

Information Retrieval

Lecture 1 Introduction

What is information retrieval?

The image is a screenshot of a Google search results page for the query "information retrieval". The search bar at the top shows the query and a magnifying glass icon. Below the search bar, there are tabs for "Web", "Books", "Videos", "Images", "News", "More", and "Search tools". The "Web" tab is selected. The search results show "About 11,200,000 results (0.31 seconds)".

The first result is from Wikipedia, titled "Information retrieval - Wikipedia, the free encyclopedia". The snippet reads: "Information retrieval is the activity of obtaining information resources relevant to an information need from a collection of information resources. Searches can be based on metadata or on full-text indexing. Wikipedia".

A red dashed box highlights a section of the page. Inside the box, the title "Information retrieval" is displayed in a large, bold, black font. Below the title, the definition is repeated: "Information retrieval is the activity of obtaining information resources relevant to an information need from a collection of information resources. Searches can be based on metadata or on full-text indexing. Wikipedia". To the right of the definition, there is a small image of a book cover titled "Introduction to Information Retrieval" by Prabhakar Raghavan. Below the book cover, the text "Book by Prabhakar ..." and "Prabhakar Raghavan" is visible.

Below the Wikipedia result, there is another result from Springer, titled "Information Retrieval - Springer". The snippet reads: "Subscription e-journal dedicated to theory and experimentation in information retrieval. Sample copy available."

Why information retrieval

- Information overload
 - “It refers to the difficulty a person can have understanding an issue and making decisions that can be caused by the presence of too much information.” - wiki



