Cross-Modal Analysis

Alex Hauptmann
School of Computer Science
Carnegie Mellon University,
Pittsburgh, PA, USA
alex@cs.cmu.edu
http://www.cs.cmu.edu/~alex

Outline - Day 1

- Introduction/Background
  - Demonstration examples
- Text, Speech, Images, Video
- Case Study 1: News Subject Monologues
- Case Study 2: Labeling Faces and Locations

- Retrieval across Modalities
  - Combination Models for Video Retrieval

  - Cross-Modal Relevance Feedback in Video Retrieval

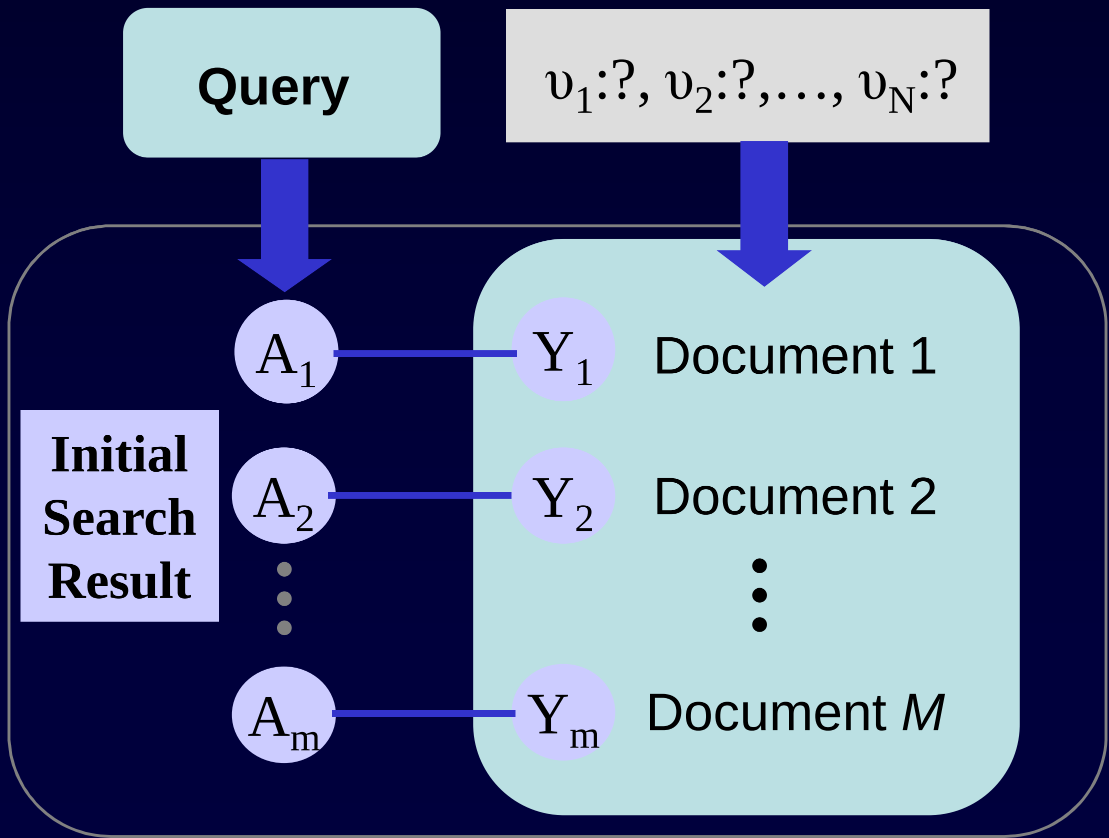
Outline - Day 2

- Bridging the Semantic Gap across Modalities
  - A Large Scale Ontology for Multimedia (LSCOM)
  - Retrieval Experiments with LSCOM

Relevance Feedback

Probabilistic Local Context Analysis (pLCA)

- **Goal:** automatically leverage useful features for current query
- **Method:** assume combination parameters υ of "un-learned" ranking features to be latent variables instead of zero



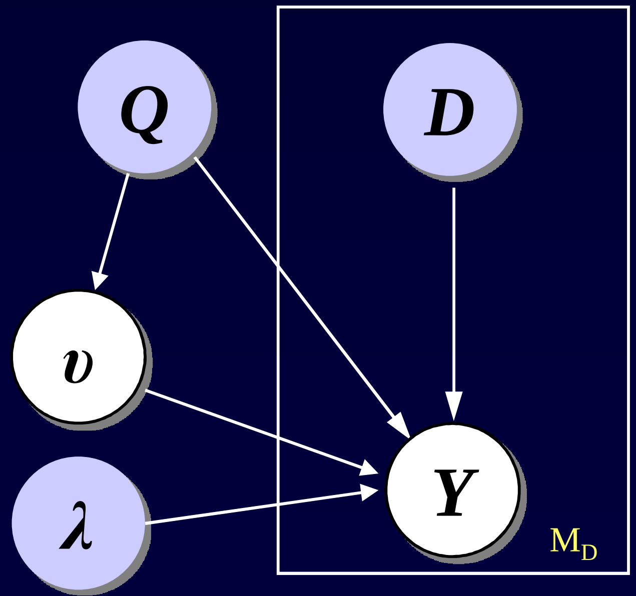
Model and Parameter Estimation

- Jointly optimize weights of query-specific sources ***u*** and answer relevance ***Y*** in initial retrieved shots

$$\max_{y,\nu} P(\nu \mid Q)\cdot\prod_{j=1}^{M_D}\sigma\left(\sum_{i=0}^{N} y_j\lambda_i P\left(S_i \mid D_j, Q\right)+\sum_{l=1}^{M} y_j\nu_l P\left(S_l^{*} \mid D_j, Q\right)\right)$$

(Prior) (Query Independent) (Query Specific)

- Iteratively maximize ***Y*** and ν
  - ***Y***: top-ranked shots as positive and others as negative
  - ν: feature selection + regularized logistic regression

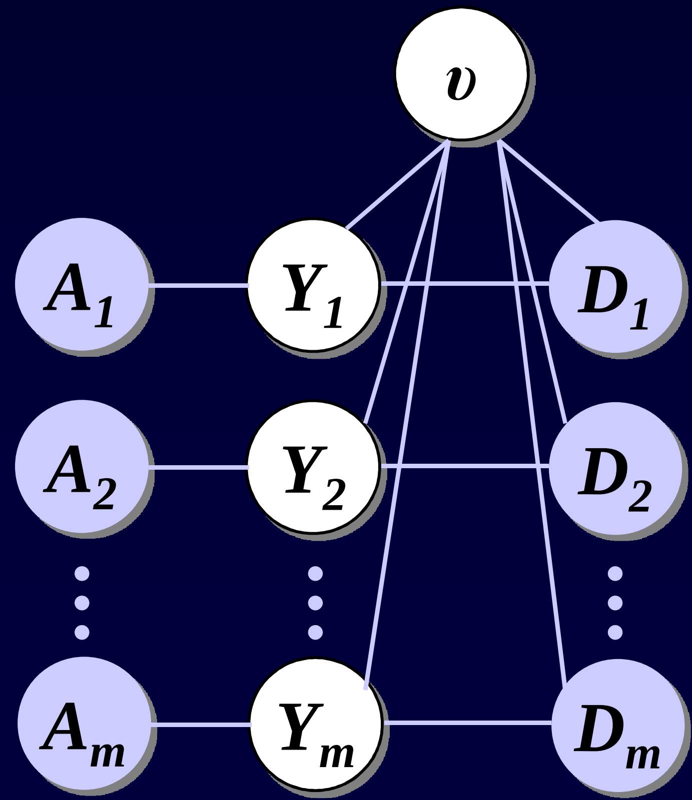


pLCA: Undirected Model and Parameter Estimation

- Compute the posterior probability of document relevance *Y* given initial results A based on an undirected graphical model

$$P(\vec{y} \mid \vec{a}; \vec{D}, Q) = \frac{1}{Z} \int_{\nu} \prod_{l} P(\nu_l \mid Q; \nu_l^0) \cdot \prod_{j=1}^{M_D} \exp\left(y_j a_j + y_j \sum_{l} \nu_l f_l (D_j, Q) \right) d\nu$$

- Variational inference
  - Approximate the posterior distribution of **p(y,ν|a)** by a family of variational distributions **q(y, ν)** where y and ν are independent
  - Iteratively maximize a variational lower bound of the log-likelihood function until it converges



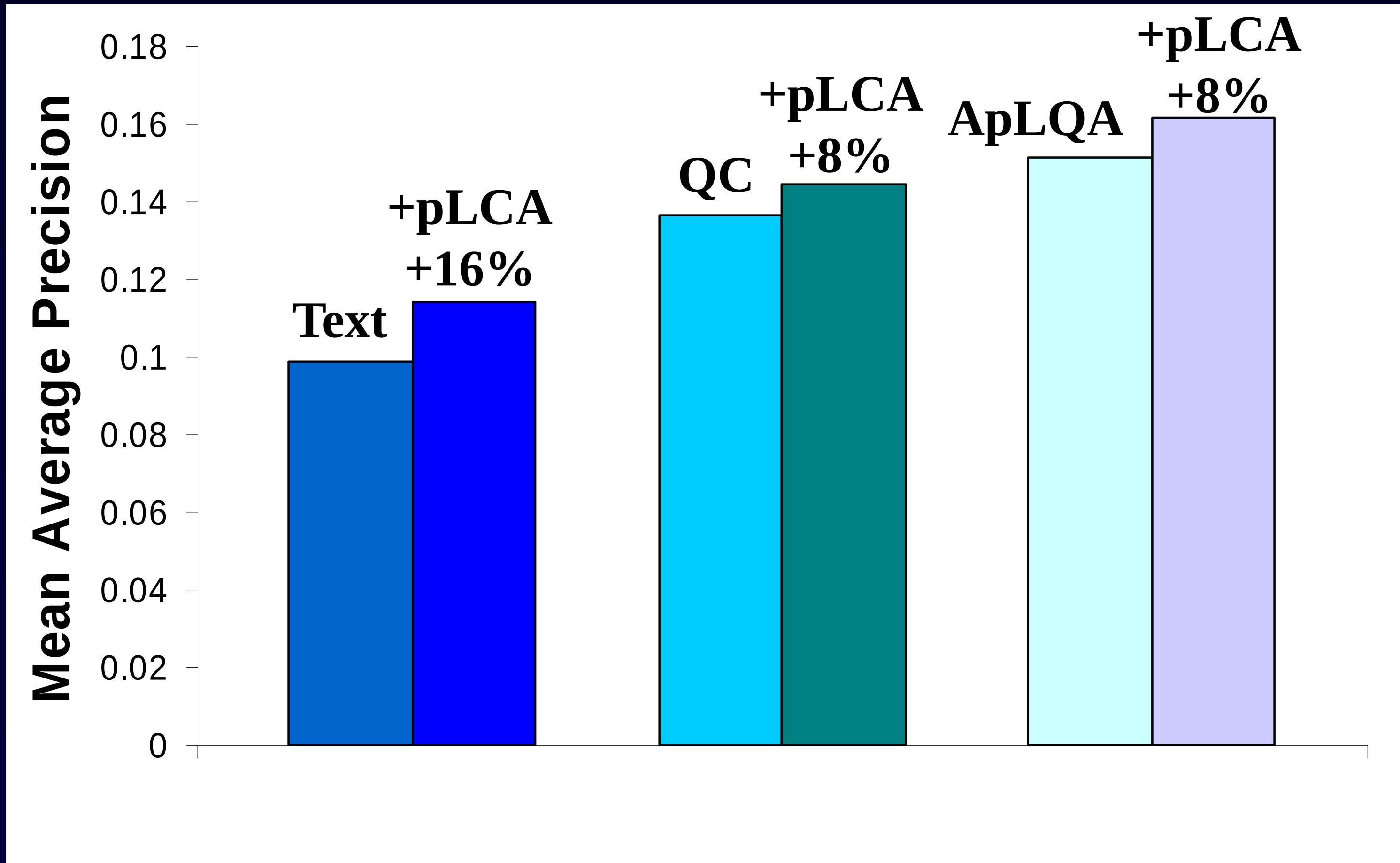
Additional Features found by pLCA in Video Retrieval

- Initial search results: adaptive pLQA
- pLCA finds useful features (concepts) for specific queries

Query	Additional features and weights found
Hu Jintao	Leader: 1.00, Crowd: 0.39, Airplane: 0.07
Tony Blair	Leader: 1.00, Commercial: -0.37,Crowd: 0.93
Condoleezza Rice	Person: 1.00, Crowd: 0.99
Helicopter	Airplane: 0.23, Sky: 1.00
Fire/flame	Crowd: 0.63, Car: 0.51, Building: 0.64, Urban: 0.93
Basketball	Crowd: 0.53, Commercial: -0.35
Map of Iraq	Maps: 1.00, Computer_Screen: 0.13

