IS2140 Information Storage and Retrieval

Unit 1: Introduction and Overview

Daqing He
School of Computing and Information
University of Pittsburgh

August 27, 2018

Instructional Staff

- Instructor: Daqing He
  - *Office: IS 618*
  - *Email: dah44@pitt.edu*
  - *Office Hour: by appointment*
- Teaching Assistants: Danchen Zhang and Ning Zou
  - *Office: 617 and 707 Information Science Building*
  - *Email: {daz45, niz19}@pitt.edu*
  - *Office hours: Wednesdays 10-noon (Danchen@707) Thursdays 3-5pm (Ning@617)*
- IRIS group is at http://crystal.exp.sis.pitt.edu:8080/iris/

Breaks

- 12:00pm-2:50pm
  - *10 minutes break after each 50 minutes*

Agenda

- Introduction to Information Retrieval
  - *What is information, and information retrieval*
  - *Types of Retrievals*
- Course Overview
  - *Why study information retrieval?*
  - *Course Materials*
  - *Grading and Exams*

Class Goals

- After this class, you should be able to
  - *know what data, information, and knowledge is*
  - *know the basic concepts of retrieval*
  - *restate the expectations and requirements of the course*
  - *make decision on whether attending the course*

Google

information storage and retrieval

All Images Videos News Shopping More Settings Tools

About 37,300,000 results (0.69 seconds)

Scholarly articles for **information storage and retrieval**

Information storage and retrieval - Korfhage - Cited by 995
Information storage and retrieval - Ikegami - Cited by 113
Information storage and retrieval - Lipetz - Cited by 43

What is IS
whatis.techta
An **informatio**
facilitates the
P2P) network

Informatic
www.springe
Chapter 1 plac
introduces ne
implemented.
realistic, allow

Informatic
https://www.
Read the lates
platform of pe

Informatic
https://encyc
Looking for in
retrieval. the s
displayed on r
retrieval.

See results about

Information storage and retrieval systems (Book by Geral...
Originally published: 2000
Author: Gerald Kowalski

- How can search engines like Google know what I want?
- Is there anything I can do to improve the results?
- Is this different to search in a database?
- How can it find information so fast?
- What exactly is information retrieval?
- What can I do in information retrieval to improve it?

People also ask

What is storage and retrieval?
What is data storage and retrieval?
What is the information storage?
What is an information retrieval system?

Feedback

What is Information?

Information Is ...

- Oxford English Dictionary
  - *Informing, telling; thing told, knowledge, items of knowledge, news*
- Random House College Dictionary
  - *Knowledge communicated or received concerning a particular fact or circumstance; news.*
- Cookie Monster
  - *News or facts about something*

Some Characters of Information

- be something
  - although the exact nature (substance, energy, or abstract concept) is not clear;
- be "new"
  - repetition of previously received messages is not informative
- be "true"
  - false or counterfactual information is "misinformation"
- be "about" something

Robert M. Losee (1997) A Discipline Independent Definition of Information. Journal of the American Society for information Science, 48(3) 254-269.

Information in Communication

- Information science is characterized by "the deliberate (purposeful) structure of the message by the sender in order to affect the image structure of the recipient"
- Information = "the structure of any text which is capable of changing the image structure of a recipient"
- Communication = transmission of information

Nicholas J. Belkin and Stephen E. Robertson. (1976) Information Science and the Phenomenon of Information. Journal of the American Society for Information Science. 27(4) , 197-204.

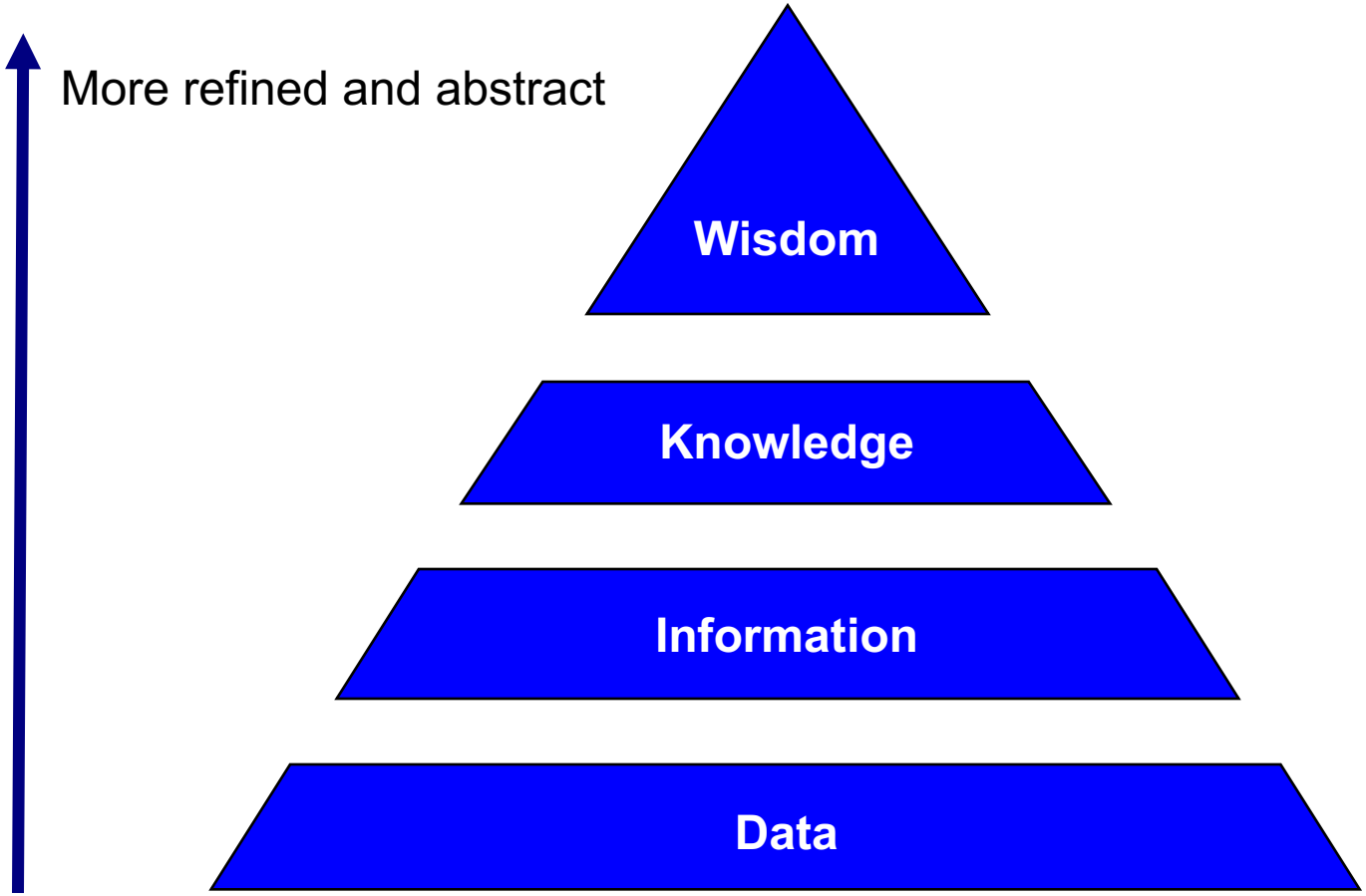
Representation of Information

- Information communication needs a representation of information
  - *Such as a set of words, a set of audio sounds, a set of video clips or even a set of numbers representing associations*
- Each representation has its advantages and limitations
  - *What is represented?*
    - E.g., written words => the content
  - *What is left out?*
    - E.g., written words => Sound and gestures fade away
- So our assumptions on representation
  - *Anything not in representation can be ignored or too difficult to represent*

Context of Information

The context where the information is generated and communicated

DIKW Hierarchy



http://www-personal.si.umich.edu/~nsharma/dikw_origin.htm

DIKW Hierarchy

- Data
  - Scientific evidence gathered by experiments, observations, and other investigations
  - Neither truth nor reality
- Information
  - Data organized and processed for a particular communication
- Knowledge
  - "justified true belief"
  - Information that can be acted on
- Wisdom
  - Distilled and integrated knowledge
  - Demonstrative of high-level "understanding"

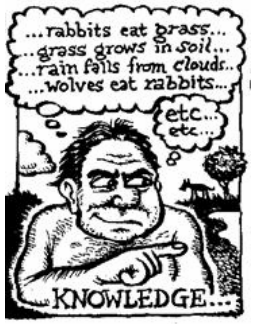


Image originally published in the December 1982 issue of THE FUTURIST.

