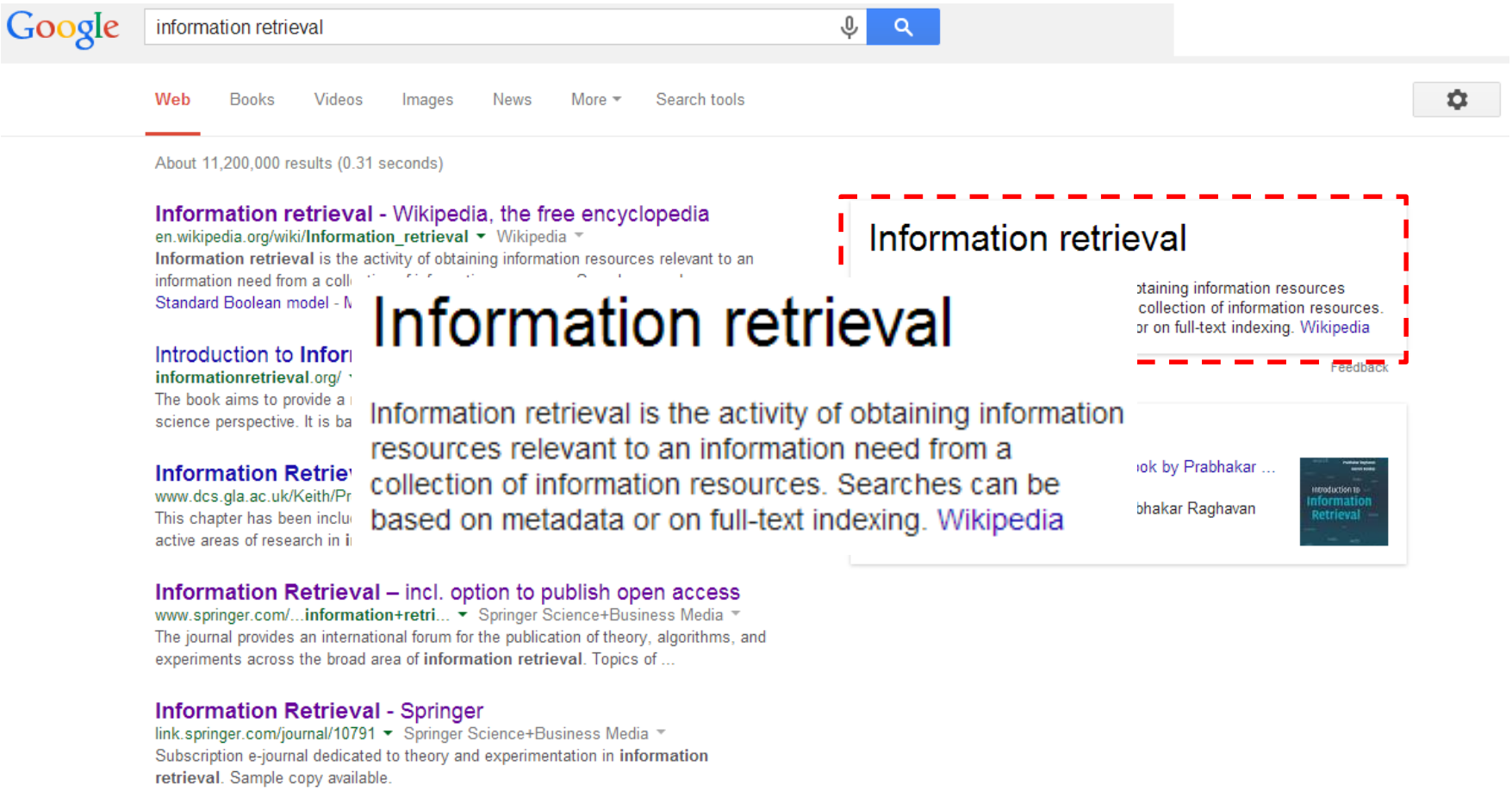
Information Retrieval

Lecture 1 Introduction

What is information retrieval?

