



College of Information Studies

University of Maryland Hornbake Library Building College Park, MD 20742-4345

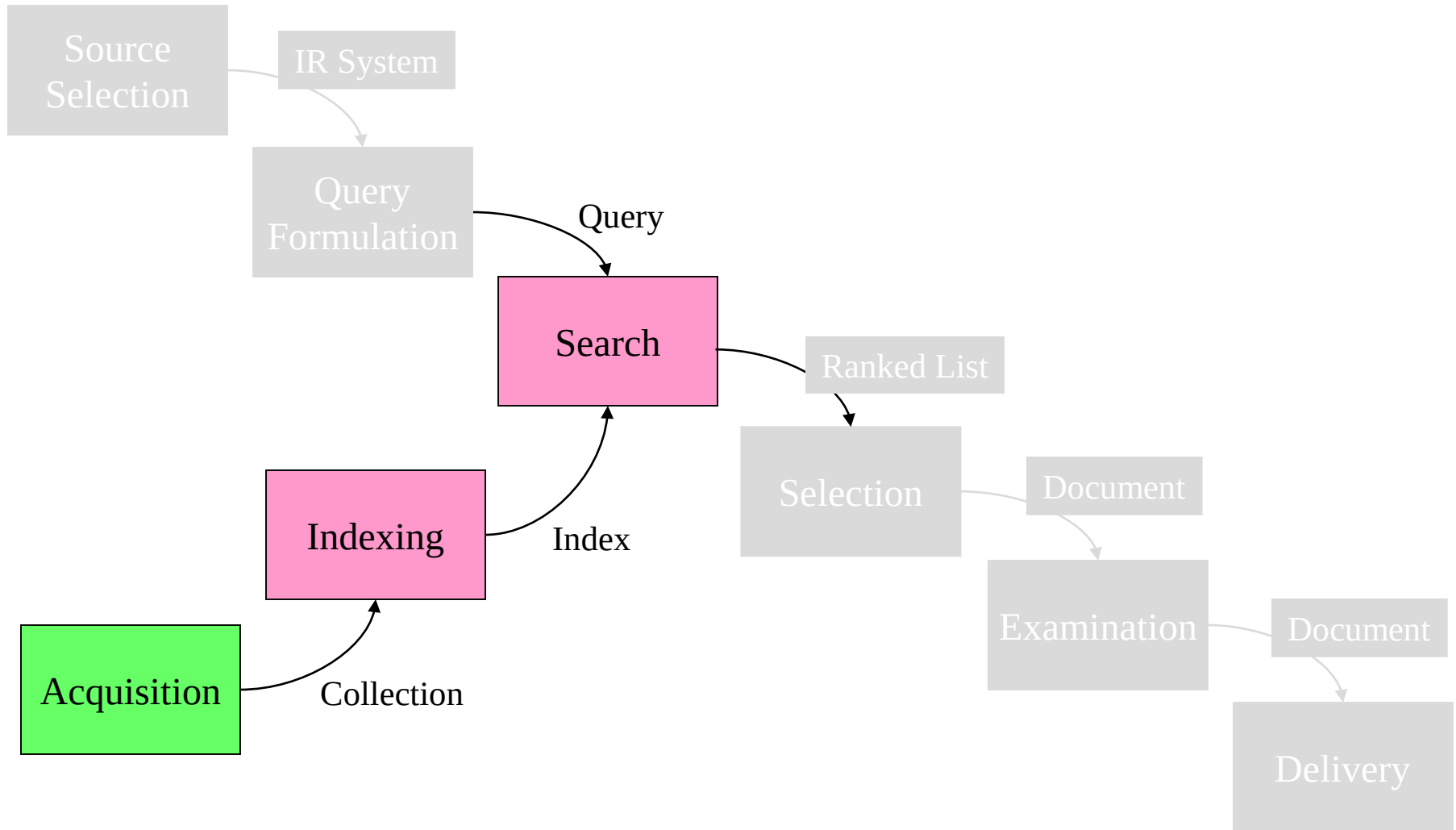
Search Strategies

Session 7

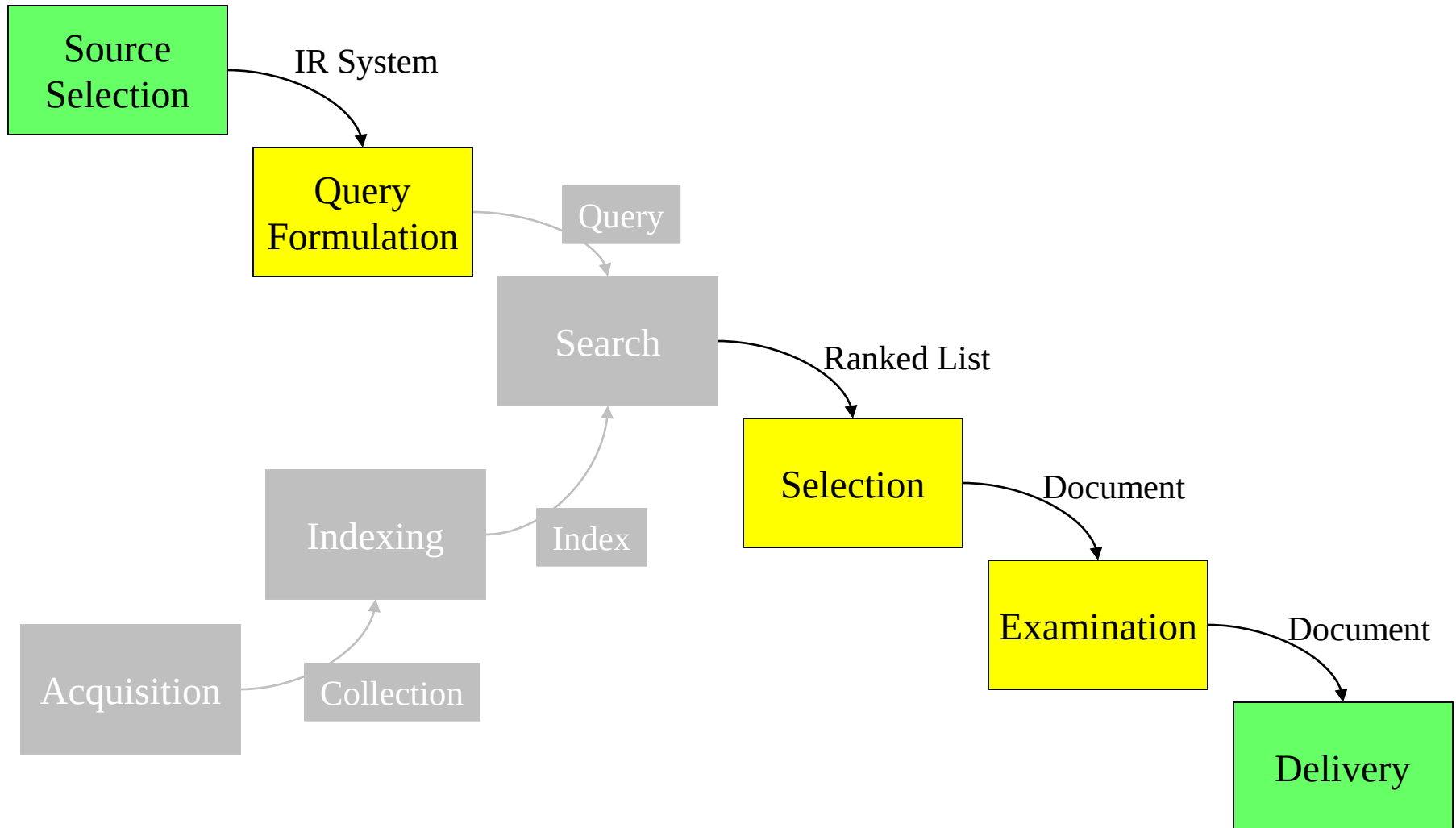
INST 301

Introduction to Information Science

