Summer School on Multimedia Semantics
Chalkidiki, Greece, Sep 2006

SEMANTIC FEATURES EXTRACTION AND REPRESENTATION

Milind R. Naphade
IBM Thomas J. Watson Research Center

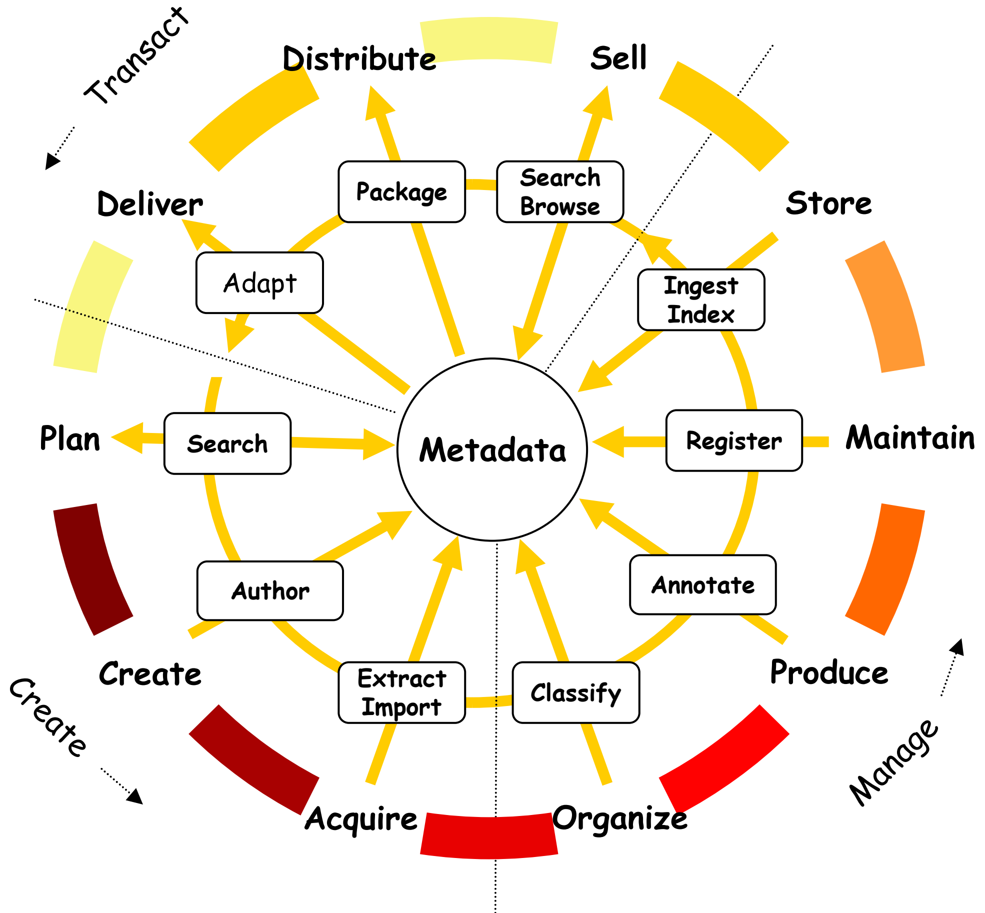
Organization

1. Motivation – Why
2. Extracting Semantics – How
   - Learning and Extraction
   - Evaluation
3. Feature Selection -- What
4. Challenges and Gaps – What next?
5. Demo & Case Study -- MARVEL

Why Bother?

Motivation

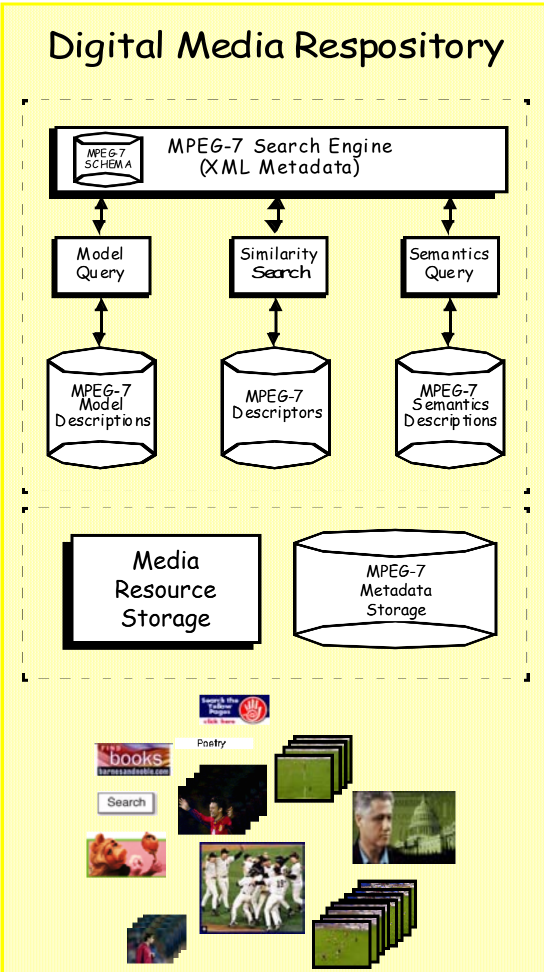
FROM DATA TO DATA+METADATA



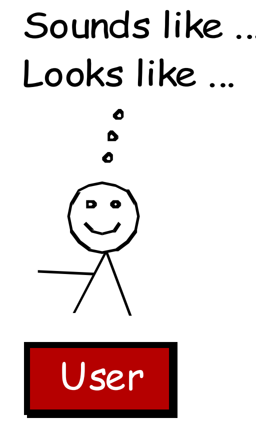
- **Metadata is critical for describing essential aspects of content:**
  - Main topics, author, language, publication, etc.
  - Events, scenes, objects, times, places, etc.
  - Rights, packaging, access control, content adaptation, etc.
- **Conformity with open metadata standards will be vital:**
  - Allows faster design and implementation
  - Interoperability with broad field of competitive standards-based tools and systems
  - Rich set of standards-based technologies for critical functions such as content extraction, advanced search, and personalization

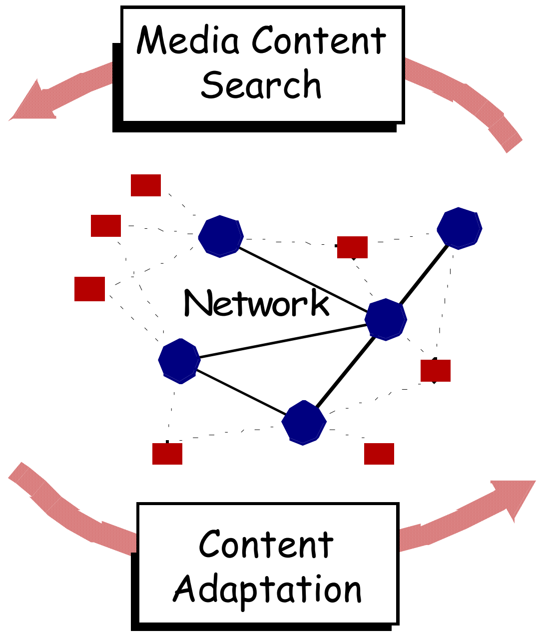
- Metadata provides solution for interoperable management throughout media content lifecycle (Create → Manage → Distribution / Transact)

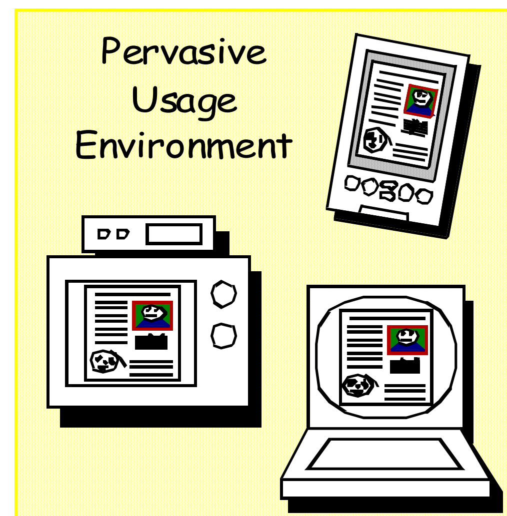
Media Content Management



- **MPEG-7 Indexing & Searching:**
  - Semantics-based (people, places, events, objects, scenes)
  - Immutable metadata (title, authors)
  - Content-based (color, texture, motion, melody, timbre)







- **MPEG-7 Access & Delivery:**
  - Media content personalization
  - Adaptation & summarization
  - Usage environment (usage context, devices, user preferences)

TRENDS

DATA

- 70,000 TB (or 101 million hours) of original TV and radio production in 2002*
- New information growing at 30% per year

METADATA

- Business value delivered when content can be leveraged meaningfully
- Manual annotation of rich media is costly, inadequate and often incomplete
- Increasing expectations of accessibility and searchability of rich media content

TECHNOLOGY TRENDS

- Cost of computation, communication and storage decreasing drastically
- Signal Processing & Machine Learning providing new capabilities for deeper analysis

INVESTMENT

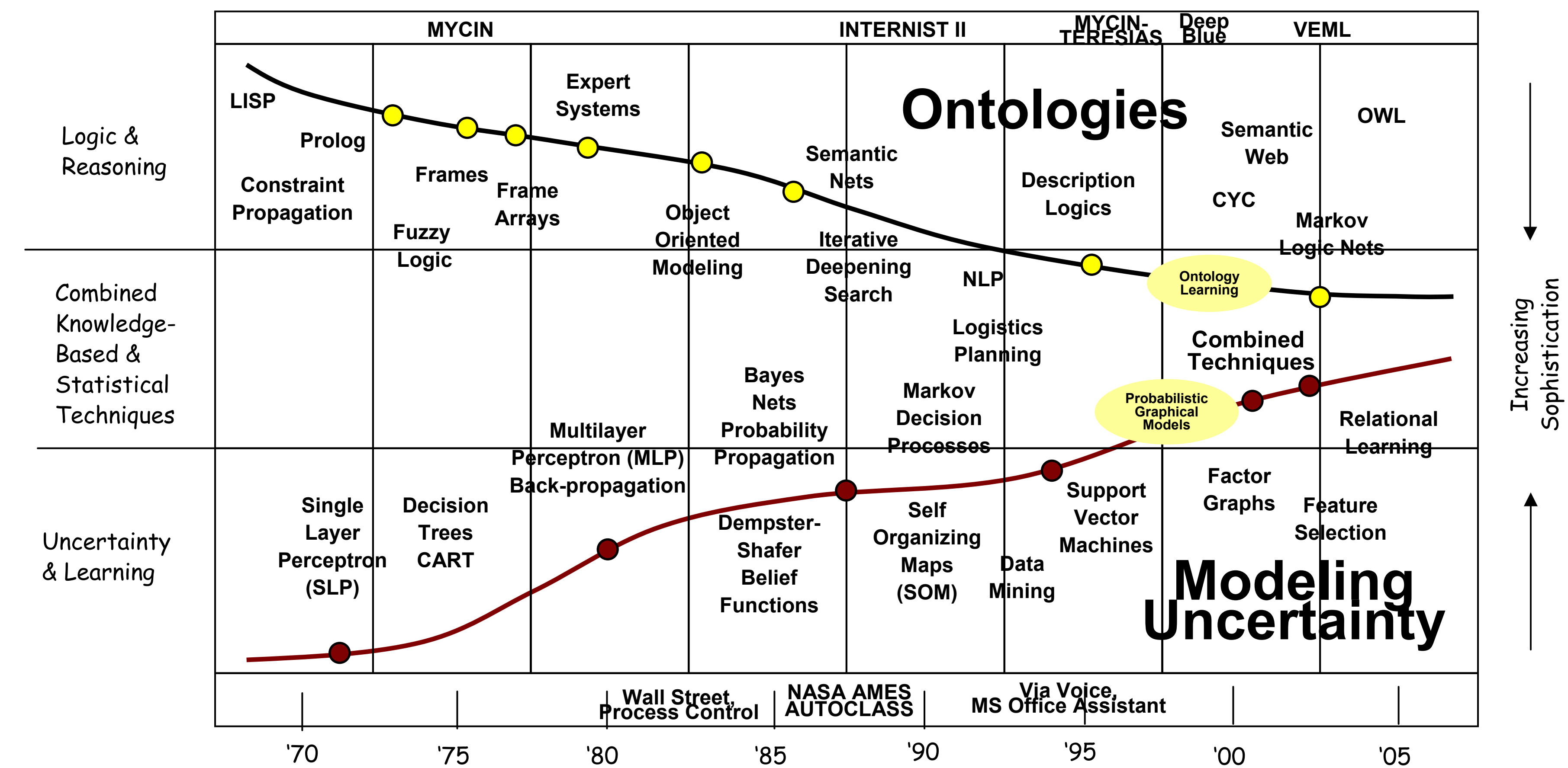
- Government agencies in America, Europe and Asia investing in several projects
- Media enterprises want to embrace promising technologies
- Web Search demands scalable technologies

ACADEMIA

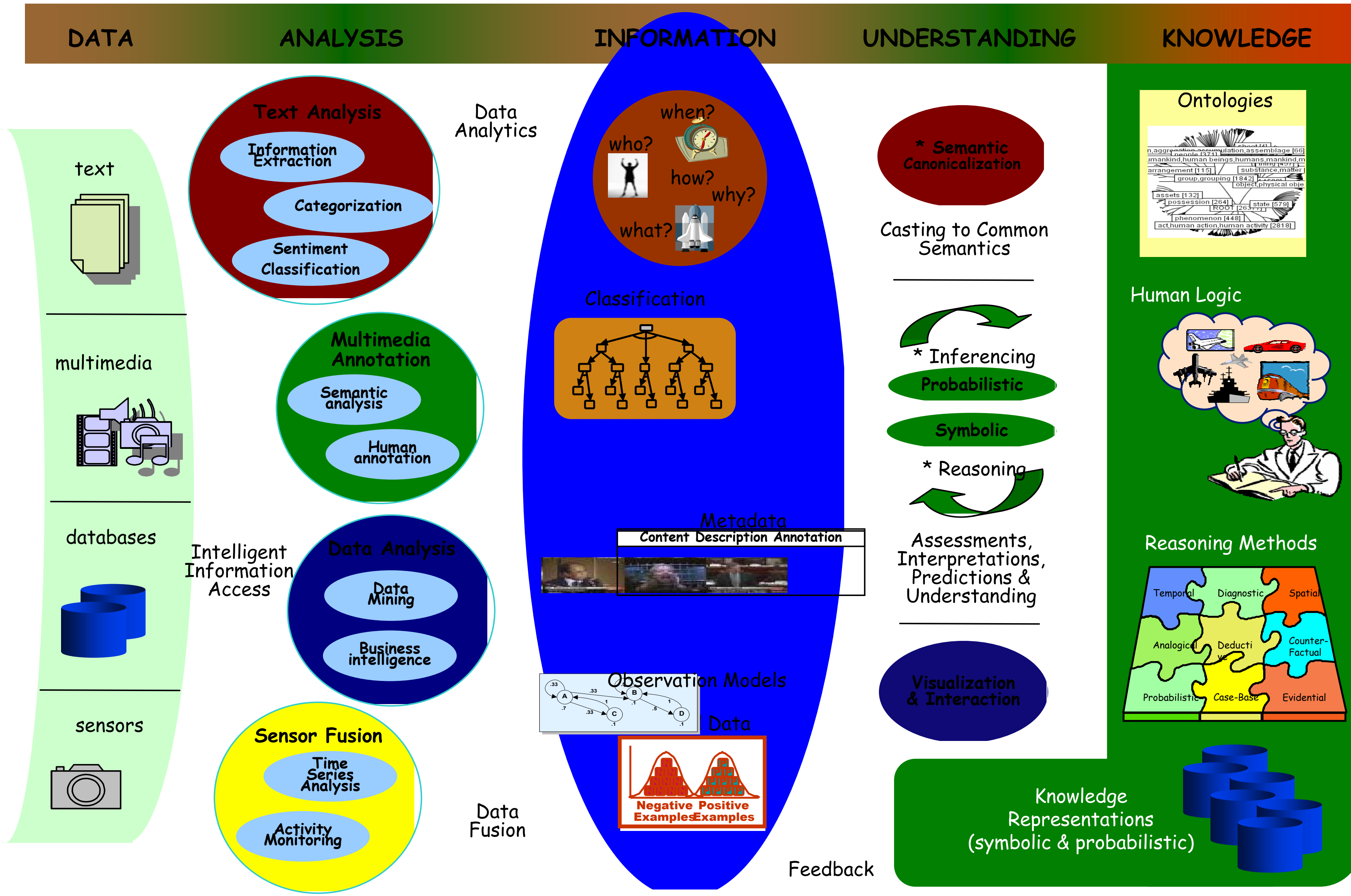
- Excellent network of academic collaboration across continents resulting in such successful joint ventures as this workshop, critical mass at TRECVID etc.

* UC Berkeley Study – "How Much Information", 2003

Confluence of Statistical Analysis & Knowledge-based Inference



- Increasing sophistication in knowledge-based and probabilistic-based inferencing & learning techniques and trends towards convergence


DATA
ANALYSIS
INFORMATION
UNDERSTANDING
KNOWLEDGE
text
multimedia
databases
sensors
Text Analysis
Information Extraction
Categorization
Sentiment Classification
Data Analytics
Multimedia Annotation
Semantic analysis
Human annotation
Intelligent Information Access
Data Analysis
Data Mining
Business intelligence
Sensor Fusion
Time Series Analysis
Activity Monitoring
Data Fusion
when?
who?
how?
why?
what?
Classification
Metadata
Content Description Annotation
Observation Models
Data
Negative Examples
Positive Examples
Feedback
* Semantic Canonicalization
Casting to Common Semantics
* Inferencing
Probabilistic
Symbolic
* Reasoning
Assessments, Interpretations, Predictions & Understanding
Visualization & Interaction
Ontologies
Human Logic
Reasoning Methods
Temporal
Diagnostic
Spatial
Analogical
Deductive
Counter-Factual
Probabilistic
Case-Based
Evidential
Knowledge Representations (symbolic & probabilistic)

Challenges in Semantic Video Management

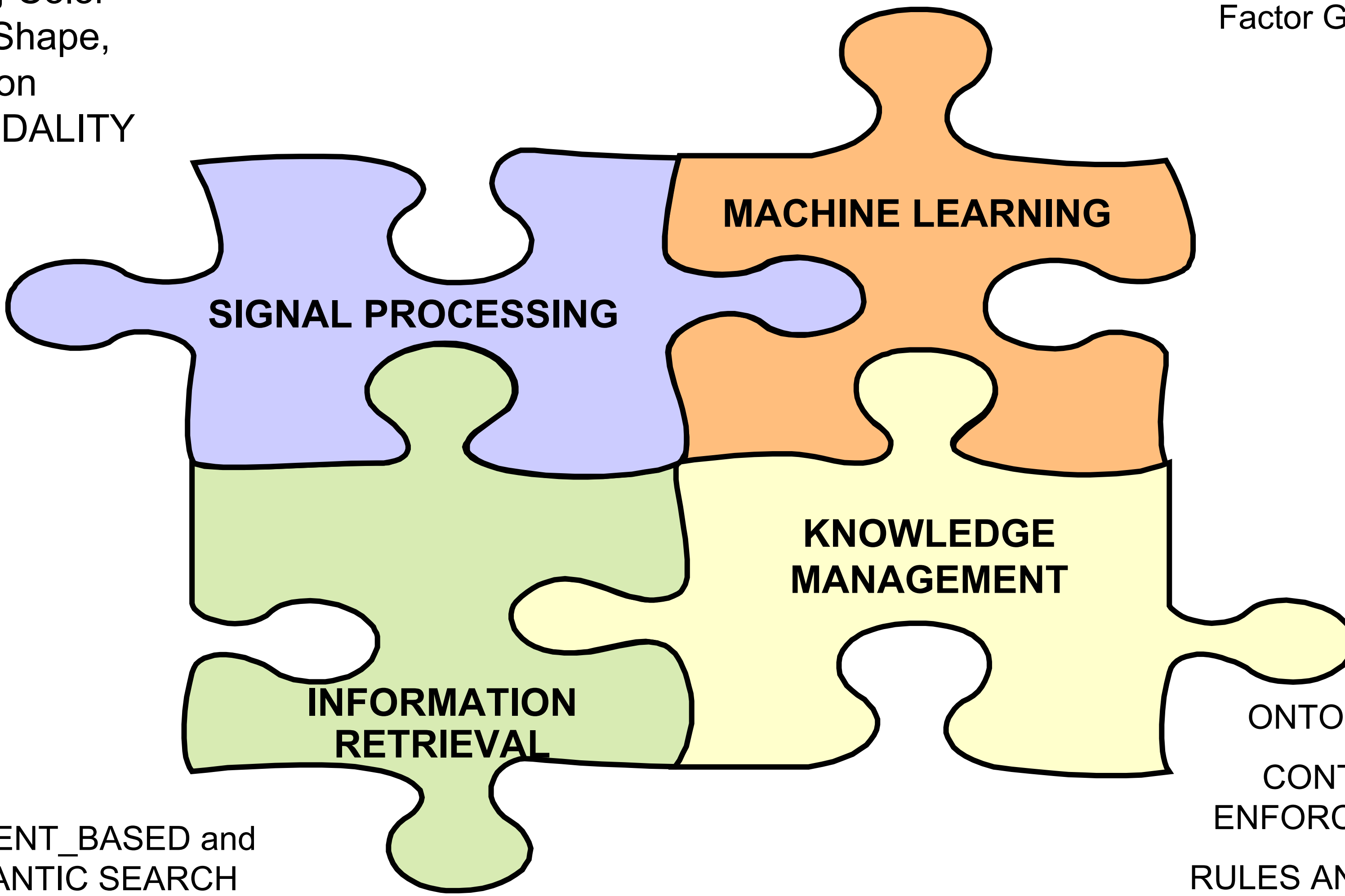
- Mapping low-level features to semantic features.
- Set of basic units that exhaust semantic space completely (as in phonemes in ASR).
- Grammar
- Fusion.
  - Modality (audio, visual, text).
  - Feature (color, texture, structure, motion).
  - Decision.
- User Interaction.
  - Minimal annotation,
  - Relevance feedback etc..
- Query Processing

MULTIMEDIA SEMANTICS: The JIGSAW PUZZLE

Time-Frequency Analysis, Color Texture Shape, Motion
MULTIMODALITY

SVMs, HMMs,

Factor Graphs



ONTOLOGY

CONTEXT ENFORCEMENT

RULES AND LOGIC

CONTENT_BASED and SEMANTIC SEARCH

Extracting Semantic Features

Challenges of Multimedia Learning

Problem	Approach
Tremendous variability and uncertainty	Framework must take uncertainty into account
Small number of training examples (relative to feature dimensionality)	Exhaustive training techniques such as those for ASR not possible
Complex distributions, highly non-linear decision boundaries, high-dimensional feature spaces	Employ feature selection and dimensionality reduction. Linear classifiers not sufficient.
Manual annotation is time-consuming expensive, human barrier	Learning needs to be user-centric
Dependence on a host of scientific disciplines for extracting good features, none of which have been perfected	Must get around imperfect segmentation, single-channel auditory non-separability
Multiple Channels with possible relationships that are unknown	Need to fuse information

Concept Modeling & Detection

TASK:

- *Learn to extract semantic labels from multimedia*

MOTIVATION:

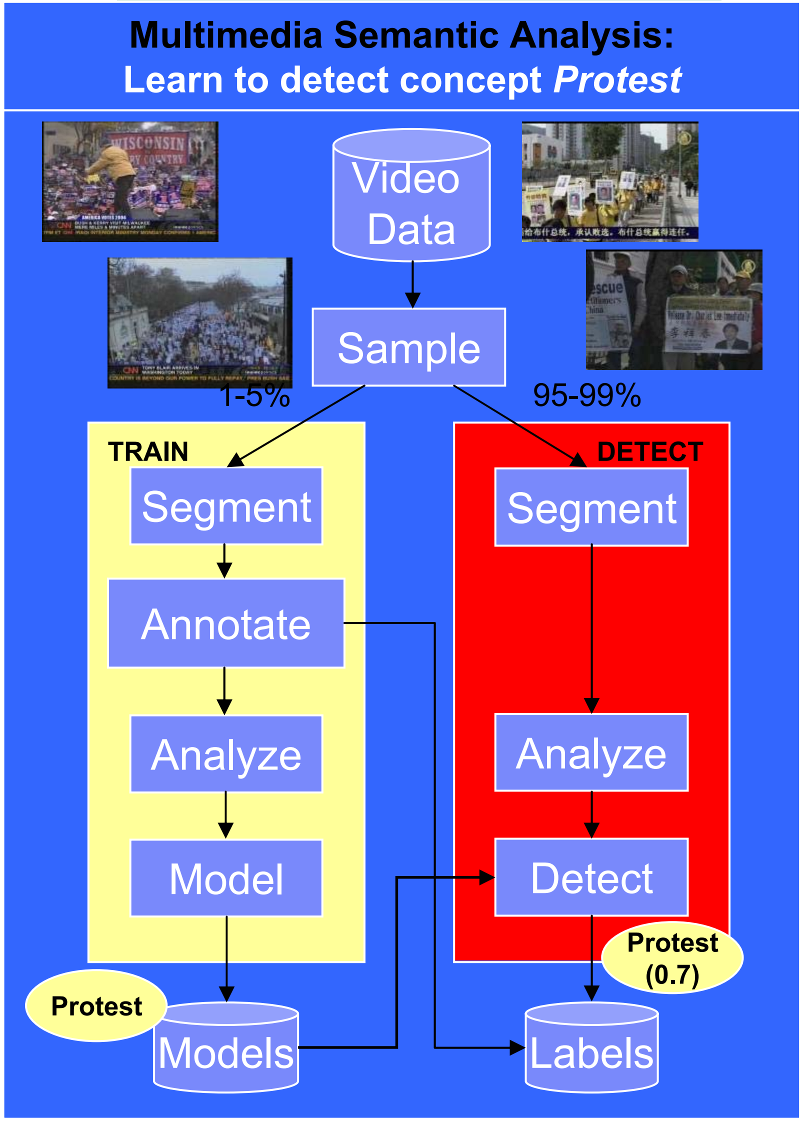
- Manual labeling is human resource intensive (10x)
- Results in incomplete & inconsistent annotations
- Traditional metadata is not enough
- Need to look at content and index semantically

APPROACH:

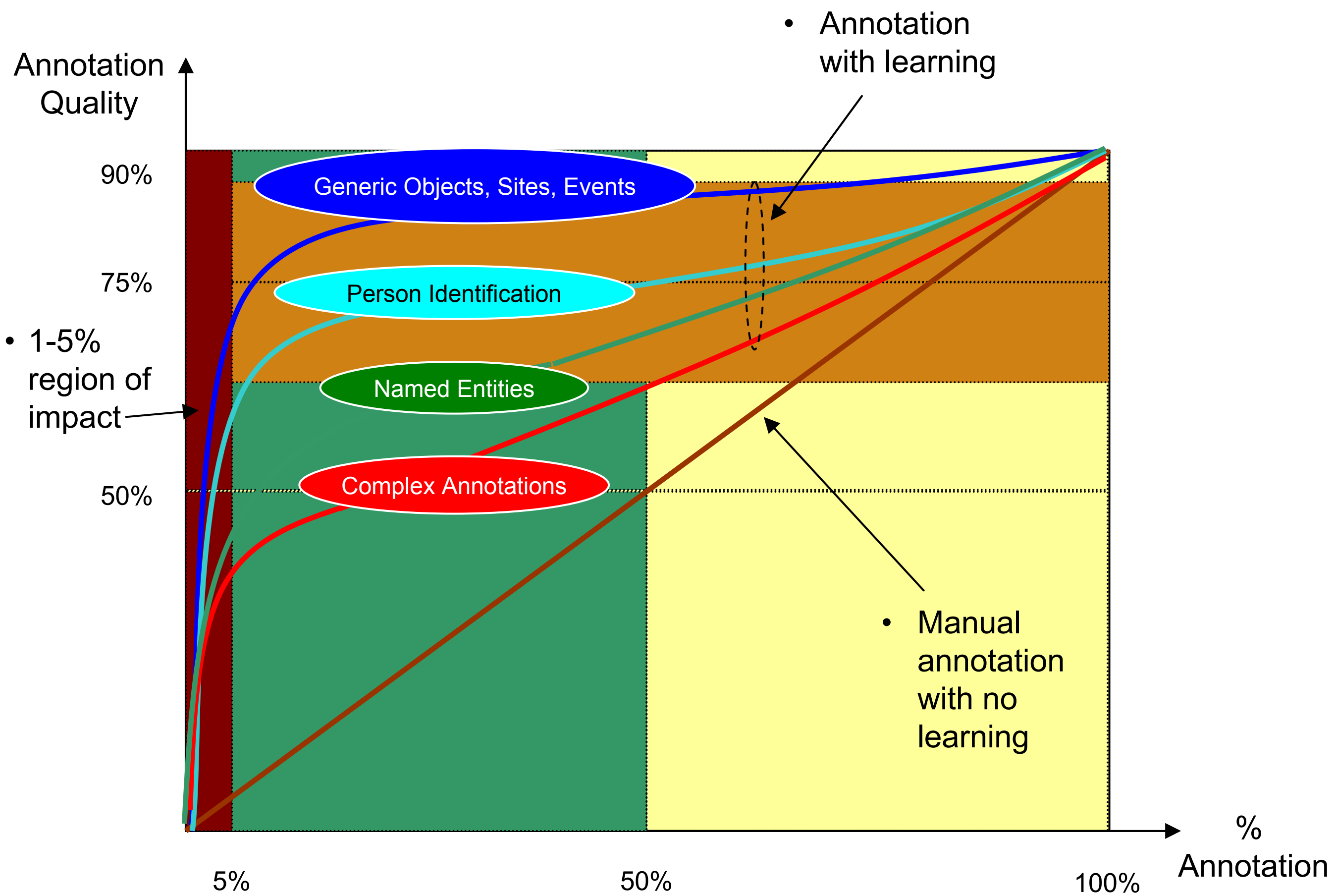
- Replace manual process with learning approach
- Annotate small sample of training data
- Learn concept models from training data
- Apply models to detect concepts in new data
- Propagate labels and confidence scores

CHALLENGES:

- Increase detection accuracy
- Reduce amount of supervision

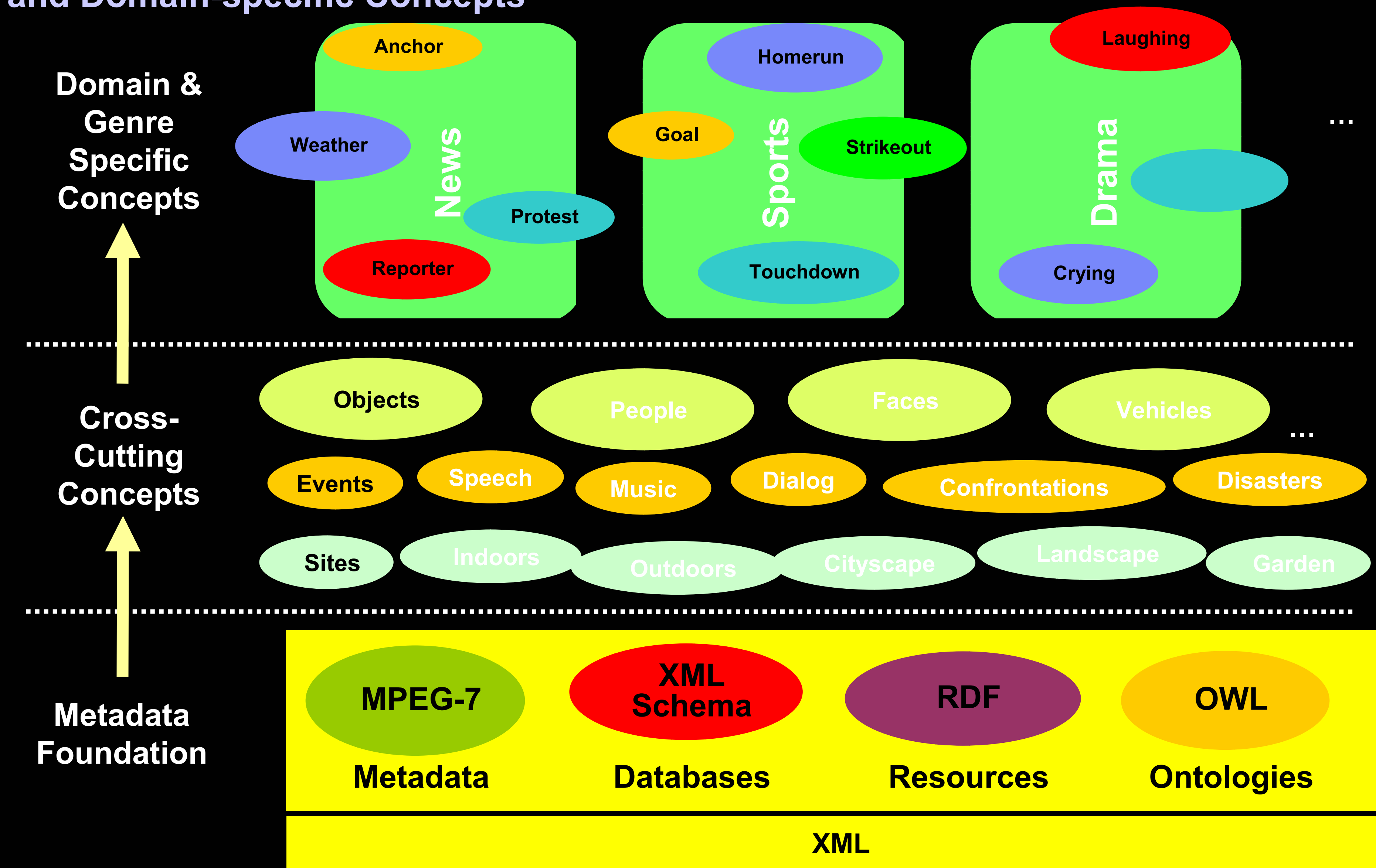


VALUE PROPOSITION

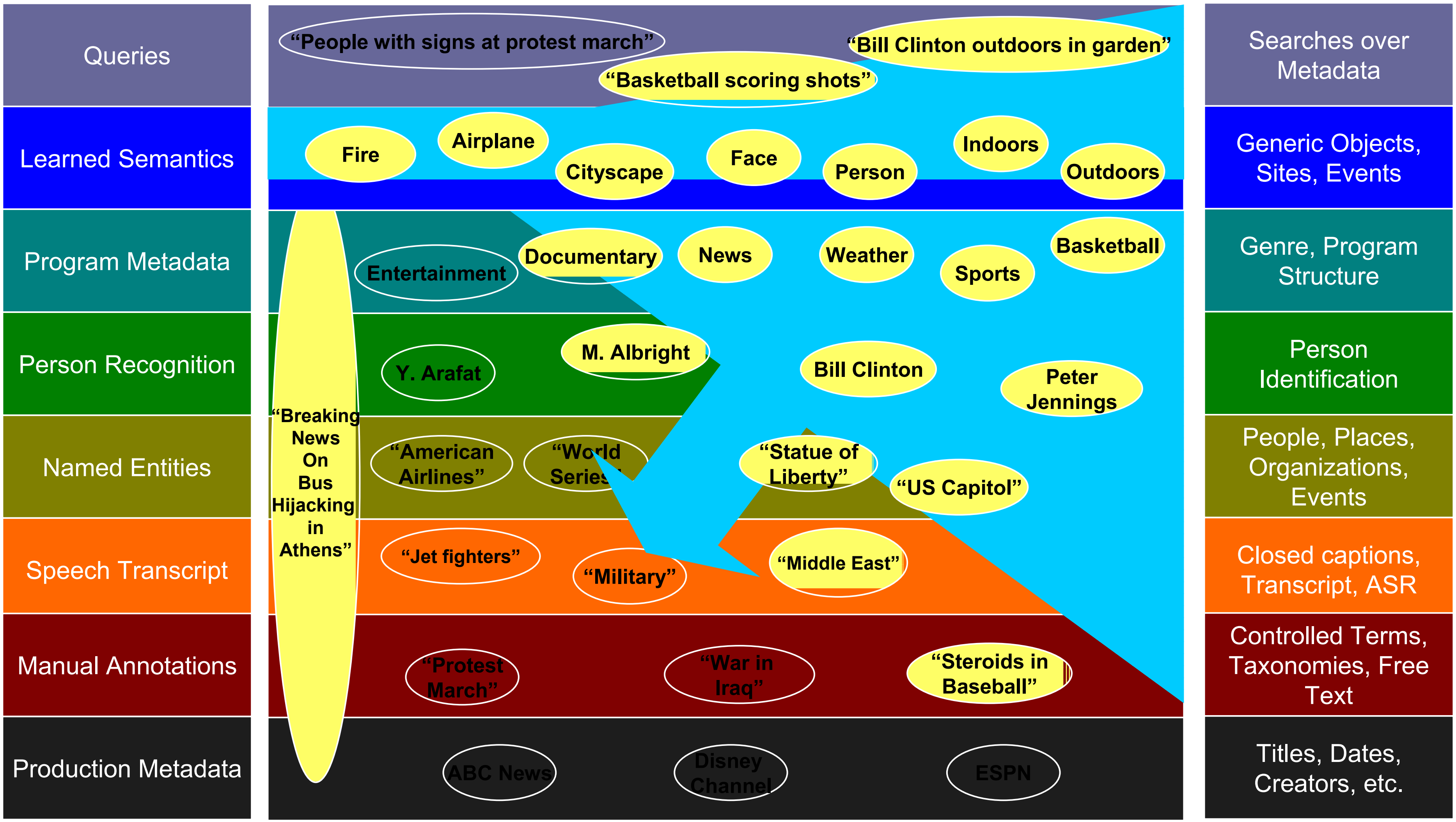


- Manual annotation achieves high annotation quality only with high completeness
- Semantics learning improves annotation quality at all levels of completeness
- Significant gain in annotation quality results from modest levels of training

Semantics Concept Ontology can be Designed to Support both Cross-cutting and Domain-specific Concepts



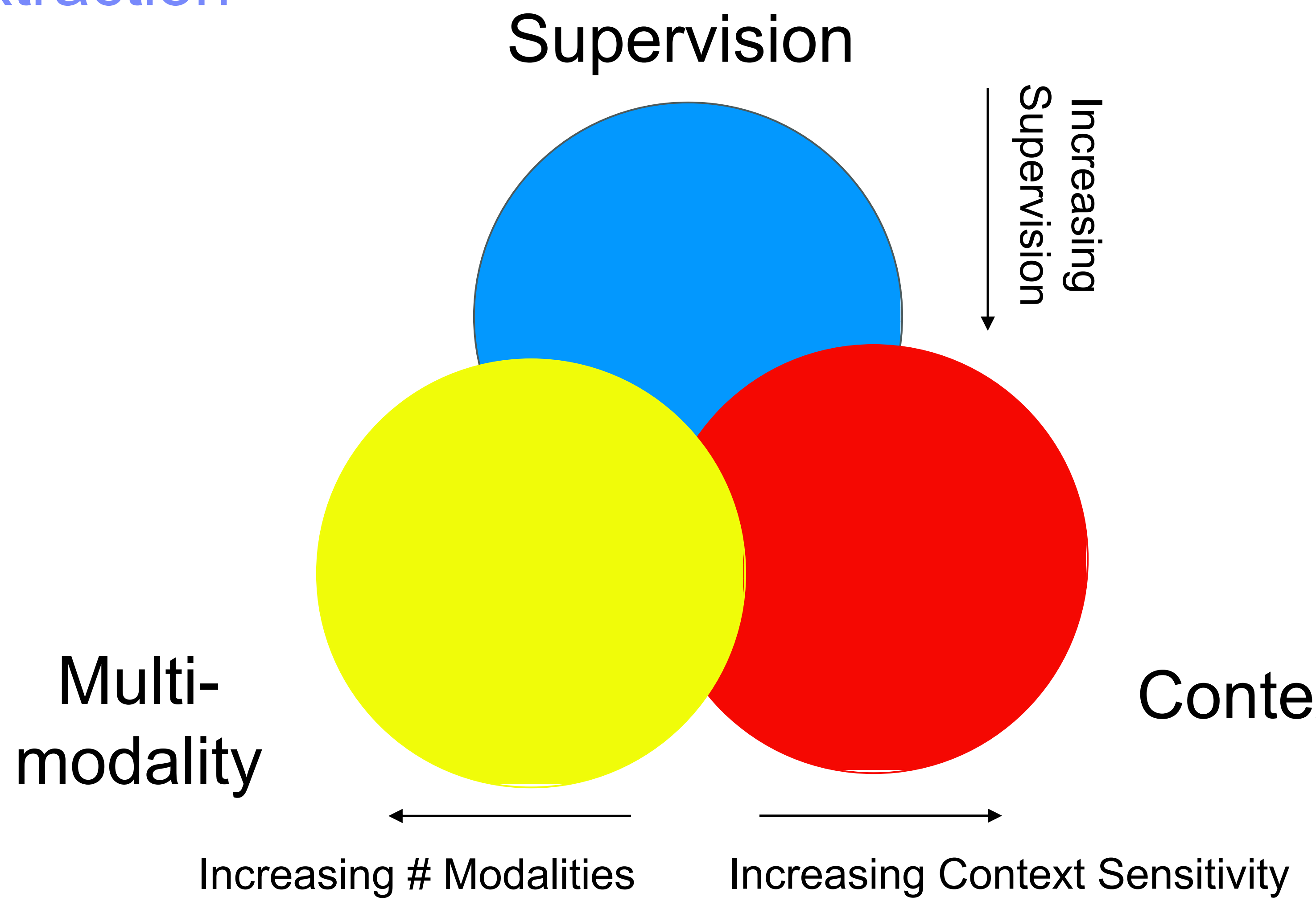
Coverage of Automation Keeps Increasing



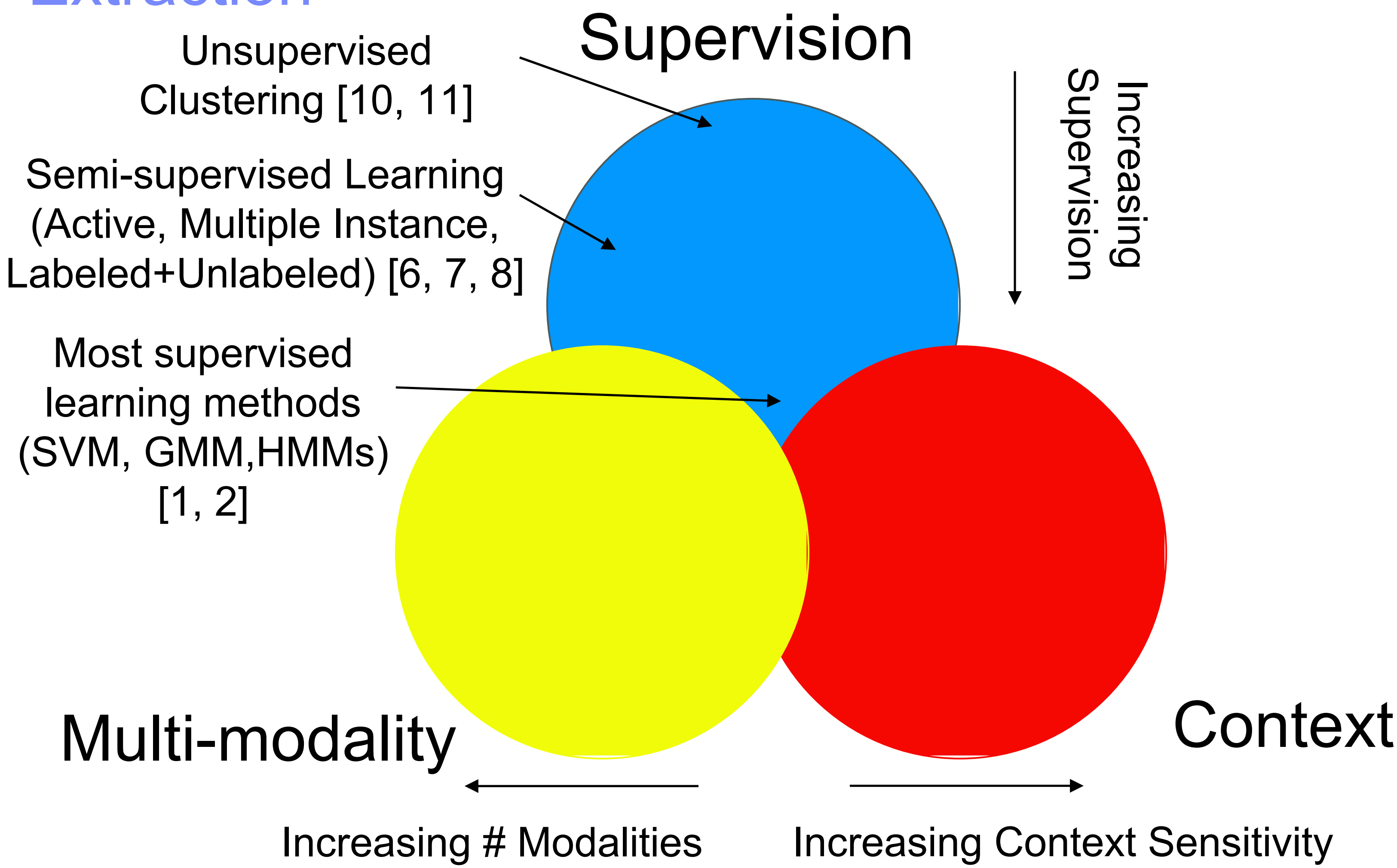
Learning Multimedia Semantics

A. Supervised Detection
   1. Static Classifiers
   2. Spatial+Temporal Classifiers
B. Multimodal Fusion
   3. Late fusion using Ensembles
   4. Intermediate Fusion for temporal evolution using graphical models
C. Enforcing Spatial, Temporal and Conceptual Context
   5. Learning Context using Multinet
D. Semi-Supervised Learning
   6. Labeled+Unlabeled Learning
   7. Active Learning
   8. Multiple Instance Learning
   9. Co-training
E. Unsupervised Clustering
   10. Spatial
   11. Spatio-temporal using hierarchical HMMs
F. Semantic Feature Extraction and Search
   12. Query Learning
   13. Leveraging detected semantic concepts for complex query answering

