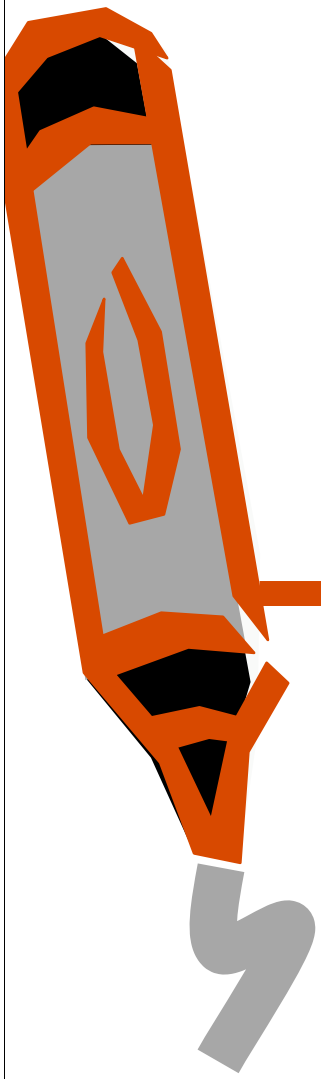
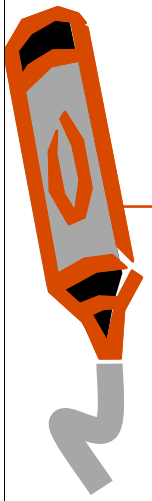


Information Retrieval

Overview

many slides courtesy James Allan @umass Amherst
some slides courtesy ChengXiang Zhai @Urbana-Champaign

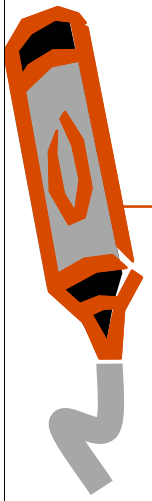


administrivia

- INSTRUCTOR
  Virgil Pavlu vip@ccs.neu.edu

- GRADER : if we get any, he/she will have final authority on homework grading issues

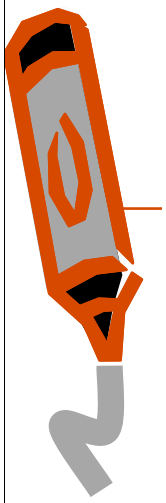
- http://ccs.neu.edu/~jaa/ISU535.06X2



overview

- what is information retrieval ?
- how does it work ?
- a simple retrieval model

what is IR?



Web Images Groups News Froogle Local Scholar more »

Google cikm 2003 Search Advanced Search Preferences

Web

12. **CIKM 2003**: New Orleans, Louisiana, USA
12. **CIKM 2003**: New Orleans, Louisiana, USA. Proceedings of the **2003** ACM **CIKM** International Conference on Information and Knowledge Management, New Orleans, ...
www.informatik.uni-trier.de/ ~ley/db/conf/**cikm**/**cikm2003**.html - 56k - Cached - Similar pages

CIKM
Proceedings of the **2003** ACM **CIKM** International Conference on Information and Knowledge Management, New Orleans, Louisiana, USA, November 2-8, **2003**. ...
www.informatik.uni-trier.de/~ley/db/conf/**cikm**/ - 10k - Jun 25, 2005 - Cached - Similar pages

CIKM'2003 review
CIKM'2003 highlights. 12th ACM International Conference on Information and Knowledge Management, 3-8 November, New Orleans ...
smi.ucd.ie/~rinat/papers/**cikm**03_rep.html - 22k - Cached - Similar pages

Collaborative Filtering Mailing List Archive: [collab@sims] CFP
ACM **CIKM 2003** Call For Papers. 12th International Conference on Information and Knowledge ... caliber papers submitted to **CIKM 2003** will be accepted. ...
www.pdesigner.net/1996/0697.html - 17k - Cached - Similar pages

TOC
Proceedings of the twelfth international conference on Information and knowledge management citation. **2003**, New Orleans, LA, USA November 03 - 08, **2003** ...
portal.acm.org/toc.cfm?id=956863&type=proceeding - Similar pages

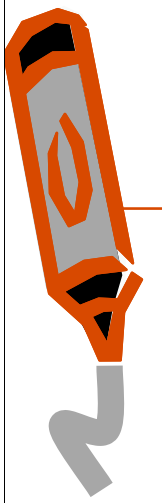
[Asis-l] **CIKM 2003**
[Asis-l] **CIKM 2003**. Padmini Srinivasan padmini@lakshmi.info-science.uiowa.edu Mon, 29 Sep **2003** 12:59:36 -0500. Previous message: [Asis-l] Re: ...
mail.asis.org/pipermail/ asis-l/**2003**-September/001024.html - 17k - Cached - Similar pages

[PDF] **CIKM 2003**
File Format: PDF/Adobe Acrobat - View as HTML
CIKM 2003. Jacob Kogan. Charles Nicholas. Marc Teboulle. –means and beyond – p.1/53. Page 2. Outline of the talk. how to build a partition ...
www.csee.umbc.edu/~nicholas/clustering/jacob.pdf - Similar pages

Tutorial on Document Clustering
CIKM 2003 Tutorial. Clustering Large and High-Dimensional Data ... Katya Pelekhov and Daniela Rus,"Using Star Clusters for Filtering", **CIKM** 2000, (pdf) ...
www.csee.umbc.edu/~nicholas/clustering/ - 9k - Cached - Similar pages

Conference on Information and Knowledge Management (**CIKM**)
CIKM has a strong tradition of workshops devoted to emerging areas of database ... The **CIKM** 2004 web page; The **CIKM 2003** Web Page; The **CIKM** 2002 Web Page ...
www.**cikm**.org/ - 7k - Cached - Similar pages

CIKM 2003, New Orleans, USA, November **2003**
Home. **CIKM 2003**, New Orleans, USA, November **2003**. << Bild 6 | Bild 7/80 | Bild 8 >>. Miniaturansicht.
www.torsten-priebe.de/showpics. php?folder=**2003**-11a_**cikm**03&picture=7 - 2k - Cached - Similar pages



evaluation

Google

Web Images Groups News Froogle Local Scholar more »

cikm 2003 Search Advanced Search Preferences

Web

12. **CIKM 2003**: New Orleans, Louisiana, USA — **R**
12. **CIKM 2003**: New Orleans, Louisiana, USA. Proceedings of the **2003** ACM **CIKM** International Conference on Information and Knowledge Management, New Orleans, ...
www.informatik.uni-trier.de/ ~ley/db/conf/**cikm**/**cikm2003**.html - 56k - Cached - Similar pages

CIKM — **R**
Proceedings of the **2003** ACM **CIKM** International Conference on Information and Knowledge Management, New Orleans, Louisiana, USA, November 2-8, **2003**. ...
www.informatik.uni-trier.de/~ley/db/conf/**cikm**/ - 10k - Jun 25, 2005 - Cached - Similar pages

CIKM'2003 review — **R**
CIKM'2003 highlights. 12th ACM International Conference on Information and Knowledge Management, 3-8 November, New Orleans ...
smi.ucd.ie/~rinat/papers/**cikm**03_rep.html - 22k - Cached - Similar pages

Collaborative Filtering Mailing List Archive: [collab@sims] CFP — **N**
ACM **CIKM 2003** Call For Papers. 12th International Conference on Information and Knowledge ... caliber papers submitted to **CIKM 2003** will be accepted. ...
www.pdesigner.net/1996/0697.html - 17k - Cached - Similar pages

TOC — **N**
Proceedings of the twelfth international conference on Information and knowledge management citation. **2003**, New Orleans, LA, USA November 03 - 08, **2003** ...
portal.acm.org/toc.cfm?id=956863&type=proceeding - Similar pages

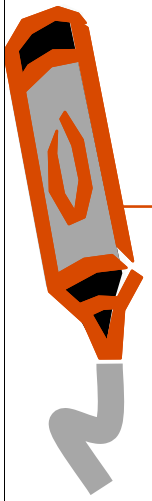
[Asis-l] **CIKM 2003** — **R**
[Asis-l] **CIKM 2003**. Padmini Srinivasan padmini@lakshmi.info-science.uiowa.edu Mon, 29 Sep **2003** 12:59:36 -0500. Previous message: [Asis-l] Re: ...
mail.asis.org/pipermail/ asis-l/**2003**-September/001024.html - 17k - Cached - Similar pages

[PDF] **CIKM 2003** — **N**
File Format: PDF/Adobe Acrobat - View as HTML
CIKM 2003. Jacob Kogan. Charles Nicholas. Marc Teboulle. –means and beyond – p.1/53. Page 2. Outline of the talk. how to build a partition ...
www.csee.umbc.edu/~nicholas/clustering/jacob.pdf - Similar pages

Tutorial on Document Clustering — **R**
CIKM 2003 Tutorial. Clustering Large and High-Dimensional Data ... Katya Pelekhov and Daniela Rus,"Using Star Clusters for Filtering", **CIKM** 2000, (pdf) ...
www.csee.umbc.edu/~nicholas/clustering/ - 9k - Cached - Similar pages

Conference on Information and Knowledge Management (**CIKM**) — **N**
CIKM has a strong tradition of workshops devoted to emerging areas of database ... The **CIKM** 2004 web page; The **CIKM 2003** Web Page; The **CIKM** 2002 Web Page ...
www.**cikm**.org/ - 7k - Cached - Similar pages

CIKM 2003, New Orleans, USA, November **2003** — **N**
Home. **CIKM 2003**, New Orleans, USA, November **2003**. << Bild 6 | Bild 7/80 | Bild 8 >>. Miniaturansicht.
www.torsten-priebe.de/showpics. php?folder=**2003**-11a_**cikm**03&picture=7 - 2k - Cached - Similar pages



altavista Web Images MP3/Audio Video News

cikm 2003 FIND Advanced Search Settings

SEARCH: Worldwide USA RESULTS IN: All languages English, Spanish

AltaVista found 19,700 results

Conference on Information and Knowledge Management (**CIKM**)
... The **CIKM** 2004 web page. The **CIKM 2003** Web Page. The **CIKM** 2002 Web Page. The **CIKM** 2001 Web page
www.cikm.org
More pages from cikm.org

Department of Computer Science at LSU
News & Events. Department of Computer Science. 298 Coates Hall. Phone: (225)578-1495
bit.csc.lsu.edu
More pages from bit.csc.lsu.edu

