Teaching computers about visual categories

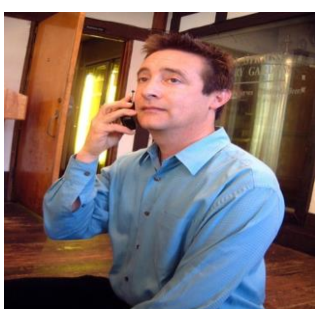
Kristen Grauman
Department of Computer Science
University of Texas at Austin

Visual category recognition

Goal: recognize and detect categories of visually and semantically related…


Objects

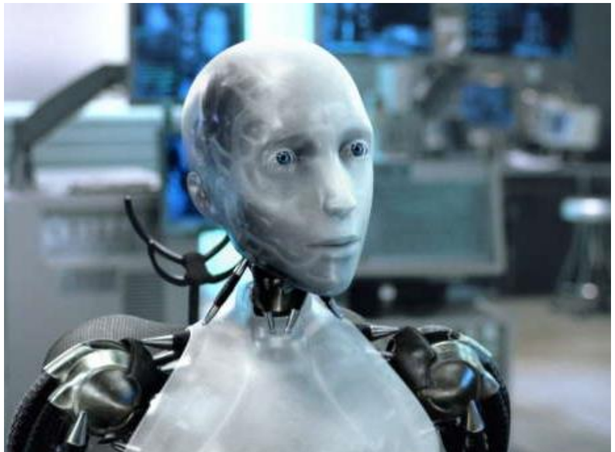

Scenes



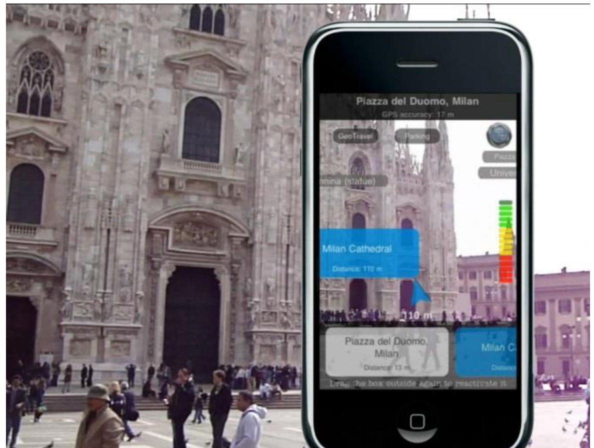

Activities

Kristen Grauman, UT Austin

The need for visual recognition



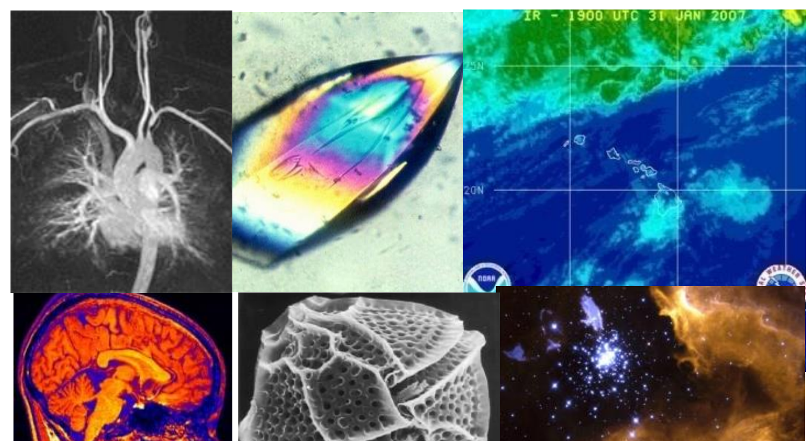
Robotics



Augmented reality

Indexing by content

Surveillance



Scientific data analysis

Difficulty of category recognition

Illumination

Object pose

Clutter

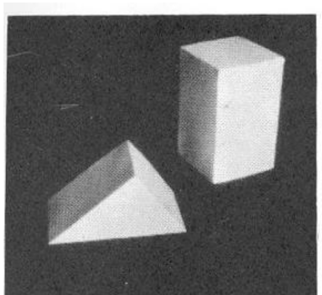
Occlusions

Intra-class appearance

Viewpoint

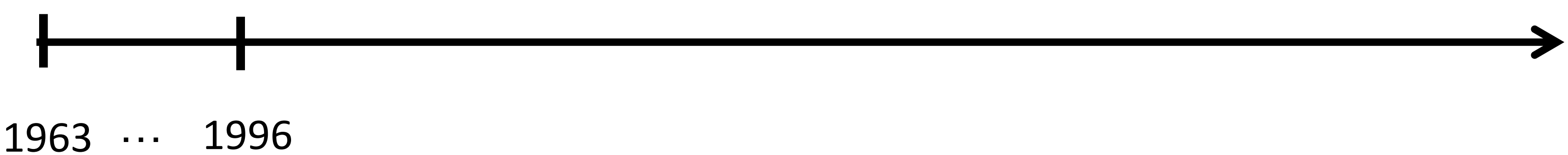
~30,000 possible categories to distinguish! [Biederman 1987]

