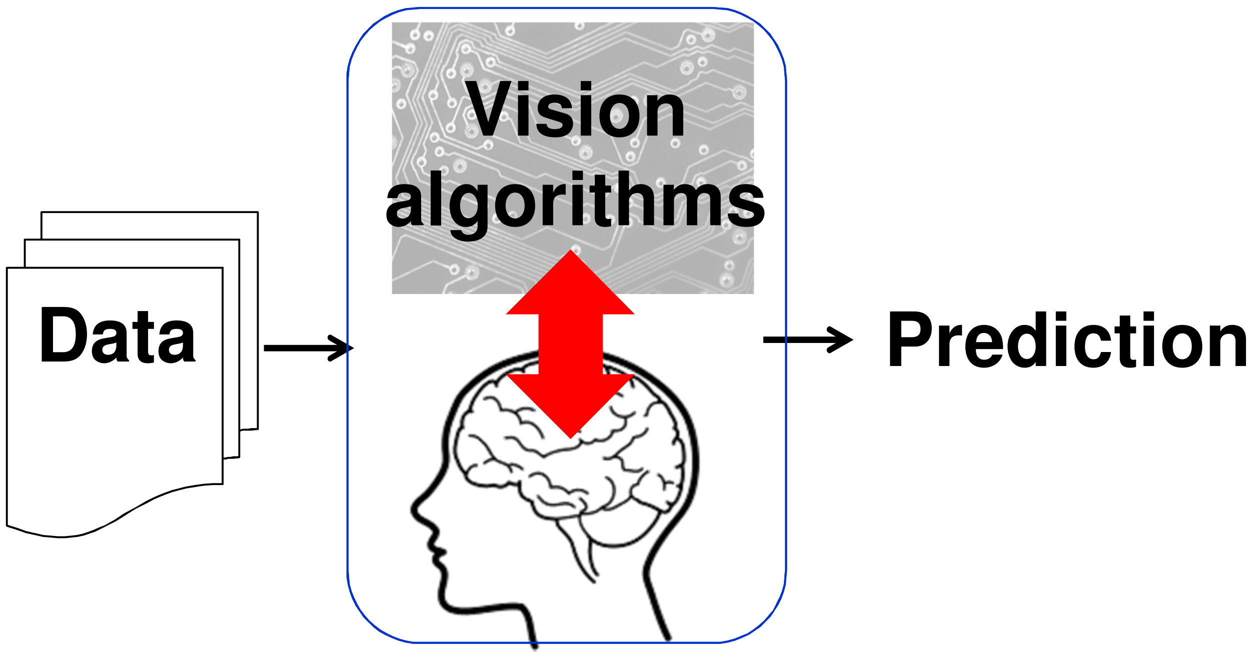
Relative Attributes: Teaching a System through Visual Comparisons

Kristen Grauman
Department of Computer Science
University of Texas at Austin

Work with Adriana Kovashka, Devi Parikh, Jeff Donahue

Interacting with computer vision systems



Semantic gap

Problem: How to teach a vision system...?

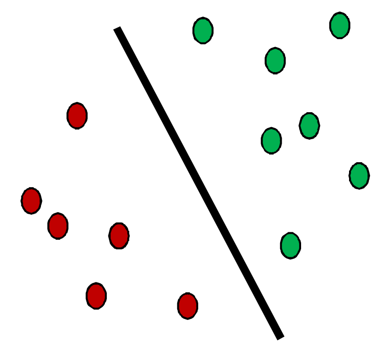
Status quo approach: teach via class labels.

…what we know about object categories?

…what kind of images we want to retrieve?

not relevant

relevant



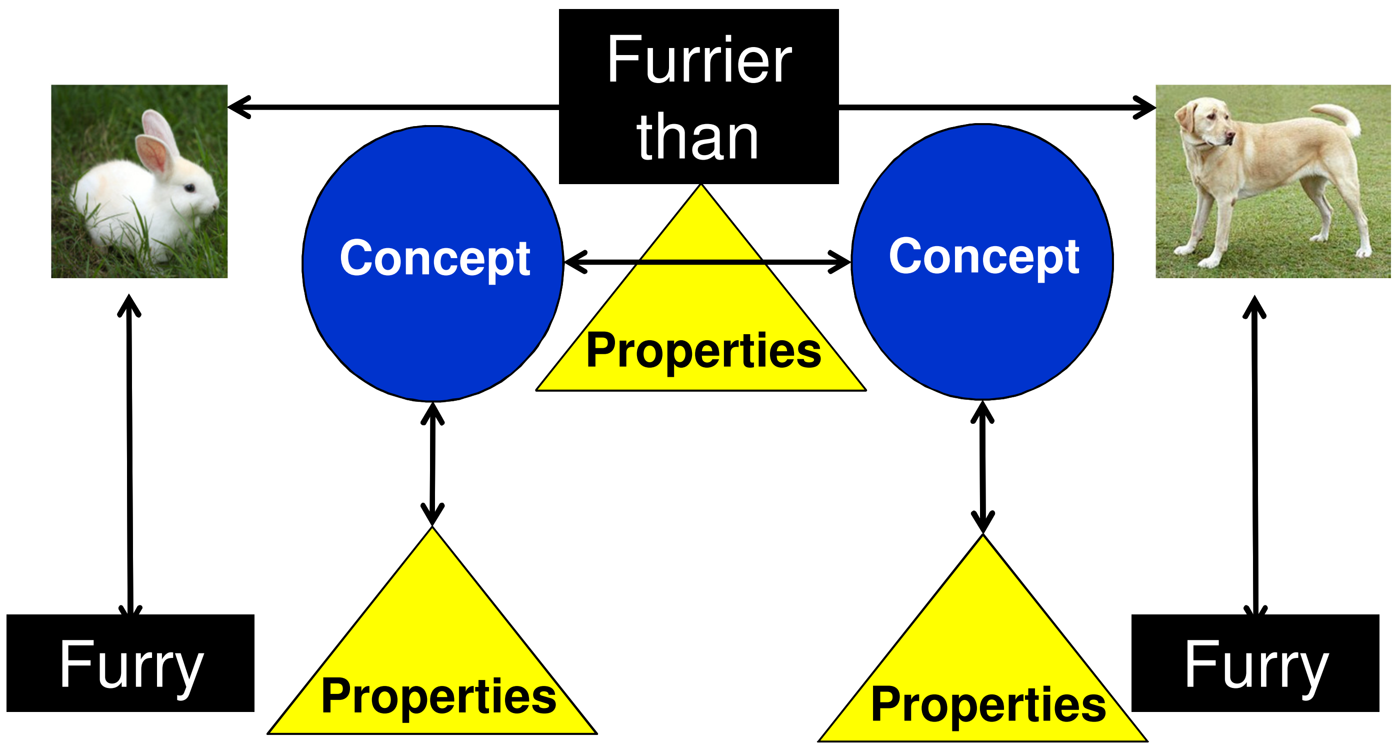
Problem: How to teach a vision system...?

Attributes offer semantic mode of communication,
yet typically restricted to another layer of labels.

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

Our idea: Teach with visual comparisons

We propose **relative attributes** to represent *relationships* between classes, images, and their properties.



[Parikh & Grauman, ICCV 2011]

Our idea: Teach with visual comparisons

We propose **relative attributes** to represent *relationships* between classes, images, and their properties.

→ Enable new modes of human-system communication

- **Training through descriptions**:

  “Rabbits are furrier than dogs.”

