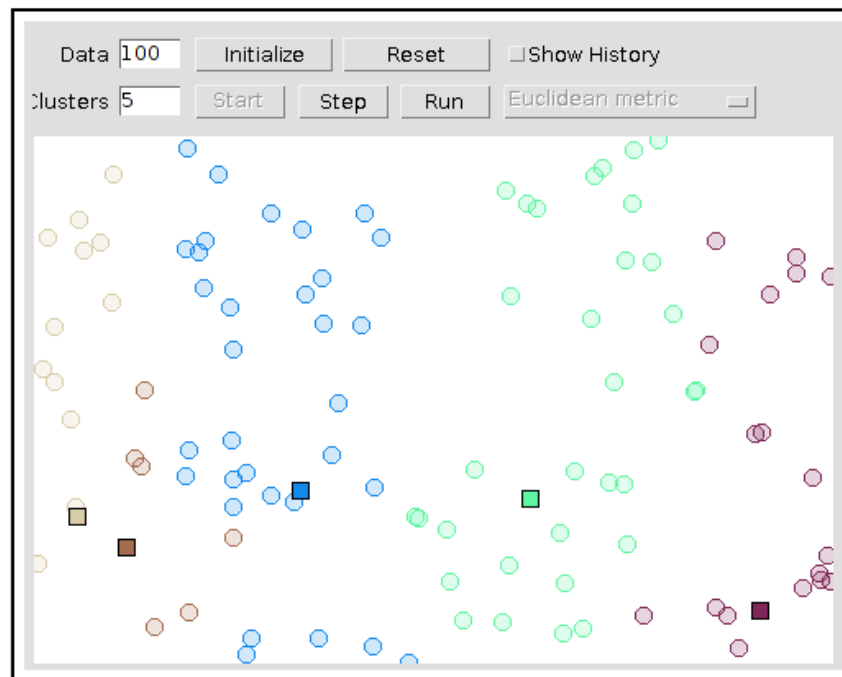
- Iterating two steps:
  - Re-assignment
    - Assigning each vector to its closest centroid
  - Re-computation
    - Computing each centroid as the average of the vectors that were assigned to it in re-assignment



# K-means

## K-means - Interactive demo

This applet requires Java Runtime Environment version 1.3 or later. You can download it from the [Sun Java website](http://www.sun.com/javase/6/docs/technotes/guides/beans/tutorial/applets.html).

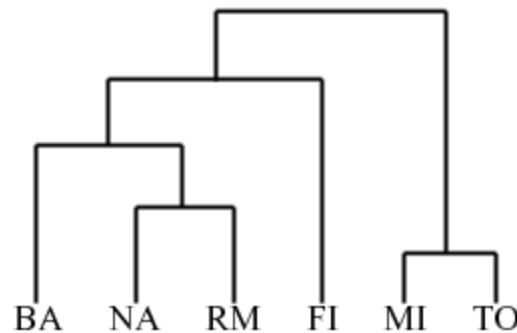


[http://home.deib.polimi.it/matteucc/Clustering/tutorial\\_html/AppletKM.html](http://home.deib.polimi.it/matteucc/Clustering/tutorial_html/AppletKM.html)