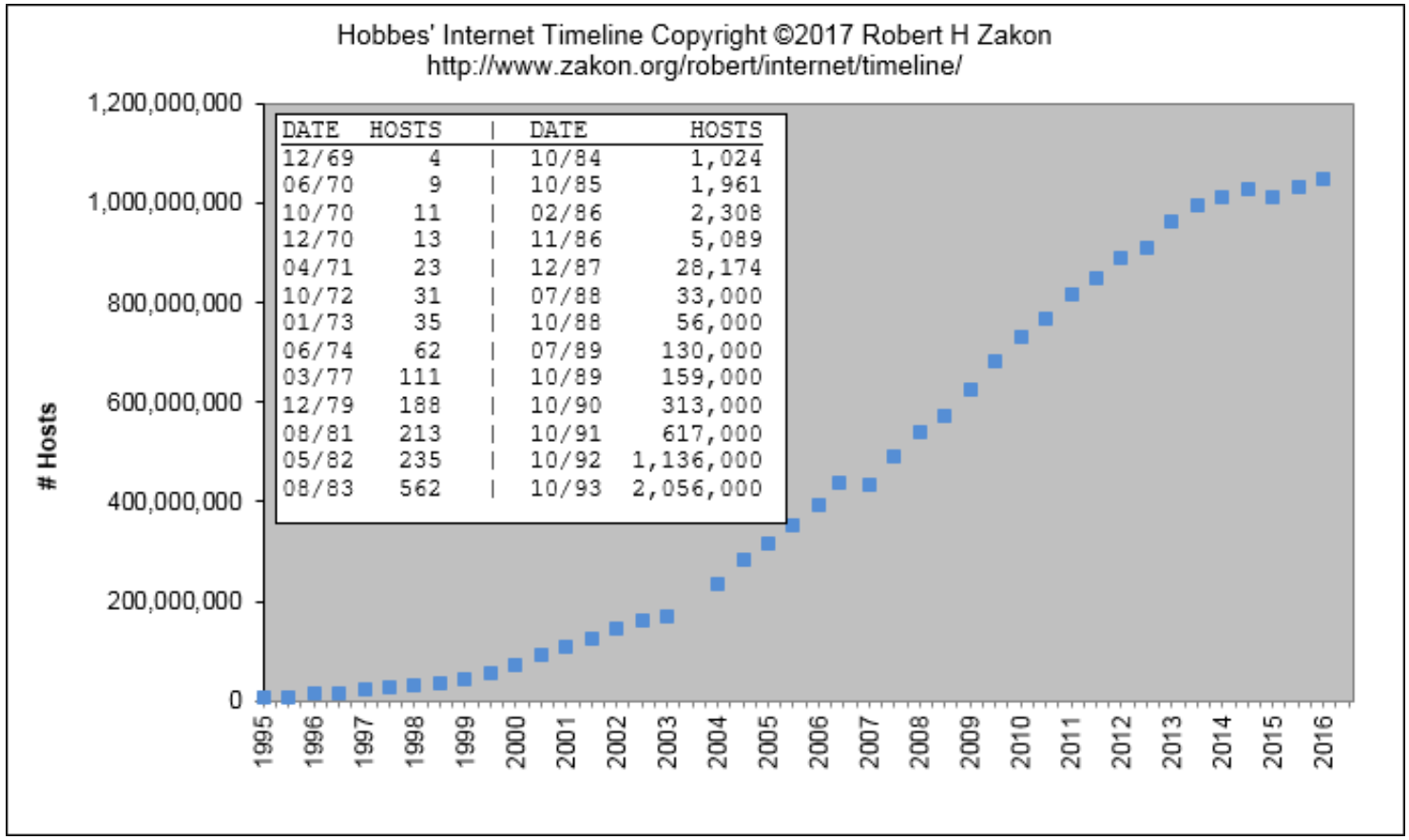


Why information retrieval

- Information overload
  - *"It refers to the <u>difficulty</u> a person can have understanding an issue and making decisions that can be caused by the presence of <u>too much</u> information."* - wiki

Why information retrieval

- Information overload

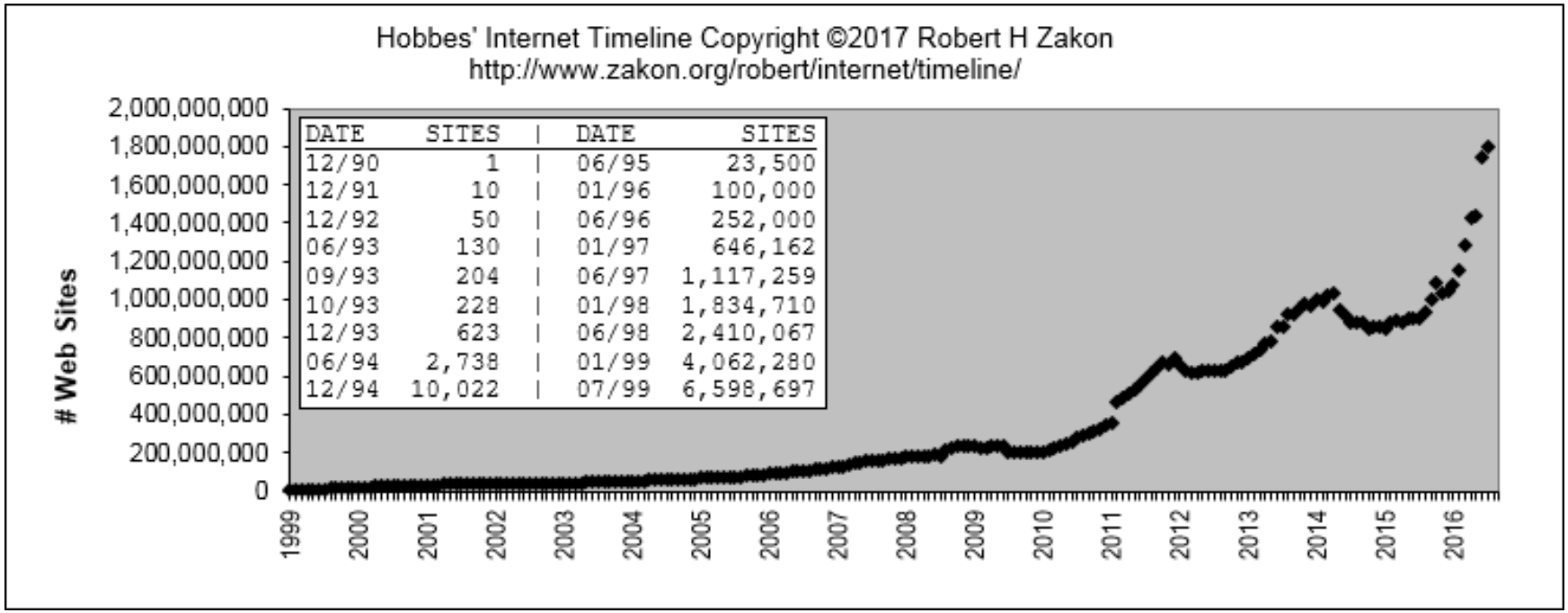


DATE	HOSTS	DATE	HOSTS
12/69	4	10/84	1,024
06/70	9	10/85	1,961
10/70	11	02/86	2,308
12/70	13	11/86	5,089
04/71	23	12/87	28,174
10/72	31	07/88	33,000
01/73	35	10/88	56,000
06/74	62	07/89	130,000
03/77	111	10/89	159,000
12/79	188	10/90	313,000
08/81	213	10/91	617,000
05/82	235	10/92	1,136,000
08/83	562	10/93	2,056,000

Figure 1: Growth of Internet

Why information retrieval

- Information overload



DATE	SITES	DATE	SITES
12/90	1	06/95	23,500
12/91	10	01/96	100,000
12/92	50	06/96	252,000
06/93	130	01/97	646,162
09/93	204	06/97	1,117,259
10/93	228	01/98	1,834,710
12/93	623	06/98	2,410,067
06/94	2,738	01/99	4,062,280
12/94	10,022	07/99	6,598,697

