COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

Towards Semantic-Level Visual Search

Prof. Shih-Fu Chang

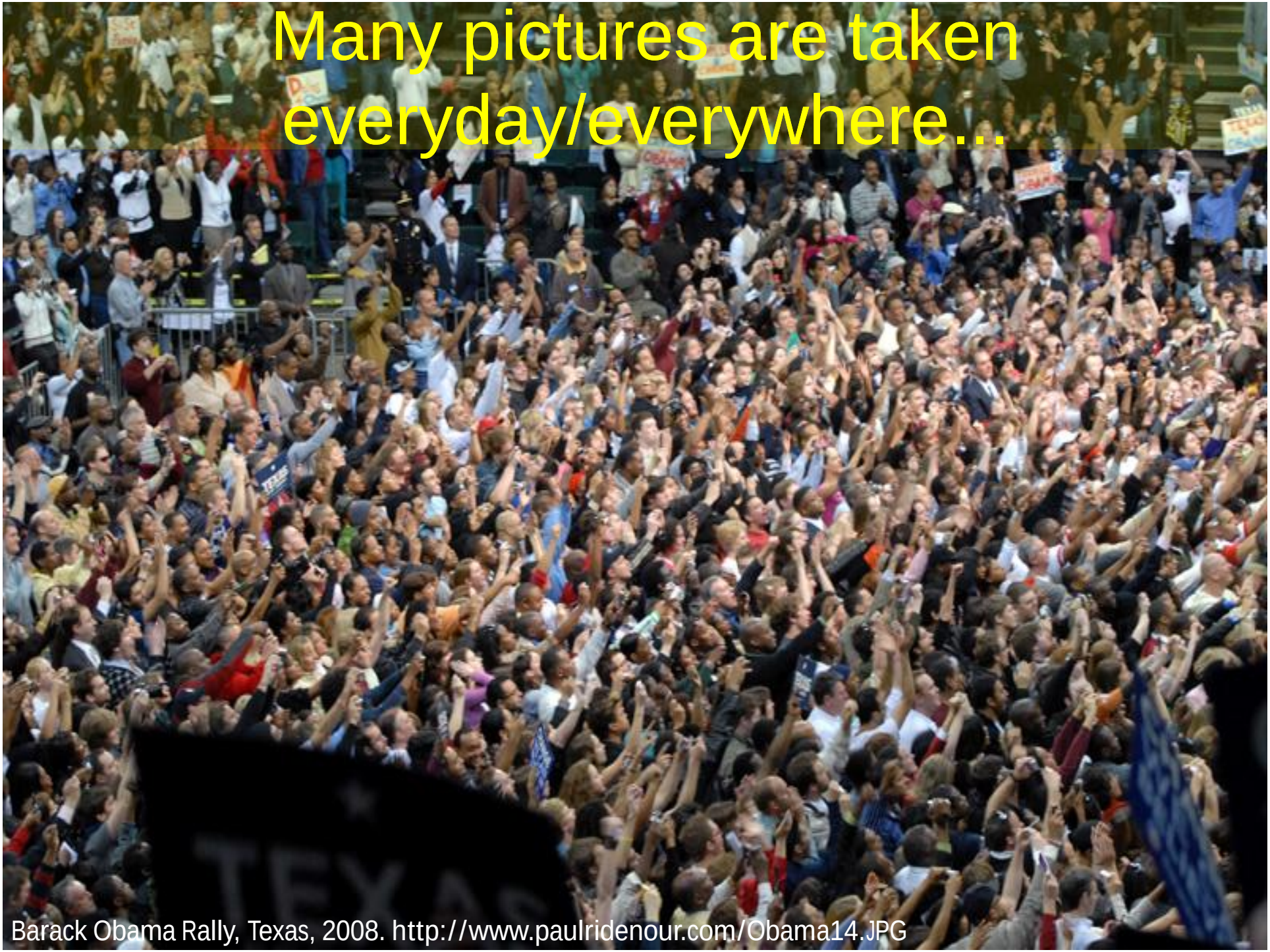
Department of Electrical Engineering

Digital Video and Multimedia Lab

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

Keynote talk at Int. Conf. on Semantic Computing, 2010

Many pictures are taken everyday/everywhere...



Barack Obama Rally, Texas, 2008. http://www.paulridenour.com/Obama14.JPG

Taking pictures everyday/everywhere...

http://cache1.asset-cache.net/xc/200545511-001.jpg?v=1&c=NewsMaker&k=2&d=CED48661B87C5DBFF5EA7A9E4218F9EE1F6F6178A68B340C
http://cache4.asset-cache.net/xc/200483959-001.jpg?v=1&c=NewsMaker&k=2&d=EDF6F2F4F969CEBD9A551141E9DF12C0E47ADEC7FBB32857DF0450484851C07200123AA3B5A18ED0

Taking pictures everyday/everywhere...

miami.eater.com

2.bp.blogspot.com

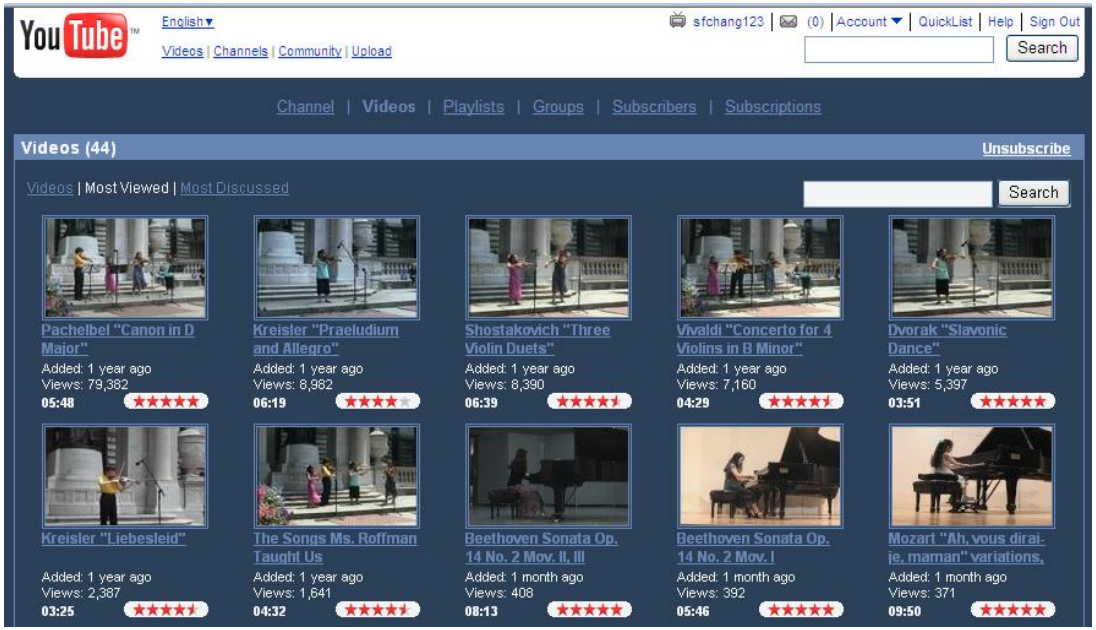
digital-photography-school.com

But we are not good at organizing …

- Most photos and videos remain in shoebox or e-shoebox
- Each uploaded photo has only 0.97 tag on average
- But we love to share them …

Example:
My family's video channel

- Sharing ~ 100 videos on Youtube
- This single video has been viewed >130,000 times
  - More popular than all of my published papers!!!



video

Challenge

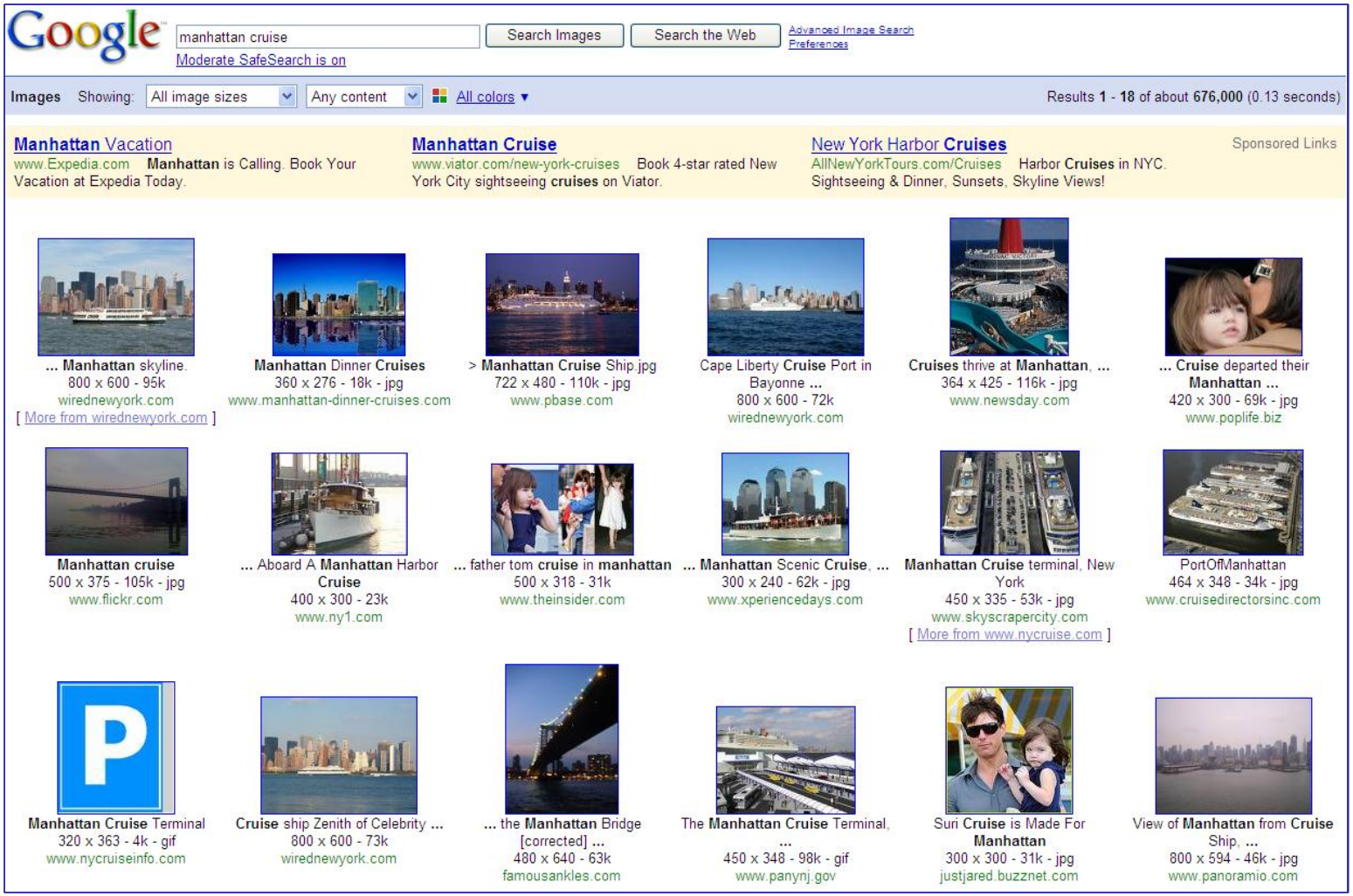
It will be nice to tag …

Performance, classical music, ensemble, concert; Pachelbel Canon in D Major, romantic; young musicians, girls, boys; outdoor, Bryant park in NYC, statue, columns, stage, flowers; Instruments, audio equipments, violins; . . .

- But tagging is boring and hard
- No wonder < 1 tag per uploaded photo

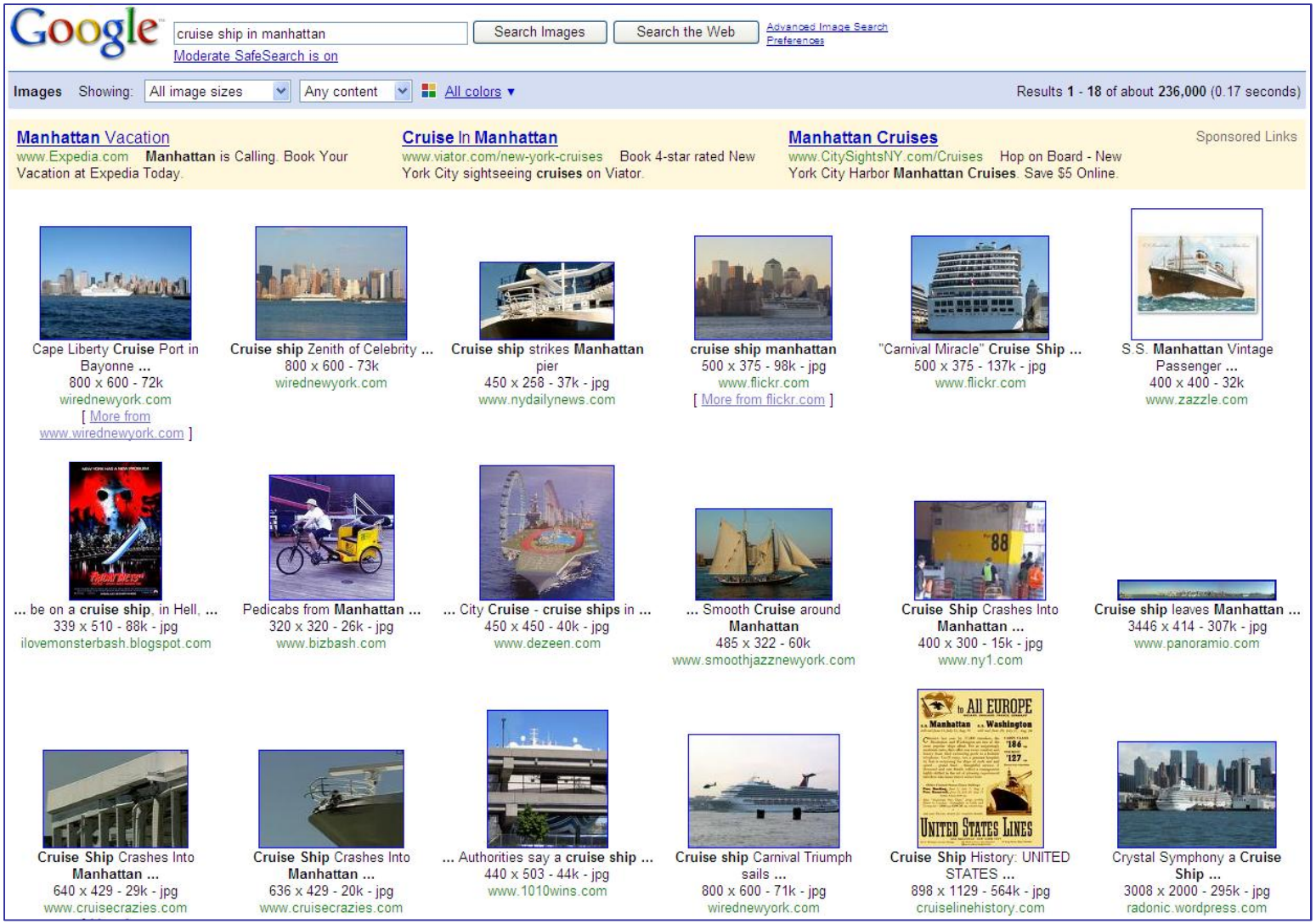
Scarce Tags → Faulty Search Engines

"Manhattan Cruise"



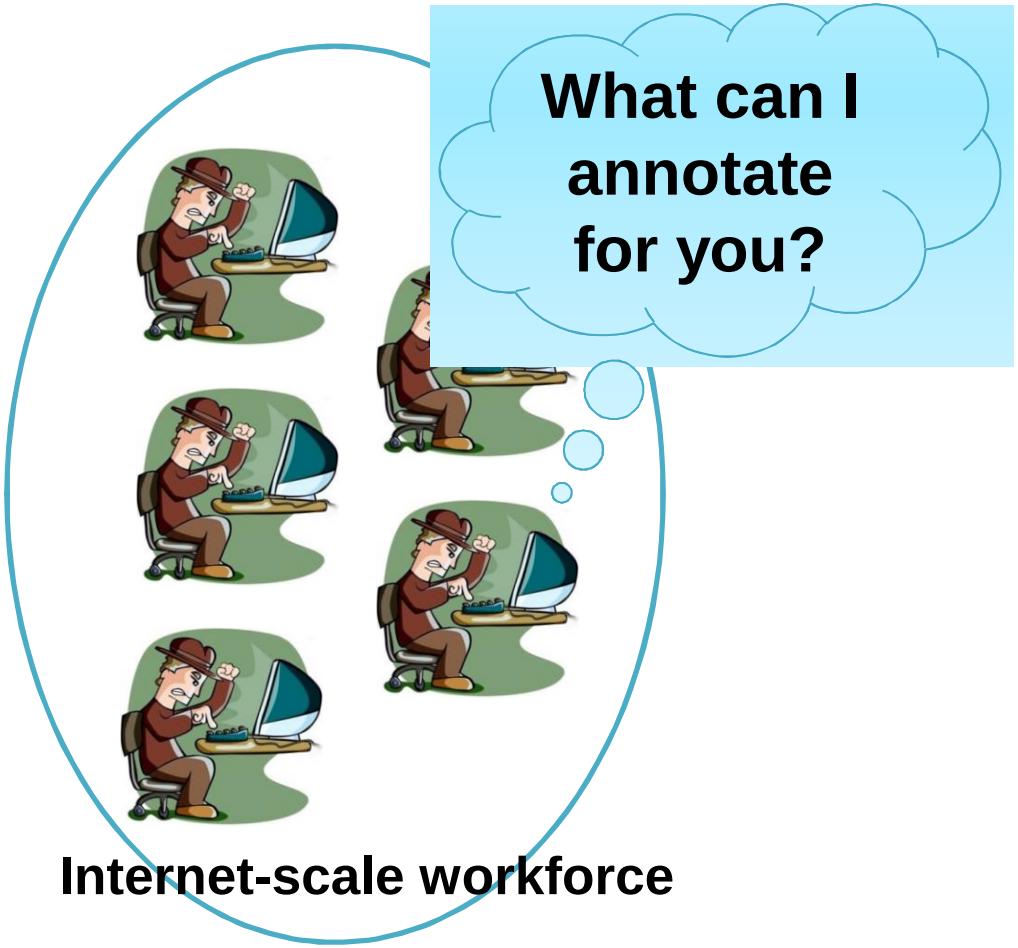
Scarce Tags → Faulty Search Engines

"Cruise ship in Manhattan"



Creative Solutions: Crowd Sourcing

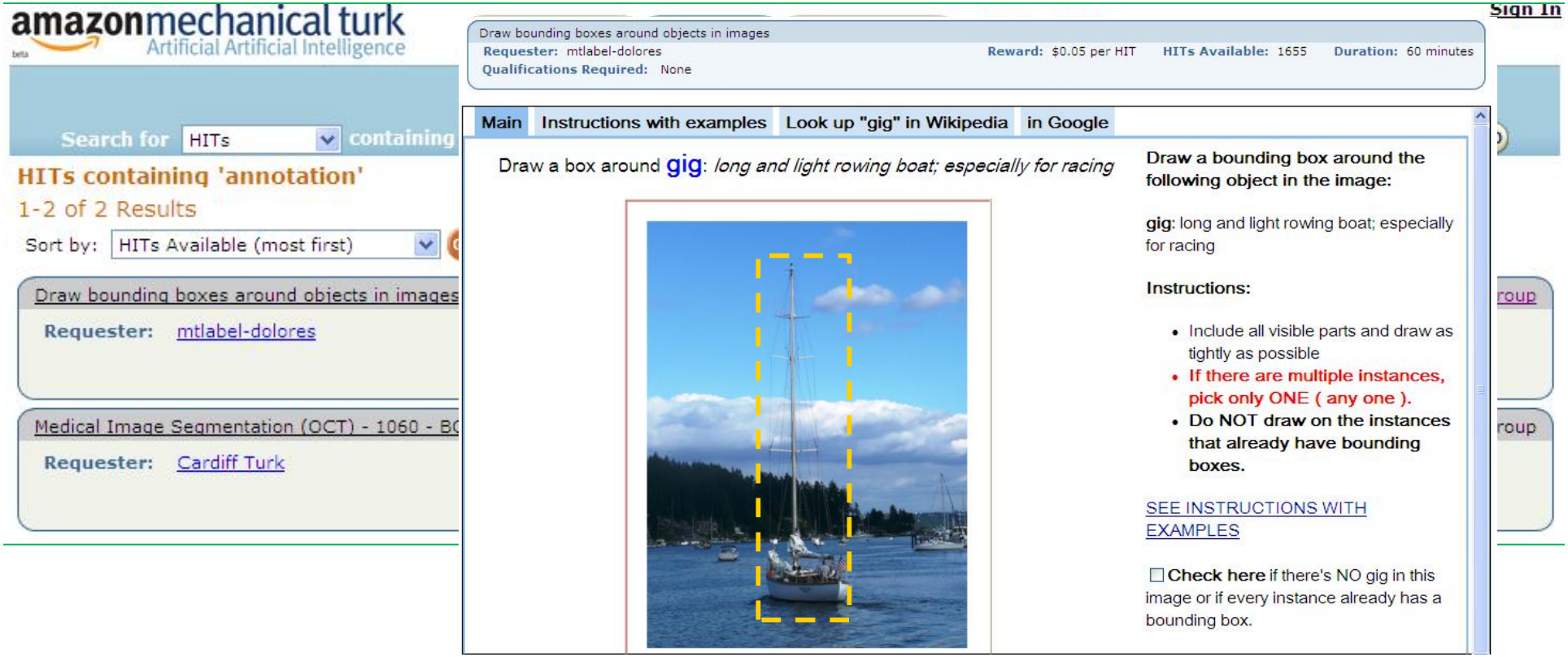
- Amazon Mechanic Turk
- Web Open Market for Human Computing



