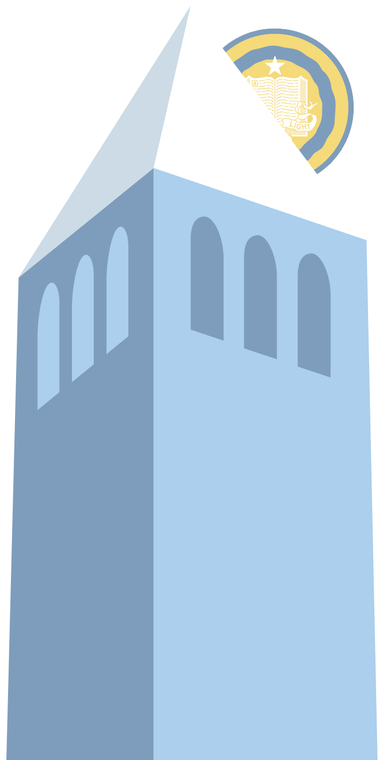
Anytime Recognition of Objects and Scenes



Sergey Karayev
UC Berkeley CS

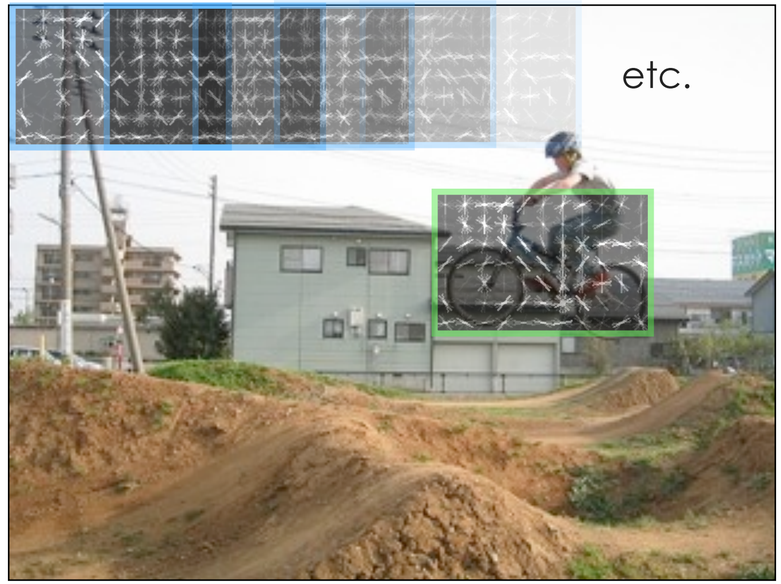
CVPR 2014
2014 May 14

Motivation: visual recognition problems

Q: Bike?

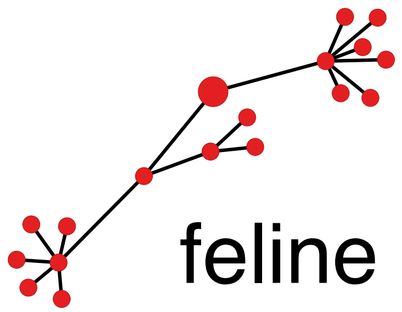
A:

