Information Retrieval

Course Summary

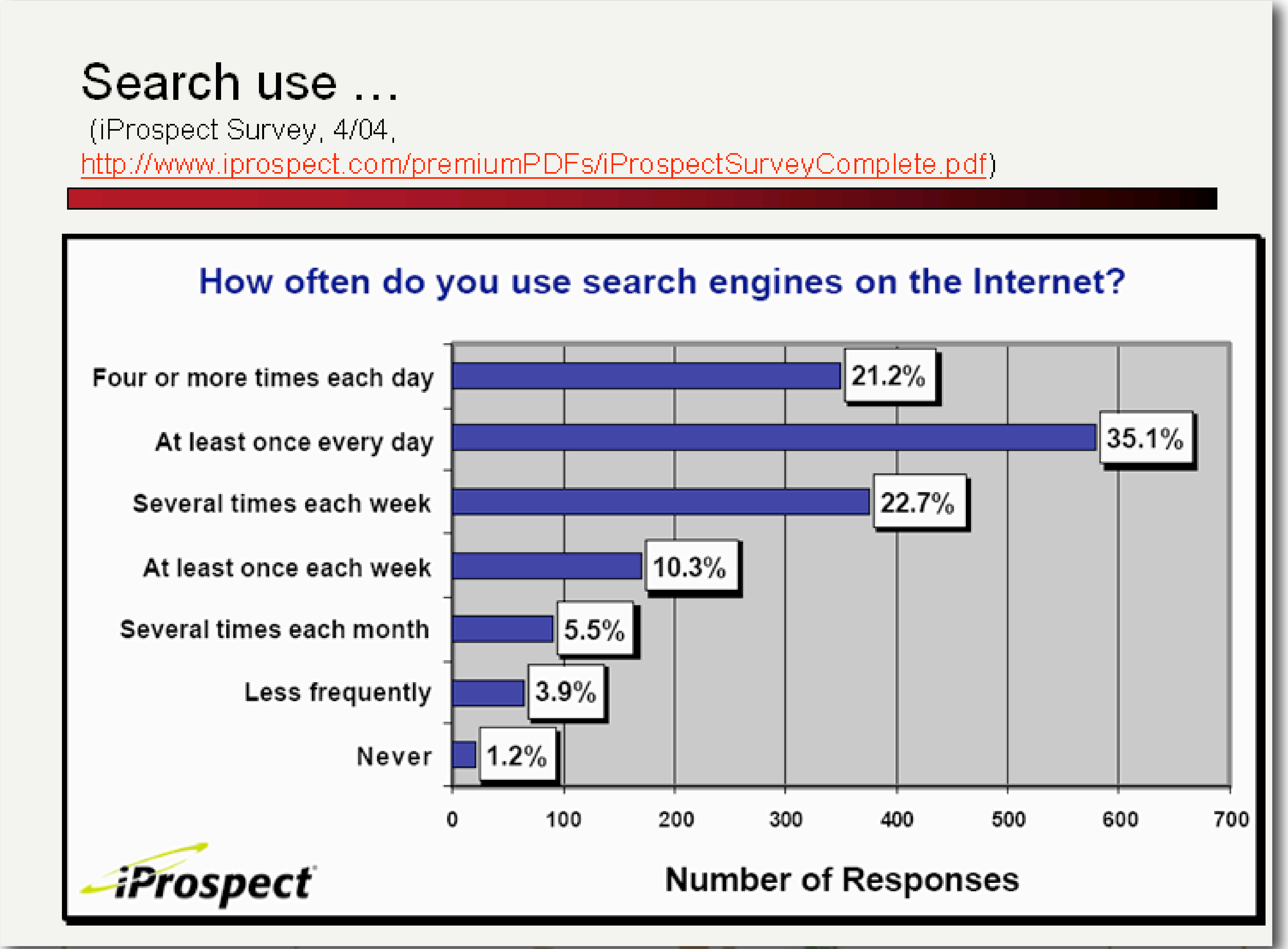
INF 141

Donald J. Patterson

Learning Objective

“Know what you know”

Introduction



Search use …

(iProspect Survey, 4/04, http://www.iprospect.com/premiumPDFs/iProspectSurveyComplete.pdf)

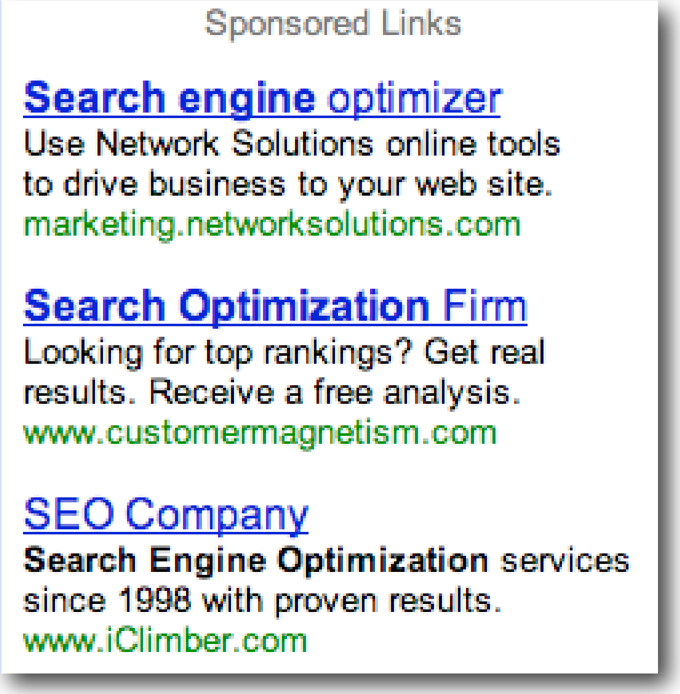
Introduction

Without search engines the web wouldn’t scale

- Search turned out to be the best mechanism for advertising on the web, a $15 billion plus industry.
    - Growing very fast (entire US advertising industry is $250 billion though)
    - Sponsored search marketing is about $10 billion

Ads vs. Search Results

- Google has maintained that ads (based on vendors bidding for search queries) do not affect vendors ranking in search results

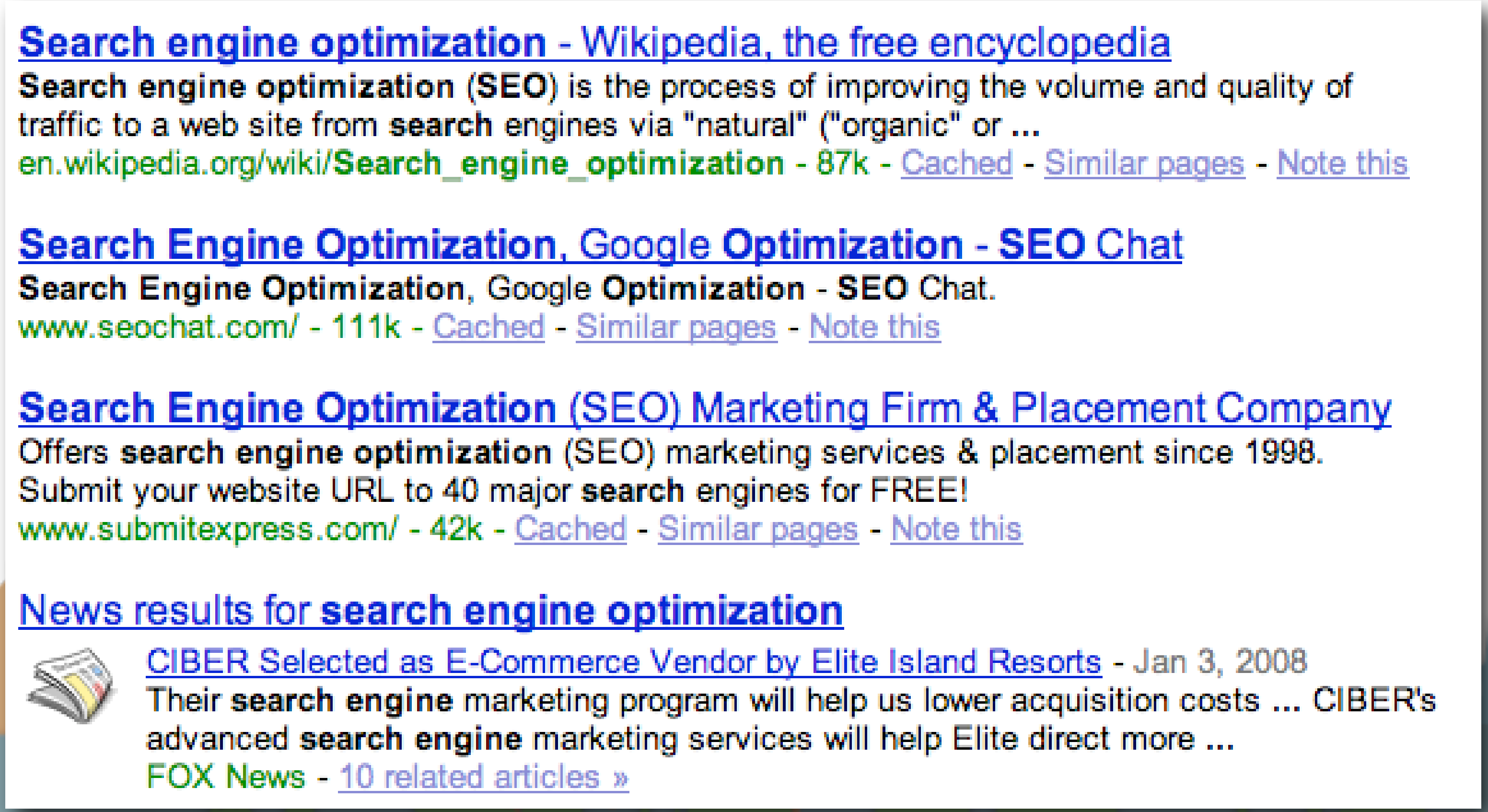


Sponsored Links

Search engine optimizer
Use Network Solutions online tools to drive business to your web site.
marketing.networksolutions.com

Search Optimization Firm
Looking for top rankings? Get real results. Receive a free analysis.
www.customermagnetism.com

SEO Company
Search Engine Optimization services since 1998 with proven results.
www.iClimber.com



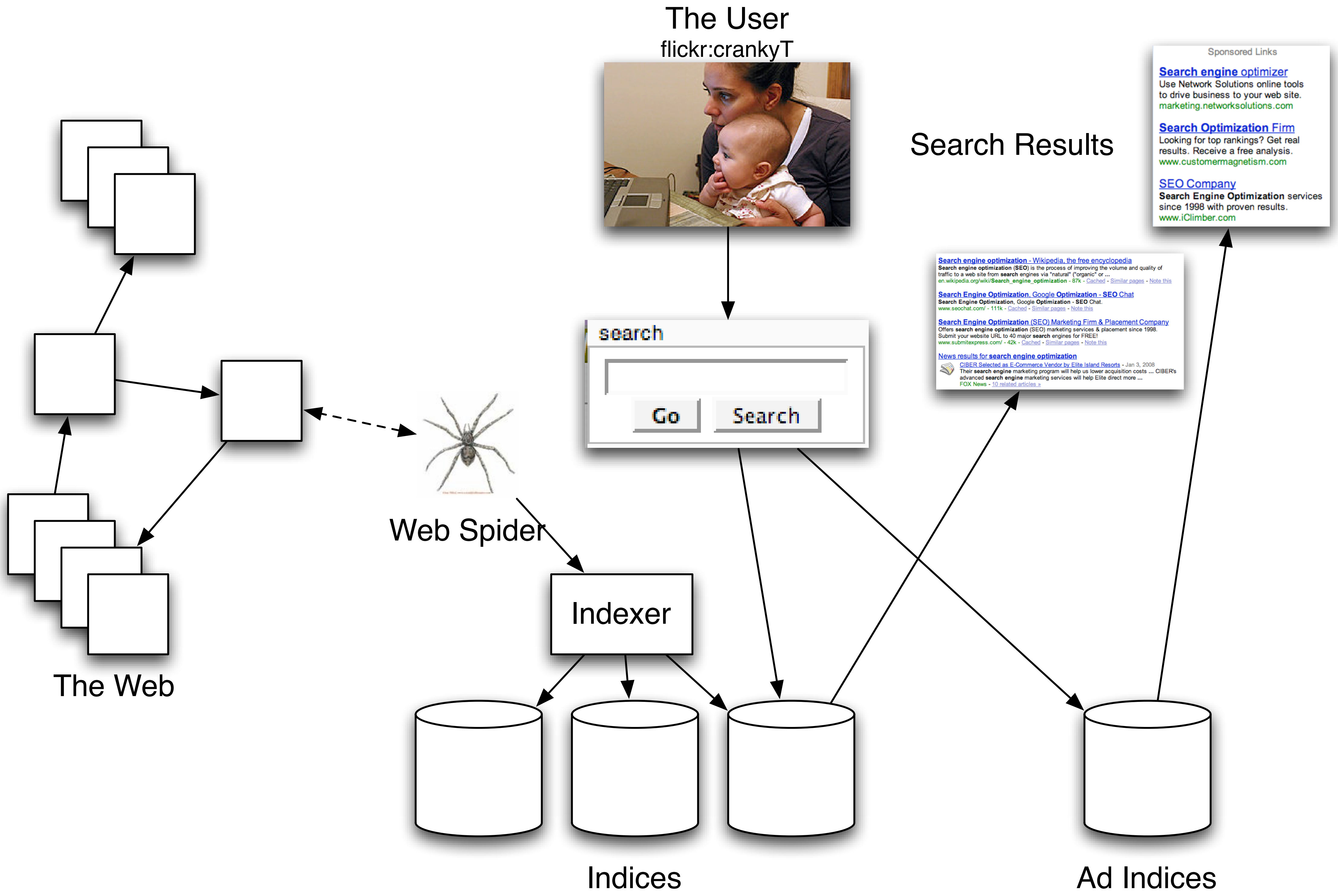
Search engine optimization - Wikipedia, the free encyclopedia
Search engine optimization (**SEO**) is the process of improving the volume and quality of traffic to a web site from **search** engines via "natural" ("organic" or ...
en.wikipedia.org/wiki/**Search_engine_optimization** - 87k - Cached - Similar pages - Note this

Search Engine Optimization, Google **Optimization** - **SEO** Chat
Search Engine Optimization, Google **Optimization** - **SEO** Chat.
www.seochat.com/ - 111k - Cached - Similar pages - Note this

Search Engine Optimization (SEO) Marketing Firm & Placement Company
Offers **search engine optimization** (SEO) marketing services & placement since 1998. Submit your website URL to 40 major **search** engines for FREE!
www.submitexpress.com/ - 42k - Cached - Similar pages - Note this

News results for **search engine optimization**
CIBER Selected as E-Commerce Vendor by Elite Island Resorts - Jan 3, 2008
Their **search engine** marketing program will help us lower acquisition costs ... CIBER's advanced **search engine** marketing services will help Elite direct more ...
FOX News - 10 related articles »

Web Search Basics



The User

flickr:crankyT

Search Results

Sponsored Links

Search engine optimizer
Use Network Solutions online tools
to drive business to your web site.
marketing.networksolutions.com

Search Optimization Firm
Looking for top rankings? Get real
results. Receive a free analysis.
www.customermagnetism.com

SEO Company
Search Engine Optimization services
since 1998 with proven results.
www.iClimber.com

Search engine optimization - Wikipedia, the free encyclopedia
Search engine optimization (**SEO**) is the process of improving the volume and quality of traffic to a web site from **search** engines via "natural" ("organic" or ...
en.wikipedia.org/wiki/**Search_engine_optimization** - 87k - Cached - Similar pages - Note this

Search Engine Optimization, Google **Optimization** - **SEO** Chat
Search Engine Optimization, Google **Optimization** - **SEO** Chat.
www.seochat.com/ - 111k - Cached - Similar pages - Note this

Search Engine Optimization (SEO) Marketing Firm & Placement Company
Offers **search engine optimization** (SEO) marketing services & placement since 1998.
Submit your website URL to 40 major **search** engines for FREE!
www.submitexpress.com/ - 42k - Cached - Similar pages - Note this

News results for **search engine optimization**
CIBER Selected as E-Commerce Vendor by Elite Island Resorts - Jan 3, 2008
Their **search engine** marketing program will help us lower acquisition costs ... CIBER's advanced **search engine** marketing services will help Elite direct more ...
FOX News - 10 related articles »