Progress charted by datasets

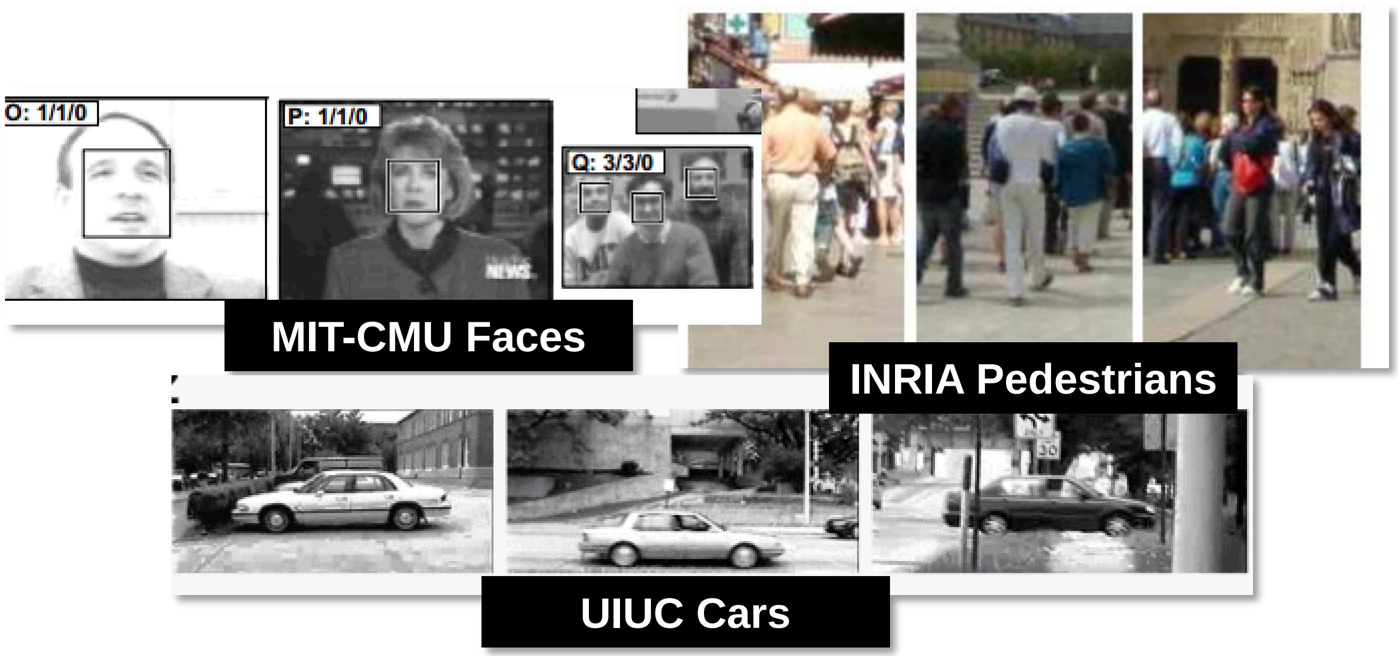


Roberts 1963

COIL

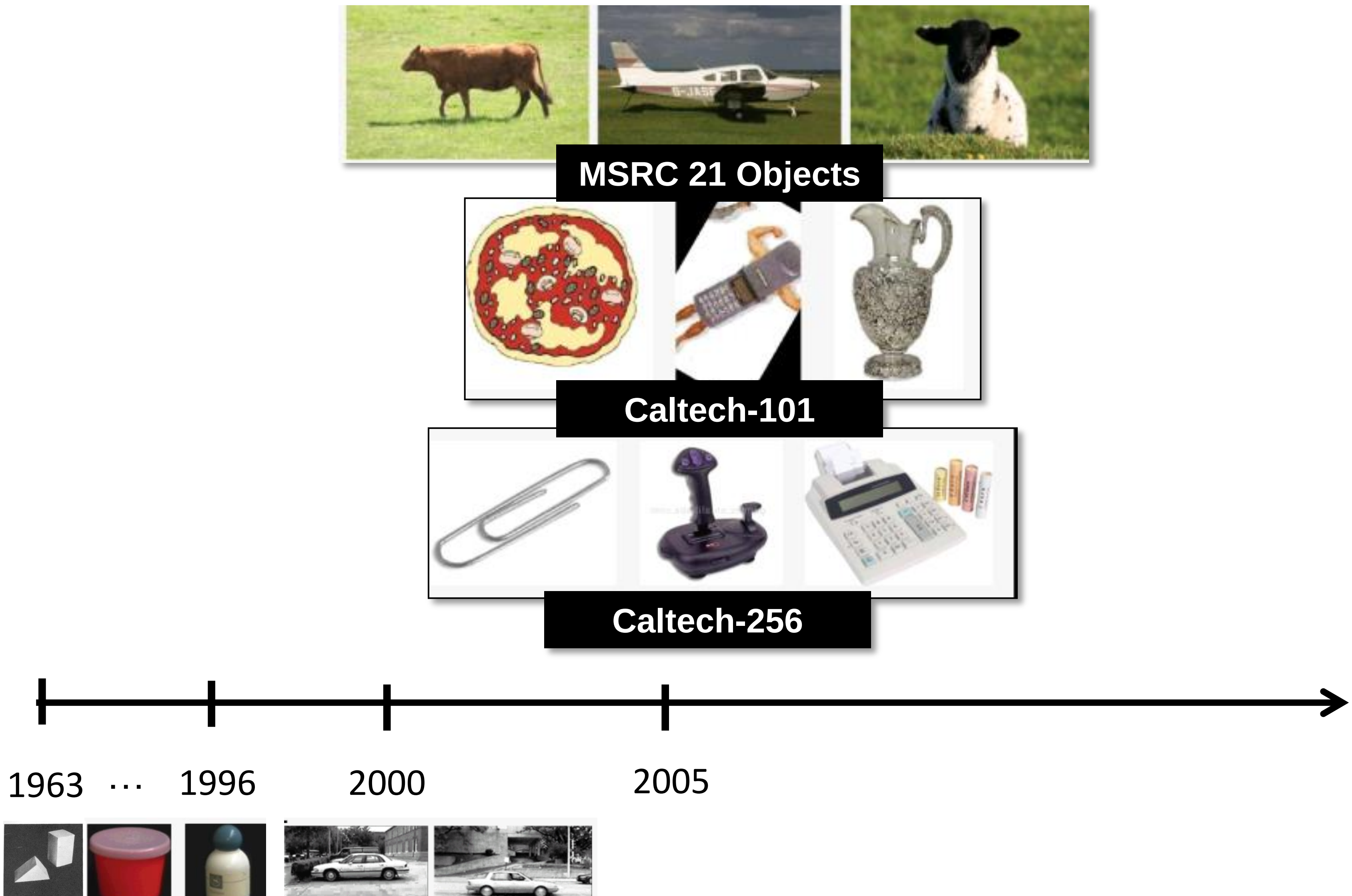


Progress charted by datasets

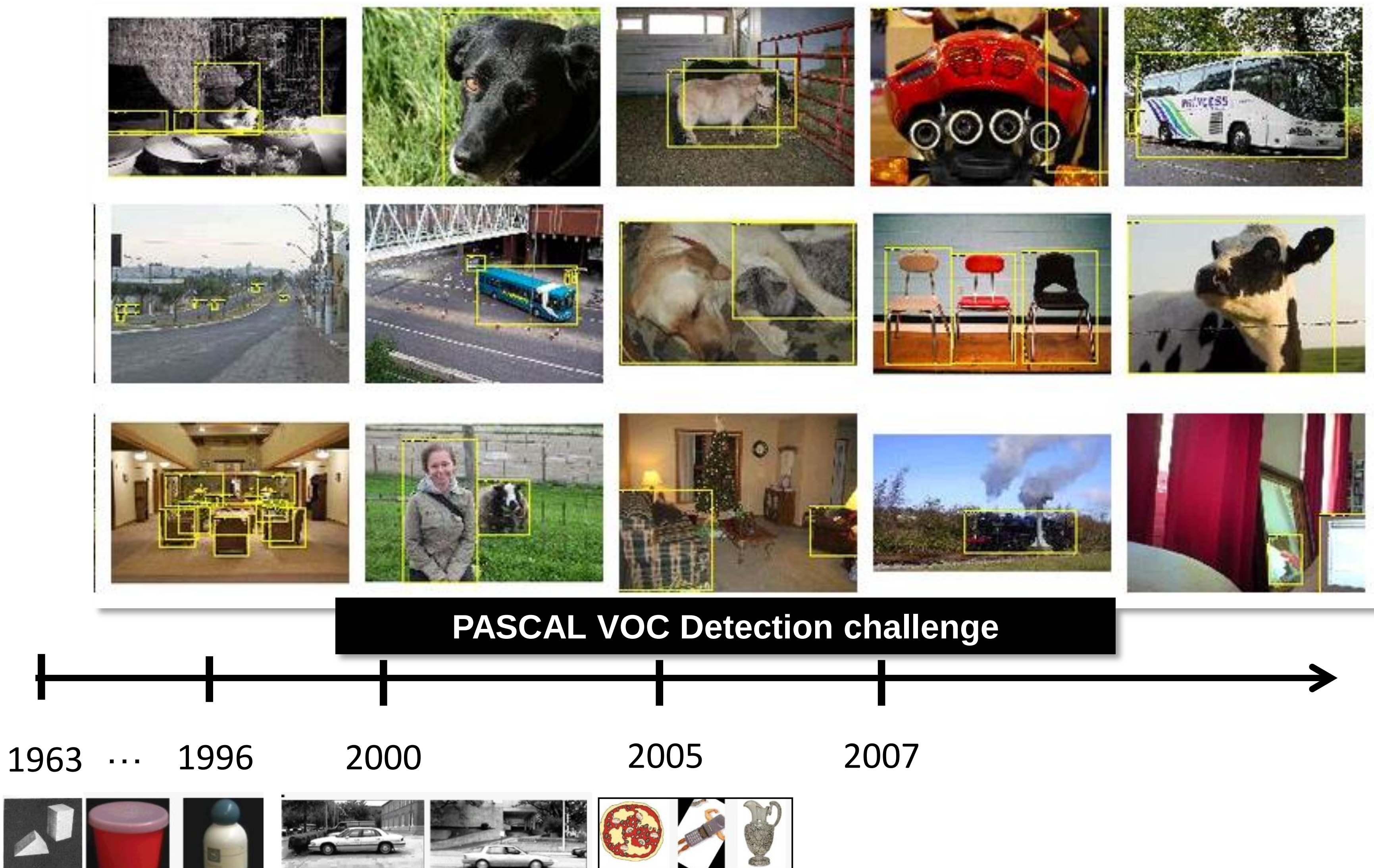


1963 ··· 1996 2000

Progress charted by datasets

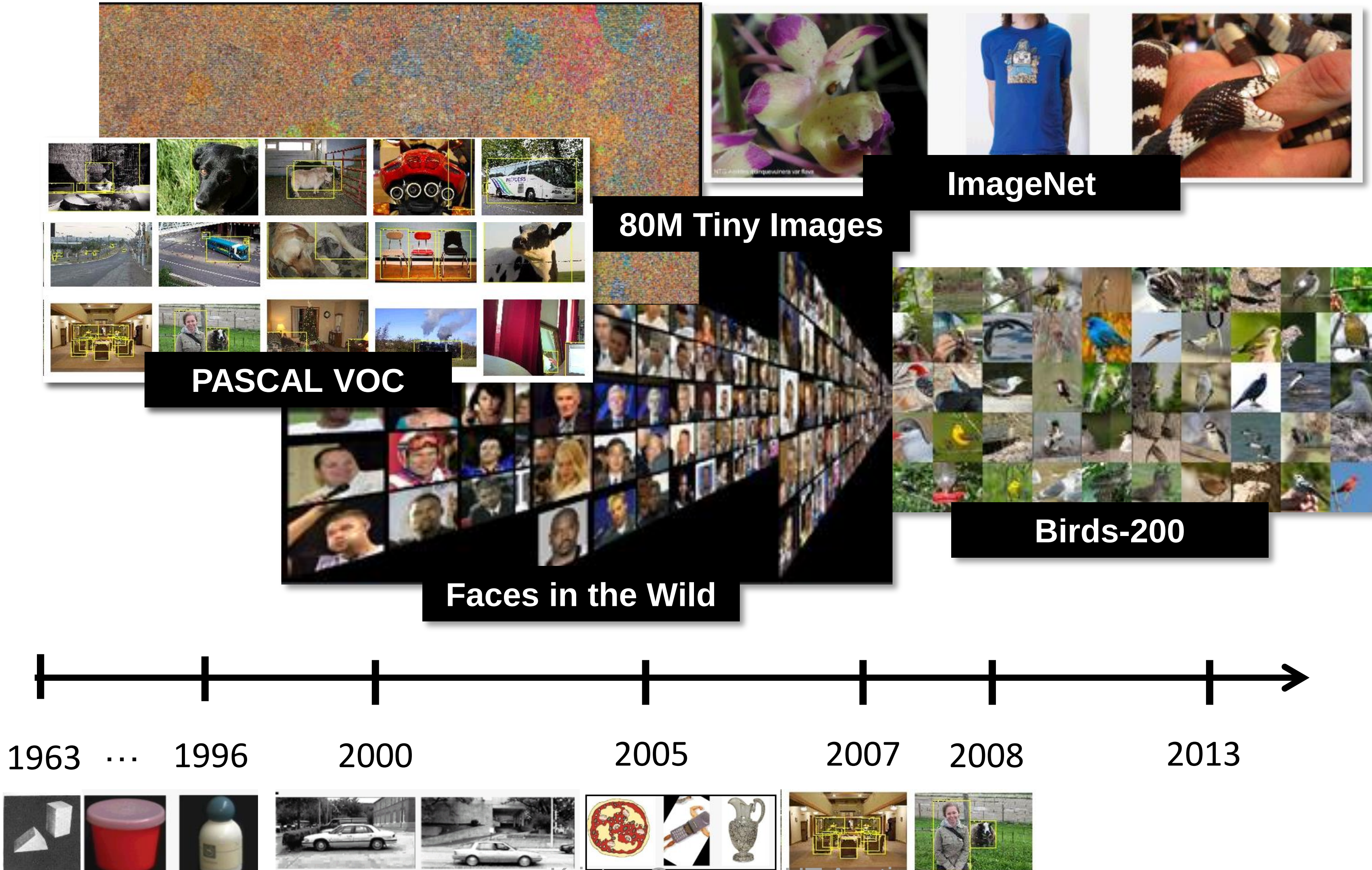


Progress charted by datasets



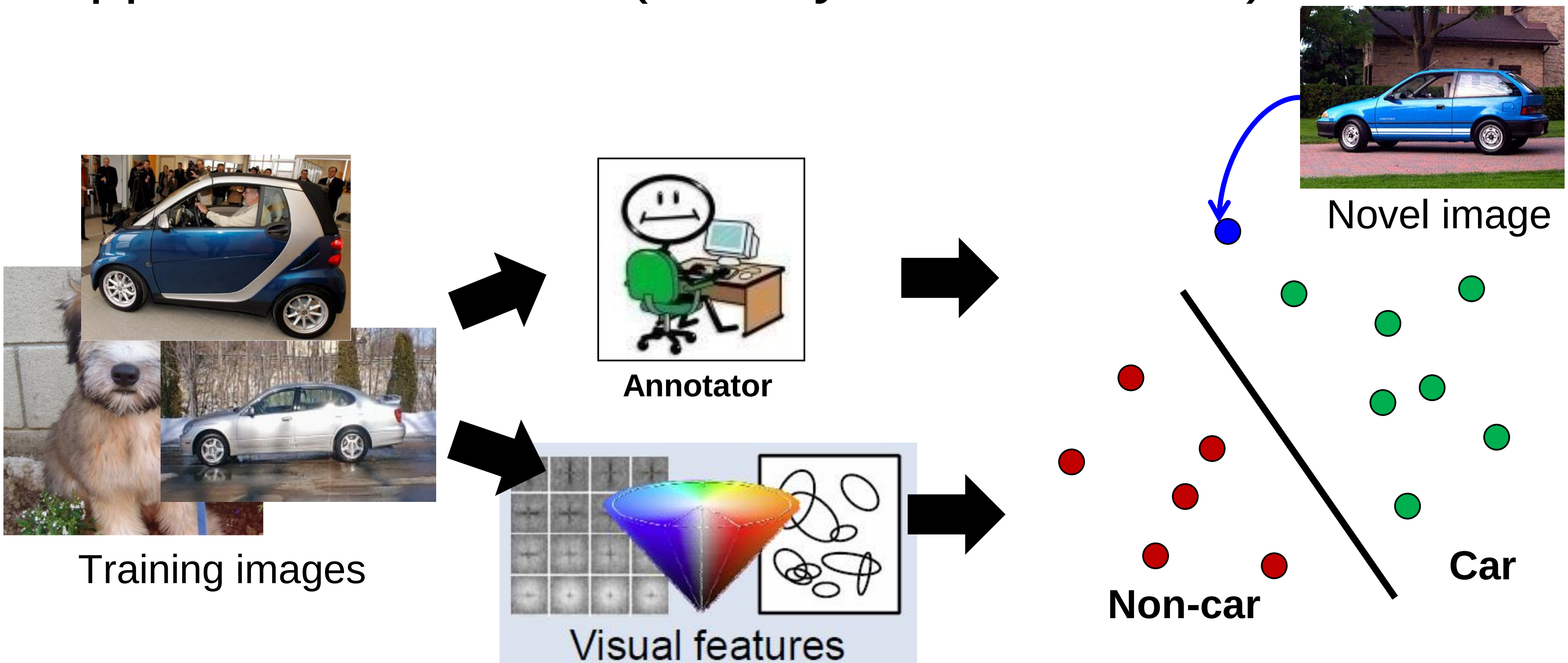
Kristen Grauman, UT Austin

Progress charted by datasets



Learning-based methods

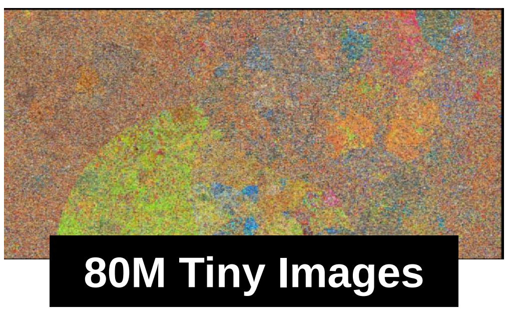
Last ~10 years: impressive strides by *learning* appearance models (usually discriminative).



[Papageorgiou & Poggio 1998, Schneiderman & Kanade 2000, Viola & Jones 2001, Dalal & Triggs 2005, Grauman & Darrell 2005, Lazebnik et al. 2006, Felzenszwalb et al. 2008,…]

Exuberance for image data (and their category labels)

14M images
1K+ labeled object categories
[Deng et al. 2009-2012]

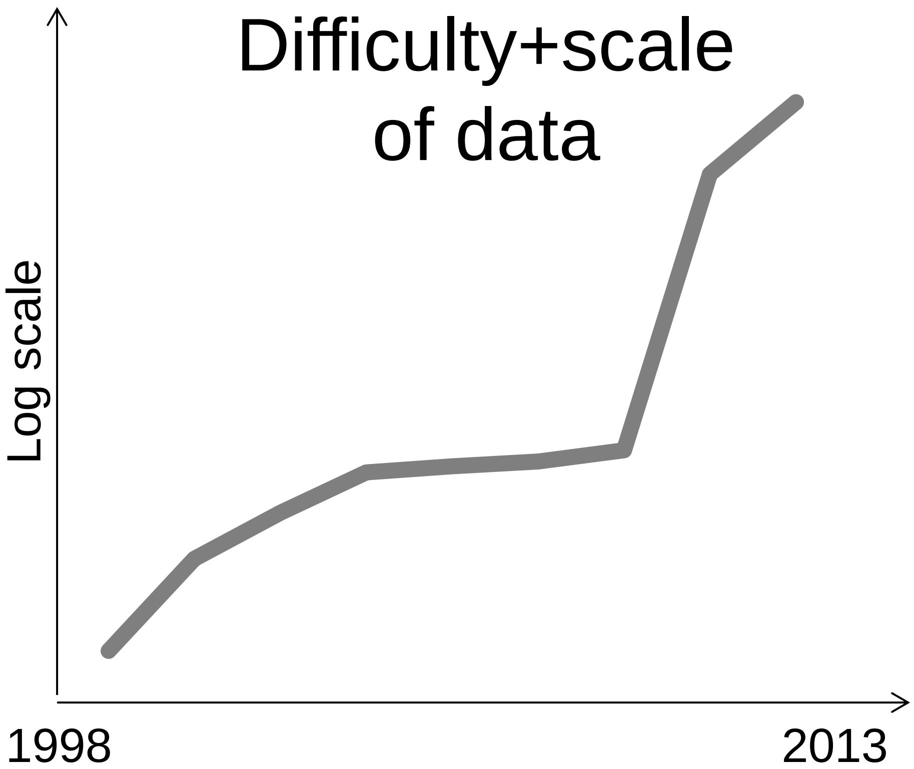


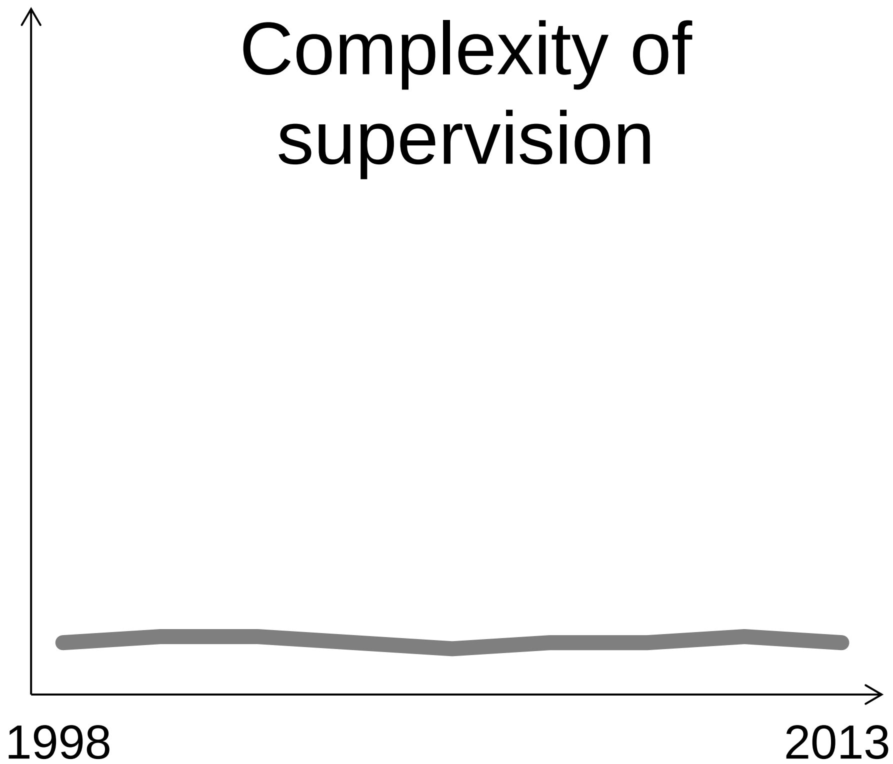
80M images
53K noisily labeled object categories
[Torralba et al. 2008]

131K images
902 labeled scene categories
4K labeled object categories
[Xiao et al. 2010]

Kristen Grauman, UT Austin

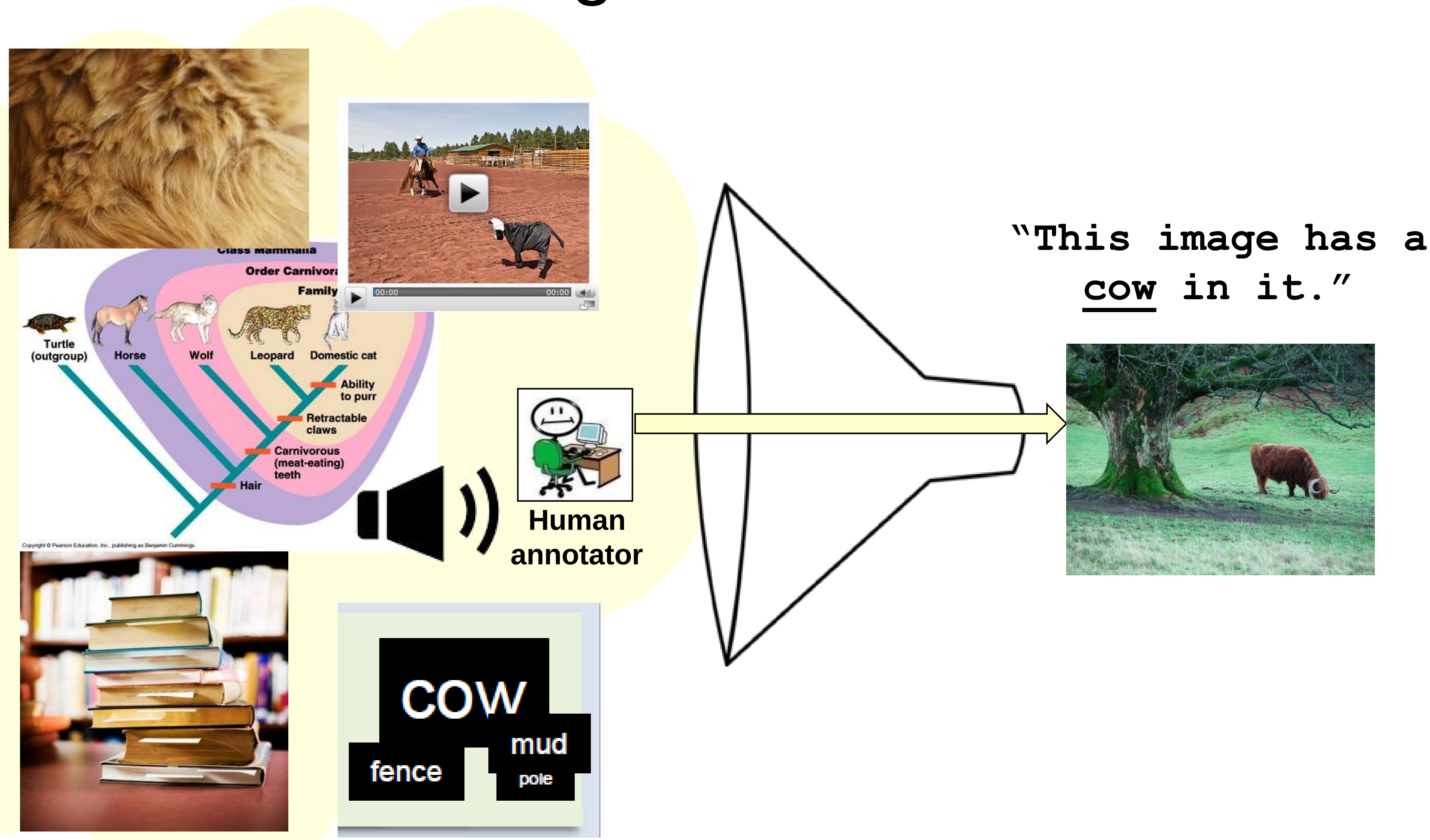
Problem





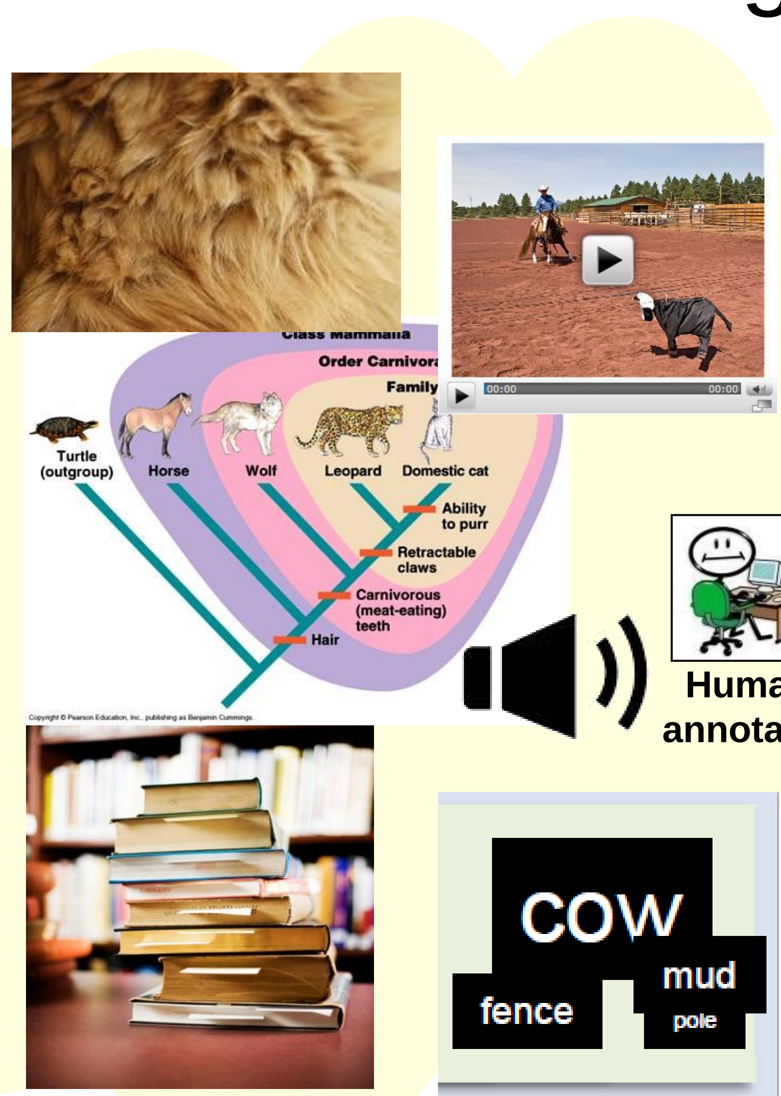
While complexity and scale of recognition task has escalated dramatically, our means of "teaching" visual categories remains shallow.

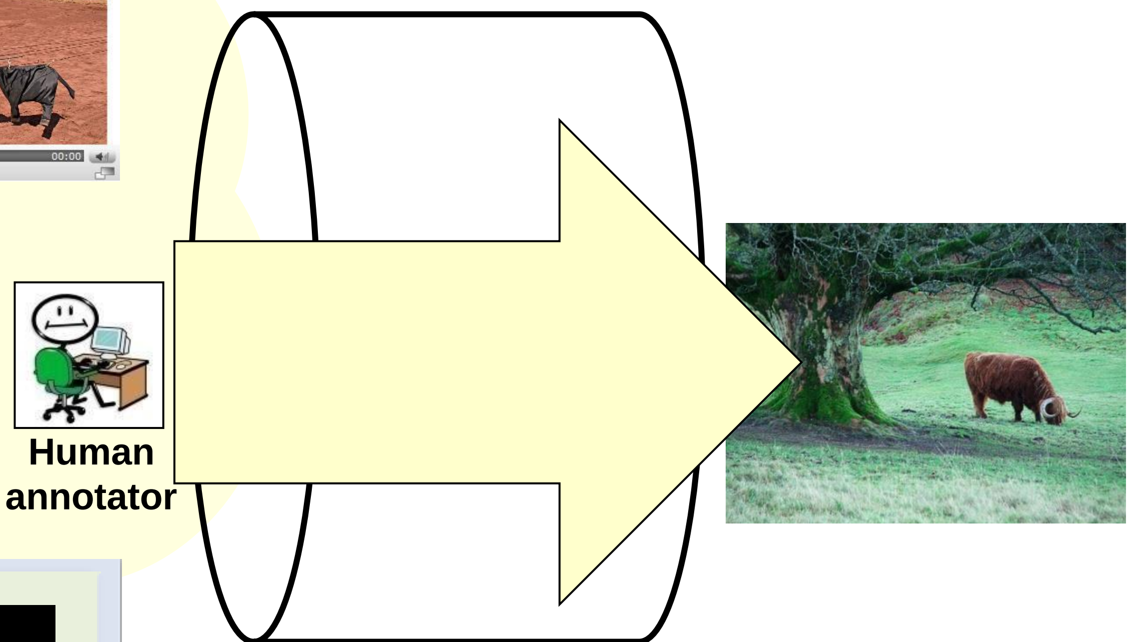
Envisioning a broader channel



More labeled images ↔ more accurate models?

