COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

Visual Search

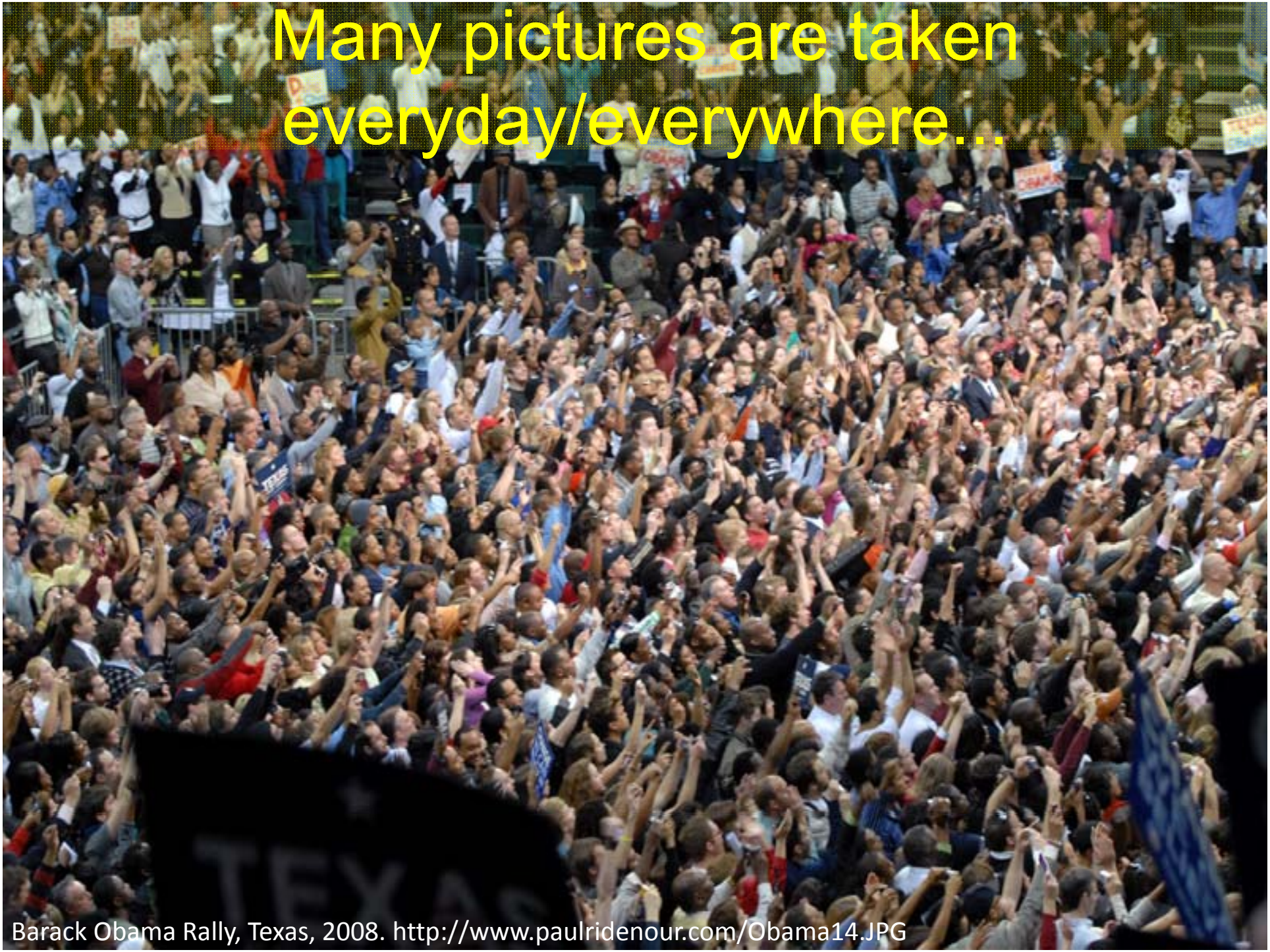
Shih-Fu Chang

Department of Electrical Engineering

Columbia University

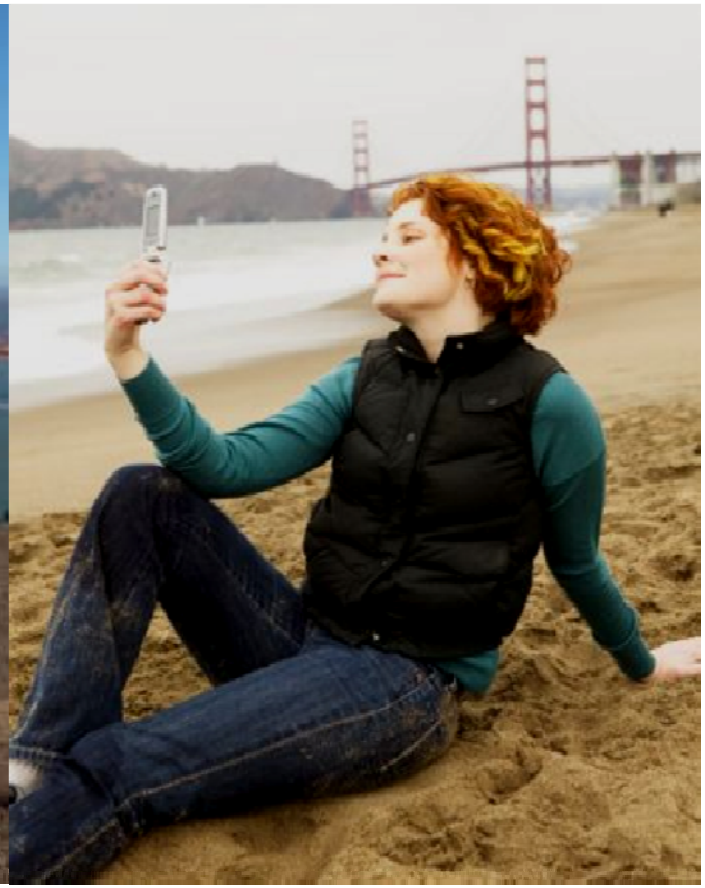
Boston University ECE Lecture, September 2010

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

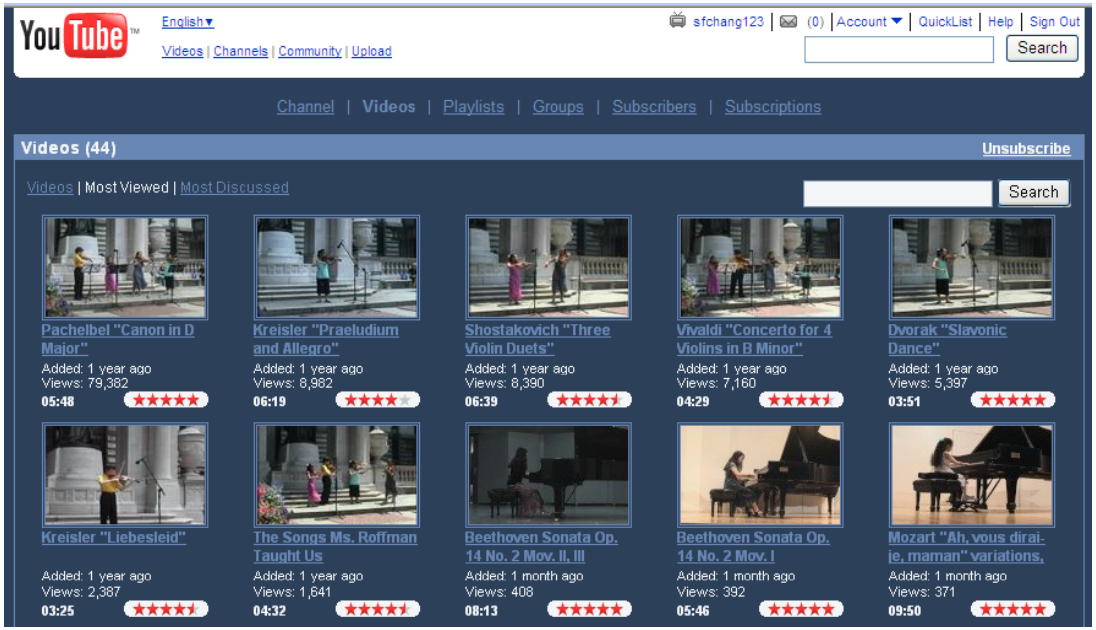
digital-photography-school.com

But we are not good at organizing them …

- Most photos and videos remain in shoebox or e-shoebox
- 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 after capture and upload

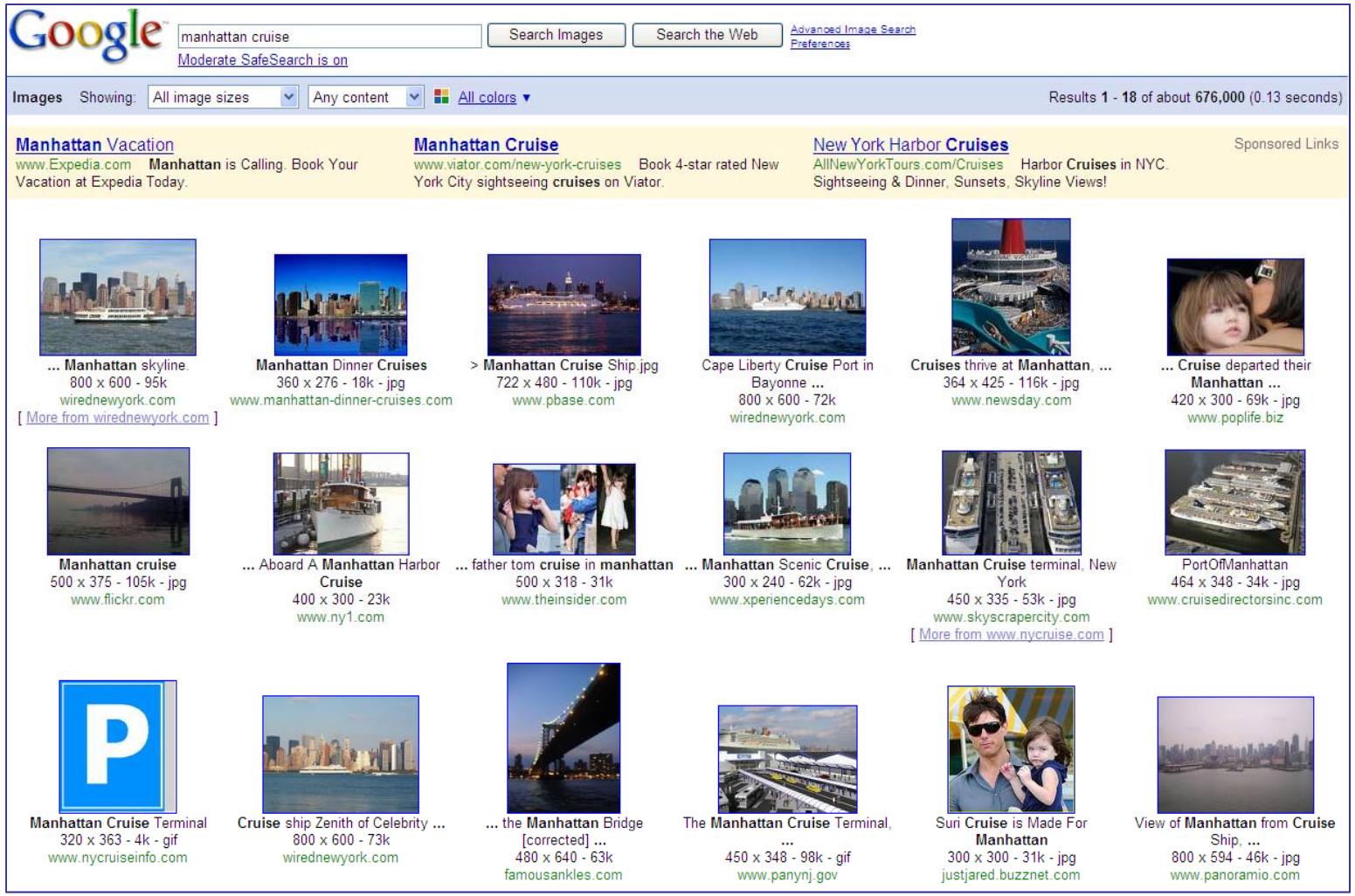
It will be nice to tag …

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; . . .

- **Tagging is boring and hard**
- **Each uploaded photo has only 0.97 tag on average**
  (Naaman, Yahoo!/Rutgers)

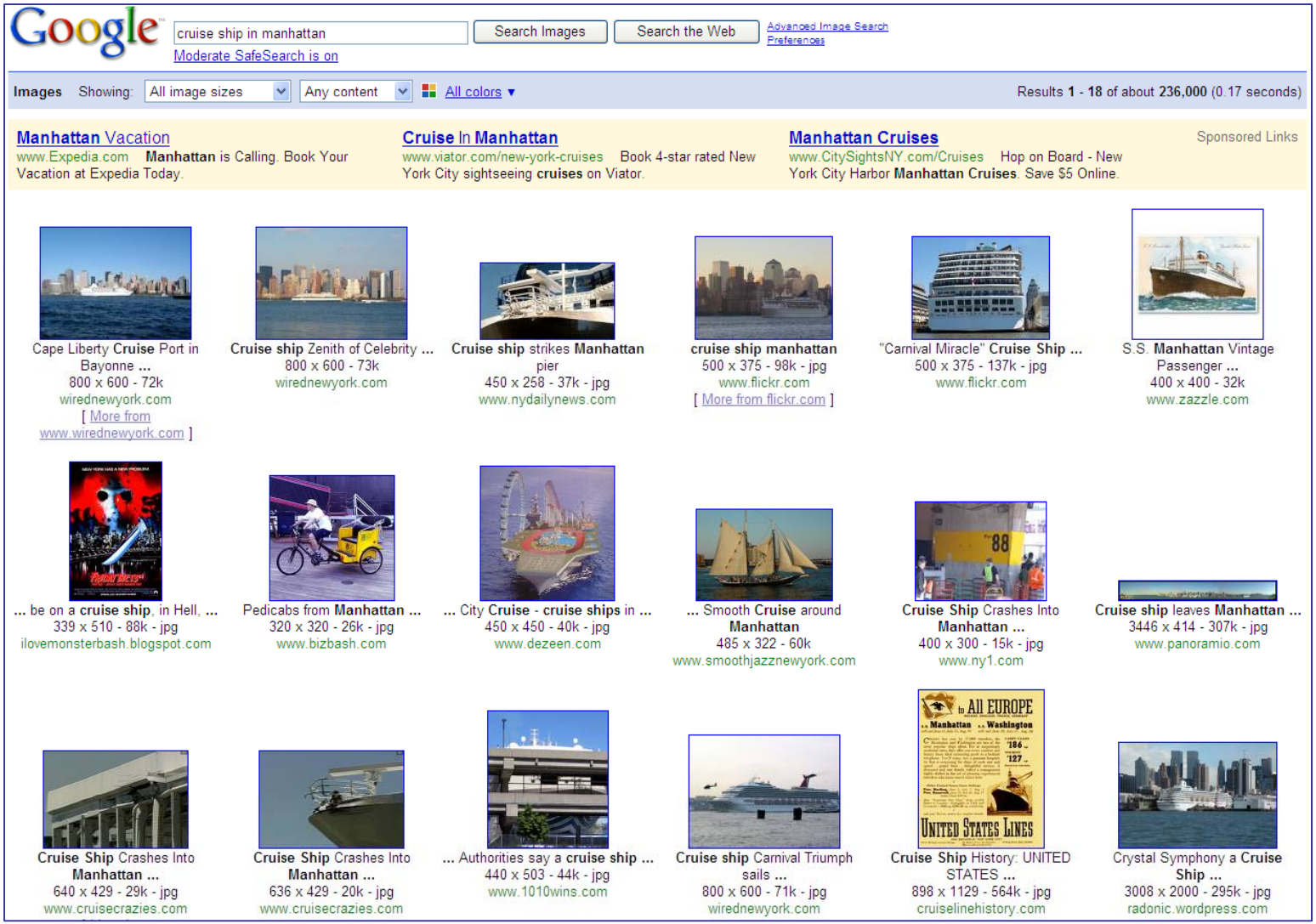
Scarce Tags → Faulty Search Engines

"Manhattan Cruise"



Scarce Tags → Faulty Search Engines

"Cruise ship in Manhattan"



The other side of coin

Visual Search: use pixels to find information

product search

landmark search

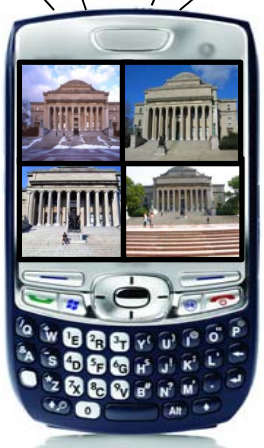
document search

Recent commercial applications:

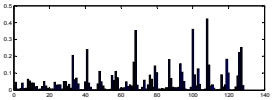
iPhone App SnapTel

Google Goggles

How Does It Work?



1. Take a picture

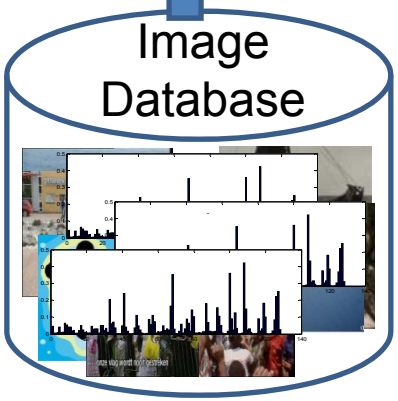


2. Image feature extraction

3. Send to server via MMS

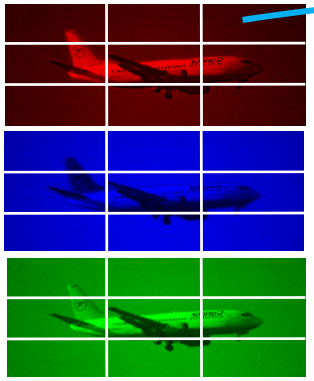
4. Feature matching with database images

5. Send most similar images back



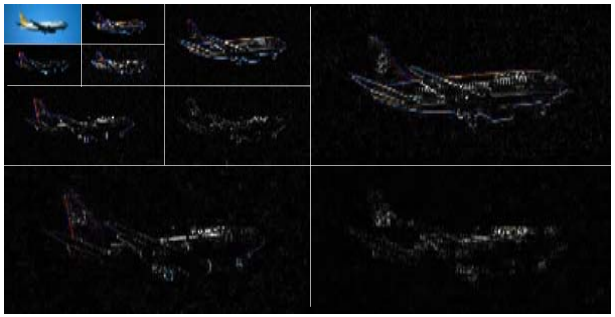
Basic Image Features

Grid Color Moment



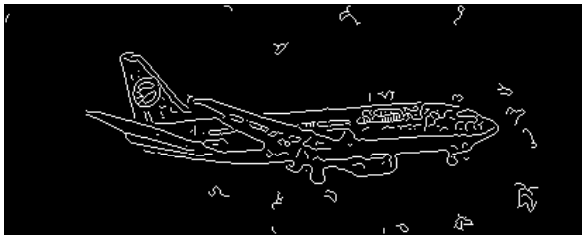
mean
variance
...

