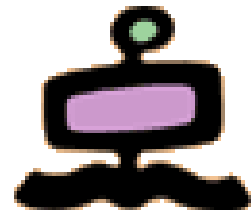


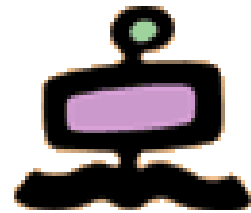


Corso di Biblioteche Digitali

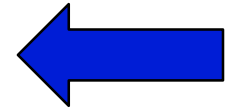


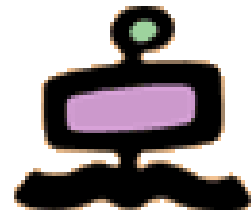
- Vittore Casarosa
 - Mail: casarosa@isti.cnr.it
 - Tel: 050 621 3115 (office) 348 397 2168 (mobile)
 - Skype: vittore1201
- “Ricevimento” at the end of the lessons or by appointment
- Final assessment
 - 70% oral examination
 - 30% project (development of a small digital library))
- Reference material:
 - Ian Witten, David Bainbridge, David Nichols, How to build a Digital Library, Morgan Kaufmann, 2010, ISBN 978-0-12-374857-7 (Second edition)
 - Material provided by the teacher
- **<http://cloudone.isti.cnr.it/casarosa/BDG/>**

Modules

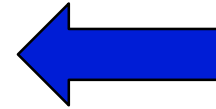


- Computer Fundamentals and Networking
- A conceptual model for Digital Libraries
- Bibliographic records and metadata
- Information Retrieval and Search Engines
- Knowledge representation
- Digital Libraries and the Web
- Hands-on laboratory: the Greenstone system

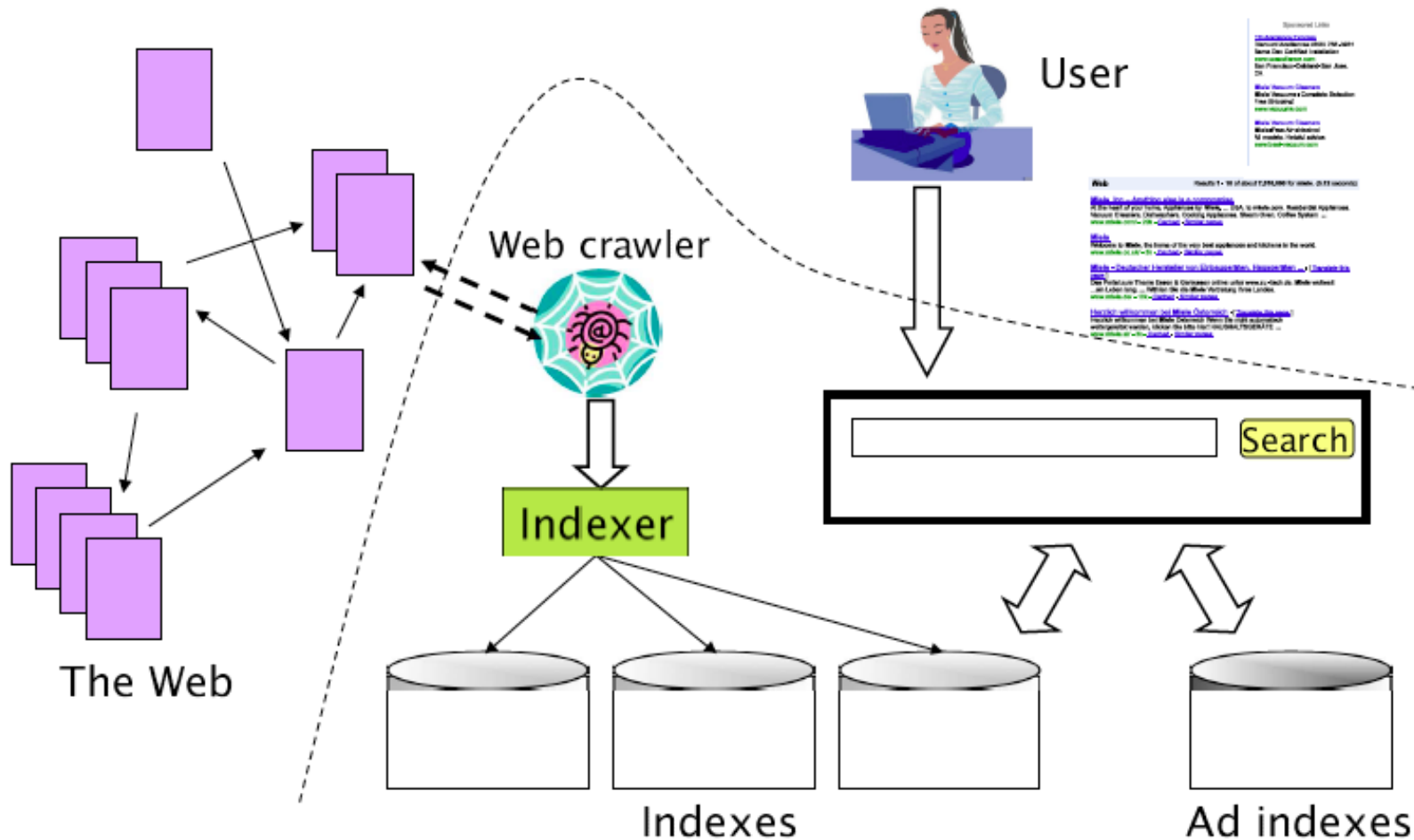
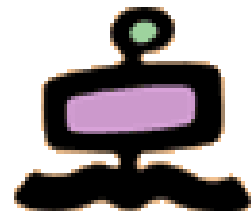




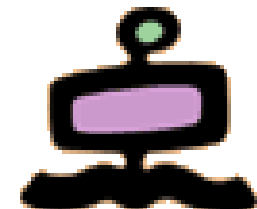
- Information Retrieval and Search Engines
 - Indexing a collection of documents
 - Ranking query results
 - Search engines in the Web
 - Ranking in Web search engines



