Applications (1 of 2): Information Retrieval

Kenneth Church

Kenneth.Church@jhu.edu

Pattern Recognition Problems in Computational Linguistics

- Information Retrieval:
  - Is this doc more like relevant docs or irrelevant docs?
- Author Identification:
  - Is this doc more like author A’s docs or author B’s docs?
- Word Sense Disambiguation
  - Is the context of this use of *bank*
    - more like sense 1’s contexts
    - or like sense 2’s contexts?
- Machine Translation
  - Is the context of this use of *drug* more like those that were translated as *drogue*
  - or those that were translated as *medicament*?

Applications of Naïve Bayes

Word Sense Disambiguation (WSD)

$$score(context) = \prod_{word\ in\ context} \frac{Pr(word|sense_1)}{Pr(word|sense_2)}$$

Author Identification

$$score(doc) = \prod_{word\ in\ doc} \frac{Pr(word|author_1)}{Pr(word|author_2)}$$

Information Retrieval (IR)

$$score(doc) = \prod_{word\ in\ doc} \frac{Pr(word|relevant)}{Pr(word|irrelevant)}$$

Sentiment Analysis

$$score(doc) = \prod_{word\ in\ doc} \frac{Pr(word|positive\ review)}{Pr(word|negative\ review)}$$

Classical Information Retrieval (IR)

- Boolean Combinations of Keywords
  - – Dominated the Market (before the web)
  - – Popular with Intermediaries (Librarians)
- Rank Retrieval (Google)
  - – Sort a collection of documents
    - (e.g., scientific papers, abstracts, paragraphs)
    - by how much they ‘‘match’’ a query
  - – The query can be a (short) sequence of keywords
    - or arbitrary text (e.g., one of the documents)

Motivation for Information Retrieval
(circa 1990, about 5 years before web)

- Text is available like never before
- Currently, N≈100 million words
  - and projections run as high as 10^{15} bytes by 2000!
- What can we do with it all?
  - It is better to do something simple,
  - than nothing at all.
- IR vs. Natural Language Understanding
  - Revival of 1950-style empiricism

How Large is Very Large?

From a Keynote to EMNLP Conference, formally Workshop on Very Large Corpora

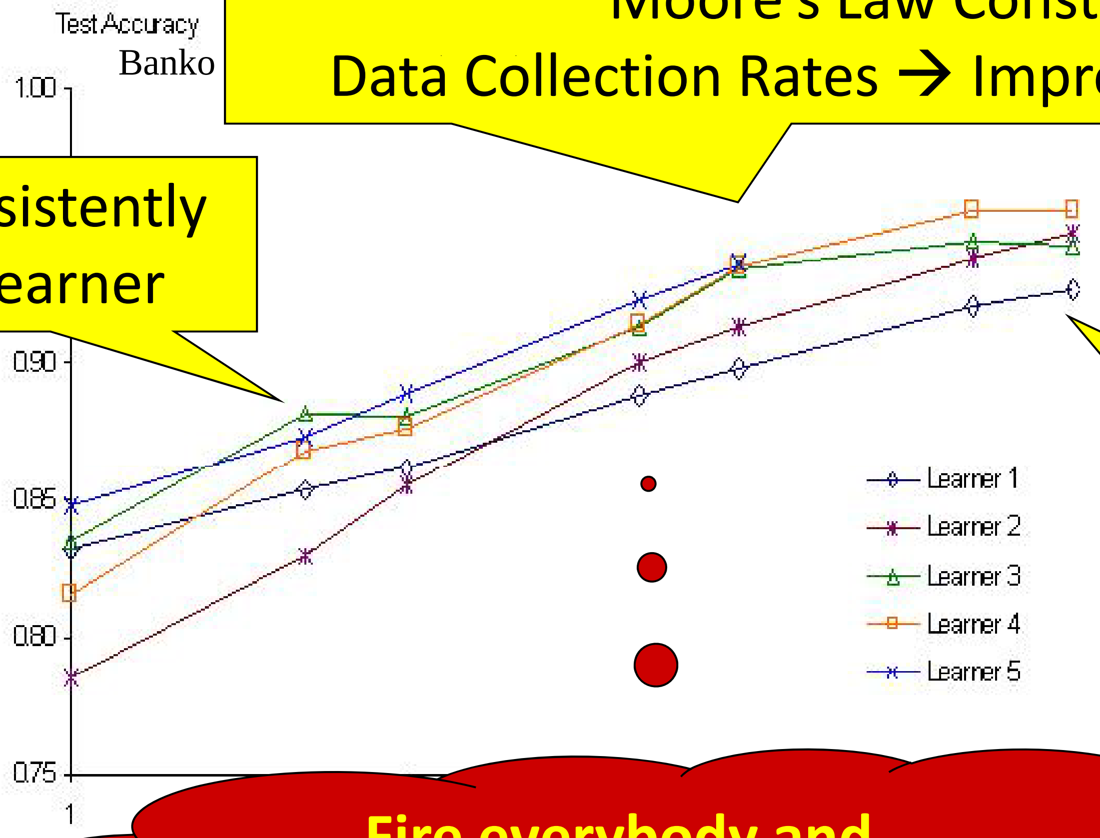
Year	Source	Size (words)
1788	Federalist Papers	1/5 million
1982	Brown Corpus	1 million
1987	Birmingham Corpus	20 million
1988-	Associate Press (AP)	50 million (per year)
1993	MUC, TREC, Tipster	

Rising Tide of Data Lifts All Boats

If you have a lot of data, then you don't need a lot of methodology

- 1985: *"There is no data like more data"*
  - Fighting words uttered by radical fringe elements (Mercer at Arden House)
- 1993 Workshop on Very Large Corpora
  - Perfect timing: Just before the web
  - Couldn't help but succeed
  - Fate
- 1995: The Web changes everything
- All you need is data (magic sauce)
  - No linguistics
  - No artificial intelligence (representation)
  - No machine learning
  - No statistics
  - No error analysis

"It never pays to think until you've run out of data" – Eric Brill



Moore's Law Constant:
Data Collection Rates → Improvement Rates

No consistently best learner

More data is better data!

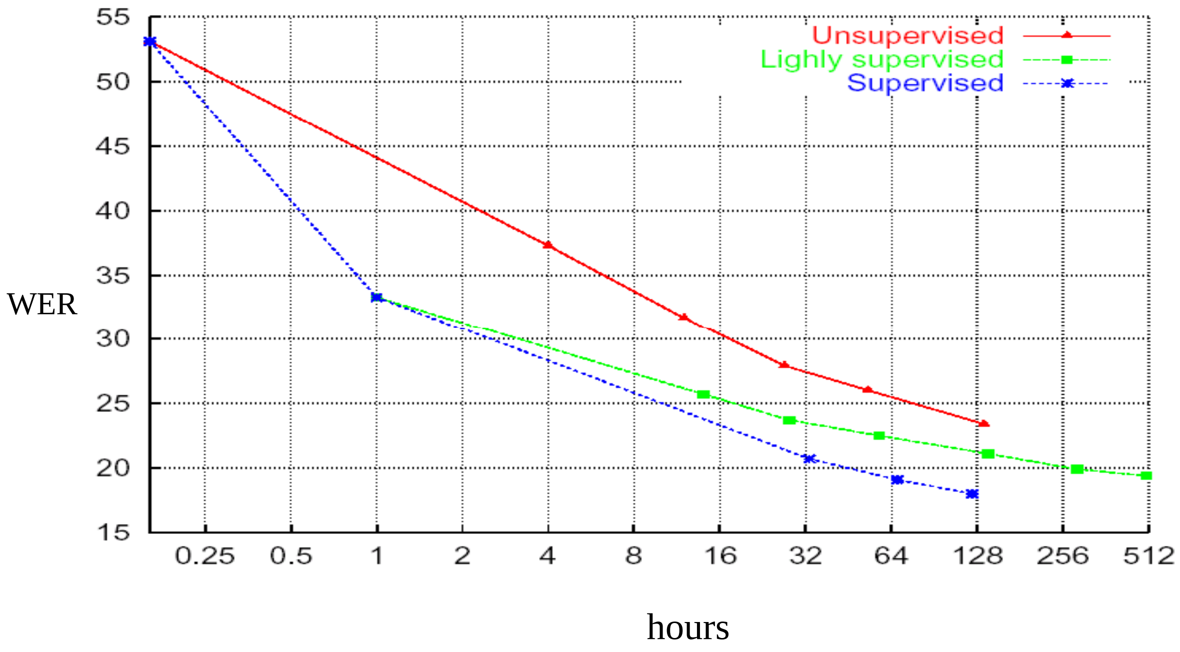
Quoted out of context

Fire everybody and spend the money on data

Borrowed Slide: Jelinek (LREC)

Benefit of Data

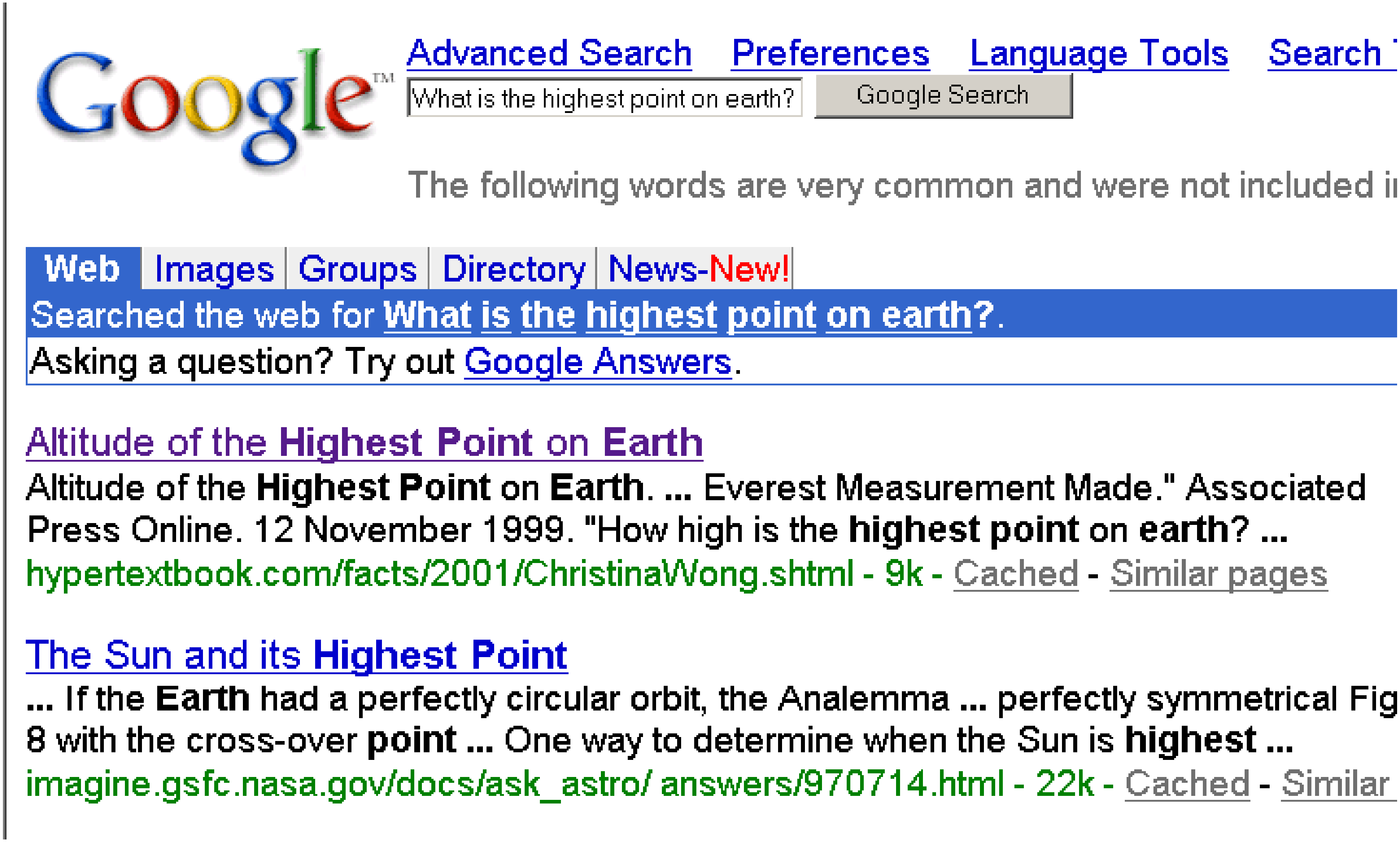
LIMSI: Lamel (2002) – Broadcast News



Supervised: transcripts
Lightly supervised: closed captions

The rising tide of data will lift all boats!
TREC Question Answering & Google:

What is the highest point on Earth?



The following words are very common and were not included i

Searched the web for What is the highest point on earth?.

Asking a question? Try out Google Answers.

Altitude of the **Highest Point** on **Earth**

Altitude of the **Highest Point** on **Earth**. ... Everest Measurement Made." Associated Press Online. 12 November 1999. "How high is the **highest point** on **earth**? ...
hypertextbook.com/facts/2001/ChristinaWong.shtml - 9k - Cached - Similar pages

The Sun and its **Highest Point**

... If the **Earth** had a perfectly circular orbit, the Analemma ... perfectly symmetrical Fig 8 with the cross-over **point** ... One way to determine when the Sun is **highest** ...
imagine.gsfc.nasa.gov/docs/ask_astro/ answers/970714.html - 22k - Cached - Similar

The rising tide of data will lift all boats!
Acquiring Lexical Resources from Data:

Dictionaries, Ontologies, WordNets, Language Models, etc.

http://labs1.google.com/sets

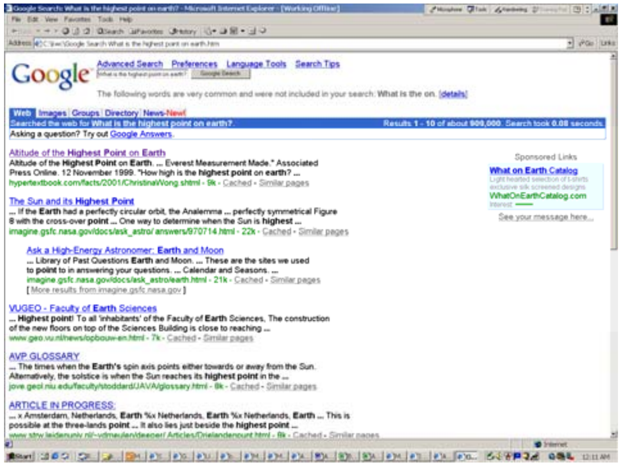
England

France

Rising Tide of Data Lifts All Boats

If you have a lot of data, then you don't need a lot of methodology

- More data → better results
  - TREC Question Answering
    - Remarkable performance: Google and not much else
      - Norvig (ACL-02)
      - AskMSR (SIGIR-02)
  - Lexical Acquisition
    - Google Sets
      - We tried similar things
        - » but with *tiny* corpora
        - » which we called *large*



The following words are very common and were not included in your search: What is the on. [details]

Searched the web for What is the highest point on earth?. Results 1 - 10 of about 909,000. Search took 0.08 seconds.

Asking a question? Try out Google Answers.

Altitude of the Highest Point on Earth
Altitude of the Highest Point on Earth. ... Everest Measurement Made." Associated Press Online. 12 November 1999. "How high is the highest point on earth? ...
hypertextbook.com/facts/2001/ChristinaWong.shtml - 9k - Cached - Similar pages

The Sun and its Highest Point
... If the Earth had a perfectly circular orbit, the Analemma ... perfectly symmetrical Figure 8 with the cross-over point ... One way to determine when the Sun is highest ...
imagine.gsfc.nasa.gov/docs/ask_astro/ answers/970714.html - 22k - Cached - Similar pages

Ask a High-Energy Astronomer: Earth and Moon
... Library of Past Questions Earth and Moon. ... These are the sites we used to point to in answering your questions. ... Calendar and Seasons. ...
imagine.gsfc.nasa.gov/docs/ask_astro/earth.html - 21k - Cached - Similar pages
[More results from imagine.gsfc.nasa.gov]

VUGEO - Faculty of Earth Sciences
... Highest point! To all 'inhabitants' of the Faculty of Earth Sciences, The construction of the new floors on top of the Sciences Building is close to reaching ...
www.geo.vu.nl/news/opbouw-en.html - 7k - Cached - Similar pages

AVP GLOSSARY
... The times when the Earth's spin axis points either towards or away from the Sun. Alternatively, the solstice is when the Sun reaches its highest point in the ...
jove.geol.niu.edu/faculty/stoddard/JAVA/glossary.html - 8k - Cached - Similar pages

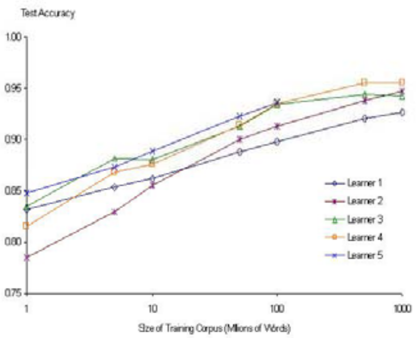
ARTICLE IN PROGRESS:
... x Amsterdam, Netherlands, Earth %x Netherlands, Earth %x Netherlands, Earth ... This is possible at the three-lands point ... It also lies just beside the highest point ...
www.strw.leidenuniv.nl/~vdmeulen/deeper/ Articles/Drielandenpunt.html - 8k - Cached - Similar pages

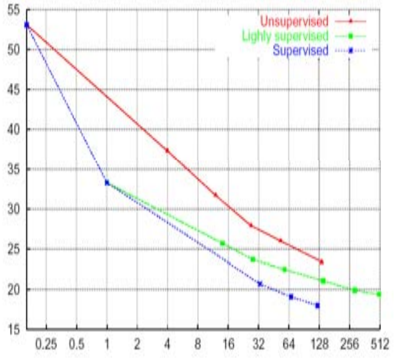
Sponsored Links
What on Earth Catalog
Light hearted selection of t-shirts exclusive silk screened designs
WhatOnEarthCatalog.com
Interest
See your message here...

http://labs1.google.com/sets

Cat	cat	England	Japan
Dog	more	France	China
Horse	ls	Germany	India
Fish	rm	Italy	Indonesia
Bird	mv	Ireland	Malaysia
Rabbit	cd	Spain	Korea
Cattle	cp	Scotland	Taiwan
Rat	mkdir	Belgium	Thailand
Livestock	man	Canada	Singapore
Mouse	tail	Austria	Australia
Human	pwd	Australia	Bangladesh

Eurospeech 2003





Applications

Don’t worry;
Be happy

5 Ian Andersons

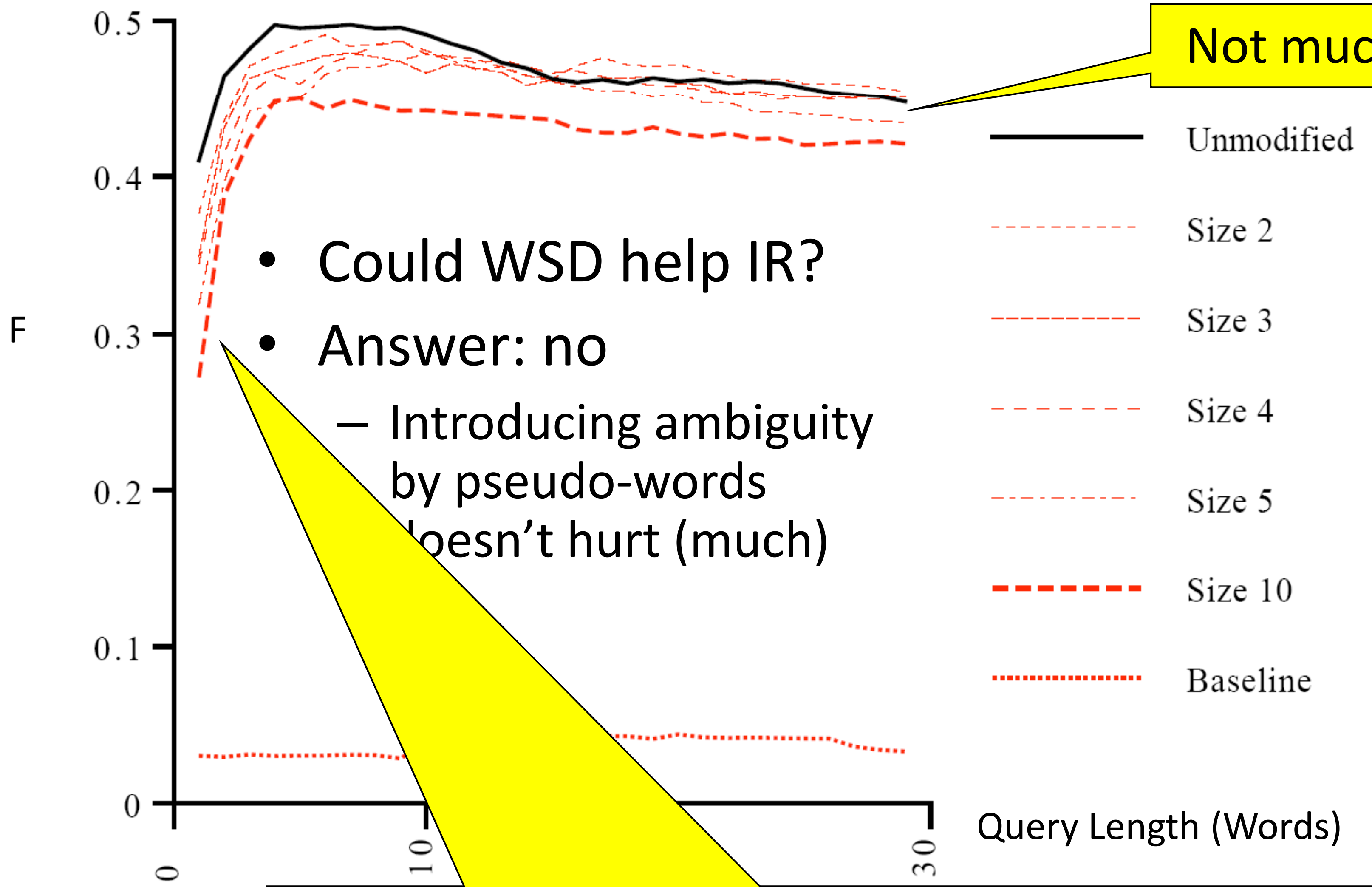
- What good is word sense disambiguation (WSD)?
  - Information Retrieval (IR)
    - Salton: Tried hard to find ways to use NLP to help IR
      - but failed to find much (if anything)
    - Croft: WSD doesn’t help because IR is already using those methods
    - Sanderson (next two slides)
  - Machine Translation (MT)
    - Original motivation for much of the work on WSD
    - But IR arguments may apply just as well to MT
- What good is POS tagging? Parsing? NLP? Speech?
- *Commercial Applications of Natural Language Processing*, CACM 1995
  - $100M opportunity (worthy of government/industry’s attention)
    1. Search (Lexis-Nexis)
    2. Word Processing (Microsoft)

ALPAC

- Warning: premature commercialization is risky

Sanderson (SIGIR-94)

http://dis.shef.ac.uk/mark/cv/publications/papers/my_papers/SIGIR94.pdf



Not much?