Wavelet Texture

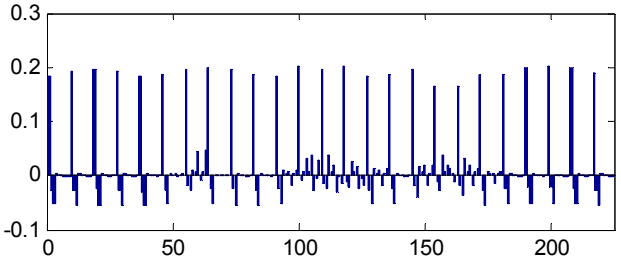


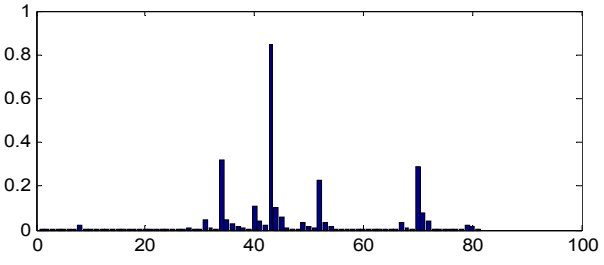
energy in different filter banks

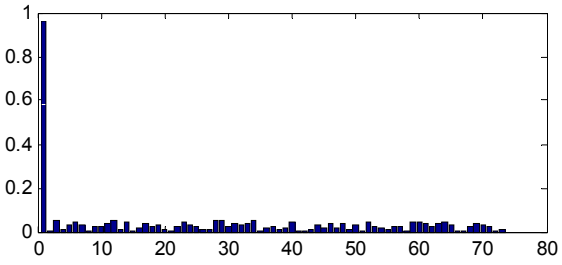
Canny Edge



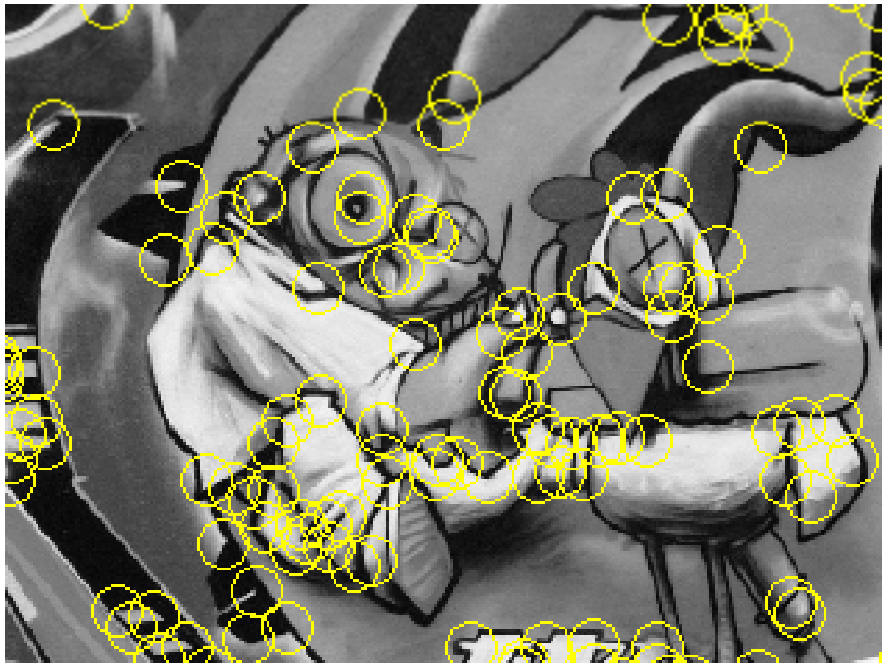
edge direction counts







Recent Popular Features: Sample keypoints in images

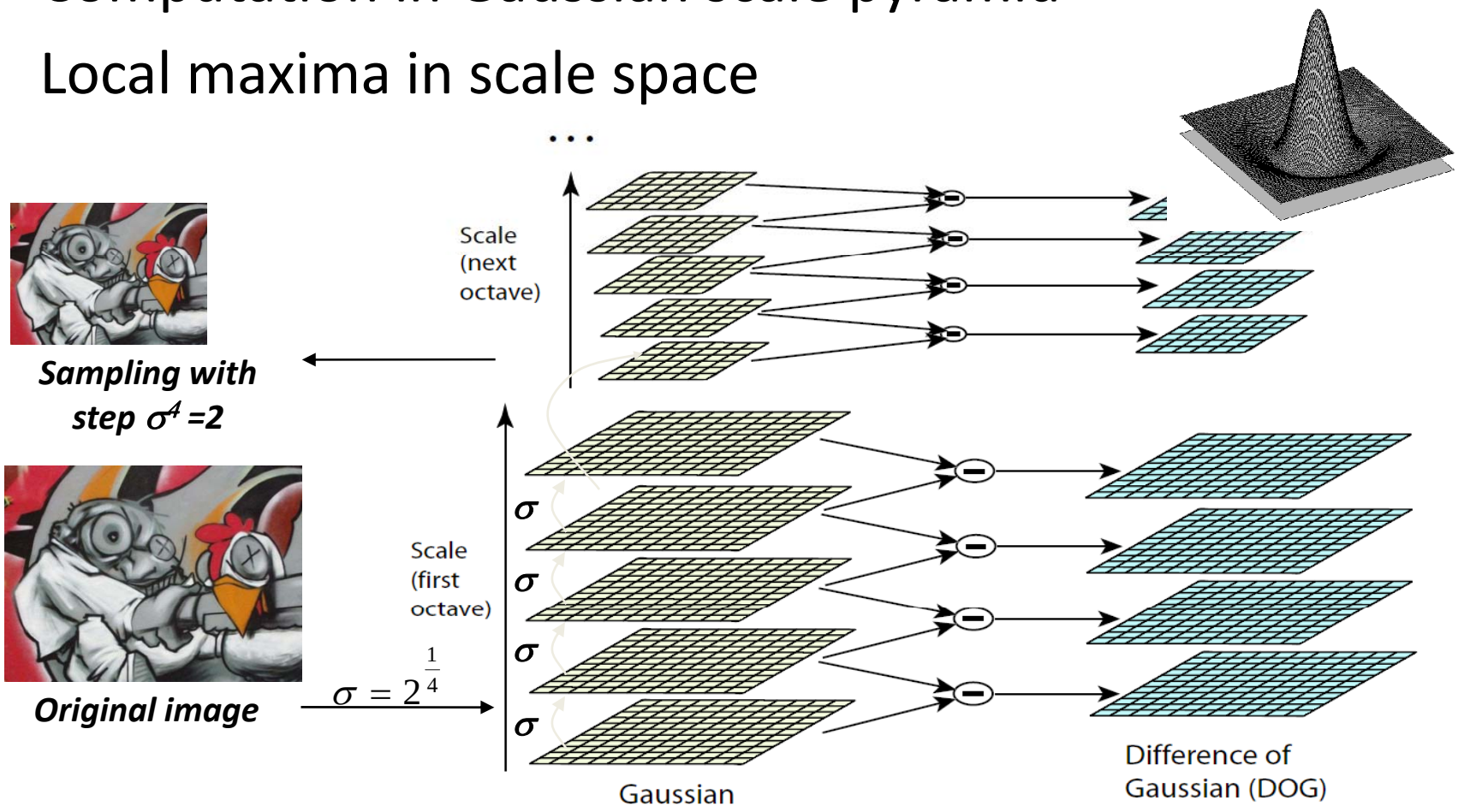


- Keypoint properties:
  - Interesting content
  - Precise localization
  - Repeatable detection under variations of scale, rotation, etc

(Slide of K. Grauman)

Example: Scale-Space

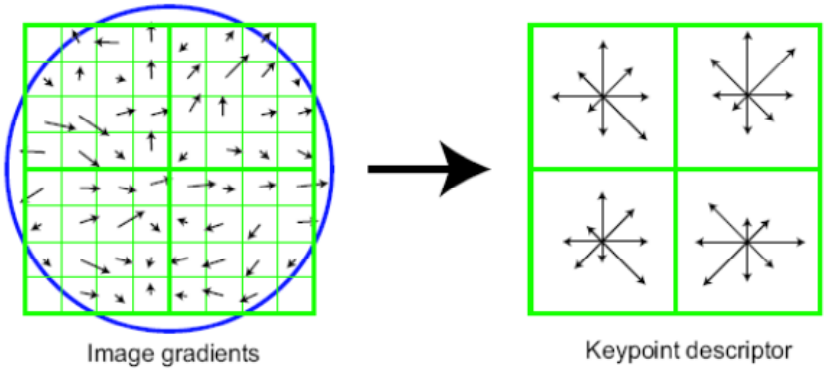
- Computation in Gaussian scale pyramid
- Local maxima in scale space



(Slide of K. Grauman)

Extract Appearance Descriptor from Keypoints

Compute gradient
in a local patch



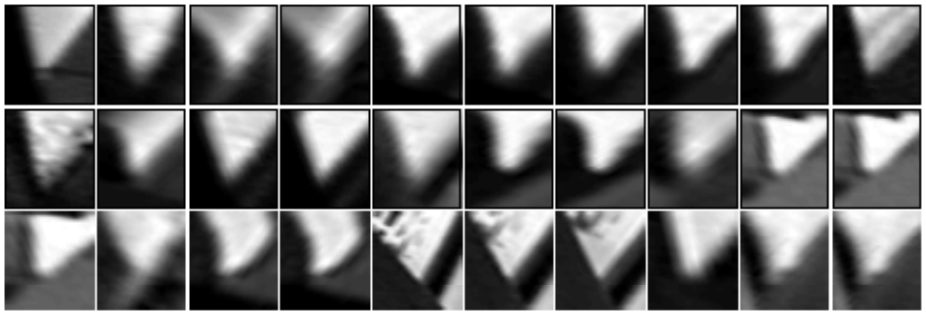
SIFT: Histogram of oriented gradients over local grids

- e.g., 4x4 grids and 8 directions
  -> 4x4x8=128 dimensions

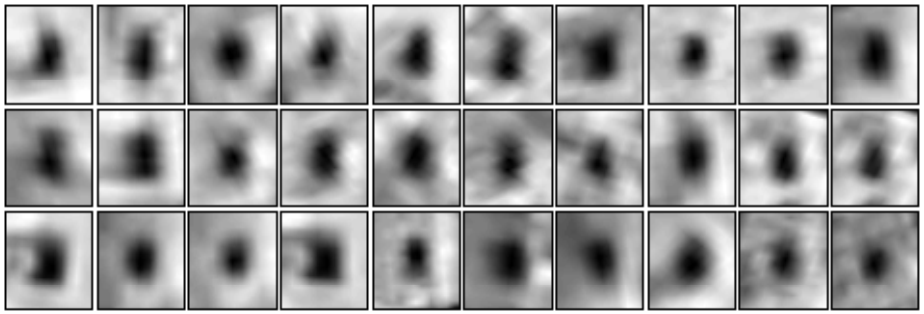
[Lowe, ICCV 1999]

- Scale invariant

Clustering of Image Patch Patterns

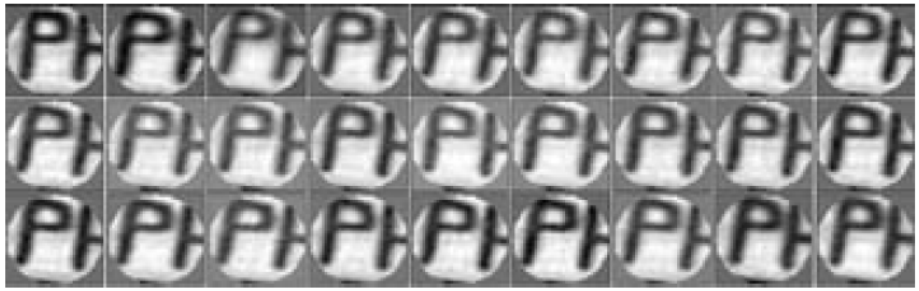


Corners



Blobs

eyes



letters

Sivic and Zisserman, “Video Google”, 2006

A New Representation: Visual Words

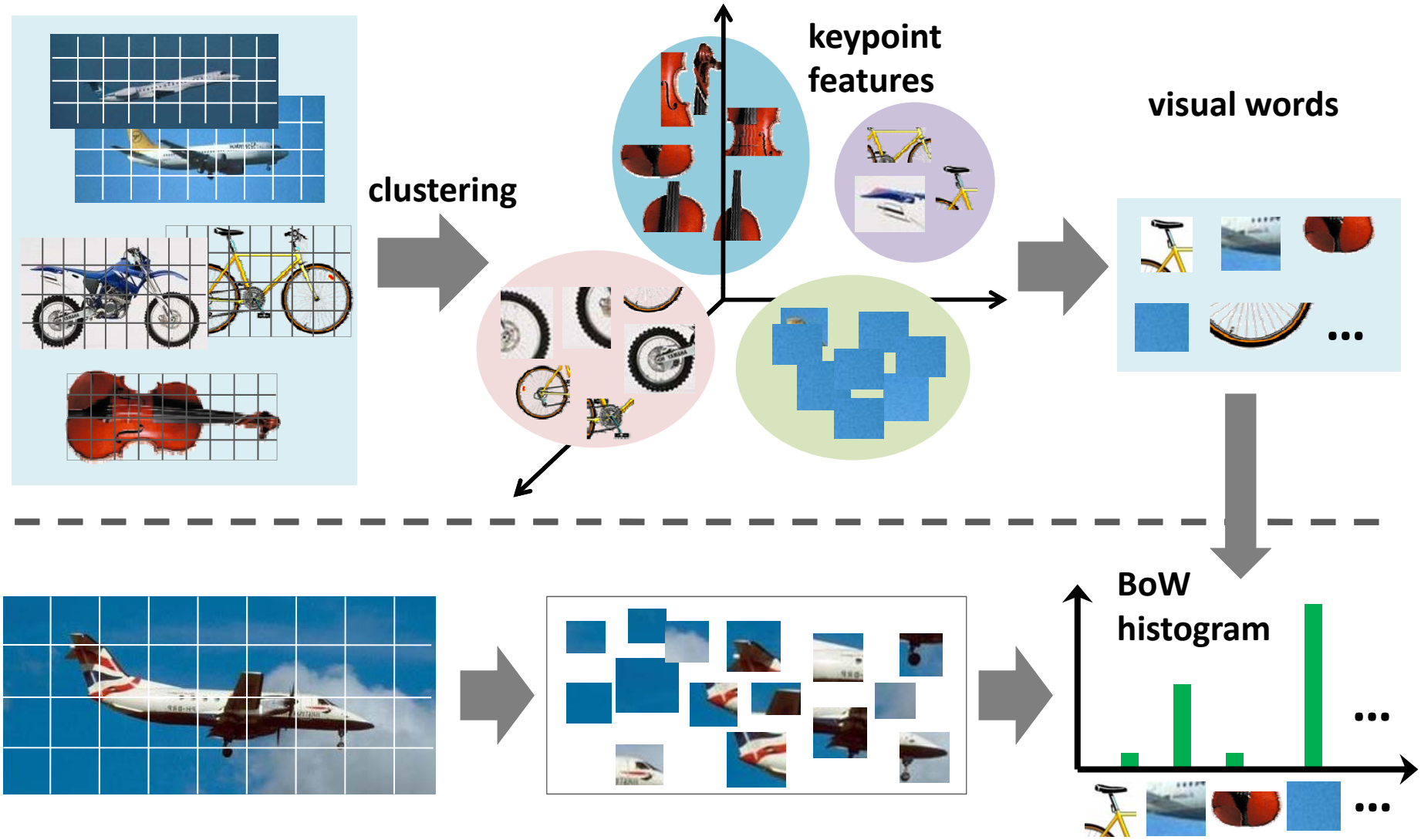
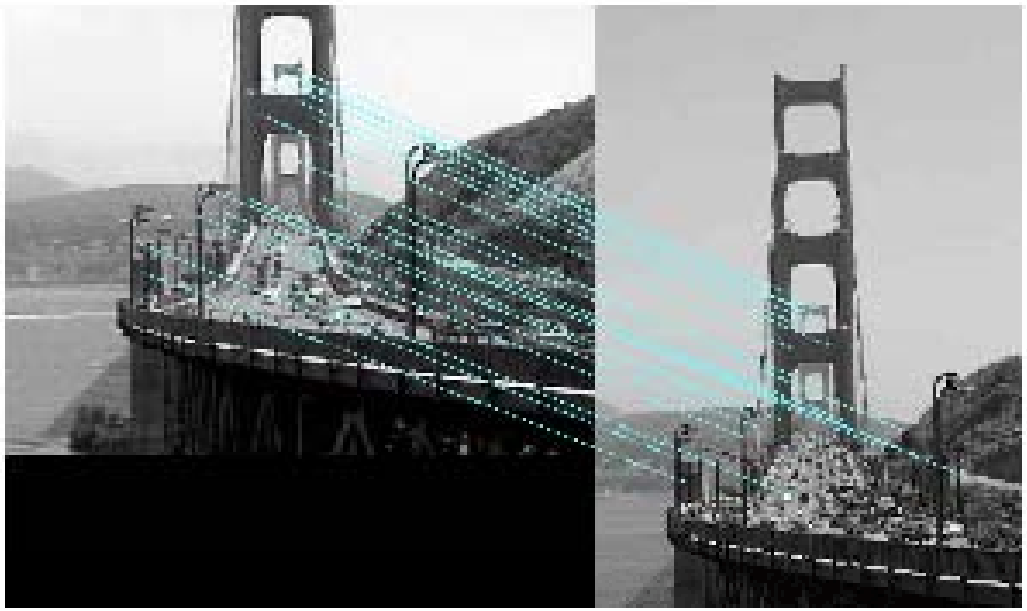


Image matching with local interest points



Matching SIFT points
[Lowe, 1999]

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

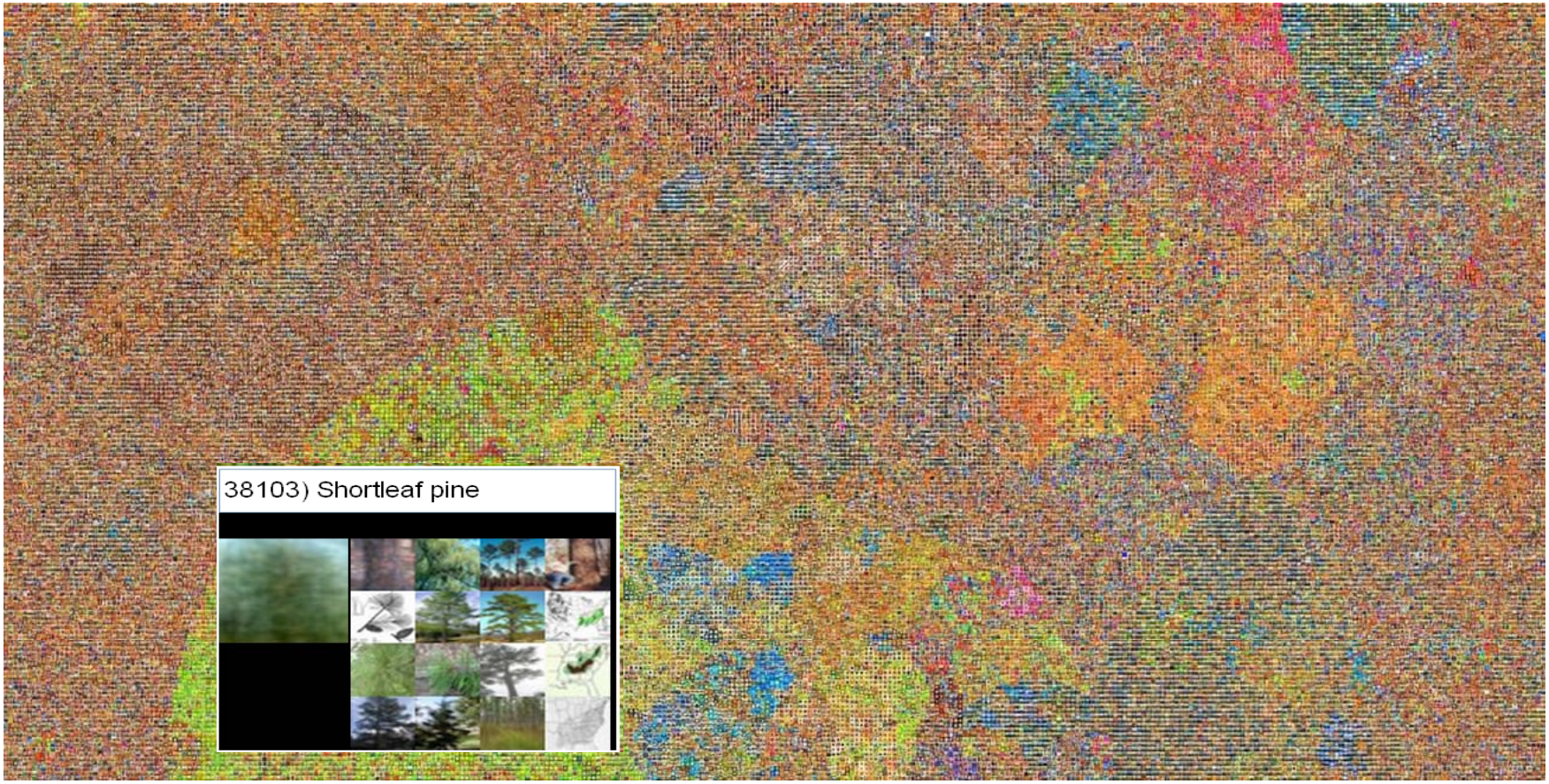
Content Based Image Search

- Demo: Object Retrieval
- Demo 2: Flickr Image Search

Scale Up: Find similar images over Internet

- Billions of images online as dense sampling of the world
- For every image taken, likely to find images that look alike

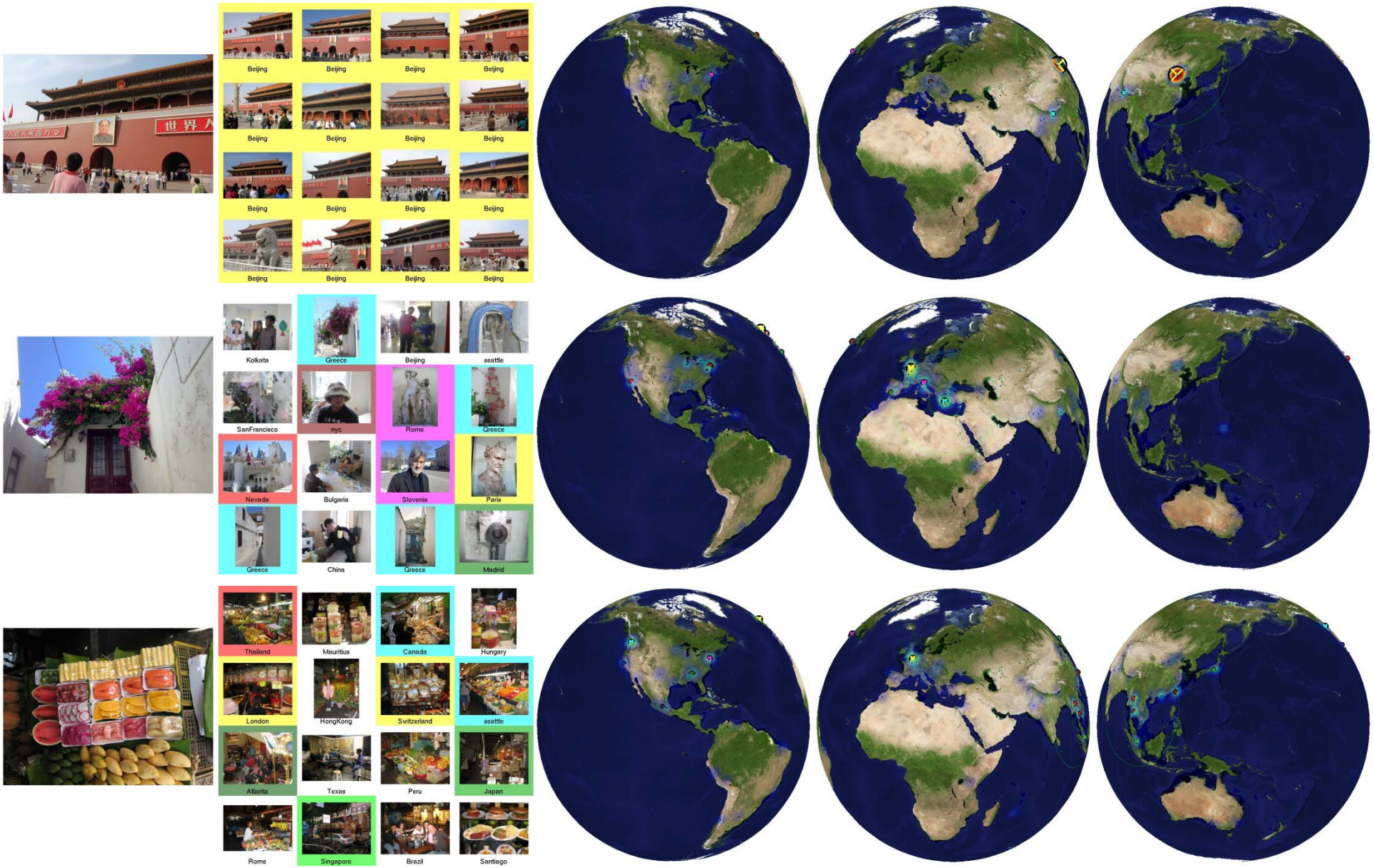
80 Million Tiny Images, Torralba, Fergus & Freeman, PAMI 2008