Internet-scale workforce

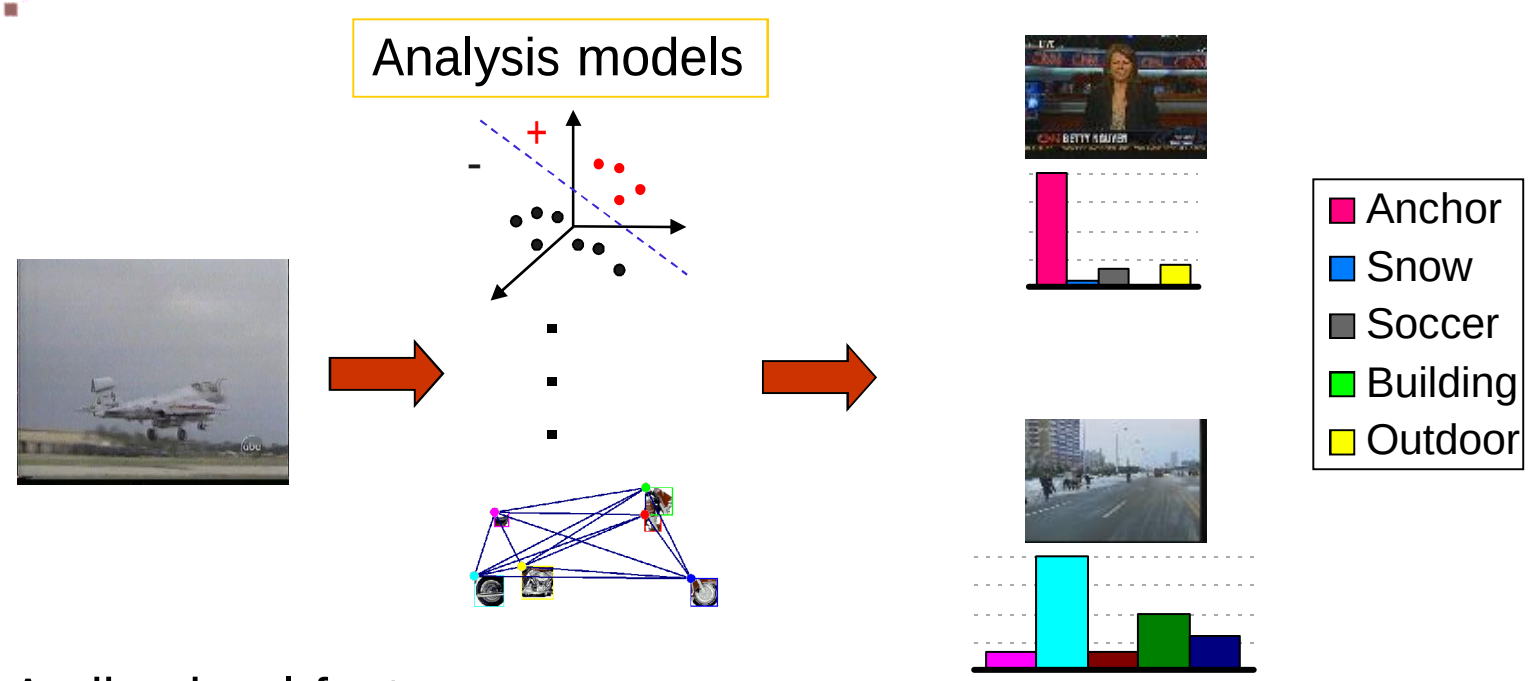
earn $0.03 per image tag

Crowd sourcing for image annotation



Such task is nicely called Human Intelligence Task (HIT)!

An ideal alternative: Automatic Visual Annotation



- Audio-visual features
- Geo, time, camera metadata
- User context

- Rich semantic labels

Hot topic ... community fast growing!

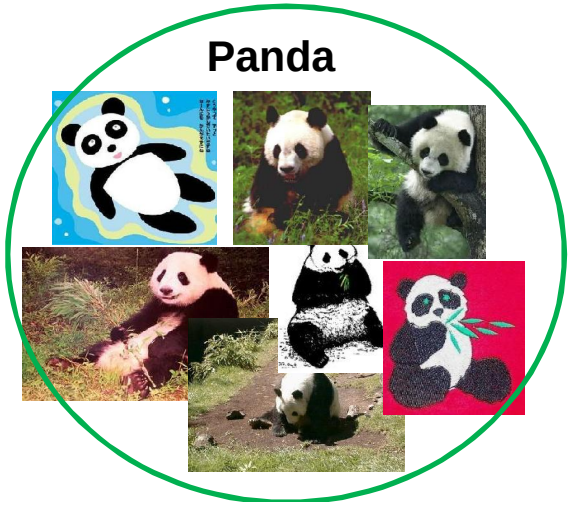
(as of Nov. 2009)

	Data domain	amount	types	Lexicon size
TRECVID	Broadcast news, documentary, Flickr, Youtube...	400 hours Sound & Vision 170 hours Television News 100 hours BBC rushes (130,000+ subshots)	video shots, keyframes	10 (2004, 2005) 39 (2006) 39 (2007) 20 (2008)
LSCOM	Broadcast news video	170 hours Television News 61901 subshots	video subshots	1000+ concepts
CalTech256	Internet Images	30,607 images	images	256 classes
PASCAL	Internet Images	9,963 images 24,640 annotated objects	images, objects	20 classes
Tiny Image	Internet Images	80,000,000 tiny images (32x32)	images	75,378 WordNet nouns
LabelMe	Internet and user uploaded images and videos	30,369 images from 183 folders	images, keyframes	111,490 object labels
ImageNet	Internet images	9,386,073 images	images	14,847 WordNet synsets
Lotus Hill Dataset	Internet Images	500,000+ images and keyframes	images, keyframes	280 object classes

Image/Video Classification

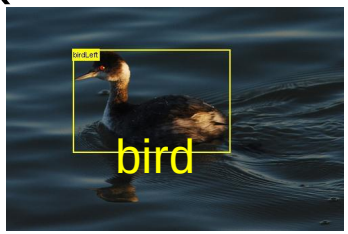
TRECVID

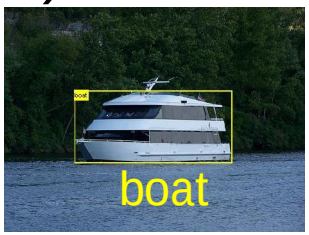
CalTech 101

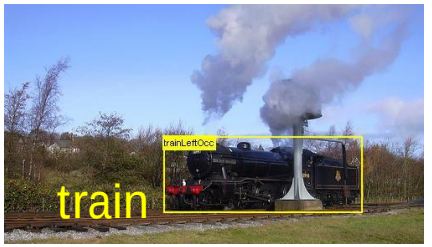


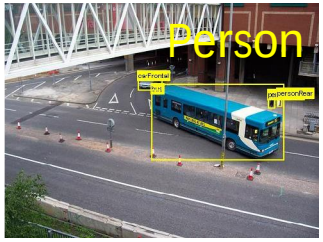
Object Localization

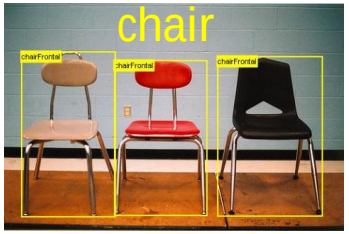
(PASCAL VOC)

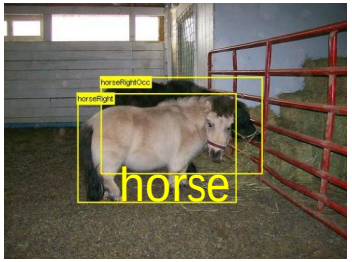


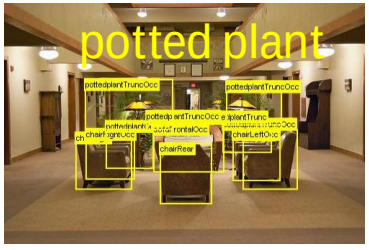


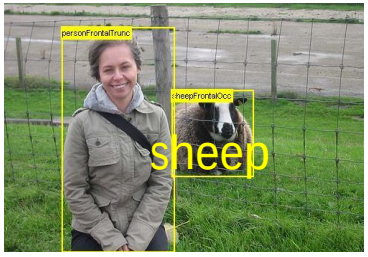




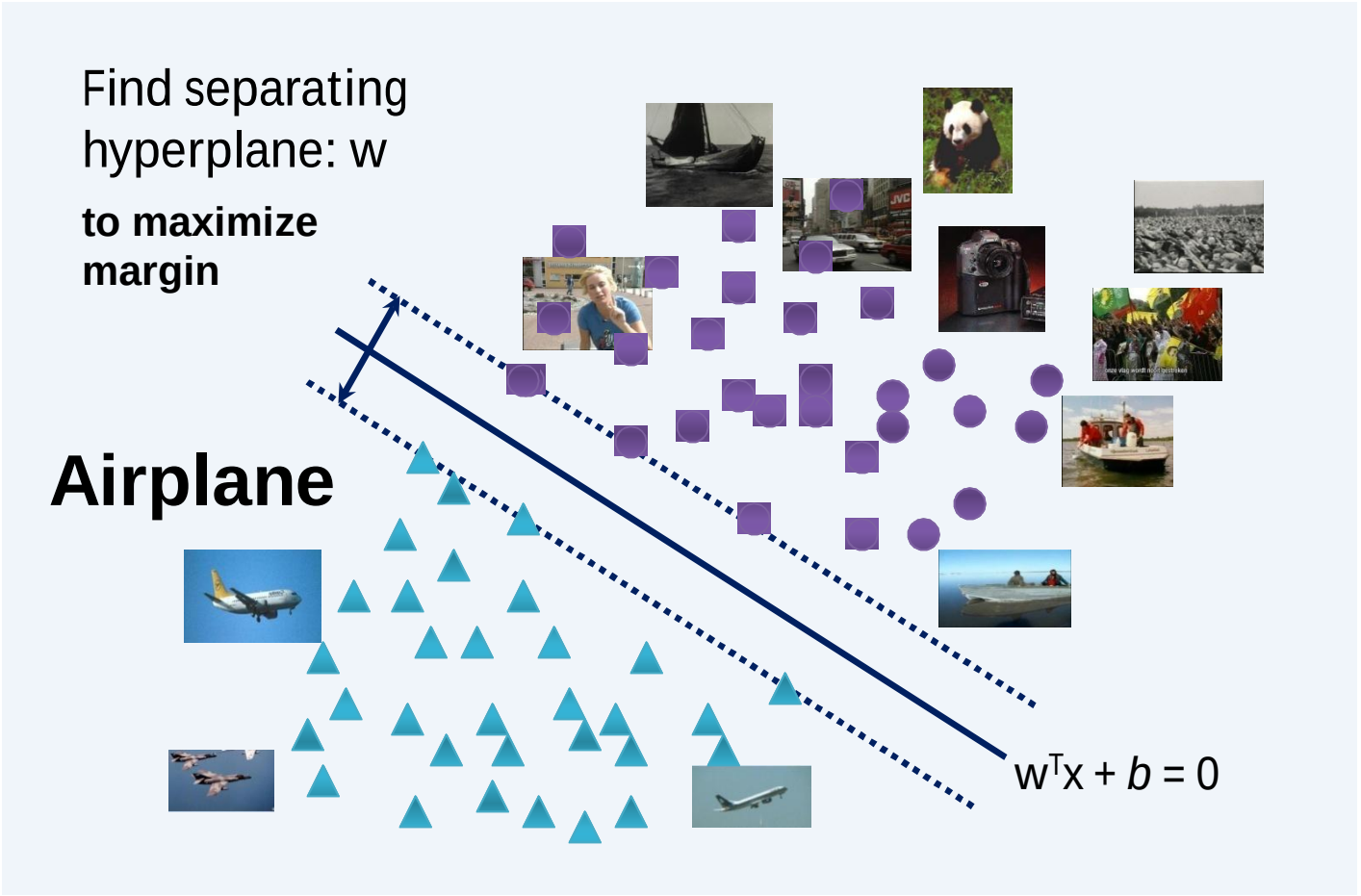








Build Classifier



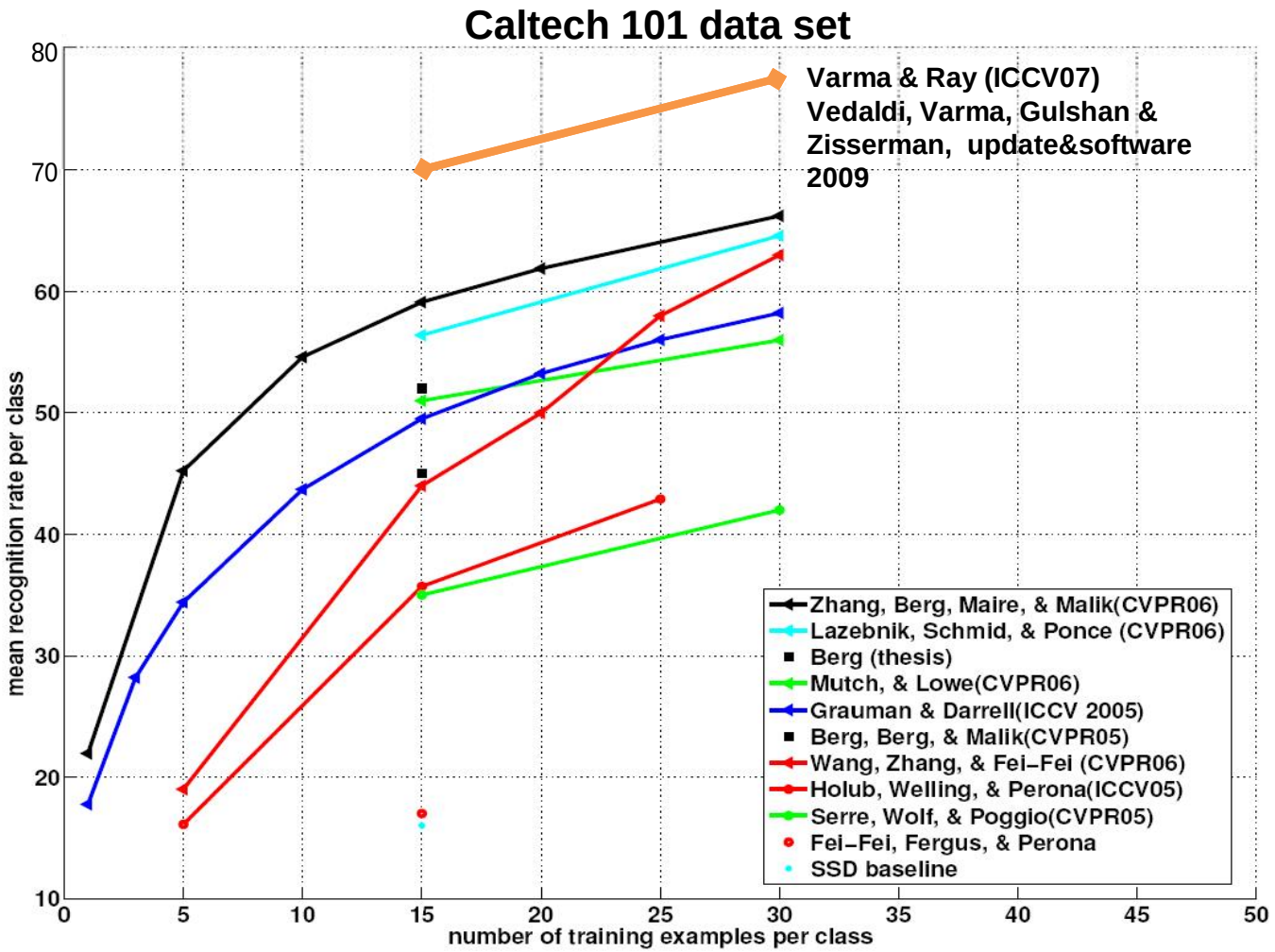
Decision function: $f(x) = \text{sign}(w^T x + b)$

$w^T x_i + b > 0$ if label $y_i = +1$

$w^T x_i + b < 0$ if label $y_i = -1$

Rapid Advances in Image Annotation

- "Moore's Law": accuracy doubles in about 2 years

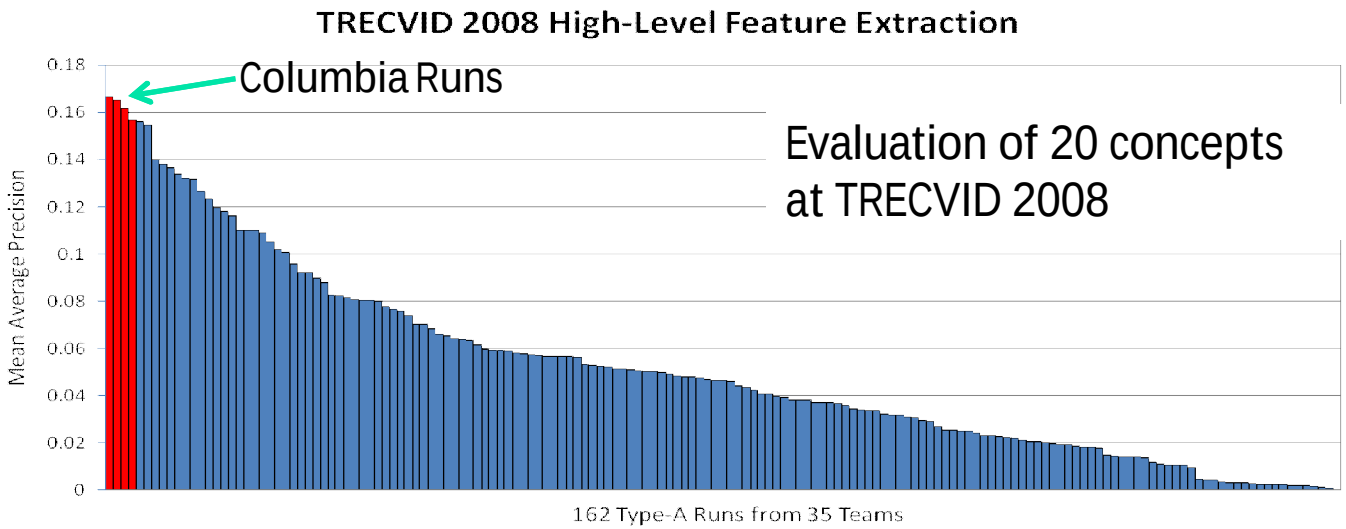


CuZero: tag videos with 400+ classifiers

COLUMBIA UNIVERSITY IN THE CITY OF NEW YORK
CuZero
digital video | multimedia laboratory

concept detection models:
objects, people, location, scenes, events, etc

airplane airplane_takeoff airport_or_airfield armed_person building car cityscape crowd desert dirt_gravel_road entertainment explosion_fire forest highway hospital insurgents landscape maps military military_base military_personnel mountain nighttime people-marching person powerplants riot river road rpg shooting smoke tanks urban vegetation vehicle waterscape_waterfront weapons weather



LSCOM Ontology Defines 1000 Concept Classes

(IBM, Columbia, CMU '06)

- Large Scale Concept Ontology for Multimedia
  - Broadcast news video
  - Defined by experts and actual intelligence users
  - Selection criteria: *useful, observable, detectable*
  - 30M+ labels for 449 concepts annotated over 60,000+ video shots
  - Download site
    - http://www.ee.columbia.edu/dvmm/lscom/

TRECVID: Detection Examples

- Top five classification results

Classroom

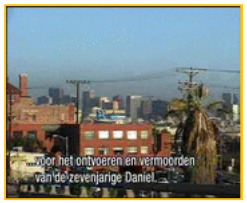



Demonstration Or Protest





Cityscape

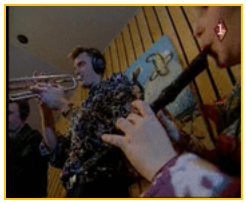


Airplane flying





Singing





What can a small recognition engine do?

