Structure of IR Systems

LBSC 796/INFM 718R

Session 1, January 26, 2011

Doug Oard

Agenda

- Teaching theater orientation
- The structure of interactive IR systems
- Course overview

Some Holistic Definitions of IR

- 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.

- A process for establishing a view on an information space from a perspective defined by the user.

Douglas W. Oard, in class, today..

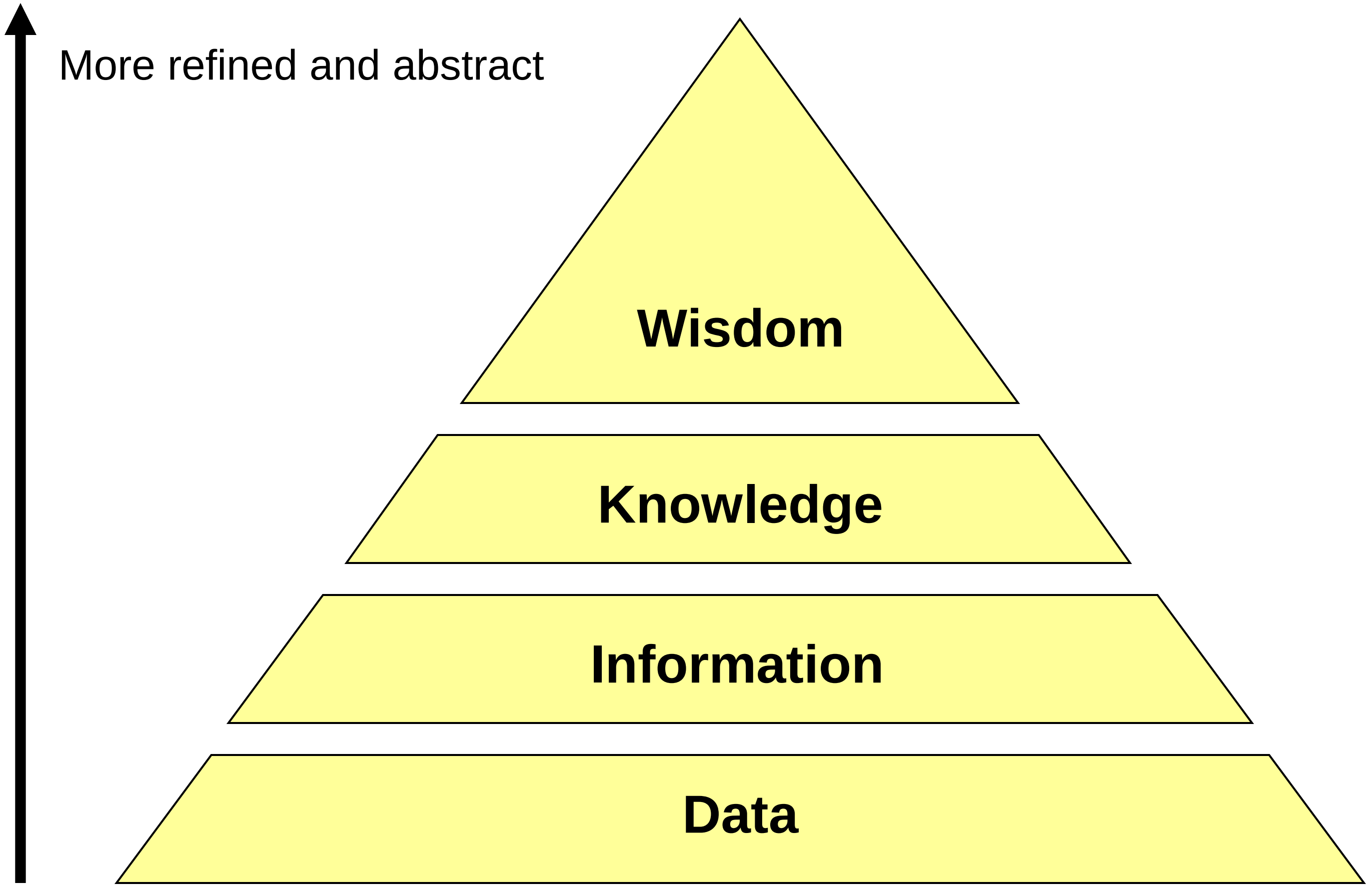
Information Retrieval Systems

- 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 do We Mean by “Information?”

- How is it different from “data”?
  - Information is **<u>data in context</u>**
    - Databases contain data and produce information
    - IR systems contain and provide information
- How is it different from “knowledge”?
  - Knowledge is a **<u>basis for making decisions</u>**
    - Many “knowledge bases” contain decision rules

Information Hierarchy



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”

An 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

What types of information?

- Text
- Structured documents (e.g., XML)
- Images
- Audio (sound effects, songs, etc.)
- Video
- Programs
- Services

What Do We Mean by "Retrieval?"

- Find something that you want
  - The information need may or may not be **explicit**
- Known item search
  - Find the class home page
- Answer seeking
  - Is Lexington or Louisville the capital of Kentucky?
- Directed exploration
  - Who makes videoconferencing systems?