Architecture of a Search Engine



The size of the indexed Web



Last Month

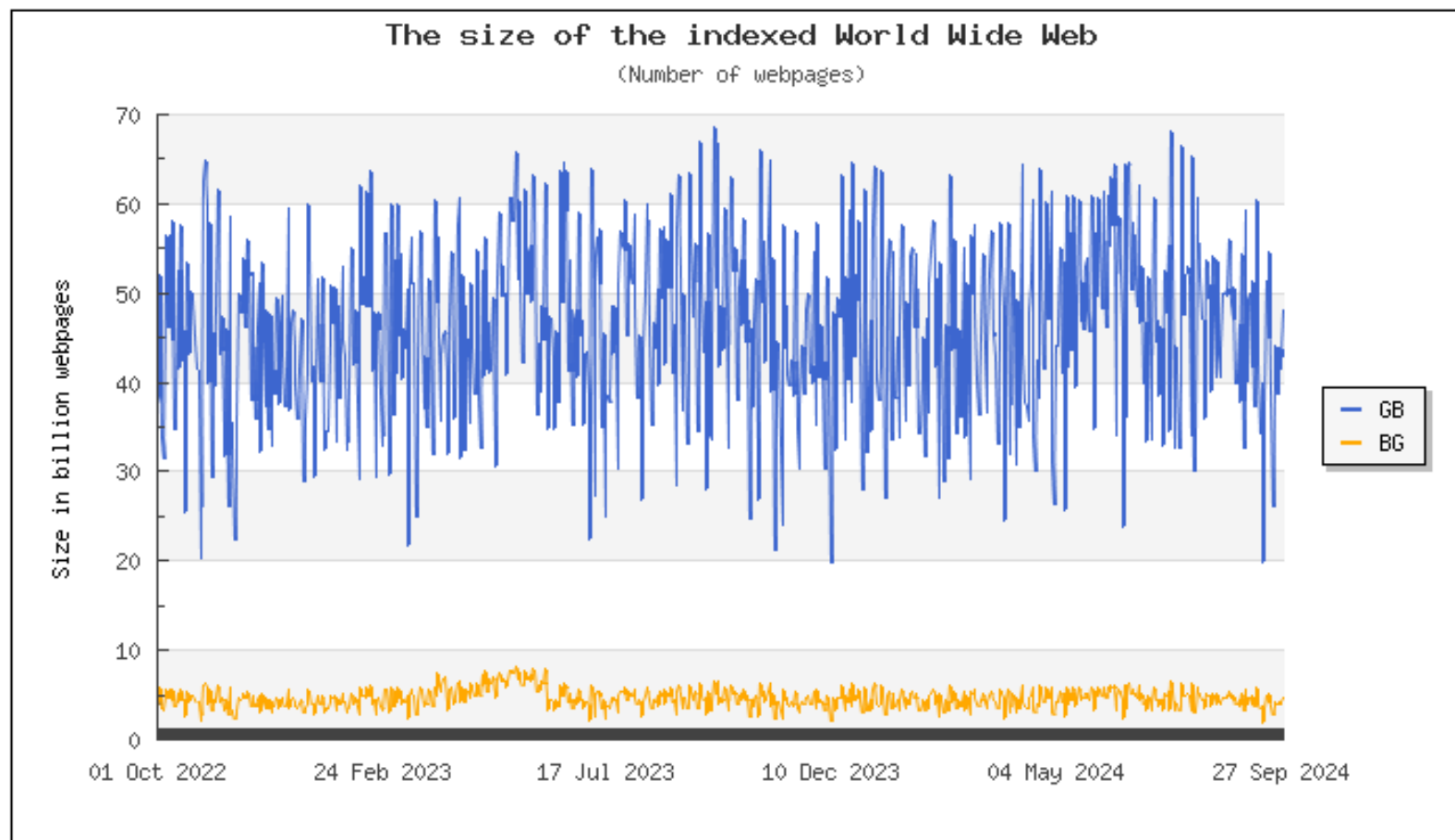
Last Three Months

Last Year

Last Two Years

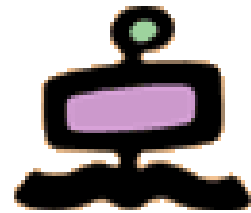
Last Five Years

Last Ten Years



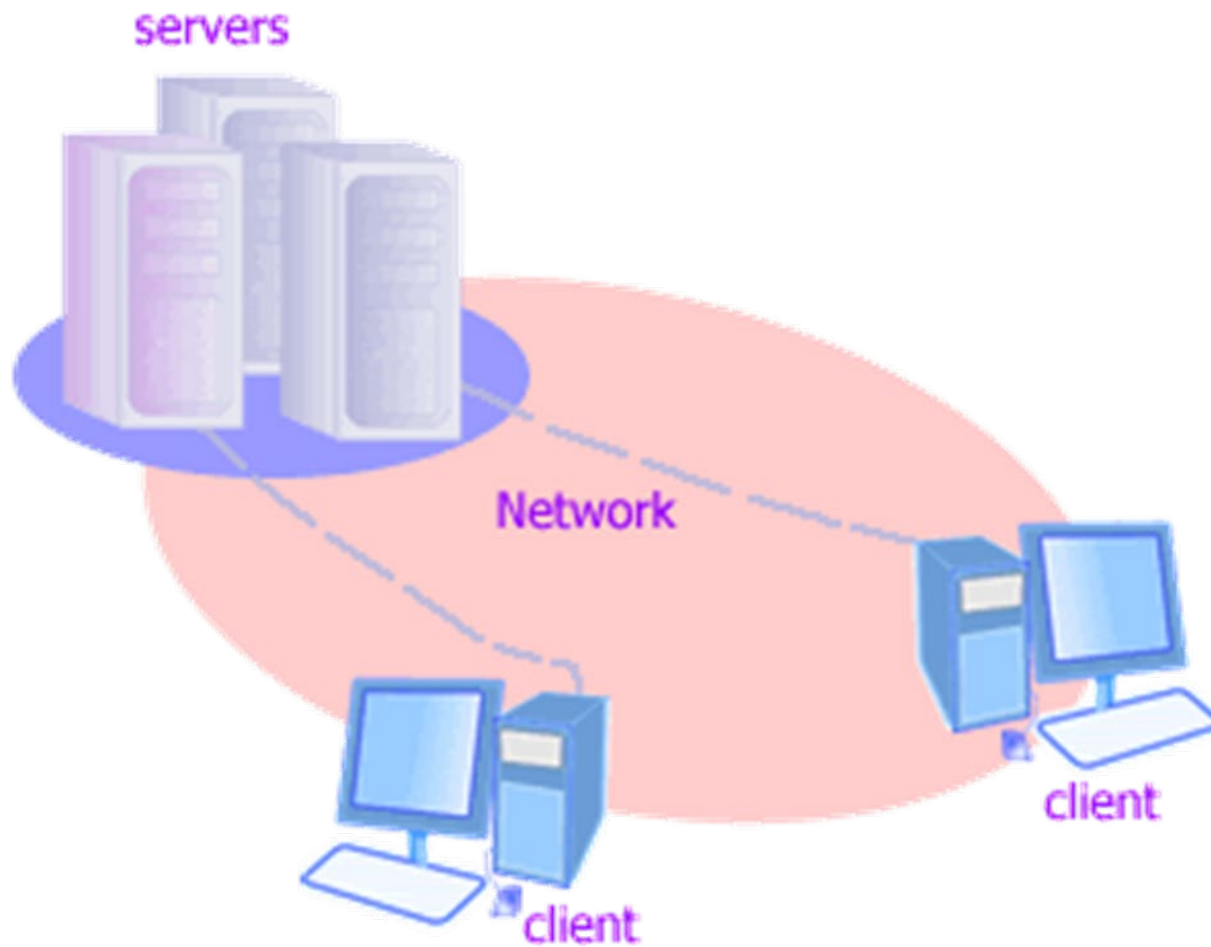
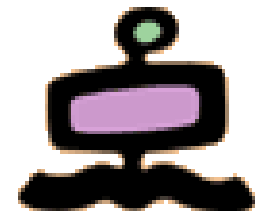
<https://www.worldwidewebsite.com/>

The Depth of the Web

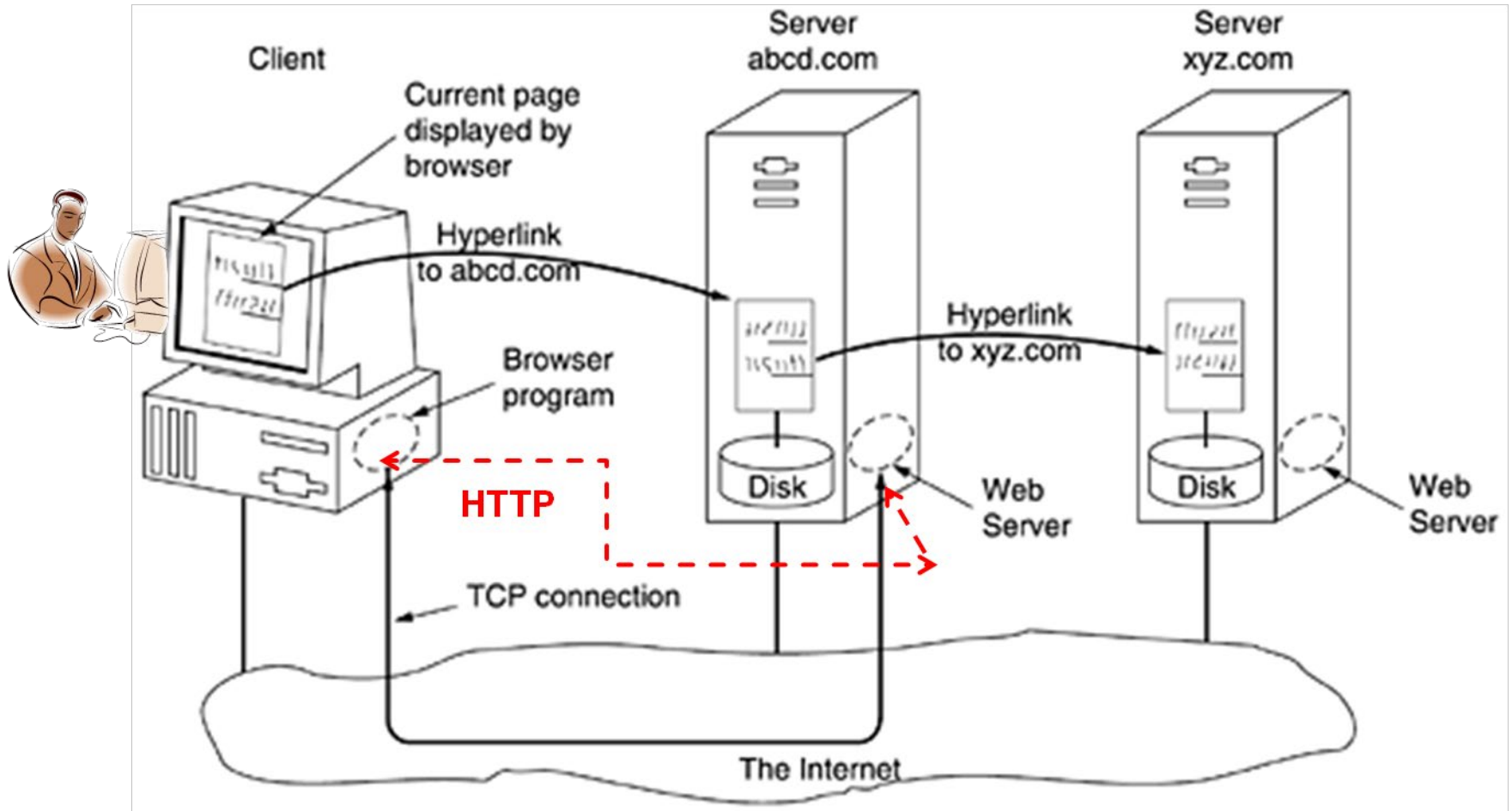
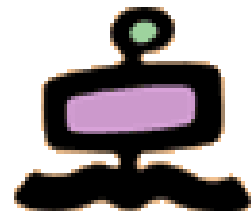


- A URL gives access to a web page.
- That page may have links to other pages (static pages). This is the **surface web**.
- Some pages (dynamic pages) are generated only when some information is provided to the web server.
- These pages cannot be discovered just by crawling. This is the deep web.
- The surface web is huge.
- The deep web is “unfathomable”.