- **Rationales to explain image labels:**

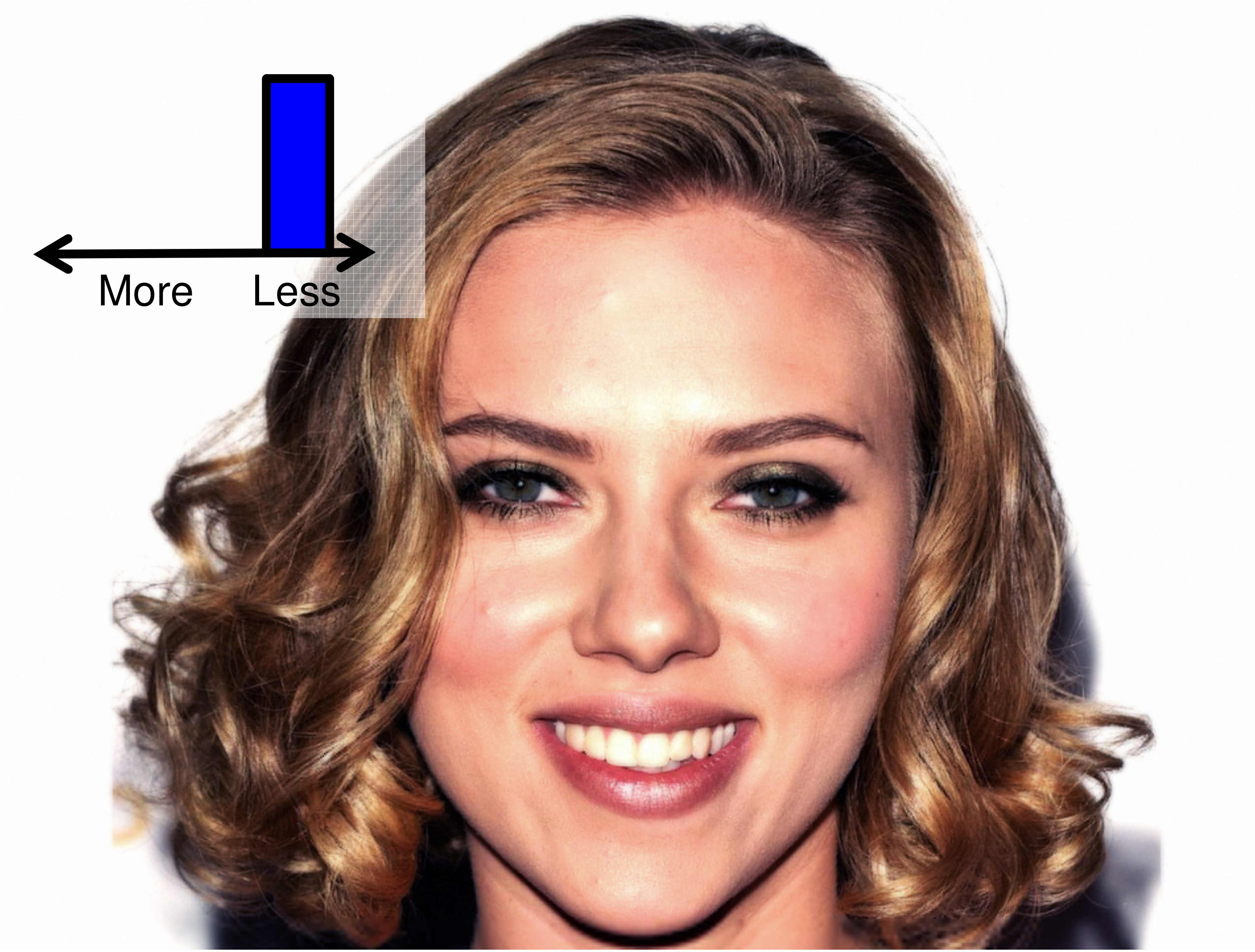
  “It’s not a coastal scene because it’s too cluttered.”

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

  “I want shoes like these, but shinier.”

How should relative attributes be learned?

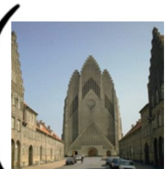
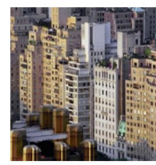
What do we need to capture from human annotators?


More
Less

Learning relative attributes

For each attribute a_m, e.g., "openness"

Supervision consists of:

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

$S_m : \{($  $\sim$  $), \cdots\}$ *Similar pairs*

Learning relative attributes

Learn a ranking function

Image features

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

Learned parameters

that best satisfies the constraints:

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

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

Learning relative attributes

Max-margin learning to rank formulation

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

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

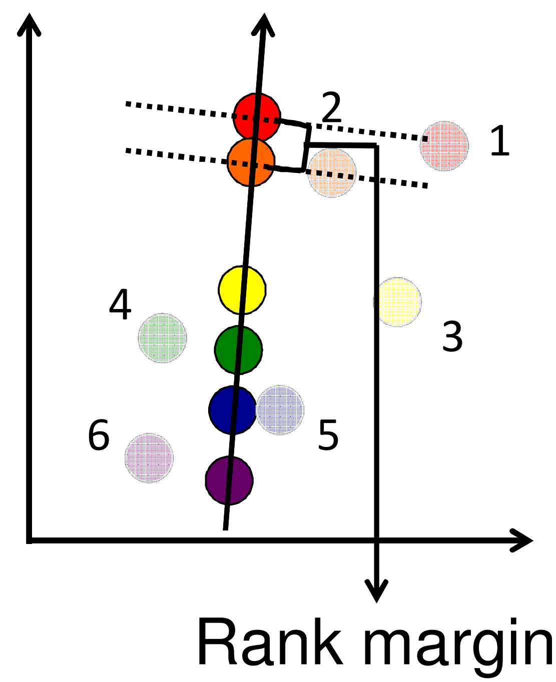


Image $\rightarrow$ Relative attribute score

Relating images

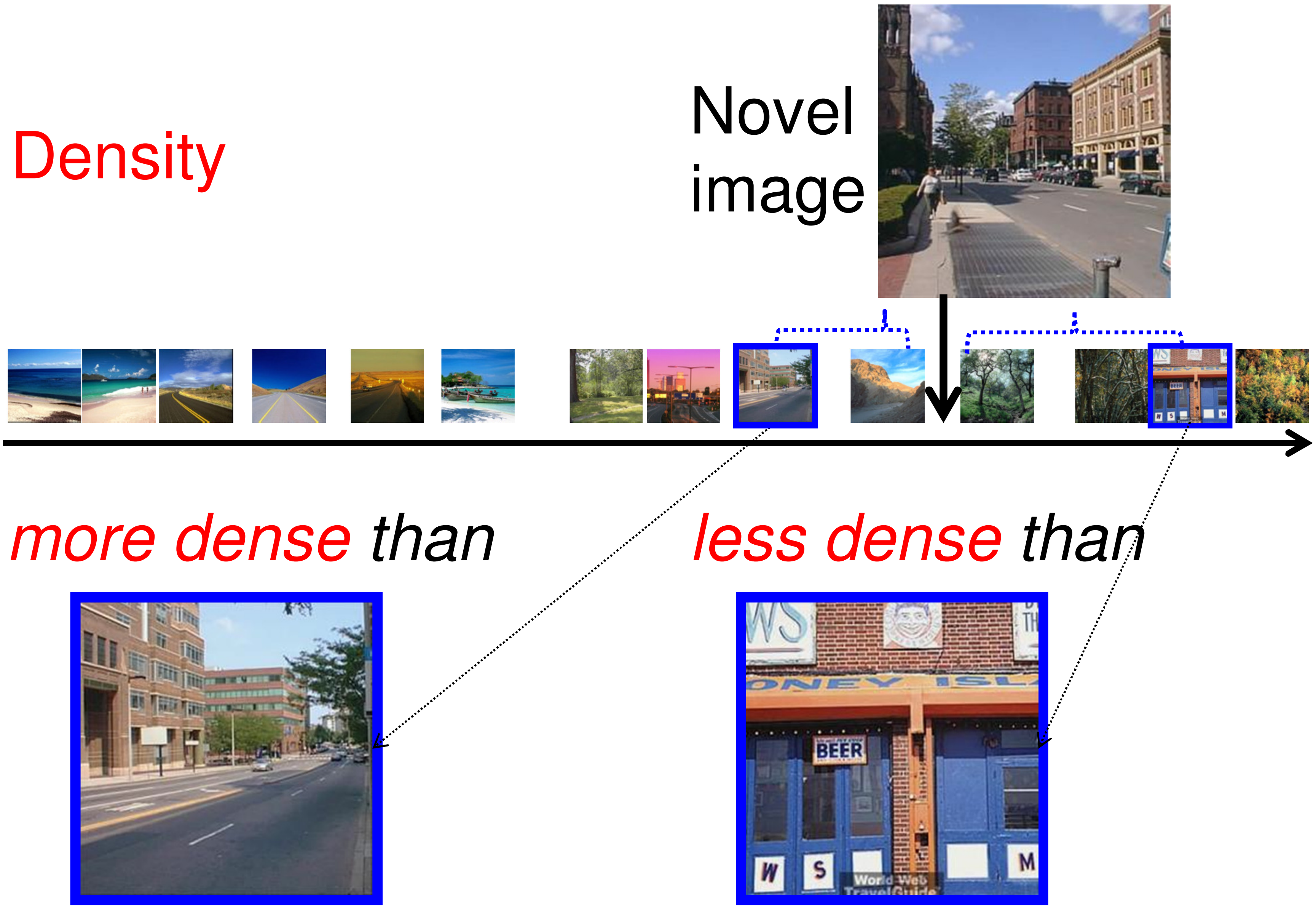
Density

Novel image

Conventional binary description: *not dense*

Relating images

Density



Relating images

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

Relating images

Binary (existing):

Not Young

BushyEyebrows

RoundFace

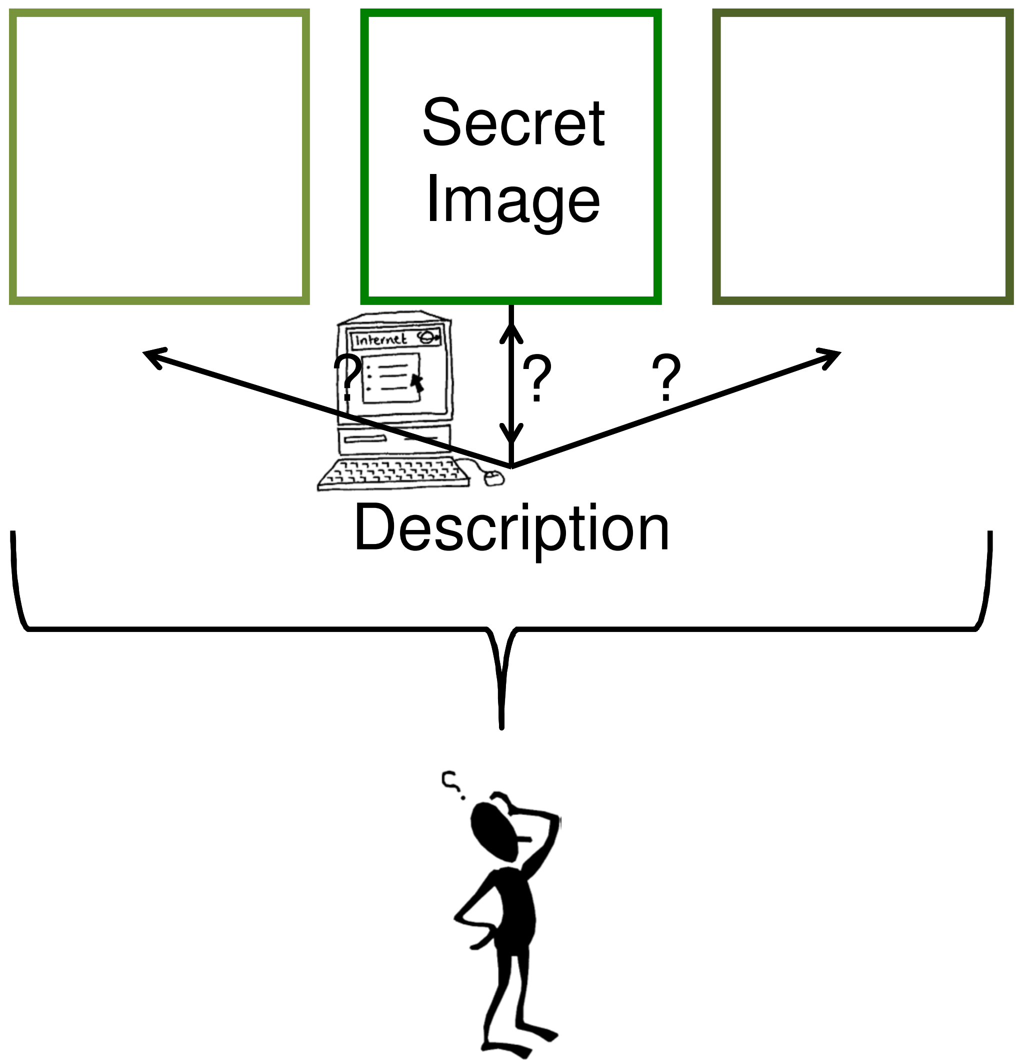
Relative (ours):

More Young than CliveOwen
Less Young than ScarlettJohansson

More BushyEyebrows than ZacEfron
Less BushyEyebrows than AlexRodriguez

More RoundFace than CliveOwen
Less RoundFace than ZacEfron

Human study:
Which image is being described?

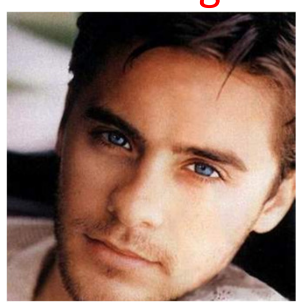


Human study:
Which image is being described?

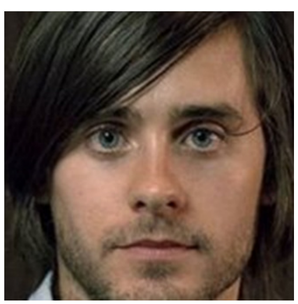
Binary: Smiling, Young

Smiling

Young



Not Smiling



Not Young

Relative

More Smiling than

Younger than

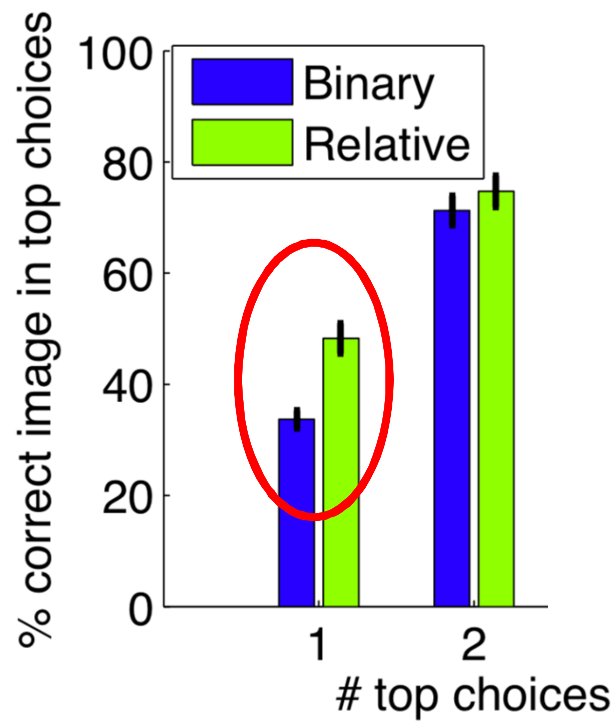
Less Smiling than

Older than

Human study:
Which image is being described?

18 subjects

Test cases:
10 Outdoor Scenes
20 PubFig Faces



Our idea: Teach with visual comparisons

We propose **relative attributes** to represent *relationships* between classes, images, and their properties.