Why information retrieval

- Information overload

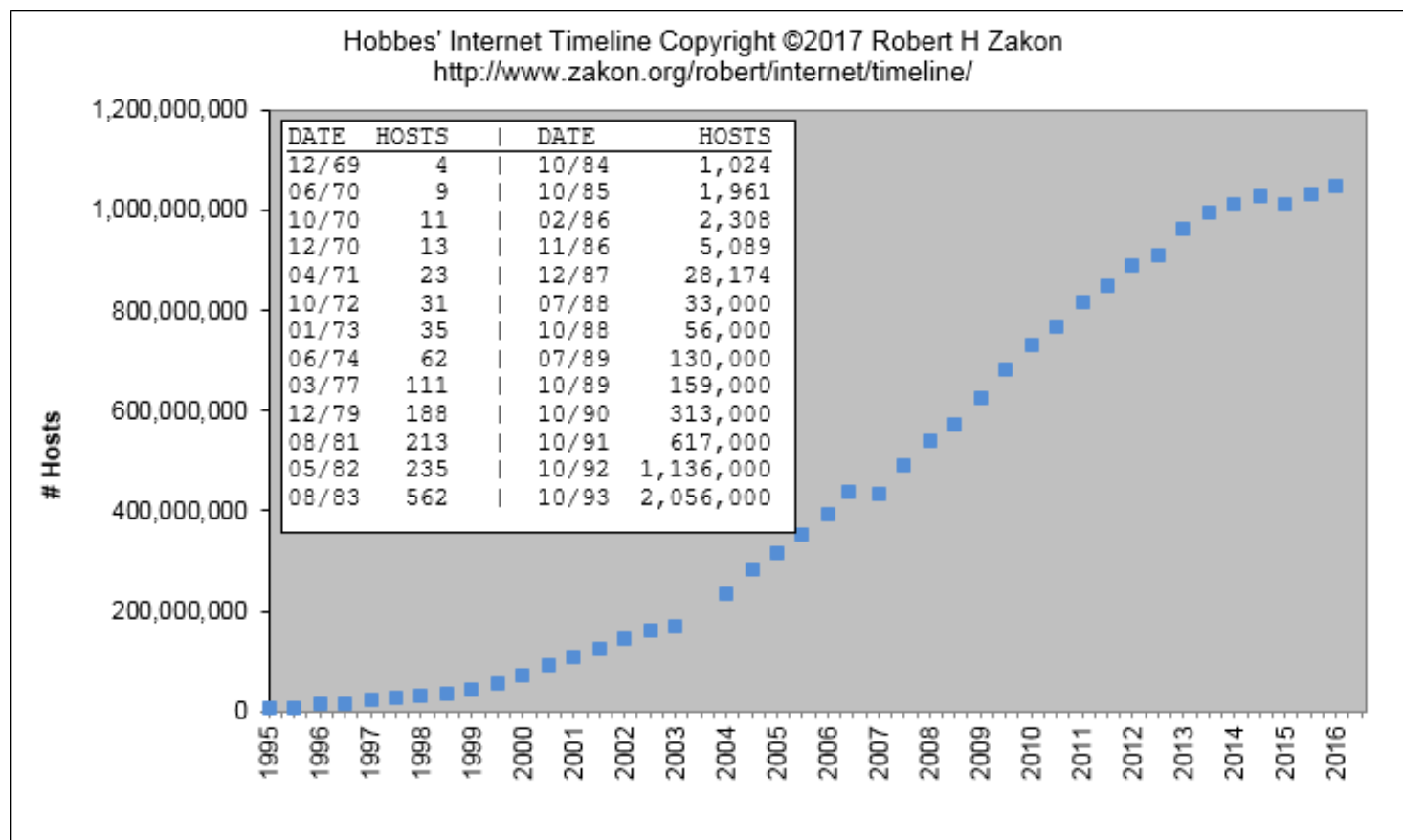


Figure 1: Growth of Internet

Why information retrieval

- Information overload

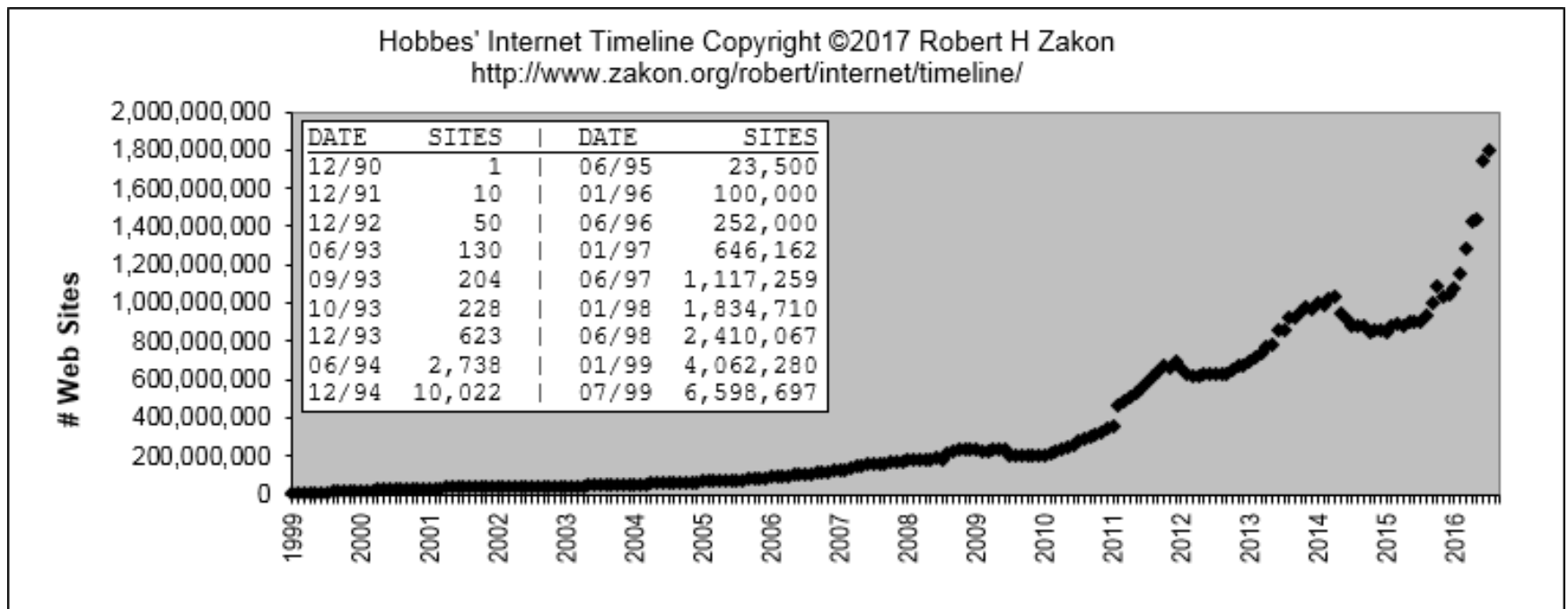


Figure 2: Growth of WWW

Why information retrieval

- Handling unstructured data
 - Structured data: database system is a good choice

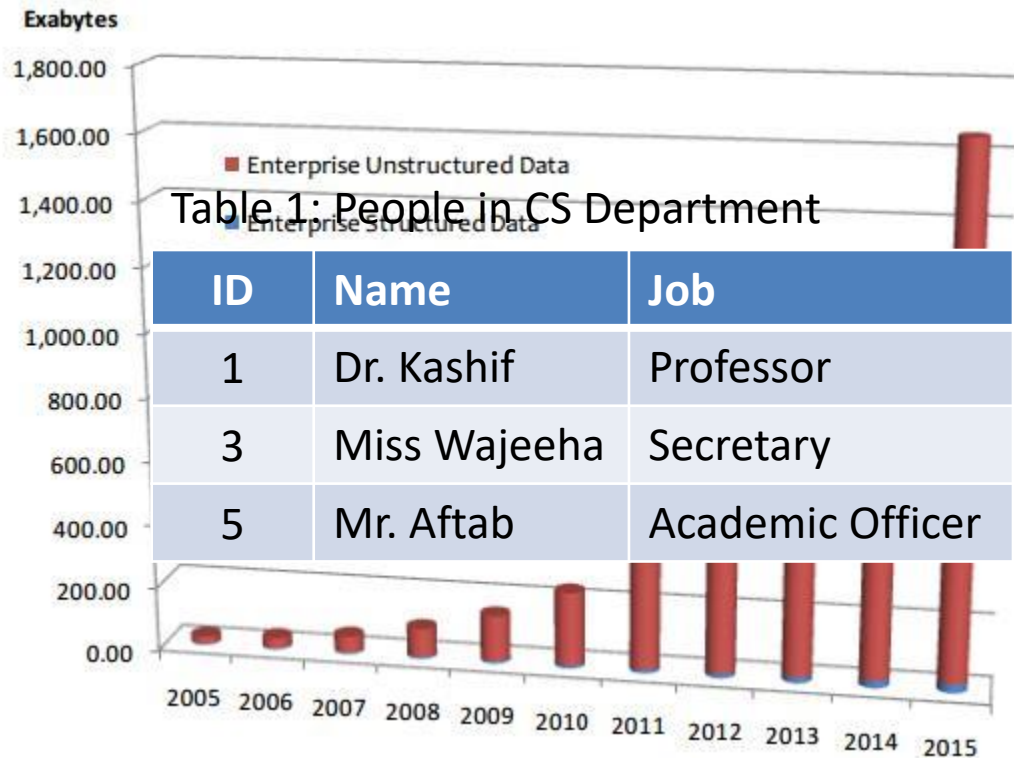
– Unstructured data

- Text

- “XML”

