COMP4901K/Math4824B
Machine Learning for Natural Language Processing

Lecture 5 and 6: Introduction to Classification

Instructor: Yangqiu Song

Slides credits: Hongning Wang, Marti Hearst, Barbara Rosario

Today

- Classification
  - Text categorization (and other applications)
- Various issues regarding classification
  - Clustering vs. classification, binary vs. multi-way, flat vs. hierarchical classification…
- Feature Selection

Classification

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

Examples:

Problem	Object	Categories
Tagging	Word	POS
Sense Disambiguation	Word	The word's senses
Information retrieval	Document	Relevant/not relevant
Sentiment classification	Document	Positive/negative
Author identification	Document	Authors

From: Foundations of Statistical Natural Language Processing. Manning and Schutze

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.

Jane Austen (1775-1817), Pride and Prejudice

or

Charles Dickens (1812-70), Bleak House ?

Language identification

- Tutti gli esseri umani nascono liberi ed eguali in dignità e diritti. Essi sono dotati di ragione e di coscienza e devono agire gli uni verso gli altri in spirito di fratellanza.

- Alle Menschen sind frei und gleich an Würde und Rechten geboren. Sie sind mit Vernunft und Gewissen begabt und sollen einander im Geist der Brüderlichkeit begegnen.

Universal Declaration of Human Rights, UN, in 363 languages

Classification

Goal: Assign ‘objects’ from a universe to two or more *classes* or *categories*

Examples:

Problem	Object	Categories
Author identification	Document	Authors
Language identification	Document	Language
Text categorization	Document	Topics

From: Foundations of Statistical Natural Language Processing. Manning and Schutze

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.

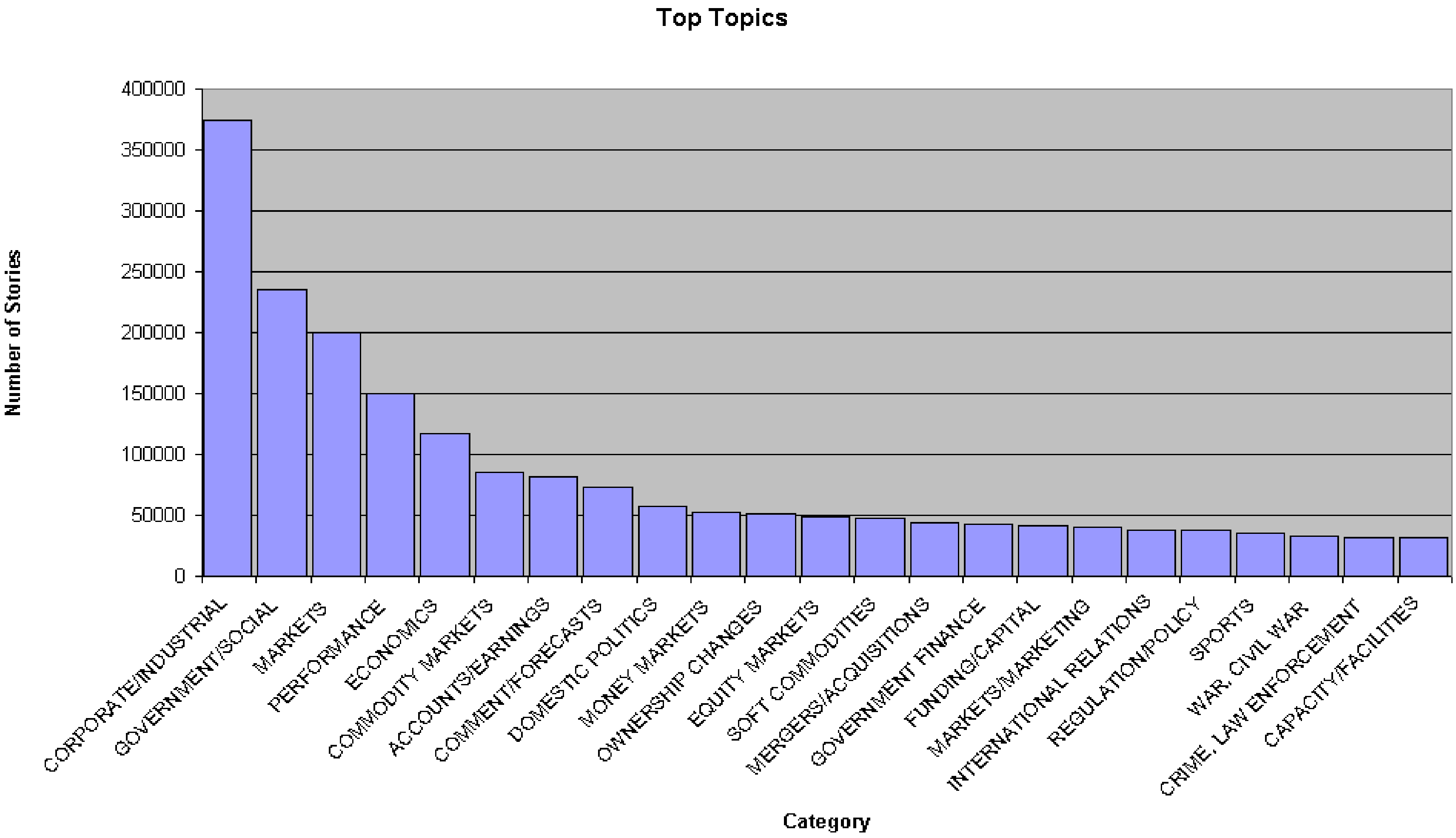
One of the strangest, most relentless hurricane seasons on record reached new bizarre heights yesterday as the plodding approach of Hurricane Jeanne prompted evacuation orders for hundreds of thousands of Floridians and high wind warnings that stretched 350 miles from the swamp towns south of Miami to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

Text categorization

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

Reuters

- 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

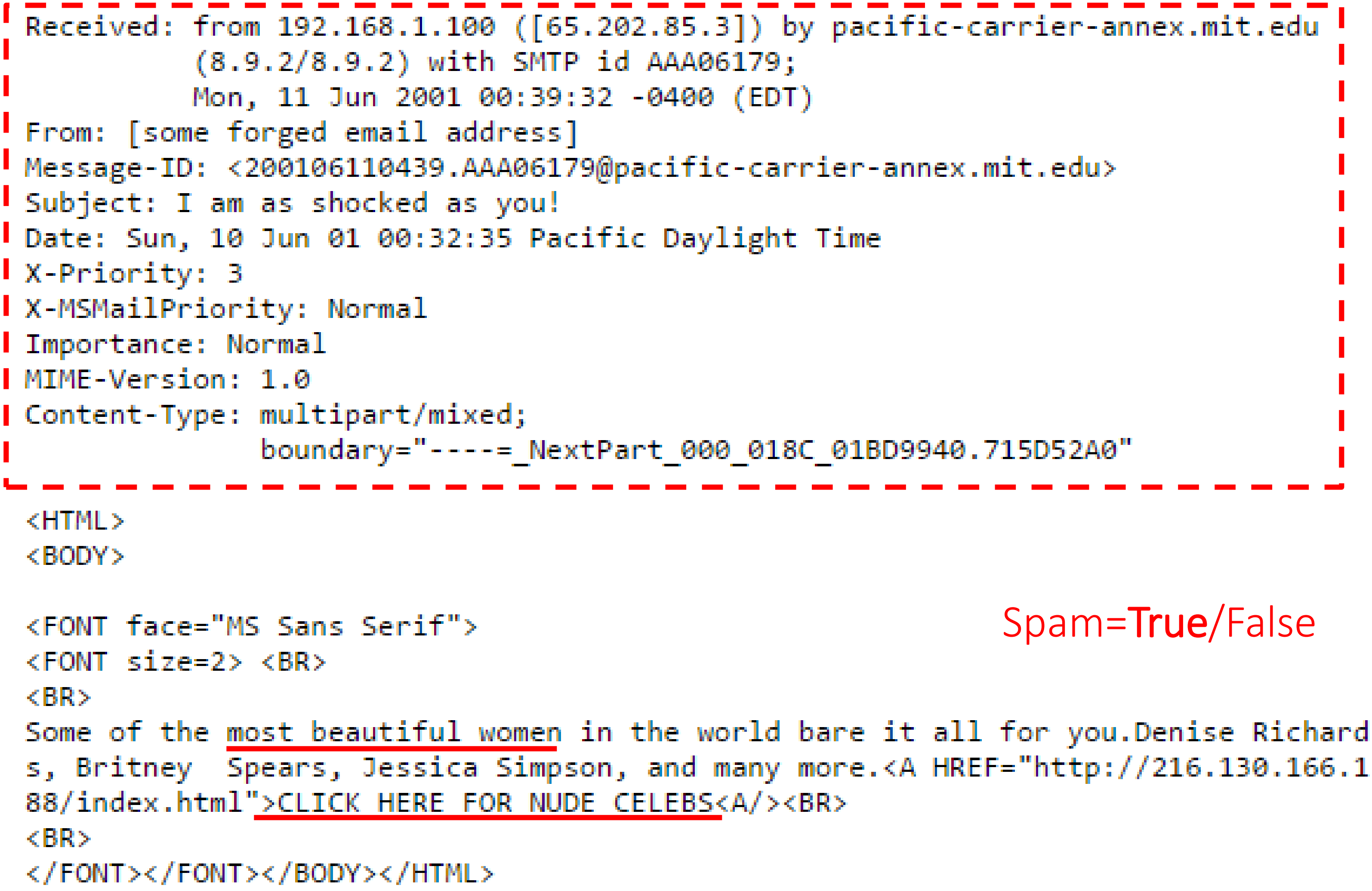
&#3;</BODY></TEXT></REUTERS>

Text categorization: examples

- Topic categorization
  - http://news.google.com/
  - Reuters.
- Spam filtering
  - Determine if a mail message is spam (or not)
- Customer service message classification

Text Categorization

- Recognizing spam emails



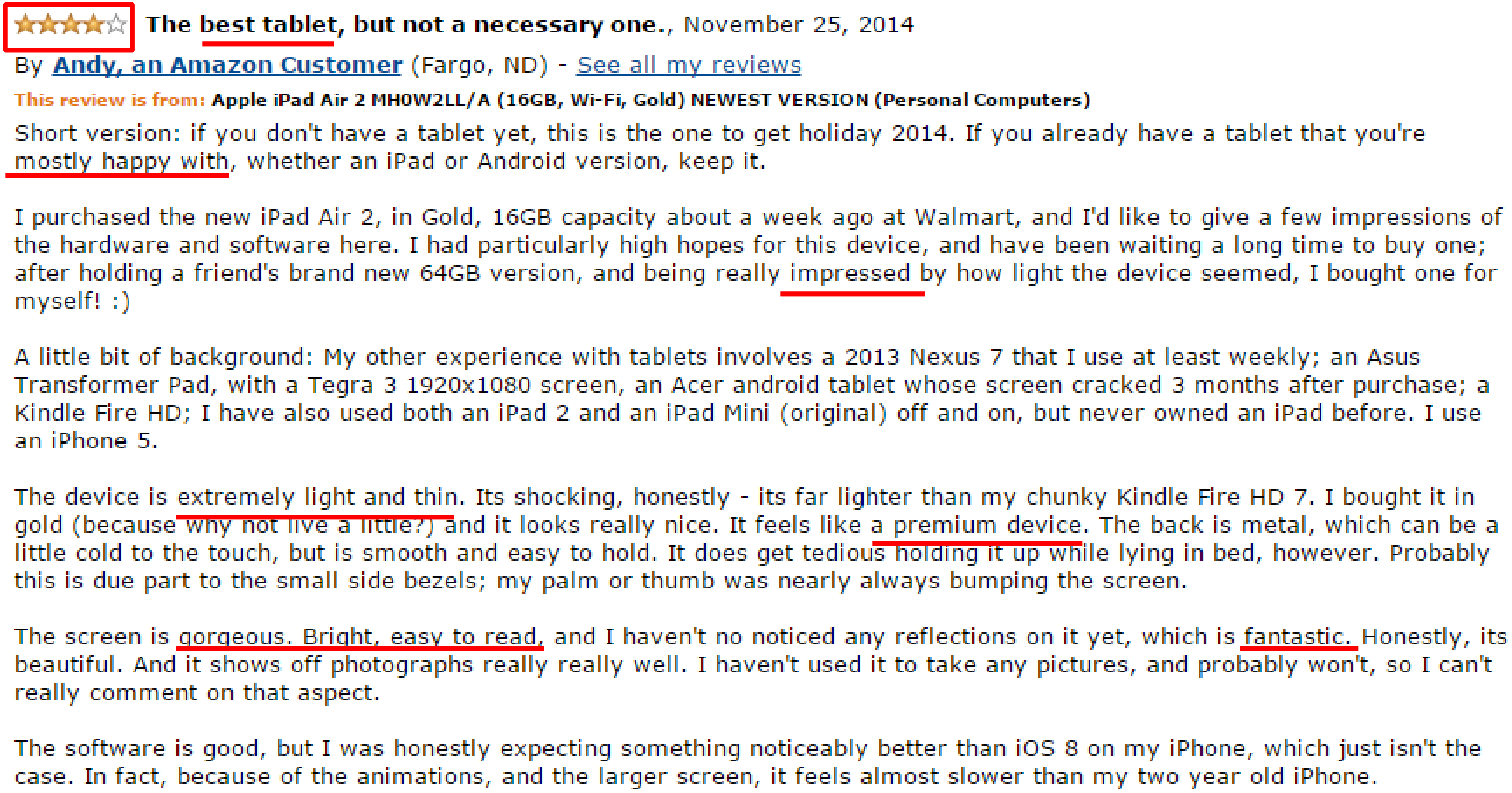
```
Received: from 192.168.1.100 ([65.202.85.3]) by pacific-carrier-annex.mit.edu
          (8.9.2/8.9.2) with SMTP id AAA06179;
          Mon, 11 Jun 2001 00:39:32 -0400 (EDT)
From: [some forged email address]
Message-ID: <200106110439.AAA06179@pacific-carrier-annex.mit.edu>
Subject: I am as shocked as you!
Date: Sun, 10 Jun 01 00:32:35 Pacific Daylight Time
X-Priority: 3
X-MSMailPriority: Normal
Importance: Normal
MIME-Version: 1.0
Content-Type: multipart/mixed;
              boundary="----=_NextPart_000_018C_01BD9940.715D52A0"
<HTML>
<BODY>

<FONT face="MS Sans Serif">
<FONT size=2> <BR>
<BR>
Some of the most beautiful women in the world bare it all for you.Denise Richard
s, Britney  Spears, Jessica Simpson, and many more.<A HREF="http://216.130.166.1
88/index.html">CLICK HERE FOR NUDE CELEBS<A/><BR>
<BR>
</FONT></FONT></BODY></HTML>
```

