College of Information Studies

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

Information Retrieval

Session 12

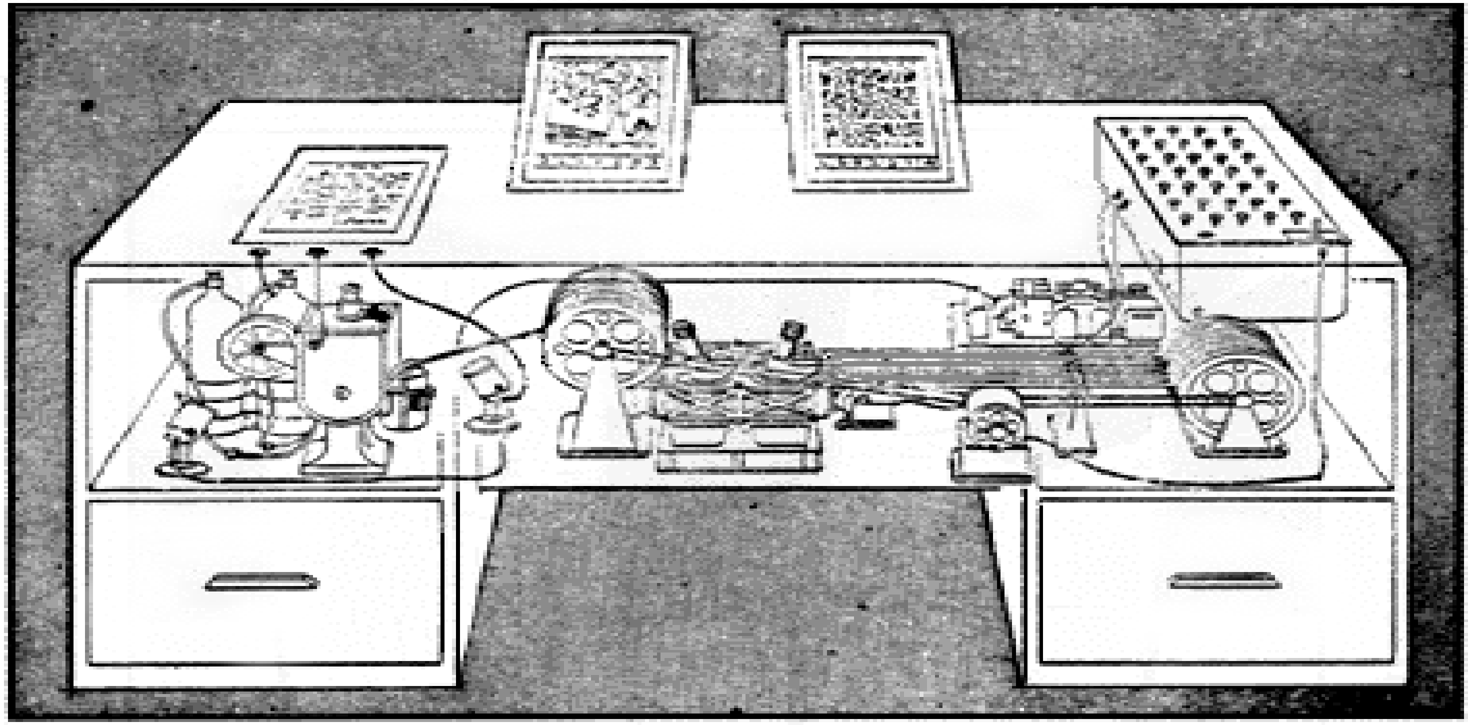
LBSC 671

Creating Information Infrastructures

Agenda

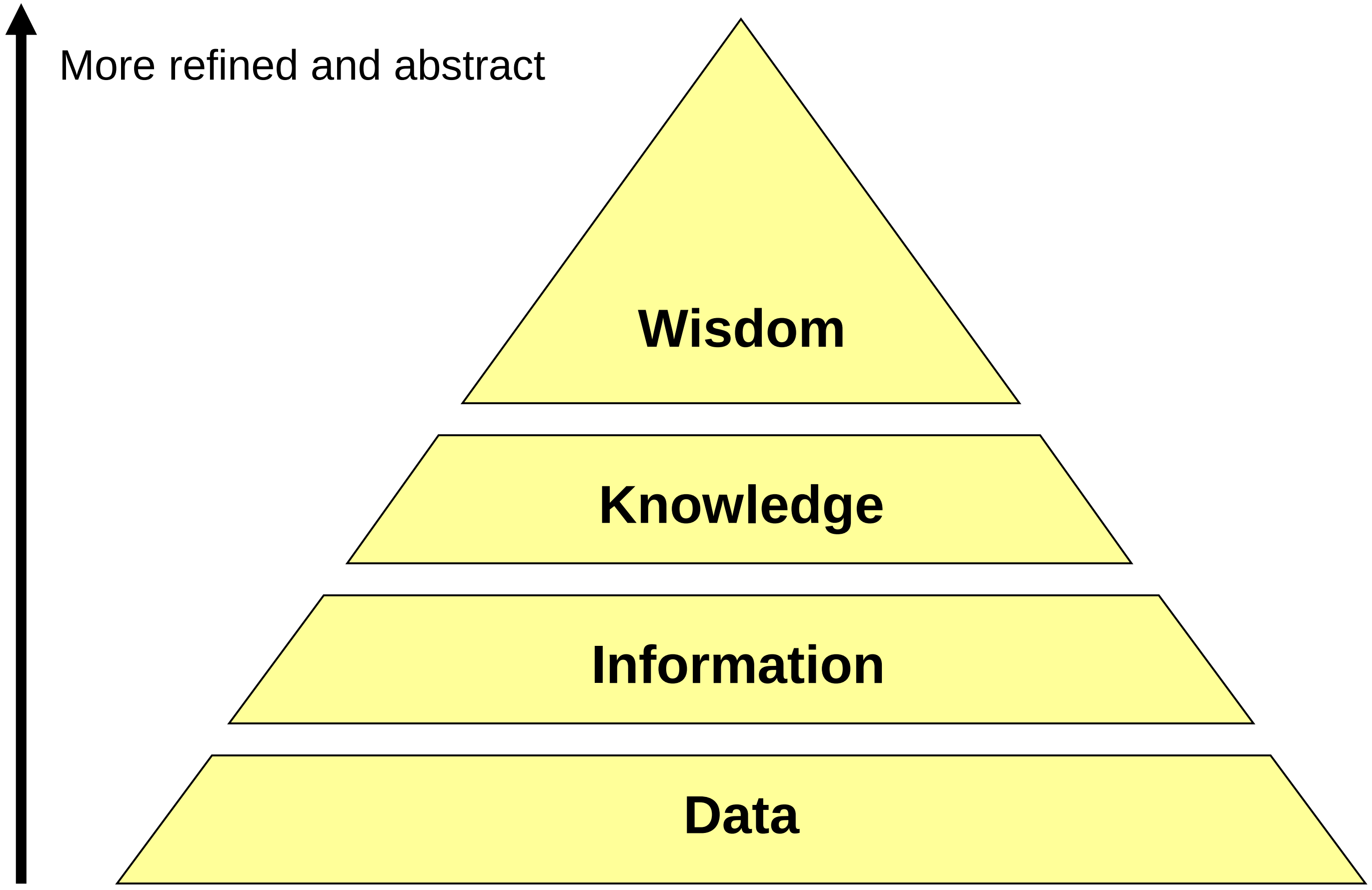
- The search process
- Information retrieval
- Recommender systems
- Evaluation

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).

Information Hierarchy



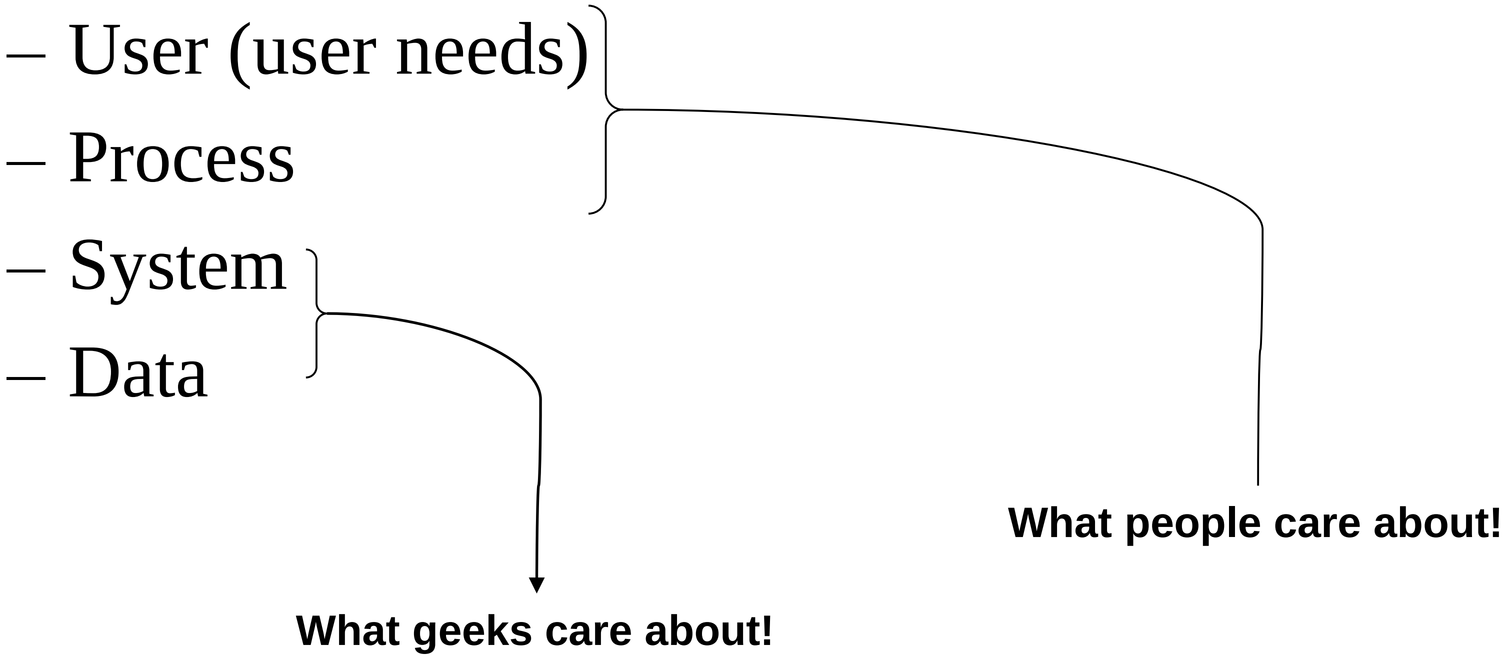
	Databases	**IR**
What we're retrieving	Structured data. Clear semantics based on a formal model.	Mostly unstructured. Free text with some metadata.
Queries we're posing	Formally (mathematically) defined queries. Unambiguous.	Vague, imprecise information needs (often expressed in natural language).
Results 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 critical.	Effectiveness and usability are critical.