- UI

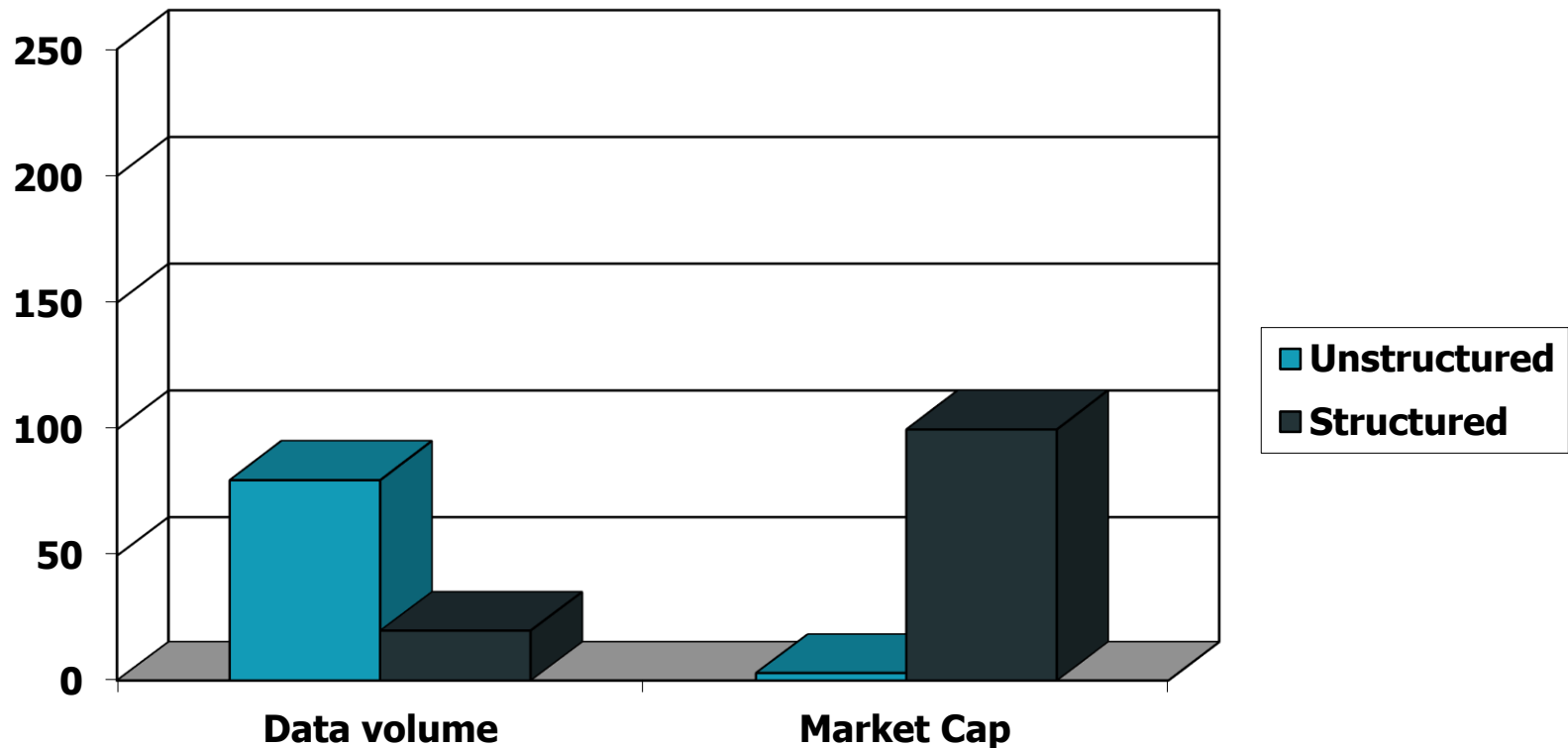
- User



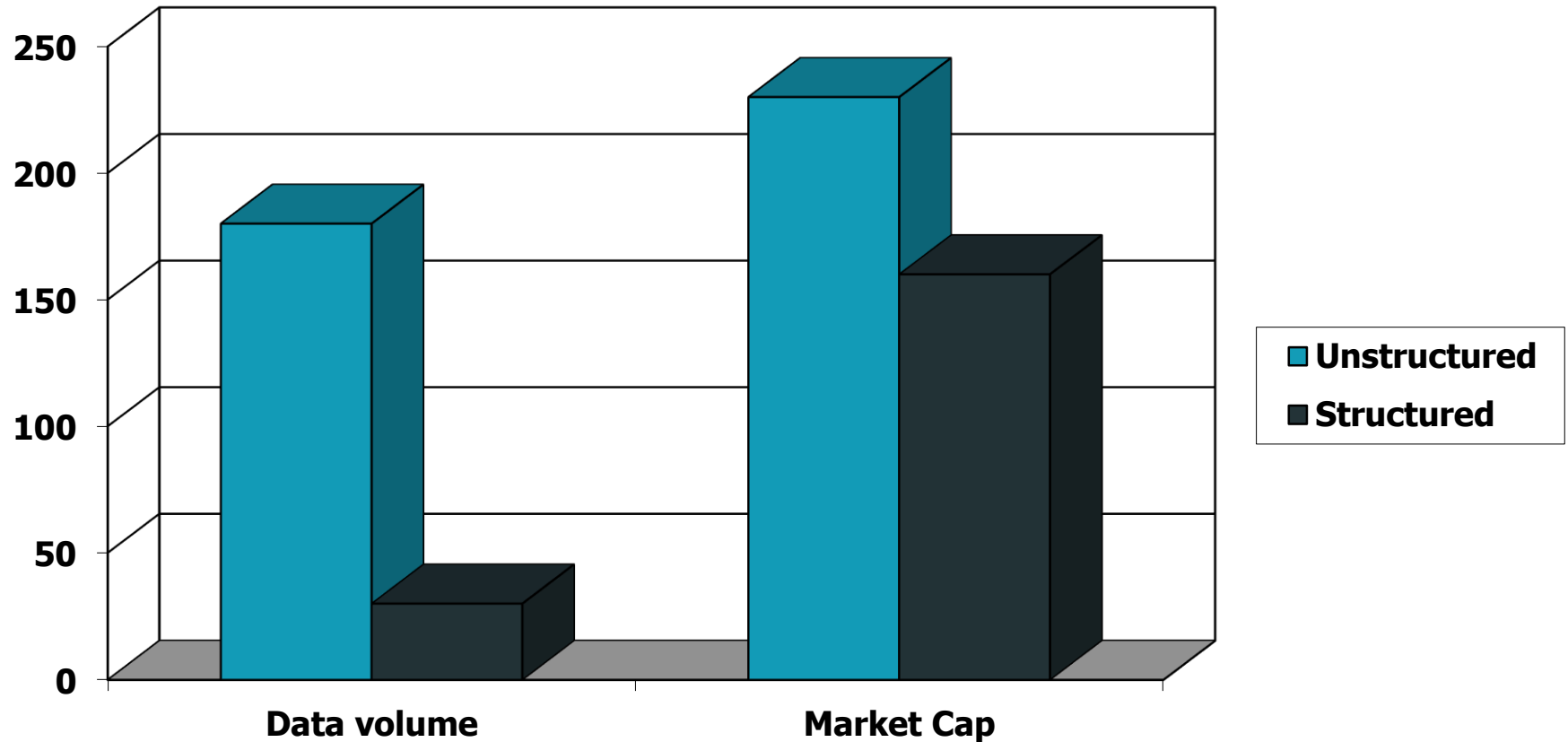
Audio, video...
as

Total Enterprise Data Growth 2005-2015, IDC 2012

Unstructured (text) vs. structured (database) data in the mid-nineties



Unstructured (text) vs. structured (database) data today



Why information retrieval

- An essential tool to deal with information overload



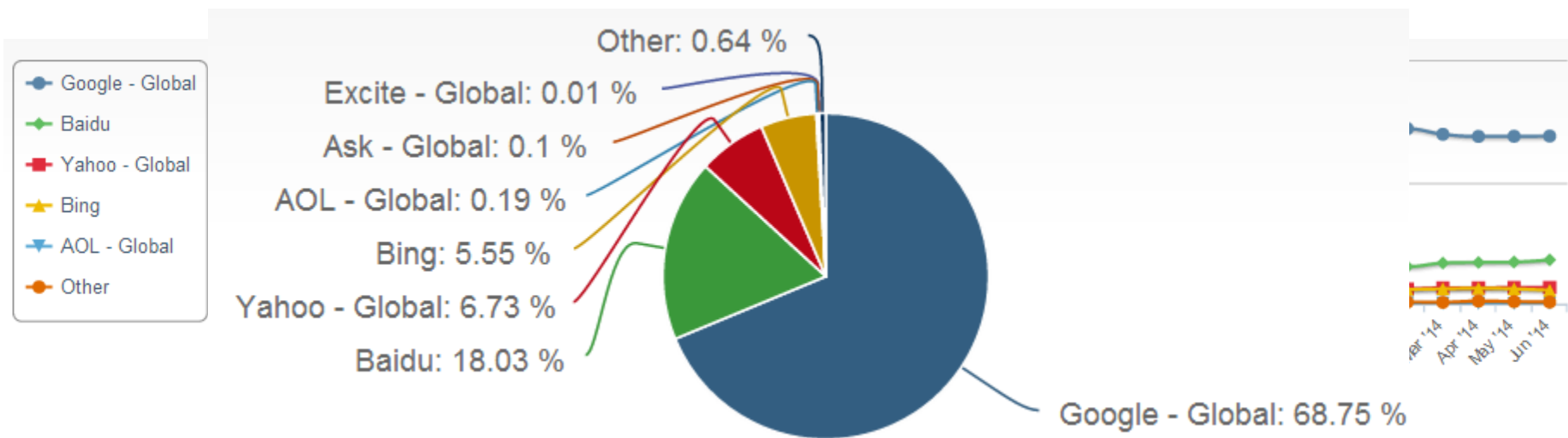
You are
here!