Relevance

- **Relevance** relates a <u>topic</u> and a document
  - Duplicates are equally relevant, by definition
  - Constant over time and across users
- **Pertinence** relates a <u>task</u> and a document
  - Accounts for quality, complexity, language, …
- **Utility** relates a <u>user</u> and a document
  - Accounts for prior knowledge

Types of Information Needs

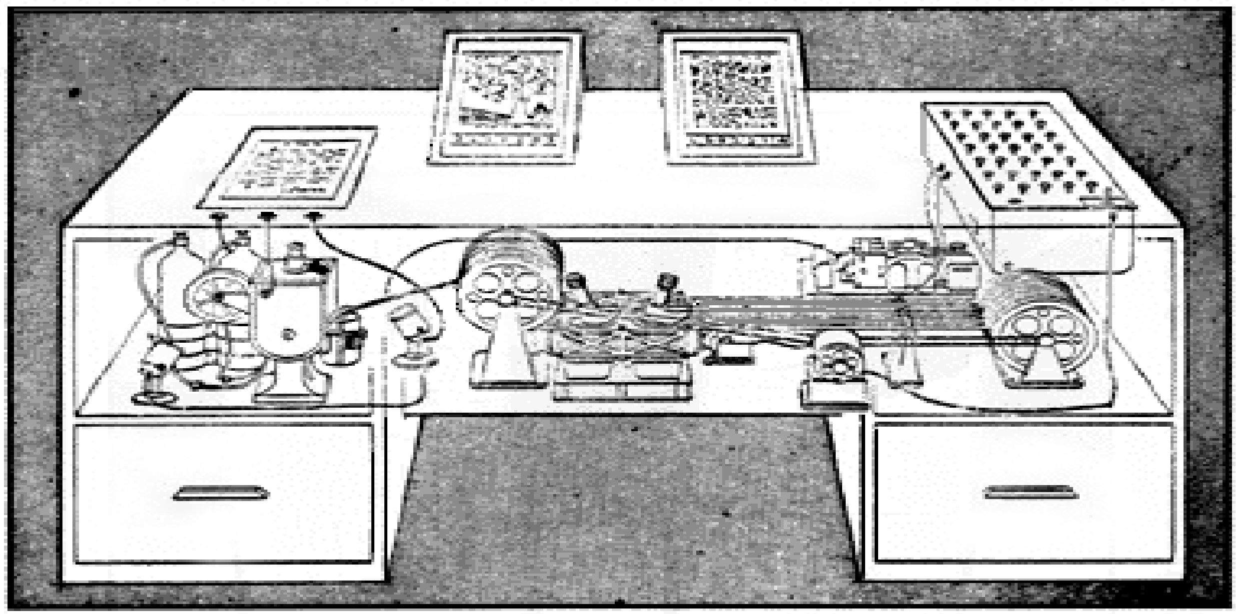
- Retrospective (“Retrieval”)
  - “Searching the past”
  - Different queries posed against a static collection
  - Time invariant

- Prospective (“Recommendation”)
  - “Searching the future”
  - Static query posed against a dynamic collection
  - Time dependent

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.

Systems: The Memex



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).

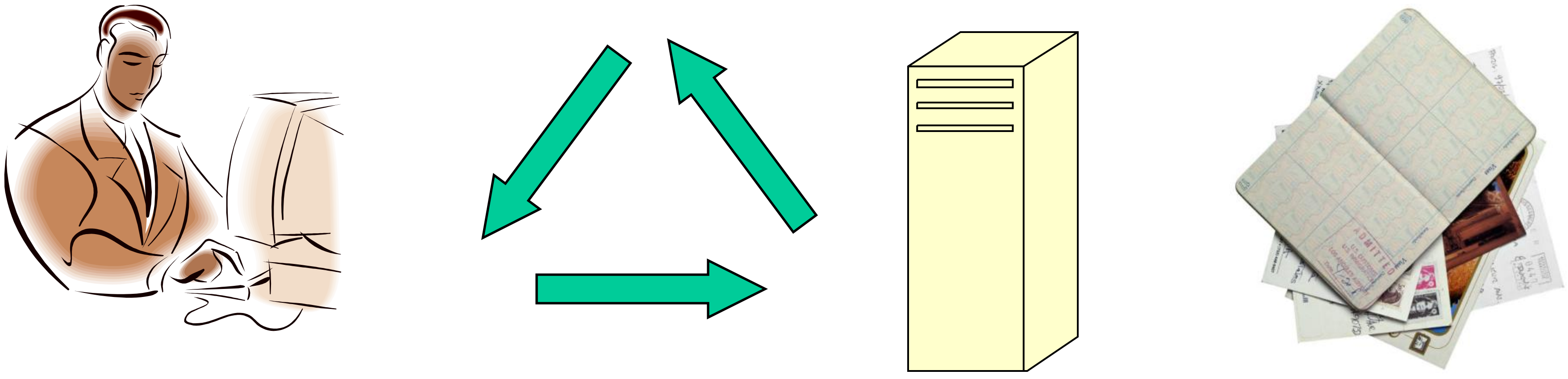
Design Strategies

- Foster human-machine synergy
  - Exploit complementary strengths
  - Accommodate shared weaknesses
- Divide-and-conquer
  - Divide task into stages with well-defined interfaces
  - Continue dividing until problems are easily solved
- Co-design related components
  - Iterative process of joint optimization