→ Enable new modes of human-system communication

- **Training 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**

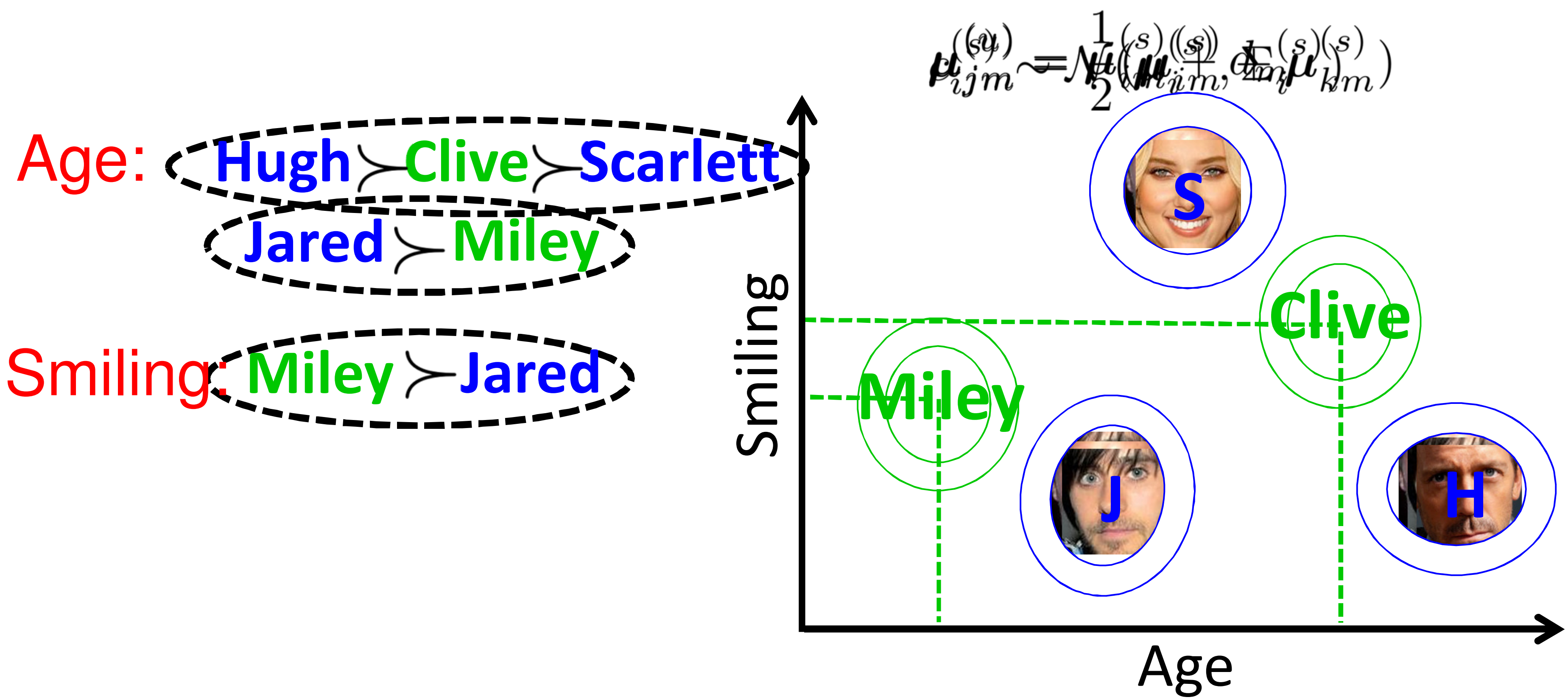
Smiling: **Miley** $\succ$ **Jared**

Need not use all attributes, nor all seen categories

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

Relative zero-shot learning

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



Infer image category using max-likelihood

Datasets

Outdoor Scene Recognition (OSR) [Oliva 2001]

8 classes, ~2700 images, Gist

6 attributes: open, natural, etc.

Public Figures Faces (PubFig) [Kumar 2009]

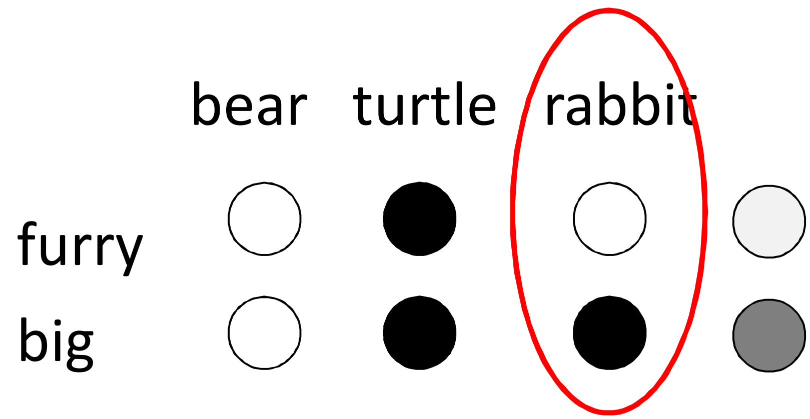
8 classes, ~800 images, Gist+color

11 attributes: white, chubby, etc.

Attributes labeled at category level

Baselines

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

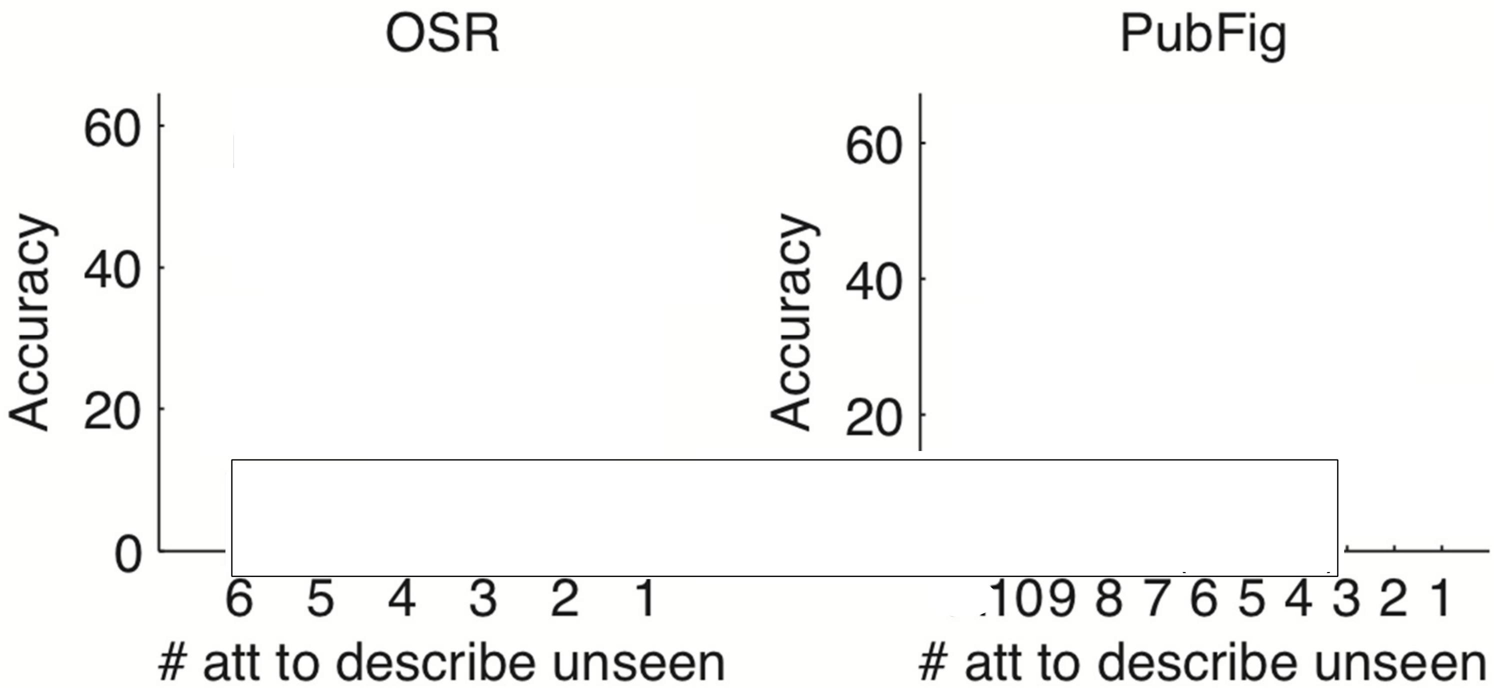


- Relative attributes via
  classifier scores

Relative zero-shot learning

- Robustness:
  - Fewer comparisons to train relative attributes
  - More unseen (fewer seen) categories
- Flexibility in supervision:
  - ‘Looseness’ in description of unseen
  - Fewer attributes used to describe unseen

Relative zero-shot learning



An attribute is more discriminative when used relatively

Our idea: Teach with visual comparisons

We propose **relative attributes** to represent *relationships* between classes, images, and their properties.

