



# Introduction: A Brief History of IR


---

Jay Aslam

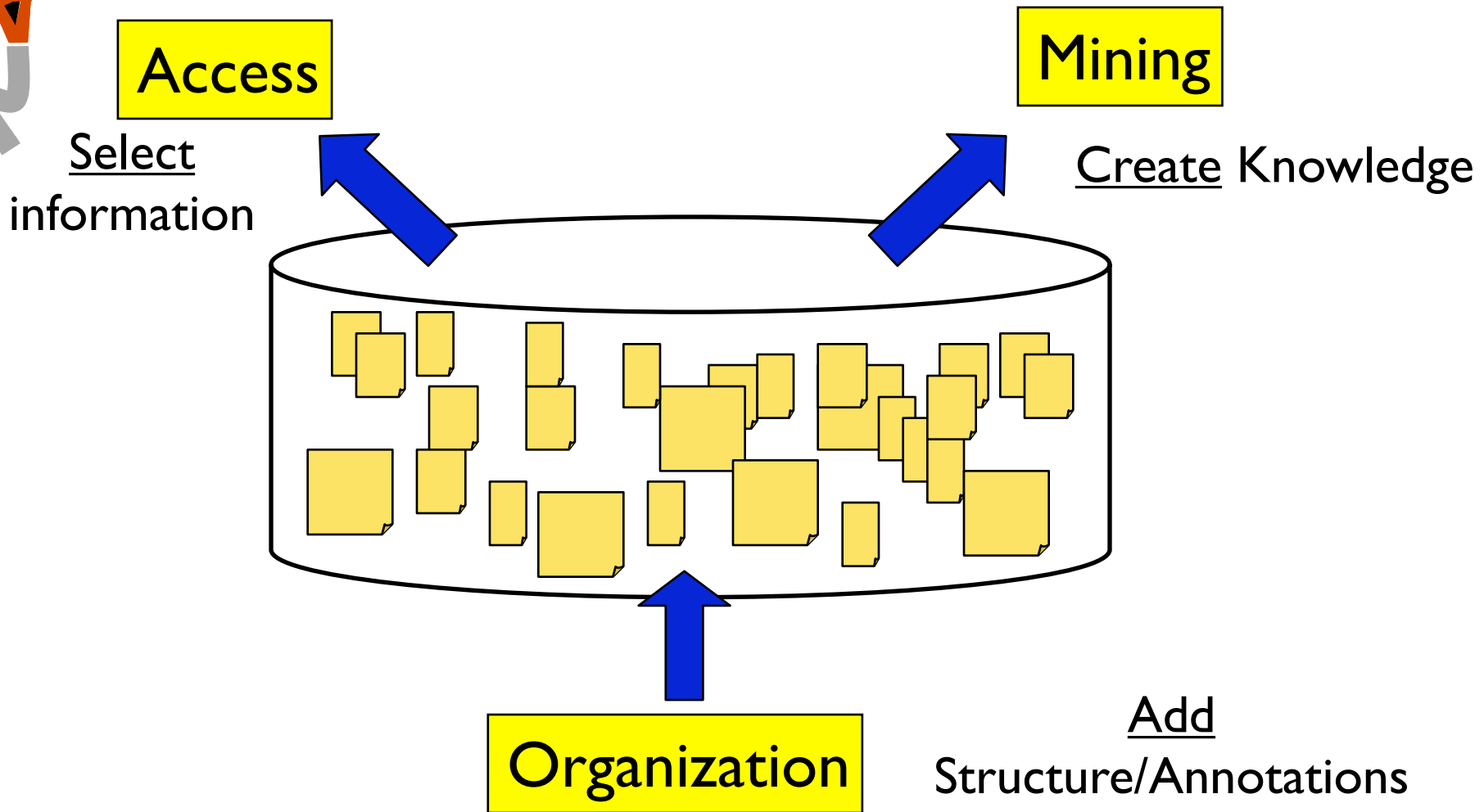
College of Computer and Information Science  
Northeastern University

# Overview

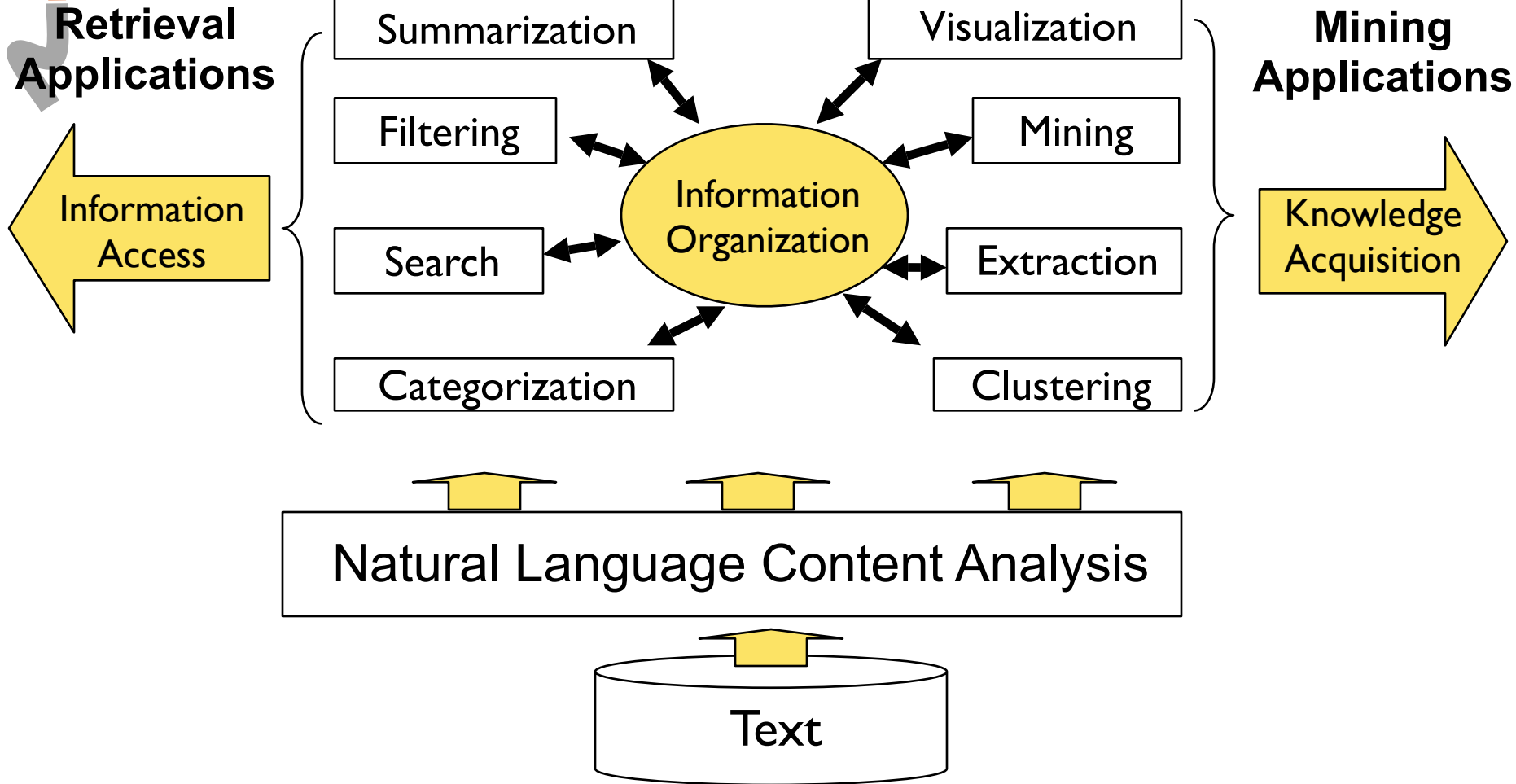
---

- 
- What is information retrieval?
  - How do search engines work?
  - The internet & web search
  - Adversarial IR