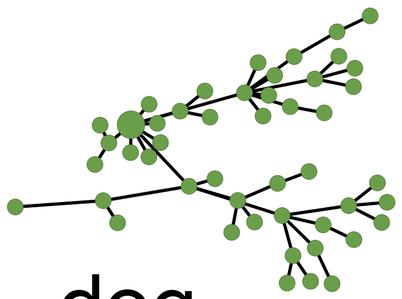
Classification

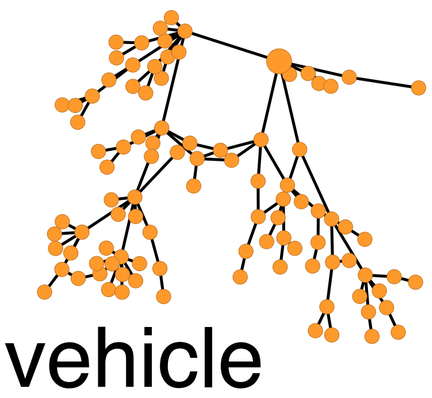


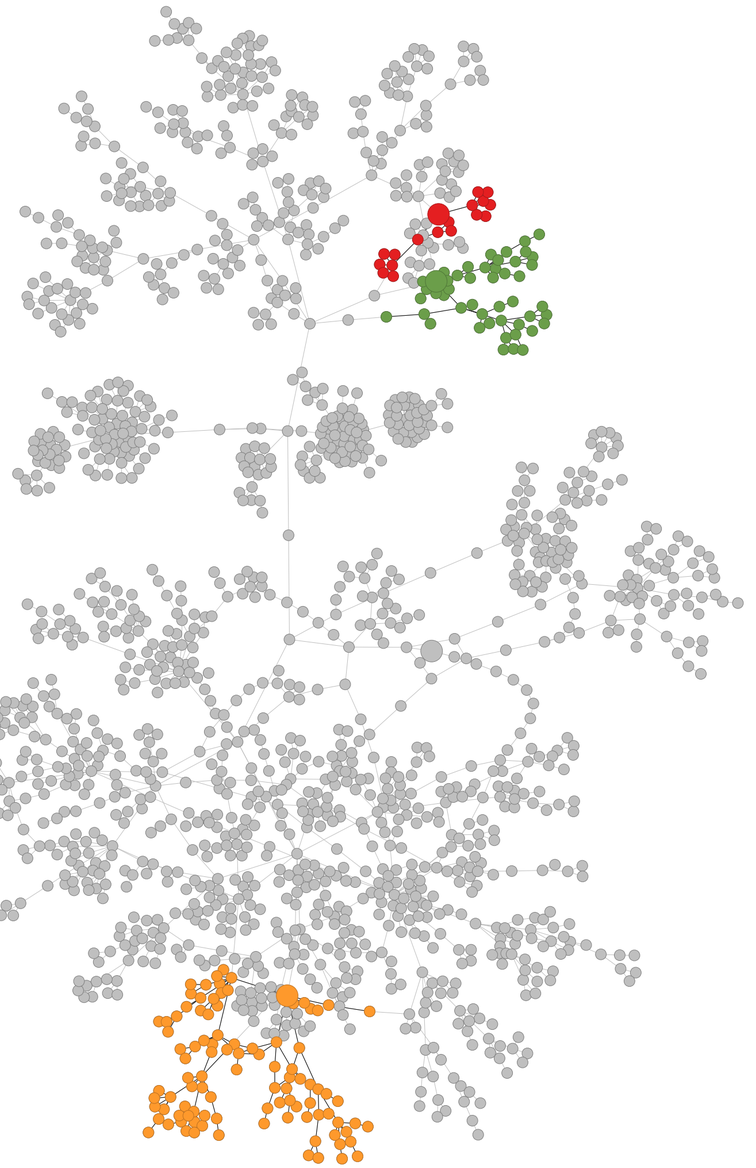
Detection

Motivation: lots of images, lots of classes





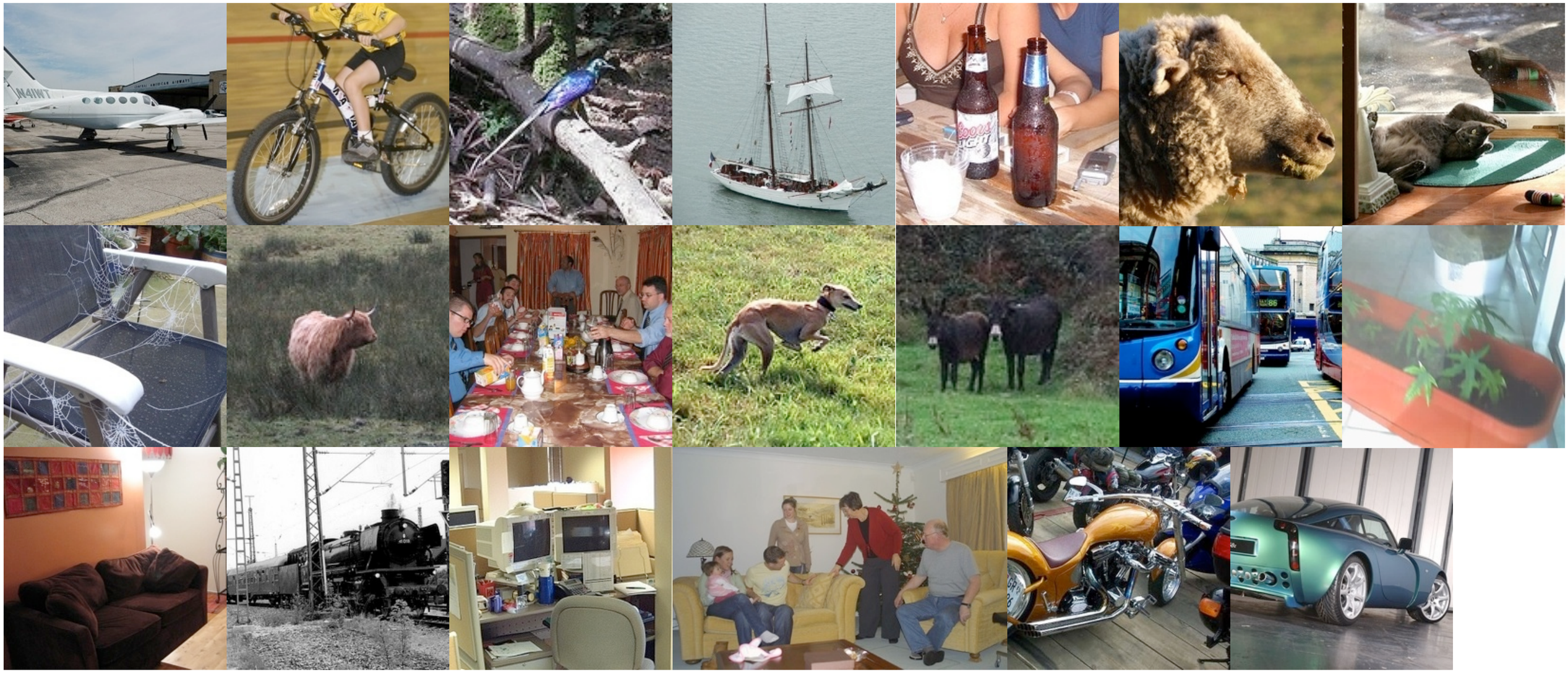




>10M images, 10K classes

IMAGENET

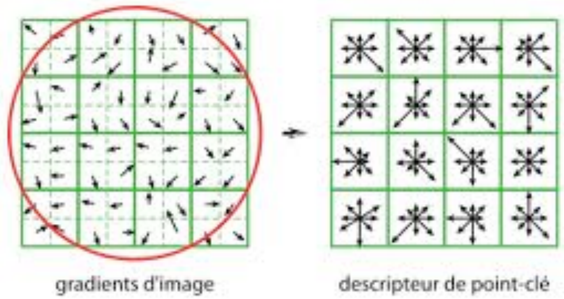
Motivation: highly varied scenes and objects



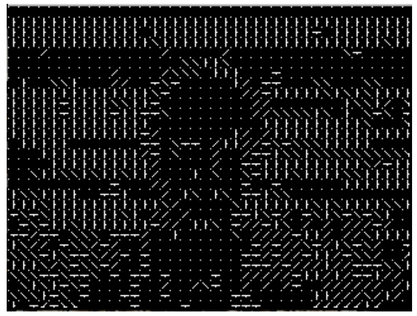
~20K images, 20 classes

PASCAL2
Pattern Analysis, Statistical Modelling and Computational Learning

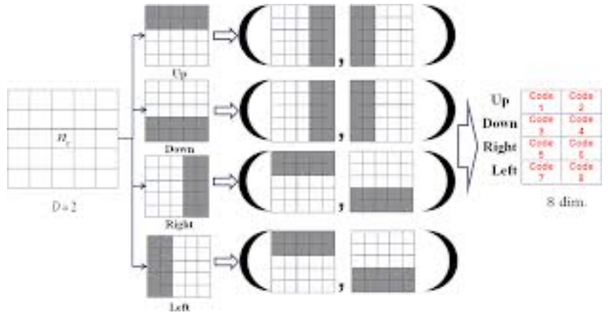
Motivation: lots of costly features



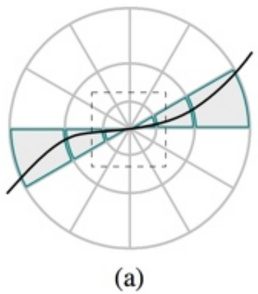
SIFT



HOG

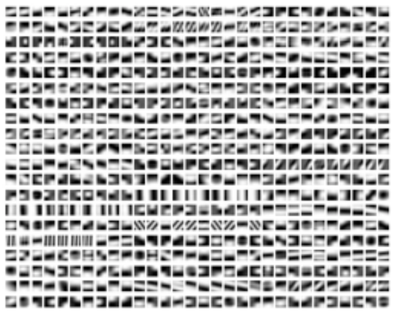


LBP

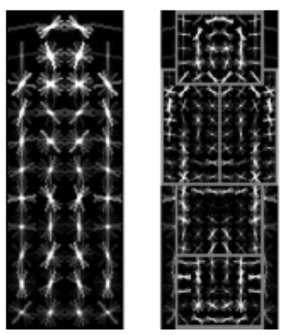


SSIM

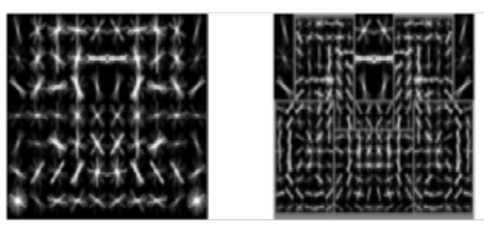
Color



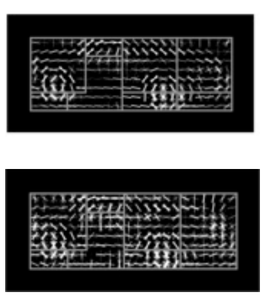
Textons



Person



Bike



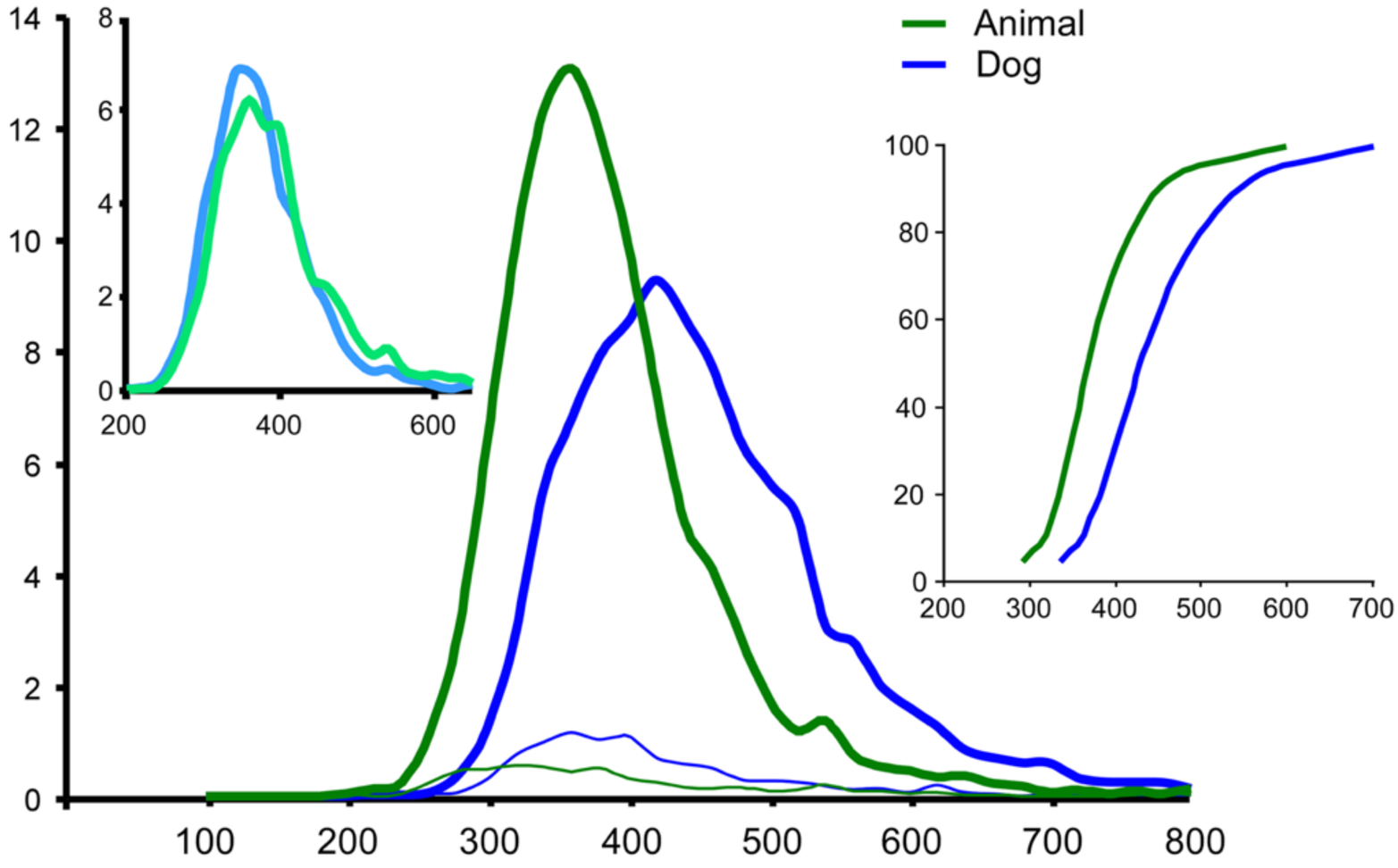
Car

Motivation: test-time budget

With a test-time budget, cannot compute everything.