History of information retrieval

- Catalyst
 - Industry: web search engines
 - WWW unleashed explosion of published information and drove the innovation of IR techniques
 - Lycos (started at CMU) was launched and became a major commercial endeavor in 1994
 - Booming of search engine industry: *Magellan, Excite, Infoseek, Inktomi, Northern Light, AltaVista, Yahoo!, Google, and Bing*

Major players in this game

- Global search engine market
 - By <http://marketshare.hitslink.com/search-engine-market-share.aspx>

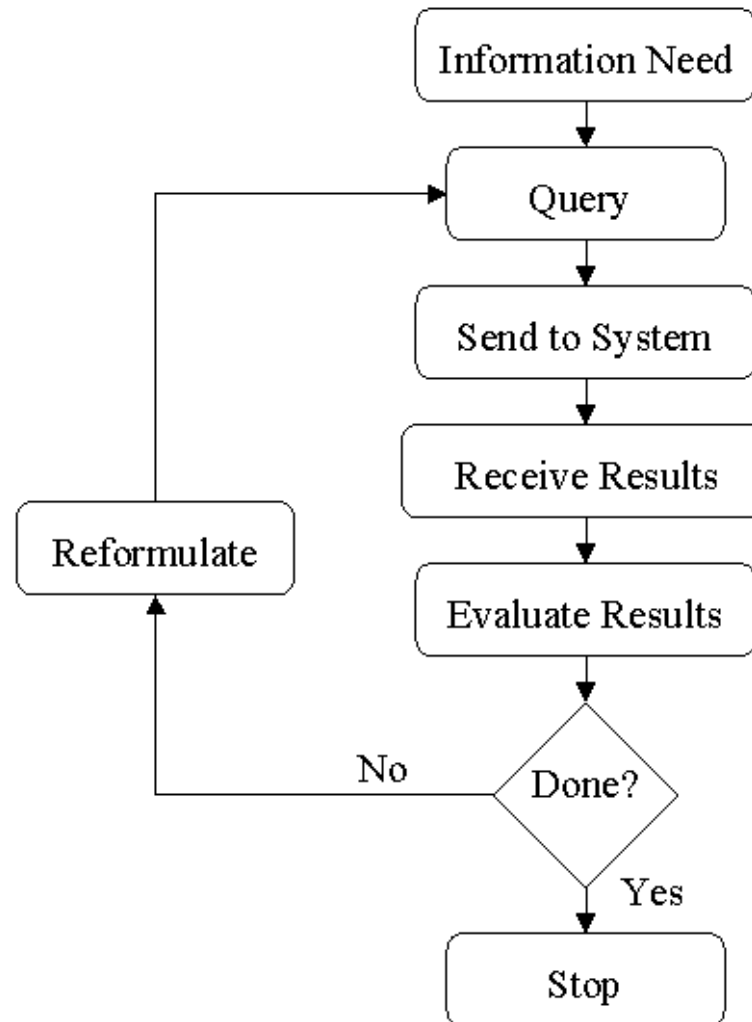


How to perform information retrieval

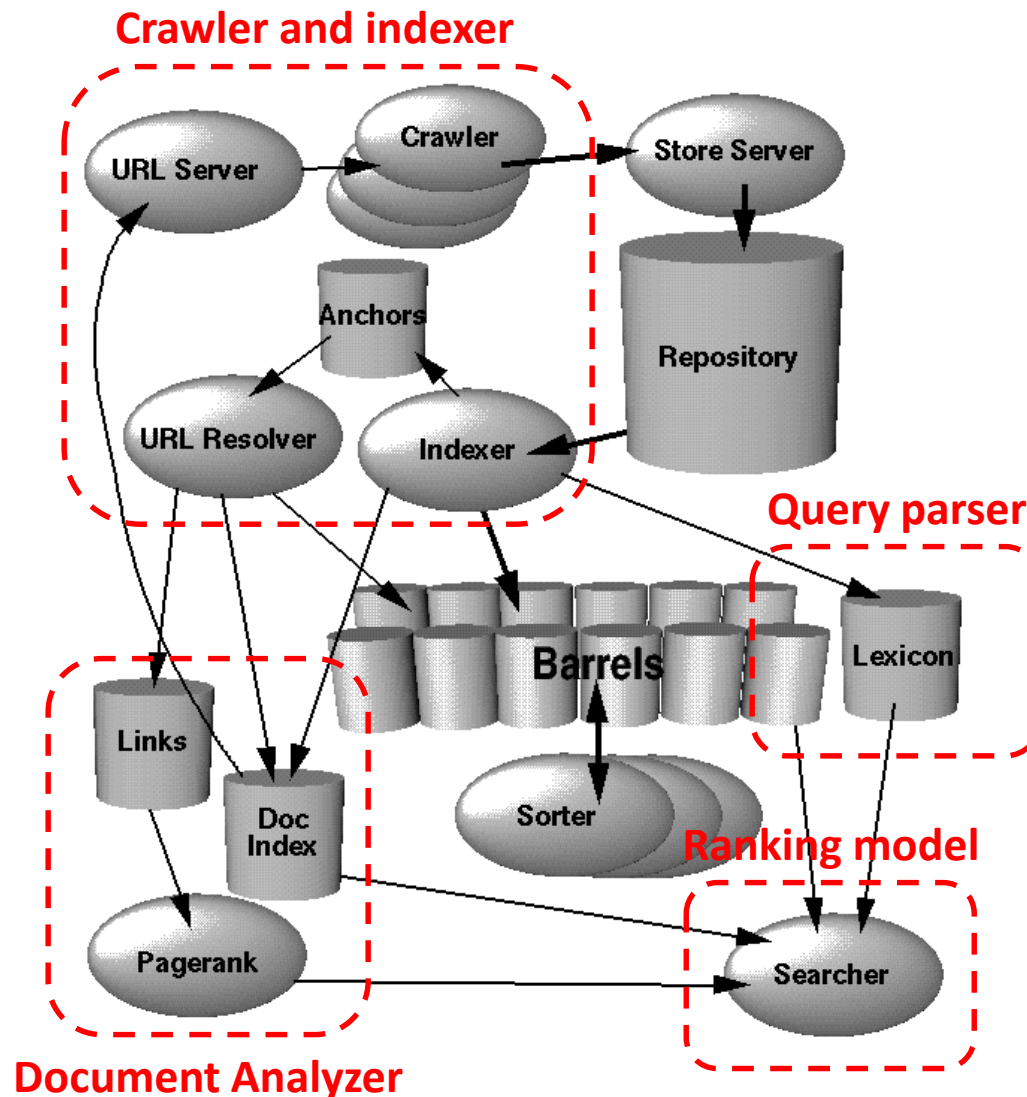
- Information retrieval when we did not have a computer



