

WEB SEARCH
An Introduction

CT102
Information
Systems











HOW OFTEN DO YOU ...

- Use a search engine?
- Use chatGPT or similar?
- Check social media?
- Check in to apps (exercise, banking, education, sports, news, shopping, etc.).?
- Stream music, videos, etc.?
- Play games online?
- Check email, whatsapp, etc.?

Could you even estimate how often you do these things per day?

TOP 10 MOST VISITED SITES

[Free Tools](#) ▾[Products](#) ▾[Our Customers](#) ▾[Our Data](#)[Pricing](#)[Resources](#)[Ranking](#)[Trending](#)

Rank ^①	Website ^①	Category ^①	Rank Change ^①
1	 google.com	Computers Electronics and Technology ▸ Search Engines	=
2	 youtube.com	Arts & Entertainment ▸ Streaming & Online TV	=
3	 facebook.com	Computers Electronics and Technology ▸ Social Media Networks	=
4	 instagram.com	Computers Electronics and Technology ▸ Social Media Networks	=
5	 twitter.com	Computers Electronics and Technology ▸ Social Media Networks	=
6	 baidu.com	Computers Electronics and Technology ▸ Search Engines	=
7	 wikipedia.org	Reference Materials ▸ Dictionaries and Encyclopedias	=
8	 yahoo.com	News & Media Publishers	=
9	 yandex.ru	Computers Electronics and Technology ▸ Search Engines	=
10	 whatsapp.com	Computers Electronics and Technology ▸ Social Media Networks	=

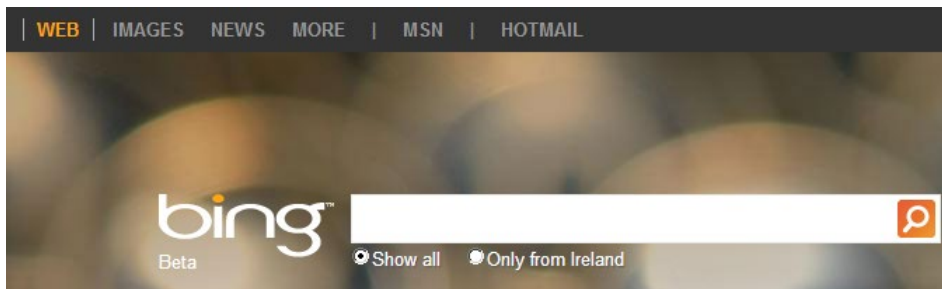
WEB SEARCH ENGINES

Is Google your
usual web search
engine?

A web search engine is an online web **information retrieval system** that, given a query, which represents a user's information need, returns a list of web pages that match that query.

YAHOO!

Search web



Google™
Ireland

Google Search

I'm Feeling Lucky

Search: ☒ the web ☐ pages from Ireland

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

SEARCH ENGINE MARKET SHARE WORLDWIDE

2023, 2022 AND 2020



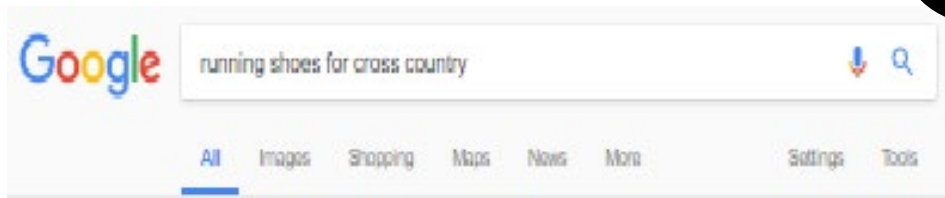
GOOGLE'S DOMINANCE ...



Worldwide, Google processes about 99,000 searches per second = approximately 8.5 billion searches per day

CLASS QUESTION:

What is
happening in
web search?



Think about this question – answer
briefly if you can

ANSWER?!

It will actually take a little time to fully answer this question!

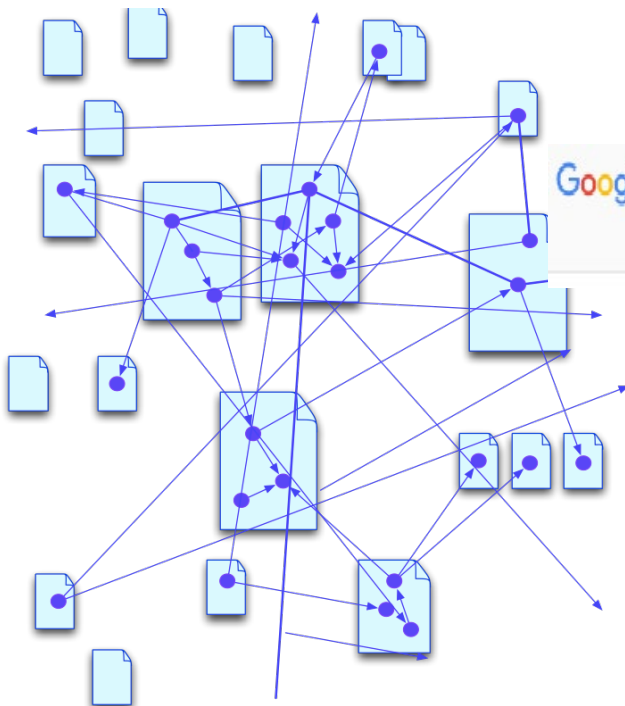
And in order to answer it, we need to first talk about the type of data web search works with and the kind of search it does