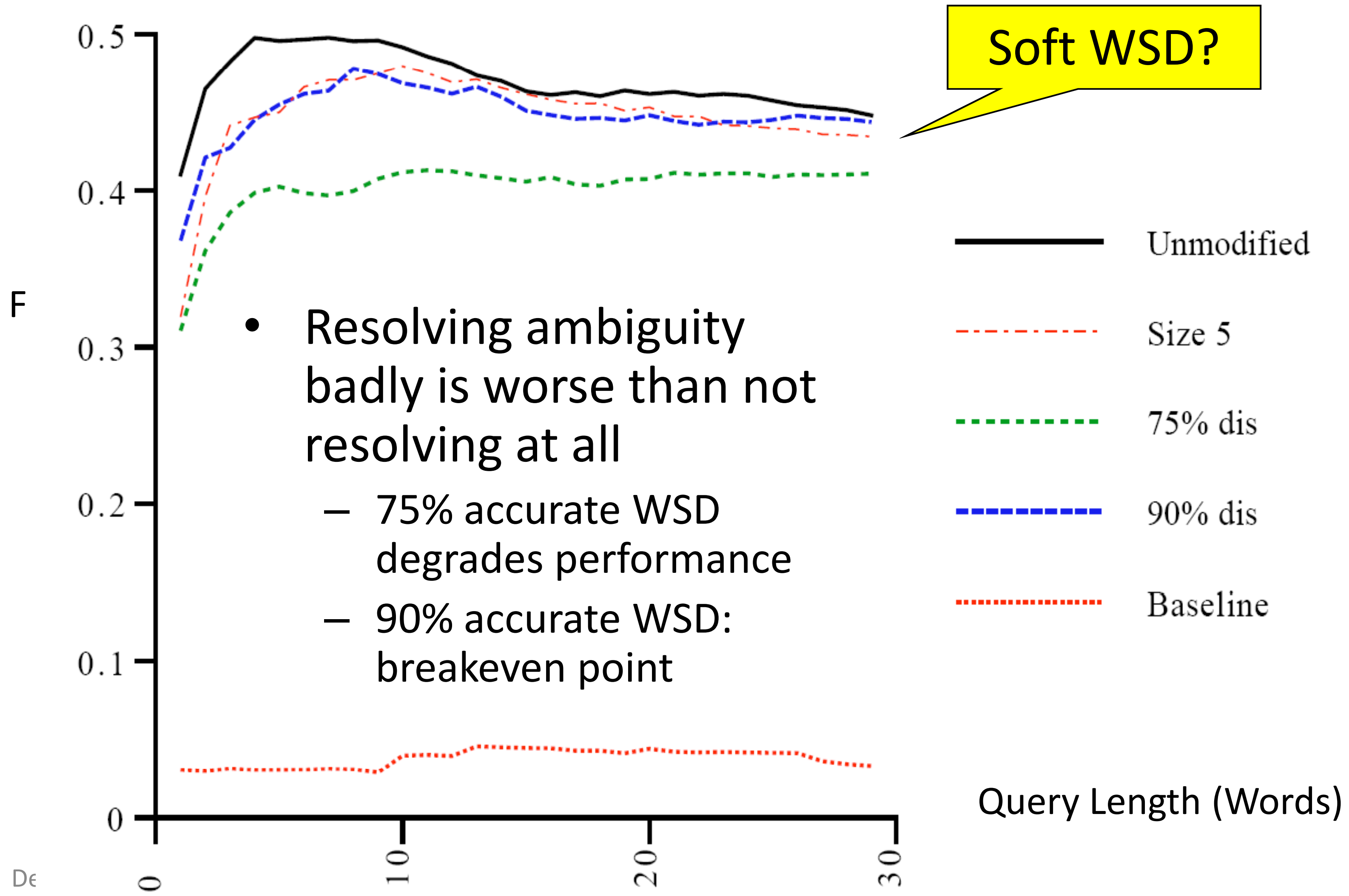
- Could WSD help IR?
- Answer: no
  - Introducing ambiguity by pseudo-words doesn't hurt (much)

5 Ian Andersons

Short queries matter most, but hardest for WSD

Sanderson (SIGIR-94)

http://dis.shef.ac.uk/mark/cv/publications/papers/my_papers/SIGIR94.pdf



Soft WSD?

- Resolving ambiguity badly is worse than not resolving at all
  - 75% accurate WSD degrades performance
  - 90% accurate WSD: breakeven point

IR Models

- Keywords (and Boolean combinations thereof)
- Vector-Space ‘‘Model’’ (Salton, chap 10.1)
  - Represent the query and the documents as V-dimensional vectors
  - Sort vectors by $$sim(x,y) = \cos(x,y) = \frac{\sum_i x_i \cdot y_i}{|x| \cdot |y|}$$
- Probabilistic Retrieval Model
  - (Salton, chap 10.3)
  - Sort documents by $$score(d) = \prod_{w \in d} \frac{\Pr(w \mid rel)}{\Pr(w \mid \overline{rel})}$$

Information Retrieval and Web Search

Alternative IR models

Instructor: Rada Mihalcea

Some of the slides were adopted from a course tought at Cornell University by William Y. Arms

Latent Semantic Indexing

Objective

Replace indexes that use **sets of index terms** by indexes that use **concepts**.

Approach

Map the term vector space into a lower dimensional space, using singular value decomposition.

Each dimension in the new space corresponds to a latent concept in the original data.

Deficiencies with Conventional Automatic Indexing

Synonymy: Various words and phrases refer to the same concept (lowers recall).

Polysemy: Individual words have more than one meaning (lowers precision)

Independence: No significance is given to two terms that frequently appear together

Latent semantic indexing addresses the first of these (synonymy),
and the third (dependence)

Bellcore’s Example

http://en.wikipedia.org/wiki/Latent_semantic_analysis

c1 Human machine *interface* for Lab ABC *computer* applications

c2 A *survey* of *user* opinion of computer *system response time*

c3 The EPS *user interface* management *system*

c4 *System* and *human system* engineering testing of EPS

c5 Relation of *user*-perceived *response time* to error measurement

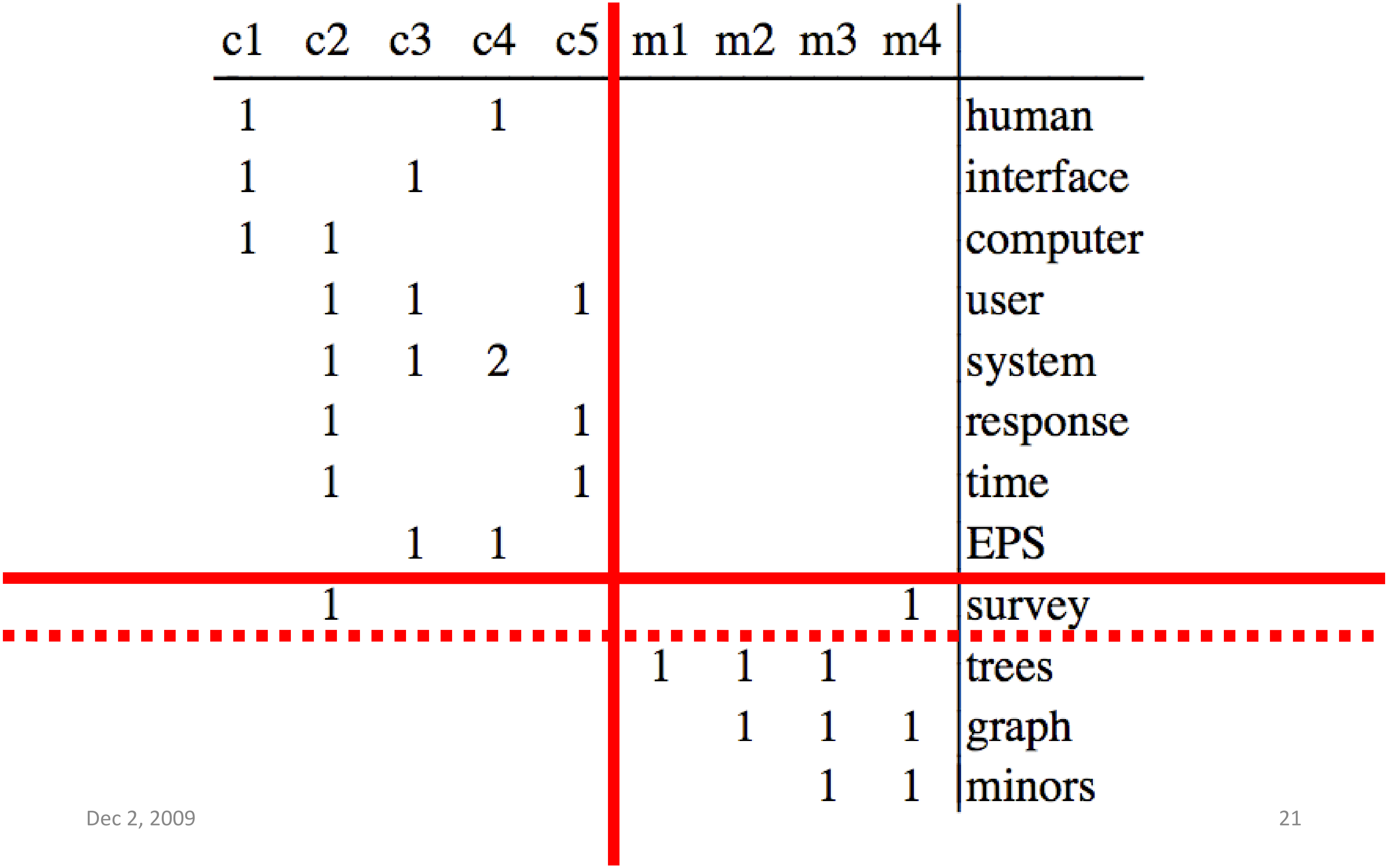
m1 The generation of random, binary, unordered *trees*

m2 The intersection *graph* of paths in *trees*

m3 *Graph minors* IV: Widths of *trees* and well-quasi-ordering

m4 *Graph minors*: A *survey*

Term by Document Matrix



c1	c2	c3	c4	c5	m1	m2	m3	m4	
1			1						human
1		1							interface
1	1								computer
	1	1		1					user
	1	1	2						system
	1			1					response
	1			1					time
		1	1						EPS
	1							1	survey
					1	1	1		trees
						1	1	1	graph
							1	1	minors

Query Expansion

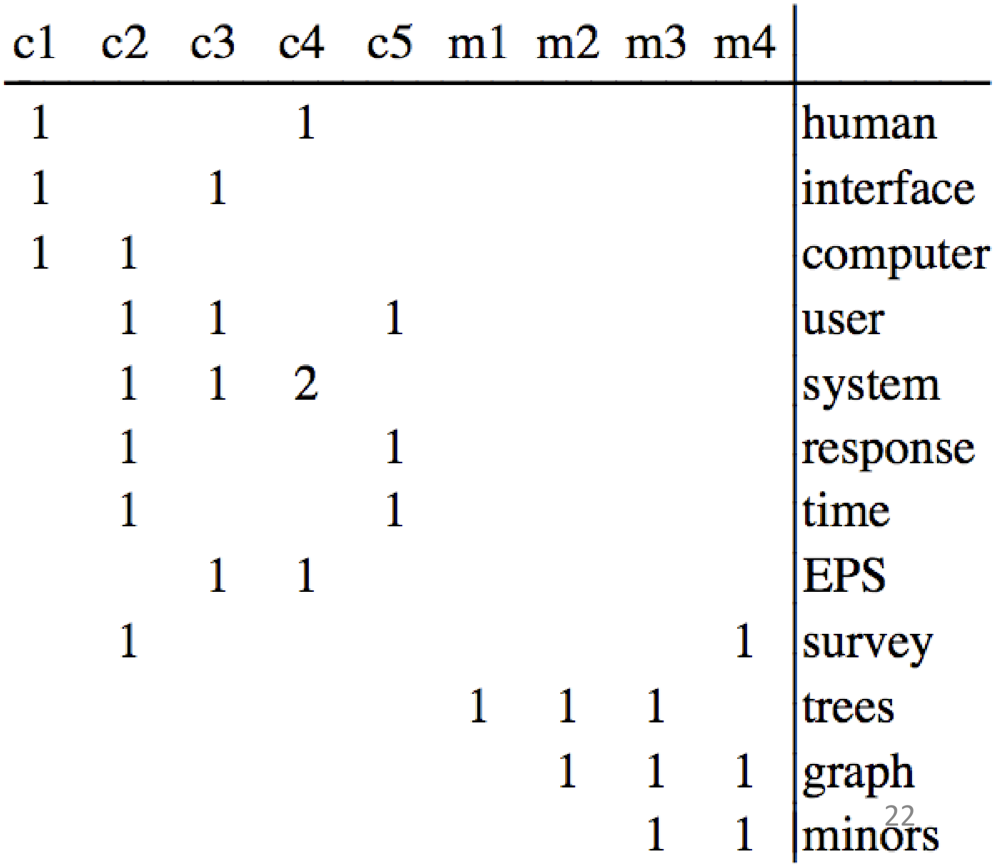
Query:

Find documents relevant to *human computer interaction*

Simple Term Matching:

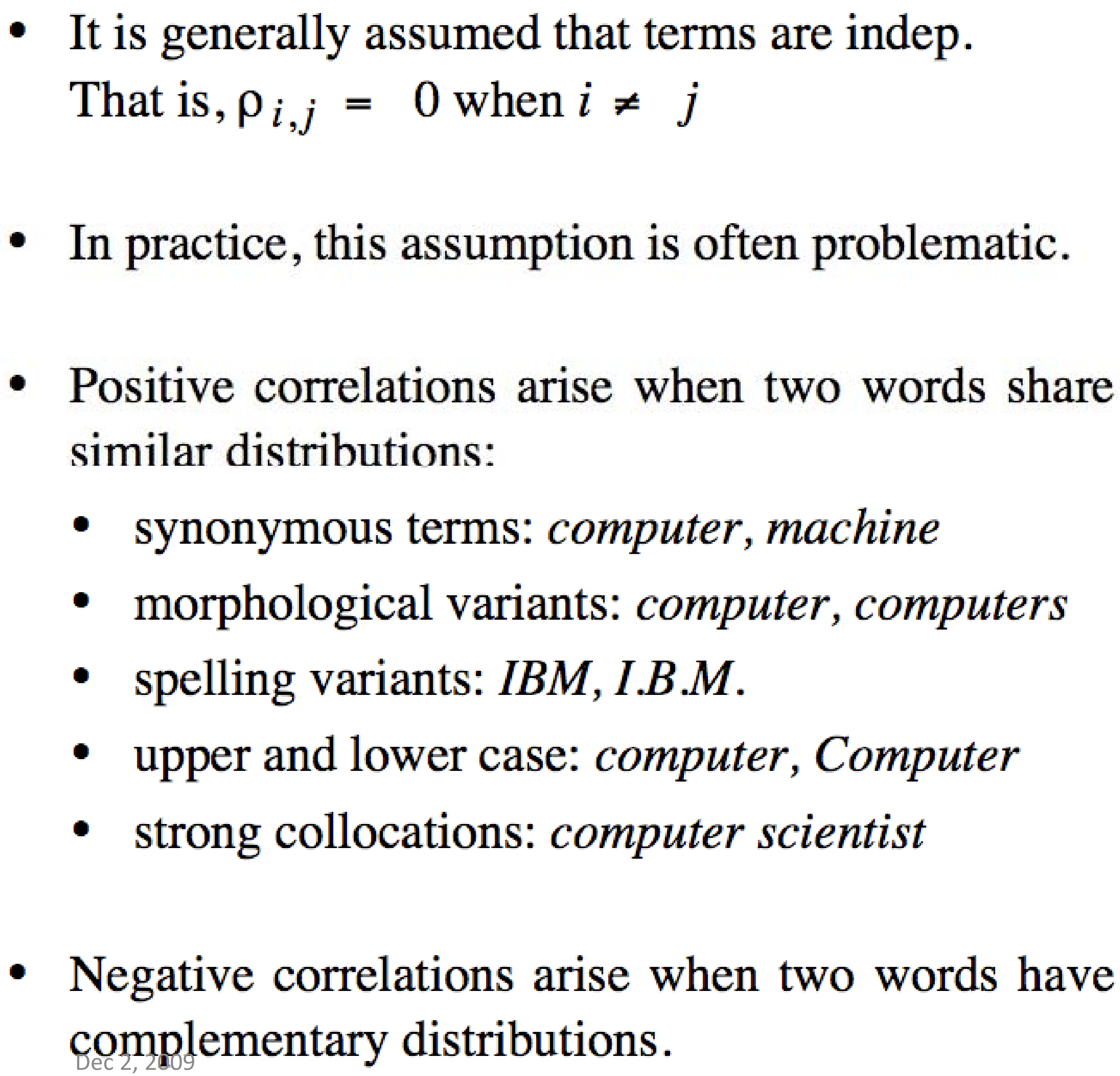
Matches c1, c2, and c4

Misses c3 and c5



c1	c2	c3	c4	c5	m1	m2	m3	m4	
1			1						human
1		1							interface
1	1								computer
	1	1		1					user
	1	1	2						system
	1			1					response
	1			1					time
		1	1						EPS
	1							1	survey
					1	1	1		trees
						1	1	1	graph
							1	1	minors

Large Correlations

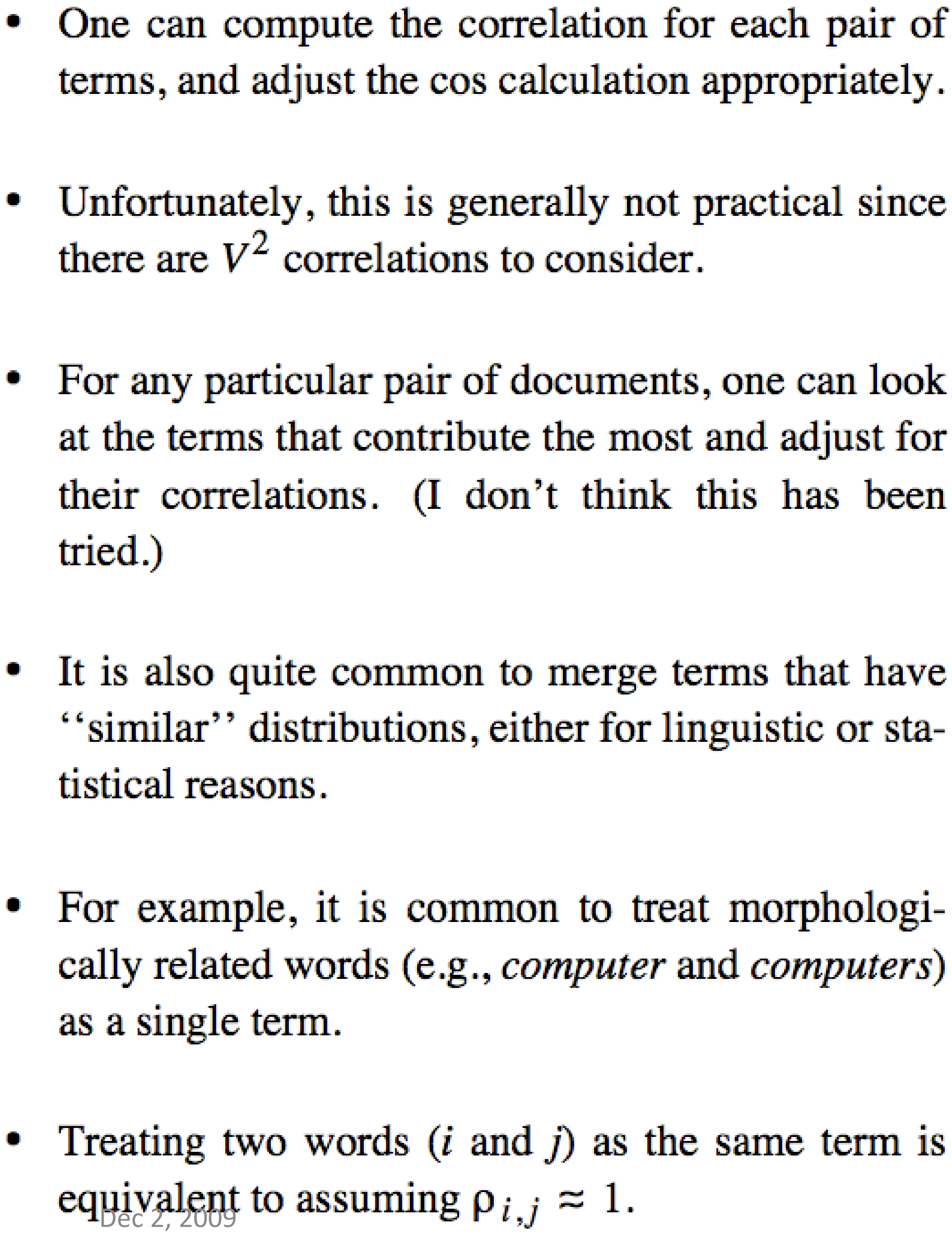


- It is generally assumed that terms are indep. That is, $\rho_{i,j} = 0$ when $i \neq j$
- In practice, this assumption is often problematic.
- Positive correlations arise when two words share similar distributions:
  - synonymous terms: *computer*, *machine*
  - morphological variants: *computer*, *computers*
  - spelling variants: *IBM*, *I.B.M*.
  - upper and lower case: *computer*, *Computer*
  - strong collocations: *computer scientist*
- Negative correlations arise when two words have complementary distributions.

