Toward richer targets in large-scale recognition

Alex Berg

Computer Science Department
Stony Brook University

Past Affiliations

EECS UC Berkeley

YAHOO! RESEARCH

CS@CU COMPUTER SCIENCE AT COLUMBIA UNIVERSITY

Co-authors so far

Long Term Collaborators

Tamara Berg's lab

Stony Brook University

Fei-Fei Li's lab

STANFORD UNIVERSITY

Please see my papers for (many!) references.

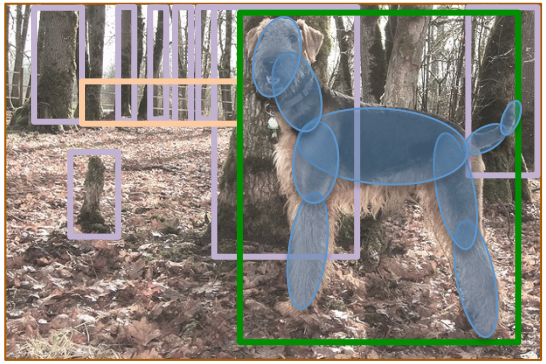
What to recognize?

http://www.flickr.com/photos/sgcallawayimages/3306849049/

Increasing structural complexity

Dog — Single label

Dog, Tree, Fence, Leaf — Multiple labels



Localization

"Happy shaggy Airedale poses in the autumn forest." — Description

Where are we?
(for large-scale recognition)

http://www.flickr.com/photos/sgcallawayimages/3306849049/

Need:

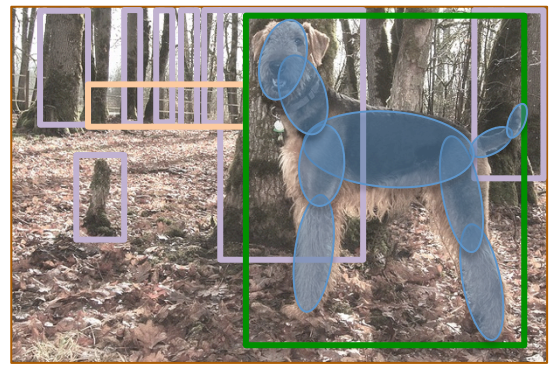
Data, Labels, Models, Algorithms

+ Computation!

Increasing structural complexity

Dog — Single label

Dog, Tree, Fence, Leaf — Multiple labels



Localization

* "Happy shaggy Airedale poses in the autumn forest." — Description

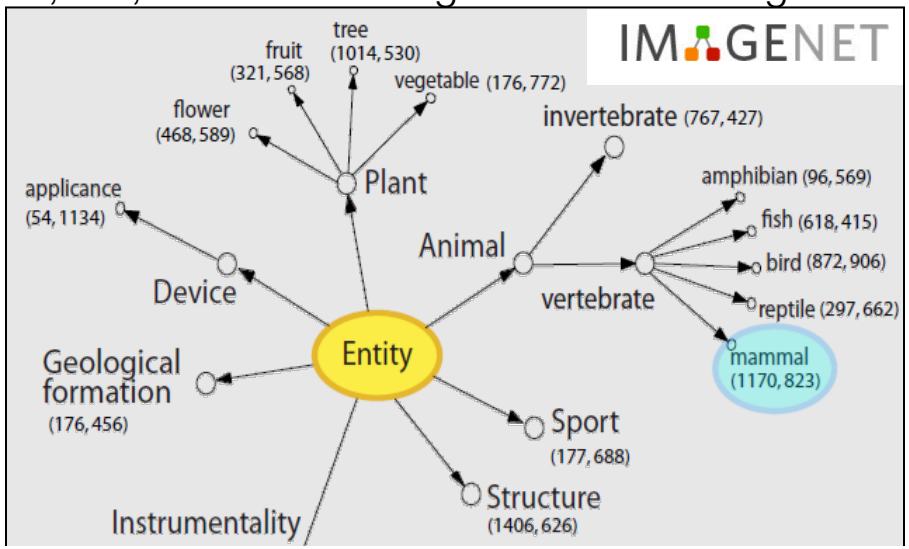
Label space?

20,000 categories (noun synsets) from WordNet

Dog

http://www.flickr.com/photos/sgcallawayimages/3306849049/

14,197,122 labeled images in 21841 categories



Dog, Tree, Fence, Leaf

Multiple labels

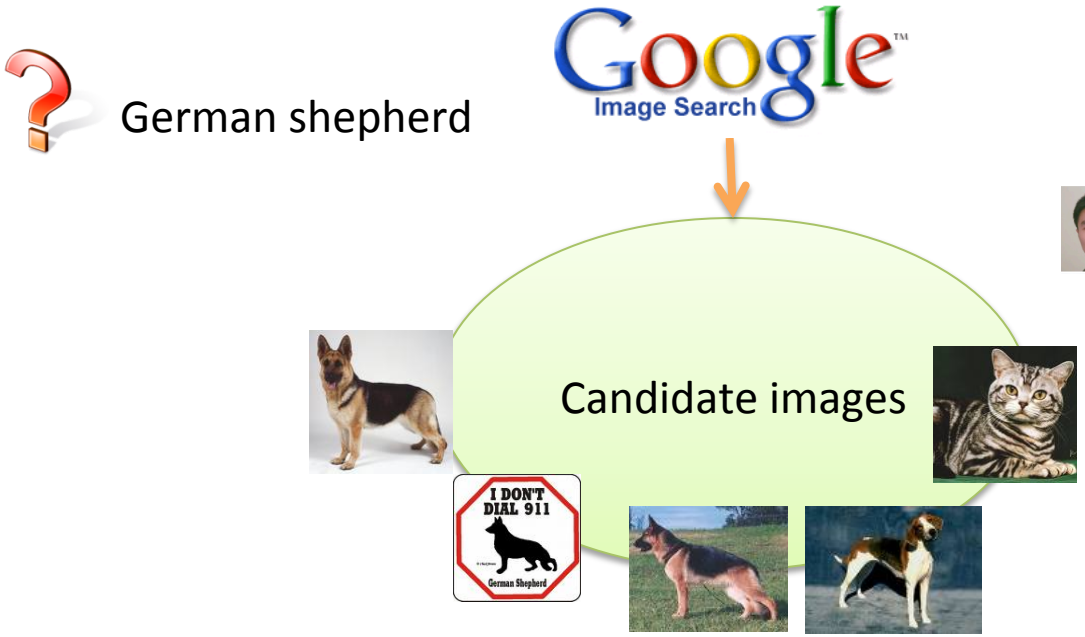
Localization

"Happy shaggy Airedale poses in the autumn forest."

Description

Step 1: Collect Images from the Internet

- Send queries to image search engines
  - Use query expansion to increase # retrievable images



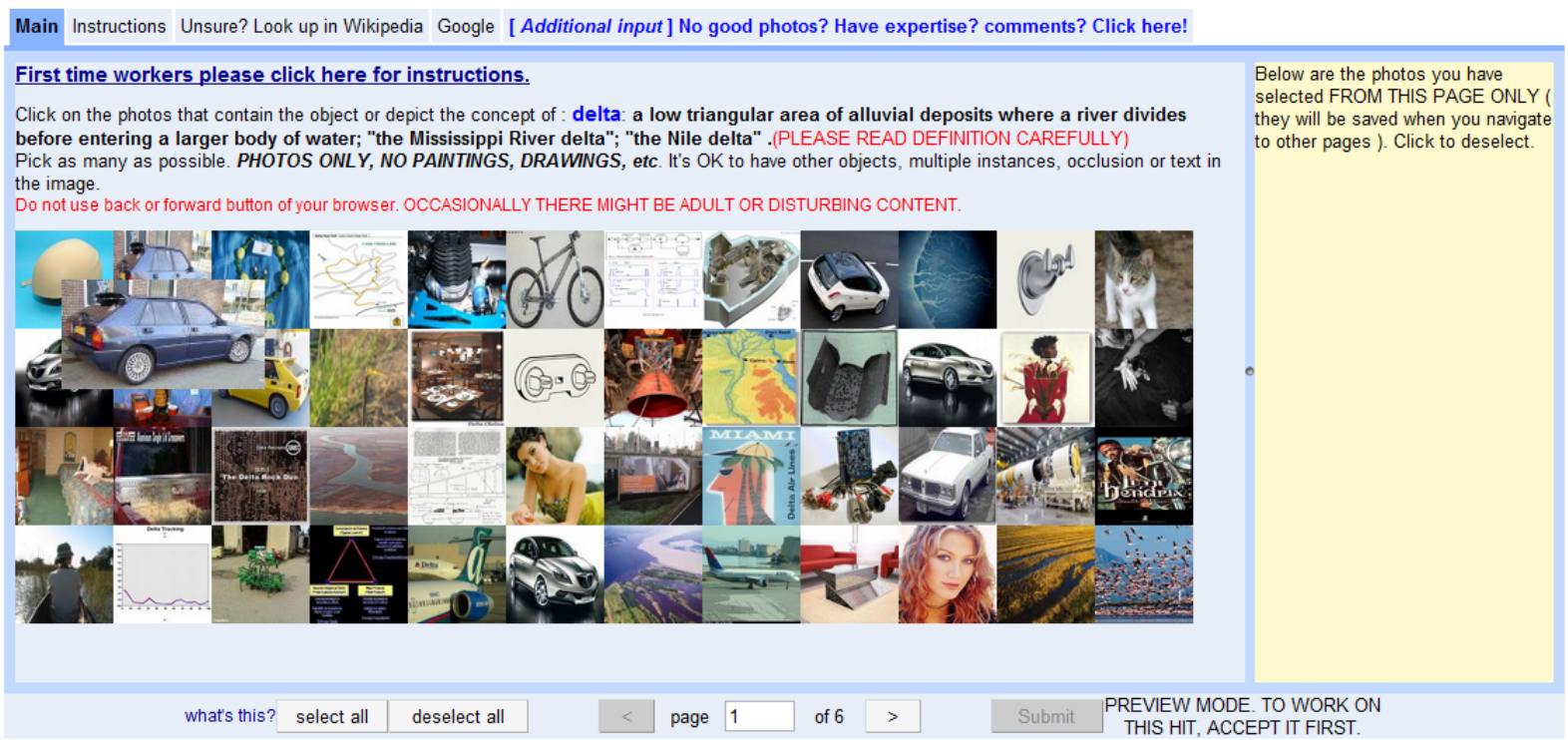
80,000,000 Tiny Images 2007
Torralba, Fergus, Freeman

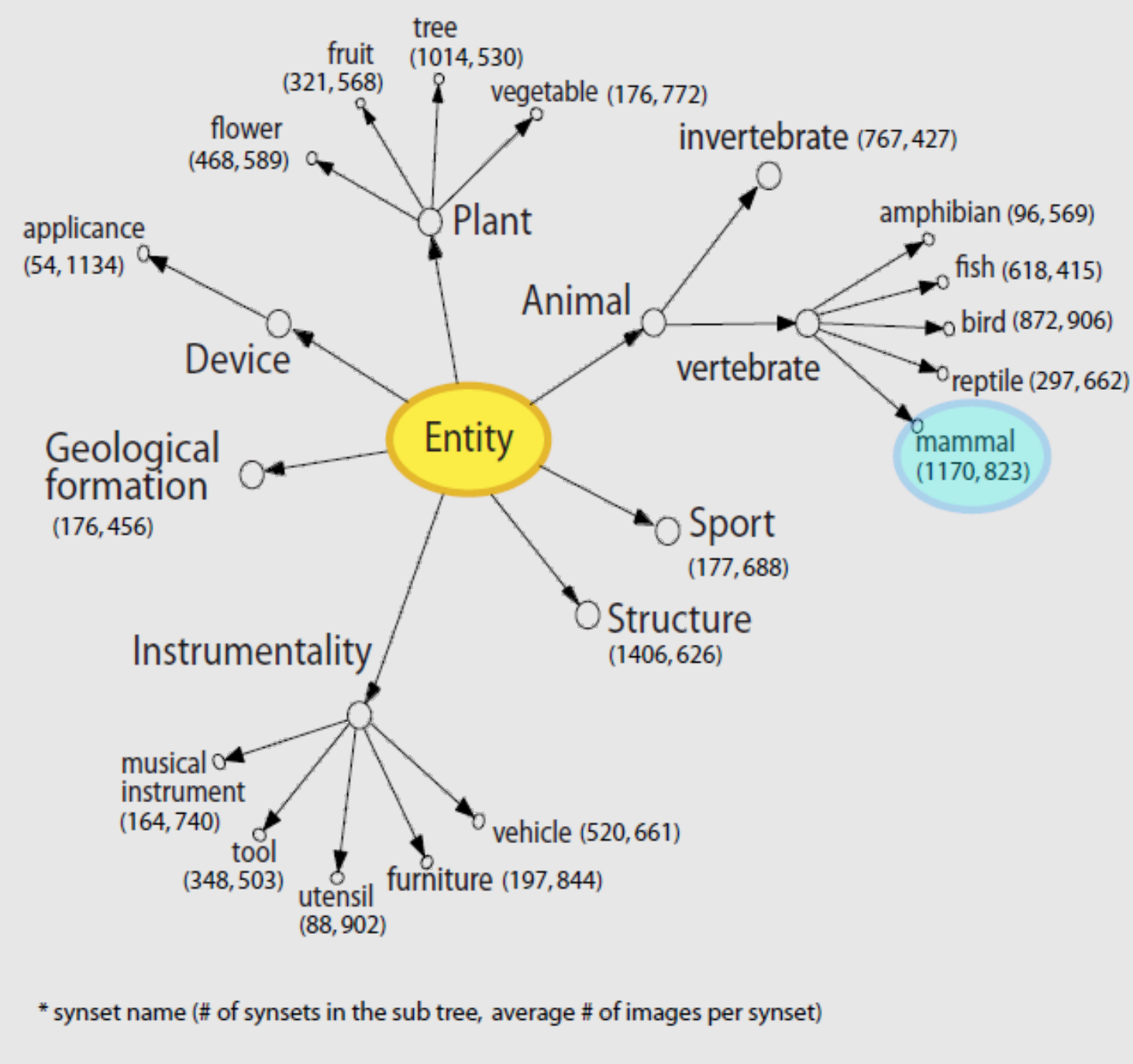
ImageNet 2009
Deng, Dong, Socher, Li, Li, Fei-Fei

Step 2: Verify the images

- Ask (pay!) people to verify presence of objects in images

ImageNet 2009
Deng, Dong, Socher, Li, Li, Fei-Fei




tree
(1014,530)
fruit
(321,568)
vegetable (176,772)
flower
(468,589)
Plant
invertebrate (767,427)
amphibian (96,569)
applicance
(54,1134)
fish (618,415)
Animal
bird (872,906)
Device
vertebrate
reptile (297,662)
Entity
mammal
(1170,823)
Geological
formation
(176,456)
Sport
(177,688)
Structure
(1406,626)
Instrumentality
musical
instrument
(164,740)
vehicle (520,661)
tool
(348,503)
utensil
(88,902)
furniture (197,844)
* synset name (# of synsets in the sub tree, average # of images per synset)

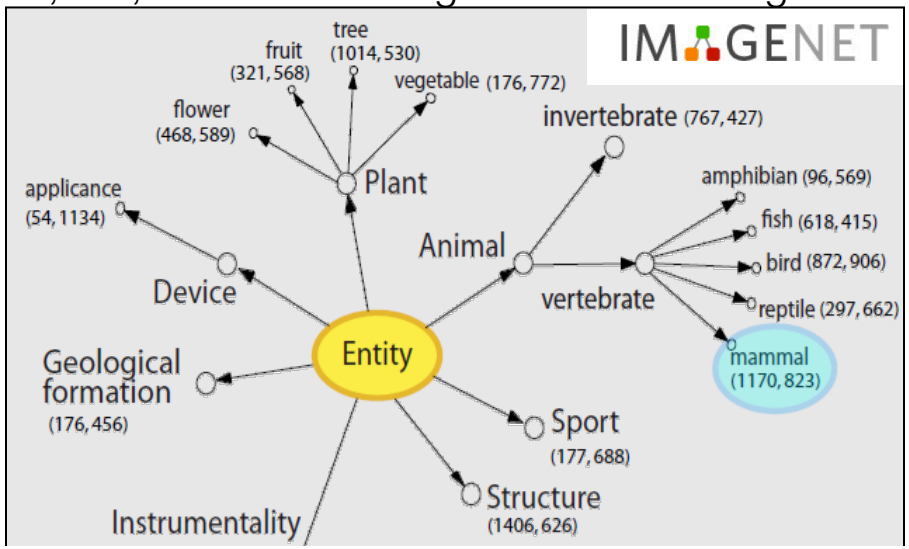
Label space?

20,000 categories (noun synsets) from WordNet

Dog

http://www.flickr.com/photos/sgcallawayimages/3306849049/

14,197,122 labeled images in 21841 categories



Dog, Tree, Fence, Leaf

Multiple labels

Localization

"Happy shaggy airedale poses in the autumn forest."

Description

Label space?

1 million+ query terms in Google image search

Bengio et al

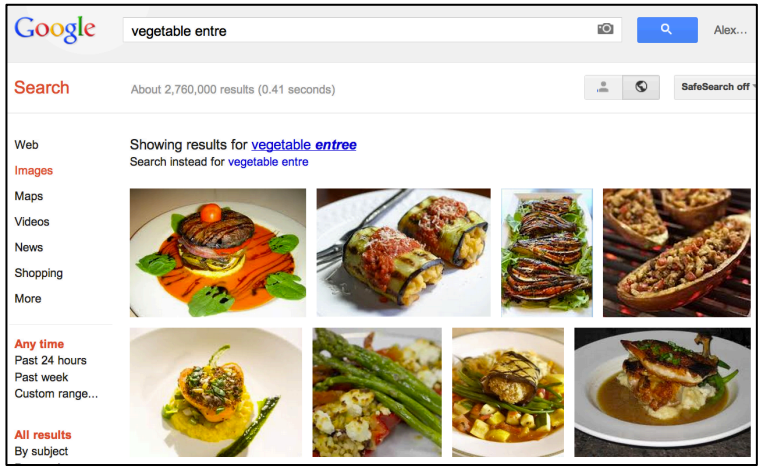
Used for ranking & retrieval experiments

Dog

http://www.flickr.com/photos/sgcallawayimages/3306849049/

Dog, Tree, Fence, Leaf

Multiple labels



Localization

"Happy shaggy airedale poses in the autumn forest."

Description

Label space?

20,000 categories
(noun synsets)
from WordNet

http://www.flickr.com/photos/sgcallawayimages/3306849049/

Dog

$\cup$

Dog, Tree, Fence, Leaf

Multiple labels

Localization

"Happy shaggy airedale poses in the autumn forest."