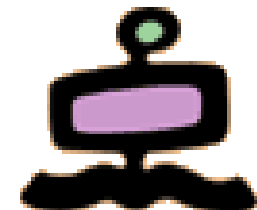
Client-server networks



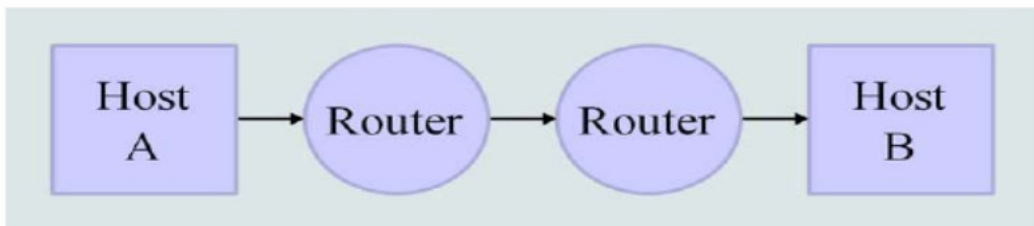
The Web architecture



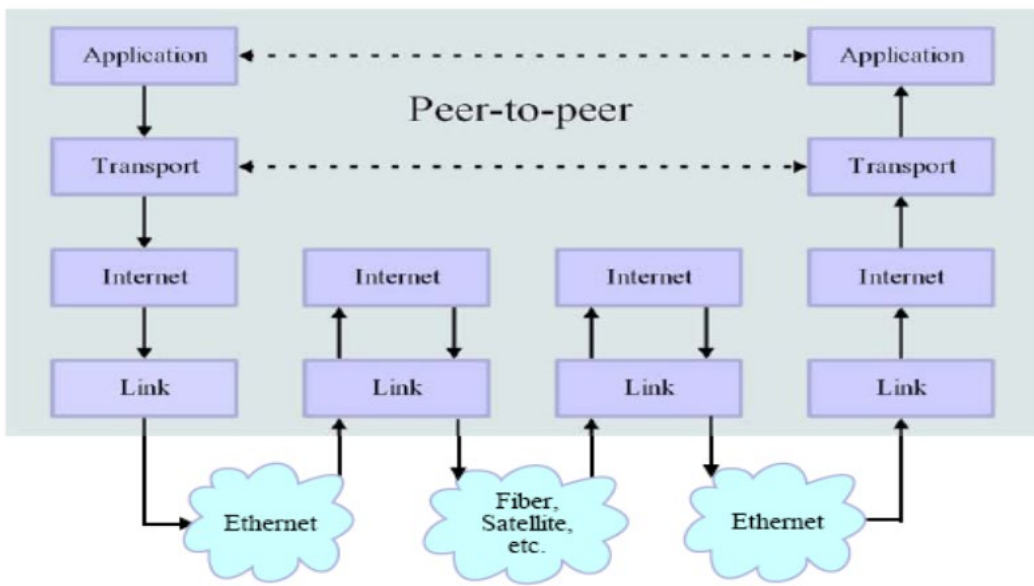
Internet protocols



Network Connections



Stack Connections



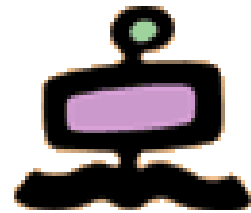
Application protocols
e.g. HTTP

TCP/UDP

IP

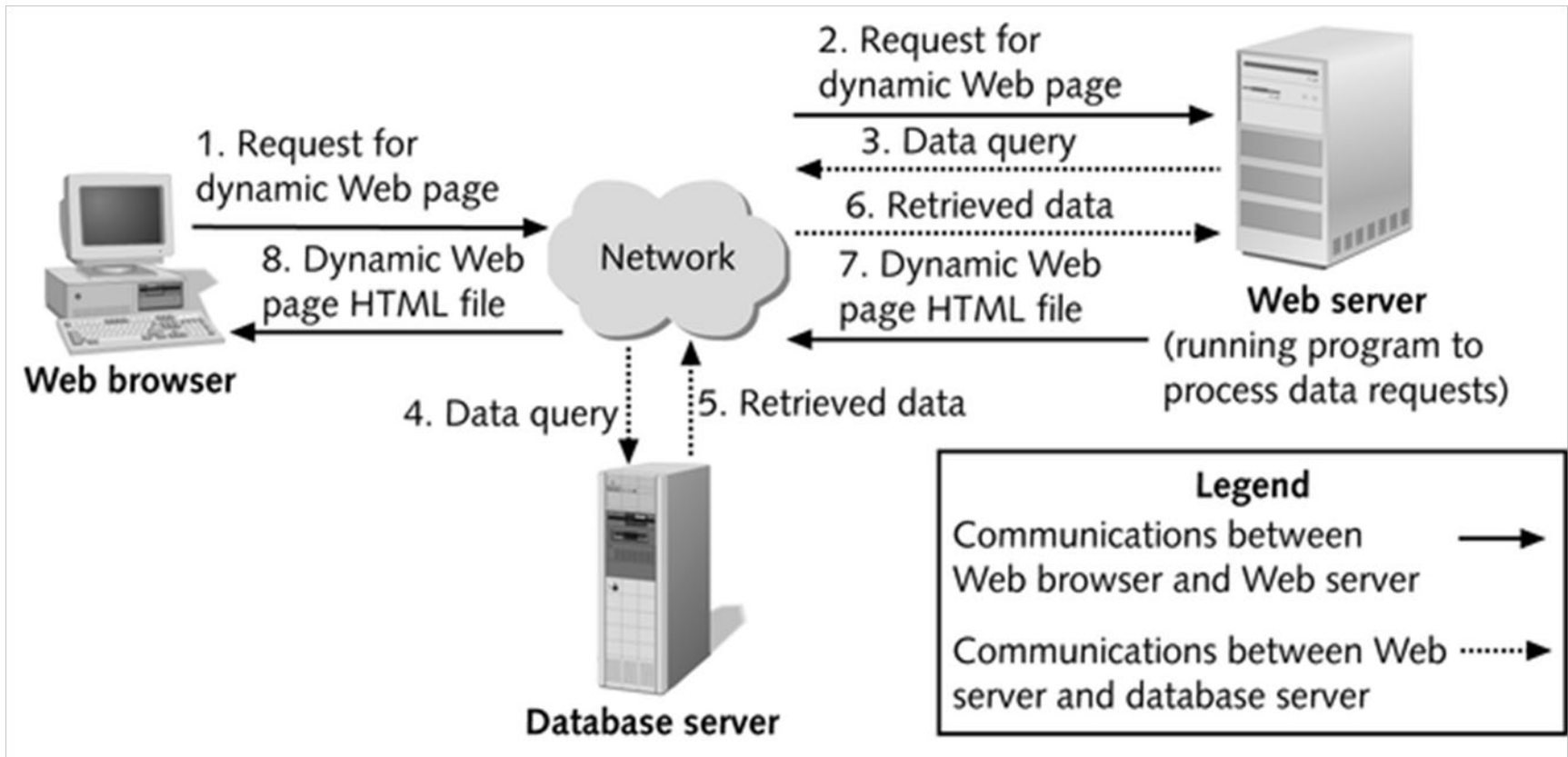
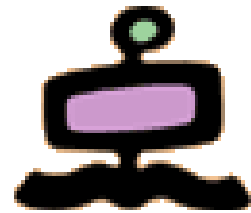
ETHERNET

The Depth of the Web

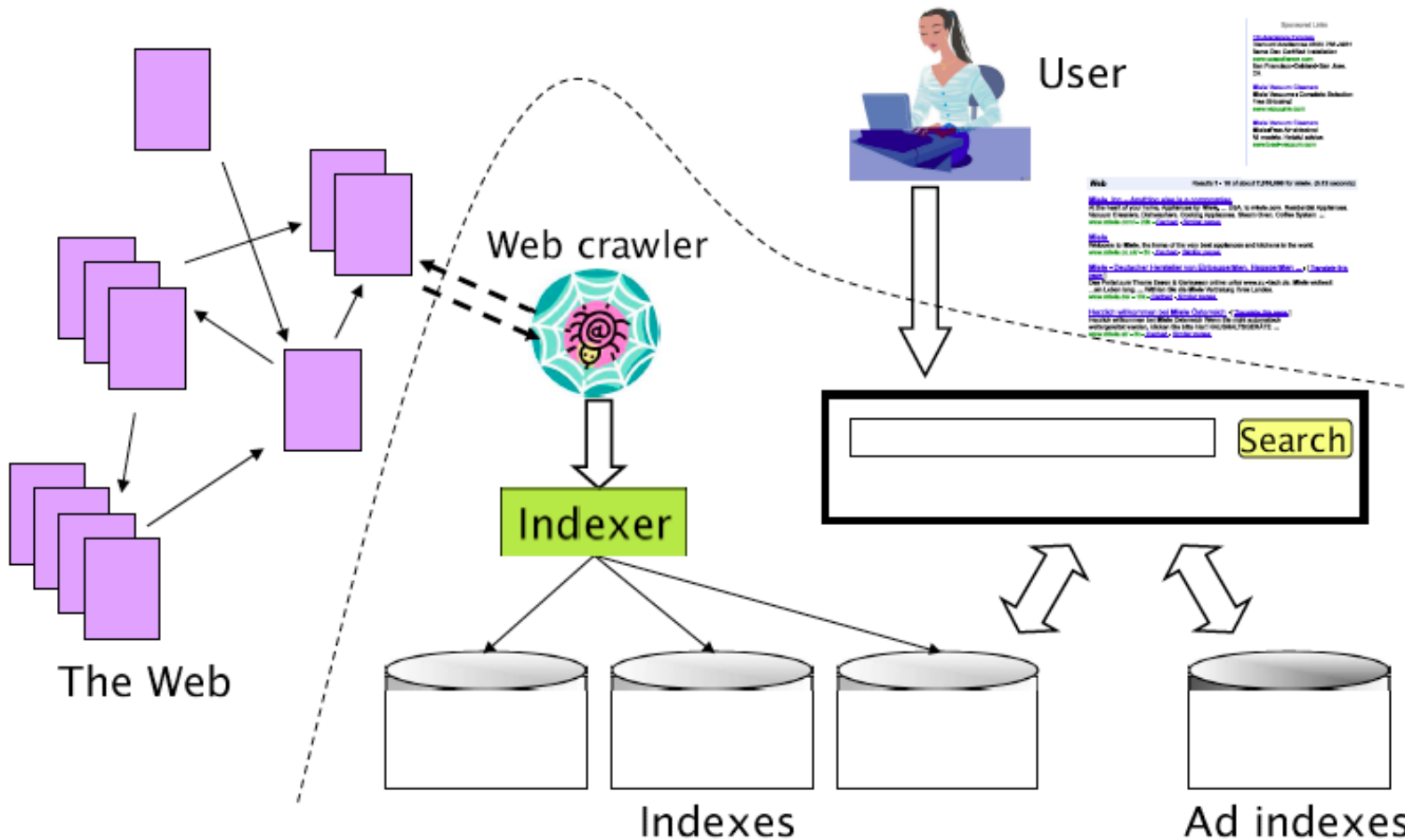
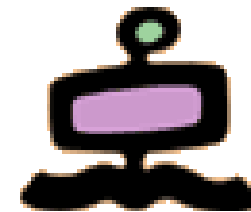


- A URL gives access to a web page.
- That page may have links to other pages (static pages). This is the surface web.
- Some pages (dynamic pages) are generated only when some information is provided to the web server.
- These pages cannot be discovered just by crawling. This is the **deep web**.
- The surface web is huge.
- The deep web is “unfathomable”.

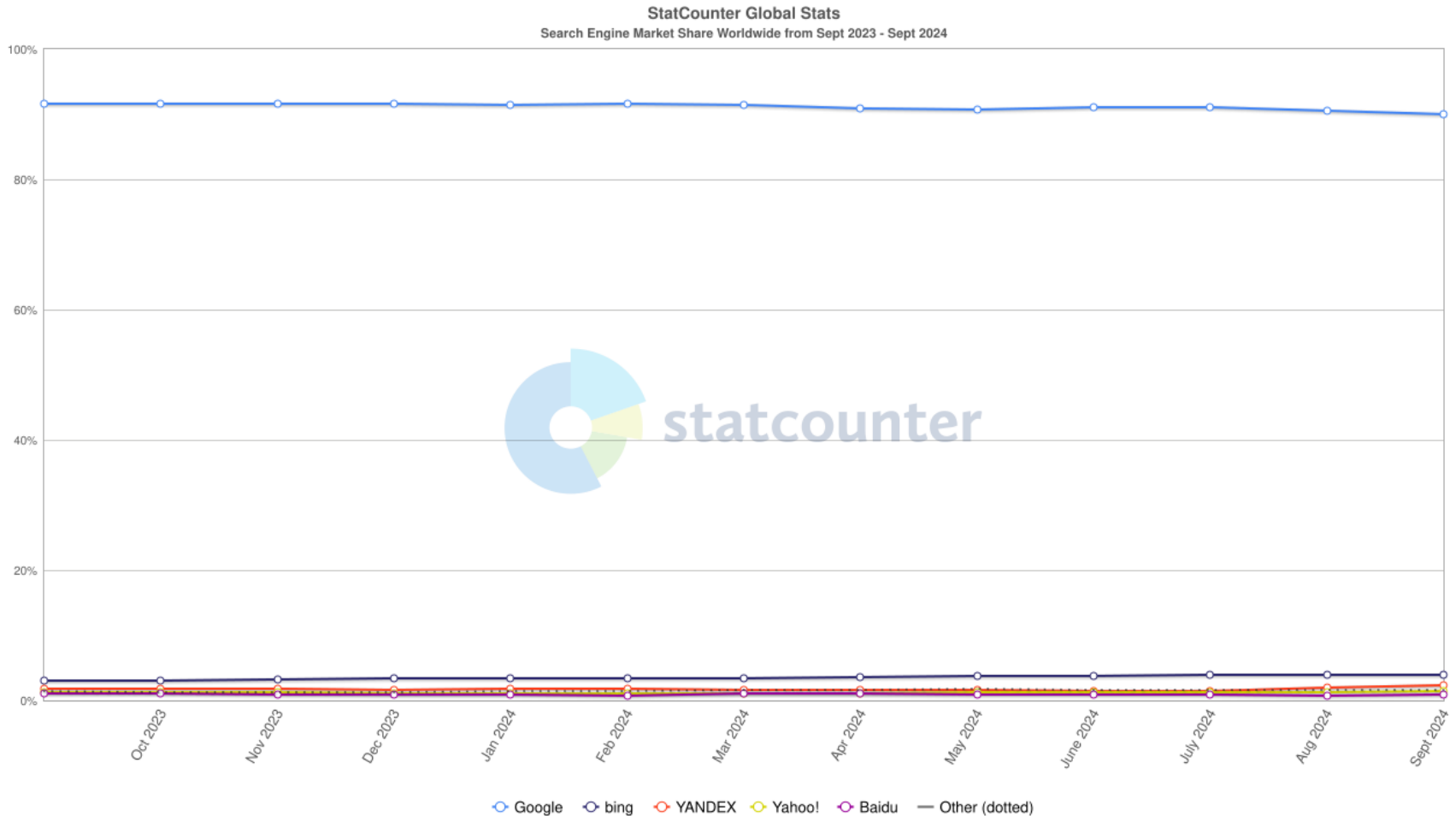
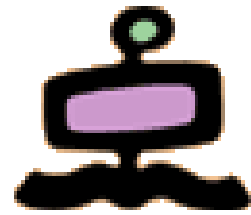
Dynamic web pages (data base driven)