Spam=**True**/False

Applications of Text Categorization

- Sentiment analysis



★★★★☆ **The best tablet, but not a necessary one.**, November 25, 2014

By **Andy, an Amazon Customer** (Fargo, ND) - See all my reviews

This review is from: **Apple iPad Air 2 MH0W2LL/A (16GB, Wi-Fi, Gold) NEWEST VERSION (Personal Computers)**

Short version: if you don't have a tablet yet, this is the one to get holiday 2014. If you already have a tablet that you're mostly happy with, whether an iPad or Android version, keep it.

I purchased the new iPad Air 2, in Gold, 16GB capacity about a week ago at Walmart, and I'd like to give a few impressions of the hardware and software here. I had particularly high hopes for this device, and have been waiting a long time to buy one; after holding a friend's brand new 64GB version, and being really impressed by how light the device seemed, I bought one for myself! :)

A little bit of background: My other experience with tablets involves a 2013 Nexus 7 that I use at least weekly; an Asus Transformer Pad, with a Tegra 3 1920x1080 screen, an Acer android tablet whose screen cracked 3 months after purchase; a Kindle Fire HD; I have also used both an iPad 2 and an iPad Mini (original) off and on, but never owned an iPad before. I use an iPhone 5.

The device is extremely light and thin. Its shocking, honestly - its far lighter than my chunky Kindle Fire HD 7. I bought it in gold (because why not live a little?) and it looks really nice. It feels like a premium device. The back is metal, which can be a little cold to the touch, but is smooth and easy to hold. It does get tedious holding it up while lying in bed, however. Probably this is due part to the small side bezels; my palm or thumb was nearly always bumping the screen.

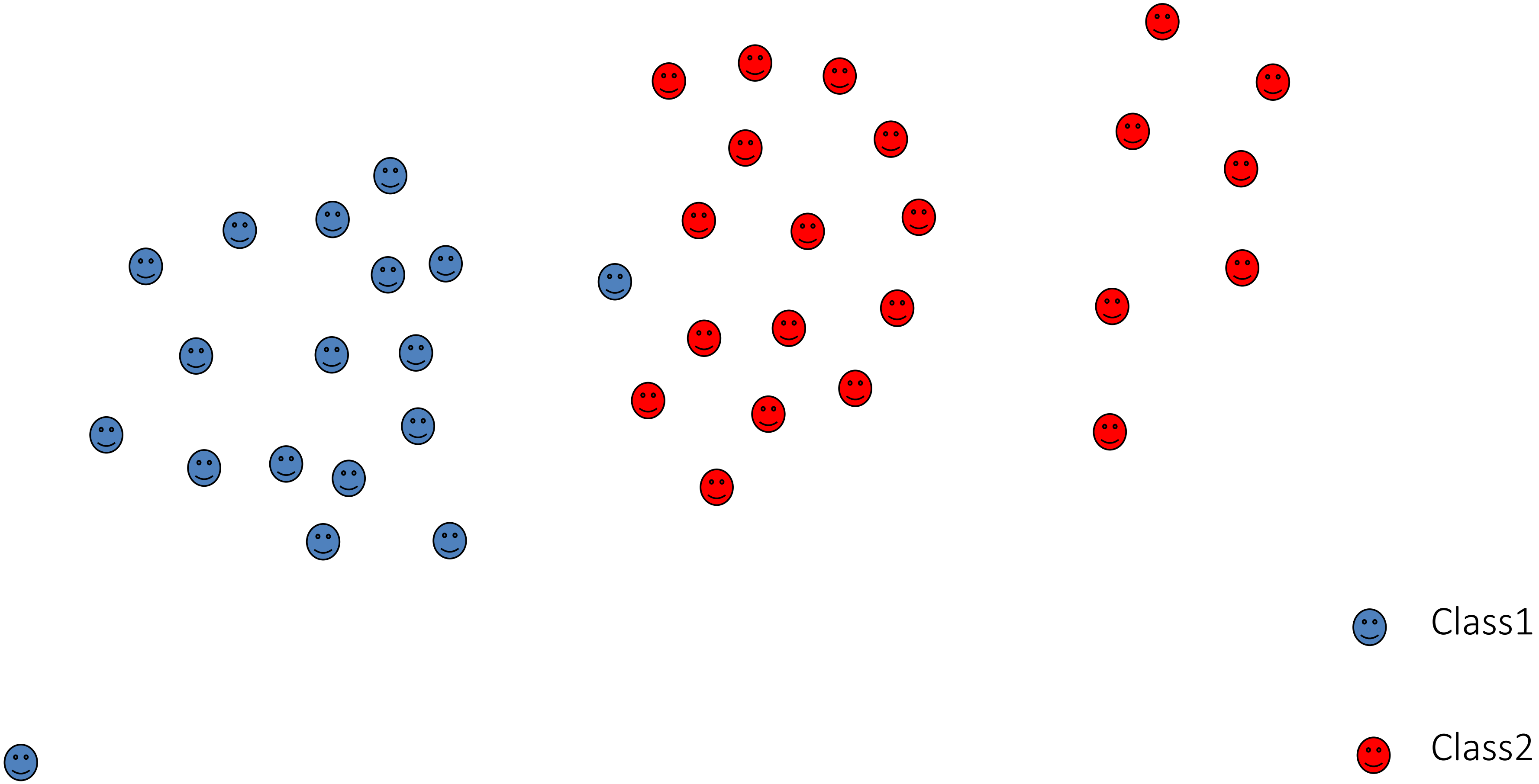
The screen is gorgeous. Bright, easy to read, and I haven't no noticed any reflections on it yet, which is fantastic. Honestly, its beautiful. And it shows off photographs really really well. I haven't used it to take any pictures, and probably won't, so I can't really comment on that aspect.

The software is good, but I was honestly expecting something noticeably better than iOS 8 on my iPhone, which just isn't the case. In fact, because of the animations, and the larger screen, it feels almost slower than my two year old iPhone.