Description

Label space?

http://www.flickr.com/photos/sgcallawayimages/3306849049/

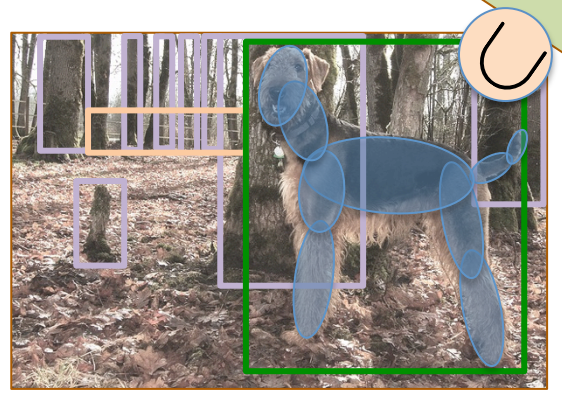
Dog

Single label

Dog, Tree, Fence

labels

Space of configurations
Big!



Localization

"Happy shaggy airedale poses in the autumn forest."

Description

But most work uses a small number of target categories, e.g. 20 from Pascal VOC, or 100s for image parsing.

ImageNet Large scale visual recognition challenge (LSVRC)

Motivation:

- Applications with millions to billions of images
- Hundreds of thousands of possible labels
- Recognition for indexing and retrieval
- Complement current Pascal VOC (20 categories)

Currently 1000 categories in the challenge, but only one class of object labeled per image.

Categorization Car

People

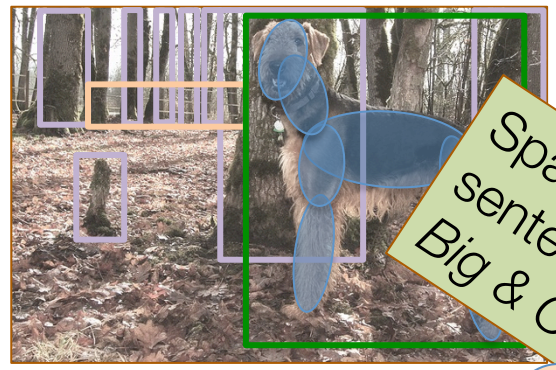
Alex Berg	-- Stony Brook
Jia Deng	-- Stanford
Sanjeev Satheesh	-- Stanford
Hao Su	-- Stanford
Fei-Fei Li	-- Stanford

Label space?

http://www.flickr.com/photos/sgcallawayimages/3306849049/

Dog — Single label

Dog, Tree, Fence, Leaf — Multiple labels



Space of sentences Big & Complex

lization

$\mathbb{U}$

"Happy shaggy airedale poses in the autumn forest." — Description

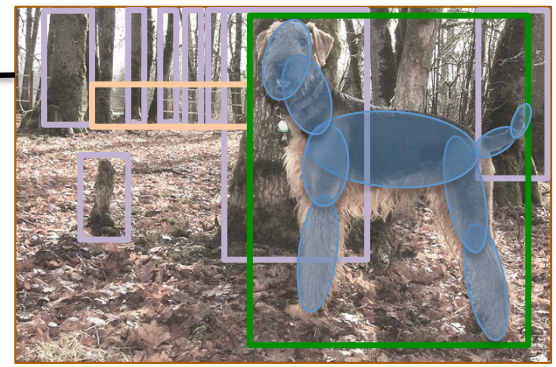
But again, work on using computer vision for this uses no more than 100s of labels.*

Why?

http://www.flickr.com/photos/sgcallawayimages/3306849049/

Dog — Single label

Dog, Tree, Fence, Leaf — Multiple labels



Localization

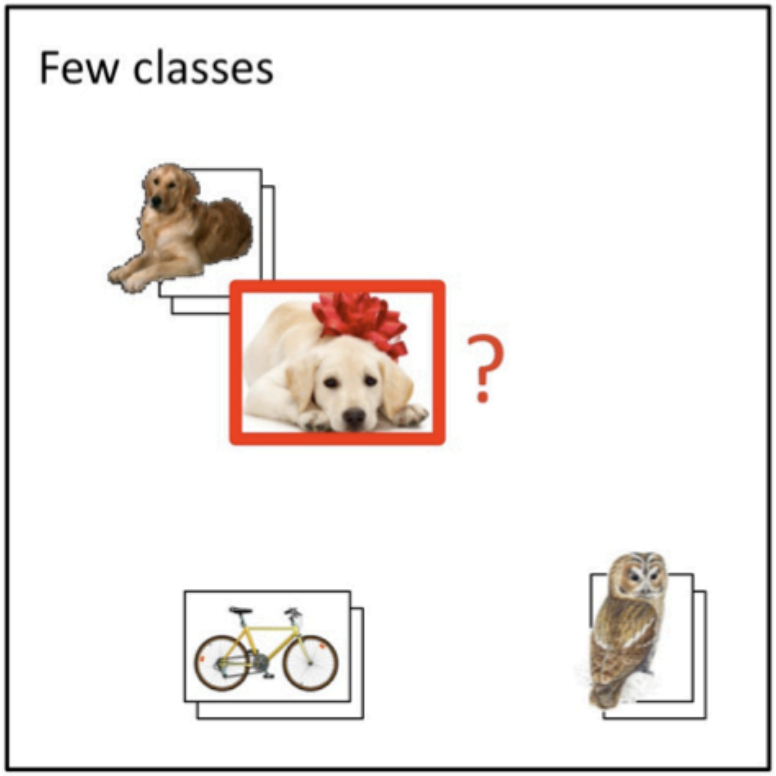
"Happy shaggy airedale poses in the autumn forest." — Description

Estimating content

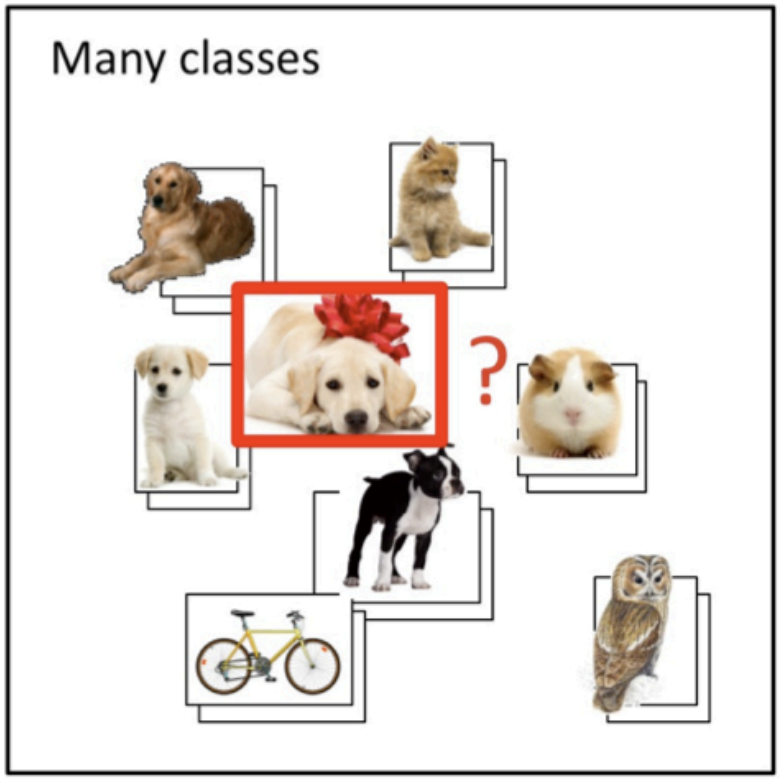
Information to index images for
- Text based retrieval
- Image based retrieval
- Visual mining
- Visual discovery

Why do we want a large-scale label space?

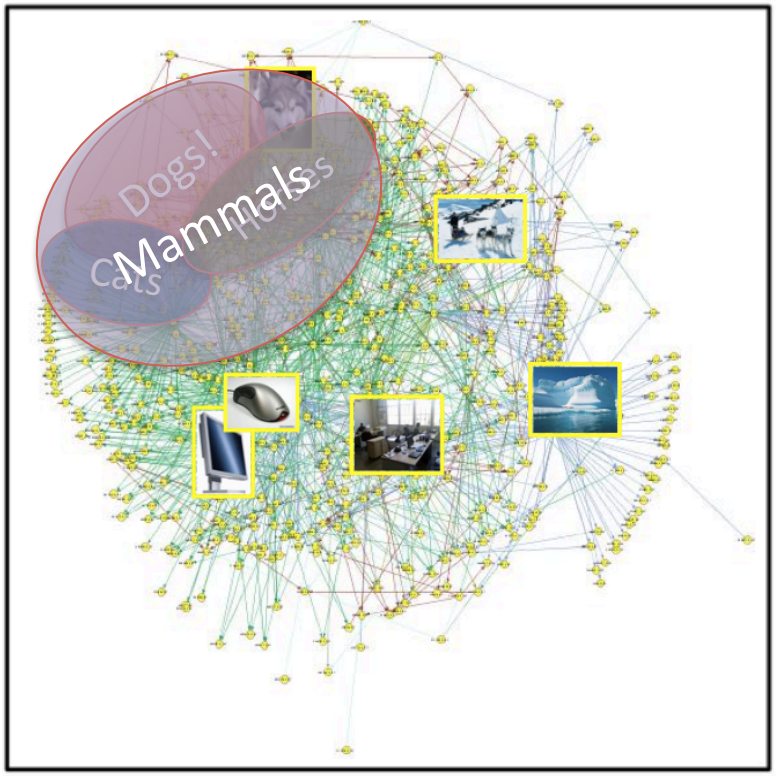
Where is the magic?



Where is the magic?



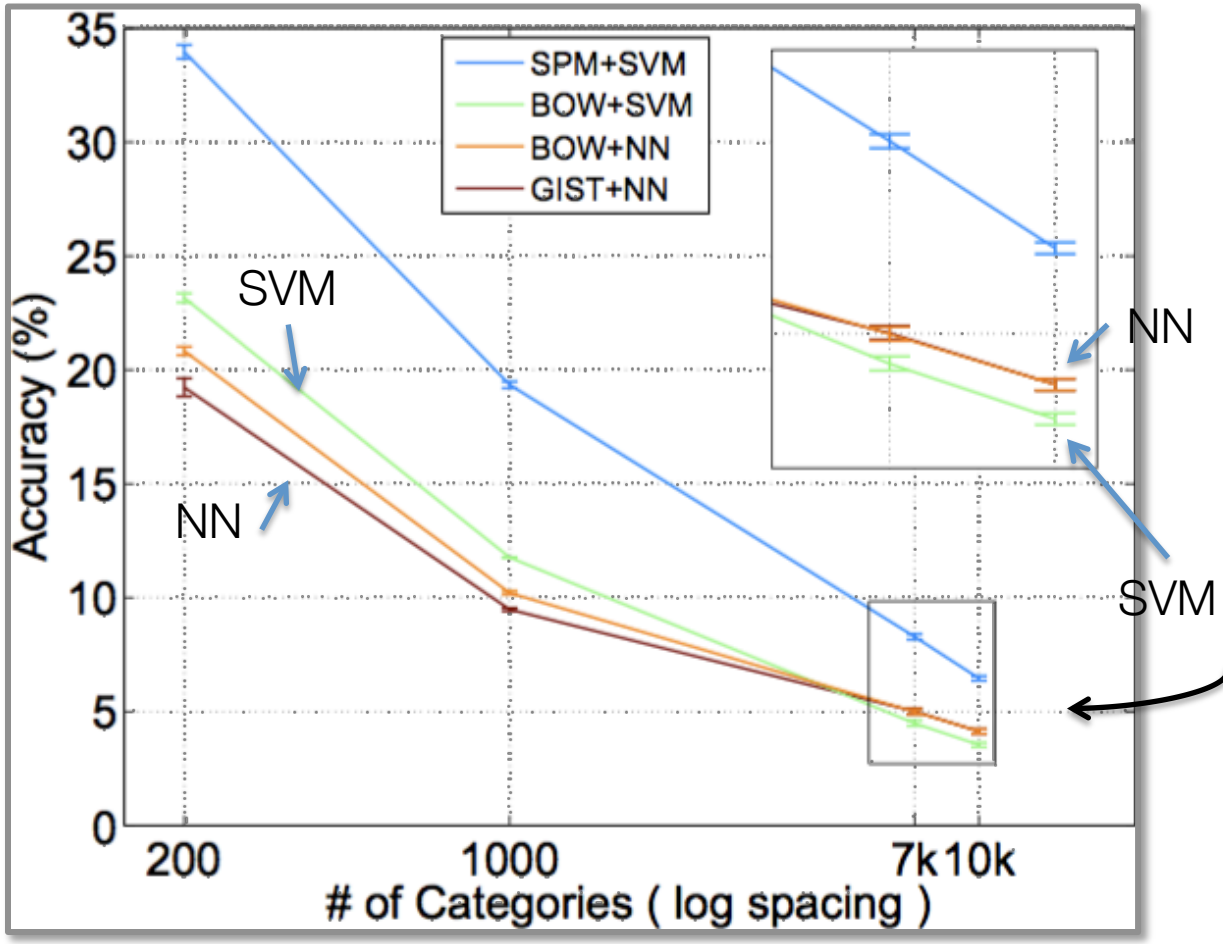
Where is the magic?



~9 Million Images in ~10,000 categories from ImageNet

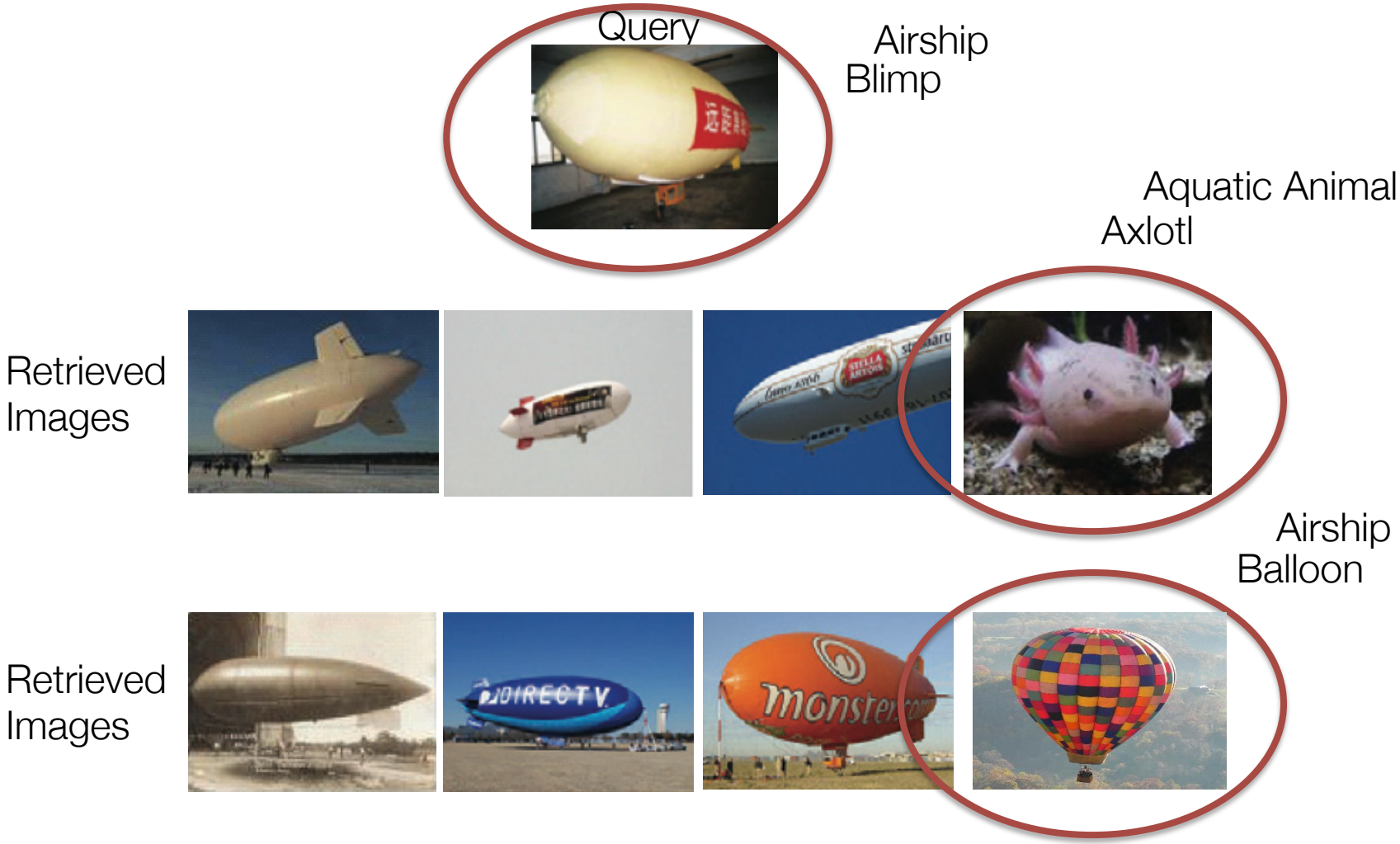
Studying large scale classification

What does classifying more than 10,000 image categories tell us?
ECCV 2010
Deng, Berg, Li, Fei-Fei



No magic…

What do we mean by similarity?