A (Facetious) Example

- Data
  - $2.95, $2.99, $3.09, $3.19 …
- Information
  - The gas prices in the past three months were $2.95, $2.99, $3.09, $3.19 …
- Knowledge
  - If the gas price is increasing, it will cost more to drive a car
- Wisdom
  - Under current gas price, do not buy a Hummer

Current Status of Information

- Lots of information is in analog format
  - *On more traditional media, like paper, films, and other physical objects*
- But much much more information is in digital format
- Lots of information is scatter around
  - *In our draws, cabinets, etc.*
- But much more information is on the Web

How Much Info is Recorded in 2002

- Magnetic 5,187,130 terabytes
  - *PC disk drives, departmental servers, camcorder tape, enterprise servers*
- Film 420,000
  - *Photograph, X-rays, cinema*
- Paper 1,630
  - *Office documents, newspapers, periodicals, books*
- Optical 103
  - *Music CDs, DVDs, Data CDs*

Grand Total ~ 5,609,121 terabytes

In comparison:

Library of Congress: about 15 Terabytes

DIALOG: about 9.5 Terabytes

Source: Lyman and Varian, UC Berkeley

Information Consumption

Total time American households spend reading, watching TV or listening to music:

1992: 3,324 hours

2000: 3,380 hours

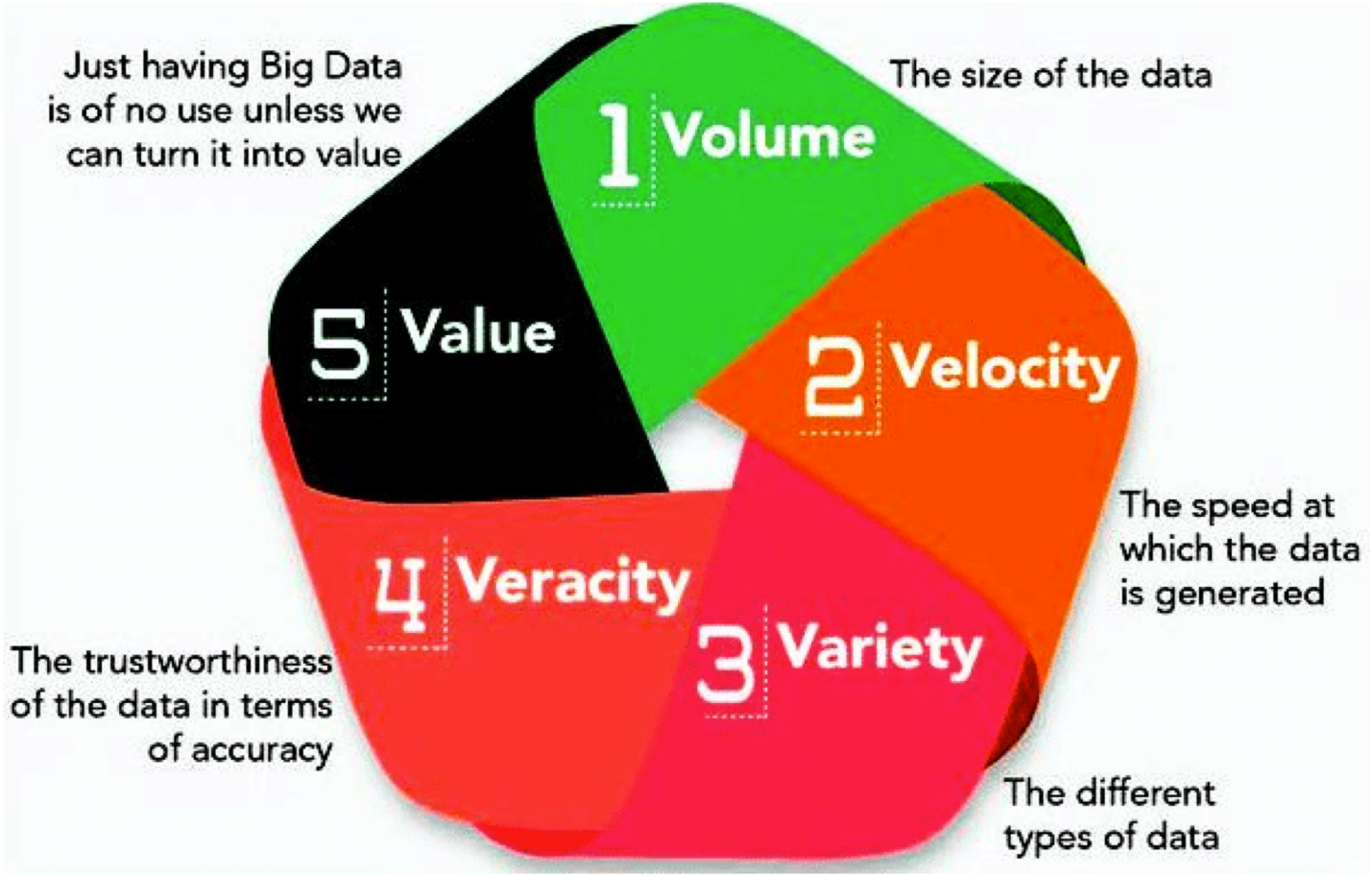
Bits consumed: 3,344,783 megabytes

or ~ 3 Terabytes

(Bits created: ~5,609,121 Terabytes)

Source: Lyman and Varian, UC Berkeley

Major Problems in Information Processing and Management?



Structured and Unstructured Data

- **Structured** Data:
  - *Data stored in databases is known as **structured** data because it is represented in a strict format.*
  - *The DBMS then checks to ensure that all data follows the structures and constraints specified in the schema.*
- **Unstructured** Data:
  - *there is very limited indication of the type of data.*
  - *A typical example would be a text document that contains information embedded within it.Web pages in HTML that contain some data are considered as unstructured data.*

Structured Data: Example

Table name: PAINTER

PTR_NUM	PTR_LASTNAME	PTR_FIRSTNAME	PTR_AREACODE	PTR_PHONE
123	Ross	Georgette	901	885-4567
126	Itero	Julio	901	346-1112
127	Geoff	George	615	221-4456

Table name: PAINTING **Database name: Ch06_Artist**

PTNG_NUM	PTNG_TITLE	PTNG_PRICE	PTR_NUM	GAL_NUM
1338	Dawn Thunder	$245.50	123	5
1339	A Faded Rose	$6,723.00	123	
1340	The Founders	$567.99	126	6
1341	Hasty Pudding Exit	$145.50	123	
1342	Plastic Paradise	$8,328.99	126	6
1343	Roamin'	$785.00	127	6
1344	Wild Waters	$999.00	127	5
1345	Stuff 'n Such 'n Some	$9,800.00	123	5

Table name: GALLERY

GAL_NUM	GAL_OWNER	GAL_AREACODE	GAL_PHONE	GAL_RATE
5	L. R. Gilliam	901	123-4456	0.37
6	G. G. Waters	405	353-2243	0.45

- Structured data : data in "tables"

Typically allows numerical range and exact match (for text) queries, e.g.,*PTNG_PRICE < $60000 AND GAL_OWNER = L.R.Gilliam*.

