INFM 603: Information Technology and Organizational Context

Session 10: Information Retrieval

Jimmy Lin
The iSchool
University of Maryland

Thursday, November 8, 2012

Information Retrieval

What you search for!

Satisfying an information need
“Scratching an information itch”

User

Process

System

Information

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

Our focus today is on textual information…

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)

- Topical search

  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.

- Open-ended exploration

  Who makes the best chocolates?

  What technologies are available for digital reference desk services?

Retrospective Searches (II)

- Known item search

  Find Jimmy Lin’s homepage.

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

- 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.

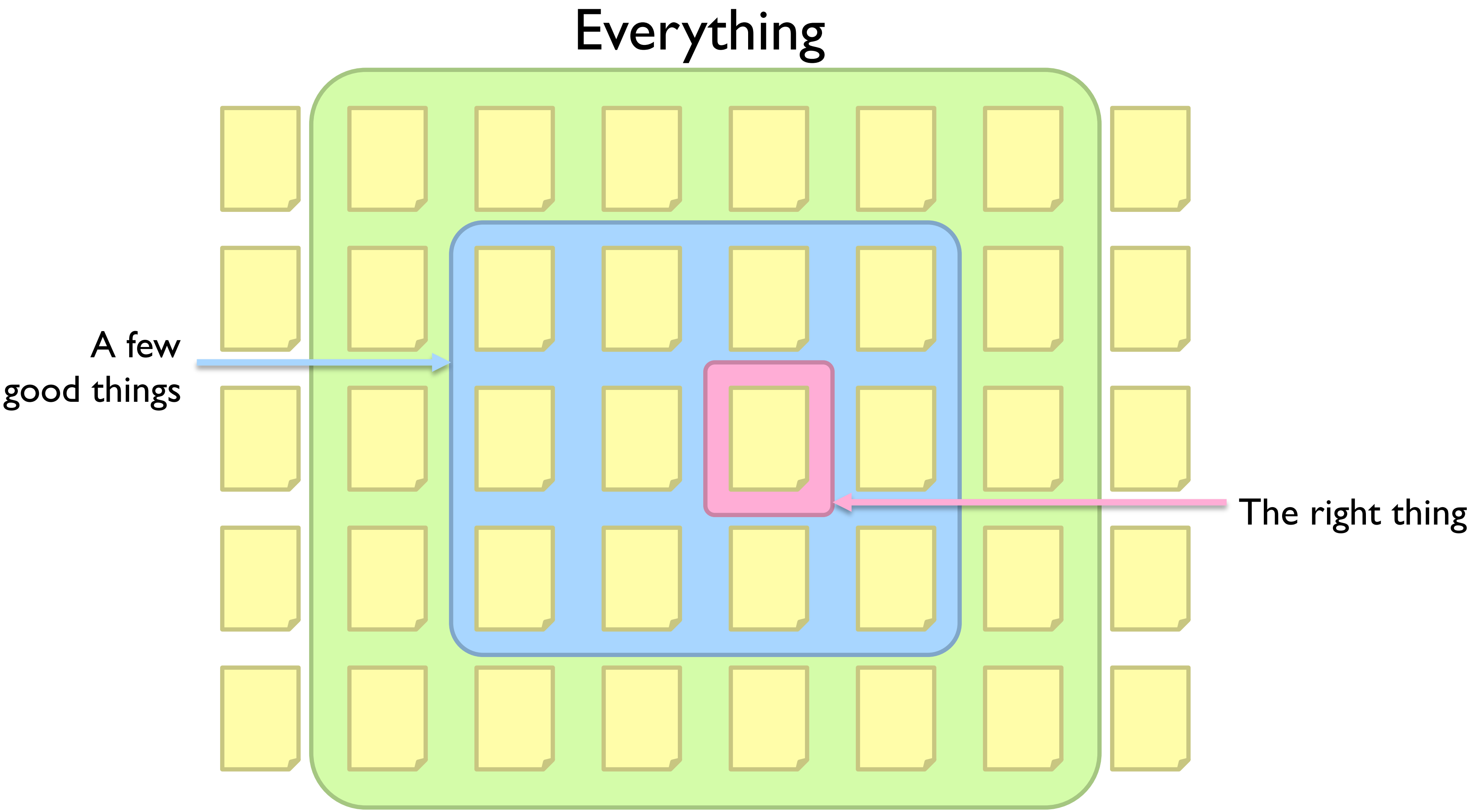
“Definition”

Who is Aaron Copland?
What is a quasar?

