Machine Learning Summer School

Beijing, 2014

Weakly-supervised Structured Learning:

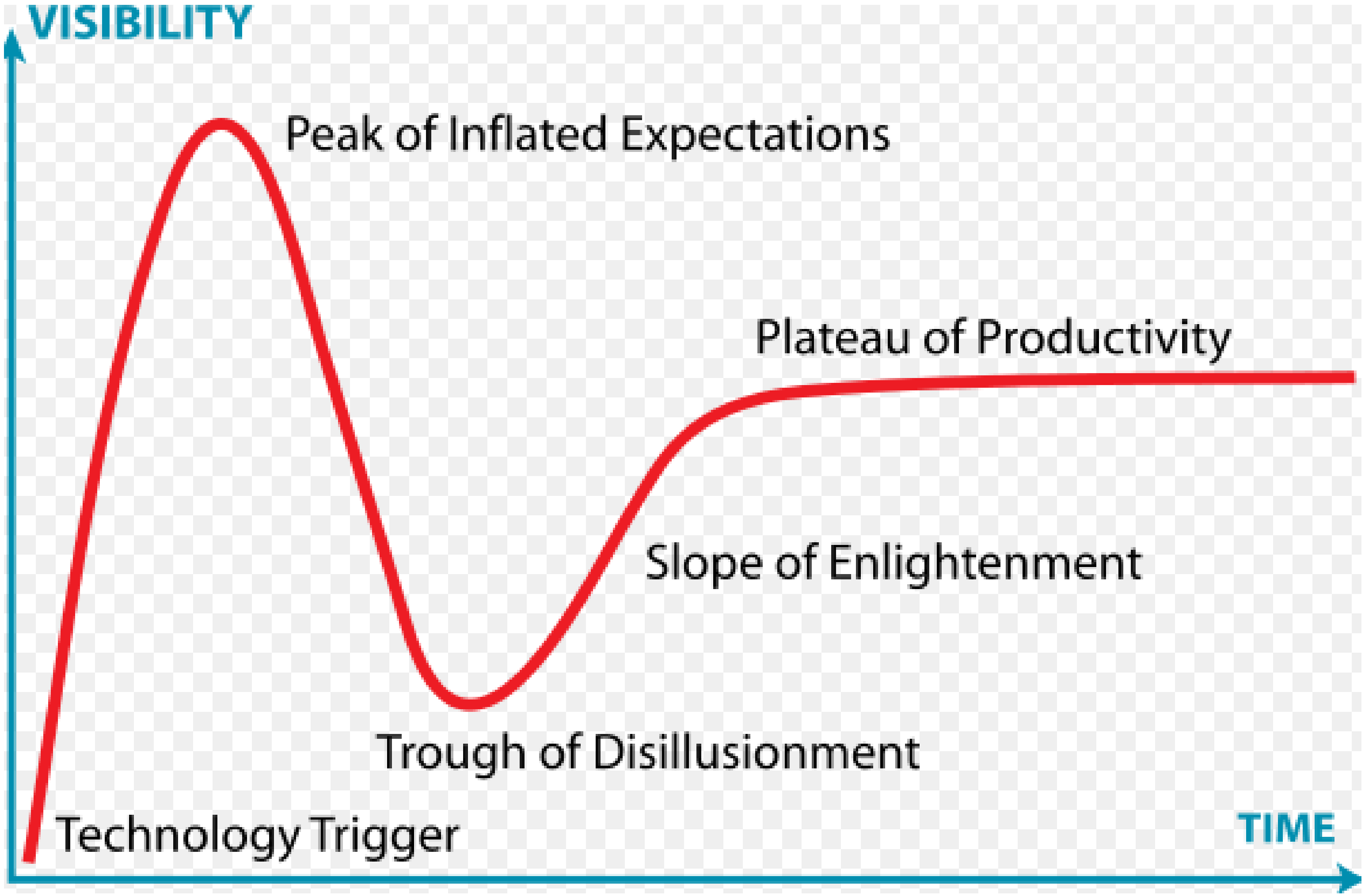
Zhuowen Tu

Department of Cognitive Science

Department of Computer Science and Engineering

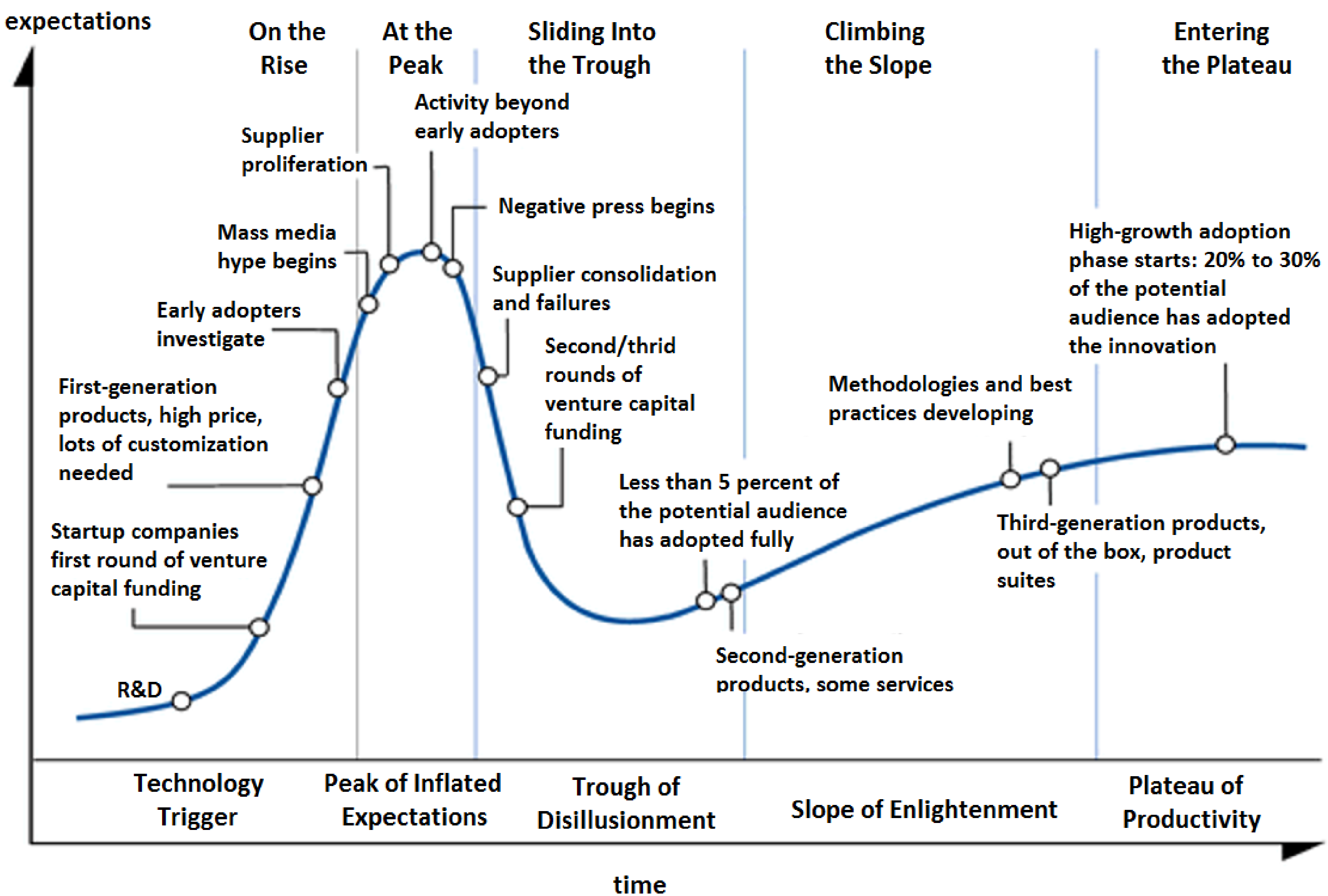
University of California, San Diego

Hype cycle



http://en.wikipedia.org/wiki/Hype_cycle

General hype cycle for technology



http://en.wikipedia.org/wiki/Hype_cycle

Hype cycle of current technologies

Courtesy of Li Deng

Part I:

Overview of some common and competing machine learning concepts

Questions to ask

Lessons about competing concepts in machine learning?

Why is learning structures so important?

Why do we emphasize weak-supervised learning?

Several pairs of competing concepts

Generative

$$p(y, x)$$

Discriminative

$$p(y|x)$$

Parametric

$$y = f(x)$$

Non-parametric

$$y = \sum_{k=1}^{K} \alpha_k f_k(x)$$

Supervised

$$\{(y_i, x_i), i = 1..N\}$$

Unsupervised

$$\{(x_i), i = 1..N\}$$

Dense

$$\|x\|_2$$

Sparse

$$\|x\|_0$$

Flat (shallow)

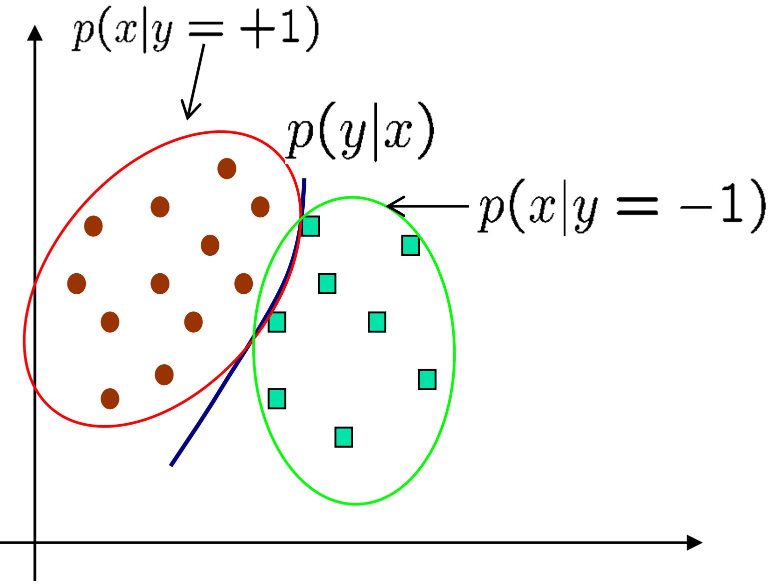
$$y = f(x)$$

Deep

$$y = f^{(n)}\left(f^{(n-1)} \ldots \left(f^{(1)}(x)\right)\right)$$

Discriminative vs. generative models

$$p(y|x)$$



Discriminative models, either explicitly or implicitly, study the posterior distribution directly.

$$p(x|y),\ p(y)$$

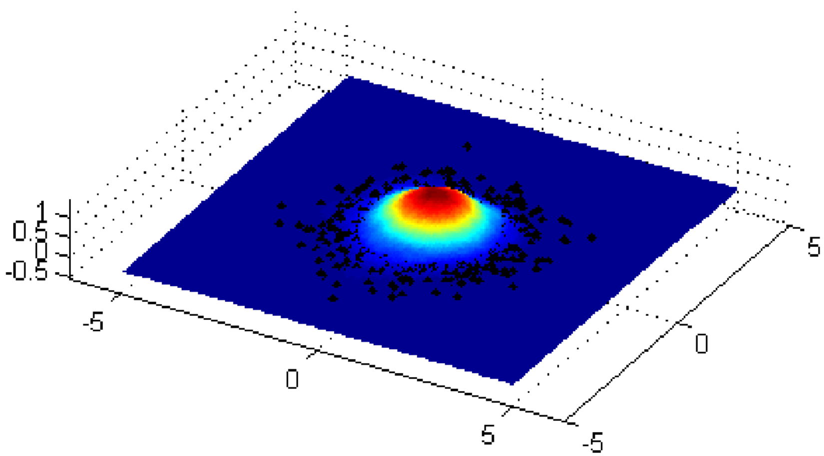
Generative approaches model the likelihood and prior separately.

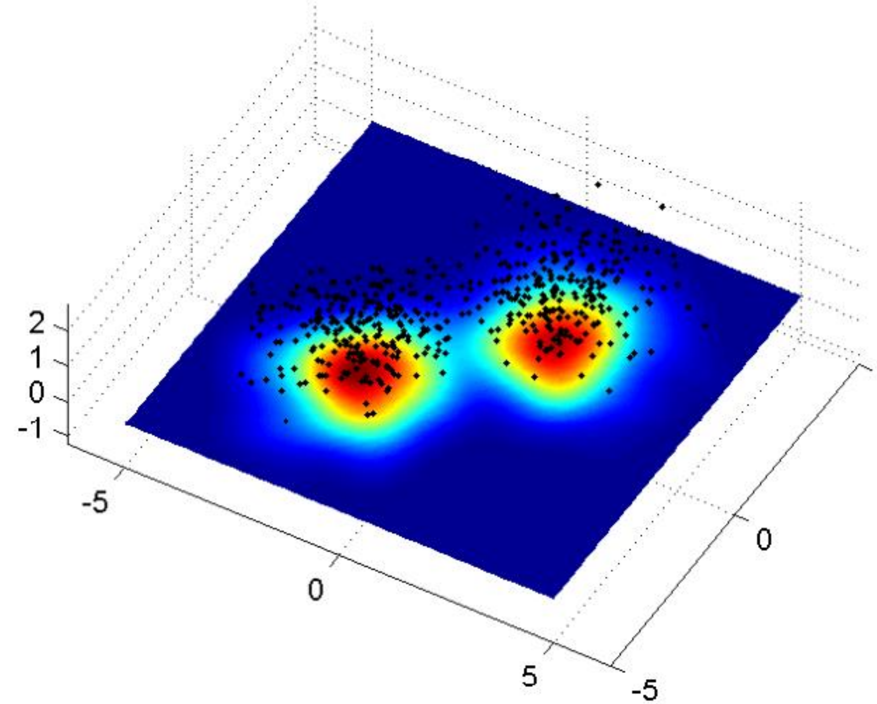
$$p(y|x) = \frac{p(x|y)p(y)}{\sum_y p(x|y)p(y)}$$

Parametric vs. non-parametric

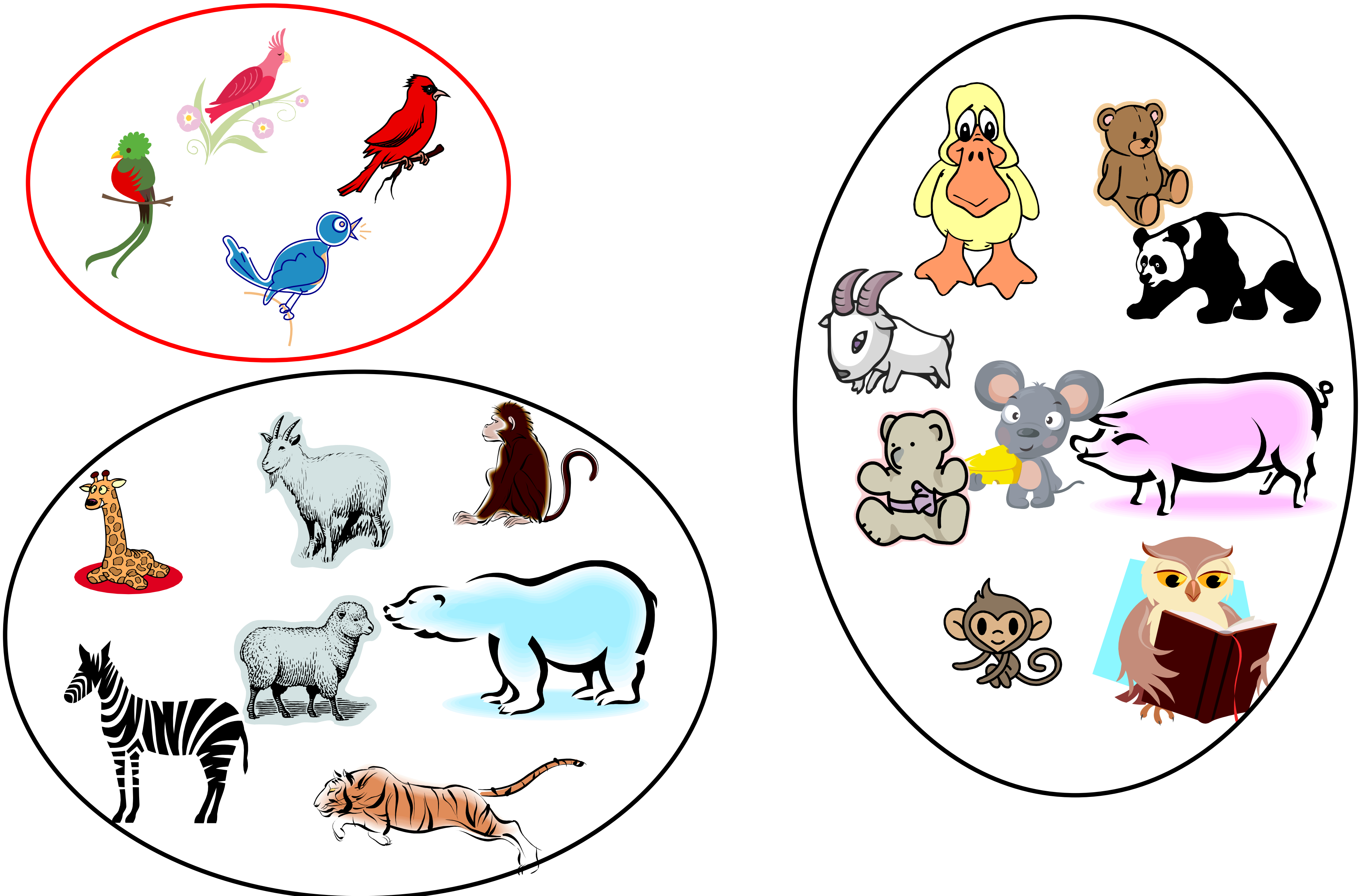
$$y = f(x)$$

$$y = \sum_{k=1}^{K} \alpha_k f_k(x)$$



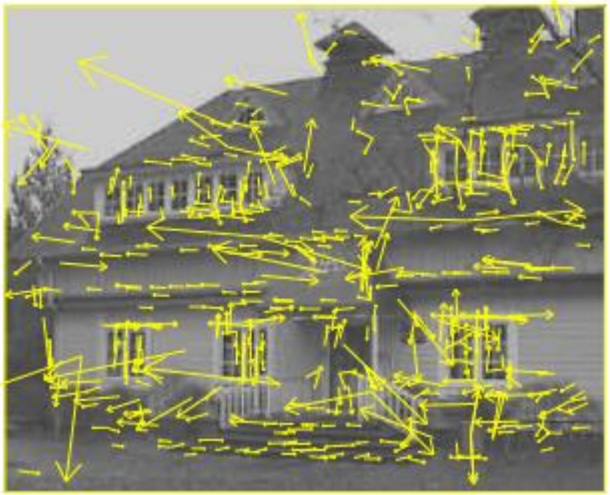


Supervised vs. unsupervised

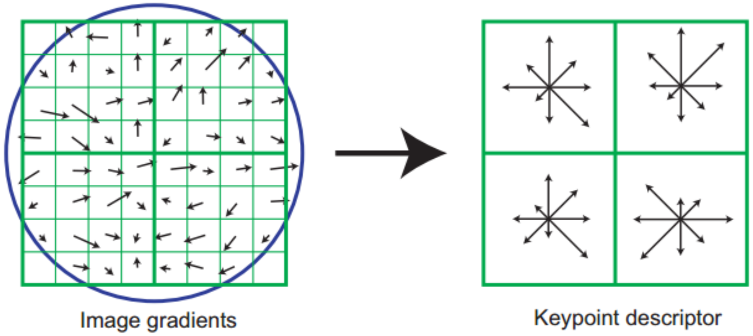


Dense vs. sparse

HOG descriptor (Dalal and Triggs)

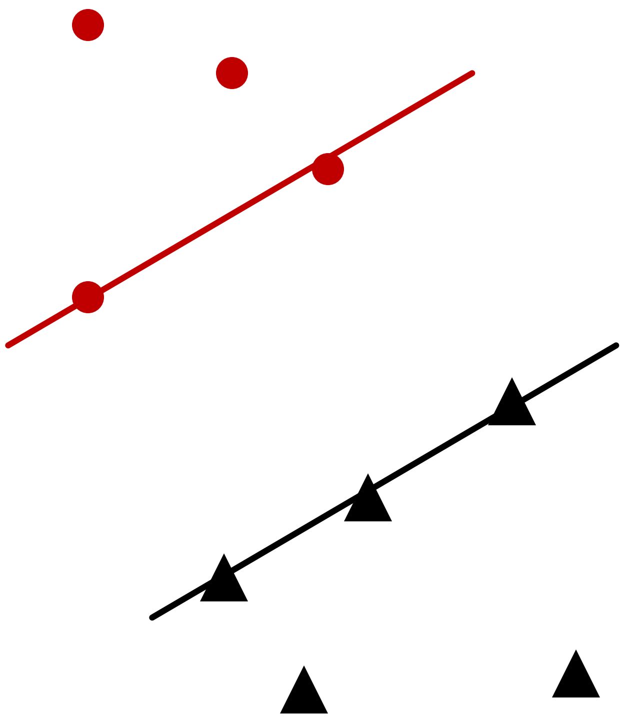


SIFT detector (D. Lowe)

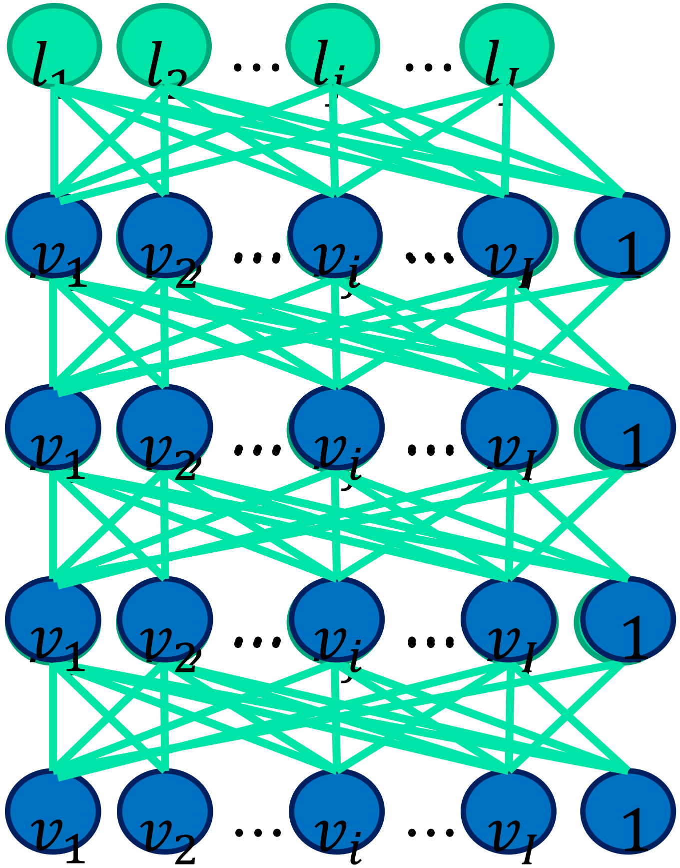


SIFT descriptor (D. Lowe)

Flat vs. deep



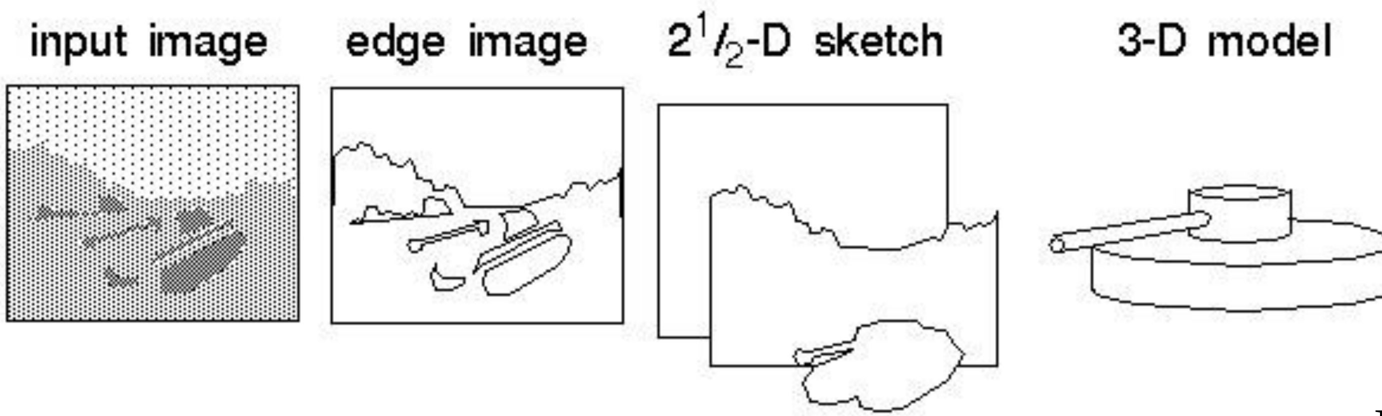
SVM (V. Vapnik)



Deep belief nets (G. Hinto)

Marr's theory

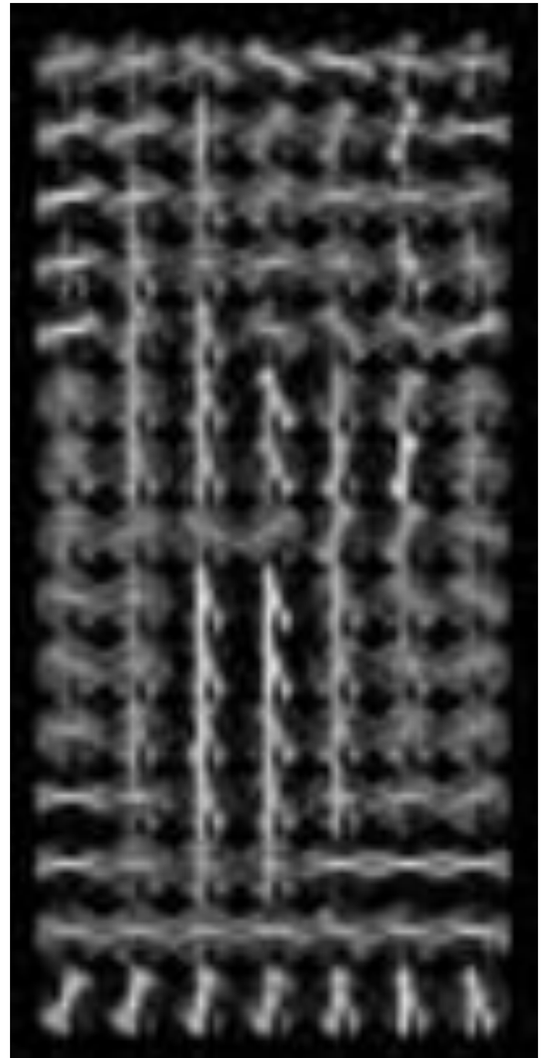
Viewer centred			*Object centred*
Input Image	Primal Sketch	2 1/2-D Sketch	3-D Model Representation
Perceived intensities	Zero crossings, blobs,edges, bars, ends, virtual lines, groups, curves boundaries.	Local surface orientation and discontinuities in depth and in surface orientation	3-D models hierarchically organised in terms of surface and volumetric primitives



D. Marr

Lessons we have learned:
(1) perfect feature extraction?

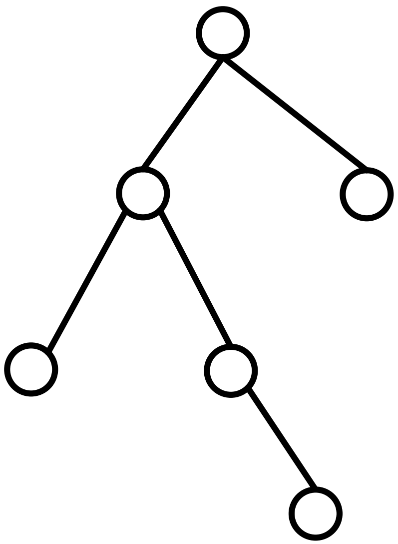
Canny edges



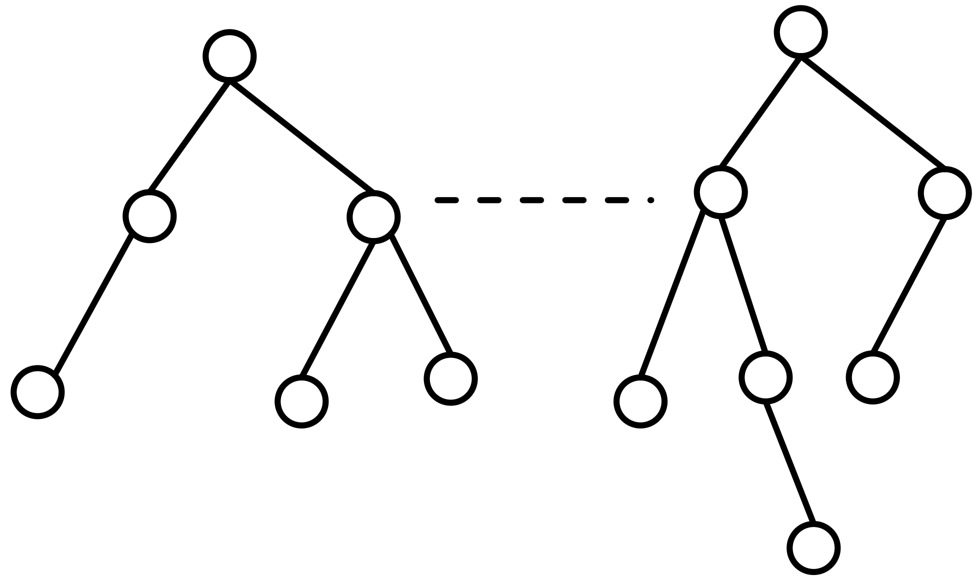
SIFT

Lessons we have learned:

(2) single decision?

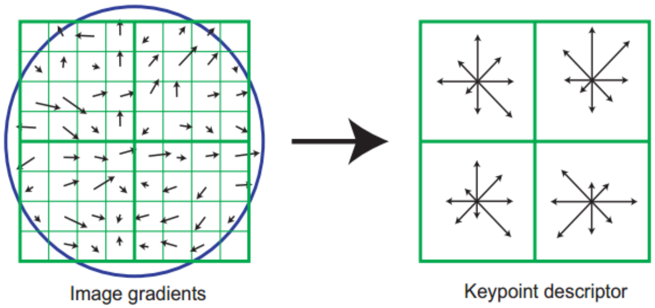


decision tree

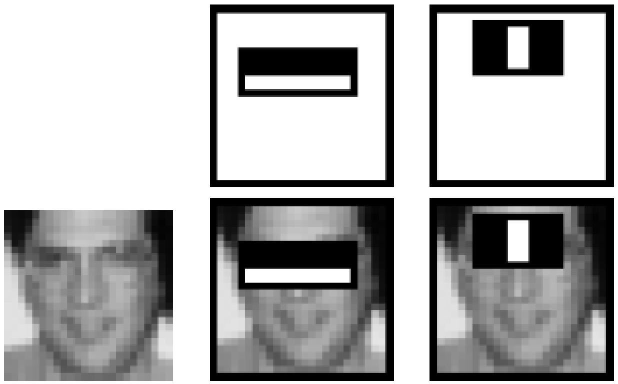


random forests

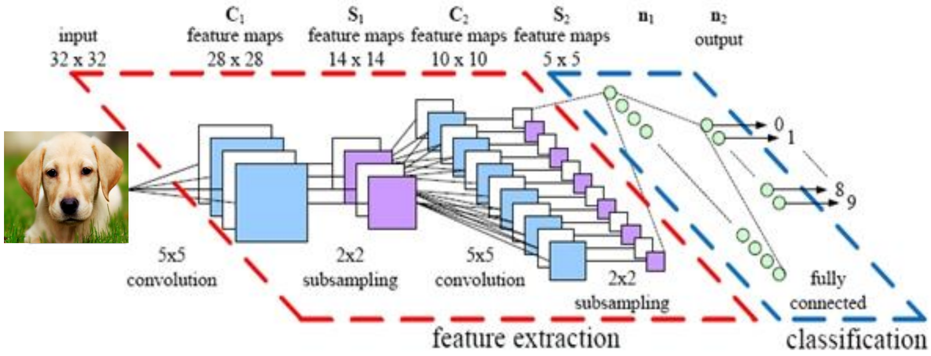
Lessons we have learned:
(3) features?

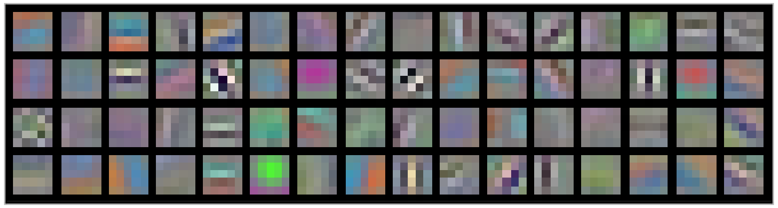


Smart human design:
SIFT descriptor (D. Lowe)



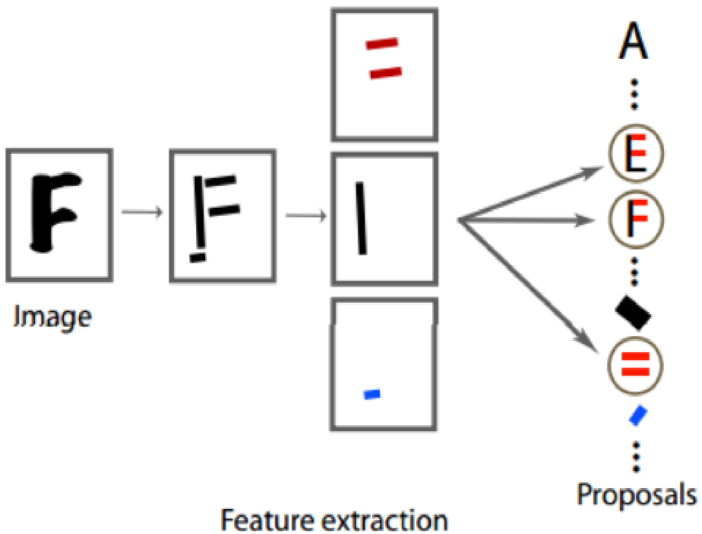
Viola and Jones

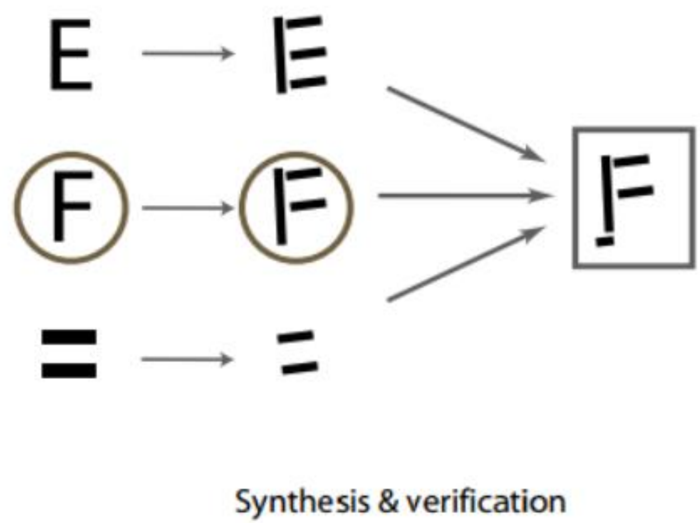




Features learned from raw data, CNN, LeCun et al.

