

# Chapter 1:

# Basic Concepts

Instructor: Dr. Mehmet S. Aktaş

---

# Road map

- World Wide Web
- Internet
- Search Engines
- Web Data Mining
  - Why? What? Where?
- Summary

---

# World Wide Web - WWW

- The Web
  - allows users of one computer to access information stored on another through a world-wide network called the Internet.
- Client-server based implementation
  - Ex: local machine browser, remote machine: Web servers
- Hypertext document based semantic structure
  - Ted Nelson in 1965
  - Hypertext, hyperlinks

---

# World Wide Web - History

## ■ The Web

- ❑ Invented by Tim Berners-Lee at CERN in 1989 .
- ❑ Distributed hypertext system – basic architecture of the Web.
- ❑ Hypertext Transfer Protocol (HTTP), Hypertext Markup Language (HTML), Universal Resource Locator (URL)

## ■ Browsers

- ❑ Mosaic, 1993, M. Andreessen, Univ. of Illinois
- ❑ Netspace, 1994, M. Andreessen and Jim Clark, Netscape Comm.
- ❑ Internet Explorer, 1995, B. Gates, Microsoft

# Internet - History

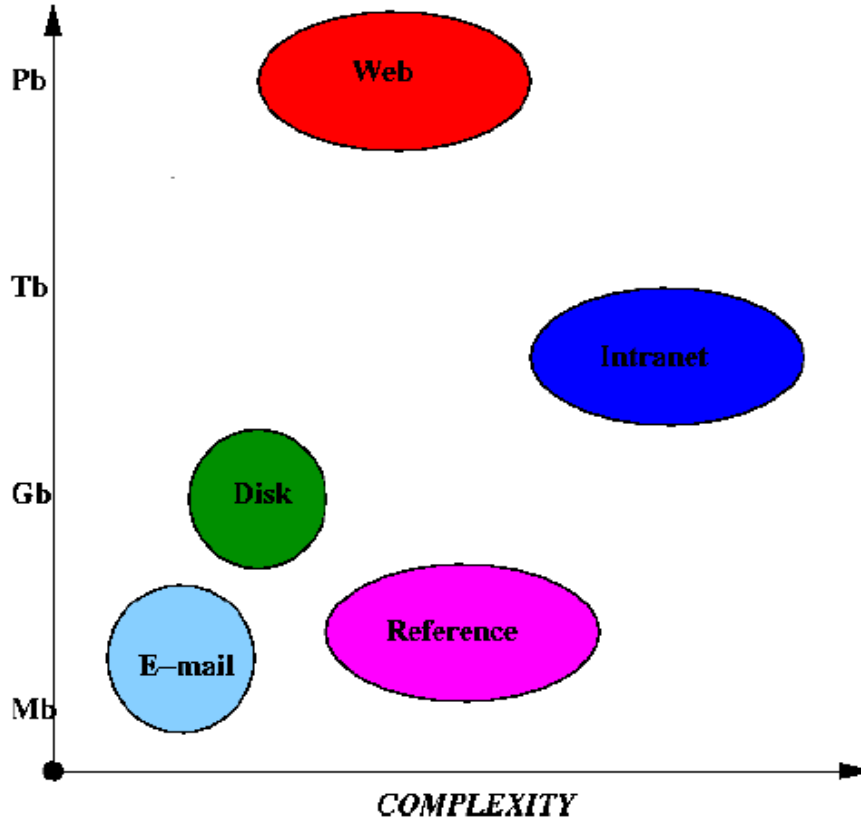
## ■ ARPANET - Internet

- Advanced Research Projects Agency (ARPA) Network, 1969-1972.
  - Connected computers from 40 different locations.
- TCP/IP (Transmission Control Protocol/Internet Protocol), 1973, Vinton Cerf, Bob Kahn.
- Internet
  - 1982 TCP/IP was widely adopted.
  - Networks were connected using TCP/IP Protocol and formed inter-net.

