Chapter 6:
Text Categorization

See chapter 16 in Manning&Schütze

Text Categorization and related Tasks

Classification

Goal:

Assign 'objects' from a universe to two or more *classes* or *categories*

Problem	Object	Categories
Text Categorization	Document	Topic
Spam Mail Detection	Document	spam/not spam
Author identification	Document	Authors
Sense Disambiguation	Word/Doc.	The word's senses
Tagging/Sequence-Labl.	Words	POS/NE
Machine translation	Sentence	Sentence
Dialog system	Sentence	Sentence
Information retrieval	Query/Doc.	Relevant/not relevant
Parsing	Sentence	Tree

Spam/junk/bulk Emails

- The messages you spend your time with just to delete them
  - Spam: do not want to get unsolicited messages
  - Junk: irrelevant to the recipient, unwanted
  - Bulk: mass mailing for business marketing (or fill-up mailbox etc.)

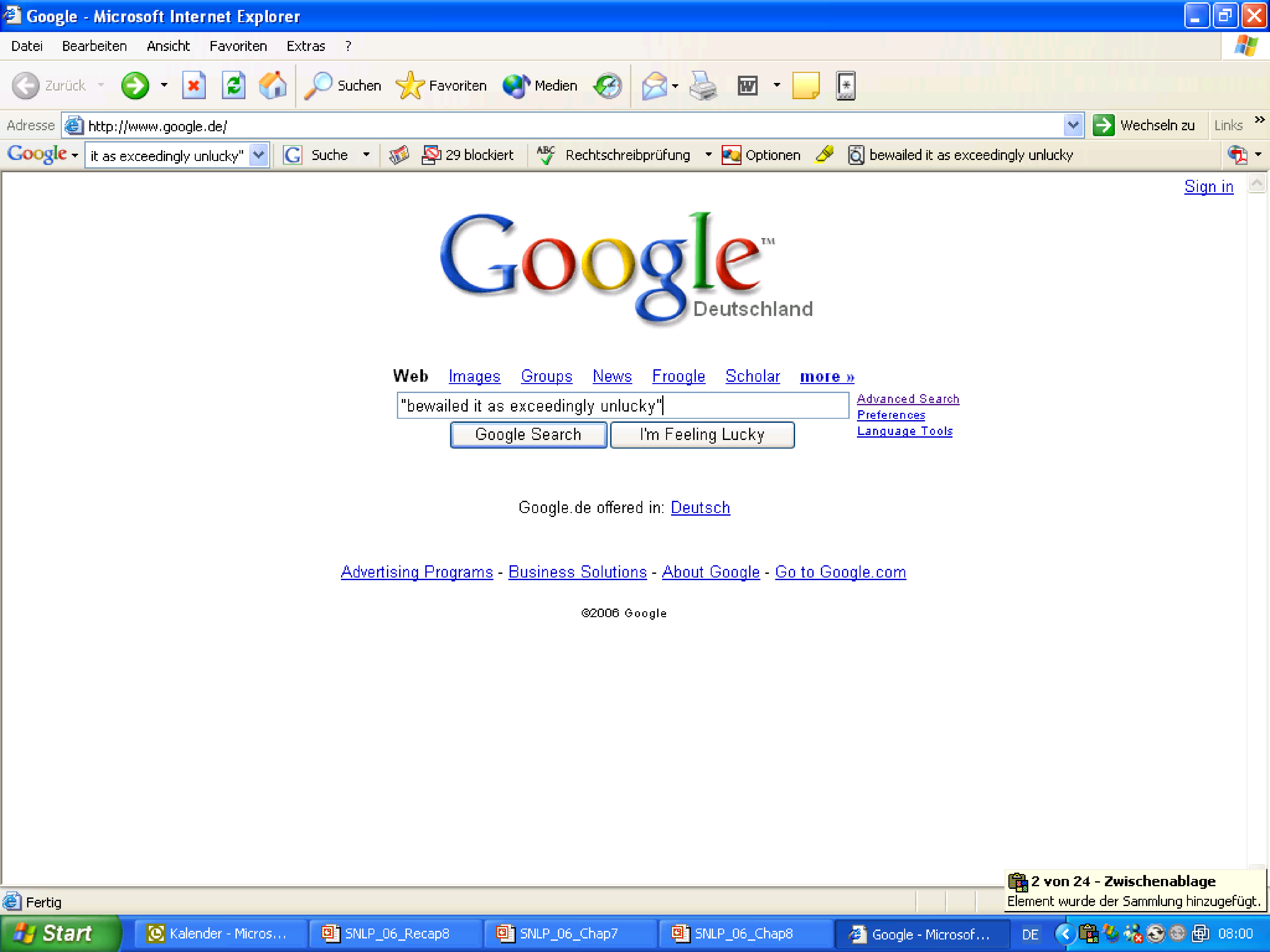
Classification task: decide for each e-mail whether it is spam/not-spam

Author identification

- They agreed that Mrs. X should only hear of the departure of the family, without being alarmed on the score of the gentleman's conduct; but even this partial communication gave her a great deal of concern, and she bewailed it as exceedingly unlucky that the ladies should happen to go away, just as they were all getting so intimate together.
- Gas looming through the fog in divers places in the streets, much as the sun may, from the spongey fields, be seen to loom by husbandman and ploughboy. Most of the shops lighted two hours before their time--as the gas seems to know, for it has a haggard and unwilling look. The raw afternoon is rawest, and the dense fog is densest, and the muddy streets are muddiest near that leaden-headed old obstruction, appropriate ornament for the threshold of a leaden-headed old corporation, Temple Bar.

Author identification

- They agreed that Mrs. X should only hear of the departure of the family, without being alarmed on the score of the gentleman's conduct; but even this partial communication gave her a great deal of concern, and she *bewailed it as exceedingly unlucky* that the ladies should happen to go away, just as they were all getting so intimate together.



Google - Microsoft Internet Explorer
Datei Bearbeiten Ansicht Favoriten Extras ?
Zurück Suchen Favoriten Medien
Adresse http://www.google.de/
Wechseln zu Links
Google it as exceedingly unlucky" Suche 29 blockiert Rechtschreibprüfung Optionen bewailed it as exceedingly unlucky
Sign in
Google Deutschland
Web Images Groups News Froogle Scholar more »
"bewailed it as exceedingly unlucky"
Advanced Search
Preferences
Language Tools
Google Search
I'm Feeling Lucky
Google.de offered in: Deutsch
Advertising Programs - Business Solutions - About Google - Go to Google.com
©2006 Google
2 von 24 - Zwischenablage
Element wurde der Sammlung hinzugefügt.
Fertig
Start Kalender - Micros... SNLP_06_Recap8 SNLP_06_Chap7 SNLP_06_Chap8 Google - Microsof... DE 08:00

"bewailed it as exceedingly unlucky" - Google Search - Microsoft Internet Explorer

Datei Bearbeiten Ansicht Favoriten Extras ?

Adresse http://www.google.de/search?hl=en&q=%22bewailed+it+as+exceedingly+unlucky%22

Sign in

Web Images Groups News Froogle Scholar more »

"bewailed it as exceedingly unlucky" Search Advanced Search Preferences

Web Results **1** - **8** of about **96** for "**bewailed** **it as** **exceedingly** **unlucky**". (**0.26** seconds)

Did you mean: "bewailed it as exceedingly ***unlikely***"

Jane Austen: Pride and Prejudice, Chapter XXI of Volume I (Chap. 21)
... and she **bewailed it as exceedingly unlucky** that the ladies should happen to go away, just as they were all getting so intimate together. ...
www.pemberley.com/janeinfo/ppv1n21.html - 19k - Cached - Similar pages

Chapter XXI. Austen, Jane. 1917. Pride and Prejudice. Vol. III ...
... and she **bewailed it as exceedingly unlucky** that the ladies should happen to go away just as they were all getting so intimate together. ...
www.bartleby.com/303/2/21.html - 33k - Cached - Similar pages

[PDF] Pride and Prejudice
File Format: PDF/Adobe Acrobat
concern, and she **bewailed it as exceedingly unlucky** that the ladies should happen to go away just as they. were all getting so intimate together. ...
www.bartleby.com/ebook/adobe/3032.pdf - Similar pages

Pride and Prejudice
... and he **bewailed it as exceedingly unlucky** that the gentlemen should happen to go away, just as they were all getting so intimate together. ...
www.lifeamgood.com/pnpchapter19_21.html - 35k - Cached - Similar pages

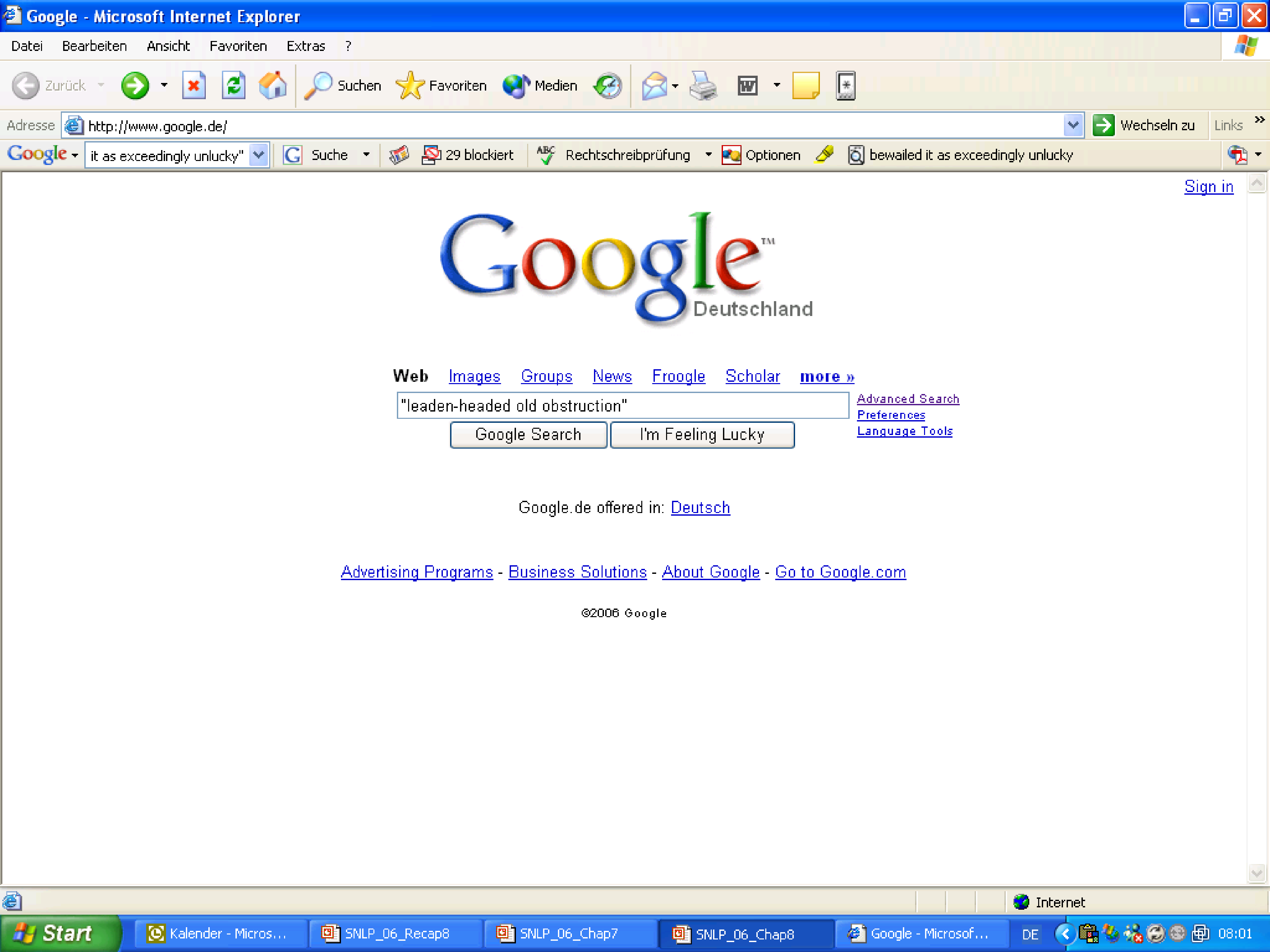
[PDF] Pride and Prejudice
File Format: PDF/Adobe Acrobat - View as HTML
cern, and she **bewailed it as exceedingly unlucky** that the ladies should. happen to go

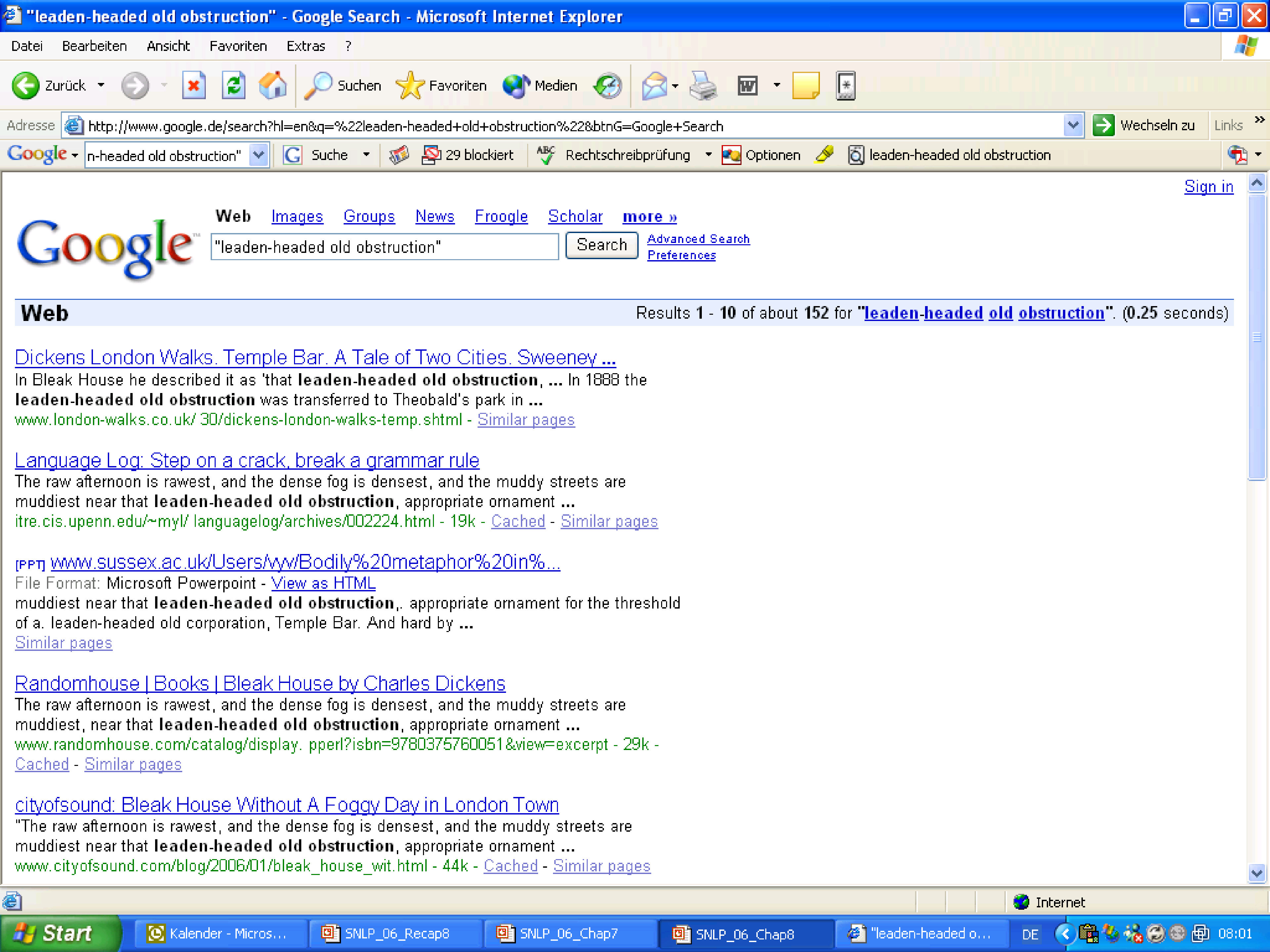
SLOX Synchronization
Appointments: Sent (2) failed item.