Correlations: Too Large to Ignore

human	interface	computer	user	system	response	time	EPS	survey	trees	graph	minors	
1.0	0.4	0.4	–0.4	0.4	–0.3	–0.3	0.4	–0.3	–0.4	–0.4	–0.3	human
0.4	1.0	0.4	0.2	0.04	–0.3	–0.3	0.4	–0.3	–0.4	–0.4	–0.3	interface
0.4	0.4	1.0	0.2	0.04	0.4	0.4	–0.3	0.4	–0.4	–0.4	–0.3	computer
–0.4	0.2	0.2	1.0	0.2	0.8	0.8	0.2	0.2	–0.5	–0.5	–0.4	user
0.4	0.04	0.04	0.2	1.0	0.04	0.04	0.8	0.04	–0.5	–0.5	–0.3	system
–0.3	–0.3	0.4	0.8	0.04	1.0	1.0	–0.3	0.4	–0.4	–0.4	–0.3	response
–0.3	–0.3	0.4	0.8	0.04	1.0	1.0	–0.3	0.4	–0.4	–0.4	–0.3	time
0.4	0.4	–0.3	0.2	0.8	–0.3	–0.3	1.0	–0.3	–0.4	–0.4	–0.3	EPS
–0.3	–0.3	0.4	0.2	0.04	0.4	0.4	–0.3	1.0	–0.4	0.2	0.4	survey
–0.4	–0.4	–0.4	–0.5	–0.5	–0.4	–0.4	–0.4	–0.4	1.0	0.5	0.2	trees
–0.4	–0.4	–0.4	–0.5	–0.5	–0.4	–0.4	–0.4	0.2	0.5	1.0	0.8	graph
–0.3	–0.3	–0.3	–0.4	–0.4	–0.3	–0.3	–0.3	0.4	0.2	0.8	1.0	minors

Correcting for Large Correlations

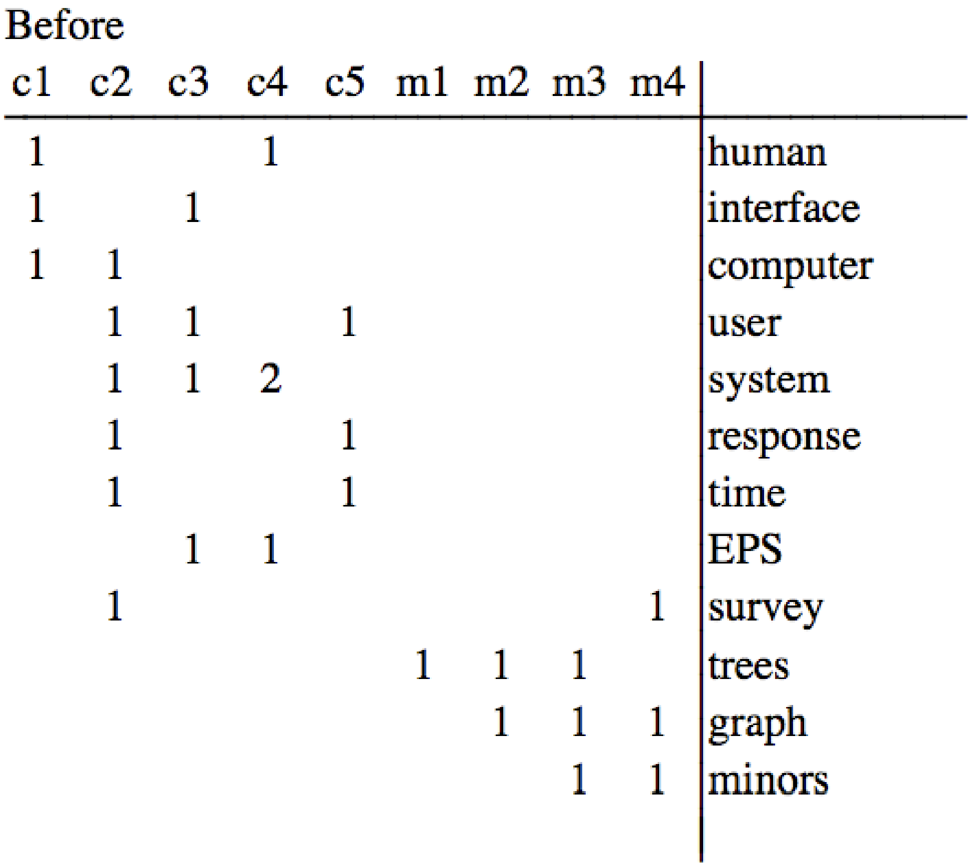


- One can compute the correlation for each pair of terms, and adjust the cos calculation appropriately.
- Unfortunately, this is generally not practical since there are V^2 correlations to consider.
- For any particular pair of documents, one can look at the terms that contribute the most and adjust for their correlations. (I don't think this has been tried.)
- It is also quite common to merge terms that have ‘‘similar’’ distributions, either for linguistic or statistical reasons.
- For example, it is common to treat morphologically related words (e.g., *computer* and *computers*) as a single term.
- Treating two words (*i* and *j*) as the same term is equivalent to assuming $\rho_{i,j} \approx 1$.

Thesaurus

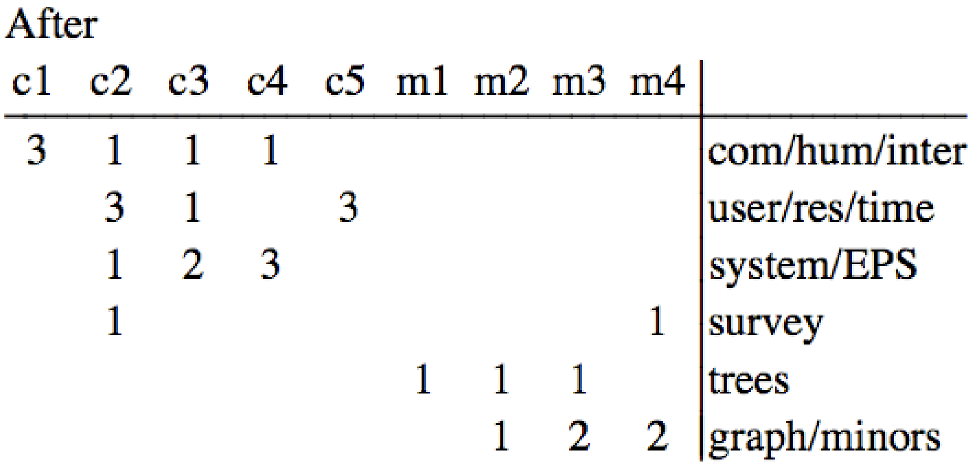
- Merge terms that cluster together
  - human/ interface/ computer
  - user/ response/ time
  - system/ EPS
  - graph/ minors

Term by Doc Matrix: Before & After Thesaurus



Before

c1	c2	c3	c4	c5	m1	m2	m3	m4	
1			1						human
1		1							interface
1	1								computer
	1	1		1					user
	1	1	2						system
	1			1					response
	1			1					time
		1	1						EPS
	1							1	survey
					1	1	1		trees
						1	1	1	graph
							1	1	minors

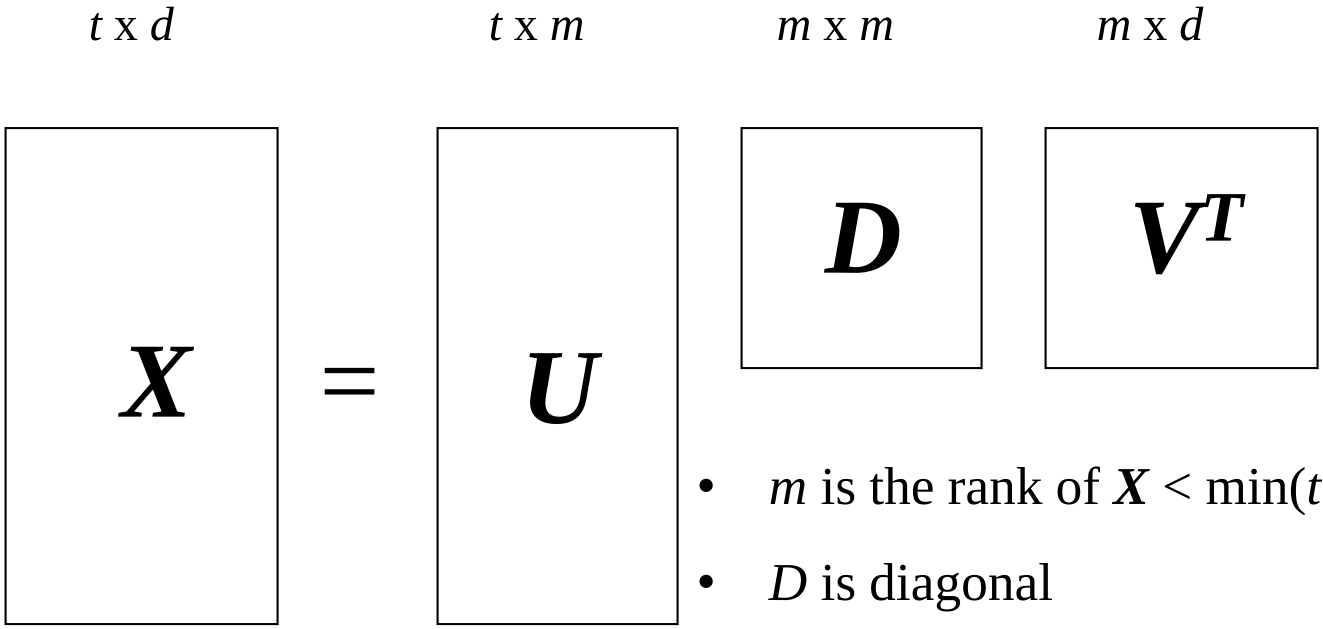


After

c1	c2	c3	c4	c5	m1	m2	m3	m4	
3	1	1	1						com/hum/inter
	3	1		3					user/res/time
	1	2	3						system/EPS
	1							1	survey
					1	1	1		trees
						1	2	2	graph/minors

Singular Value Decomposition (SVD)

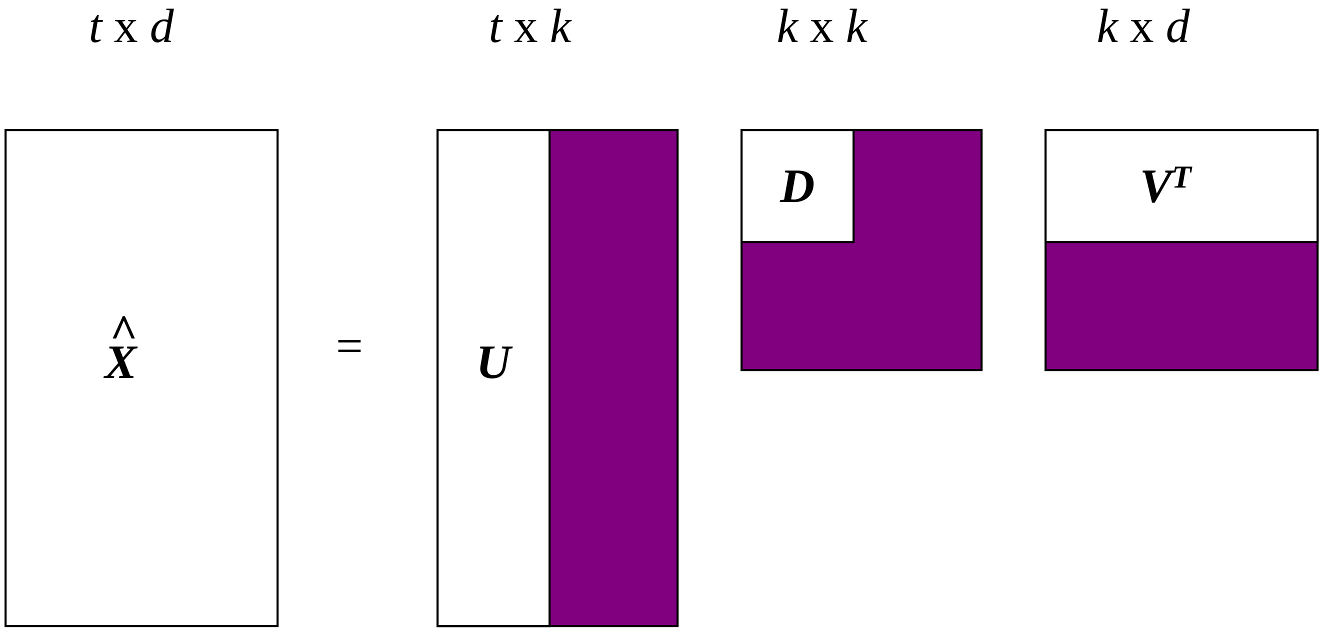
$$X = UDV^T$$


t x d $\quad$ t x m $\quad$ m x m $\quad$ m x d

$$\boldsymbol{X} = \boldsymbol{U} \; \boldsymbol{D} \; \boldsymbol{V^T}$$

- m is the rank of $\boldsymbol{X} < \min(t, d)$
- D is diagonal
  - D^2 are eigenvalues (sorted in descending order)
- $U U^T = I$ and $V V^T = I$
  - Columns of U are eigenvectors of $X X^T$
  - Columns of V are eigenvectors of $X^T X$

Dimensionality Reduction



k is the number of latent concepts

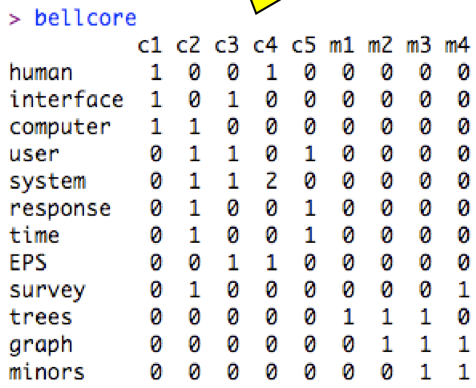
(typically 300 ~ 500)

SVD

$B\,B^T = U\,D^2\,U^T$

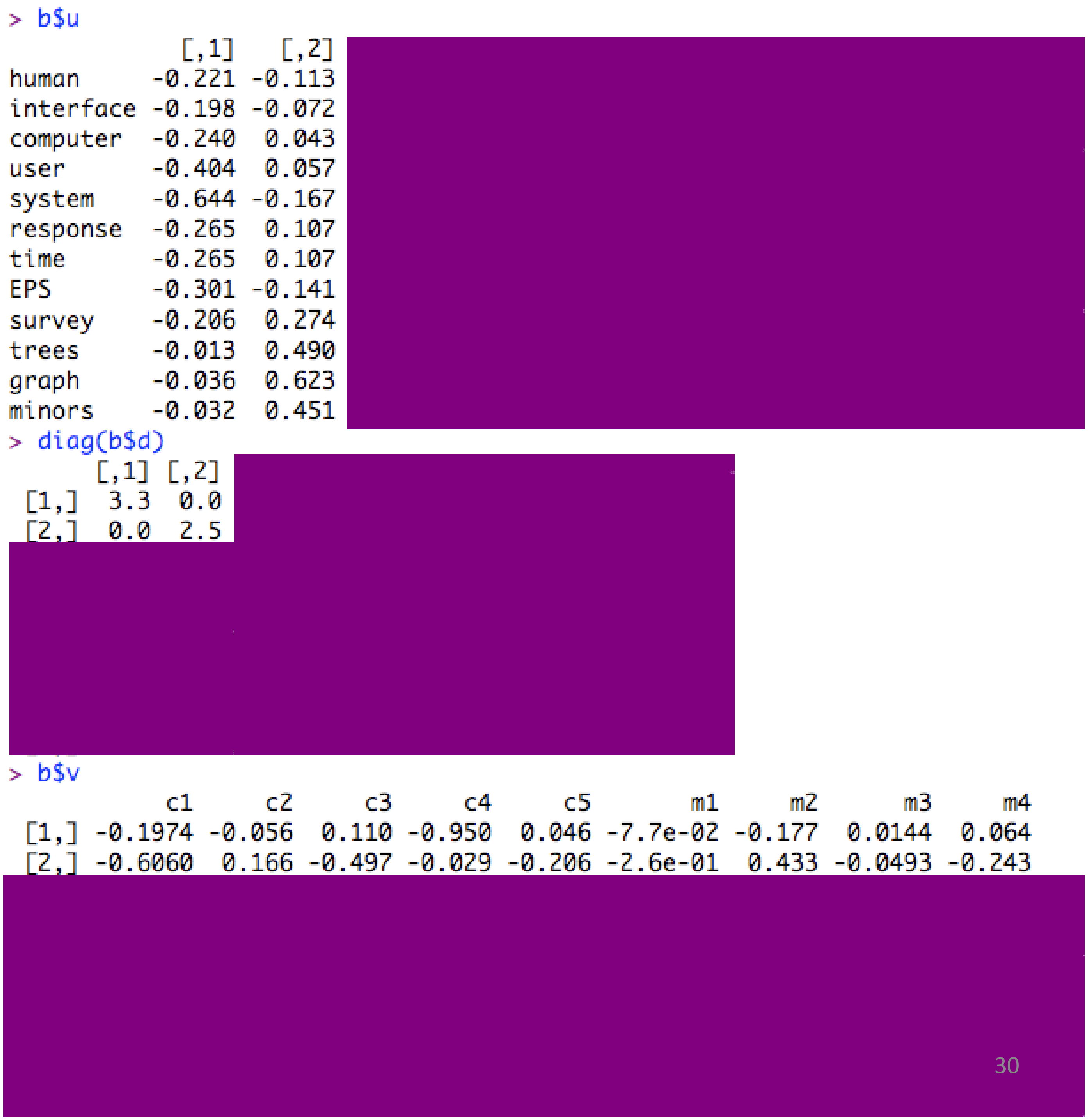
$B^T\,B = V\,D^2\,V^T$

Doc



```
> bellcore
          c1 c2 c3 c4 c5 m1 m2 m3 m4
human      1  0  0  1  0  0  0  0  0
interface  1  0  1  0  0  0  0  0  0
computer   1  1  0  0  0  0  0  0  0
user       0  1  1  0  1  0  0  0  0
system     0  1  1  2  0  0  0  0  0
response   0  1  0  0  1  0  0  0  0
time       0  1  0  0  1  0  0  0  0
EPS        0  0  1  1  0  0  0  0  0
survey     0  1  0  0  0  0  0  0  1
trees      0  0  0  0  0  1  1  1  0
graph      0  0  0  0  0  0  1  1  1
minors     0  0  0  0  0  0  0  1  1
```

Term



```
> b$u
            [,1]   [,2]
human     -0.221 -0.113
interface -0.198 -0.072
computer  -0.240  0.043
user      -0.404  0.057
system    -0.644 -0.167
response  -0.265  0.107
time      -0.265  0.107
EPS       -0.301 -0.141
survey    -0.206  0.274
trees     -0.013  0.490
graph     -0.036  0.623
minors    -0.032  0.451
```

[redacted]

```
> diag(b$d)
     [,1] [,2]
[1,]  3.3  0.0
[2,]  0.0  2.5
```

[redacted]

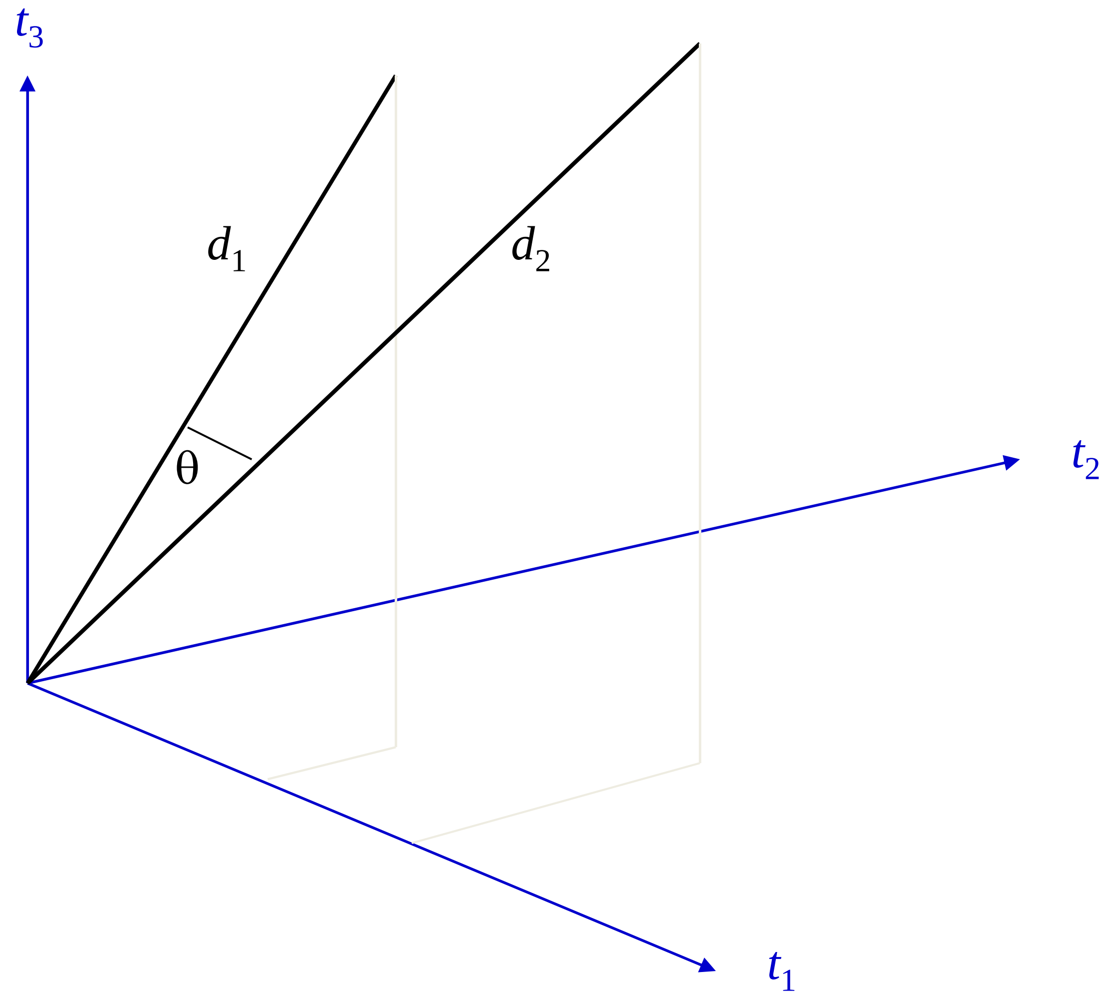
```
> b$v
          c1     c2     c3     c4     c5       m1     m2      m3     m4
[1,] -0.1974 -0.056  0.110 -0.950  0.046 -7.7e-02 -0.177  0.0144  0.064
[2,] -0.6060  0.166 -0.497 -0.029 -0.206 -2.6e-01  0.433 -0.0493 -0.243
```