And what if the budget is flexible?

Motivation: time course of human perception



PT
27 ms

There was a range of dark sploches in the middle of the picture, running from most of the way on the left side, to all the way on the right side. This was surrounded primarily by a white or light gray color. (Subject: KM)

PT
40 ms

I saw a very brght object, shaped in a pyramidal shape. There was somethin black in the front, but I couldn't tell what it was. (Subject: JB)

PT
67 ms

Possibly outdoors. maybe a few ducks, or geese. Water in the background. (Subject: JL)

PT
500 ms

It was definately on a coast byt hte ocean with a large [r]ock in the forground and atleast three bird sitting on the rock. (Subject: CC)

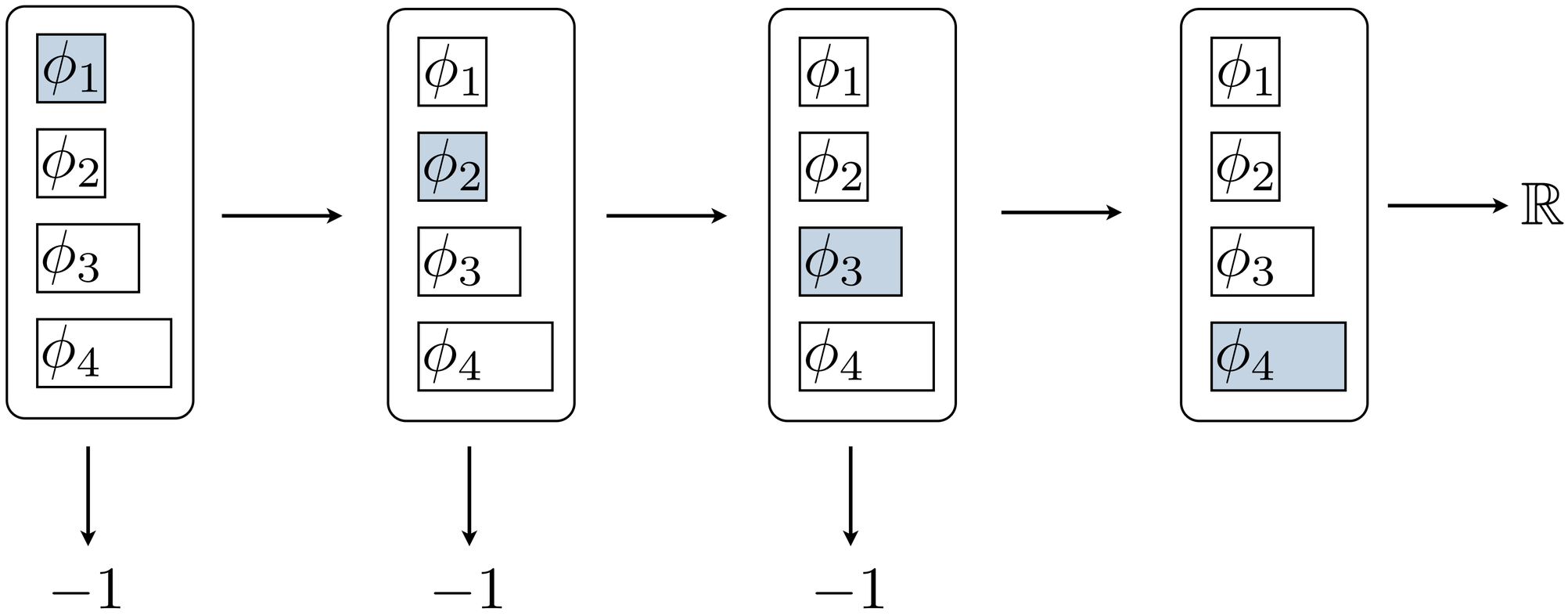
[1] M. J.-M. Macé, O. R. Joubert, J.-L. Nespoulous, and M. Fabre-Thorpe, “The time-course of visual categorizations: you spot the animal faster than the bird.,” PLoS One, vol. 4, no. 6, p. e5927, Jan. 2009.

[2] L. Fei-Fei, A. Iyer, C. Koch, and P. Perona, “What do we perceive in a glance of a real-world scene?,” J. Vis., Jan. 2007.

A: Anytime recognition.

Review: cascades

Viola & Jones (CVPR 2001), Bourdev & Brandt (CVPR 2005)

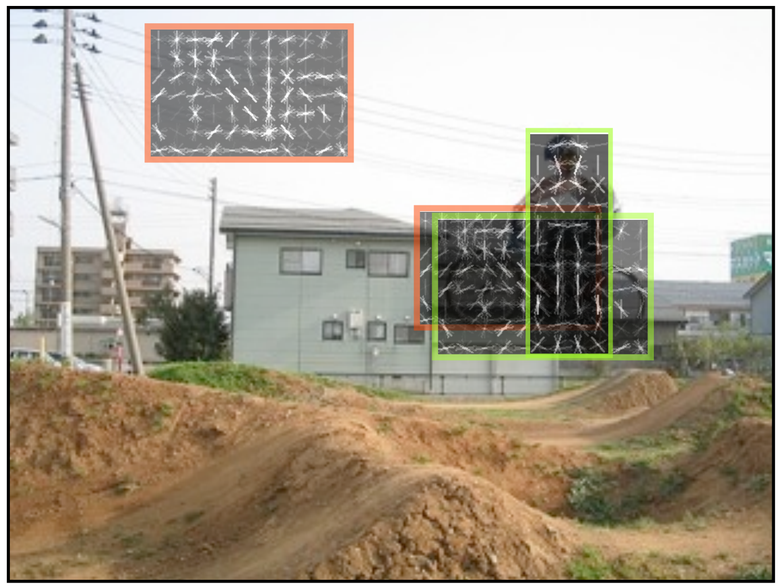


- Two actions: Reject and Continue.

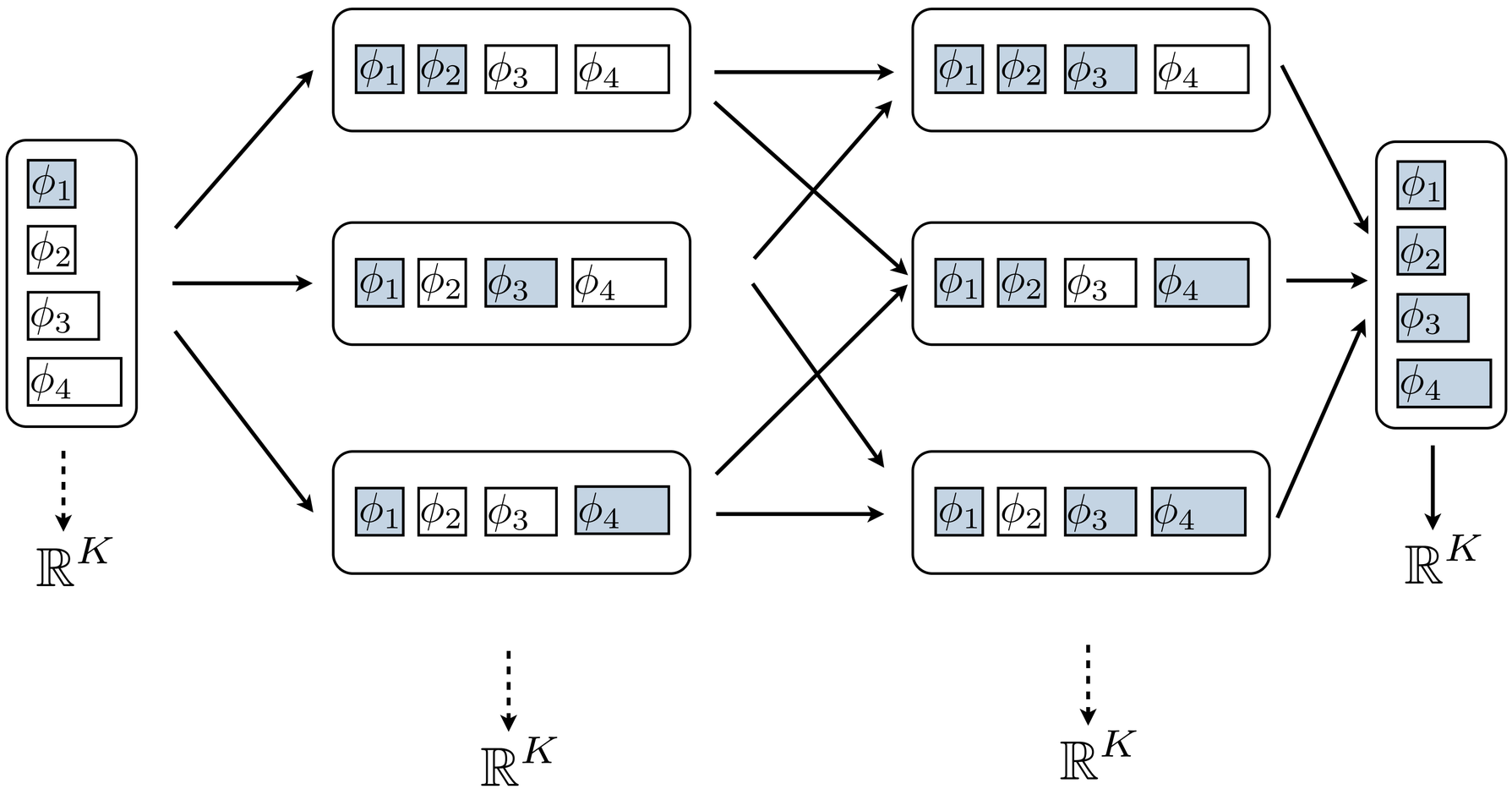
Review: cascades are not enough

Different instances benefit from different features.