Lessons we have learned:
(4) bottom-up and top-down?





bottom-up

top-down

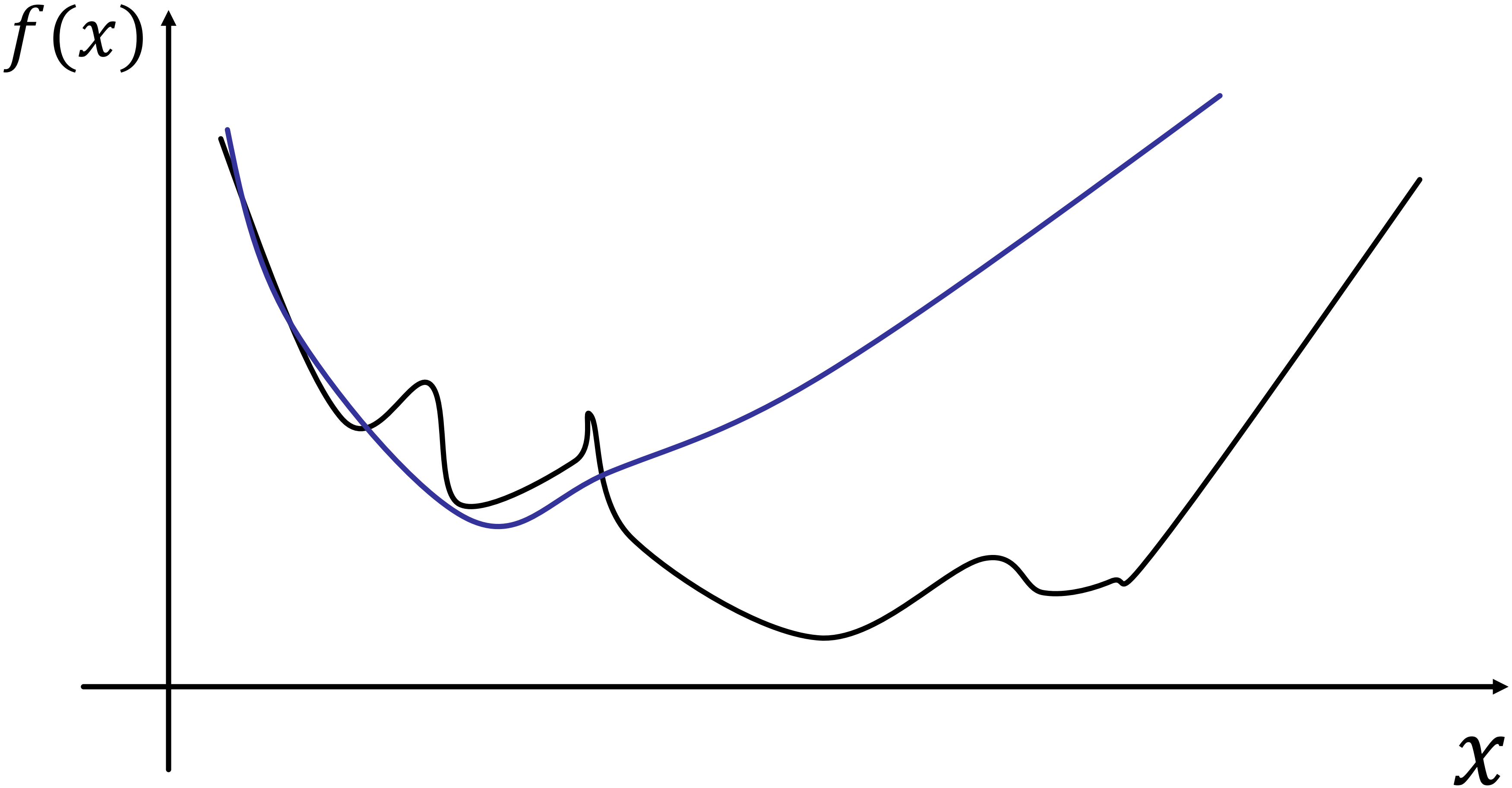
Yuille and Kersten

Lessons we have learned:
(5) bottom-up and top-down?

	Sparse Vector	*Low-Rank Matrix*
Degeneracy of	**individual signal**	**correlated signals**
Measure	**L_0 norm** $\|x\|_0$	$\mathrm{rank}(X)$
Convex Surrogate	**L_1 norm** $\|x\|_1$	**Nuclear norm** $\|X\|_*$
Compressed Sensing	$y = Ax$	$Y = A(X)$
Error Correction	$y = Ax + e$	$Y = A(X) + E$
Domain Transform	$y \circ \tau = Ax + e$	$Y \circ \tau = A(X) + E$
Mixed Structures	$Y = A(X) + B(E) + Z$	

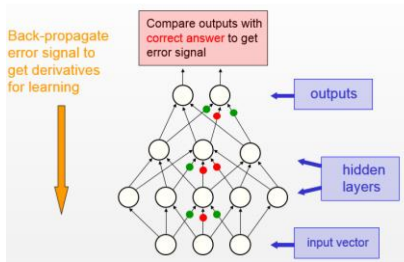
Ma and Wright

Lessons we have learned:
(6) convex vs. non-convex?



Some general notes about discriminative and generative models

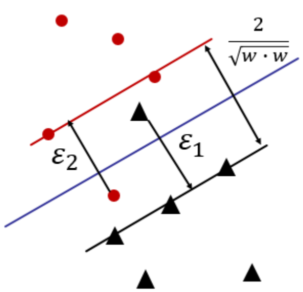
Neural networks, SVM, and Boosting



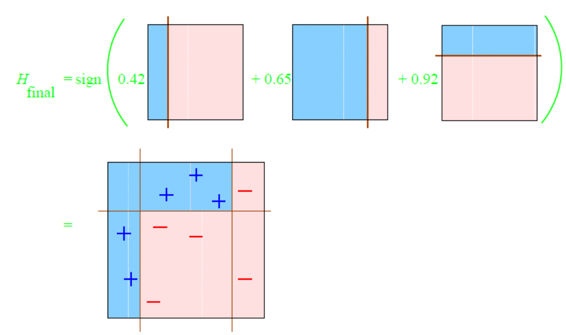
F. Rosenblatt
J. Hopfield
G. Hinton

Neural Networks
SVM
Boosting
Deep Learning

- Explicit formulation
- Explicit loss function
- Convex optimization
- Back-propagation
- Hierarchical representation
- Greedy and well-behaved
- Linear and Kernel-based
- Efficiency in testing
- Robustness
- Large-scale computing



V. Vapnik



Y. Freund and R. Schapire

Empirical comparisons of different algorithms

Caruana and Niculesu-Mizil, ICML 2006

MODEL	1ST	2ND	3RD	4TH	5TH	6TH	7TH	8TH	9TH	10TH
BST-DT	0.580	0.228	0.160	0.023	0.009	0.000	0.000	0.000	0.000	0.000
RF	0.390	0.525	0.084	0.001	0.000	0.000	0.000	0.000	0.000	0.000
BAG-DT	0.030	0.232	0.571	0.150	0.017	0.000	0.000	0.000	0.000	0.000
SVM	0.000	0.008	0.148	0.574	0.240	0.029	0.001	0.000	0.000	0.000
ANN	0.000	0.007	0.035	0.230	0.606	0.122	0.000	0.000	0.000	0.000
KNN	0.000	0.000	0.000	0.009	0.114	0.592	0.245	0.038	0.002	0.000
BST-STMP	0.000	0.000	0.002	0.013	0.014	0.257	0.710	0.004	0.000	0.000
DT	0.000	0.000	0.000	0.000	0.000	0.000	0.004	0.616	0.291	0.089
LOGREG	0.000	0.000	0.000	0.000	0.000	0.000	0.040	0.312	0.423	0.225
NB	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.030	0.284	0.686

Overall rank by mean performance across problems and metrics (based on bootstrap analysis).

BST-DT: boosting with decision tree weak classifier

RF: random forest

BAG-DT: bagging with decision tree weak classifier

SVM: support vector machine

ANN: neural nets

KNN: k nearest neighboorhood

BST-STMP: boosting with decision stump weak classifier

DT: decision tree

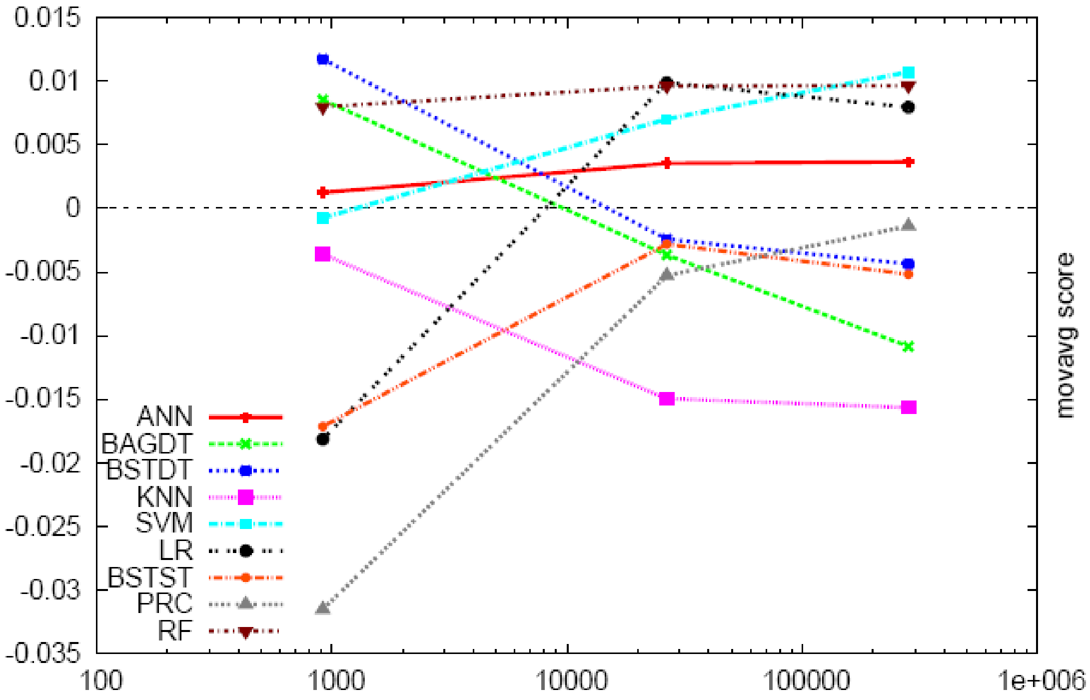
LOGREG: logistic regression

NB: naïve Bayesian

It is informative, but by no means final.

Empirical study on high-dimension

Caruana et al., ICML 2008



Moving average standardized scores of each learning algorithm as a function of the dimension.

The rank for the algorithms to perform consistently well:

(1) random forest (2) neural nets (3) boosted tree (4) SVMs

Some literature

Discriminative Approaches:

Perceptron and Neural networks (Rosenblatt 1958, Windrow and Hoff 1960, Hopfiled 1982, Rumelhart and McClelland 1986)

Nearest neighborhood classifier (Hart 1968)

Fisher linear discriminant (Fisher)

Support Vector Machine (Vapnik 1995)

AdaBoost and its variants (Freund and Schapire 1995, Friedman et al. 1998, Breiman 1994)

…

Generative Approaches:

PCA, TCA, ICA (Karhunen and Loeve 1947, H´ erault et al. 1980, Frey and Jojic 1999)

MRFs, Particle Filtering (Ising, Geman and Geman 1994, Isard and Blake 1996)

Maximum Entropy Model (Della Pietra et al. 1997, Zhu et al. 1997, Hinton 2002)

DBN (Hinton 2006)….

Max entropy principle and boosting

$$p_\lambda(I|y) = \frac{1}{\sum_I exp\{-\sum_{j=1}^{T} \lambda_j h_j(I)\}} exp\{-\sum_{j=1}^{T} \lambda_j h_j(I)\}$$

Della Pietra et al. 997,
Zhu, Wu, and Mumford 1997
Hinton 2002

$$p_\lambda(y|I) = \frac{1}{\sum_y exp\{\sum_{j=1}^{T} \lambda_j f_j(I,y)\}} exp\{\sum_{j=1}^{T} \lambda_j f_j(I,y)\}$$

Freund and Schapire 1995,
Friedman et al. 1998

- Both have the feature selection procedure.
- Both follow a exponential probabilistic model (arguable).

▪Although generative model is always preferred, if we can, we are forced to use discriminative models in many cases.

From discriminative to generative (Tu 2008)

We are given a set of training samples (positive), S, and we want to learn a corresponding generative model. We can turn a single class learning problem into a two-class learning problem. Let x be a data vector and $y \in \{-1, +1\}$ its label.

Bayes rule:

$$p(y=+1|x) = \frac{p(x|y=+1)p(y=+1)}{p(x|y=-1)p(y=-1) + p(x|y=+1)p(y=+1)}$$

$$\longrightarrow \quad p(x|y=+1) = \frac{p(y=+1|x)p(y=-1)}{p(y=-1|x)p(y=+1)} p(x|y=-1)$$

Drop p(y) for simplicity:

$$p(x|y=+1) = \frac{p(y=+1|x)}{p(y=-1|x)} p(x|y=-1)$$

The above equation says that a generative model for the positives p(x|y=+1) can be obtained from the discriminative model p(y|x) and a generative model p(x|y=-1) for the negatives.

From discriminative to generative

Instead, we learn the model recursively.

$$p(x|y=+1) = \frac{p(y=+1|x)}{p(y=-1|x)} p_1^r(x)$$

$$q_1(x) \sim p(y|x)$$

$$p_2^r(x) = \frac{1}{Z_1} \frac{q_1(y=+1|x)}{q_1(y=-1|x)} p_1^r(x)$$

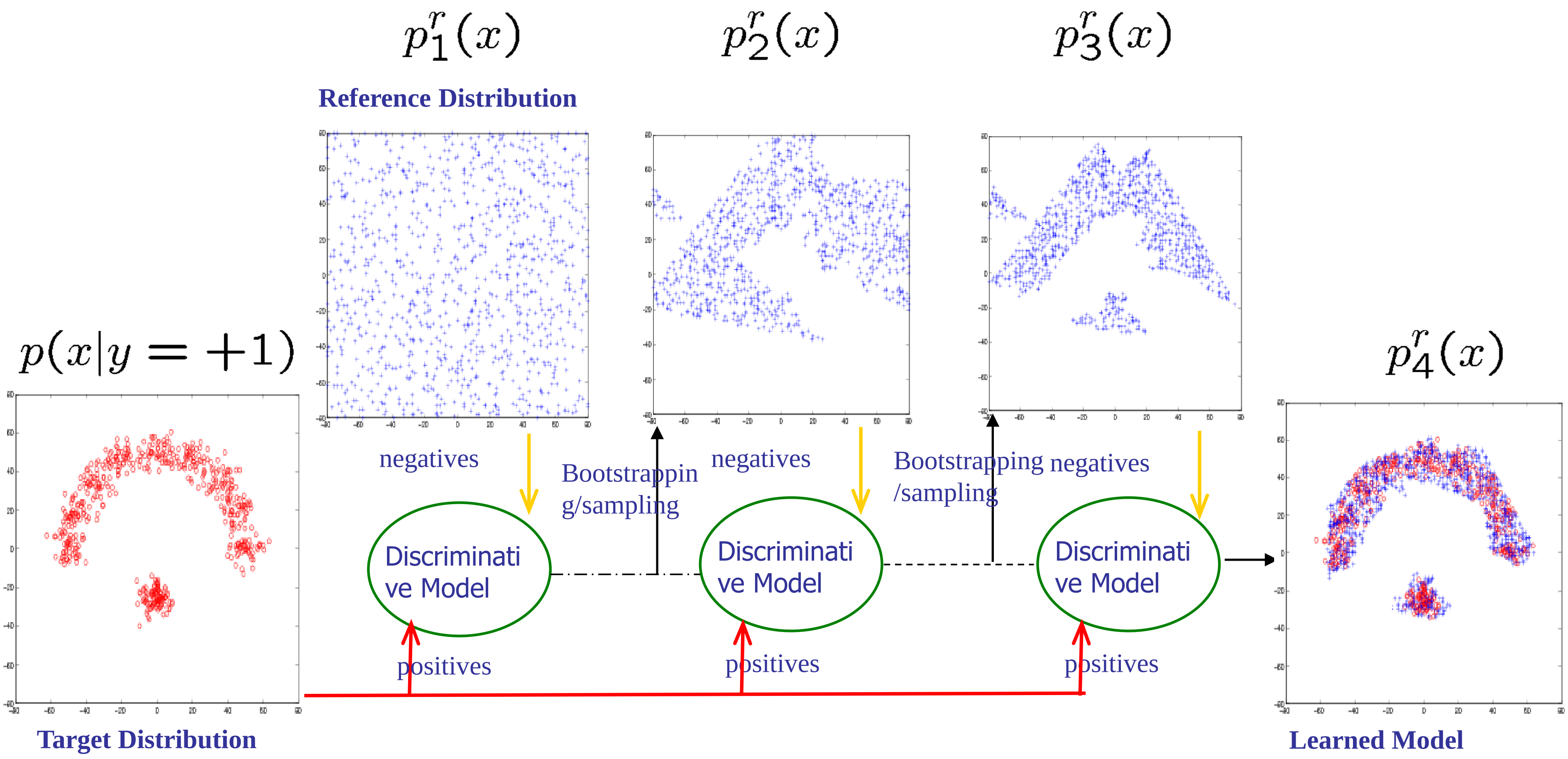
$$q_k(x)$$

$$p_{n+1}^r(x) = \prod_{k=1}^{n} \frac{1}{Z_k} \frac{q_k(y=+1|x)}{q_k(y=-1|x)} p_1^r(x)$$

Goal: $p_{n+1}^r(x) \rightarrow p(x|y=+1)$

A toy example

$$p^r_{n+1}(x) = \prod_{k=1}^{n} \frac{1}{Z_k} \frac{q_k(y=+1|x)}{q_k(y=-1|x)} p^r_1(x)$$



From discriminative to generative

$$p^r_{n+1}(x) = \prod_{k=1}^{n} \frac{1}{Z_k} \frac{q_k(y=+1|x)}{q_k(y=-1|x)} p^r_1(x)$$

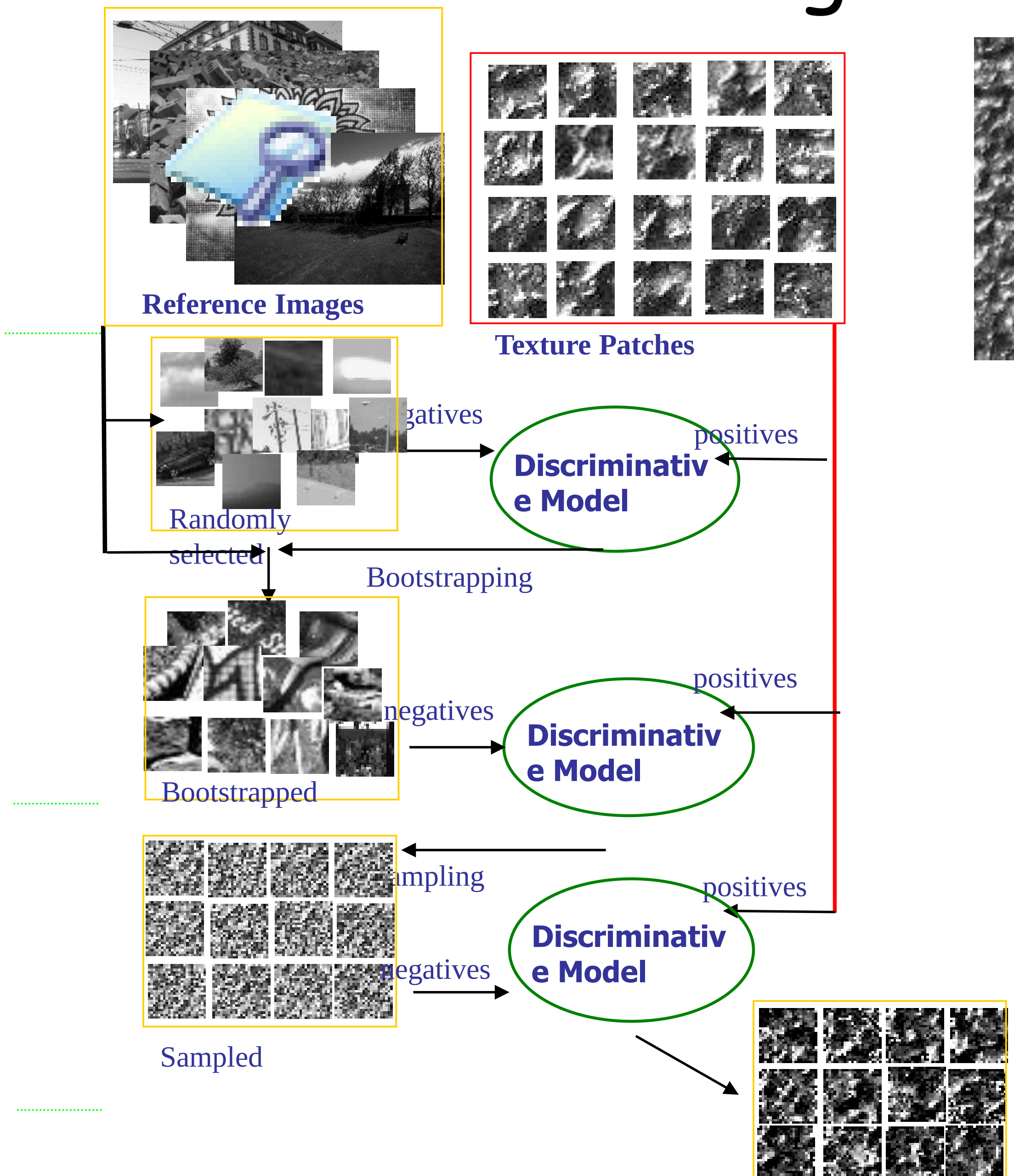
Theory: $p^r_{n+1}(x)$ asymptotically approaches p(x|y=+1). We write $p^+(x) = p(x|y=+1)$

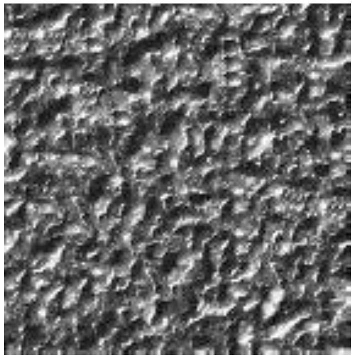
$$KL[p^+(x)||p^r_{n+1}(x)] \leq KL[p^+(x)||p^r_n(x)]$$

Proof:

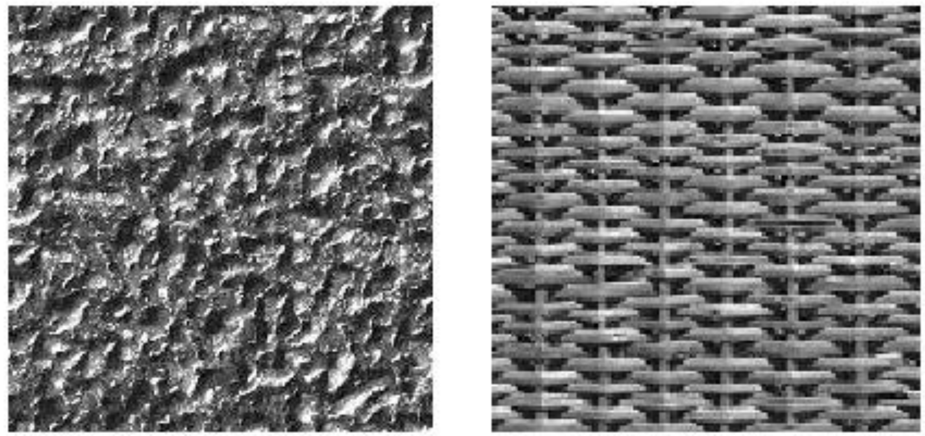
$$\begin{aligned}
& KL[p^+(x)||p^r_n(x)] - KL[p^+(x)||p^r_{n+1}(x)] \\
= & \int p^+(x) \log \left(\frac{1}{Z_n} \frac{q(y=+1|x)}{q(y=-1|x)} p^r_n(x) \right) dx - \int p^+(x) \log[p^r_n(x)] dx \\
= & \int p^+(x) \log \frac{1}{Z_n} dx + \int p^+(x) \log \frac{q(y=+1|x)}{q(y=-1|x)} dx \\
= & \log \frac{1}{Z_n} + \int p^+(x) \log \frac{q(y=+1|x)}{q(y=-1|x)} dx \geq 0
\end{aligned}$$

Texture modeling

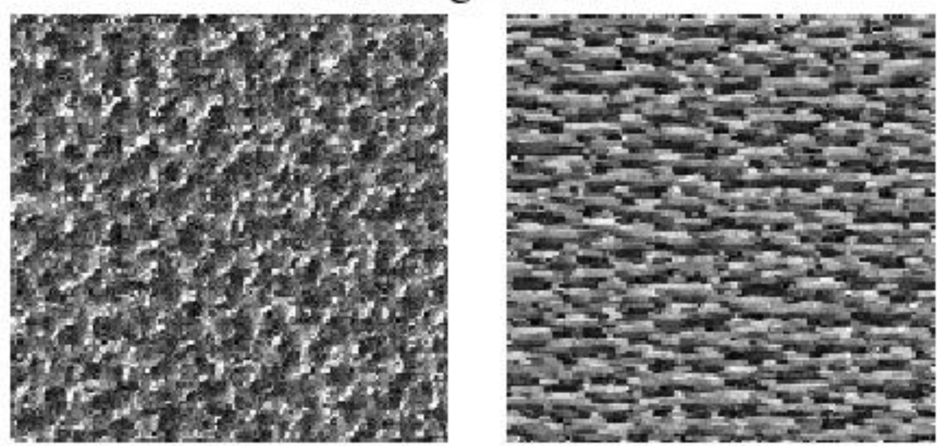




Texture modeling



Training textures

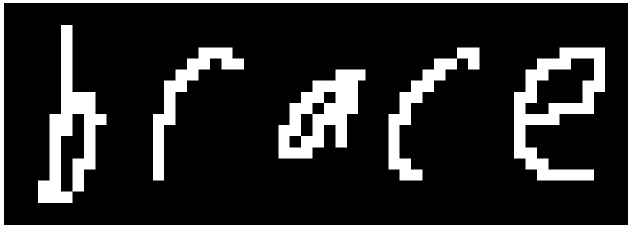


Synthesized textures

Importance of structural information.

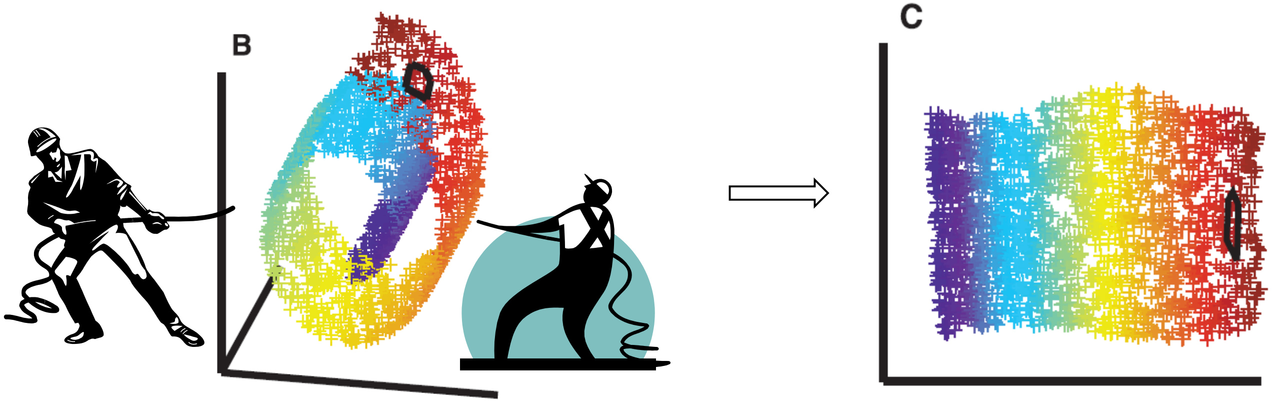
Importance of structured data

- Structured information within data sample.



OCR

- Structured information in-between data samples.



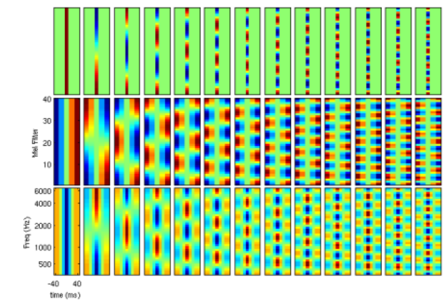
ISOMAP (Tenenbaum et al.), LLE (Roweis and Saul)

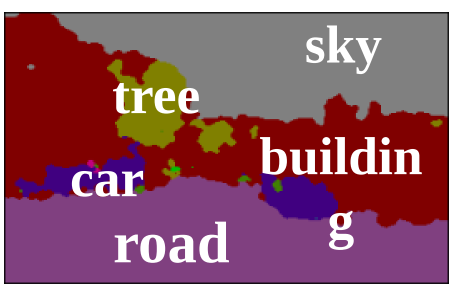
Structural prediction-overview

15 emotion categories (Anger, Abuse, Blame, Fear, Forgiveness, Guilt, Happiness_peacefulness, Hopefulness, Hopelessness, Love, Pride, Sorrow, and Thankfulness)

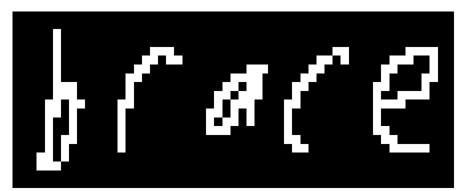
Hopelessness/Sorrow/Fear

John : I am going to tell you this at the last . You and John and Mother are what I am thinking - I ca n't go on - my life is ruined. I am ill and heart - broken . Always I have felt alone and never more alone than now . John . Please God forgive me for all my wrong doing . I am lost and frightened . God help me , Bless my son and my mother .





Depth data, Shotton et al.



OCR

Structural information

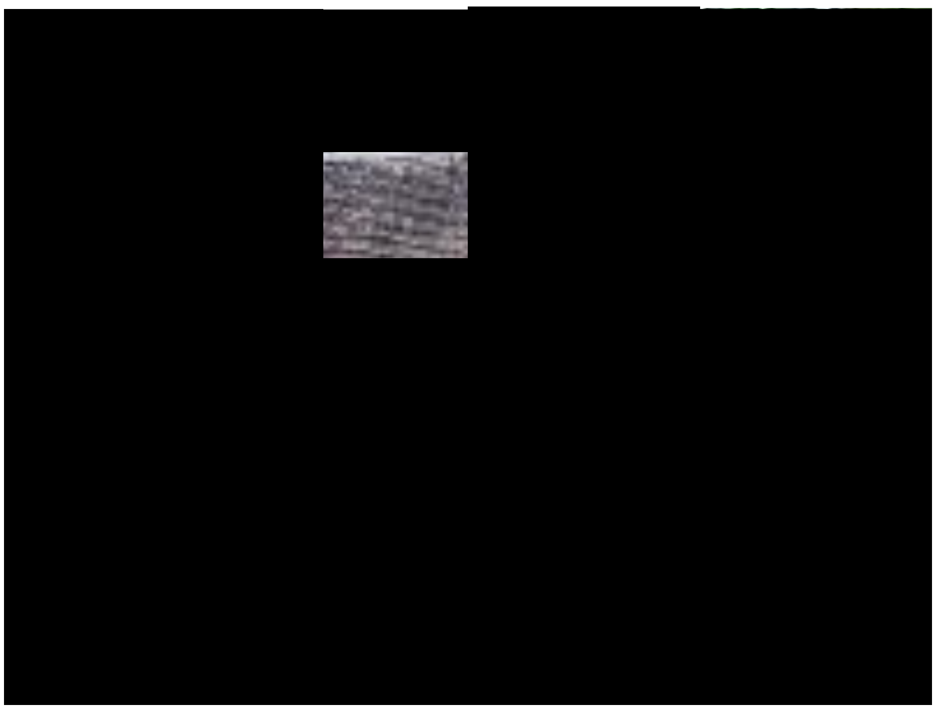


Image from PASCAL

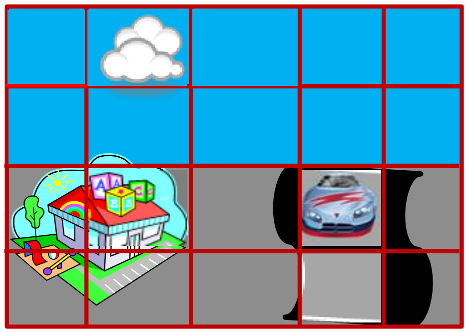
Structural prediction literature

- Hidden Markov Models (Markov 1922, Baum and Petrie 1966,..)
- Bayesian Network (Peral 1986,…)
- Neural Networks (Rosenblatt 1958, Werbos 1975, Hinton 2006)
- Markov Random Fields (Ising 1924, Geman and Geman 1984, …)
- Structural Support Vector Machine (Vapnik 1992, Tsochantaridis et al. 2005,…)
- Conditional Random Fields (Lafferty et al. 2001,…)

Graphical models…

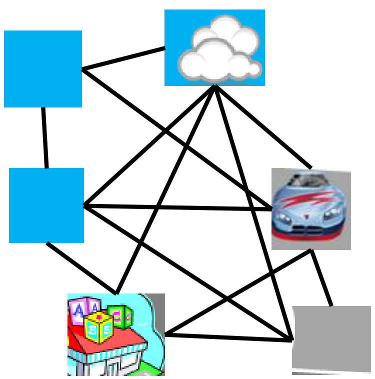
Typical inference methods include Belief/message propagation, MCMC (Gibbs sampling, Metropolis-Hasting), EM, Graph Cuts, Stochastic descent...

Problem formulation



$$X = \{ \qquad \}$$

$$Y = \{sky, cloud, sky, building, car, road\}$$



$$p(Y|X) \propto p(Y)p(X|Y)$$

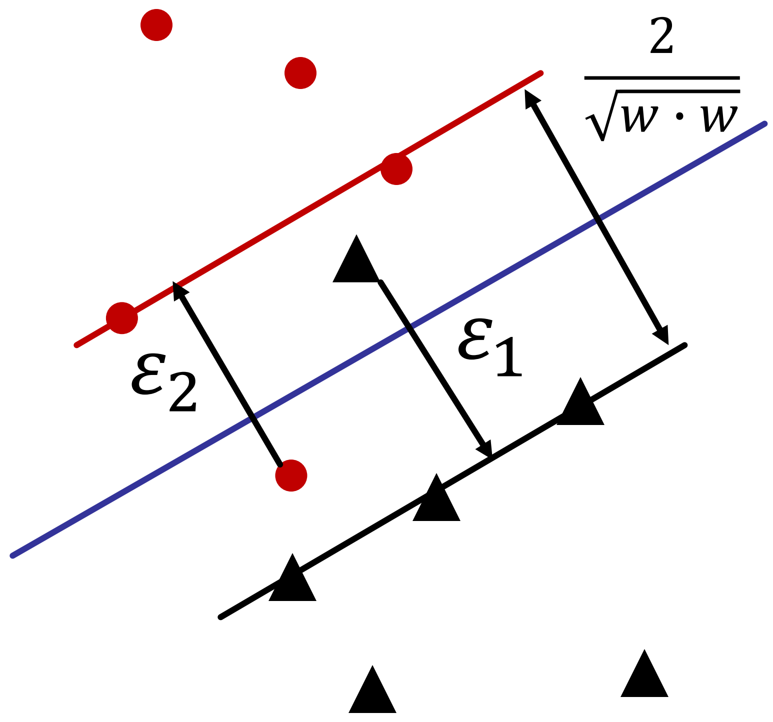
$$\rightarrow \prod_{(i,j)\in N} p(y_i, y_j) \prod p(x_i|y_i)$$

MRF

$$p(Y|X) \rightarrow \prod_{(i,j)\in N} p(y_i, y_j|x_i, x_j) \prod p(y_i|x_i)$$

CRF

Binary SVM (V. Vapnik)



$$\min ||w||^2 + C \sum \varepsilon_i$$

$$s.t.\ y_i(w \cdot x_i + b) \geq 1 - \varepsilon_i$$

Multi-Class SVM (Crammer & Singer, 2001)

- Training Examples: $(\vec{x}_1, y_1), \dots, (\vec{x}_n, y_n) \;\; \vec{x} \in \Re^N \;\; y \in \{1, \dots, k\}$
- Inference:
- Training: Find $\underset{<\vec{w}_1, \dots, \vec{w}_k>}{h(\vec{x}) = argmax_{i \in \{1,\dots,k\}} \left[\vec{w}_i^T \vec{x}\right]}$ that solve

$$\min_{\vec{w}_1,\dots,\vec{w}_n,\vec{\xi}} \quad \sum_{i=1}^{k} \vec{w}_i^2 + \frac{C}{n} \sum_{i=1}^{n} \xi_i$$

$$\begin{aligned} s.t. \quad & \forall j \neq y_1 : \vec{w}_{y_1}^T \vec{x}_1 \geq \vec{w}_j^T \vec{x}_1 + 1 - \xi_1 \\ & \dots \\ & \forall j \neq y_n : \vec{w}_{y_n}^T \vec{x}_n \geq \vec{w}_j^T \vec{x}_n + 1 - \xi_n \end{aligned}$$

Structured SVM (Tsochantaridis et al.)