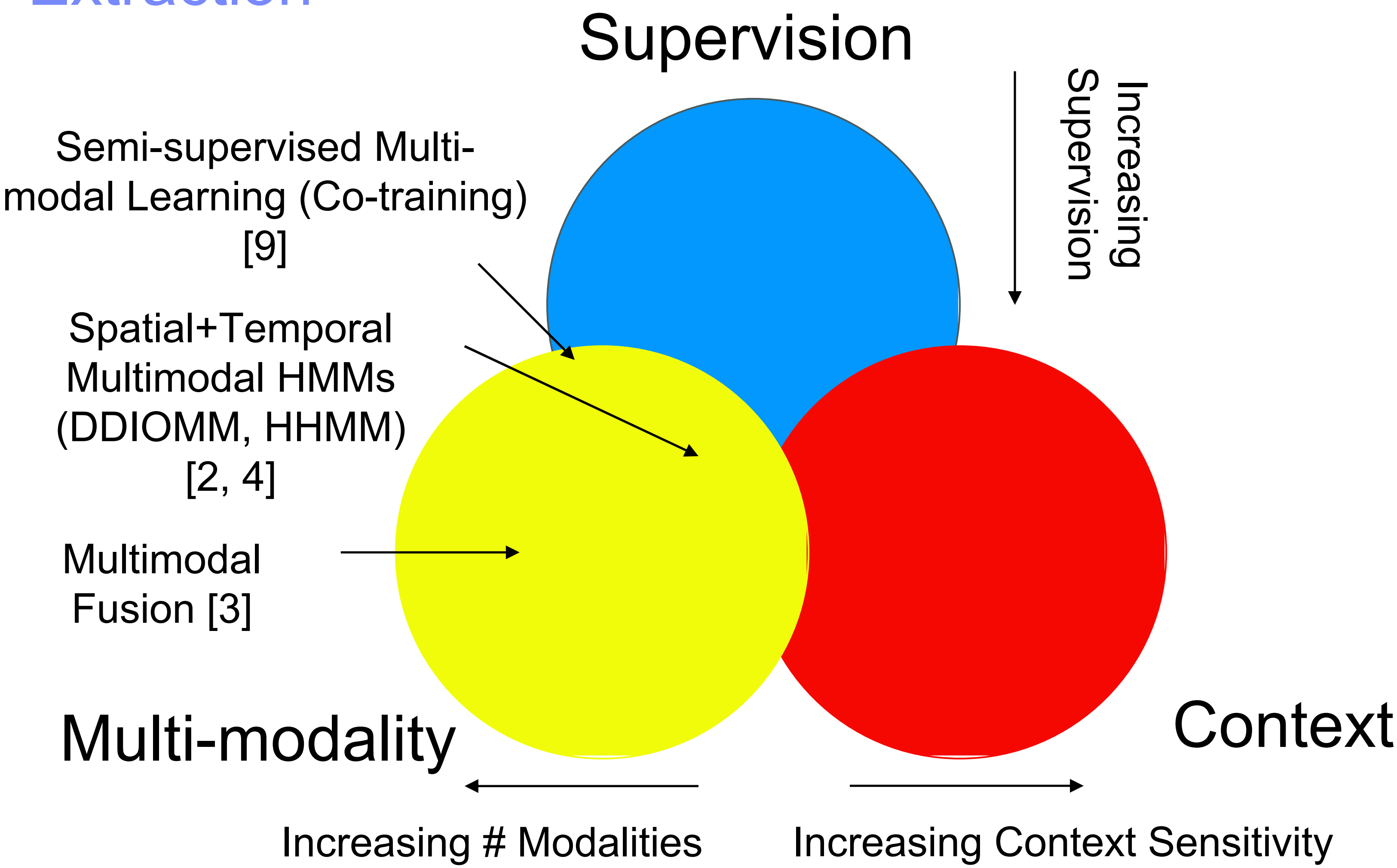
The Landscape of Multimedia Semantic Feature Extraction



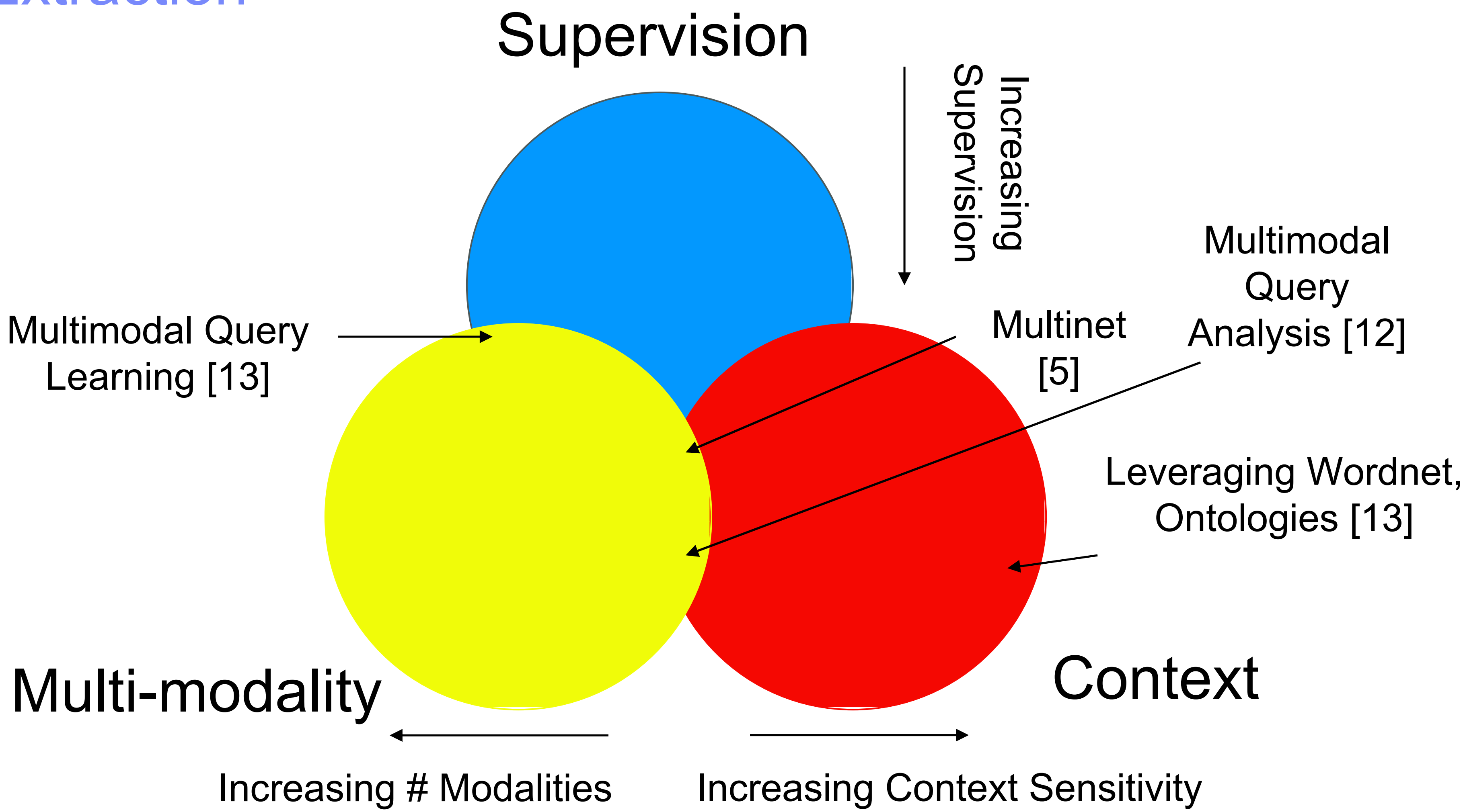
The Landscape of Multimedia Semantic Feature Extraction



The Landscape of Multimedia Semantic Feature Extraction



The Landscape of Multimedia Semantic Feature Extraction



Learning Multimedia Semantics

A. **Supervised Detection**
   1. **Static Classifiers**
   2. **Spatial+Temporal Classifiers**

B. Multimodal Fusion
   3. Late fusion using Ensembles
   4. Intermediate Fusion for temporal evolution using graphical models

C. Enforcing Spatial, Temporal and Conceptual Context
   5. Learning Context using Multinet

D. Semi-Supervised Learning
   6. Labeled+Unlabeled Learning
   7. Active Learning
   8. Multiple Instance Learning
   9. Co-training

E. Unsupervised Clustering
   10. Spatial
   11. Spatio-temporal using hierarchical HMMs

F. Semantic Feature Extraction and Search
   12. Query Learning
   13. Leveraging detected semantic concepts for complex query answering

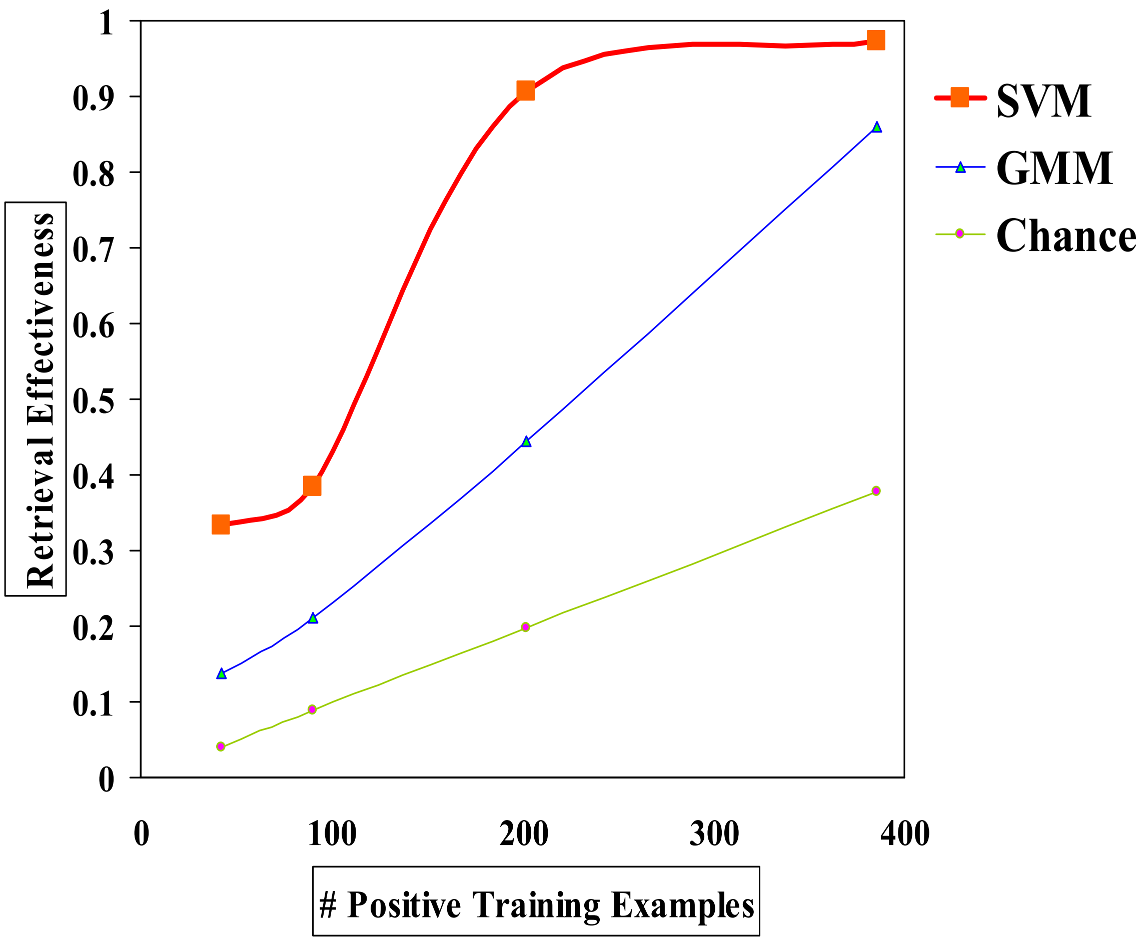
Supervised Learning for Concept Modeling: Nutshell

- **Problem:**
  - Automatically detect concepts and extract semantic labels from video
- **Approach**:
  - **Training**: Assume multimodal examples for each semantic concept
  - **Feature Extraction**: Automatically extract visual and auditory features
  - **Statistical Learning**: Learn parametric models to represent concepts in terms of distribution of features. Use validation set to select optimal model settings.
  - **Detection**: Use the trained model for detecting semantic concepts
- **Result Summary**:
  - Discriminant Learning better suited to problem of multimedia concept detection than Density Modeling.
  - Over 100 semantic concept models built for TRECVID benchmark corpora.
  - SVM-based detection approach results in the highest mean average precision in five years of the benchmark concepts including visual concepts such as Outdoors, Indoors, People, Cityscape, etc.
  - Statistical model-based approach improves retrieval effectiveness over content-based approaches
  - Enables semantic filtering, access, search and retrieval

Popular Modeling Approaches

Density Modeling	Decision Boundary Modeling
Aim is to model the distribution of features under multiple hypotheses	Aim is to maximize classification accuracy
Graphical Models: Bayesian Nets, Markov Random Fields, Factor Graphs etc.	Discriminant Classifiers Neural Networks, Kernel machines etc.
Learning is based on maximizing likelihood of data given model parameters. EM most popular for this optimization.	Learning based on minimizing empirical risk. Non-linear optimization solved mostly using gradient-based methods.
Robust when corpus for training is large.	Suffers from the threat of over-fitting on the training set.
Model selection uses MDL and such principles	Model selection is ad-hoc

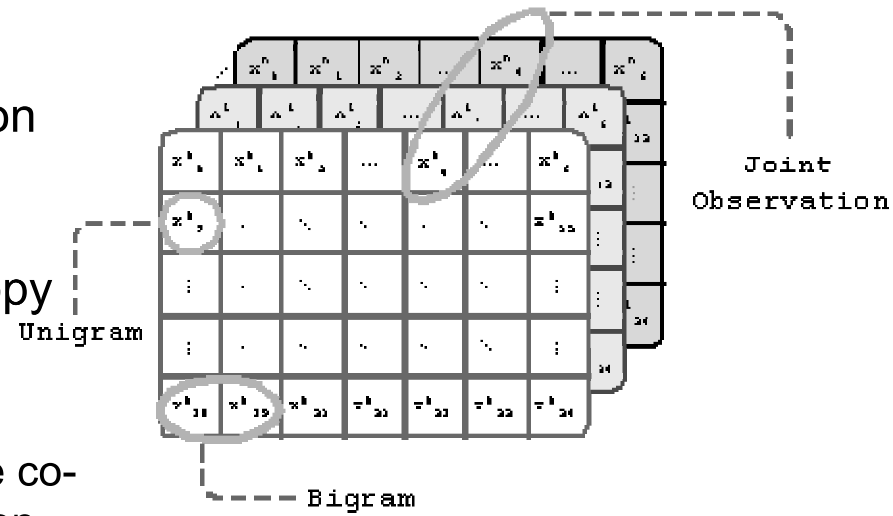
Number of Training Samples and Performance Comparison



- SVM needs fewer training examples to than GMMs to ramp up performance
- When sufficient training samples available, both algorithms perform similarly.
- Each data-point on the curves is a different semantic concept.

Maximum Entropy Approach for Concept detection w/o regional annotation

- Key-frame partitioned into regular grid
- Low-level features extracted from each region
- Extracted features are tokenized using K-means.
- Statistical information to the Maximum Entropy model is presented via specially designed predicates:
  - o*Unigram* predicates are defined to capture the co-occurrence statistics between manual annotation and tokenized feature.
  - o*Bigram* predicates capture the relationships between horizontal and vertical neighboring region.
  - o*Place Dependent* predicates are defined to capture location specific statistics.
  - o*Joint Observation* predicates are defined to capture interactions between the visual low-level features.



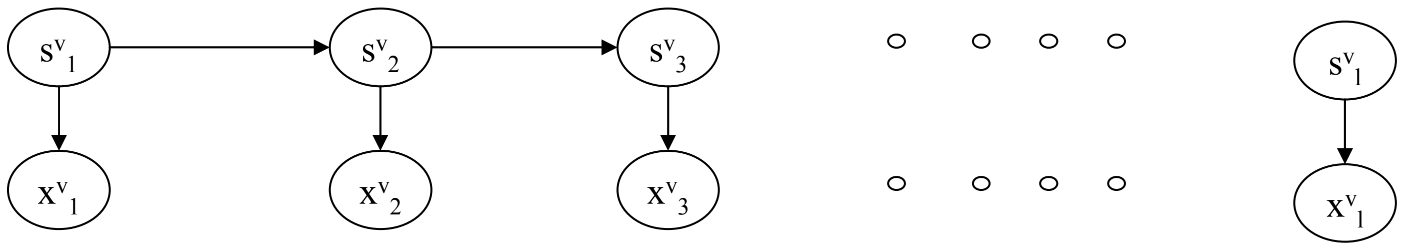
$$p(y|x) = \frac{\exp\left(\sum_i \lambda_i f_i(x,y)\right)}{Z_\lambda(x)}$$

$$Z_\lambda(x) = \sum_y \exp\left(\sum_i \lambda_i f_i(x,y)\right)$$

$$\lambda^* = \arg\max_\lambda \Psi(\lambda)$$

Hidden Markov Models for Event Modeling

- Hidden Markov models used for temporal event detection based on their successful application in Speech Recognition
- Application of HMMs for modeling events in various domains including movie events (explosion etc.), sports, aural events, news videos, surveillance, etc.
- Composed of states with observation densities and transitions between states to capture change of active state in events.
- Several variants for hierarchical processing, and multi-modal fusion



Learning Multimedia Semantics

A. Supervised Detection
   1. Static Classifiers
   2. Spatial+Temporal Classifiers

B. <u>Multimodal Fusion</u>
   3. <u>Late fusion using Ensembles</u>
   4. <u>Intermediate Fusion for temporal evolution using graphical models</u>

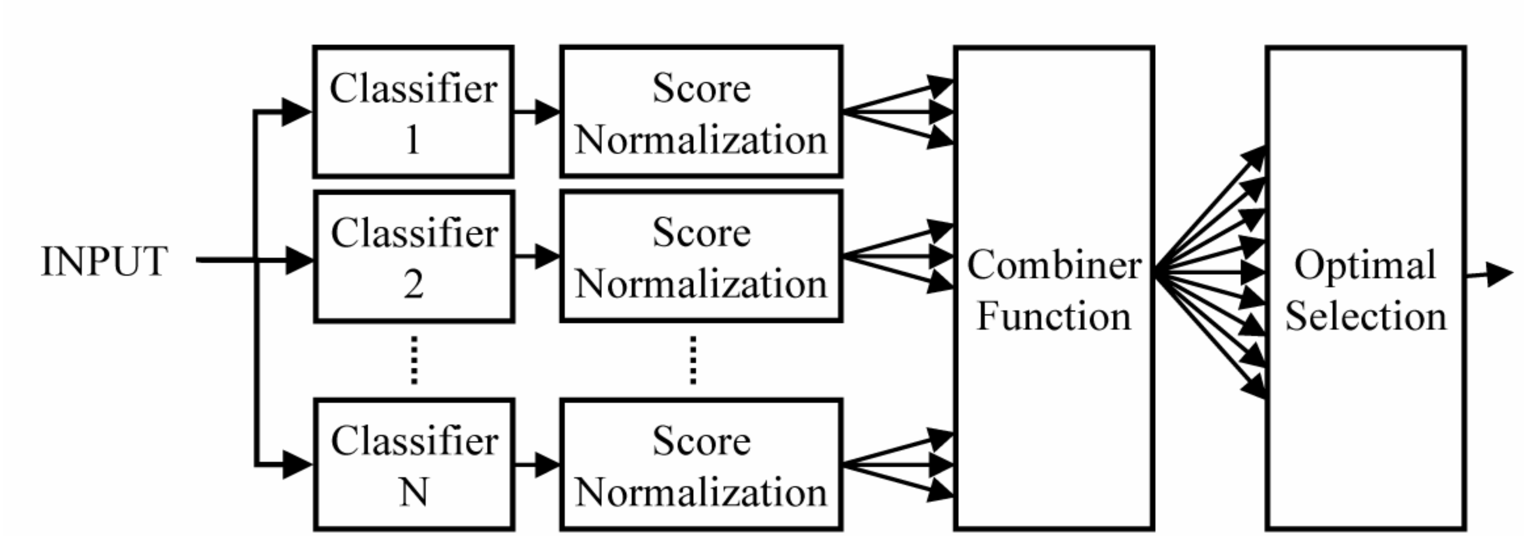
C. Enforcing Spatial, Temporal and Conceptual Context
   5. Learning Context using Multinet

D. Semi-Supervised Learning
   6. Labeled+Unlabeled Learning
   7. Active Learning
   8. Multiple Instance Learning
   9. Co-training

E. Unsupervised Clustering
   10. Spatial
   11. Spatio-temporal using hierarchical HMMs

F. Semantic Feature Extraction and Search
   12. Query Learning
   13. Leveraging detected semantic concepts for complex query answering

Multi-Modality/ Multi-Concept Fusion Methods



Ensemble Fusion:

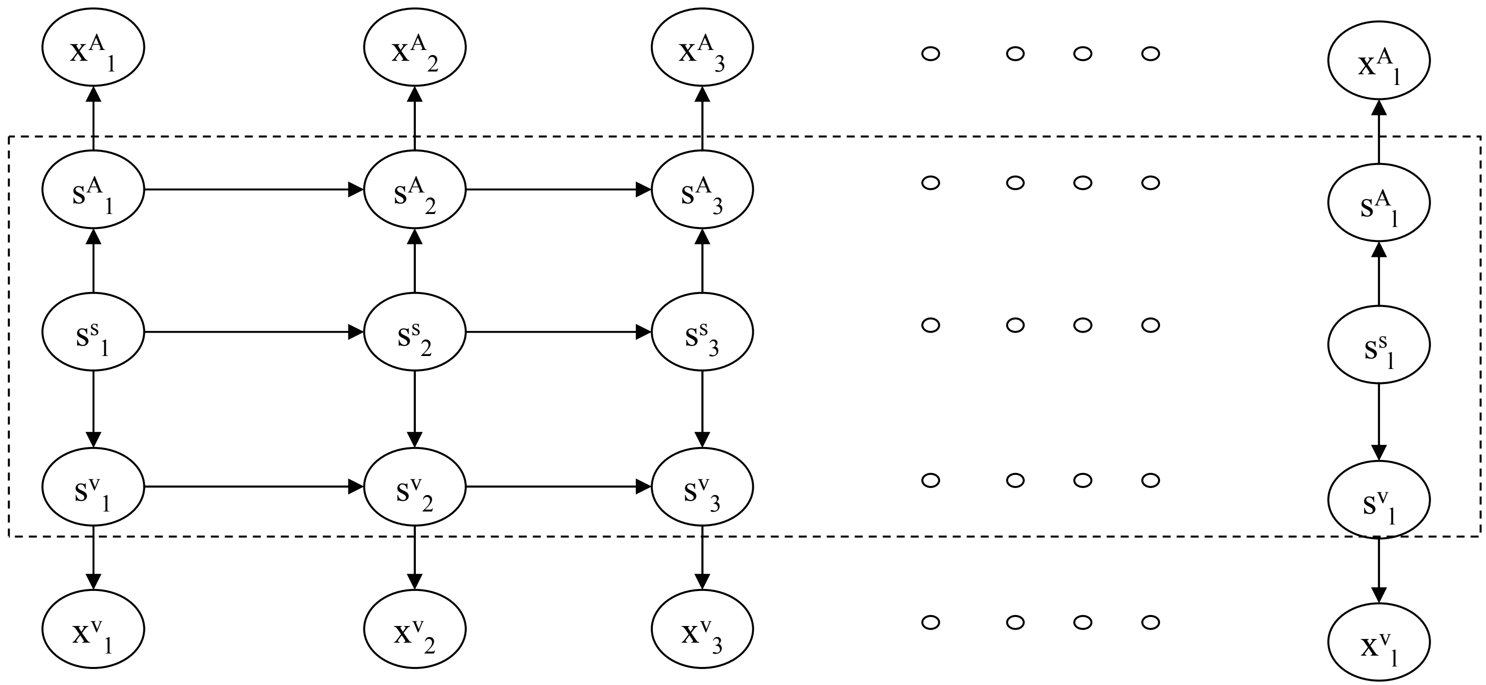
- Normalization: rank, Gaussian, linear.
- Combination: average, product, min, max
- Works well for uni-modal concepts with few training examples
- Computationally low-cost method of combining multiple classifiers.

Fusion as a classification problem

•Similar approach as in classification except that now the supervised scheme uses detection results of different models and learns based on joint predicates

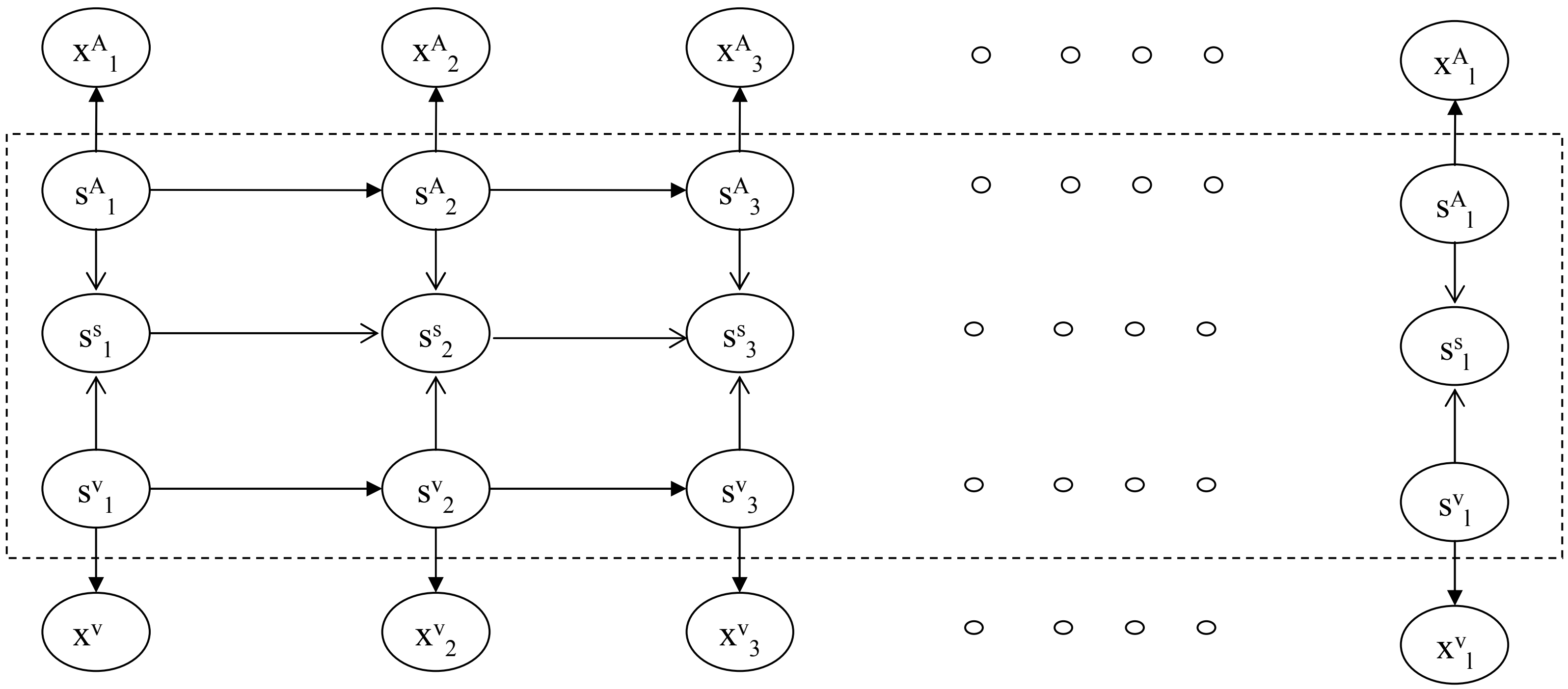
Multimodal Fusion
Hierarchical hidden Markov models

Late integration of audio and video through the sequences of the hidden states of the audio and video HMM. The decoded state sequences are treated as observations of the supervisor HMM.



Multimodal Fusion Duration density input output Markov model

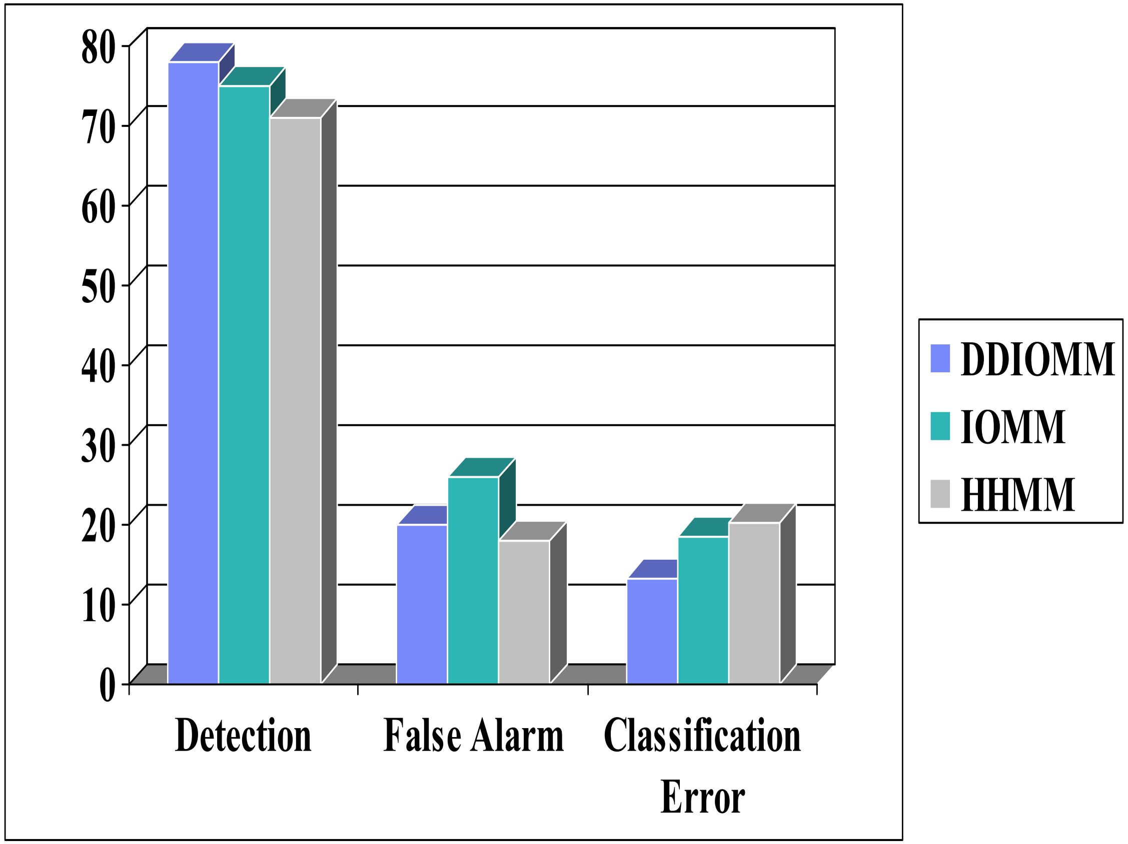
The decoded state sequences are treated as input sequences and the multimodal decisions are considered the output sequence. Using explicit duration models, the output sequence is predicted based on the input sequences.



Performance Comparison for Event Detection

Visual Features
 Color: HSV histogram, Moments.
 Texture: Edge direction histogram.
 Gray-level Co-occurrence
 Shape: Moment Invariants
Audio Features
 15 MFCC coefficients,
 15 delta coefficients
 2 energy coefficients

- We use 9 clips with a leave one out strategy and compare performance of HHMM with IOMM and DDIOM for the event **explosion**



Learning Multimedia Semantics

A. Supervised Detection
   - Static Classifiers
   - Spatial+Temporal Classifiers
B. Multimodal Fusion
   - Late fusion using Ensembles
   - Intermediate Fusion for temporal evolution using graphical models
C. **Enforcing Spatial, Temporal and Conceptual Context**
   - **Learning Context using Multinet**
D. Semi-Supervised Learning
   - Labeled+Unlabeled Learning
   - Active Learning
   - Multiple Instance Learning
   - Co-training
E. Unsupervised Clustering
   - Spatial
   - Spatio-temporal using hierarchical HMMs
F. Semantic Feature Extraction and Search
   - Query Learning
   - Leveraging detected semantic concepts for complex query answering

Modeling and Enforcing Semantic Context: Nutshell

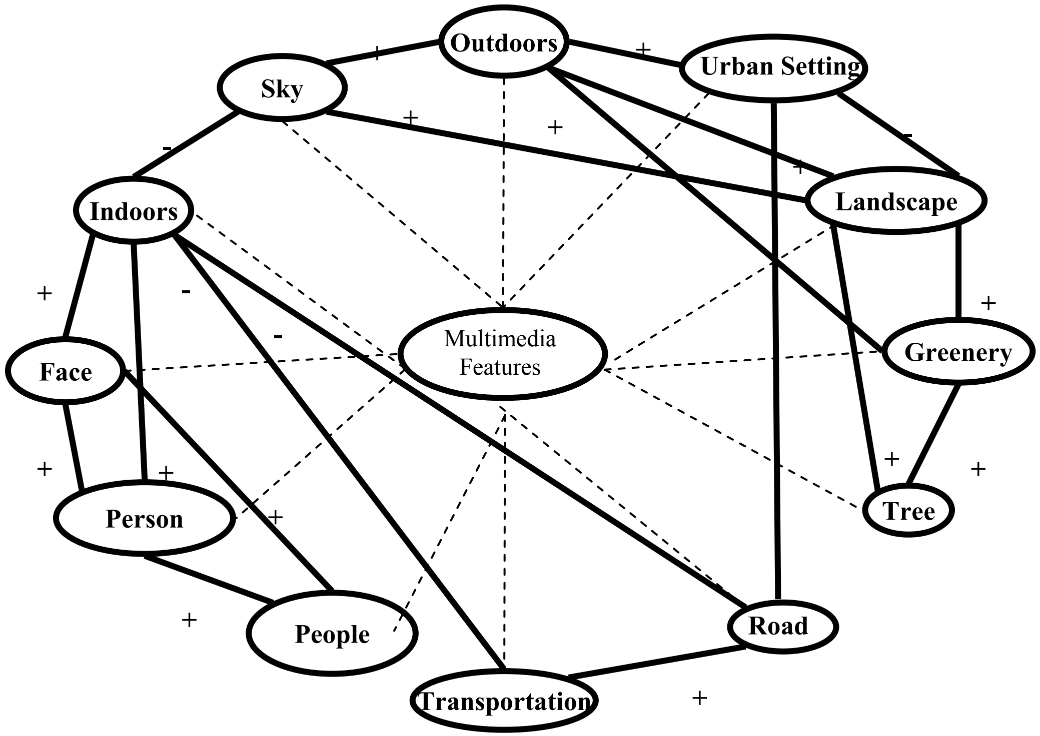
- **Problem:**

  Learn and Utilize Spatial, Temporal and Conceptual Context
- **Approaches**:

  Multinet: Network of Multijects or Concept Models represented as a graph with undirected edges. Use of probabilistic graphical models to encode and enforce context.

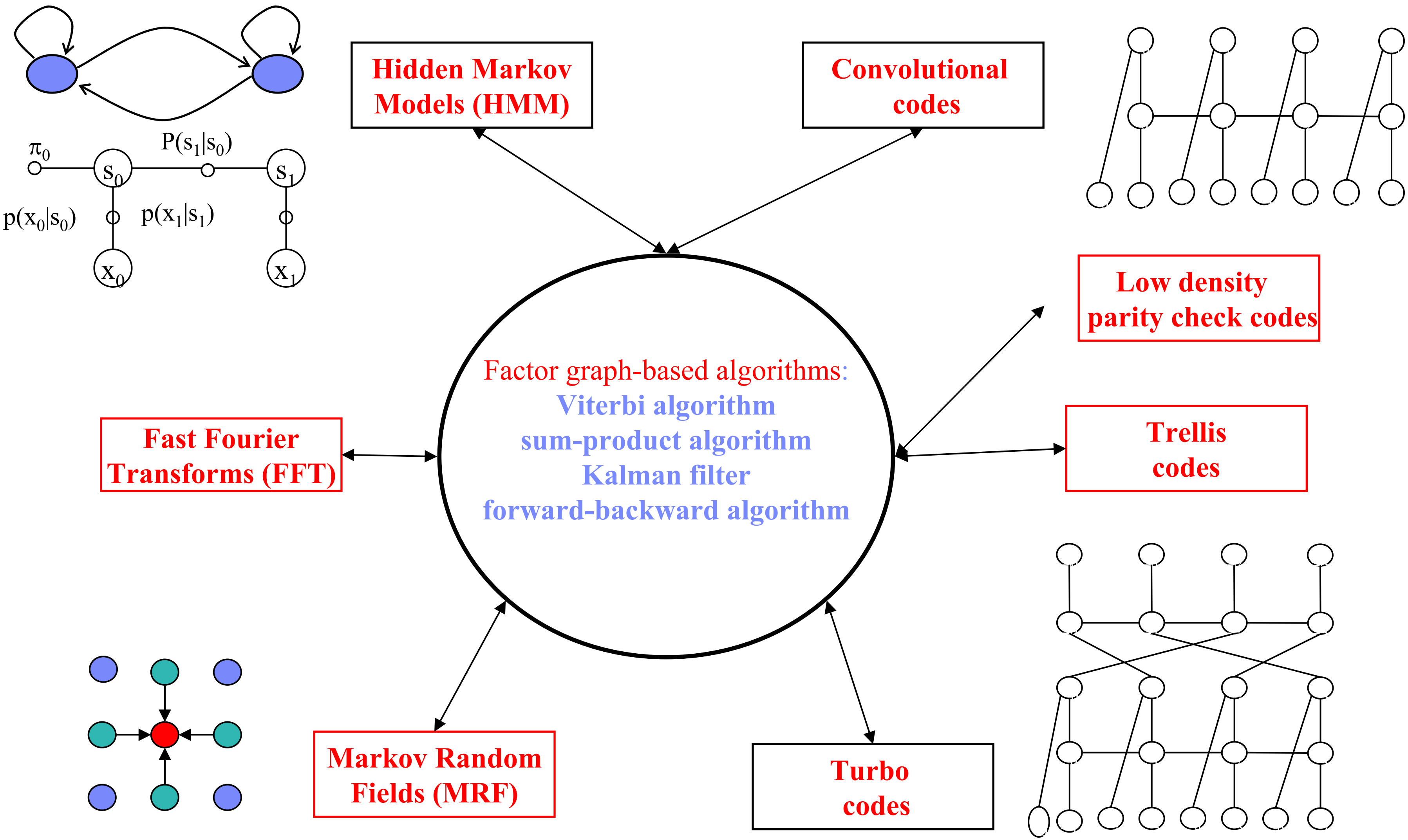
  Hierarchical Classification: Use baseline models' concept detection confidences as features and train another layer of classifiers.
- **Result**:

  Factor-graph multinet with Markov chain temporal models reduced error rates by more than 27 % .