# Text management applications

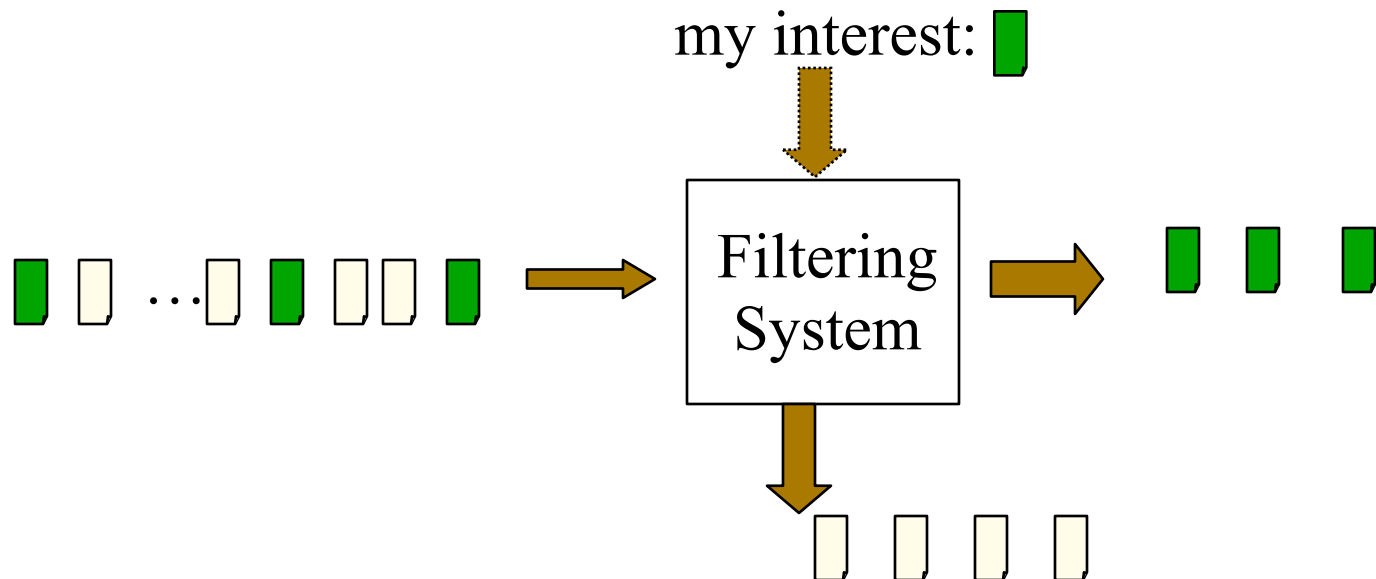


# Text management applications



# Information filtering

- Stable & long term interest, dynamic info source
- System must make a delivery decision immediately as a document “arrives”



# Collaborative filtering




amazon.com Your Store Books See All 31 Product Categories Your Account | Cart | Wish List | Help

Search | Browse Subjects | Bestsellers | The New York Times® Best Sellers | Magazines | Corporate Accounts | E-books & Docs | Bargain Books | Used Books

Search Books  GO Advanced Search




Your order qualifies for free shipping! (some restrictions apply)  
Make sure to select **FREE Super Saver Shipping** as your shipping speed at checkout.

You could save \$30 today with the Amazon Visa® Card:




 Your current subtotal: \$109.76  
**Amazon Visa discount: - \$30.00**  
Your new subtotal: **\$79.76** [Find out how](#)

Save \$30 off your first purchase, earn 3% rewards, get a 0% APR\*, and pay no annual fee.

Customers who bought *Managing Gigabytes* also bought:

 <p><a href="#">Mining the Web</a> by Soumen Chakrabarti Price: <b>\$57.95</b> Used &amp; new from \$41.42 <a href="#">Add to cart</a> <a href="#">Explore similar items</a></p>	 <p><a href="#">Foundations of Statistical Natural Language Processing</a> by Christopher D. Manning, Hinrich Schtze Price: <b>\$67.32</b> Used &amp; new from \$42.98 <a href="#">Add to cart</a></p>	 <p><a href="#">Natural Language Processing for Online Applications</a> by Peter Jackson, Isabelle Moulinier Price: <b>\$126.00</b> Used &amp; new from \$142.44 <a href="#">Add to cart</a></p>
---	---	---

Customers who shopped for *Managing Gigabytes* also shopped for:

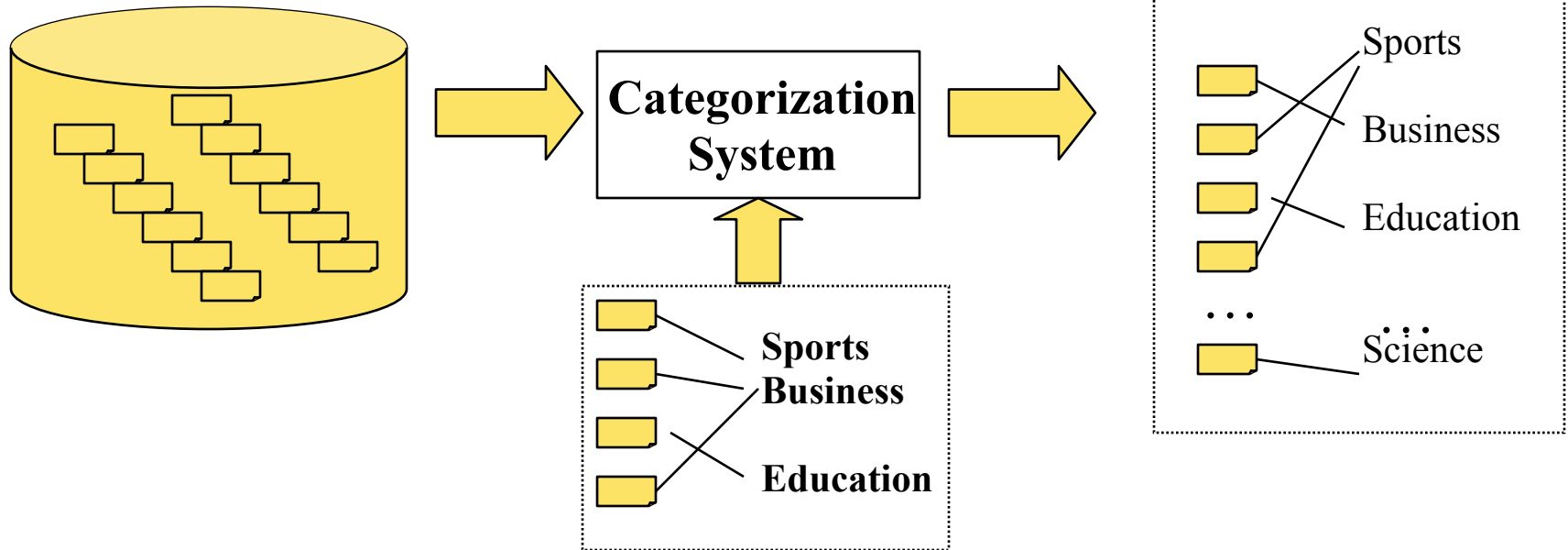
 <p><a href="#">Information Retrieval</a> by William B. Frakes, Ricardo Baeza-Yates Price: <b>\$69.67</b> Used &amp; new from \$37.03 <a href="#">Add to cart</a> <a href="#">Explore similar items</a></p>	 <p><a href="#">Understanding Search Engines</a> by Michael W. Berry, Murray Browne Price: <b>\$41.50</b> Used &amp; new from \$28.50 <a href="#">Add to cart</a></p>	 <p><a href="#">Survey of Text Mining</a> by Michael W. Berry (Editor) Price: <b>\$58.93</b> Used &amp; new from \$54.59 <a href="#">Add to cart</a></p>
--	--	---

Want free shipping? You're almost there! Add a recommended item and qualify now. [Some restrictions apply.](#)

 <p><a href="#">Natural Language Processing for Online Applications</a> by Peter Jackson, Isabelle Moulinier Price: <b>\$39.95</b> Used &amp; new from \$44.89</p>	 <p><a href="#">Lucene in Action (In Action series)</a> by Erik Hatcher, Otis Gospodnetic Price: <b>\$29.67</b> Used &amp; new from \$28.00</p>	 <p><a href="#">Speech and Language Processing</a> by Daniel Jurafsky, James H. Martin Price: <b>\$83.28</b> Used &amp; new from \$63.53</p>
---	--	---

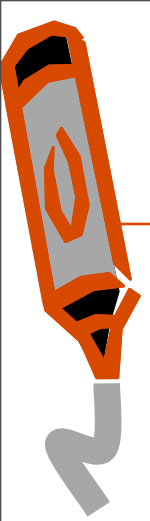
# Categorization

- Pre-given categories and labeled document examples (Categories may form hierarchy)
- Classify new documents
- A standard supervised learning problem



# Clustering

---

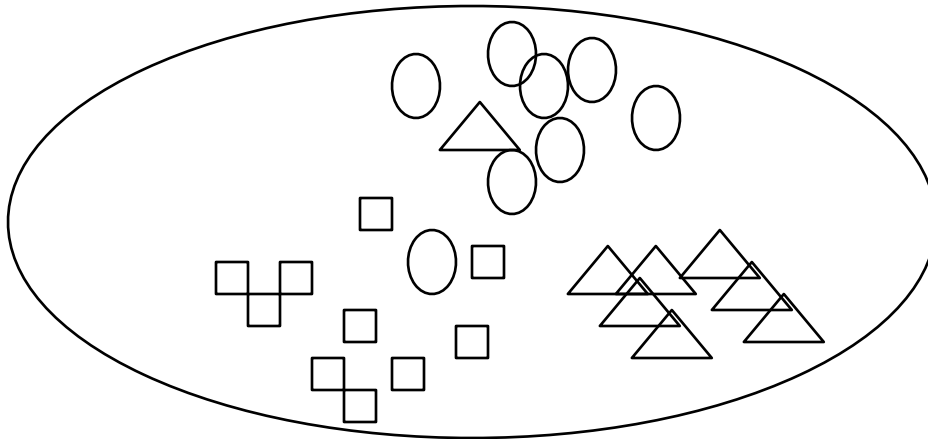


- Discover “natural structure”
- Group similar objects together
- Object can be document, term, passages