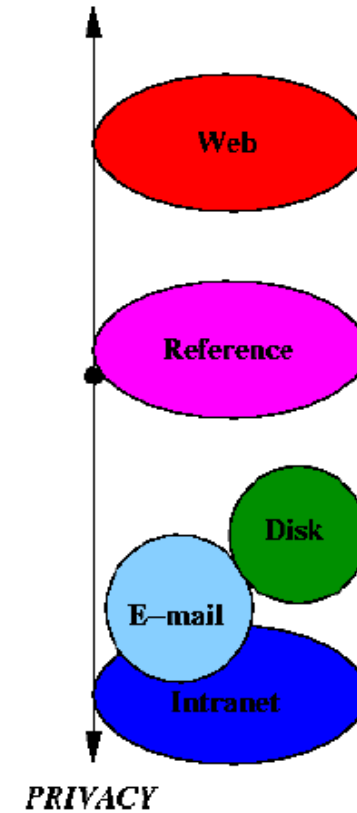


# Different Views on Data

*VOLUME*



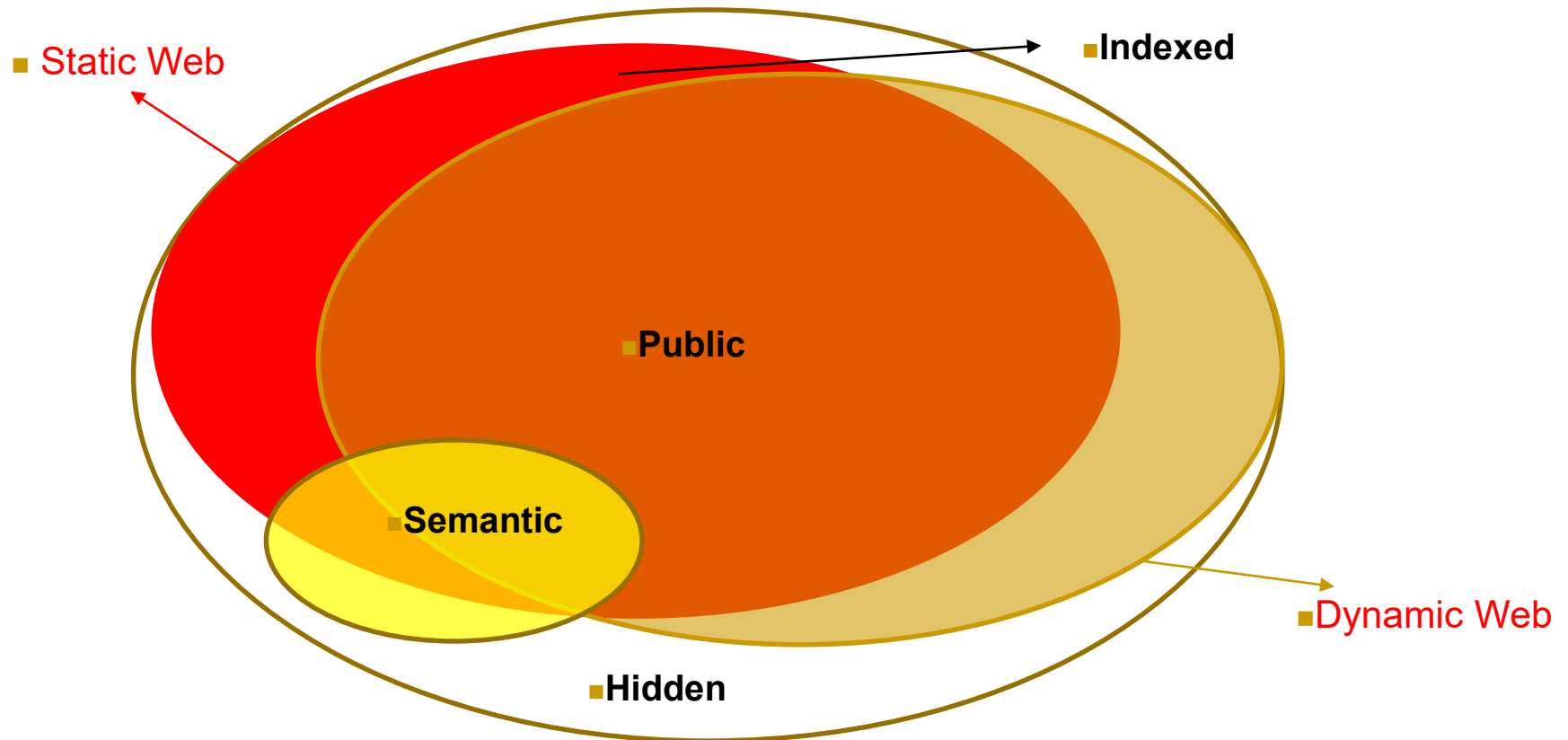
*ADVERSARIAL*



■ <http://ijcai-11.iiia.csic.es/files/proceedings/T16-Web-Mining.pdf>



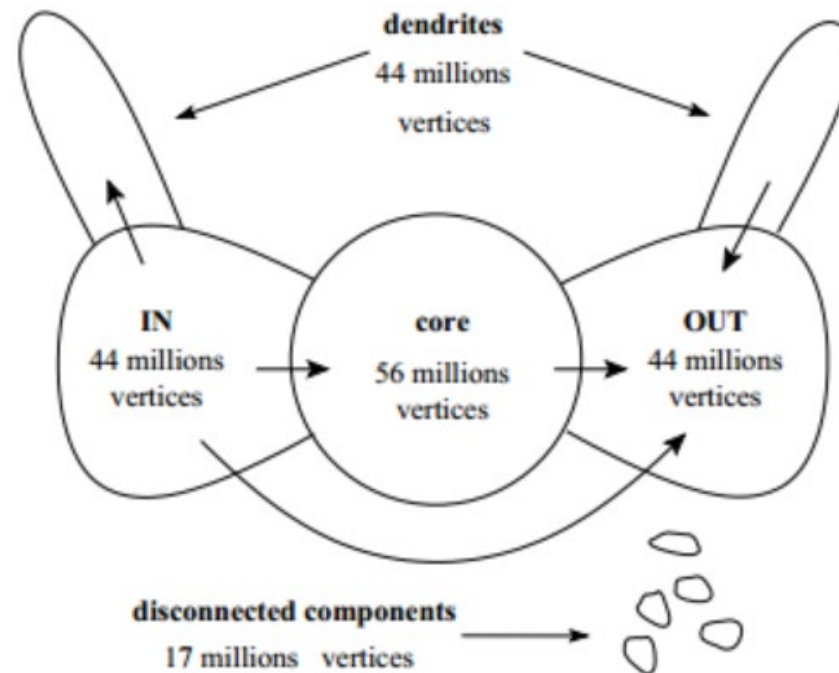
# The Different Facets of the Web



■ <http://ijcai-11.iia.csic.es/files/proceedings/T16-Web-Mining.pdf>

■ <http://hal.archives-ouvertes.fr/docs/00/05/44/58/PDF/webgraph.pdf>

Jean-Loup Guillaume and Matthieu Latapy



**Fig. 3:** The bow-tie macroscopic structure of the Web graph [BKM<sup>+</sup>00]: the core, the IN component, the OUT component and the dendrites. Each of these parts contains around one quarter of the pages, the disconnected part being reduced to less than 10% of the whole.





## What for?

---

- The Web as an object
- User-driven Web design
- Improving Web applications
- Social mining
- .....



## The Wisdom of Crowds

---

- James Surowiecki, a *New Yorker* columnist, published this book in 2004
  - “Under the **right** circumstances, groups are remarkably intelligent”
- Importance of diversity, independence and decentralization **Aggregating data**

*“large groups of people are smarter than an elite few, no matter how brilliant—they are better at solving problems, fostering innovation, coming to wise decisions, even predicting the future”.*

## Tags / jaguar / clusters

SEARCH

(Or, try an [advanced search](#).)