Prospective “Searches”

- Filtering
  - Make a binary decision about each incoming document
- Routing
  - Sort incoming documents into different bins

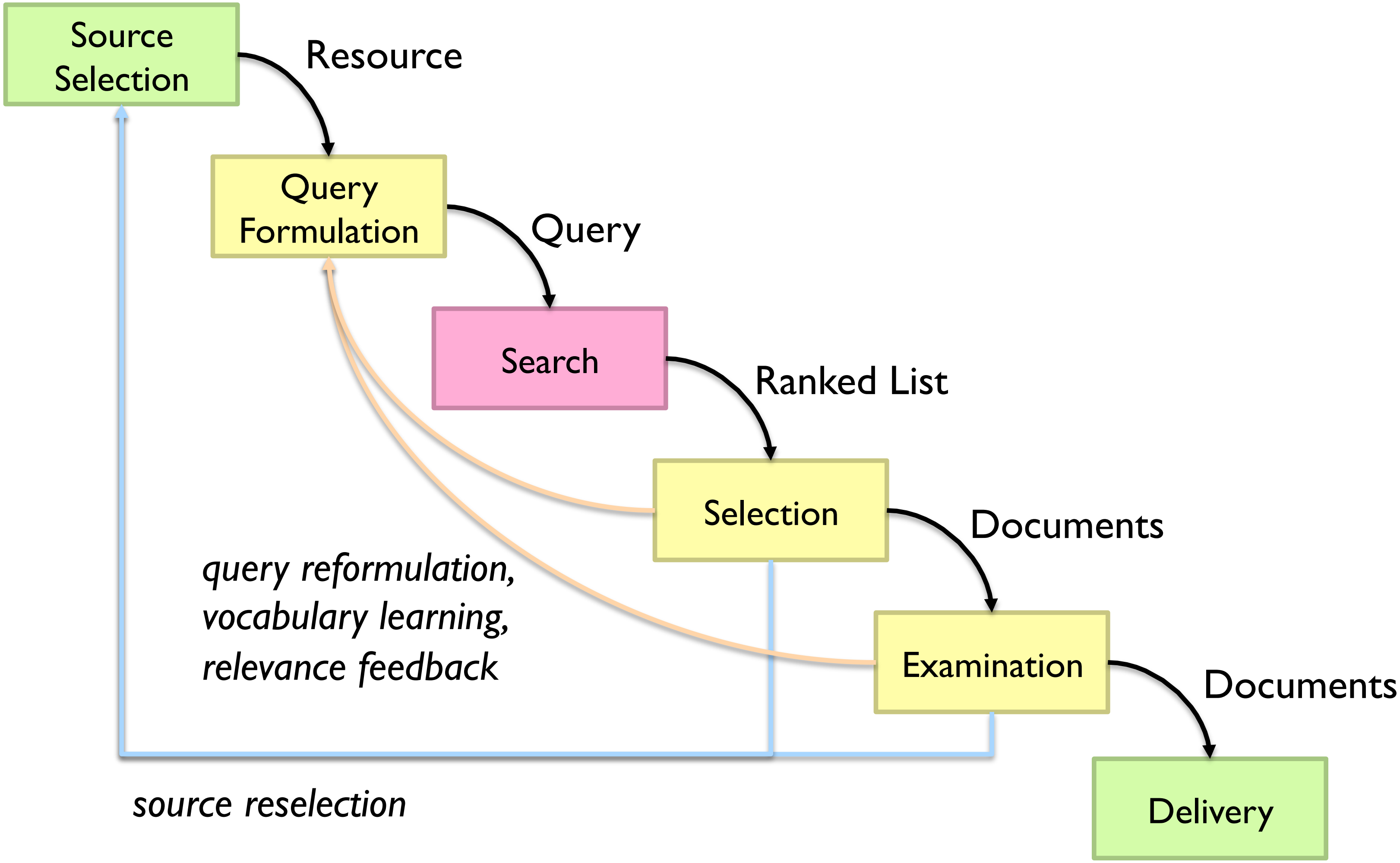
Scope of Information Needs



Relevance

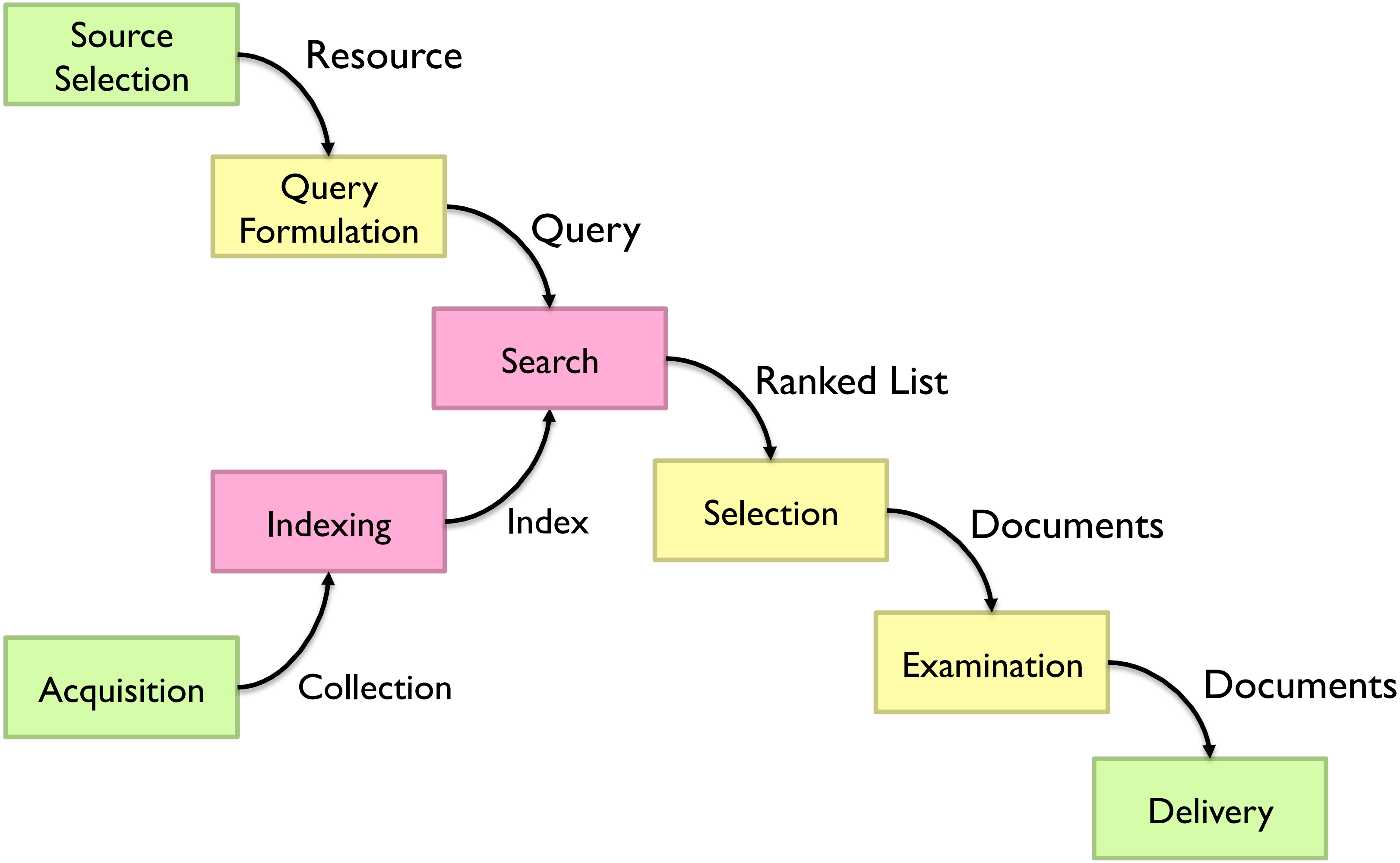
- How well information addresses your needs
  - Harder to pin down than you think!
  - Complex function of user, task, and context
- Types of relevance:
  - Topical relevance: is it about the right thing?
  - Situational relevance: is it useful?

The Information Retrieval Cycle



source reselection

Supporting the Search Process



Spiders, Crawlers, and Robots: Oh My!

The Central Problem in Search

Searcher

Author

Concepts ↓ **Query Terms**

"tragic love story"

⟷

Concepts ↓ **Document Terms**

"fateful star-crossed romance"

Do these represent the same concepts?

Ambiguity
Synonymy
Polysemy
Morphology
Paraphrase
Anaphora
Pragmatics

How do we represent documents?

- Remember: computers don’t “understand” anything!
- “Bag of words” representation:
  - Break a document into words
  - Disregard order, structure, meaning, etc. of the words
  - Simple, yet effective!

Boolean Text Retrieval

- Keep track of which documents have which terms
- Queries specify constraints on search results
  - a AND b: document must have both terms “a” and “b”
  - a OR b: document must have either term “a” or “b”
  - NOT a: document must not have term “a”
  - Boolean operators can be arbitrarily combined
- Results are not ordered!