Unstructured Data: Example

The New York Times

January 4, 2013

New Rules for the New Year

By BILL MAHER

2012: I call it the year in "meh." Not the worst we've ever experienced, but nothing particularly great to say about it either. Like being a socialite, but in Tampa.

I am looking forward to 2013, however, because I love the odd-numbered years — they're the ones without congressional elections, Olympics, World Cups or weird extra days tacked onto the calendar by so-called scientists. Odd-numbered years are chill. They're the 3 p.m. of years — that small sliver of time when lunch is digested and it's too early to think about dinner and you stand at least a fighting chance of getting something done.

In that spirit, here are the New Rules for the new year:

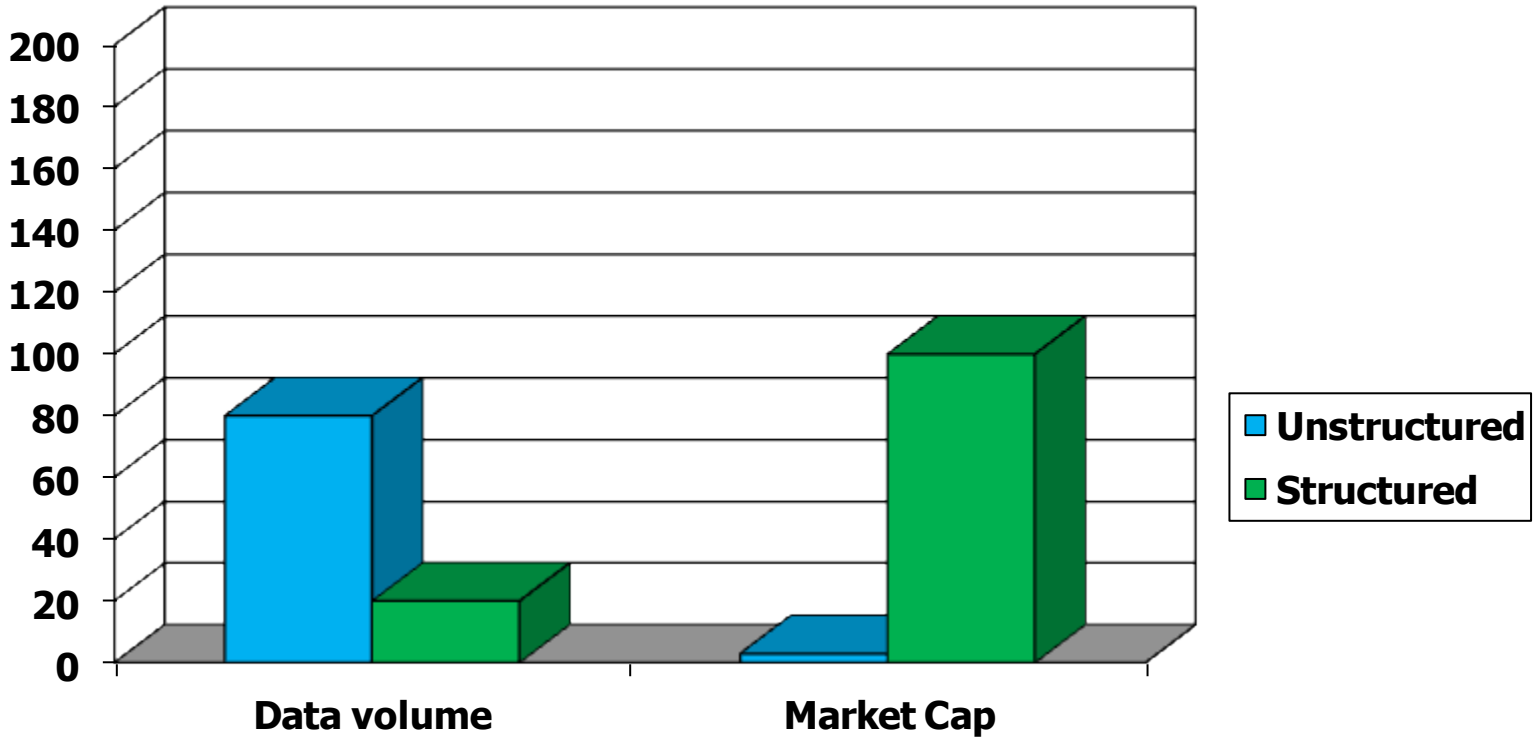
NEW RULE Now that their end-of-the-world prophecy has proved to be complete baloney, the Mayans must be given a job predicting election results for Fox News.

NEW RULE Sometime during the 2013 awards show season, "Gangnam Style" must be given an award for the shortest amount of time between my finding out what something is to my being completely sick of it. Besting the time of 7 hours, 12 minutes, set by "The Macarena" in 1996.

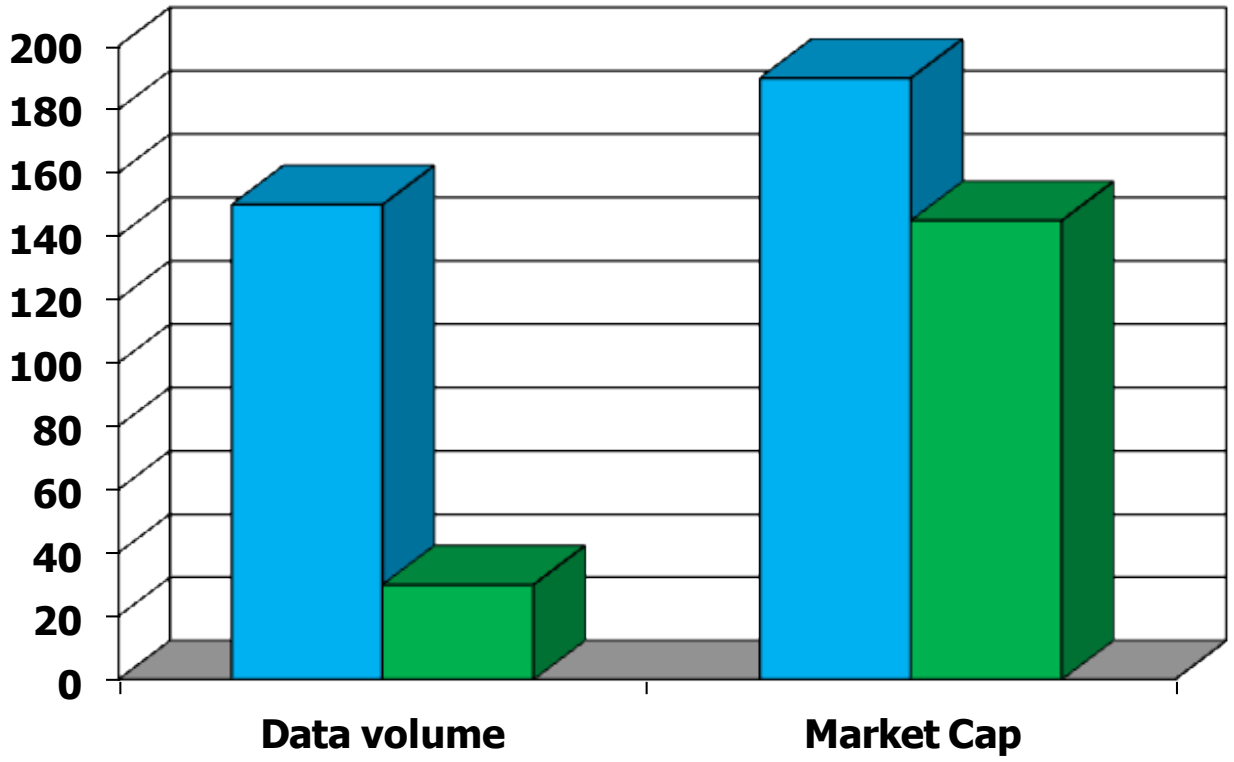
NEW RULE Congress must make it a tradition to drive off the fiscal cliff every year. And I mean really off the cliff, like Toonces the cat drove that car. This way Republicans can learn that lower military spending won't lead to China invading. And Democrats can learn that no one cares what the Commerce Department does anyway.