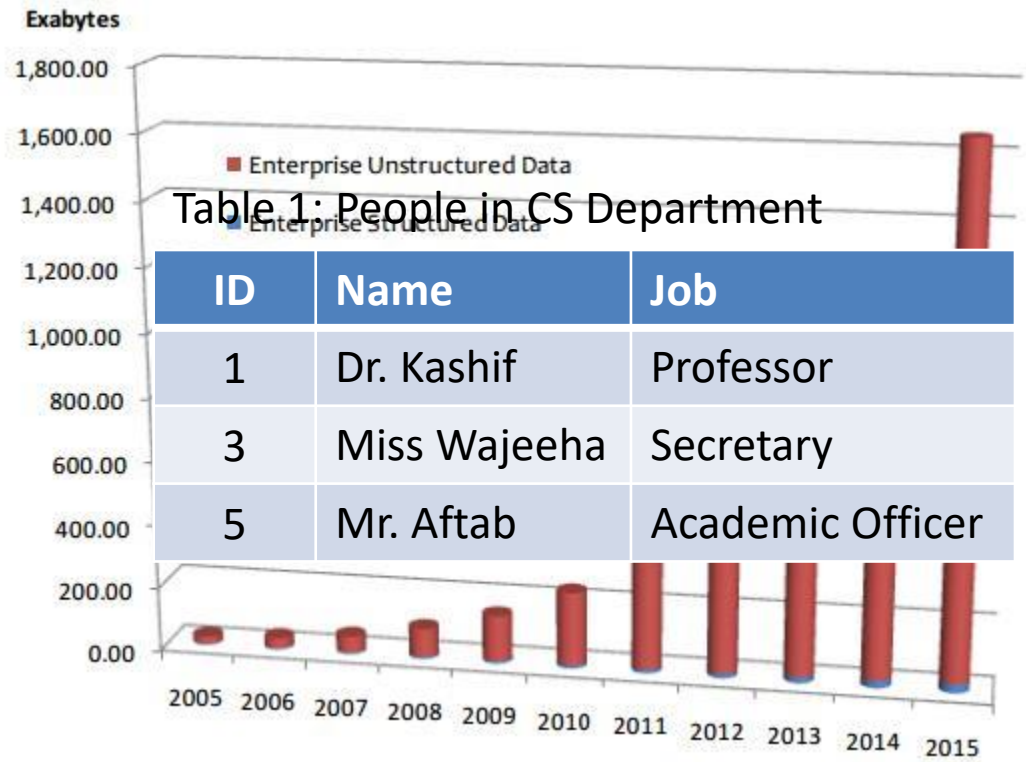
Figure 2: Growth of WWW

Why information retrieval

- Handling unstructured data
  - Structured data: database system is a good choice
  - Unst
    - Te lio, video…
    - "g *as*
      ur
    - U

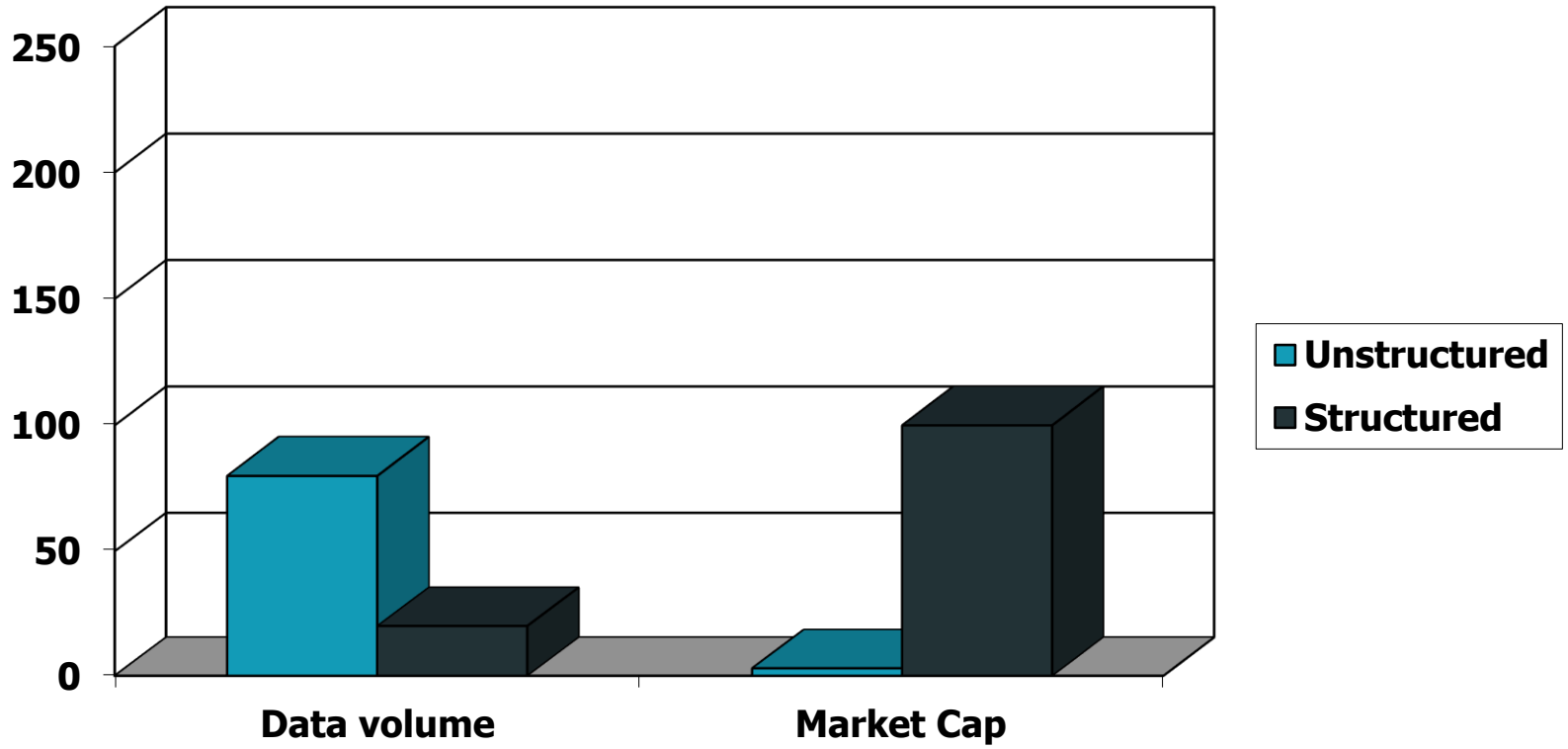
Table 1: People in CS Department

ID	Name	Job
1	Dr. Kashif	Professor
3	Miss Wajeeha	Secretary
5	Mr. Aftab	Academic Officer