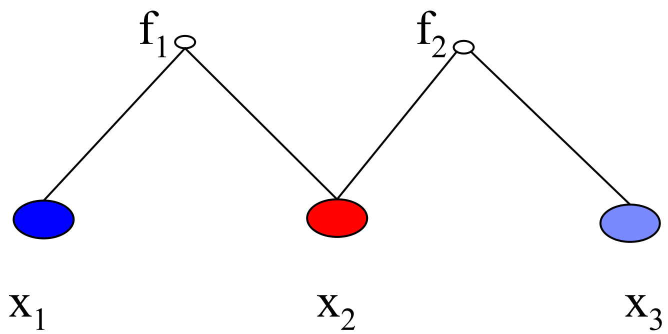
Multinet: Modeling the interaction between semantic concepts using a probabilistic graphical network of multijects (Naphade et al ICIP 98, Naphade et al NIPS 00, Naphade et al, T. CSVT 2002

Factor Graphs: A Glimpse



Factor Graphs: Notation

$f(x_1,x_2,x_3)=f_1(x_1,x_2)*f_2(x_2,x_3)$



2 types of nodes:

Function nodes (f_1, f_2)

Variable nodes (x_1, x_2, x_3)

$$f(x) = \prod_{i=1}^{m} f_i(x^{(i)})$$

where $x^{(i)} \subset x$

is the set of variables of local function $f_i(x^{(i)})$

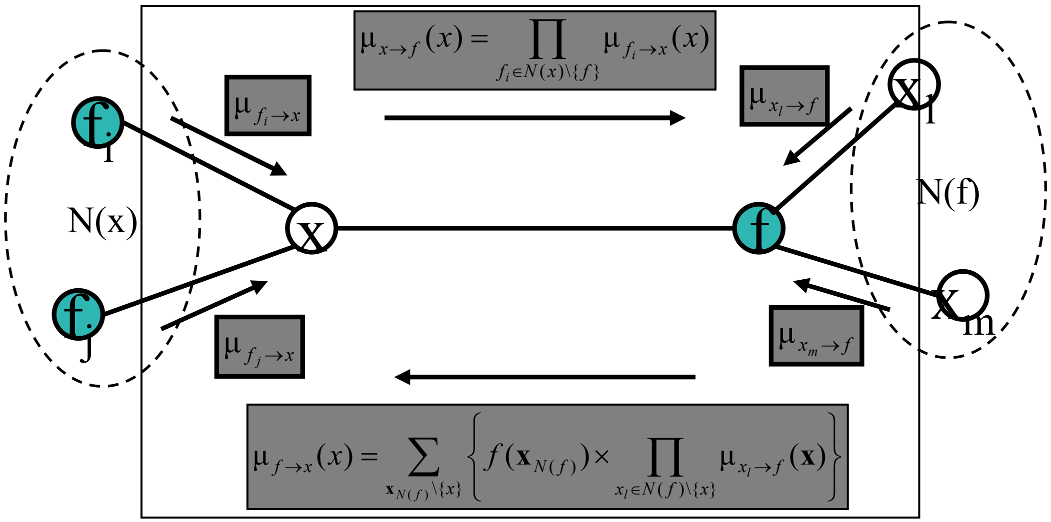
A function node is connected only to those variable nodes, which are its arguments.

Why a Factor Graph for the Multinet?

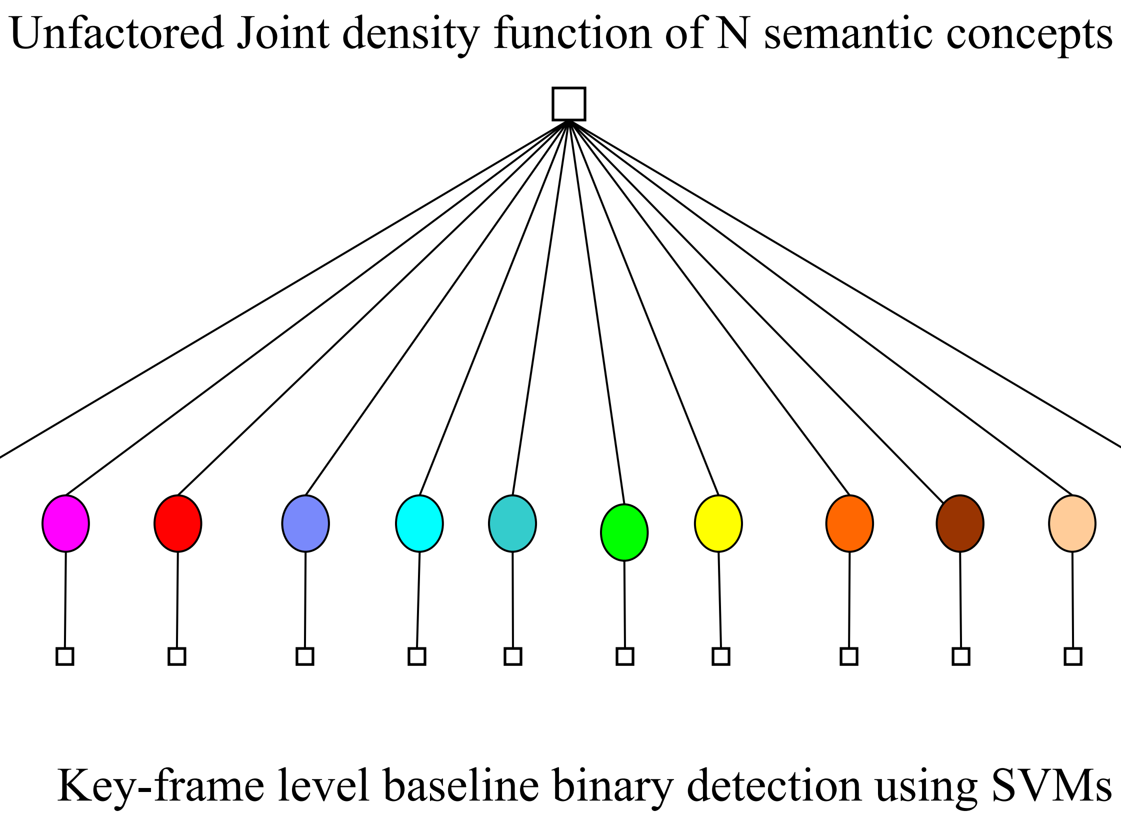
- No causality assumptions are necessary in FG
- Cycles are allowed and graphs are undirected
- Semantics may not adhere to the causality assumptions
- The multinet is bound to have cycles and loops due to complex inter-conceptual relations.
- When Factor Graph is Tree, exact inference possible with the sum-product message passing algorithm.
- When Factor Graph is not a Tree, loopy propagation leads to approximate inference.

<u>Variable Node -> Function Node:</u>
Product of all messages coming in to variable node from other function nodes connected to it.

<u>Function Node -> Variable Node:</u>
Product of all messages coming in to function node with the local function itself, marginalized for the variable associated with the variable node

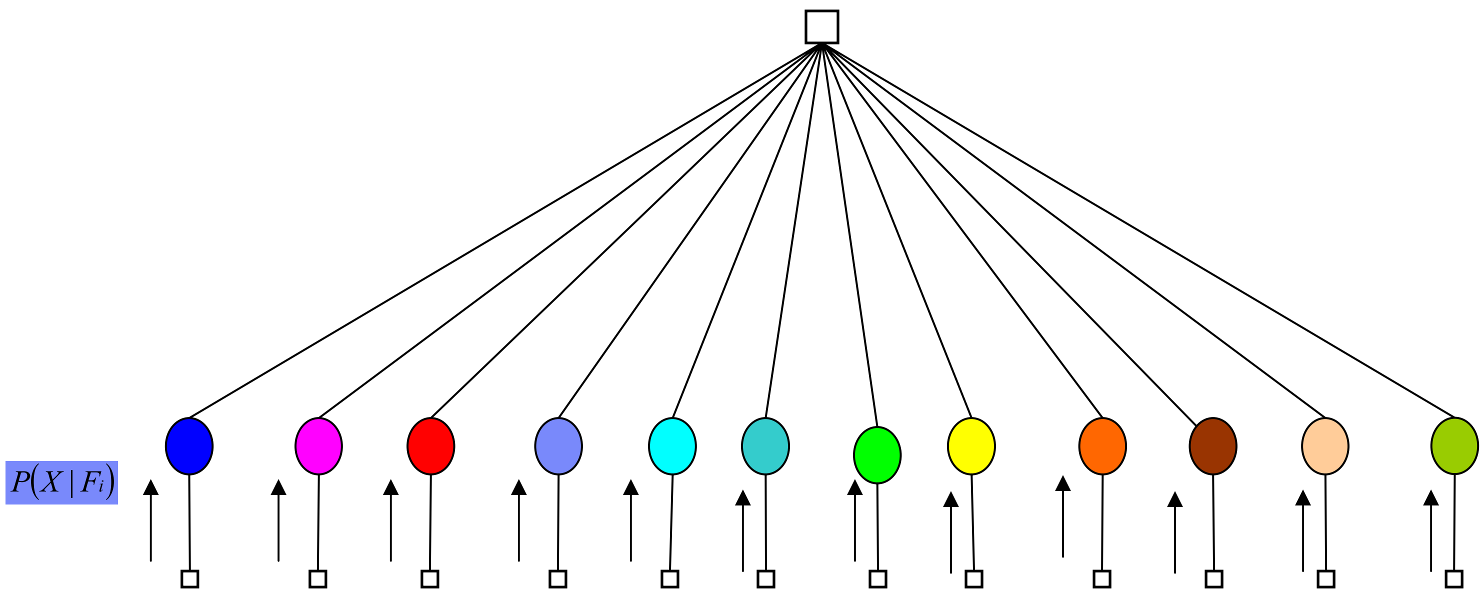


Learning and Using the factor graph: Unfactored Global Distribution



Message Passing
From frame-level classifiers to variables

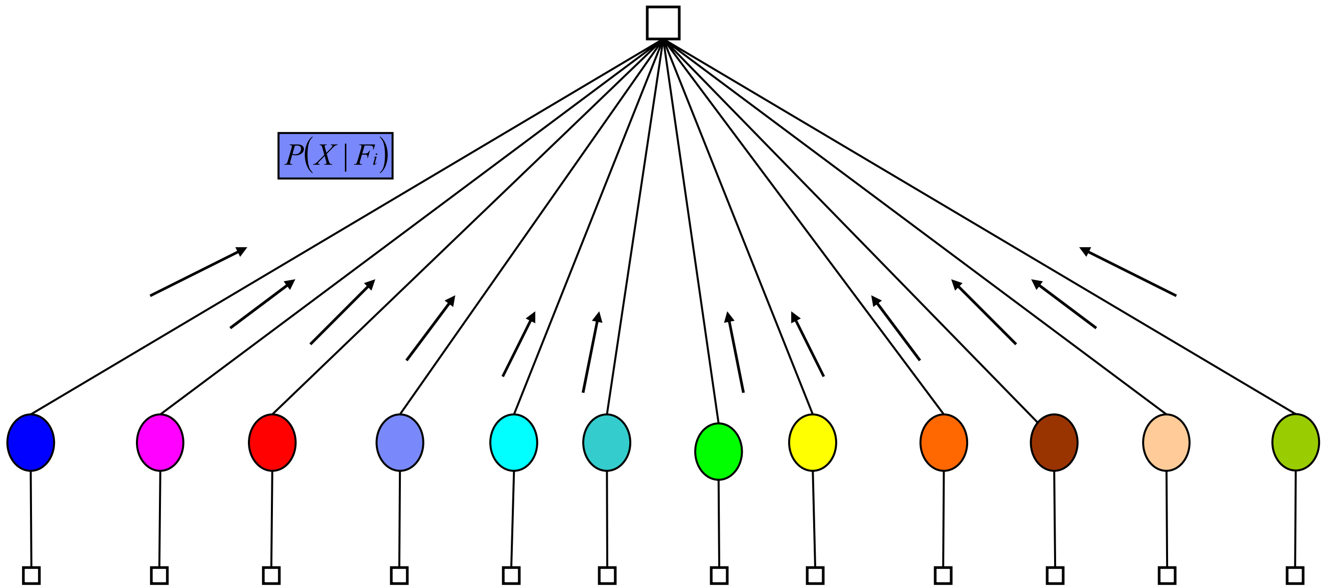
Unfactored Joint density function of N semantic concepts



Key-frame level baseline binary detection using SVMs

Message Passing
From variables to global function

Unfactored Joint density function of N semantic concepts

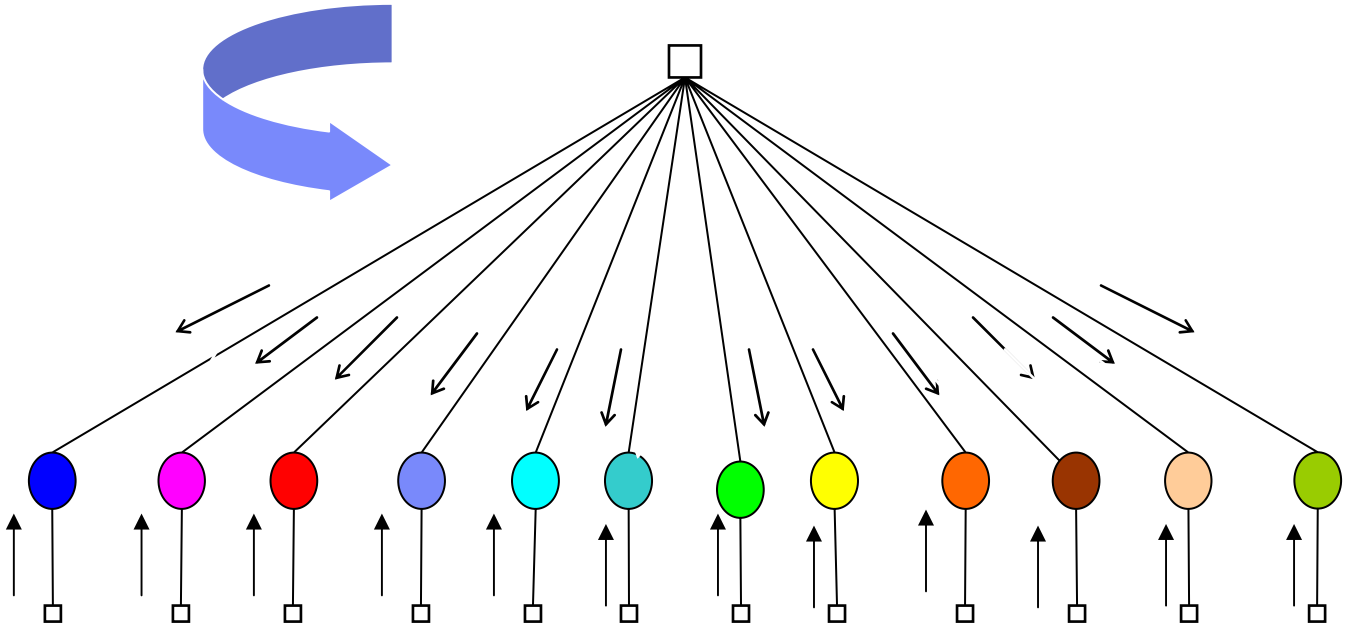


Key-frame level baseline binary detection using SVMs

Message Passing: Global function to variables

Unfactored Joint density function of N semantic concepts

$$P_w(X \mid F_1 = v_1) = \sum_{v_2}\sum_{v_3}\sum_{v_4}\dots\sum_{v_{12}}\prod_{j=2}^{12} P(X \mid F_j = v_j)P(F_1 = v_1, F_2 = v_2, ..., F_{12} = v_{12})$$

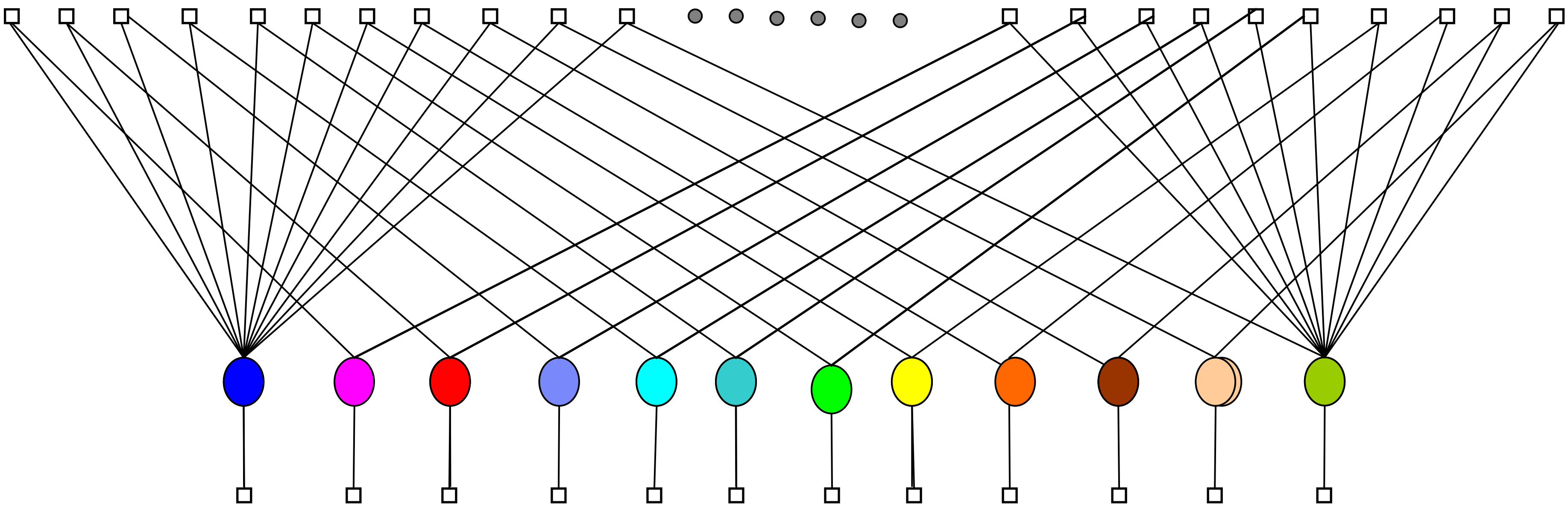


$P(X \mid F_1 = v_1)$

Key-frame level baseline binary detection using SVMs

Factoring the Global Function

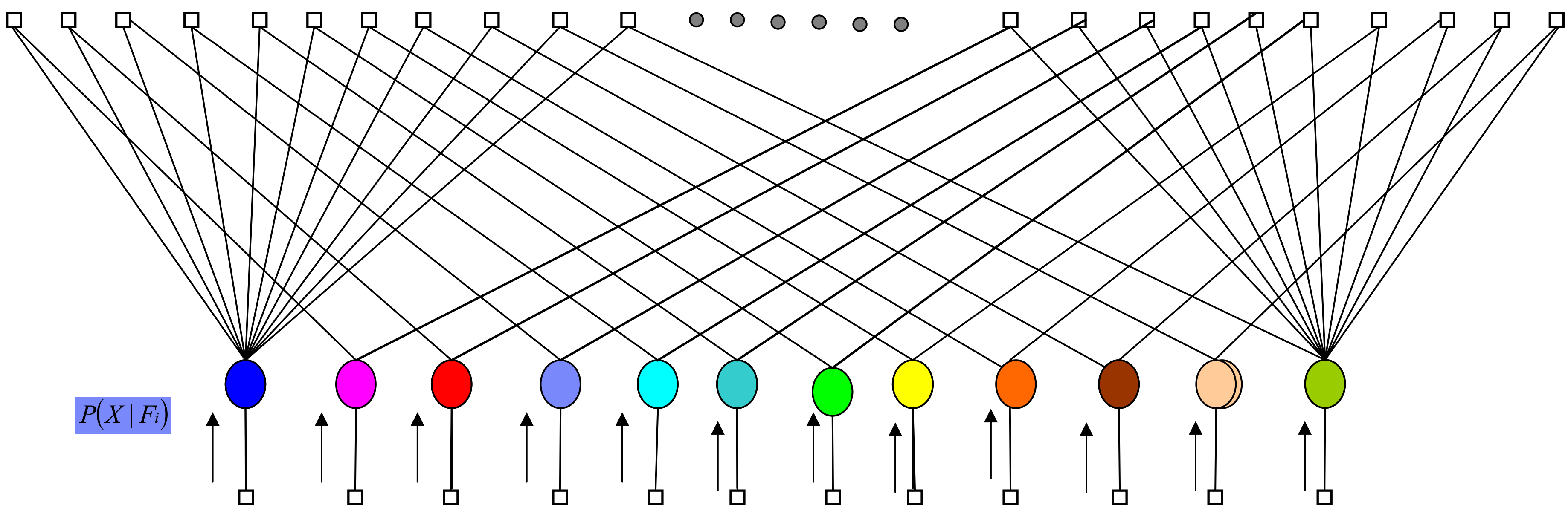
Factored joint density function of N (N=12) semantic concepts



Key-frame level baseline binary detection using SVMs

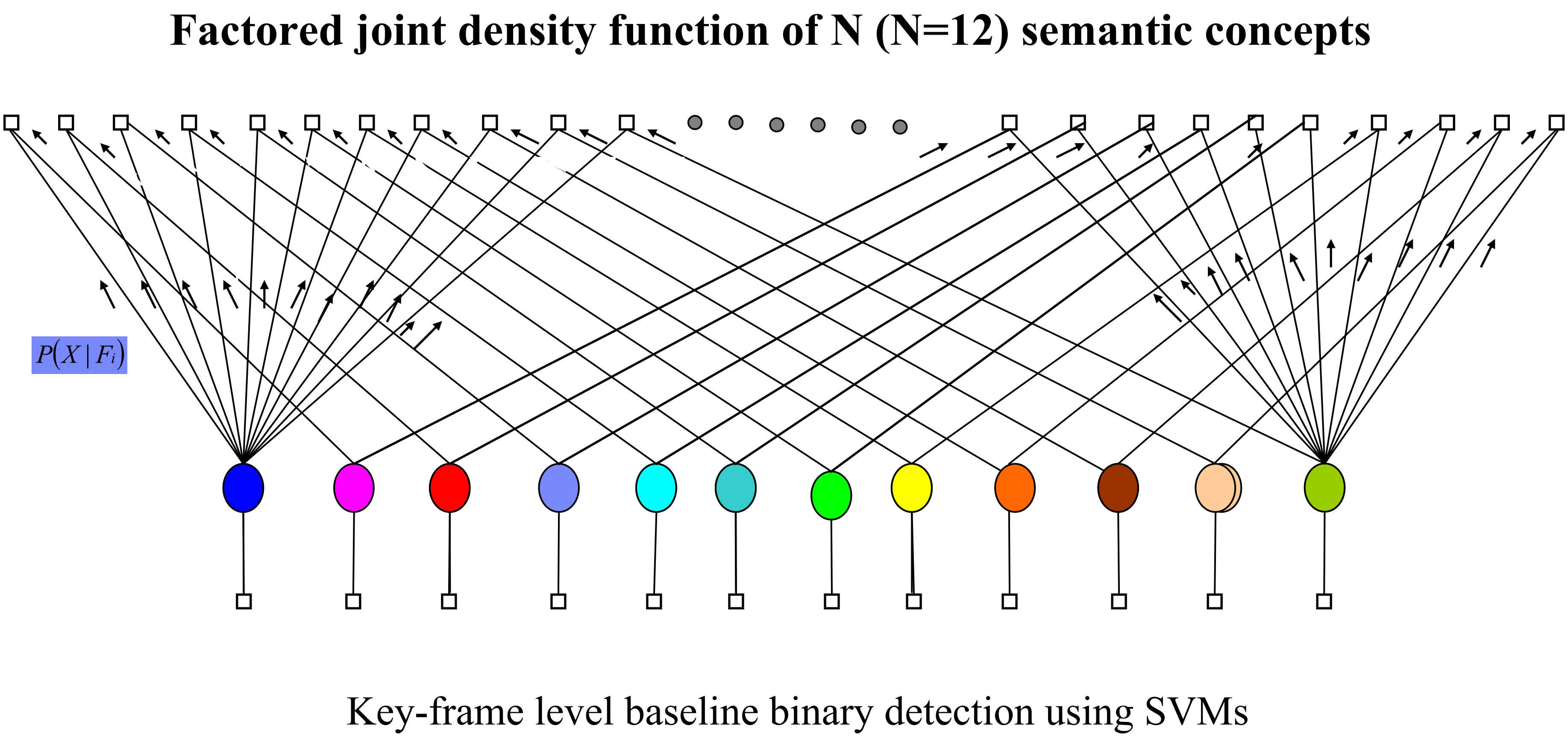
Factored Global Function

Factored joint density function of N (N=12) semantic concepts



Key-frame level baseline binary detection using SVMs

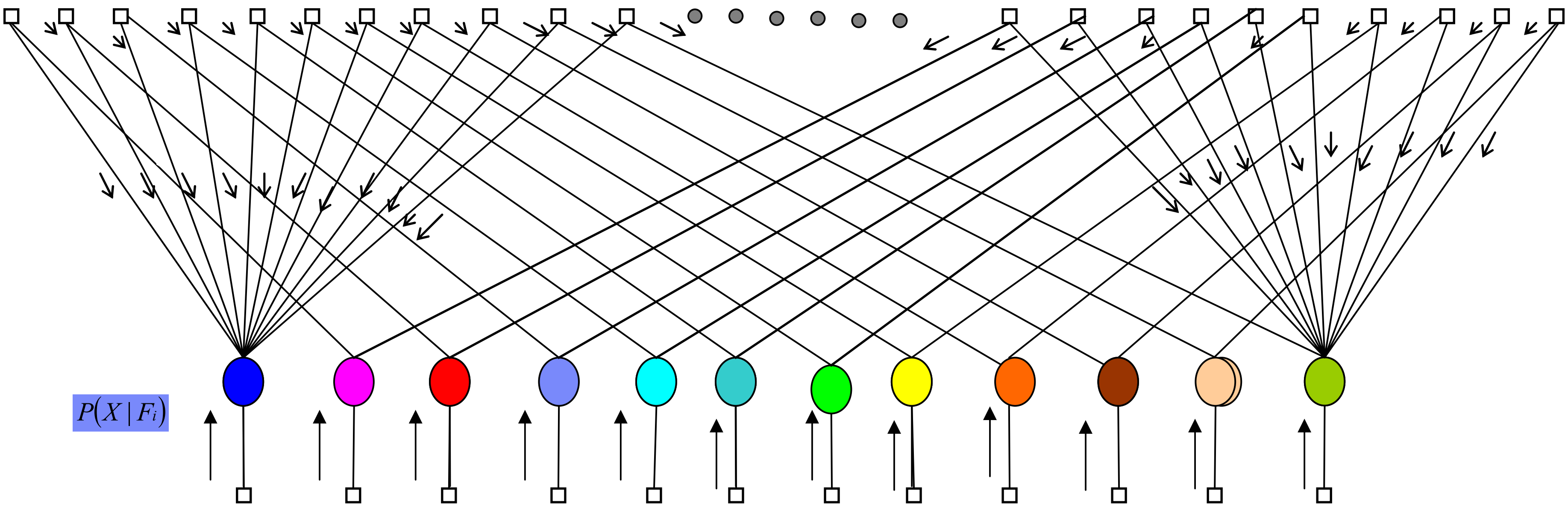
Factored Global Function



Factored Global Function

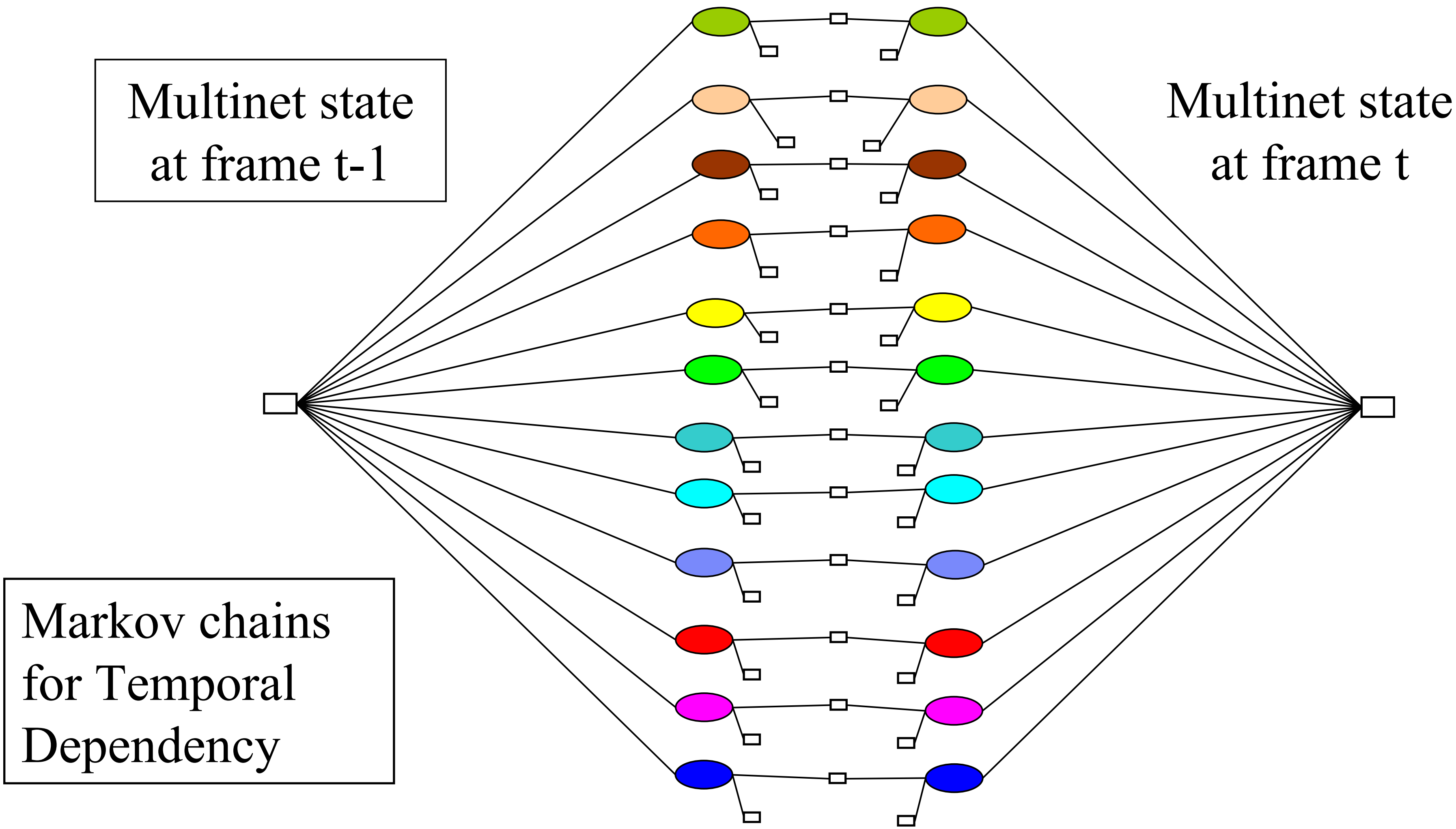
$$P_w(X \mid F_k = v_k) = \sum_{v_l} P(X \mid F_l = v_l) P(F_k = v_k, F_l = v_l)$$

Factored joint density function of N (N=12) semantic concepts

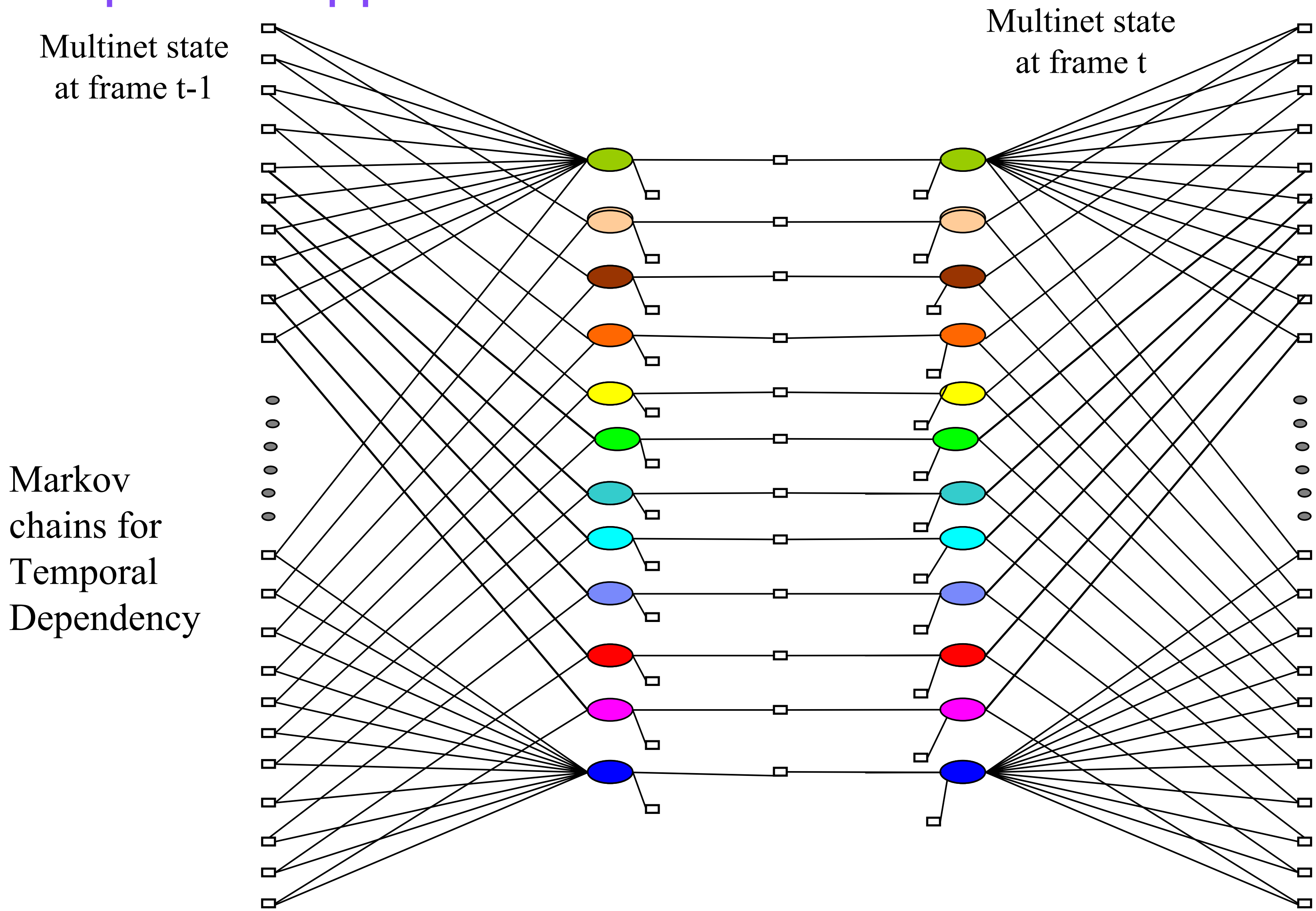


Key-frame level baseline binary detection using SVMs

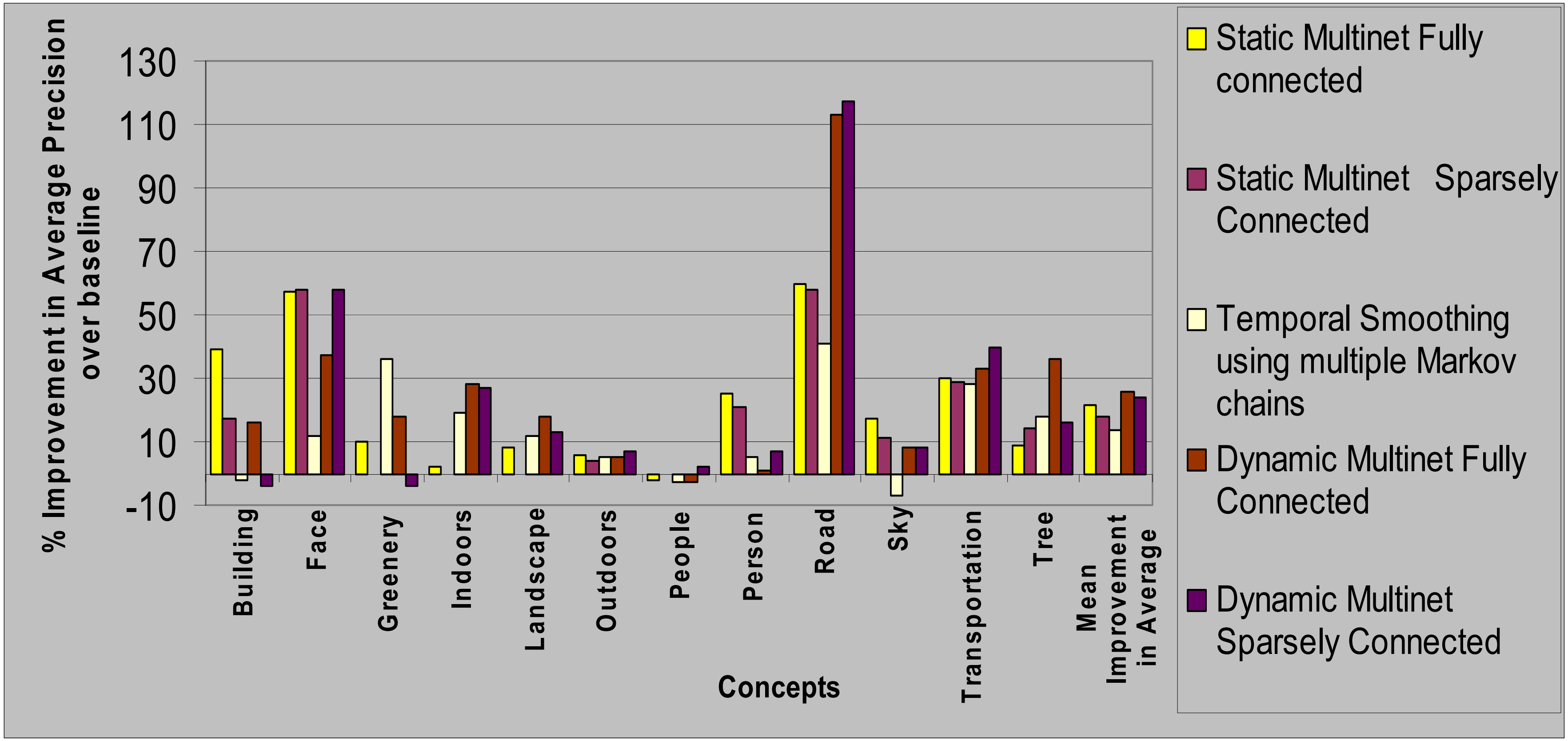
Temporal support + global function



Temporal Support + Factored Global Function

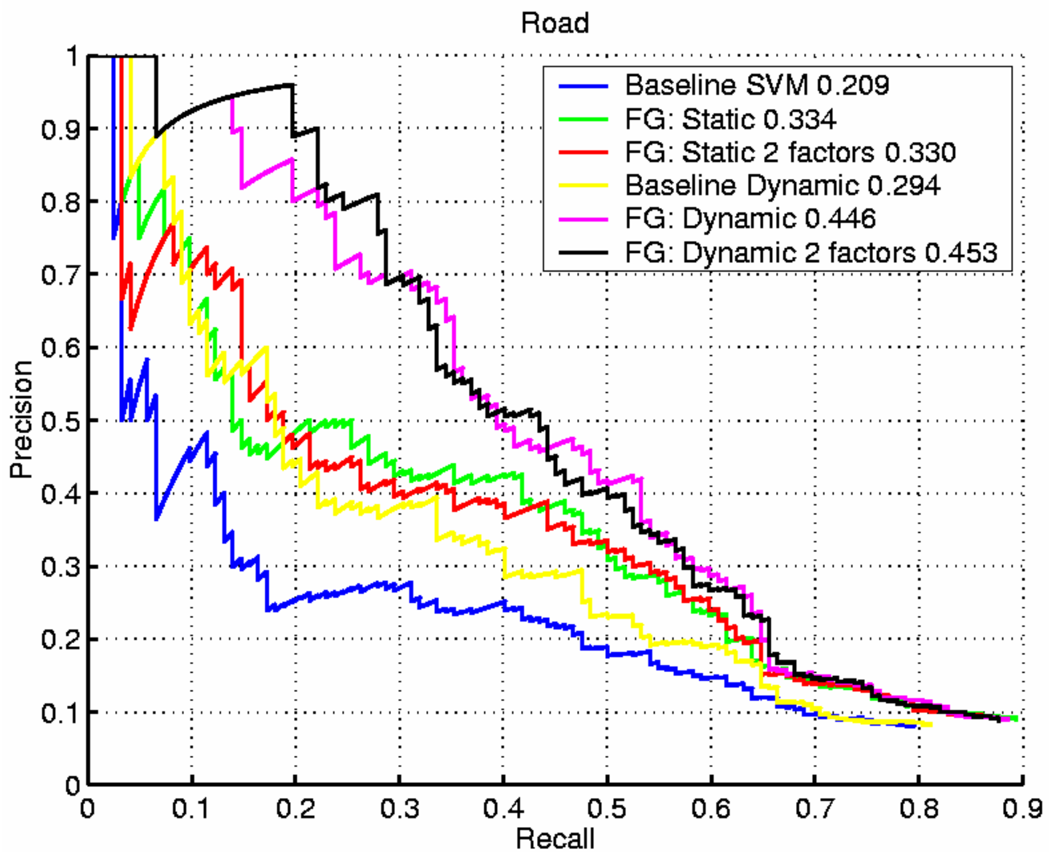


Improvement due to Context Modeling



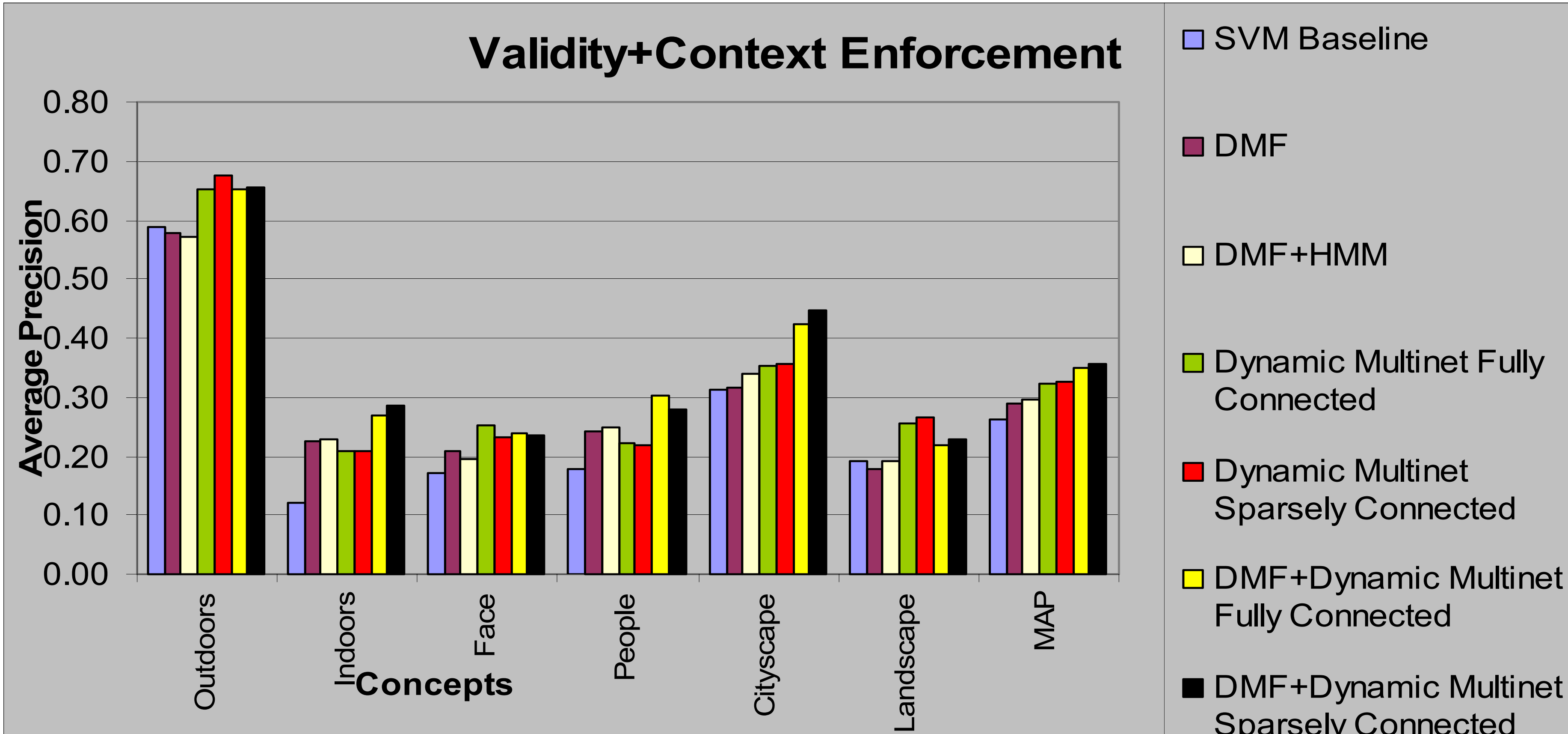
- Mean improvement in average precision by Modeling Conceptual Context: 21 %,
- Mean improvement in average precision by Modeling Temporal Context: 13 %
- Mean improvement in average precision by Modeling Conceptual & Temporal Context: 26 %

Precision Recall Curves: Road (Validation Set)



Temporal modeling: 41 %
Static multinet: 60 %
Static Factored multinet: 58 %

Multinet Improves Even Results that have been post-processed to improve detection by other methods



DMF+DFMN Improvement over baseline 37 %

DMF Improvement over baseline 12 %
DFMN Improvement over baseline 27 %
DFMN Improvement over DMF 24 %

Learning Multimedia Semantics

A. Supervised Detection
   - Static Classifiers
   - Spatial+Temporal Classifiers
B. Multimodal Fusion
   - Late fusion using Ensembles
   - Intermediate Fusion for temporal evolution using graphical models
C. Enforcing Spatial, Temporal and Conceptual Context
   - Learning Context using Multinet
D. **Semi-Supervised Learning**
   - **Labeled+Unlabeled Learning**
   - **Active Learning**
   - **Multiple Instance Learning**
   - **Co-training**
E. Unsupervised Clustering
   - Spatial
   - Spatio-temporal using hierarchical HMMs
F. Semantic Feature Extraction and Search
   - Query Learning
   - Leveraging detected semantic concepts for complex query answering

Partial Supervision and Unsupervised Approaches

- **Problem:**

  Using the inherent clusters in data space, semantic space and the relationships between different samples and concepts to reduce the amount of user supervision needed to learn concept models.