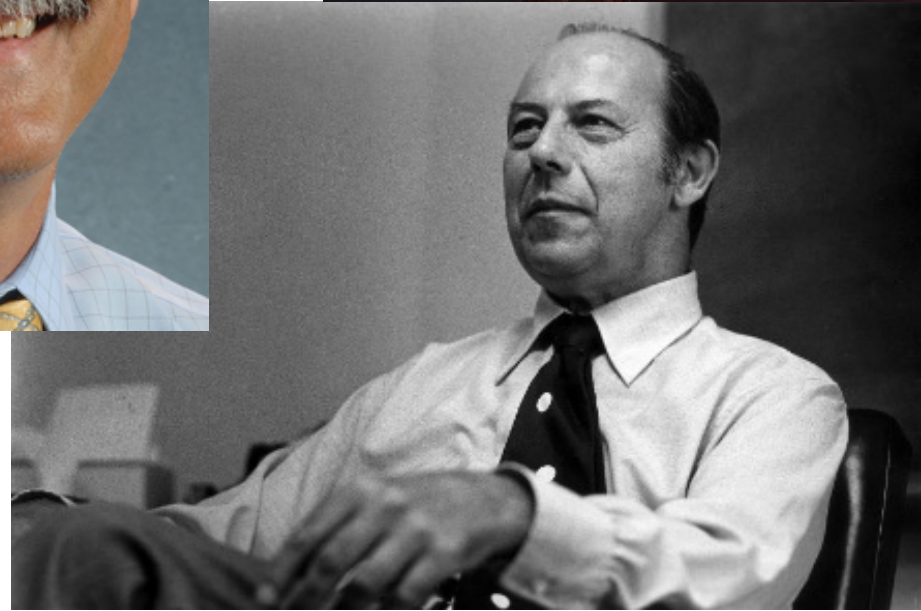
The Search Engine



“The Search”



Search Strategies



Marchionini's Factors Affecting Information Seeking

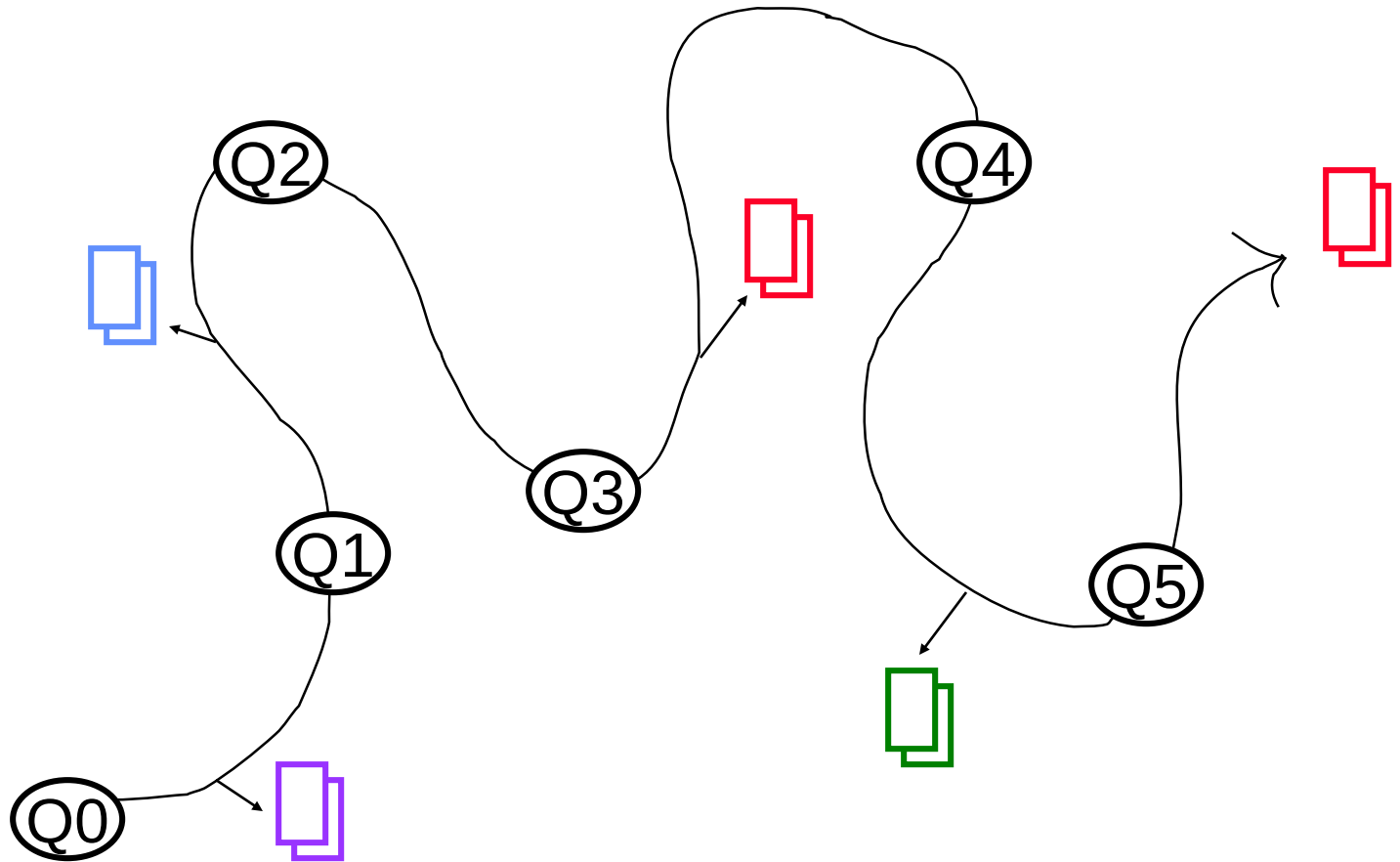
- Information seeker
- Task
- Search system
- Domain
- Setting
- Outcomes