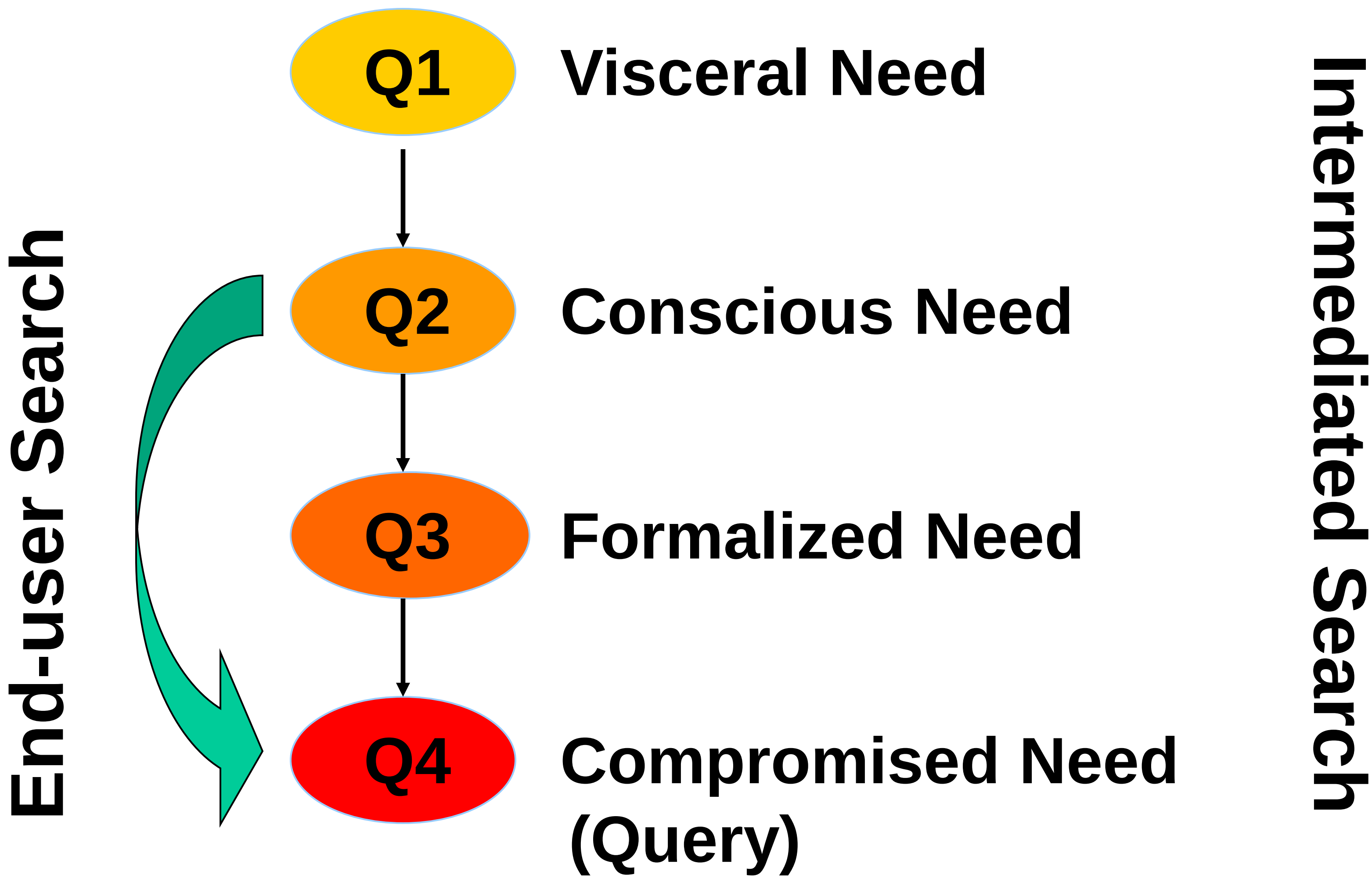
Human-Machine Synergy

- Machines are good at:
  - Doing simple things accurately and quickly
  - Scaling to larger collections in sublinear time
- People are better at:
  - Accurately recognizing what they are looking for
  - Evaluating intangibles such as “quality”
- Both are pretty bad at:
  - Mapping consistently between words and concepts

Process/System Co-Design



Taylor’s Model of Question Formation

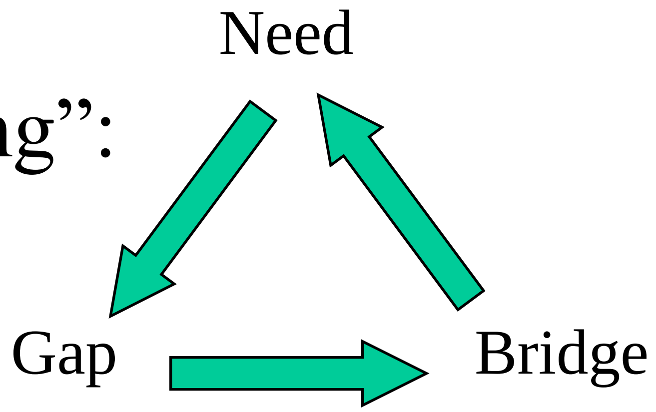


Iterative Search

- Searchers often don't clearly understand
  - The problem they are trying to solve
  - What information is needed to solve the problem
  - How to ask for that information