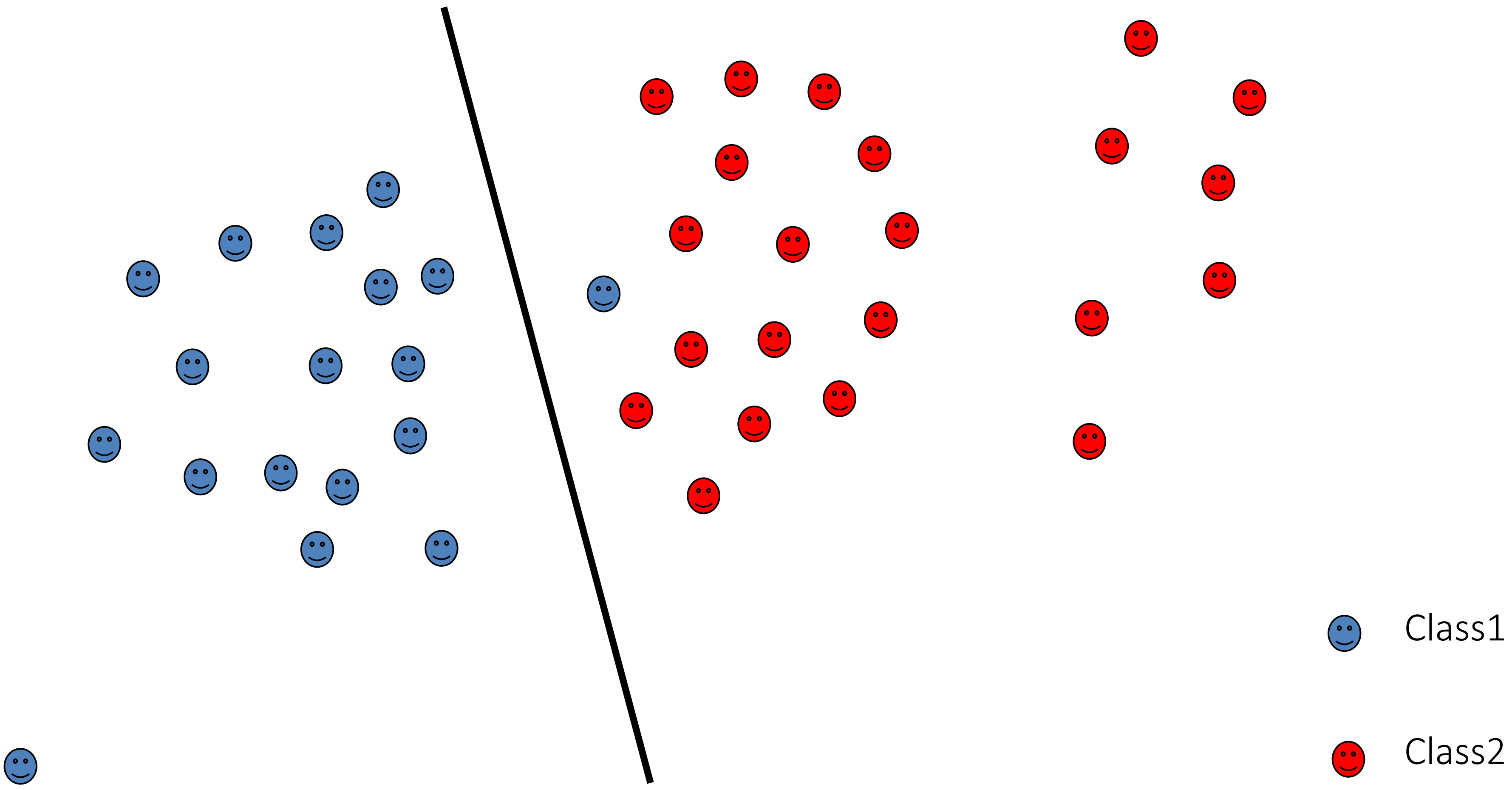
Classification vs. Clustering

- 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

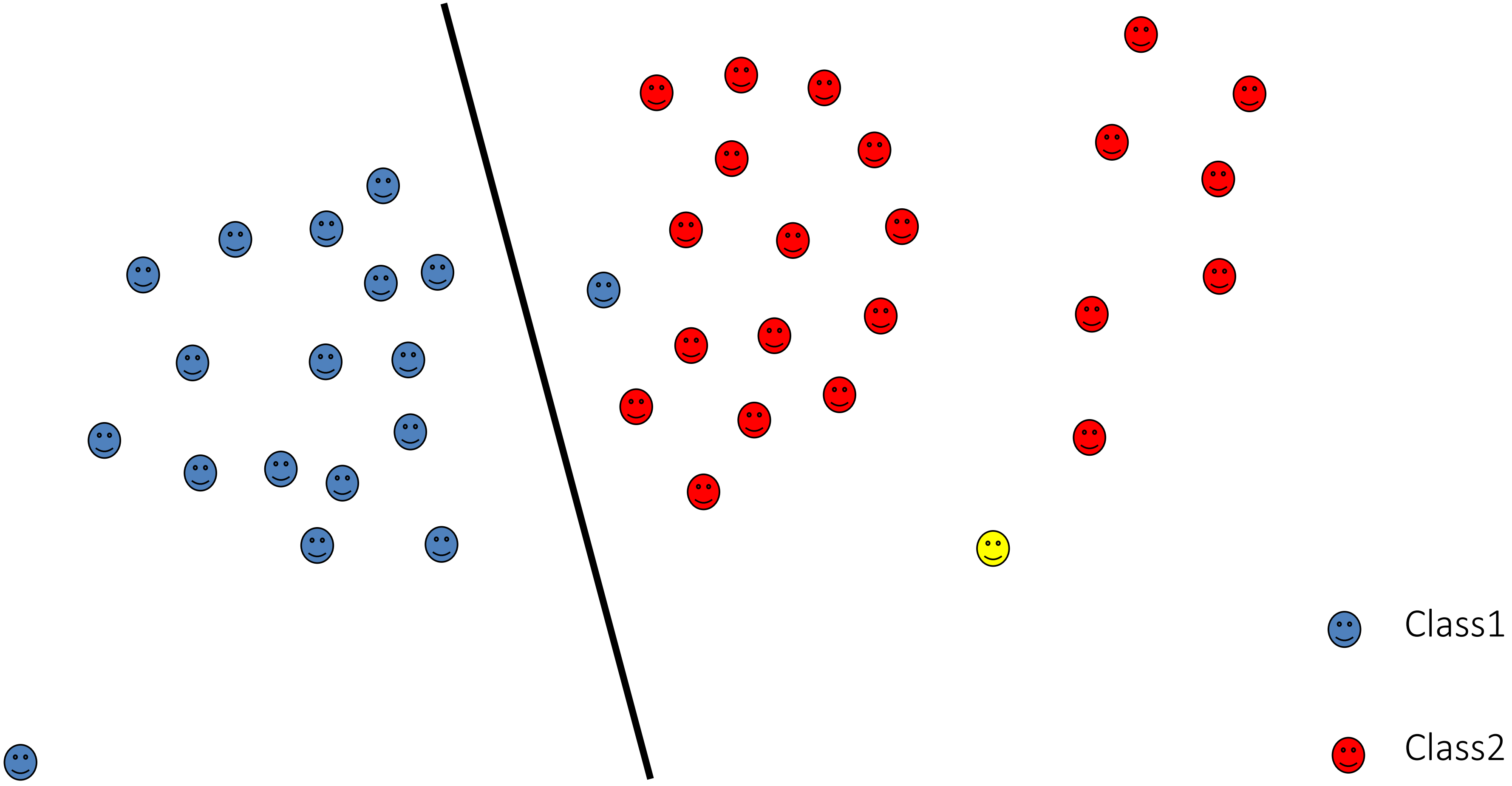
Classification



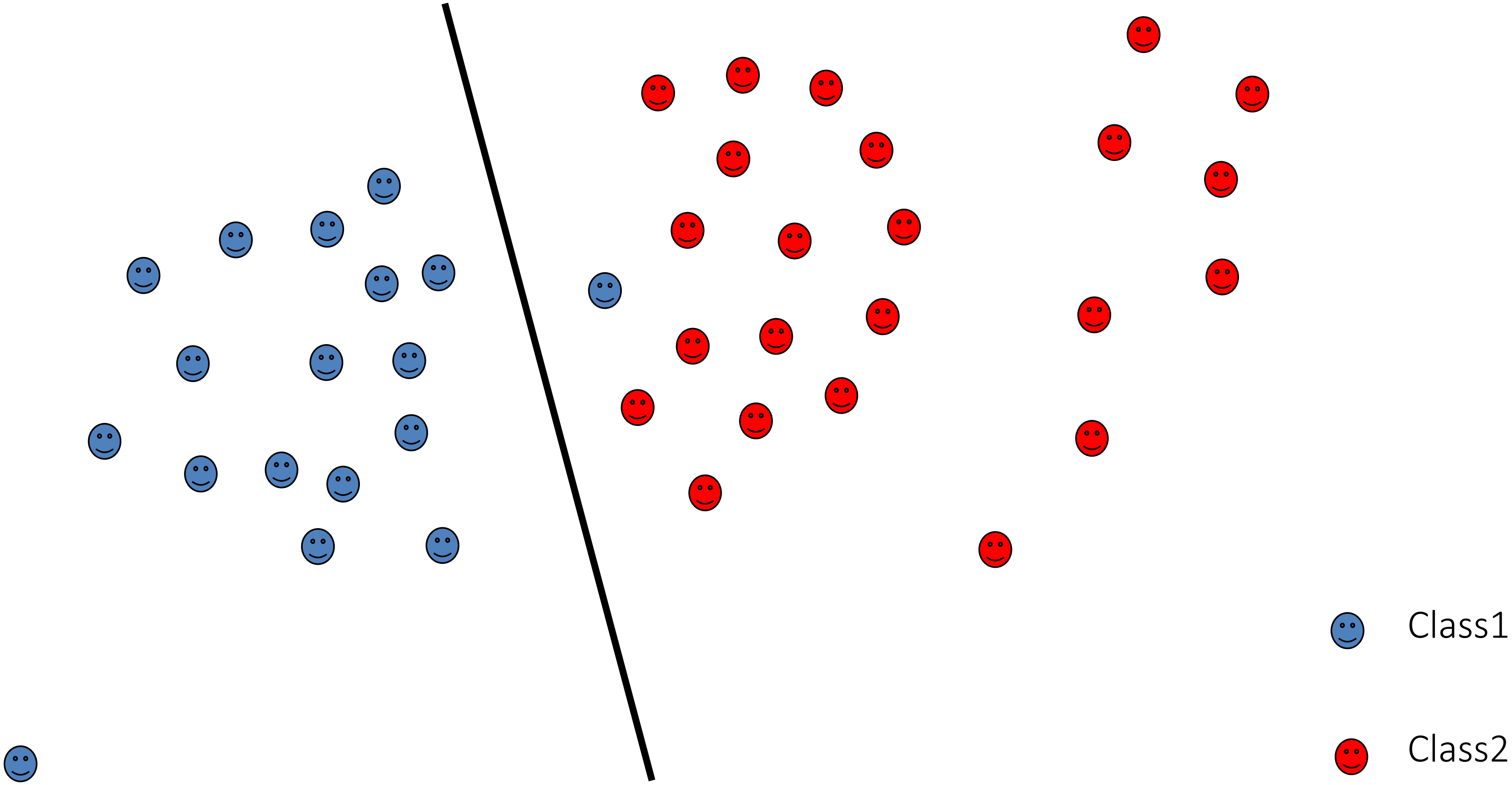
Classification



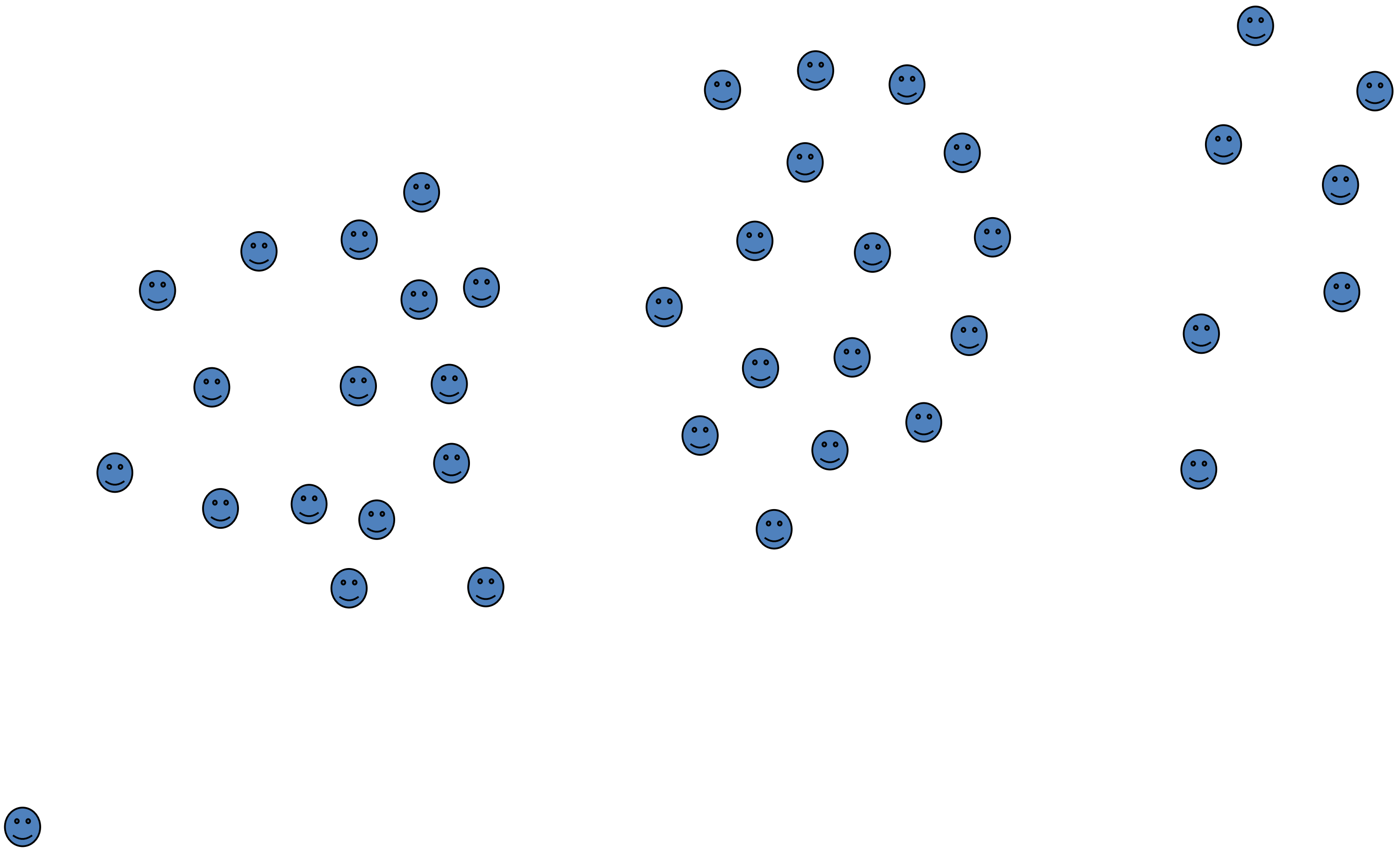
Classification



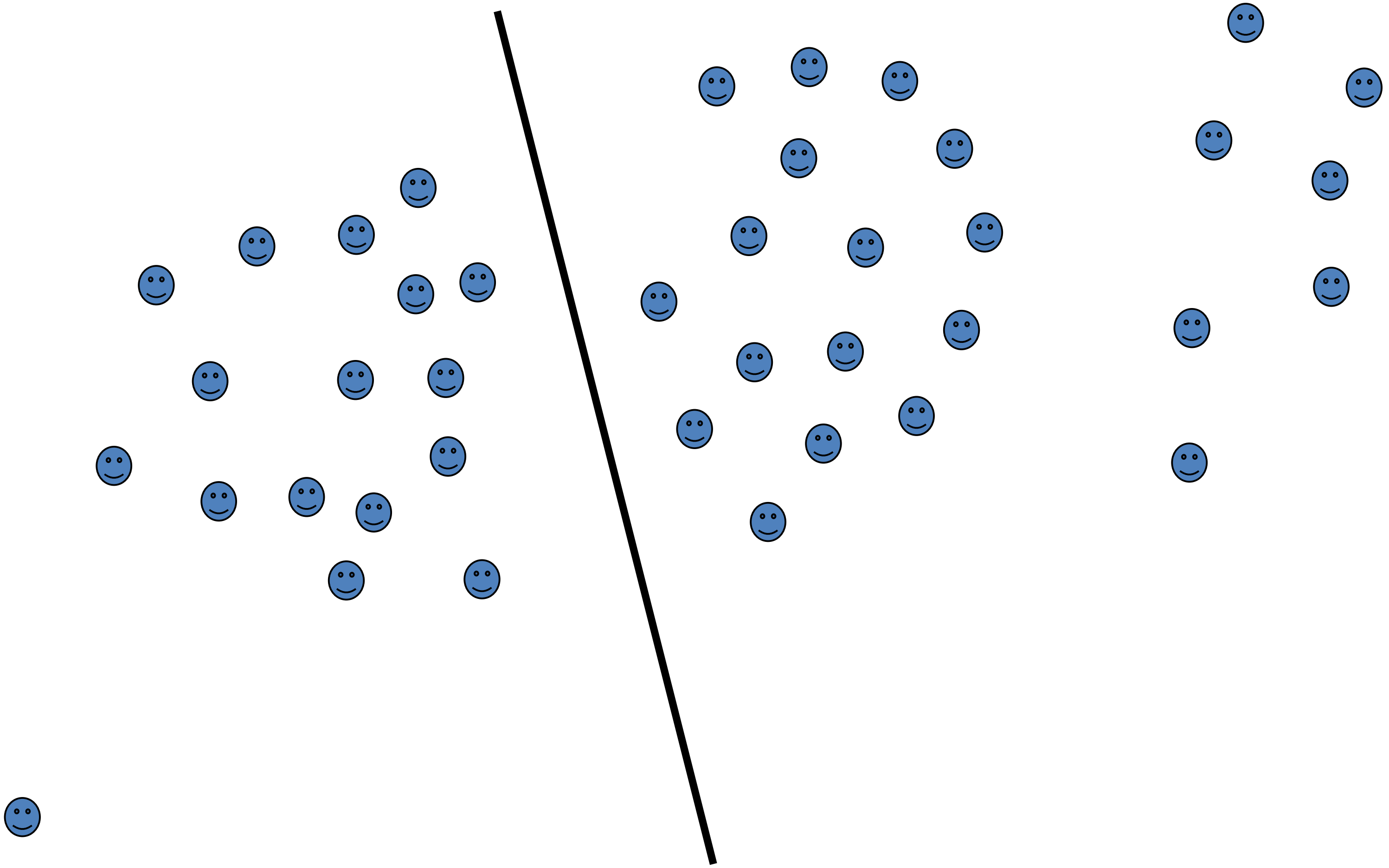
Classification



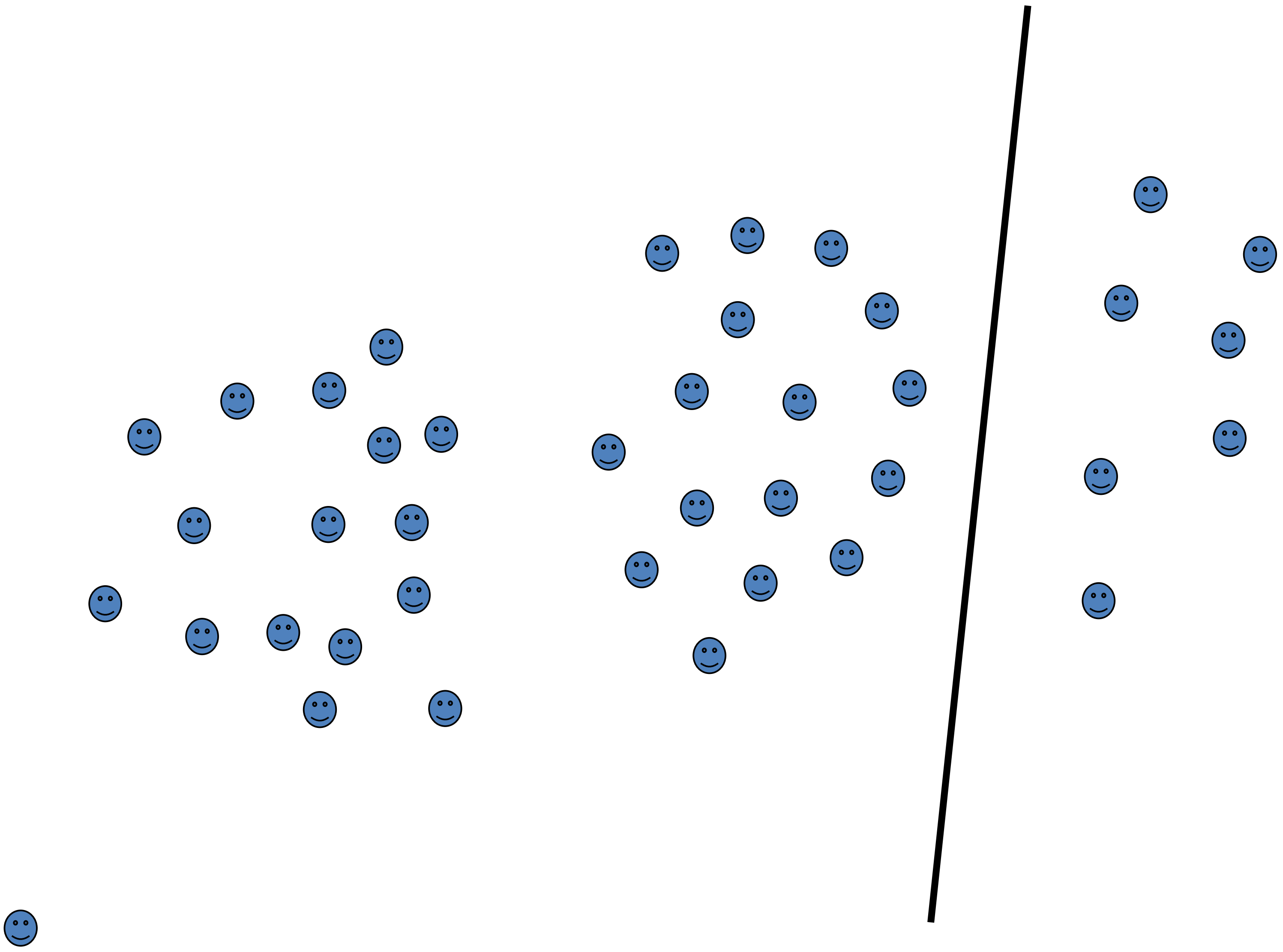
Clustering



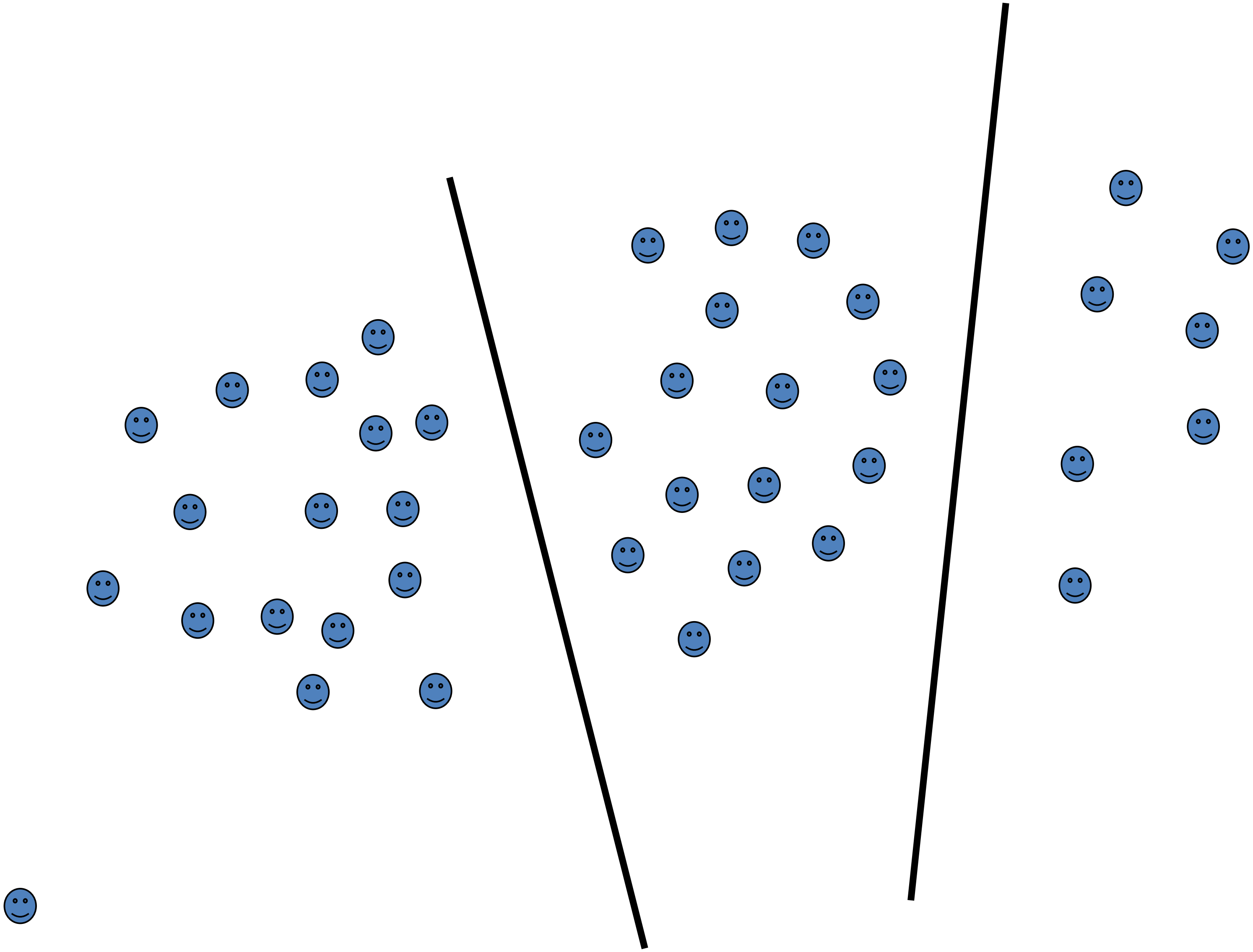
Clustering



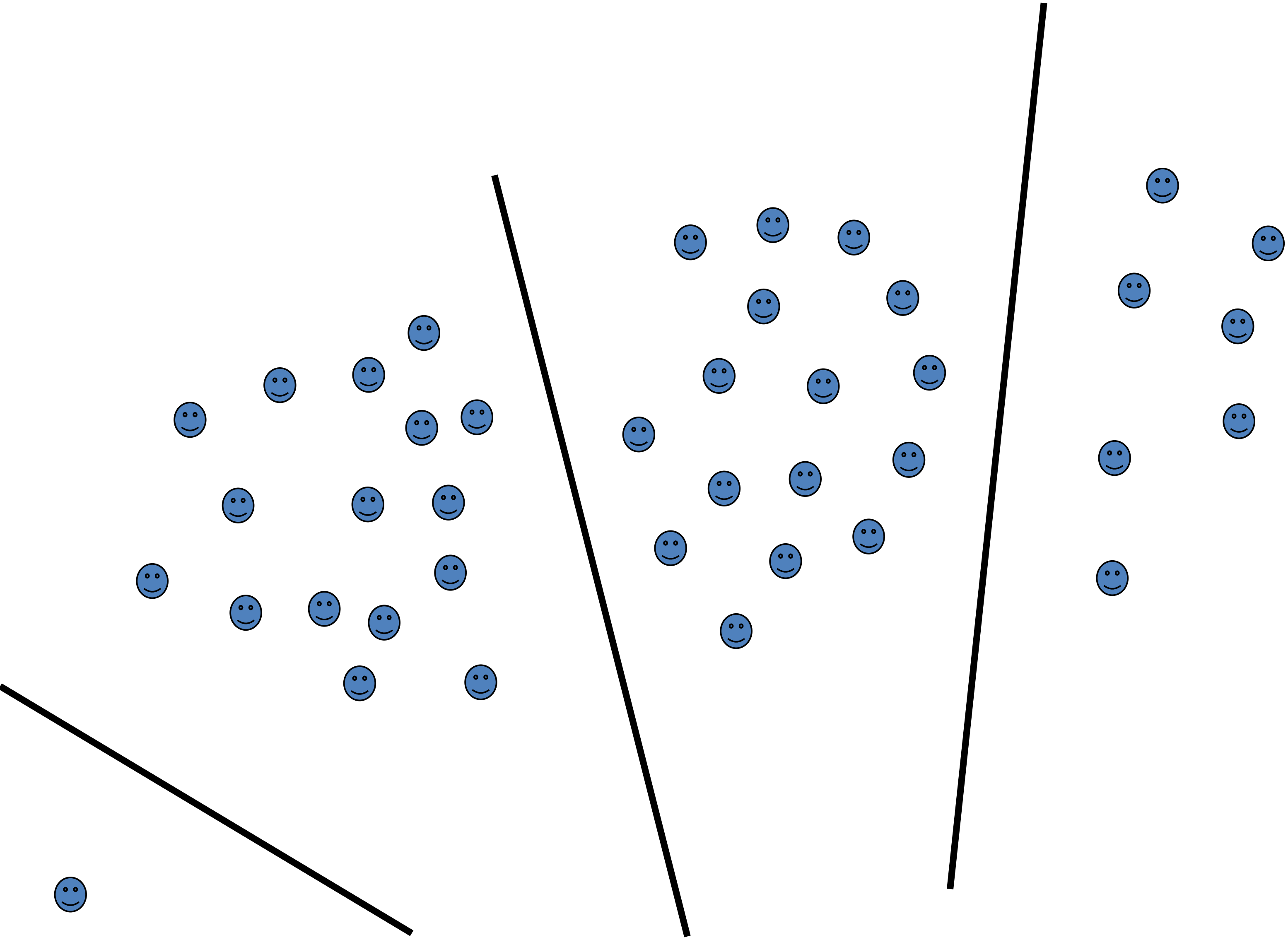
Clustering



Clustering



Clustering



Categories (Labels, Classes)

- Labeling data
- 2 problems:
- Decide the possible classes (which ones, how many)
  - Domain and application dependent
  - http://news.google.com
- Label text
  - Difficult, time consuming, inconsistency between annotators

Binary vs. multi-way classification

- Binary classification: two classes

- Multi-way classification: more than two classes

- Sometime 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

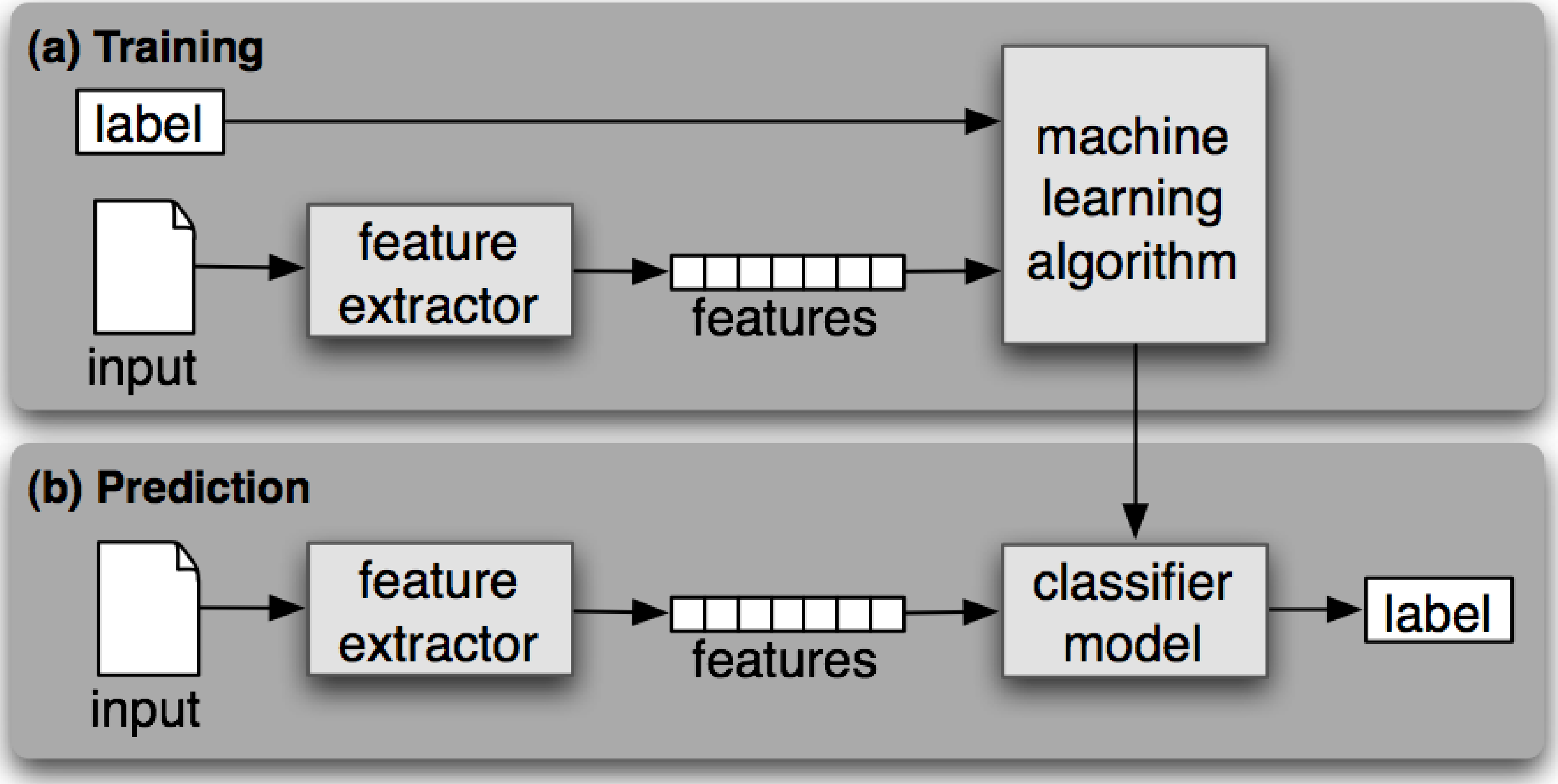
- Hierarchical classification: hierarchy where each node is the sub-class of its parent’s node

Single- vs. multi-category classification

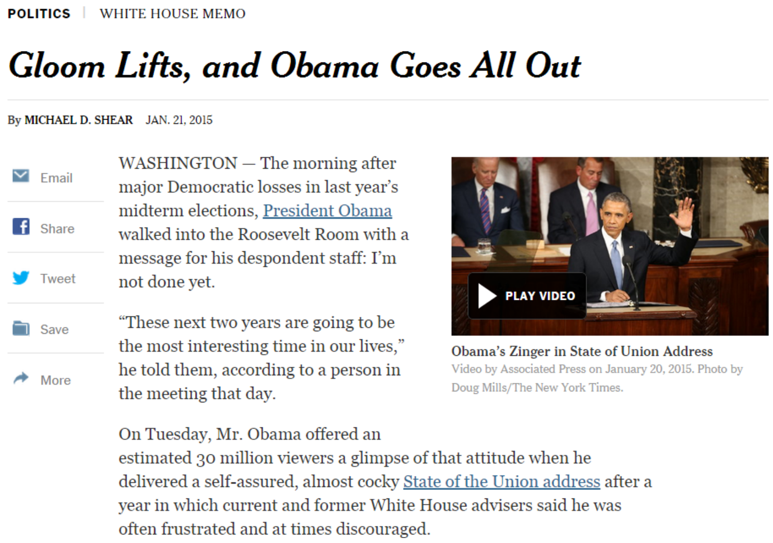
- In single-category text classification each text belongs to exactly one category

