Plan for Today's Lecture(s)

- Computational classification
  - Topic identification
  - Author identification
  - Spam detection
  - Sentiment analysis
- Question answering & Watson

UNIVERSITY OF CALIFORNIA, BERKELEY
SCHOOL OF INFORMATION

INFO 202
“Information Organization & Retrieval”
Fall 2013

Robert J. Glushko
glushko@berkeley.edu
@rjglushko

3 December 2013
Lecture 27.1 – Computational Classification

Computational Classification

- Some classification tasks can't be done by people at the needed scale or speed
- Some of the tasks can be done by computational approaches like N-gram analysis that are very simple yet very useful
- Some tasks are inherently difficult and require techniques of “machine learning”, which are not simple

Text Classification Problems

- Classification assigns objects in some domain to one of at least two classes or categories
  - words - determine part of speech
  - words - disambiguate polysemy
  - document retrieval - relevant/not relevant?
  - author identification - Shakespeare or not?
  - sentiment classification - positive or negative affect? urgent or not urgent?
  - language - English, Spanish, whatever?

Computational Classification

- CLASSIFICATION assumes a system of categories and some labeled instances so we can train a system to assign new instances to the appropriate categories
- In contrast, CLUSTERING techniques don't assume pre-existing categories - they create them (usually to maximize similarity within categories and maximize it between them)

Classifiers

- A *classifier* is a system whose input is a vector of discrete or continuous feature values and whose output is a single discrete value, the name of the class
- There are many learning algorithms; the choice among them depends on the domain and the kinds of features that resources in it have
- The more complex the domain – the more features it takes to describe each instance – the more examples are needed to train the classifier

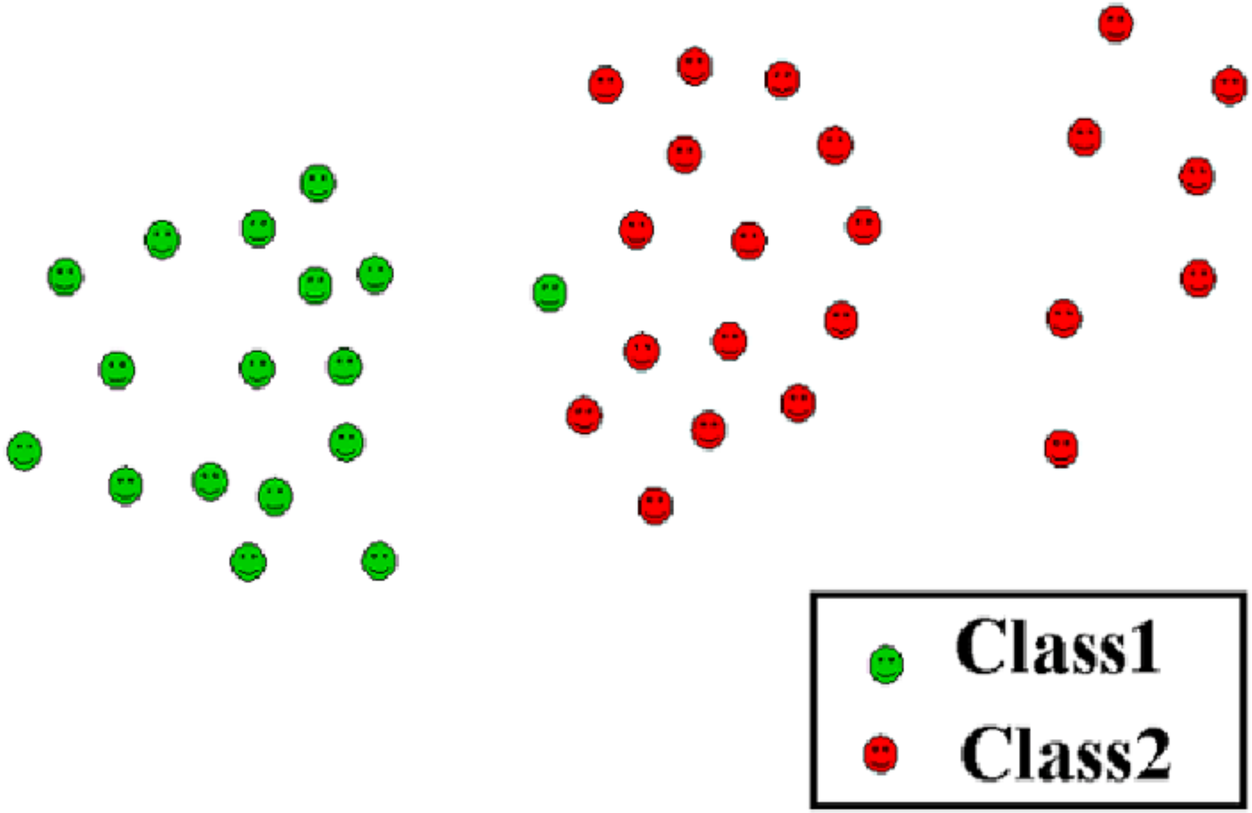
The Classification Process

- Specify classes
  - Sounds straightforward, but it isn't. How many categories?
- Label examples
  - Are the examples representative?
- Extract features / Choose a classifier algorithm
  - These are interdependent. Choose some set of features and run an algorithm; try some other features with the same algorithm; try other algorithms with the same features…

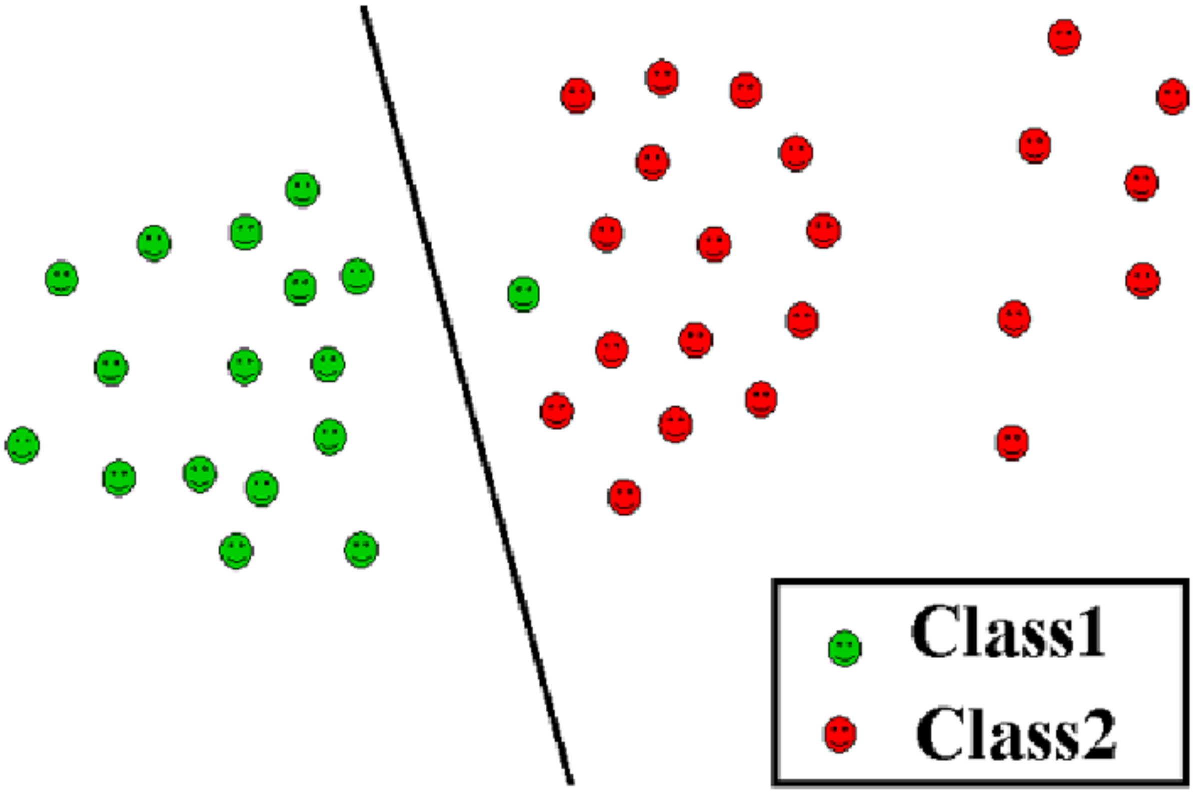
The Classification Process

- Train and test
  - after you've chosen features and an algorithm, you let it "learn" - adjusting the weights it gives to each feature....
  - (usually better to have a “slow” learner and lots of data than a “smart” one with limited data)
- Classify new examples

A Simple Classification Problem



A Classification Solution



Easy Problem: Language Identification

Dr. Ted Mazer is one of the few ear, nose and throat specialists in this region who treat low-income people on Medicaid, so many of his patients travel long distances to see him.