Index Structure

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.

Stopword List

for
is
of
the
to

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

Boolean Searching

Term	Document							
	1	2	3	4	5	6	7	8
aid	0	0	0	1	0	0	0	1
all	0	1	0	1	0	1	0	0
back	1	0	1	0	0	0	1	0
brown	1	0	1	0	1	0	1	0
come	0	1	0	1	0	1	0	1
dog	0	0	1	0	1	0	0	0
fox	0	0	1	0	1	0	1	0
good	0	1	0	1	0	1	0	1
jump	0	0	1	0	0	0	0	0
lazy	1	0	1	0	1	0	1	0
men	0	1	0	1	0	0	0	1
now	0	1	0	0	0	1	0	1
over	1	0	1	0	1	0	1	1
party	0	0	0	0	0	1	0	1
quick	1	0	1	0	0	0	0	0
their	1	0	0	0	1	0	1	0
time	0	1	0	1	0	1	0	0

- dog AND fox
  - Doc 3, Doc 5
- dog NOT fox
  - Empty
- fox NOT dog
  - Doc 7
- dog OR fox
  - Doc 3, Doc 5, Doc 7
- good AND party
  - Doc 6, Doc 8
- good AND party NOT over
  - Doc 6

Extensions

- Stemming (“truncation”)
  - Technique to handle morphological variations
  - Store word stems: love, loving, loves … → lov
- Proximity operators
  - More precise versions of AND
  - Store a list of positions for each word in each document

Why Boolean Retrieval Works

- Boolean operators approximate natural language
- AND can specify relationships between concepts
  - good party
- OR can specify alternate terminology
  - excellent party
- NOT can suppress alternate meanings
  - Democratic party

Why Boolean Retrieval Fails

- Natural language is way more complex
- AND “discovers” nonexistent relationships
  - Terms in different paragraphs, chapters, …
- Guessing terminology for OR is hard
  - good, nice, excellent, outstanding, awesome, …
- Guessing terms to exclude is even harder!
  - Democratic party, party to a lawsuit, …

Strengths and Weaknesses

- Strengths
  - Precise, if you know the right strategies
  - Precise, if you have an idea of what you're looking for
  - Implementations are fast and efficient
- Weaknesses
  - Users must learn Boolean logic
  - Boolean logic insufficient to capture the richness of language
  - No control over size of result set: either too many hits or none
  - When do you stop reading? All documents in the result set are considered "equally good"
  - What about partial matches? Documents that "don't quite match" the query may be useful also

Ranked Retrieval Paradigm

- Pure Boolean systems provide no ordering of results
  - … but some documents are more relevant than others!
- "Best-first" ranking can be superior
  - Select *n* documents
  - Put them in order, with the "best" ones first
  - Display them one screen at a time
  - Users can decided when they want to stop reading

"Best-first"? Easier said than done!

Extending Boolean retrieval:
Order results based on number of matching terms

a AND *b* AND *c*

What if multiple documents have the same number of matching terms?
What if no single document matches the query?

Similarity-Based Queries

- Treat both documents and queries as "bags of words"
  - Assign a weight to each word
- Find the similarity between the query and each document
  - Compute similarity based on weights of the words
- Rank order the documents by similarity
  - Display documents most similar to the query first

Surprisingly, this works pretty well!

