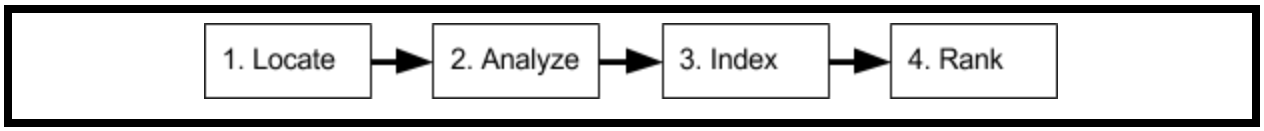
Search II: Web Search

A. (10 min) Web Search	start time:

As shown in the diagram above, the web is **huge**.
As of 2017, there were over 1,000,000,000 (1×10^9) web sites,
and around 3,000,000,000 (3×10^9) Google searches **per day**.

The web and other large data sets present new challenges, since a search system must:

1. **Locate** the web pages and other documents.
2. **Analyze** (read) the documents to get useful content.
3. **Index** the content so it can be searched efficiently.
4. **Rank** the search results to provide the best content.



1. To **locate** new documents, search engines **follow links**. More specifically,
a search engine has programs that step through a list of documents,
look for links to new documents, and add the new documents to the list.
Explain why these are called **web crawlers** or **spiders**.

2. Visiting a web page uses (a small amount of) computer processor time,
network bandwidth, and electricity. Also, most web sites and pages change over time.
What problems could arise if web crawlers visit a site **too often** or **too rarely**?

3. The first web search engines listed **every** document that used the search terms.

This is no longer a good idea. Why?

4. Why would it help to count the **number of times** search terms appear in each document, and list documents with the **highest counts** first?

5. A developer modifies the search engine code so that it will:

- count 1 for each time the search term appears in the text
- count 5 for each time the search term appears in a heading
- count 10 if the search term appears in the title

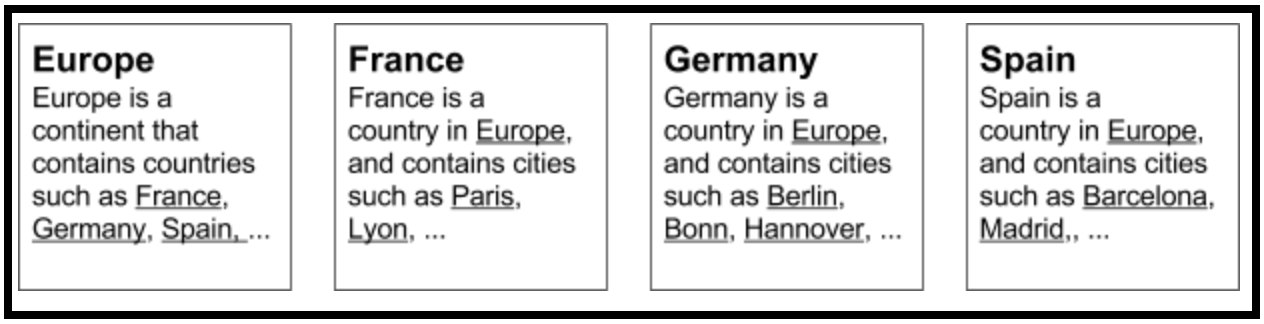
Thus, the final count would be calculated:

final count = 1 * (count in text) + 5 * (count in headings) + 10 * (count in title)

Explain why this might be a good way to rank search results.

6. What other similar factors might be useful?

7. If advertisers know that a search engine will count how many times a search term appears, how could they modify a web page to achieve **better rankings**?



7. The figure above represents 4 web pages about places in Europe.
Each underlined word represents a link to another web page (with that name).

a.	How many links are shown **from** the page for Europe?	
b.	How many links are shown **from** the page for Germany?	
c.	How many pages have links **to** the page for Europe?	
d.	How many pages have links **to** the page for France?	

8. Each link is **outbound** from one page and **inbound** to the other page.
For a given web page, is it **easier** to count inbound or outbound links? Explain.

9. Explain why the most useful web pages tend to have many inbound links.

10. Explain how we could use **inbound links** to predict how **useful** a web page is.
(This is a key idea behind Google's **pagerank** algorithm).