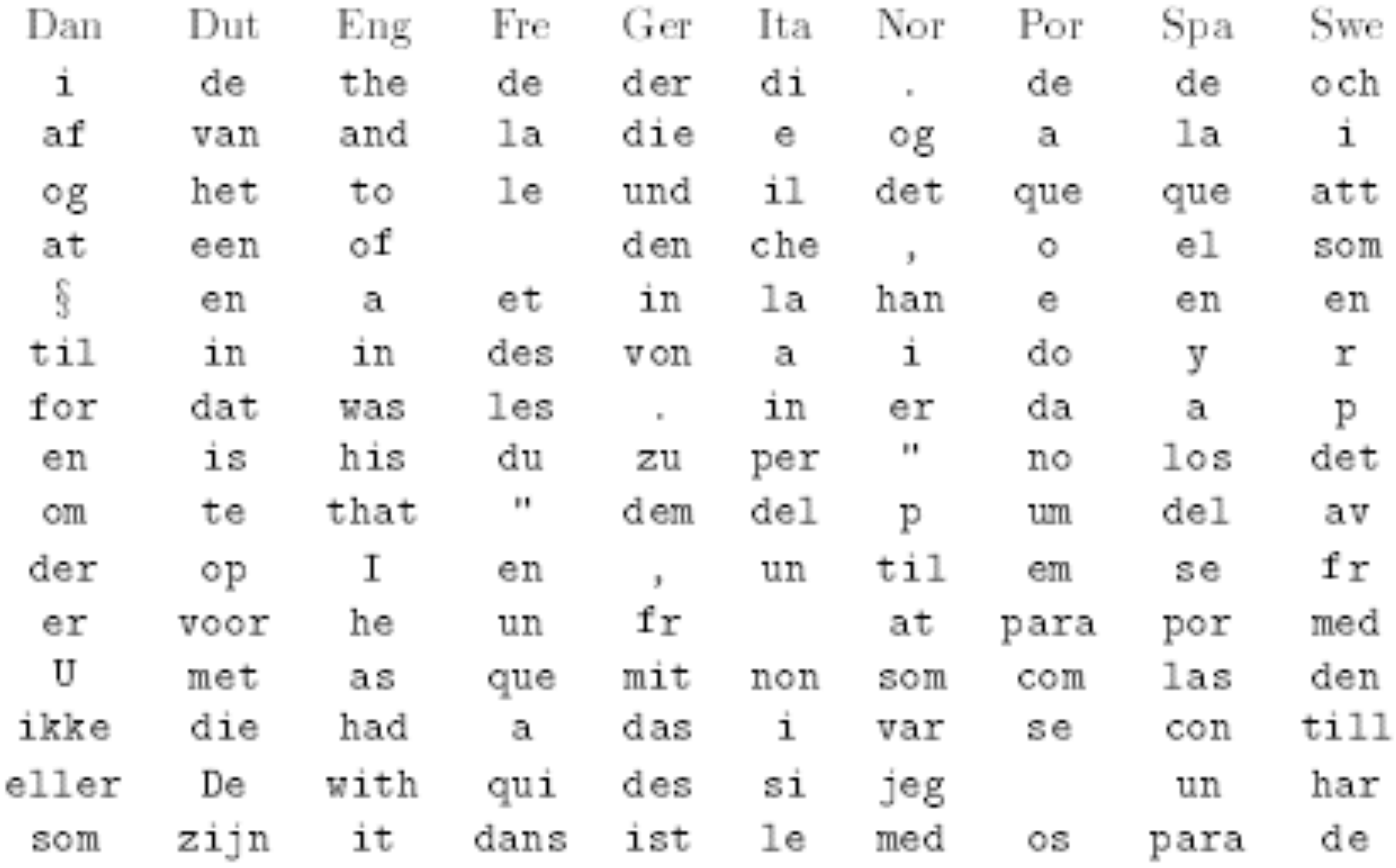
Dr. Ted Mazer ist einer der wenigen Hals-Nasen-Ohrenärzte in dieser Region, die Menschen mit niedrigem Einkommen auf Medicaid zu behandeln, so dass viele seiner Patienten lange Strecken, um ihn zu sehen.

N-Grams

- A text can be sliced into a set of overlapping N-grams, an N-character contiguous "slice"
- The word "TEXT" can be composed of these N-grams:
  - Uni-grams: _, T, E, X, T, _
  - Bi-grams: _T, TE, EX, XT, T_
  - Tri-grams: _TE, TEX, EXT, XT_, T__
  - Quad-grams: _TEX, TEXT, EXT_, XT__, T___

(The Google Books "N-gram viewer" enables the "quantitative analysis of culture" via analysis of usage trends for words and grammatical constructions)

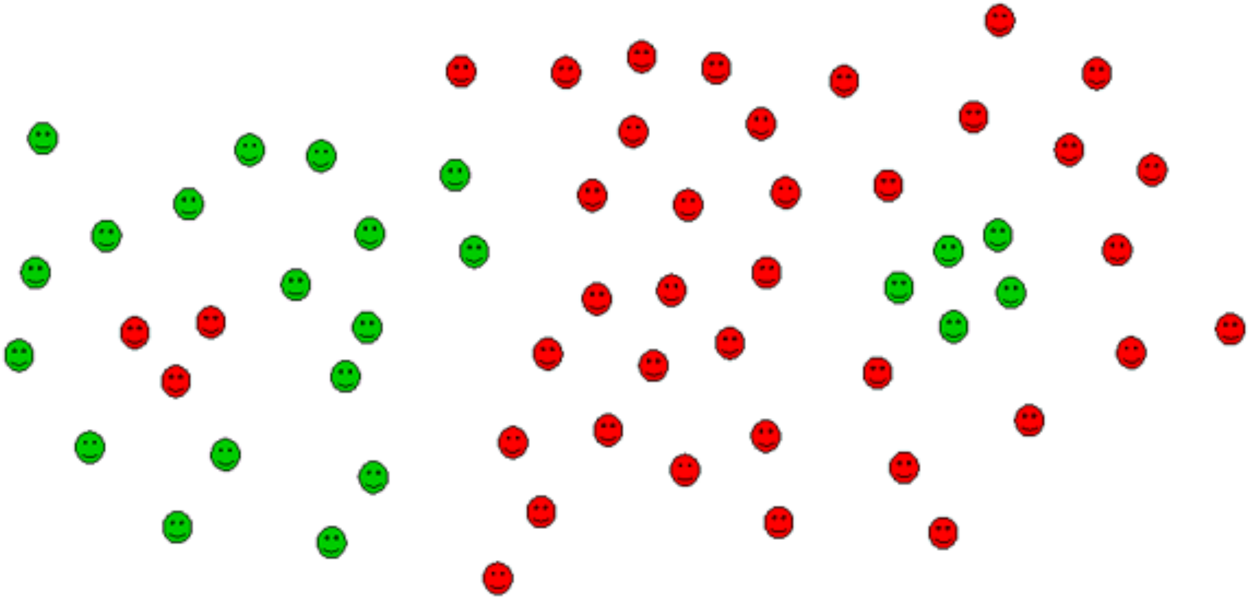
N-Gram Frequency (or Probability) Identifies a Language



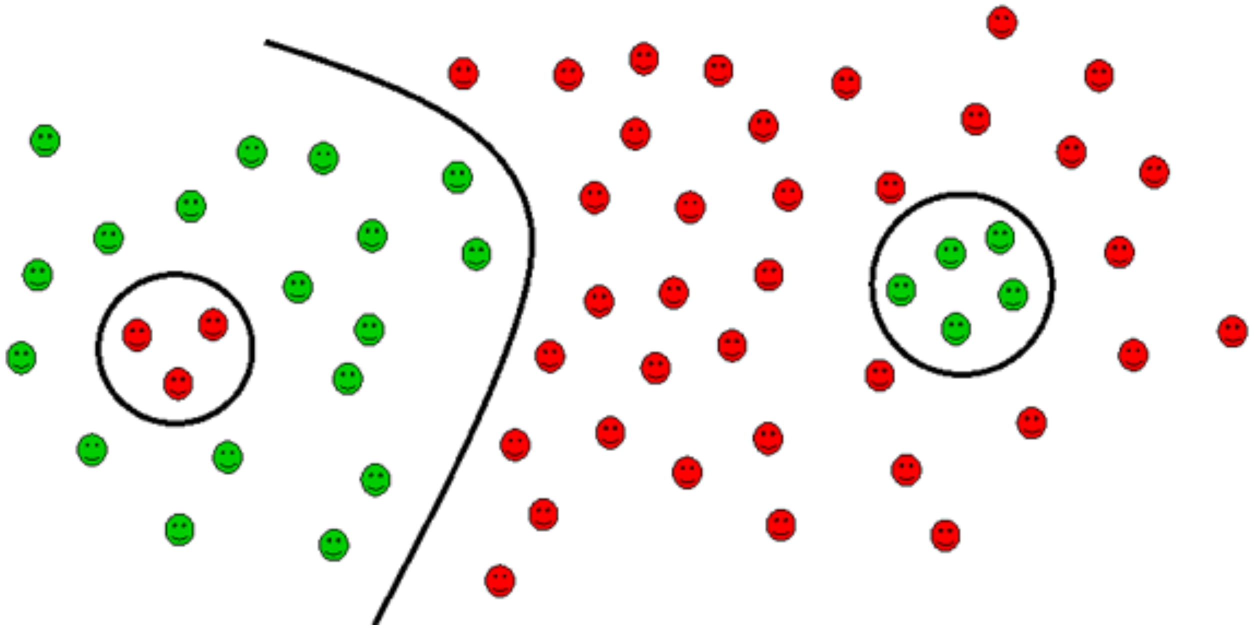
Dan	Dut	Eng	Fre	Ger	Ita	Nor	Por	Spa	Swe
i	de	the	de	der	di	.	de	de	och
af	van	and	la	die	e	og	a	la	i
og	het	to	le	und	il	det	que	que	att
at	een	of		den	che	,	o	el	som
§	en	a	et	in	la	han	e	en	en
til	in	in	des	von	a	i	do	y	r
for	dat	was	les	.	in	er	da	a	p
en	is	his	du	zu	per	"	no	los	det
om	te	that	"	dem	del	p	um	del	av
der	op	I	en	,	un	til	em	se	fr
er	voor	he	un	fr		at	para	por	med
U	met	as	que	mit	non	som	com	las	den
ikke	die	had	a	das	i	var	se	con	till
eller	De	with	qui	des	si	jeg		un	har
som	zijn	it	dans	ist	le	med	os	para	de

Most frequent tokens in different European languages (Grefenstette, 1995)