ACM **CIKM 2003** Call For Papers
ACM **CIKM 2003** –Call for Industry Track Presentations. We solicit high-quality Industry Track presentati management products and marketplace trends ... Web site: http://www.**cikm**.org/**2003**. Sponsored by ACM SIG
www.cs.wisc.edu/dbworld/messages/2003-06/1057176548.html
More pages from cs.wisc.edu

MMDB'03
... conjunction with The 12th International Conference on Information and Knowledge Management (ACM **CIK** Call for ...
www.cs.fiu.edu/mmdb03
More pages from cs.fiu.edu

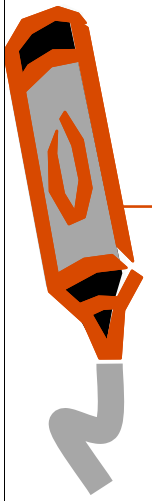
CIKM 2002 Homepage
The **CIKM** 2002 Technical Program has been updated. Hotel information has been updated in the Local Arran international forum for presentation ... The 12th International Conference on Information and Knowledge Mana
cikm.org/2002
More pages from cikm.org

DBWorld Message
Only the highest caliber papers submitted to **CIKM 2003** will be accepted. We have a special interest in papers For Papers 12th International Conference on Information and Knowledge Management (**CIKM**'03 ...
www.cs.wisc.edu/dbworld/messages/2003-03/1051827399.html
More pages from cs.wisc.edu

Past Events
... July 28 - August 1, **2003**. SIGIR **2003** (Toronto) - 26th Annual International ACM SIGIR Conference ... May 28,
www.sigir.org/events/events-past.html
More pages from sigir.org

Upcoming Conference Deadlines
... 5/20/**2003**. ACM **CIKM 2003**. Electronic abstracts deadline [Website] ... 5/28/**2003**. **CIKM 2003**. Paper submis
www-users.cs.umn.edu/~gladmin/conferences
More pages from www-users.cs.umn.edu

web search



web search

alltheweb
find it all

advanced search :: customize preferences :: submit site :: help

cikm 2003 SEARCH

Results in: Any Language English

Web | News | Pictures | Video | Audio

1 - 10 of **18,400** Results for **cikm 2003**

Web Results (What's this?)

Conference on Information and Knowledge Management (**CIKM**)
... The **CIKM** 2004 web page. The **CIKM 2003** Web Page. The **CIKM** 2002 Web Page. The **CIKM** 2001 Web page. Th
more hits from: http://www.cikm.org/ - 6 KB

12. **CIKM 2003**: New Orleans, Louisiana, USA
CIKM 2003: New Orleans, Louisiana, USA. Proceedings of the **2003** ACM **CIKM** International Conference on Inform
ISBN 1-58113-723-0
more hits from: http://www.informatik.uni-trier.de/~ley/db/conf/cikm/cikm2003.html - 55 KB

Department of Computer Science at LSU
News & Events. Department of Computer Science. 298 Coates Hall. Phone: (225)578-1495
more hits from: http://bit.csc.lsu.edu/ - 9 KB

CIKM
International Conference on Information and Knowledge Management (**CIKM**) 14. **CIKM** 2005: Bremen, Germany.
CIKM 2003 Home Page. 11. **CIKM** 2002: McLean, Virginia, USA ...
more hits from: http://sunsite.informatik.rwth-aachen.de/dblp/db/conf/cikm - 10 KB

ACM WIDM'**2003**
CIKM 2003. Past WIDM's. Important Dates: Submission of abstract. June 28, **2003**. Submission full paper : July 5,
Past WIDM's ... Conference on Information and Knowledge Management (**CIKM 2003**) Hotel Inter-Continental, Nev
more hits from: http://www.cais.ntu.edu.sg/widm2003 - 12 KB

ACM **CIKM 2003** Call For Papers
ACM **CIKM 2003** Call for Industry Track Presentations. We solicit high-quality Industry Track presentations which p
marketplace trends ... Web site: http://www.**cikm**.org/**2003**. Sponsored by ACM SIGIR and ACM SIGMIS (pending a
more hits from: http://www.cs.wisc.edu/dbworld/messages/2003-06/1057176548.html - 39 KB

ACM **CIKM 2003** :: Agents Portal :: Agent and Multi-agent Technology Resources
... PRELIMINARY CALL FOR PAPERS. http://**cikm**.org/**2003**/ Twelfth International Conference on Information and ...
more hits from: http://aose.ift.ulaval.ca/modules.php?op=modload&name=News&file=article&sid=86 - 27 KB

MMDB'03
... conjunction with The 12th International Conference on Information and Knowledge Management (ACM **CIKM 20**
more hits from: http://www.cs.fiu.edu/mmdb03 - 6 KB

CIKM 2002 Homepage
The **CIKM** 2002 Technical Program has been updated. Hotel information has been updated in the Local Arrangem
for presentation ... The 12th International Conference on Information and Knowledge Management (**CIKM 2003**) ...
more hits from: http://cikm.org/2002 - 15 KB

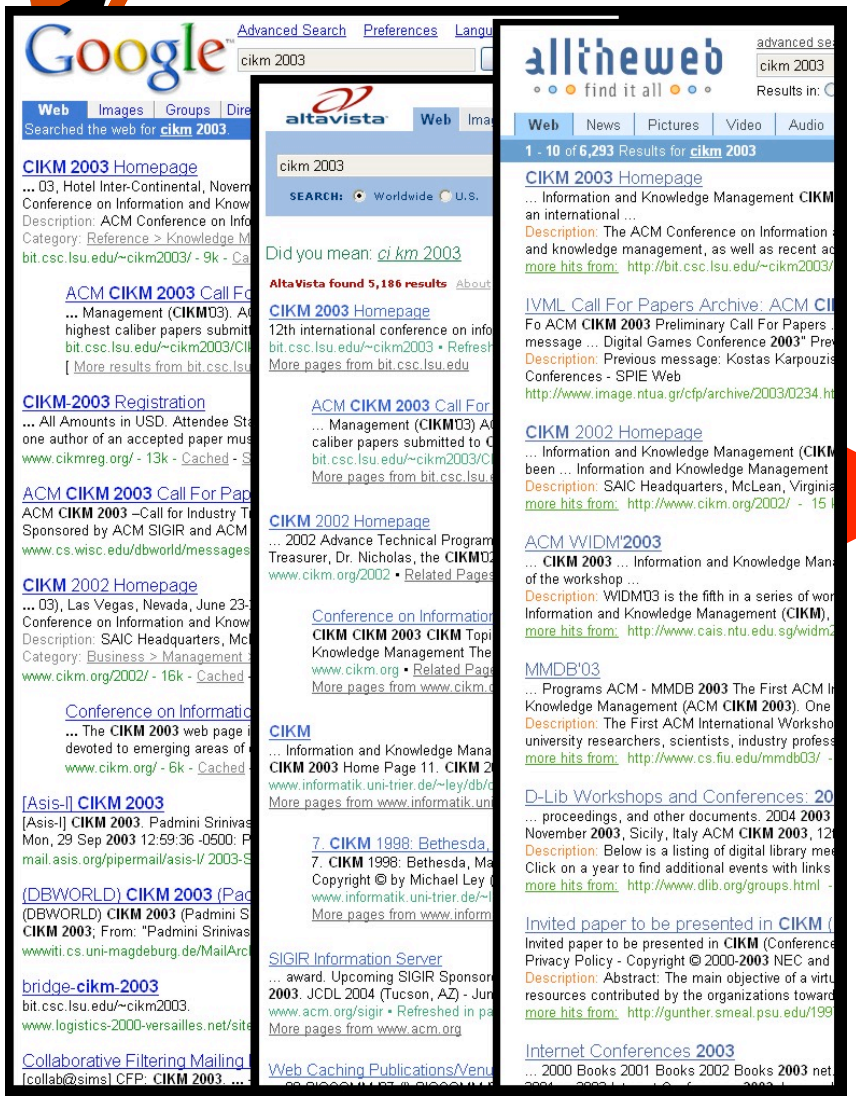
DBWorld Message
Only the highest caliber papers submitted to **CIKM 2003** will be accepted. We have a special interest in papers tha
International Conference on Information and Knowledge Management (**CIKM**'03 ...
more hits from: http://www.cs.wisc.edu/dbworld/messages/2003-03/1051827399.html - 4 KB

metasearch


Google
Advanced Search
Preferences
cikm 2003
Web
Images
Groups
CIKM 2003 Homepage
CIKM-2003 Registration
ACM CIKM 2003 Call For Pap
CIKM 2002 Homepage
[Asis-l] CIKM 2003
(DBWORLD) CIKM 2003 (Pac
bridge-cikm-2003
Collaborative Filtering Mailing
altavista
Web
cikm 2003
SEARCH:
Worldwide
U.S.
Did you mean: ci km 2003
AltaVista found 5,186 results
CIKM 2003 Homepage
ACM CIKM 2003 Call For
CIKM 2002 Homepage
Conference on Information
CIKM
7. CIKM 1998: Bethesda,
SIGIR Information Server
Web Caching Publications/Venu
alltheweb
find it all
advanced se
cikm 2003
Results in:
Web
News
Pictures
Video
Audio
1 - 10 of 6,293 Results for cikm 2003
CIKM 2003 Homepage
IVML Call For Papers Archive: ACM CI
CIKM 2002 Homepage
ACM WIDM'2003
MMDB'03
D-Lib Workshops and Conferences: 20
Invited paper to be presented in CIKM (
Internet Conferences 2003


metasearch
CIKM 2003 Homepage
ACM CIKM 2003 Call For Pape
CIKM–2003 Registration
ACM CIKM 2003 Call For Pape
CIKM 2002 Homepage
Conference on Information an
[Asis–l] CIKM 2003
(DBWORLD) CIKM 2003 (Padm
bridge–cikm–2003
Collaborative Filtering Mailing
CIKM Home Page ACM DL: CIK
Yahoo! Groups : webir Messag
dBforums – Cfp: Cikm 03
Mailing List ARL–ERESERVE@a
Received: from cni.org by b.cr
ACM WIDM 2003
ACM – MMDB 2003
ACM CIKM 2003 PRELIMINARY
Selected Publications
Web Caching Publications/Ver
cikm '03
Call for papers
Mario A. Nascimento – Person
Iceved
Conferences and Journals on T
Conferences On Information V
Calendrier des manifestations
(DBWORLD) Final Call for ACM
Collaborative Filtering Mailing