# Hierarchical Agglomerative Clustering (HAC)

- Creating a hierarchy in the form of a binary tree

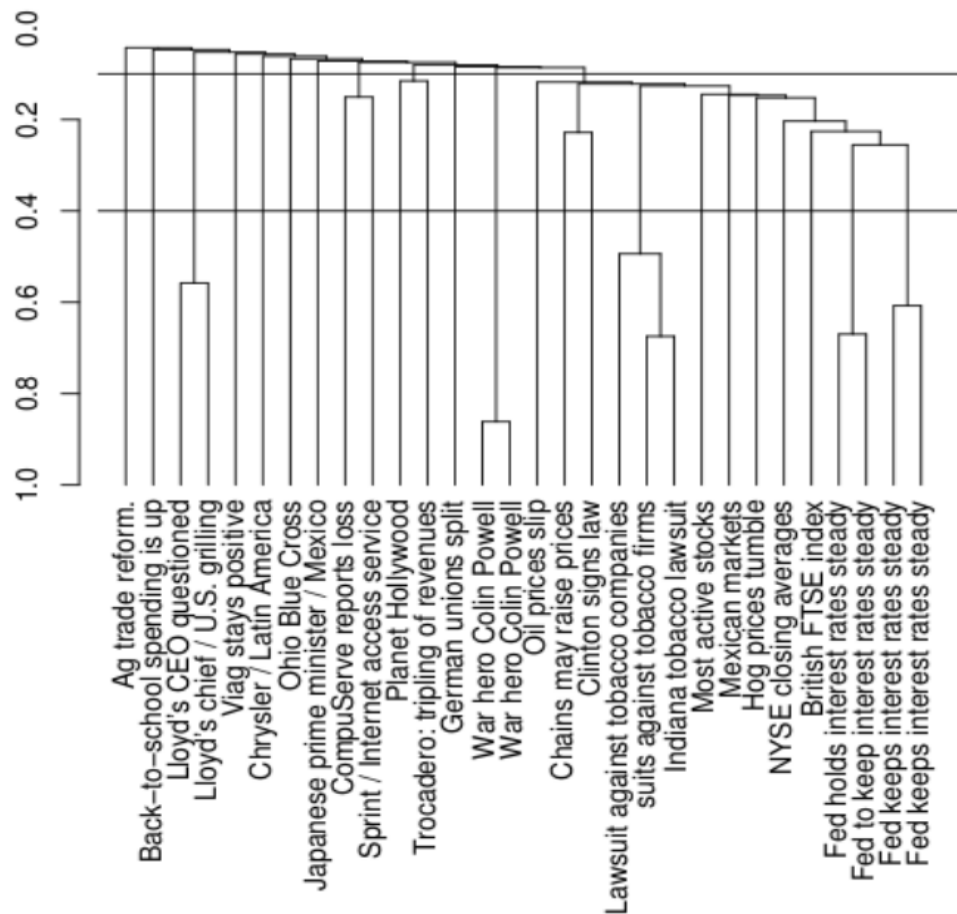
|           | <b>BA</b> | <b>FI</b> | <b>MI</b> | <b>NA</b> | <b>RM</b> | <b>TO</b> |
|-----------|-----------|-----------|-----------|-----------|-----------|-----------|
| <b>BA</b> | 0         | 662       | 877       | 255       | 412       | 996       |
| <b>FI</b> | 662       | 0         | 295       | 468       | 268       | 400       |
| <b>MI</b> | 877       | 295       | 0         | 754       | 564       | 138       |
| <b>NA</b> | 255       | 468       | 754       | 0         | 219       | 869       |
| <b>RM</b> | 412       | 268       | 564       | 219       | 0         | 669       |
| <b>TO</b> | 996       | 400       | 138       | 869       | 669       | 0         |



[http://home.deib.polimi.it/matteucc/Clustering/tutorial\\_html/hierarchical.html](http://home.deib.polimi.it/matteucc/Clustering/tutorial_html/hierarchical.html)