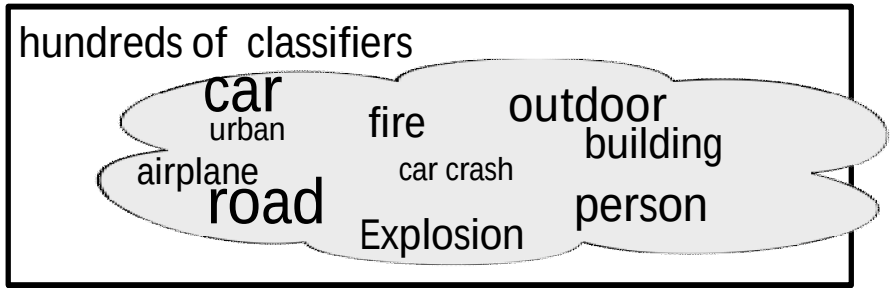
- Several video search engines incorporate ~1000 visual concepts
  - IBM IMAR
  - U. Amsterdam MediaMill
  - CMU Informedia
  - Columbia CuZero
- But this is still a noisy and relatively small vocabulary

Cope with Small Noisy Vocabulary

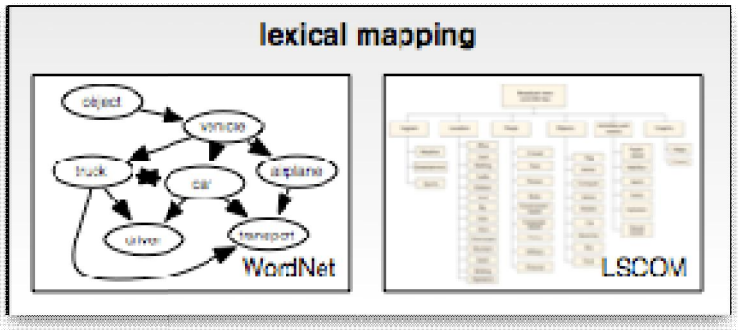
- Given a search topic, users often have difficulty in choosing matched concept classifiers

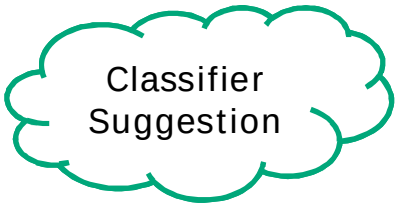
Find shots of something burning with flames visible

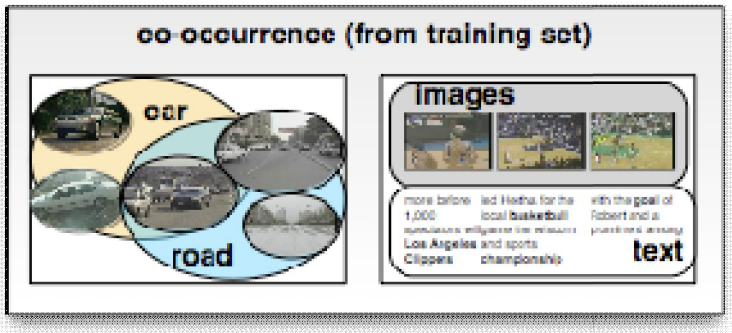
Cannot find matched classifiers!
Which classifiers work?

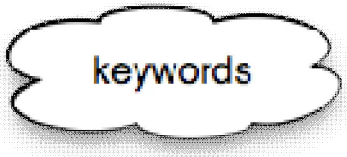


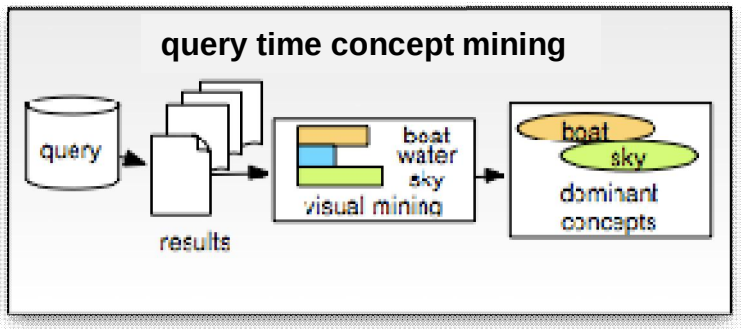
Lessons from IR: word-concept query expansion





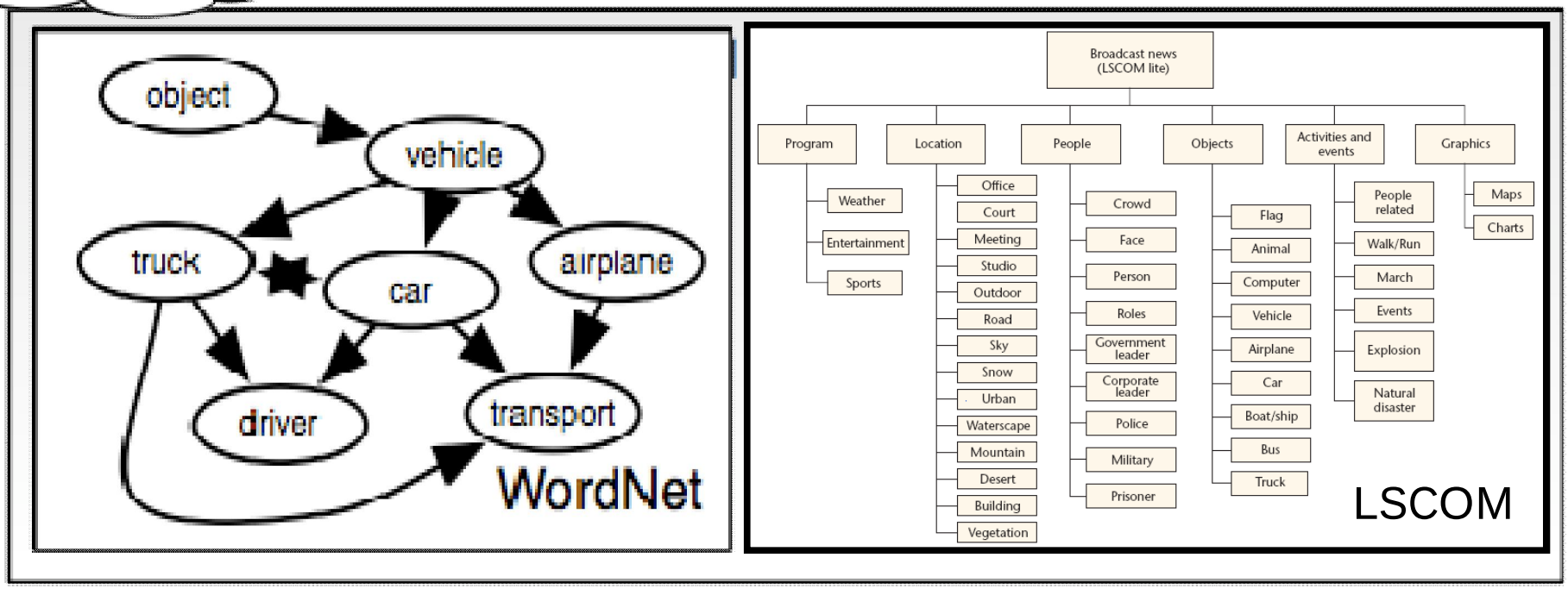






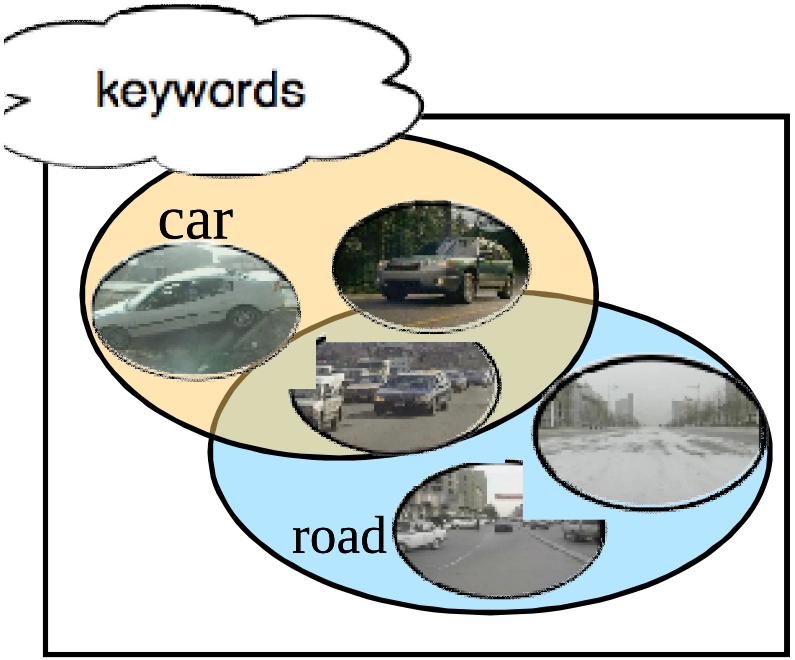
Lexical mapping

keywords

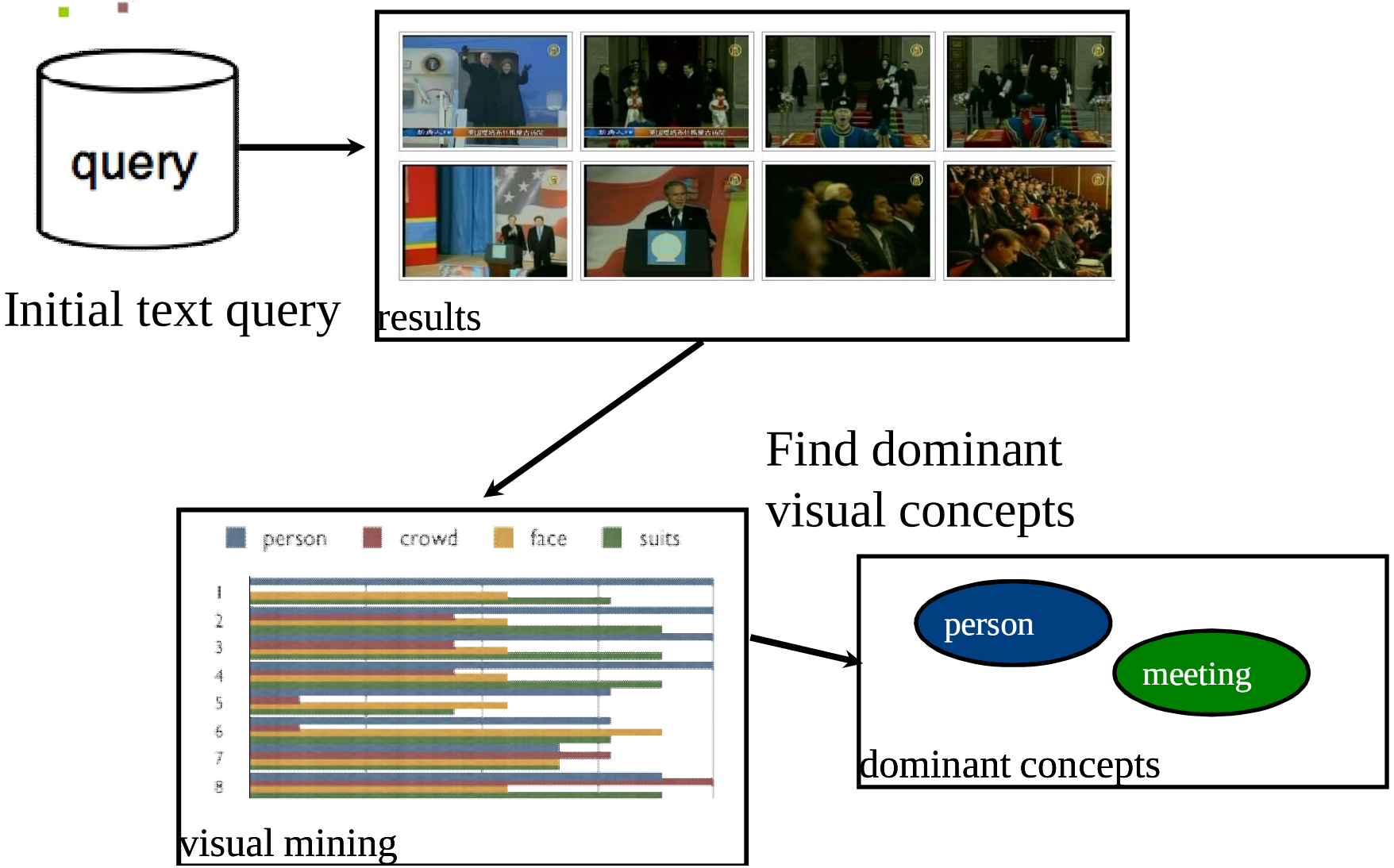


Mapping keywords to concept definition, synonyms, sense context, etc

Co-occurrent concepts

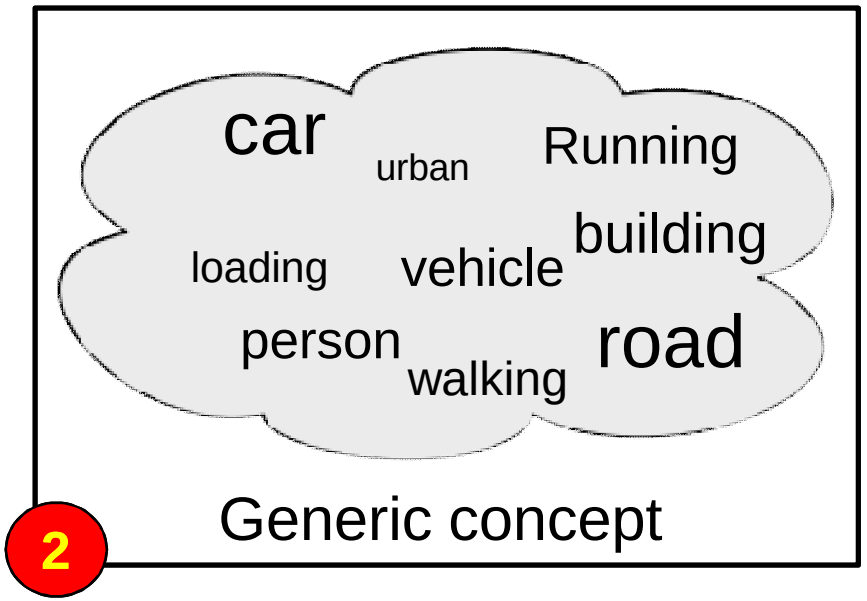


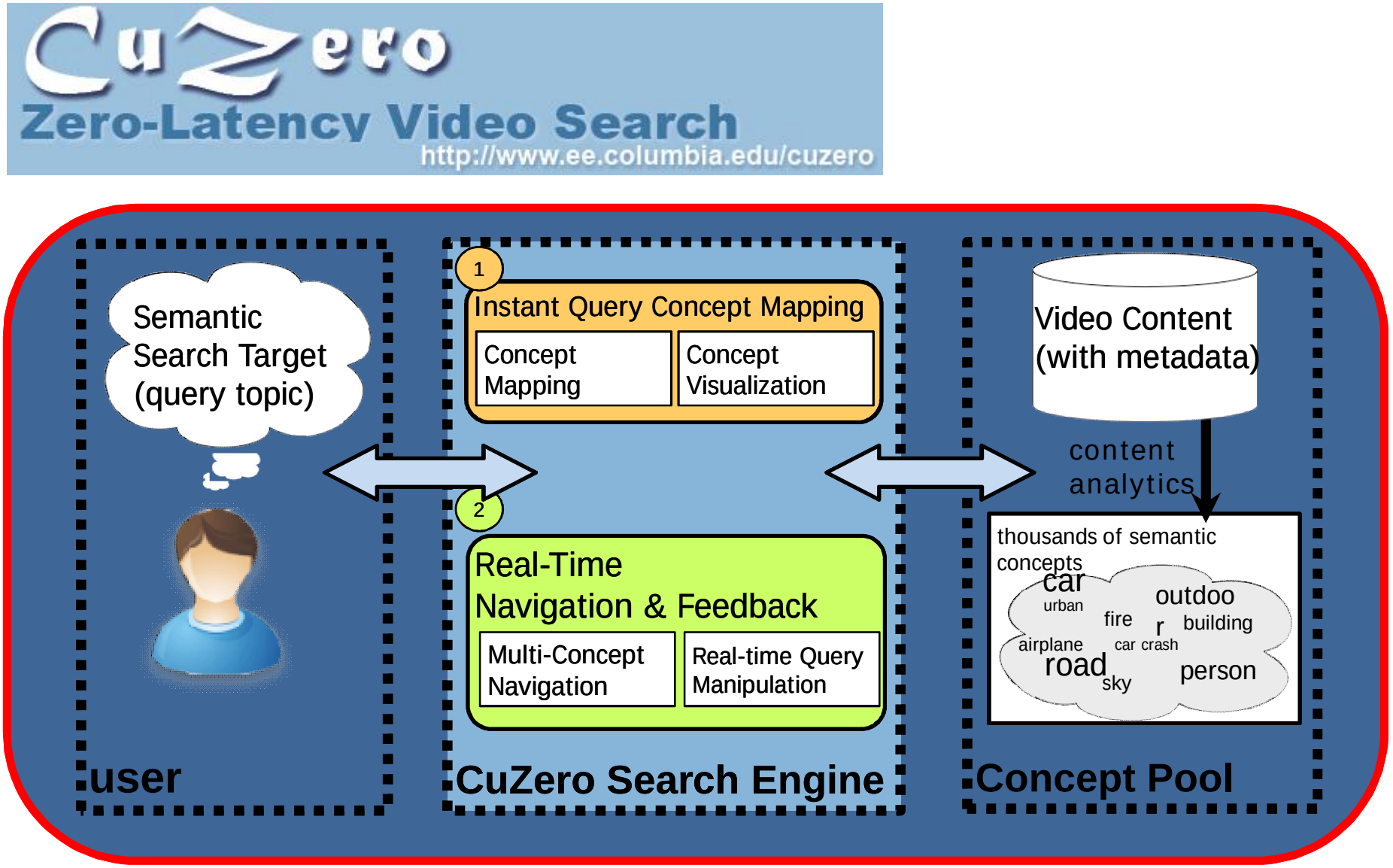
Query-Time Concept Mining



Use visual examples to compensate the difficiency

Query: “find person running around a building”