- In multi-category text classification, each text can have zero or more categories

General Pipeline of Classification



General steps for text categorization



POLITICS | WHITE HOUSE MEMO

Gloom Lifts, and Obama Goes All Out

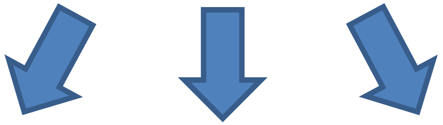
By MICHAEL D. SHEAR JAN. 21, 2015

WASHINGTON — The morning after major Democratic losses in last year's midterm elections, President Obama walked into the Roosevelt Room with a message for his despondent staff: I'm not done yet.

"These next two years are going to be the most interesting time in our lives," he told them, according to a person in the meeting that day.

On Tuesday, Mr. Obama offered an estimated 30 million viewers a glimpse of that attitude when he delivered a self-assured, almost cocky State of the Union address after a year in which current and former White House advisers said he was often frustrated and at times discouraged.

Obama's Zinger in State of Union Address
Video by Associated Press on January 20, 2015. Photo by Doug Mills/The New York Times.



Political News

Sports News

Entertainment News

1. Feature construction and selection
2. Model specification
3. Model estimation and selection
4. Evaluation

Consider:
1.1 How to represent the text documents?
1.2 Do we need all those features?

Feature construction for text categorization

