ifis

Institut für Informationssysteme
Technische Universität Braunschweig

Information Retrieval and Web Search Engines

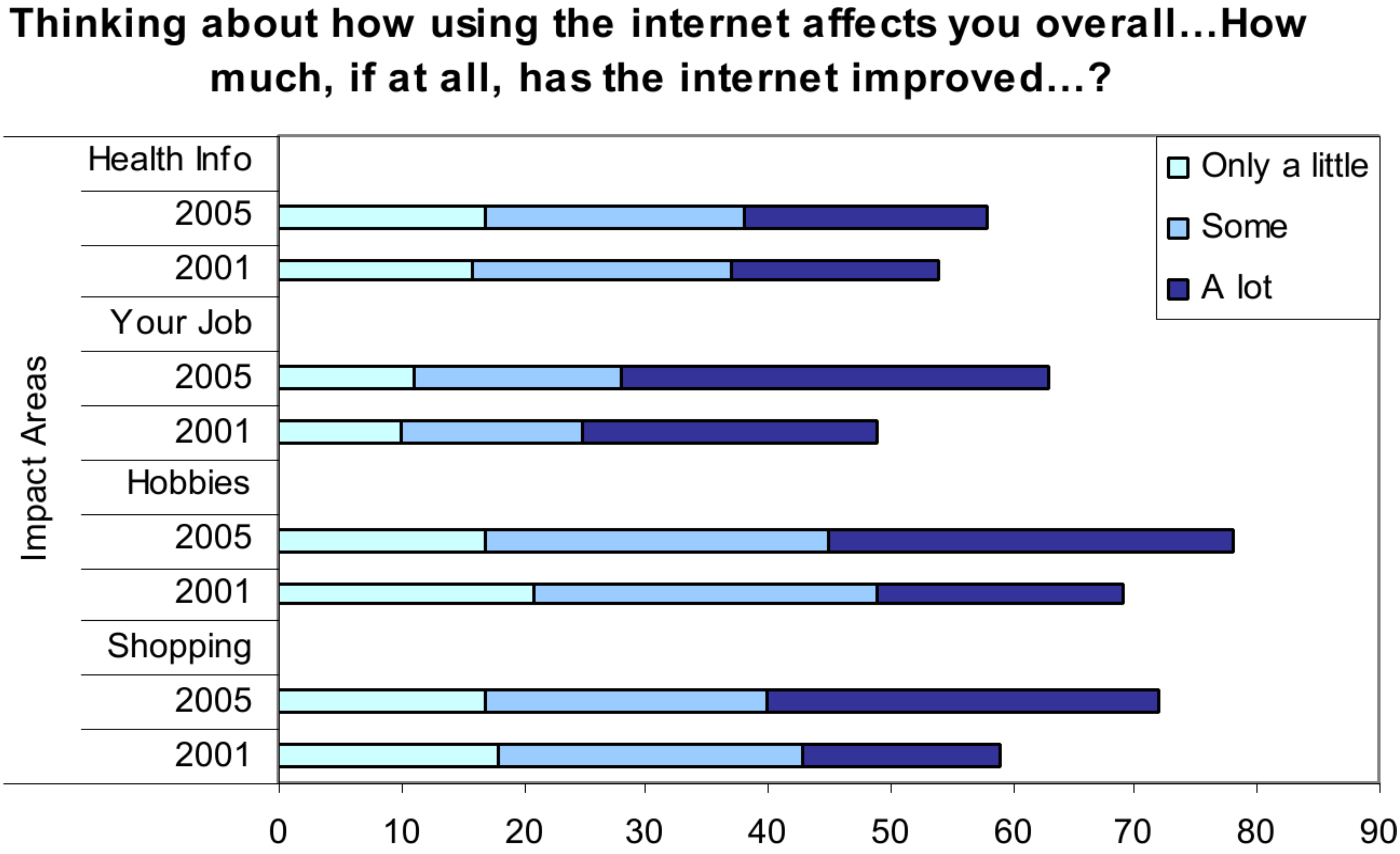
Lecture 10: Introduction to Web Retrieval

Wolf-Tilo Balke

Muhammad Usman

Institut für Informationssysteme
Technische Universität Braunschweig

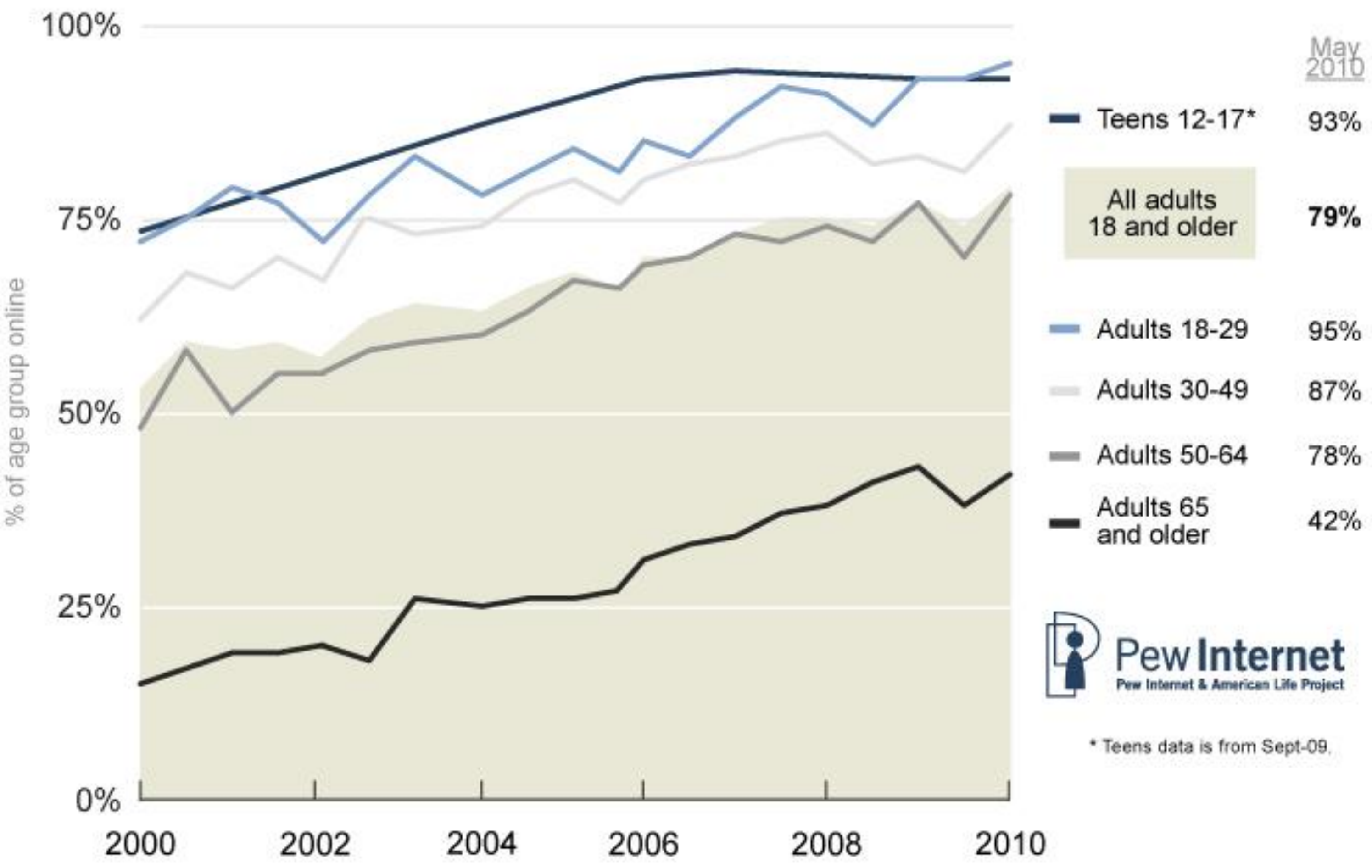
The Web is Important



Source: pewinternet.org

Most People Use the Web

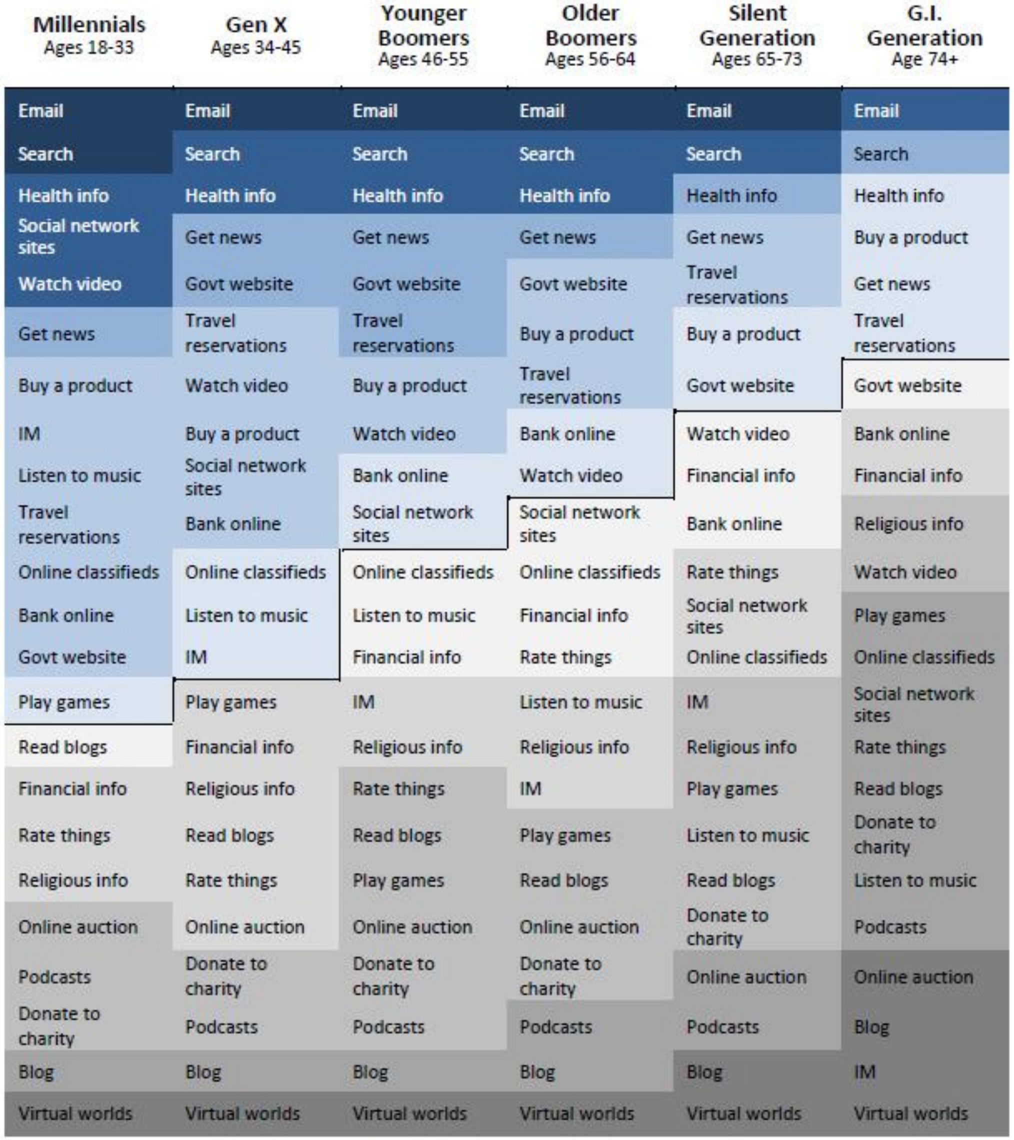
Change in internet use by age, 2000-2010 (US)



What Do People Do?