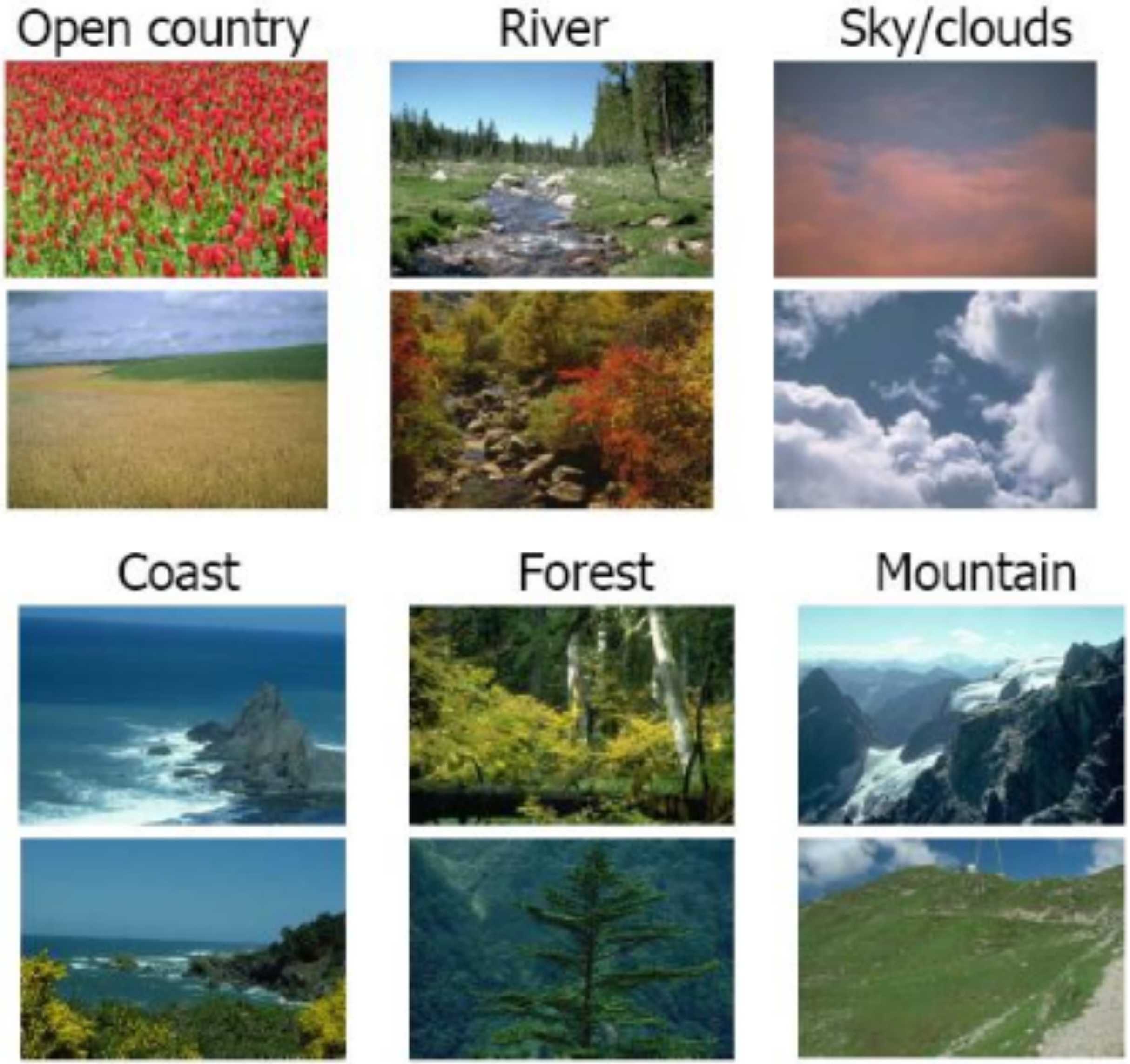
A More Complex Classification Problem



A More Complex Classification Solution



A Very Hard Classification Problem: Scenes

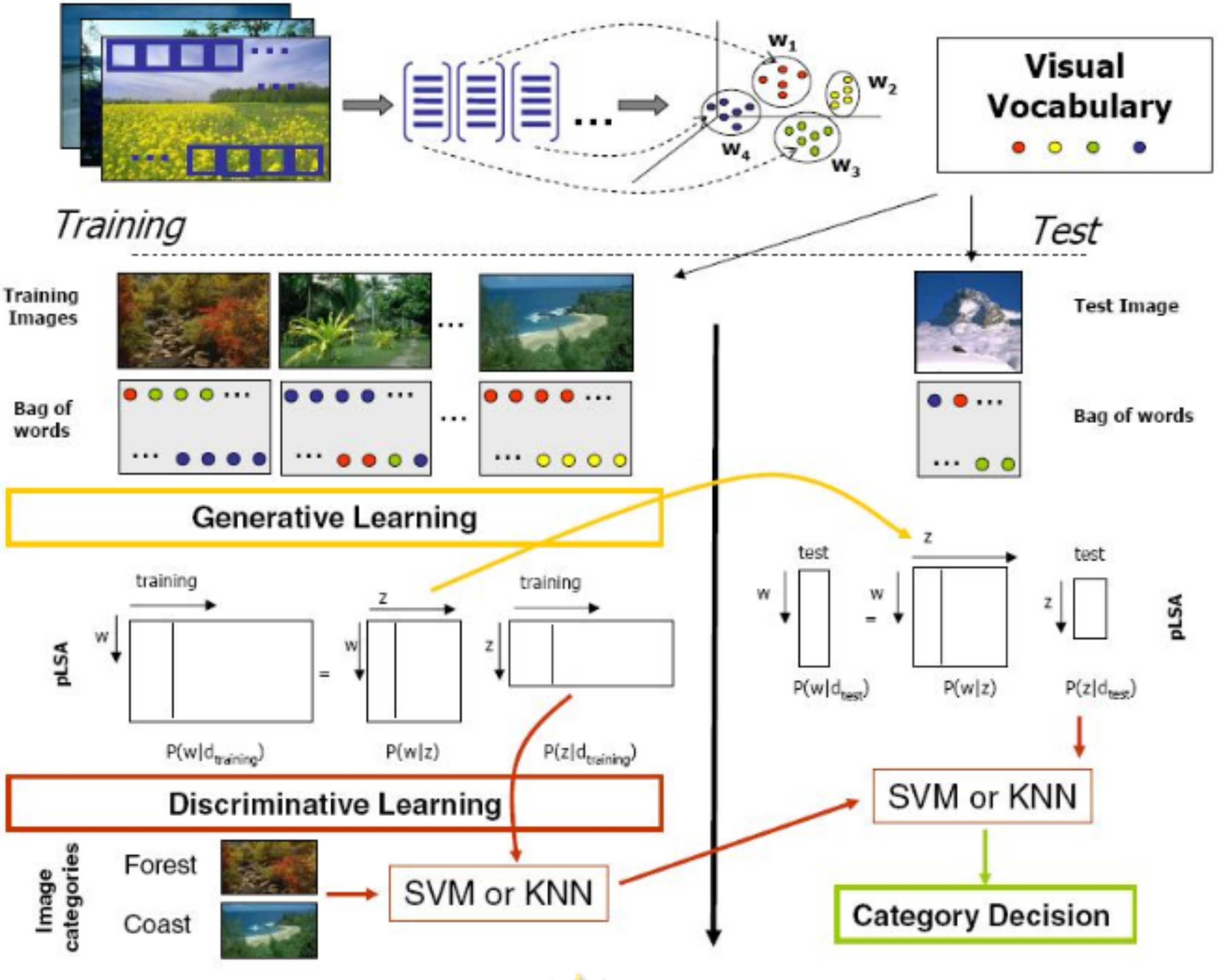


Joshi et al, Machine Learning in Computer Vision: A Tutorial

Machine Learning

- “Supervised” learning algorithms create classifiers using labeled examples
- “Unsupervised” learning algorithms create the categories in a classification system by discovering the correlations between features or properties of the things to be classified (also called “statistical pattern recognition”)
- In either case, the goal is to generalize beyond the examples used to train the classifier

Scene Classification Solution



Joshi et al, Machine Learning in Computer Vision: A Tutorial

UNIVERSITY OF CALIFORNIA, BERKELEY
SCHOOL OF INFORMATION

INFO 202
"Information Organization & Retrieval"
Fall 2013

Robert J. Glushko
glushko@berkeley.edu
@rjglushko

3 December 2013
Lecture 27.2 – Topic Identification and Sentiment Analysis

Features for Text Classification

- Linguistic Features (lexical and syntactic)
  - Words (stems?)
  - Phrases
  - Word and character level "N-grams"
  - Punctuation
  - Part of speech
- Non-linguistic features (structure and formatting)

Topic Categorization in Google News

- Google News http://news.google.com/ gathers stories from more than 4,500 English-language news sources worldwide, and automatically arranges by relevance
- "Google News has no human editors selecting stories or deciding which ones deserve top placement. Our headlines are selected by computer algorithms…"
- "Our grouping technology examines numerous data points for each article including the titles, text and publication time. We then use *clustering* algorithms to identify closely related articles."

Going Beyond Topic Analysis

- Google's news application analyzes the content and (some of) the metadata for news stories to categorize them on the basis of the topic or event
- However, news sources and authors have different points of view on the same topic or event
- Should these be treated as the "same" story?
  - Police Arrest Student Protestors, Ending Illegal Occupation of UC's Wheeler Hall
  - Students Protest Huge Tuition Hikes with Symbolic Occupation on UC Campus

Sentiment Analysis

- Sentiment analysis (aka "opinion mining") can be thought of a three-stage classification problem
  - Entity extraction to locate text of interest
  - Classifying texts as opinions or facts
  - Classifying the opinions according to polarity - positive vs. negative (or on some numerical scale)
- This is challenging because these classes are really continuaa without sharp boundaries
- ...and because sarcasm, slang, cliches, and cultural norms obscure the content used to make the classification

