Big Data and Data Mining

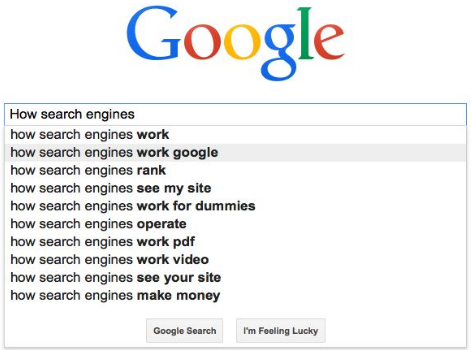
Web Information Retrieval

Flavio Bertini

flavio.bertini@unipr.it

How Search Engines Work

The user inputs a query



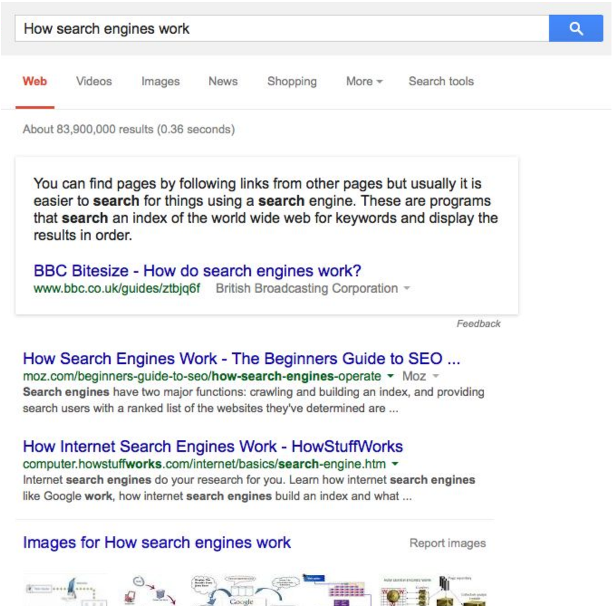
How search engines

how search engines **work**
how search engines **work google**
how search engines **rank**
how search engines **see my site**
how search engines **work for dummies**
how search engines **operate**
how search engines **work pdf**
how search engines **work video**
how search engines **see your site**
how search engines **make money**

The search engine search the query terms in its very large **index**

About 83,900,000 results (0.36 seconds)

Millions of results found are **ranked** accordingly to a notion of **relevance** and are shown as a sorted list



About 83,900,000 results (0.36 seconds)

You can find pages by following links from other pages but usually it is easier to **search** for things using a **search** engine. These are programs that **search** an index of the world wide web for keywords and display the results in order.

BBC Bitesize - How do search engines work?
www.bbc.co.uk/guides/ztbjq6f British Broadcasting Corporation

How Search Engines Work - The Beginners Guide to SEO ...
moz.com/beginners-guide-to-seo/**how-search-engines**-operate Moz
Search engines have two major functions: crawling and building an index, and providing search users with a ranked list of the websites they've determined are ...

How Internet Search Engines Work - HowStuffWorks
computer.howstuff**works**.com/internet/basics/**search**-engine.htm
Internet **search engines** do your research for you. Learn how internet **search engines** like Google **work**, how internet **search engines** build an index and what ...

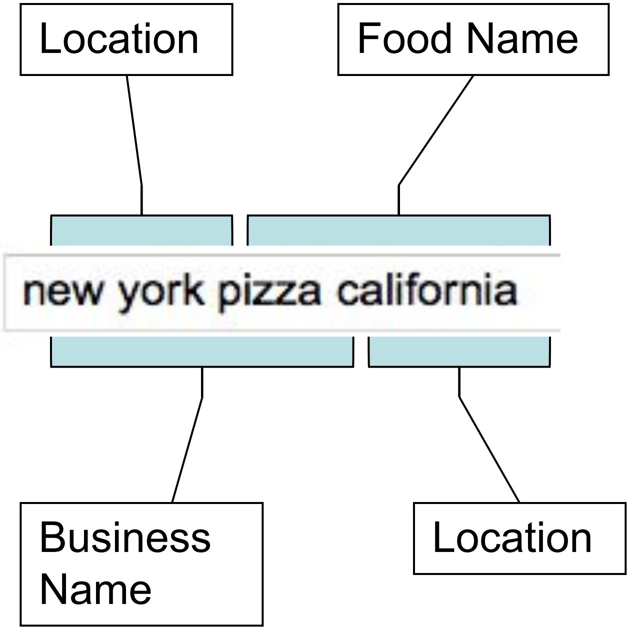
Images for How search engines work

Automated programs (a.k.a. spiders or **crawlers**) scan all pages in the web to build an index

What is Relevant?

- How can a search engine know what the user is looking for?
  - **Information need:** *a desire to locate and obtain information to satisfy a conscious or unconscious need* [Wikipedia]
  - **User query:** the set of consecutive terms formulated by a user to express his information need
  - **Query intent:** the task, goal or intent of a user, expressed by a query. The same query formulated by different users may be the result of different intents

Query intent can be ambiguous:



Location

Food Name

new york pizza california

Business Name

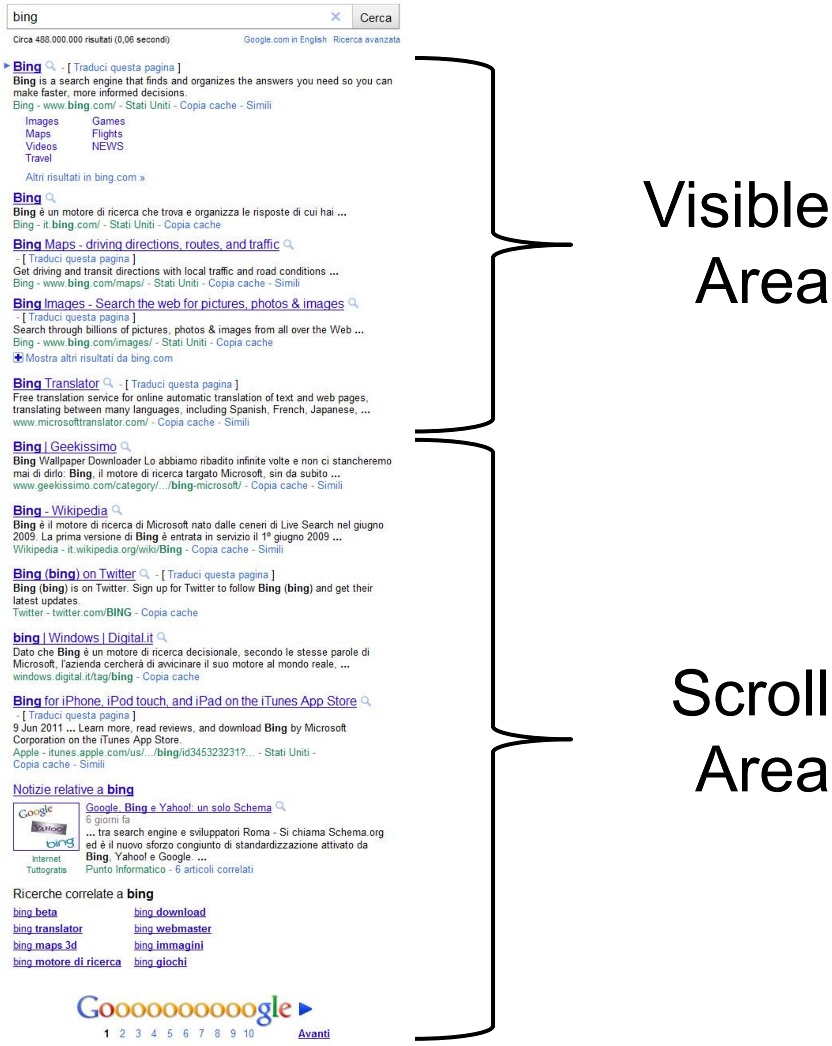
Location

Ranking: Google PageRank

- A relevance notion based solely on term frequencies is not enough to rank billions of documents
- Complex measures are applied to evaluate the quality, reliability and authority of web pages:
    - Measures based **on topological network properties** (such as Google PageRank)
    - Measures based **on different field boosting** (Title, subtitle, body have different weight)
    - Measures based **on semantics and time-space properties** (freshness of a page)
- PageRank set up a **rich-get-richer loop**, whereby few sites dominate the top ranks

Search Engine Results Page

- A Search Engine Results Page (SERP) shows by default **10 results**
- Of these results, on average, the first **5 are visible** without scrolling down
- We will take in consideration the **user behavior on the first SERP** of search engines (the first 10 results)



Circa 488.000.000 risultati (0,06 secondi) Google.com in English Ricerca avanzata

Bing - [Traduci questa pagina]
Bing is a search engine that finds and organizes the answers you need so you can make faster, more informed decisions.
Bing - www.**bing**.com/ - Stati Uniti - Copia cache - Simili

Images Maps Videos Travel Games Flights NEWS

Altri risultati in bing.com »

Bing
Bing è un motore di ricerca che trova e organizza le risposte di cui hai ...
Bing - it.**bing**.com/ - Stati Uniti - Copia cache

Bing Maps - driving directions, routes, and traffic - [Traduci questa pagina]
Get driving and transit directions with local traffic and road conditions ...
Bing - www.**bing**.com/maps/ - Stati Uniti - Copia cache - Simili

Bing Images - Search the web for pictures, photos & images - [Traduci questa pagina]
Search through billions of pictures, photos & images from all over the Web ...
Bing - www.**bing**.com/images/ - Stati Uniti - Copia cache

Mostra altri risultati da bing.com

Bing Translator - [Traduci questa pagina]
Free translation service for online automatic translation of text and web pages, translating between many languages, including Spanish, French, Japanese, ...
www.microsofttranslator.com/ - Copia cache - Simili

Bing | Geekissimo
Bing Wallpaper Downloader Lo abbiamo ribadito infinite volte e non ci stancheremo mai di dirlo: **Bing**, il motore di ricerca targato Microsoft, sin da subito ...
www.geekissimo.com/category/.../**bing**-microsoft/ - Copia cache - Simili

