Teaching visual recognition systems

Kristen Grauman
Department of Computer Science
University of Texas at Austin

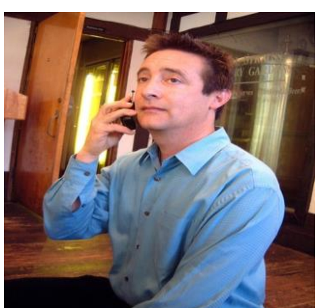
Work with Sudheendra Vijayanarasimhan, Prateek Jain, Devi Parikh, Adriana Kovashka, and Jeff Donahue

Visual categories

Beyond instances, need to recognize and detect *classes* of visually and semantically related…


Objects


Scenes

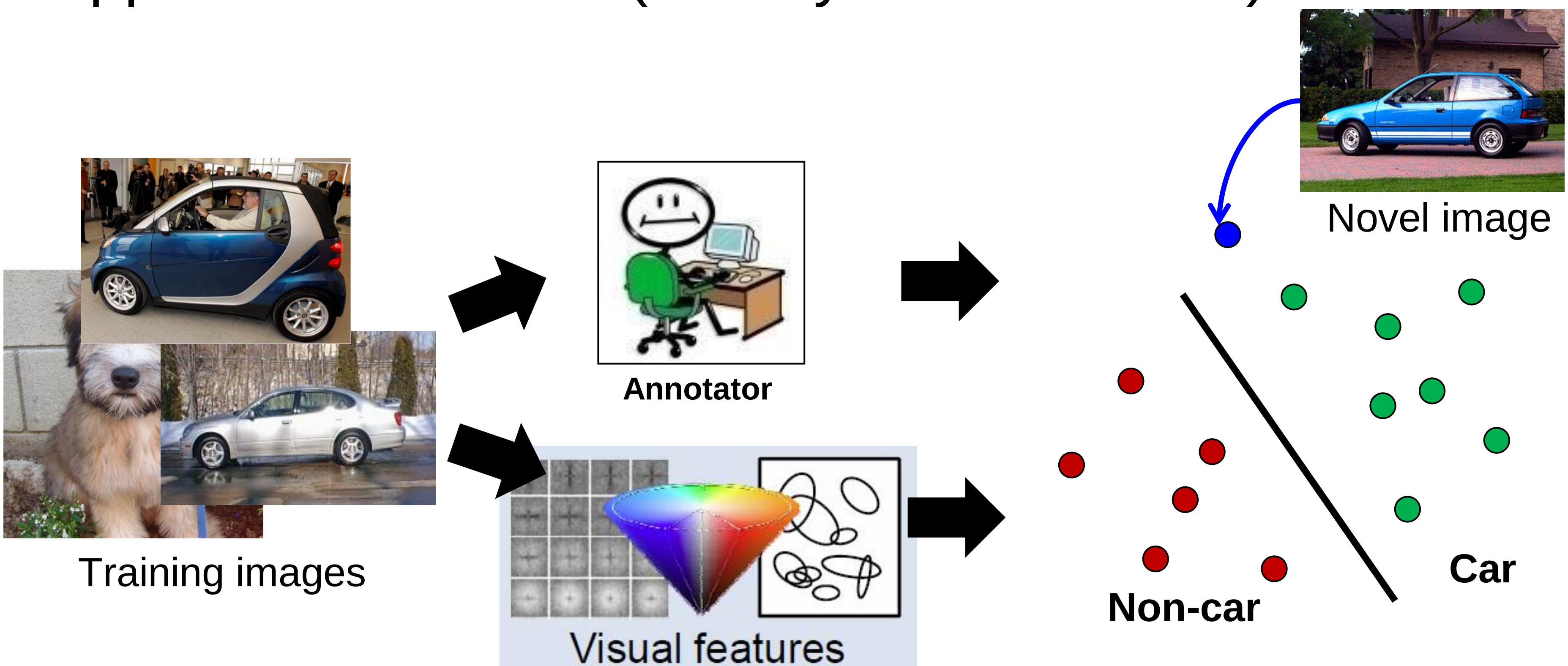



Activities

Kristen Grauman, UT-Austin

Learning-based methods

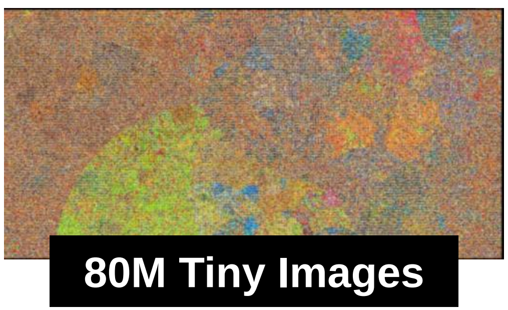
Last ~10 years: impressive strides by *learning* appearance models (usually discriminative).



Kristen Grauman, UT-Austin

Exuberance for image data (and their category labels)

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



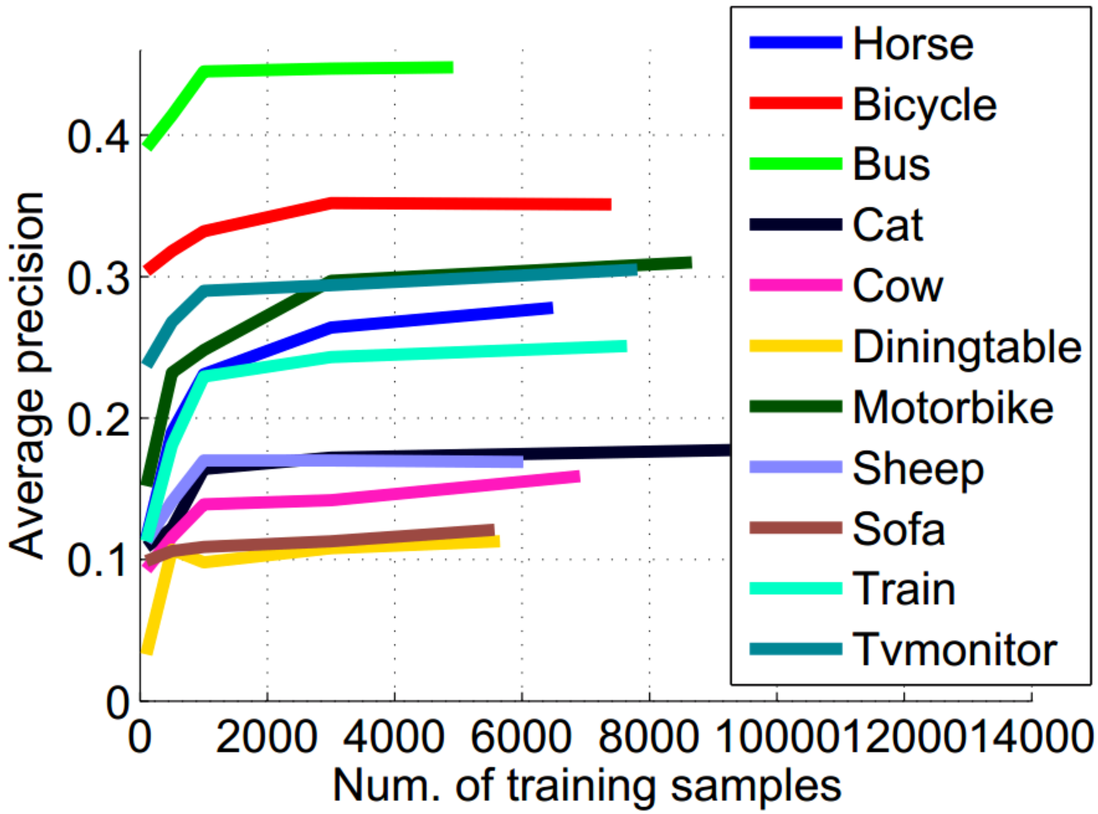
80M images
53K noisily labeled object categories
[Torralba et al. 2008]

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

Kristen Grauman, UT-Austin

And yet…

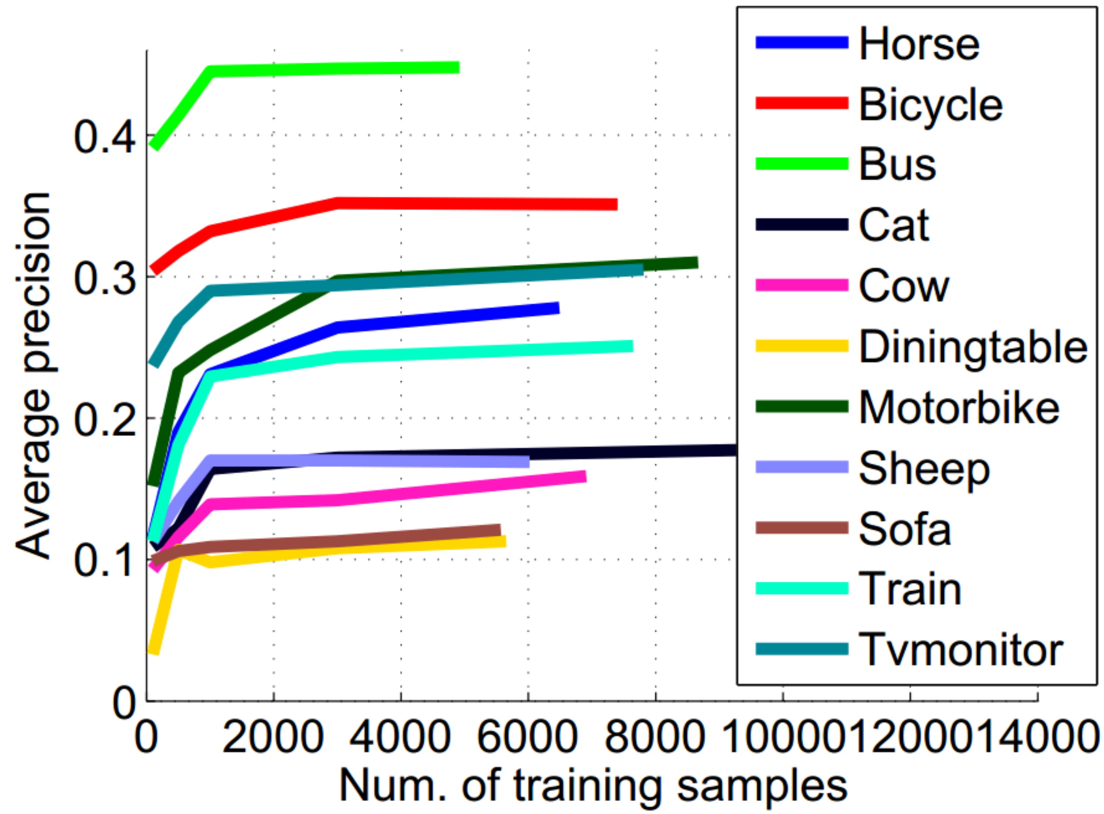
- More data ↔ more accurate visual models?
- Which



X. Zhu, C. Vondrick, D. Ramanan and C. Fowlkes. Do We Need More Training Data or Better Models for Object Detection? BMVC 2012.

And yet…

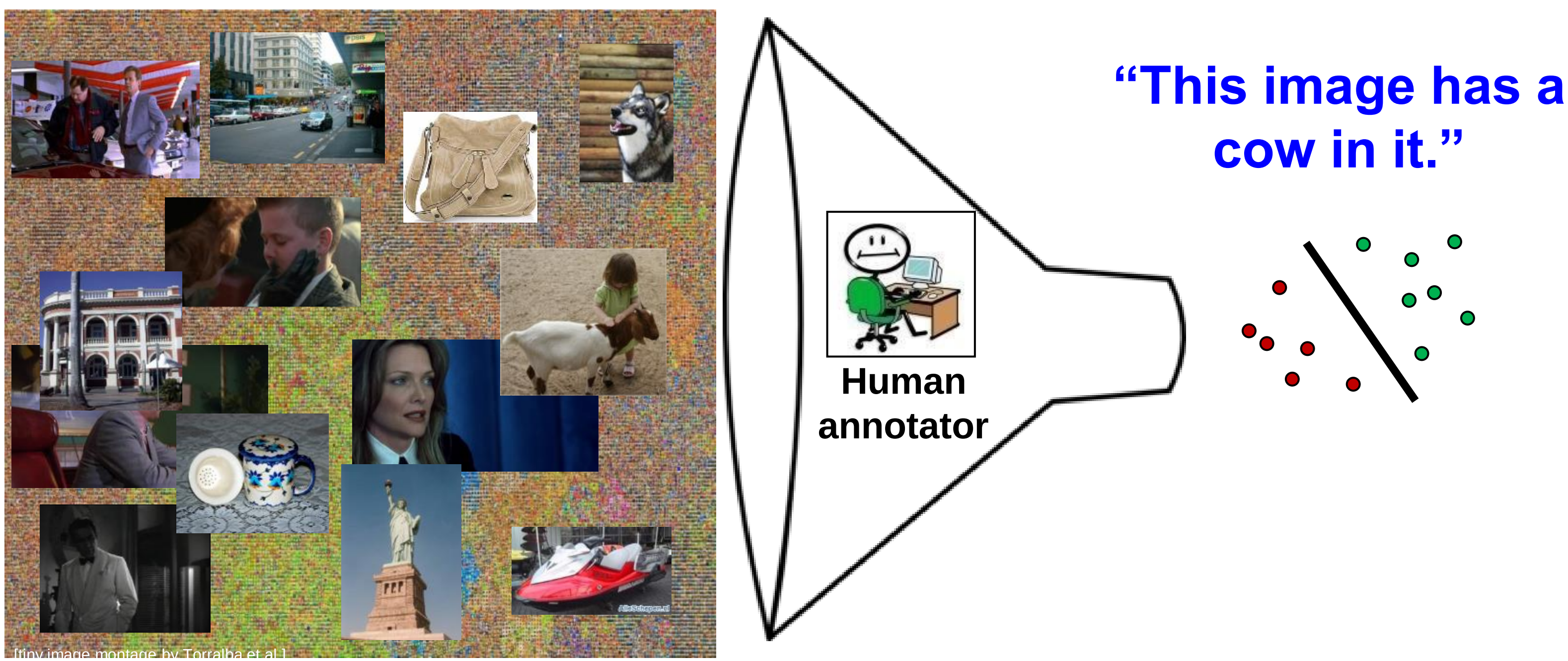
- More data ↔ more accurate visual models?



X. Zhu, C. Vondrick, D. Ramanan and C. Fowlkes. Do We Need More Training Data or Better Models for Object Detection? BMVC 2012.

And yet…

- More data ↔ more accurate visual models?
- Which images should be labeled?
- Are labels enough to teach visual concepts?



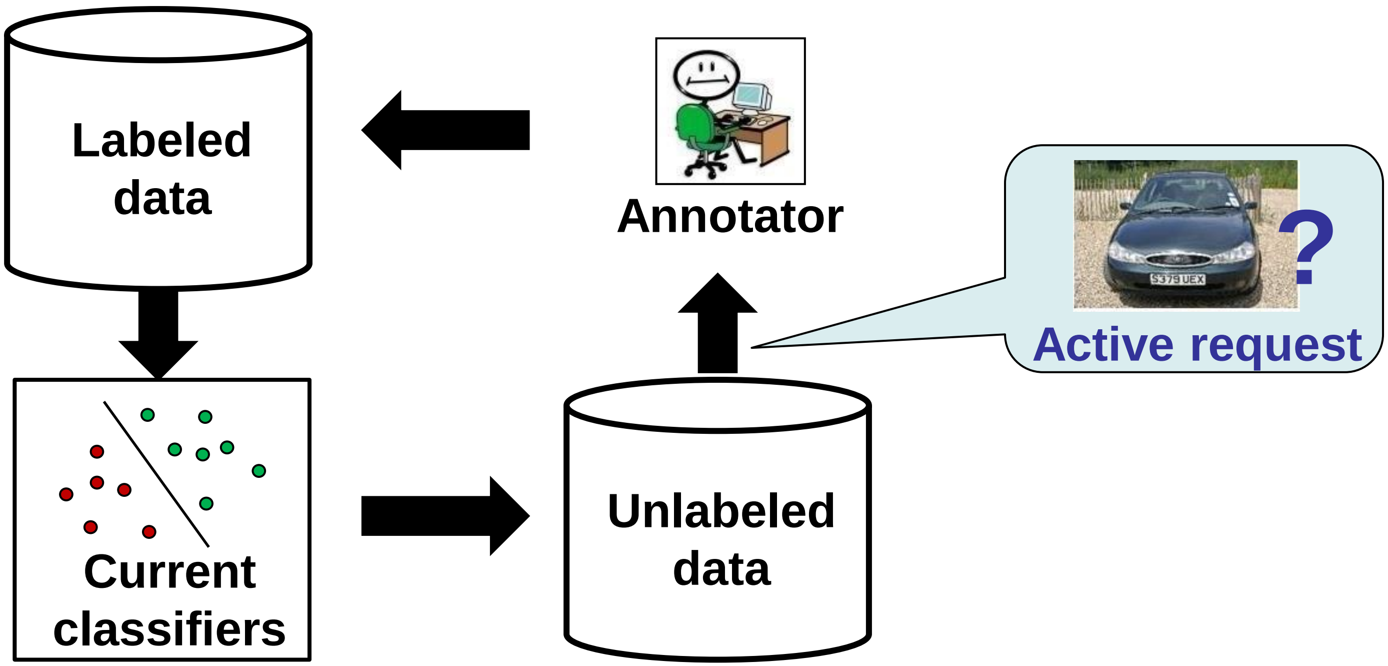
Kristen Grauman, UT-Austin

This lecture

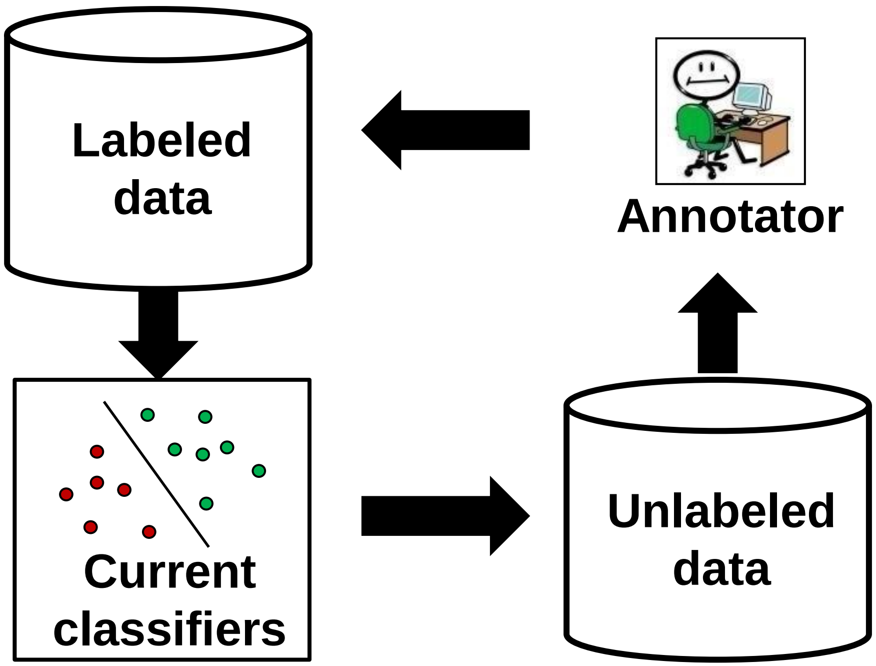
Teaching machines visual categories

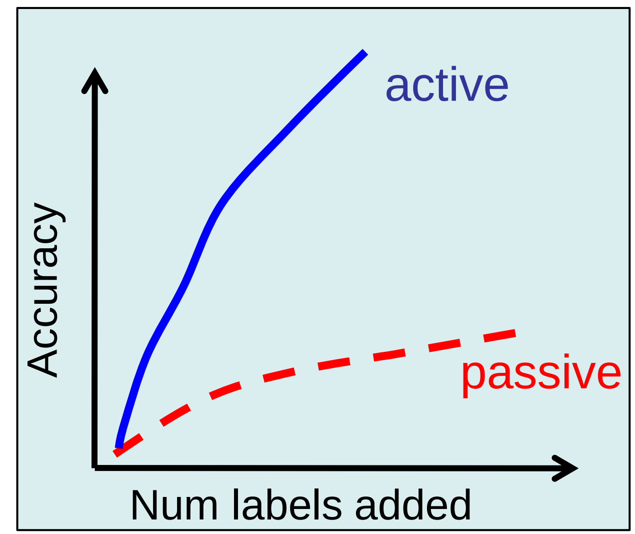
- Active learning to prioritize informative annotations
- Relative attributes to learn from visual comparisons

Active learning for image annotation



Active learning for image annotation



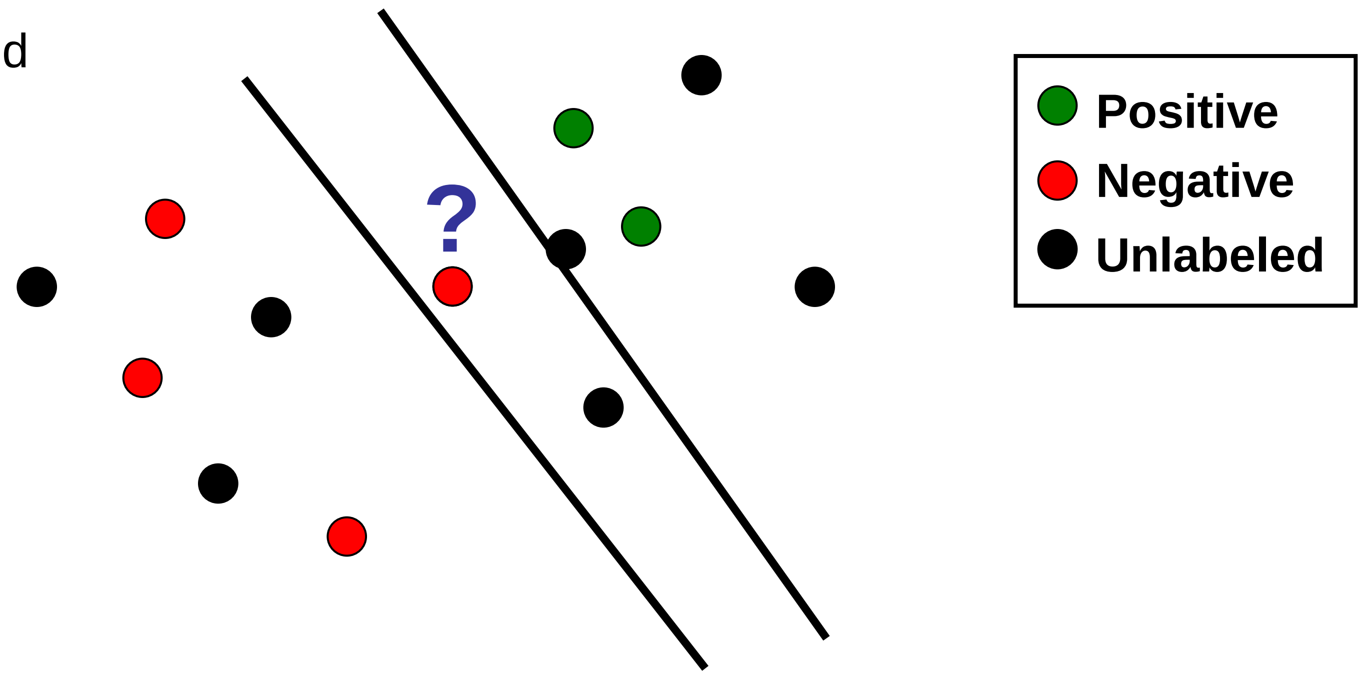


Intent: better models, faster/cheaper

Active selection

- **Traditional active learning**: obtain most informative labels first.