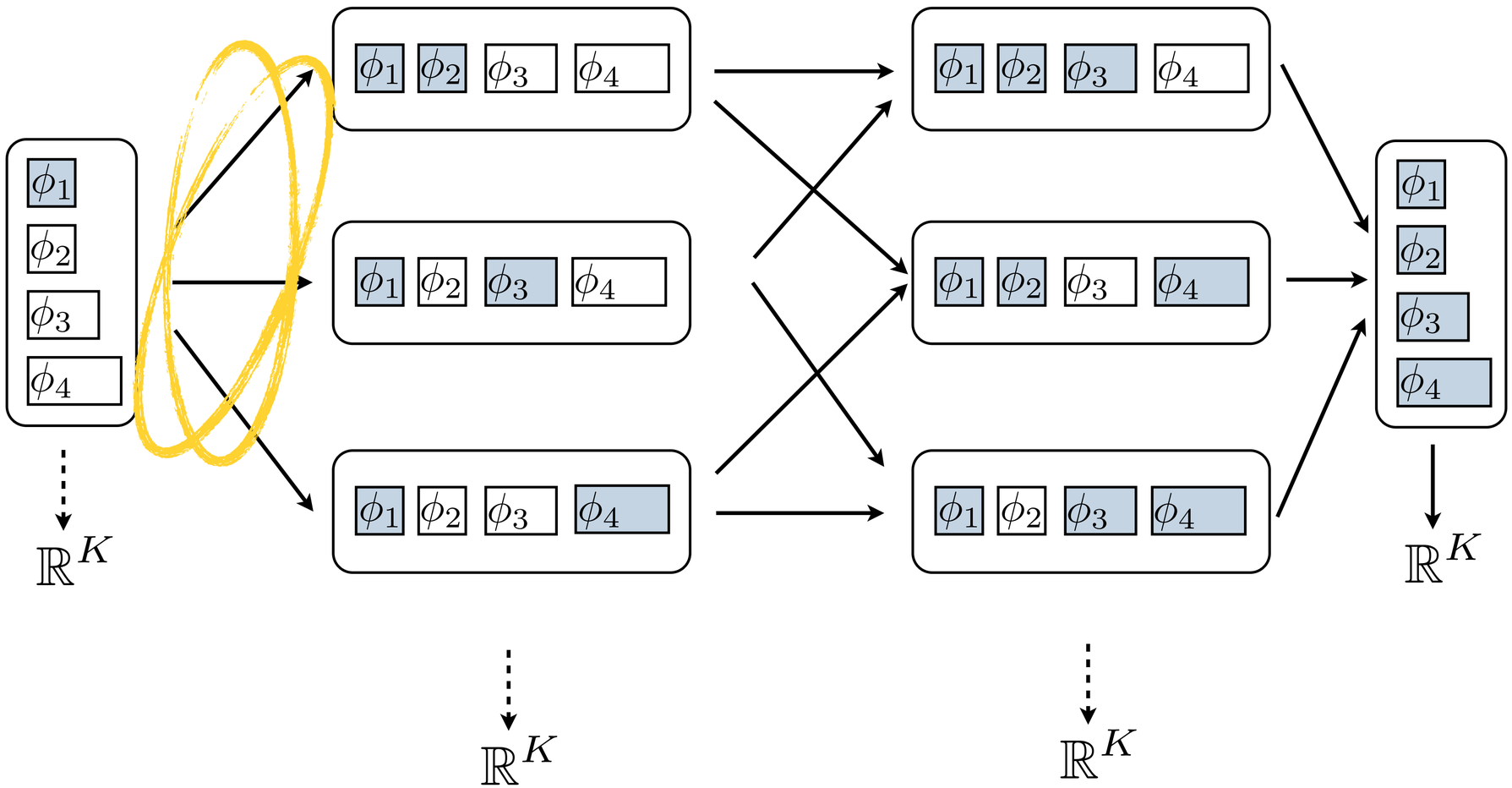
Co-occurrence signal



This work

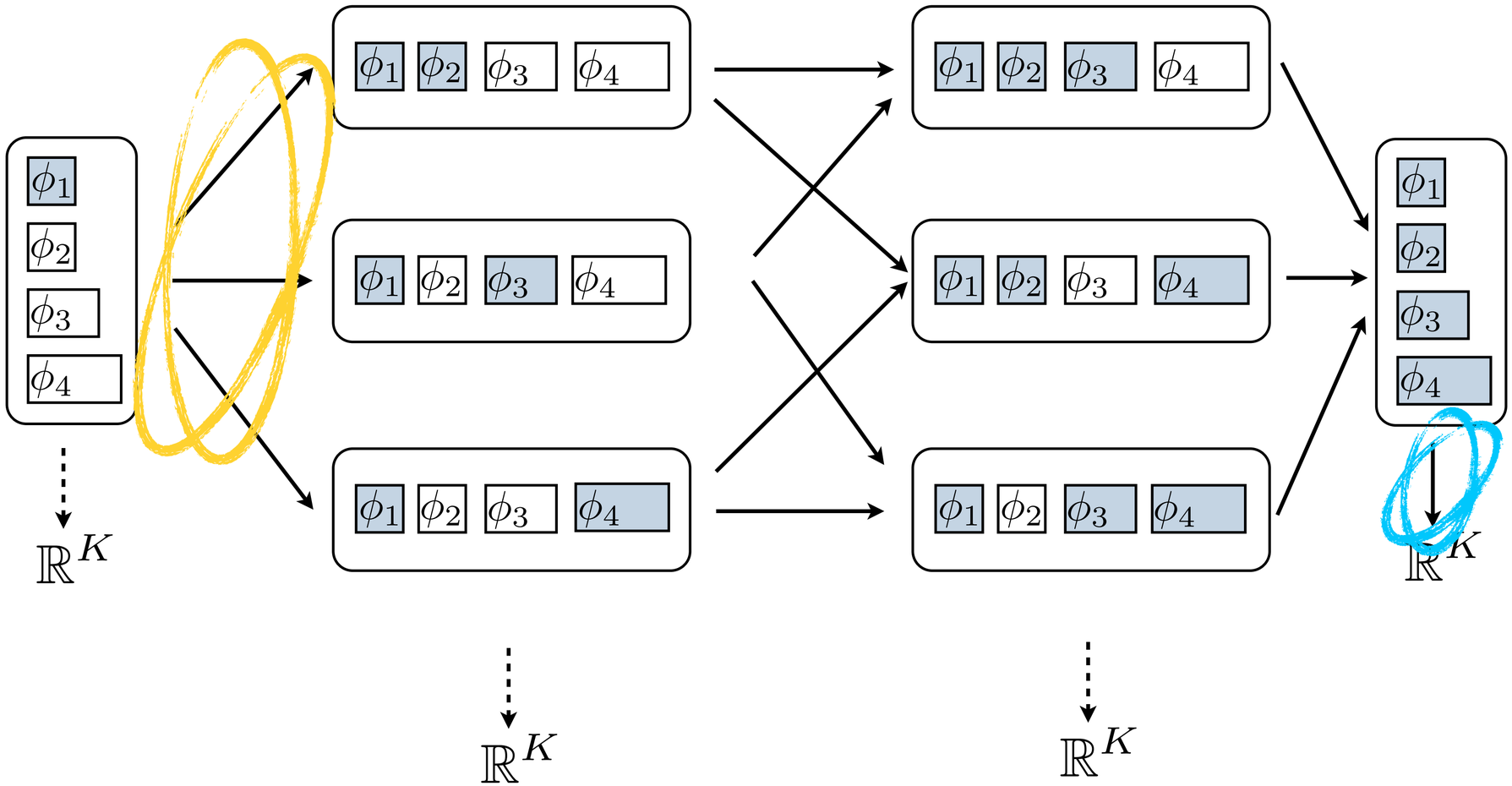


This work



- Action selection: non-myopic policy learned by MDP.

This work



- Action selection: non-myopic policy learned by MDP.
- Feature combination: linear.

