LBSC 796/INFM 718R: Week 1

Introduction to Information Retrieval

Jimmy Lin
College of Information Studies
University of Maryland

Monday, January 30, 2006

Information Retrieval Systems

- Information
  - What is “information”?
- Retrieval
  - What do we mean by “retrieval”?
  - What are different types information needs?
- Systems
  - How do computer systems fit into the ***human*** information seeking process?

What is Information?

- What do you think?
- There is no “correct” definition
- Cookie Monster’s definition:
  - “news or facts about something”
- Different approaches:
  - Philosophy
  - Psychology
  - Linguistics
  - Electrical engineering
  - Physics
  - Computer science
  - Information science

Dictionary says…

- Oxford English Dictionary
  - **information**: informing, telling; thing told, knowledge, items of knowledge, news
  - **knowledge**: knowing familiarity gained by experience; person’s range of information; a theoretical or practical understanding of; the sum of what is known
- Random House Dictionary
  - **information**: knowledge communicated or received concerning a particular fact or circumstance; news

Intuitive Notions

- Information must
  - 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 “mis-information”
  - Be “about” something

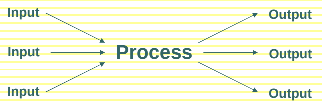
Robert M. Losee. (1997) A Discipline Independent Definition of Information. *Journal of the American Society for Information Science*, 48(3), 254-269.

Three Views of Information

- Information as process
- Information as communication
- Information as message transmission and reception

One View

- Information = characteristics of the output of a process
  - Tells us something about the process and the input



- Information-generating process do not occur in isolation

Input → Process$_1$ → Process$_2$ → … → Output

Ibid.

Where's the human?

- If a tree falls in the forest, and no one is around to hear it, is information transmitted?
- In the "information as process": Yes, but that's not very interesting to us
- We're concerned about information for human consumption
  - Transmission of information from one person to another
  - Recording of information
  - Reconstruction of stored information

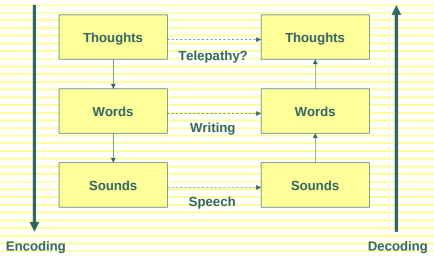
Another View

- 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"
  - This implies that the sender has knowledge of the recipient's structure
- Text = "a collection of signs purposefully structured by a sender with the intention of changing image-structure of a recipient"
- Information = "the structure of any text which is capable of changing the image-structure of a recipient"

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.

Transfer of Information

- Communication = transmission of information

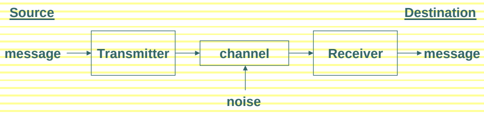


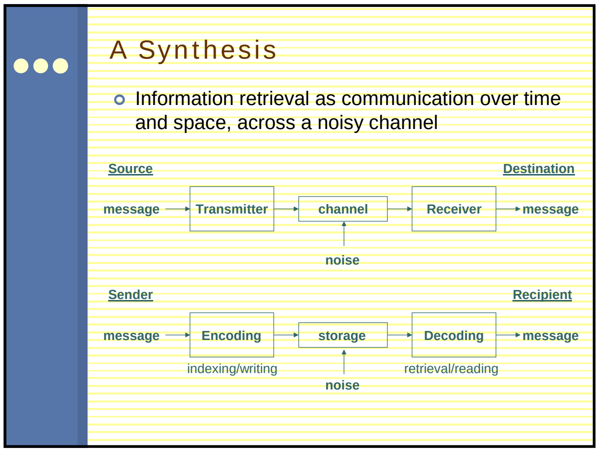
Information Theory

- Better called "communication theory"
- Developed by Claude Shannon in 1940's
  - Concerned with the transmission of electrical signals over wires
  - How do we send information quickly and reliably?
- Underlies modern electronic communication:
  - Voice and data traffic…
  - Over copper, fiber optic, wireless, etc.
- Famous result: Channel Capacity Theorem
- Formal measure of information in terms of entropy
  - Information = "reduction in surprise"

The Noisy Channel Model

- Communication = producing the same message at the destination that was sent at the source
  - The message must be encoded for transmission across a medium (called channel)
  - But the channel is noisy and can distort the message
- Semantics (meaning) is irrelevant





A Synthesis

- Information retrieval as communication over time and space, across a noisy channel