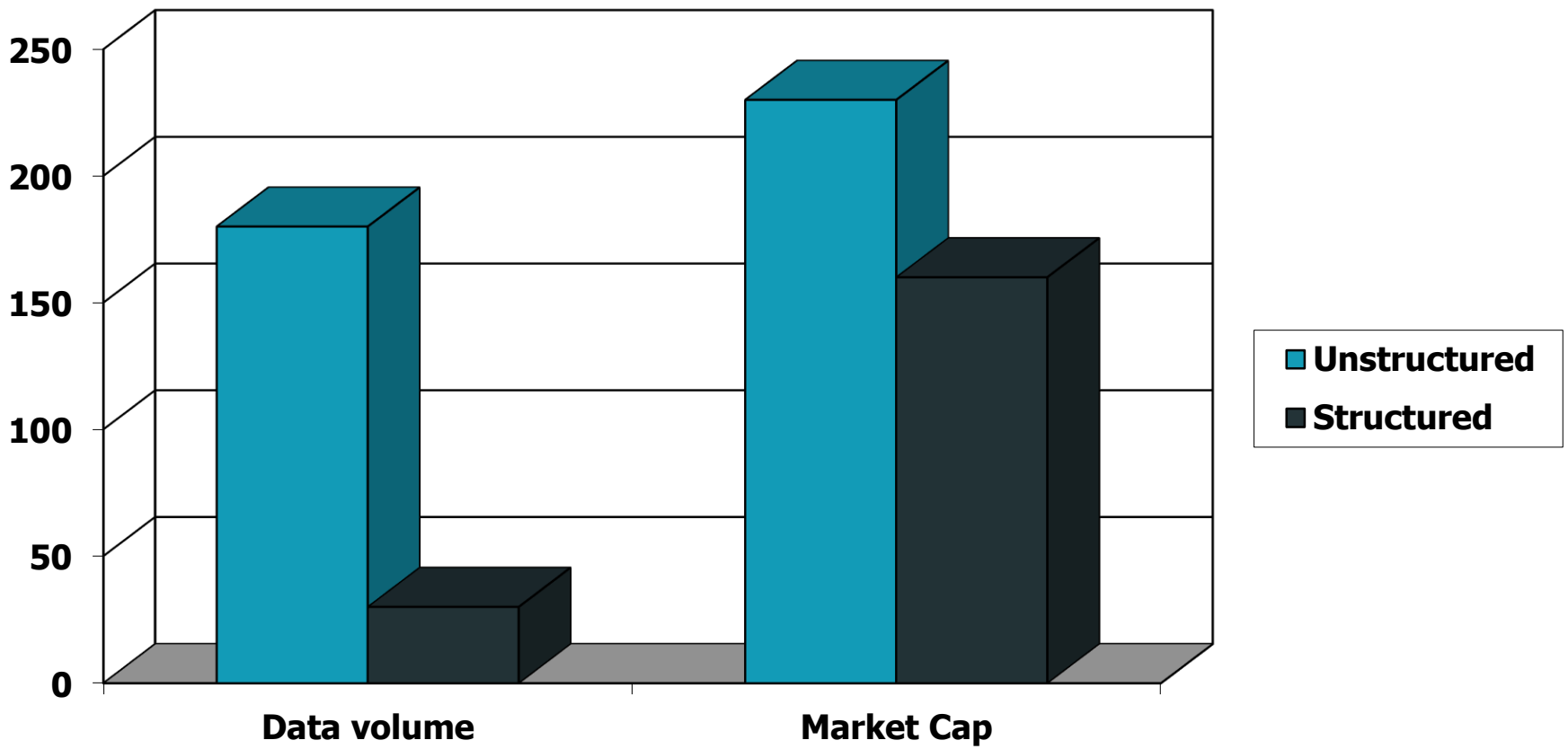


Total Enterprise Data Growth 2005-2015, IDC 2012

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

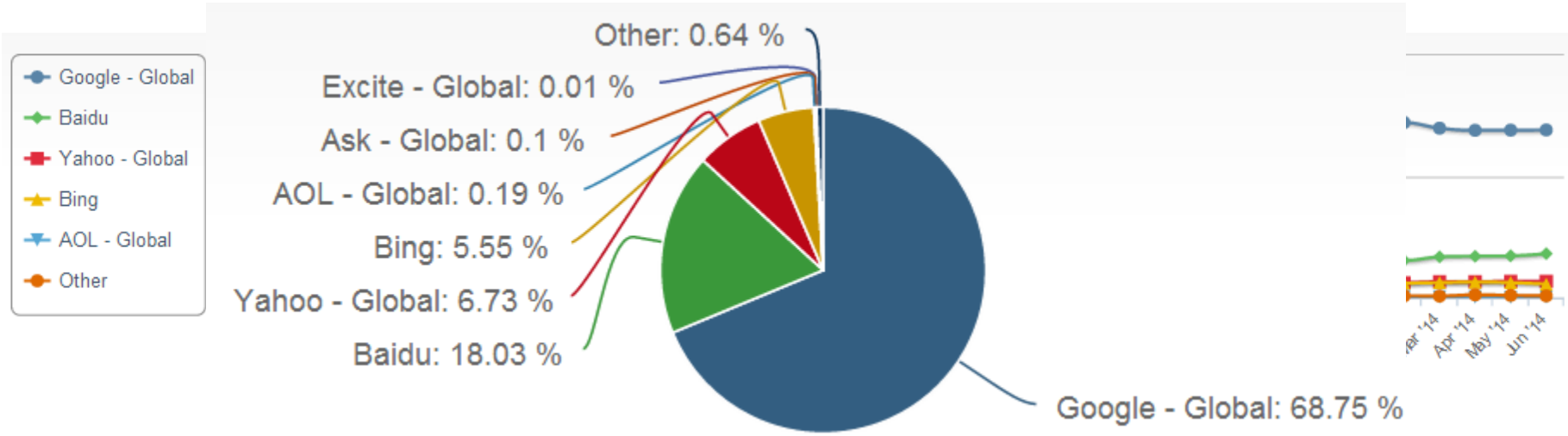
You are here!

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

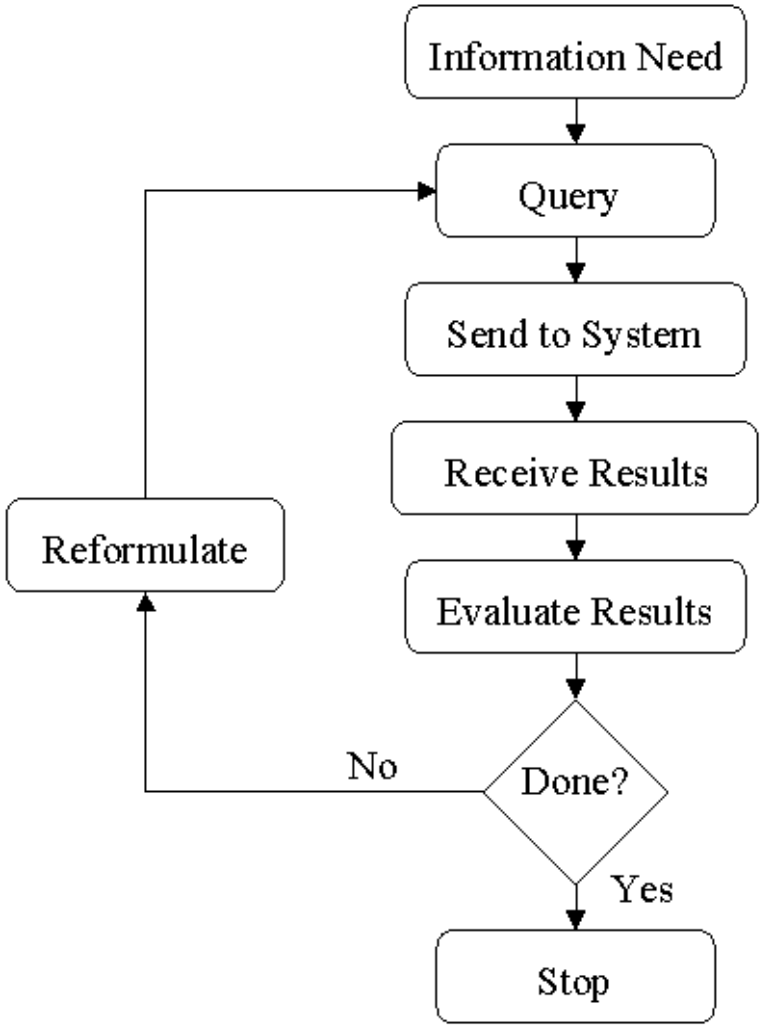
- Global search engine market
  - By http://marketshare.hitslink.com/search-engine-market-share.aspx