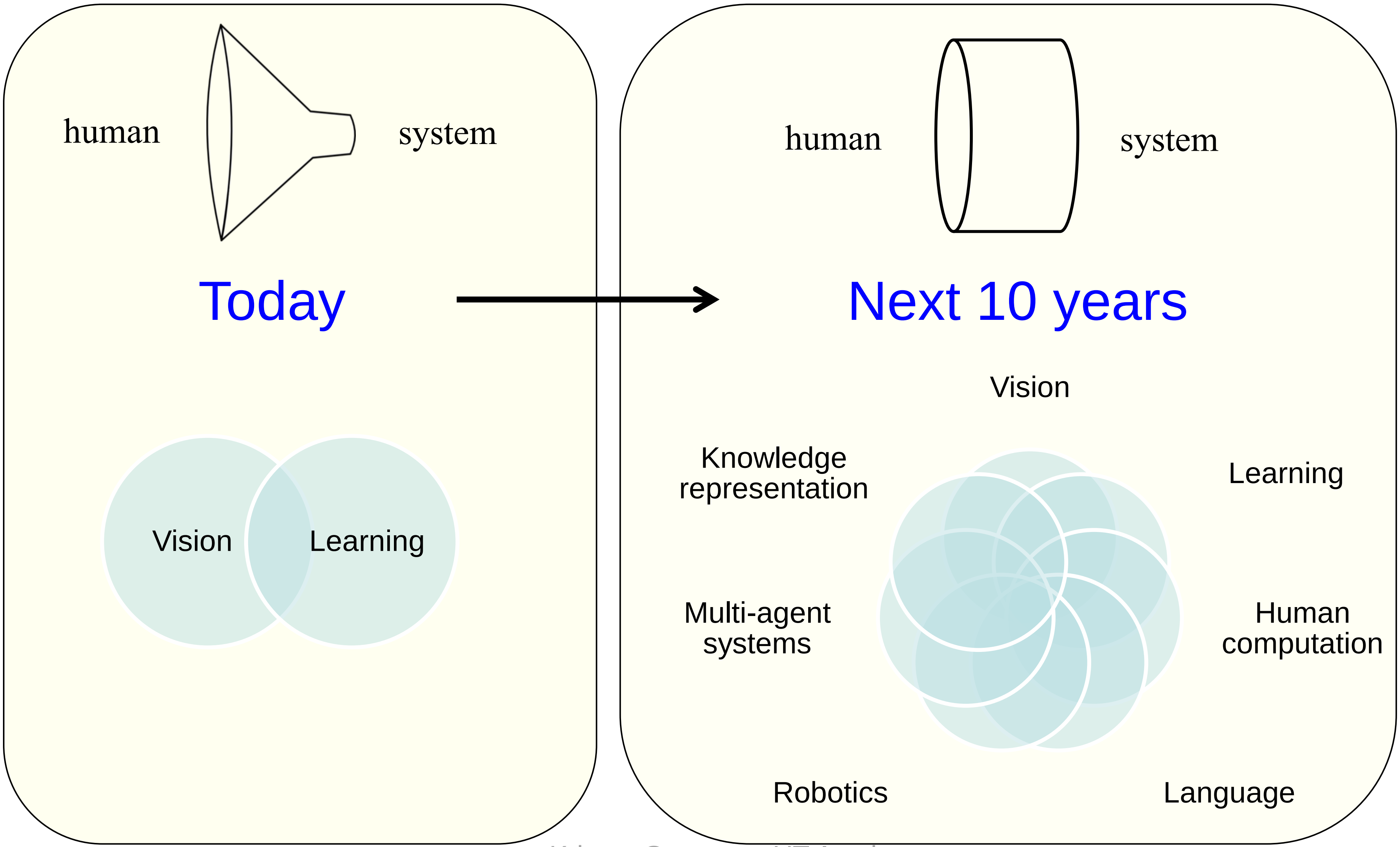
Envisioning a broader channel





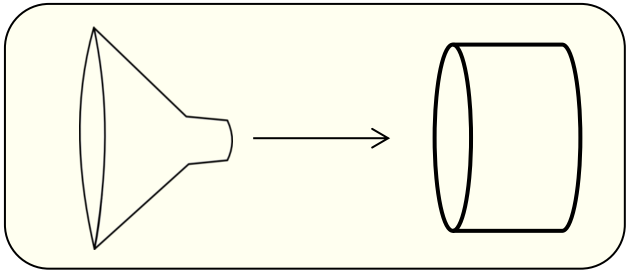
Need richer means to teach system about visual world

Envisioning a broader channel



Our goal

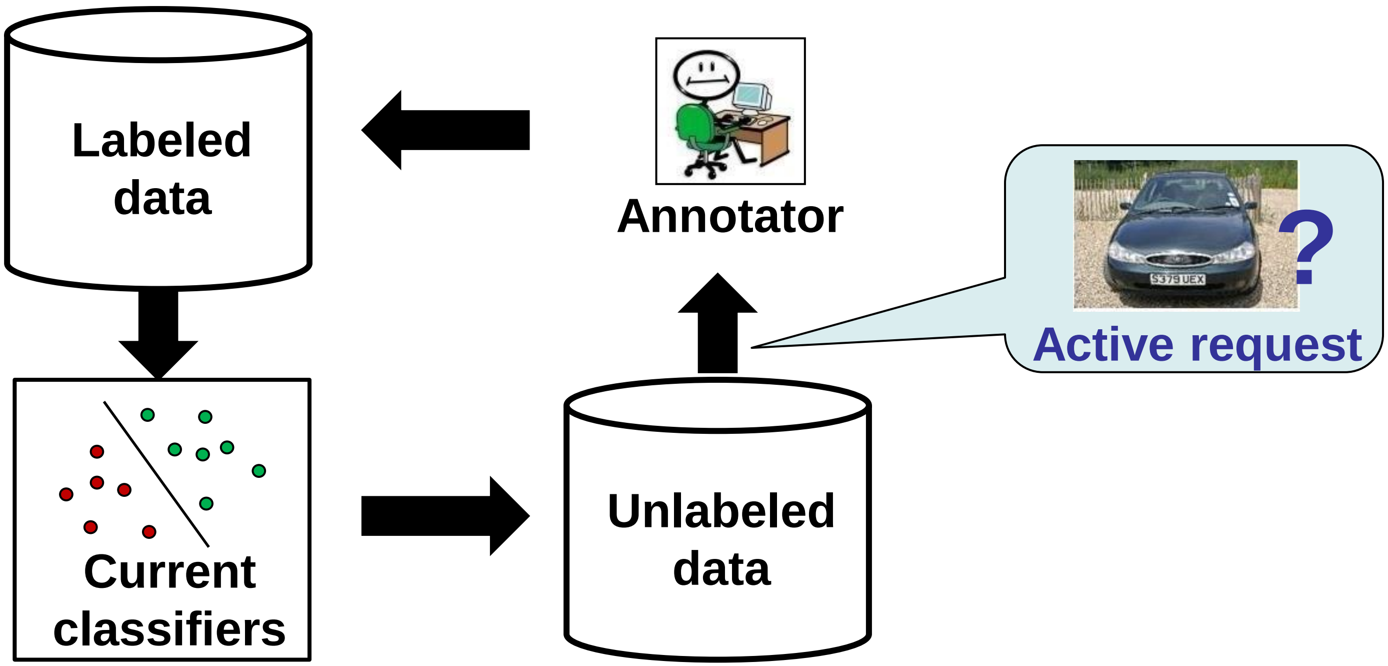
Teaching computers about visual categories must be an ongoing, interactive process, with communication that goes beyond labels.



This talk:

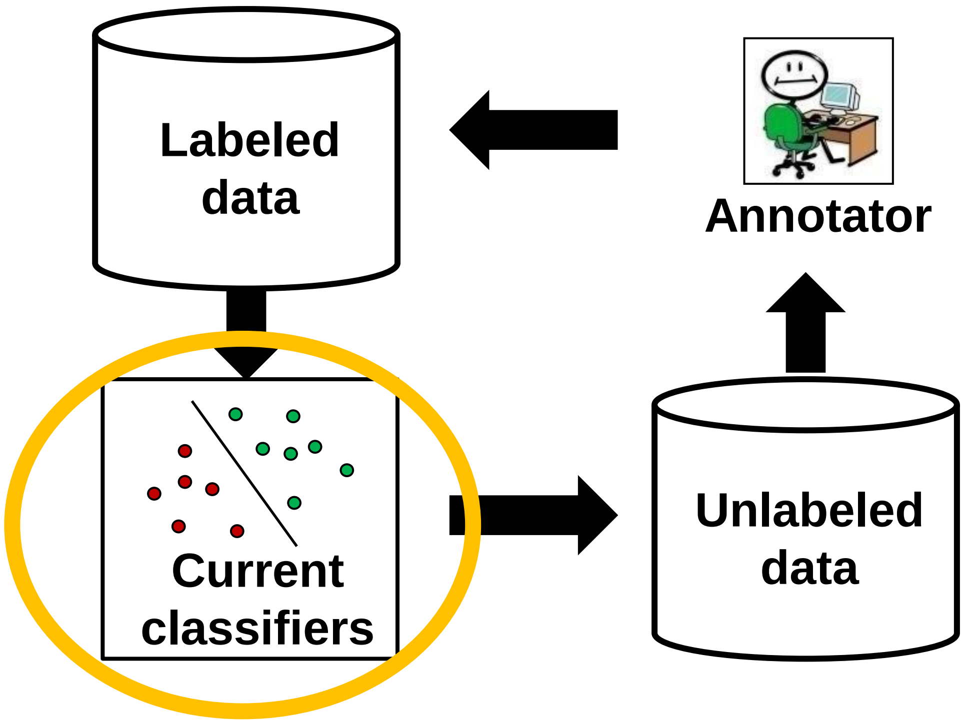
1. Active visual learning
2. Learning from visual comparisons

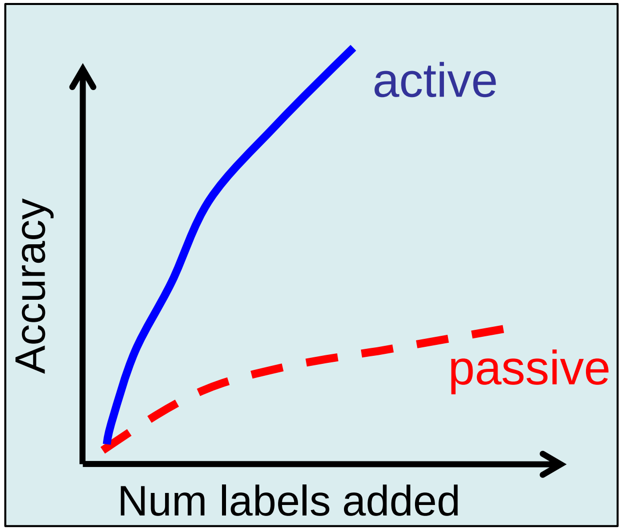
Active learning for visual recognition



[Mackay 1992, Cohn et al. 1996, Freund et al. 1997, Lindenbaum et al. 1999, Tong & Koller 2000, Schohn and Cohn 2000, Campbell et al. 2000, Roy & McCallum 2001, Kapoor et al. 2007,…]

Active learning for visual recognition





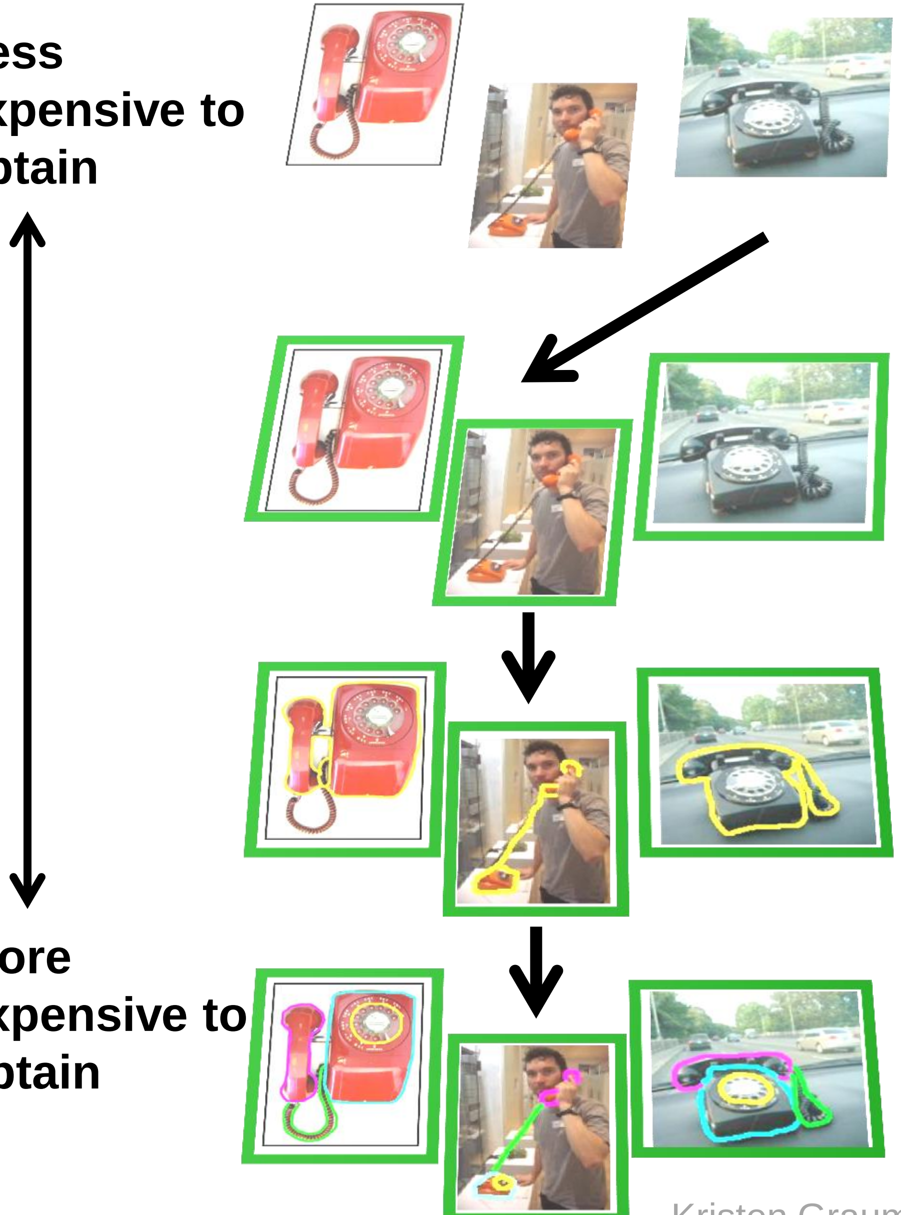
Intent: better models, faster/cheaper

Kristen Grauman, UT Austin

Problem: Active selection and recognition

Less expensive to obtain

More expensive to obtain



- **Multiple levels** of annotation are possible
- **Variable cost** depending on level *and* example

Our idea: Cost-sensitive multi-question active learning

- Compute decision-theoretic active selection criterion that weighs both:
  - which *example* to annotate, and
  - what *kind* of annotation to request for it

  as compared to
  - the *predicted effort* the request would require

[Vijayanarasimhan & Grauman, NIPS 2008, CVPR 2009]

Decision-theoretic multi-question criterion

$$\underbrace{\mathrm{VALUE}(O,Q)}_{\text{Value of asking given question about given data object}} = \underbrace{\mathrm{RISK}(\mathcal{X}_L, \mathcal{X}_U)}_{\text{Current misclassification risk}} - \underbrace{\widehat{\mathrm{RISK}}(\mathcal{X}_L \cup O_A, \mathcal{X}_U \setminus O)}_{\text{Estimated risk if candidate request were answered}} - \underbrace{\mathrm{COST}(O,Q)}_{\text{Cost of getting the answer}}$$

Three “levels” of requests to choose from:

1. Label a region

2. Tag an object in the image

3. Segment the image, name all objects.

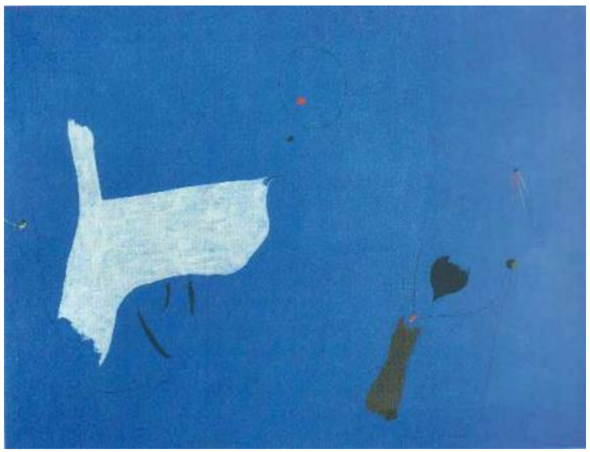
Predicting effort

- What manual effort cost would we expect to pay for an unlabeled image?

Which image would you rather annotate?

Predicting effort

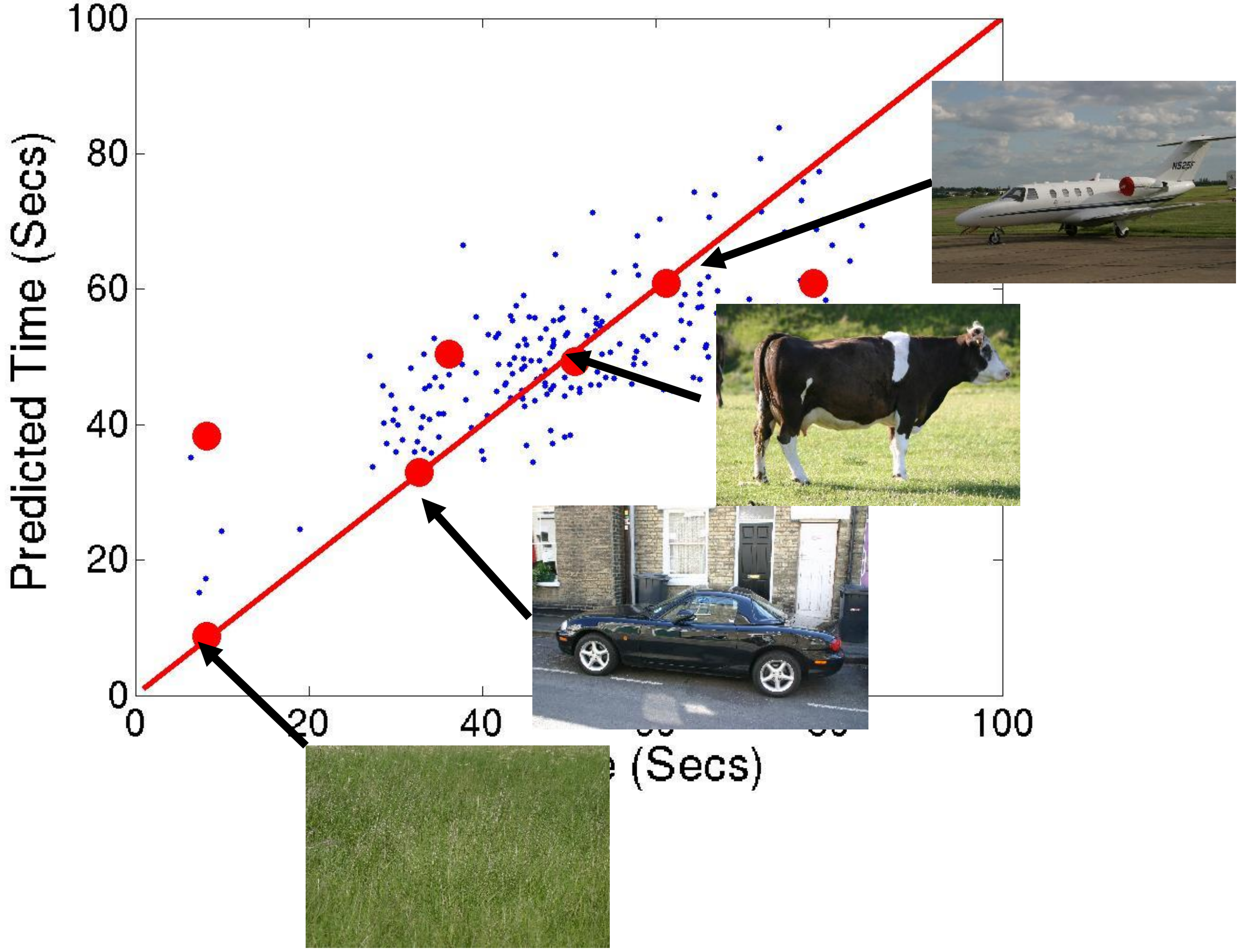
- What manual effort cost would we expect to pay for an unlabeled image?



Which image would you rather annotate?

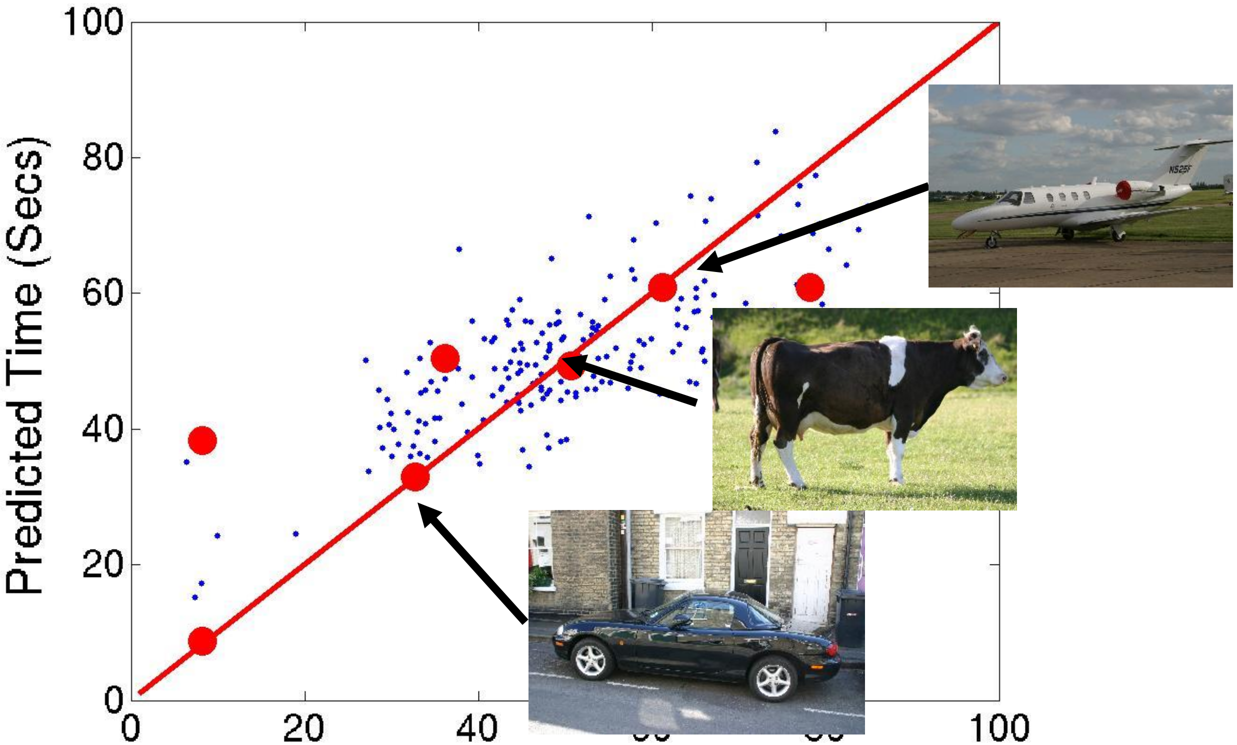
Predicting effort

We estimate labeling difficulty from visual content.