[car](#), [cars](#), [auto](#), [etype](#), [automobile](#), [classic](#),  
[vintage](#), [autoshow](#), [red](#), [show](#)

➔ [See more in this cluster...](#)



[zoo](#), [animal](#), [cat](#), [animals](#), [bigcat](#), [seattle](#),  
[woodlandparkzoo](#), [sleep](#), [edinburgh](#), [caged](#)

➔ [See more in this cluster...](#)



[guitar](#), [fender](#)

➔ [See more in this cluster...](#)



[aircraft](#), [raf](#)

➔ [See more in this cluster...](#)

These are the *most recent* photos tagged with **jaguar**. [See more...](#)



# Flickr: Geo-tagged pictures

flickr®

[Inicio](#)

[La visita](#)

[Crear cuenta](#)

[Explorar](#)

No has iniciado sesión

[Iniciar sesión](#)

[Ayuda](#)

Buscar un lugar

[Buscar](#)

**232128 elementos con geotiquetas**

Ordenar por: [Interesante](#) • [Recientes](#)

[Buscar en el mapa](#)





# The Wisdom of Crowds

---

- Popularity
- Diversity
- Quality
- Coverage



**Long tail**



## The Long Tail

---

Explore Flickr through tags

architecture **art** australia **beach** birthday blue bw **california** canada  
**canon** china christmas **city** concert england europe **family** festival flower  
flowers food france friends fun germany green italy japan london  
music **nature** new newyork night **nikon** nyc paris park **party**  
people portrait red sanfrancisco sky snow spain street **summer** sunset taiwan  
**travel** trip uk usa vacation water **wedding** white winter

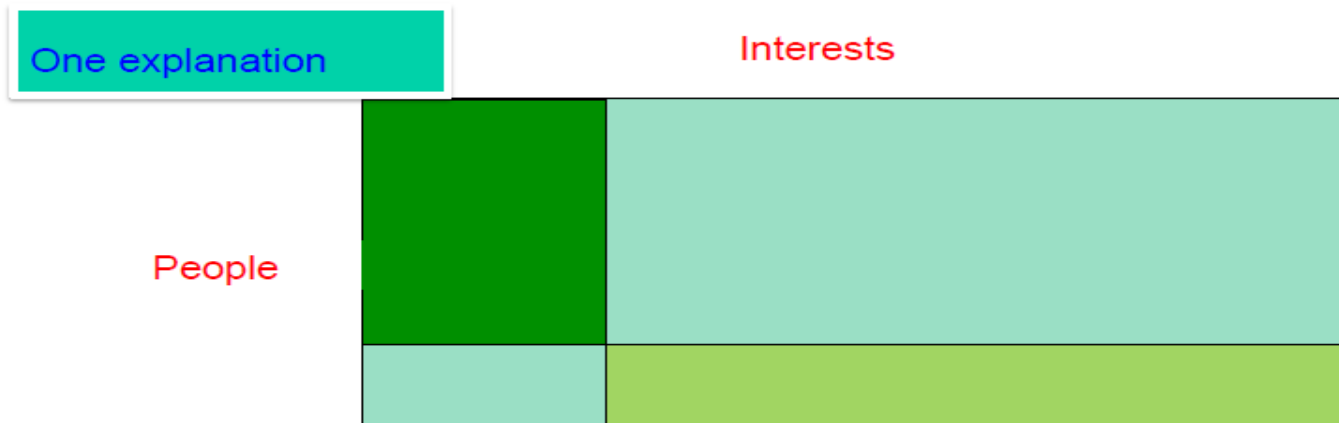
■ <http://ijcai-11.iiia.csic.es/files/proceedings/T16-Web-Mining.pdf>

---



## Heavy tail of user interests

- **Many queries, each asked very few times, make up a large fraction of all queries**
  - Movies watched, blogs read, words used ...





## The Wisdom of Crowds

---

- Crucial for Search Ranking
- Text: Web Writers & Editors
  - not only for the Web!
- Links: Web Publishers
- Tags: Web Taggers
- Queries: All Web Users!
  - Queries and actions (or no action!)



# Search Engines

- Excite search system
  - 1993, started by six Stanford Univ. students
- Yahoo
  - 1994, started out as a listing of favorite web sites
  - 2004, provided general search capability with Inktomi
- Many Search Engines
  - 1994 - present, Lycos, Infoseek, Alta Vista, Ask Jeeves etc.
- Google
  - 1998 by Sergey Brin and Larry Page at Stanford Univ.
- MSN
  - 2005 by Microsoft, which used others search engines before

---

# Characteristics of the Web

- The Web has huge amount of available data
- The Web has all types of data
- The Web has heterogeneous information
- The Web has a hyperlink structure
- The Web has noisy information
- The Web is about services
- The Web is very dynamic
- The Web is a virtual society

---

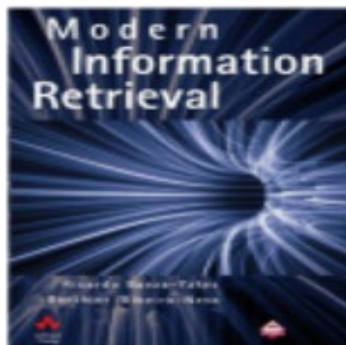
# Web Data Mining

## ■ Web Mining

- aims to discover useful information from the Web hyperlink structure, page content and usage data.
- Web structure mining
  - Discovers knowledge from hyperlinks
- Web content mining
  - Mines useful information from Web page contents
- Web usage mining
  - Mines useful information from Web usage logs



## What is Information Retrieval (IR)?



IR: Part of computer science which studies the **retrieval of information (not data)** from a collection of **written documents**. The retrieved documents aim at satisfying a **user information need** usually expressed in **natural language**.

- Documents, unstructured, text, large
- Information need
- Store, search, find
- The World Wide Web?
- Relational databases?