90-100%	40-49%
80-89%	30-39%
70-79%	20-29%
60-69%	10-19%
50-59%	0-9%

Key: % of internet users in each generation who engage in this online activity



Millennials Ages 18-33	Gen X Ages 34-45	Younger Boomers Ages 46-55	Older Boomers Ages 56-64	Silent Generation Ages 65-73	G.I. Generation Age 74+
Email	Email	Email	Email	Email	Email
Search	Search	Search	Search	Search	Search
Health info	Health info	Health info	Health info	Health info	Health info
Social network sites	Get news	Get news	Get news	Get news	Buy a product
Watch video	Govt website	Govt website	Govt website	Travel reservations	Get news
Get news	Travel reservations	Travel reservations	Buy a product	Buy a product	Travel reservations
Buy a product	Watch video	Buy a product	Travel reservations	Govt website	Govt website
IM	Buy a product	Watch video	Bank online	Watch video	Bank online
Listen to music	Social network sites	Bank online	Watch video	Financial info	Financial info
Travel reservations	Bank online	Social network sites	Social network sites	Bank online	Religious info
Online classifieds	Online classifieds	Online classifieds	Online classifieds	Rate things	Watch video
Bank online	Listen to music	Listen to music	Financial info	Social network sites	Play games
Govt website	IM	Financial info	Rate things	Online classifieds	Online classifieds
Play games	Play games	IM	Listen to music	IM	Social network sites
Read blogs	Financial info	Religious info	Religious info	Religious info	Rate things
Financial info	Religious info	Rate things	IM	Play games	Read blogs
Rate things	Read blogs	Read blogs	Play games	Listen to music	Donate to charity
Religious info	Rate things	Play games	Read blogs	Read blogs	Listen to music
Online auction	Online auction	Online auction	Online auction	Donate to charity	Podcasts
Podcasts	Donate to charity	Donate to charity	Donate to charity	Online auction	Online auction
Donate to charity	Podcasts	Podcasts	Podcasts	Podcasts	Blog
Blog	Blog	Blog	Blog	Blog	IM
Virtual worlds	Virtual worlds	Virtual worlds	Virtual worlds	Virtual worlds	Virtual worlds

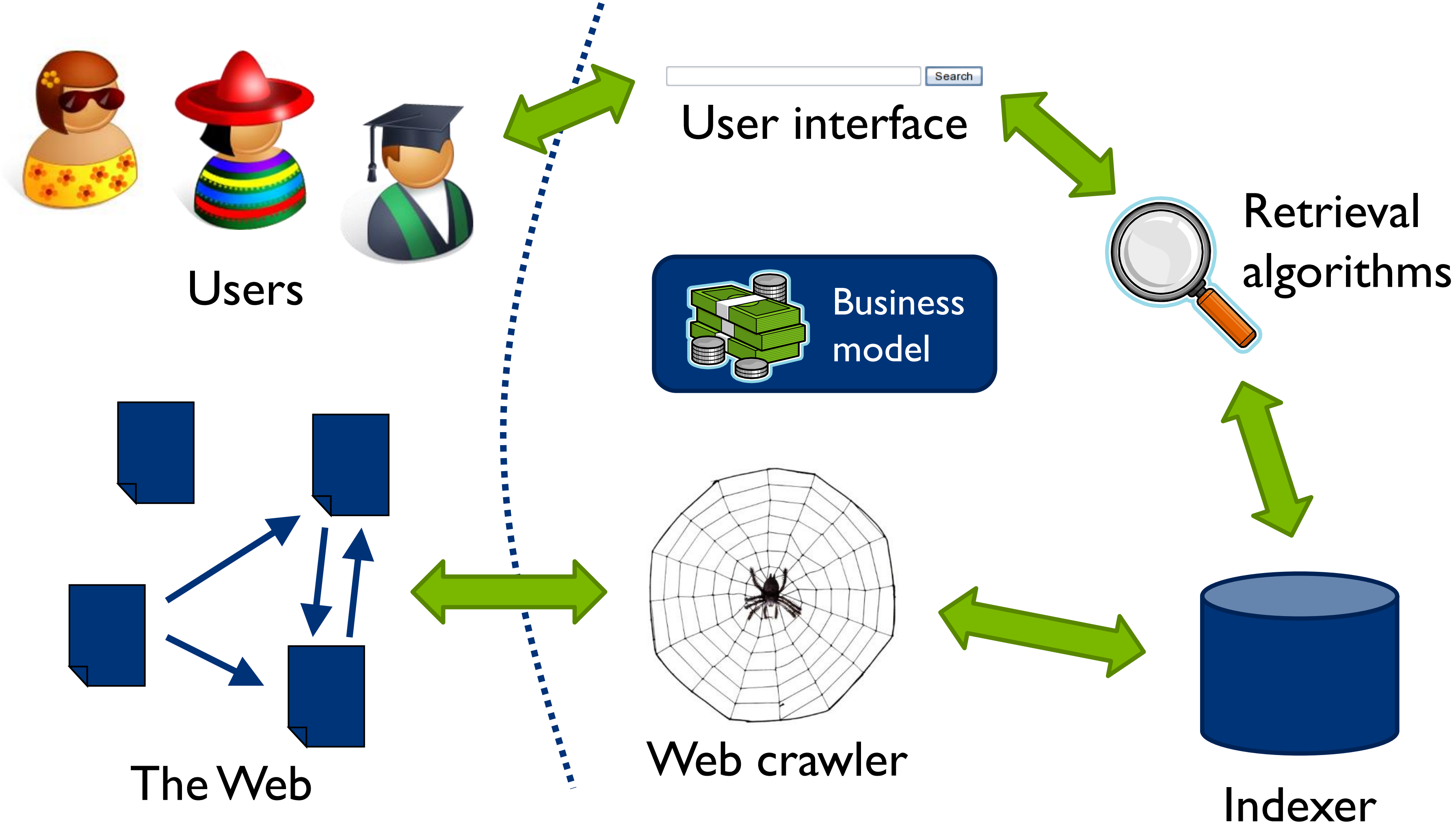
Source: pewinternet.org

Web Search is Essential

- Without Web search, **content** cannot be found
  - Why create online content if nobody will read it?
  - Only for very popular topics, Web search can be replaced by Web directories like DMOZ
- Without Web search, there would be less **collaboration**
  - How to find people with similar interests and problems?
  - What open source projects would be possible without Web search? What about the Social Web?
- Without Web search, **bills** cannot be paid
  - Infrastructure, servers, and content cost a lot of money
  - This is largely paid by search ads

An Overview of Web Retrieval

A typical Web search engine:



Introduction to Web Retrieval

1. **Web Retrieval vs. Classical IR**
2. What Does the Web Look Like?
3. How Do Users Use the Web?

Web Retrieval vs. Classical IR

- **Heterogeneity**
  - Many different users, topics, languages, document types, …
  - Websites are not classical documents (dynamic content, …)
  - Open platform: variety of authors, opinions, writing styles, …
- **Hyperlinks**
  - Documents are connected and refer to each other
- **Problem size**
  - Many documents, many queries, high percentage of volatile data
- **Spam**
  - Evil forces are around
- **Business model**
  - Web search is expensive

World Internet Usage and Population Statistics

June 30, 2014

Region	Population 2014	Internet Users in 2000	Internet Users latest update	Penetration (% Population)	Growth 2000-2014
Africa	1,125,721,038	4,514,400	**297,885,898**	26.5 %	6,498.6 %
Asia	3,996,408,007	114,304,000	**1,386,188,112**	34.7 %	1,112.7 %
Europe	825,824,883	105,096,093	**582,441,059**	70.5 %	454.2 %
Middle East	231,588,580	3,284,800	**111,809,510**	48.3 %	3,303.8 %
North America	353,860,227	108,096,800	**310,322,257**	87.7 %	187.1 %
Latin America	612,279,181	18,068,919	**320,312,562**	52.3 %	1,672.7 %
Oceania	36,724,649	7,620,480	**26,789,942**	72.9 %	251.6 %
World	**7,182,406,565**	**360,985,492**	**3,035,749,340**	**42.3 %**	**741.0 %**

Heterogeneity of Users

- **Web users are not all alike**
- Demographics of US Internet users (2014):

	Use the Internet
Total adults	87%
Women	86%
Men	87%

Education	Use the Internet
High school grad or less	76%
Some college	91%
College+	97%

Age	Use the Internet
18–29	97%
30–49	93%
50–64	88%
65+	57%

Household income (per year)	Use the Internet
Less than $30,000	77%
$30,000–$49,999	85%
$50,000–$74,999	93%
$75,000+	99%

Source: pewinternet.org

Heterogeneity of Languages

Some statistics about the Web's languages:

Language	Web sites (2013)	Wikipedia articles (2014)
English	54.9%	4,420,454
German	5.3%	1,673,551
Spanish	4.8%	1,070,597
French	4.3%	1,464,427
Japanese	4.2%	889,993
Polish	1.8%	1,021,375
Italian	1.5%	1,090,207
Dutch	1.1%	1,717,560
Swedish	0.6%	1,607,434
Vietnamese	0.4%	885,729

Website language statistics are based on the 1,000,000 most viewed websites

Sources: wikipedia.org

Heterogeneity of Document Types

Some file types a search engine should be able to process:

application/ms-excel (different versions), application/ms-powerpoint (different versions), application/msword (different versions), application/pdf (different versions), application/postscript, application/x-dvi, application/x-tar, application/x-zip-compressed, text/html (different versions and encodings), text/plain (different encodings), text/rtf, application/xml, text/xml, application/xhtml+xml, application/docbook+xml, application/x-shockwave-flash, …

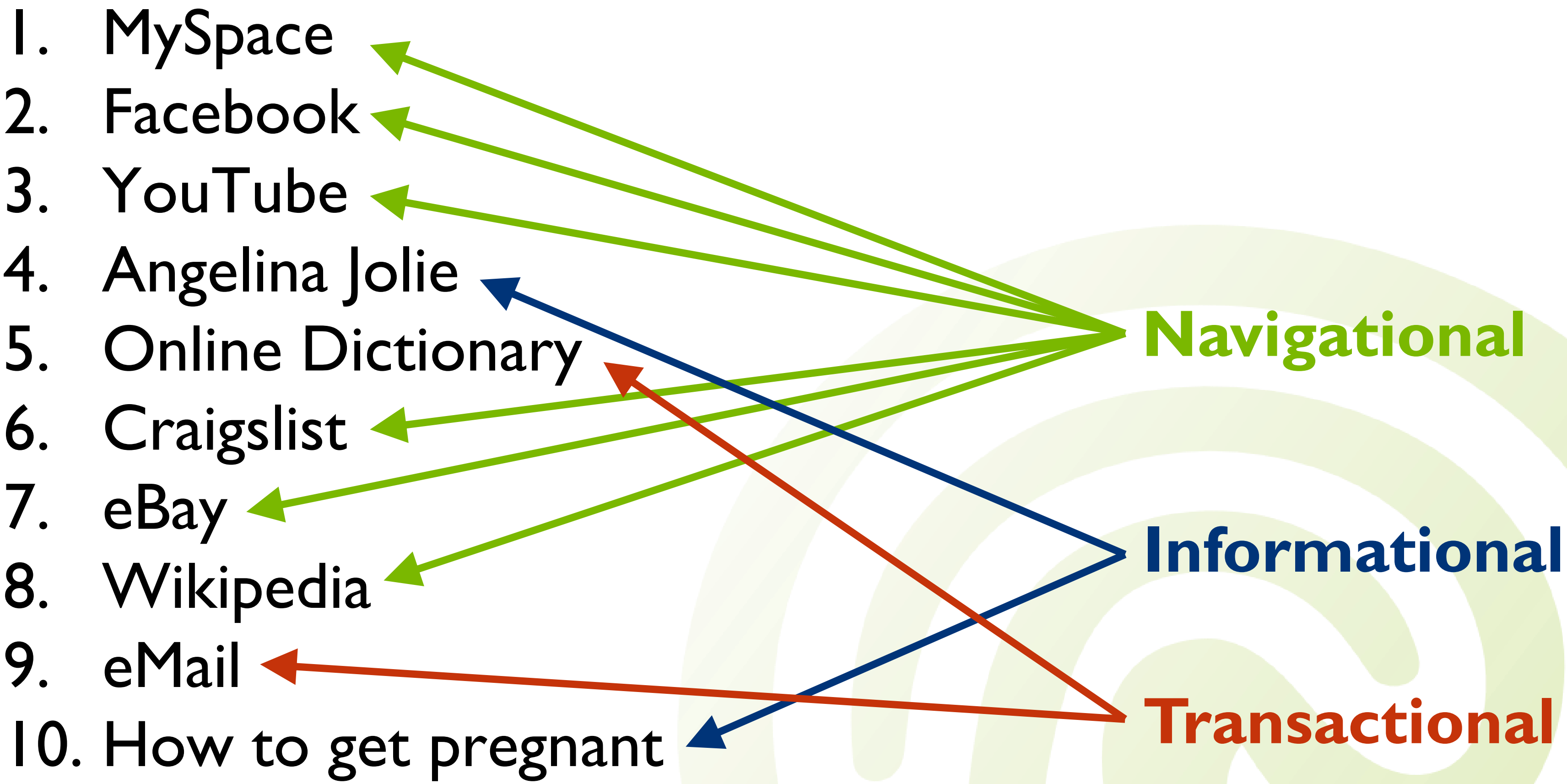
- – Images, videos, audio, executable code?