user feedback



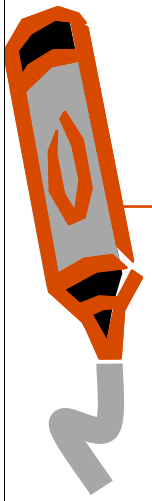
metasearch

- CIKM 2003 Homepage
- ACM CIKM 2003 Call For Pape
- CIKM-2003 Registration
- ACM CIKM 2003 Call For Pape
- CIKM 2002 Homepage
- Conference on Information an
- [Asis-l] CIKM 2003
- (DBWORLD) CIKM 2003 (Padm
- bridge-cikm-2003
- Collaborative Filtering Mailing
- CIKM Home Page ACM DL: CIK
- Yahoo! Groups : webir Messag
- dBforums - Cfp: Cikm 03
- Mailing List ARL-ERESERVE@a
- Received: from cni.org by b.cr
- ACM WIDM 2003
- ACM - MMDB 2003
- ACM CIKM 2003 PRELIMINARY
- Selected Publications
- Web Caching Publications/Ver
- cikm '03
- Call for papers
- Mario A. Nascimento - Person
- Iceved
- Conferences and Journals on T
- Conferences On Information V
- Calendrier des manifestations
- (DBWORLD) Final Call for ACM
- Collaborative Filtering Mailing

metasearch

- CIKM 2003 Homepage
- CIKM-2003 Registration
- CIKM 2002 Homepage
- ACM CIKM 2003 Call For Pape
- ACM CIKM 2003 Call For Pape
- CIKM Home Page ACM DL: CIK
- Conference on Information an
- ACM CIKM 2003 PRELIMINARY
- CIKM Home Page ACM DL: CIK
- ACM CIKM 2003 PRELIMINARY
- [Asis-l] CIKM 2003
- (DBWORLD) CIKM 2003 (Padm
- bridge-cikm-2003
- Collaborative Filtering Mailing
- Yahoo! Groups : webir Messag
- dBforums - Cfp: Cikm 03
- Mailing List ARL-ERESERVE@a
- Received: from cni.org by b.cr
- ACM WIDM 2003
- ACM - MMDB 2003
- Selected Publications
- Web Caching Publications/Ver
- cikm '03
- Call for papers
- Mario A. Nascimento - Person
- Iceved
- Conferences and Journals on T
- Conferences On Information V
- Calendrier des manifestations