# DIKW

- **Data:** Raw web pages
  - **Information:** Result of query
  - **Knowledge:** Result of processing query result by user
  - **Wisdom:** Synthesis of many such actions by a set of users
- 
- One possible classification of steps in process



# Information Retrieval vs. Databases

| Information retrieval  | Data retrieval  |
|--|---|
| Retrieve all objects <b>relevant</b> to some <b>information need</b> | Retrieve all objects satisfying some <b>clearly defined conditions</b>                |
| Find all documents about the <b>topic</b> “semantic web”!            | <b>SELECT</b> id <b>FROM</b> document <b>WHERE</b> title <b>LIKE</b> ‘%semantic web%’ |
| <b>Result list</b>   | <b>Well-defined result set</b>  |



```
[selke@tddb ~]$ db2 "SELECT id FROM document WHERE title LIKE '%semantic web%' FETCH FIRST 3 ROWS ONLY"
```

ID

```
-----
45489
9635899
98556
```

3 record(s) selected.



# Web Search

- Very similar to information retrieval
- Main differences:
  - **Links** between Web pages can be exploited
  - **Collecting**, storing, and **updating** documents is more difficult
  - Usually, the **number of users** is very large
  - **Spam** is a problem







# IR Models

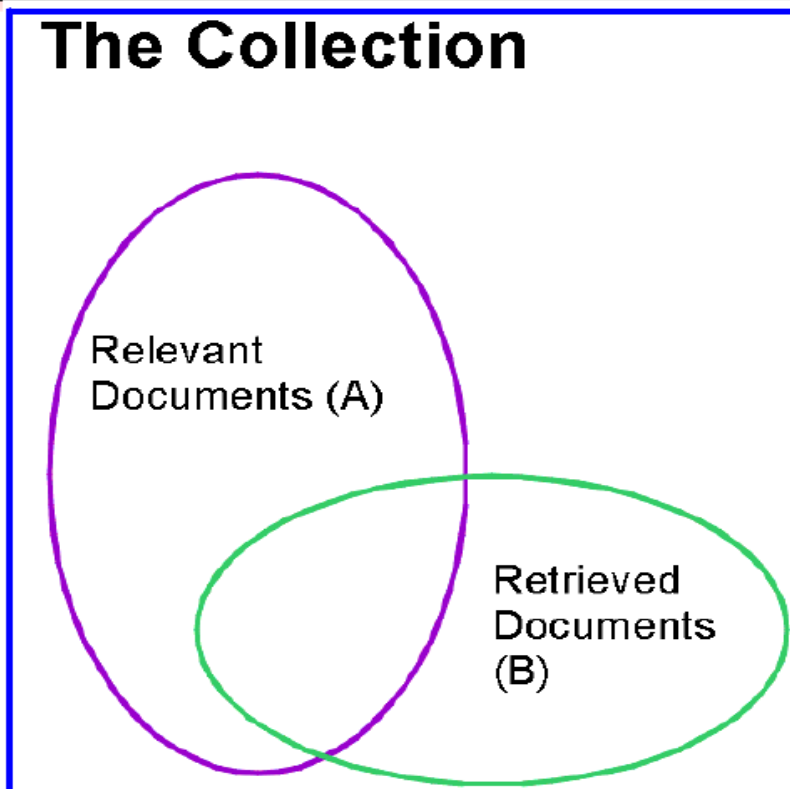
- Any IR system is based on an **IR model**
- The model defines ...
  - ... a **query language**,
  - ... an internal **representation of queries**,
  - ... an internal **representation of documents**,
  - ... a **ranking function** which associates a real number with each query–document pair.
- Optional: A mechanism for **relevance feedback**





# "Classic" Information Retrieval

## The Collection



- Given a query, the system retrieves a set B of documents
- Every retrieved document is either relevant or irrelevant to the query

Quality metrics:

- Recall :  $(A \cap B) / A$
- Precision :  $(A \cap B) / B$



# Recall and Precision on the Web

- Relevance of document to queries is not binary – there are many shades of gray
- Broad-topic queries:
  - abundance problem
  - Precision is the dominating factor: users mostly satisfied with a few good results (a few *authoritative* pages)
- Narrow-topic queries:
  - Find a needle in an enormous haystack
  - Recall demands engines cover significant portions of the Web
- Common measure: precision@10
- Nowadays larger emphasis on *diversity*
  - Positive recall for many aspects of the query

## New Google Update “Improves Diversity” Of Domains In Results

Sep 14, 2012 at 5:03pm ET by [Danny Sullivan](#)

[Facebook](#) Share 118 [Twitter](#) Tweet 993 [Google+](#) +1 215 [LinkedIn](#) Share 110 [Buffer](#) 46

There's been [chatter](#) about Google having done some type of algorithm update this week that's having an impact on rankings. Now it's finally confirmed, a new change meant to allow more domains to appear in search results.



The head of Google's spam fighting team Matt Cutts [tweeted](#) this about it:

Just fyi, we rolled out a small algo change this week that improves the diversity of search results in terms of different domains returned.

Lately, more and more people have been noting that Google's search results can sometimes be dominated by pages that all come from the same domain. In other words, do a search, and all the listings seem to come from the same web site.

Here's an example of this from last month, [when we wrote](#) about the problem:

christopher jagmin plates

About 5,570 results (0.20 seconds)

**All from this domain**

[home \\* christopher jagmin](#)  
[christopherjagmin.com/](#)  
**Christopher Jagmin** Design. Living well isn't about excess. A peanut butter and jelly sandwich tastes better on a beautiful **plate**. An escapist novel is more ...  
[SHOP](#) - [Stores](#) - [Press](#) - [Custom](#)

[shop \\* christopher jagmin](#)  
[christopherjagmin.com/shop/](#)  
Odd Number Dinner **Plates** Four 10.5" **plates**. Even Number Dinner **Plates** Four 10.5" **plates**. Odd Little Number **Plates** Four 7.5" **plates**. Even Little Number ...

[buy > Little Silhouette Plates \\* christopher jagmin](#)  
[www.christopherjagmin.com/shop/buy.phtml?id=31](#)  
These four 7 1/2 inch porcelain **plates** are very friendly, and can also be used for salads, for bread, for snacks, or just for fun. They would be very happy to hang ...