e.g., margin-based
criterion

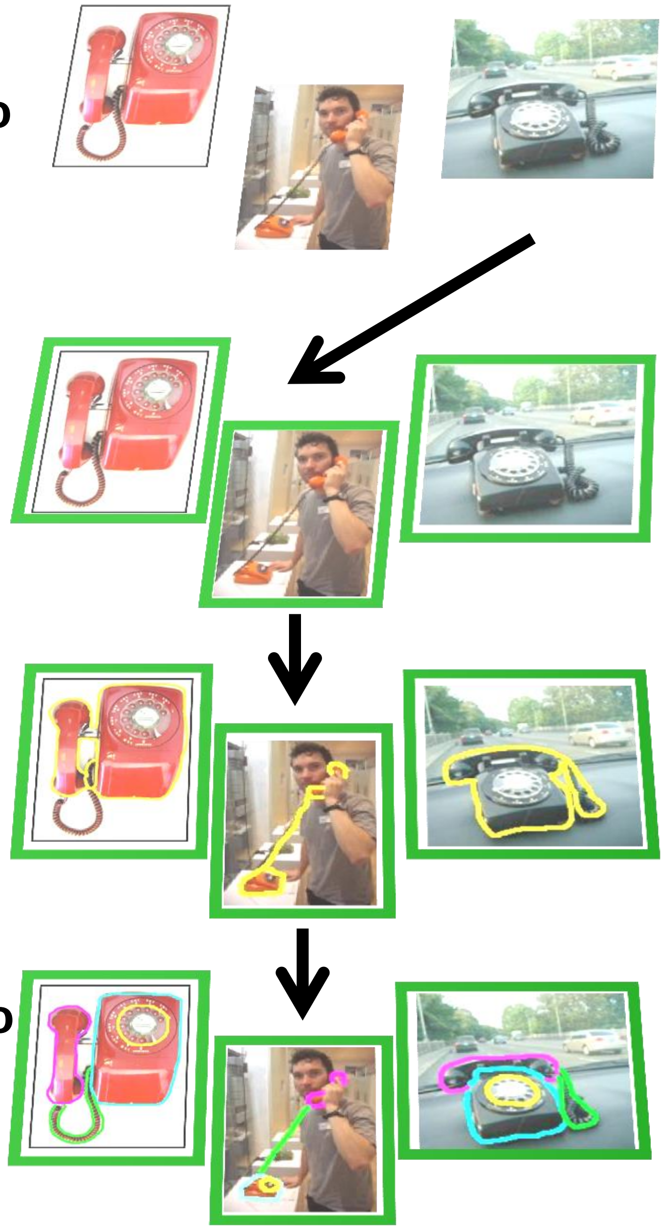


[Mackay 1992, Cohn et al. 1996, Freund et al. 1997, Lindenbaum et al. 1999, Tong & Koller 2000, Schohn and Cohn 2000, Campbell et al. 2000, Roy & McCallum 2001, Kapoor et al. 2007,…]

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
- **Many annotators** working simultaneously

Kristen Grauman, UT-Austin

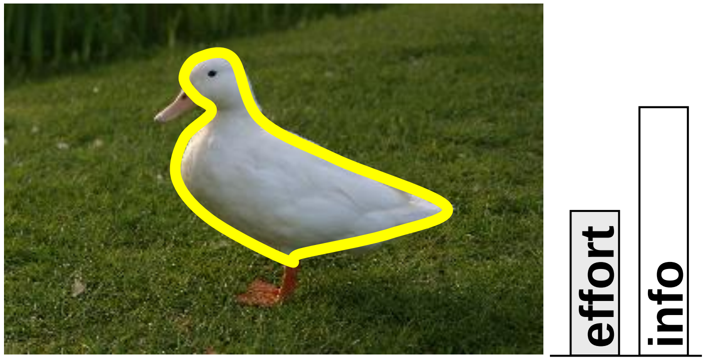
Our idea: Cost-sensitive multi-question active learning

- 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]

Our idea: Cost-sensitive multi-question active learning



Most regions are understood, but this region is unclear.

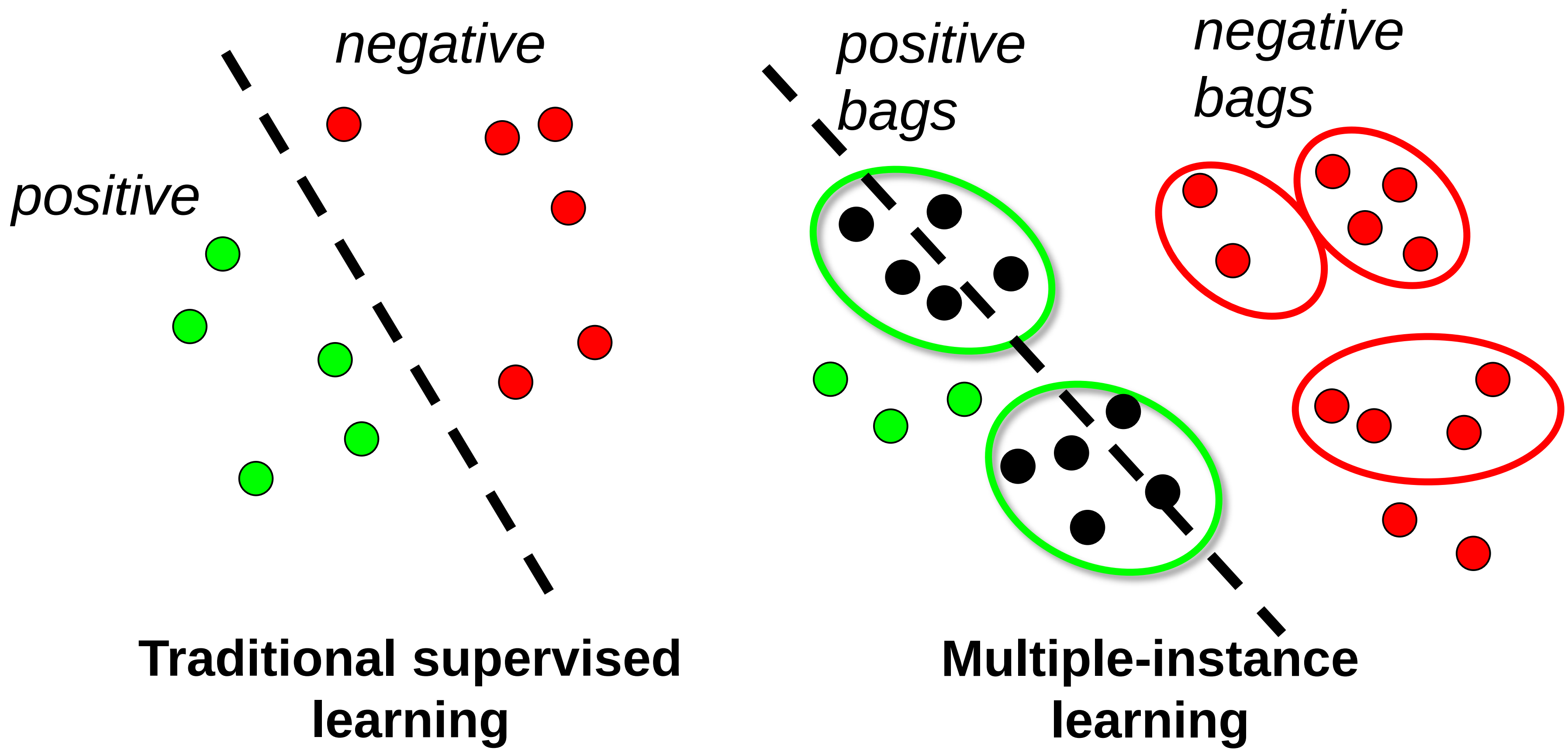
This looks expensive to annotate, and it does not seem informative.

This looks expensive to annotate, but it seems very informative.

This looks easy to annotate, but its content is already understood.

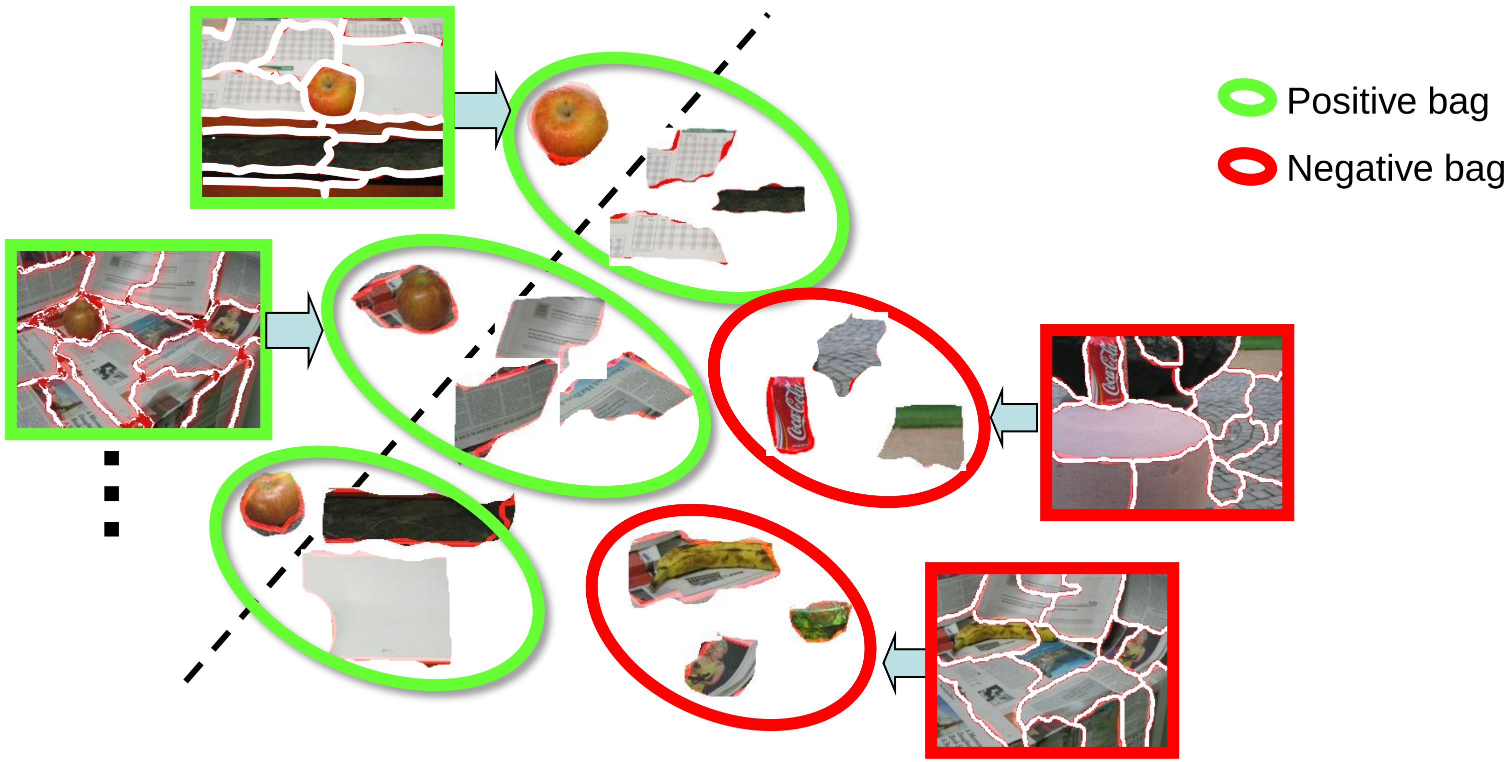
Kristen Grauman, UT-Austin

Multiple-instance learning (MIL)



[Dietterich et al. 1997]

Multiple-instance learning (MIL)



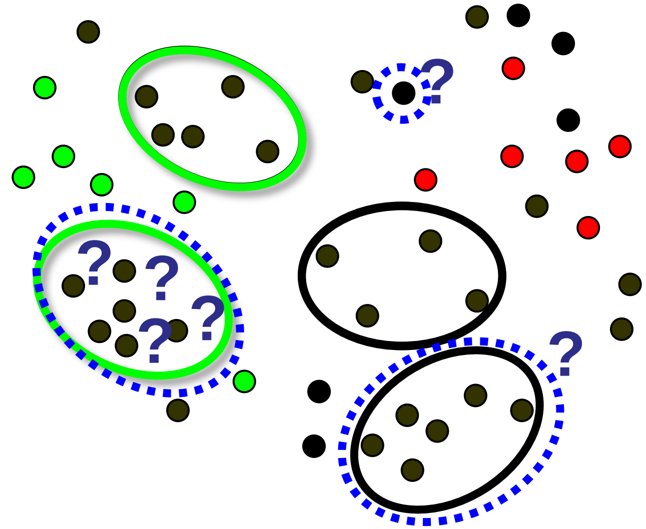
- **Positive instance**: Segment belonging to class
- **Negative instance**: Segment not in class
- **Positive bag**: Image containing class
- **Negative bag**: Image not containing class

[Dietterich et al.; Maron & Ratan, Yang & Lozano-Perez, Andrews et al.,…]

Kristen Grauman, UT-Austin

Multi-question active queries

- Predict which query will be most informative, given the cost of obtaining the annotation.
- Three levels (types) to choose from:



1. Label a region

2. Tag an object in the image

3. Segment the image, name all objects.

Kristen Grauman, UT-Austin

Decision-theoretic multi-question criterion

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

Estimate risk of incorporating the candidate before obtaining true answer A by computing expected value:

$$\widehat{\textsc{Risk}}(\mathcal{X}_L \cup O_A, \mathcal{X}_U \setminus O) = \sum_{\ell \in \mathbb{L}} \textsc{Risk}(\mathcal{X}_L \cup O_\ell, \mathcal{X}_U \setminus O)\ \ p(\ell|O)$$

where $\mathbb{L}$ is set of all possible answers.

For *M* regions $O = \{o_1, \ldots, o_M\}$

$$\approx \frac{1}{S} \sum_{k=1}^{S} \textsc{Risk}\left(\mathcal{X}_L \cup \{o_1^{(a_1)_k}, \ldots, o_M^{(a_M)_k}\}, \mathcal{X}_U \setminus O\right)$$

Kristen Grauman, UT-Austin

Decision-theoretic multi-question criterion

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

Estimate risk of incorporating the candidate before obtaining true answer A by computing expected value:

$$\widehat{\text{RISK}}(\mathcal{X}_L \cup O_A, \mathcal{X}_U \setminus O) = \sum_{\ell \in \mathbb{L}} \text{RISK}(\mathcal{X}_L \cup O_\ell, \mathcal{X}_U \setminus O) \;\; p(\ell|O)$$

where $\mathbb{L}$ is set of all possible answers.

Cost of the answer: domain knowledge, or directly predict.

Predicting effort

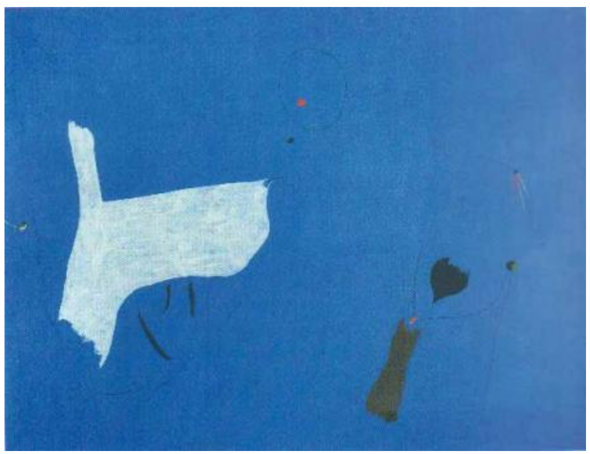
- What manual effort cost would we expect to pay for an unlabeled image?

Which image would you rather annotate?

Kristen Grauman, UT-Austin

Predicting effort

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

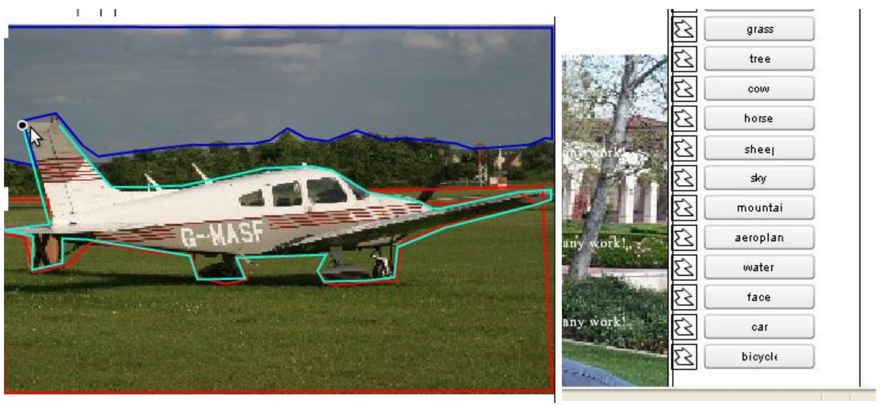


Which image would you rather annotate?

Other forms of annotation cost: expertise required, resolution of data, length of video clips,…

Learning from annotation examples

Extract cost-indicative image features, train regressor to map features to times.



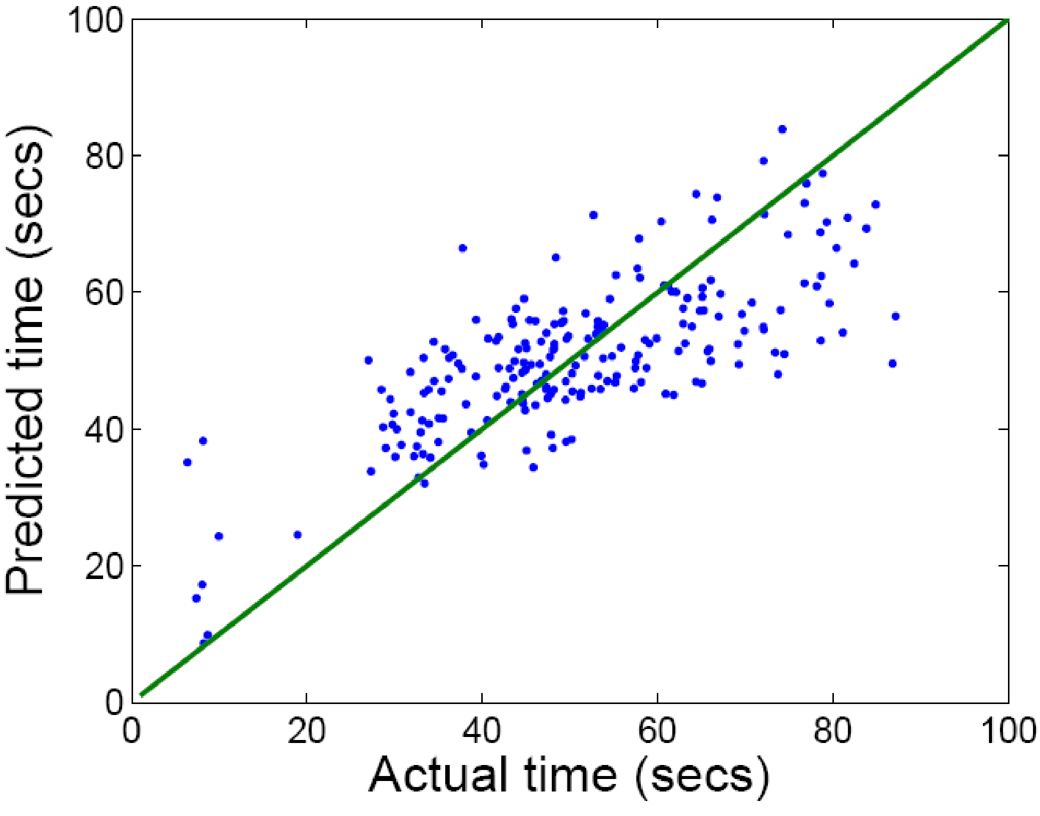
Interface on Mechanical Turk

…

…
32 s
24 s
48 s

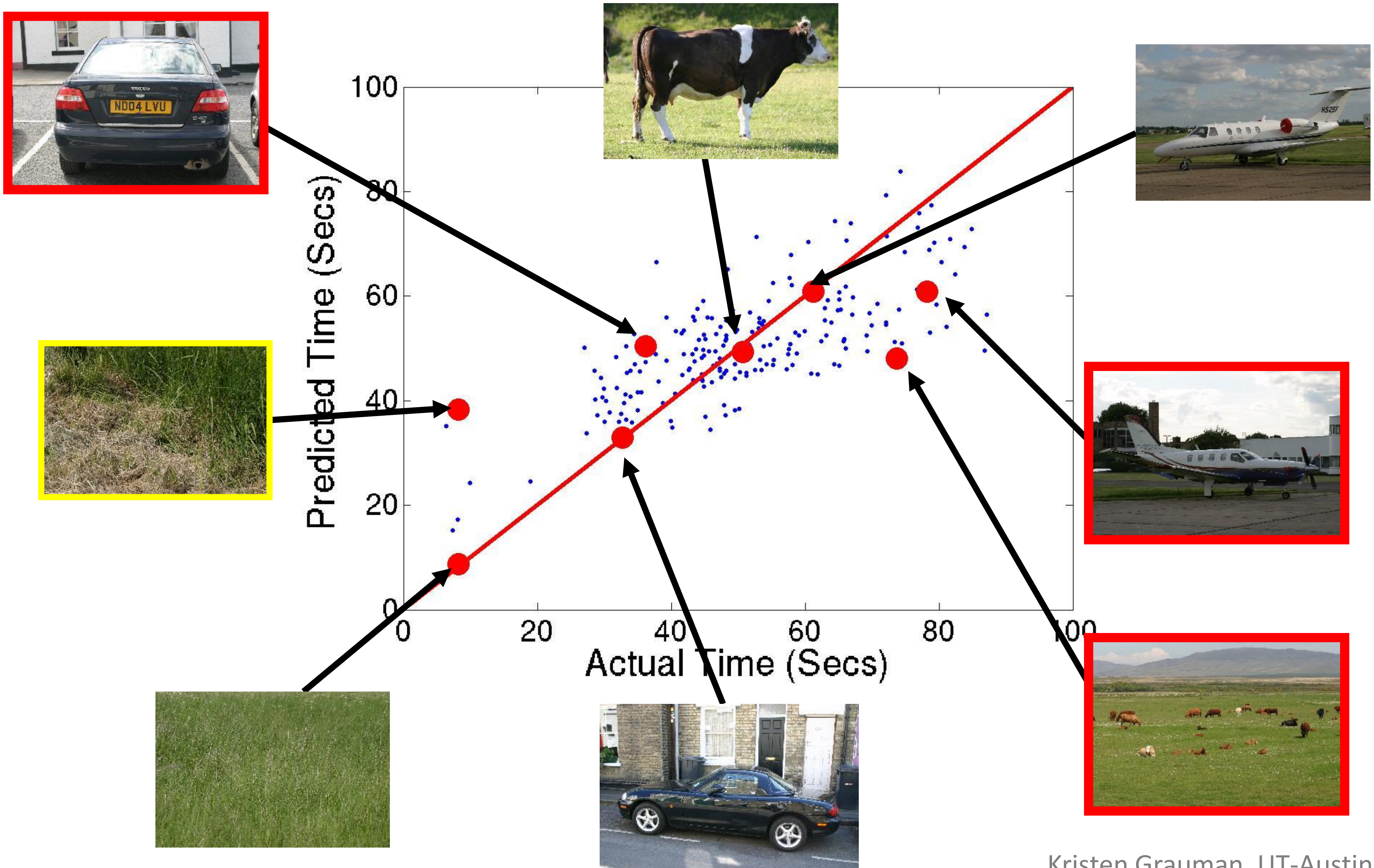
Collect about 50 responses per training image.