Bing - Wikipedia
Bing è il motore di ricerca di Microsoft nato dalle ceneri di Live Search nel giugno 2009. La prima versione di **Bing** è entrata in servizio il 1º giugno 2009 ...
Wikipedia - it.wikipedia.org/wiki/**Bing** - Copia cache - Simili

Bing (**bing**) on Twitter - [Traduci questa pagina]
Bing (**bing**) is on Twitter. Sign up for Twitter to follow **Bing** (**bing**) and get their latest updates.
Twitter - twitter.com/**BING** - Copia cache

bing | Windows | Digital.it
Dato che **Bing** è un motore di ricerca decisionale, secondo le stesse parole di Microsoft, l'azienda cercherà di avvicinare il suo motore al mondo reale, ...
windows.digital.it/tag/**bing** - Copia cache

Bing for iPhone, iPod touch, and iPad on the iTunes App Store - [Traduci questa pagina]
9 Jun 2011 ... Learn more, read reviews, and download **Bing** by Microsoft Corporation on the iTunes App Store.
Apple - itunes.apple.com/us/.../**bing**/id345323231?... - Stati Uniti - Copia cache - Simili

Notizie relative a **bing**

Google, **Bing** e Yahoo!: un solo Schema
6 giorni fa
... tra search engine e sviluppatori Roma - Si chiama Schema.org ed è il nuovo sforzo congiunto di standardizzazione attivato da **Bing**, Yahoo! e Google. ...
Punto Informatico - 6 articoli correlati

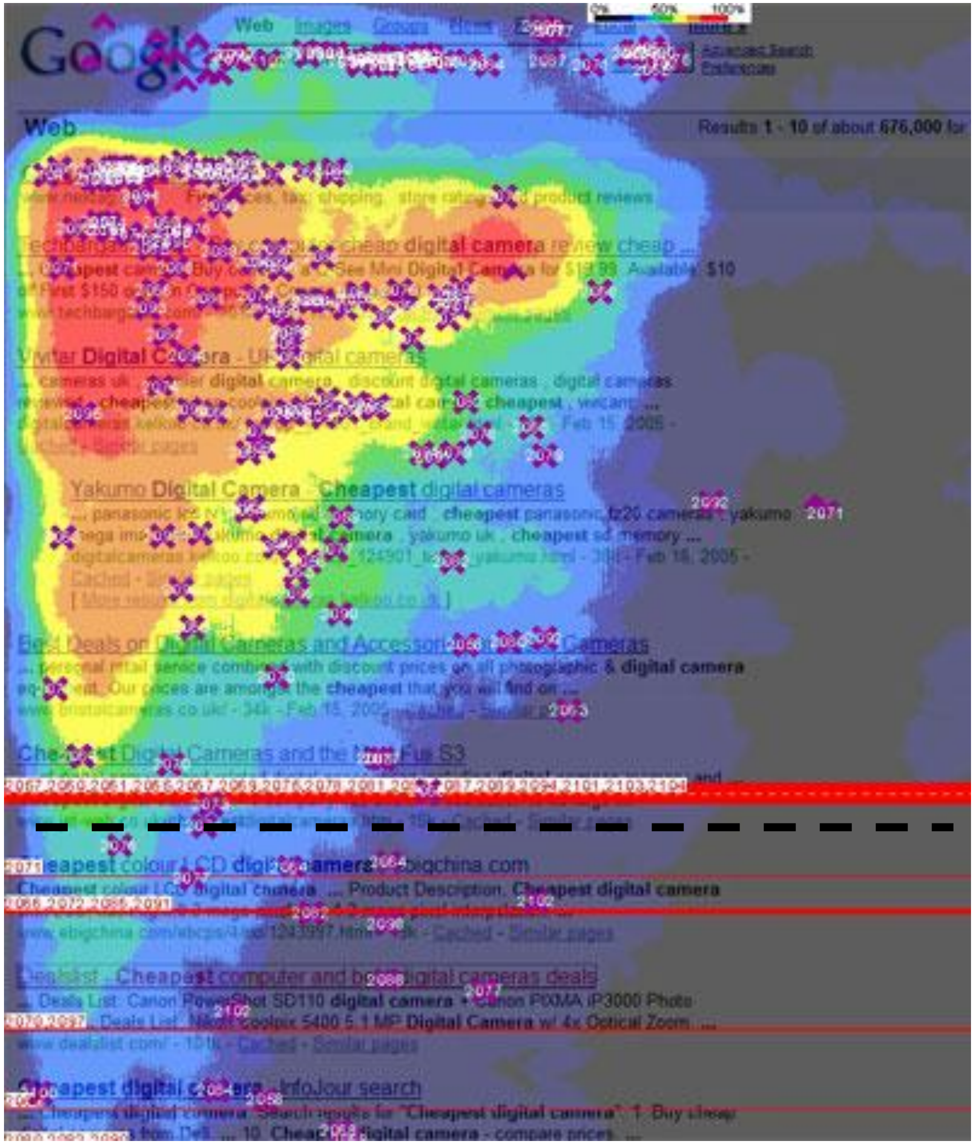
Internet Tuttogratis

Ricerche correlate a **bing**

bing **beta**	bing **download**
bing **translator**	bing **webmaster**
bing **maps 3d**	bing **immagini**
bing **motore di ricerca**	bing **giochi**

Gooooooooooogle ►
1 2 3 4 5 6 7 8 9 10 Avanti

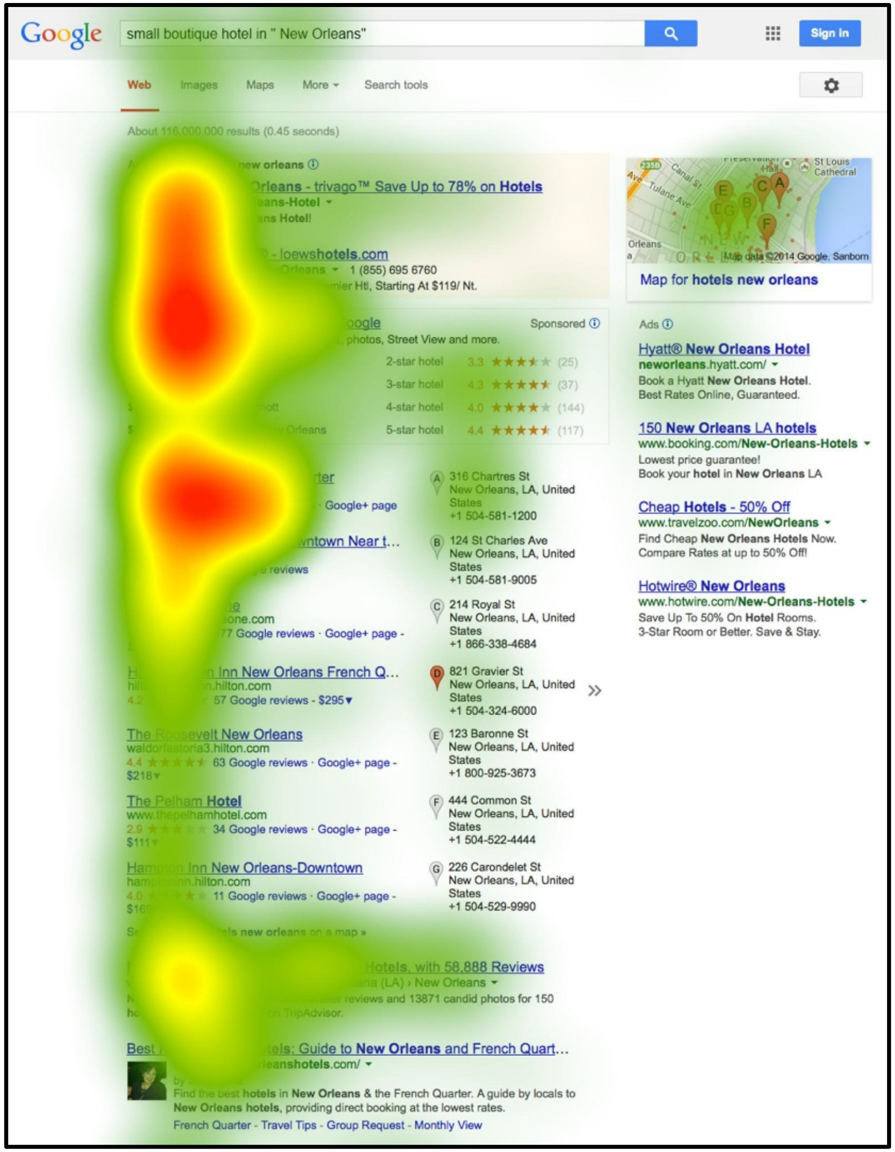
Eye-tracking in 2005: Golden Triangle


Google
Web Images Groups News
0% 50% 100%
Web
Results 1 - 10 of about 676,000 for
Yakumo Digital Camera - Cheapest digital cameras
Best Deals on Digital Cameras and Accessories
Cheapest Digital Cameras and the
Dealslist - Cheapest computer and best digital cameras deals
Cheapest digital camera - InfoJour search

- Color 0% 50% 100% represents percentage of time spent looking the area
- Purple x represents a mouse click on the page
- Dotted line - - - represents where the page breaks on the computer screen (Visible area/Scroll area)
- Red lines ▬ indicate how far down the page scrolled before leaving the page

Results in 1st position are clicked more than 10 times results on 6th position

Eye-tracking in 2014: F-shape Scan



About 116,000,000 results (0.45 seconds)

new orleans ⓘ

Orleans - trivago™ Save Up to 78% on Hotels
eans-Hotel
ans Hotel!

- loewshotels.com
Orleans 1 (855) 695 6760
mier Htl, Starting At $119/ Nt.

oogle — Sponsored ⓘ
, photos, Street View and more.