[buy > Odd Poco Number Plates \\* christopher jagmin](#)  
[www.christopherjagmin.com/shop/buy.phtml?id=17](#)  
Odd Poco Number **Plates**. ... No need for a passport with these little 7.5" Spanish **plates**. Perfect for tapas, tacos, with a little guacamole on top! 4 **plates** to a set ...

## ■ Diversity of Search Results



# The Bag of Words Representation

- A very popular **representation of documents** is the **bag of words model**
- Each document is represented by a bag (= multiset) of **terms** from a predefined **vocabulary**
- Standard case:
  - Vocabulary  
= set of all the words occurring in the collection's documents
  - Each document is represented by the words it contains



That's one small step for a man,  
a giant leap for mankind

→ { that's, one, small, step,  
for (2), a (2), man, giant,  
leap, mankind }



## The Bag of Words Model (2)

- Cons:
  - Word order gets lost
  - Very different documents could have similar representations
  - Document structure (e.g. headings) and metadata is ignored
- Pros:
  - Simple set-theoretic representation of documents
  - Efficient storage and retrieval of individual terms
  - IR models using the bag of words representation work well!







## The Bag of Words Model (3)

- Any document can be represented by an **incidence vector**:

vocabulary (aka index terms) → that's one small step for a man giant leap mankind taikonaut Zhai's is China

That's one small step for a man,  
a giant leap for mankind

Taikonaut Zhai's small step is a  
giant leap for China

|   |   |   |   |   |   |   |   |   |   |   |   |   |   |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 2 | 2 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|

**incidence matrix**  
(aka term-document matrix)

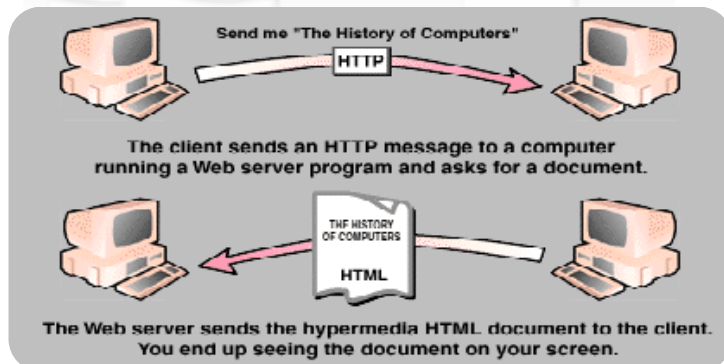
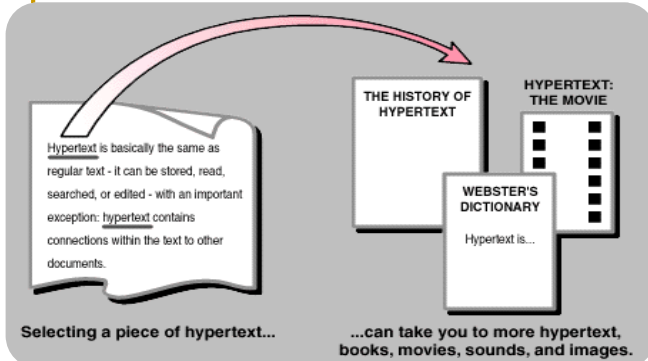
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 1 | 1 | 1 | 1 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|



# Web Terminology

- ❑ Internet – Web
- ❑ Link - Hyperlink
- ❑ Hypertext - Hypermedia
- ❑ HTTP - HTML
- ❑ Anchor Text - URL