text management applications

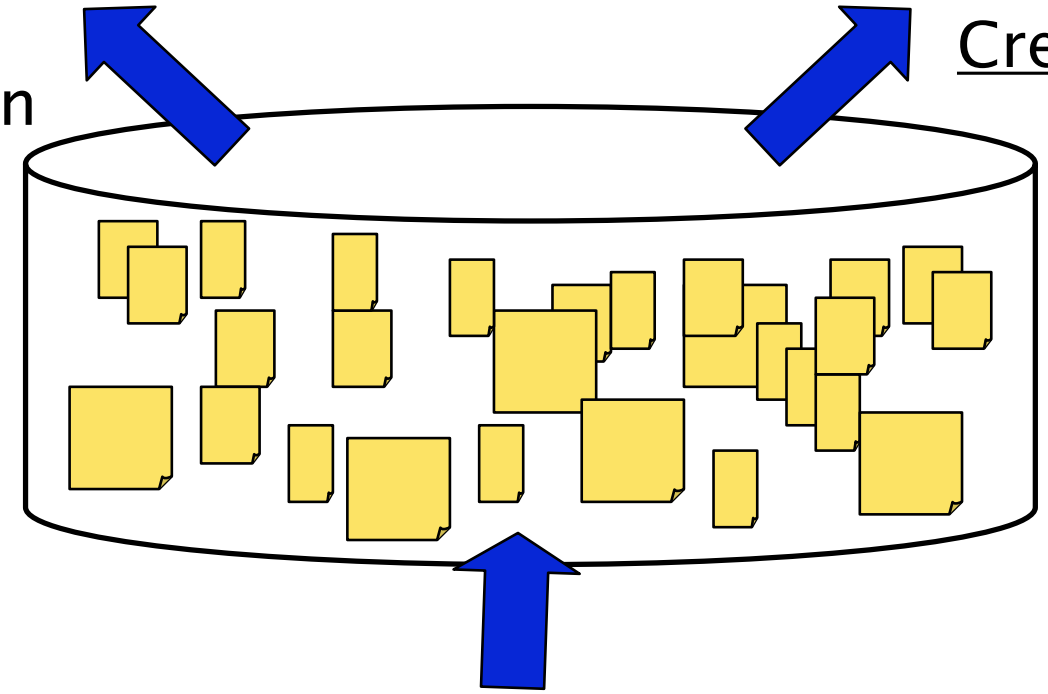


Access

Select information

Mining

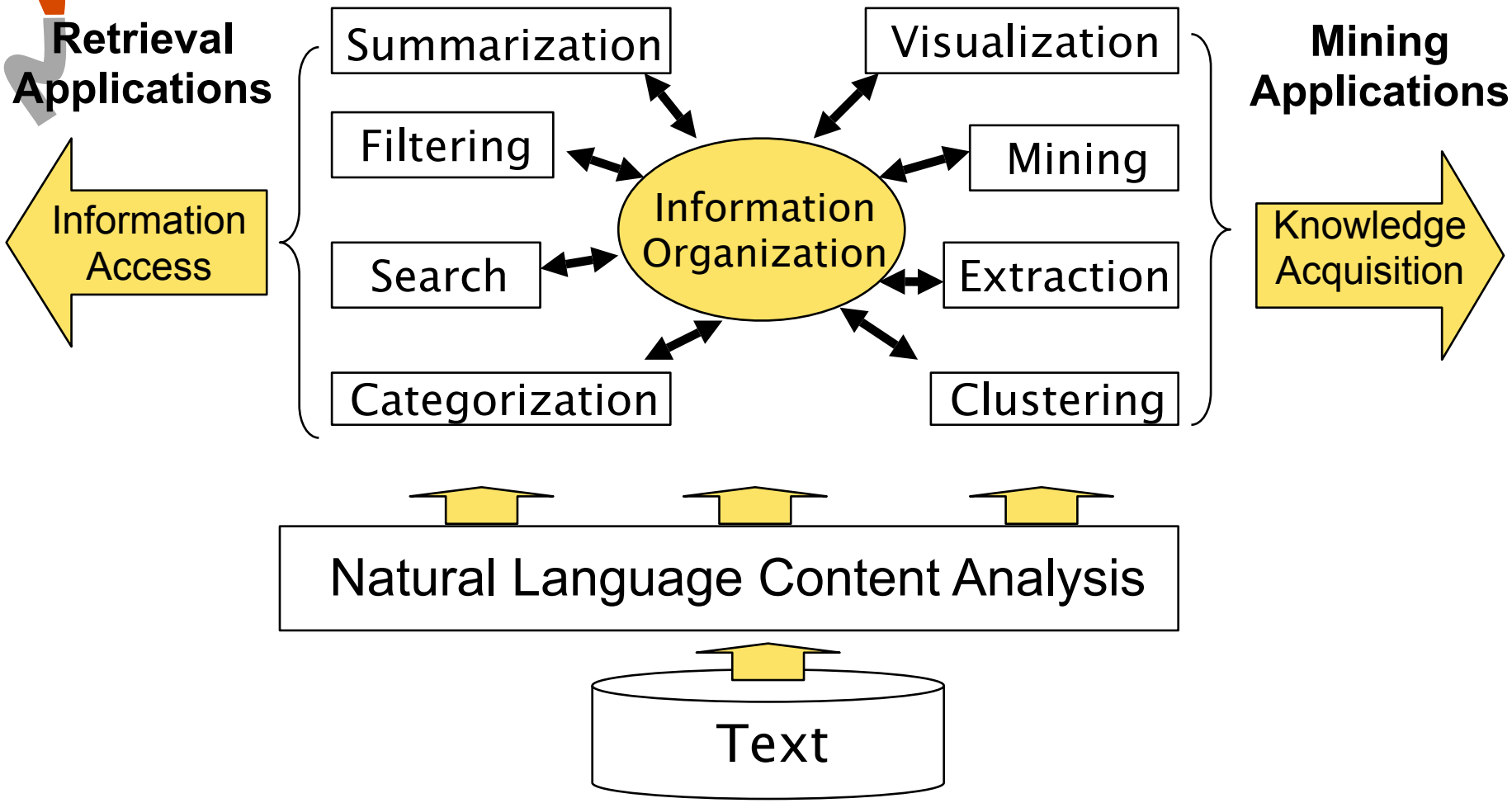
Create Knowledge



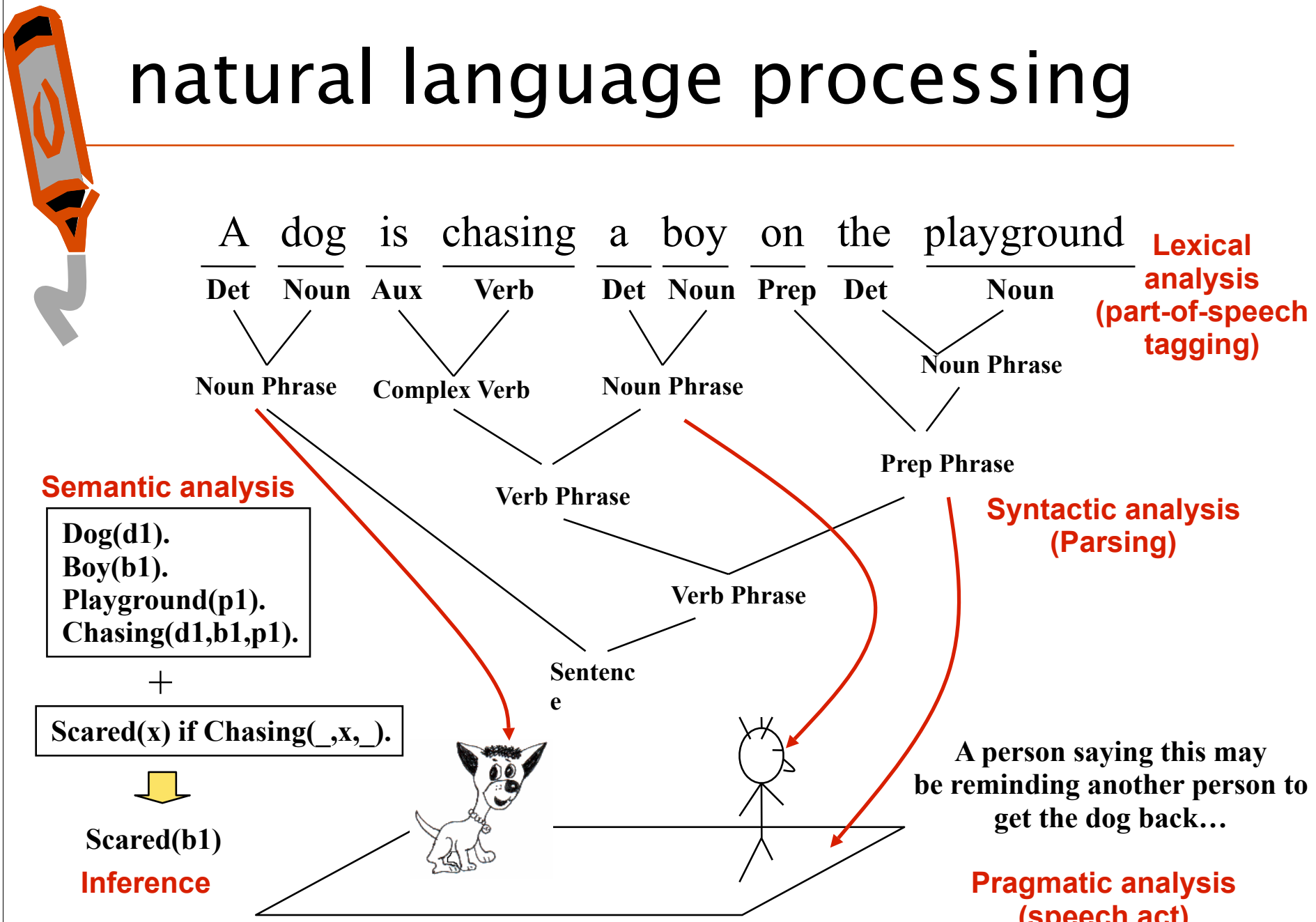
Organization

Add Structure/Annotations

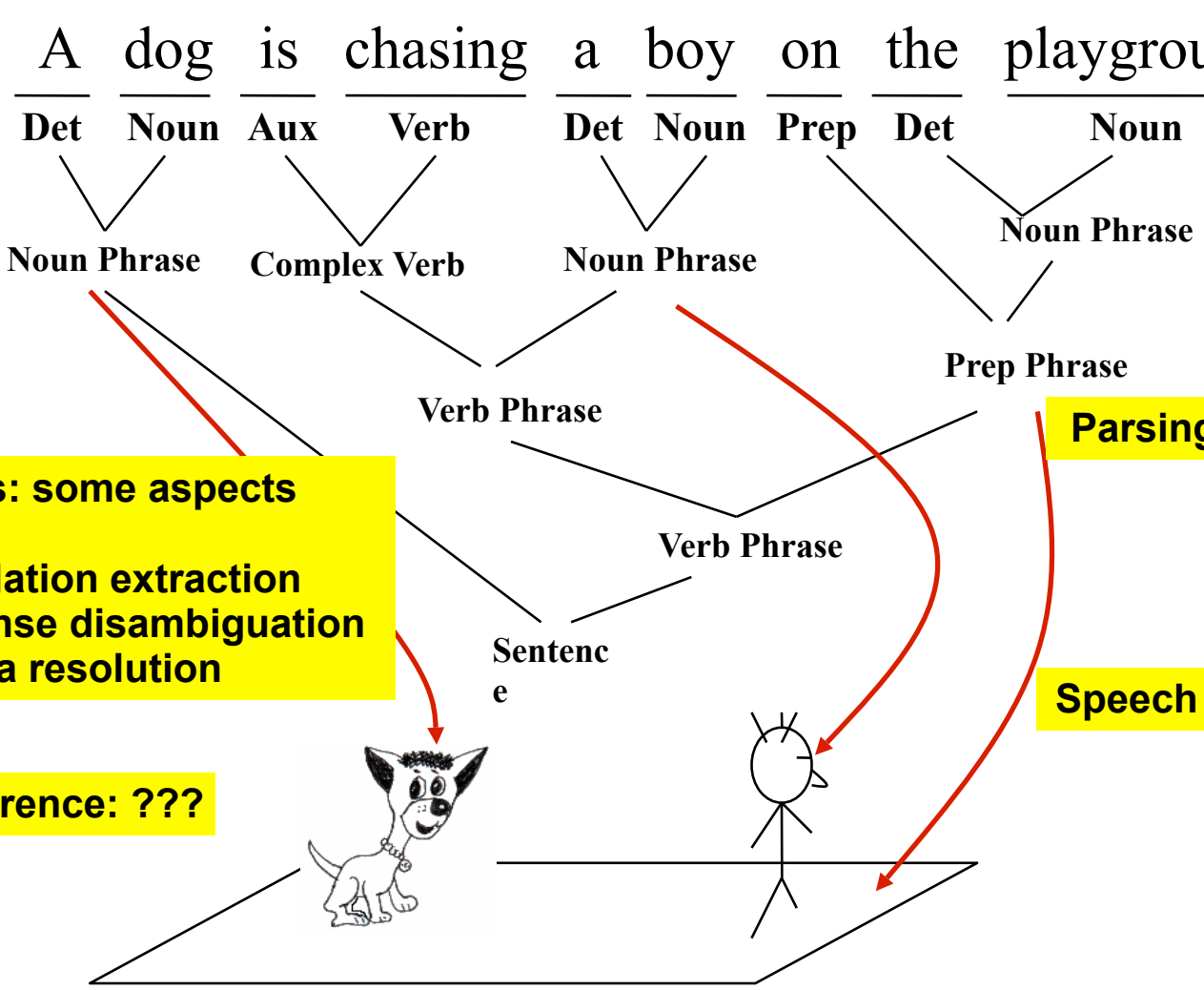
text management applications



natural language processing



what we can do in NLP



POS Tagging: 97%

Parsing: partial >90%(?)

Semantics: some aspects

- **Entity/relation extraction**
- **Word sense disambiguation**
- **Anaphora resolution**

Speech act analysis: ???

Inference: ???

natural language processing

Arabic text

... يَجِبُ عَلَى الإِنْسَانِ أن يَكُونَ أمِيْنَاً وَصَادِقَاً مَعَ نَفْسِهِ وَمَعَ أَهْلِهِ وَجِيْرَانِهِ وَأَنْ يَبْذُلَ كُلَّ جُهْدٍ فِي إِعْلاءِ شَأْنِ الوَطَنِ وَأَنْ يَعْمَلَ عَلَى ما ...

How can a computer make **sense** out of this **string**?

Morphology - What are the basic units of meaning (words)?
- What is the meaning of each word?

Syntax - How are words related with each other?

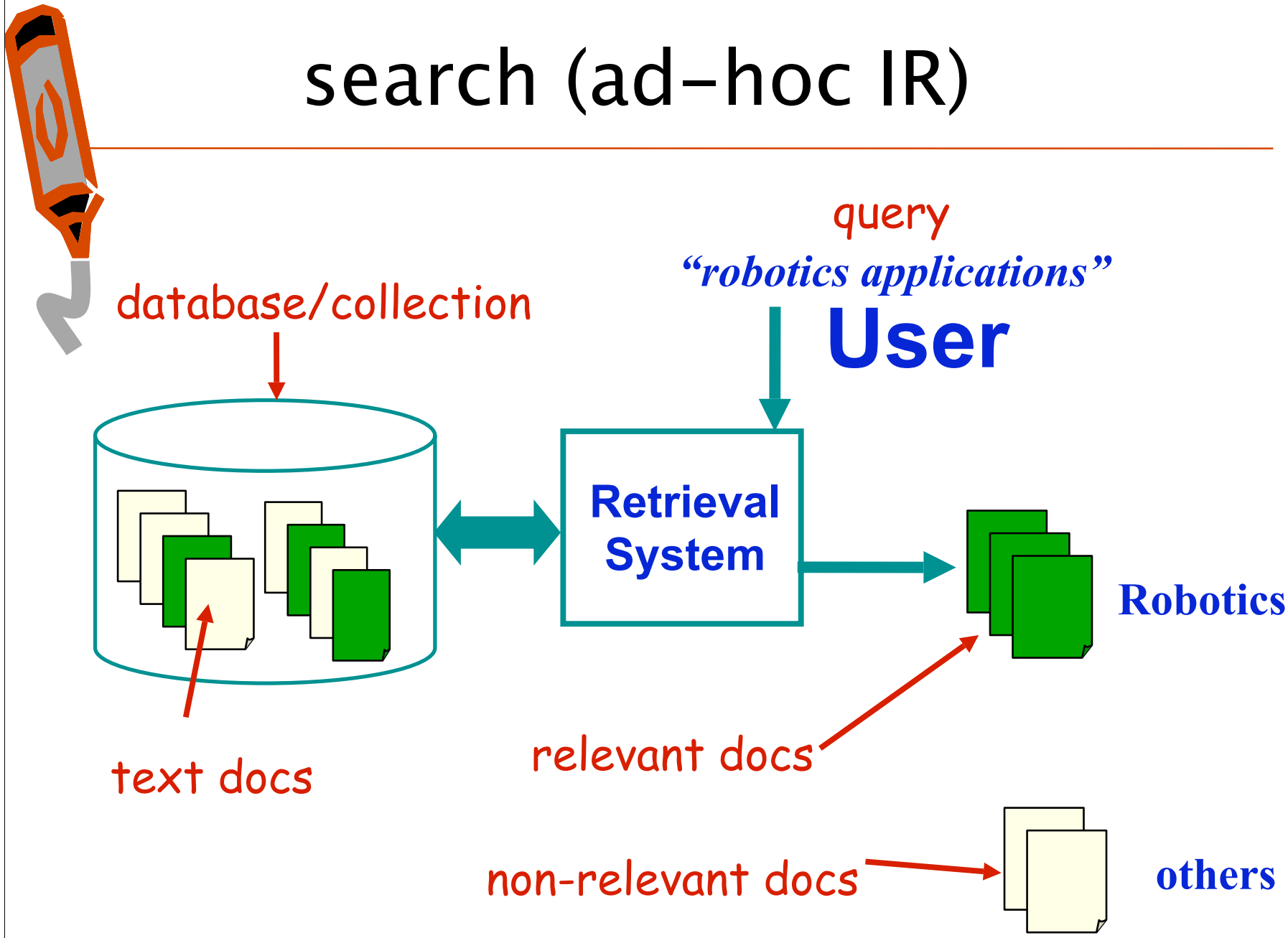
Semantics - What is the "combined meaning" of words?