NEW RULE No more mixing politics with pizza. The filthy rich founder of Papa John's, John Schnatter, said he'd cut his employees' hours to avoid the costs of Obamacare. This is where I'd normally suggest boycotting Papa John's, but that's like telling people to boycott sadness. Nobody eats Papa John's because they like it. They eat it because Domino's won't deliver to crack houses.

Unstructured vs. Structured Data in 1996

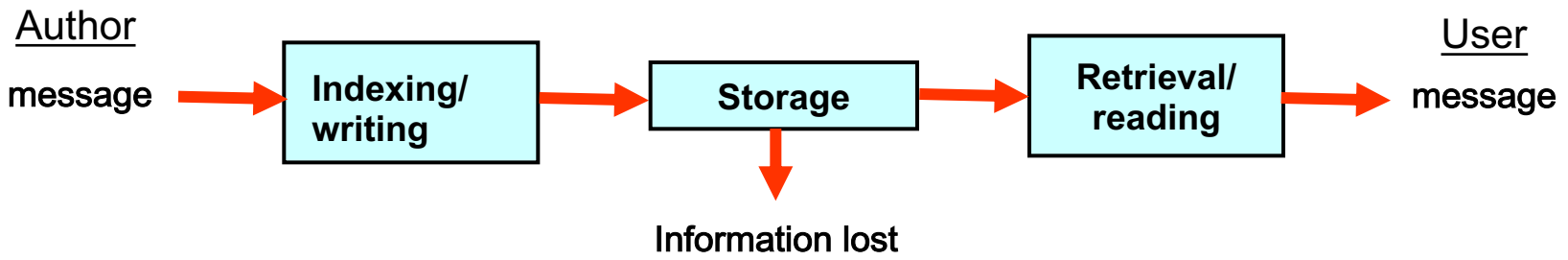


Unstructured vs. Structured Data in 2009



Retrieval is

- "Fetch something" that has been stored
- Recover a stored stated of knowledge
- Search through stored messages to find some messages relevant to the task at hand



Information Retrieval

- Information retrieval is a field concerned with the structure, analysis, organization, storage, searching, and retrieval of information

Salton, G. (1968). Automatic information organization and retrieval. New York:McGraw-Hill.

- Information retrieval is a problem-oriented discipline, concerned with the problem of the effective and efficient transfer of desired information between human generator and human user

Nicholas J. Belkin. (1980) Anomalous States of Knowledge as a Basis for information Retrieval. Canadian Journal of Information Science. 5 , 133-143.

Scopes of IR

- Early days, IR is focused on
  - *Indexing text and searching for relevant documents in a collection*
- Nowadays, IR includes
  - *Modeling search process, web search, text classification / clustering, system architecture, user interfaces, data visualization, adaptive retrieval / personalization, multiple languages*