- Formulation

$$\min_{\vec{w}} \quad \frac{1}{2}\vec{w}^T\vec{w}$$

$$s.t. \quad \forall y \in Y \setminus y_1 : \vec{w}^T\Phi(x_1, y_1) \geq \vec{w}^T\Phi(x_1, y) + 1$$

$$\dots$$

$$\forall y \in Y \setminus y_n : \vec{w}^T\Phi(x_n, y_n) \geq \vec{w}^T\Phi(x_n, y) + 1$$

Achieve:

argmax$_{\textbf{word}}$ **w**^{T} **f(****,word)** = **"brace"**

Such that:

w^{T} **f(****,"brace")** > **w**^{T} **f(****,"aaaaa")**

w^{T} **f(****,"brace")** > **w**^{T} **f(****,"aaaab")**

...

w^{T} **f(****,"brace")** > **w**^{T} **f(****,"zzzzz")**

B. Taskar

A unified view of binary, multi-class, and structured SVM

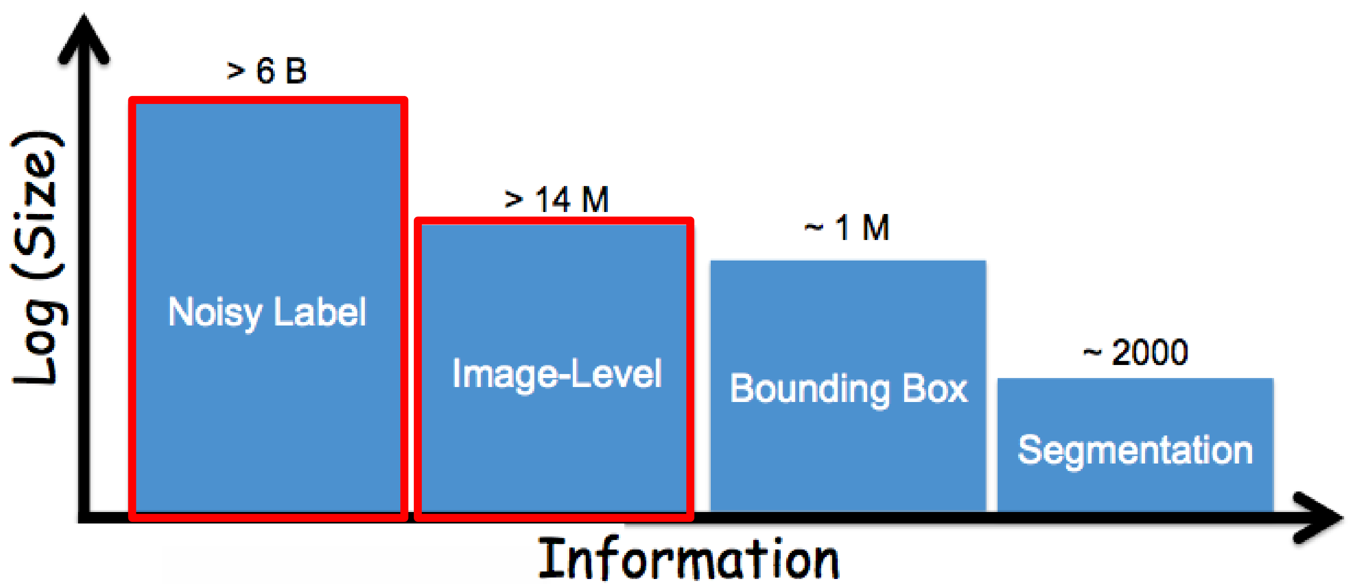
	Binary	Multi-class	Structured
Specific	$S = \{(X_m, Y_m), m = 1..M\}$ $X \in \mathcal{R}^L$ $Y \in \{-1, +1\}$ Compute feature (explicitly or implicitly through kernels) $\Phi(X)$ $Y = \begin{cases} +1, if\ W \cdot \Phi(X) \geq 0 \\ -1, \quad otherwise \end{cases}$	$S = \{(X_m, Y_m), m = 1..M\}$ $X \in \mathcal{R}^L$ $Y \in \{1,..,k\}$ Compute feature (explicitly or implicitly through kernels) $\Phi(X)$ $Y^* = arg \max_{Y \in \{1,..,k\}} W_Y \cdot \Phi(X)$	$X = (x_1, .., x_n), x_i \in \mathcal{R}^L$ $Y = (y_1, .., y_n), y_i \in \{1..k\}$ Compute feature (explicitly or implicitly through kernels) $\Phi(X, Y)$ $Y^* = arg \max_{Y} W \cdot \Phi(X, Y)$
Unified	$Y^* = arg \max_{Y \in \{-1,+1\}} YW \cdot \Phi(X)$	$\Phi(X, Y) = (\Phi(X) \cdot \delta(1 = Y), .., \Phi(X) \cdot \delta(k = Y))$ $Y^* = arg \max_{Y \in \{1,..,k\}} W \cdot \Phi(X, Y)$	$\Phi(X, Y) = (\Phi(X) \cdot \delta(1 = Y), .., \Phi(X) \cdot \delta(k = Y), y_1, .., y_n)$ $Y^* = arg \max_{Y} W \cdot \Phi(X, Y)$

Part II:

Why weakly-supervised learning?

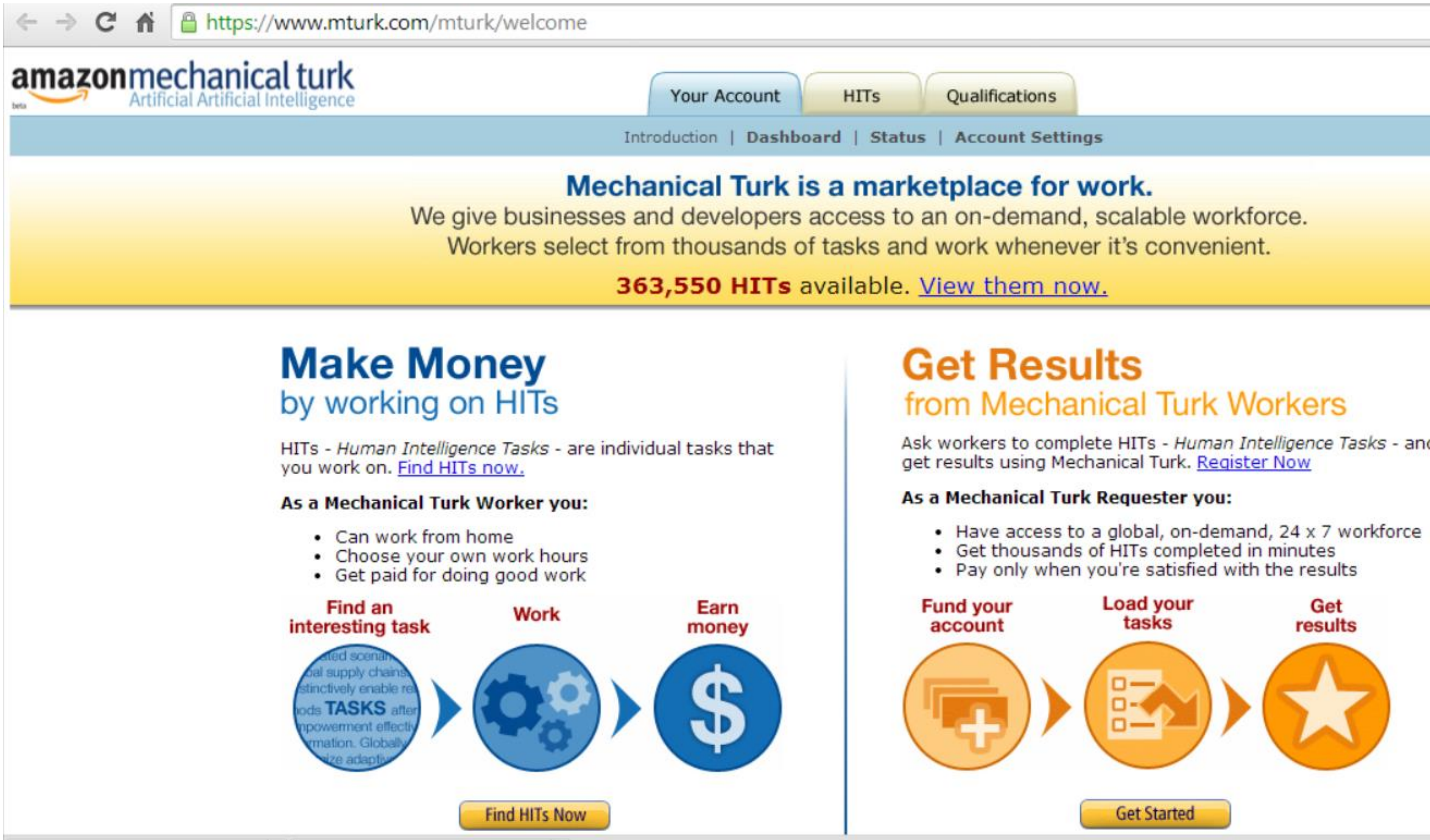
Data and supervision (images)

- Social Networking Sites (e.g. Facebook, MySpace)
- Image Search Engines (e.g. Google, Bing)
- Photo Sharing Sites (e.g. Flickr, Picasa)
- Computer Vision Datasets (e.g. LabelMe, SUN, ImageNet)

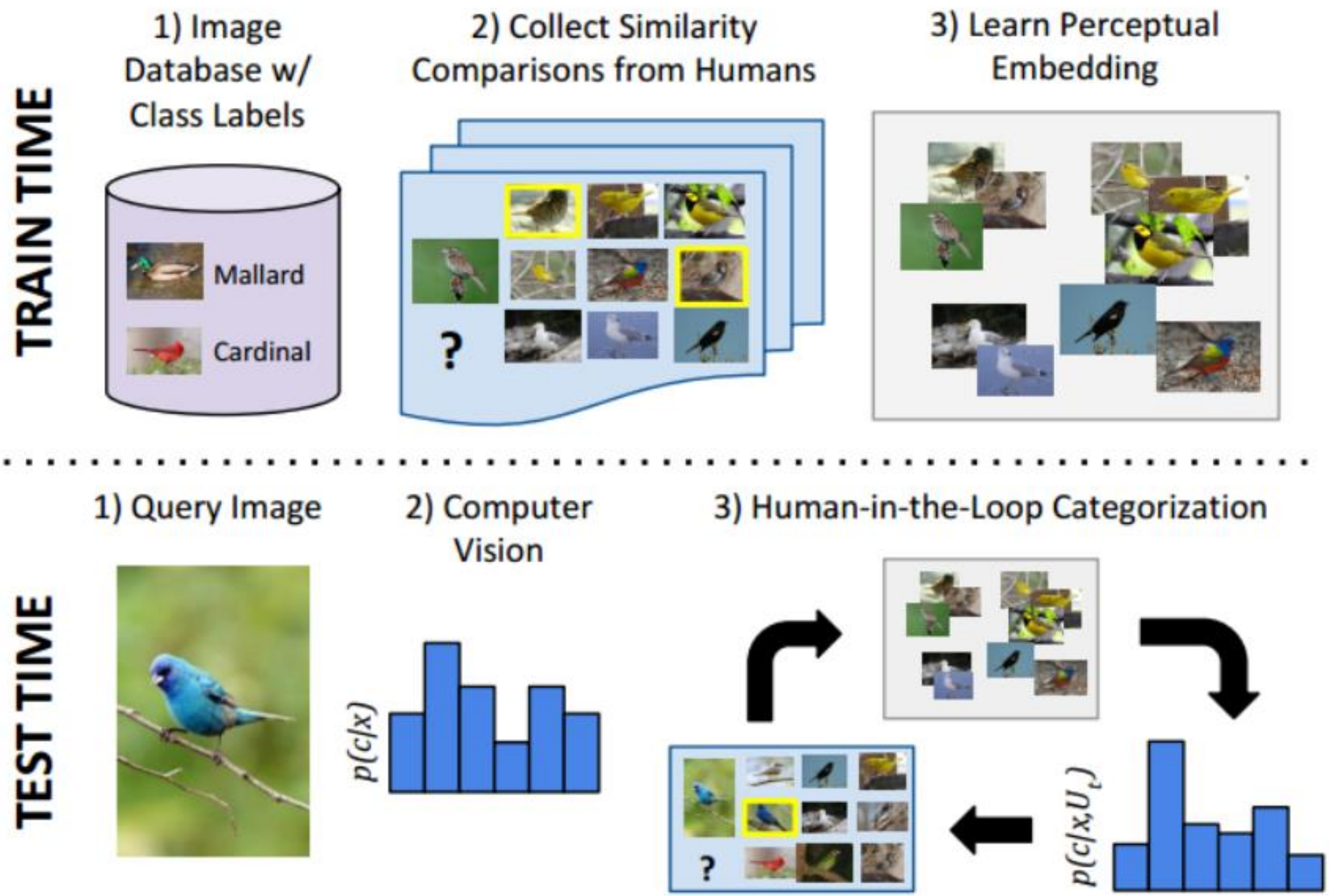


CVPR 2013 workshop “Visual Learning with Weak Supervision”

Crowdsourcing: gross labels are easier to get

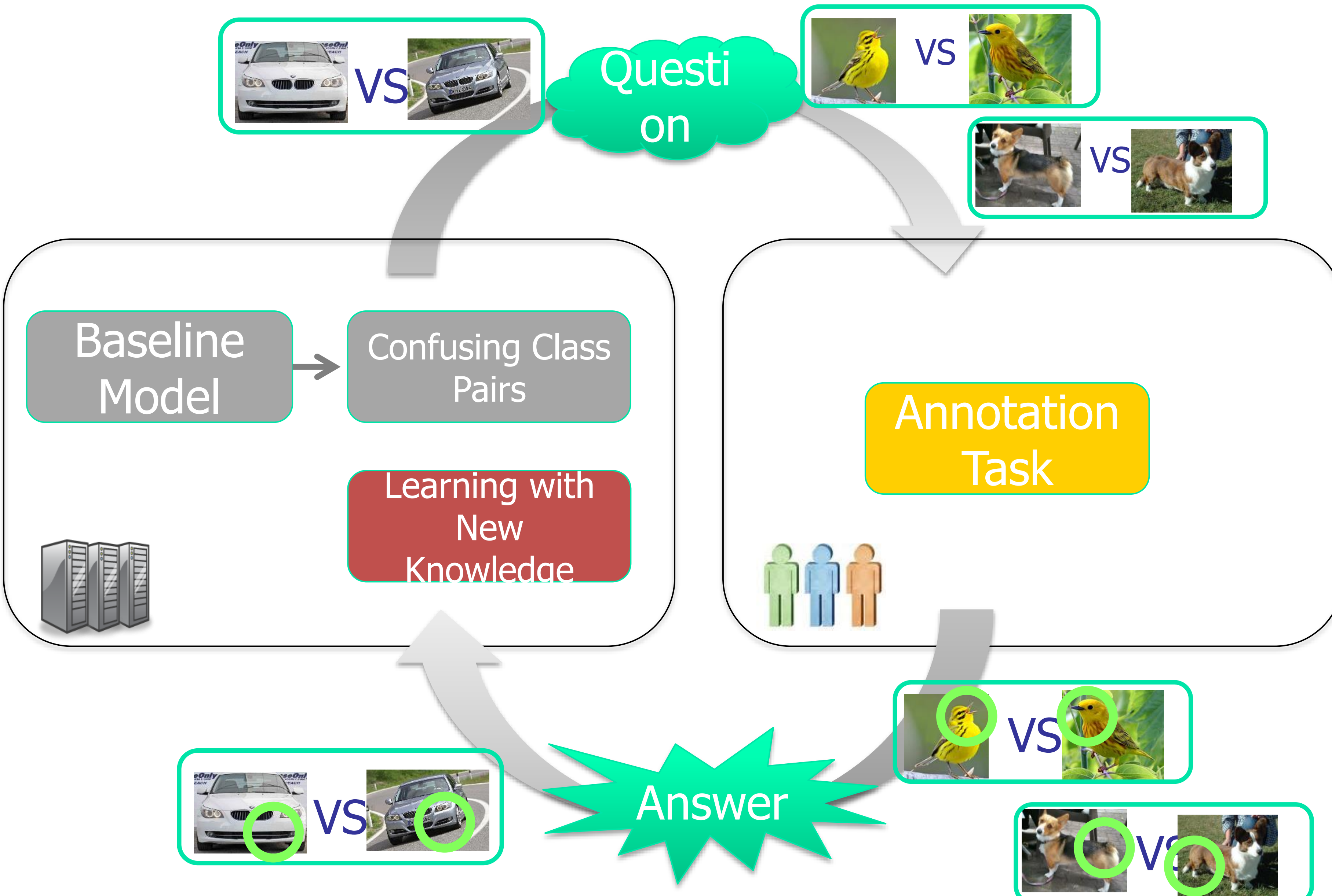


Fine-grained classification



Wah et al. cvpr 2014

Machine-crowd collaboration



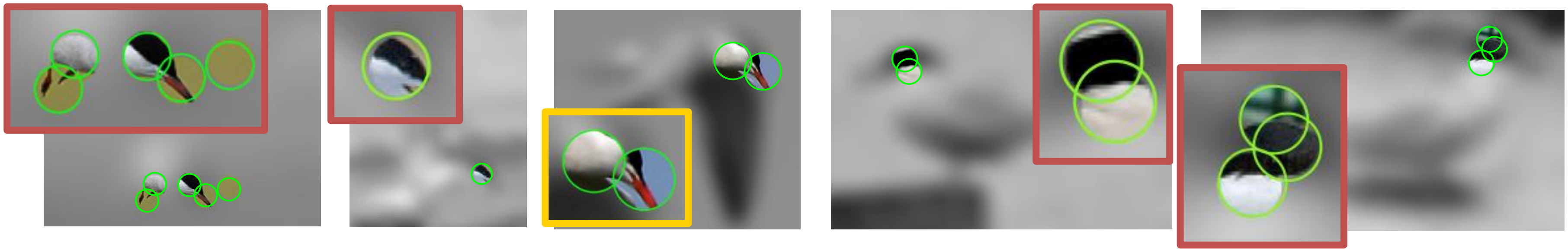
Deng, Krause, & Fei-Fei, *CVPR2013*

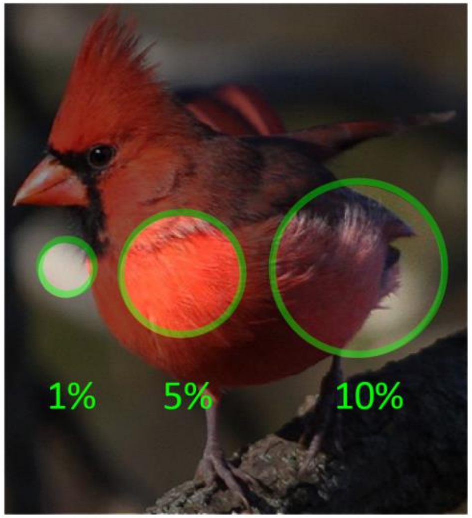
Crowd picked bubbles (AMT)

200 classes from Caltech-UCSD-Bird [Welinder et al. 2010]

800 top confusing class pairs (via cross-validation)

90K games on Amazon Mechanical Turk





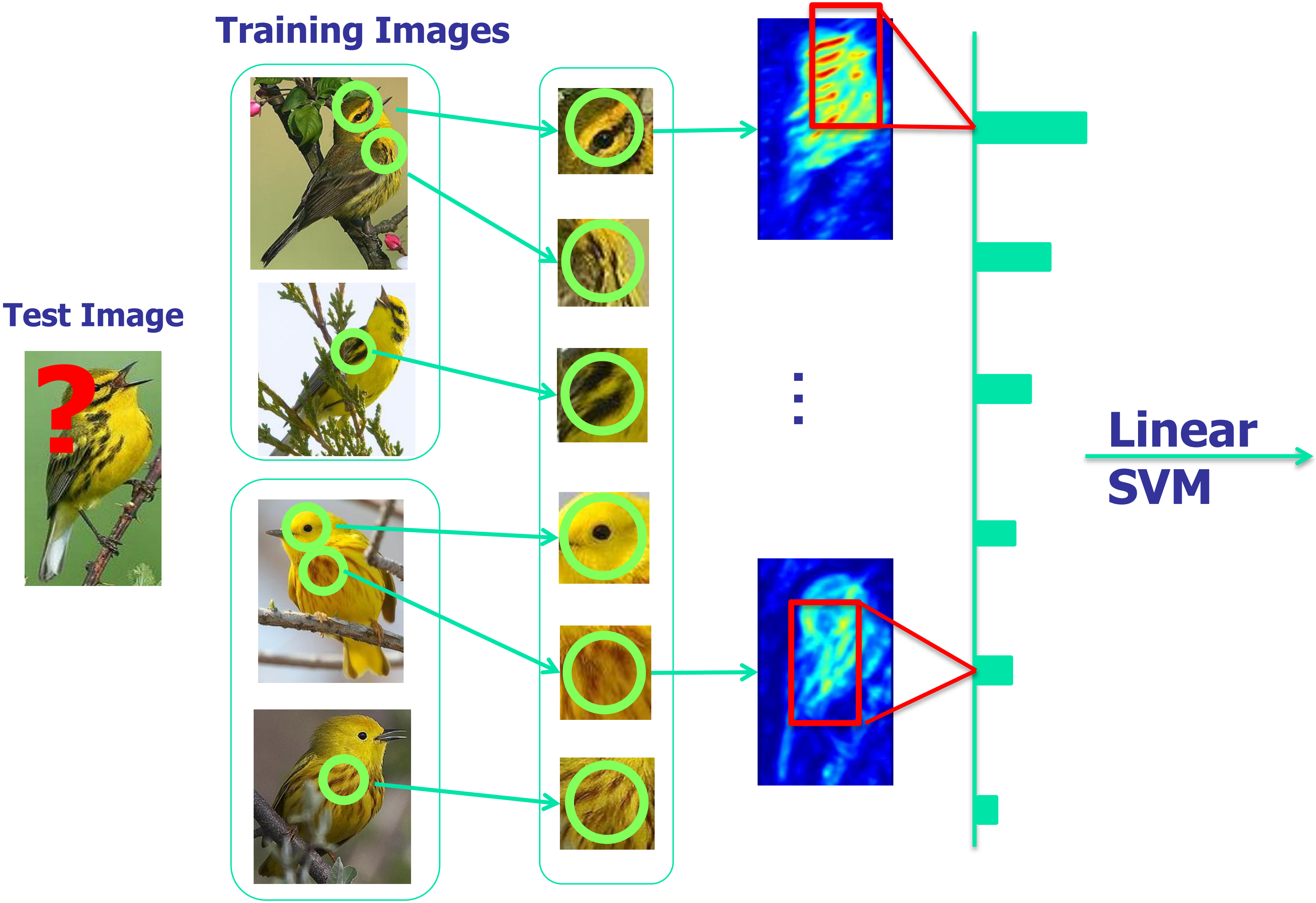
Bubble sizes as proportions of image

70% of games are successful

>90% of successful games use <10% of the bounding box

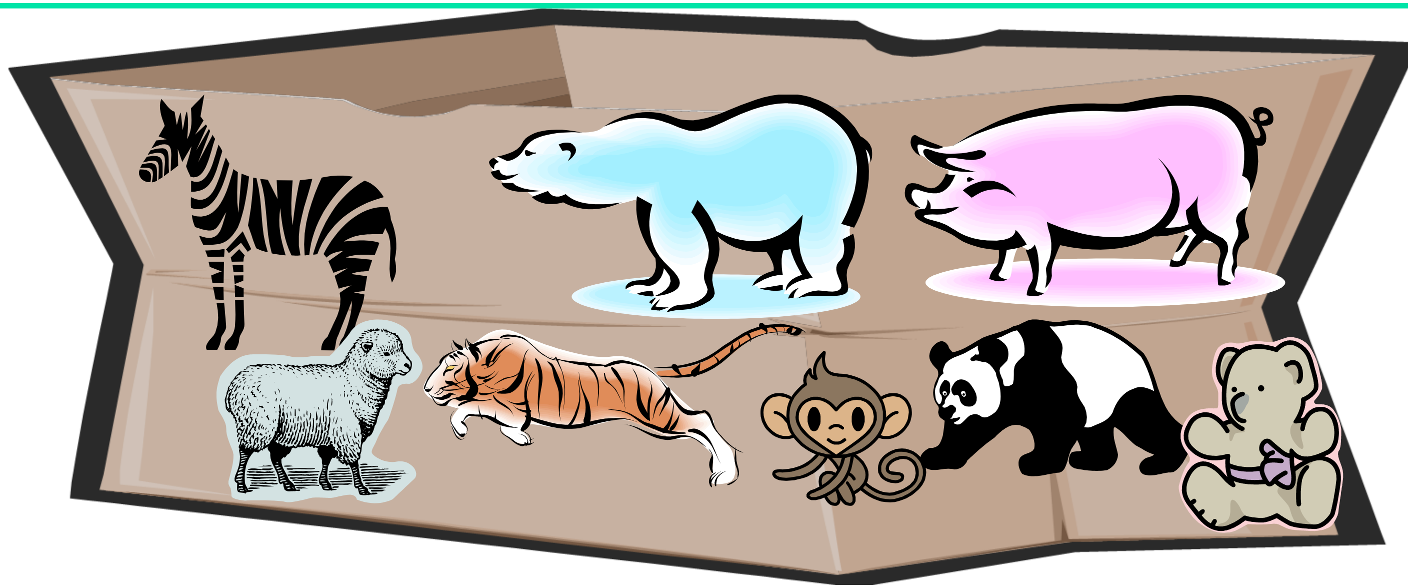
Deng, Krause, & Fei-Fei, *CVPR2013*

BubbleBank representation



Deng, Krause, & Fei-Fei, *CVPR2013*

Multiple instance learning (discriminative)



Multiple instance learning (generative)

Multiple instance learning

[Dietterich 97]

- Training data given in sets/bags [weakly supervised]
  - If all instances in set are negative, set is negative
  - Set is positive if at least 1 instance in set is positive

- Goal is to learn <u>instance</u> classifier *f*:

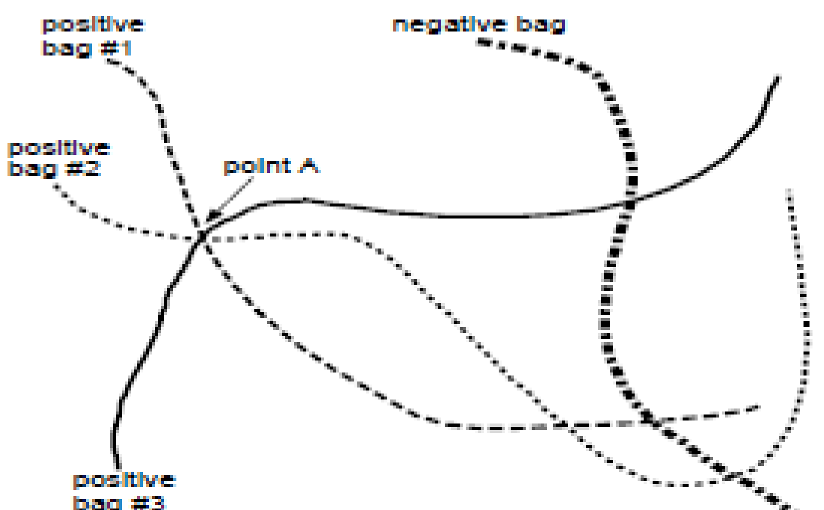
$$F(X_i) = \begin{cases} 1 & \text{if } \exists j \text{ s.t. } f(x_{ij}) = 1 \\ 0 & \text{otherwise} \end{cases}$$

- If oracle gave positive instance *j* for each positive set, could train *f* using standard supervised learning

MIL example

Drug Activity Prediction

- Molecule can take on multiple shapes
- Representation ambiguous, use MIL to find most consistent state



[Dietterich 97]

Multiple instance learning (MIL)

- Supervised Learning Training Input

$$\{x_1, \ldots, x_n\}, x_i \in \mathcal{X}$$
$$\{y_1, \ldots, y_n\}, y_i \in \mathcal{Y}$$

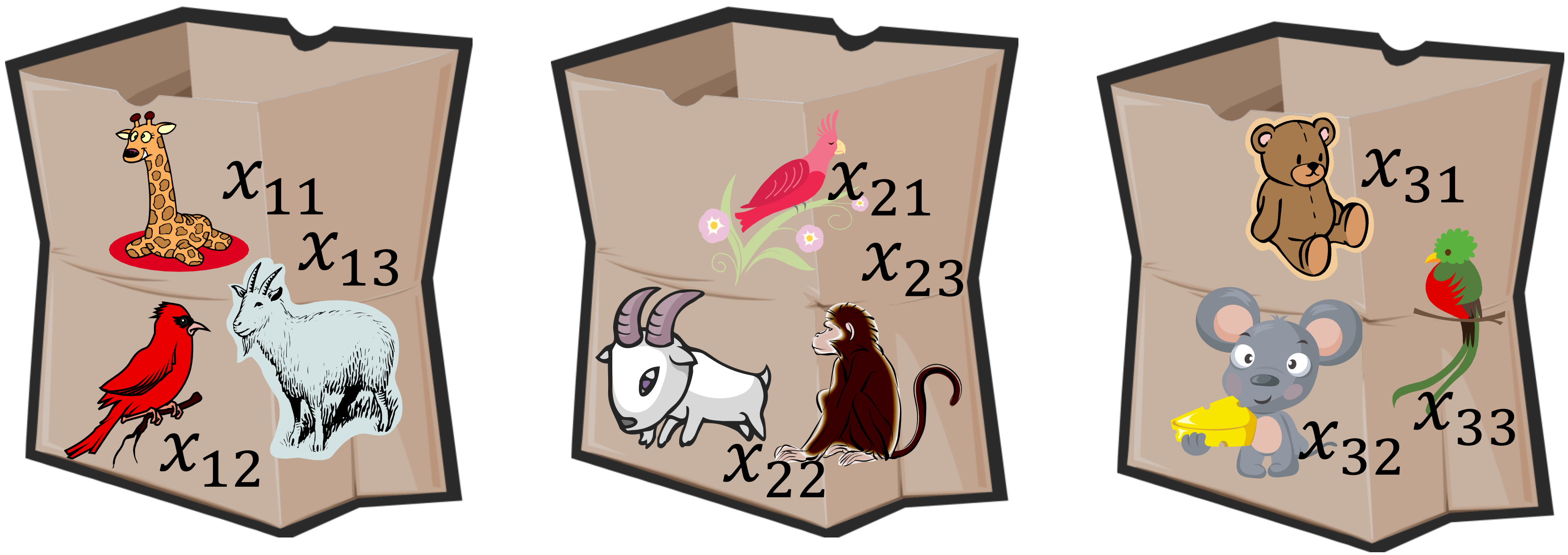
- MIL Training Input

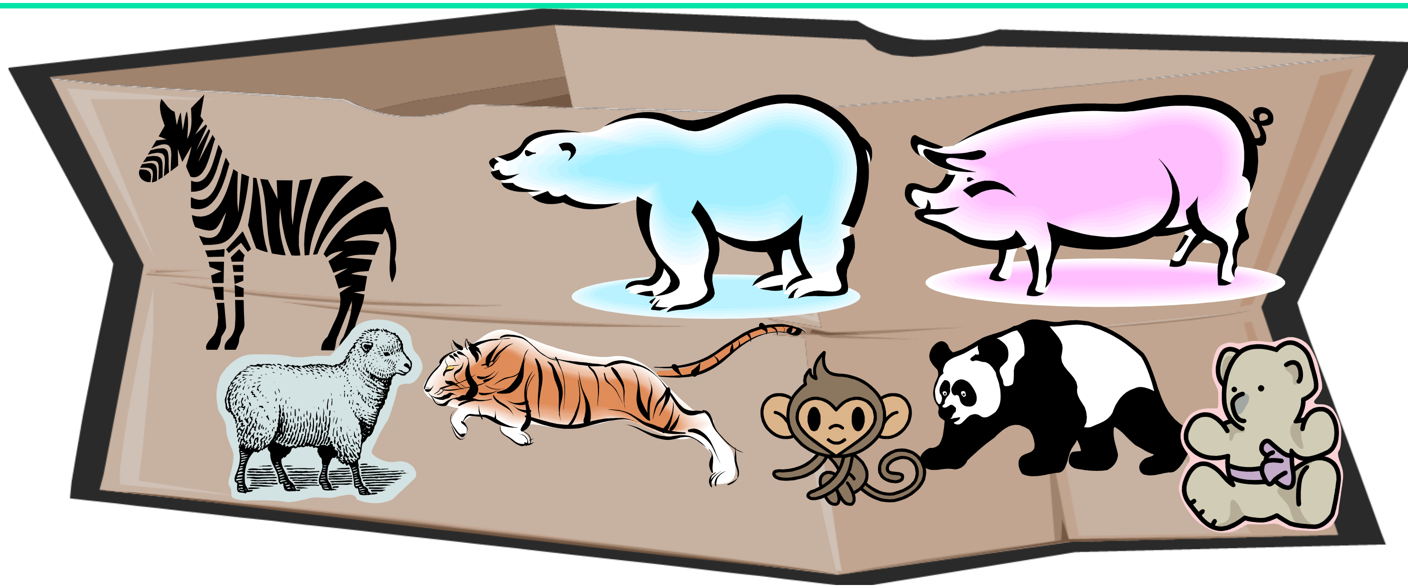
$$\{X_1, \ldots, X_n\}, X_i = \{x_{i1}, \ldots, x_{im}\}, x_{ij} \in \mathcal{X}$$
$$\{y_1, \ldots, y_n\}, y_i \in \mathcal{Y}$$

- Goal: learning instance classifier

$$h : \mathcal{X} \rightarrow \mathcal{Y}$$
$$\{h_{i1}, \ldots, h_{im}\}$$

Multiple instance learning





Bags vs. instances

$$\mathcal{L}_{MIL} = -\sum_{i=1}^{n} \left(1(y_i = 1) \log p_i + 1(y_i = -1) \log (1 - p_i) \right)$$

$$p_i = \Pr(y_i = 1 | x_i; h) = 1 - \prod_{j=1}^{m} (1 - p_{ij})$$

$$p_{ij} = \Pr(y_{ij} = 1 | x_{ij}; h) = \frac{1}{1 + \exp(-h_{ij})}$$

$$w_{ij} = -\frac{\partial \mathcal{L}_{\text{MIL}}}{\partial h_{ij}} = \begin{cases} -\dfrac{1}{1 - p_{ij}} \dfrac{\partial p_{ij}}{\partial h_{ij}} & \text{if } \; y_i = -1; \\[2ex] \dfrac{1 - p_i}{p_i(1 - p_{ij})} \dfrac{\partial p_{ij}}{\partial h_{ij}} & \text{if } \; y_i = 1. \end{cases}$$

Optimization: discriminative EM

- Perform the discriminative learning in the presence of hidden variables.

$$\frac{d}{d\theta}\mathcal{L}(Y|X;\theta) = E_{H\sim Pr(H|Y,X;\theta)}\frac{d}{d\theta}\mathcal{L}(Y,H|X;\theta)$$

E-step: Update the hidden variable (label) of each sample in positive bags.

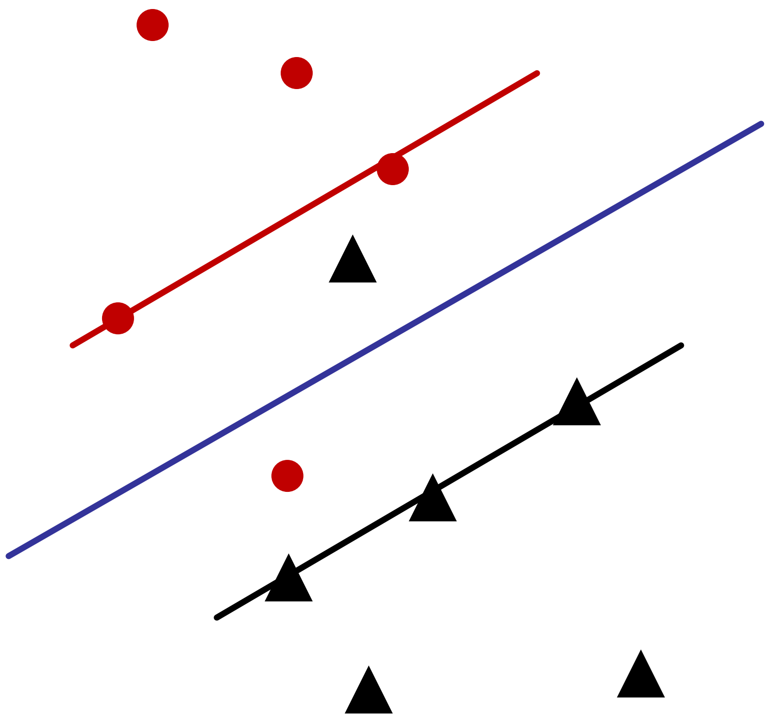
M-step: train discriminative models based on the estimated labels.

EM-DD (Zhang and Goldman, 2001)

- In the MIL setting, the label of a bag is determined by the "most positive" instance in the bag, i.e., the one with the highest probability of being positive among all the instances in that bag. The difficulty of MIL comes from the ambiguity of not knowing which instance is the most likely one.

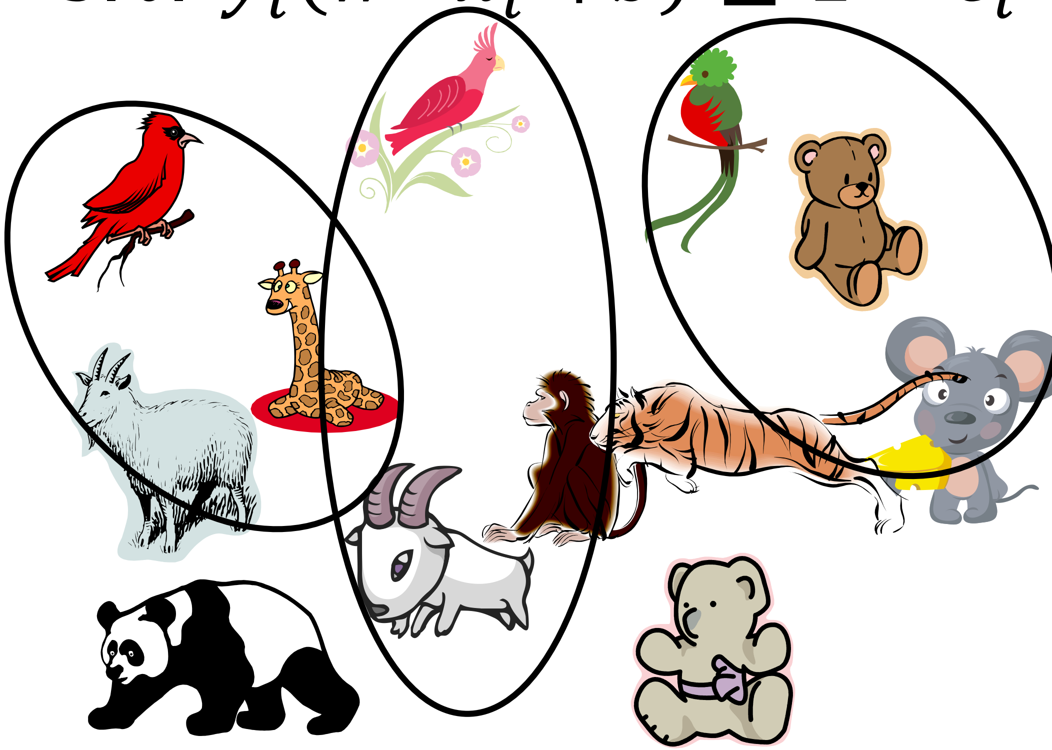
- The knowledge of which instance determines the label of the bag is modeled using a set of ***hidden variables***, which are estimated using the Expectation Maximization style approach. This results in an algorithm called EM-DD, which combines this EM-style approach with the DD algorithm.