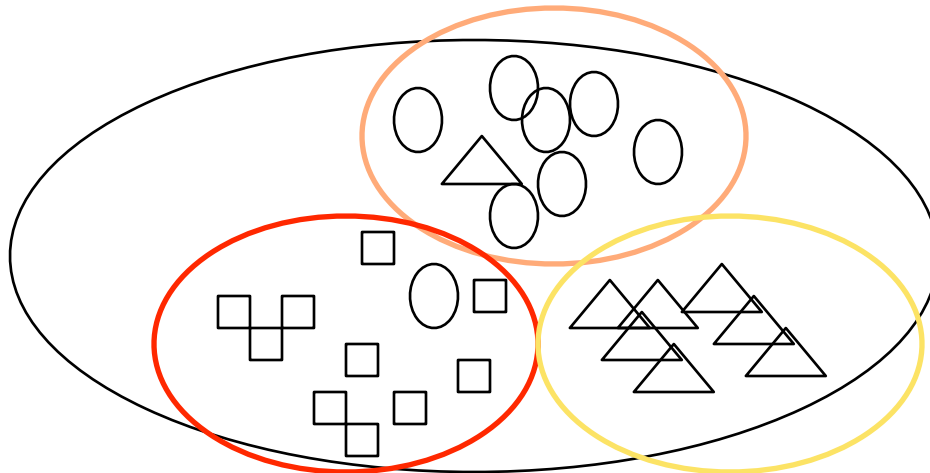
# Clustering

- Discover “natural structure”
- Group similar objects together
- Object can be document, term, passages

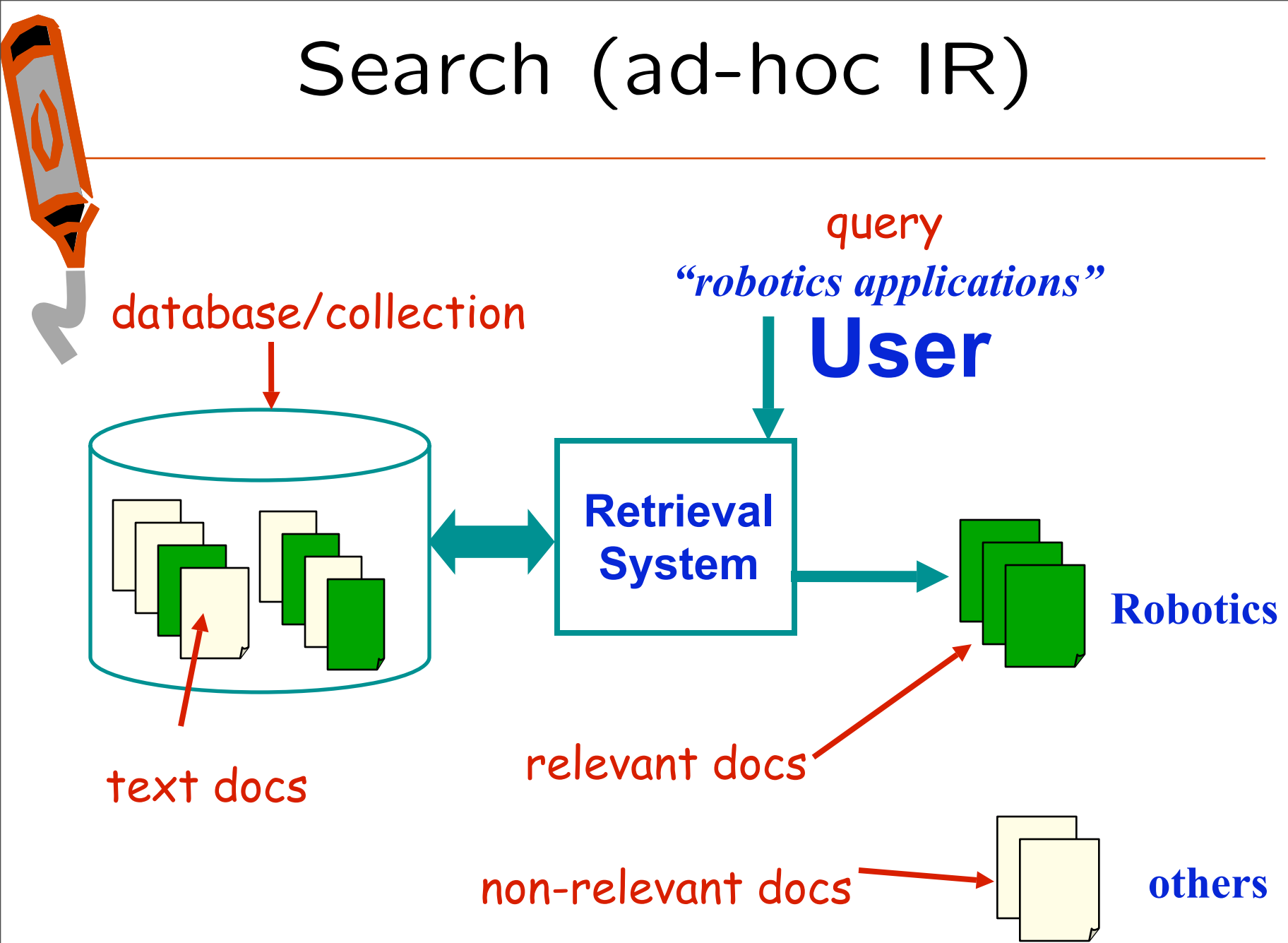


# Clustering

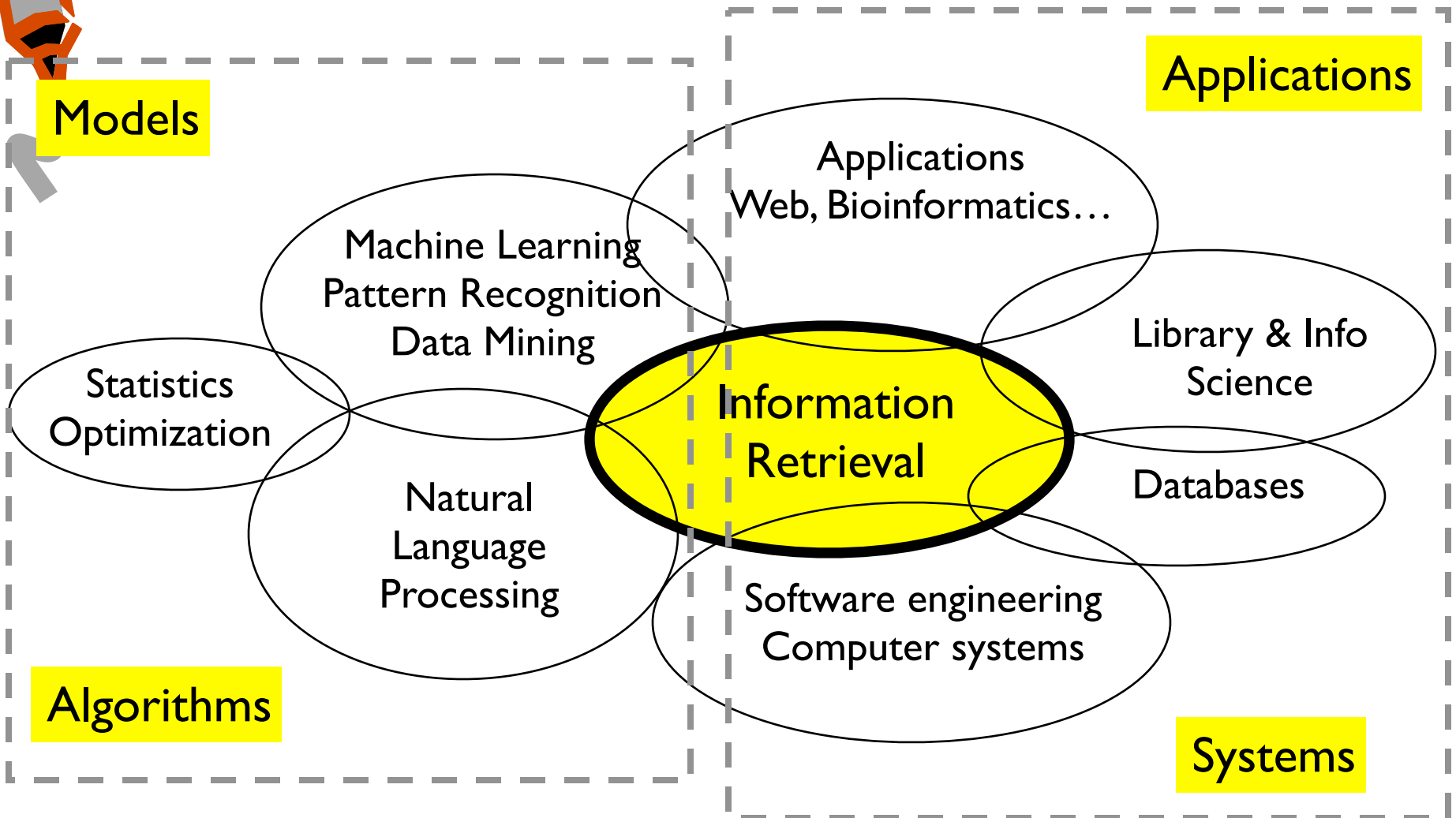
- Discover “natural structure”
- Group similar objects together
- Object can be document, term, passages



# Search (ad-hoc IR)




# Related areas




# Overview

---

- 
- What is information retrieval ?
  - How do search engines work ?
  - The internet & Web Search
  - Adversarial IR

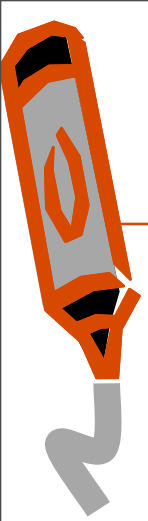
# Basic Idea

---

- 
- Much of IR depends upon idea that **similar vocabulary → similar “meaning”**
  - Usually look for documents matching query words
  - “Similar” can be measured in many ways...

