COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

Content-Based Multimedia Retrieval

- Lessons Learned from Two Decades of Research

Shih-Fu Chang

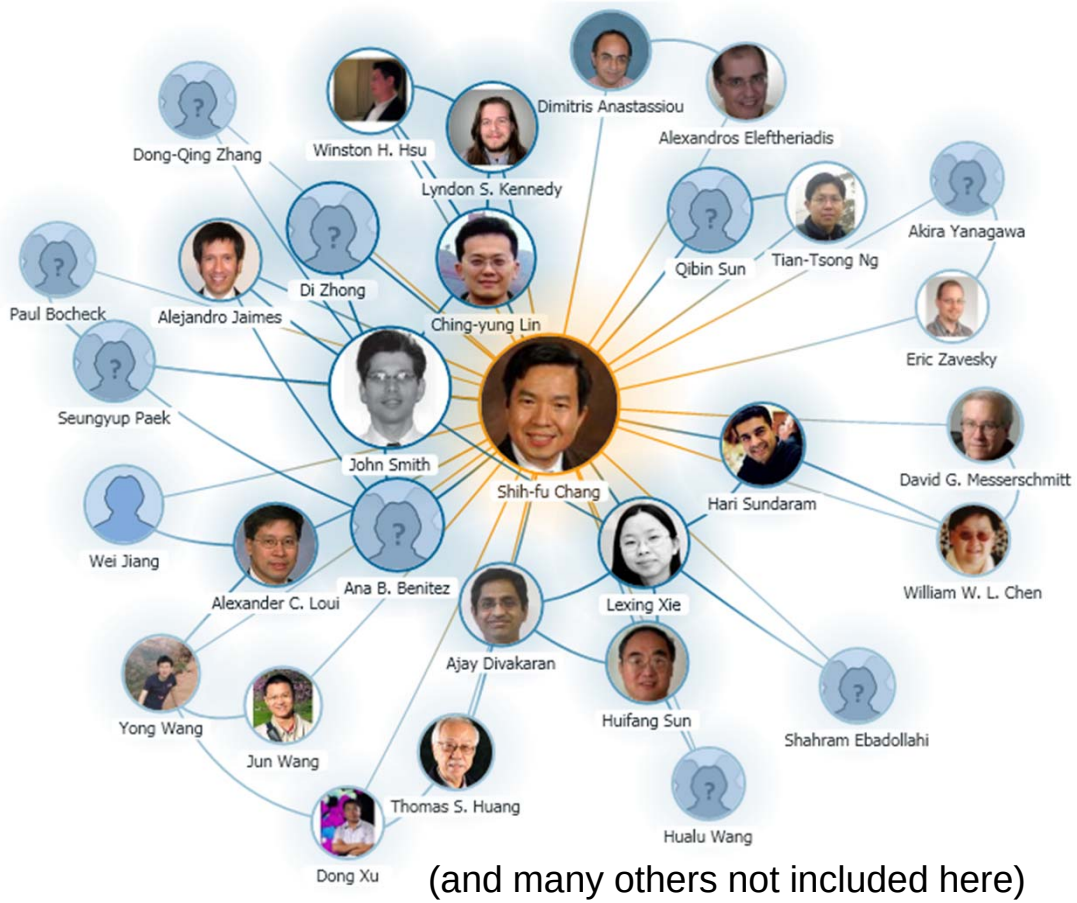
ACM SIGMM Technical Achievement Award Talk

Scottsdale, AZ, November 2011

Acknowledgement

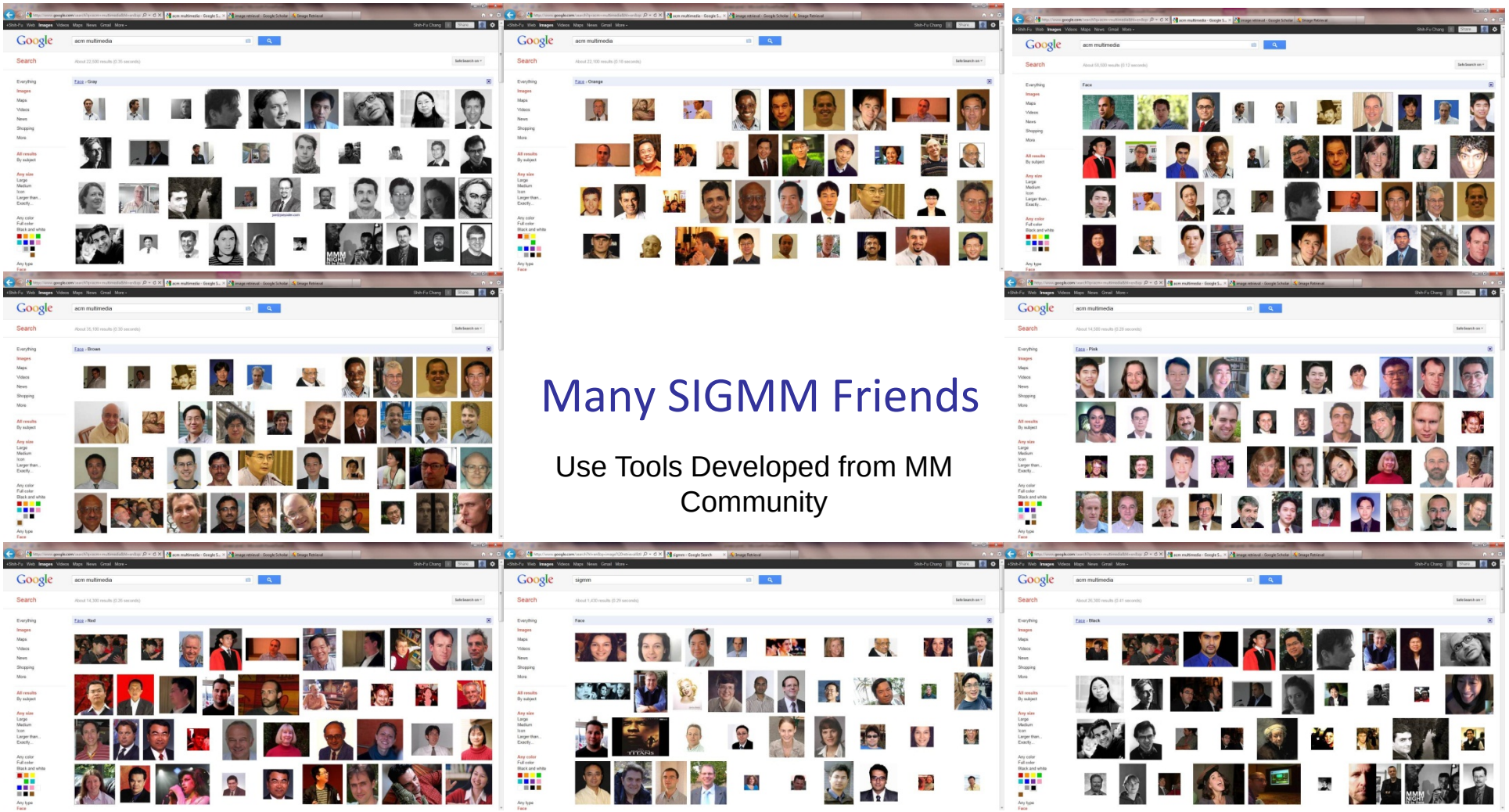
Collaborators

Microsoft Academic Search



(and many others not included here)

Sponsors

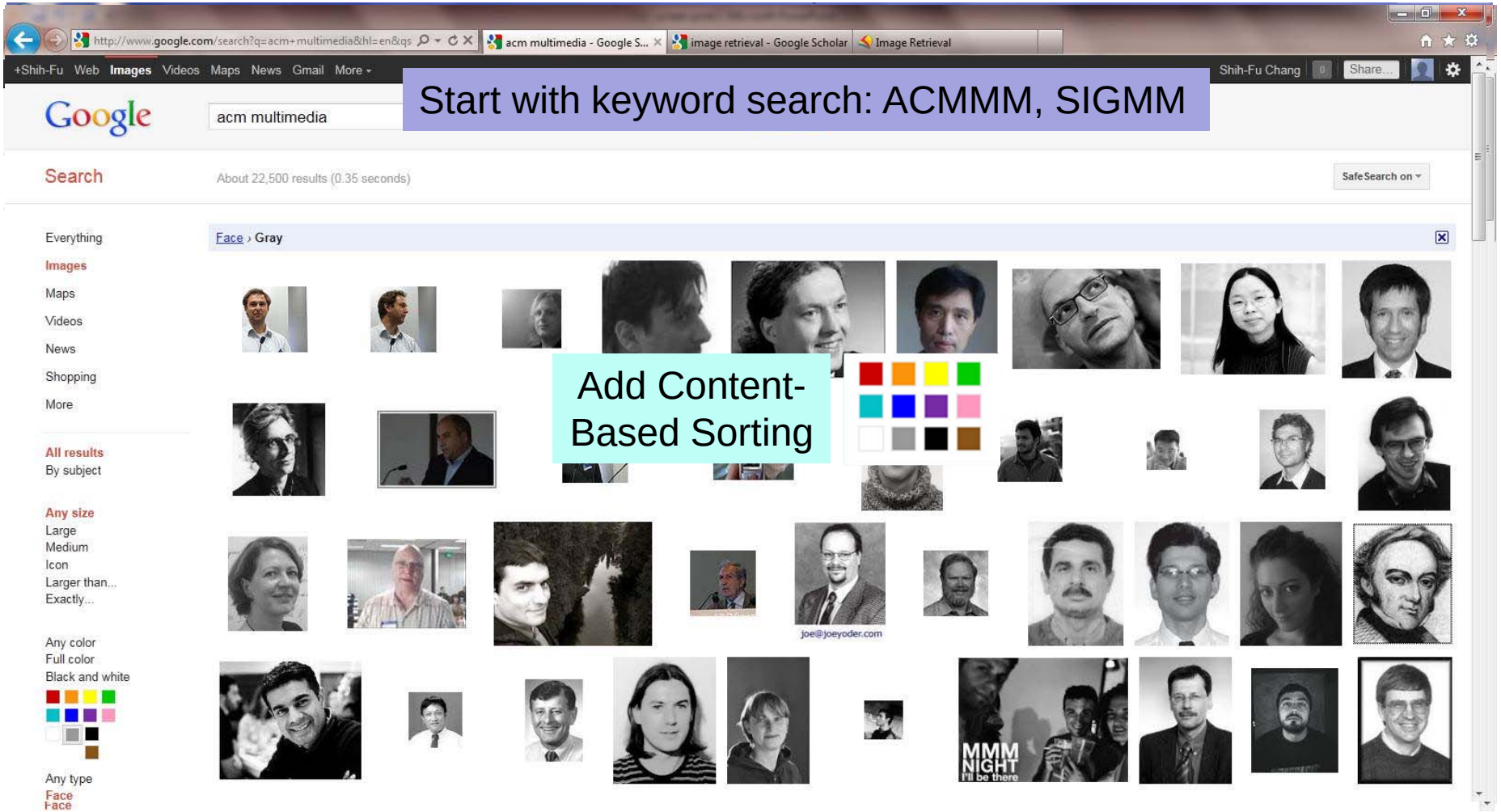


Many SIGMM Friends

Use Tools Developed from MM Community

Compiling the Community List

(a use case of multimedia search tools)



COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

Content-Based Multimedia Retrieval

- Lessons Learned from Two Decades of Research

Shih-Fu Chang

ACM SIGMM Technical Achievement Award Talk

Scottsdale, AZ, November 2011

First Digital Camera in 1975

New York Times Bits, 8/26/2010

by Steve Sassan of Kodak

- 16 batteries, new CCD array, A/D converter
- 23 seconds to record a photo to cassette
- A customized reader on a B/W TV for viewing

First Digital Camera in 1975

Questions asked by audience in 1975

- Why would anyone ever want to view his or her pictures on a TV?
- How would you store these images?
- What does an electronic photo album look like?
- When would this type of approach be available to the consumer?

