

---

# CS506/606 - Topics in Information Retrieval

---

Instructors: Steven Bedrick, Brian Roark,  
Emily Prud'hommeaux

Class time: Tu/Th 11:00 a.m. - 12:30 p.m.  
September 25 - December 6, 2012

Class location: WCC 403

Office hours: By appointment

Required texts: Readings will come from papers and texts online

<http://www.cs1u.ogi.edu/people/roark/courses/cse506-TIR/cse506TIRf12.html>

---

## Course overview

---

- Variety of topics in general area of information retrieval (IR).
- Initial lectures: Fundamentals of IR, with particular emphasis on applications of modern NLP techniques and evaluation.
- Remainder of course:
  - Seminar-style review of papers from the IR literature.
  - Visiting speakers.
- Example topics to be covered:
  - practical issues related to web-crawling
  - indexing of raw text and other data, e.g., ASR word lattices
  - query expansion and suggestion methods
  - issues in IR evaluation

---

## Readings

---

- Most readings will be articles available on the website.
- In addition, some online book chapters will be assigned:

Christopher Manning, Prabhakar Raghavan P, and Hinrich Schütze H. *Introduction to Information Retrieval*. Cambridge University Press, 2008. <http://www-nlp.stanford.edu/IR-book/>

Marti Hearst. *Search User Interfaces*. Cambridge University Press, 2009. <http://www.searchuserinterfaces.com>

---

## Required coursework

---

- Active participation in class discussions of research papers (10%)
- Lead discussion of papers in one or more sessions (30%).
- One or more homework assignments involving implementation and evaluation of components of IR systems (20%).
- Term project involving IR (to be approved by one of us, 40%).
  - Cross-language IR
  - Distributed IR
  - Speech IR
  - IR applications

---

# Information retrieval

---

Manning et al. “Information retrieval (IR) is finding material ... of an **unstructured** nature ... that satisfies an information need from within large collections ...”

# A note on structured information



Find People   Find a Business   Reverse Phone   Address & Neighbors

nicole snooki   \* polizzi   new jersey   Find

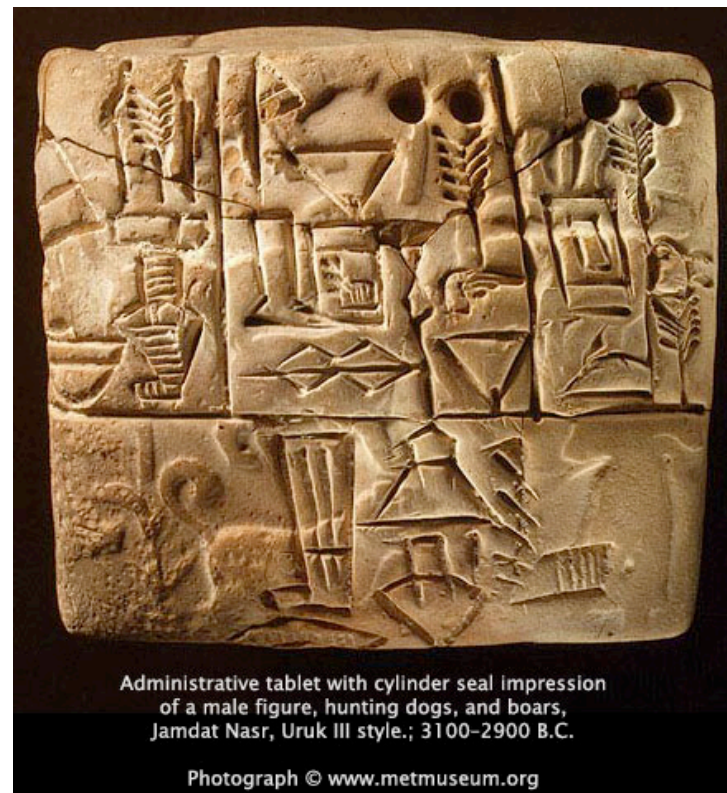
```
SELECT *  
FROM whitepages  
WHERE lastname='polizzi'  
AND firstname='nicole snooki'  
AND state='new jersey'
```

In this class, we'll be talking about unstructured information rather than structured information (e.g., information stored in a database).



# Information retrieval of yore

Manning et al. “Information retrieval (IR) is finding material ... of an unstructured nature ... that satisfies an information need from within large collections ...”



Administrative tablet with cylinder seal impression of a male figure, hunting dogs, and boars, Jemdat Nasr, Uruk III style, 3100–2900 B.C.

Photograph © [www.metmuseum.org](http://www.metmuseum.org)



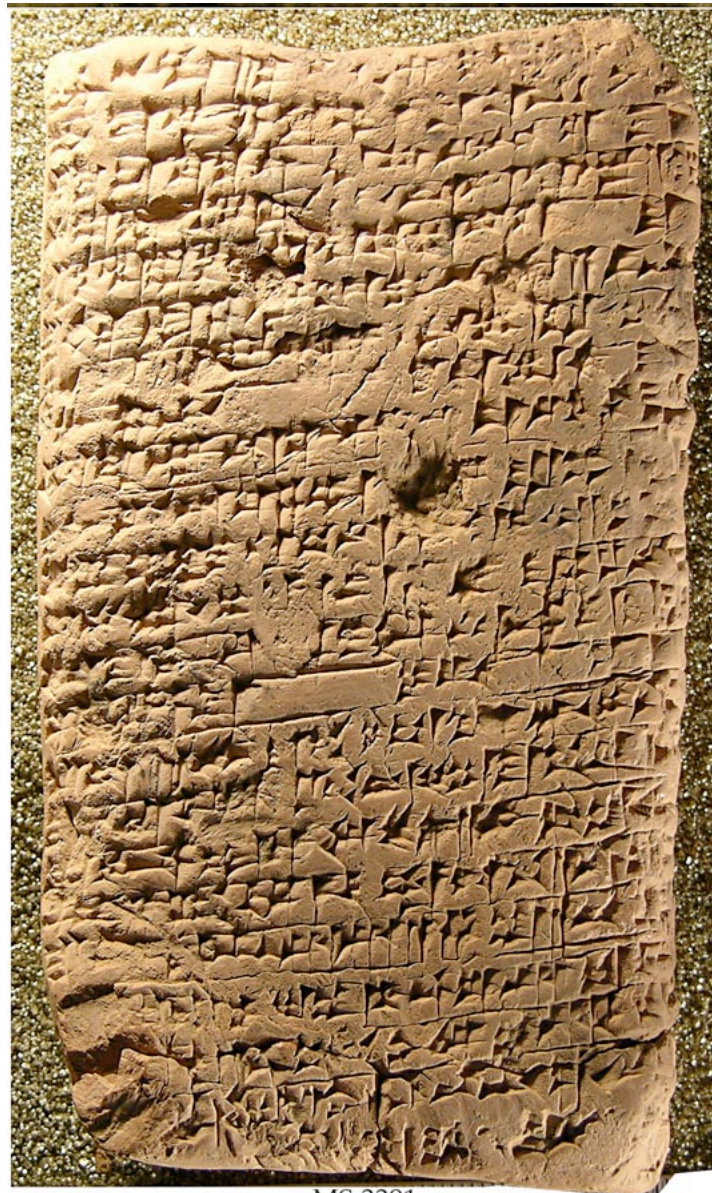


---

# Information retrieval of yore

---

Manning et al. “Information retrieval (IR) is finding material ... of an unstructured nature ... that satisfies an information need from within large collections ...”



MS 3391

Library catalogue. Babylonia, 2000-1600 BC



---

# Information retrieval of yore

---

Manning et al. “Information retrieval (IR) is finding material ... of an unstructured nature ... that satisfies an information need from within large collections ...”



---

# Information retrieval of yore

---

Manning et al. “Information retrieval (IR) is finding material ... of an unstructured nature ... that satisfies an information need from within large collections ...”



---

## Information retrieval today

---

Manning et al. “Information retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers).”

The Google logo, featuring the word "Google" in its characteristic multi-colored font (blue, red, yellow, blue, green, red).The Bing logo, featuring the word "bing" in a blue sans-serif font with a small orange dot above the 'i' and a trademark symbol (TM) to the right.



---

# Information retrieval today

---

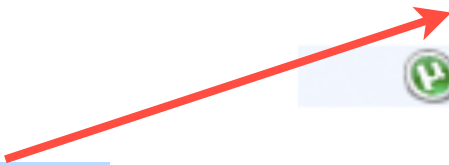
Manning et al. “Information retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers).”