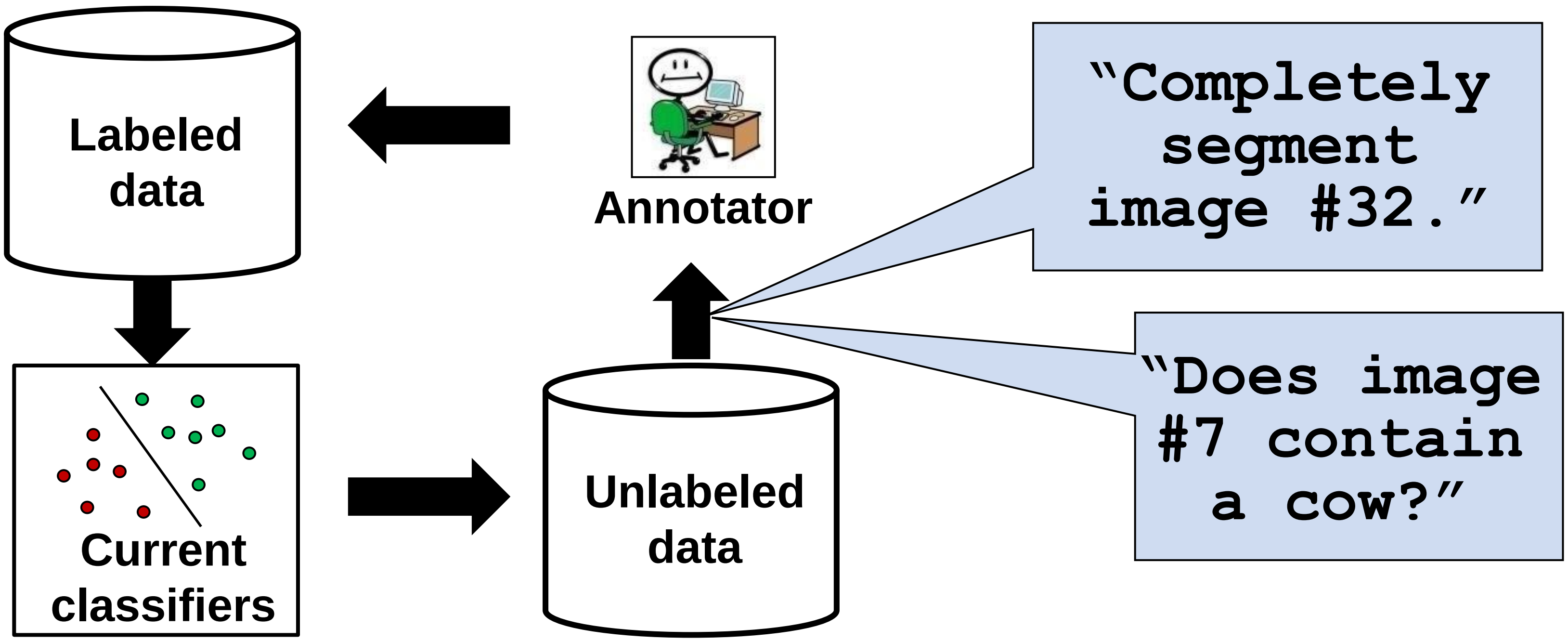
Predicting effort

We estimate labeling difficulty from visual content.



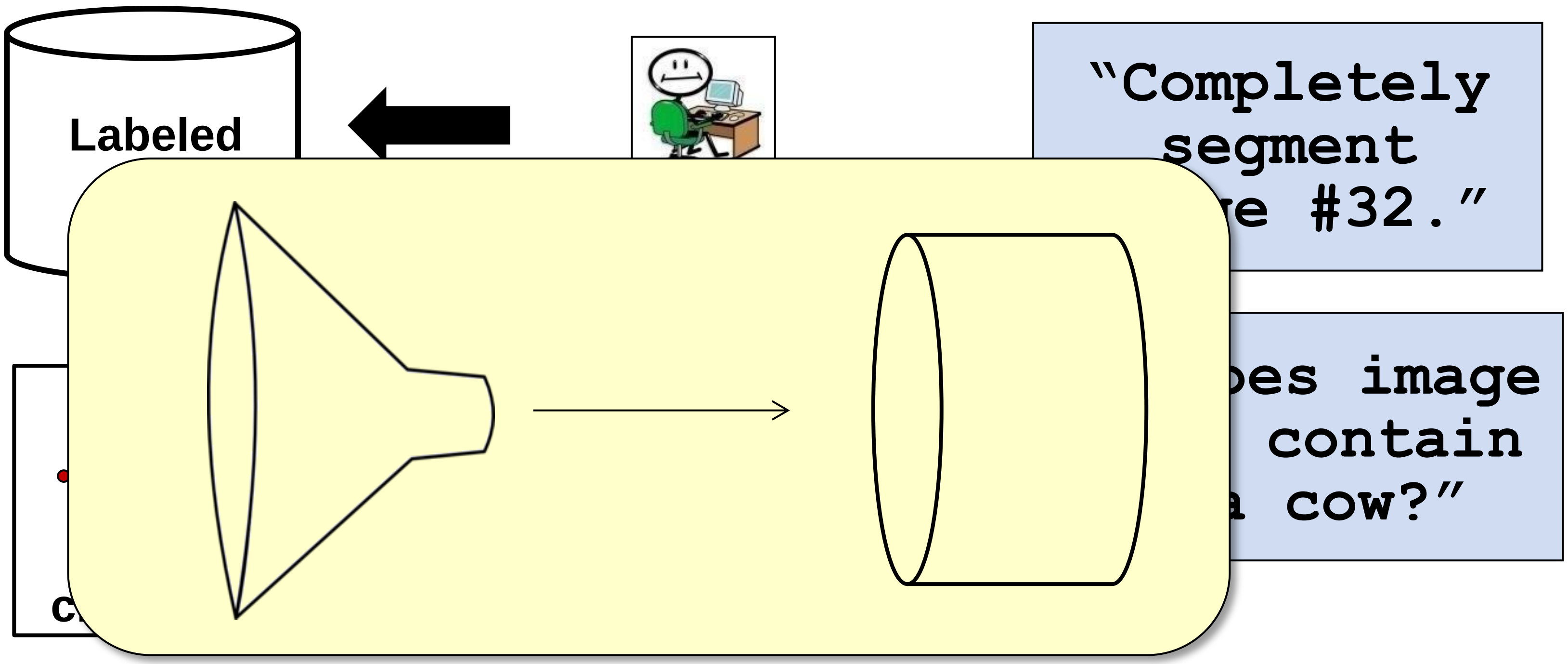
Other forms of effort cost: expertise required, resolution of data, how far the robot must move, length of video clip,…

Multi-question active learning



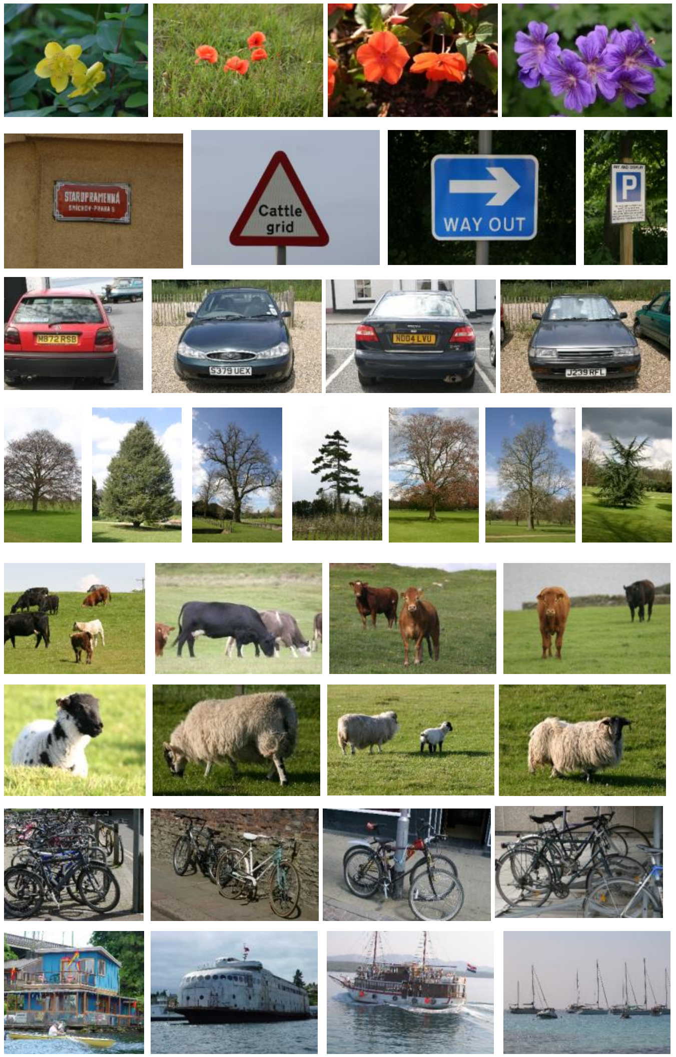
[Vijayanarasimhan & Grauman, NIPS 2008, CVPR 2009]

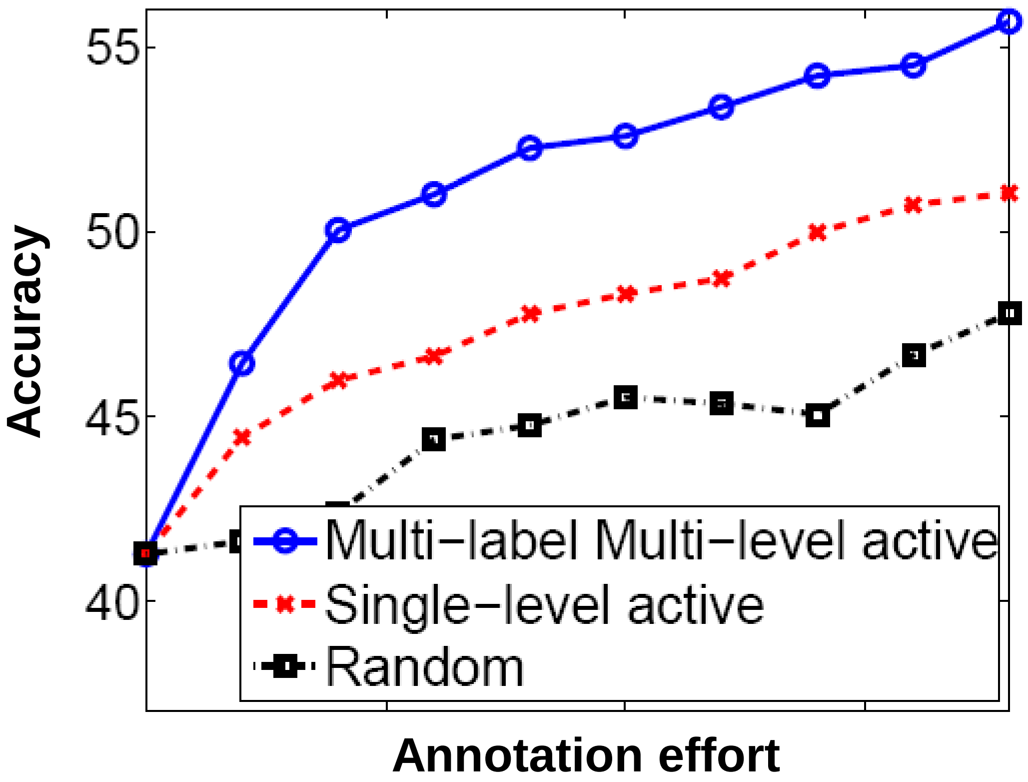
Multi-question active learning



[Vijayanarasimhan & Grauman, NIPS 2008, CVPR 2009]

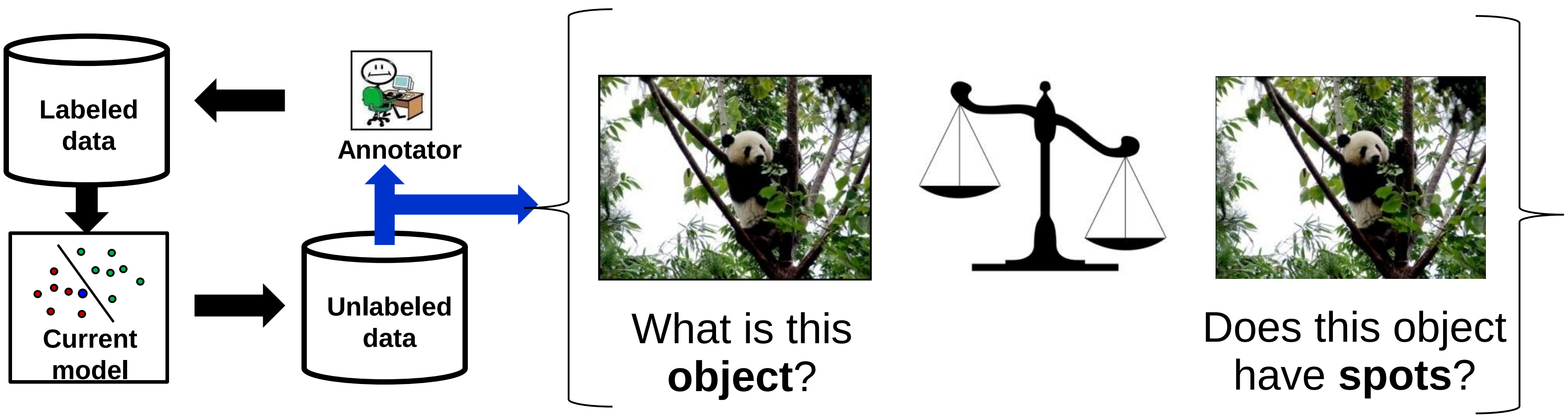
Multi-question active learning curves





Multi-question active learning with objects and attributes

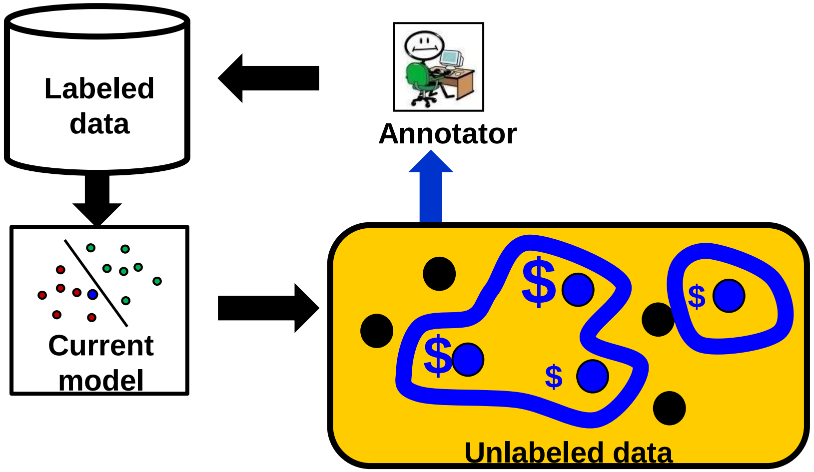
[Kovashka et al., ICCV 2011]



Weigh relative impact of an object label or an attribute label, at each iteration.

Budgeted batch active learning

[Vijayanarasimhan et al., CVPR 2010]



$$S^* = \operatorname{argmax}\ \text{Pred.Gain}(S)$$

$$s.t. \sum_{x \in S} \text{LabelCost}(x) \leq \text{Budget}$$

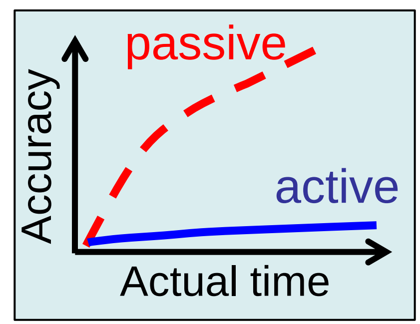
Select *batch* of examples that together improves classifier objective *and* meets annotation *budget*.

Problem: “Sandbox” active learning

Thus far, tested only in artificial settings:

- Unlabeled data already fixed, small scale, biased

- Computational cost ignored

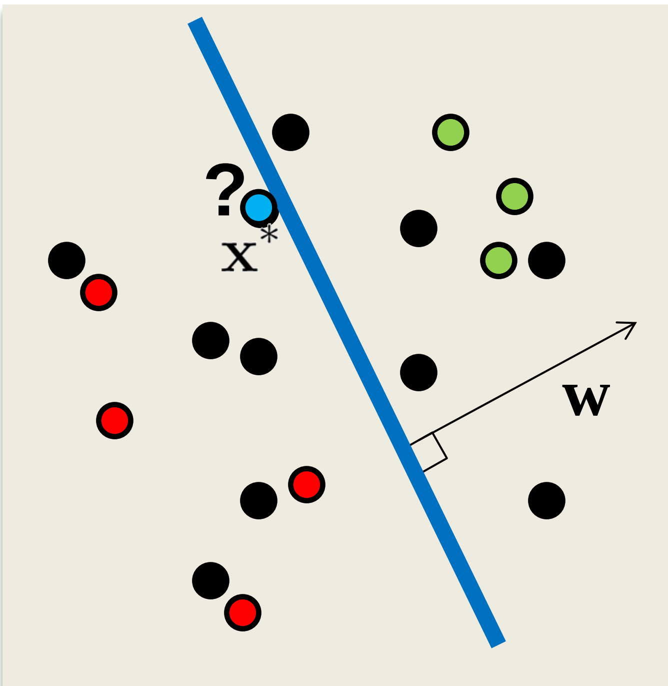


Our idea: Live active learning

Large-scale active learning of object detectors with crawled data and crowdsourced labels.

How to scale active learning to massive unlabeled pools of data?