Predicting effort



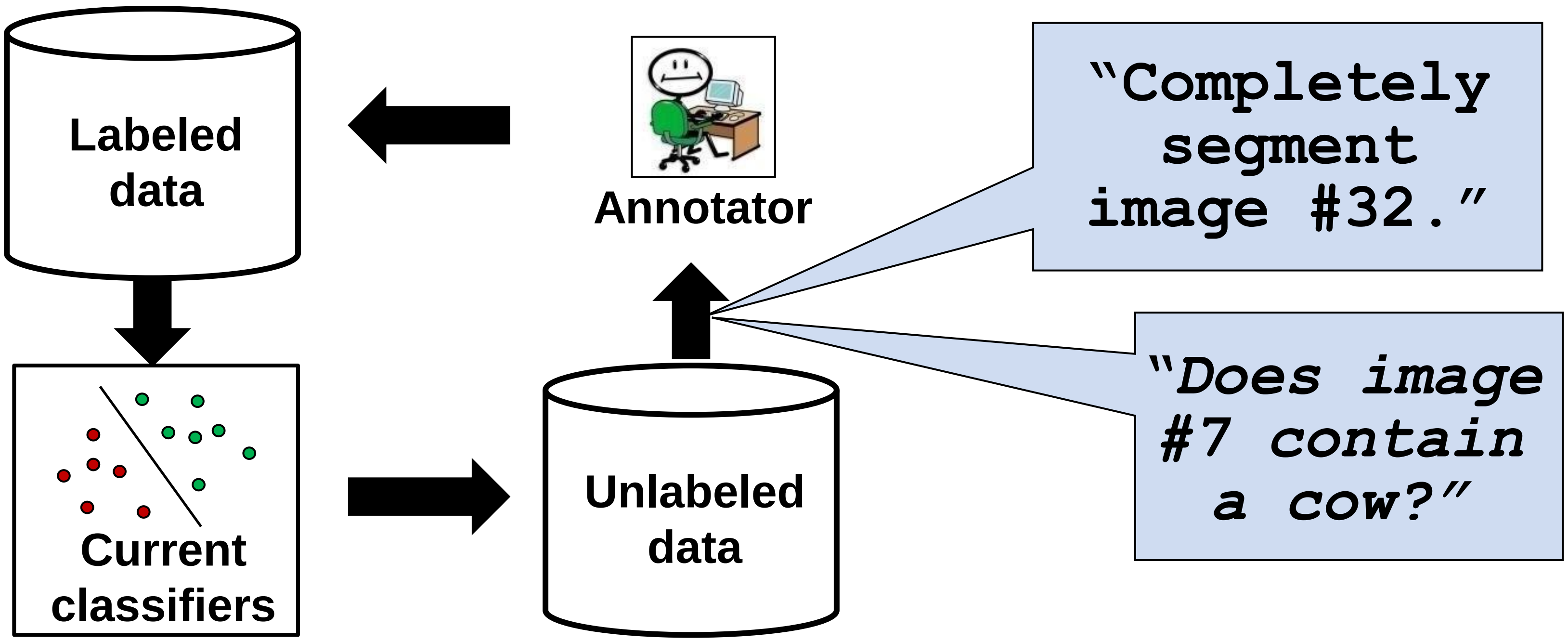
Kristen Grauman, UT-Austin

Predicting effort



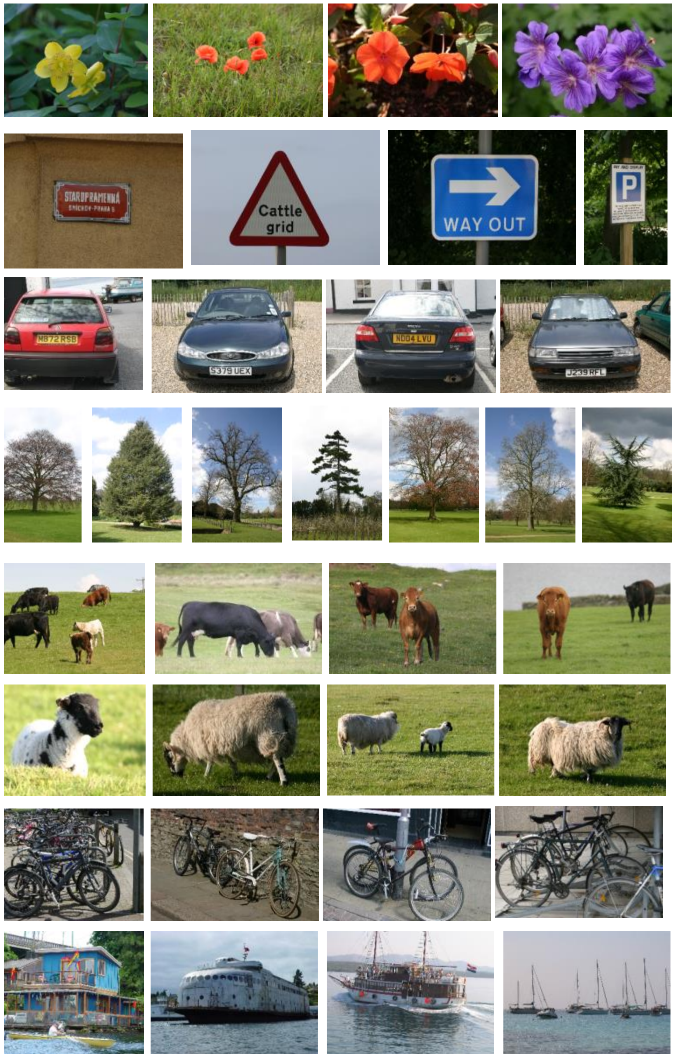
Kristen Grauman, UT-Austin

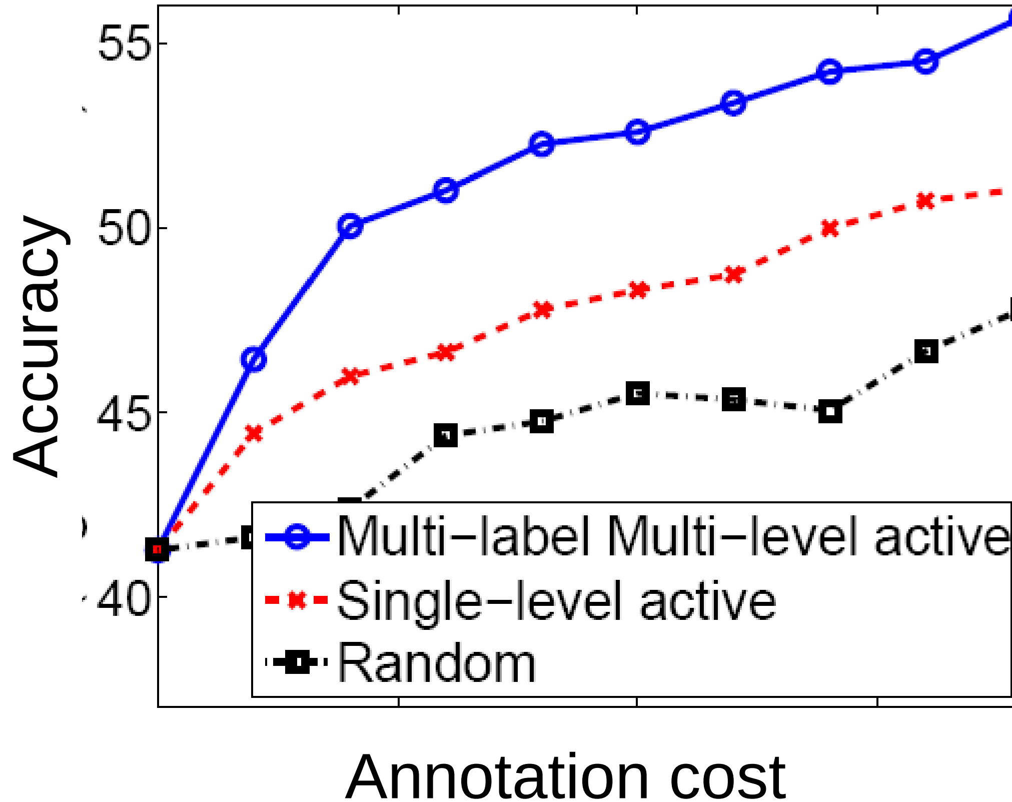
Multi-question active learning



Kristen Grauman, UT-Austin

Multi-question active learning curves

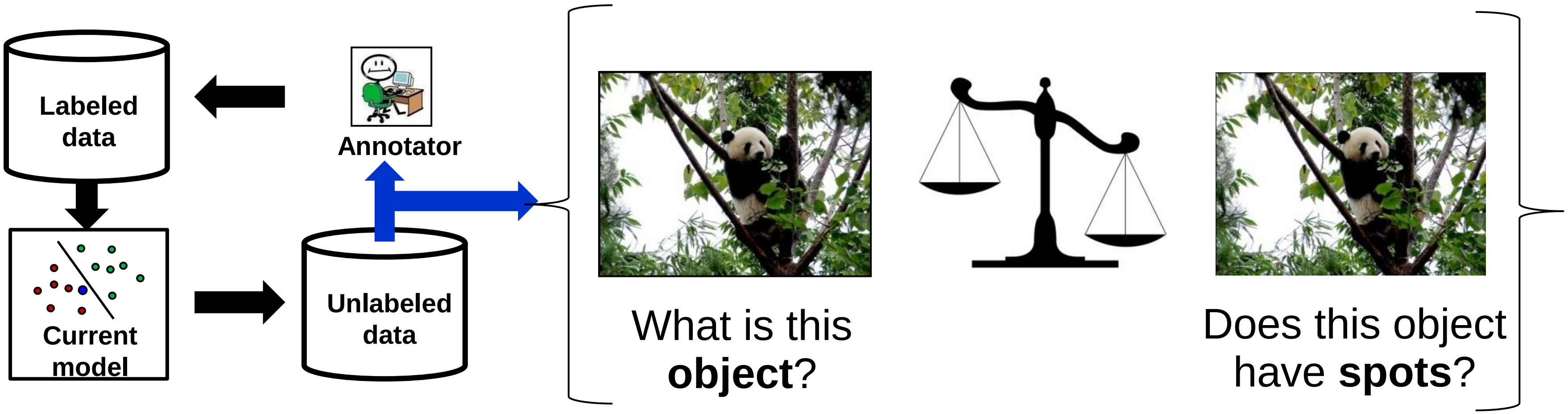




Region features: texture and color

Multi-question active learning with objects and attributes

[Kovashka et al., ICCV 2011]

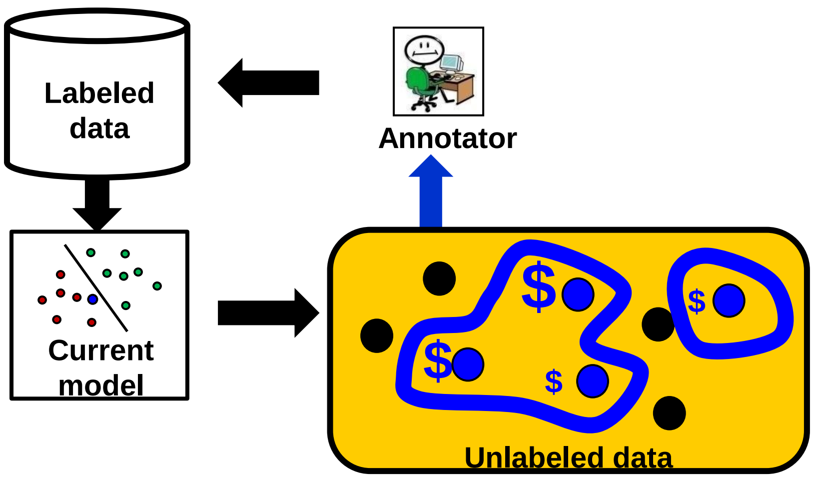


Weigh relative impact of an object label or an attribute label, at each iteration.

Kristen Grauman, UT-Austin

Budgeted batch active learning

[Vijayanarasimhan et al., CVPR 2010]



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

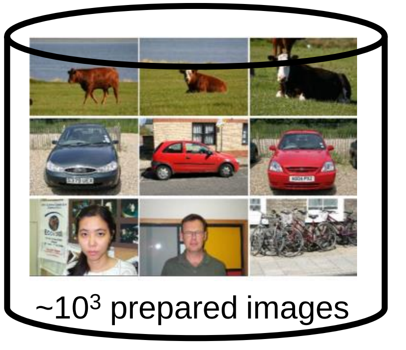
$$s.t. \sum_{x \in S} \text{LabelCost}(x) \leq \text{Budget}$$

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

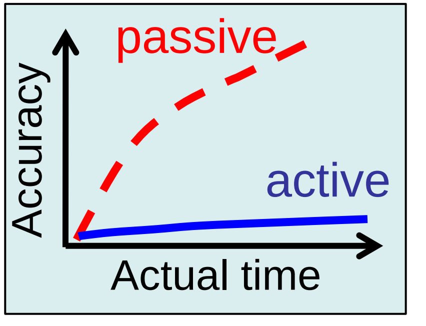
Problem: “Sandbox” active learning

Thus far, tested only in artificial settings:

- Unlabeled data already fixed, small scale, biased



- Computational cost ignored

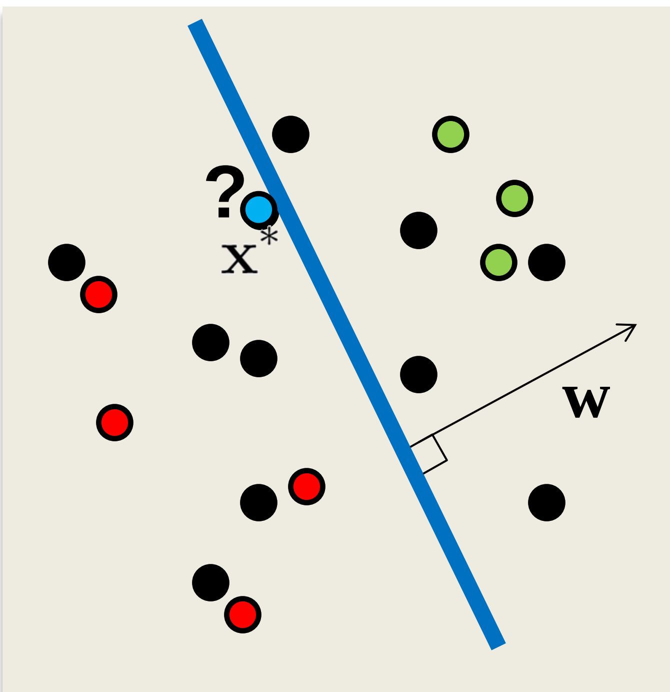


Our idea: Live active learning

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

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

SVM margin criterion for active selection



Select point nearest to hyperplane decision boundary for labeling.

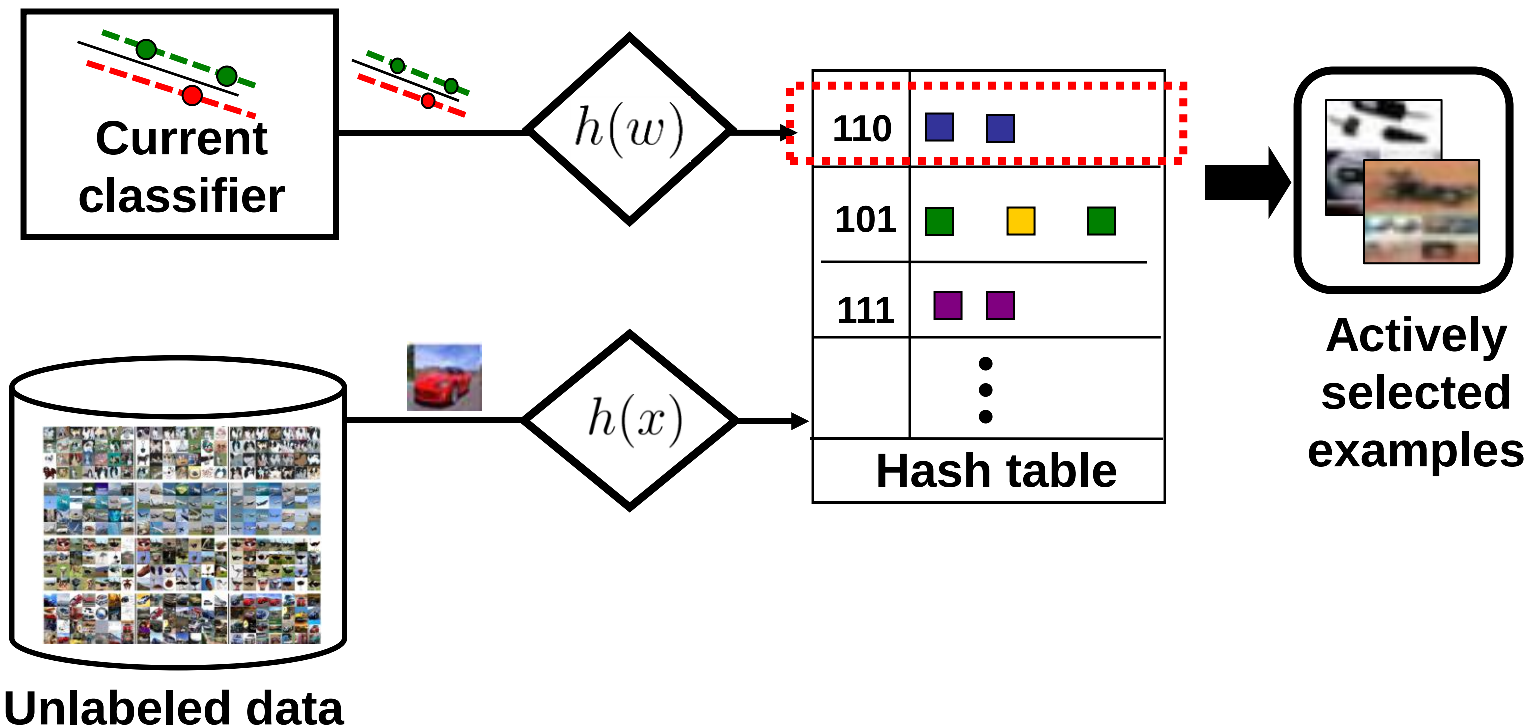
$$\mathbf{x}^* = \operatorname{argmin}_{\mathbf{x}_i \in \mathcal{U}} |\mathbf{w}^T \mathbf{x}_i|$$

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

Sub-linear time active selection

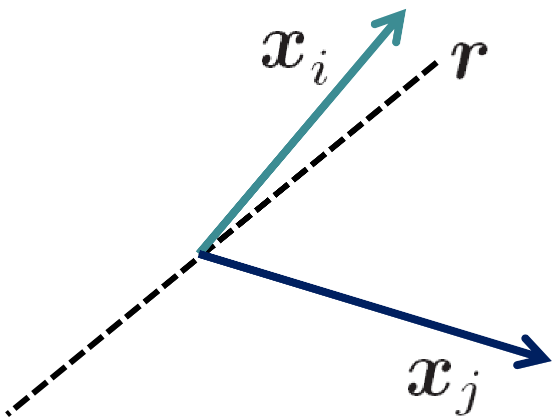
[Jain, Vijayanarasimhan, Grauman, NIPS 2010]

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



Background: Locality-Sensitive Hashing

Probability a *random hyperplane* separates two unit vectors depends on the angle between them:



Bigger angle: likely to split

Corresponding hash function:

$$h_{\boldsymbol{r}}(\boldsymbol{x}) = \begin{cases} 1, & \text{if } \boldsymbol{r}^T\boldsymbol{x} \geq 0 \\ 0, & \text{otherwise} \end{cases}$$

$$r_i \sim \mathcal{N}(0, 1)$$

Probability of collision:

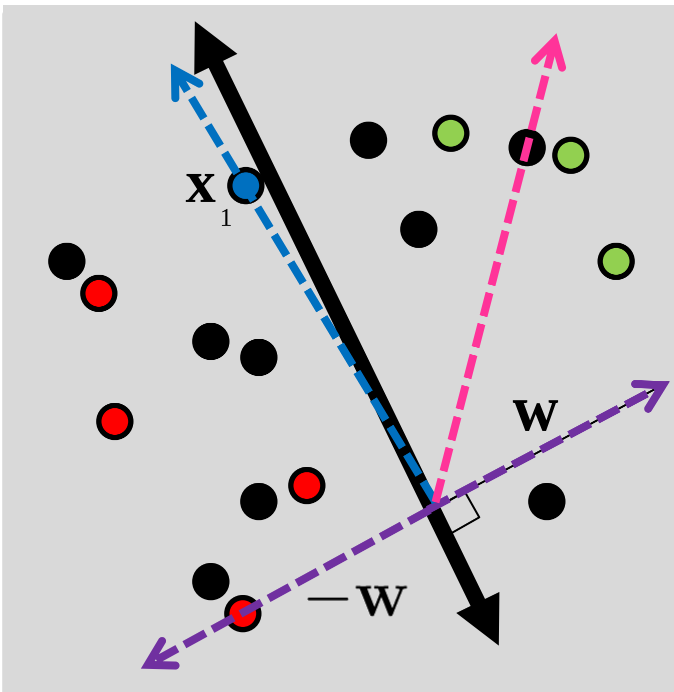
$$\Pr(h_{\boldsymbol{r}}(x_i) = h_{\boldsymbol{r}}(x_j)) = 1 - \frac{1}{\pi}\cos^{-1}(x_i^T x_j)$$

[Goemans and Williamson 1995, Charikar 2004]

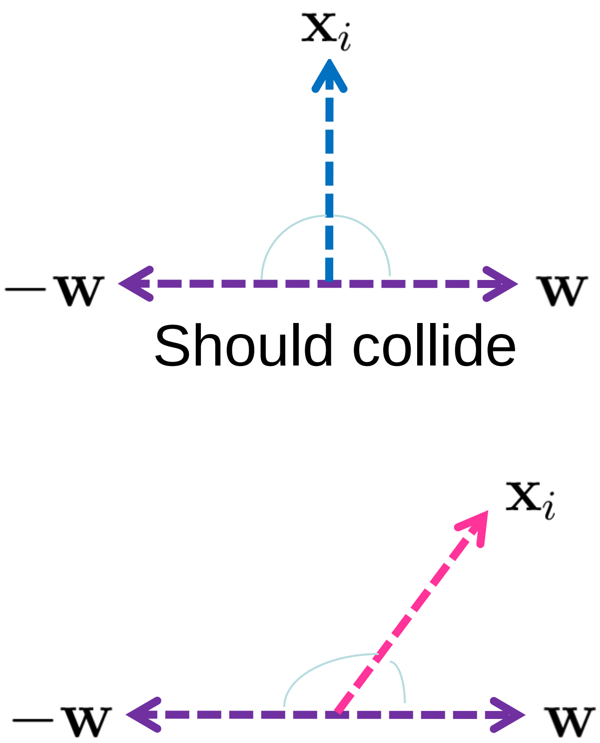
Kristen Grauman, UT-Austin

Hashing a hyperplane query

To retrieve those points for which $|\mathbf{w}^T \mathbf{x}_i|$ is small, want probable collision for **perpendicular** vectors:



Assuming normalized data.