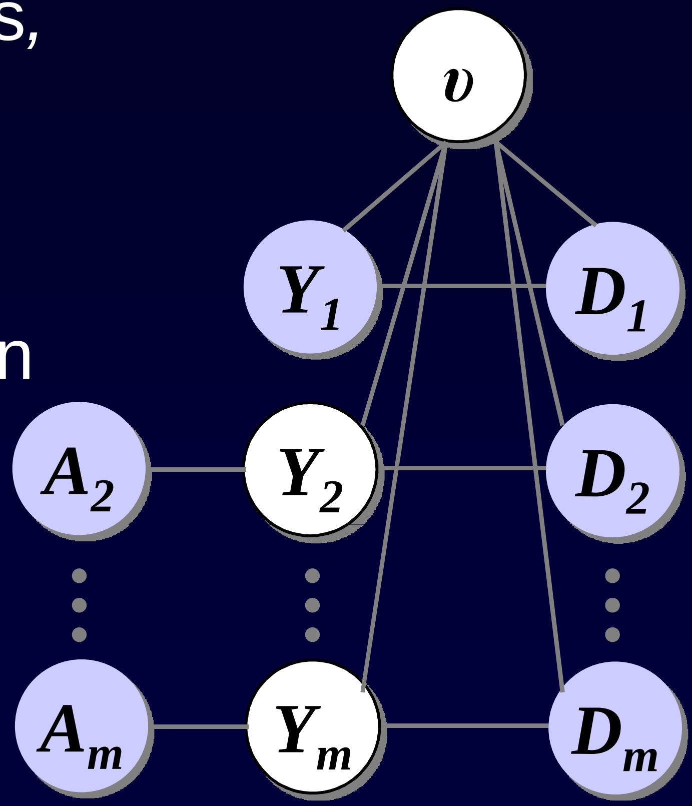
Video Retrieval Performance

- Retrieval performance of pLCA on TREC'03-'05
  - pLCA on 3 retrieval algorithms w. ~50 additional ranking features



pLCA with Relevance Feedback

- Users could provide their own relevance judgments in an interactive retrieval interface.
  - Given these relevance feedbacks, we can use a similar variational inference technique to update parameters except that some relevance variables $\mathbf{Y_i}$ are fixed in this case.



Undirected Model and Parameter Estimation

- Compute the posterior probability of document relevance ***Y*** given initial results A based on an

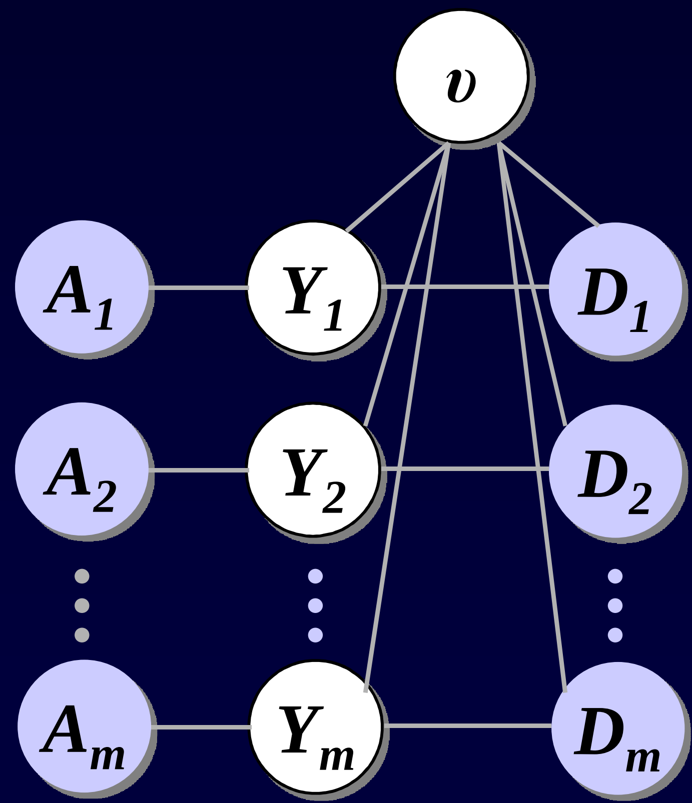
$$P(\vec{y}\,|\,\vec{a};\vec{D},Q)=\frac{1}{Z}\int_{v}\prod_{l}P(v_l\,|\,Q;v_l^0)\cdot\prod_{j=1}^{M_D}\exp\left(y_j a_j + y_j\sum_l v_l f_l\left(D_j,Q\right)\right)dv$$

- Variational inference, i.e., iterate the following steps until convergence and approximate $P(y_j|a_j)$ by q_{yj}
  - Maximize w.r.t. variational para. of ***Y***,

$$q_{yj}=\left[1+\exp\left(a_j+\sum_l q_{vl} f_l(D_j,Q)\right)\right]^{-1}$$

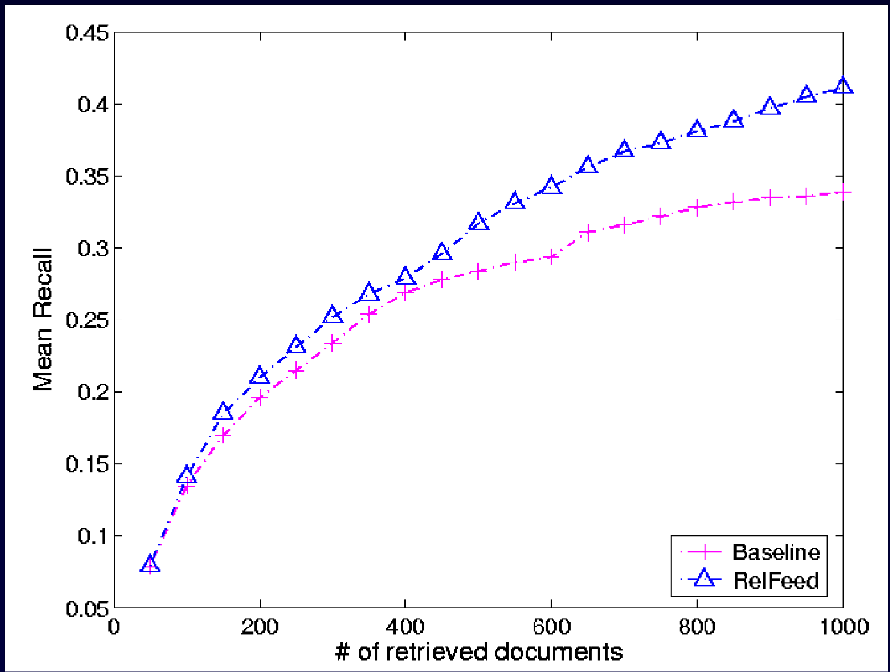
  - Maximize w.r.t. variational para. of ***u***

$$q_{vl}=v_l^0+\sum_j q_{yj} f_l(D_j,Q)$$

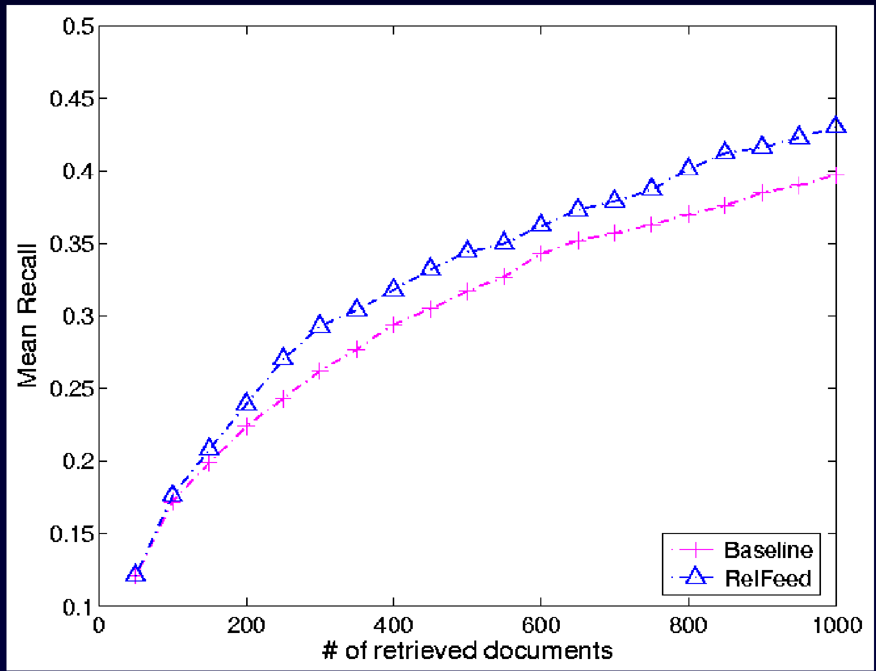


Results of Relevance Feedback

- Video retrieval performance w. relevance feedback
  - Performance in terms of average recall
  - Update combination parameters every 50 shots



Baseline: **text retrieval**

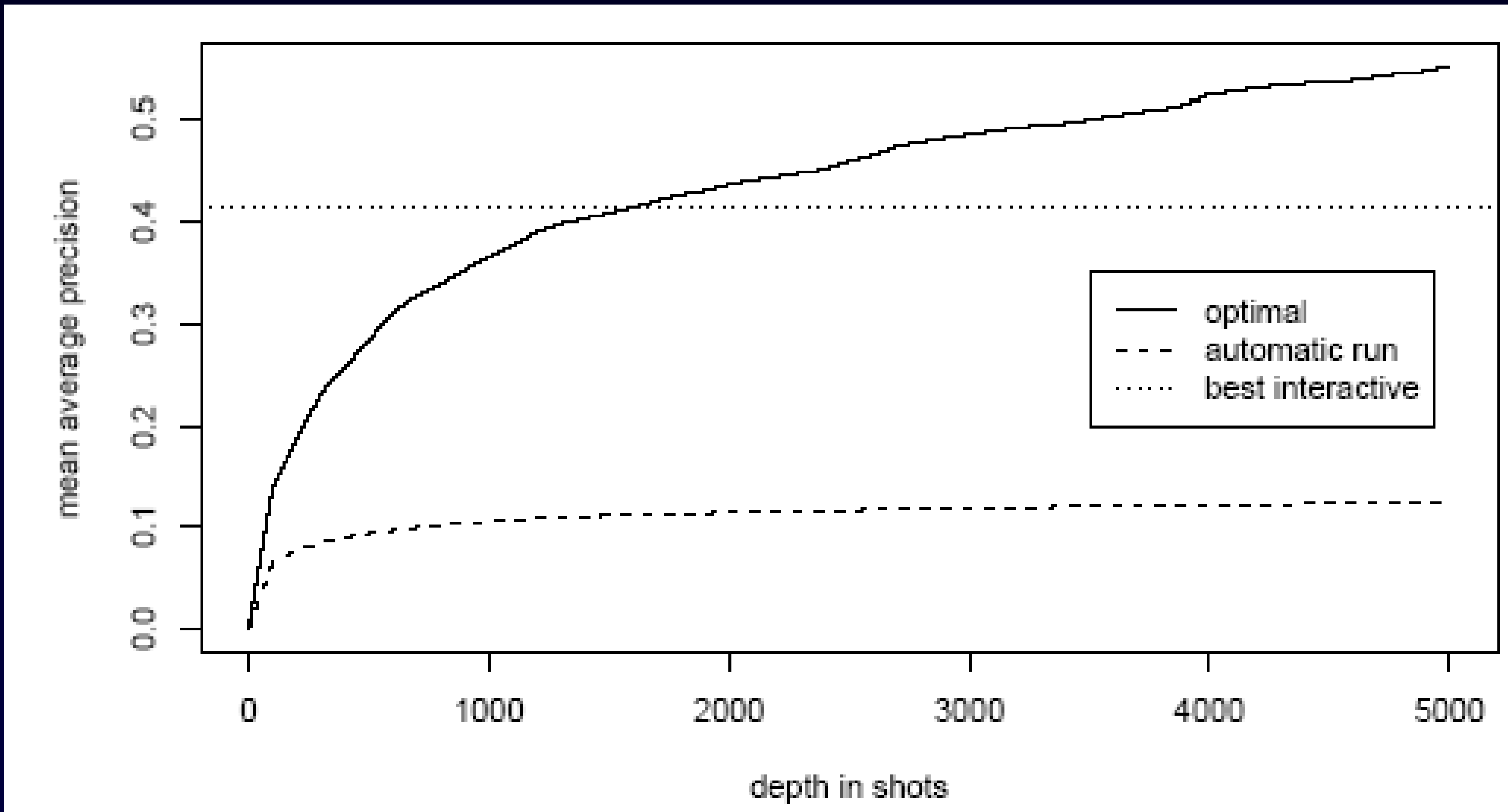


Baseline: **ApLQA**

Extreme Video Retrieval

- Maximize the synergy between humans and machines
- Make efforts efficient!