# Information retrieval today

Has this ever happened to you?



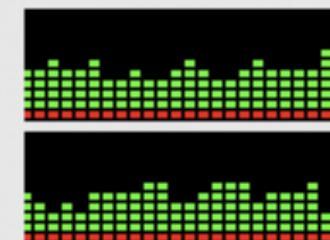
Process Name	User	CPU	# Threads	Real Memory
mdworker	Ryan	68.6	3	3.16 MB
uTorrent	Ryan	4.1	7	23.86 MB
SystemUIServer	Ryan	3.7	6	12.45 MB
Activity Monitor	Ryan	1.5	6	12.32 MB
screencapture	Ryan	0.1	1	2.28 MB
QUICKSILVER	Ryan	0.0	5	14.55 MB
Finder	Ryan	0.0	7	9.81 MB
loginwindow	Ryan	0.0	3	5.86 MB
iStat menus Helper	Ryan	0.0	1	2.55 MB
Dock	Ryan	0.0	2	6.64 MB
launchd	Ryan	0.0	3	528.00 KB

mdworker is part of the indexer for Mac OS X, which is indexing your stuff so that it can be easily searched with Spotlight, the desktop search feature on OS X.

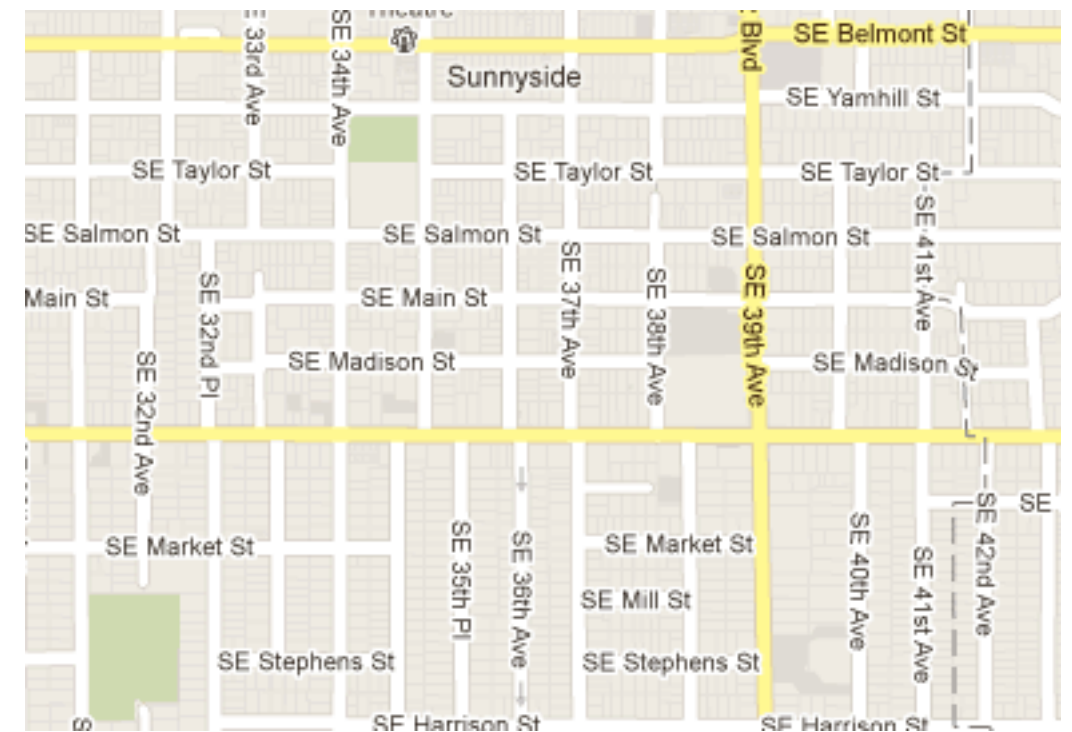
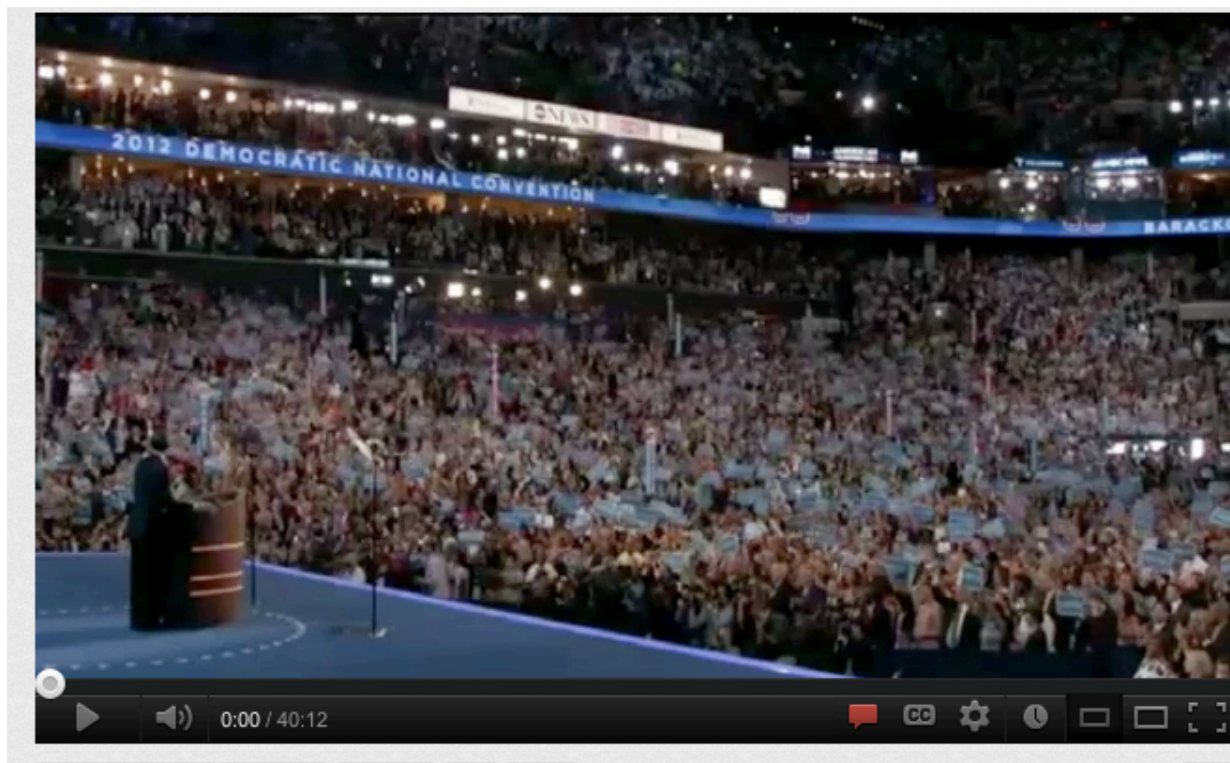
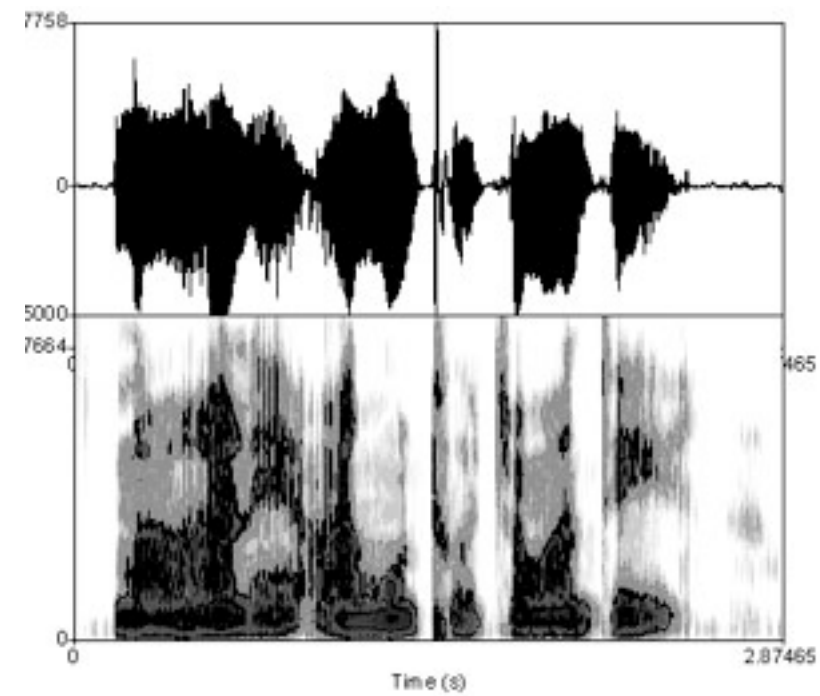
6 User: 35.50  
ystem: 6.00  
6 Nice: 0.00  
% Idle: 58.50