Start Kalender - Micros... SNLP_06_Recap8 SNLP_06_Chap7 SNLP_06_Chap8 "bewailed it as ex... DE 08:00

Author identification

- Gas looming through the fog in divers places in the streets, much as the sun may, from the spongey fields, be seen to loom by husbandman and ploughboy. Most of the shops lighted two hours before their time--as the gas seems to know, for it has a haggard and unwilling look. The raw afternoon is rawest, and the dense fog is densest, and the muddy streets are muddiest near that *leaden-headed old obstruction*, appropriate ornament for the threshold of a leaden-headed old corporation, Temple Bar.


Google - Microsoft Internet Explorer
Datei Bearbeiten Ansicht Favoriten Extras ?
Zurück
Suchen
Favoriten
Medien
Adresse http://www.google.de/
Wechseln zu
Links
Google
it as exceedingly unlucky"
Suche
29 blockiert
Rechtschreibprüfung
Optionen
bewailed it as exceedingly unlucky
Sign in
Google
Deutschland
Web Images Groups News Froogle Scholar more »
"leaden-headed old obstruction"
Google Search
I'm Feeling Lucky
Advanced Search
Preferences
Language Tools
Google.de offered in: Deutsch
Advertising Programs - Business Solutions - About Google - Go to Google.com
©2006 Google
Internet
Start
Kalender - Micros...
SNLP_06_Recap8
SNLP_06_Chap7
SNLP_06_Chap8
Google - Microsof...
DE
08:01


"leaden-headed old obstruction" - Google Search - Microsoft Internet Explorer
Datei Bearbeiten Ansicht Favoriten Extras ?
Zurück Suchen Favoriten Medien
Adresse http://www.google.de/search?hl=en&q=%22leaden-headed+old+obstruction%22&btnG=Google+Search
Wechseln zu Links
Suche 29 blockiert Rechtschreibprüfung Optionen leaden-headed old obstruction
Sign in
Web Images Groups News Froogle Scholar more »
"leaden-headed old obstruction"
Search
Advanced Search
Preferences
Web
Results 1 - 10 of about 152 for "leaden-headed old obstruction". (0.25 seconds)
Dickens London Walks. Temple Bar. A Tale of Two Cities. Sweeney ...
In Bleak House he described it as 'that leaden-headed old obstruction, ... In 1888 the leaden-headed old obstruction was transferred to Theobald's park in ...
www.london-walks.co.uk/ 30/dickens-london-walks-temp.shtml - Similar pages
Language Log: Step on a crack, break a grammar rule
The raw afternoon is rawest, and the dense fog is densest, and the muddy streets are muddiest near that leaden-headed old obstruction, appropriate ornament ...
itre.cis.upenn.edu/~myl/ languagelog/archives/002224.html - 19k - Cached - Similar pages
[PPT] www.sussex.ac.uk/Users/vyv/Bodily%20metaphor%20in%...
File Format: Microsoft Powerpoint - View as HTML
muddiest near that leaden-headed old obstruction,. appropriate ornament for the threshold of a. leaden-headed old corporation, Temple Bar. And hard by ...
Similar pages
Randomhouse | Books | Bleak House by Charles Dickens
The raw afternoon is rawest, and the dense fog is densest, and the muddy streets are muddiest, near that leaden-headed old obstruction, appropriate ornament ...
www.randomhouse.com/catalog/display. pperl?isbn=9780375760051&view=excerpt - 29k - Cached - Similar pages
cityofsound: Bleak House Without A Foggy Day in London Town
"The raw afternoon is rawest, and the dense fog is densest, and the muddy streets are muddiest near that leaden-headed old obstruction, appropriate ornament ...
www.cityofsound.com/blog/2006/01/bleak_house_wit.html - 44k - Cached - Similar pages
Internet
Start Kalender - Micros... SNLP_06_Recap8 SNLP_06_Chap7 SNLP_06_Chap8 "leaden-headed o... DE 08:01

Author identification

- Jane Austen (1775-1817), Pride and Prejudice
- Charles Dickens (1812-70), Bleak House

Author identification

- Federalist papers
  - 77 short essays written in 1787-1788 by Hamilton, Jay and Madison to persuade NY to ratify the US Constitution; published under a pseudonym
  - The authorships of 12 papers was in dispute (*disputed papers*)
  - In 1964 Mosteller and Wallace* solved the problem
  - They identified 70 *function* words as good candidates for authorships analysis
  - Using statistical inference they concluded the author was Madison

Function words for Author Identification

1	*a*	15	*do*	29	*is*	43	*or*	57	*this*
2	*all*	16	*down*	30	*it*	44	*our*	58	*to*
3	*also*	17	*even*	31	*its*	45	*shall*	59	*up*
4	*an*	18	*every*	32	*may*	46	*should*	60	*upon*
5	*and*	19	*for*	33	*more*	47	*so*	61	*was*
6	*any*	20	*from*	34	*must*	48	*some*	62	*were*
7	*are*	21	*had*	35	*my*	49	*such*	63	*what*
8	*as*	22	*has*	36	*no*	50	*than*	64	*when*
9	*at*	23	*have*	37	*not*	51	*that*	65	*which*
10	*be*	24	*her*	38	*now*	52	*the*	66	*who*
11	*been*	25	*his*	39	*of*	53	*their*	67	*will*
12	*but*	26	*if*	40	*on*	54	*then*	68	*with*
13	*by*	27	*in*	41	*one*	55	*there*	69	*would*
14	*can*	28	*into*	42	*only*	56	*things*	70	*your*

Table 1: Function Words and Their Code Numbers

Function words for Author Identification

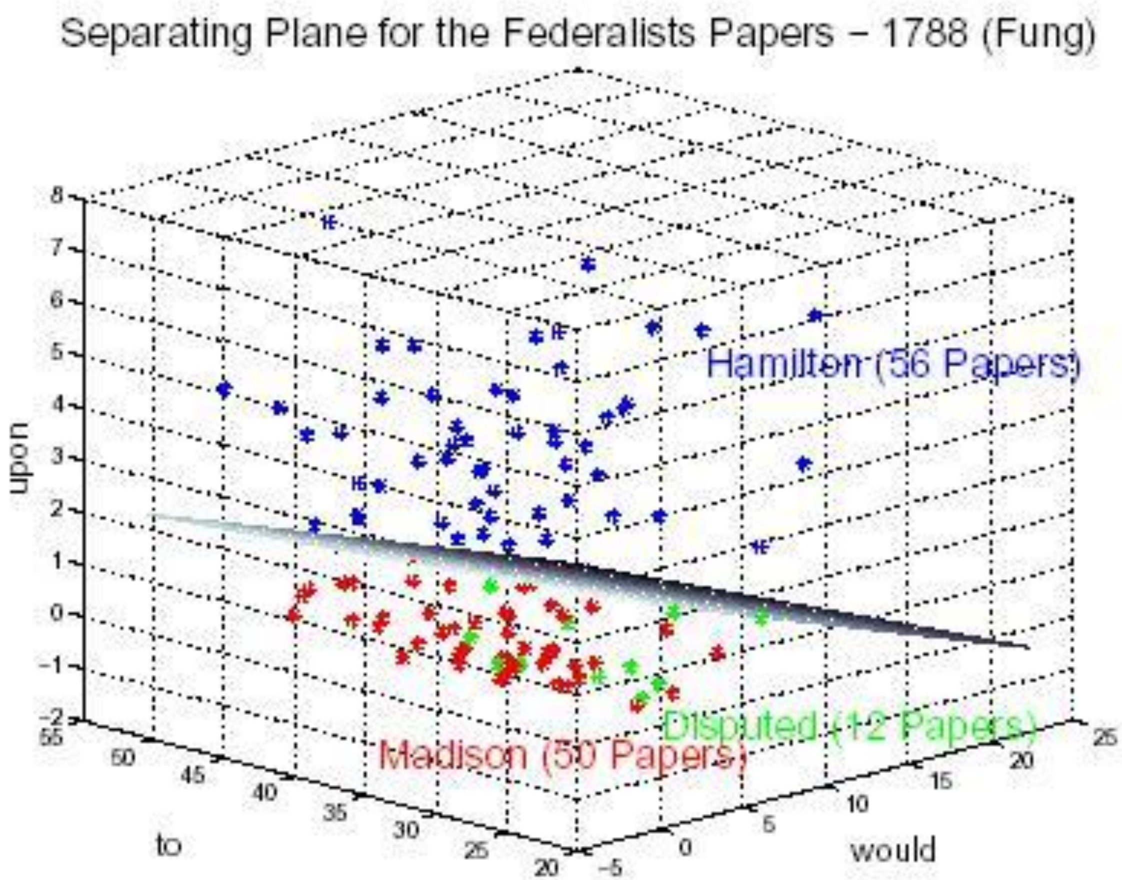
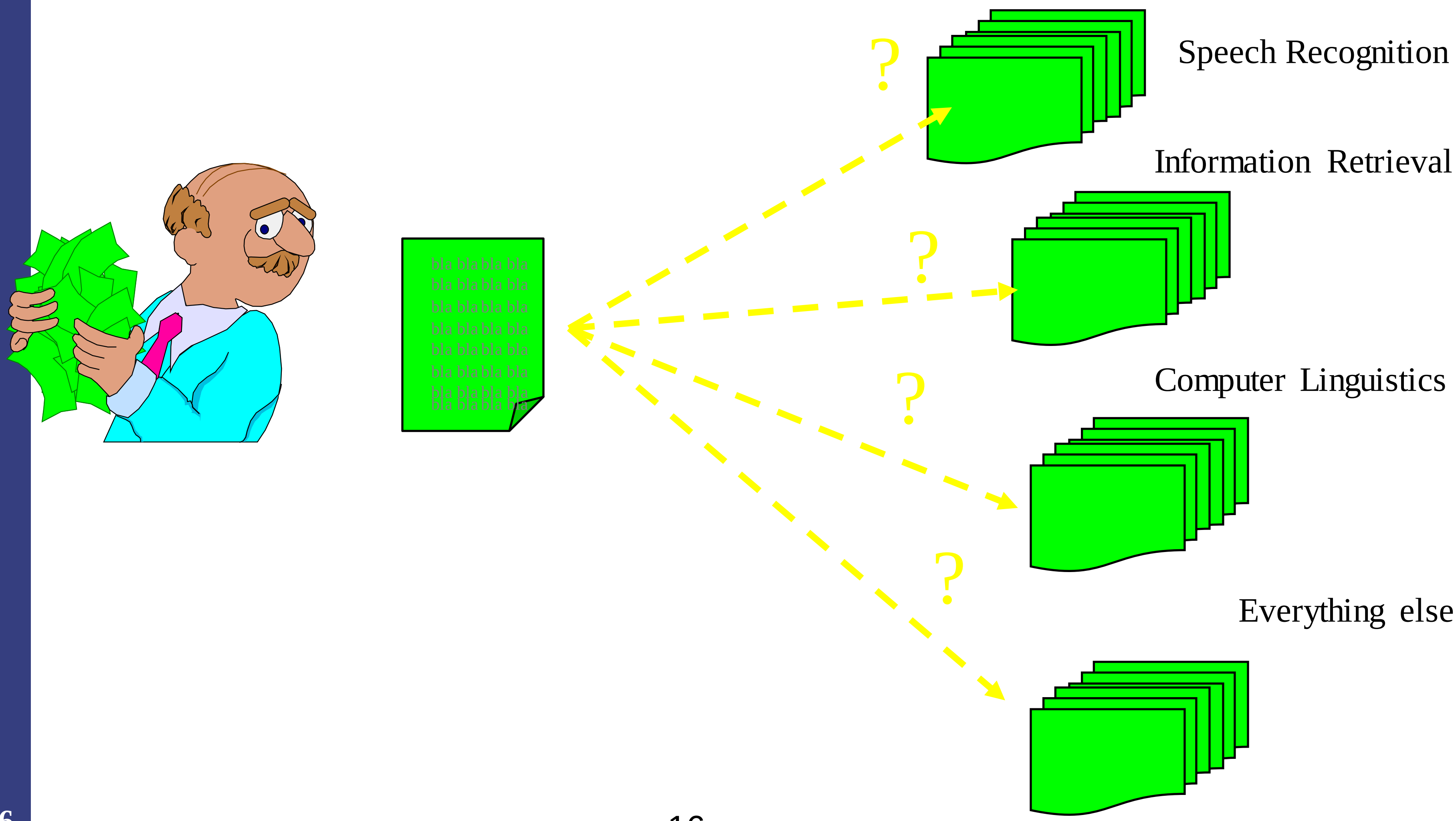


Figure 1: Obtained Hyperplane in 3 dimensions

15

Text Categorization



Text Categorization

- Topic categorization: classify the document into semantics topics

The U.S. swept into the Davis Cup final on Saturday when twins Bob and Mike Bryan defeated Belarus's Max Mirnyi and Vladimir Voltchkov to give the Americans an unsurmountable 3-0 lead in the best-of-five semi-final tie.

One of the strangest, most relentless hurricane seasons on record reached new bizarre heights yesterday as the plodding approach of Hurricane Jeanne prompted evacuation orders for hundreds of thousands of Floridians and high wind warnings that stretched 350 miles from the swamp towns south of Miami to the historic city of St. Augustine.

17

Google News - Microsoft Internet Explorer

Datei Bearbeiten Ansicht Favoriten Extras ?

Zurück Suchen Favoriten Medien

Adresse http://news.google.com/ Wechseln zu Links

Google n-headed old obstruction" Suche 29 blockiert Rechtschreibprüfung Optionen leaden-headed old obstruction

Sign in

Google News

Web Images Groups **News** Froogle Maps **more »** Advanced News Search

Search News | Search the Web

Search and browse 4,500 news sources updated continuously.