- **Approaches**:

  Labeled+Unlabeled Learning

  Active Learning

  Multiple Instance Learning

  Co-training

Using Labeled *AND* Unlabeled Examples

Using Unlabeled Examples

- Seems counter-intuitive
- What can unlabeled examples tell us ?
- How can we use unlabeled examples ?
- Suppose we did use the unlabeled examples in some way
- Can we guarantee improvement in performance ?
- If so under what conditions will there be no loss in performance ?

Hypothesis

- If labeled and unlabeled samples contradict each other strongly, there is no guarantee that performance will not degrade
- If labeled and unlabeled samples are in harmony what is the need of using unlabeled samples ?

  Refining estimation

  Performance will not degrade in general

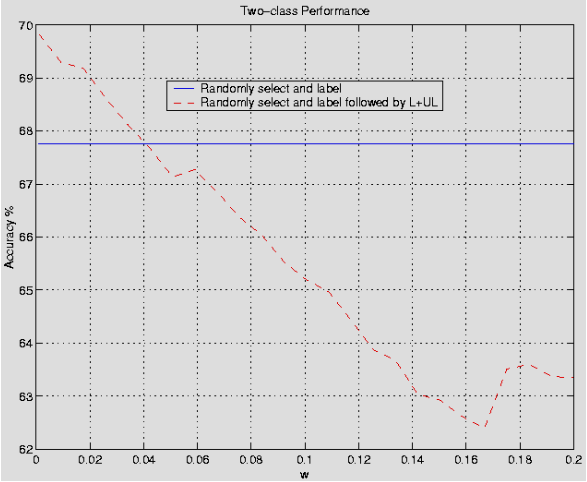
  No harm in using the unlabeled samples which come at no extra cost

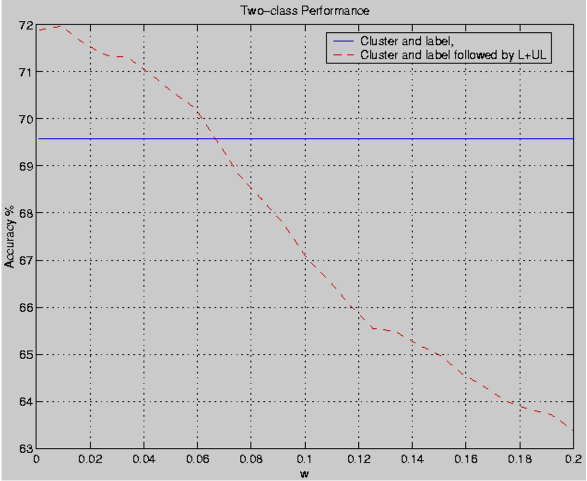
Prior Art

- Shahshahani and Landgerbe (IEEE T. Geoscience & Remote Sensing '94): "Effects of unlabeled samples in small sample size problem and mitigating the Hughes phenomenon."
- Nigam, McCallum, Thrun and Mitchell "Text Classification …" (Machine Learning '99). Extension of Shahshahani's work to mass functions instead of continuous densities.
- In all these cases the "EM" algorithm forms the basis of the classification algorithms.

Strategy for Enforcing Consistency (Naphade et al Photonics East 2000)

- Algorithm
  - Begin with a completely unlabeled set
  - Unsupervised Clustering of unlabeled samples into as many clusters as the number of examples to be labeled
  - Prompt user to provide the label for one sample from each cluster
- Observations
  - Local consistency is necessary for global consistency so intra-cluster consistency is more likely than global consistency
  - The weighting of the unlabeled samples w.r.t to the labeled samples plays an important role in performance.
  - Figures on right show accuracy of classification on a test set using 500 training samples for the concept “Sky”. Figure on top shows performance with random selection of 500 samples for annotation. Figure at the bottom shows K-means clustering used to select 500 samples for annotation.
  - Clustering as a pre-processing step for sample selection results in better performance unless the dataset is uniformly randomly distributed.
  - Using unlabeled data along with labeled samples always helps over using only labeled samples as long as the relative weight to the two sets is well controlled.





Active Learning Sample Selection for Media Annotation

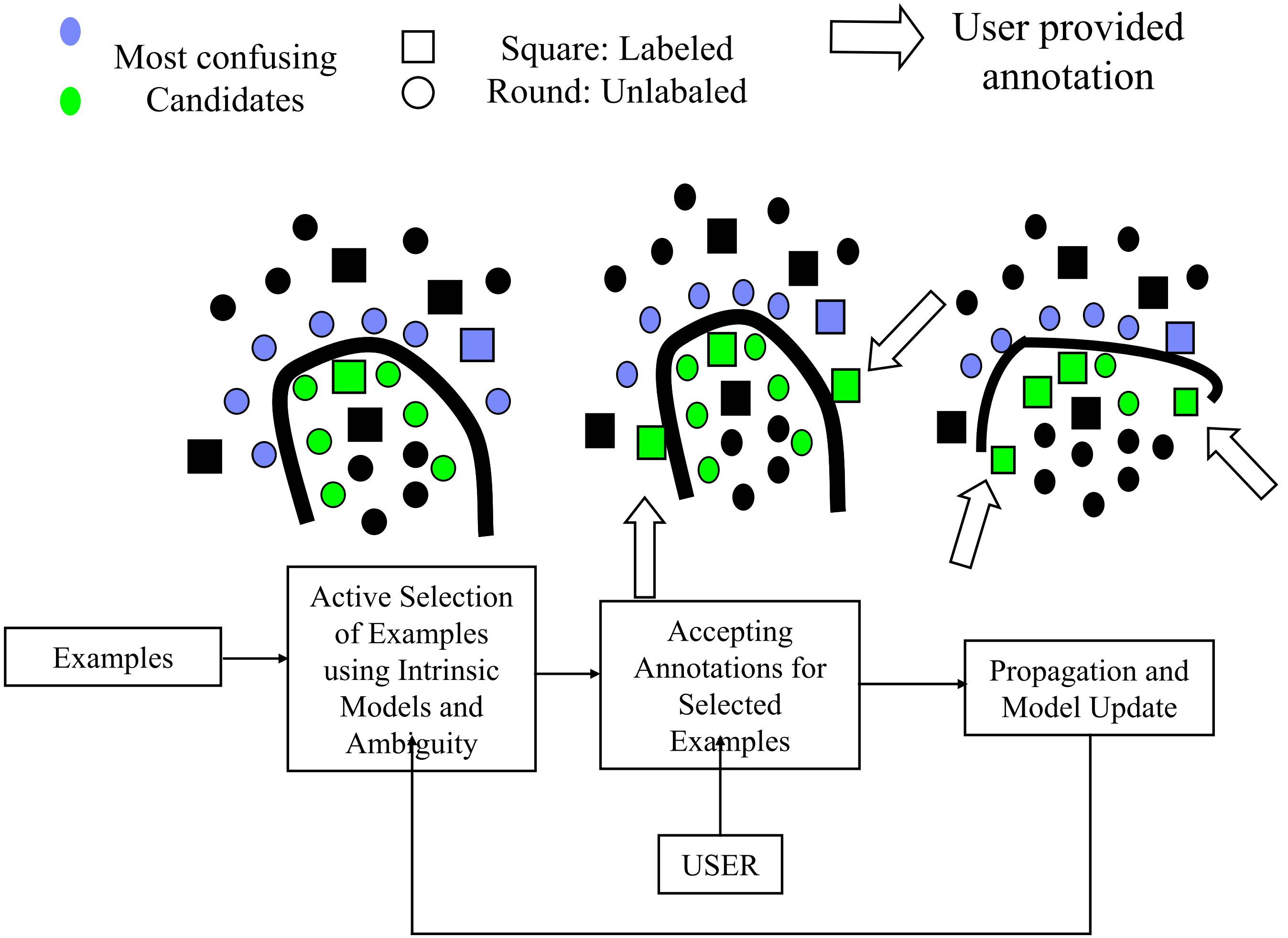
STRATEGY:

Instead of waiting passively for user to annotate, help user by selecting the most difficult examples to be annotated.

RESULT:

By learning how to resolve conflict in the case of difficult examples:

- Reduce the number of examples (and annotation time) that need to be manually annotated by orders of magnitude.
- Automatically pass on annotation to the remaining samples that are easier to annotate.



Evaluation: Performance does not drop despite dropping 90 % training samples!

- **Setup**

  In each annotation experiment a warp-up set with as many as 1 % of the total number of examples to be annotated was assumed to be labeled.

  Beyond this continue to annotate up to 10 % of the total number of examples using the above different approaches

  The aim is to investigate how many examples need annotation before the rest can be automatically annotated

  Tried 3 schemes of sample selection using an SVM-based active learner with the distance from the hyperplane as a measure of ambiguity.

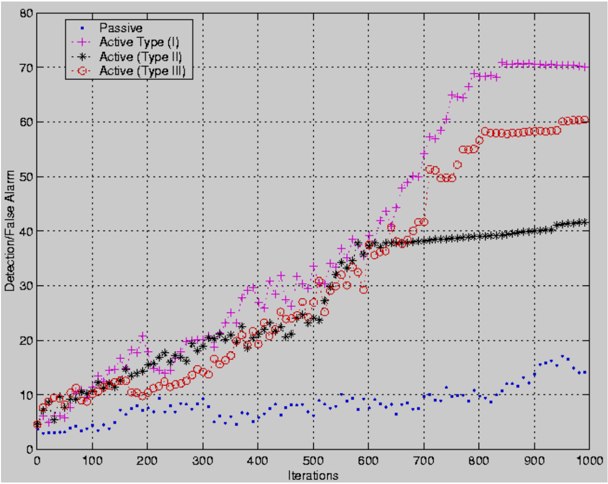
  Chose “Outdoor” to test the algorithm.

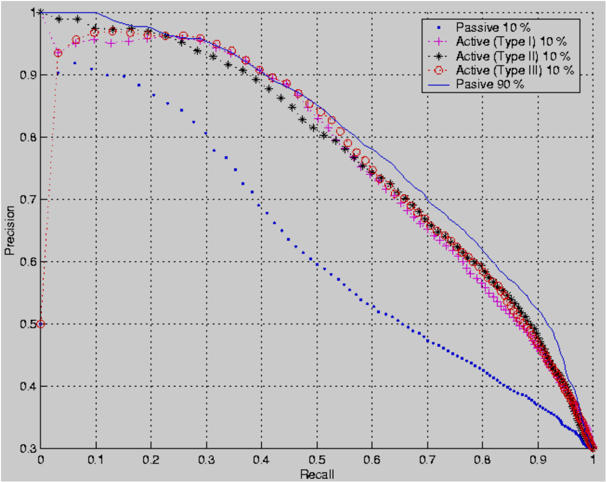
- **Observations**

  A ratio of detection/false alarms indicates that most of the information can be captured by actively selecting up to 10% of the total number of examples.

  Law of diminishing returns ? Improvement starts diminishing beyond 10% of the total number of examples.

  Same Performance With 90% less annotations needed





Cross-Granular Disambiguation using Multiple Instance Learning

- **Problem:**

  Supervision is extremely expensive especially for regional concepts

  Improving regional ground truth by accepting coarse labels can be in general beneficial to any conventional learning algorithm

FACE

FACE

NO FACE

REGIONAL ANNOTATION

Cross-Granular Disambiguation using Multiple Instance Learning

Problem:

- Supervision is extremely expensive especially for regional concepts
- Improving regional ground truth by accepting coarse labels can be in general beneficial to any conventional learning algorithm

FACE

FACE

NO FACE

Approach: Allow users to supervise at coarse granularity and learn the implicit coarse to fine granularity mapping

GLOBAL ANNOTATION

Cross-Granular Disambiguation using Multiple Instance Learning

Problem:

- Supervision is extremely expensive especially for regional concepts
- Improving regional ground truth by accepting coarse labels can be in general beneficial to any conventional learning algorithm

FACE

FACE

MODEL

NO FACE

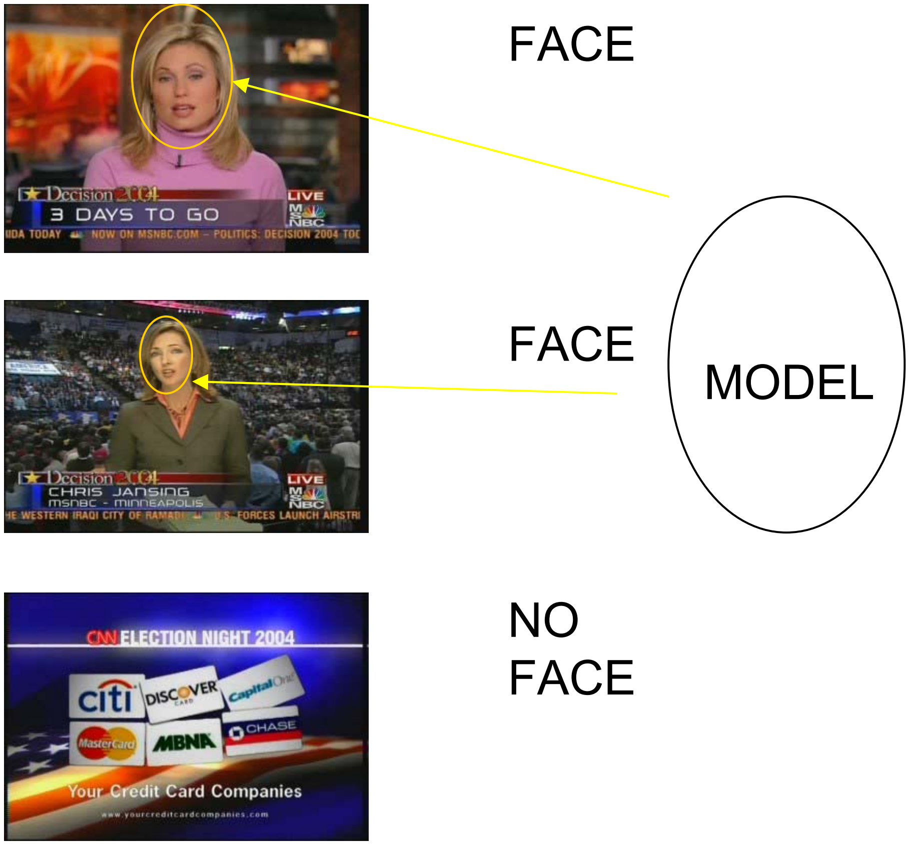
Approach: Allow users to supervise at coarse granularity and learn the implicit coarse to fine granularity mapping

LEARN from GLOBAL ANNOTION

Cross-Granular Disambiguation using Multiple Instance Learning

Problem:

- Supervision is extremely expensive especially for regional concepts
- Improving regional ground truth by accepting coarse labels can be in general beneficial to any conventional learning algorithm



Approach: Allow users to supervise at coarse granularity and learn the implicit coarse to fine granularity mapping

THEN APPLY MODEL TO DERIVE REGIONAL ANNOTATION

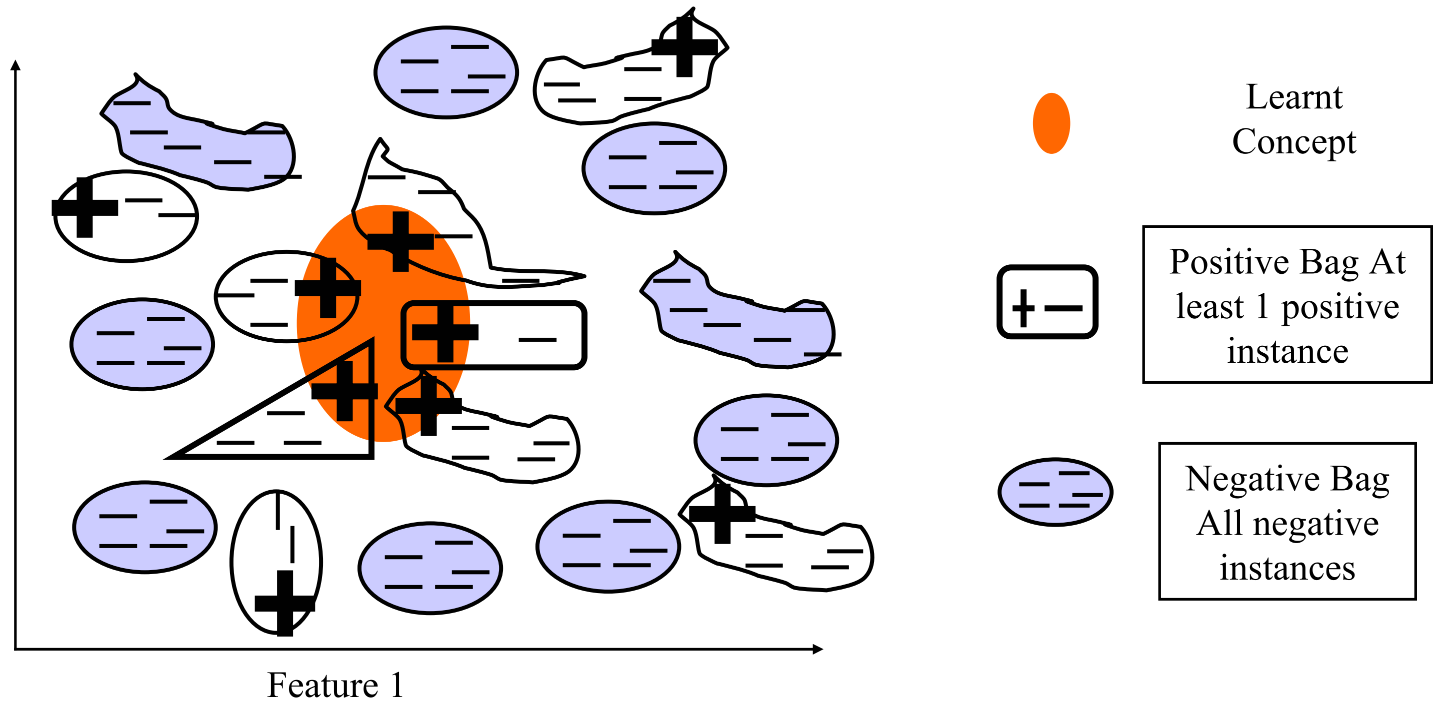
Multiple Instance Learning for Granularity Resolution

PROBLEM

Ask user only for <u>coarse annotations</u>
Resolve ambiguity in propagating annotations from coarse-to-fine using discriminant learning algorithms

RESULT:

Strategy <u>propagates</u> annotations from coarser granularity <u>to finer granularity</u> with excellent accuracy.
Strategy <u>reduces annotation time significantly</u>.



Bag=Image; Instance=Region in Image.
Bag= Shot; Instance=Tracked Region
Bag= Video Clip; Instance=Tracked Regions

State of the Art: MIL for Image Annotation

- Diverse Density.

  Idea: How many positive bags and how far from negative bags.

  Oded Maron Aparna Lakshmi Ratan, **Multiple-Instance Learning for Natural Scene Classification, Proceedings of the Fifteenth International Conference on Machine Learning, 1998.**

  Cheng Yang and Tomas Lozano-Perez, **Image Database Retrieval with Multiple-Instance Learning Techniques, Proceedings of the 16th International Conference on Data Engineering,2000.**

- Extension of SVM.

  Idea: Bag's margin in addition to instance's margin.

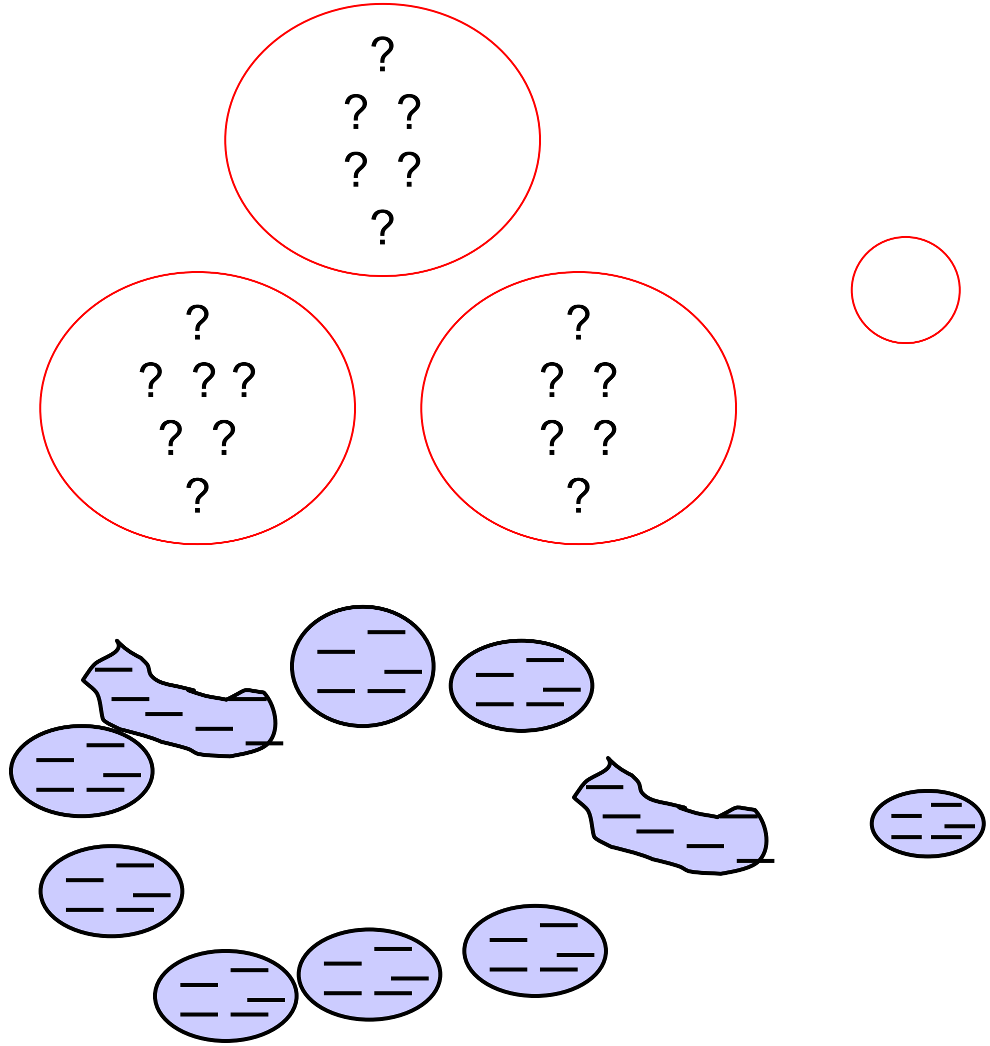
  Stuart Andrews, Ioannis Tsochantaridis and Thomas Hofmann, **Multiple instance learning with generalized support vector machines, Advances in Neural Information Processing Systems (NIPS), 2003.**

- General framework to pick the positive one .

  Idea: select the point far away from negative one as positive.

  Milind Naphade, John Smith, **A Generalized Multiple Instance Learning Algorithm for Large Scale Modeling of Multimedia Semantics, 2005 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005**

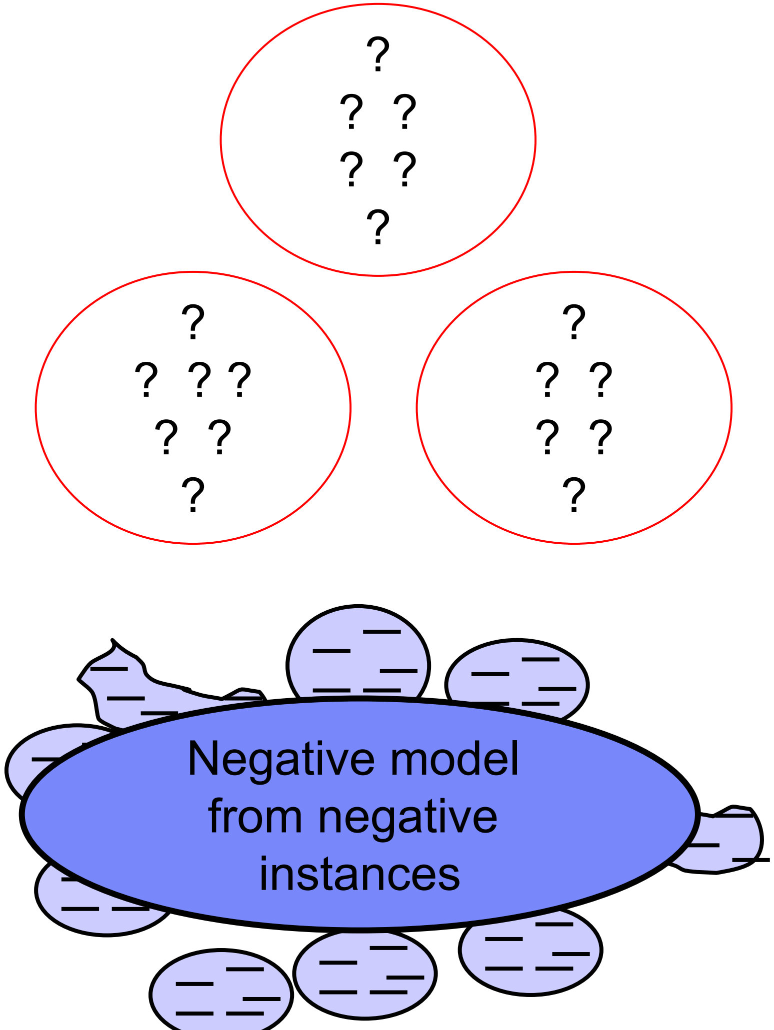
Extending Generalized Multiple Instance Learning with New Selection Strategies

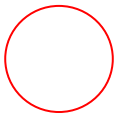


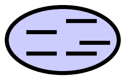
Positive bag. At least one of the instances is positive, but which one?

Negative Bag
All negative instances

Extending Generalized Multiple Instance Learning with New Selection Strategies



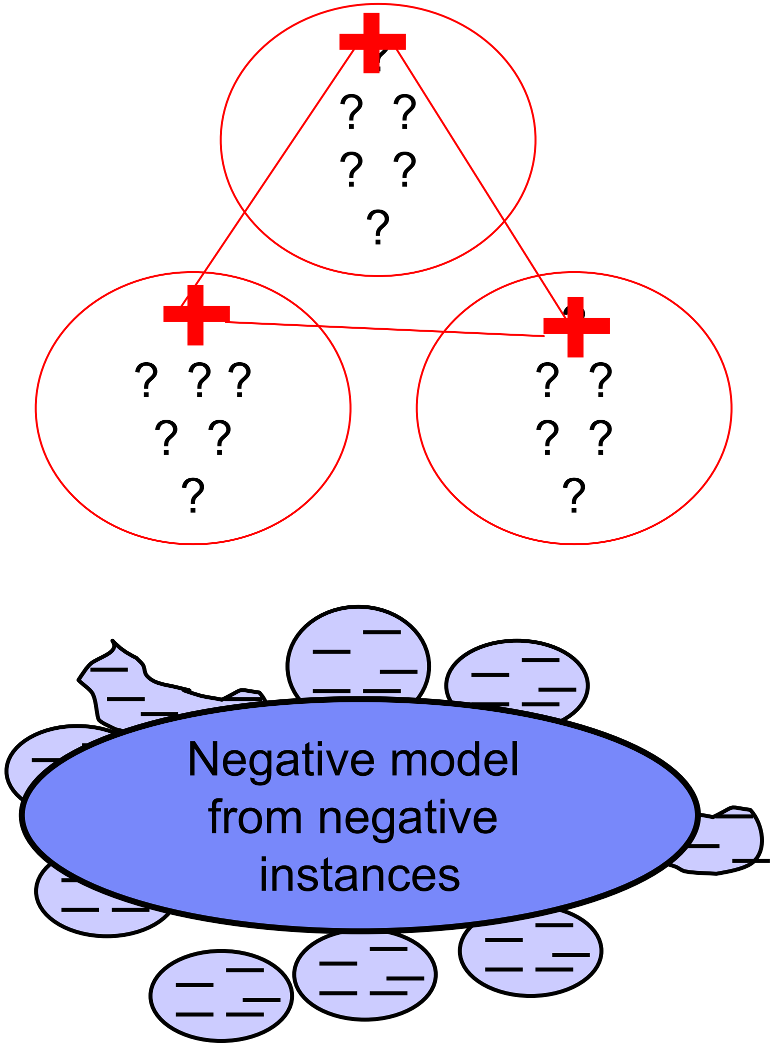


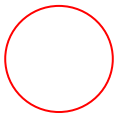


LEAST NEGATIVE SELECTION STRATEGY

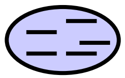
STRATEGY

Use negative model to rank instances in each positive bag and select the least likely negative instance as the most likely positive instance.





Positive bag. At least one of the instances is positive, but which one?

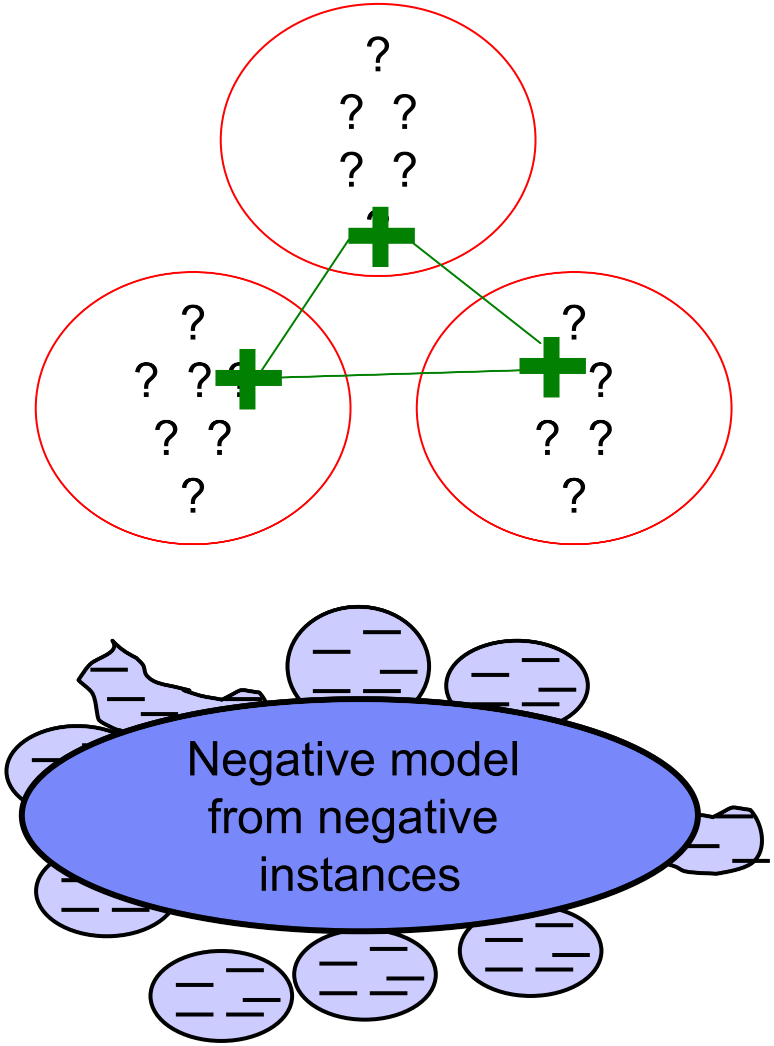


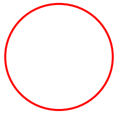
Negative Bag
All negative instances

MOST POSITIVE SELECTION STRATEGY

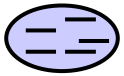
STRATEGY

Use all instances in positive bags to create a positive model and apply it to select the most positive instance from each positive bag





Positive bag. At least one of the instances is positive, but which one?



Negative Bag
All negative instances

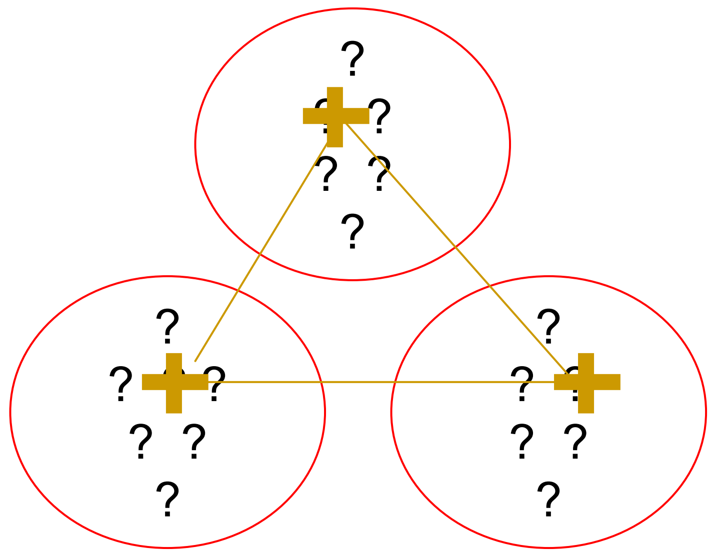
LIKELIHOOD RATIO SELECTION STRATEGY

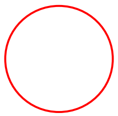
STRATEGY

Use all instances in positive bags to create a positive model

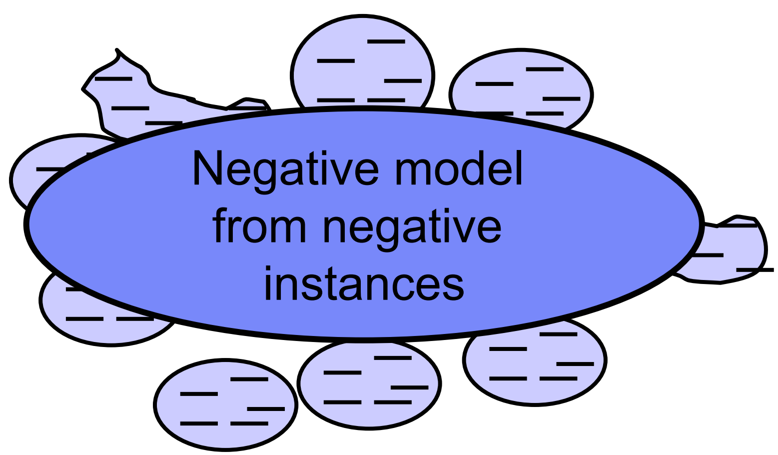
Use all instances in negative bags to create a negative model

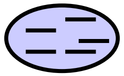
Use likelihood ratio to select most likely positive instance which is also least likely negative instance





Positive bag. At least one of the instances is positive, but which one?