Information "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?

The Big Picture

- The four components of the information retrieval environment:

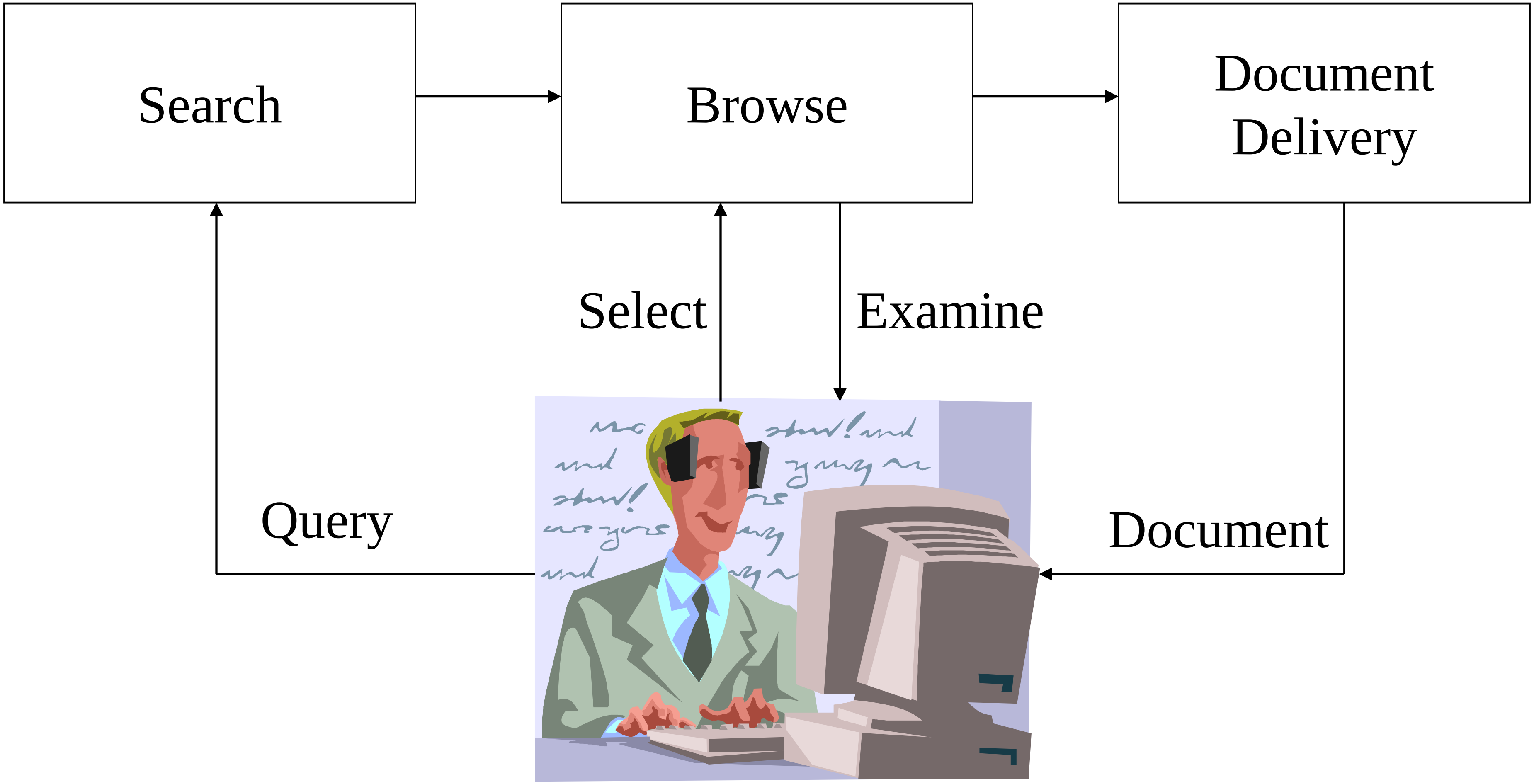


  - User (user needs)
  - Process
  - System
  - Data

What people care about!

What geeks care about!