Architecture of a Search Engine

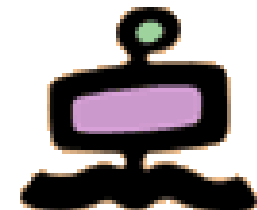


Worldwide queries to search engines (2024)

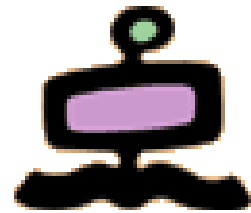


<https://gs.statcounter.com/search-engine-market-share>

Distribution of queries to Search Engines



Date	Google	bing	YAND EX	Yahoo!	Baidu	DuckDuckGo
2023-09	91.58	3.01	1.81	1.24	1,00	0.58
2023-10	91.56	3.10	1.84	1.2	1.02	0.54
2023-11	91.54	3.20	1.78	1.22	0.93	0.55
2023-12	91.62	3.37	1.65	1.12	0.96	0.51
2024-01	91.47	3.43	1.78	1.1	0.83	0.54
2024-02	91.61	3.32	1.83	1.08	0.77	0.53
2024-03	91.38	3.35	1.65	1.1	1,00	0.53
2024-04	90.91	3.64	1.61	1.13	1.15	0.55
2024-05	90.80	3.72	1.58	1.2	0.92	0.56
2024-06	91.06	3.72	1.45	1.26	0.87	0.60
2024-07	91.02	3.88	1.37	1.24	0.91	0.62
2024-08	90.50	3.91	1.95	1.32	0.79	0.64
2024-09	90.00	3.96	2.35	1.35	0.81	0.65



GENERAL GOOGLE SEARCH STATISTICS



Google is the most visited website



92% of all search volume globally is from Google



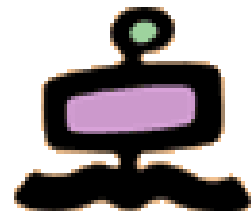
There are at least **8.5 billion** Google searches per day



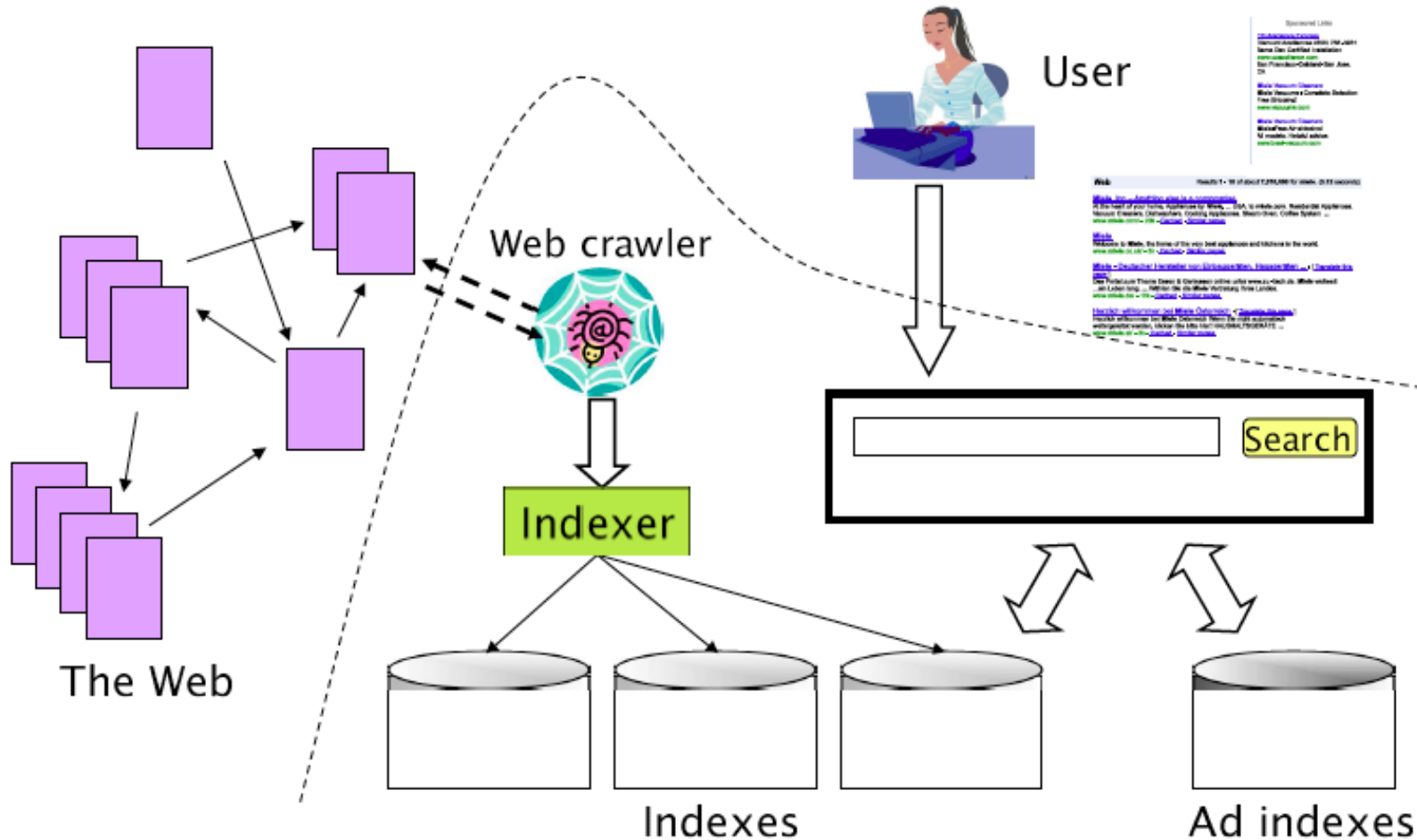
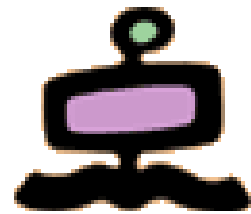
The Google Index has up to **40 to 60 billion** active web pages

<https://fitsmallbusiness.com/google-search-statistics/>

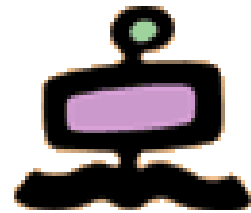
To google



Architecture of a Search Engine

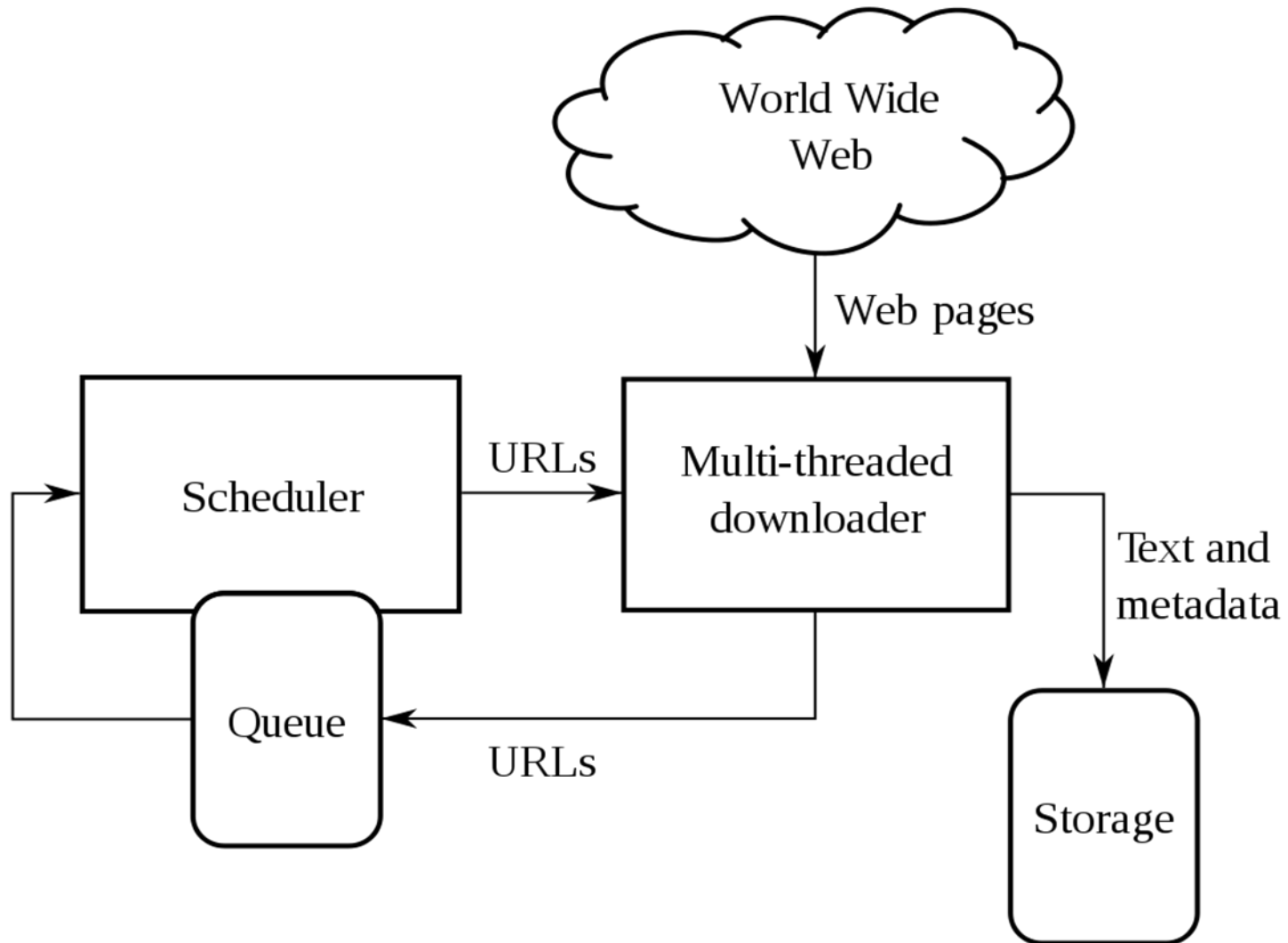
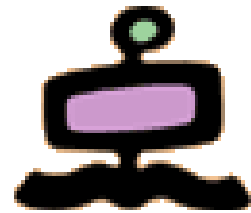