Hierarchical Semantic Indexing, CVPR 2010, Deng, Berg, Fei-Fei

Similarity functions

Scale of the label space matters here!

$$s_i = \text{semantic}\left(\text{features}\left(I_i\right)\right)$$

Ours

$$\text{Sim}\left(I_1, I_2\right) = s_1^T I s_2$$

$$\text{Sim}\left(I_1, I_2\right) = s_1^T H s_2$$

$$\text{Sim}\left(I_1, I_2\right) = \widehat{\text{Sim}}\left(\text{semantic}\left(\text{features}\left(I_1\right)\right), \text{semantic}\left(\text{features}\left(I_2\right)\right)\right)$$

Learn

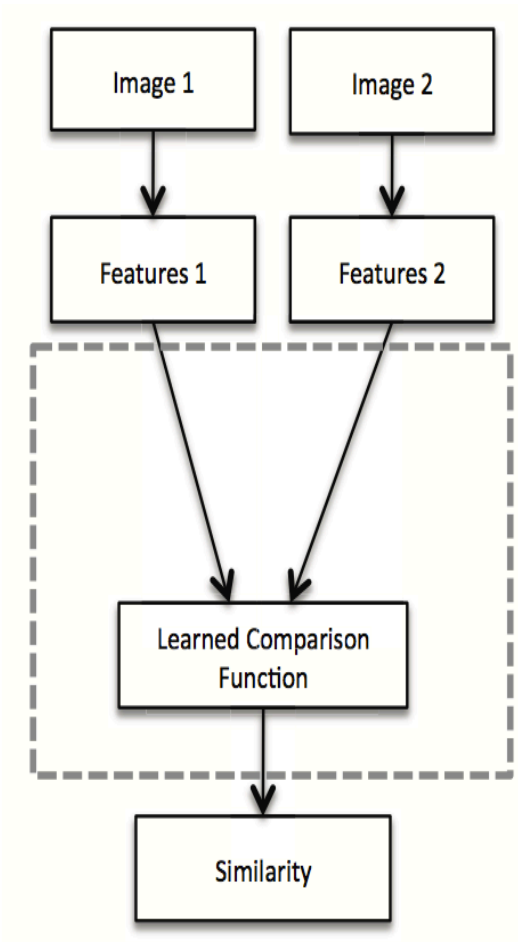
Fixed Mapping

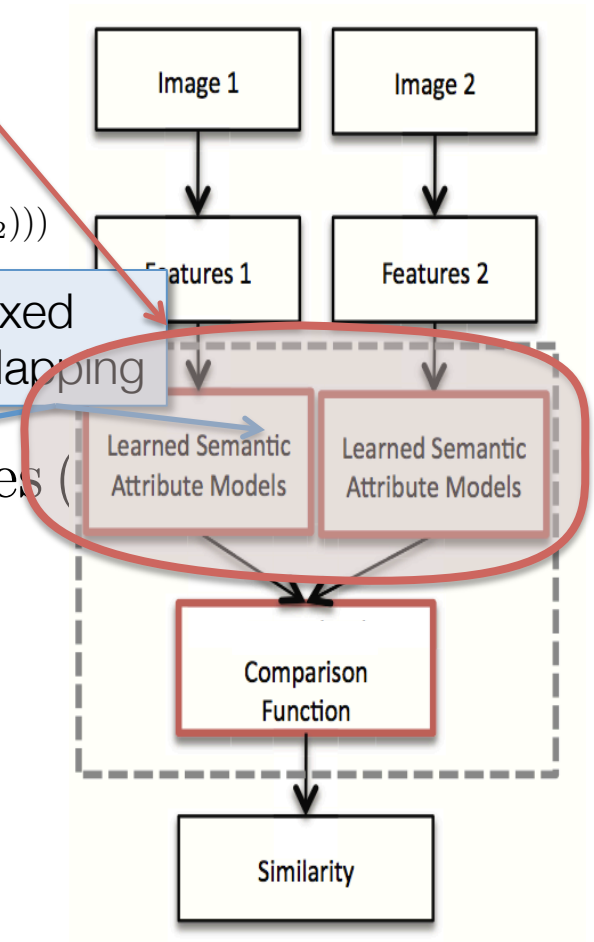
Fixed Mapping

$$\text{Sim}\left(I_1, I_2\right) = \text{Sim}\left(\text{features}\left(\right.\right.$$

Fixed or Learn

$$\text{Sim}\left(I_1, I_2\right) = f_1^T M f_2$$

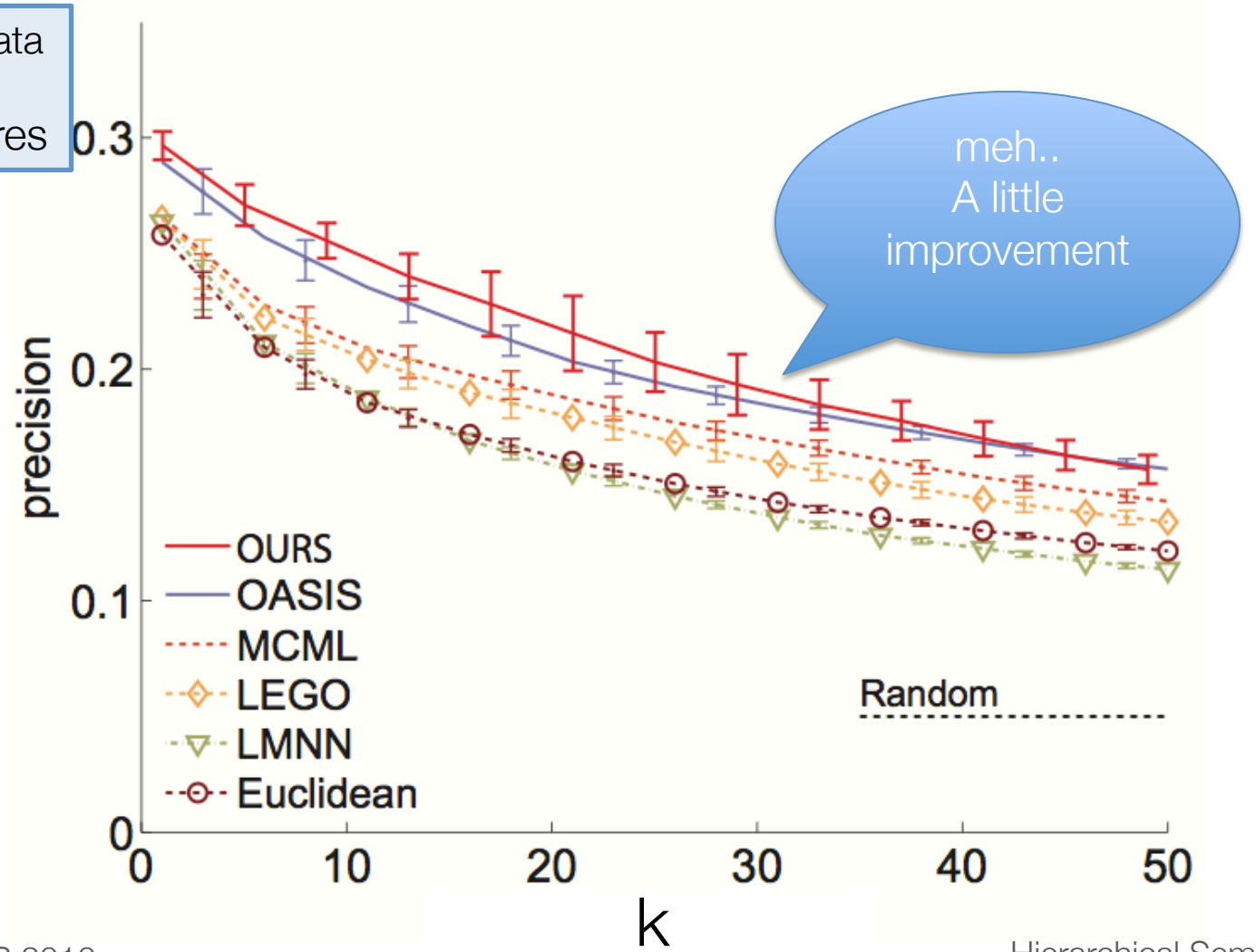




$$\text{Sim}\left(I_1, I_2\right) = \widehat{\text{Sim}}\left(\text{features}\left(I_1\right), \text{features}\left(I_2\right)\right)$$

Similar image retrieval small # of classes

Caltech 256 Data
20 classes
1,000d features

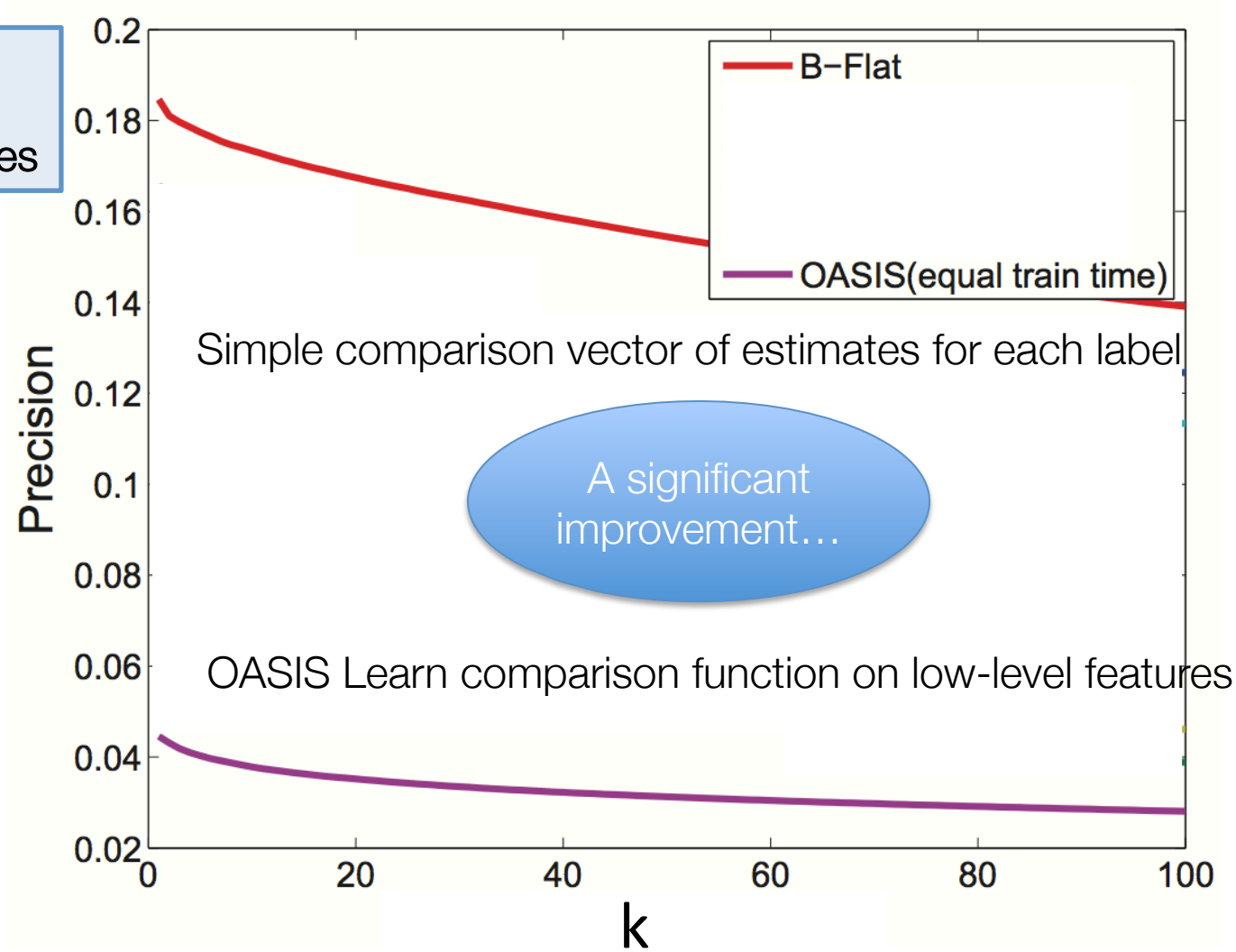


OASIS
NIPS 2009, JMLR 2010
Chechik, Shalit, Bengio, Sonnenburg, Franc, Yom-tov, Sebag

Hierarchical Semantic Indexing
CVPR 2010
Deng, Berg, Fei-Fei

Similar image retrieval large # of classes

ImageNet Data
1000 classes
21,000d features



Random is 0.001

Similar approach for face similarity

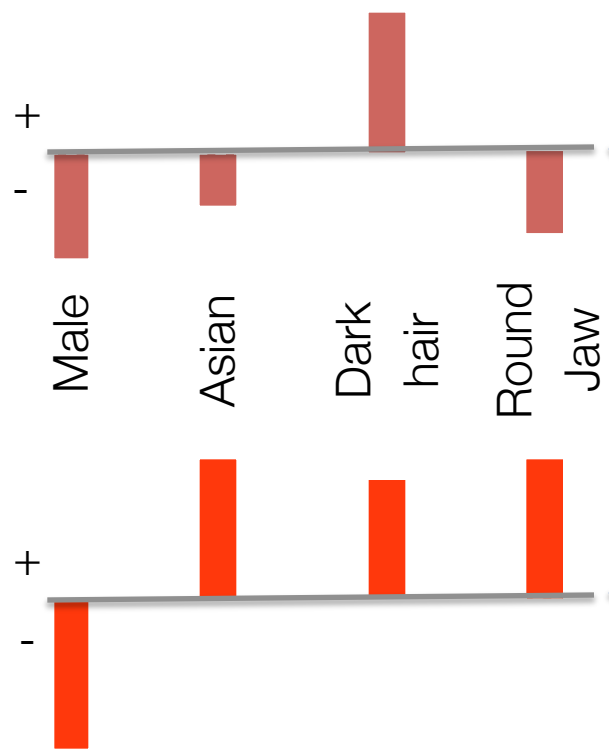
Images

Low-level features

RGB
HOG
LBP
SIFT
…

RGB
HOG
LBP
SIFT
…

Attributes



Verification

Different

Attribute and Simile Classifiers for Face Verification
ICCV 2009, PAMI 2011
Kumar, Berg, Belhumeur, Nayar

Similar approach for face similarity

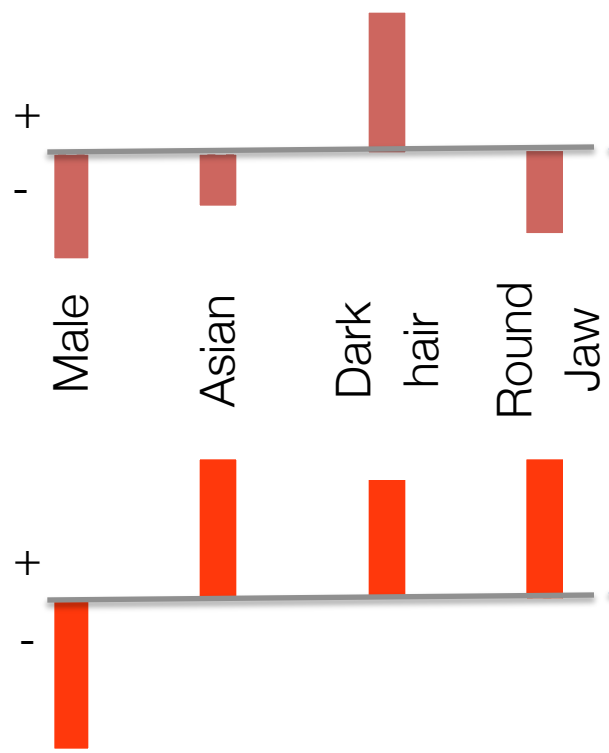
Images

Low-level features

RGB
HOG
LBP
SIFT
…

RGB
HOG
LBP
SIFT
…

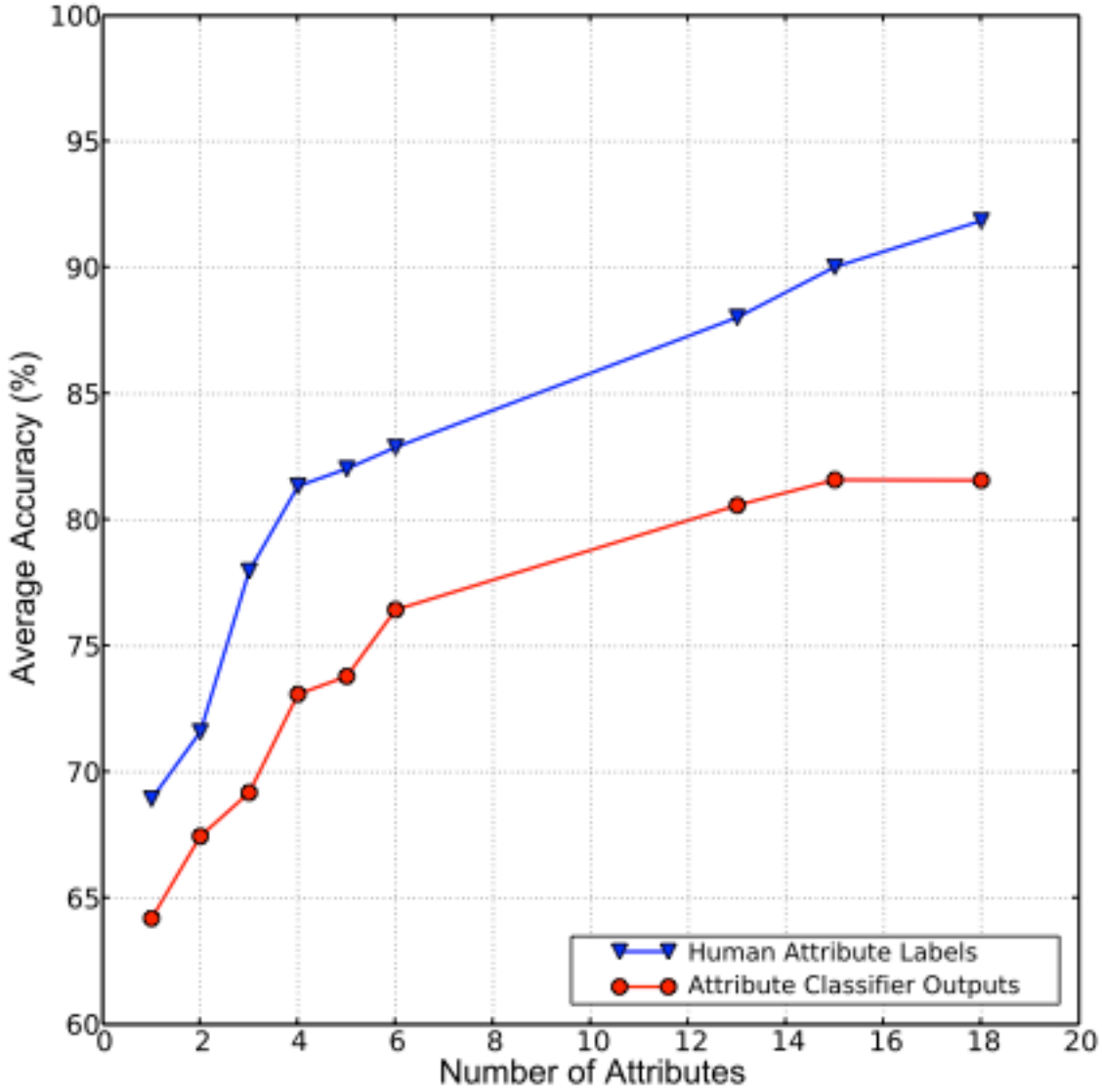
Attributes



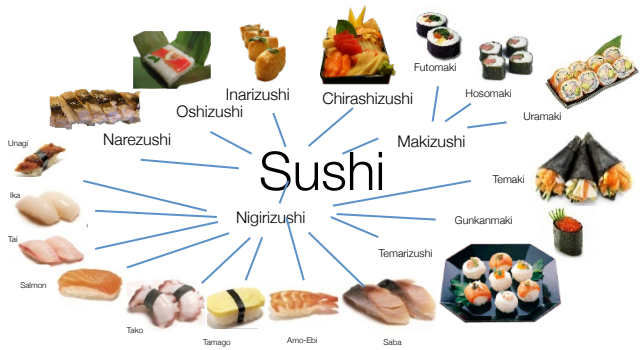
Verification

Different

Attribute and Simile Classifiers for Face Verification
ICCV 2009, PAMI 2011
Kumar, Berg, Belhumeur, Nayar


Average Accuracy (%)
Number of Attributes
Human Attribute Labels
Attribute Classifier Outputs
100
95
90
85
80
75
70
65
60
0
2
4
6
8
10
12
14
16
18
20

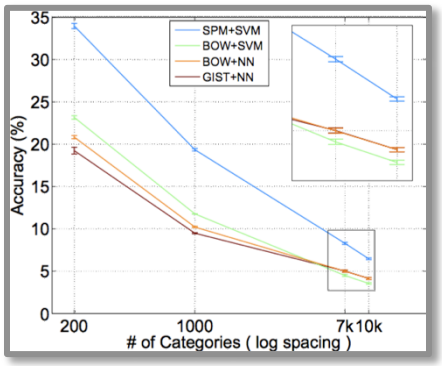
Large scale allows us to use structure over labels!