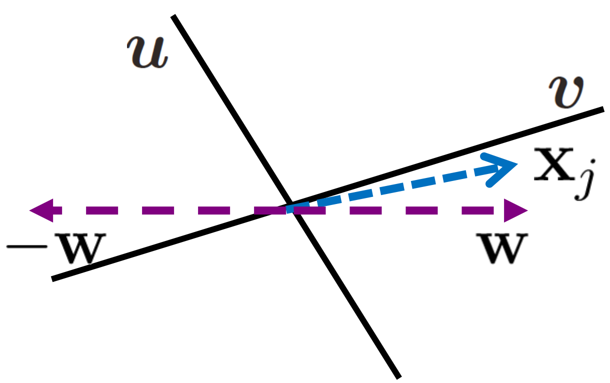
Hashing a hyperplane query

We generate two independent random vectors **u** and **v:**

- one to constrain angle between **x** and **w**
- one to constrain angle between **x** and **–w**

Collision likely only if <u>neither</u> vector splits

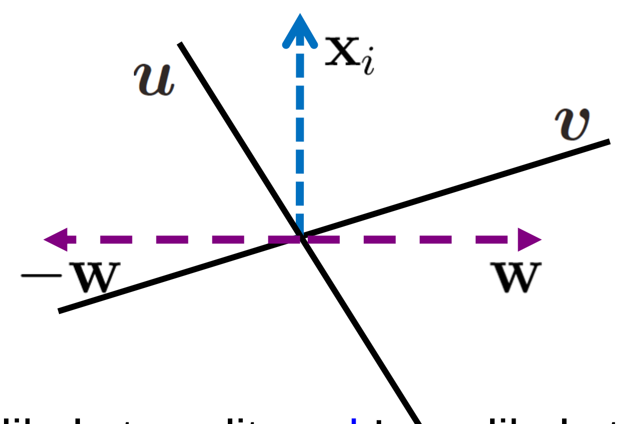
For parallel vectors



Unlikely to split and Likely to split

= Likely to split

For perpendicular vectors



Less likely to split and Less likely to split

= Unlikely to split

Kristen Grauman, UT-Austin

Hashing a hyperplane query

- We define an asymmetric 2-bit hash function:

H-Hash family:

$$h_{\mathcal{H}}(\boldsymbol{z}) = \begin{cases} h_{\boldsymbol{u},\boldsymbol{v}}(\boldsymbol{z},\boldsymbol{z}), & \text{if } \boldsymbol{z} \text{ is a database point vector,} \\ h_{\boldsymbol{u},\boldsymbol{v}}(\boldsymbol{z},-\boldsymbol{z}), & \text{if } \boldsymbol{z} \text{ is a query hyperplane vector.} \end{cases}$$

where $h_{\boldsymbol{u},\boldsymbol{v}}(\boldsymbol{a},\boldsymbol{b}) = [h_{\boldsymbol{u}}(\boldsymbol{a}), h_{\boldsymbol{v}}(\boldsymbol{b})] = [\text{sign}(\boldsymbol{u}^T\boldsymbol{a}), \text{sign}(\boldsymbol{v}^T\boldsymbol{b})]$

- We prove necessary conditions for locality sensitivity:

$$\Pr[h_{\mathcal{H}}(\boldsymbol{w}) = h_{\mathcal{H}}(\boldsymbol{x})] = \Pr[h_{\boldsymbol{u}}(\boldsymbol{w}) = h_{\boldsymbol{u}}(\boldsymbol{x})]\ \Pr[h_{\boldsymbol{v}}(-\boldsymbol{w}) = h_{\boldsymbol{v}}(\boldsymbol{x})]$$

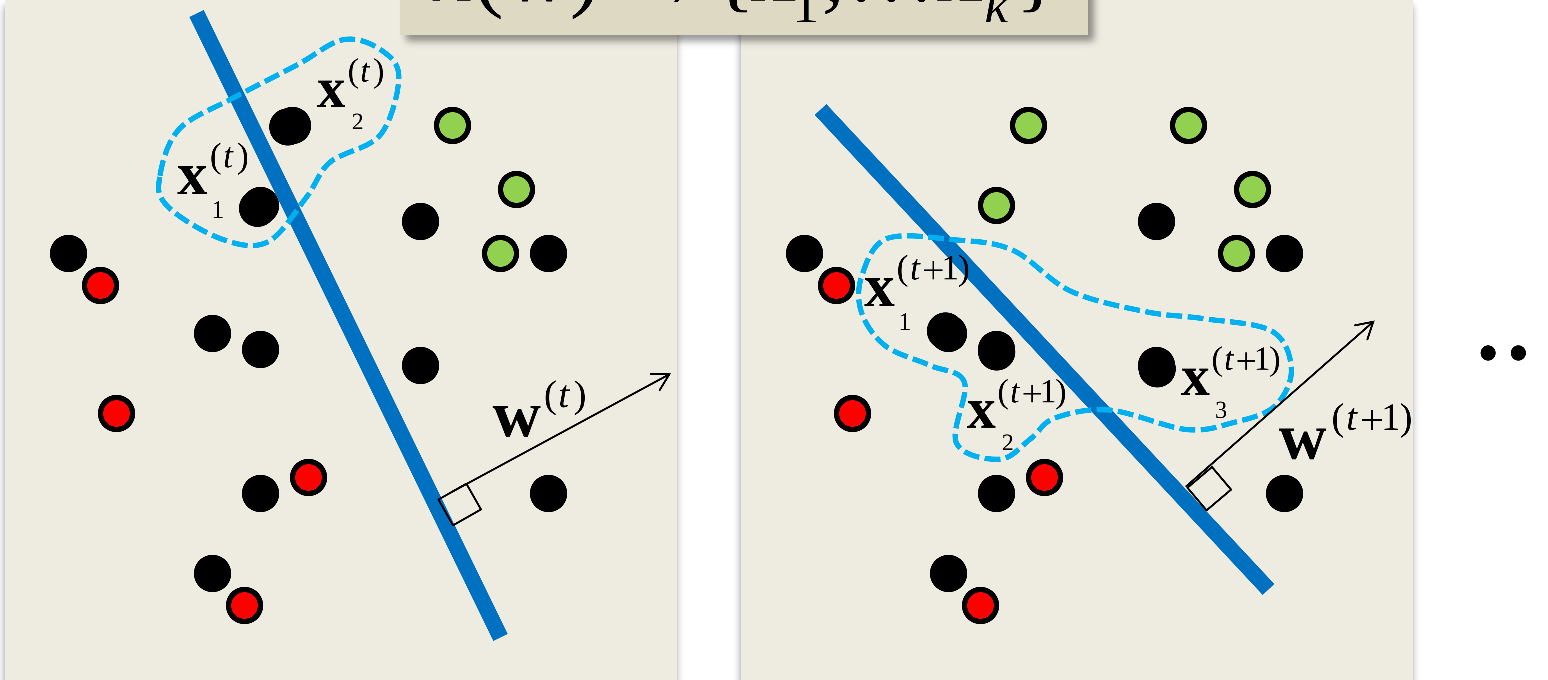
$$= \frac{1}{4} - \frac{1}{\pi^2}\left(\theta_{\boldsymbol{x},\boldsymbol{w}} - \frac{\pi}{2}\right)^2$$

[Jain, Vijayanarasimhan & Grauman, NIPS 2010]

Kristen Grauman, UT-Austin

Hashing a hyperplane query

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

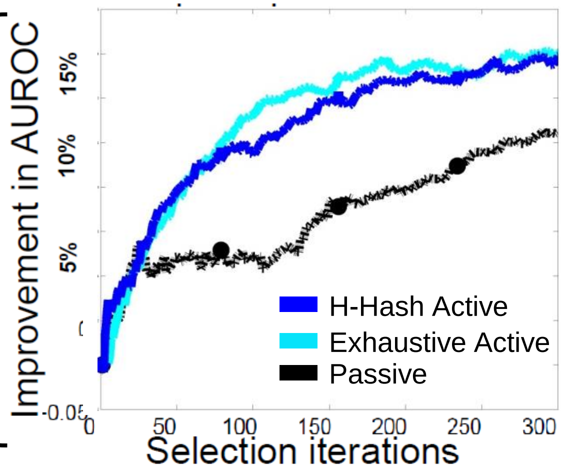


At each iteration of the learning loop, our hash functions map the current hyperplane directly to its nearest unlabeled points.

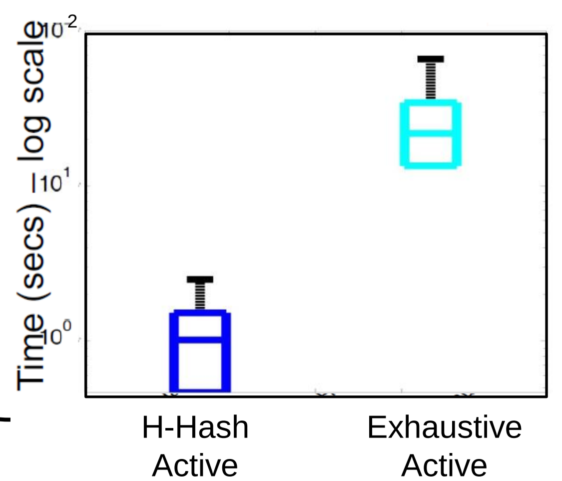
Kristen Grauman, UT-Austin

Sub-linear time active selection

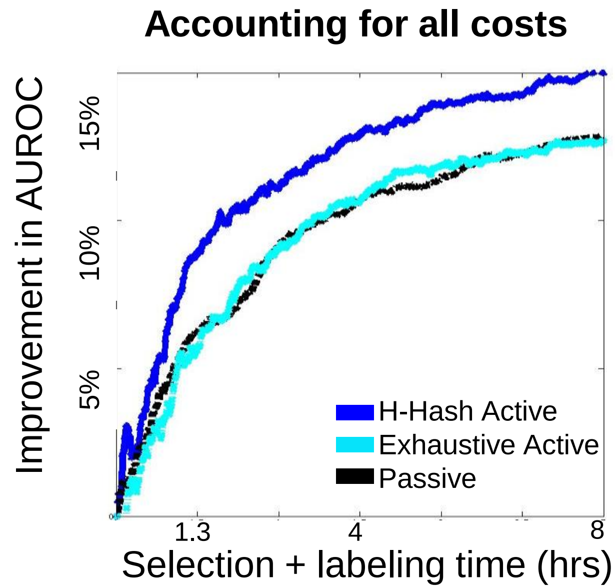
Accuracy improvements as more data labeled



Time spent searching for selection



H-Hash result on 1M Tiny Images

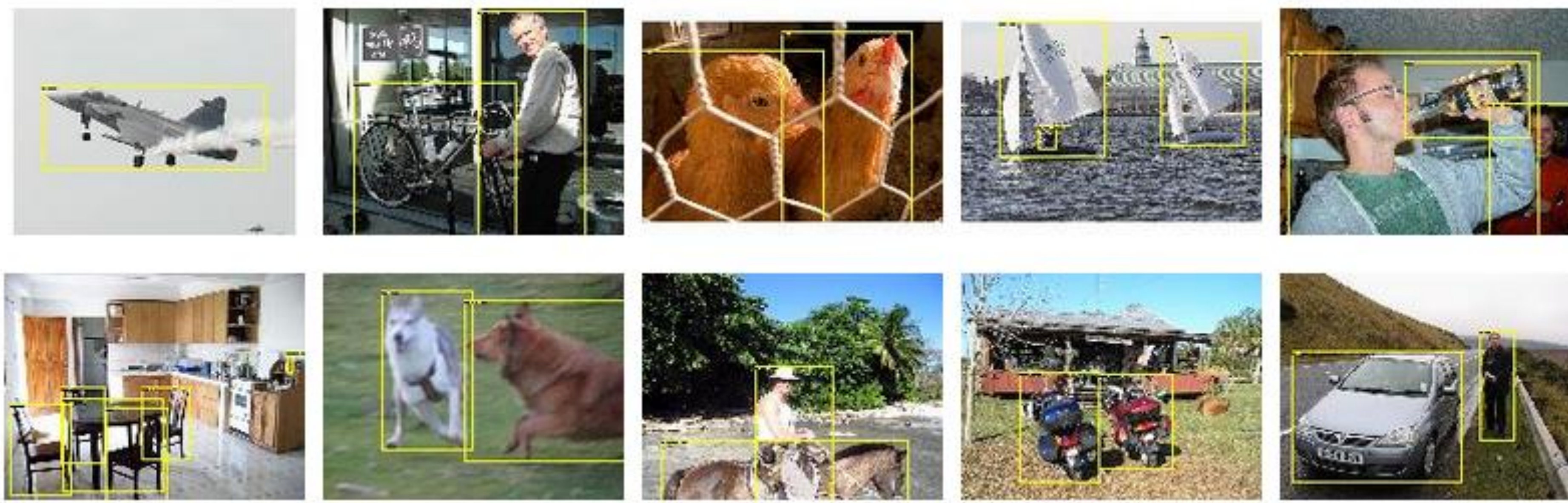


By minimizing **both** selection and labeling time, obtain the best accuracy per unit time.

Kristen Grauman, UT-Austin

PASCAL Visual Object Categorization

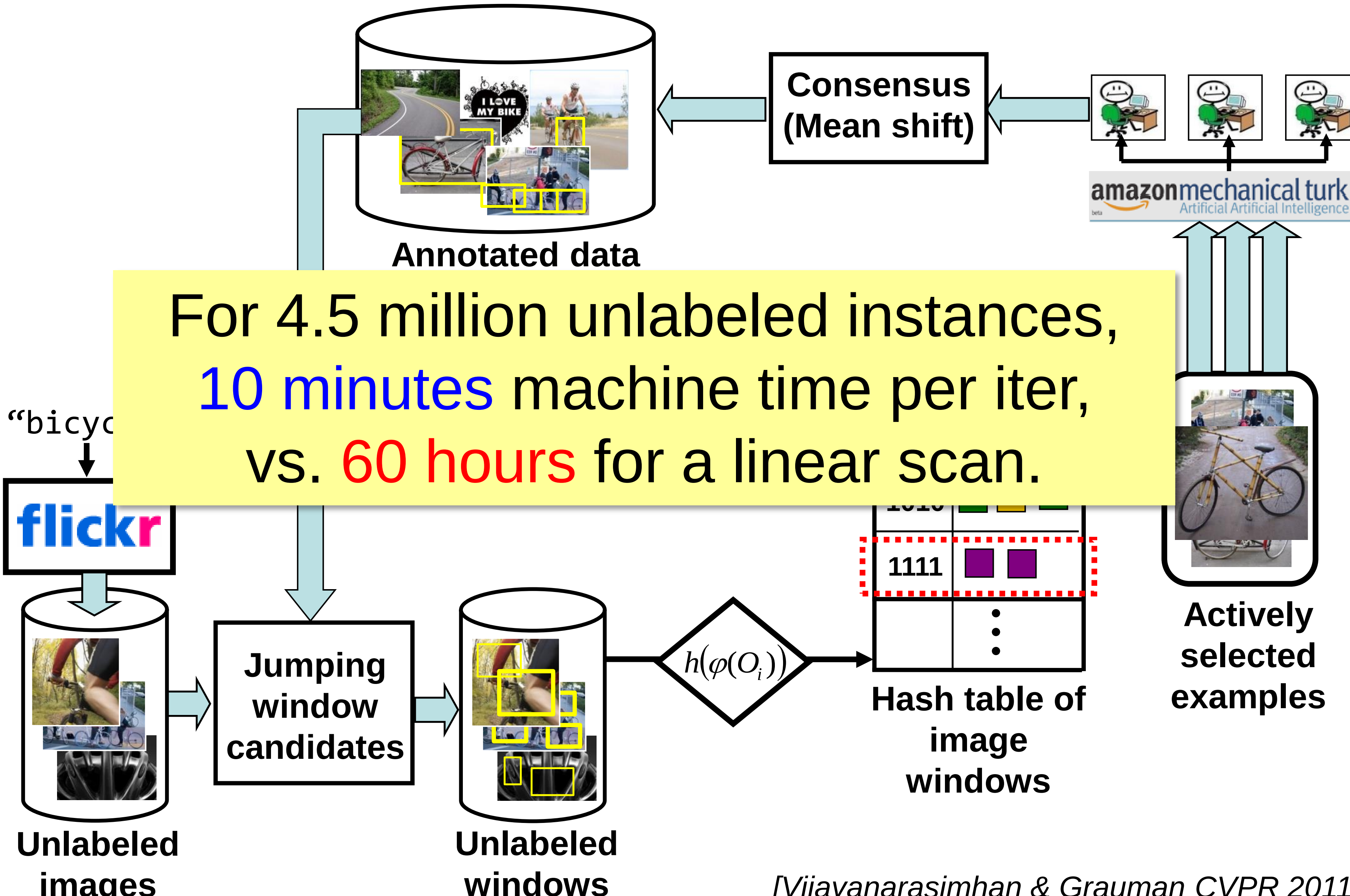
- Closely studied object detection benchmark
- Original image data from Flickr



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

Kristen Grauman, UT-Austin

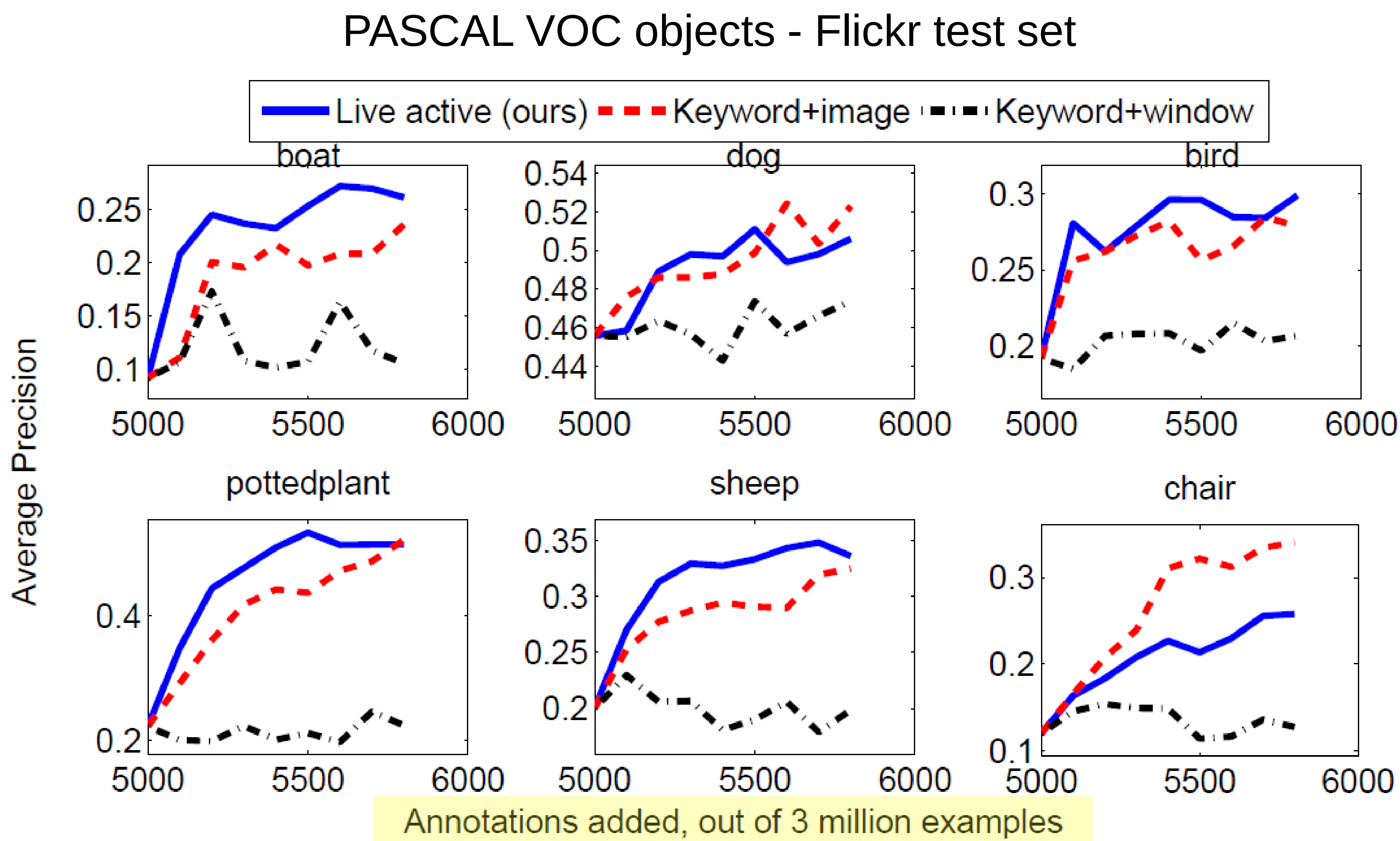
Live active learning



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”

Kristen Grauman, UT-Austin

PASCAL Live active learning results

Live learning improves some of most difficult PASCAL VOC categories:

	bird	boat	dog	potted plant	sheep	chair
Ours	**15.8****	**18.9****	**25.3****	11.6*	**28.4****	9.1*
Previous best	15.3	16.8	21.5	**14.6**	23.9	**17.9**

Our approach's efficiency makes live learning feasible

	Active selection	Training	Detection per image
Ours + active	10 mins	5 mins	150 secs
LSVM [Felzenszwalb et al. 2009]	3 hours	4 hours	2 secs
SP+MKL [Vedaldi et al. 2009]	93 hours	> 2 days	67 secs

Previous best : [Vedaldi et al. ICCV 2009] or [Felzenszwalb et al. PAMI 2009]

Summary so far

Actively eliciting human insight for visual recognition algorithms.