Input: $\mathcal{D} = \{x_n, y_n\}_{n=1}^{N}$; $\mathcal{L}_{\mathcal{B}}$
Result: Trained π, g

$\pi_0 \leftarrow$ random;
for $i \leftarrow 1$ **to** *max_iterations* **do**
    States, Actions, Costs, Labels $\leftarrow$ `GatherSamples(`$\mathcal{D}$, π_{i-1}`)`;
    $g_i \leftarrow$ `UpdateClassifier(`*States, Labels*`)`;
    Rewards $\leftarrow$ `ComputeRewards(`*States, Costs, Labels,* $g_i, \mathcal{L}_{\mathcal{B}}, \gamma$`)`;
    $\pi_i \leftarrow$ `UpdatePolicy(`*States, Actions, Rewards*`)`;
end

Algorithm 1: Because reward computation depends on the classifier, and the distribution of states depends on the policy, g and π are trained iteratively.

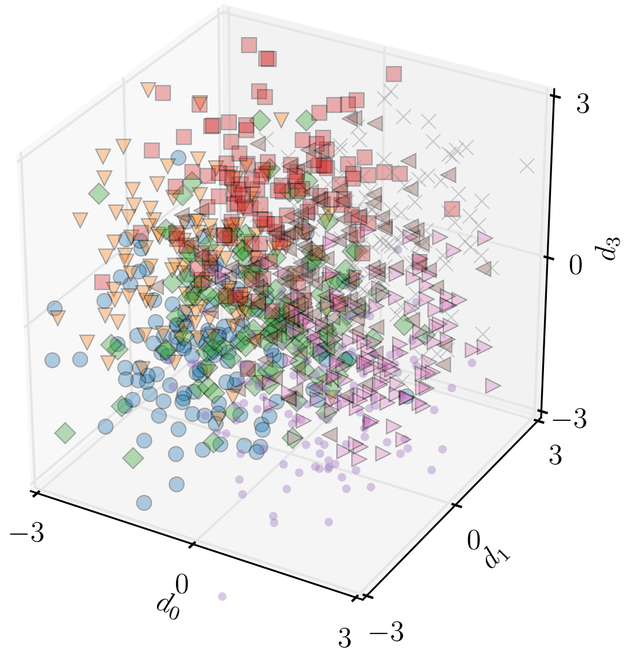
- Skipping over:
  - reinforcement learning
  - reward function definition
  - combining arbitrary subsets of features.
- Please see paper: http://sergeykarayev.com/recognition-on-a-budget/

We evaluate the following baselines:

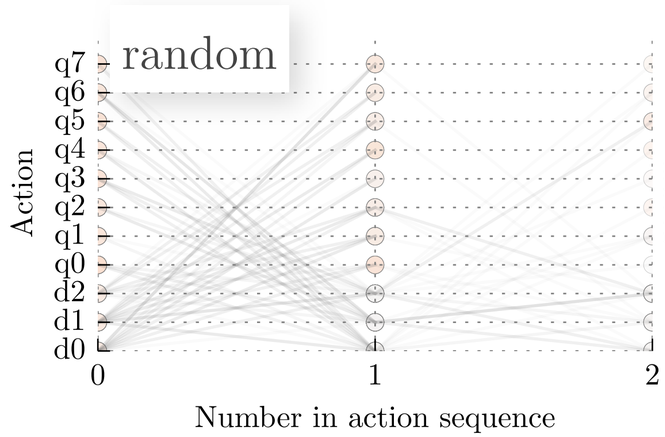
- **Static, greedy**: corresponds to best performance of a policy that does not observe feature values and selects actions greedily ($\gamma = 0$).
- **Static, non-myopic**: policy that does not observe values but considers future action rewards ($\gamma = 1$).
- **Dynamic, greedy**: policy that observes feature values, but selects actions greedily.

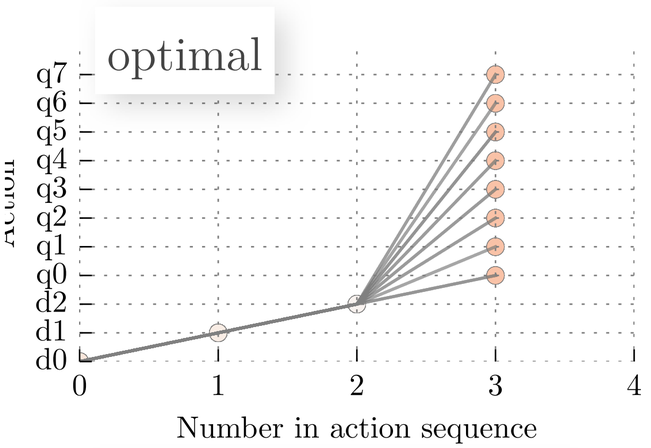
Our method is the **Dynamic, non-myopic** policy: feature values are observed, with full lookahead.

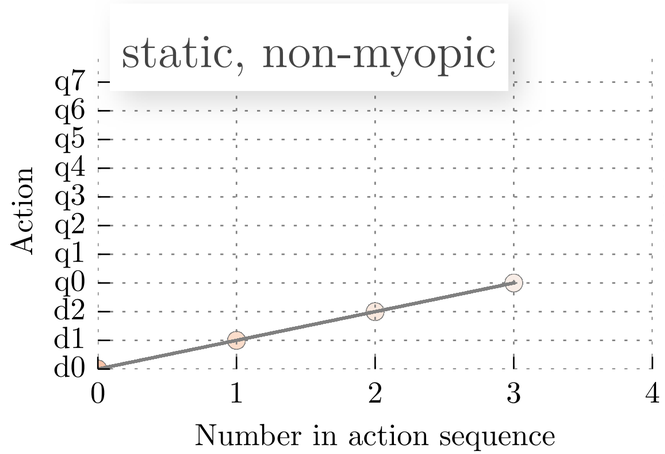
Synthetic Example

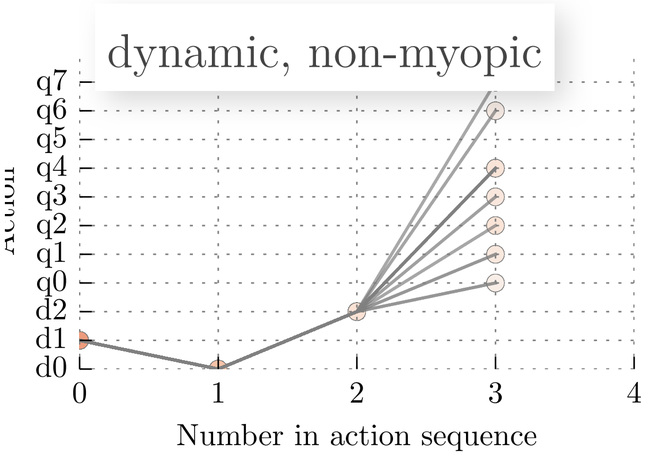


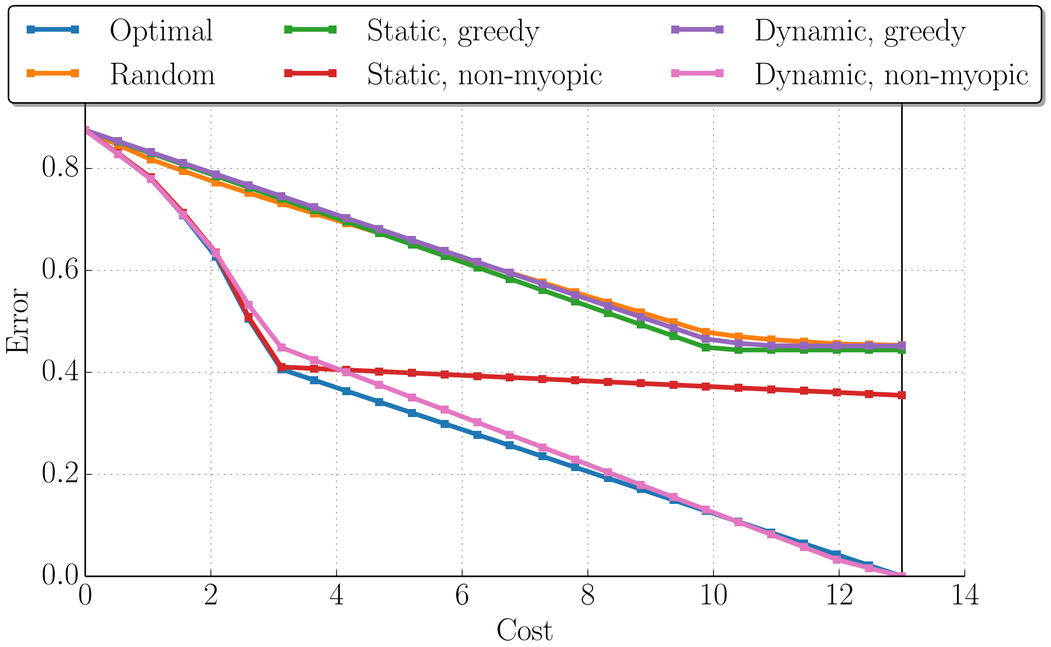
Feature	Number	Cost
d_i: sign of dimension i	D	1
q_o: label of datapoint, if in quadrant o	2^D	10