What happens in 2010?

- **Images**
  - **36 billion** – Rate of photos uploaded to Facebook per year.
- **Videos**
  - **2 billion** – Number of videos watched per day on YouTube.
  - **35** – Hours of video uploaded to YouTube every minute.
- **Internet users**
  - **1.97 billion** – Internet users worldwide (June 2010).
- **Social media**
  - **30 billion** – Pieces of content (links, notes, photos, etc.) shared on Facebook per month.

A Personal Sharing Experience

- a single video shared by my family on Youtube
- 150,000+ views in 3 years
- higher than all of citations to my papers published over 20 years!

Image Storage/Retrieval: Finding Needle in Haystack

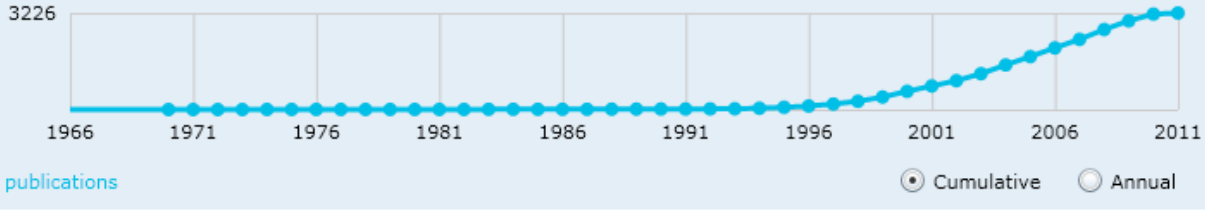
Google Scholar Search “Image Retrieval”: 2.1 million results

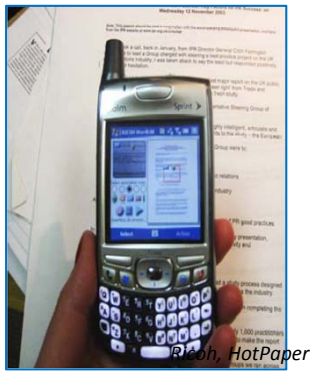
Academic > Keyword > Image Retrieval Subscribe

Image Retrieval - IR

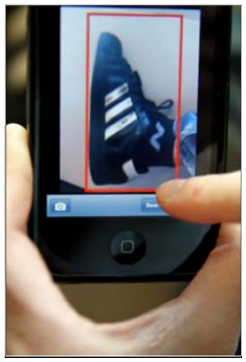
Publications: 3,830 | Citation Count: 27,052

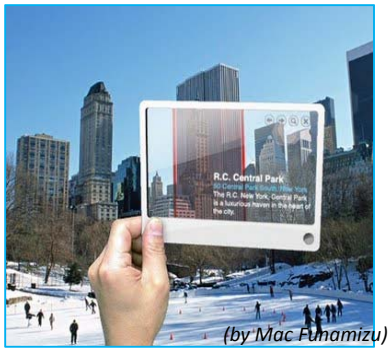
Stemming Variations: image retrieving, images retrieved, images Retrieval, image retrievals, image retrieved





Ricoh, HotPaper





(by Mac Funamizu)

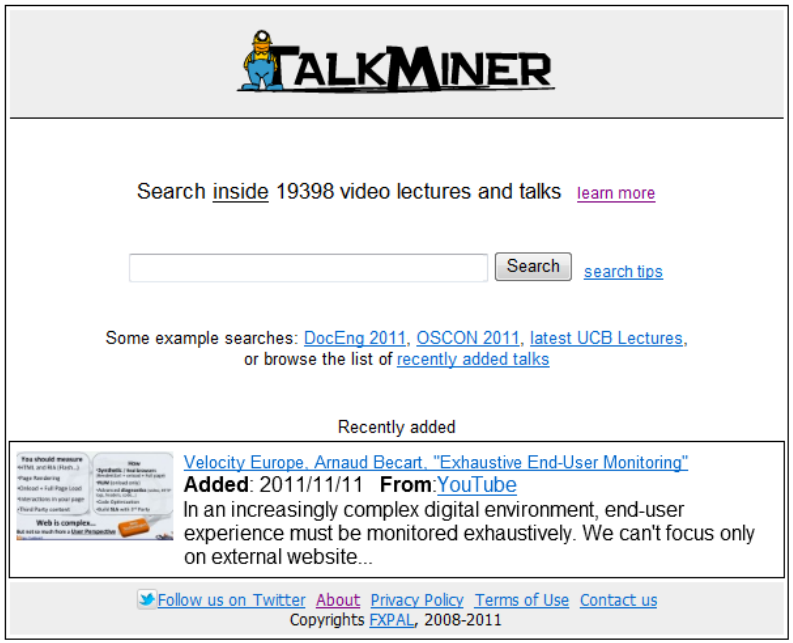
Tineye.com

NSF Workshop on Visual Information Management Systems (1993)

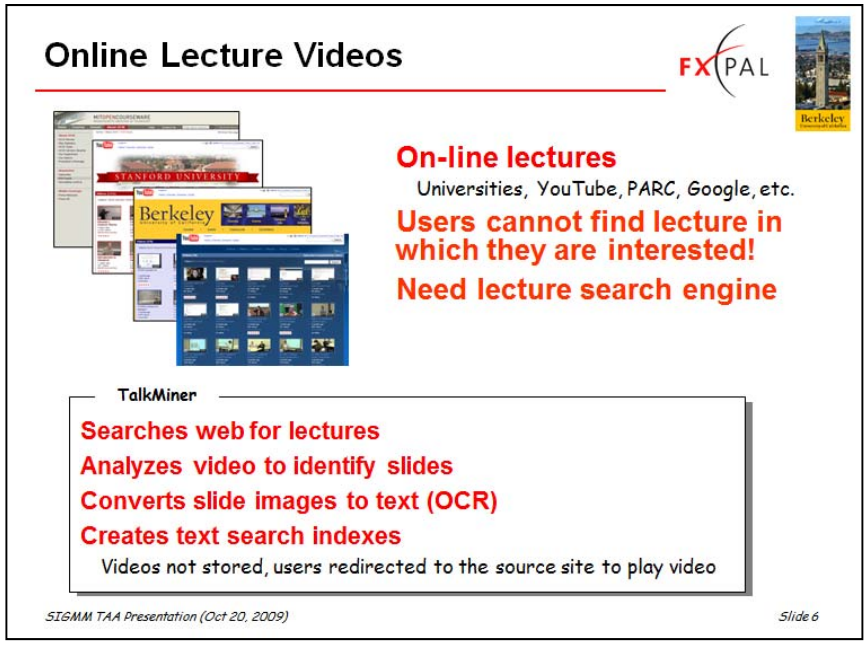
- Require New Technologies in
  - Database
    - data model, indexing, memory management, query processing
  - Computer Vision
    - Interactive image understanding
  - Knowledge Representation and Reasoning

- Four Grand Challenge Applications
  - A Nation-Wide Educational Network
    - provide a visual repository of the best available lectures, videos, interactive classes.
  - Engineering/Scientific Visualization System
    - increase engineering productivity
  - Medical Information System
    - assist diagnosis and treatment
  - Geographic/Environment Information System

Response to First Grand Challenge System



FXPAL TalkMiner Search Engine, ACM MM 10
J. Adcock, M. Cooper, L. Denoue, H. Pirsiavash, L. A. Rowe



Larry's SIGMM TAA talk in 2009

A Few Survey Papers in This Field

(each has been cited more than 1000 times)

- *Image Retrieval: Current Techniques, Promising Directions, and Open Issues*
  [Rui, Huang, and Chang, J. of Vis. Comm. And Image Rep., 1999]

- *Content-Based Image Retrieval at the End of the Early Years*
  [Smeulders, Worring, Santini, Gupta, Jain, T-PAMI, 2000]

- *Image Retrieval: Ideas, Influences, and Trends of the New Age*
  [Datta, Joshi, Li, and Wang, ACM Comp. Survey, 2008]

Key Issues Identified

- *Rui et al 1999*
  - Incorporate Human in the Loop
  - Link Low-Level Features to High-Level Concepts
  - Understand Human Perception of Media Content
  - Support High Dimensional Indexing
  - Provide Web Resources (Taxonomy, Standard)
  - Facilitate Evaluation Testbed
- *Smeulders et al 2000*
  - Address Sensory Gap and Semantic Gap
    - Visual data vs. real-world object vs. human interpretation
  - Use Domain Knowledge to Bridge Gaps
    - Syntactic, perceptual, and topological patterns
  - Consider Different Search Types
    - Aimed, browsing, category search
  - Other issues: User in the Loop, Visualization, Evaluation
- *Datta et al 2008*
  - Discuss Increasingly Diverse Features, Including Regions
  - Identify the Strong Influence of Machine Learning and Statistical Techniques
  - Predict Paradigm Shift to Application-Oriented Domain Specific Work

Example products and systems

informedia
digital video understanding

Autonomy
VIRAGE

MediaMill
Semantic Video Search Engine

Sixth Sense

QBIC™

VideoGoogle

CuZero

bing

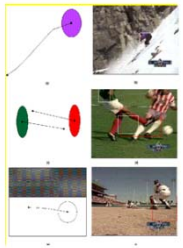
VisualSEEk

IBM
Multimedia Analysis and Retrieval System

snaptell

TALKMINER

VideoQ



blinkx

2008

Google goggles
labs

Google™
Video
BETA

IQ Engines

VisionGo

TinEye

kooaba
IMAGE RECOGNITION

SINGING FISH
the audio/video search engine

like.com
visual search

leafsnap

VIDEOSURF

ALIPR™
Automatic Photo Tagging and Visual Image Search

Stanford Mobile Visual Search

Mealsnap

NOT

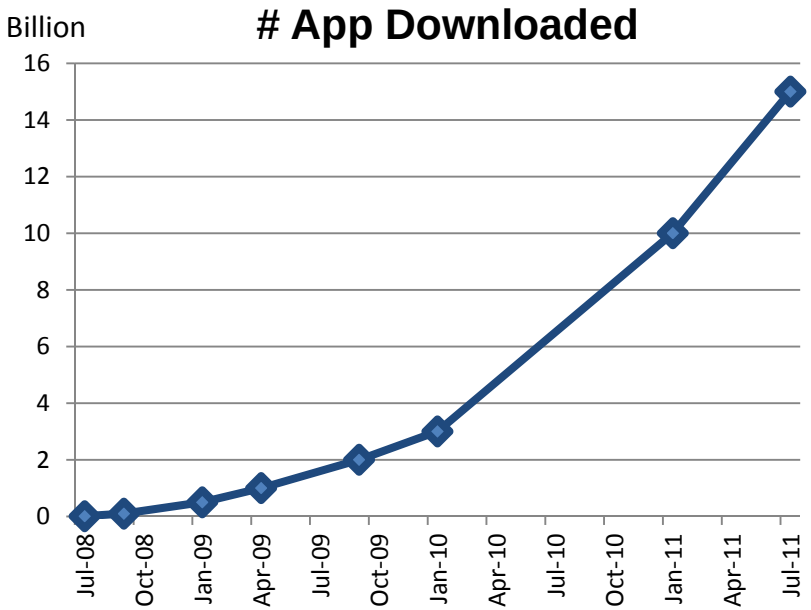
Key Issues Identified

- *Rui et al 1999*
  - Incorporate Human in the Loop
  - Link Low-Level Features to High-Level Concepts
  - Understand Human Perception of Visual Content
  - Support High Dimensional Indexing
  - Provide Web Resources (Taxonomy, Standard)
  - Facilitate Evaluation Testbed
- *Smeulders et al 2000*
  - Address Sensory Gap and Semantic Gap
    - Visual data vs. real-world object vs. human interpretation
  - Use Domain Knowledge to Bridge Gaps
    - Syntactic, perceptual, and topological patterns
  - Consider Different Search Types
    - Aimed, browsing, category search
  - Other issues: User in the Loop, Visualization, Evaluation
- *Datta et al 2008*
  - Discuss Increasingly Diverse Features, Including Regions
  - Identify the Strong Influence of Machine Learning and Statistical Techniques
  - Predict Paradigm Shift to Application-Oriented Domain Specific Work