Pool-based active learning



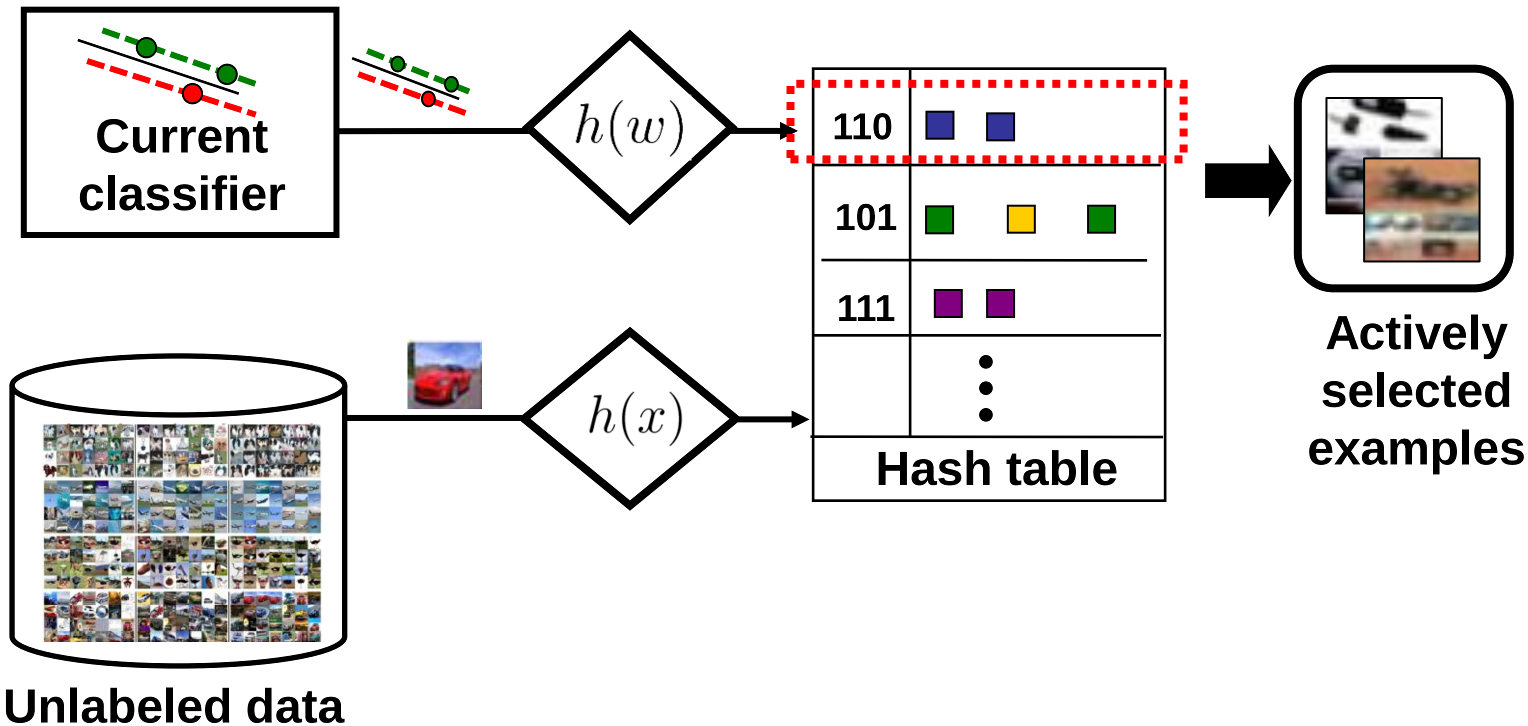
e.g., select point nearest to hyperplane decision boundary for labeling.

$$\mathbf{x}^* = \operatorname{argmin}_{\mathbf{x}_i \in \mathcal{U}} |\mathbf{w}^T \mathbf{x}_i|$$

[Tong & Koller, 2000; Schohn & Cohn, 2000; Campbell et al. 2000]

Sub-linear time active selection

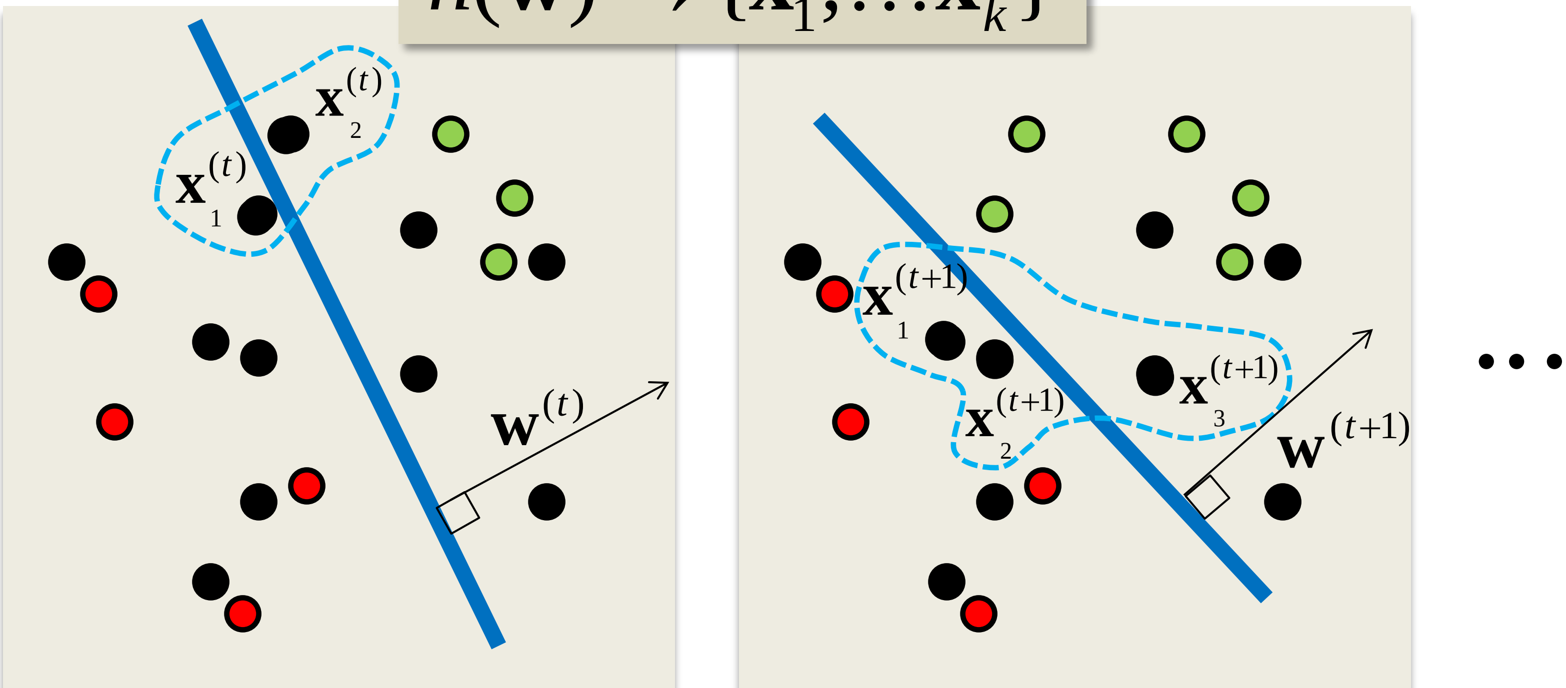
We propose a novel hashing approach to identify the most uncertain examples in sub-linear time.



[Jain, Vijayanarasimhan, Grauman, NIPS 2010]

Hashing a hyperplane query

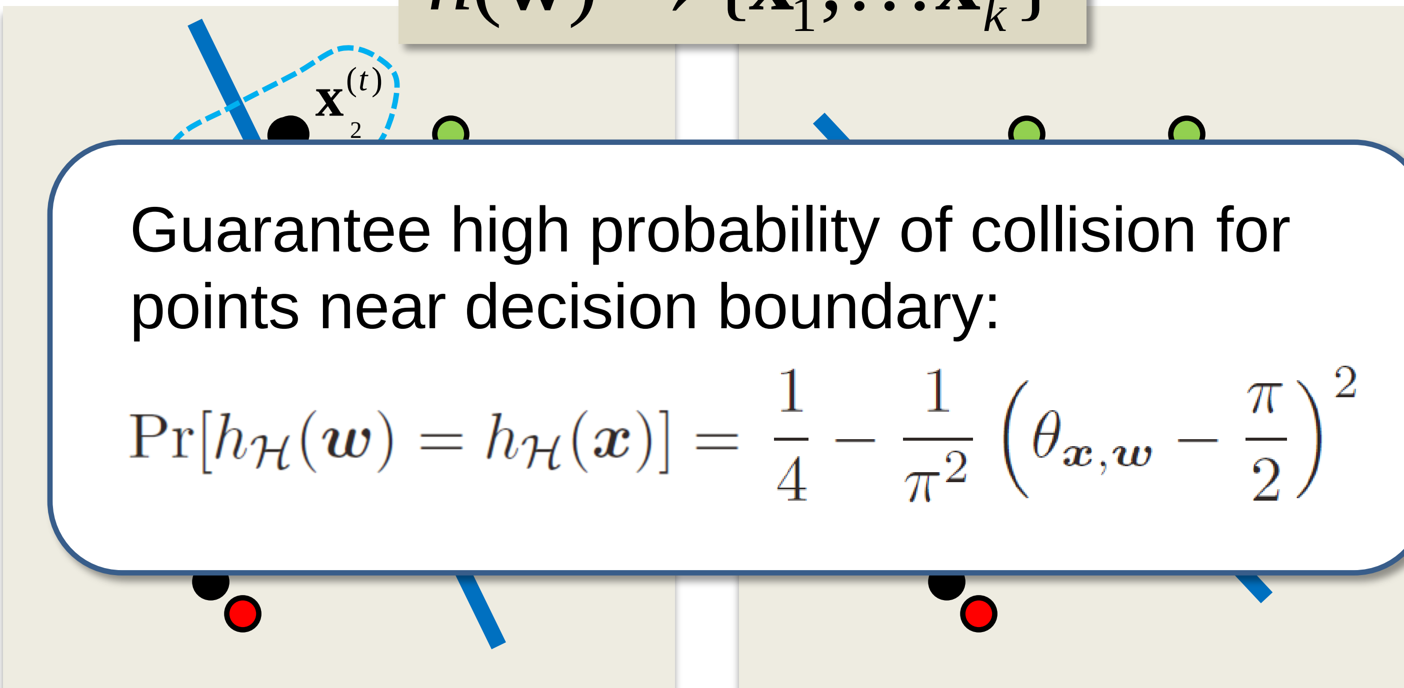
$$h(\mathbf{w}) \rightarrow \{\mathbf{x}_1, \ldots \mathbf{x}_k\}$$



At each iteration of the learning loop, our hash functions map the current hyperplane directly to its nearest unlabeled points.

Hashing a hyperplane query

$$h(\mathbf{w}) \rightarrow \{\mathbf{x}_1, \ldots \mathbf{x}_k\}$$



Guarantee high probability of collision for points near decision boundary:

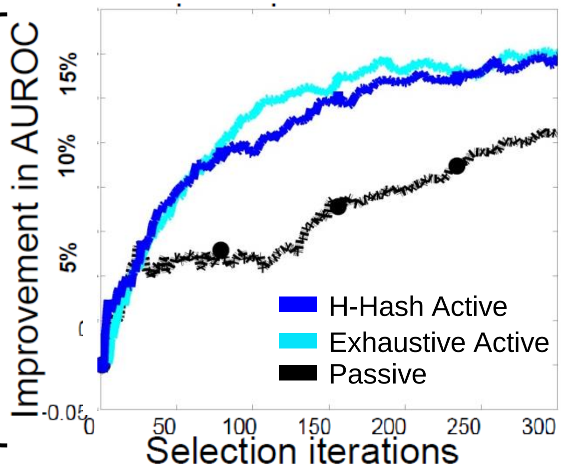
$$\Pr[h_{\mathcal{H}}(\boldsymbol{w}) = h_{\mathcal{H}}(\boldsymbol{x})] = \frac{1}{4} - \frac{1}{\pi^2}\left(\theta_{\boldsymbol{x},\boldsymbol{w}} - \frac{\pi}{2}\right)^2$$

…

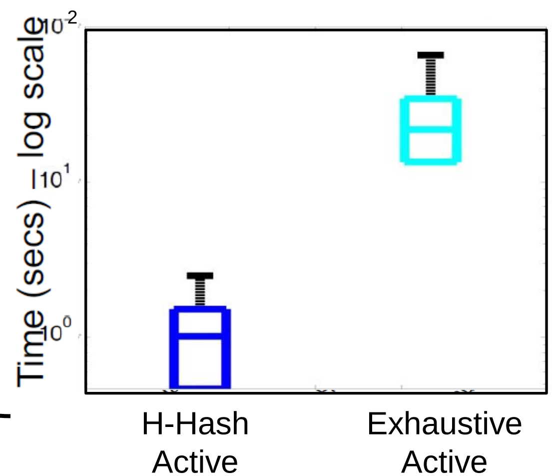
At each iteration of the learning loop, our hash functions map the current hyperplane directly to its nearest unlabeled points.

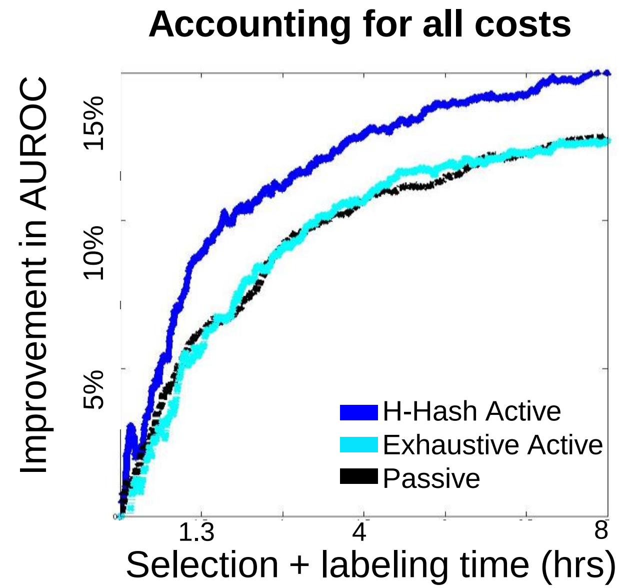
Sub-linear time active selection

Accuracy improvements as more data labeled



Time spent searching for selection



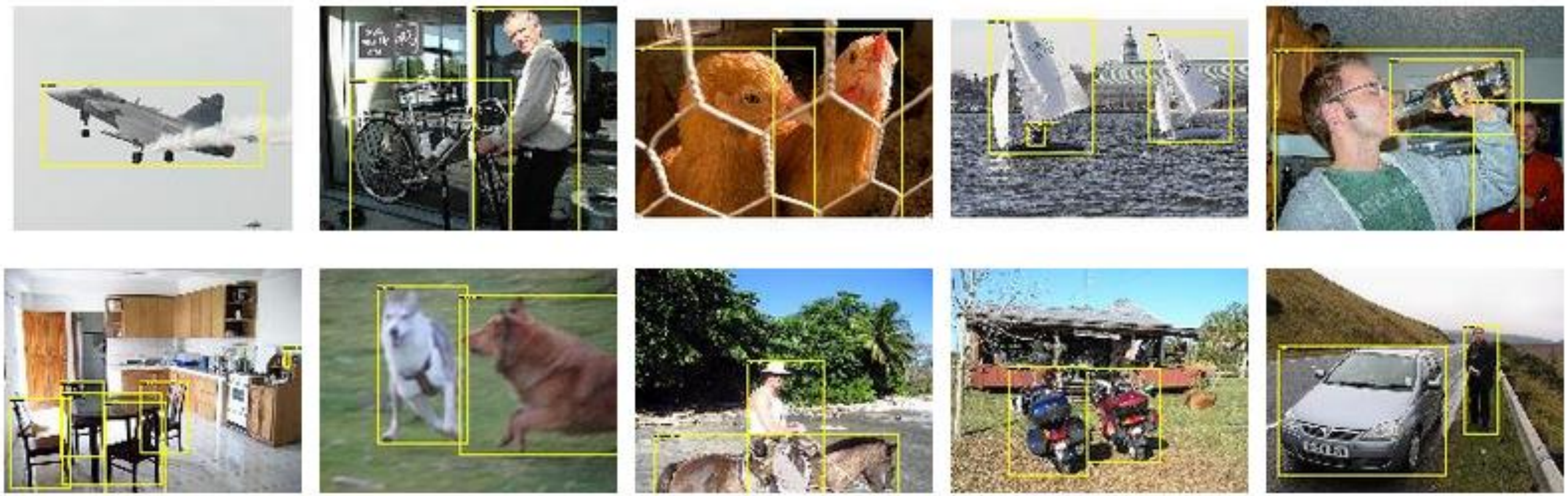


By minimizing **both** selection and labeling time, obtain the best accuracy per unit time.

H-Hash result on 1M Tiny Images

PASCAL Visual Object Categorization

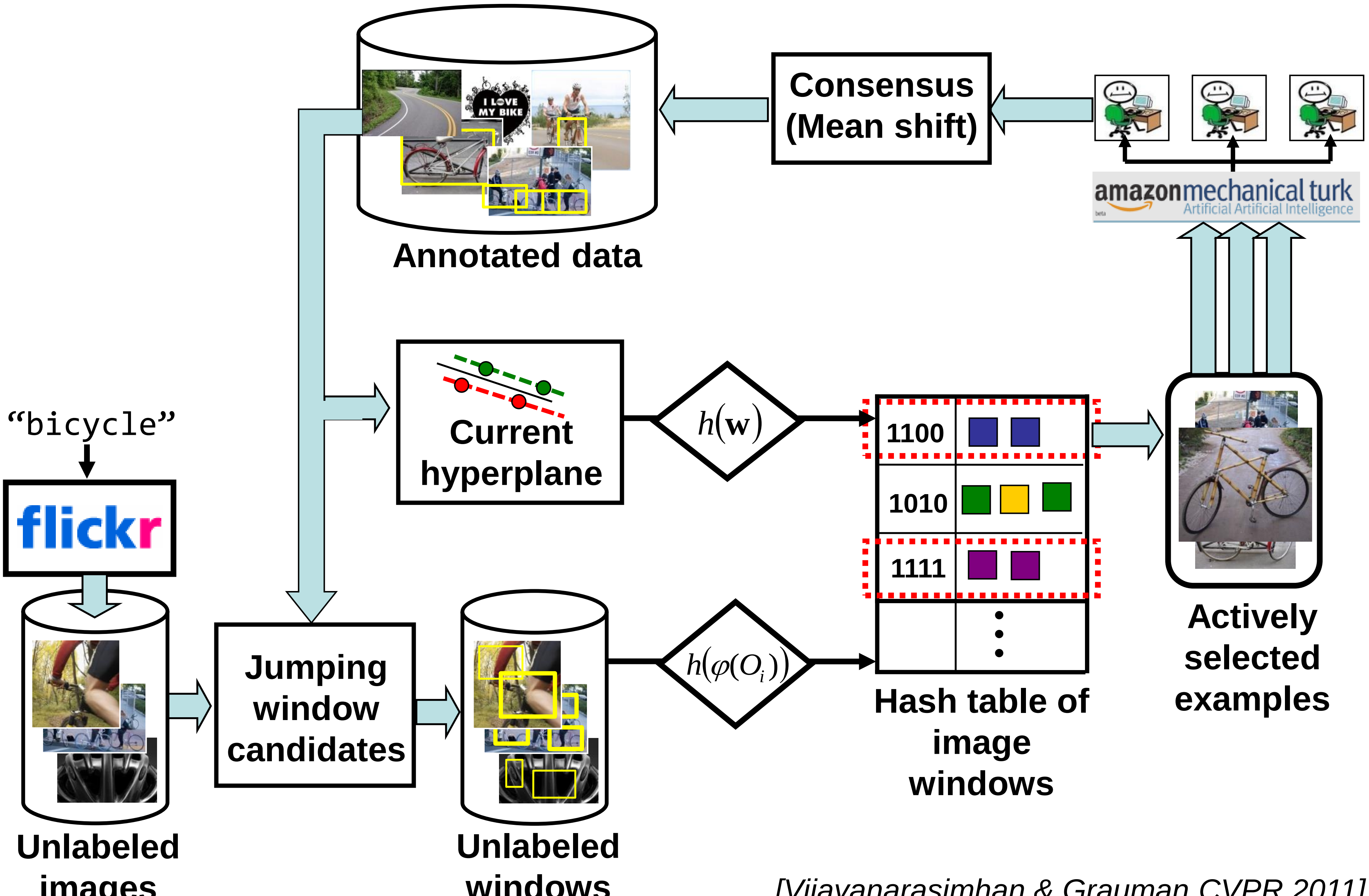
- Closely studied object detection benchmark
- Original image data from Flickr