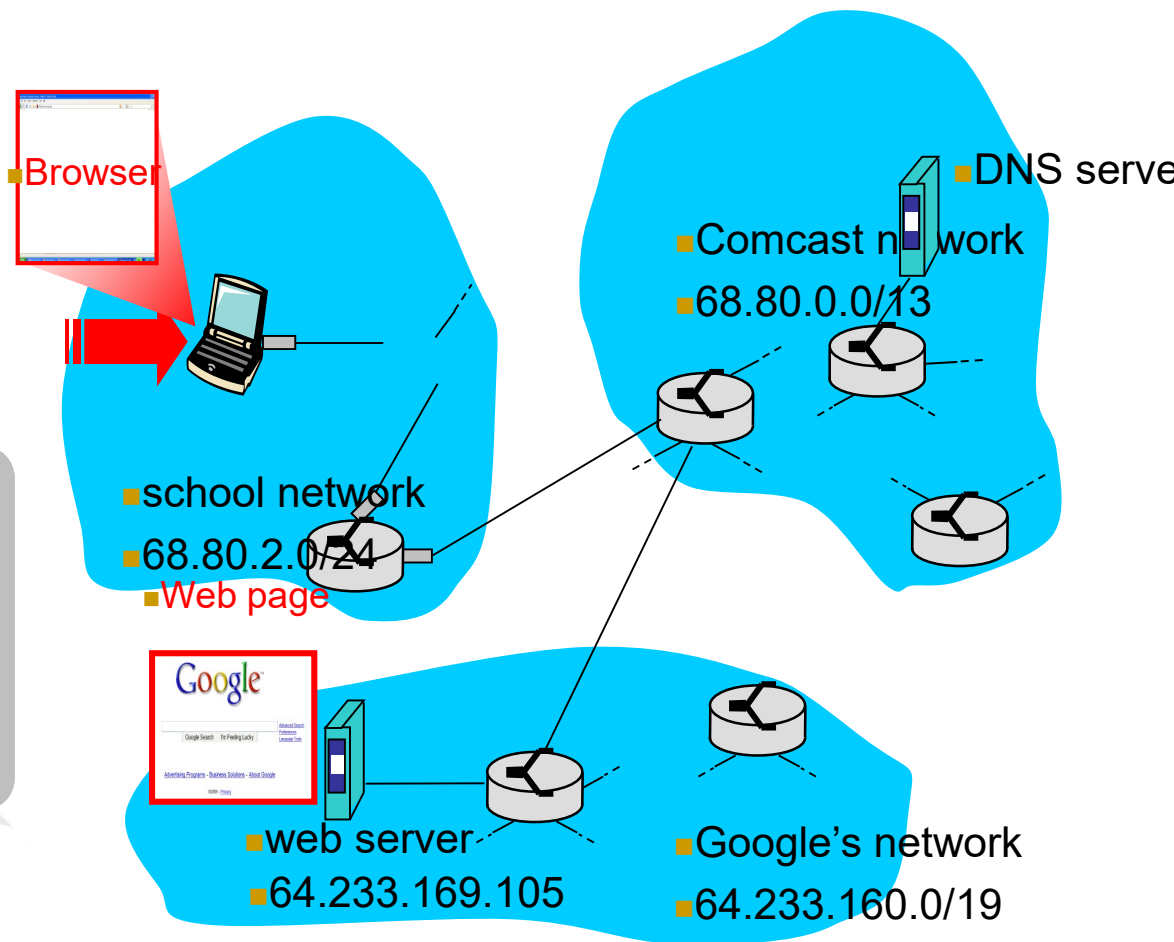
# Web Terminology



A hyperlink is a link you can click on or activate with the keyboard or other device in order to go somewhere else. The hyperlinks are used on the pages and they navigate from one page to another between same or different Sites/servers.

A URL/Link can be thought of as the "address" of a web page and is sometimes referred to informally as a "web address."

Like if we write "http://google.com" in the address bar, it is a link and if we apply it on the page so that, anyone can navigate from the site to Google, it is called hyperlink.



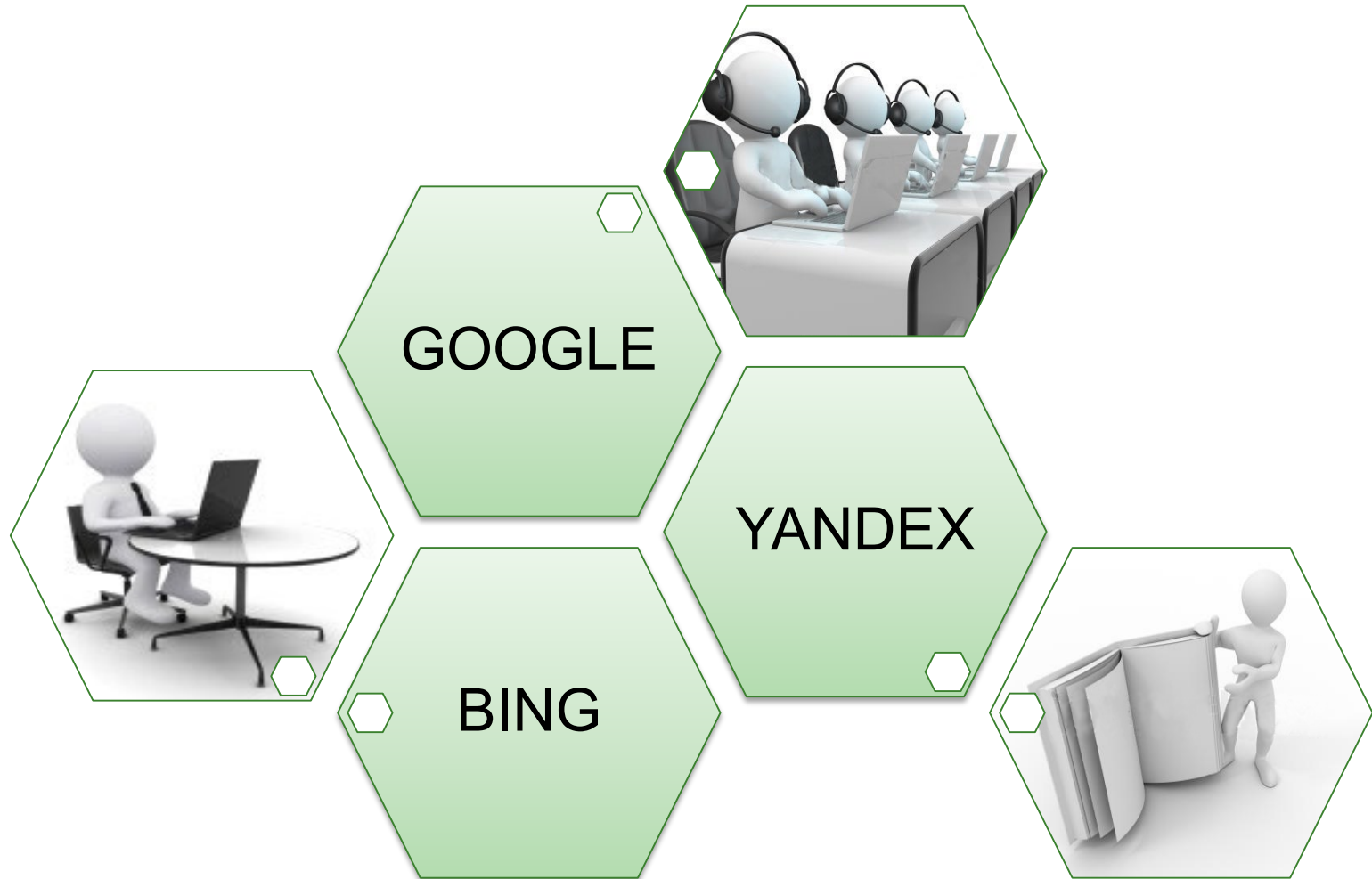
Anchor text usually gives the user relevant descriptive or contextual information about the content of the link's destination. The anchor text may or may not be related to the actual text of the URL of the link. the English-language Wikipedia's homepage might take this form:

```
<a href="http://en.wikipedia.org/wiki/Main_Page">Wikipedia</a>
```

The anchor text in this example is "Wikipedia"; the unwieldy URL [http://en.wikipedia.org/wiki/Main\\_Page](http://en.wikipedia.org/wiki/Main_Page) displays on the web page as Wikipedia, contributing to clean, easy-to-read text.