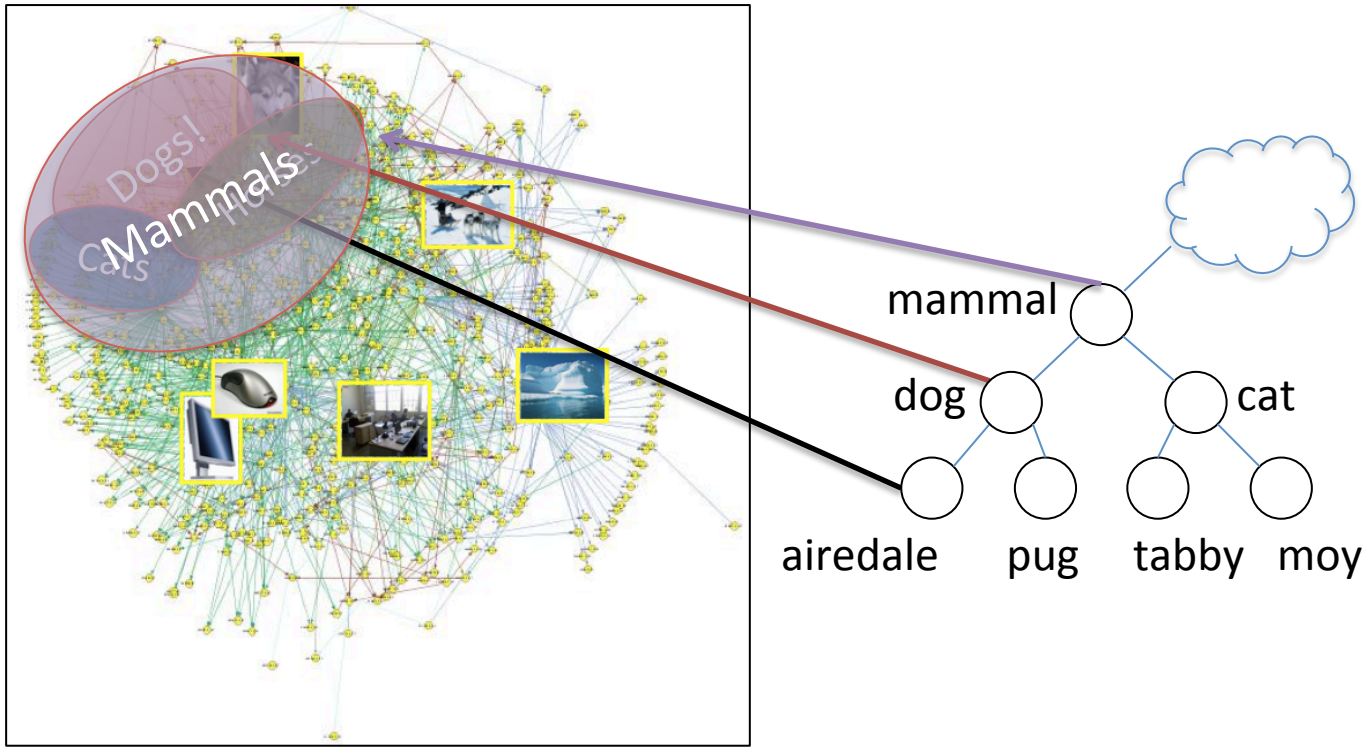


Multi-level categorization

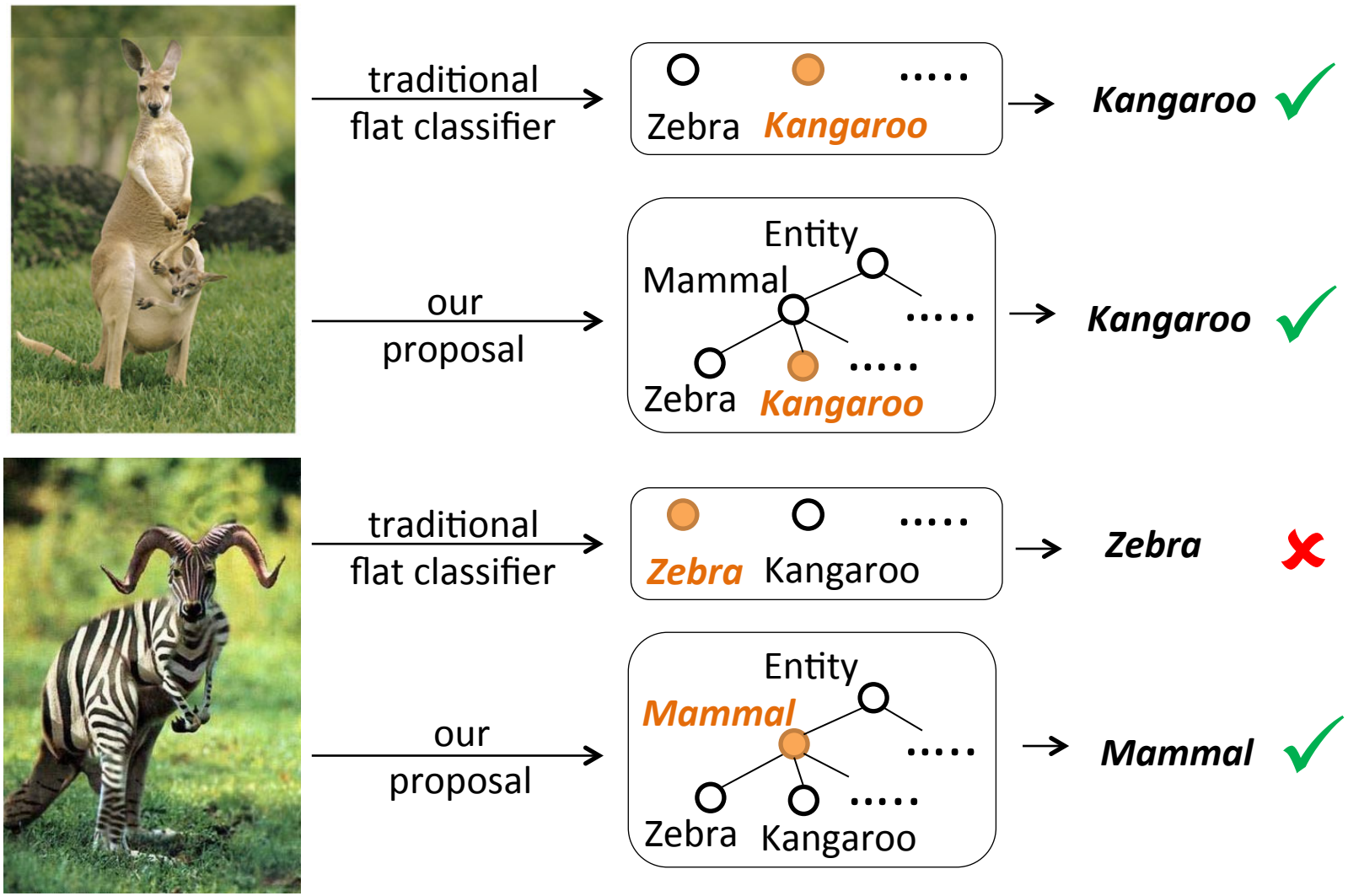
Recognizing new things

Where is the magic?





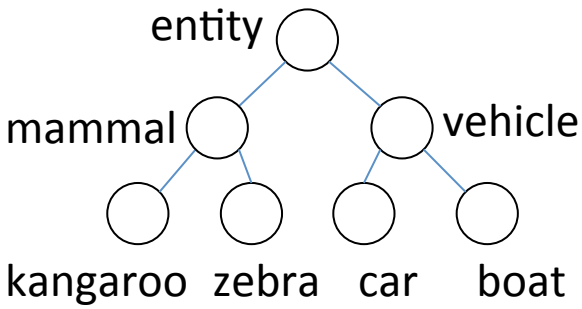
How can we do this well?



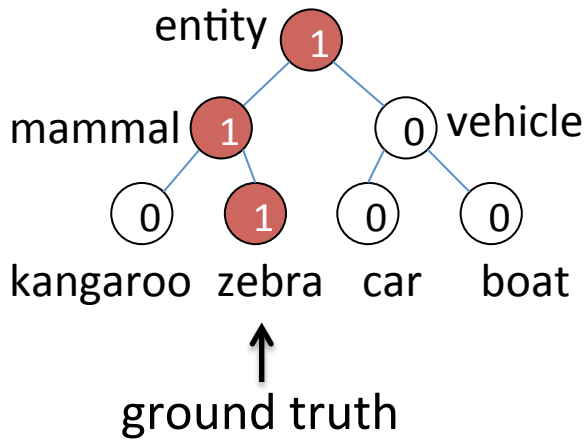
"Hedging your bets" Deng, Krause, Berg, Fei-Fei, CVPR2012

Formulation

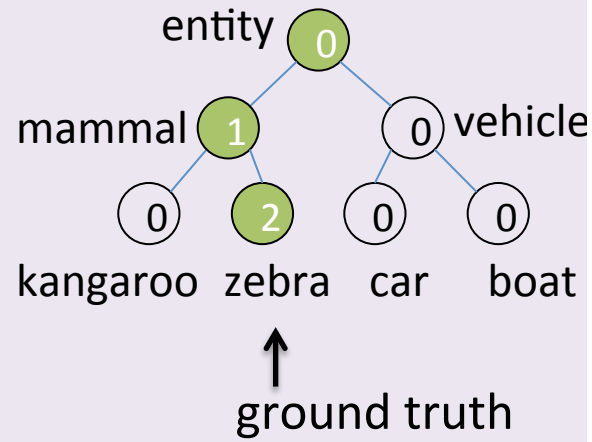
Semantic hierarchy



Accuracy



Reward

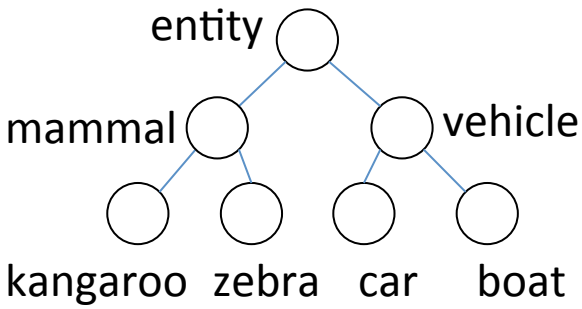


Reward: amount of correct *information gain (i.e. decrease of uncertainty)*

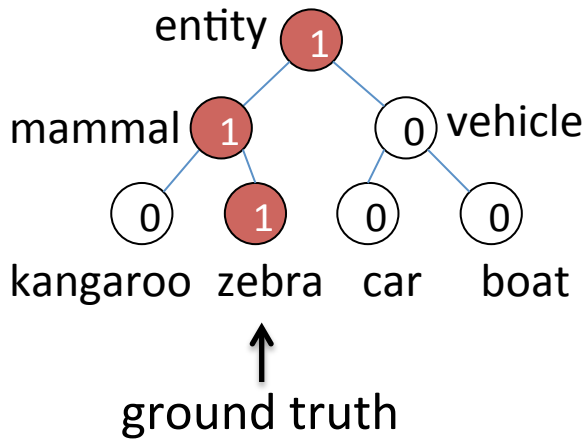
Deng, Krause, Berg, Fei-Fei, CVPR2012

Formulation

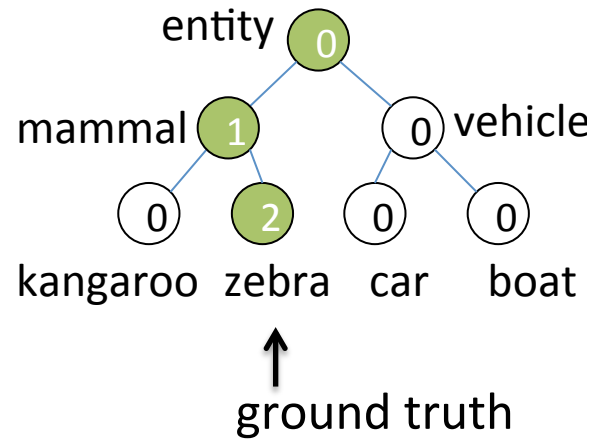
Semantic hierarchy



Accuracy



Reward



Training images

Reward of classifier

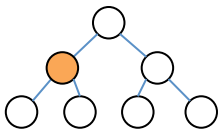
Rewards on nodes

$$\text{Maximize over } f, \quad R(f,r)$$
$$\text{Subject to} \quad \Phi(f) \geq 1-\epsilon$$

Accuracy of classifier

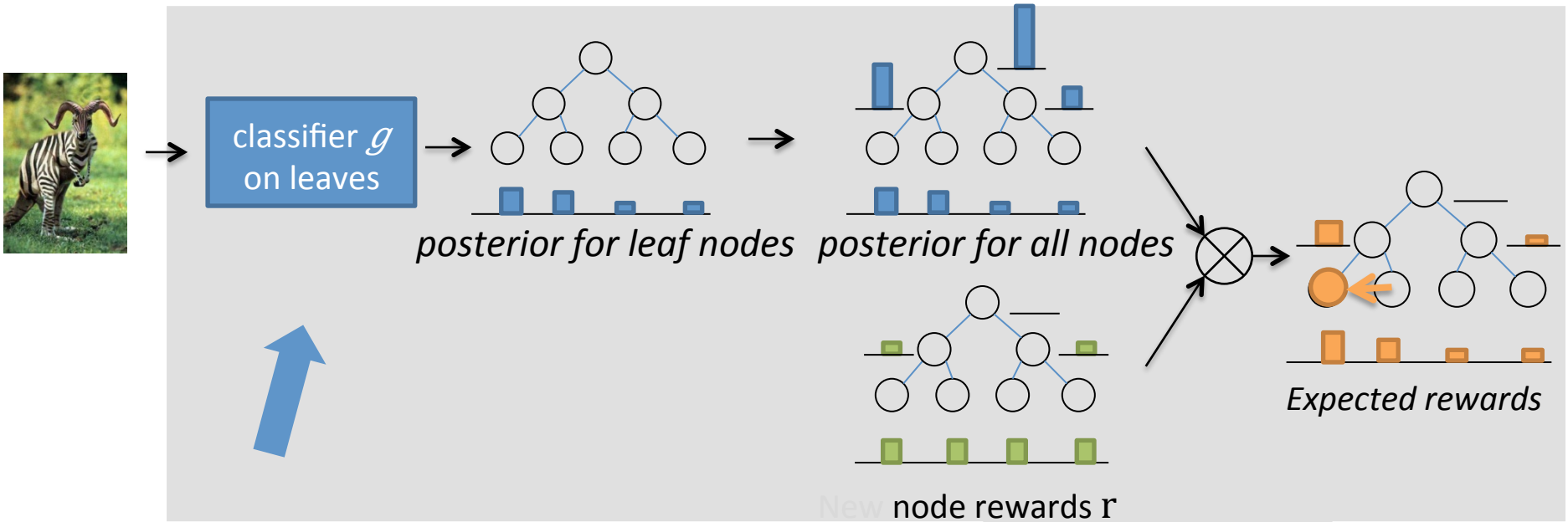
Accuracy guarantee

classifier f



Deng, Krause, Berg, Fei-Fei, CVPR2012

For now, assume no accuracy guarantee.



Training images

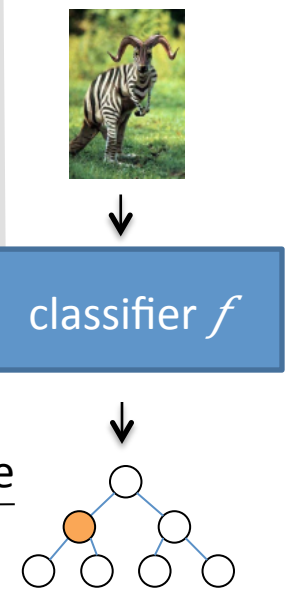
Reward of classifier

Rewards on nodes

$$\text{Maximize over } f, \quad R(f,r)$$
$$\text{Subject to} \quad \Phi(f) \geq 1-\varepsilon$$

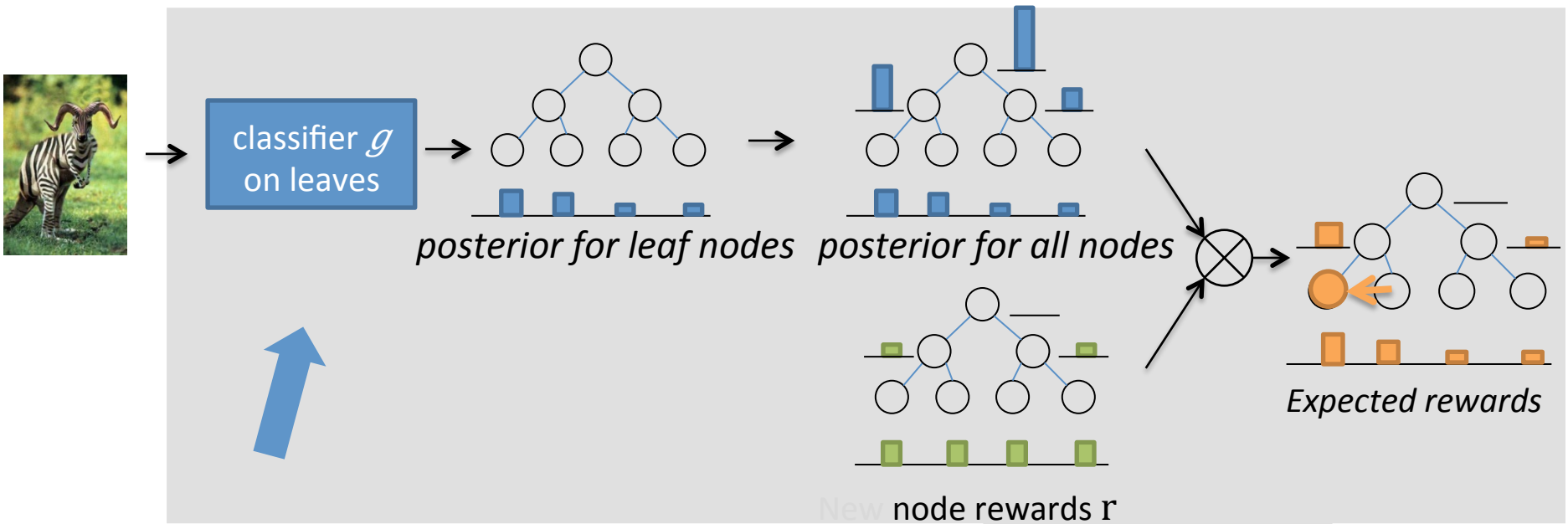
Accuracy of classifier

Accuracy guarantee



Deng, Krause, Berg, Fei-Fei, CVPR2012

What about the accuracy guarantee $1-\epsilon$?