# Bag of words

---



- An effective and popular approach
- Compares words without regard to order
- Consider reordering words in a headline:
  - **Random:** beating takes points falling another Dow 355
  - **Alphabetical:** 355 another beating Dow falling points
  - **“Interesting”:** Dow points beating falling 355 another
  - **Actual: Dow takes another beating, falling 355 points**

# What is this about?

---

16 × said

14 × McDonalds

12 × fat

11 × fries

8 × new

6 × company french nutrition

5 × food oil percent reduce taste Tuesday

4 × amount change health Henstenburg make obesity

3 × acids consumer fatty polyunsaturated US

2 × amounts artery Beemer cholesterol clogging director

down eat estimates expert fast formula impact initiative

moderate plans restaurant saturated trans win

1 × ...

added addition adults advocate affect afternoon age

Americans Asia battling beef bet brand Britt Brook Browns

calorie center chain chemically ... crispy customers cut ...

vegetable weapon weeks Wendys Wootan worldwide years

York



# The text

---



## **McDonald's slims down spuds**

Fast-food chain to reduce certain types of fat in its french fries with new cooking oil.

**NEW YORK (CNN/Money) - McDonald's Corp. is cutting the amount of "bad" fat in its french fries nearly in half, the fast-food chain said Tuesday as it moves to make all its fried menu items healthier.**

But does that mean the popular shoestring fries won't taste the same? The company says no. "It's a win-win for our customers because they are getting the same great french-fry taste along with an even healthier nutrition profile," said Mike Roberts, president of McDonald's USA.

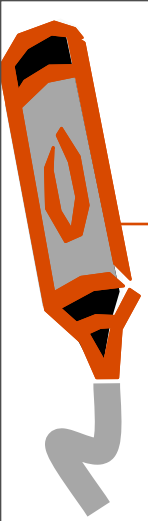
But others are not so sure. McDonald's will not specifically discuss the kind of oil it plans to use, but at least one nutrition expert says playing with the formula could mean a different taste.

Shares of Oak Brook, Ill.-based McDonald's (MCD: down \$0.54 to \$23.22, Research, Estimates) were lower Tuesday afternoon. It was unclear Tuesday whether competitors Burger King and Wendy's International (WEN: down \$0.80 to \$34.91, Research, Estimates) would follow suit.

Neither company could immediately be reached for comment.

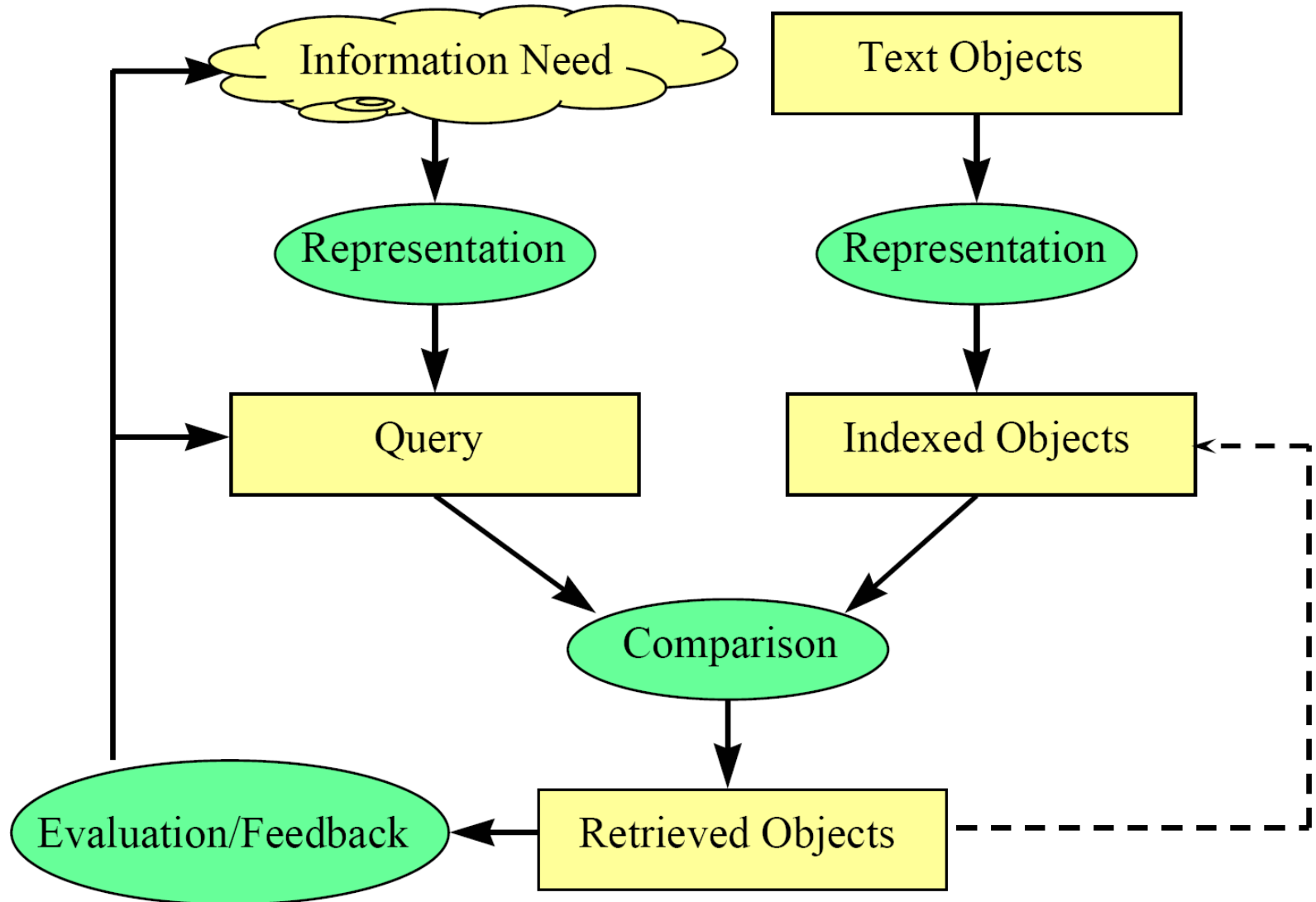
# Text representation

---



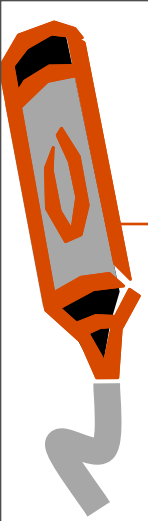
- Text representation
  - what makes a “good” representation?
  - how is a representation generated from text?
  - what are retrievable objects and how are they organized?
- Representing information needs
  - what is an appropriate query language?
- Comparing representations
  - what is a “good” model of retrieval?

# Retrieval process



# Basic approaches

---



- Boolean: exact match vs. best match
- Geometric: vector space model
- Probabilistic: language models
- Graph-based: PageRank