Explosion of Mobile Apps

Explosion of Mobile Apps

- July 2008 – 10 million apps downloaded in the first weekend
- Jan. 2011 – 10 billion apps downloaded (1000 apps every 3 seconds)
- July 2011 – 15 billion iPhone apps downloaded



Jan. 2009, askiphone.net

The expanded “senses” in the mobile age

- Expanded visual sense

Stanford Mobile Visual Search

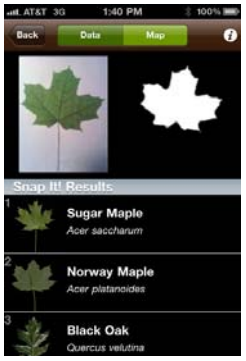
MIT Sixth Sense

kooaba IMAGE RECOGNITION

- Expanded audio sense

shazam

- Expanded sense of food/nature



leafsnap

mealsnap

Augmented Reality

- Create virtual worlds at finger tip and interact
  - Examples: Smart AR from Qualcomm and SONY (2011)
  - Easy creation of 3D virtual space
  - Real-time interaction between characters in physical & virtual worlds

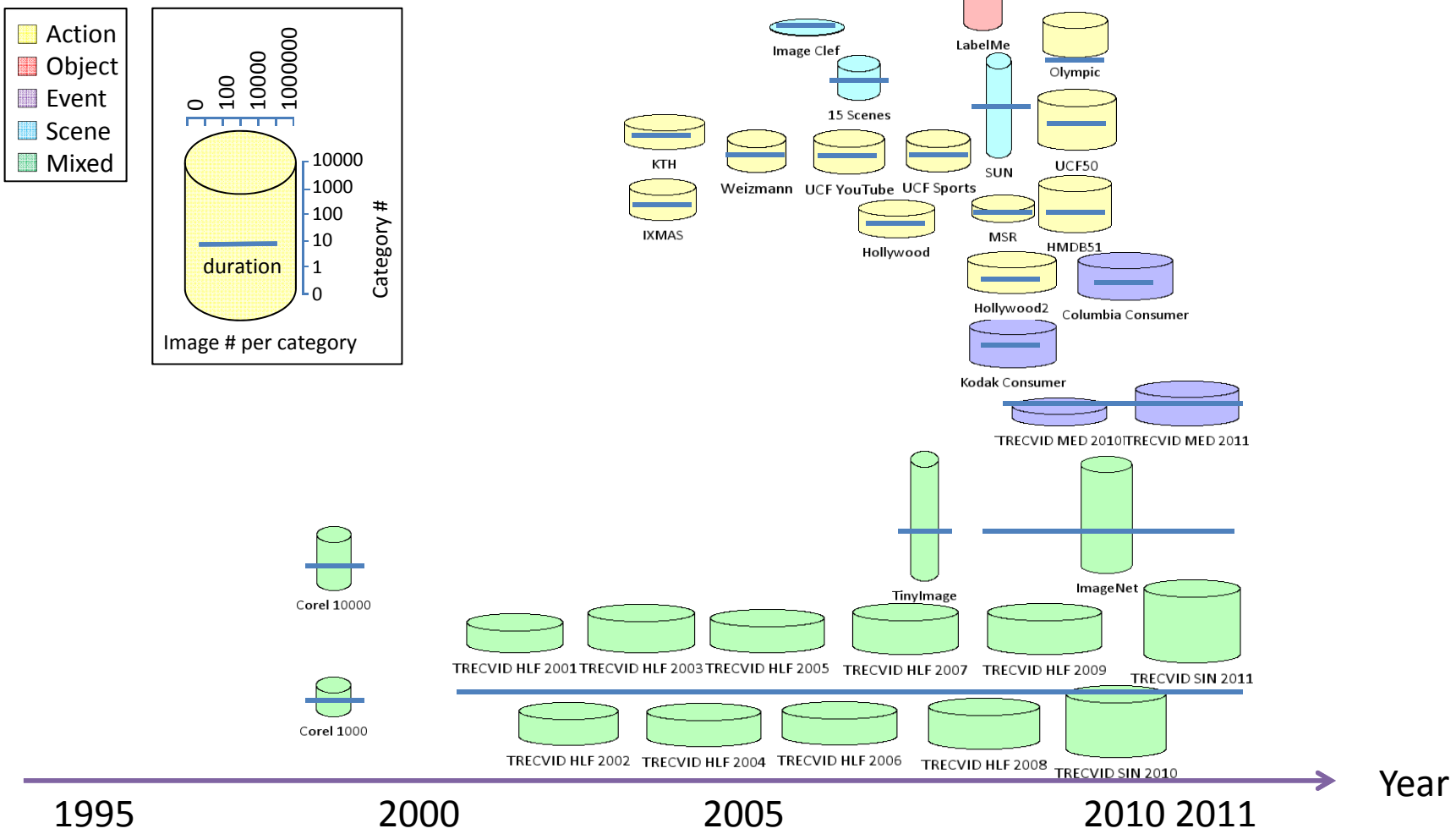
tech.philbuzz.com, 0:40, 1:31

bookmarkblogs.com, 0:41, 1:48, 2:35

Looking Ahead: Challenges & Opportunities

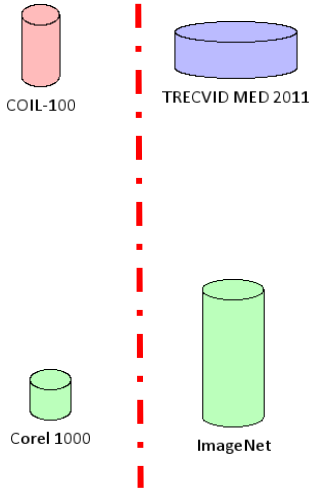
- Data
  - Beyond sample catalogue data
  - Handle real-world gigantic, noisy, complex data
- Content
  - Beyond domain specific solutions
  - Deep multimodal analysis and knowledge representation
  - Return to general large-scale semantic modeling
- User Dimension
  - Beyond human in the loop and relevance feedback
  - Understand user intention and behavior

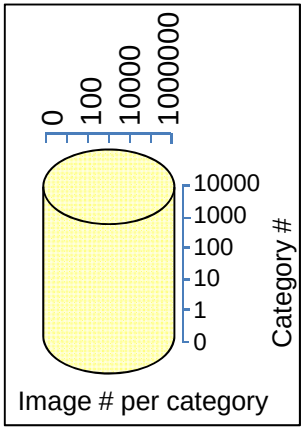
MM Dataset Evolution



A Gigantic Leap in both Data and Semantics

1996 → 2011





- 10-100 primitive categories ~100 images per category (COREL, COIL)

Batting a run in

Making a cake

Assembling a shelter

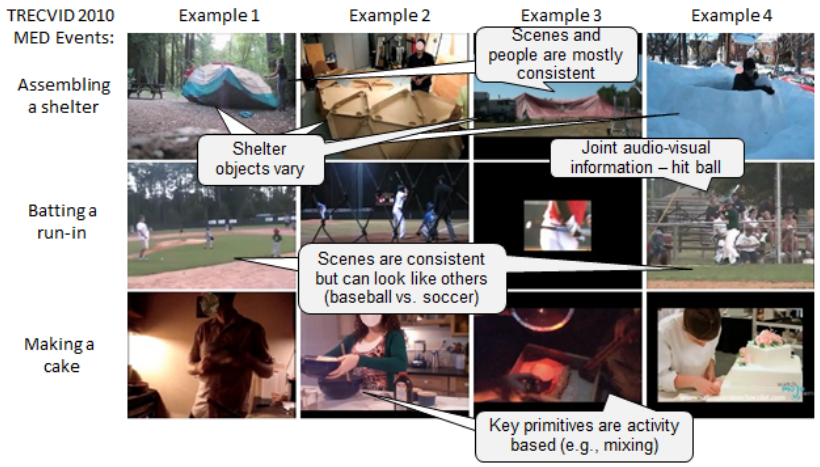
- ~ 1 million video frames per event (IARPA ALADDIN MED)
- 15,000 noun categories (ImageNet)

Looking Ahead: Challenges & Opportunities

- Data
  - Beyond sample catalogue data
  - Handle real-world gigantic, noisy, complex data
- Content
  - Beyond domain specific solutions
  - Deep multimodal analysis and knowledge representation
  - Return to general large-scale semantic modeling
- User Dimension
  - Beyond human in the loop and relevance feedback
  - Understand user intention and behavior

Challenges/Opportunities in ALADDIN MED

Semantic Complexity

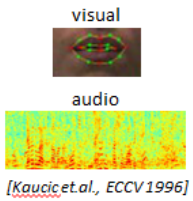


Need discriminative semantic bases for composite event modeling.

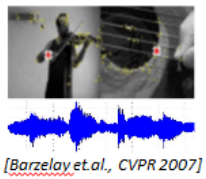
Event Context

Understanding contexts is helpful for event detection.