# Hierarchical Agglomerative Clustering (HAC)

- Creating a hierarchy in the form of a binary tree



## Hierarchical Agglomerative Clustering (HAC)

**Initial Mapping:** Put a single item in each cluster

**while** reaching the predefined number of clusters **do**

**for** each pair of clusters **do**

    Measure the similarity of two clusters

**end for**

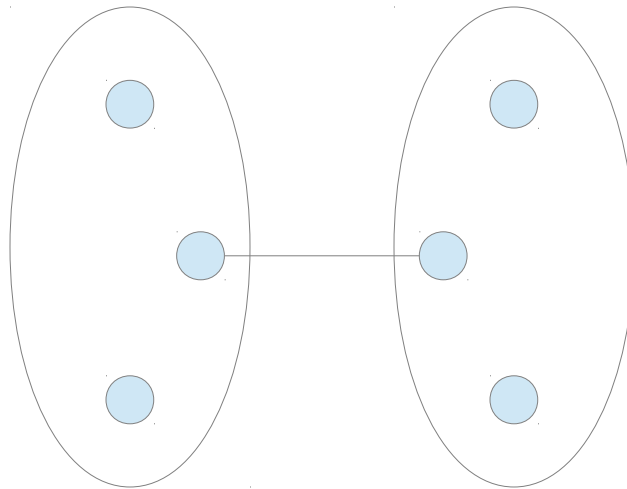
    Merge the two clusters that are most similar

**end while**

- Measuring the similarity in three ways:
  - Single-link
  - Complete-link
  - Average-link

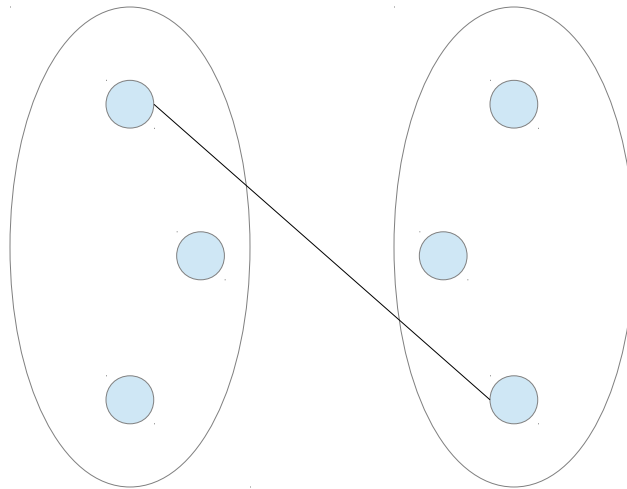
# Hierarchical Agglomerative Clustering (HAC)

- Single-link / single-linkage clustering
  - Based on the similarity of the most similar members



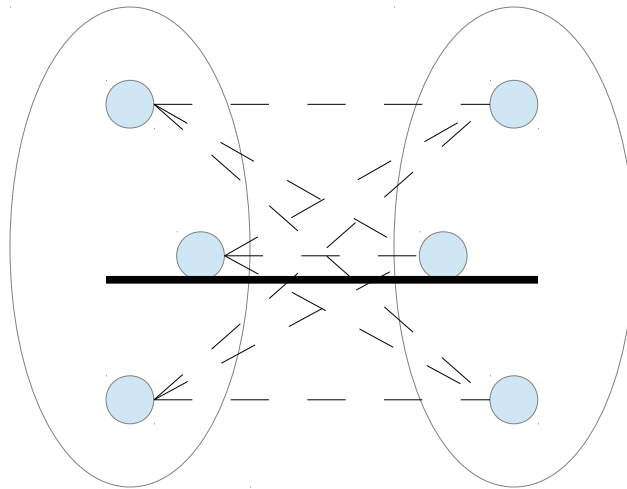
# Hierarchical Agglomerative Clustering (HAC)

- Complete-link / complete-linkage clustering
  - Based on the similarity of the most dissimilar members