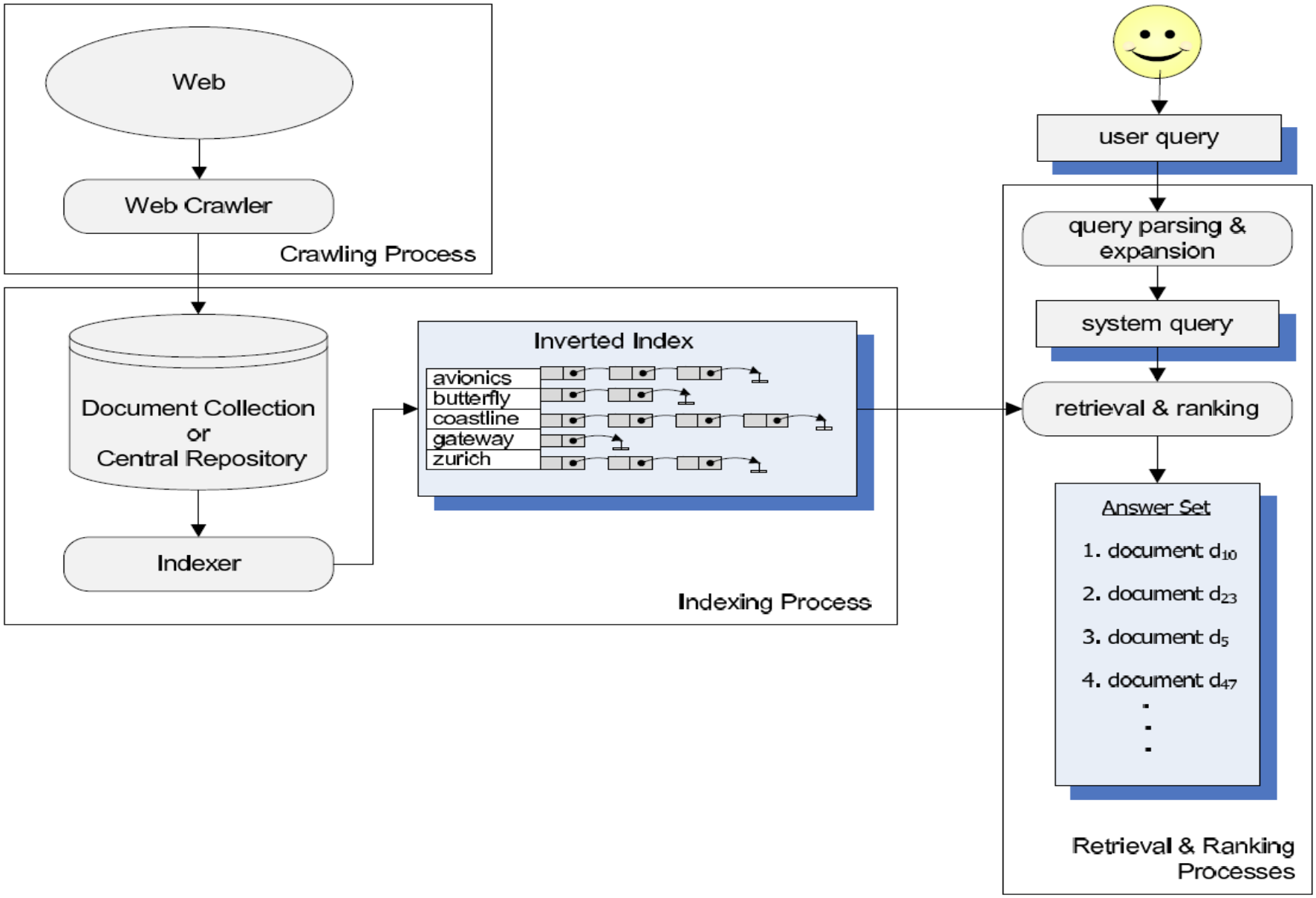
A System Oriented View of IR

Query

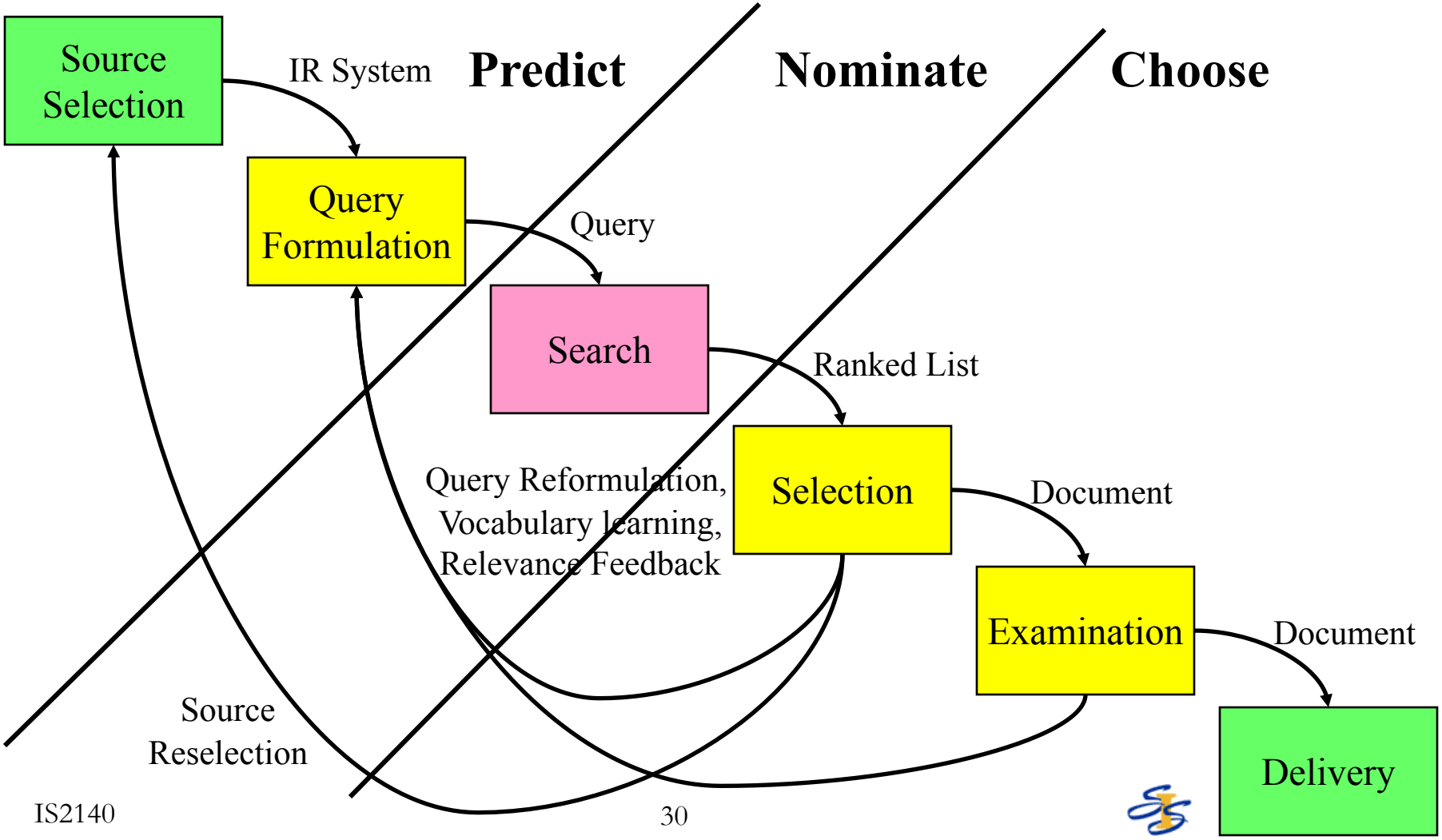
Search Engine

Ranked List

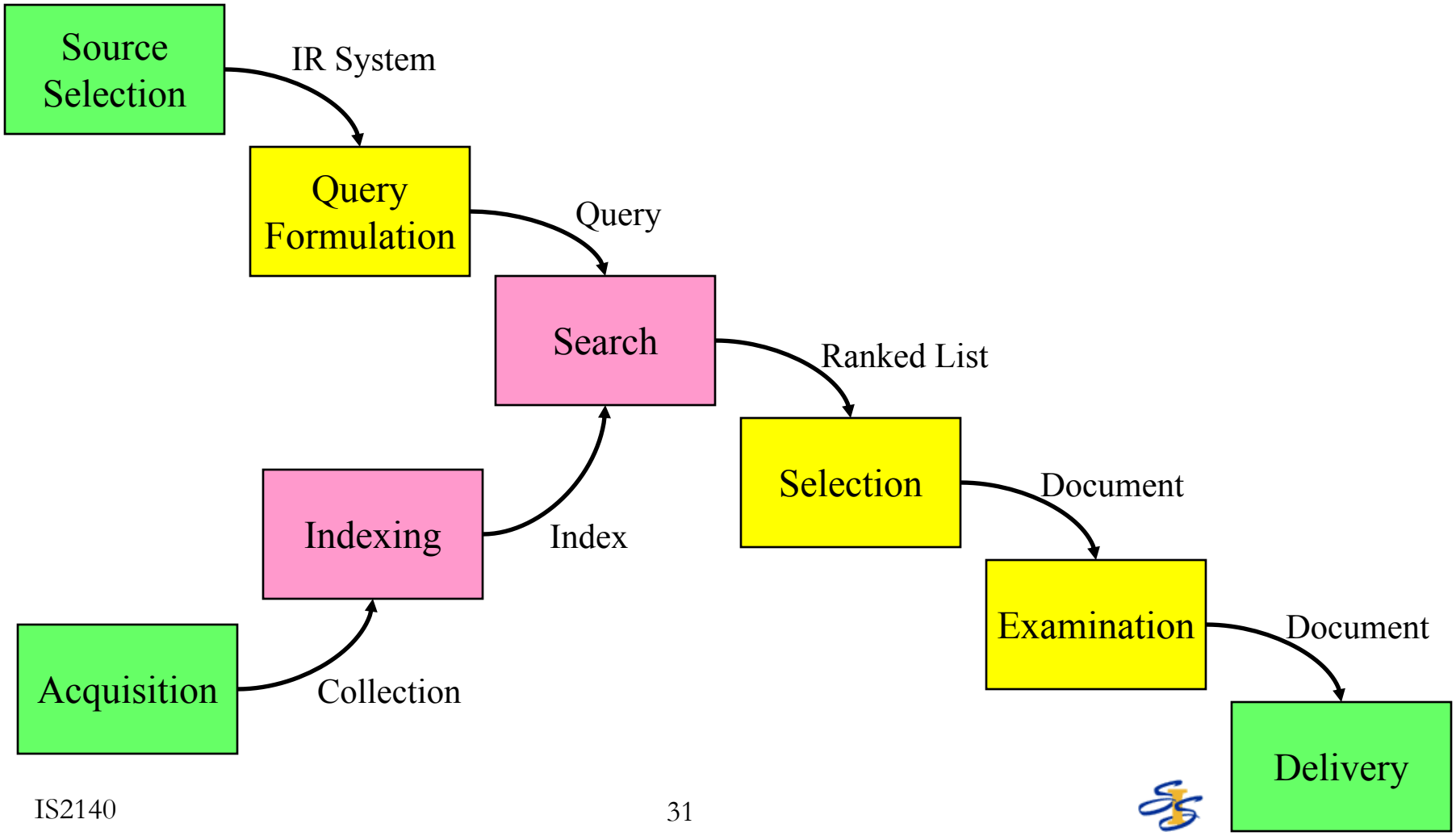
Whole View of System Oriented IR



Information Retrieval Cycle



Supporting the Search Process



Types of Information Needs

- Short term information needs
  - *"Temporary need", e.g., info about hotels at NYC*
  - *Information source is relatively static*
  - *User "pulls" information*
  - *Application example: library search, Web search*