Term Weighting

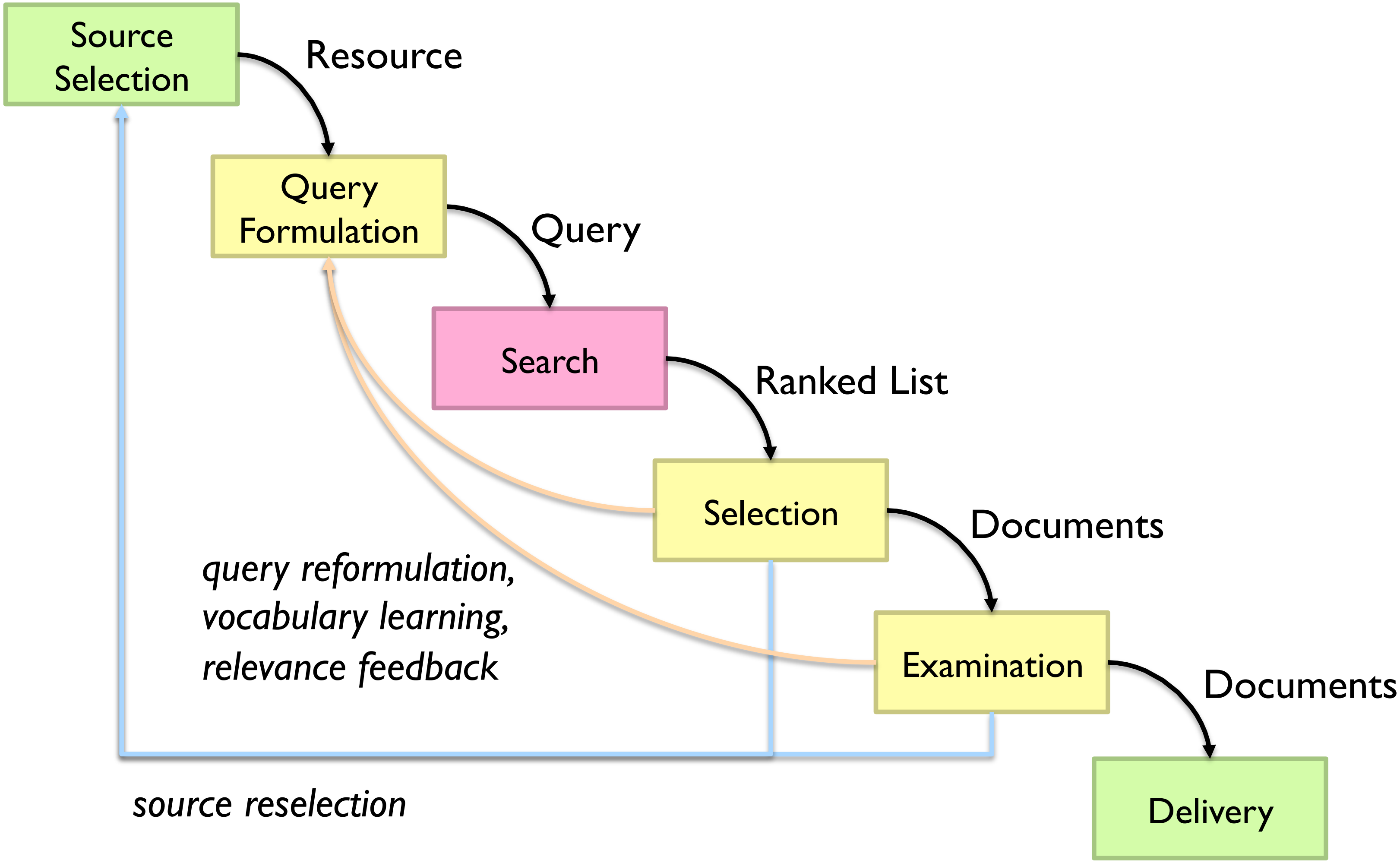
- Term weights consist of two components
  - Local: how important is the term in this doc?
  - Global: how important is the term in the collection?
- Here’s the intuition:
  - Terms that appear often in a document should get high weights
  - Terms that appear in many documents should get low weights
- How do we capture this mathematically?
  - Term frequency (local)
  - Inverse document frequency (global)

TF.IDF Term Weighting

$$w_{i,j} = \mathrm{tf}_{i,j} \cdot \log \frac{N}{n_i}$$

$w_{i,j}$	weight assigned to term *i* in document *j*
$\mathrm{tf}_{i,j}$	number of occurrence of term *i* in document *j*
N	number of documents in entire collection
n_i	number of documents with term *i*

The Information Retrieval Cycle



Search Output

- What now?
  - User identifies relevant documents for “delivery”
  - User issues new query based on content of result set
- What can the system do?
  - Assist the user to identify relevant documents
  - Assist the user to identify potentially useful query terms

Selection Interfaces

- One dimensional lists
  - What to display? title, source, date, summary, ratings, ...
  - What order to display? similarity score, date, alphabetic, ...
  - How much to display? number of hits
  - Other aids? related terms, suggested queries, …
- Two+ dimensional displays
  - Clustering, projection, contour maps, VR
  - Navigation: jump, pan, zoom

Query Enrichment

- Relevance feedback
  - User designates “more like this” documents
  - System adds terms from those documents to the query
- Manual reformulation
  - Initial result set leads to better understanding of the problem domain
  - New query better approximates information need
- Automatic query suggestion