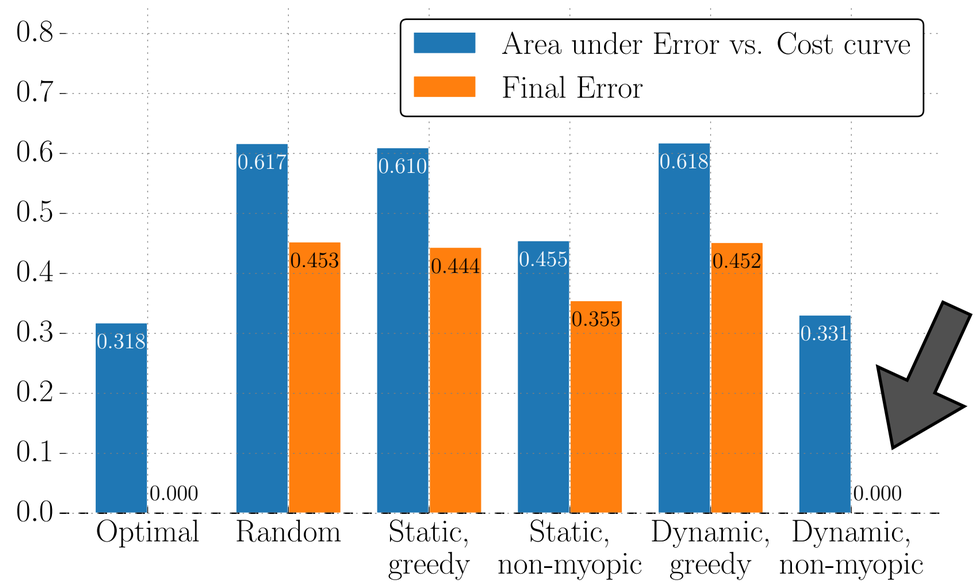




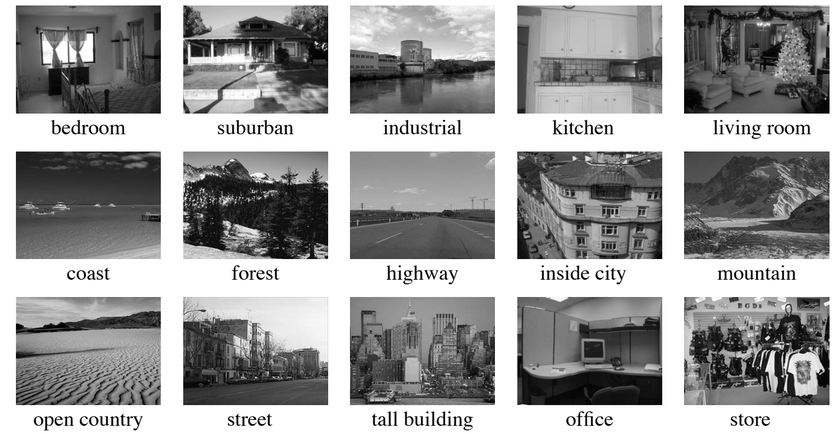


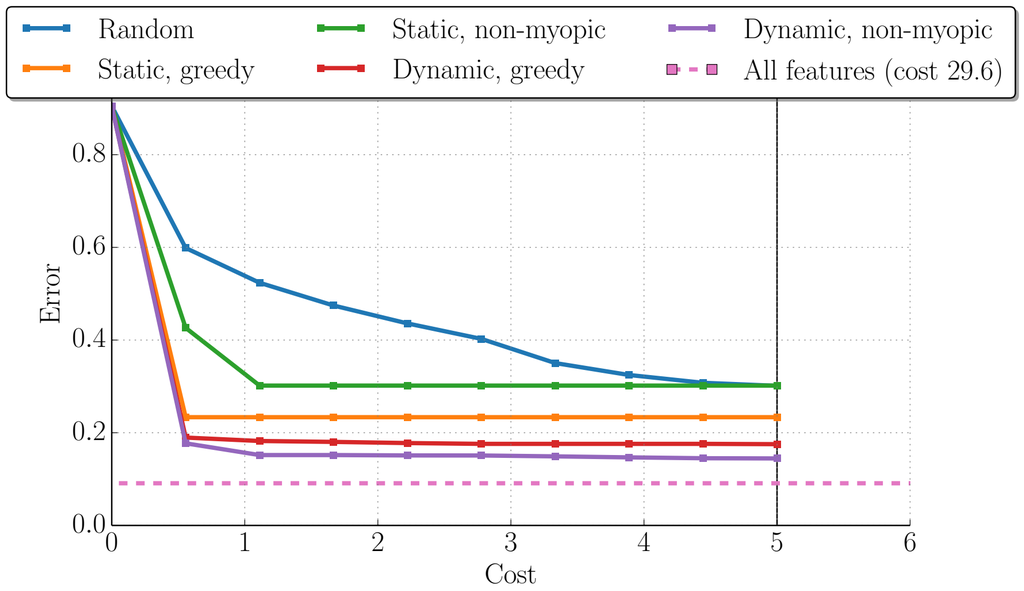


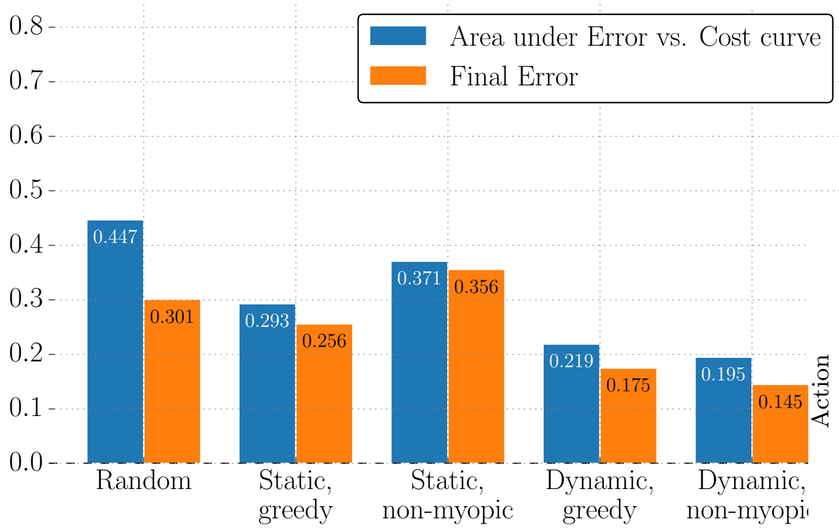


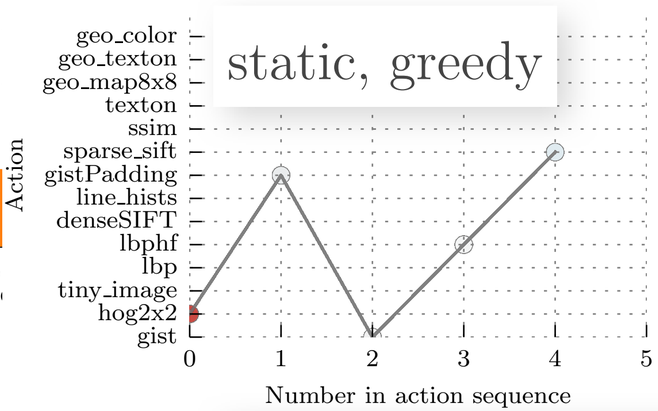


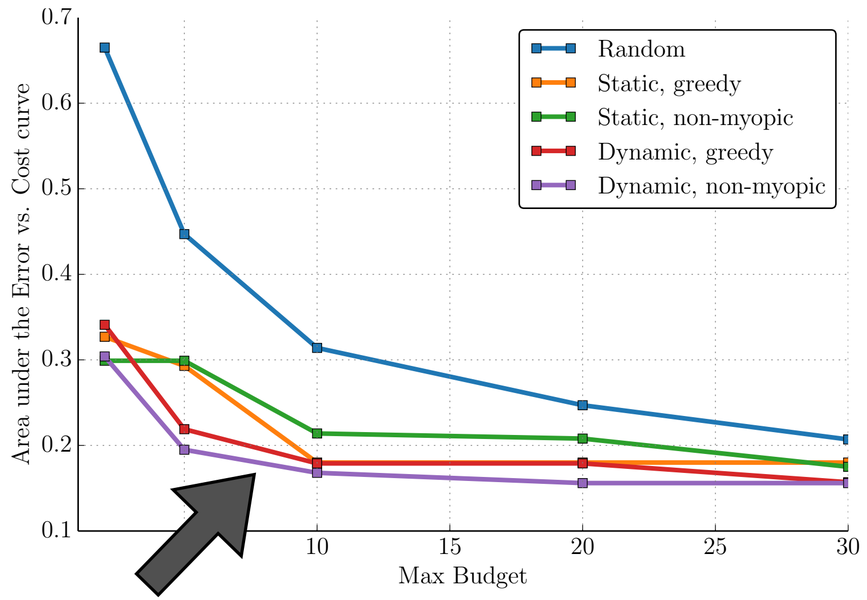
Scenes-15

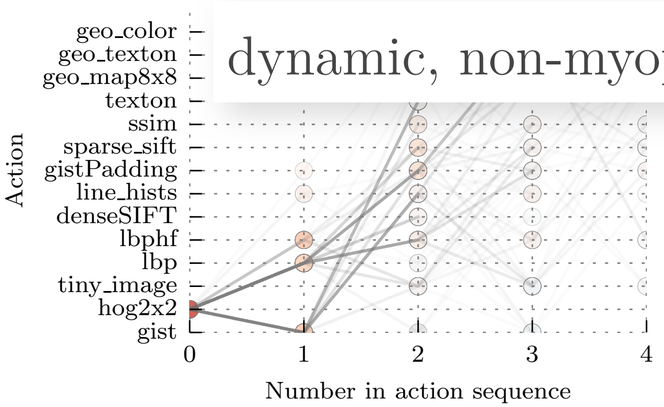

bedroom
suburban
industrial
kitchen
living room
coast
forest
highway
inside city
mountain
open country
street
tall building
office
store