Standard News | Text Version

- **>Top Stories**
- **World**
- **U.S.**
- **Business**
- **Sci/Tech**
- **Sports**
- **Entertainment**
- **Health**
- **Most Popular**

Make Google News Your Homepage

News Alerts

RSS | Atom
About Feeds

Mobile News

Top Stories U.S. Go

Auto-generated **18 minutes ago**

Israel targets Hamas leaders

Peninsula On-line

The Standard - 1 hour ago
Israeli tanks and troops massed near Gaza for a threatened offensive against the Palestinians, and the Israeli government said it would target Hamas leaders if a captured soldier was not freed. Israeli tanks **...**
Hamas-Fatah to Implicitly Recognize Israel ABC News
Olmert defends W.Bank pullout plan amid Gaza crisis Reuters AlertNet
Ireland Online - United Press International - Times Online - Belfast Telegraph -
all 2,966 related »

The New York Times

Buffalo News

National Review Online Blogs - 20 hours ago
The New York Times' decision to disclose the Terrorist Finance Tracking Program, a robust and classified effort to map terrorist networks through the use of financial data, was irresponsible and harmful to the security of Americans and freedom-loving **...**
Bush condemns disclosure of secret anti-terror program CNN
The media vs. the president -- again Town Hall
Los Angeles Times - New York Times - San Jose Mercury News - Bloomberg -
all 612 related »

Get recommended stories
Sign in to get recommended stories by using search history

Personalize this page

Buffett: Gates' charity 'surest way' to helping
TMCnet - all 1,667 related »

Intel Unveils Xeon 5100 Processors
Techtree.com - all 268 related »

UNC Throws Away National Title
NBC 17.com - all 1,247 related »

EW review: 'Superman' is only average, man
CNN International - all 262 related »

Sexual orientation of men determined before birth
Reuters - all 411 related »

In The News

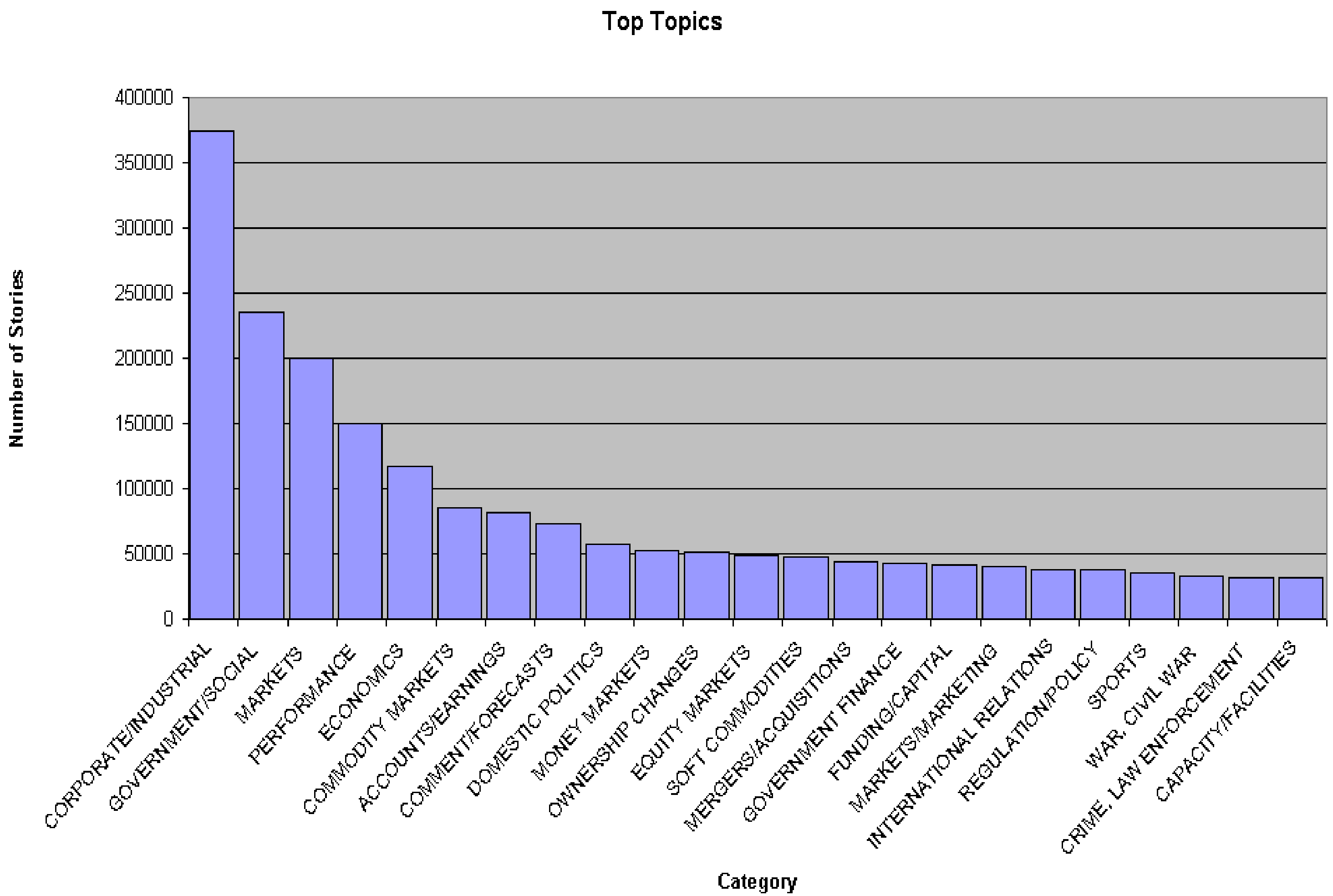
Keith Urban	Jeff Gordon
Harry Potter	Boy George
Knight Ridder	College World Series
Tamil Tiger	David Beckham
Roger Federer	Superman Returns

Start SNLP_06_Chap8 FoundationsSN... Foundations_0... lecture8_Text_... Index of /cours... Google News - ... DE 20:12

Text categorization

- Reuters
  - Collection of (21,578) newswire documents.
  - For research purposes: a standard text collection to compare systems and algorithms
  - 135 valid topics categories

Top topics in Reuters



Reuters

<REUTERS TOPICS="YES" LEWISSPLIT="TRAIN" CGISPLIT="TRAINING-SET" OLDID="12981" NEWID="798">

<DATE> 2-MAR-1987 16:51:43.42</DATE>

<TOPICS><D>livestock</D><D>hog</D></TOPICS>

<TITLE>AMERICAN PORK CONGRESS KICKS OFF TOMORROW</TITLE>

<DATELINE> CHICAGO, March 2 - </DATELINE><BODY>The American Pork Congress kicks off tomorrow, March 3, in Indianapolis with 160 of the nations pork producers from 44 member states determining industry positions on a number of issues, according to the National Pork Producers Council, NPPC.

Delegates to the three day Congress will be considering 26 resolutions concerning various issues, including the future direction of farm policy and the tax law as it applies to the agriculture sector. The delegates will also debate whether to endorse concepts of a national PRV (pseudorabies virus) control and eradication program, the NPPC said.

A large trade show, in conjunction with the congress, will feature the latest in technology in all areas of the industry, the NPPC added. Reuter

&#3;</BODY></TEXT></REUTERS>

Classification vs. Clustering

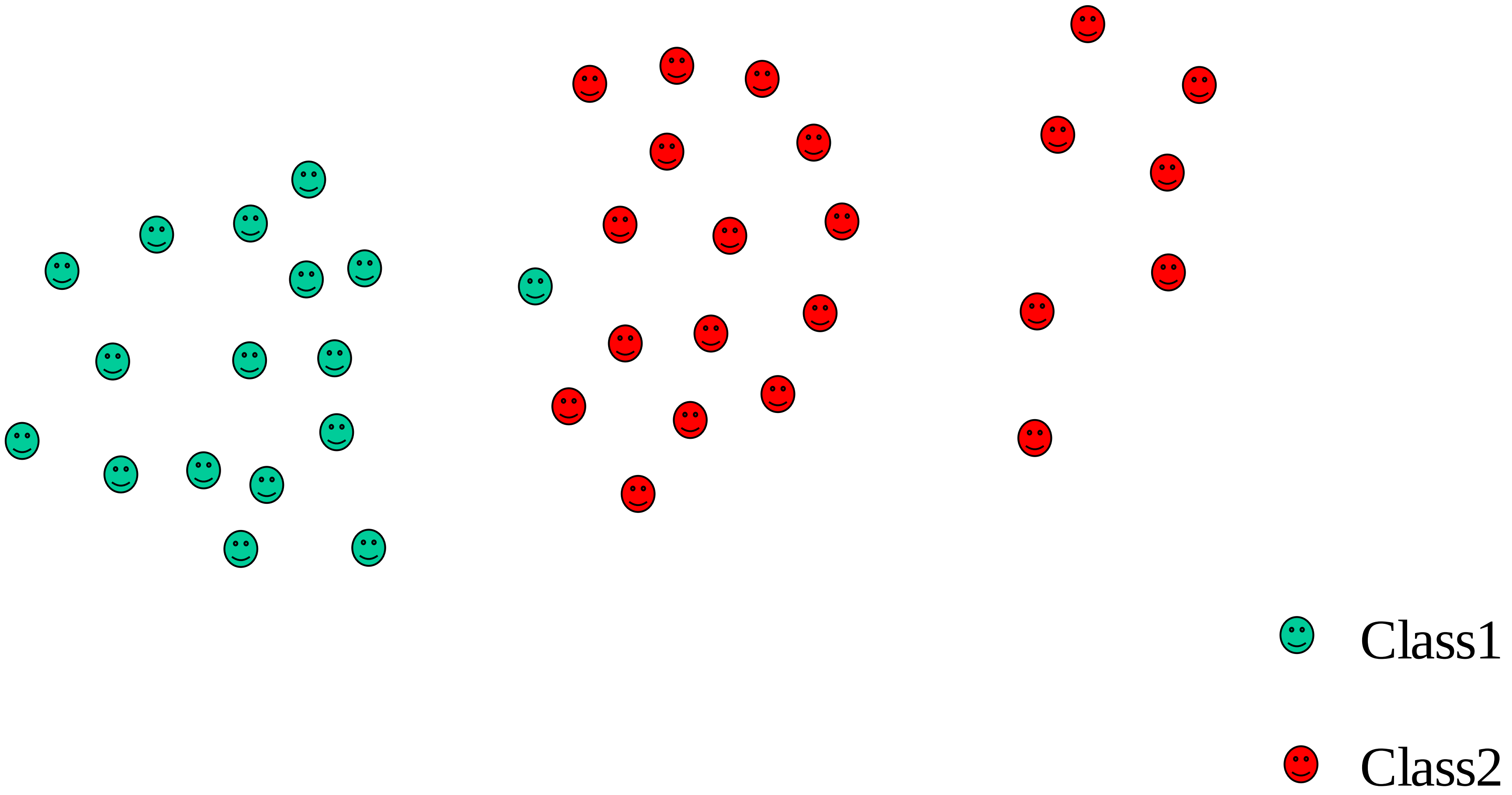
22

Classification vs. Clustering

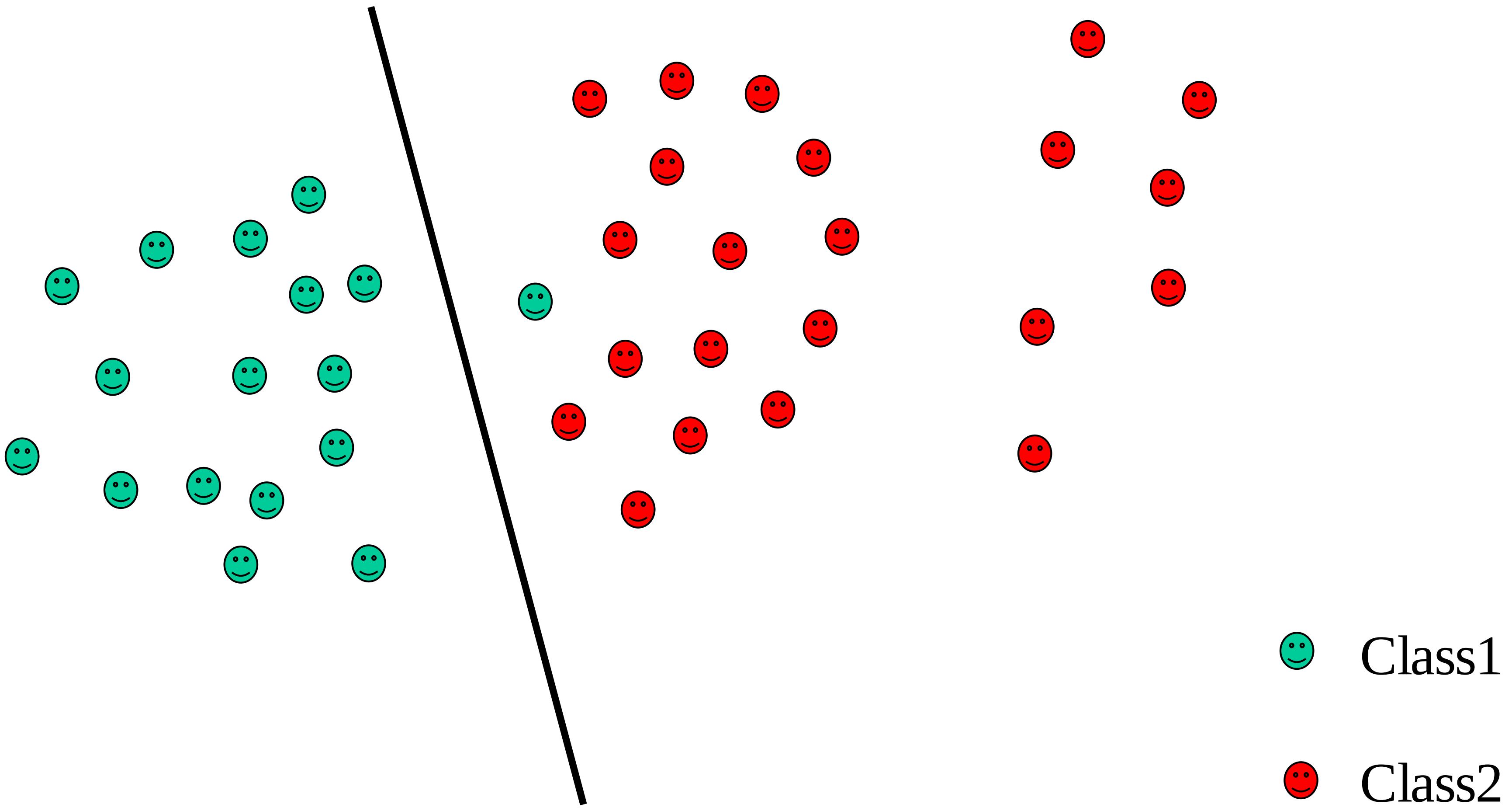
- Classification assumes labeled data: we know how many classes there are and we have examples for each class (labeled data).
- Classification is supervised
- In Clustering we don’t have labeled data; we just assume that there is a natural division in the data and we may not know how many divisions (clusters) there are
- Clustering is unsupervised