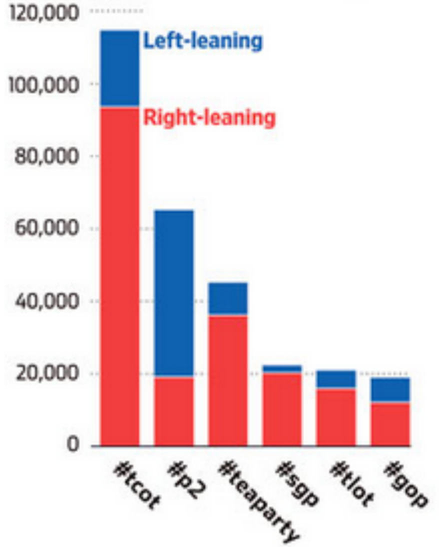
Sentiment Analysis in Twitter

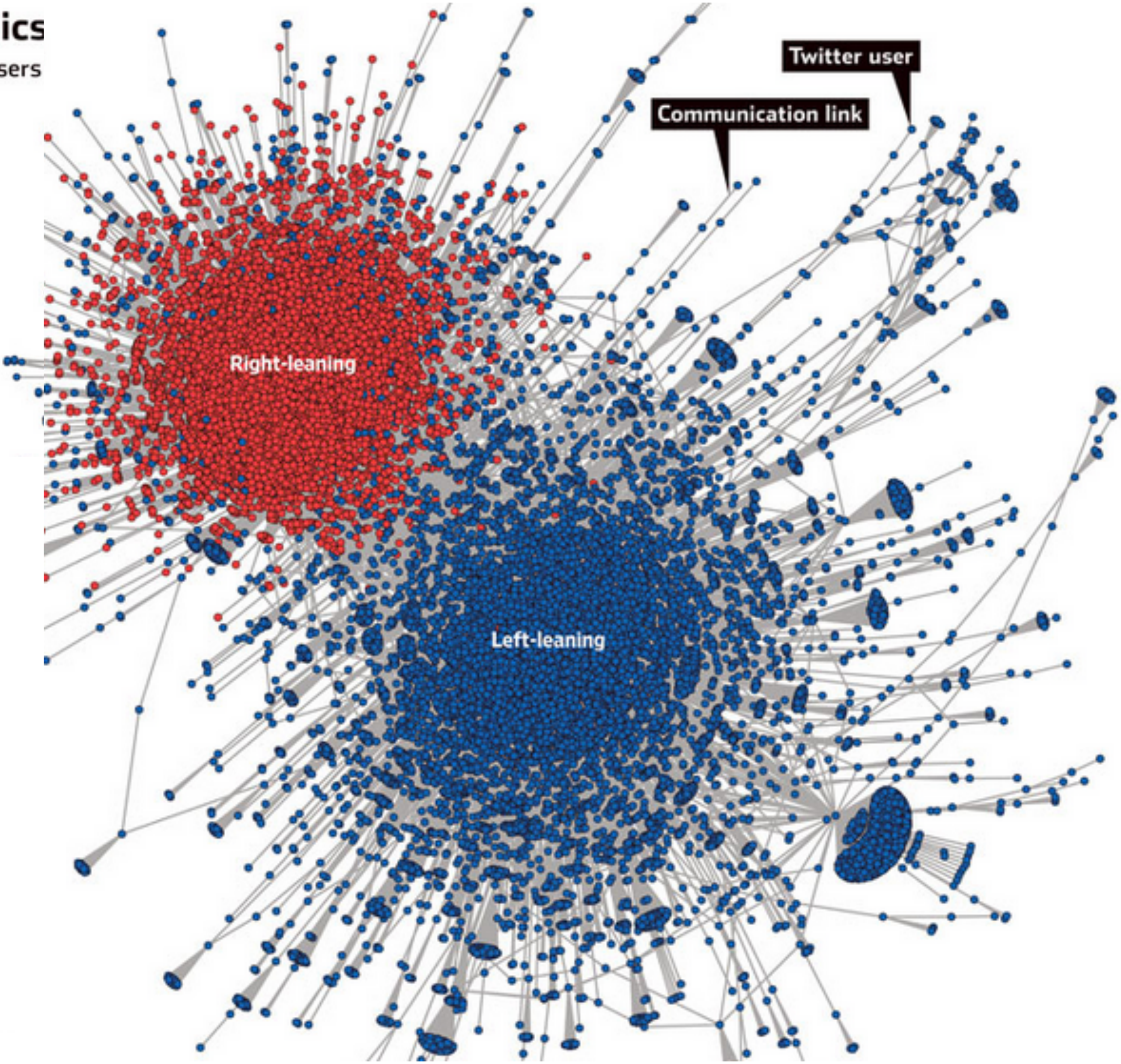
- Twitter messages are being analyzed to:
  - Monitor political activity
  - Assess employee morale or customer sentiment
  - Track outbreaks of flu or food poisoning
  - Predict box-office receipts for new movies
  - Measure "moods" and their fluctuations
- Twitter is a good domain to analyze because…
- Twitter is a poor domain to analyze because…

Twitter's Divided Politics

Political Twitter traffic reveals that users polarized along party lines.*

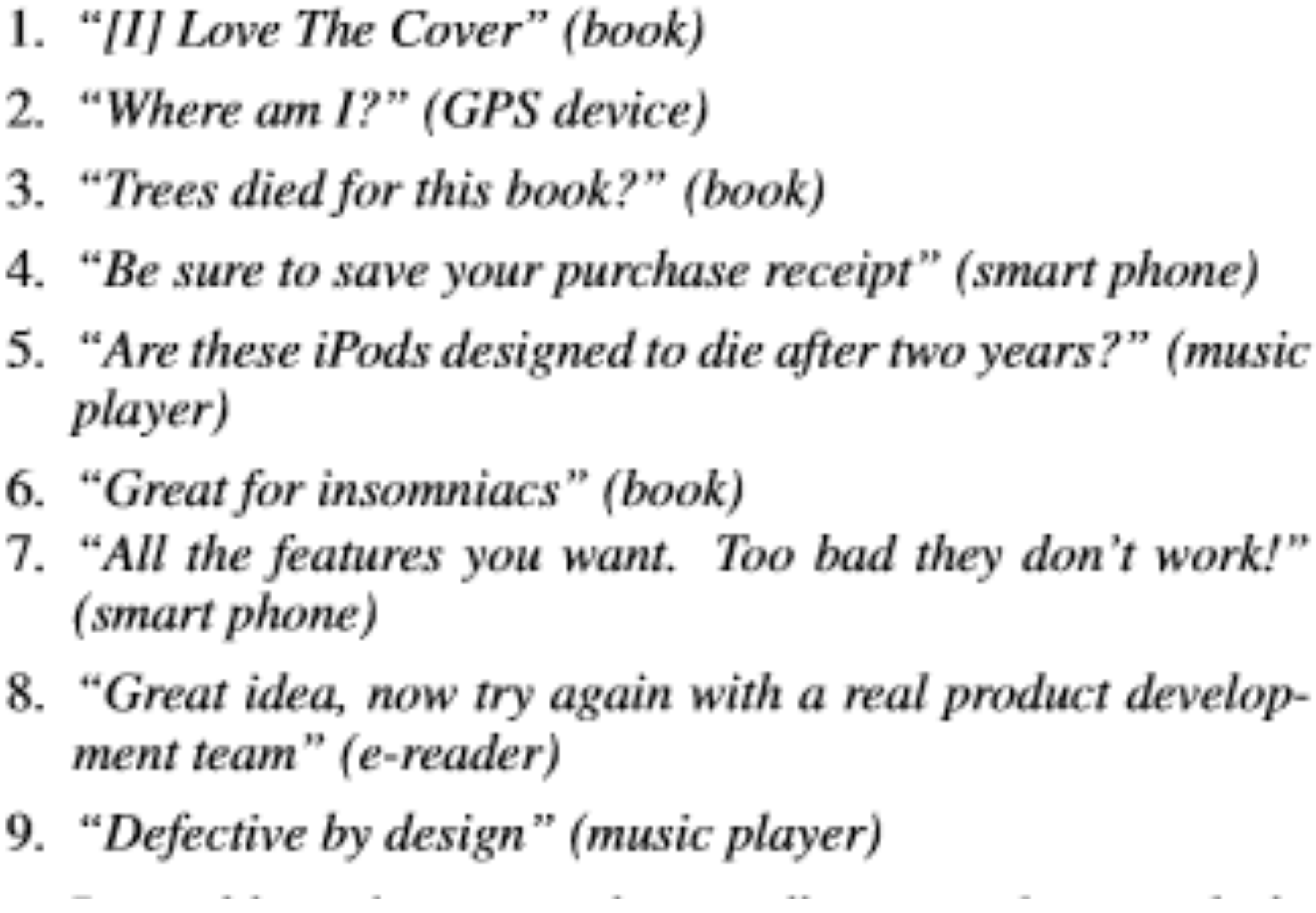
Researchers at Indiana University analyzed 250,000 Twitter messages on political topics exchanged by 45,000 people during the 2010 mid-term congressional elections. This chart of 'retweets'—in which one user forwards another's message—shows that, though there were more left-leaning users, right-leaning users were more densely connected to one another. (Each dot is a Twitter user, and the lines show retweets.) Even so, as the chart illustrates, lines of communication do sometimes reach across the political divide.





Decoding Our Chatter Wall Street Journal 1 October 2011

The Challenge of Sarcasm



1. *"[I] Love The Cover" (book)*
2. *"Where am I?" (GPS device)*
3. *"Trees died for this book?" (book)*
4. *"Be sure to save your purchase receipt" (smart phone)*
5. *"Are these iPods designed to die after two years?" (music player)*
6. *"Great for insomniacs" (book)*
7. *"All the features you want. Too bad they don't work!" (smart phone)*
8. *"Great idea, now try again with a real product development team" (e-reader)*
9. *"Defective by design" (music player)*

The Challenge of Fake Accounts

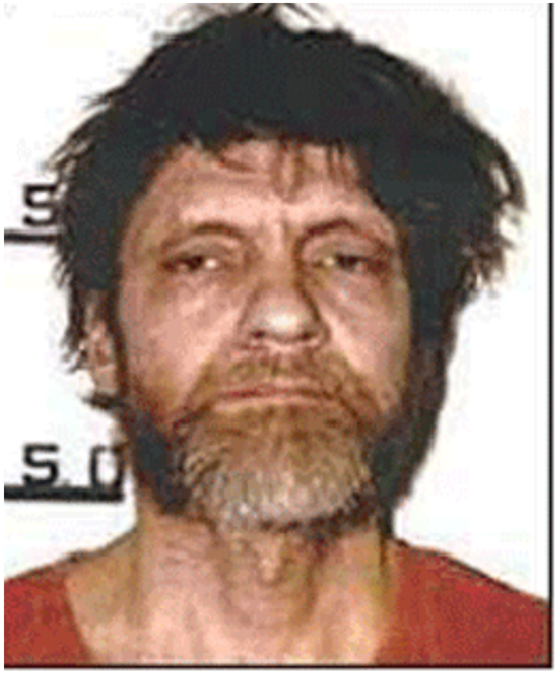
- The analysis of tweets to measure popularity or sentiment is compromised by the huge proportion or twitter accounts that are fake – "tweetbots"
- Tweetbots are programmed to follow people and retweet them, which greatly exaggerates the actual popularity or sentiment about some topic
- See "Millions of Fake Accounts Dog Twitter"