[redacted]

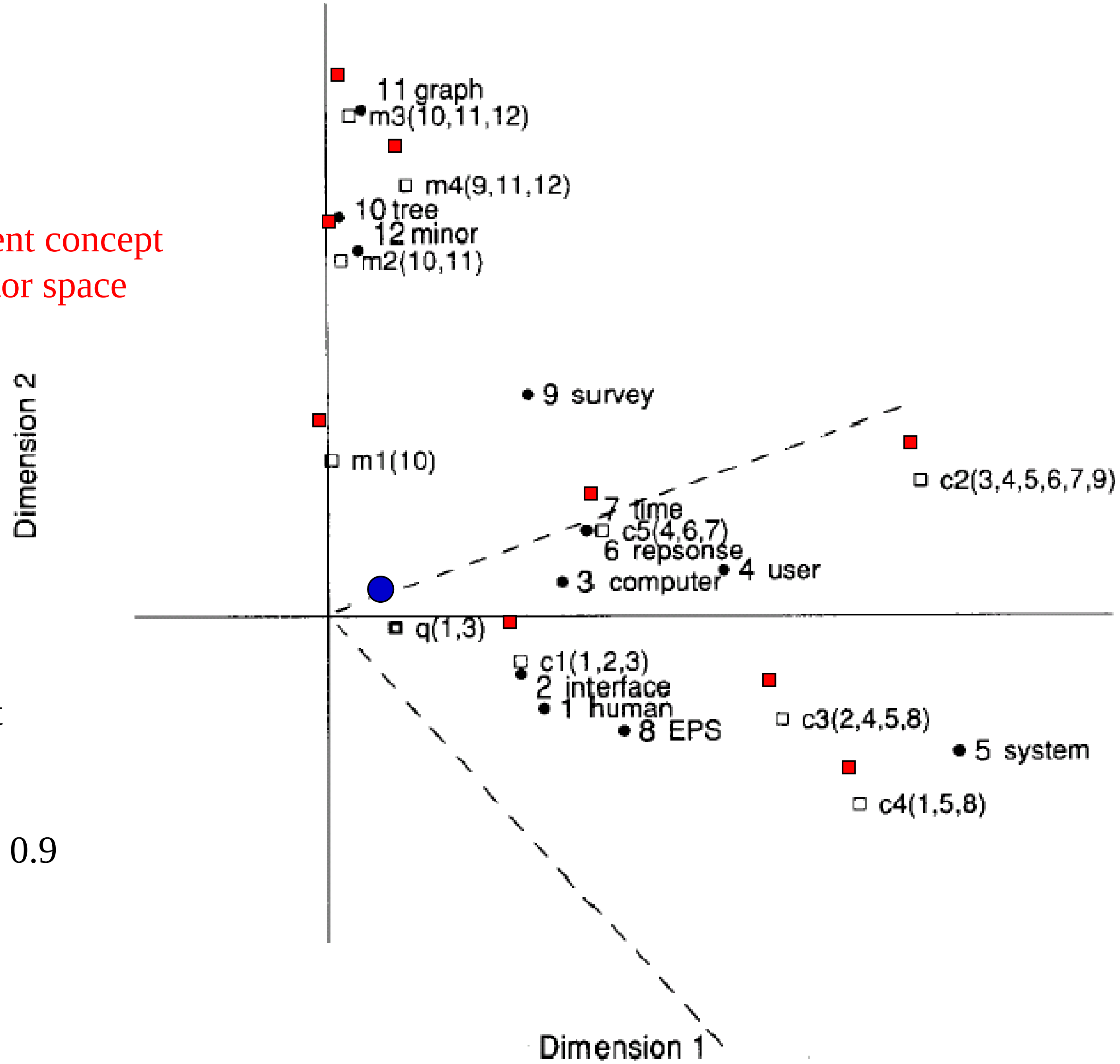
Latent

The term vector space

The space has as many dimensions as there are terms in the word list.



Latent concept
vector space



- term

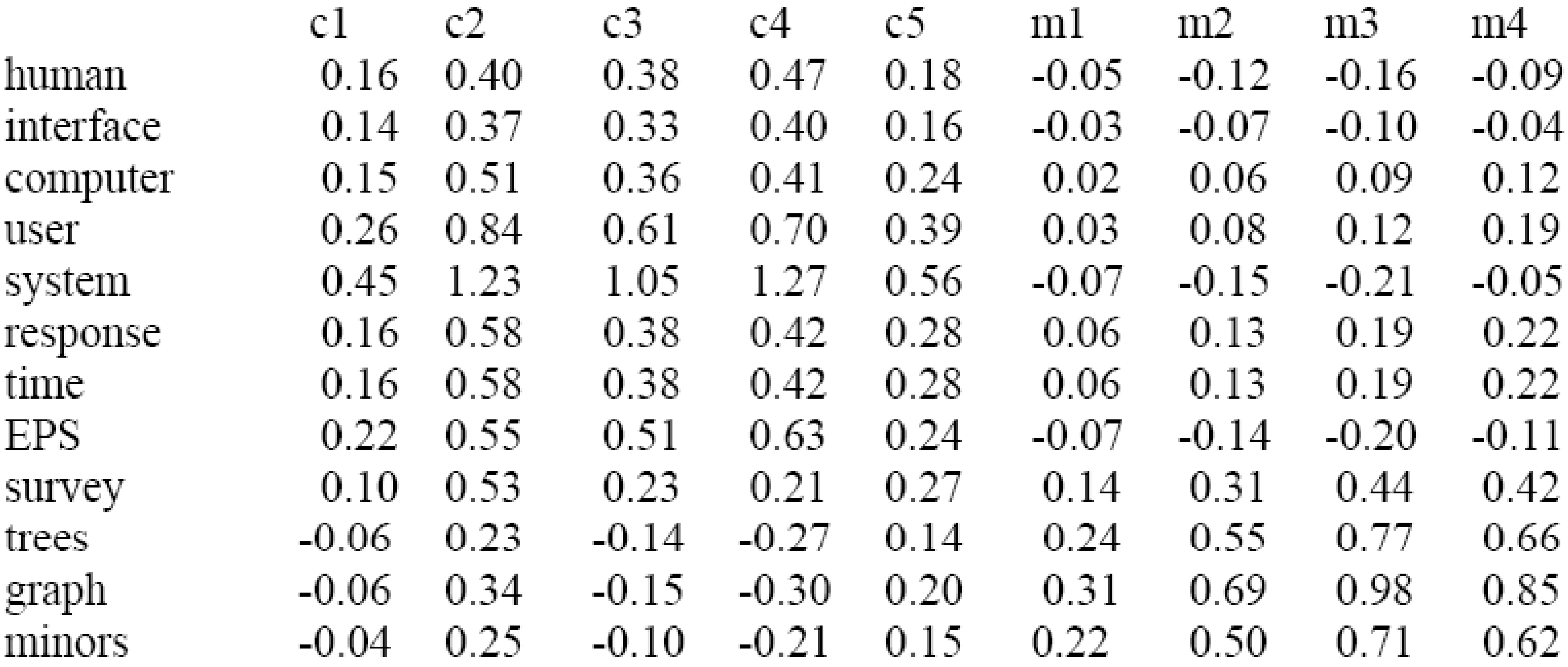
document

🟥

query

● --- cosine > 0.9

Recombination after Dimensionality Reduction



	c1	c2	c3	c4	c5	m1	m2	m3	m4
human	0.16	0.40	0.38	0.47	0.18	-0.05	-0.12	-0.16	-0.09
interface	0.14	0.37	0.33	0.40	0.16	-0.03	-0.07	-0.10	-0.04
computer	0.15	0.51	0.36	0.41	0.24	0.02	0.06	0.09	0.12
user	0.26	0.84	0.61	0.70	0.39	0.03	0.08	0.12	0.19
system	0.45	1.23	1.05	1.27	0.56	-0.07	-0.15	-0.21	-0.05
response	0.16	0.58	0.38	0.42	0.28	0.06	0.13	0.19	0.22
time	0.16	0.58	0.38	0.42	0.28	0.06	0.13	0.19	0.22
EPS	0.22	0.55	0.51	0.63	0.24	-0.07	-0.14	-0.20	-0.11
survey	0.10	0.53	0.23	0.21	0.27	0.14	0.31	0.44	0.42
trees	-0.06	0.23	-0.14	-0.27	0.14	0.24	0.55	0.77	0.66
graph	-0.06	0.34	-0.15	-0.30	0.20	0.31	0.69	0.98	0.85
minors	-0.04	0.25	-0.10	-0.21	0.15	0.22	0.50	0.71	0.62

Document Cosines

(before dimensionality reduction)

$$sim(x, y) = cos(x, y) = \frac{\sum_{i=1}^{V} x_i \, y_i}{|x| \, |y|}$$

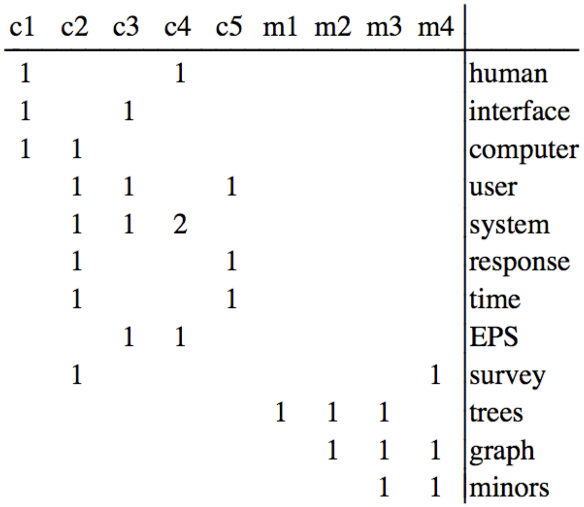
$$cos(c1, c2) = \frac{1}{\sqrt{3} \times \sqrt{6}} = 0.2$$

c1	c2	c3	c4	c5	m1	m2	m3	m4	
1			1						human
1		1							interface
1	1								computer
	1	1		1					user
	1	1	2						system
	1			1					response
	1			1					time
		1	1						EPS
	1							1	survey
					1	1	1		trees
						1	1	1	graph
							1	1	minors

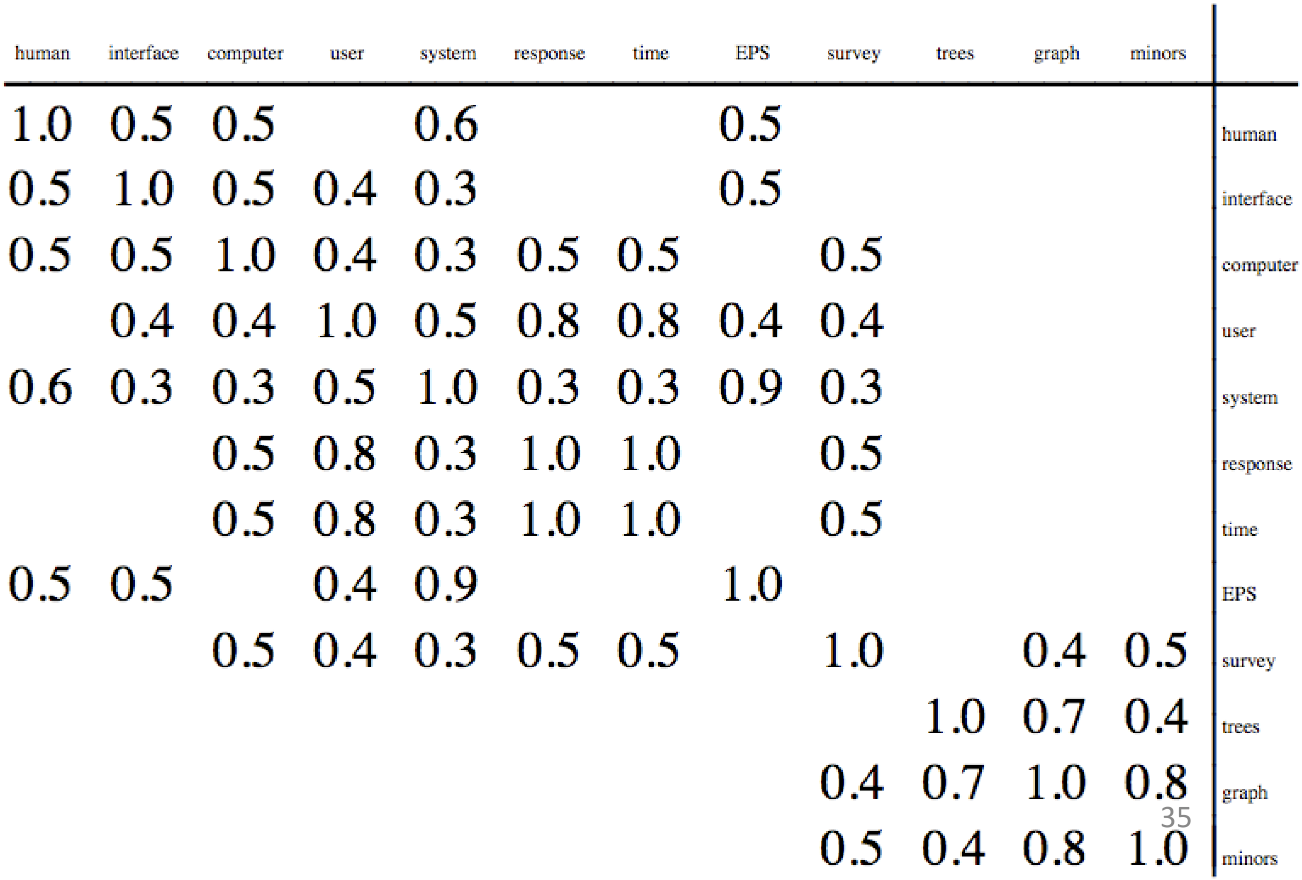
	c1	c2	c3	c4	c5	m1	m2	m3	m4
c1	1.0	0.2	0.3	0.2					
c2	0.2	1.0	0.4	0.3	0.7				0.2
c3	0.3	0.4	1.0	0.6	0.3				
c4	0.2	0.3	0.6	1.0					
c5		0.7	0.3		1.0				
m1						1.0	0.7	0.6	
m2						0.7	1.0	0.8	0.4
m3						0.6	0.8	1.0	0.7
m4		0.2					0.4	0.7	1.0

Term Cosines

(before dimensionality reduction)

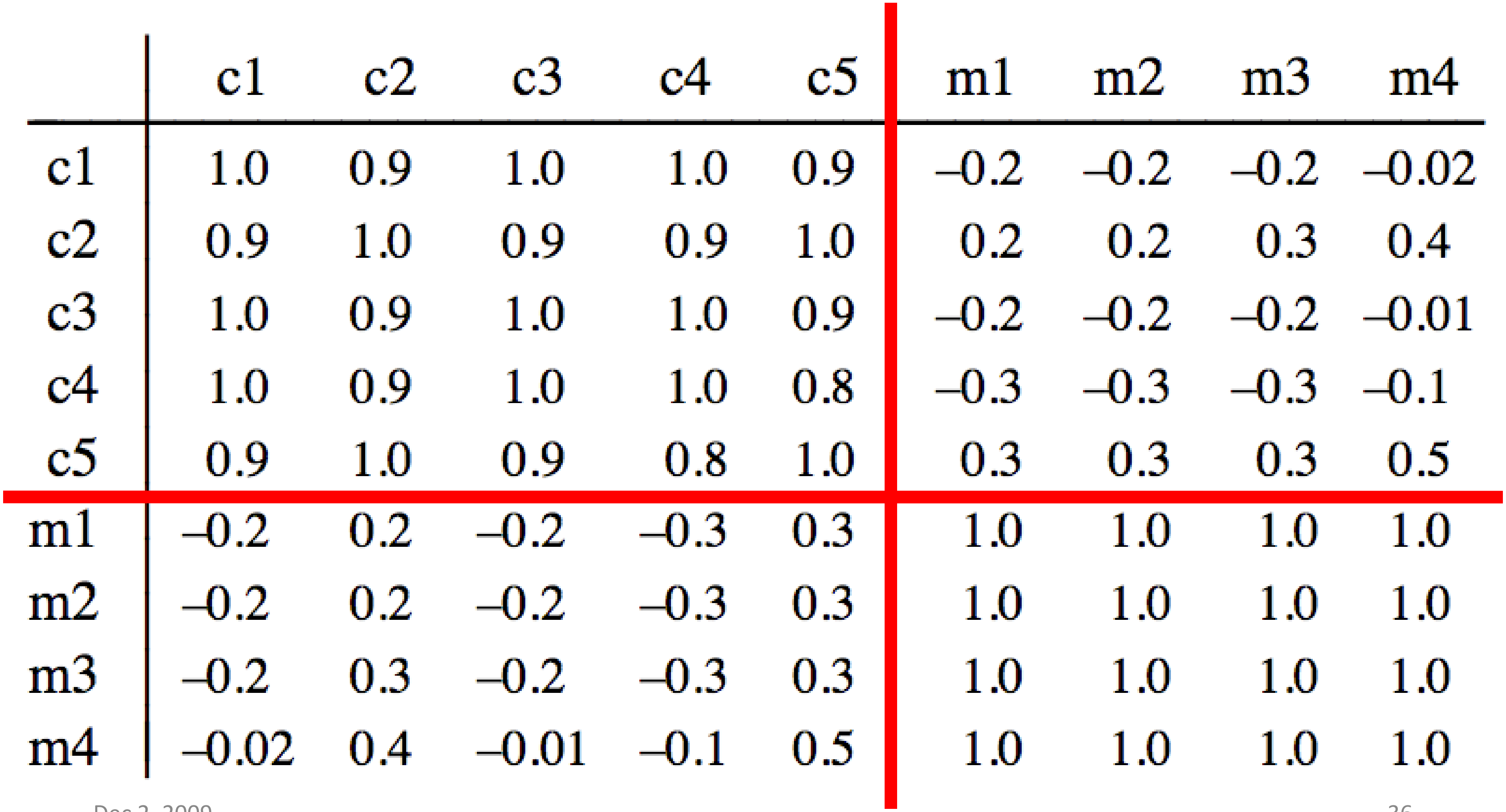


c1	c2	c3	c4	c5	m1	m2	m3	m4	
1			1						human
1		1							interface
1	1								computer
	1	1		1					user
	1	1	2						system
	1			1					response
	1			1					time
		1	1						EPS
	1							1	survey
					1	1	1		trees
						1	1	1	graph
							1	1	minors



human	interface	computer	user	system	response	time	EPS	survey	trees	graph	minors	
1.0	0.5	0.5		0.6			0.5					human
0.5	1.0	0.5	0.4	0.3			0.5					interface
0.5	0.5	1.0	0.4	0.3	0.5	0.5		0.5				computer
	0.4	0.4	1.0	0.5	0.8	0.8	0.4	0.4				user
0.6	0.3	0.3	0.5	1.0	0.3	0.3	0.9	0.3				system
		0.5	0.8	0.3	1.0	1.0		0.5				response
		0.5	0.8	0.3	1.0	1.0		0.5				time
0.5	0.5		0.4	0.9			1.0					EPS
		0.5	0.4	0.3	0.5	0.5		1.0		0.4	0.5	survey
									1.0	0.7	0.4	trees
								0.4	0.7	1.0	0.8	graph
								0.5	0.4	0.8	1.0	minors

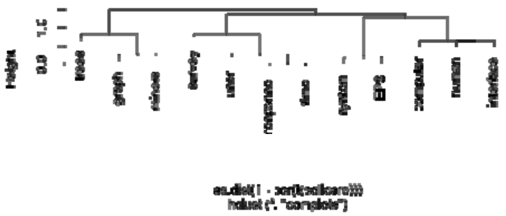
Document Cosines
(after dimensionality reduction)

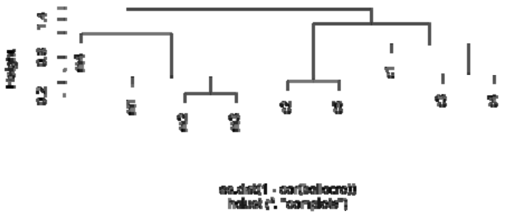


	c1	c2	c3	c4	c5	m1	m2	m3	m4
c1	1.0	0.9	1.0	1.0	0.9	–0.2	–0.2	–0.2	–0.02
c2	0.9	1.0	0.9	0.9	1.0	0.2	0.2	0.3	0.4
c3	1.0	0.9	1.0	1.0	0.9	–0.2	–0.2	–0.2	–0.01
c4	1.0	0.9	1.0	1.0	0.8	–0.3	–0.3	–0.3	–0.1
c5	0.9	1.0	0.9	0.8	1.0	0.3	0.3	0.3	0.5
m1	–0.2	0.2	–0.2	–0.3	0.3	1.0	1.0	1.0	1.0
m2	–0.2	0.2	–0.2	–0.3	0.3	1.0	1.0	1.0	1.0
m3	–0.2	0.3	–0.2	–0.3	0.3	1.0	1.0	1.0	1.0
m4	–0.02	0.4	–0.01	–0.1	0.5	1.0	1.0	1.0	1.0