# Hierarchical Agglomerative Clustering (HAC)

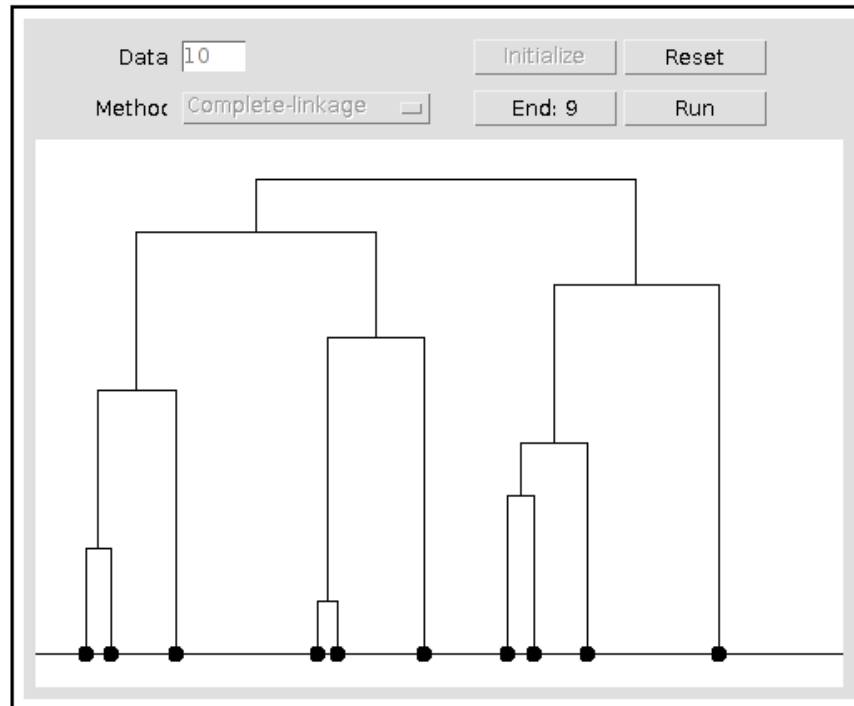
- Average-link / average-linkage clustering
  - Based on the average of all similarities between the members



# Hierarchical Agglomerative Clustering (HAC)

## Hierarchical Clustering - Interactive demo

This applet requires Java Runtime Environment version 1.3 or later. You can download it from the [Sun Java website](http://www.sun.com).



[http://home.deib.polimi.it/matteucc/Clustering/tutorial\\_html/AppletH.html](http://home.deib.polimi.it/matteucc/Clustering/tutorial_html/AppletH.html)



# This is no clustering...just word frequencies

Wordle™ Home Create Gallery Credits News Forum FAQ Advanced

"English notebook cover" by Ace Acedemic! 4 years, 8 months ago

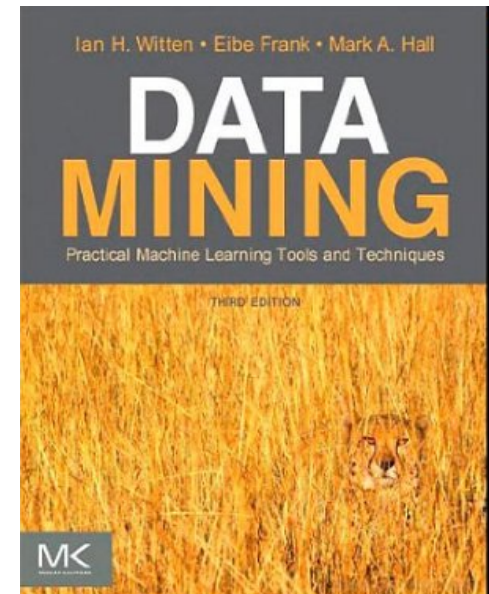
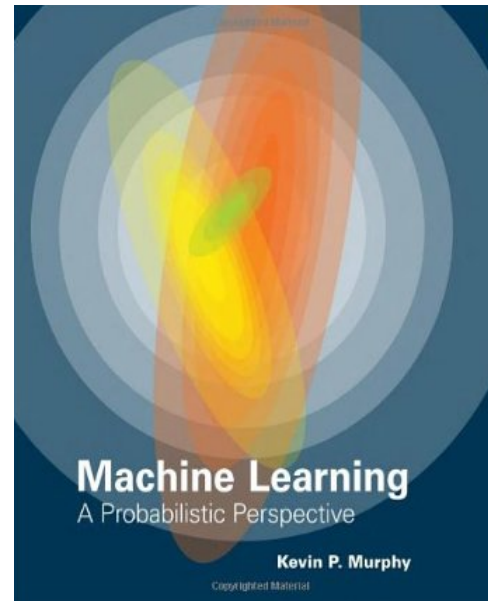
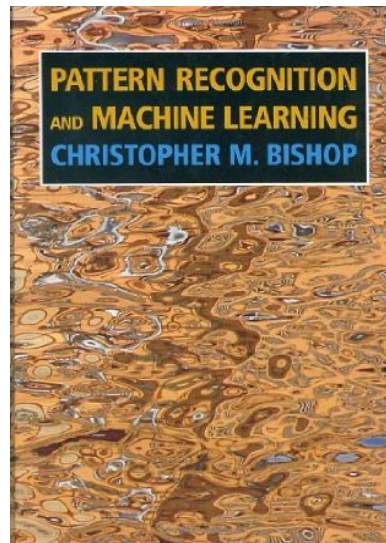
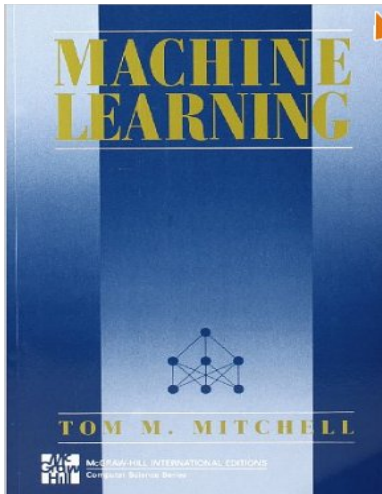