Web Spider

Indexer

The Web

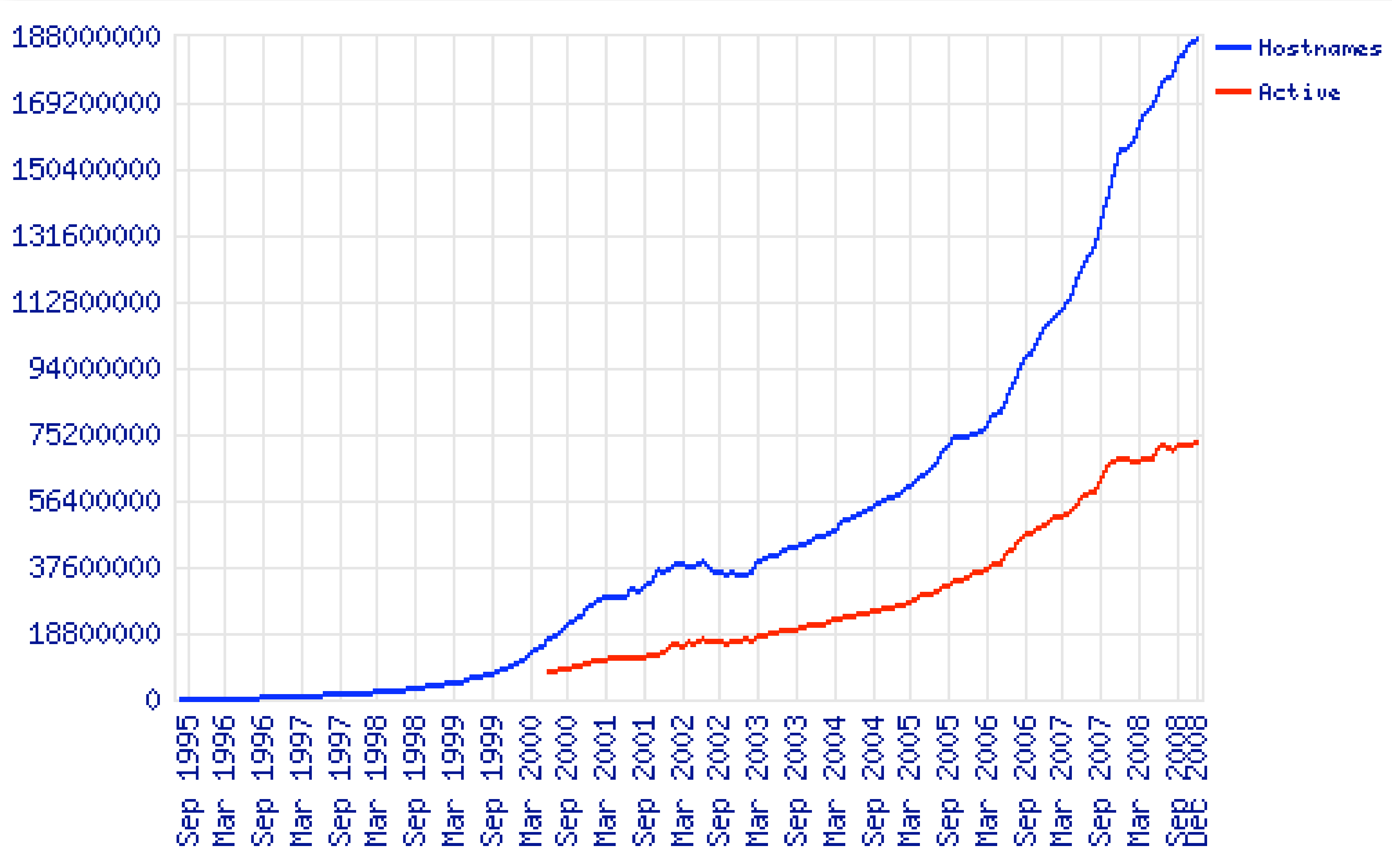
Indices

Ad Indices

Size of the Web

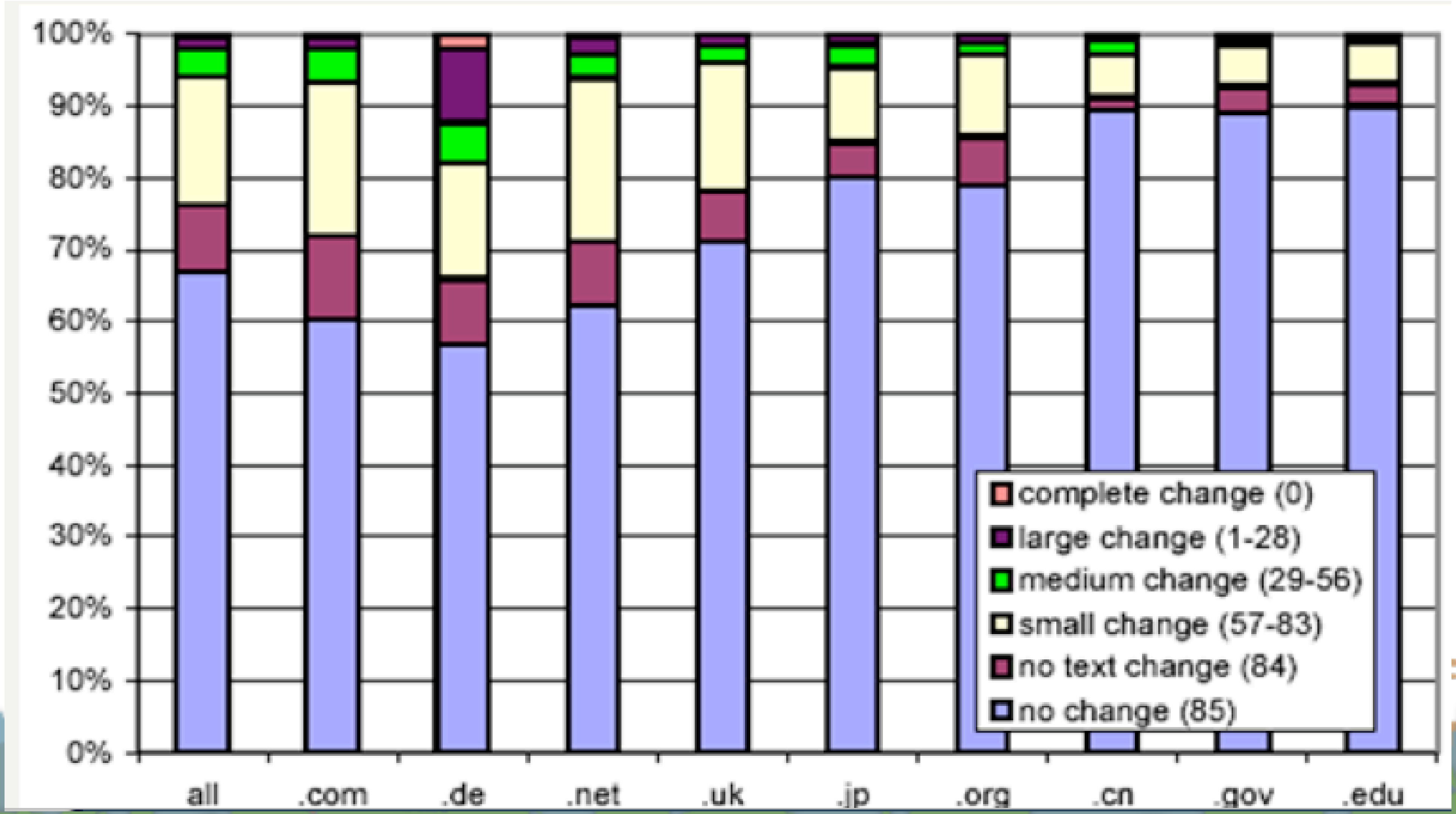
How big is the web?

- Netcraft Web Server Survey



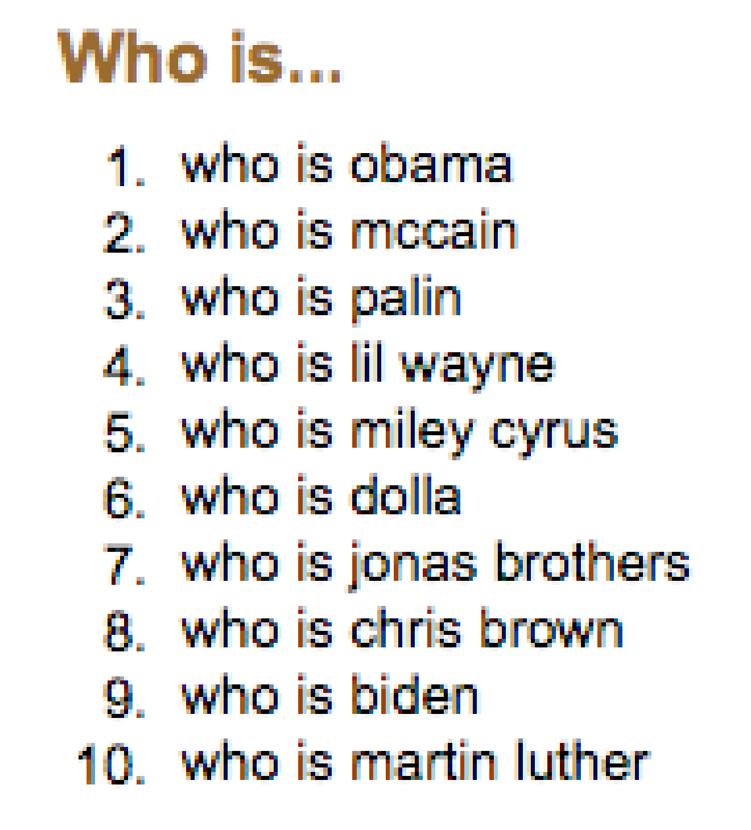
Rate of change

- Fetterly et al. study in 2002
  - 150 million pages over 11 weekly crawls
  - Bucketed into 85 groups according to amount of change



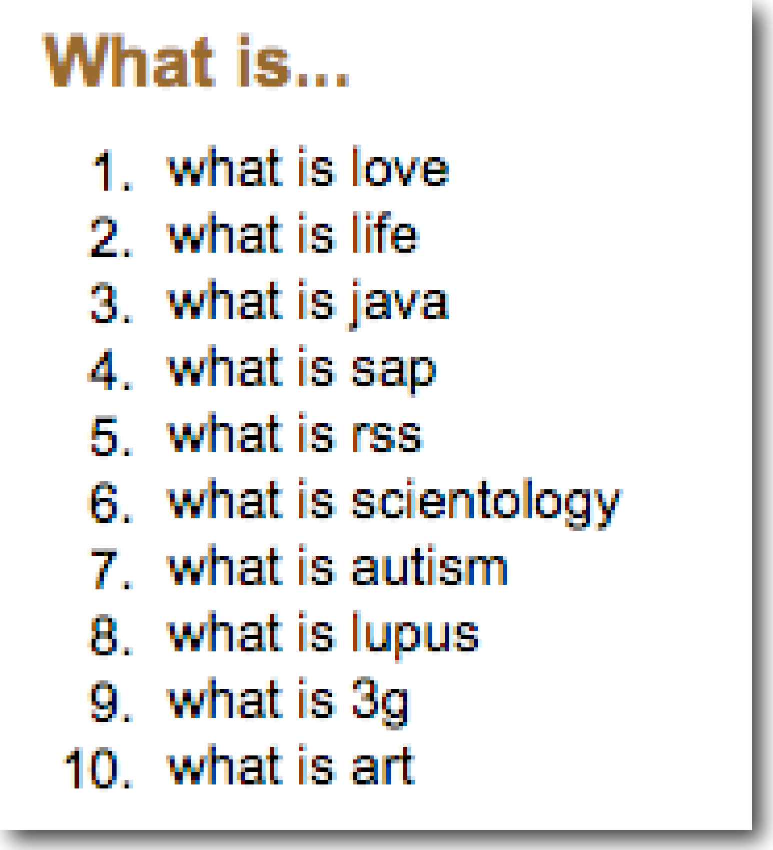
Top queries

- Most are related to sex
- 2008 Who What How (Google)



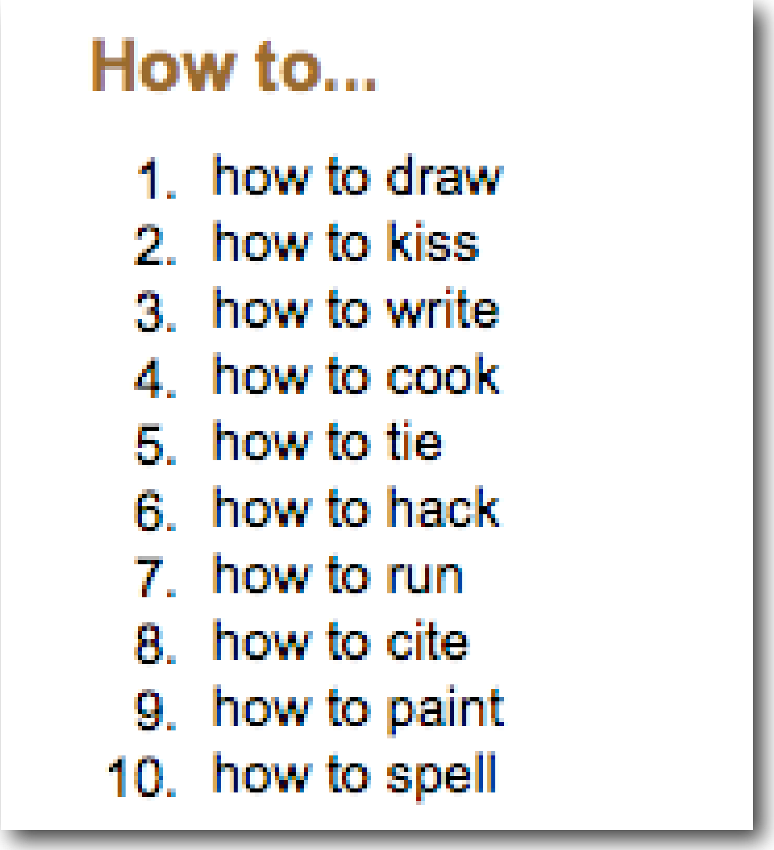
Who is...

1. who is obama
2. who is mccain
3. who is palin
4. who is lil wayne
5. who is miley cyrus
6. who is dolla
7. who is jonas brothers
8. who is chris brown
9. who is biden
10. who is martin luther



What is...

1. what is love
2. what is life
3. what is java
4. what is sap
5. what is rss
6. what is scientology
7. what is autism
8. what is lupus
9. what is 3g
10. what is art

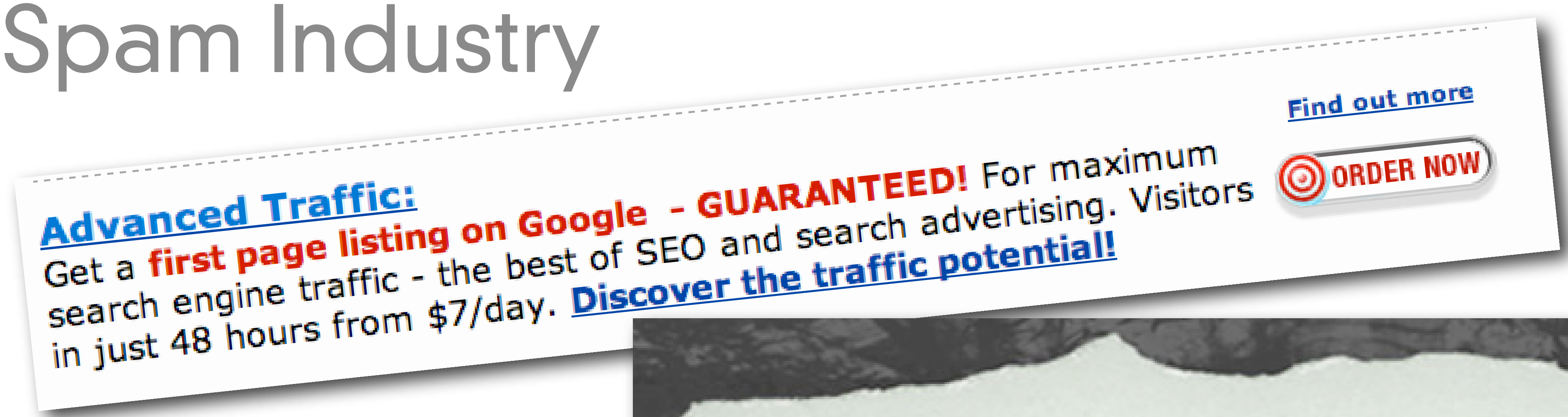


How to...