Clustering

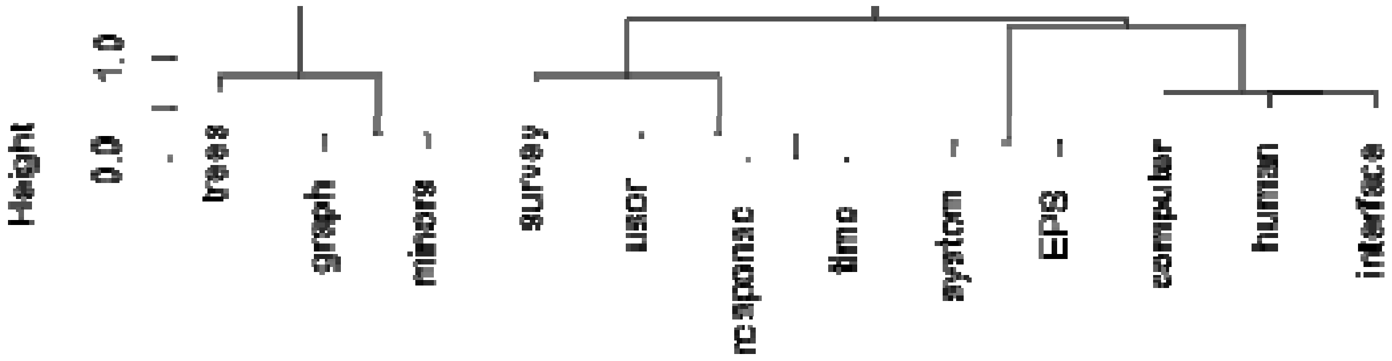
- Useful display for eye-balling a similarity matrix, e.g., cos of terms, cos of docs
- Can also play a role in IR
  - Term clusters can be used to reduce dimensionality (V → number of clusters)
  - Doc clusters can be used to reduce search space (score clusters, rather than documents)
- Input: a similiarity matrix (square, symmetric)
- Output: a tree of clusters
- Methods: single linkage, complete linkage, k-means, and many more



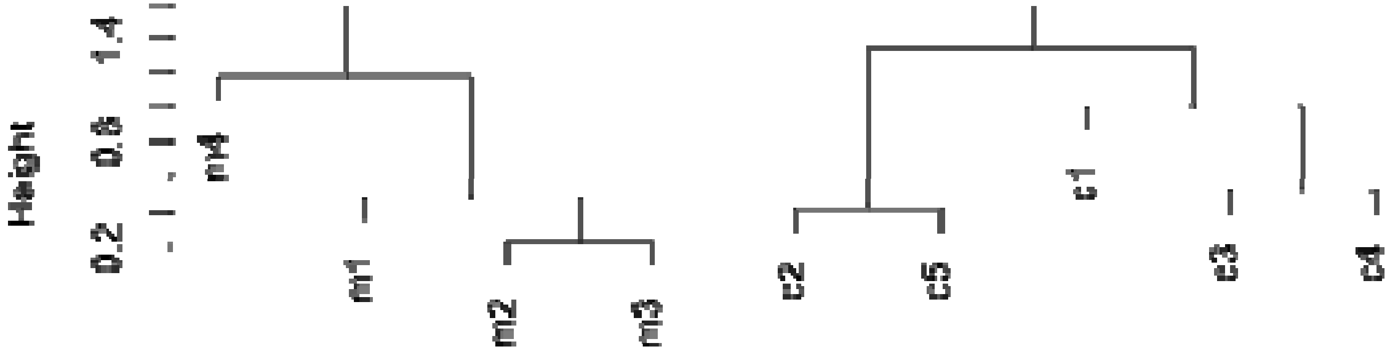


Clustering

(before dimensionality reduction)



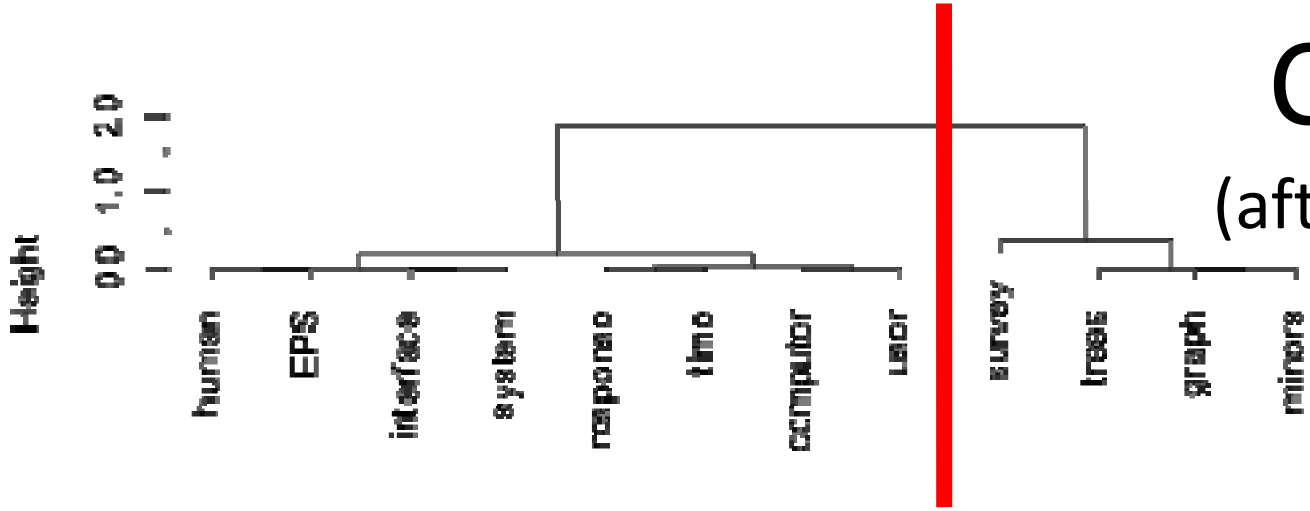
as.dist(1 - cor(t(bellcore)))
hclust (*, "complete")



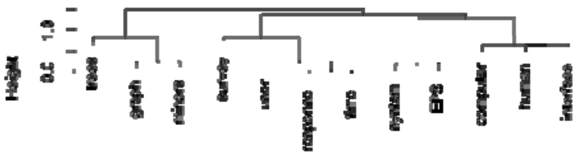
as.dist(1 - cor(bellcore))
hclust (*, "complete")

Clustering

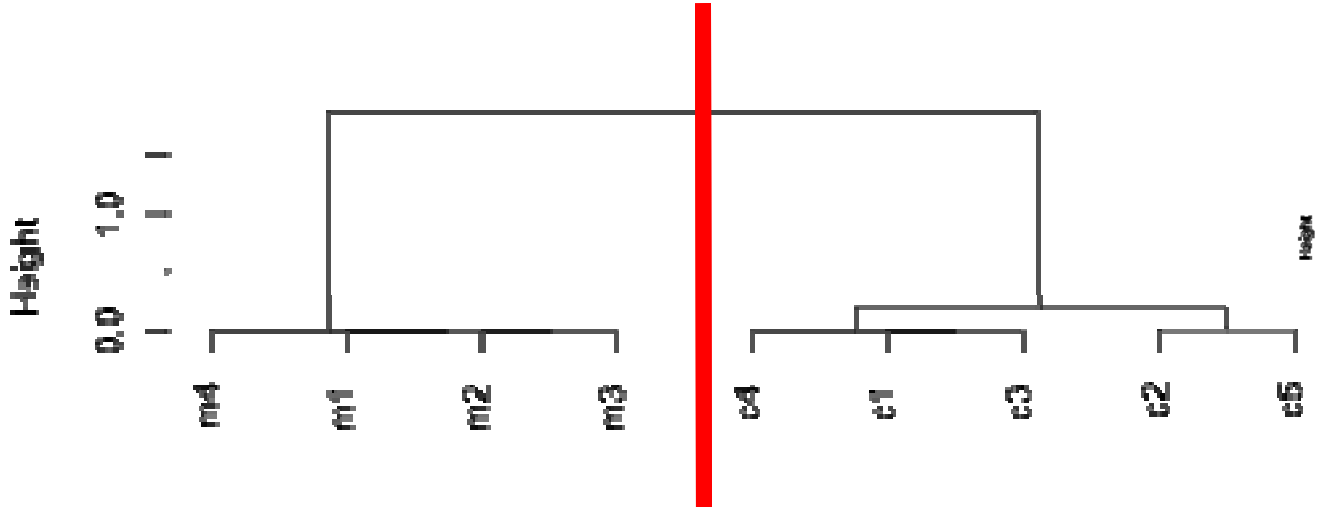
(after dimensionality reduction)



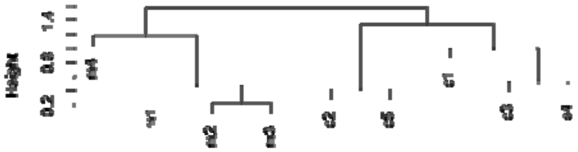
as.dist(1 - cor(t(bellcore2)))
hclust (*, "complete")



as.dist(1 - cor(t(bellcore)))
hclust (*, "complete")



as.dist(1 - cor(bellcore2))
hclust (*, "complete")



as.dist(1 - cor(bellcore))
hclust (*, "complete")

Stop Lists & Term Weighting

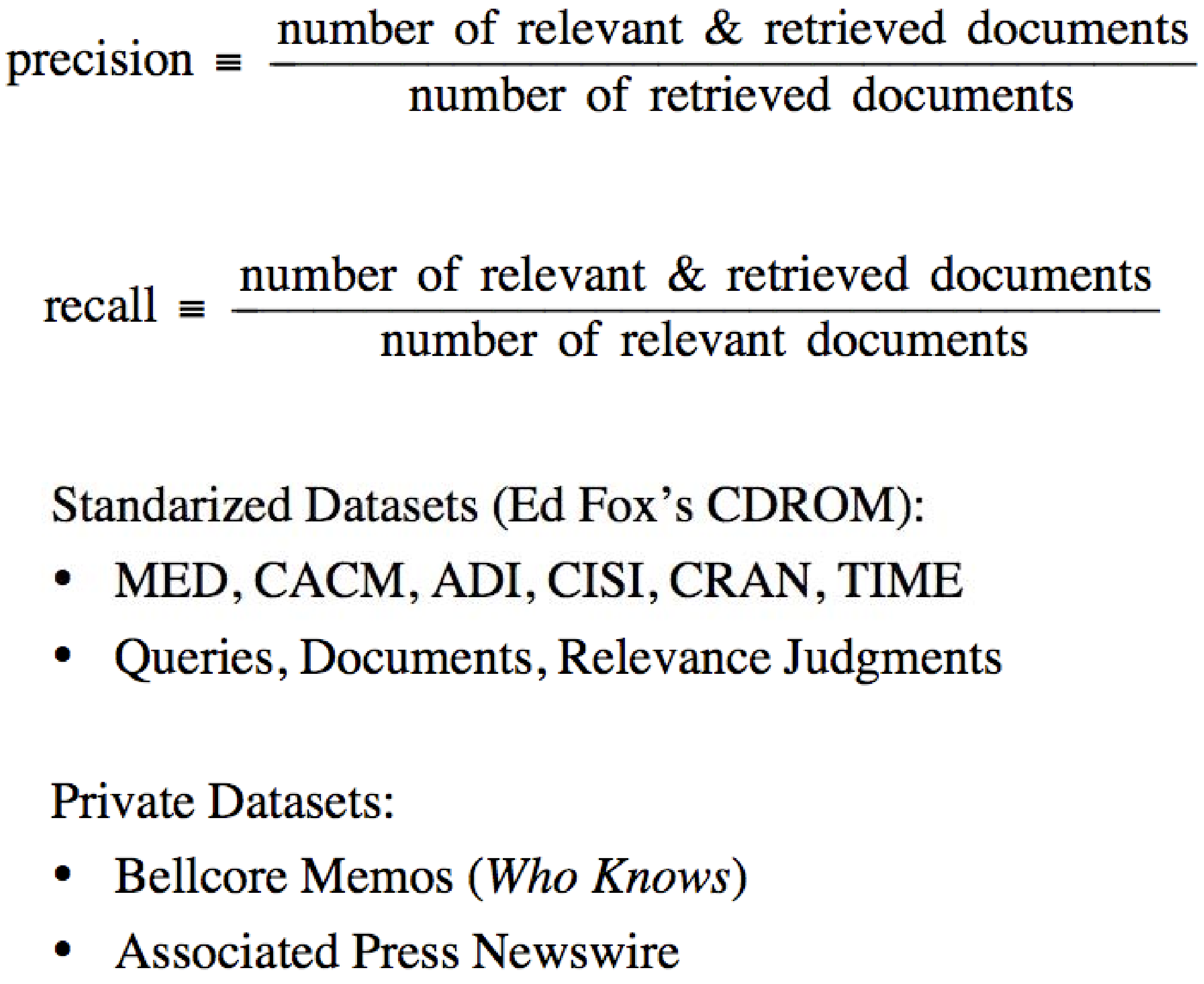
- Emphasize content words and de-emphasize function words

$$sim(x, y) = \frac{\sum_{t=1}^{V} (w_t\ x_t)\ (w_t\ y_t)}{|w\ x|\ |w\ y|}$$

- IDF (inverse document freq)

$$w_t = -\log_2 \frac{\text{number of documents with term } t}{\text{number of documents } (= N)}$$

Evaluation



$$\text{precision} \equiv \frac{\text{number of relevant \& retrieved documents}}{\text{number of retrieved documents}}$$

$$\text{recall} \equiv \frac{\text{number of relevant \& retrieved documents}}{\text{number of relevant documents}}$$

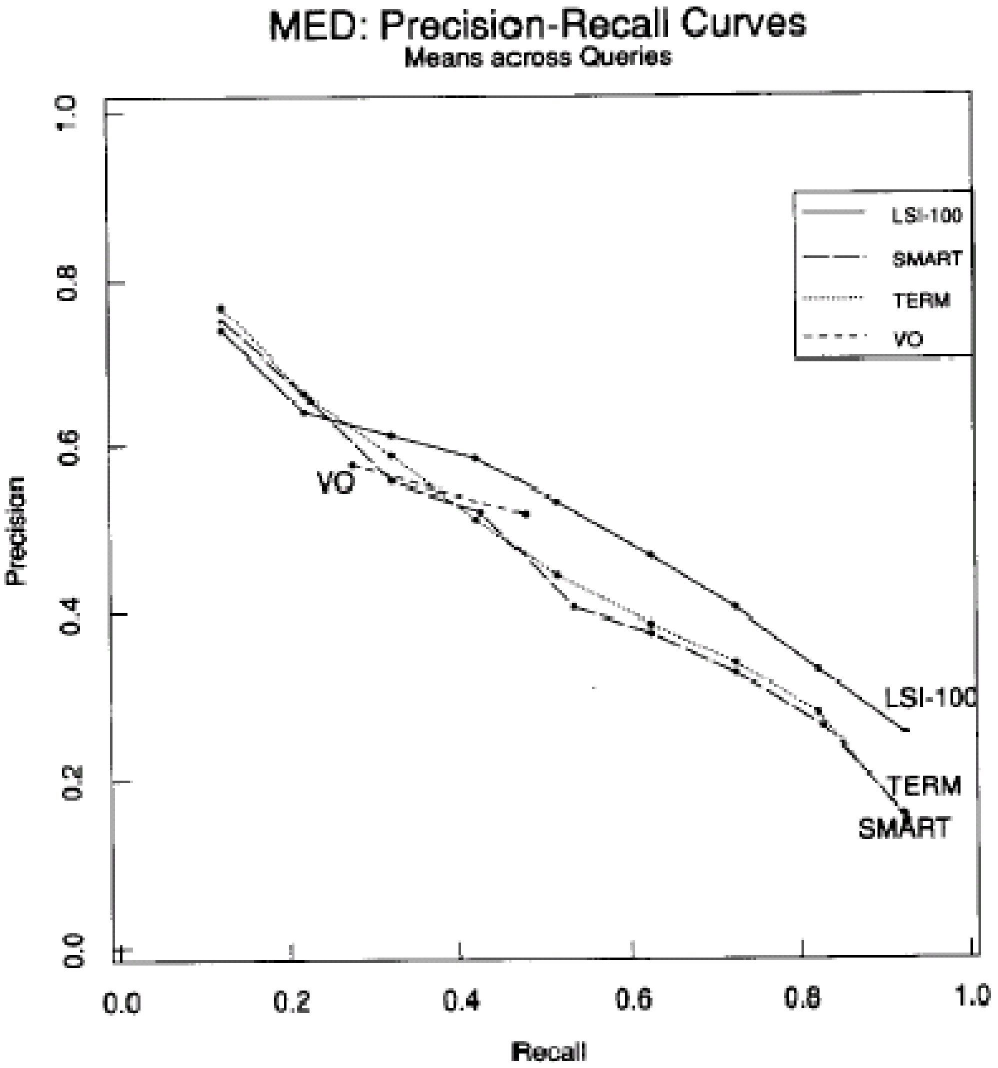
Standarized Datasets (Ed Fox's CDROM):

- MED, CACM, ADI, CISI, CRAN, TIME
- Queries, Documents, Relevance Judgments

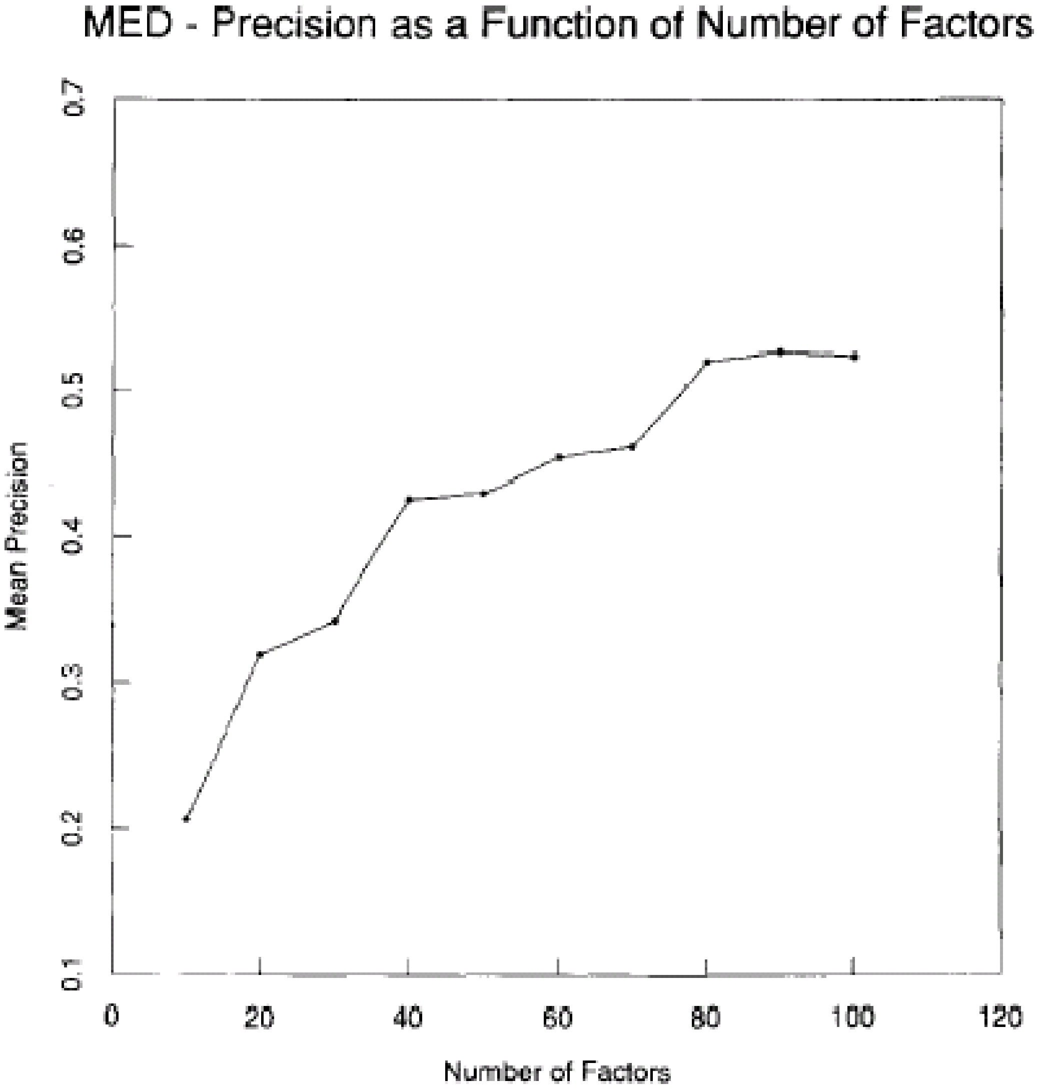
Private Datasets:

- Bellcore Memos (*Who Knows*)
- Associated Press Newswire

Experimental Results: 100 Factors



Experimental Results: Number of Factors



Summary

- IR Problem: sort docs by $sim(d,q)$

- Vector-space ''Model'' (cosine similarity)
- Probabilistic Retrieval Model

- Clustering

- Correlations ''Fixes''
  - Merge morphologically related words
  - Merge synonymous words using a thesaurus
  - Singular Value Decomposition (SVD)

- Evaulation

- Term Weighting: IDF & Entropy

Entropy of Search Logs

- How Big is the Web?
- How Hard is Search?
- With Personalization? With Backoff?