Threads: 190  
Processes: 43



## IR for non-textual media



---

## Beyond textual queries

---





---

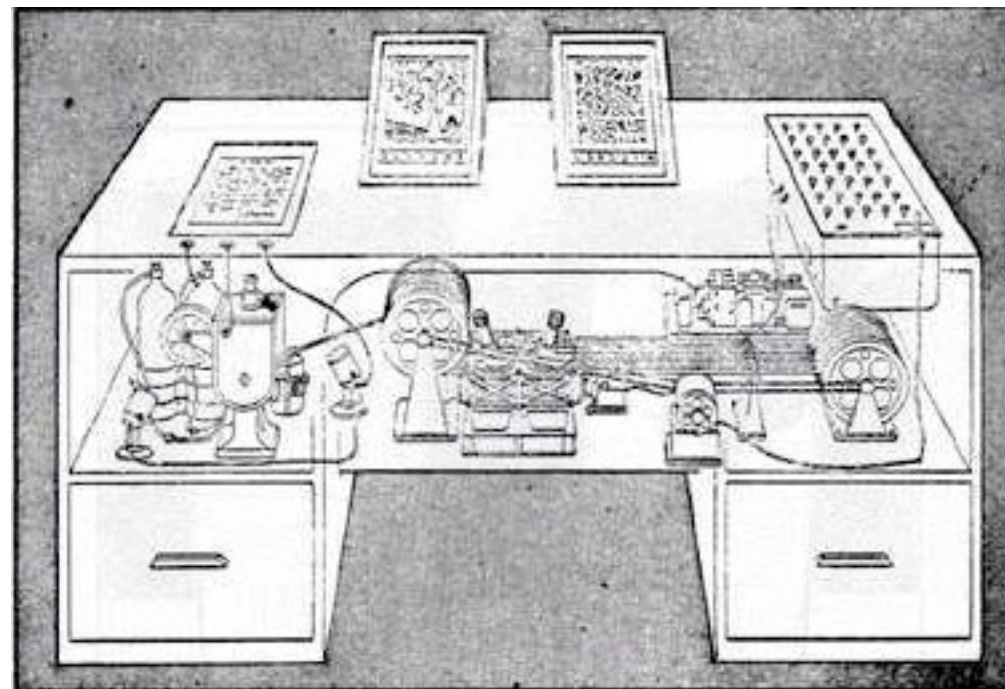
# Highlights from the history of IR

---

+

1945

Memex: “a device in which an individual stores all his books, records, and communications, and which is mechanized so that it may be consulted with exceeding speed and flexibility”



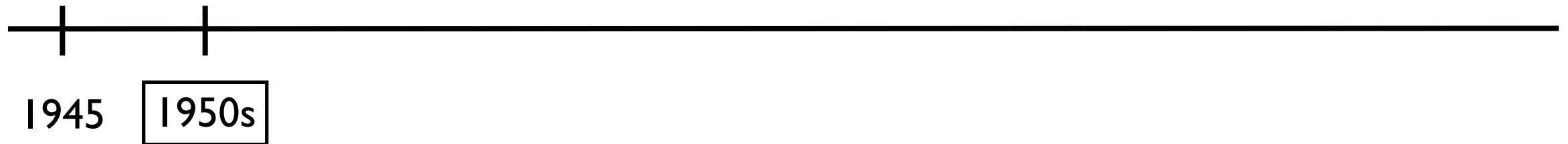
*As we may think*  
Vannevar Bush  
1945



---

## Highlights from the history of IR

---



Alan Kent: literature searching systems, precision/recall evaluation.

Calvin Mooers: allegedly coins the term *information retrieval*.

---

## Highlights from the history of IR

---



Cranfield experiments: computer-based information retrieval evaluation experiments.

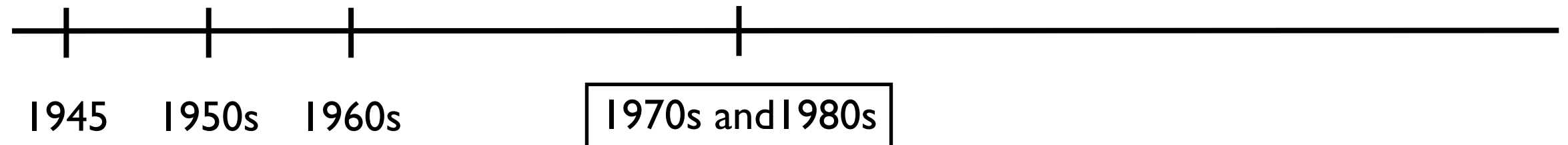
MEDLARS: machine readable database at NLM.

SMART: information retrieval system by Gerard Salton.

---

## Highlights from the history of IR

---



Gerard Salton: indexing, term weighting, vector space model.

SIGIR: special interest research group on IR.

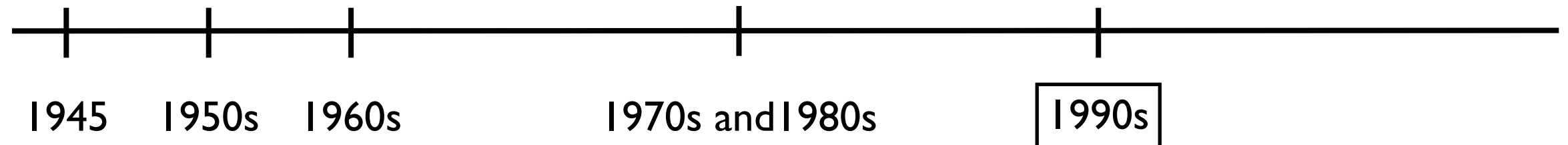
Online IR systems: MEDLINE, AIM-TWX, Dialog.

LexisNexis: database for legal cases and news articles.

---

## Highlights from the history of IR

---



NIST TREC: Text REtrieval Conference

Archie: early search engine for FTP archives.

Altavista, Yahoo, Excite: first WWW search engines.

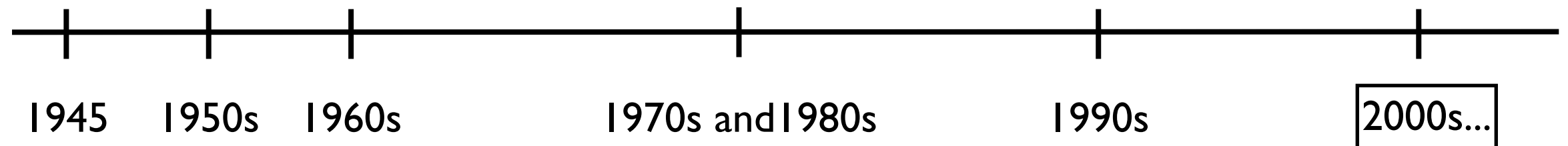
PageRank: Google's web page ranking algorithm



---

## Highlights from the history of IR

---



Google: web search and specialized search (scholar, news).

Multimedia IR: images, video, audio.