1. how to draw
2. how to kiss
3. how to write
4. how to cook
5. how to tie
6. how to hack
7. how to run
8. how to cite
9. how to paint
10. how to spell

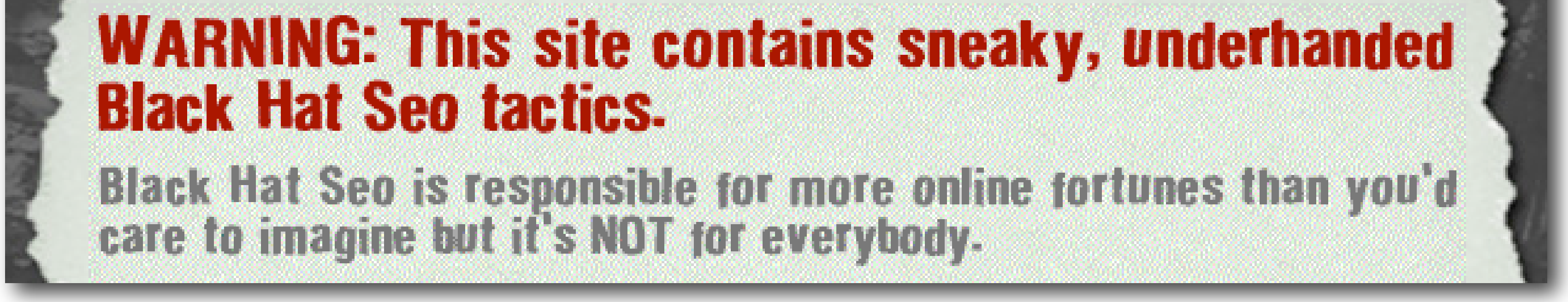
- http://www.google.com/intl/en/press/zeitgeist2008/mind.html

Spam Industry



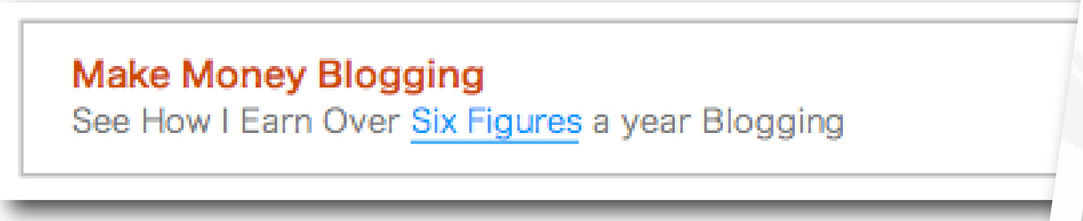
Advanced Traffic:

Get a **first page listing on Google - GUARANTEED!** For maximum search engine traffic - the best of SEO and search advertising. Visitors in just 48 hours from $7/day. **Discover the traffic potential!**



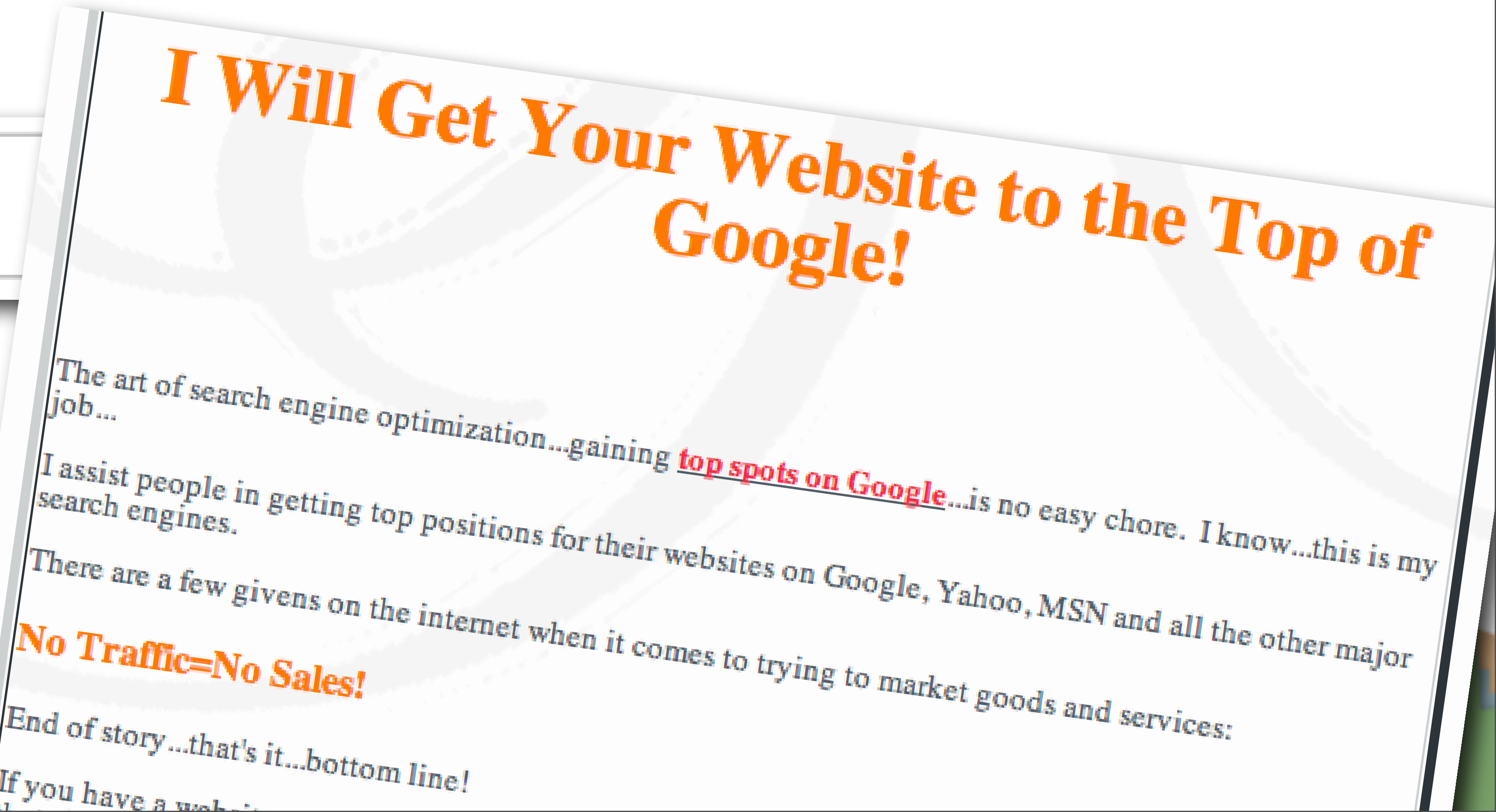
WARNING: This site contains sneaky, underhanded Black Hat Seo tactics.

Black Hat Seo is responsible for more online fortunes than you'd care to imagine but it's NOT for everybody.



Make Money Blogging

See How I Earn Over Six Figures a year Blogging



I Will Get Your Website to the Top of Google!

The art of search engine optimization...gaining top spots on Google...is no easy chore. I know...this is my job...

I assist people in getting top positions for their websites on Google, Yahoo, MSN and all the other major search engines.

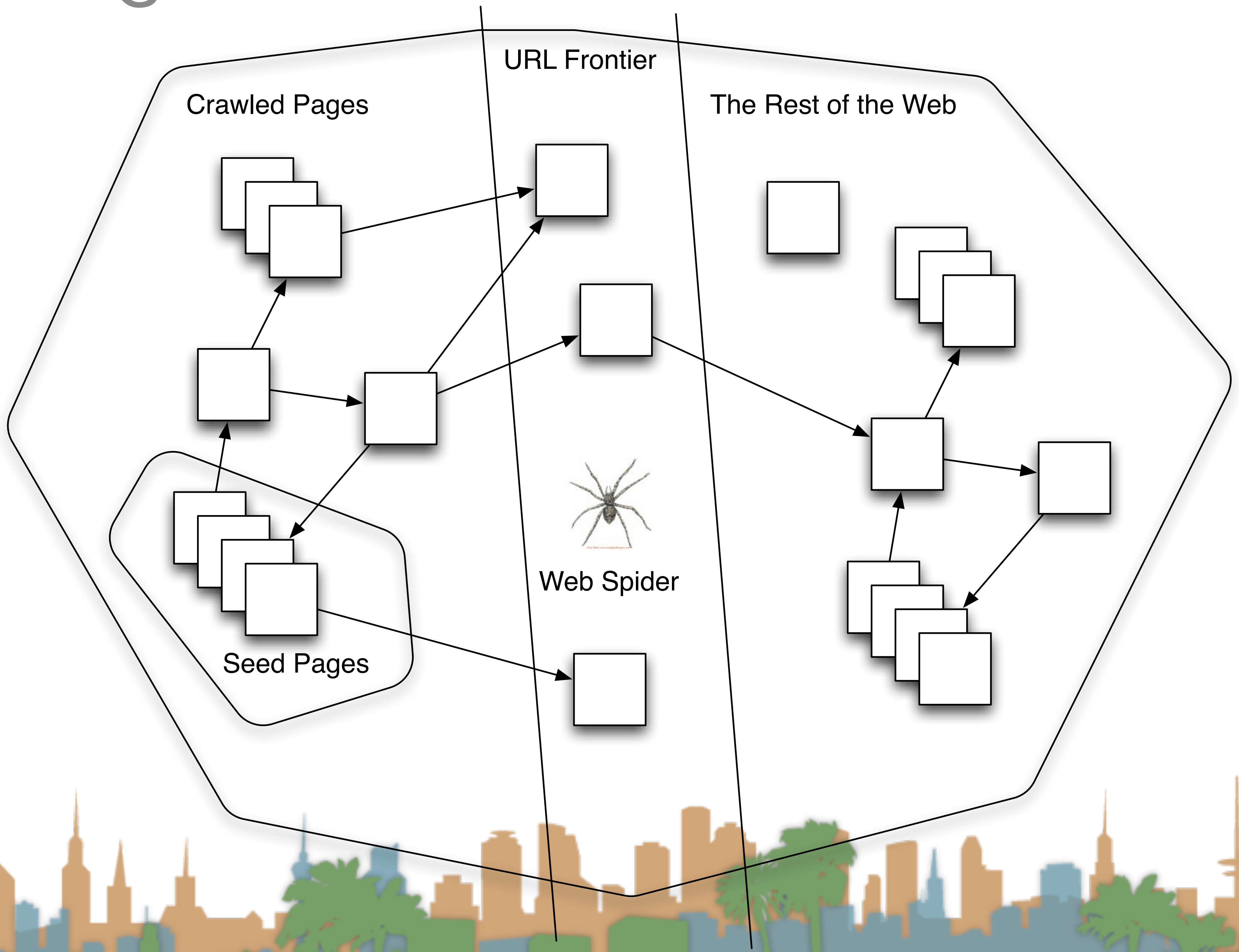
There are a few givens on the internet when it comes to trying to market goods and services:

No Traffic=No Sales!

End of story...that's it...bottom line!

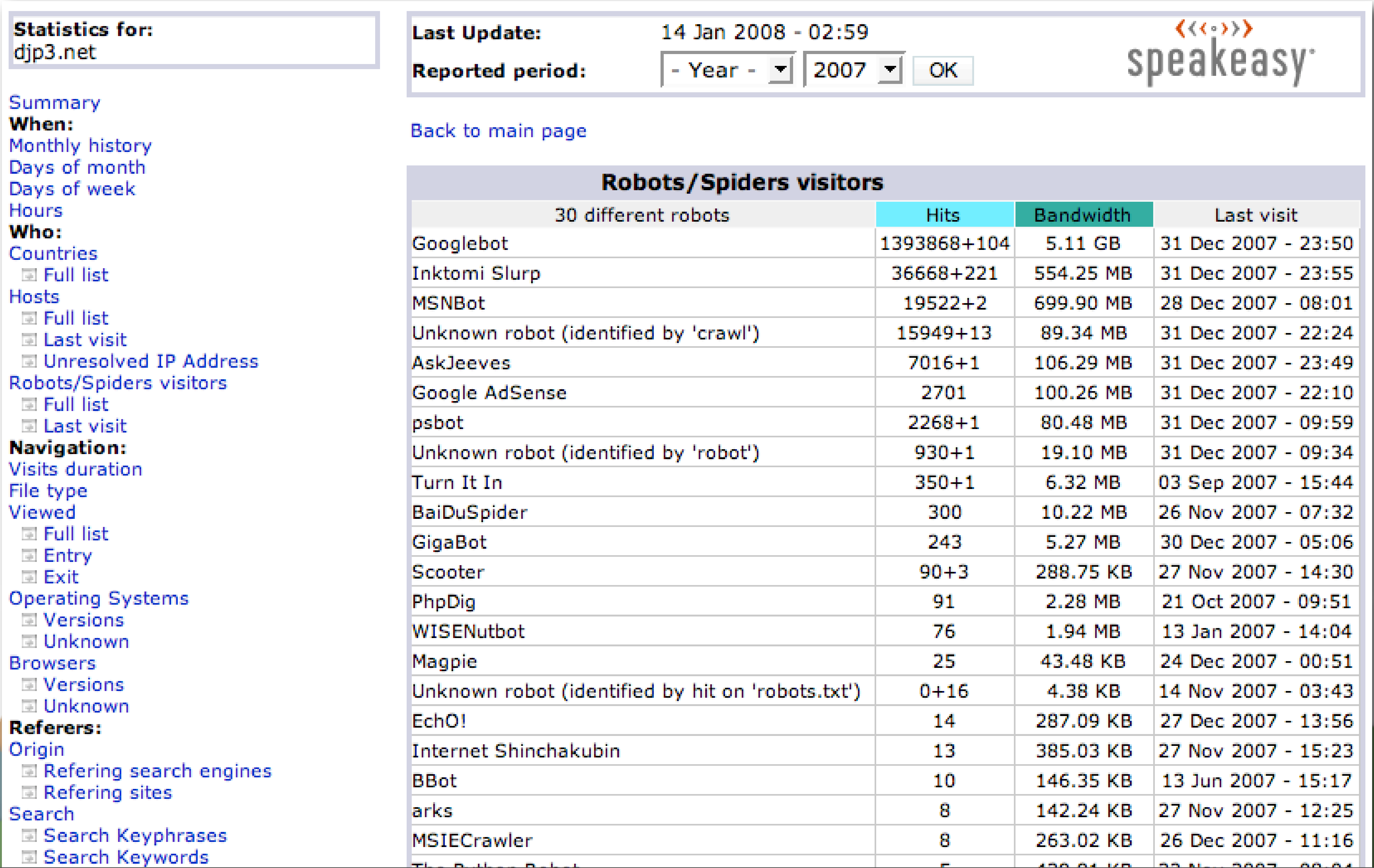
If you have a

Crawling the web



URL Frontier

Politeness?



Statistics for:
djp3.net

Last Update: 14 Jan 2008 - 02:59

Reported period: - Year - ▾ 2007 ▾ OK

Summary
When:
Monthly history
Days of month
Days of week
Hours
Who:
Countries
Full list
Hosts
Full list
Last visit
Unresolved IP Address
Robots/Spiders visitors
Full list
Last visit
Navigation:
Visits duration
File type
Viewed
Full list
Entry
Exit
Operating Systems
Versions
Unknown
Browsers
Versions
Unknown
Referers:
Origin
Refering search engines
Refering sites
Search
Search Keyphrases
Search Keywords

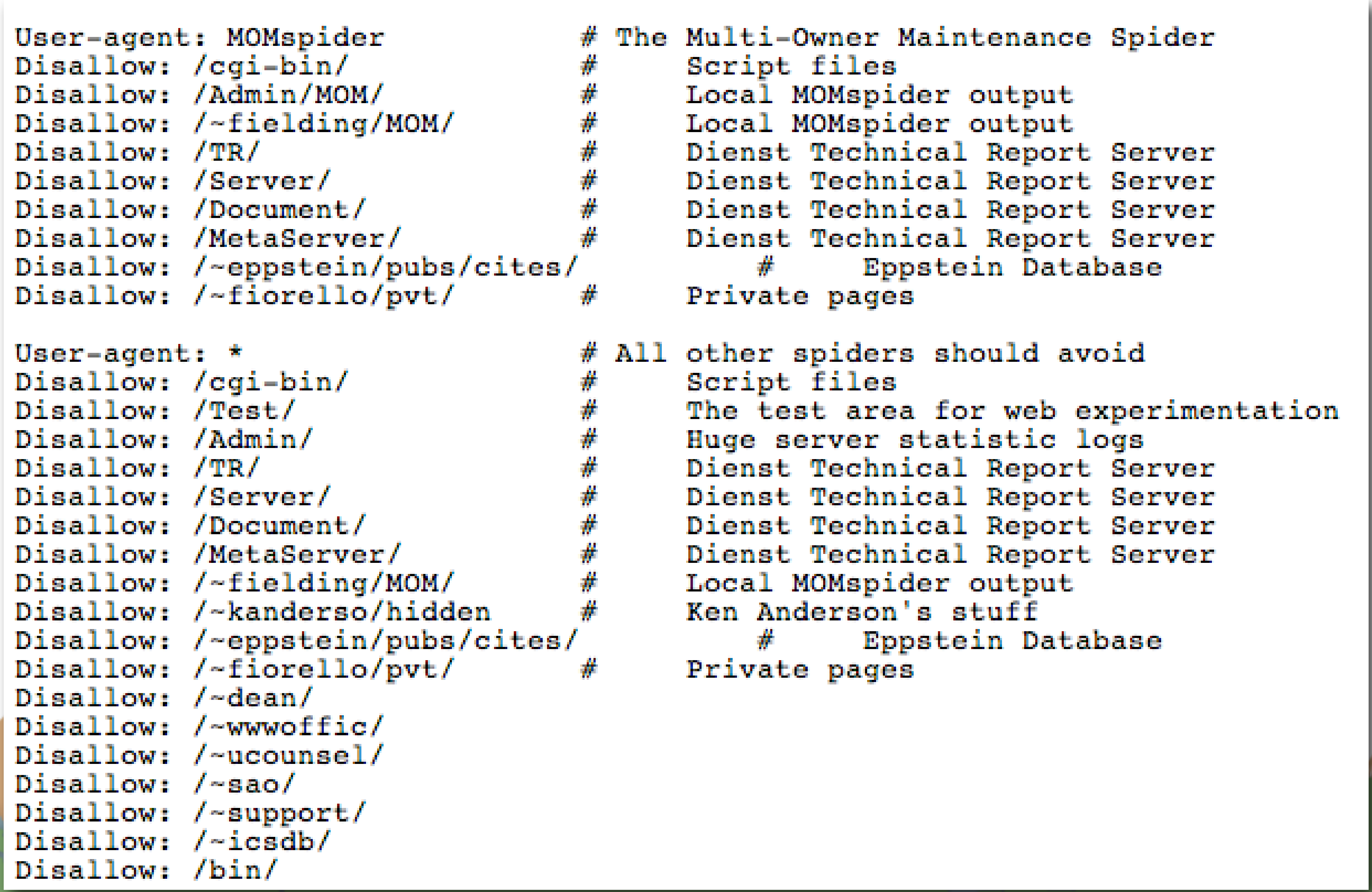
Back to main page

Robots/Spiders visitors

30 different robots	Hits	Bandwidth	Last visit
Googlebot	1393868+104	5.11 GB	31 Dec 2007 - 23:50
Inktomi Slurp	36668+221	554.25 MB	31 Dec 2007 - 23:55
MSNBot	19522+2	699.90 MB	28 Dec 2007 - 08:01
Unknown robot (identified by 'crawl')	15949+13	89.34 MB	31 Dec 2007 - 22:24
AskJeeves	7016+1	106.29 MB	31 Dec 2007 - 23:49
Google AdSense	2701	100.26 MB	31 Dec 2007 - 22:10
psbot	2268+1	80.48 MB	31 Dec 2007 - 09:59
Unknown robot (identified by 'robot')	930+1	19.10 MB	31 Dec 2007 - 09:34
Turn It In	350+1	6.32 MB	03 Sep 2007 - 15:44
BaiDuSpider	300	10.22 MB	26 Nov 2007 - 07:32
GigaBot	243	5.27 MB	30 Dec 2007 - 05:06
Scooter	90+3	288.75 KB	27 Nov 2007 - 14:30
PhpDig	91	2.28 MB	21 Oct 2007 - 09:51
WISENutbot	76	1.94 MB	13 Jan 2007 - 14:04
Magpie	25	43.48 KB	24 Dec 2007 - 00:51
Unknown robot (identified by hit on 'robots.txt')	0+16	4.38 KB	14 Nov 2007 - 03:43
EchO!	14	287.09 KB	27 Dec 2007 - 13:56
Internet Shinchakubin	13	385.03 KB	27 Nov 2007 - 15:23
BBot	10	146.35 KB	13 Jun 2007 - 15:17
arks	8	142.24 KB	27 Nov 2007 - 12:25
MSIECrawler	8	263.02 KB	26 Dec 2007 - 11:16
The Python Robot	5	[illegible]	[illegible]