Observations about Automatic vs Interactive Search



Extreme Video Retrieval

- Automatic retrieval baseline for ranked shot order

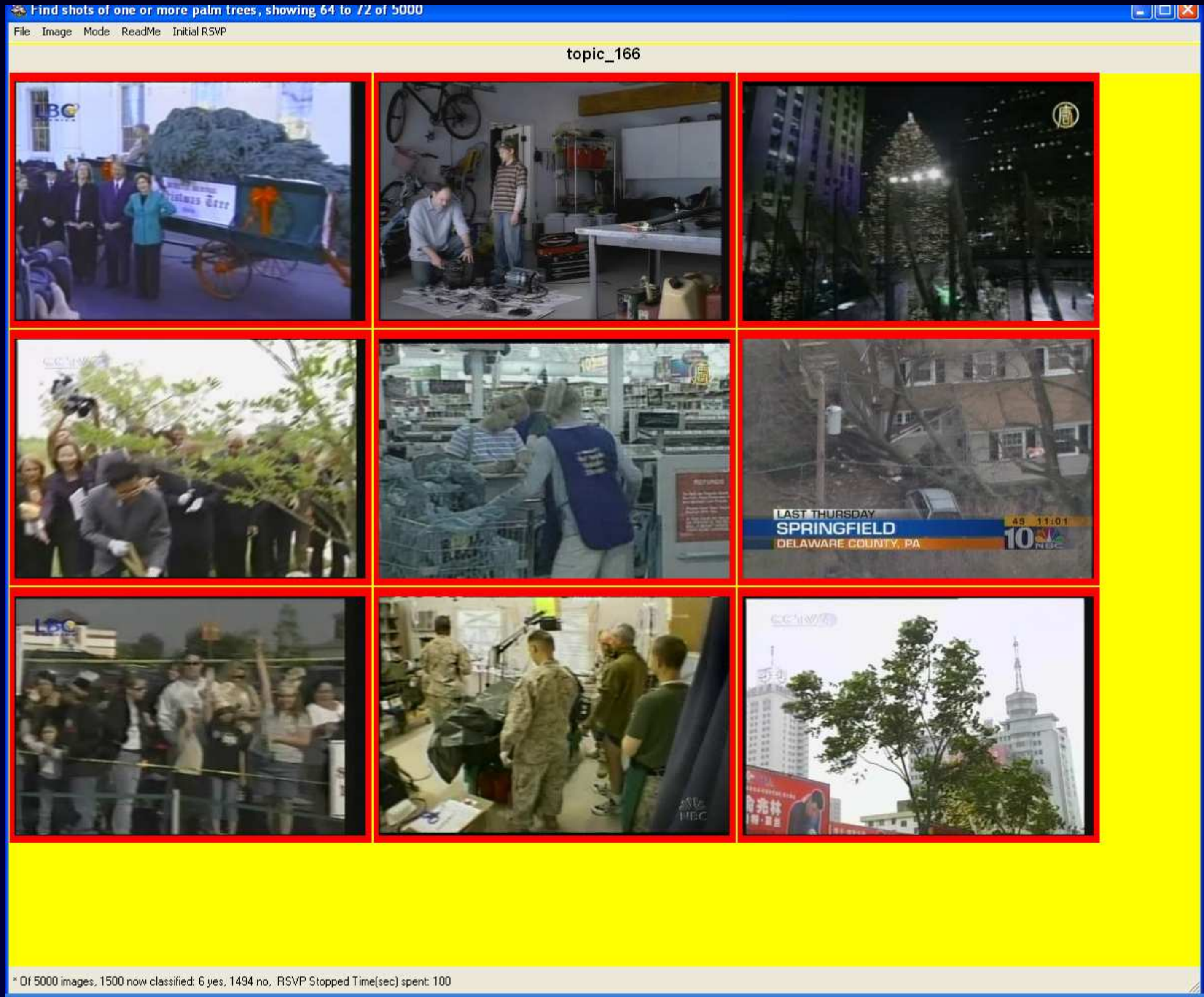
- Two methods of presentation:
  User-controlled or *System-controlled* time interval
  - System-controlled Presentation - Rapid Serial Visual Presentation (RSVP)
  - User-controlled Presentation – Manual Browsing with Resizing of Pages

System-controlled Presentation

- Rapid Serial Visual Presentation (RSVP)
  - Minimizes eye movements
    - All images in same location
  - Maximizes information transfer: System → Human
    - Up to 10 key images/second
    - 1 or 2 images per page
    - Presentation intervals are dynamically adjustable by the user
      - Slower initially (or when “breaks” are needed)
        - Many relevant images, user needs habituation
      - Faster after a few minutes (100 msec/page increments)
        - Few relevant images, accommodation
  - Click when relevant shot is seen
    - Mark previous page also as relevant

- A final verification step (~3 min) is necessary
  - Should be related to the number of relevant shots

Extreme QA with RSVP



3x3 display

1 page/second

Numpad chording to select shots

User-controlled presentations

- Manual Browsing with Resizing of Pages
  - Manually page through images
    - User decides to view next page
  - Vary the number of images on a page (2, 4, 9, 16)
  - Allow chording on the keypad to identify shots of interest
- Also tried clustering by story and without resizing of pages
  - Not as effective
- A very brief final verification step (1 min)

MBRP - Manual Browsing with Resizable Pages

Active Learning for Extreme Video Retrieval

Eventually, we envision the computer will observe the user and LEARN!

The system can learn:

- What object and image characteristics are relevant
- What text characteristics (words) are relevant to the query
- What combination weights should be used to combine them

Based on shots that have just been marked as relevant

- As learning improves, the human has to do less and less work

We exploit the human's ability to quickly mark relevant shots and the computer's ability to learn from given examples

Questions?

A Large Scale Concept Ontology for Multimedia Understanding

Joint work with Milind Naphade[2], John R. Smith[2], Shih-Fu Chang[3], Rong Yan[2], Wei-Hao Lin[1] and Michael Christel[1]
Carnegie Mellon University[1], IBM Research[2], Columbia University[3]

Research in Semantic Concepts

"High-Level" or Semantic Features

Faces, cars, outdoors, people, buildings, vegetation, rivers, animals, …

- Can be detected automatically
  from color, texture, shape, style
- Intense competition for good classifiers
  (TRECVID)
- Useful for retrieval and summarization

LSCOM Goal

Problem:

- Users and analysts require richly annotated video content for search and retrieval
- We don't know how to translate video content into words
- Manual annotation is prohibitively expensive and slow

Solution:

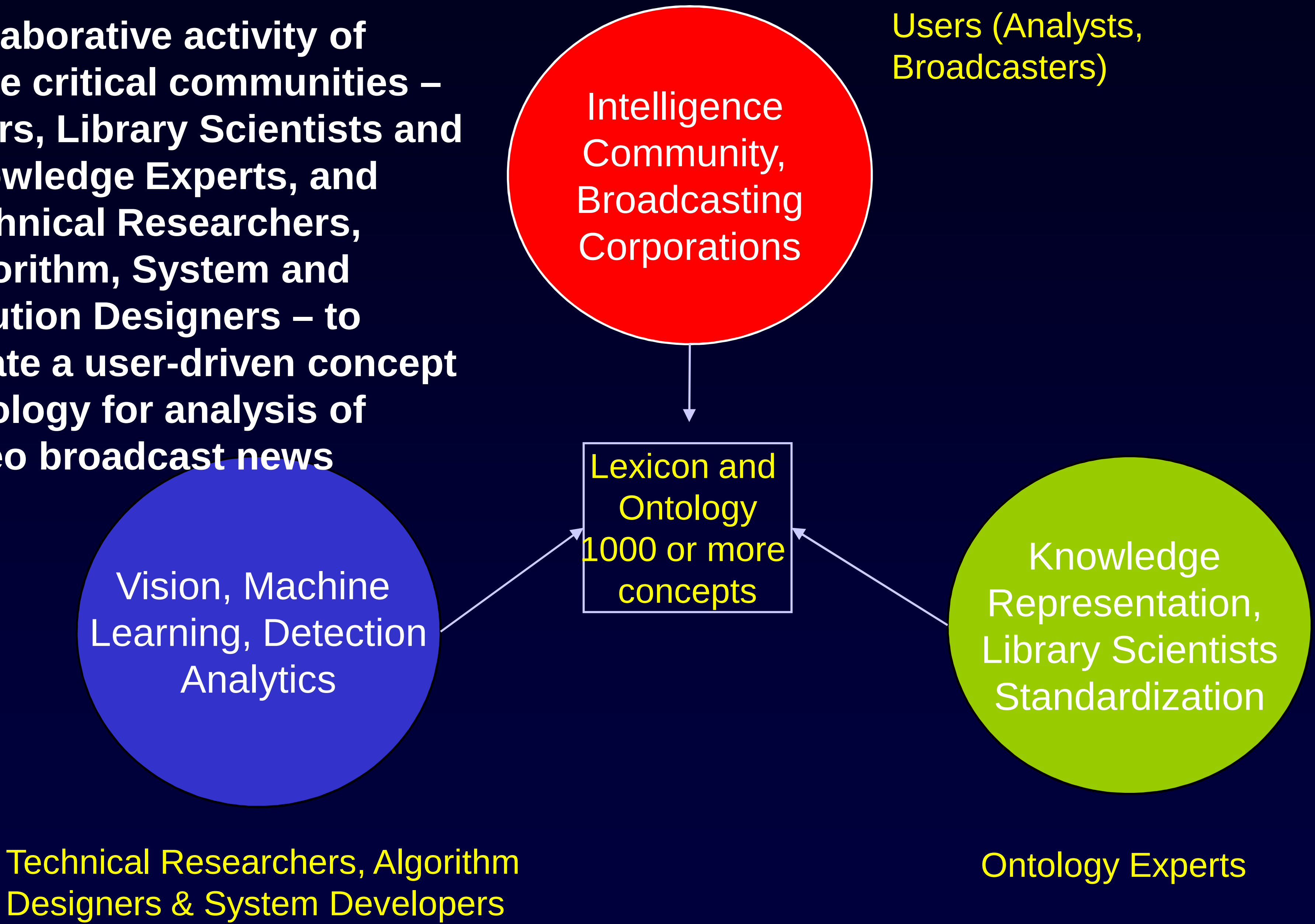
- Find a restricted (controlled) concept vocabulary which can be used to (automatically) describe broadcast news video content
  - Start with 1000 concepts grouped into a taxonomy/ontology
  - Evaluate if these concepts are useful for retrieval
  - Test if they can be automatically detected
  - Iterate

Impact:

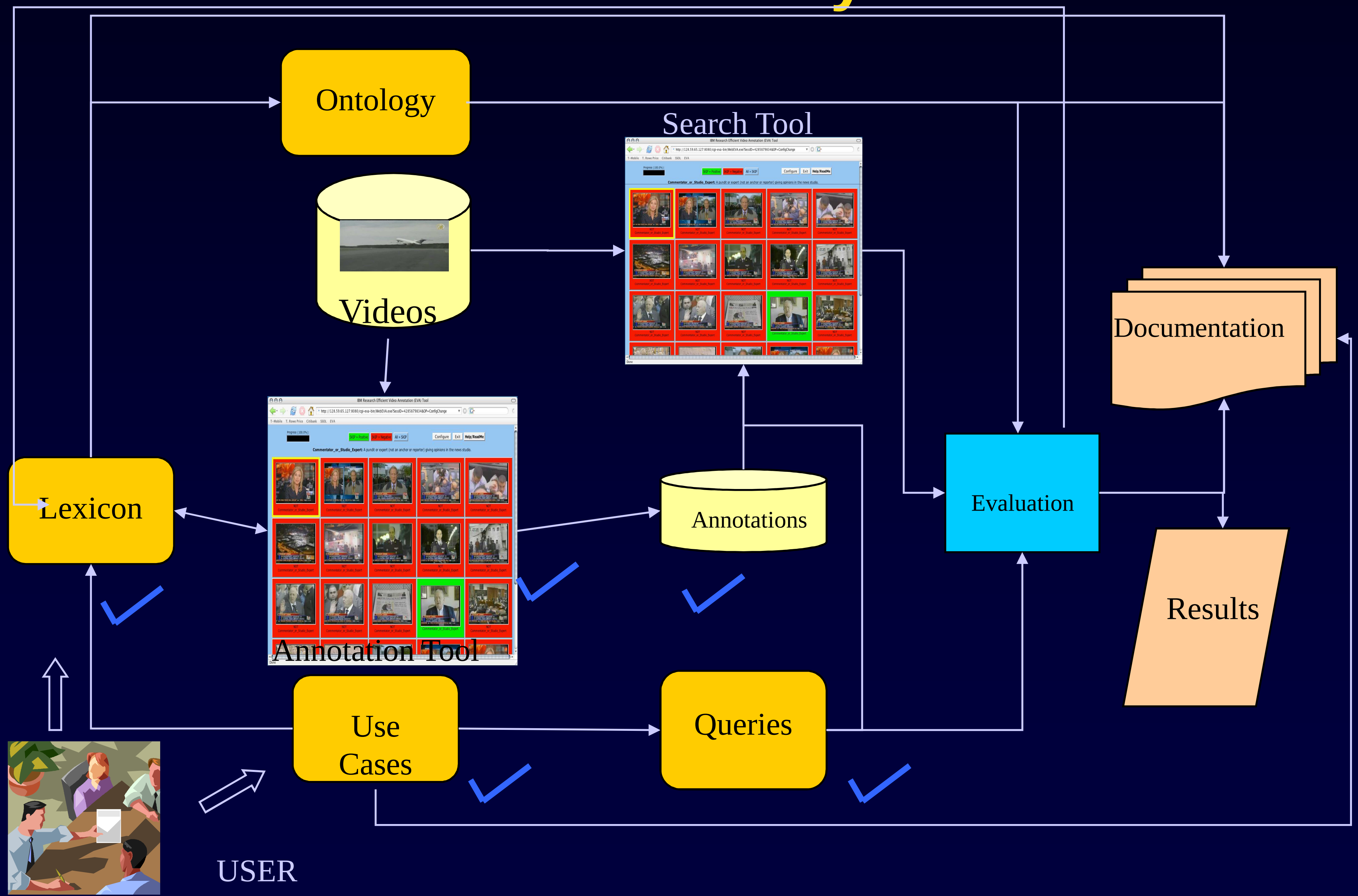
- Allow useful classification of multilingual broadcast video
- Provide an extensible framework and procedures for video analysis, beyond the 1000 concepts

Central Idea

- **Collaborative activity of three critical communities – Users, Library Scientists and Knowledge Experts, and Technical Researchers, Algorithm, System and Solution Designers – to create a user-driven concept ontology for analysis of video broadcast news**



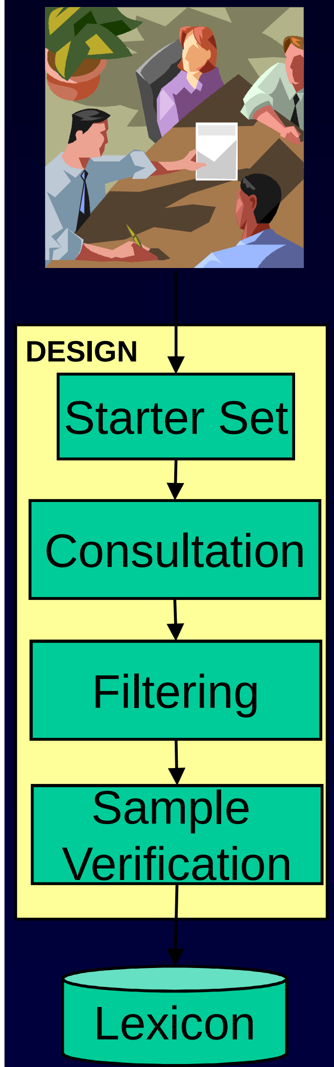
LSCOM Creation Summary



Multimedia Controlled Vocabularies

- MPEG-7 Multimedia Description Schemes standardizes classification schemes comprising 754 controlled terms
- TV-Anytime standardizes 954 terms for broadcast TV content
- Escort 2.4 defines 115 terms
- Thesaurus of Graphical Materials (TGM-I) defines index terms for cataloging graphical (image and video) content
- SMPTE Metadata Registry from Society of Motion Picture and Television Engineers has a metadata dictionary defining a registry of metadata element descriptions for association with essence or other metadata. (Full explanation in SMPTE 335M)
- IPTC Core Metadata, NewsML, SportsML and ProgramGuide ML from International Press Telecommunications Council. The versatile News Markup Language for global news exchange. NewsML 1 is designed to provide a media-independent structural framework for multimedia news.

LSCOM Lexicon Design



More than 30 Media Analytics Experts, 10 User Community Experts and 6 Knowledge Representation Experts met twice

More than 10,000 concepts
TGM, Time Life, TV Anytime, Comstock, WordNet

More than 600 concepts from media companies, intelligence analysts

Filtered down to 834 concepts (so far)
based on Usability, Feasibility and Observability

Manual annotation over sample video corpus led to
Annotation of 449 unique concepts (so far) based on availability of
concept in corpus and inter-annotator agreement

Sample Lexicon

	A	B	C
20	1	Activities	Talking
21	1	Activities	Throwing
22	1	Activities	Walking
23	1	Animals	Bird
24	1	Animals	Camel
25	1	Animals	Cat
26	1	Animals	Cow
27	1	Animals	Dog
28	1	Animals	Donkey
29	1	Animals	Elephant
30	1	Animals	Fish
31	1	Animals	Horse
32	1	boat/ship	Aircraft Carrier
33	1	boat/ship	Barge
34	1	boat/ship	Battleship
35	1	boat/ship	Canoe
36	1	boat/ship	Cigar boat
37	1	boat/ship	Cruise liner
38	1	boat/ship	Cutter
39	1	boat/ship	Freighter
40	1	boat/ship	Frigate
41	1	boat/ship	Houseboat
42	1	boat/ship	Landing Craft
43	1	boat/ship	Raft
44	1	boat/ship	Rowboat
45	1	boat/ship	Sailboat
46	1	boat/ship	Ship
47	1	boat/ship	Submarine
48	1	boat/ship	Tanker
49	1	boat/ship	Tugboat
50	1	Building	Airport
51	1	Building	Barn
52	1	Building	Barracks
53	1	Building	Castle
54	1	Building	College
55	1	Building	Conference/conventi

Design Criteria

- Utility
- Feasibility
- Observability

Additional Annotation Criteria

- Availability in the Corpus
- Inter-annotator Agreement

Annotated Data

- TRECVID 2005 development set
  - 80 hours of broadcast news
    - 11 programs
    - 6 channels
    - 3 languages
  - 61901 shots
  - Each concept (449, total) labeled over each shot.

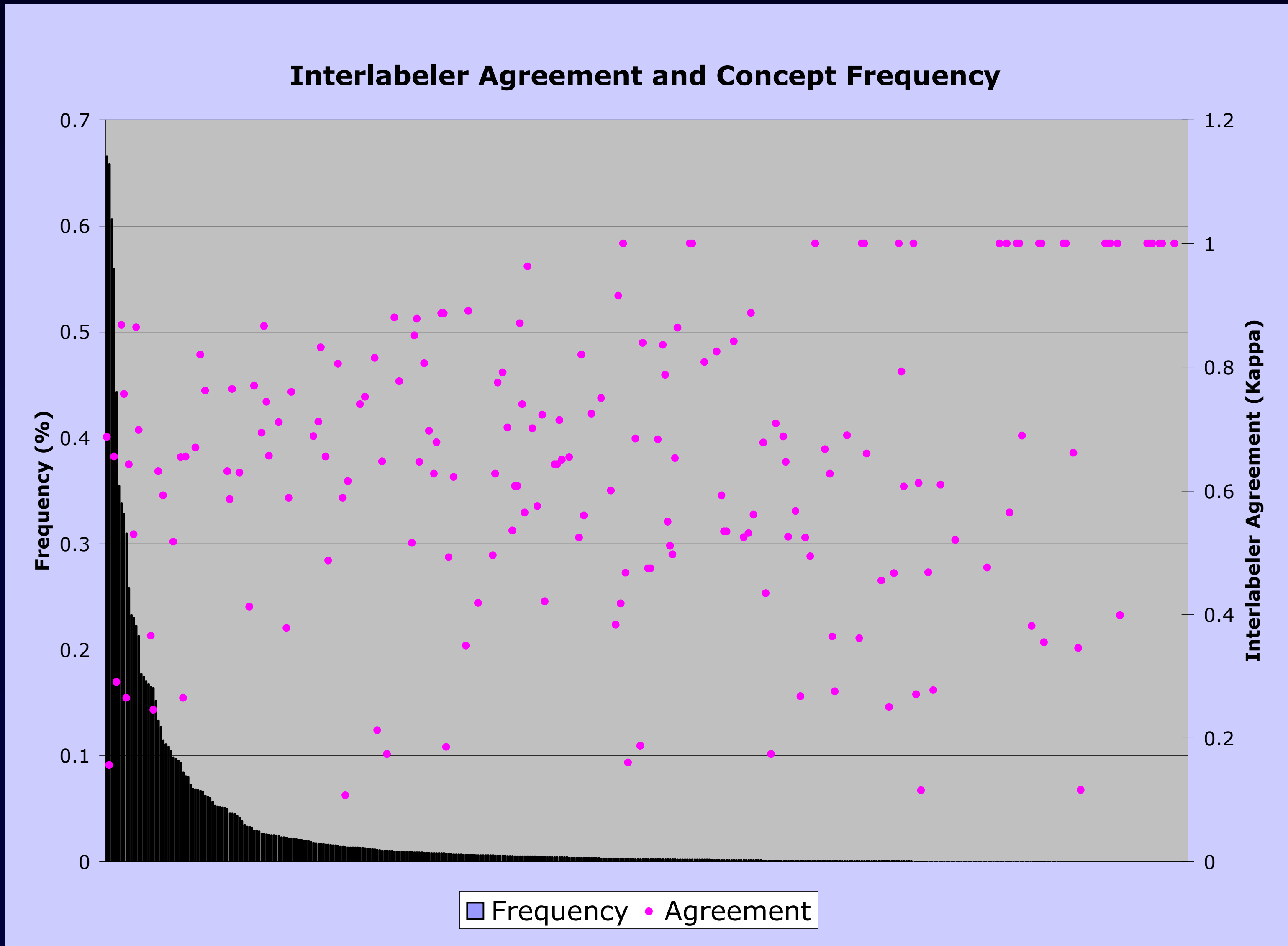
Kappa Coefficient for Quality Assessment

- Compare level of agreement between annotators
- Avoid bias towards less frequent concepts

- P(A) = Proportion of agreement
- P(E) = Proportion of agreement by chance

$$\kappa = \frac{P(A) - P(E)}{1 - P(E)}$$

Annotation Quality



Connecting Concept to an Ontology

- Cyc – more than 20 years building an ontology of everything
- Specialized and more useful for specific domains

Mapping the LSCOM Taxonomy into CYC

HL Formula :

Mt : LSCOMObjectAndSituationOntologyMappingMt
(synonymousExternalConcept Place LSCOMObjectAndSituationOntology "locations")

Mt : LSCOMObjectAndSituationOntologyMappingMt
(synonymousExternalConcept Island LSCOMObjectAndSituationOntology "island")

Then map the graphical structure.