Main functions of a search engine

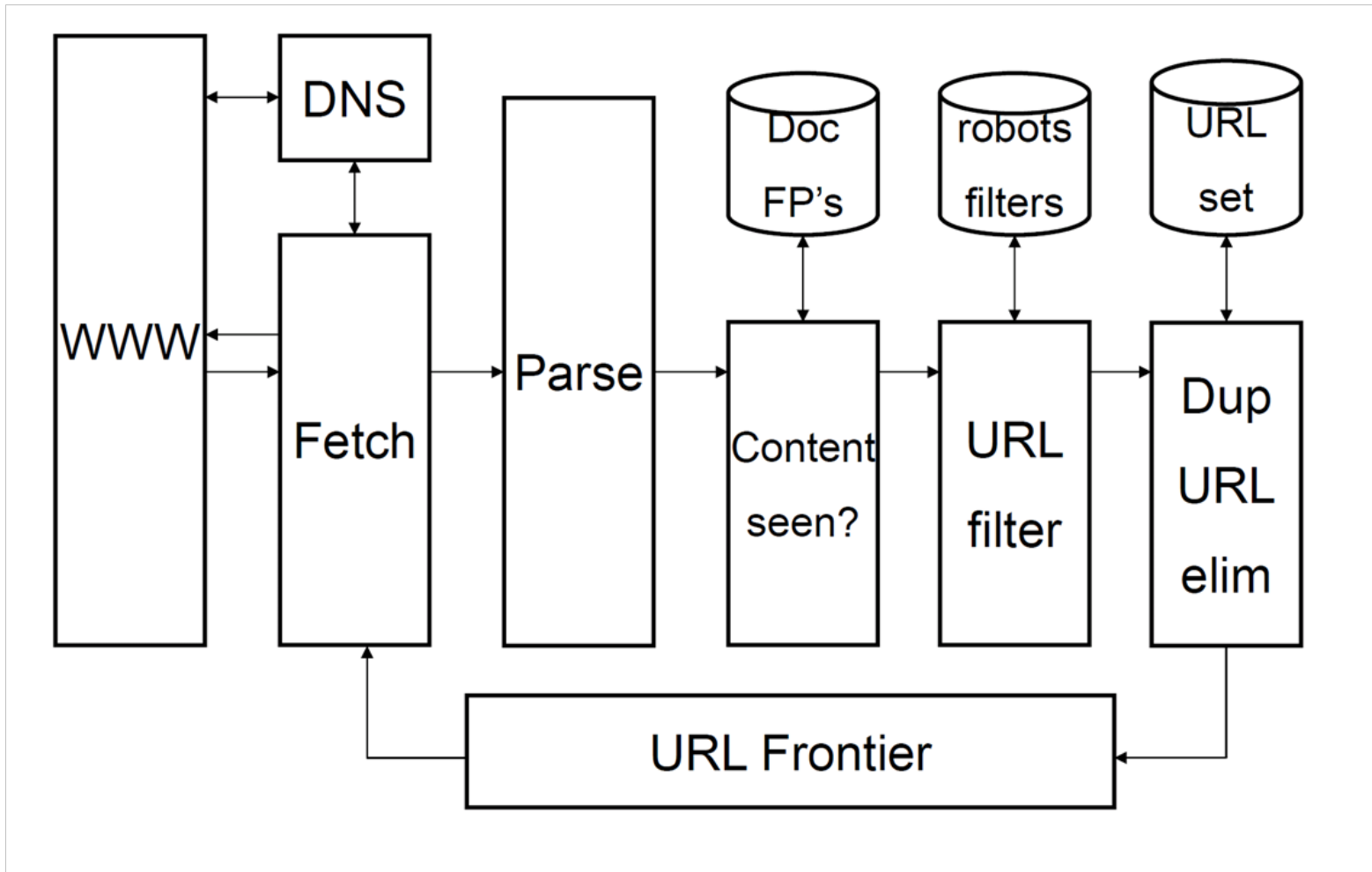
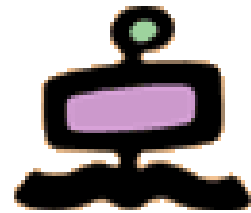


- Crawling ←
- Indexing (in parallel with crawling)
- Ranking based on page content
- Ranking based on Web considerations
- Display of results

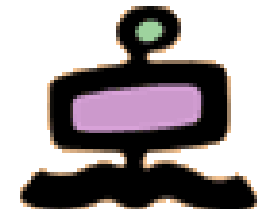
Basic architecture of a crawler (spider, bot)



Crawler architecture

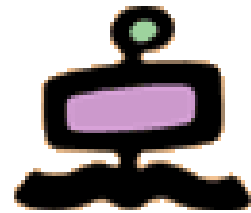


Crawling



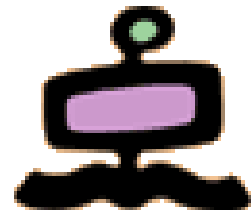
- A web crawler (aka a *spider* or a *bot*) is a program
 - Starts with one or more URL – the *seed*
 - Other URLs will be found in the pages pointed to by the seed URLs. They will be the starting point for further crawling
 - Uses the standard protocols (HTTP, FTP) for requesting a resource from a server
 - Requirements for respecting server policies
 - Politeness
 - Parses the resource obtained
 - Obtains additional URLs from the fetched page
 - Provides the fetched page to the indexer
 - Implements policies about content
 - Recognizes and eliminates duplicate or unwanted URLs
 - Adds the URLs found in the fetched page to the queue and continues requesting pages

What any crawler must do



- A crawler must be
 - Robust: Survive *spider traps*. Websites that fool a spider into fetching large or limitless numbers of pages within the domain.
 - Some deliberate; some errors in site design
 - Polite: Crawlers can interfere with the normal operation of a web site. Servers have policies, both implicit and explicit, about the allowed frequency of visits by crawlers. Responsible crawlers obey these policies.

Politeness



- Explicit
 - Specified by the web site owner
 - What portions of the site may be crawled and what portions may not be crawled
 - robots.txt file
- Implicit
 - If no restrictions are specified, still restrict how often you hit a single site.
 - You may have many URLs from the same site. Too much traffic can interfere with the site's operation. Crawler hits are much faster than ordinary traffic – could overtax the server. (Constitutes a denial of service attack) Good web crawlers do not fetch multiple pages from the same server at one time.

robots.txt example



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

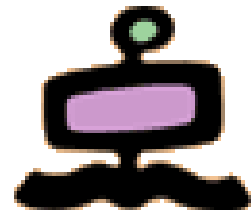
```
User-agent: *
```

```
Disallow: /yoursite/temp/
```

```
User-agent: searchengine
```

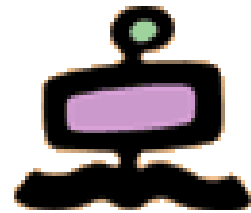
```
Disallow:
```


Scale of crawling



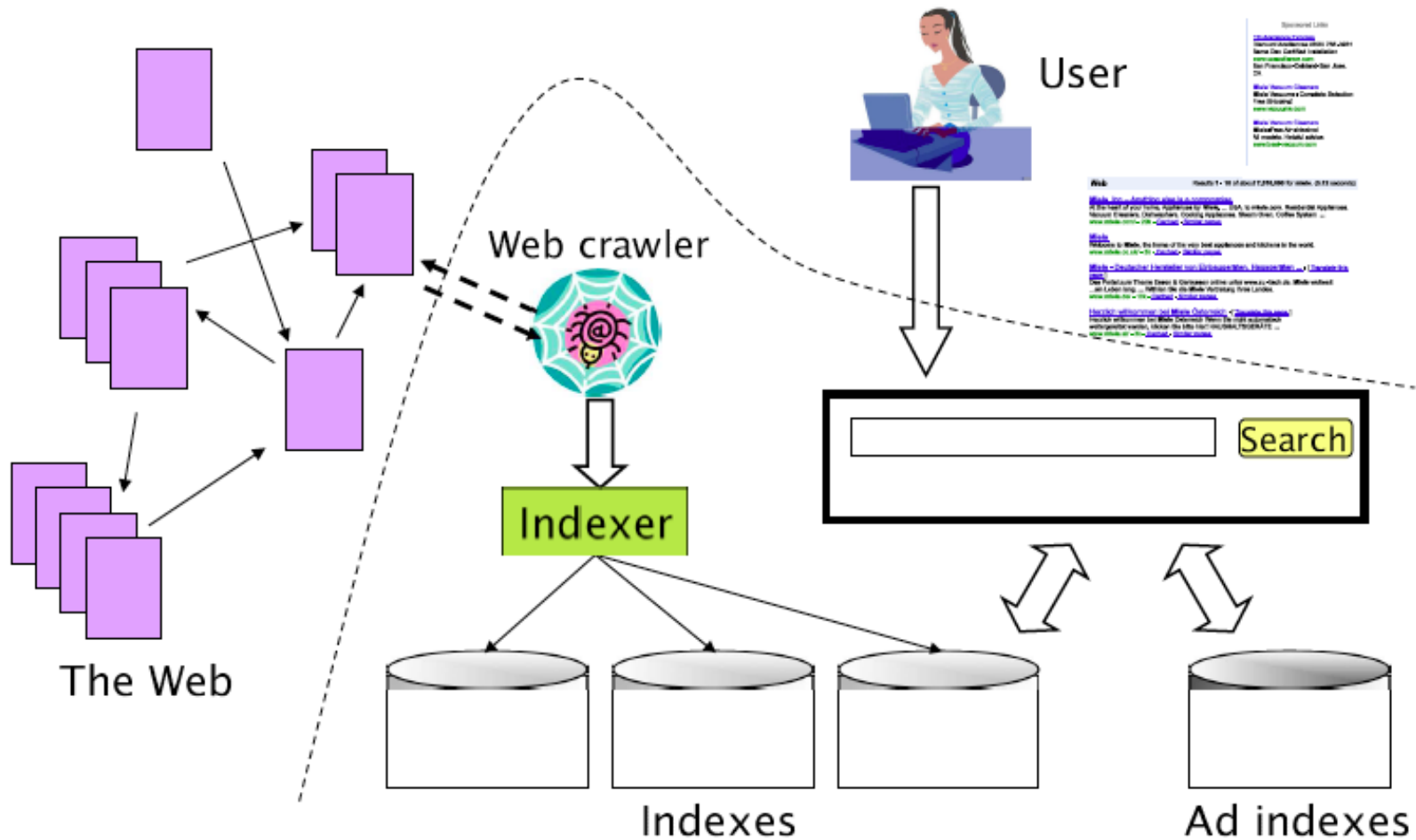
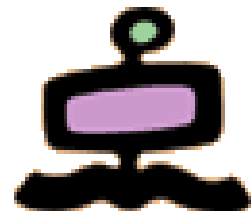
- A one month crawl of a billion pages requires fetching several hundred pages per second
- It is easy to lose sight of the numbers when dealing with data sources on the scale of the Web.
 - $30 \text{ days} * 24 \text{ hours/day} * 60 \text{ minutes/hour} * 60 \text{ seconds/minute} = 2,592,000 \text{ seconds}$
 - $1,000,000,000 \text{ pages} / 2,592,000 \text{ seconds} = \mathbf{386} \text{ pages/second}$
- Note that those numbers assume that the crawling is continuous