Belkin's ASK: Anomalous State of Knowledge

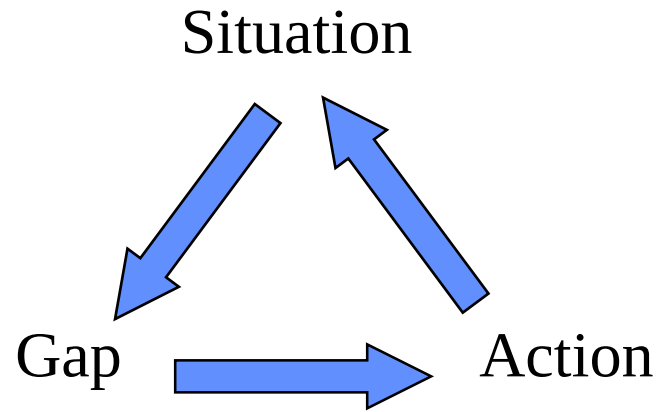
- Searchers do not clearly understand
 - The problem itself
 - What information is needed to solve the problem
- The query results from a clarification process

Bates' “Berry Picking” Model

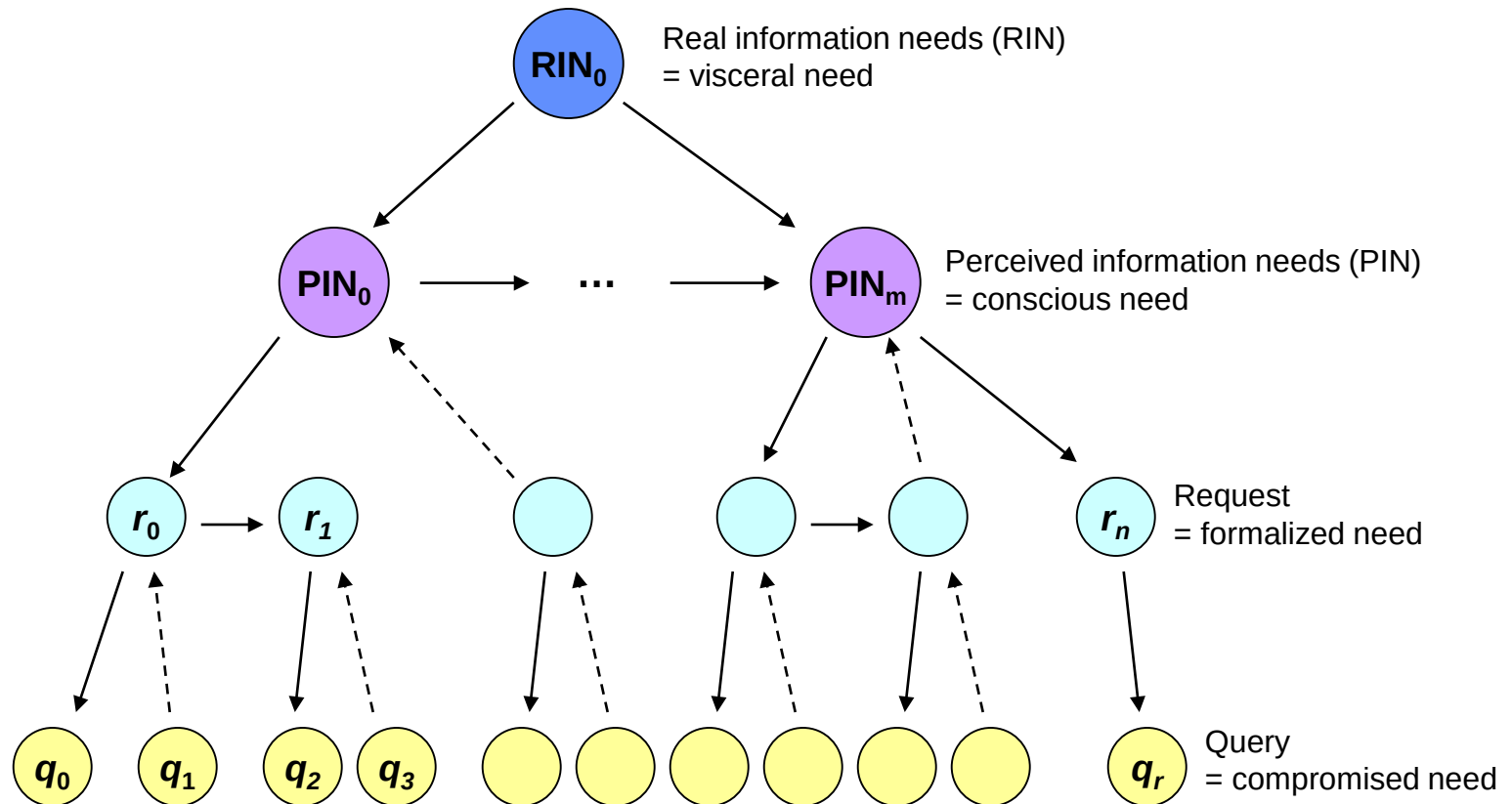
A sketch of a searcher... “moving through many actions towards a general goal of satisfactory completion of research related to an information need.”