→ Enable new modes of human-system communication

- **Training through descriptions**:

  "Rabbits are furrier than dogs."

- **Rationales to explain image labels:**

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

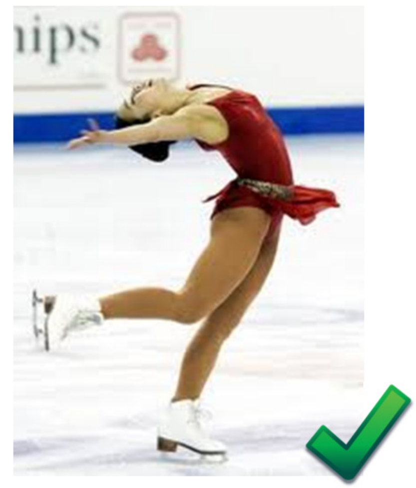
- **Semantic relative feedback for image search:**

  "I want shoes like these, but shinier."

Complex visual recognition tasks

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?

Our idea:

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

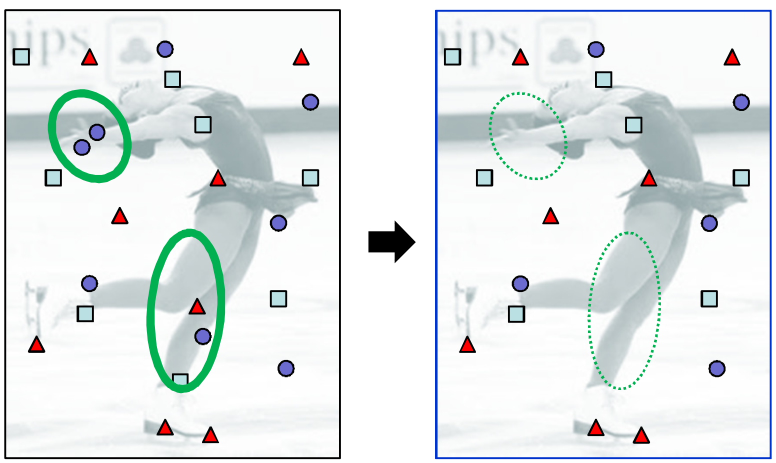
[Donahue & Grauman, ICCV 2011]

Soliciting visual rationales

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

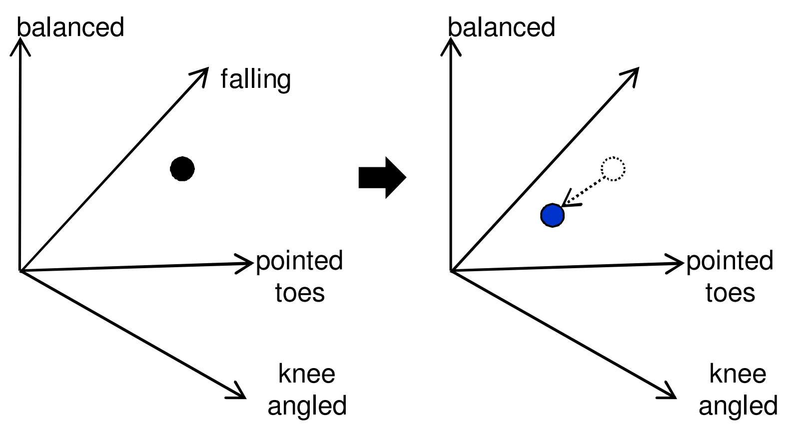
- [x] pointed toes
- [x] balanced
- [ ] falling
- [ ] knee angled

Spatial rationale



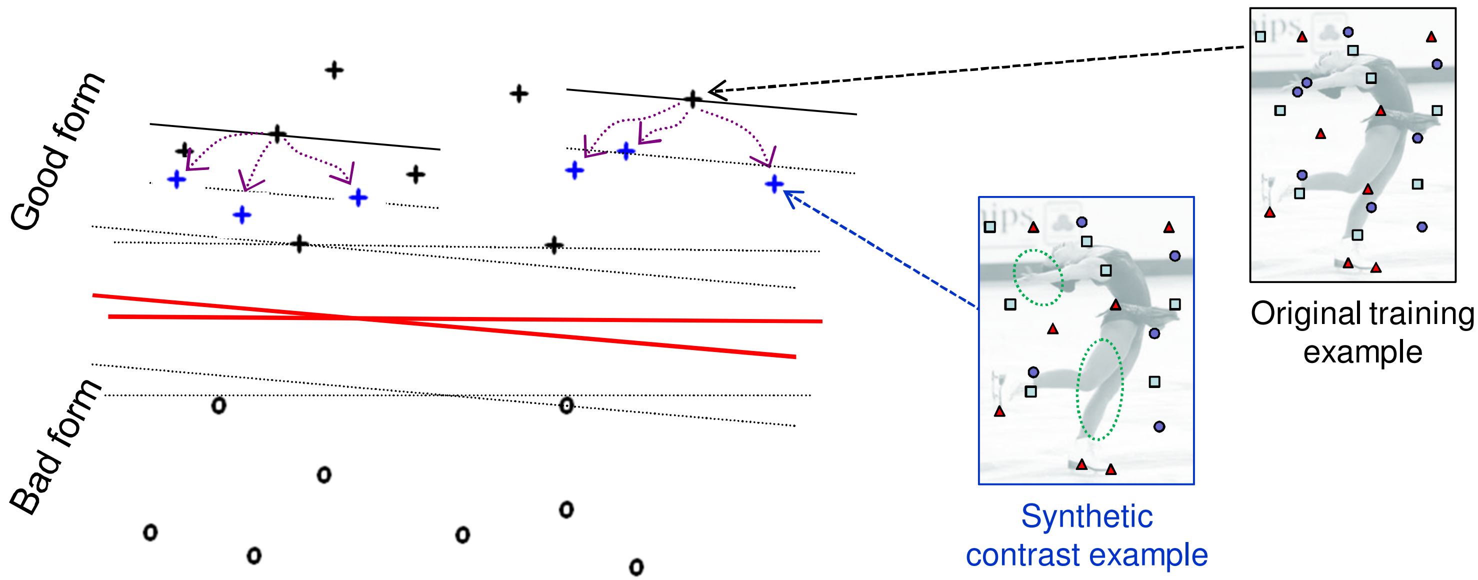
Synthetic contrast example

Attribute rationale



Synthetic contrast example

Rationales' influence on the classifier



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

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

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

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

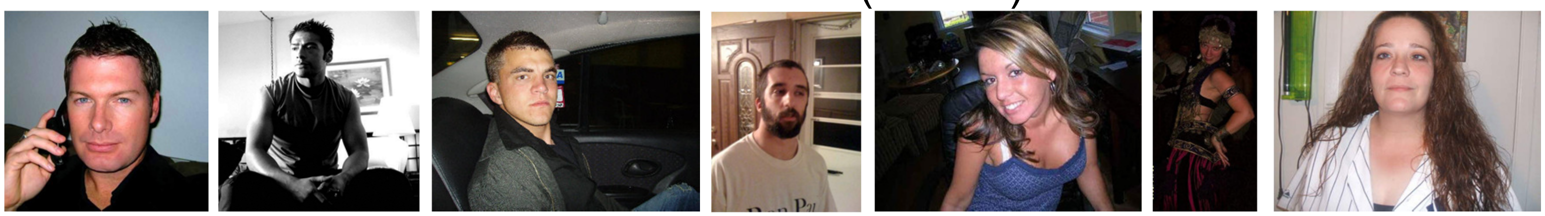
$$\xi_i \geq 0; \gamma_i \geq 0,$$

[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)?