SEARCH ENGINE OVERVIEW:

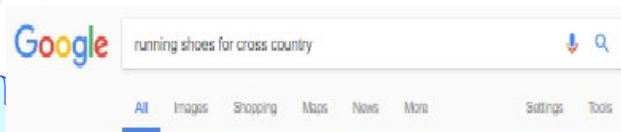
Inputs, processing, outputs

Data → Information

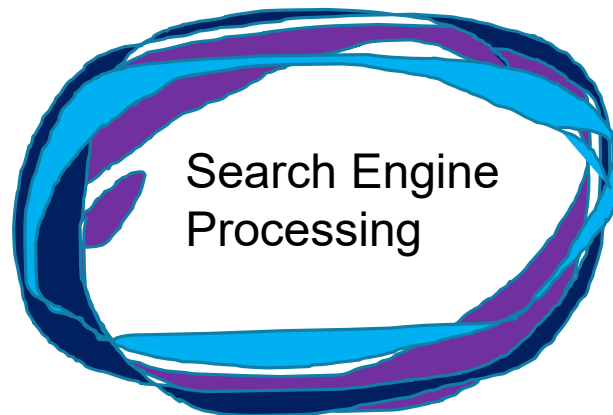
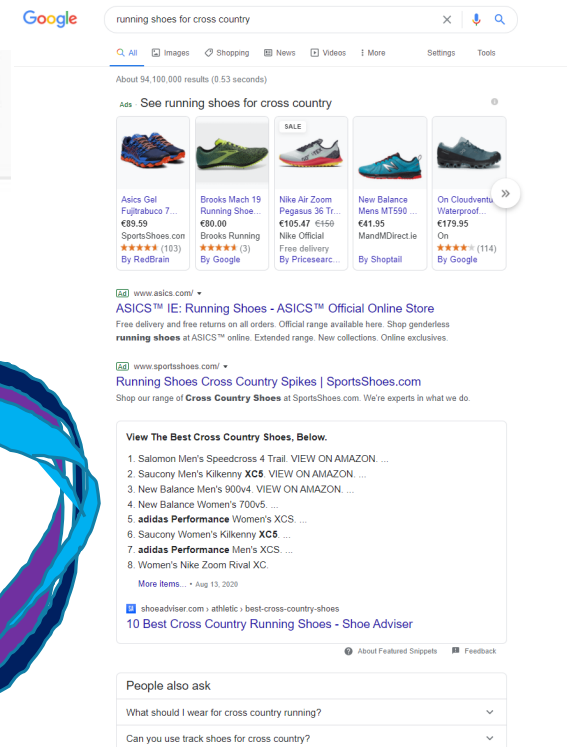
The data?



Input: The query



The output: The results



DATA ON THE WEB

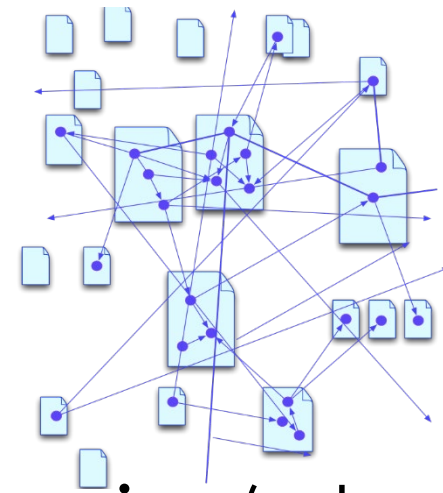
tic shows the global market share of leading internet search engines. In July 2018, Chinese search engine Baidu had a market share of 1.08 percent.

```
r>
earch in 1997, the worldwide market share of all search engines has been rather
re as of July 2018. The majority of Google revenues are generated through <a hr
as also expanded its services to mail, productivity tools, enterprise products,
googles-annual-global-revenue/">highest tech company revenues in 2017 with roug
class="js-readingAid__gradient readingAid__gradient readingAid__gradient--invi
ute"
ReadSup" data-gtm="descriptionReadMore"
de"><span>Show more</span><i class="fa fa-caret-down margin-left-5"></i></span>
t--link link hideMobile"
tatistic216573" , "name": "Popup: Premium Account", "creative": "Global&#x20;ma
="paywall_c2a--sources--1"
    data-modal="#popupOverlay"
    data-file="sources" data-gtm="paywall_c2a--sources">
    </dd><dd href="/accounts/" class="text--link link"
tatistic216573" , "name": "Popup: Premium Account", "creative": "Global&#x20;ma
="paywall_c2a--publish"
    data-modal="#popupOverlay"
    data-file="publisher" data-gtm="showPublishLink">
    </dd><dt>Release date</dt><dd>

v id="info" class="tabContent tabContent--noPadding js_hidden"><dl><dt>Region</
e period</dt><dd>
0 to July 2018
operties</dt><dd>

iv><div class="actionBar float-right margin-top-10"><button id="statisticBrowse
here to be redirected to see all your recently viewed statistics." data-toolti
="fa fa-history" aria-hidden="true"></i></button><button class="button button--
tatistic216573" , "name": "Popup: Premium Account", "creative": "Global&#x20;ma
```

DATA ON THE WEB



Typically **general-purpose web search engines** (such as Google) deal with data that:

- Has large portions of **unstructured** data which have (hyper)**links** to other web pages.
- Has smaller portions of structured data (ad data) and semi-structured **semantically-tagged** data (e.g., dbpedic similar).
- Often much of the data used in searching is in **natural language** (e.g., English), even if the results returned are videos, images, etc.