23

Classification

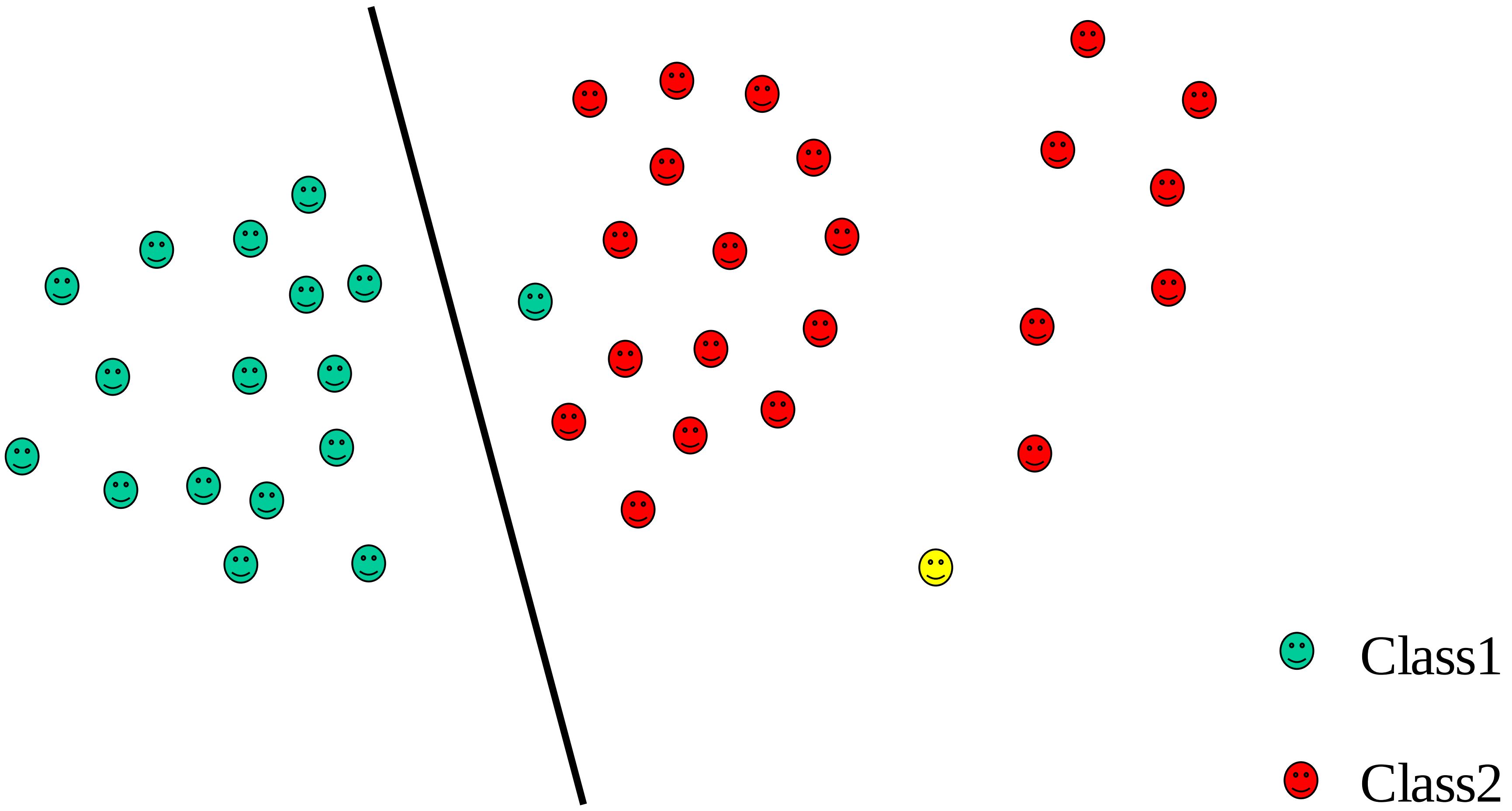


Classification

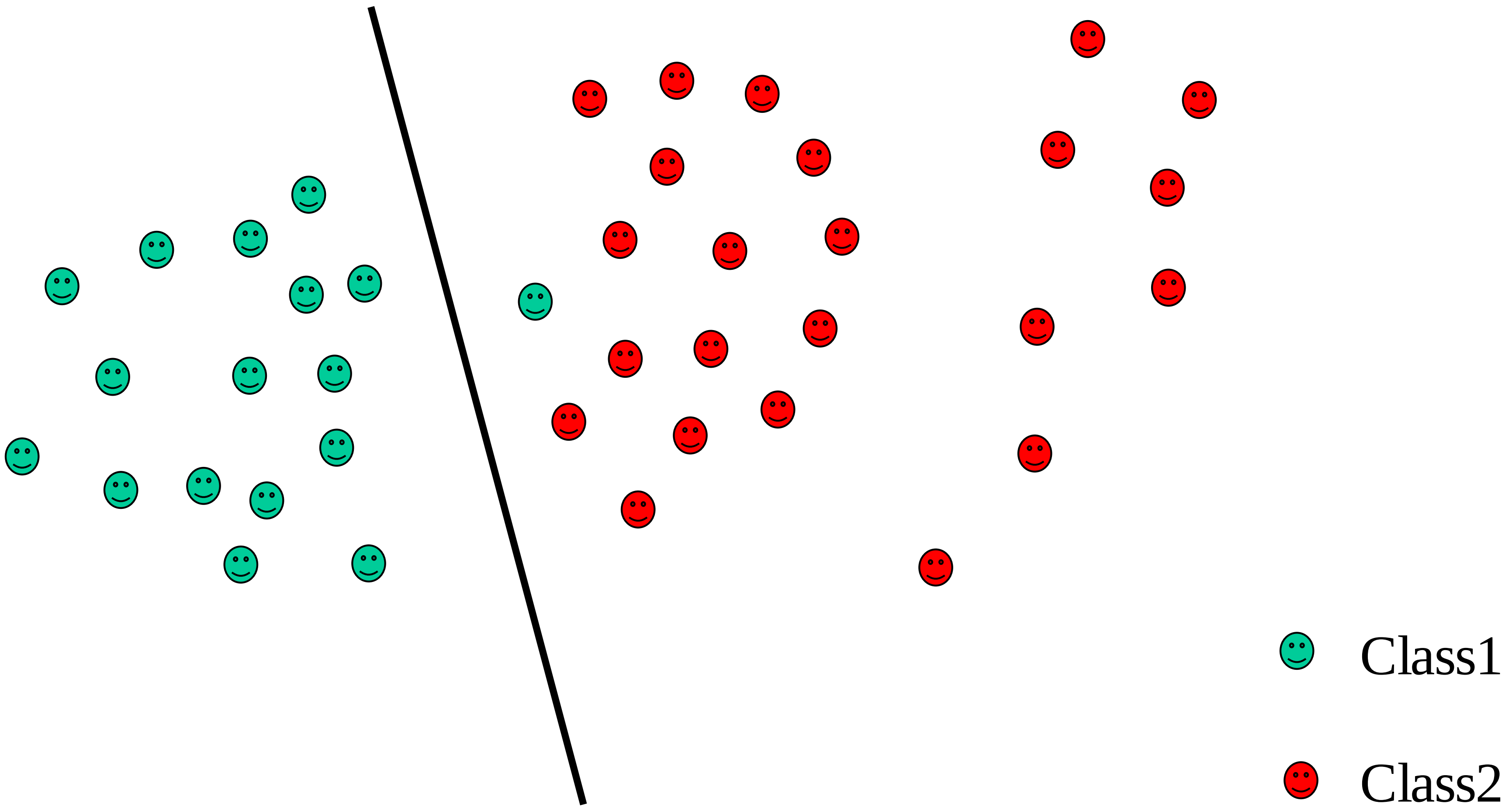


25

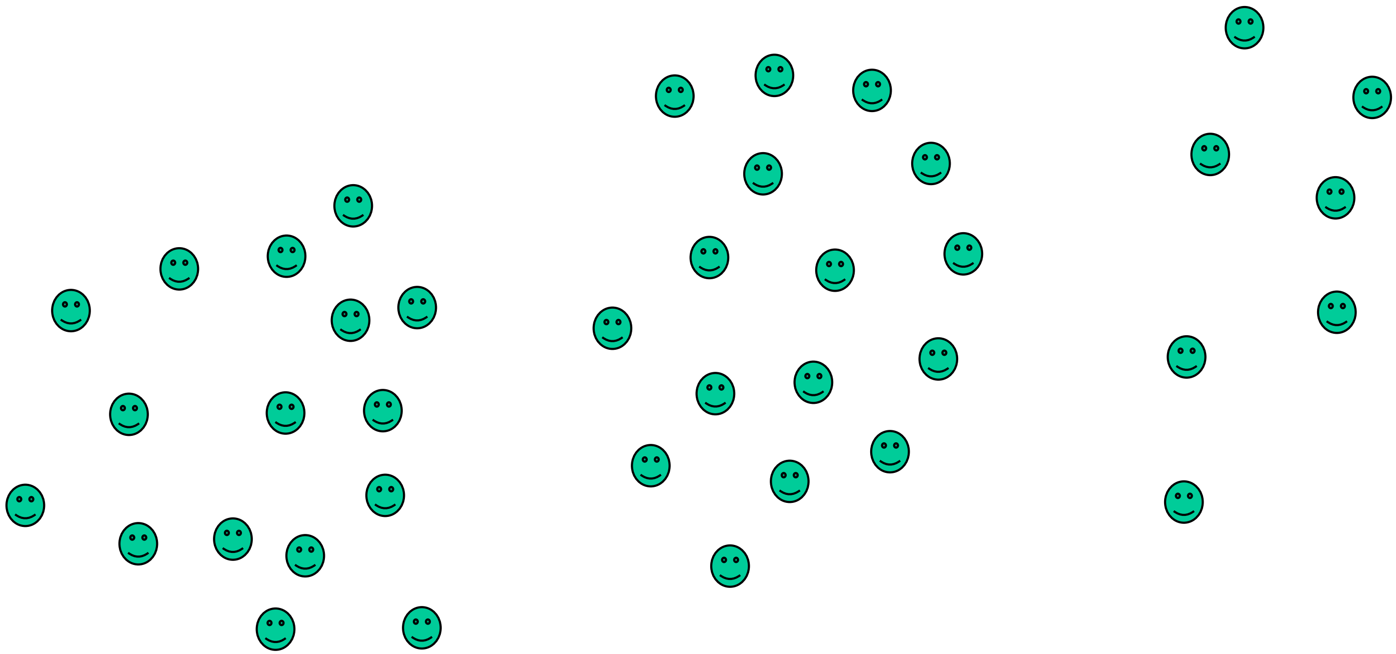
Classification



Classification

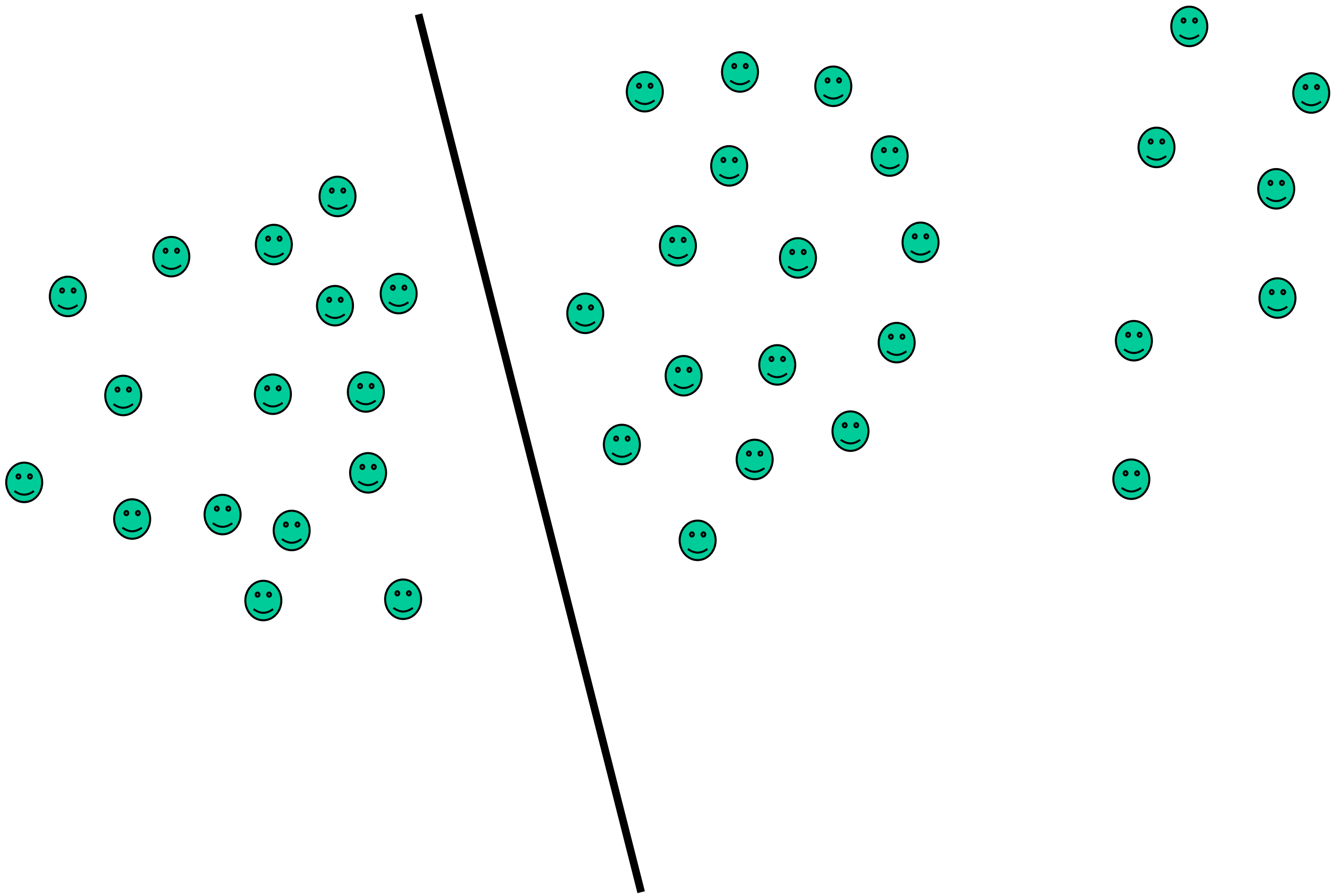


Clustering



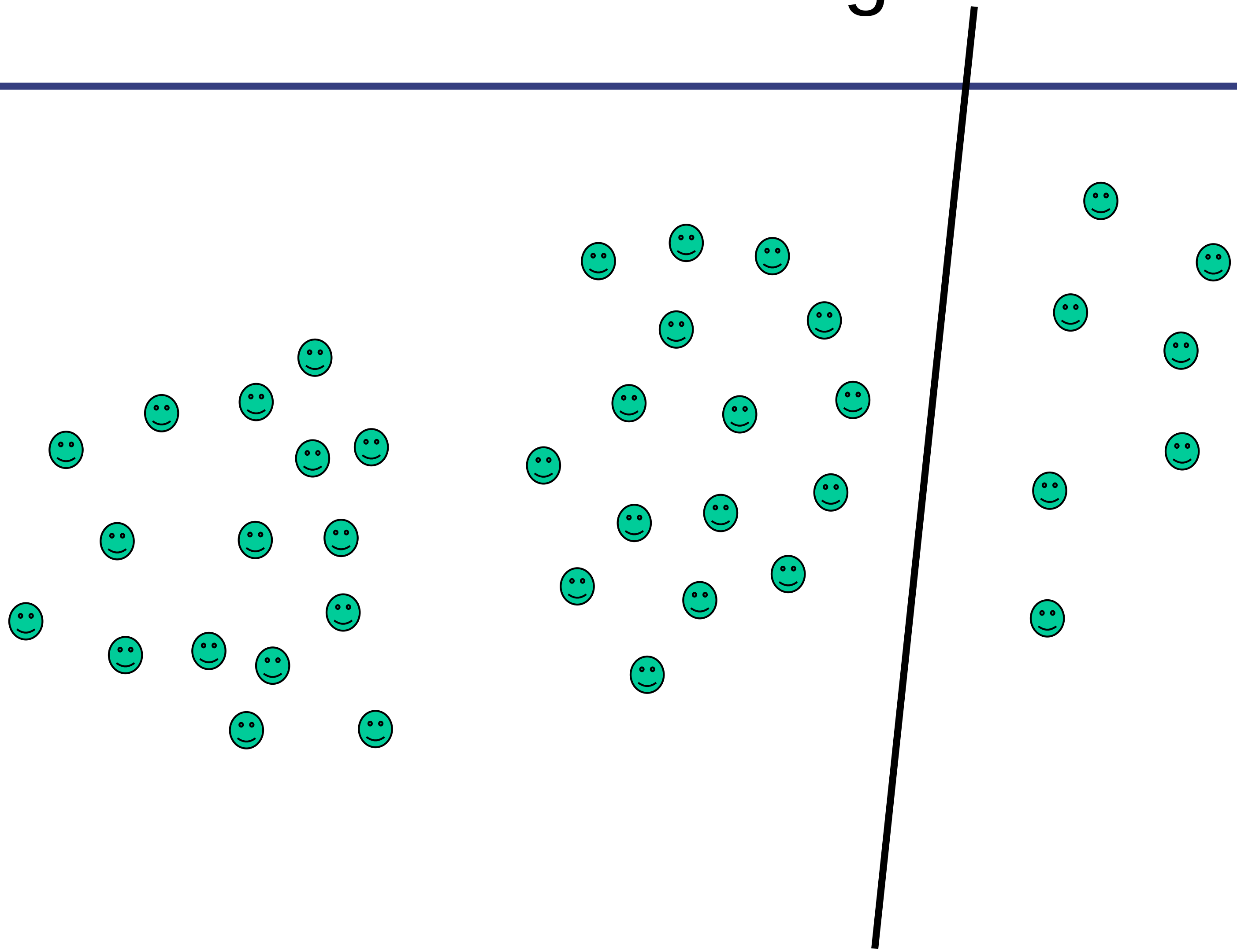
28

Clustering

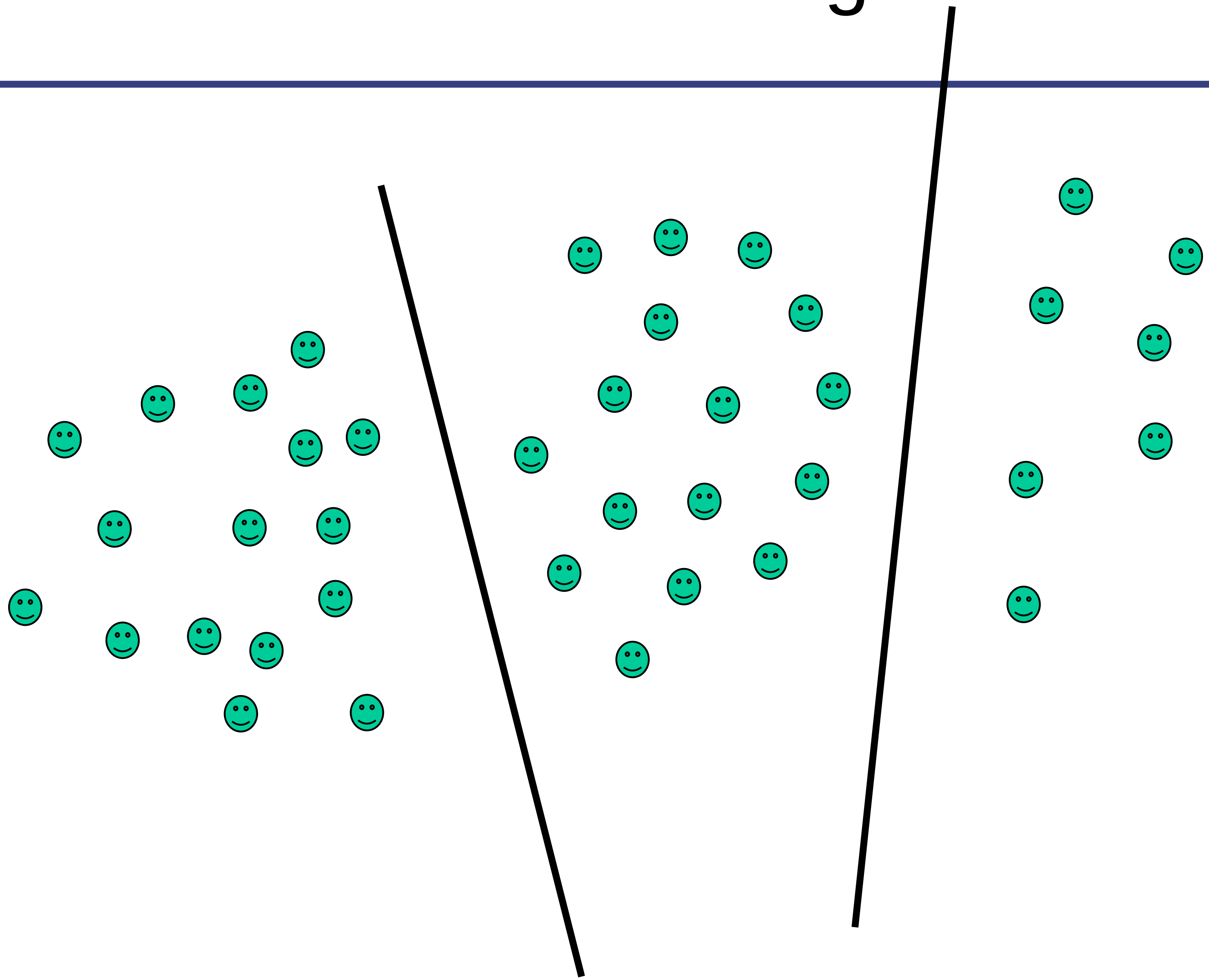


29

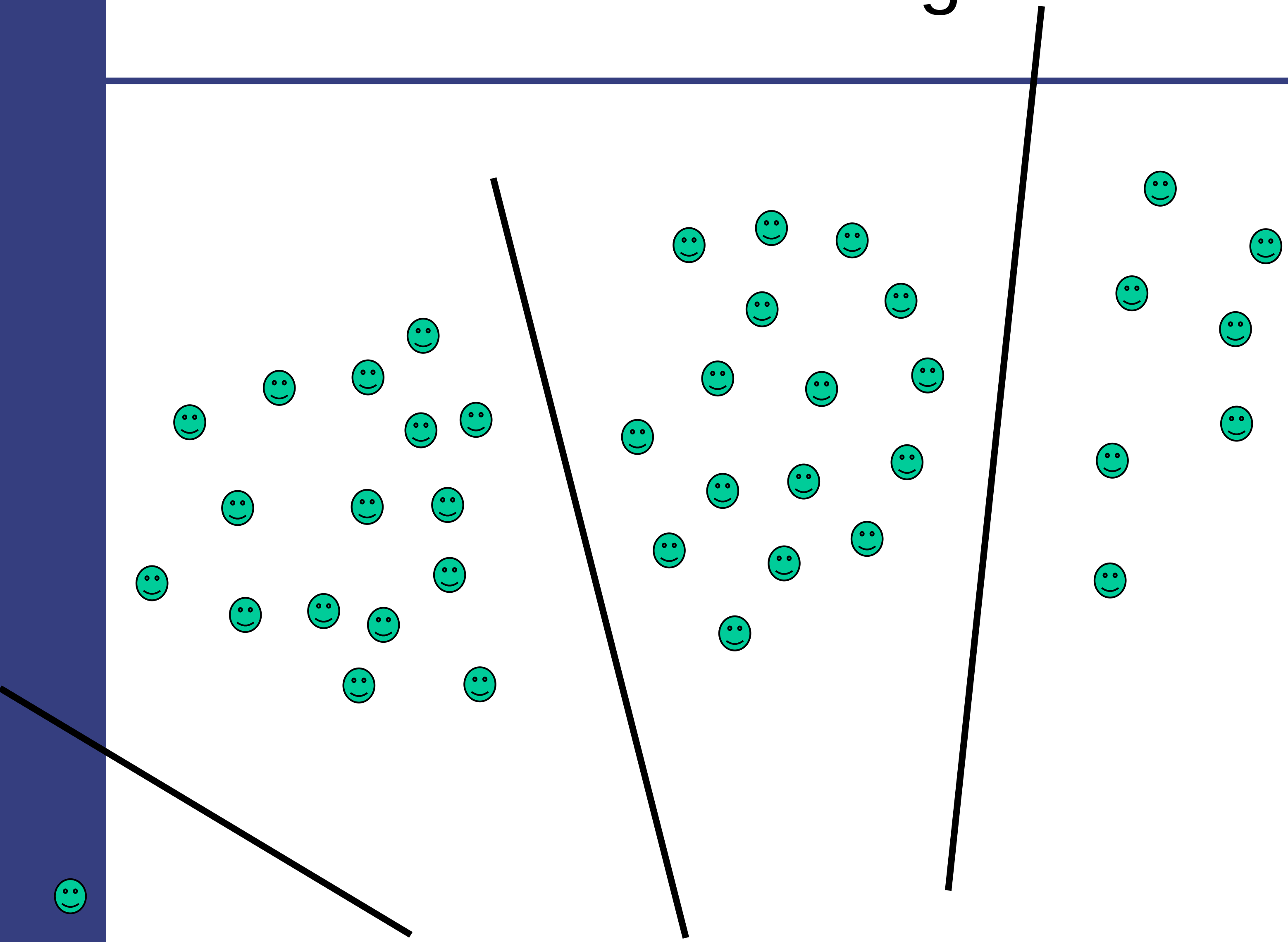
Clustering



Clustering



Clustering



Binary vs. multi-way classification

- Binary classification: two classes
- Multi-way classification: more than two classes
- Sometimes it can be convenient to treat a multi-way problem like a binary one: one class versus all the others, for all classes

Flat vs. Hierarchical classification

- Flat classification: relations between the classes undetermined
- Hierarchical classification: hierarchy where each node is the sub-class of its parent's node

34

Single- vs. multi-category classification

- In single-category text classification each text belongs to exactly one category
- In multi-category text classification, each text can have zero or more categories

Getting Features for Text Categorization

36

Feature terminology

- Feature: An aspect of the text that is relevant to the task
- Feature value: the realization of the feature in the text
  - Words present in text : Clinton, Schumacher, China...
  - Frequency of word: Clinton(10), Schumacher(1)...
  - Are there dates? Yes/no
  - Are there PERSONS? Yes/no
  - Are there ORGANIZATIONS? Yes/no
  - WordNet: Holonyms (China is part of Asia), Synonyms(China, People's Republic of China, mainland China)

Feature Types

- **Boolean (or Binary) Features**
- Features that generate Boolean (binary) values.
- Boolean features are the simplest and the most common type of feature.

  - f_1(text) = 1 if text contain "Clinton"

    0 otherwise
  - f_2(text) = 1 if text contain PERSON

    0 otherwise

Feature Types

- **Integer Features**
- Features that generate integer values.
- Integer features can be used to give classifiers access to more precise information about the text.

  - f_1(text) = Number of times text contains “Clinton”
  - f_2(text) = Number of times text contains PERSON

39

When Do We Need Feature Selection?

- If the algorithm cannot handle all possible features
  - e.g. language identification for 100 languages using all words
  - text classification using *n*-grams

- Good features can result in higher accuracy
  - But! Why feature selection?
  - What if we just keep all features?
    - Even the unreliable features can be helpful.
    - But we need to weight them:
      - In the extreme case, the bad features can have a weight of 0 (or very close), which is… a form of feature selection!

Why Feature Selection?

- Not all features are equally good!
  - Bad features: best to remove
    - Infrequent
      - unlikely to be be met again
      - co-occurrence with a class can be due to chance
    - Too frequent
      - mostly function words
    - Uniform across all categories
  - Good features: should be kept
    - Co-occur with a particular category
    - Do not co-occur with other categories
  - The rest: good to keep

Types Of Feature Selection?

- Feature selection reduces the number of features
  - Usually:
    - Eliminating features
    - Weighting features
    - Normalizing features
  - Sometimes by transforming parameters
    - e.g. Latent Semantic Indexing using Singular Value Decomposition
- Method may depend on problem type
  - For classification and filtering, may use information from example documents to guide selection

42

Feature Selection

- Task independent methods
  - Document Frequency (DF)
  - Term Strength (TS)

- Task-dependent methods
  - Information Gain (IG)
  - Pointwise Mutual Information (PMI; just called MI in Yang&Pedersen)
  - χ^2 statistic (CHI)

Empirically compared by Yang & Pedersen (1997)

A Comparative Study on Feature Selection in Text Categorization

Yiming Yang
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213-3702, USA
yiming@cs.cmu.edu

Jan O. Pedersen
Verity, Inc.
894 Ross Dr.
Sunnyvale, CA 94089, USA
jpederse@verity.com

- Compared feature selection methods for text categorization
  - 5 feature selection methods:
    - DF, PMI, CHI, IG, TS
    - Features were just words
  - 2 classifiers:
    - kNN: *k*-Nearest Neighbour (to be covered next week)
    - LLSF: Linear Least Squares Fit
  - 2 data collections:
    - Reuters-22173
    - OHSUMED: subset of MEDLINE (1990&1991 used)

Document Frequency (DF)

DF: number of documents a term appears in

- Based on Zipf's Law
- Remove the **rare** terms: (met 1-2 times)
  - Non-informative
  - Unreliable – can be just noise
  - Not influential in the final decision
  - Unlikely to appear in new documents
- Plus
  - Easy to compute
  - Task independent: do not need to know the classes
- Minus
  - Ad hoc criterion