Pragmatics - What is the "meta-meaning"? (speech act)

Discourse - Handling a large chunk of text

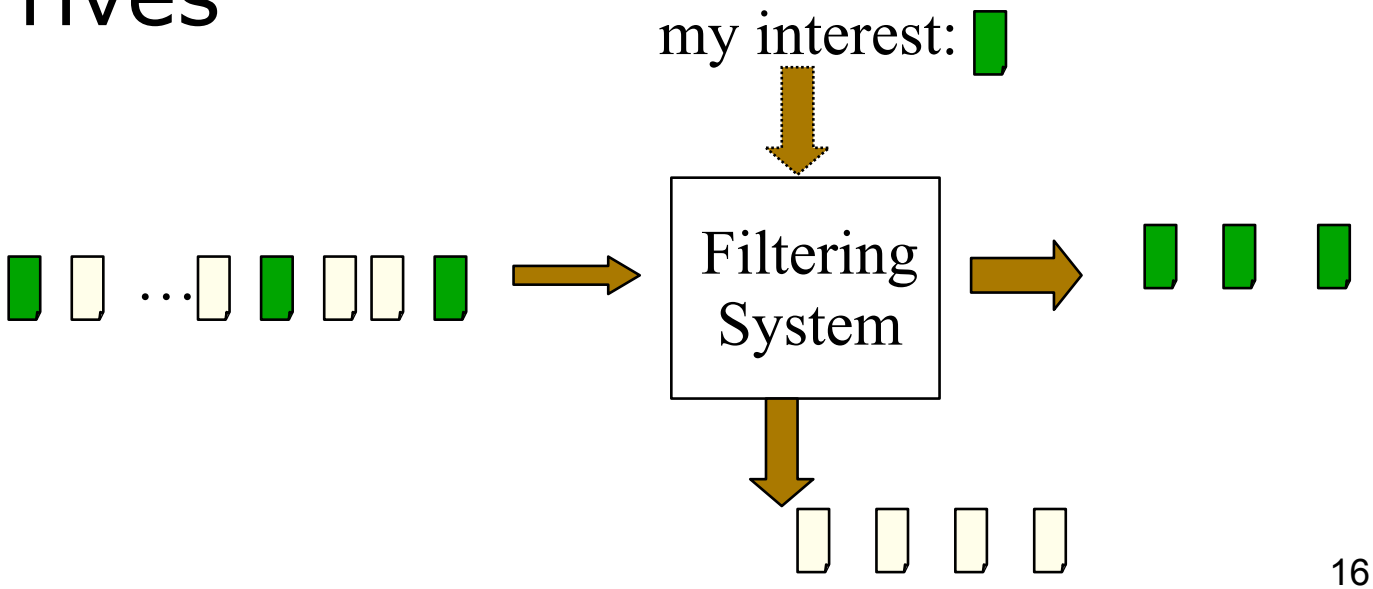
Inference - Making sense of everything

search (ad-hoc IR)

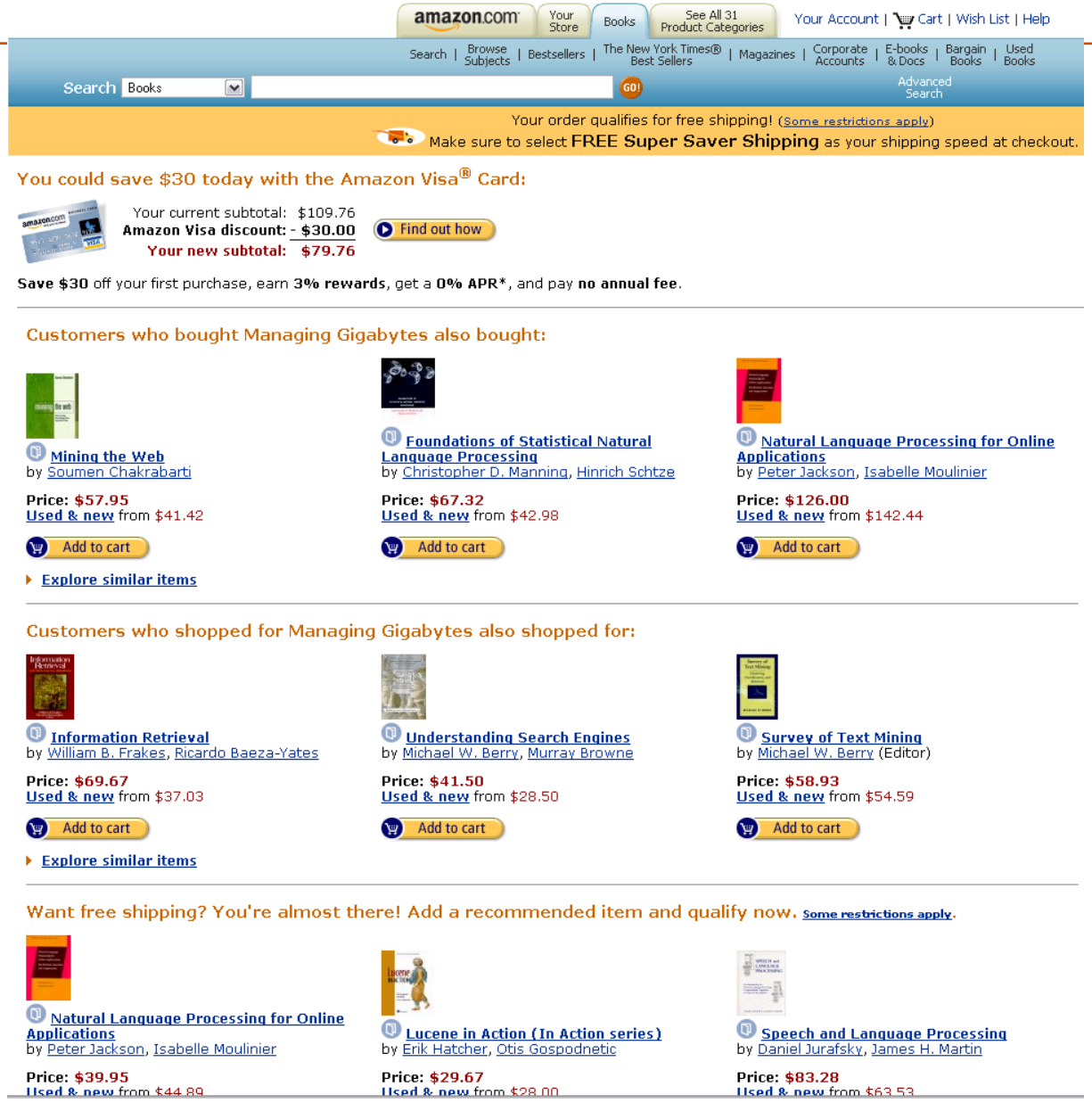


information filtering

- Stable & long term interest, dynamic info source
- System must make a delivery decision immediately as a document "arrives"

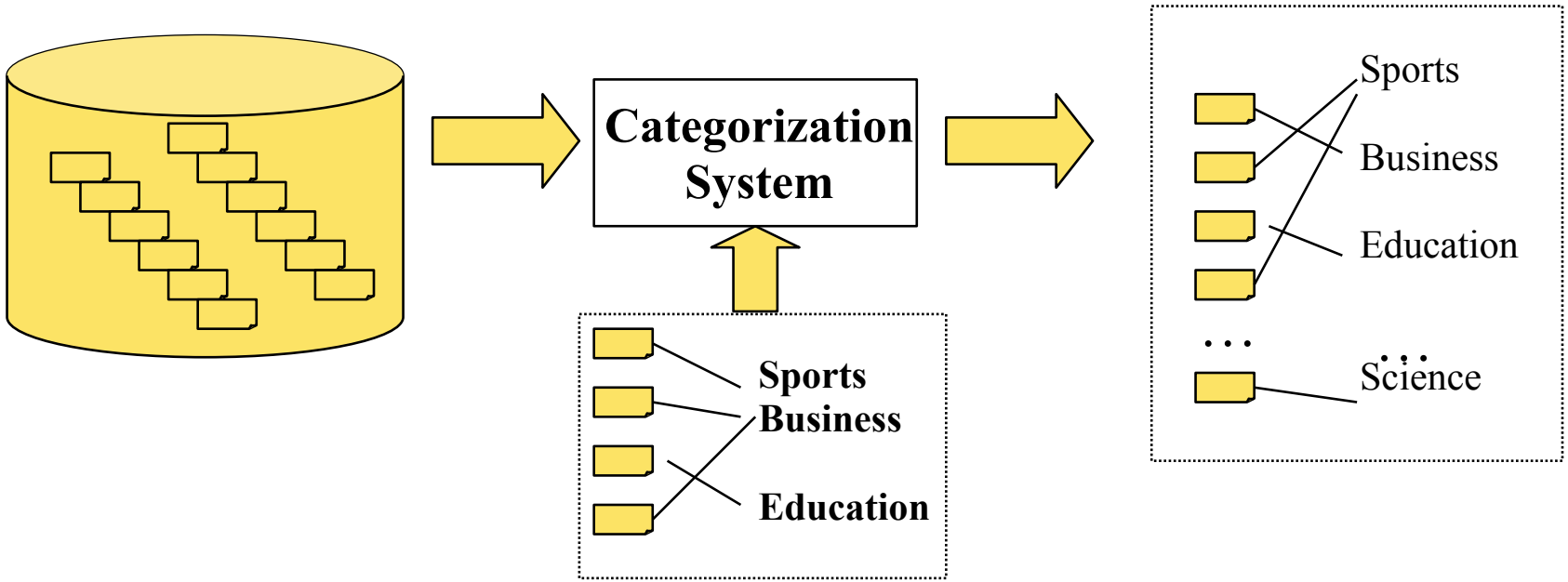


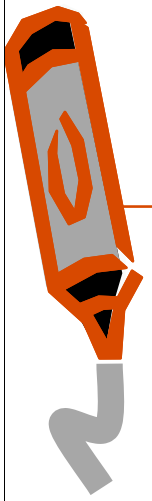
collaborative filtering



categorization

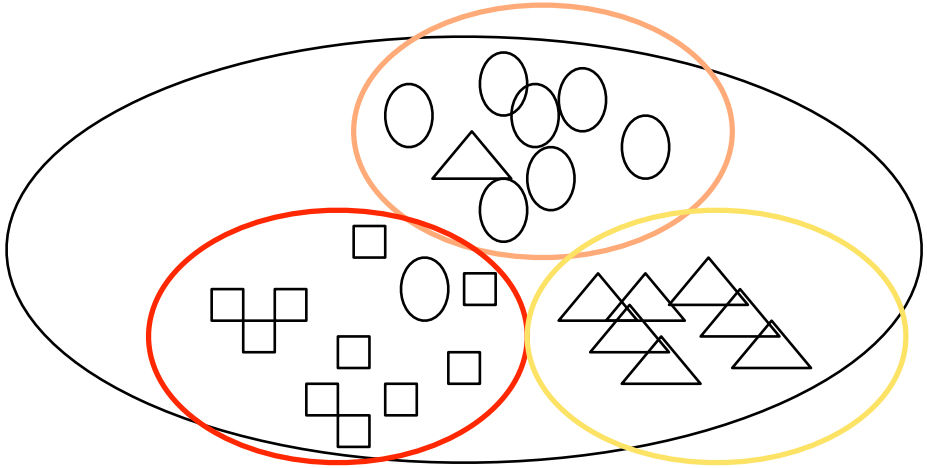
- Pre-given categories and labeled document examples (Categories may form hierarchy)
- Classify new documents
- A standard supervised learning problem