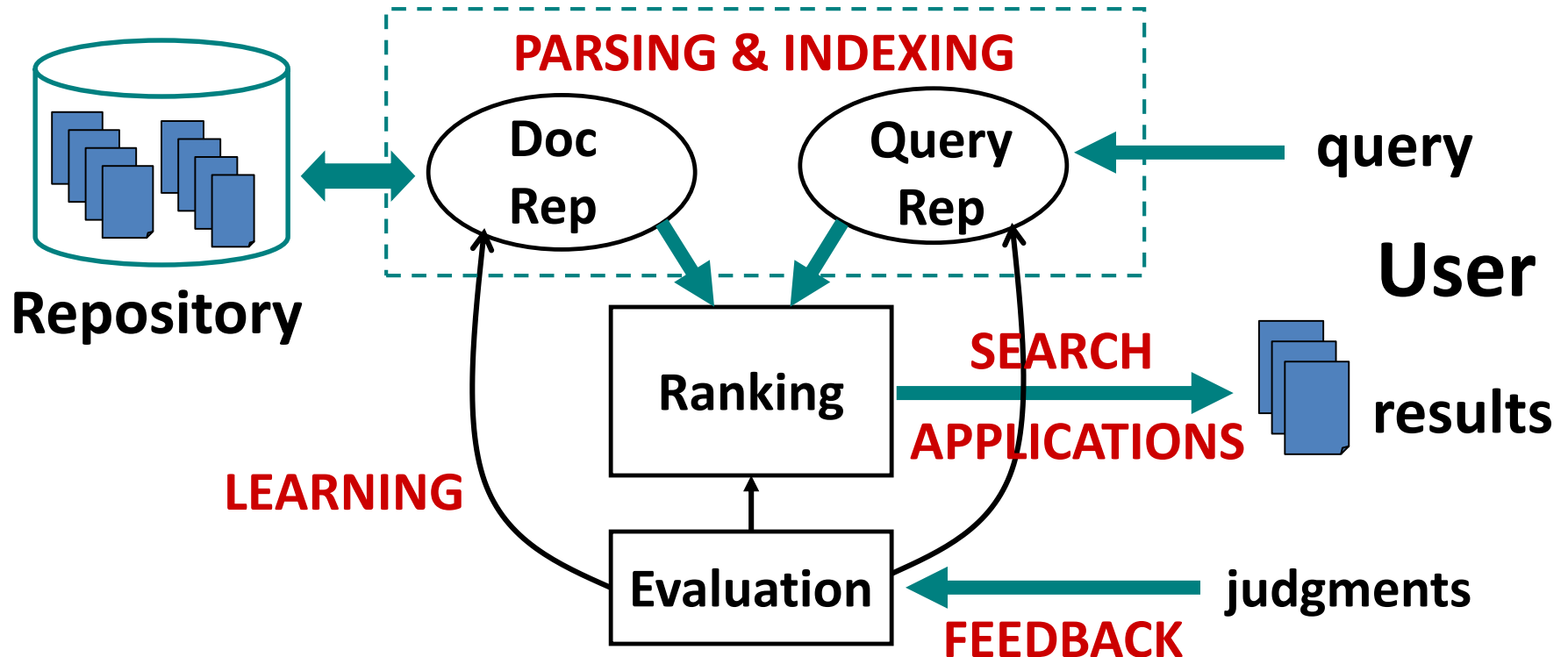
The Standard Retrieval Interaction Model



How to perform information retrieval



How to perform information retrieval



We will cover:

- 1) Search engine architecture;
- 2) Retrieval models;
- 3) Retrieval evaluation;
- 4) Relevance feedback;
- 5) Link analysis;
- 6) Search applications.

Core concepts in IR

- Query representation
 - Lexical gap: say v.s. said
 - Semantic gap
- Document representation
 - Specific data structure for efficient access
- Retrieval model
 - Algorithms that find the most relevant documents for the given information need

A glance of modern search engine

- In old times



A glance of modern search engine

In modern time



A glance of modern search engine

Google

NU pakistan

Demand of understanding

All News Videos Images Maps More

About 20,000,000 results (0.90 seconds)

Demand of efficiency

Demand of accuracy

FAST-NU
www.nu.edu.pk/
Our vision is to become a globally recognized research university of Pakistan within the next ... FAST-NU, Islamabad Campus is organizing convocation 2017.
Lahore Campus · Admission Schedule · How To Apply · Islamabad Campus

National University of Computer and Emerging Sciences - Wikipedia
https://en.wikipedia.org/.../National_University_of_Computer_and_Emerging_Scienc...
The National University of Computer and Emerging Sciences is a private research university in Pakistan. It has multiple campuses based in cosmopolitan cities of Pakistan and has Geek even was held in FAST-NU Lahore. more than 1,000 students from all over the country are participating in the Geek Week 2016 at the ...

pakistan | NU - Het laatste nieuws het eerst op NU.nl
www.nu.nl/tag/pakistan Translate this page
NU.nl; pakistan ... 16 uur geleden Buitenland Zeker vijftien doden bij bomaanslag in Pakistan Een aanslag op een militair voertuig in Pakistan heeft zaterdag ...

Tiden i Karachi, Pakistan nu - Time.is

Demand of diversity

Demand of convenience

See photos

National University of Computer and Emerging Sciences ★

Private university in Islamabad, Pakistan

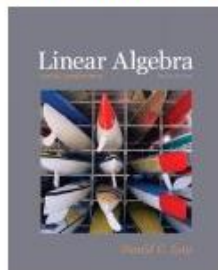
Website Directions

The National University of Computer and Emerging Sciences is a private research university in Pakistan. It has multiple campuses based in cosmopolitan cities of Pakistan and has distinction of being the first multi-campus university in Pakistan. [Wikipedia](#)

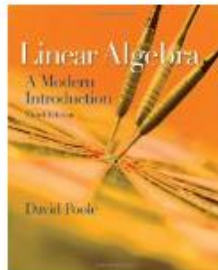
IR is not just about web search

- Web search is just one important area of information retrieval, but not all
- Information retrieval also includes
 - Recommendation

Recommended Based on Your Browsing History



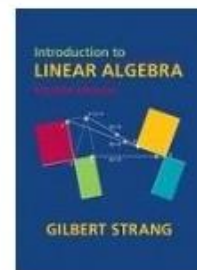
Linear Algebra and Its Applications...
› David C. Lay
Hardcover
★★★★☆ (84)
~~\$183.33~~ **\$141.16**



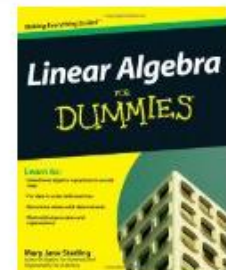
Linear Algebra: A Modern Introduction
› David Poole
Hardcover
★★★★☆ (41)
~~\$316.95~~ **\$289.88**



Linear Algebra
› G. E. Shilov
Paperback
★★★★☆ (34)
~~\$48.95~~ **\$12.65**



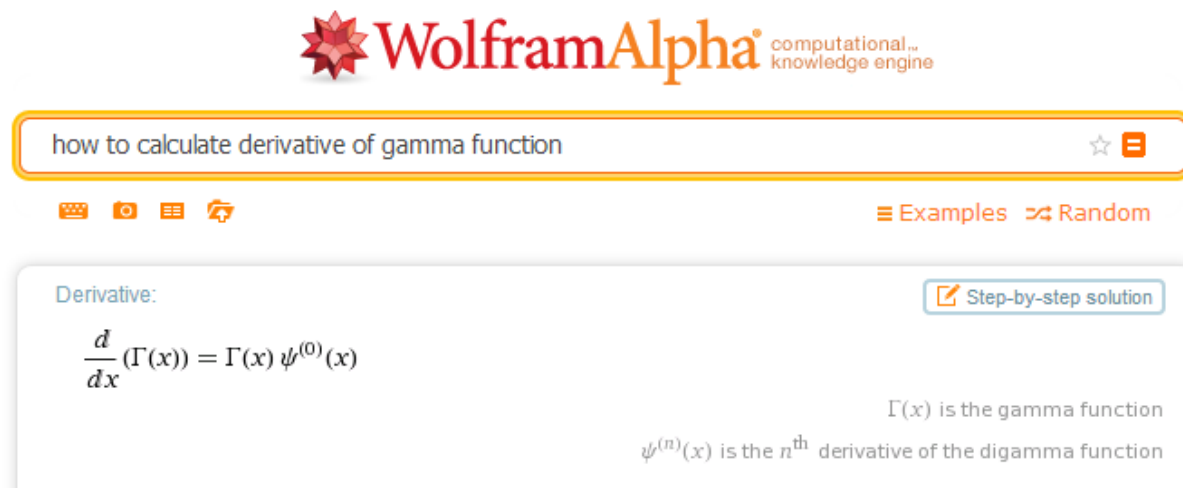
Introduction to Linear Algebra...
› Gilbert Strang
Hardcover
★★★★☆ (57)
~~\$87.50~~ **\$83.13**