Heterogeneity of Queries

- Web search engines are used for **different purposes** and within **different contexts**
- There are **four main types of queries:**
  - **Informational queries:**
    Find general information about some topic, e.g., “Web search”
  - **Navigational queries:**
    Find a specific website, e.g., “Facebook”
  - **Transactional queries:**
    Find websites providing some service,
    e.g., “Adobe Reader download”
  - **Connectivity queries:**
    Find connected pages, e.g., “link:www.tu-bs.de”
    (finds all pages that link to http://www.tu-bs.de)

Heterogeneity of Queries

Ask.com's **top searches** for the week ending Jan 16th, 2008:



1. MySpace
2. Facebook
3. YouTube
4. Angelina Jolie
5. Online Dictionary
6. Craigslist
7. eBay
8. Wikipedia
9. eMail
10. How to get pregnant

Navigational

Informational

Transactional

Heterogeneity of Queries

Again, some statistics…

	%of ADULT internet users in the U.S. who do this on a typical day
Use a search engine to find information	59%
Send or read e-mail	59%
Use a social networking site	48%
Get news	45%
Go online just for fun or to pass the time	44%
Look for info on a hobby or interest	35%
Check the weather	34%
Play online games	13%
Look online for info about a job	11%

Source: pewinternet.org

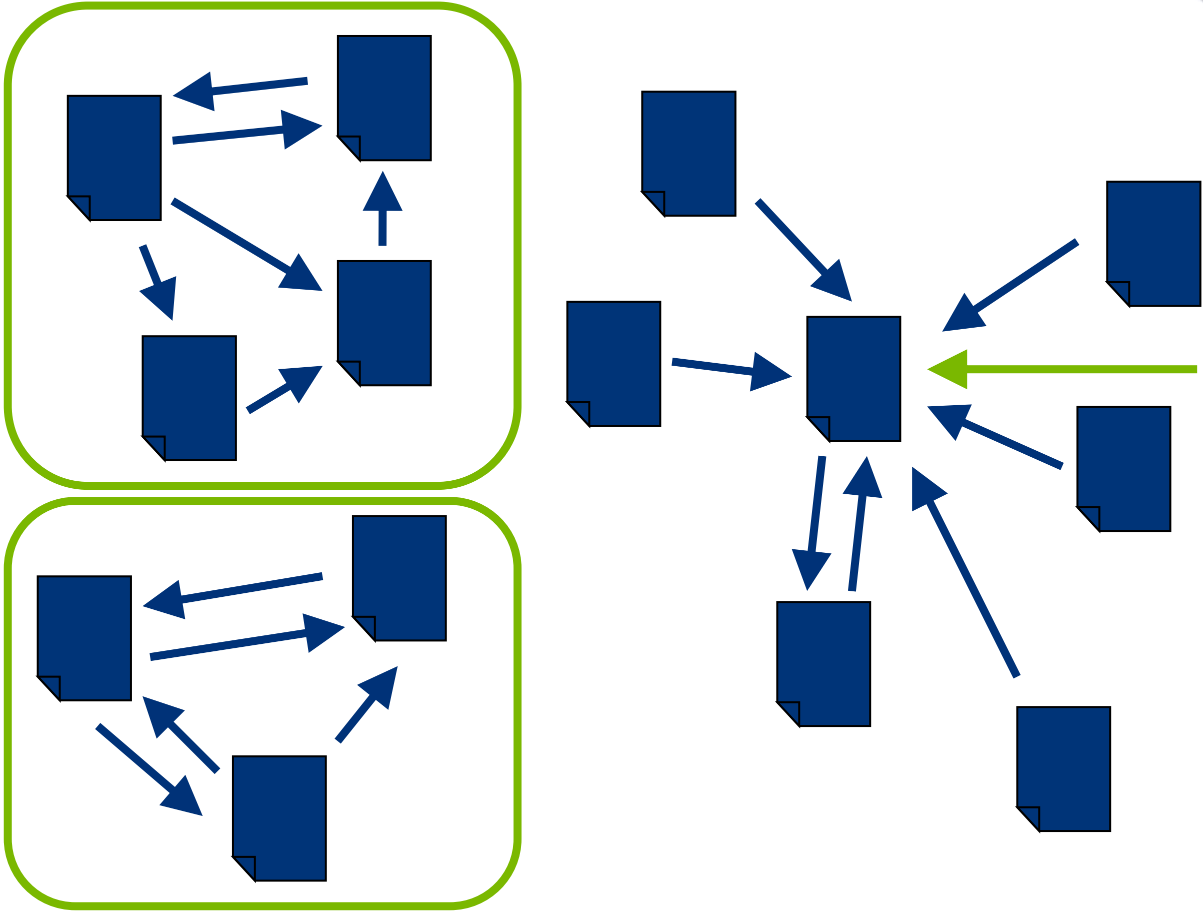
Heterogeneity of Queries

	% of **TEEN** internet users in the U.S. who do this on a typical day
Use a social networking site	80%
Get news about current events or politics	62%
Buy things online e.g. Books, clothing, music	48%
Share something online that you created yourself e.g. Artwork, photos, stories,..	38%
Have a video chat conversation e.g. Skype	37%
Look online for health, dieting or physical fitness information	31%
Use Twitter	16%
Create or work on you own online blog	14%

Source: pewinternet.org

Link Structure

- Web documents can **link** to each other
- **Links are not created randomly**



Two different topics?

This page seems to be interesting

Number of Queries

- **How many queries** a search engine has to process?
- Here are some numbers from 2023 :

	Average number of queries per second
Google	51666
Bing	10115
Yahoo	6517

- **51666 queries per second** are…
  - …around 186 million queries per day
  - …around 1,628 billion queries per year

Index Size

- **How large is a typical Web search engine's index?**
- Here are some recent estimates from worldwidewebsize.com

	Number of indexed Web pages
Google (January 2014)	~15,000,000,000,000
Bing (January 2014)	~9,000,000,000,000
Yahoo (June 2010)	50,000,000,000
Ask (June 2010)	1,700,000,000

  - Both Yahoo and Ask have stopped showing their total number of results, so no recent estimates are available.
- By the way:
  Where did they get these numbers from?

Index Size: Estimation

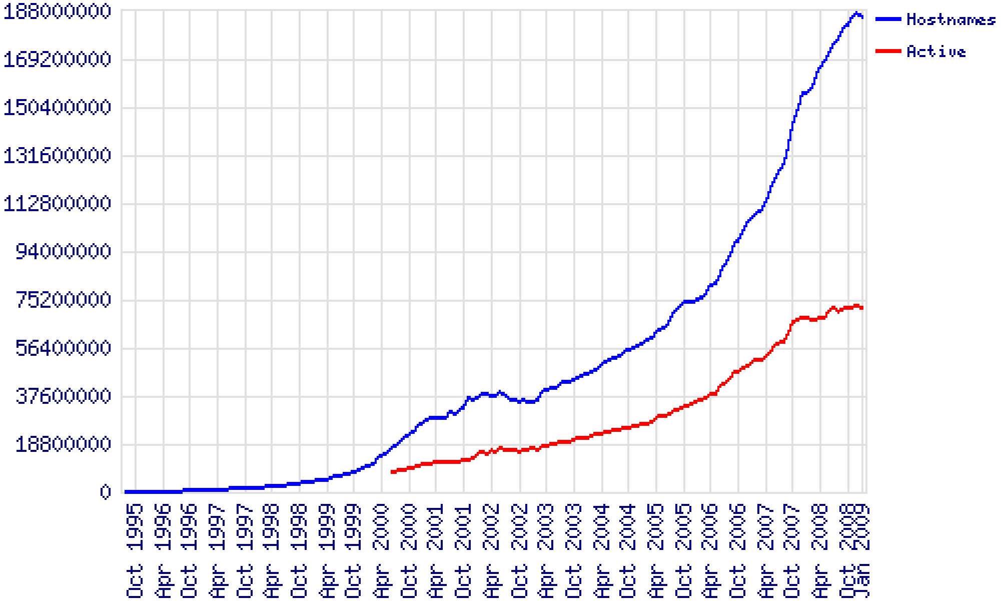
- The authors of worldwidewebsize.com describe their estimation method as follows:
  - Obtain **word frequencies** from a large offline text collection
    - More than 1 million web pages from DMOZ
    - Can be considered a representative sample of the World Wide Web
  - **Send 50 randomly chosen words** to the search engine
    - "Randomly" = selected evenly across logarithmic frequency intervals
  - For each word, **record the number of Web pages found**
  - **Estimate the index size** using these numbers by exploiting the **relative word frequencies** of the background corpus

Web Traffic and Bandwidth

- When operating a search engine, you need a **crawler**
- The crawler must continuously feed the indexer with **new or updated information**
  - New Web pages
  - Deleted Web pages
  - Updated Web pages
- **How much data** must be transferred for doing this?
- Some recent numbers from netcompetition.org:
  - Within the US part of the Internet, Google transfers around **60 petabytes per month:** 60,000,000,000 megabytes!
- Now you know why **Web search is expensive…**

Scalability

- **The Web grows fast (exponentially?)…**
- **The total number of hostnames:**



Source: netcraft.com

- A Web search engine must **scale well** to keep up

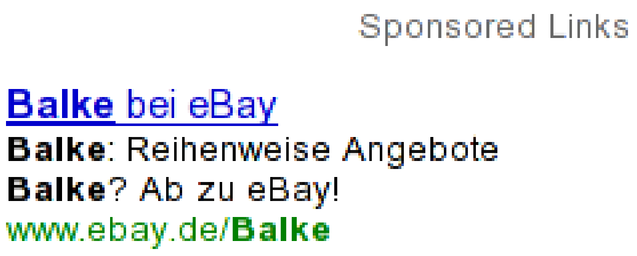
Business Models

Business model:
The method of doing business by which
a company can sustain itself, i.e., generate revenue

- We have seen: Web search is complicated and expensive
  - Exception: Local search functionality for a single Web site
- You cannot run a Web search engine for free
  - Hardware, traffic, development, …
- What could be a reasonable **business model** here?
  - Advertising model
  - Subscription model
  - Community model
  - Infomediary model

Business Models

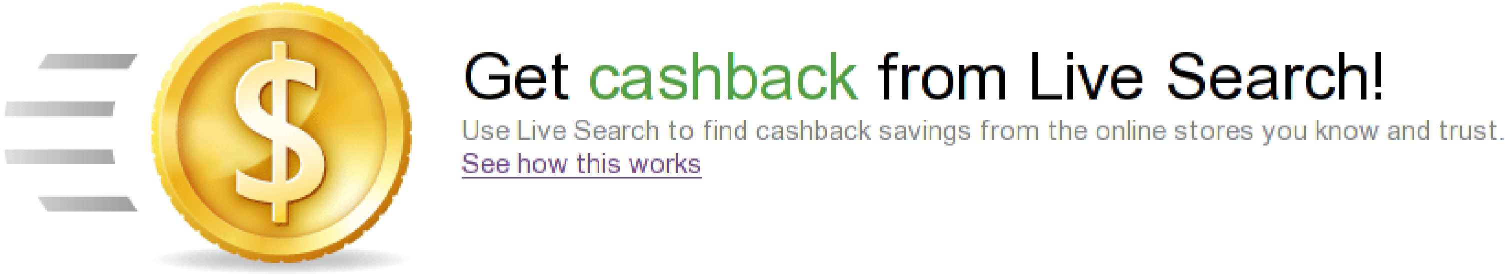
- **The advertising model**
  - You get paid for showing other people's ads on your search result pages



Sponsored Links

[Balke bei eBay]
Balke: Reihenweise Angebote
Balke? Ab zu eBay!
www.ebay.de/**Balke**

  - Used by Google and most other search engines
  - To make this work, your search engine must attract a lot of people and placement of ads must be personalized
  - If your search engine fails at the former, there are other ways: In Microsoft's "Live Search cashback" program, people earn some money if they buy products found via Live Search's ads



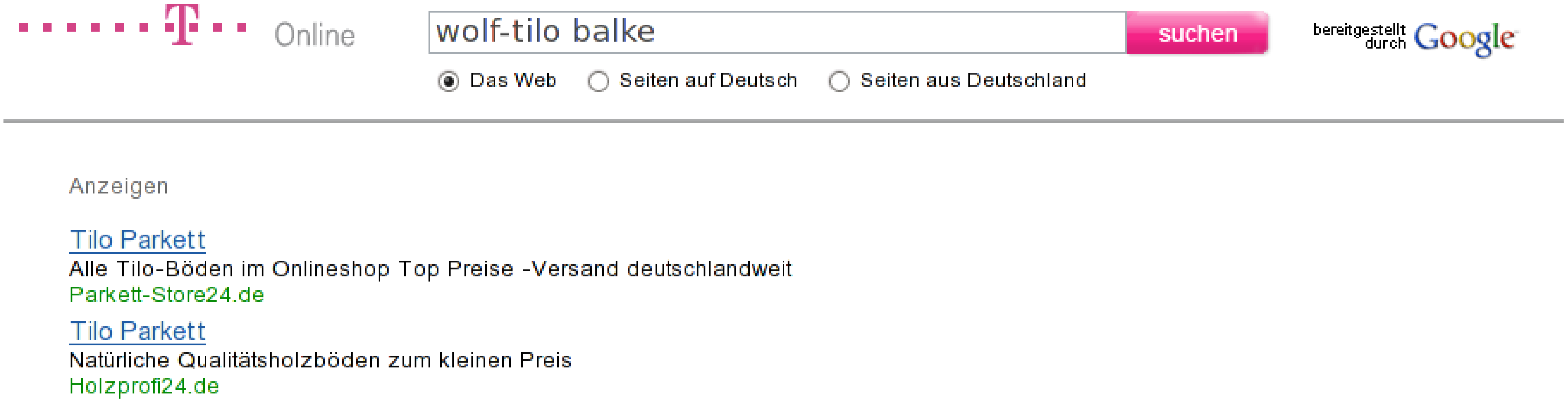
Get cashback from Live Search!