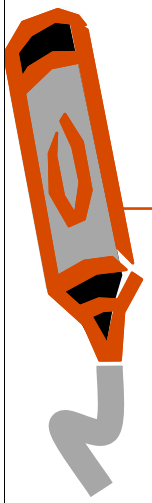




clustering

- Discover “natural structure”
- Group similar objects together
- Object can be document, term, passages

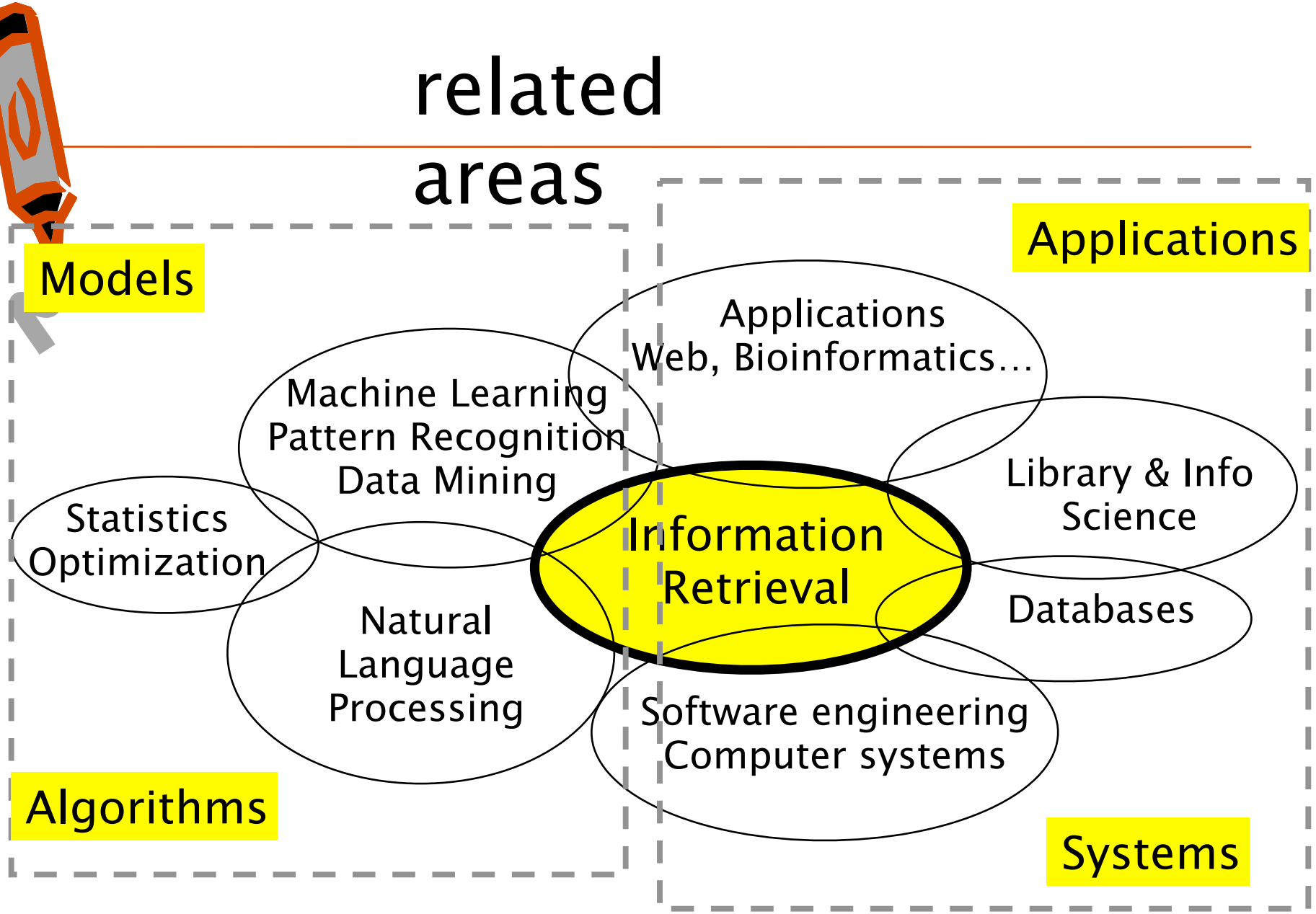




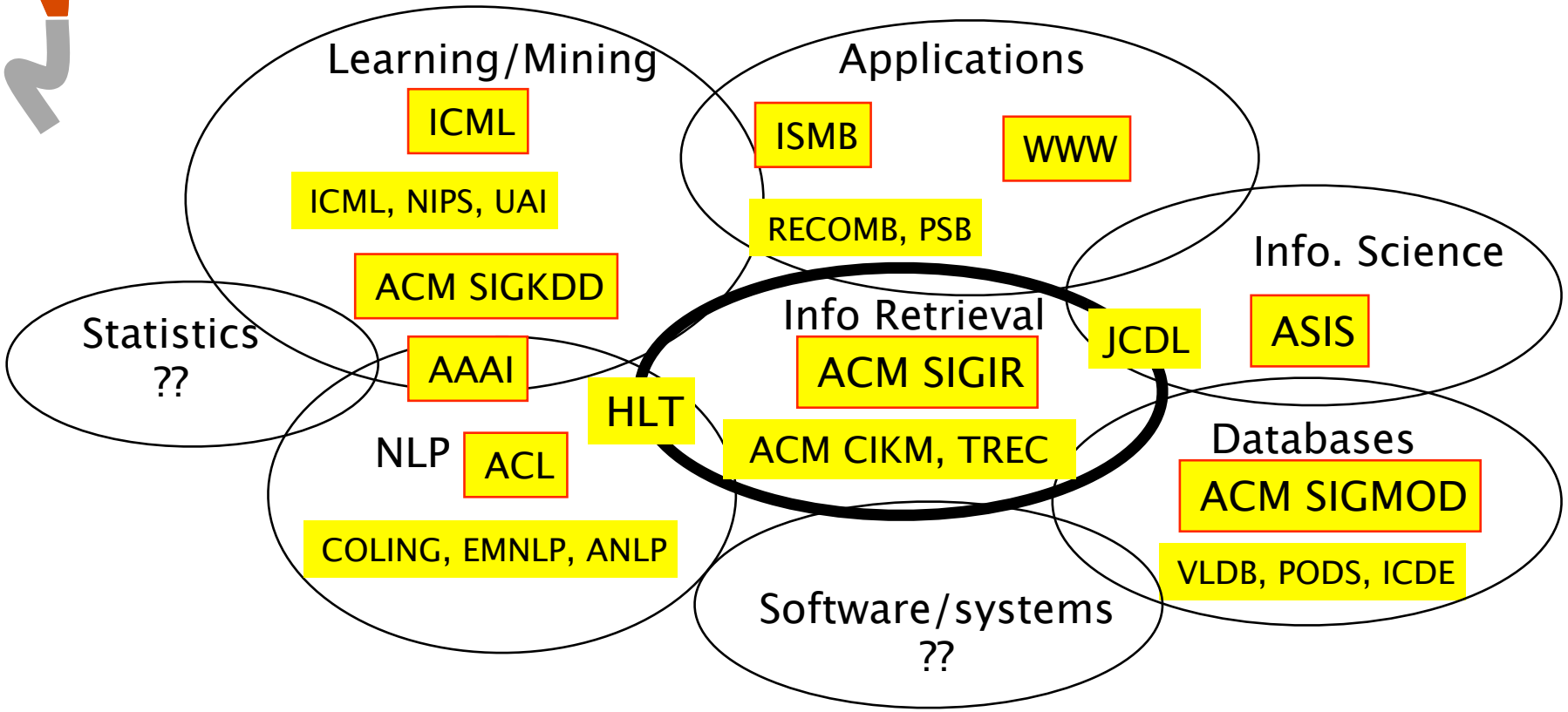
IR vs Databases

	Databases	IR
Data	**Structured**	**Unstructured**
Fields	**Clear semantics** (SSN, age)	**No fields** (other than text)
Queries	**Defined** (relational algebra, SQL)	**Free text** ("natural language"), Boolean
Recoverability	**Critical** (concurrency control, recovery, atomic operations)	**Downplayed**, though still an issue
Matching	**Exact** (results are *always* "correct")	**Imprecise** (need to measure effectiveness)