Source — Destination

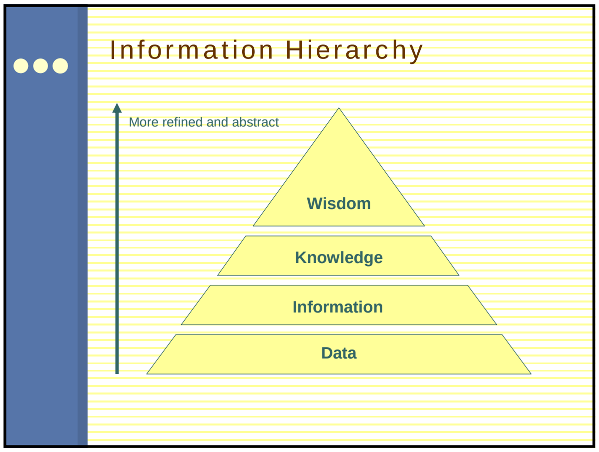
message → Transmitter → channel → Receiver → message

noise

Sender — Recipient

message → Encoding → storage → Decoding → message

indexing/writing — noise — retrieval/reading



Information Hierarchy

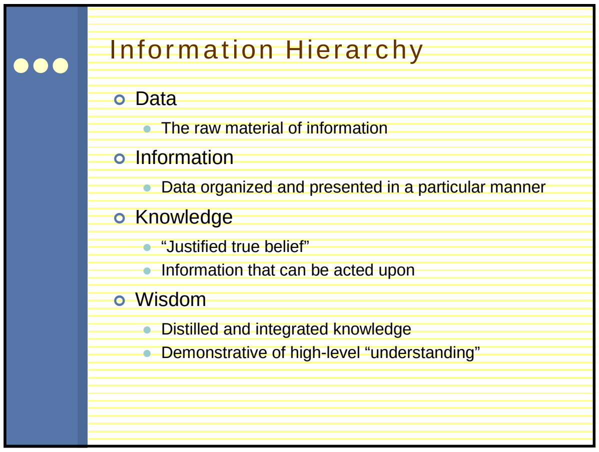
More refined and abstract

Wisdom

Knowledge

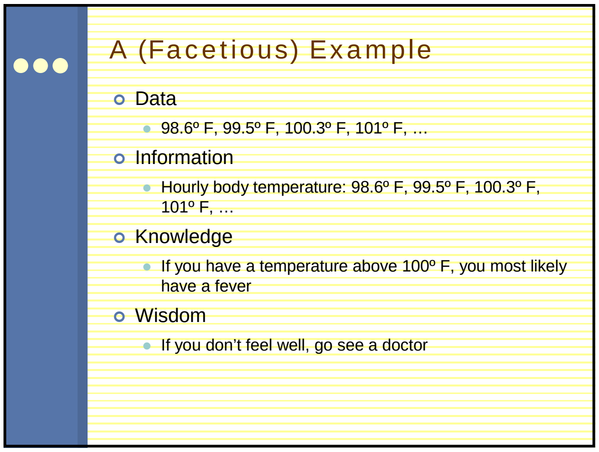
Information

Data



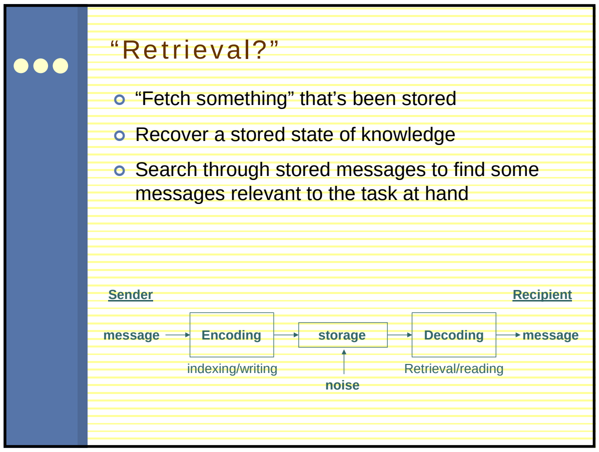
Information Hierarchy

- Data
  - The raw material of information
- Information
  - Data organized and presented in a particular manner
- Knowledge
  - "Justified true belief"
  - Information that can be acted upon
- Wisdom
  - Distilled and integrated knowledge
  - Demonstrative of high-level "understanding"



A (Facetious) Example

- Data
  - 98.6º F, 99.5º F, 100.3º F, 101º F, …
- Information
  - Hourly body temperature: 98.6º F, 99.5º F, 100.3º F, 101º F, …
- Knowledge
  - If you have a temperature above 100º F, you most likely have a fever
- Wisdom
  - If you don't feel well, go see a doctor