- Vector space representation
  - Standard procedure in document representation
  - Features
    - N-gram, POS tags, named entities, topics
  - Feature value
    - Binary (presence/absence)
    - TF-IDF (many variants)

A Typical Corpus

- How many unigram+bigram are there in our controlled vocabulary?
  - 130K on Yelp_small
- How many review documents do we have there for training?
  - 629K Yelp_small

Very sparse feature representation!

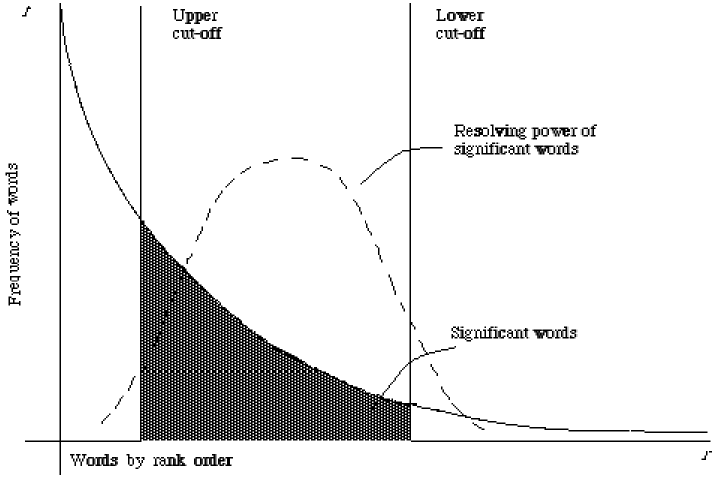


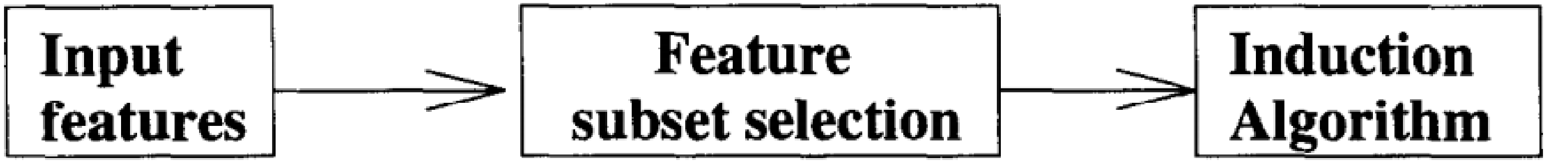
Figure 2.1. A plot of the hyperbolic curve relating f, the frequency of occurrence and r, the rank order (Adaped from Schultz[44] page 120)

Feature Selection

- A small corpus can have millions of unigram+bigram features
- Select the most informative features for model training
  - Reduce the feature space
  - Improve the final classification performance
  - Improve training/testing efficiency
    - Less tuning time
    - Fewer training data

Feature selection methods

- Filter method
  - Evaluate the features independently from the classifier and other features
    - No indication of a classifier’s performance on the selected features
    - No dependency among the features
  - Feasible for very large feature set
    - Usually used as a preprocessing step