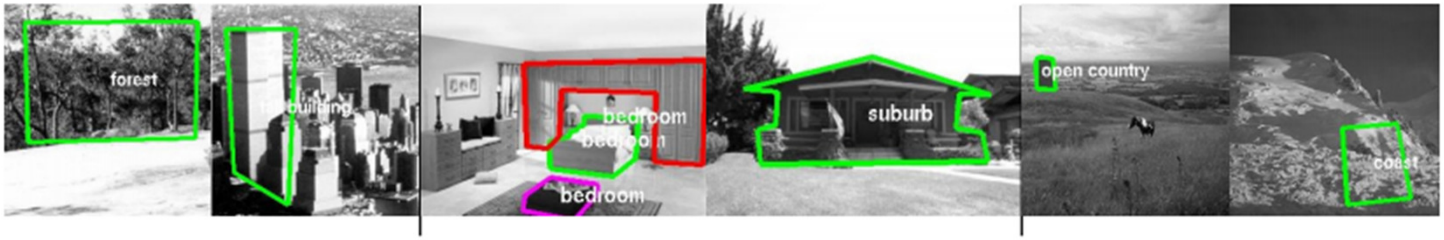
- **Public Figures**: What attributes make them (un)attractive?

Collect rationales from hundreds of MTurk workers.

[Donahue & Grauman, ICCV 2011]

Example rationales from MTurk

Scene categories

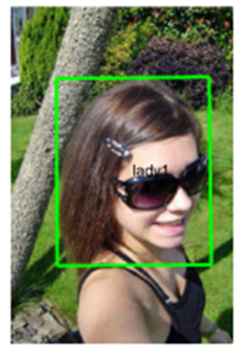


Typical | Tight | "Artistic"

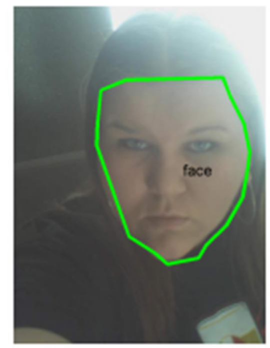
Hot or Not

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

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

How do spatial rationales differ from foreground segmentation?

Scenes	Originals	+Rationales	Rationales only
Kitchen	0.1196	**0.1395**	0.1277
Living Rm	0.1142	**0.1238**	0.1131
Inside City	0.1299	**0.1487**	0.1394
Coast	0.4243	**0.4513**	0.4205
Highway	0.2240	**0.2379**	0.2221
Bedroom	0.3011	**0.3167**	0.2611
Street	0.0778	**0.0790**	0.0766
Country	0.0926	**0.0950**	0.0946
Mountain	0.1154	**0.1158**	0.1151
Office	0.1051	**0.1052**	0.1051
Tall Building	0.0688	**0.0689**	0.0689
Store	0.0866	**0.0867**	0.0857
Forest	0.3956	**0.4006**	0.4004

Mean AP

[Donahue & Grauman, ICCV 2011]

Rationale results




How do spatial rationales differ from foreground segmentation?

Why not just use discriminative feature selection?

Scenes	Originals	+Rationales	Rationales only	Mutual information
Kitchen	0.1196	**0.1395**	0.1277	0.1202
Living Rm	0.1142	**0.1238**	0.1131	0.1159
Inside City	0.1299	**0.1487**	0.1394	0.1245
Coast	0.4243	**0.4513**	0.4205	0.4129
Highway	0.2240	**0.2379**	0.2221	0.2112
Bedroom	0.3011	**0.3167**	0.2611	0.2927
Street	0.0778	**0.0790**	0.0766	0.0775
Country	0.0926	**0.0950**	0.0946	0.0941
Mountain	0.1154	**0.1158**	0.1151	0.1154
Office	0.1051	**0.1052**	0.1051	0.1048
Tall Building	0.0688	**0.0689**	0.0689	0.0686
Store	0.0866	**0.0867**	0.0857	0.0866
Forest	0.3956	**0.4006**	0.4004	0.3897

Mean AP

[Donahue & Grauman, ICCV 2011]

Our idea: Teach with visual comparisons

We propose **relative attributes** to represent *relationships* between classes, images, and their properties.

→ Enable new modes of human-system communication

- **Training through descriptions**:

  "Rabbits are furrier than dogs."

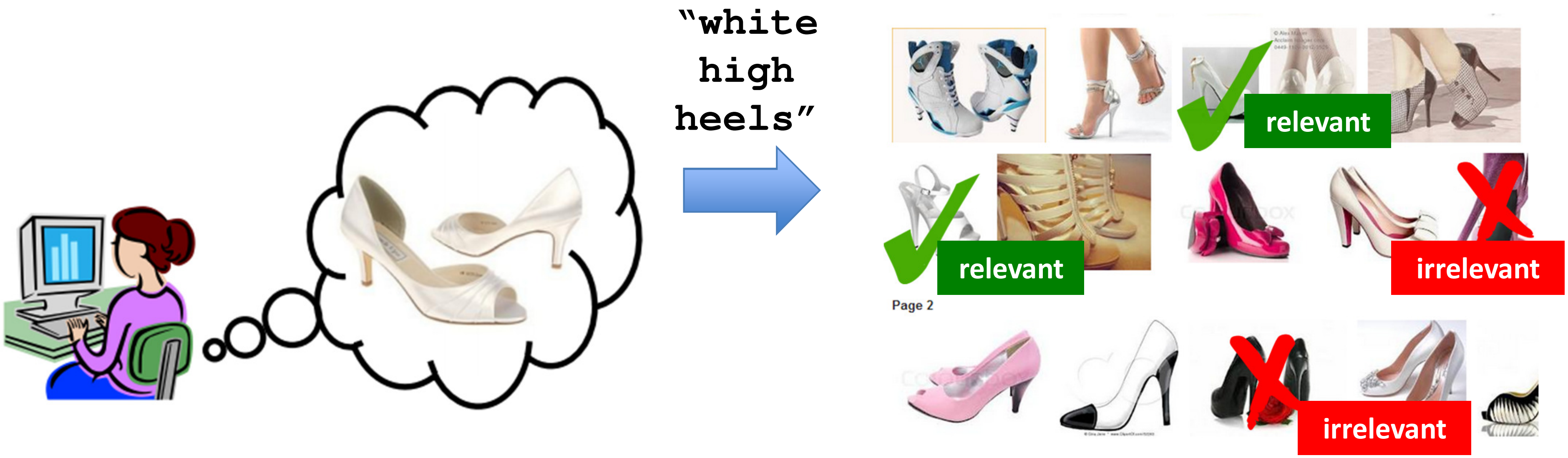
- **Rationales to explain image labels:**

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

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

  "I want shoes like these, but shinier."

Content-based image search



- **Semantic gap** between low-level visual features and high-level user concepts→ impedes search

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

Our idea: Image search refinement via relative attribute feedback

Initial top search results

User communicates target visual concept precisely in semantic terms---***feedback beyond labels***.