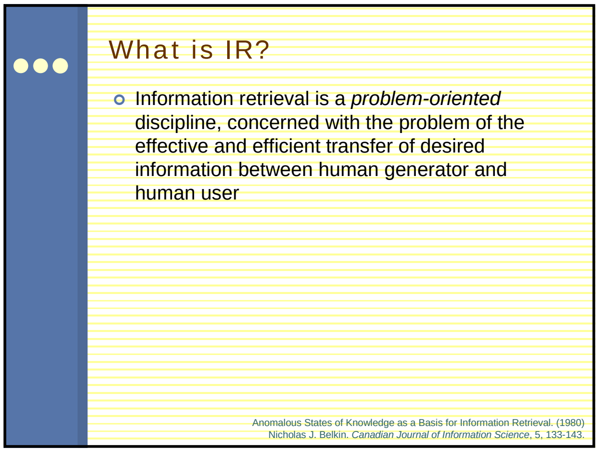
"Retrieval?"

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

Sender — Recipient

message → Encoding → storage → Decoding → message

indexing/writing — noise — Retrieval/reading



What is IR?

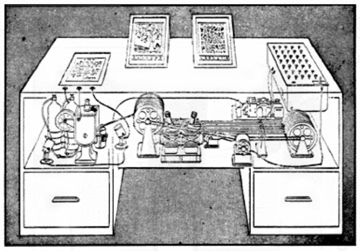
- Information retrieval is a *problem-oriented* discipline, concerned with the problem of the effective and efficient transfer of desired information between human generator and human user

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

Modern History

- The “information overload” problem is much older than you may think
- Origins in period immediately after World War II
  - Tremendous scientific progress during the war
  - Rapid growth in amount of scientific publications available
- The “Memex Machine”
  - Conceived by Vannevar Bush, President Roosevelt's science advisor
  - Outlined in 1945 Atlantic Monthly article titled “As We May Think”
  - Foreshadows the development of hypertext (the Web) and information retrieval system

The Memex Machine



Memex in the form of a desk would instantly bring files and material on any subject to the operator's fingertips. Slanting translucent viewing screens magnify supermicrofilm filed by code numbers. At left is a mechanism which automatically photographs longhand notes, pictures and letters, then files them in the desk for future reference (*LIFE 19*(11), p. 123).

Types of Information Needs

- Retrospective
  - “Searching the past”
  - Different queries posed against a static collection
  - Time invariant
- Prospective
  - “Searching the future”
  - Static query posed against a dynamic collection
  - Time dependent

Retrospective Searches (I)

- *Ad hoc* retrieval: find documents “about this”

  Identify positive accomplishments of the Hubble telescope since it was launched in 1991.

  Compile a list of mammals that are considered to be endangered, identify their habitat and, if possible, specify what threatens them.

- Known item search

  Find Jimmy Lin's homepage.

  What's the ISBN number of “Modern Information Retrieval”?

- Directed exploration

  Who makes the best chocolates?

  What video conferencing systems exist for digital reference desk services?

Retrospective Searches (II)

- Question answering

“Factoid”	Who discovered Oxygen? When did Hawaii become a state? Where is Ayer's Rock located? What team won the World Series in 1992?
“List”	What countries export oil? Name U.S. cities that have a “Shubert” theater.
“Definition”	Who is Aaron Copland? What is a quasar?

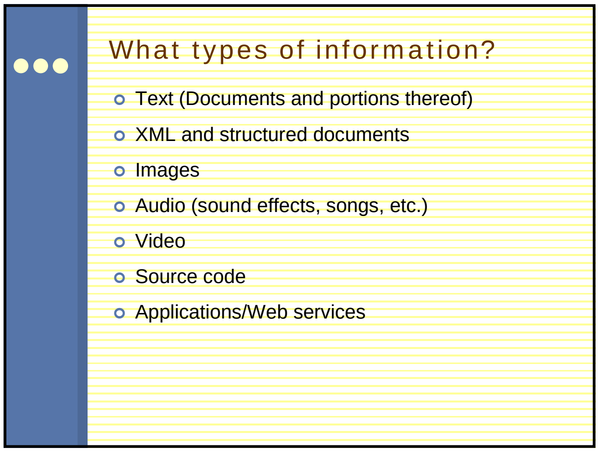
Prospective “Searches”

- Filtering
  - Make a binary decision about each incoming document

  Spam or not spam?