- Long-term information need (Filtering)
  - *"Stable need", e.g., new data mining algorithms*
  - *Information source is dynamic*
  - *System "pushes" information to user*
  - *Applications: news filter, spam filter, email classifier*

Searches with short term needs

- Ad hoc retrieval: find documents "about this"
  - *E.g., Look for information about green energy projects*
- Comprehensive exploration
  - *E.g., List all available drugs with names similar to "claritin"*
- Null existence search
  - *E.g., a search to proof that no drug is called "claritin" or with a similar name*
- Known item search
  - *E.g., Find the School of Information Sciences web page*
  - *E.g., Find the price of "online retrieval" book in amazon.com*
- Question answering
  - *"factoid": who was the 42nd president of USA?*
  - *"list": which countries are in NATO*
  - *"definition": what is information science*

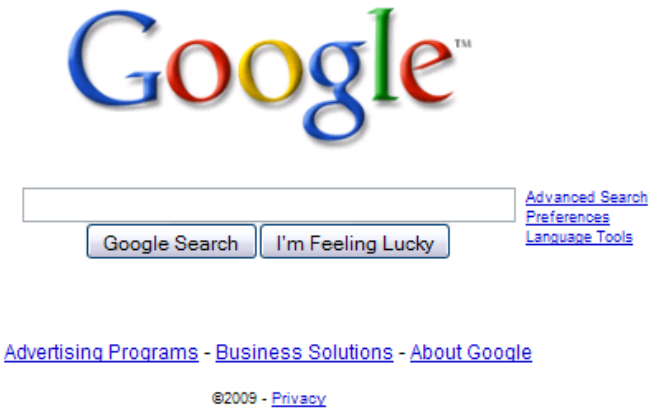
IR = Database Search?

- Database—shared, integrated computer structure that contains data (raw facts) and their associated metadata
  - *Student information stored in university database*
  - *account information stored in banks' databases*
- Relational database
  - *Collection of "tables"*
  - *Model some aspects of the "world"*

Database Queries vs IR Queries

- What would you want to know from a database?
  - *Which paints did George Geoff paint?*
  - *Which is the most expensive painting?*
  - *What is the average price of the painting in the gallery owned by G. G.Waters*

- Would you type these queries to



IR vs Database

	Databases	IR
Data	**Structured**	**Unstructured**
Fields	**Clear semantics** (SSN, age)	**No fields** (other than text)
Queries	**Defined** (relational algebra, SQL)	**Free text** ("natural language"), Boolean
Recoverability	**Critical** (concurrency control, recovery, atomic operations)	**Downplayed**, though still an issue
Matching	**Exact** (results are *always* "correct")	**Imprecise** (need to measure effectiveness)

Table is extracted from Allan 2004 IR course

Information Retrieval is Hard!

- Under/over-specified query
  - *Ambiguous: "visiting Harbin" (business or leisure?)*
  - *Incomplete: "hotels in Washington", which Washington?*
  - *What if document just mention "DC"?*

- Vague semantics of documents
  - *Ambiguity:*
    - E.g., word-sense: bank
    - E.g., structural: I look at the house on the hill
  - *Incomplete: Inferences required: e.g., "see you at the usual place"*

- Even hard for people!
  - *80% agreement in human judgments*

From Chengxiang Zhai 2008 IR course

Daqing He

Information Retrieval is "Easy"!

- Information Retrieval CAN be easy in a particular case
  - *Ambiguity in query/document is RELATIVE to the database*
  - *So, if the query is SPECIFIC enough, just one keyword may get all the relevant documents*

- PERCEIVED Information Retrieval performance is usually better than the actual performance
  - *Users can NOT judge the completeness of an answer*

From Chengxiang Zhai 2008 IR course

Daqing He

History of IR on One Slide

- Birth of IR
  - *1945: V. Bush's article "As we may think"*
  - *1957: H. P. Luhn's idea of word counting and matching*
- Indexing & Evaluation Methodology (1960's)
  - *Smart system (G. Salton's group)*
  - *Cranfield test collection (C. Cleverdon's group)*
  - *Indexing: automatic can be as good as manual (controlled vocabulary)*
- IR Models (1970's & 1980's) …
- Large-scale Evaluation & Applications (1990's-Present)
  - *TREC (D. Harman & E.Voorhees, NIST)*
  - *Web search, PubMed, …*
  - *Boundary with related areas are disappearing*

From Chengxiang Zhai 2008 IR course

Basic Approach to IR

- Most successful approaches are statistical
  - *Direct or indirect use of probabilities*
- Natural Language Processing has tried
  - *Let computer understands the meaning of query and documents, and match them*
  - *Works in restricted domains*
    - Unstable in unrestricted domains
- Rely on automatic methods
  - *Manual indexing cannot scale, and*
  - *agreement is a problem*

Inspired by James Allan's 2004 IR slides

Basic Assumption

- Similar vocabulary <=> relevance, and "similar" can mean
  - *String matching / comparison*
  - *Same vocabulary used*
  - *Probability that documents arise from*
  - *Same meaning of text*
- Bag of Word Representation
  - *Index Terms are:Words in documents without ordering, relation*

4: Wii
3: Amazon, Selling
2: 17, 2007, games, retailer, second, sold, top, video
1: additionally, com, console, item, division, entire, helped, item, make, most, online, product, profitable, recently, stock, year, ...

Original Document

Amazon Sold 17 Wiis Per Second In 2007

Online retailer Amazon.com recently announced that Wii systems were among its top-selling items in 2007, and sales of the console helped make the year its most profitable one to date.

Amazon said it sold about 17 Wiis per second when the product was in stock. Additionally, the retailer noted that along with Wii sales, its best-selling video games, *Super Mario Galaxy* and *Call of Duty 4: Modern Warfare*, were top sellers in its entire video games and hardware division.

Extracted from www.gamasutra.com/php-bin/news_index.php?story=16762

Information to be Retrieved

- Types of documents
  - *Articles, web pages, online catalogs, structured records, multimedia objects, user-generated contents, etc.*
- Types of document representations
  - *Bibliographic representation*
  - *Full-text representation*
  - *Directory representation*

Bibliographic Representation

- Title: Online Retrieval
- Author(s): Geraldene Walker, Joseph Janes
- Publisher: Libraries Unlimited
- Publication Date: August 15, 1999
- Subject: Reference
- Format: Paperback
- Pages: 312
- Dimensions: 8.10 x 11.52 x 0.81 in
- ISBN: 1563086573

Full-text Representation

1 THE SEARCH FOR INFORMATION IN THE ONLINE AGE

This chapter identifies some of the problems facing the information-seeker in a world of information overload. It introduces the concept of online searching and attempts to show the usefulness of such searching by giving examples of the wide variety of information-bearing materials that are available online today. The chapter then explains the basic mechanics of information retrieval, which apply equally to both manual and computerized information systems of all types—library catalogs, CD-ROMs, online database systems, and the Internet.