Distributed crawler

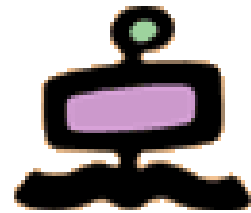


- For big crawls,
 - Many processes, each doing part of the job
 - Possibly on different nodes
 - Geographically distributed
 - How to distribute
 - Give each node a set of hosts to crawl
 - Use a hashing function to partition the set of hosts
 - How do these nodes communicate?
 - Need to have a common index

Architecture of a Search Engine



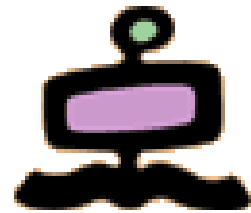
Main functions of a search engine



- Crawling
- Indexing (in parallel with crawling)
- Ranking based on page content
- Ranking based on Web considerations
- Display of results

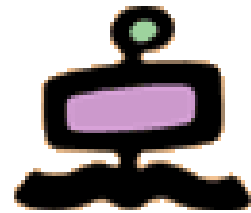


Indexing



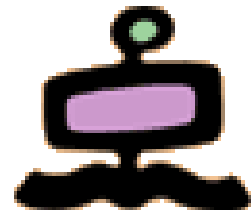
- Retrieved web page sent also to indexer to scan text (ignoring links)
- Build the index and the document (pages) representations (bag of words represented as vectors)
- Use of HTML information to improve the index and the “weight vectors”
- At query time, use the index and the weight vectors to get an initial ranking of relevant web pages, based on their content

Summary of retrieval and ranking



- Build a “term-document matrix”, assigning a weight to each term in a document (instead of just a binary value as in the simple approach)
 - Usually the weight is *tf.idf*, i.e. the product of the “term frequency” (number of occurrences of the term in the document) and the “inverse of the “term document frequency” (number of documents in which the term appears)
- Consider each document as a vector in n -space (n is the number of distinct terms, i.e. the size of the lexicon)
 - The non-zero components of the vector are the weights of the terms appearing in the document
 - Normalize each vector to “unit length” (divide each component by the modulus – the “length” – of the vector)
- Consider also the query as a vector in n -space
 - The non-zero components are just the terms appearing in the query (possibly with a weight)
 - Normalize also the query vector
- Define the similarity measure between the query and a document as the cosine of the “angle” between the two vectors
 - If both vectors are normalized, the computation is just the inner product of the two vectors

Final weight: $tf \times idf$ (or $tf.idf$)



- In conclusion, the weight of each term i in each document d ($w_{i,d}$) is usually given by the following formula (or very similar variations), called the *tf.idf* weight

$$w_{i,d} = tf_{i,d} \times \log(n / df_i)$$

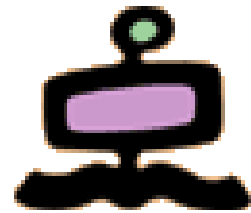
$tf_{i,d}$ = frequency of term i in document d

n = total number of documents

df_i = the number of documents that contain term i

- Increases with the number of occurrences *within* a doc
- Increases with the rarity of the term *across* the whole corpus

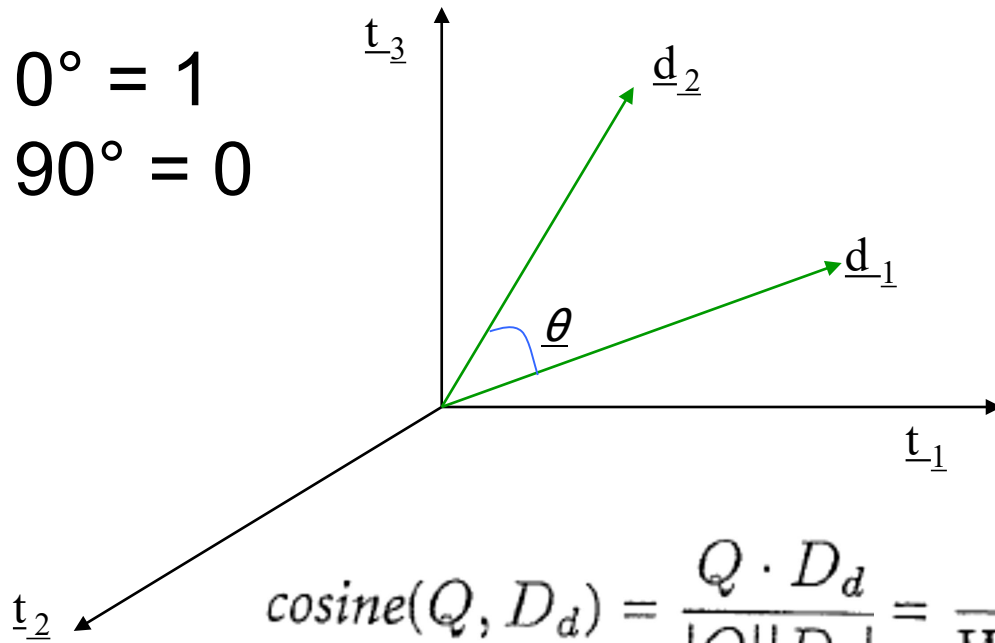
Similarity in vector space



- Similarity between vectors d_1 and d_2 is *captured* by the **cosine** of the angle x between them.
- Note – this is *similarity*, not distance

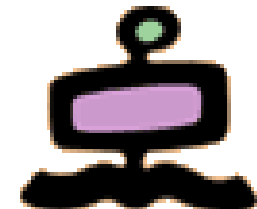
$$\cos 0^\circ = 1$$

$$\cos 90^\circ = 0$$



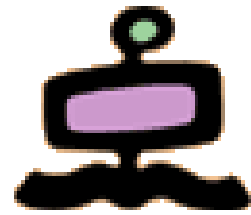
$$\text{cosine}(Q, D_d) = \frac{Q \cdot D_d}{|Q||D_d|} = \frac{1}{W_q W_d} \sum_{t=1}^n w_{q,t} \cdot w_{d,t}$$

“Boosting” of terms

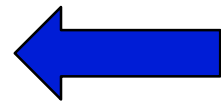


- A term coming from an HTML page is “more important” if it is:
 - In the title tag
 - In the page URL
 - In an HTML heading
 - In capital letters
 - Larger font
 - Early on in the page
 - In an HTML metatag
 - in the anchor text of a link pointing to that page
- A set of query terms is more important if they appear in the page:
 - Close together
 - In the right order
 - As a phrase

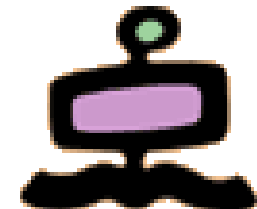
Main functions of a search engine



- Crawling
- Indexing (in parallel with crawling)
- Ranking based on page content
- Ranking based on Web considerations
- Display of results

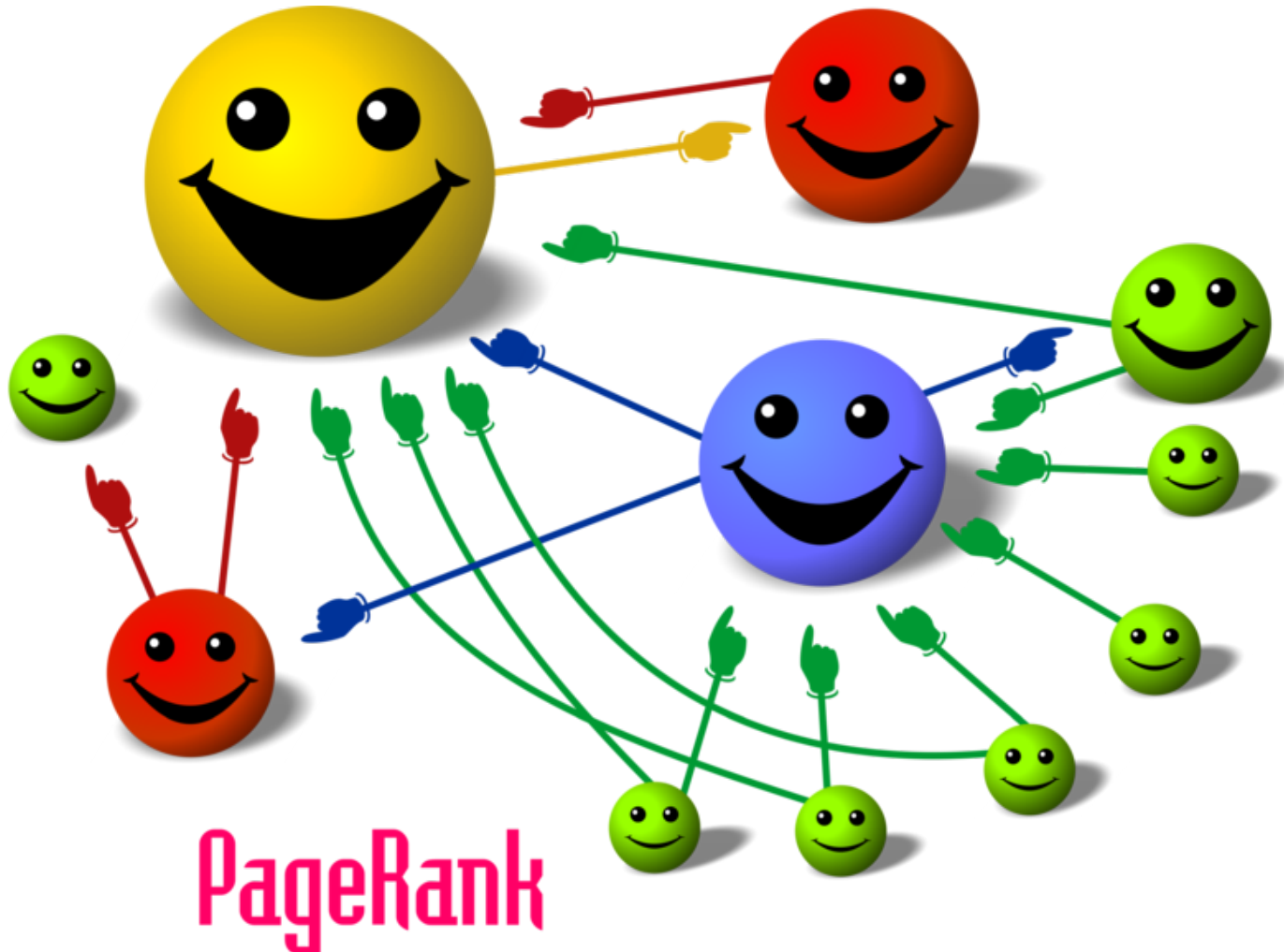
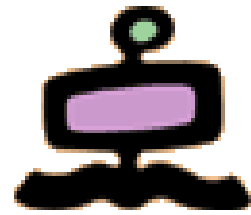