Deep Multimodal Correlation



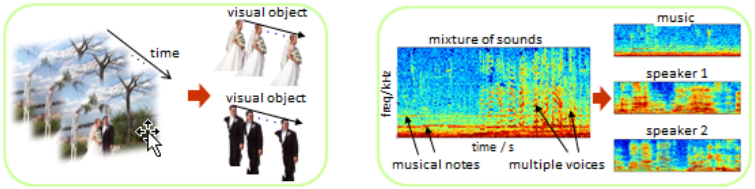
[Kaucic et.al., ECCV 1996]



[Barzelay et.al., CPVR 2007]

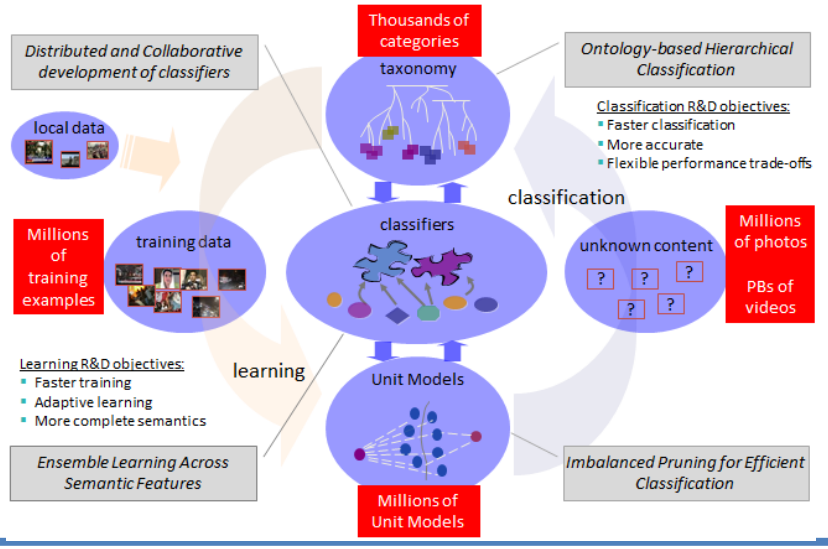
(Cross-media synchrony)

(Causal dynamics across media: human motion -> horse footstep)



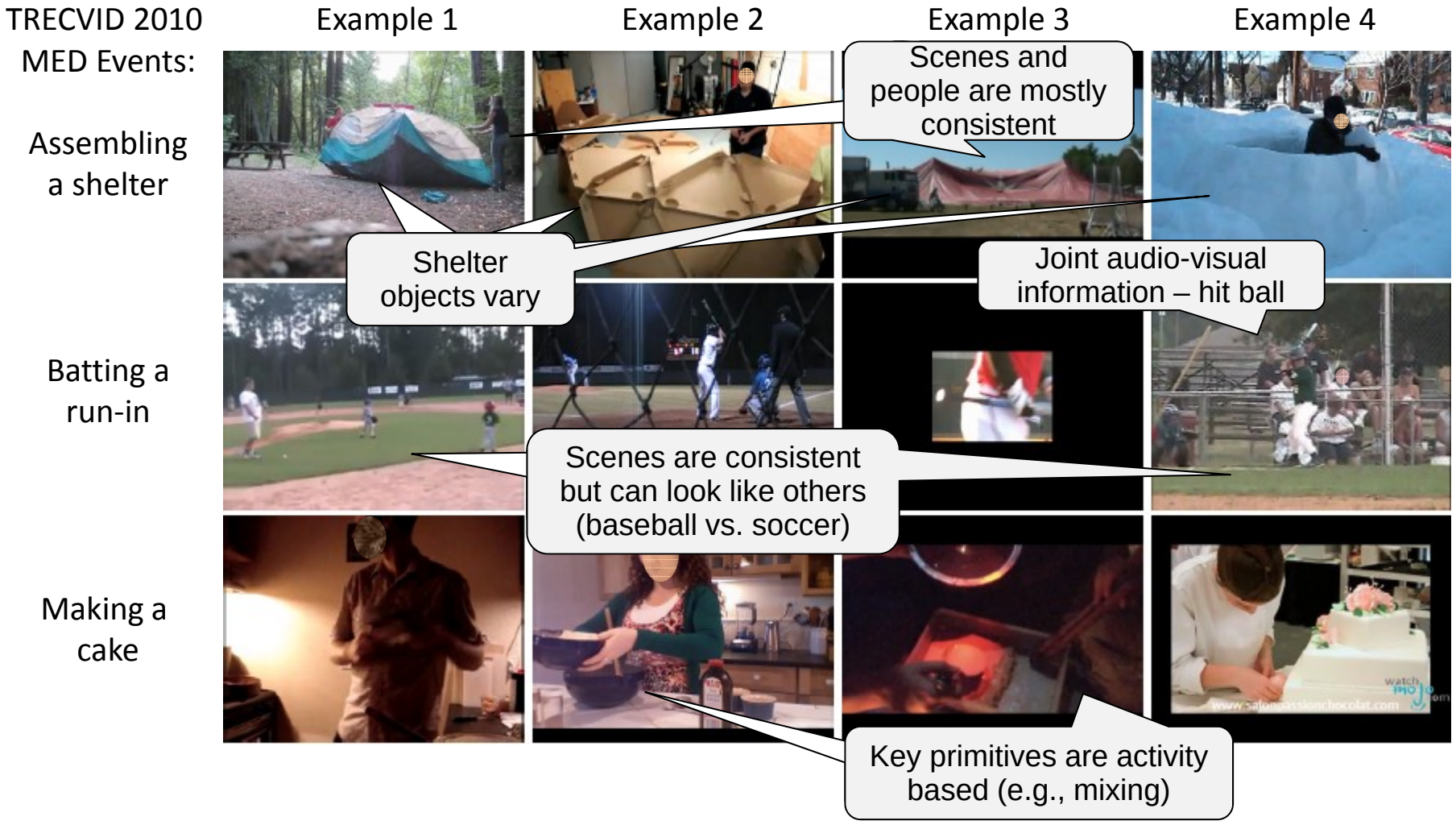
Audio Visual Atoms : multimodal correlative codewords [Jiang, Chang, Ellis, et al ACMMM 09]

Large-Scale Semantic Modeling (IBM IMARS)



Research supported by the IARPA ALADDIN program

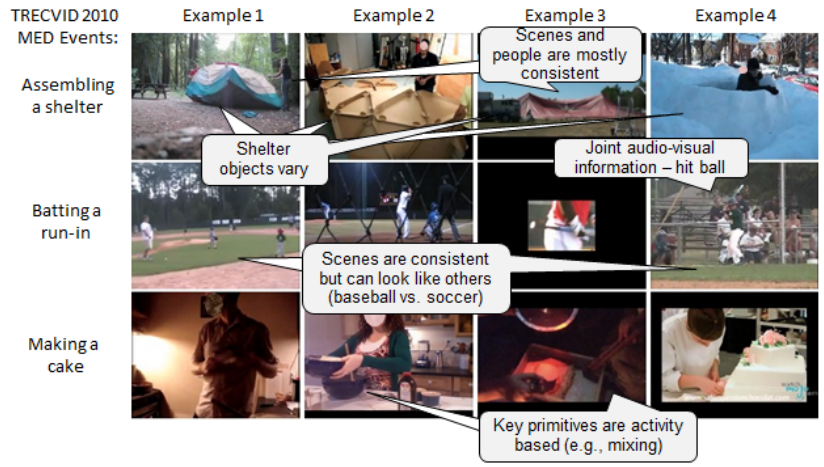
Open-Source Semantic Complexity



Need discriminative semantic bases for composite event modeling.

Challenges/Opportunities in ALADDIN MED

Semantic Complexity

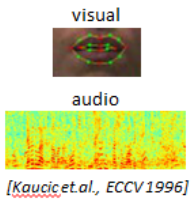


Need discriminative semantic bases for composite event modeling.

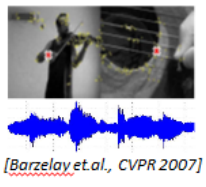
Event Context

Understanding contexts is helpful for event detection.

Deep Multimodal Correlation



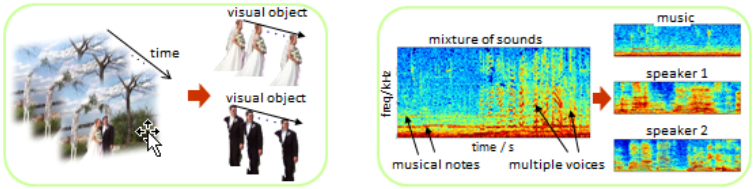
[Kaucic et.al., ECCV 1996]



[Barzelay et.al., CPVR 2007]

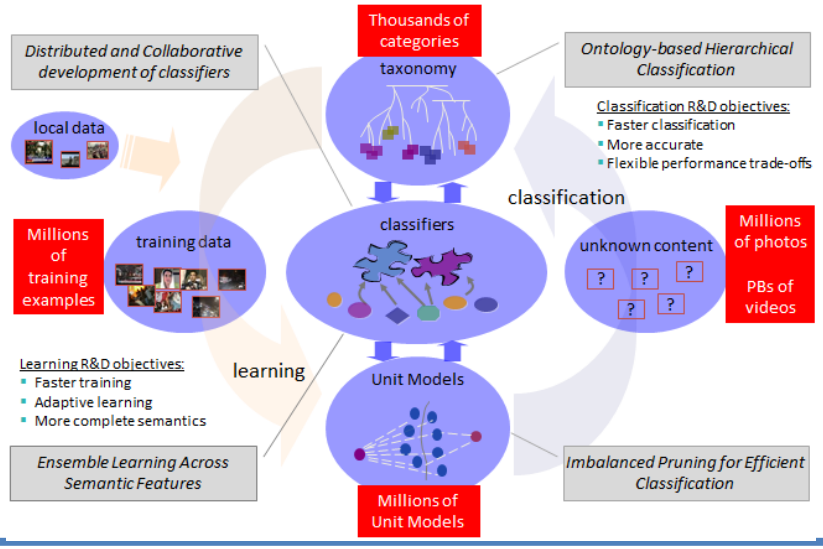
(Cross-media synchrony)

(Causal dynamics across media: human motion -> horse footstep)



Audio Visual Atoms : multimodal correlative codewords [Jiang, Chang, Ellis, et al ACMMM 09]

Large-Scale Semantic Modeling (IBM IMARS)



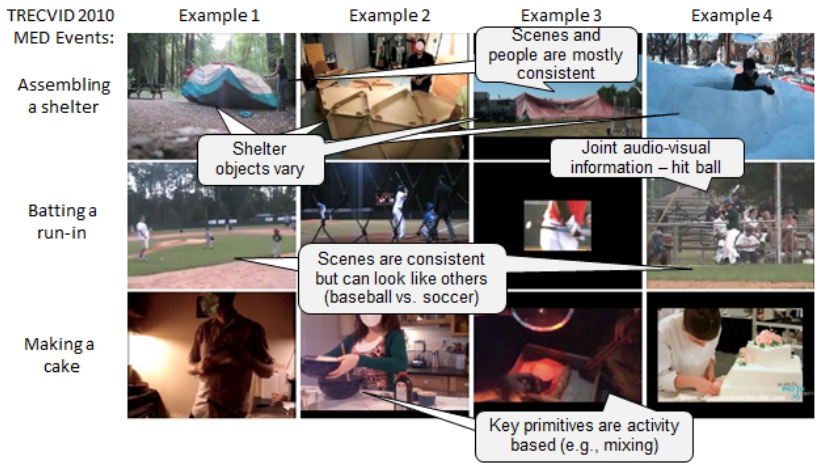
Research supported by the IARPA ALADDIN program

Event Context

Understanding contexts is critical for event modeling.

Challenges/Opportunities in ALADDIN MED

Semantic Complexity

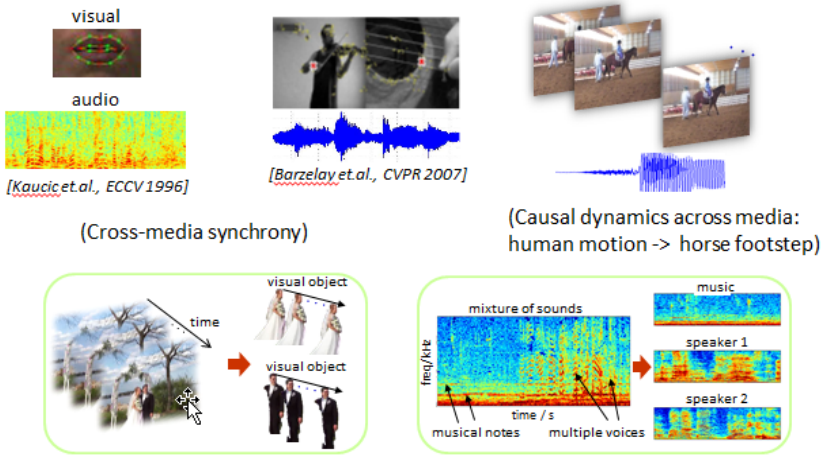


Need discriminative semantic bases for composite event modeling.