2-star hotel	3.3 ★★★★★	(25)
3-star hotel	4.3 ★★★★★	(37)
4-star hotel	4.0 ★★★★★	(144)
5-star hotel	4.4 ★★★★★	(117)

rter · Google+ page
ntown Near t…
reviews
one.com
77 Google reviews · Google+ page -
n Inn New Orleans French Q…
n.hilton.com
4.2 57 Google reviews - $295▾

The Roosevelt New Orleans
waldorfastoria3.hilton.com
4.4 ★★★★★ 63 Google reviews · Google+ page -
$218▾

The Pelham Hotel
www.thepelhamhotel.com
2.9 ★★★★★ 34 Google reviews · Google+ page -
$111▾

Hampton Inn New Orleans-Downtown
hamptoninn.hilton.com
4.0 ★★★★★ 11 Google reviews · Google+ page -
$16

ls new orleans on a map »

A	316 Chartres St New Orleans, LA, United States +1 504-581-1200
B	124 St Charles Ave New Orleans, LA, United States +1 504-581-9005
C	214 Royal St New Orleans, LA, United States +1 866-338-4684
D	821 Gravier St New Orleans, LA, United States +1 504-324-6000
E	123 Baronne St New Orleans, LA, United States +1 800-925-3673
F	444 Common St New Orleans, LA, United States +1 504-522-4444
G	226 Carondelet St New Orleans, LA, United States +1 504-529-9990

Hotels, with 58,888 Reviews
ana (LA) › New Orleans
reviews and 13871 candid photos for 150
on TripAdvisor.

Best tels: Guide to New Orleans and French Quart…
leanshotels.com/
by
Find the best hotels in New Orleans & the French Quarter. A guide by locals to New Orleans hotels, providing direct booking at the lowest rates.
French Quarter - Travel Tips - Group Request - Monthly View

Map for hotels new orleans

Ads ⓘ

Hyatt® New Orleans Hotel
neworleans.hyatt.com/
Book a Hyatt New Orleans Hotel.
Best Rates Online, Guaranteed.

150 New Orleans LA hotels
www.booking.com/New-Orleans-Hotels
Lowest price guarantee!
Book your hotel in New Orleans LA

Cheap Hotels - 50% Off
www.travelzoo.com/NewOrleans
Find Cheap New Orleans Hotels Now.
Compare Rates at up to 50% Off!

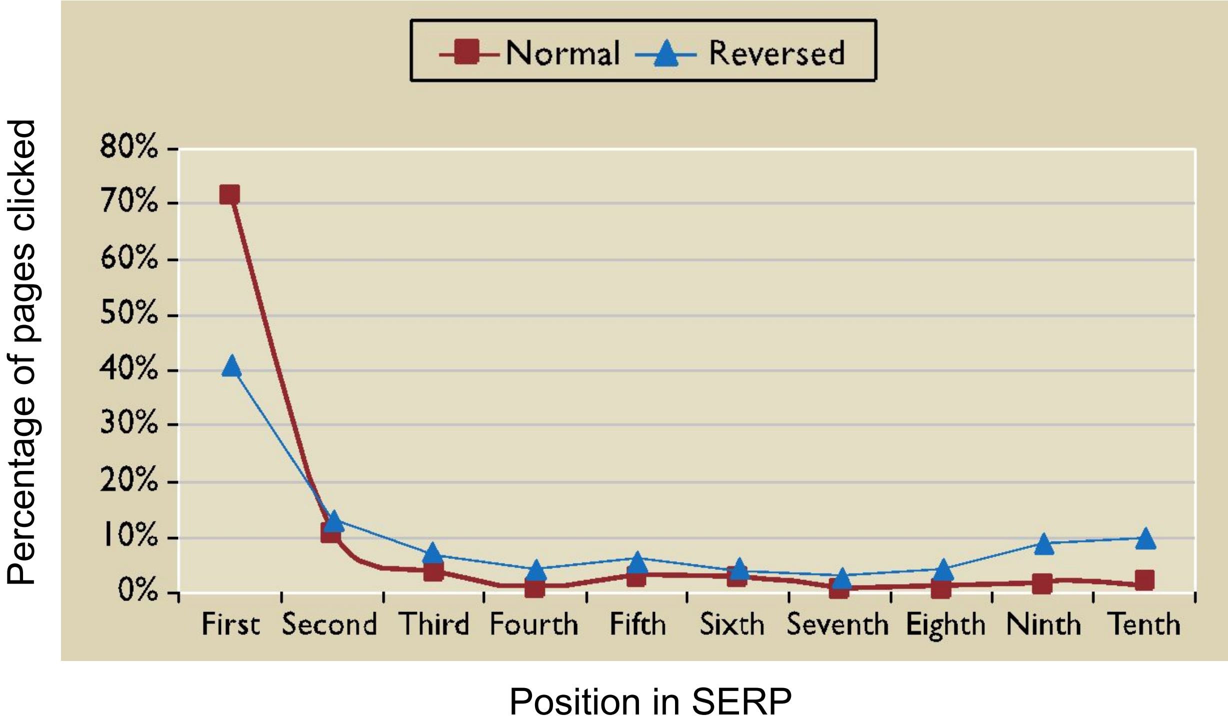
Hotwire® New Orleans
www.hotwire.com/New-Orleans-Hotels
Save Up To 50% On Hotel Rooms.
3-Star Room or Better. Save & Stay.

- In 2005, all results were links to a website. Now, we have a variety of results, and the results page layout varies from search to search.
- The first (vertical) glance allows to determine what categories Google has decided to show.
- Then, in a second (horizontal) glance, we go back and deep to the most relevant chunks of results.

The time spent scanning each result is halved: in 2005 was 2.5 seconds, in 2014 is 1.2 seconds.

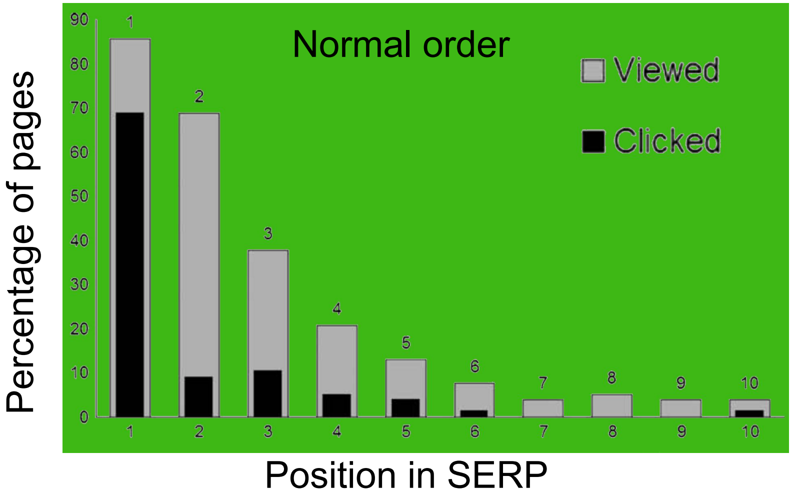
Result Positioning Bias

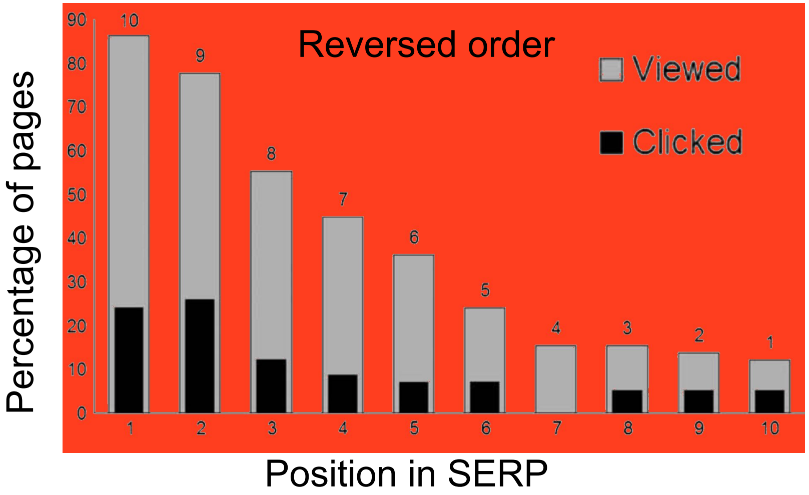
- When searching for information, the average searchers do not judge systematically all the results, instead they **simply click on top results**
- Studies have been conducted comparing the users’ responses when they received results lists in **normal ordering versus a reversed order**