Information Retrieval Paradigm



Search

Browse

Document Delivery

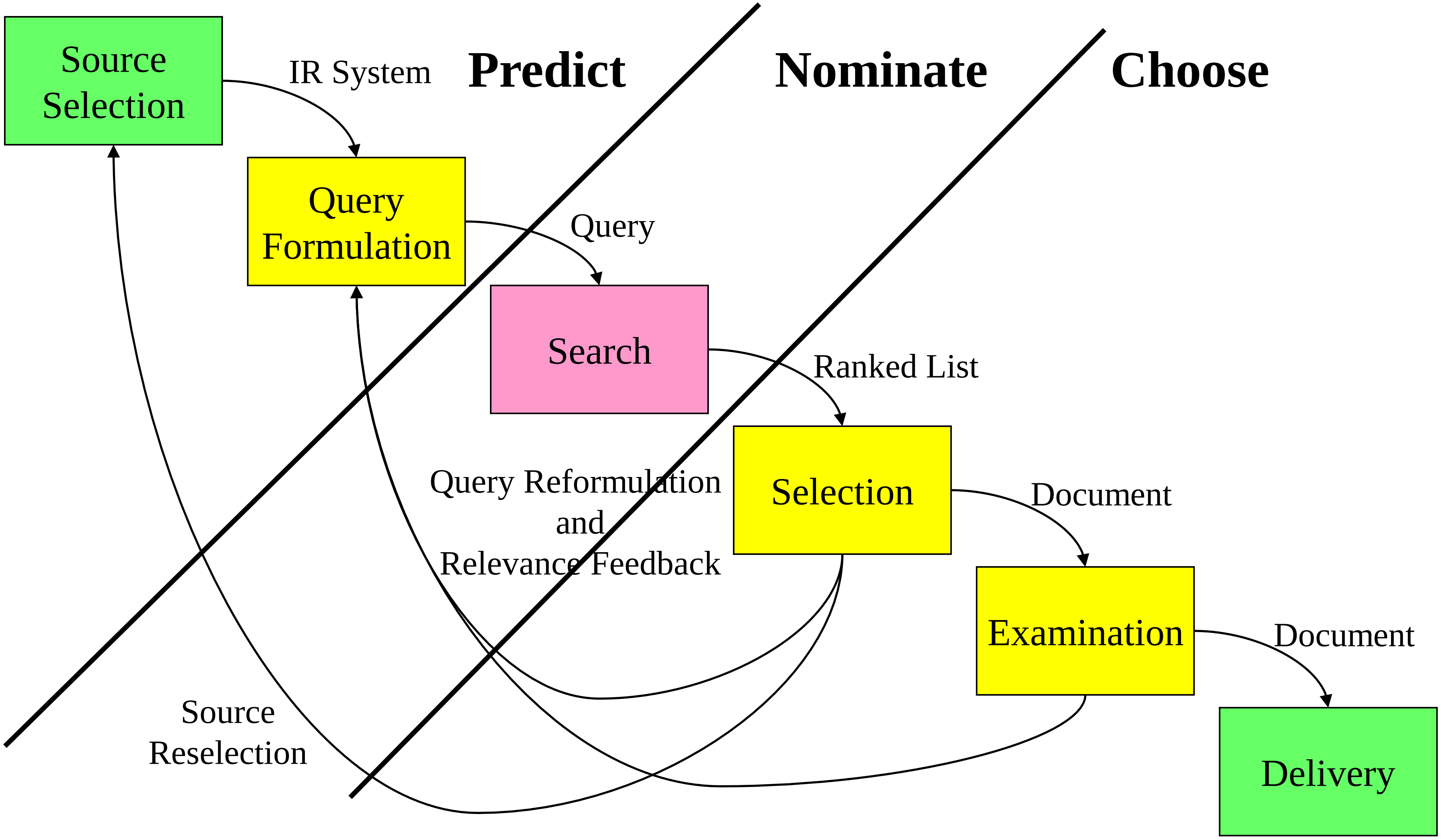
Select

Examine

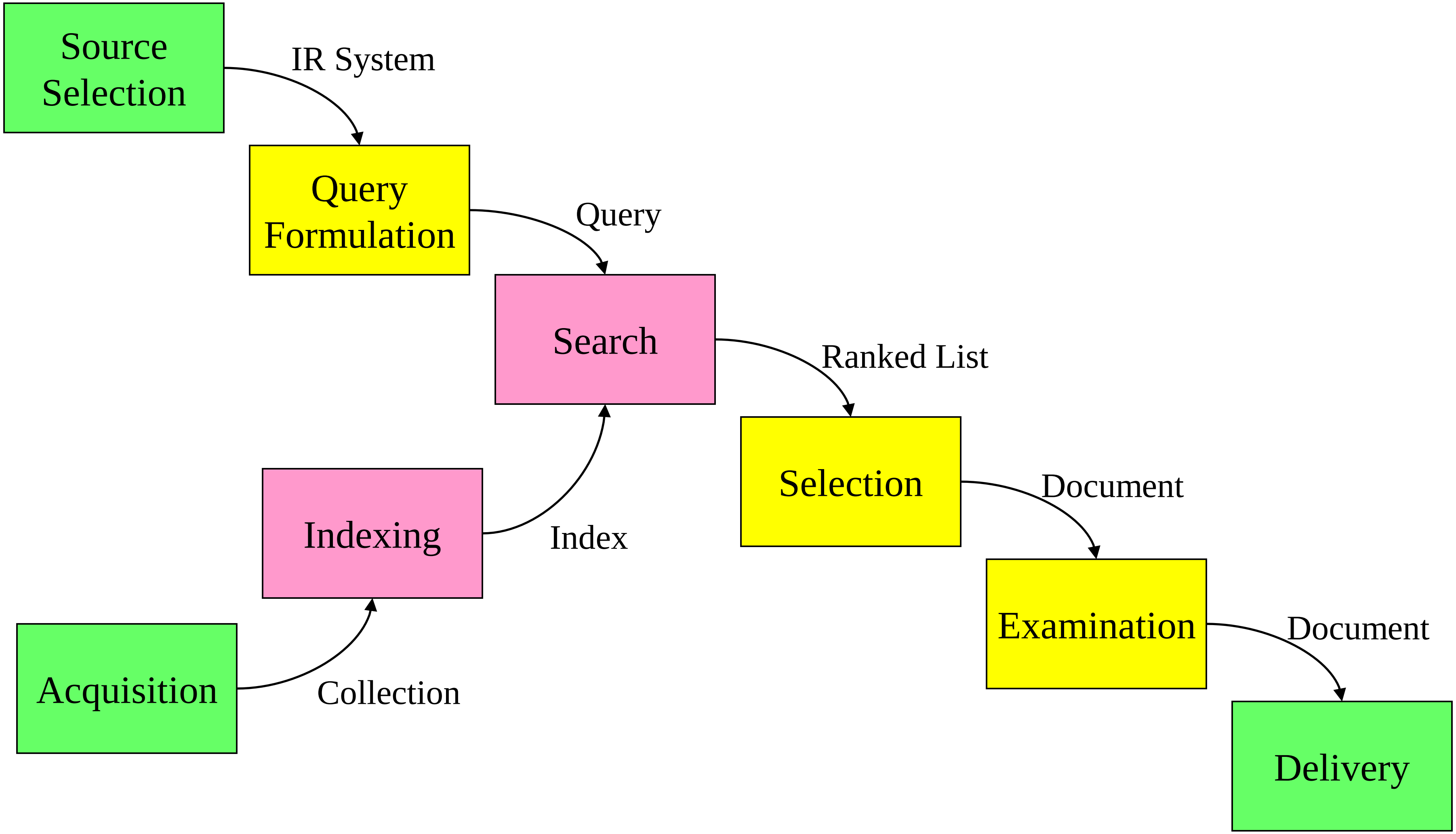
Query

Document

Supporting the Search Process



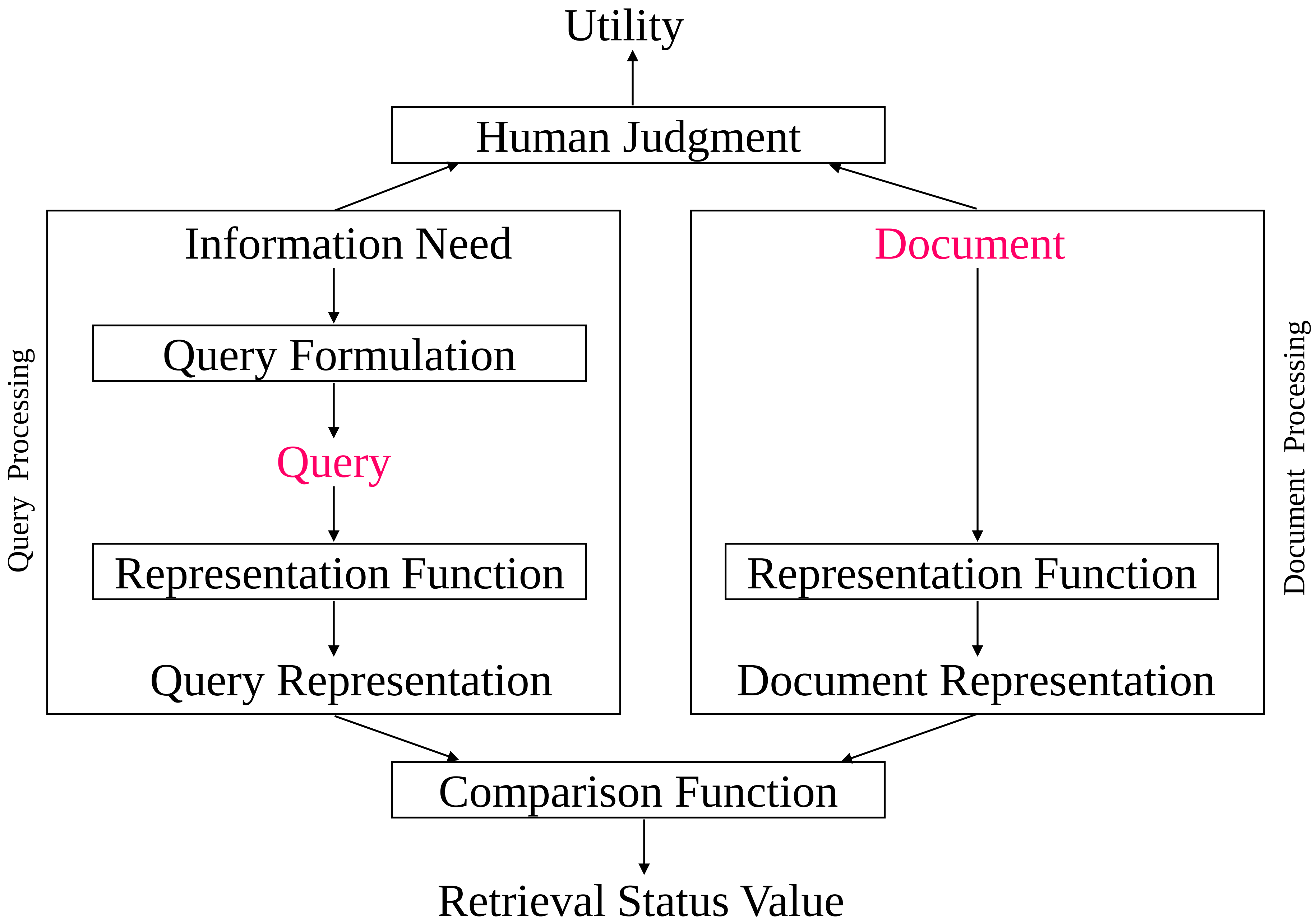
Supporting the Search Process



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

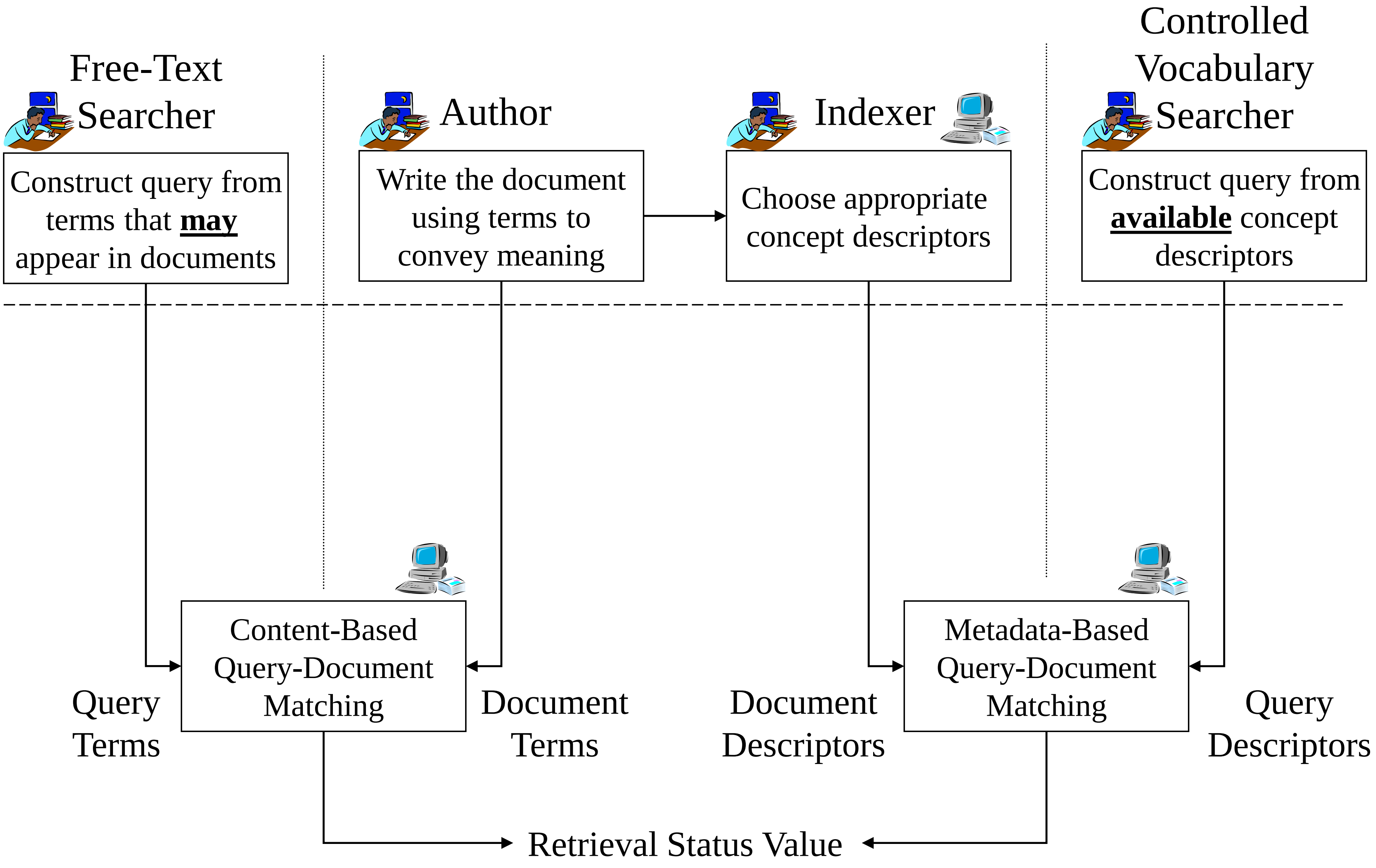
Search Component Model



Ways of Finding Text

- Searching metadata
  - Using controlled or uncontrolled vocabularies
- Searching content
  - Characterize documents by the words the contain
- Searching behavior
  - User-Item: Find similar users
  - Item-Item: Find items that cause similar reactions

Two Ways of Searching



Free-Text Searcher

Construct query from terms that **may** appear in documents

Author

Write the document using terms to convey meaning

Indexer

Choose appropriate concept descriptors

Controlled Vocabulary Searcher

Construct query from **available** concept descriptors

Query Terms

Content-Based Query-Document Matching

Document Terms

Document Descriptors

Metadata-Based Query-Document Matching

Query Descriptors

Retrieval Status Value

“Exact Match” Retrieval

- Find all documents with some characteristic
  - Indexed as “Presidents -- United States”
  - Containing the words “Clinton” and “Peso”
  - Read by my boss
- A set of documents is returned
  - Hopefully, not too many or too few
  - Usually listed in date or alphabetical order

The Perfect Query Paradox

- Every information need has a perfect document ste
  - 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!

Queries on the Web (1999)

- Low query construction effort
  - 2.35 (often imprecise) terms per query
  - 20% use operators
  - 22% are subsequently modified

- Low browsing effort
  - Only 15% view more than one page
  - Most look only “above the fold”
    - One study showed that 10% don’t know how to scroll!

Types of User Needs

- Informational (30-40% of queries)
  - What is a quark?
- Navigational
  - Find the home page of United Airlines
- Transactional
  - Data: What is the weather in Paris?
  - Shopping: Who sells a Viao Z505RX?
  - Proprietary: Obtain a journal article

Ranked Retrieval

- Put most useful documents near top of a list
  - Possibly useful documents go lower in the list
- Users can read down as far as they like
  - Based on what they read, time available, ...
- Provides useful results from weak queries
  - Untrained users find exact match harder to use

Similarity-Based Retrieval

- Assume “most useful” = most similar to query
- Weight terms based on two criteria:
  - Repeated words are good cues to meaning
  - Rarely used words make searches more selective
- Compare weights with query
  - Add up the weights for each query term
  - Put the documents with the highest total first

Simple Example: Counting Words

Query: recall and fallout measures for information retrieval

Documents:

1: Nuclear fallout contaminated Texas.