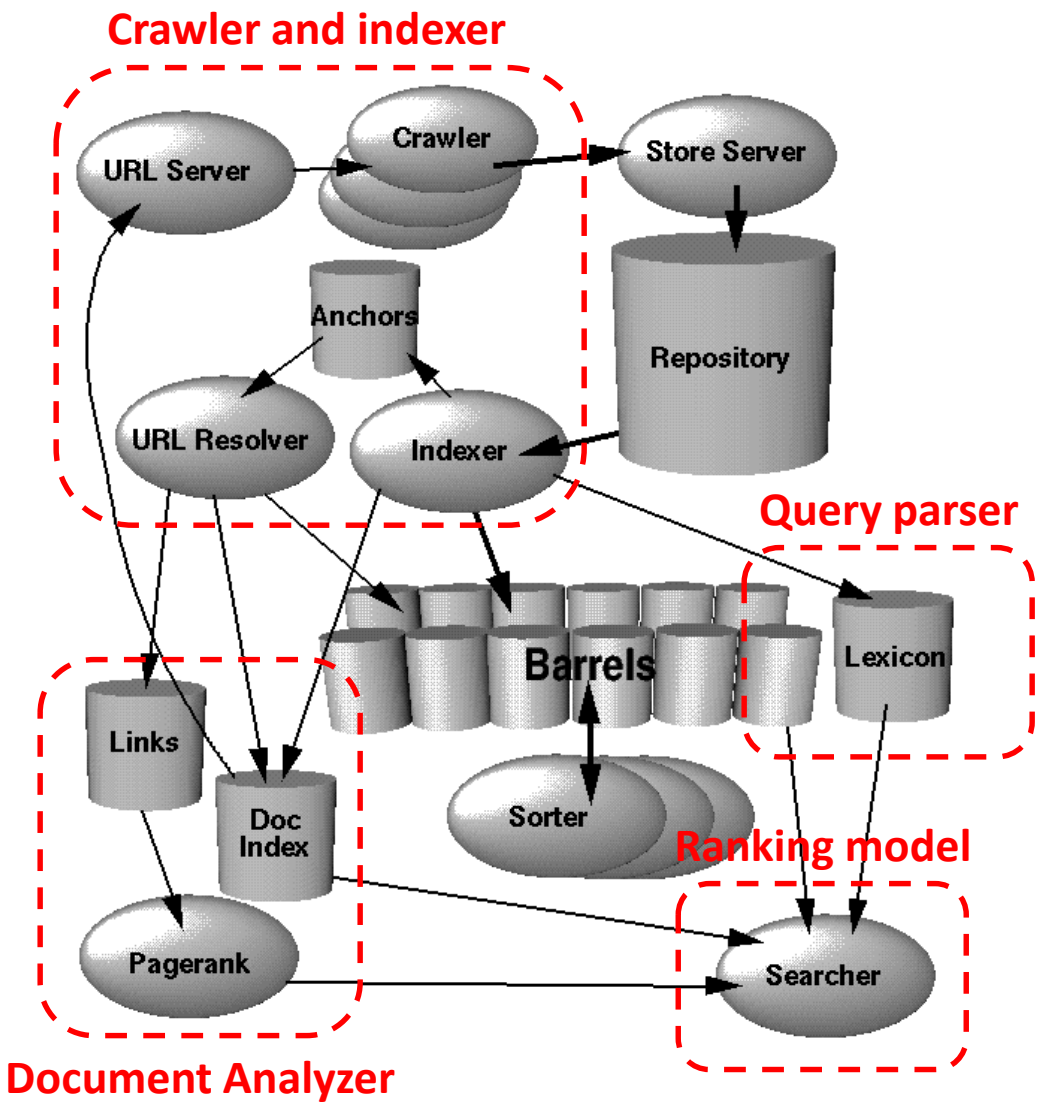
How to perform information retrieval

- Information retrieval when we did not have a computer

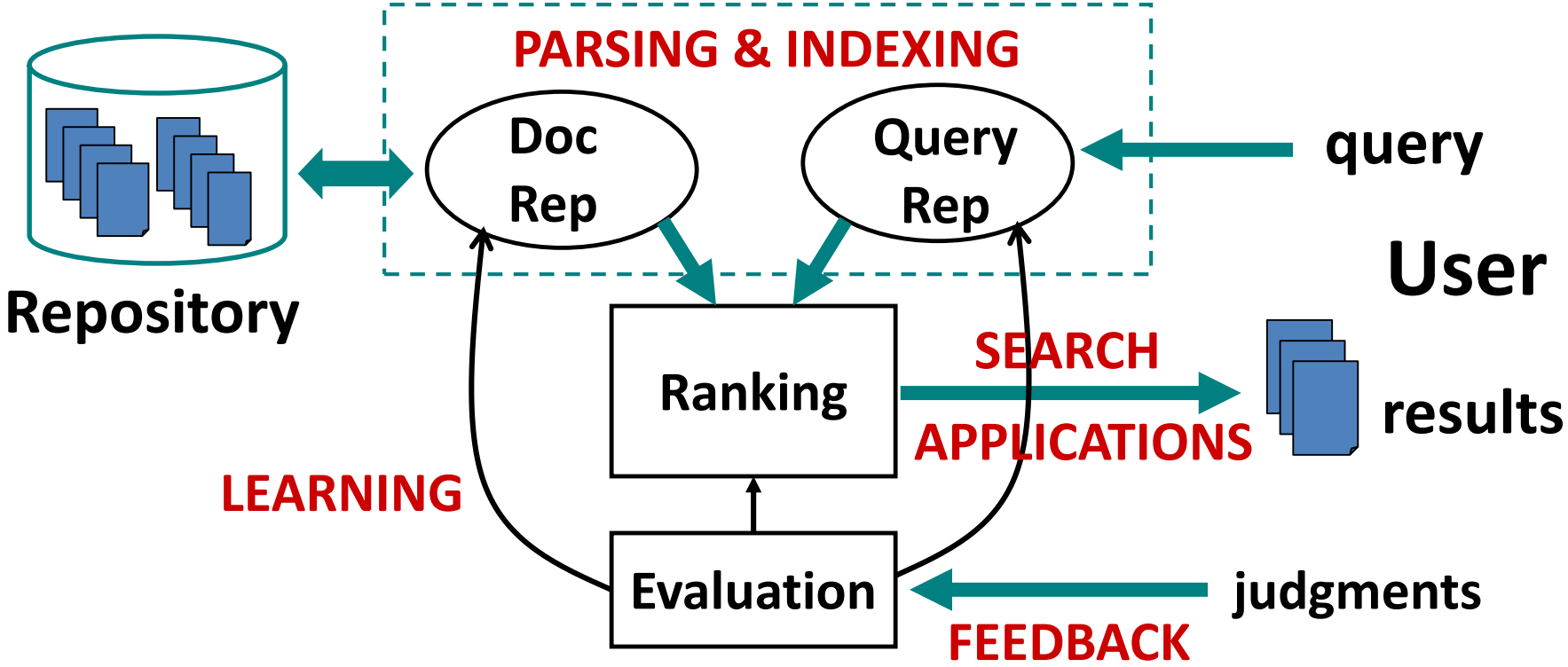
The Standard Retrieval Interaction Model