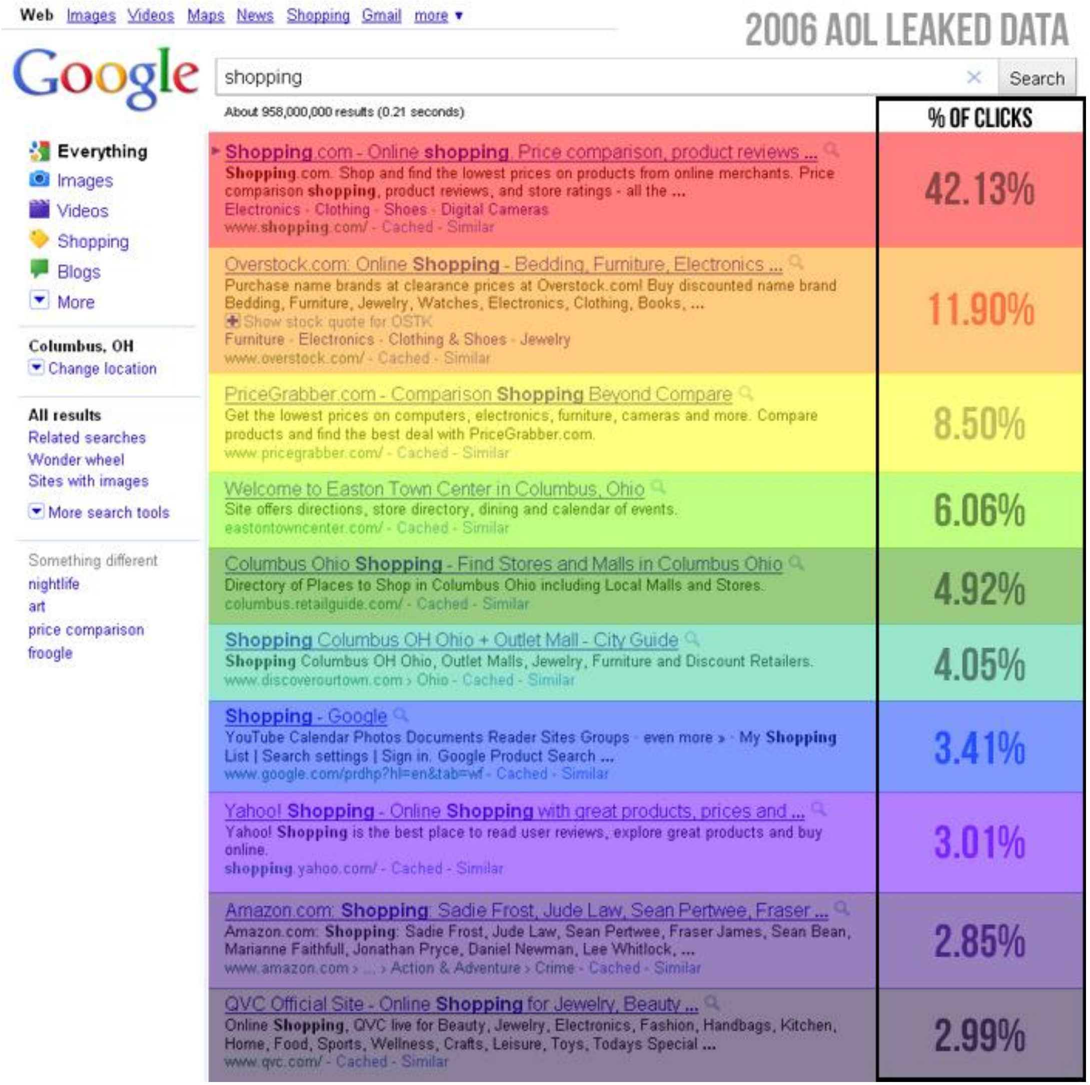
Click and Eye-tracking Bias

- View percentage and mouse clicks are compared in normal vs. reversed order of the first 10 results
- Even though results show some user awareness, **excessive trust in ranking** algorithms may negatively affect smaller positioned websites





Position Impact on Business



2006 AOL LEAKED DATA

Web Images Videos Maps News Shopping Gmail more ▾

Google

shopping × Search

About 958,000,000 results (0.21 seconds)

Everything
Images
Videos
Shopping
Blogs
More

Columbus, OH
Change location

All results
Related searches
Wonder wheel
Sites with images
More search tools

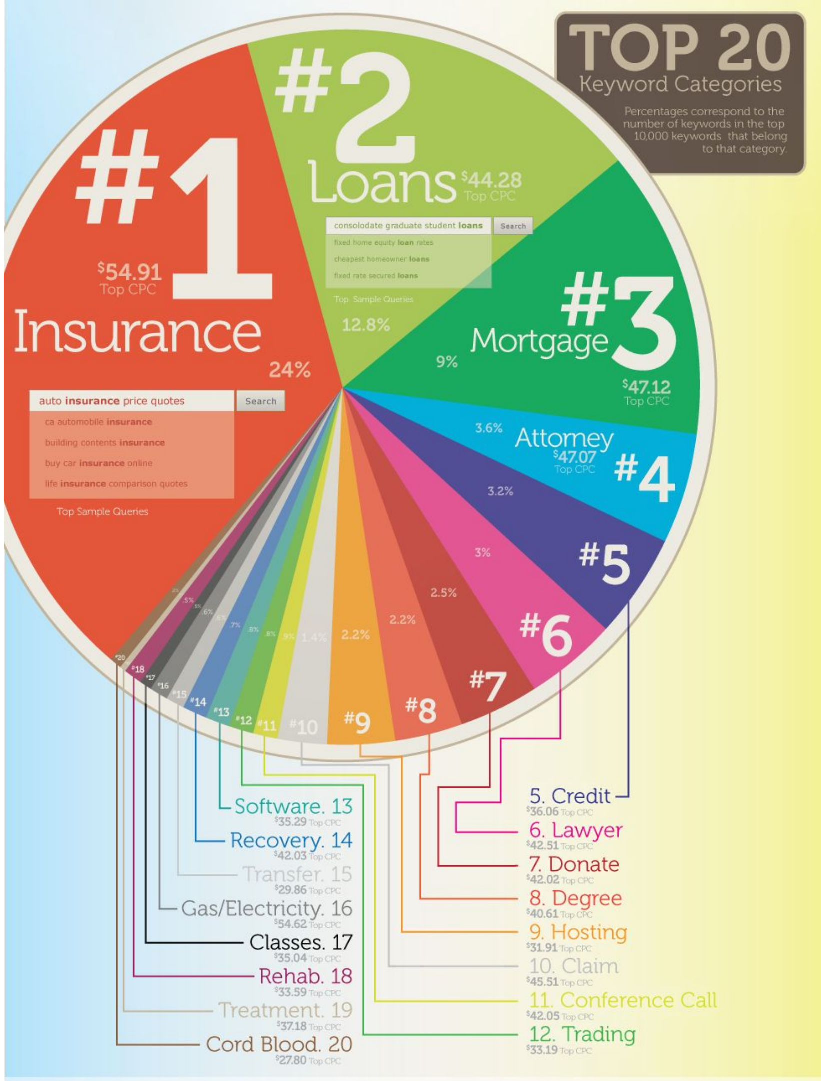
Something different
nightlife
art
price comparison
froogle

	% OF CLICKS
Shopping.com - Online **shopping**, Price comparison, product reviews ... **Shopping**.com. Shop and find the lowest prices on products from online merchants. Price comparison **shopping**, product reviews, and store ratings - all the ... Electronics - Clothing - Shoes - Digital Cameras www.**shopping**.com/ - Cached - Similar	42.13%
Overstock.com: Online **Shopping** - Bedding, Furniture, Electronics ... Purchase name brands at clearance prices at Overstock.com! Buy discounted name brand Bedding, Furniture, Jewelry, Watches, Electronics, Clothing, Books, ... Show stock quote for OSTK Furniture - Electronics - Clothing & Shoes - Jewelry www.overstock.com/ - Cached - Similar	11.90%
PriceGrabber.com - Comparison **Shopping** Beyond Compare Get the lowest prices on computers, electronics, furniture, cameras and more. Compare products and find the best deal with PriceGrabber.com. www.pricegrabber.com/ - Cached - Similar	8.50%
Welcome to Easton Town Center in Columbus, Ohio Site offers directions, store directory, dining and calendar of events. eastontowncenter.com/ - Cached - Similar	6.06%
Columbus Ohio **Shopping** - Find Stores and Malls in Columbus Ohio Directory of Places to Shop in Columbus Ohio including Local Malls and Stores. columbus.retailguide.com/ - Cached - Similar	4.92%
Shopping Columbus OH Ohio + Outlet Mall - City Guide **Shopping** Columbus OH Ohio, Outlet Malls, Jewelry, Furniture and Discount Retailers. www.discoverourtown.com › Ohio - Cached - Similar	4.05%
Shopping - Google YouTube Calendar Photos Documents Reader Sites Groups - even more » - My **Shopping** List | Search settings | Sign in. Google Product Search ... www.google.com/prdhp?hl=en&tab=wf - Cached - Similar	3.41%
Yahoo! **Shopping** - Online **Shopping** with great products, prices and ... Yahoo! **Shopping** is the best place to read user reviews, explore great products and buy online. **shopping**.yahoo.com/ - Cached - Similar	3.01%
Amazon.com: **Shopping**: Sadie Frost, Jude Law, Sean Pertwee, Fraser ... Amazon.com: **Shopping**: Sadie Frost, Jude Law, Sean Pertwee, Fraser James, Sean Bean, Marianne Faithfull, Jonathan Pryce, Daniel Newman, Lee Whitlock, ... www.amazon.com › ... › Action & Adventure › Crime - Cached - Similar	2.85%
QVC Official Site - Online **Shopping** for Jewelry, Beauty ... Online **Shopping**, QVC live for Beauty, Jewelry, Electronics, Fashion, Handbags, Kitchen, Home, Food, Sports, Wellness, Crafts, Leisure, Toys, Todays Special ... www.qvc.com/ - Cached - Similar	2.99%

- Results in 1st position are clicked more than 10 times results on 6th position
- Suppose company A and company B are respectively on the 1st and 6th position for a valuable business keyword
- B should buy 10 times the clicks that A gets to obtain the same traffic!
- **How much does a click cost?**

[BMW given Google 'death penalty', 2006]

Click Economic Value



- In web advertising, the cost of a single click, also known as Cost-Per-Click (CPC), is **1$ on average**. However...
  - On Google some keywords can be really expensive (Google most expensive keywords 2011):
    - **Insurance: 54.91$**
    - Mortgage: 47.12$
    - Attorney: 47.07
    - Claim: 45.51$
    - Loans: 44.28$
    - Lawyer: 42.51$
  - Bing is even more expensive (Bing most expensive keywords 2015):
    - **Lawyers: 109.21$**
    - Attorney: 101.77$
    - Structured settlements: 78.39$