Use Live Search to find cashback savings from the online stores you know and trust.

See how this works

Business Models

- **The subscription model**
  - Customers pay for using your search engine
  - To make this work, your search engine must be really good
  - More popular: Rent your technology to other companies; many search engines use this model
  - Example: t-online.de's search functionality is provided by Google



wolf-tilo balke [suchen]

bereitgestellt durch Google

● Das Web ○ Seiten auf Deutsch ○ Seiten aus Deutschland

Anzeigen

Tilo Parkett
Alle Tilo-Böden im Onlineshop Top Preise -Versand deutschlandweit
Parkett-Store24.de

Tilo Parkett
Natürliche Qualitätsholzböden zum kleinen Preis
Holzprofi24.de

Business Models

- **The community model**
  - Let users participate in product development
  - This lowers costs and often increases product quality
  - Pay your bills by ads and donations
  - Example: Wikia Search, in which users can directly annotate or even modify search results (discontinued in May 2009)

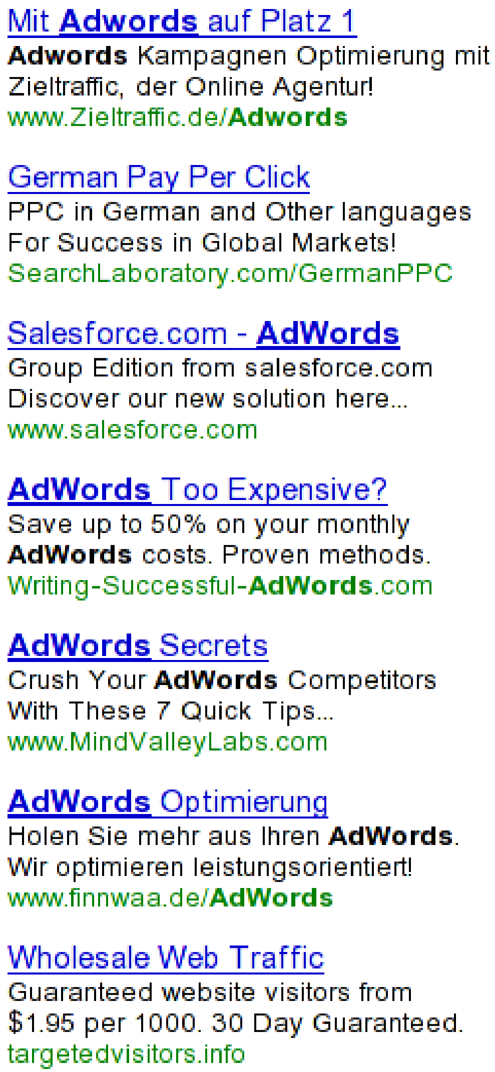
Business Models

- **The infomediary model**
  - Users can use your search engine for free but agree to participate in “market studies”
  - The users’ search behavior is analyzed to yield individual “user profiles” and to distill overall search trends
  - This information is sold to other companies, which can use it to optimize their own advertizing strategies
  - This model usually comes along with severe legal issues regarding the users’ privacy
  - Examples: No search engine would tell about…

Google's Business Model

Detour

- Google's ad program is called **AdWords**
- It's very successful
  - 99% of Google's revenue is derived from its advertising programs
  - In 2007, Google had 1 million advertisers
    2003: 89,000 2005: 360,000
    2004: 201,000 2006: 600,000
  - In 2007, on average, each advertiser spent $16,000 a year on Google ads
  - In 2012, Google earned $42.5 billion with ads



Mit **Adwords** auf Platz 1
Adwords Kampagnen Optimierung mit Zieltraffic, der Online Agentur!
www.Zieltraffic.de/**Adwords**

German Pay Per Click
PPC in German and Other languages For Success in Global Markets!
SearchLaboratory.com/GermanPPC

Salesforce.com - **AdWords**
Group Edition from salesforce.com Discover our new solution here...
www.salesforce.com

AdWords Too Expensive?
Save up to 50% on your monthly **AdWords** costs. Proven methods.
Writing-Successful-**AdWords**.com

AdWords Secrets
Crush Your **AdWords** Competitors With These 7 Quick Tips...
www.MindValleyLabs.com

AdWords Optimierung
Holen Sie mehr aus Ihren **AdWords**. Wir optimieren leistungsorientiert!
www.finnwaa.de/**AdWords**

Wholesale Web Traffic
Guaranteed website visitors from $1.95 per 1000. 30 Day Guaranteed.
targetedvisitors.info

https://adwords.google.com/select/KeywordToolExternal

Google's Business Model

Detour

- How it works…
  - Advertisers:
    1. Identify bidding keywords and price
    2. Create groupings of keywords and ads
  - Upon a search query, google initiates an auction with:
    1. Most relevant keyword
    2. Maximum specified bid
    3. Associated Ad

Google AdWords

Ad Rank

Detour

- During **Auction**, google looks at:
  1. Maximum Bid
  2. Quality Score

- Ranking is given as follows

Ad Rank = Maximum Bid x Quality Score

- Advertiser is charged with the second highest bid.
- As of November 2013, formula was updated

Ad Rank = Max. Bid x Quality Score x Expected Impact from Ad extensions

Google’s Business Model

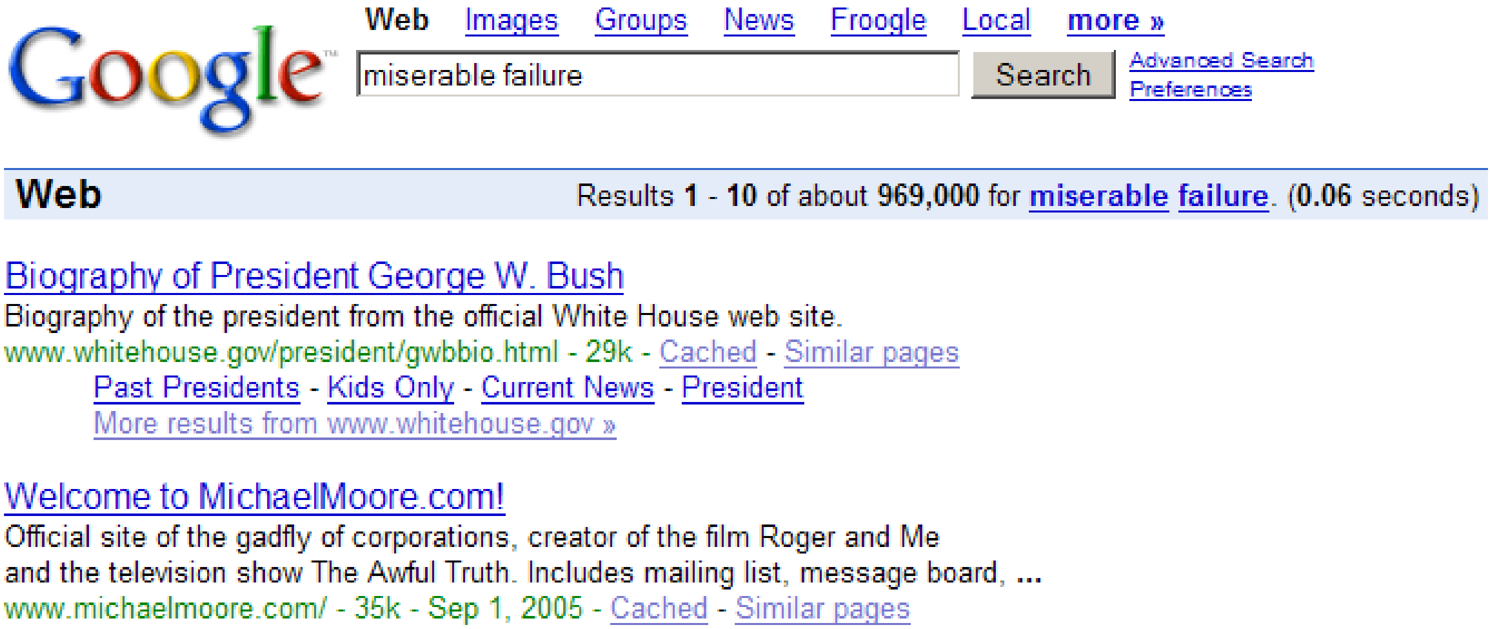
Detour

Most expensive Adwords in 2016 in the USA (according to searchenginewatch):

Bid	Keywords
$935.71	best mesothelioma lawyer
$425.70	dallas truck accident lawyer
$411.04	truck accident lawyer houston
$333.79	louisville car accident lawyer
$388.84	houston wheeler accident lawyer
$381.65	san diego water damage
$377.70	are personal injury settlements taxable
$361.34	baltimore auto accident lawyer
$358.11	accident lawyer sacramento
$358.03	car accident lawyer phoenix

Spam

- There are **cheaper ways than AdWords** to get your page on Google's result pages…
- Just let your page look as if it would be highly relevant…
- The general term for such techniques is **"spamdexing"**



Web Results **1** - **10** of about **969,000** for **miserable failure**. (**0.06** seconds)

Biography of President George W. Bush
Biography of the president from the official White House web site.
www.whitehouse.gov/president/gwbbio.html - 29k - Cached - Similar pages
Past Presidents - Kids Only - Current News - President
More results from www.whitehouse.gov »

Welcome to MichaelMoore.com!
Official site of the gadfly of corporations, creator of the film Roger and Me and the television show The Awful Truth. Includes mailing list, message board, ...
www.michaelmoore.com/ - 35k - Sep 1, 2005 - Cached - Similar pages