Dervin's Sensemaking



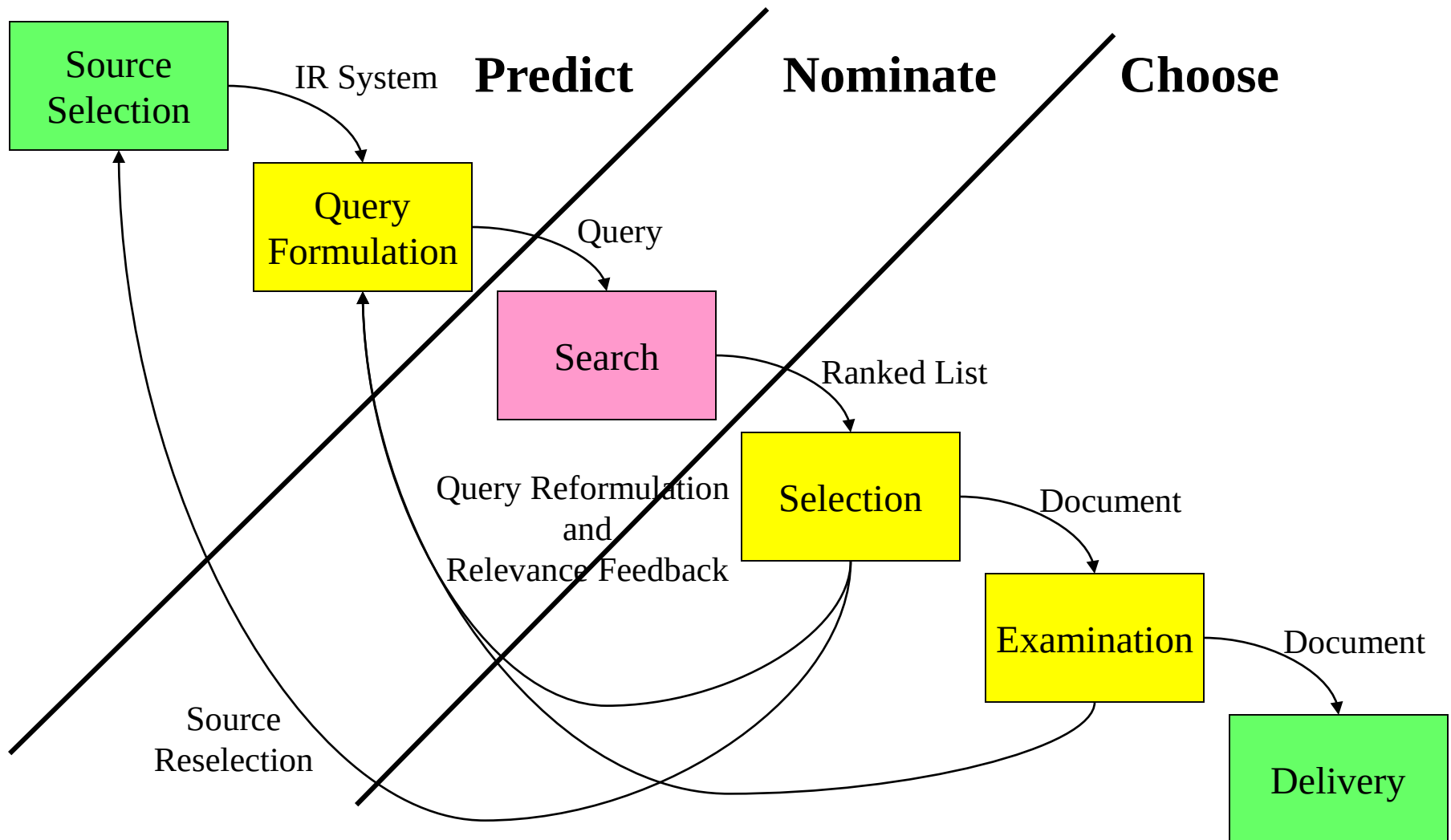
Four Levels of Information Needs



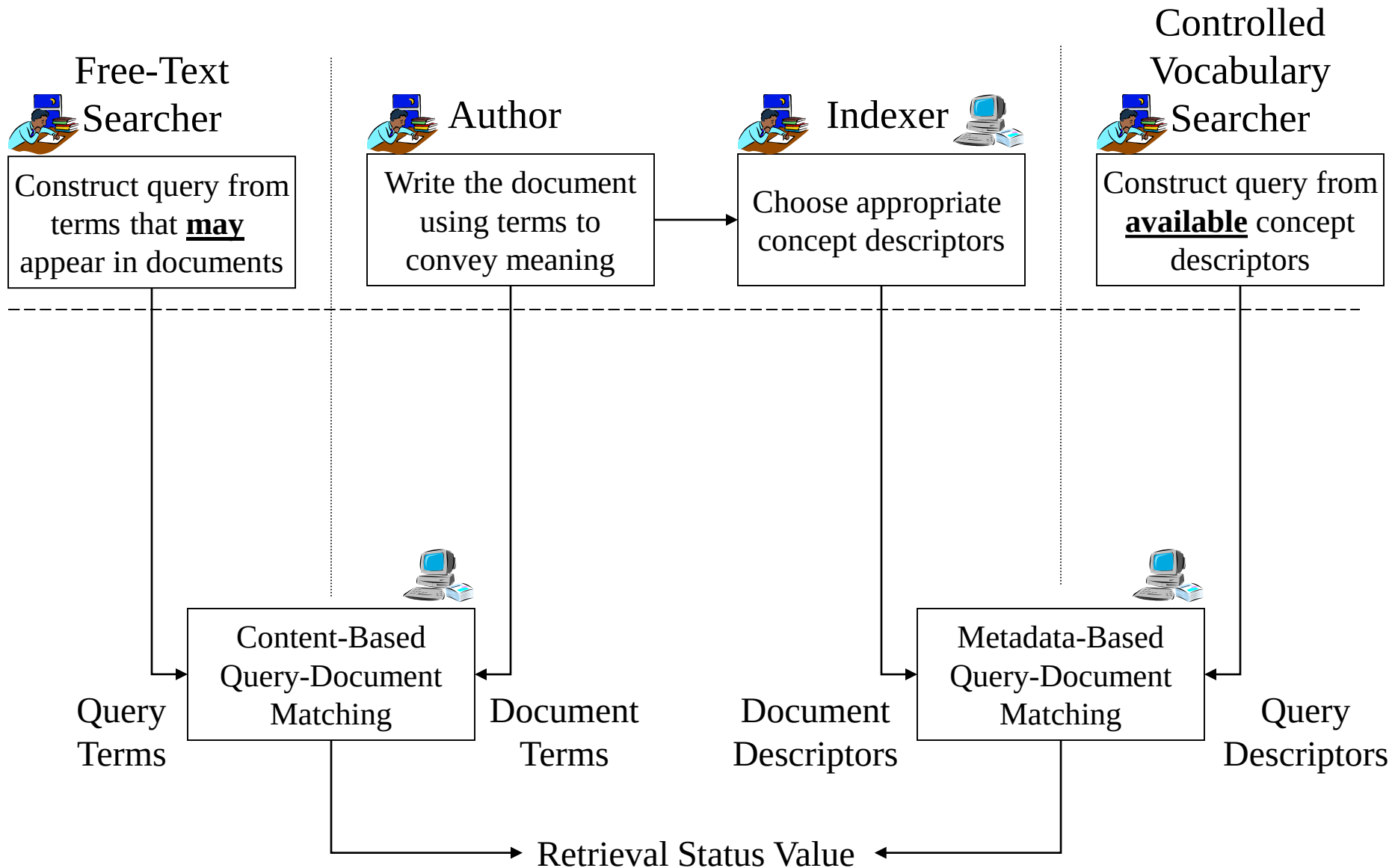
Broder's Web Query Taxonomy

- Informational (~50%)
 - Acquire static information (“topical”)
- Navigational (~20%)
 - Reach a particular site (“known item”)
- Transactional (~30%)
 - Perform a Web-mediated activity (“service”)

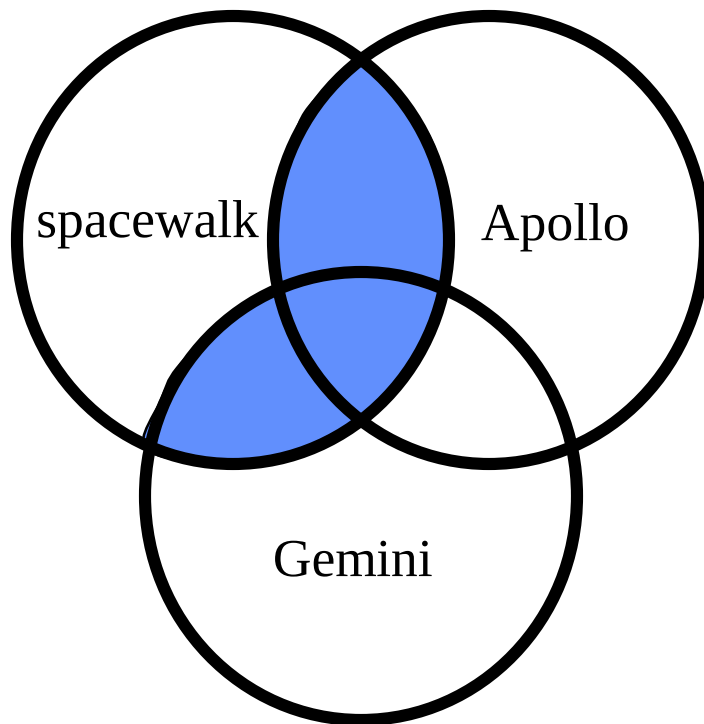
Supporting the Search Process



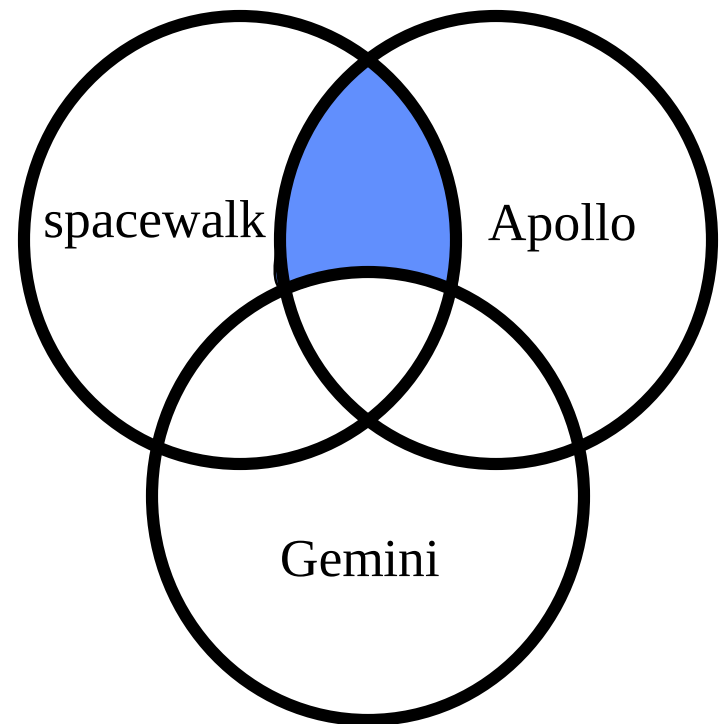
Two Ways of Searching



Boolean Operators



spacewalk AND
(Apollo OR Gemini)



spacewalk AND
Apollo AND
(NOT Gemini)

The Perfect Query Paradox

- Every information need has a perfect document set
 - Finding that set is the goal of search
- Every document set has a perfect query
 - AND every word to get a query for document 1
 - Repeat for each document in the set
 - OR every document query to get the set query
- The problem isn't the system ... it's the query!