http://pascallin.ecs.soton.ac.uk/challenges/VOC/

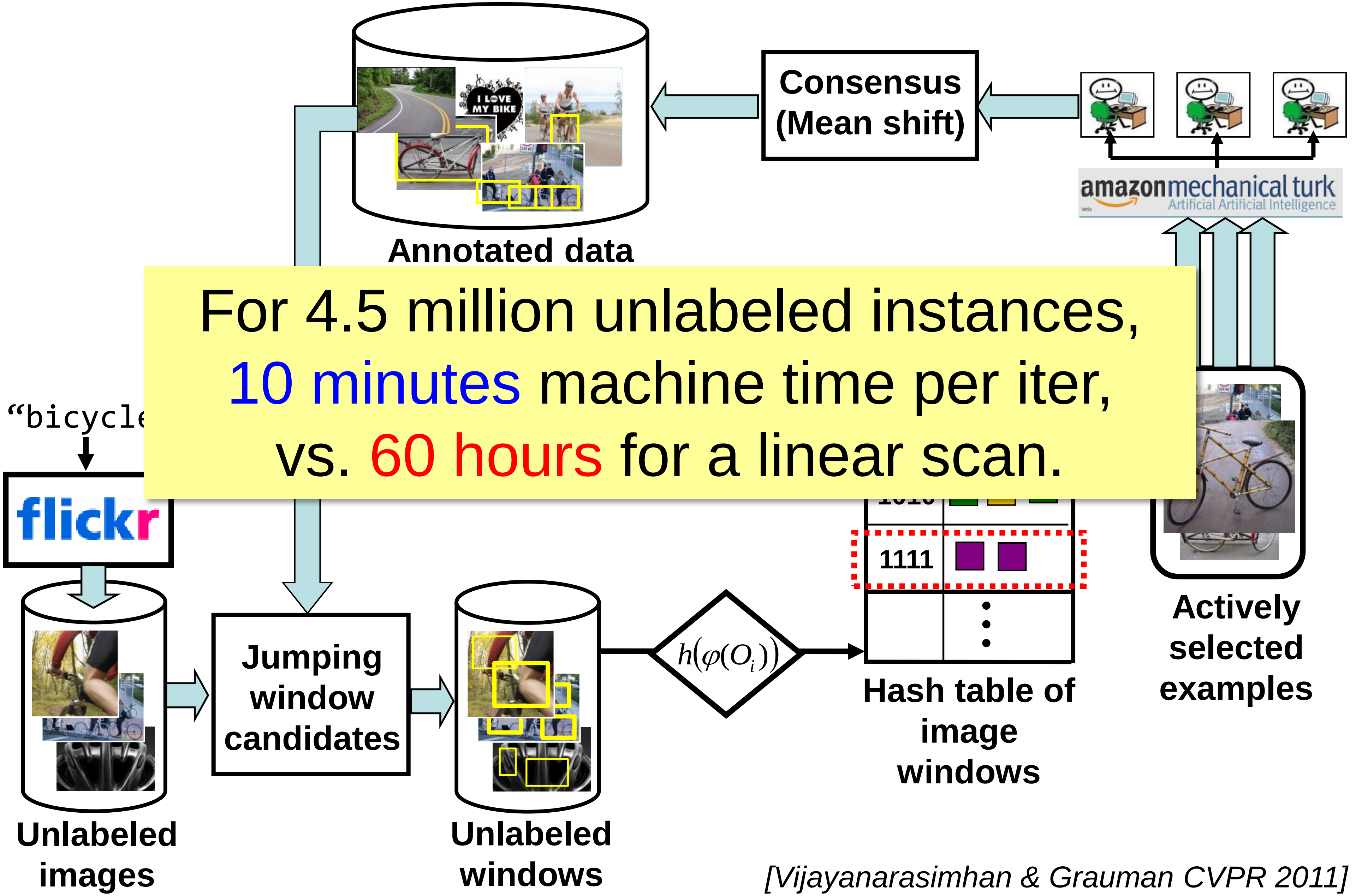
Kristen Grauman, UT Austin

Live active learning



[Vijayanarasimhan & Grauman CVPR 2011]

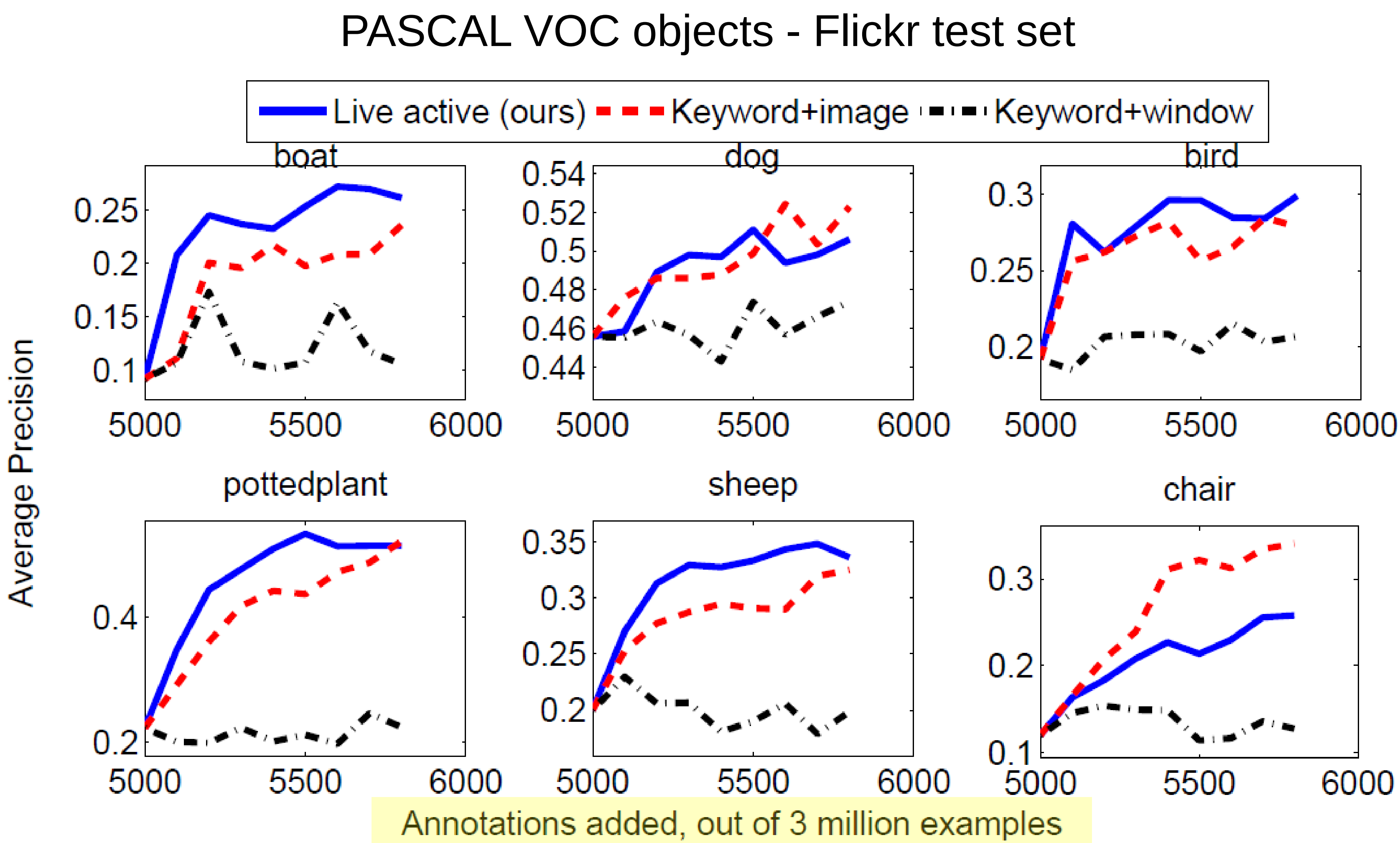
Live active learning



For 4.5 million unlabeled instances,
10 minutes machine time per iter,
vs. 60 hours for a linear scan.

[Vijayanarasimhan & Grauman CVPR 2011]

Live active learning results



Outperforms status quo data collection approach

Live active learning results

What does the live learning system ask first?

Live active learning (ours)

Keyword+image baseline

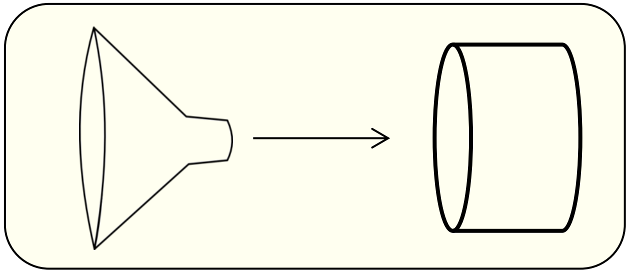
First selections made when learning “boat”

Ongoing challenges in active visual learning

- Exploration vs. exploitation
- Utility tied to specific classifier or model
- Joint batch selection (“non-myopic”) expensive, remains challenging
- Crowdsourcing: reliability, expertise, economics
- Active annotations for objects/activity in video

Our goal

Teaching computers about visual categories must be an ongoing, interactive process, with communication that goes beyond labels.



This talk:

1. Active visual learning
2. Learning from visual comparisons

Visual attributes

- High-level semantic properties shared by objects
- Human-understandable and machine-detectable

[Oliva et al. 2001, Ferrari & Zisserman 2007, Kumar et al. 2008, Farhadi et al. 2009, Lampert et al. 2009, Endres et al. 2010, Wang & Mori 2010, Berg et al. 2010, Branson et al. 2010, Parikh & Grauman 2011, …]

Mule

Attributes

A mule…

Is furry

Has four legs

Has a tail

Binary attributes

A mule…

Is furry

Has four legs

Has a tail

[Ferrari & Zisserman 2007, Kumar et al. 2008, Farhadi et al. 2009, Lampert et al. 2009, Endres et al. 2010, Wang & Mori 2010, Berg et al. 2010, Branson et al. 2010, …]

Relative attributes

A mule…

Is furry

Has four legs

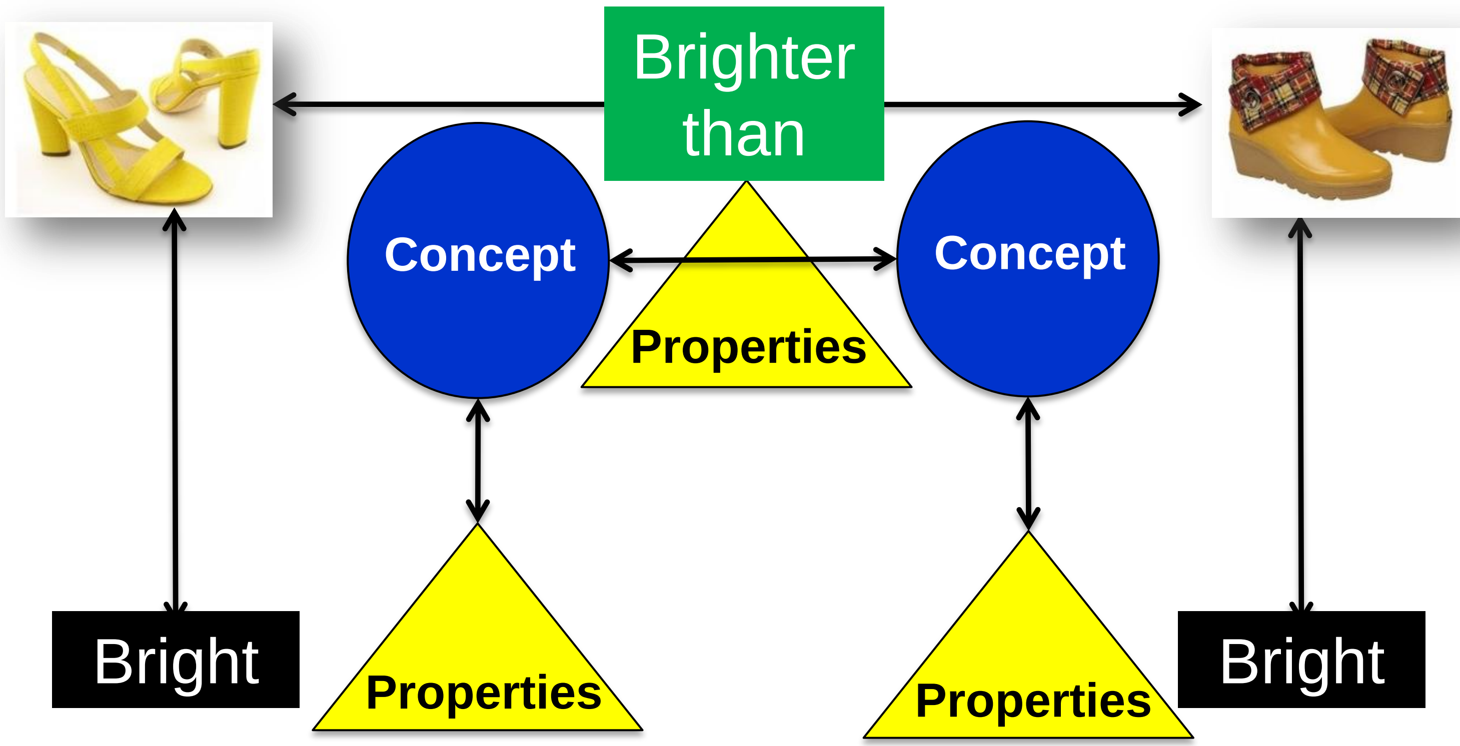
Has a tail

Legs **shorter** than horses'

Tail **longer** than donkeys'

Relative attributes

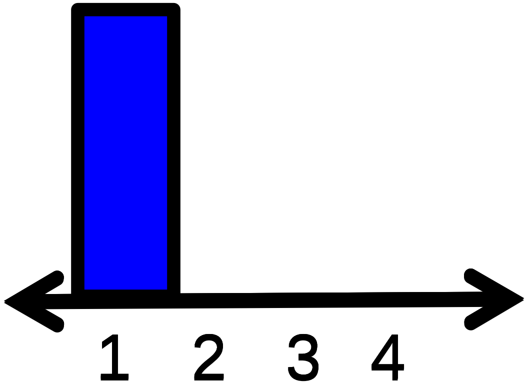
Idea: represent ***visual comparisons*** between classes, images, and their properties.



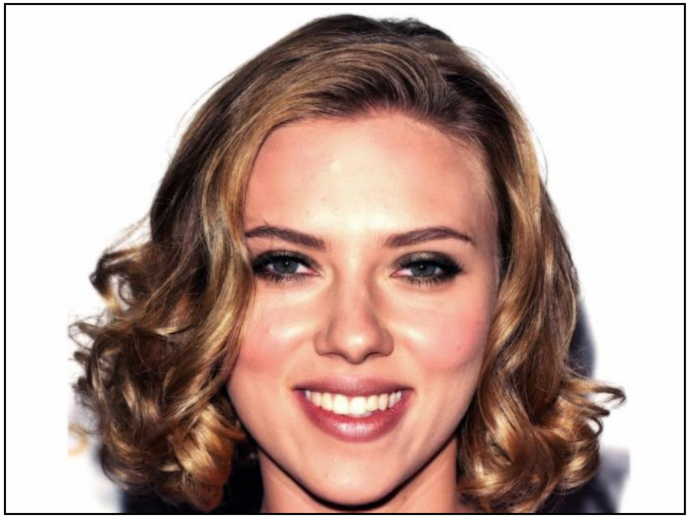
[Parikh & Grauman, ICCV 2011]

How to teach relative visual concepts?

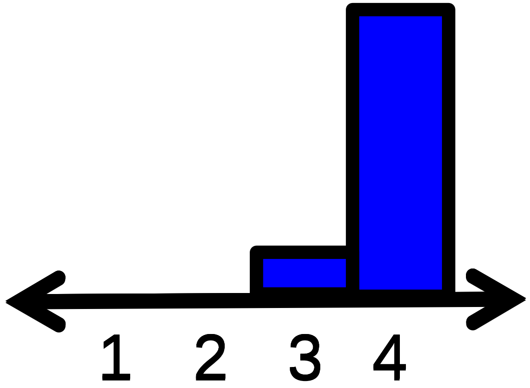
How much is the person smiling?



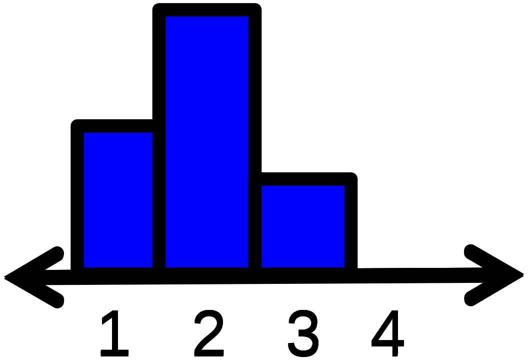
How to teach relative visual concepts?



How much is the person smiling?



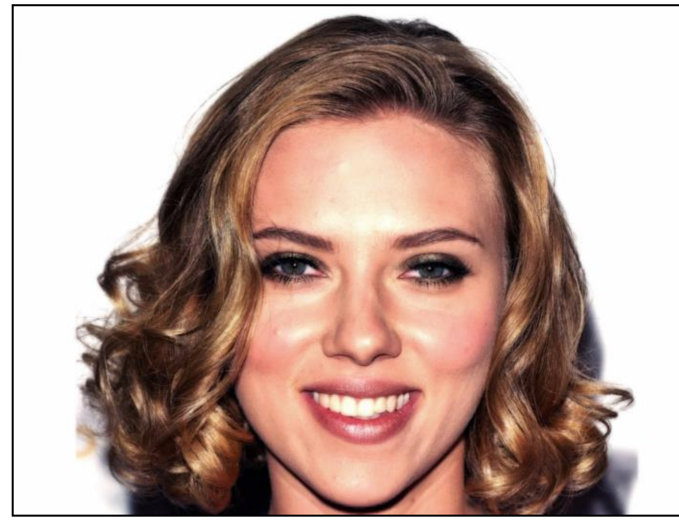
How to teach relative visual concepts?

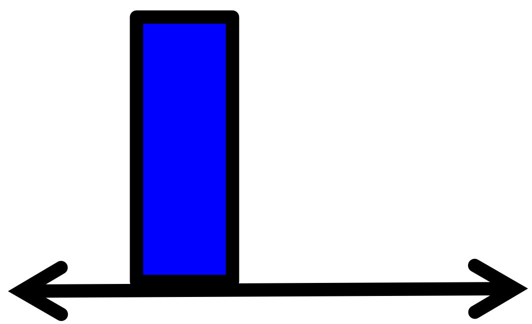


How much is the person smiling?

How to teach relative visual concepts?

> ?