2: Information retrieval is interesting.

3: Information retrieval is complicated.

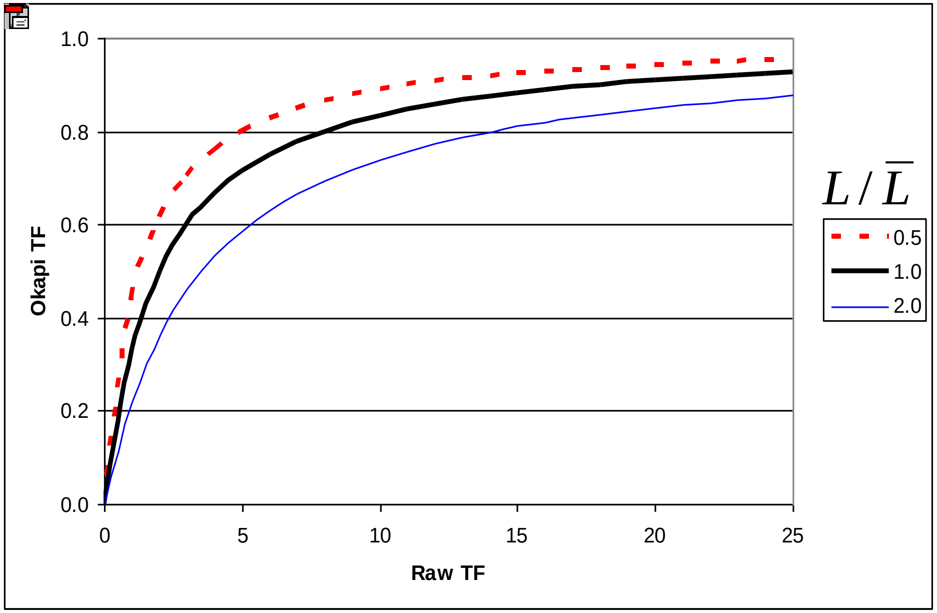
	1	2	3	Query
complicated			1	
contaminated	1			
fallout	1			1
information		1	1	1
interesting		1		
nuclear	1			
retrieval		1	1	1
Texas	1			

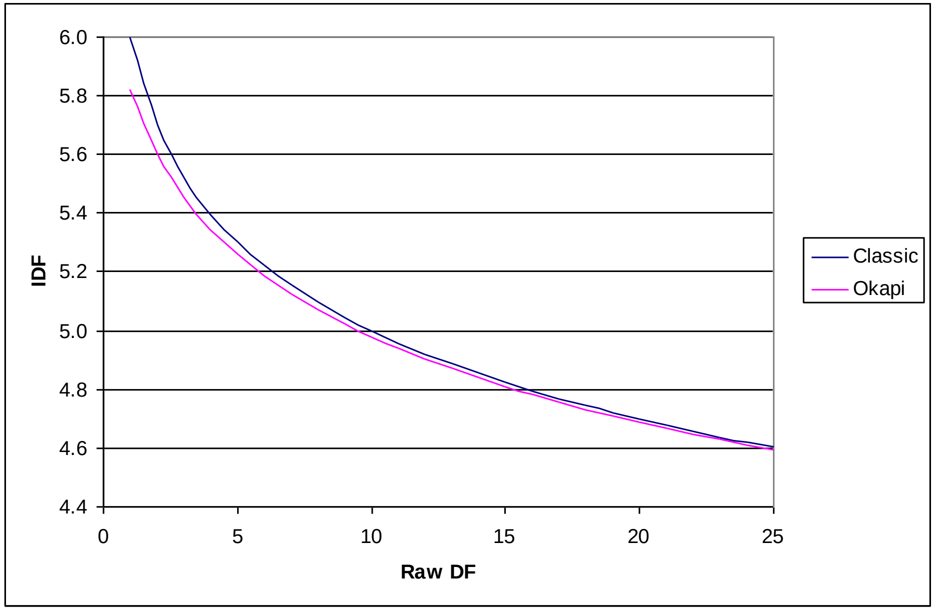
Discussion Point: Which Terms to Emphasize?

- Major factors
  - Uncommon terms are more selective
  - Repeated terms provide evidence of meaning
- Adjustments
  - Give more weight to terms in certain positions
    - Title, first paragraph, etc.
  - Give less weight each term in longer documents
  - Ignore documents that try to “spam” the index
    - Invisible text, excessive use of the “meta” field, …

"Okapi" Term Weights

$$w_{i,j} = \underbrace{\frac{TF_{i,j}}{1.5\frac{L_i}{\overline{L}} + TF_{i,j} + 0.5}}_{\text{TF component}} * \underbrace{\log\left(\frac{N - DF_j + 0.5}{DF_j + 0.5}\right)}_{\text{IDF component}}$$





Index Quality

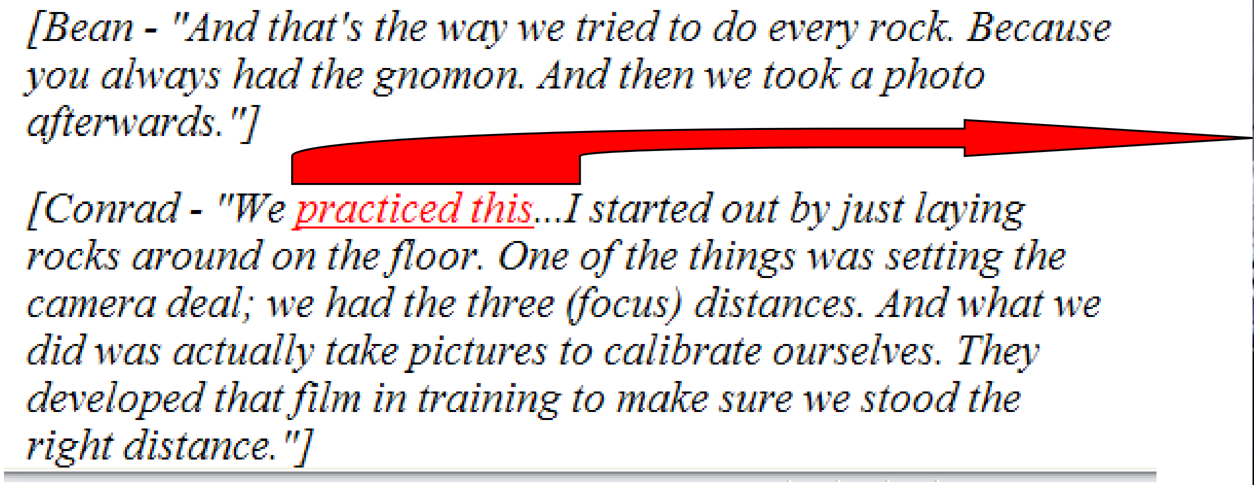
- Crawl quality
  - Comprehensiveness, dead links, duplicate detection
- Document analysis
  - Frames, metadata, imperfect HTML, …
- Document extension
  - Anchor text, source authority, category, language, …
- Document restriction (ephemeral text suppression)
  - Banner ads, keyword spam, …

Other Web Search Quality Factors

- Spam suppression
  - “Adversarial information retrieval”
  - Every source of evidence has been spammed
    - Text, queries, links, access patterns, …
- “Family filter” accuracy
  - Link analysis can be helpful

Indexing Anchor Text

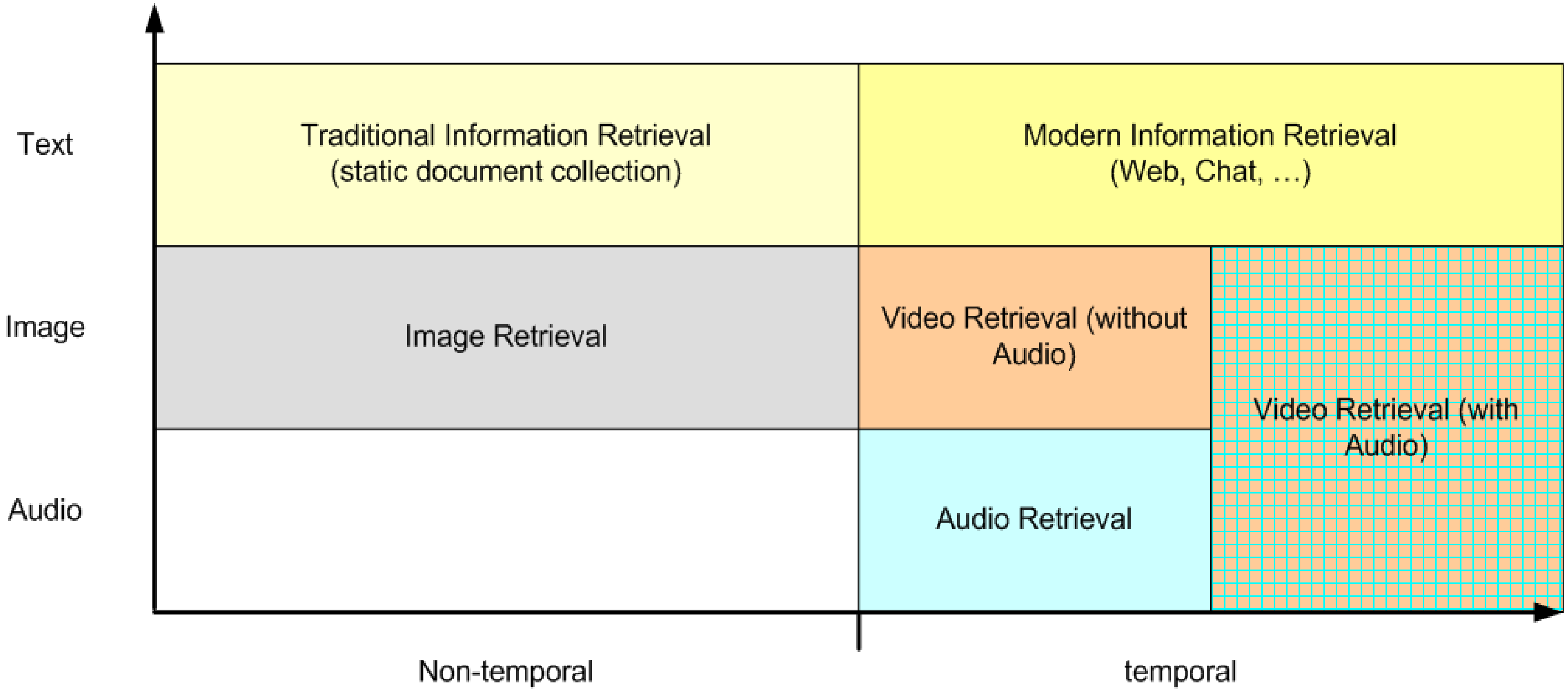
- A type of “document expansion”
  - – Terms near links describe content of the target
- Works even when you can’t index content
  - – Image retrieval, uncrawled links, …



[Bean - "And that's the way we tried to do every rock. Because you always had the gnomon. And then we took a photo afterwards."]