Introduction to Web Retrieval

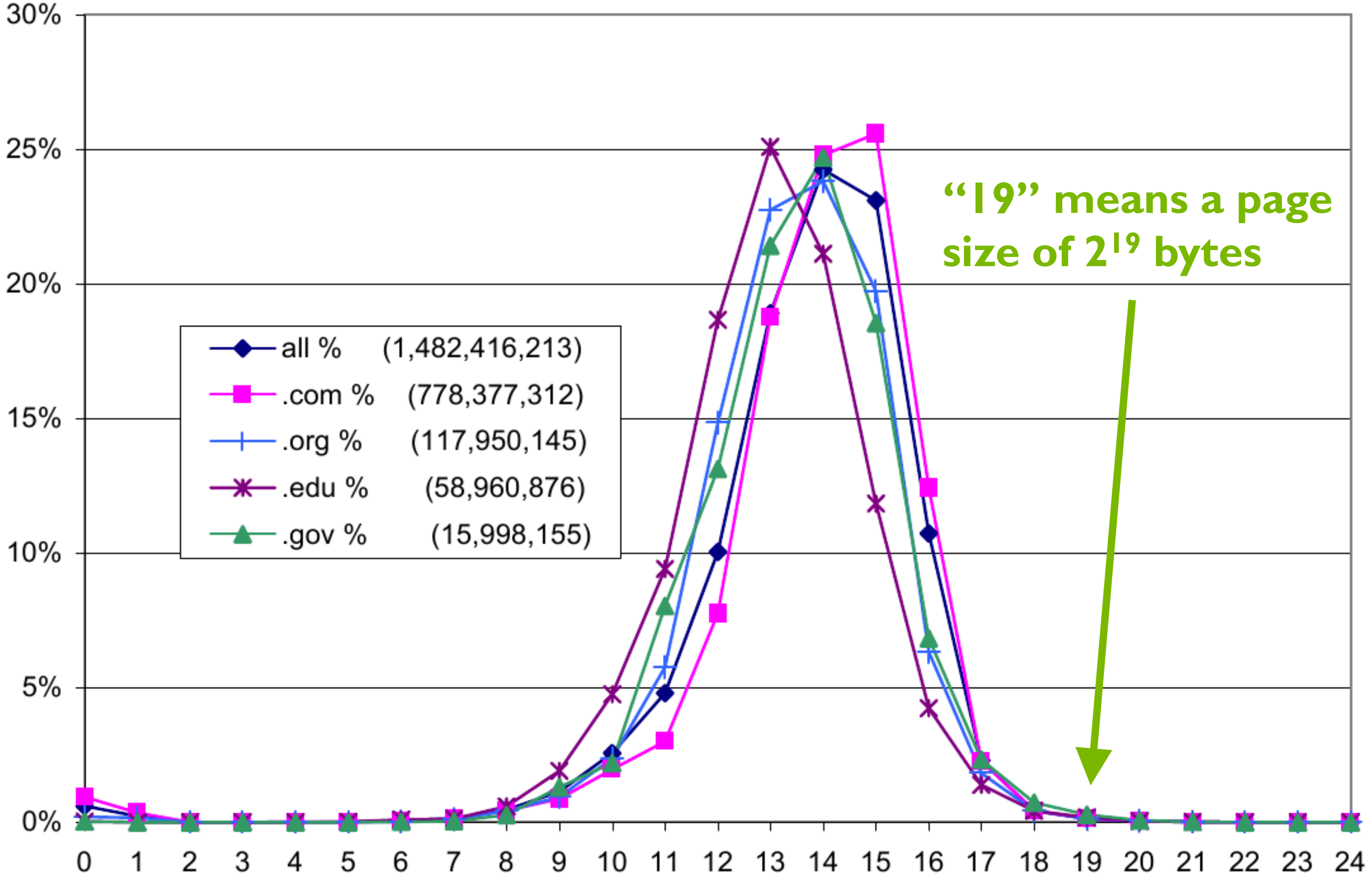
1. Web Retrieval vs. Classical IR
2. **What Does the Web Look Like?**
3. How Do Users Use the Web?

Properties of Web Pages

- In 2002, (Fetterly *et al.*, 2004) crawled a set of around 151 million HTML pages once every week, over a span of 11 weeks
- Amongst others, they tried to answer the following questions:
  - **How large is a Web page (measured in bytes)?**
  - **How large is a Web page (measured in words)?**
  - **How much does a Web page change (within a week)?**

Properties of Web Pages

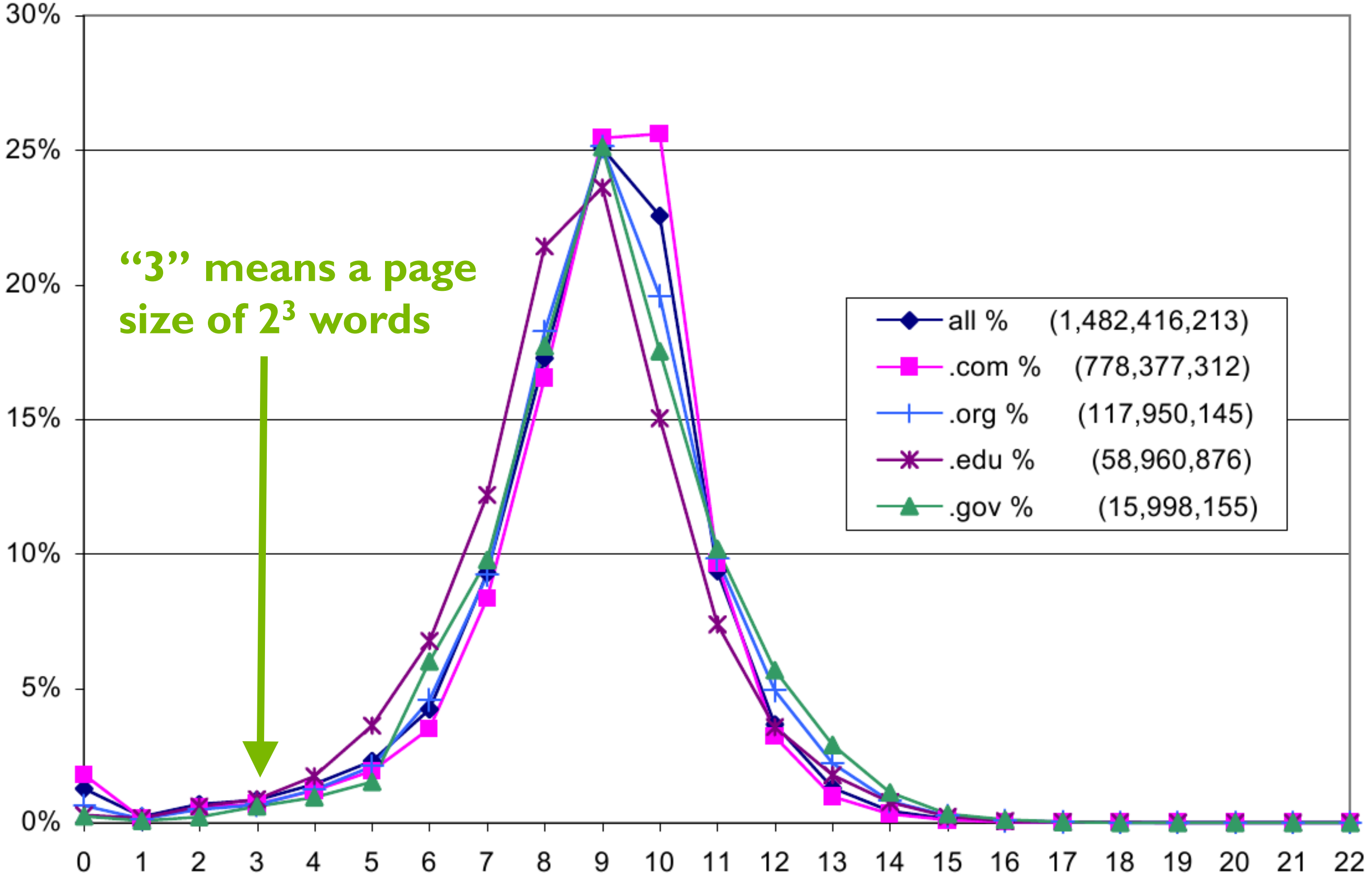
How large is a Web page (measured in bytes)?



"19" means a page size of 2^{19} bytes

Properties of Web Pages

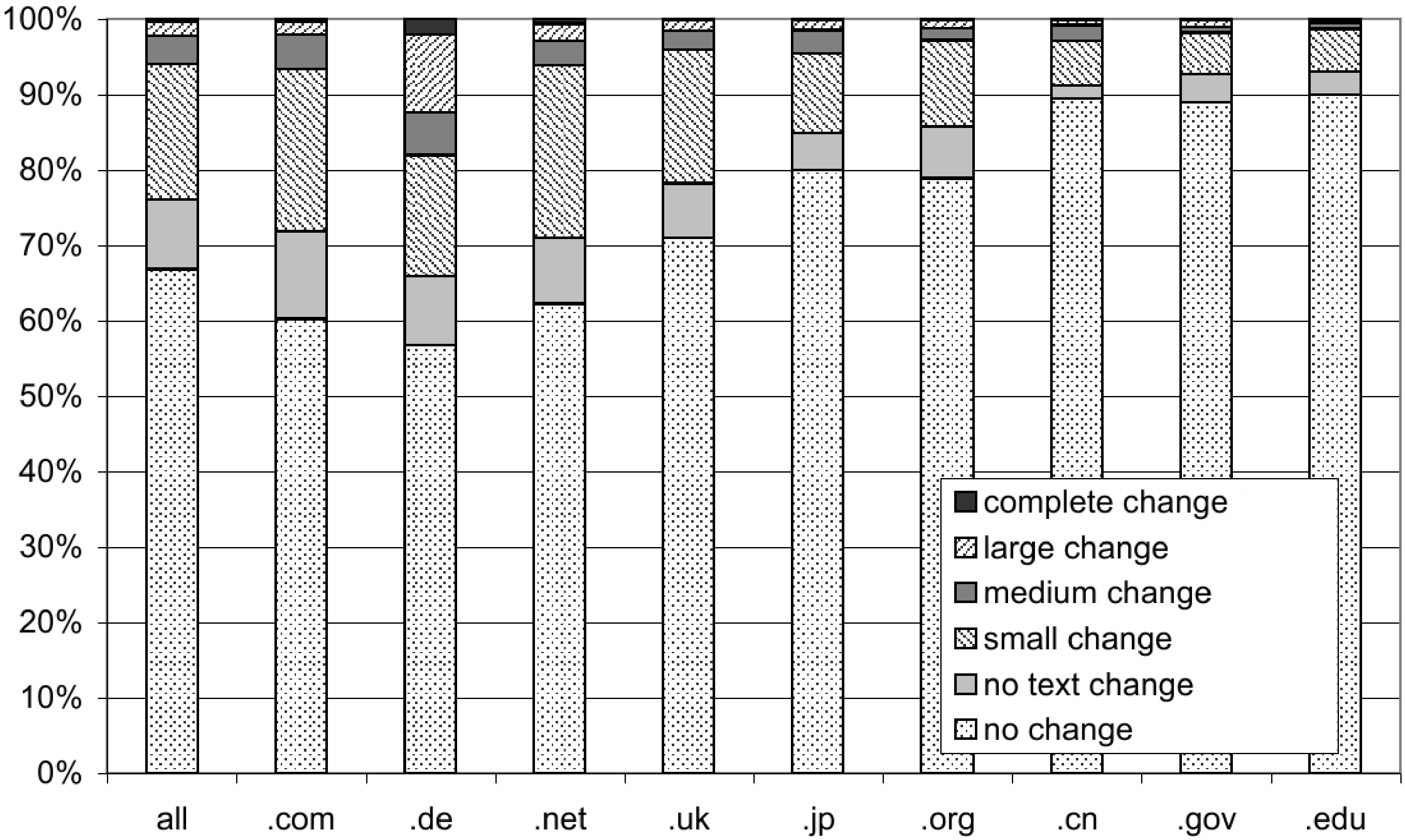
How large is a Web page (measured in words)?



"3" means a page size of 2^3 words

Properties of Web Pages

How much does a Web page change (within a week)?

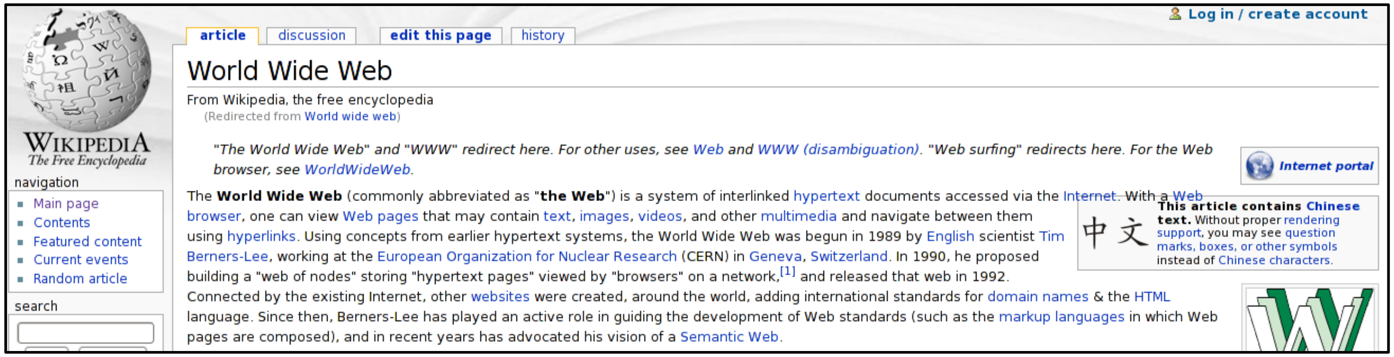


How Large is the Web?