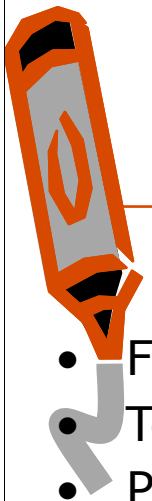
related areas



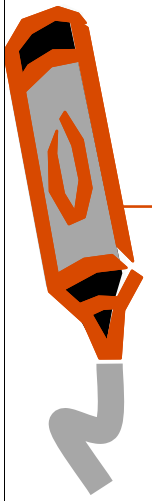
publications/societies



SIGIR 2003 topics

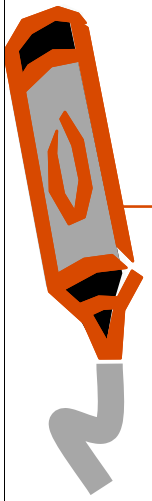


- Formal Models, Language Models, Fusion/Combination
- Text Representation and Indexing, XML and Metadata
- Performance, Compression, Scalability, Architectures, Mobile Applications
- Web IR, Intranet/Enterprise Search, Citation and Link Analysis, Digital Libraries, Distributed IR
- Cross-language Retrieval, Multilingual Retrieval, Machine Translation for IR
- Video and Image Access, Audio and Speech Retrieval, Music Retrieval
- Machine Learning for IR, Text Data Mining, Clustering, Text Categorization
- Topic Detection and Tracking, Content-Based Filtering, Collaborative Filtering, Agents
- Summarization, Question Answering, Natural Language Processing for IR, Information Extraction, Lexical Acquisition
- Interfaces, Visualization, Interactive IR, User Models, User Studies
- Specialized Applications of IR, including Genomic IR, IR in Software Engineering, and IR for Chemical Structures



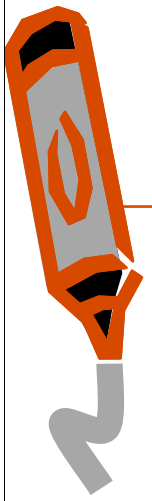
overview

- what is information retrieval ?
- how does it work ?
- a simple retrieval model



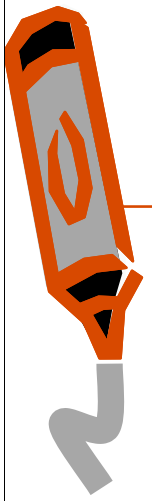
basic approaches

- boolean
- geometric : vector space model
- probabilistic : language models
- statistical : bayesian networks
- graph like : page rank



relevant items are similar

- Much of IR depends upon idea that
  similar vocabulary → relevant to same queries
- Usually look for documents matching query words
- "Similar" can be measured in many ways
  - String matching/comparison
  - Same vocabulary used
  - Probability that documents arise from same model
  - Same meaning of text

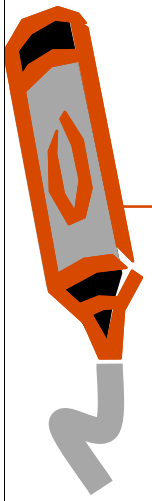


bag of words

- An effective and popular approach
- Compares words without regard to order
- Consider reordering words in a headline

– **Random**: beating takes points falling another Dow 355
– **Alphabetical**: 355 another beating Dow falling points
– "**Interesting**": Dow points beating falling 355 another

– **Actual**: **Dow takes another beating, falling 355 points**



what is this about ?

16 × said
14 × McDonalds
12 × fat
11 × fries
8 × new
6 × company french nutrition
5 × food oil percent reduce taste Tuesday
4 × amount change health Henstenburg make obesity
3 × acids consumer fatty polyunsaturated US
2 × amounts artery Beemer cholesterol clogging director down eat estimates expert fast formula impact initiative moderate plans restaurant saturated trans win
1 × …
added addition adults advocate affect afternoon age Americans Asia battling beef bet brand Britt Brook Browns calorie center chain chemically … crispy customers cut … vegetable weapon weeks Wendys Wootan worldwide years York

the text

McDonald's slims down spuds

Fast-food chain to reduce certain types of fat in its french fries with new cooking oil.

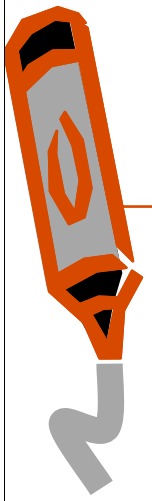
NEW YORK (CNN/Money) - McDonald's Corp. is cutting the amount of "bad" fat in its french fries nearly in half, the fast-food chain said Tuesday as it moves to make all its fried menu items healthier.

But does that mean the popular shoestring fries won't taste the same? The company says no. "It's a win-win for our customers because they are getting the same great french-fry taste along with an even healthier nutrition profile," said Mike Roberts, president of McDonald's USA.

But others are not so sure. McDonald's will not specifically discuss the kind of oil it plans to use, but at least one nutrition expert says playing with the formula could mean a different taste.

Shares of Oak Brook, Ill.-based McDonald's (MCD: down $0.54 to $23.22, Research, Estimates) were lower Tuesday afternoon. It was unclear Tuesday whether competitors Burger King and Wendy's International (WEN: down $0.80 to $34.91, Research, Estimates) would follow suit.

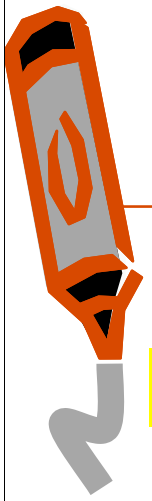
Neither company could immediately be reached for comment.



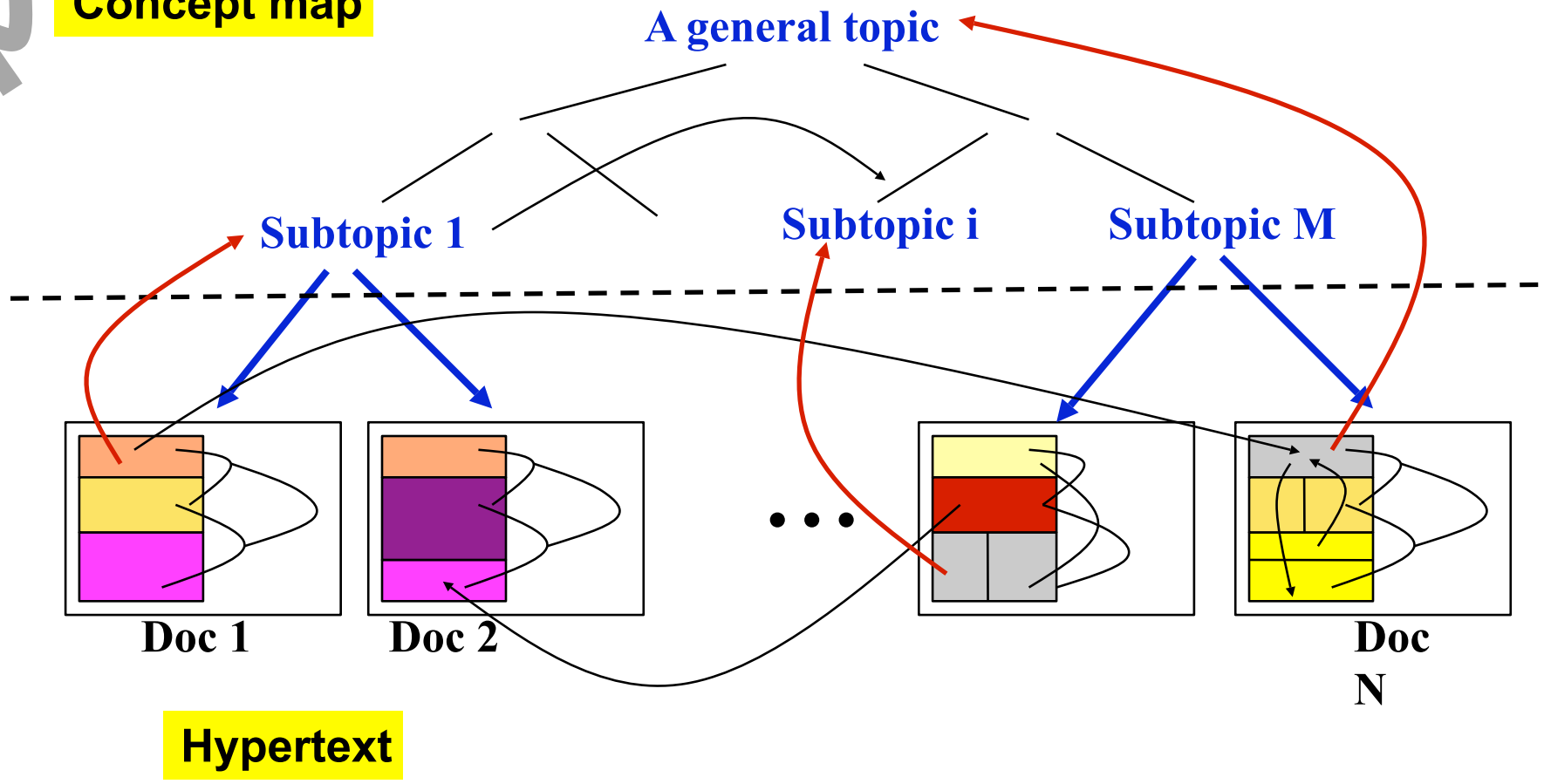
text representation

- Text representation
  - what makes a "good" representation?
  - how is a representation generated from text?
  - what are retrievable objects and how are they organized?
- Representing information needs
  - what is an appropriate query language?
  - how can interactive query formulation and refinement be supported?
- Comparing representations
  - what is a "good" model of retrieval?
  - how is uncertainty represented?