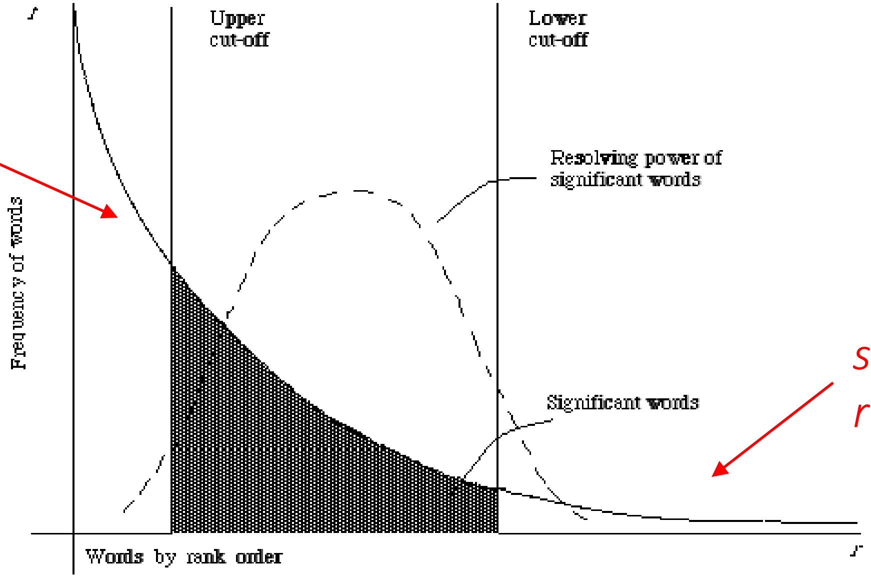


R. Kohavi, G.H. John/Artijicial Intelligence 97 (1997) 273-324

Feature scoring metrics

- Document frequency
  - Rare words: non-influential for global prediction, reduce vocabulary size

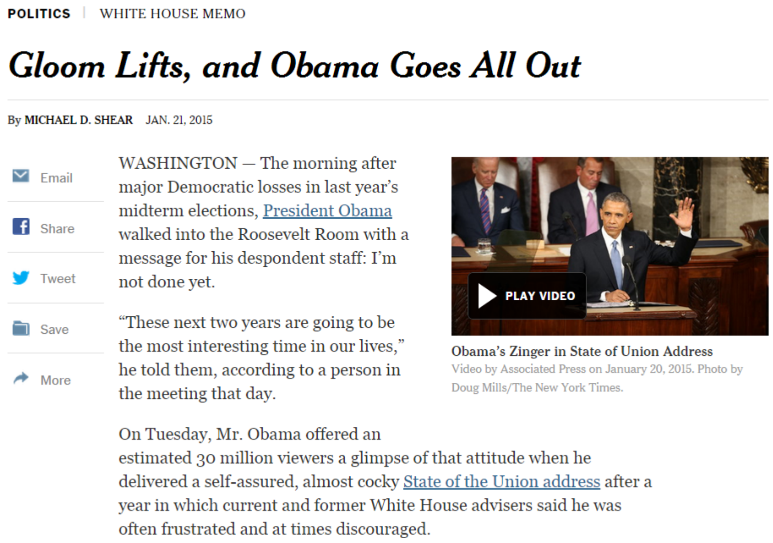
riskier to remove head words



safer to remove rare words

Figure 2.1. A plot of the hyperbolic curve relating f, the frequency of occurrence and r, the rank order (Adapted from Schultz[44] page 120)

General steps for text categorization



POLITICS | WHITE HOUSE MEMO

Gloom Lifts, and Obama Goes All Out

By MICHAEL D. SHEAR JAN. 21, 2015

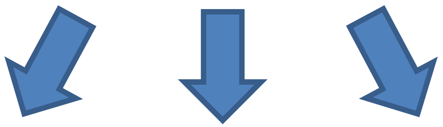
WASHINGTON — The morning after major Democratic losses in last year's midterm elections, President Obama walked into the Roosevelt Room with a message for his despondent staff: I'm not done yet.

"These next two years are going to be the most interesting time in our lives," he told them, according to a person in the meeting that day.

On Tuesday, Mr. Obama offered an estimated 30 million viewers a glimpse of that attitude when he delivered a self-assured, almost cocky State of the Union address after a year in which current and former White House advisers said he was often frustrated and at times discouraged.

Obama's Zinger in State of Union Address

Video by Associated Press on January 20, 2015. Photo by Doug Mills/The New York Times.



Political News

Sports News

Entertainment News

1. Feature construction and selection
2. Model specification
3. Model estimation and selection
4. Evaluation

Consider:

2.1 What is the unique property of this problem?

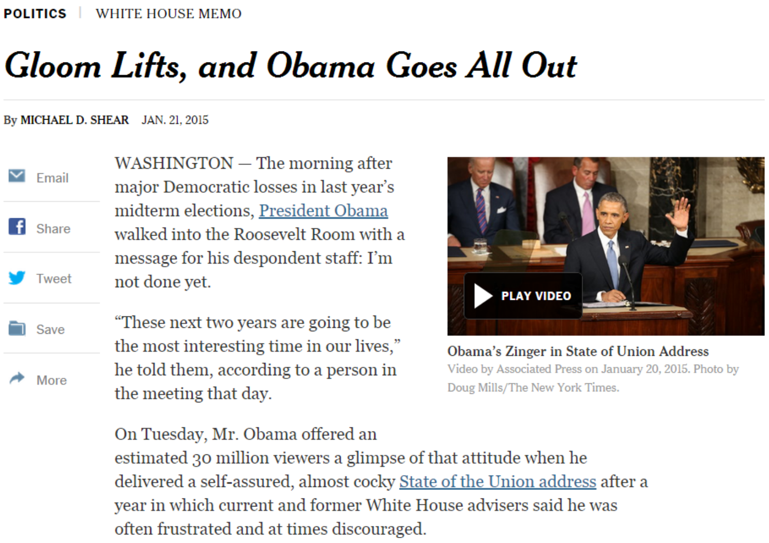
2.2 What type of classifier we should use?

Model specification

- Specify dependency assumptions
  - Linear relation between feature vector x and label y
    - $w^T x \rightarrow y$
    - Features are independent among each other
      - Naïve Bayes, perceptron
  - Non-linear relation between x and y
    - $f(x) \rightarrow y$, where $f(\cdot)$ is a non-linear function of x
    - Features are not independent among each other
      - Decision tree, kernel SVM, mixture model
- Choose based on our domain knowledge of the problem

We will discuss these choices later

General steps for text categorization



POLITICS | WHITE HOUSE MEMO

Gloom Lifts, and Obama Goes All Out

By MICHAEL D. SHEAR JAN. 21, 2015

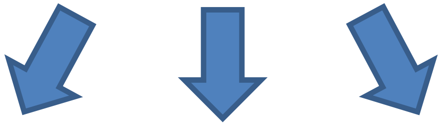
WASHINGTON — The morning after major Democratic losses in last year's midterm elections, President Obama walked into the Roosevelt Room with a message for his despondent staff: I'm not done yet.

"These next two years are going to be the most interesting time in our lives," he told them, according to a person in the meeting that day.

Obama's Zinger in State of Union Address

Video by Associated Press on January 20, 2015. Photo by Doug Mills/The New York Times.

On Tuesday, Mr. Obama offered an estimated 30 million viewers a glimpse of that attitude when he delivered a self-assured, almost cocky State of the Union address after a year in which current and former White House advisers said he was often frustrated and at times discouraged.



Political News

Sports News

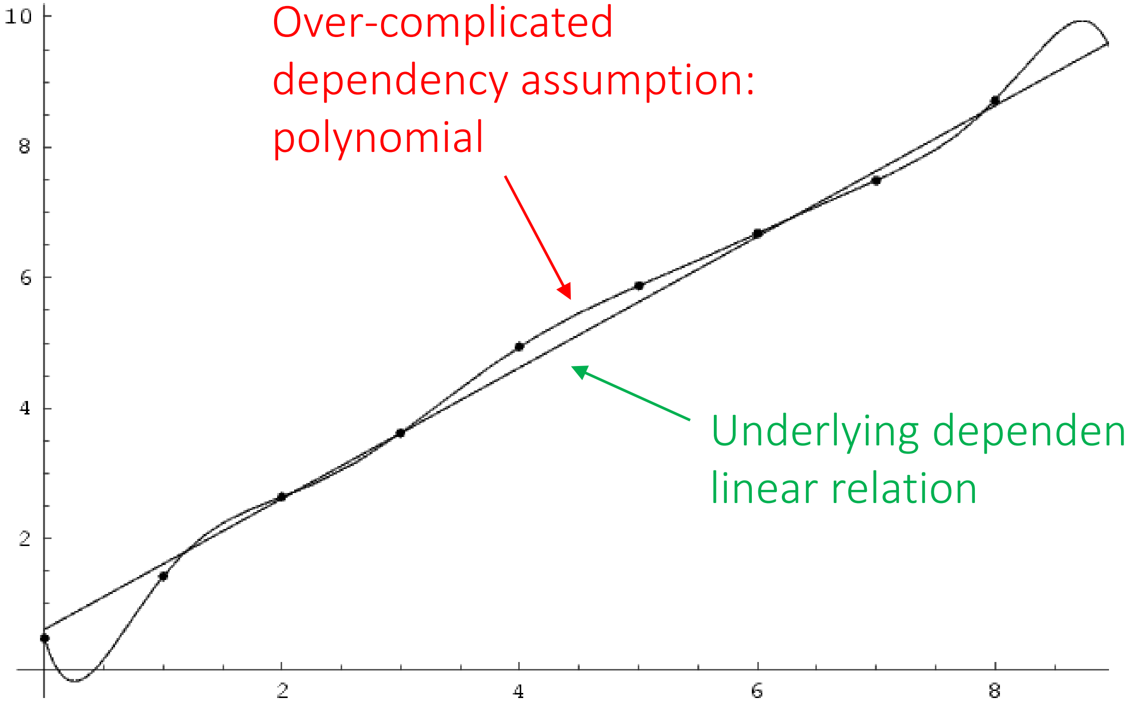
Entertainment News

1. Feature construction and selection
2. Model specification
3. Model estimation and selection
4. Evaluation

Consider:
3.1 How to estimate the parameters in the selected model?
3.2 How to control the complexity of the estimated model?

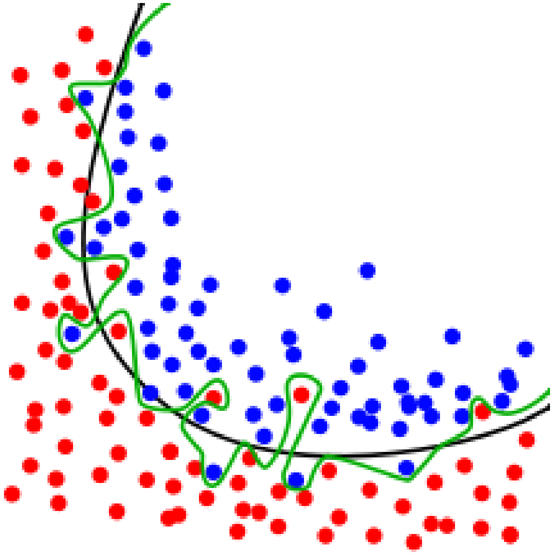
Empirical loss minimization

- Overfitting
  - Good empirical loss, terrible generalization loss
  - High model complexity -> prune to overfit noise