  - Rare terms can be good discriminators (e.g. in IR)

Stop Word Removal

- Common words from a predefined list
  - Mostly from closed-class categories:
    - unlikely to have a new word added
    - include: auxiliaries, conjunctions, determiners, prepositions, pronouns, articles
  - But also some open-class words like numerals

- Bad discriminators
  - uniformly spread across all classes
  - can be safely removed from the vocabulary
    - *Is this always a good idea? (e.g. author identification)*

Examples of Frequent Words:

Most Frequent Words in Brown Corpus

	Word	Instances	% Frequency
1.	The	69970	6.8872
2.	of	36410	3.5839
3.	and	28854	2.8401
4.	to	26154	2.5744
5.	a	23363	2.2996
6.	in	21345	2.1010
7.	that	10594	1.0428
8.	is	10102	0.9943
9.	was	9815	0.9661
10.	He	9542	0.9392
11.	for	9489	0.9340
12.	it	8760	0.8623
13.	with	7290	0.7176
14.	as	7251	0.7137
15.	his	6996	0.6886
16.	on	6742	0.6636
17.	be	6376	0.6276

	Word	Instances	% Frequency
18.	at	5377	0.5293
19.	by	5307	0.5224
20.	I	5180	0.5099
21.	this	5146	0.5065
22.	had	5131	0.5050
23.	not	4610	0.4538
24.	are	4394	0.4325
25.	but	4381	0.4312
26.	from	4370	0.4301
27.	or	4207	0.4141
28.	have	3942	0.3880
29.	an	3748	0.3689
30.	they	3619	0.3562
31.	which	3561	0.3505
32.	one	3297	0.3245
33.	you	3286	0.3234
34.	were	3284	0.3232

47

Information Gain

- A measure of importance of the feature for predicting the presence of the class.
- Defined as:
  - The number of “bits of information” gained by knowing the term is present or absent
  - Based on Information Theory
- Plus:
  - sound information theory justification
- Minus:
  - computationally expensive

Information Gain (IG)

IG: number of bits of information gained by knowing the term is present or absent

$$G(t) = -\sum_{i=1}^{m} P(c_i)\log P(c_i)$$

$$+ P(t)\sum_{i=1}^{m} P(c_i \mid t)\log P(c_i \mid t)$$

$$+ P(\bar{t})\sum_{i=1}^{m} P(c_i \mid \bar{t})\log P(c_i \mid \bar{t})$$

t is the term being scored,

c_i is a class variable

Pointwise Mutual Information (PMI)

See https://en.wikipedia.org/wiki/Pointwise_mutual_information

Logarithmic version of correlation to term t with category c

$$pmi(t,c) = \log\left(\frac{P(t,c)}{P(t)P(c)}\right)$$

$$= \log\left(\frac{P(t \mid c)}{P(t)}\right)$$

$$= \log\left(\frac{P(c \mid t)}{P(c)}\right)$$

50

Using Pointwise Mutual Information

- Compute PMI for each category and then combine
  - If we want to discriminate well *across all categories*, then we need to take the expected value of PMI:

$$pmi_{avg}(t) = \sum_{i=1}^{m} P(c_i)\, pmi(t, c_i)$$

  - To discriminate well for a *single* category, we take the maximum:

$$pmi_{\max}(t) = \max_{i=1\ldots m} pmi(t, c_i)$$

Pointwise Mutual Information

- Plus
  - *pmi*(*t*,*c*) is 0, when *t* and *c* are independent
  - Sound information-theoretic interpretation

- Minus
  - Small numbers produce unreliable results
  - No weighting with frequency of a pair (t,c)

χ^2 statistic

- The most commonly used method of comparing proportions.

- **Example:** Let us measure the dependency between a term *t* and a category *c.*
  - the groups would be:
    - 1) the documents from a category c_i
    - 2) all other documents
  - the characteristic would be:
    - "document contains term *t*"

χ^2 statistic

Is "jaguar" a good predictor for the "auto" class?

	Term = jaguar	*Term ≠ jaguar*
Class = auto	2	500
Class ≠ auto	3	9500

We want to compare:

- the observed distribution above; and
- null hypothesis: that *jaguar* and *auto* are independent

54

χ^2 statistic

Under the null hypothesis: (*jaguar* and *auto* – independent):

How many co-occurrences of *jaguar* and *auto* do we expect?

- We would have: $P(j,a) = P(j)\ P(a)$
- $P(j) = (2+3)/N$; $P(a) = (2+500)/N$; $N=2+3+500+9500$
- Num. co-occur. :
- $N \times P(j,a) = N \times P(j) \times P(a)$

$=N\times(5/N)\times(502/N)=2510/N=2510/10005 \approx 0.25$

	Term = jaguar	*Term ≠ jaguar*
Class = auto	2 (0.25)	500
Class ≠ auto	3	9500

55

χ^2 statistic

	Term = jaguar	*Term ≠ jaguar*
Class = auto	2 *(0.25)*	500 *(502)*
Class ≠ auto	3 *(4.75)*	9500 *(9498)*

χ^2 statistic

χ^2 is interested in $(f_o - f_e)^2/f_e$ summed over all table entries:

$$\chi^2(j,a) = \sum (O-E)^2/E = (2-.25)^2/.25 + (3-4.75)^2/4.75$$
$$+ (500-502)^2/502 + (9500-9498)^2/9498 = 12.9$$

	Term = jaguar	*Term ≠ jaguar*
Class = auto	2 *(0.25)*	500 *(502)*
Class ≠ auto	3 *(4.75)*	9500 *(9498)*

χ^2 statistic

Alternatives:

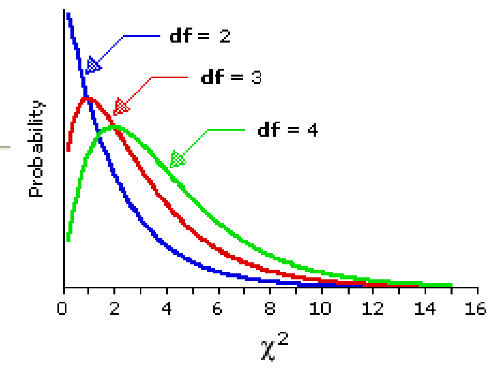
- Look up value for χ^2 in a table
- Calculate it from

$$f(x,k) = \frac{(1/2)^{k/2}}{\Gamma(k/2)} x^{k/2-1} e^{-x/2}$$

- Look it up on the internet

Chi-Square to P Calculator

For values of df between 1 and 20, inclusive, this section will calculate the proportion of the relevant sampling distribution that falls to the right of a particular value of chi-square. To proceed, enter the values of chi-square and df in the designated cells and click «Calculate».