Using SVM for MIL directly



$$\min ||w||^2 + C\sum \varepsilon_i$$

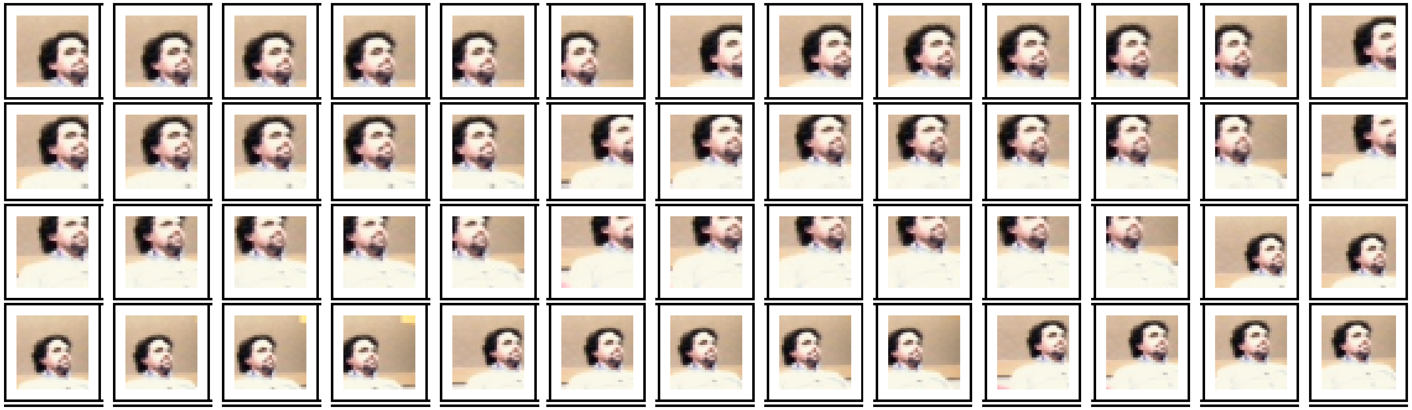
$$s.t.\ y_i(w \cdot x_i + b) \geq 1 - \varepsilon_i$$



MIL example

Object detection with weak supervision

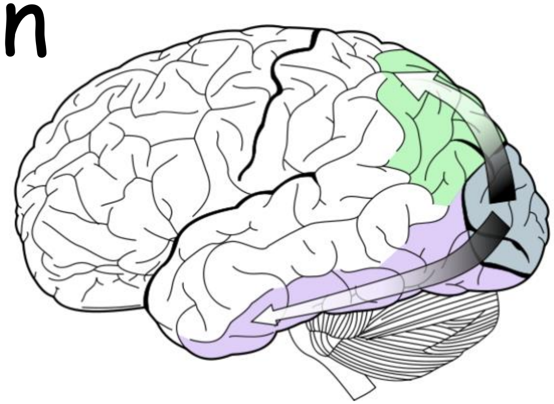
- Positive set: image contains object
- Goal to train standard object detector
- Example positive set:

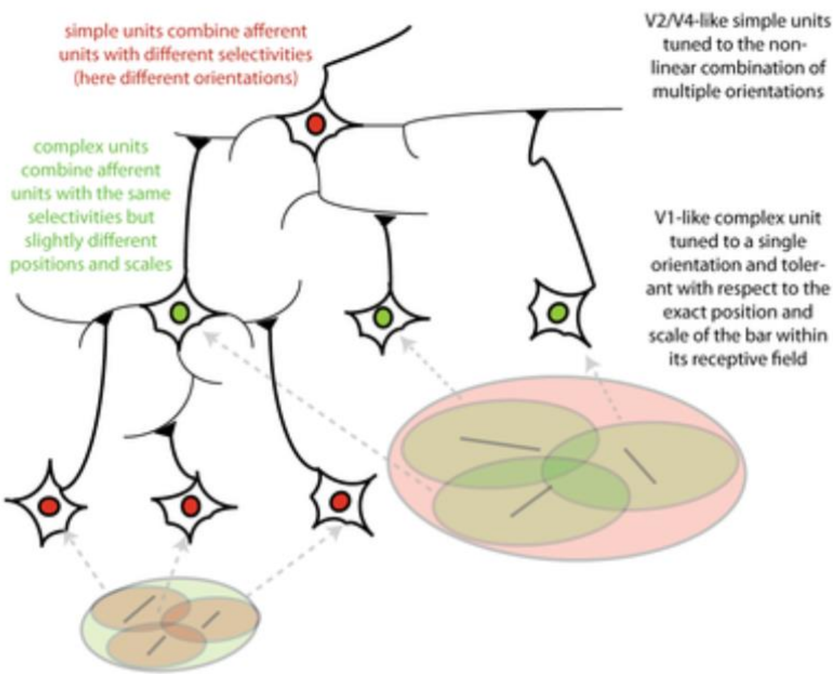


[Viola 05]

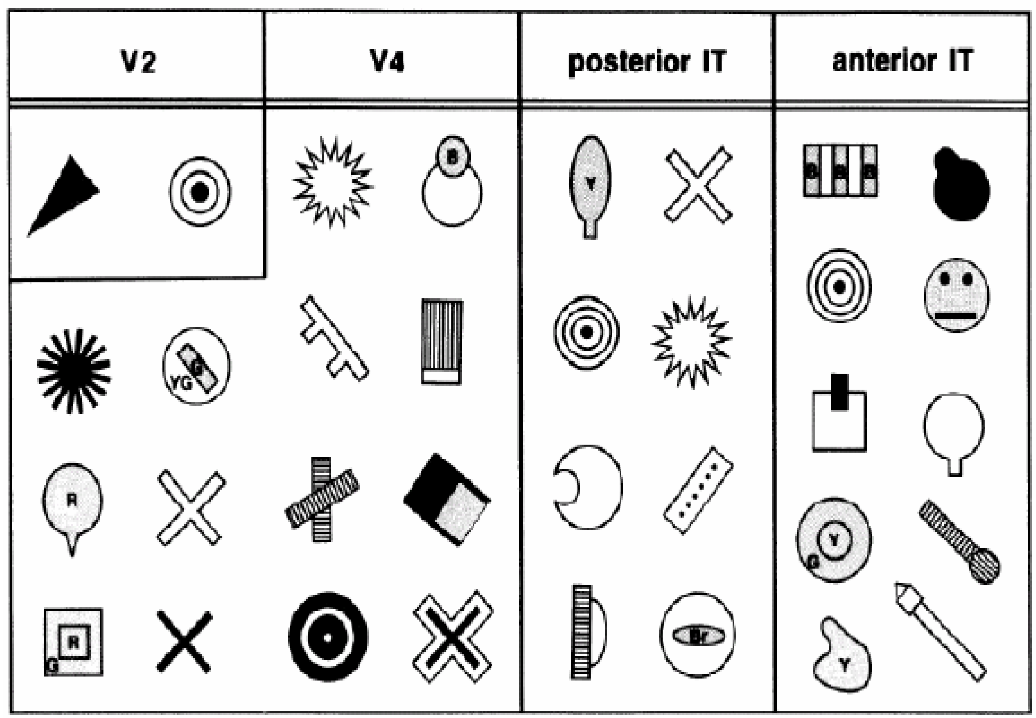
Weakly-supervised learning for structured data.

Visual representation



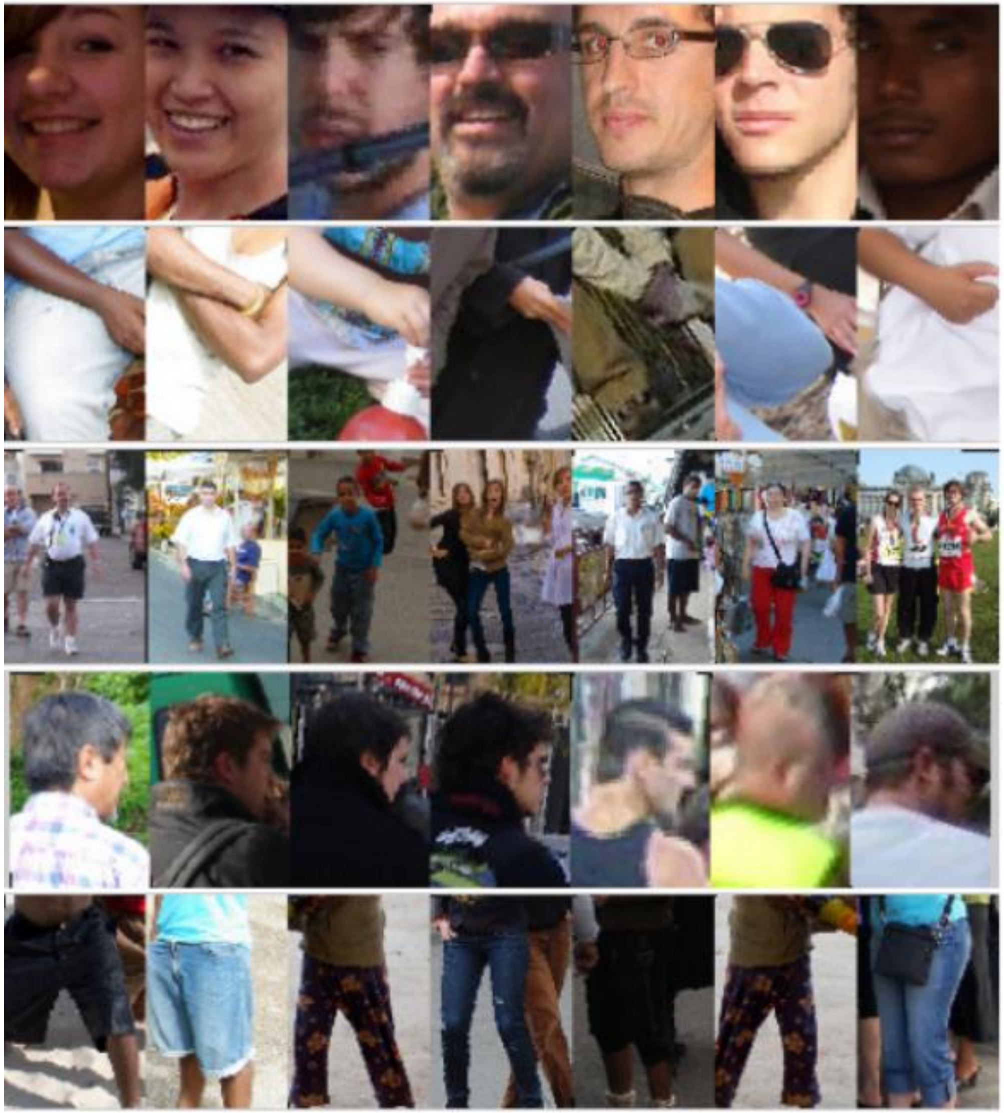


Hubel and Wiesel Model



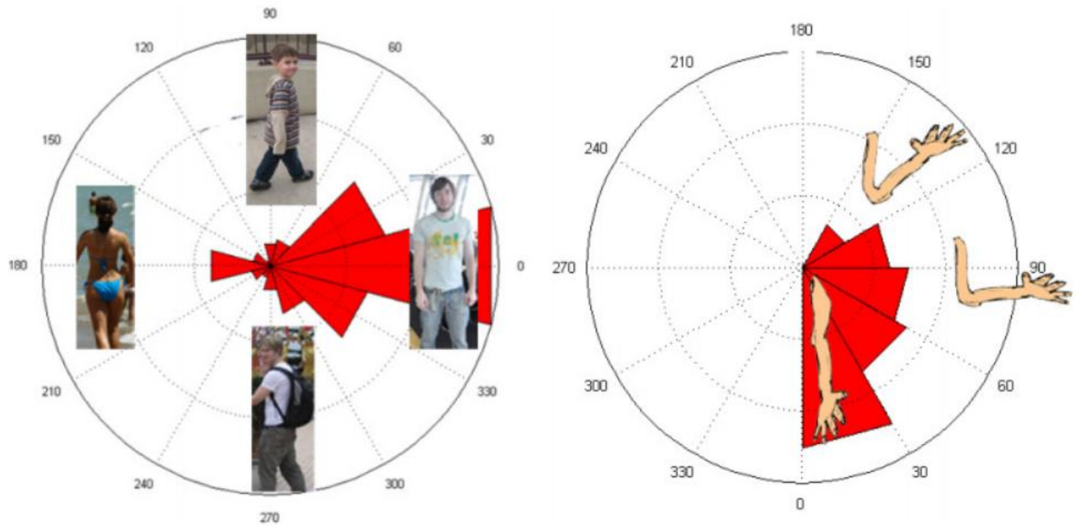
Kobatake and Tanaka, 1994

Poselets: a fully supervised approach



Specific body parts with full supervision
(Bourdev and Malik, 2010)

3D poselets

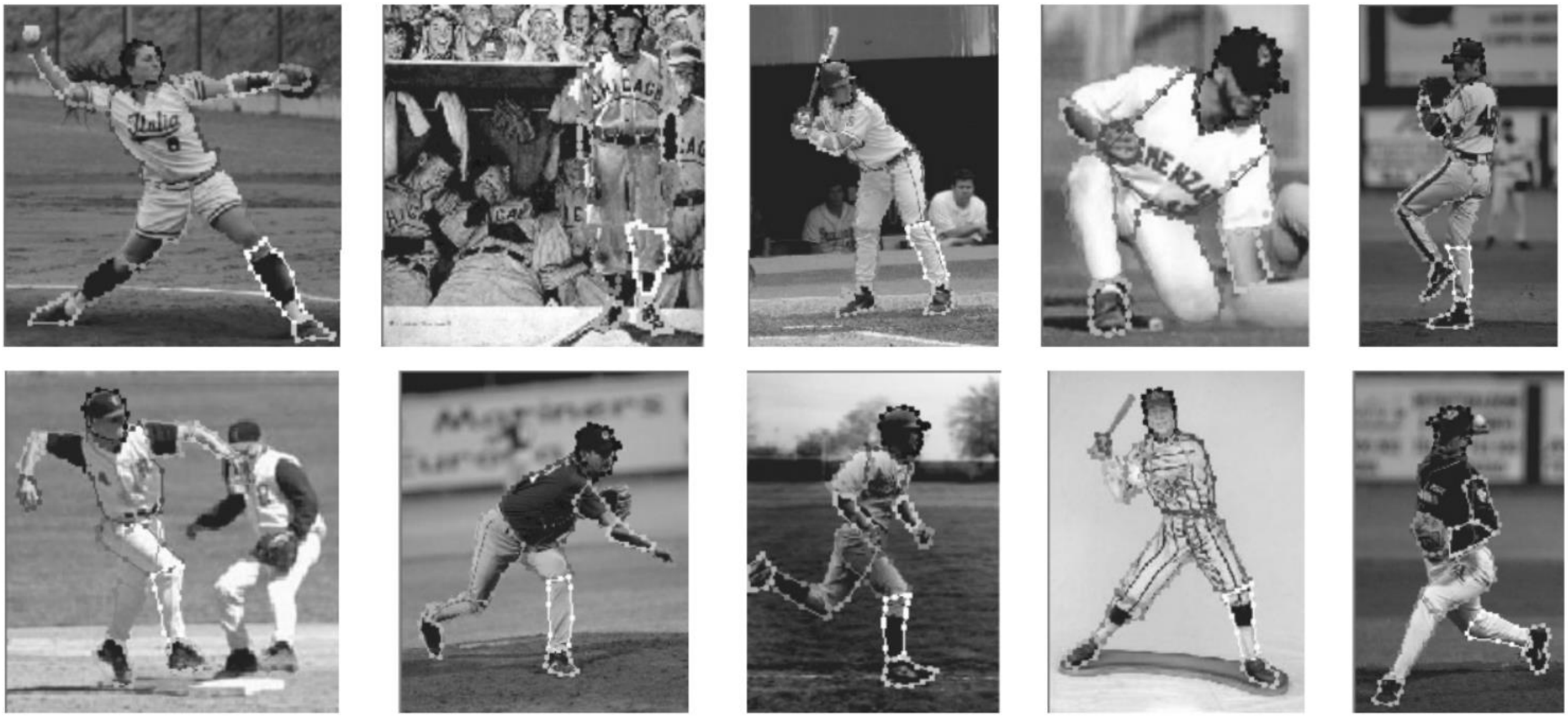


3D Postelets

Torso detection using poselets

(Bourdev and Malik, 2010)

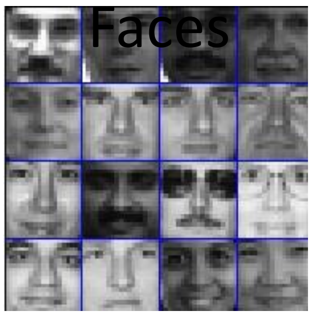
Body parts are hard to define in presence of occlusion



Object detection

object vs. background*

Frontal Faces



Motorbikes

Spotted Cats

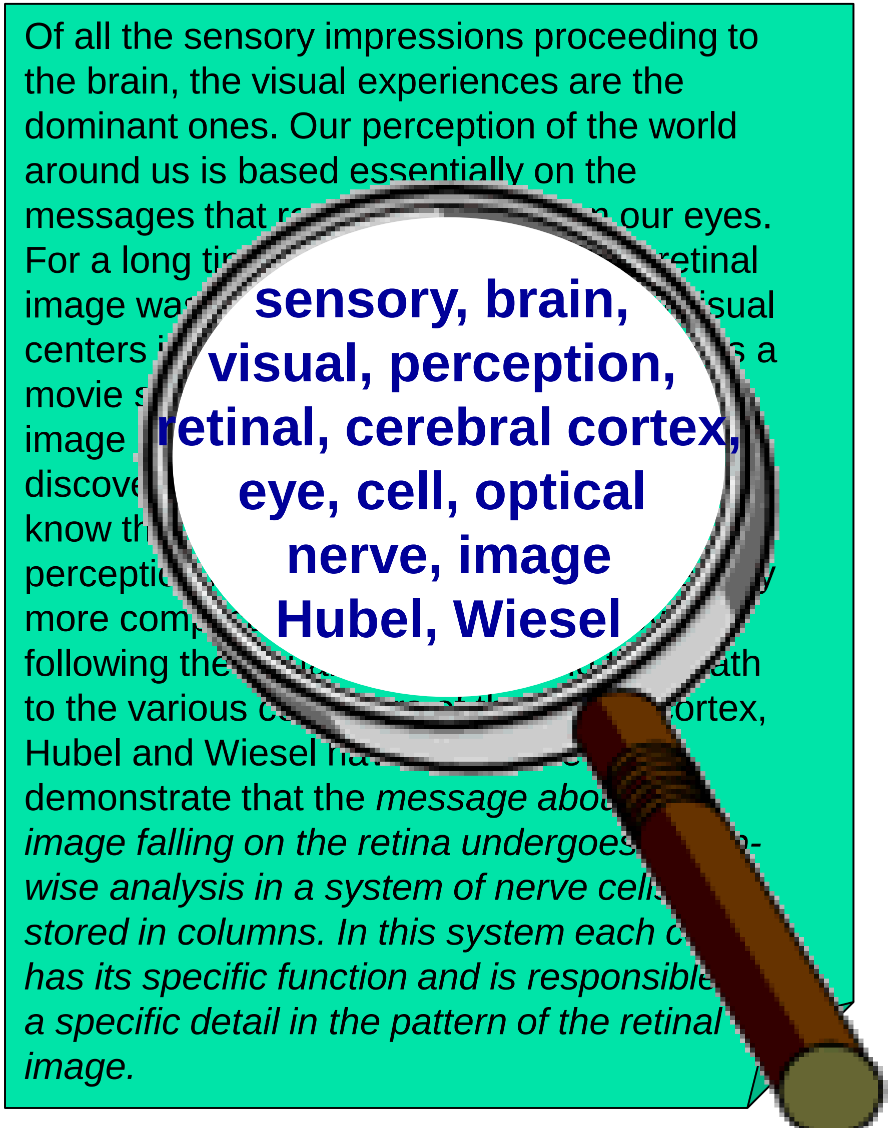
Rigid ⟷ **Articulated**

Dollar et al. 2008

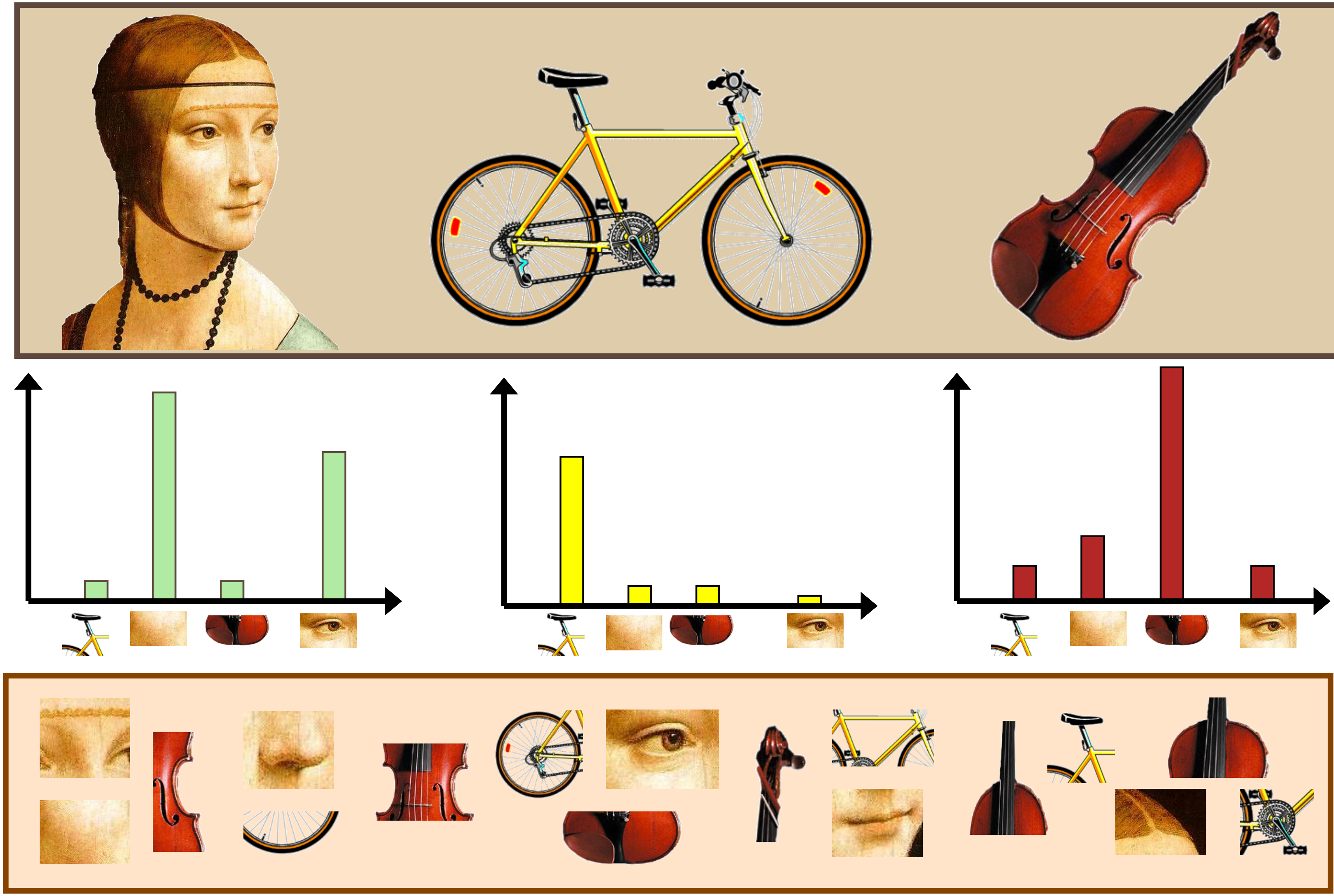
Object → Bag of ‘words’

L. Fei-fei

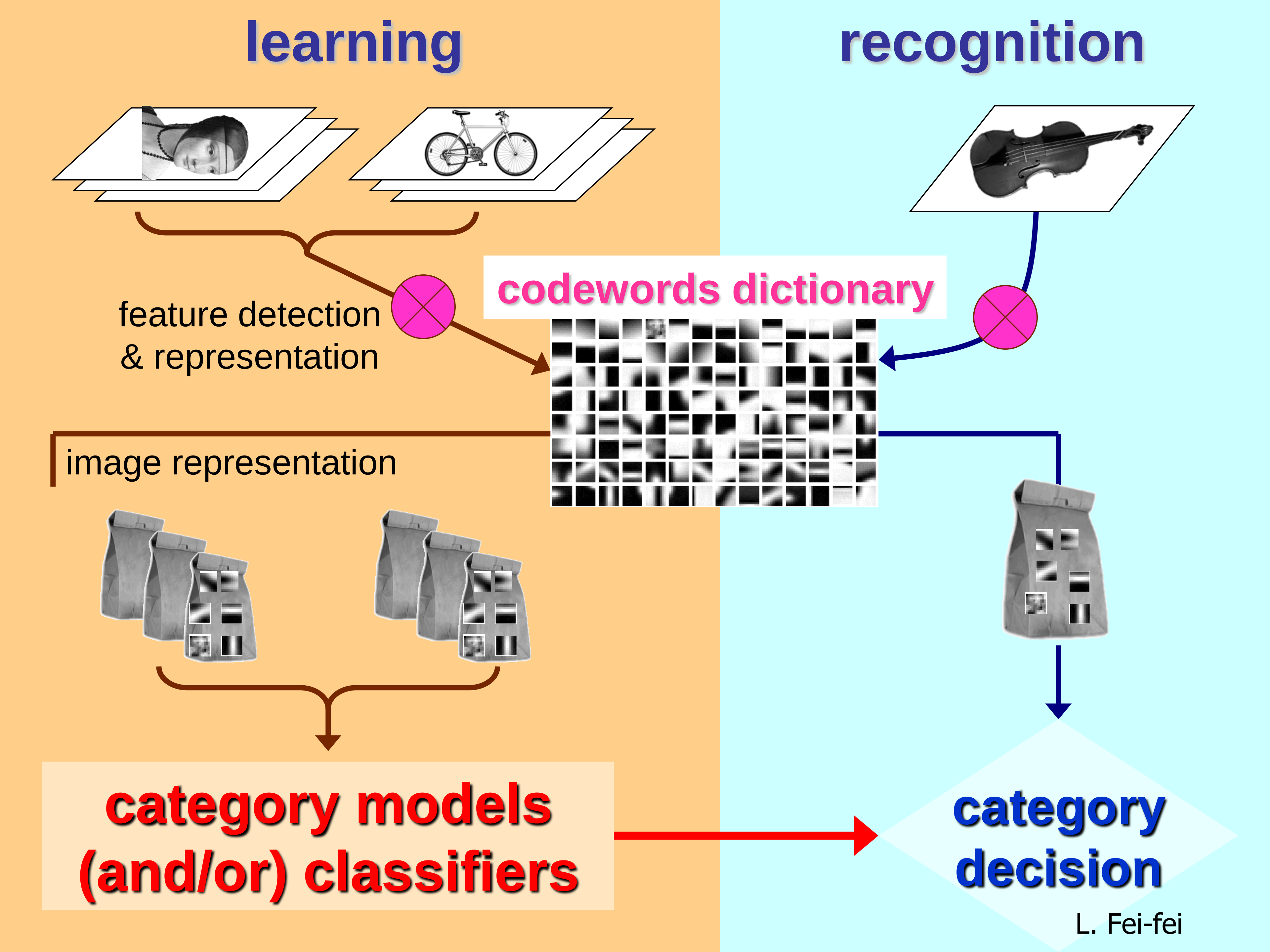
Analogy to documents



L. Fei-fei



L. Fei-fei

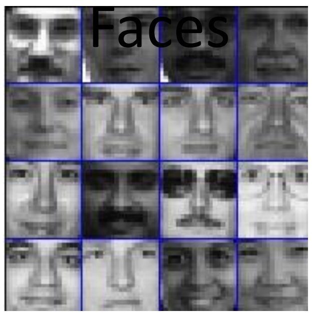

learning
recognition
feature detection
& representation
codewords dictionary
image representation
category models
(and/or) classifiers
category
decision
L. Fei-fei

To learn parts with weak-supervision.

Object detection

object vs. background*

Frontal Faces



Motorbikes

Spotted Cats

Rigid ⟷ **Articulated**

Dollar et al. 2008

Standard vs MIL vs MCL

Dollar et al. 2008

MCL Definition (1)

- Most general definition of a set/bag classifier:

$$\boxed{\mathcal{F}^k(X_i) = \begin{cases} 1 & \text{if } \exists j_1 \ldots, j_k \text{ s.t. } g([x_{ij_1} \ldots, x_{ij_k}]) = 1 \\ 0 & \text{otherwise} \end{cases}}$$

  Note defined $\mathcal{F}^k$ in terms of regular function *g*

- To compute $\mathcal{F}^k(X_i)$:
  - For every sequence $j_1, \ldots, j_k$ test $g([x_{ij_1} \ldots, x_{ij_k}])$
  - Computation time exponential in *k:* $O(m^k)$ (*m* is set size)

- Model exponential in number of components

MCL definition (2)

- This leads to the second MCL formulation:

Sets $\tilde{\mathcal{F}}(X_i) = \tilde{g}\left(\mathcal{F}_1^k(X_i), \ldots, \mathcal{F}_T^k(X_i)\right)$

$\mathcal{F}(\mathbf{X}_i) = \tilde{g}\left(\mathcal{F}_1^k(X_i^1), \ldots, \mathcal{F}_p^k(X_i^p)\right)$

Sequence of sets

$\tilde{g}$ a standard function

$\leftarrow$ T "components"

use small *k*

- $\tilde{\mathcal{F}}(X_i)$ depends on up to Tk instances
  - Computation time is $O(Tm^k)$ + the running time of $\tilde{g}$
  - For *k=1*, running time is linear in *T* and *m*
- But, is training tractable?

Learning: single component

- Note:

$$\mathcal{F}^1(X_i) = \begin{cases} 1 & \text{if } \exists j \text{ s.t. } g([x_{ij}]) = 1 \\ 0 & \text{otherwise} \end{cases} \quad \leftarrow \text{MCL } (k\text{=1})$$

$$F(X_i) = \begin{cases} 1 & \text{if } \exists j \text{ s.t. } f(x_{ij}) = 1 \\ 0 & \text{otherwise} \end{cases} \quad \leftarrow \text{MIL}$$

- So first formulation of MCL with *k*=1 equivalent to MIL
  - Can also show reduction for *k*>1, but training exponential in *k*
- Therefore existing MIL algorithms provide mechanism to learn single components

Learning multiple components

- Additive Formulation:

$$\mathcal{F}(X_i) = \text{sign}\Big(\sum_{t=1}^{T} \alpha_t [2F_t(X_i) - 1]\Big)$$

- Additive models are simple but powerful
  - Prevalent in statistics, rich theory
  - Can use boosting to train additive model

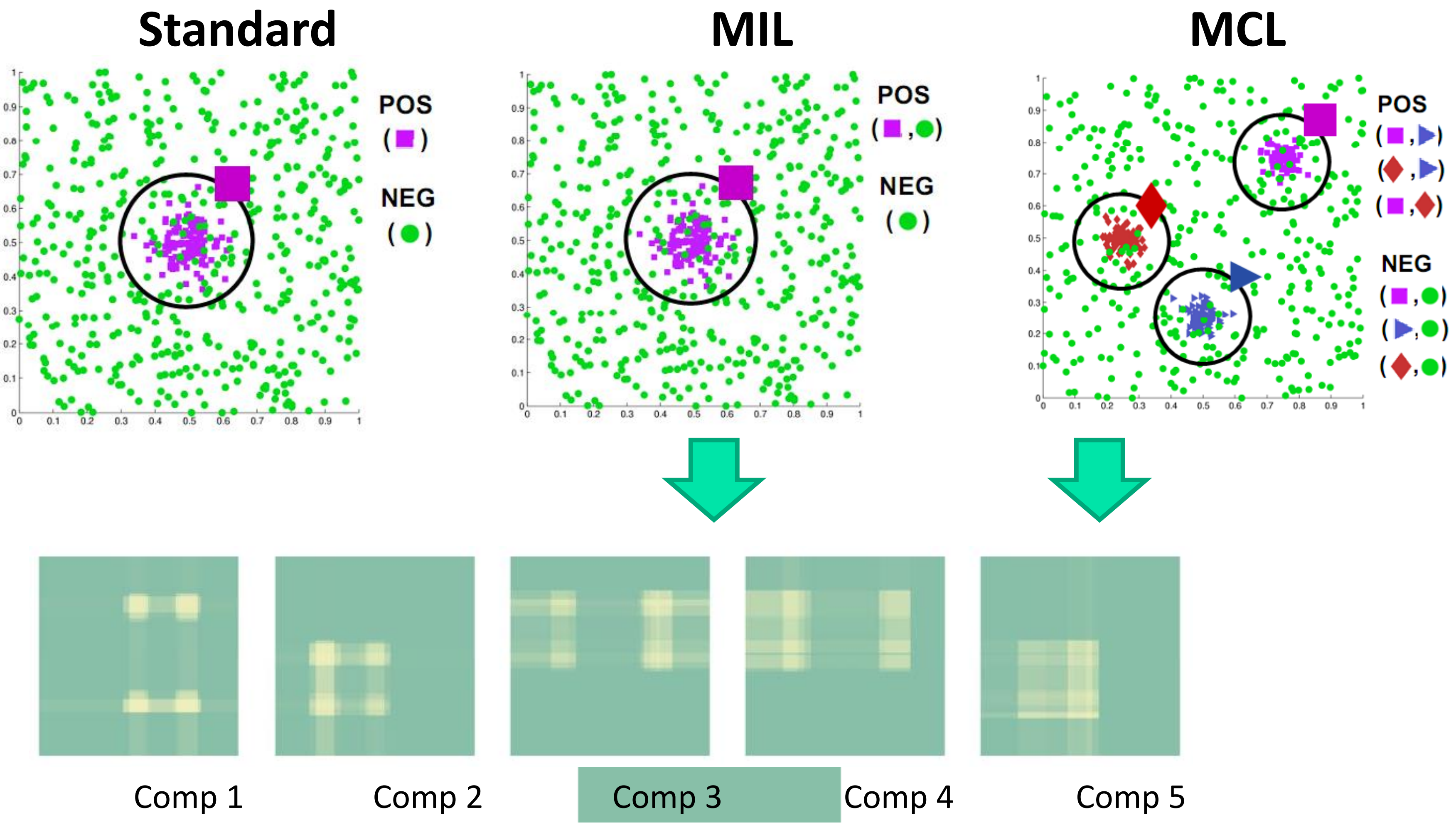
Learning multiple components

- General algorithm:
  - Use MIL to obtain weak classifiers (components)
  - Use boosting to combine components into strong classifier

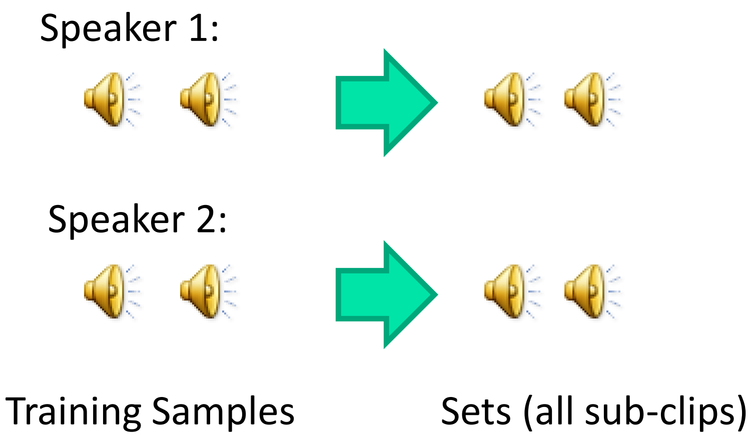
- RealBoost for MCL:

$$\mathcal{F}(X_i) = \mathrm{sign}\Big(\sum_{t=1}^{T} \frac{1}{2} \log \frac{\hat{F}_t(X_i)}{1 - \hat{F}_t(X_i)}\Big)$$

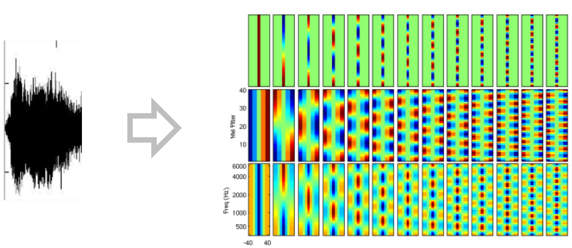
Standard vs MIL vs MCL

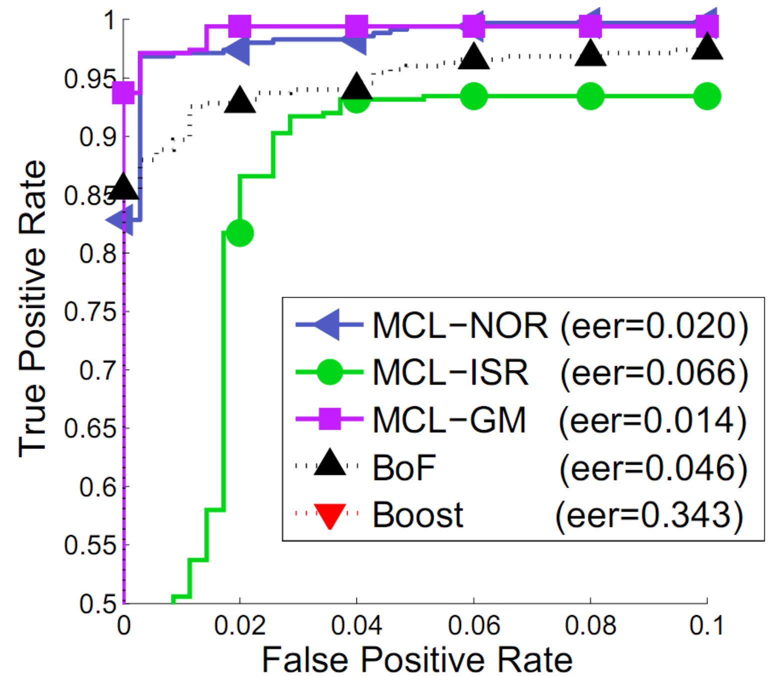


Speaker identification



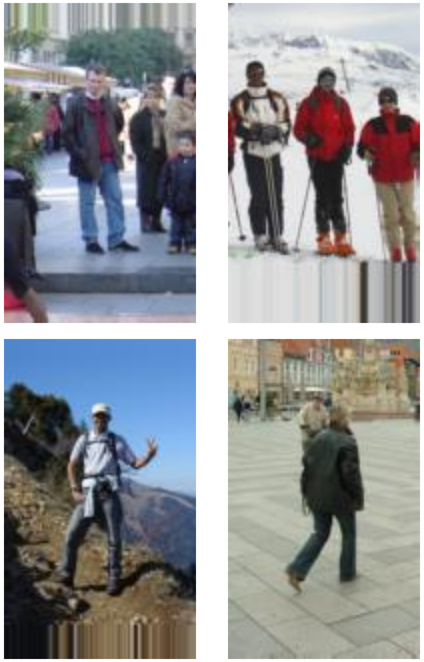
VoiceBox Matlab Toolbox (MFCC features)





Results

Pedestrian detection



- Inria Dataset [Dalal & Triggs 2005]
  - 1213 Training Positives (+ reflections)
  - O(2000) background training images
  - Test dataset about ½ as big
- Verification task:
  - Does window contain pedestrian?
- Challenging dataset, much recent work

Specialized version of MCL:

1. Optimize MIL training
2. Incorporate spatial model

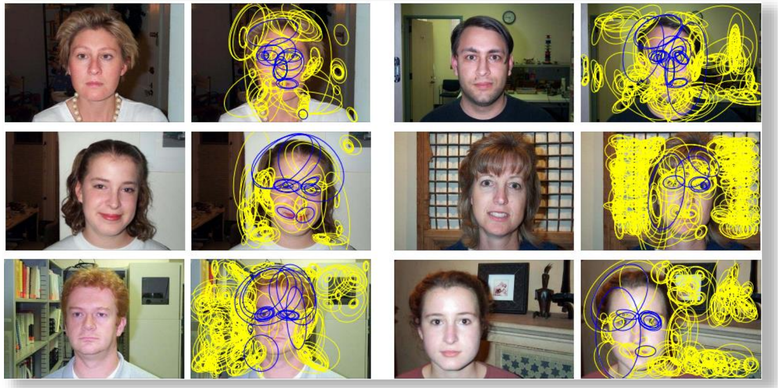
Learning from Less Supervision

- Learn object class models from unlabeled/weakly labeled images.
- Unsupervised/Weakly Supervised Learning.