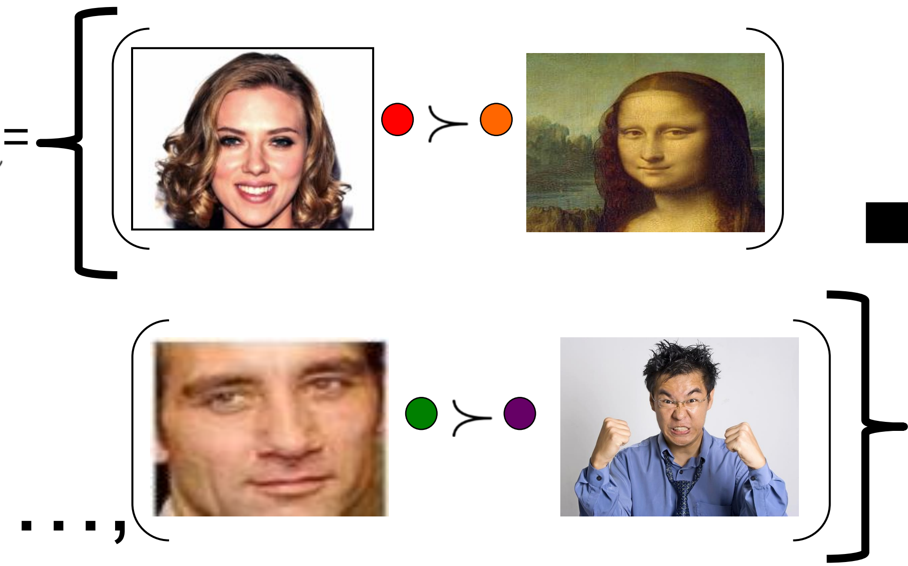


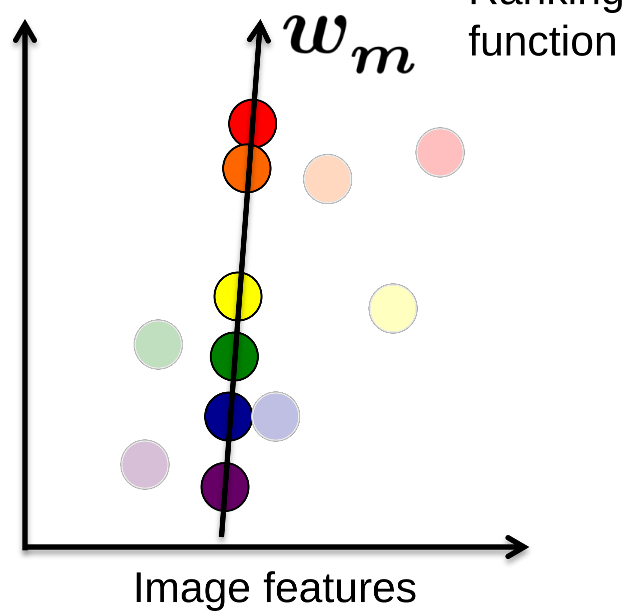
Less More

Learning relative attributes

For each attribute, use ordered image pairs to train a ranking function:

$O_m =$





$$\boldsymbol{w_m^T x_i} > \boldsymbol{w_m^T x_j}$$

$$\forall (i, j) \in O_m$$

[Parikh & Grauman, ICCV 2011; Joachims 2002]

Relating images

Rather than simply label images with their properties,

Not bright

Smiling

Not natural

Relating images

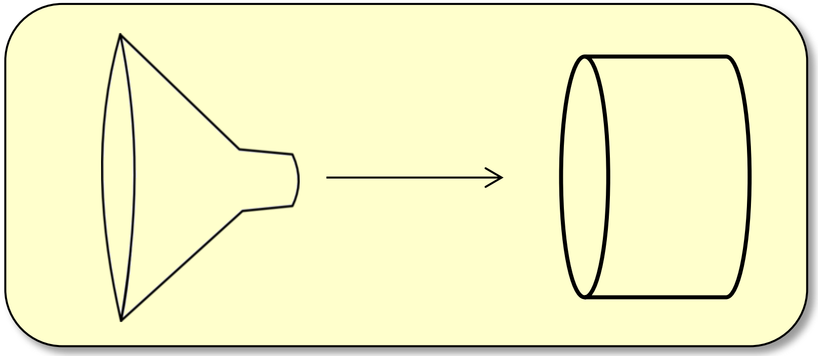
Now we can compare images by attribute's "strength"

Kristen Grauman, UT Austin

Learning with visual comparisons

Enable new modes of human-system communication

- Training category models through descriptions
- Rationales to explain image labels
- Semantic relative feedback for image search
- Analogical constraints on feature learning



Relative zero-shot learning

Training: Images from **S seen** categories and
Descriptions of **U unseen** categories



Age: **Hugh** $\succ$ **Clive** $\succ$ **Scarlett** **Jared** $\succ$ **Miley**

Smiling: **Miley** $\succ$ **Jared**

Need not use all attributes, nor all seen categories

Testing: Categorize image into one of **S**+**U** classes

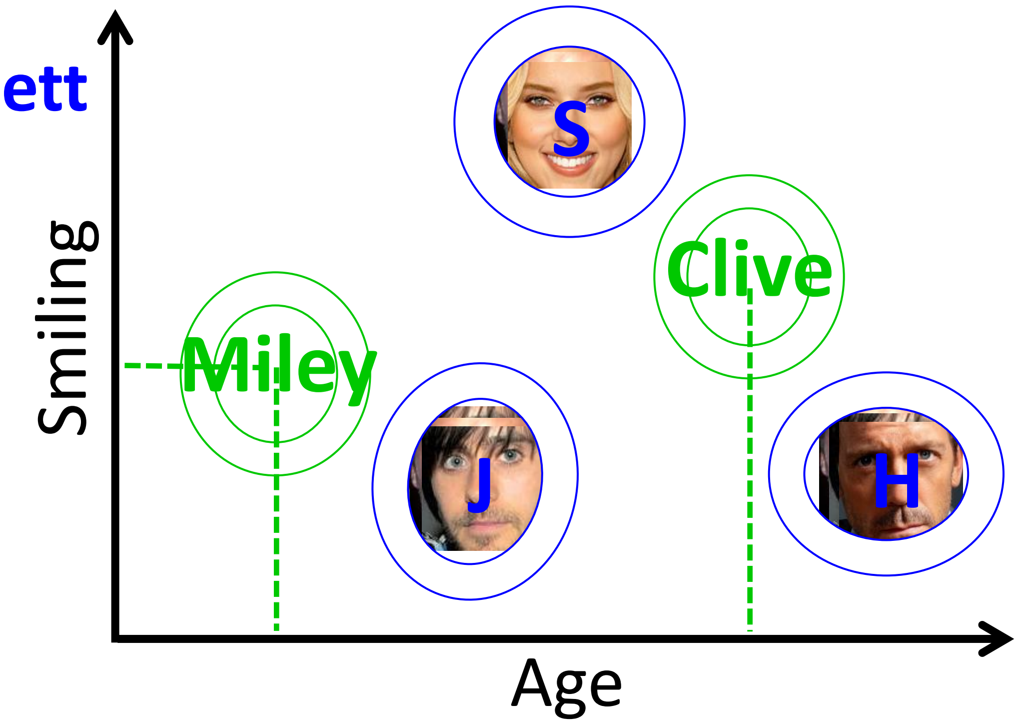
Relative zero-shot learning

Predict new classes based on their **relationships** to existing classes – even without training images.

Age: Hugh $\succ$ Clive $\succ$ Scarlett

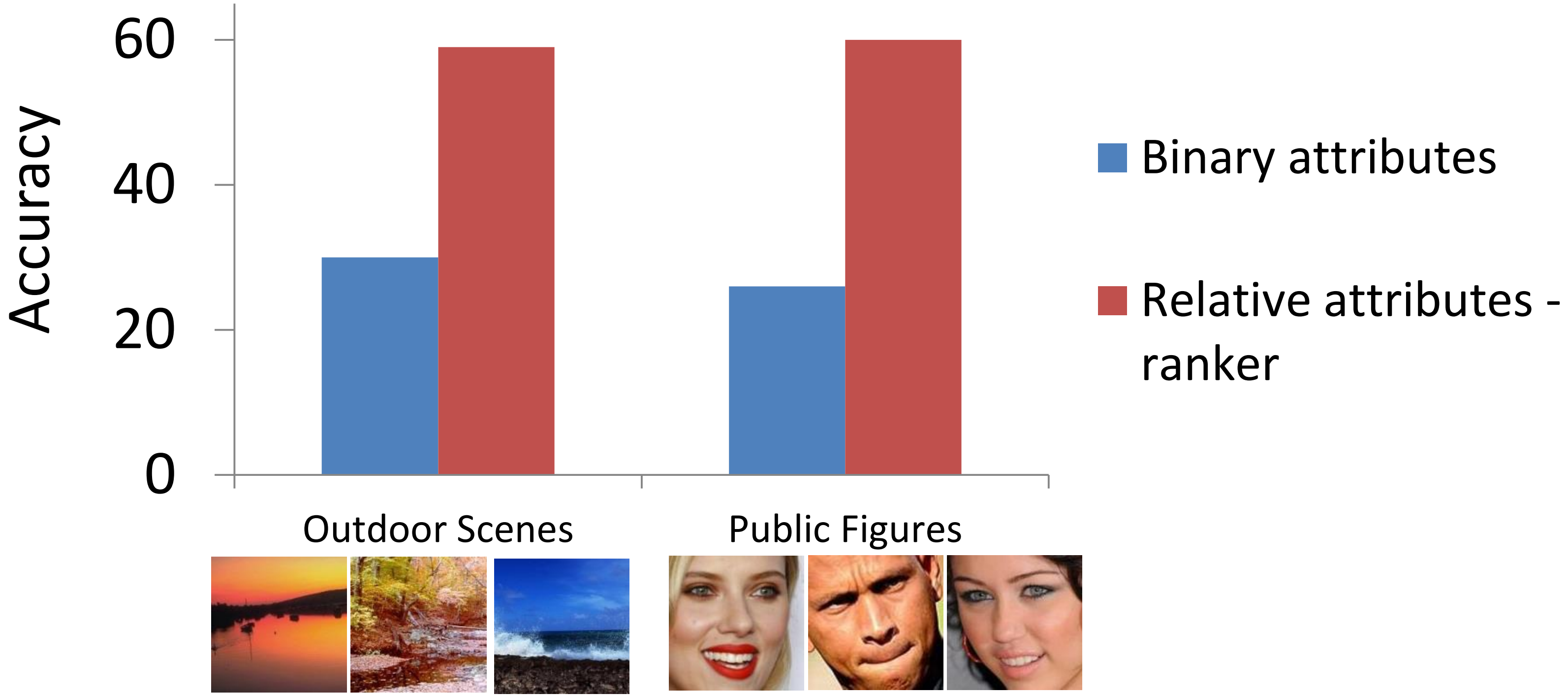
Jared $\succ$ Miley

Smiling: Miley $\succ$ Jared



$$c_i^{(s)} \sim \mathcal{N}(\boldsymbol{\mu}_i^{(s)}, \Sigma_i^{(s)})$$

Relative zero-shot learning

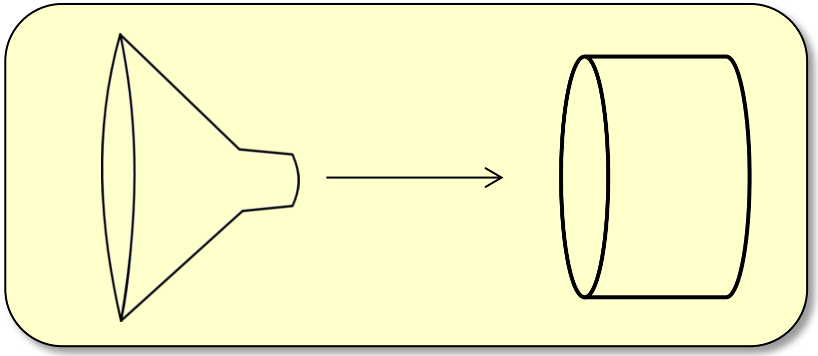


Comparative descriptions are more discriminative than categorical descriptions.

Learning with visual comparisons

Enable new modes of human-system communication

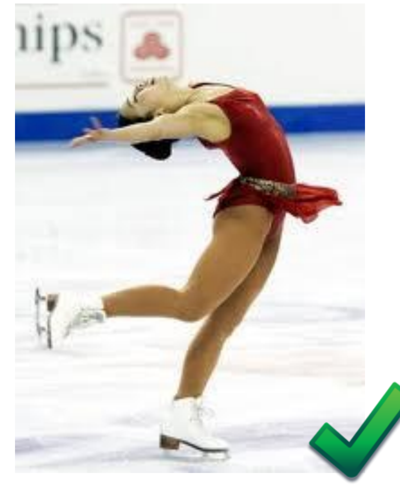
- Training category models through descriptions
- Rationales to explain image labels
- Semantic relative feedback for image search
- Analogical constraints on feature learning



Soliciting visual rationales

Is the team winning?
How can you tell?

Is it a safe route?
How can you tell?



Is her form good?
How can you tell?

Main idea:

- Ask the annotator not just *what*, but also *why.*

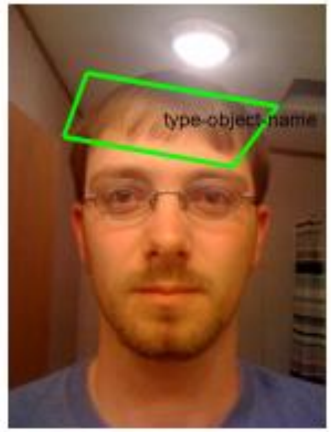
[Donahue and Grauman, ICCV 2011; Zaidan et al. NAACL HLT 2007]

Soliciting visual rationales

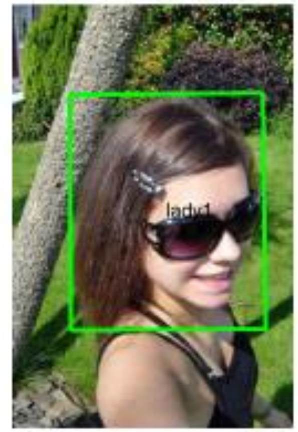
Hot or Not? *How can you tell?*

Spatial rationales

Hot, Male



Not, Male

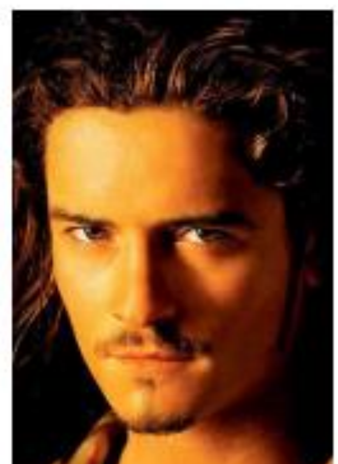


Hot, Female

Not, Female

Attribute rationales

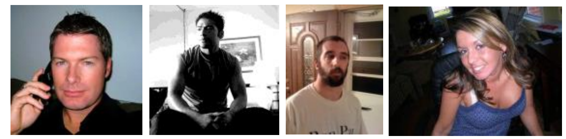
Youth
Smiling
Straight Hair
Narrow Eyes



Youth
Black Hair
Goatee
Square Face
Shiny Skin
High Cheekbones

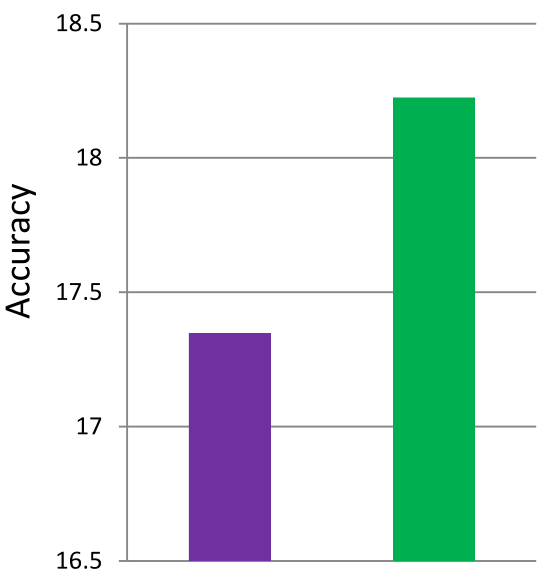
Soliciting visual rationales

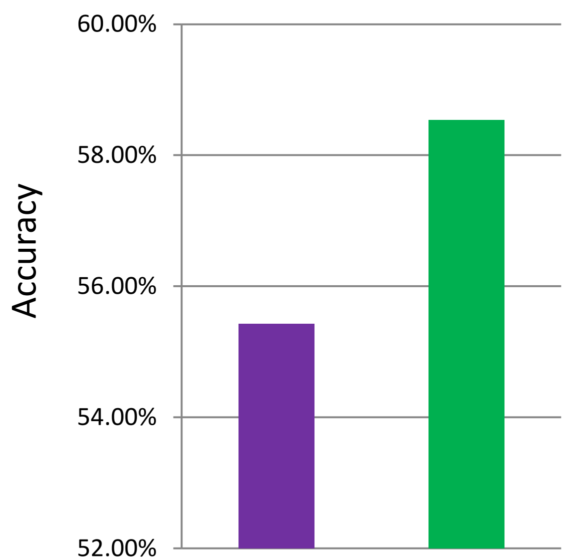
Scene categories

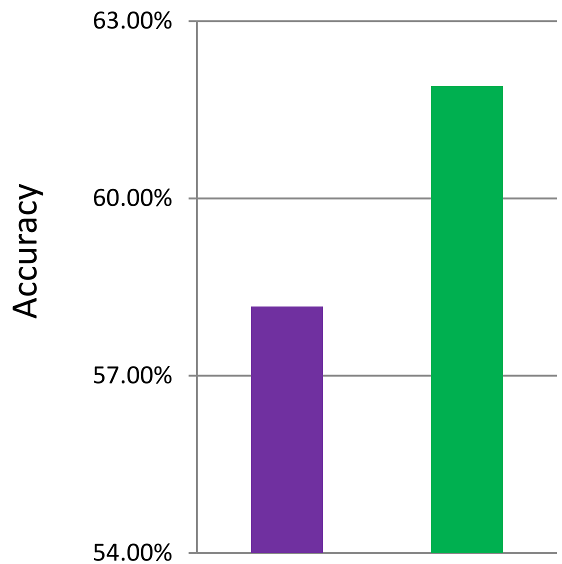


Hot or Not

Attractiveness







Original labels only

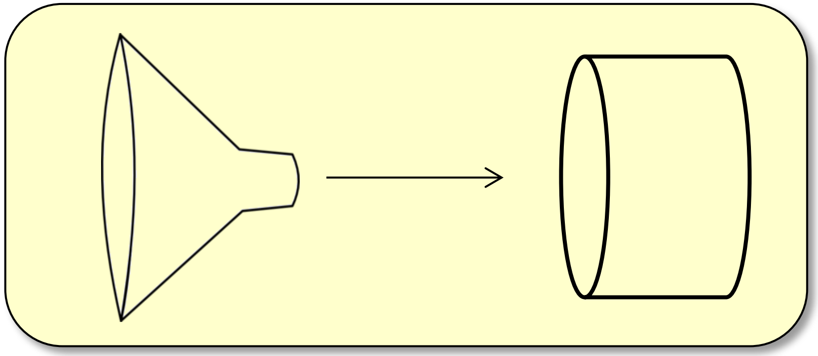
+ Rationales

[Donahue & Grauman, ICCV 2011]

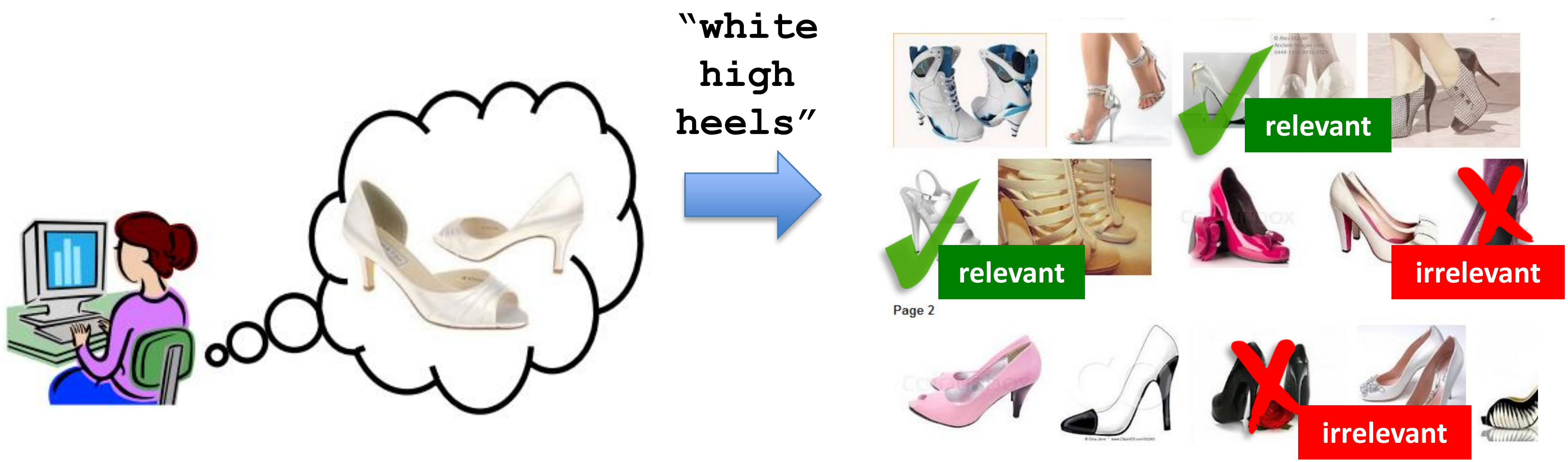
Learning with visual comparisons

Enable new modes of human-system communication

- Training category models through descriptions
- Rationales to explain image labels
- Semantic relative feedback for image search
- Analogical constraints on feature learning