Example Interfaces

- Google
- Amazon
- Clusty
- PubMed

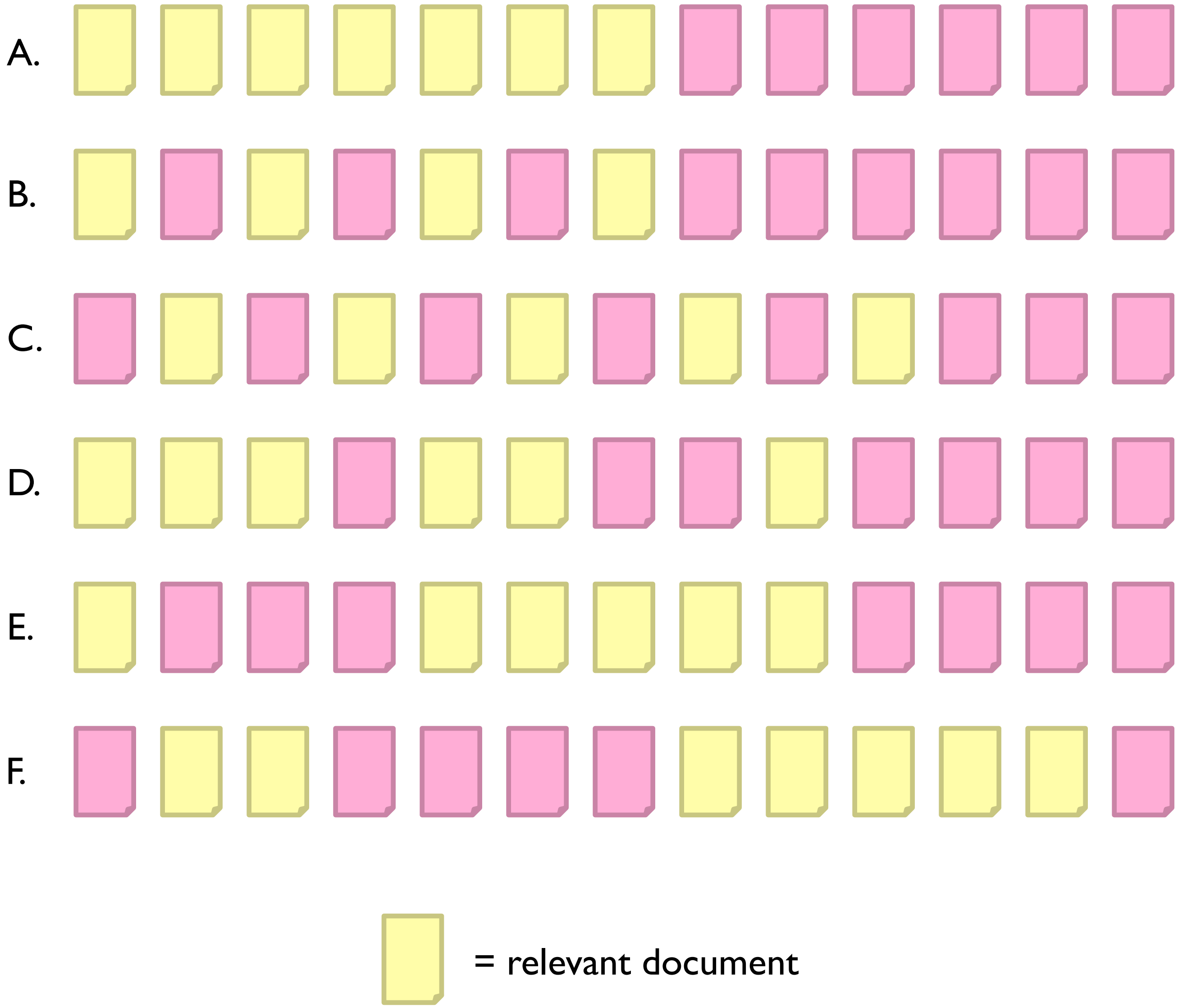
Evaluating IR Systems

- User-centered strategy
  - Recruit several users
  - Observe each user working with one or more retrieval systems
  - Measure which system works the “best”
- System-centered strategy
  - Given documents, queries, and relevance judgments
  - Try several variant of the retrieval method
  - Measure which variant is more effective

Good Effectiveness Measures

- Capture some aspect of what the user wants
- Have predictive value for other situations
- Easily replicated by other researchers
- Easily compared

Which is the Best Rank Order?