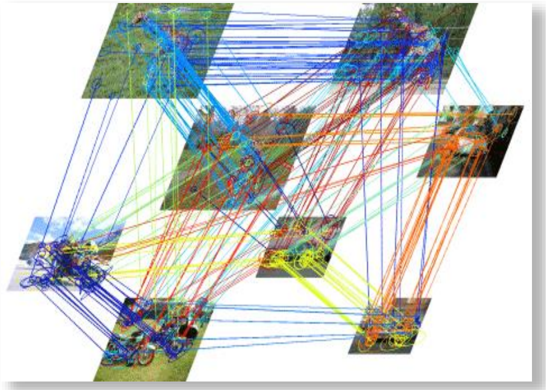
R. Girshick

“Is it possible to learn visual object classes simply from looking at images?” - [Josef Sivic et al. ICCV 2005]

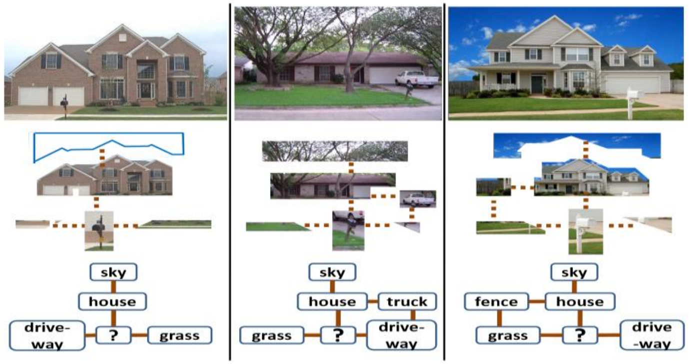
Topic Models (pLSA/LDA)
[Sivic et al. ICCV 05]
[Russell et al. CVPR 06]



Pyramid Match Kernel + Normalized Cut
[Grauman and Darrell. CVPR 06']
[Lee and Grauman. IJCV 09']



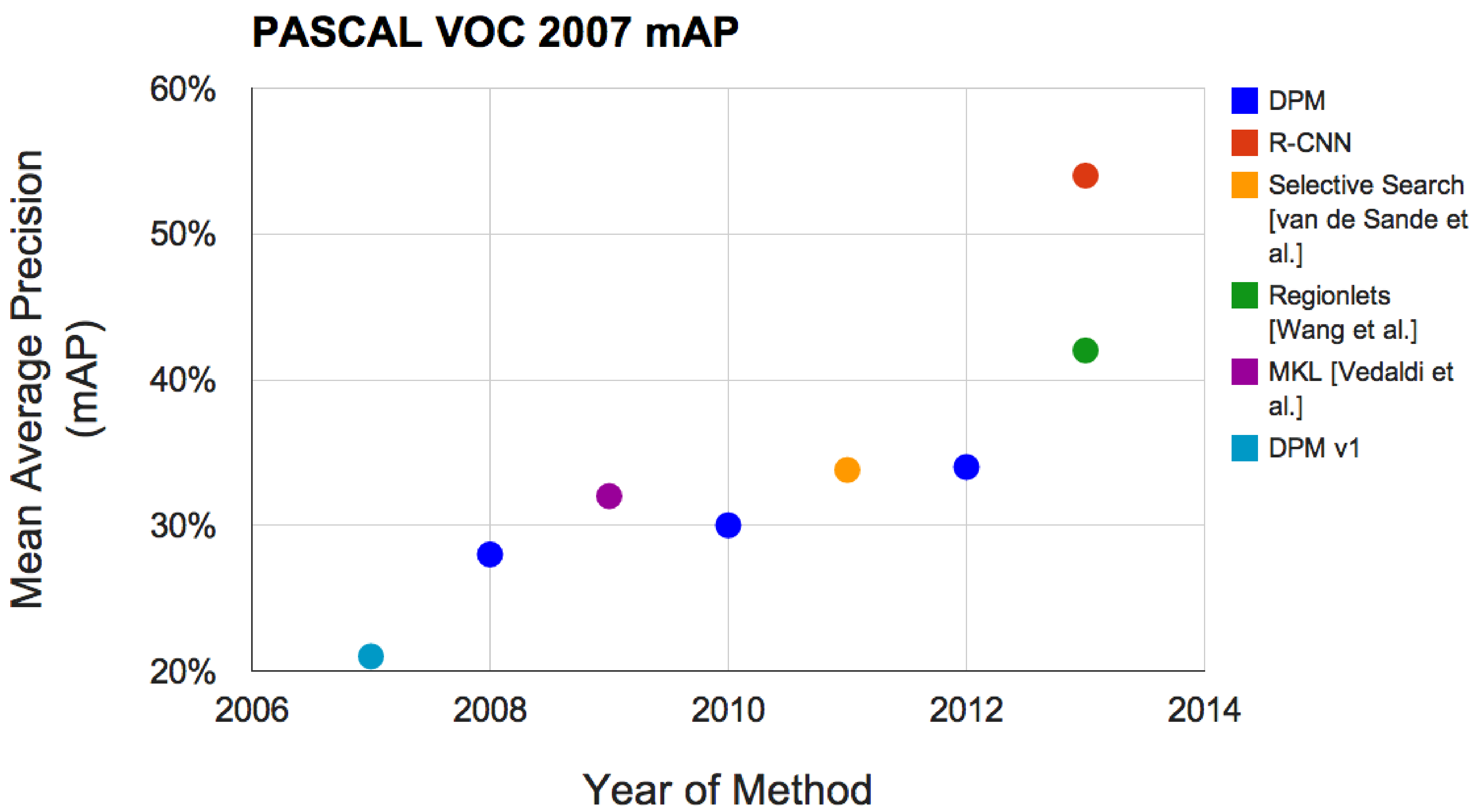
Link Analysis Technique
[Kim et al. CVPR 05']
[Kim and Torralba. NIPS 09']



Context-Aware Discovery
[Lee and Grauman. CVPR 10']
[Deselaers et al. IJCV 12']

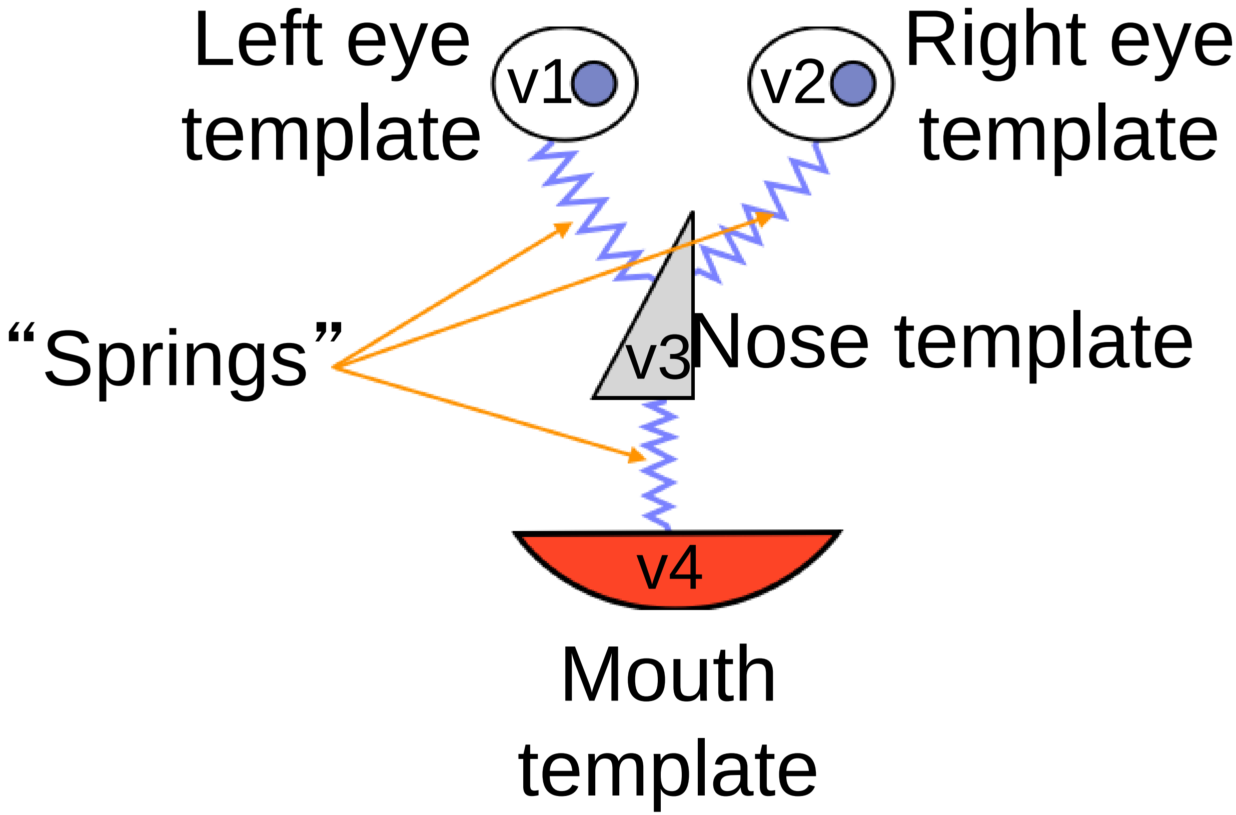
K. Grauman

PASCAL VOC results over time



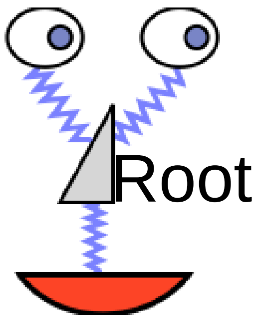
R. Girshick

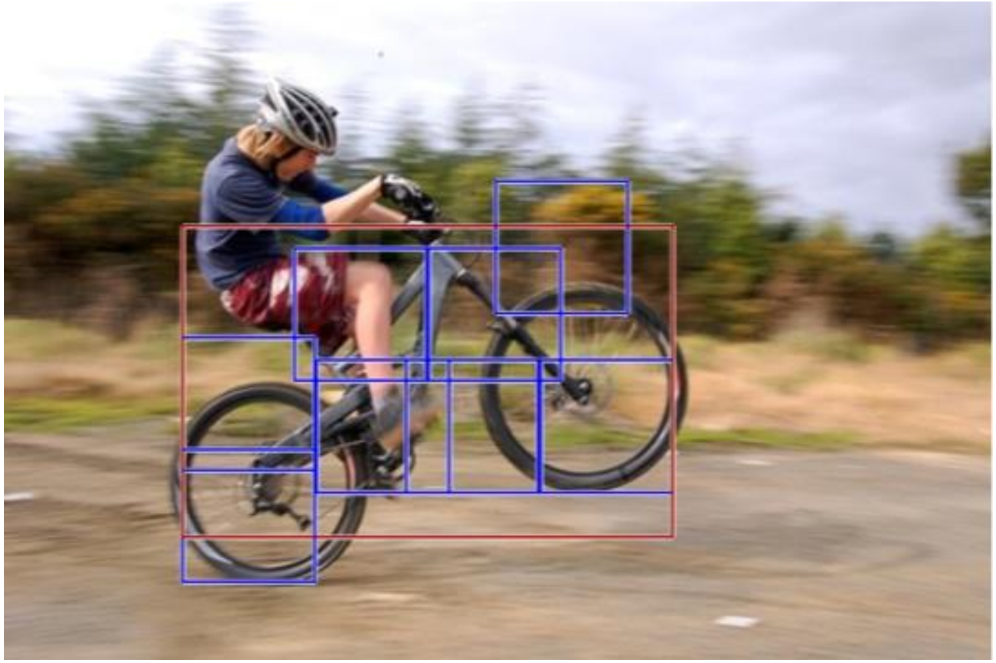
Deformable models

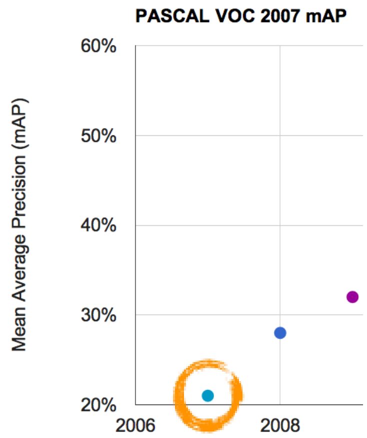


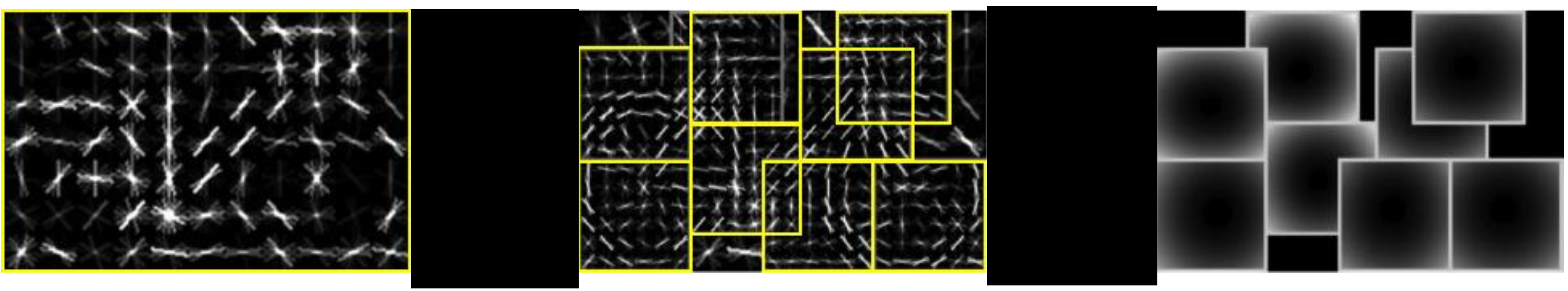
Fischler and Elschlager 1973 "The Representation and Matching of Pictorial Structures"

34 years later









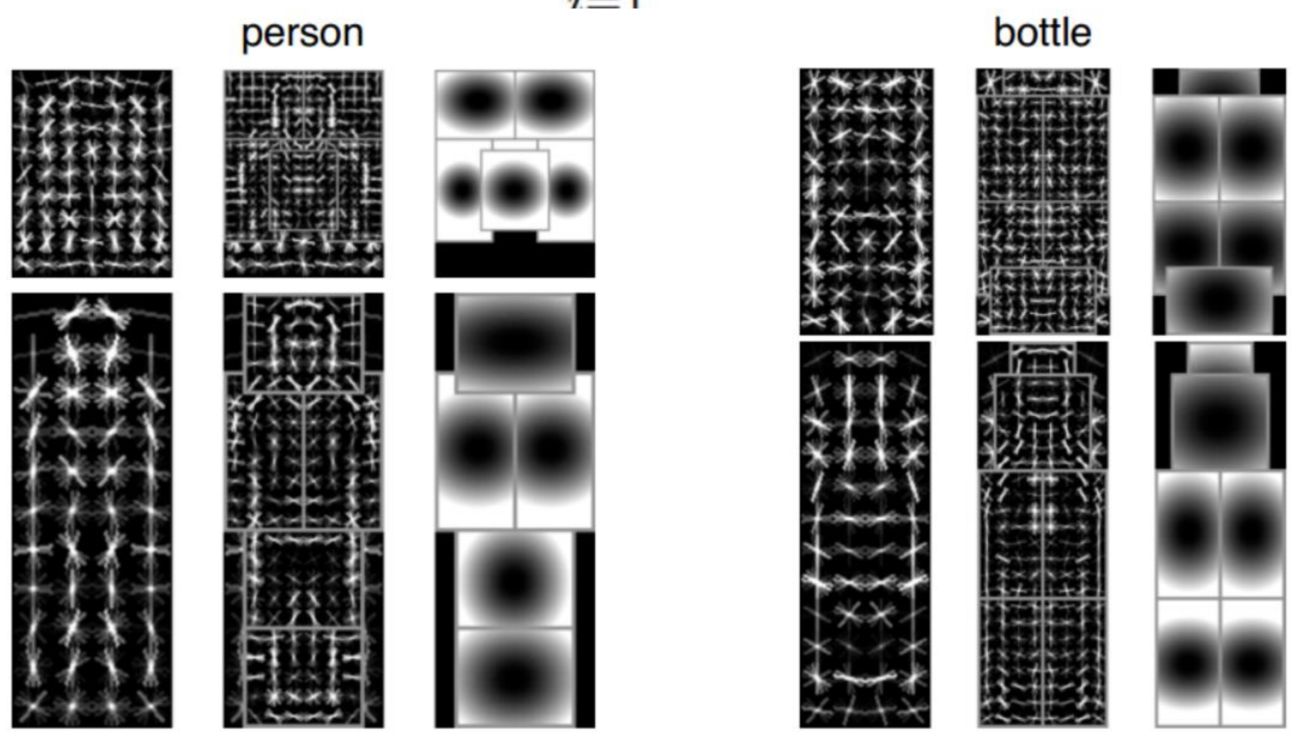
Root template Part templates Spring costs

(Felzenszwalb, McAllester, Ramanan ’08)

Discriminative-trained part-based models

$$f_\beta(x) = \max_{z \in Z(x)} \beta \cdot \Phi(x, z)$$

$$L_D(\beta) = \frac{1}{2}||\beta||^2 + C \sum_{i=1}^{n} \max(0, 1 - y_i f_\beta(x_i))$$



(Felzenszwalb, McAllester, Ramanan '08)

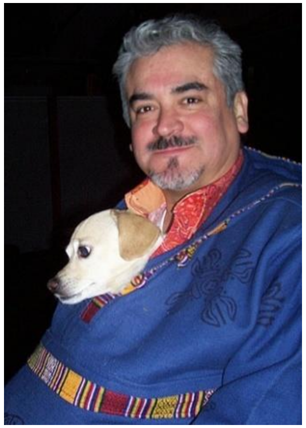
Deformation is not enough

Viewpoint

Subclasses

R. Girshick

Deformation is not enough

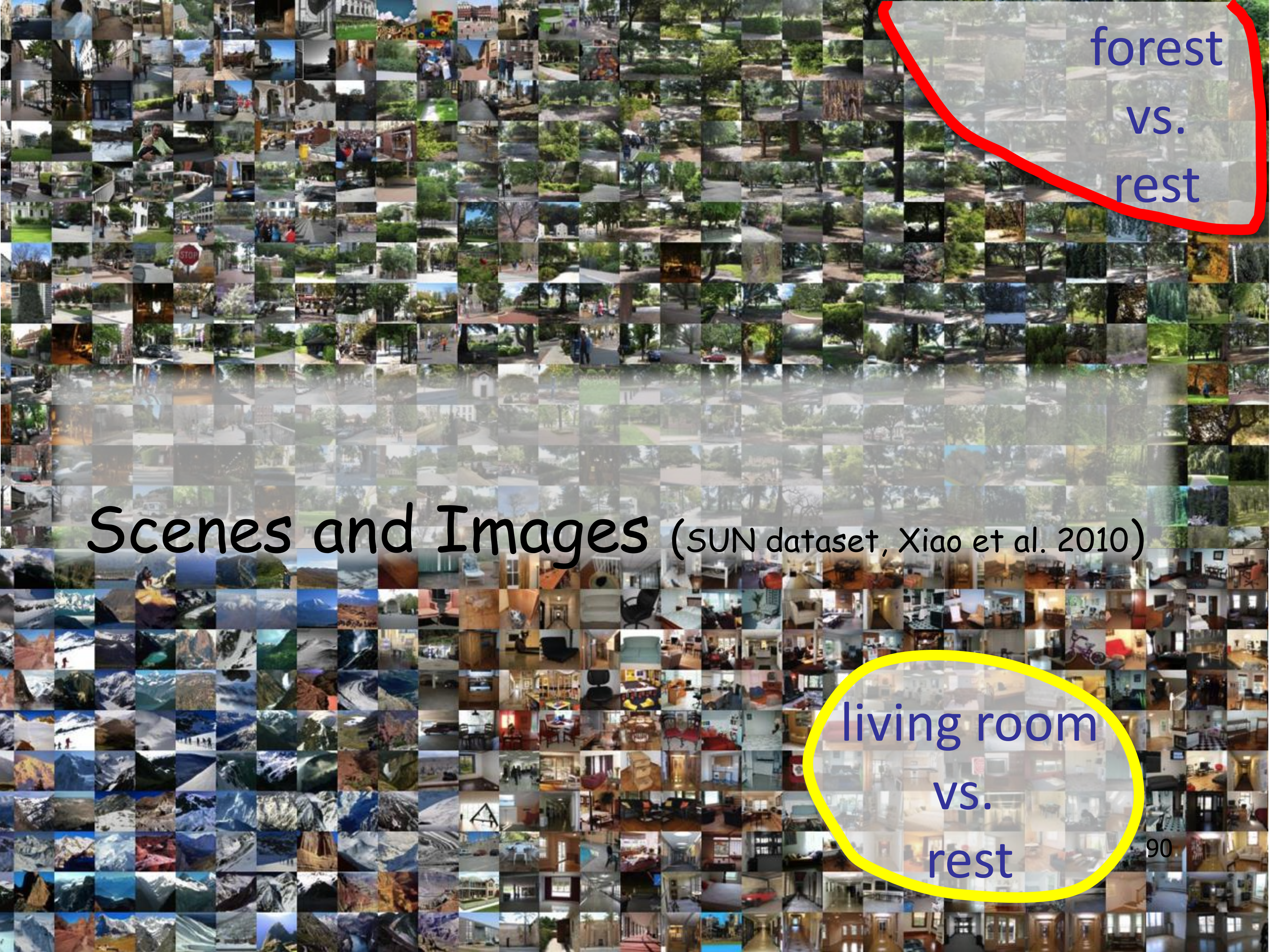



Occlusion/truncation


Symmetries

Compositional structure (kid with bucket hat and scuba goggles)

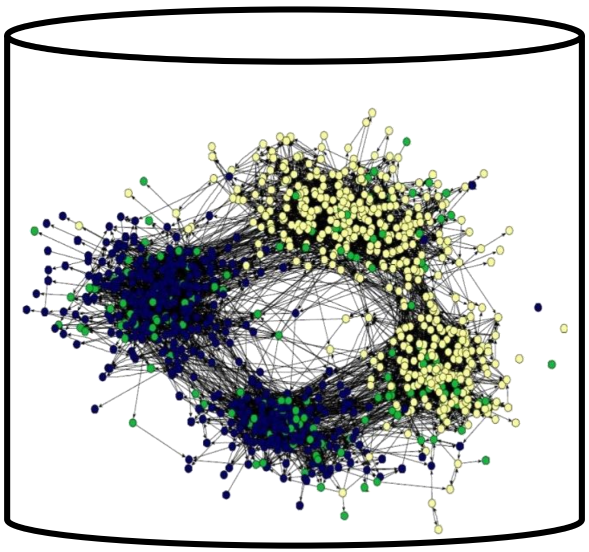
R. Girshick


forest
vs.
rest
Scenes and Images (SUN dataset, Xiao et al. 2010)
living room
vs.
rest
90

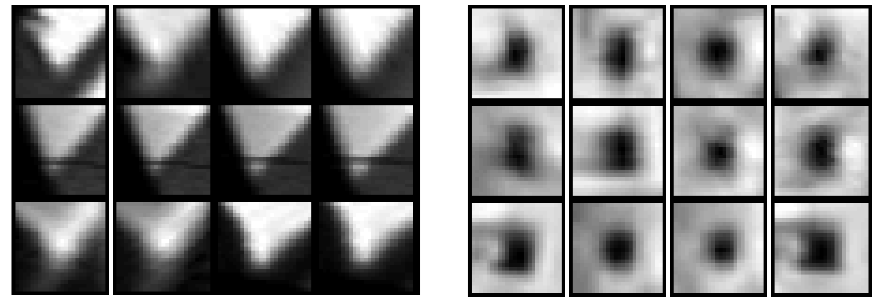
when? 1972

Lee, Efros, and Hebert

Visual data mining in computer vision



Visual world



Low-level "visual words"

[Sivic & Zisserman 2003, Laptev & Lindeberg 2003, Czurka et al. 2004, …]

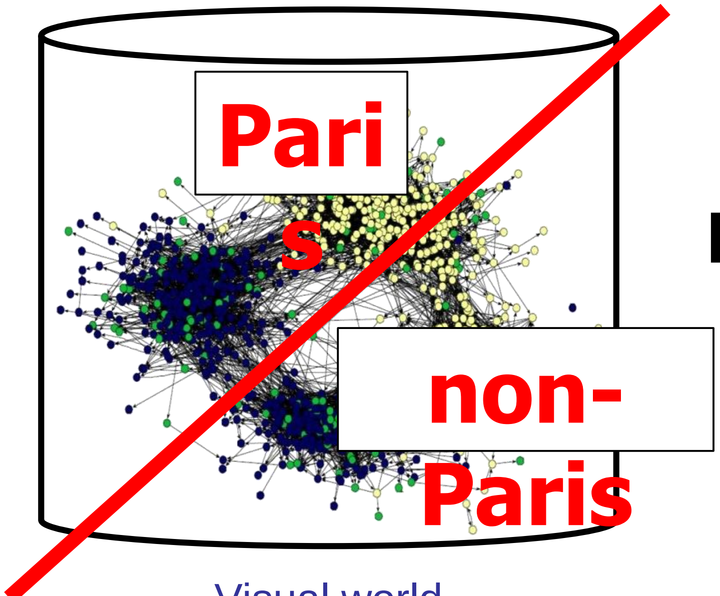
Object category discovery

[Sivic et al. 2005, Grauman & Darrell 2006, Russell et al. 2006, Lee & Grauman 2010, Payet & Todorovic, 2010, Faktor & Irani 2012, Kang et al. 2012, …]

- Most approaches mine *globally consistent* patterns

Lee, Efros, and Hebert

Visual data mining in computer vision



Visual world

Mid-level visual elements

[Doersch et al. 2012, Endres et al. 2013, Juneja et al. 2013, Fouhey et al. 2013, Doersch et al. 2013]

- Recent methods discover *specific* visual patterns

Lee, Efros, and Hebert

Problem

- Much in our visual world undergoes a *gradual change*

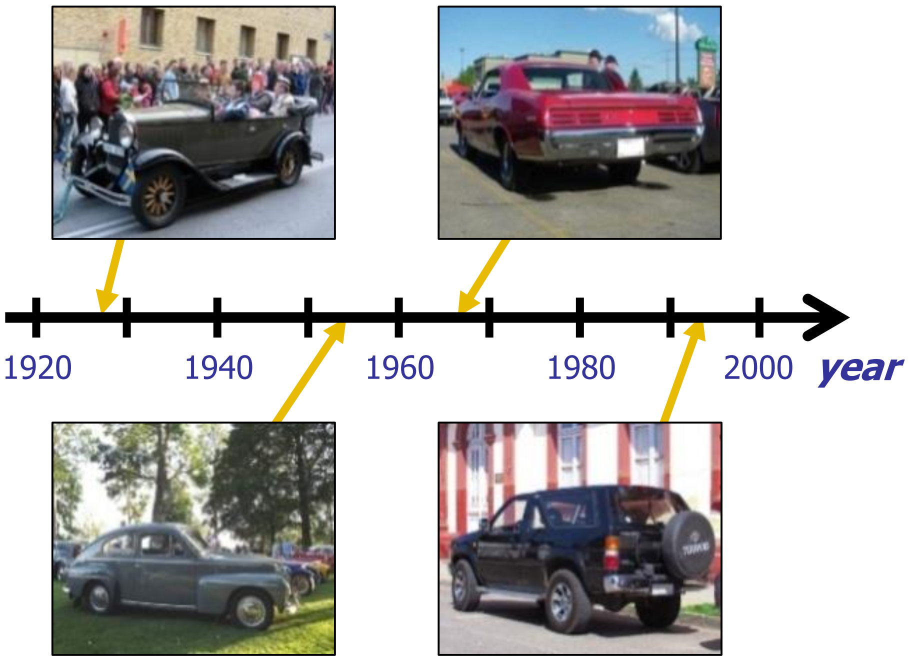
Temporal:

1887-1900	1900-1941	1941-1969	1958-1969	1969-1987

Lee, Efros, and Hebert

Goal

- Mine mid-level visual elements in temporally- and spatially-varying data and model their “visual style”



when?

Historical dating of cars

[Kim et al. 2010, Fu et al. 2010, Palermo et al. 2012]

where?

Geolocalization of StreetView images

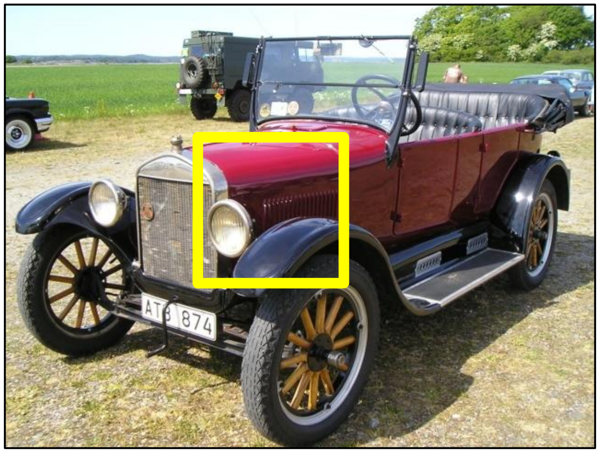
[Cristani et al. 2008, Hays & Efros 2008, Knopp et al. 2010, Chen & Grauman. 2011, Schindler et al. 2012]

Lee, Efros, and Hebert

Key Idea

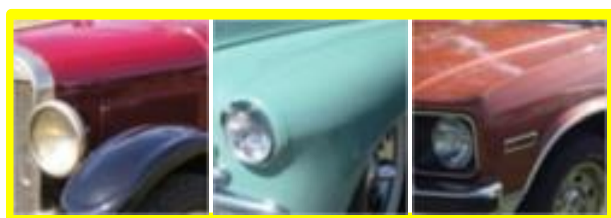
1) Establish connections

1926



1947

1975



1926 **1947** **1975**

"closed-world"

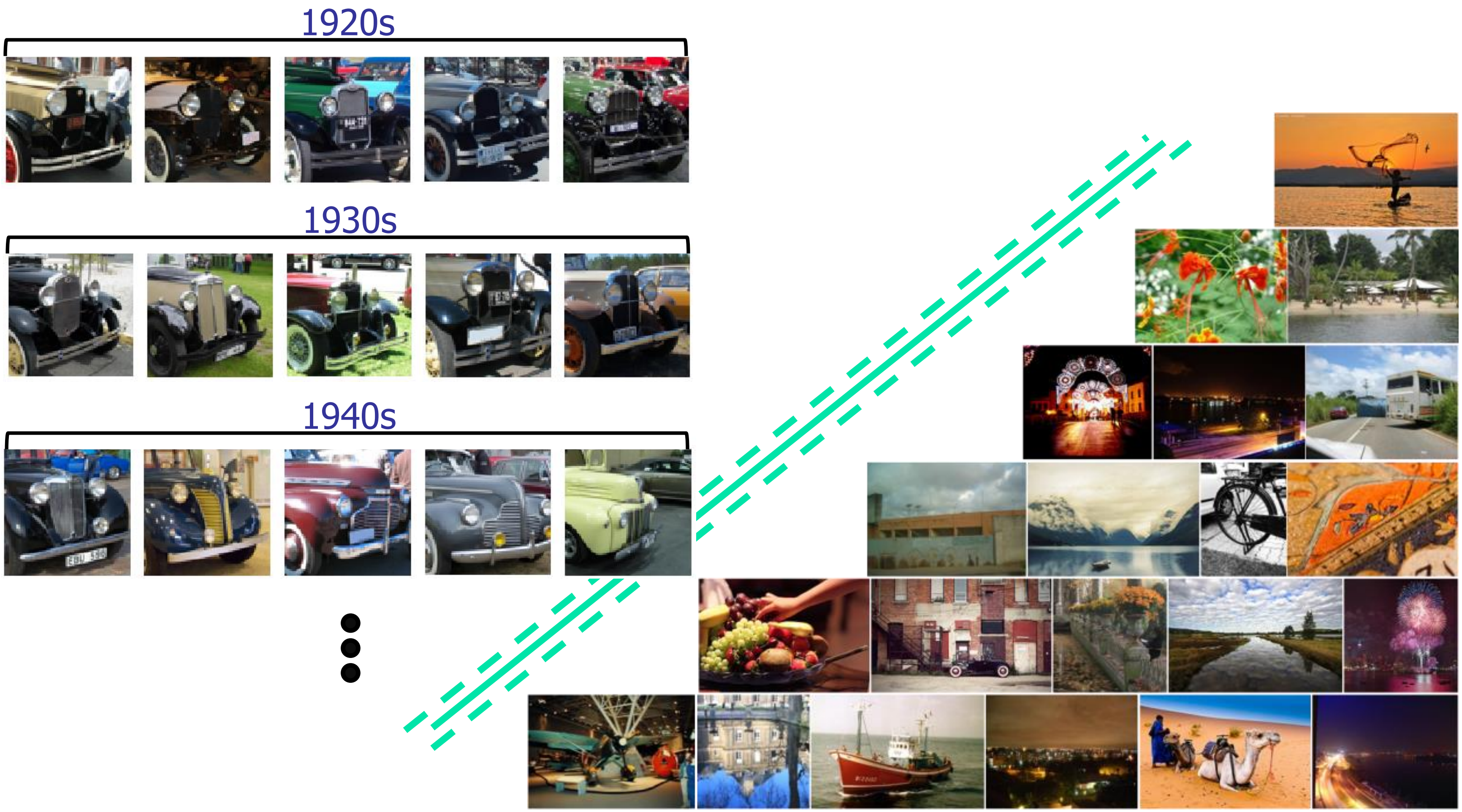
2) Model style-specific differences

Lee, Efros, and Hebert

Making visual connections

Expect style to change gradually...