Open in Window Print...

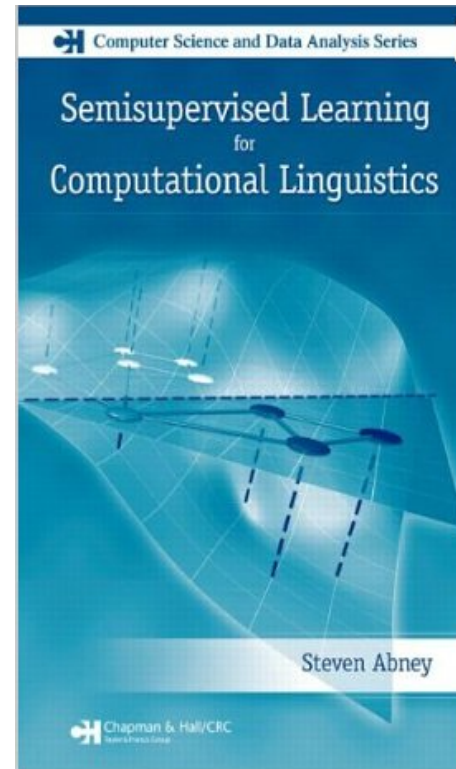
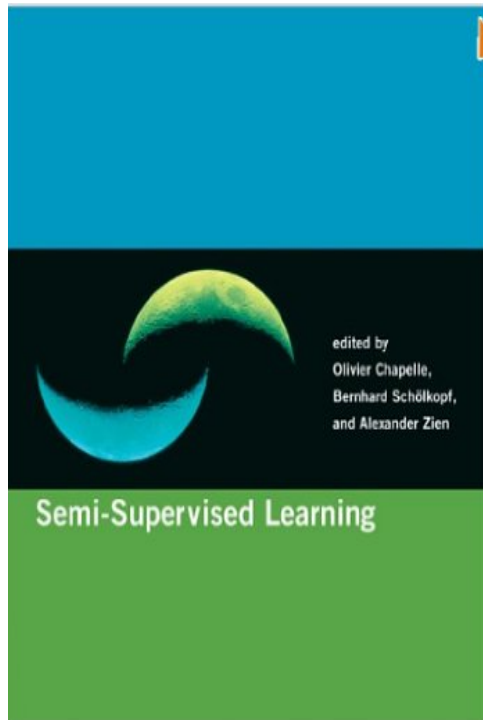
© 2013 Jonathan Feinberg [Terms of Use](#)  
build #1421

[http://www.wordle.net/display/wrdl/1059224/English\\_notebook\\_cover](http://www.wordle.net/display/wrdl/1059224/English_notebook_cover)

## Further reading



## Further reading



# Further reading