Pearl Growing

- Start with a set of relevant documents
- Use them to learn new vocabulary
 - Related terms to help broaden the search
 - Terms that help to remove unrelated senses
- Repeat until you converge on a useful set

Pearl Growing Example

- What is the Moon made of?
- Query: moon
- Initial search reveals:
 - Adding “Apollo” might help focus the search
 - Rejecting “Greek” might avoid unrelated page
- Revised query: +moon -Greek Apollo

28% of Web Queries are Reformulations

**Timeline
(mm:ss)**

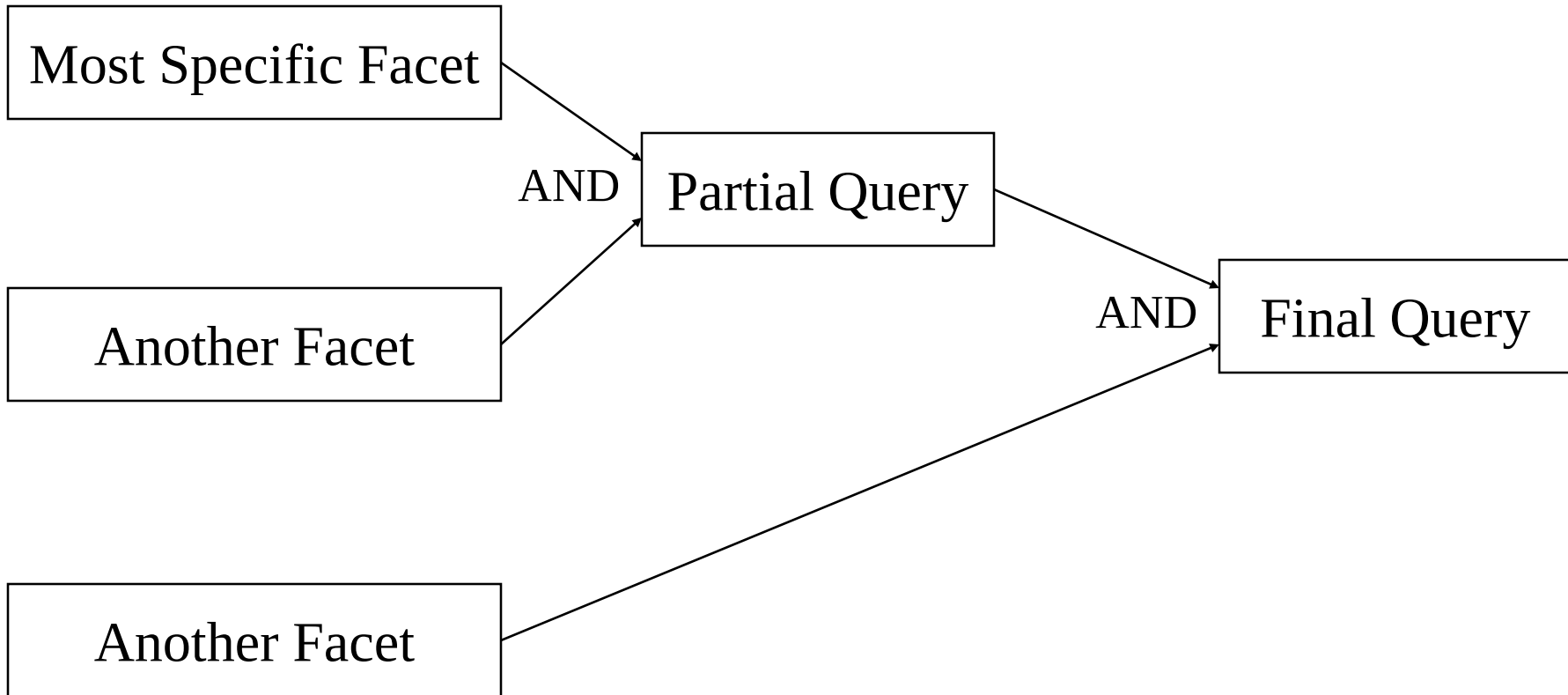
Query

00:00	○	nursing registry
04:18	Ⓢ	certified nursing assistant 1
08:48	Ⓢ	nursing assistant registry
09:48	Ⓢ	license look up for nursing assistants
10:06	Ⓢ	nursing assistant 1 certification
11:42	Ⓢ	nursing assistant 1 license look ups
12:18	Ⓢ	nursing assistant 1 expiration look up
12:30	Ⓢ	nursing registry in Raleigh
13:24	Ⓢ	nursing aide registry of Raleigh
15:00	⊕	nursing aide registry of Raleigh website
16:06	⏪	nursing aide registry of Raleigh
19:48	Ⓢ	north carolina board of nursing information for nursing assistant 1
22:24	Ⓢ	license look up for nursing assistant 1
24:36	Ⓢ	license information for nursing assistant 1 expiration
28:30	Ⓢ	north carolina nursing assistant 1 license information

Concept Analysis

- Identify facets of your question
 - What entities are involved?
 - What attributes of those entities are important?
 - What attribute values do you seek?
- Choose the appropriate search terms
 - What terms might an author have used?
 - Perhaps by using a thesaurus
 - Use initial searches to refine your term choices

Building Blocks



Building Blocks Example

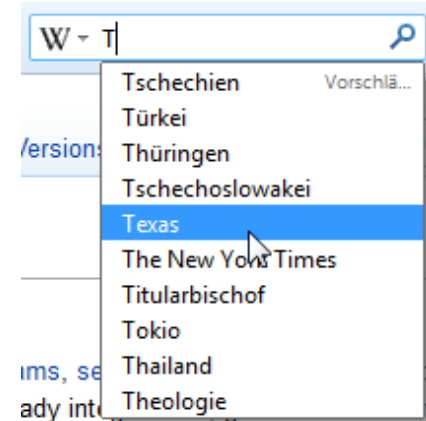
- What is the history of Apartheid?
- Facets?
 - Entity: Racial segregation
 - Location: South Africa
 - Time: Before 1990
- Query construction:
 - (Apartheid OR segregation) AND
("South Africa" OR Pretoria) AND
(history OR review)