UNIVERSITY OF CALIFORNIA, BERKELEY
SCHOOL OF INFORMATION

INFO 202
"Information Organization & Retrieval"
Fall 2013

Robert J. Glushko
glushko@berkeley.edu
@rjglushko

3 December 2013
Lecture 27.3 – Author Identification



Motivation and Applications

Authorship Classification Use Cases

- Authorship IDENTIFICATION determines the likelihood of a particular author having written a piece of work by examining other works produced by that author
- Authorship CHARACTERIZATION is aimed at inferring an author's background characteristics rather than identity
- Similarity detection compares multiple pieces of writing without identifying the author
- (related use case: plagiarism detection)

The Authorship Identification NLP Model

- Goal is to identify a set of features that remain relatively constant among a number of writings by a particular author
- Given n predefined features, each piece of writing can be represented by an n-Dimensional feature vector.
- Supervised learning techniques can train and generate a classifier that can to determine the category of a new vector to identify the authorship of an anonymous (or disputed) writing

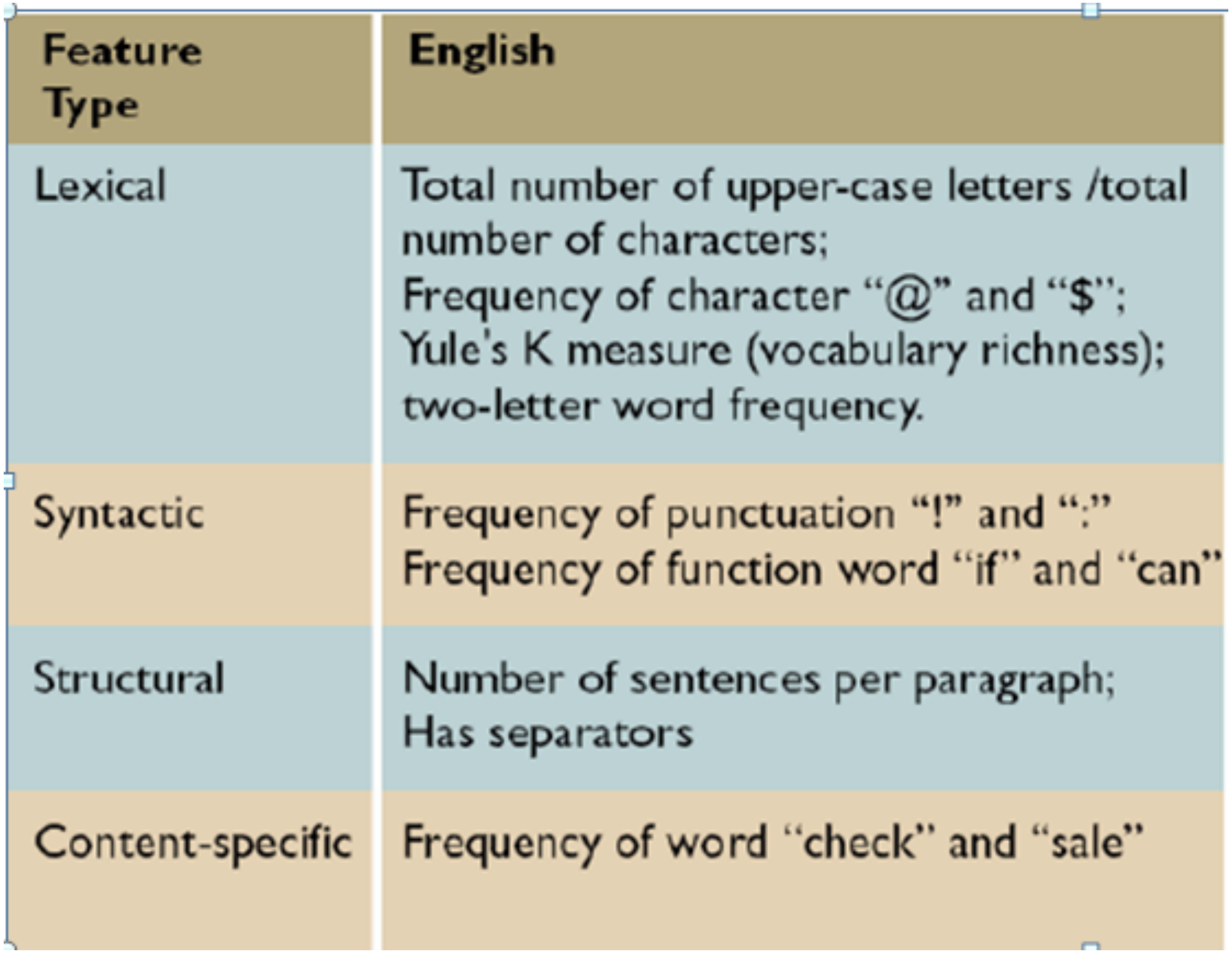
Authorship Identification

- Given:
  - A text with unknown author
  - A list of possible authors
  - A sample of their writing
- Can we automatically determine which person wrote the text?

From Fingerprint to "Writeprint"

- A *writeprint* is composed of multiple features, such as vocabulary richness, length of sentence, use of function words, layout of paragraphs, and keywords
- These writeprint features can represent an author's writing style, which is usually consistent across his or her writings, and further become the basis of authorship analysis

Features used in Writeprints

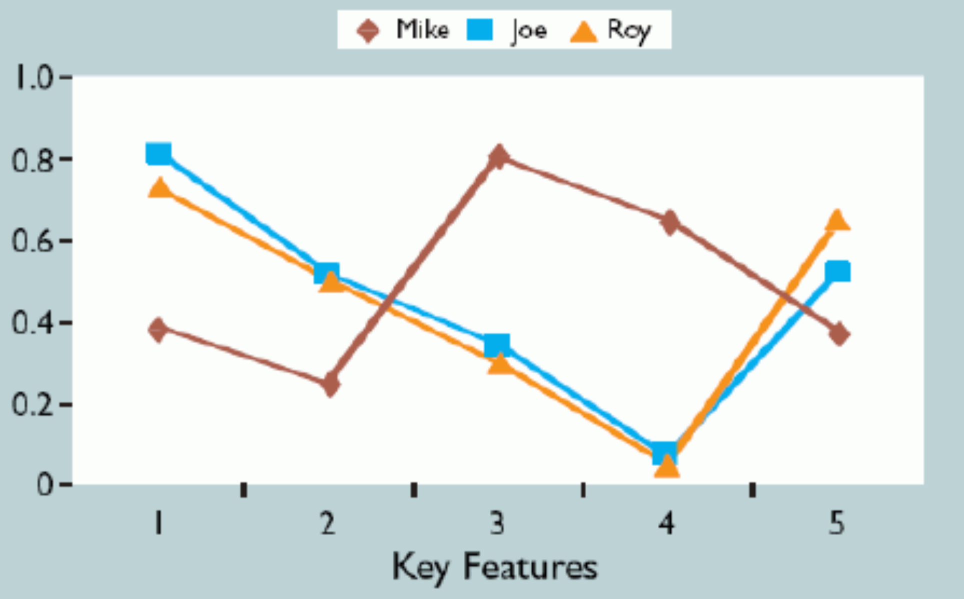


Feature Type	English

Lexical	Total number of upper-case letters /total number of characters;
	Frequency of character "@" and "$";
	Yule's K measure (vocabulary richness);
	two-letter word frequency.

Syntactic	Frequency of punctuation "!" and ":" Frequency of function word "if" and "can"
Structural	Number of sentences per paragraph; Has separators
Content-specific	Frequency of word "check" and "sale"

How Many Authors?



Feature	Description
1	Yule's K
2	Frequency of "$"
3	Frequency of "!"
4	No. of sentences/paragraph
5	Frequency of "sale"

Who wrote "The Cukoo's Calling?"

The Disputed Federalist Papers

- The Federalist papers were 77 short essays written in 1787-1788 by Hamilton, Jay and Madison to persuade NY to ratify the US Constitution; published under a pseudonym
- Historians disputed the authorship of 12 of the papers
- Two statisticians (Mosteller and Wallace, 1964) solved the problem by identifying 70 words whose usage patterns distinguished the papers with known authors
- Their statistical classifier concluded that the author was Madison