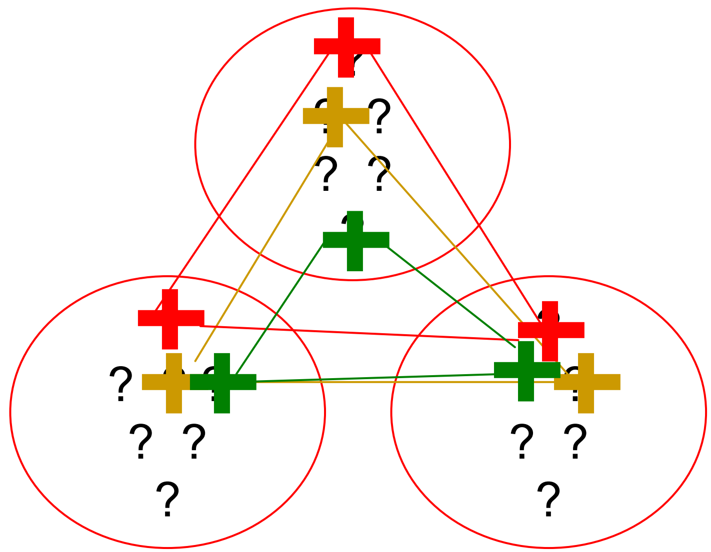


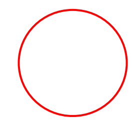
Negative Bag
All negative instances

Comparing LIKELIHOOD RATIO SELECTION STRATEGY

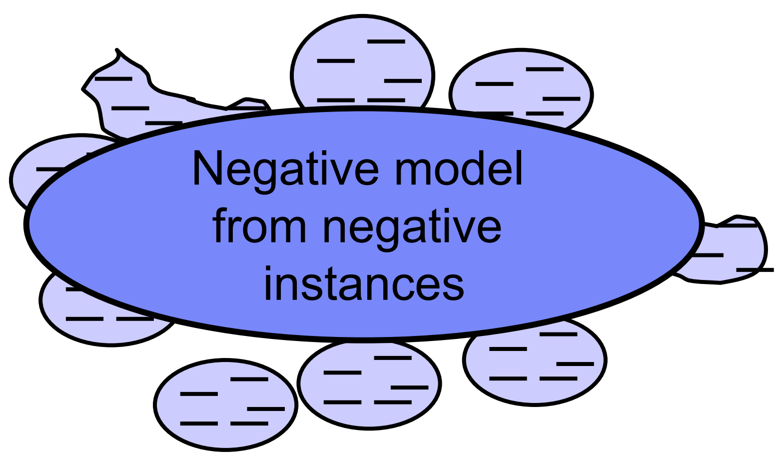
FUSION STRATEGY

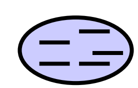
Use all three selection strategies and perform late fusion across the three resulting models





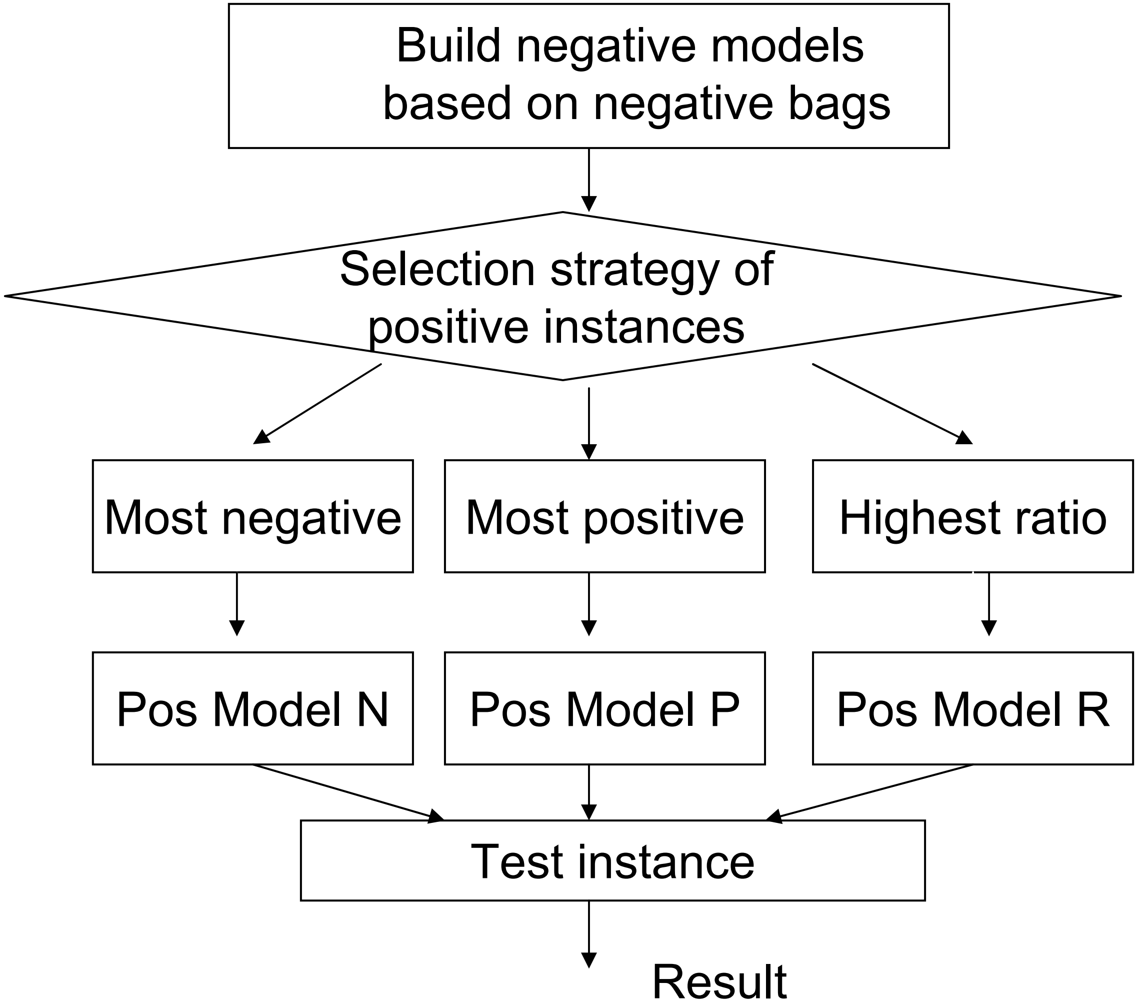
Positive bag. At least one of the instances is positive, but which one?





Negative Bag All negative instances

Overall Algorithm



Experimental Results on TRECVID corpus

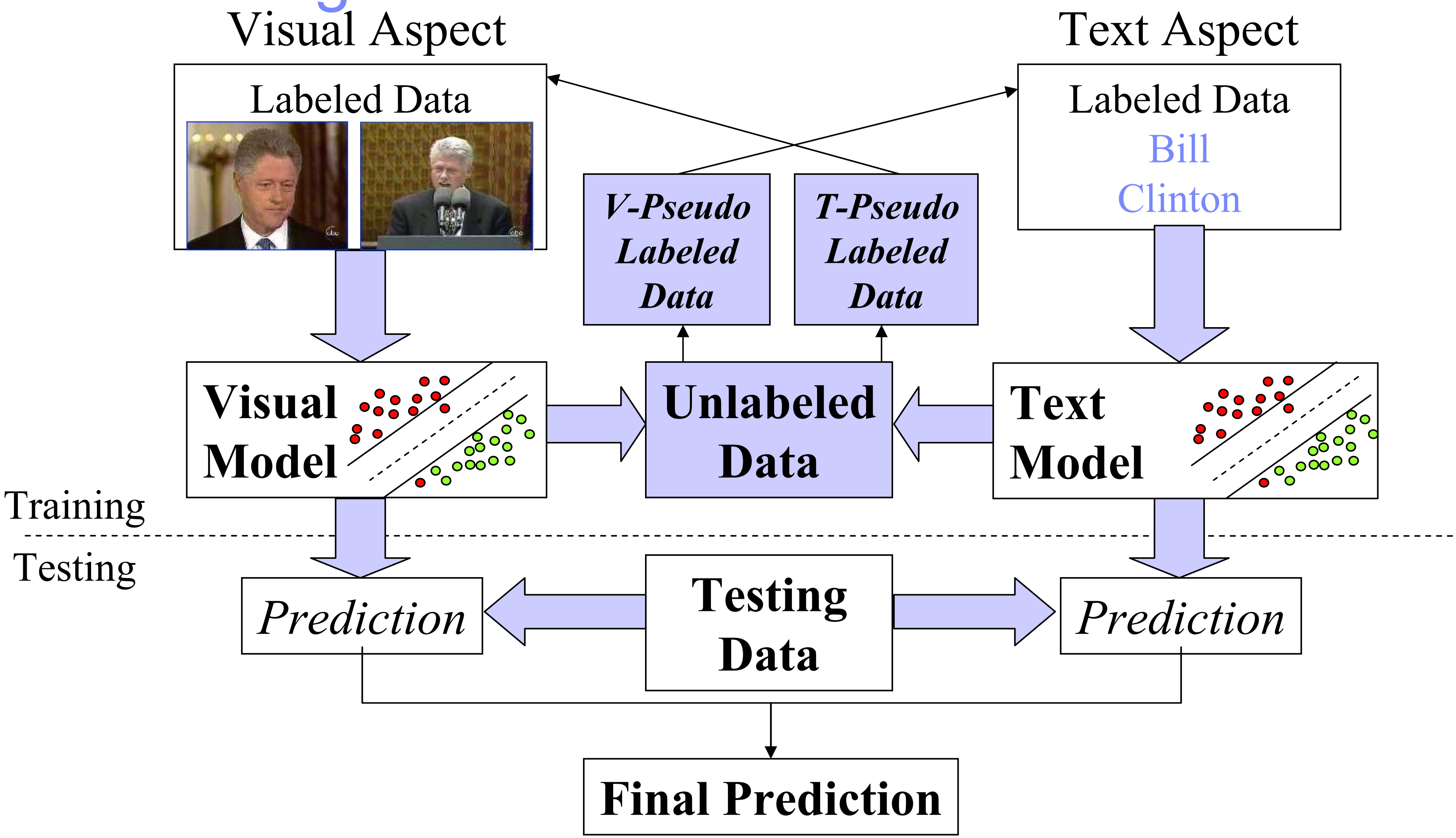
Mean average precision for five concepts

	Road	Sky	Building	Person	MAP
pos (p)	0.109	0.499	0.095	0.11	0.203
neg (n)	0.109	0.487	0.079	0.138	0.203
ratio (r)	0.105	0.482	0.087	0.146	0.206
avg(p,n,r)	0.137	0.532	0.119	0.149	0.234

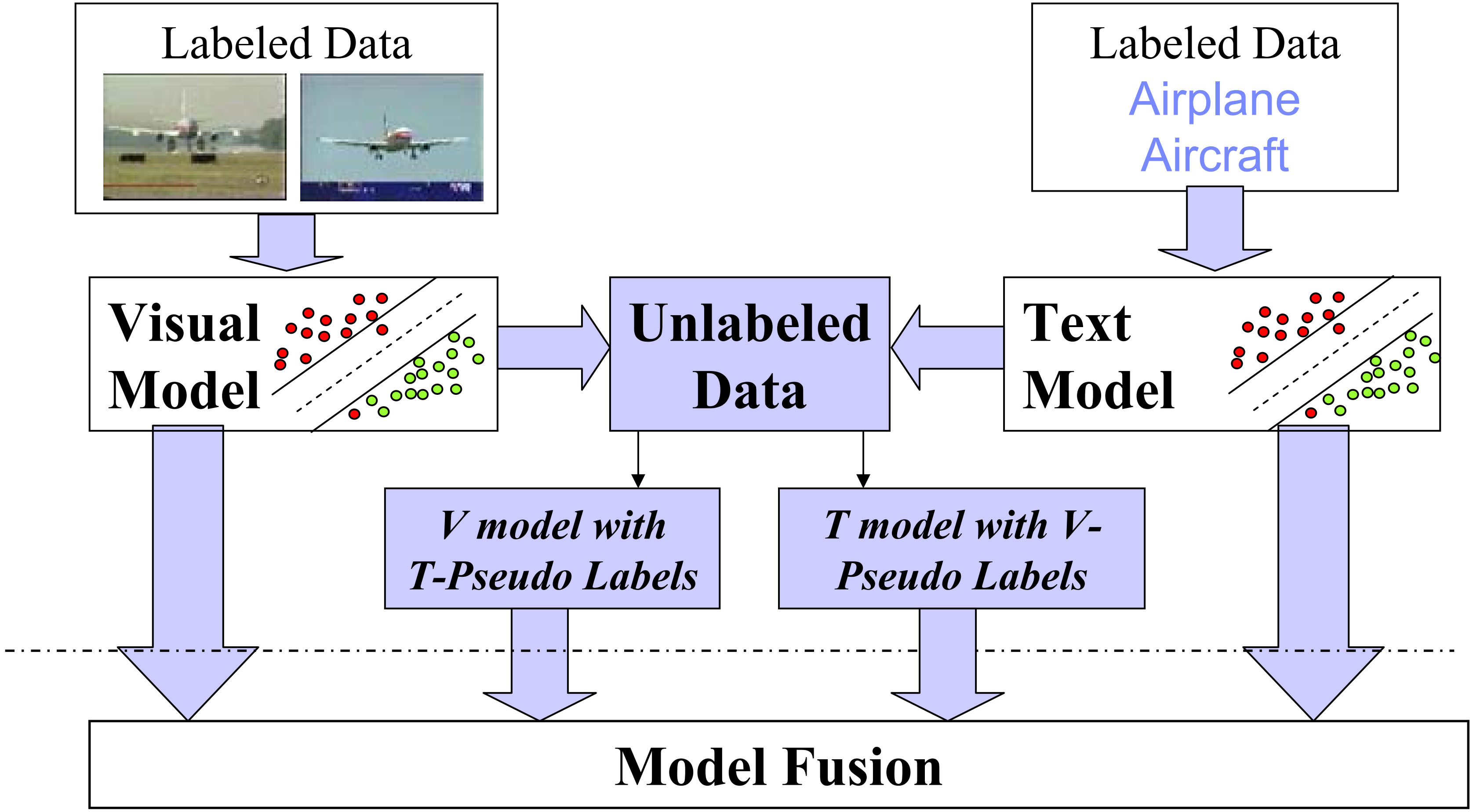
OBSERVATIONS:

- **Individual selection strategies perform optimally for different concepts**
- **Fusion across selection strategies always improves performance**
- **Improvement is between 5% and 30%**

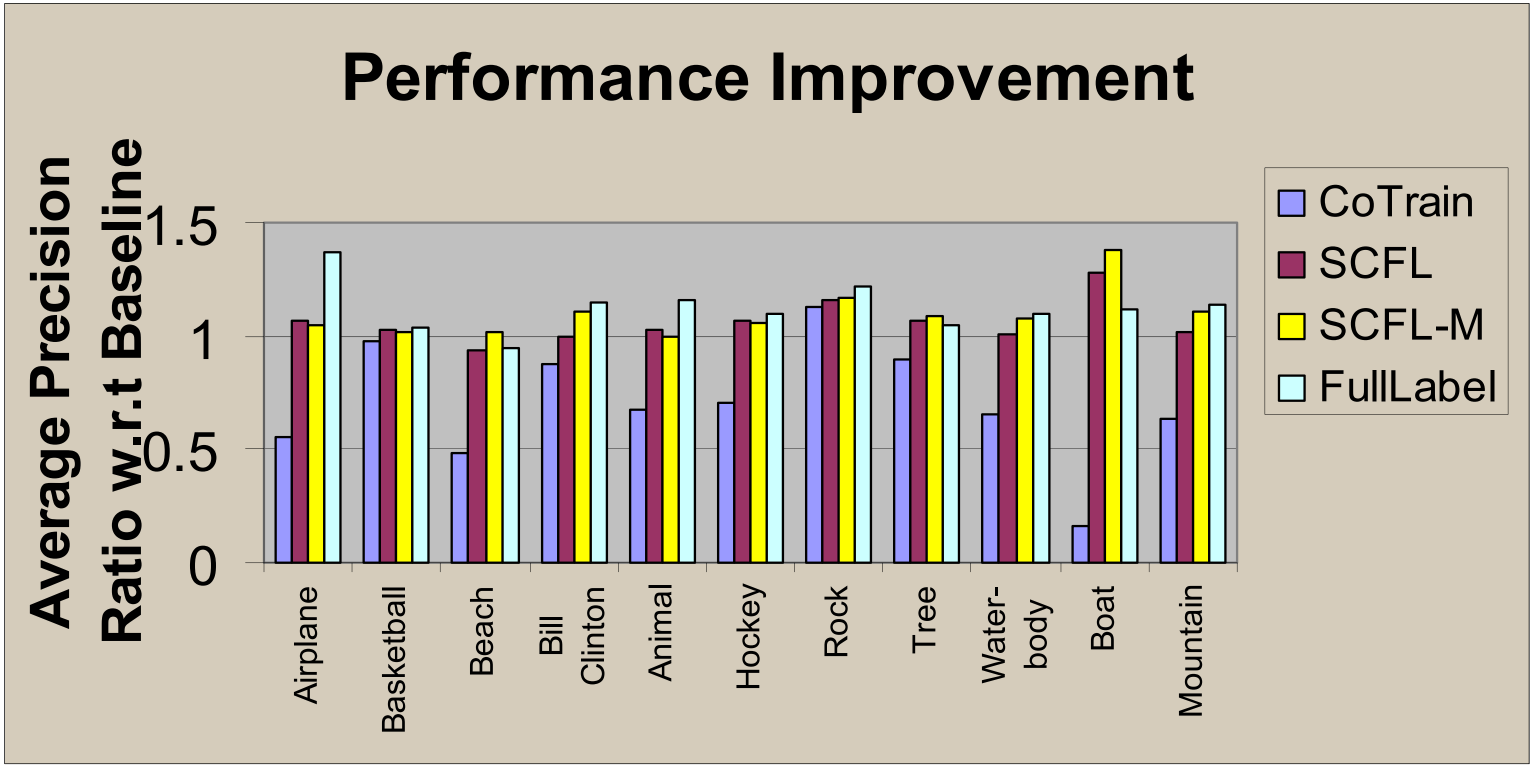
Co-training


Visual Aspect
Text Aspect
Labeled Data
Labeled Data
Bill
Clinton
V-Pseudo Labeled Data
T-Pseudo Labeled Data
Visual Model
Unlabeled Data
Text Model
Training
Testing
Prediction
Testing Data
Prediction
Final Prediction

Extending Co-training: Semi supervised Cross Feature Learning



SCFL performs better than Co-training



- CoTraining range: **0.55 to 1.12**. Average: **30 % worse**
- SCFL range: **0.93 to 1.27**. Average: **6 % better**
- SCFL-M: **1.0 to 1.38**. Average: **10 % better**
- Fully-Labeled range: **0.95 to 1.36**. Average: **12 % better**

Learning Multimedia Semantics

A. Supervised Detection
   1. Static Classifiers
   2. Spatial+Temporal Classifiers
B. Multimodal Fusion
   3. Late fusion using Ensembles
   4. Intermediate Fusion for temporal evolution using graphical models
C. Enforcing Spatial, Temporal and Conceptual Context
   5. Learning Context using Multinet
D. Semi-Supervised Learning
   6. Labeled+Unlabeled Learning
   7. Active Learning
   8. Multiple Instance Learning
   9. Co-training
E. **Unsupervised Clustering**
   10. **Spatial**
   11. **Spatio-temporal using hierarchical HMMs**
F. Semantic Feature Extraction and Search
   12. Query Learning
   13. Leveraging detected semantic concepts for complex query answering

Discovering Recurring Patterns and Structure

Problem Statement

Short-term structure and long-term relationship are common in broadcast videos like talk shows, sport videos, news etc.

Examples: anchor (news), pitch (baseball), laughter (Late-night with DL.)

Can we capture short term and long term structure and discover recurring patterns in **unsupervised fashion.**

Prior Art

Early Use of HMMs for capturing stationarity and transition and its application to clustering: A. B. Poritz, Levenson et al.

Scene Segmentation (using HMMs): Wolf, Ferman & Tekalp; Kender & Yeo; Liu, Huang & Wang; Sundaram and Chang, Divakaran & Chang.

Multimodal scene similarity: Nakamura & Kanade; Nam Cetin & Tewfik; Naphade, Wang & Huang; Adams et al.

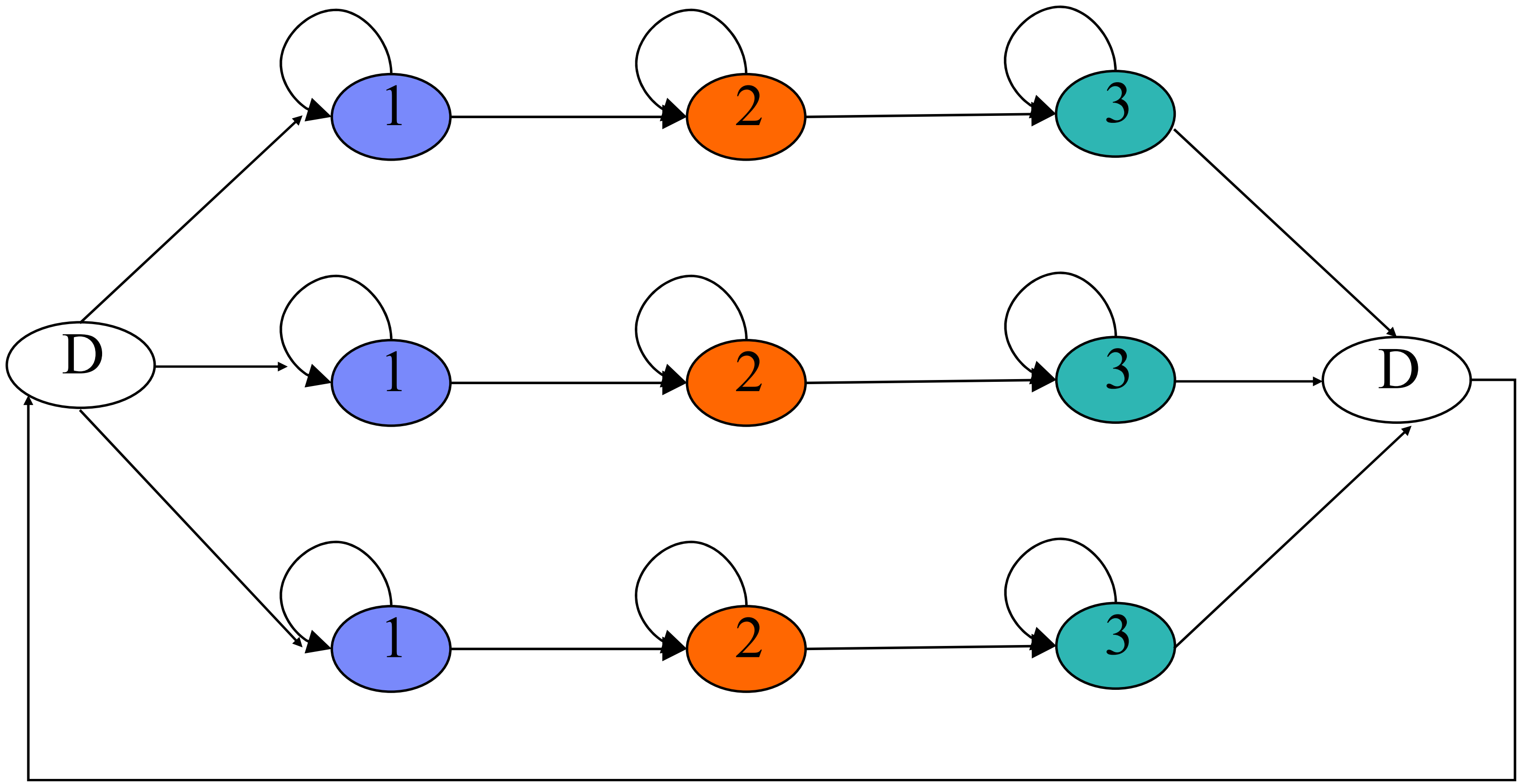
Strategy

- Given a set of examples and the knowledge that they contain multiple instances of <u>recurring temporal patterns</u>, attempt to extract the recurring patterns.
- Use <u>unsupervised temporal clustering</u> using a hierarchical ergodic model with non-ergodic temporal pattern models.
- <u>User</u> then needs to <u>analyze only</u> the <u>extracted</u> recurring set to quickly propagate annotation.

Result

- Successfully detects and extracts recurring patterns (laughter, explosion, monologue etc.) and regular structure.
- <u>Substantially reduces time</u> needed for semantic <u>annotation</u>.

Capturing Short Term Stationarity and Long-Term Structure

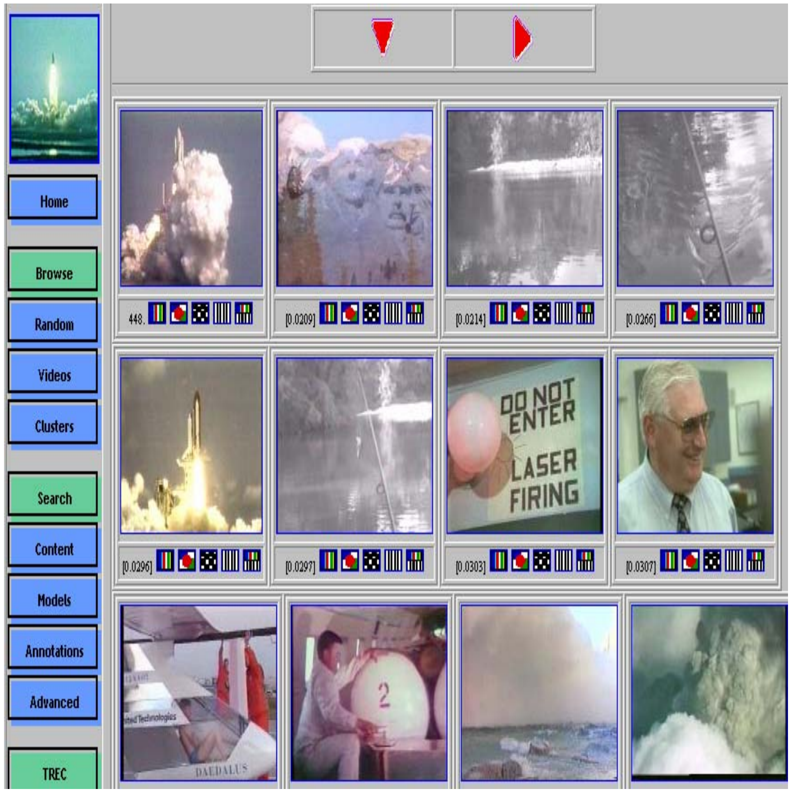


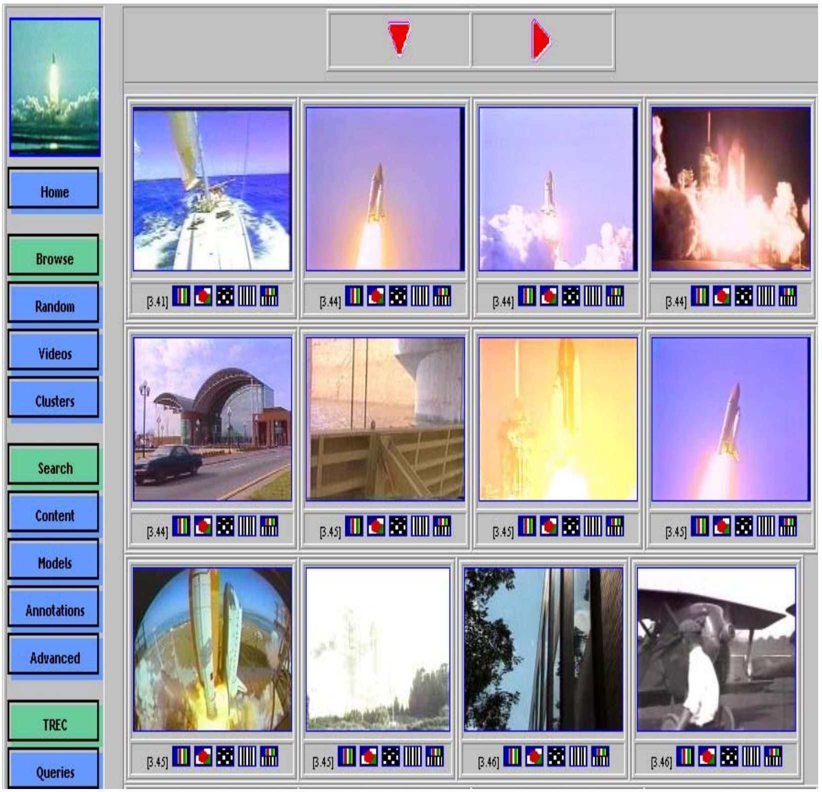
- Each branch: non-ergodic
- All branches embedded in a hierarchical ergodic structure

Learning Multimedia Semantics

A. Supervised Detection
   1. Static Classifiers
   2. Spatial+Temporal Classifiers
B. Multimodal Fusion
   3. Late fusion using Ensembles
   4. Intermediate Fusion for temporal evolution using graphical models
C. Enforcing Spatial, Temporal and Conceptual Context
   5. Learning Context using Multinet
D. Semi-Supervised Learning
   6. Labeled+Unlabeled Learning
   7. Active Learning
   8. Multiple Instance Learning
   9. Co-training
E. Unsupervised Clustering
   10. Spatial
   11. Spatio-temporal using hierarchical HMMs
F. **Semantic Feature Extraction and Search**
   12. **Query Learning**
   13. **Leveraging detected semantic concepts for complex query answering**

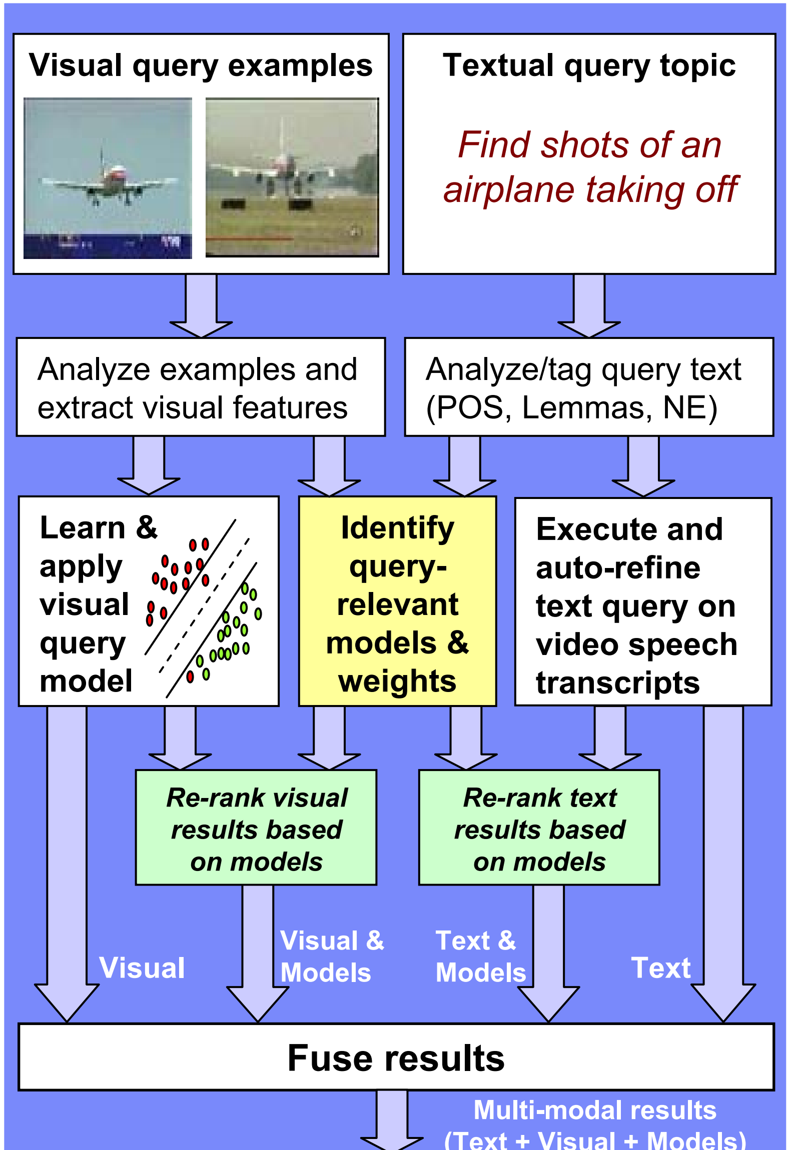
Retrieval Results: QBE vs. QBK
Light-weight vs. Heavy-weight Classification





- Model-based retrieval improves retrieval effectiveness
- Up to 200 % higher precision for same recall compared to CBR

Automatic Search with Multimodal Concepts & Context

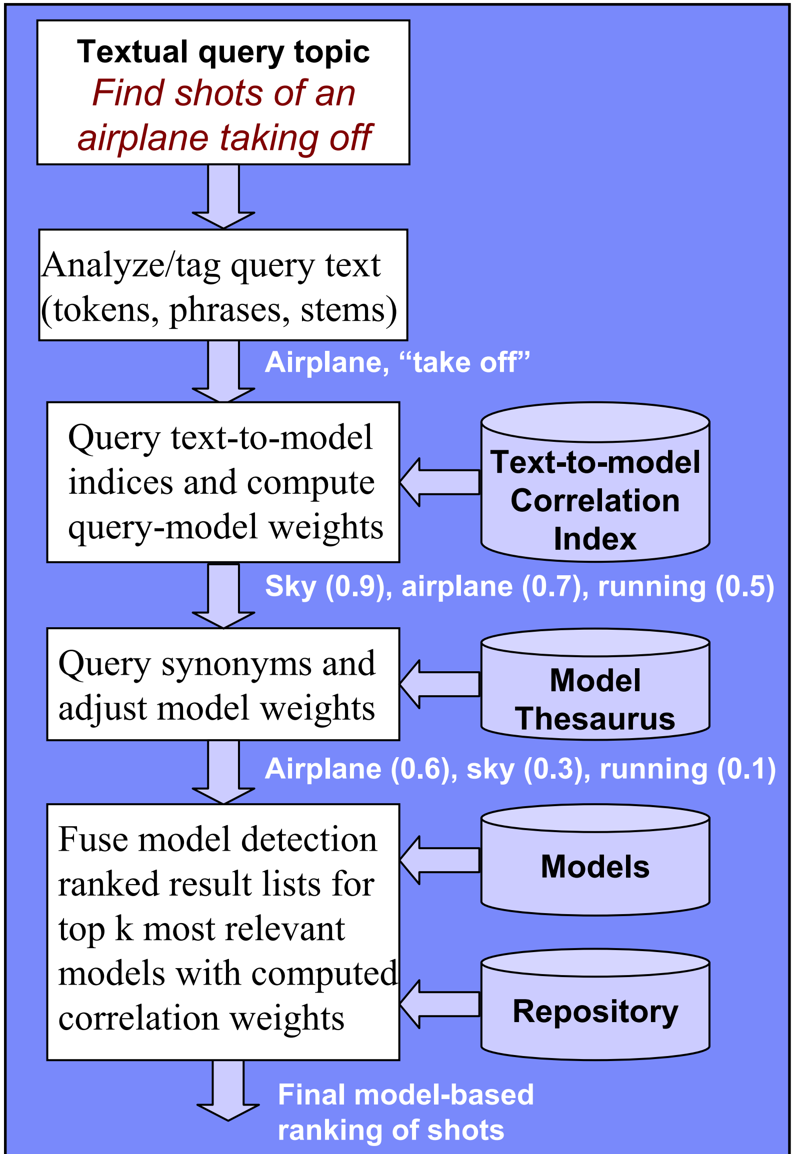


Automatic search approaches

- Text retrieval with automatic query expansion
- Visual retrieval with light-weight learning—nearest neighbor & discriminative models
- Model-based retrieval with automatic query- to-model mapping and weight determination

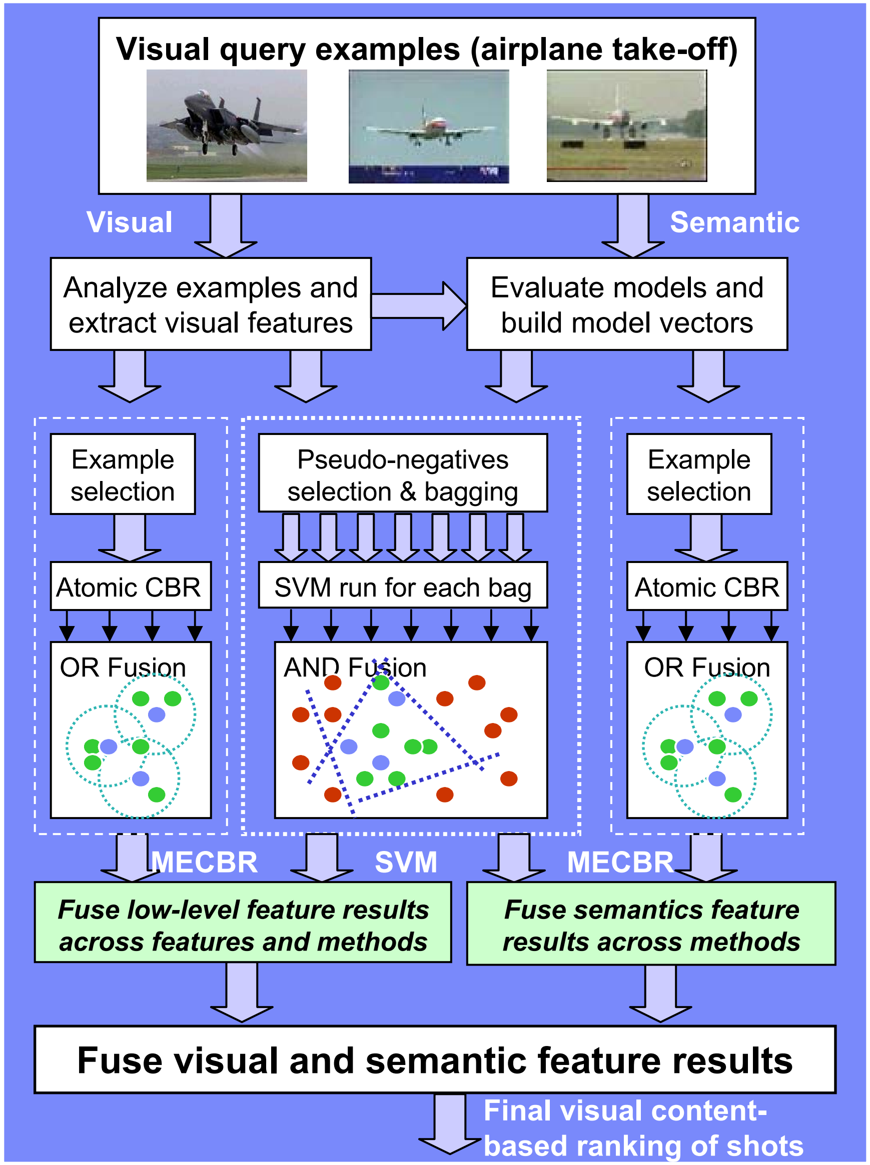
- **Query-independent fusion approach**
  - Simple score averaging within modality:
    - •Statistical normalization for visual runs
    - •Rank normalization for text runs
  - Round-Robin fusion across modalities
    - •OR fusion of rank normalized lists
  - Model-based re-ranking of text & visual runs
- **Results:** Highest MAP for automatic type A search at TRECVID

Model-Based Retrieval



- **Problem**
  - Given query text, identify relevant semantic models and use to retrieve relevant content
- **Challenges**
  - Expanding query in one modality (text) with models built from different modality (visual)
- **Approaches**
  - **Corpus-based statistical approach**
    - Use co-occurrence statistics between ASR tokens and detected concepts:
    - Supervised—learn correlations using concept ground truth on training set
    - Unsupervised—learn correlations using concept detection confidence on test set
  - **Language-driven lexical approach**
    - Use model thesaurus for synonym-based query expansion
- **Highlights**
  - Used to re-rank text and visual baselines
  - Improved both baselines by 10-20%

Visual Query Retrieval using Query Learning



- **Problem**
  - Given few positive visual examples, retrieve similar video content
- **Challenges**
  - Complex query topics (high semantics)
  - Very small number of query examples
  - No negative examples
- **Approach**
  - Modeled as light-weight learning problem
    - Sample pseudo-negative examples
    - Use bagging-like approach to address imbalanced learning problem
  - Fusion of two synergistic approaches:
    - Support Vector Machines
    - MECBR (Nearest Neighbor)
  - Low-level and semantic visual features
- **Highlights**
  - Dominated speech-based retrieval results
  - Outperformed all other automatic type A search approaches at TRECVID 05

Performance Evaluation

Milind Naphade

NIST TRECVID Benchmark at a Glance

- **TRECVID:**
  - NIST benchmark for evaluating state of the art in video retrieval
- **Benchmark tasks:**
  - Shot Boundary Determination
  - **Semantic Concept Detection**
  - Story Segmentation
  - **Search**

***Topic 101**: Find shots of a basket being made - the basketball passes down through the hoop and net*

***Topic 129**: Find shots zooming in on the US Capitol dome.*

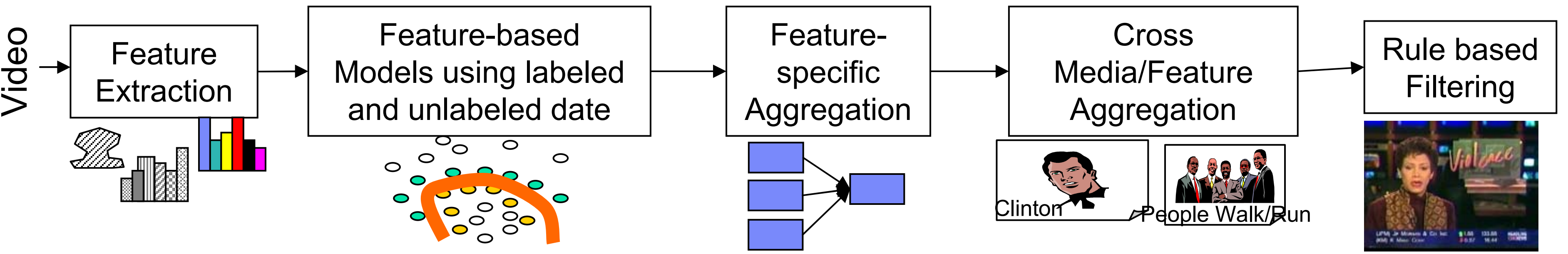
***Topic 104 and 167**: Find shots of an airplane taking off*

Growing Participation

TRECVID 2001	TRECVID 2002	TRECVID 2003	TRECVID 2004	TRECVID 2005
12 Participants	17 Participants	24 Participants	38 Participants	62 Participants
11 Hours of NIST video	73 Hours of Video from Prelinger archives	133 Hours of 1998 ABC, CNN news & C-SPAN	173 Hours of 1998 ABC, CNN news & C-SPAN	220 Hours of 2004 news from U.S., Arabic, Chinese sources, BBC stock shots

Growing Data Sets

TRECVID Systems: A Canonical View



▪Visual ▪Aural ▪ASR/CC ▪VOCR ▪Metadata	▪Classifiers ▪Feature Reduction ▪Granularity of Modeling	▪Late vs. Early Aggregation ▪Supervised vs. Unsupervised Aggregation	▪Synchronization ▪Late vs. Early Aggregation ▪Supervised vs. Unsupervised Aggregation	▪Classifiers ▪Context Modeling	▪Domain Filters ▪ Domain independent filters
Necessary	Necessary	Common	Common	Rare	Common

The TRECVID Benchmark Concepts

QBE era

- Outdoors
- Indoors
- Face
- Music
- Speech

TREC 01

- Outdoors
- Sky
- Greenery
- Rock
- Sand
- Face
- Water
- Boat
- Rocket
- Vehicle
- News Anchor
- Airplane
- Rocket Launch
- Music
- Speech
- Explosion
- Smoke
- Fire
- Zoom In

TREC 02

- Outdoors
- Indoors
- Landscape
- Cityscape
- Face
- People
- Text Overlay
- Instrument Sound
- Speech
- Monologue

TREC 03

- Outdoors
- Non Studio Setting
- Vegetation
- Building
- Road
- News Subject Face
- People
- Sport Event
- Weather News
- Car/Truck/Bus
- Aircraft
- Animal
- Person X (M. Albright)
- Female Speech
- News Subject Monologue
- Physical Violence
- Zoom In

TREC 04

- Beach
- Road
- People Walking or Running
- Basket Scored
- Train
- Boat
- Airplane Takeoff
- Person X (M. Albright)
- Person X (Bill Clinton)
- Physical Violence

Site
Object
Event
Specific Entity

- Increasing Specificity
- Increasing Complexity
- Increasing Events and Objects over Sites
- Decreasing Training Set Support in terms of number of examples

Feature-based Modeling

Generic vs. Specific Modeling

Generic Classifiers

- KNN
- SVM
- GMM
- HMM
- MAXENT
- Shape, Motion and Appearance Templates
- Boosting
- Trees

Features Modeled

- Keyframe-based
- Multi-frame based

Validation-Based Optimization

Opinions:

- **Generic concept modeling is necessary to push the envelop although for each concept, it may be possible to perform better with a specific approach fine tuned for that concept.**
- **Generic Machine Learning Techniques are responsible for the siginificant advances that we are seeing in concept detection and modeling**
- **A combination of better computing power and better algorithms**

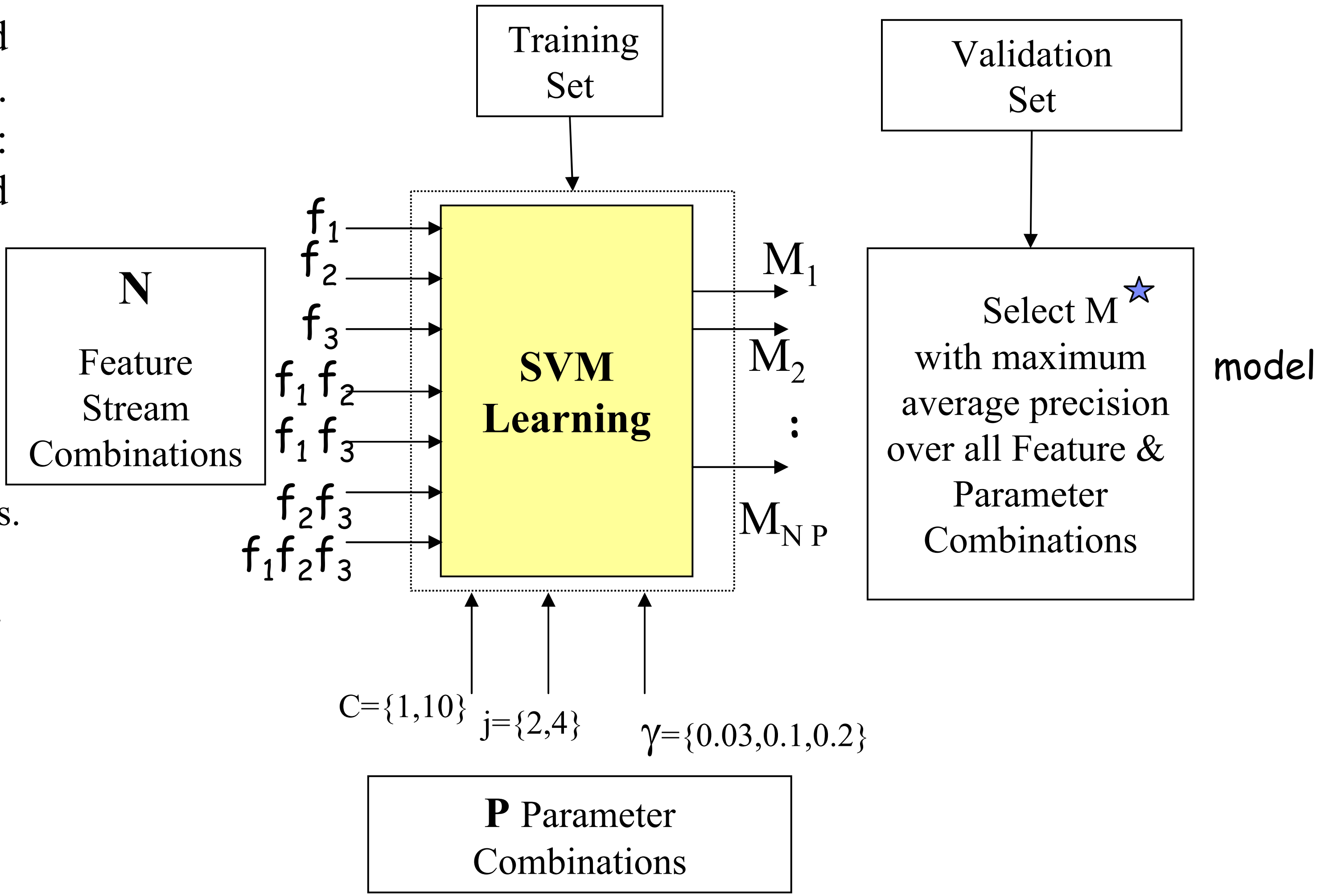
SVM Models: Minimizing Sensitivity

Global Support: Features extracted from entire frame.