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

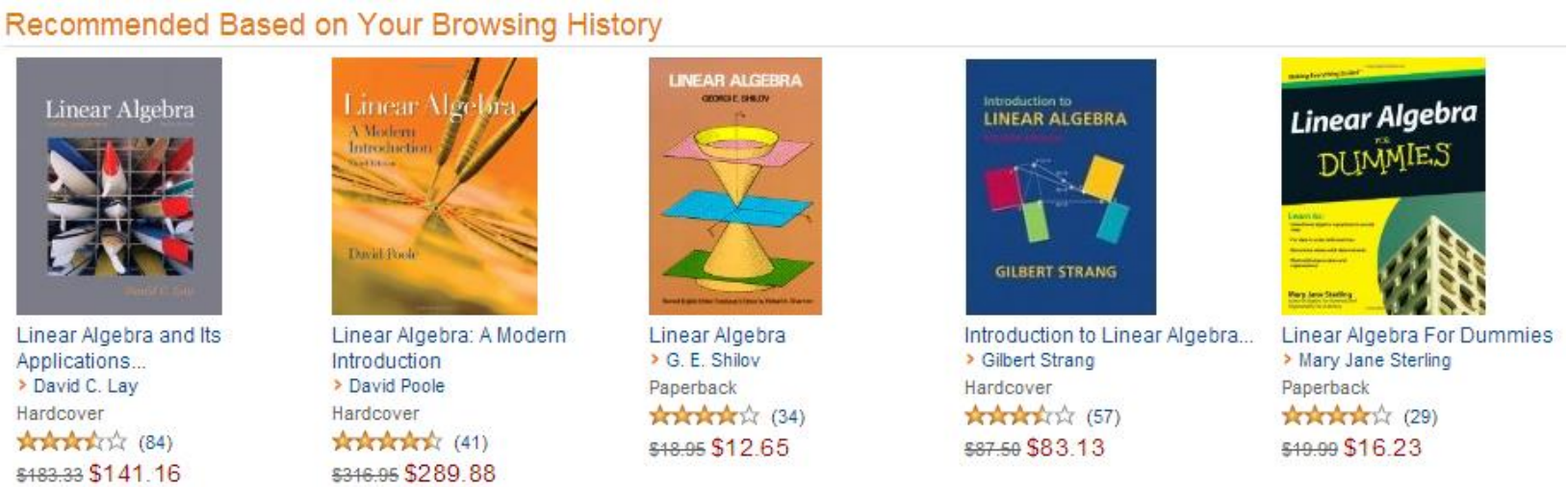
A glance of modern search engine

In modern time

A glance of modern search engine

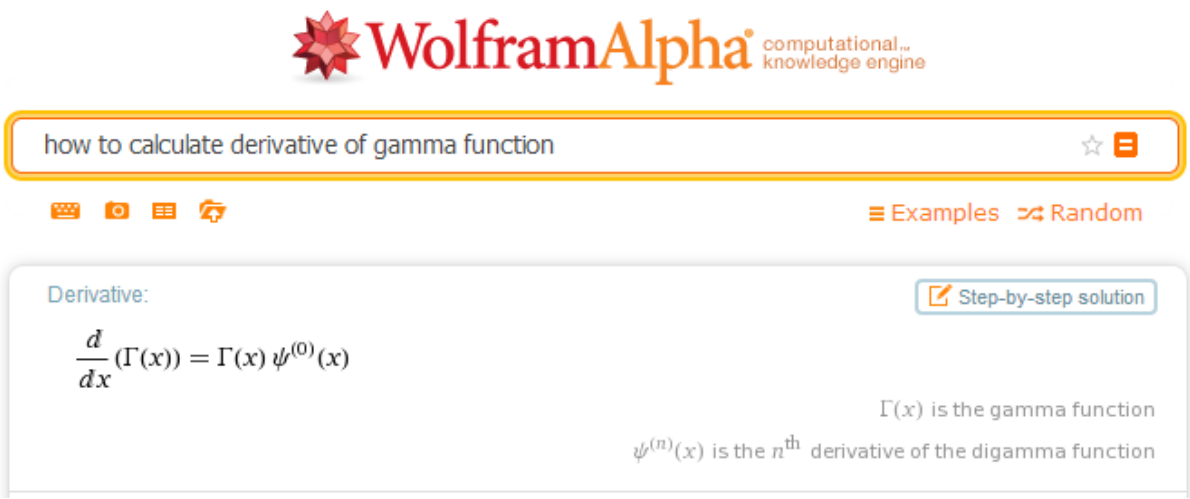
IR is not just about web search

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



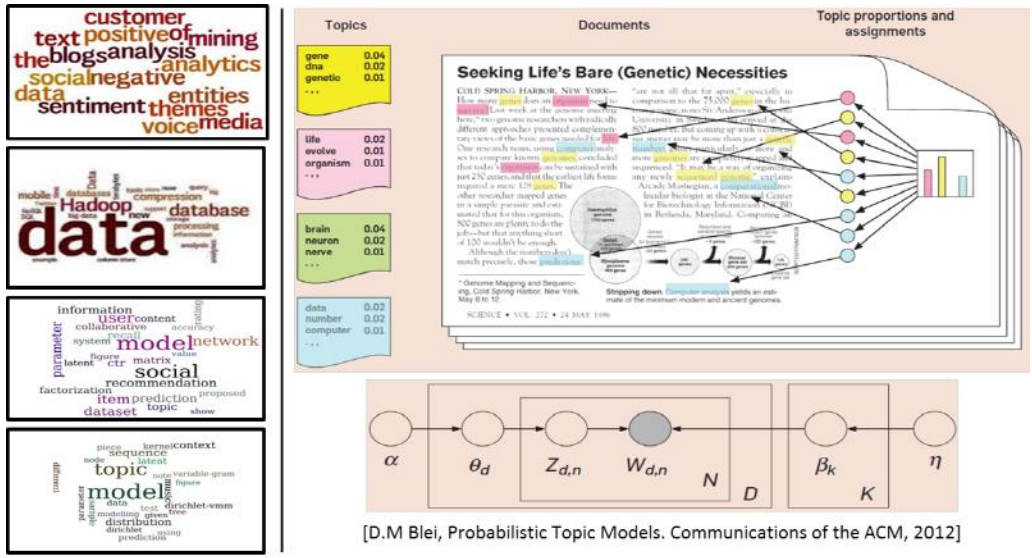
IR is not just about web search

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



IR is not just about web search

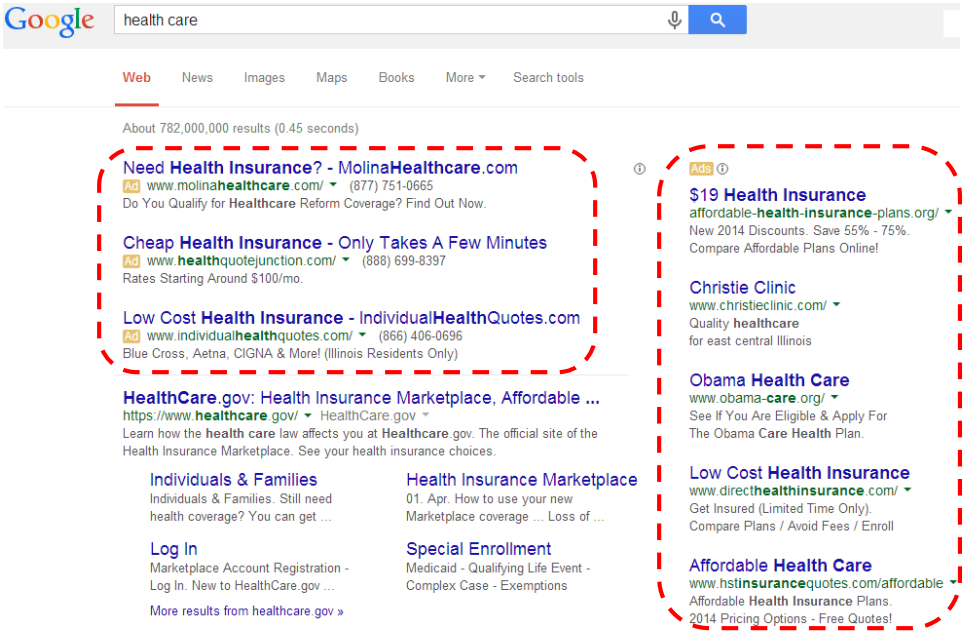
- Web search is just one important area of information retrieval, but not all
- Information retrieval also includes
  - Text mining



[D.M Blei, Probabilistic Topic Models. Communications of the ACM, 2012]