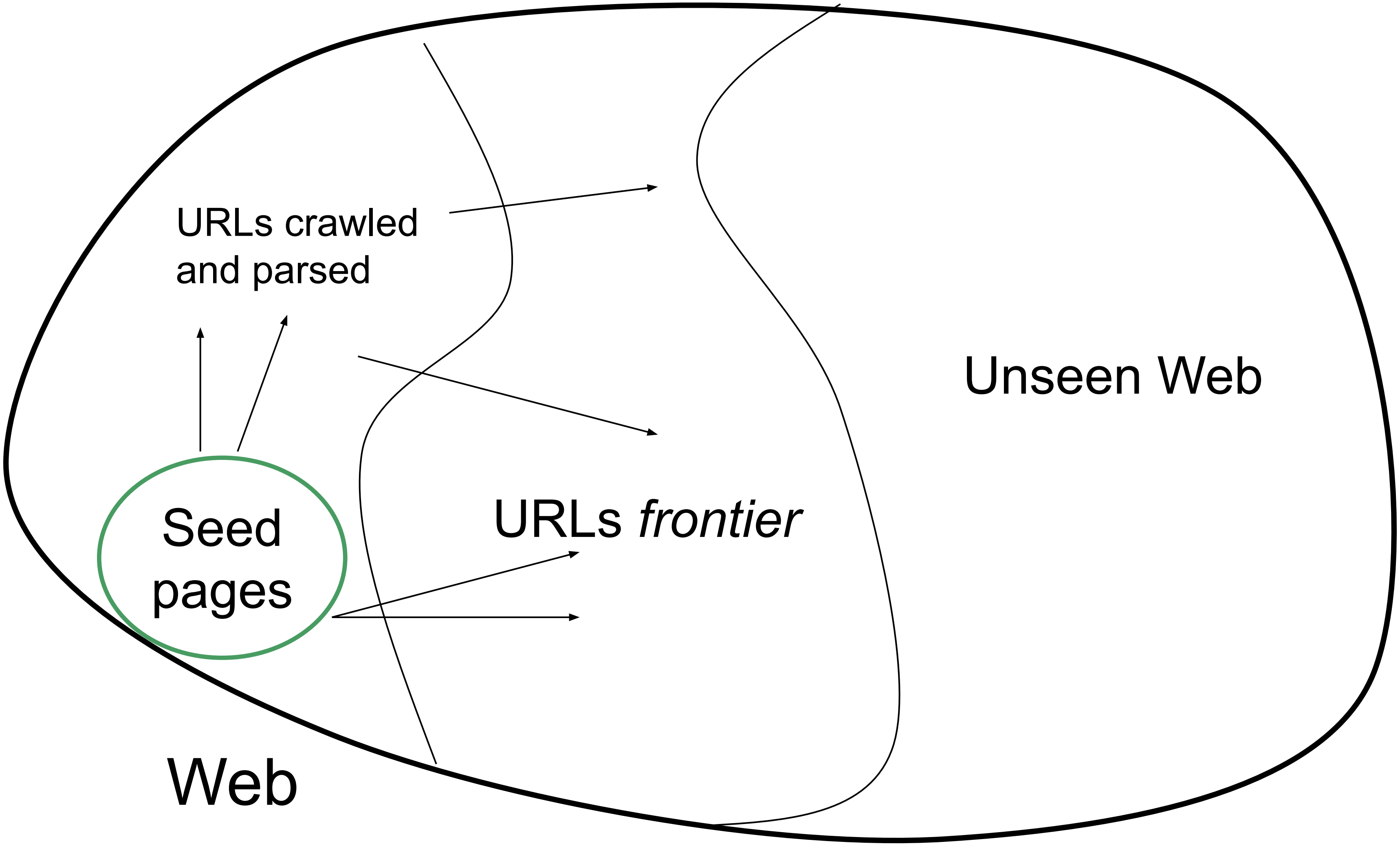
Web Crawling

- Web crawling is the process by which **we gather pages from the Web**
- Goal: **quickly** and efficiently gather as many useful Web pages as possible, together with the **link structure** that interconnects them

Basic Crawler Operation

- A crawler (a.k.a. spider)
  - Begin with known “seed” URLs
  - Fetch and parse them
    - Extract URLs they point to
    - Place the extracted URLs on a queue (the URLs *frontier*)
  - Fetch each URL on the frontier and repeat

Crawling Picture



URLs crawled and parsed

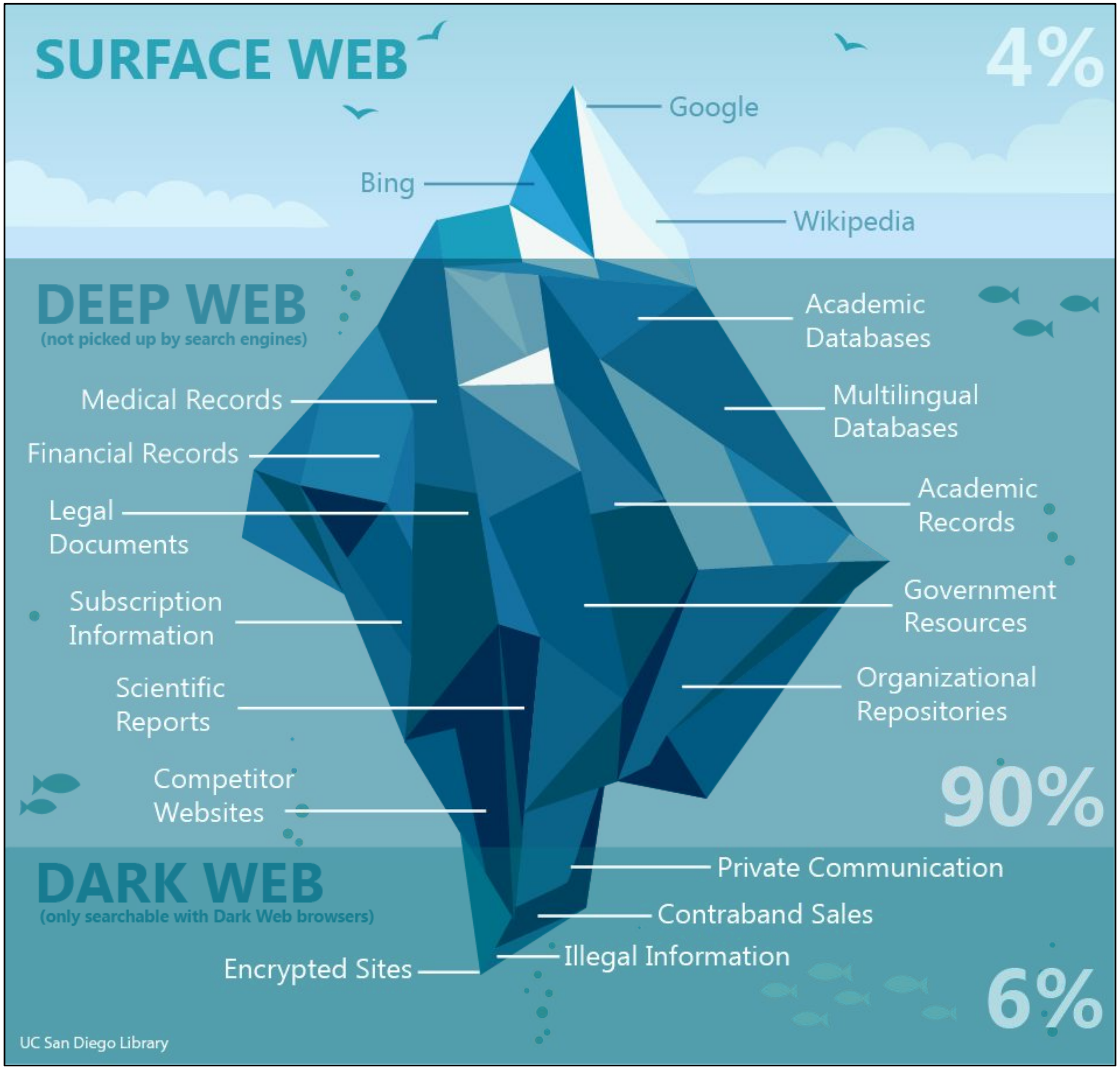
Unseen Web

URLs *frontier*

Seed pages

Web

Web, Deep Web & Dark Web



Crawler requirements

- A crawler must respect some requirements:
  - **Robustness**: MUST avoid spider traps (fetching an infinite number of pages in a particular domain)
  - **Politeness**: MUST respect Web servers policies, regulating the rate at which crawlers can visit them

Robustness

- Web crawling isn’t feasible with one machine
    - All of the above steps are distributed
- Malicious pages
    - Spam pages
    - Spider traps – including dynamically generated
- Even non-malicious pages pose challenges
    - Latency/bandwidth to remote servers vary
    - Webmasters specific guidelines
        - How “deep” should you crawl a site’s URL hierarchy?
    - Site mirrors and duplicate pages

Politeness

- **Explicit politeness**: specifications from webmasters on what portions of site can/cannot be crawled
  - `robots.txt`
- **Implicit politeness**: even with no specification, avoid hitting any site too often

robots.txt

- Protocol for giving spiders ("robots") limited access to a website, originally from 1994
    - www.robotstxt.org
- Website announces its request on what can (or cannot) be crawled
    - For a server, create a file named `robots.txt`
    - This file specifies access restrictions
- `Robots.txt` contains set of rules that **should** be followed by clients

`robots.txt`: an Example

- Example:
  - No robot should visit any URL starting with "/yoursite/temp/", except the robot called “searchengine":

```
User-agent: *
Disallow: /yoursite/temp/
```

For **all** user-agents (client names) forbid access to directory */yoursite/temp*

```
User-agent: searchengine
Disallow:
```

For user-agents named “searchengine” do not forbid nothing, everything is thus accessible

`llms.txt`

- A proposal to standardise a dedicated file for providing guidance to LLMs on interpreting website content during inference
  - https://llmstxt.org
- A markdown file to websites to provide LLM-friendly content
  - For a server, create a file named `llms.txt`
  - This file is human and LLM readable that allows fixed processing methods (*i.e.*, parsers and regex)

llms.txt: an Example

```
# Title

> Optional description goes here

Optional details go here

## Section name

- [Link title](https://link_url): Optional link details

## Optional

- [Link title](https://link_url)
```