Precision and Recall

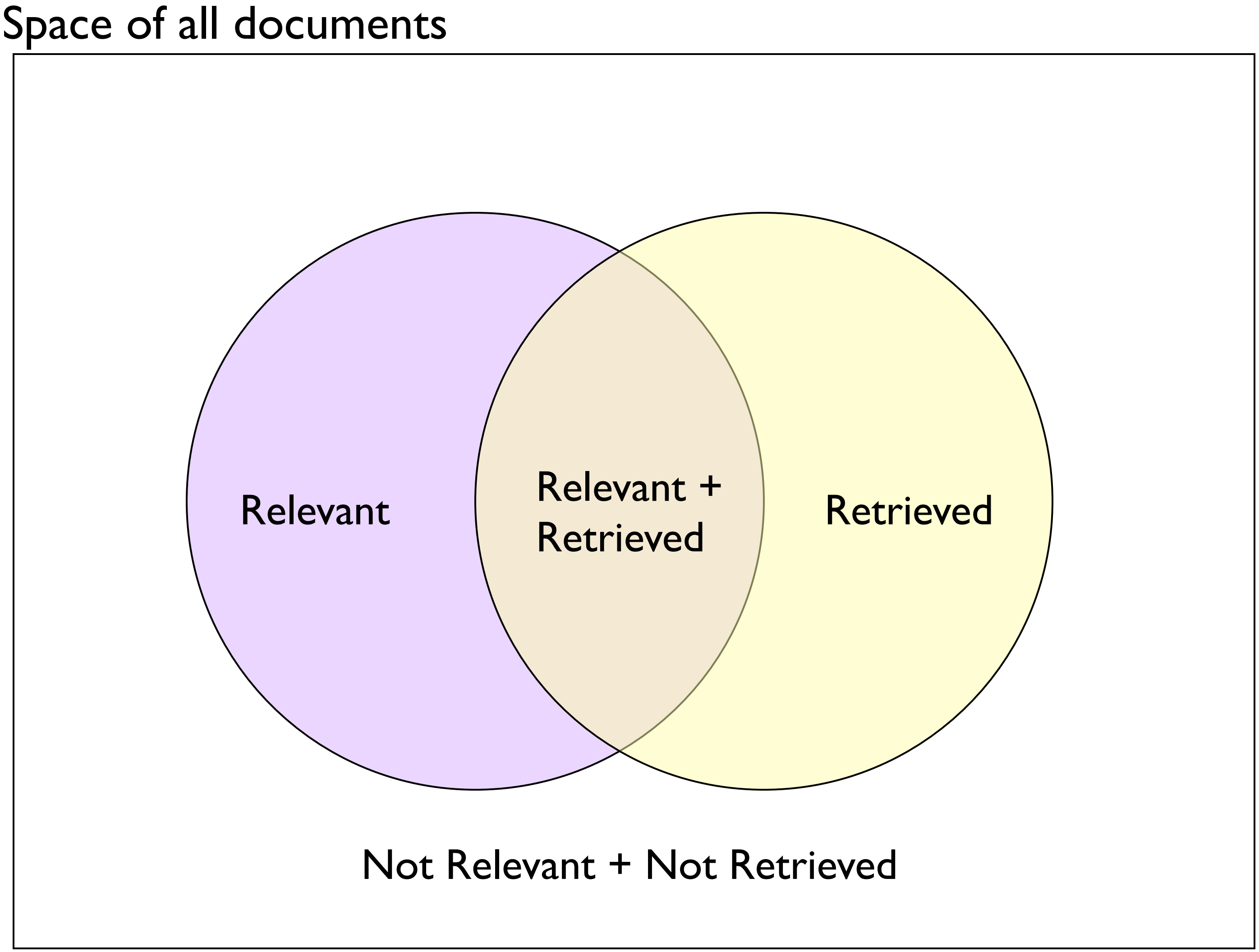
	Relevant	Not relevant
Retrieved	A	B
Not retrieved	C	D

Collection size = A+B+C+D
Relevant = A+C
Retrieved = A+B

Precision = A / (A+B)
Recall = A / (A+C)

When is precision important?
When is recall important?