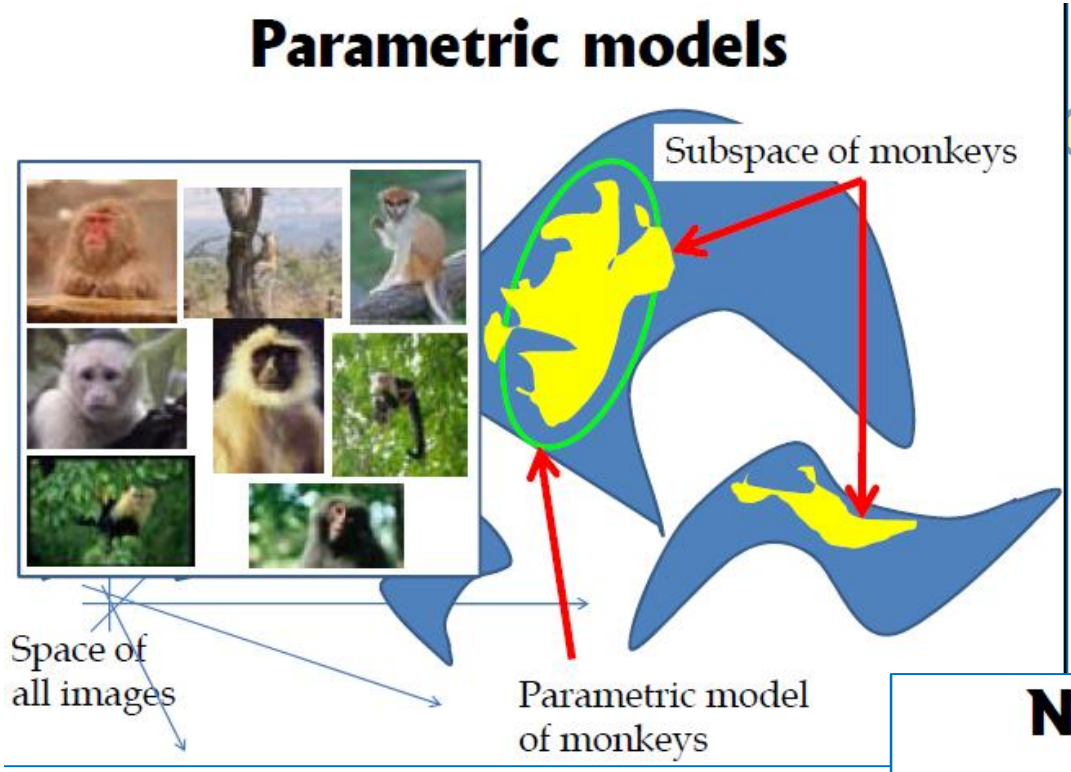
(Zavesky and Chang ,
Multimedia Info Retrieval MIR '08)

Demos

- Find lake front buildings in the park
- Find person walking around building
- Find a car on a road in a snowy condition

Scalability:
Robust recognition is still developing

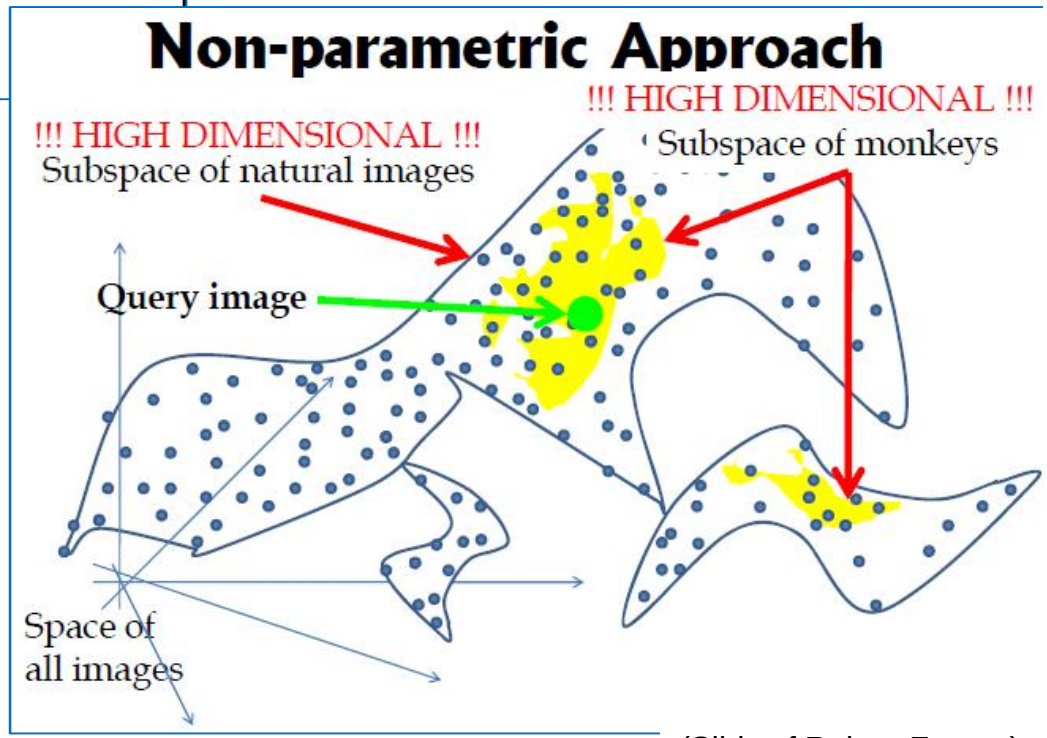
- Many concept classes are very challenging
- Face, background scene, OCR are relatively mature compared to others



Some classifiers are really hard to build

(Slide from Robert Fergus' site)

Instead, try visual matching with A LOT of data



(Slide of Robert Fergus)

Image matching with local interest points

Matching SIFT points
[Lowe, 1999]

- Measure useful information
  - Image similarity
  - Copy/source identification
  - Discover possible transformation between images
- Can be done efficiently

Visual Matching Spurs Interesting Applications

use image pixels to find additional useful information

product search

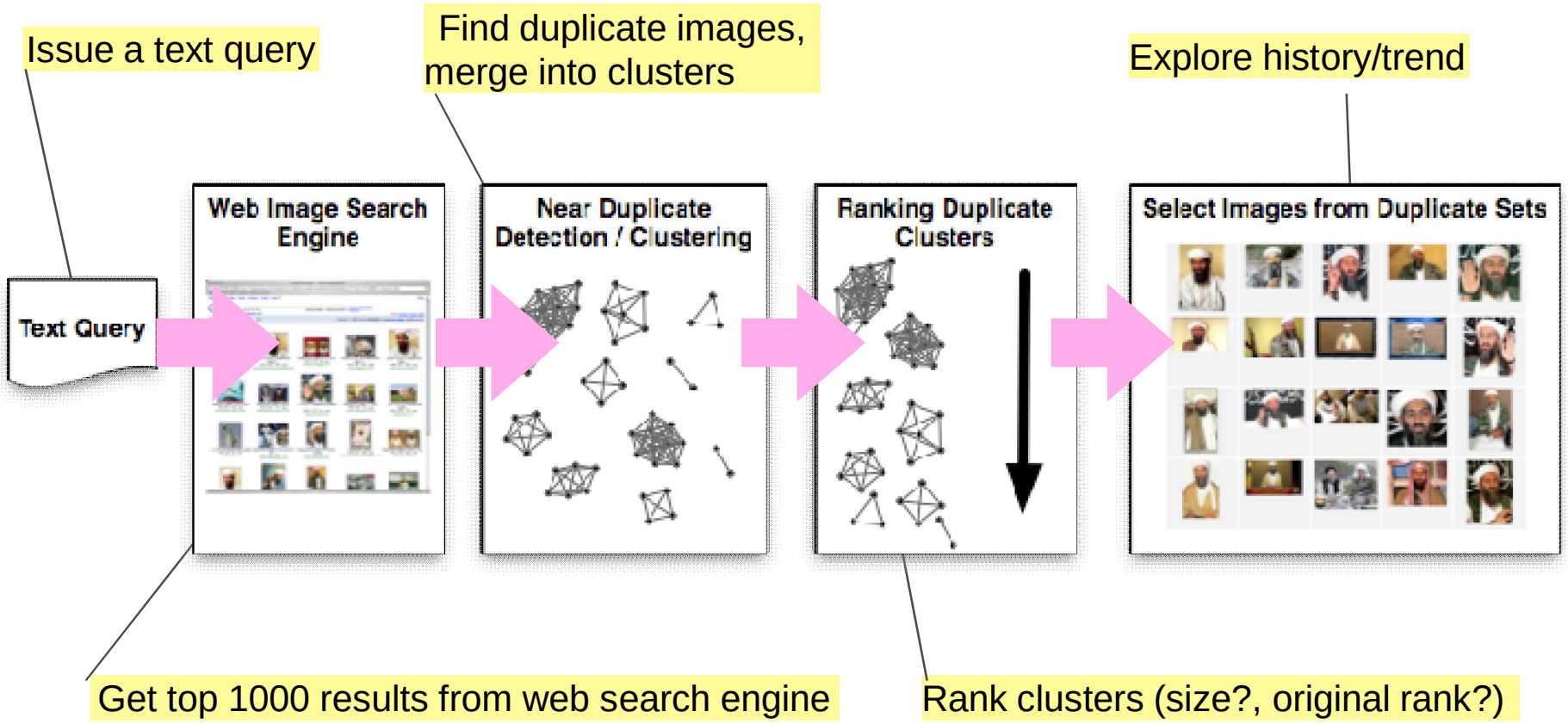
landmark search

document search

Recent applications: iPhone App SnapTel Google Goggles

Demo: Flickr Object Retrieval

Image matching for search result summary



Matching Reveals Image Provenance

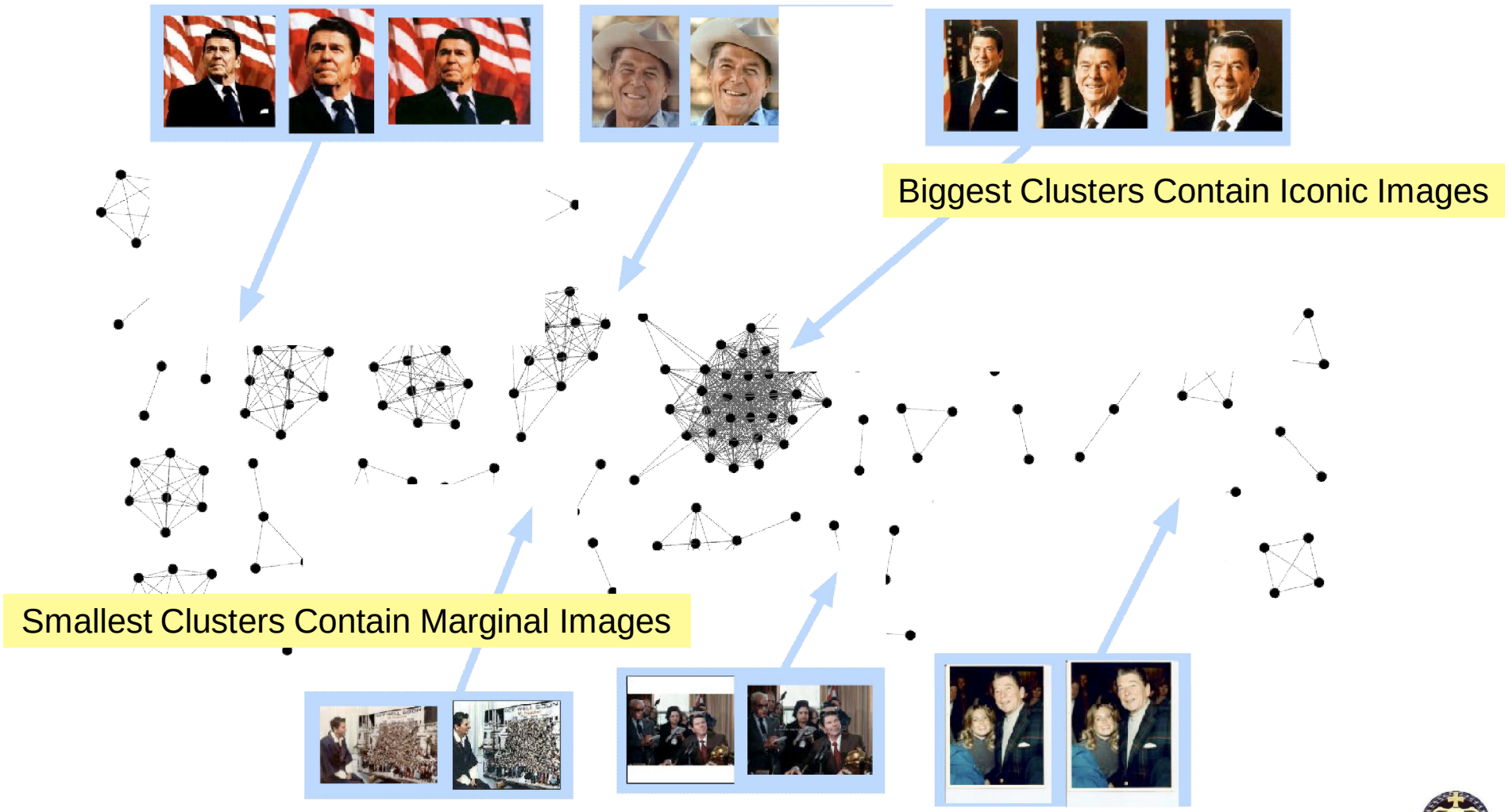


Image Matching Reveals Possible Manipulations


Original


Scaled


Cropped


Gray

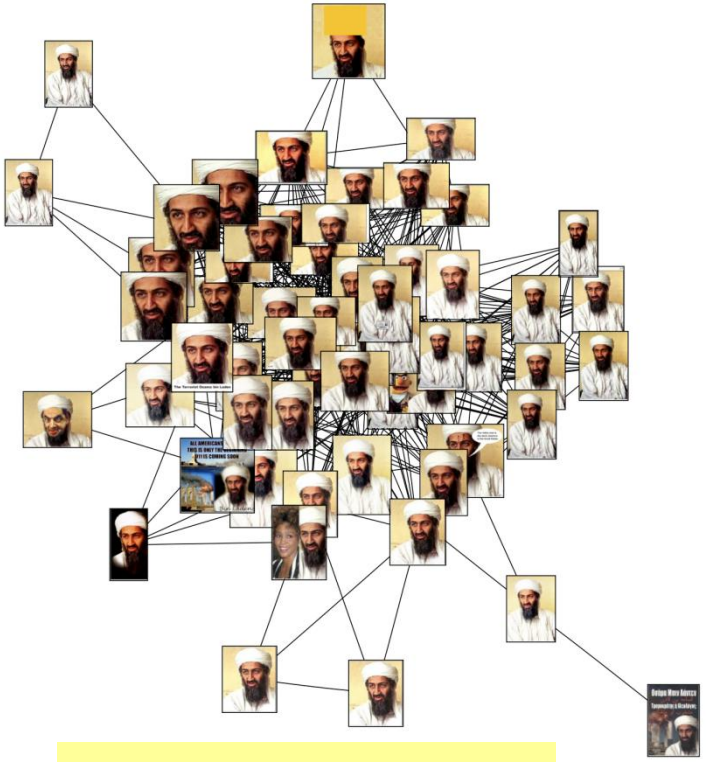
Overlay


Insertion

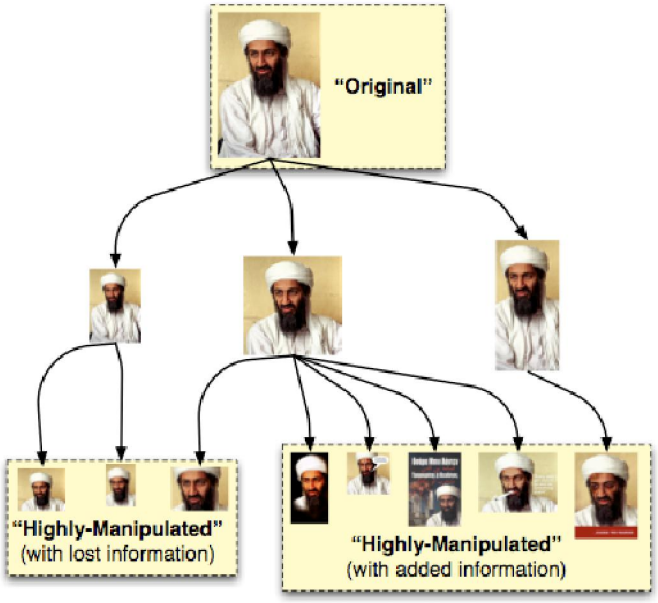
- Given a returned image set, detect possible manipulations
- Each implies editing direction (one image derived from other)
- Other possible manipulations: color correction, multiple compression, sharpening, blurring

Deeper Analysis of Search Results: Visual Migration Map (VMM)

Kenndy and Chang, ACMMM 08

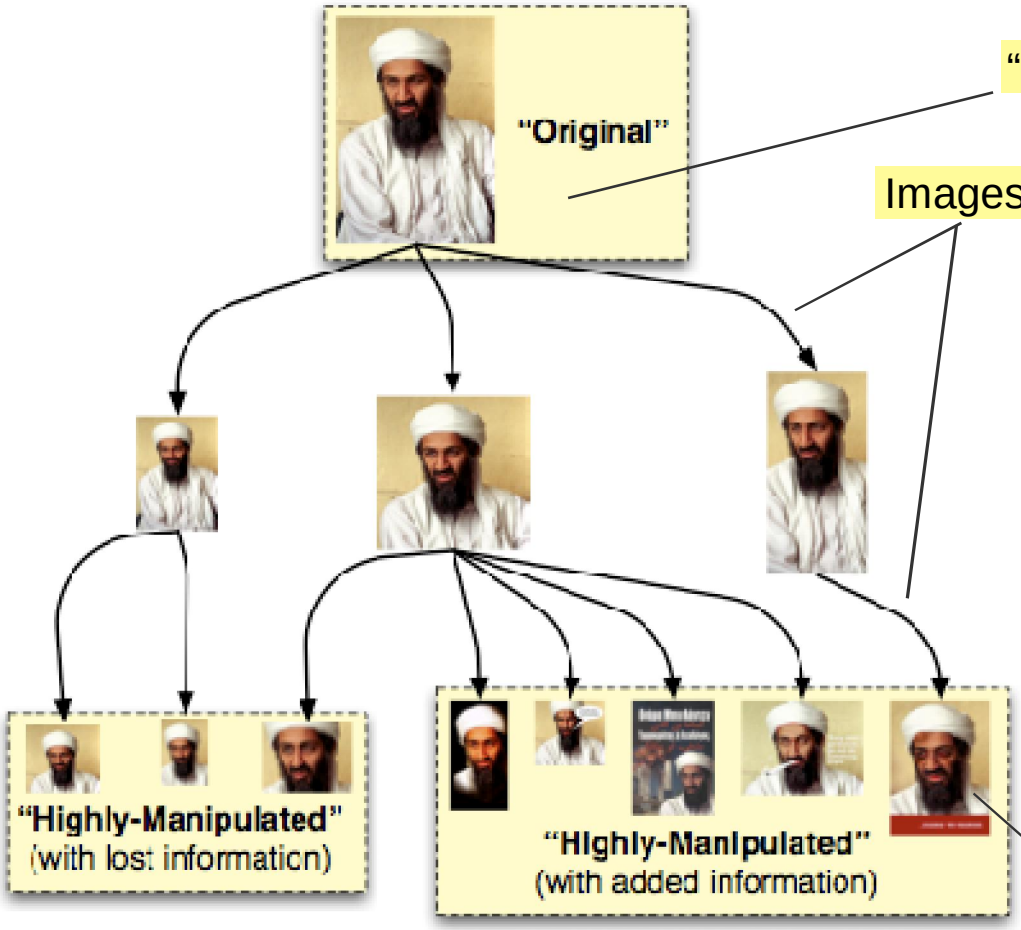


Duplicate Cluster



Visual Migration Map

Visual Migration Map (VMM)



"Most Original" at the root

Images Derived through Series of Manipulations

VMM uncovers history of image manipulation and plausible dissemination paths

"Most Divergent" at the leaves

VMM Applications: Geographic/Cultural Dispersion

Manipulations Often Correlate with Perspectives

Joke Website:
“Every time I get stoned, I go and do

Democratic National Committee Site:
“Capture Osama Bin Laden!”