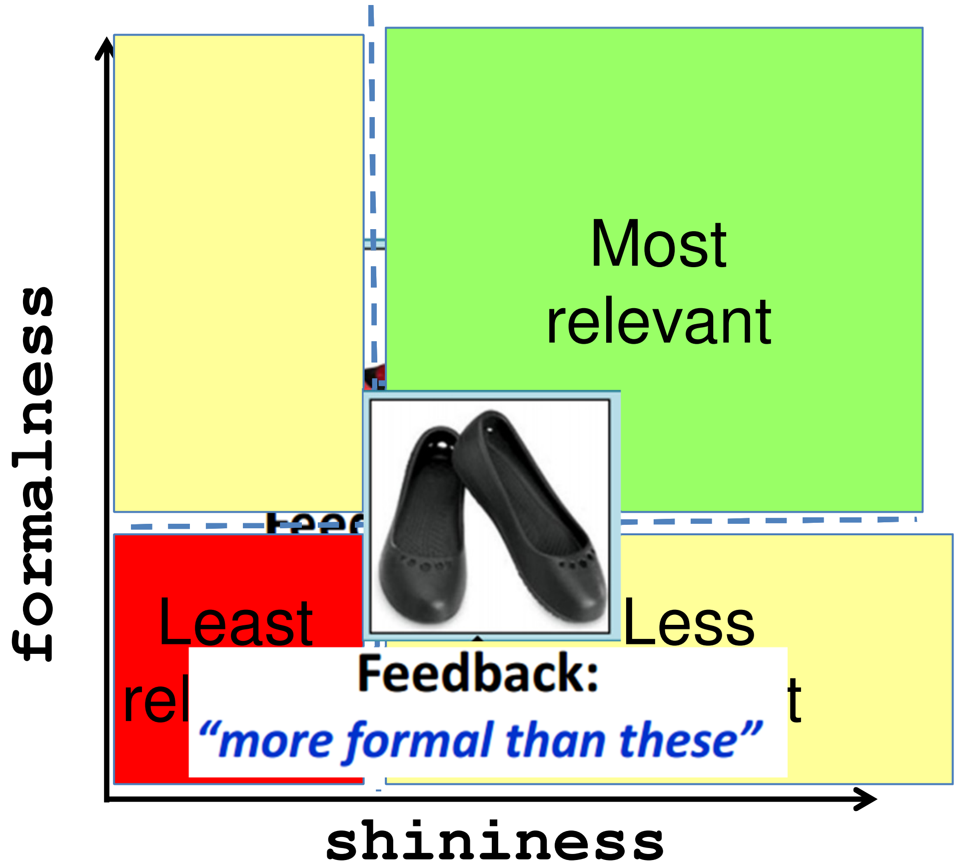
[Kovashka, Parikh,& Grauman, CVPR 2012]

Approach: Whittle Search

1. Rank images by their number of satisfied constraints

2. Iterate, displaying top-ranked images as new reference examples

(To integrate *both* binary and relative feedback, learn relevance ranker.)



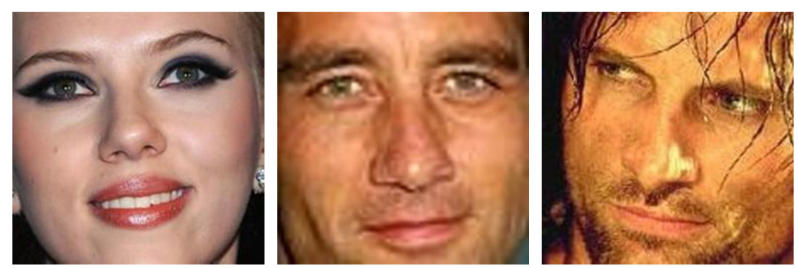
[Kovashka, Parikh,& Grauman, CVPR 2012]

Datasets

Shoes – 14,658 images from Attribute Discovery dataset [Berg et al.]
10 attributes (we added)

Scenes – 2,688 images from Outdoor Scene Recognition [Oliva et al.]
6 attributes

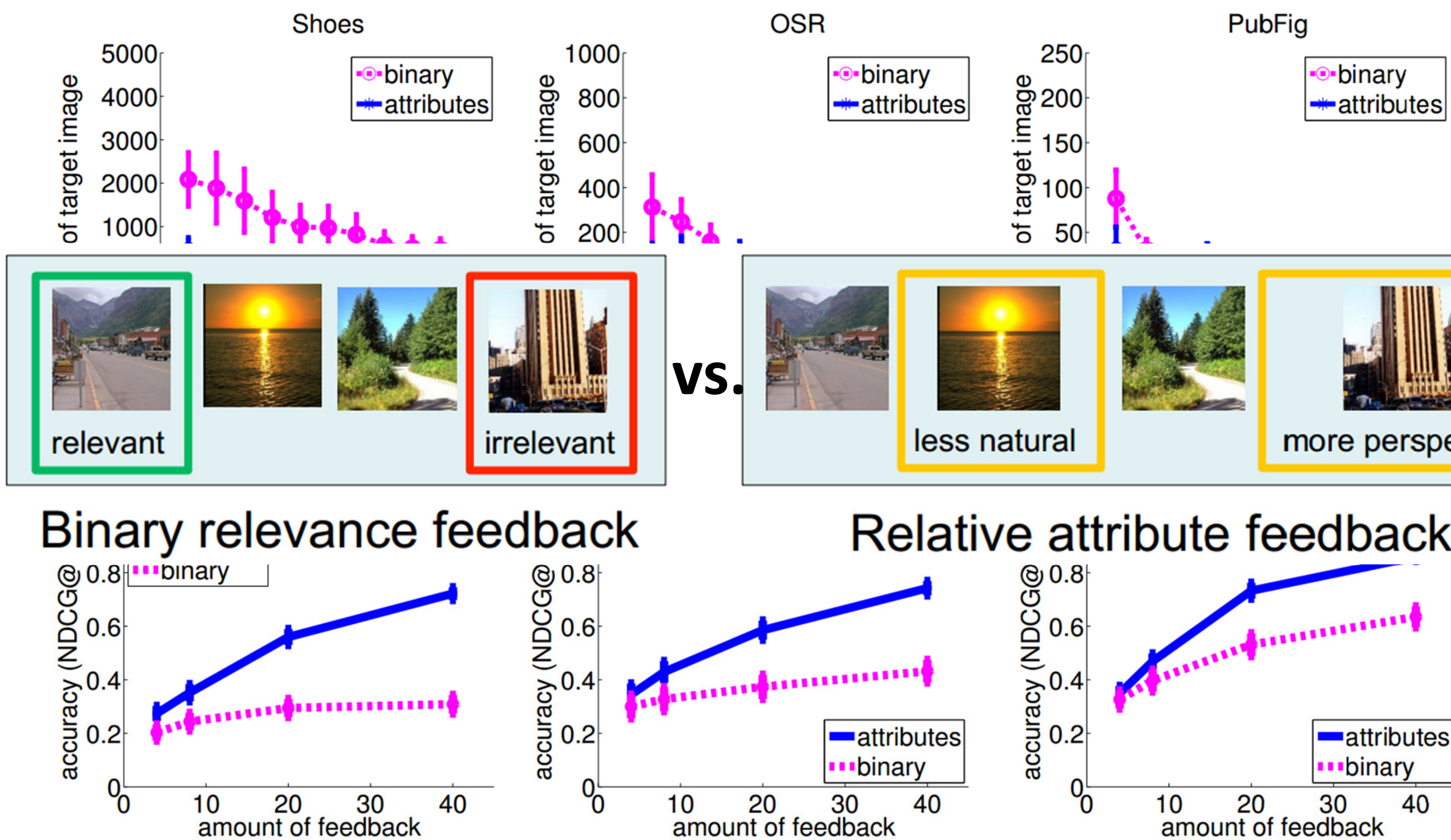
Faces– 772 images from Public Figures [Kumar et al.]
11 attributes;



Features: GIST+color

[Kovashka, Parikh,& Grauman, CVPR 2012]

Results: Whittle Search



Richer feedback → faster gains per unit of user effort.