B. (15 min) Refining Search	start time:

Most web searches produce many results, but the results may not be **useful**.
In this section we will explore some ways to **refine** and **improve** search queries.
Use the table below to answer each question on the next page. Specifically:

- Perform the indicated search(s) and record the number of results.
  Round to 2 or 3 significant digits, and use M (million) and B (billion),
  so 1,234,567,890 would be 1.2B or 1200M.
- Answer the specific question(s) to describe the pattern of results,
  including the effects of quotes, ampersand (&), vertical bar (|), and minus (-).

	Query	**# Results**	**Notes**
1a.	computer science		
1b.	science computer		
2a.	computer		
2b.	science		
3.	"computer science"		
4a.	computer & science		
4b.	computer | science		
5a.	computer -science		
5b.	science -computer		

Note that search result pages often include paid advertisements,
which may be formatted to look like search results. **Ignore ads** in your analysis.

(If you want to use a search engine other than Google, check with the instructor first.)

1. Search **twice**, for *computer science* and *science computer*.
In the table, record the (approximate) number of results for each,
and describe any differences in the lists of results.

2. Search **twice**, for *computer* and *science*. In the table, record the number of results,
and describe how these results differ from those in question 1.

3. Search for *"computer science"* (with quotes). In the table,
record the number of results, and describe how quotes affect search results.

4. Search for *computer & science* and *computer | science*. In the table,
record the number of results, and describe how "&" and "|" affects search results.
Hint: In computing, & often means AND and | often means OR.

5. Search for *computer -science* and *science -computer*. In the table,
record the number of results, and describe how "-" affects search results.

Check with the instructor before you start the following questions.

6. Choose a two word phrase of your own.
Repeat questions 1-5 and answer in the table below.

Hint	Query	# Hits	Description and/or Explanation
w1 w2			
w2 w1			
w1			
w2			
"w1 w2"			
w1 & w2			
w1 | w2			
w1 -w2			
w2 -w1			

7. Summarize similarities or differences between the two sets of results.

C. Metadata	start time:

1. When a search engine finds a document, it reads the content, but it also reads information that **describes** the content. This is called **metadata** (data about data). Describe how each type of metadata might influence the **relevance** of text.

a. the **website** (e.g. nytimes.com, tumblr.com)
b. the **date** it was last updated (e.g. 5 days ago, 5 years ago)
c. the **language** (e.g. Chinese, English)
d. the **file format** (e.g. HTML, PDF, Microsoft Word, plain text)
e. the **formatting or markup** (e.g. title, heading, text, sidebar, footnote)

2. Assign one team member to each **set of queries** (a,b,c, …) in the table below. **Individually** (or **in pairs**, depending how many computers you have), **search** for each query in the set, **record** the **number of results** in the table, look for patterns in the result sets, and describe the effects of the **search operator(s)**.

Set	Query	# Hits	Description and/or Explanation
a.	computer science site:amazon.com		
	computer science -site:amazon.com		
b.	computer science filetype:pdf		
	computer science -filetype:pdf		
c.	intitle:computer science		
	science intitle:computer		
	computer intitle:science		
	intitle: computer intitle:science		
	allintitle: computer science		
d.	inurl: computer science		
	computer inurl: science		
	allinurl: computer science		

3. When all team members have finished, report your findings to each other and discuss.

4. To see how these metadata search operators are useful, consider 3 example searches:
   i. Pictures of "dogs" in JPG files from the Spanish Wikipedia (es.wikipedia.org)
   ii. PDFs containing "proposal", from the National Science Foundation (nsf.gov)
   iii. Pages with "biscuit" in the title, only from businesses in the UK (co.uk).

For each example:
   a. Record the number of results for the basic search query.
   b. Design and test a better search query that use metadata search operators.
   c. Record the number of results with the better search query.

	Basic Search	**a. # Hits**	**b. Better Search with Operators**	**c. # Hits**
i.	dogs			
ii.	proposal			
iii.	biscuit			

In this activity we have explored some of the principles and capabilities used in most search engines. Different search engines (e.g. Amazon, Bing, Google) may have slightly different syntax and different options, and you should invest time to learn how to use them more efficiently.