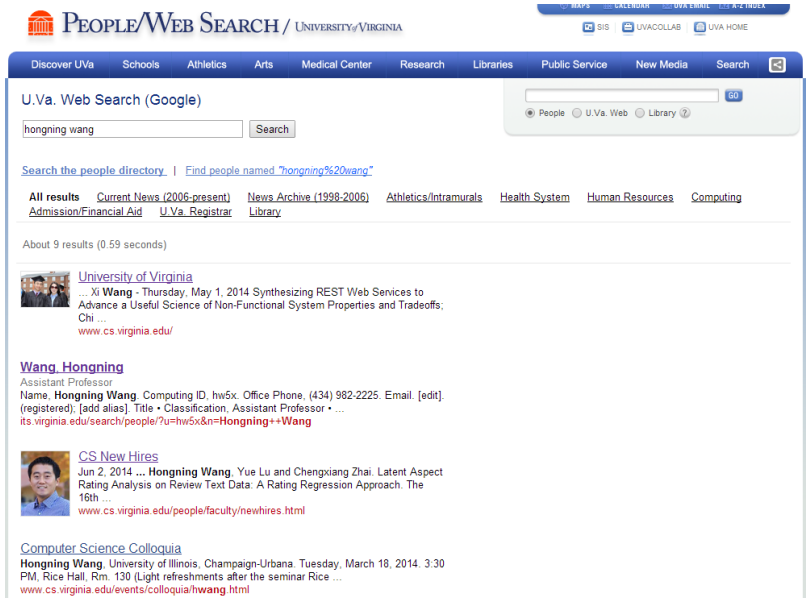
IR is not just about web search

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

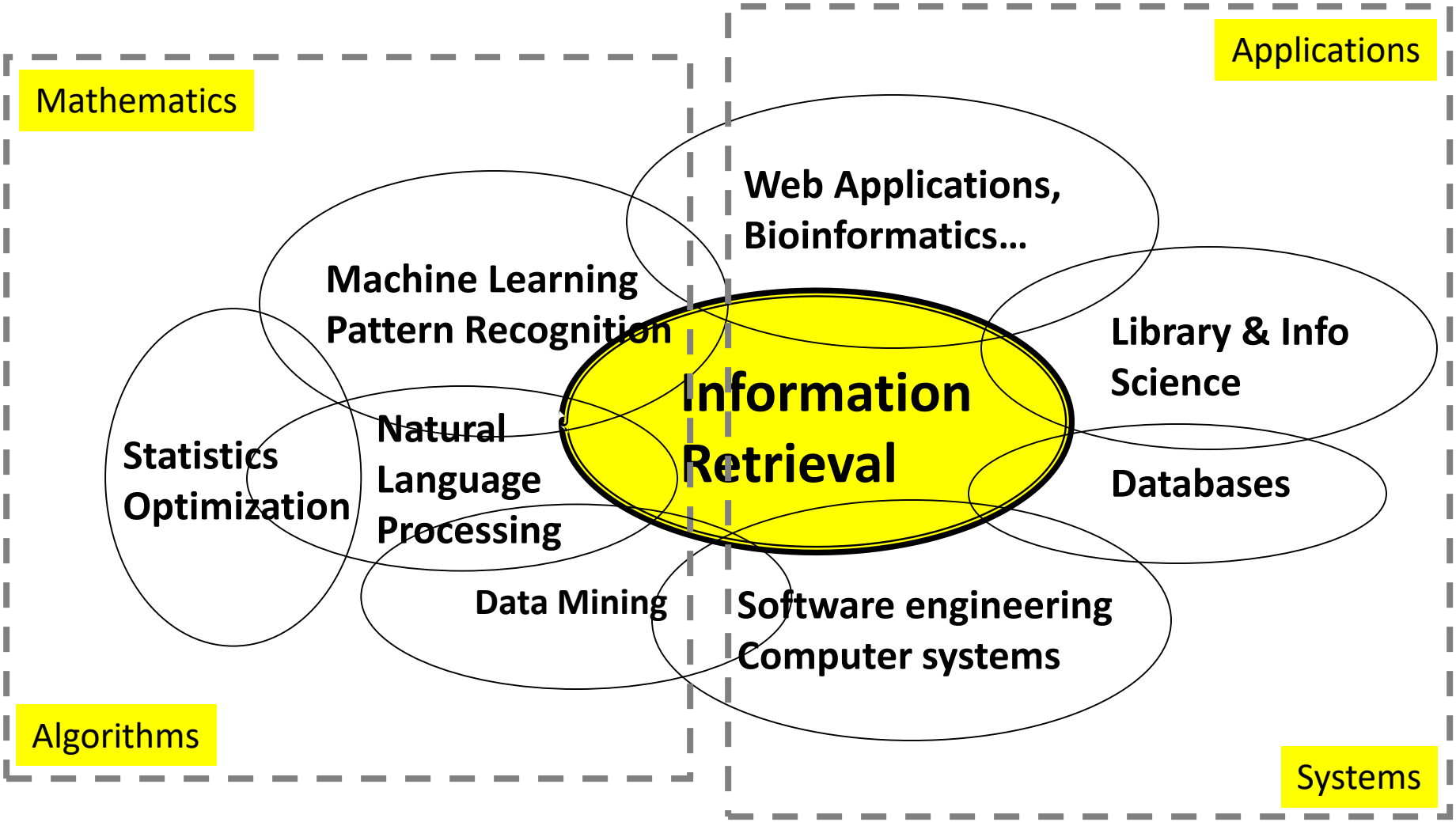


IR is not just about web search

- 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

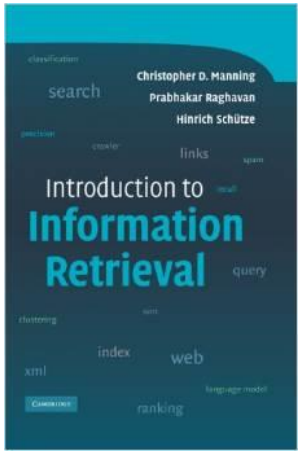
- IR => NLP
  - Larger data collections
  - Scalable/robust NLP techniques, e.g., translation models
- NLP => IR
  - Deep analysis of text documents and queries
  - Information extraction for structured IR tasks

Course Learning Objectives

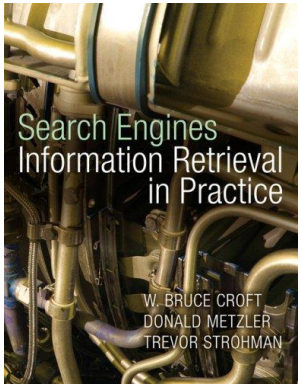
- Enable students to understand the common algorithms and techniques for information retrieval (document indexing and retrieval, query processing, etc)
- Introduce the quantitative evaluation methods for the IR systems and data mining techniques