Robots.txt Example

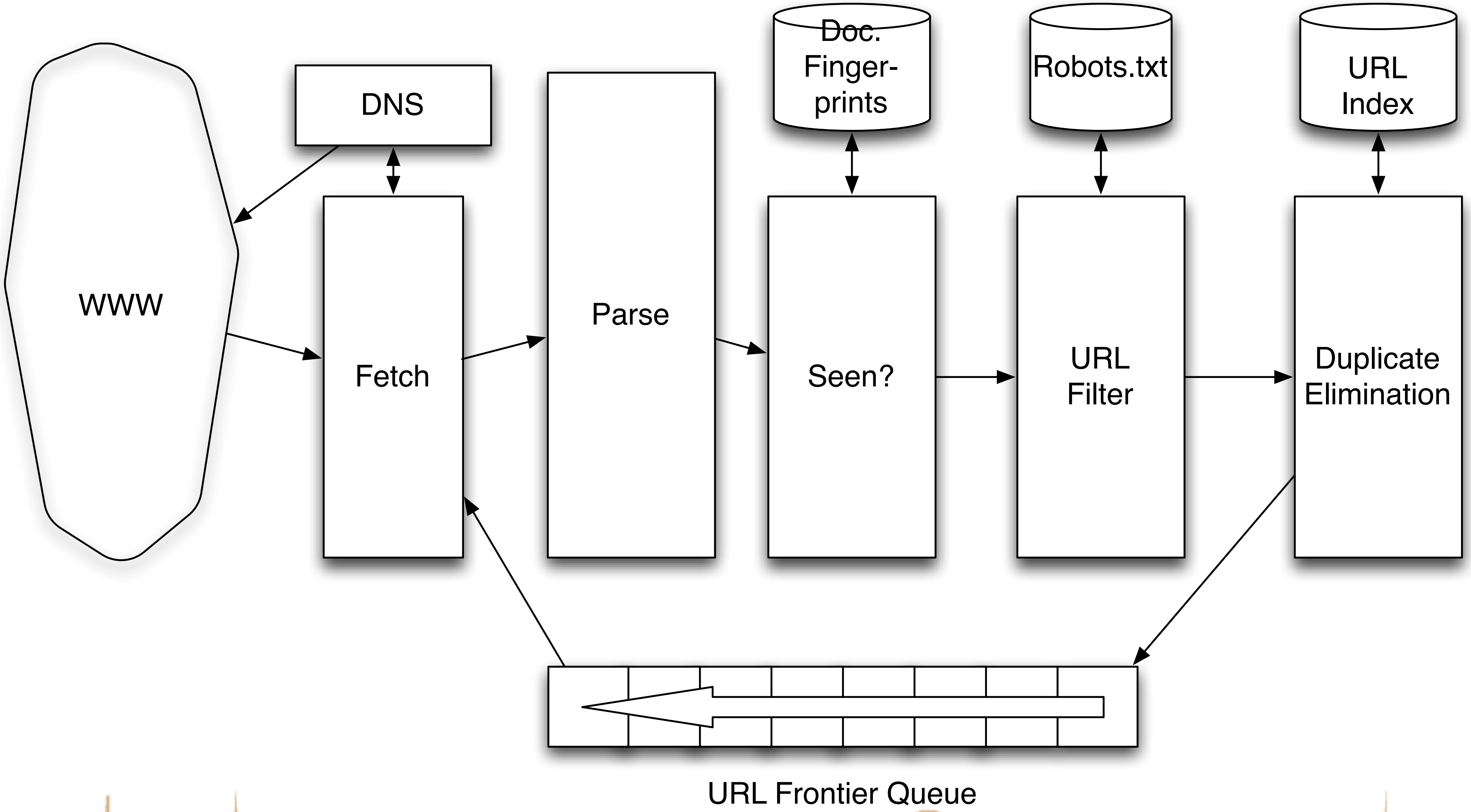
- http://www.ics.uci.edu/robots.txt



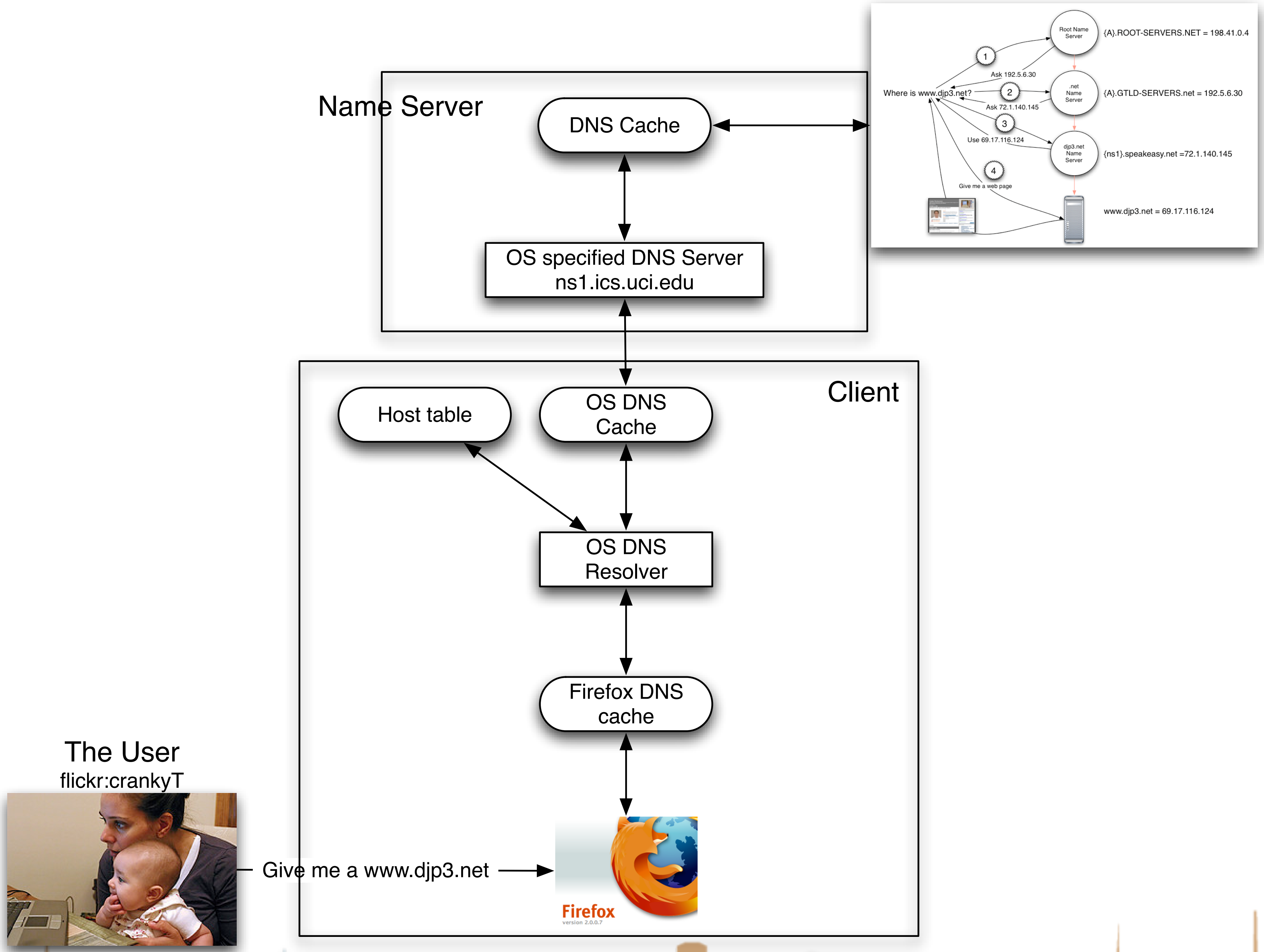
```
User-agent: MOMspider           # The Multi-Owner Maintenance Spider
Disallow: /cgi-bin/             #     Script files
Disallow: /Admin/MOM/           #     Local MOMspider output
Disallow: /~fielding/MOM/       #     Local MOMspider output
Disallow: /TR/                  #     Dienst Technical Report Server
Disallow: /Server/              #     Dienst Technical Report Server
Disallow: /Document/            #     Dienst Technical Report Server
Disallow: /MetaServer/          #     Dienst Technical Report Server
Disallow: /~eppstein/pubs/cites/          #     Eppstein Database
Disallow: /~fiorello/pvt/       #     Private pages

User-agent: *                   # All other spiders should avoid
Disallow: /cgi-bin/             #     Script files
Disallow: /Test/                #     The test area for web experimentation
Disallow: /Admin/               #     Huge server statistic logs
Disallow: /TR/                  #     Dienst Technical Report Server
Disallow: /Server/              #     Dienst Technical Report Server
Disallow: /Document/            #     Dienst Technical Report Server
Disallow: /MetaServer/          #     Dienst Technical Report Server
Disallow: /~fielding/MOM/       #     Local MOMspider output
Disallow: /~kanderso/hidden     #     Ken Anderson's stuff
Disallow: /~eppstein/pubs/cites/          #     Eppstein Database
Disallow: /~fiorello/pvt/       #     Private pages
Disallow: /~dean/
Disallow: /~wwwoffic/
Disallow: /~ucounsel/
Disallow: /~sao/
Disallow: /~support/
Disallow: /~icsdb/
Disallow: /bin/
```