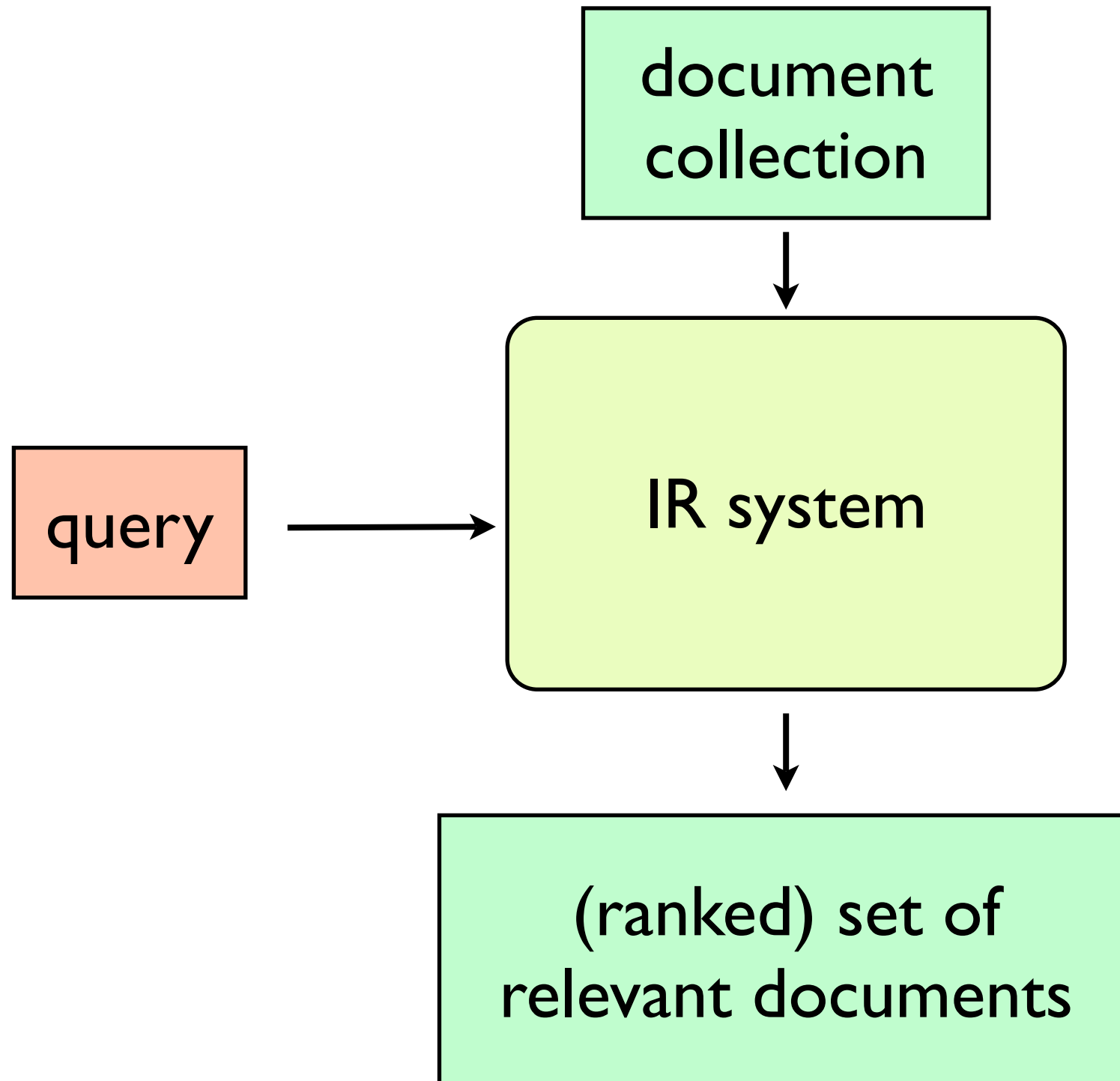
Multilingual IR: CLEF initiative.

Recommendation systems: Amazon, Netflix challenge.