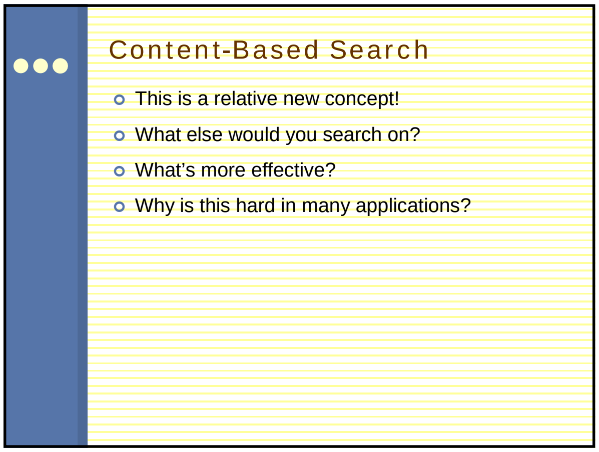
- Routing
  - Sort incoming documents into different bins?

  Categorize news headlines: World? Nation? Metro? Sports?



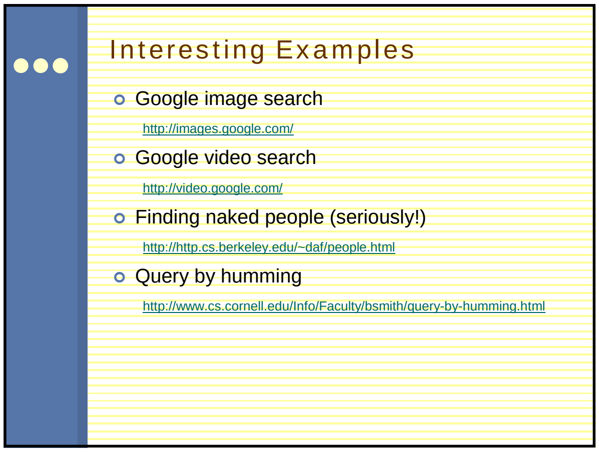
What types of information?

- Text (Documents and portions thereof)
- XML and structured documents
- Images
- Audio (sound effects, songs, etc.)
- Video
- Source code
- Applications/Web services



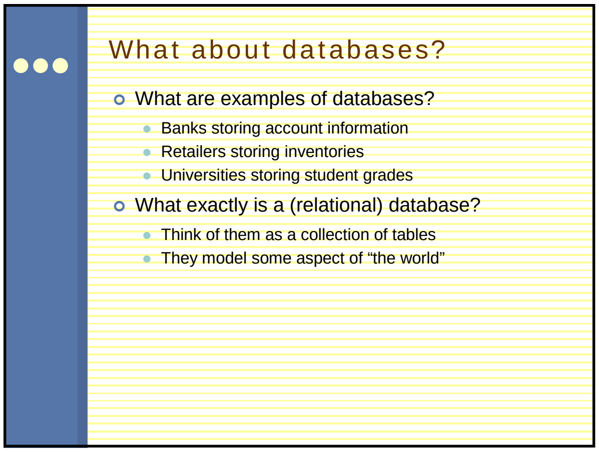
Content-Based Search

- This is a relative new concept!
- What else would you search on?
- What's more effective?
- Why is this hard in many applications?



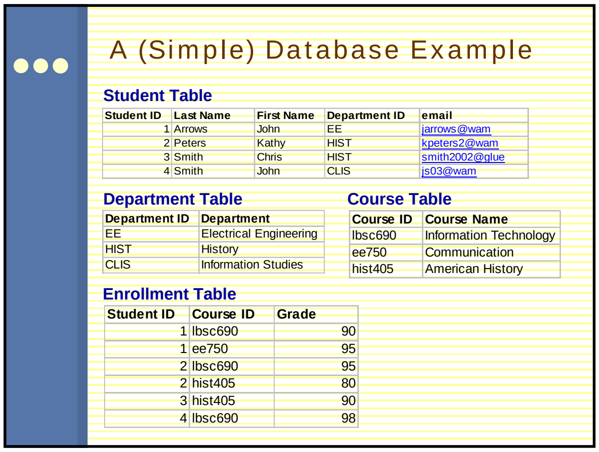
Interesting Examples

- Google image search
  - http://images.google.com/
- Google video search
  - http://video.google.com/
- Finding naked people (seriously!)
  - http://http.cs.berkeley.edu/~daf/people.html
- Query by humming
  - http://www.cs.cornell.edu/Info/Faculty/bsmith/query-by-humming.html



What about databases?