Training images

Reward of classifier

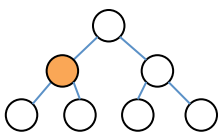
Rewards on nodes

$$\text{Maximize over } f, \quad R(f,r)$$
$$\textbf{Subject to} \quad \Phi(f) \geq 1-\epsilon$$

Accuracy of classifier

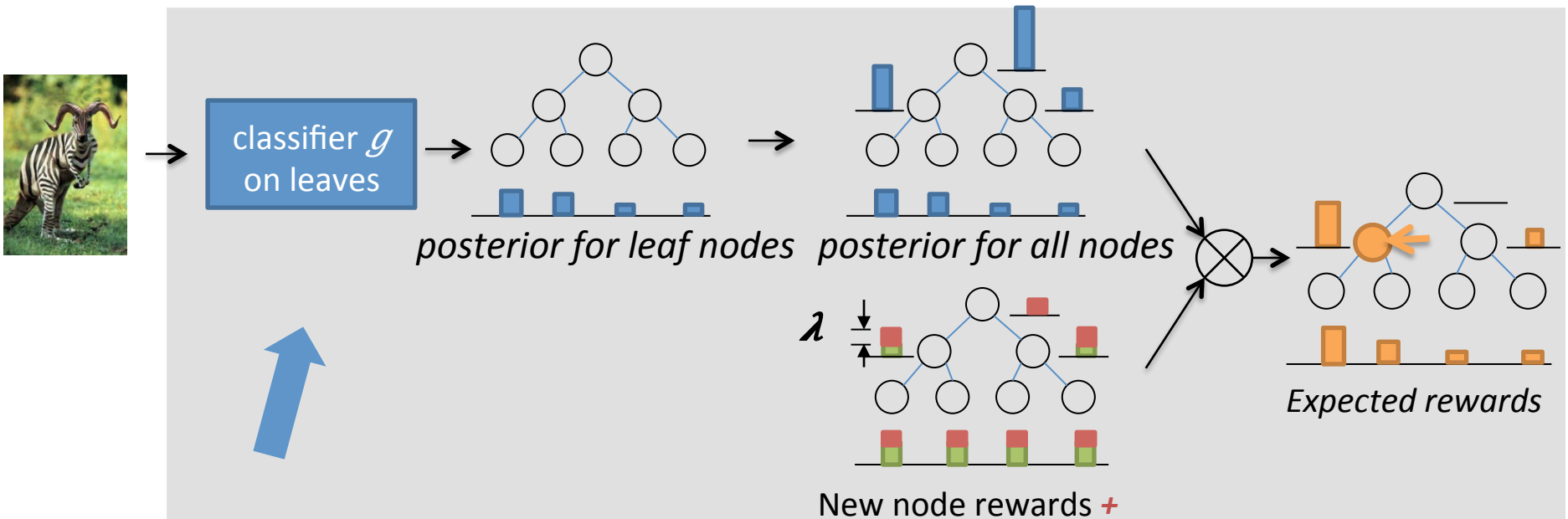
Accuracy guarantee

classifier f



Deng, Krause, Berg, Fei-Fei, CVPR2012

What about the accuracy guarantee $1-\epsilon$?



Training images

Reward of classifier

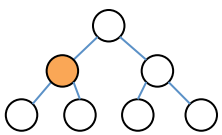
Rewards on nodes

$$\text{Maximize over } f, \quad R(f,r)$$
$$\text{Subject to} \quad \Phi(f) \geq 1-\epsilon$$

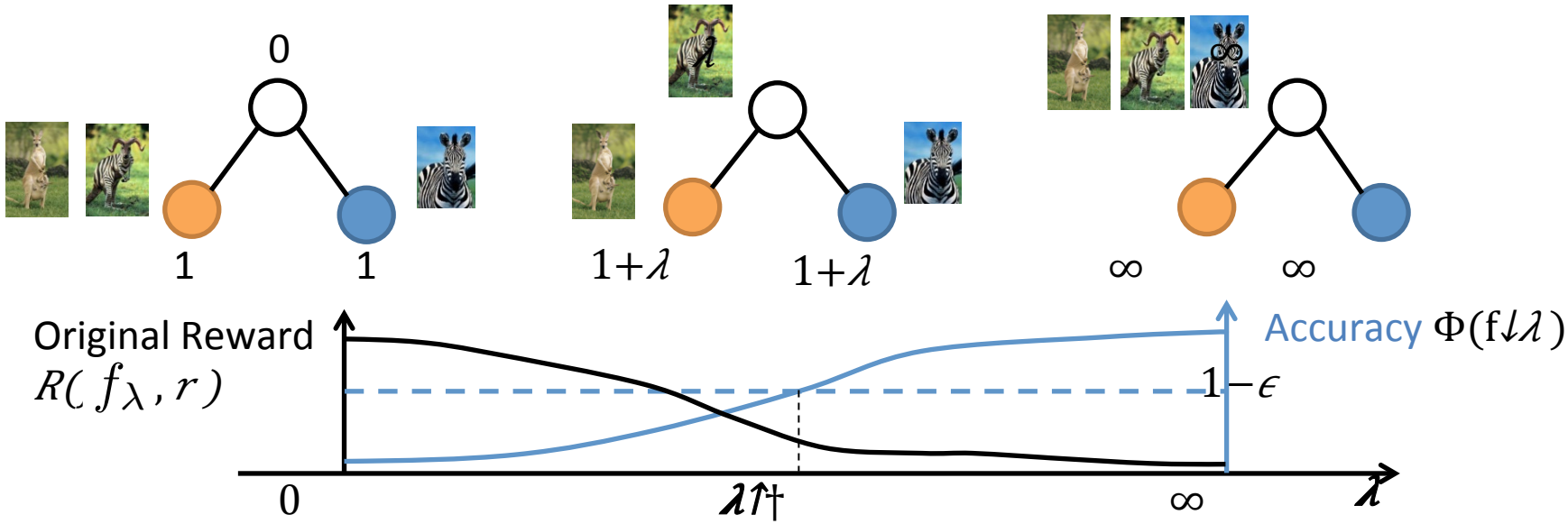
Accuracy of classifier

Accuracy guarantee

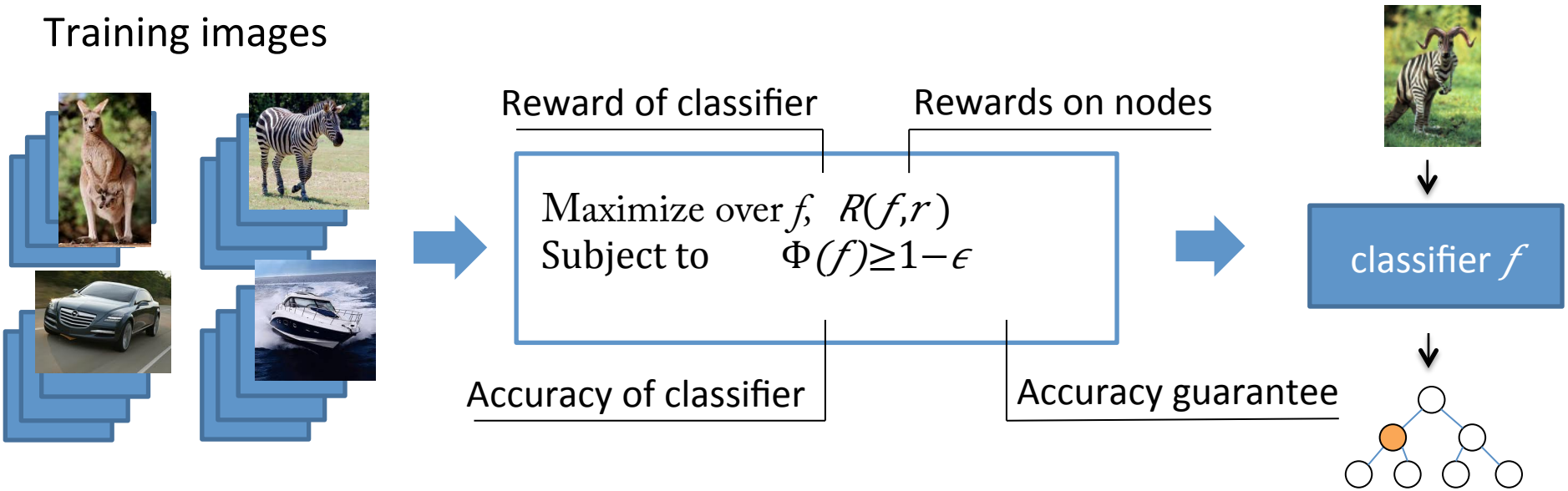
classifier f



Deng, Krause, Berg, Fei-Fei, CVPR2012



The optimal $\lambda^\dagger$ is where the accuracy is exactly $1-\epsilon$: binary search



Deng, Krause, Berg, Fei-Fei, CVPR2012

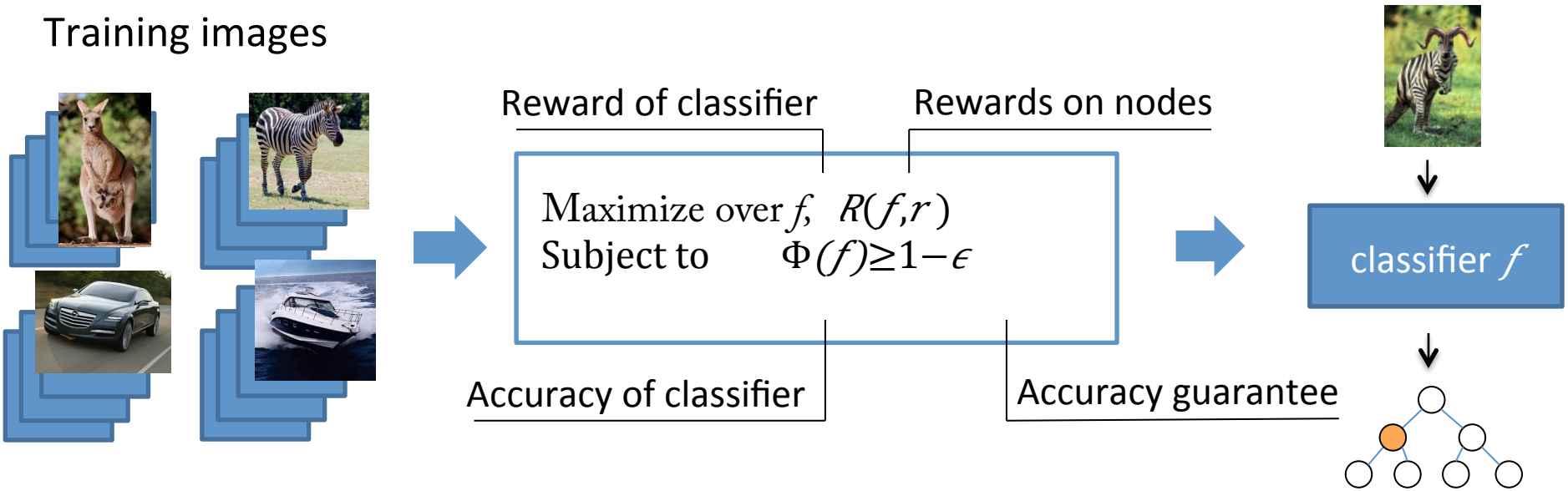
The DARTS algorithm

Dual **A**ccuracy **R**eward **T**rade-off **S**earch

- Train a flat classifier that gives probability estimates on the leaf nodes.
- $f_\lambda \leftarrow$ a classifier that maximizes the expected *new node rewards* $(r + \lambda)$
- Binary search to find the optimal f_λ such that f_λ is $1-\epsilon$ accurate

- λ is the dual variable in the Lagrange function
- Theorem: for any $1-\epsilon$, DARTS converges to an optimal solution except for artificial cases (no worries in practice).