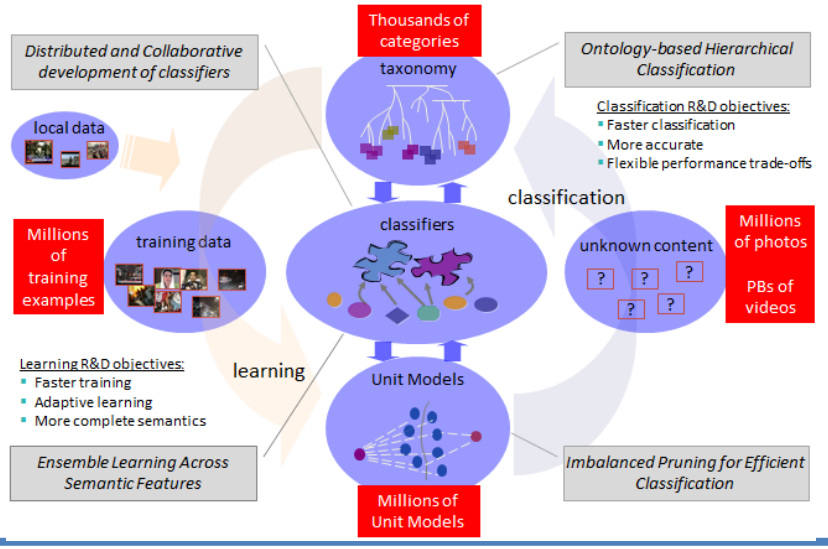
Event Context

Understanding contexts is helpful for event detection.

Deep Multimodal Correlation



Large-Scale Semantic Modeling (IBM IMARS)

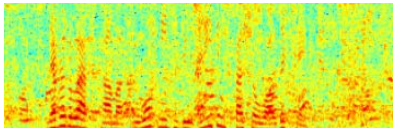


Research supported by the IARPA ALADDIN program

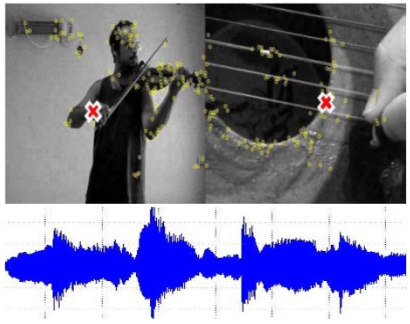
Deep Multimodal Correlation

visual

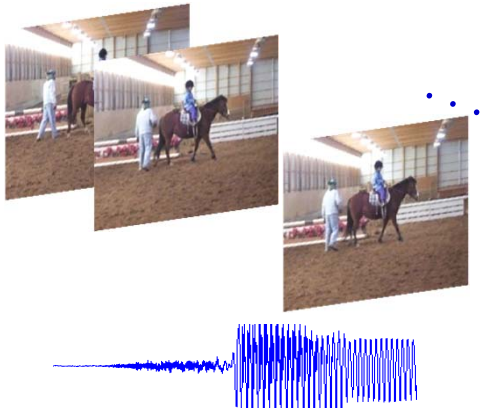
audio



[Kaucic et.al., ECCV 1996]

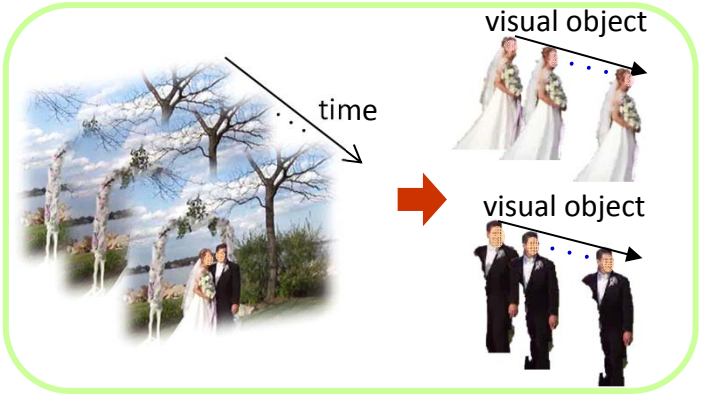


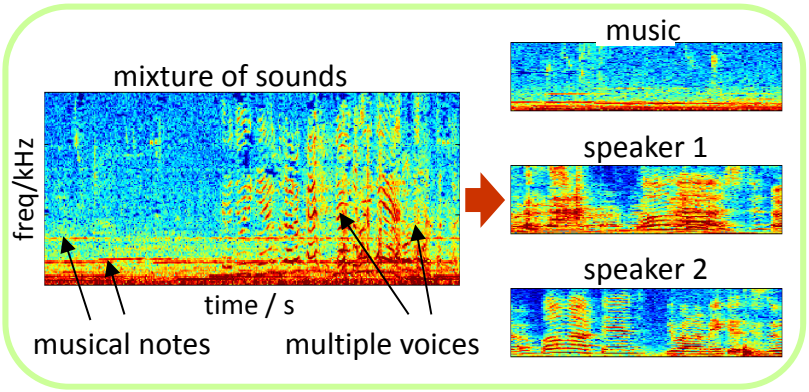
[Barzelay et.al., CVPR 2007]



(Cross-media synchrony)

(Causal dynamics across media:
human motion -> horse footstep)

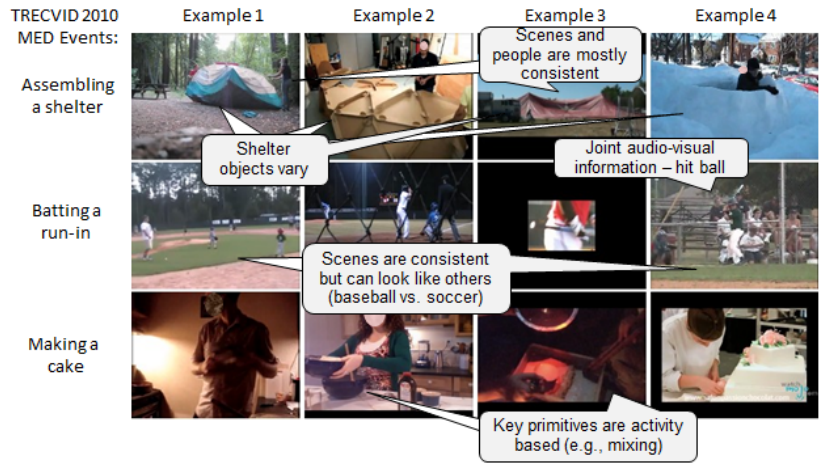




Audio Visual Atoms : Joint multimodal codewords for event detection [Jiang, et al, ACMMM 09]

Challenges/Opportunities in ALADDIN MED

Semantic Complexity

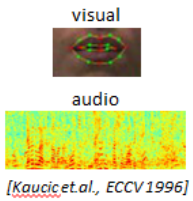


Need discriminative semantic bases for composite event modeling.

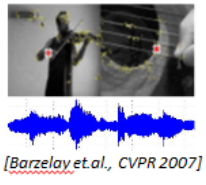
Event Context

Understanding contexts is helpful for event detection.

Deep Multimodal Correlation



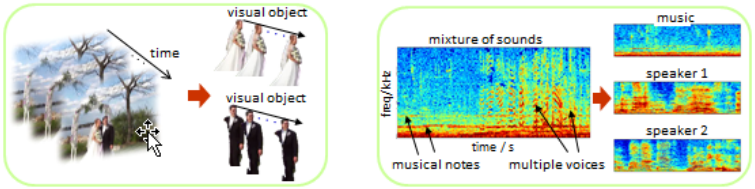
[Kaucic et.al., ECCV 1996]



[Barzelay et.al., CPVR 2007]

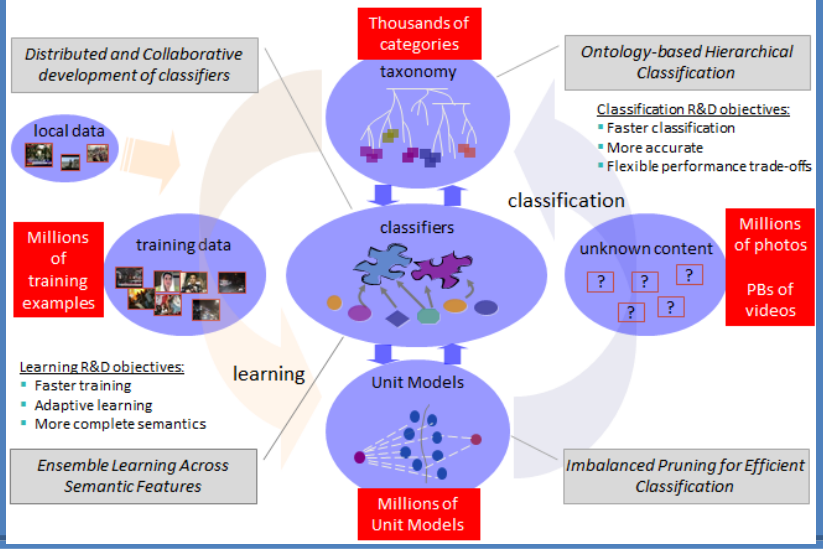
(Cross-media synchrony)

(Causal dynamics across media: human motion -> horse footstep)



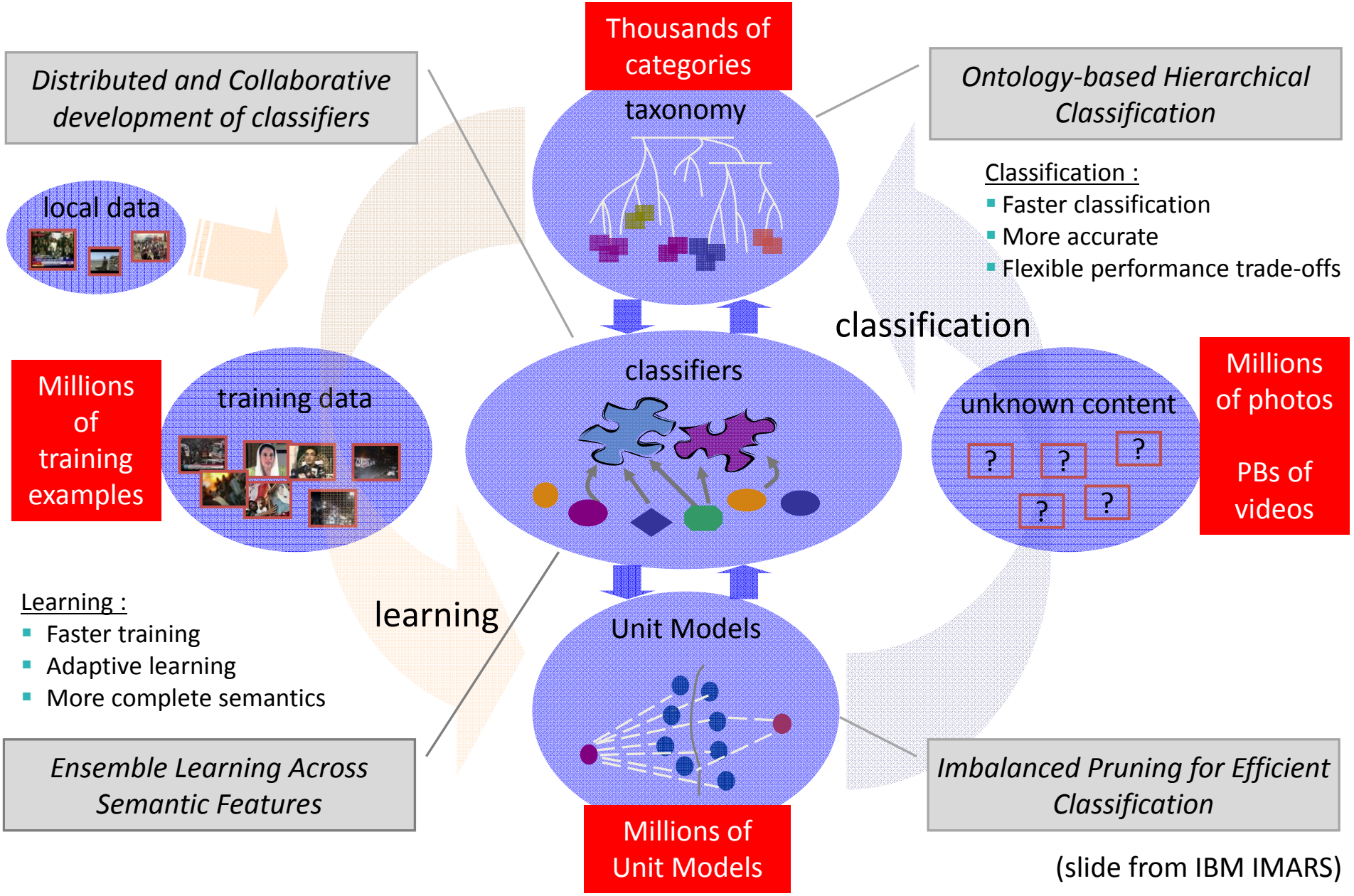
Audio Visual Atoms : multimodal correlative codewords [Jiang, Chang, Ellis, et al ACMMM 09]