# Vector-space Model

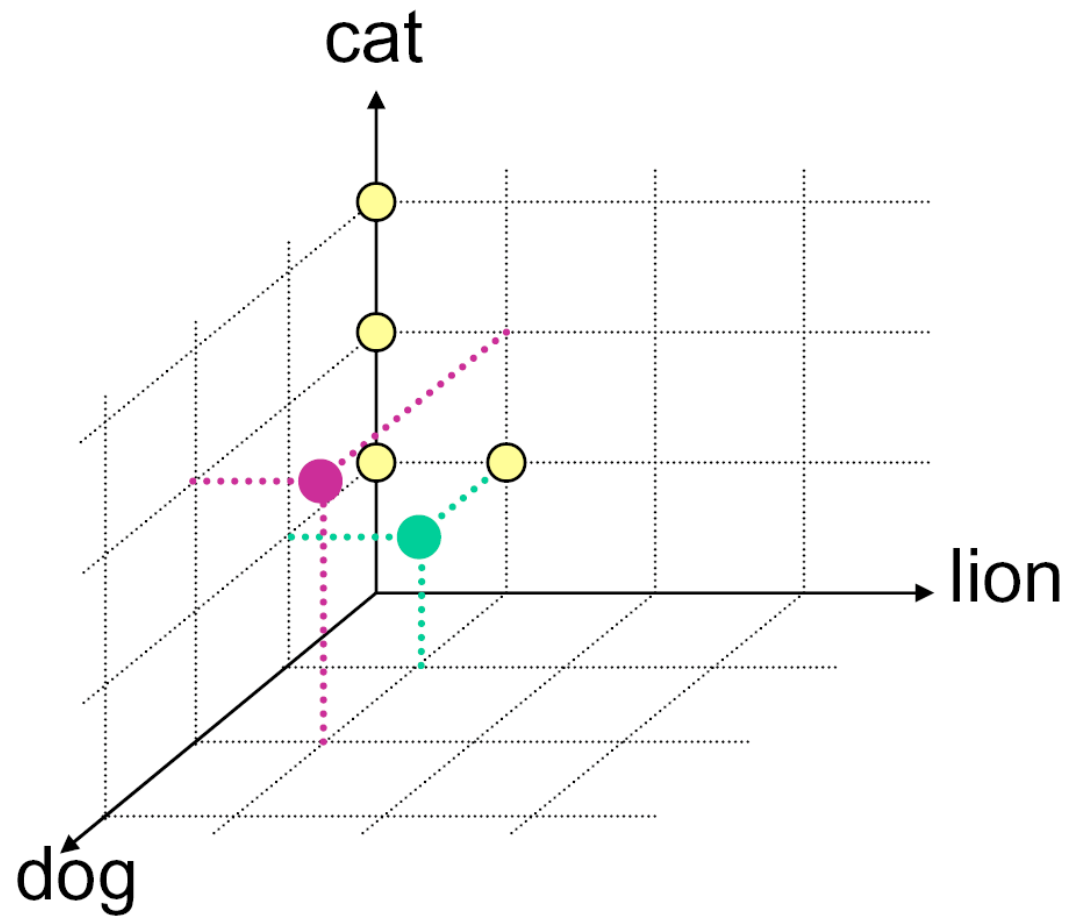
---

- Represent documents and queries as vectors in the term space
- Issue: find the right coefficients...
- Use a geometric similarity measure, often angle-related

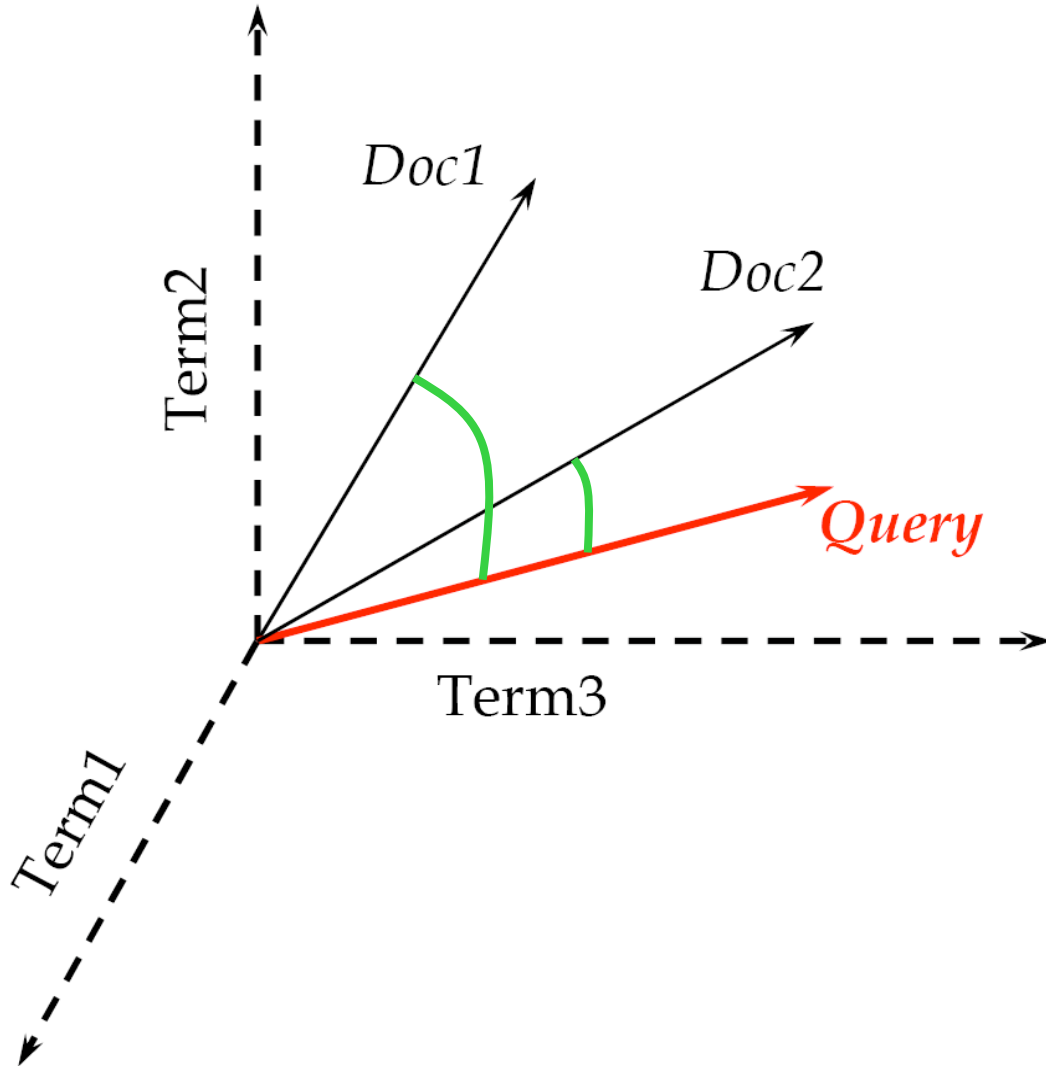
# Example



- cat
- cat cat
- cat cat cat
- cat lion
- lion cat
- cat lion dog
- cat cat lion dog dog



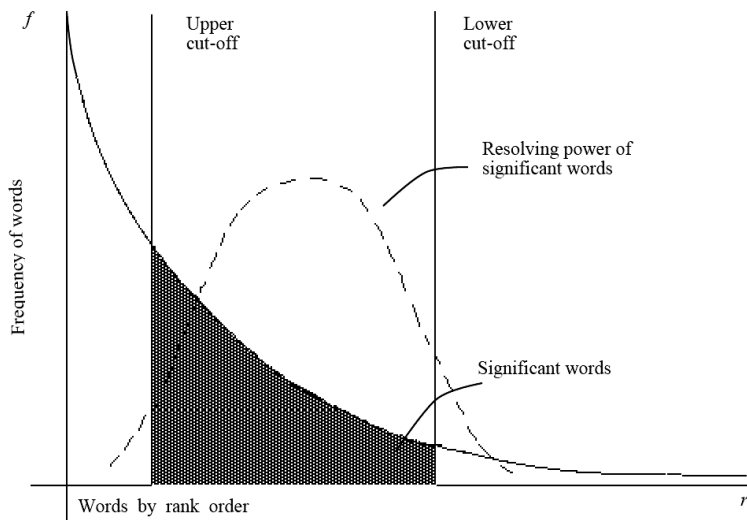
# Vector similarity: angles



# Weights



collection



documents

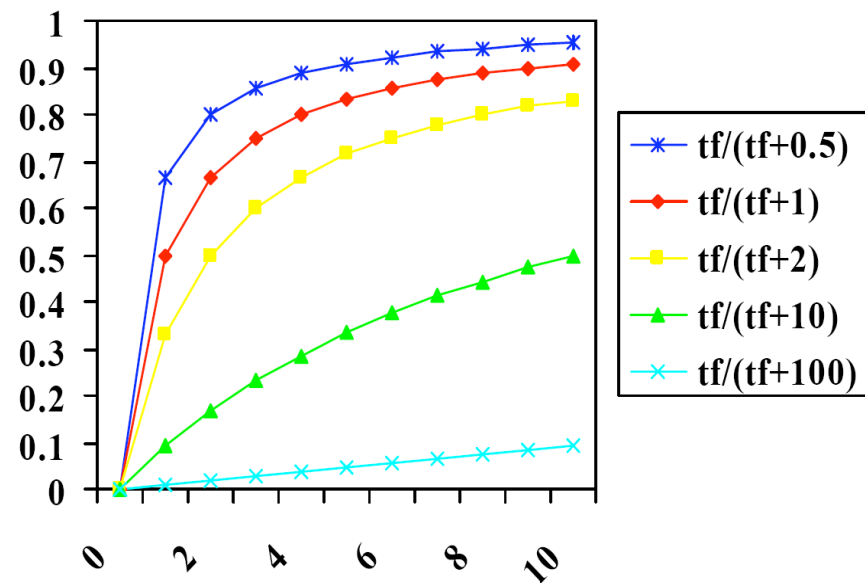



Figure 2.1. A plot of the hyperbolic curve relating  $f$ , the frequency of occurrence and  $r$ , the rank order (Adapted from Schultz <sup>4</sup>page 120)