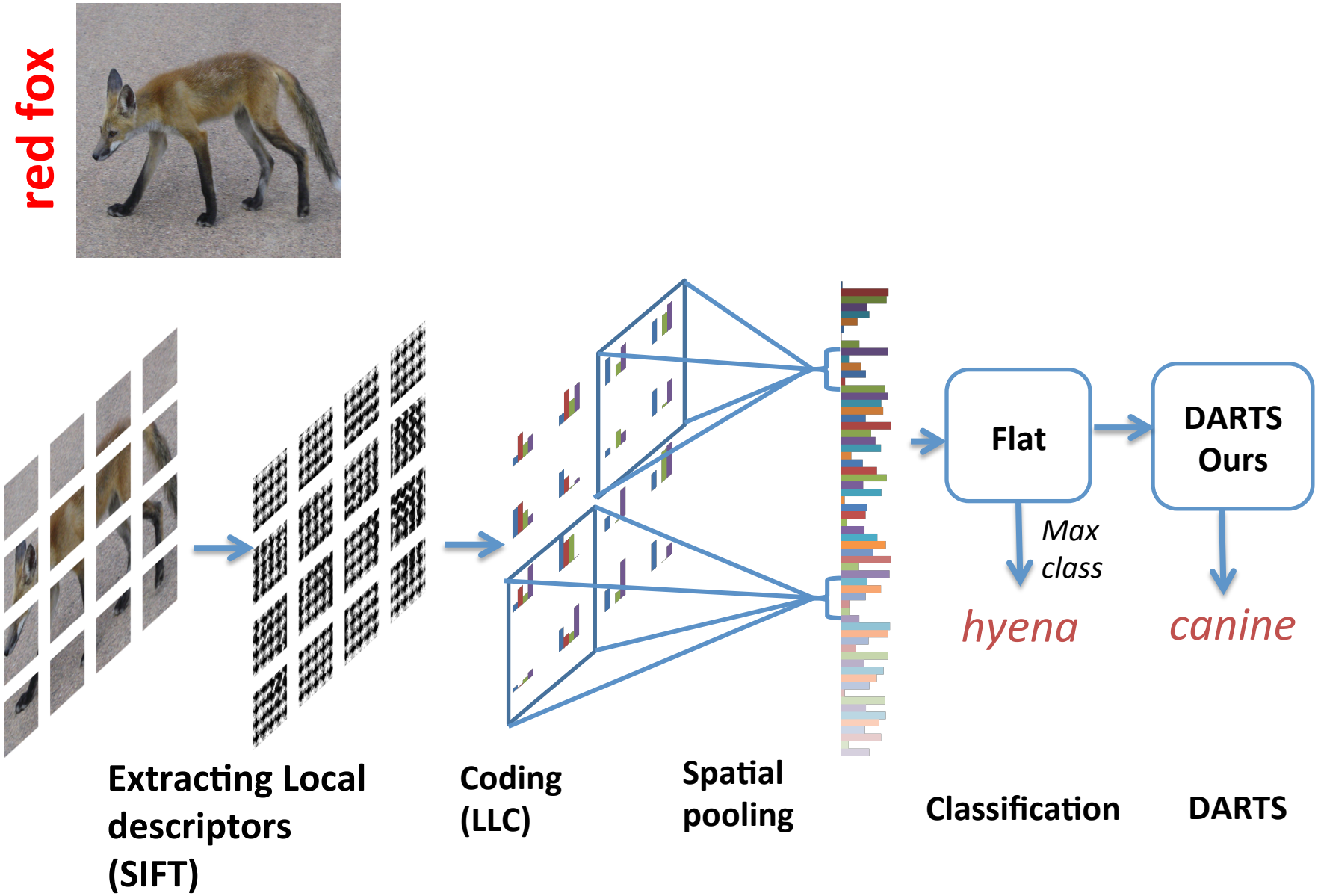
Training images



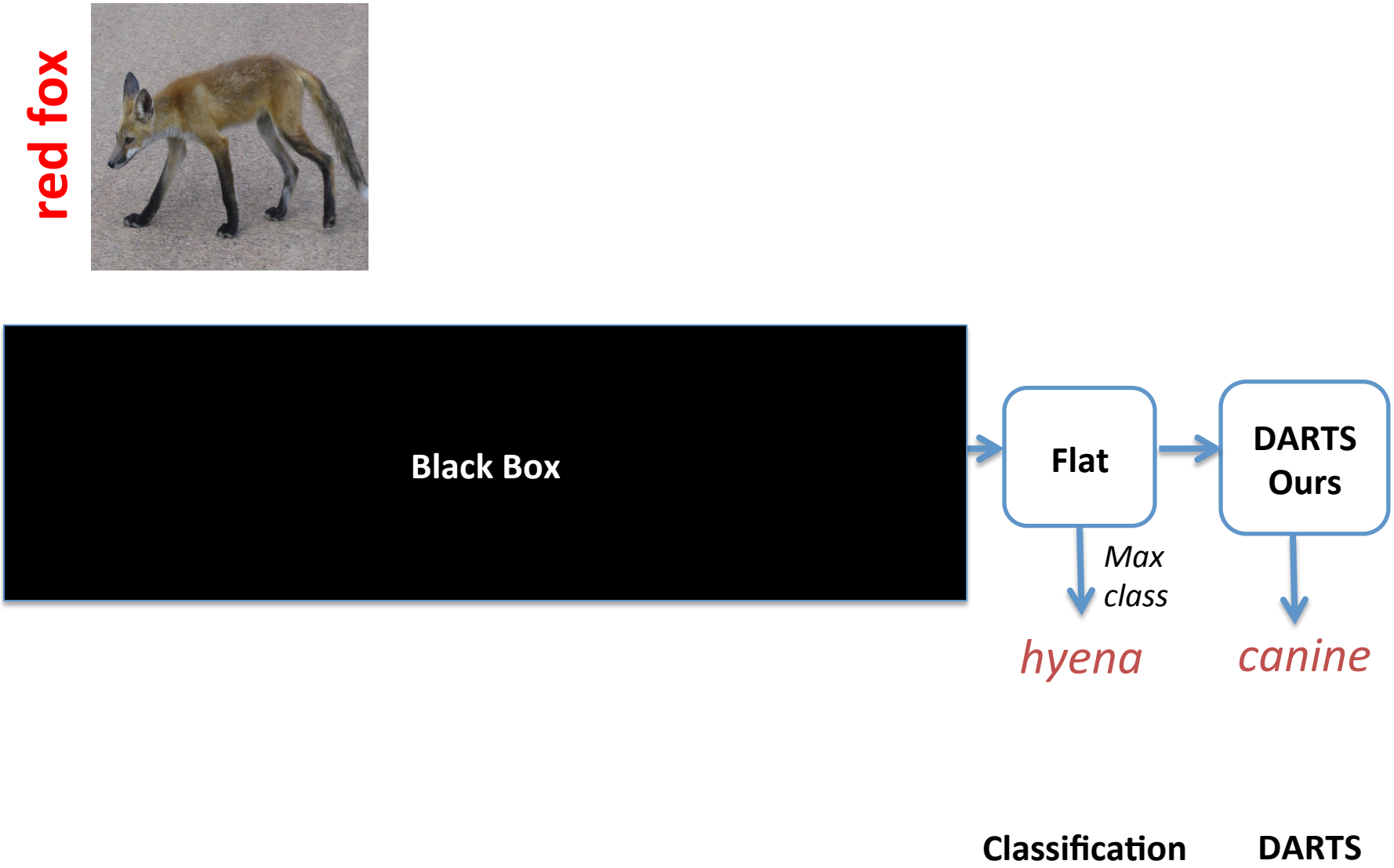
Deng, Krause, Berg, Fei-Fei, CVPR2012

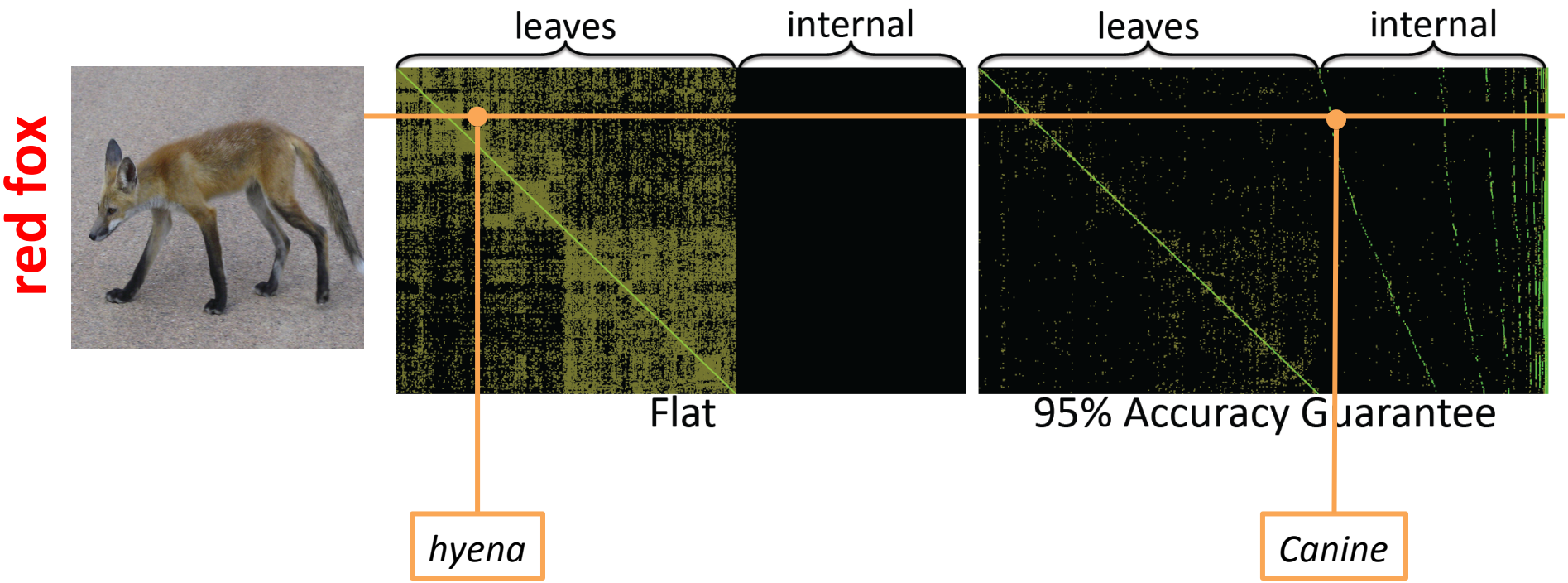
Results

Recognition Pipeline



Recognition Pipeline





Deng, Krause, Berg, Fei-Fei, CVPR2012

Some more examples: flat classifier vs. ours

red fox

Flat	hyena	Egyptian cat	orangutan	mantis	jelly fungus
Ours	**canine**	**carnivore**	**mammal**	**animal**	**living thing**

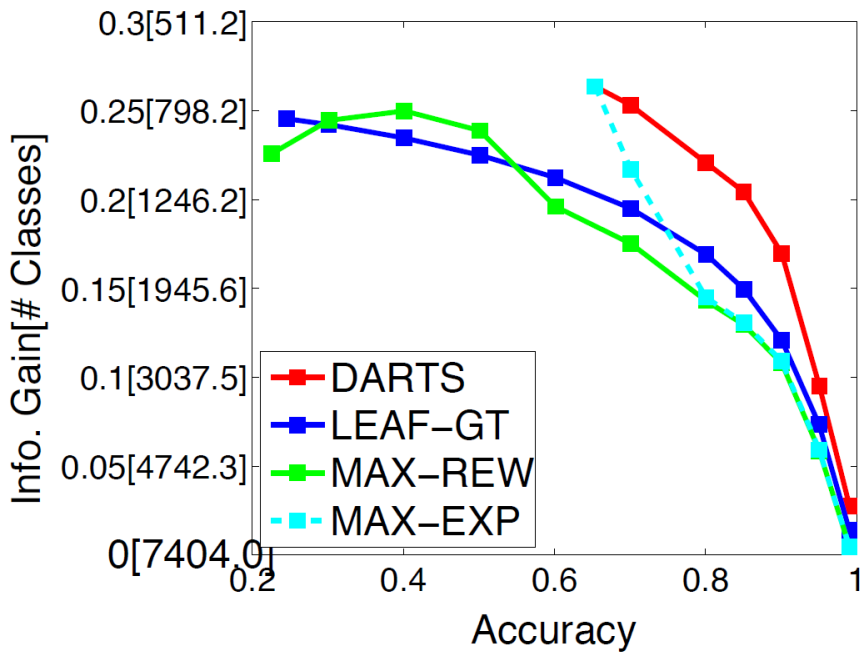
trimaran

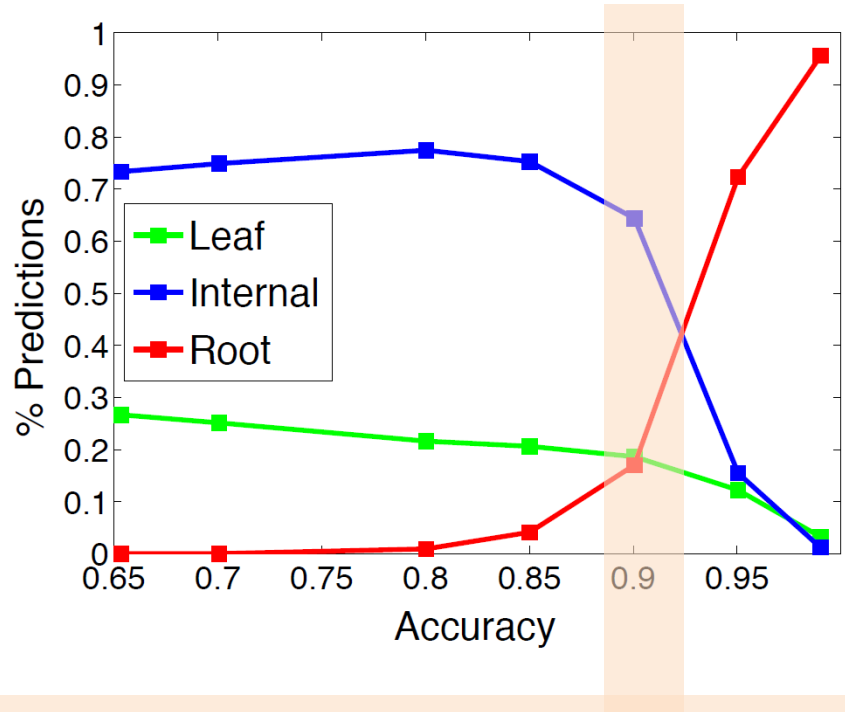
Flat	catamaran	submarine	airship	iron	electric guitar
Ours	**sailboat**	**watercraft**	**craft**	**artifact**	**artifact**

Deng, Krause, Berg, Fei-Fei, CVPR2012

Results

Datasets: 10,000 image classes from ImageNet (~9million images)
Baselines: Flat classifier with a reject option, etc.





10K classes: 90% accurate, 19% on leaf nodes, 64% non-root internal nodes, 17% "entity"

Deng, Krause, Berg, Fei-Fei, CVPR2012

Let's get extreme…



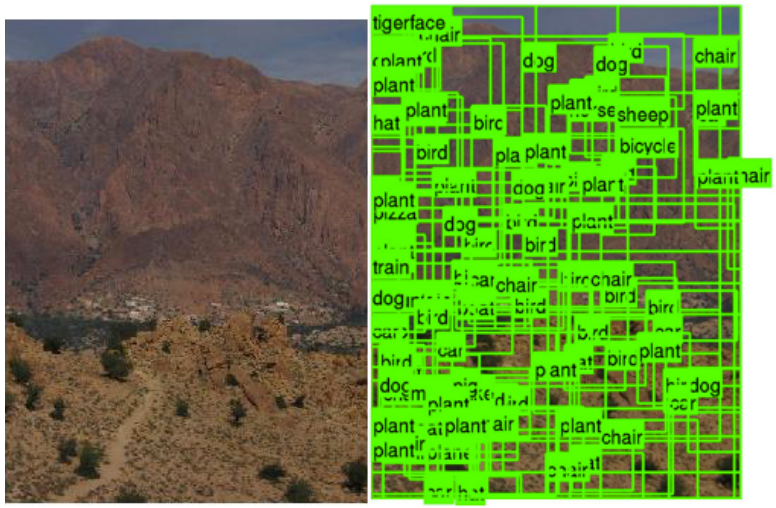

Flat	bobsled	pheasant	mortar	canoe
Ours	**vehicle**	**animal**	**edible fruit**	**watercraft**

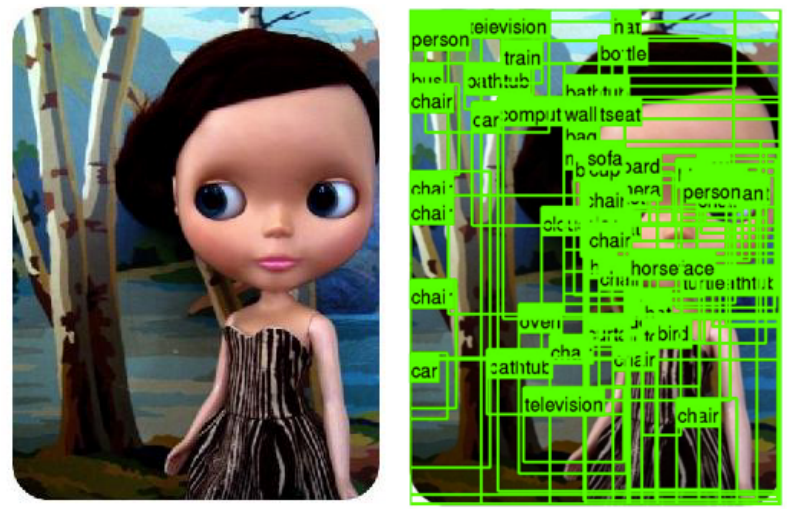



Flat	loggerhead	cannon	Bouvier des Flandres	grapefruit
Ours	**animal**	**animal**	**living thing**	**citrus fruit**

Deng, Krause, Berg, Fei-Fei, CVPR2012

Detection with many categories?

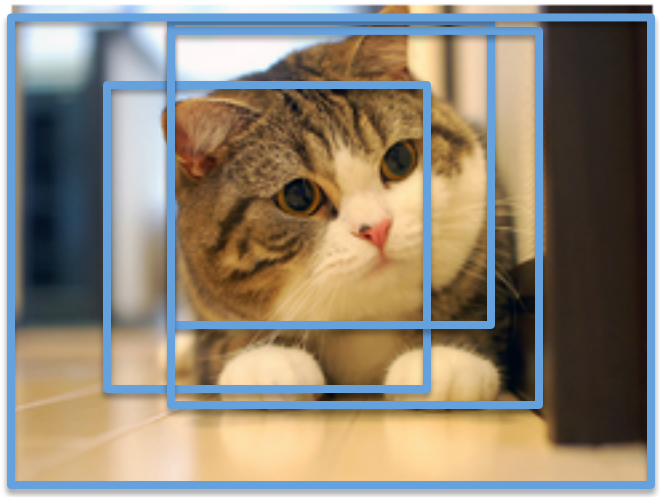




Ouch!

Use “hedging” approach for each bounding box

Input



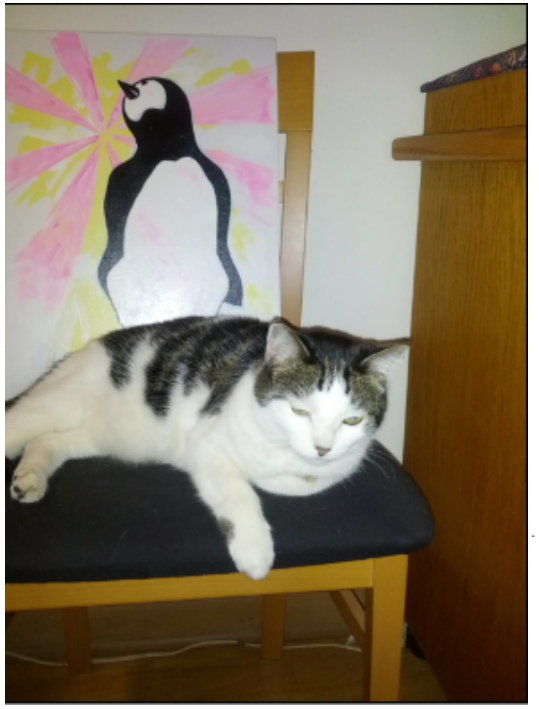
Hypothesize bounding boxes

e.g. from segmentation
or saliency operators

0.70	cat
0.80	cat
0.85	cat
0.90	domestic animal
0.95	entity
0.99	entity

- Evaluate hedging system for each bbox
- Output unique labels

Example detections and classification from “Hedging” system tuned to 70% expected accuracy on ImageNet data



NOT ImageNet data!

Output

0.70 face, covering, domestic animal, instrumentality, king penguin

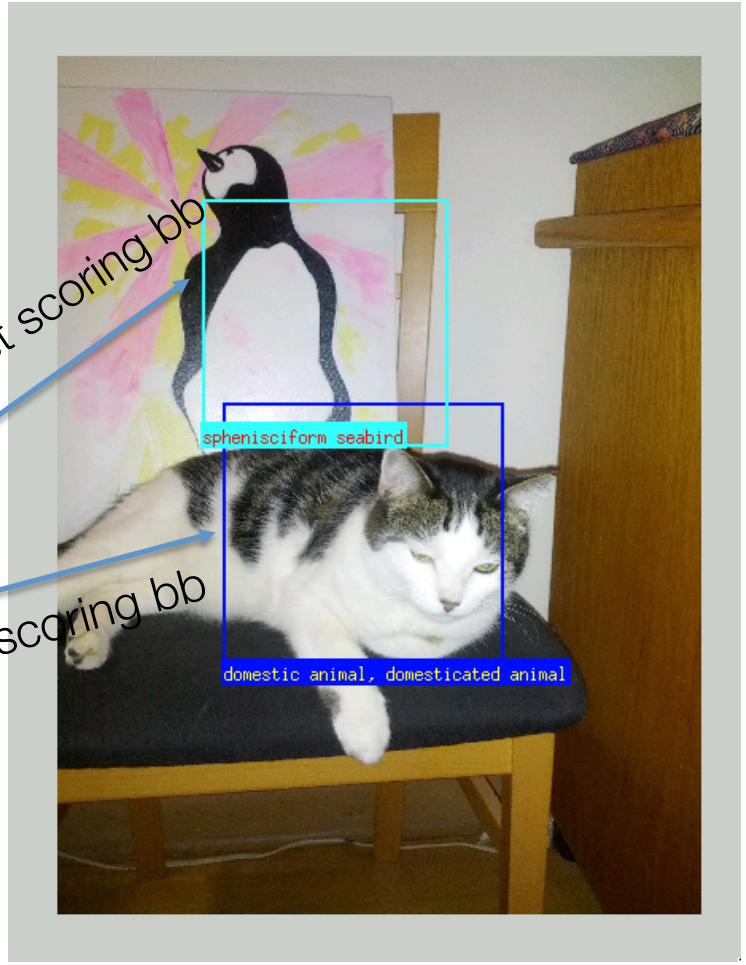
0.80 face, artifact, domestic animal, sphenisciform seabird

0.85 face, artifact, domestic animal, sphenisciform seabird

0.90 face, artifact

Highest scoring bb

Highest scoring bb



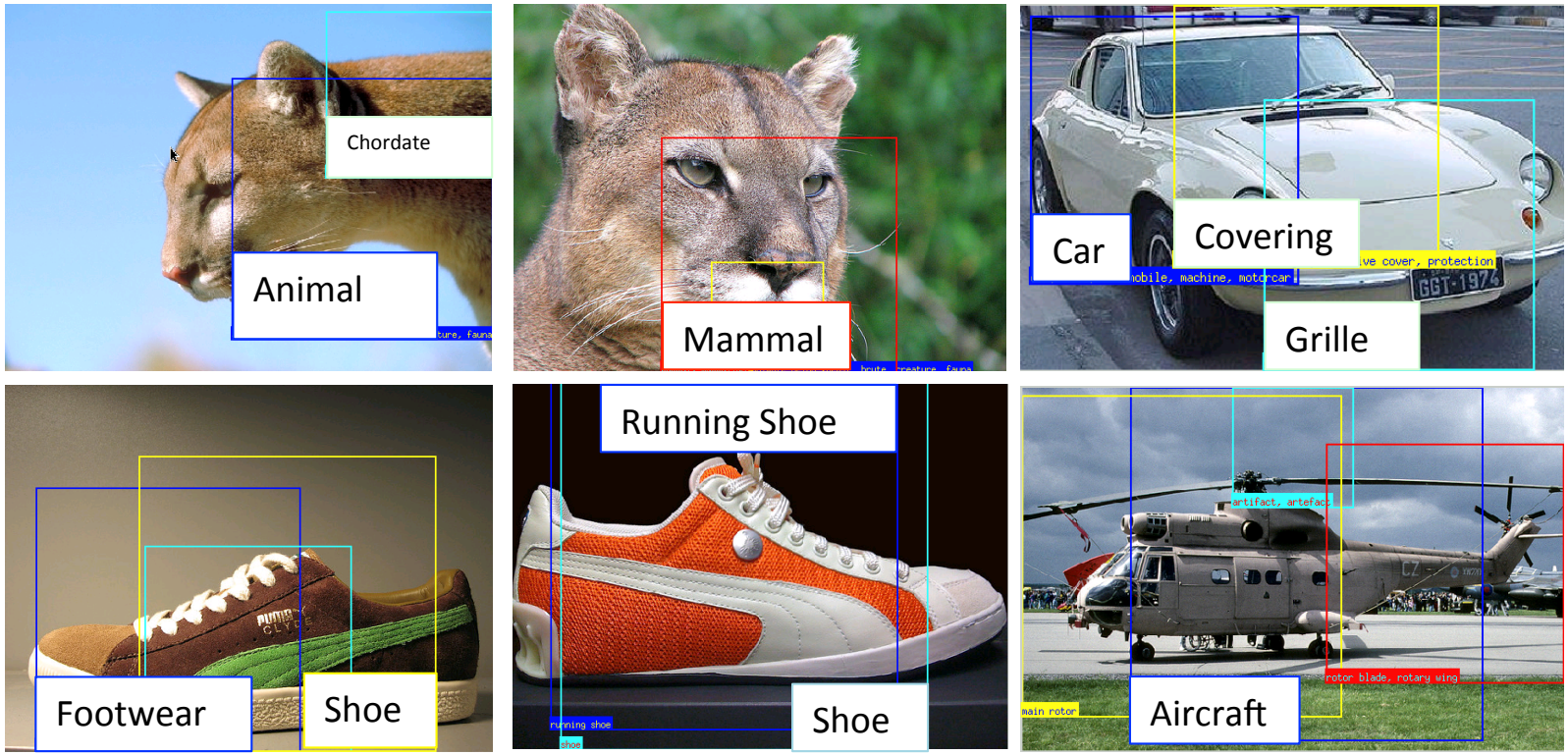
Google similar images baseline

Image size:
1944 × 2592

No other sizes of this image found.