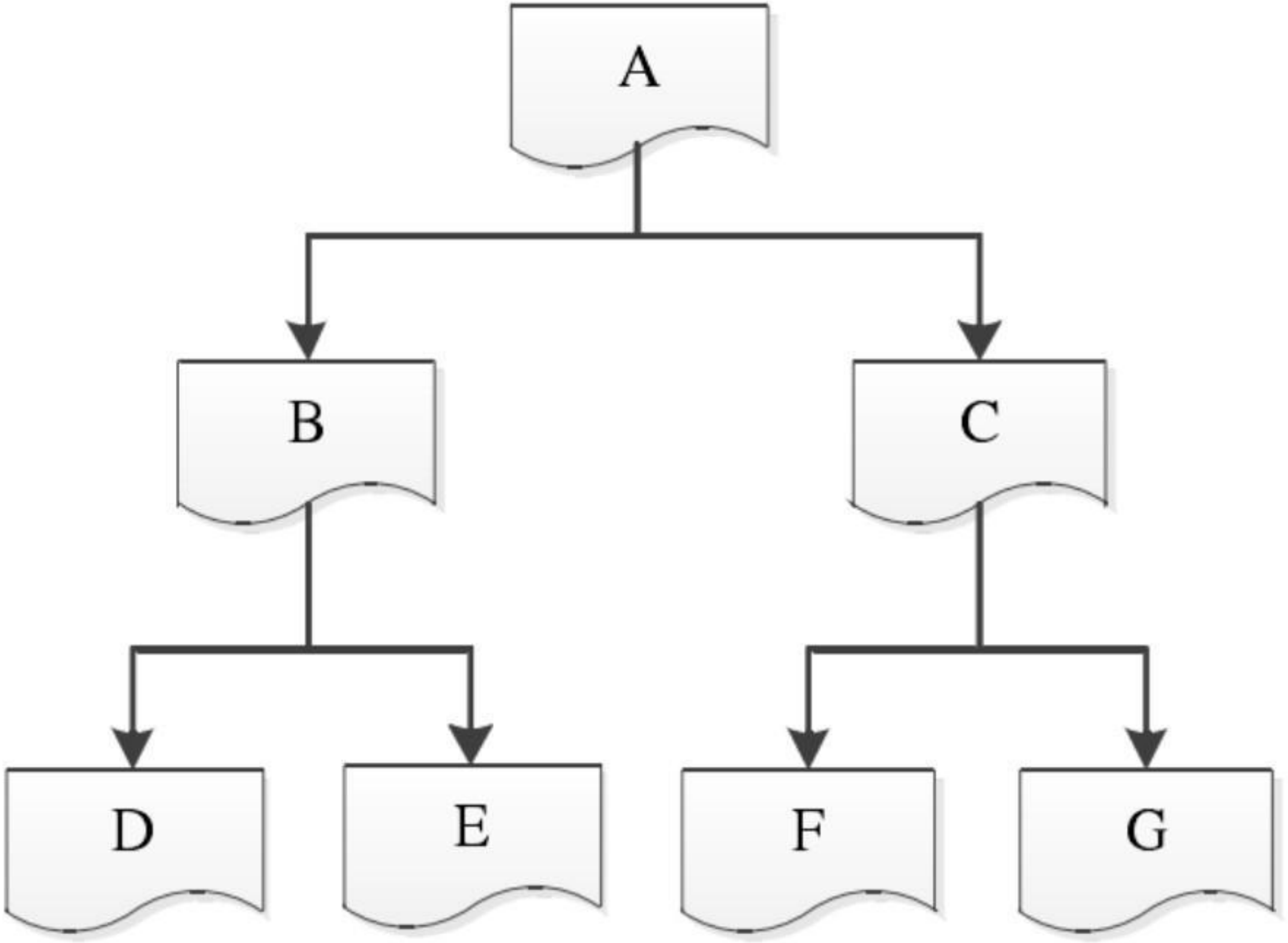
Rage Against the Machine

- Cloudflare has developed AI Labyrinth to **mitigate the impact of unauthorised web scraping** by bots that collect data for AI training
- The system generates **AI-created decoy pages** that appear authentic but are designed to **mislead bots and drain their resources**
- By monitoring interactions with these decoy pages, Cloudflare can accurately **detect and blacklist malicious web scrapers**

URL frontier

- Pages are **added** to the URL frontier according to the following strategies:
  - **Breadth first** strategy: given a Web page in the URL frontier, add **all pages linked by the current page**. Coverage is wide but superficial
  - **Depth first** strategy: given a Web page in the URL frontier, **follow the first link in the current page** until the first page without links