- What are examples of databases?
  - Banks storing account information
  - Retailers storing inventories
  - Universities storing student grades
- What exactly is a (relational) database?
  - Think of them as a collection of tables
  - They model some aspect of "the world"



A (Simple) Database Example

Student Table

Student ID	Last Name	First Name	Department ID	email
1	Arrows	John	EE	jarrows@wam
2	Peters	Kathy	HIST	kpeters2@wam
3	Smith	Chris	HIST	smith2002@glue
4	Smith	John	CLIS	js03@wam

Department Table

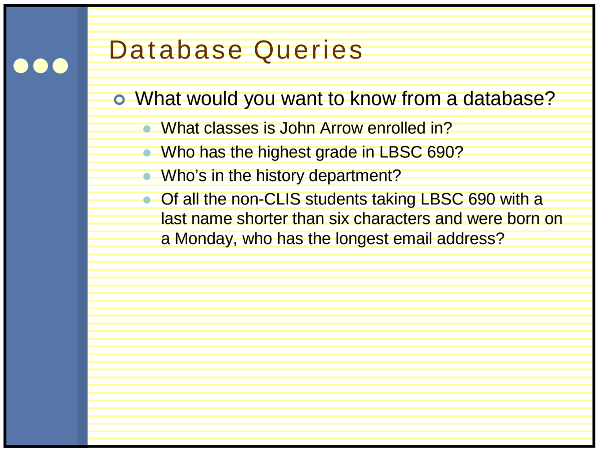
Department ID	Department
EE	Electrical Engineering
HIST	History
CLIS	Information Studies

Course Table

Course ID	Course Name
lbsc690	Information Technology
ee750	Communication
hist405	American History

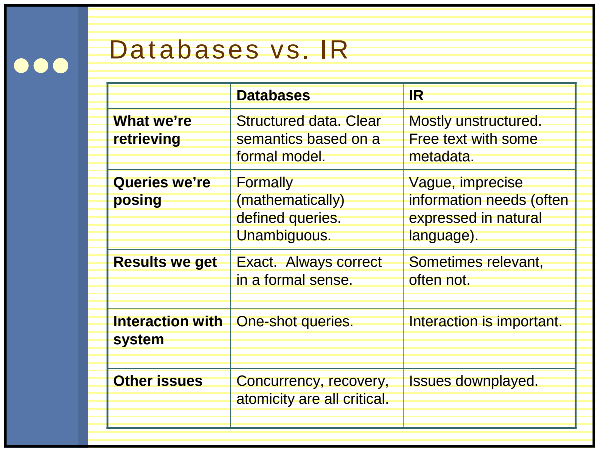
Enrollment Table

Student ID	Course ID	Grade
1	lbsc690	90
1	ee750	95
2	lbsc690	95
2	hist405	80
3	hist405	90
4	lbsc690	98



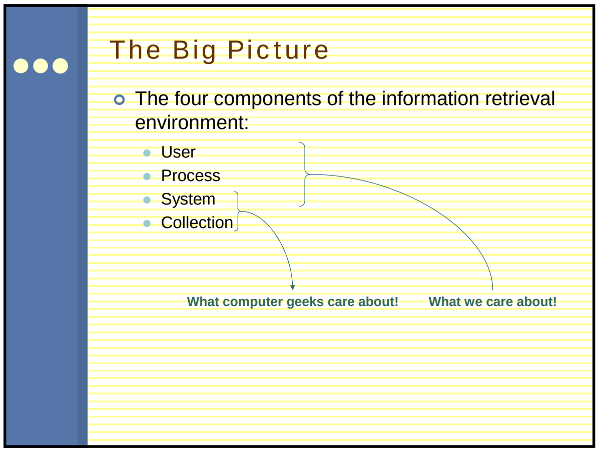
Database Queries

- What would you want to know from a database?
  - What classes is John Arrow enrolled in?
  - Who has the highest grade in LBSC 690?
  - Who's in the history department?
  - Of all the non-CLIS students taking LBSC 690 with a last name shorter than six characters and were born on a Monday, who has the longest email address?



Databases vs. IR

	Databases	IR
What we're retrieving	Structured data. Clear semantics based on a formal model.	Mostly unstructured. Free text with some metadata.
Queries we're posing	Formally (mathematically) defined queries. Unambiguous.	Vague, imprecise information needs (often expressed in natural language).
Results we get	Exact. Always correct in a formal sense.	Sometimes relevant, often not.
Interaction with system	One-shot queries.	Interaction is important.
Other issues	Concurrency, recovery, atomicity are all critical.	Issues downplayed.

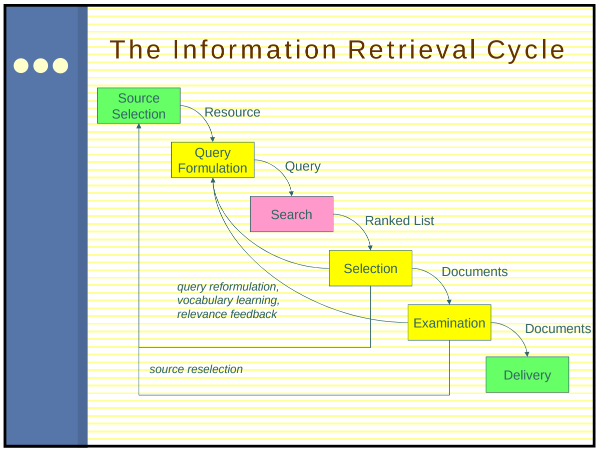


The Big Picture

- The four components of the information retrieval environment:
  - User
  - Process
  - System
  - Collection

What computer geeks care about!