IM2GPS: where is this photo taken? (Hays & Efros, 2008)

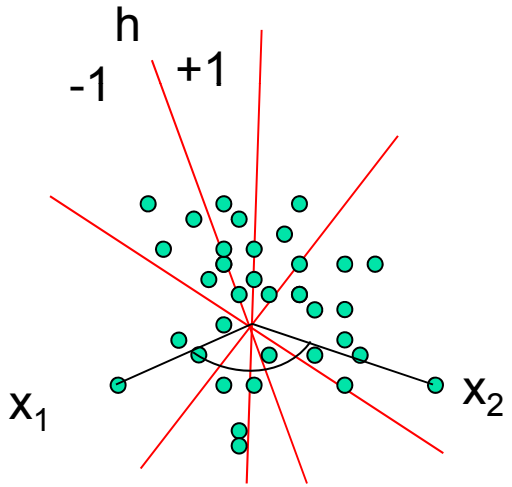
Similar images

Most likely locations



Search over Billions: Scalability is a Big Issue

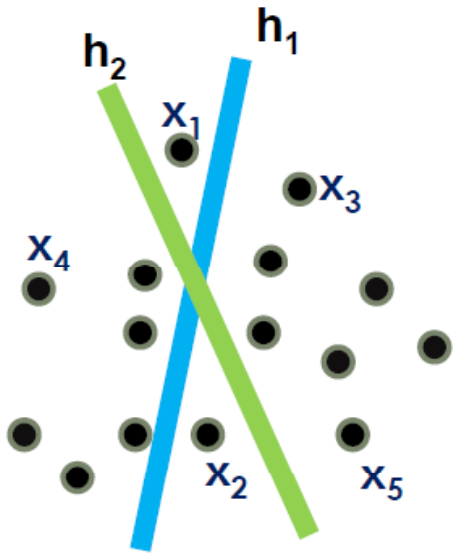
- Similarity Search: traditional tree-based methods (e.g., kd-tree) not suitable in high dimension
- Need accurate, sublinear solutions (o(N), O(log(N)), O(1))
- Recent trends:
  projection based hashing
  - Random projection:
    Locality Sensitive Hash (LSH)
    [Indyk & Motwani 98, Charikar 02]
  - Principal projection:
    Spectral Hashing [Weiss et al 08]
  - Restricted boltzman machines
    [Hinton et al. 06, Torralba et al. 08]
  - Kernel LSH
    [Kulis et al. 09 & Mu et al. 10]



$P(h(x_1) = h(x_2)) = 1 - \cos^{-1}(x_1 \cdot x_2)/\pi$
$= Sim(x_1, x_2)$
random projection h with *N(0,1)*

Binary Codes

Linear projection (hyperplane) based partitioning



X	x_1	x_2	x_3	x_4	x_5
y_1	0	1	1	0	1
y_2	1	0	1	0	1
…	…	…	…	…	…
y_k	…	…	…	…	…
	010…	100…	111…	001…	110…

Linear Projection based hashing

$$h_k(\mathbf{x}) = \mathrm{sgn}(f(\mathbf{w}_k^\top \mathbf{x} + b_k))$$
$$y_k(\mathbf{x}) = (1 + h_k(\mathbf{x}))/2$$

Very efficient training and testing

Probabilistic guarantee of finding true targets within ε distance range

[Indyk & Motwani 98]

Slide of Sanjiv Kumar, Google

Beyond Randomness: Semi-Supervised Learned Projection

- Given Pair-Wise Relations

$$(\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{M}\text{: neighbor pair}$$
$$(\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{C}\text{ nonneighbor-pair}$$

- Measure empirical fit of hash bits
- Are the partitions balanced? Measure hash bit variance

$$J(\mathbf{H}) = \sum_k \left\{ \sum_{(\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{M}} h_k(\mathbf{x}_i) h_k(\mathbf{x}_j) - \sum_{(\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{C}} h_k(\mathbf{x}_i) h_k(\mathbf{x}_j) \right\}$$

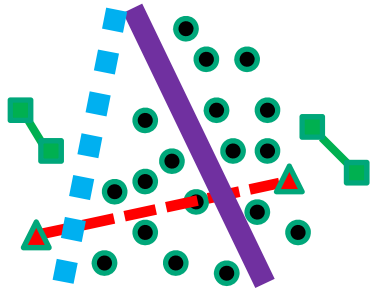
$$\sum_k E[\|h_k(\mathbf{x}) - \mu_k\|^2]$$

- Elegant eigen-decomposition solution

$$J(\mathbf{W}) = \frac{1}{2}\mathrm{tr}\left[\mathbf{W}^\top \mathbf{X}_l \mathbf{S} \mathbf{X}_l^\top \mathbf{W}\right] + \frac{\eta}{2} \sum_k E[\|\mathbf{w}_k^\top \mathbf{x}\|^2]$$

$$S_{ij} = \begin{cases} 1 & : \quad (\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{M} \\ -1 & : \quad (\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{C} \\ 0 & : \quad otherwise. \end{cases}$$

[Wang, Kumar, Chang CVPR10, ICML10]

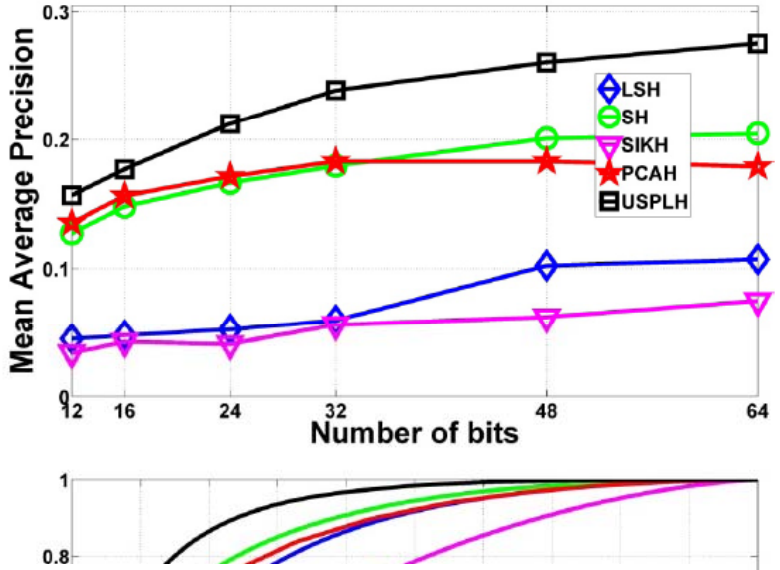


- Incremental learning via AdaBoosting

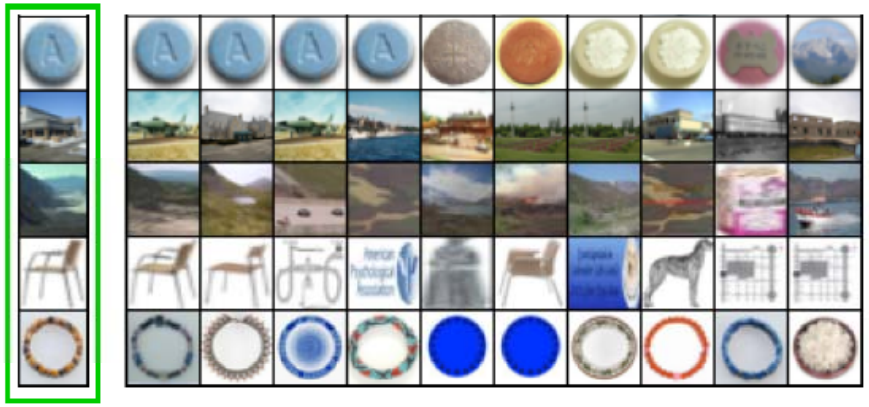
- Learned projection hash increases accuracy > 2X
- Query time: a few seconds
- Compact code - 48 bits vs. 128 bytes per sample)

- Challenge: scale to billions?

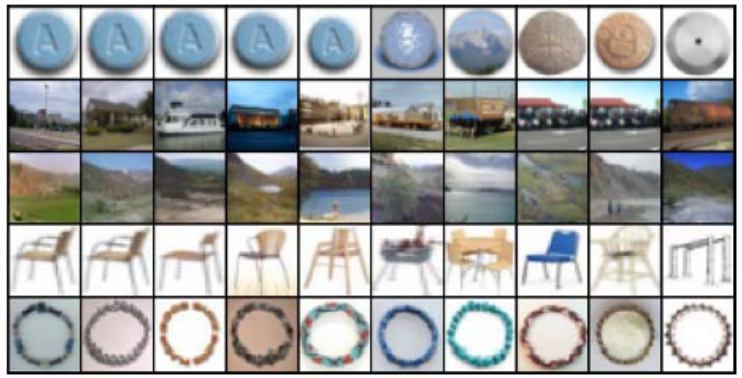
SIFT-1M



Tiny Image – 80M



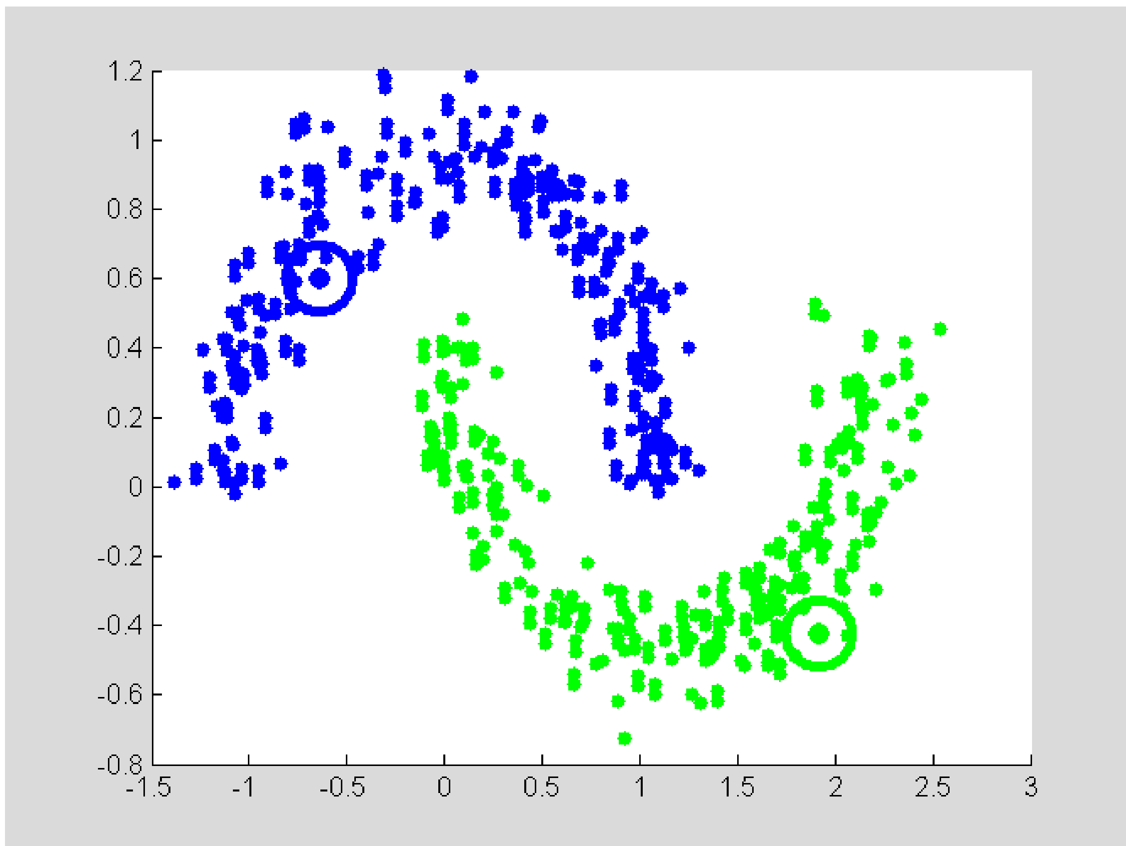
Binary Reconstructive Embedding



Learned Projection Hash

Beyond pair-wise similarity: Image Manifold

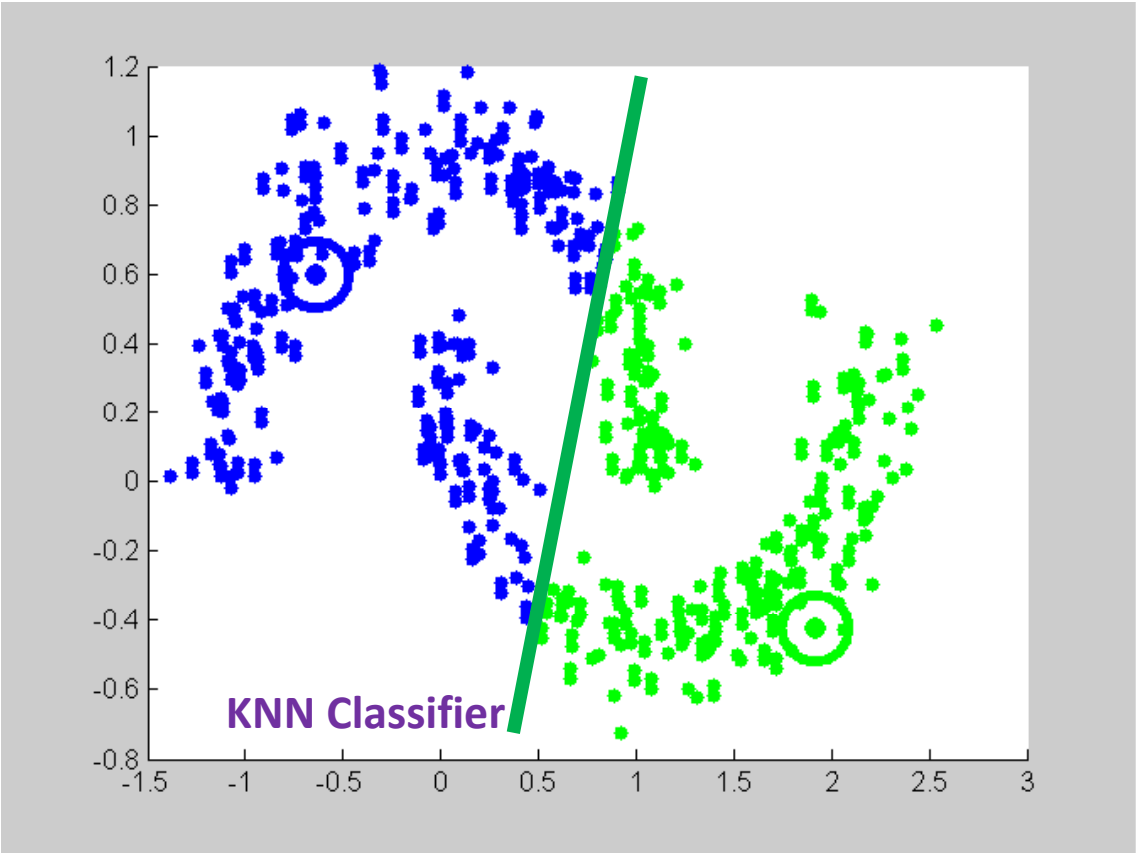
- Recover manifold distributions in high-dimensional space
- Each point represents an image, construct sparse graph
- Graph for information propagation, like PageRank



cons[illegible]d data

Manifold vs. Supervised Learning

- Poor classifier when ignoring manifold distribution

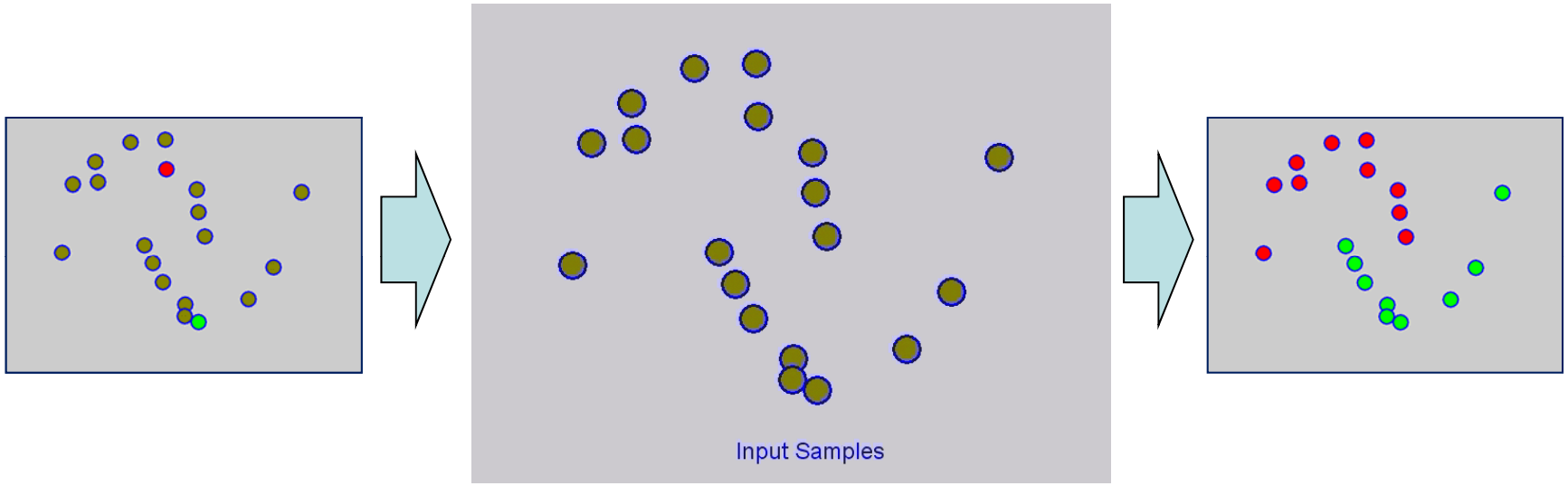


Only consider labeled data

train classifier

Graph-based Semi-Supervised Learning

- Construct Sparse for Label Propagation



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

An active area in Machine Learning

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

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*)

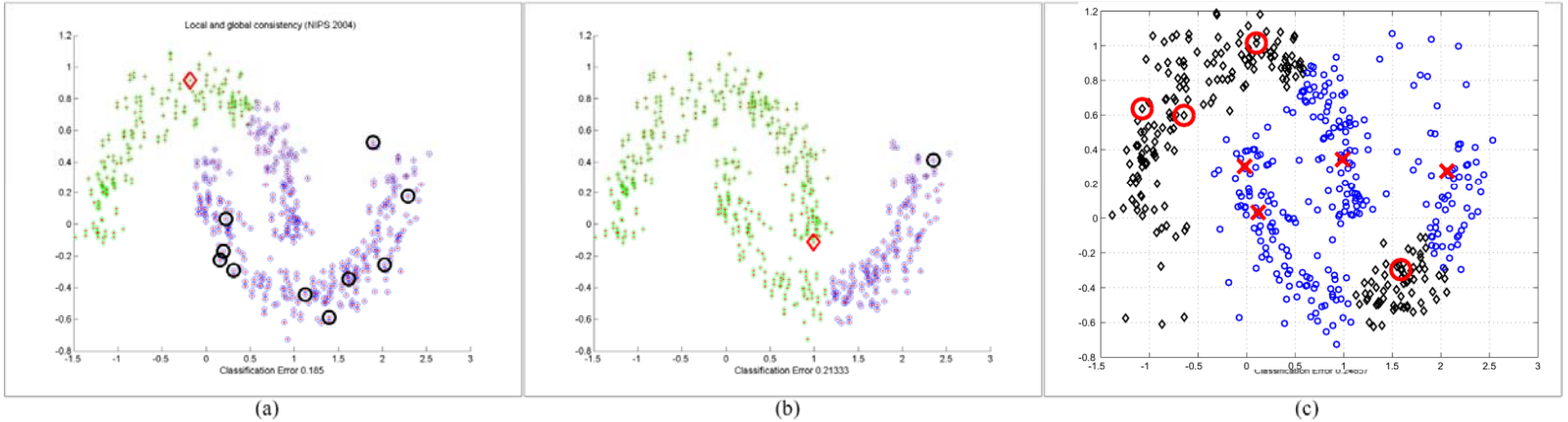
$$\mathcal{Q}(F) = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\|F_{i.}-F_{j.}\|^2$$

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

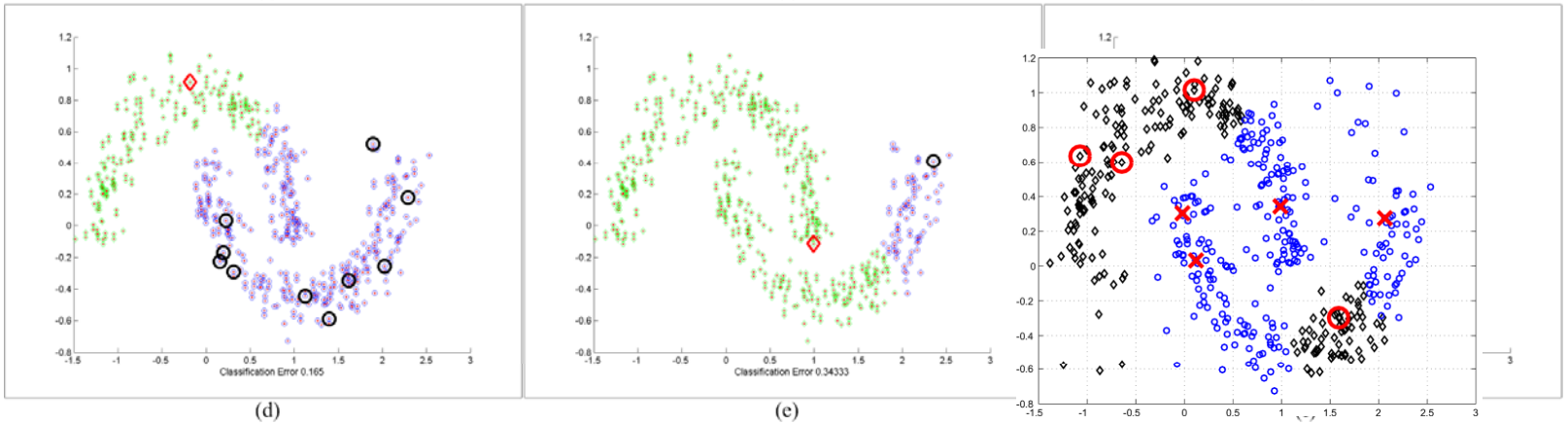
2) $F_{i.} = Y_{i.}$ on labeled data.