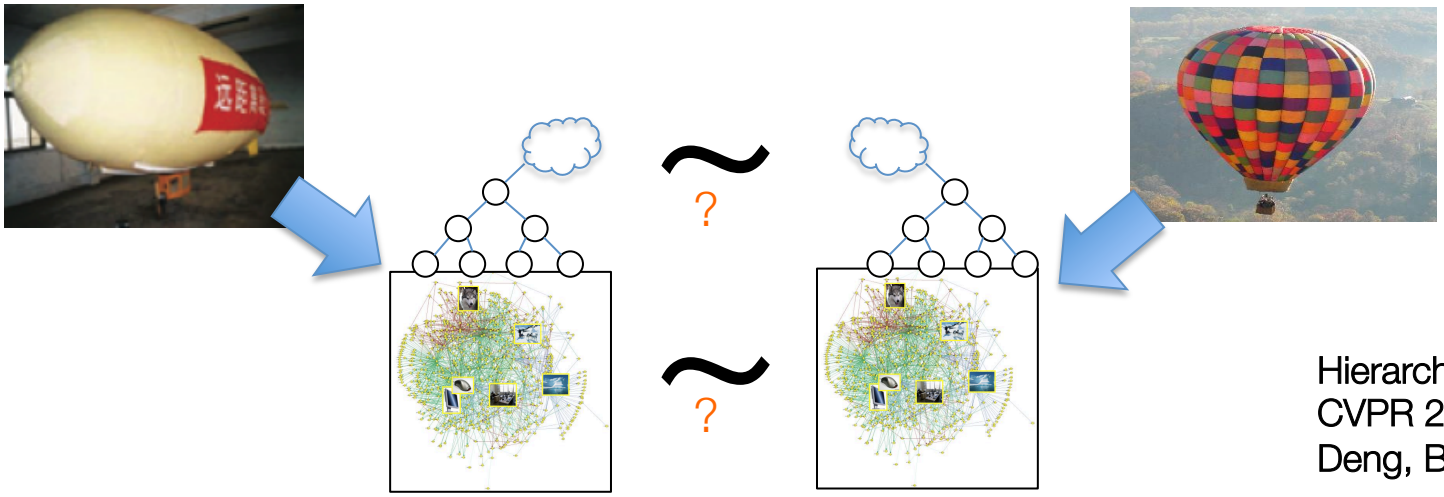
Visually similar images - Report images

Example detections and classification from “Hedging” system tuned on ImageNet data to 70% expected accuracy



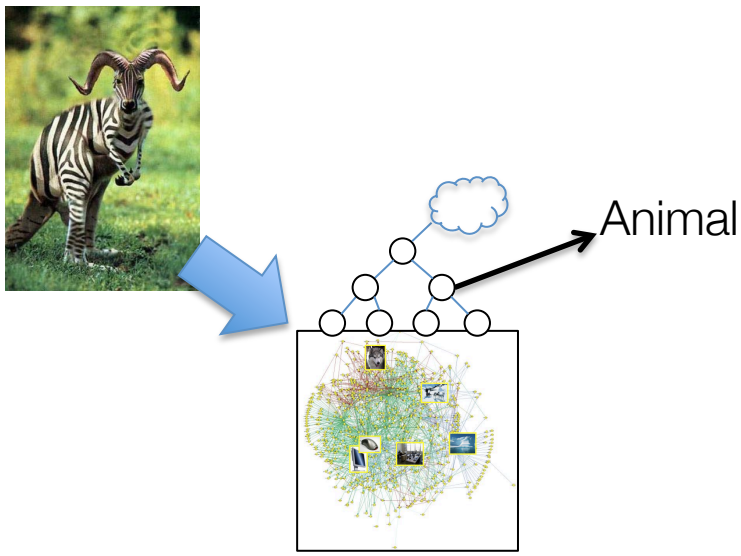
All google image search results for “Puma”

Using Large-Scale Structured Label Spaces



Hierarchical Semantic Indexing
CVPR 2010
Deng, Berg, Fei-Fei

Using Large-Scale Structured Label Spaces

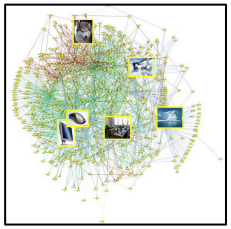


First 10,000 class object detection

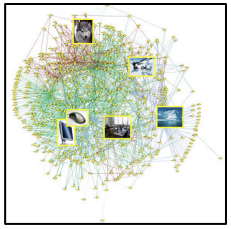
Hedging your bets
CVPR 2012
Deng, Krause, Berg, Fei-Fei

Using Large-Scale Structured Label Spaces

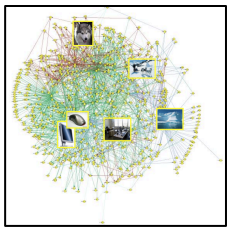
The old castle's tower was crumbling



New white paint on the bell tower



The view downtown toward the BCA tower.



Train a linear model to recognize images with "Tower" in their caption, using the 8k classifier outputs as features

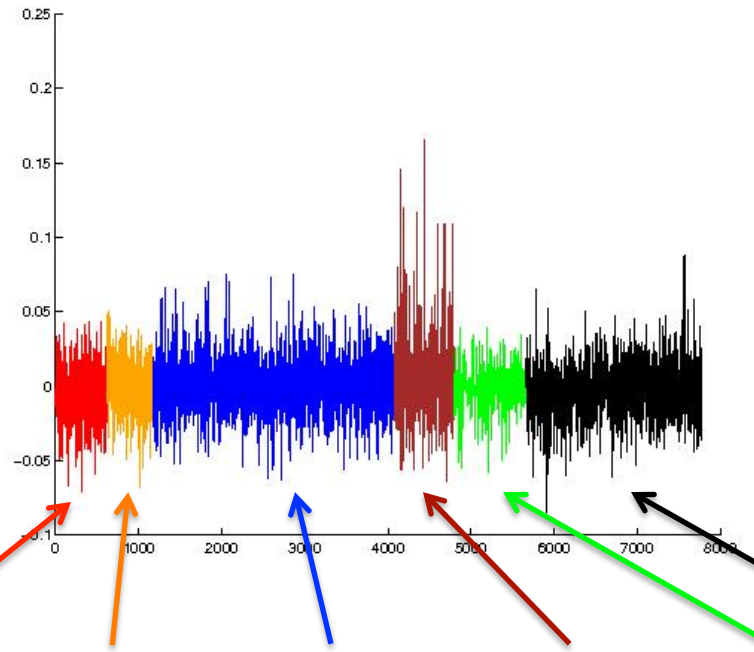
Look at the learned weights…

WORK IN PROGRESS
Contact Alex Berg

Work in progress, contact Alex Berg

Meaning from large-scale computer vision

Weights learned to recognize images with “Tower” in caption



Mammals Birds Instruments Structures Plants Other

Top weighted classifier outputs

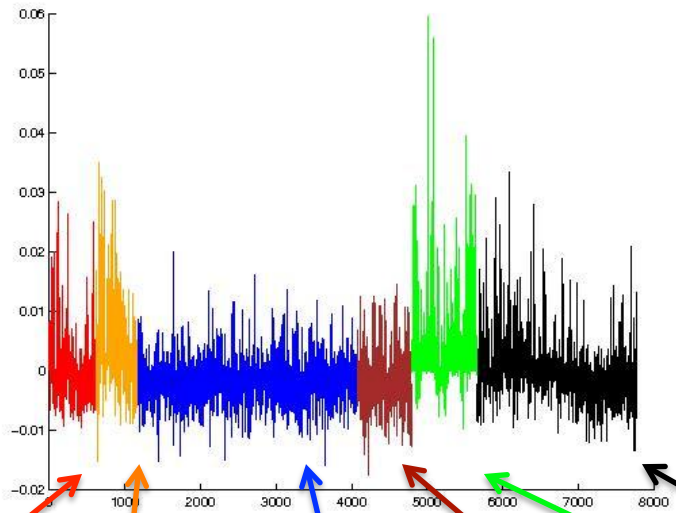
campanile, belfry (0.165488)
skyscraper (0.145514)
church tower (0.119316)
high-rise, tower block (0.118820)
battlement, crenelation, crenellation (0.108992)
alcazar (0.108717)
control tower (0.106499)

Weights learned over outputs of ~8k classifiers

Work in progress, contact Alex Berg

Meaning from large-scale computer vision

Weights learned to recognize images with “Tree” in caption



Mammals Birds Instruments Structures Plants Other

Top weighted classifier outputs

Brazilian rosewood, caviuna wood, jacaranda, Dalbergia nigra (0.059477)
redbud, Cercis canadensis (0.055979)
bristlecone pine, Rocky M
pine, Pinus aristata (0.039
mangrove, Rhizophora ma
frogmouth (0.035104)
bracket fungus, shelf fungu
barred owl, Strix varia (0.0
Japanese apricot, mei, Pru
mume (0.031360)
snag (0.031075)

A pair of Tawny Frogmouths resting in a tree fork during the day

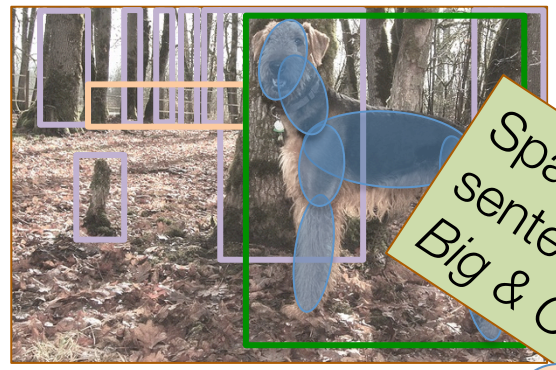
Weights learned over outputs of ~8k classifiers

Label space?

http://www.flickr.com/photos/sgcallawayimages/3306849049/

Dog — Single label

Dog, Tree, Fence, Leaf — Multiple labels



Space of sentences Big & Complex

lization

$\mathbb{U}$

"Happy shaggy airedale poses in the autumn forest." — Description

But again, work on using computer vision for this uses no more than 100s of labels.*

Composing captions
guessing game

a) monkey playing in the tree canopy, Monte Verde in the rain forest

b) capuchin monkey in front of my window

c) monkey spotted in Apenheul Netherlands under the tree

d) a white-faced or capuchin in the tree in the garden

e) the monkey sitting in a tree, posing for his picture

Composing captions
guessing game

a) monkey playing in the tree canopy, Monte Verde in the rain forest

b) capuchin monkey in front of my window

c) monkey spotted in Apenheul Netherlands under the tree

d) a white-faced or capuchin in the tree in the garden

e) the monkey sitting in a tree, posing for his picture

Collective Phrase Fusion for Natural Image Descriptions

Kuznestova et al ACL 2012
With Tamara Berg's lab at Stony Brook

Object appearance → NP: the dirty sheep

Object pose → VP: meandered along a desolate road

Scene appearance → PP: in the highlands of Scotland

Region appearance & relationship → PP: through frozen grass

↓

Example Composed Description: the dirty sheep meandered along a desolate road in the highlands of Scotland through frozen grass

1 million captioned photos!

SBU Captioned Photo Dataset

The Egyptian cat statue by the floor clock and perpetual motion machine in the pantheon

Little girl and her dog in northern Thailand. They both seemed interested in what we were doing

Interior design of modern white and brown living room furniture against white wall with a lamp hanging.

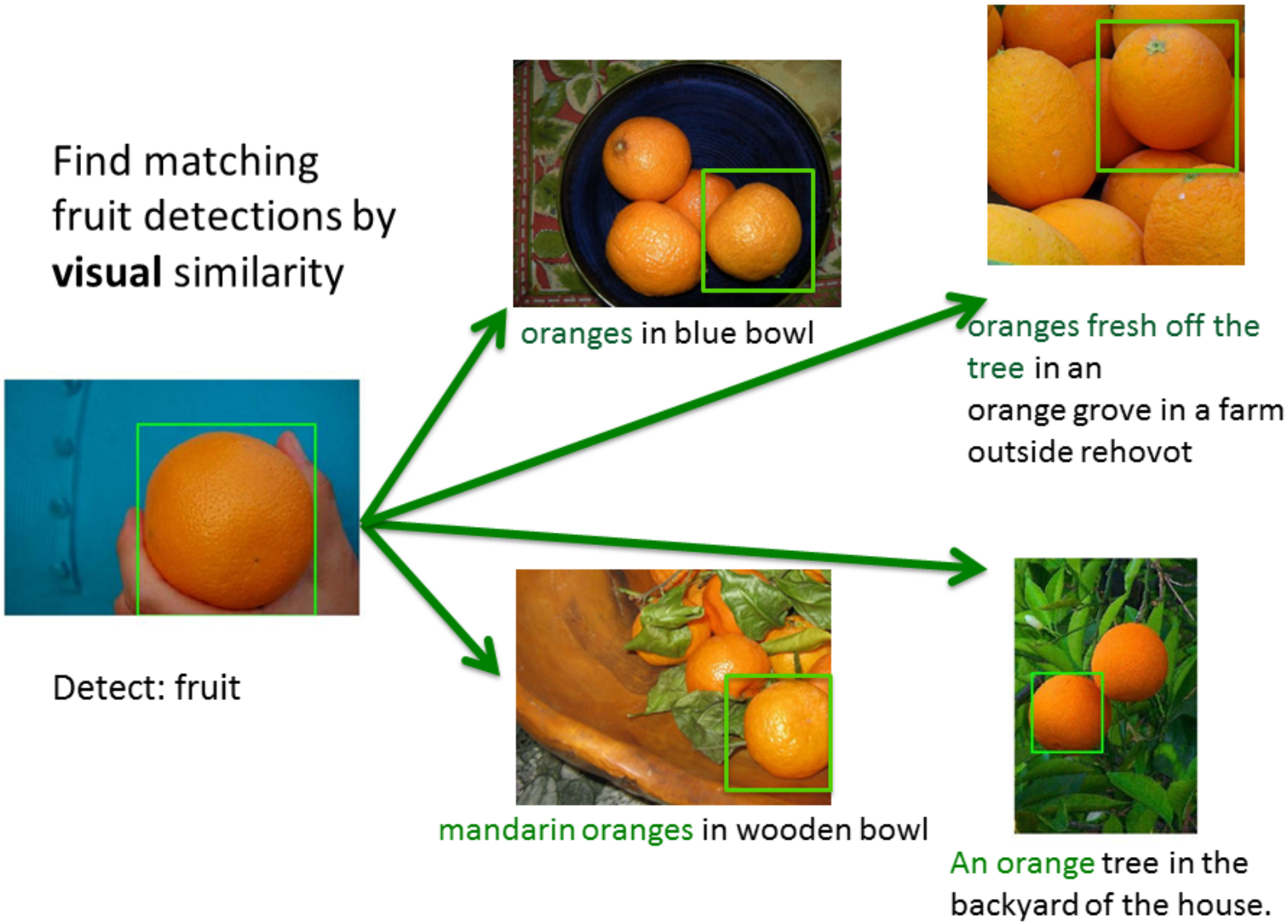
Man sits in a rusted car buried in the sand on Waitarere beach

Our dog Zoe in her bed

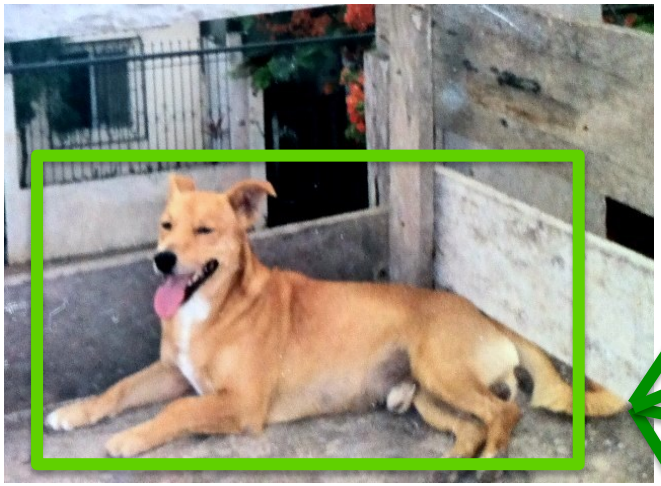
Emma in her hat looking super cute

http://tamaraberg.com/sbucaptions

Retrieving noun phrases from similar object detections

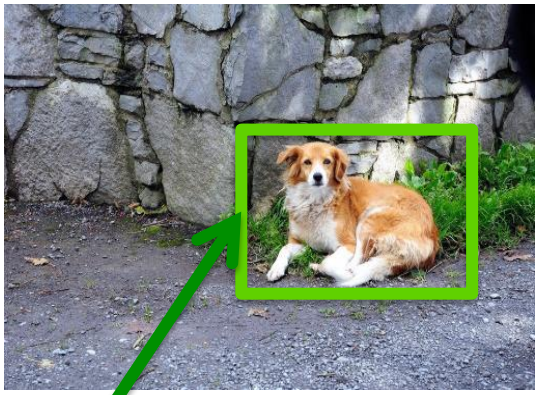


Retrieving verb phrases from similar object detections

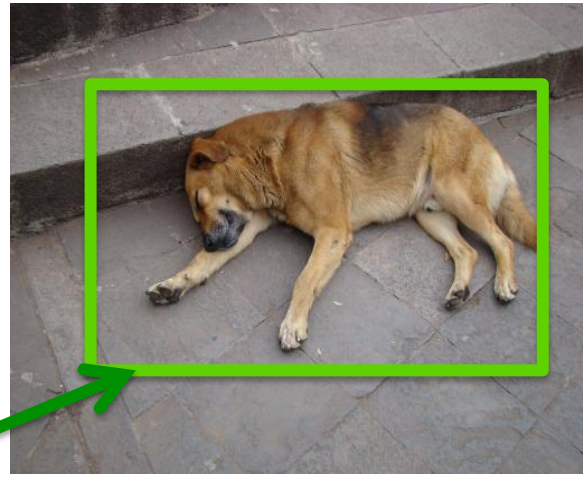


Detect: dog

Find matching dog detections by visual similarity



Contented dog just laying on the edge of the road in front of a house..