[Conrad - "We practiced this...I started out by just laying rocks around on the floor. One of the things was setting the camera deal; we had the three (focus) distances. And what we did was actually take pictures to calibrate ourselves. They developed that film in training to make sure we stood the right distance."]

Information Retrieval Types



Text

Image

Audio

Traditional Information Retrieval
(static document collection)

Modern Information Retrieval
(Web, Chat, …)

Image Retrieval

Video Retrieval (without Audio)

Video Retrieval (with Audio)

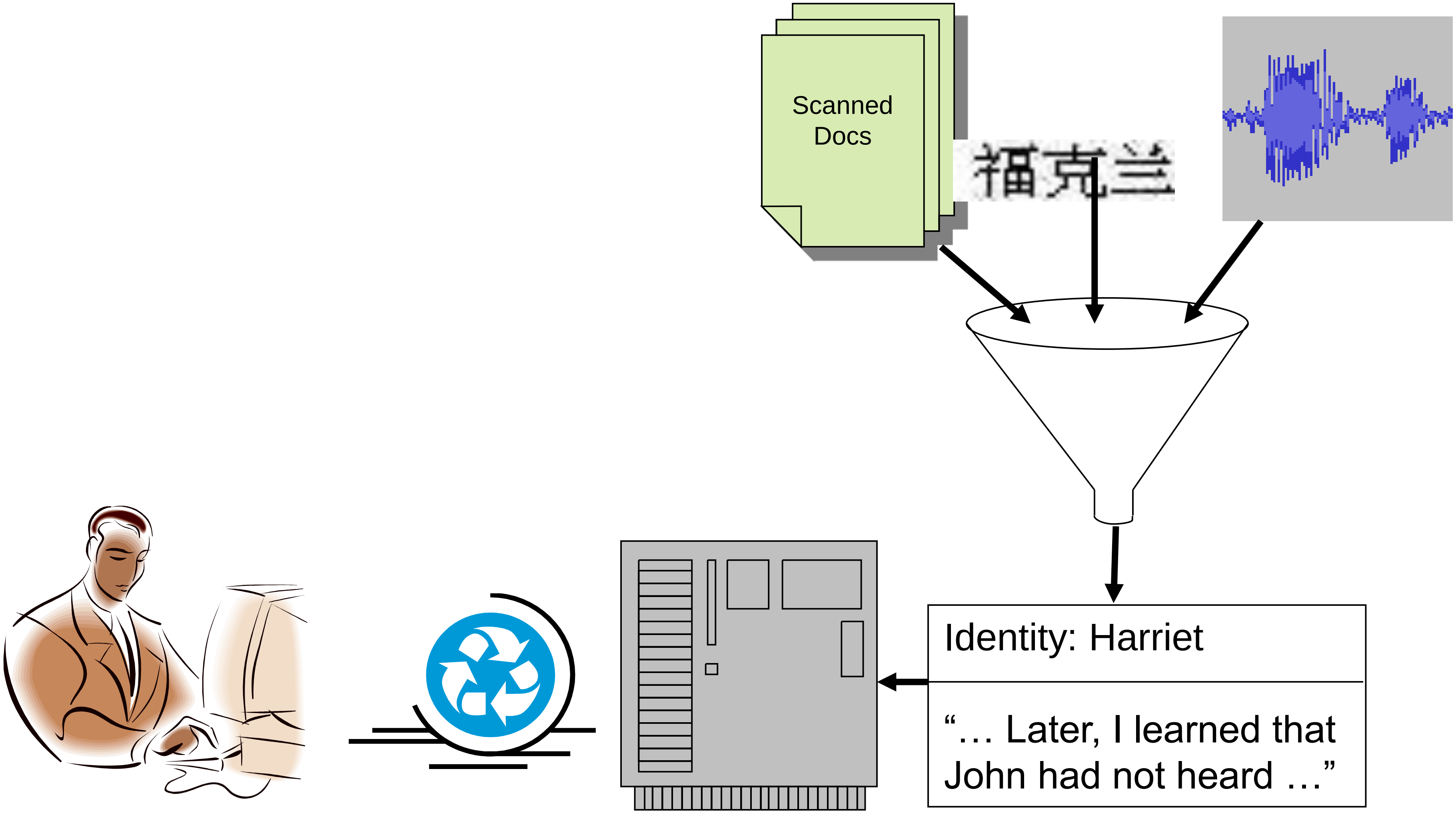
Audio Retrieval

Non-temporal

temporal

Source: Ayse Goker

Expanding the Search Space



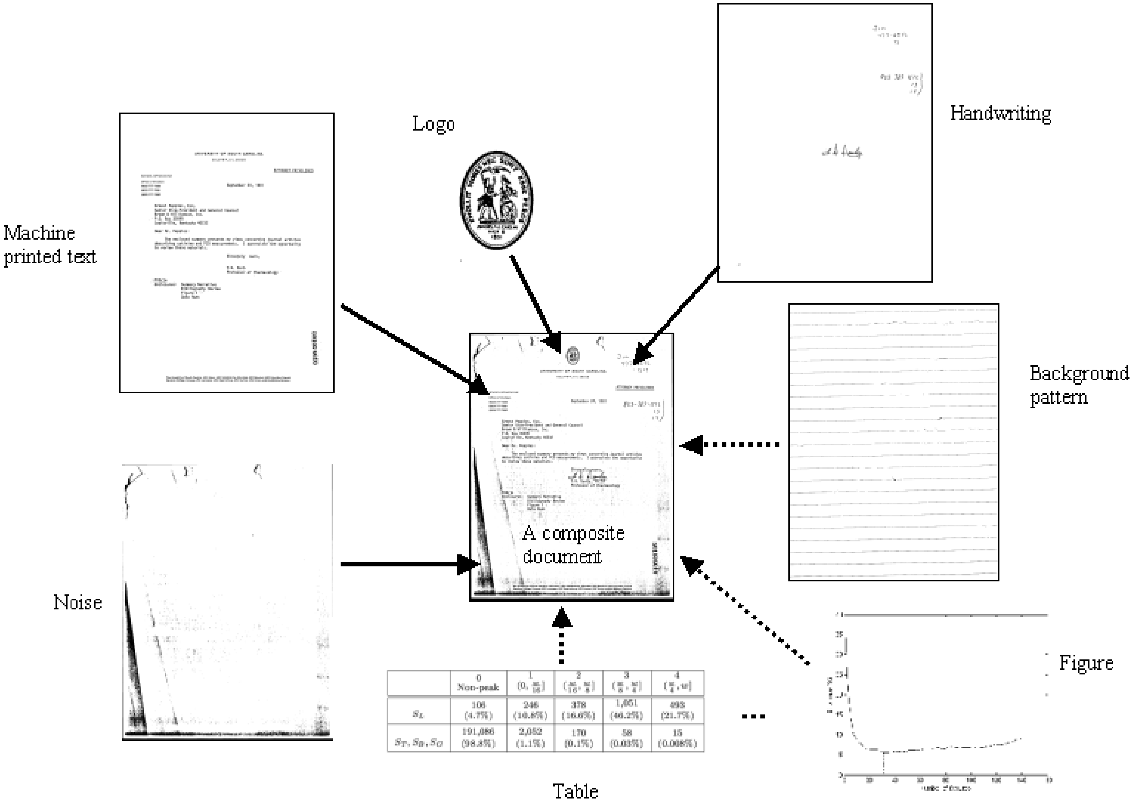
Scanned Docs

Identity: Harriet

“… Later, I learned that John had not heard …”