- Multi-question active learning to formulate annotation requests that specify the example *and* the task.
- Budgeted batch selection for effective joint selection of multiple requests suited for online annotators.
- Live active learning shows large-scale practical impact.

Ongoing challenges in active visual learning

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

This lecture

Teaching machines visual categories

- Active learning to prioritize informative annotations
- Relative attributes to learn 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, …]

Kristen Grauman, UT-Austin


Mule

Attributes

A mule…

Is furry

Has four-legs

Legs shorter than horses’

Tail longer than donkeys’

Has tail

Binary attributes

A mule…

Is furry

Has four-legs

Legs shorter than horses'

Tail longer than donkeys'

Has 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

Legs **shorter** than horses'

Tail **longer** than donkeys'

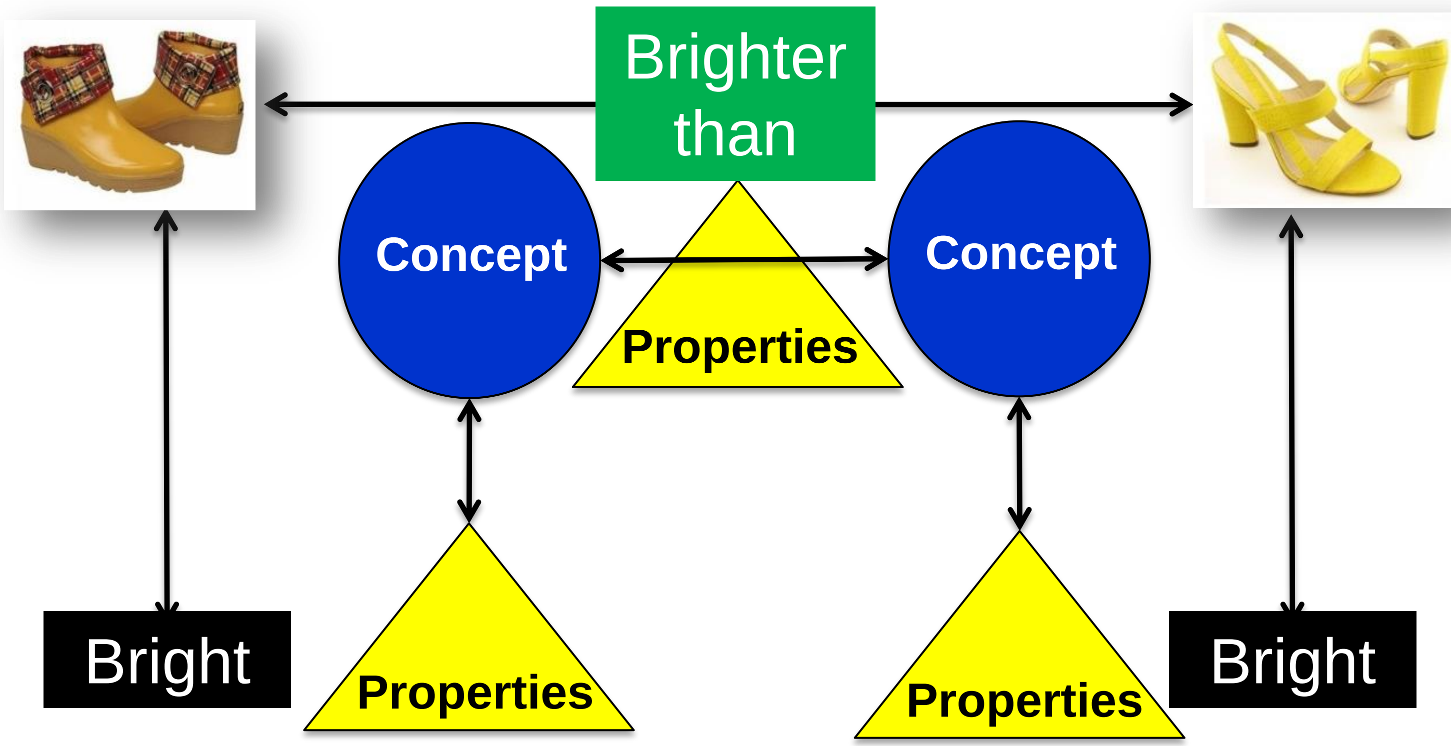
Has tail

Kristen Grauman, UT-Austin

Relative attributes

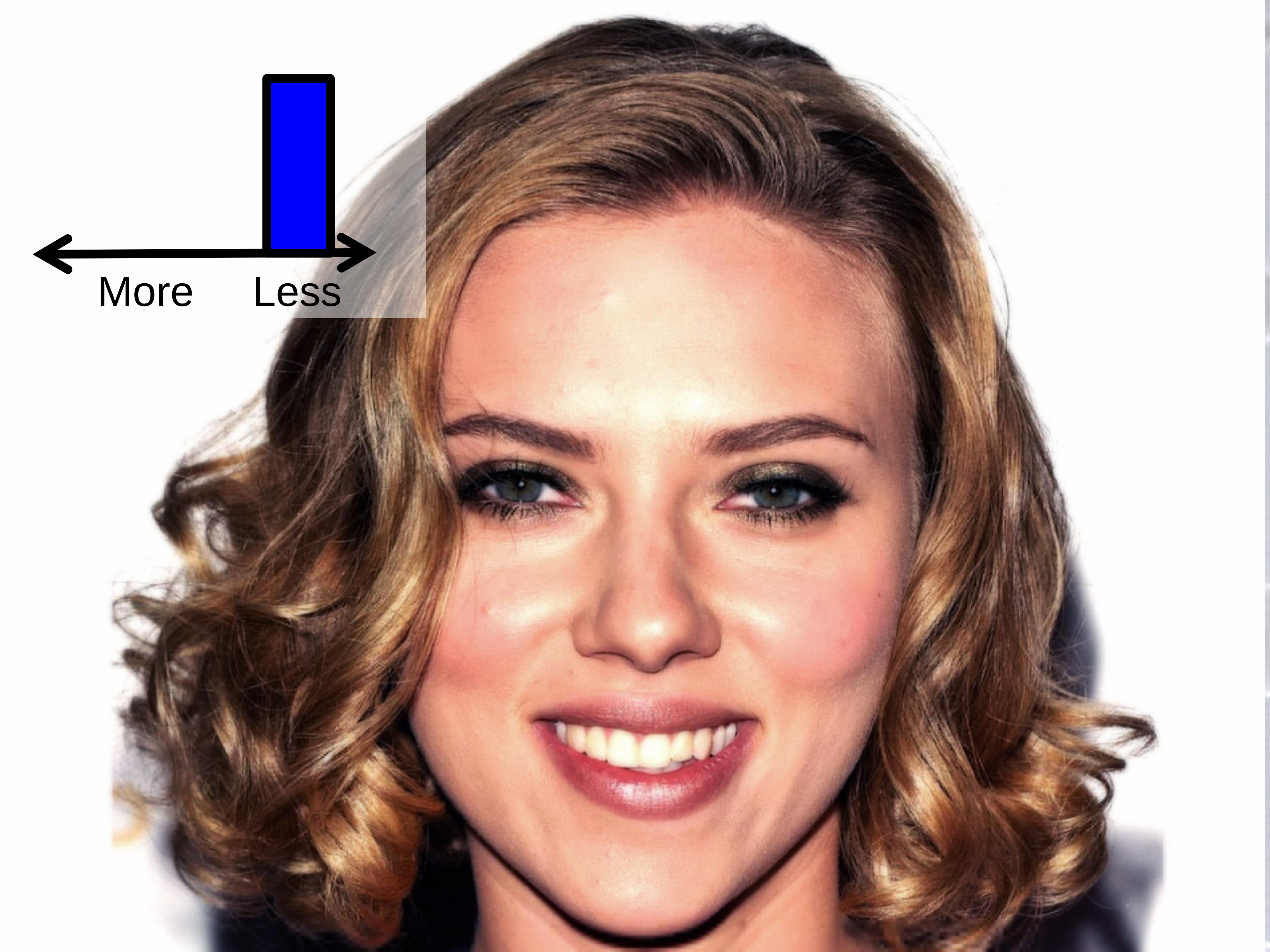
[Parikh & Grauman, ICCV 2011]

- Represent ***visual comparisons*** between classes, images, and their properties.



How should relative attributes be learned?

What do we need to capture from human annotators?


More
Less

Learning relative attributes

- Learn a ranking function for each attribute, e.g. "brightness".
- Supervision consists of:

O_m: $\Big\{\Big($  $\succ$  $\Big),$ $\Big($  $\succ$  $\Big), \ldots \Big\},$ *Ordered pairs*

E_m: $\Big\{\Big($  $\sim$  $\Big),$ $\Big($  $\sim$  $\Big), \ldots \Big\}$ *Similar pairs*

Parikh and Grauman, ICCV 2011

Kristen Grauman, UT-Austin

Learning relative attributes

Learn a ranking function

$$a_m(\boldsymbol{x_i}) = \boldsymbol{w_m^T x_i}$$

Image features

Learned parameters

that best satisfies the constraints:

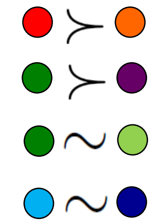
$$\forall (i,j) \in O_m : \boldsymbol{w_m^T x_i} > \boldsymbol{w_m^T x_j}$$

$$\forall (i,j) \in E_m : \boldsymbol{w_m^T x_i} = \boldsymbol{w_m^T x_j}$$

Parikh and Grauman, ICCV 2011

Kristen Grauman, UT-Austin

Learning relative attributes



Max-margin learning to rank formulation

$$\min \quad \left(\frac{1}{2}||\boldsymbol{w}_{\boldsymbol{m}}^T||_2^2 + C\left(\sum \xi_{ij}^2 + \sum \gamma_{ij}^2\right)\right)$$

$$s.t. \quad \boldsymbol{w}_{\boldsymbol{m}}^T(\boldsymbol{x_i} - \boldsymbol{x_j}) \geq 1 - \xi_{ij}$$

$$|\boldsymbol{w}_{\boldsymbol{m}}^T(\boldsymbol{x_i} - \boldsymbol{x_j})| \leq \gamma_{ij}$$

$$\xi_{ij} \geq 0; \gamma_{ij} \geq 0$$

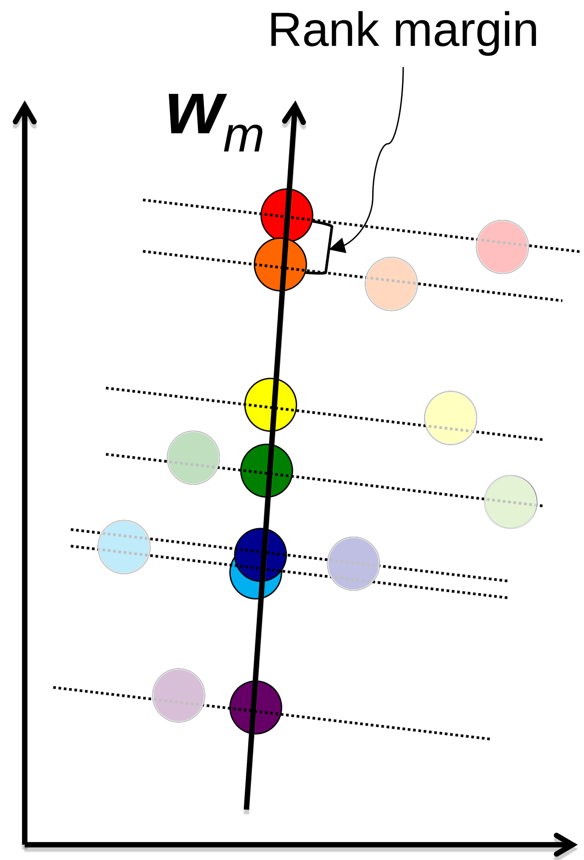


Image $\longrightarrow$ Relative attribute score

Joachims, KDD 2002; Parikh and Grauman, ICCV 2011

Kristen Grauman, UT-Austin

Relating images

- We can rank images according to attribute strength

Kristen Grauman, UT-Austin

Relating images

Density

Novel image

Conventional binary description: *not dense*

Kristen Grauman, UT-Austin

Relating images

Density

Novel image

more dense than

less dense than

Kristen Grauman, UT-Austin

Relating images

more dense than ***Highways***,
less dense than ***Forests***

Relating images

Multi-attribute descriptions offer greater precision when they are relative

Binary (existing):

Not Young

BushyEyebrows

RoundFace

Relative (ours):

More Young than CliveOwen
Less Young than ScarlettJohansson

More BushyEyebrows than ZacEfron
Less BushyEyebrows than AlexRodriguez

More RoundFace than CliveOwen
Less RoundFace than ZacEfron

Applications of relative attributes

Enable new modes of human-system communication

- **Training category models through descriptions**:

  "Rabbits are furrier than dogs."

- **Rationales to explain image labels:**

  "It's not a coastal scene because it's too cluttered."

- **Semantic relative feedback for image search:**

  "I want shoes like these, but shinier."

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

Kristen Grauman, UT-Austin

Relative zero-shot learning

We can predict new classes based on their **relationships** to existing classes – even without training images.

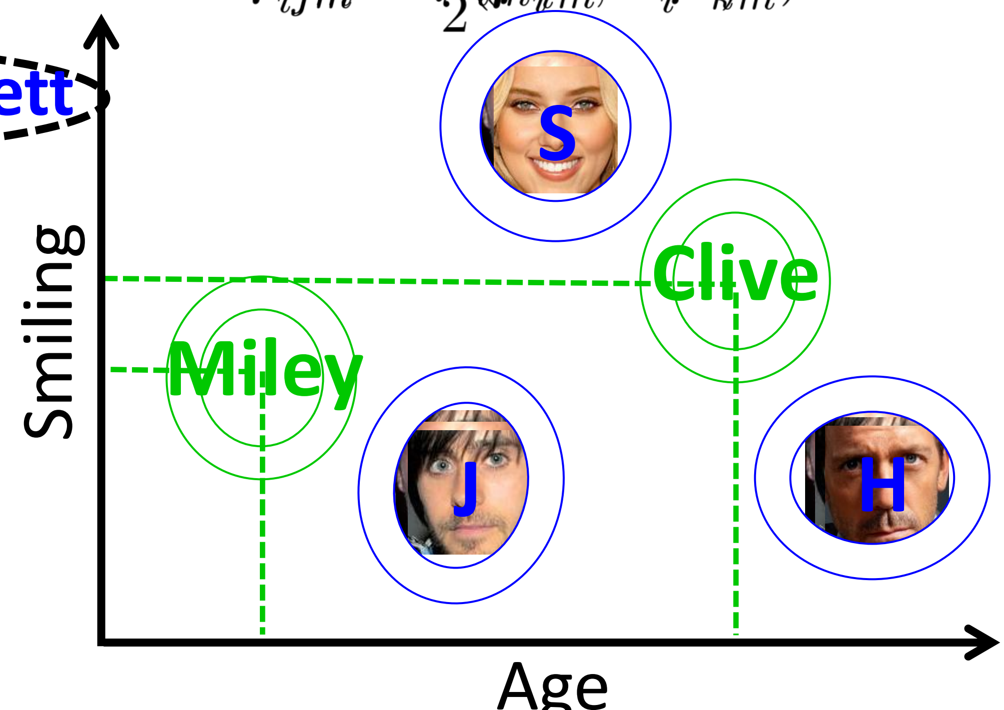
$$\mu_{jm}^{(u)} = \frac{1}{2}\left(\mu_{im}^{(s)} + \mu_{km}^{(s)}\right)$$

$$\mu_{jm}^{(s)} \sim \mathcal{N}\left(\mu_{jm}^{(s)}, \Sigma_m^{(s)}\right)$$

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

Jared $\succ$ Miley

Smiling: Miley $\succ$ Jared



Infer image category using max-likelihood

Kristen Grauman, UT-Austin

Datasets

Outdoor Scene Recognition (OSR) [Oliva 2001]

8 classes, ~2700 images, Gist

6 attributes: open, natural, etc.

Public Figures Faces (PubFig) [Kumar 2009]

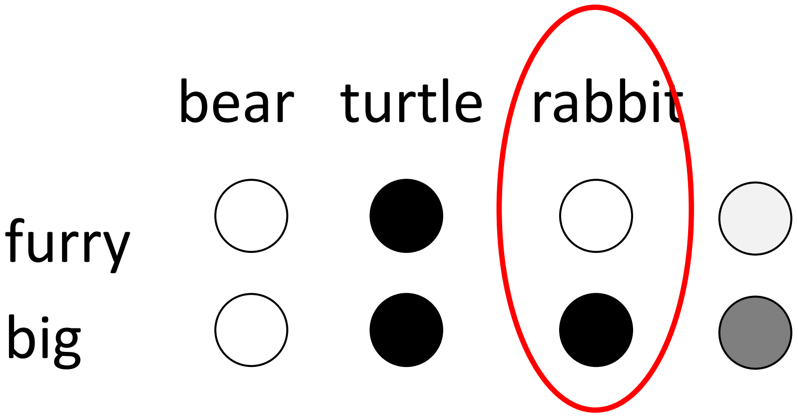
8 classes, ~800 images, Gist+color

11 attributes: white, chubby, etc.

Kristen Grauman, UT-Austin

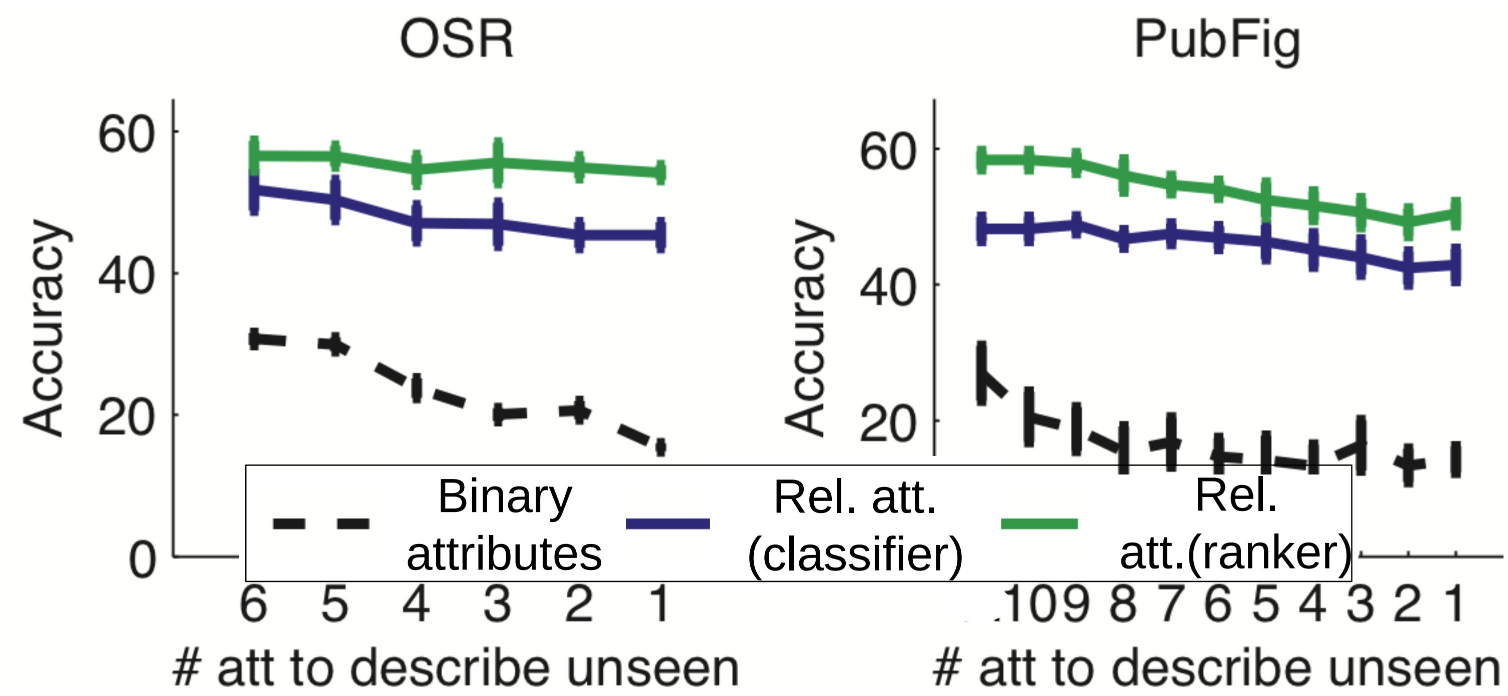
Baselines

- Binary attributes:
  Direct Attribute Prediction
  [Lampert et al. 2009]



- Relative attributes via
  classifier scores

Relative zero-shot learning



An attribute is more discriminative when used relatively

Kristen Grauman, UT-Austin

Bootstrapped scene learning with relative attribute constraints

[Gupta et al. ECCV 2012]

Semantic supervision:

Is More Open

Amphitheatre > Barn

Amphitheatre > Conference Room

Desert > Barn

Has Taller Structures

Church (Outdoor) > Cemetery

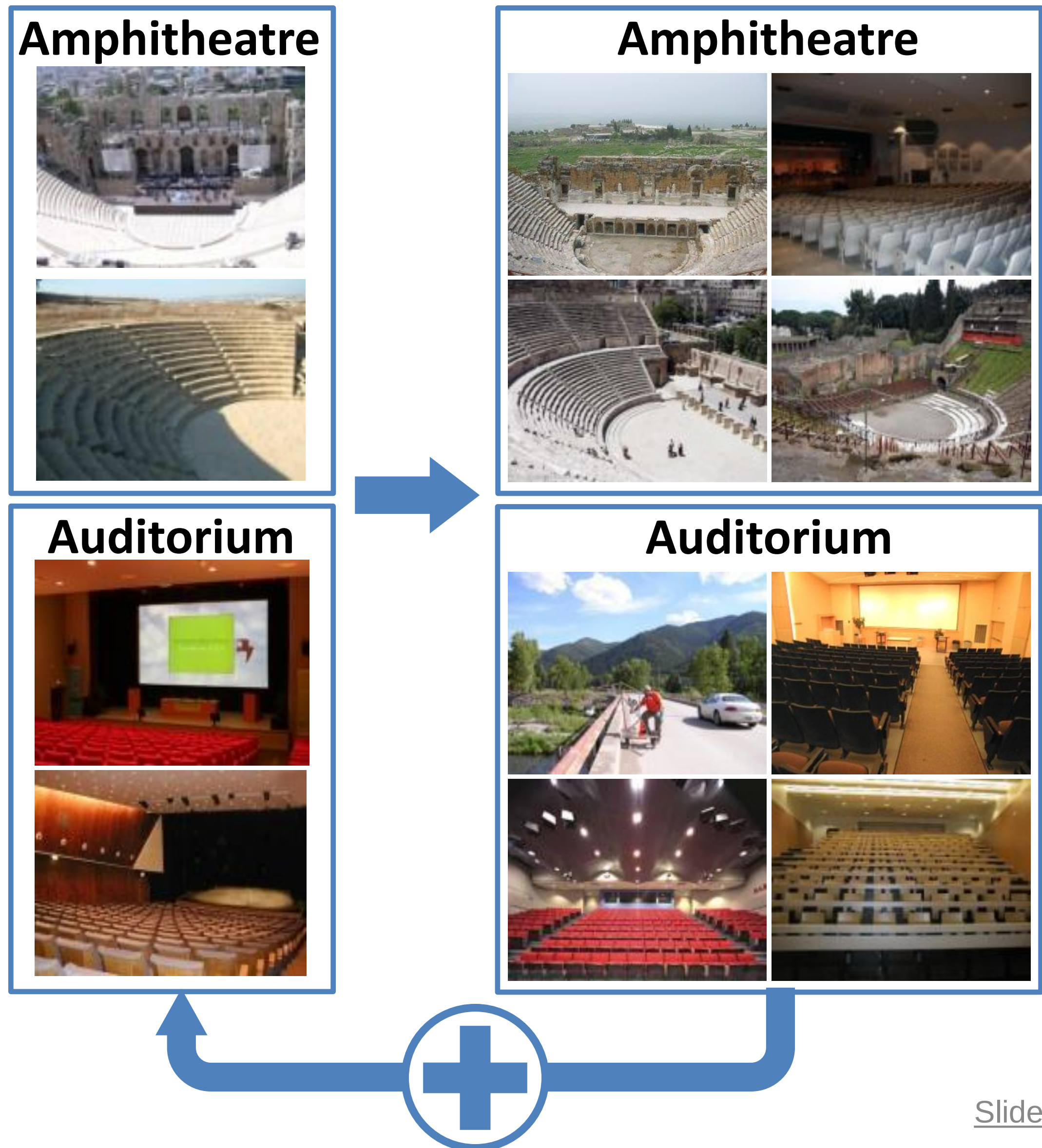
Barn > Cemetery

Slide Credit: Abhinav Gupta

Bootstrapped scene learning

Labeled Seed Examples

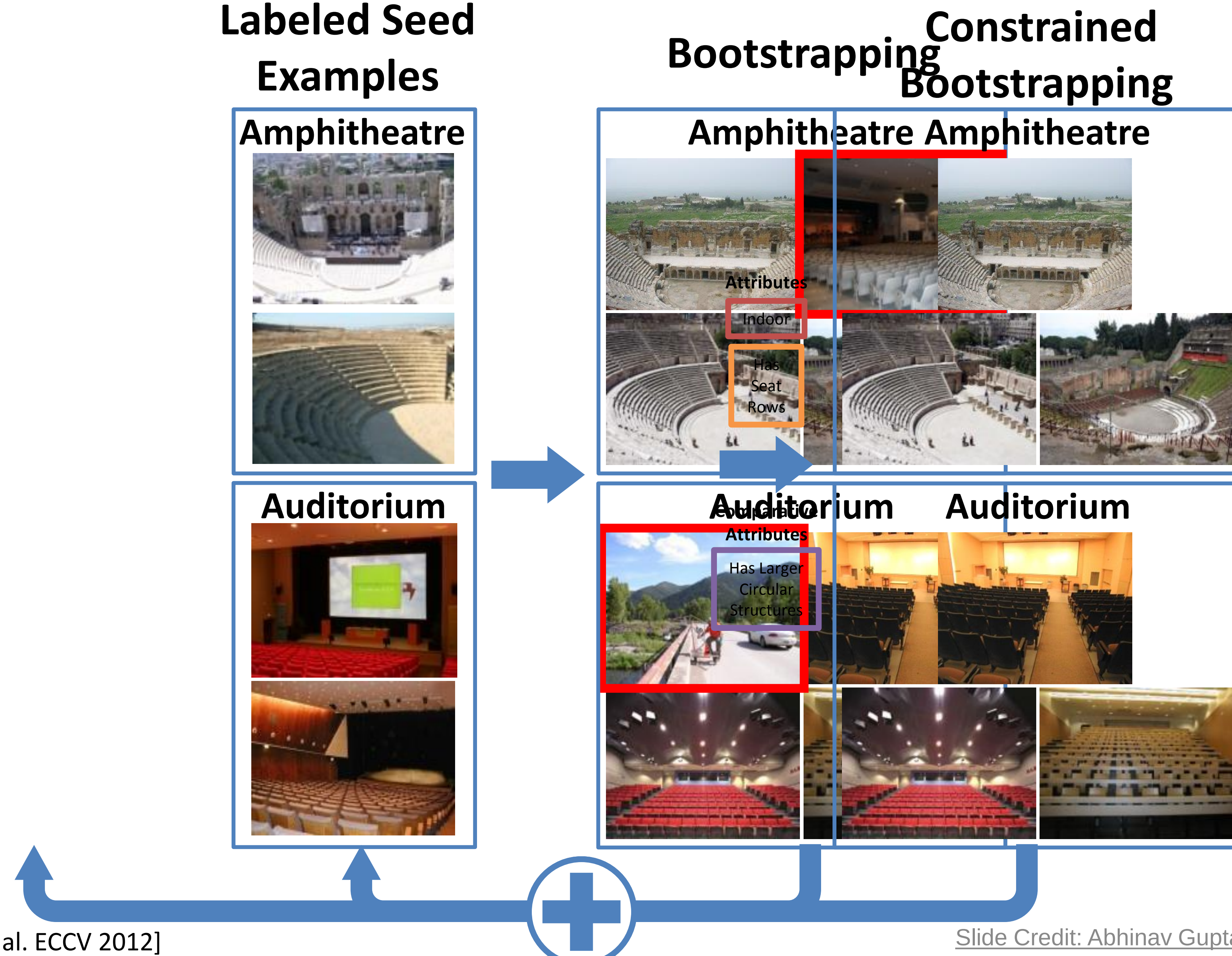
Bootstrapping



[Gupta et al. ECCV 2012]

Slide Credit: Abhinav Gupta

Bootstrapped scene learning



[Gupta et al. ECCV 2012]

Slide Credit: Abhinav Gupta

Applications of relative attributes

Enable new modes of human-system communication

- **Training category models through descriptions**:

  "Rabbits are furrier than dogs."

- **Rationales to explain image labels:**

  "It's not a coastal scene because it's too cluttered."

- **Semantic relative feedback for image search:**

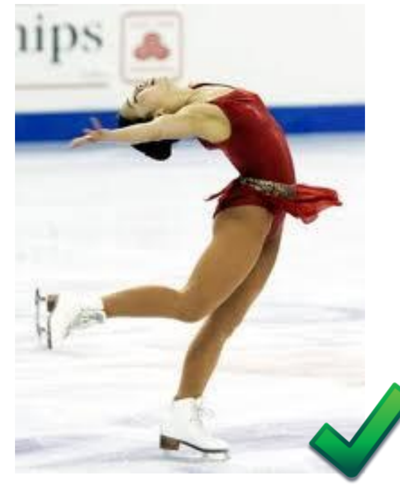
  "I want shoes like these, but shinier."

Complex visual recognition tasks

[Donahue and Grauman, ICCV 2011]

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:

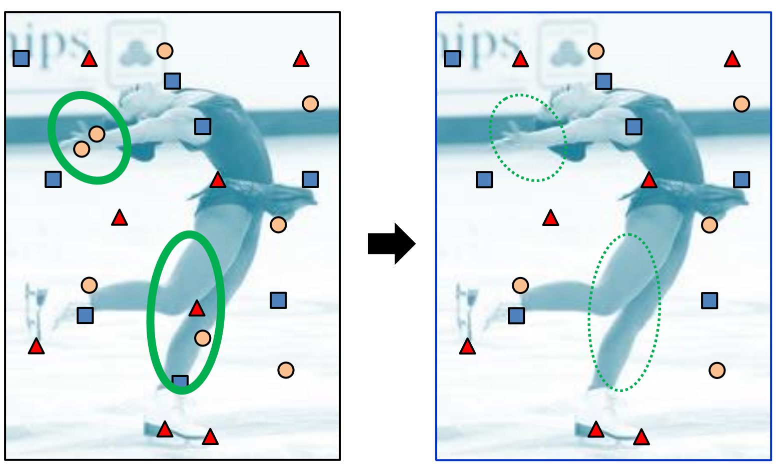
- Solicit a visual rationale for the label.
- Ask the annotator not just *what*, but also *why.*

Soliciting visual rationales

Annotation task: Is her form good? How can you tell?

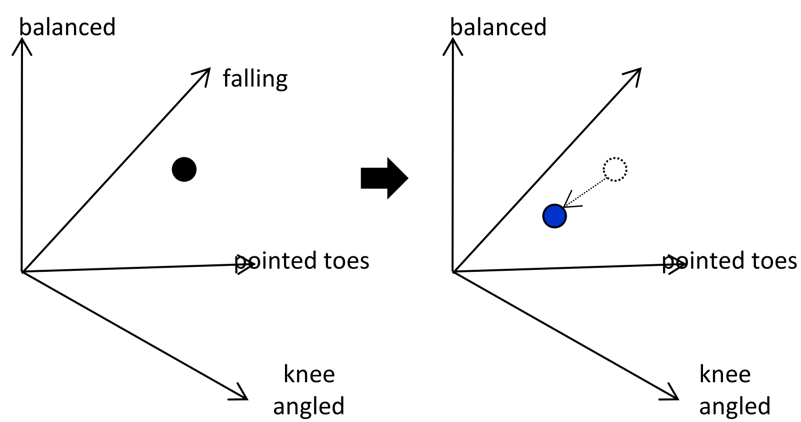
☑ pointed toes
☑ balanced
☐ falling
☐ knee angled

Spatial rationale



Synthetic contrast example

Attribute rationale

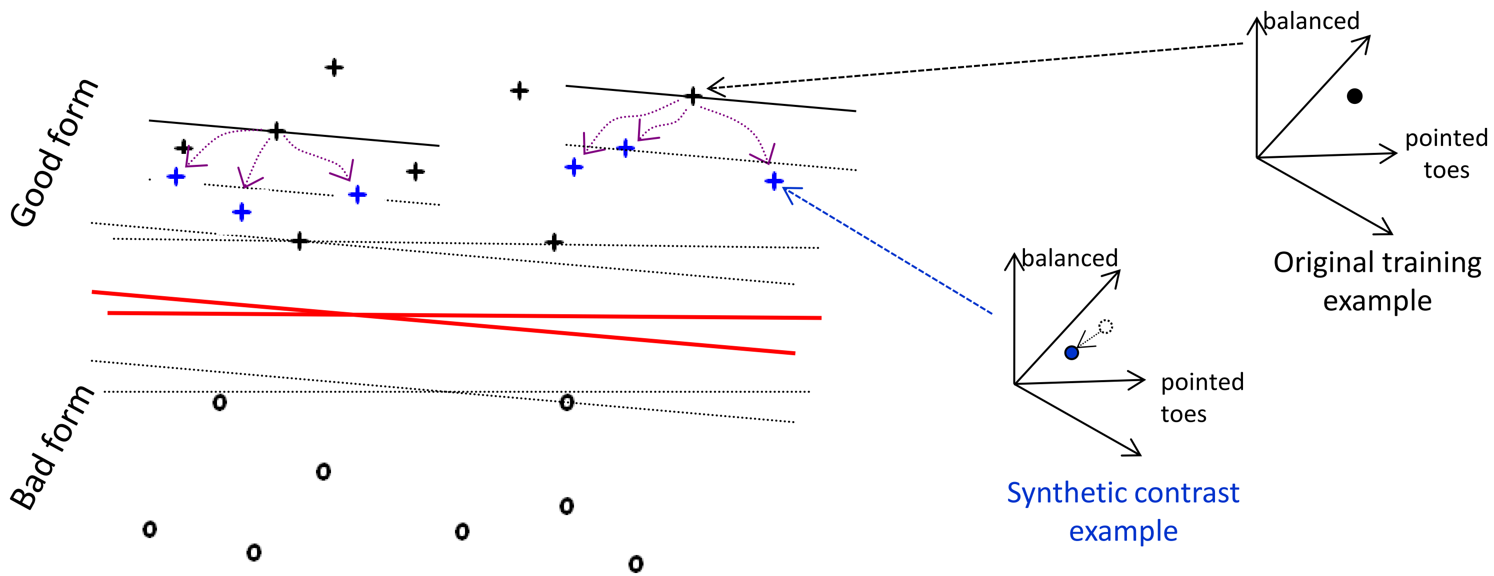


Synthetic contrast example

[Annotator Rationales for Visual Recognition. J. Donahue and K. Grauman, ICCV 2011]

Kristen Grauman, UT-Austin

Rationales' influence on the classifier



Decision boundary *refined* in order to satisfy "secondary" margin

$$\text{minimize} \quad \left(\frac{1}{2}||\boldsymbol{w}||^2 + C\sum_i \xi_i + C_c\sum_i \gamma_i\right)$$

$$s.t. \quad y_i \boldsymbol{w}^T \boldsymbol{x_i} \geq 1 - \xi_i; \quad \forall i \in \mathcal{L}$$

$$\boxed{y_i(\boldsymbol{w}^T\boldsymbol{x_i} - \boldsymbol{w}^T\boldsymbol{v}_i) \geq \mu(1-\gamma_i); \quad \forall i \in \mathcal{C}}$$

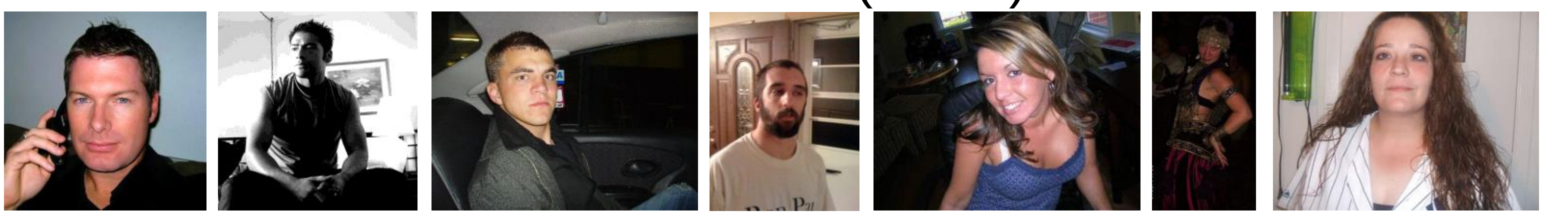
$$\xi_i \geq 0; \gamma_i \geq 0,$$

[Zaidan et al. Using Annotator Rationales to Improve Machine Learning for Text Categorization, NAACL HLT 2007]

Rationale results

- **Scene Categories**: How can you tell the scene category?

- **Hot or Not**: What makes them hot (or not)?



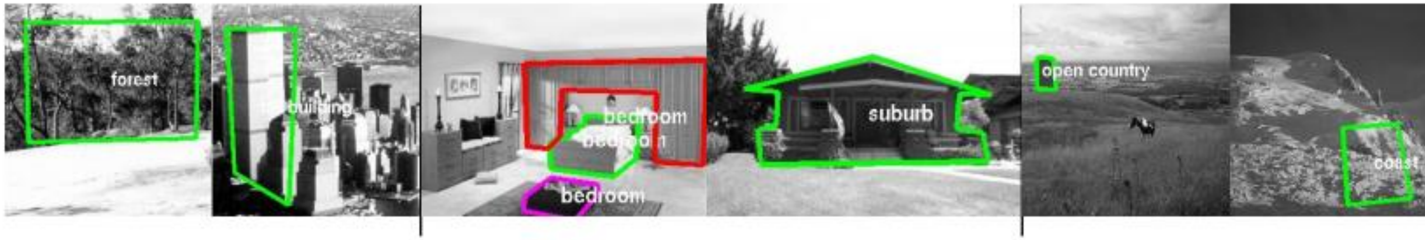
- **Public Figures**: What attributes make them (un)attractive?

Collect rationales from hundreds of MTurk workers.

[Annotator Rationales for Visual Recognition. J. Donahue and K. Grauman, ICCV 2011]

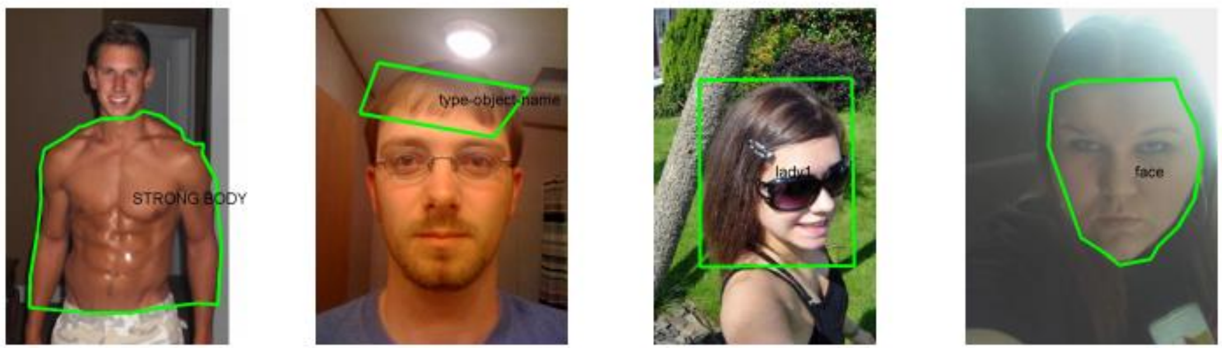
Example rationales from MTurk

Scene categories



Typical | Tight | "Artistic"

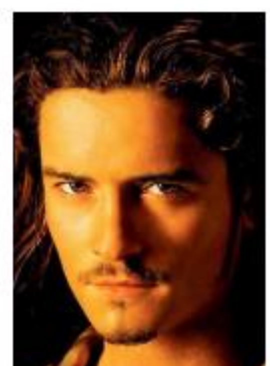
Hot or Not



Hot, Male | Not, Male | Hot, Female | Not, Female

PubFig Attractiveness

Youth
Smiling
Straight Hair
Narrow Eyes



Youth
Black Hair
Goatee
Square Face
Shiny Skin
High Cheekbones

Kristen Grauman, UT-Austin

Rationale results

Mean AP

Scenes	Originals	+Rationales
Kitchen	0.1196	**0.1395**
Living Rm	0.1142	**0.1238**
Inside City	0.1299	**0.1487**
Coast	0.4243	**0.4513**
Highway	0.2240	**0.2379**
Bedroom	0.3011	**0.3167**
Street	0.0778	**0.0790**
Country	0.0926	**0.0950**
Mountain	0.1154	**0.1158**
Office	0.1051	**0.1052**
Tall Building	0.0688	**0.0689**
Store	0.0866	**0.0867**
Forest	0.3956	**0.4006**

Hot or Not	Originals	+Rationales
Male	54.86%	**60.01%**
Female	55.99%	**57.07%**

PubFig	Originals	+Rationales
Male	64.60%	**68.14%**
Female	51.74%	**55.65%**

[Donahue & Grauman, ICCV 2011]

Rationale results

Why not just use discriminative feature selection?

Scenes	Originals	+Rationales	Mutual information
Kitchen	0.1196	**0.1395**	0.1202
Living Rm	0.1142	**0.1238**	0.1159
Inside City	0.1299	**0.1487**	0.1245
Coast	0.4243	**0.4513**	0.4129
Highway	0.2240	**0.2379**	0.2112
Bedroom	0.3011	**0.3167**	0.2927
Street	0.0778	**0.0790**	0.0775
Country	0.0926	**0.0950**	0.0941
Mountain	0.1154	**0.1158**	0.1154
Office	0.1051	**0.1052**	0.1048
Tall Building	0.0688	**0.0689**	0.0686
Store	0.0866	**0.0867**	0.0866
Forest	0.3956	**0.4006**	0.3897

Mean AP

[Donahue & Grauman, ICCV 2011]

Relative feedback for object learning

[Parkash & Parikh, ECCV 2012]

Current belief

I think this is a giraffe. What do you think?

No, its neck is too short for it to be a giraffe.

Knowledge of the world

[Animals with even shorter necks]

Ah! These must not be giraffes either then.

……

Feedback on one, transferred to many

Biswas & Parikh, CVPR 2013; Parkash & Parikh, ECCV 2012]

Slide credit: Devi Parikh

Applications of relative attributes

Enable new modes of human-system communication

- **Training category models through descriptions**:

  "Rabbits are furrier than dogs."

- **Rationales to explain image labels:**

  "It's not a coastal scene because it's too cluttered."

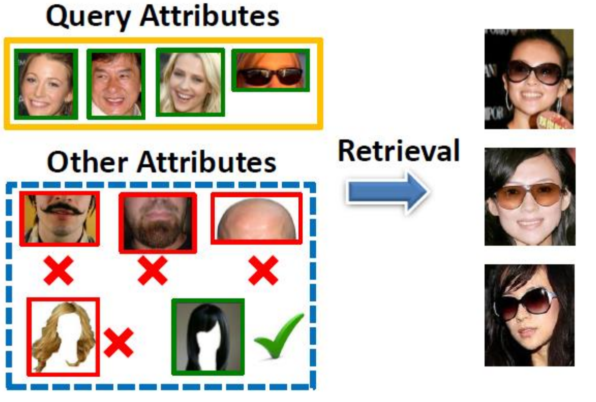
- **Semantic relative feedback for image search:**

  "I want shoes like these, but shinier."

Attributes for search

Previously, attributes serve as keywords for one-shot search

Kumar et al. 2008



Siddiquie et al. 2011

Vaquero et al. 2009

Kristen Grauman, UT-Austin

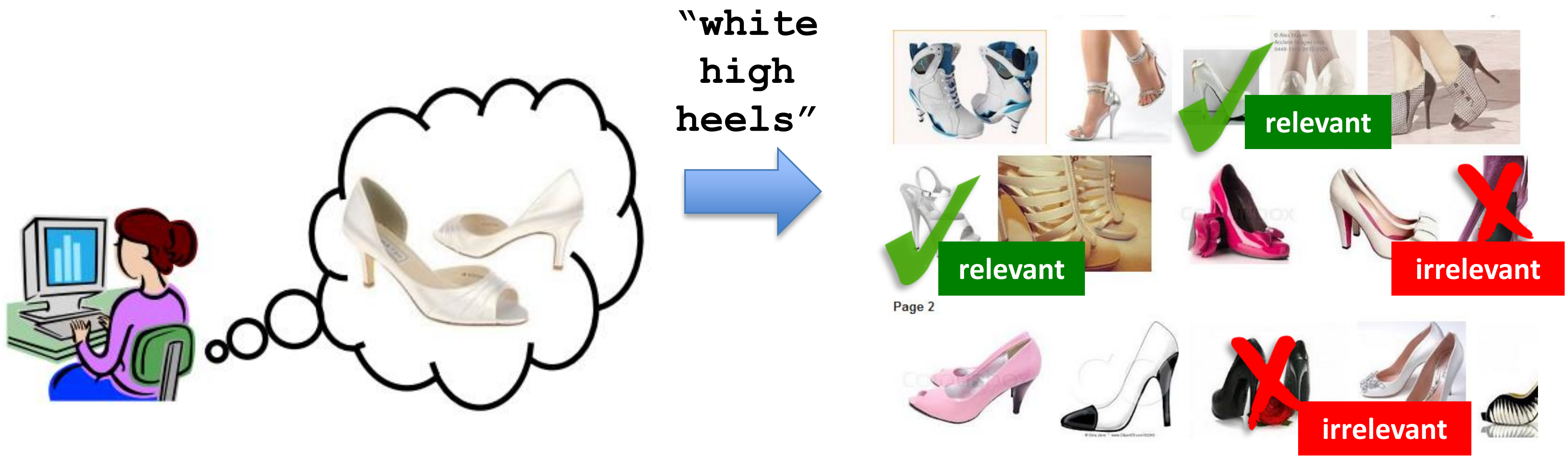
Problem with one-shot visual search

- But keywords (including attributes) can be insufficient to capture target in one shot.