Non-Trivial Issues

LGC Method



GFHF Method

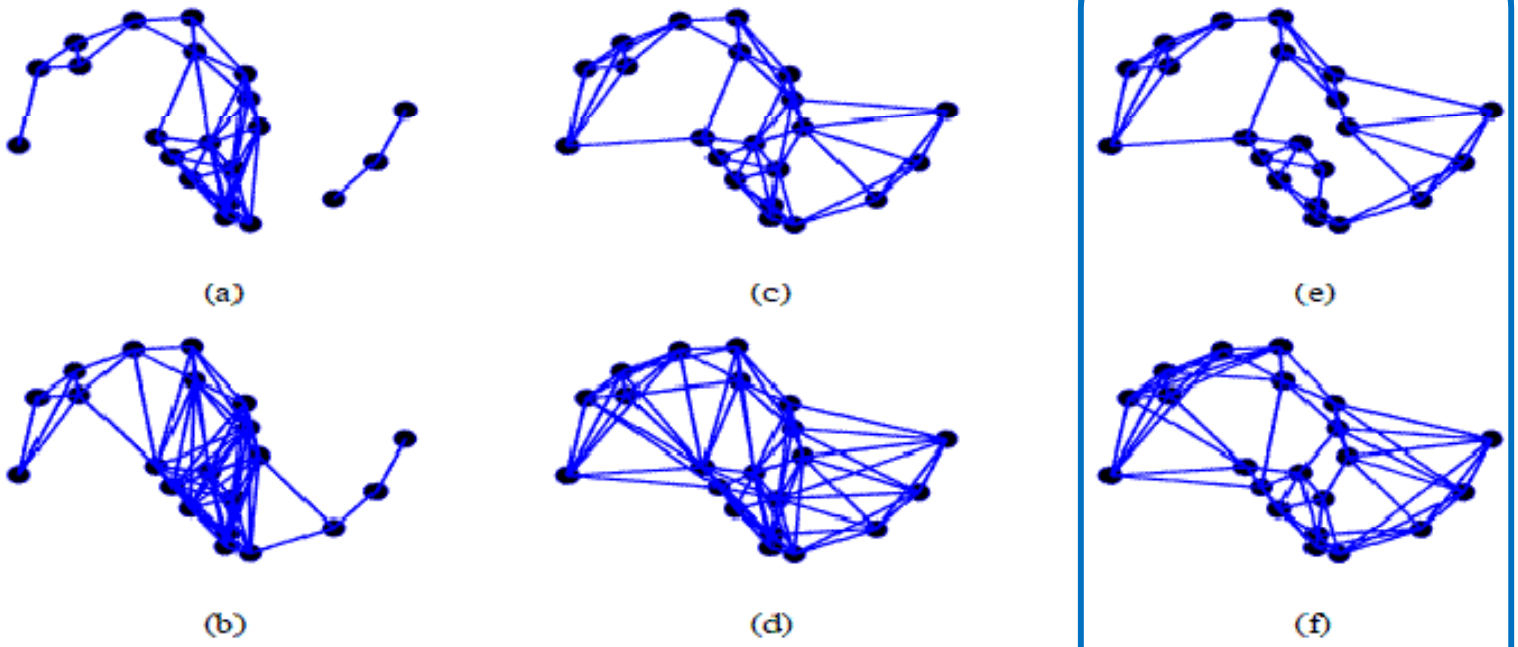


Unbalanced Labels

Bad Label Locations

Noisy Labels

A Key Problem: How to Construct a Good Sparse Graph from Massive Data?



Distance threshold | Rank threshold (kNN) | B-Match

- **B-Match preserves sparse structures**
- **Efficient method exists much faster than O(N^3)**

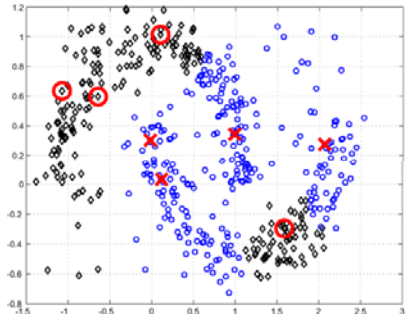
$$\min_{P \in \mathbb{B}} \sum_{ij} P_{ij} D_{ij}$$

$$s.t. \sum_{j} P_{ij} = b, P_{ii} = 0, P_{ij} = P_{ji},$$

P: connectivity
D: distance

(Jebara, Wang, Chang ICML 2009)

Another Key Issue: Noisy Labels



→ Treat both labels & prediction as optimization variables

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

- Diagnose and add/remove labels
  - Select the most reliable and informative label

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

- Propagate trusted labels to the whole graph
  - given label (Y), propagate over graph, predict F

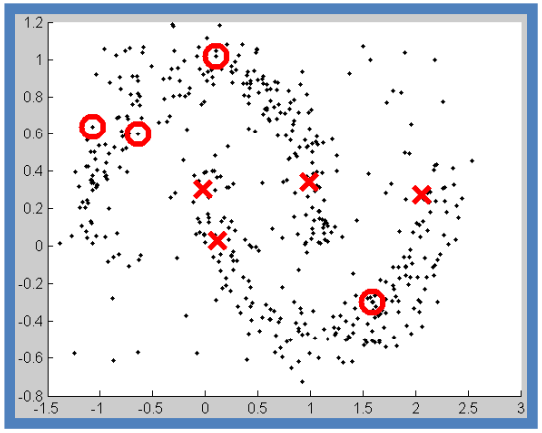
$$\frac{\partial \mathcal{Q}}{\partial \mathbf{F}^*} = 0 \Rightarrow \mathbf{F}^* = (\mathbf{L}/\mu + \mathbf{I})^{-1}\mathbf{V}\mathbf{Y} = \mathbf{P}\mathbf{V}\mathbf{Y}$$

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

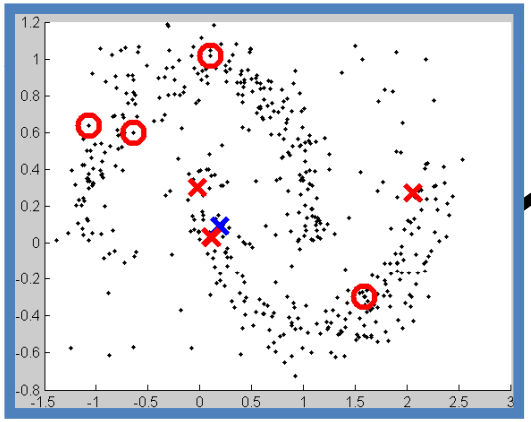
Iteratively Tune and Propagate Label Information

(Wang and Chang CVPR'09)

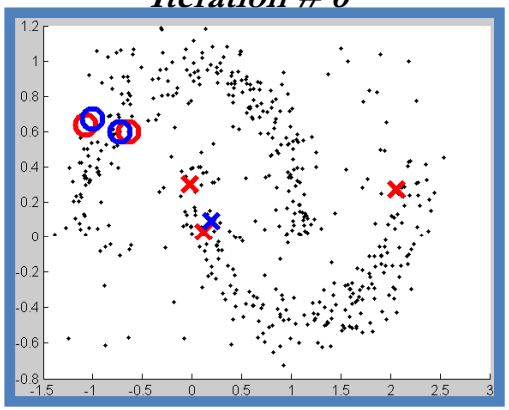
Initial Labels

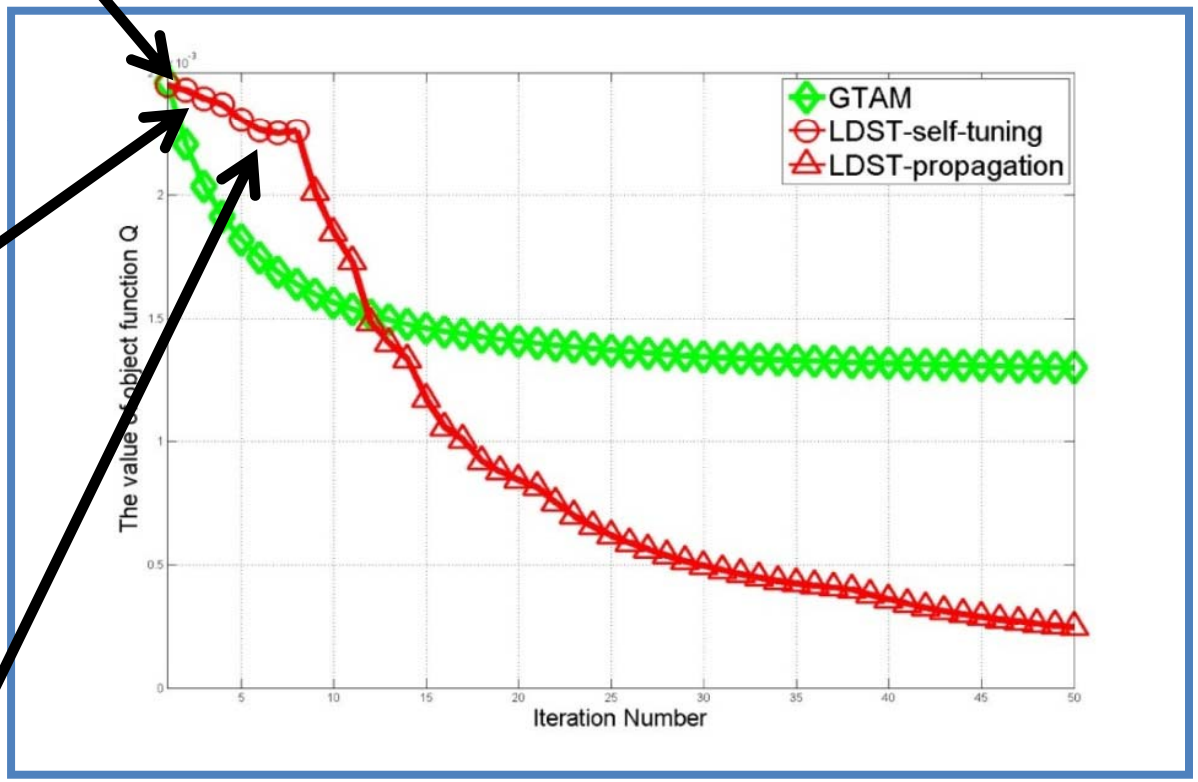


Iteration # 2



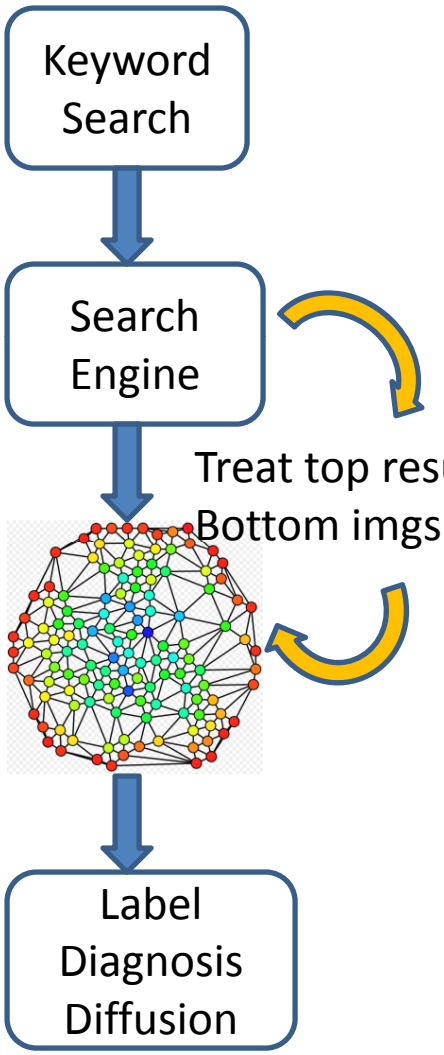
Iteration # 6





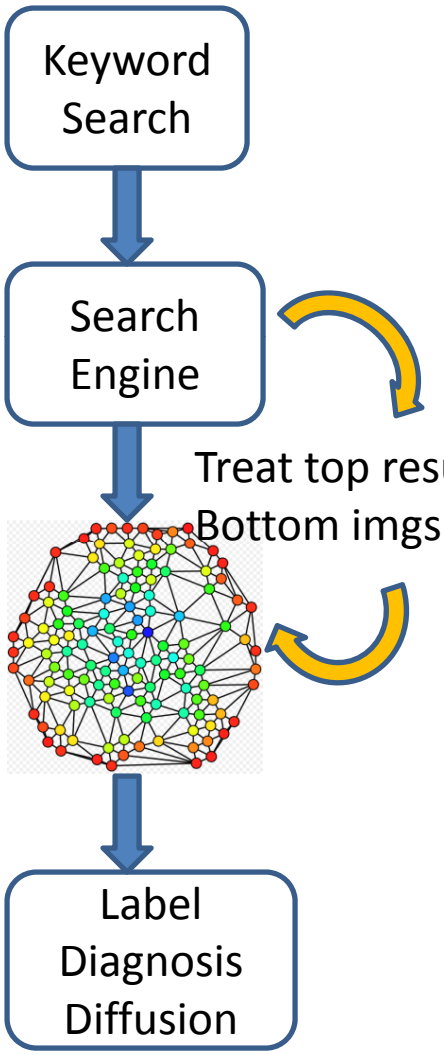
Decline of the cost function Q over iterations (with vs. without label tuning)

Application: Internet Search Result Reranking



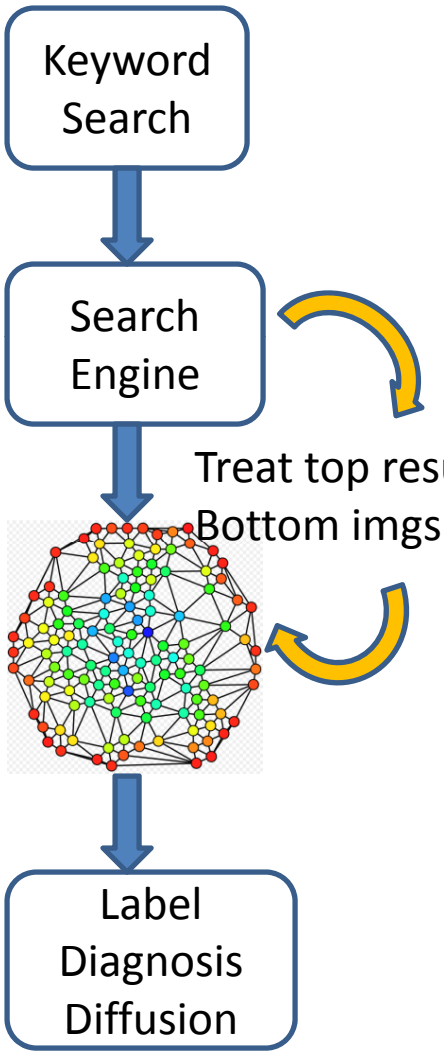
Google Search "Tiger"

Application: Internet Search Result Reranking

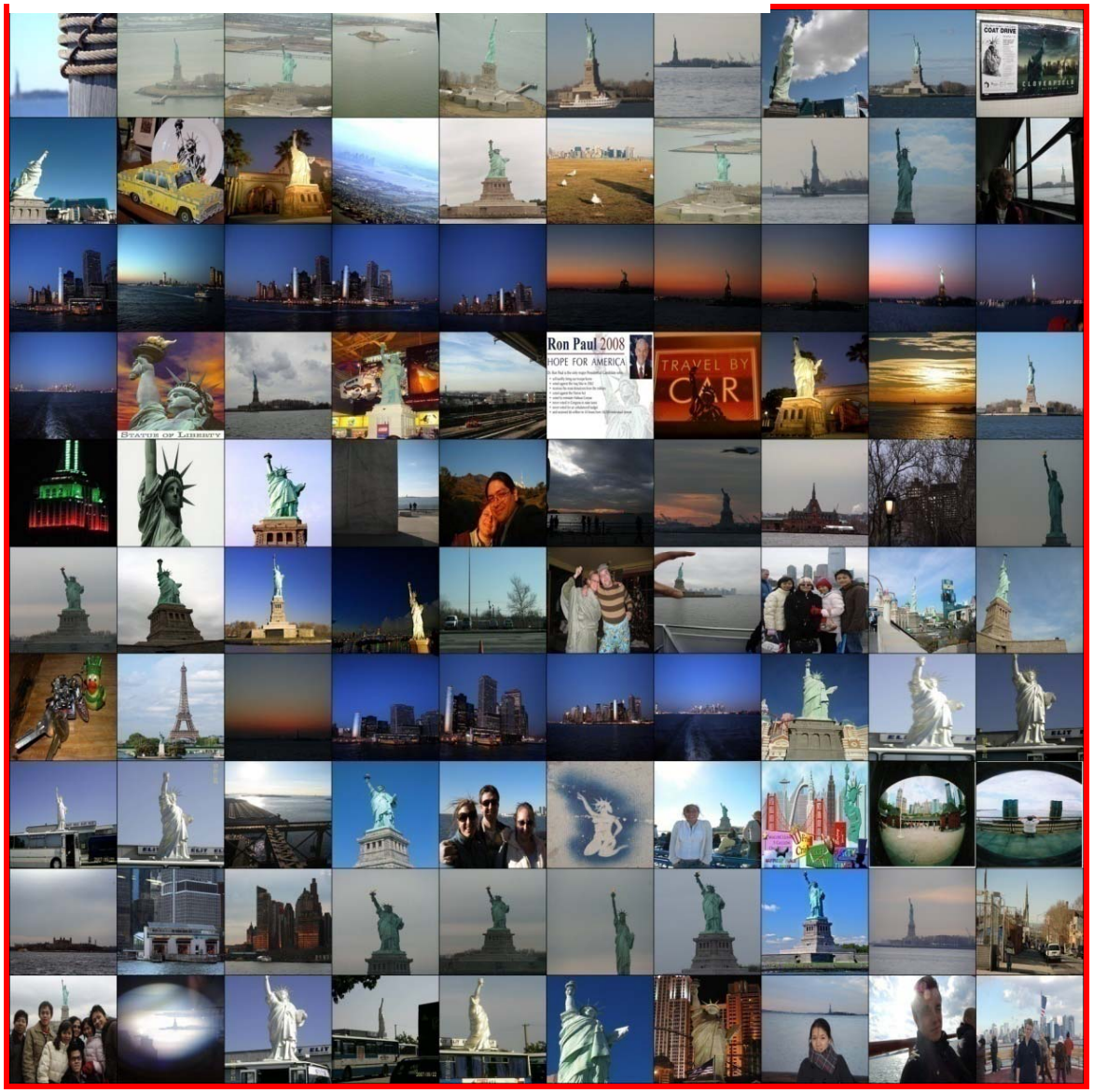


Reranked Results

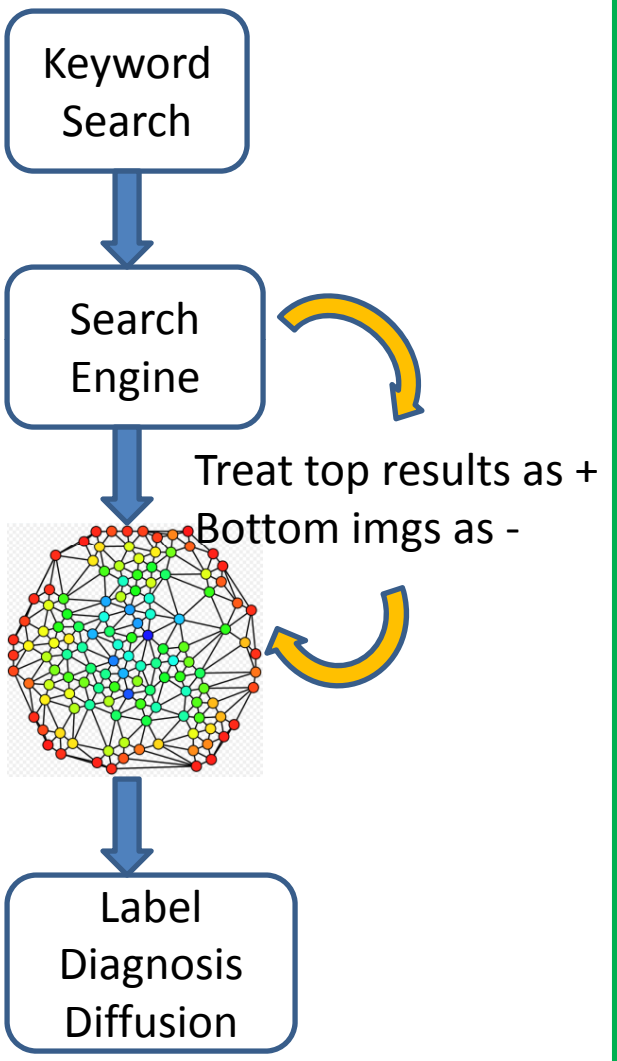
Application: Internet Search Result Reranking