Another View



Space of all documents

Relevant

Relevant +
Retrieved

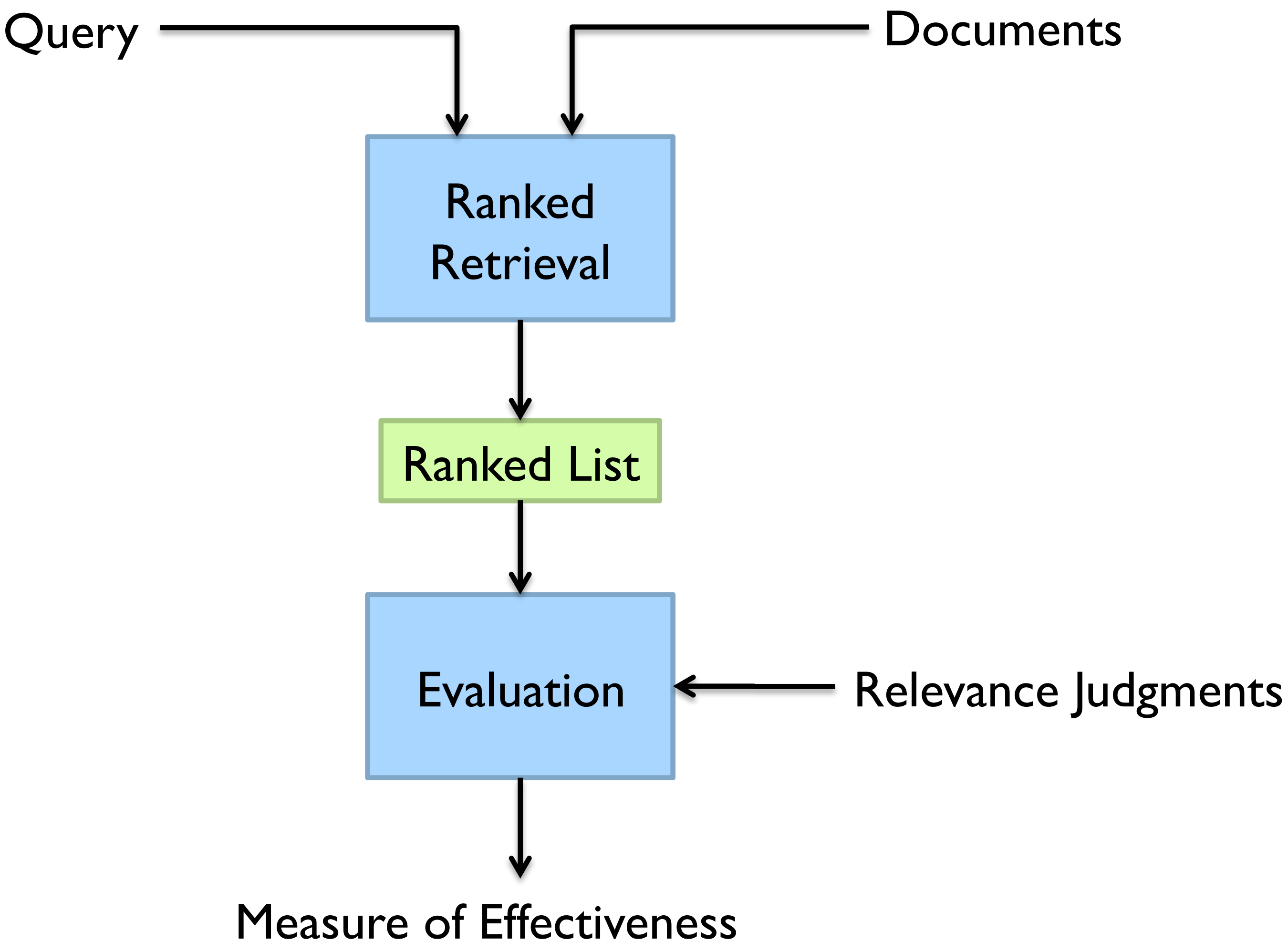
Retrieved

Not Relevant + Not Retrieved

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

Abstract Evaluation Model



User Studies

- Goal is to account for interface issues
  - By studying the interface component
  - By studying the complete system
- Formative evaluation
  - Provide a basis for system development
- Summative evaluation
  - Designed to assess effectiveness

Quantitative User Studies

- Select independent variable(s)
  - E.g., what info to display in selection interface
- Select dependent variable(s)
  - E.g., time to find a known relevant document
- Run subjects in different orders
  - Average out learning and fatigue effects
- Compute statistical significance
  - Null hypothesis: independent variable has no effect

Qualitative User Studies

- Direct observation
- Think-aloud protocols

Objective vs. Subjective Data

- Subjective self-assessment
  - Which did they think was more effective?
- Preference
  - Which interface did they prefer? Why?

Often at odds with objective measures!

Take-Away Messages

- Search engines provide access to unstructured textual information
- Searching is fundamentally about bridging the gap between words and meaning
- Information seeking is an iterative process in which the search engine plays an important role