Chi-Square	df	P
12.9	1	0.0003
Reset	Calculate	

Click here to see the details of the sampling distribution to which any particular value of chi-square belongs. At the prompt, enter the appropriate value of df.

Return to Top

The null hypothesis is rejected with confidence 0.9997

t to P Calculator

This section will calculate the one-tail and two-tail probabilities of t for any

χ^2 statistic

Collect all the terms to calculate χ^2 directly from contingency table

$$\chi^2(t,c) = \frac{N(AD - CB)^2}{(A+B)(A+C)(B+D)(C+D)}$$

$A = \#(t,c)$	$C = \#(\neg t,c)$
$B = \#(t,\neg c)$	$D = \#(\neg t, \neg c)$

$$N = A + B + C + D$$

χ^2 statistic

How to use χ^2 for multiple categories?

Compute χ^2 for each category and then combine:

- we can require to discriminate well across all categories, then we need to take the expected value of χ^2:

$$\chi^2_{avg}(t) = \sum_{i=1}^{m} P(c_i)\chi^2(t, c_i)$$

- or to discriminate well for a single category, we take the maximum:

$$\chi^2_{\max}(t) = \max_{i=1\ldots m} \chi^2(t, c_i)$$

χ^2 statistic

- Plus
  - normalized and thus comparable across terms
  - $\chi^2(t,c)$ is 0, when *t* and *c* are independent
  - sound theoretical background
- Minus
  - unreliable for low frequency terms
  - computationally expensive

Term strength

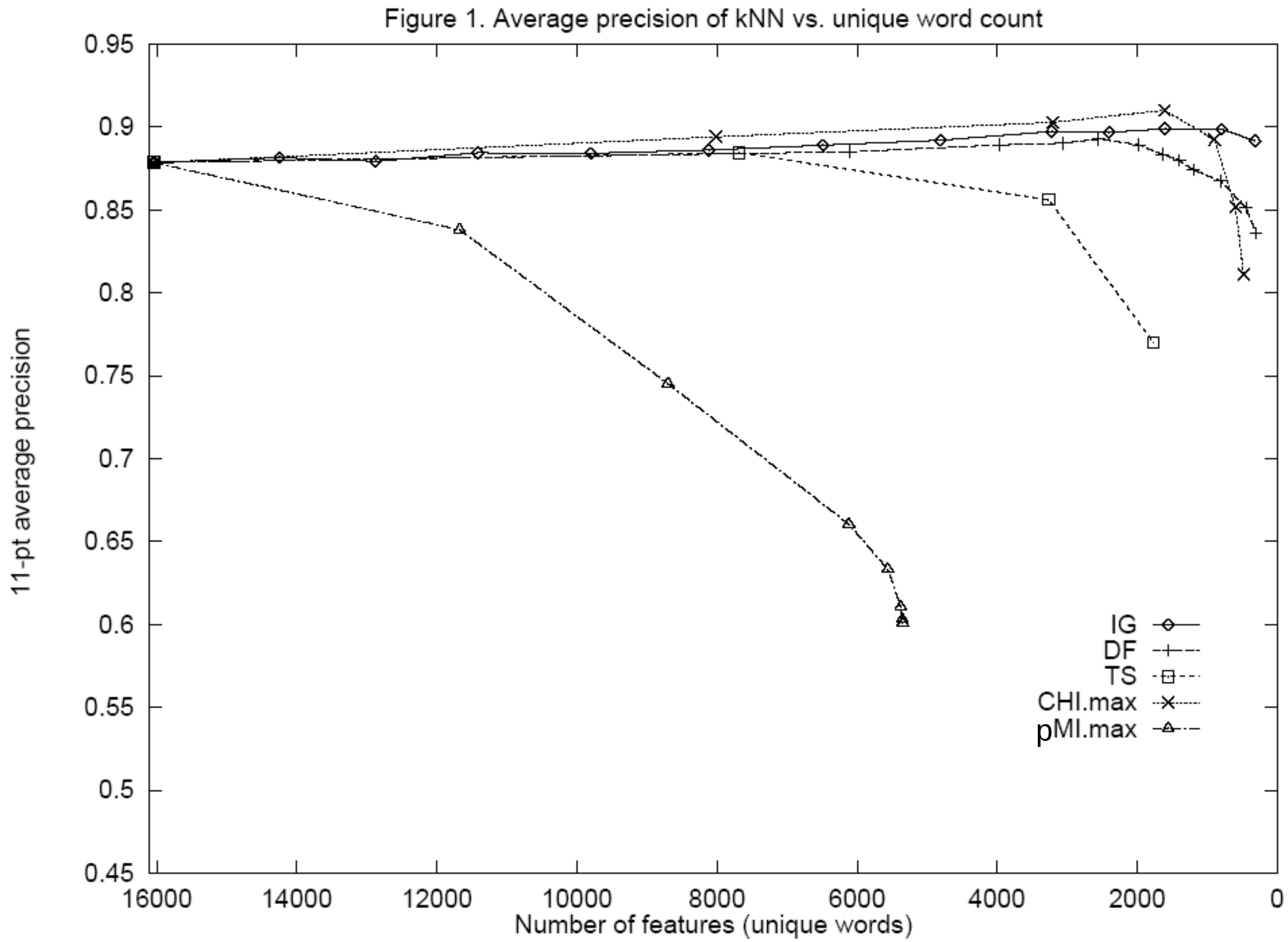
Term strength:

$$s(t) = p(t \in y \mid t \in x)$$

x,y: topically related document
(e.g. from a clustering algorithm)

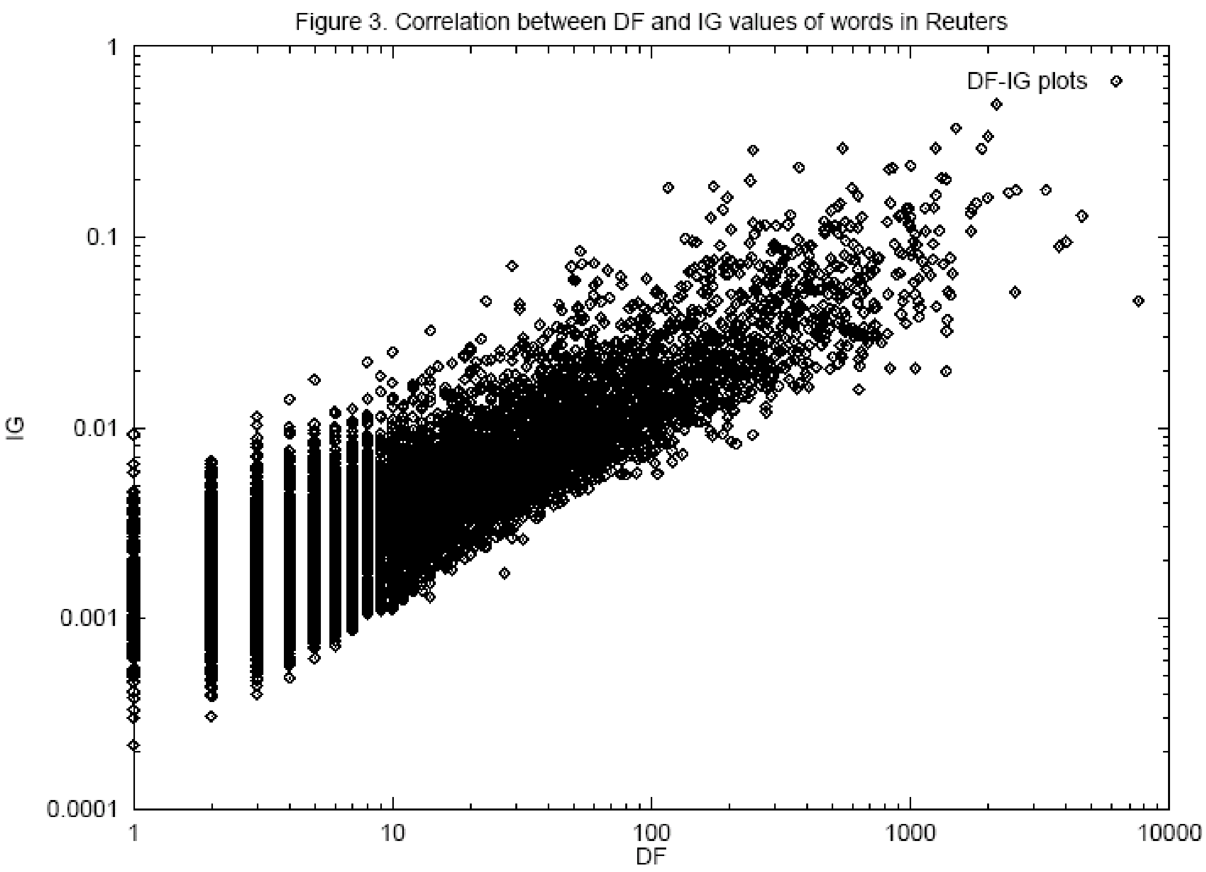
- measures co-occurrence of terms (unlike idf)
- For more details see:
  Wilbur and Sorotkin
  The automatic identification of stop words

Comparison on Reuters



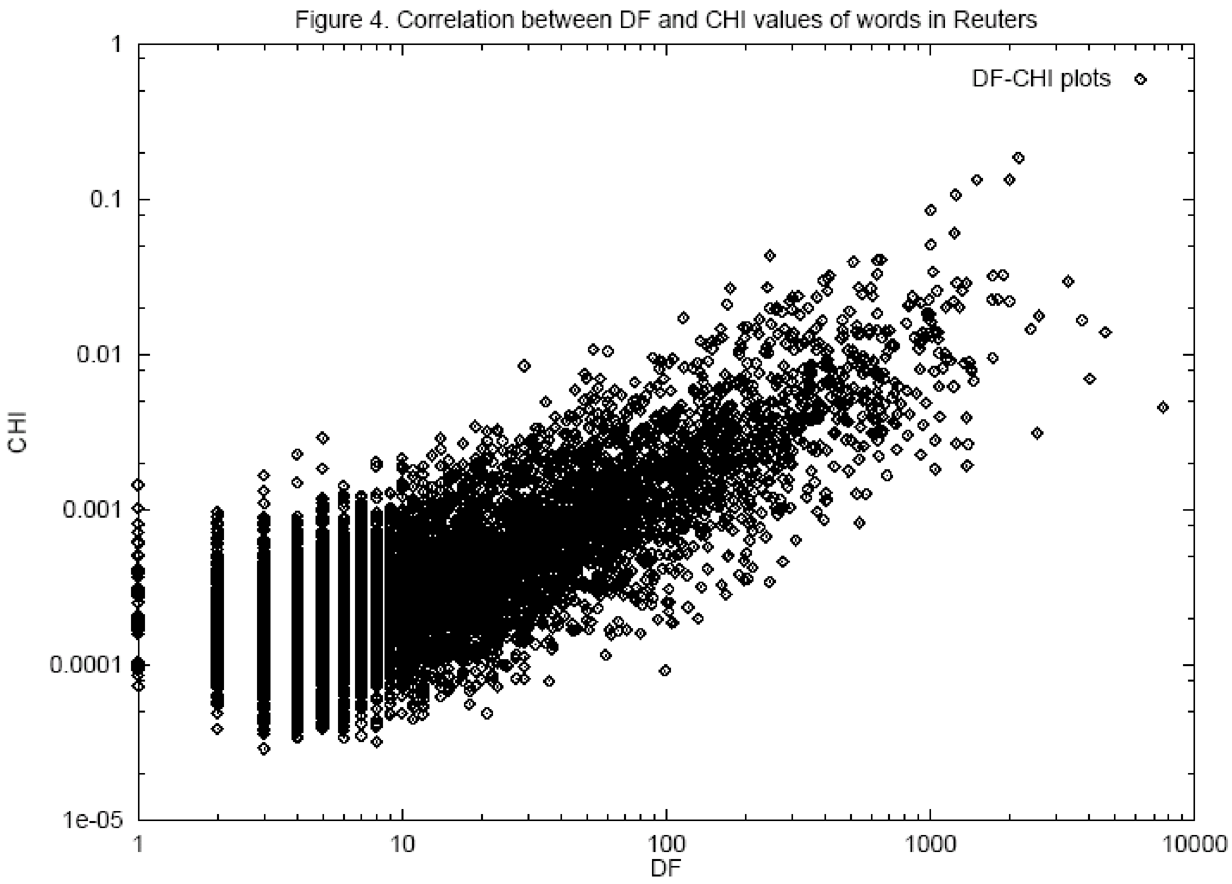
64

Correlation of feature selection criteria



65

Correlation of feature selection criteria



Feature Selection Summary (From Yang and Pedersen)

Table 1. Criteria and performance of feature selection methods in kNN & LLSF

Method	DF	IG	CHI	PMI	TS
favoring common terms	Y	Y	Y	N	Y/N
using categories	N	Y	Y	Y	N
using term absence	N	Y	Y	N	N
performance in kNN/LLSF	excellent	excellent	excellent	poor	ok

67

Das Bild kann zurzeit nicht angezeigt werden.