Information Overload

The problems of organizing information in order to find it when it is required are nothing new. From as early as the seventh century BC, collections of information-bearing artifacts existed in the form of stone tablets, and later parchment scrolls. However, for many centuries, the emphasis was on preservation of the cultural heritage of society in an uncertain world, rather than on its retrieval in response to a request.[1] Documents were relatively rare, and so few people could read that there was little call for access to individual items. Today, the situation has totally changed. The pressures brought about by the explosion in the amount of material being published annually over the last 50 years have highlighted the problems involved in retrieving a single desired item from the proverbial "haystack" of available information. The focus has moved from collection and preservation to retrieval and selection.

It is all too easy to assume that modern technology can solve this retrieval problem, or at least simplify the process. The use of computers and communication technologies has had an enormous influence on the way that information is produced, organized, stored, searched, and transmitted, and has certainly made more information accessible to more

Directory Representation

So !

- IR is based on very simple approach
  - *Count words in documents*
  - *Compare them to the words in a query*
  - *This approach is very effective*
- Can also consider other types of features
  - *Phrases*
  - *Named entities (people, locations, or organizations)*

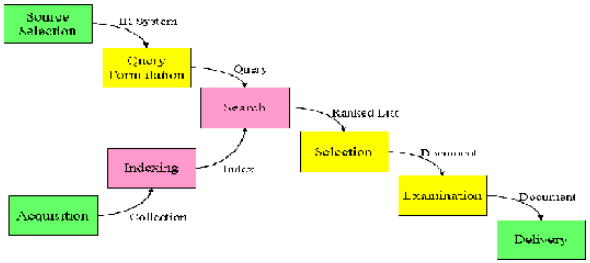
Course Overview

Do you know these?

- Bag of word representation, information needs
- Tokenization, Stemming, inverted index, gamma code
- Exact match vs best match models
- tf, idf, vector space model
- Statistical language models, query likelihood, document likelihood, KL divergence
- Relevance feedback, interactive RF, pseudo RF, Rocchio algorithm
- Cranfield evaluation methodology, TREC, precision, recall, mean average precision, NDCG
- Cross language retrieval, parallel corpus, query translation, document translation,

Why Study Information Retrieval

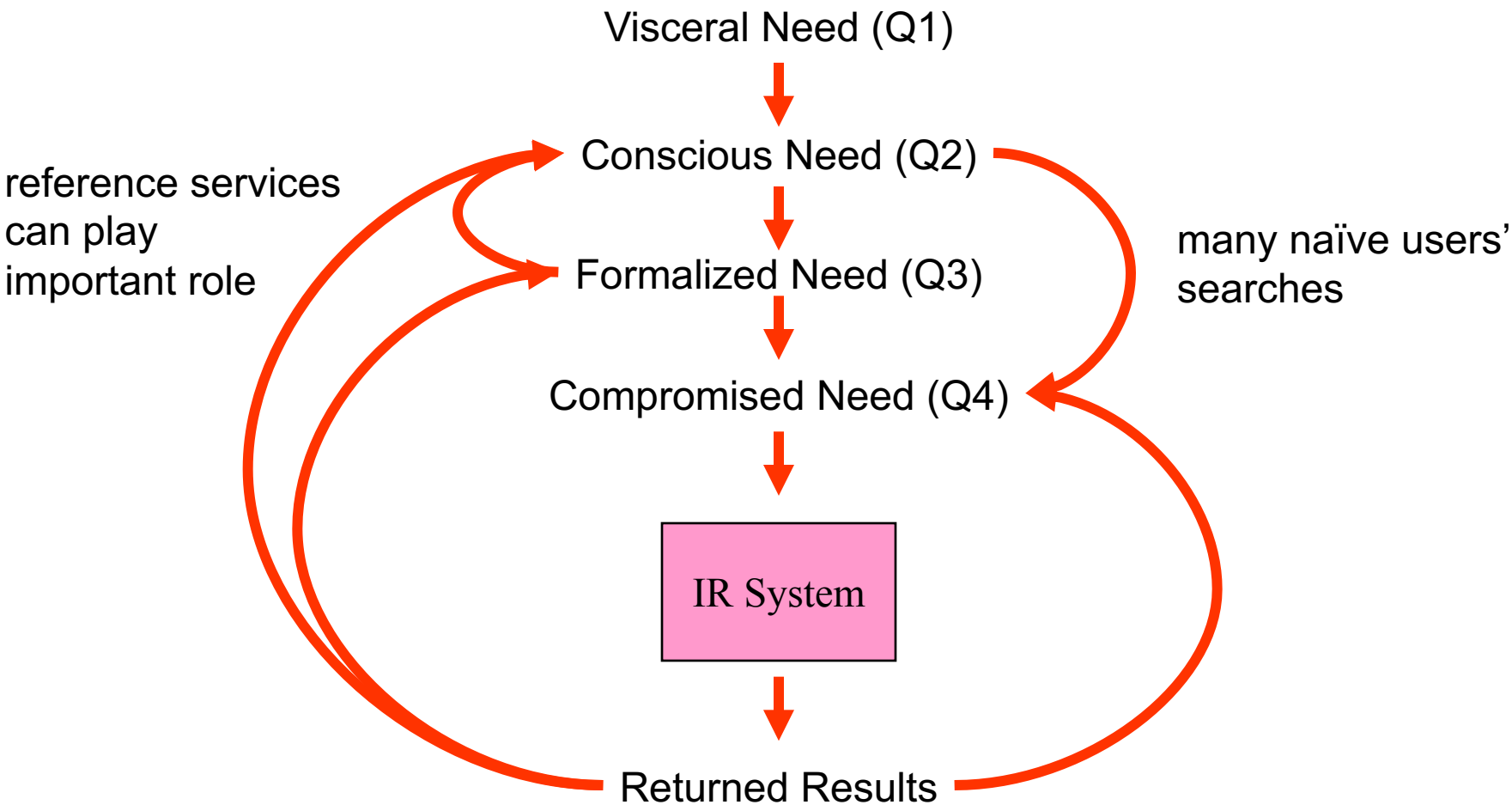
- Information Retrieval Process is complex
  - *Many stages in the process*



  - *Multiple stages in people's information need (Taylor's model)*
    - Visceral need (Q1): the actual, but unexpressed need
    - Conscious need (Q2): the within brain description of the need
    - Formalized need (Q3): the formal statement of the need
    - Compromised need (Q4): the query presented to the IR system

Robert S. Taylor. (1968) Question-negotiation and information seeking in libraries. College and Research libraries 29:179-194.