STRUCTURED, SEMI-STRUCTURED AND UNSTRUCTURED DATA

Structured data: data that resides in a fixed field within a record or file, e.g., often relational (or other) database approach.

Semi-structured data: does not have a formal structure but does have tags or other information that convey meaning of data, e.g., XML or RDF documents with headings/sections, email, etc.

Unstructured data: data is not organised in any obviously meaningful way.

Approximately, 80-90% of the data we generate today is unstructured.

WHICH IS WHICH?

Fair Daffodils, we weep to see
You haste away so soon;
As yet the early-rising sun
Has not attain'd his noon

```
<quiz>
<qanda seq="1">
  <question>
    Who was the forty-second
    president of the U.S.A.?
  </question>
  <answer>
    William Jefferson Clinton
  </answer>
</qanda>
```

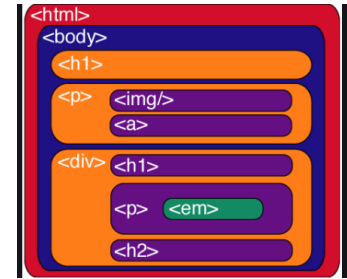
```
<a name="d.en.83884"></a>
<div class="column-content richTextBlock">
<h3>Welcome Message from the Head of Computer Science</h3>
<p class="person"> The School of Computer Science is the largest academic disci-
of&nbsp; Science and Engineering and Informatics</a> and one of the largest in NUI Galway overall. In 2
target=" blank">QS World University Rankings</a> put Computer Science & Information Systems @NUIG i
<p>There are <a href="/engineering-informatics/information-technology/people/">22 academic staff and 5
over 80 full-time researchers at M.Sc., Ph.D. and postdoctoral level. We are actively engaged in a wide
technology/research/researchtopics/">research topics</a> in areas such as Artificial Intelligence; Mach
Communications; Internet of Things; Image Processing; Simulation; Evolutionary Computation; and Informa
Health Informatics, Energy Informatics, Enterprise Systems, Cyber-Security, Social Network Analysis, Di
research awards from Science Foundation Ireland, Irish Research Council, Enterprise Ireland, Health Res
<p>We have close to 700 students on our comprehensive suite of taught and research programmes at undergrad
<a href="/courses/undergraduate-courses/computer-science-and-information-technology.html">BSc in Comput
informatics/information-technology/programmes/undergraduateprogrammes/itasasubjectforartstudents/">Bac
programmes include the <a href="http://www.it.nuigalway.ie/engineering-informatics/information-technolo
courses/softwaredesignanddevelopmentmscexternalstream/">MSc in Software Design and Development</a>, the
technology/programmes/postgraduate-courses/softwaredesignanddevelopmentmscexternalstream/">Higher Diplo
informatics/information-technology/programmes/postgraduate-courses/softwaredesignanddevelopmenthdipapps
Stream)</a>, the <a href="/engineering-informatics/information-technology/programmes/postgraduate-cours
Analytics/>&nbsp;&nbsp;&, <a href="http://www.it.nuigalway.ie/engineering-informatics/information-technology/
Artificial Intelligence</a>&nbsp;&nbsp;& and two programmes that are delivered entirely online in conjunction w
informatics/information-technology/programmes/postgraduate-courses/softwareengineeringanddatabasetechno
href="/engineering-informatics/information-technology/programmes/postgraduate-courses/softwareengineeri
href="http://www.nuigalway.ie/science/undergraduate-courses/science-undenominated.html">Bachelor of Sci
Science</a>), and the MA in Digital Media (with the <a href="http://www.filmsschool.ie/filmsschool/">Hust
of the Bachelor of Engineering degrees.</p>
<p>In addition to our research and teaching, our staff and students are heavily engaged with the commun
(free introductory computer classes for the digitally excluded), CoderDojo (an international movement o
(the Galway makerspace) and Galway Games Group.</p>
<p>We are located in the Information Technology Building, a dedicated 4100 square metre building in mid
students on taught programmes, dedicated research office space for our researchers, state-of-the-art eq
```

ou,
y,
y Away,

employee									
	FName	MII	LName	SSN					
+	John	B	Smith	123456789					
+	Franklin	T	Wong	333445555	08/12/1955	638 Voss, Houston, TX	M	€40,000.00	888665555
+	Joyce	A	English	453453453	31/07/1972	5631 Rice, Houston, TX	F	€25,000.00	333445555
+	Ramesh	K	Narayan	666884444	15/09/1959	675 Rice, Houston, TX	M	€38,000.00	333445555
+	James	E	Borg	888665555	10/11/1955	638 Voss, Houston, TX	M	€40,000.00	888665555
+	Jennifer	S	Wallace	987654321	20/06/1955	638 Voss, Houston, TX	M	€40,000.00	888665555
+	Ahmad	V	Jabbar	987987987	29/03/1955	638 Voss, Houston, TX	M	€40,000.00	888665555
+	Alicia	J	Zelaya	999887777	19/07/1955	638 Voss, Houston, TX	M	€40,000.00	888665555