Ranking in the Web

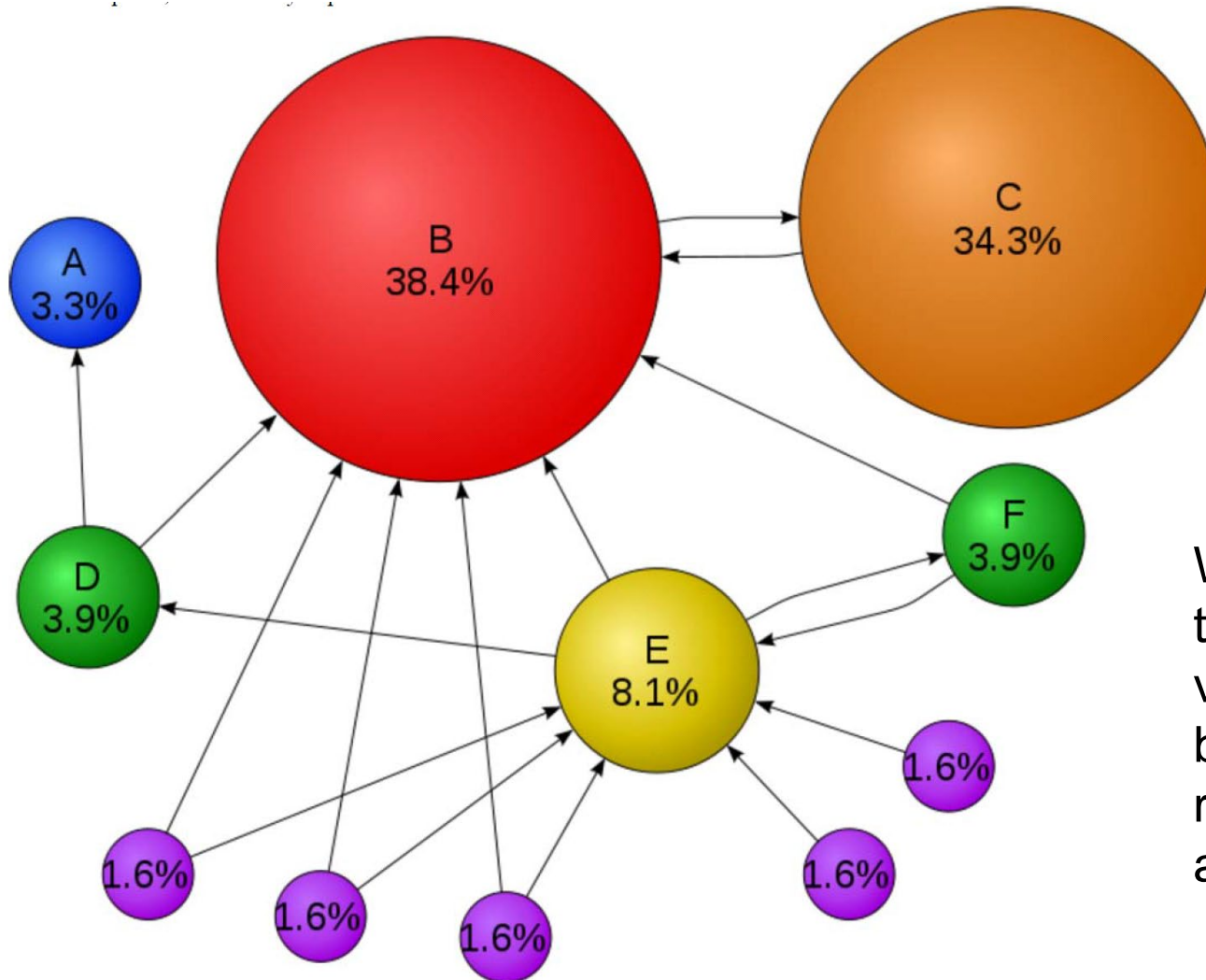
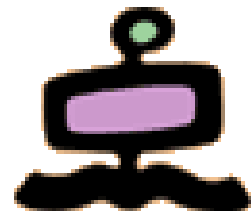


- Searching and ranking for “relevant documents” in a collection depends only on the content of the documents (free text search/retrieval, “bag of words” model)
- In the web, however, in addition to the page content there is the information provided by the hyperlinks from one web page to another
- The idea is therefore to rank the relevance of a web page based also on its “popularity” in the web, i.e. the number of links pointing to it from other web pages

The PageRank idea

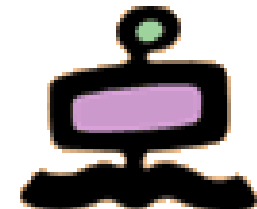


The PageRank values



We can consider the PageRank value as a number between 0 and 1, represented here as a percentage

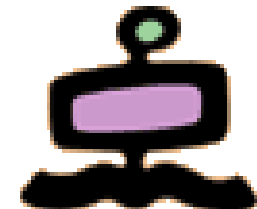
The PageRank algorithm



- The PageRank algorithm was published in 1996 by two students at Stanford University (Larry Page and Sergey Brin, the founders of Google)
 - the patent belongs to the University of Stanford and Google has the exclusive right to it
- The PageRank of a page is the sum of the values of the links pointing to it
- The value of an outgoing link is the PageRank of the page containing the link divided by the total number of outgoing links from that page
- Simple example for a “Web” of four pages, where pages B, C and D contain a link to page A:

$$PR(A) = \frac{PR(B)}{L(B)} + \frac{PR(C)}{L(C)} + \frac{PR(D)}{L(D)}.$$

The PageRank algorithm

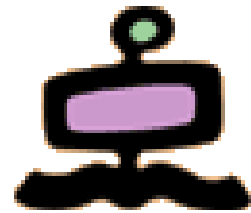


- More in general:

$$PR(u) = \sum_{v \in B_u} \frac{PR(v)}{L(v)}$$

- where B_u is the set of pages pointing to page u and $L(v)$ is the number of outgoing links in page v
- In the mathematical model behind the PageRank algorithm, the rank of a page represents the probability that a **random surfer** sooner or later will land on that page
 - a random surfer starts navigation from random page of the web
 - clicks at random a link on that page
 - goes on “forever”
- The above is valid for a web completely connected
 - What if a page does not have outgoing links ?
 - What if a page does not have incoming links ?

Complete PageRank algorithm

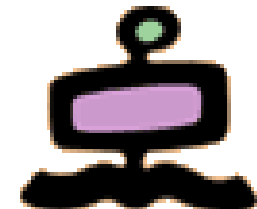


- To take into account “dangling pages”, the random surfer model is modified
 - At each page, the surfer can choose between clicking a link on that page, or jumping to a new page at random
 - The probability that the surfer clicks a link on that page is called the “damping factor”
- The final formula is (d is the damping factor, between 0 and 1, usually set at 0,85):

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

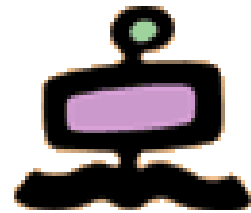
N is the total number of pages

Calculating the PageRank



- PageRank is a “normal” problem of linear algebra
 - a system of N equations in N unknowns
- For big (huge) systems, mathematicians have developed “iterative” ways to solve the system
 - all the pages are assigned an initial value (usually the same, $1/N$)
 - the system is solved to get new values
 - the new values are assigned to the pages
 - the process is repeated until the difference with the previous step is negligible
- In the real Web, the number of iterations is in the order of 100, and the computation of the PageRank for all the pages may take several days

Search Engines considerations



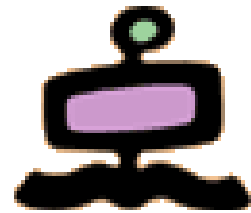
- Collection of query data (for statistics)
 - topics
 - time and location
 - number of clicks
- Search Engines Optimization (SEO)
 - objective is to increase the PageRank of a page
 - increase the PageRank of the pages pointing to it
 - divide a Web site into many pages
- Advertising on search engines
 - high volume of visitors
 - “knowledge” of web page content
 - targeted advertising

Just

0.63%

of people click on the second page of Google search results.

Table of SEO ranking factors



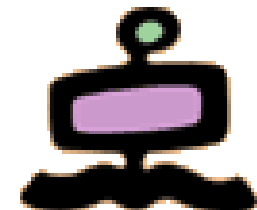
On-The-Page SEO			Off-The-Page SEO			
Content	Architecture	HTML	Trust	Links	Personal	Social
Cq ⁺³ Quality	Ac ⁺³ Crawl	Ht ⁺³ Titles	Ta ⁺³ Authority	Lq ⁺³ Quality	Pc ⁺³ Country	Sr ⁺² Reputation
Cr ⁺³ Research	Am ⁺³ Mobile	Hd ⁺² Description	Te ⁺² Engage	Lt ⁺² Text	Pl ⁺³ Locality	Ss ⁺¹ Shares
Cw ⁺² Words	Ad ⁺² Duplicate	Hs ⁺² Structure	Th ⁺¹ History	Ln ⁺¹ Numbers	Ph ⁺² History	
Cf ⁺² Fresh	As ⁺² Speed	Hh ⁺¹ Headers	Vd ⁻¹ Piracy	Vp ⁻³ Paid		
Cv ⁺² Vertical	Au ⁺¹ URLs	Vs ⁻² Stuffing	Va ⁻¹ Ads	Vi ⁻³ Spam		
Ca ⁺² Answers	Ah ⁺¹ HTTPS	Vh ⁻¹ Hidden				
Vt ⁻² Thin	Vc ⁻³ Cloaking					

Factors Work Together

All factors on the table are important, but those marked 3 carry more weight than 1 or 2. No single factor guarantees top rankings or success, but having several favorable ones increases the odds. Negative "violation" factors shown in red harm your chances.