Taylor's Model and IR Process



Belkin's ASK Model

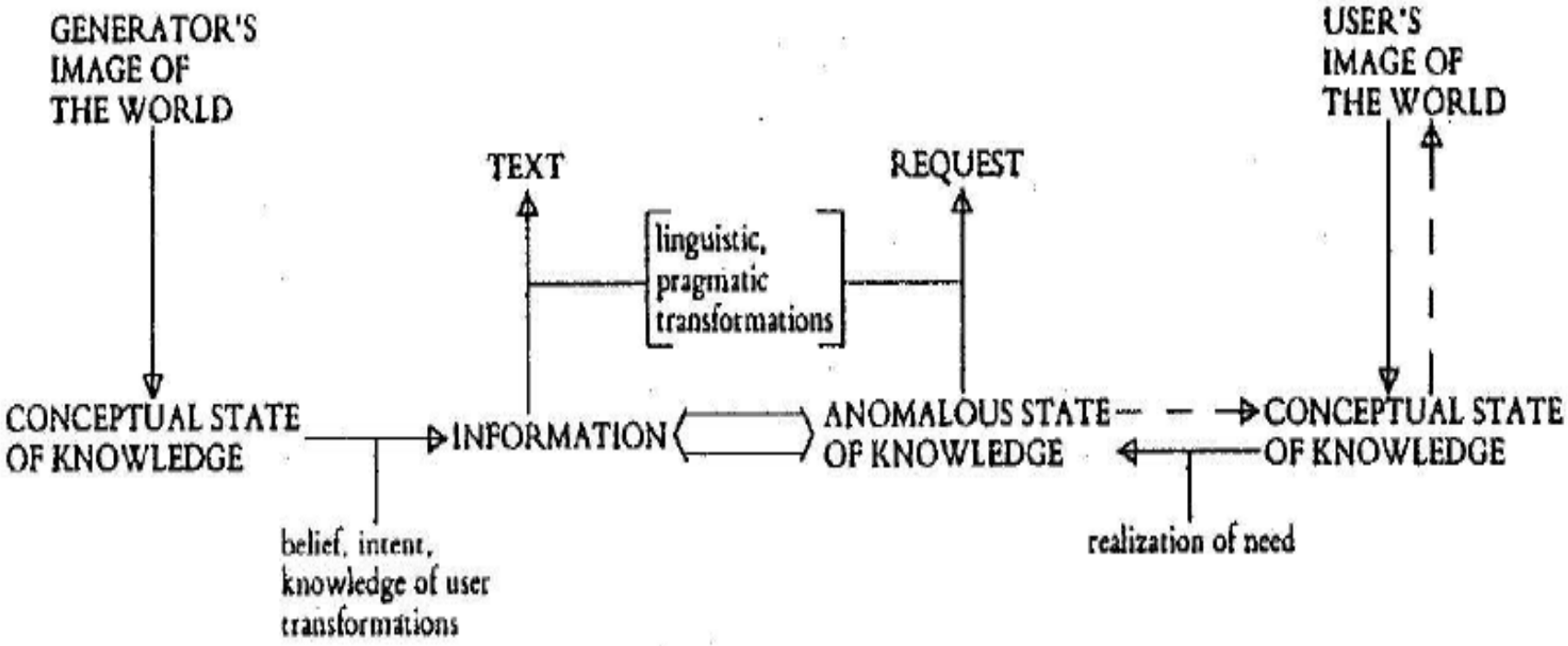


FIG. 1. *A cognitive communication system for information retrieval* (from Belkin,[9] p. 135)

Course Description

- an examination of problems and techniques related to storing and accessing unstructured information with an emphasis on textual information.
- The content of the course includes
  - *Overview of several approaches to information access with a primary focus on search-based information access.*
  - *Covers automated retrieval system design, content analysis, retrieval models, result presentation, and system evaluation.*
  - *Examines applications of retrieval techniques on theWeb, in multimedia and multilingual environments, and in text classification and event tracking.*
- Prerequisites: introduction to logic and statistical analysis, familiarity with a high-level programming language

Course Goals

- Goals
  - *to understand the dimensions of the information retrieval "problem";*
  - *to master the analysis and design of information retrieval systems;*
  - *to consider the factors which optimize the information retrieval process;*
  - *to examine current issues in information retrieval*

Objectives

- Upon satisfactory completion of this course, students will:
  - *be able to explain core concepts and terms of information retrieval*
  - *be able to explain different retrieval models and basic algorithms*
  - *be able to evaluate existing information retrieval systems and suggest how the systems can be improved*
  - *Be able to apply theories to effectively solve information retrieval problems in real world situations*

Approach

- Textbooks and Readings
  - *Provide background and detail*
- Class sessions
  - *Provide conceptual structure*
    - Outline notes provided in class
    - Slides are available online
- Assignments, project
  - *Provide hands-on experience*
- Exam
  - *Measure progress*

Textbooks

- C. Manning, P. Raghavan, H. Schutze, "Introduction to Information Retrieval". Cambridge University Press. 2008. Available at http://nlp.stanford.edu/IR-book/.
- Stefan Büttcher, Charles L. A. Clarke, Gordon V. Cormack, "Information Retrieval: Implementing and Evaluating Search Engines." MIT Press. 2010. Sample chapters are available at http://www.ir.uwaterloo.ca/book/.
- Ricardo Baeza-Yates, Berthier Riberiro-Neto, "Modern Information Retrieval", 2nd Edition. Addison Wesley, 2011. ISBN-10: 9780321416919. http://www.mir2ed.org/.
- Richard K. Belew, "Finding Out About: A Cognitive Perspective on Search Engine Technology and the WWW", Cambridge University Press, 2000. Referred as "FOA" subsequently.

Readings

- 3 to 4 required readings per week
  - *Some of them are long chapters*
- Read before class on a reading system
  - *Must complete the reading before the Saturday 9:00pm before the class*

Muddiest Points

- A question about the most unclear point among all topics discussed in the class
- Create a 2140 course blog on blogger.com so that each week update it with an entry for muddiest points
  - *Post an entry in the discussion board section of the courseweb to tell TA the URL of your blog*

The Grand Plan

Text Retrieval Basics

- unit 1: introduction and overview
- unit 2: document and query processing
- unit 3: index construction and compression
- unit 4: matching models: Boolean and Vector Space
- unit 5: matching models: probabilistic and language model
- unit 6: evaluation
- unit 7: relevance feedback and query expansion
- Unit 8: user interaction and visualization
- Unit 9: middle term exam

Better and Beyond

- unit 11: Web information retrieval
- unit 12: intelligent information retrieval
- unit 13: text classification and clustering
- unit 14: latent semantic and neural-based IR

Assignments

- Total 4 assignment
  - *Assignment 1: Document Processing*
  - *Assignment 2: Index Construction*
  - *Assignment 3: Retrieval Model*
  - *Assignment 4: Relevance Feedback*
- Submission Deadline: before come to class on the due date
- Submission place: Assignment sections in Course Web

Exam

- One exam during the term
- Total 90 minutes for the exam
- Exam coverage:
  - *All topics discussed in the class up to the exam date*
- Question types:
  - *Multiple choices, short definition, short discussion, and long discussion*

Term Projects

- Group project
  - *3 persons per team*
- Task
  - *Predefined projects: projects proposed by faculty and students*
- Requirements:
  - *improve existing retrieval algorithms*
  - *designing an innovative interface,*
  - *focus on applying the existing retrieval system into innovative application domains*
- More in Unit 2

Grading

- 88% academic work
  - *Exam 28%:*
  - *Assignments 32%: 8% for each assignment*
  - *Term project 28%*
- 12% class participation
  - *Pre-class readings 5%*
    - 10 submissions would be counted, .5% each
  - *Post-class muddiest points 5%*
    - 10 submissions would be counted, .5% each
  - *Class attendance (2%)*

Grading (continued)

- 90-92.9 => A-, 93-97.9 => A, 98-100 => A+
- 80-82.9 => B-, 83-87.9 => B, 88-89.0 => B+
- 70-72.9 => C-, 73-77.9 => C, 78-79.9 => C+
- 60-69.9 => D,
- Score < 60% => F,

The Fine Print

- Group work is encouraged in class
  - *But you must personally write what you turn in*
- Deadlines are firm and sharp
- Academic integrity is a serious matter
  - *No group work during the exams!*
  - *Don't discuss exam until all students finish*

CourseWeb at Pitt

- CourseWeb site: courseweb.pitt.edu
  - *Virtual environment for course materials*
  - *Accessible 24h everyday*
  - *Login with your university computer account*
    - Contact 412-624-HELP for help
    - Set up the forward function if you do not use it

Assignment 1

- See CourseWeb Assignment Section for instructions
- Due September 17 noon

Final Words

- Course loads
  - *Lots of activities*
  - *Not easy course*
- Programming
  - *Intermediate JAVA programming,*
  - *or you want to try harder*
- Expectations
  - *Whole commitment*
  - *High and practical*
  - *So you will learn a lot*