---

# IR basics

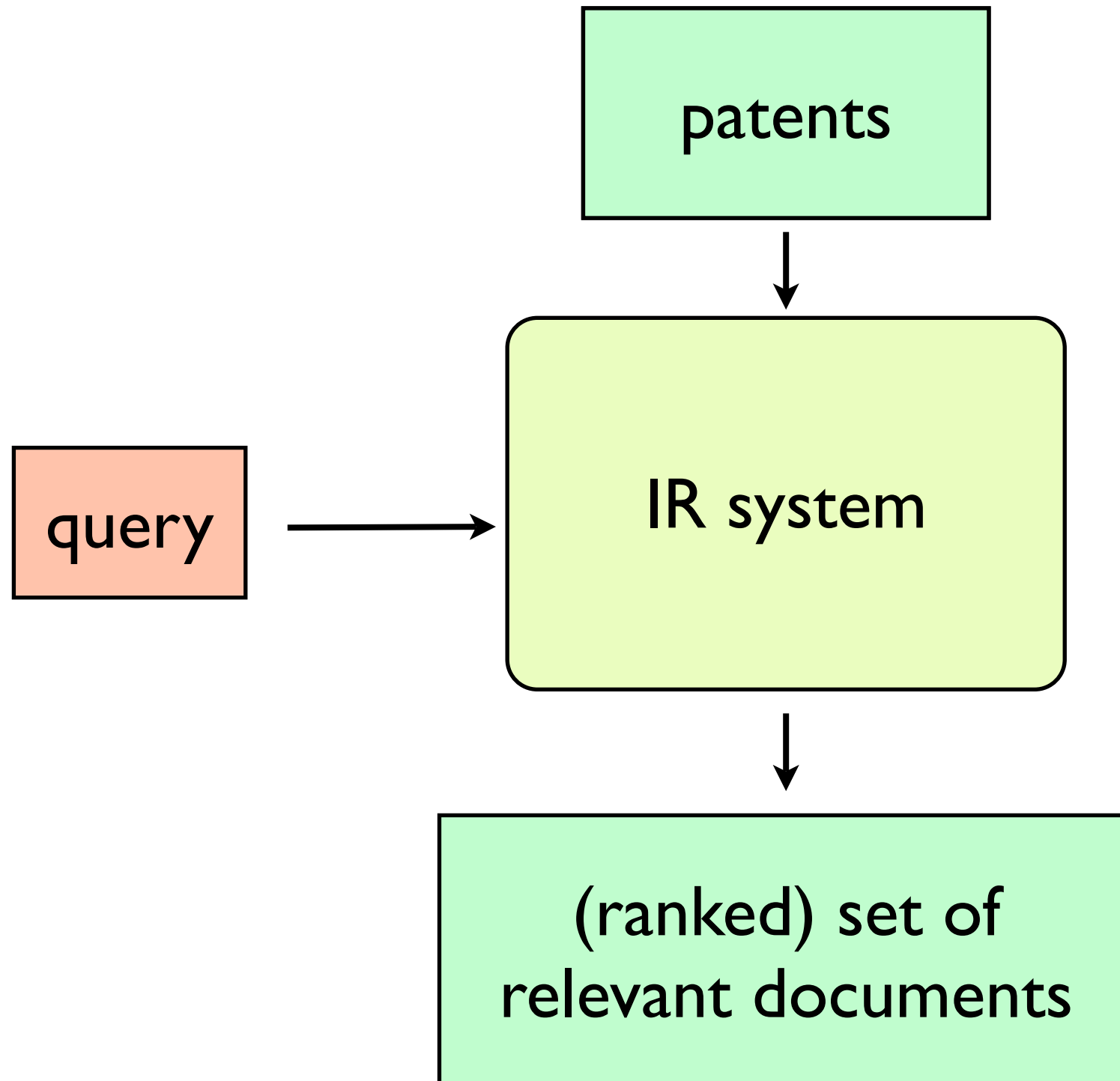
---



---

# IR basics

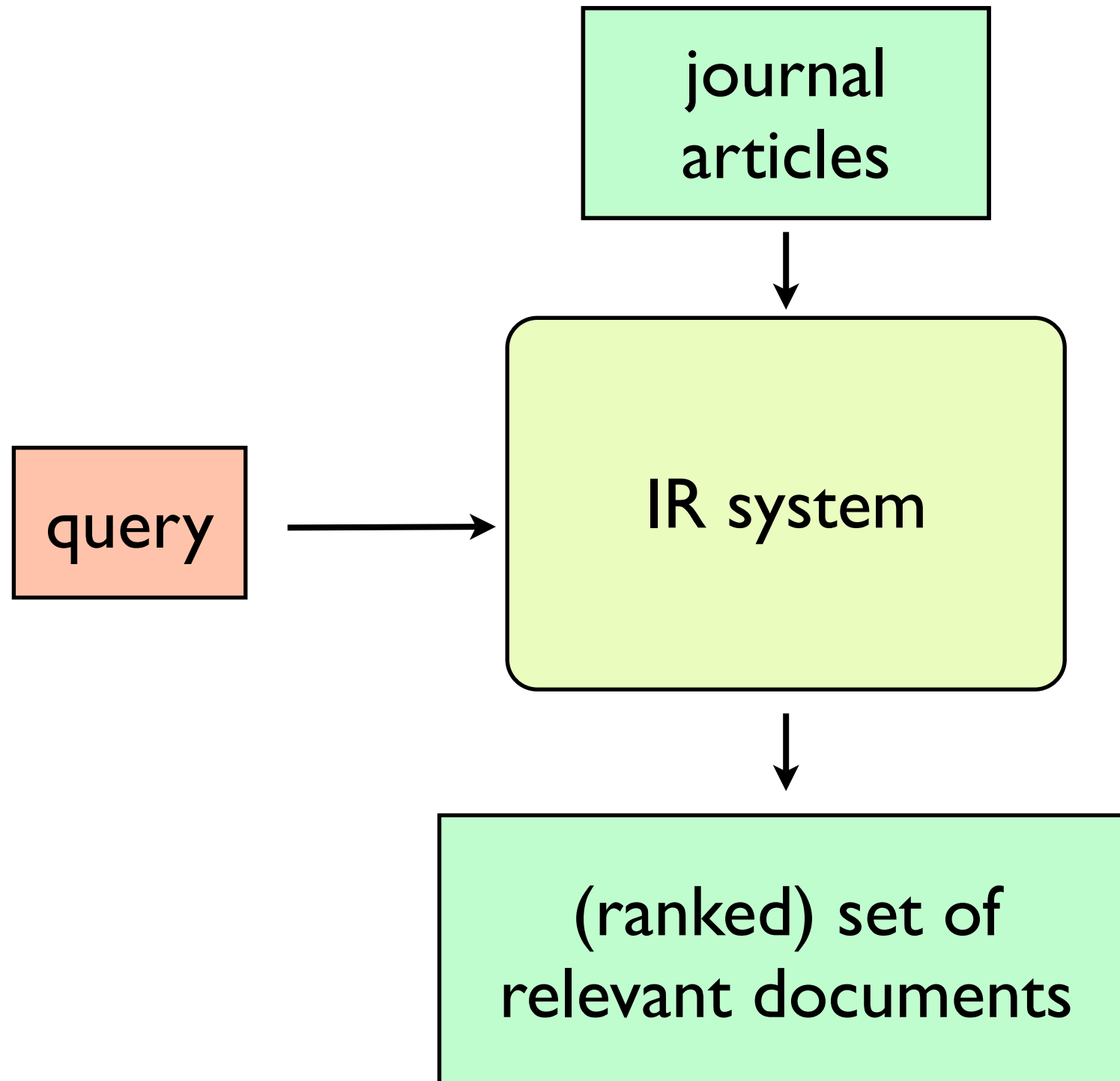
---



---

# IR basics

---

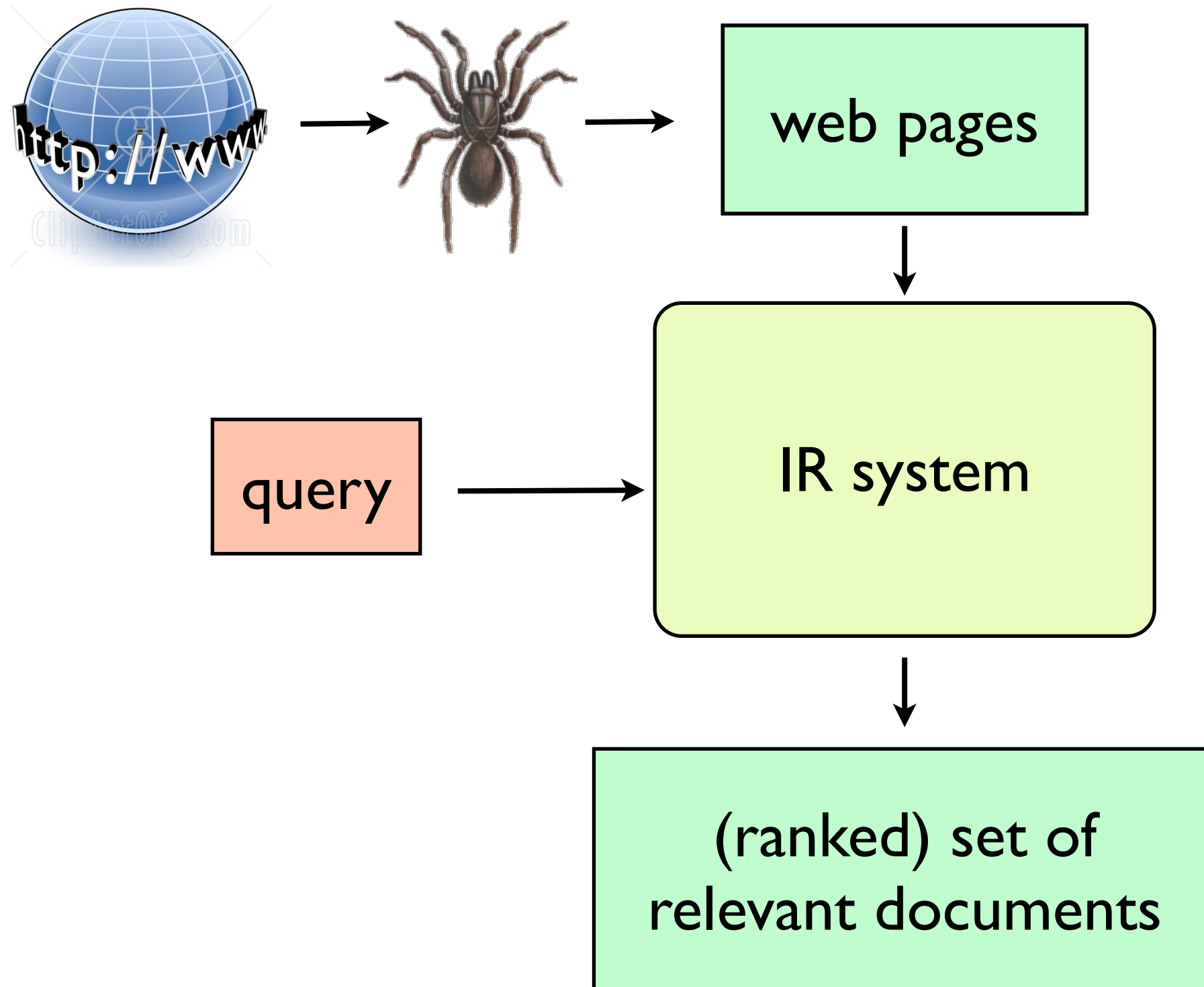




---

# IR basics

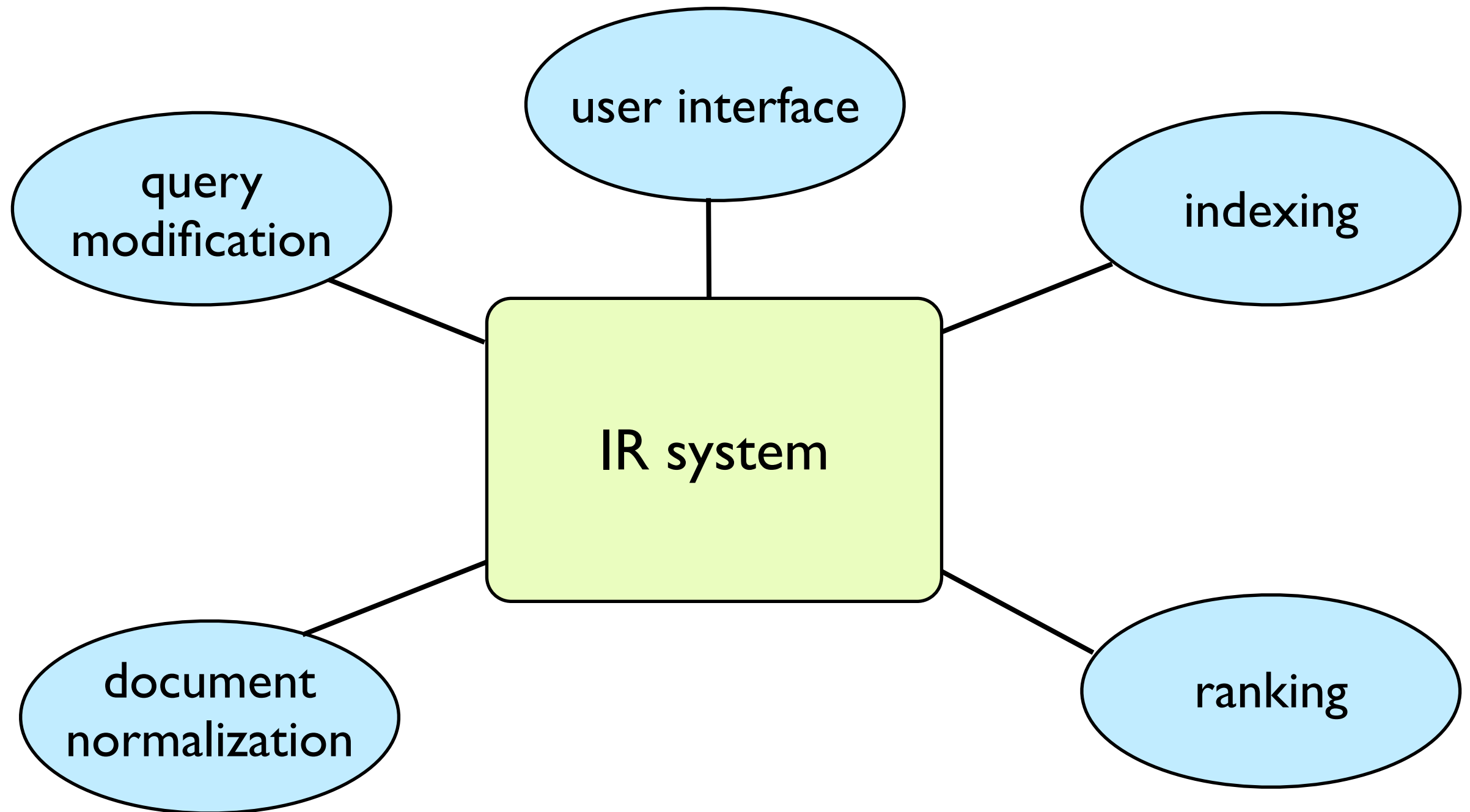
---



---

# IR system components

---



---

## Important considerations in IR: Relevance

---

- Are the retrieved documents...
  - about the target subject?
  - timely and up-to-date?
  - from a trusted source?
  - satisfying the user's needs and goals?
- How do we rank documents in terms of these features?
- How do we decide which of these are more important and which are less important?
- More on this in a lecture soon.

---

## Important considerations in IR: Evaluation

---

- Precision: Proportion of retrieved documents that are relevant.
- Recall: Proportion of relevant documents that are retrieved.
- What is the best balance between the two?
  - Easy to get perfect recall: just retrieve everything.
  - Easy to get good precision: return only the most relevant.
- Best balance may depend on the context:
  - patent search: high recall
  - ad hoc web search: high precision
- Evaluation will be covered in detail in a later lecture.

---

# Information needs

---

- Fact finding

The screenshot shows a Google search interface. At the top, the Google logo is on the left, and the search query 'inches in a mile' is in the search bar. Below the search bar, the word 'Search' is on the left, and 'About 10,800,000 results (0.30 seconds)' is on the right. A horizontal line separates the search results from the conversion tool. On the left side of the conversion tool, there is a vertical list of links: 'Web' (highlighted with a red bar), 'Images', 'Maps', 'Videos', and 'News'. The conversion tool itself consists of a dropdown menu labeled 'Length' with a double-headed arrow. Below this, there are two input fields. The left field contains the number '1' and a dropdown menu labeled 'Mile' with a double-headed arrow. The right field contains the number '63360' and a dropdown menu labeled 'Inch' with a double-headed arrow. An equals sign '=' is positioned between the two input fields.

Google inches in a mile

Search About 10,800,000 results (0.30 seconds)

Web  
Images  
Maps  
Videos  
News

Length

1 = 63360

Mile Inch



---

# Information needs

---

- Fact finding