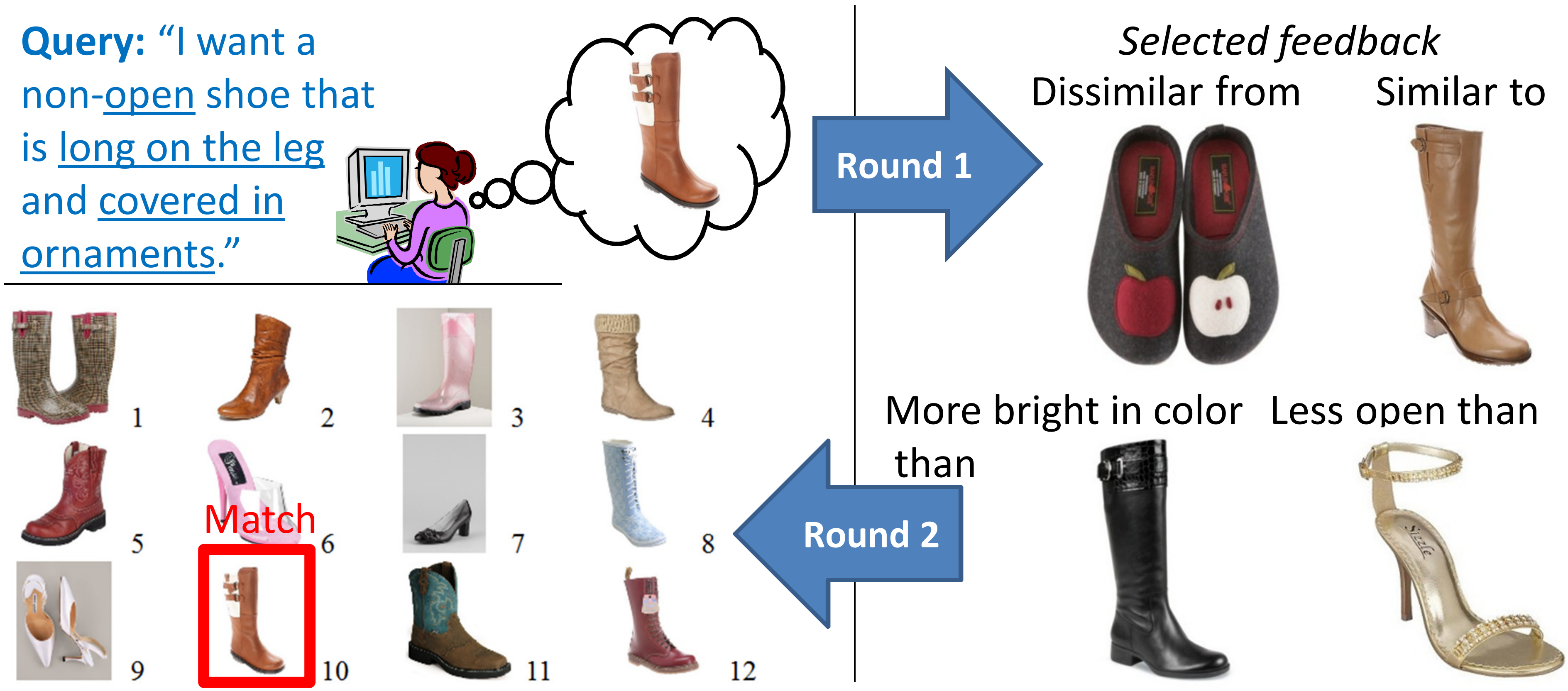
[Kovashka, Parikh,& Grauman, CVPR 2012]

Results: Whittle Search

[Kovashka, Parikh,& Grauman, CVPR 2012]

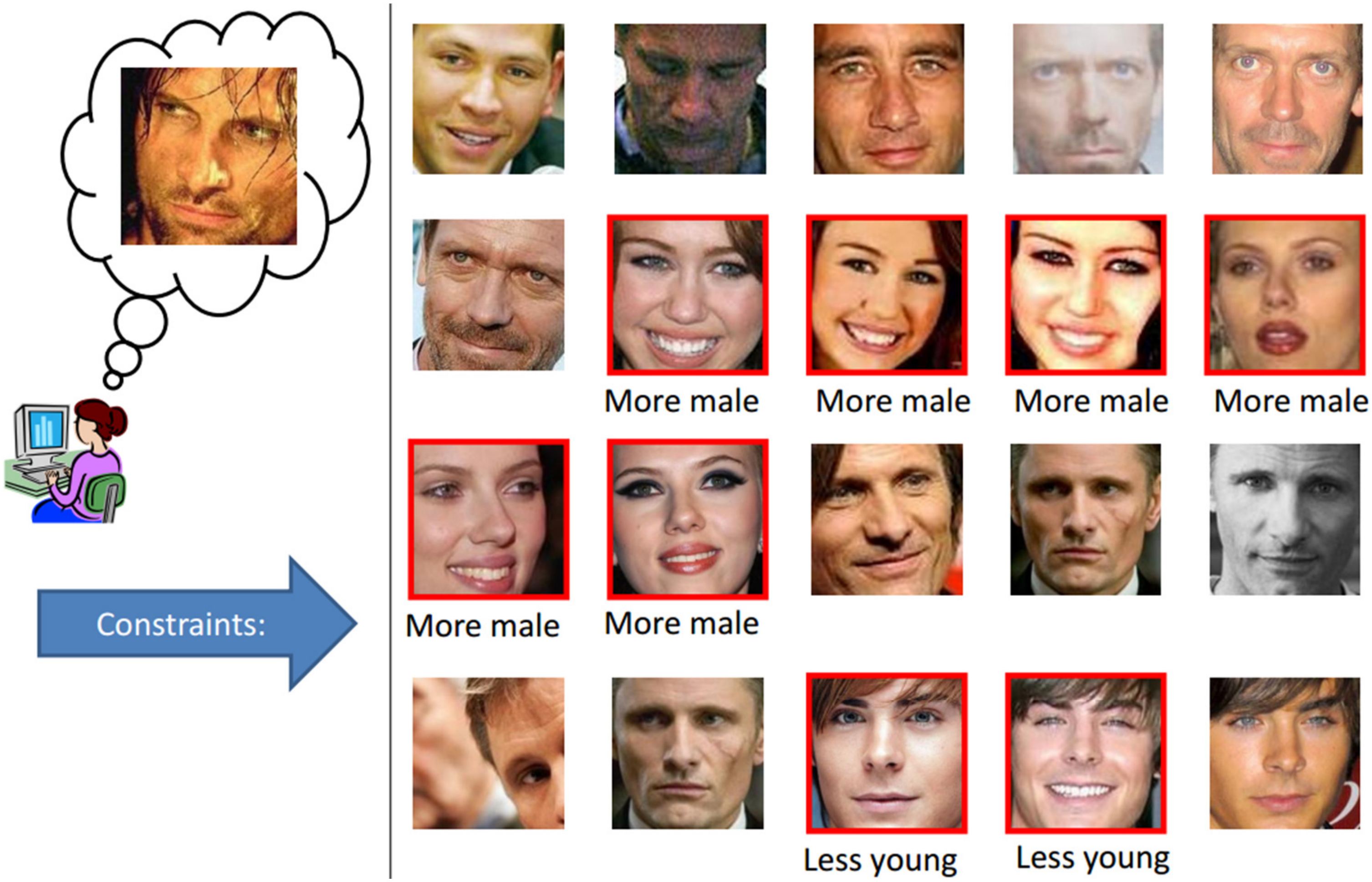
Results: Whittle Search

Hybrid feedback example



[Kovashka, Parikh,& Grauman, CVPR 2012]

Results: Whittle Search



[Kovashka, Parikh,& Grauman, CVPR 2012]

Summary

- Humans are not simply “label machines”
- Widen access to visual knowledge by modeling visual comparisons
- Relative attributes enable new applications for recognition and visual search