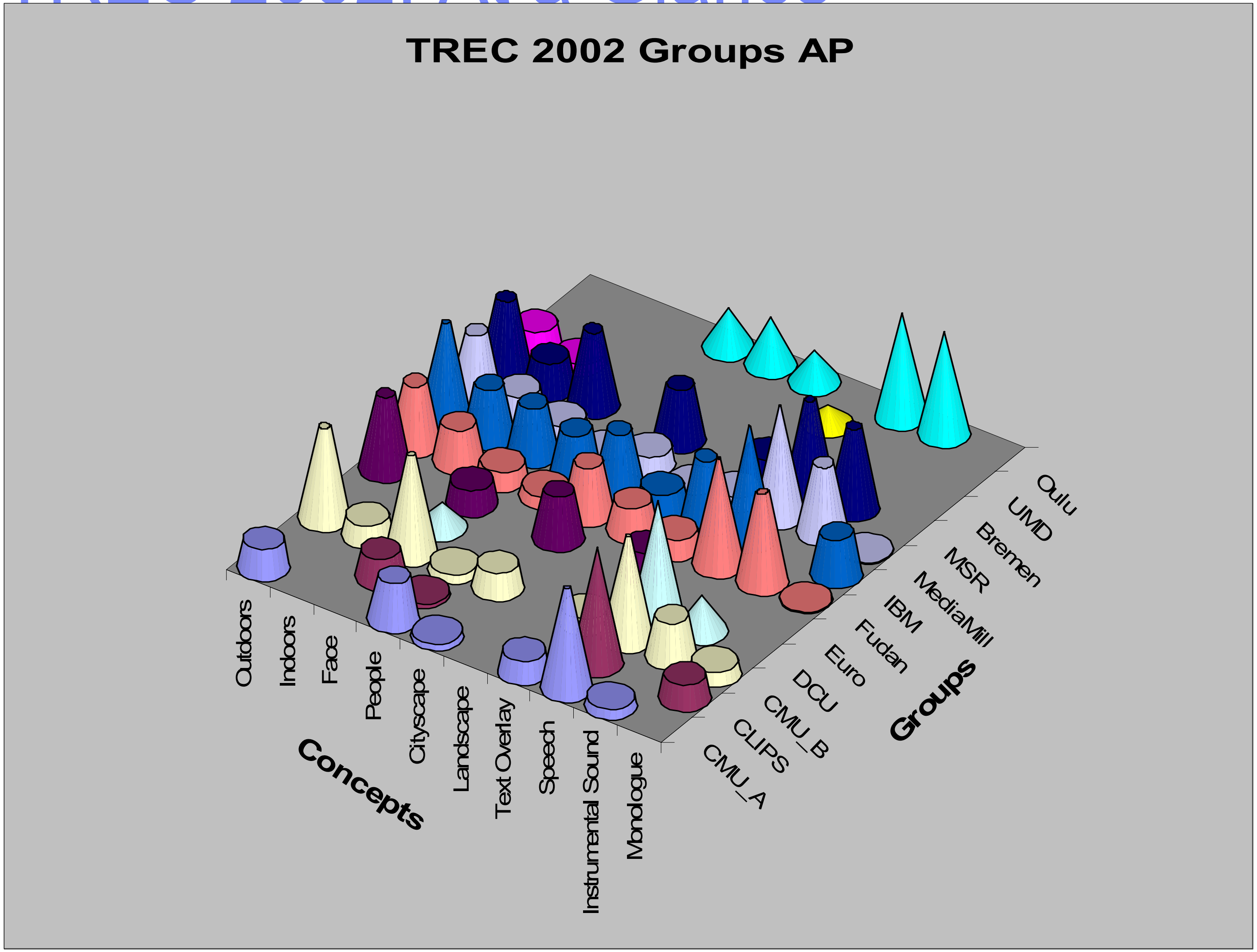
Regional Support: Features extracted from regions.

Features

- Color: HSV histogram, Moments.
- Texture: Edge direction histogram.
- Gray-level Co-occurrence
- Shape: Moment Invariants

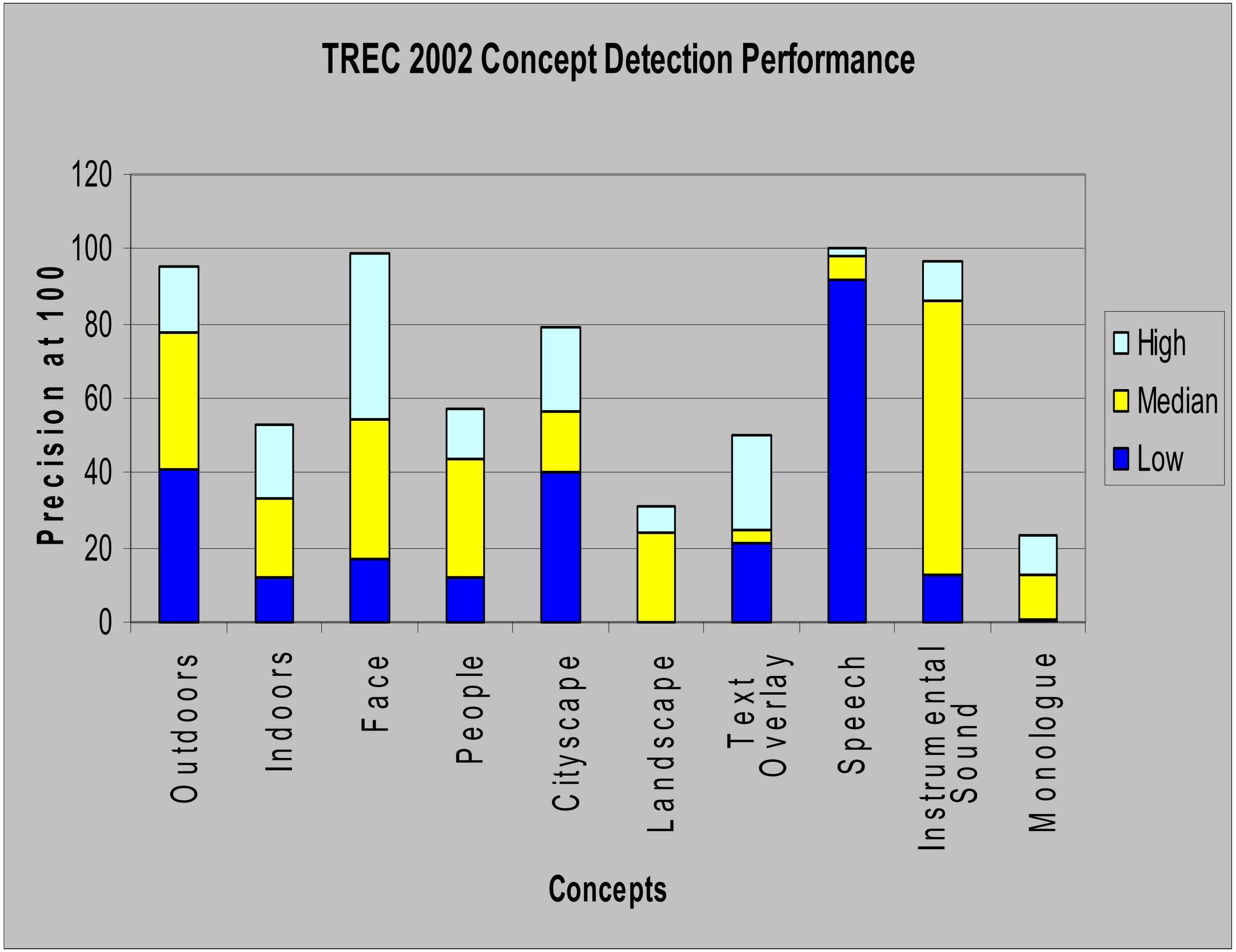


TREC 2002: At a Glance



- 10 Concepts
- 11 Teams
- 24 hours of training data
- 5 hours of test data
- All runs evaluated to full depth
- MAP evaluated at depth of 1000 shots
- Most concepts were frequent

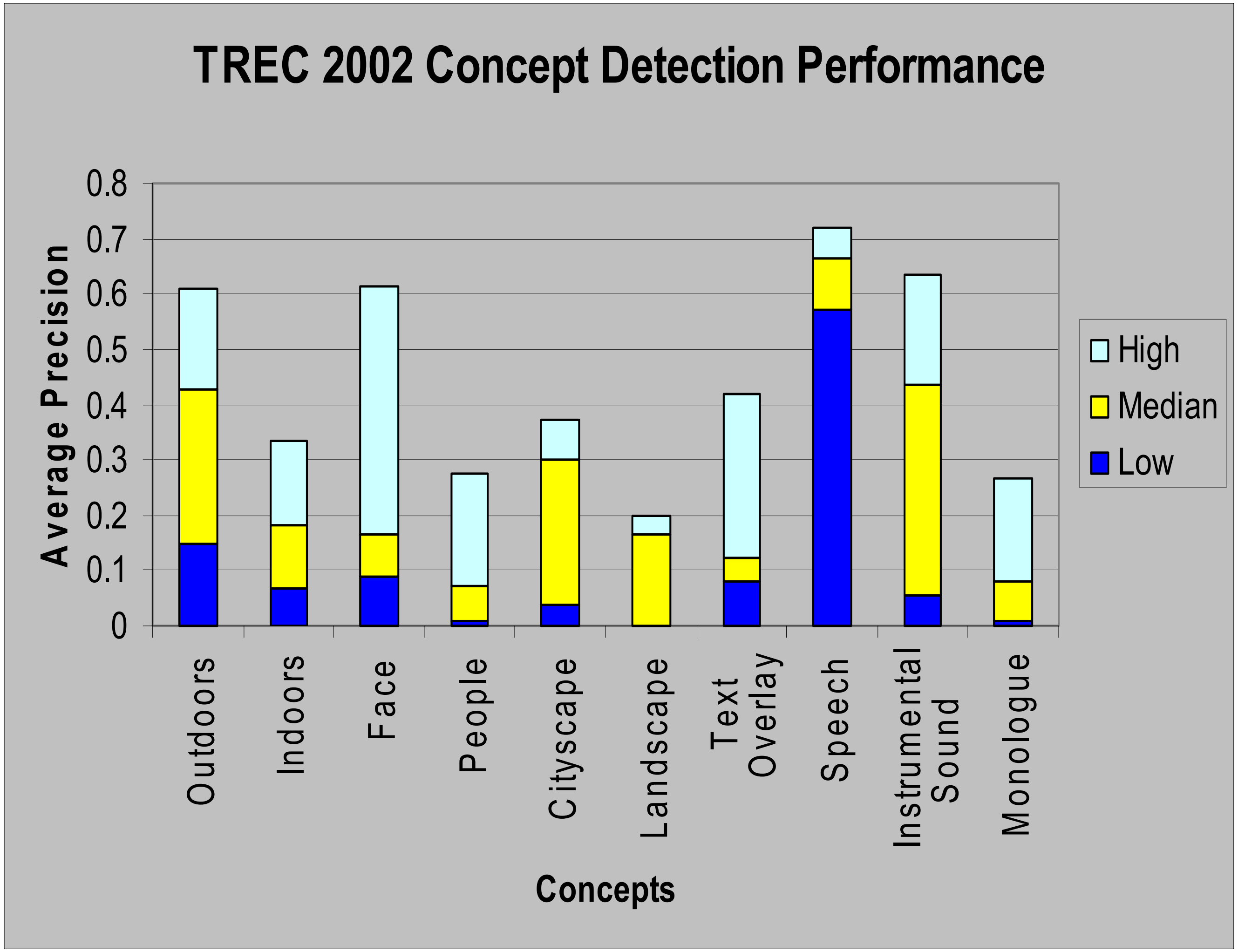
TRECVID Concept Detection 2002



- Assuming a Median of 50 hits in the top 100 as a measure of the maturity of the detector, 5/10 fared well
- Assuming high values as a measure of feasibility of detection, 8/10 fared well

Opinion: Generic and frequent concepts seem feasible candidates for robust detection

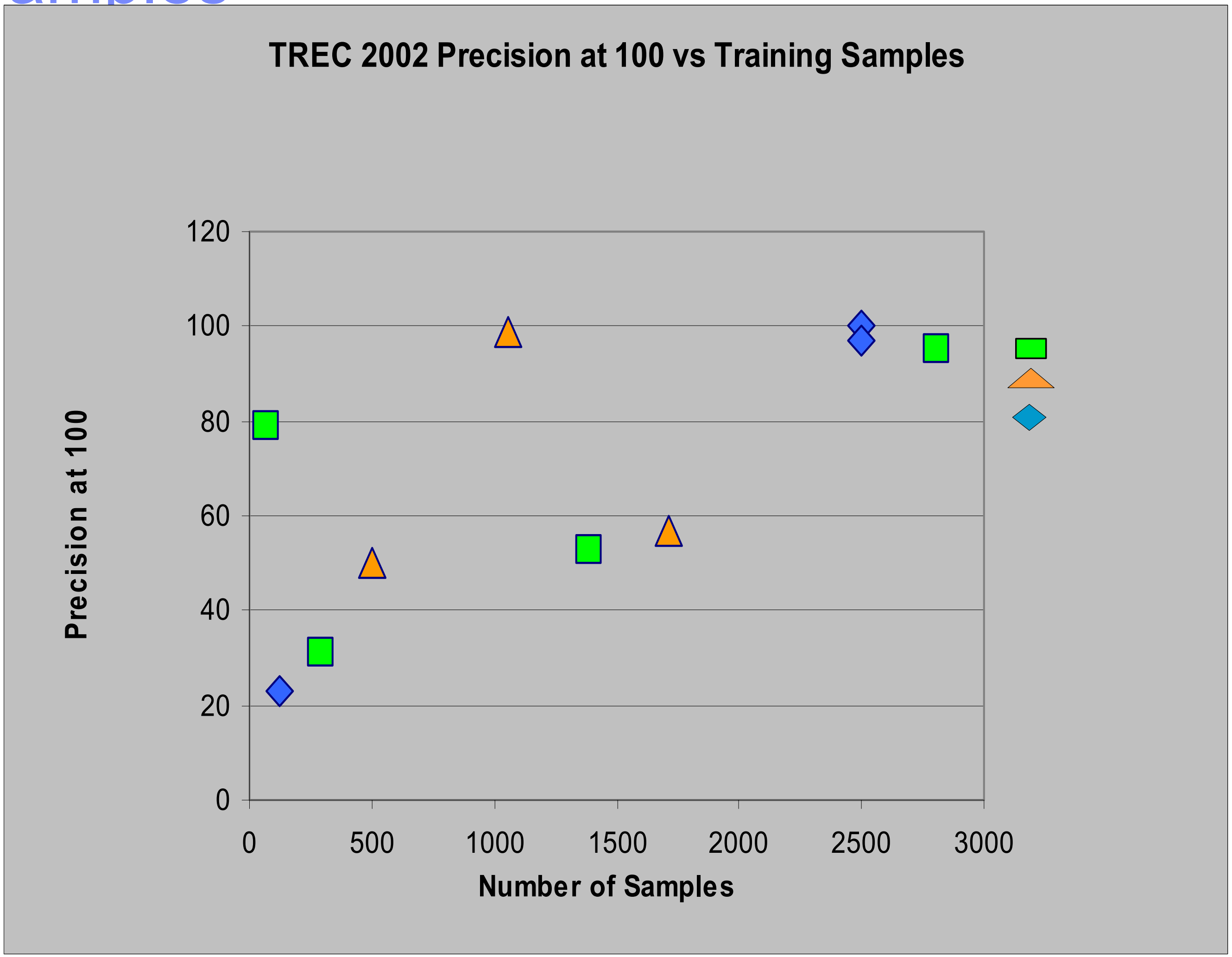
TRECVID Concept Detection 2002



- AP in 2002 did not account for presence of more true hits than evaluation depth. So the AP for Speech and Instrumental Sound should have been in the mid nineties.
- All concepts returned decent average precisions for the best performing systems and the median AP was below 0.1 only for 2 of the 10 concepts

Opinion: Generic and frequent concepts seem feasible candidates for robust detection

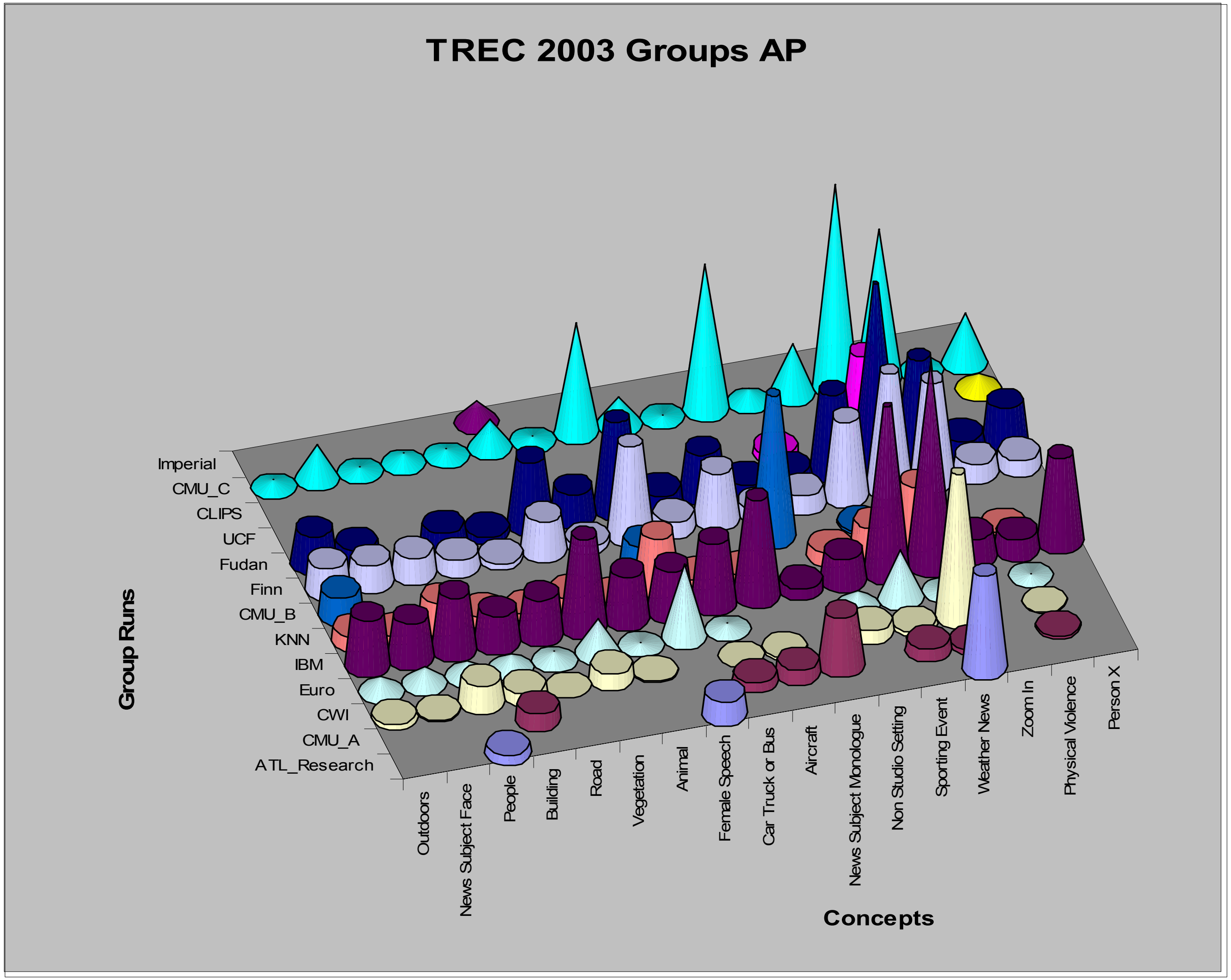
2002 Concept Complexity as a Function of Training Samples



- Frequent concepts were easier to detect as robust models could be built based on training set.
- For its relative rarity, Cityscape fared well.
- Hard to determine if difficulty in detection rose from concept being object/site or event

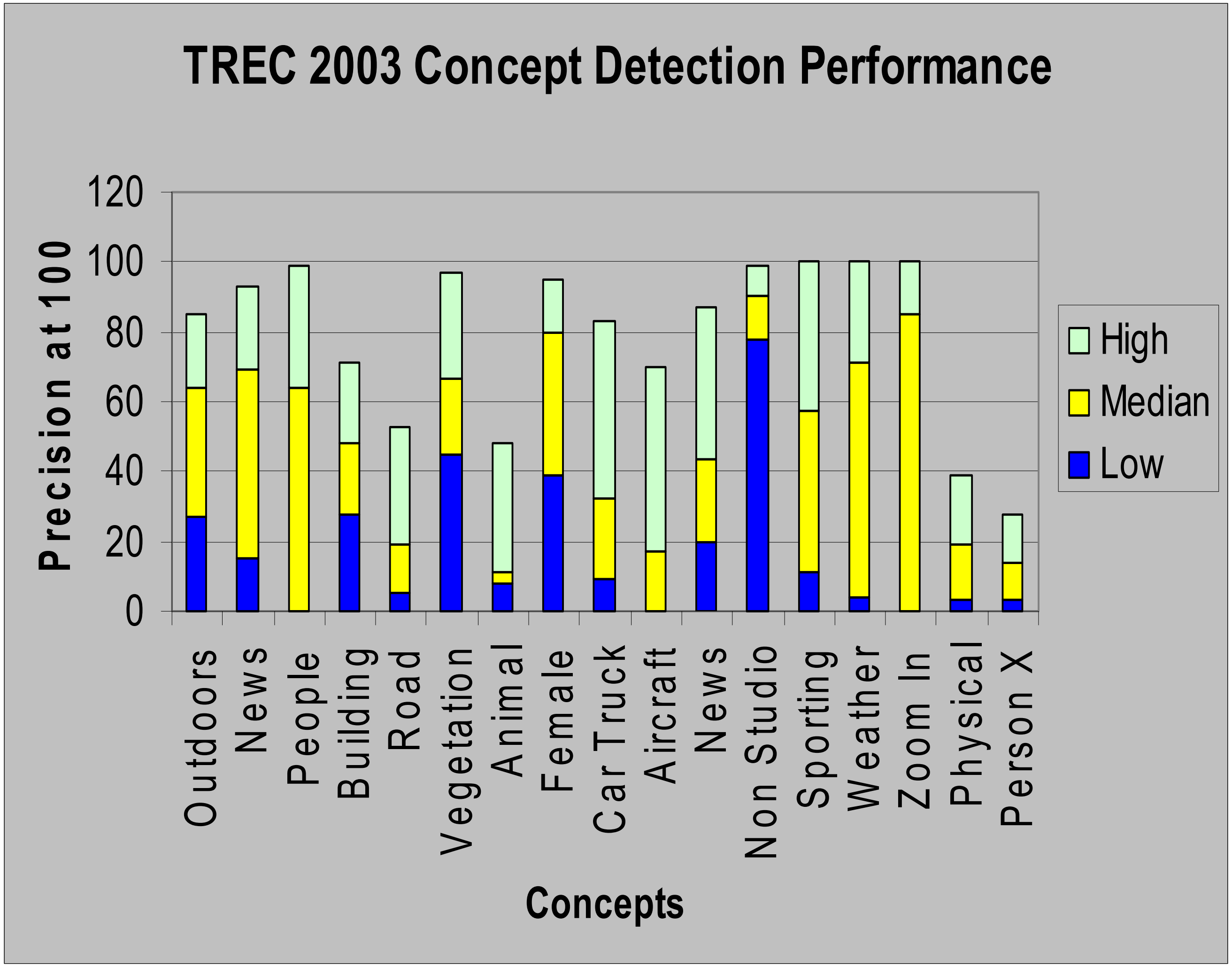
Opinion: Rapid detection improvement when # training samples increase, law of diminishing returns later

TREC 2003: At a Glance



- 17 Concepts
- 11 Teams
- 60 hours of training data
- 60 hours of test data
- Ground truth pooled by using top 100 items from runs
- MAP evaluated at depth of 1000 shots
- Mix of frequent and infrequent concepts

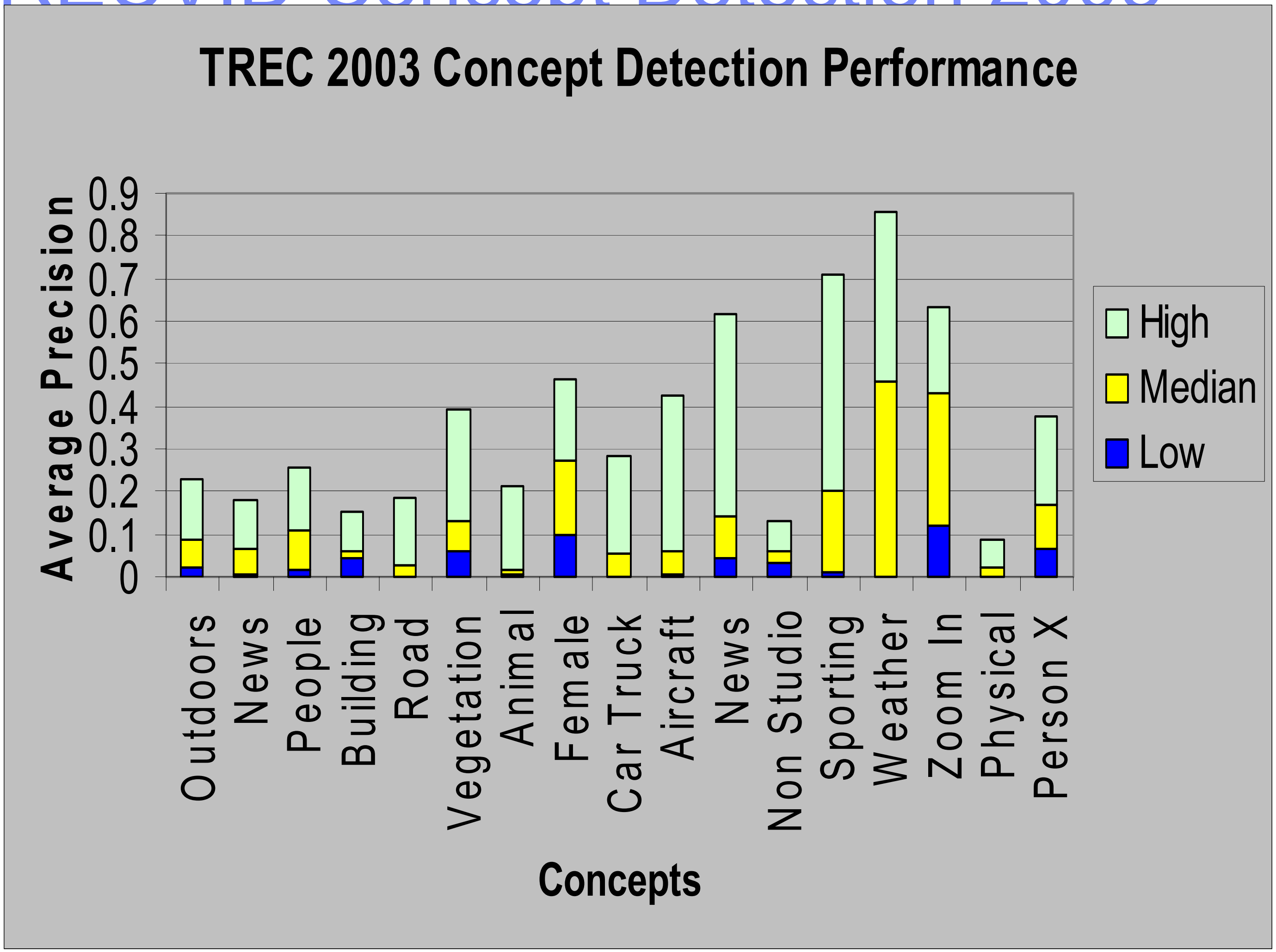
TRECVID Concept Detection 2003



- Assuming a Median of 50 hits in the top 100 as a measure of the maturity of the detector, 10/17 fared well
- Physical Violence and Specific Person Detection fared poorly
- Assuming high values as measure of feasibility of detection 14/17 fared well.

Opinion: Generic and frequent concepts feasibility validated on a larger number of concepts

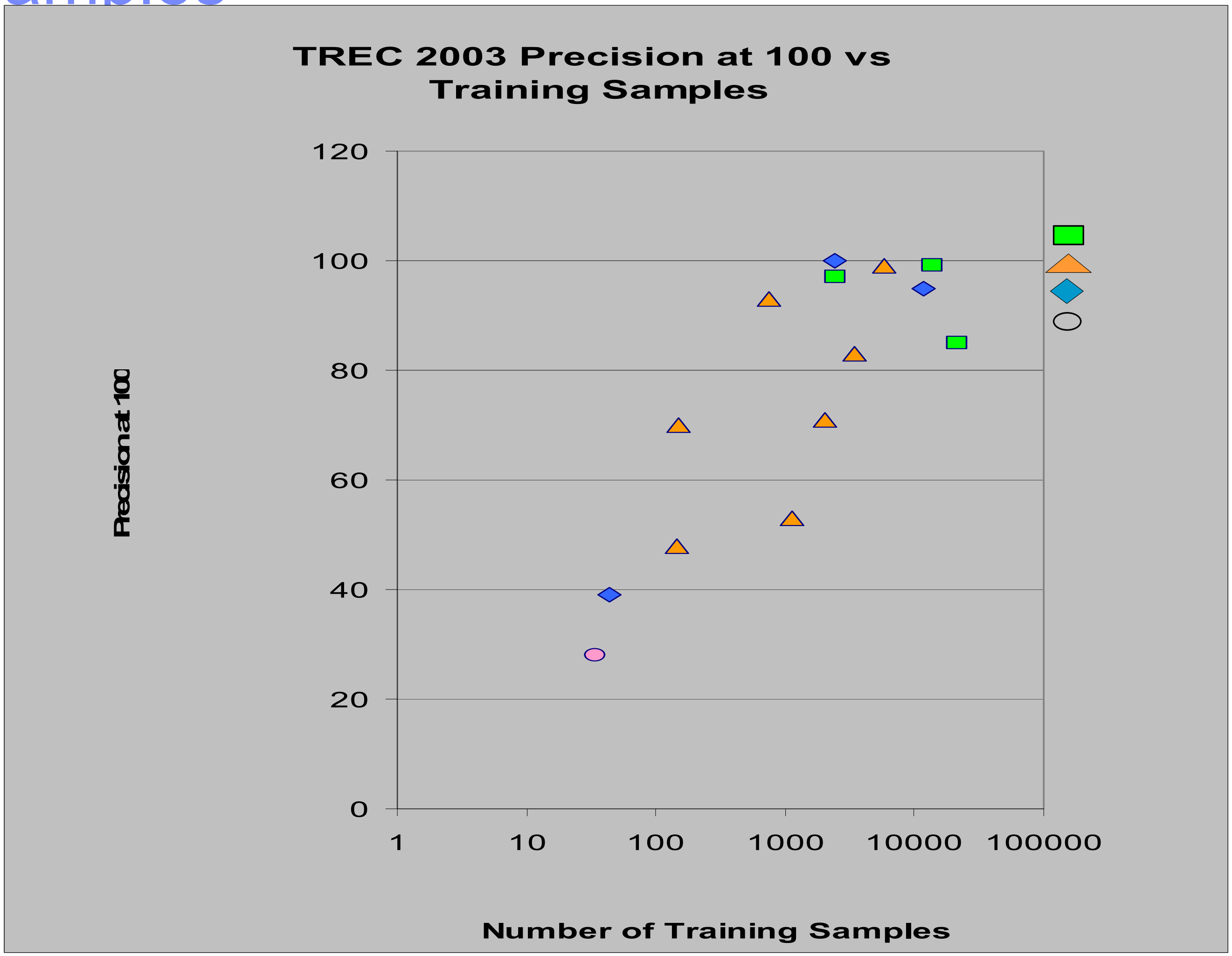
TRECVID Concept Detection 2003



- Impact of pooling with 100 shot depth felt by low AP values of frequent concepts such as Outdoors (80+ in top 100 but AP is only 0.2+ due to pooling problem)
- Some detectors seem better than they may be due to pooling (the denominator effect.. If no one got it, no one got penalized..)

Opinion: Use of AP for frequent concepts misleading. Infrequent concepts fared badly

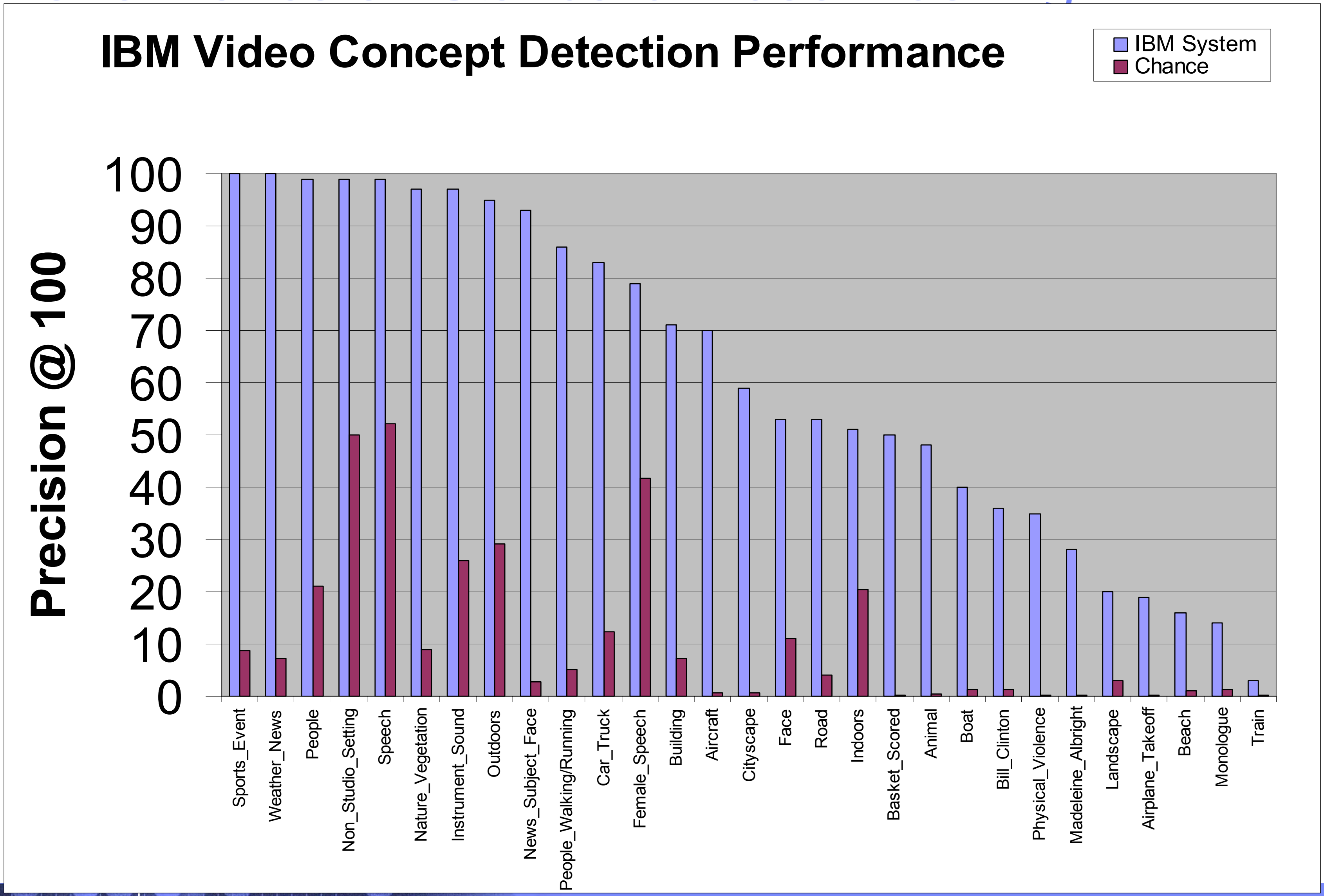
2003 Concept Complexity as a Function of Training Samples



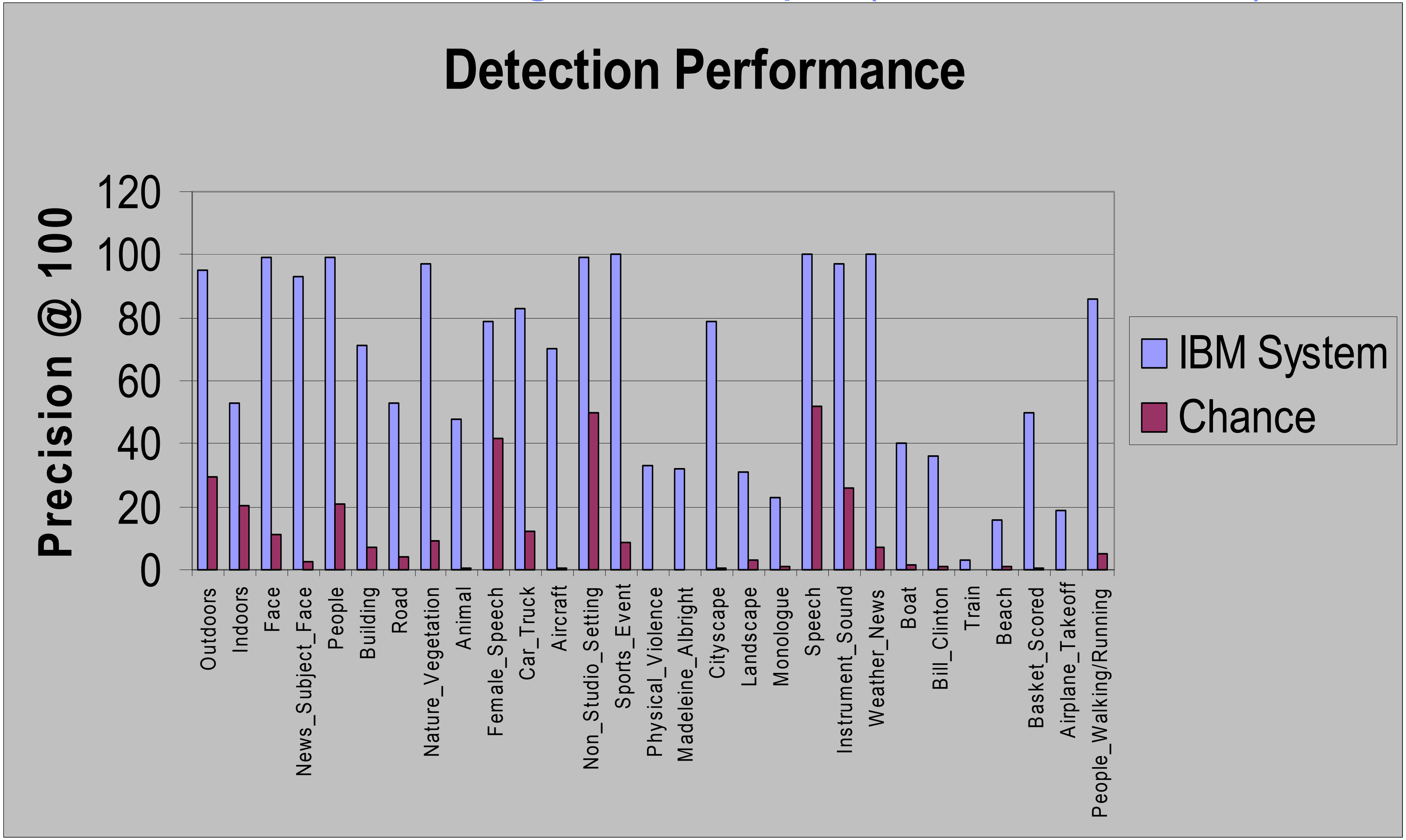
- Aural events easier to detect than visual events
- Objects harder to detect than sites
- Events related to objects thereby harder to detect also

Hypothesis: Log-Linear relationship between performance and positive training sample size?