What we care about!



The Information Retrieval Cycle

Source Selection

Resource

Query Formulation

Query

Search

Ranked List

Selection

Documents

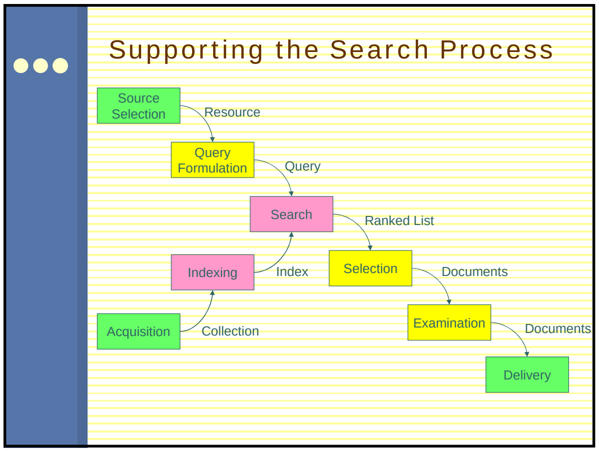
Examination

Documents

Delivery

query reformulation, vocabulary learning, relevance feedback

source reselection



Supporting the Search Process

Source Selection

Resource

Query Formulation

Query

Search

Ranked List

Selection

Documents

Examination

Documents

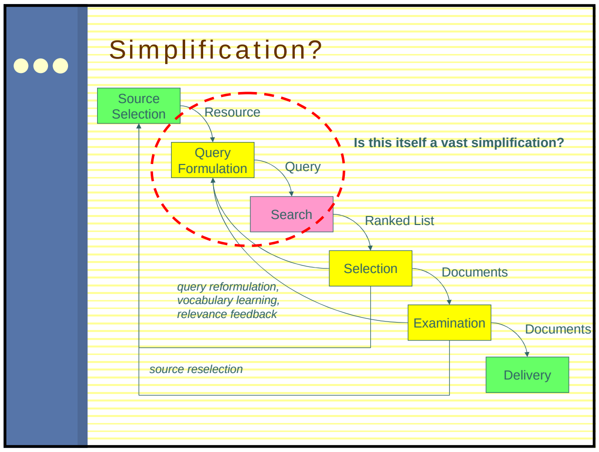
Delivery

Indexing

Index

Acquisition

Collection



Simplification?

Source Selection

Resource

Query Formulation

Query

Search

Ranked List

Selection

Documents

Examination

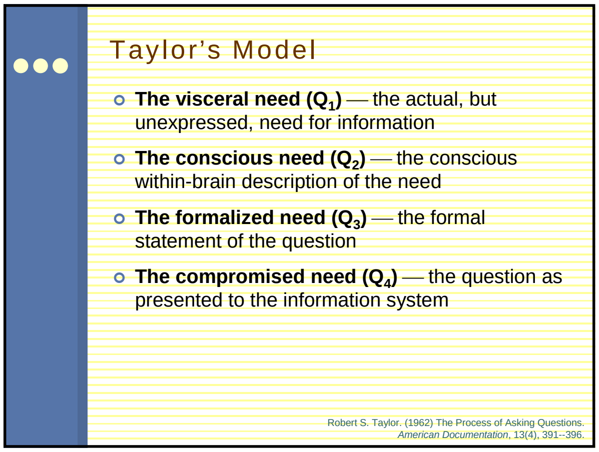
Documents

Delivery

Is this itself a vast simplification?

query reformulation, vocabulary learning, relevance feedback

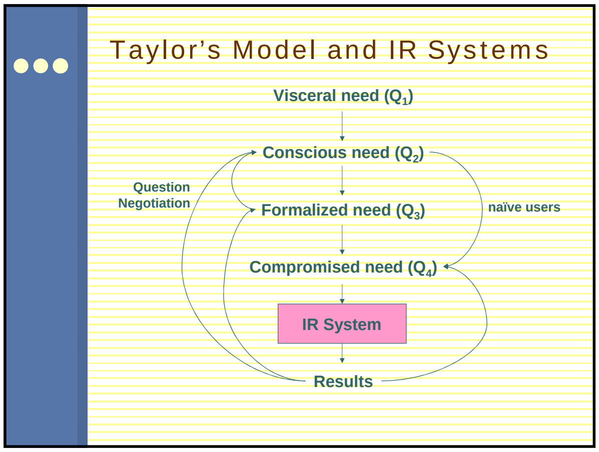
source reselection



Taylor's Model

- **The visceral need (Q_1)** — the actual, but unexpressed, need for information
- **The conscious need (Q_2)** — the conscious within-brain description of the need
- **The formalized need (Q_3)** — the formal statement of the question
- **The compromised need (Q_4)** — the question as presented to the information system

Robert S. Taylor. (1962) The Process of Asking Questions. *American Documentation*, 13(4), 391--396.