# Global Search Engines



# Mission Impossible?

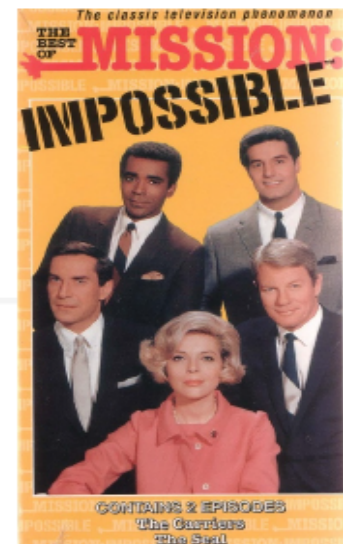
## Search engines:

- Crawl and index **tens of billions** of documents
- Answer **hundreds of millions** of queries per day
- Devote less than **1 second** of processing time per query execution

## Users:

- Submit **very short** queries (averaging about 2.6 terms)
- Expect to receive **the most relevant** results on the Web
- In a **blink of an eye**

In terms of 1990 Information Retrieval research - almost unimaginable!





# The Static Web “Corpus”

- **Bulk** ..... 10s of billions of “static” pages; doubling every 8-24 months
- **Lack of stability**..... Estimates vary widely (half life of page is several months)
- **Lexicon Size** ..... 10s-100s of millions of words
- **Heterogeneity**
  - Type of documents ..... Text, pictures, audio, video, scripts,...
  - Quality ..... From junk to the best content on earth
  - Authorship ..... From masterfully edited content to user-generated content
  - Languages ..... 100+
- **Duplication**
  - Syntactic..... 30%-40% (near) duplicates
  - Semantic..... ??
- **Complex graph topology**.....  $\geq 8$  links/page in the average; bow-tie structure
- **Spam**..... 100s of millions of pages



# Web Searchers - Observations

- Make ill defined queries
  - Short (2.54 terms average, 80% contain less than 3 words)
  - Use imprecise (and often misspelled) terms
  - Unfamiliar with query syntax (80% queries without operator)
- Wide variance in information needs, expectations, education/knowledge, screen sizes, IP bandwidth, patience
  - Different modalities (mobile, desktop) = different needs and expectations, even with same person
- Specific behavior (desktop and laptop)
  - 85% look **over one result screen only** (mostly “above the fold”)
  - 78% of queries are not modified
- Overall, we as users are investing low cognitive effort per query (formulating and looking at results)



Web Images Groups News Froogle more »

spears

Search

[Advanced Search](#)  
[Preferences](#)

## Web

Results 1 - 10 of about 9,440,000 for **spears** [\[definition\]](#). (0.14 seconds)[News results for spears](#) - [View today's top stories](#)

[Knee Injury Closes Spears' Onyx Hotel](#) - [Billboard](#) - 1 hour ago  
[Britney Spears' tour is canceled](#) - [San Diego Union Tribune](#) - 7 hours ago  
[As fall approaches, Spears may start to smell Curious](#) - [Houston Chronicle](#) - Jun 14, 2004

[Britney Spears :: The Official Web Site](#)

The Official Web Site of Britney Spears. Your official source for all things Britney. ...  
 Remember, proceeds benefit the Britney Spears Foundation. ...  
[www.britneyspears.com/](#) - 41k - Jun 14, 2004 - [Cached](#) - [Similar pages](#)

[Britney Spears - britney.com - Jive Records](#)

iTunes. Real/Rhapsody. Napster. Under 11.  
[www.britney.com/](#) - 10k - [Cached](#) - [Similar pages](#)

[Britney Spears Portal - pics, lyrics, MP3s and more!](#)

Britney Spears pics, lyrics, MP3s, news, gossip, fan sites, forums, and much more!  
 Britney Spears Portal, ... ): Britney Spears Portal. ...  
[www.britney-spears-portal.com/](#) - 25k - [Cached](#) - [Similar pages](#)

[Britney Spears guide to Semiconductor Physics: semiconductor ...](#)

Britney Spears lectures on semiconductor physics, radiative and non-radiative transitions, edge emitting lasers and VCSELs. ...  
[britneyspears.ac/lasers.htm](#) - 13k - [Cached](#) - [Similar pages](#)

[BritneySpears.org: Your online guide to Britney!](#)

A comprehensive Britney Spears fansite which pays tribute to Britney with the most active message board, daily news, many pictures, desktop media and more. ...  
[www.britneyspears.org/](#) - 78k - Jun 14, 2004 - [Cached](#) - [Similar pages](#)

[Britney-Spears.To You! - The Britney Spears Community](#)

Britney Spears : biography, discography, musics, real, mp3, videos, pictures, clips, guestbook, www board, free page, search engine, links and more. ...  
[www.britney-spears.to/](#) - 9k - [Cached](#) - [Similar pages](#)

[The Mystery of Britney's Breasts](#)

[www.liquidgeneration.com/poptoons/britneys\\_breasts.asp](#) - 2k - [Cached](#) - [Similar pages](#)

[Britney Spears spelling correction](#)

The data below shows some of the misspellings detected by our spelling correction system for the query [ britney spears ], and the count of how many different ...  
[www.google.com/jobs/britney.html](#) - 40k - [Cached](#) - [Similar pages](#)

[Britney Spears pictures news music Britney Spears lyrics](#)

Britney Spears pictures mp3 sites gallery photos images music fun games chat lyrics. ...  
 Britney Spears Forum Come see what is inside the Britney Spears forum! ...  
[www.britney-spears.com/](#) - 42k - Jun 14, 2004 - [Cached](#) - [Similar pages](#)

[Britney Spears Zone - Your Guide to Britney Pictures and News](#)

Britney Spears, Britney Spears, Britney Spears, ... Britney Spears, ...  
[www.britneyzone.com/](#) - 101k - Jun 14, 2004 - [Cached](#) - [Similar pages](#)