The screenshot shows a Google search interface. The search bar contains the text "weather in portland". Below the search bar, the word "Search" is displayed in red, followed by the text "About 84,900,000 results (0.32 seconds)". On the left side, there is a vertical list of search filters: "Web", "Images", "Maps", "Videos", and "News". The "Web" filter is selected. The main content area displays the "Weather for Portland, OR" widget. The current temperature is 60°F | °C, with a "Partly Cloudy" condition. Wind is "N at 0 mph" and humidity is "69%". A weather icon shows a sun partially obscured by a cloud. To the right, a 4-day forecast is shown for Saturday (Sat), Sunday (Sun), Monday (Mon), and Tuesday (Tue). Each day has a weather icon and a temperature range. Below the forecast, a link for "Detailed forecast" points to "The Weather Channel - Weather Underground - AccuWeather".

Day	Weather	Temperature Range
Sat	Partly Cloudy	77° 50°
Sun	Sunny	61° 52°
Mon	Partly Cloudy	64° 50°
Tue	Partly Cloudy	57° 46°

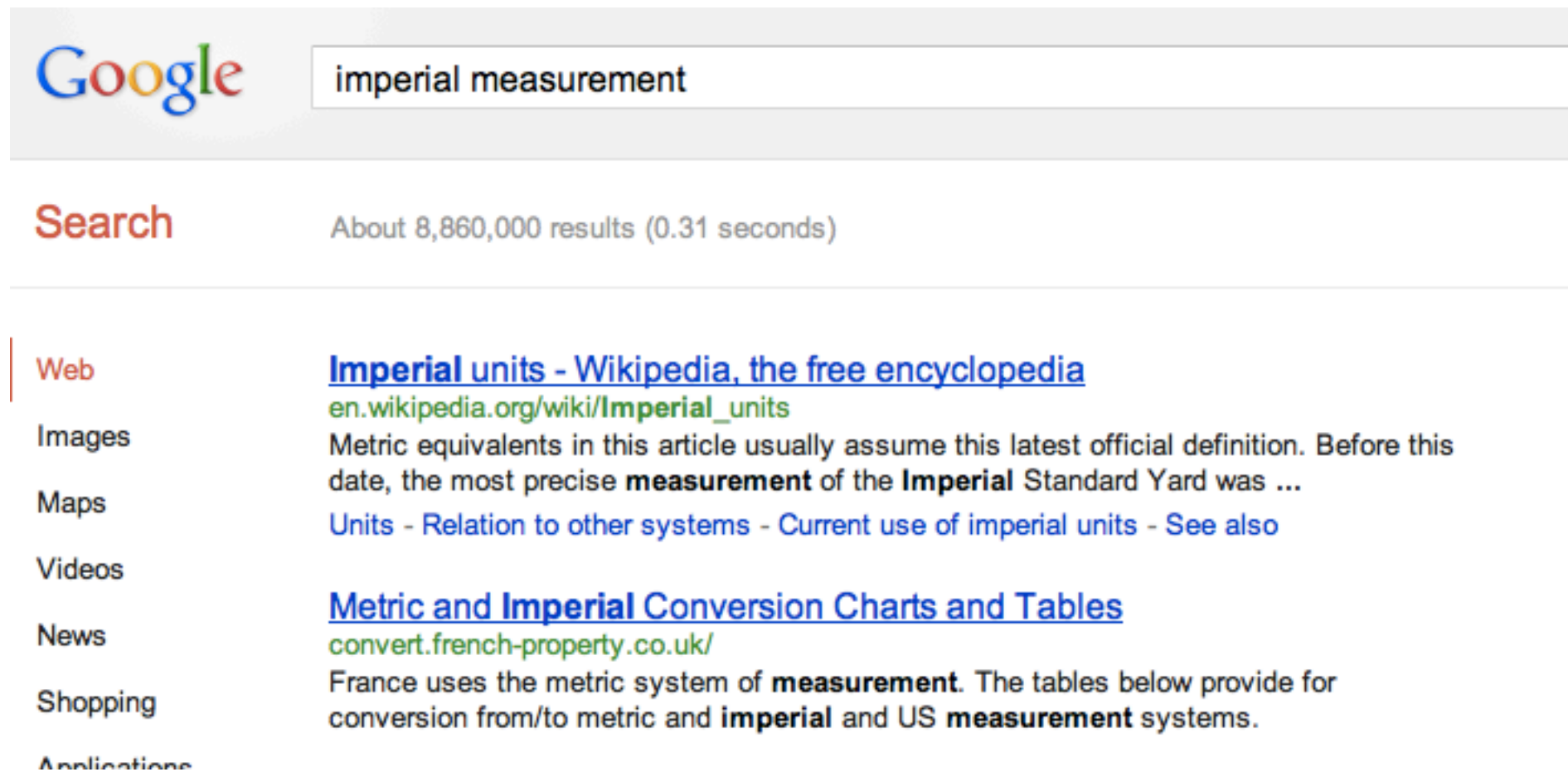
Detailed forecast: [The Weather Channel](#) - [Weather Underground](#) - [AccuWeather](#)

---

# Information needs

---

- General knowledge



The screenshot shows a Google search interface. The search bar contains the text "imperial measurement". Below the search bar, the word "Search" is displayed in red, followed by the text "About 8,860,000 results (0.31 seconds)". On the left side, there is a vertical menu with links to "Web", "Images", "Maps", "Videos", "News", "Shopping", and "Applications". The "Web" link is selected. The search results are displayed on the right. The first result is titled "Imperial units - Wikipedia, the free encyclopedia" and includes the URL "en.wikipedia.org/wiki/Imperial\_units". The snippet for this result reads: "Metric equivalents in this article usually assume this latest official definition. Before this date, the most precise **measurement** of the **Imperial** Standard Yard was ...". Below this, there are links: "Units - Relation to other systems - Current use of imperial units - See also". The second result is titled "Metric and Imperial Conversion Charts and Tables" and includes the URL "convert.french-property.co.uk/". The snippet for this result reads: "France uses the metric system of **measurement**. The tables below provide for conversion from/to metric and **imperial** and US **measurement** systems."