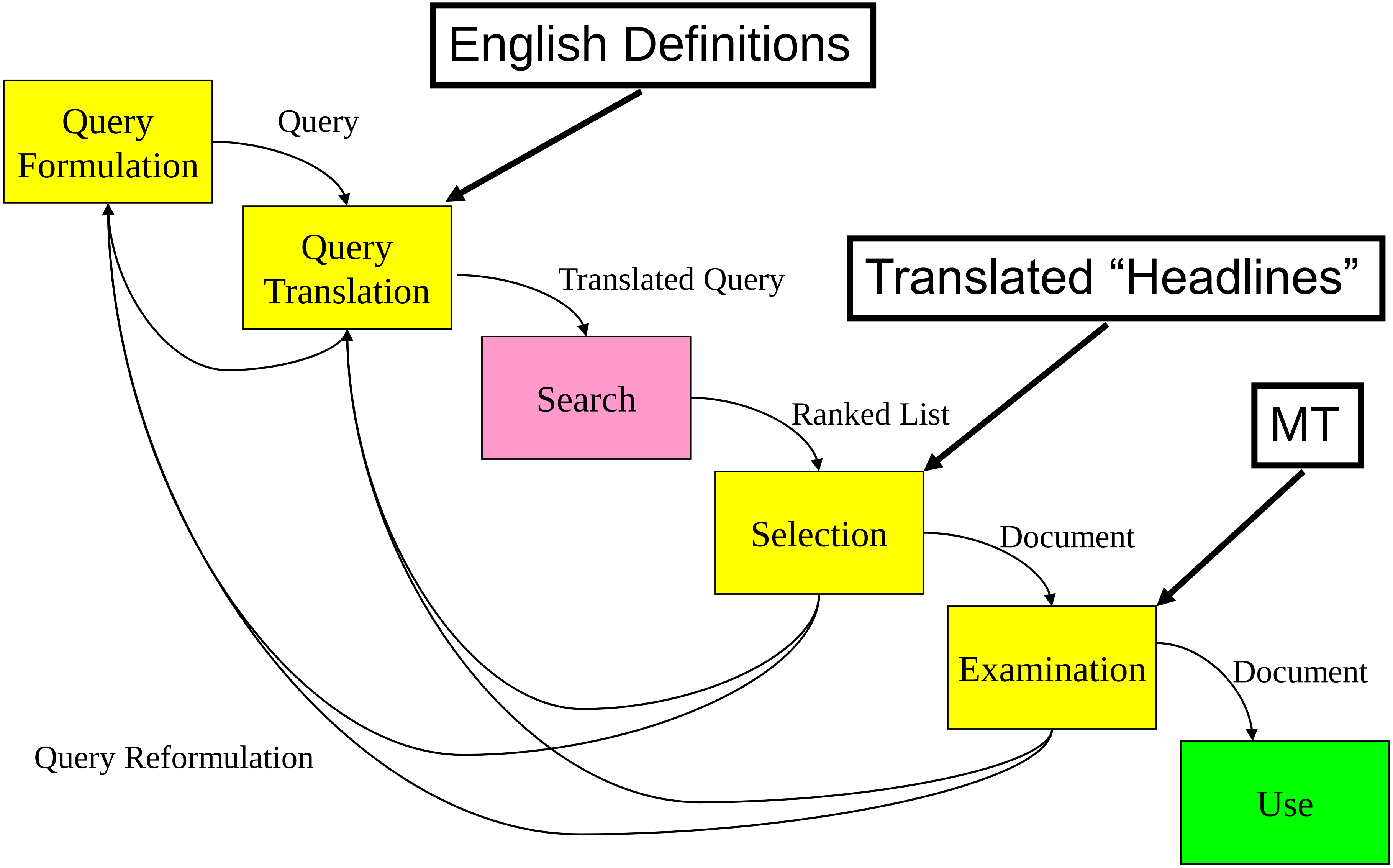
Page Layer Segmentation

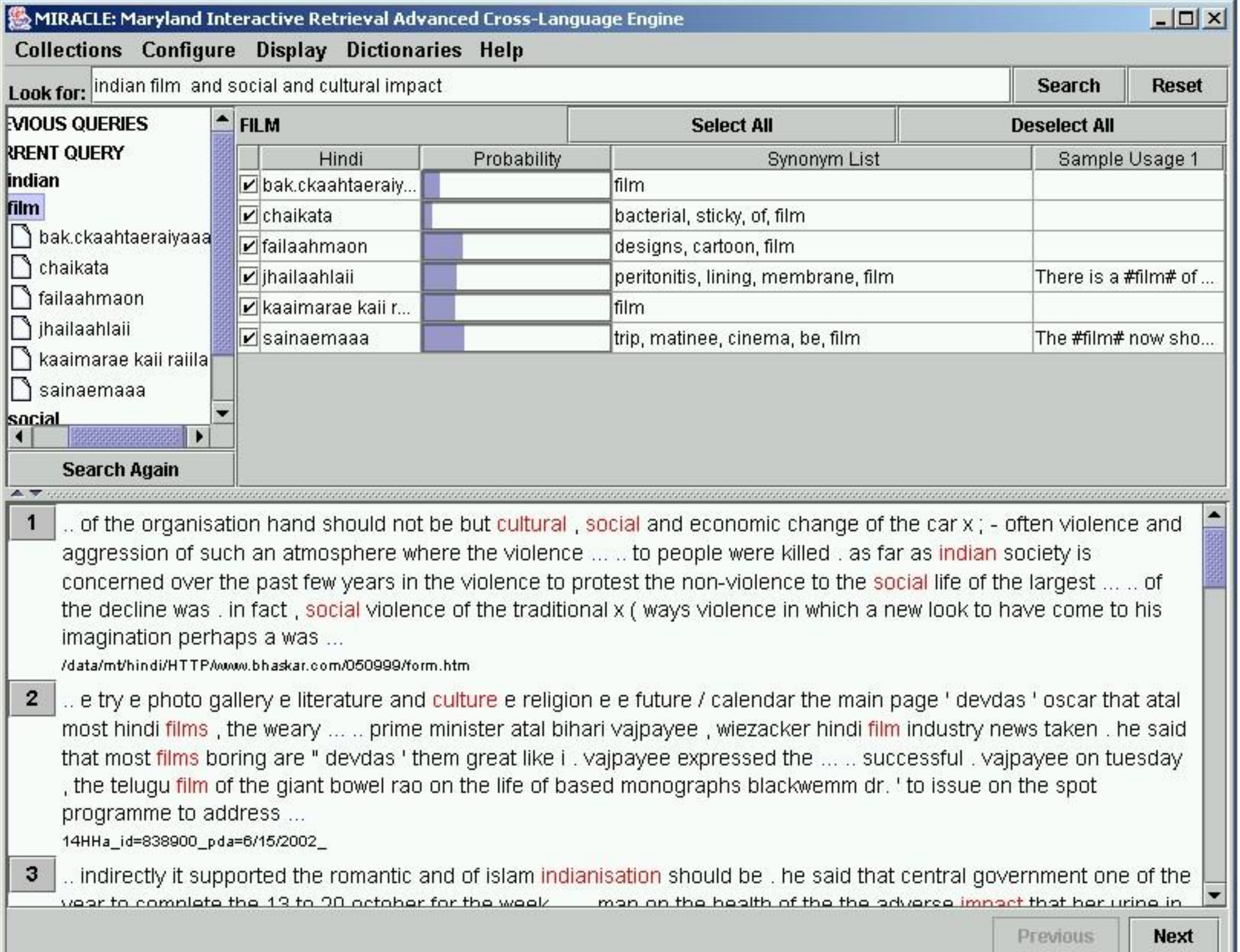
- Document image generation model
  - A document consists many layers, such as handwriting, machine printed text, background patterns, tables, figures, noise, etc.



	0 Non-peak	1 $(0, \frac{w}{16}]$	2 $(\frac{w}{16}, \frac{w}{8}]$	3 $(\frac{w}{8}, \frac{w}{4}]$	4 $(\frac{w}{4}, w]$
S_L	106 (4.7%)	246 (10.8%)	378 (16.6%)	1,051 (46.2%)	493 (21.7%)
S_T, S_B, S_G	191,086 (98.8%)	2,052 (1.1%)	170 (0.1%)	58 (0.03%)	15 (0.008%)

Searching Other Languages





Look for: indian film and social and cultural impact

EVIOUS QUERIES
RRENT QUERY
indian
film
- bak.ckaahtaeraiyaaa
- chaikata
- failaahmaon
- jhailaahlaii
- kaaimarae kaii raiila
- sainaemaaa

social

FILM

	Hindi	Probability	Synonym List	Sample Usage 1
☑	bak.ckaahtaeraiy...		film	
☑	chaikata		bacterial, sticky, of, film	
☑	failaahmaon		designs, cartoon, film	
☑	jhailaahlaii		peritonitis, lining, membrane, film	There is a #film# of ...
☑	kaaimarae kaii r...		film	
☑	sainaemaaa		trip, matinee, cinema, be, film	The #film# now sho...

1 .. of the organisation hand should not be but cultural , social and economic change of the car x ; - often violence and aggression of such an atmosphere where the violence ... .. to people were killed . as far as indian society is concerned over the past few years in the violence to protest the non-violence to the social life of the largest ... .. of the decline was . in fact , social violence of the traditional x (ways violence in which a new look to have come to his imagination perhaps a was ...

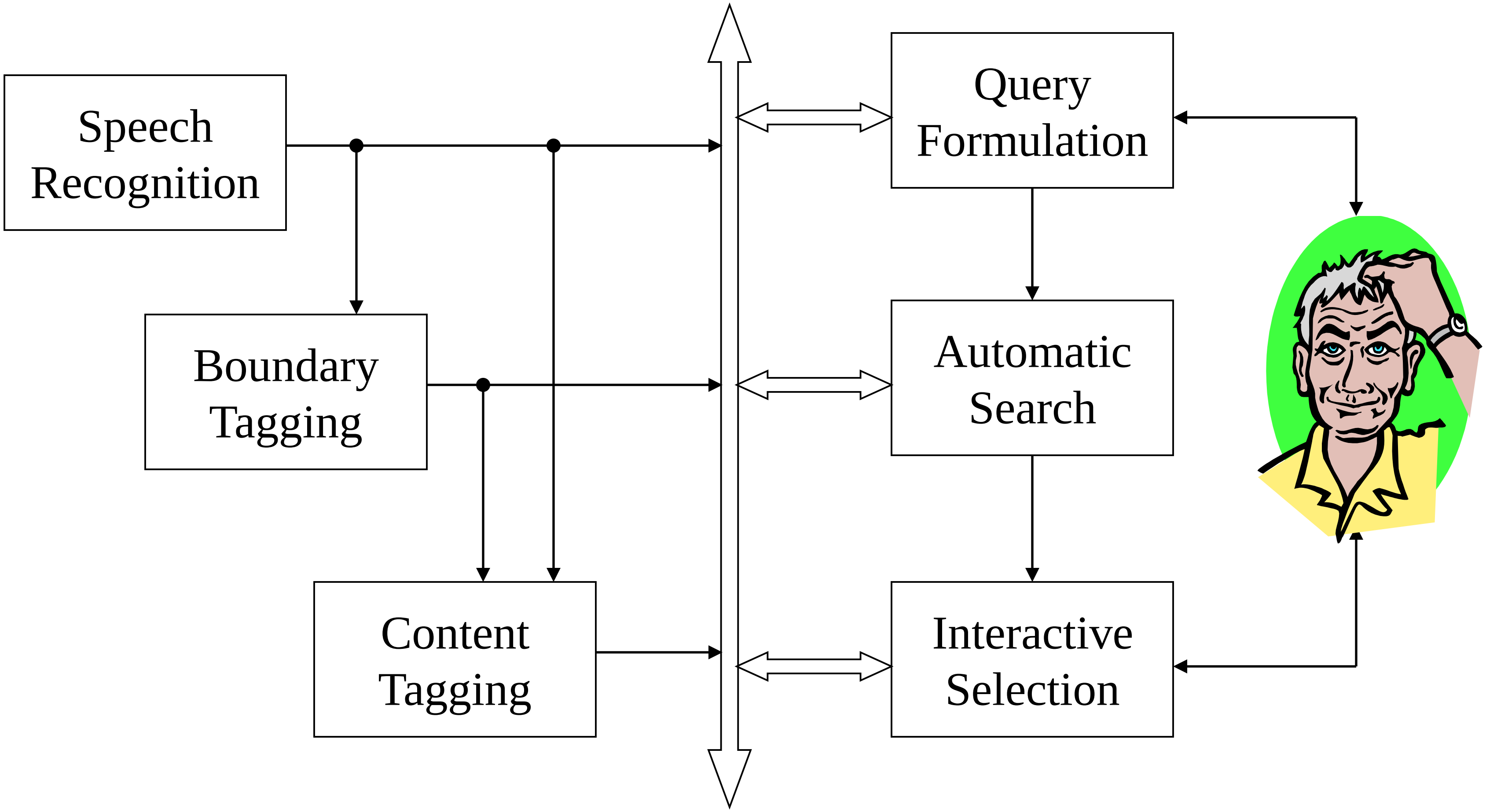
/data/mt/hindi/HTTP/www.bhaskar.com/050999/form.htm

2 .. e try e photo gallery e literature and culture e religion e e future / calendar the main page ' devdas ' oscar that atal most hindi films , the weary ... .. prime minister atal bihari vajpayee , wiezacker hindi film industry news taken . he said that most films boring are " devdas ' them great like i . vajpayee expressed the ... .. successful . vajpayee on tuesday , the telugu film of the giant bowel rao on the life of based monographs blackwemm dr. ' to issue on the spot programme to address ...