- In 1993, measuring the Web's size has been easy
  - **Every web page corresponded to a file** on some server
  - There was almost **no duplicate content**
  - There was **no spam**
  - Most **Web servers have been known explicitly**
- Estimation of 1993:
  - 100 servers
  - 200,000 documents
  - 4,000,000 pages
- **Today, estimating the Web's size is more difficult**

How Large is the Web?

First problem: **What pages counts as "the Web"?**



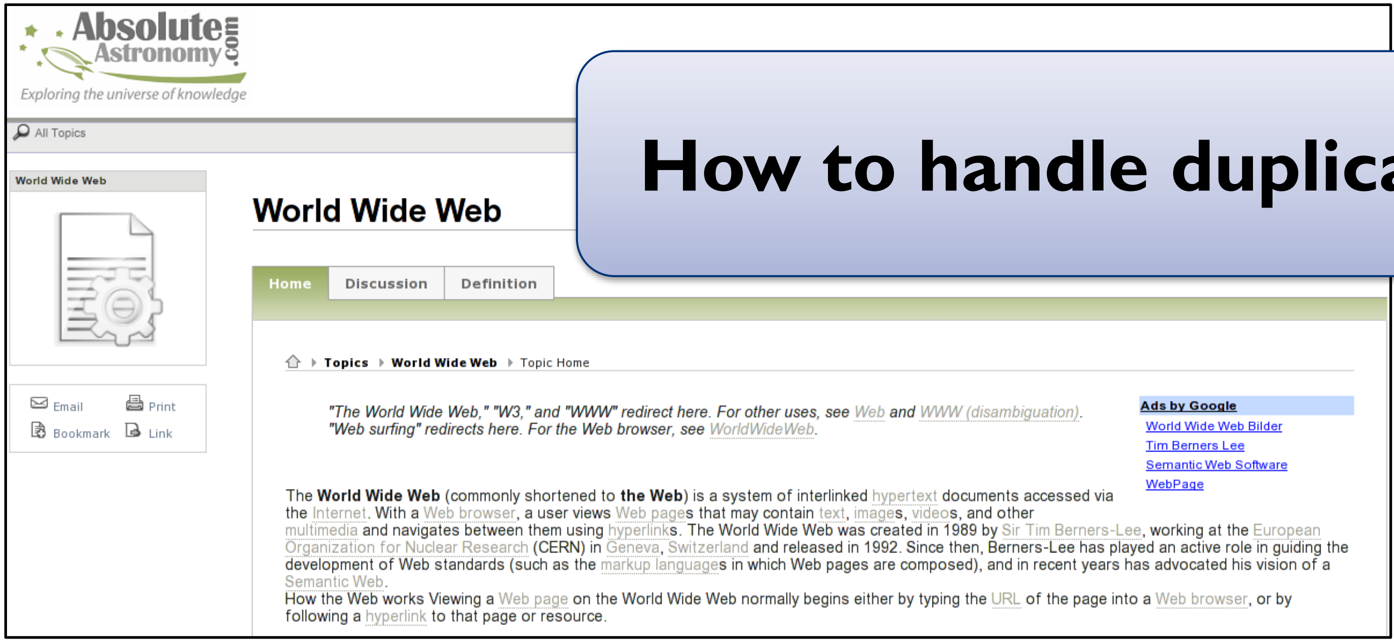
World Wide Web

From Wikipedia, the free encyclopedia
(Redirected from World wide web)

"The World Wide Web" and "WWW" redirect here. For other uses, see Web and WWW (disambiguation). "Web surfing" redirects here. For the Web browser, see WorldWideWeb.

This article contains Chinese text. Without proper rendering support, you may see question marks, boxes, or other symbols instead of Chinese characters.

The **World Wide Web** (commonly abbreviated as "**the Web**") is a system of interlinked hypertext documents accessed via the Internet. With a Web browser, one can view Web pages that may contain text, images, videos, and other multimedia and navigate between them using hyperlinks. Using concepts from earlier hypertext systems, the World Wide Web was begun in 1989 by English scientist Tim Berners-Lee, working at the European Organization for Nuclear Research (CERN) in Geneva, Switzerland. In 1990, he proposed building a "web of nodes" storing "hypertext pages" viewed by "browsers" on a network,[1] and released that web in 1992. Connected by the existing Internet, other websites were created, around the world, adding international standards for domain names & the HTML language. Since then, Berners-Lee has played an active role in guiding the development of Web standards (such as the markup languages in which Web pages are composed), and in recent years has advocated his vision of a Semantic Web.



World Wide Web

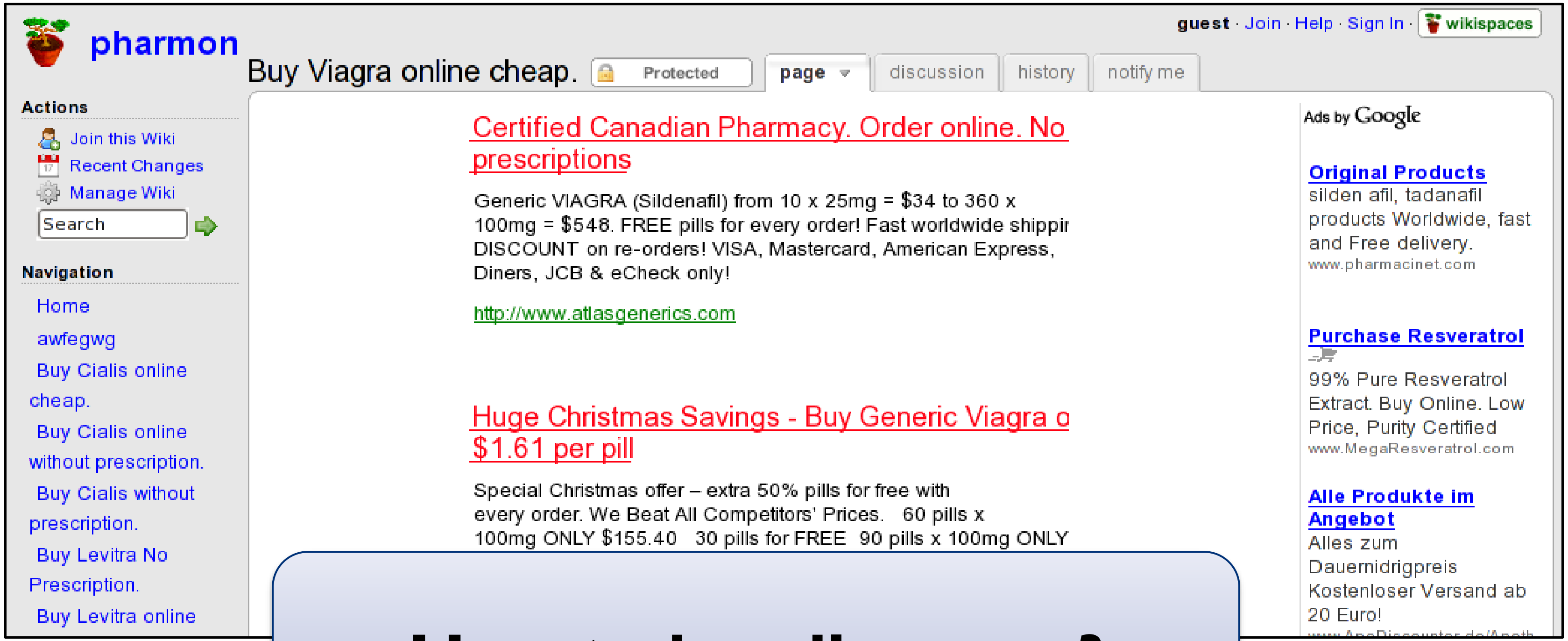
"The World Wide Web," "W3," and "WWW" redirect here. For other uses, see Web and WWW (disambiguation). "Web surfing" redirects here. For the Web browser, see WorldWideWeb.

The **World Wide Web** (commonly shortened to **the Web**) is a system of interlinked hypertext documents accessed via the Internet. With a Web browser, a user views Web pages that may contain text, images, videos, and other multimedia and navigates between them using hyperlinks. The World Wide Web was created in 1989 by Sir Tim Berners-Lee, working at the European Organization for Nuclear Research (CERN) in Geneva, Switzerland and released in 1992. Since then, Berners-Lee has played an active role in guiding the development of Web standards (such as the markup languages in which Web pages are composed), and in recent years has advocated his vision of a Semantic Web.
How the Web works Viewing a Web page on the World Wide Web normally begins either by typing the URL of the page into a Web browser, or by following a hyperlink to that page or resource.

How to handle duplicates?

How Large is the Web?

What pages counts as “the Web”?



pharmon

guest · Join · Help · Sign In · wikispaces

Buy Viagra online cheap.

Actions

- Join this Wiki
- Recent Changes
- Manage Wiki

Navigation

- Home
- awfegwg
- Buy Cialis online cheap.
- Buy Cialis online without prescription.
- Buy Cialis without prescription.
- Buy Levitra No Prescription.
- Buy Levitra online

Certified Canadian Pharmacy. Order online. No prescriptions

Generic VIAGRA (Sildenafil) from 10 x 25mg = $34 to 360 x 100mg = $548. FREE pills for every order! Fast worldwide shippir DISCOUNT on re-orders! VISA, Mastercard, American Express, Diners, JCB & eCheck only!

http://www.atlasgenerics.com

Huge Christmas Savings - Buy Generic Viagra o $1.61 per pill

Special Christmas offer – extra 50% pills for free with every order. We Beat All Competitors' Prices. 60 pills x 100mg ONLY $155.40 30 pills for FREE 90 pills x 100mg ONLY

Ads by Google

Original Products
silden afil, tadanafil products Worldwide, fast and Free delivery.
www.pharmacinet.com

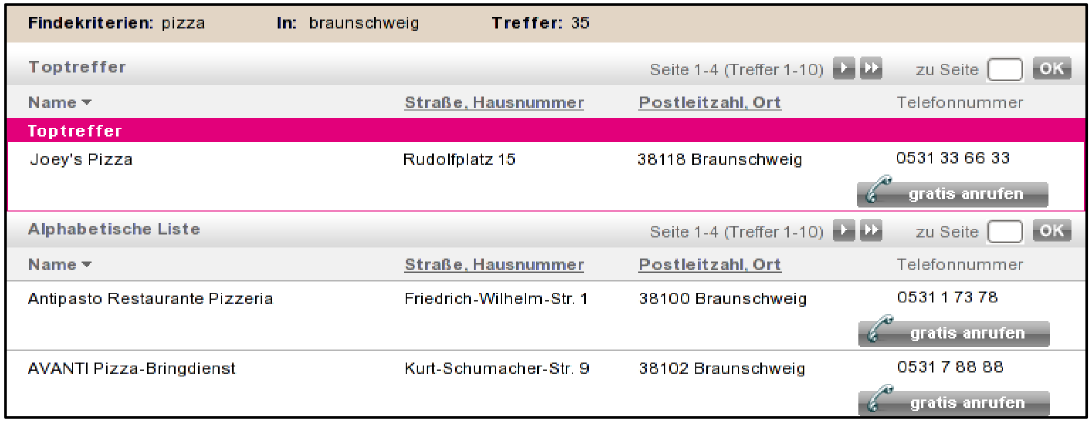
Purchase Resveratrol
99% Pure Resveratrol Extract. Buy Online. Low Price, Purity Certified
www.MegaResveratrol.com

Alle Produkte im Angebot
Alles zum Dauerniedrigpreis Kostenloser Versand ab 20 Euro!
www.ApoDiscounter.de/Apoth

How to handle spam?

How Large is the Web?

What content counts as "the Web"?