A Robust Crawl Architecture



What really happens



Name Server

DNS Cache

OS specified DNS Server
ns1.ics.uci.edu

Client

Host table

OS DNS
Cache

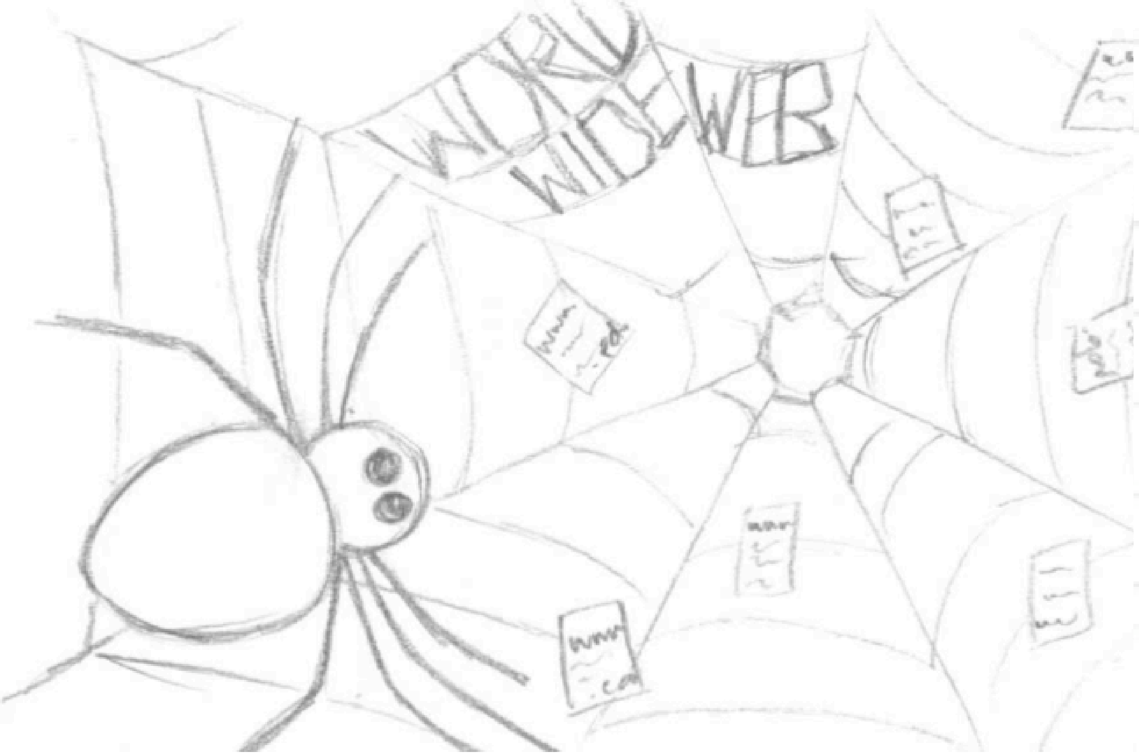
OS DNS
Resolver

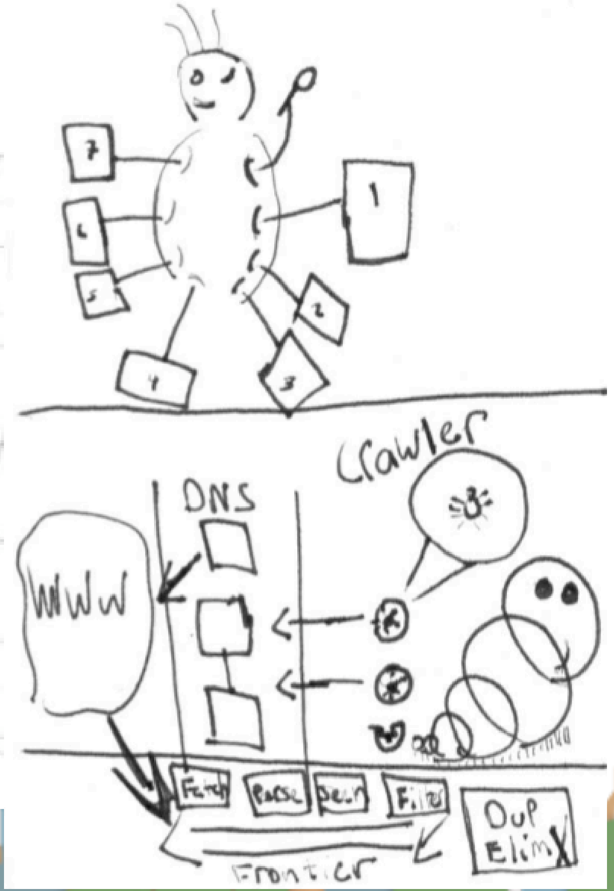
Firefox DNS
cache

The User
flickr:crankyT

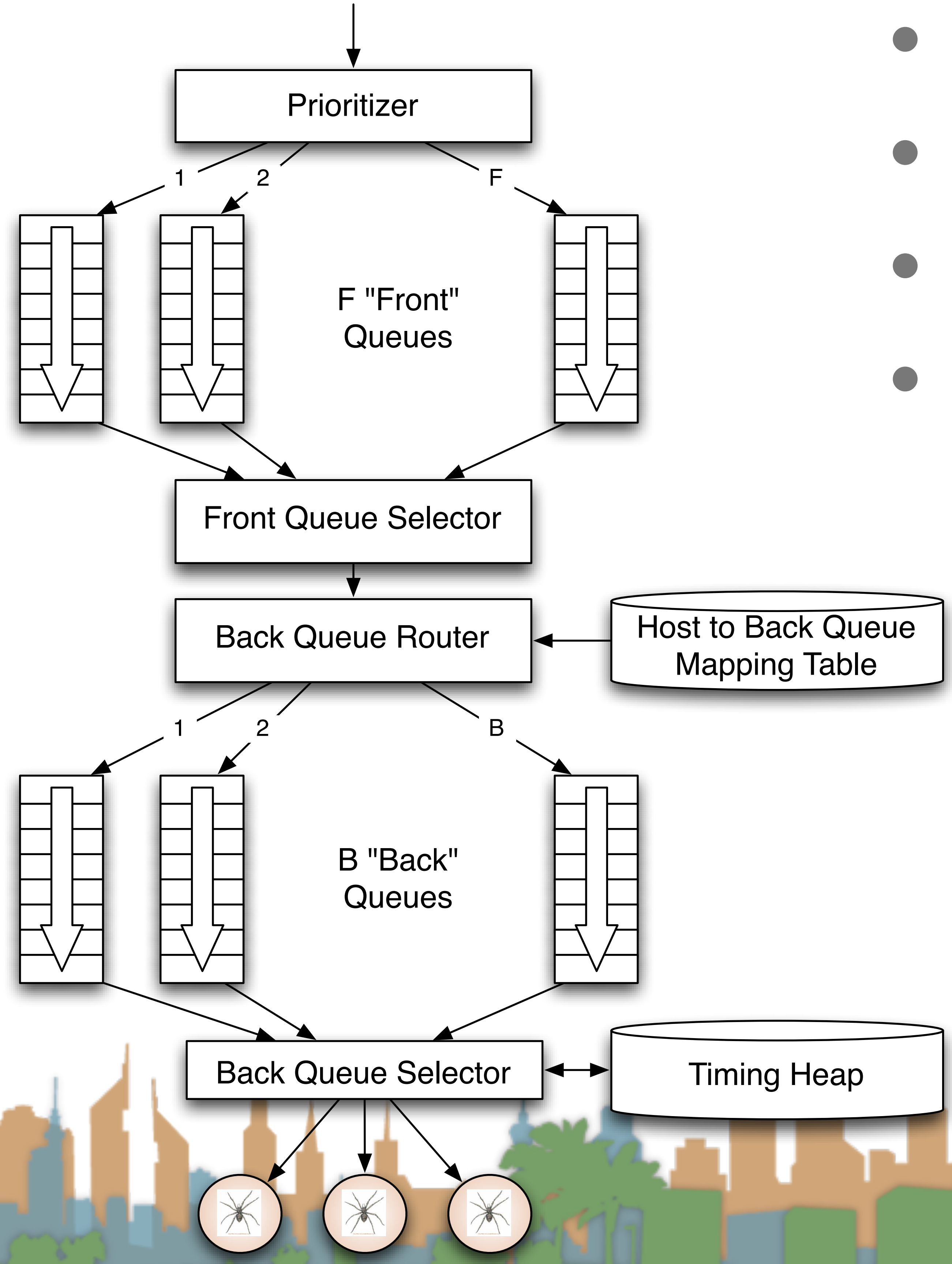
Give me a www.djp3.net

Web Crawlers





URL Frontier Implementation - Mercator

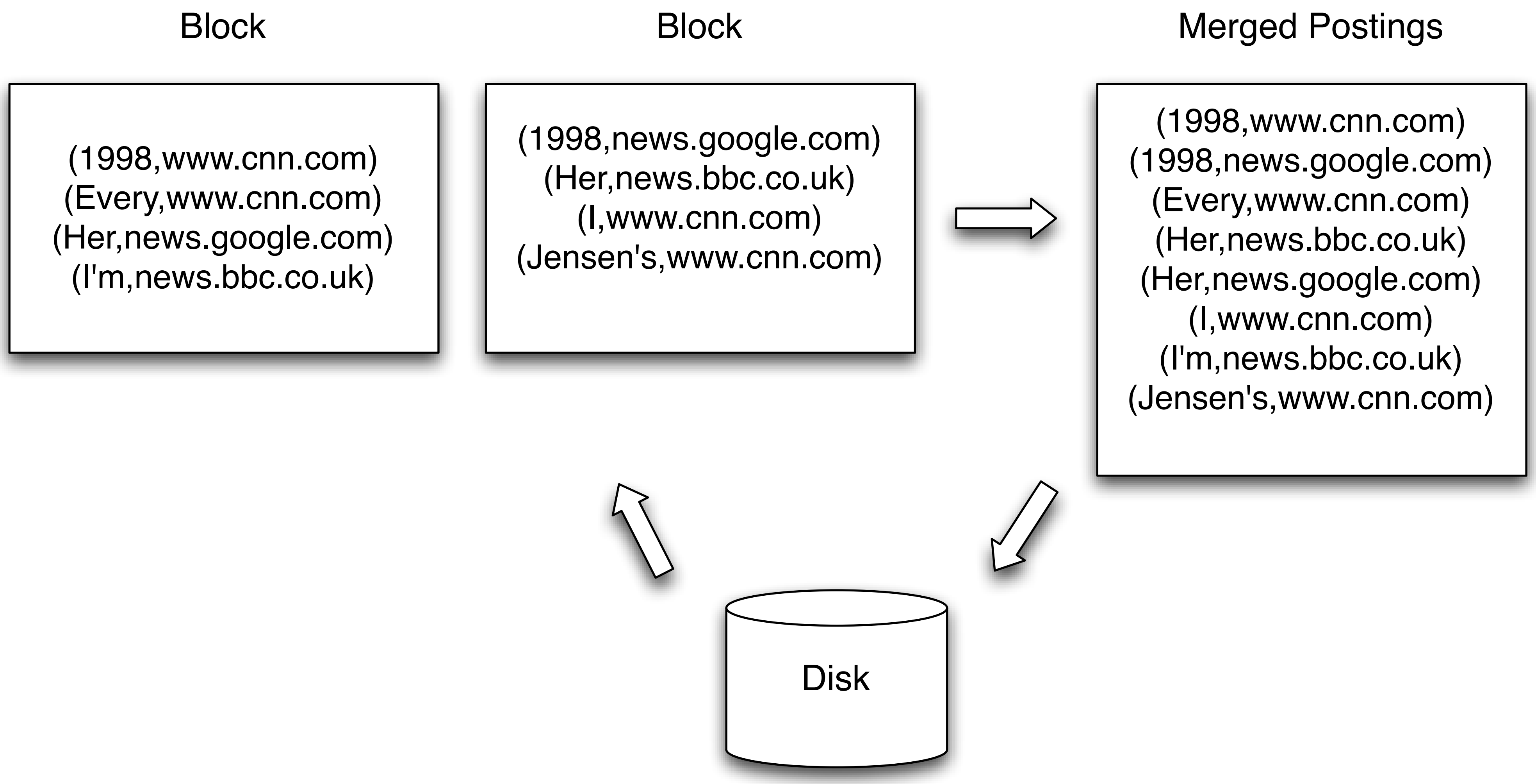


- URLs flow from top to bottom
- Front queues manage priority
- Back queue manage politeness
- Each queue is FIFO

http://research.microsoft.com/~najork/mercator.pdf

BSBI - Block sort-based indexing

Different way to sort index



Block

(1998,www.cnn.com)
(Every,www.cnn.com)
(Her,news.google.com)
(I'm,news.bbc.co.uk)

Block

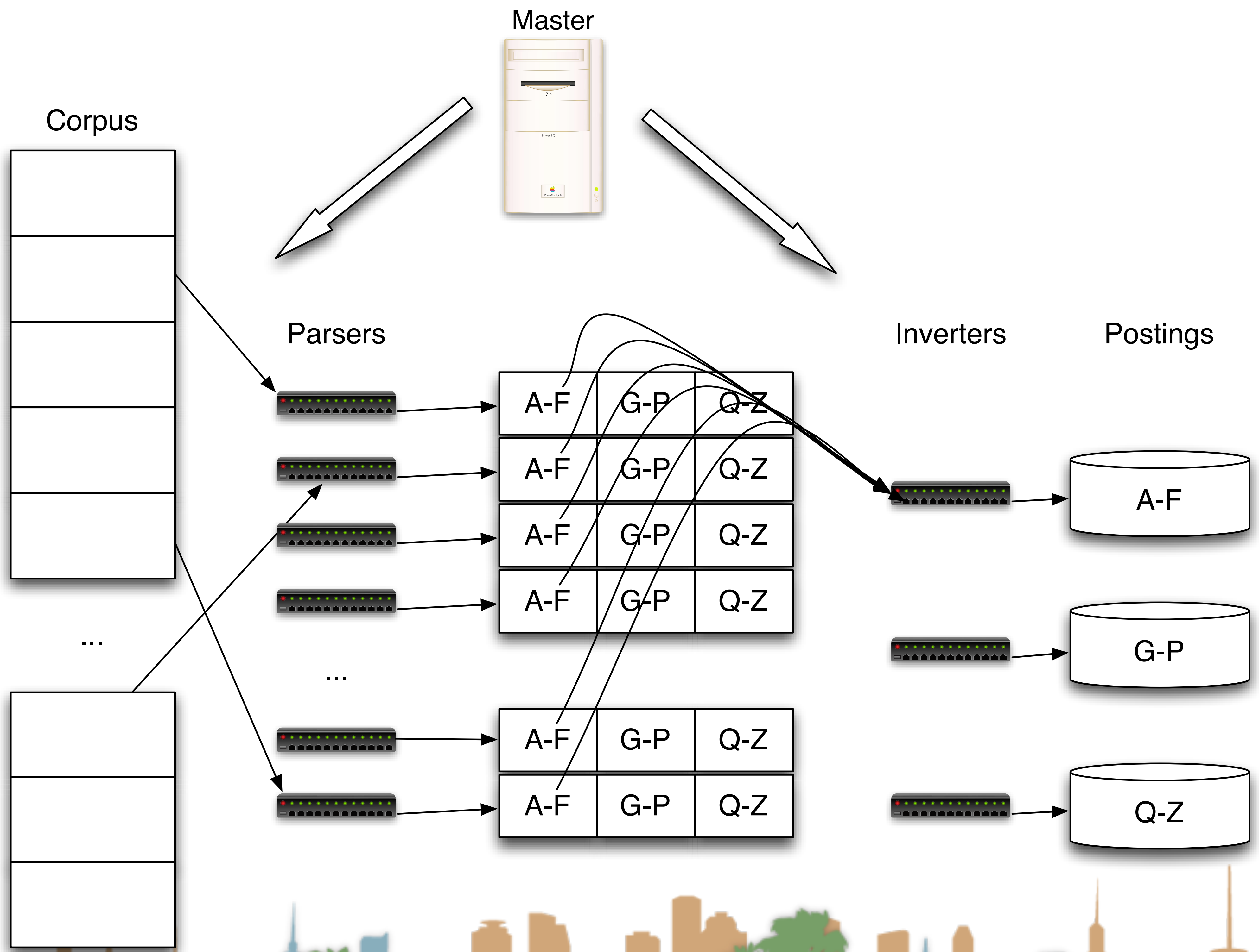
(1998,news.google.com)
(Her,news.bbc.co.uk)
(I,www.cnn.com)
(Jensen's,www.cnn.com)

Merged Postings

(1998,www.cnn.com)
(1998,news.google.com)
(Every,www.cnn.com)
(Her,news.bbc.co.uk)
(Her,news.google.com)
(I,www.cnn.com)
(I'm,news.bbc.co.uk)
(Jensen's,www.cnn.com)

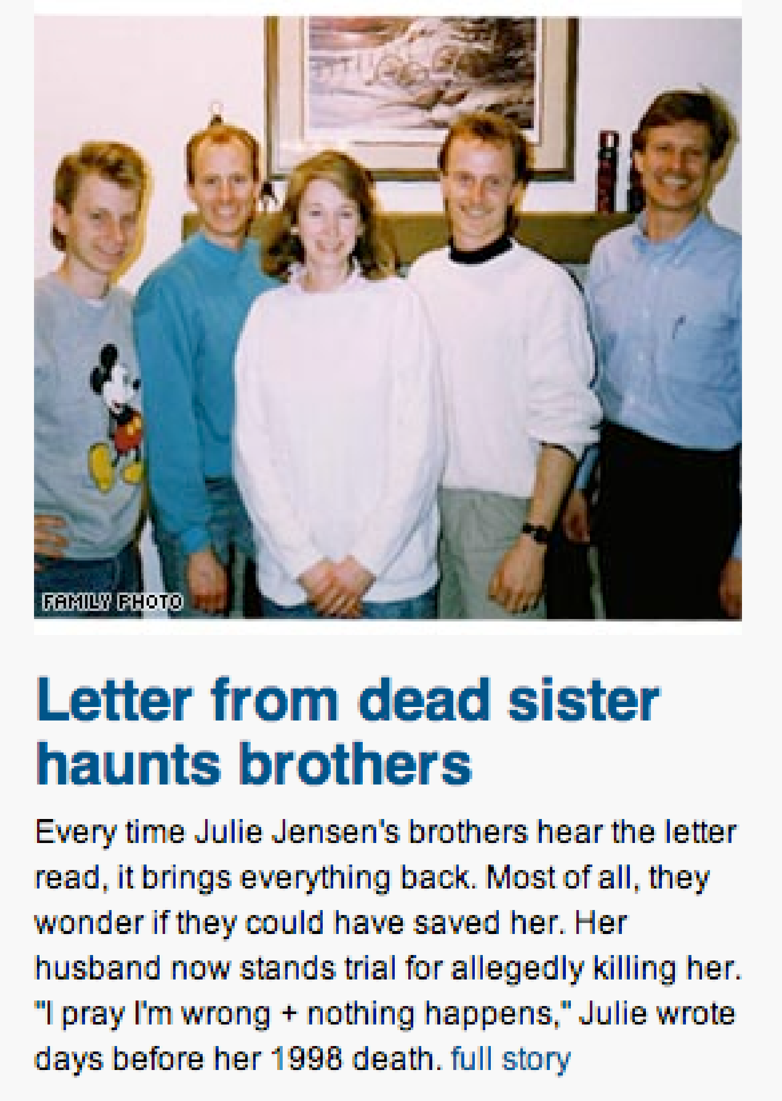
Disk

Distributed Indexing - Architecture



Indices

The index has a list of vector space models



Letter from dead sister haunts brothers

Every time Julie Jensen's brothers hear the letter read, it brings everything back. Most of all, they wonder if they could have saved her. Her husband now stands trial for allegedly killing her. "I pray I'm wrong + nothing happens," Julie wrote days before her 1998 death. full story

1	1998
1	Every
1	Her
1	I
1	I'm
1	Jensen's
2	Julie
1	Letter
1	Most
1	all
1	allegedly
1	back
1	before
1	brings
2	brothers
1	could
1	days
1	dead
1	death
1	everything
1	for
1	from
1	full
1	happens
1	haunts
1	have
1	hear
3	her
1	husband
1	if
1	it
1	killing
1	letter
1	nothing
1	now
1	of
1	pray
1	read,
1	saved
1	sister
1	stands
1	story
1	the
2	they
1	time
1	trial
1	wonder
1	wrong
1	wrote

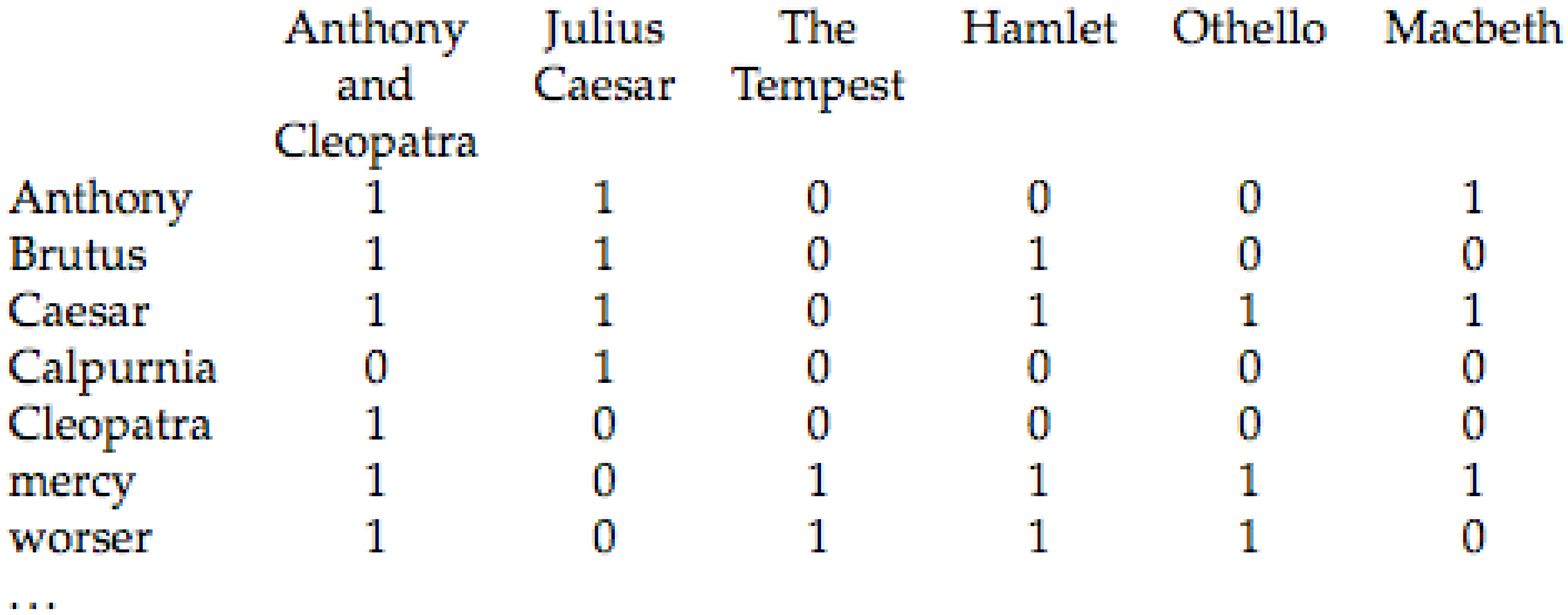
1 1 1 1 1 1 2 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 3 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1

Querying

Our inverted index is a 2-D array or Matrix

A Column For Each Document

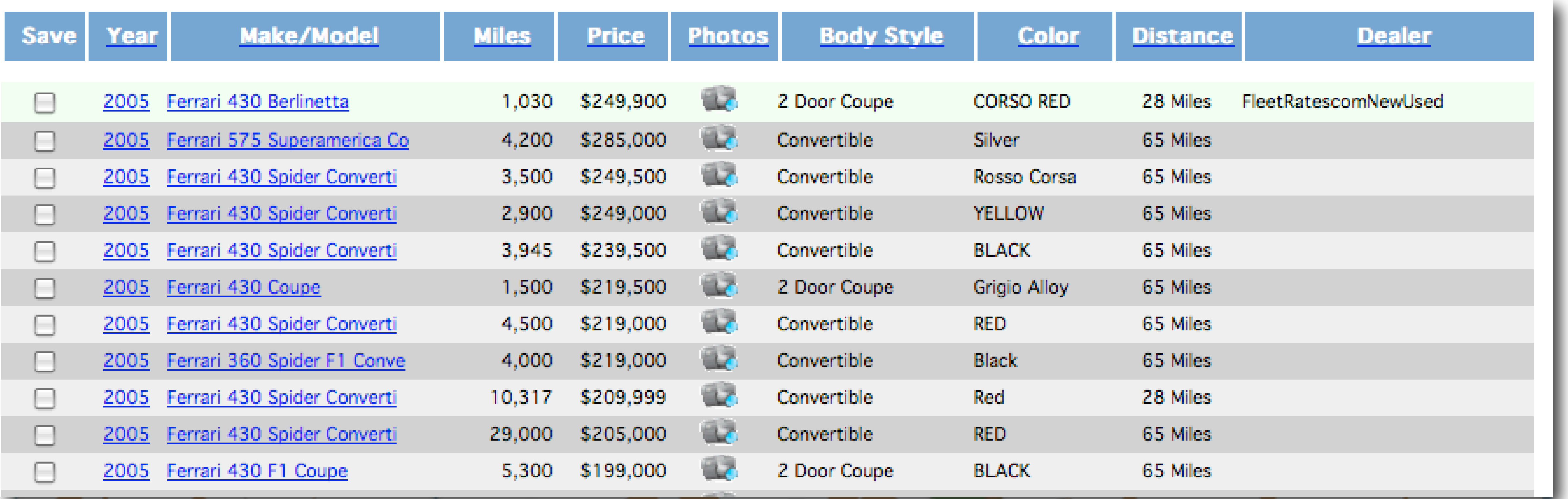
A Row for Each Word (or "Term")



	Anthony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth
Anthony	1	1	0	0	0	1
Brutus	1	1	0	1	0	0
Caesar	1	1	0	1	1	1
Calpurnia	0	1	0	0	0	0
Cleopatra	1	0	0	0	0	0
mercy	1	0	1	1	1	1
worser	1	0	1	1	1	0
...						