Classification Algorithms

Overview

- There is a large zoo of classification algorithms
  - Decision Trees
  - Naïve Bayes
    - Maximum Entropy methods
  - k Nearest Neighbor Classifiers
  - Neural networks
  - Support Vector Machines
- Many of them have been covered in other lectures

Decision Tree for Reuter classification

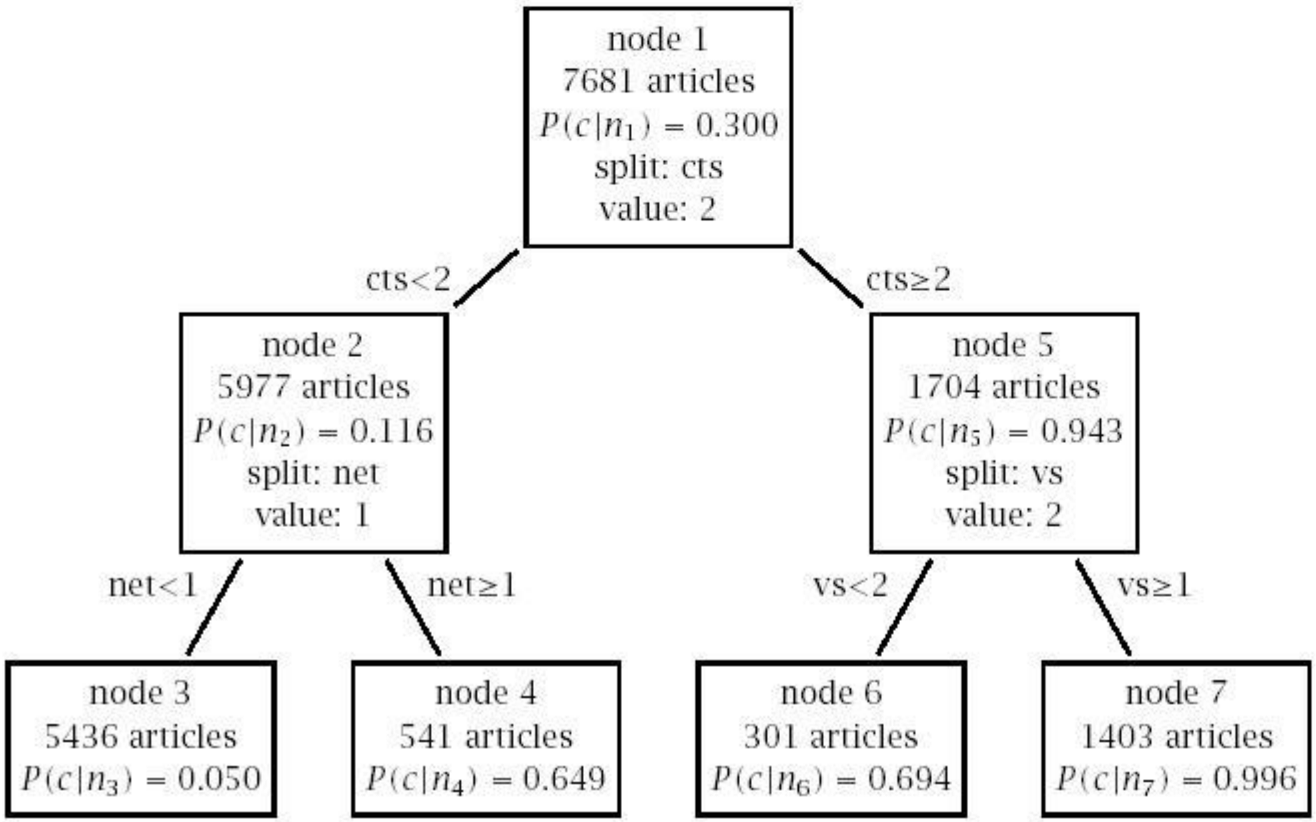
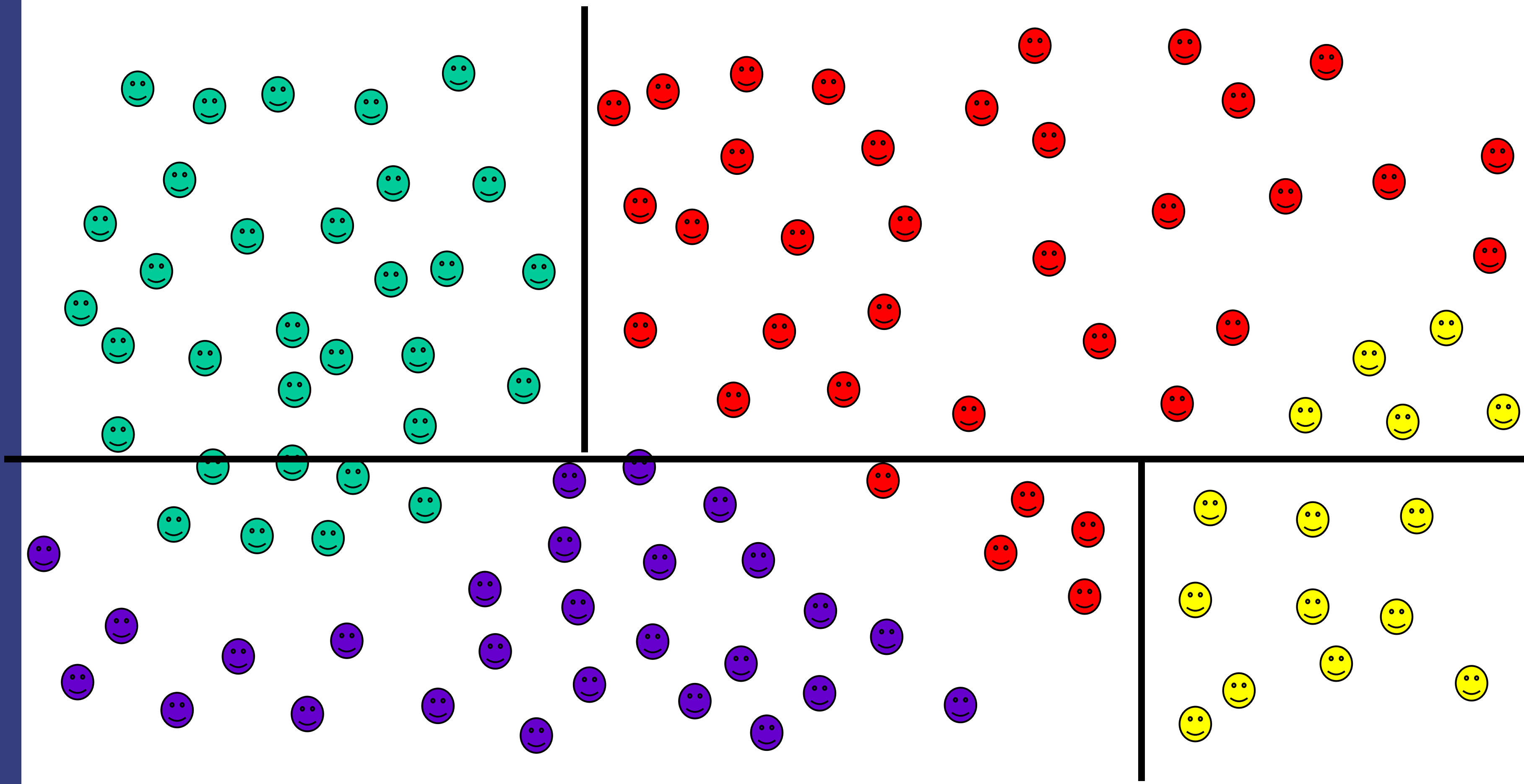


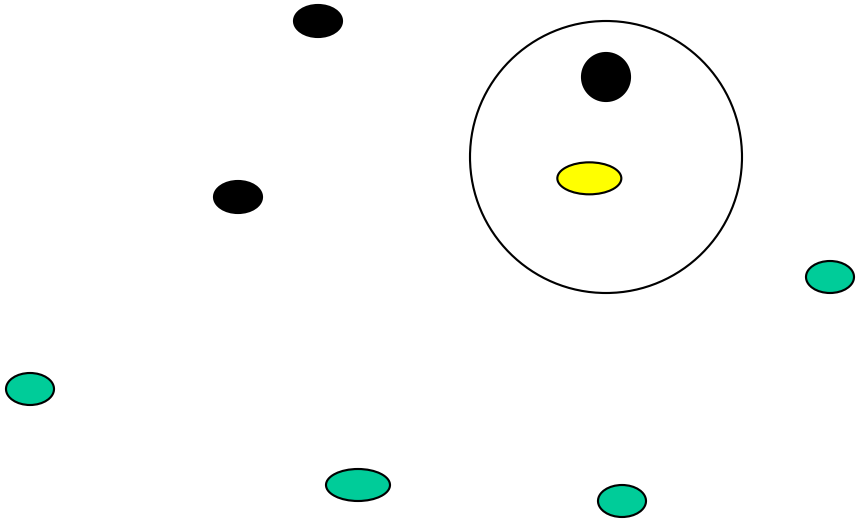
Figure 16.1 A decision tree. This tree determines whether a document is part of the topic category "earnings" or not. $P(c|n_i)$ is the probability of a document at node n_i to belong to the "earnings" category c.

From Manning&Schütze 70

Decision Boundaries for Decision Trees

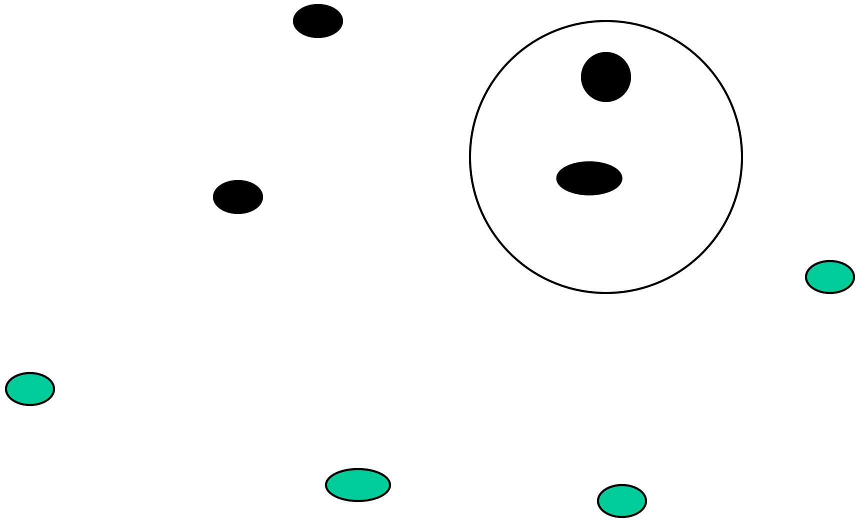


1-Nearest Neighbor

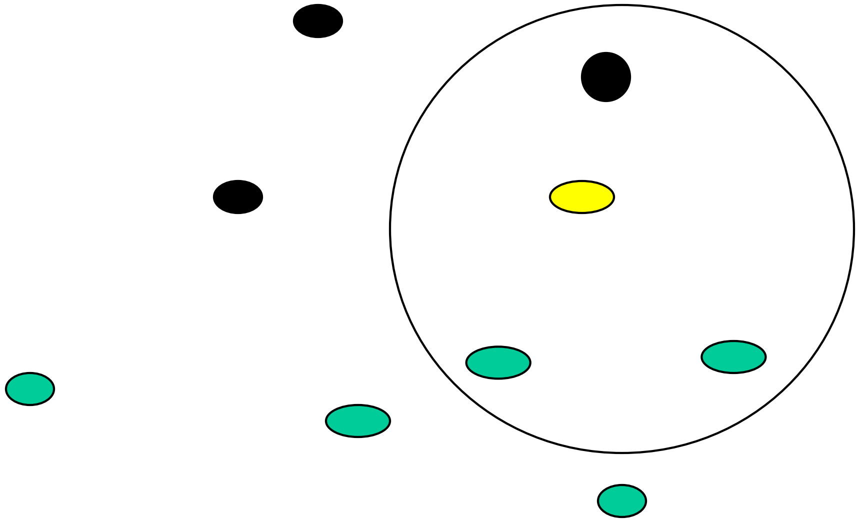


72

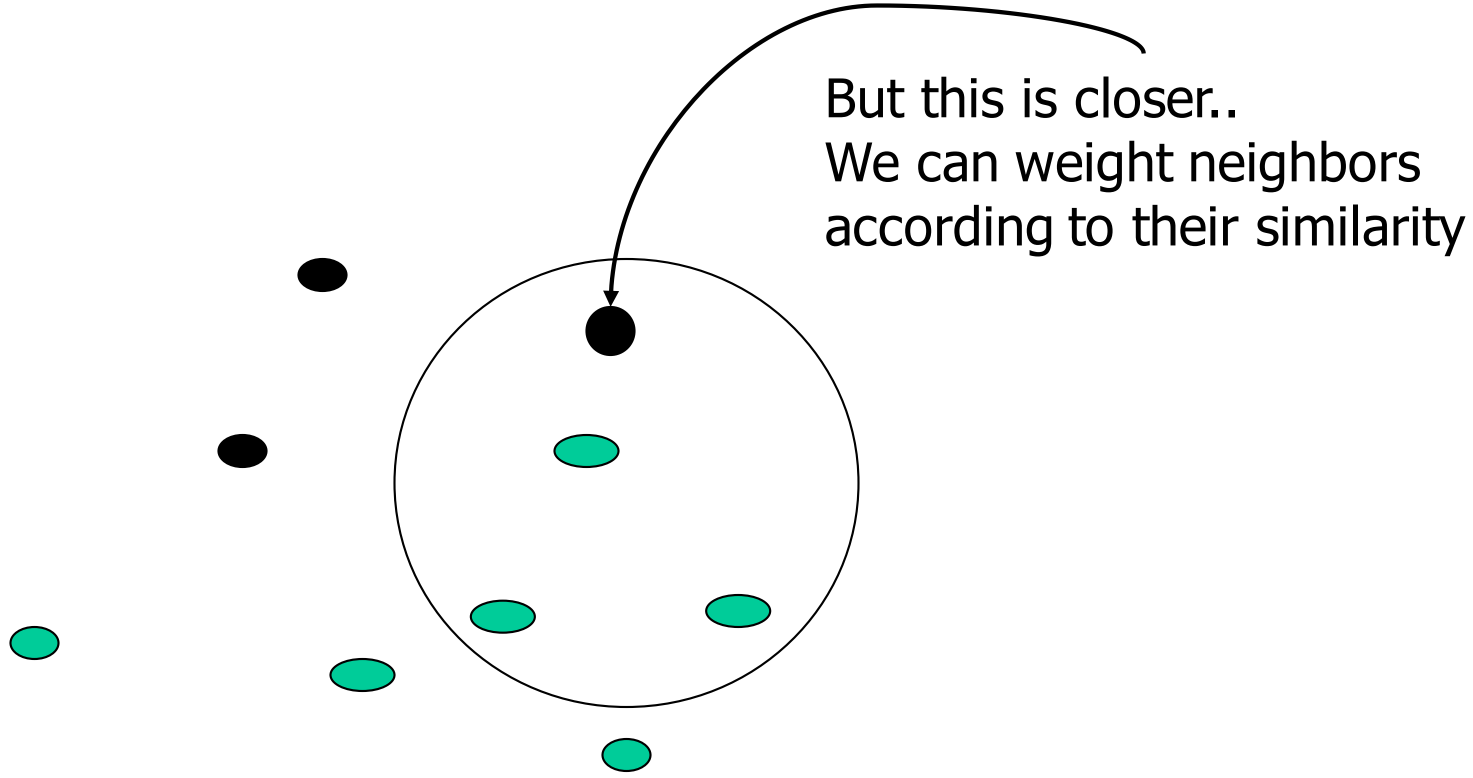
1-Nearest Neighbor



3-Nearest Neighbor



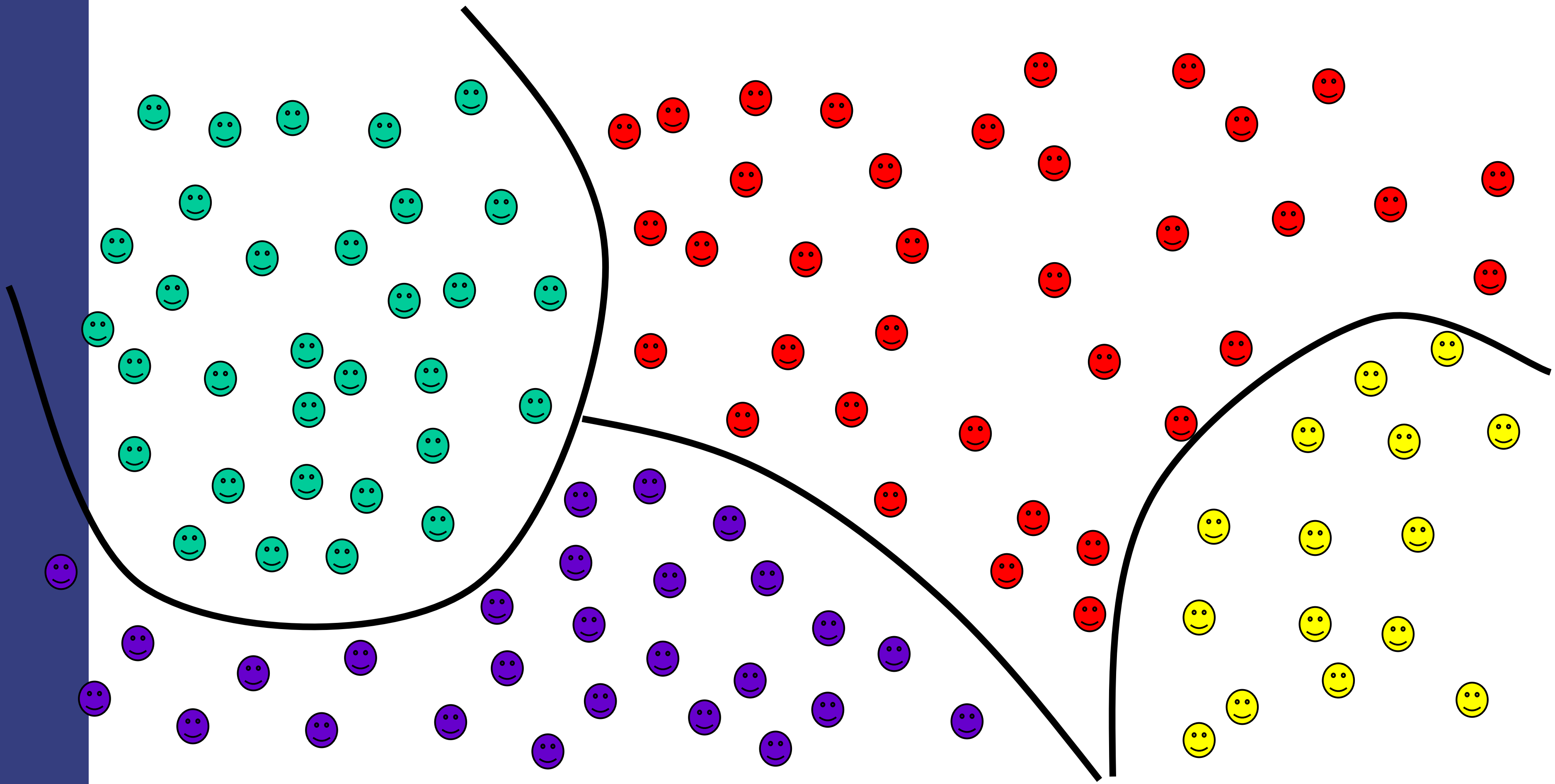
3-Nearest Neighbor



Assign the category of the majority of the neighbors

75

Decision Boundaries for *k* Nearest Neighbor (schematic)