Author Identification for the Federalist Papers

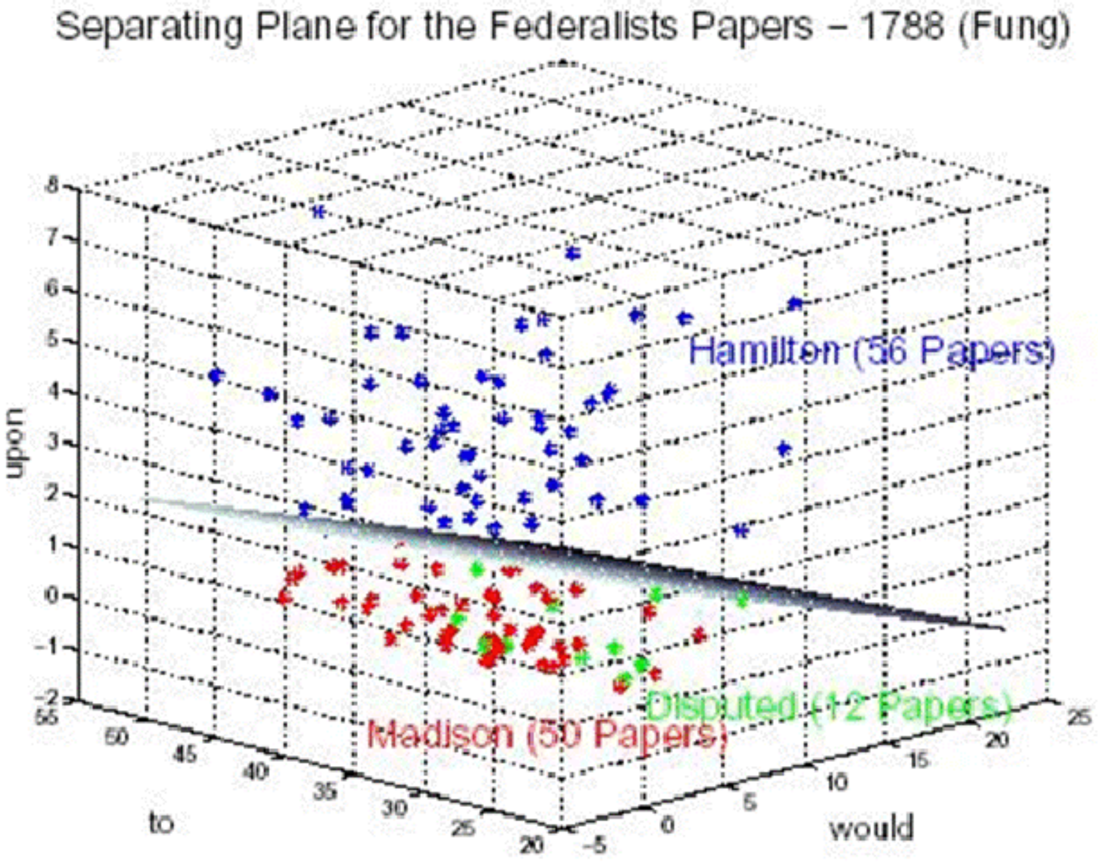


Figure 1: Obtained Hyperplane in 3 dimensions

UNIVERSITY OF CALIFORNIA, BERKELEY
SCHOOL OF INFORMATION

INFO 202
"Information Organization & Retrieval"
Fall 2013

Robert J. Glushko
glushko@berkeley.edu
@rjglushko

3 December 2013
Lecture 27.4 – Spam Classification

Classifying Spam

- Spam can be defined as “unsolicited, unwanted mail that was sent by automated means, directly or indirectly, by a sender having no current relationship with the recipient”
- (Fake reviews on Yelp, Trip Advisor, etc can also be viewed as spam, but they are not usually automated)
- Classifying email as "spam" or "not spam" using the simple and obvious approach of classifying messages as "spam" when they contain words most often contained in spam messages yields many false positives
- But if you are conservative you have too many misses

Classifying Spam

- Miss? False positives?
- These two kinds of classification mistakes are not equal in costs: what is worse, missing so that something that is spam gets through, or calling something spam that isn't?
- Answering these questions requires that we make a slight detour into probability theory and hypothesis testing

Hypothesis Testing [1]

- We assume that there is some "true" state or value - called the "null hypothesis" - and we conduct some tests or make some observations to determine whether to believe it or to instead reject it and accept an "alternative hypothesis"
- Example null hypotheses - this message isn't spam, the patient doesn't have the disease, the defendant is innocent, the graduation rate for starting football players is 90%
- Alternative hypotheses - this message is spam, the patient has the disease, the defendant is guilty, the graduation rate isn't 90%
- We conduct experiments / make observations to determine if we should reject the null hypothesis

Hypothesis Testing [2]

- No test is perfect
- The number of observations we make and their variability gives us more or less confidence about the hypotheses
- Our experiments or observations may suggest that the null hypothesis is false - that is, a "positive" test that the patient has the disease, the defendant is guilty, the message is spam, the graduation rate for starting football players isn't 90%
- Or the results might be "negative" and not provide enough evidence for the disease, conviction, etc.

Type I and Type II Errors

- A *Type I error* or *false positive* is the error of rejecting a null hypothesis when it is in fact true; the supposedly positive evidence was observed due to chance (classifying a message as spam when it isn't)
- A *Type II error* or *false negative* is the error of not rejecting a null hypothesis when the alternative hypothesis is the true state of nature; the test or observations made weren't powerful enough to detect the evidence that was there (failing to catch spam)
- http://www.intuitor.com/statistics/CurveApplet.html shows how differences in power and confidence levels affect the proportions of Type I and Type II errors

Thinking About Probabilities

- Most people think of probability using a frequentist approach, which focuses on identifying the "true" probability of some event, defined as the limit of its relative frequency in a large number of trials or samples
- In contrast, the Bayesian approach is a more subjective interpretation of probability, defined as a person's degree of belief about some event
- This degree of belief, called the *prior probability*, is then changed by any data or observations - i.e., your opinion can change if you get new information
- You updated degree of belief, the *posterior probability*, is computed using Bayes' Theorem

Bayes' Theorem

- Proposes how a subjective degree of belief should rationally change to account for evidence
- For Proposition A and Evidence B:

$$p(A|B) = \frac{p(A)p(B|A)}{p(B)}$$

- p(A) is the initial belief or *prior probability*
- P(A|B) is the *posterior probability*, is the revised belief after accounting for B

Accounting for “False Positives”

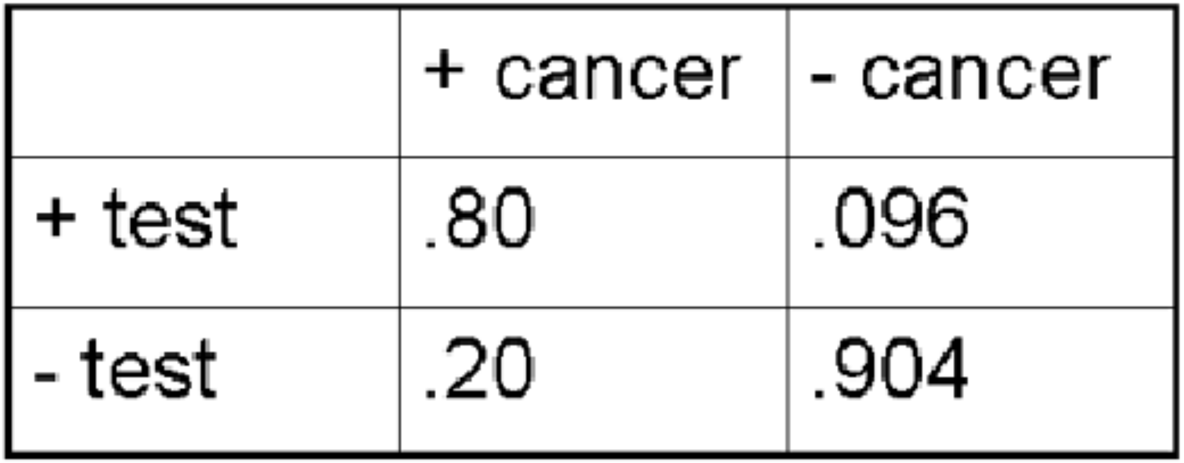
- We can rewrite the denominator to make it clear when we calculate that we are considering all the possible outcomes

$$p(B) = p(B|A)\,p(A) + p(B|{\sim}A)\,p({\sim}A)$$

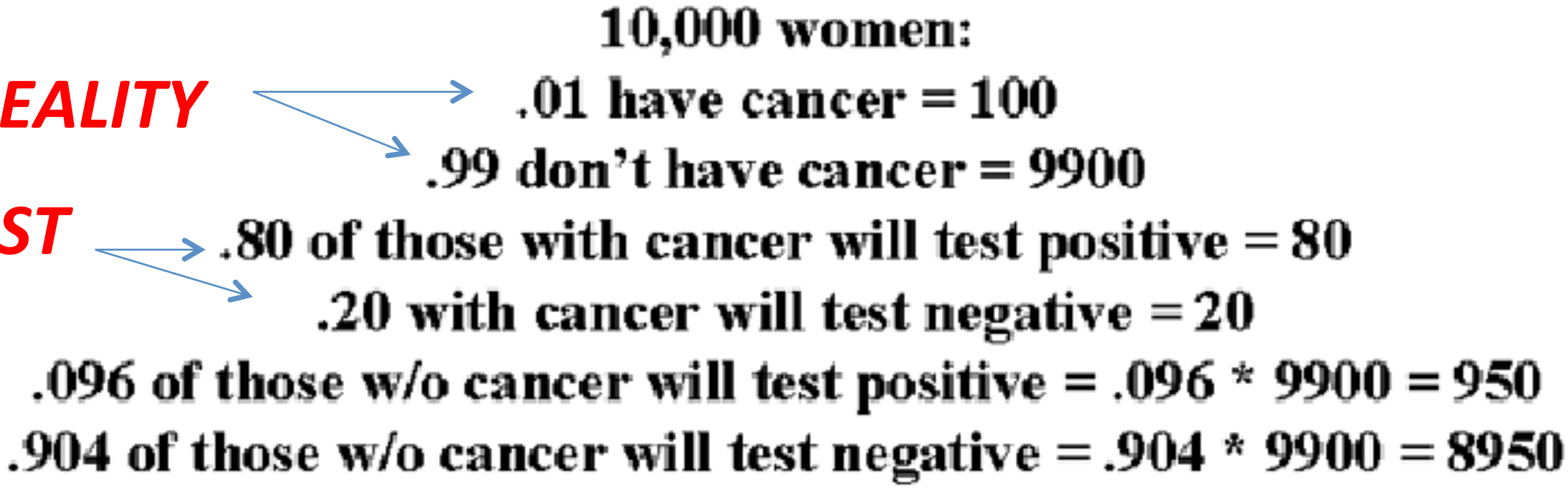
True Positive: A is true and B happened

False Positive: A is not true and B happened

Bayes’ Theorem Example



	+ cancer	- cancer
+ test	.80	.096
- test	.20	.904



10,000 women:

REALITY

.01 have cancer = 100

.99 don’t have cancer = 9900

TEST

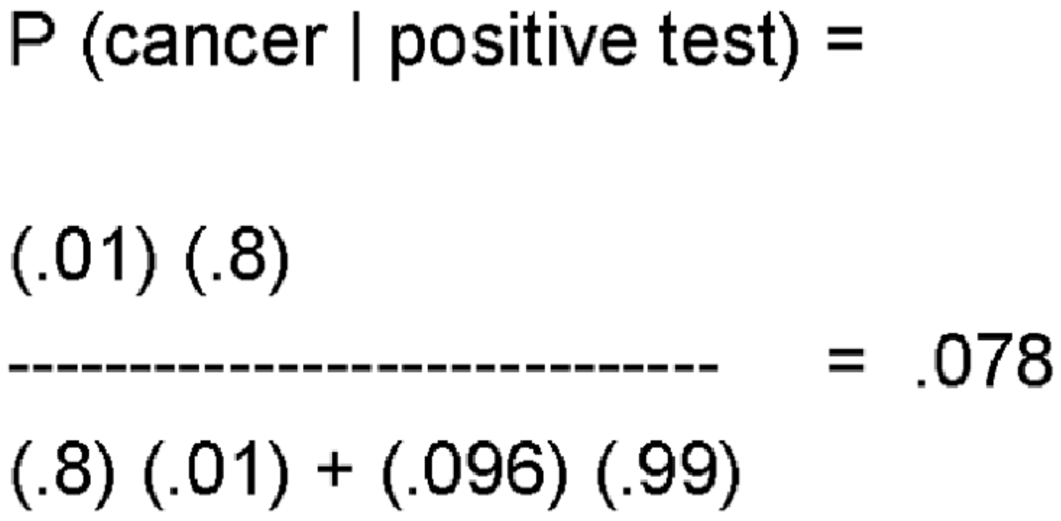
.80 of those with cancer will test positive = 80

.20 with cancer will test negative = 20

.096 of those w/o cancer will test positive = .096 * 9900 = 950

.904 of those w/o cancer will test negative = .904 * 9900 = 8950

Bayes' Theorem Example



P (cancer | positive test) =

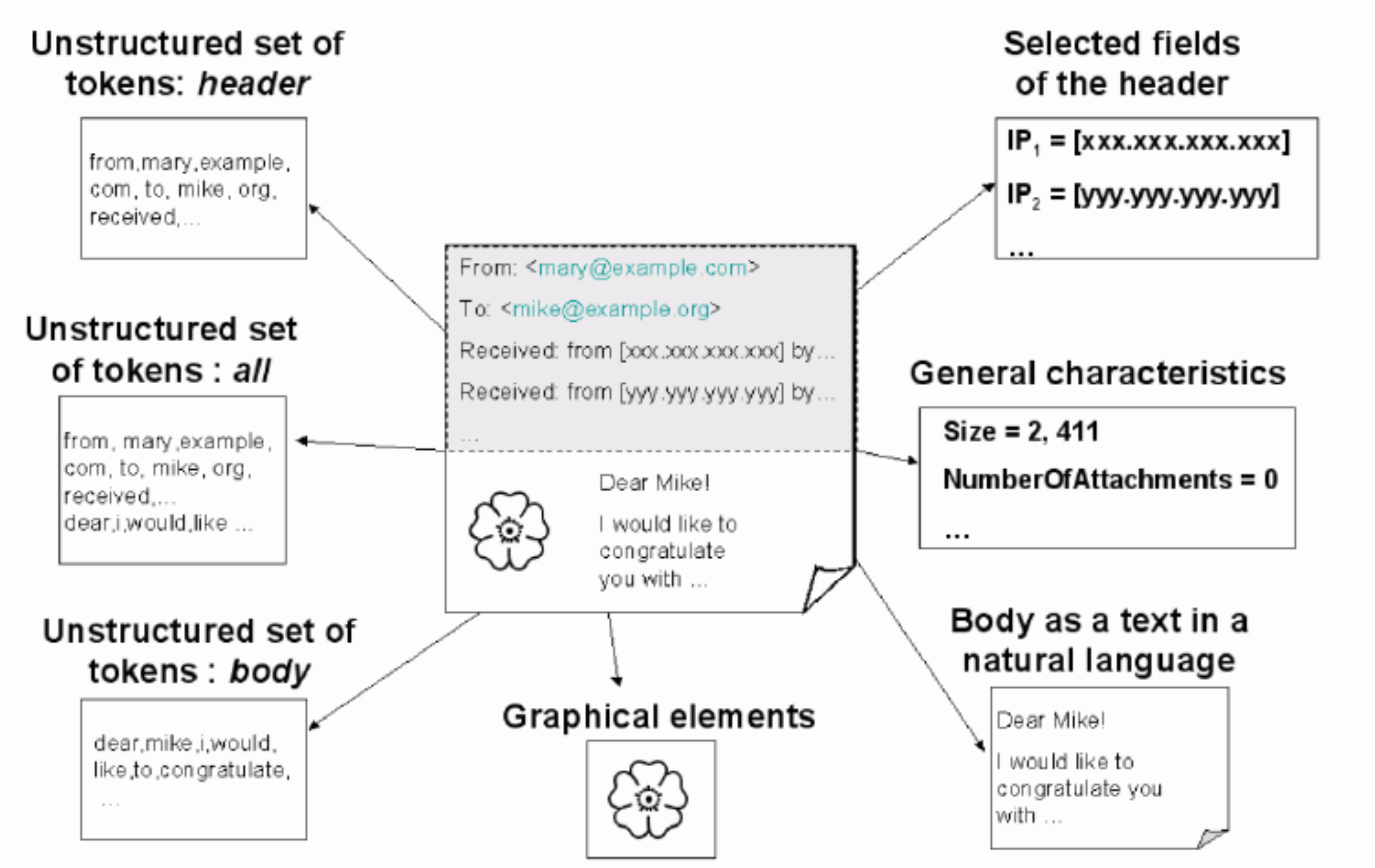
$$\frac{(.01)\ (.8)}{(.8)\ (.01) + (.096)\ (.99)} = .078$$

- So even though the test is 80% accurate at detecting cancer, cancer is rare (1%) – which means that a 9.6% false positive rate is a substantial concern

Baysian Spam Classifiers

- Bayesian approaches to spam classification assign a "spam probability" to each word, then combines them into a single probability for the email. This combined score considers the good and bad words in an email
- This approach evolves with spam as it learns new words and considers their probabilities
- Trying to trick a Bayesian filter with misspelled words like "V1AG RA" just trains it to be more reliable because that string has 0 probability in non-spam messages
- Very sophisticated spam classifiers use multiple features of messages, not just body content

What to Analyze to Classify Spam?



From: <mary@example.com>

To: <mike@example.org>

Received: from [xxx.xxx.xxx.xxx] by…

Received: from [yyy.yyy.yyy.yyy] by…

…

Dear Mike!

I would like to congratulate you with …

Unstructured set of tokens: *header*

from,mary,example, com, to, mike, org, received,…

Unstructured set of tokens : *all*

from, mary,example, com, to, mike, org, received,… dear,i,would,like …

Unstructured set of tokens : *body*

dear,mike,i,would, like,to,congratulate, …

Graphical elements

Selected fields of the header

IP$_1$ = [xxx.xxx.xxx.xxx]

IP$_2$ = [yyy.yyy.yyy.yyy]

…

General characteristics

Size = 2, 411

NumberOfAttachments = 0

…

Body as a text in a natural language

Dear Mike!

I would like to congratulate you with …

UNIVERSITY OF CALIFORNIA, BERKELEY
SCHOOL OF INFORMATION

INFO 202
"Information Organization & Retrieval"
Fall 2013

Robert J. Glushko
glushko@berkeley.edu
@rjglushko

3 December 2013
Lecture 27.5 – "Answer Machines"

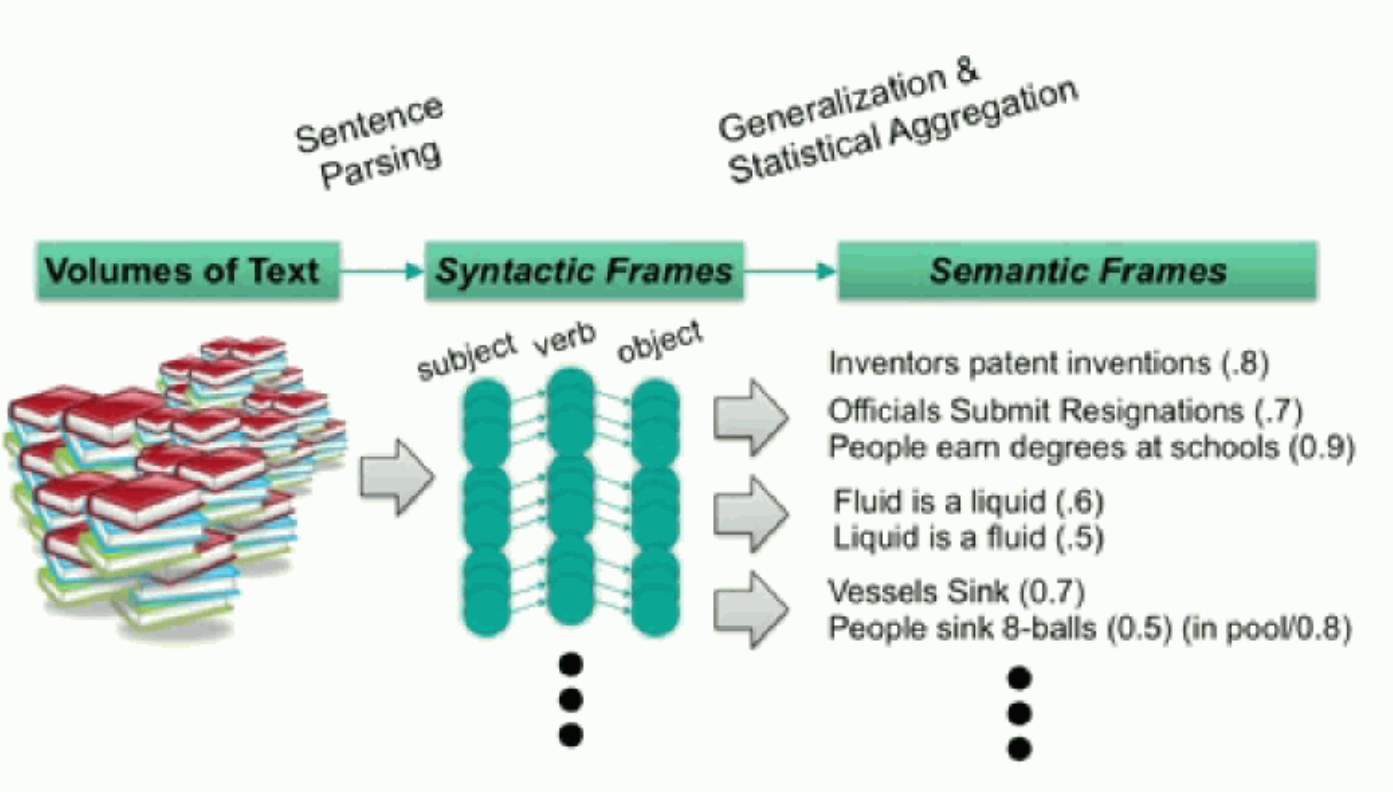
Question Answering Systems

- QA systems have been built on both ends of the language vs statistical learning dimension
- QA systems use IE techniques to identify the parts of the retrieved documents where the questions are most likely to be answered
- Statistical systems rewrite the question into multiple queries in which the keywords occur in different orders
- This increases the probability of finding the answers, but is very inefficient, so they use Bayes’ Rule to learn which query rewrites are best and stop doing useless ones