Lee, Efros, and Hebert



Natural world “background” dataset

Mining style-sensitive elements

Lee, Efros, and Hebert

Making visual connections

1920s 1930s 1940s 1950s 1960s 1970s 1980s 1990s

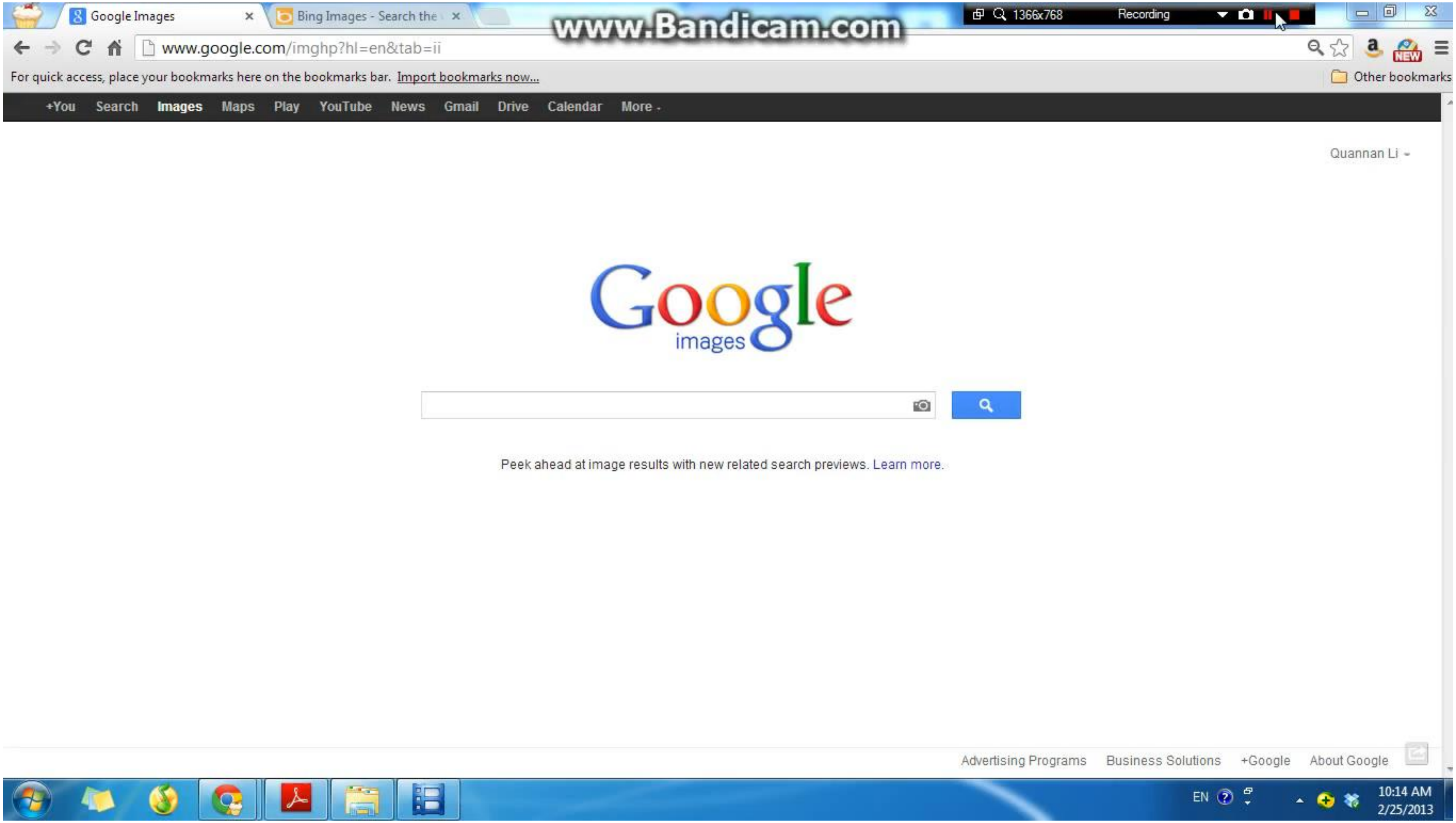
Top detection per decade

Lee, Efros, and Hebert

Mid-Level visual knowledge discovery

- Doersch et al. What Makes Paris Look like Paris? SIGGRAPH 2012
- Singh et al. Unsupervised Discovery of Mid-Level Discriminative Patches. ECCV 2012

Image search



Harvesting mid-level visual concepts from large-scale internet images

700 words ⇨ 450,000 images

goggles
propeller
swing
streetlight
shelf
public
cross
sail
rock
chair
desk
table
backboard
drum
basket
bouquet
pen
glove
horse
seashore
soil
wing
cell
kangaroo
wheel
coaster
candle
bookshelf
balloon
garage
wall
homo
bench
writing
pool
shield
veil

spectacle
fruit
mirror
room
umbrella
toilet
fence
rack
pool
sofa
dressing
attire
basketball
guitar
blind
blanket
bathtub
towel
squash
jersey
cesspool
aqualung
loudspeaker
goggles
propeller
swing
streetlight
shelf
public
cross
sail
rock
chair
desk
table
backboard
drum

key
wheel
button
light
plate
cupboard
door
shower
ball
toilet
table
table-tennis
court
horn
floor
bridal
rug
mouse
racket
boot
duck
oxygen
filter
spectacle
fruit
mirror
room
umbrella
toilet
fence
rack
pool
sofa
dressing
attire
basketball
guitar

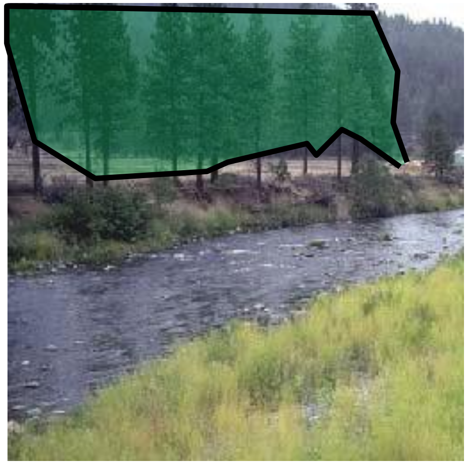
faucet
roller
hook
saddle
snail
drawer
railing
curtain
bed
seat
gravel
table
face
suit
bear
gown
curtain
stick
box
fork
turtle
mask
stove
key
wheel
button
light
plate
cupboard
door
shower
ball
toilet
table
table-tennis
court
horn

wheel
coaster
candle
bookshelf
balloon
garage
wall
homo
bench
writing
pool
shield
veil
shoe
grass
vase
baseball
male
glove
flipper
snake
lion
monkey
faucet
roller
hook
saddle
snail
drawer
railing
curtain
bed
seat
gravel
table
face
snail

Supervised learning for visual concepts

Sky

Trees

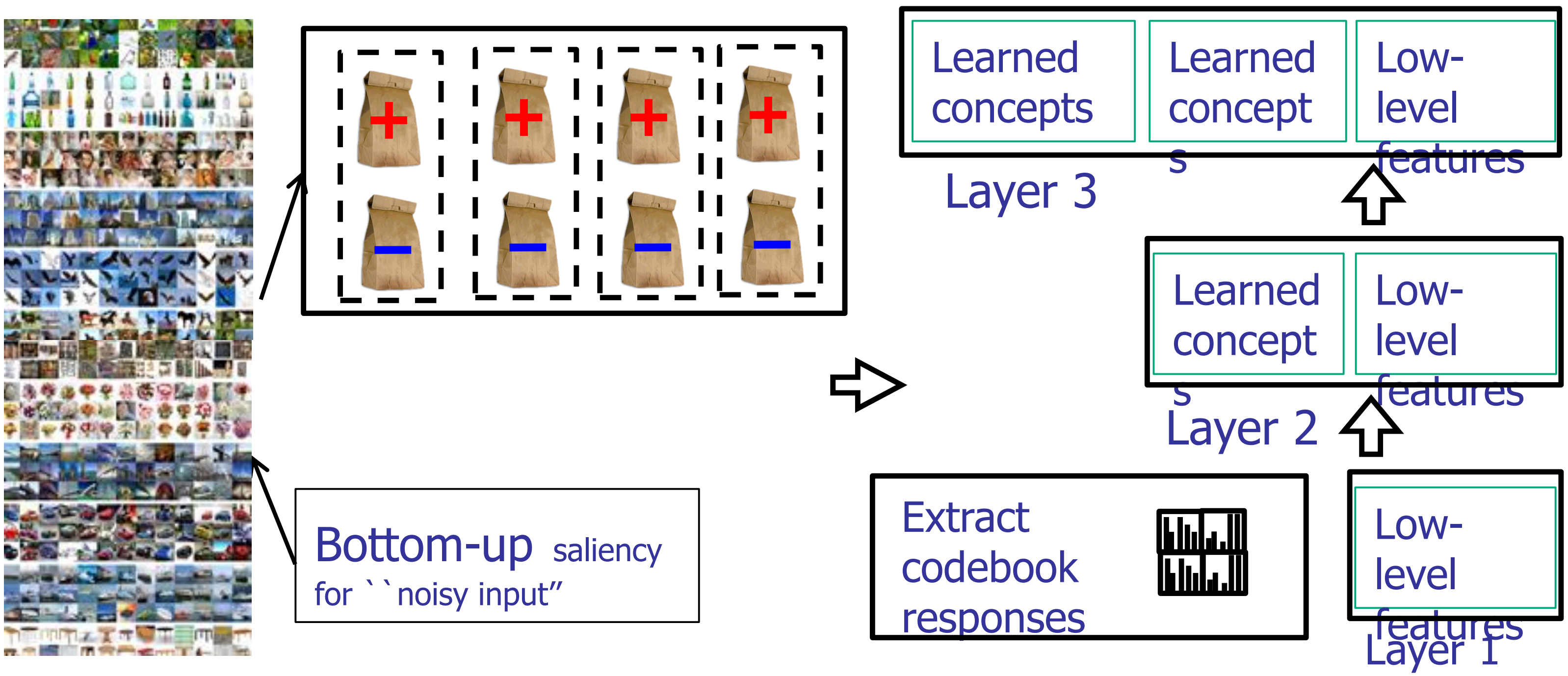


Difficulty with supervised learning

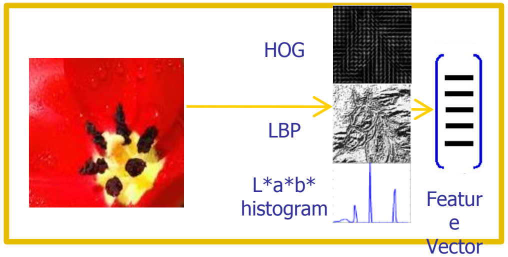
Sky

- Scalability
- Intrinsic ambiguity in human annotations
- Inconsistency across different subjects

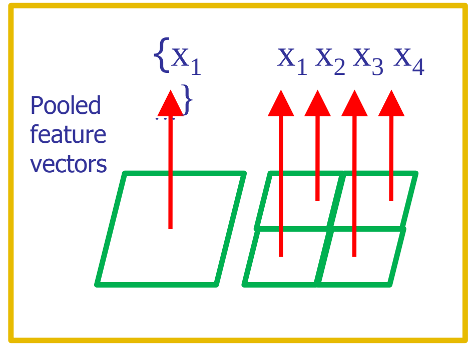
Weakly-Supervised Visual Concept Learning

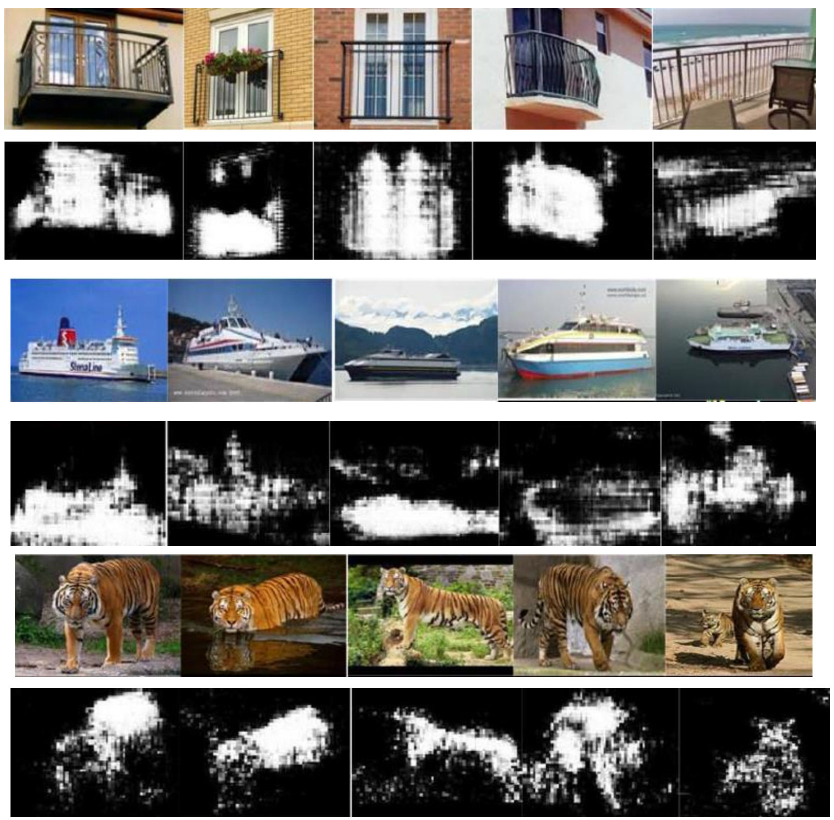


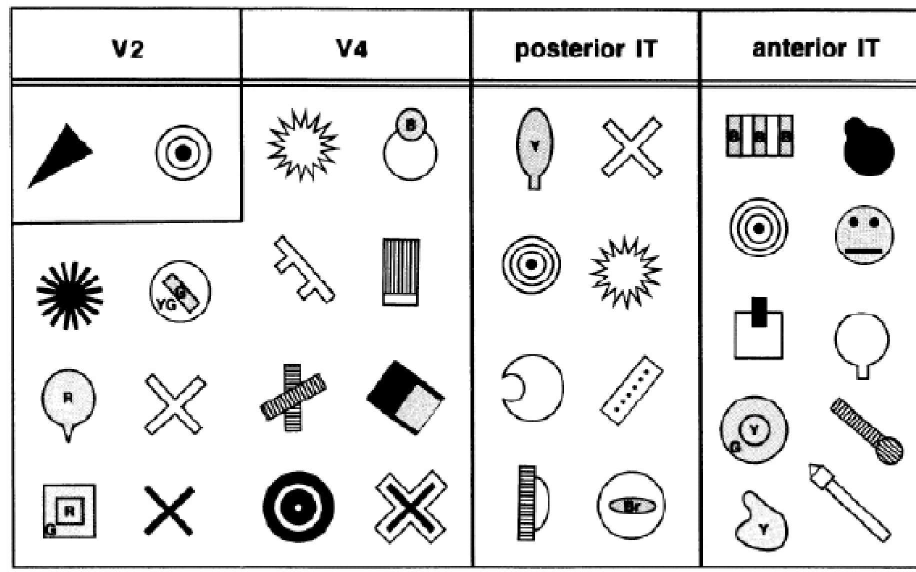
Learned visual concepts



Visual Concepts

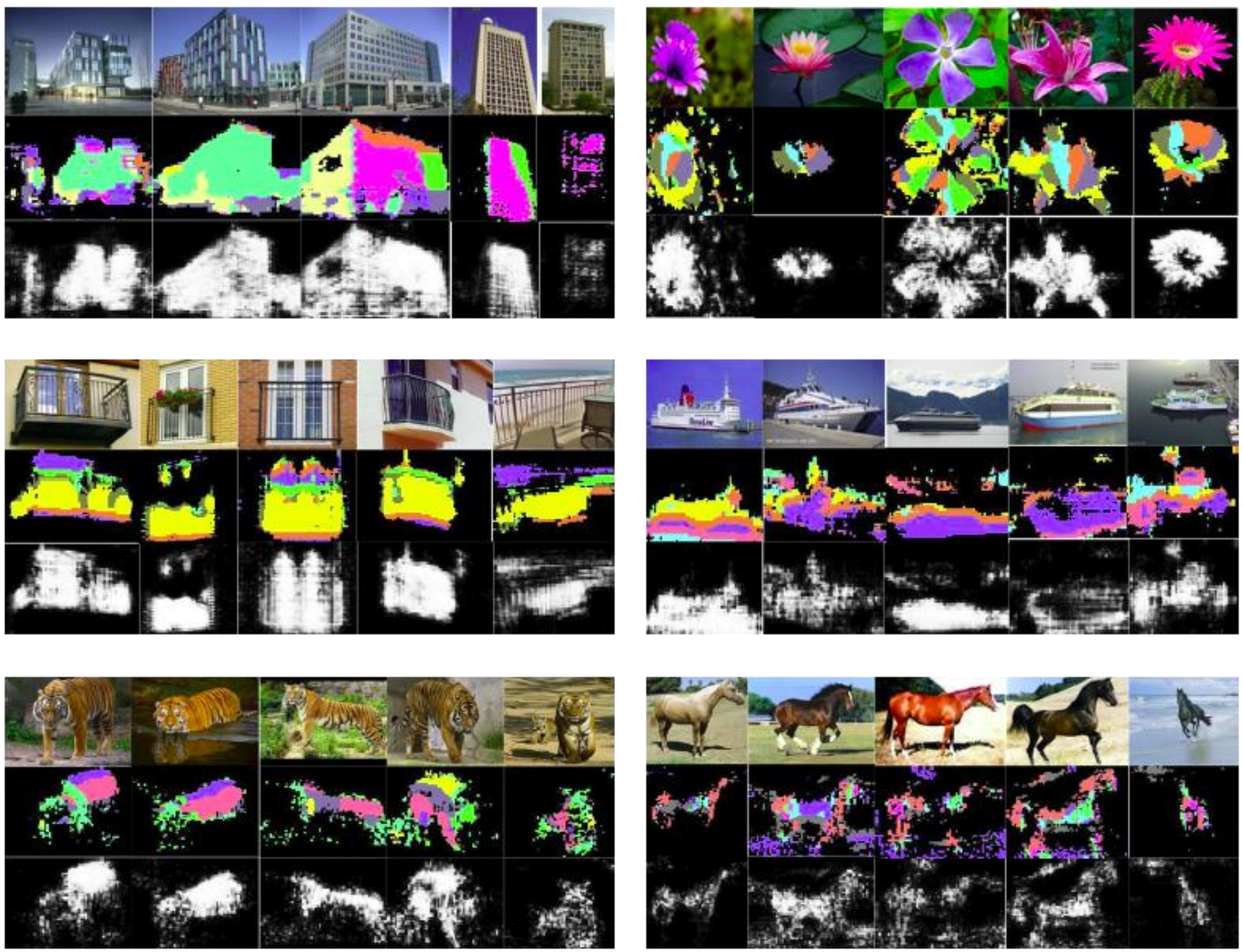




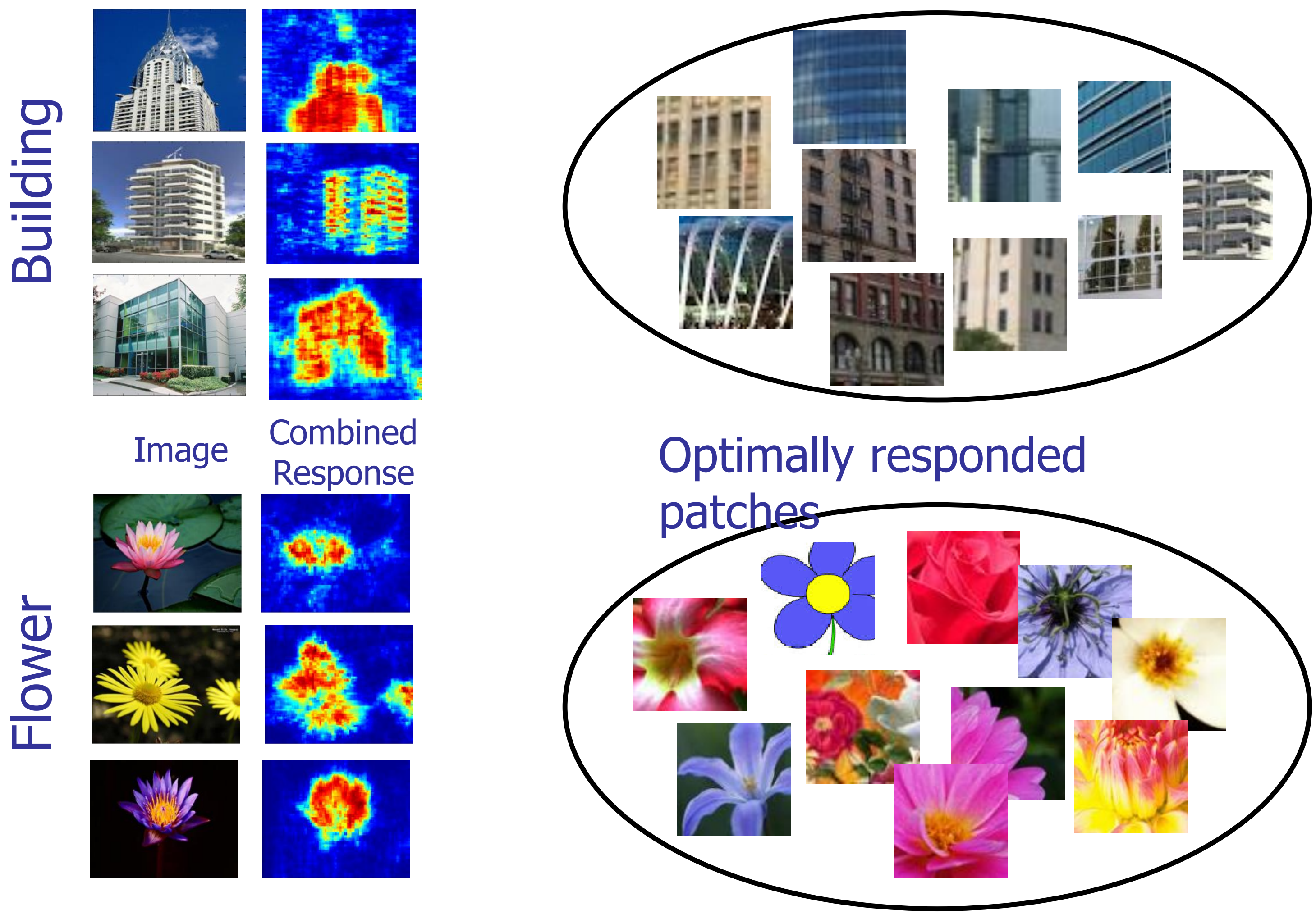


Kobatake and Tanaka, 1994

Learned response maps



Learned mid-level visual concepts

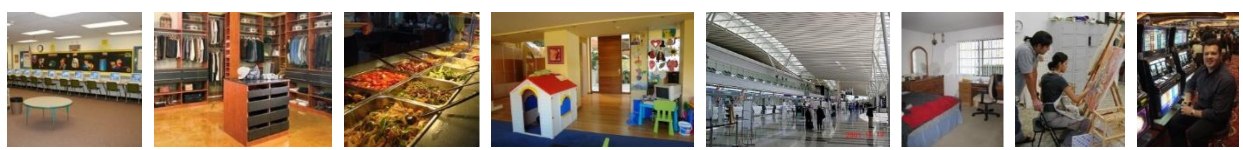


Classification using visual concepts

15 Scene

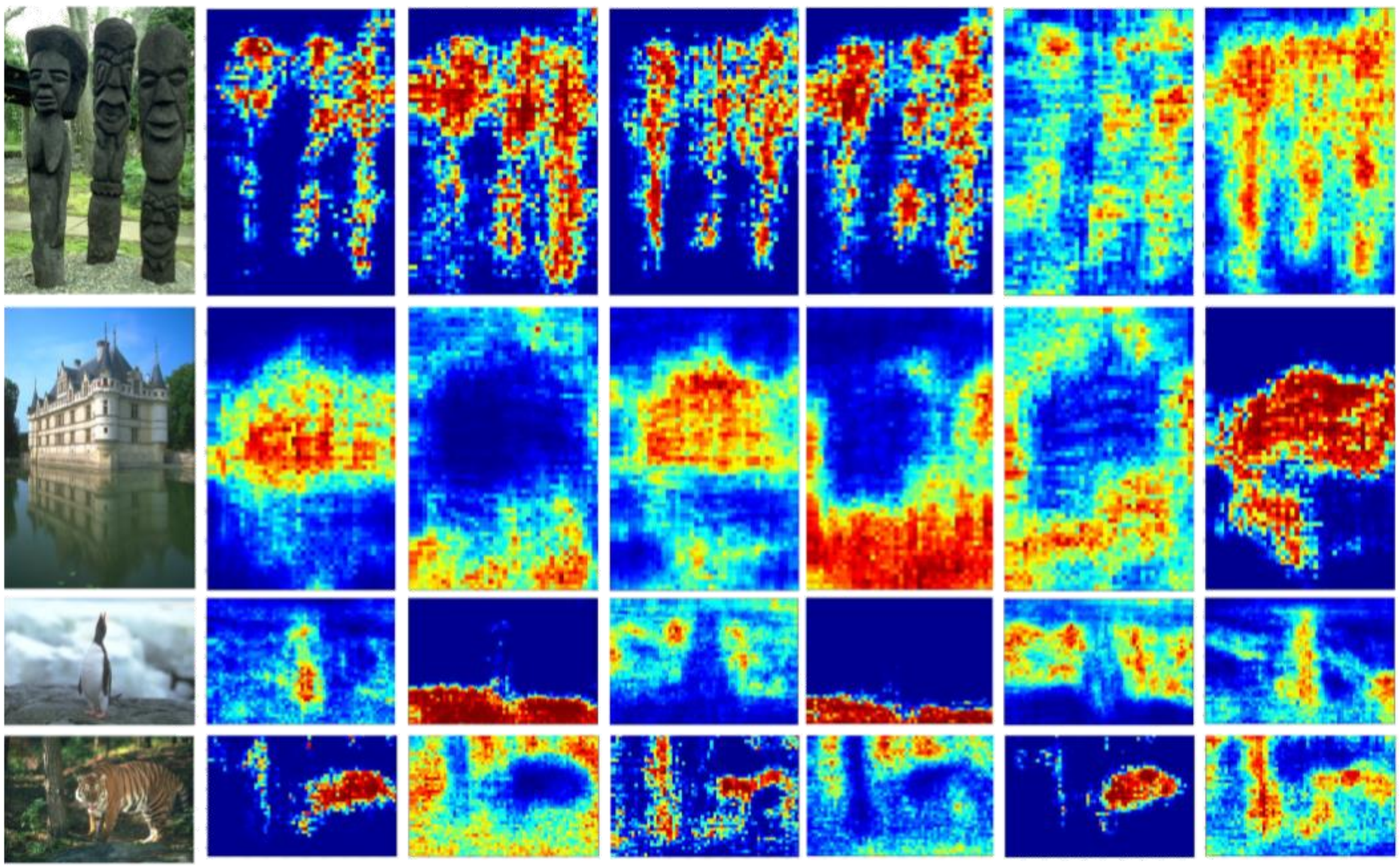
UIUC Sport

MIT Indoor Scene

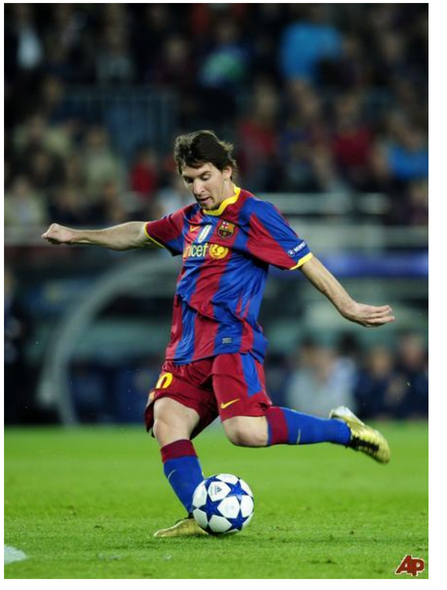


Pascal VOC 2007

Response maps of mid-level concepts



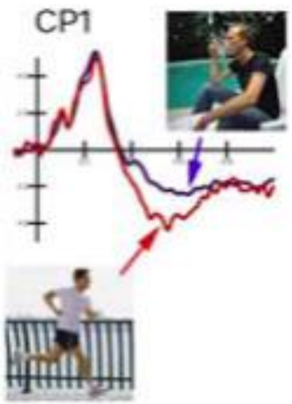
Extension

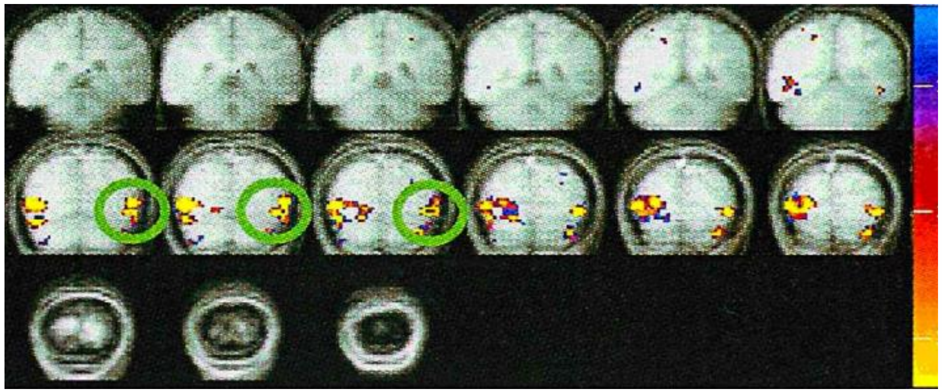


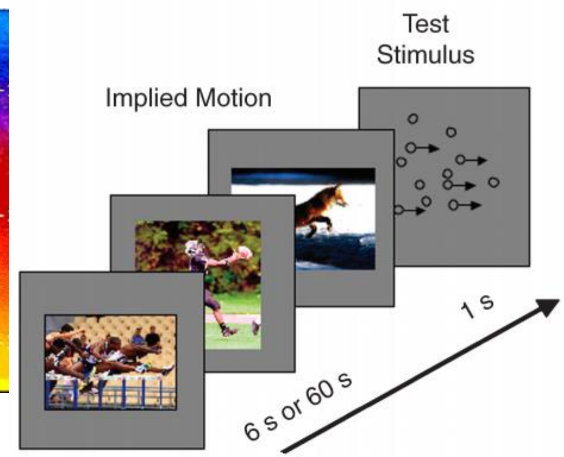
Objects
Nouns

Motions
Verbs

Views in cognitive science





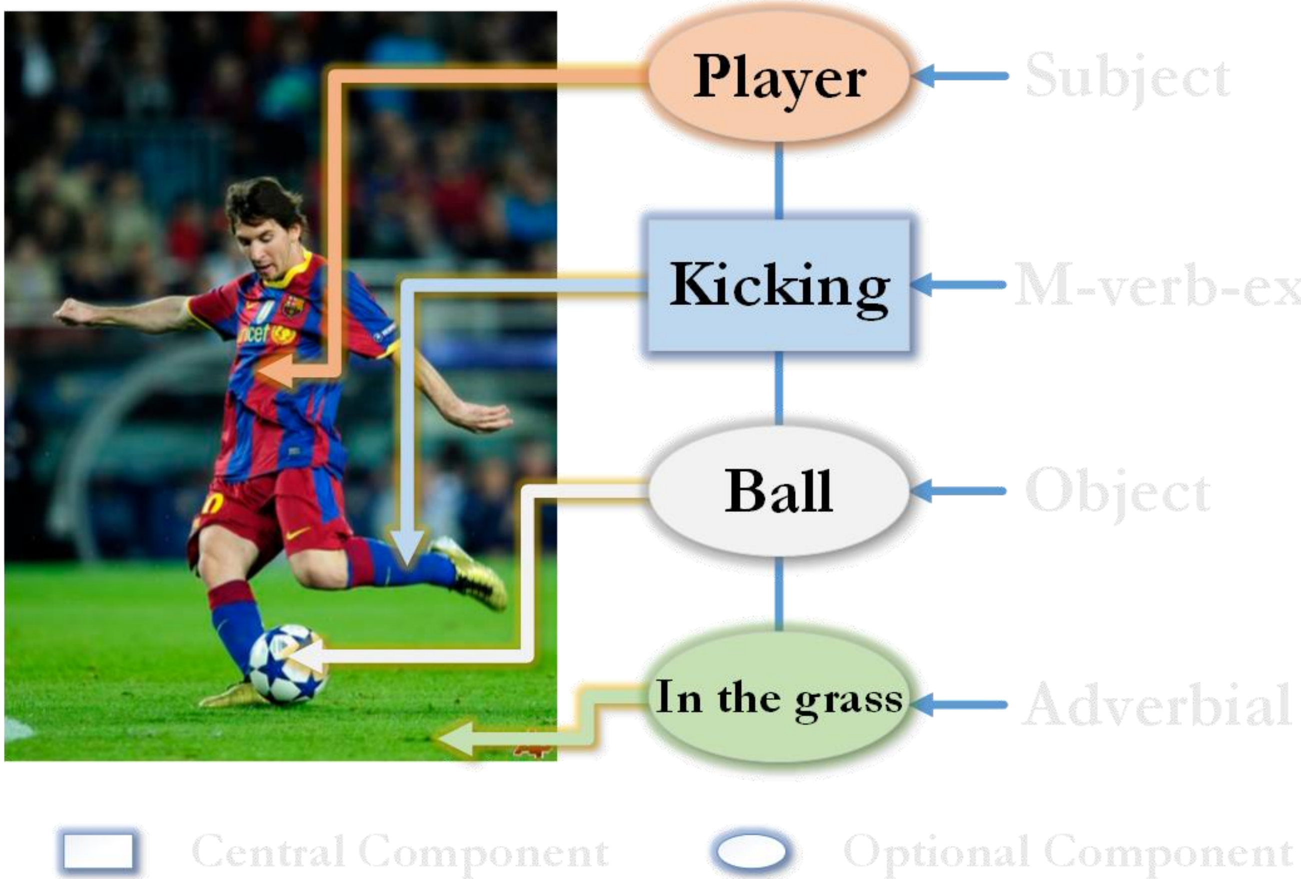


- Activation in human MT/MST
  - Kourtzi et al., J. Cog Neurosci, 2000, Proverbio et al., PLoS One, 2009
- A series of findings from Boroditsky
  - Still images of actions ⇔ human cognition ⇔ Visual imagery of motion ⇔ motion language, Psych Sci, 2008, Cognition, 2010, PNAS, 2010
- Experiments in Computer Vision (a MIL demonstration)

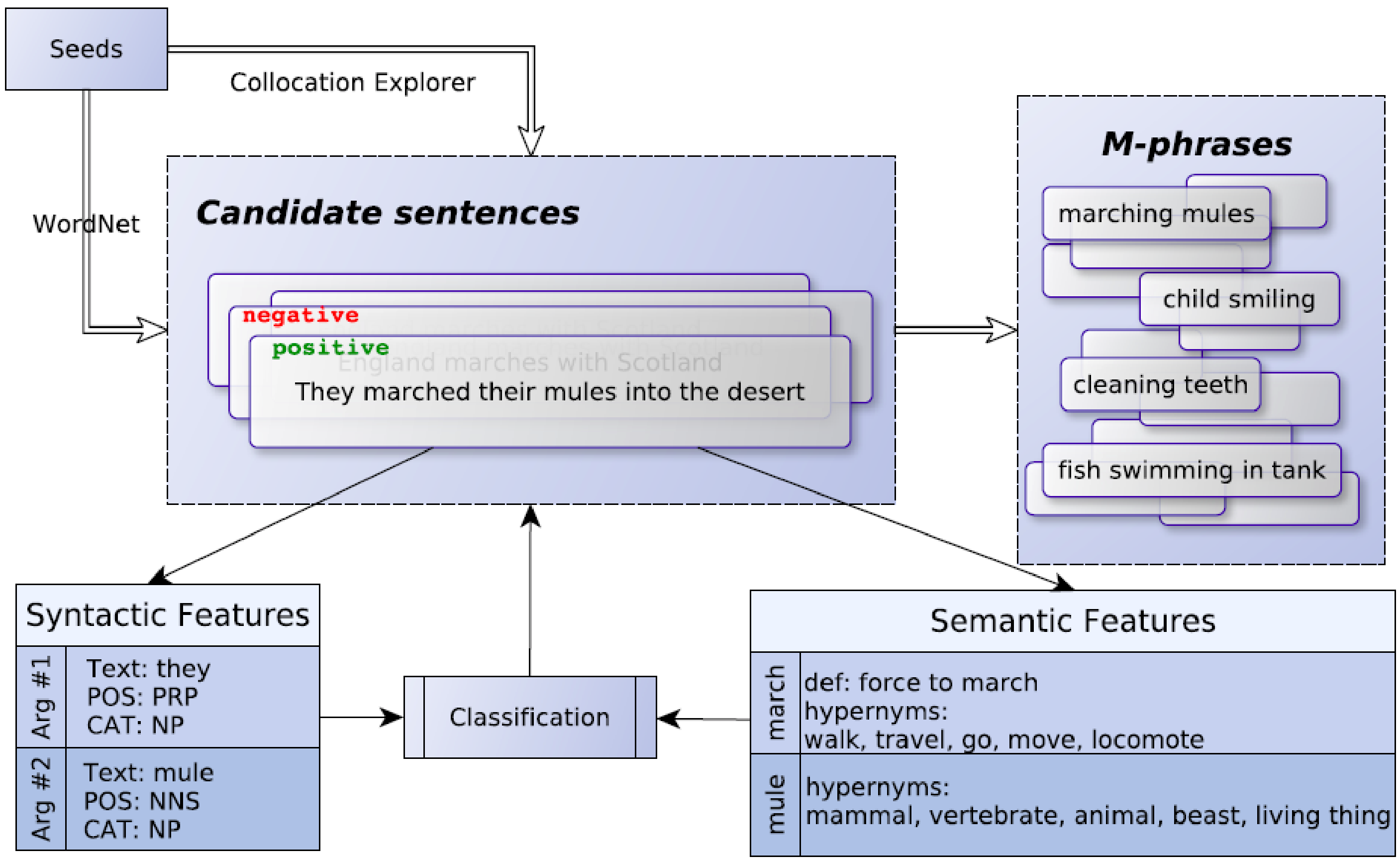
Action concepts from still images

We are crawling ~1000 action categories, e.g. *brushing teeth, bowling*, from Google and Bing image search engines.