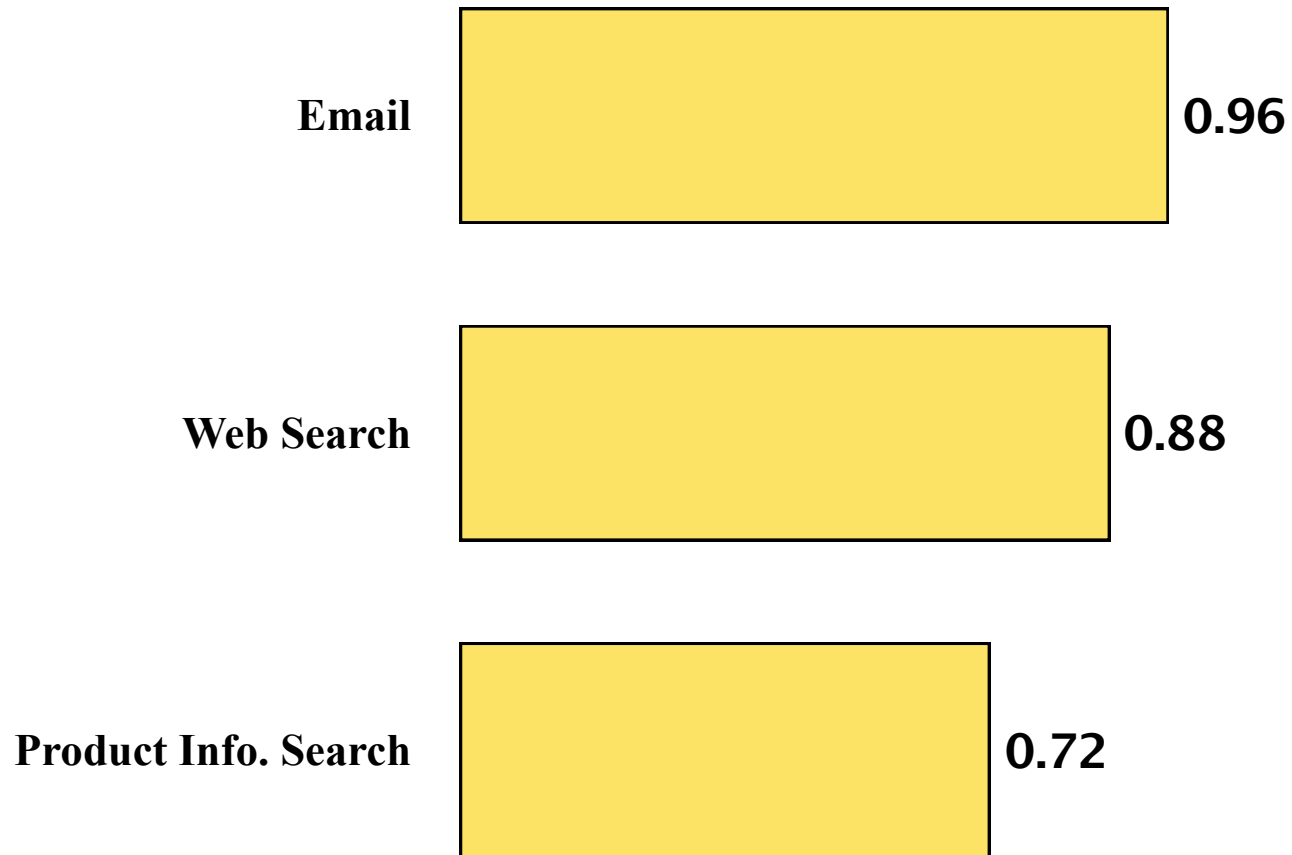
# Overview

---

- 
- What is information retrieval ?
  - How do search engines work ?
  - The internet & web search
  - Adversarial IR

# Top online activities

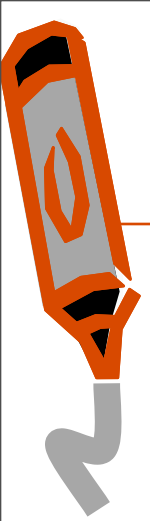
---



# US users (2002)


---

- Total Internet users = 111 M
- Do a search on any given day = 33 M
- Have used Internet to search = 85%




# Search on the web

---

- 
- Corpus: The publicly accessible Web: static + dynamic
  - Goal: Retrieve high quality results relevant to the user's need
    - (not docs!)
  - Need
    - Informational - want to learn about something (~40%)
    - Navigational - want to go to that page (~25%)
    - Transactional - want to do something (web-mediated) (~35%)
      - Access a service
      - Downloads
      - Shop
    - Gray areas
      - Find a good hub
      - Exploratory search “see what’s there”


# Search on the web

---


- 
- Corpus: The publicly accessible Web: static + dynamic
  - Goal: Retrieve high quality results relevant to the user's need
    - (not docs!)
  - Need
    - Informational - want to learn about something (~40%) **Low hemoglobin**
    - Navigational - want to go to that page (~25%)
    - Transactional - want to do something (web-mediated) (~35%)
      - Access a service
      - Downloads
      - Shop
    - Gray areas
      - Find a good hub
      - Exploratory search “see what’s there”

# Search on the web


---

- 
- Corpus: The publicly accessible Web: static + dynamic
  - Goal: Retrieve high quality results relevant to the user's need
    - (not docs!)
  - Need
    - Informational - want to learn about something (~40%) **Low hemoglobin**
    - Navigational - want to go to that page (~25%) **United Airlines**
    - Transactional - want to do something (web-mediated) (~35%)
      - Access a service
      - Downloads
      - Shop
    - Gray areas
      - Find a good hub
      - Exploratory search “see what’s there”

# Search on the web

- 
- Corpus: The publicly accessible Web: static + dynamic
  - Goal: Retrieve high quality results relevant to the user's need
    - (not docs!)
  - Need
    - Informational - want to learn about something (~40%) **Low hemoglobin**
    - Navigational - want to go to that page (~25%) **United Airlines**
    - Transactional - want to do something (web-mediated) (~35%)
      - Access a service **Tampere weather**
      - Downloads **Mars surface images**
      - Shop **Nikon CoolPix**
    - Gray areas
      - Find a good hub
      - Exploratory search “see what’s there”

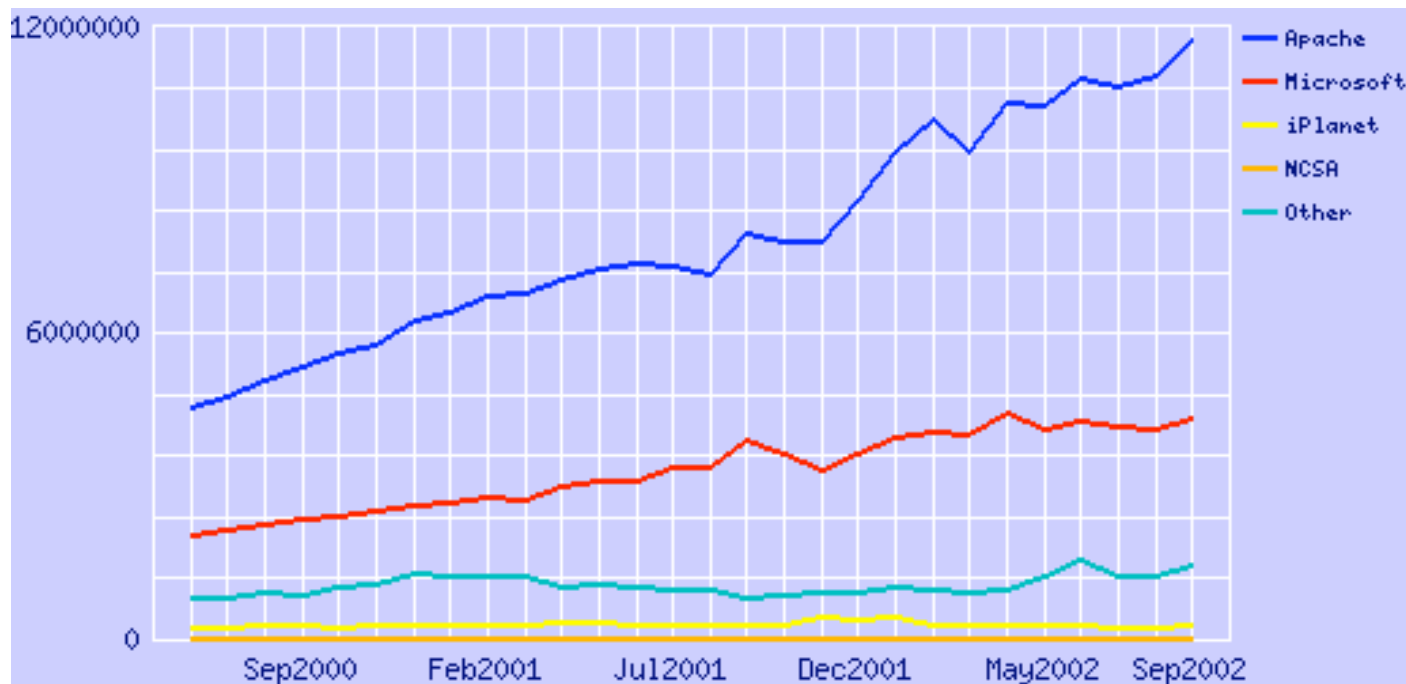
# Search on the web

- 
- Corpus: The publicly accessible Web: static + dynamic
  - Goal: Retrieve high quality results relevant to the user's need
    - (not docs!)
  - Need
    - Informational - want to learn about something (~40%) **Low hemoglobin**
    - Navigational - want to go to that page (~25%) **United Airlines**
    - Transactional - want to do something (web-mediated) (~35%)
      - Access a service **Tampere weather**
      - Downloads **Mars surface images**
      - Shop **Nikon CoolPix**
    - Gray areas
      - Find a good hub
      - Exploratory search “see what’s there”  
**Car rental Finland**



# Scale

- Immense amount of content
  - 10-20B static pages, doubling every 8-12 months
  - Lexicon Size: 10s-100s of millions of words
- Authors galore (1 in 4 hosts run a web server)



# Diversity

- Languages/Encodings
  - Hundreds (thousands ?) of languages, W3C encodings: 55
  - Home pages (1997): English 82%, Next 15: 13%
  - Google (mid 2001): English: 53%
- Popular Query Topics (from 1M Google queries, 06/2000)

Arts	14.6%	Arts: Music	6.1%
Computers	13.8%	Regional: North America	5.3%
Regional	10.3%	Adult: Image Galleries	4.4%
Society	8.7%	Computers: Software	3.4%
Adult	8%	Computers: Internet	3.2%
Recreation	7.3%	Business: Industries	2.3%
Business	7.2%	Regional: Europe	1.8%
...	...	...	...