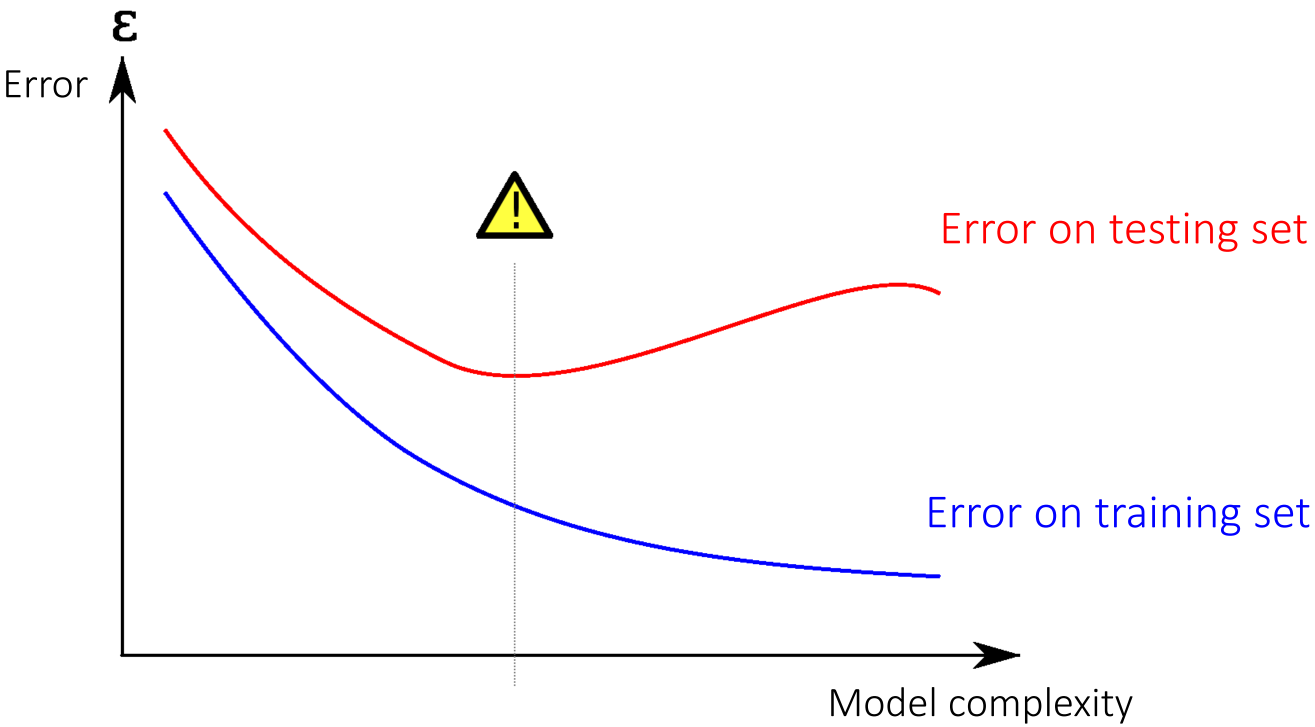
Over-complicated dependency assumption: polynomial

Underlying dependency: linear relation



Generalization loss minimization

- Avoid overfitting
  - Measure model complexity as well
  - Model selection and regularization

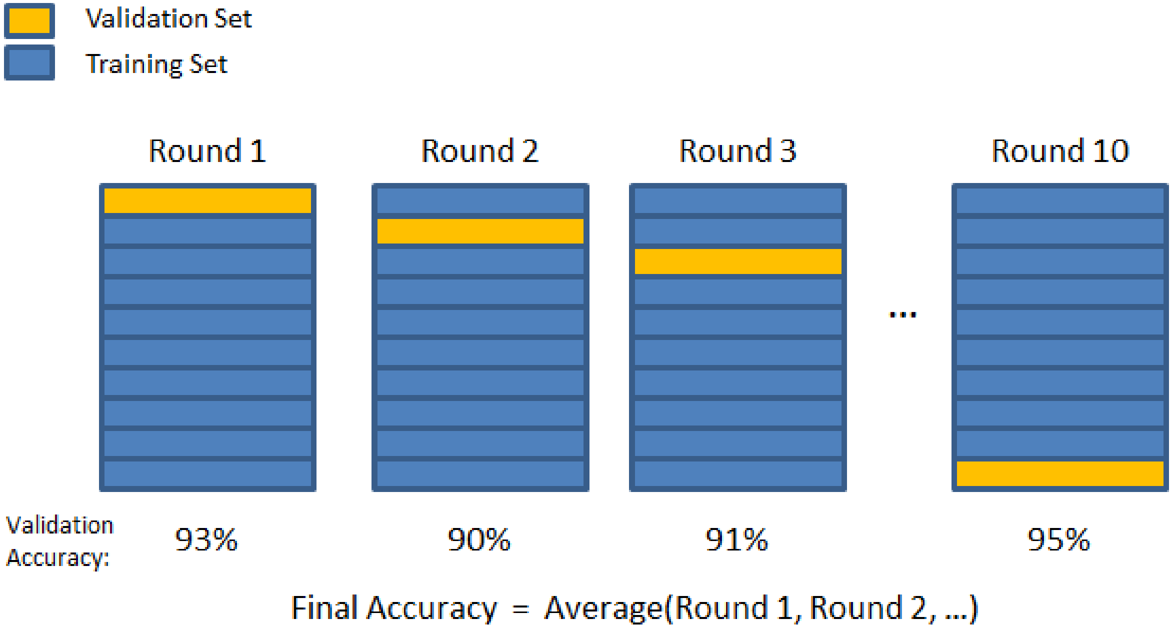


Generalization loss minimization

- Cross validation
  - Avoid noise in train/test separation
  - *k*-fold cross-validation
    1. Partition all training data into *k* equal size disjoint subsets;
    2. Leave one subset for validation and the other *k*-1 for training;
    3. Repeat step (2) *k* times with each of the *k* subsets used exactly once as the validation data.

Generalization loss minimization

- Cross validation
  - Avoid noise in train/test separation
  - *k*-fold cross-validation

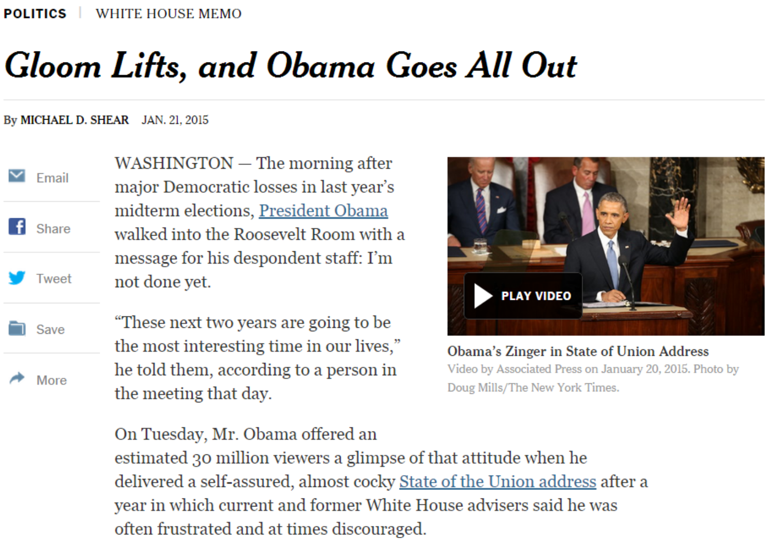


Generalization loss minimization

- Cross validation
  - Avoid noise in train/test separation
  - *k*-fold cross-validation
    - Choose the model (among different models or same model with different settings) that has the best average performance on the validation sets
    - Some statistical test is needed to decide if one model is significantly better than another

Will cover it shortly

General steps for text categorization



POLITICS | WHITE HOUSE MEMO

Gloom Lifts, and Obama Goes All Out

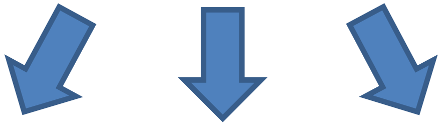
By MICHAEL D. SHEAR JAN. 21, 2015

WASHINGTON — The morning after major Democratic losses in last year's midterm elections, President Obama walked into the Roosevelt Room with a message for his despondent staff: I'm not done yet.

"These next two years are going to be the most interesting time in our lives," he told them, according to a person in the meeting that day.

Obama's Zinger in State of Union Address
Video by Associated Press on January 20, 2015. Photo by Doug Mills/The New York Times.

On Tuesday, Mr. Obama offered an estimated 30 million viewers a glimpse of that attitude when he delivered a self-assured, almost cocky State of the Union address after a year in which current and former White House advisers said he was often frustrated and at times discouraged.



Political News

Sports News

Entertainment News

1. Feature construction and selection
2. Model specification
3. Model estimation and selection
4. Evaluation

Consider:
4.1 How to judge the quality of learned model?
4.2 How can you further improve the performance?

Testing, evaluation of the classifier

- After choosing the parameters of the classifiers (i.e. after training it) we need to test how well it’s doing on a test set (not included in the training set)
- Calculate misclassification on the test set

Baselines

- First step: get a baseline
  - Help determine how hard the task is
  - Help know what a “good” accuracy is

- Weak baseline: most frequent label classifier
  - Gives all test instances whatever label was most common in the training set
  - E.g. for spam filtering, might label everything as “not spam”
  - Accuracy might be very high if the problem is skewed
  - E.g. calling everything “not spam” gets 66%, so a classifier that gets 70% isn’t very good…

- For real research, usually use previous work as a (strong) baseline

Classification evaluation

- Accuracy
  - Percentage of correct prediction over all predictions, i.e., $p(y^* = y)$
  - Limitation
    - Highly skewed class distribution
      - $p(y^* = 1) = 0.99$
        - Trivial solution: all testing cases are positive
      - Classifiers' capability is only differentiated by 1% testing cases

Evaluating classifiers

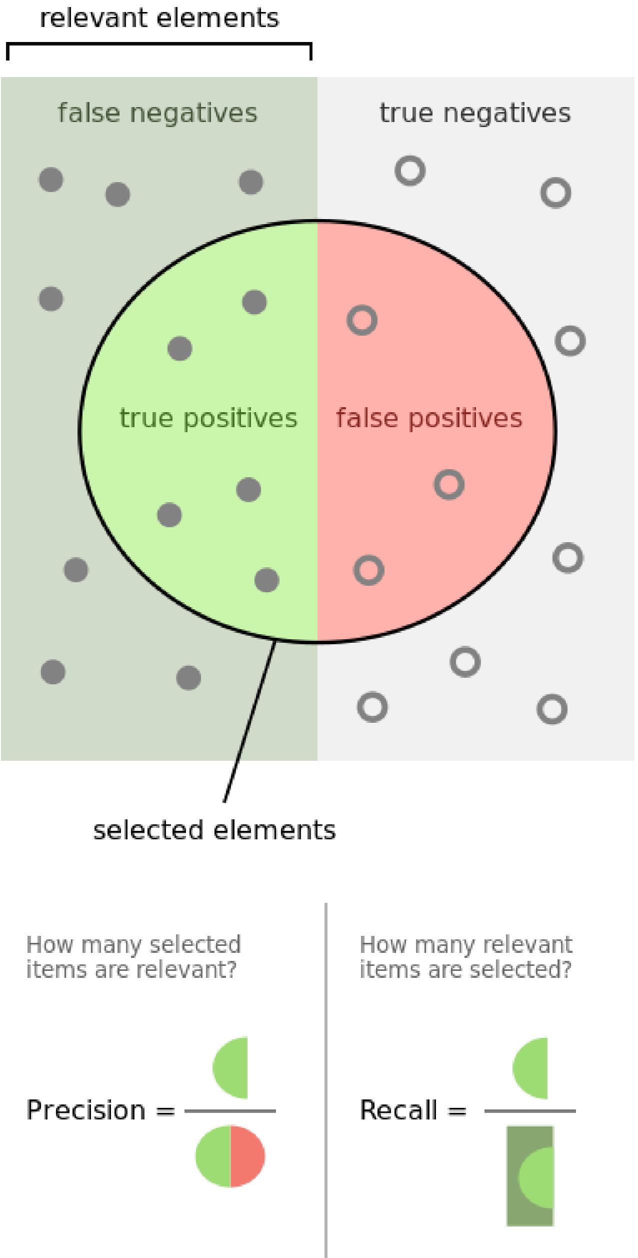
- Contingency table for the evaluation of a binary classifier

	GREEN is correct	RED is correct
GREEN was assigned	a	b
RED was assigned	c	d

- Accuracy = $(a+d)/(a+b+c+d)$
- Precision: P_GREEN = $a/(a+b)$, P_ RED = $d/(c+d)$
- Recall: R_GREEN = $a/(a+c)$, R_ RED = $d/(b+d)$