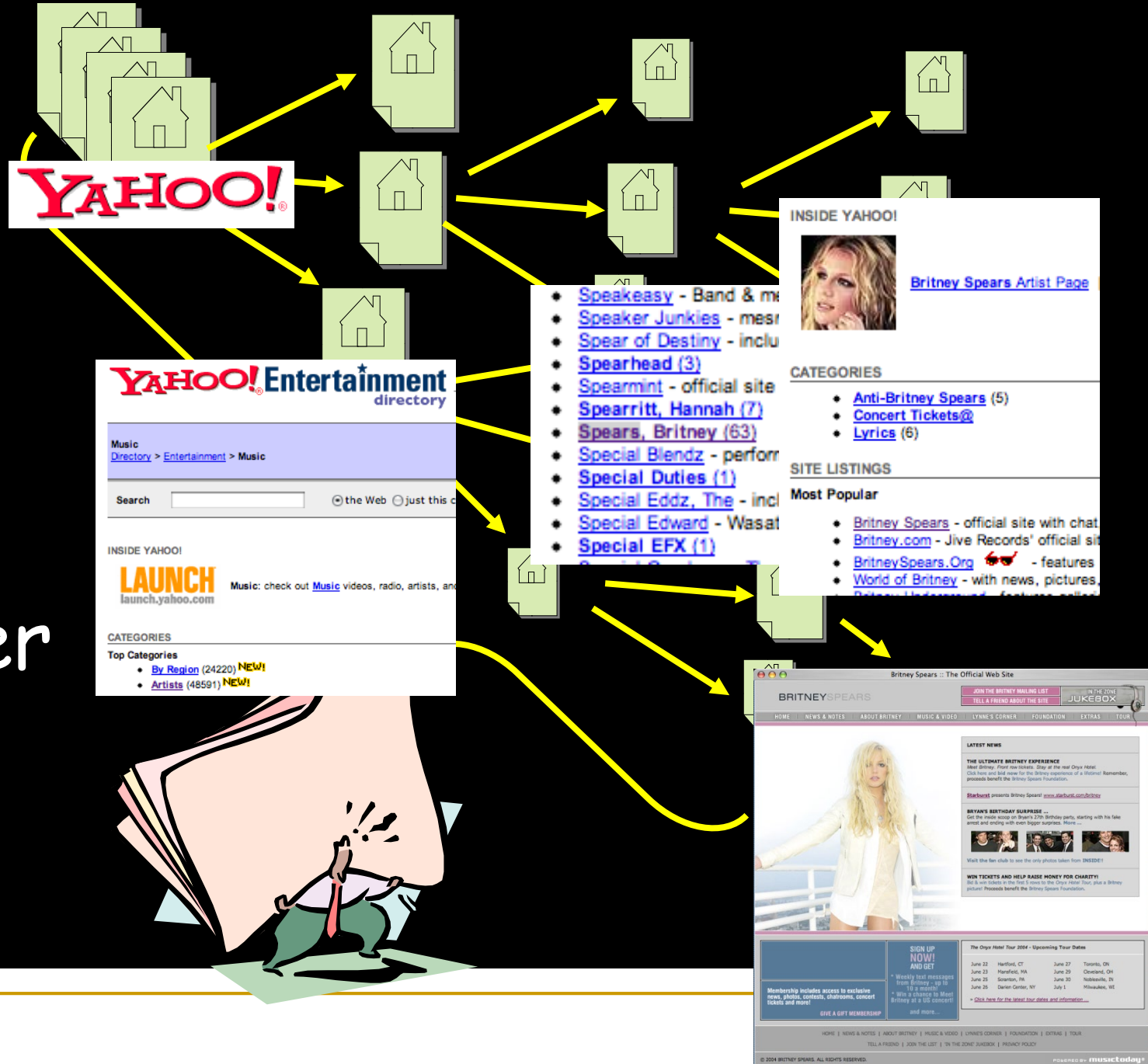
## Sponsored Links

[Financial FAQ](#)

An online loan may be exactly what you need.  
[www.tv-guide.ws](#)

[See your message here...](#)

# Crawler





[illegible]

COMMUNITIES RISE AND FALL, AND TOTAL MEMBERSHIP NUMBERS ARE NO LONGER A GOOD MEASURE OF A COMMUNITY'S CURRENT SIZE AND HEALTH. THIS UPDATED MAP USES SIZE TO REPRESENT TOTAL SOCIAL ACTIVITY IN A COMMUNITY—THAT IS, HOW MUCH TALKING, PLAYING, SHARING, OR OTHER SOCIALIZING HAPPENS THERE. THIS MEANT SOME COMPARING OF APPLES AND ORANGES, BUT I DID MY BEST AND TRIED TO BE CONSISTENT.

ESTIMATES ARE BASED ON THE BEST NUMBERS I COULD FIND, BUT INVOLVED A GREAT DEAL OF GUESSWORK, STATISTICAL INFERENCE, RANDOM SAMPLING, NONRANDOM SAMPLING, A 20,000-CELL SPREADSHEET, EMAILING, CADDING, TEA-LEAF READING, GOAT SACRIFICES, AND GUT INSTINCT (I.E. MAKING THINGS UP).

SOURCES OF DATA INCLUDE GOOGLE AND BING, WIKIPEDIA, ALEXA, GIG-BOARDS.COM, STUMBLEUPON, WORDPRESS, AKUSMET, EVERY WEBSITE STATISTICS PAGE I COULD FIND, PRESS RELEASES, NEWS ARTICLES, AND INDIVIDUAL SITE EMPLOYEES. THANKS IN PARTICULAR TO: ROKS AT LAST.FM, LIVEJOURNAL, REDDIT, AND THE NEW YORK TIMES, AS WELL AS SYSADMIN AT A NUMBER OF SITES WHO SHARED STATISTICS ON CONDITION OF ANONYMITY.





## Why Should I Know about All This?

### Gartner

- “80% of business is conducted on **unstructured** information”
- “85% of all data stored is held in an **unstructured** format”
- “7 million **Web pages** are being added every day”

### Butler Group a Datamonitor Company

- “**Unstructured** data doubles every three months”