Google

imperial measurement

Search About 8,860,000 results (0.31 seconds)

Web

[Imperial units - Wikipedia, the free encyclopedia](#)  
[en.wikipedia.org/wiki/Imperial\\_units](http://en.wikipedia.org/wiki/Imperial_units)  
Metric equivalents in this article usually assume this latest official definition. Before this date, the most precise **measurement** of the **Imperial** Standard Yard was ...  
[Units - Relation to other systems - Current use of imperial units - See also](#)

Images

Maps

Videos

News

[Metric and Imperial Conversion Charts and Tables](#)  
[convert.french-property.co.uk/](http://convert.french-property.co.uk/)  
France uses the metric system of **measurement**. The tables below provide for conversion from/to metric and **imperial** and US **measurement** systems.

Shopping

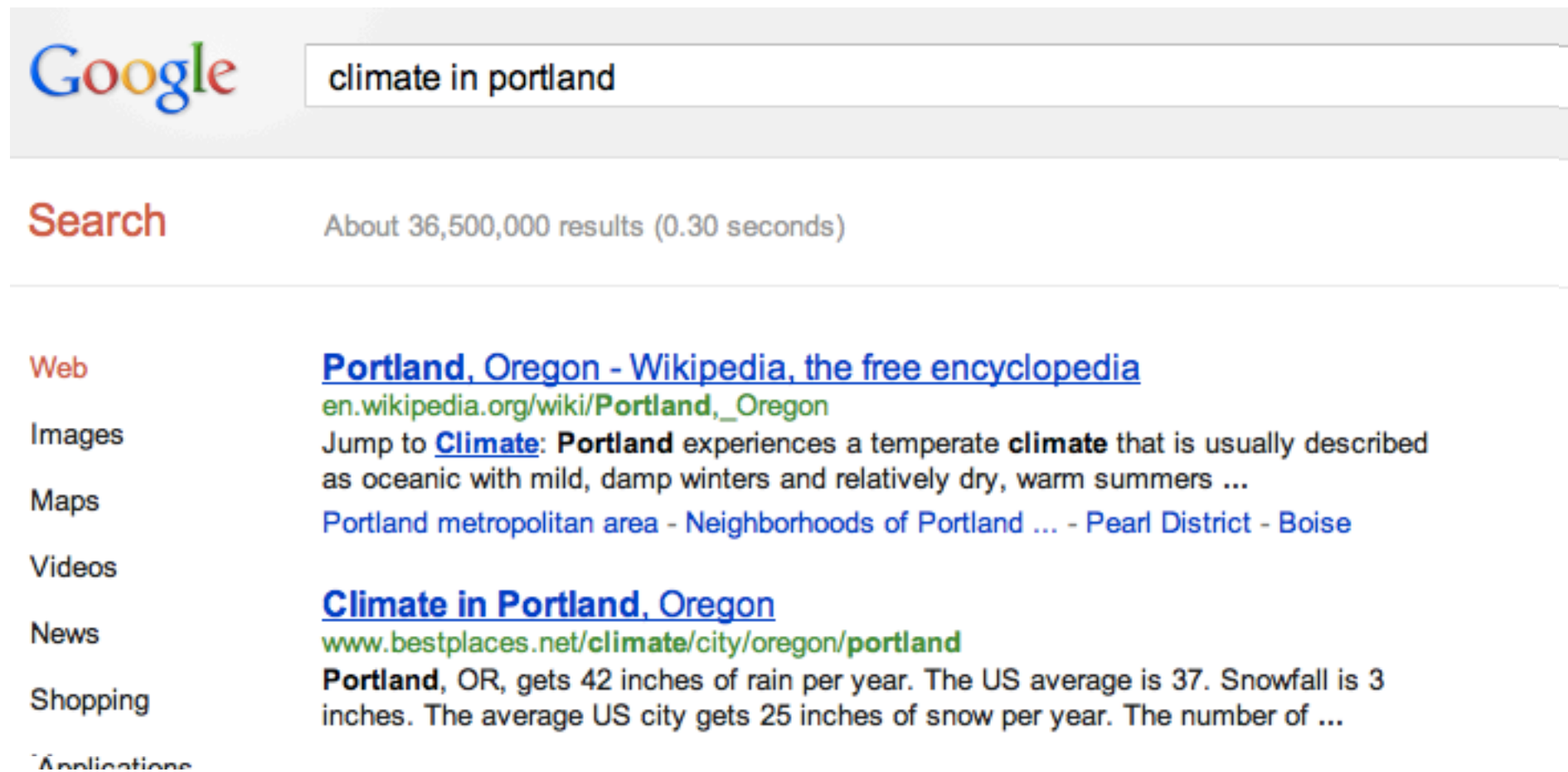
Applications

---

# Information needs

---

- General knowledge



The screenshot shows a Google search interface. At the top left is the Google logo. To its right is a search bar containing the text "climate in portland". Below the search bar, the word "Search" is displayed in red, followed by the text "About 36,500,000 results (0.30 seconds)". On the left side, there is a vertical list of search categories: "Web", "Images", "Maps", "Videos", "News", "Shopping", and "Applications". The "Web" category is selected, and the search results are displayed to its right. The first result is titled "Portland, Oregon - Wikipedia, the free encyclopedia" and includes the URL "en.wikipedia.org/wiki/Portland,\_Oregon". The snippet for this result reads: "Jump to **Climate**: **Portland** experiences a temperate **climate** that is usually described as oceanic with mild, damp winters and relatively dry, warm summers ...". Below this is a link to "Portland metropolitan area - Neighborhoods of Portland ... - Pearl District - Boise". The second result is titled "Climate in Portland, Oregon" and includes the URL "www.bestplaces.net/climate/city/oregon/portland". The snippet for this result reads: "Portland, OR, gets 42 inches of rain per year. The US average is 37. Snowfall is 3 inches. The average US city gets 25 inches of snow per year. The number of ...".

Google

climate in portland

Search About 36,500,000 results (0.30 seconds)

Web [Portland, Oregon - Wikipedia, the free encyclopedia](#)  
en.wikipedia.org/wiki/Portland,\_Oregon  
Jump to **Climate**: **Portland** experiences a temperate **climate** that is usually described as oceanic with mild, damp winters and relatively dry, warm summers ...  
[Portland metropolitan area - Neighborhoods of Portland ... - Pearl District - Boise](#)

Images

Maps

Videos

News [Climate in Portland, Oregon](#)  
www.bestplaces.net/climate/city/oregon/portland  
**Portland**, OR, gets 42 inches of rain per year. The US average is 37. Snowfall is 3 inches. The average US city gets 25 inches of snow per year. The number of ...

Shopping

Applications

---

## **A few models of information seeking**

---

- Standard model
- Dynamic (berry-picking) model
- Search as a strategic process
- Sensemaking