14HHa_id=838900_pda=6/15/2002_

3 .. indirectly it supported the romantic and of islam indianisation should be . he said that central government one of the year to complete the 13 to 20 october for the week ... man on the health of the the adverse impact that her urine in

Speech Retrieval Architecture



Speech Recognition

Boundary Tagging

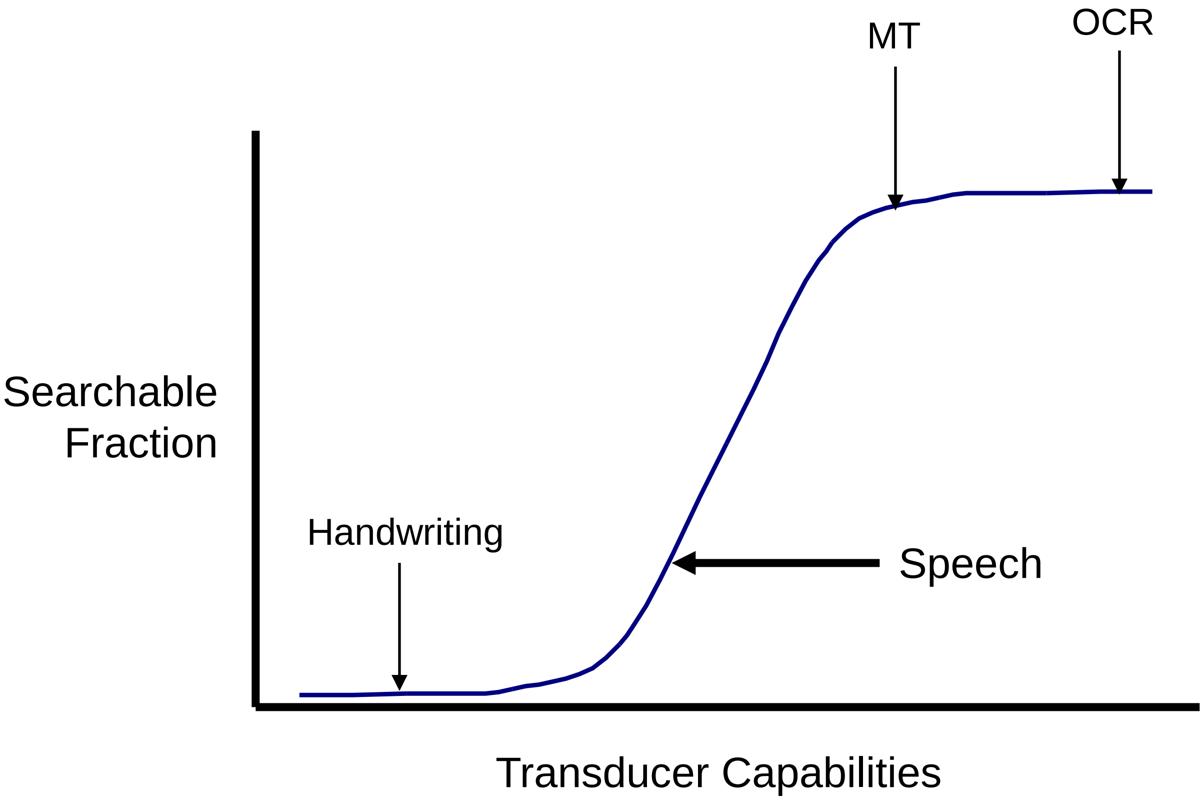
Content Tagging

Query Formulation

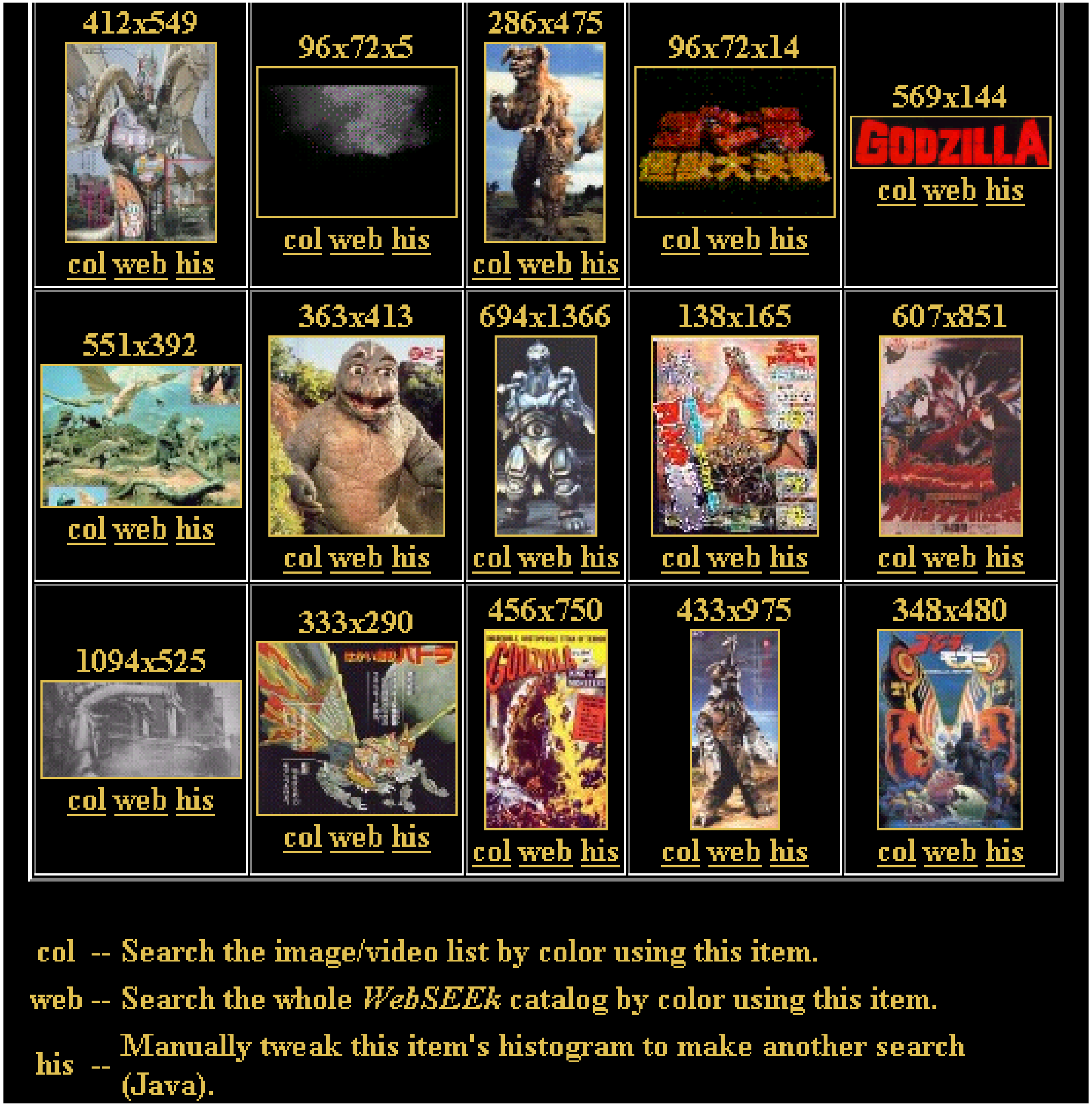
Automatic Search

Interactive Selection

High Payoff Investments



$$\frac{\textit{accurately recognized words}}{\textit{words produced}}$$



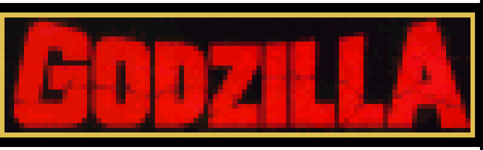
col -- Search the image/video list by color using this item.

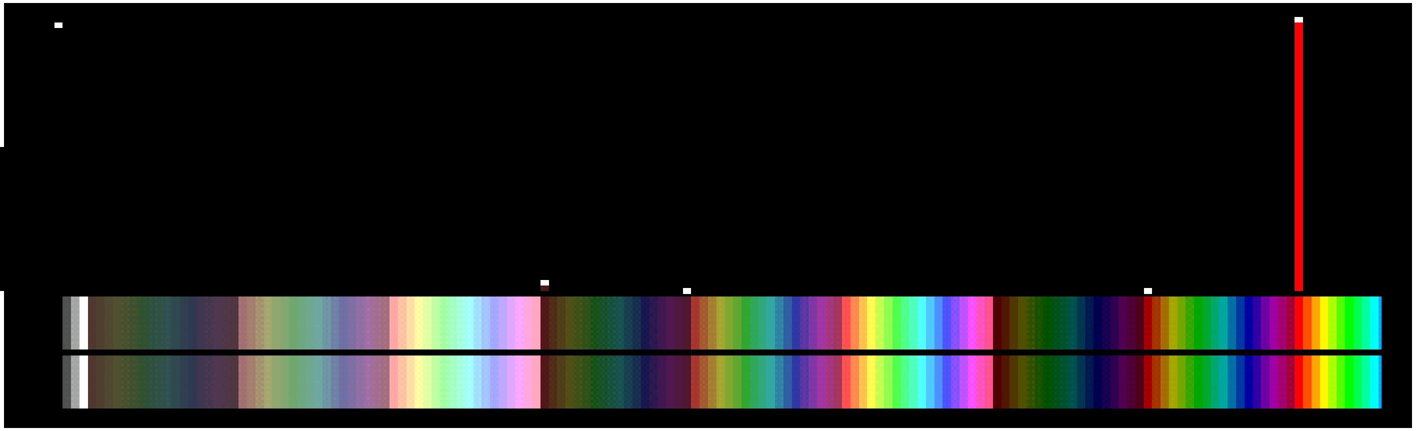
web -- Search the whole *WebSEEk* catalog by color using this item.

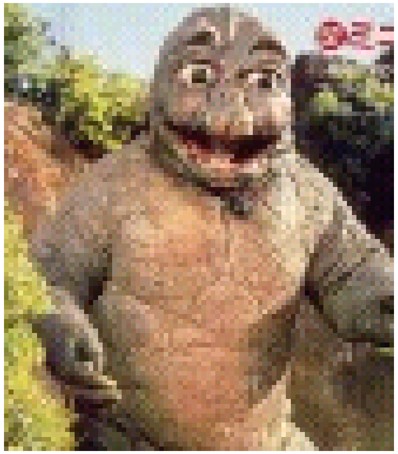
his -- Manually tweak this item's histogram to make another search (Java).