Qiaozhu Mei[†], Kenneth Church[‡]

† University of Illinois at Urbana-Champaign

‡ Microsoft Research

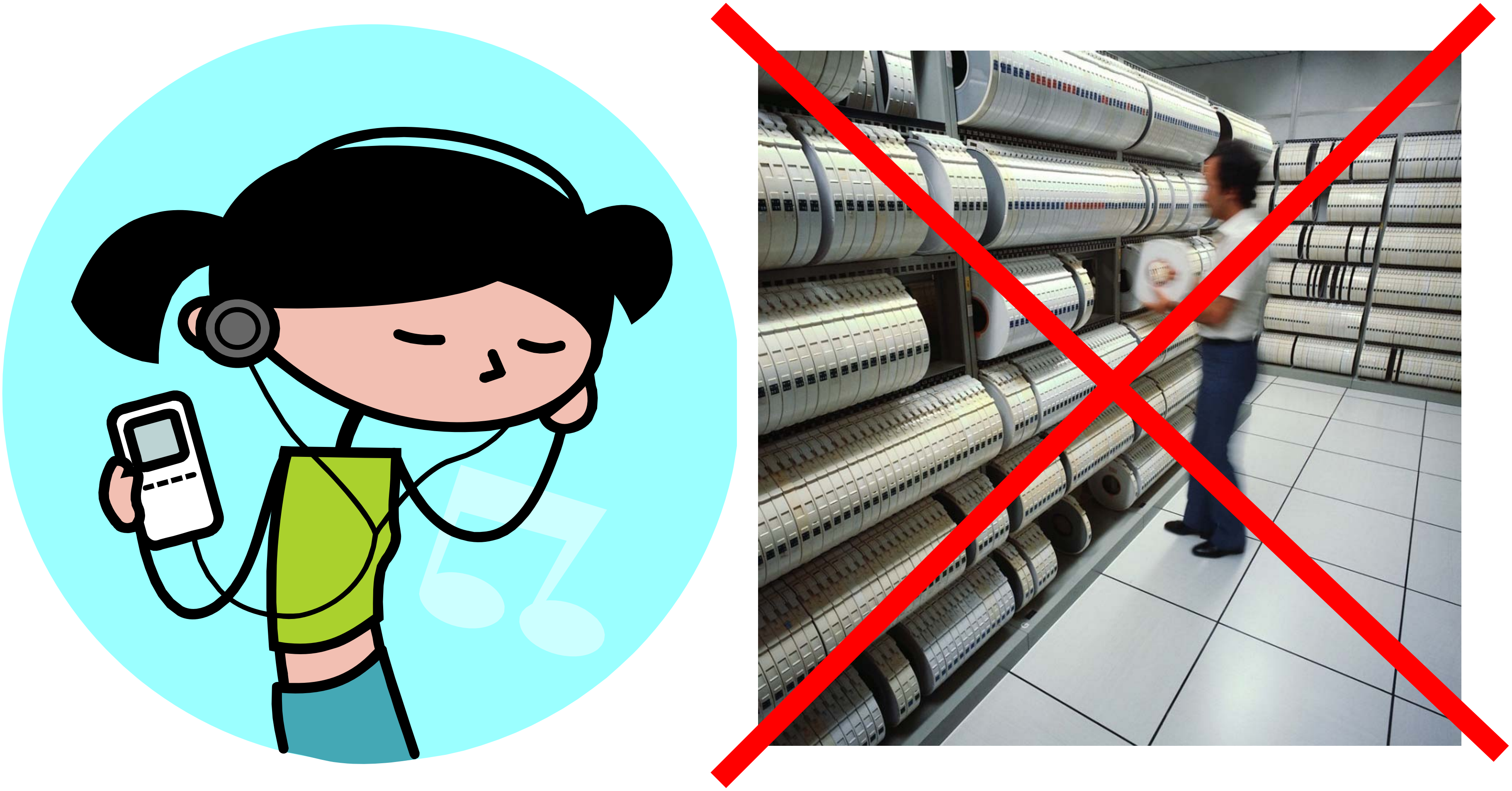
How ~~***Big***~~ is the Web?

Small

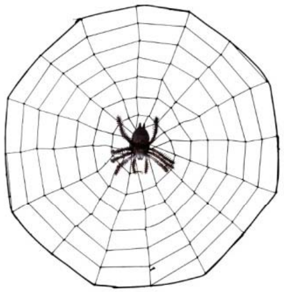
5B? 20B? More? Less?

- What if a small cache of millions of pages
  - Could capture much of the value of billions?
- Could a ***Big*** bet on a cluster in the clouds
  - Turn into a big liability?
- Examples of Big Bets
  - Computer Centers & Clusters
    - Capital (Hardware)
    - Expense (Power)
    - Dev (Mapreduce, GFS, Big Table, etc.)
  - Sales & Marketing >> Production & Distribution

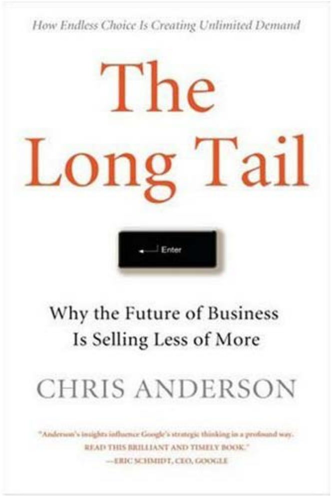
Millions (Not Billions)



Population Bound



- With all the talk about the Long Tail
  - You'd think that the Web was astronomical
  - Carl Sagan: Billions and Billions…
- Lower Distribution $$ → Sell Less of More
- But there are limits to this process
  - NetFlix: 55k movies (not even millions)
  - Amazon: 8M products
  - Vanity Searches: Infinite???
    - Personal Home Pages << Phone Book < Population
    - Business Home Pages << Yellow Pages < Population
- Millions, not Billions (until market saturates)



How Endless Choice Is Creating Unlimited Demand

The Long Tail

Enter

Why the Future of Business Is Selling Less of More

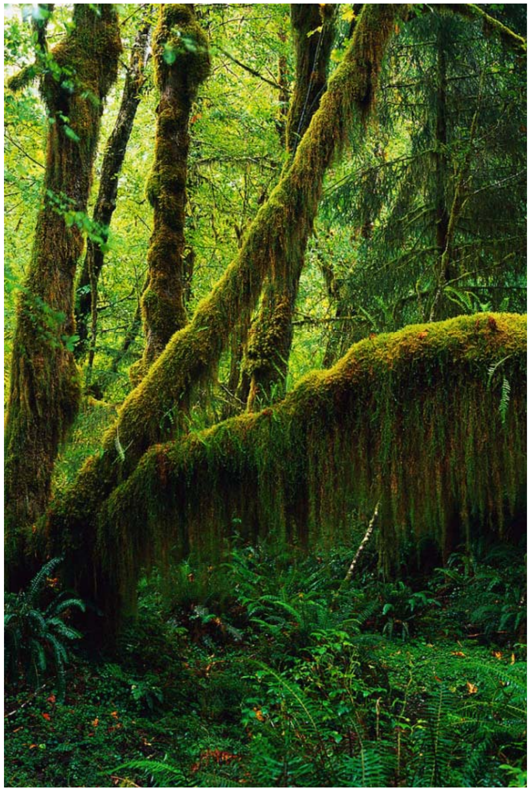
CHRIS ANDERSON

"Anderson's insights influence Google's strategic thinking in a profound way. READ THIS BRILLIANT AND TIMELY BOOK."
—ERIC SCHMIDT, CEO, GOOGLE

It Will Take Decades to Reach Population Bound

- Most people (and products)
  - don’t have a web page (yet)
- Currently, I can find famous people
    - (and academics)
    - but not my neighbors
  - There aren’t that many famous people
    - (and academics)…
  - Millions, not billions
    - (for the foreseeable future)

Equilibrium: Supply = Demand



- If there is a page on the web,
  - And no one sees it,
  - Did it make a sound?
- How big is the web?
  - Should we count “silent” pages
  - That don’t make a sound?
- How many products are there?
  - Do we count “silent” flops
  - That no one buys?

Demand Side Accounting

- Consumers have limited time
  - Telephone Usage: 1 hour per line per day
  - TV: 4 hours per day
  - Web: ??? hours per day
- Suppliers will post as many pages as consumers can consume (and no more)
- Size of Web: O(Consumers)

How Big is the Web?

- Related questions come up in language
- How big is English?
  - – Dictionary Marketing
  - – Education (Testing of Vocabulary Size)
  - – Psychology
  - – Statistics
  - – Linguistics
- Two Very Different Answers
  - – Chomsky: language is infinite
  - – Shannon: 1.25 bits per character

How many words do people know?

What is a word? Person? Know?

Chomskian Argument: Web is Infinite

- One could write a malicious spider trap
  - http://successor.aspx?x=0 → http://successor.aspx?x=1 → http://successor.aspx?x=2
- Not just academic exercise
- Web is full of benign examples like
  - http://calendar.duke.edu/
  - Infinitely many months
  - Each month has a link to the next

How ***Big*** is the Web?

5B? 20B? More? Less?

Entropy (H)

- More (Chomsky)
  - http://successor?x=0
- Less (Shannon)

MSN Search Log
1 month

Query	21.1
URL	22.1
IP	22.1

More Practical Answer

Comp Ctr ($$$$) → Walk in the Park ($)

Cluster in Cloud → Desktop → Flash

Millions (not Billions)

Dec 2, 2009

Entropy (H)

- $H(X) = -\sum_{x \in X} p(x) \log p(x)$
  - Size of search space; difficulty of a task
- H = 20 → 1 million items distributed uniformly
- Powerful tool for sizing challenges and opportunities
  - How hard is search?
  - How much does personalization help?

How Hard Is Search?
Millions, not Billions

- Traditional Search
  - H(URL | Query)
  - 2.8 (= 23.9 – 21.1)
- Personalized Search
  - H(URL | Query, ***IP***)
  - ***1.2*** (= 27.2 – 26.0)

	Entropy (H)
Query	21.1
URL	22.1
IP	22.1

Personalization cuts H in Half!

Difficulty of Queries

- Easy queries (low H(URL|Q)):
  - google, yahoo, myspace, ebay, …
- Hard queries (high H(URL|Q)):
  - dictionary, yellow pages, movies,
  - “what is may day?”

How Hard are Query Suggestions?

The Wild Thing? C* Rice → Condoleezza Rice

- Traditional Suggestions
  - H(Query)
  - 21 bits
- Personalized
  - H(Query | ***IP***)
  - **5** bits (= 26 – 21)

Personalization cuts H in Half!

Twice

	Entropy (H)
Query	21.1
URL	22.1
IP	22.1
All But IP	23.9
All But URL	26.0
All But Query	27.1
All Three	27.2

Personalization with Backoff

- Ambiguous query: MSG
  - Madison Square Garden
  - Monosodium Glutamate
- Disambiguate based on user's prior clicks
- When we don't have data
  - Backoff to classes of users
- Proof of Concept:
  - Classes defined by IP addresses
- Better:
  - Market Segmentation (Demographics)
  - Collaborative Filtering (Other users who click like me)

Backoff

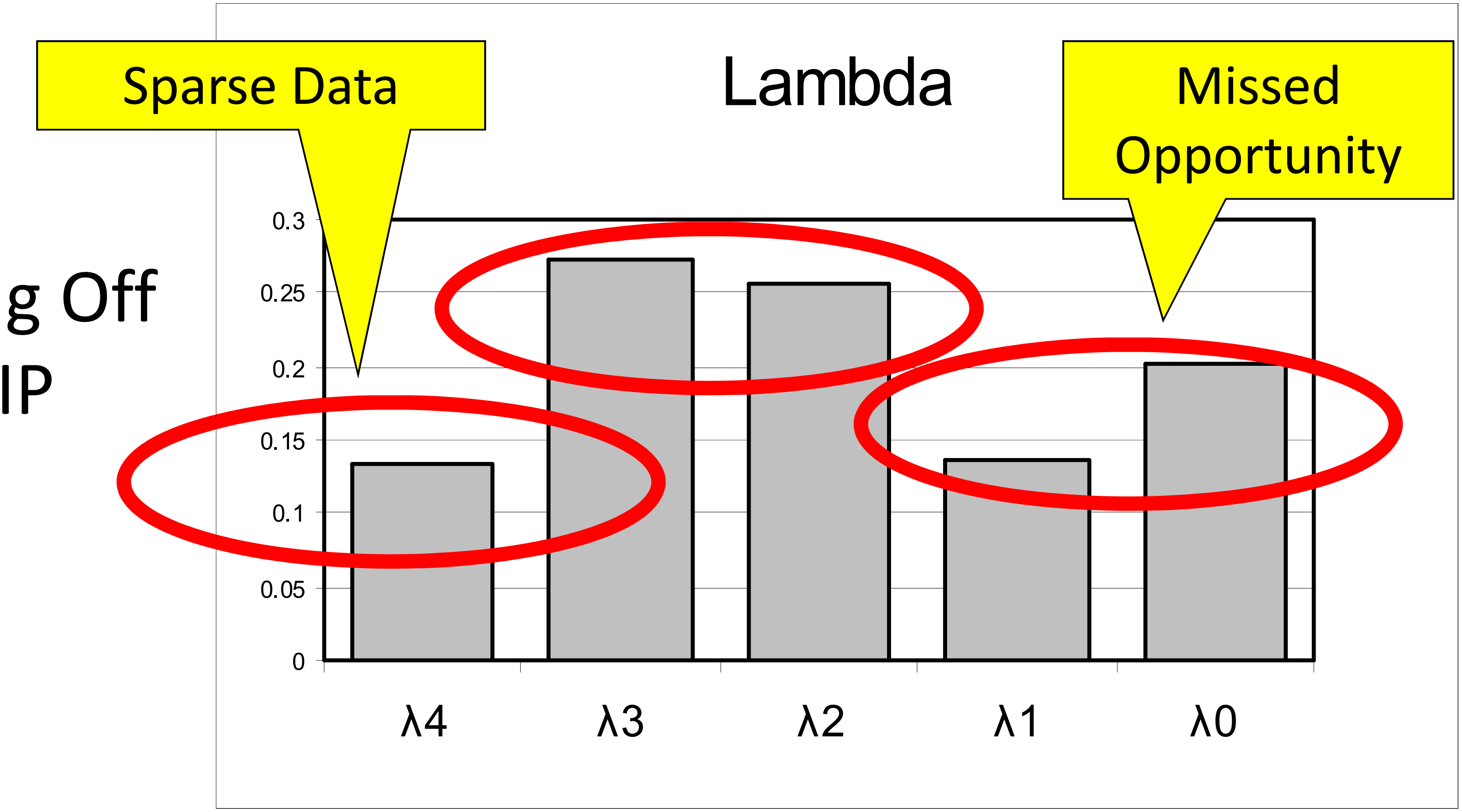
- Proof of concept: bytes of IP define classes of users
- If we only know some of the IP address, does it help?

Bytes of IP addresses	H(URL| IP, Query)
156.111.188.243	1.17
156.111.188.*	1.20
156.111.*.*	1.39
156.*.*.*	1.95
..*.*	2.74

Some of the IP is better than none

Cuts H in half even if using the first two bytes of IP

Backing Off by IP


Lambda

Sparse Data

Missed Opportunity

- Personalization with Backoff
- **λs** estimated with EM and CV
- A little bit of personalization
  - Better than too much
  - Or too little

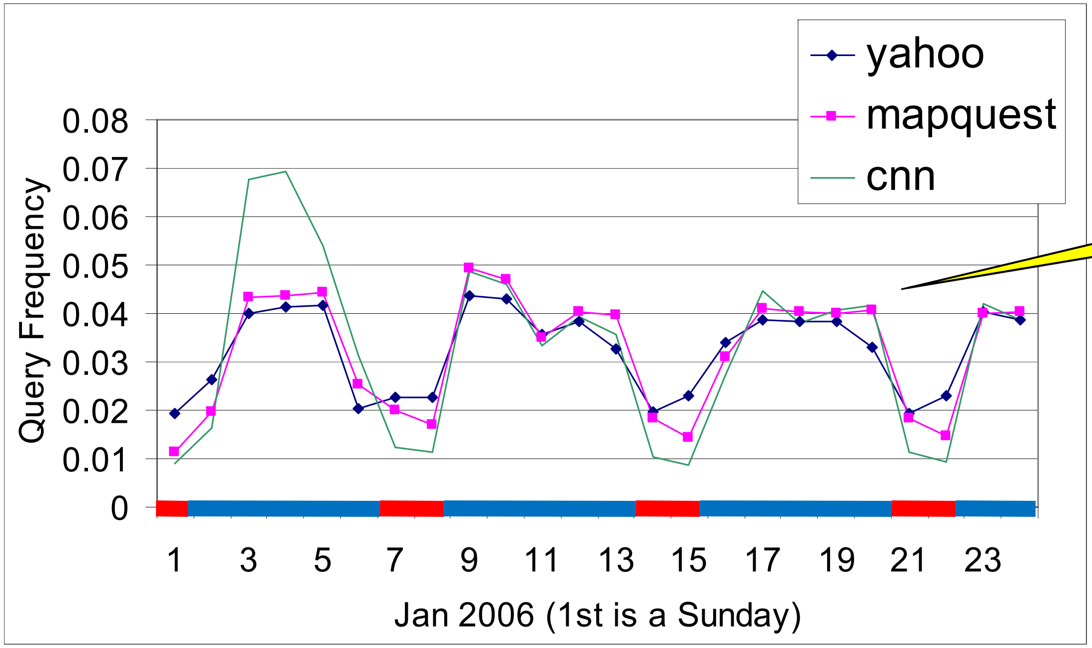
$$P(Url \mid IP, Q) = \sum_{i=0}^{4} \lambda_i P(Url \mid IP_i, Q)$$

λ_4 : weights for first 4 bytes of IP
λ_3 : weights for first 3 bytes of IP
λ_2 : weights for first 2 bytes of IP

……

Personalization with Backoff → Market Segmentation

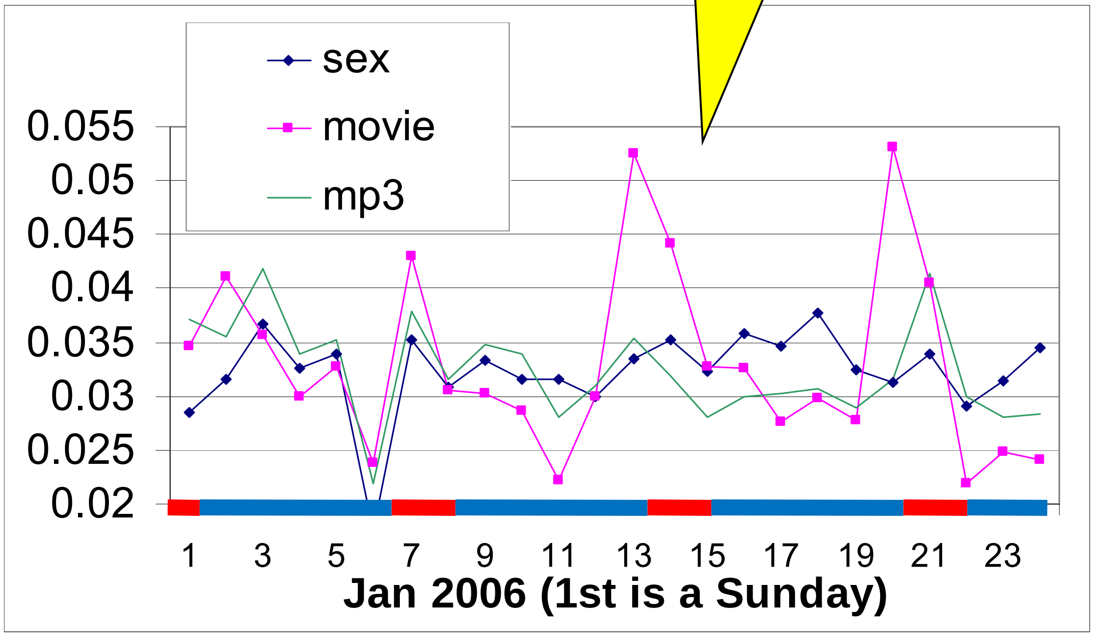
- Traditional Goal of Marketing:
  - Segment Customers (e.g., Business v. Consumer)
  - By Need & Value Proposition
    - Need: Segments ask different questions at different times
    - Value: Different advertising opportunities
- Segmentation Variables
  - Queries, URL Clicks, IP Addresses
  - Geography & Demographics (Age, Gender, Income)
  - Time of day & Day of Week



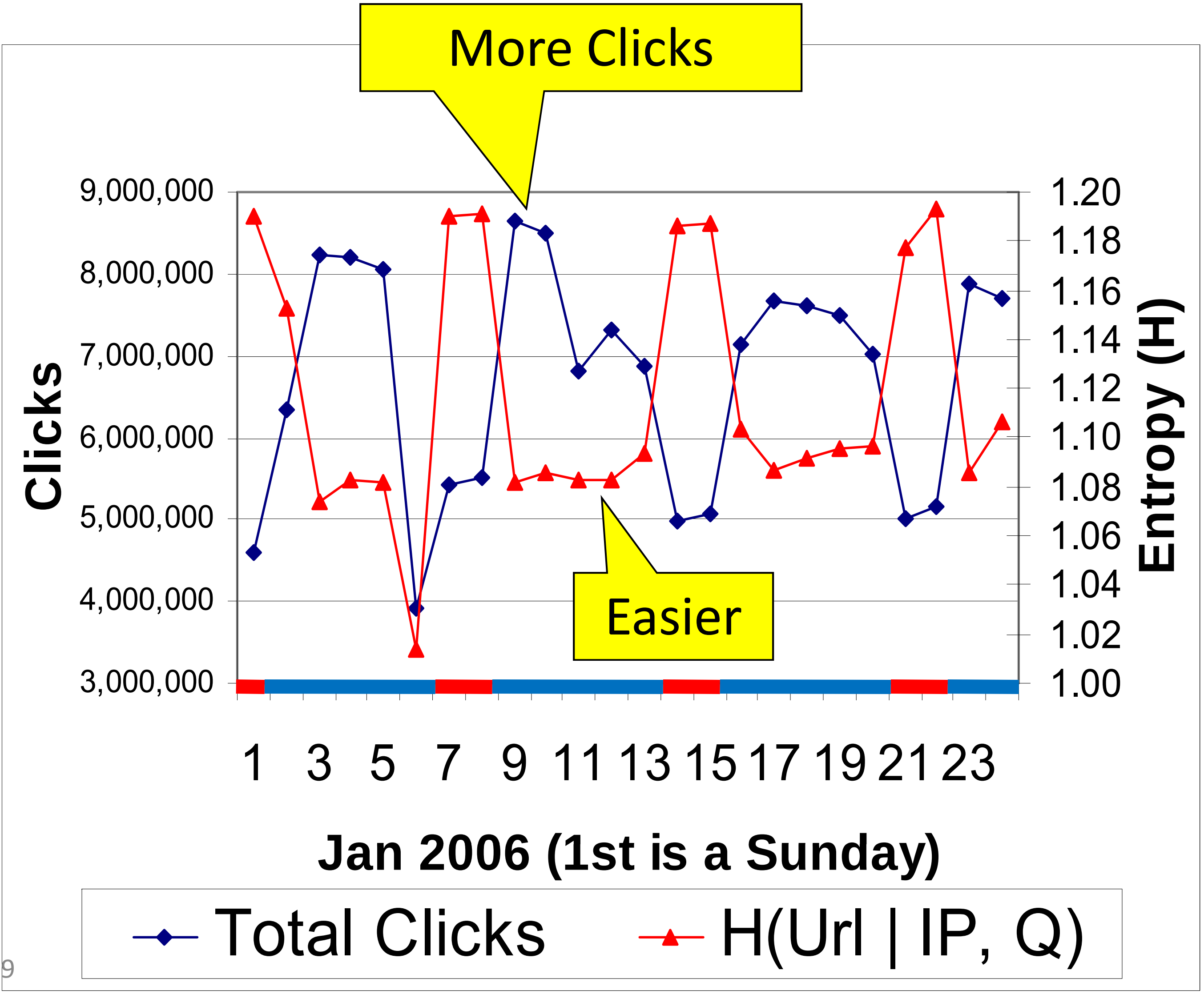
Business Queries on Business Days

Consumer Queries

(Weekends & Every Day)