Linear Algebra For Dummies
› Mary Jane Sterling
Paperback
★★★★☆ (29)
~~\$49.99~~ **\$16.23**

IR is not just about web search

- Web search is just one important area of information retrieval, but not all
- Information retrieval also includes
 - Question answering



The screenshot shows the WolframAlpha interface. At the top is the WolframAlpha logo with the tagline "computational... knowledge engine". Below the logo is a search bar containing the text "how to calculate derivative of gamma function". To the right of the search bar are a star icon and a menu icon. Below the search bar are several icons: a keyboard, a camera, a list, and a refresh icon. To the right of these icons are the links "Examples" and "Random". Below the search bar is a box containing the text "Derivative:" and a button labeled "Step-by-step solution". The main result is the equation $\frac{d}{dx}(\Gamma(x)) = \Gamma(x) \psi^{(0)}(x)$. Below this equation, there is a note: " $\Gamma(x)$ is the gamma function" and " $\psi^{(n)}(x)$ is the n^{th} derivative of the digamma function".

WolframAlpha computational... knowledge engine

how to calculate derivative of gamma function

Examples Random

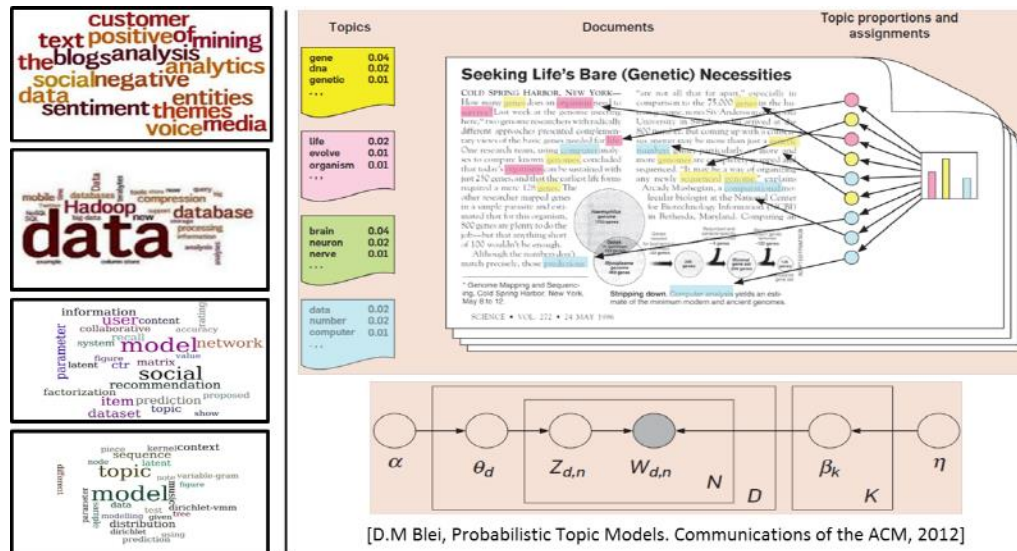
Derivative: [Step-by-step solution](#)

$$\frac{d}{dx}(\Gamma(x)) = \Gamma(x) \psi^{(0)}(x)$$

$\Gamma(x)$ is the gamma function
 $\psi^{(n)}(x)$ is the n^{th} derivative of the digamma function

IR is not just about web search

- Web search is just one important area of information retrieval, but not all
- Information retrieval also includes
 - Text mining



IR is not just about web search

- Web search is just one important area of information retrieval, but not all
- Information retrieval also includes
 - Online advertising

The screenshot shows a Google search for "health care". The search bar at the top contains the text "health care" and a magnifying glass icon. Below the search bar, the navigation tabs include "Web", "News", "Images", "Maps", "Books", "More", and "Search tools". The search results indicate "About 782,000,000 results (0.45 seconds)".

The organic search results include:

- HealthCare.gov: Health Insurance Marketplace, Affordable ...**
<https://www.healthcare.gov/> HealthCare.gov
Learn how the health care law affects you at HealthCare.gov. The official site of the Health Insurance Marketplace. See your health insurance choices.
- Individuals & Families**
Individuals & Families. Still need health coverage? You can get ...
- Log In**
Marketplace Account Registration - Log In. New to HealthCare.gov ...
- More results from healthcare.gov »**
- Health Insurance Marketplace**
01. Apr. How to use your new Marketplace coverage ... Loss of ...
- Special Enrollment**
Medicaid - Qualifying Life Event - Complex Case - Exemptions

Two groups of paid advertisements are highlighted with red dashed boxes:

Left Ad Group:

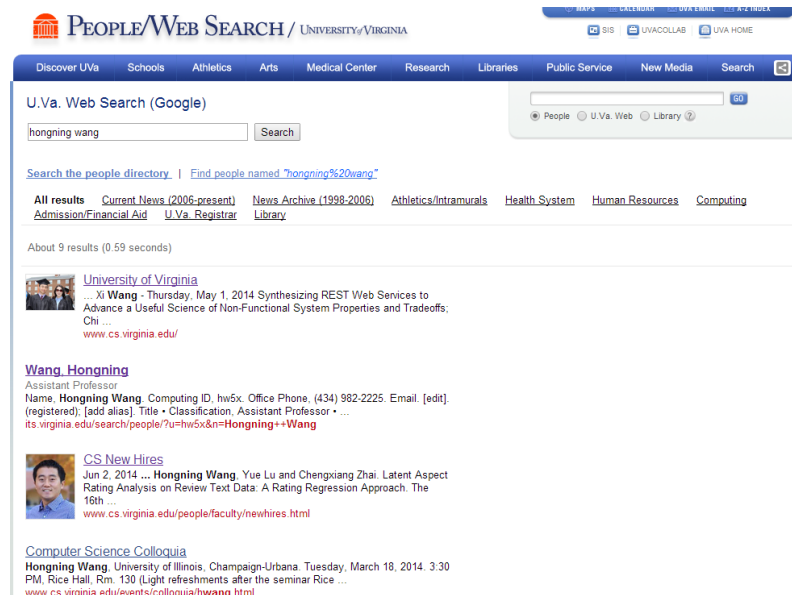
- Need Health Insurance? - MolinaHealthcare.com**
Ad www.molinahealthcare.com/ (877) 751-0665
Do You Qualify for Healthcare Reform Coverage? Find Out Now.
- Cheap Health Insurance - Only Takes A Few Minutes**
Ad www.healthquotejunction.com/ (888) 699-8397
Rates Starting Around \$100/mo.
- Low Cost Health Insurance - IndividualHealthQuotes.com**
Ad www.individualhealthquotes.com/ (866) 406-0696
Blue Cross, Aetna, CIGNA & More! (Illinois Residents Only)