Evaluation of Binary Classification

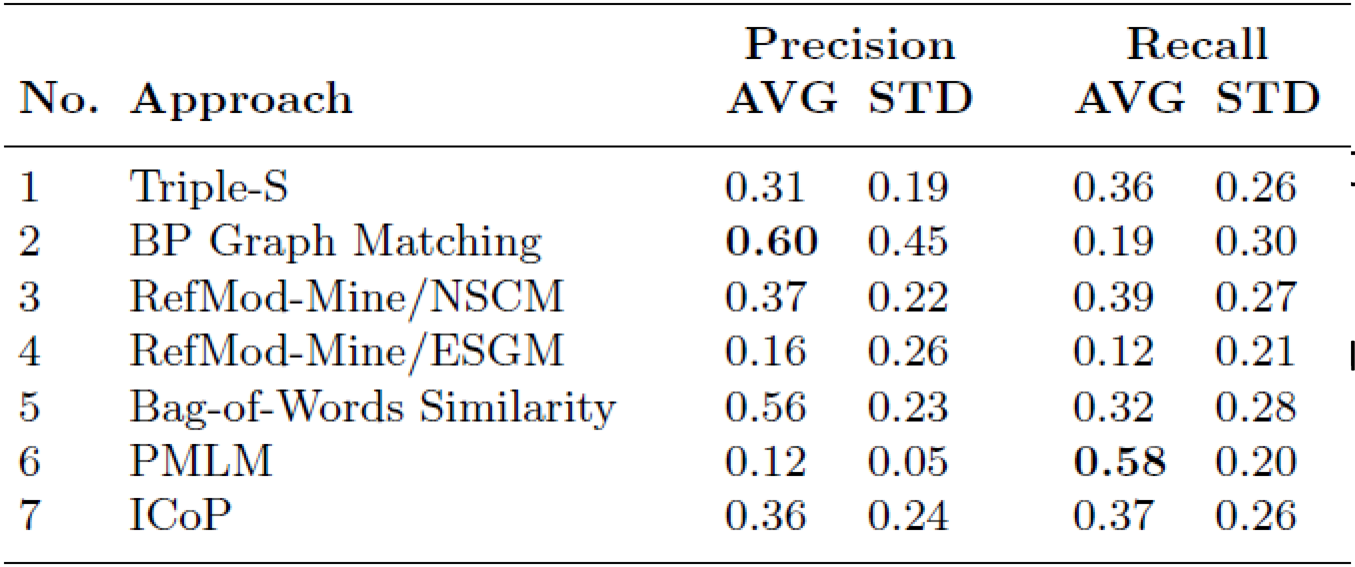
- Precision
  - Fraction of predicted positive documents that are indeed positive, i.e., P(gold = 1 | prediction = 1)
- Recall
  - Fraction of positive documents that are predicted to be positive, i.e., P(prediction = 1 | gold = 1)



https://en.wikipedia.org/wiki/Precision_and_recall
https://en.wikipedia.org/wiki/F1_score

Precision and recall trade off

- Precision decreases as the number of documents predicted to be positive increases (unless in perfect cla
- Th tives of
  - – ıments,
  - – nts



No.	Approach	Precision AVG	Precision STD	Recall AVG	Recall STD
1	Triple-S	0.31	0.19	0.36	0.26
2	BP Graph Matching	**0.60**	0.45	0.19	0.30
3	RefMod-Mine/NSCM	0.37	0.22	0.39	0.27
4	RefMod-Mine/ESGM	0.16	0.26	0.12	0.21
5	Bag-of-Words Similarity	0.56	0.23	0.32	0.28
6	PMLM	0.12	0.05	**0.58**	0.20
7	ICoP	0.36	0.24	0.37	0.26

Summarizing precision and recall

- With a single value
  - In order to compare different classifiers
  - F-measure: weighted harmonic mean of precision and recall, α balances the trade-off

$$F = \frac{1}{\alpha \frac{1}{P} + (1-\alpha)\frac{1}{R}} \qquad \left(F_1 = \frac{2}{\frac{1}{P} + \frac{1}{R}} \right)$$

Equal weight between precision and recall

  - Why harmonic mean?
    - Classifier1: P:0.53, R:0.36
    - Classifier2: P:0.01, R:0.99

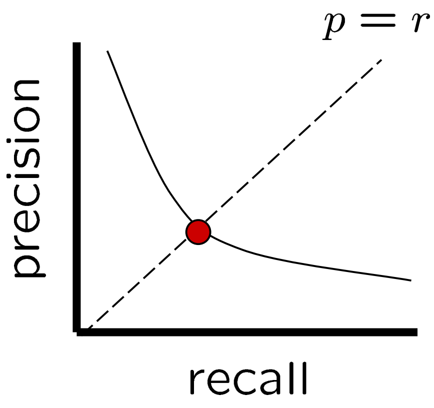
Harmonic	Average
0.429	0.445
0.019	0.500

Summarizing precision and recall

- With a curve
  1. Order all the testing cases by the classifier's prediction score (assuming the higher the score is, the more likely it is positive);
  2. Scan through each testing case: treat all cases above it as positive (including itself), below it as negative; compute precision and recall;
  3. Plot precision and recall computed for each testing case in step (2).

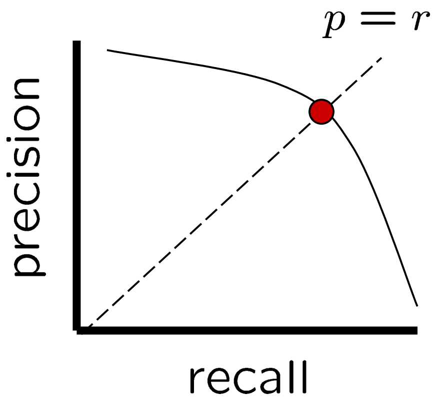
Summarizing precision and recall

- With a curve
  - A.k.a., precision-recall curve



- F-measure: harmonic mean of p and r:

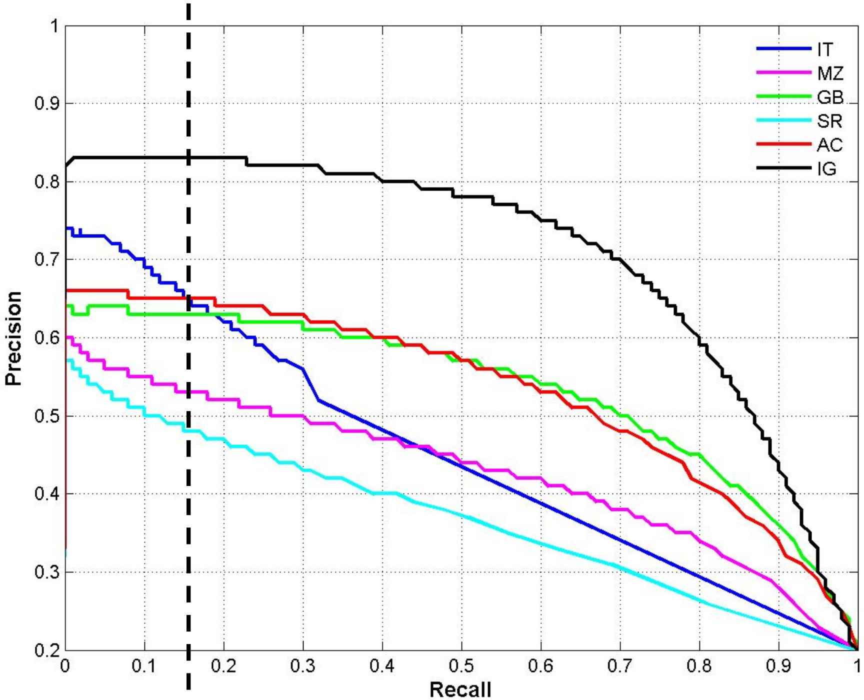
$$F_1 = \frac{2}{1/p + 1/r}$$



- Break-even point: precision value when p = r

Summarizing precision and recall

- With a curve
  - A.k.a., precision-recall curve
  - Area Under Curve (AUC)

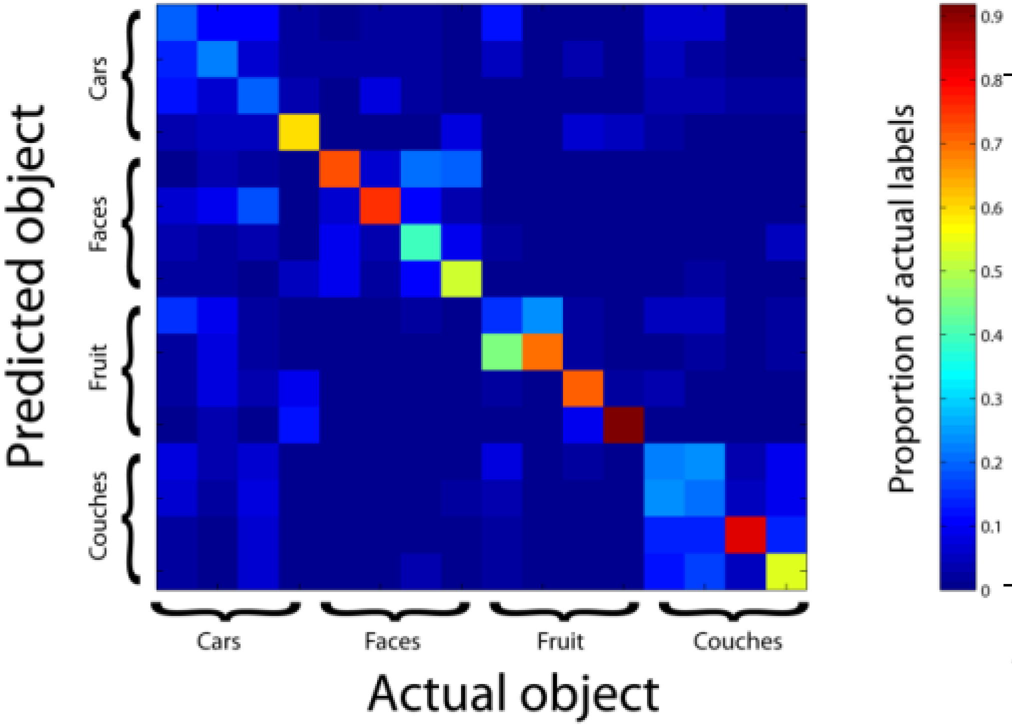


Under each recall level, we prefer a higher precision

Multi-class Categorization

- Confusion matrix
  - A generalized contingency table for precision and recall

Actual/ Predicted	Class 1	10	Total	Recall
Class 1	**9.06**	0.14	10	90.60
Class 2		0.42	10	82.00
Class 3	0.03	0.22	10	95.20
Class 4	0.01	0.06	10	90.10
Class 5		0.90	10	26.70
Class 6		0.62	10	77.50
Class 7	0.02	0.67	10	36.60
Class 8	0.20	0.03	10	97.00
Class 9		1.02	10	68.40
Class 10		**5.37**	10	53.70
Total	9.32	9.45	100	
Precision	97.21	56.83		



Training size

- The more the better! (usually)
- Results for text classification*

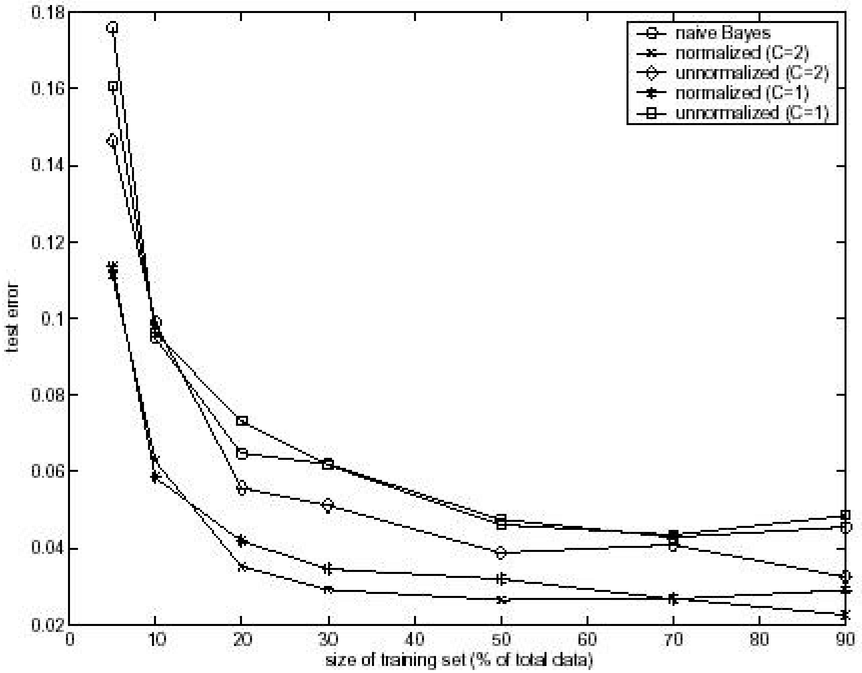


Figure 1: Test error vs training size on the newsgroups rec.sport.baseball and rec.sport.hockey

*From: Improving the Performance of Naive Bayes for Text Classification, Shen and Yang

What you should know

- General steps for text categorization
  - Text feature construction
  - Feature selection methods
  - Model specification and estimation
  - Evaluation metrics