$\neq$

brown strappy heels

Interactive visual search

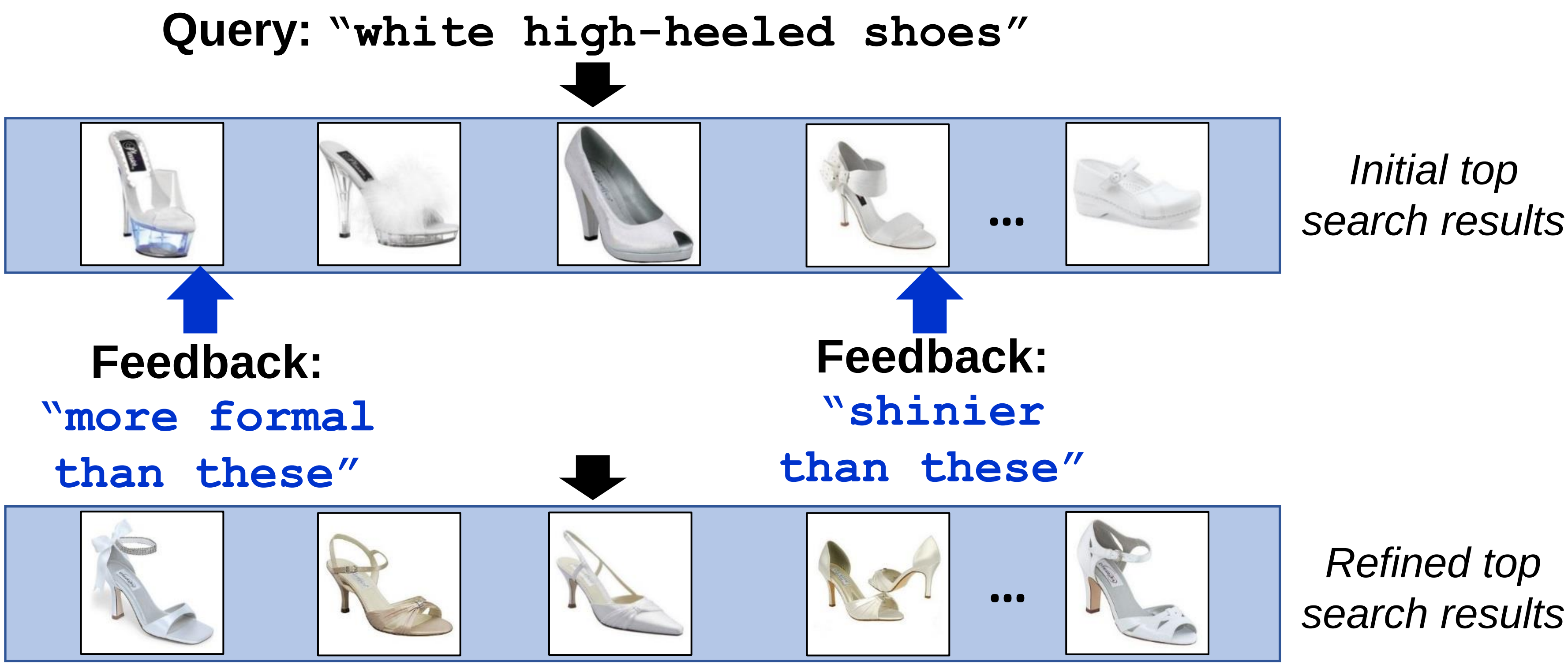


- Interactive search can help iteratively refine
- …but traditional **binary relevance feedback** offers only coarse communication between user and system

Kristen Grauman, UT-Austin

WhittleSearch: Relative attribute feedback

[Kovashka et al. CVPR 2012]



Whittle away irrelevant images via precise semantic feedback

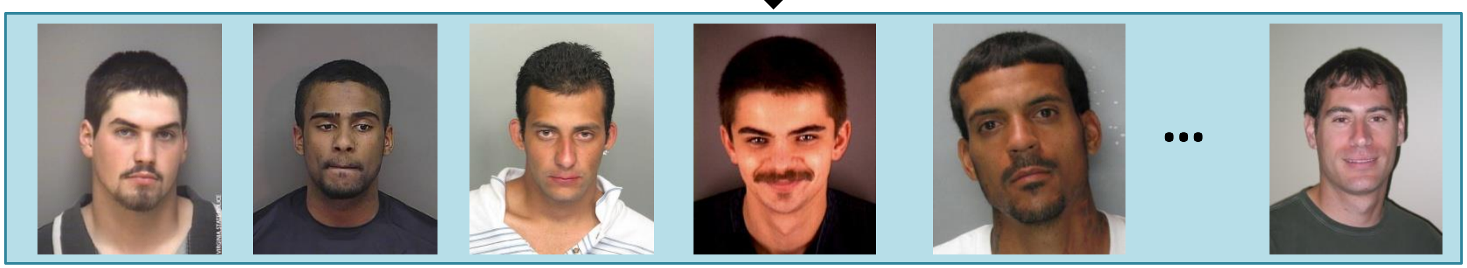
WhittleSearch: Relative attribute feedback

[Kovashka et al. CVPR 2012]

Initial reference images

Feedback: **"similar hair style"**

Feedback: **"broader nose"**



Refined top search results

Whittle away irrelevant images via precise semantic feedback

Kovashka, Parikh, and Grauman, CVPR 2012

WhittleSearch with relative attribute feedback

<u>Offline:</u>

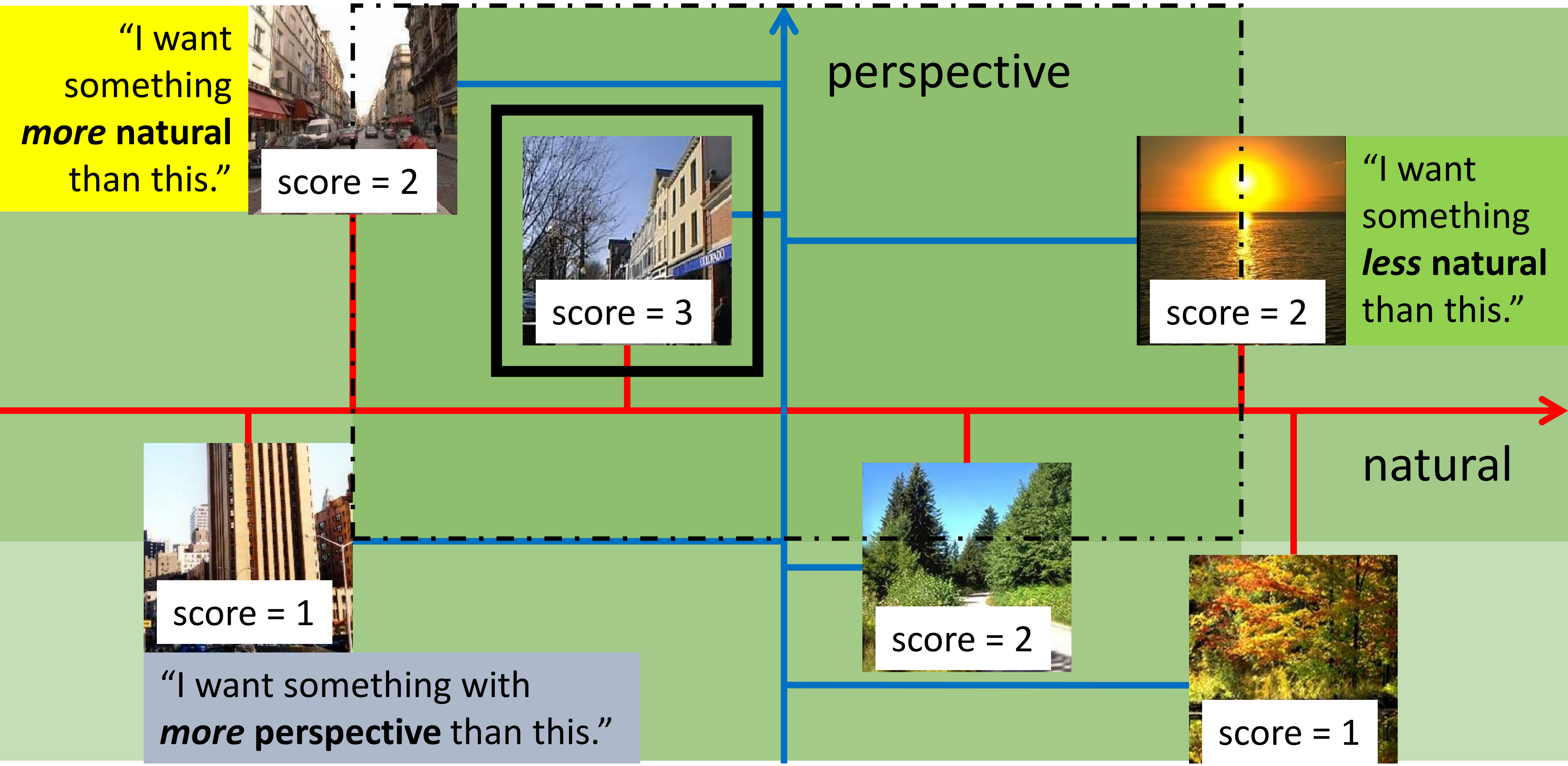
We learn a spectrum for each attribute

<u>During search:</u>

1. User selects some reference images and marks *how they differ from the* desired target
2. We update the scores for each database image

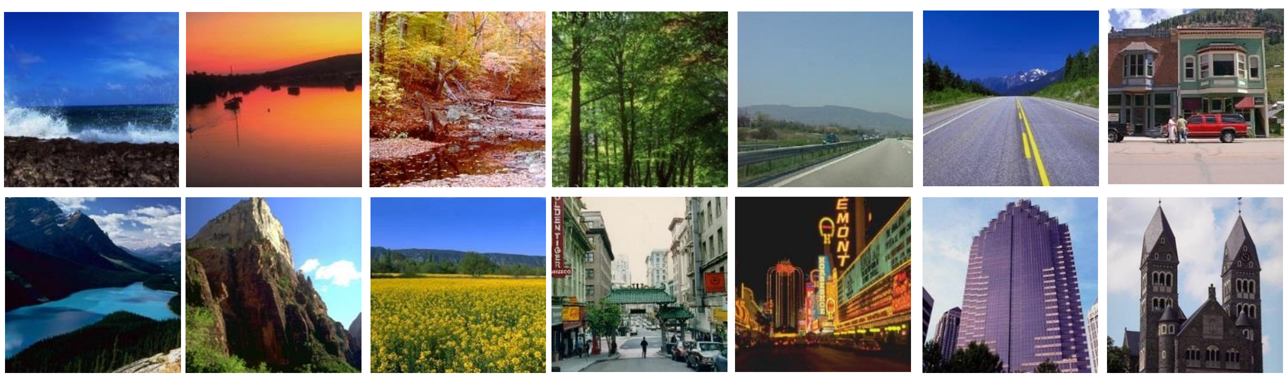
Kristen Grauman, UT-Austin

WhittleSearch with relative attribute feedback



Datasets

Shoes: [Berg; Kovashka]
14,658 shoe images;
10 attributes:
“pointy”, “bright”, “high-heeled”, “feminine” etc.

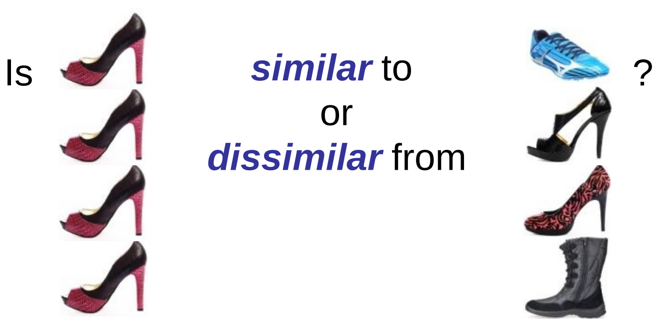


OSR: [Oliva & Torralba]
2,688 scene images;
6 attributes:
“natural”, “perspective”, “open-air”, “close-depth” etc.

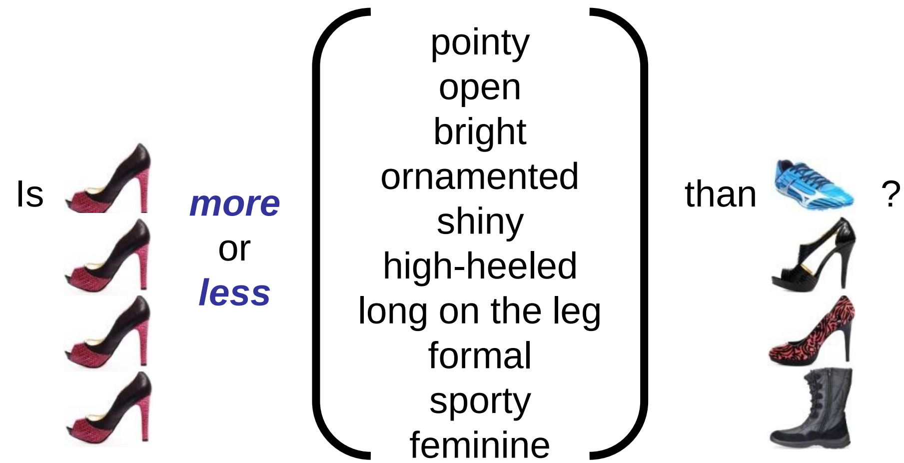
PubFig: [Kumar et al.]
772 face images;
11 attributes:
“masculine”, “young”, “smiling”, “round-face”, etc.

Experimental setup

- Give the user the target image to look for
- Pair each target image with 16 reference images
- Get judgments on pairs from users on MTurk



Binary feedback baseline



Relative attribute feedback

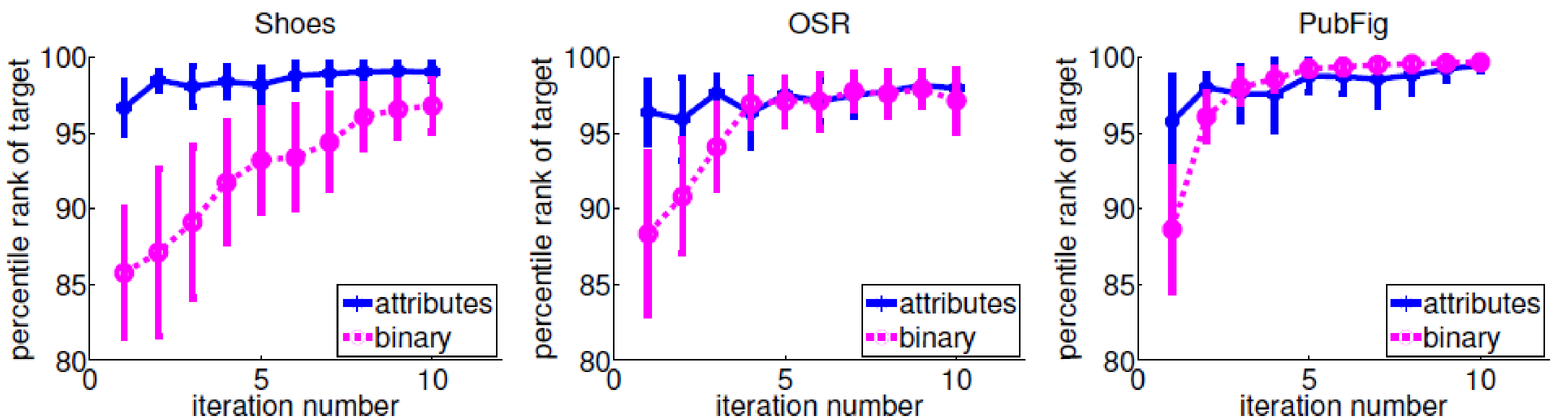
WhittleSearch Results

Binary relevance feedback

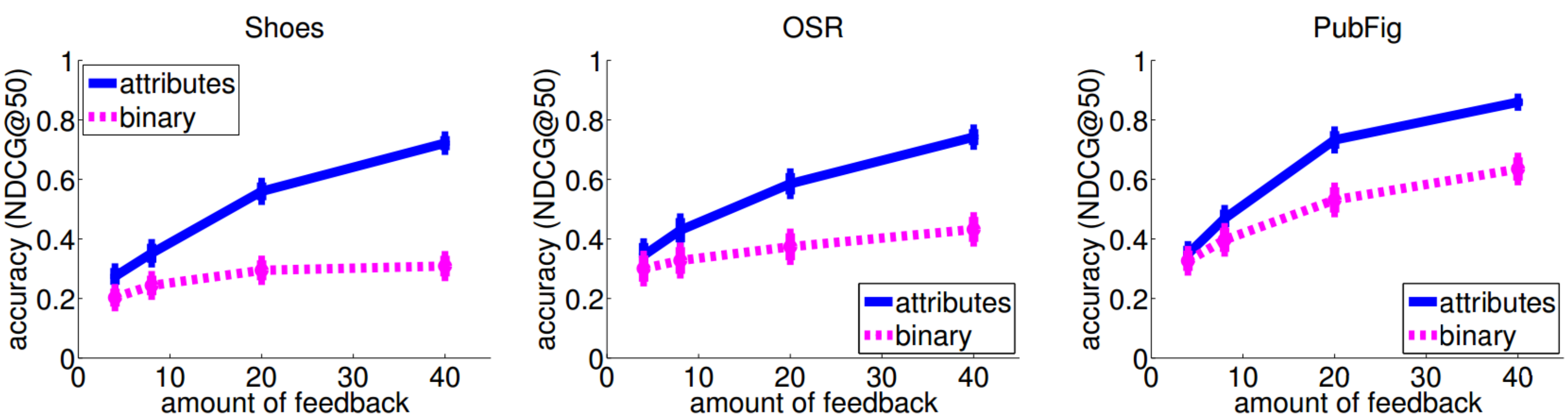
Relative attribute feedback

[Kovashka et al., CVPR 2012]

WhittleSearch Results



We more rapidly converge on the envisioned visual content.

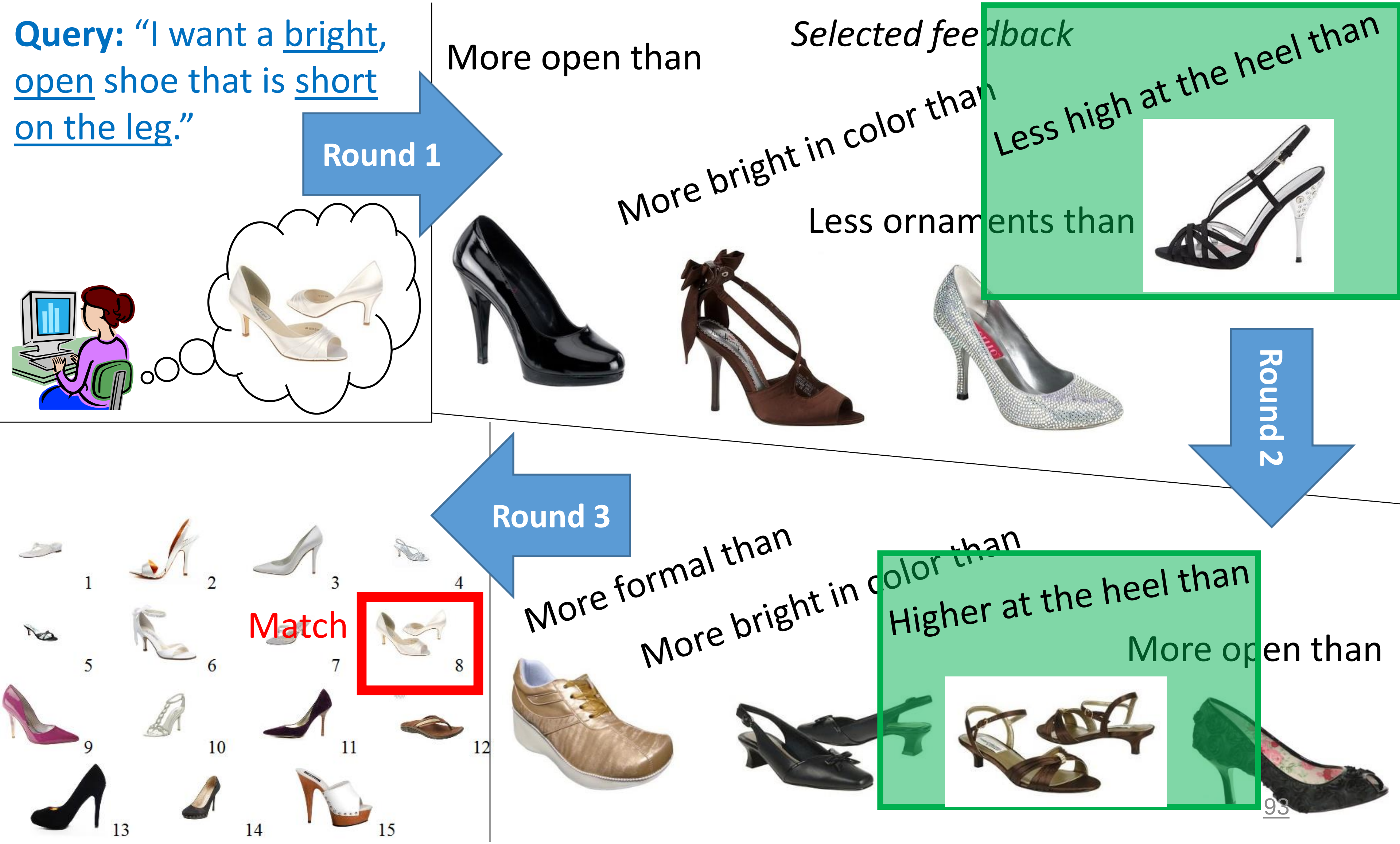


Richer feedback → faster gains per unit of user effort.