- Enable students to implement a basic textual information 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.

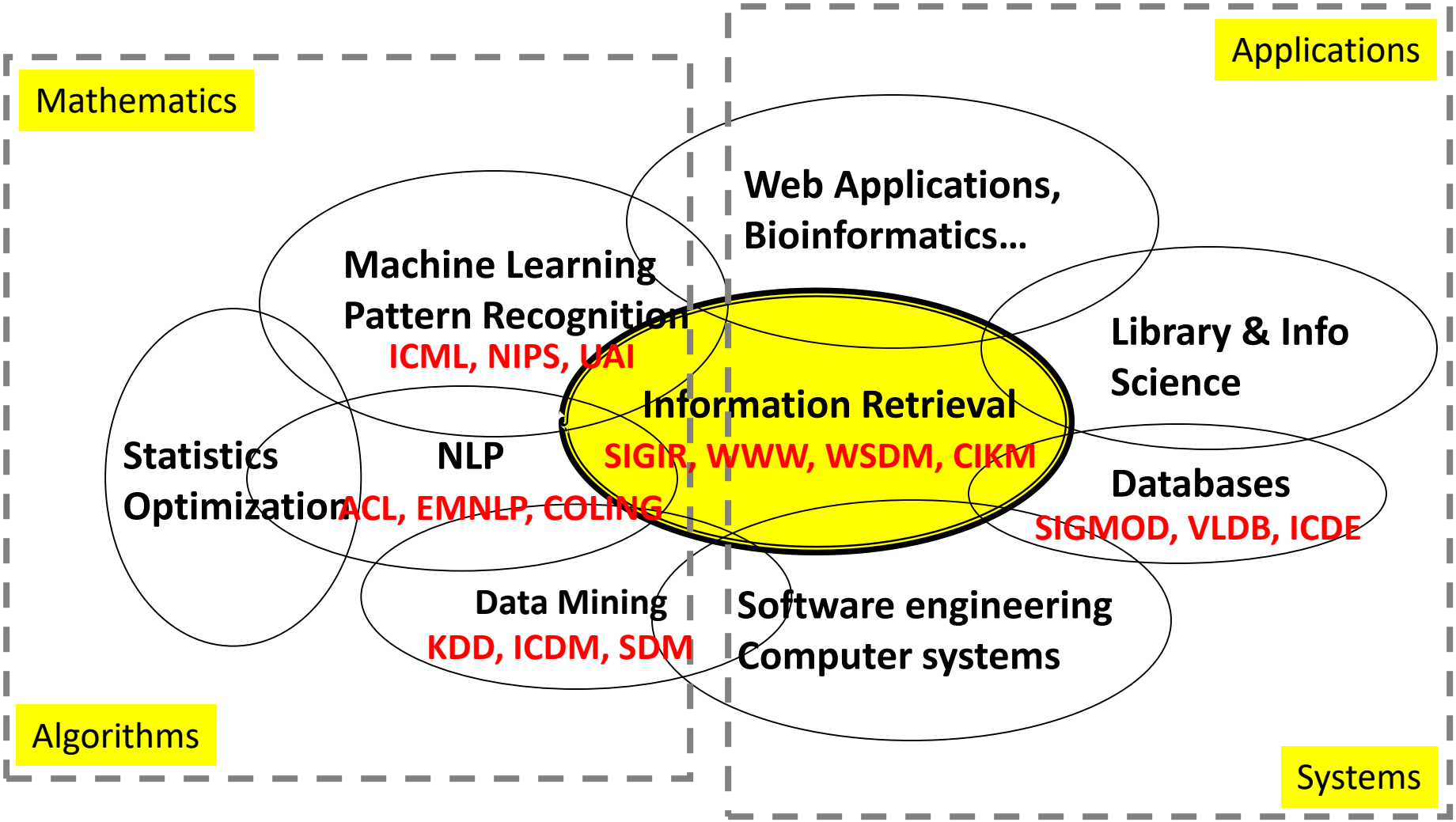


- ***Search Engines: Information Retrieval in Practice***. Bruce Croft, Donald Metzler, and Trevor Strohman, Pearson Education, 2009.

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?


Mathematics
Applications
Web Applications,
Bioinformatics…
Machine Learning
Pattern Recognition
ICML, NIPS, UAI
Library & Info
Science
Information Retrieval
SIGIR, WWW, WSDM, CIKM
Statistics
Optimization
NLP
ACL, EMNLP, COLING
Databases
SIGMOD, VLDB, ICDE
Data Mining
KDD, ICDM, SDM
Software engineering
Computer systems
Algorithms
Systems

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 (Jave-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