Right Ad Group:

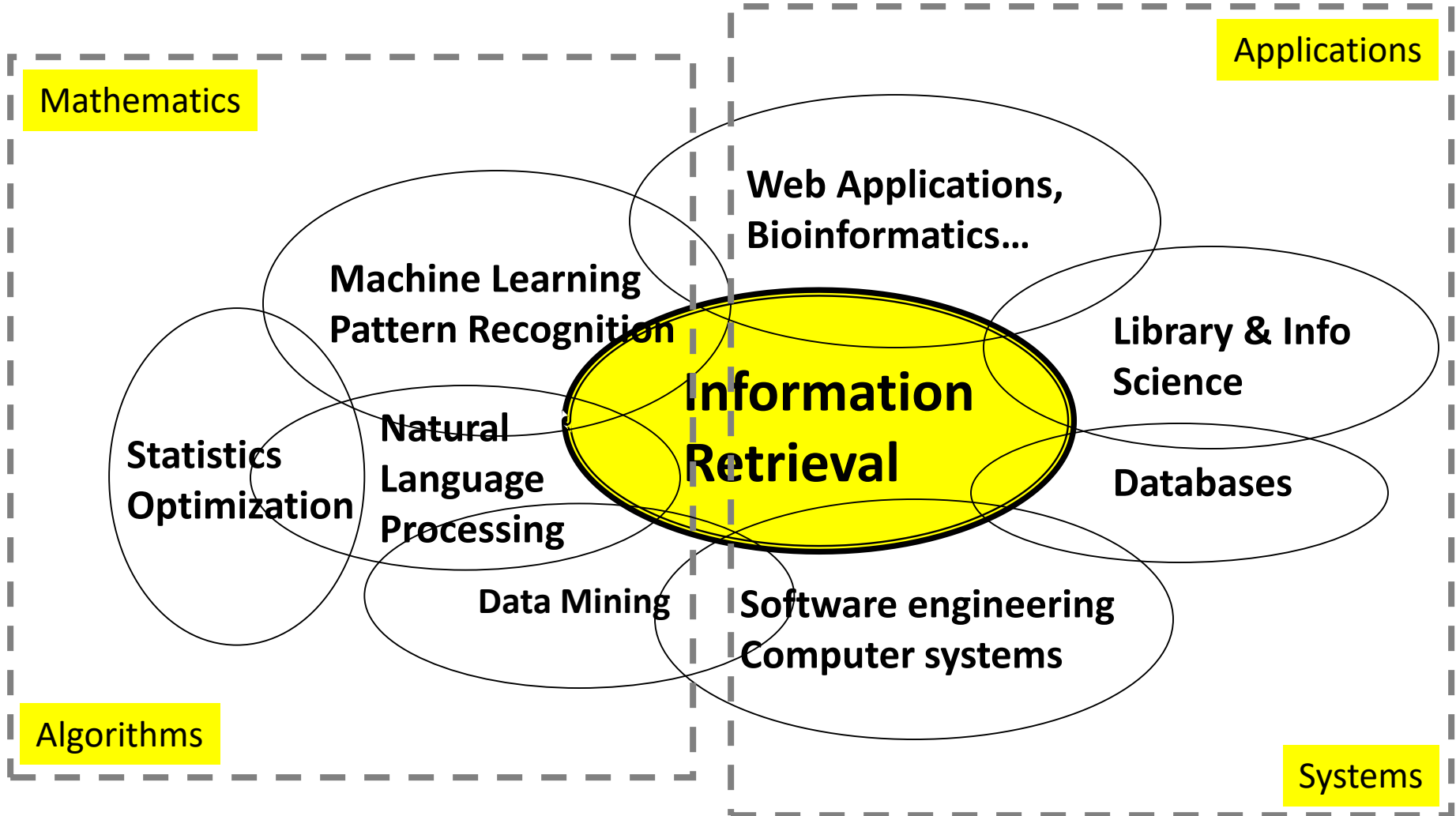
- \$19 Health Insurance**
Ad affordable-health-insurance-plans.org/
New 2014 Discounts. Save 55% - 75%. Compare Affordable Plans Online!
- Christie Clinic**
www.christieclinic.com/
Quality healthcare for east central Illinois
- Obama Health Care**
www.obama-care.org/
See If You Are Eligible & Apply For The Obama Care Health Plan.
- Low Cost Health Insurance**
www.directhealthinsurance.com/
Get Insured (Limited Time Only). Compare Plans / Avoid Fees / Enroll
- Affordable Health Care**
www.hstinsurancequotes.com/affordable
Affordable Health Insurance Plans. 2014 Pricing Options - Free Quotes!

IR is not just about web search

- Web search is just one important area of information retrieval, but not all
- Information retrieval also includes
 - Enterprise search: web search + desktop search



Related Areas



IR v.s. DBs

- Information Retrieval:
 - Unstructured data
 - Semantics of object are subjective
 - Simple key work queries
 - Relevance-driven retrieval
 - Effectiveness is primary issue, though efficiency is also important
- Database Systems:
 - Structured data
 - Semantics of each object are well defined
 - Structured query languages (e.g., SQL)
 - Exact retrieval
 - Emphasis on efficiency

IR and DBs are getting closer

- IR => DBs

- Approximate search is available in DBs
- Eg. in MySQL

```
mysql> SELECT * FROM articles  
-> WHERE MATCH (title,body)  
    AGAINST ('database');
```

- DBs => IR

- Use information extraction to convert unstructured data to structured data
- Semi-structured representation: XML data; queries with structured information

IR v.s. NLP

- Information retrieval
 - Computational approaches
 - Statistical (shallow) understanding of language
- Natural language processing
 - Cognitive, symbolic and computational approaches
 - Semantic (deep) understanding of language

IR and NLP are getting closer

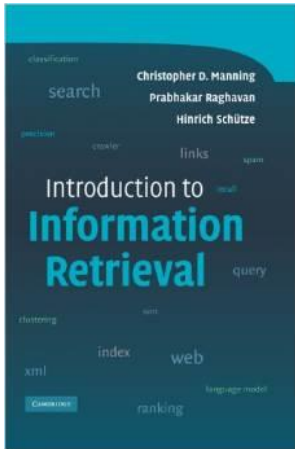
- IR => NLP
 - Larger data collections
 - Scalable/robust NLP techniques, e.g., translation models
- NLP => IR
 - Deep analysis of text documents and queries
 - Information extraction for structured IR tasks

Course Learning Objectives

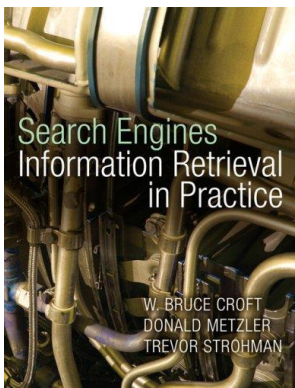
- Enable students to understand the common algorithms and techniques for information retrieval (document indexing and retrieval, query processing, etc)
- Introduce the quantitative evaluation methods for the IR systems and data mining techniques
- Enable students to implement a basic textual information retrieval system using Java or Python
- Introduce the popular probabilistic retrieval methods and ranking principles
- Apply information retrieval techniques to the problems of text clustering, text classification etc.