Web-specific Strategies

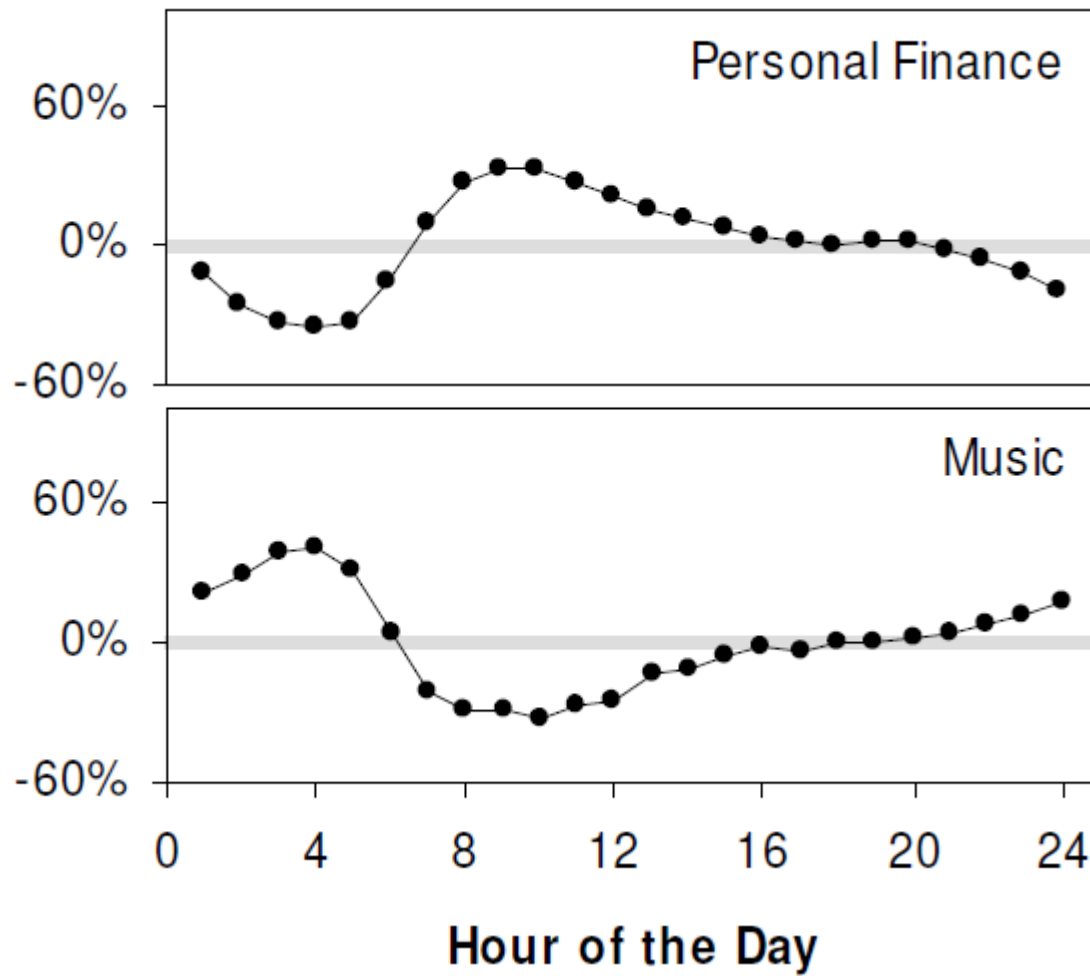
- Using outward links
 - Find good hubs, follow their links
- Using inward links
 - Find good authorities, see who links to them
 - +url:http://terpconnect.umd.edu/~oard/
- URL pruning
 - Discover related work, authors, organization, ...
 - Some servers provide raw directory listings

Query Suggestion

- Predict what the user might type next
 - Learned from behavior of many users
 - Can be customized to a user or group
- Helps w/typos, limited vocabulary, ...
- Provides a basis for auto-completion
 - Particularly useful in difficult input settings



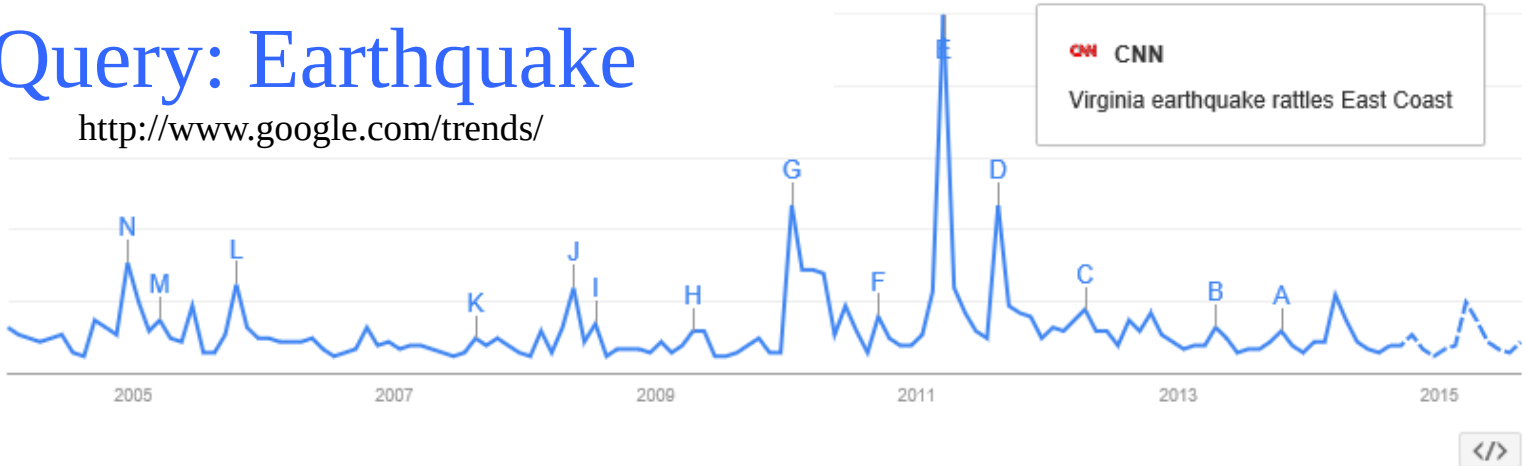
Difference from Daily
Mean Query Frequency



Burstiness

Query: Earthquake

<http://www.google.com/trends/>



Regional interest ?