Bayes Decision Rule

$$\overline{\omega_k} = \arg\max_{\omega_k} \left[P(x \mid \omega_k) P(\omega_k) \right]$$

ω_k: class label

x: features

Naïve Bayes

- x is not a single feature, but a bag of features

  e.g. different key-words for your spam-mail detection system

- Assume statistical independence of features

$$P(\{x_1 \ldots x_N\} \mid \omega_k) \approx \prod_{i=1}^{N} P(x_i \mid \omega_k)$$

Maximum Entropy Methods

- A way to estimate probabilities
- Features are taken into account as constraints for the probabilities
- Otherwise as “unbiased” probability estimate as possible

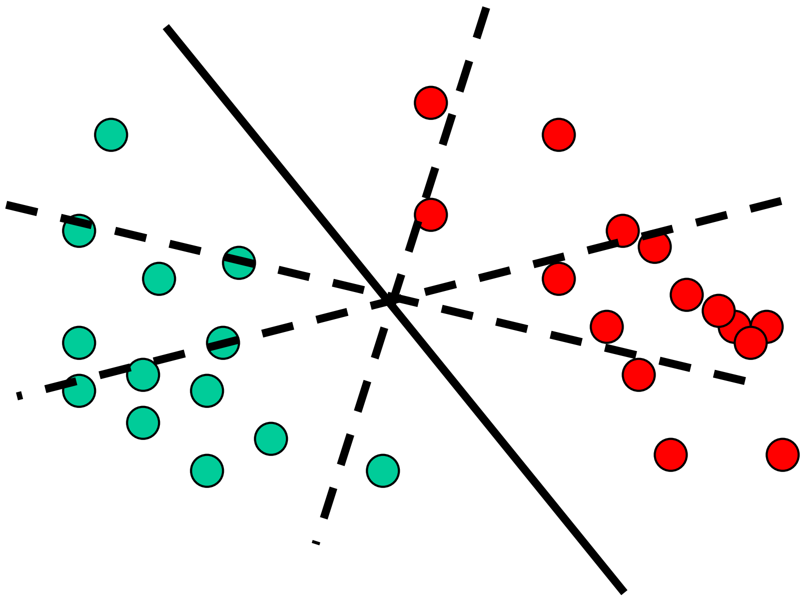
Linear binary classification using a Perceptron (Simplest Neural Network)

- Data: $\{(x_i,y_i)\}_{i=1\ldots n}$
  - x in R^d (x is a vector in d-dimensional space)
    → feature vector
  - y in {-1,+1}
    → label (class, category)
- Question:
  - Design a linear decision boundary: **wx + b** (equation of hyperplane) such that the classification rule associated with it has minimal probability of error
  - **classification rule**:
    - **y = sign(w x + b)** which means:
    - if wx + b > 0 then y = +1
    - if wx + b < 0 then y = -1

Linear binary classification

- Find a good hyperplane
  (w,b) in R^{d+1}
  that correctly classifies data points as much as possible

- In online fashion: one data point at the time, update weights as necessary



wx + b = 0

Classification Rule:
$y = \mathrm{sign}(wx + b)$

81

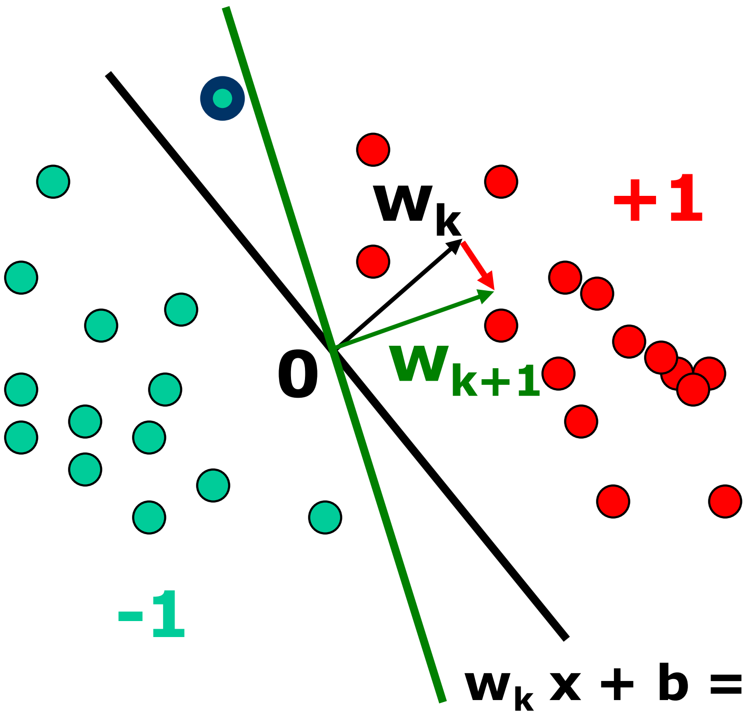
Perceptron algorithm

- Initialize: $w_1 = 0$
- Updating rule For each data point x
  - If class(x) != decision(x,w)
  - then

    $w_{k+1} \leftarrow w_k + y_i x_i$

    $k \leftarrow k + 1$
  - else

    $w_{k+1} \leftarrow w_k$



$w_k\, x + b = 0$

$w_{k+1}\, x + b = 0$

- Function **decision(x, w)**
  - If wx + b > 0 return +1
  - Else return -1

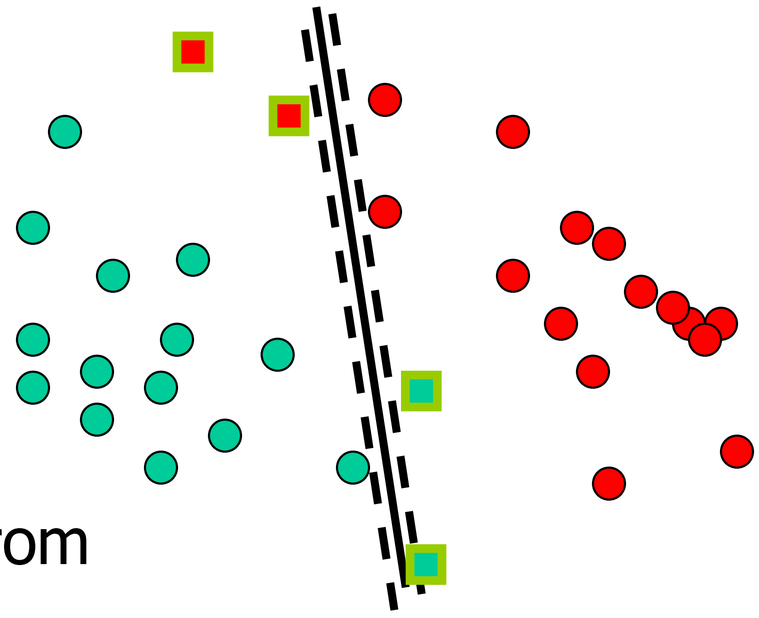
Drawing does not correspond to algorithm with respect to the treatment of b

Perceptron algorithm

- Online: can adjust to changing target, over time
- Advantages
  - Simple and computationally efficient
  - Guaranteed to learn a linearly separable problem (convergence, global optimum)
- Limitations
  - Only linear separations
  - Only converges for linearly separable data
  - Not really "efficient with many features"

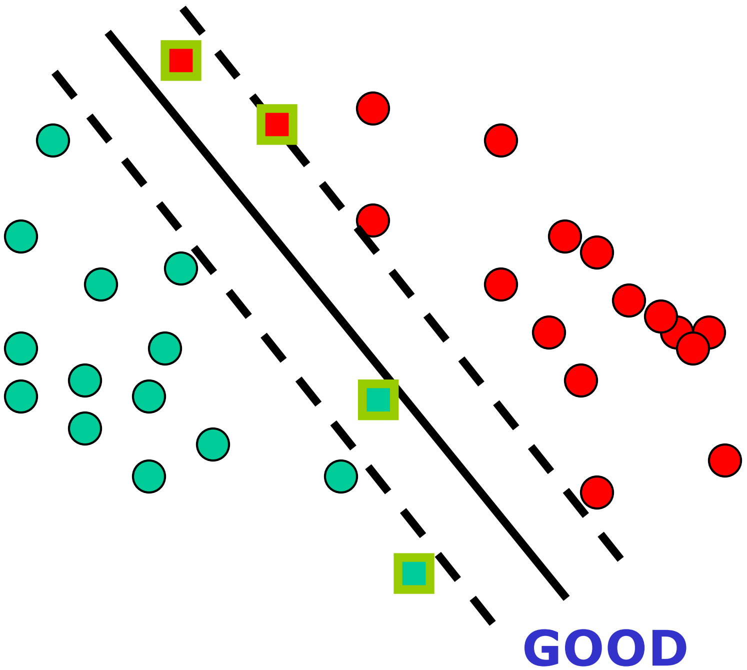
Large margin classifier

- **Another family of linear algorithms**
- **Intuition** (Vapnik, 1965)
- If the classes are linearly separable:
  - Separate the data
  - Place hyper-plane “far” from the data: **large margin**
  - Statistical results guarantee **good generalization**



Large margin classifier

- **Intuition** (Vapnik, 1965) if linearly separable:
  - Separate the data
  - Place hyperplane “far” from the data: **large margin**
  - Statistical results guarantee **good generalization**

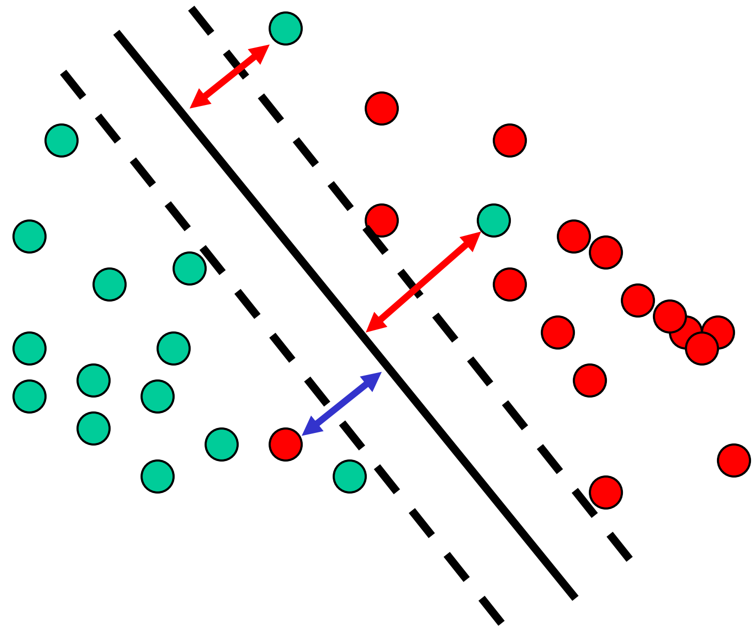


→ **Maximal Margin Classifier**

Large margin classifier

If **not linearly separable**

- **Allow** some **errors**
- Still, try to place hyperplane “far” from each class

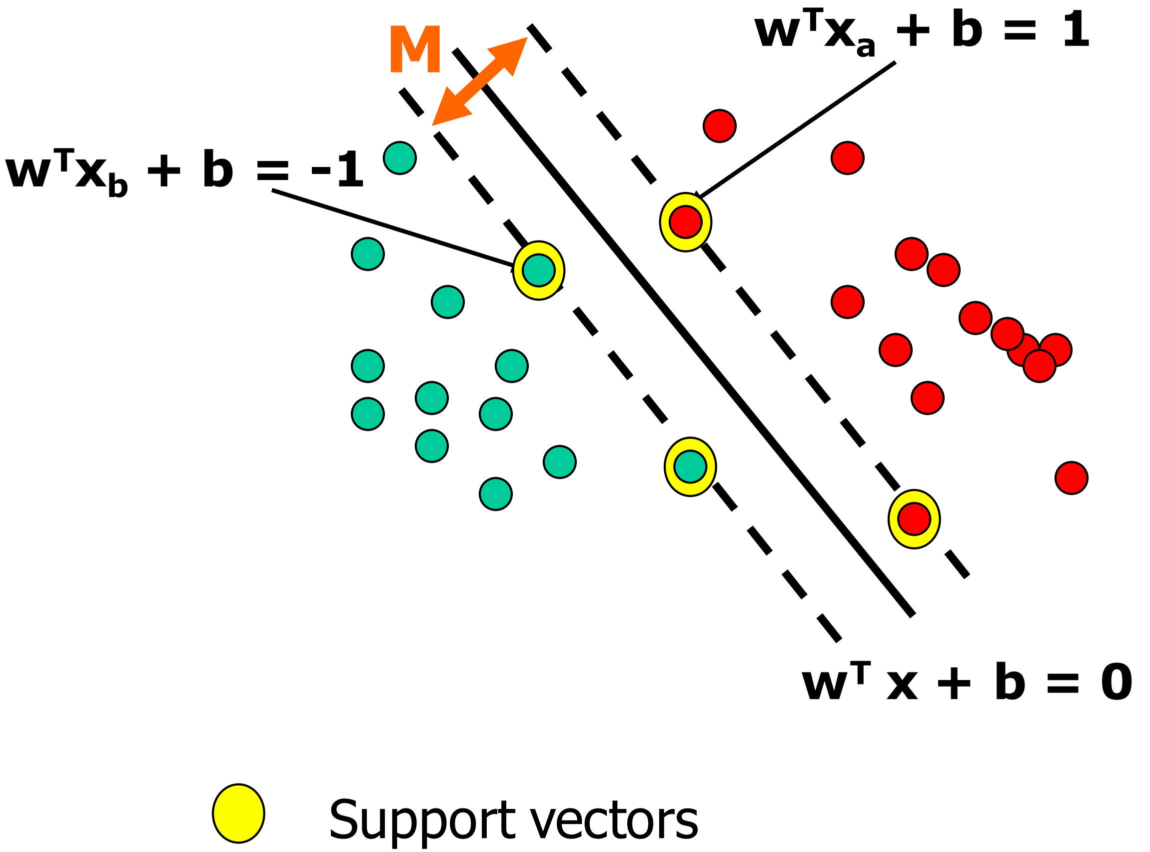


Large Margin Classifiers

- Advantages
  - Theoretically better (better error bounds)
- Limitations
  - Computationally more expensive, large quadratic programming

Support Vector Machine (SVM)

- Large Margin Classifier
- Linearly separable case
- Goal: find the hyperplane that maximizes the margin



Summary

- Types of text classification
- Features and feature selection
- Classification algorithms