Interactive visual search

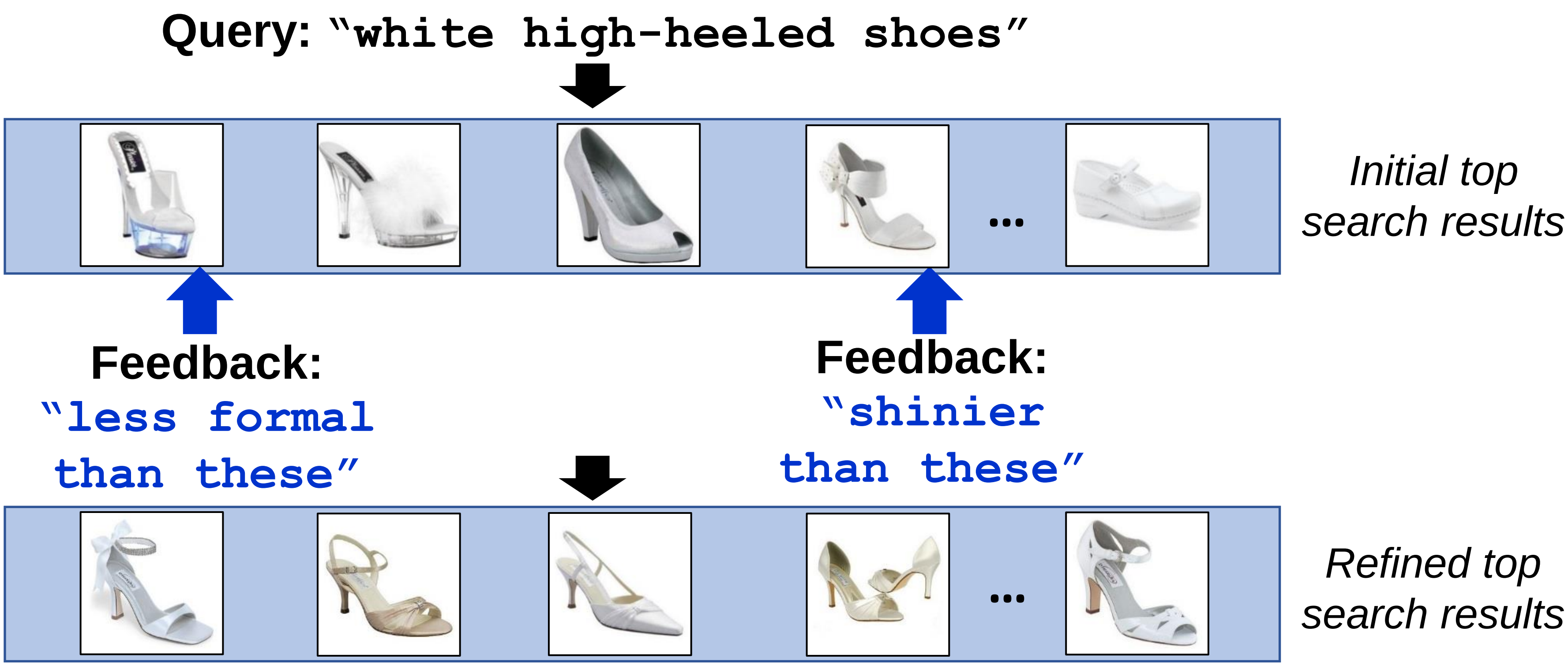


Traditional **binary relevance feedback** offers only coarse communication between user and system

[Rui et al. 1998, Zhou et al. 2003, …]

WhittleSearch: Relative attribute feedback

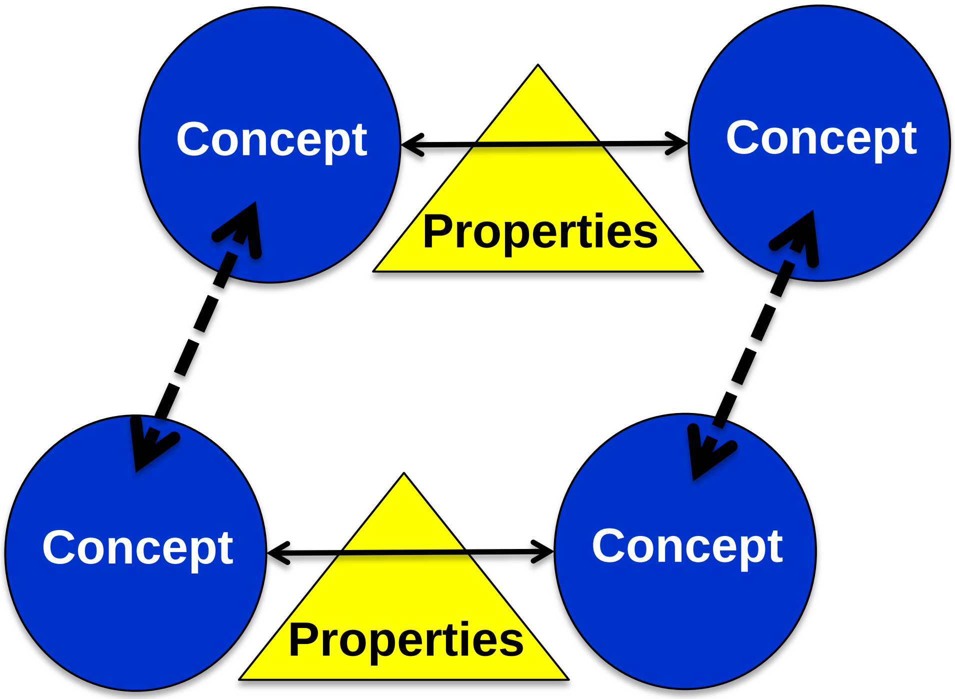
[Kovashka, Parikh, and Grauman, CVPR 2012]



Whittle away irrelevant images via precise semantic feedback

Visual analogies

Beyond pairwise comparisons …



[Hwang, Grauman, & Sha, ICML 2013]

Learning with visual analogies

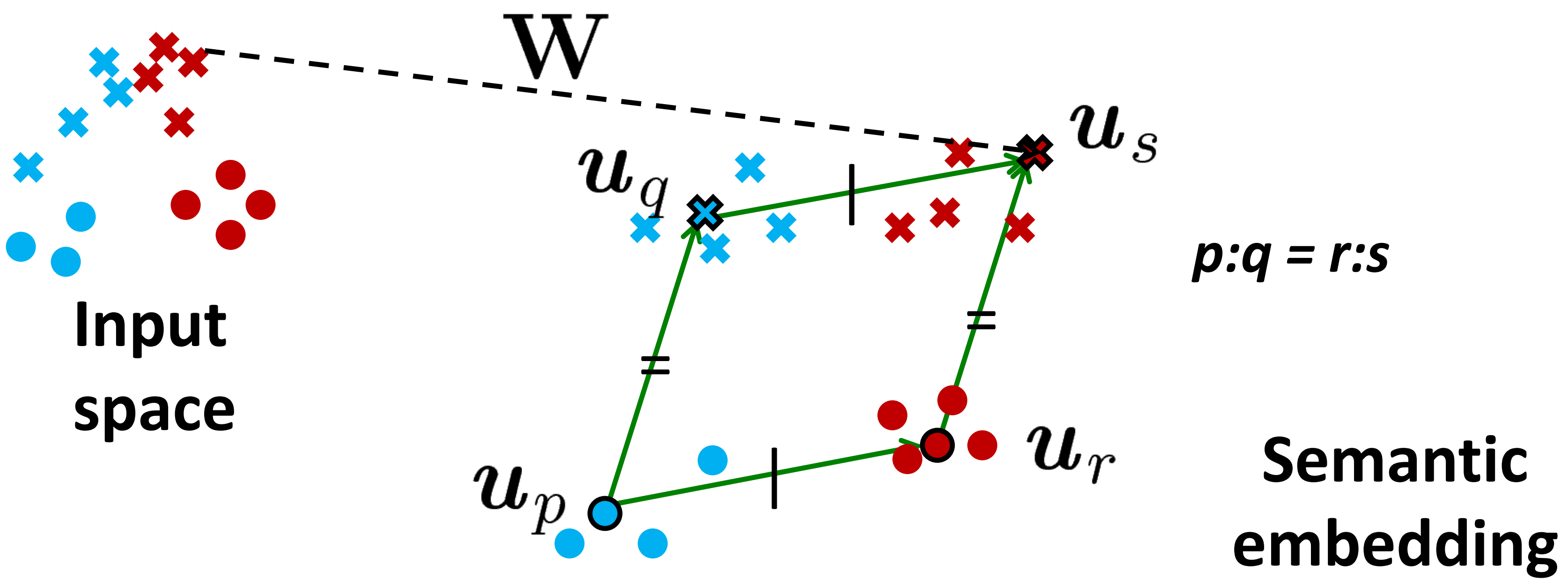
Regularize object models with analogies

planet : sun = electron : nucleus

[Hwang, Grauman, & Sha, ICML 2013]

Learning with visual analogies

Regularize object models with analogies

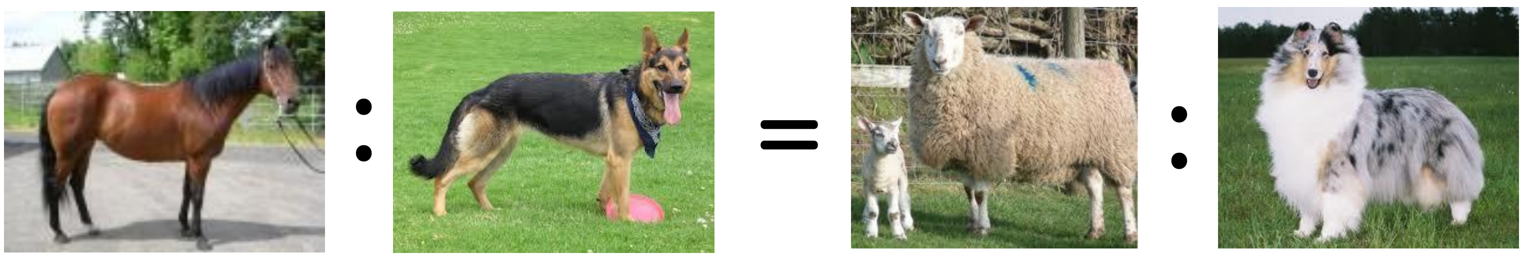


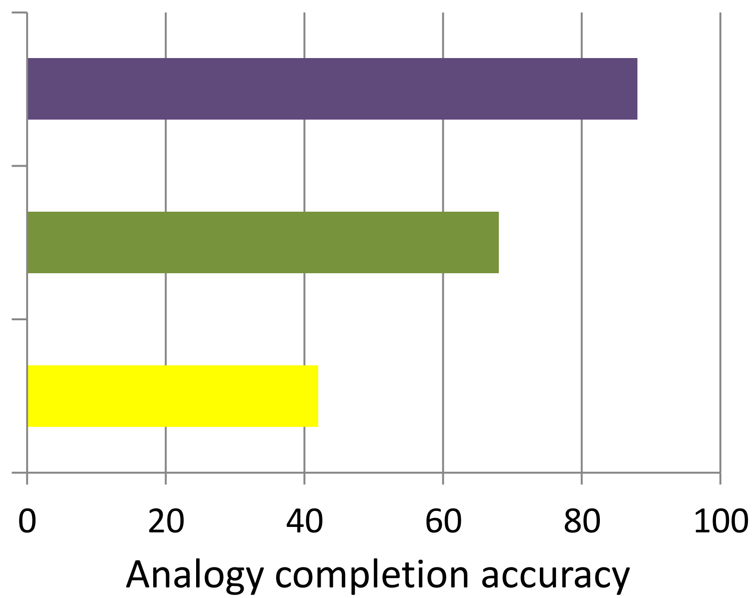
[Hwang, Grauman, & Sha, ICML 2013]

Kristen Grauman, UT Austin

Visual analogies

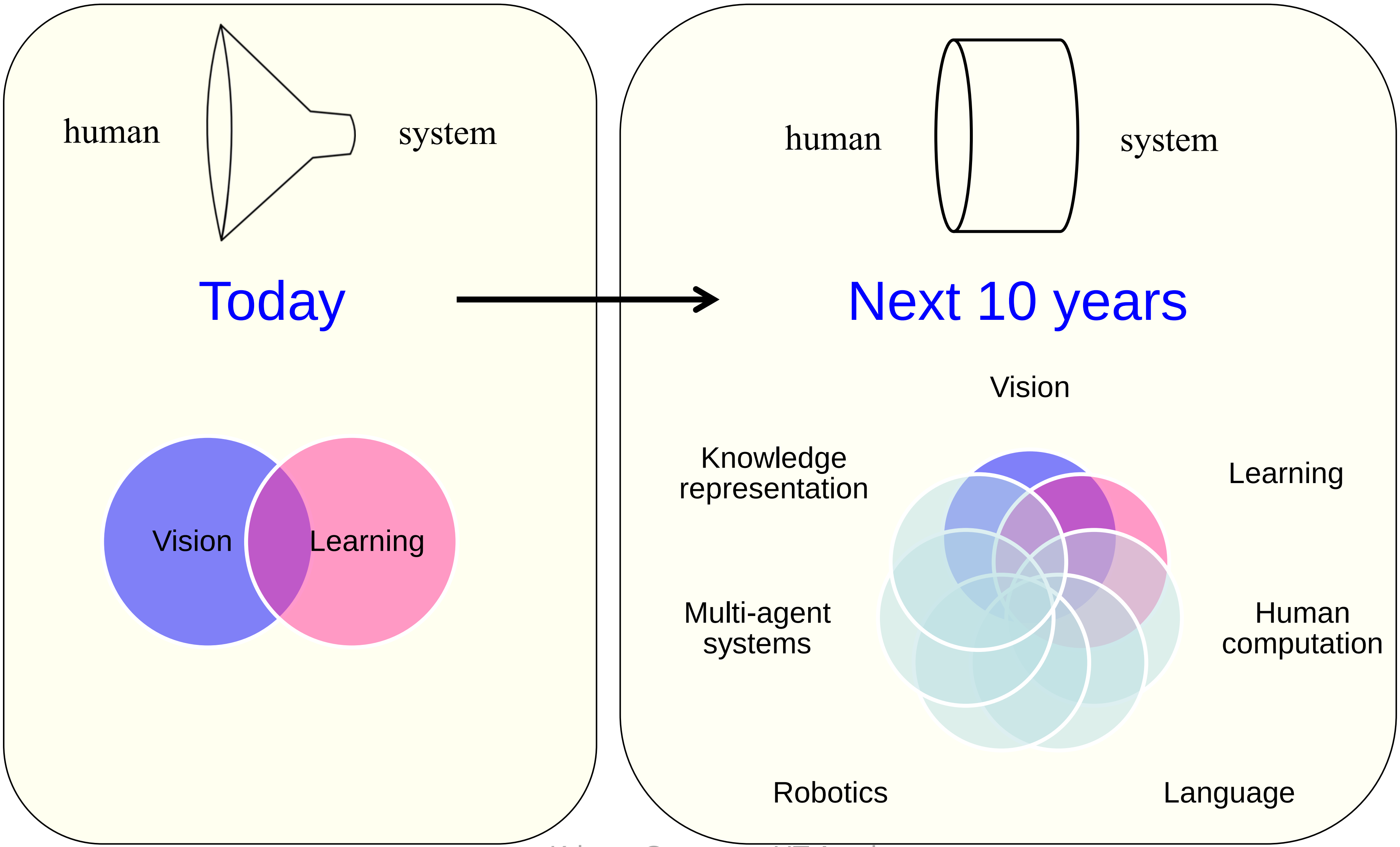
GRE-like visual analogy tests





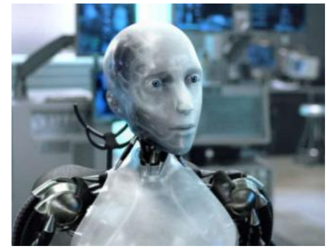
[Hwang, Grauman, & Sha, ICML 2013]

Teaching visual recognition systems



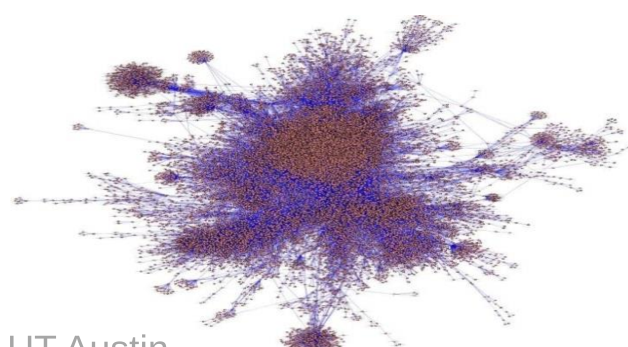
Important next directions

- Recognition in action: embodied, egocentric



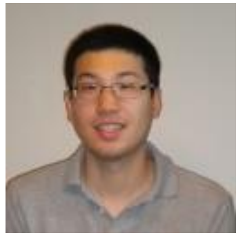
- Activity understanding: objects & actions

- Scale: many classes, fine-grained recognition

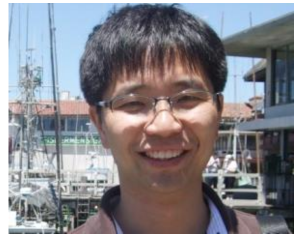


Acknowledgements

Sudheendra Vijayanarasimhan



Yong Jae Lee



Jaechul Kim

Adriana Kovashka

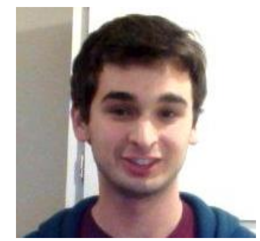
Sung Ju Hwang

Chao-Yeh Chen

Suyog Jain

Dinesh Jayaraman

Aron Yu



Jeff Donahue

Devi Parikh (Virginia Tech), Fei Sha (USC), Prateek Jain (MSR), Trevor Darrell (UC Berkeley)

J. K. Aggarwal, Ray Mooney, Peter Stone, Bruce Porter, and all my UT colleagues

NSF, ONR, DARPA, Luce Foundation, Google, Microsoft

Summary

- Humans are not simply “label machines”
- More data need not mean better learning
- Widen access to visual knowledge through
  - Large-scale interactive/active learning systems
  - Representing relative visual comparisons
- Visual recognition offers new AI challenges, and progress demands that more AI ideas convene

THE UNIVERSITY OF TEXAS AT AUSTIN

References

- WhittleSearch: Image Search with Relative Attribute Feedback. A. Kovashka, D. Parikh, and K. Grauman. CVPR 2012.
- Hashing Hyperplane Queries to Near Points with Applications to Large-Scale Active Learning. P. Jain, S. Vijayanarasimhan, and K. Grauman. NIPS 2010.
- Annotator Rationales for Visual Recognition. J. Donahue and K. Grauman. ICCV 2011.
- Actively Selecting Annotations Among Objects and Attributes. A. Kovashka, S. Vijayanarasimhan, and K. Grauman. ICCV 2011.
- Large-Scale Live Active Learning: Training Object Detectors with Crawled Data and Crowds. S. Vijayanarasimhan and K. Grauman. CVPR 2011.
- Cost-Sensitive Active Visual Category Learning. S. Vijayanarasimhan and K. Grauman. *International Journal of Computer Vision* (IJCV), Vol. 91, Issue 1 (2011), p. 24.
- What’s It Going to Cost You?: Predicting Effort vs. Informativeness for Multi-Label Image Annotations. S. Vijayanarasimhan and K. Grauman. CVPR 2009.
- Multi-Level Active Prediction of Useful Image Annotations for Recognition. S. Vijayanarasimhan and K. Grauman. NIPS 2008.
- Far-Sighted Active Learning on a Budget for Image and Video Recognition. S. Vijayanarasimhan, P. Jain, and K. Grauman. CVPR 2010.
- Relative Attributes. D. Parikh and K. Grauman. ICCV 2011.
- Analogy-Preserving Semantic Embedding for Visual Object Categorization. S. J. Hwang, K. Grauman, and F. Sha. ICML 2013.