Taylor’s Model and IR Systems

Visceral need (Q_1)

Conscious need (Q_2)

Formalized need (Q_3)

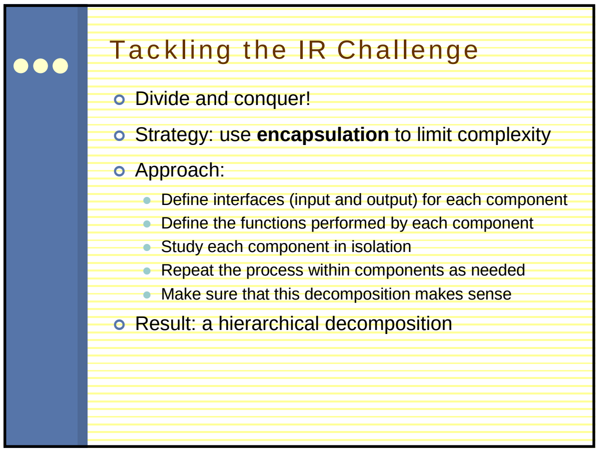
Compromised need (Q_4)

IR System

Results

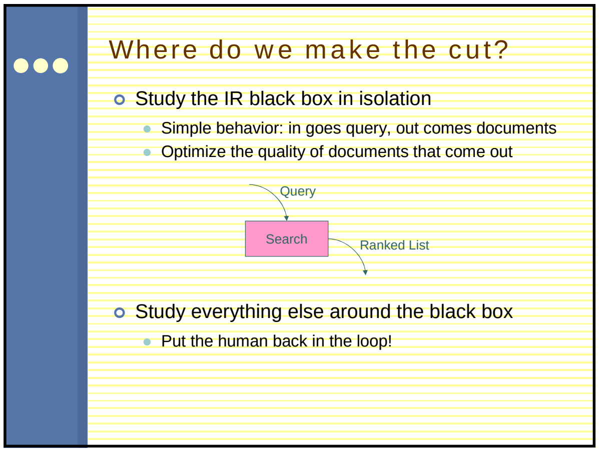
Question Negotiation

naïve users



Tackling the IR Challenge

- Divide and conquer!
- Strategy: use **encapsulation** to limit complexity
- Approach:
  - Define interfaces (input and output) for each component
  - Define the functions performed by each component
  - Study each component in isolation
  - Repeat the process within components as needed
  - Make sure that this decomposition makes sense
- Result: a hierarchical decomposition



Where do we make the cut?

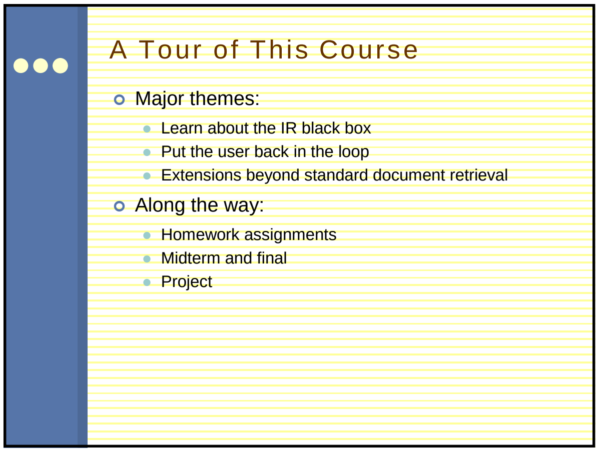
- Study the IR black box in isolation
  - Simple behavior: in goes query, out comes documents
  - Optimize the quality of documents that come out

Query

Search

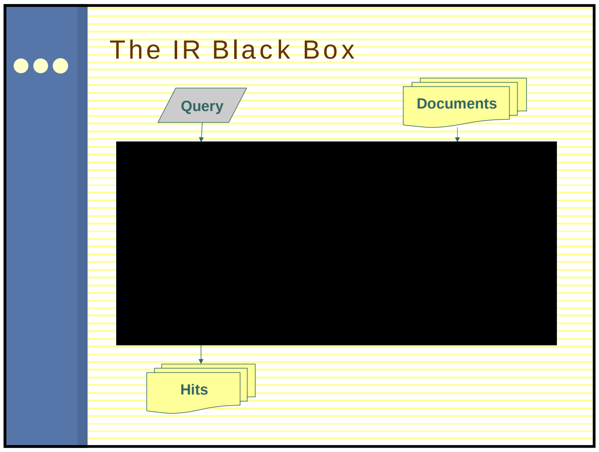
Ranked List

- Study everything else around the black box
  - Put the human back in the loop!