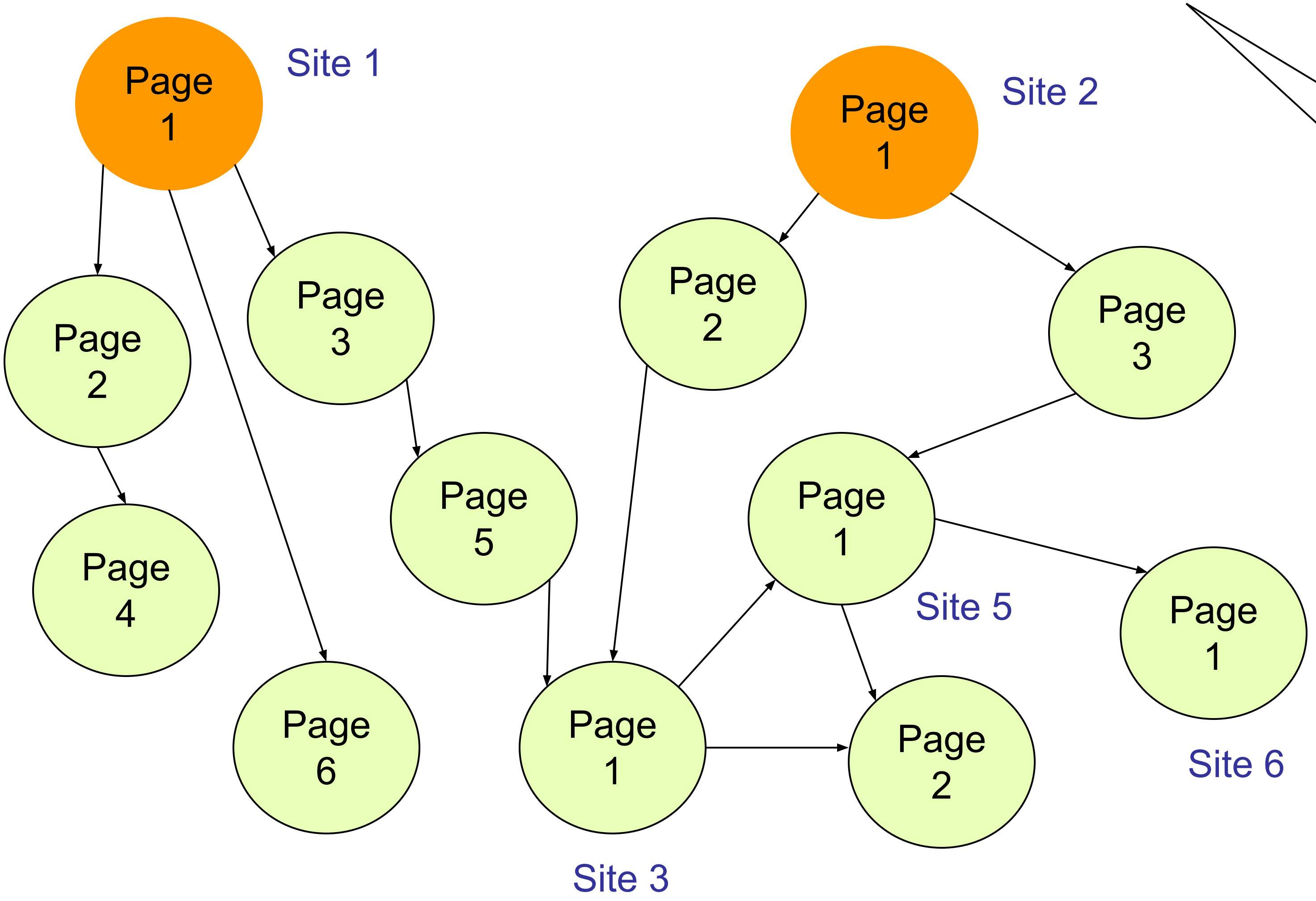
Breadth first strategy



Pages initially available in the URL frontier

A
B
C
D
E
F
G

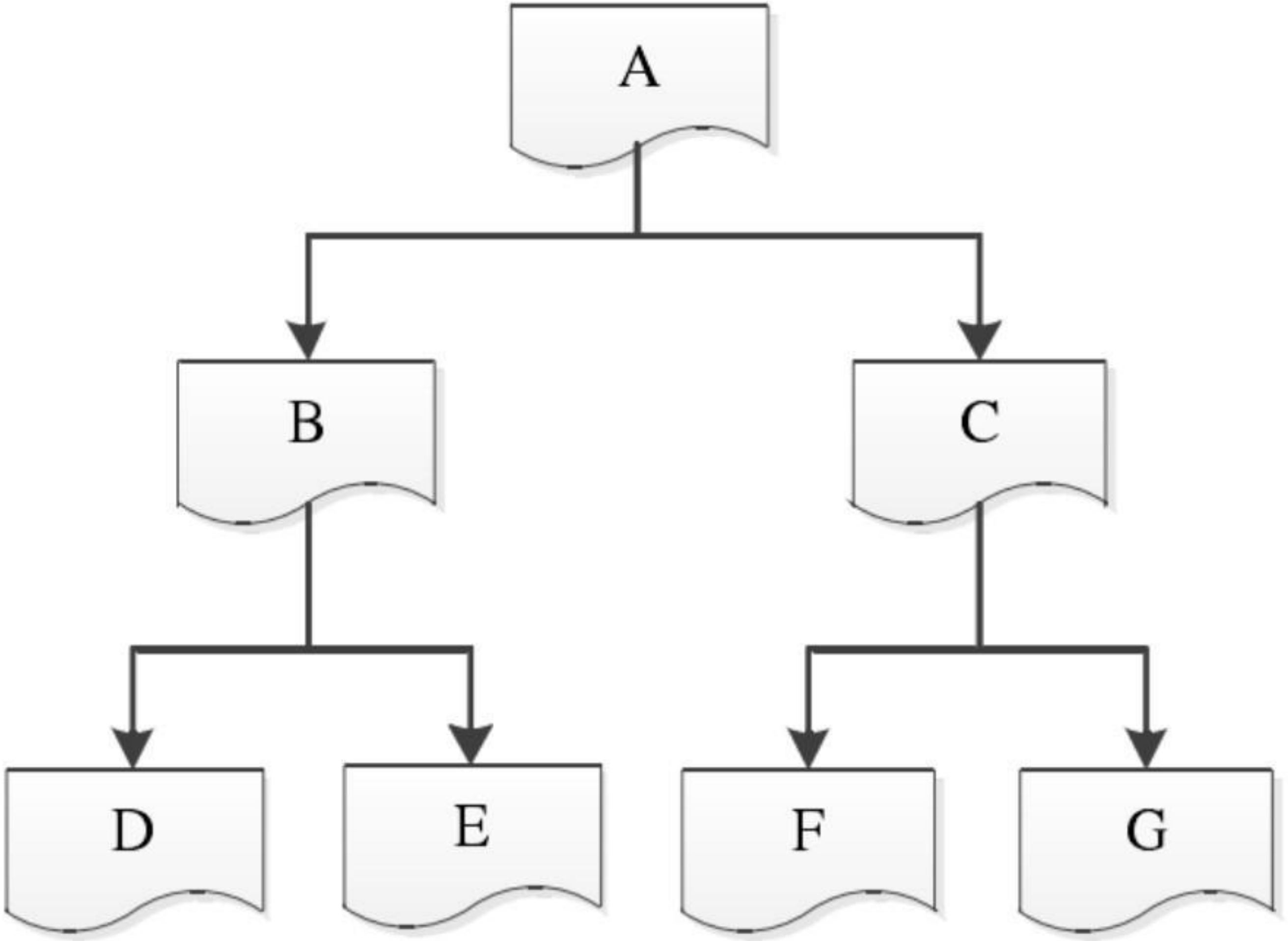
URL frontier with BFS



Pages initially available
in the URL frontier

Site	Page
1	1
2	1
1	2
1	6
1	3
2	2
2	3
1	4
1	5
3	1
5	1
5	2
6	1

Depth first strategy



Pages initially available in the URL frontier

A

B

D

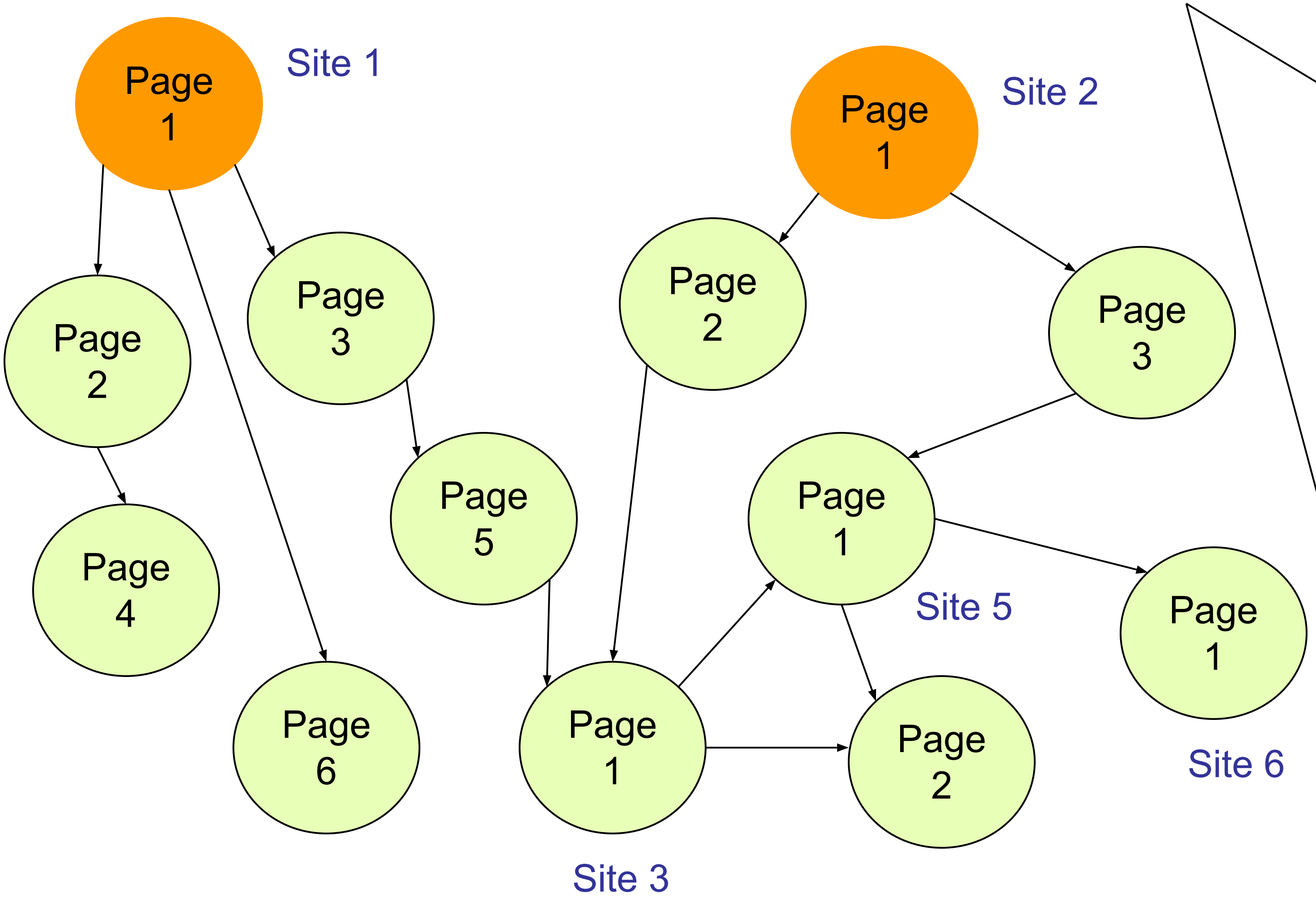
E

C

F

G

URL frontier with DFS



Pages initially available in the URL frontier

Site	Page
1	1
1	2
1	4
1	6
1	3
1	5
3	1
5	2
5	1
6	1
2	1
2	2
2	3

BFS vs DFS

BFS frontier

Site	Page
1	1
2	1
1	2
1	6
1	3
2	2
2	3
1	4
1	5
3	1
5	1
5	2
6	1

DFS frontier

Site	Page
1	1
1	2
1	4
1	6
1	3
1	5
3	1
5	2
5	1
6	1
2	1
2	2
2	3

Searching the web

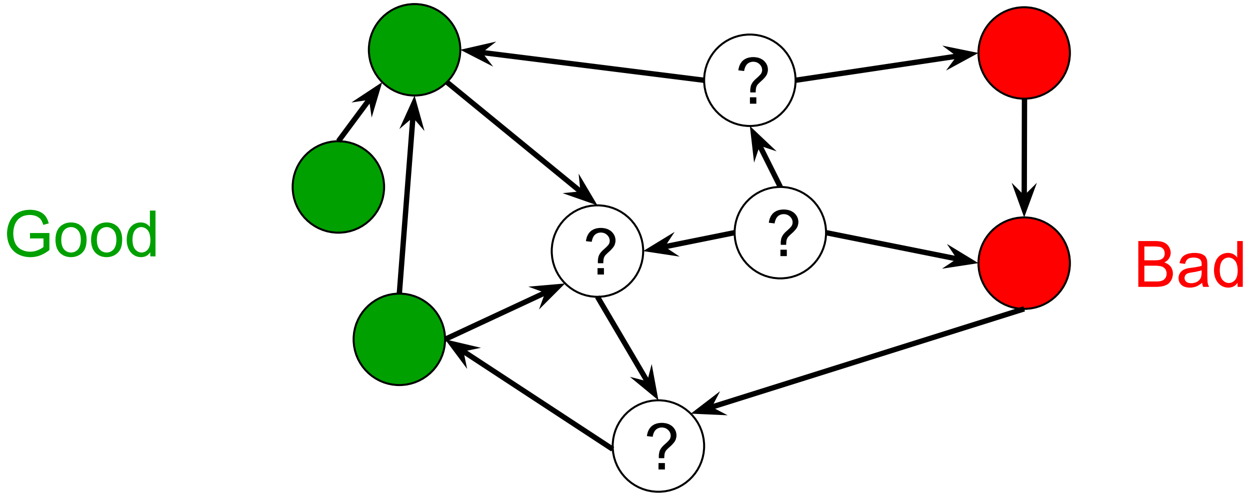
- There are thousands of billions of pages on the web, but most of them are not very interesting
- Suppose you have to visit the site for eBay and you don’t know that www.ebay.com is the URL
  - There are millions of web pages that contain the term “eBay”
  - There can be websites with more frequency on the term “eBay” than eBay itself
- We need a notion of **popularity**, together with a notion of relevance

Web information retrieval

- With respect to traditional textual search engines, **Web information retrieval** systems build **ranking** by combining at least two evidences of relevance:
  - the degree of matching of a page: the **content score**
  - the degree of importance of a page: the **popularity score**
- While the **content score** can be calculated using one of the information retrieval models described so far
- The **popularity score** can be calculated from an analysis of the indexed pages' **hyperlink structure** using one or more ***link analysis*** models
  - Do the links represent a conferral of authority to some pages? Is this useful for ranking?

Simple link analysis

- Links are powerful sources of authenticity and authority
- The **Good**, The **Bad** and The **Unknown**, simple iterative logic
  - Good nodes won’t point to Bad nodes
  - If you point to a Bad node, you’re Bad
  - If a Good node points to you, you’re Good



Citation Analysis

- Citation frequency is an estimation of a researcher popularity
- Bibliographic coupling frequency
  - Articles that co-cite the same articles are related
- Citation indexing: as a tool in journal evaluation
  - Who is this author cited by? (Garfield 1972)
- PageRank preview: Pinski and Narin ‘70s*
  - Asked: which journals are authoritative?