.ml

http://www.democrats.org/page/petition/osama

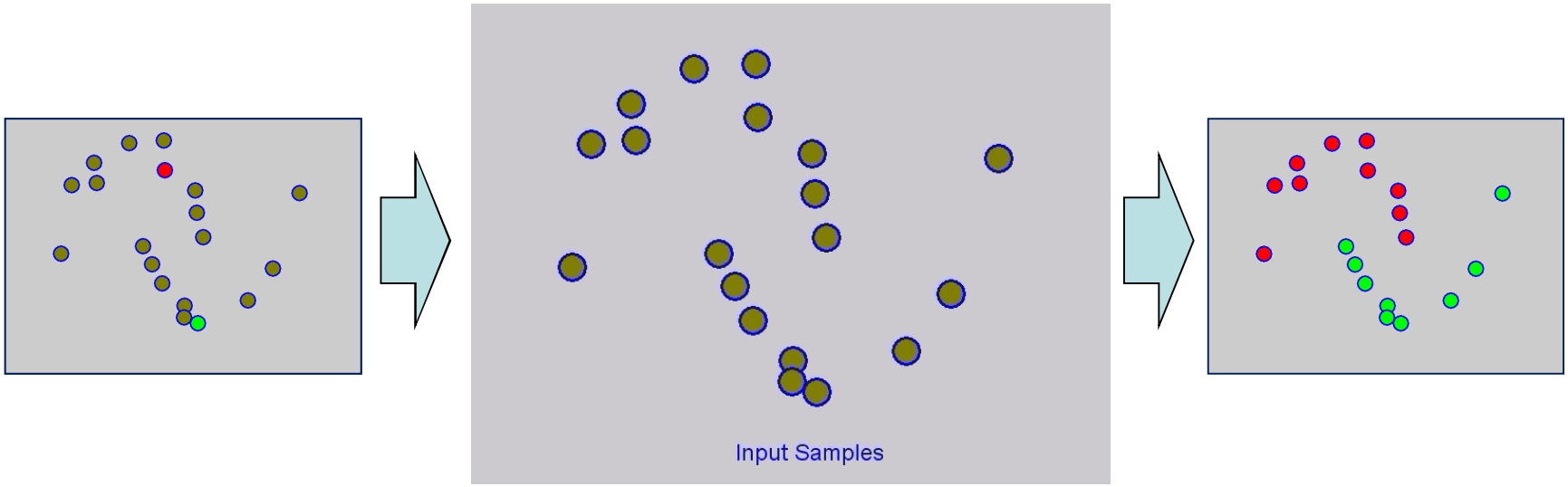
Daily Excelsior Newspaper:
“Further Details of Bin Laden Plot Unearthed: ABC Report.”

http://www.dailyexcelsior.com/00jan31/inter.htm

e:
.den - My Idol of All Time!”

http://www.myspace.com/mamu_potnoi

Image Matching + Manifold Graph Allows Useful Propagation of Information



Input samples with sparse labels

Label propagation on graph

Label inference results

Positive Negative

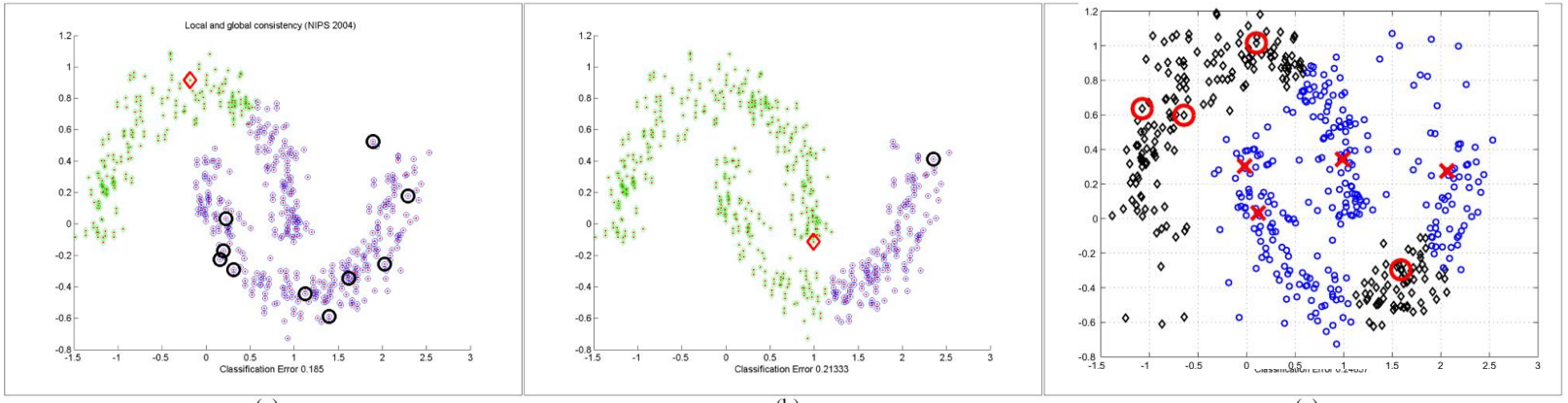
Unlabeled Positive Negative

$$f^* = \min_f Q(f, y, \mathcal{G}(V, W))$$

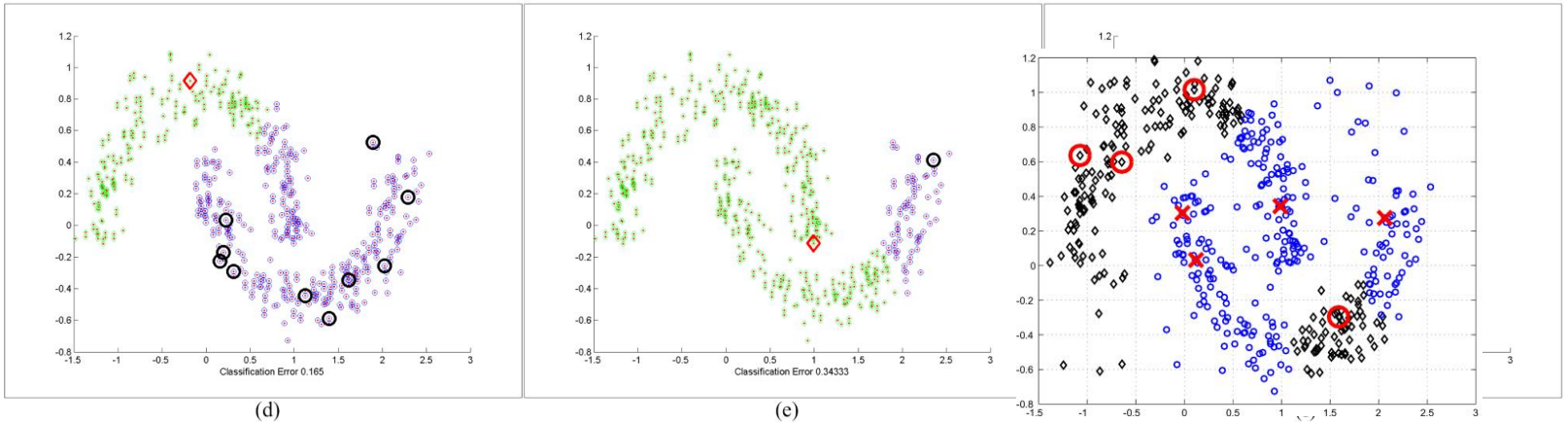
$\mathcal{G}$ -- graph
V -- graph node
W -- weight matrix
$\mathcal{Q}$ -- risk function
f -- classification
y -- label matrix

Non-Trivial Issues

LGC Method



GFHF Method



Unbalanced Labels

Bad Label Locations

Noisy Labels

An active area in Machine Learning

- Given initial labels, Y, find classification function F over graph nodes

Label smoothness

Fit known labels

$$\begin{aligned} \mathcal{Q}(F) &= \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\left\|\frac{F_{i.}}{\sqrt{D_{ii}}}-\frac{F_{j.}}{\sqrt{D_{jj}}}\right\|^2 + \mu\sum_{i=1}^{l}\|F_{i.}-Y_{i.}\|^2 \\ &= \mathrm{tr}\{F^\top LF + \mu(F-Y)^\top(F-Y)\} \end{aligned}$$

(Zhou, et al NIPS04)

- Gaussian fields & Harmonic functions (*Zhu et al ICML03*)

$$\mathcal{Q}(F) = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\|F_{i.}-F_{j.}\|^2$$

1) $\triangle F = 0$ on unlabeled data, where $\triangle = D - W$ is the graph Laplacian;

2) $F_{i.} = Y_{i.}$ on labeled data.

Graph Transduction via Alternating Minimization (GTAM)

(Wang, Jebara, Chang, ICML08) (Wang and Chang, CVPR09)

-- Optimization over both Labels (Y) and Prediction (F)

$$\mathcal{Q}(\mathbf{F},\mathbf{Y}) = \frac{1}{2}\text{tr}\left\{\mathbf{F}^T\mathbf{L}\mathbf{F} + \mu(\mathbf{F}-\mathbf{V}\mathbf{Y})^T(\mathbf{F}-\mathbf{V}\mathbf{Y})\right\}$$

- Propagation Step
  - given label (Y), propagate over graph, predict F

$$\frac{\partial \mathcal{Q}}{\partial \mathbf{F}^*} = 0 \Rightarrow \mathbf{F}^* = (\mathbf{L}/\mu + \mathbf{I})^{-1}\mathbf{V}\mathbf{Y} = \mathbf{P}\mathbf{V}\mathbf{Y}$$

- Label Diagnosis and Selection Step
  - add good labels or remove bad labels

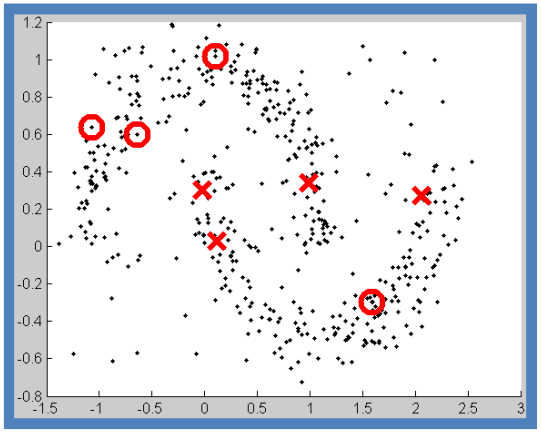
$$\mathcal{Q}(\mathbf{Y}) = \frac{1}{2}\text{tr}\left(\mathbf{Y}^T\mathbf{V}^T\left[\mathbf{P}^T\mathbf{L}\mathbf{P} + \mu(\mathbf{P}^T-\mathbf{I})(\mathbf{P}-\mathbf{I})\right]\mathbf{V}\mathbf{Y}\right)$$

(TAG demo)

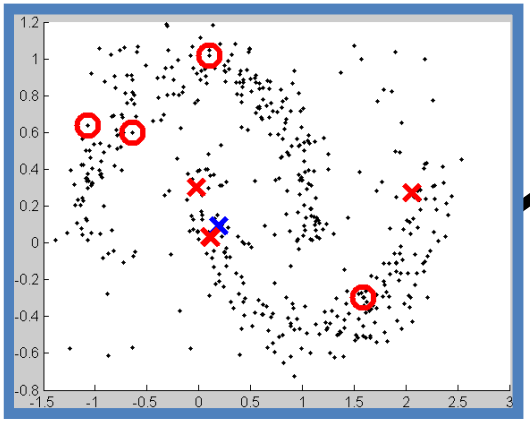
Iteratively Tune and Propagate Label Information

(Wang and Chang CVPR'09)

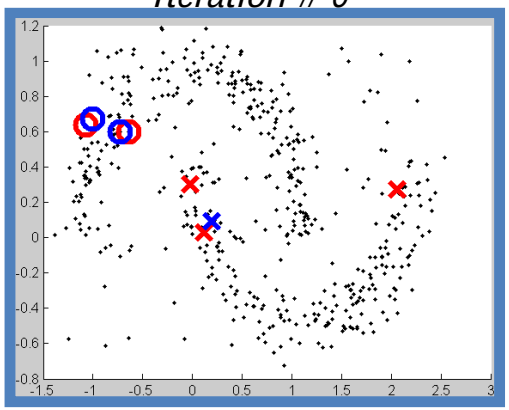
Initial Labels

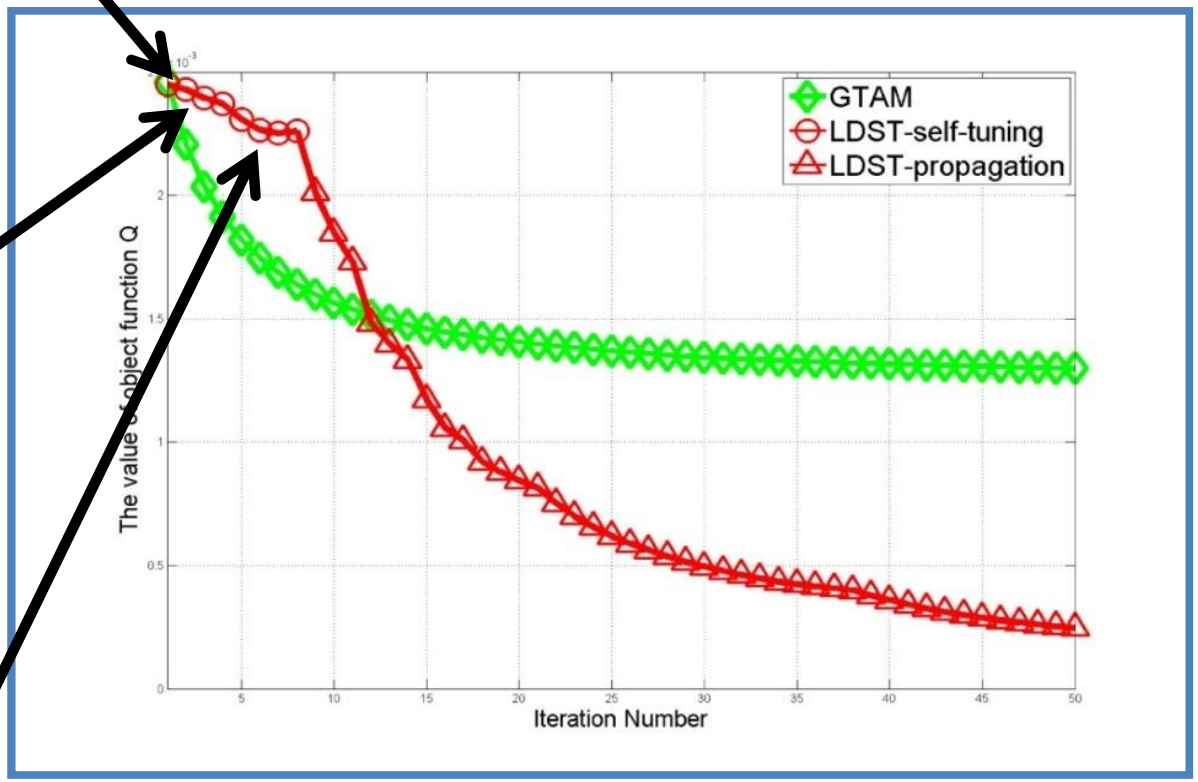


Iteration # 2



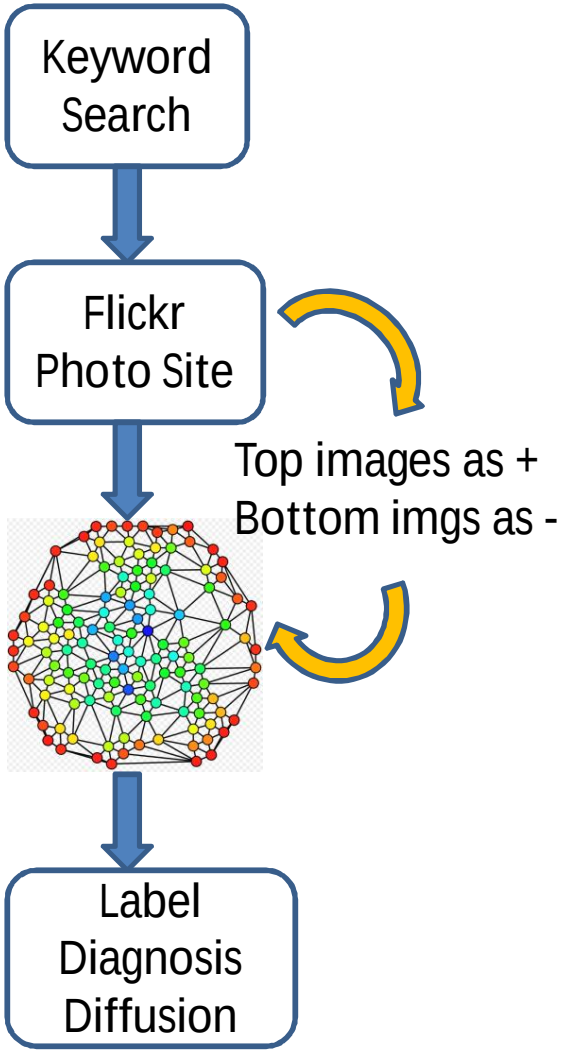
Iteration # 6



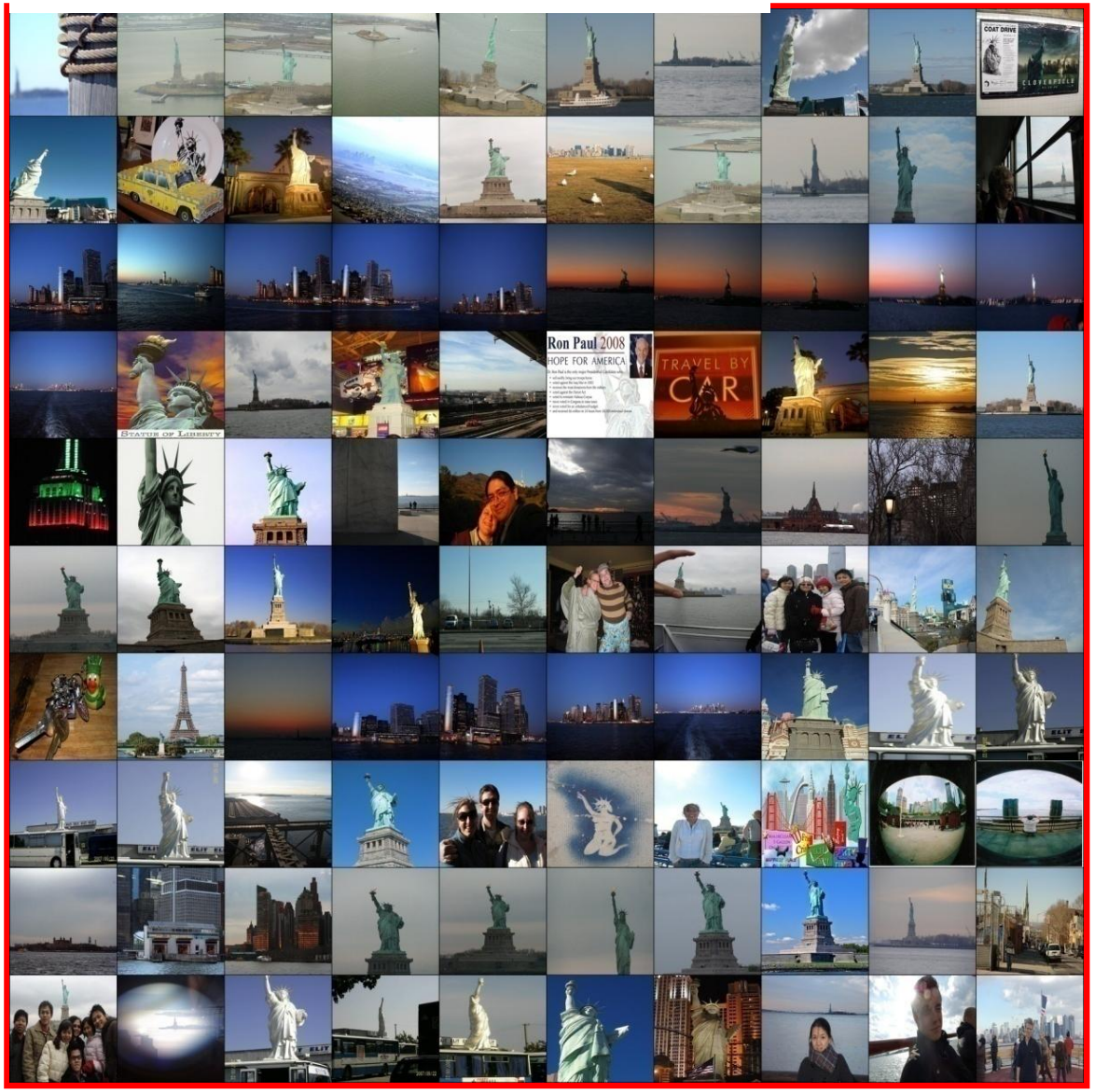


The values of the cost function Q during optimization procedure of LDST and GTAM methods.