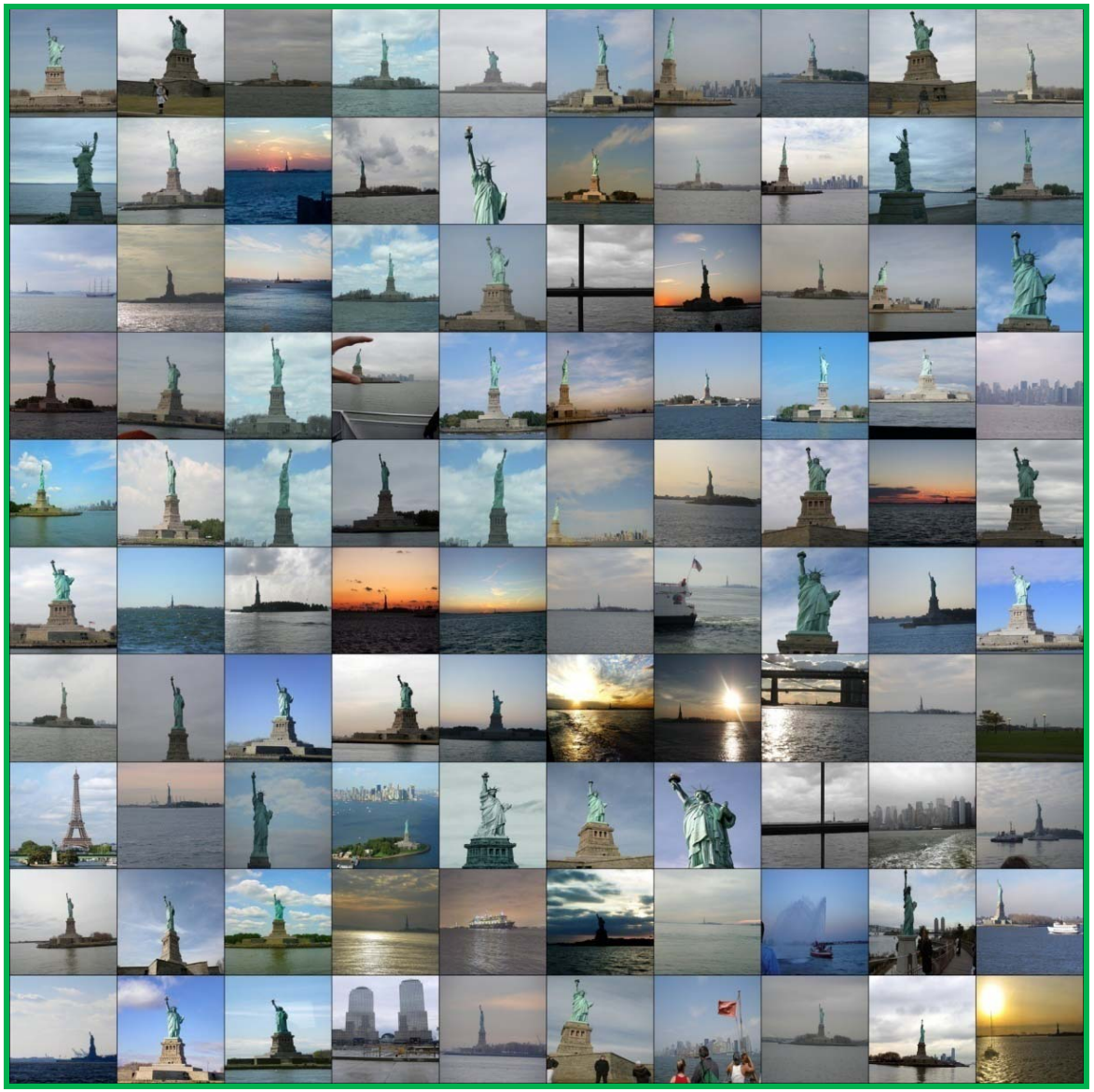
Google Search "Statue of Liberty"



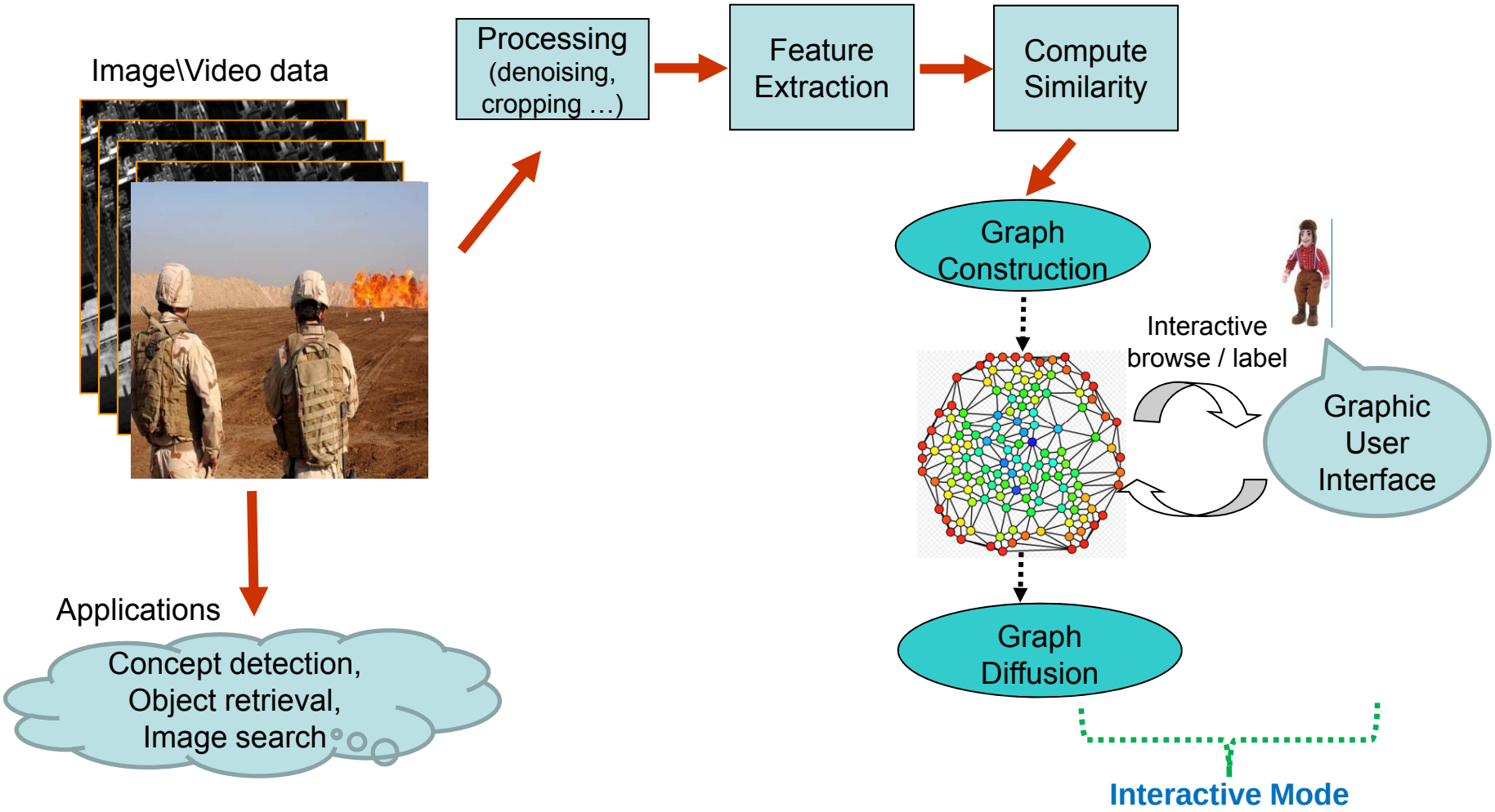
Application: Internet Search Result Reranking



Reranked Results



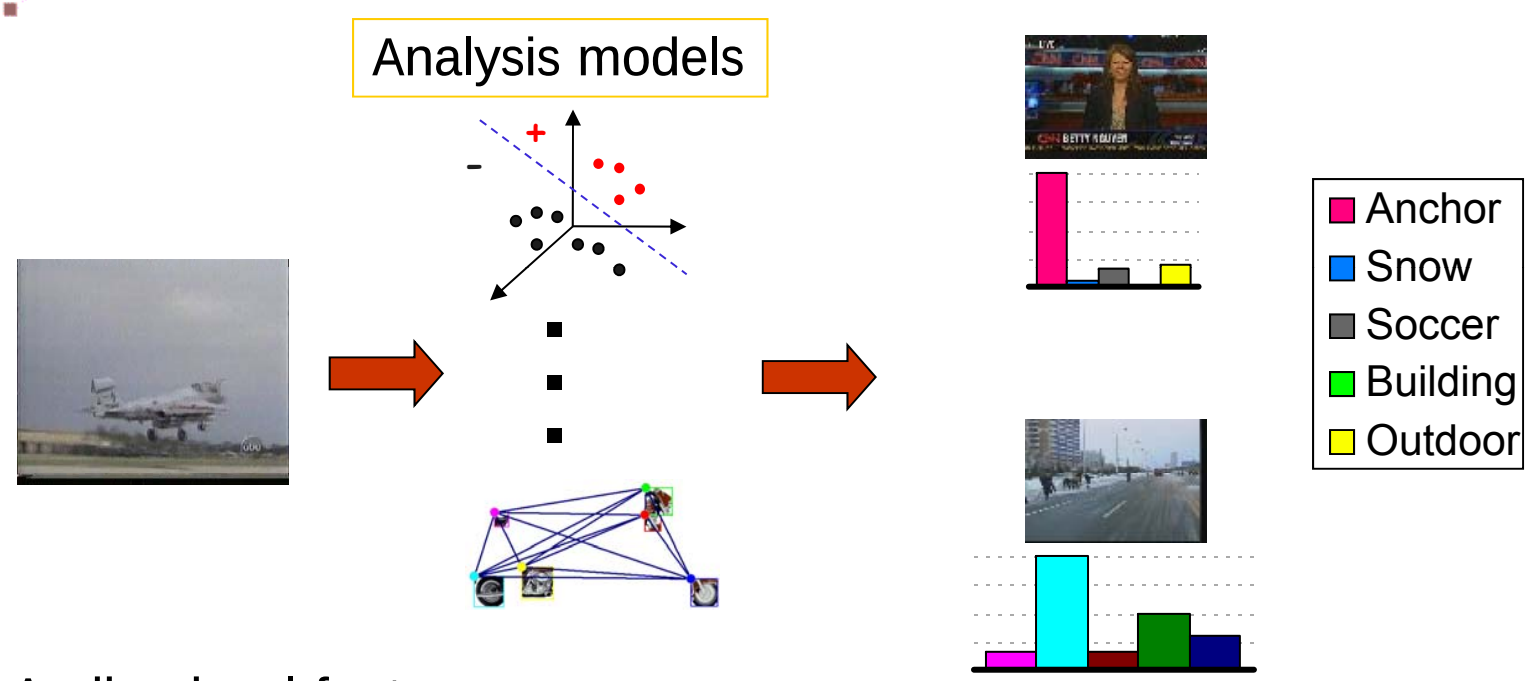
Interactive Image Annotation: Columbia TAG System



Columbia TAG Image Search System

- Demo:
  Rapid Image Annotation with
  User Interaction

Beyond Matching: Automatic Visual Recognition



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

- Rich semantic labels

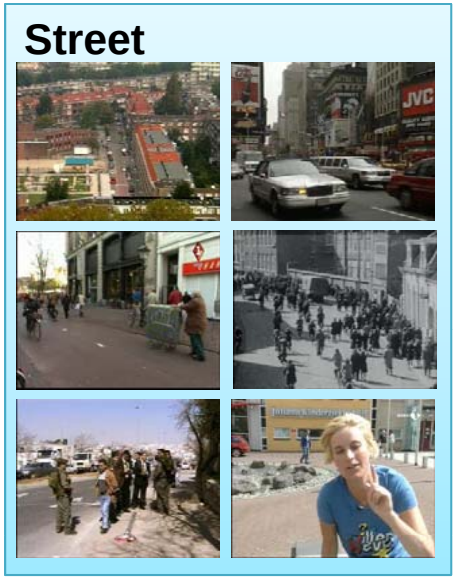
Hot topic ... community fast growing!

(as of Nov. 2009)

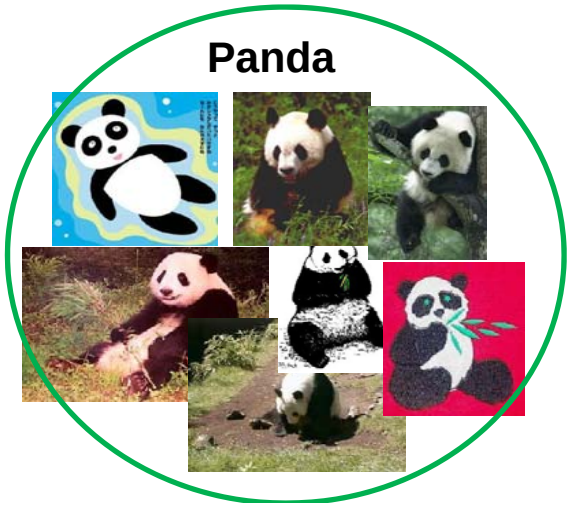
	Data domain	amount	types	Lexicon size
TRECVID	Broadcast news, documentary, Flickr, Youtube...	400 hours Sound & Vision 170 hours Television News 100 hours BBC rushes (130,000+ subshots)	video shots, keyframes	10 (2004, 2005) 39 (2006) 39 (2007) 20 (2008)
LSCOM	Broadcast news video	170 hours Television 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

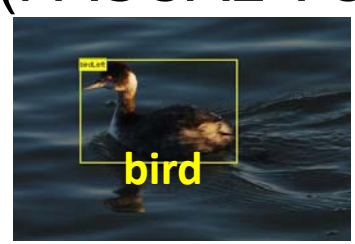


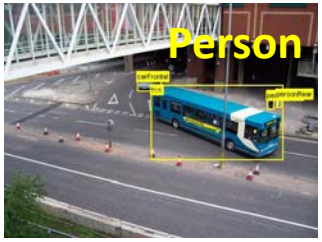
CalTech 101

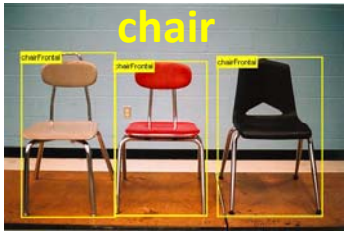


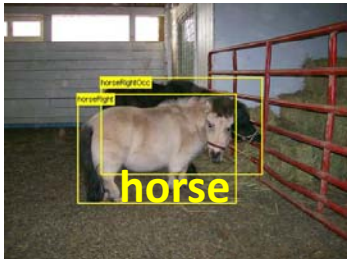
Object Localization

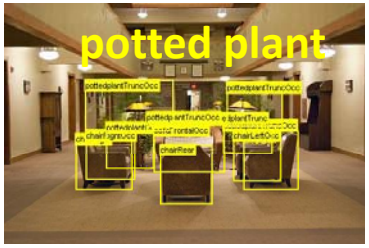
(PASCAL VOC)

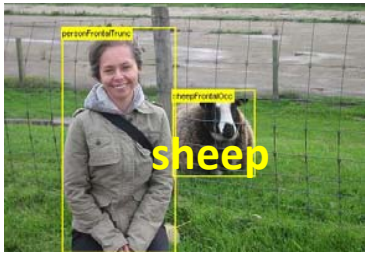






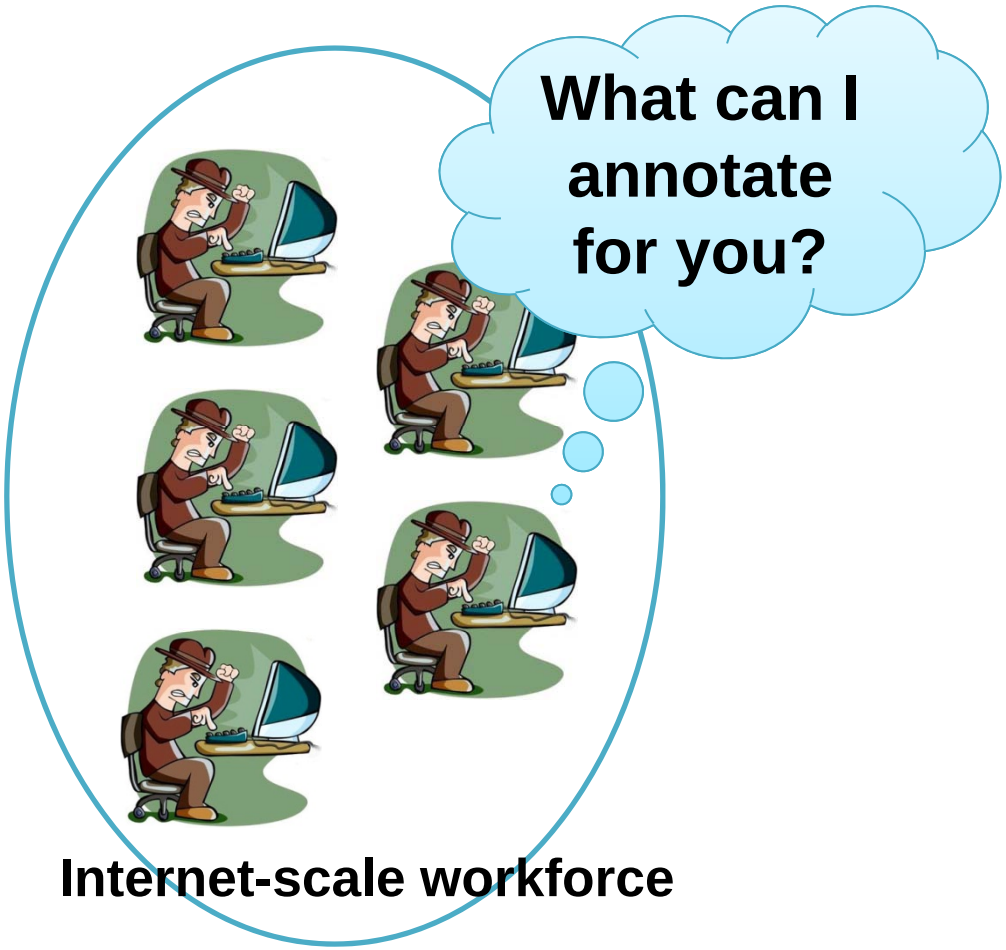






Create Large Corpora: Crowd Sourcing

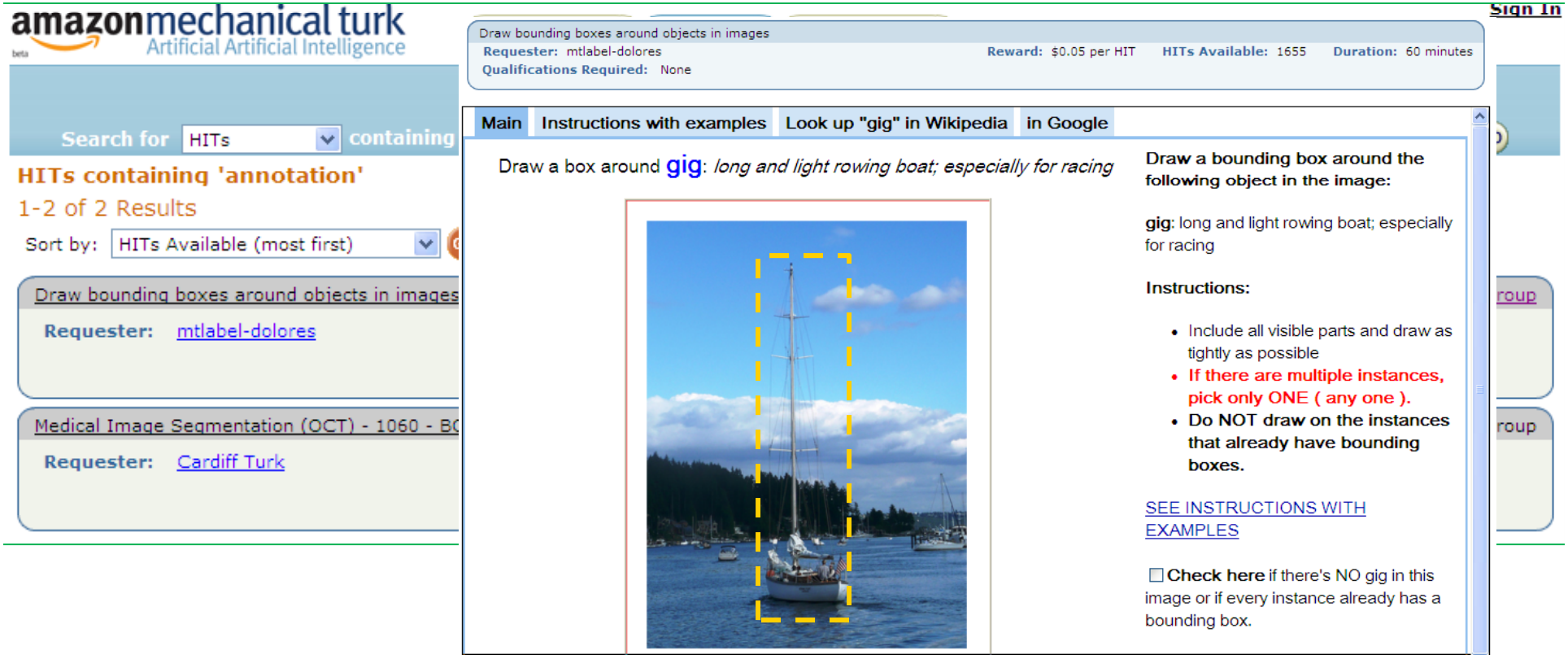
- Amazon Mechanic Turk
- Web Open Market for Human Computing