000000000000000043300000300040000000000
00435034000000000000000000000000000000
000000000003000000000000000000000030004040
050000000000000000000000000000000040404300000
004050000000000200000000000000000030000300

Characteristics of HTML files

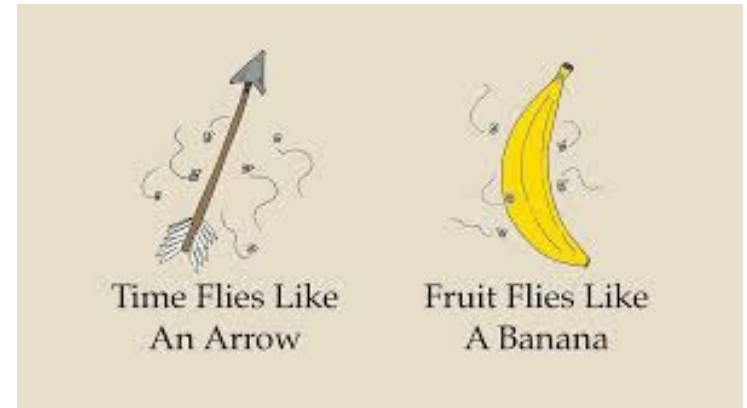


- The files can contain natural language text, audio, images, video, etc.
- HTML tags define the format of the content (headings, bullet points, etc.) From these tags we can sometimes infer importance of certain text, e.g. `<title>` indicates the title and if correct will give the words that are probably most important in the page but they do not give us *meaning*.
- HTML files contain A LOT of formatting tags which must be stripped away by search engines.
- One important HTML tag is the `href` tag
- HTML files are displayed/rendered by a browser but this is not the view a program (spider or scraper) sees – it sees the raw HTML file.

Go to your favourite web page, right click on page and choose 'View page source'

Natural language is generally *unstructured* and *meaning* is not easy to determine

- Writing programs to “read”/process, decipher, “understand” and make sense of (analyse) human languages is a difficult task.
- “Language is compositional”, i.e., letters form words, words form phrases and sentences and the meaning of a phrase can be “larger” than the individual words that comprise it.



Fair Daffodils, we weep to see
You haste away so soon;
As yet the early-rising sun
Has not attain'd his noon.
Stay, stay, Until the hasting day
Has run But to the even-song;
And, having pray'd together, we
Will go with you along.

Natural language is generally *unstructured* and *meaning* is not easy to determine

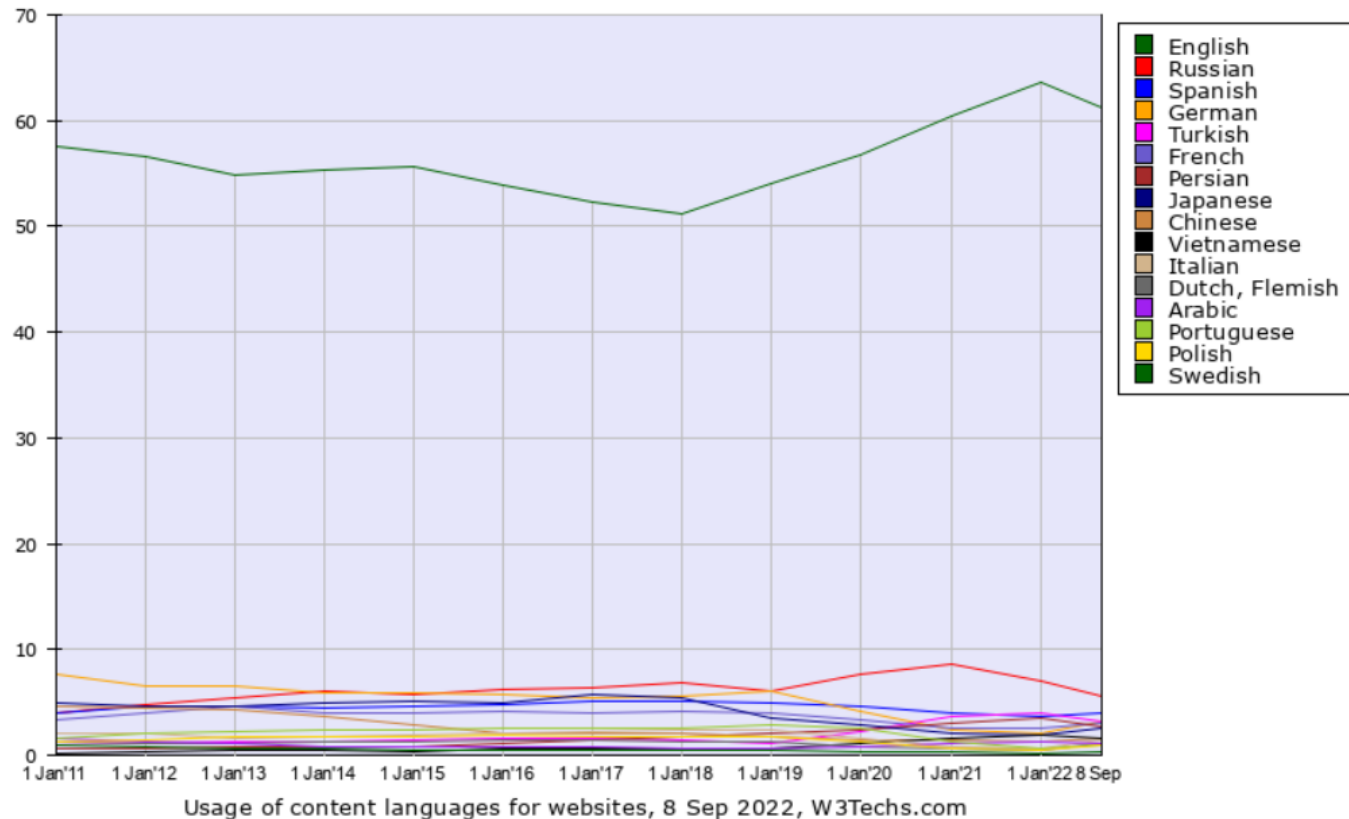
- Recent success has been due to “deep learning” (machine learning) techniques that learn from huge amounts of data and create “LLMs” – large language models – such as chatGPT etc.
- Many traditional and existing techniques still use *statistical approaches* which derive meaning from frequencies of letters, symbols, words, etc. Essentially large language models also try to infer meaning using the same ideas.