# Rate of change

720K pages from 270 popular sites sampled daily from Feb 17 - Jun 14, 1999

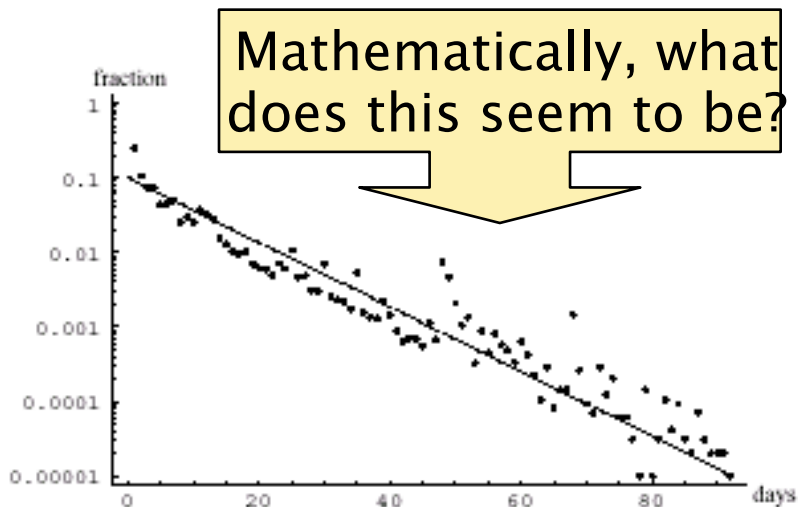


Figure 11: Change intervals for pages with the average change interval of 10 days

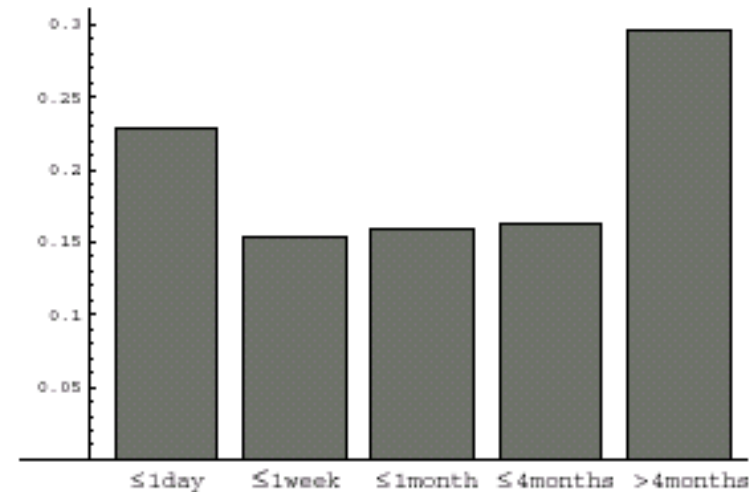
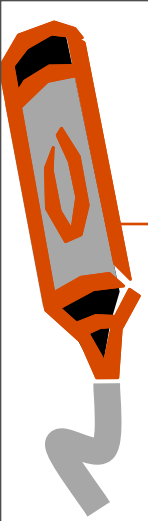


Figure 12: Percentage of pages with given average interval of change

# Web idiosyncrasies

---



- Distributed authorship
  - Millions of people creating pages with their own style, grammar, vocabulary, opinions, facts, falsehoods ...
  - Not all have the purest motives in providing high-quality information - commercial motives drive “spamming” - 100s of millions of pages.



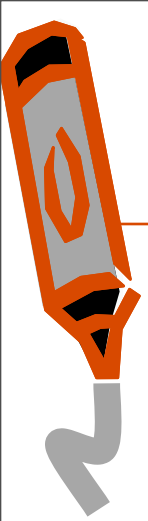
# Web search users

---

- Ill-defined queries
  - Short
    - AV 2001: 2.54 terms avg, 80% 3 words or less)
  - Imprecise terms
  - Sub-optimal syntax (80% queries without operator)
  - Low effort
- Specific behavior
  - 85% look over one result screen only (mostly above the fold)
  - 78% of queries are not modified (one query/session)
  - Follow links - “the scent of information” ...
- Wide variance in
  - Needs
  - Expectations
  - Knowledge
  - Bandwidth

# Evolution of search engines

---



1995-1997 AV,  
Excite, Lycos, etc

From 1998-2003.  
Made popular by  
Google

present



# Evolution of search engines

---

- First generation -- use only “on page”, text data
  - Vector-space model

1995-1997 AV,  
Excite, Lycos, etc

From 1998-2003.  
Made popular by  
Google

present



# Evolution of search engines

---

- First generation -- use only “on page”, text data
  - Vector-space model

1995-1997 AV,  
Excite, Lycos, etc

- Second generation -- use off-page, web-specific data
  - Link (or connectivity) analysis
  - Click-through data (What results people click on)
  - Anchor-text (How people refer to this page)

From 1998-2003.  
Made popular by  
Google

present





# Evolution of search engines

---

- First generation -- use only “on page”, text data
  - Vector-space model

1995-1997 AV,  
Excite, Lycos, etc

- Second generation -- use off-page, web-specific data
  - Link (or connectivity) analysis
  - Click-through data (What results people click on)
  - Anchor-text (How people refer to this page)

From 1998-2003.  
Made popular by  
Google

- Third generation -- answer “the need behind the query”
  - Semantic analysis -- what is this about?
  - Focus on user need, rather than on query
  - Context determination
  - Helping the user
  - Integration of search and text analysis

present



# Second generation

---

- Ranking -- use off-page, web-specific data
  - Link (or connectivity) analysis
  - Click-through data (results people click on)
  - Anchor-text (how people refer to this page)
- Crawling
  - Algorithms to create the best possible corpus



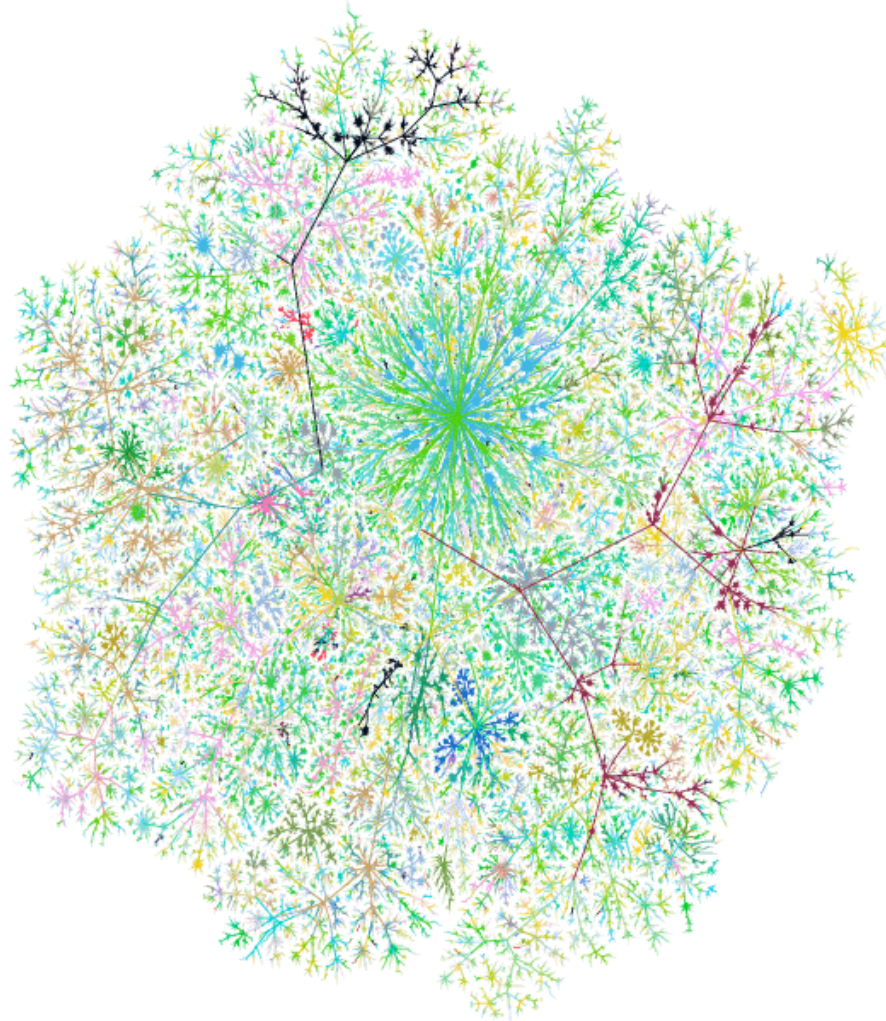
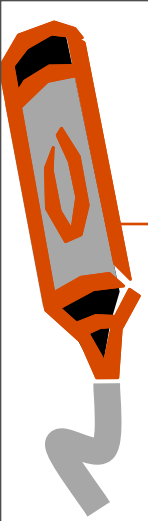
# Connectivity analysis