Course Outline

Text books



- ***Introduction to Information Retrieval.*** Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schuetze, Cambridge University Press, 2007.

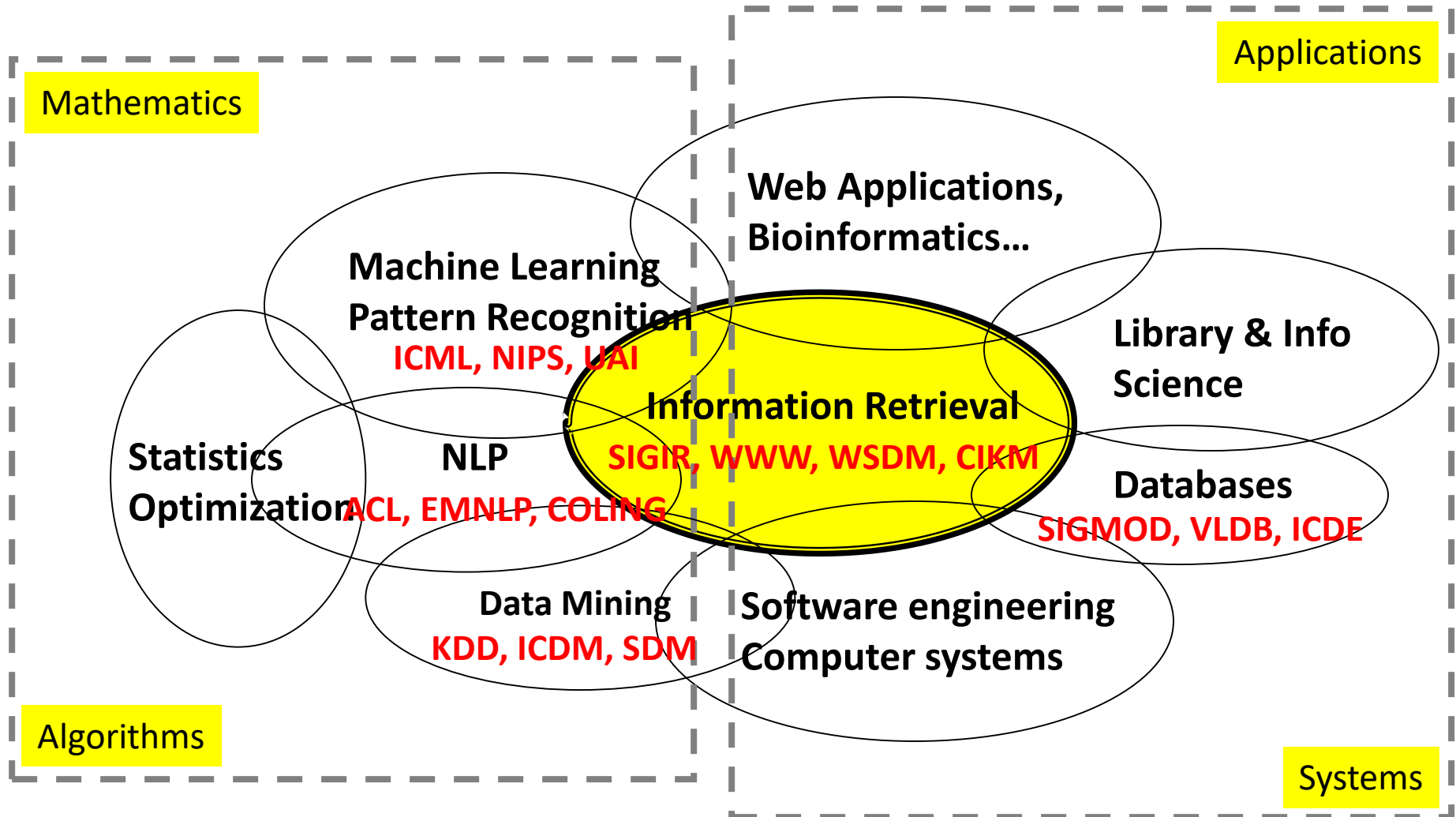


- ***Search Engines: Information Retrieval in Practice.*** Bruce Croft, Donald Metzler, and Trevor Strohman, Pearson Education, 2009.

You should know

- IR originates from library science for handling unstructured data
- IR has many important application areas, e.g., web search, recommendation, and question answering
- IR is a highly interdisciplinary area with DBs, NLP, ML, HCI

What to read?



Top Conferences and Journals in IR Field

- [SIGIR](#): One of the most important and influential conference in IR field (attract more attention from academia), proceedings of publications can be found [here](#).
- [WWW](#): Another most important and influential conference in IR field (attract more attention from industry), proceedings of publications can be found [here](#).
- [WSDM](#): A new but quickly raising conference in the field, attracting attentions from both industry and academia. Proceedings of publications can be found [here](#).
- [CIKM](#): A major conference in IR field. Proceedings of publications can be found [here](#).
- [ECIR](#) Conference Proceedings

- [TOIS](#): One of major journals for IR field.
- Information Processing and Management (Journal)
- Knowledge and Data Engineering (Journal)
- Information Retrieval (Journal)
- Information Science (Journal)
- Knowledge Based systems (Journal)

IR Toolkits

- [ElasticSearch](#)
- [Lucene](#) (Apache)
- [Lemur & Indri](#) (CMU/Univ. of Massachusetts)
- [Terrier](#) (Glasgow)
- [MeTA](#) (University of Illinois)
- [RankLib](#) (A collection of learning-to-rank algorithms University of Massachusetts Amherst)
- [General Information Retrieval Systems](#)

NLP-related Resources

- [Statistical natural language processing and corpus-based computational linguistics: An annotated list of resources](#)
- [Stanford NLP parser](#) (Stanford University NLP group)
- [OpenNLP](#) (Apache)
- [LingPipe](#) (Java-based)
- [NLTK](#) (Python-based)

Machine Learning Toolkits

- [Weka](#) (A rich collection of machine learning algorithms, Machine Learning Group at the University of Waikato)
- [Mallet](#) (An alternative package for Weka, developed by Andrew McCallum at University of Massachusetts Amherst)
- [LibSVM](#) (A collection of SVMs, developed by Chih-Chung Chang and Chih-Jen Lin at National Taiwan University)
- [SVM-light](#) (Another collection of SVMs, developed by Thorsten Joachims at Cornell University)
- [GraphLab](#) (Large-scale machine learning package)
- [mahout](#) (Apache large-scale machine learning package)
- [Topic Models](#) (David Blei's collection of various topic models)

Plagiarism Policy

You are not allowed to copy code for programming assignments from internet or any other student. Penalty of plagiarism in programming assignments will be from one of the following depending on severity of case:

- -1 absolute from final grade
- Final grade is lowered
- F in course

Slide Credits

- Dr. ChengXiang Zhai
- Lecture Notes, Text Retrieval and Mining by Christopher Manning and Prabhakar Raghavan, Stanford University