Which (natural) language do you think is most predominantly used for HTML text content?

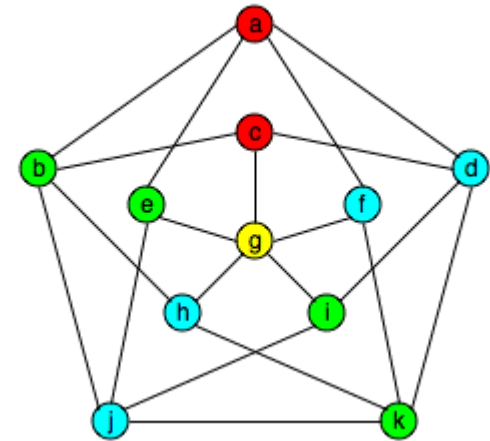
LANGUAGES USED FOR WEBSITES

(https://w3techs.com/technologies/history_overview/content_language/ms/y)

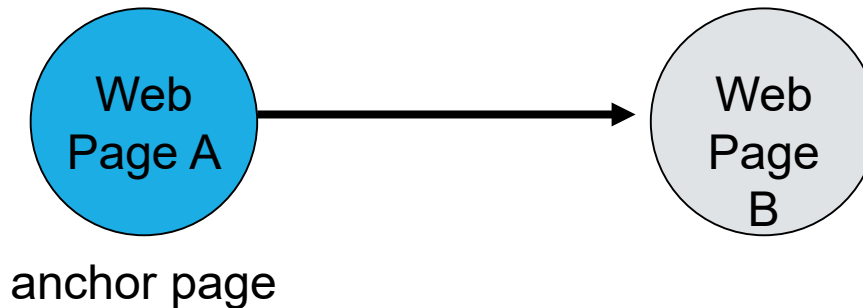
The diagram shows only content languages with more than 1% usage.



ANOTHER IMPORTANT ASPECT OF HTML DOCUMENTS: Linked Data



Can view the static Web as consisting of static HTML pages (containing text, images, video etc.) and **in addition** the hyperlinks between pages



e.g. page A has HTML:

` University of Galway `

where B = <http://www.universityofalway.ie>

FROM OUR WEBSITE

HOME > COLLEGES & SCHOOLS > COLLEGE OF SCIENCE AND ENGINEERING > SCHOOL OF COMPUTER SCIENCE > WELCOME

Welcome

Overview
Welcome
Strategic Plan
Sharepoint (staff only)
Contact Us
Current Students
Prospective Students
Research

Welcome Message from the Head of Computer Science



The School of Computer Science is one of the largest academic disciplines in the [College of Science and Engineering](#) and one of the largest in University of Galway overall. In 2017 the [QS World University Rankings](#) put Computer Science & Information Systems @NUIG internationally in the 201-250 bracket, and second in Ireland.

There are [22 academic staff and 5 technical & administrative staff](#) in Computer Science. We also have over 80 full-time researchers at M.Sc., Ph.D. and postdoctoral level. We are actively engaged in a wide range of [research topics](#) in areas such as

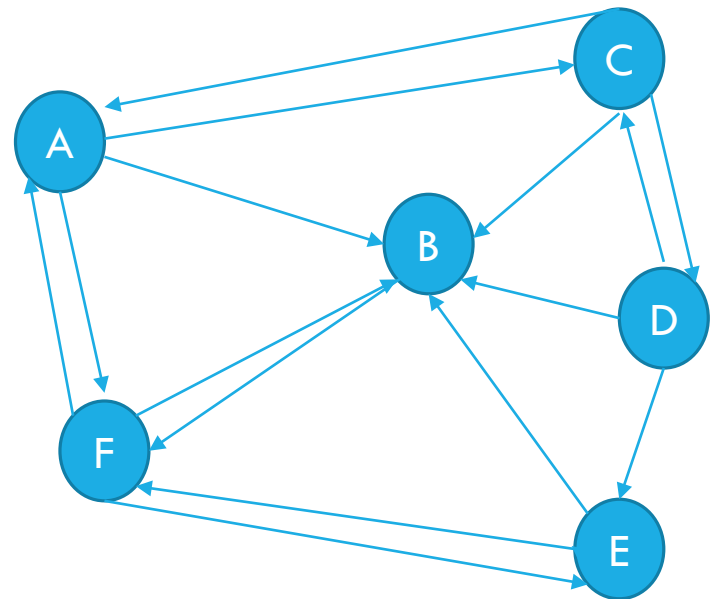
Artificial Intelligence; Machine Learning; Human-Computer Interaction; Medical Informatics; Networks & Communications; Internet of Things; Image Processing; Simulation; Evolutionary Computation; and Information Retrieval. We apply our research expertise in application areas such as Health Informatics, Energy Informatics, Enterprise Systems, Cyber Security, Social

```
421 <div class="column-content richTextBlock">
422 <h3>Welcome Message from the Head of Computer Science</h3>
423 <p class="person">22 ac
425 <p>We have close to 700 students on our comprehensive suite of taught and research progra
426 <p>All of our taught programmes provide a good balance of theoretical and applied content
427 <p>In partnership with other disciplines in the University, we contribute substantially t
428 <p>In addition to our research and teaching, our staff and students are heavily engaged w
429 <p>We are located in the Information Technology Building, a dedicated 4100 square metre b
430 <table border="0">
```