Querying

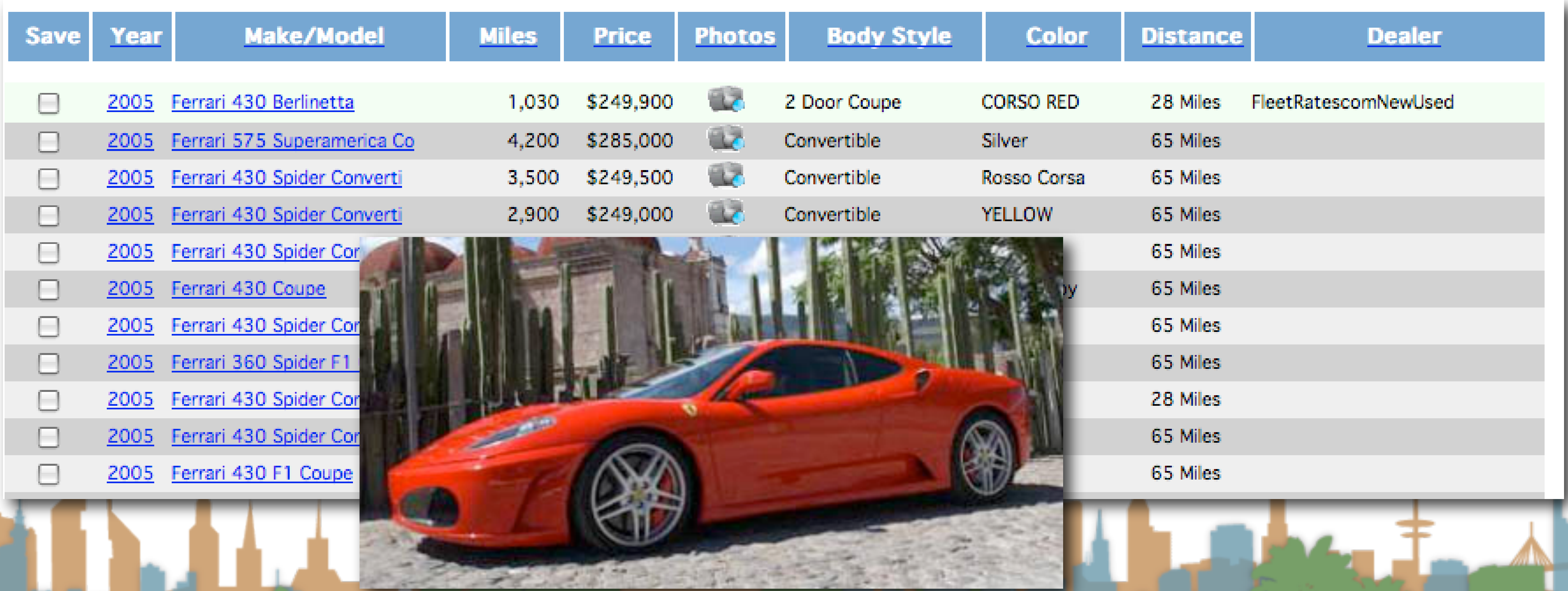
- Parametric Search
  - Example:
    - Result is a large table
    - Columns are fields
    - Searching for "2005" only applied to year field



Save	Year	Make/Model	Miles	Price	Photos	Body Style	Color	Distance	Dealer
□	2005	Ferrari 430 Berlinetta	1,030	$249,900	📷	2 Door Coupe	CORSO RED	28 Miles	FleetRatescomNewUsed
□	2005	Ferrari 575 Superamerica Co	4,200	$285,000	📷	Convertible	Silver	65 Miles	
□	2005	Ferrari 430 Spider Converti	3,500	$249,500	📷	Convertible	Rosso Corsa	65 Miles	
□	2005	Ferrari 430 Spider Converti	2,900	$249,000	📷	Convertible	YELLOW	65 Miles	
□	2005	Ferrari 430 Spider Converti	3,945	$239,500	📷	Convertible	BLACK	65 Miles	
□	2005	Ferrari 430 Coupe	1,500	$219,500	📷	2 Door Coupe	Grigio Alloy	65 Miles	
□	2005	Ferrari 430 Spider Converti	4,500	$219,000	📷	Convertible	RED	65 Miles	
□	2005	Ferrari 360 Spider F1 Conve	4,000	$219,000	📷	Convertible	Black	65 Miles	
□	2005	Ferrari 430 Spider Converti	10,317	$209,999	📷	Convertible	Red	28 Miles	
□	2005	Ferrari 430 Spider Converti	29,000	$205,000	📷	Convertible	RED	65 Miles	
□	2005	Ferrari 430 F1 Coupe	5,300	$199,000	📷	2 Door Coupe	BLACK	65 Miles	

Querying

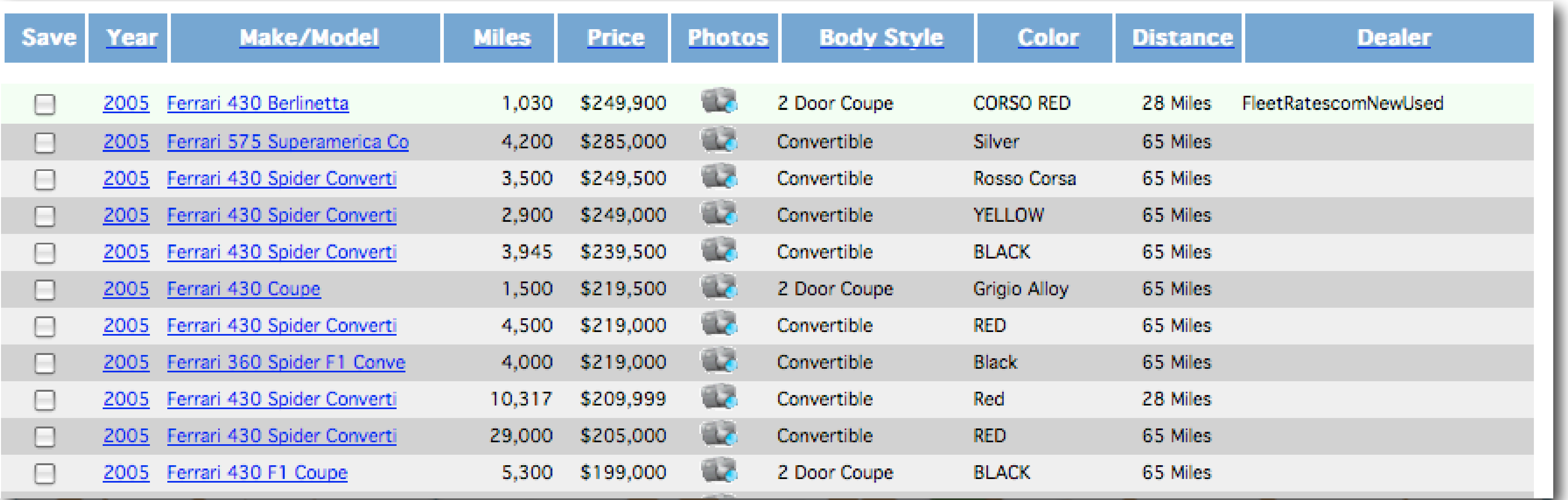
- Parametric Search
  - Example:
    - Result is a large table
    - Columns are fields
    - Searching for "2005" only applied to year field



Save	Year	Make/Model	Miles	Price	Photos	Body Style	Color	Distance	Dealer
□	2005	Ferrari 430 Berlinetta	1,030	$249,900	📷	2 Door Coupe	CORSO RED	28 Miles	FleetRatescomNewUsed
□	2005	Ferrari 575 Superamerica Co	4,200	$285,000	📷	Convertible	Silver	65 Miles	
□	2005	Ferrari 430 Spider Converti	3,500	$249,500	📷	Convertible	Rosso Corsa	65 Miles	
□	2005	Ferrari 430 Spider Converti	2,900	$249,000	📷	Convertible	YELLOW	65 Miles	
□	2005	Ferrari 430 Spider Cor						65 Miles	
□	2005	Ferrari 430 Coupe					by	65 Miles	
□	2005	Ferrari 430 Spider Cor						65 Miles	
□	2005	Ferrari 360 Spider F1						65 Miles	
□	2005	Ferrari 430 Spider Cor						28 Miles	
□	2005	Ferrari 430 Spider Cor						65 Miles	
□	2005	Ferrari 430 F1 Coupe						65 Miles	

Querying

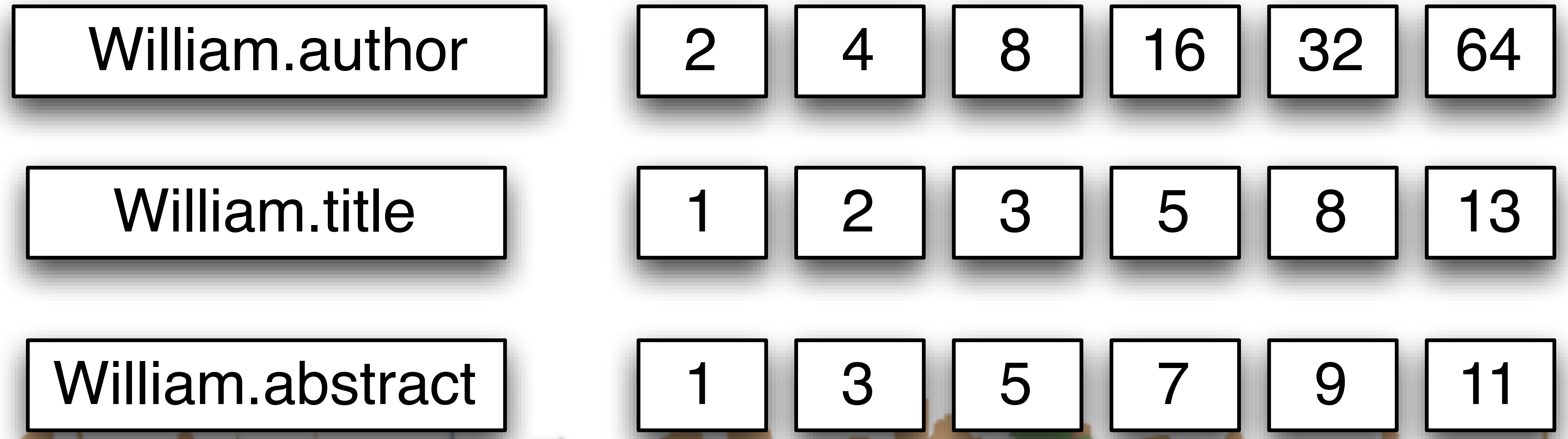
- Parametric Search
  - Example:
    - Result is a large table
    - Columns are fields
    - Searching for “2005” only applied to year field



Save	Year	Make/Model	Miles	Price	Photos	Body Style	Color	Distance	Dealer
☐	2005	Ferrari 430 Berlinetta	1,030	$249,900	📷	2 Door Coupe	CORSO RED	28 Miles	FleetRatescomNewUsed
☐	2005	Ferrari 575 Superamerica Co	4,200	$285,000	📷	Convertible	Silver	65 Miles	
☐	2005	Ferrari 430 Spider Converti	3,500	$249,500	📷	Convertible	Rosso Corsa	65 Miles	
☐	2005	Ferrari 430 Spider Converti	2,900	$249,000	📷	Convertible	YELLOW	65 Miles	
☐	2005	Ferrari 430 Spider Converti	3,945	$239,500	📷	Convertible	BLACK	65 Miles	
☐	2005	Ferrari 430 Coupe	1,500	$219,500	📷	2 Door Coupe	Grigio Alloy	65 Miles	
☐	2005	Ferrari 430 Spider Converti	4,500	$219,000	📷	Convertible	RED	65 Miles	
☐	2005	Ferrari 360 Spider F1 Conve	4,000	$219,000	📷	Convertible	Black	65 Miles	
☐	2005	Ferrari 430 Spider Converti	10,317	$209,999	📷	Convertible	Red	28 Miles	
☐	2005	Ferrari 430 Spider Converti	29,000	$205,000	📷	Convertible	RED	65 Miles	
☐	2005	Ferrari 430 F1 Coupe	5,300	$199,000	📷	2 Door Coupe	BLACK	65 Miles	

Parametric Search

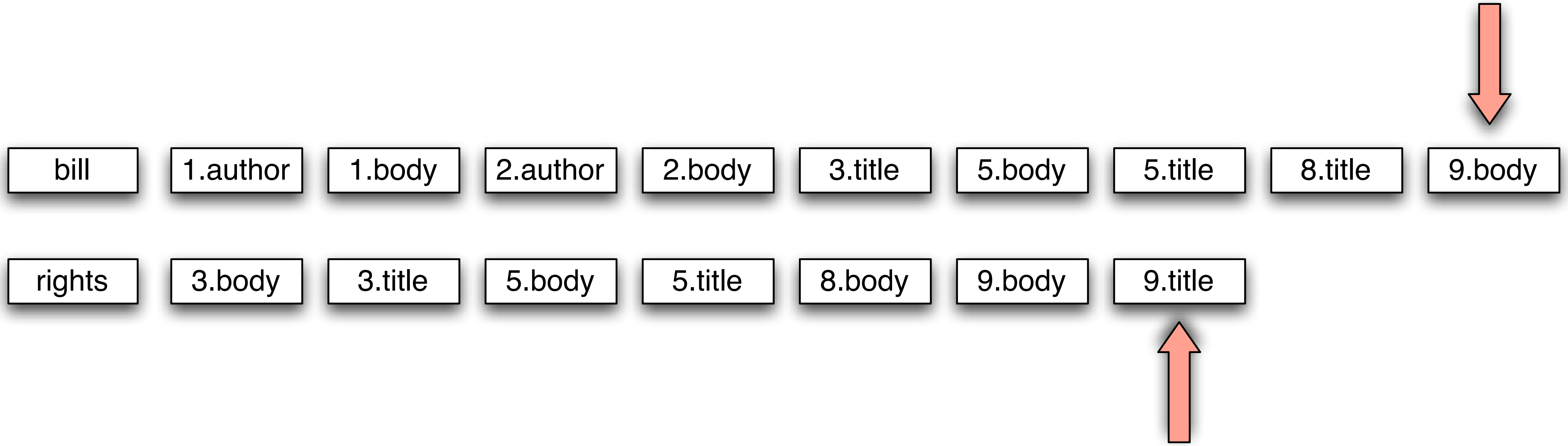
- Now, we crawl the corpus
- We parse the document keeping track of terms, fields and docIDs
- Instead of building just a (term, docID) pair
- We build (term, field, docID) triples
- These can then be combined into postings like this:



William.author	2	4	8	16	32	64
William.title	1	2	3	5	8	13
William.abstract	1	3	5	7	9	11

Zone scoring with zones combination index

"bill OR rights" (0.1 author), (0.3 body), (0.6 title)