- The query results from a clarification process

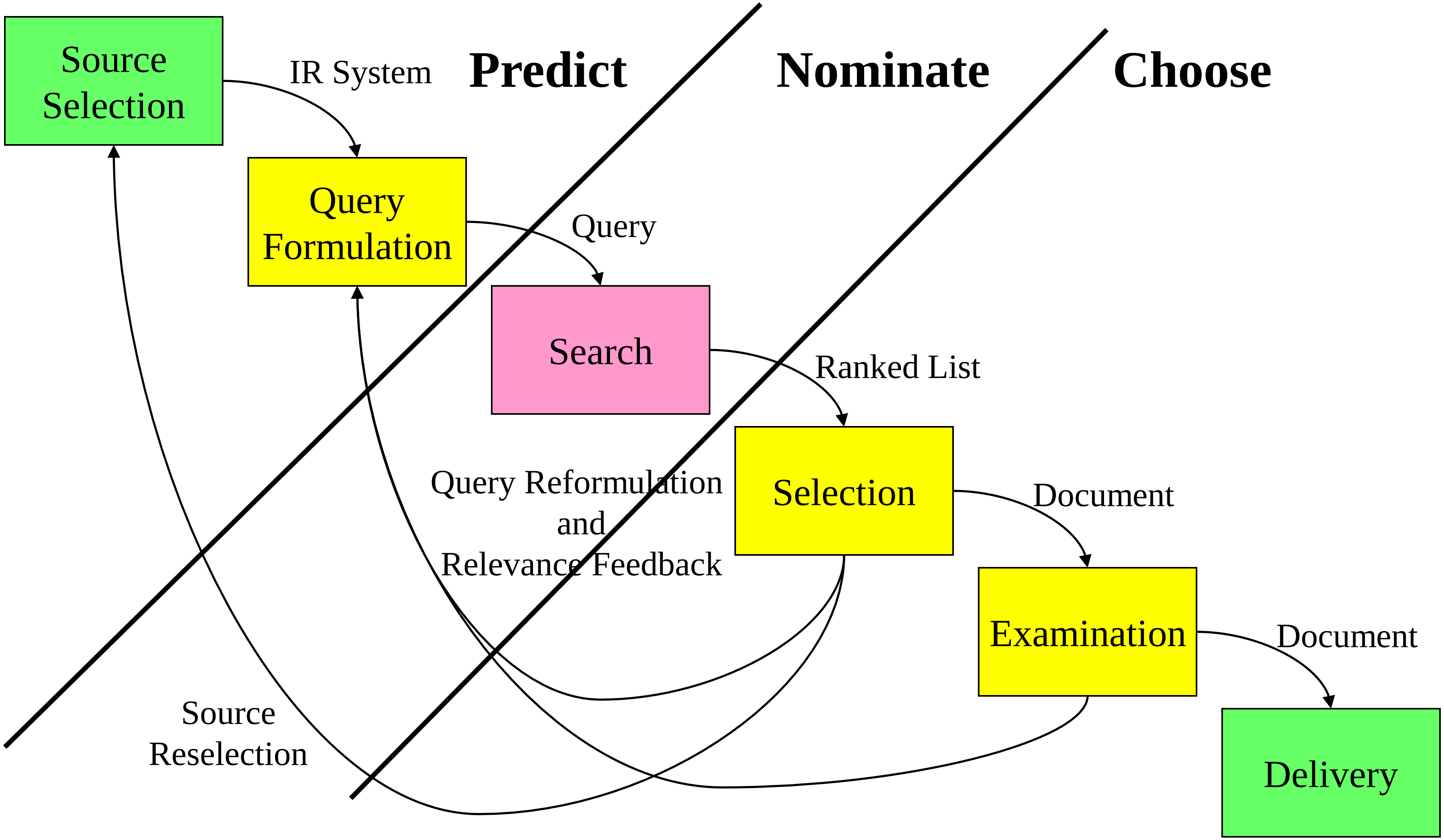
- Dervin's "sense making":



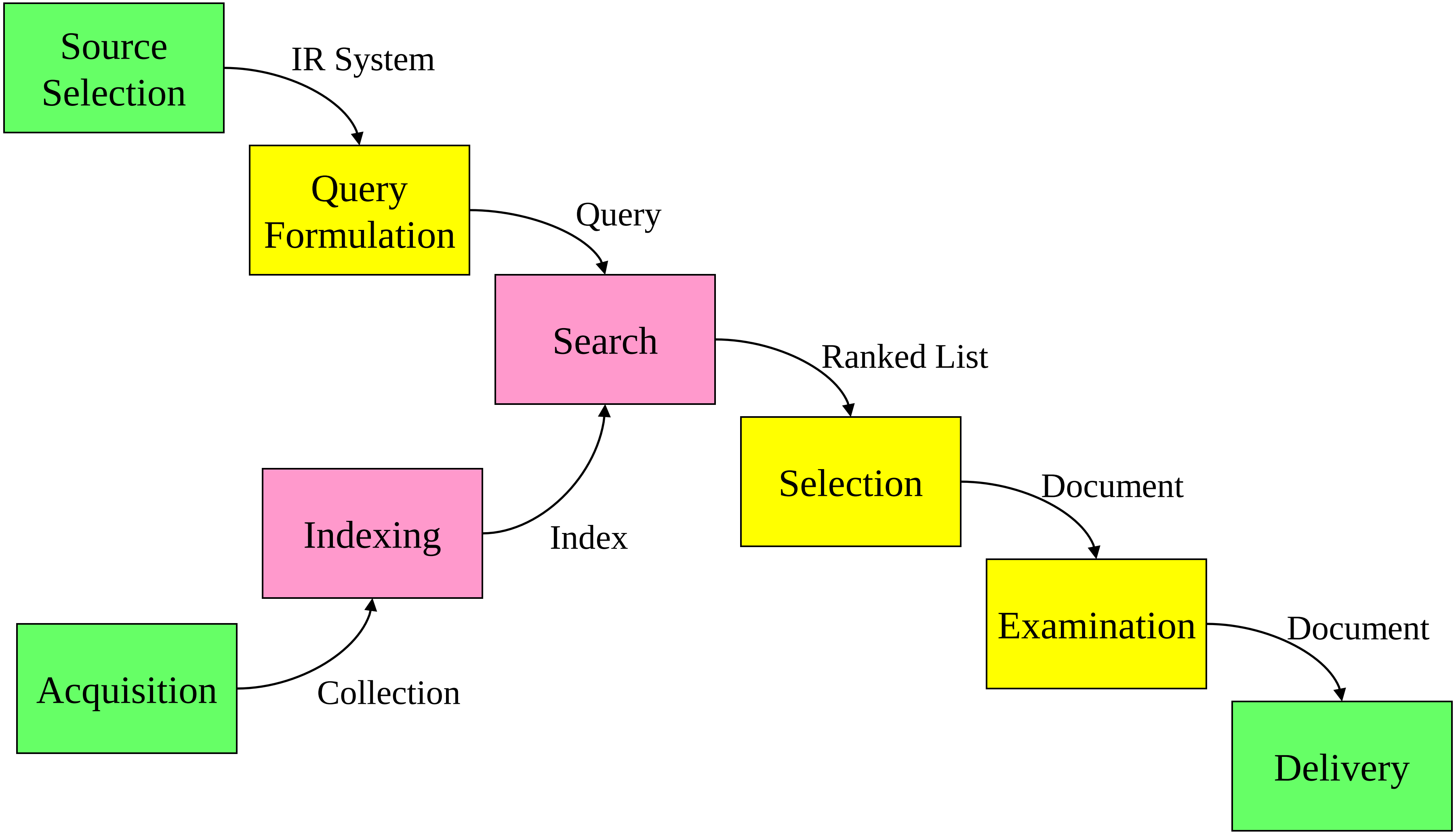
Divide and Conquer

- Strategy: use encapsulation to limit complexity
- Approach:
  - Define interfaces (input and output) for each component
  - Define the functions performed by each component
  - Build each component (in isolation)
  - See how well each component works
    - Then redefine interfaces to exploit strengths / cover weakness
  - See how well it all works together
    - Then refine the design to account for unanticipated interactions
- Result: a hierarchical decomposition