*_Citation influence for journal aggregates of scientific publications: Theory, with application to the literature of physics_

PageRank

- The PageRank technique for link analysis assigns a **numerical score between 0 and 1** to every node in the web graph
- The PageRank score of a node depends on the **link structure** of the web graph
- Given a query, a web search engine computes a **composite score** for each web page that combines hundreds of features such as cosine similarity, together with the PageRank score
- This composite score is used to provide a **ranked list of results** for the query

The random surfer (1/3)

- Consider a *random surfer* Alice who begins a random walk on the web, starting from a page
  - Alice is extremely bored, she wanders aimlessly between web pages
  - Her browser has a special “**surprise me**” button at the top that will jump to a random web page when clicked
  - Each time a web page loads she chooses whether to
    - Click on a **random** link on the page
    - Click the surprise me button
  - Alice is sufficiently bored that she intends to keep browsing the Web like this forever

The random surfer (2/3)

- Let provide a more formally definition: Alice browses the Web using this algorithm:

  1. Choose a random number *r* between 0 and 1
  2. If $r > \lambda$:

     ⇒ Click the “surprise me” button
  3. If $r \leq \lambda$:

     ⇒ Click a link at random on the current page
  4. Start again

- Because of Alice’s special “surprise me” button, we can be guaranteed that eventually she will reach every page on the Internet

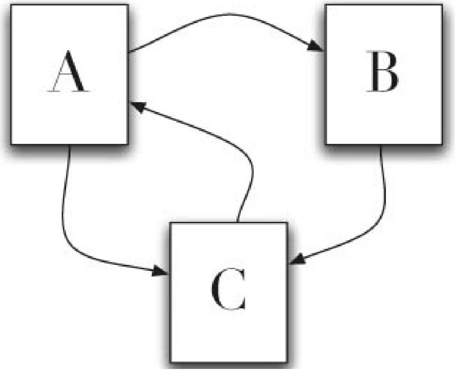
The random surfer (3/3)

- Now suppose that while Alice is browsing, you walk in and glance at the web page on her screen. **What is the probability** that she will be looking at the eBay website?

- That probability is eBay's **PageRank**

PageRank calculation

- The PageRank calculation corresponds to finding the stationary probability distribution of a *random walk* on the graph of the Web. A random walk is a special case of a *Markov chain* in which the next state depends solely on the current state



- If the web consists of the 3 pages in figure (A,B,C), the PageRank of C depends on the PageRank of A and B:

$$PR(C) = \frac{PR(A)}{2} + \frac{PR(B)}{1}$$

- The PageRank conferred by an outbound link is equal to the document's own PageRank score divided by the number of outbound links
- We start by assuming that the PageRank values for all pages are the same, then we iterate the calculation. After few iterations, the PageRank values converge to the final values of
  - PR(C) = 0.4
  - PR(A) = 0.4
  - PR(B) = 0.2

Use of PageRank in Google

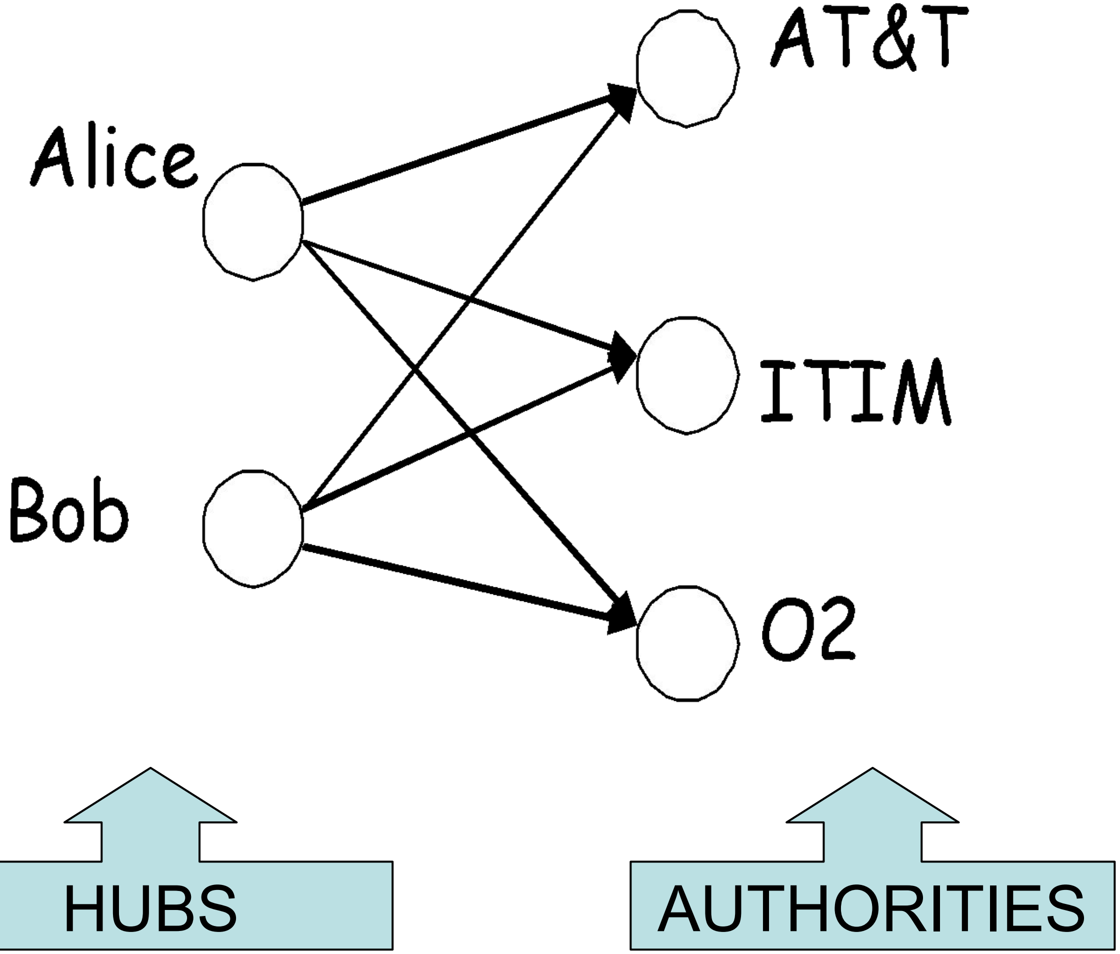
- PageRank is now **only one of the many** factors that determine the final score of a Web page in Google
- It is now a part of a much larger ranking system that it is believed to account for more than **200 different "signals"** (ranking variables):
  - **language features** (phrases, synonyms, spelling mistakes, etc.)
  - **query features** that relate to language features, trending terms/phrases
  - **time-related features** (e.g., "news" related queries might be best answered by recently indexed documents, while factual queries are better answered by more "resilient" pages)
  - **personalization features**, which relate to one's search history, behavior, and social surrounding

Hyperlink-Induced Topic Search (HITS)

- In response to a query, instead of an ordered list of pages each meeting the query, find **two sets of inter-related pages**:
  - ***Hub pages*** are good lists of links on a subject
    - e.g., “Bob’s list of cancer-related links.”
  - ***Authority pages*** occur recurrently on good hubs for the subject
- Best suited for “broad topic” queries rather than for page-finding queries
- Gets at a broader slice of common *opinion*

HITS Example

Query: “***Mobile telecom companies***”



HUBS

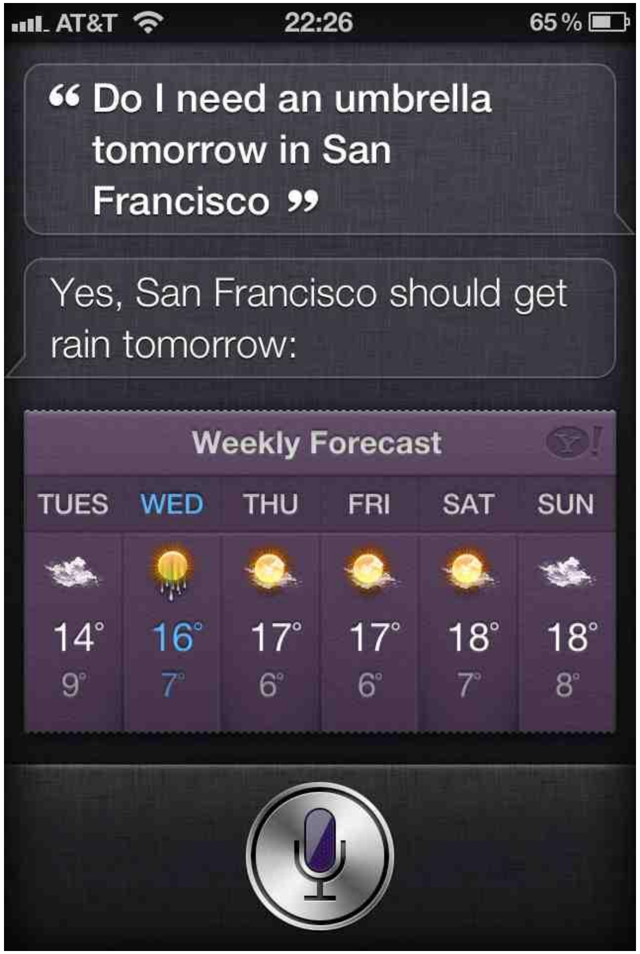
AUTHORITIES

Hubs and Authorities

- Thus, a **good hub page** for a topic *points* to many authoritative pages for that topic
- A **good authority page** for a topic is *pointed* to by many good hubs for that topic
- Circular definition - will turn this into an iterative computation

Semantic Search

- The name “information retrieval” is standard, but as traditionally practiced, it’s not really right
- All you get is **document retrieval**, and beyond that the job is up to you
- **Semantic Search**: doing graph search over structured knowledge rather than traditional text search:
  - Google Knowledge Graph
  - Facebook Graph Search
  - Bing’s Satori
  - Things like Wolfram Alpha



AT&T 22:26 65 %

“ Do I need an umbrella tomorrow in San Francisco ”

Yes, San Francisco should get rain tomorrow:

Weekly Forecast

TUES	WED	THU	FRI	SAT	SUN
14°	16°	17°	17°	18°	18°
9°	7°	6°	6°	7°	8°

References

- Bruce Croft, Donald Metzler, Trevor Strohman **Search Engines: Information Retrieval in Practice** Pearson (2010)

- Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze **Introduction to Information Retrieval** Cambridge University Press. (2008)

- Stefano Ceri, Alessandro Bozzon **Web Information Retrieval** Springer. (2013)