---

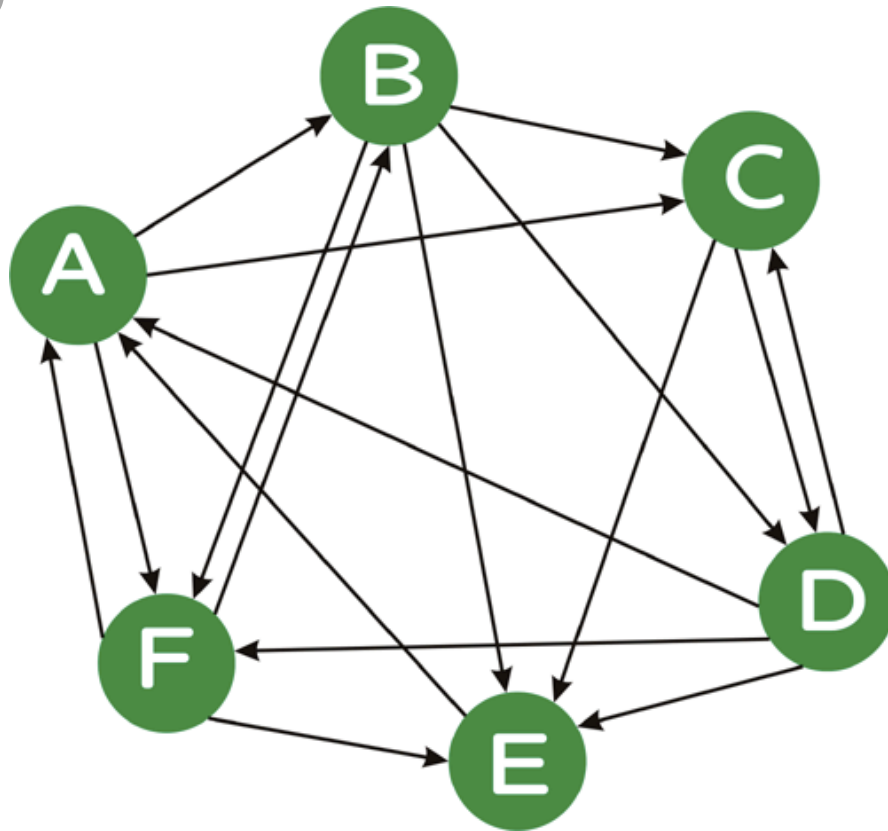
- Idea: Mine hyperlink information
- Assumptions:
  - Links often connect related pages
  - A link between pages is a recommendation  
“people vote with their links”

# PageRank scoring

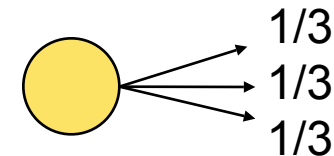
---



# PageRank scoring




- Imagine a browser doing a random walk on web pages...



- “In the steady state” each page has a long-term visit rate - the PageRank score


# PageRank summary

---

- 
- Preprocessing:
    - Crawl web & create graph
    - Compute PageRank
    - Recompute often...
  - Query processing:
    - Retrieve pages meeting query.
    - Rank them by PageRank.
    - Order is query-independent!
  - Pagerank is a global property
    - Your pagerank score depends on “everybody” else
    - Harder to spam than simple popularity counting
  - In reality: Hundreds of features (e.g., anchor text)

# Overview

---

- 
- What is information retrieval ?
  - How do search engines work ?
  - The internet & web search
  - Adversarial IR



# Adversarial IR (spamdexing)

---

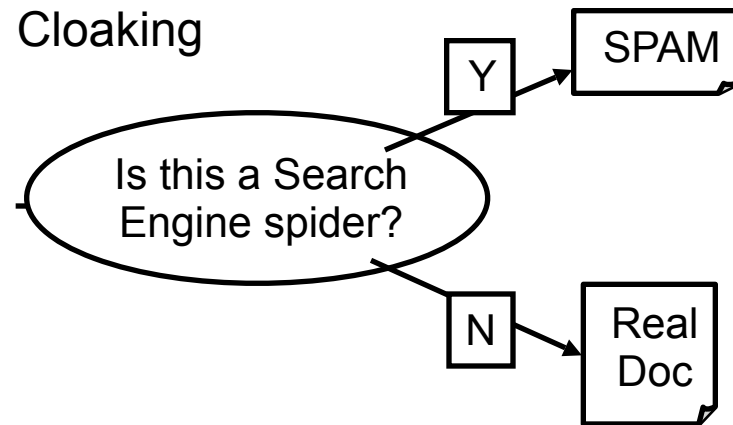
- Motives
  - Commercial, political, religious, lobbies
  - Promotion funded by advertising budget
- Operators
  - Contractors (Search Engine Optimizers)
  - Web masters
  - Hosting services
- Forum
  - Web master world ( [www.webmasterworld.com](http://www.webmasterworld.com) )
    - Search engine specific tricks
    - Discussions about academic papers ☺



# A few spam technologies

- Cloaking
  - Serve fake content to search engine robot
  - DNS cloaking: Switch IP address. Impersonate
- Doorway pages
  - Pages optimized for a single keyword that re-direct to the real target page
- Keyword Spam
  - Misleading meta-keywords, excessive repetition of a term, fake “anchor text”
  - Hidden text with colors, CSS tricks, etc.
- Link spamming
  - Mutual admiration societies, hidden links, awards
  - Domain flooding: numerous domains that point or re-direct to a target page
- Robots
  - Fake click stream
  - Fake query stream
  - Millions of submissions via Add-Url

## Cloaking



## Meta-Keywords =

"... London hotels, hotel, holiday inn, hilton, discount, booking, reservation, sex, mp3, britney spears, viagra, ..."

# The war against spam

- Quality signals - Prefer authoritative pages based on:
  - Votes from authors (linkage signals)
  - Votes from users (usage signals)
- Policing of URL submissions
  - Anti robot test
- Limits on meta-keywords
- Robust link/text analysis
  - Ignore statistically implausible linkage (or text)
  - Use link analysis to detect spammers (guilt by association)
- Spam recognition by machine learning
  - Training set based on known spam
- Family friendly filters
  - Linguistic analysis, general classification techniques, etc.
  - For images: flesh tone detectors, source text analysis, etc.
- Editorial intervention
  - Blacklists
  - Top queries audited
  - Complaints addressed



# Google Bombs

Anchor text “link” spam...



**Web** Results 1 - 10 of about 969,000 for [miserable failure](#). (0.06 seconds)

## [Biography of President George W. Bush](#)

Biography of the president from the official White House web site.

[www.whitehouse.gov/president/gwbbio.html](http://www.whitehouse.gov/president/gwbbio.html) - 29k - [Cached](#) - [Similar pages](#)

[Past Presidents](#) - [Kids Only](#) - [Current News](#) - [President](#)

[More results from www.whitehouse.gov »](#)

## [Welcome to MichaelMoore.com!](#)

Official site of the gadfly of corporations, creator of the film Roger and Me and the television show The Awful Truth. Includes mailing list, message board, ...

[www.michaelmoore.com/](http://www.michaelmoore.com/) - 35k - [Sep 1, 2005](#) - [Cached](#) - [Similar pages](#)

## [BBC NEWS | Americas | 'Miserable failure' links to Bush](#)

Web users manipulate a popular search engine so an unflattering description leads to the president's page.

[news.bbc.co.uk/2/hi/americas/3298443.stm](http://news.bbc.co.uk/2/hi/americas/3298443.stm) - 31k - [Cached](#) - [Similar pages](#)

## [Google's \(and Inktomi's\) Miserable Failure](#)

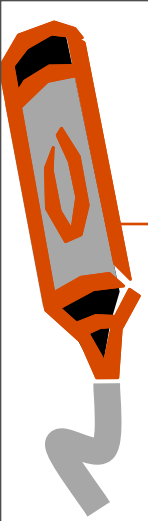
A search for **miserable failure** on Google brings up the official George W.

Bush biography from the US White House web site. Dismissed by Google as not a ...

[searchenginewatch.com/sereport/article.php/3296101](http://searchenginewatch.com/sereport/article.php/3296101) - 45k - [Sep 1, 2005](#) - [Cached](#) - [Similar pages](#)

# Google Bombs Live Demo...

---





# Conclusions

---

- Web search is hard:
  - Web is vast, growing, and changing constantly
  - Bottleneck in specification of information need
- NextGen IR:
  - Multimedia (all info, all the time)
  - NLP & specification of information needs
  - Spam, spam, spam...