Motion phrase



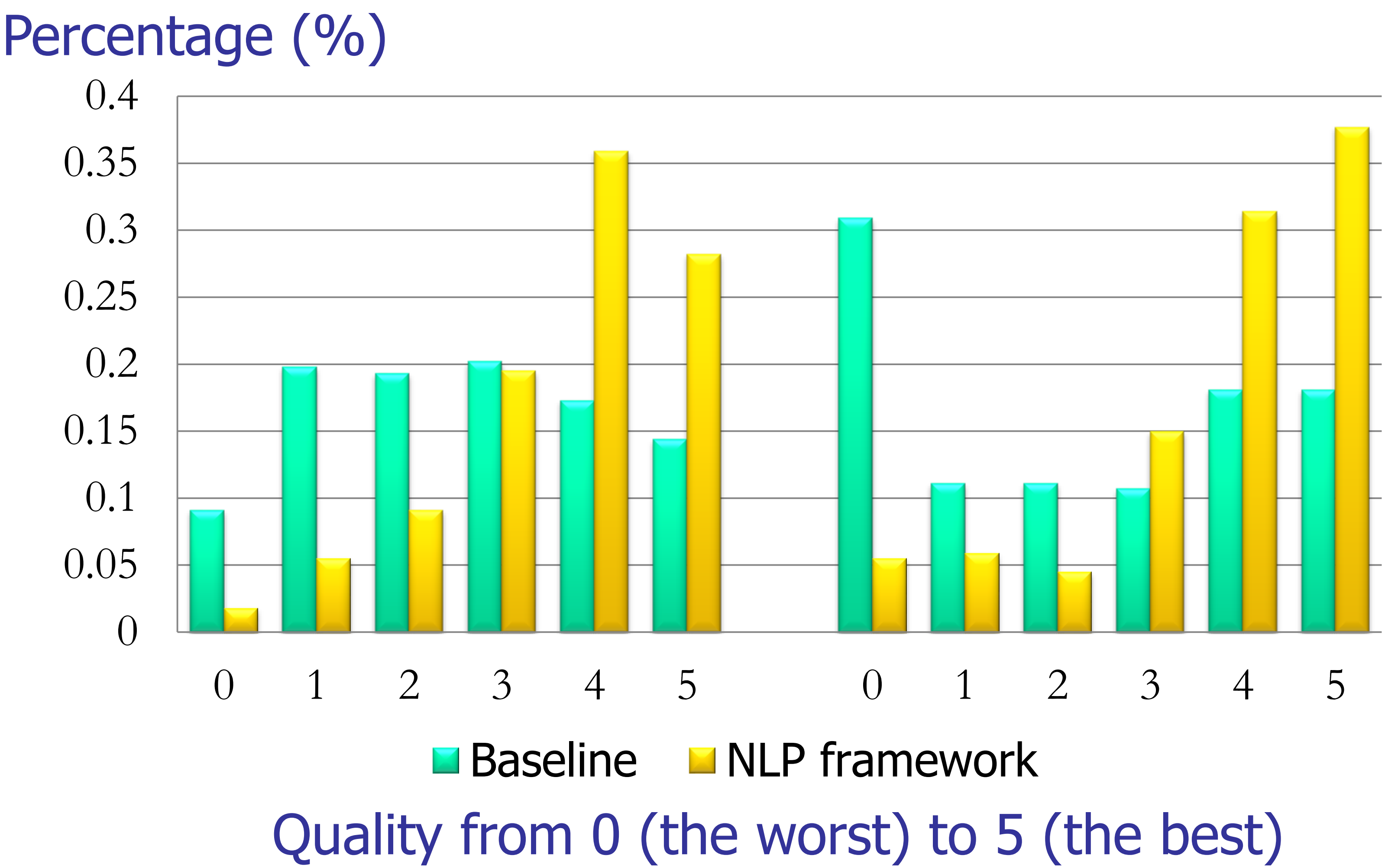
Expansion


Seeds
Collocation Explorer
WordNet
Candidate sentences
negative
positive
England marches with Scotland
They marched their mules into the desert
M-phrases
marching mules
child smiling
cleaning teeth
fish swimming in tank
Syntactic Features
Arg #1
Text: they
POS: PRP
CAT: NP
Arg #2
Text: mule
POS: NNS
CAT: NP
Classification
Semantic Features
march
def: force to march
hypernyms:
walk, travel, go, move, locomote
mule
hypernyms:
mammal, vertebrate, animal, beast, living thing

M-phrases

M-verb		S + M-verb-ex		M-verb-ex + O / Ad		S + M-verb-ex + O / Ad
crawling	throwing	whistle blowing	child running	raising hands	applying eye makeup	wind blowing leaf
marching	applauding	water flowing	man smoking	pushing against wall	applying lipstick	boat driting on water
brushing	diving	leaf swirling	fish swimming	delivering ball	blowing dry hair	fish swimming in tank
pushing	walking	cat running	kid skiing	brushing hair	blowing bubbles	feather drifting past window
cycling	smiling	dog barking	woman smoking	cooking dinner	blowing candles	dog licking hand
jogging	dancing	man sailing	baby crawling	lifting box	brushing teeth	bird clapping wing
archering	dunking	military marching	band playing	clibming rock	cutting trees	face being angry
bowling	drinking	car running	baby wailing	closing eyes	raising eyebrows	face being disgusted
boxing	fishing	child clinging	child writing	fixing car	fixing bike	face being surprised
kayaking	bathing	dog baying	dog eating	playing badminton	playing football	dentish cleaning tooth
coughing	decanting	fish swimming	girl dancing	playing guitar	playing cello	people crowding street
dabbling	harvesting	girl walking	girl pouting	ascending mountain	assembling car	parent protecting child
refueling	spinning	man leaping	man sitting	bonding with child	brushing wall	pitcher delivering ball
spitting	telephoning	potato sprouting	train derailing	cheering child	conditioning hair	squirrel learping from tree
undressing	yelling	tree swaying	water bubbling	cleaning fingernail	cleaning stove	smoke rising from fire
dancing	crawling	water pouring	woman biting	disciplining child	drinking soda	spider spinning web
kneading	hugging	balloon popping	child bathing	feeding child	filleting fish	teacher teaching child
jibing	sledging	dog barking	child coloring	holding bowl	holding nose	veteran prading street
parying	quarreling	dog snapping	bomb exploding	harpooning whales	jumping over fence	wind blowing leaf
injecting	leaping	mammal predating	hair greying	riding camel	riding motorcycle	lightning striking tree
rushing	roaring	hair falling	horse galloping	skating along canal	veiling face	parent disciplining child
shampooing	shaving	woman nibbling	ship sinking	shuffling card	sifting flour	bird perching in tree
photographing	melting	woman jumping	infant suckling	raising hand	spiking volleyball	wind howling in tree
mining	migrating	whale blowing	snake swimming	riding bull	playing table tennis	mushroom growing under tree
nibbling	mopping	patient walking	flower withering	cutting vegetables	driving car	face being shy
chanting	bullying	smoke rising	sheep eating	clapping hands	decanting wine	caterpillar feeding on leaf

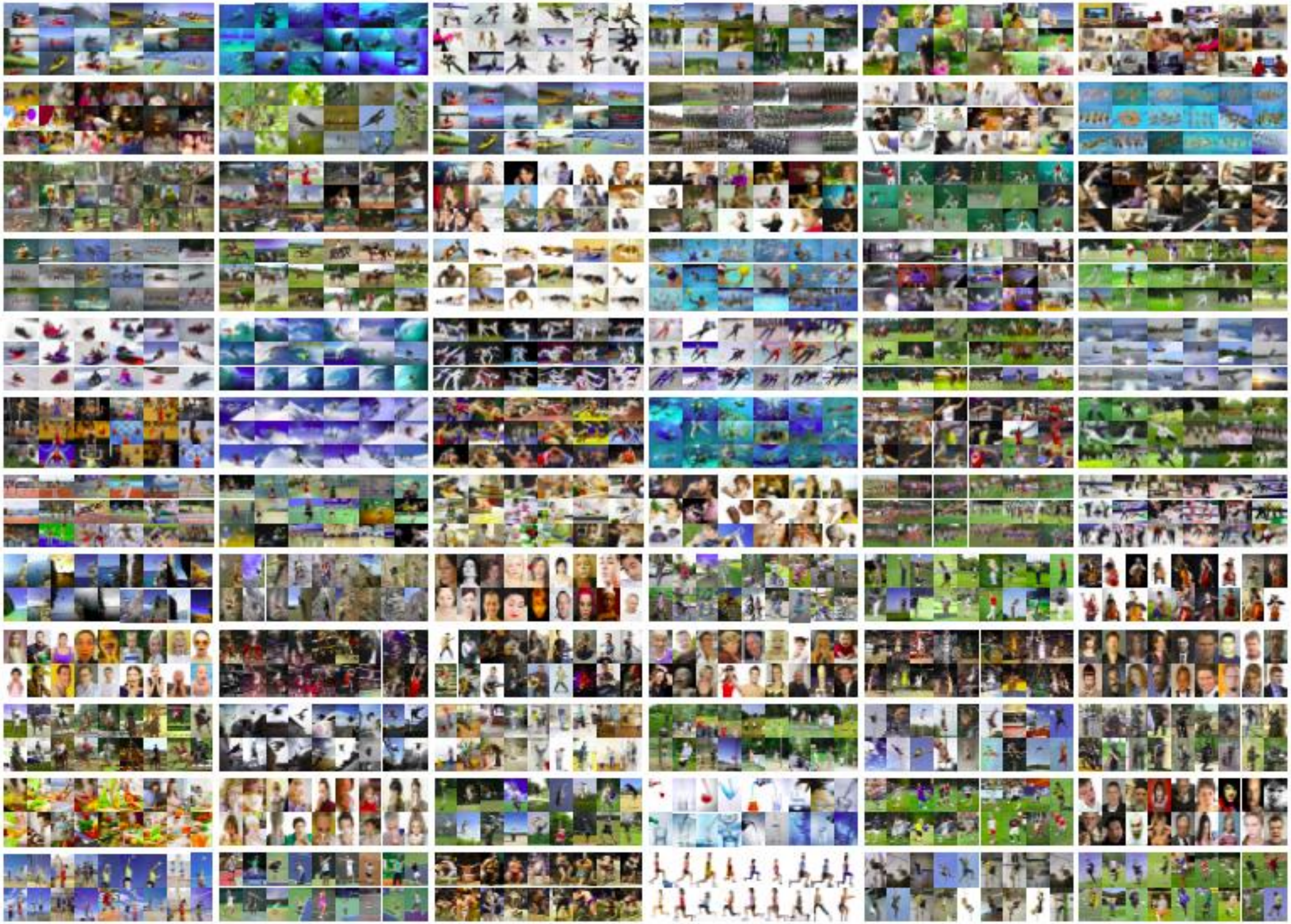
Distribution of images



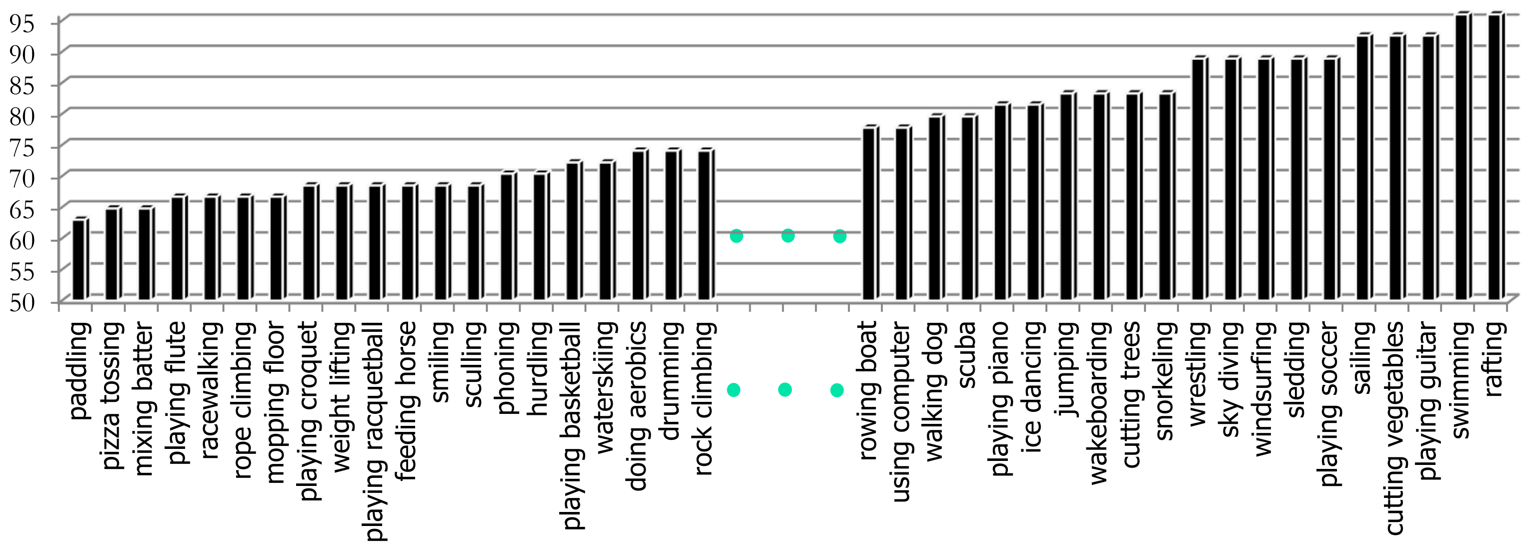
Motions in still images

1,024 categories of motions from Google and Bing

UCSD-1024



Inner-category consistency



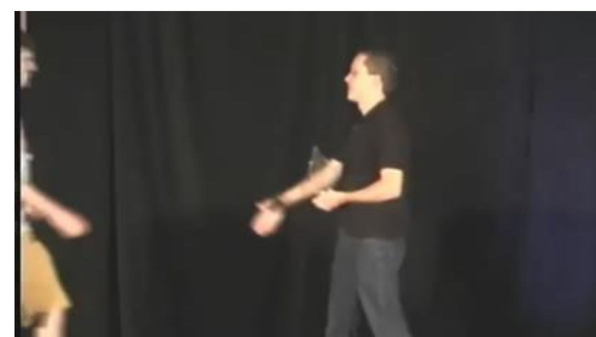
Recognizing human actions in videos





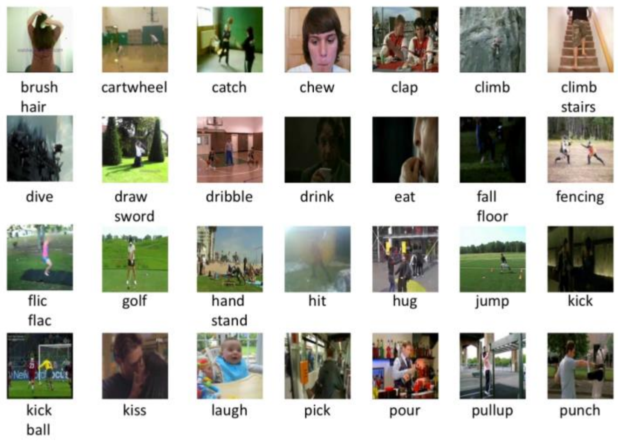




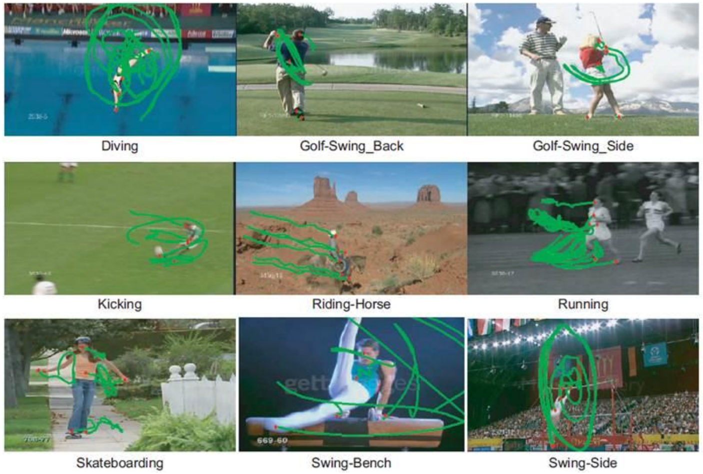


Background & applications

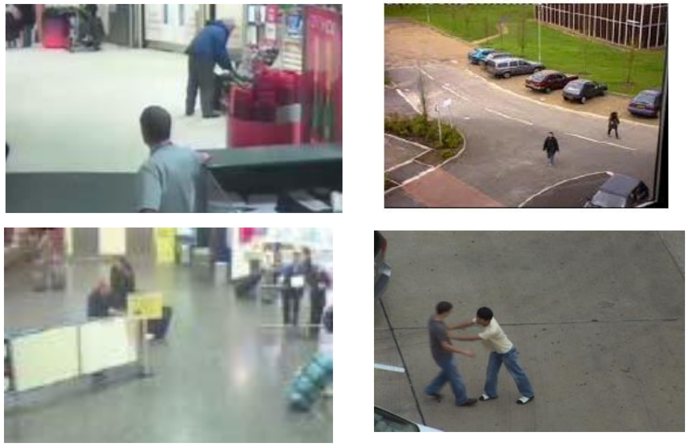
Content-based video



Sports video analysis

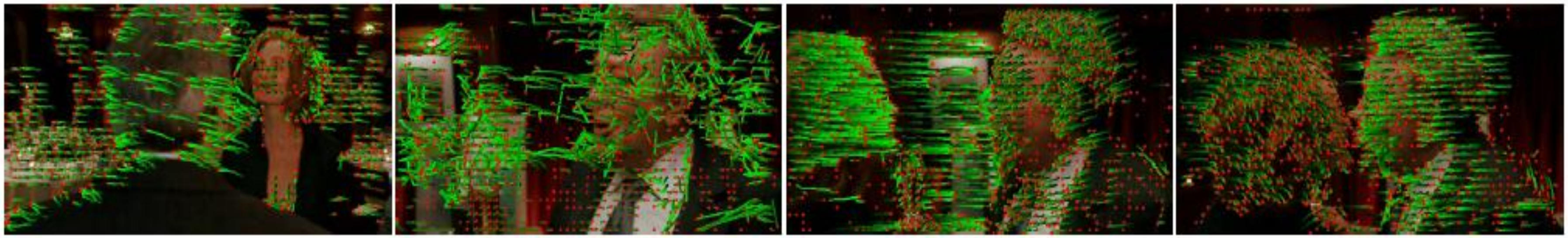


Surveillance event detection

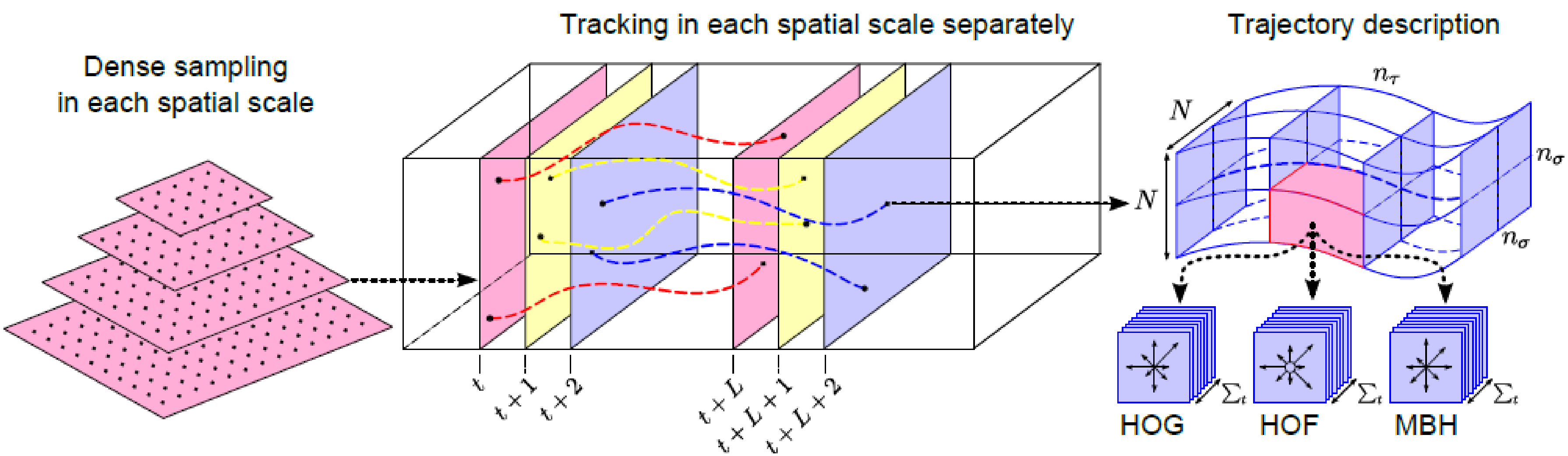


Human-machine interface / Gaming

Spatial-temporal video features by dense trajectory

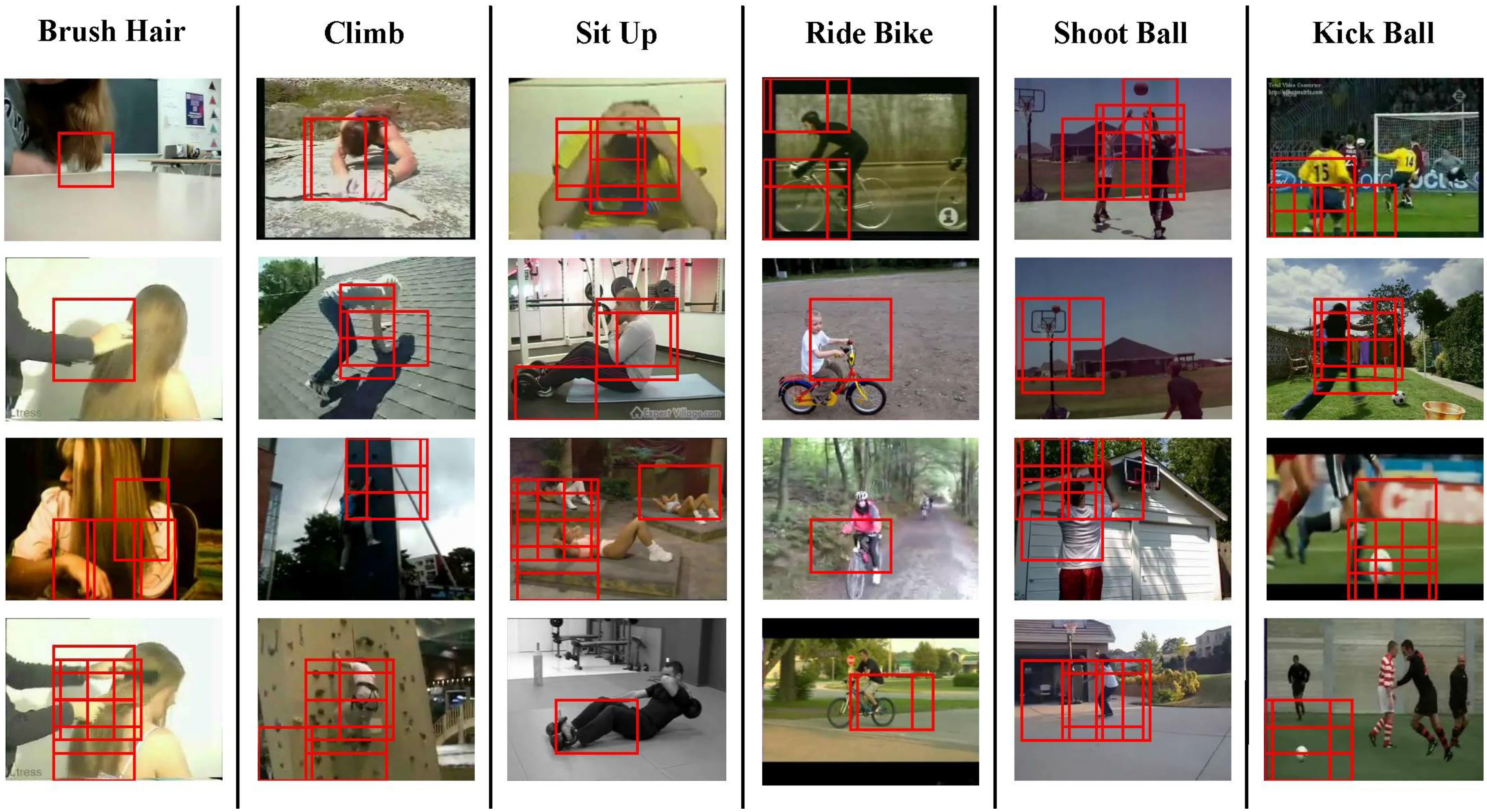


Dense trajectories

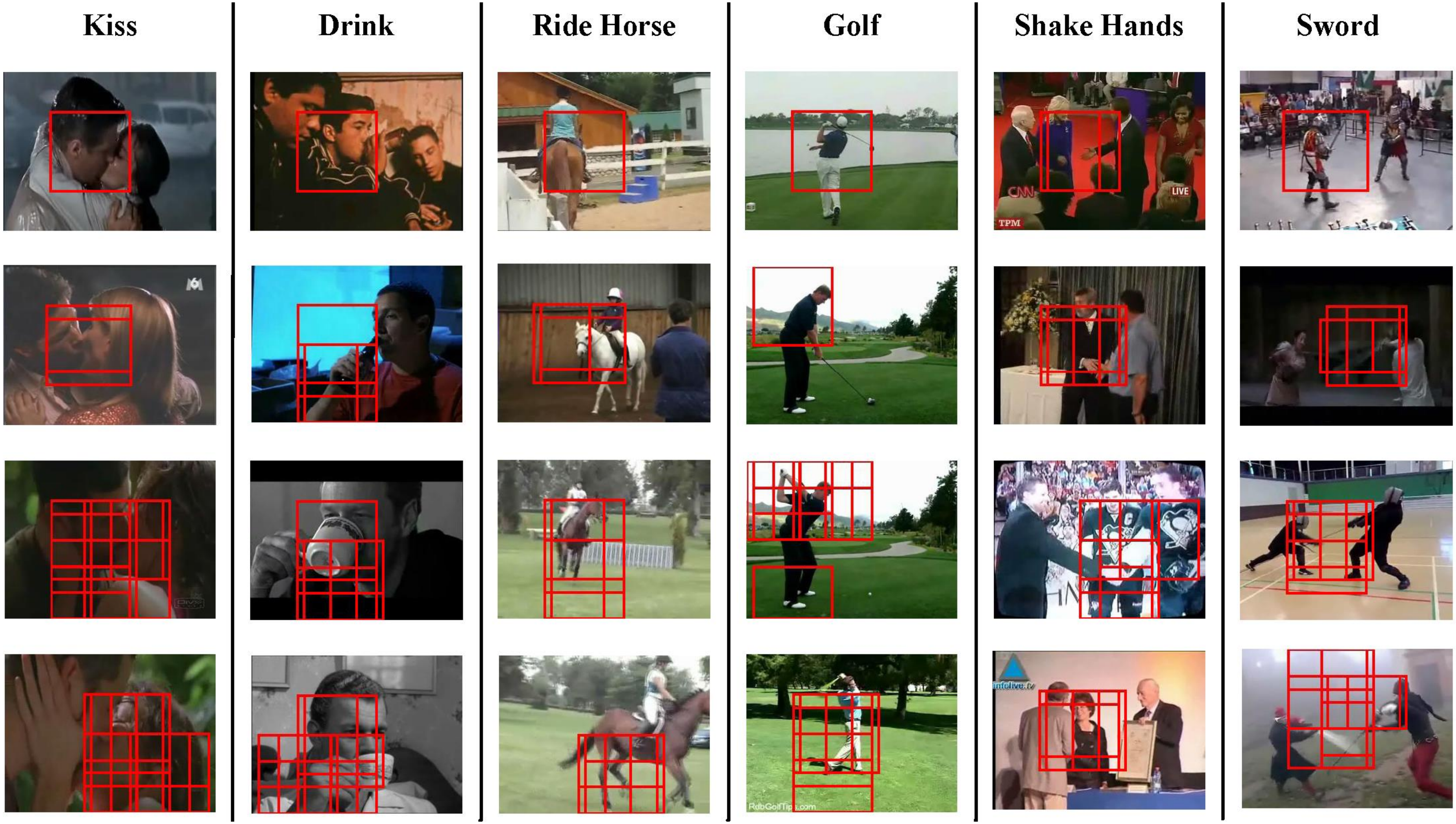


Wang et al. 2013

Learned video patches from action video clips



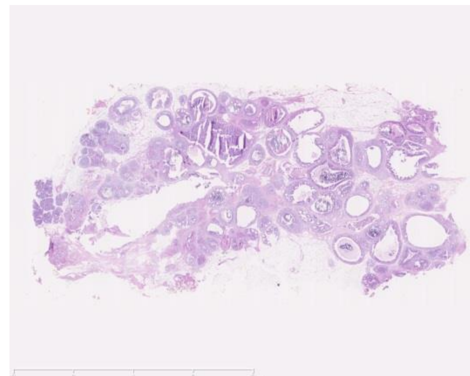
Learned video patches from action video clips

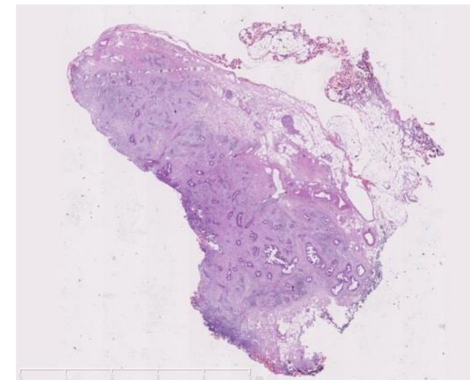


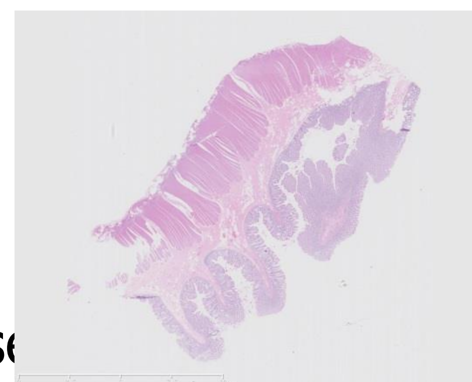
Weakly-Supervised Learning for Microscopic Image Segmentation, Clustering, and Classification

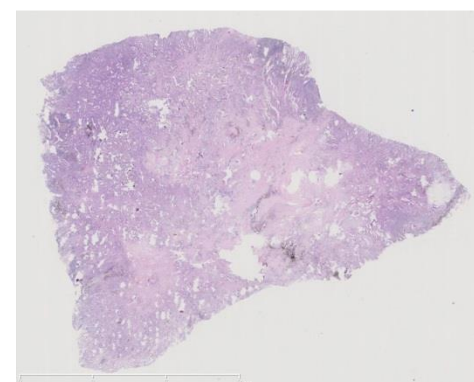
- Colon cancer
- Lung cancer
- Liver cancer
- Breast cancer
- Nasopharyngeal cancer
- Kidney cancer
- Esophagus cancer
- Gastric cancer

2000 pathology reports of colon cancer including disease information and image information









Weakly-Supervised Learning

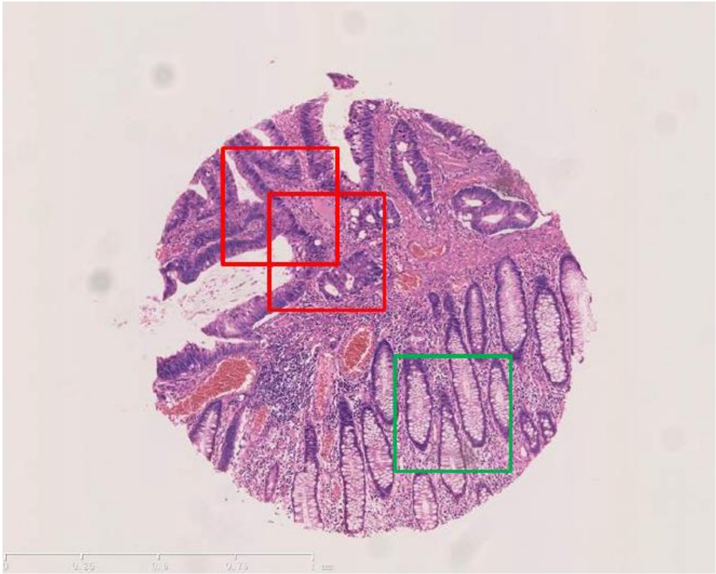
1. It is relatively easy to identify cancer/non-cancer histopathology images.

2. The detailed segmentation however requires careful manual annotations.

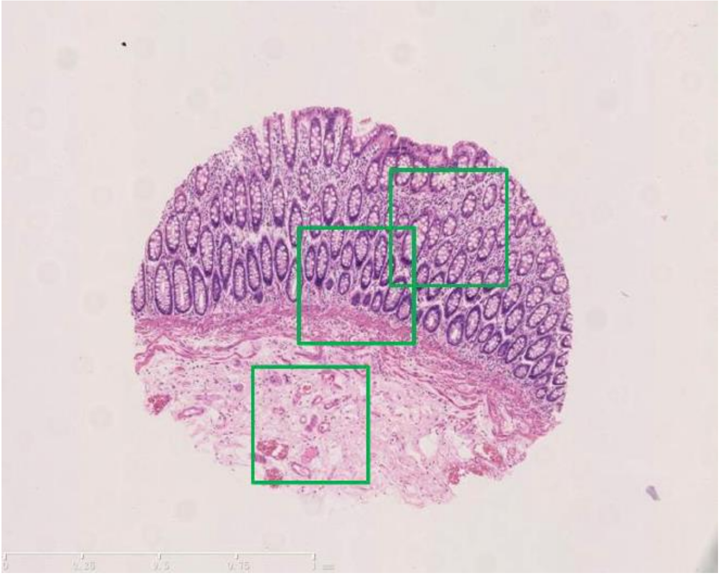
3. It is an ambiguous task to identify/recognize the subclasses of the cancer type.