---

## Standard model of information seeking

---

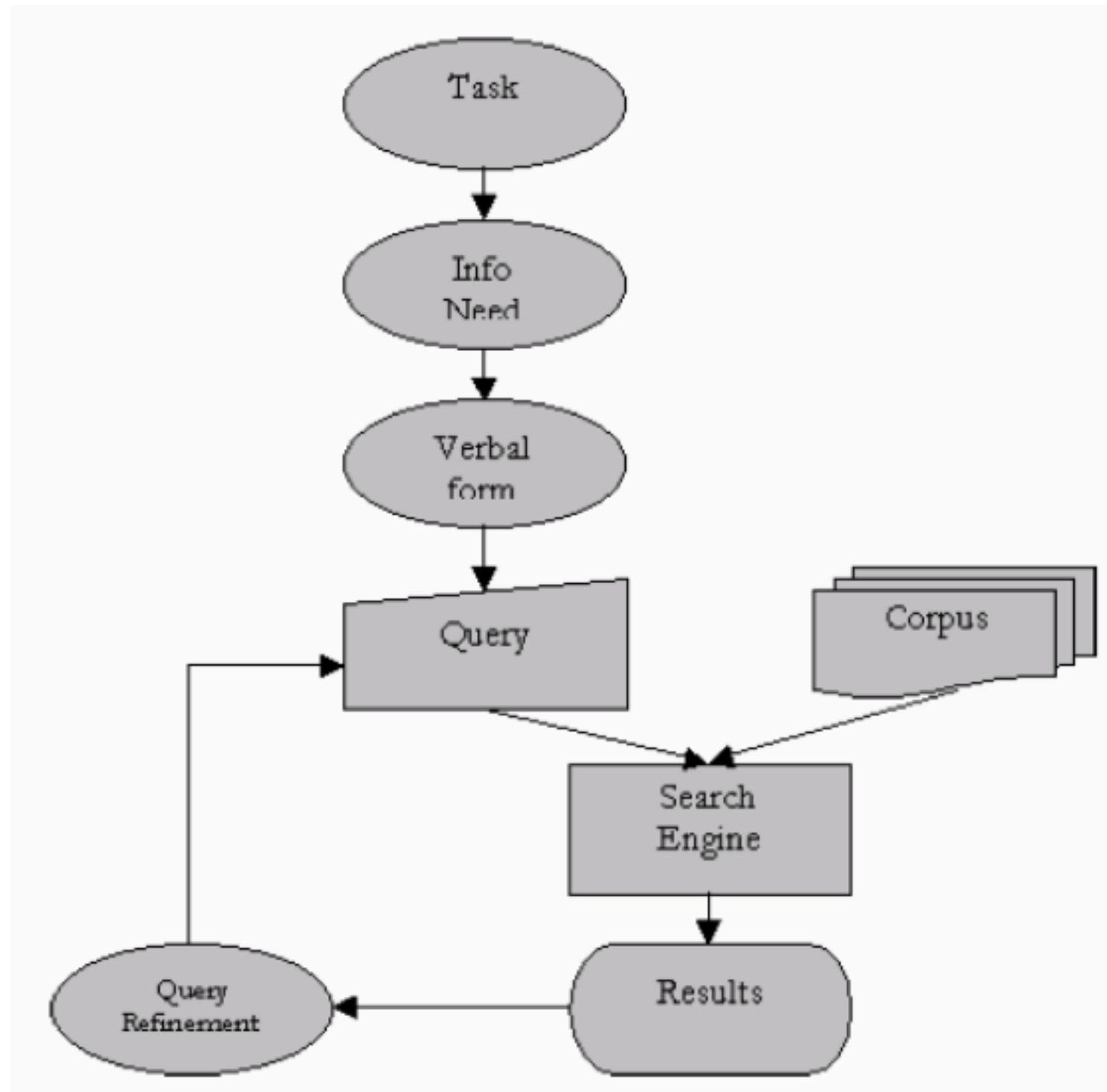
- Recognizing a need for information,
- Accepting the challenge to take action to fulfill the need,
- Formulating the problem,
- Expressing the information need in a search system,
- Examination of the results,
- Reformulation of the problem and its expression,
- Admitting you are powerless over the internet, and
- Use of the results.

Slightly modified summary of information-seeking process outlined in Marchionini and White (2008), described in M. Hearst (2009) *Search User Interfaces*.

---

# Standard model of information seeking

---



---

## **“Berry-picking” model of info seeing**

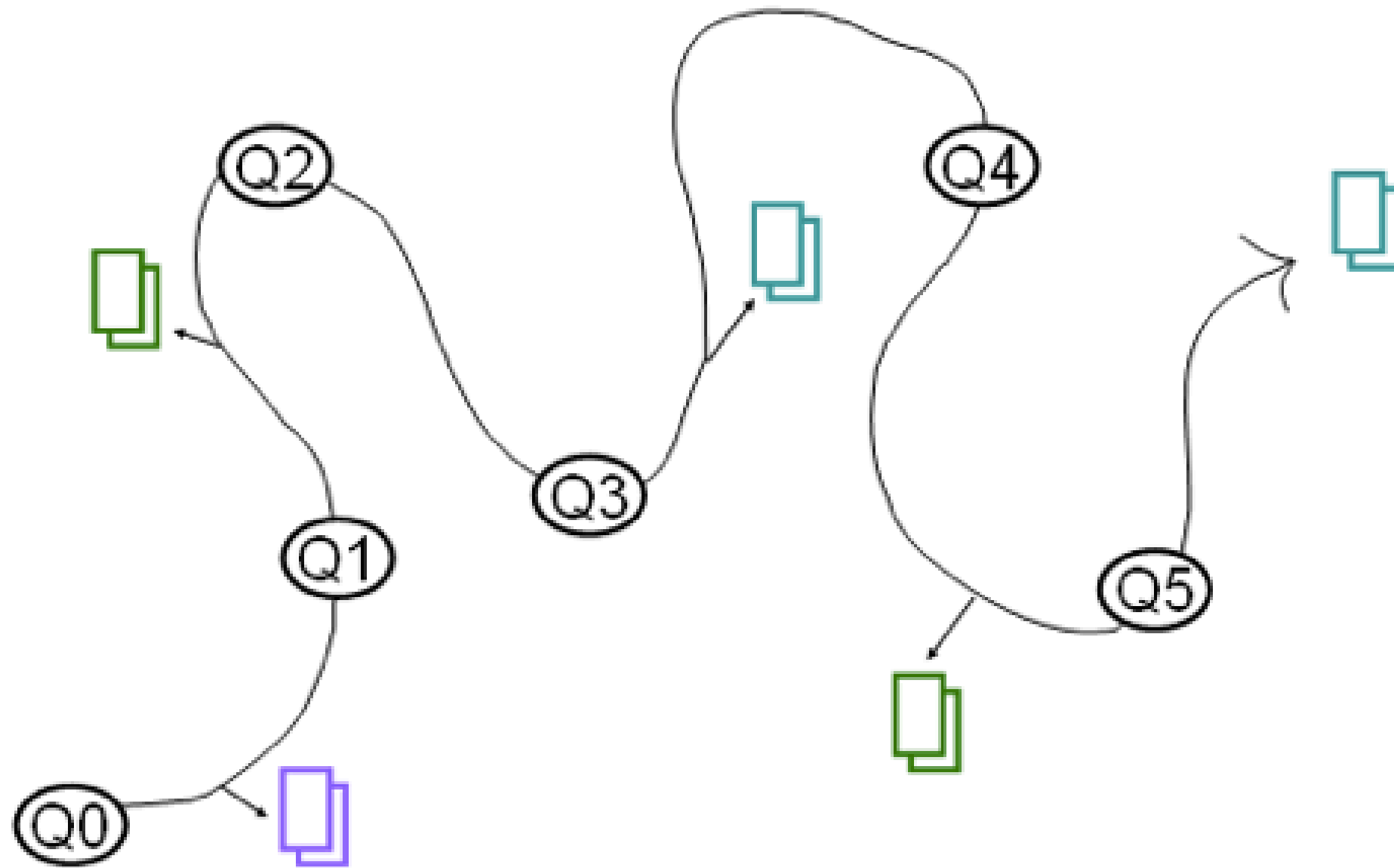
---

- General model assumes a user’s information is static.
- Turns out that information needs change as results are retrieved.
- As an information seeker gets results, he learns more about what he’s looking for.
  - Reformulated queries may veer in a different direction.
  - One search goal may take precedence over another.
  - Results may lead to entirely new search needs.
- For some information seekers, the value of the search is in the total acquisition of knowledge rather than in the final set of search results.

---

## “Berry-picking” model of info seeing

---



---

## Information seeking as a strategic process

---

- Querying/searching vs. browsing/navigating.
  - search queries → ad hoc collections of documents
  - browsing → following links to predefined collections
- Less work required to browse than formulate a query, but...
  - if too many links, takes too long to find what you want
  - desired links may not be available or labeled appropriately
- Some strategies incorporate both querying and navigating.
  - To find web-page for this class, you might Google Brian Roark, and then follow the link to the class page from Brian's personal web page.



---

# Sensemaking

---

- “an iterative process of formulating a conceptual representation from of a large volume of information”
  - information retrieval via searching, browsing
  - analysis and synthesis of results (analyzing and saving data, organizing documents)
- Process involved in scientific research, legal discovery.
- Interestingly, persistence on the part of the information seeker (i.e., time spent searching and browsing) results in more relevant documents retrieved -- not the kinds of queries issued.

---

## Next lecture

---

- Information retrieval basics:
  - Term-document matrix
  - Inverted indices
  - Boolean retrieval, index intersection
- Additional topics on terms and postings
  - Faster intersection of posting lists
  - Positional indices
  - Tokenization and normalization