Random
Static, greedy
Static, non-myopic
Dynamic, greedy
Dynamic, non-myopic
All features (cost 29.6)
Error
Cost

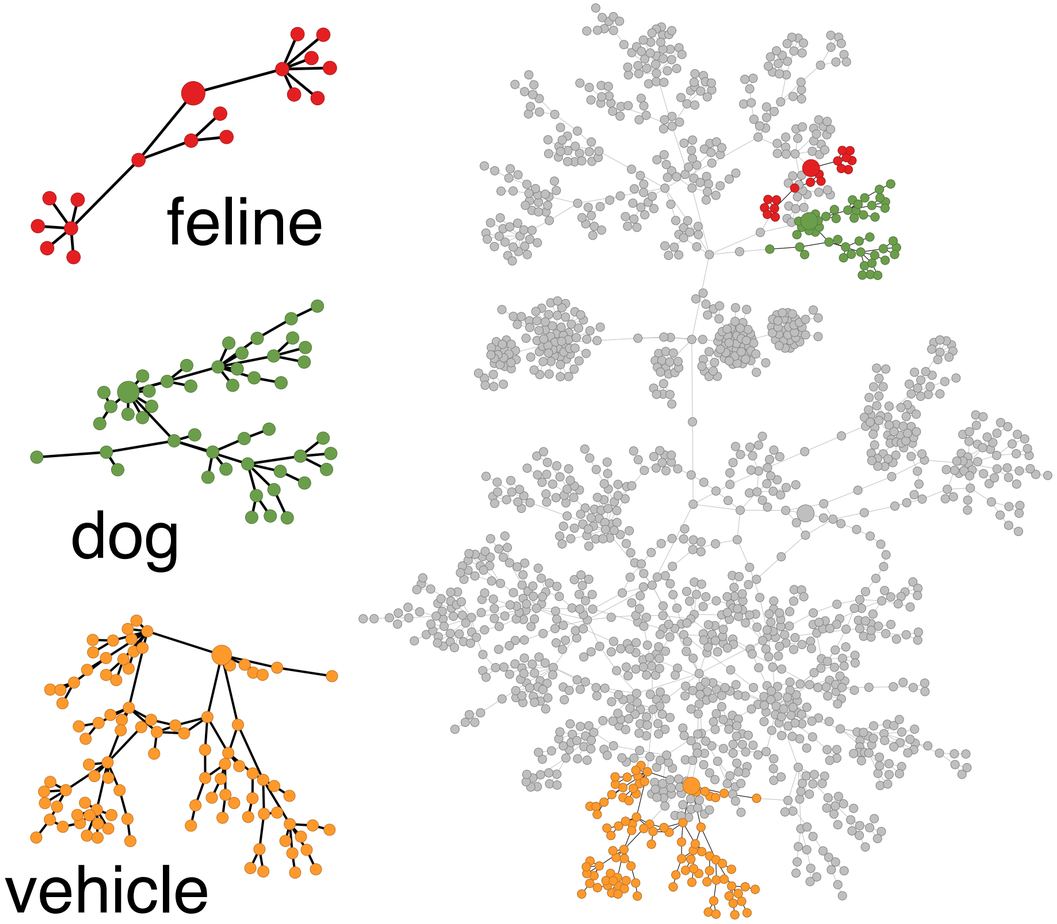

Area under Error vs. Cost curve
Final Error
0.447
0.301
0.293
0.256
0.371
0.356
0.219
0.175
0.195
0.145
Random
Static, greedy
Static, non-myopic
Dynamic, greedy
Dynamic, non-myopi
Action

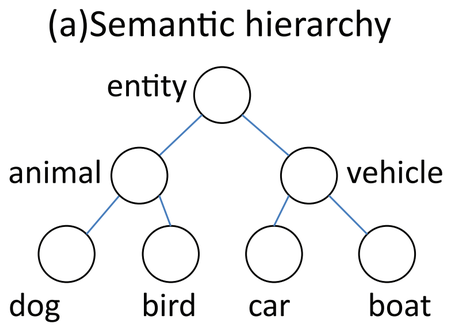

static, greedy
geo_color
geo_texton
geo_map8x8
texton
ssim
sparse_sift
gistPadding
line_hists
denseSIFT
lbphf
lbp
tiny_image
hog2x2
gist
Number in action sequence

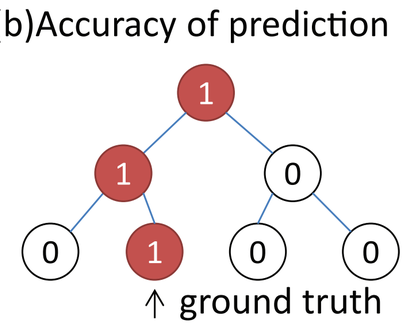

Random
Static, greedy
Static, non-myopic
Dynamic, greedy
Dynamic, non-myopic
Area under the Error vs. Cost curve
Max Budget

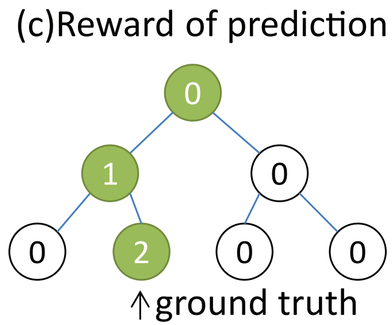

dynamic, non-myopic
geo_color
geo_texton
geo_map8x8
texton
ssim
sparse_sift
gistPadding
line_hists
denseSIFT
lbphf
lbp
tiny_image
hog2x2
gist
Action
Number in action sequence