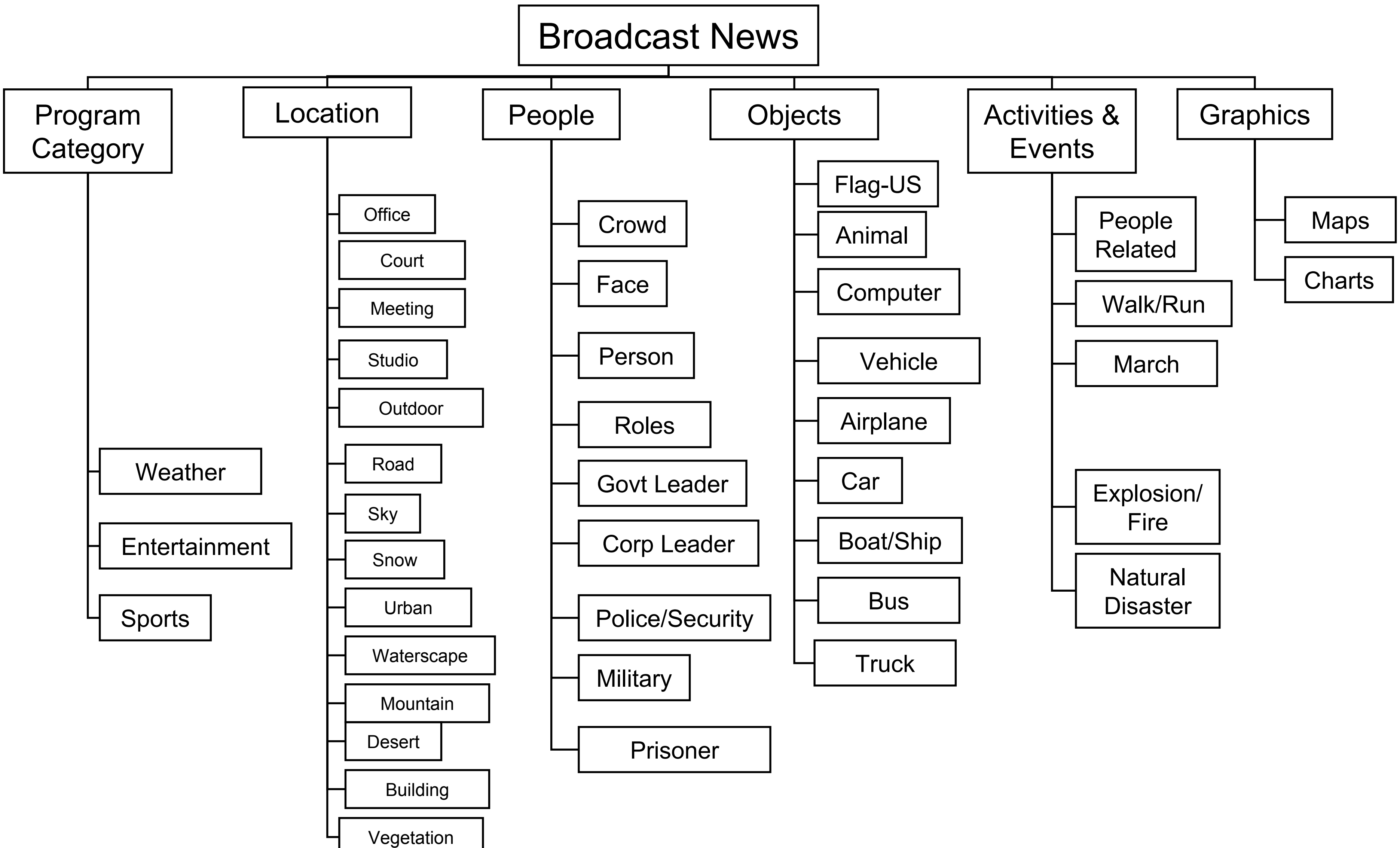
Semantic Concept Detection Achieves High Performance on Standard Video Indexing



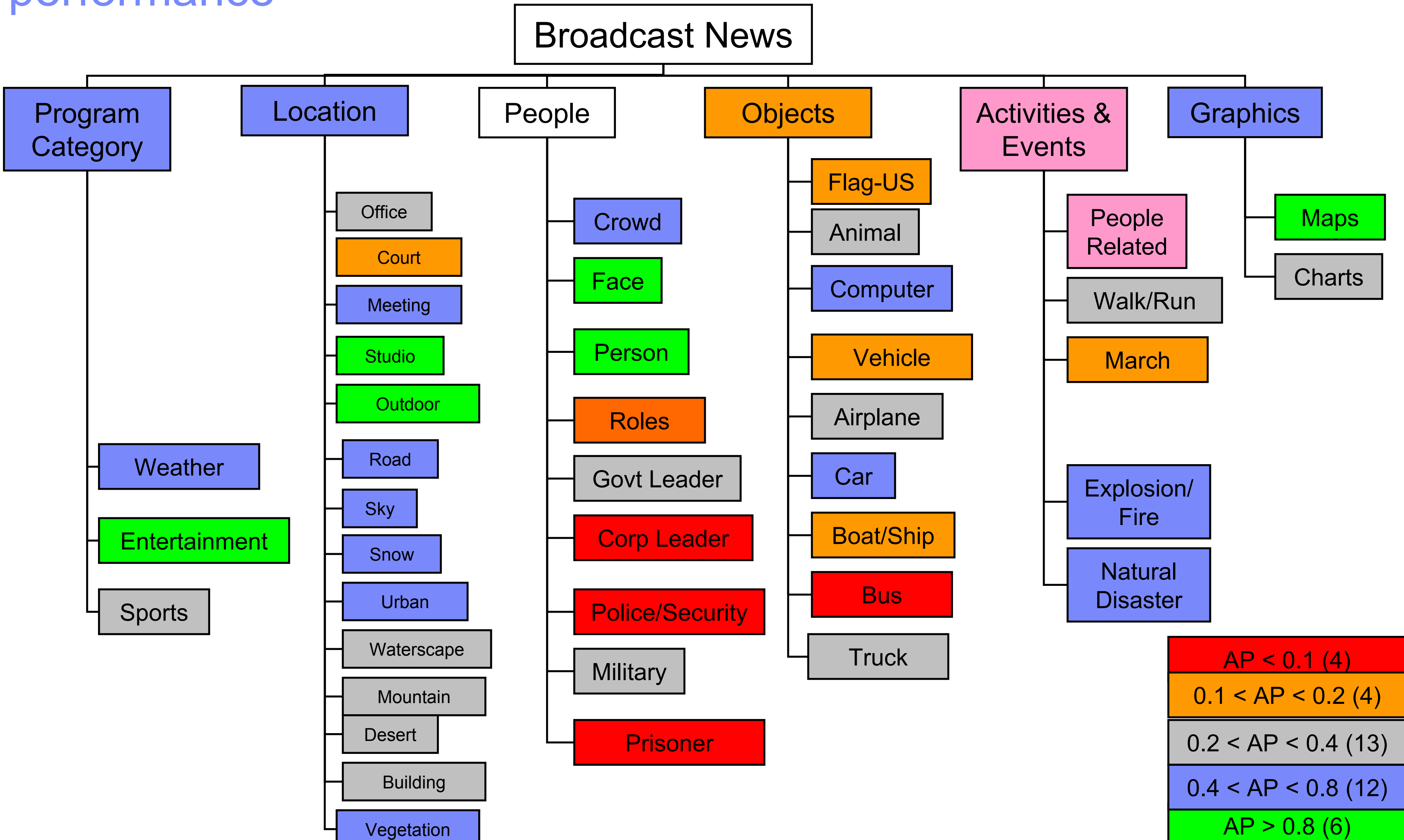
Detection over a wide range of concepts (70 h. news video)



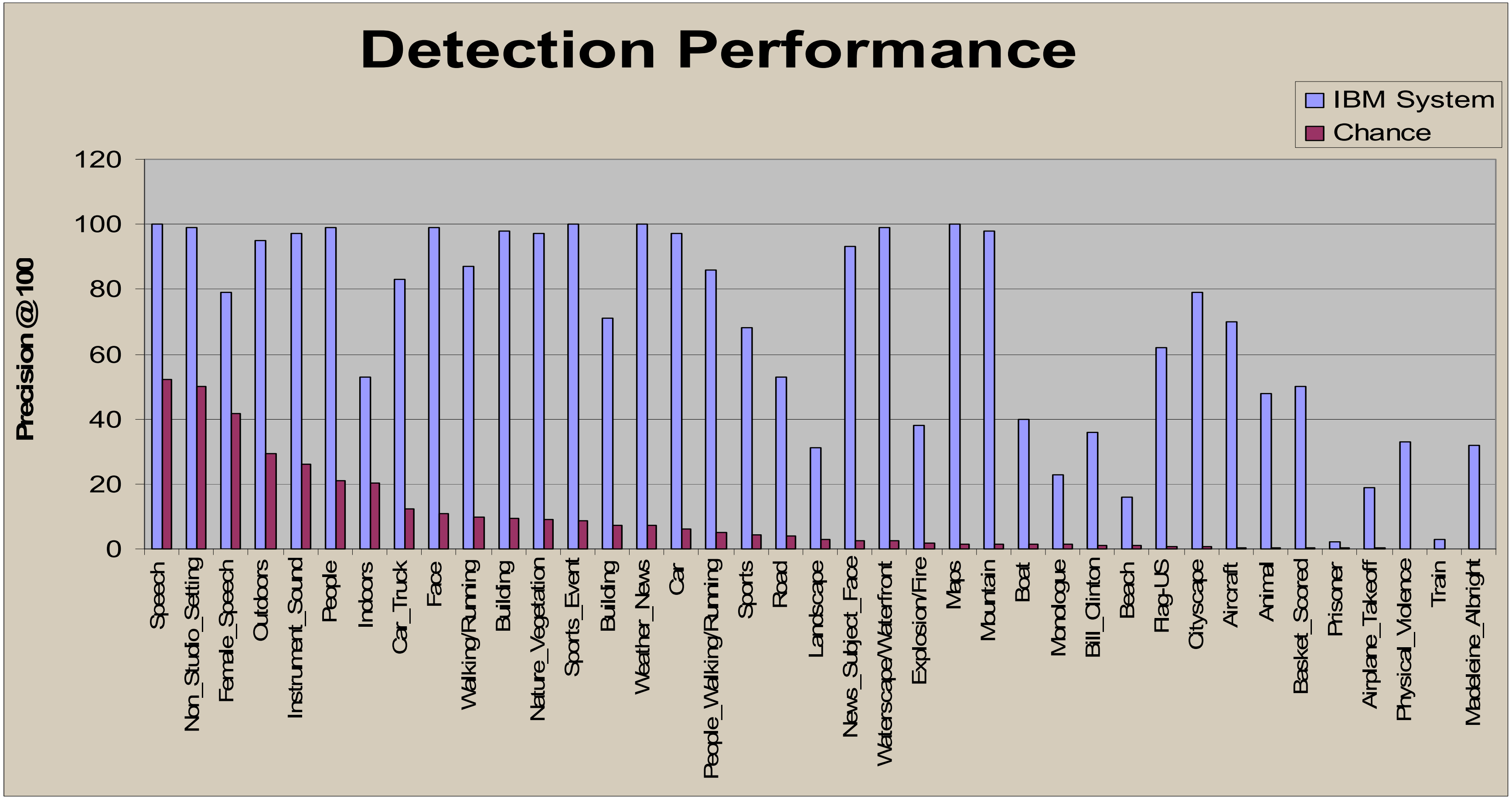
Semantic Concept Lexicon (LSCOM-lite)



Semantic Concept Lexicon (LSCOM-lite) – concept detection performance



MARVEL Overall Semantic Concept Detection Performance TRECVID 2002-2005



- Average hits in top 100 by Chance: **9**
- Average hits in top 100 for MARVEL: **68**

Some Lessons

- The formula of annotating a training set and using this to build concept models works.
- Generic methods worked better than specific methods.
- SVM classifiers worked better in general than other classifiers
- Multimodality helps. In fact it almost always is necessary
- Filtering improves retrieval effectiveness but not very significantly.
- Context helps.
  - Deterministic context enforcement helps improve performance especially when in “Composition mode”
  - Non Studio Setting was enforced with Outdoors,
  - Madeleine Albright had to be detected with a Face and Speech
  - Probabilistic Context helps when deterministic rules cannot be designed.
- Multiple layers of processing helps.
- There are still too many free parameters and knobs in detection systems to understand where the maximum gains are made but combination of multiple detectors for the same concept, whether across features or across modalities seems to provide biggest improvement over individual detectors.

A Picture worth thousand words….
Which Thousand?

DTO

Video Analysis
Content Extraction

A Large Scale Concept Ontology for Multimedia Understanding

Challenge Workshop

Co-PIs: Milind R. Naphade, John R. Smith, Alexander Hauptmann, Shih-Fu Chang

IBM Research, Carnegie Mellon University, Columbia University, CyC Corp.

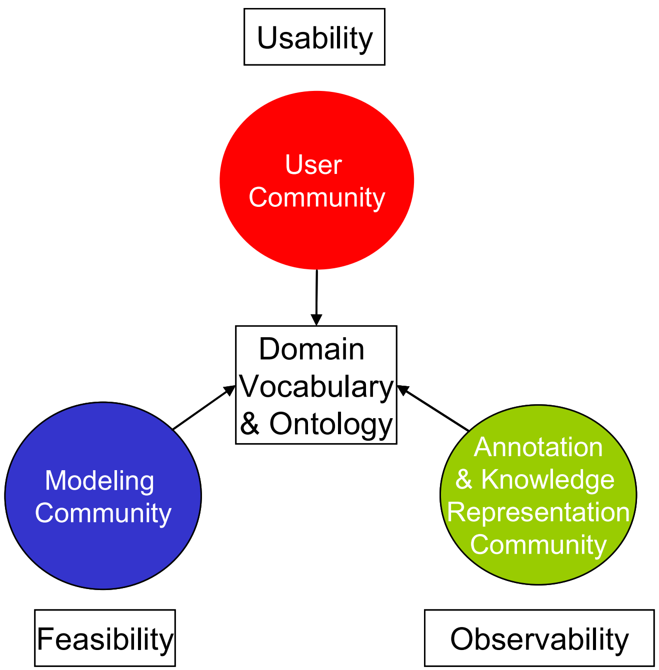
naphade@us.ibm.com jsmith@us.ibm.com alex@cs.cmu.edu sfchang@ee.columbia.edu

June 27 2006

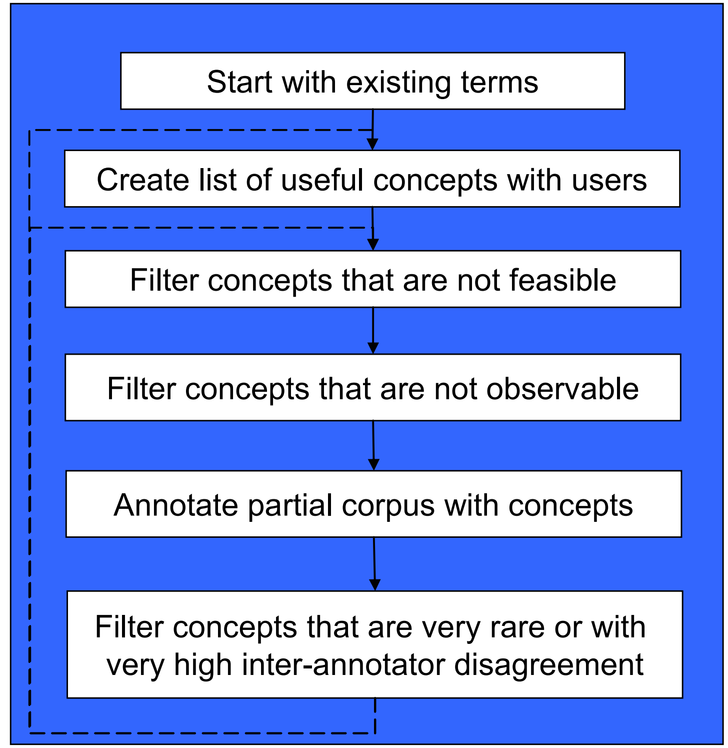
NRRC

MITRE

Goal and Vision



Workflow



Video Analysis Content Extraction

Deliverables

- 1000+ concept lexicon
- Annotated corpus
- 39 Use Cases and 250 + Queries
- Ontology
- Experimental Evaluation

Impact

- Largest annotated video corpus
- Leveraged at TRECVID and other fora
- LSCOM mapped into openCyC and ResearchCyC
- Dissemination at various fora for optimizing utilization leading to collaboration opportunities

Team

- 40+ experts from Multimedia Analytics, Knowledge Representation and User Community
- IBM: Milind R. Naphade, John R. Smith, Jelena Tesic
- Columbia University: Shih-Fu Chang, Lyndon Kennedy, John Kender
- CMU: Alex Hauptmann, Rong Yan
- CyC Corporation: Jon Curtis, Michael Witbrock
- Several student annotators
- DTO Champions: Dennis Moellman, Randy Paul, Paul Matthews

Mission

Video Analysis
Content Extraction

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

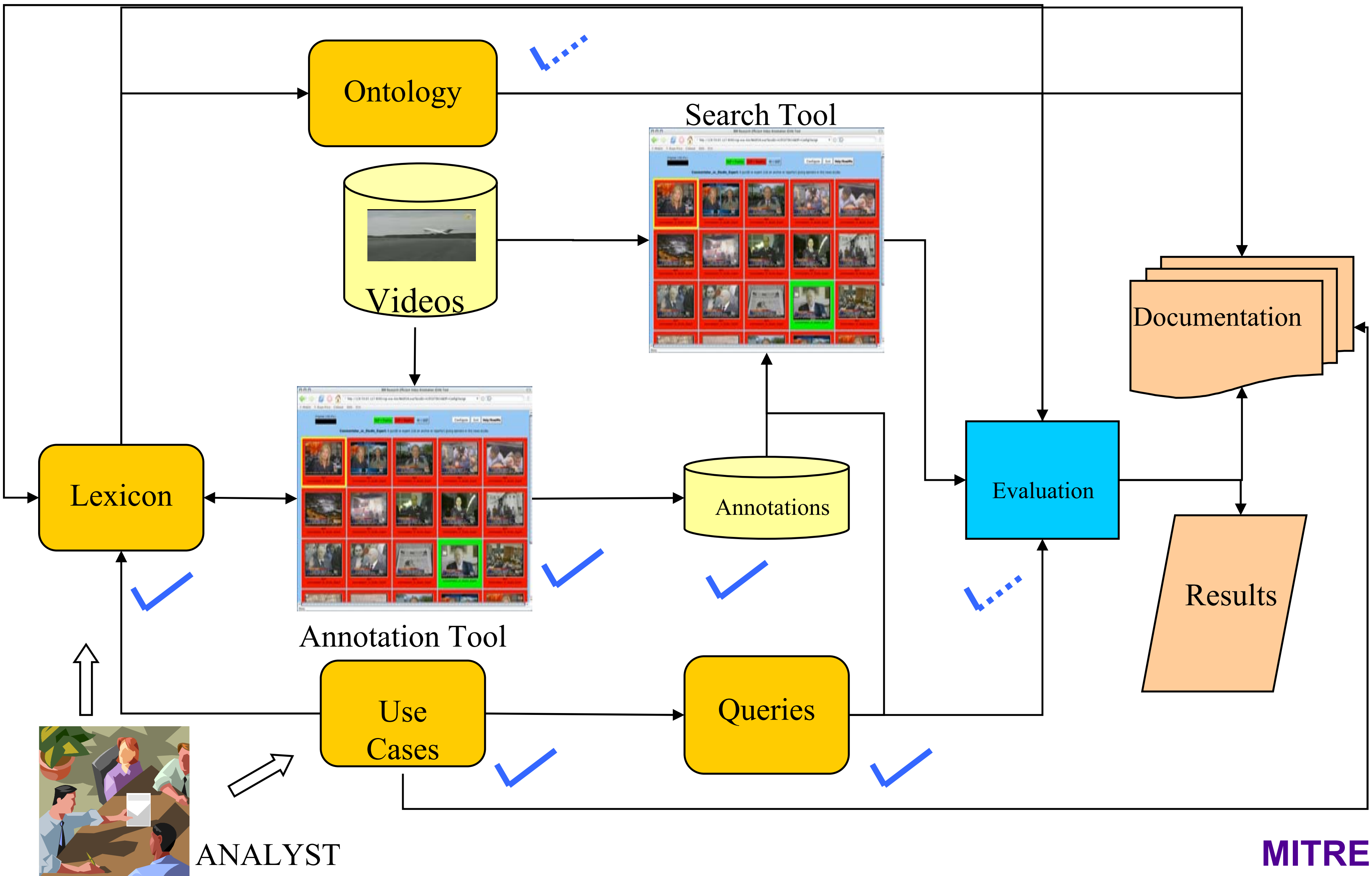
Solution:

- 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

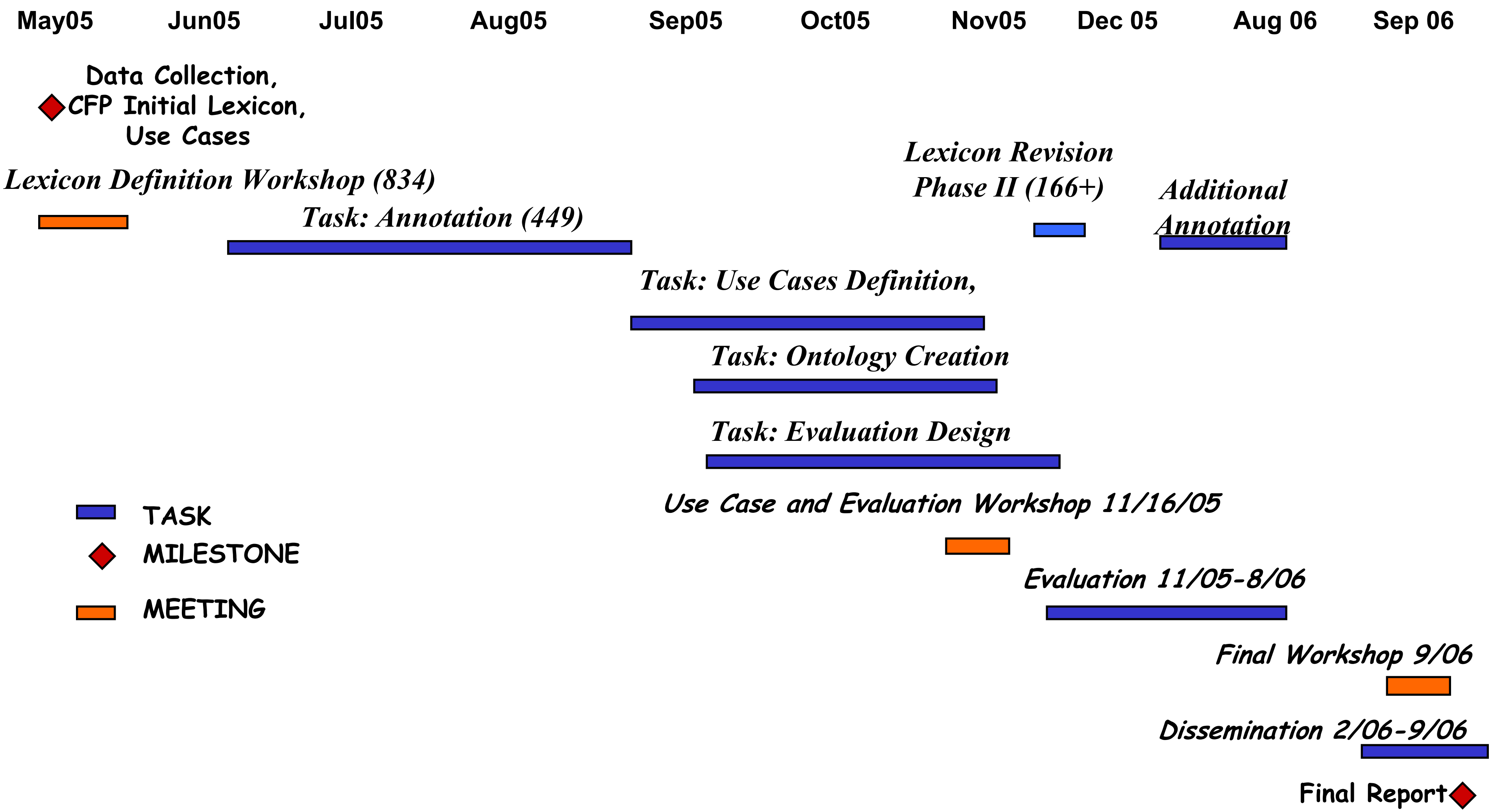
Impact:

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

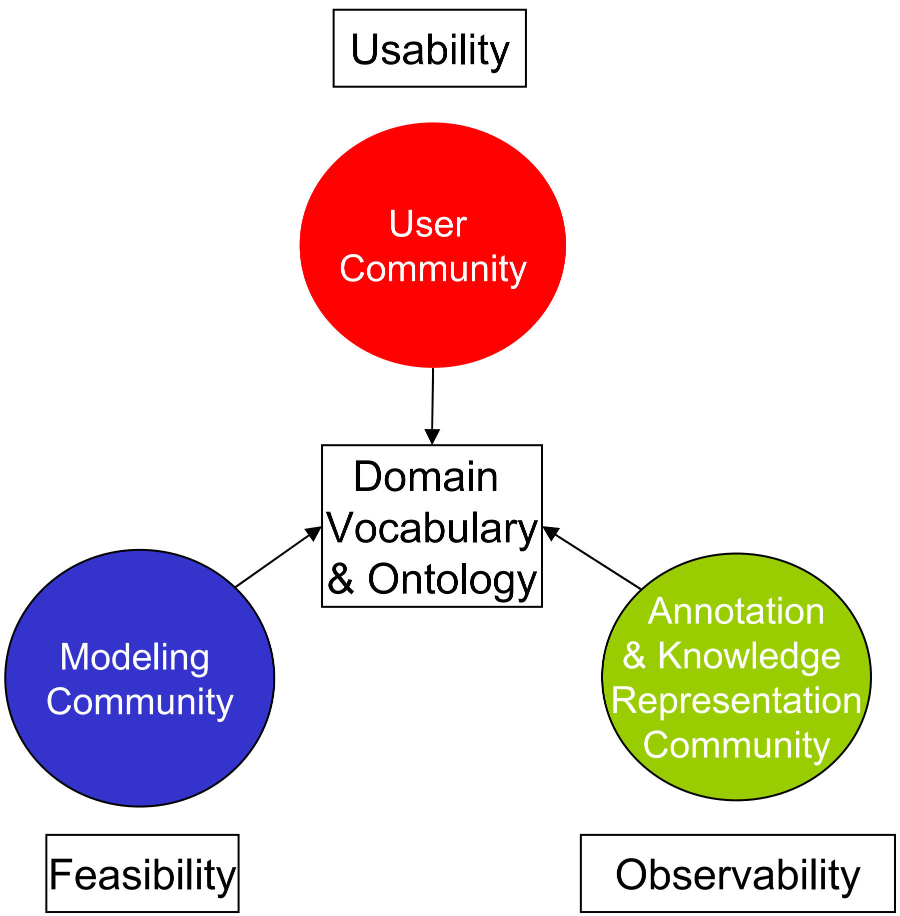
Workflow Summary

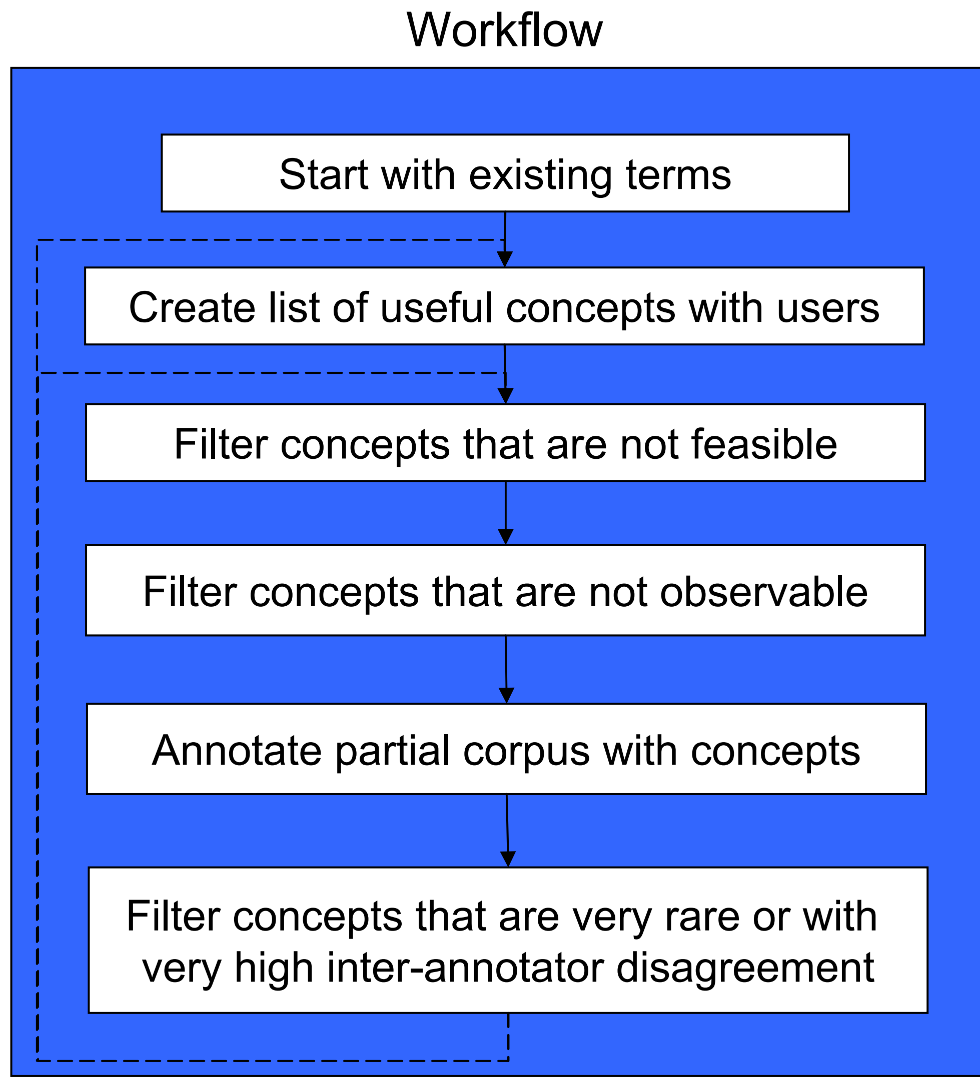


Timeline


May05
Jun05
Jul05
Aug05
Sep05
Oct05
Nov05
Dec 05
Aug 06
Sep 06
Data Collection,
CFP Initial Lexicon,
Use Cases
Lexicon Definition Workshop (834)
Task: Annotation (449)
Lexicon Revision
Phase II (166+)
Additional
Annotation
Task: Use Cases Definition,
Task: Ontology Creation
Task: Evaluation Design
Use Case and Evaluation Workshop 11/16/05
Evaluation 11/05-8/06
Final Workshop 9/06
Dissemination 2/06-9/06
Final Report
TASK
MILESTONE
MEETING

Lexicon Design Methodology





LSCOM Lexicon Design

Video Analysis
Content Extraction

DESIGN

Starter Set

Consultation

Filtering

Sample Verification

Ontology-based Expansion

Lexicon

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 corpus led to annotation of 449 unique concepts based on availability of concept in corpus and inter-annotator agreement

Mapping of LSCOM concepts into CyC and using CyC's knowledge-base for filling gaps and eliminate redundant concepts led to > 2600 concepts

LSCOM Annotation

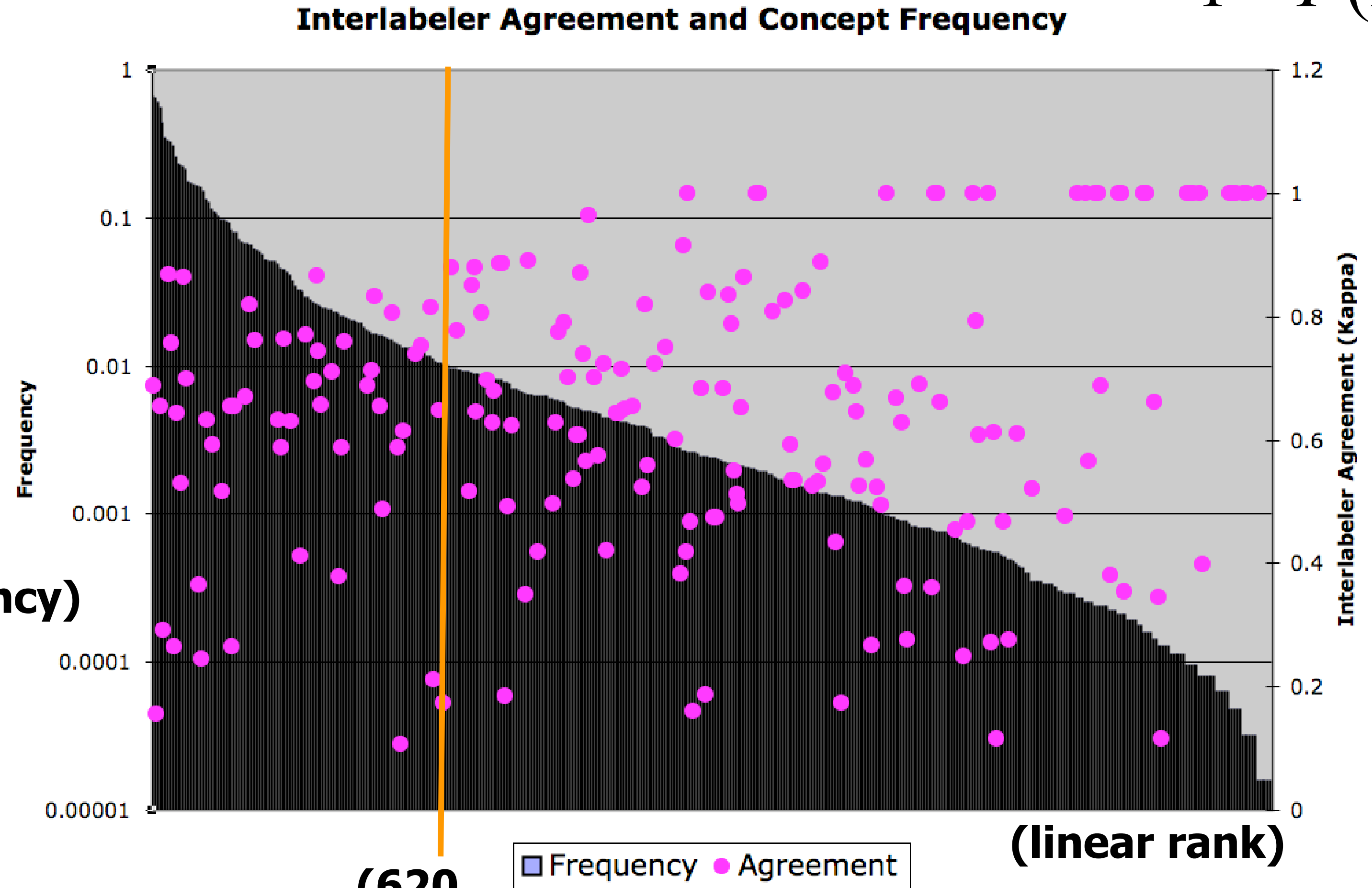
- Annotated 449 concepts using CMU and IBM annotation tool that had some presence in evaluation corpus
- Each of 74,000 shot keyframes from 80+ hours of video in the broadcast news corpus was examined for presence/absence of the concept
- Refinement of annotation for events is ongoing at CU
- Also annotated queries defined based on use cases

Annotation Quality

Video Analysis Content Extraction

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

Total # subshots ~62K



(log frequency)

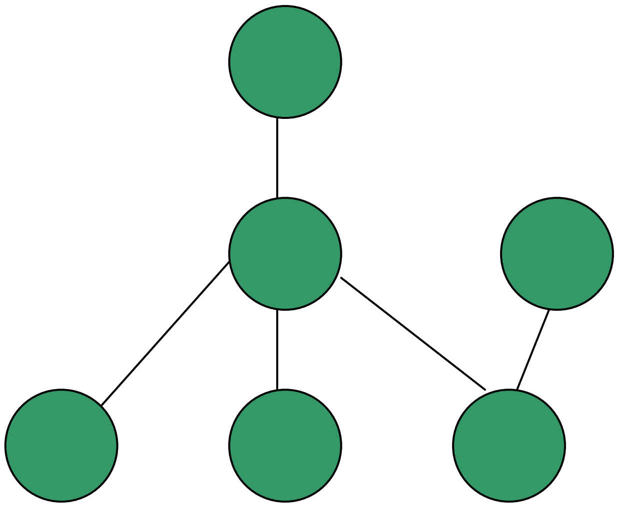
(linear rank)

(620 subshots)

MITRE

Ontology Design with CyC

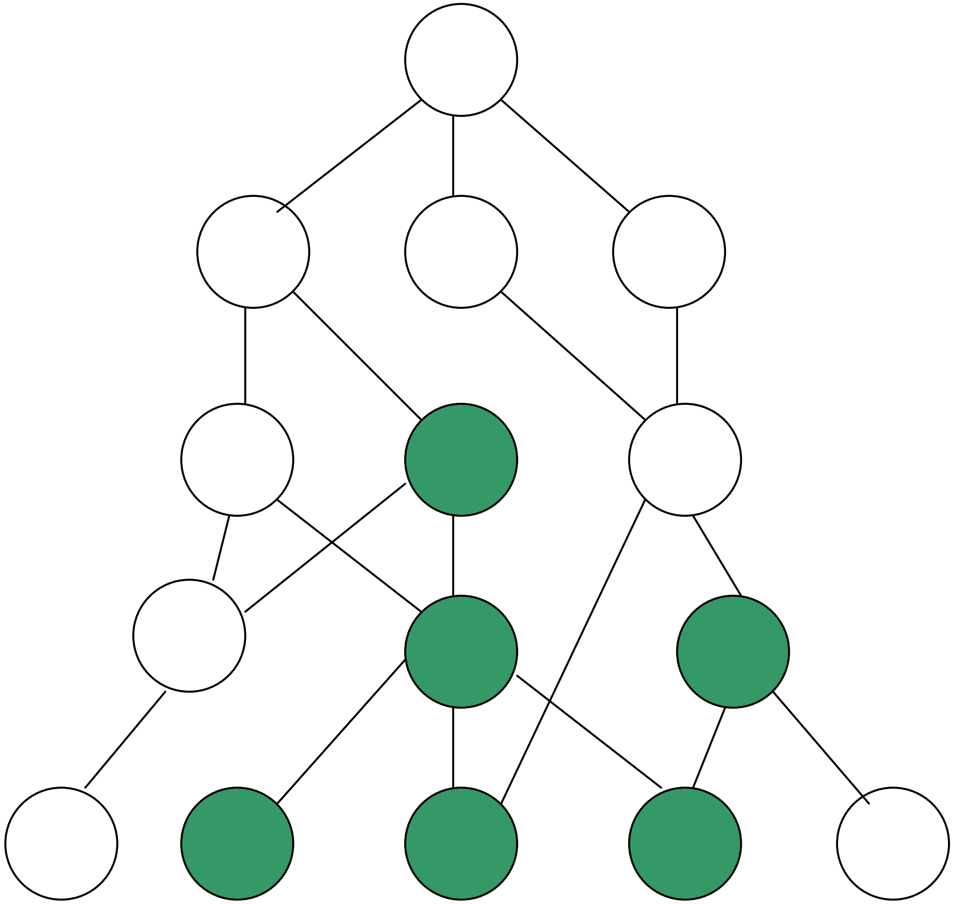
Use Cyc to extend **Breadth** and **Depth**



- More and Richer Distinctions
- Achieved Semi-Automatically
- Result: LSCOM = Cyc's First-Order Upward Closure of the Leaf Nodes