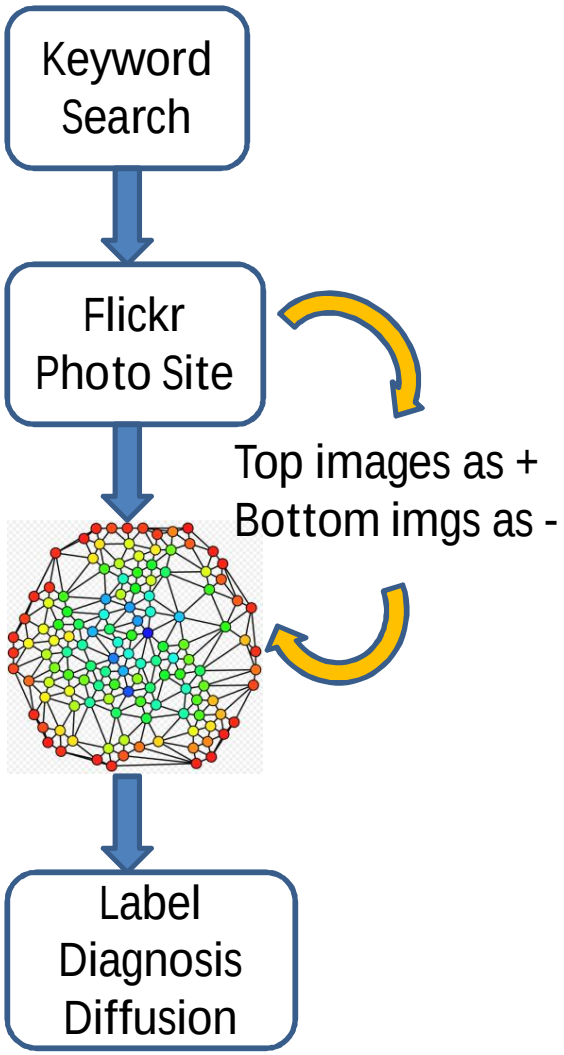
Tune and propagate top scores through image graph



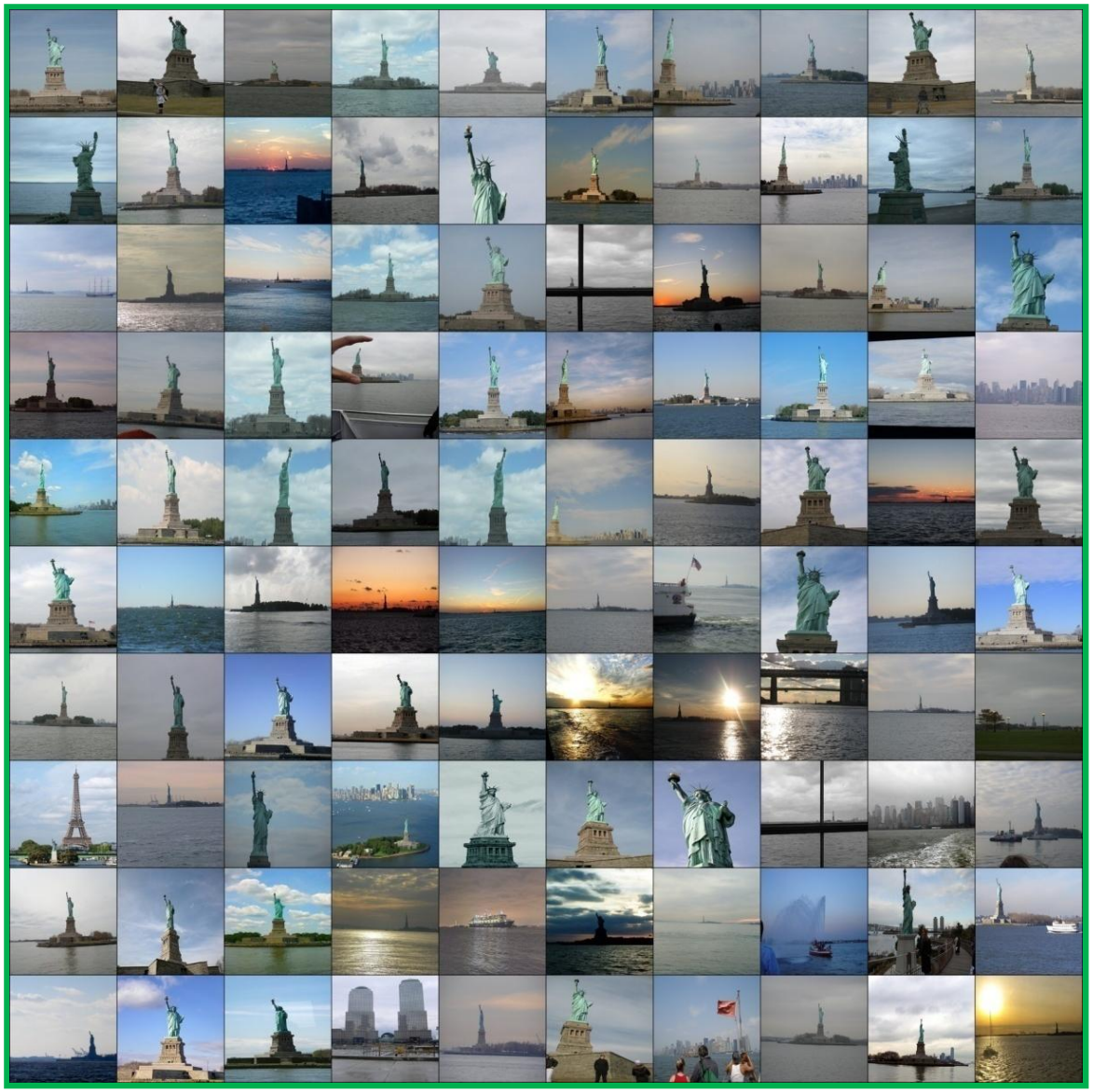
Google Search “Statue of Liberty”



Tune and propagate top scores through image graph

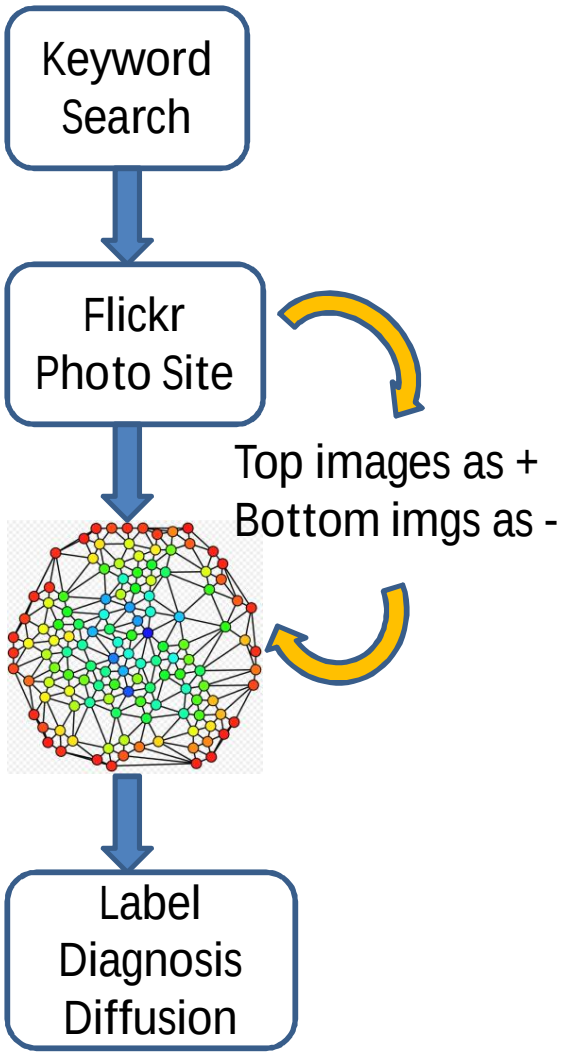


Rerank

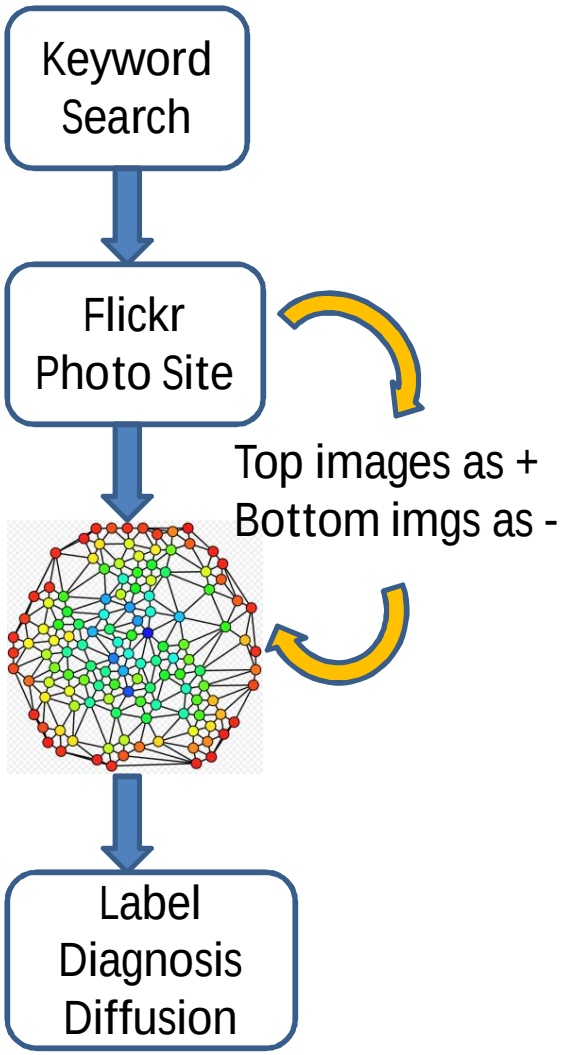


Tune and propagate top scores through image graph

Google Search “Tiger”



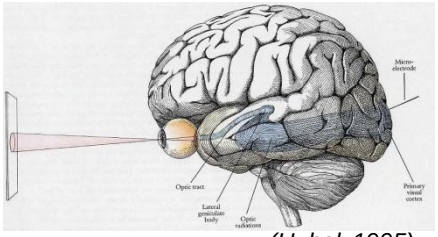
Tune and propagate top scores through image graph



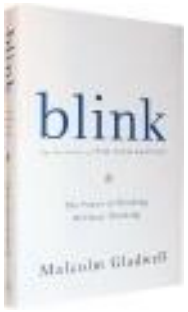
Rerank

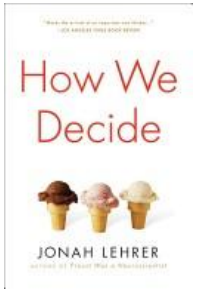
Detect Image Semantics via Brain State Decoding

- Human Vision is Superb by quick "gist" in the "Blink of an Eye"



(Hubel, 1995)



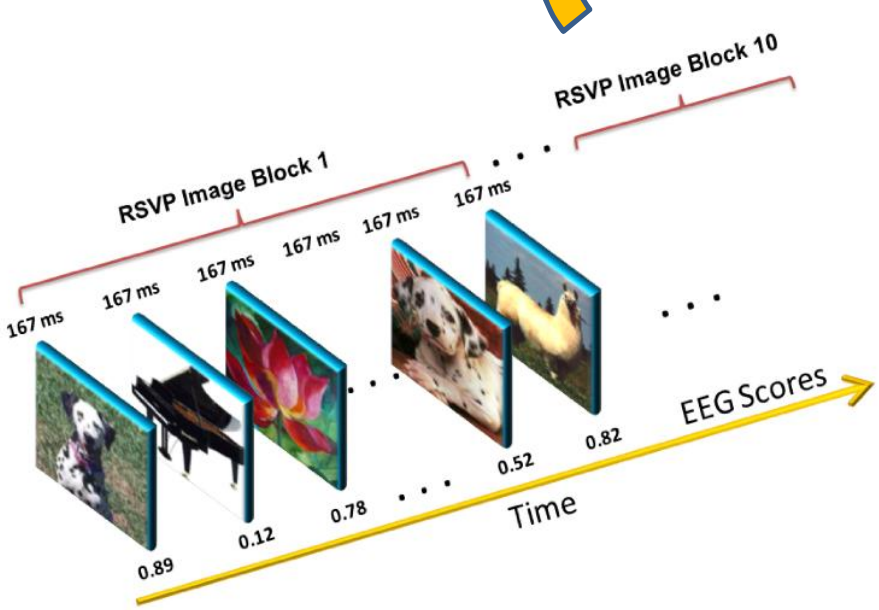


Joint work with
Prof. Paul Sajda's group

Brain Machine Interface for Image Retrieval

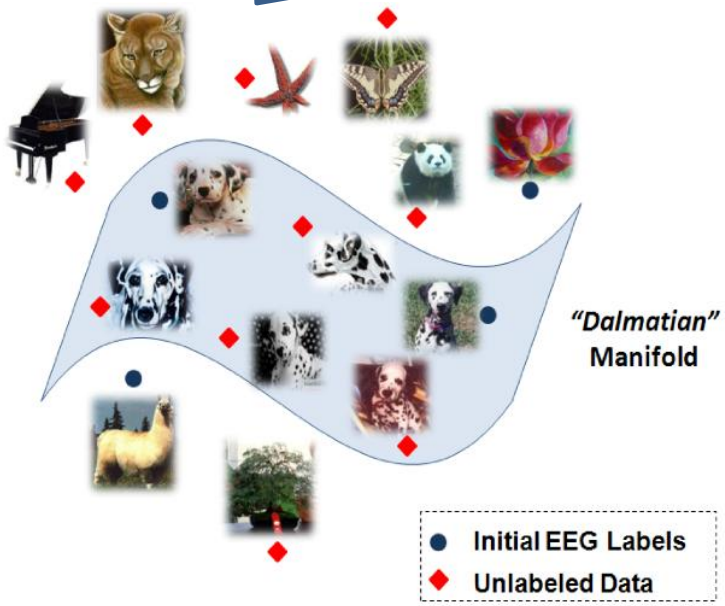
Use EEG brain signals
to detect target of interest
(video)

Use image graph to tune
& propagate information



Rapid Serial Presentation of Caltech 101 Images

C3Vision System
by Sajda et al

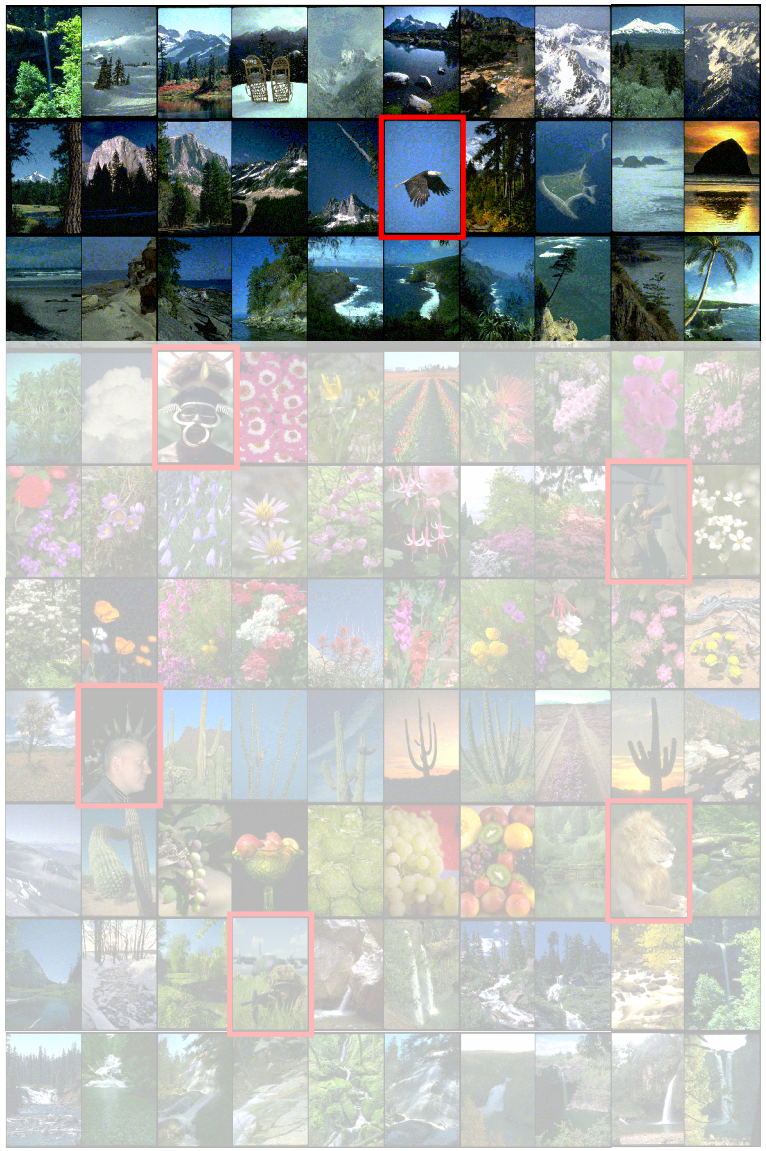


Graph-Based Visual Pattern Discovery

TAG System
by Wang and Chang

(Wang , Pohlmeyer, Hanna, Jiang, Sajda, Chang, ACMMM09)

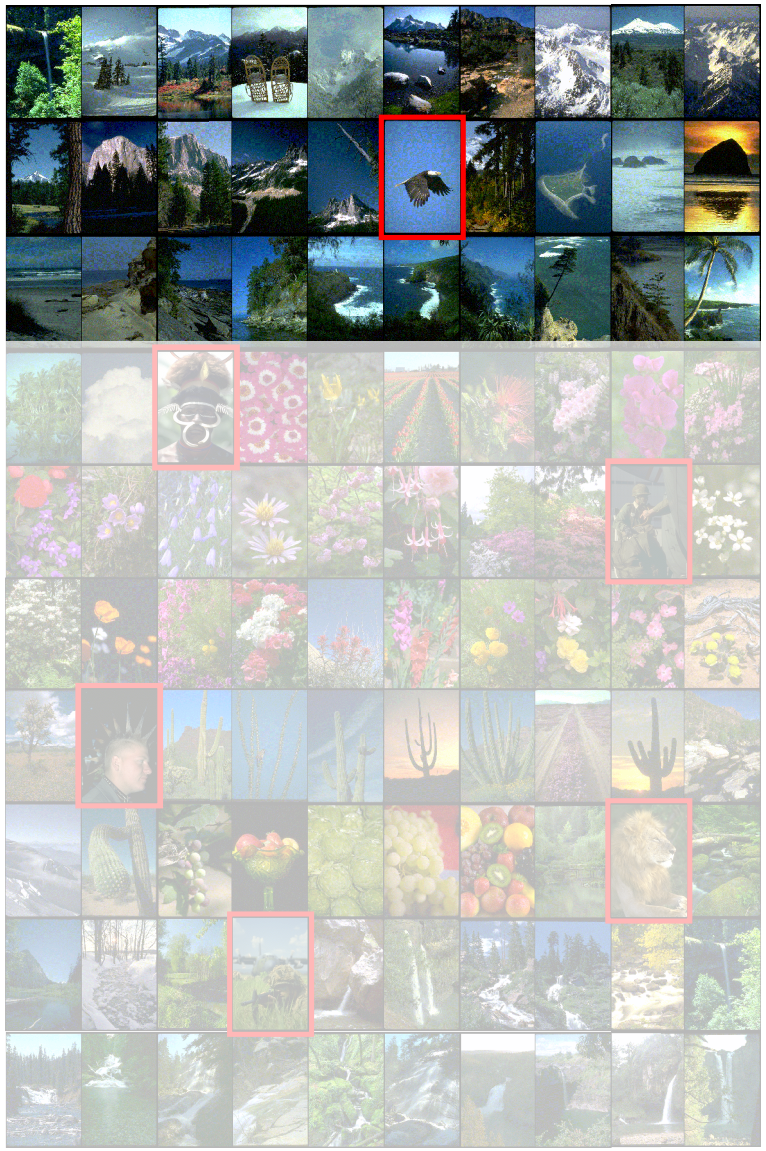
The Visual Interest Readout Experiment



Database (any target that may interest users)