Internet-scale workforce

earn $0.03 per image tag

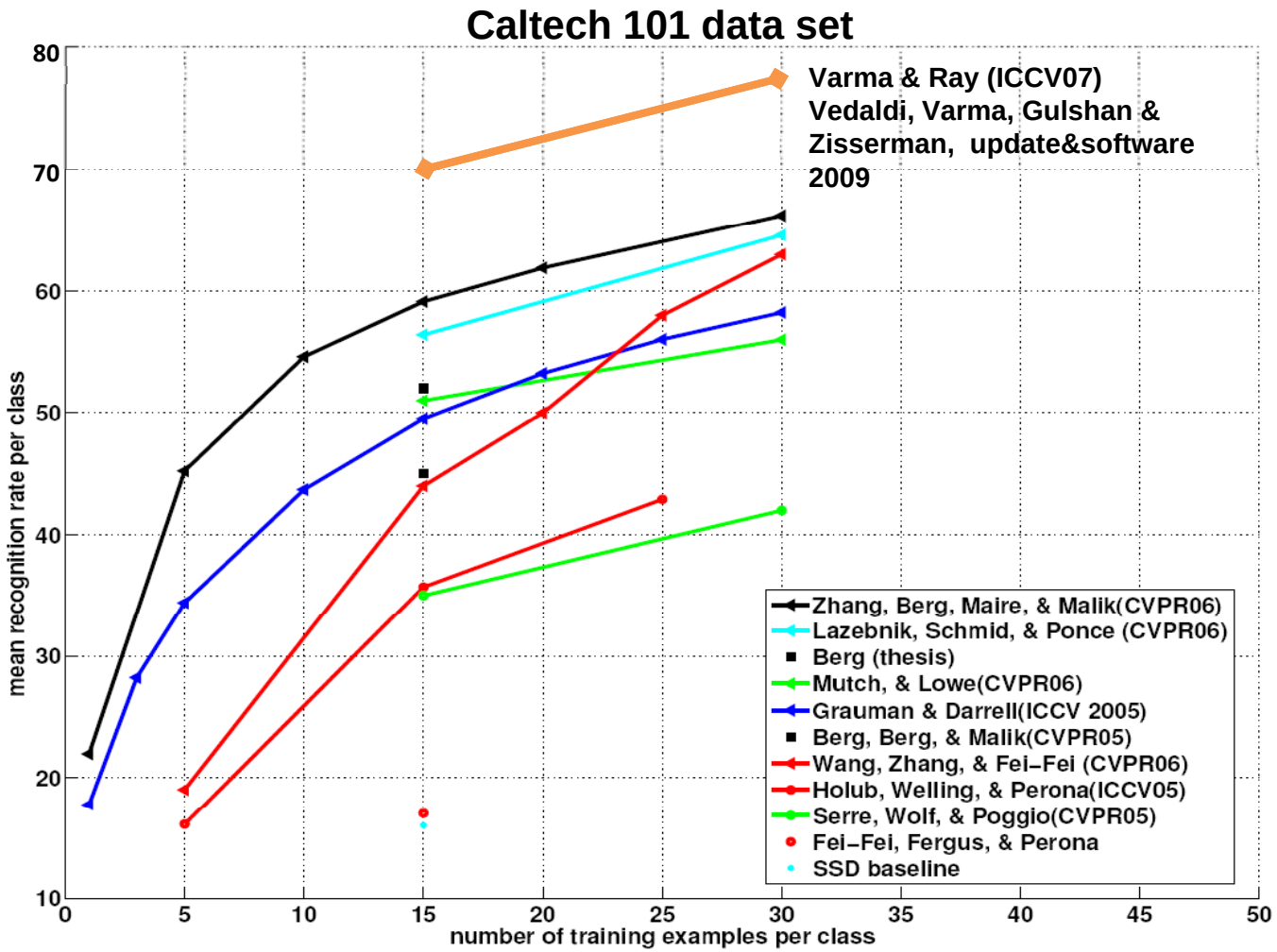
Crowd sourcing for image annotation



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

Rapid Advances in Image Annotation

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

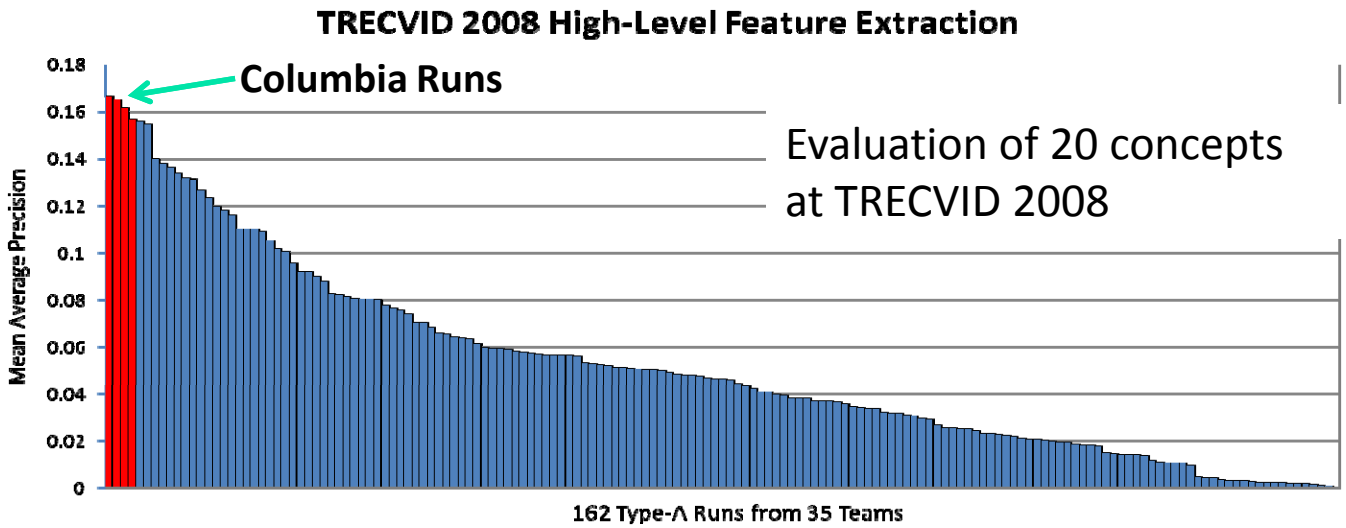


CuZero: tag videos with 400+ classifiers

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



TRECVID: Detection Examples

- Top five classification results

Classroom

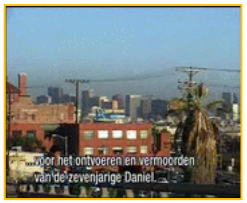



Demonstration Or Protest





Cityscape

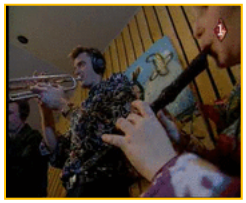


Airplane flying





Singing





What can a small recognition engine do?

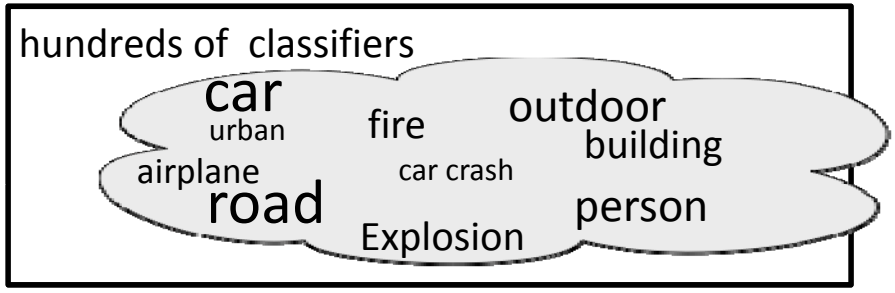
- How to leverage a noisy and small visual vocabulary?
- State of the Art video search engines have ~1000 visual concepts
  - IBM IMAR
  - U. Amsterdam MediaMill
  - CMU Informedia
  - Columbia CuZero

Obvious Problems with Small Vocabulary

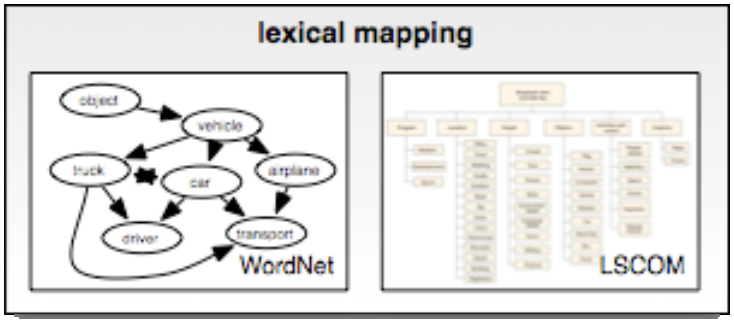
- Given a search topic, users often have difficulty in choosing matched concept classifiers

Find shots of something burning with flames visible

Cannot find matched classifiers!
Which classifiers work?

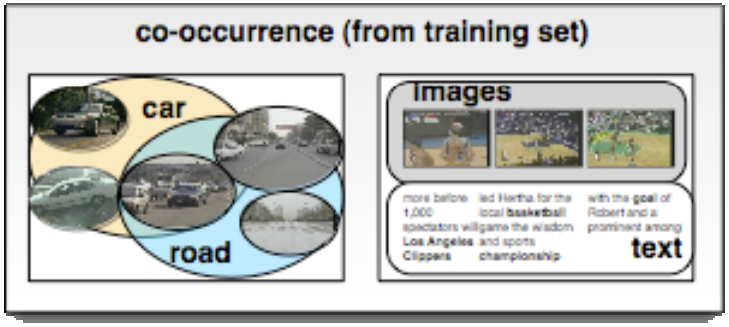


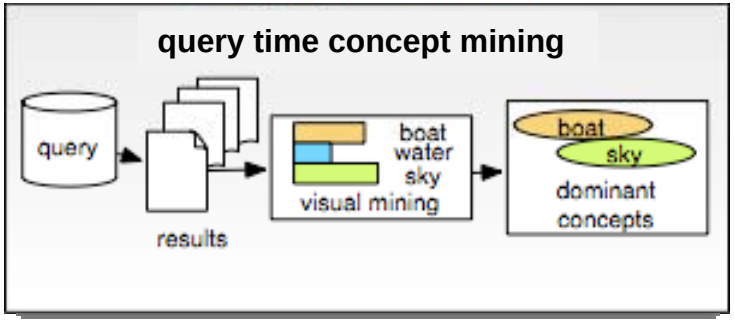
Ideas from Information Retrieval: Query Expansion



Classifier Suggestion

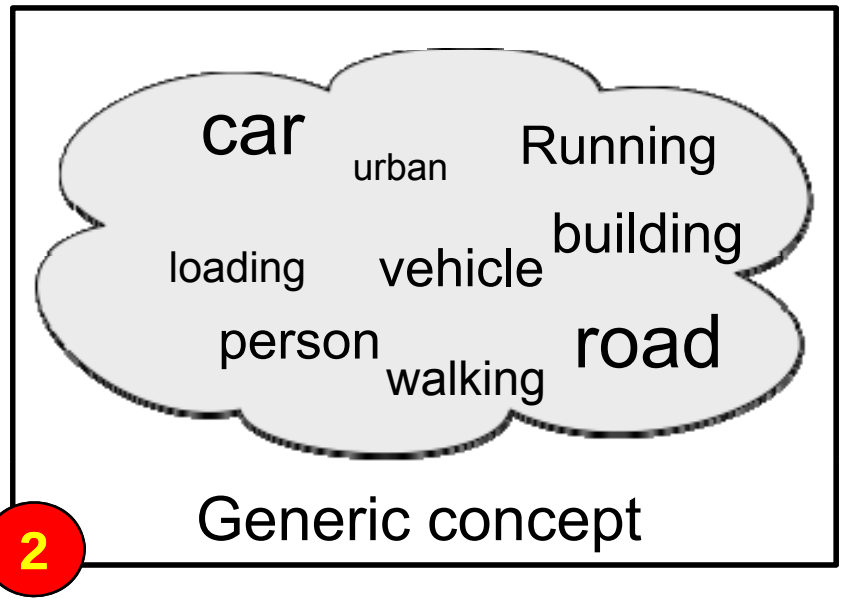
keywords

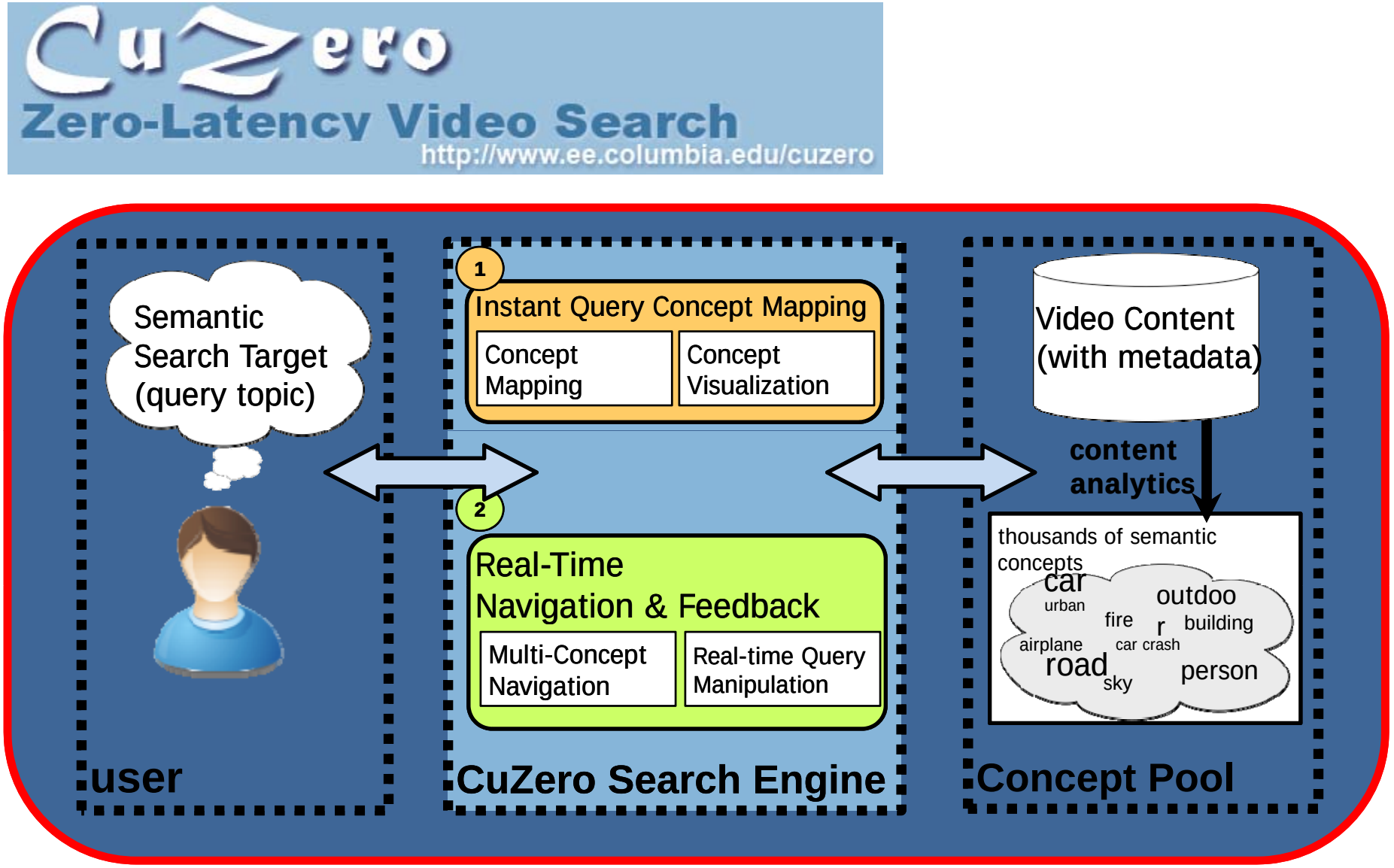




Use visual examples to compensate the deficiency

Query: “find person running around a building”





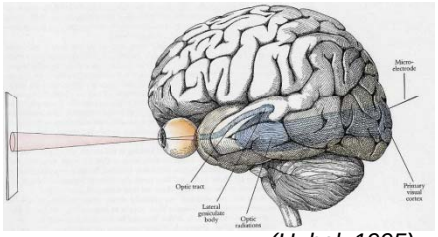
(Zavesky and Chang ,
Multimedia Info Retrieval MIR '08)

Demos

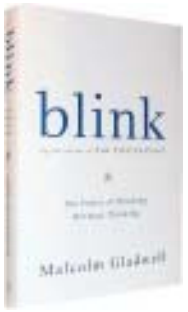
- Find lake front buildings in the park
- Find person walking around building
- Find a car on a road in a snowy condition

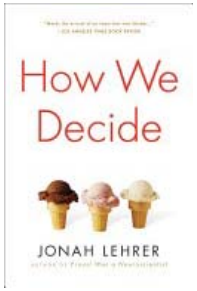
Visual Search via Brain State Decoding

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



(Hubel, 1995)



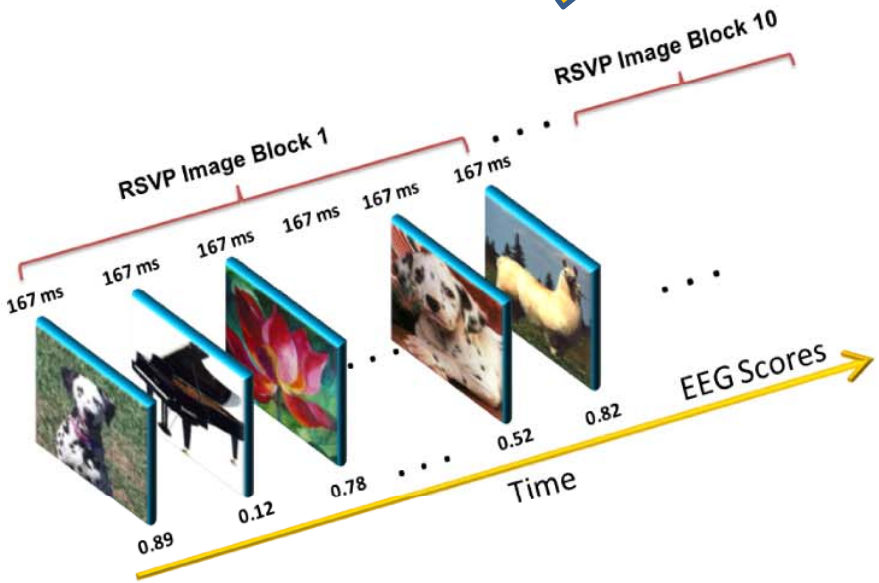


Joint work with Paul Sajda's group

Brain Machine Interface for Image Retrieval

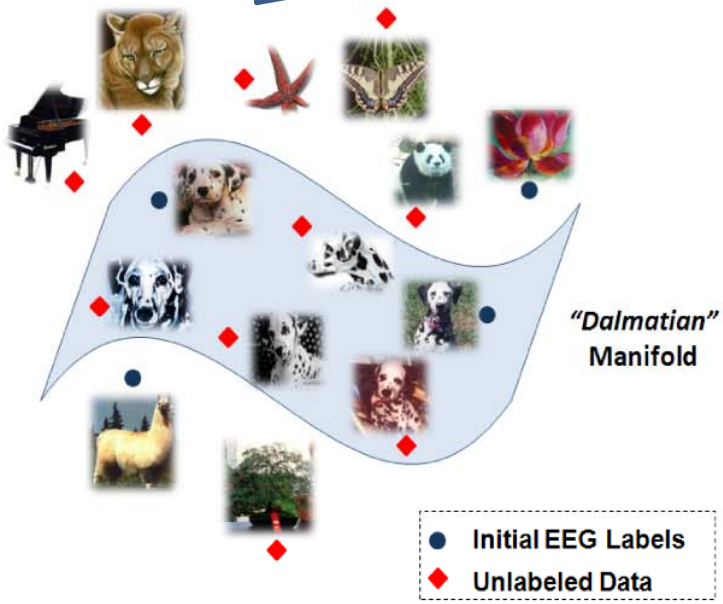
Use EEG brain signals to detect target of interest
(video)