+3	
+2	
+1	
-1	
-2	
-3	

On-page SEO ranking factors



On-The-Page Factors

These elements are in the direct control of the publisher

Content

Cq

QUALITY

Are pages well written & have substantial quality content?

Cr

RESEARCH

Have you researched the keywords people may use to find your content?

Cw

WORDS

Do pages use words & phrases you hope they'll be found for?

Cf

FRESH

Are pages fresh & about "hot" topics?

Cv

VERTICAL

Do you have image, local, news, video or other vertical content?

Ca

ANSWERS

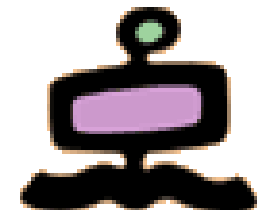
Is your content turned into direct answers within search results?

Vt

THIN

Is content "thin" or "shallow" & lacking substance?

On-page SEO ranking factors



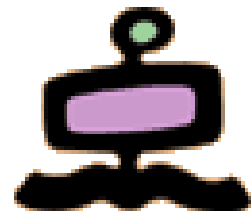
On-The-Page Factors

These elements are in the direct control of the publisher

Architecture

Ac	CRAWL	Can search engines easily "crawl" pages on site?
Am	MOBILE	Does your site work well for mobile devices?
Ad	DUPLICATE	Does site manage duplicate content issues well?
As	SPEED	Does site load quickly?
Au	URLS	Do URLs contain meaningful keywords to page topics?
Ah	HTTPS	Does site use HTTPS to provide secure connection for visitors?
Vc	CLOAKING	Do you show search engines different pages than humans?

On-page SEO ranking factors



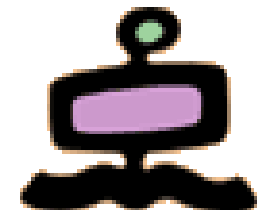
On-The-Page Factors

These elements are in the direct control of the publisher

HTML

Ht	TITLES	Do HTML title tags contain keywords relevant to page topics?
Hd	DESCRIPTION	Do meta description tags describe what pages are about?
Hs	STRUCTURE	Do pages use structured data to enhance listings?
Hh	HEADERS	Do headlines & subheads use header tags with relevant keywords?
Vs	STUFFING	Do you excessively use words you want pages to be found for?
Vh	HIDDEN	Do colors or design "hide" words you want pages to be found for?

Off-page SEO ranking factors



Off-The-Page Factors

Elements influenced by readers, visitors & other publishers

Trust

Ta

AUTHORITY

Do links, shares & other factors make pages trusted authorities?

Te

ENGAGE

Do visitors spend time reading or "bounce" away quickly?

Th

HISTORY

Has site or its domain been around a long time, operating in same way?

Vd

PIRACY

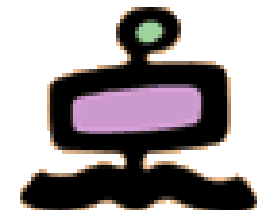
Has site been flagged for hosting pirated content?

Va

ADS

Is content ad-heavy? Do you make use of intrusive interstitials?

Off-page SEO ranking factors



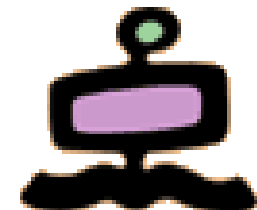
Off-The-Page Factors

Elements influenced by readers, visitors & other publishers

Links

Lq	QUALITY	Are links from trusted, quality or respected web sites?
Lt	TEXT	Do links pointing at pages use words you hope they'll be found for?
Ln	NUMBER	Do many links point at your web pages?
Vp	PAID	Have you purchased links in hopes of better rankings?
VI	SPAM	Have you created links by spamming blogs, forums or other places?

Off-page SEO ranking factors



Off-The-Page Factors

Elements influenced by readers, visitors & other publishers

Personal

Pc COUNTRY What country is someone located in?

Pl LOCALITY What city or local area is someone located in?

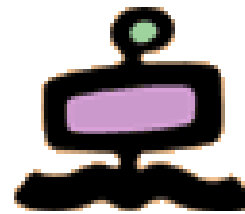
Ph HISTORY Has someone regularly visited your site?

Social

Sr REPUTATION Do those respected on social networks share your content?

Ss SHARES Do many share your content on social networks?

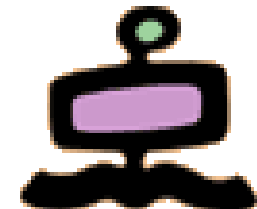
Table of SEO ranking factors



Cr CLIQUE Search engines must be able to identify your page.	Qu QUALITY Pages must be well written and have substantial quality.	Mo MOBILE USE Optimize for smartphones and tablets.	Rs REWARDS Choose keywords people may use to find your content.	St STRUCTURE Make sure the page is structured to be found.	Tt TITLE Make keywords into your title.	Cl CLIPPING Don't abuse the engine's different content type page allowing variations.	Sc SCHEMATA Using links, opening tags and so on are signs of quality to get your content.	Bc BAD CONTENT Google penalizes bad content, content, and doorway pages.	Ux USER EXPERIENCE Use your website's user experience that makes your user want to visit and return.	Mb MOBILE-BEHAVIOR Planning, design, content and loading will get your mobile use of the device.	GMB GOOGLE MY BUSINESS A Google My Business listing is essential to get based on the local map.	Ts TOPIC SIGNAL Content you also is able to be found on the website of your niche.	At AUTHORITY Factor on how much the website and page have any content that they need.		
An ANALYSIS Create content that can be used for analysis in the SERP.	Pr PRICING They determine value and bidding opportunities.	Sc SCHEMATA Document like structured data and other things.	Dt DATA Make content data for your audience.	Ur URL Make keywords into your page's URL address.	Hd HEADINGS Make keywords into your page's local address.	It INTENT Consider why someone is searching to guide the search.	Sf SERIALS Don't be overzealous with page frequency (SERP only).	Lc LOCAL The website location gets into what results show up in a local search.	De DEVELOPMENT Google's device that provides used to make the user more comfortable with.	Mc MEASUREMENT Apply to your website.	Lc LOCAL The website location gets into what results show up in a local search.	De DEVELOPMENT Google's device that provides used to make the user more comfortable with.	Mc MEASUREMENT Apply to your website.		
Kw KEYWORDS Make sure keywords are on your page.	Dd DEVELOPMENT Be smart. Don't overdo it.	CLS CONTENT LENGTH Shorter content will open up more space.	Ex EXPERIENCE Content should be your top priority on your page.	Lq LINK QUALITY Only links from quality websites.	Ly LOCALITY Consider location, user intent and other things.	Hi HISTORY Check your history in a search engine.	Ci CONTENT Check your content in a search engine.	Av AUTHORITY A strong, well optimized website will be a good place in your industry.	Sd SERIALIZED DATA Include structured data for products, offers, reviews.	NAP NAME, ADDRESS, AND PHONE Number must be consistent across all digital channels.	Ir IMAGE RICHNESS Use tags to help in finding images on the page.	URL URL Make the URL easy to read and include your brand name.	Rv REVIEW RICHNESS If your content is rich in reviews, it's a good sign.	Fs FEATURED SNIPPETS If your content is rich in featured snippets, it's a good sign.	Pg PAGE RICHNESS If your content is rich in page elements, it's a good sign.
Fr FRAGMENTATION Create small content, add it to other pages.	Sp SPINNING Use your content to create other pages.	Ds DESCRIPTIONS Use tags to help in finding images on the page.	Au AUTHORITY Content should be your top priority on your page.	Ac ANCHORS Use your link text to help in finding images on the page.	Cy CONTENT Check your content in a search engine.	Ar ARCHIVE Check your history in a search engine.	Rv REVIEW RICHNESS If your content is rich in reviews, it's a good sign.	Fs FEATURED SNIPPETS If your content is rich in featured snippets, it's a good sign.	Pg PAGE RICHNESS If your content is rich in page elements, it's a good sign.						
Mm MULTIMEDIA Use tags to help in finding images on the page.	Ps PAGES Use your content to create other pages.	ALT ALTERNATIVE TEXT Use tags to help in finding images on the page.	Tr TRUST Content should be your top priority on your page.	Qt QUALITY Use your link text to help in finding images on the page.	Eg EMBEDDED CONTENT Check your content in a search engine.	Iv IMAGE RICHNESS Use tags to help in finding images on the page.	Fs FEATURED SNIPPETS If your content is rich in featured snippets, it's a good sign.	Pg PAGE RICHNESS If your content is rich in page elements, it's a good sign.							

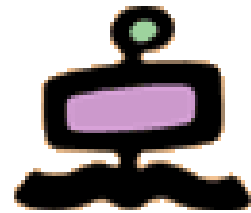
NICHES

Violations



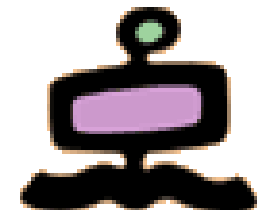
- Link spam
 - Link farms
 - Hidden links
 - Sybil attack
 - Page hijacking
 - Buying expired domains
 - Cookie stuffing
 - Using world-writable pages
 - Blog spam
 - Comment spam
 - Wiki spam
 - Referrer log spamming
- Content spam
 - Keyword stuffing
 - Hidden or invisible text
 - Meta-tag stuffing
 - Doorway pages
 - Scraper sites
 - Article spinning
- Other types
 - Mirror websites
 - URL redirection
 - Cloaking

Search Engines considerations



- Search Engines Optimization (SEO)
 - increase the number of incoming links (link farms)
 - increase the PageRank of the pages pointing to it
 - divide a Web site into many pages
- Collection of query data (for statistics)
 - topics
 - time and location
 - number of clicks
- Advertising on search engines
 - high volume of visitors
 - “knowledge” of web page content
 - targeted advertising

Advertising on Google



Your ads appear beside related search results...

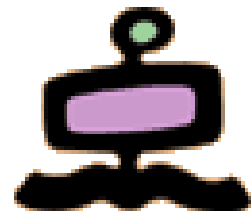
People click your ads...

...And connect to your business



- advertising is associated to “key words” (Google AdWords)
- ads are published on the result page of a query containing a keyword
- ads are paid “per click”
- ads may be published also on “partner sites” (Google AdSense)

Google advertising revenues



Advertising revenue of Google from 2001 to 2023 (in billion U.S. dollars)