User thinks about what he/she wants to search

The Paradigm

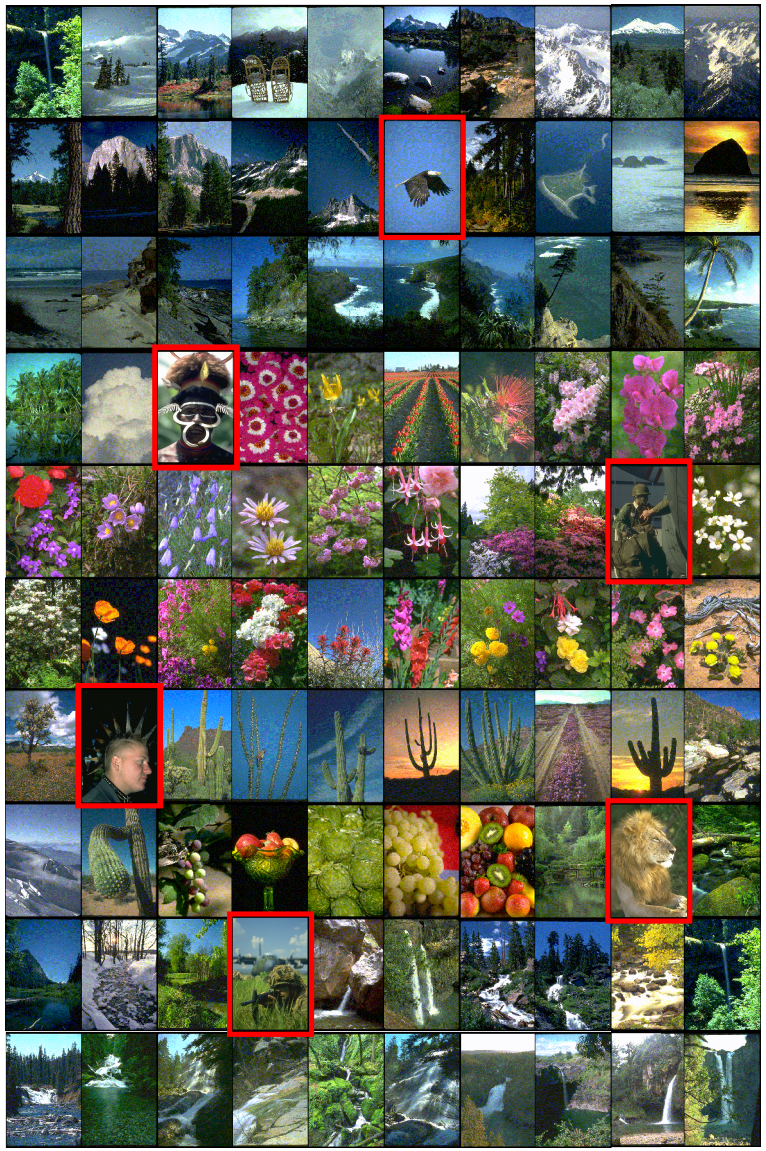


Database

Neural (EEG) decoder

Interest-scores

The Paradigm



Database

Neural (EEG) decoder

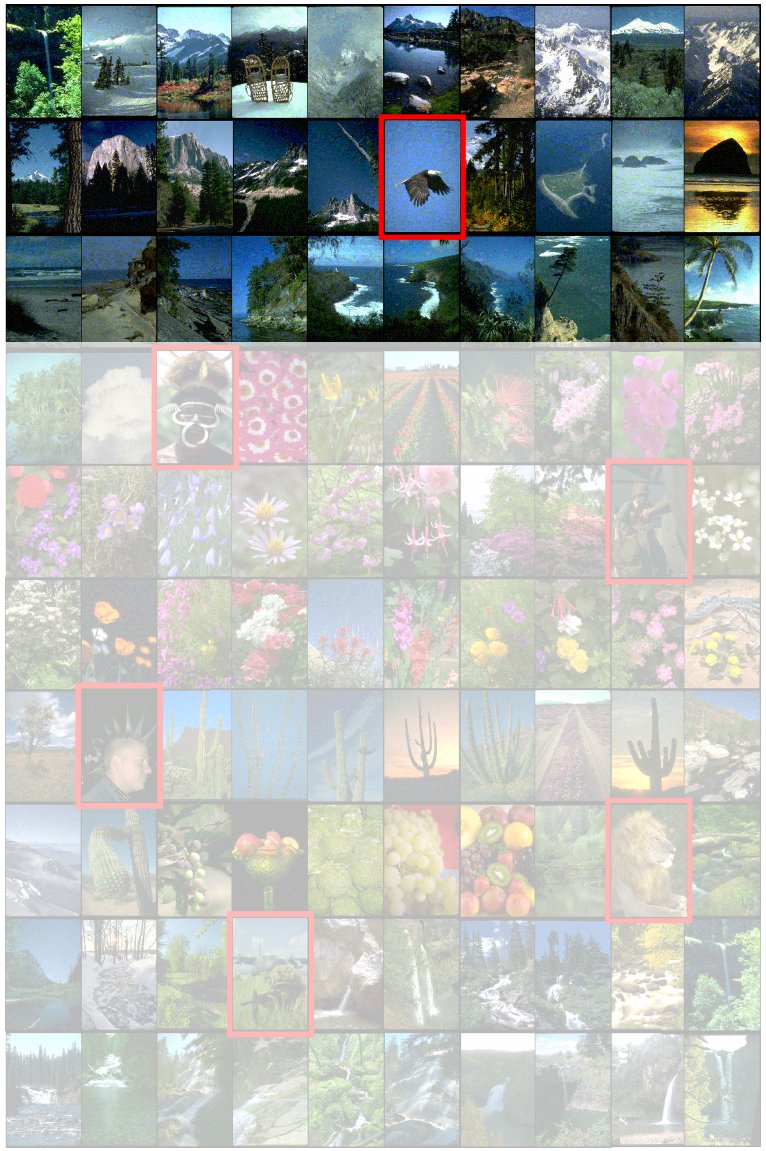
Exemplar labels (noisy)

Semi-supervised
Graph-based propagation

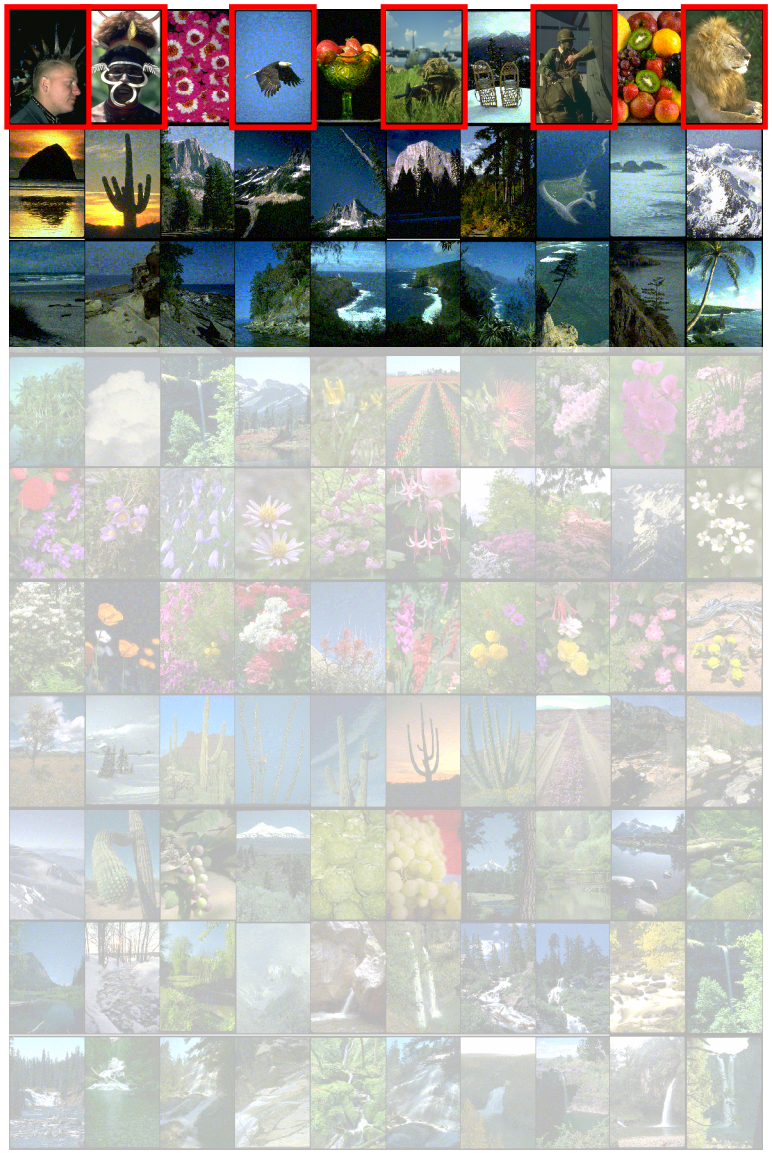
Features from the entire DB

prediction score

The Paradigm

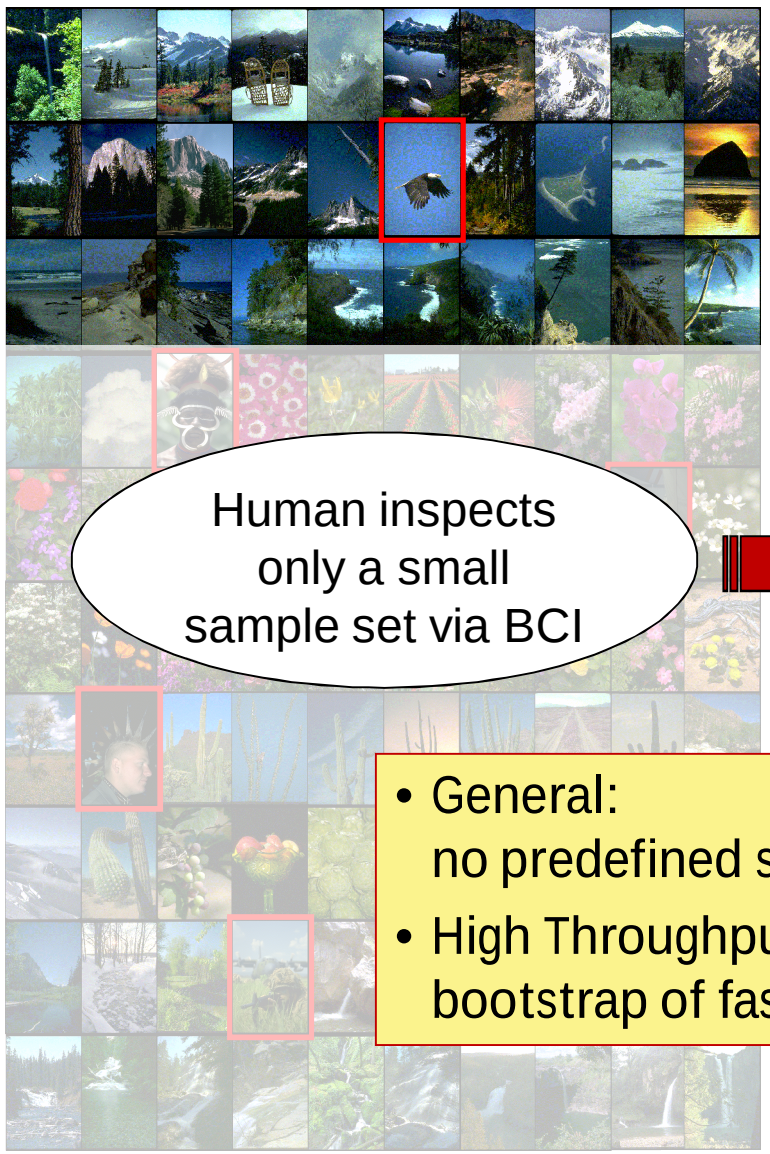


Pre-triage

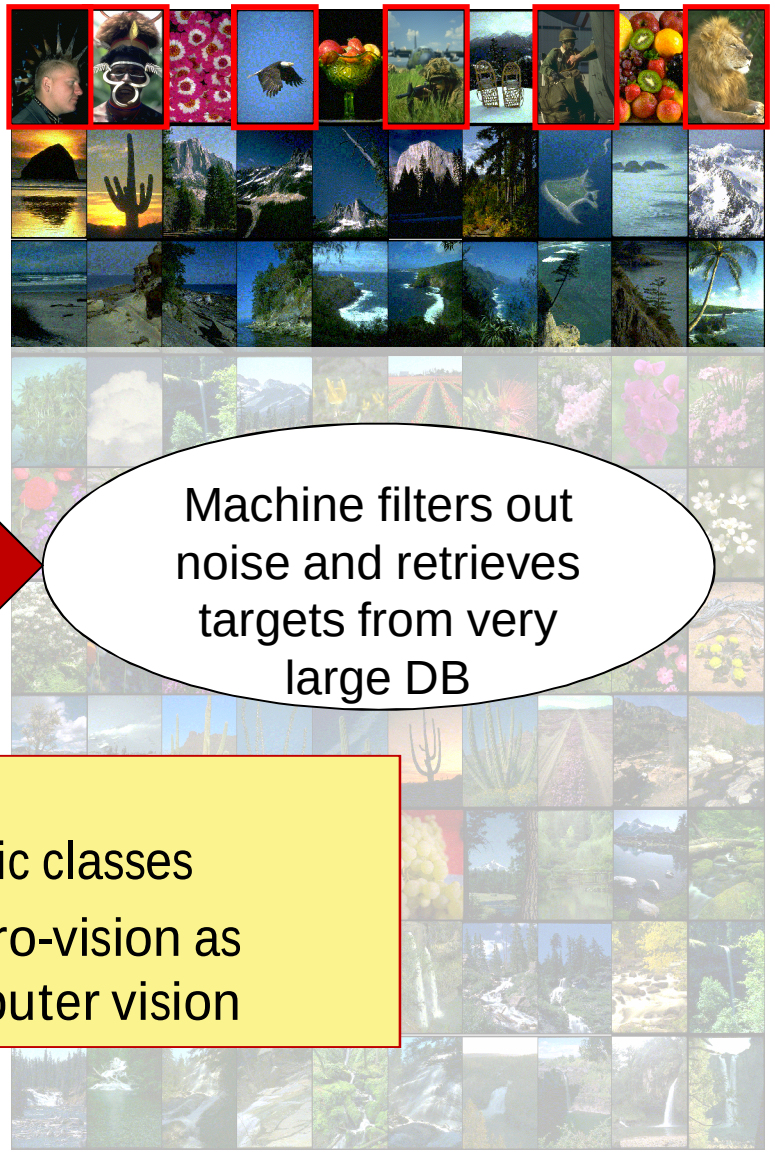


Post-triage

The Paradigm



Human inspects only a small sample set via BCI



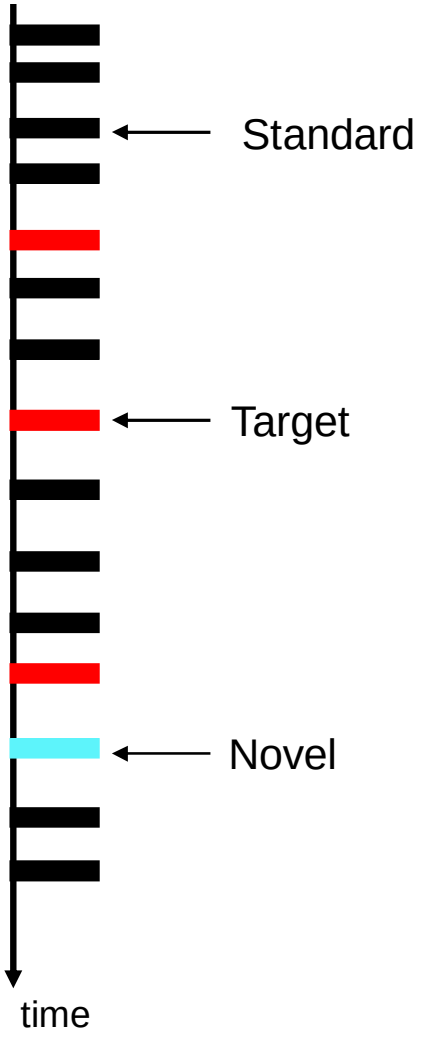
Machine filters out noise and retrieves targets from very large DB

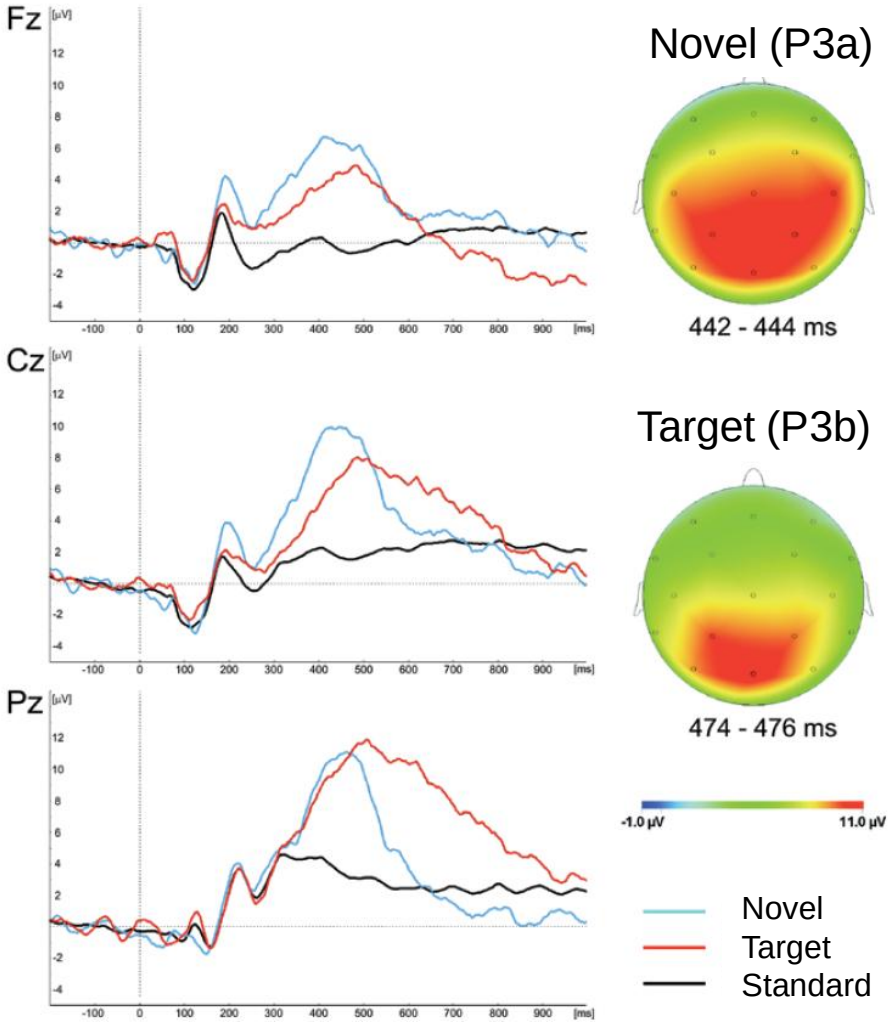
- General: no predefined semantic classes
- High Throughput: neuro-vision as bootstrap of fast computer vision

Pre-triage

Post-triage

The Neural Signatures of “Recognition”