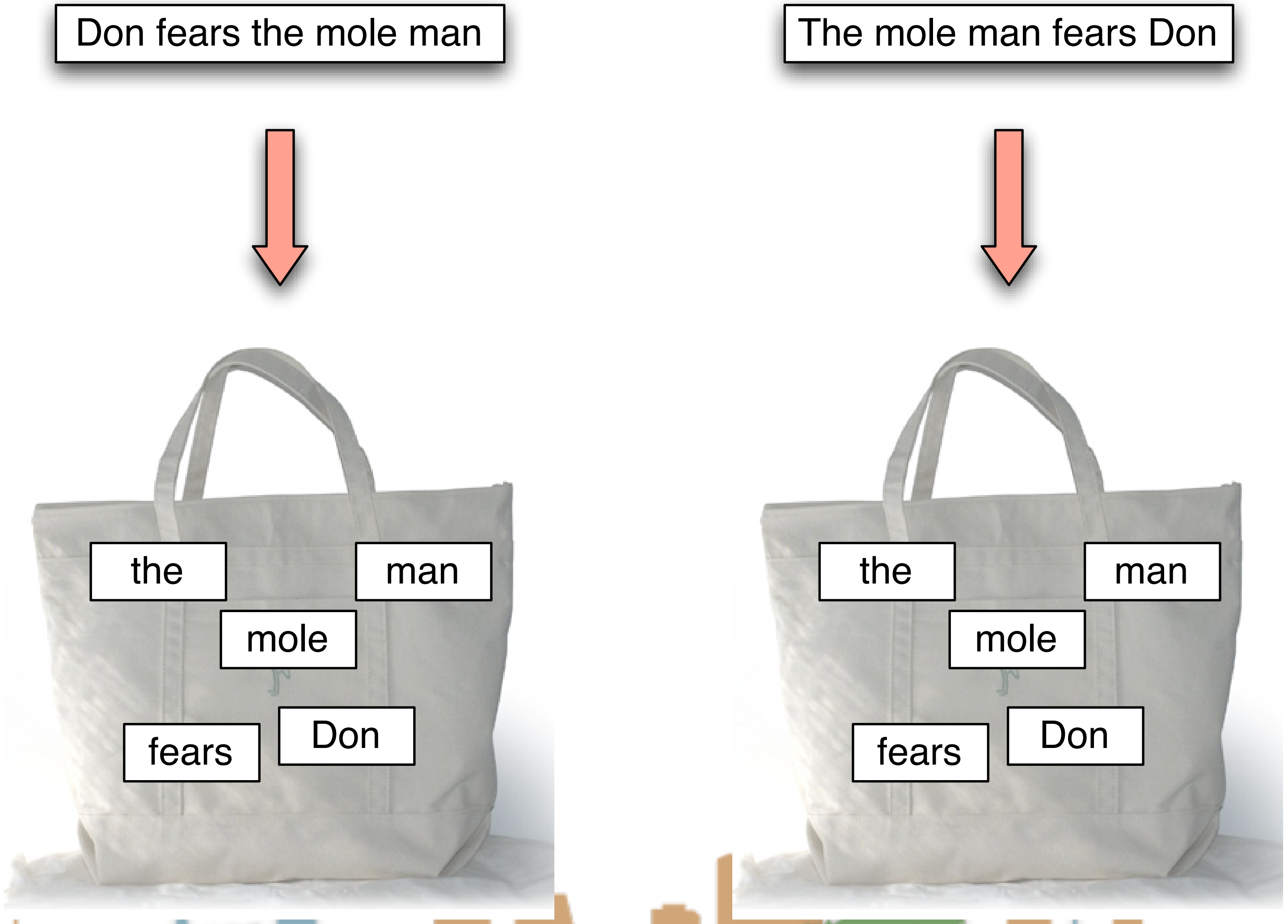
bill → 1.author → 1.body → 2.author → 2.body → 3.title → 5.body → 5.title → 8.title → 9.body

rights → 3.body → 3.title → 5.body → 5.title → 8.body → 9.body → 9.title

1: 0.4   5: 0.9
2: 0.4   8: 0.9
3: 0.9   9: 0.9

Bag of Words Model

- “Don fears the mole man” equals “The mole man fears Don”
- The incidence matrix for both looks the same

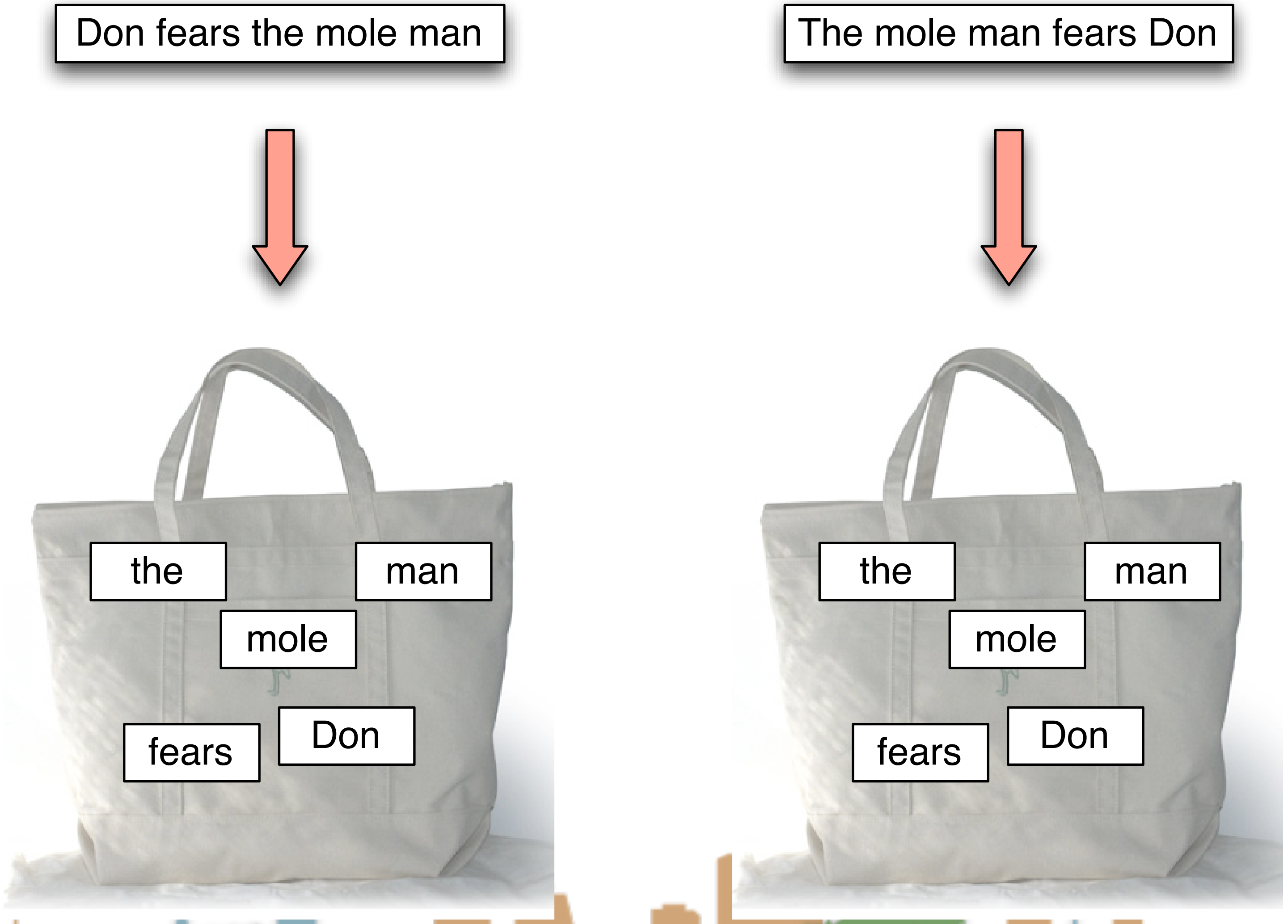


Don fears the mole man

The mole man fears Don

Bag of Words Model

- “Don fears the mole man” equals “The mole man fears Don”
- The incidence matrix for both looks the same



Don fears the mole man

the

man

mole

fears

Don

The mole man fears Don

the

man

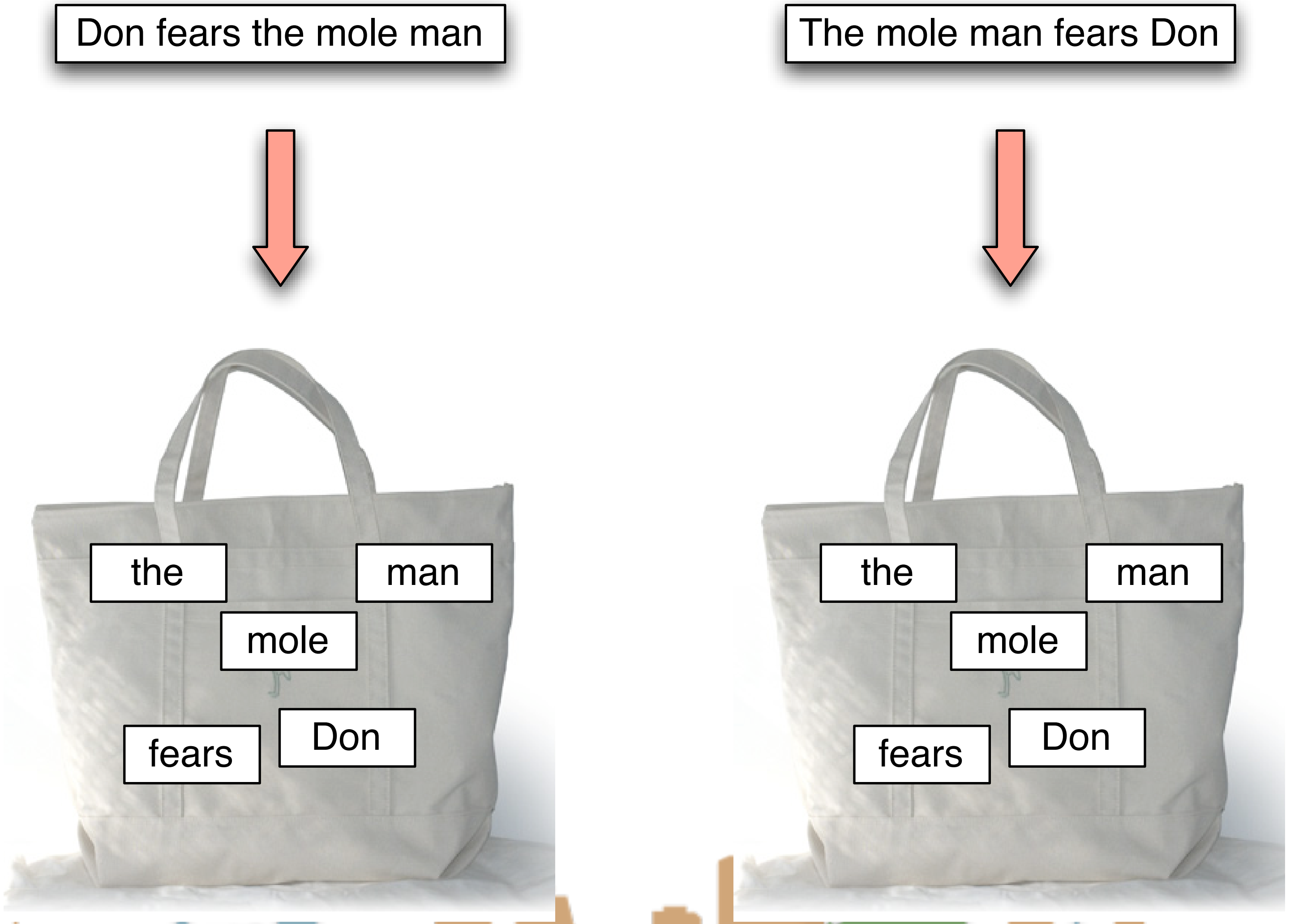
mole

fears

Don

Bag of Words Model

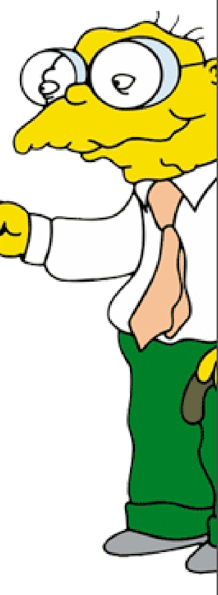
- “Don fears the mole man” equals “The mole man fears Don”
- The incidence matrix for both looks the same



Don fears the mole man

The mole man fears Don

	d_1	d_2
Don	1	1
fears	1	1
man	1	1
mole	1	1
mule	0	0
the	1	1
zoo	0	0



Weighting Term Frequency - WTF

$$Score_{WTF}(q,d) = \sum_{t \in q}(WTF(t,d))$$

Weighting Term Frequency - WTF

$$Score_{WTF}(q,d) = \sum_{t \in q} (WTF(t,d))$$

$$\begin{aligned}
Score_{WTF}(\text{"}bill\ rights\text{"}, declarationOfIndependence) &= \\
WTF(\text{"}bill\text{"}, declarationOfIndependence) &+ \\
WTF(\text{"}rights\text{"}, declarationOfIndependence) &= \\
0 + 1 + log(3) &= \quad 1.48
\end{aligned}$$

Weighting Term Frequency - WTF

$$Score_{WTF}(q,d) = \sum_{t \in q} (WTF(t,d))$$

$$\begin{aligned}
Score_{WTF}(\text{"}bill\ rights\text{"}, declarationOfIndependence) &= \\
WTF(\text{"}bill\text{"}, declarationOfIndependence) &+ \\
WTF(\text{"}rights\text{"}, declarationOfIndependence) &= \\
0 + 1 + log(3) &= \quad 1.48
\end{aligned}$$

$$\begin{aligned}
Score_{WTF}(\text{"}bill\ rights\text{"}, constitution) &= \\
WTF(\text{"}bill\text{"}, constitution) &+ \\
WTF(\text{"}rights\text{"}, constitution) &= \\
1 + log(10) + 1 + log(1) &= \quad 3
\end{aligned}$$

Vector Space Model

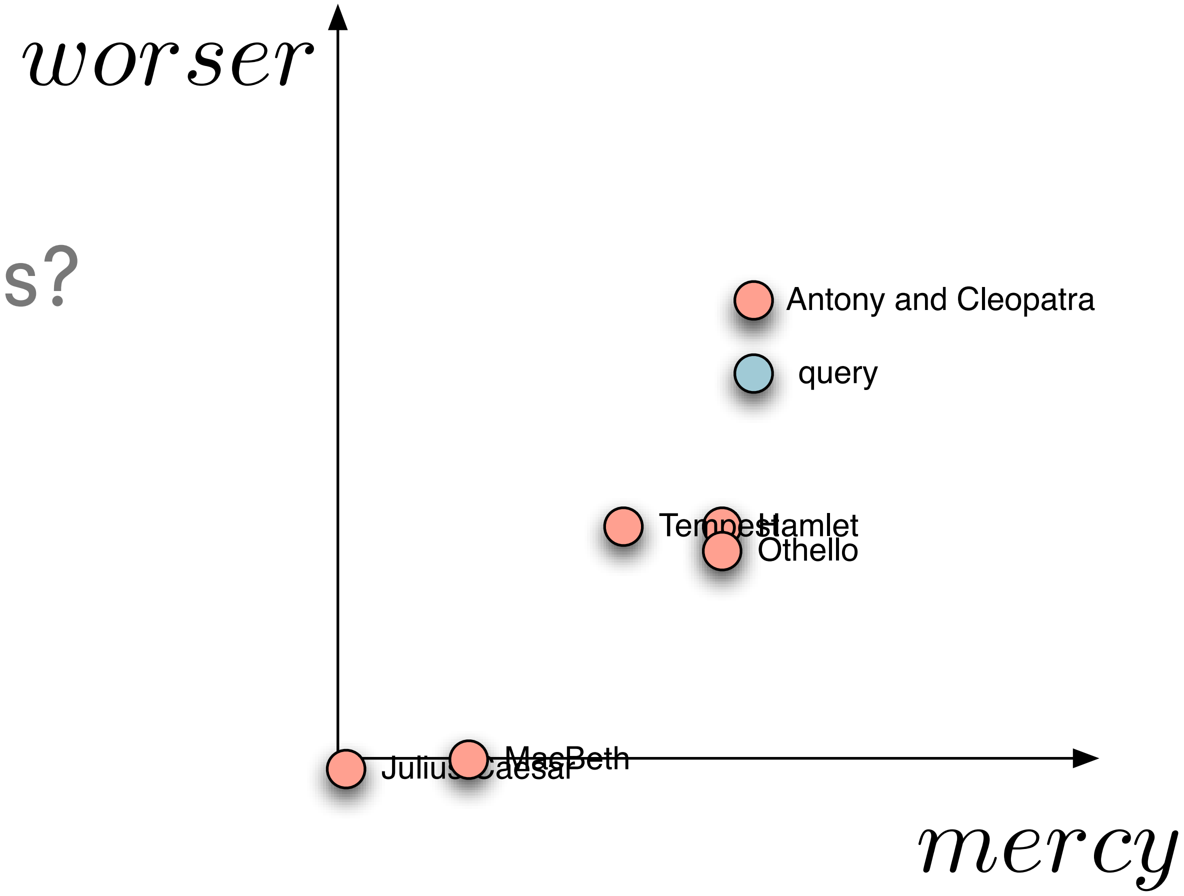
- Recall our Shakespeare Example:

	$\vec{V}(d_1)$	$\vec{V}(d_2)$				$\vec{V}(d_6)$
	Antony and Cleopatra	*Julius Caesar*	*The Tempest*	*Hamlet*	*Othello*	*Macbeth*
Antony	13.1	11.4	0.0	0.0	0.0	0.0
Brutus	3.0	8.3	0.0	1.0	0.0	0.0
Caesar	2.3	2.3	0.0	0.5	0.3	0.3
Calpurnia	0.0	11.2	0.0	0.0	0.0	0.0
Cleopatra	17.7	0.0	0.0	0.0	0.0	0.0
mercy	0.5	0.0	0.7	0.9	0.9	0.3
worser	1.2	0.0	0.6	0.6	0.6	0.0