Region City	
New Zealand	100
Bhutan	57
Fiji	48
Cayman Islands	42
Antigua and Barbuda	36
Philippines	35
St. Vincent & Grenadines	35

▼ View change over time ?



Jan-Jun 2011

</>

</>

Diversity Ranking

- Query ambiguity
 - UPS: United Parcel Service
 - UPS: Uninterruptible power supply
 - UPS: University of Puget Sound
- Query aspects
 - United Parcel Service: store locations
 - United Parcel Service: delivery tracking
 - United Parcel Service: stock price

Try Some Searches

- Using building blocks:
 - Which cities in the former country of Czechoslovakia have pollution problems?
- Using pearl growing:
 - What event in the early 1900's is Noel Davis famous for?

Some Good Advice

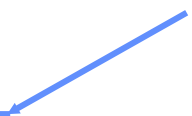
Human-Computer Interaction

- User in control
 - Anticipatable outcomes
 - Explainable results
 - Browsable content
 - Informative feedback
 - Easy reversal
- Limit working memory load
 - Show query context
- Support for learning
 - Novice and expert alternatives
 - Scaffolding

Interactive IR

- Support human reasoning
 - Show actual content
 - Depict uncertainty
 - Be fast
- Use familiar metaphors
 - Timelines, ranked lists, maps, ...
- Some system initiative
 - Loosely guide the process
 - Expose structure of knowledge
- Co-design w/search strategies

Evaluation

- What can be measured that reflects the searcher's ability to use a system? (Cleverdon, 1966)
 - Coverage of Information
 - Form of Presentation
 - Effort required/Ease of Use
 - Time and Space Efficiency
 - Recall
 - Precision
- Effectiveness
- 

Evaluating IR Systems

- User-centered strategy
 - Given several users, and at least 2 retrieval systems
 - Have each user try the same task on both systems
 - Measure which system works the “best”
- System-centered strategy
 - Given documents, queries, and relevance judgments
 - Try several variations on the retrieval system
 - Measure which ranks more good docs near the top

Which is the Best Rank Order?

A.             

B.             

C.             

D.             

E.             

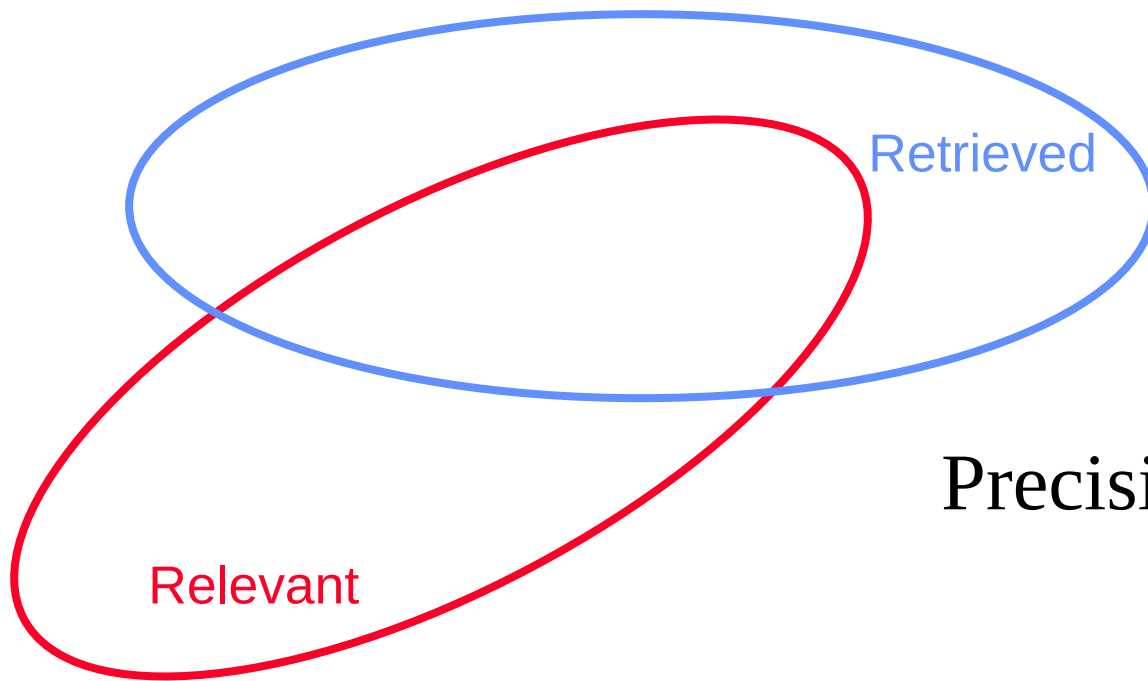
F.             

 = relevant document

Precision and Recall

- Precision
 - How much of what was found is relevant?
 - Often of interest, particularly for interactive searching
- Recall
 - How much of what is relevant was found?
 - Particularly important for law, patents, and medicine

Measures of Effectiveness



$$\text{Precision} = \frac{|\text{Ret} \cap \text{Rel}|}{|\text{Ret}|}$$

$$\text{Recall} = \frac{|\text{Ret} \cap \text{Rel}|}{|\text{Rel}|}$$

Affective Evaluation

- Measure stickiness through frequency of use
 - Non-comparative, long-term
- Key factors (from cognitive psychology):
 - Worst experience
 - Best experience
 - Most recent experience
- Highly variable effectiveness is undesirable
 - Bad experiences are particularly memorable

Before You Go

On a sheet of paper, answer the following (ungraded) question (no names, please):

What was the muddiest point in today's class?