Use image graph to tune & propagate information



Rapid Serial Presentation of Caltech 101 Images

C3Vision System
by Sajda et al

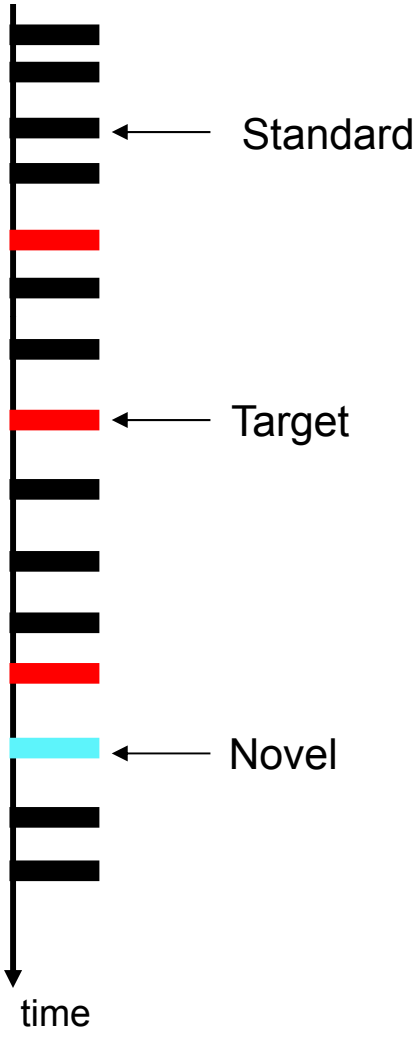


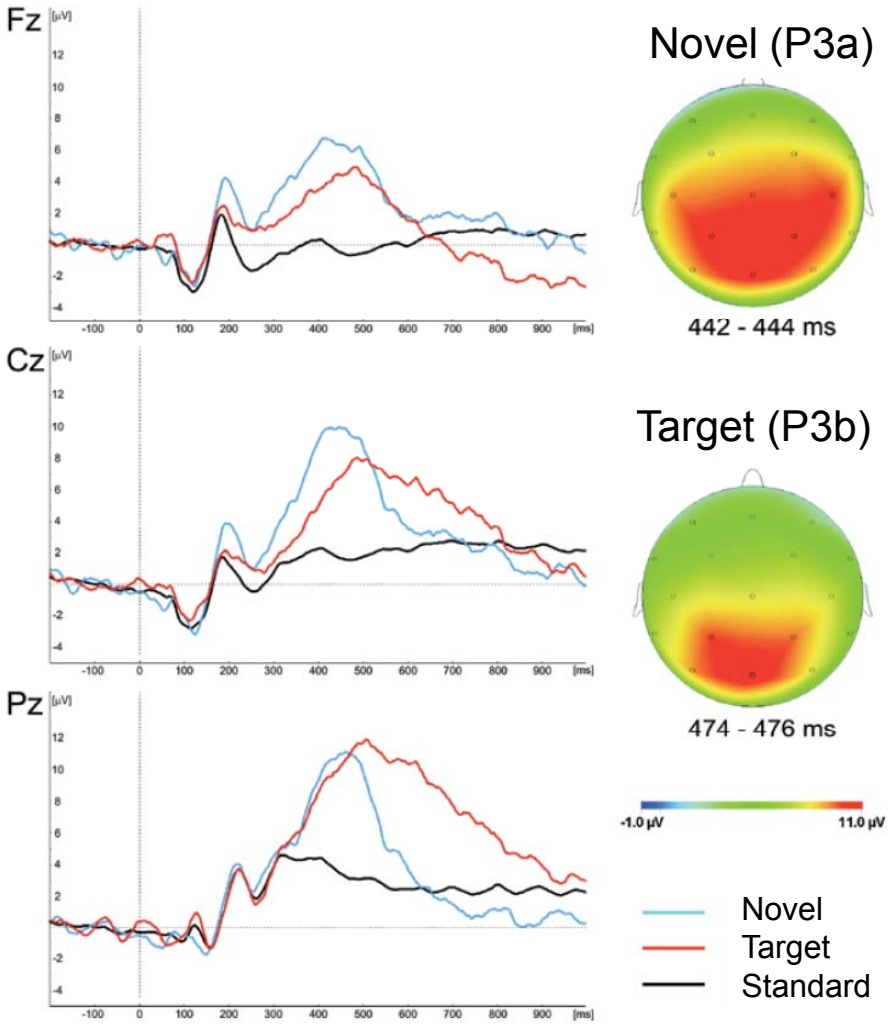
Graph-Based Visual Pattern Discovery

TAG System
by Wang and Chang

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

The Neural Signatures of "Visual Attention"

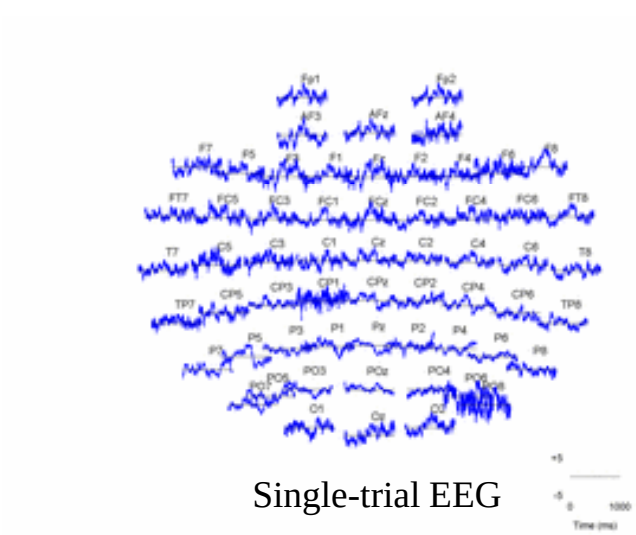




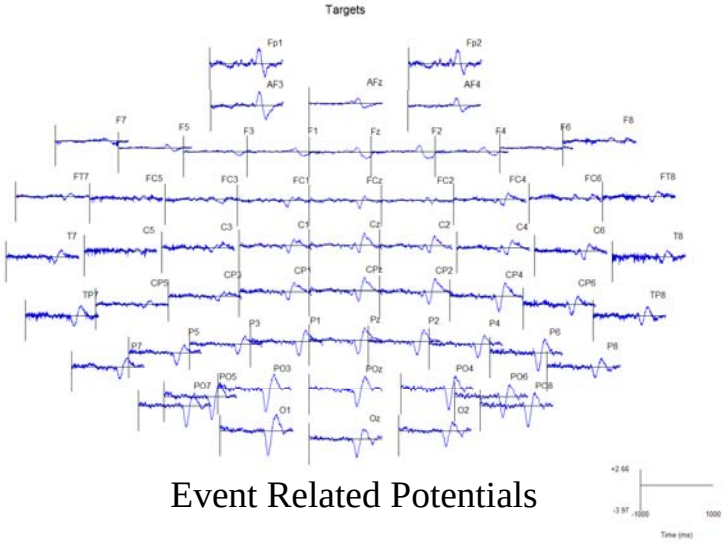
From D. Linden, 2005

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

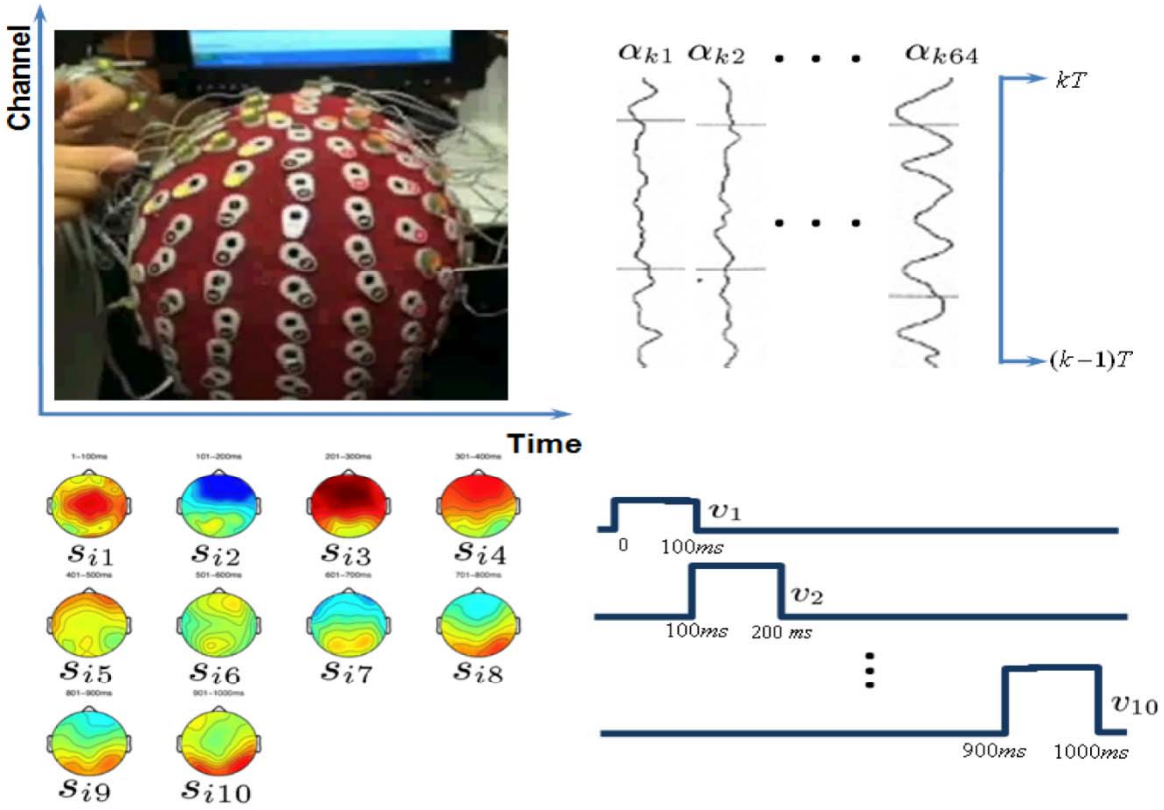


Event Related Potentials

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

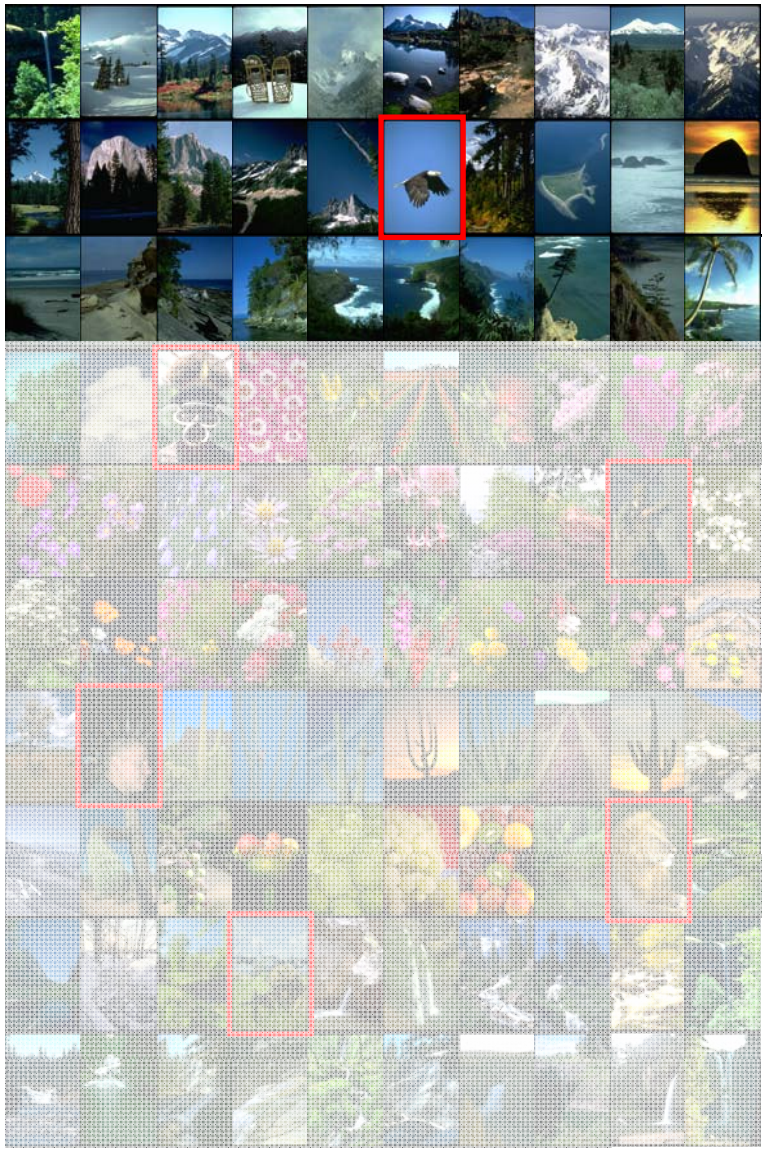
(Parra, Sajda et al. 2002, 2003)



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

Optimal temporal filtering over time windows after onset

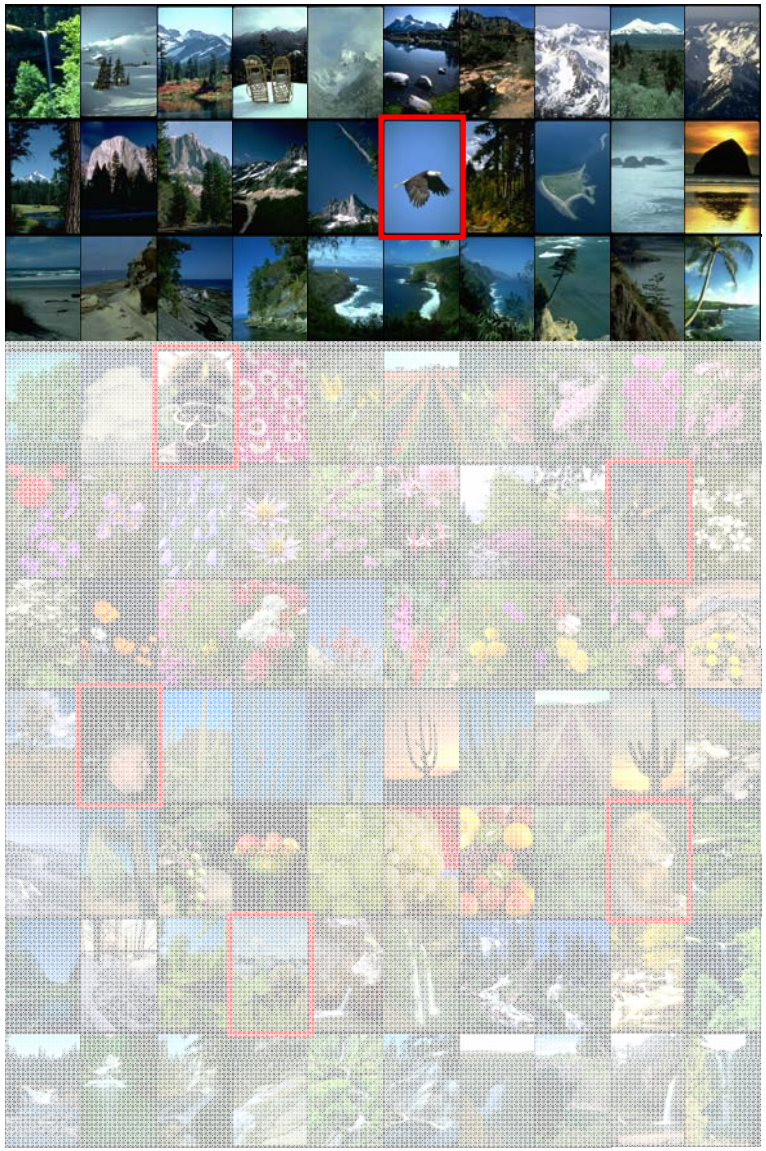
The Visual Interest Readout Experiment



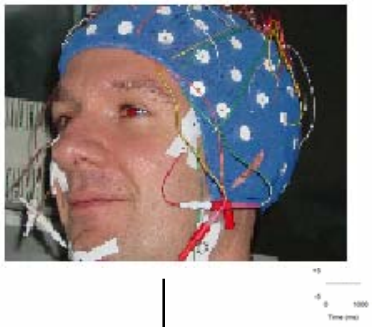
Database (any target that may interest users)

User thinks about what he/she wants to search

The Paradigm



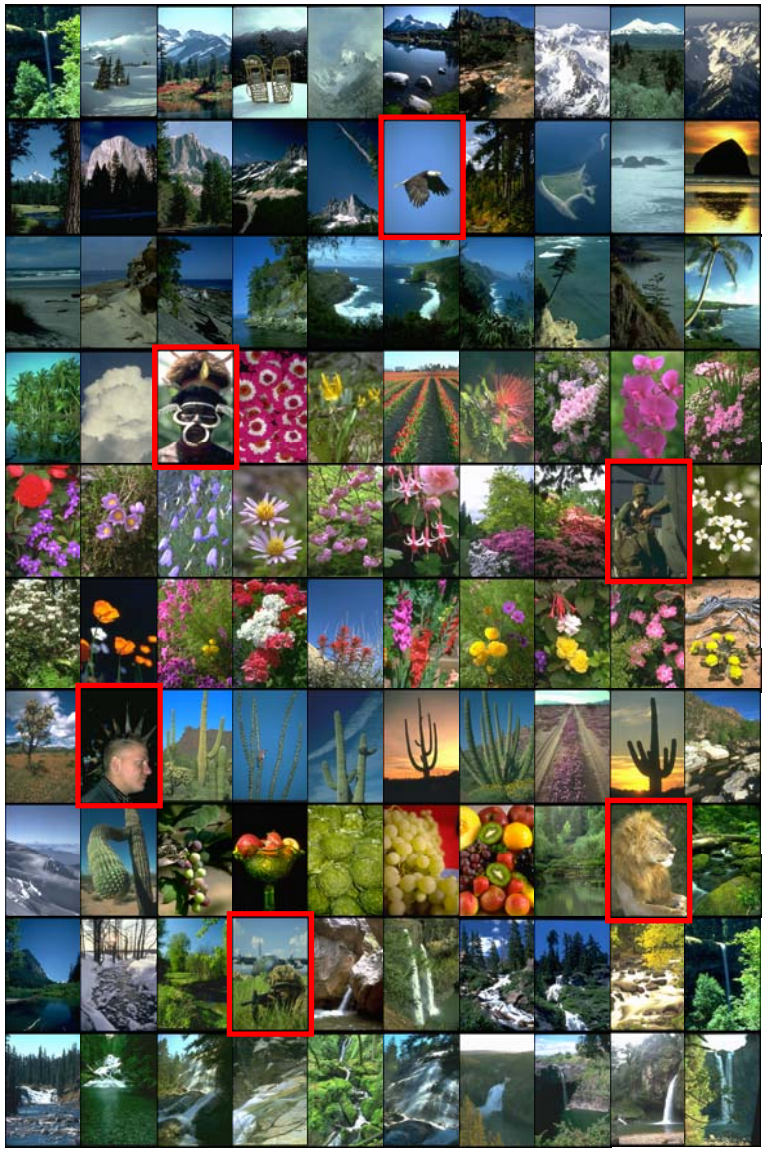
Database



Neural (EEG) decoder

Interest-scores

The Paradigm



Database

Neural (EEG) decoder

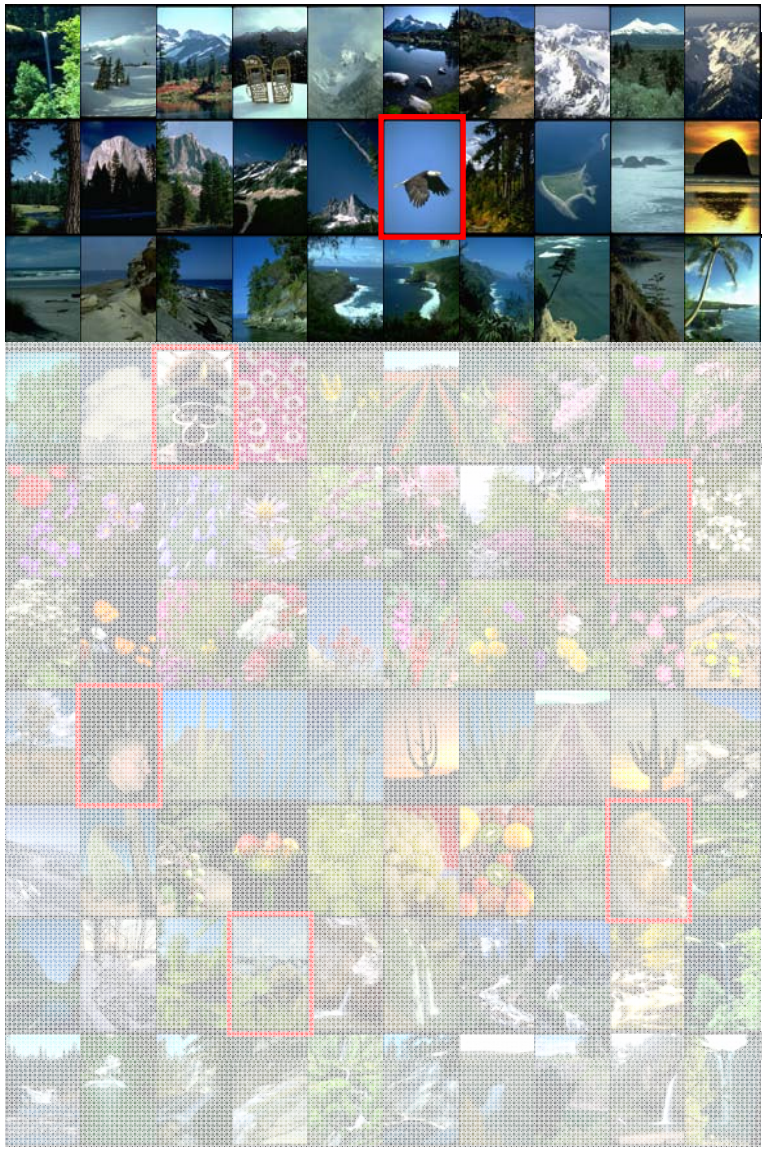
Exemplar labels (noisy)