THE WEB GRAPH

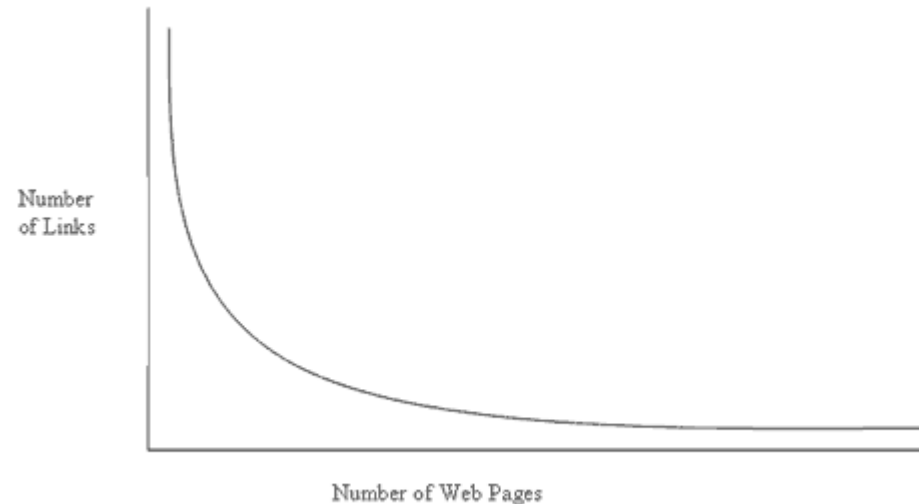
The hyperlink connections between pages can be viewed as a (**directed**) graph

An example of a web graph representation of 6 web pages



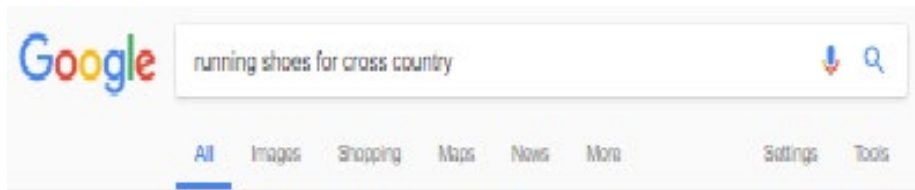
WEB LINK DISTRIBUTION

- Web page links are not randomly distributed.
- Distribution is widely reported to be a *power law*, in which the total number of web pages with in-degree i is proportional to $1/i^c$ (c a constant)
- i.e. only a small portion of web pages have a huge number of links



BACK TO QUESTION AGAIN:

How do web search engines work?

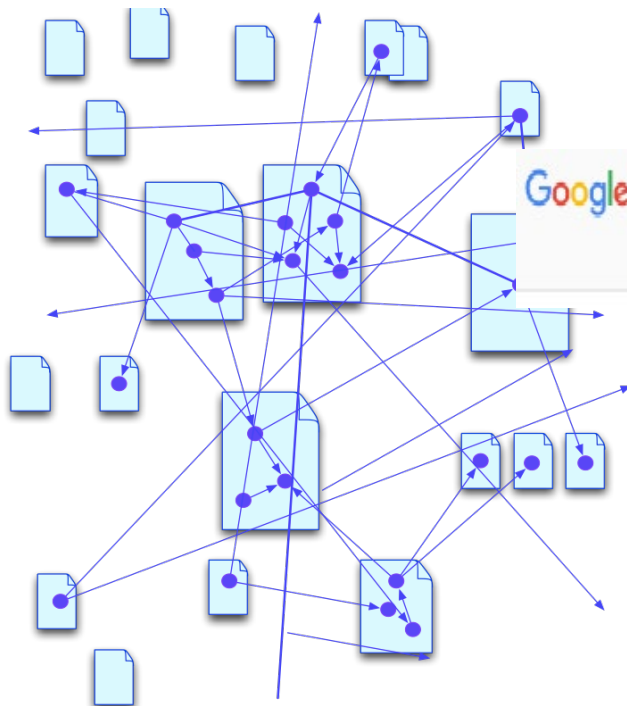


SEARCH ENGINE OVERVIEW:

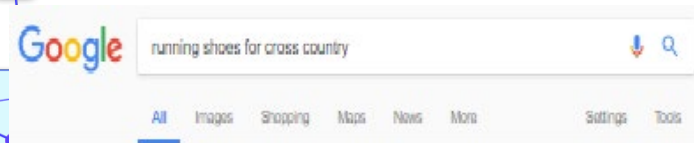
Inputs, processing, outputs

Data → Information

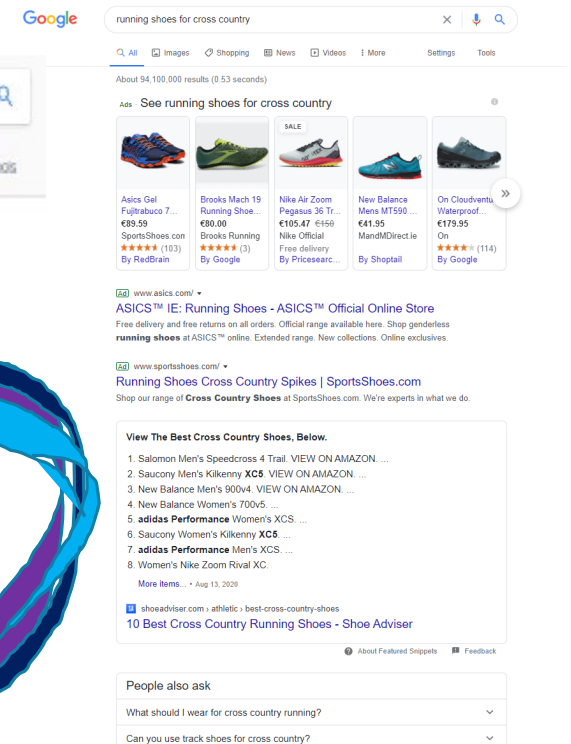
The data?



Input: The query



The output: The results



Input: generally, web search begins with an “information need” ... what is this?

- Information needs are related to **problems**
- Part of “Information seeking behaviour” that a person will engage in given some problem
- The process of “asking” a question of an information system
- Often non-trivial to map or translate an information need in to a query

information need -> query

PROPERTIES OF INFORMATION NEEDS

May be well-defined

May be vague

May contain no text (e.g. an image or tune)

A single correct answer may not exist

The answer may be surprising or not

The answer may be believable or not

Many solutions may match an information need, but a user's tastes and preferences may be the deciding factor in which solution the user deems relevant.

Are some of these “easier” queries than others (from a search engine point of view)

today's weather forecast

pizza delivery

conspiracy theories

accommodation in galway

emploi d'été en france

best route to Rosslare from Galway

SEARCH ENGINE RESULTS: SPONSORED/AD AND ORGANIC CONTENT

In the results returned we can distinguish between **organic** and **sponsored** content in the results window – SERP – Search Engine Results Page

- Sponsored specifically refers to ad data, i.e., paid-for-data
- Organic content refers to data found on web pages “for free”
- **Sponsored** content, if available, is listed (**ranked**) above **organic** content.

The screenshot shows a Google search for "running shoes for cross country". The search bar at the top shows the query and the Google logo. Below the search bar, there are tabs for "All", "Images", "Shopping", "News", "Videos", "More", "Settings", and "Tools". The search results show "About 94,100,000 results (0.53 seconds)".

The first section is labeled "Ads" and "See running shoes for cross country". It displays five sponsored product listings:

- Asics Gel Fujitrabuco 7... €89.59 SportsShoes.com ★★★★★ (103) By RedBrain
- Brooks Mach 19 Running Shoe... €80.00 Brooks Running ★★★★★ (3) By Google
- Nike Air Zoom Pegasus 36 Tr... €105.47 €149 Nike Official Free delivery By Pricesearc...
- New Balance Mens MT590 ... €41.95 MandMDirect.ie By Shoptail
- On Cloudventu Waterproof... €179.95 On ★★★★★ (114) By Google

Below the ads, there are organic search results:

- ASICS™ IE: Running Shoes - ASICS™ Official Online Store**
Free delivery and free returns on all orders. Official range available here. Shop genderless **running shoes** at ASICS™ online. Extended range. New collections. Online exclusives.
- Running Shoes Cross Country Spikes | SportsShoes.com**
Shop our range of **Cross Country Shoes** at SportsShoes.com. We're experts in what we do.

Below the organic results, there is a section titled "View The Best Cross Country Shoes, Below." with a list of 8 shoes:

1. Salomon Men's Speedcross 4 Trail. VIEW ON AMAZON. ...
2. Saucony Men's Kilkenny **XC5**. VIEW ON AMAZON. ...
3. New Balance Men's 900v4. VIEW ON AMAZON. ...
4. New Balance Women's 700v5. ...
5. **adidas Performance** Women's XCS. ...
6. Saucony Women's Kilkenny **XC5**. ...
7. **adidas Performance** Men's XCS. ...
8. Women's Nike Zoom Rival XC.

Below the list, there is a link "More items..." and a date "Aug 13, 2020".

Below the "More items..." link, there is a section titled "10 Best Cross Country Running Shoes - Shoe Adviser" with a link to "shoeadviser.com > athletic > best-cross-country-shoes".