Large-Scale Semantic Modeling (IBM IMARS)



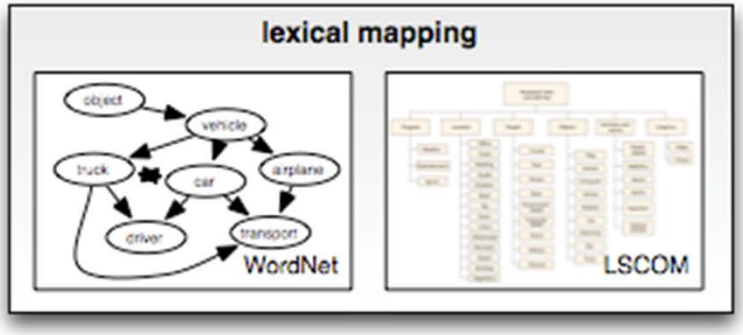
Research supported by the IARPA ALADDIN program

Large-Scale Semantic Modeling



(slide from IBM IMARS)

Facilitate High-Level Multimedia Search

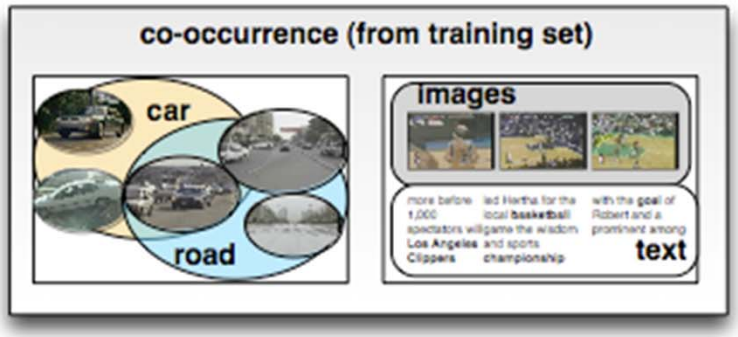


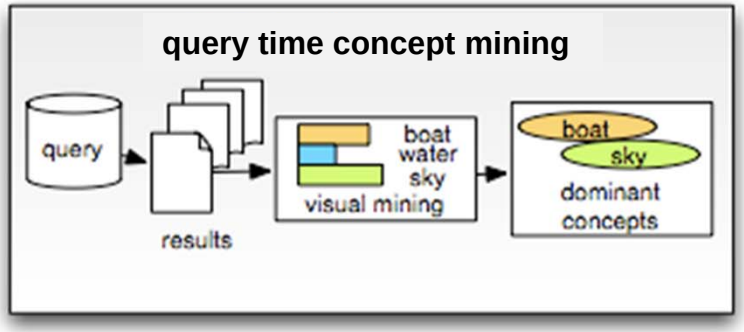
Online Concept Mapping

explosion?

smoke? **vehicle?** **road?**

Query Attributes





P. Natsev, et al, Semantic Concept Based Query Expansion, ACM Multimedia 2007.
W. Hsu, et al, Reranking Methods for Visual Search, Multimedia, 2007.

Encouraging Progress Made

- Taxonomy:
  - LSCOM (2006), ImageNet (2009-11), TRECVID (2001-11)
- Concept Detectors:
  - Columbia374, IMARS, MediaMill, Informedia, Classemes2600
- Example: 126 filtered attributes from TRECVID 2011

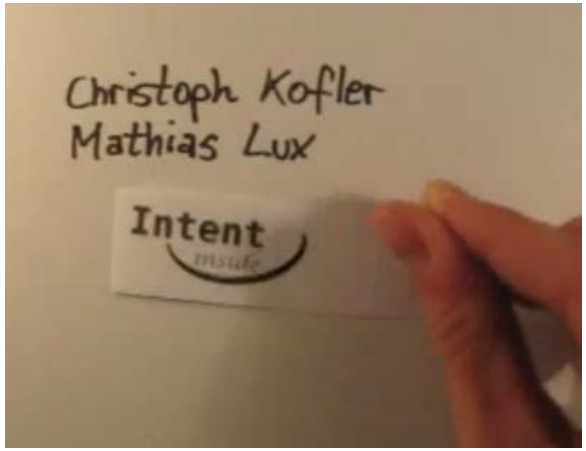
Boy	Airplane	Airplane_Flying	Trees	Classroom	Government-Leader	Highway	Politicians	Dark-skinned_People
Car	Bicycles	Daytime_Outdoor	Child	Reporters	Animation_Cartoon	Kitchen	Black_Frame	Domesticated_Animal
Gun	Building	Ground_Vehicles	Lakes	Teenagers	Apartment_Complex	Meeting	Blank_Frame	Eukaryotic_Organism
Face	Cheering	People_Marching	Animal	Carnivore	Female_Human_Face	Outdoor	Wild_Animal	Female_News_Subject
Girl	Mountain	Walking_Running	Beards	Herbivore	Head_And_Shoulder	Running	Anchorperson	Construction_Vehicles
Hand	Suburban	Civilian_Person	Driver	Quadruped	Human_Young_Adult	Singing	Asian_People	Instrumental_Musician
Road	Swimming	Female_Reporter	Indoor	Old_People	Male_News_Subject	Streets	Sitting_Down	Waterscape_Waterfront
City	US_Flags	Hispanic_Person	Person	Scene_Text	Military_Aircraft	Walking	Urban_Scenes	Residential_Buildings
News	Clearing	Male_Human_Face	Forest	Body_Parts	Adult_Female_Human	Glasses	Female_Person	Celebrity_Entertainment
Room	Speaking	Office_Building	Hockey	Caucasians	Man_Wearing_A_Suit	Skating	Overlaid_Text	Male-Face-Closeup
Actor	Standing	House_Of_Worship	Mammal	Junk_Frame	Military_Personnel	Talking	Single_Person	Demonstration
Adult	Bicycling	Press_Conference	Athlete	Urban_Park	Religious_Building	Traffic	Amateur_Video	Studio_Anchorperson
Beach	Boat_Ship	Roadway_Junction	Dancing	Vertebrate	Single_Person_Male	Valleys	Male_Reporter	Female-Face-Closeup
Birds	Cityscape	Adult_Male_Human	Flowers	Male_Person	Speaking_To_Camera	Windows	Man_Made	Text_On_Artificial_Bk

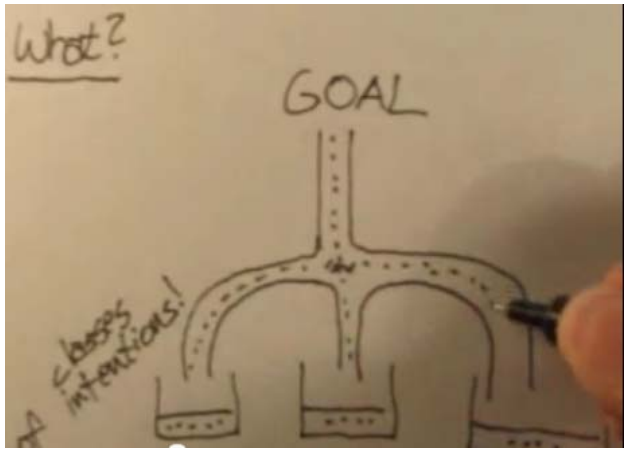
Table 1. 126 query attributes of a-TRECVID, selected from a pool of 346 attributes, by discarding attributes with too few positive images.

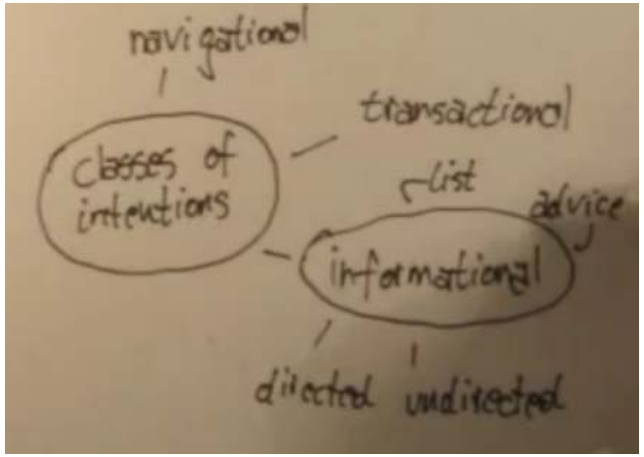
Looking Ahead: Challenges & Opportunities

- Data
  - Beyond sample catalogue data
  - Handle real-world gigantic, noisy, complex data
- Content
  - Beyond domain specific solutions
  - Deep multimodal analysis and knowledge representation
  - Return to general large-scale semantic modeling
- User Dimension
  - Beyond human in the loop and relevance feedback
  - Understand user intention and behavior

User Gap:
User's intention in MM search?