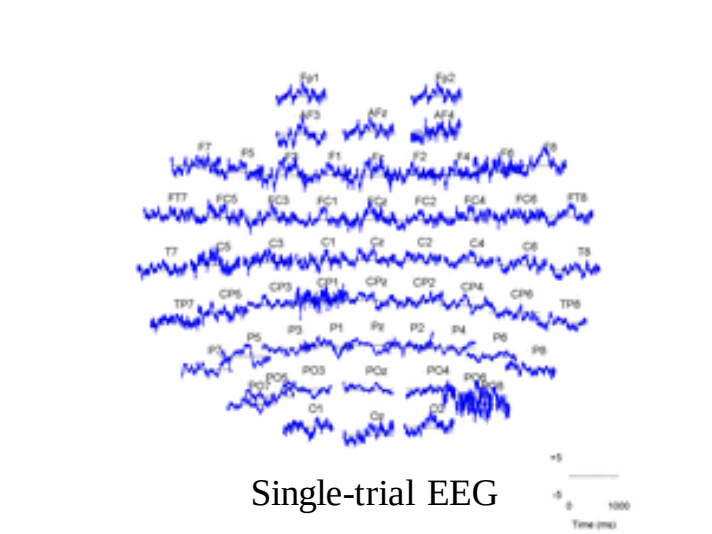




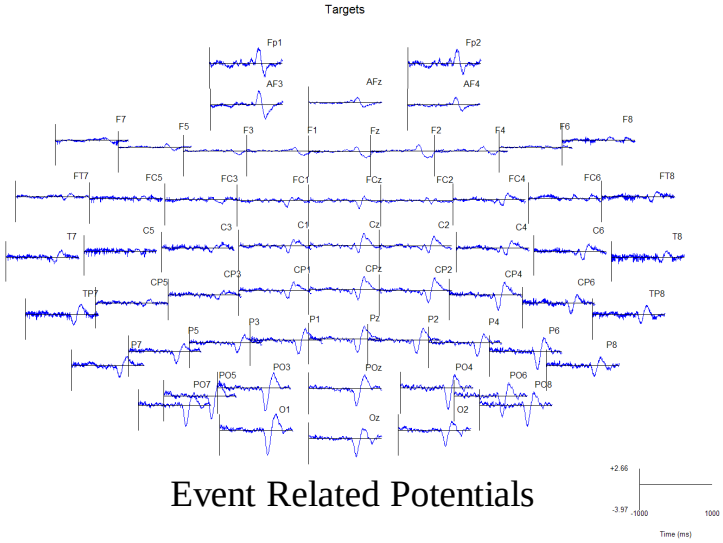
From D. Linden, 2005

Single-trial EEG Analysis

- Typically EEG is averaged over trials to increase the amplitude of the signal correlated with cortical processes relative to artifacts (very low SNR)
- High-density EEG systems were designed without a principled approach to handling the volume of information provided by simultaneously sampling from large electrode arrays.
- Our solution: identifying neural correlates with individual stimuli via single trial EEG analysis.
- We apply principled methods to find optimal ways for combining information over electrodes and moments in time contained in individual trials



Single-trial EEG

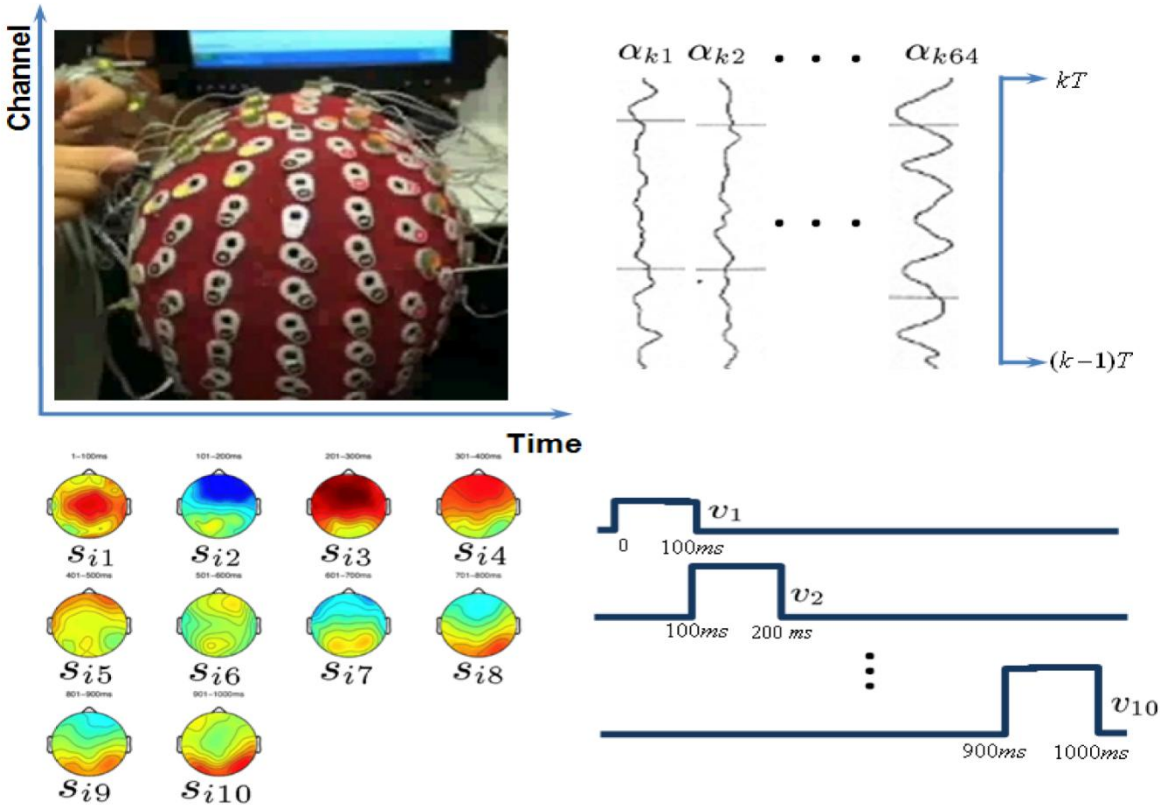


Event Related Potentials

Identifying Discriminative Components in the EEG Using Single-Trial Analysis

LDA or Logistic Regression is used to learn the contributions of EEG signal components at different spatial-temporal locations

(Parra, Sajda et al. 2002, 2003)

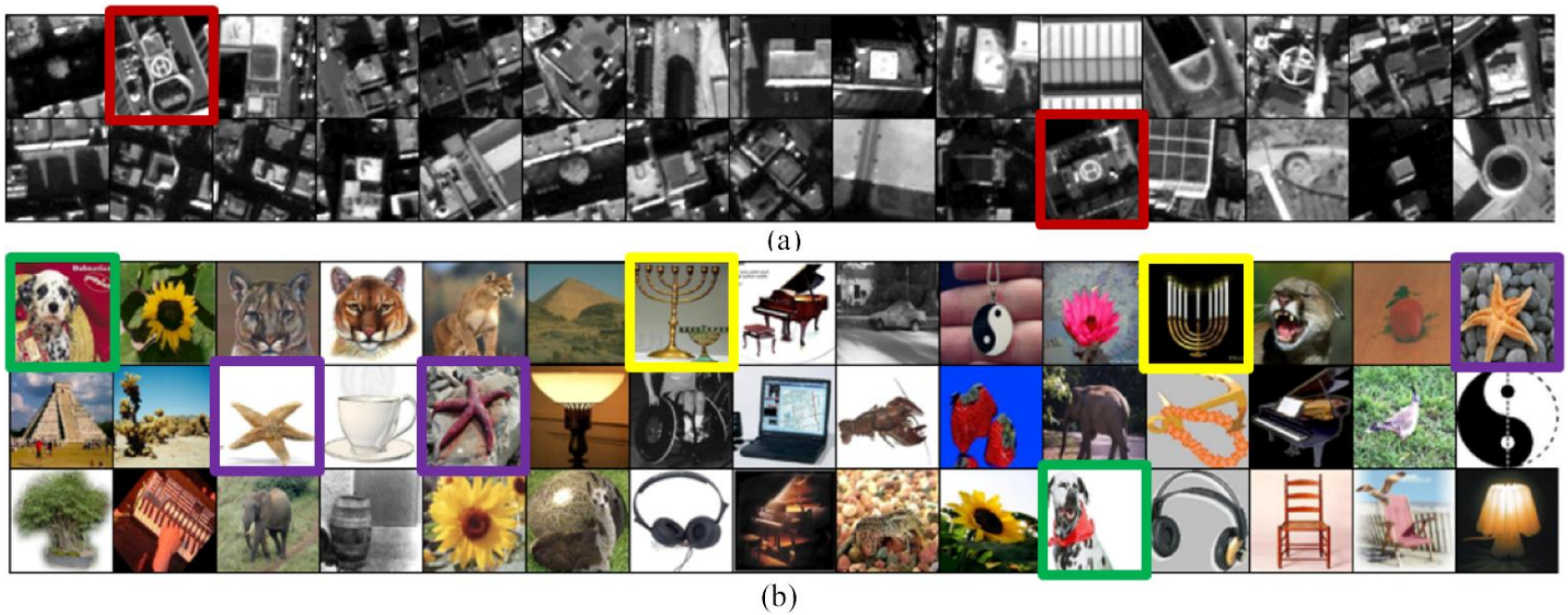


Optimal spatial filtering across electrodes within each short window (e.g., 100ms)

Optimal temporal filtering over time windows after onset

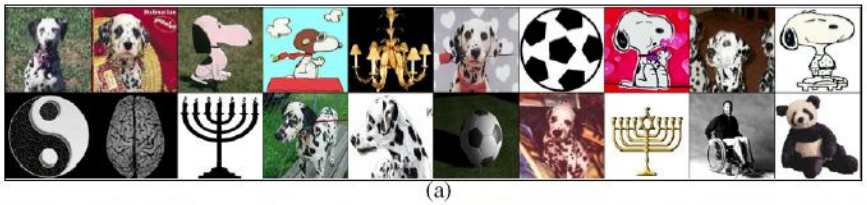
Experiments

- CalTech101: 3798 images from 62 categories
  Satellite images
- Generic neural decoder trained per user using images *(Soccer Ball* or *Baseball Gloves)* from Caltech256
- A subset images randomly sampled to construct 6-Hz RSVP sequence
- Initial Trials: 4 subjects, 3 targets (*Dalmatian*, *Chandelier/Menorah, & Starfish)*

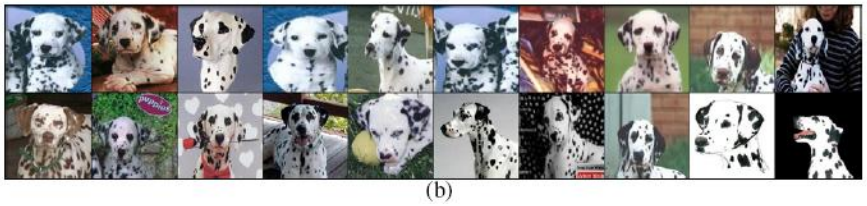


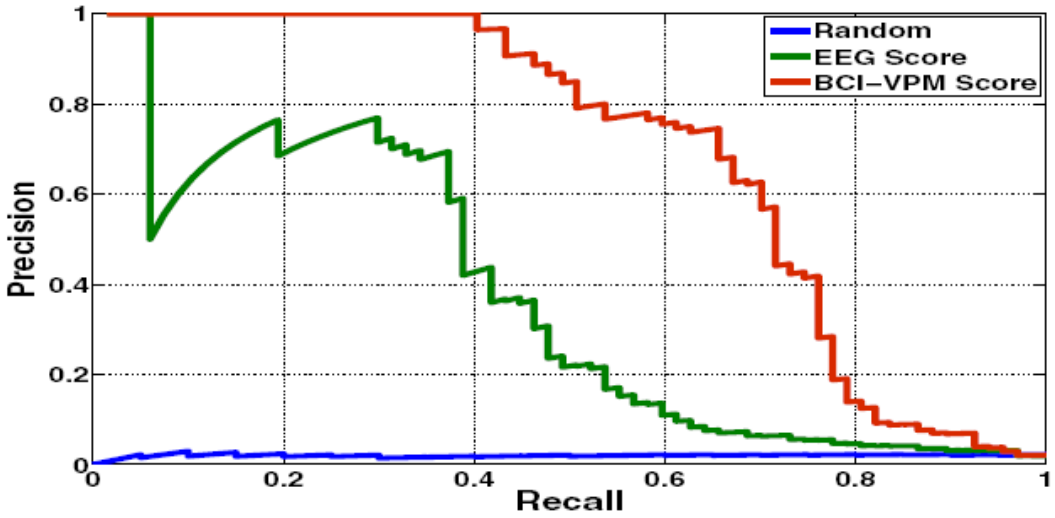
Example results

Top 20 results of
Neural EEG detection



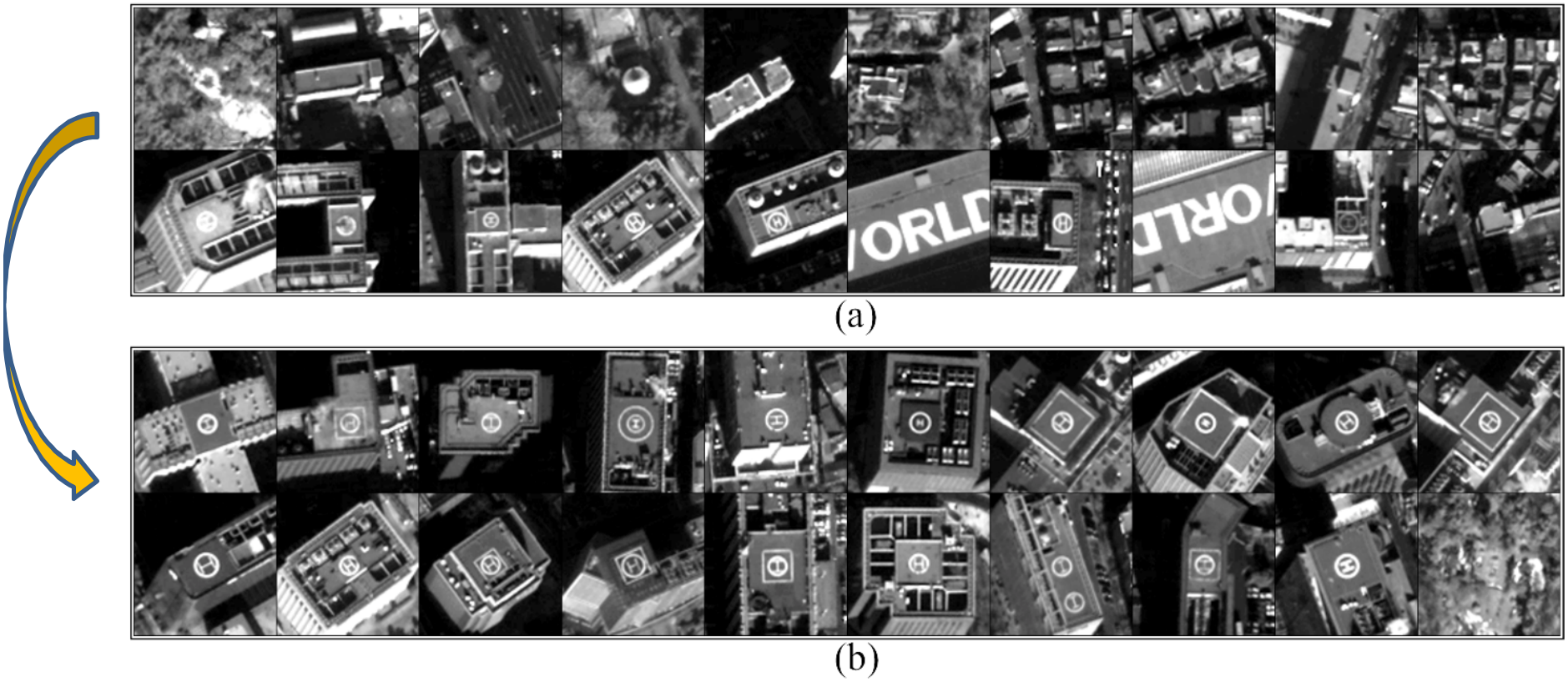
Top 20 results of
Hybrid System (BCI-VPM)





Application in target searching in satellite images:

Initial EEG neural signal detection:

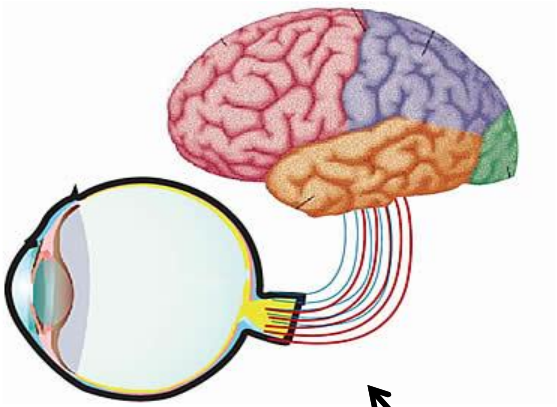


After graph refinement and diffusion

Images from DigiGlobe)

Summary: Cross Fertilization of Several Fields

Human Vision



Internet/Data Vision

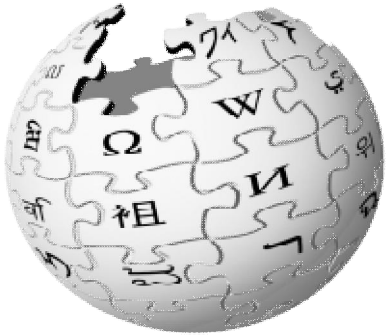


Image Recognition/
Machine Learning

Image video
search engines

Conclusions

- Great opportunity for video search research
- Exciting topics
  - Semantic Search:
    Large-scale visual ontology and intuitive search
  - Machine Learning and Computer Vision:
    Robust classification and image understanding
  - Matching of Billions of Images or More
    Robust features and fast matching
  - Internet Vision:
    Explore new applications on Internet
  - Neuro-Computer Vision:
    Synergistic integration with neural vision systems

CuZero
Zero-Latency Video Search
http://www.ee.columbia.edu/cuzero

Acknowledgments

- Columbia University
  - Eric Zavesky, Yu-Gang Jiang, Jun Wang, Junfeng He, Wei Liu, Wei Jiang, Akira Yanagawa
  - Paul Sajda, Barbara Hanna, Eric Pohlmeyer
- Yahoo! Research
  - Lyndon Kennedy
- City University, Hong Kong
  - Chong-Wah Ngo
- IRIT, France
  - Elie El Khoury

References

(many papers can be found at http://www.ee.columbia.edu/dvmm)

- E. Zavesky, S.-F. Chang, "**CuZero**: Embracing the Frontier of Interactive Visual Search for Informed Users," ACM Multimedia Information Retrieval, 2008.
- Y.-G. Jiang, A.Yanagawa, S.-F.Chang, C.-W.Ngo,
  "**CU-VIREO374**: Fusing Columbia374 and VIREO374 for Large Scale Semantic Concept Detection," ADVENT Technical Report #223-2008-1 Columbia University, August 2008.
  (http://www.ee.columbia.edu/ln/dvmm/CU-VIREO374/)
- J. Wang, E. Pohlmeyer, B. Hanna, Y.-G. Jiang, P. Sajda, and S.-F. Chang,
  "**Brain State Decoding** for Rapid Image Retrieval," ACM Multimedia Conference, 2009.
- Jun Wang, Yu-Gang Jiang, Shih-Fu Chang, "**Label Diagnosis** through Self Tuning for Web Image Search," CVPR 2009.
- Jun Wang, Sanjiv Kumar, Shih-Fu Chang,
  "**Semi-Supervised Hashing** for Scalable Image Retrieval", CVPR 2010.
- Junfeng He, Wei Liu, Shih-Fu Chang,
  "Scalable Similarity Search with **Optimized Kernel Hashing**," ACM SIGKDD, 2010.
- Wei Liu, Junfeng He, Shih-Fu Chang,
  "**Large Graph Construction** for Scalable Semi-Supervised Learning," ICML 2010.
- M. Naphade, J.R. Smith, J.Tesic, S.-F. Chang, W. Hsu, L. Kennedy, A. Hauptmann, J. Curtis,
  "**Large-Scale Concept Ontology for Multimedia**," *IEEE MultiMedia* magazine, 2006.
- Lyndon Kennedy, Shih-Fu Chang, "Internet image archaeology: automatically tracing the manipulation history of photographs on the web," ACM Multimedia, 2008.