A Tour of This Course

- Major themes:
  - Learn about the IR black box
  - Put the user back in the loop
  - Extensions beyond standard document retrieval
- Along the way:
  - Homework assignments
  - Midterm and final
  - Project

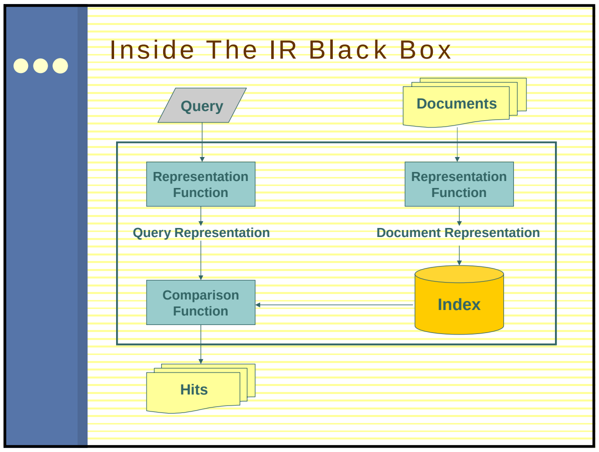


The IR Black Box

Query

Documents

Hits



Inside The IR Black Box

Query

Documents

Representation Function

Representation Function

Query Representation

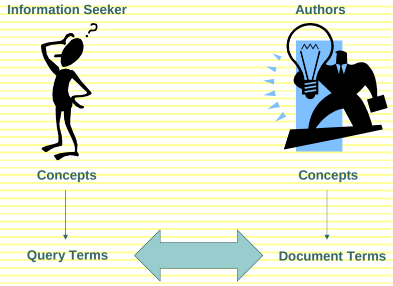
Document Representation

Comparison Function

Index

Hits

The Central Problem in IR



Do these represent the same concepts?

What makes IR “experimental”?

- **Week 2:** Evaluation
  - How do design experiments that answer our questions?
  - How do we assess the quality of the documents that come out of the IR black box?
  - Can we do this automatically?

Building the IR Black Box

- **Week 3 and 4:** Different models of information retrieval
  - Boolean model
  - Vector space model
  - Languages models
- **Week 5:** Representing the meaning of documents
  - How do we capture the meaning of documents?
  - Is meaning just the sum of all terms?
- **Week 6:** Indexing
  - How do we actually store all those words?
  - How do we access indexed terms quickly?

Beyond the IR Black Box

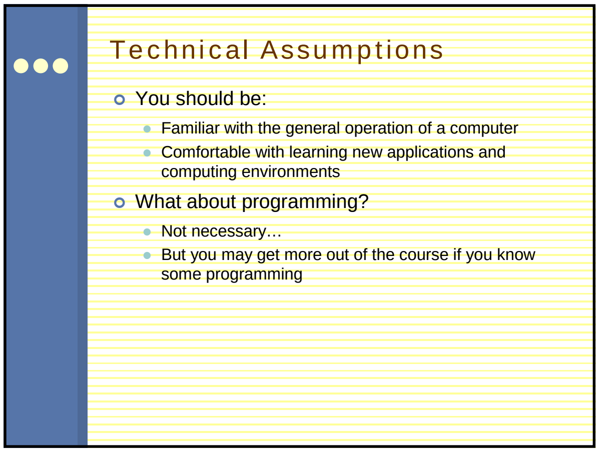
- Studying the IR black box in isolation: Is this realistic?
- What are the assumptions of this methodology?

The User in the Loop

- **Week 8:** Relevance Feedback
  - How do humans (and machines) modify queries based on retrieved results?
- **Week 9:** User Interaction
  - Information retrieval meets computer-human interaction
  - How do we present search results to users in an effective manner?
  - What tools can systems provide to aid the user in information seeking?

Extensions

- **Week 10:** Filtering and Categorization
  - Traditional information retrieval: static collection, dynamic queries
  - What about static queries against dynamic collections?
- **Week 11:** Multimedia Retrieval
  - Thus far, we’ve been focused on text…
  - What about images, sounds, video, etc.?
- **Week 12:** Question Answering
  - We want answers, not just documents!



Technical Assumptions

- You should be:
  - Familiar with the general operation of a computer
  - Comfortable with learning new applications and computing environments
- What about programming?
  - Not necessary…
  - But you may get more out of the course if you know some programming