Semi-supervised
Graph-based propagation

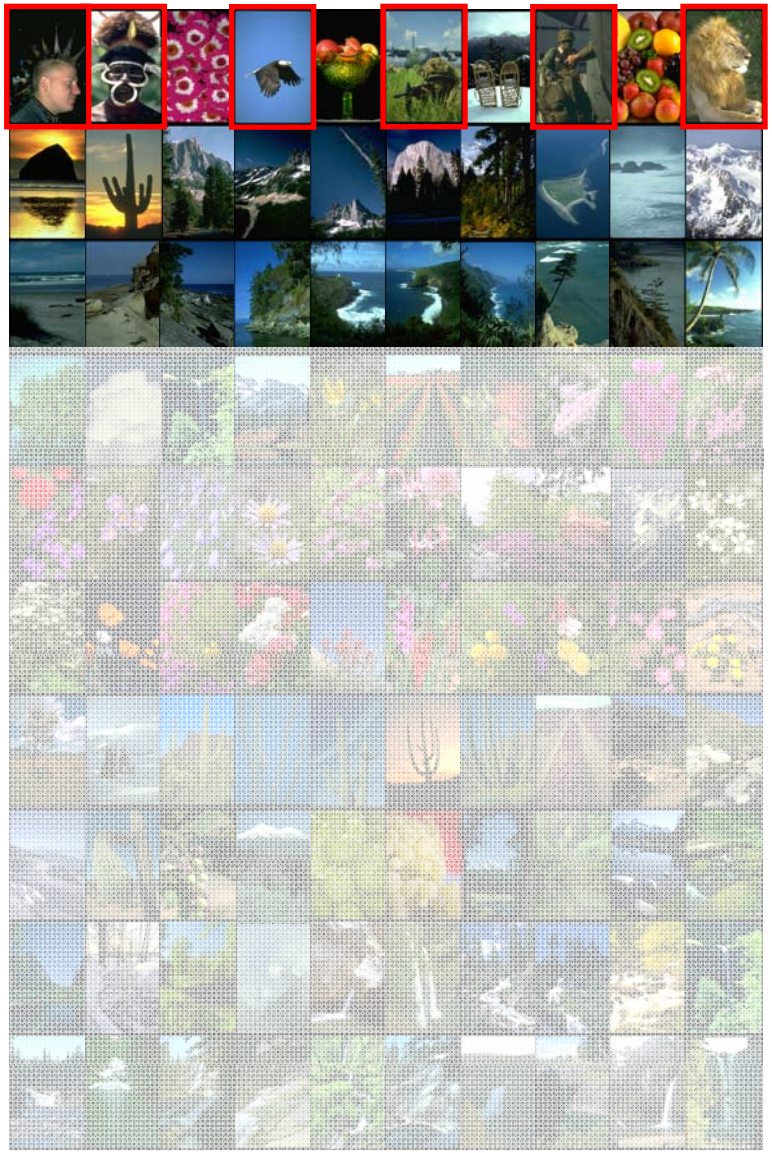
Features from the entire DB

prediction score

The Paradigm

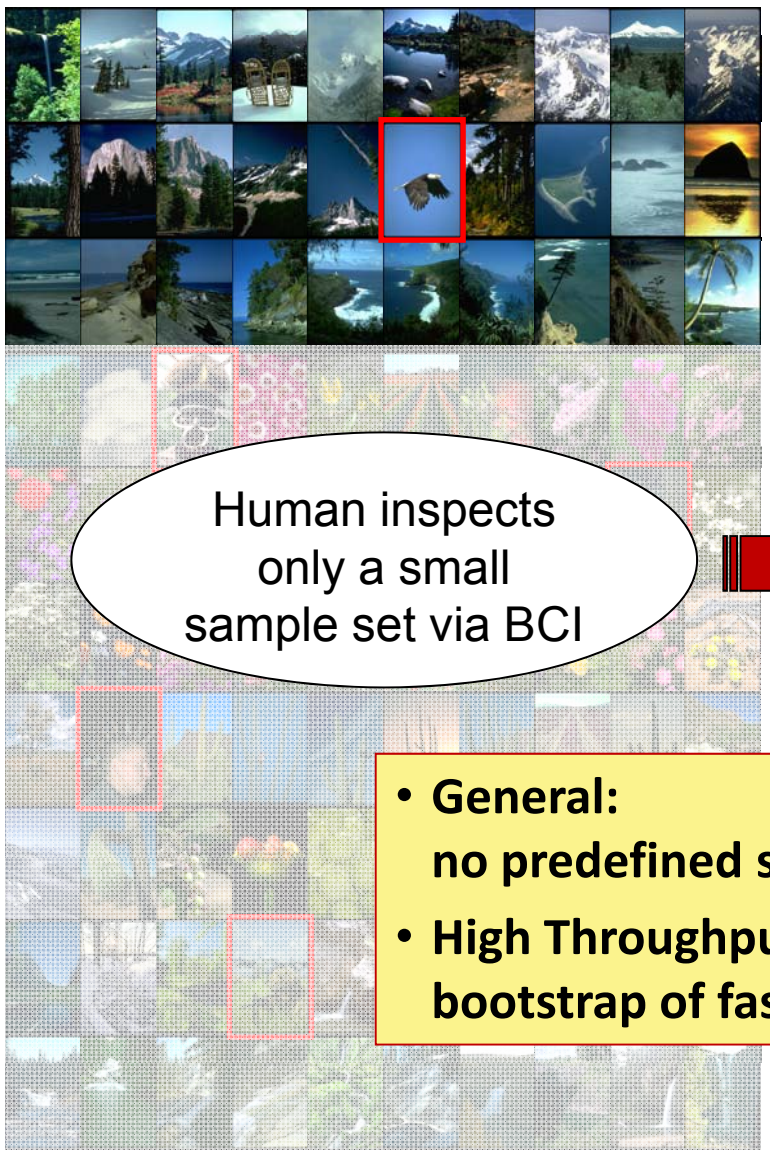


Pre-triage

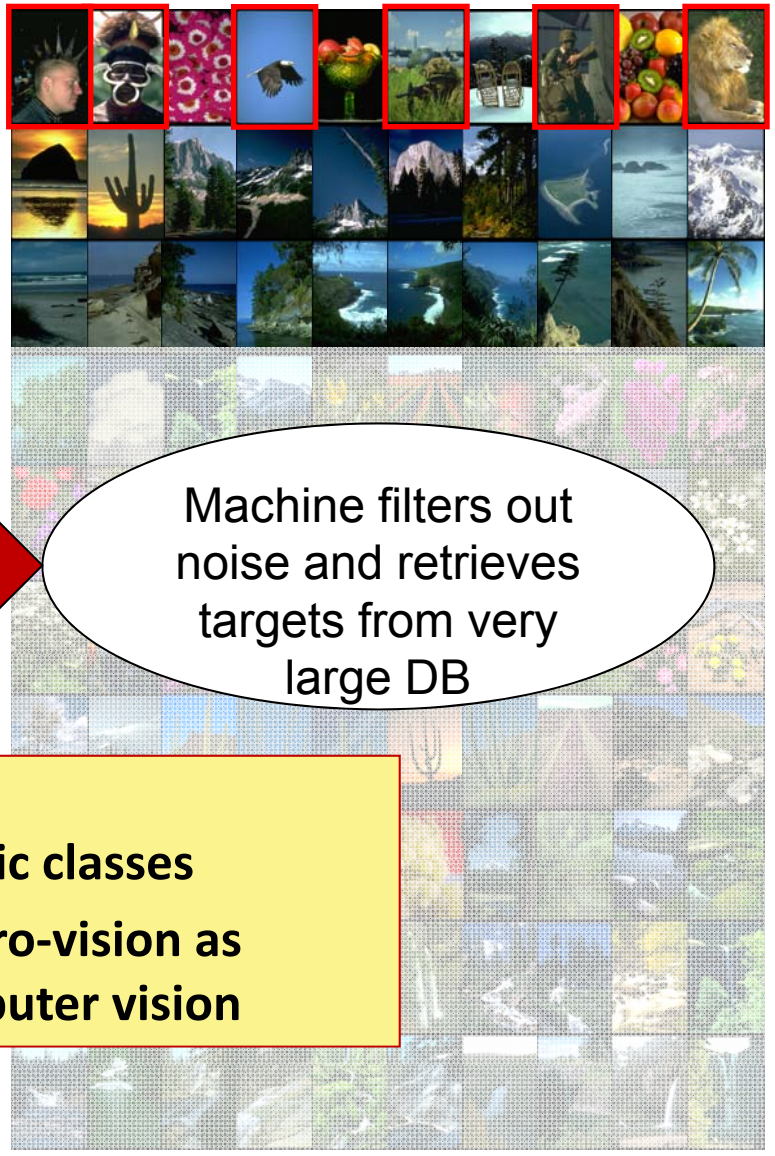


Post-triage

The Paradigm



Human inspects only a small sample set via BCI



Machine filters out noise and retrieves targets from very large DB

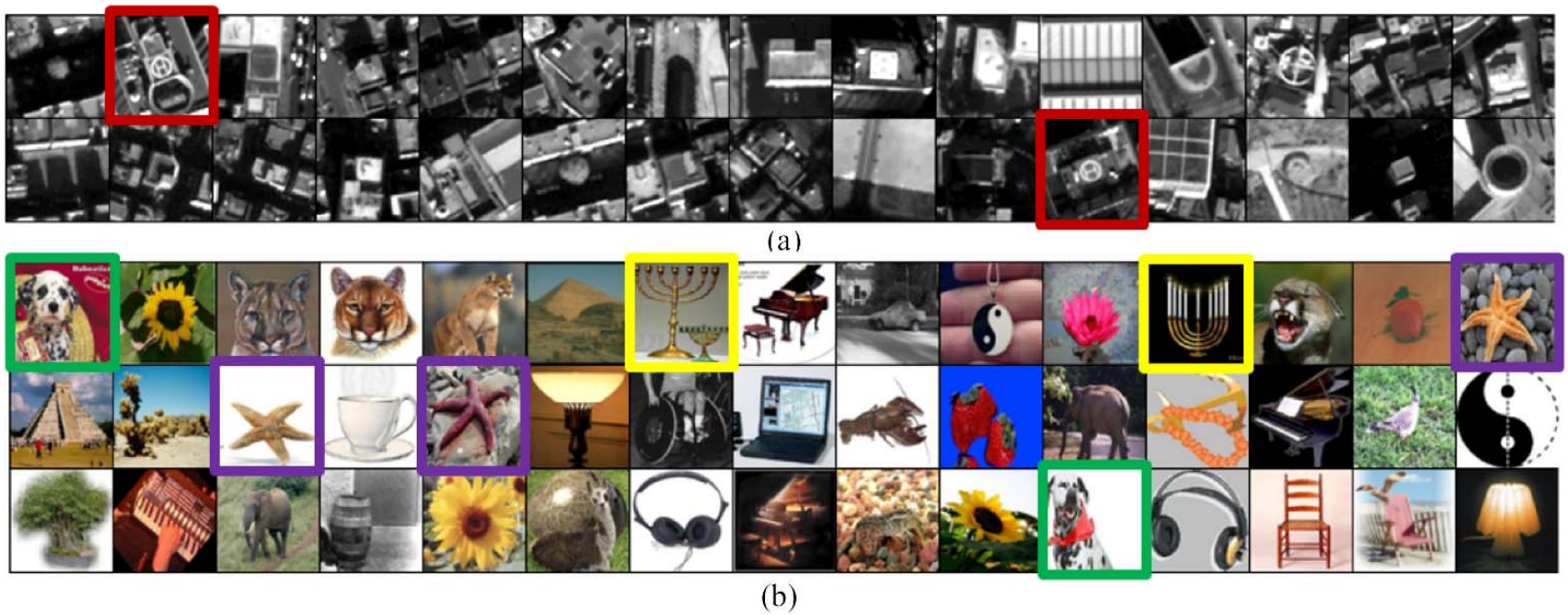
- **General: no predefined semantic classes**
- **High Throughput: neuro-vision as bootstrap of fast computer vision**

Pre-triage

Post-triage

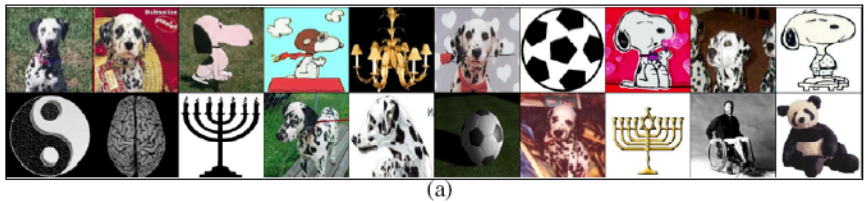
Experiments

- 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)*

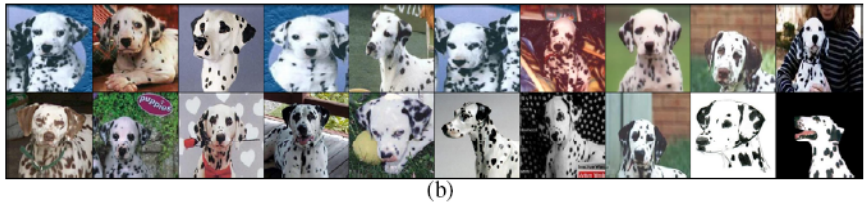


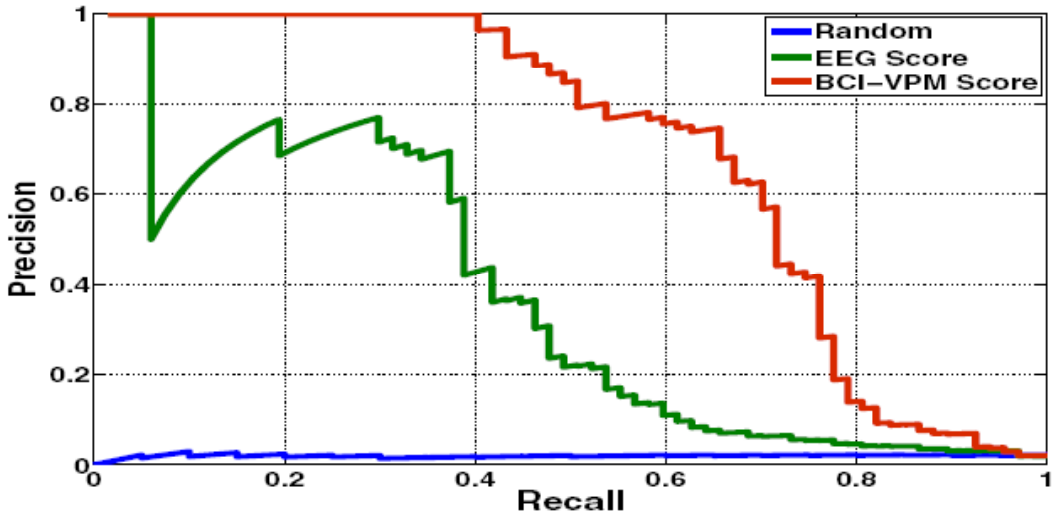
Example results

Top 20 results of
Neural EEG detection



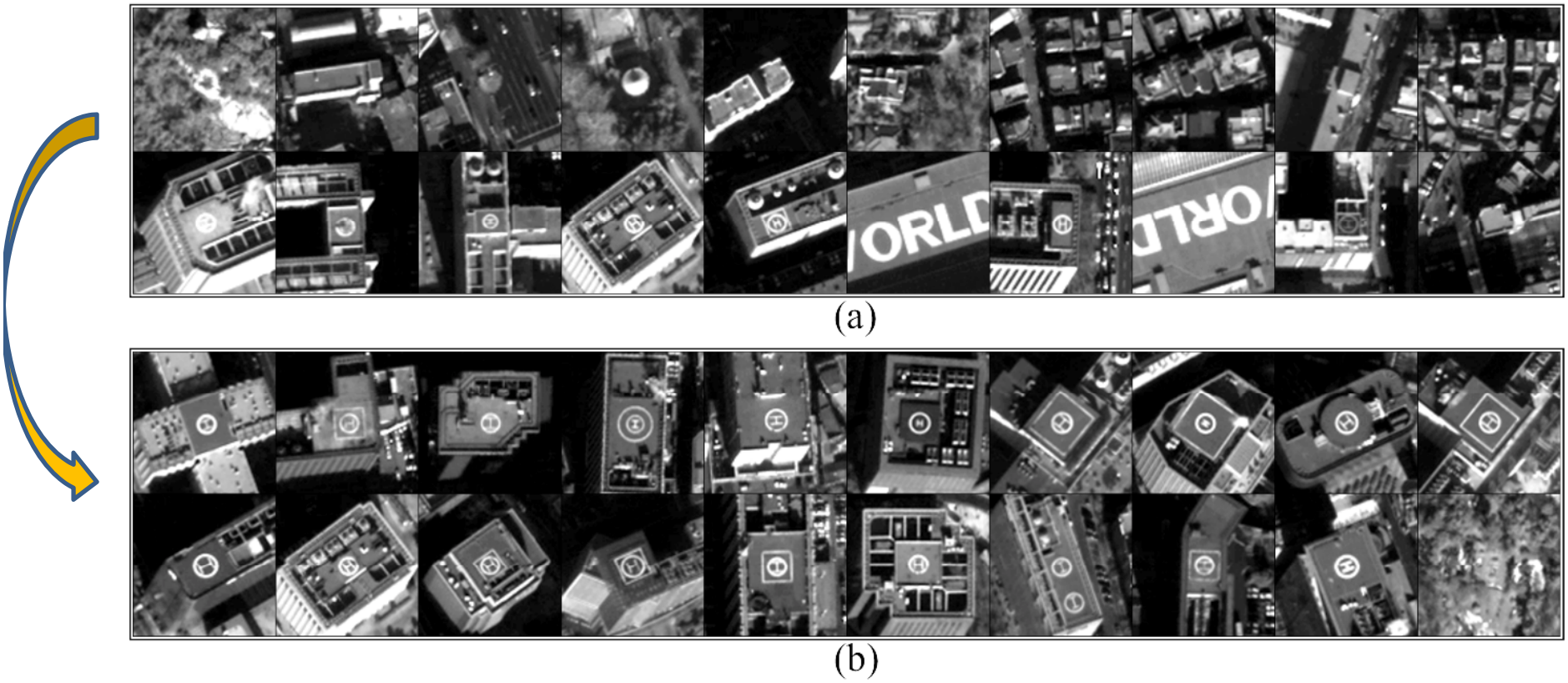
Top 20 results of
Hybrid System (BCI-VPM)





DARPA NIA Program: Remote Sensing Target Search

Initial EEG neural signal detection:



(a)

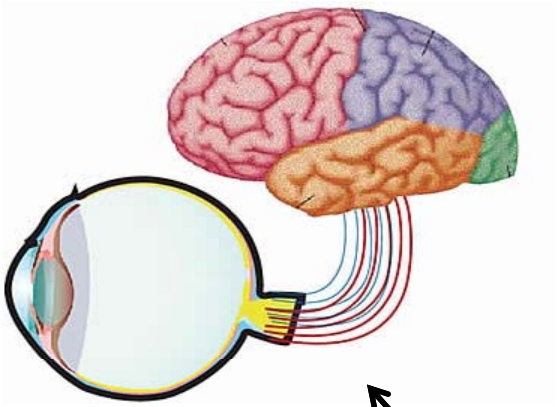
(b)

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

Acknowledgments

- Columbia University
  - Eric Zavesky, Yu-Gang Jiang, Jun Wang, Junfeng He, Wei Liu, Wei Jiang, Akira Yanagawa
- 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.