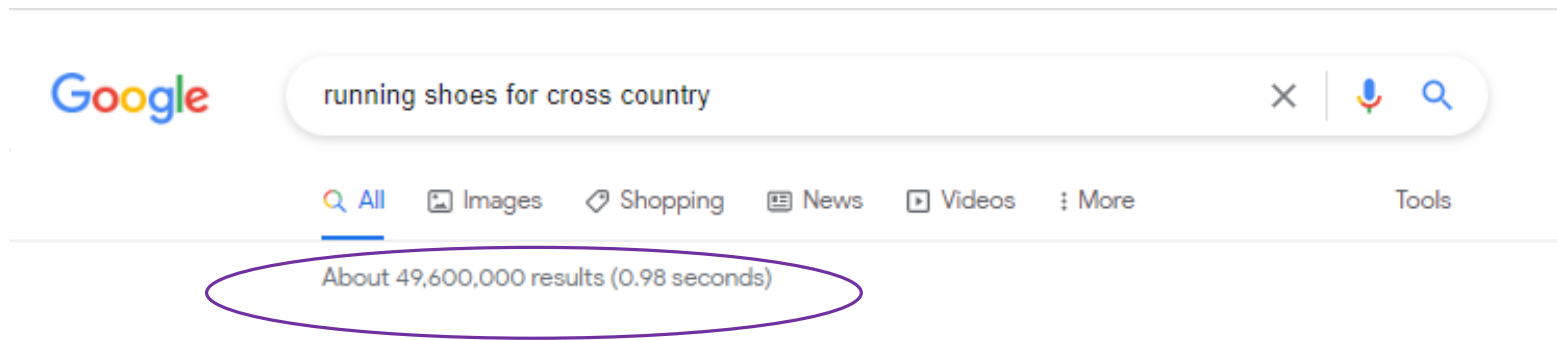
At the bottom, there is a section titled "People also ask" with two questions:

- What should I wear for cross country running?
- Can you use track shoes for cross country?

ORGANIC CONTENT



Returning (usually millions) of web pages in response to a user query:



(Maybe) looks like: matching user query to web pages

But need something like this that is **very scalable** (**works for millions of queries and millions of results**) and **very efficient** (**quick**)

SPONSORED CONTENT

Sponsored ⓘ

Ad

“Sponsored content” is essentially “paid-for-ads”

An additional search occurs independent of the organic search

This search uses a repository (database) of ad words. If any ad words match the query words and the ad passes some “quality tests” then the associated ad is ranked above the web documents returned.

Examples: Compare searching for “nike runners” and “running shoes review”

SEARCH ENGINE RESULTS: Results are ranked

Ranking involves ordering the results
returned in response to a user query

Ranking is based on:

- Business model (e.g., ads first but not all ads)
- Similarity scores (between web pages and ads and query)
- “Importance of Web page” - Page rank scores (using web links)
- Search Engine Optimisation (SEO)
- Ad word scores
- Personalisation scores: Location, Language, profile settings, past search information, etc. (if used by the search engine)

Cross Country Running Shoes - 7 Things High School Runners ...

<https://www.runnersworld.com/.../7-tips-to-help-high-school-runners-choose-the-right-...>

Aug 9, 2018 - As cross-country season begins, it's time for a fresh pair of running shoes. But we know it can be easy to get overwhelmed by the ...

The Best Sneakers for Cross-Country Running | Fitness Magazine

<https://www.fitnessmagazine.com/.../Exercise Equipment/Running Shoes>

The Best Sneakers for Cross-Country Running. Brian Maranan Pineda. New Balance 840. Brian Maranan Pineda. Adidas Supernova Riot. Pearl Izumi Peak XC. These lightweight sneakers are great for fast-paced trail races or training runs. Asics GEL-Trabuco 11 WR. Brooks Cascadia 3. Mizuno Wave Ascend 3. Saucony ProGrid Xodus. ...

Best Cross Country Shoes Reviewed & Compared in 2018 | RunnerClick

<https://runnerclick.com/10-best-cross-country-shoes-reviewed/>

★★★★★ Rating: 5 - Review by Tess Bercan

Jump to Brooks Running Mach 15 - 10 Best Cross Country Shoes. Salomon Speedcross 4. See more images. ASICS GEL Kayano 25. See more images. Brooks Running Mach 15. See more images. Adidas Supernova Riot M. See more images. La Sportiva Wildcat 2.0 GTX. See more images. Saucony Shay XC4 Flat Shoe. See more images. Pearl Izumi Peak 2. See more images. New ... Best Cross Country Shoes - Salomon Speedcross 4 - Adidas Supernova Riot M

What to Wear For Cross Country Running | Run and Become

<https://www.runandbecome.com/running.../what-to-wear-for-cross-country-running>

Cross Country Spikes. Adidas XCS. Women's Adidas XCS. Brooks Mach 18. Women's Brooks Mach 18. Saucony Havok XC. New Balance XC700 V5. Nike Zoom Rival D 9. Junior Adidas Allroundstar.

Men's Cross Country Shoes - Running Warehouse

<https://www.runningwarehouse.com/catpage-MXC.html>

IMPORTANT TO KNOW

N.B. Summary

1. What is web search?
2. What kind data is used in web search?
3. What do we mean by web search links?
4. What is the difference between structured, unstructured and semi-structured data ... give examples
5. Why are some queries more difficult/easy than others?
6. Explain what is meant by sponsored and organic content
7. What does ranking mean? What is web search ranking?