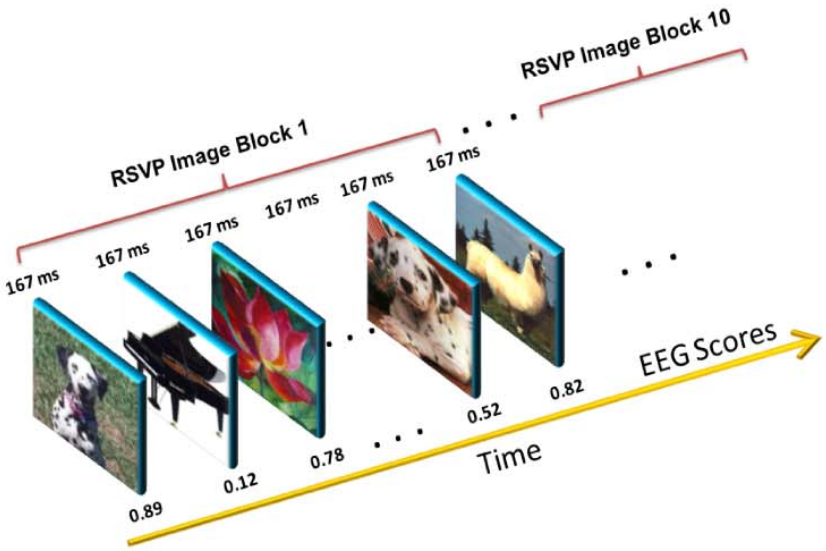
Kofler and Lux, ACM Multimedia 2009, Grand Challenge, Best Presentation

Understand User Intention via Brain State Decoding

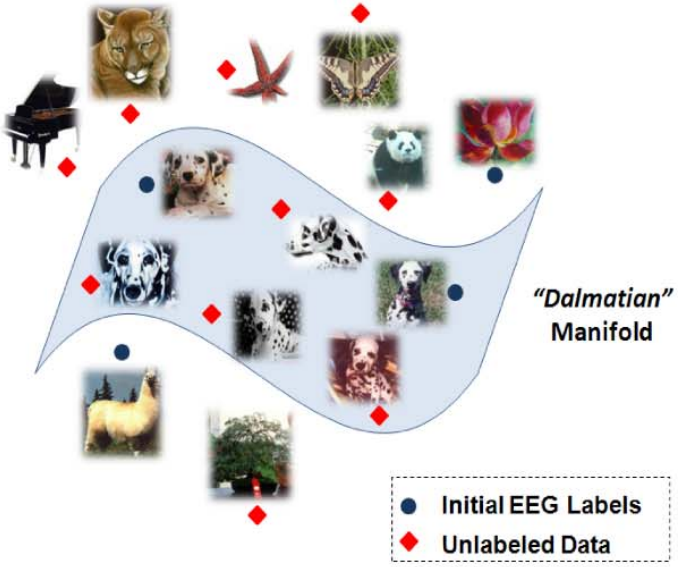
(Wang , Pohlmeyer, Hanna, Jiang, Sajda, Chang, ACM Multimedia 09)

Use EEG brain signals to detect target of interest

Use manifold model to propagate interest labels

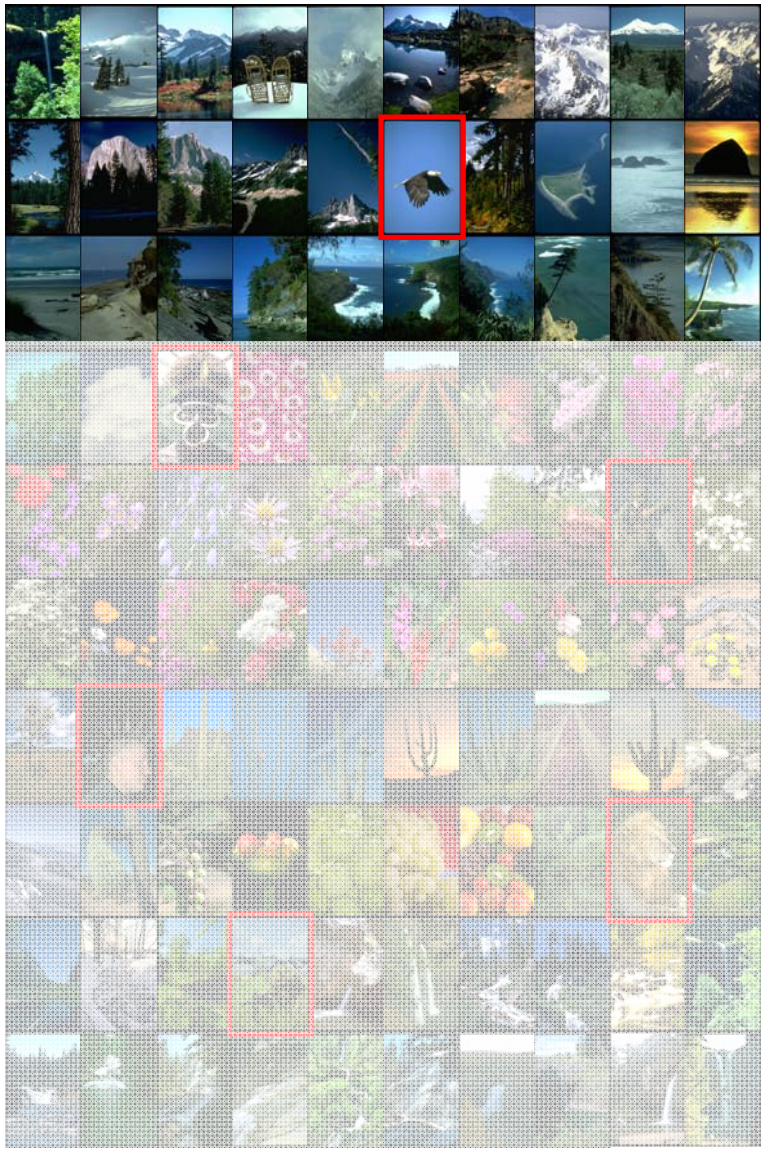


Rapid Serial Presentation of Caltech 101 Images



Graph-Based Visual Pattern Discovery

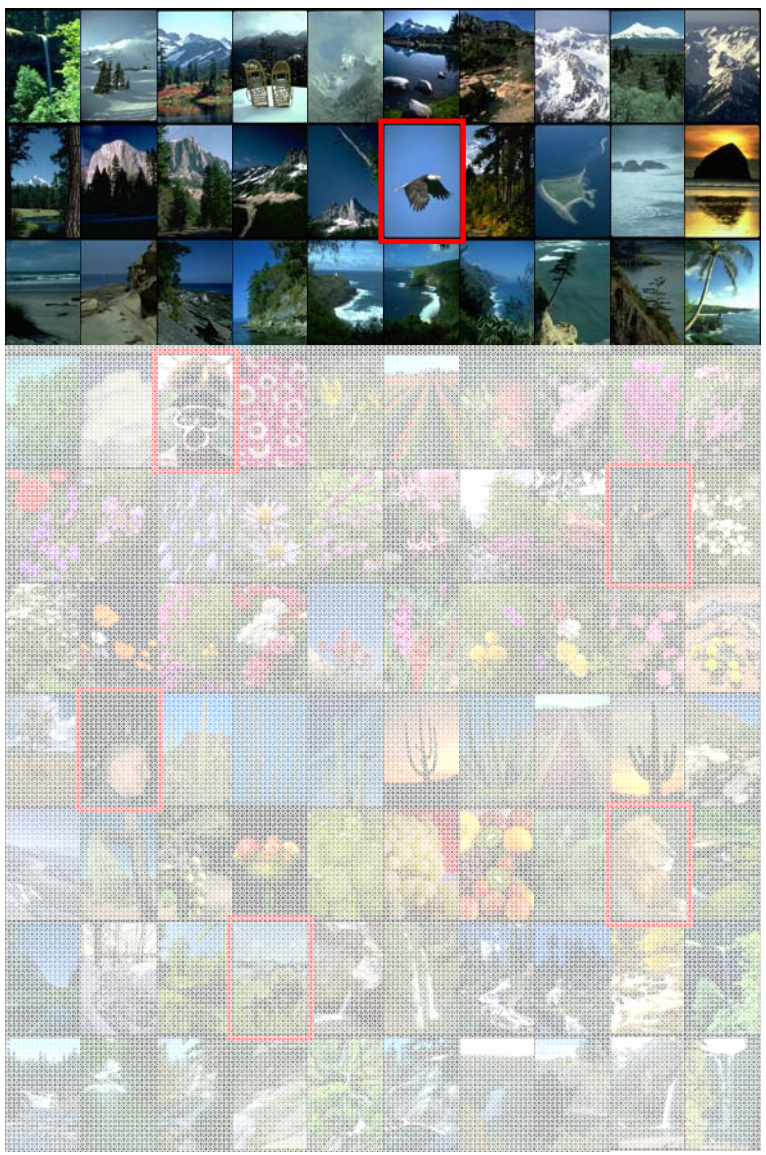
BCI for Reading User Search Intention



Database (any target that may interest users)