[Kovashka et al., CVPR 2012]

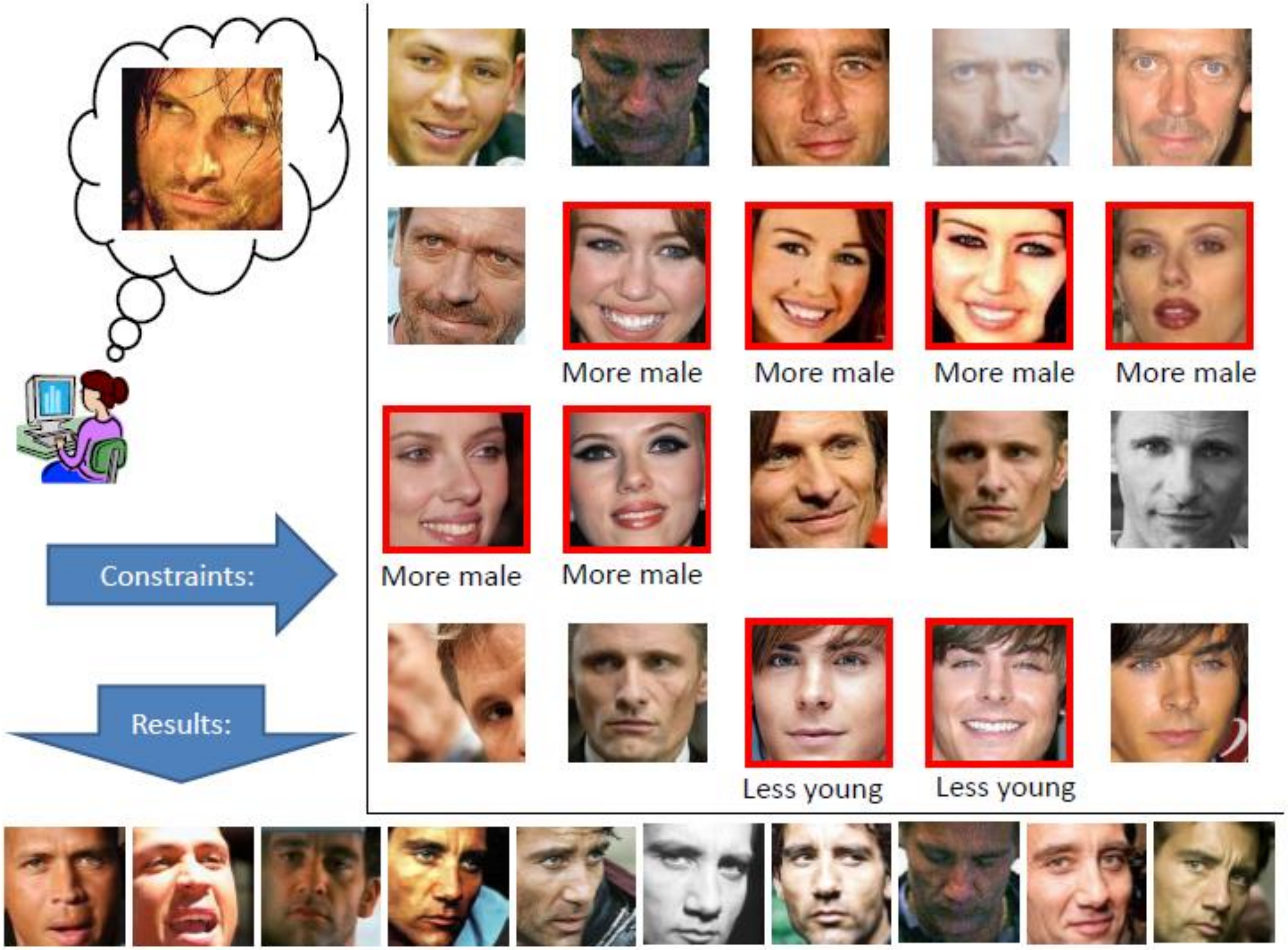
Kristen Grauman, UT-Austin

Example WhittleSearch



[Kovashka et al., CVPR 2012]

Failure case (?)

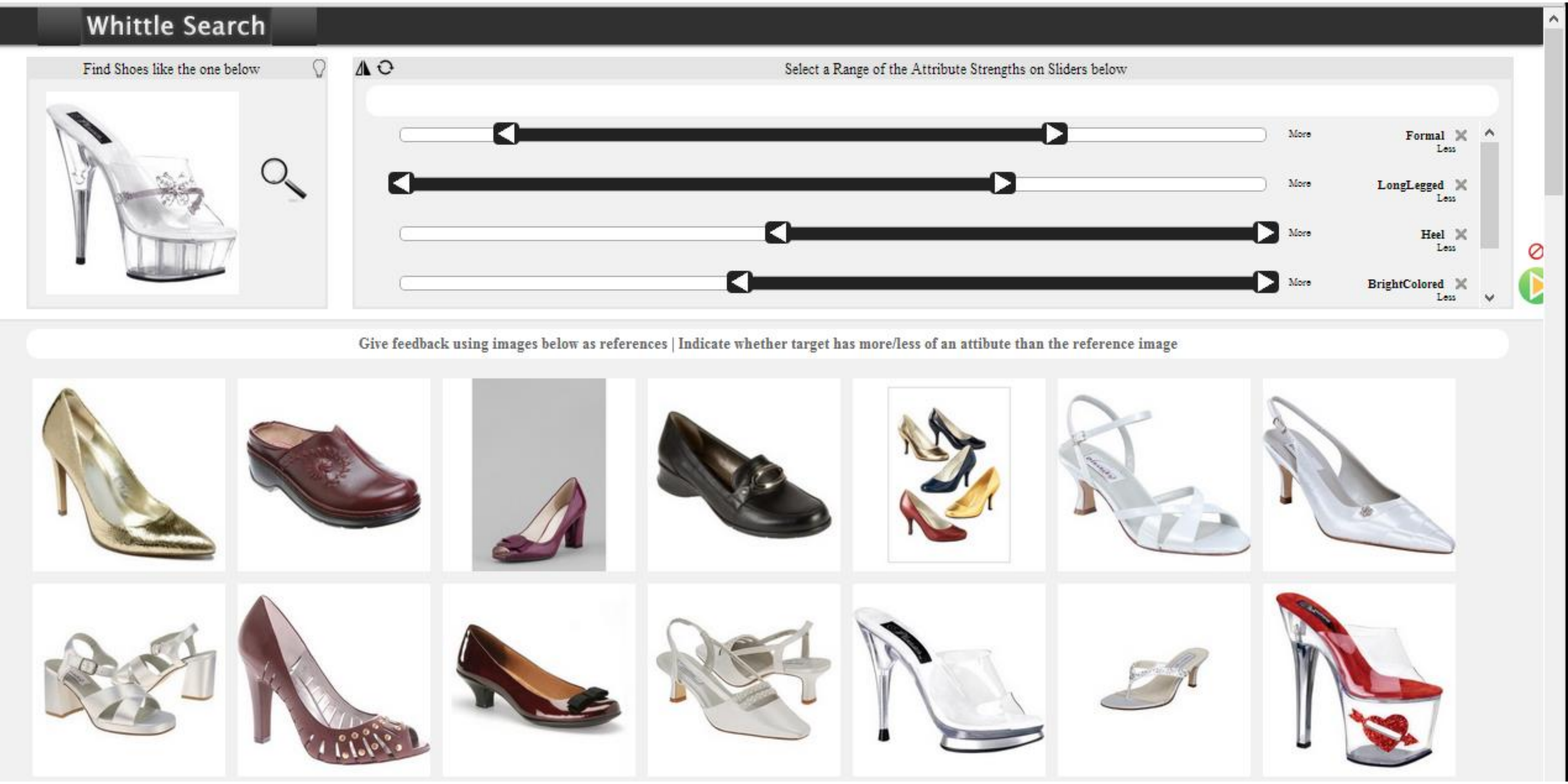


Is the user searching for a specific person (identity), or an image similar to the specific target image?

Kristen Grauman, UT-Austin

WhittleSearch Demo

http://godel.ece.vt.edu/whittle/



Kristen Grauman, UT-Austin

Problem: Where is feedback most useful?

- The most *relevant* images might not be most *informative*
- Existing active methods focus on binary relevance, expensive selection procedures

  [Tong & Chang 2001, Li et al. 2001, Cox et al. 2000, Ferecatu & Geman 2007, …]

Kristen Grauman, UT-Austin

Idea: Attribute Pivots for Guiding Feedback

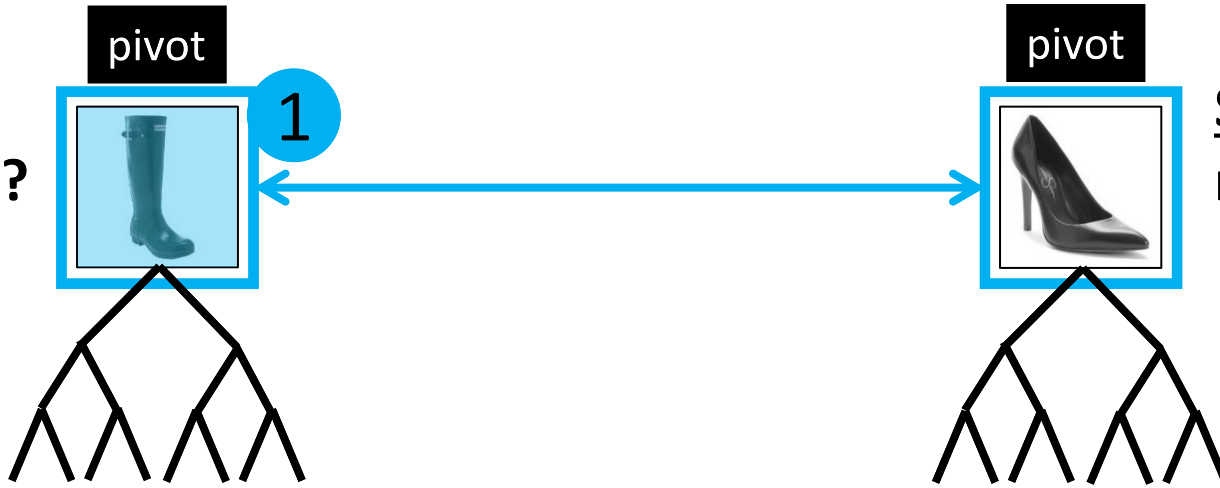
[Kovashka and Grauman, 2013]

- Select series of most informative *visual comparisons* that user should make to help deduce target
- Use binary search trees in attribute space for rapid selection

Kristen Grauman, UT-Austin

Selecting a Series of Informative Comparisons

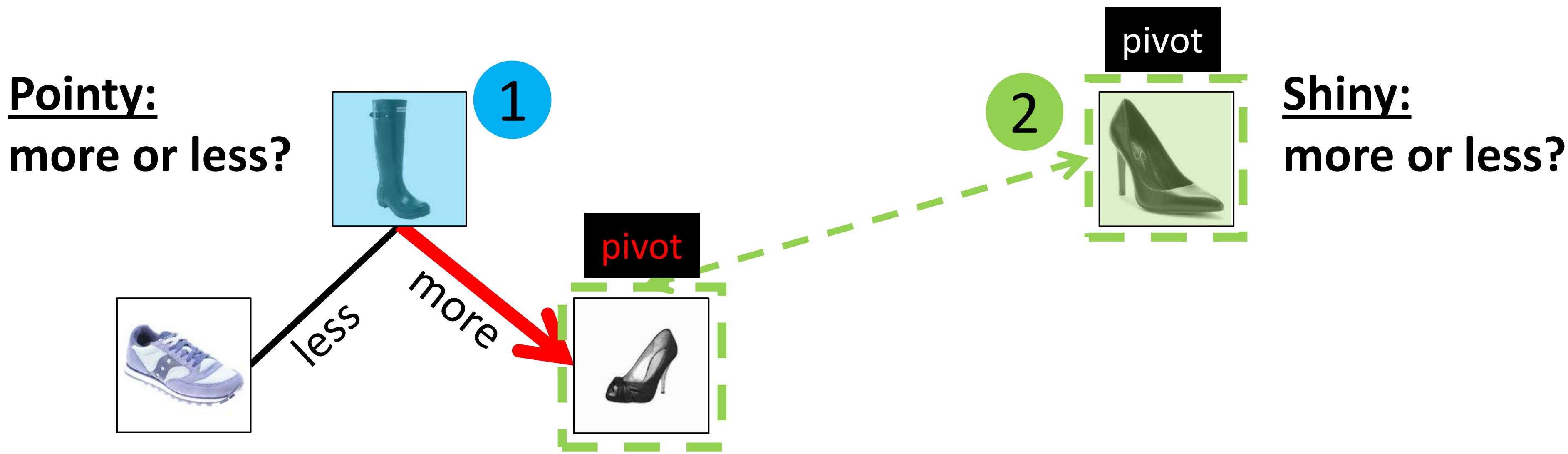
<u>Pointy:</u>
more or less?



<u>Shiny:</u>
more or less?

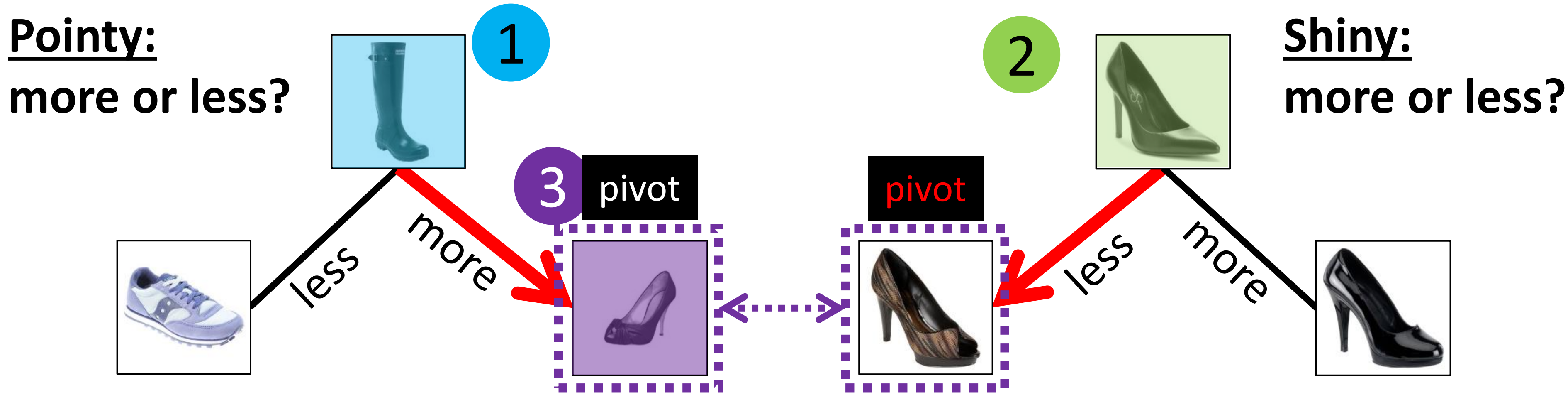
Kristen Grauman, UT-Austin

Selecting a Series of Informative Comparisons

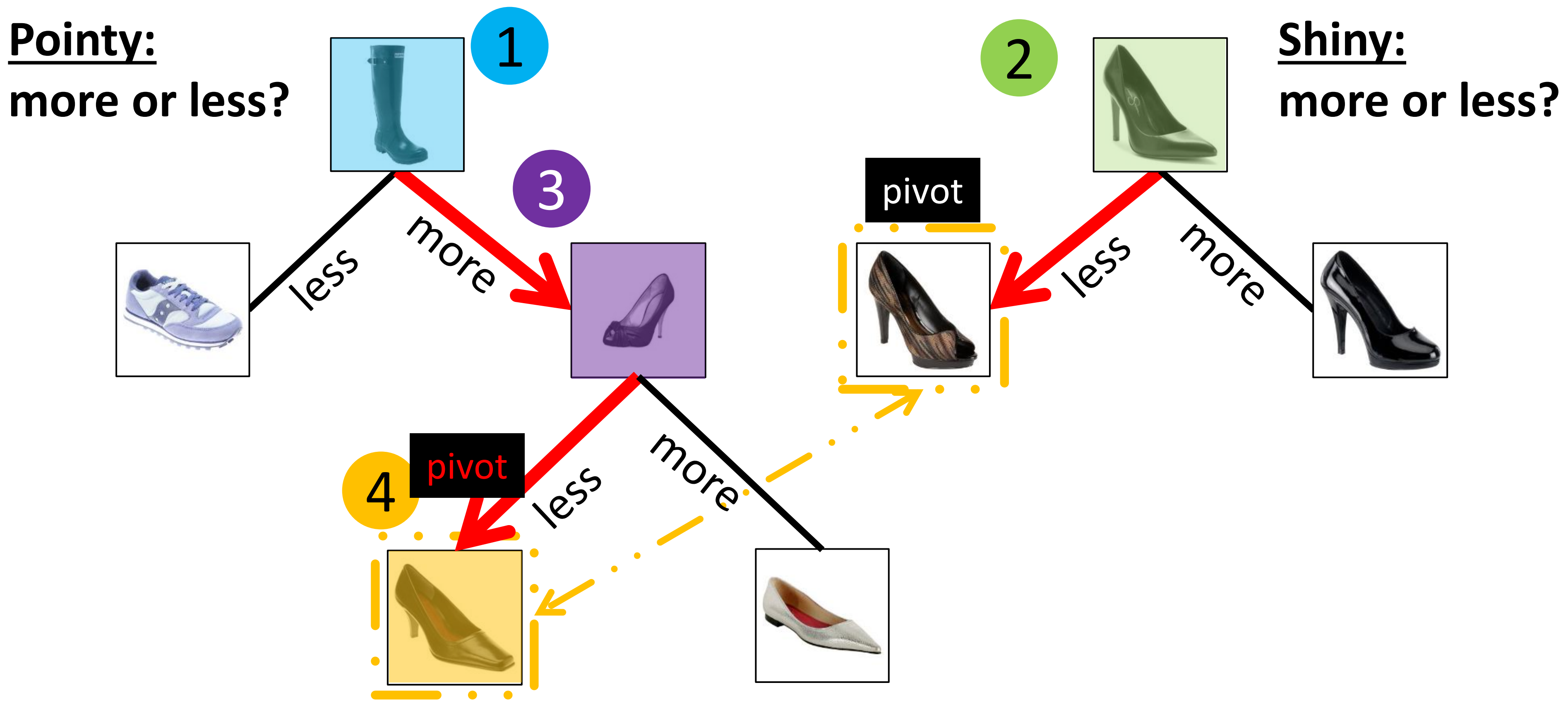


Kristen Grauman, UT-Austin

Selecting a Series of Informative Comparisons



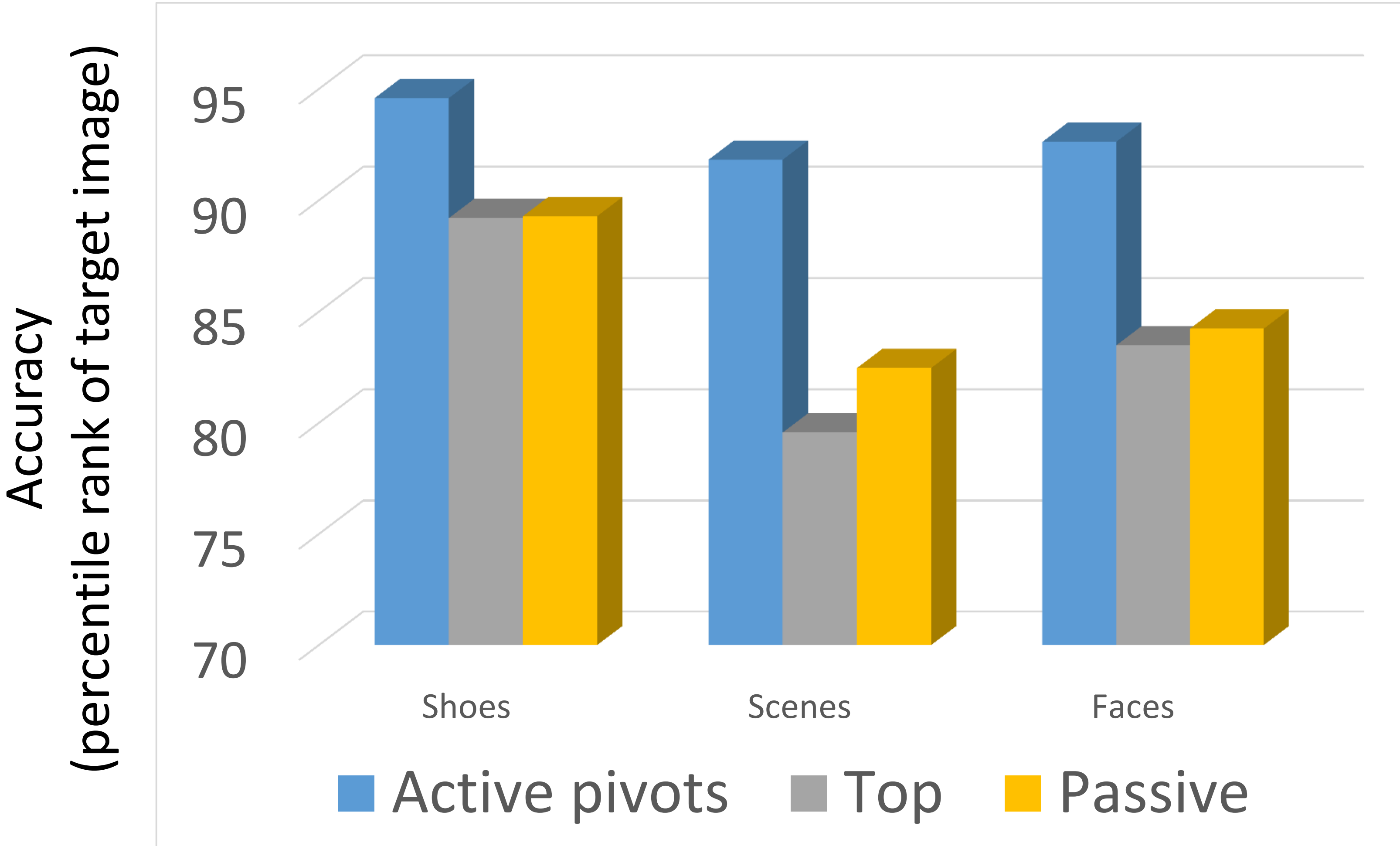
Selecting a Series of Informative Comparisons



Kristen Grauman, UT-Austin

Attribute Pivots for Active WhittleSearch

Active feedback requests zero in on target more quickly



Ongoing issues with attributes

- What attributes should be in the vocabulary?
- How to align user's attribute language with the visual attribute models?
- Joint learning of multiple attributes?
- Category-based vs. image-based comparative constraints?
- Class-specific attributes?
- How do we make sure we're learning the "right" thing?

Summary

- Humans are not simply “label machines”
- More data need not mean better learning
- Active learning focuses annotator effort
- Widen access to visual knowledge by modeling visual comparisons
- Relative attributes enable new applications for recognition and visual search

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.