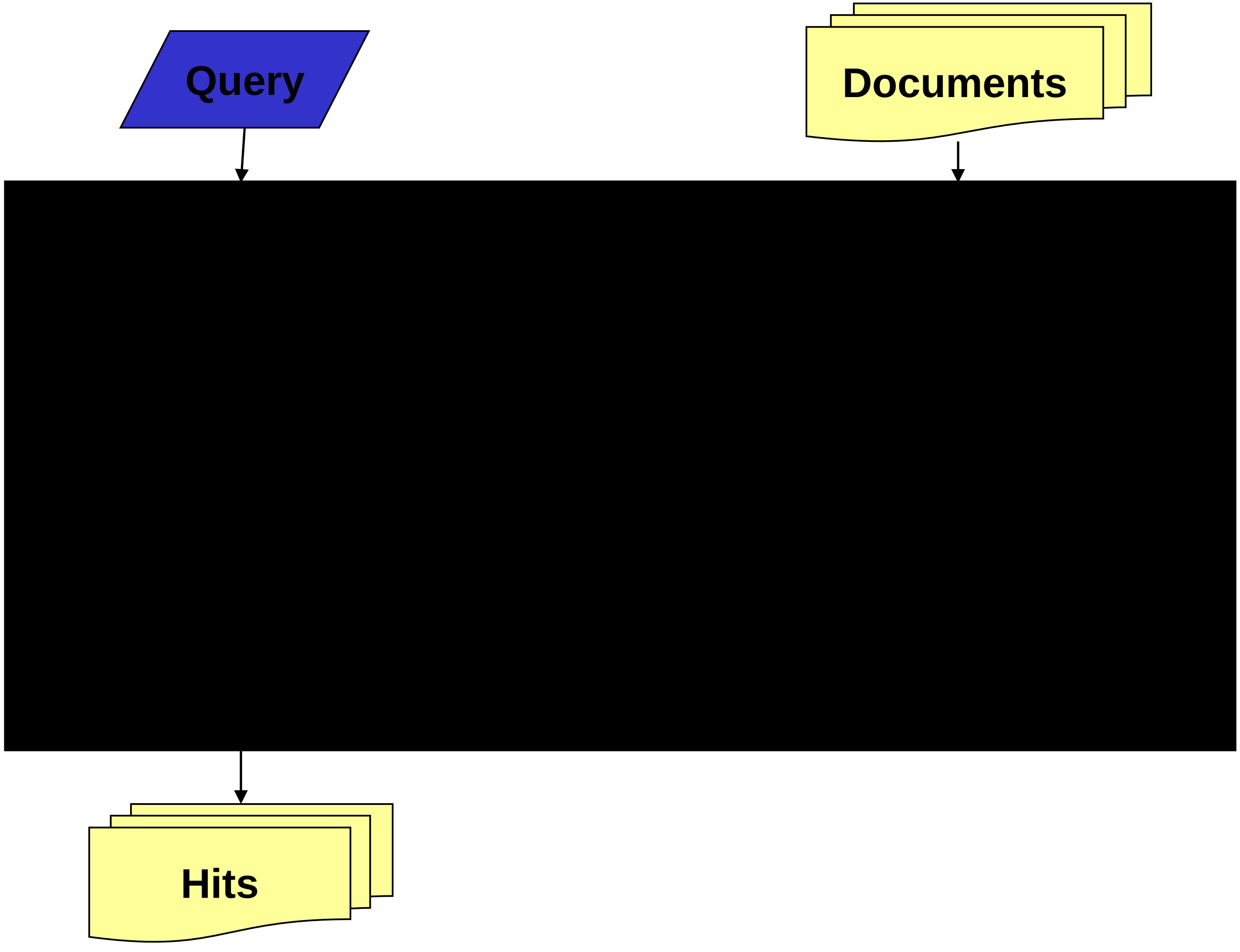
Supporting the Search Process



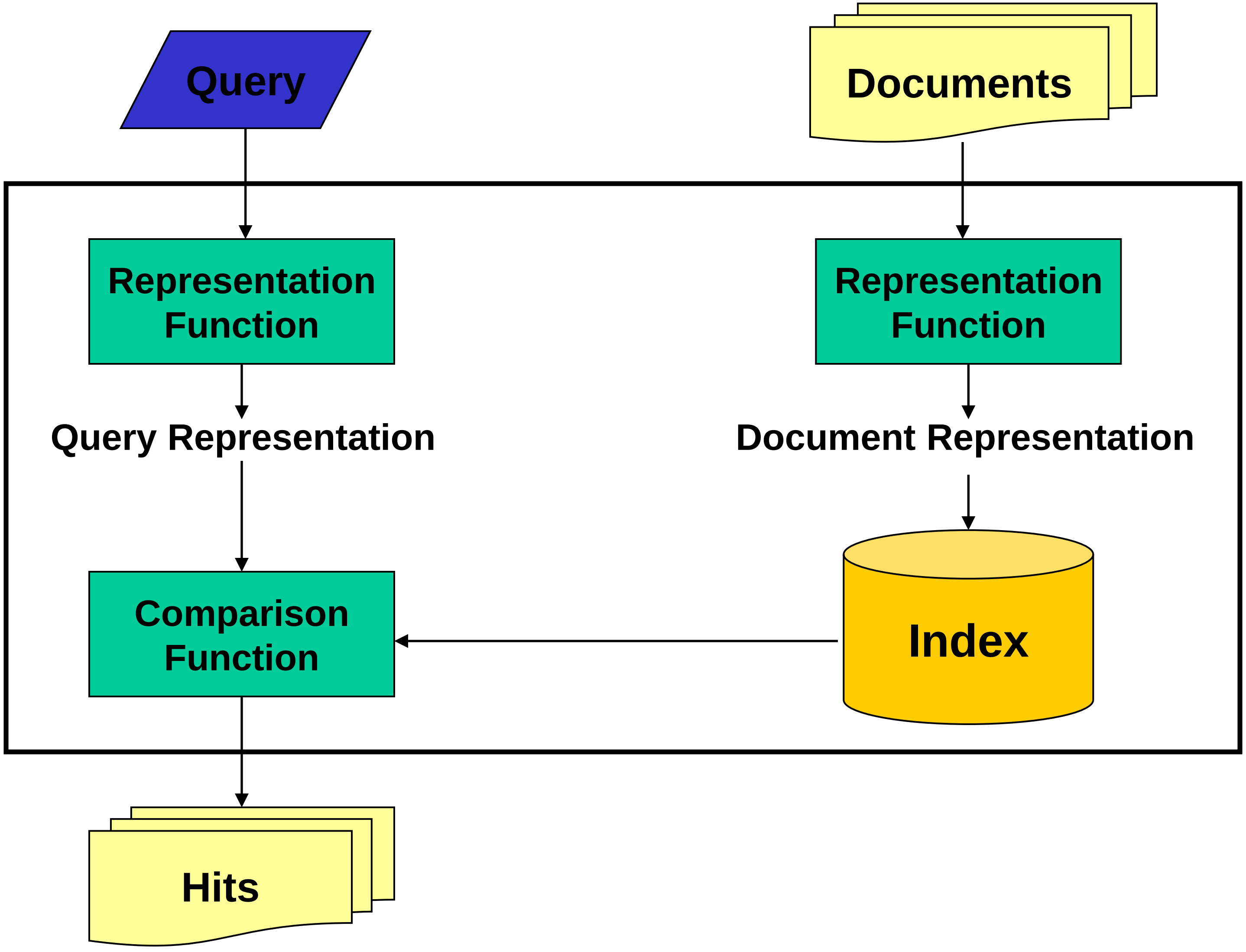
Supporting the Search Process



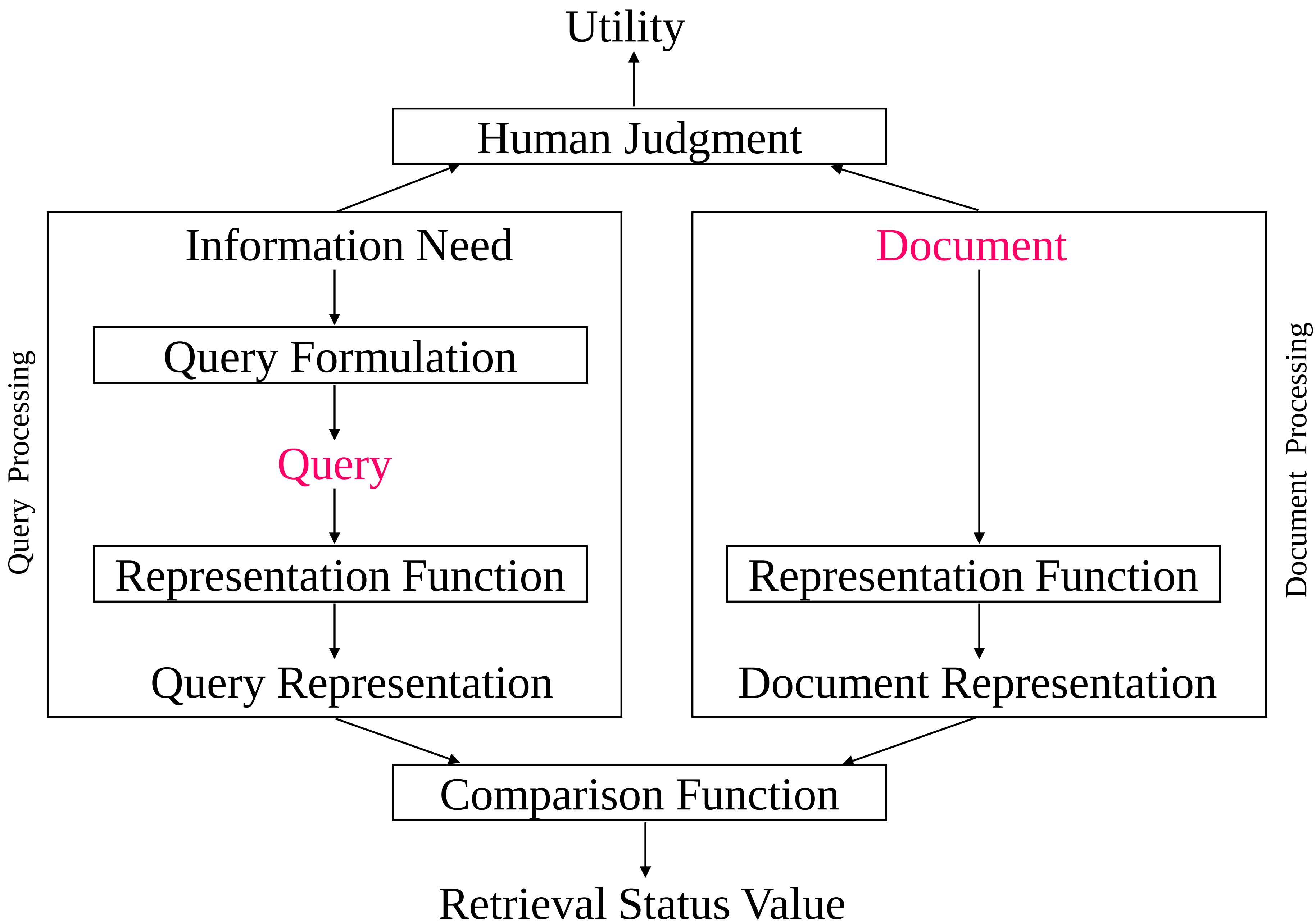
The IR Black Box



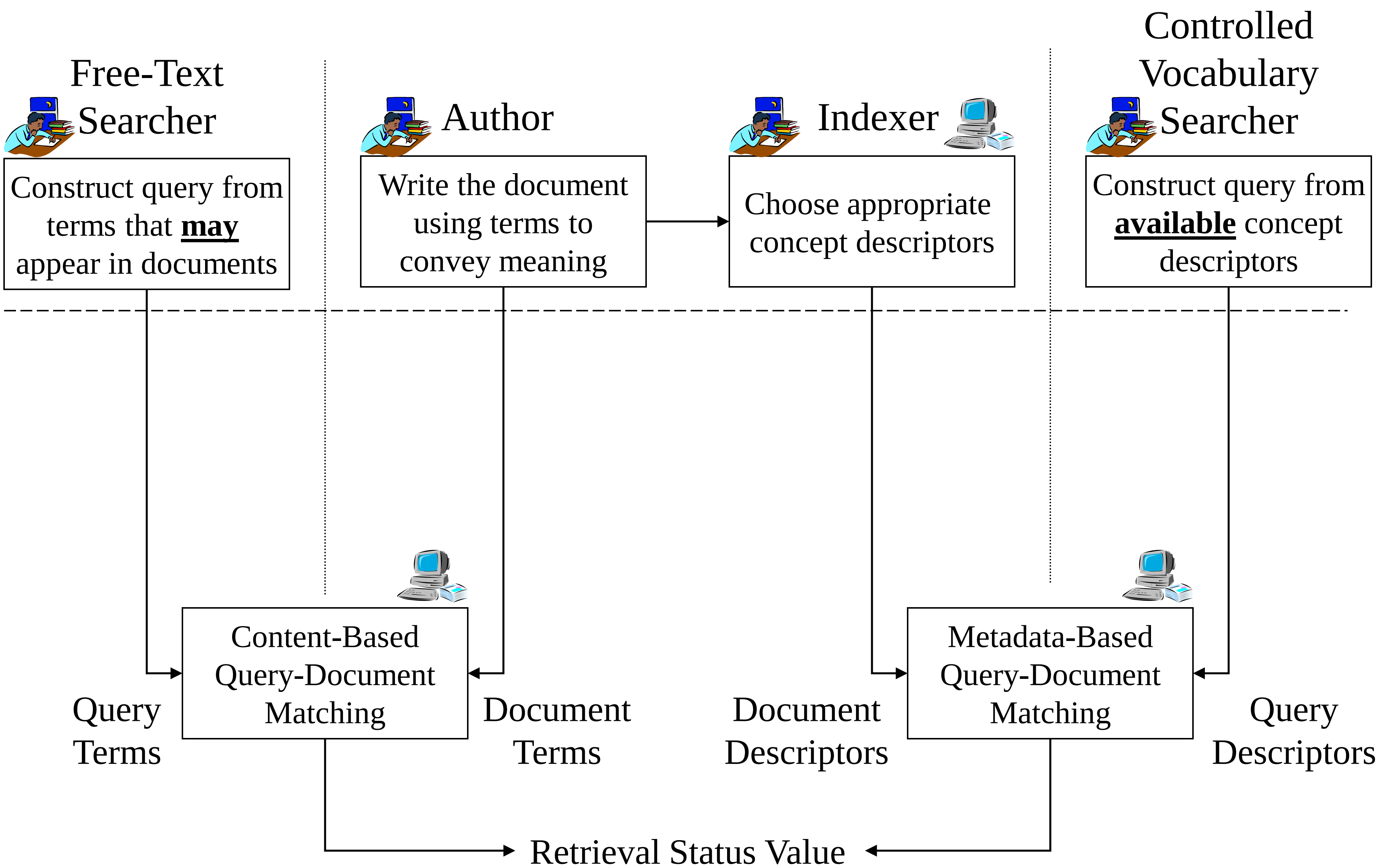
Inside The IR Black Box



Search Component Model



Two Ways of Searching



Counting Terms

- Terms tell us about documents
  - If “rabbit” appears a lot, it may be about rabbits
- Documents tell us about terms
  - “the” is in every document -- not discriminating
- Documents are most likely described well by <u>rare</u> terms that occur in them <u>frequently</u>
  - Higher “term frequency” is stronger evidence
  - Low “document frequency” makes it stronger still

“Bag of Terms” Representation

- Bag = a “set” that can contain duplicates
  - “The quick brown fox jumped over the lazy dog’s back” → *{back, brown, dog, fox, jump, lazy, over, quick, the, the}*
- Vector = values recorded in any consistent order
  - *{back, brown, dog, fox, jump, lazy, over, quick, the, the}* → [1 1 1 1 1 1 1 1 2]

Bag of Terms Example

Document 1

The quick brown fox jumped over the lazy dog's back.

Document 2

Now is the time for all good men to come to the aid of their party.