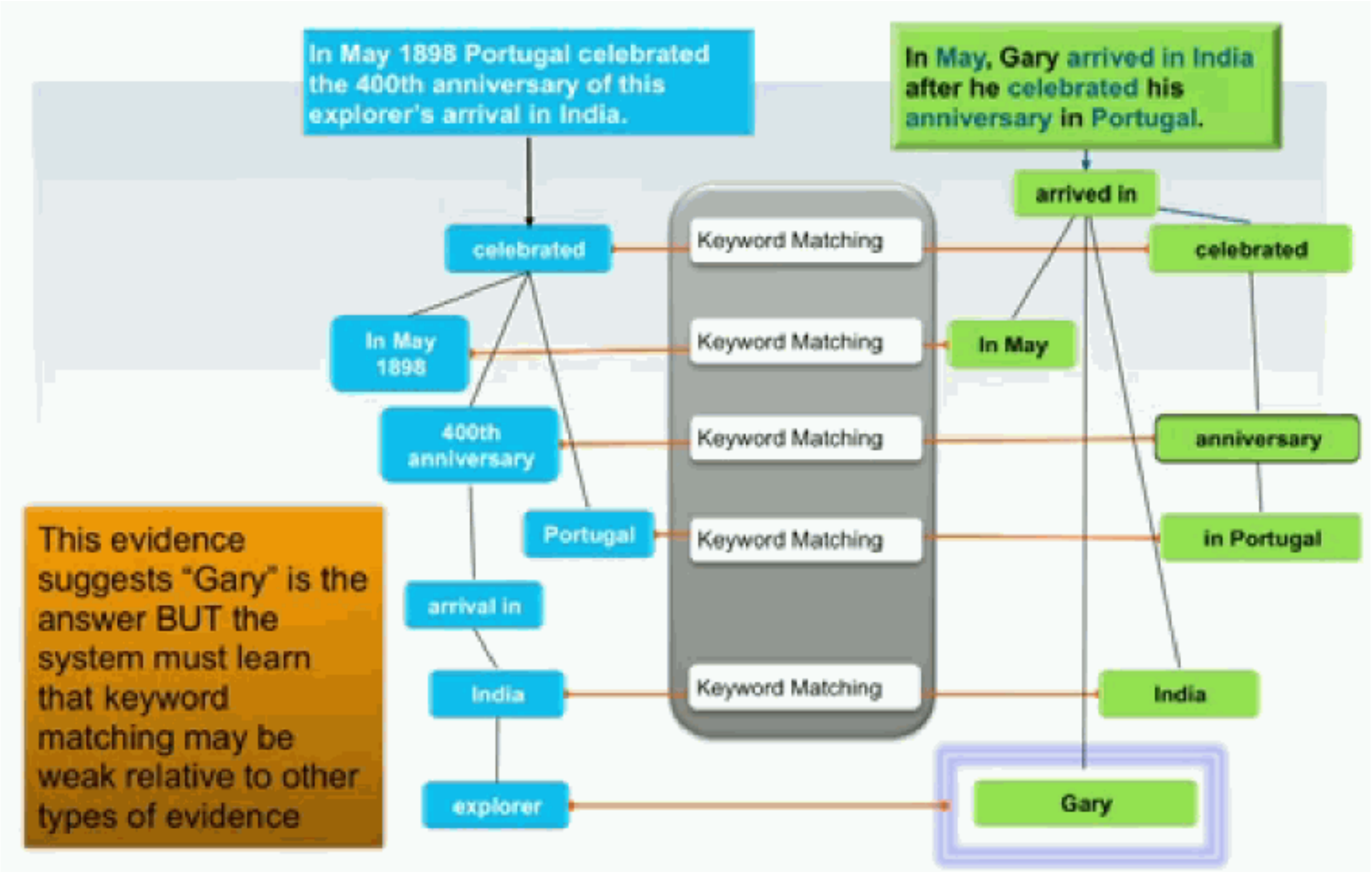
Watson Plays Jeopardy

- Watson Beats the Human Champs
- Jeopardy uses a broad and open knowledge domain, uses complexly (with puns and abbreviations) worded clues, demands precise answers, and you have to be quick!
- A compelling technology (and marketing) demonstration for IBM

Watson – Learning from Reading

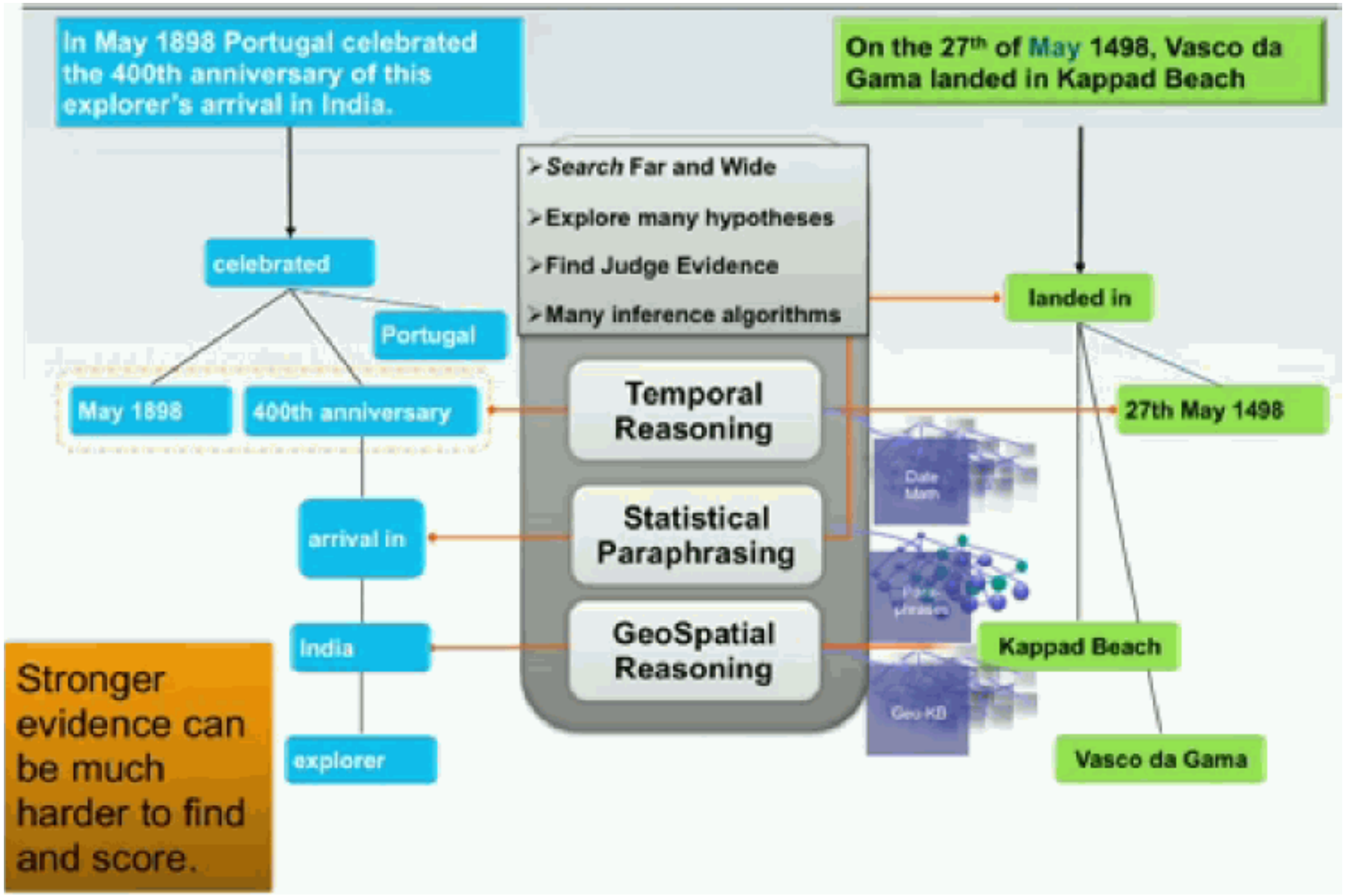

Sentence Parsing
Generalization & Statistical Aggregation
Volumes of Text
Syntactic Frames
Semantic Frames
subject
verb
object
Inventors patent inventions (.8)
Officials Submit Resignations (.7)
People earn degrees at schools (0.9)
Fluid is a liquid (.6)
Liquid is a fluid (.5)
Vessels Sink (0.7)
People sink 8-balls (0.5) (in pool/0.8)

Watson- Why Keywords Won’t Work



This evidence suggests “Gary” is the answer BUT the system must learn that keyword matching may be weak relative to other types of evidence

Watson – Using Deeper Evidence



In May 1898 Portugal celebrated the 400th anniversary of this explorer's arrival in India.

On the 27th of May 1498, Vasco da Gama landed in Kappad Beach

- *Search* Far and Wide
- Explore many hypotheses
- Find Judge Evidence
- Many inference algorithms

Stronger evidence can be much harder to find and score.

Reading For Next Lecture

- TDO Chapter 10,
  “The Organizing System Roadmap”