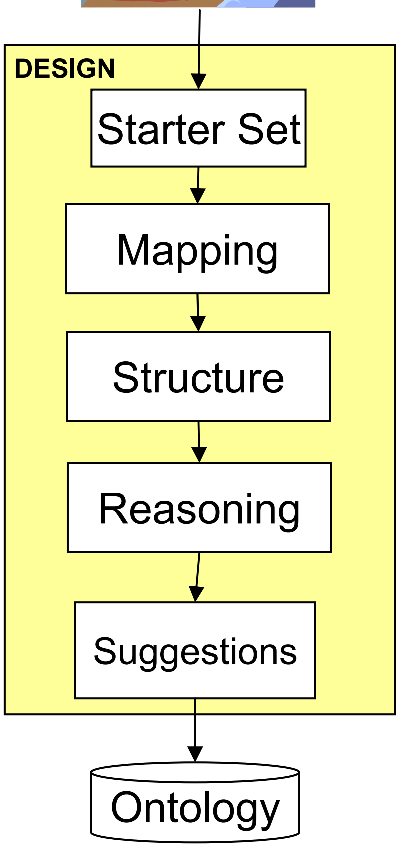
Ontology Design with CyC

Use Cyc to extend **Breadth** and **Depth**



- More and Richer Distinctions
- Achieved Semi-Automatically
- Result: LSCOM = Cyc's First-Order Upward Closure of the Leaf Nodes

Ontology Design with CyC



Starting with 834 concepts from first LSCOM Workshop

834 Concepts from original LSCOM lexiocn

Mapped LSCOM starter set into CyC

Then Mapped the graphical structure

CyC can now reason about the taxonomy

CyC also makes post-mapping suggestions for better alignment of nodes of the ontology and fills gaps and removing duplicates

CyCling LSCOM we went from **784** concepts with **763** leaf nodes to **2556** nodes with **1284** leaf nodes

Use Cases

- Needed to factor in user requirements without being too specific and capture broad user context with examples
- Needed to drive the evaluation through the expansion of use cases into TRECVID like topical queries
- Designed over 39 use cases based on events that occurred in the time-frame corresponding to the corpus capture dates provided by senior DIA Analyst
- Worked with senior DIA Analyst to validate utiity of the use cases
- Manually expanded the 39 use cases into 400+ TRECVID like queries which look to find specific information content.
- Use cases were mapped to a number of queries ranging from 5 to 30+
- Combined and collapsed the 400+ queries defined into 250+ distinct queries
- Of the 250+ distinct queries, partially annotated 50+ queries with maximum support in the corpus for evaluation

Use Cases

- Need to assess LSCOM with respect to usefulness
- "Use cases" provide a scenario of information need
- Anticipate up to twenty scenarios
- Users:
  - Intelligence analysts (see later slides)
  - Broadcast clients video archive
- Types of video archive use
  - Repeated stories ("evergreens")
  - Going back to file footage
    - How often was this shown, when/who showed it first
    - Were there multiple feeds from different perspectives
    - When was the last time this person was seen
- Examples:
  - Housing starts
    - Need buildings, construction, about this "theme"
  - Press announcements prompt search of archives
    - Get (e.g. Pentagon) stock footage before/after some event
    - Get weapon systems footage

Use Cases

- Need to assess LSCOM with respect to usefulness
- "Use cases" provide a scenario of information need
- Anticipate up to twenty scenarios
- Users:
  - Intelligence analysts (see later slides)
  - Broadcast clients video archive
- Types of video archive use
  - Repeated stories ("evergreens")
  - Going back to file footage
    - How often was this shown, when/who showed it first
    - Were there multiple feeds from different perspectives
    - When was the last time this person was seen
- Examples:
  - Housing starts
    - Need buildings, construction, about this "theme"
  - Press announcements prompt search of archives
    - Get (e.g. Pentagon) stock footage before/after some event
    - Get weapon systems footage

Possible Scenarios from ’04 Events

- Military/Terrorism:
  - Afghan – Battles; Disarm-Demob-ReIntegrate
  - Iraq – Fallujah; car bombs/IEDs; assassinations; collateral damage
  - GWOT – Oil LOC attacks to increase
  - Eritrea – War by Proxy
  - Africa – lots of conflicts
  - Pakistan – Terrorist attacks
  - Cote d’Ivoire – Internal conflict
  - Saudi Arabia – terrorism attacks mount
  - Egypt – Taba suicide bombers
  - Israel – Hezbollah fly UAVs

Possible Scenarios from '04 Events

- Iran – weapons testing; fast boats; nuclear dispute
- Syria/Lebanon – conflicts
- China – Taiwan conflicts; force extensions
- Russia – Return to Sea; Chechnya conflicts
- Balkans – force handovers
- Sudan – Darfur conflicts
- Israeli-Palestine – forever war
- Congo - Civil war
- India – AKULA-class attack sub purchase
- OBL tape promises severe US et al violence

• Political [lots of elections]:
- Ramadan timeframe activities
- Afghan – 1st direct Presidential election
- Somalia – New interim President chosen > warlords

Possible Scenarios from ’04 Events

- Cambodia – New King chosen
- Myanmar – Lt Gen replaces Gen as PM Iraq – Election prep
- Palestine – Arafat dies; successorship turmoil
- Indonesia – 1^{st} direct Presidential election
- USA – President re-elected
- Uruguay – leftist President elected
- Ukraine – Turmoil over elections
- Belarus – Presidential timeframe extended
- Burundi – elections postponed
- Argentina – China offers $$$$ influence
- Chile – Compensation promised
- Australia – PM re-elected
- Africa – Great Lakes Regional Leadership

Few Scenarios from '04 Events

- Military/Terrorism:
  - Afghan – Battles; Disarm-Demob-ReIntegrate
  - Iraq – Fallujah; car bombs/IEDs; assassinations; collateral damage
  - GWOT – Oil LOC attacks to increase
  - Eritrea – War by Proxy
  - Africa – lots of conflicts
  - Pakistan – Terrorist attacks
  - Cote d'Ivoire – Internal conflict
  - Saudi Arabia – terrorism attacks mount
  - Egypt – Taba suicide bombers
  - Israel – Hezbollah fly UAVs

Use Case to Queries Expansion: Aghan battles, demobilization and disarmament

Video Analysis
Content Extraction

Battles/Violence in Mountains	Convoy of several vehicles on makeshift roads
Landmines exploding in barren landscapes	Empty Streets with buildings in state of dilapidation
Masked Gunmen	Groups of People commenting on the terrorism
Camps with Masked Gumen without uniforms	Map of Afghanistan with Kandahar and Kabul shown
Armored Vehicles driving through barren landscapes	Afghan flag atop building
Mountainous scenes with openings of caves visible	Scenes from the meetings of political leaders
People wearing turbans with Missile Launchers	Militia with guns firing across mountains
Group of People with Pile of Weapons	Men in black Afghan dresses with weapons exercising with bunkers in the background
Refugee Camps with women and children visible	Military personnel watching battlefield with binoculars
Political Leaders making speeches or meeting with people	Series of explosions in hilly terrain
Predator Drone flying over mountainous landscape	Man firing soldier fired missile in air
Munitions being dropped from aircrafts over landscape	Incarcerated people in makeshift jail
Munitions being dropped from aircrafts in mountains	Funeral procession of young victims of bombing
Dead People and Injured people	Afghan warlords with weapon carrying bodyguards in a village meeting discussing strategy and tactics
Bearded Man speaking on Satellite phones in mountainous landscape	

DTO

Video Analysis Content Extraction

Annotation Use-case Queries with multi-modal search

http://www.ee.columbia.edu/cuvidsearch

Evaluation

- Evaluation of lexicon coverage through expansion of use case queries into LSCOM for coverage analysis and gap analysis
- Evaluation of retrieval effectiveness using baseline search for benchmark queries and comparison with baseline + LSCOM search.
- Evaluation of lexicon by mapping LSCOM into openCyC and querying the openCyC to find redundancy/gaps and help fill these gaps
- Evaluation of the lexicon using tests such as Zipf’s law, collocation and negative mutual information analysis

Evaluation Methods II

- Evaluating confirmity with Zipf's law about mature vocabulary: Probability of use inversely proportional to rank. Violations of Zipf's law will show if a set of concepts has too many or too few generic concepts relative to more specific ones. It can also indicate how many generic concepts to delete or how many specific concepts to add.
- Collocation: Indicated by higher than chance co-occurance of two concepts in same frame or episode. Usually one concept out of the pair can be dropped, or the two can be combined into a single new one
- Negative mutual information helps find what true variability does occur, by showing the opposite sides of some dimension, or two non-mergable branches of the semantic tree (e.g. "text"-"outdoors", "face"-"graphics", "vegetation"-"indoors", etc.)
- Descriptions of settings usually most useful for episode discrimination relative to categorization of episodes. Missing background descriptions can be found by noting episodes having no background description at all

Evaluation Methods

- Require benchmarks and metrics for evaluating:
  - Utility of ontology – coverage of queries in terms of quality and quantity
- Metrics of Retrieval Effectiveness
  - Precision & Recall Curves, Average Precision, Precision at Fixed Depth
- Metrics of Lexicon Effectiveness
  - Number of Use Case Queries that are answered by lexicon successfully
  - Mean average precision across the set of use case queries
- This will be achieved by automatic/semi-automatic mapping of use case queries into LSCOM lexicon
- The expanded concepts will then be used to return shot lists that can be evaluated for retrieval effectiveness

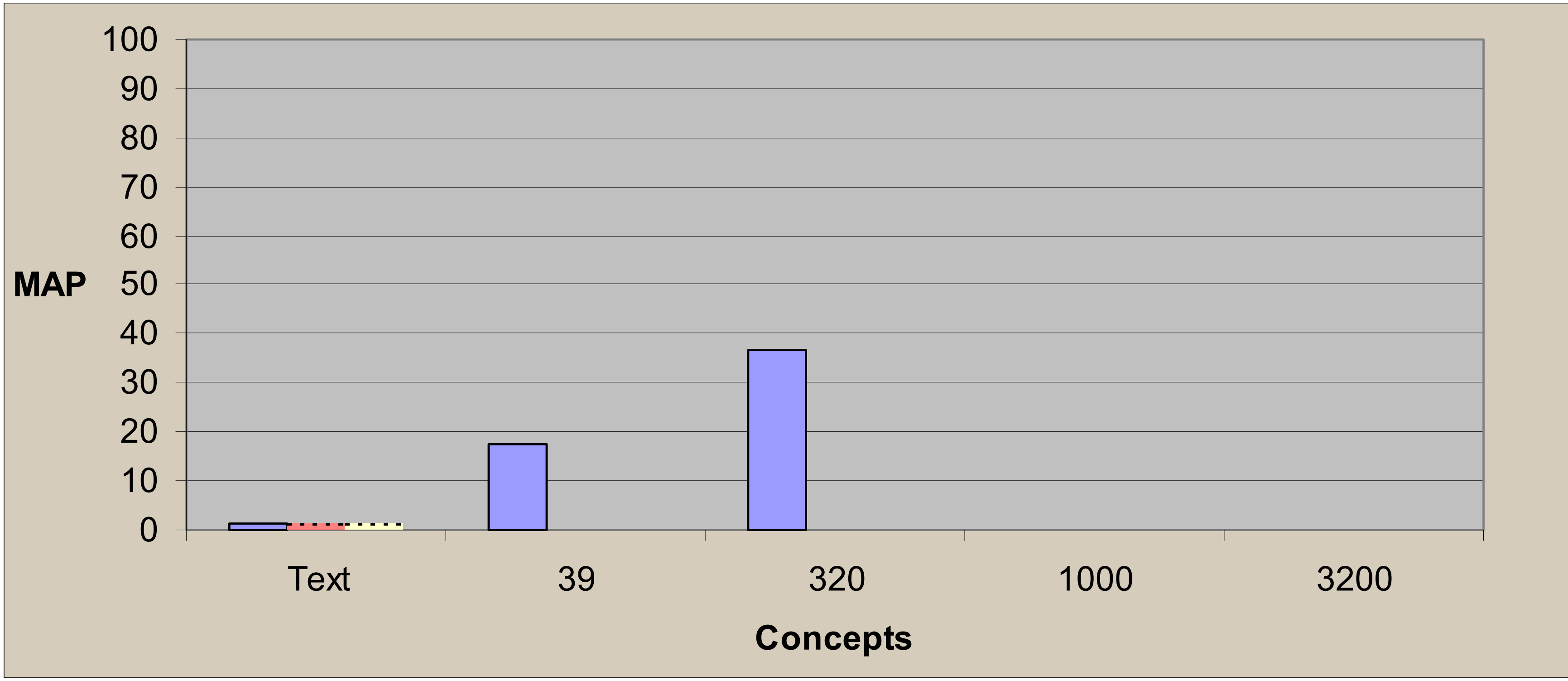
ARDA

Long Term Evaluation Goal

- For exhaustive judgements on whether each lexicon entry deserves its place in the list extensive testing will be required
- Wish List
  - Large Query Log from DIA/FBIS of around 10000 queries
  - Trial use of lexicon for annotation at one of the agencies to validate utility and coverage
  - Use of lexicon in other tasks and domains to analyze cross domain utility
  - Leveraging the TRECVID community to build detectors for various concepts in the lexicon
  - Iterative refinement of the lexicon based on on at least one cycle of definition->validation->utility measurement

MITRE

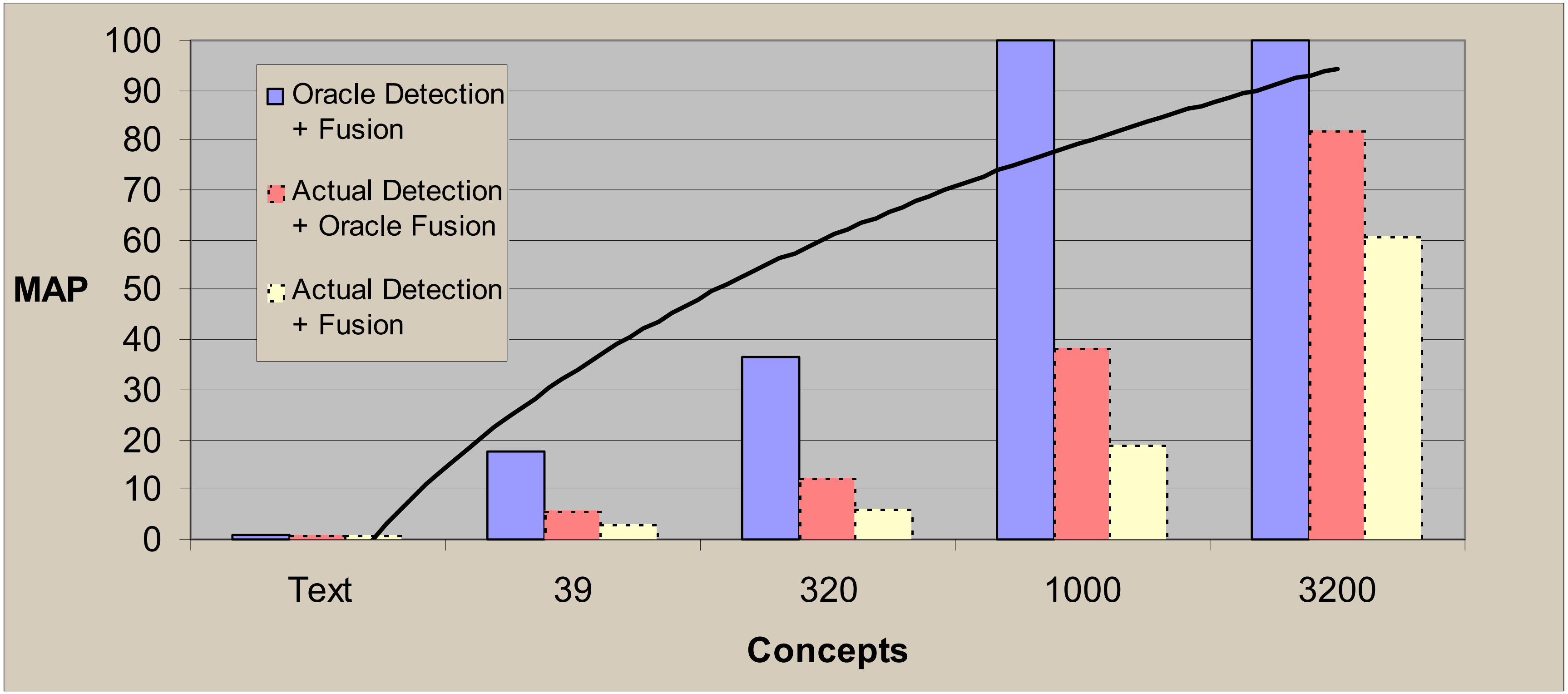
Preliminary Evaluation



LSCOM-based retrieval (based on ~ 75 annotated queries) using oracle detection and fusion is significantly (30x) better than baseline (text) as well as LSCOM-lite

Preliminary Evaluation & Emerging Trends

Video Analysis
Content Extraction



Trends* indicate that a few thousand concepts with state of the art detection and fusion can get very high search accuracy

Assumptions * Auto detection 1/3 as good as manual
Assumptions* Auto query expansion & fusion ½ as good as manual

Impact

- Adoption of LSCOM by TRECVID (already achieved for TRECVID 2006 cycle) thus opening the experimentation cycle to hundreds of researchers worldwide. 60 downloads so far
- Synergy with research networks such as DELOS-MUSCLE European network of excellence. Efforts underway to share LSCOM annotations and ontology
- LSCOM will be part of OpenCyC and ResearchCyC thus creating a win-win for both LSCOM and CyC and making LSCOM available to the CyC user community

Future Directions

- Baseline maintenance and update site being discussed
- Trial use of lexicon for annotation at one of the agencies to validate utility and coverage will be beneficial
- Work to realize LSCOM potential is just starting

Case Study: MARVEL

MARVEL in a NUTSHELL

- **What is Marvel?**
  - Novel system for indexing and search of digital media content
- **How does it work?**
  - Models semantic concepts using visual, audio and speech modalities
  - Applies models to extract semantic concepts (scenes, objects, events, people, sites)
  - Builds models from training examples (can exploit pre-existing catalogs and taxonomies)
- **What are the benefits?**
  - Enhances traditional metadata- and speech-based indexing and search
  - Reduces costs of semantic-based indexing of digital media content
  - Increases asset reuse by providing standards-based semantic search capabilities
  - Enables new models of consumer-oriented content distribution
- **What does deployment require**?
  - One-time efforts:
    - Definition of concept ontology for domain(s) of interest (e.g., news, sports, movies)
    - Building of models from training examples
  - On-going processing:
    - Automated indexing of new content using models

THE MARVEL STACK

MULTIMEDIA SOLUTIONS- MARVEL

[Search Engines, Repositories, Filters, Personalization, Content-based Routing Mining, Benchmarking]

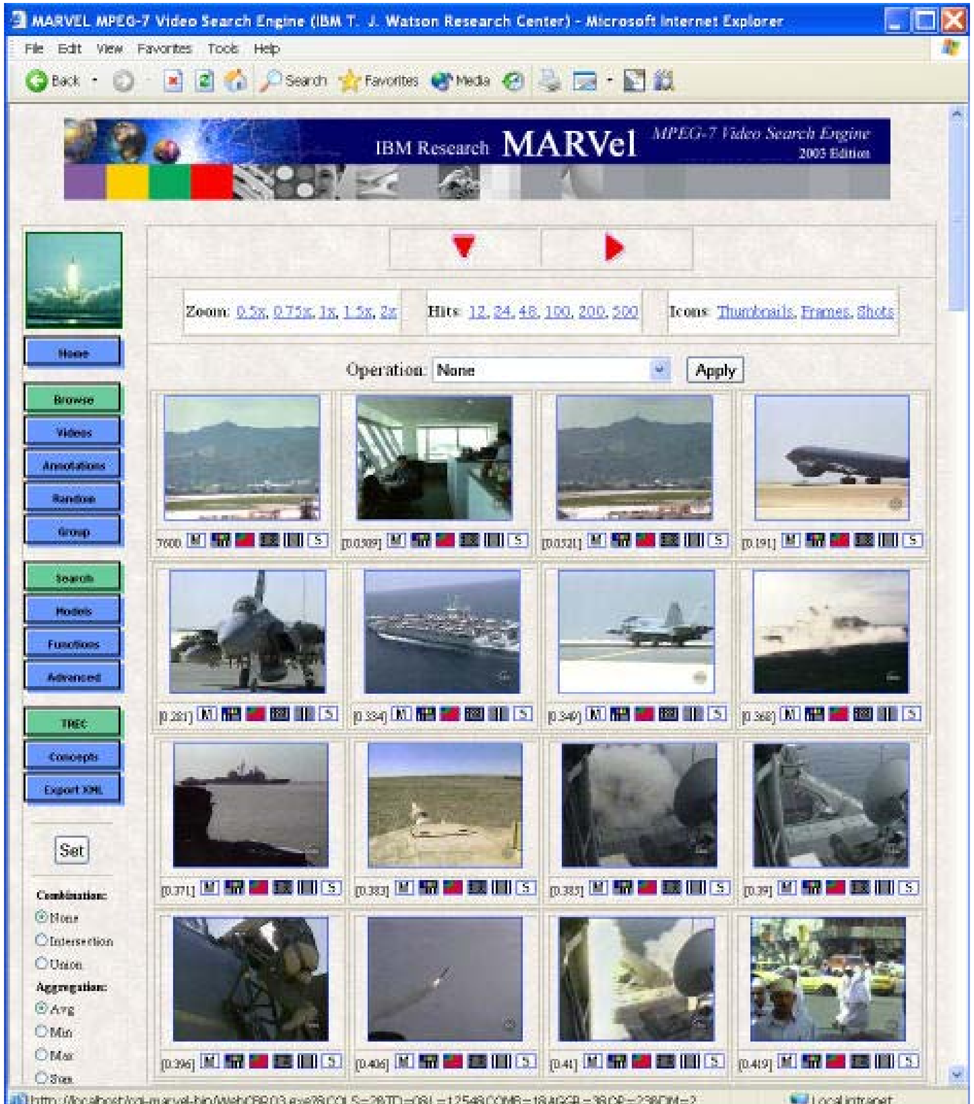
STANDARDIZED LEXICON & ONTOLOGY [LSCOM]

STANDARDIZED METADATA DESCRIPTION LANGUAGE [MPEG-7, VEML, SMPTE, etc.]

STANDARDIZED STRUCTURES FOR ACCESS
[Keyframes, Shots, etc.]

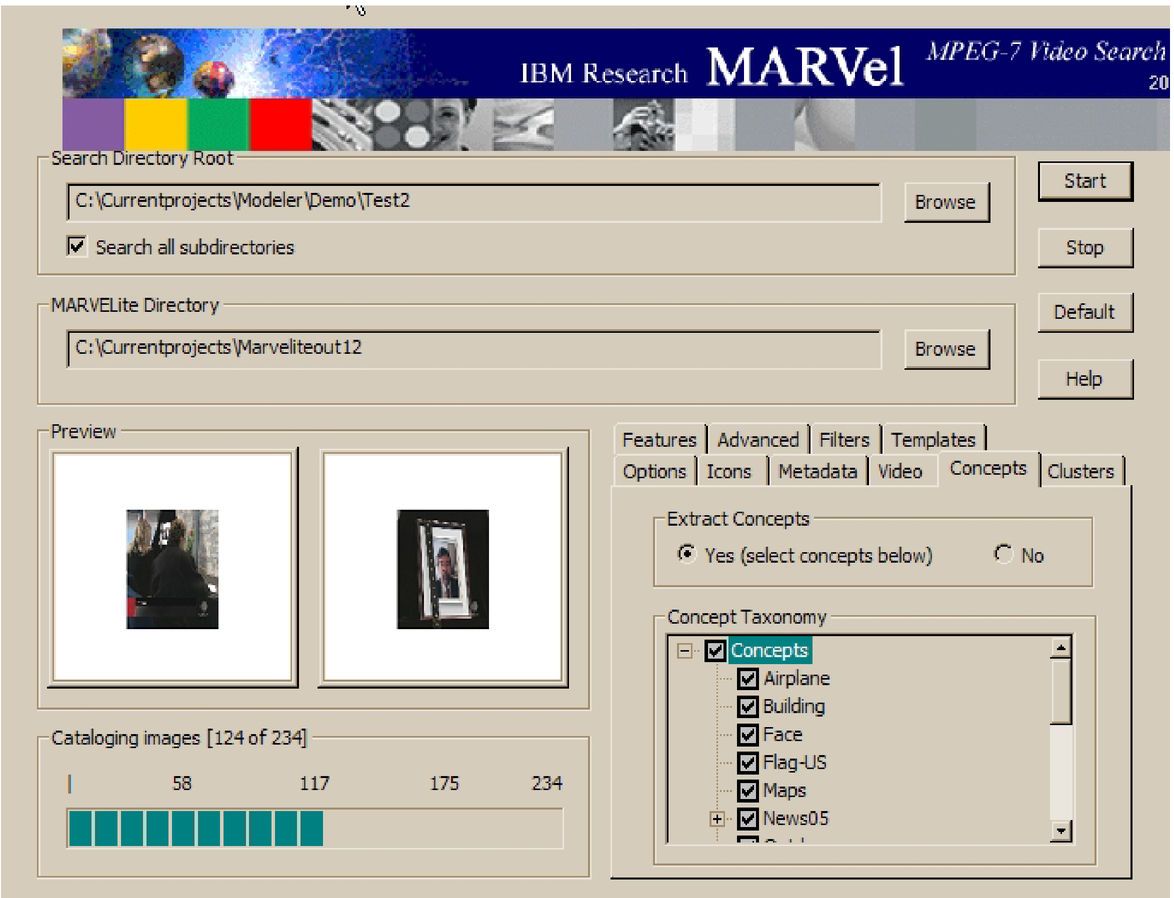
UNSTRUCTURED MULTIMEDIA CONTENT [Broadcast News, Movies, Handheld Videos, Web Video Blogs, Surveillance, etc.]

MARVEL



- MPEG-7 Video Search Engine
- Automatic indexing:
  - Shot detection/key-frame extraction
  - Feature Extraction
  - Semantic Concept Detection
- Search methods:
  - **Model-based retrieval (MBR)** – statistical modeling and detection of semantic concepts - faces, people, outdoors, etc.
  - **Content-based retrieval (CBR)** - color, texture, edges, etc.
  - **Text-based retrieval (TBR)** – textual metadata, annotations, speech transcript
  - **Model-vector based retrieval (MVBR)** = MBR + CBR
- Interaction:
  - Multi-example relevance feedback searching
  - * Iterative searching (combination methods and aggregation functions)
- On-line demo:
  - http://mp7.watson.ibm.com

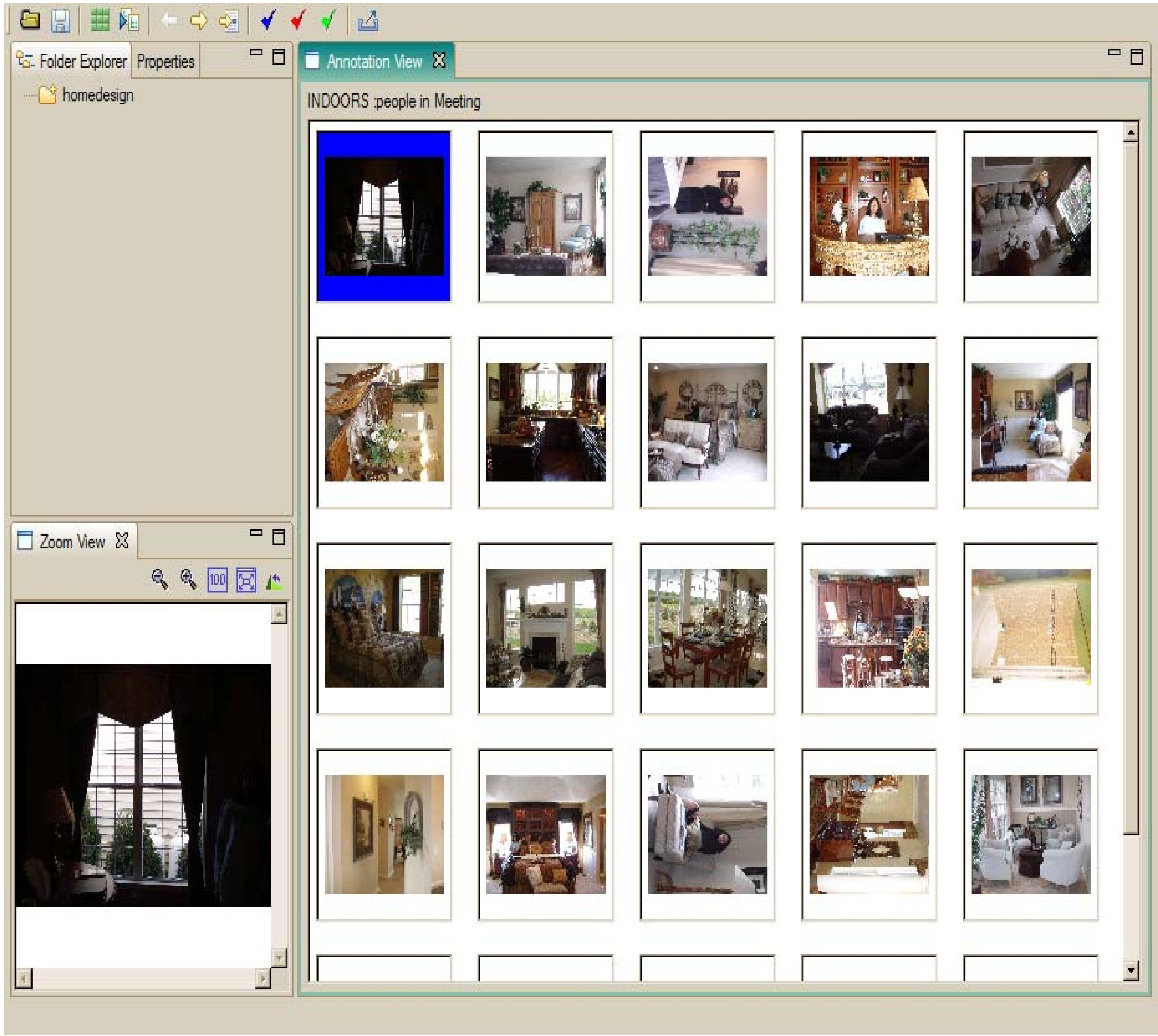
MARVELITE



http://www.alphaworks.ibm.com/tech/marvel

- Interactive image and video analysis tool available for free trial usage
- Comes with several low level features and a few high level semantic features such as Outdoors, Face, Sky, etc.
- 150 Downloads so far

MARVEL MODELER



- Interactive annotation and modeling tool to be released later this year

MARVEL References

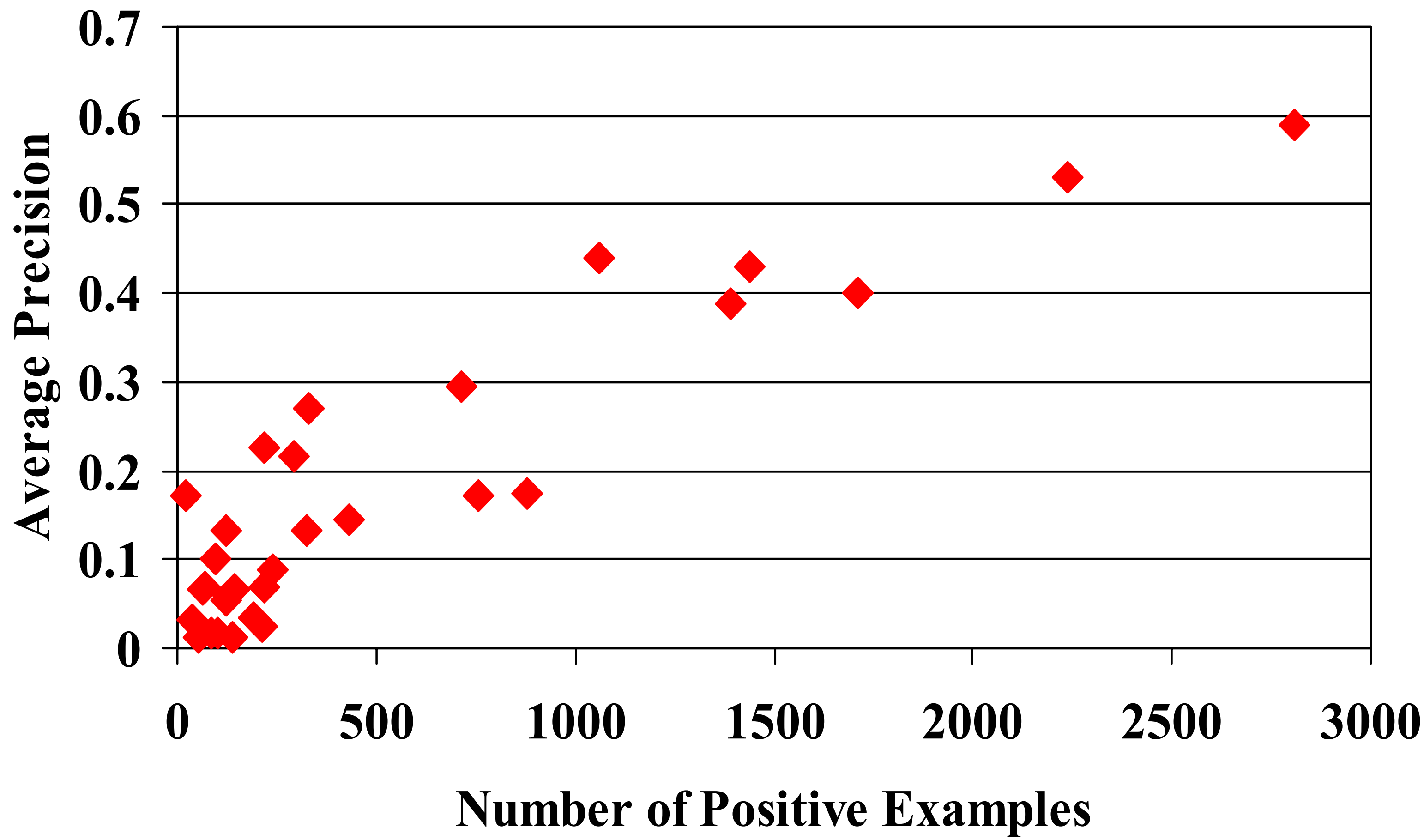
- **Awards and News:**
  - IBM MARVEL received Wall Street Journal 2004 Innovation Award (Nov. 2004):
    - http://www.wsj.com
  - "Search Looks at the Big Picture", Wired News (Jan. 6, 2005)
    - http://www.wired.com/news/technology/0,1282,66185,00.html
  - Article in c/Net and zdnet (Sept. 2004):
    - http://news.com.com/IBMs+Marvel+to+scour+Net+for+video%2C+audio/2100-1025_3-5388718.html
    - http://news.zdnet.com/2100-9596_22-5388718.html
  - Information Week article (Aug. 2004):
    - http://www.informationweek.com/story/showArticle.jhtml?articleID=43200005

- **Demos and Tools:**
  - IBM Research Marvel “lite”
    - http://www.alphaworks.ibm.com/tech/marvel
  - IBM MARVEL MPEG-7 Multimedia Analysis and Retrieval System:
    - http://www.research.ibm.com/marvel
  - IBM MPEG-7 Video Annotation Tool:
    - http://www.alphaworks.ibm.com/tech/videoannex

- **Links:**
  - IBM Research Intelligent Information Management Department:
    - http://www.research.ibm.com/iim
  - IBM Research Marvel project page:
    - http://www.research.ibm.com/marvel
  - IBM Research SLAM - Semantic Learning and Analysis of Multimedia:
    - http://www.research.ibm.com/slam

Challenges and Gaps

Learning Rare Concepts is a Challenge



Bridging of the semantic gap depends on trai

Madeleine Albright
Newt Gingrich
Physical Violence
Riot

Bill Clinton
Sam Donaldson
Flower

Car
Cloud
Crowd
Rock
Truck
Beach
Mountain
Road

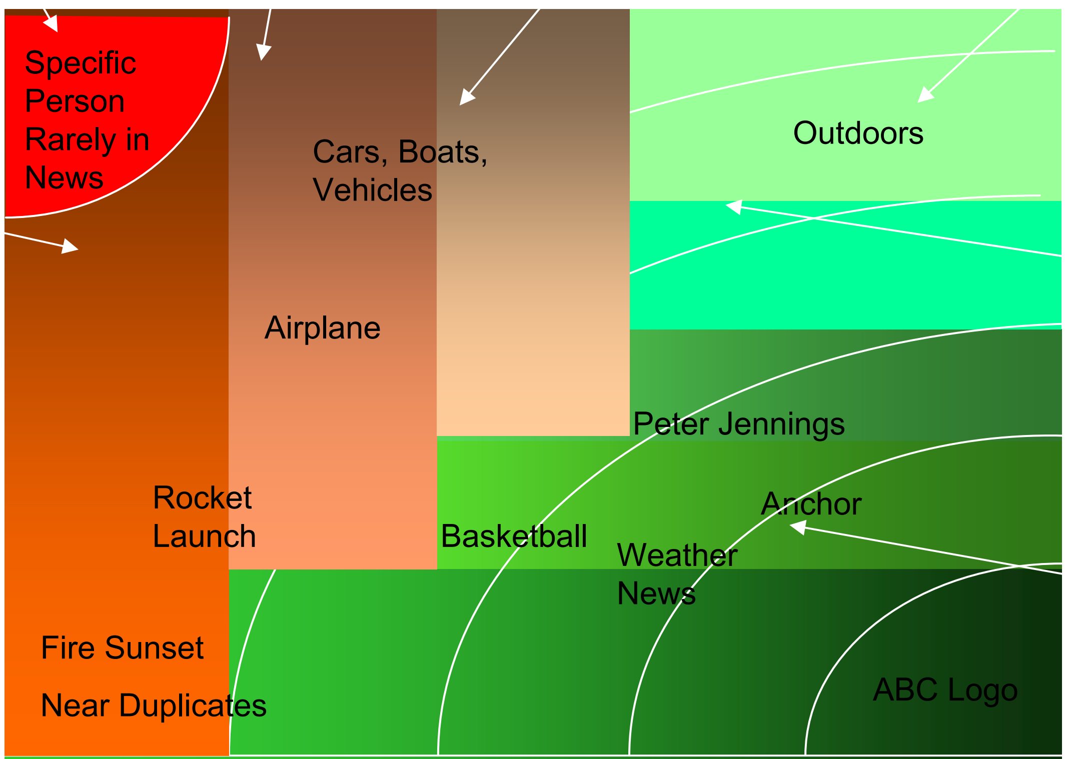
Outdoors,
Indoors
People
Human
Graphics & Text
Nature Vegetation
Person
Face
Male Face
Female Face
Sky
Cartoon

Train

Wide

Semantic Gap

Narrow



Rare Training Sample Availability **Frequent**

Cityscape
Man Made Scene
Building
People Event

Basketball
Baseball
Tennis
Weather News
Sports Event
Studio Setting
Golf
Desert

INDEX

- Easy and Abundant
- Easy but Rare
- Hard but Abundant
- Hardest and Rare

Requirements

- Accuracy:

  Need to Capture Spatial, Temporal, Multimodal, Conceptual dependencies
- Rare-Classes

  Need to account for few positive samples
- Active Role

  Passive is inefficient. Active is the way to go
- User-friendliness

  Help the user select, annotate, propagate retrieve and learn constantly from the user's interaction with the system at different levels.
- Knowledge Integration

  Systematic ways of incorporating domain and other knowledge/knowledge bases, interaction with NLP, ASR

Future Directions

- Need to expand the set of multimodal concepts that can be detected with greater reliability
- Learning can play a far greater role than it currently is playing in extracting semantic features
- Need to work on multimedia grammar
- Need to work on a common (perhaps open source) architecture that allows for easy plug and play of different analytics so that not every group has to reinvent every wheel and build systems from scratch.
- Need to encourage standardization of best of breed algorithms/sub-systems and focus on extracting significantly differentiating performance
- Benchmark has helped remove misconceptions and established that
  - Text analysis is not sufficient. We do need visual analysis
  - Concept detectors can be used for more complex search
  - Fixing lexica, experiments, corpora reveal significant information about what works, and more importantly, what does not..

Shifting Emphasis also apparent in Publications

- **Content Analysis:**
  - **Image/Video Classification**: Naphade (UIUC, IBM), Vailaya (Michigan State), Iyengar & Vasconcelos (MIT), Bertini and del Bimbo (Firenze), Smith (IBM), Hauptmann (CMU), Wang and Li (Penn State), Alan Hanjalic (TU Delft), Nicu Sebe (UVA), Marcel Worring (UVA)
  - Semantic Audiovisual Analysis: Naphade (UIUC), Chang (Columbia). Lienhart (U. Augsburg), Slaney (IBM, Yahoo)
- **Learning and Multimedia:**
  - **Applied Statistical Media Learning**: Frey (U Toronto), Naphade (UIUC), Forsyth (Berkeley), Fisher & Jebara (MIT), V. Iyengar (IBM).
  - **Learning in Image Retrieval**: Chang et al. (UCSB, Google), Zhang et al (Microsoft Research), Naphade et al. (UIUC) Viola et al. (MIT, MERL).
  - **Linking Clusters in Media Feature**: Barnard & Forsyth (Berkeley), Slaney (IBM).
  - Theoretical Learning: M. Jordan (UCB), Michael Kearns (U Penn), B. Frey (U Toronto), T. Jakkola (MIT)
- **Vision and Speech:**
  - **Computer Vision in Media Analysis**: Bolle (IBM), Mallik (Berkeley)
  - **Auditory Scene Analysis & Discriminant ASR Models**: Ellis (MIT), Nadas et al. (IBM), Gopalkrishnan et al (IBM), Woodland et al. (Cambridge), Naphade et al (UIUC) Wang et al (NYU), Kuo et al. (USC)
- **Learning for Retrieval:**
  - **62 groups at TRECVID led by Paul Over, Alan Smeaton and Wessel Kraaij**

Acknowledgements

- SSMS Organizers
- IBM: John R. Smith (Senior Mgr.), Chung-Sheng Li, Murray Campbell, Paul Natsev,Dipankar Datta, Jelena Tesic, Ching-Yung Lin, Giridharan Iyengar, Neti Chalapathy, Harriet Nock, Arnon Amir, Janne Argillander
- Summer interns: Rong Yan, Feng Kang, Dhiraj Joshi, Alexander Haubold,
- LSCOM: Dennis Moellman, Timothy Stormer, Randy Paul, Paul Matthews, Shih-Fu Chang, Alexander, Hauptmann, John Kender, Jonathan Curtis, Lyndon Kennedy
- UIUC: Thomas Huang, Roy Wang, Ashutosh Garg, Xian Zhou
- NIST: Paul Over, Alan Smeaton, Wessel Kraiij, Tzveta Ianeva

EFHARISTO