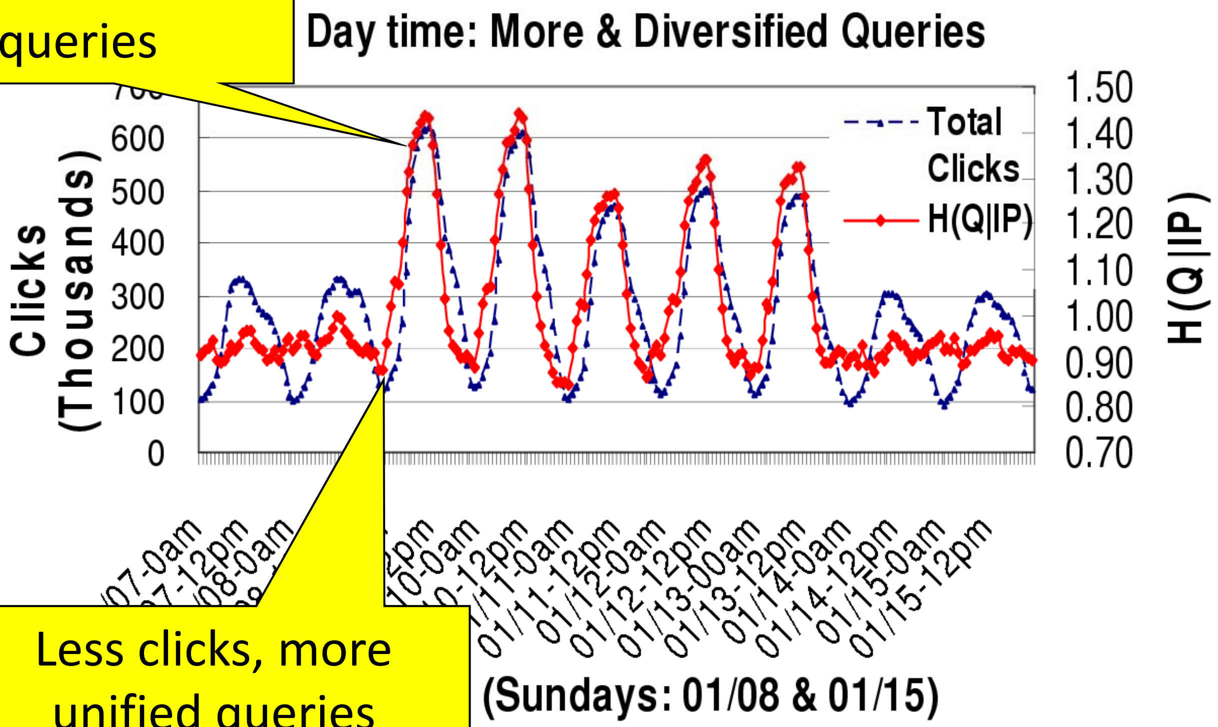
Business Days v. Weekends:
More Clicks and Easier Queries



Day v. Night:

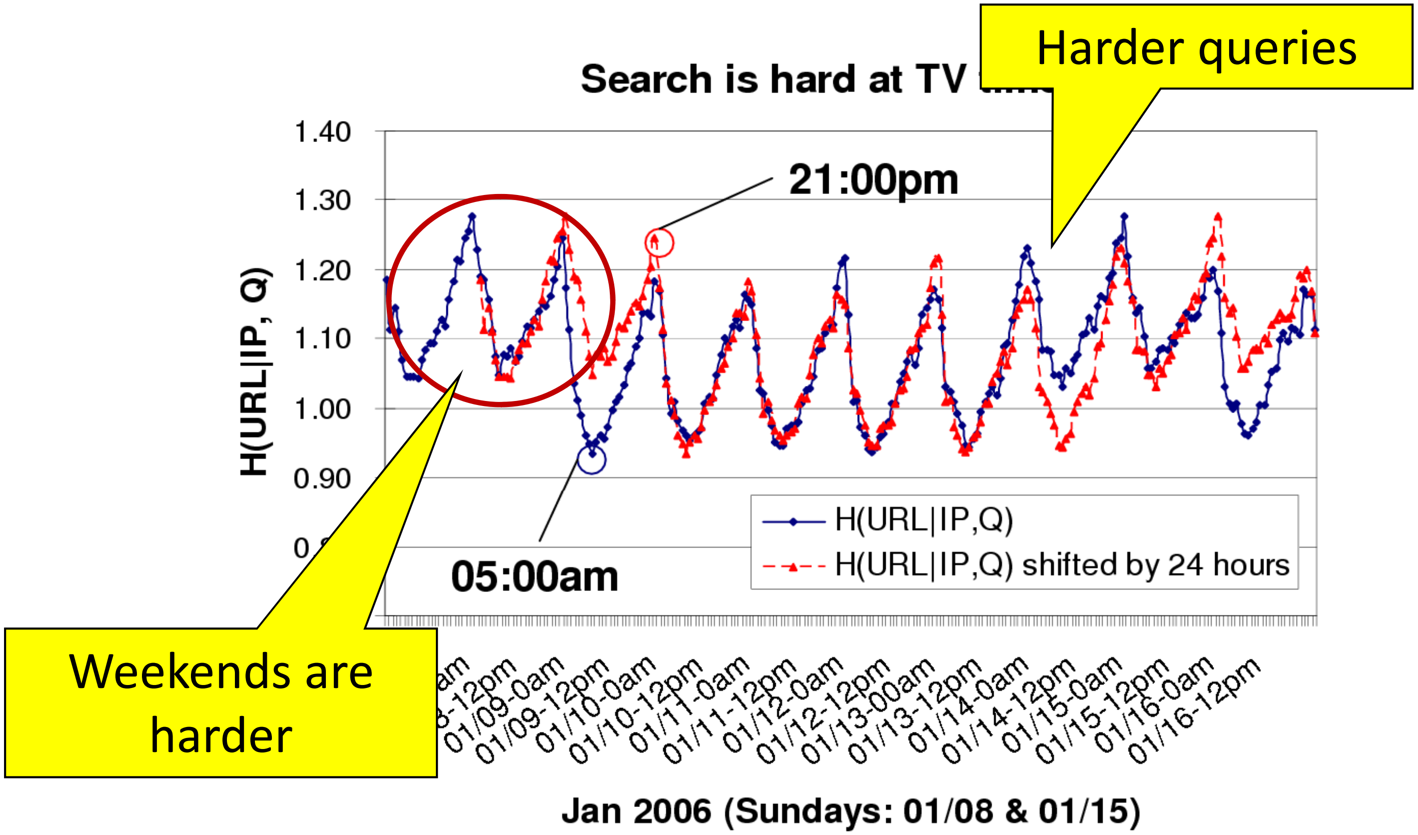
More queries (and easier queries) during business hours

More clicks and diversified queries



Less clicks, more unified queries

Harder Queries during Prime Time TV



Harder queries

Weekends are harder

Conclusions: Millions (not Billions)

- How Big is the Web?
  - Upper bound: O(Population)
    - Not Billions
    - Not Infinite
- Shannon >> Chomsky
  - How hard is search?
  - Query Suggestions?
  - Personalization?
- Cluster in Cloud ($$$$) → Walk-in-the-Park ($)

Entropy is a great hammer

Conclusions:
Personalization with Backoff

- Personalization with Backoff
  - Cuts search space (entropy) in half
  - Backoff → Market Segmentation
    - Example: Business v. Consumer
      - Need: Segments ask different questions at different times
      - Value: Different advertising opportunities
- Demographics:
  - Partition by *ip, day, hour, business/consumer query*…
- Future Work:
  - Model combinations of surrogate variables
  - Group users with similarity → collaborative search

Noisy Channel Model for Web Search

Michael Bendersky

- Input → Noisy Channel → Output
  - Input’ ≈ $\text{ARGMAX}_{\text{Input}}$ Pr(Input) * Pr(Output | Input)

Prior

Channel Model

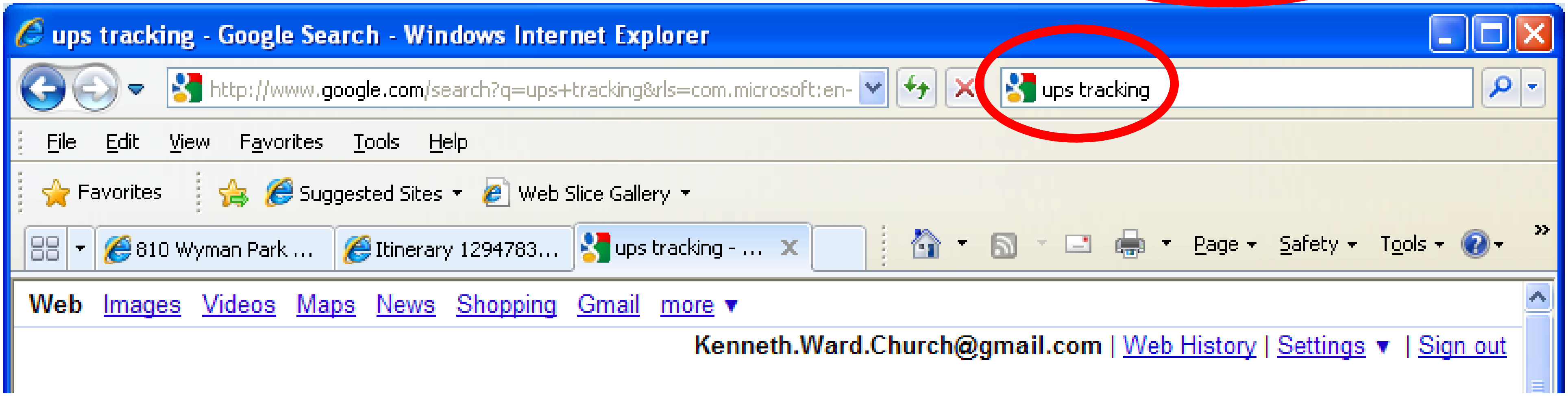
- Speech
  - Words → Acoustics
  - Pr(Words) * Pr(Acoustics | Words)
- Machine Translation
  - English → French
  - Pr(English) * Pr (French | English)
- Web Search
  - Web Pages → Queries
  - Pr(Web Page) * Pr (Query | Web Page)

Prior

Channel Model

Document Priors

- Page Rank *(Brin & Page, 1998)*
  - Incoming link votes
- Browse Rank *(Liu et al., 2008)*
  - Clicks, toolbar hits
- Textual Features *(Kraaij et al., 2002)*
  - Document length, URL length, anchor text
  - Wikipedia



ups tracking - Google Search - Windows Internet Explorer

http://www.google.com/search?q=ups+tracking&rls=com.microsoft:en-

ups tracking

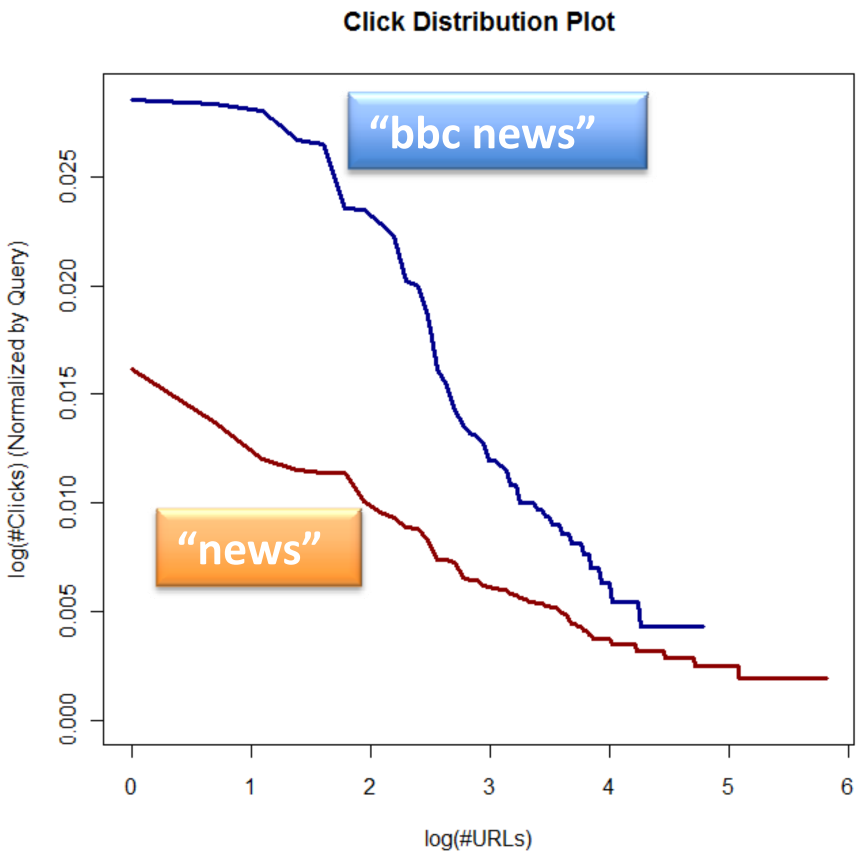
Kenneth.Ward.Church@gmail.com | Web History | Settings ▾ | Sign out

Query Priors: Degree of Difficulty

- **Some queries are easier than others**
  - Human Ratings (HRS): Perfect judgments → *easier*
  - Static Rank (Page Rank): higher → *easier*
  - Textual Overlap: match → *easier*
      - "cnn" → www.cnn.com (match)
  - Popular: lots of clicks → *easier* (toolbar, slogs, glogs)
  - Diversity/Entropy: fewer plausible URLs → *easier*
  - Broder's Taxonomy:
    - Navigational/Transactional/Informational
    - Navigational tend to be easier:
      - "cnn" → www.cnn.com (navigational)
      - "BBC News" (navigational) easier than "news" (informational)

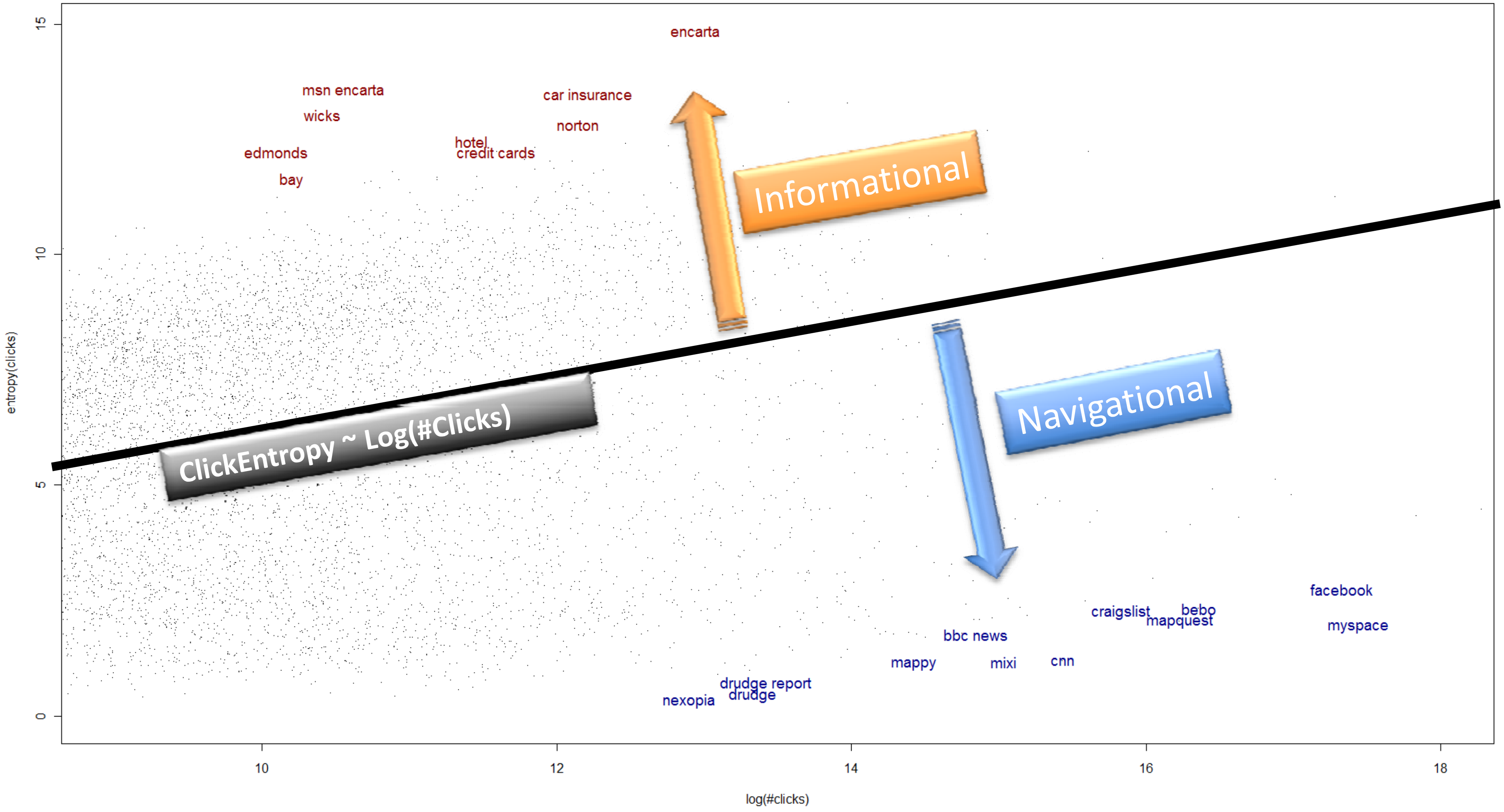
Informational vs. Navigational Queries

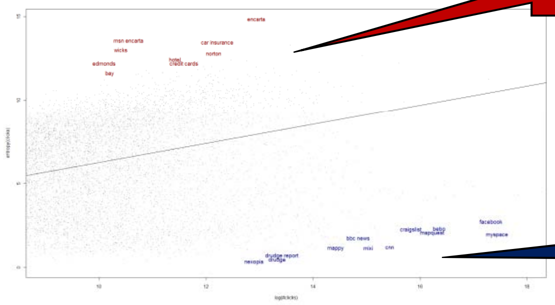
- **Fewer plausible URL's → easier query**
  - Click Entropy
    - Less is easier
  - Broder's Taxonomy:
    - Navigational / Informational
    - Navigational is easier:
      - "BBC News" (navigational) easier than "news"
  - Less opportunity for personalization
    - (Teevan et al., 2008)



Navigational queries have smaller entropy

Informational/Navigational by Residuals





Informational

Navigational

Informational Vs. Navigational Queries

Residuals – Highest Quartile

"bay" *"car insurance "*
"carinsurance" *"credit cards"*
"date" *"day spa"*
“dell computers" *"dell laptops“*
"edmonds" *"encarta"*
"hotel" *"hotels"*
"house insurance" *"ib"*
"insurance" *"kmart"*
"loans" *"msn encarta"*
"musica" *"norton"*
"payday loans" *"pet insurance "*
"proactive" *"sauna"*

Residuals – Lowest Quartile

"accuweather" *"ako"*
"bbc news" *"bebo"*
"cnn" *"craigs list"*
"craigslist" *"drudge“*
“drudge report" *"espn"*
"facebook" *"fox news"*
"foxnews" *"friendster"*
"imdb" *"mappy"*
"mapquest" *"mixi“*
“msnbc" *"my"*
"my space" *"myspace"*
"nexopia" *"pages jaunes"*
"runescape" *"wells fargo"*

Alternative Taxonomy: Click Types

- Classify queries by type
  - Problem: query logs have no "informational/navigational" labels

- Instead, we can use logs to categorize queries
  - **Commercial Intent** → more ad clicks
  - **Malleability** → more query suggestion clicks
  - **Popularity** → more future clicks (anywhere)
    - Predict future clicks (anywhere)
      - Past Clicks: February – May, 2008
      - Future Clicks: June, 2008

0 of about **181,000,000** r **digit cmera**. (**0.28** seconds)

Canon **Digital** Cameras
Sponsored Link
www.BestBuy.com A Canon **Digital Camera** Makes The Perfect Gift. Shop Best Buy® Today!