Term	Document 1	Document 2
aid	0	1
all	0	1
back	1	0
brown	1	0
come	0	1
dog	1	0
fox	1	0
good	0	1
jump	1	0
lazy	1	0
men	0	1
now	0	1
over	1	0
party	0	1
quick	1	0
their	0	1
time	0	1

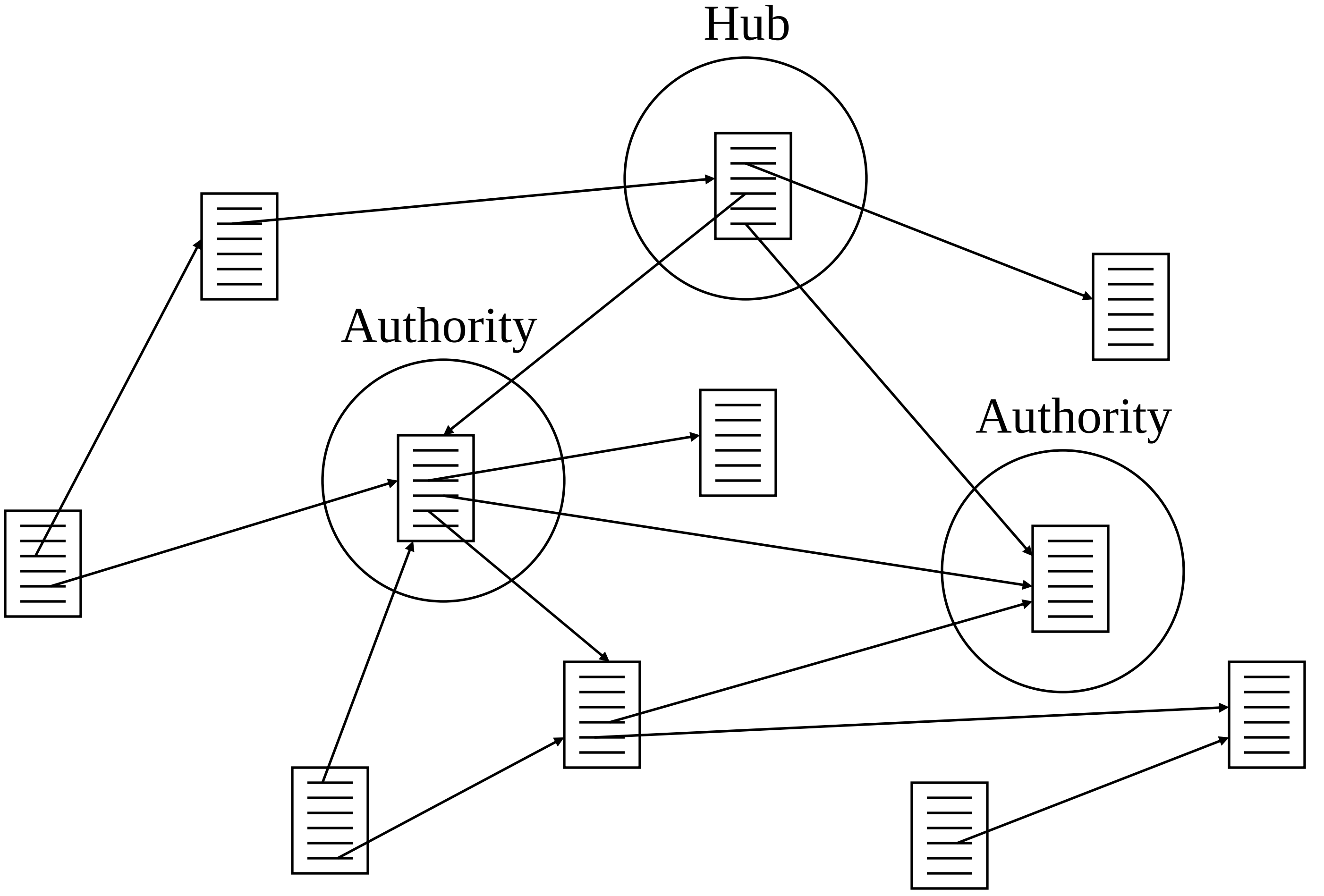
Stopword List

for
is
of
the
to

Representing Behavior

Behavior Category	Minimum Scope: Segment	Minimum Scope: Object	Minimum Scope: Class
Examine	View Listen	Select	
Retain	Print	Bookmark Save Purchase Delete	Subscribe
Reference	Copy / paste Quote	Forward Reply Link Cite	
Annotate	Mark up	Rate Publish	Organize

Learning From Linking Behavior



Putting It All Together

	Free Text	**Behavior**	**Metadata**
Topicality			
Quality			
Reliability			
Cost			
Flexibility			

Course Goals

- Appreciate IR system capabilities and limitations
- Understand IR system design & implementation
  - For a broad range of applications and media
- Evaluate IR system performance
- Identify current IR research problems

Course Design

- Readings provide background and detail
  - At least one recommended reading is required
- Class provides organization and direction
  - We will <u>not</u> cover every detail
- Assignments and project provide experience
- Final exam helps focus your effort

Assumed Background

- Everyone:
  - LBSC 690 or INFM 603 or equivalent
  - Comfortable with learning about technology
- MIM Students:
  - Basic systems analysis, scripting languages
  - Some programming is helpful
- MLS students:
  - LBSC 650 and LBSC 670
  - LBSC 750 or a subject access course is helpful

Grading

- Assignments (20%)
  - Mastery of concepts and experience using tools
- Term project (50%)
  - Options are described on course Web page
- Final exam (30%)
  - In-class exam

Handy Things to Know

- Classes will (hopefully!) be recorded
- Office hours: 5 PM Wednesdays
  - Or schedule by email, or ask after class
- Everything is on the Web
  - http://terpconnect.umd.edu/~oard
- I am most easily reached by email
  - oard@umd.edu

Some Things to Do This Week

- Assignment 1
  - Due at 6 PM next Wednesday!!
- Do the reading **<u>before</u>** class
  - Read for ideas, not detail
  - Don’t fall behind!
- Explore the Web site
  - Start thinking about the term project