User freely thinks about what he/she wants to search

BCI for Reading User Search Intention

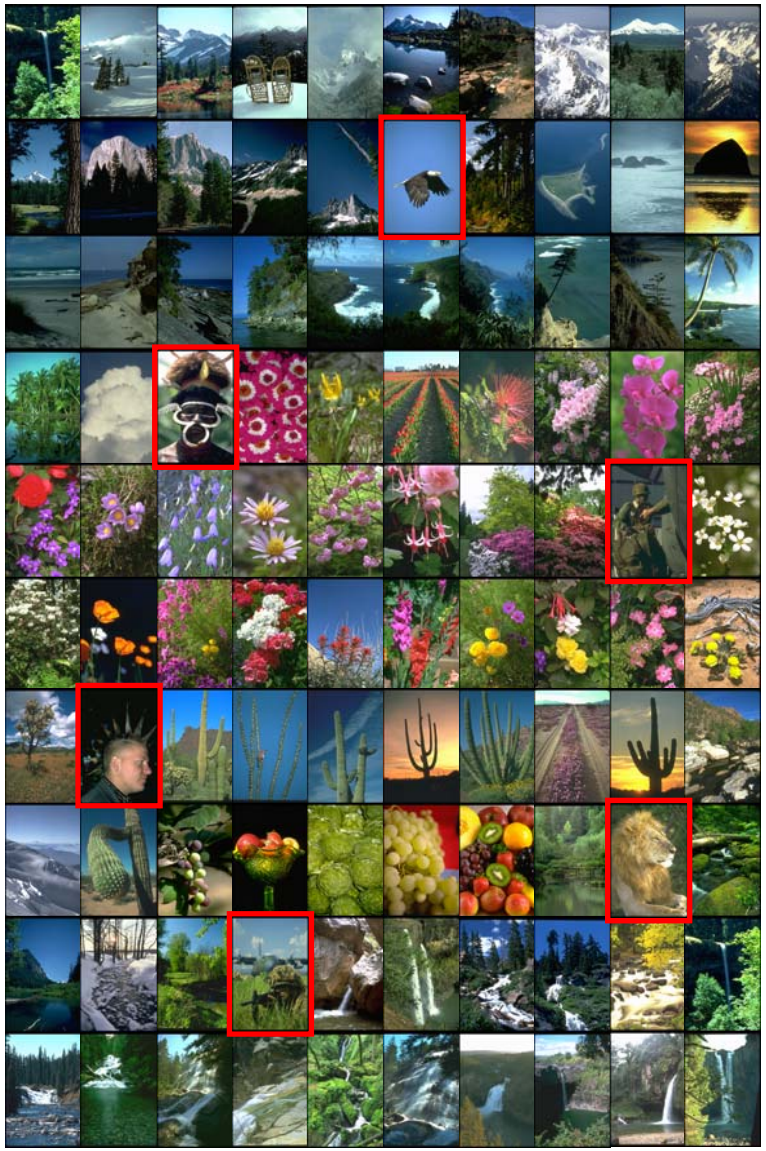


Database

Neural (EEG) decoder

Interest-scores

BCI for Reading User Search Intention



Database

Neural (EEG) decoder

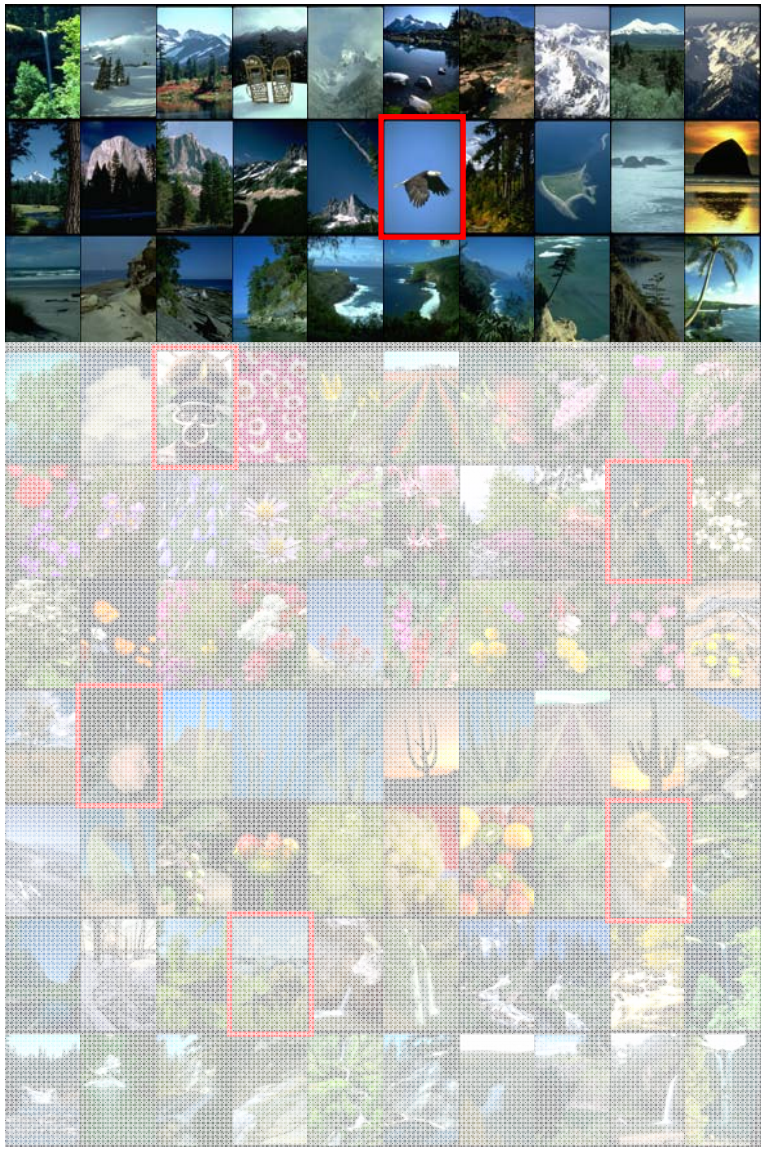
Exemplar labels (noisy)

Semi-supervised
Graph-based propagation

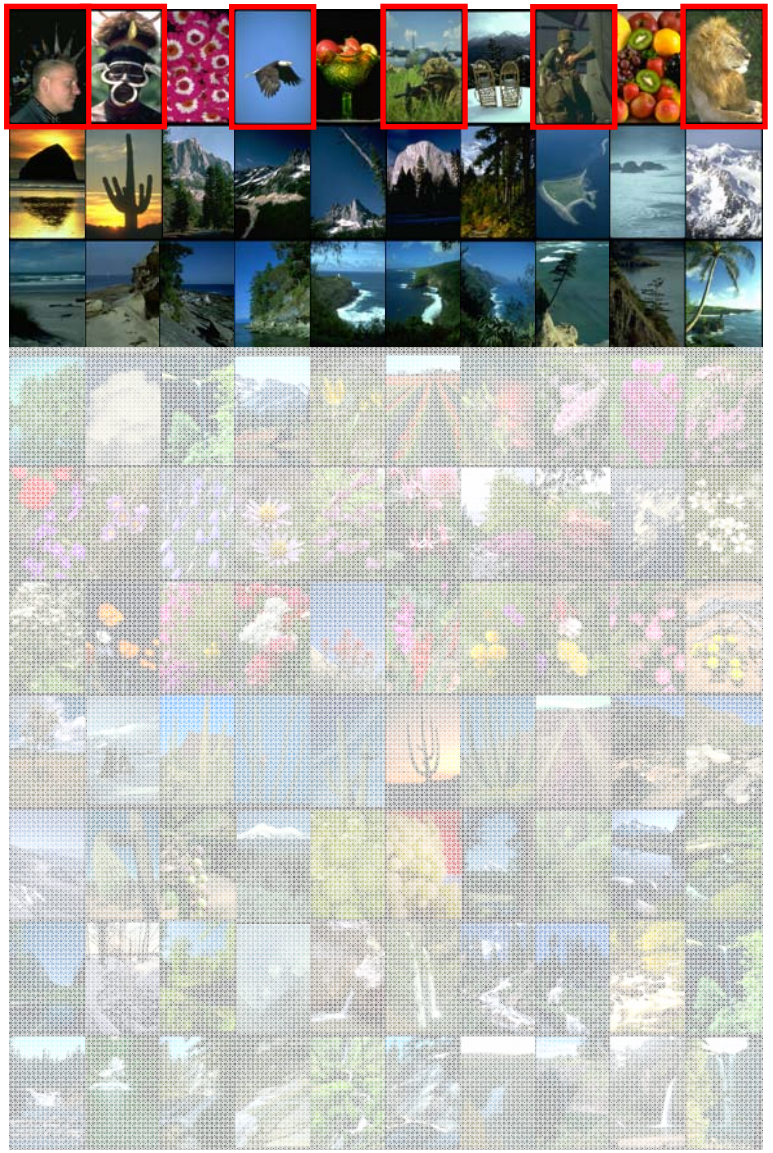
Features from the entire DB

prediction score

BCI for Reading User Search Intention



Pre-triage

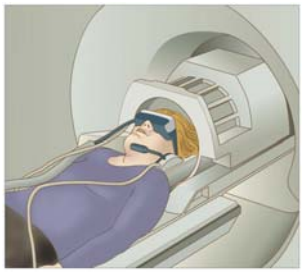


Post-triage

More Blue Skies? Reading Picture in User Mind

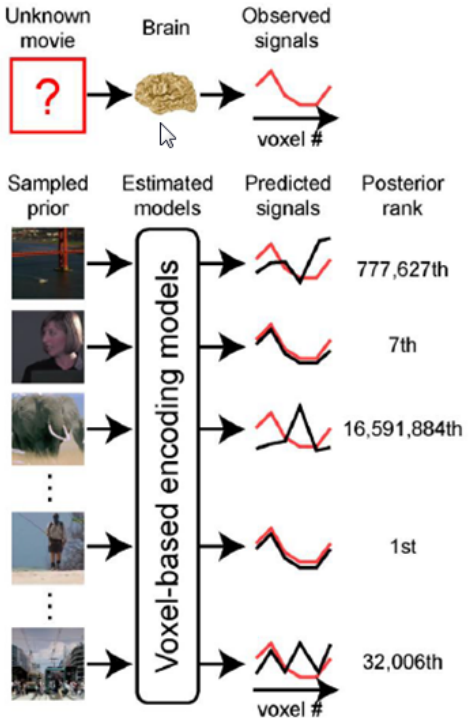
(Nishimoto, Vu, Naslaris, Benjamini, Yu, Gallant, Current Biology, 2011)

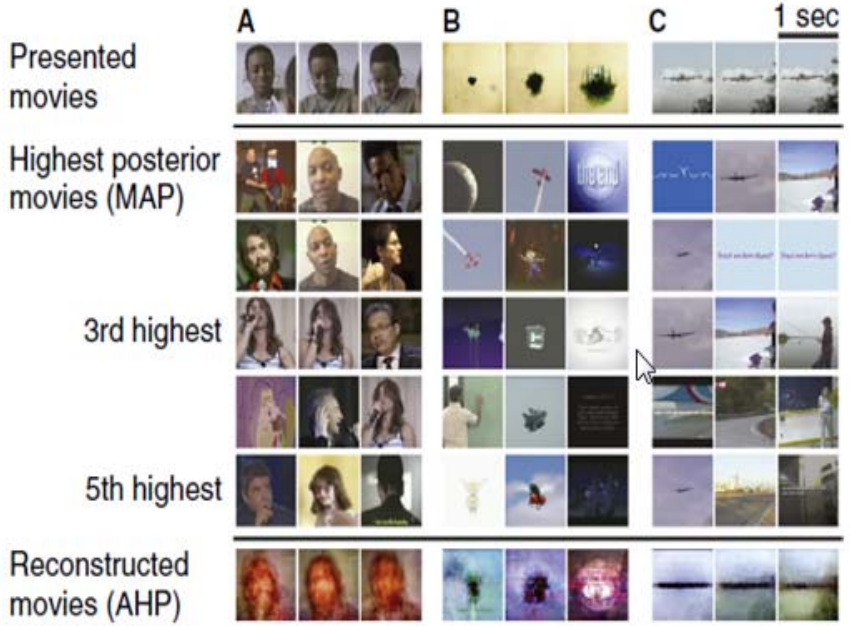
presented movie



reconstructed movie

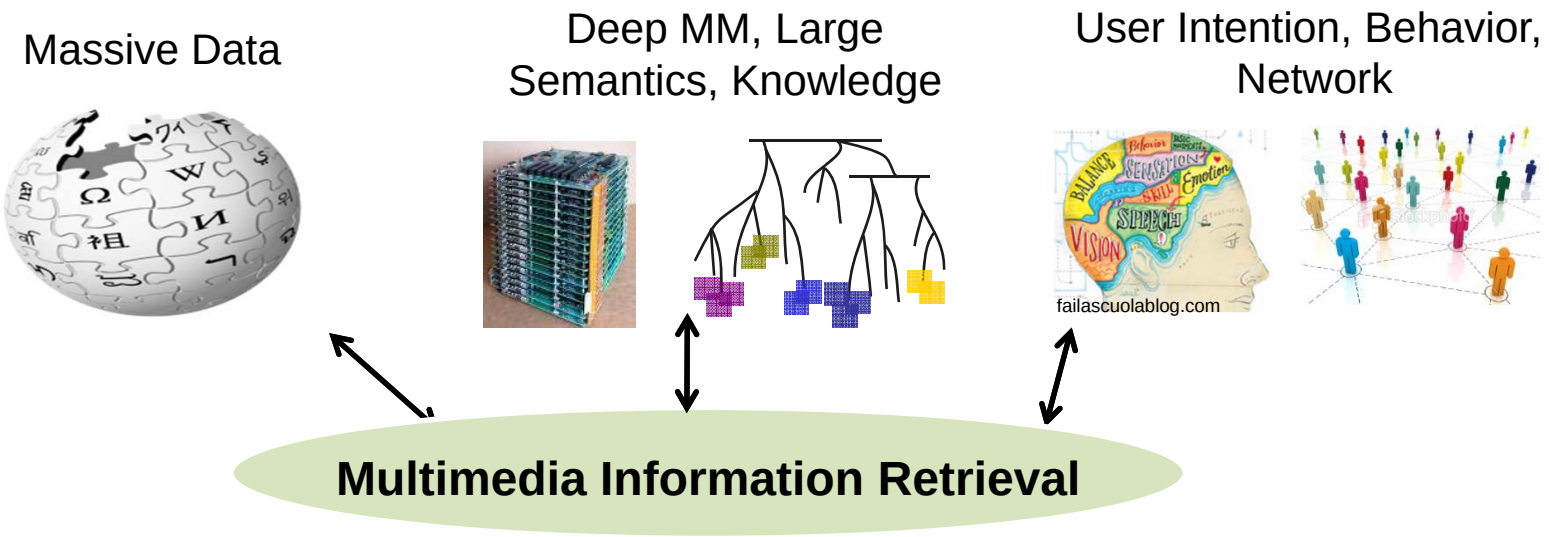
Content-Based Retrieval Framework





An Exciting Time for Multimedia Research

- Future:



- Many exciting research problems
  - New theoretical foundations, tools, and data resources
  - Broad participation and support from government & industry
- Many key contributions made by ACMMM community!