Mt : (ContextOfPCWFn LSCOMObjectAndSituationOntology)
(broaderThan Island Place)

Cyc can use the mappings to reason about the LSCOM taxonomy

EL Query :

```
(and
  (synonymousExternalConcept ?X LSCOMObjectAndSituationOntology ?S)
  (sourceNodeInSystem ?X LSCOMObjectAndSituationOntology))
```

Status : Suspended, Exhaust

Hide **Answers** *(11 new)* **Actions :** [Save Answers] [Graph Answers] [

Answer	?X	?S
*[Explain #10]	HumanScaleObject	"Objects"
*[Explain #9]	Person	"People"
*[Explain #8]	Roadway	"Road"
*[Explain #7]	AnimateActivity	"activities"
*[Explain #6]	Barrier-ConstructionArtifact	"Barrier"
*[Explain #5]	Event	"Events"
*[Explain #4]	AirTransportationDevice	"flying-object"
*[Explain #3]	Place	"locations"
*[Explain #2]	PlayingOf-SportsEvent	"sports"
*[Explain #1]	Building	"Building"
*[Explain #0]	SewerSystem	"Sewage infrastructure"

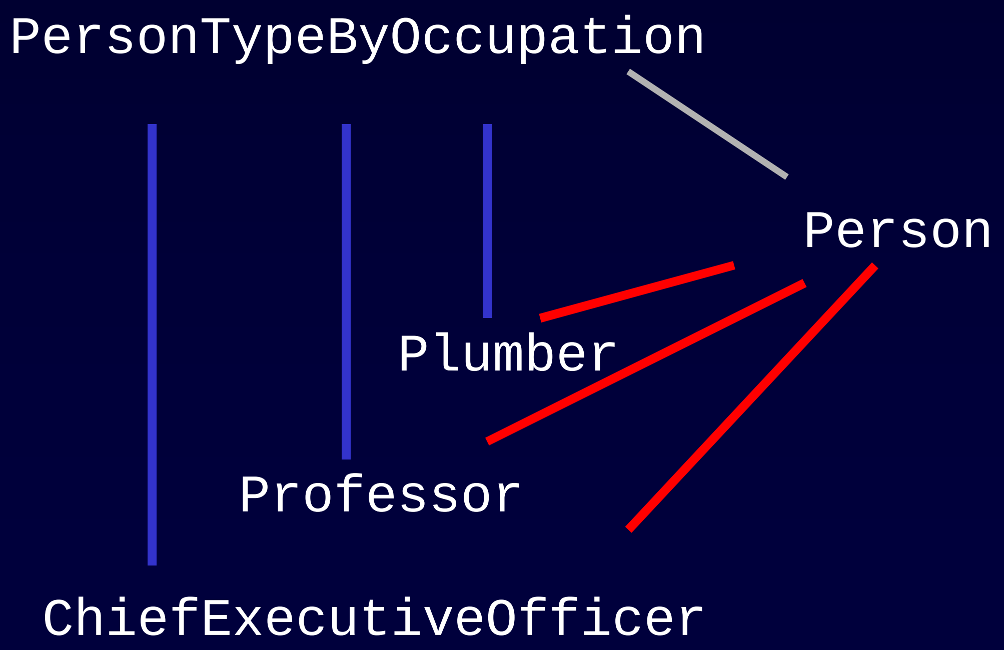
Results of a query for the top (source) nodes in the LSCOM taxonomy, as authored by workshop participants.

Cyc can Suggest Extensions/Improvements

Use of Faceting Collections

- Higher-order collections that facet a class into sub-classes whose instances share a property

- Example:



Final Statistics of the Ontology

Size: 784 → 2638

3.36x increase

164% over 1000-node goal

Depth: 763/97.3% → 1305/49.5%

47.8% reduction

Distribution

LSCOM ontology will be included in all ResearchCyc and OpenCyc releases.

OpenCyc: All terms, plus all terms in Cyc ontology, plus isa and genls links among nodes

ResearchCyc: All of Cyc, including rules, lexicon, and inference engine.

You can also get the LSCOM ontology from me or from Shih-Fu Chang at Columbia University

What are the Limits of Semantic Concepts

- How much can semantic concepts improve video search? What is the upper bound?
- How many do we need for good retrieval?
- Are there enough concepts out there?
- Can we find them?

Concept Sets with Annotations

- LSCOM Lite
  - 39 concepts
- Media Mill
  - 101 concepts available
  - 75 used that have overlap with LSCOM
- LSCOM
  - 300 concepts
  - Minimal frequency cutoff

Best Case Scenario

- Perfect concept detection
  - Oracle Detection (OD)
- Perfect concept combination
  - Oracle Combination (OC)

- How well can you find shots relevant to the query

Computing the Optimal Weights

The locally linear bound (LLB) is the best weighted combination of concepts

$$LLB = \max_{\lambda_i} AP\left(\sum_{i=1}^{k} \lambda_i f_i(D, Q)\right)$$

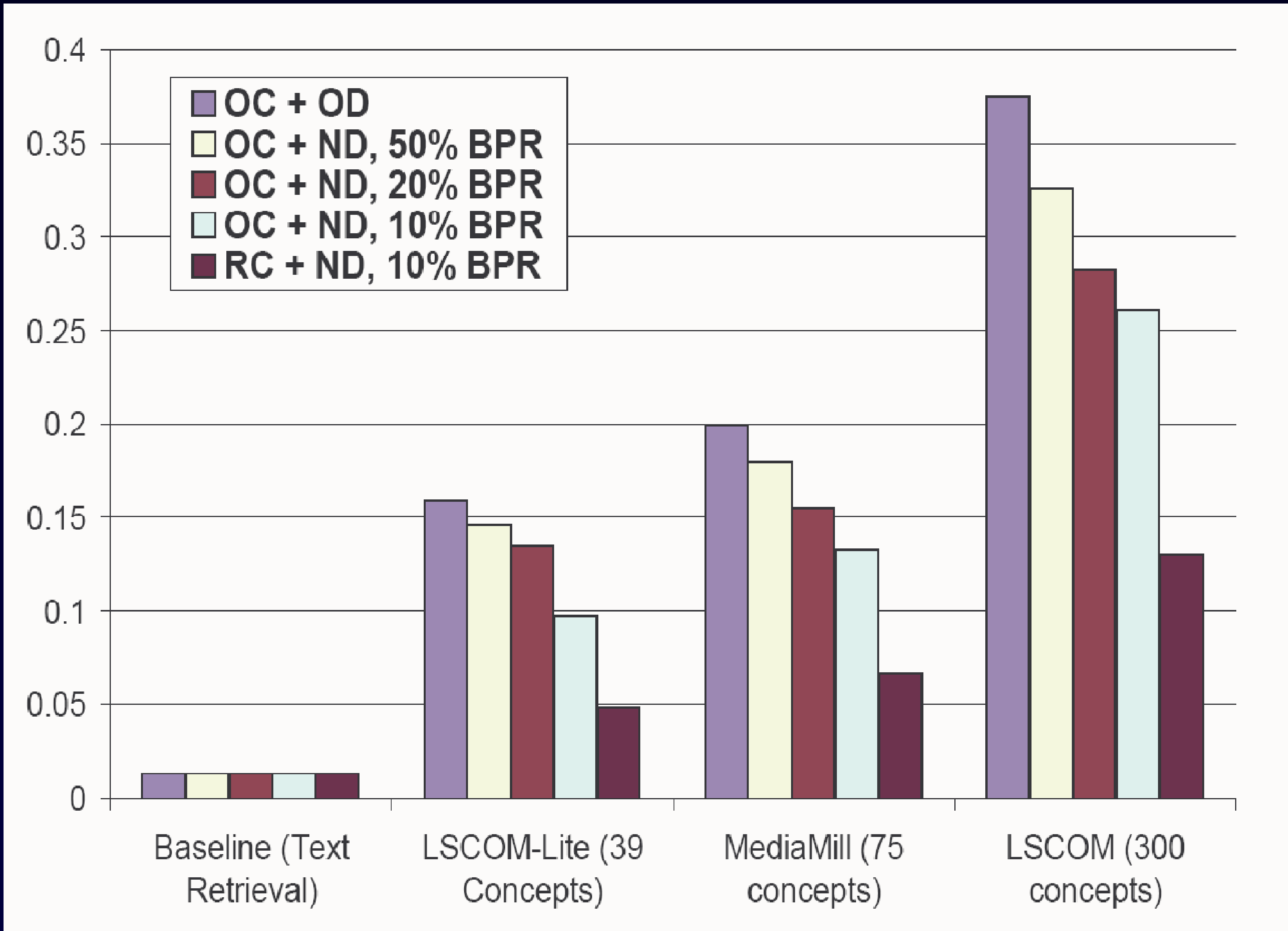
More Realistic Assumptions

- Noisy concept detection (ND)
  - MAP at 0.5, 0.2, 0.1
  - MAP of 0.2 is current TRECVID 2006 semantic feature detection results

- Approximate MAP through breakeven precision/recall (BPR) point at 50%, 20%, 10%

- Realistic combination (RC)
  - 50% effectiveness of Oracle combination

Oracle Experiments



Extrapolation into the Future

- How many concepts do we need?
- Exponential function to fit curve
  - Only 3 points

It is far better to foresee even without certainty than not to foresee at all.
-- Henri Poincaré, *The Foundations of Science*

Extrapolation Assumptions

Improvement is harder to achieve as you add more concepts

Proportional to difference between current MAP and the upper limit 1

At higher MAP, we get less benefit from a new concept

$$\frac{dm}{dx} \propto (1 - m)$$

where m is the value of MAP,

x is the number of concepts

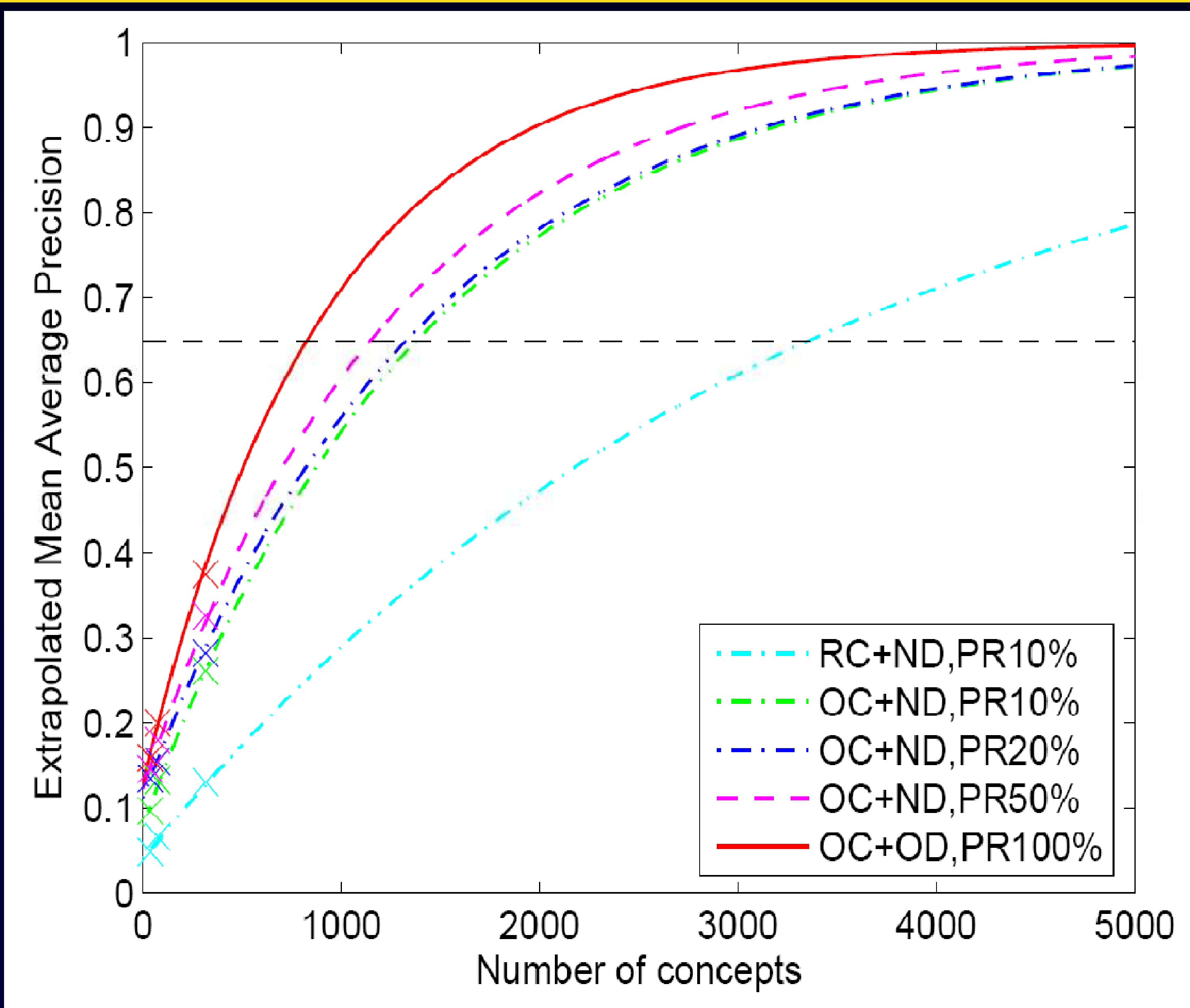
Bridging the Semantic Gap
Experiments Using High-Level Concepts

Extrapolating from video retrieval with up to 300 concepts

Assumptions:

- Oracle Concept Combination and Oracle Detection (**OC+OD 100%**)
  - Perfect concept detection and perfect use of concepts for retrieval
- Oracle Combination and Noisy Detection w. X% accuracy (**OC+ND X%**)
  - Concept detection with X% equal precision/recall and perfect use of concept combination for retrieval
- Realistic Combination and Noisy Detection w. 20% accuracy (**RC+ND 20%**)
  - Concept detection with 20% equal precision/recall as achieved in practice, with realistic use of concept combination for retrieval, as obtained in practice

Extrapolation Results



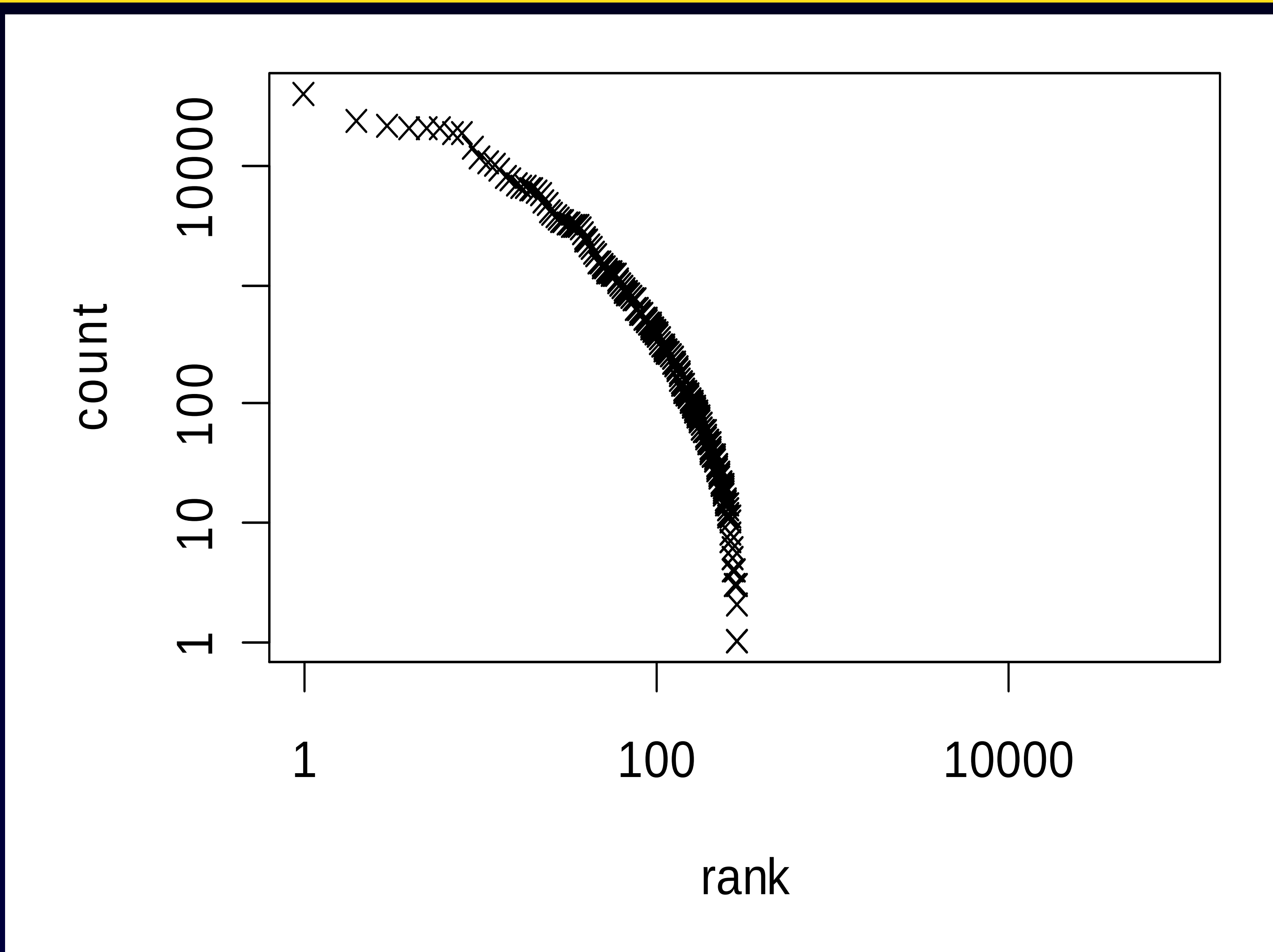
Is this achievable?

3000+ concepts promise effective video retrieval

- Comparable to current text results

But, are there that *many* concepts in a video archive?

Counts and ranks of visual concepts basically follow Zipf’s Law



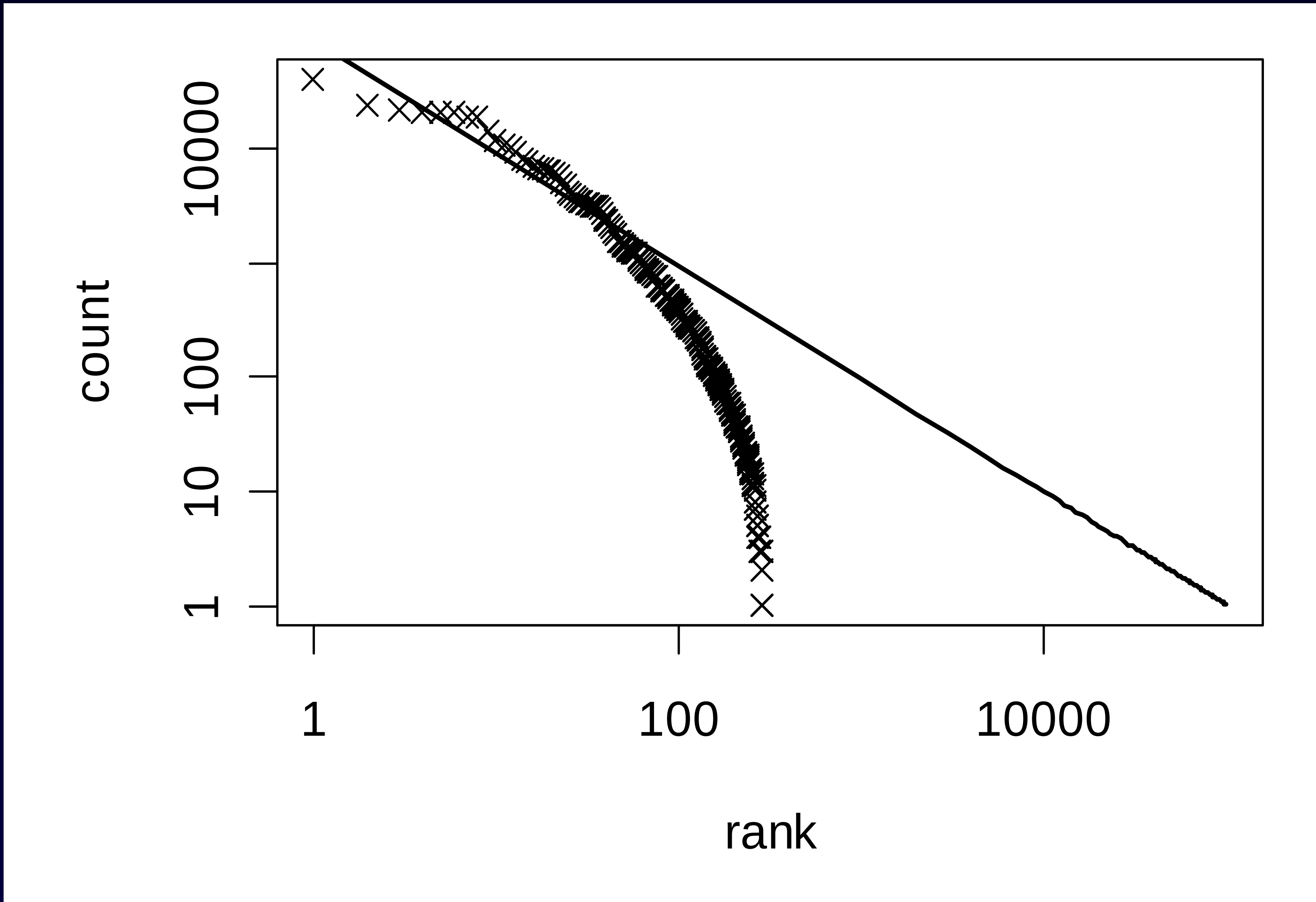
Zipf’s law states the product of the count and rank is constant:

count(i) * rank(i) = c

Zipf’s Law
Estimate the Number of Concepts

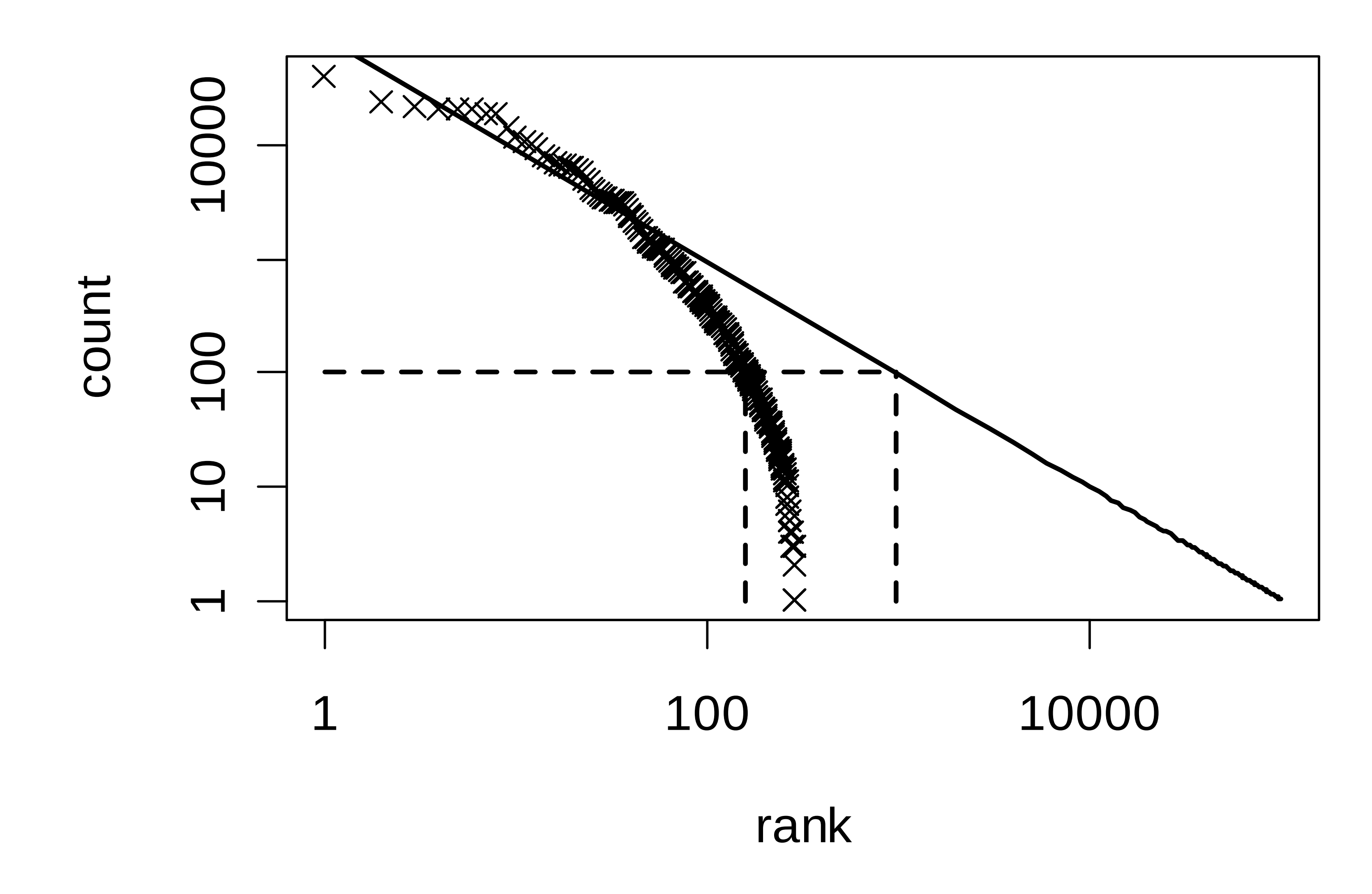
- Linear relationship holds for the top concepts
- Low-frequency concepts deviate from Zipf’s Law
  - Limited annotation resources and utility
- Fit a linear curve using weighted linear regression, and exponential decay weights to overcome the concept selection bias of LSCOM
- Reconstruct the linear Zipf curve
  - As if one could annotate all concepts from the video collection