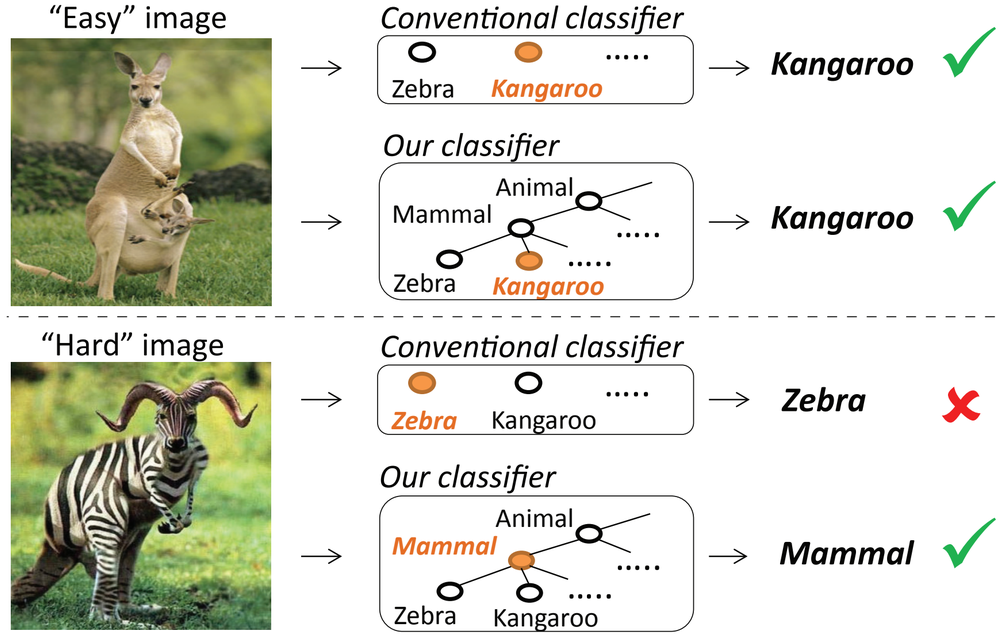
IMAGENET





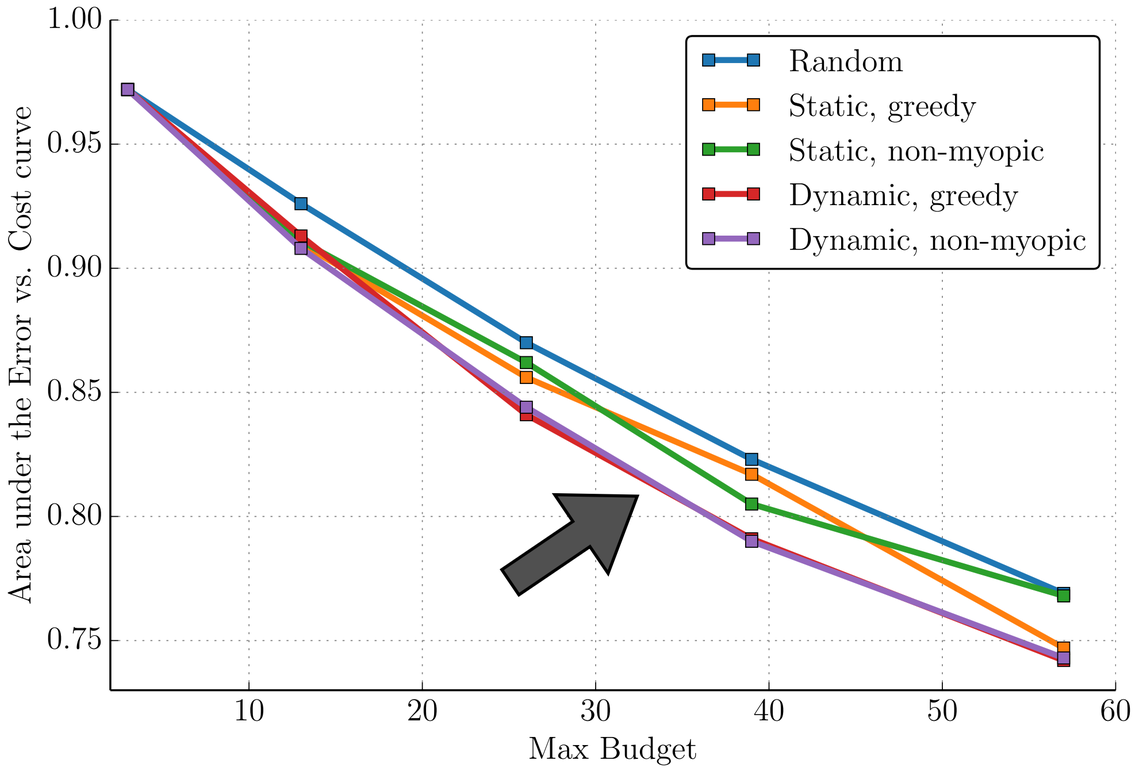


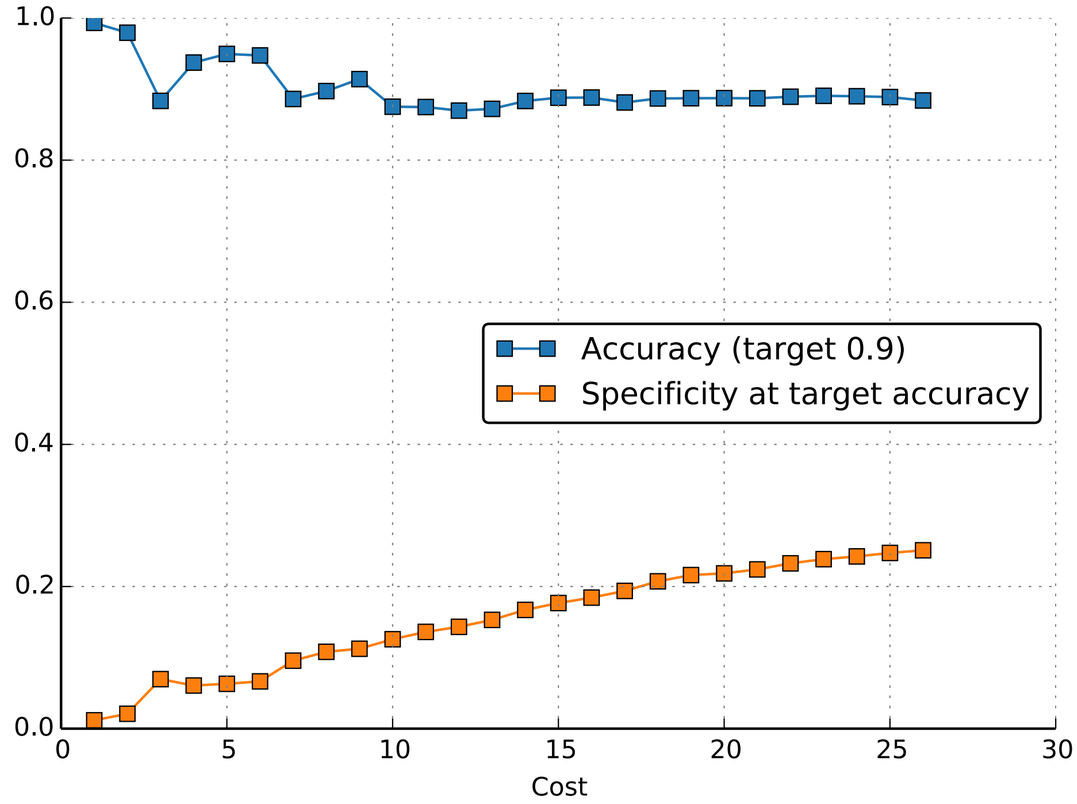


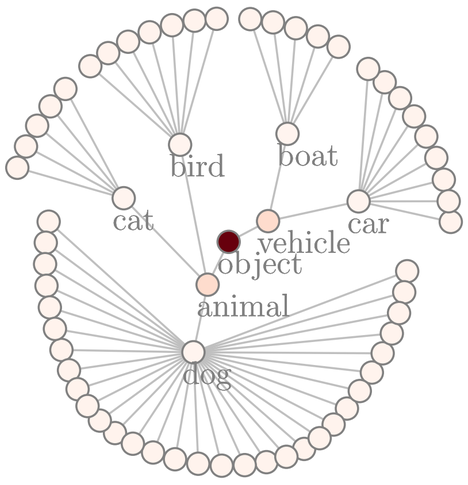


[1] J. Deng, J. Krause, A. C. Berg, and L. Fei-fei, "Hedging Your Bets: Optimizing Accuracy-Specificity Trade-offs in Large Scale Visual Recognition," in CVPR, 2012.

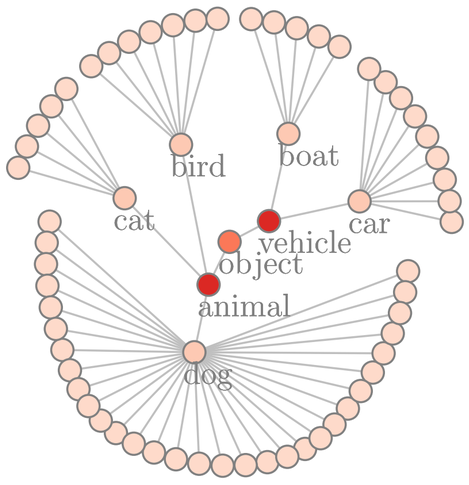
ILSVRC-65



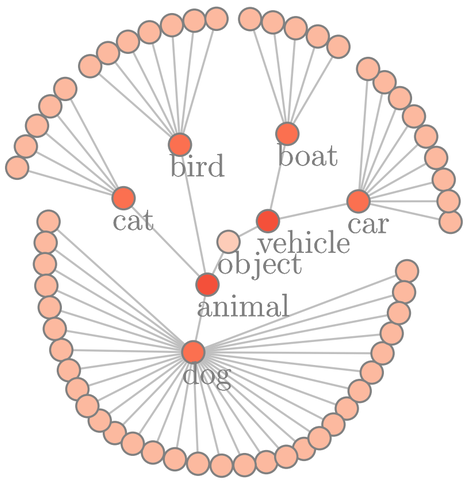




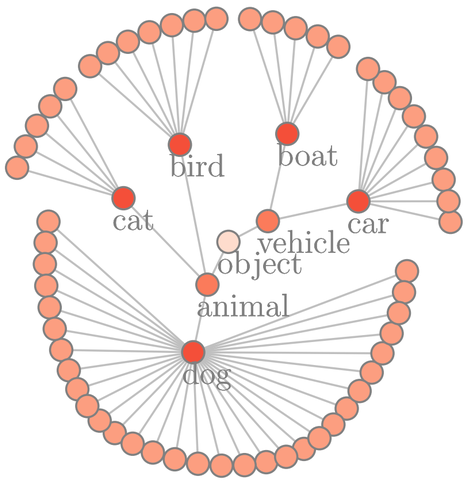
Cost: 1



Cost: 9



Cost: 17



Cost: 26