Did you mean: ***digital camera*** Top 2 results shown

Digital Camera Reviews and News: **Digital** Photography Review
Nov 20, 2009 ... **Digital** Photography Review: All the latest **digital camera** reviews and
digital imaging news. Lively discussion forums.
Reviews - Canon EOS 7D / 50D - Most popular cameras
www.dpreview.com/ - Cached - Similar

Unbiased **Digital Camera** Reviews and News | **Digital Camera** Resource ...
The **Digital Camera** Resource Page has been providing unbiased **digital camera** reviews,
news, discussion forums, buyers guides, and frequently asked questions ...
www.dcresource.com/ - Cached - Similar

Results for: **digit cmera**

Digital cameras: compare **digital camera** reviews - CNET Reviews
Digital camera reviews and ratings, video reviews, user opinions, most popular **digital**
cameras, **camera** buying guides, prices, and comparisons.
Editors - Jet lag - Canon PowerShot SD880 IS (gold)
reviews.cnet.com/**digital**-cameras/ - Cached - Similar

Digital Cameras – Best Latest **Digital Camera** Reviews | Features ...
digital camera - compare n best latest **digital** cameras prices and **digital** cameras reviews,
thinkdigit.com provides best **digital** cameras reviews and features ...
www.thinkdigit.com/Digital-Cameras-ca-35.php - Cached - Similar

Sponsored Links

Samsung® **Digital** Cameras
Record Videos & Watch in HD w/
a New Samsung **Digital Camera**.
www.Samsung.com

Low Low Prices on Brand Names
Ships Free, Save More Today!
www.TigerDirect.com
Google Checkout

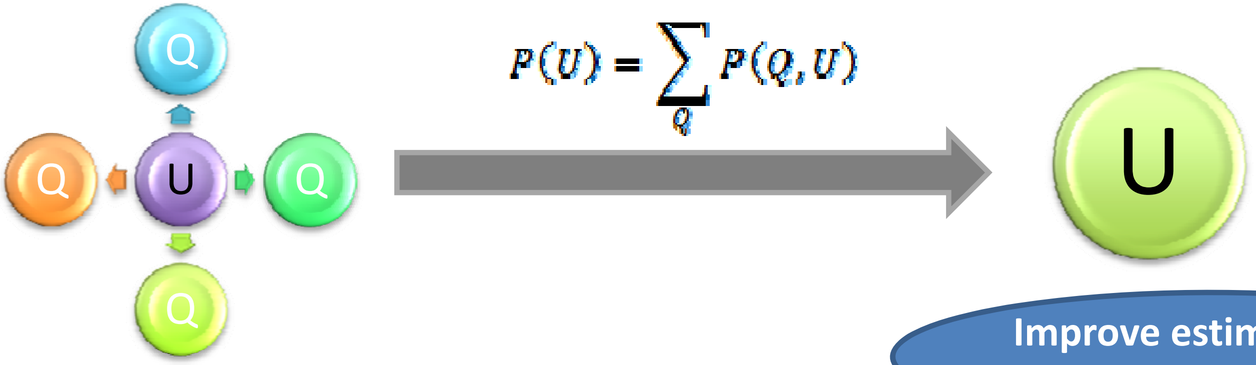
ls

dealnews.com

Digital Camera Sale
Awesome Deals on Top Brand **Digital**
Cameras only at Newegg.com!
www.Newegg.com
Show products from this advertiser

Digital Camera

Aggregates over *(Q,U)* pairs



$$P(U) = \sum_{Q} P(Q,U)$$

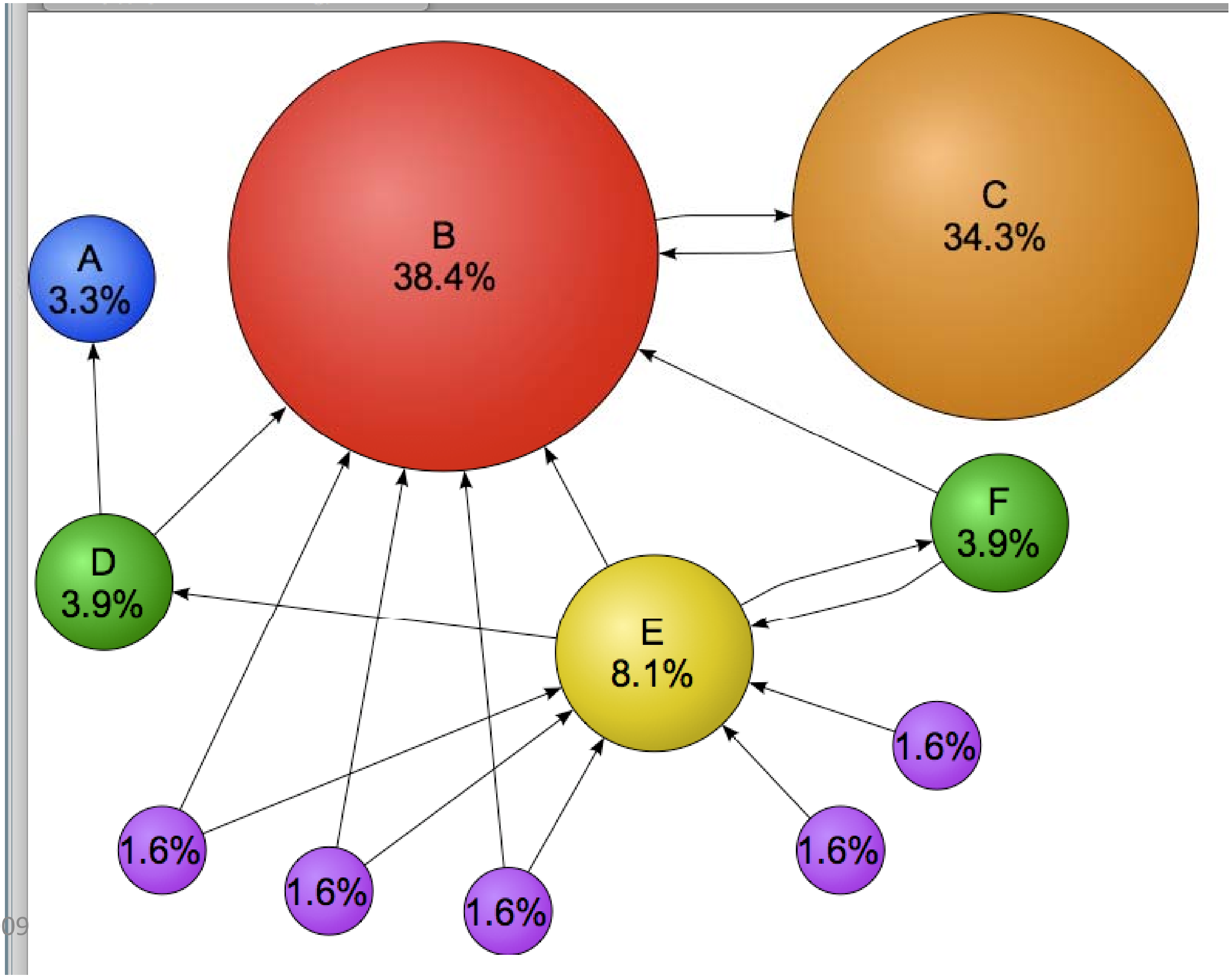
Improve estimation by adding ***features***

MODEL		**Q/U Features**				
		Static Rank	*Toolbar Counts*	*BM25F*	*Words In URL*	*Clicks*
Aggregates	*max*					
	median					
	sum					
	count					
	entropy					

Prior(U)

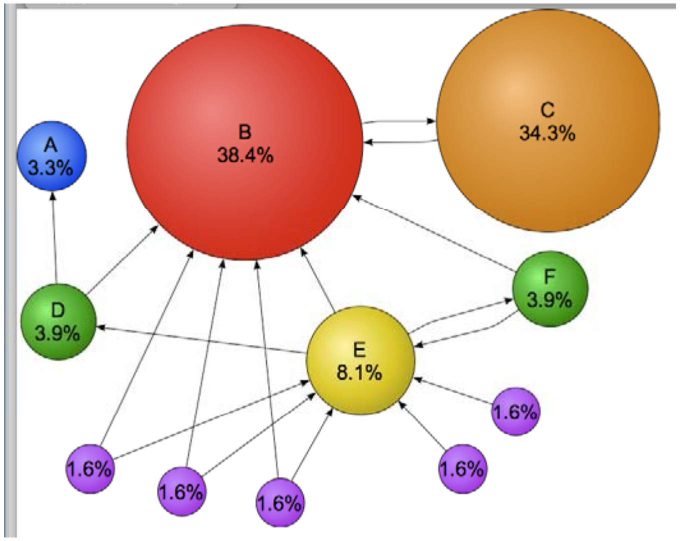
Improve estimation by adding ***aggregates***

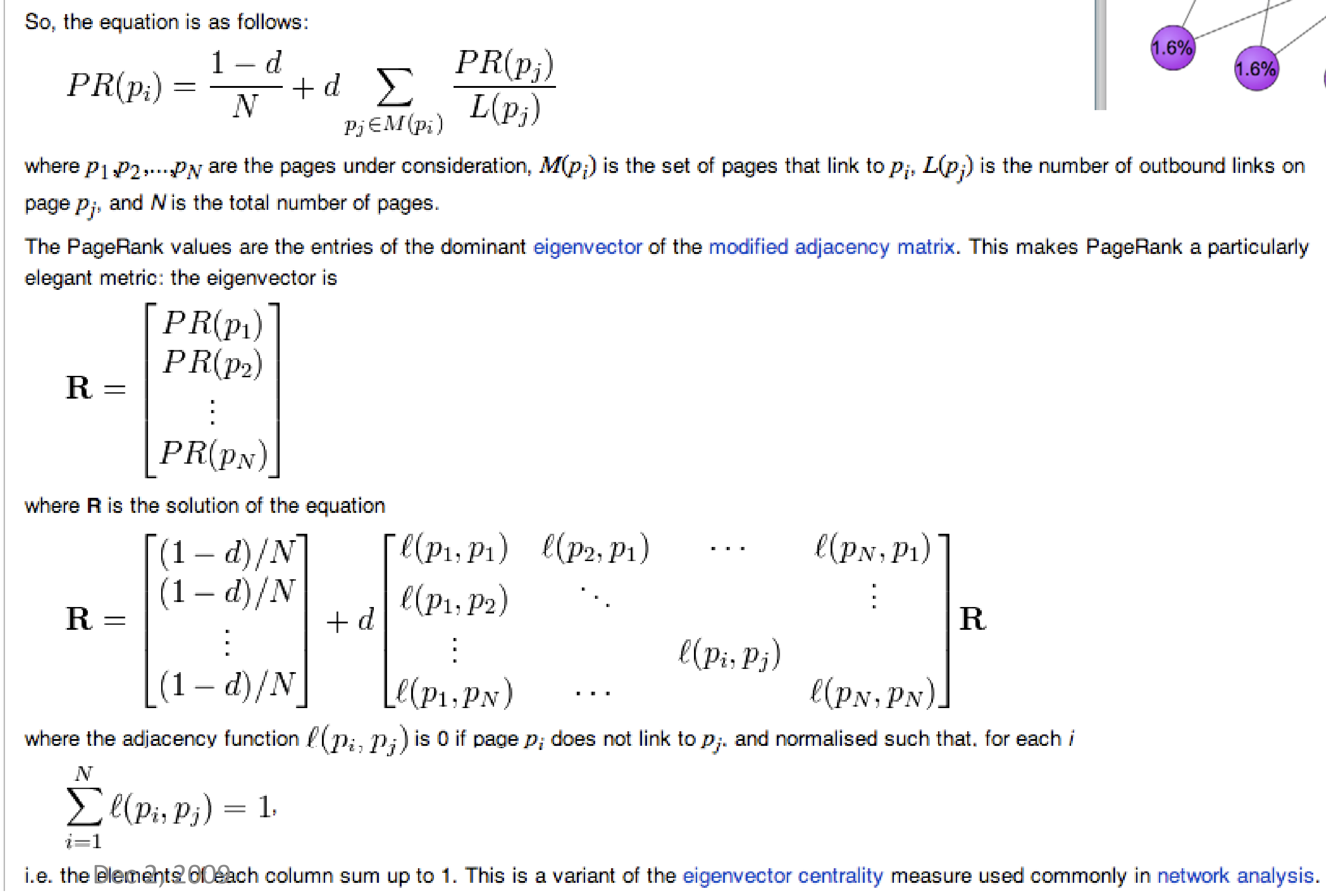
Page Rank (named after Larry Page) aka Static Rank & Random Surfer Model



Page Rank = 1st Eigenvector

http://en.wikipedia.org/wiki/PageRank





So, the equation is as follows:

$$PR(p_i) = \frac{1-d}{N} + d \sum_{p_j \in M(p_i)} \frac{PR(p_j)}{L(p_j)}$$

where $p_1, p_2, ..., p_N$ are the pages under consideration, $M(p_i)$ is the set of pages that link to p_i, $L(p_j)$ is the number of outbound links on page p_j, and N is the total number of pages.

The PageRank values are the entries of the dominant eigenvector of the modified adjacency matrix. This makes PageRank a particularly elegant metric: the eigenvector is

$$\mathbf{R} = \begin{bmatrix} PR(p_1) \\ PR(p_2) \\ \vdots \\ PR(p_N) \end{bmatrix}$$

where **R** is the solution of the equation

$$\mathbf{R} = \begin{bmatrix} (1-d)/N \\ (1-d)/N \\ \vdots \\ (1-d)/N \end{bmatrix} + d \begin{bmatrix} \ell(p_1,p_1) & \ell(p_2,p_1) & \cdots & \ell(p_N,p_1) \\ \ell(p_1,p_2) & \ddots & & \vdots \\ \vdots & & \ell(p_i,p_j) & \\ \ell(p_1,p_N) & \cdots & & \ell(p_N,p_N) \end{bmatrix} \mathbf{R}$$

where the adjacency function $\ell(p_i, p_j)$ is 0 if page p_i does not link to p_j, and normalised such that, for each i

$$\sum_{i=1}^{N} \ell(p_i, p_j) = 1,$$

i.e. the elements of each column sum up to 1. This is a variant of the eigenvector centrality measure used commonly in network analysis.

Document Priors are like Query Priors

- Human Ratings (HRS): Perfect judgments → *more likely*
- Static Rank (Page Rank): higher → *more likely*
- Textual Overlap: match → *more likely*
  - "cnn" → www.cnn.com (match)
- Popular:
  - lots of clicks → *more likely* (toolbar, slogs, glogs)
- Diversity/Entropy:
  - fewer plausible queries → *more likely*
- Broder's Taxonomy
  - Applies to documents as well
  - "cnn" → www.cnn.com (navigational)

Task Definition

- What will determine future clicks on the URL?
  - Past Clicks ?
  - High Static Rank ?
  - High Toolbar visitation counts ?
  - Precise Textual Match ?
  - All of the Above ?

- ~3k queries from the extracts
  - 350k URL’s
  - Past Clicks: February – May, 2008
  - Future Clicks: June, 2008

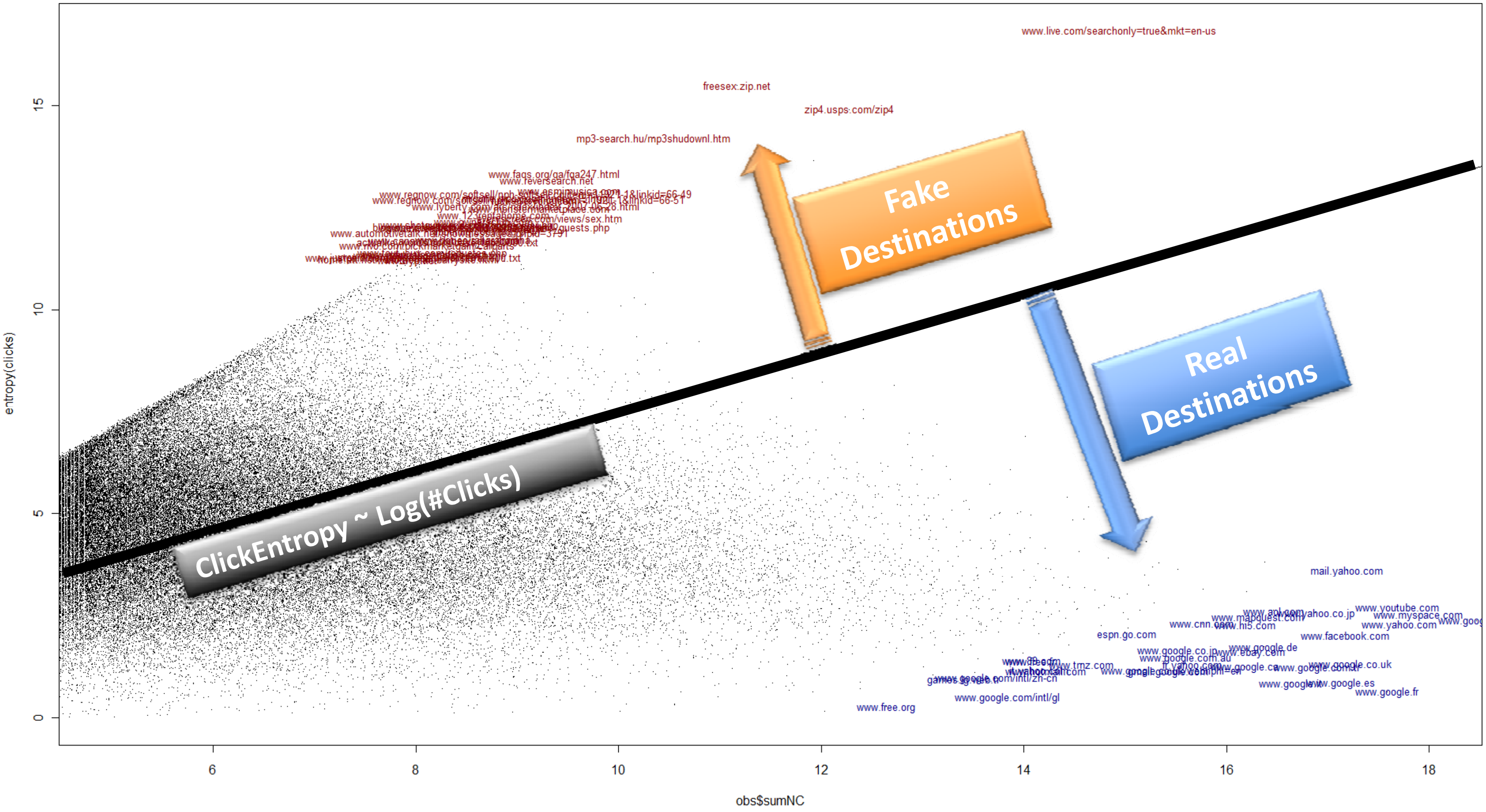
Estimating URL Popularity

URL Popularity	**Normalized RMSE Loss**		
	Extract	**Clicks**	**Extract + Clicks**
Linear Regression			
A: Regression	.619	.329	**.324**
B: Classification + Regression	-	.324	**.319**
Neural Network (3 Nodes in the Hidden Layer)			
C: Regression	.619	.311	**.300**

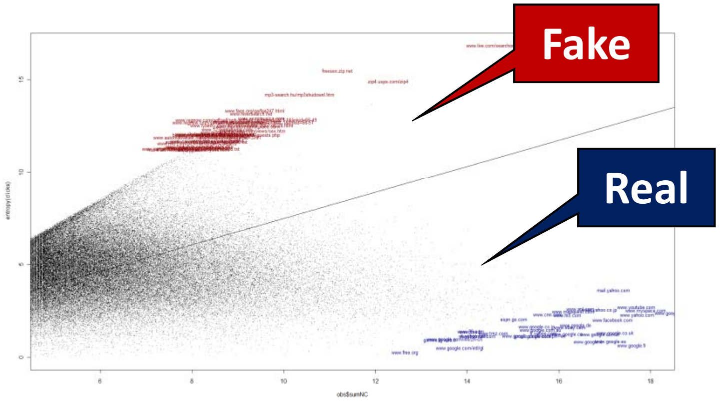
B is better than A

Extract + Clicks:
Better Together

Destinations by Residuals



Real and Fake Destinations



Residuals – Highest Quartile

actualkeywords.com/base_top50000.txt
blog.nbc.com/heroes/2007/04/wine_and_guests.php
everyscreen.com/views/sex.htm
freesex.zip.net
fuck-everyone.com
home.att.net/~btuttleman/barrysite.html
jibbering.com/blog/p=57
migune.nipox.com/index-15.html
mp3-search.hu/mp3shudownl.htm
www.123rentahome.com
www.automotivetalk.net/showmessages.phpid=3791
www.canammachinerysales.com
www.cardpostage.com/zorn.htm
www.driverguide.com/drilist.htm
www.driverguide.com/drivers2.htm
www.esmimusica.com

Residuals – Lowest Quartile

espn.go.com
fr.yahoo.com
games.lg.web.tr
gmail.google.com
it.yahoo.com
mail.yahoo.com
www.89.com
www.aol.com
www.cnn.com
www.ebay.com
www.facebook.com
www.free.fr
www.free.org
www.google.ca
www.google.co.jp
www.google.co.uk

Fake Destination Example

```
free
new
school
home
county
online
lyrics
download
video
car
city
sale
texas
music
de
pictures
florida
hotel
real
state
sex
high
mp3
center
uk
2007
california
movie
software
best
estate
black
```

actualkeywords.com/base_top50000.txt

Clicked ~110,000 times

In response to ~16,000 unique queries

Dictionary Attack

Learning to Rank with Document Priors

- **Baseline: Feature Set A**
  - Textual Features **(5 features)**

- **Baseline: Feature Set B**
  - Textual Features + Static Rank **(7 features)**

- **Baseline: Feature Set C**
  - All features, with click-based features filtered **(382 features)**

- **Treatment: Baseline + 5 Click Aggregate Features**
  - Max, Median, Entropy, Sum, Count

Summary: Information Retrieval (IR)

- Boolean Combinations of Keywords
  - Popular with Intermediaries (Librarians)
- Rank Retrieval
  - Sort a collection of documents
    - (e.g., scientific papers, abstracts, paragraphs)
    - by how much they ‘‘match’’ a query
  - The query can be a (short) sequence of keywords
    - or arbitrary text (e.g., one of the documents)
- Logs of User Behavior (Clicks, Toolbar)
  - Solitaire → Multi-Player Game:
    - Authors, Users, Advertisers, Spammers
  - More Users than Authors → More Information in Logs than Docs
  - Learning to Rank:
    - Use Machine Learning to combine doc features & log features