There are 100K unique visual concepts in the TRECVID'05 video collection



TRECVid'05 has enough frequent concepts

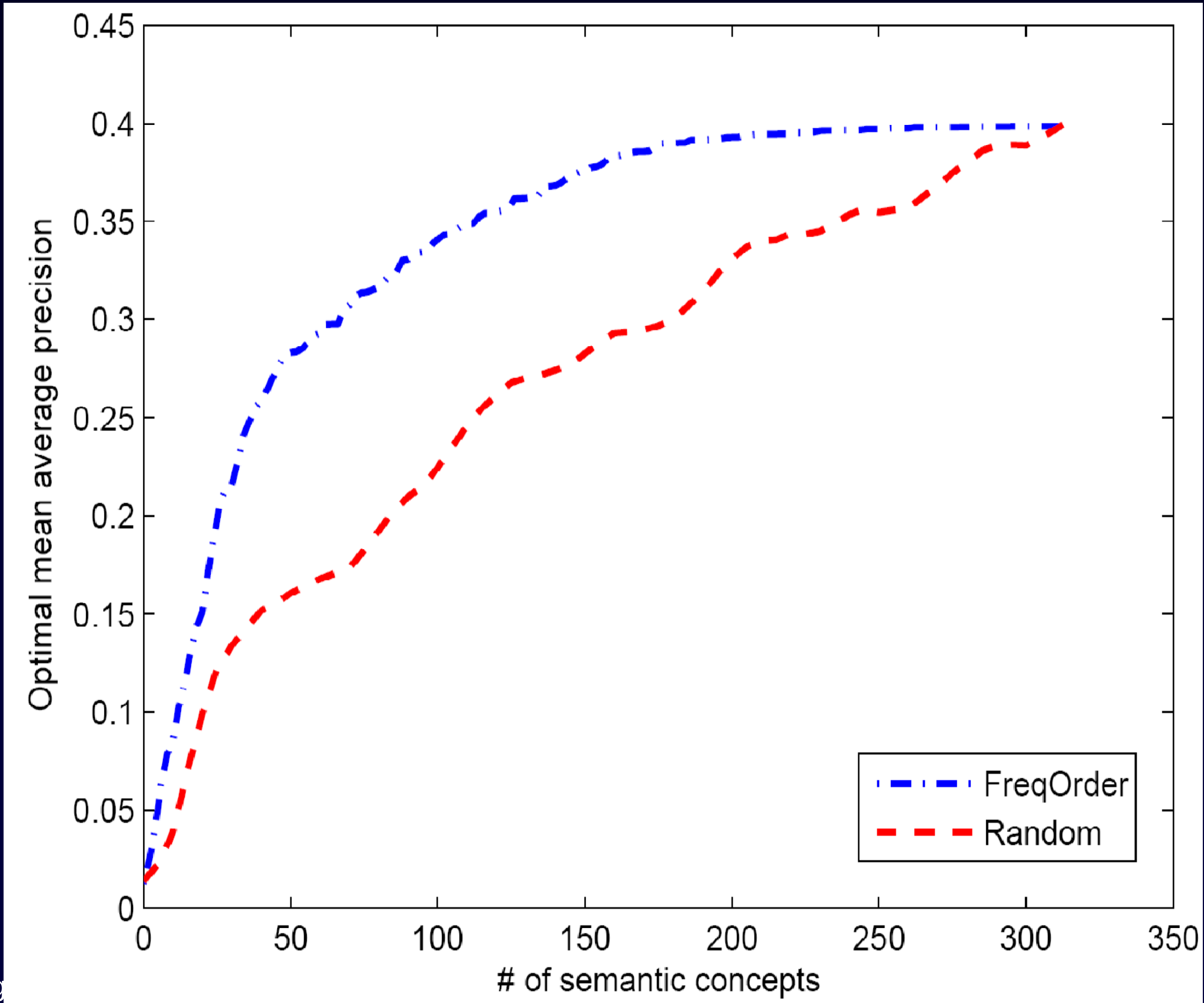
- Correlate to higher detection accuracy and usefulness



Concept Frequencies

			Frequency		
Concepts	Size	# Used	> 1000	>500	> 100
LSCOM	0.3K	0.3K	62	89	159
LSCOM-Zipf	100K	3K	97	192	945

Random Concept Selection vs Frequency Ordered (OC+OD)

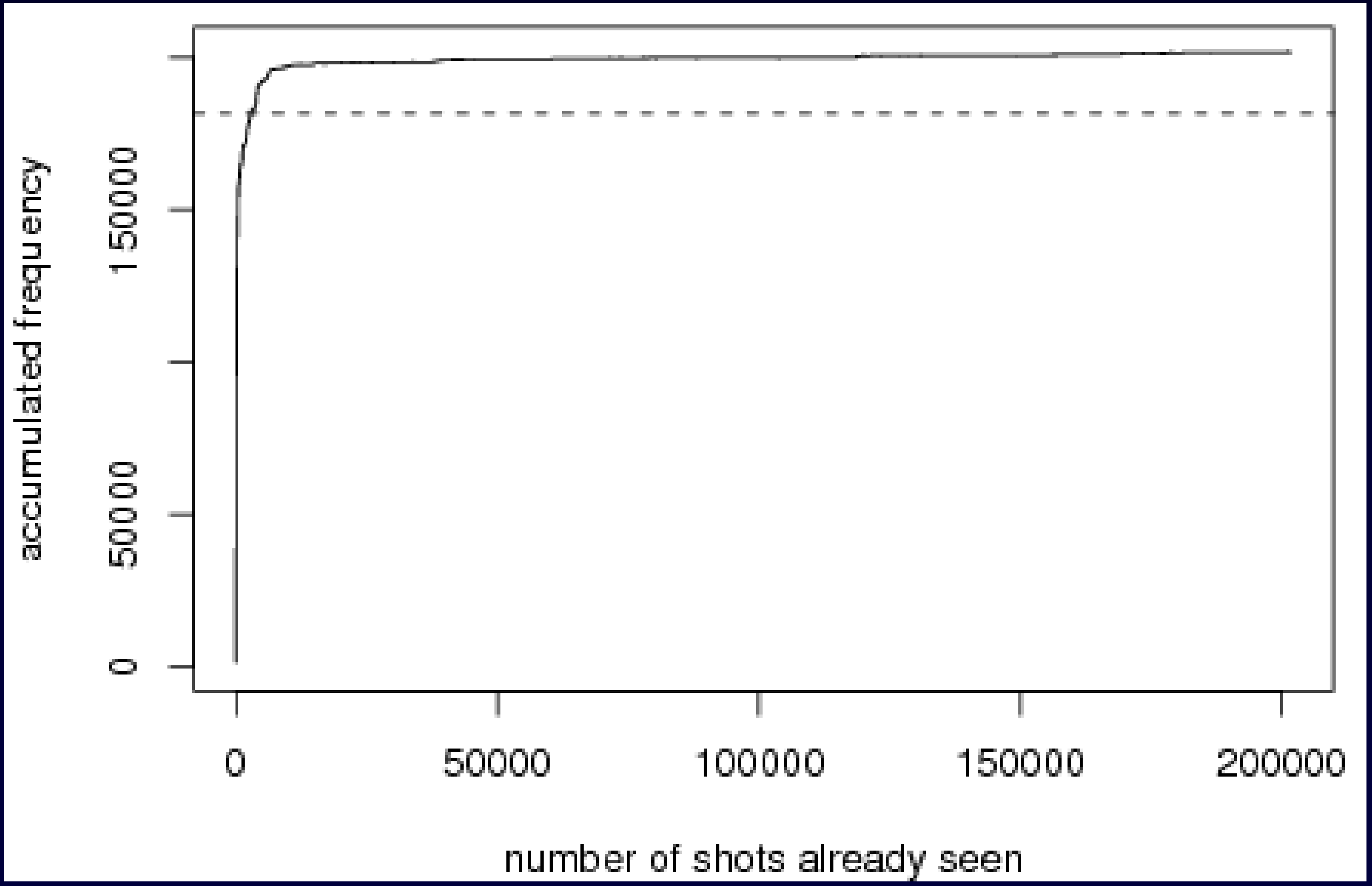


Finding These Concepts in the Archive

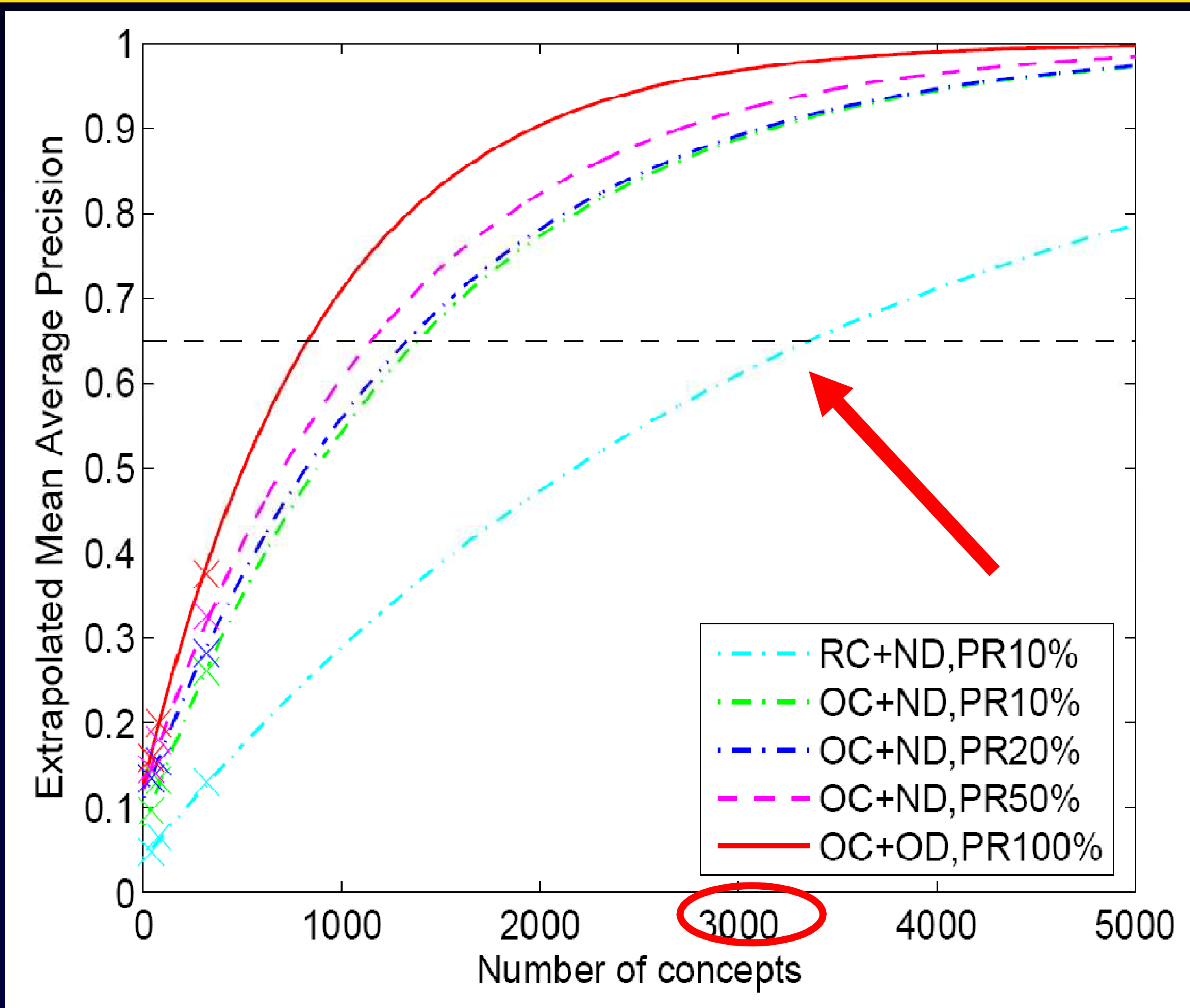
- How much video do annotators have to watch before enough concepts have been identified?