Histopathology Images (extremely large: around 1TB per image)

Motivation for Weakly-Supervised Learning



Cancer histopathology image

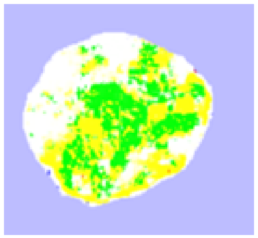


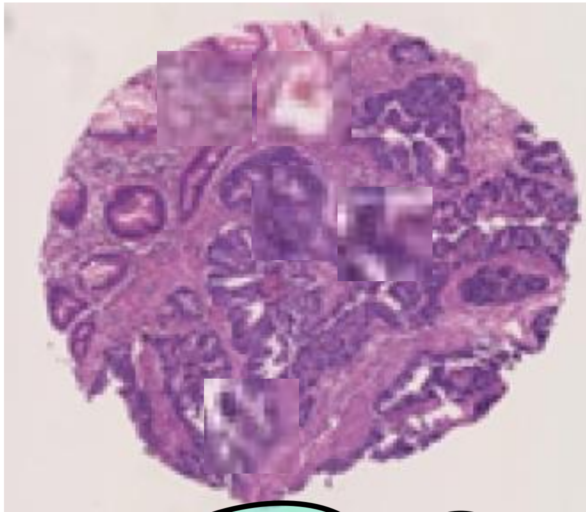
Non-cancer histopathology image

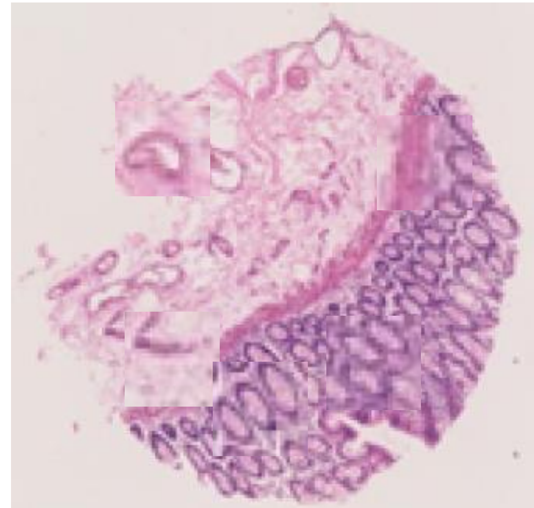
Motivation for Weakly-Supervised Learning

Cancer Image

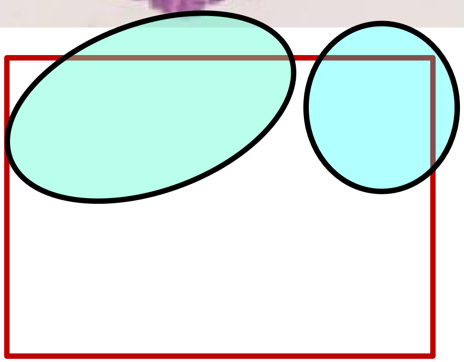
Non-cancer Image







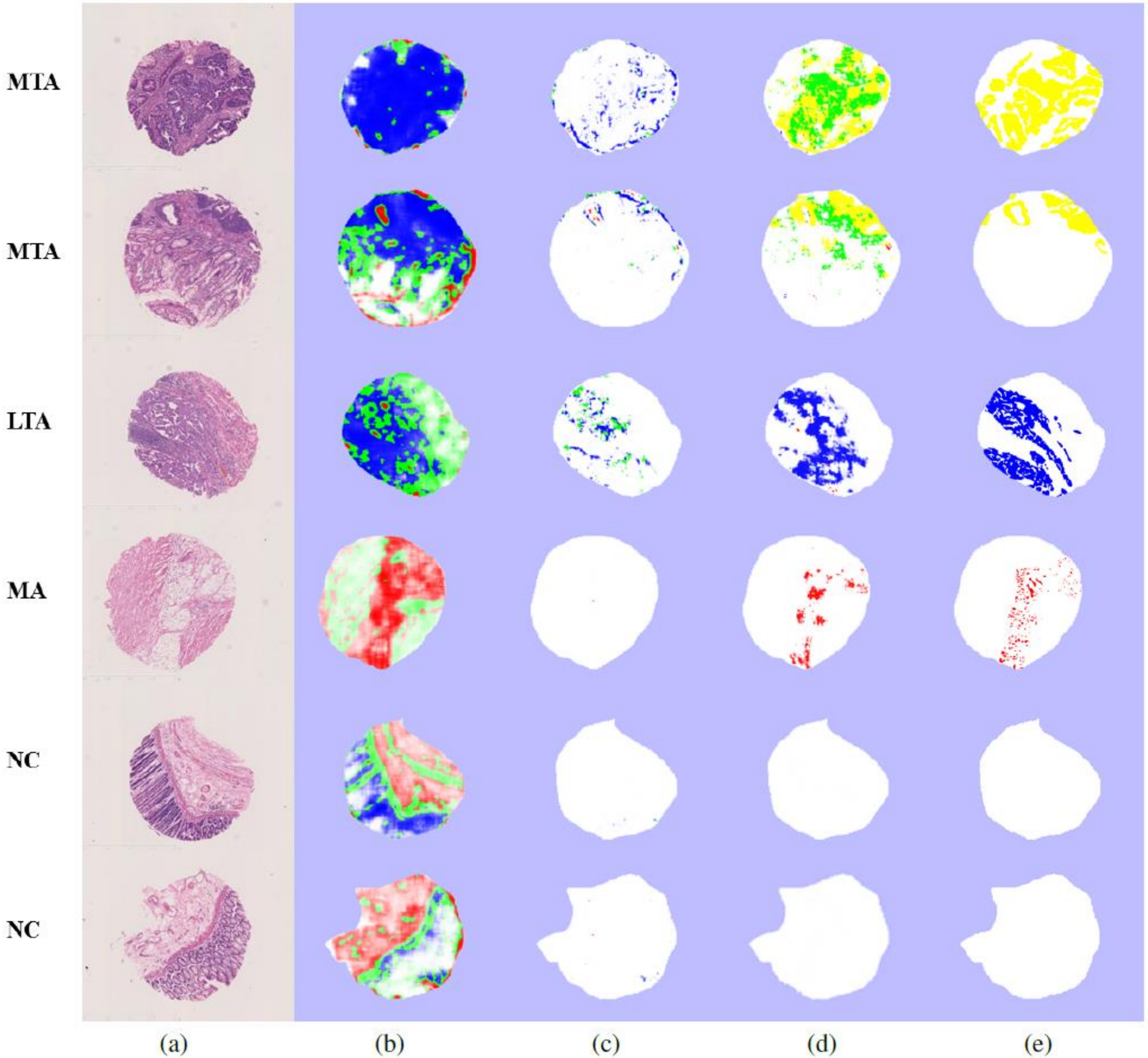
Positive bags



Negative bags

An integrated formulation to perform pixel-level segmentation, patch-level clustering, and image-level classification with image-level labels as supervision, Multiple Clustered Instance Learning (Xu et al. cvpr 2012, Xu et al. MICCAI 2012).

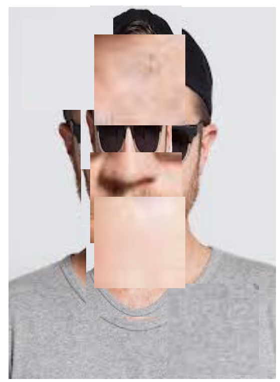
Results- Test Images (Xu et al. CVPR 2012)



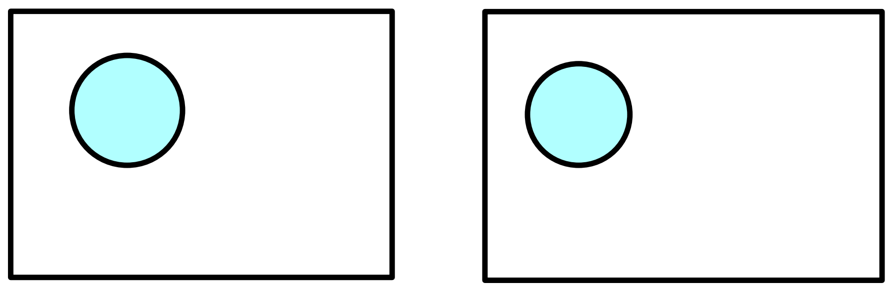
(a): The original images.
(b): The pixel-level segmentation and clustering for standard Boosting + K-means
(c): MIL + K-means, and our MCIL.
(d): MCIL
(e): The instance-level ground truth labeled by three pathologists.

Unsupervised object discovery

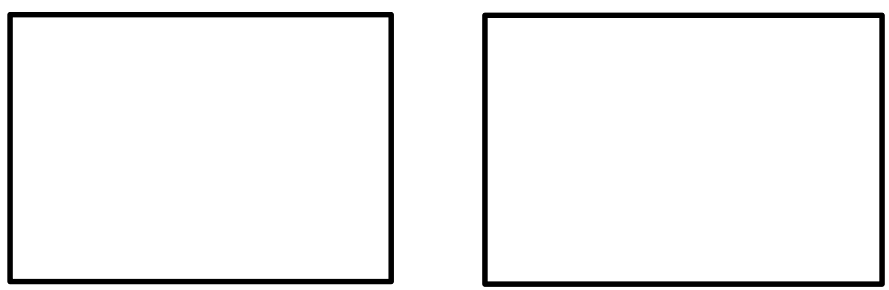
Illustration



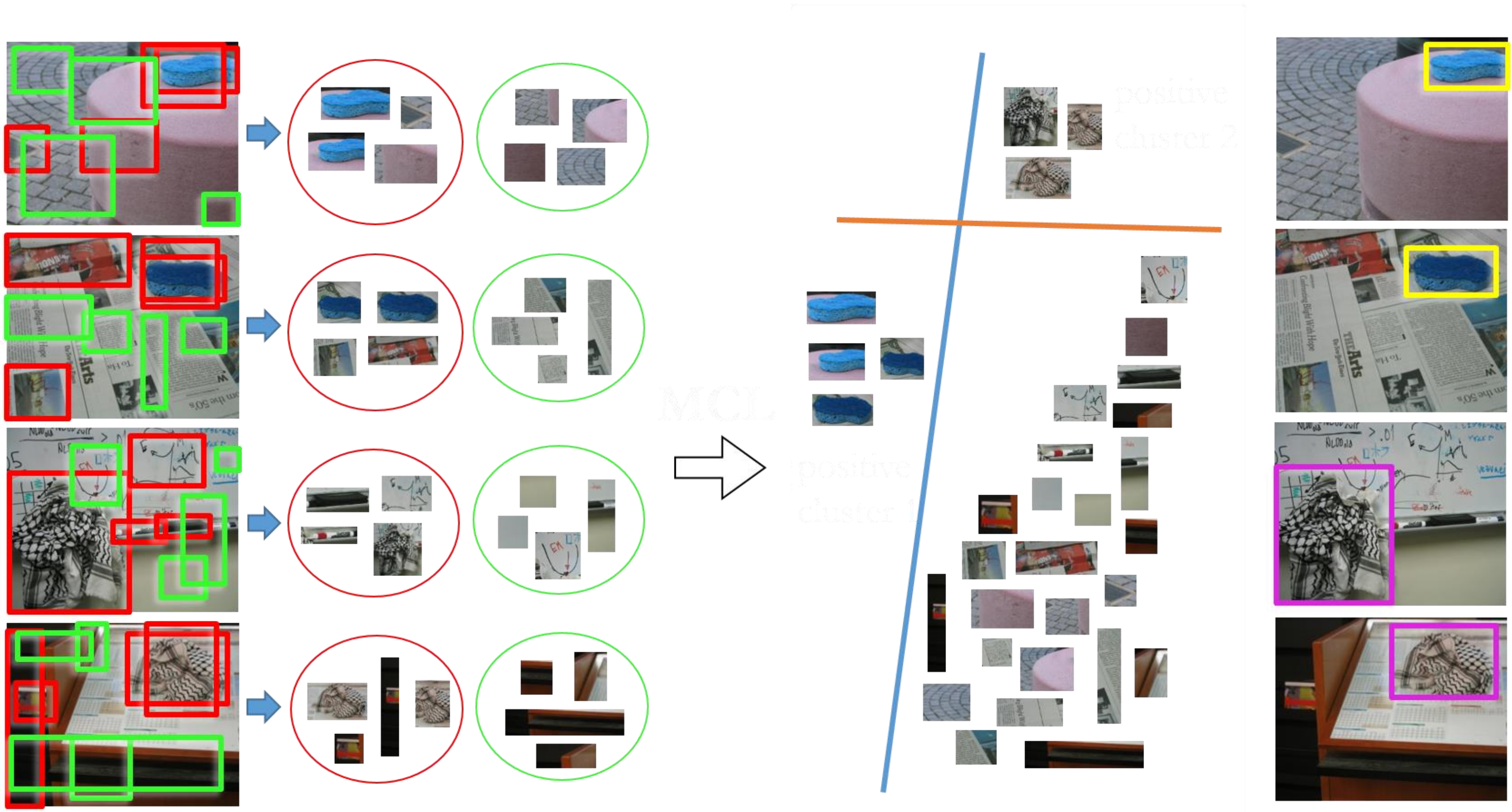
Positive bags



Negative bags

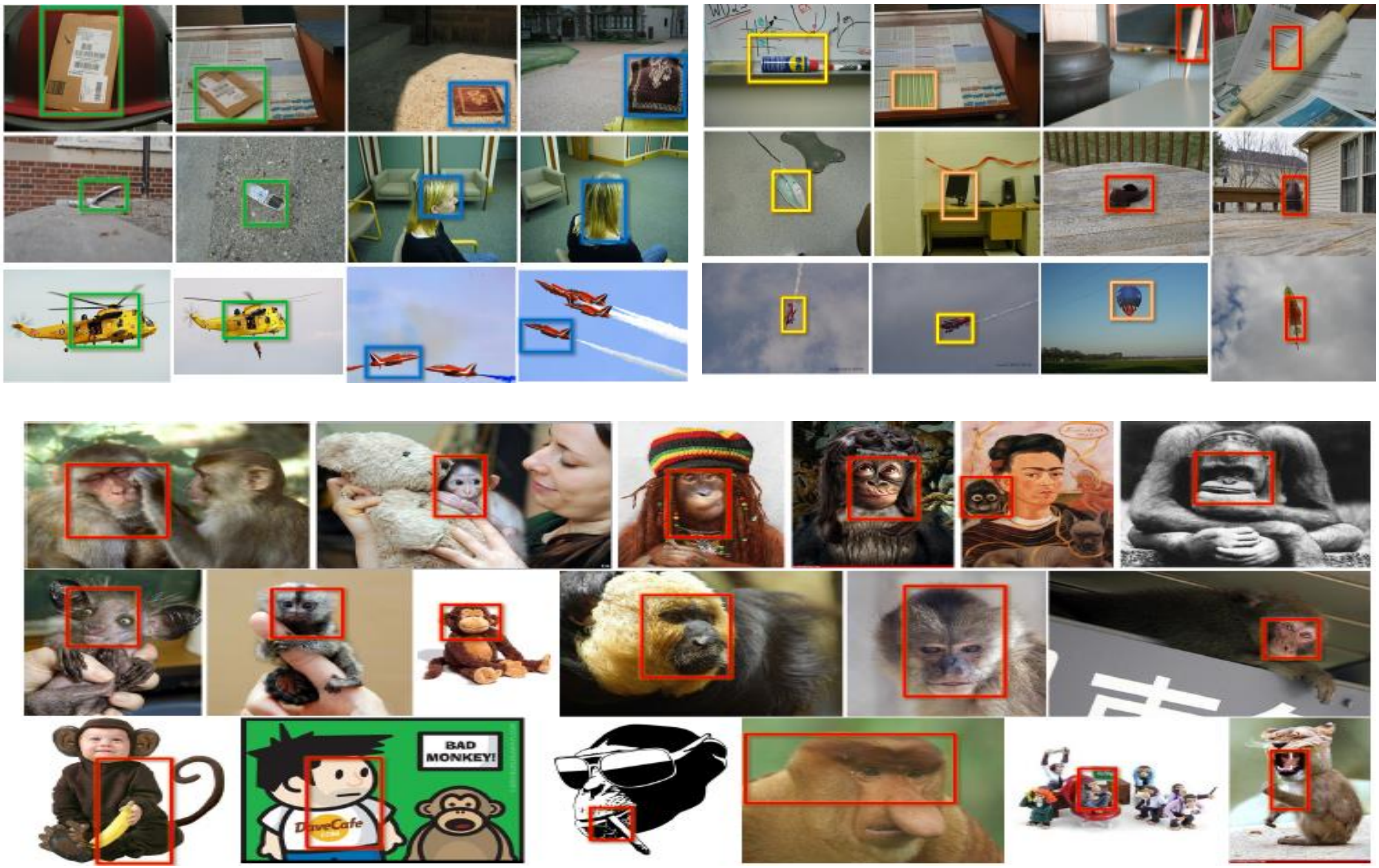


Bottom-up multiple class learning



Zhu et al., CVPR 2012

Object discovery results

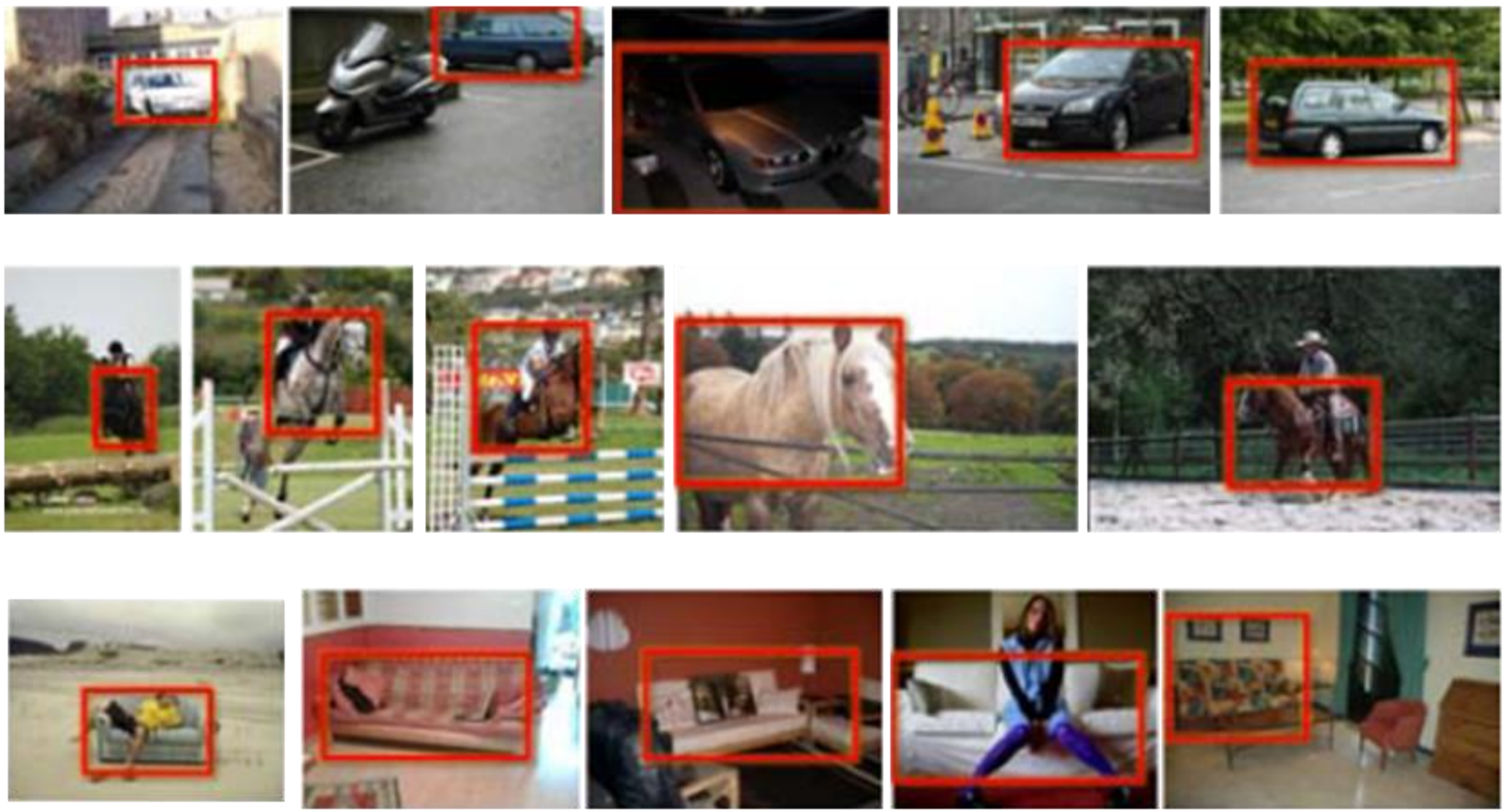


Weakly supervised modeling of single object class

	Ours	[32]	[11]	[9]	[41]	[38]	[25]
PASCAL 06-subset	45	36	**49**	34	27	N/A	43
PASCAL 07-subset	**31**	25	28	19	14	30	30

- [32] Leistner, et al. ECCV 11′
- [11] Deselaers et al. IJCV 12′
- [9] Chum Zisserman. CVPR 07′
- [41] Russell et al. CVPR 06′
- [38] Pandey and Lazebnik. ICCV 11′
- [25] Joulin et al. CVPR 12′

PASCAL results:



Unsupervised object discovery-a low-rank approach (Wang et al. 2014)

Previous work

RPCA: E. Candes, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? Journal of the ACM, 58(3), May 2011.

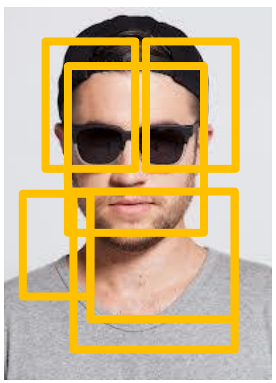
RASL: Y. Peng, A. Ganesh, J. Wright, W. Xu, and Y. Ma, RASL: Robust Alignment by Sparse and Low-rank Decomposition for Linearly Correlated Images, PAMI 2011.

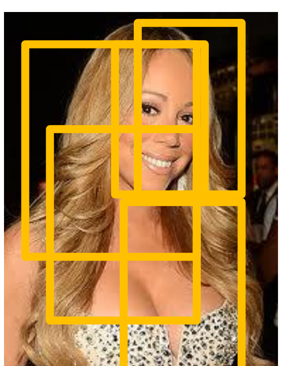
$$\min_{A,E,Z} ||A||_* + \gamma ||E||_1 \; s.t. \; D \cdot \tau = A + E$$

A Low-rank solution

$$\min_{A,E,Z} rank(A) + \gamma ||E||_0 \; s.t. \; X \cdot diag(Z) = A + E, \forall k \in [K] \bigcup_{i=1}^{n_k} z_i^k = 1$$

X Z A E



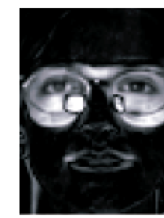


1
0
0
0

0
1
0
0

=

+



Relaxing the conditions

$$\min_{A,E,Z} rank(A) + \gamma ||E||_0 \; s.t. \, X \; diag(Z) = A + E, \forall k \in [K] \bigcup_{i=1}^{n_k} z_i^k = 1$$

$$\downarrow$$

$$\min_{A,E,Z} ||A||_* + \gamma ||E||_1 \; s.t. \, X \; diag(Z) = A + E, \forall k \in [K] \bigcup_{i=1}^{n_k} z_i^k = 1$$

$$\downarrow$$

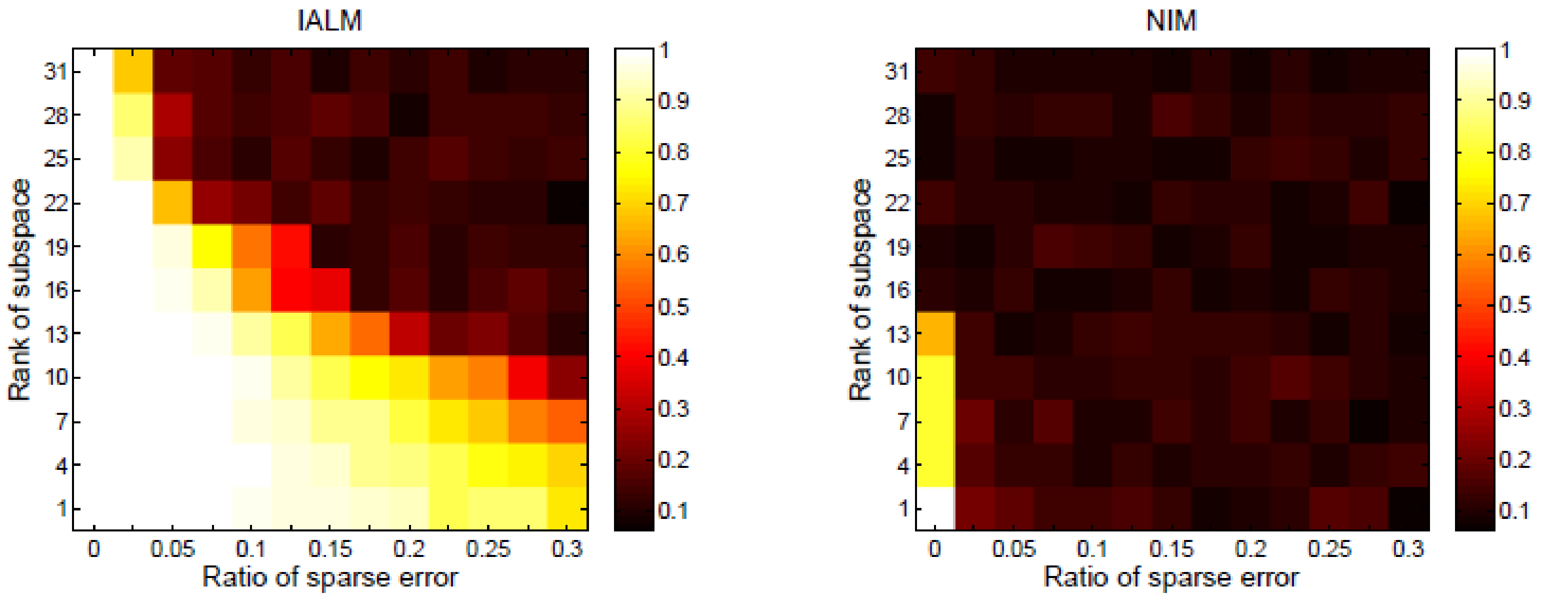
$$\min_{A,E,Z} ||A||_* + \gamma ||E||_1 \; s.t. \, X \; diag(Z) = A + E, \forall k \in [K] 1^T Z^{(k)} = 1$$

Now a convex optimization which can be solved by e.g. Inexact Augmented Lagrange Multiplier.

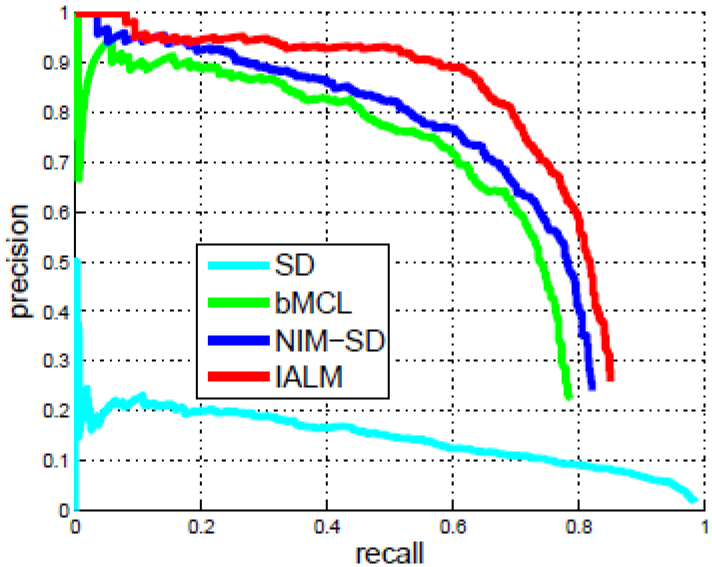
Inexact augmented Lagrange multiplier

$$L(A, E, Z, Y_0, Y_1, \ldots, Y_K) \doteq \|A\|_* + \lambda \|E\|_1 + \langle Y_0, X \operatorname{diag}(Z) - A - E \rangle + \frac{\mu}{2} \|X \operatorname{diag}(Z) - A - E\|_F^2$$

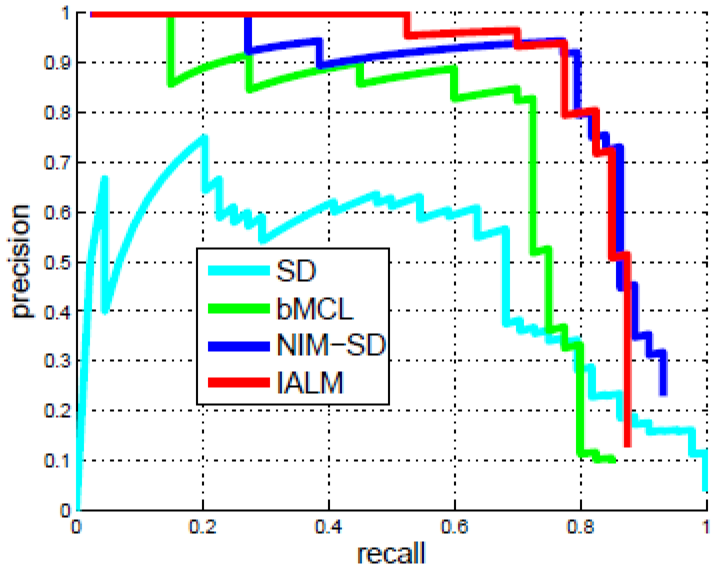
$$\ldots + \sum_{k=1}^{K} \left(\langle Y_k, \mathbf{1}^T Z^{(k)} - 1 \rangle + \frac{\mu}{2} \|\mathbf{1}^T Z^{(k)} - 1\|_F^2 \right).$$



Results

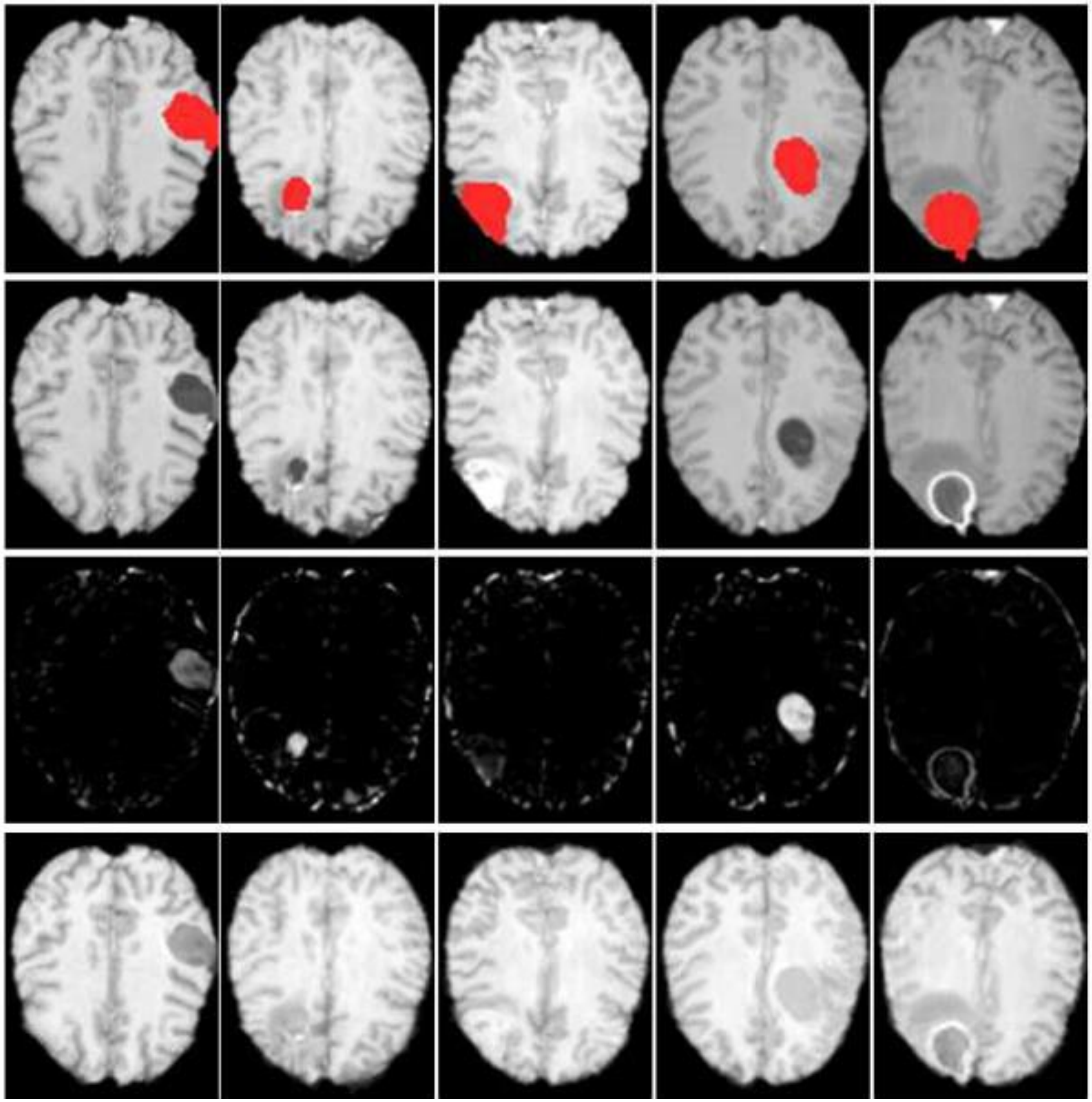


precision
recall
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
0.2
0.4
0.6
0.8
1
SD
bMCL
NIM−SD
IALM

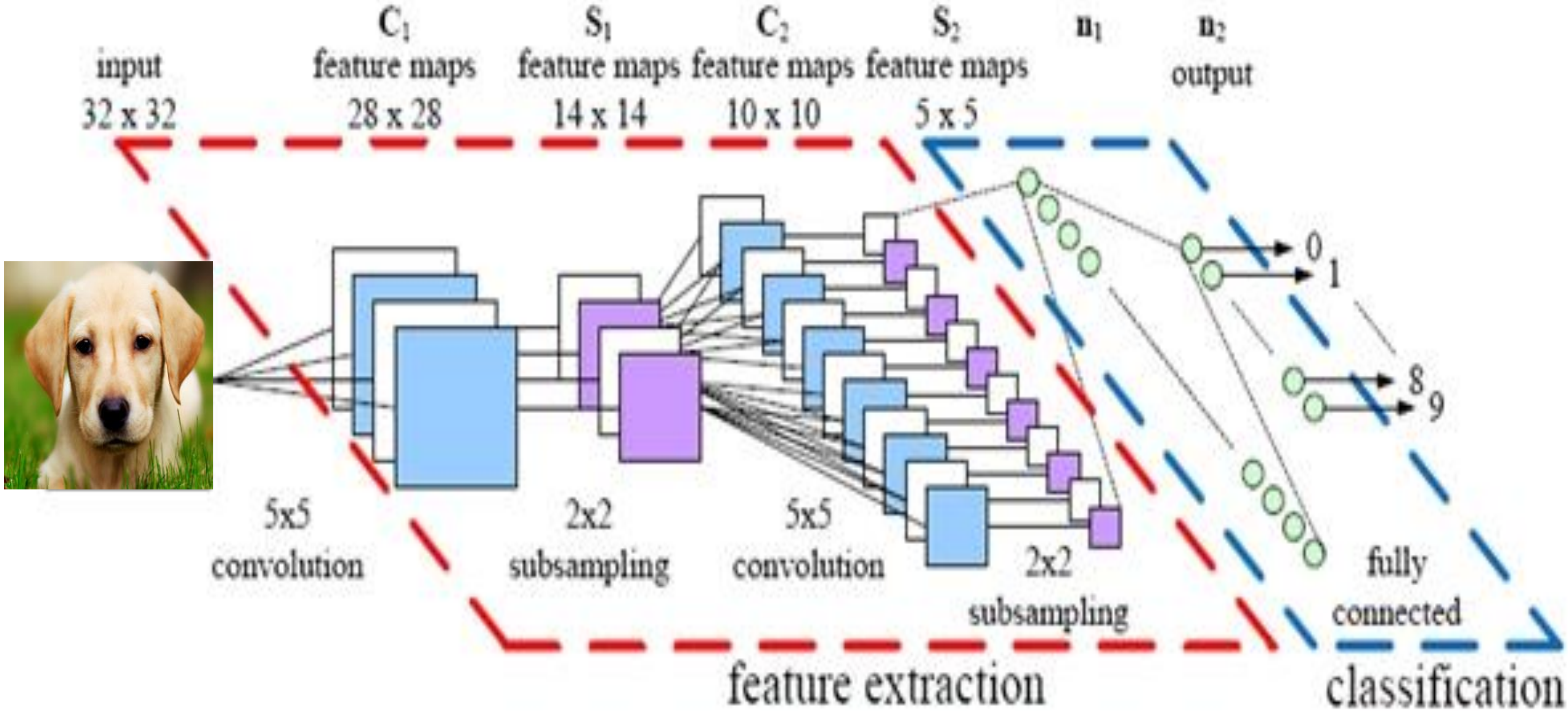


precision
recall
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
0.2
0.4
0.6
0.8
1
SD
bMCL
NIM−SD
IALM

MRF tumor discovery



Connection with deep learning



LeCun,et al.

Conclusion

- There are rich mathematical/statistical/computational models which become increasingly convenient to use.
- The availability of ever increasing data cohort provides a golden opportunity to exploit rich and intrinsic data representation.
- Gross label information is much easier to obtain which can be viewed as "noisy" input which allows us to explore structural information which might be hard to specify at the first place.
- Weakly-supervised learning allows us to greatly automate and scale up the learning process, which is strongly tied with the development of human cognition.
- There are still a lot of open questions, so as great opportunities ahead.

Abstraction, Composition, Competition, and Computation

Thanks! Questions?