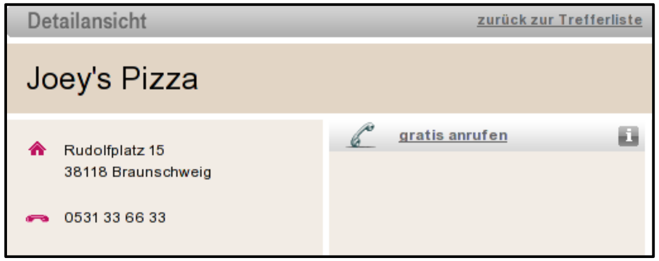
Findekriterien: pizza **In:** braunschweig **Treffer:** 35

Toptreffer Seite 1-4 (Treffer 1-10) zu Seite OK

Name	Straße, Hausnummer	Postleitzahl, Ort	Telefonnummer
Toptreffer			
Joey's Pizza	Rudolfplatz 15	38118 Braunschweig	0531 33 66 33 gratis anrufen

Alphabetische Liste Seite 1-4 (Treffer 1-10) zu Seite OK

Name	Straße, Hausnummer	Postleitzahl, Ort	Telefonnummer
Antipasto Restaurante Pizzeria	Friedrich-Wilhelm-Str. 1	38100 Braunschweig	0531 1 73 78 gratis anrufen
AVANTI Pizza-Bringdienst	Kurt-Schumacher-Str. 9	38102 Braunschweig	0531 7 88 88 gratis anrufen



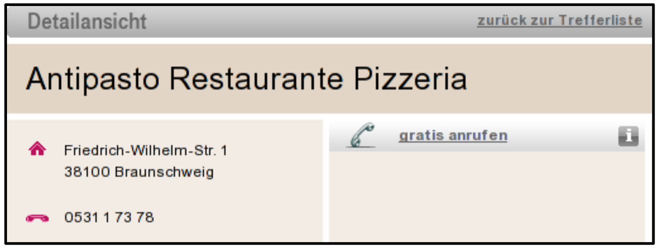
Detailansicht zurück zur Trefferliste

Joey's Pizza

Rudolfplatz 15
38118 Braunschweig

0531 33 66 33

gratis anrufen



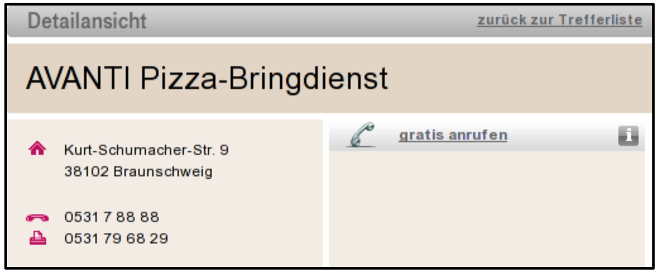
Detailansicht zurück zur Trefferliste

Antipasto Restaurante Pizzeria

Friedrich-Wilhelm-Str. 1
38100 Braunschweig

0531 1 73 78

gratis anrufen



Detailansicht zurück zur Trefferliste

AVANTI Pizza-Bringdienst

Kurt-Schumacher-Str. 9
38102 Braunschweig

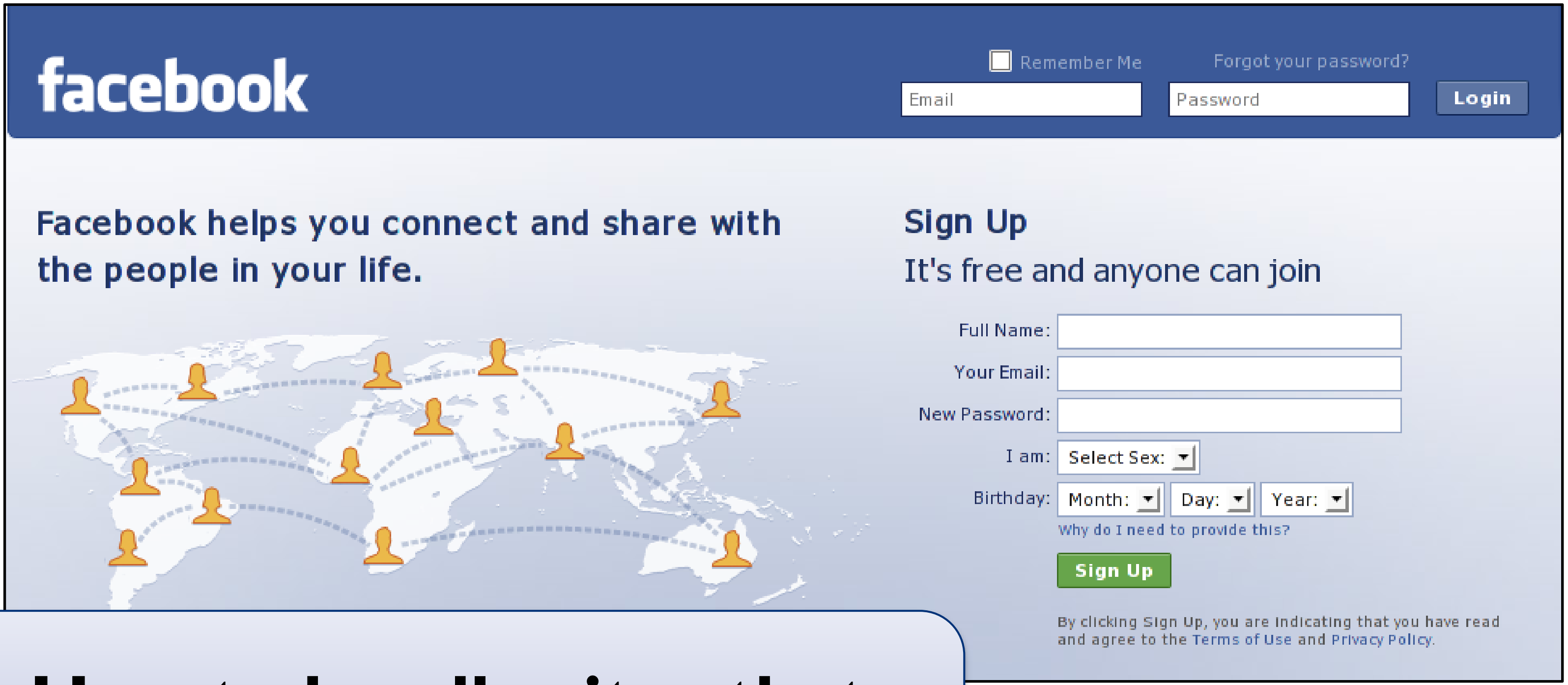
0531 7 88 88
0531 79 68 29

gratis anrufen

How many different pages should we count in this case?

How Large is the Web?

What content counts as "the Web"?



facebook

□ Remember Me

Forgot your password?

Email

Password

Login

Facebook helps you connect and share with the people in your life.

Sign Up

It's free and anyone can join

Full Name:

Your Email:

New Password:

I am: Select Sex:

Birthday: Month: Day: Year:

Why do I need to provide this?

Sign Up

By clicking Sign Up, you are indicating that you have read and agree to the Terms of Use and Privacy Policy.

How to handle sites that require users to login?

How Large is the Web?

- **Now, what pages should be counted?**
  - **Duplicates:**
    Ignore them!
  - **Spam:**
    Ignore it!
  - **Dynamic Web pages (e.g. database interfaces):**
    Count them but try to focus on the actual information;
    maybe it is better to count in megabytes instead of pages…
  - **(More or less public) private pages:**
    Count them if they can be accessed by a large number of people
- Well, now we have defined what should be counted
- **But… How to do it?**

How Large is the Web?

- **How to find all Web pages?**
  - Just follow the links…
- What about pages nobody links to?
- How to detect duplicates?
- How to detect spam?
- How to crawl Web sites with dynamic pages?
- How to access (more or less public) private pages?

A lot of interesting questions to be solved by
Web crawlers and indexers!
Let’s answer them next week…

How Large is the Web?

- Let's assume for now, that we have some Web crawler that can automatically solve all these problems as good as currently possible
- **Then, calculating the Web's size is easy:** Simply crawl the complete Web and count its number of pages or its size in megabytes!
- **Bad news:** **This doesn't work** due to the Web's enormous size
  - It would either take forever or require an enormous effort
  - The Web has changed completely until the crawl is finished
- Any better ideas?

How Large is the Web?

- A better approach is called "mark and recapture": Take two (large) **random samples** of the Web and compute the Web's total size by looking at the **overlap**
- **Idea:**
  - Let *f* be the number of pages found in the **first crawl**
  - Let s be the number of pages found in the **second crawl**
  - Let *b* be the number of pages found in **both crawls**
  - Then, the estimation of size is:
  - $\frac{Web\ Pages\ in\ first\ Crawl\ (f)}{Total\ Size\ (t)} = \frac{Pages\ in\ Both\ Crawls\ (b)}{Web\ Pages\ in\ Second\ Crawl\ (S)}$
  - Taken together, we get ***t* = *f* · *s* / *b***

How Large is the Web?

- In practice, one takes **random samples** from the **index** of different search engines
- Of course, we cannot assume anymore that these draws have been **independent**
- There are more advanced methods to account for this…

- In 2005, the Web has been estimated to contain at least 11.5 billion pages
- **Nobody knows exactly…**

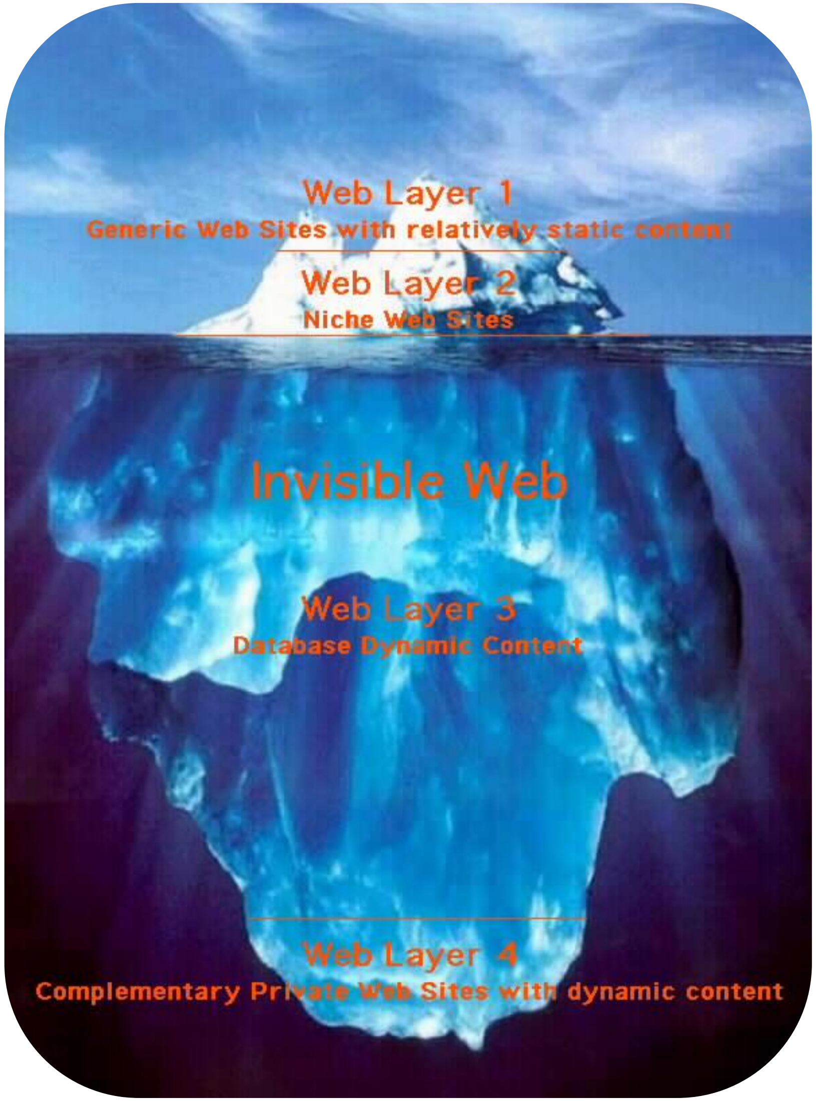
How Large is the Web?

- Of course, these estimates only cover the so-called **"surface web,"** i.e., the part of the Web that can be accessed automatically by **current Web crawlers**
  - Even today's best Web crawlers cannot find pages without in-links or all pages that have been generated dynamically…
- The term **"Deep Web"** refers to all web pages that currently are not indexed by any Web search engine
- There are different estimates on the Deep Web's size
  - **The Deep Web is 15−500x as large as the surface Web**

How Large is the Web?