Common Concepts are Identified Quickly

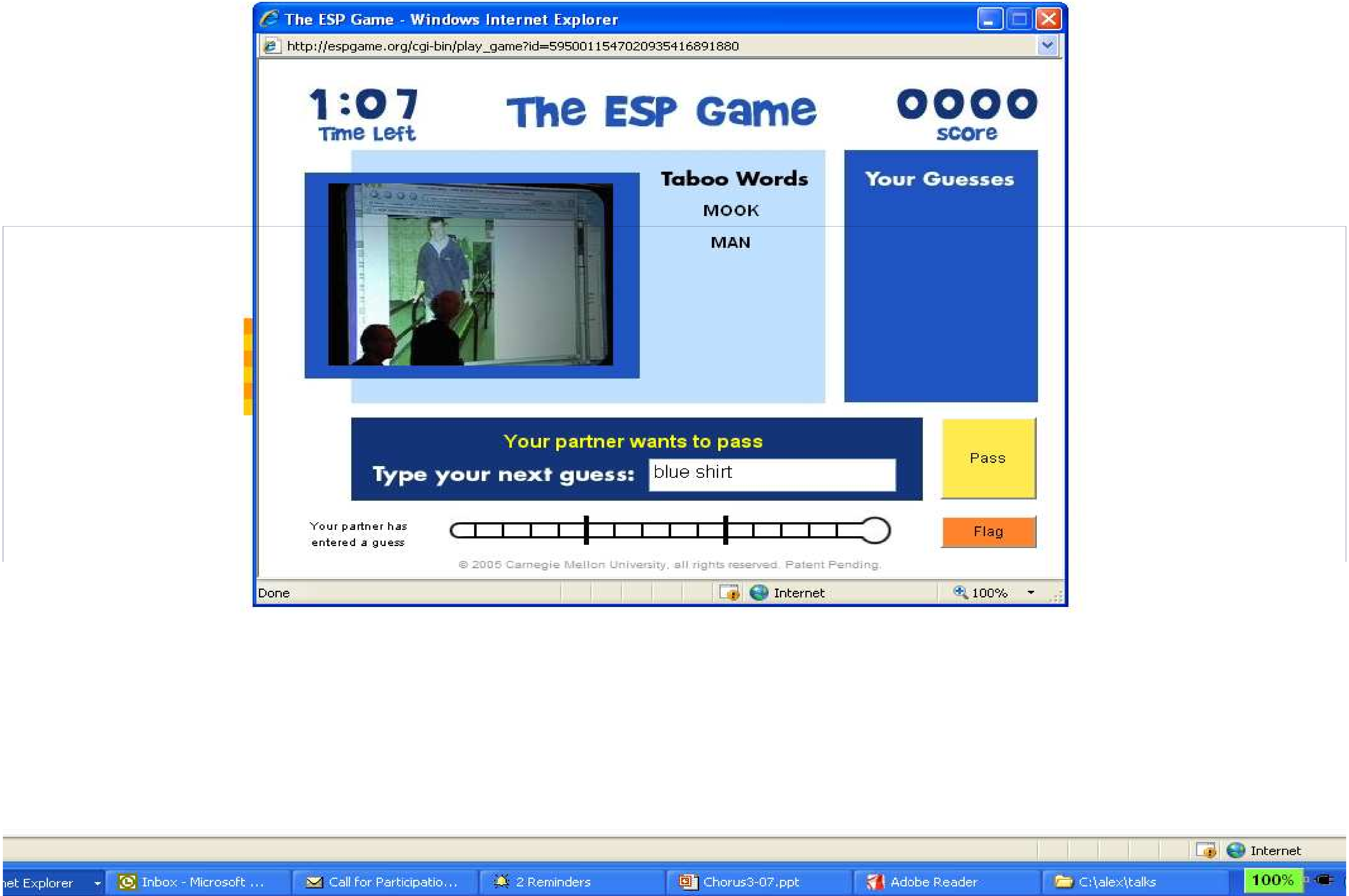
Watching 1.2% of the archive will reveal a set of concepts that account for 90% of concepts occurrences



The Value of 3000+ Concepts

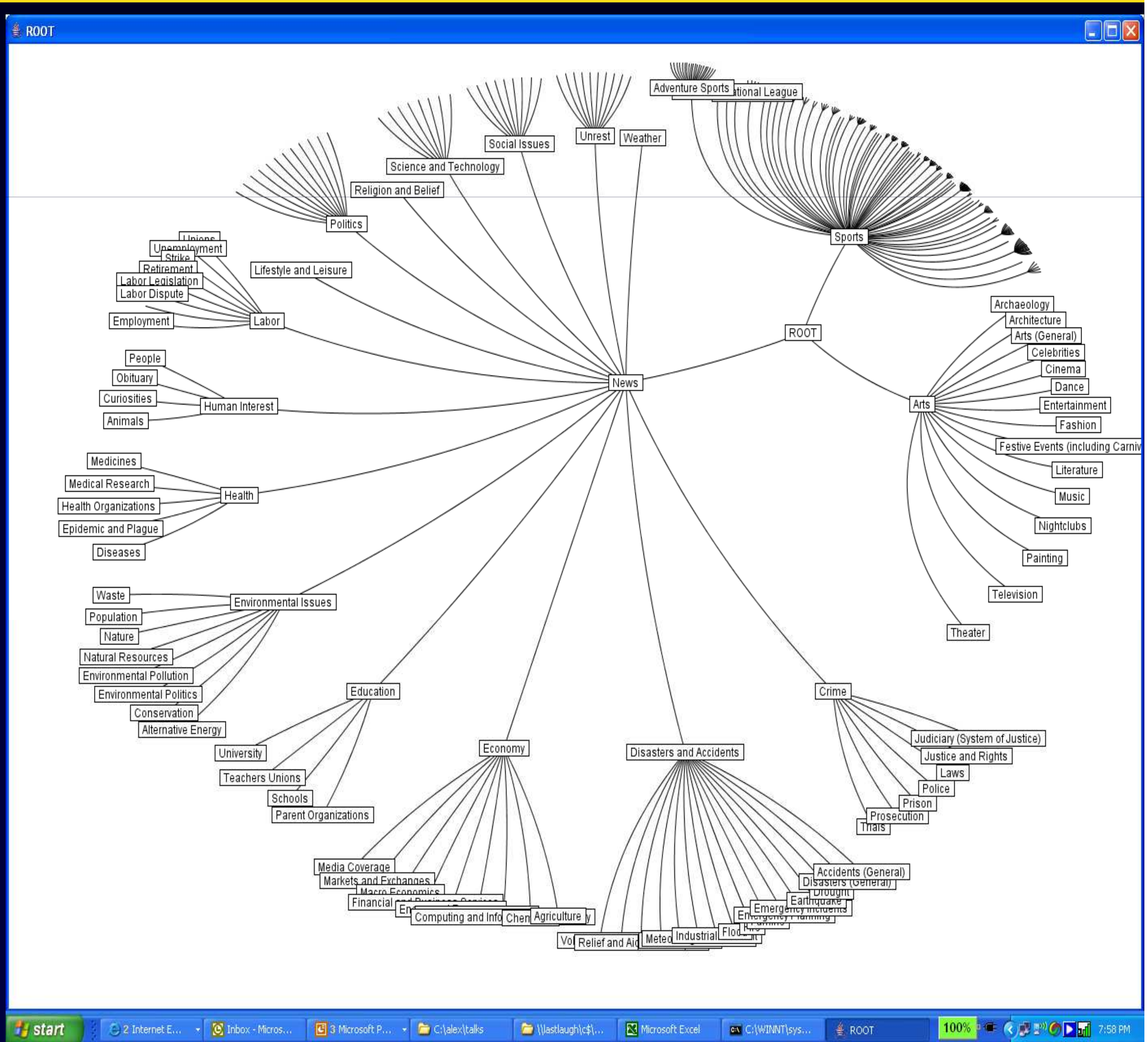


Human Computer Interaction


The ESP Game - Windows Internet Explorer
http://espgame.org/cgi-bin/play_game?id=5950011547020935416891880
1:07
Time Left
The ESP Game
0000
score
Taboo Words
MOOK
MAN
Your Guesses
Your partner wants to pass
Type your next guess:
blue shirt
Pass
Your partner has entered a guess
Flag
© 2005 Carnegie Mellon University, all rights reserved. Patent Pending.
Done
Internet
100%
Internet
Inbox - Microsoft ...
Call for Participatio...
2 Reminders
Chorus3-07.ppt
Adobe Reader
C:\alex\talks
100%

Ontologies Won't Solve All Problems

- Hand-constructed
  - As good as person building it
- Brittle with missing and extraneous information
- Don't match the data in the archive
- Reasoning can't deal with the high error rates from AV analysis



Conclusions

We are speculating, under “reasonable” assumptions:

- A few thousand concepts can lead to fairly high retrieval accuracy
- We don’t need perfect concept detection
- We don’t need perfect combination weights
- There are enough concepts out there

How Much Progress is Realistic?

- We are getting closer to asking the right questions (not to answers)
- Machine learning is close to knee in an asymptotic curve

- Certain, targeted solutions will succeed well
  - Careful combination of multimedia analysis
  - Understanding of needs and opportunities

- I doubt many of the challenges will be solved
  - Most will become irrelevant

Recipe for Success:
Some research, much clever engineering, then more research!
Think of user needs.

Questions & Comments?

Acknowledgements:

Rong Yan, Cees Snoek, Howard Wactlar, Mike Christel, Milind Naphade, John Smith, Shih-Fu Chang, Jun Yang, Ming-yu Chen, Wei-Hao Lin, and the members from Informedia project

References (unsorted)

- Zinger S., Millet, C., Mathieu, B., Grefenstette, G., Hède, P., Moëllic, P.-A., Extracting an Ontology of Portrayable Objects from Wordnet, MUSCLE / ImageCLEF workshop on Image and Video retrieval evaluation, Vienna, Austria, pp. 17-23, 2005.
- C. Lin, B. Tseng, and J. Smith, “VideoAnnEx: IBM MPEG-7 annotation tool for multimedia indexing and concept
- learning,” in IEEE International Conference on Multimedia and Expo, 2003.
- J. Yang, M. Y. Chen, and A. G. Hauptmann, “Finding person x: Correlating names with visual appearances,” in Intl. Conf. on Image and Video Retrieval (CIVR’04), Ireland, 2004.
- Arden Alexander and Tracy Meehleib, “The thesaurus for graphic materials: Its history, use, and future,” Cataloging & Classification Quarterly, vol. 31, no. 3/4, pp. 189 – 212, 2001.
- Cees G.M. Snoek, Marcel Worring, Jan C. van Gemert, Jan-Mark Geusebroek, and Arnold W.M. Smeulders, “The challenge problem for automated detection of 101 semantic concepts in multimedia,” in Proceedings of ACM Multimedia, 2006.
- R. Yan, Probabilistic Models for Combining Diverse Knowledge Sources in Multimedia Retrieval, Ph.D. thesis, Carnegie Mellon University, 2006.
- Wei-Hao Lin and Alexander Hauptmann, “Which thousand words are worth a picture? experiments on video retrieval using a thousand concepts,” in 2006 IEEE International Conference on Multimedia and Expo (ICME), July 9–12, 2006.
- George Kingsley Zipf, Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology, Hafner Pub. Co, 1972.
- John R. Kender and Milind R. Naphade, “Visual concepts for news story tracking: Analyzing and exploiting the NIST TRECVID video annotation experiment,” in Proceedings of the 2005 Conference on Computer Vision and Pattern Recognition, 2005, pp. 1174–1181.
- Alexander G. Hauptmann, Wei-hao Lin, Rong Yan, Jun Yang, and Robert Chen, “Extreme video retrieval: Joint maximization of human and computer performance,” in Proceedings of the 14th ACM international conference on Multimedia, 2006.
- J. J. Rocchio, “Relevance feedback in information retrieval,” in The SMART Retrieval System: Experiments in Automatic Document Processing. Prentice Hall, Englewood Cliffs, NJ, 1971.
- R. Yan, J. Yang, and A. G. Hauptmann, “Learning query-class dependent weights in automatic video retrieval,” in Proceedings of the 12th annual ACM international conference on Multimedia, 2004, pp. 548–555.
- M. Christel and A. G. Hauptmann, “The use and utility of high-level semantic features,” in Intl. Conf. on Image and Video Retrieval (CIVR’05), Singapore, 2005.
- R. Yan and A. G. Hauptmann, “The combination limit in multimedia retrieval,” in Proc. of the eleventh ACM international conference on Multimedia, 2003, pp. 339–342.
- M. Naphade, J. R. Smith, J. Tesic, S.-F. Chang, W. Hsu, L. Kennedy, A. Hauptmann, and J. Curtis, Large-scale concept ontology for multimedia," IEEE Multi-Media, vol. 13, no. 3, pp. 86-91, 2006.
- R. Yan and A. G. Hauptmann. Efficient margin-based rank learning algorithms for information retrieval. In International Conference on Image and Video Retrieval (CIVR), 2006.
- R. Yan and A. G. Hauptmann. Probabilistic latent query analysis for combining multiple retrieval sources. In Proceedings of the 29th international ACM SIGIR conference, Seattle, WA, 2006.