Vector Space Scoring

Query as a vector

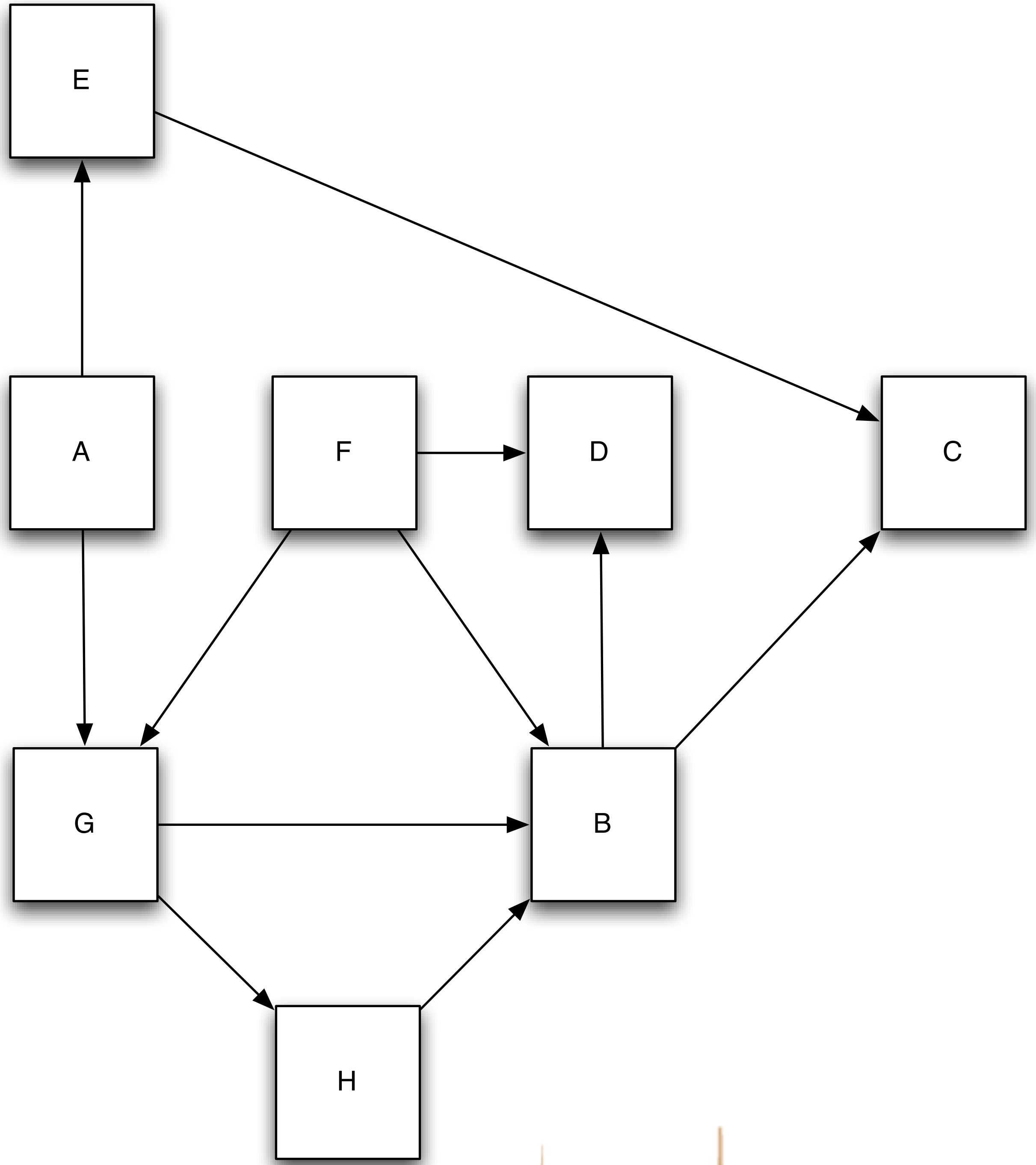
- So a query can also be plotted in the same space
  - “worser mercy”
  - To score, we ask:
    - How similar are two points?
  - How to answer?



Markov Chains

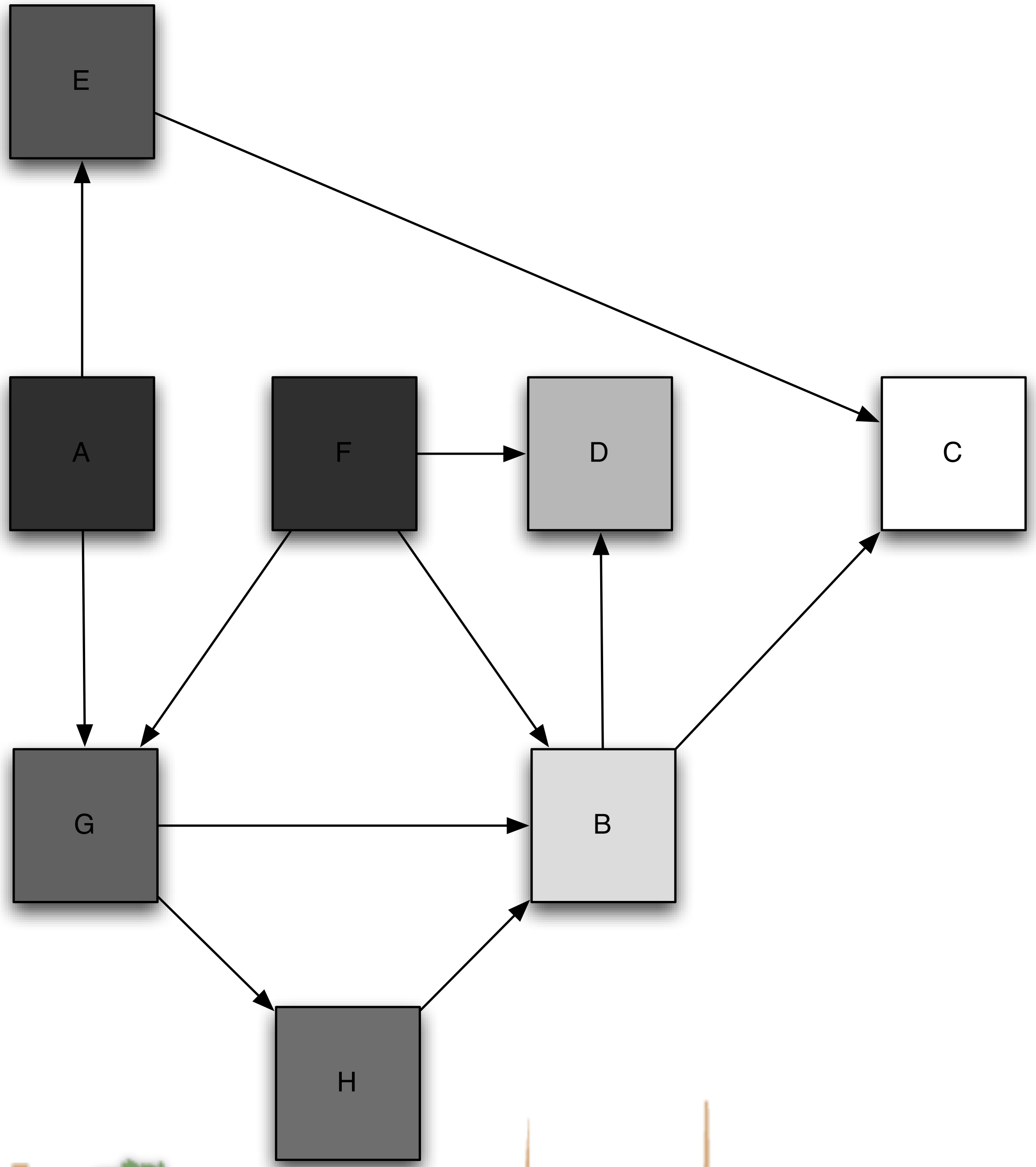
- Example:
  - 8 states
    - (web pages or whatever)
  - 8 by 8 transition prob. matrix

	A	*B*	*C*	*D*	*E*	*F*	*G*	*H*
A	0	0	0	0	0.5	0	0.5	0
B	0	0	0.5	0.5	0	0	0	0
C	0	0	0	0	0	0	0	0
D	0	0	0	0	0	0	0	0
E	0	0	1.0	0	0	0	0	0
F	0	0.33	0	0.33	0	0	0.33	0
G	0	0.5	0	0	0	0	0	0.5
H	0	1.0	0	0	0	0	0	0



Long-Term visit rate

- A: 5%
- B: 21%
- C: 23%
- D: 18%
- E: 8%
- F: 5%
- G: 9%
- H: 10%

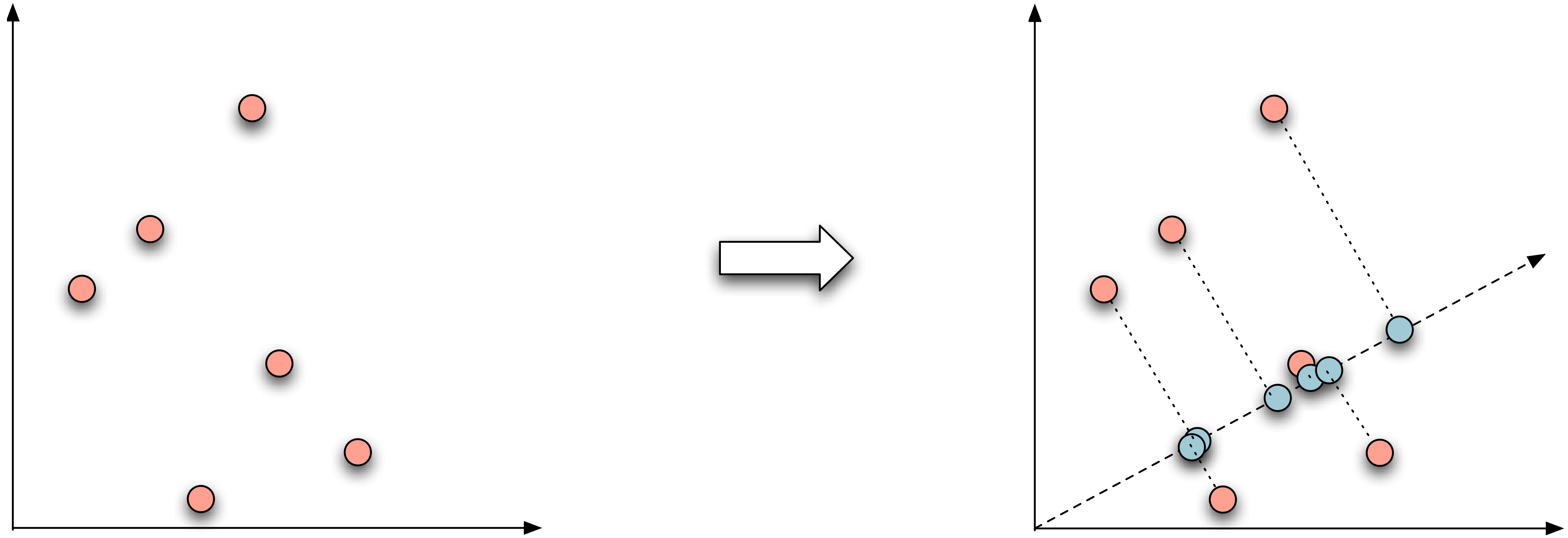


Latent Semantic Indexing - Introduction

Star Cluster NGC 290 - ESA & NASA

Mathematically speaking

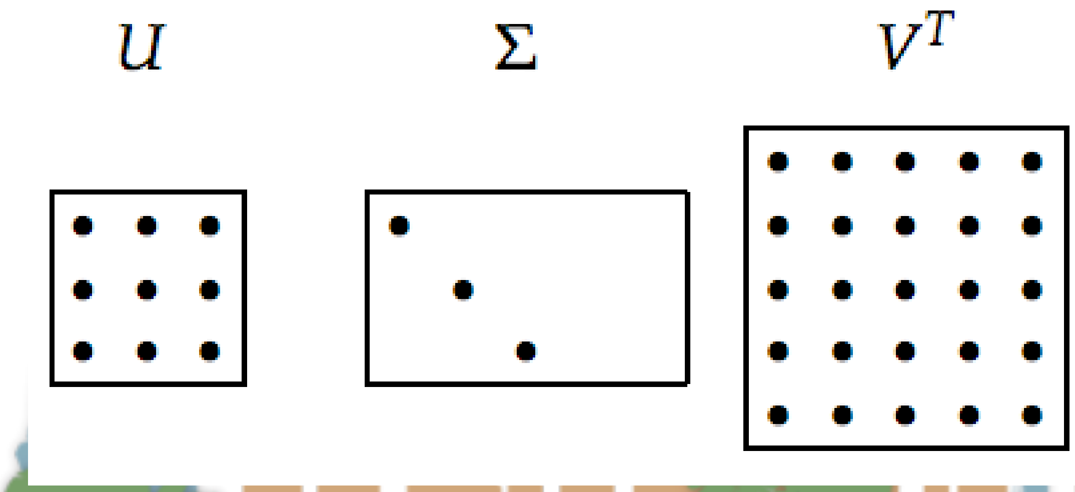
- Latent Semantic Indexing can project on an arbitrary axis, not just a principal axis



Latent Semantic Indexing - Linear Algebra Refresher

Matrix Decomposition

- Singular Value Decomposition
  - SVD enables lossy compression of your term-document matrix
    - reduces the dimensionality or the rank
    - you can arbitrarily reduce the dimensionality by putting zeros in the bottom right of sigma
    - this is a mathematically optimal way of reducing dimensions

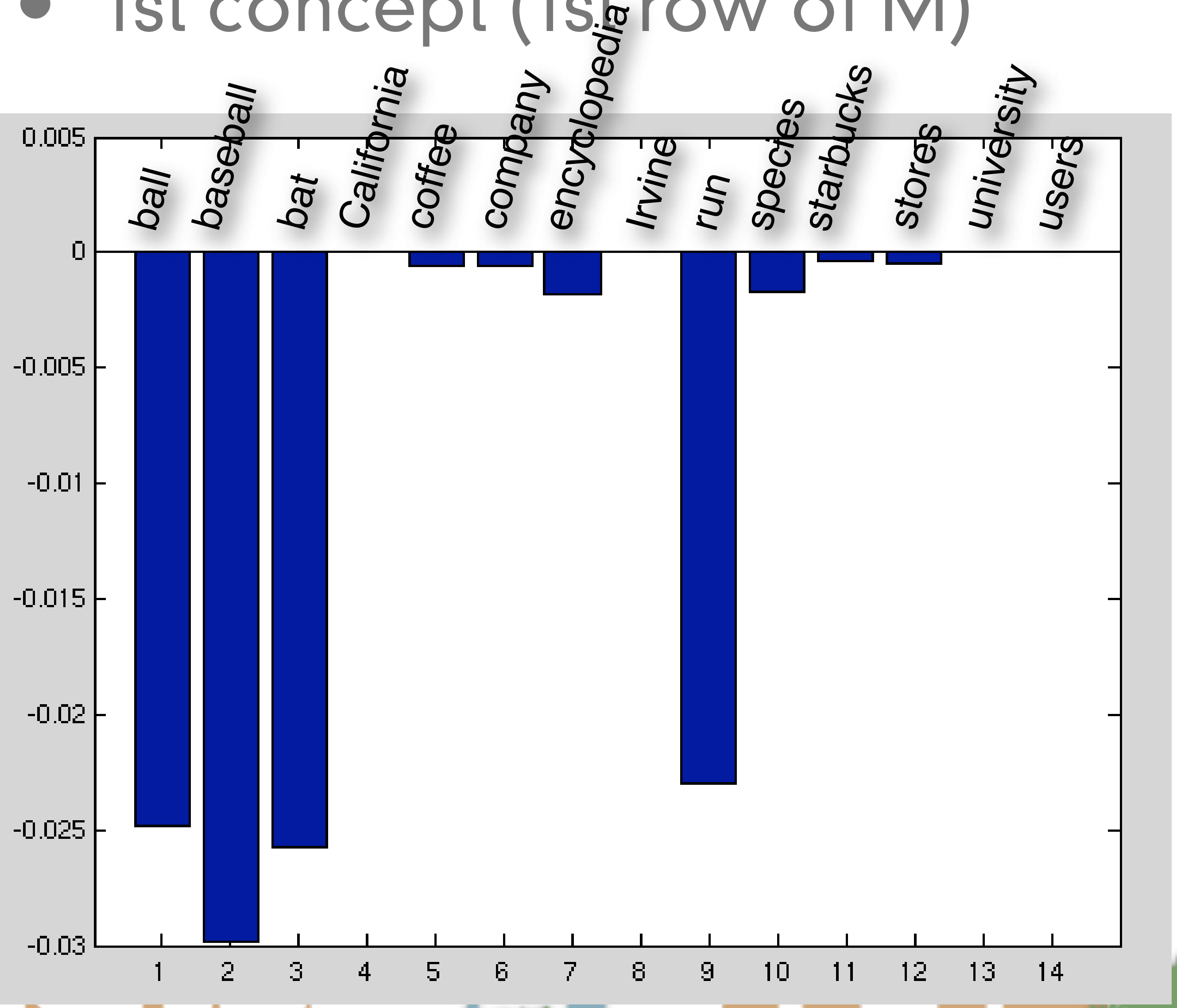


Using MATLAB For LSA

First concept is selecting for wikipedia?

Demo

- Demonstrate what SVD is capturing
  - 1st concept (1st row of M)



First concept is selecting for baseball?

Finally .... I promised it would be hard

- 19 Lectures
- 7 Discussions
- 4 quizzes - 8 chapters - 6 (+2) papers
- 7 assignments
  - Built (building) a web search engine from scratch
  - Used cutting edge architecture (hadoop)
- 2 web pages - a trip to Google
- Hopefully had fun, were challenged and learned something...
  - you can sleep when you are dead, until then coffee.

Congratulations!

LUCI