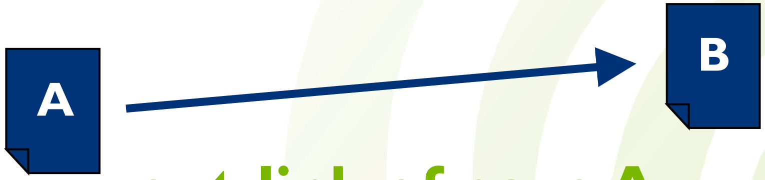
Some types of “deep resources”:

- Dynamic content that cannot be accessed automatically, e.g. pages that are generated dynamically after filling out **Web forms**
- Unlinked or private content
- “Scripted” content, which requires code execution (e.g. Java, JavaScript, or Flash)
- “Strange” file formats not handled by current search engines


Web Layer 1
Generic Web Sites with relatively static content
Web Layer 2
Niche Web Sites
Invisible Web
Web Layer 3
Database Dynamic Content
Web Layer 4
Complementary Private Web Sites with dynamic content

The Web Graph

- We can view the **static Web** consisting of static HTML pages together with the hyperlinks between them as a **directed graph**
  - Each Web page is a node
  - Each hyperlink is a directed edge
- The hyperlinks into a page are called **in-links**
- The hyperlinks out of a page are called **out-links**



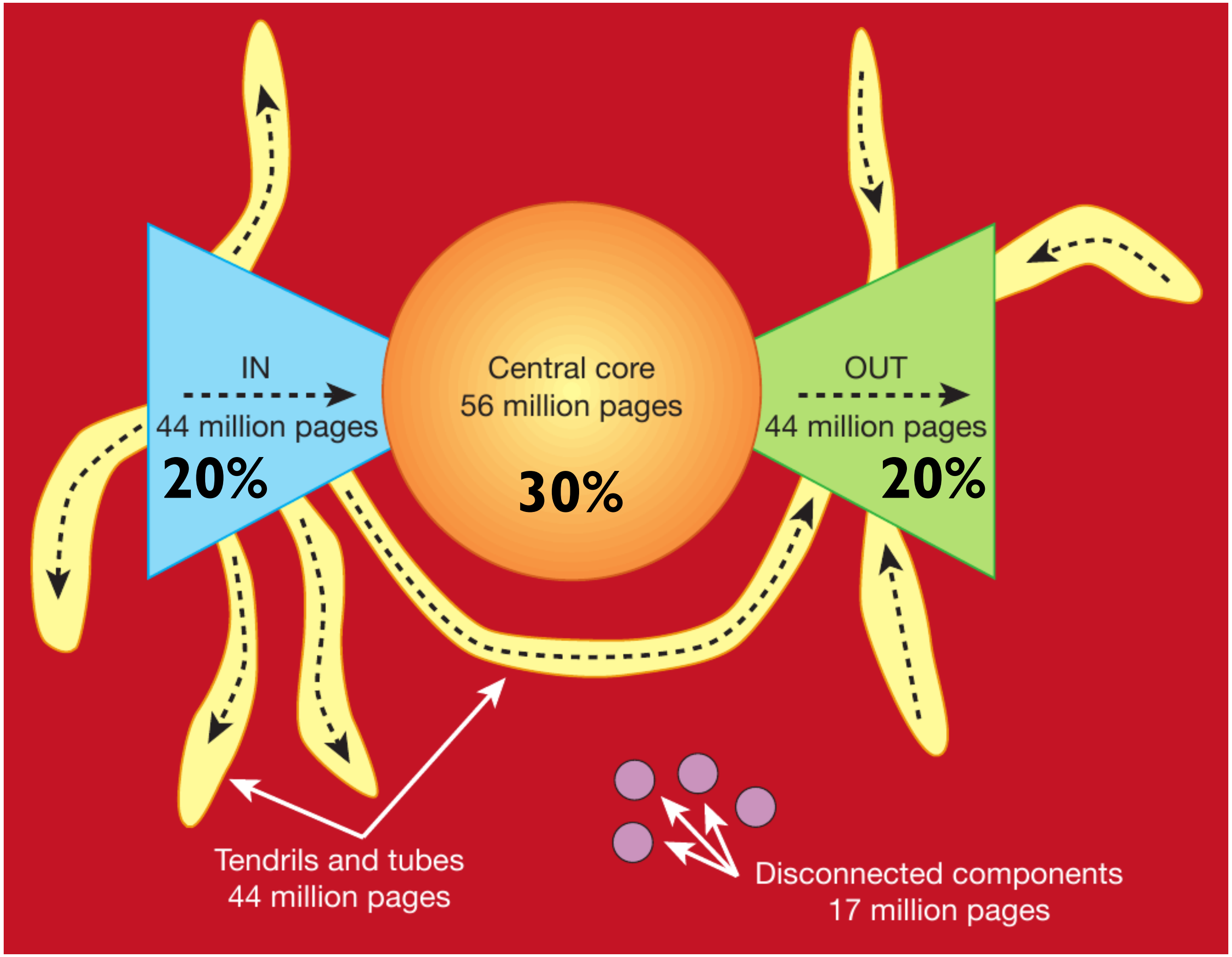
out-link of page A

in-link of page B

The Web Graph

- There is evidence that these links are not randomly distributed
- The distribution of in-links seems to follow a **power law**
  - The total number of pages having exactly k in-links is proportional to $1 / k^{2.1}$
- Furthermore, several studies have suggested that the Web graph has a **bowtie shape:**

The Web Graph



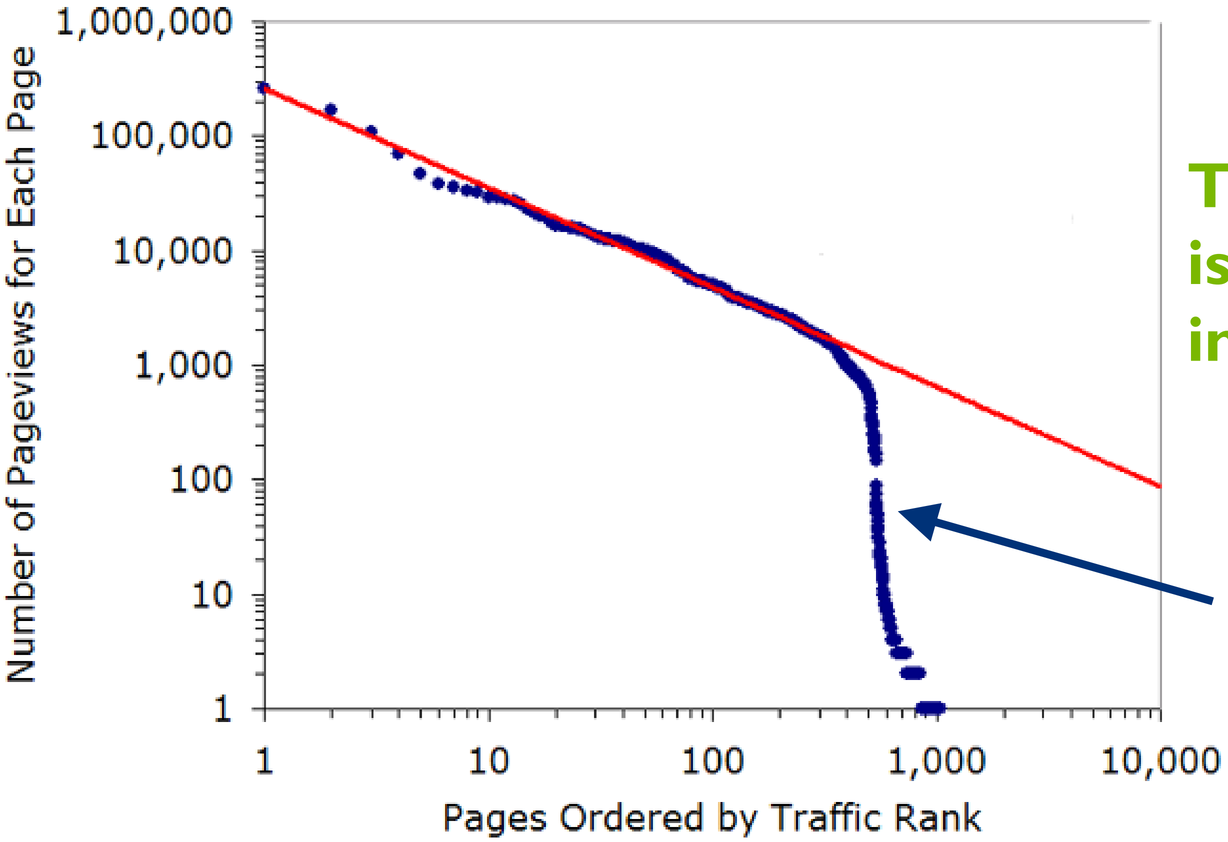
Note: the exact numbers given are as of 2000

Introduction to Web Retrieval

1. Web Retrieval vs. Classical IR
2. What Does the Web Look Like?
3. **How Do Users Use the Web?**

Page Popularity

Page popularity is approximately Zipf distributed:



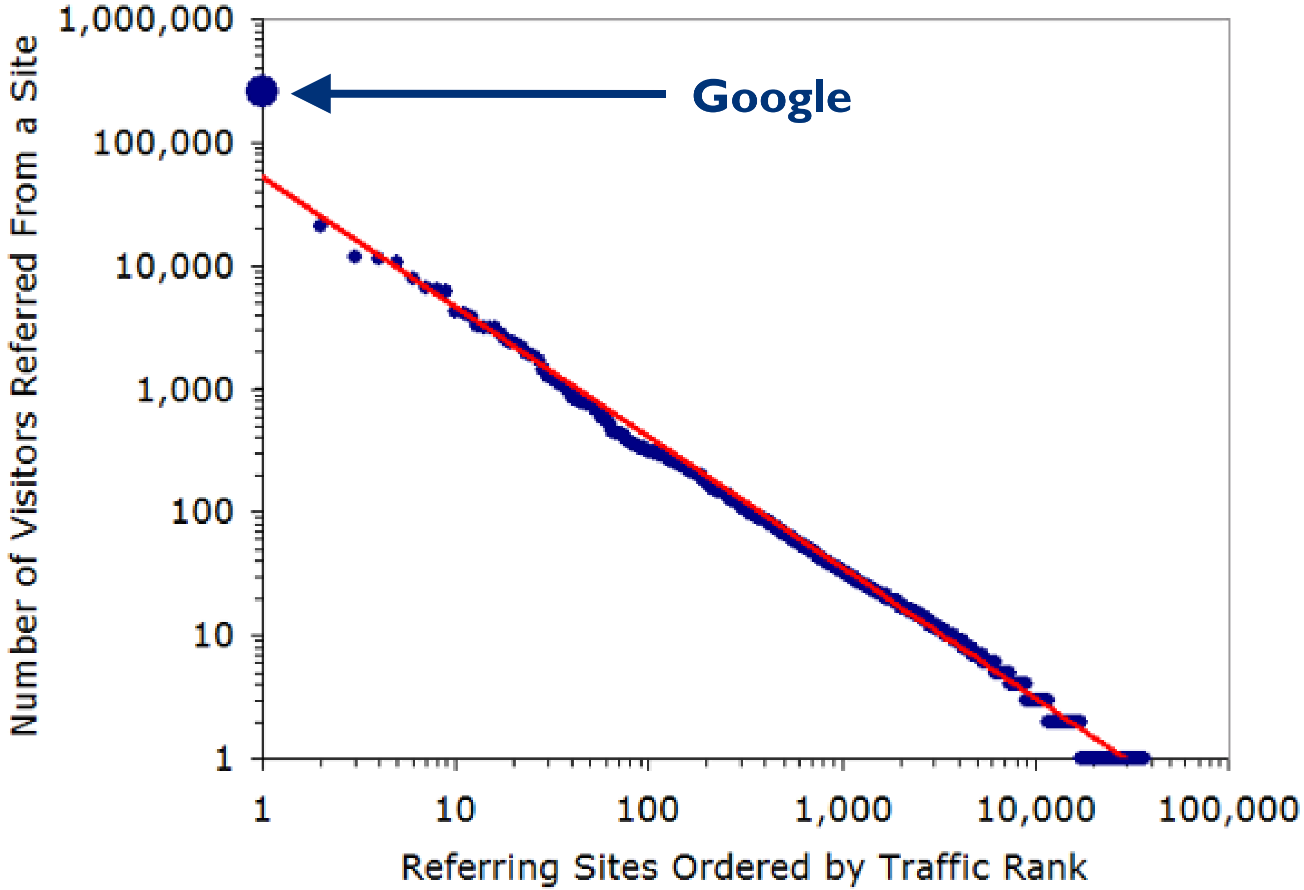
Source: useit.com

The Zipf curve is a straight line in log–log scale

The end of the "long tail" is absent

Incoming Traffic

Incoming traffic from other sites follows Zipf's law:



Source: useit.com

Search Engine Queries

- **Several studies analyzed users' query behavior:**
  - The **average length** of a query is **2.4 terms**
  - About **half of all queries** consist of a **single term**
  - About **half of the users** looked only at the **first 20 results**
  - Less than 5% of users use advanced search features (e.g., Boolean operators)
  - About **20%** of all queries contain a **geographic term**
  - About **a third of the queries** from the same user were **repeated queries;** about 90% of the time the user would click on the same result
  - **Term frequency distributions** conform to the **power law**

Next Lecture

- Web crawling
- Duplicate detection