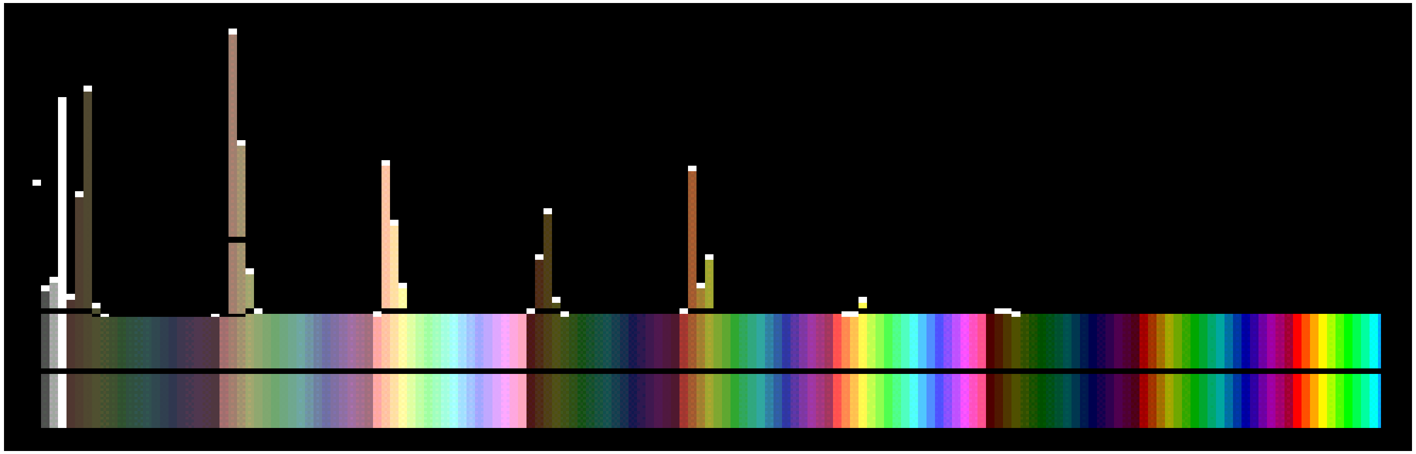
http://www.ctr.columbia.edu/webseek/

Color Histogram Example









Rating-Based Recommendation

- Use ratings as to describe objects
  - Personal recommendations, peer review, …
- Beyond topicality:
  - Accuracy, coherence, depth, novelty, style, …
- Has been applied to many modalities
  - Books, Usenet news, movies, music, jokes, beer, …

Using Positive Information

	Small World	Space Mtn	Mad Tea Pty	Dumbo	Speedway	Cntry Bear
Joe	D	A	B	D	?	?
Ellen	A	F	D		F	
Mickey	A	A	A	A	A	A
Goofy	D	A		C		
John	A	C	A	C		A
Ben	F	A				F
Nathan	D		A		A	

Using Negative Information

	Small World	Space Mtn	Mad Tea Pty	Dumbo	Speedway	Cntry Bear
Joe	D	A	B	D	?	?
Ellen	A	F	D		F	
Mickey	A	A	A	A	A	A
Goofy	D	A		C		
John	A	C	A	C		A
Ben	F	A				F
Nathan	D		A		A	

Problems with Explicit Ratings

- Cognitive load on users -- people don’t like to provide ratings
- Rating sparsity -- needs a number of raters to make recommendations
- No ways to detect new items that have not rated by any users

Putting It All Together

	Free Text	**Behavior**	**Metadata**
Topicality	🟩	🟨	🟩
Quality	🟥	🟩	🟩
Reliability	🟩	🟩	🟨
Cost	🟩	🟨	🟥
Flexibility	🟩	🟥	🟥

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?

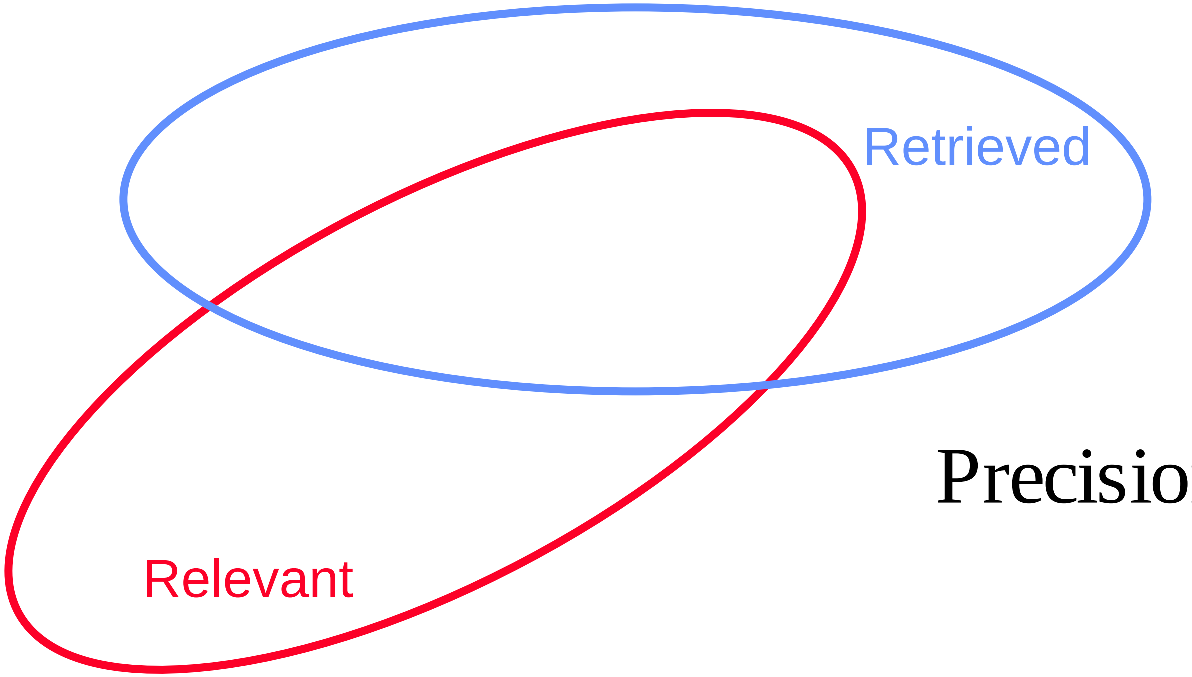
A.	🟦	🟦	🟦	🟦	🟦	🟦	🟦	⬜	⬜	⬜	⬜	⬜	⬜
B.	🟦	⬜	🟦	⬜	🟦	⬜	🟦	⬜	⬜	⬜	⬜	⬜	⬜
C.	⬜	🟦	⬜	🟦	⬜	🟦	⬜	🟦	⬜	🟦	⬜	⬜	⬜
D.	🟦	🟦	🟦	⬜	🟦	🟦	⬜	⬜	🟦	⬜	⬜	⬜	⬜
E.	🟦	⬜	⬜	⬜	🟦	🟦	🟦	🟦	🟦	⬜	⬜	⬜	⬜
F.	⬜	🟦	🟦	⬜	⬜	⬜	⬜	🟦	🟦	🟦	🟦	🟦	⬜

🟦 = 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

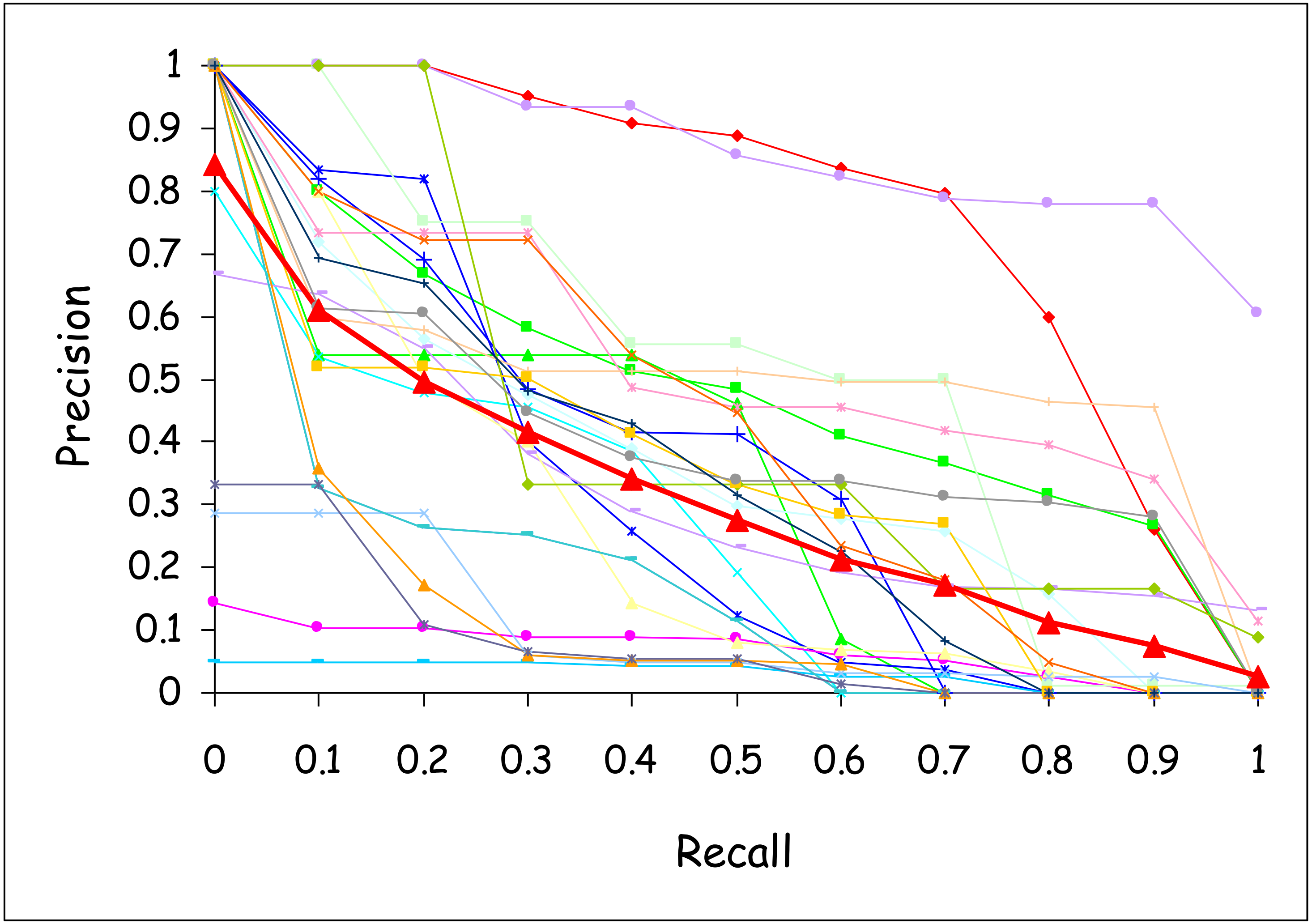
Measures of Effectiveness



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

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

Precision-Recall Curves



Source: Ellen Voorhees, NIST

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

Summary

- Search is a process engaged in by people
- Human-machine synergy is the key
- Content and behavior offer useful evidence
- Evaluation must consider many factors

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?