hypertext

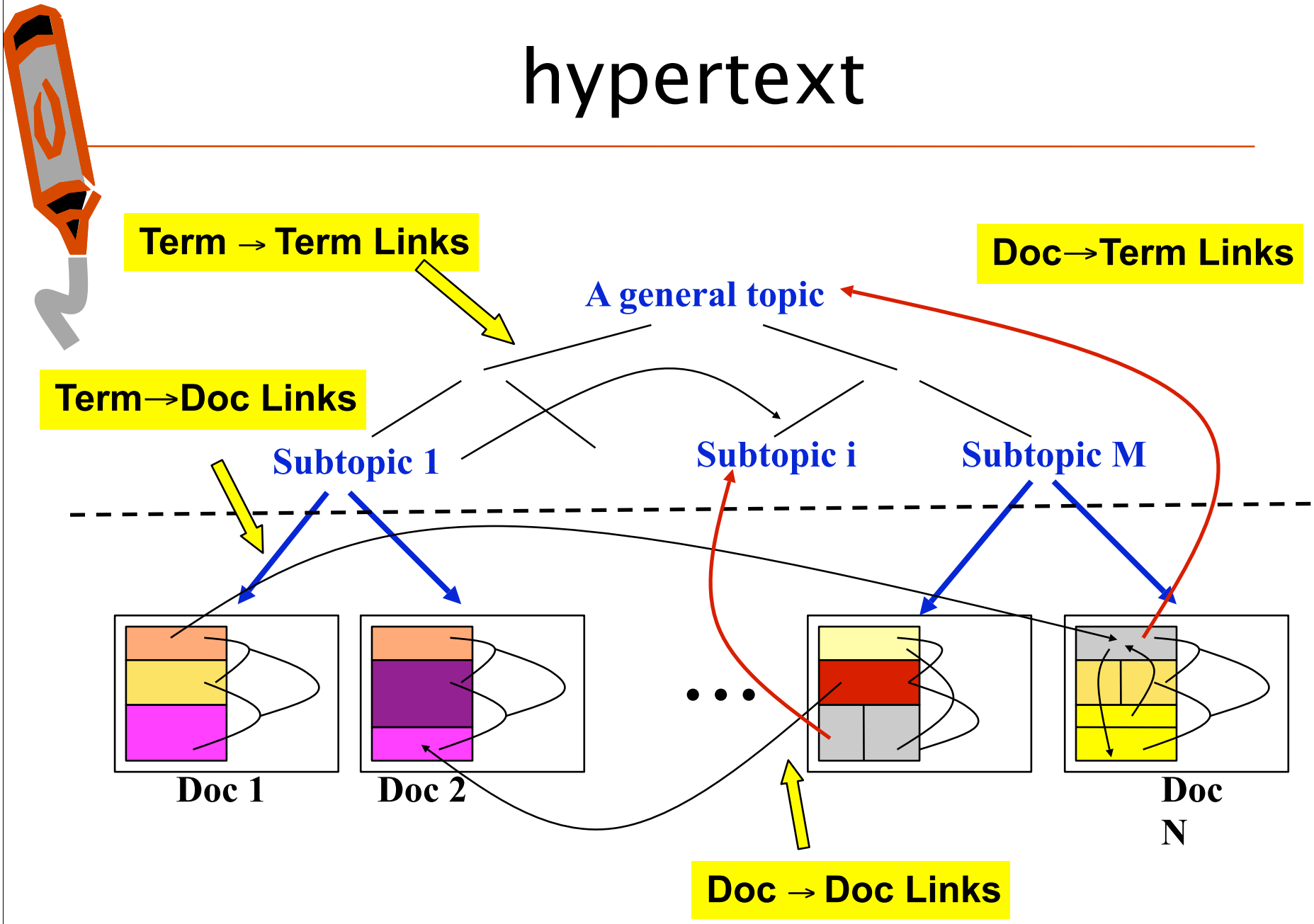


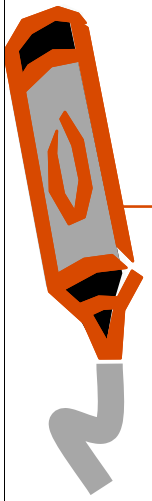
Concept map



Hypertext

hypertext

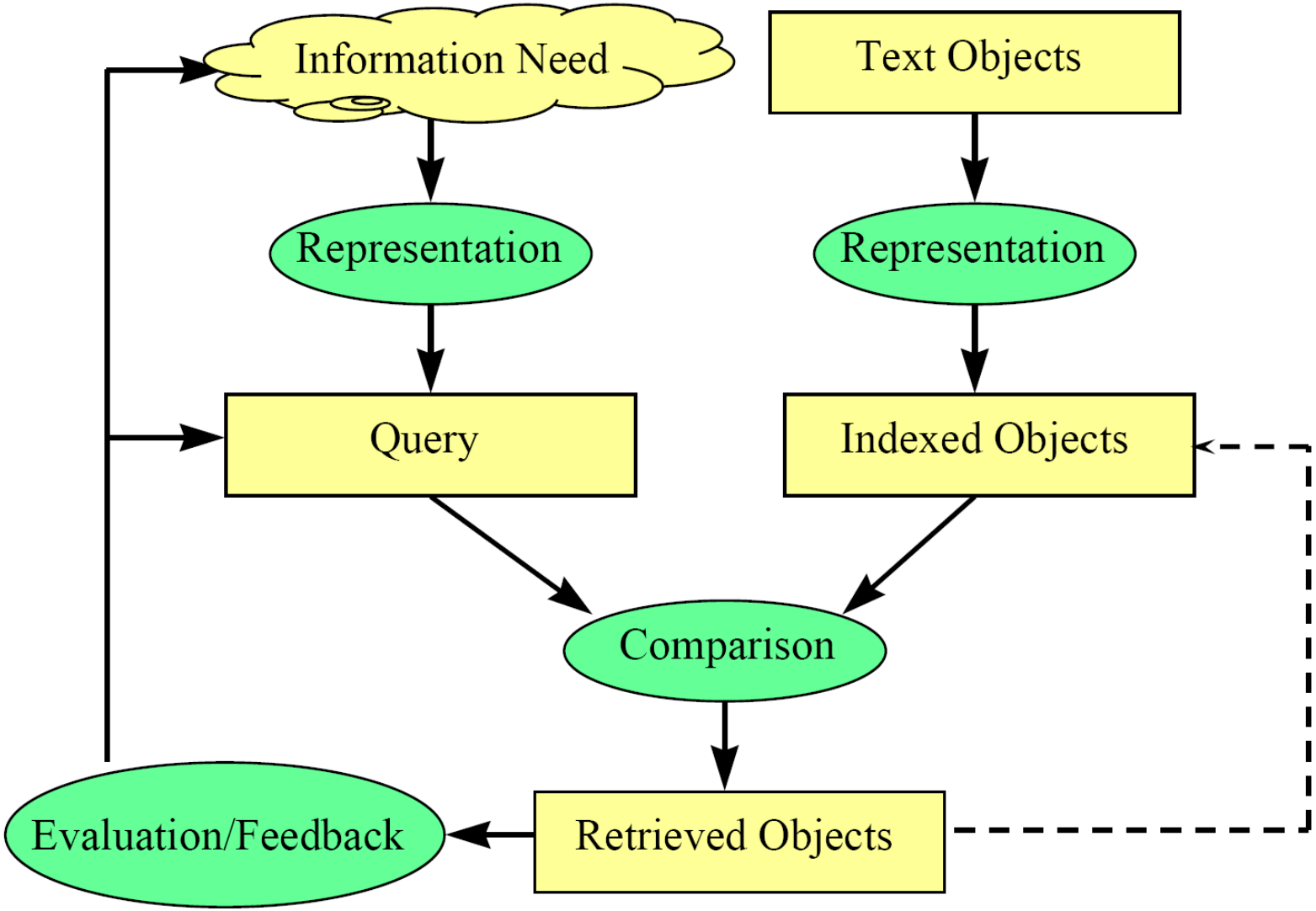

Term → Term Links
Doc→Term Links
A general topic
Term→Doc Links
Subtopic 1
Subtopic i
Subtopic M
Doc 1
Doc 2
Doc N
Doc → Doc Links

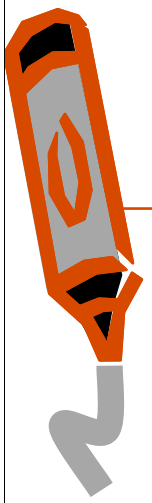


overview

- what is information retrieval ?
- how does it work ?
- a simple retrieval model

retrieval process





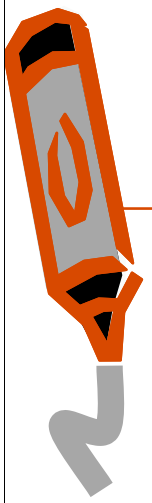
statistical language model

D = { One fish, two fish, red fish, blue fish.
Black fish, blue fish, old fish, new fish.

len(D) = 16

P(fish|D) = 8/16 = 0.5
P(blue|D) = 2/16 = 0.125
P(one|D) = 1/16 = 0.0625
…
P(eggs|D) = 0/16 = 0
…

A “topic”



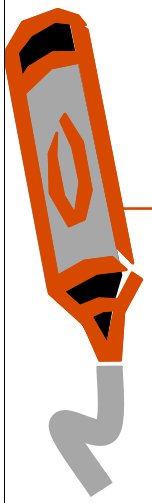
statistical language model

- Document came from a topic
- Did query come from *this* document's topic?

- For each document, find probability its topic could have generated the query

$$
\begin{aligned}
P(Q|T_D) &\approx P(Q|D) \\
&= P(q_1, \ldots, q_t|D) \\
&= \prod_{i=1}^{t} P(q_i|D)
\end{aligned}
$$

Independence assumption (Naïve Bayes)



statistical language model

$D_1 =$ { This one, I think, is called a Yink.
He likes to wink, he likes to drink.

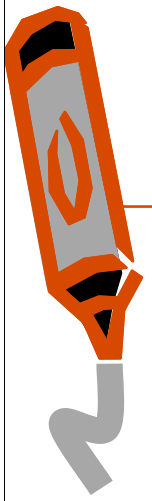
$D_2 =$ { He likes to drink, and drink, and drink.
The thing he likes to drink is ink.

$D_2 =$ { The ink he likes to drink is pink.
He links to wink and drink pink ink.

Query "drink"
- $P(drink|D_1) = 1/16$
- $P(drink|D_2) = 4/16$
- $P(drink|D_3) = 2/16$

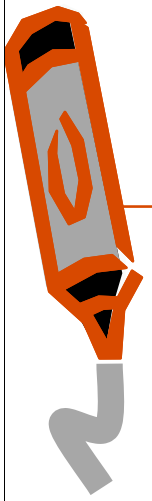
Query "pink ink"
- $P(Q|D_1) = 0 \cdot 0 = 0$
- $P(Q|D_2) = 0 \cdot 1/16 = 0$
- $P(Q|D_3) = 2/16 \cdot 2/16 = 0.016$

Query "wink drink"
- $P(Q|D_1) = 0.004$
- $P(Q|D_2) = 0$
- $P(Q|D_3) = 1/16 \cdot 2/16 = 0.008$



conclusion

- Information Retrieval?
  - Indexing, retrieving, and organizing text by probabilistic or statistical techniques that reflect semantics without actually understanding
  - Search engines
- Core idea
  - Bag of words captures much of the "meaning"
  - Objects that use vocabulary the same way are related
- Statistical language model
  - Documents used to estimate a topic model
  - Query reflects a topic, too
  - Documents of topics that are likely to produce the query are most likely to be relevant



in this course

- http://ccs.neu.edu/~jaa/ISU535.06X2/schedulen.html