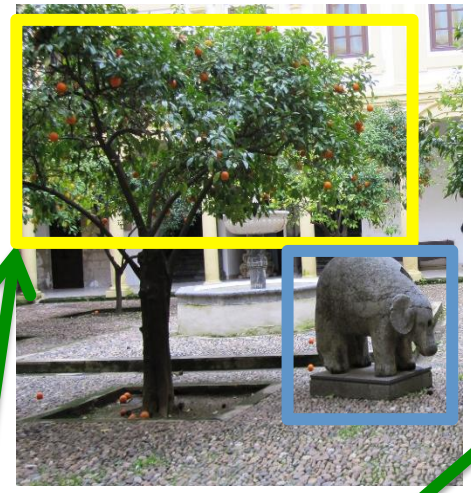
Peruvian dog sleeping on city street in the city of Cusco, (Peru)

this dog was laying in the middle of the road on a back street in jaco

Closeup of my dog sleeping under my desk.

Retrieving prepositional phrases from region + detection matches

Find matching region detections using appearance + arrangement

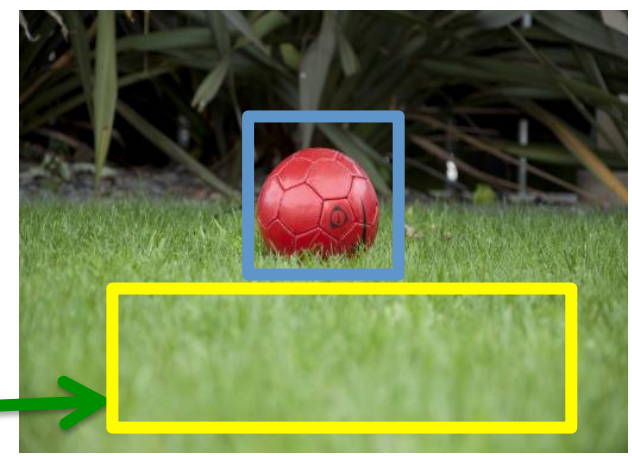


Cordoba - lonely elephant under an orange tree...

I positioned the chairs around the lemon tree -- it's like a shrine

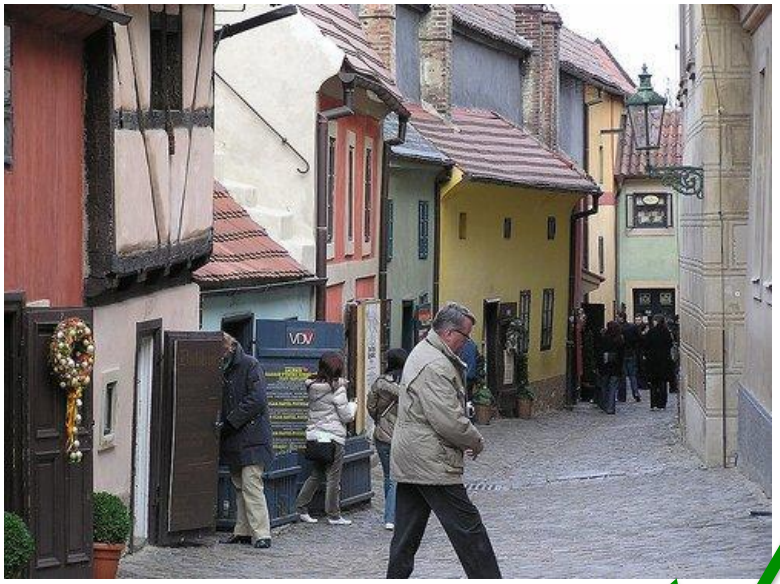
Object: car

Comfy chair under a tree.



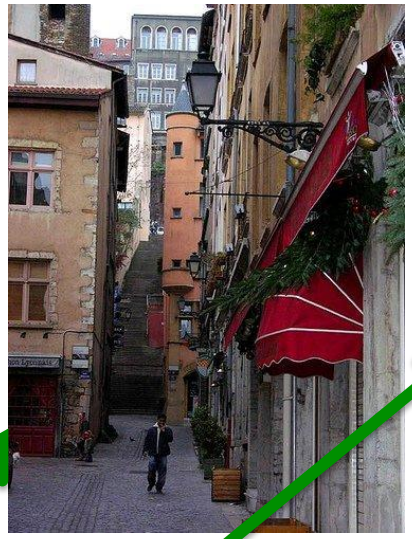
Mini Nike soccer ball all alone in the grass

Retrieving prepositional phrases from scene matches

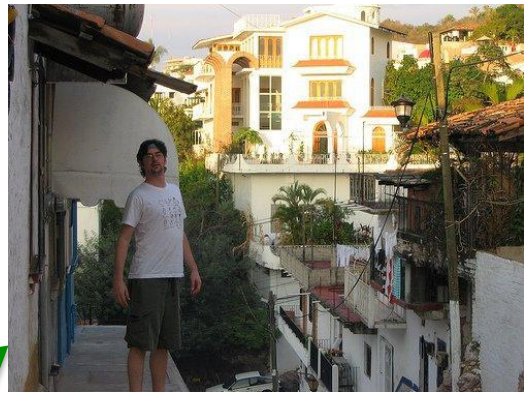


Extract scene descriptor

Find matching images by **scene similarity**

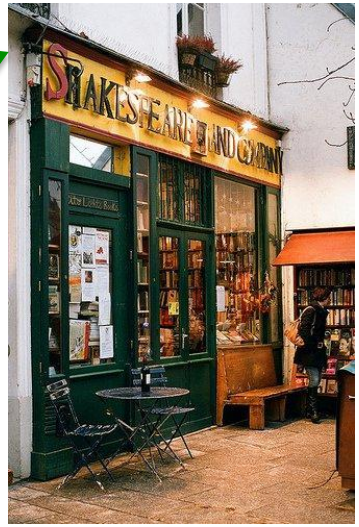


Pedestrian street in the Old Lyon with stairs to climb up the hill of fourviere



I'm about to blow the building across the street over with my massive lung power.

View from our B&B in this photo



Only in Paris will you find a bottle of wine on a table outside a bookstore

Composing Descriptions

Compose descriptions from phrases with ILP approach

- Vision constraints
  - Include visual detection/classification scores for phrase selection
- Linguistic constraints
  - Allow only one phrase of each type
  - Enforce plural/singular agreement between NP and VP
- Discourse constraints
  - Prevent inclusion of repeated phrasing
- Phrase cohesion constraints
  - n-gram statistics between phrases
  - Co-occurrence statistics between head words of phrases (last word or main verb) to encourage longer range cohesion

Bad Results

Grammar Problems

Found MIT boy ***gave*** me this quizical expression.

One of the ***most shirt*** in the wall of the house.

Cognitive Absurdity

Here you can see ***a cross by the frog in the sky***.

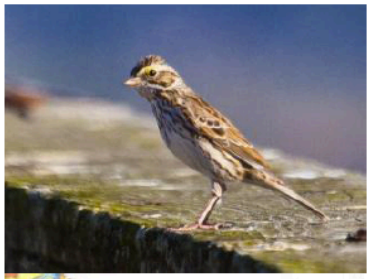
Content Irrelevance

Here you can see ***a bright red flower*** taken near our apartment in Torremolinos the Costa Del Sol.

This is ***a shoulder bag*** with a blended rainbow effect.

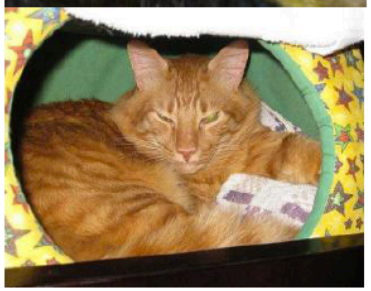
Visual Turing Test

Proposed approach vs Original Human Written Caption



ILP: This is a photo of this bird hopping around eating things off of the ground by river.

Human: IMG_6892 Lookn up in the sky its a bird its a plane its ah..... you



ILP: Taken in front of my cat sitting in a shoe box. Cat likes hanging around in my recliner.

Human: H happily rests his armpit on a warm Gatorade bottle of water (a small bottle wrapped in a rag)

In some cases (16%), ILP generated captions were preferred over human written ones!

ILP: This is a sporty little red convertible made for a great day in Key West FL. This car was in the 4th parade of the apartment buildings.

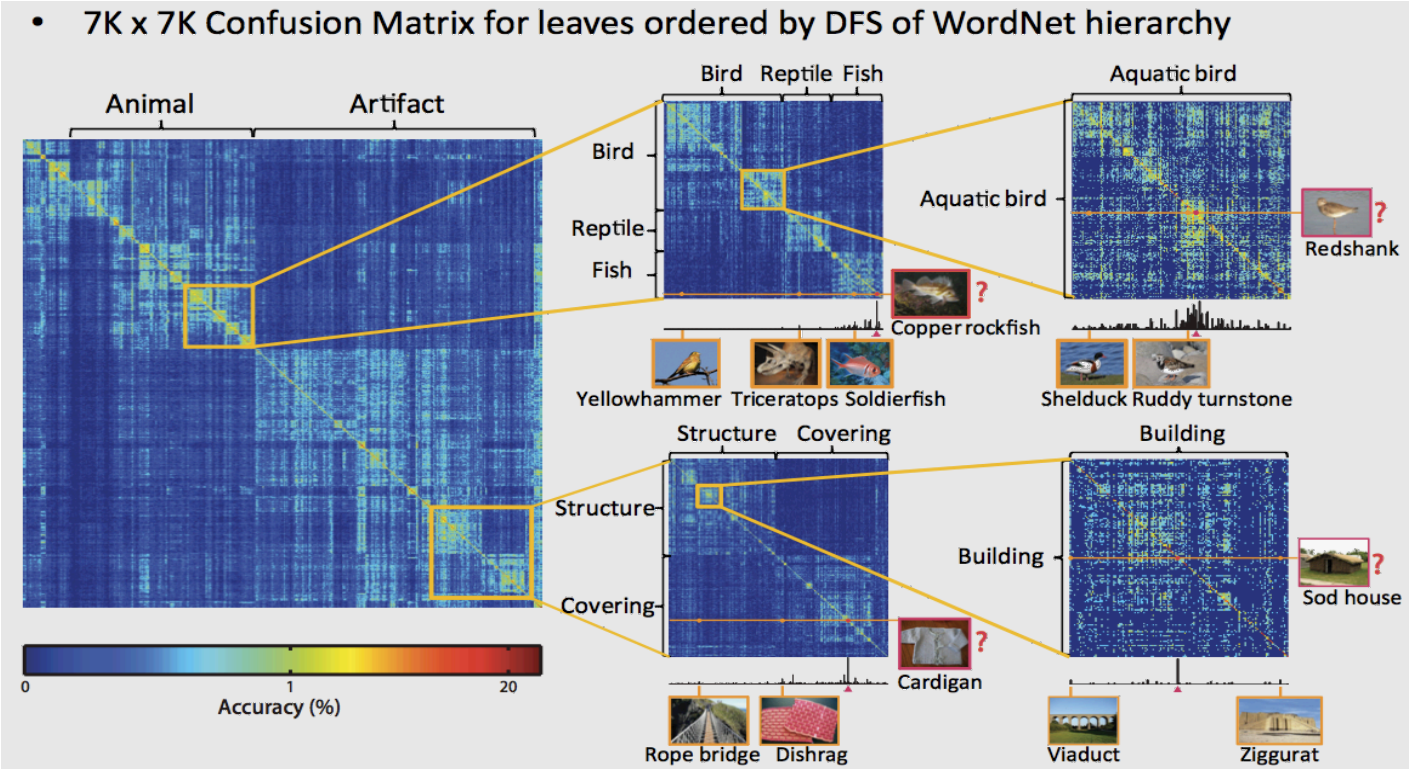
Human: Hard rock casino exotic car show in June

ILP: I like the way the clouds hanging down by the ground in Dupnitsa of Avikwalal.

Human: Car was raised on the wall over a bridge facing traffic..paramedics were attending the driver on the ground

So far – all about using the output of classifiers…

- Avoid wasting the **CPU millennia** invested.
- What about training the classifiers themselves?
- Structure between classes -> structure in learning



Scalability of computation

Multiclass classification

A function H_c is learned for each class c. Then at test time, an item x is classified as class $\text{argmax}_c H_c(x)$.

One-vs-all training constrains $H_c(x) > H_c(z)$, single-machine training constrains $H_c(x) > H_d(x)$,

for any item x of class c, and item z of class d, $d \neq c$.

-State of the art for large scale classification / detection
-We cannot afford this for large label spaces
-Our recent results indicate it is already undesirable computationally

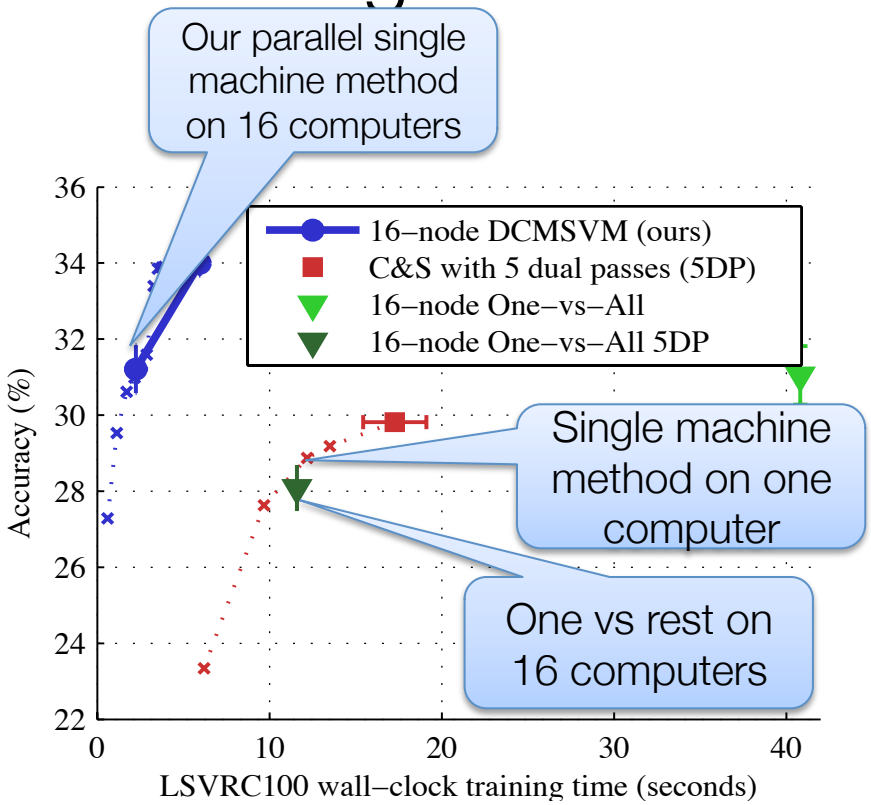
Basically structured prediction across classes

LIBLINEAR's binary solver
(dual coordinate descent)

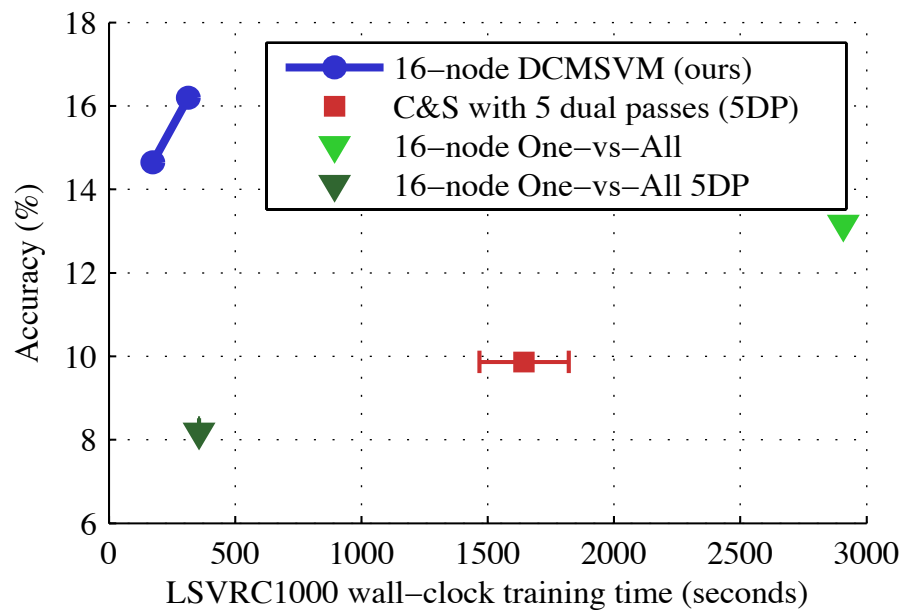
Cramer & Singer
(LIBLINEAR's multi-class)

DCMSVM: Distributed Parallel Training For Single-Machine Multiclass Classifiers

DCMSVM
CVPR 2012
Han, Berg



100 Class Problem



1000 Class Problem

Summary

- We are moving past image categorization
- Toward detection, parsing, attributes, & **description**
- Large structured label spaces are useful
  - Don't start over from scratch every time!
- We will never have fully labeled data
- Significant computational bottlenecks in training

Long Term Collaborators

Tamara Berg's lab
Stony Brook University

Fei-Fei Li's lab
STANFORD UNIVERSITY

Please see my papers for (many!) references.