



Inferring What's Important in Image Search

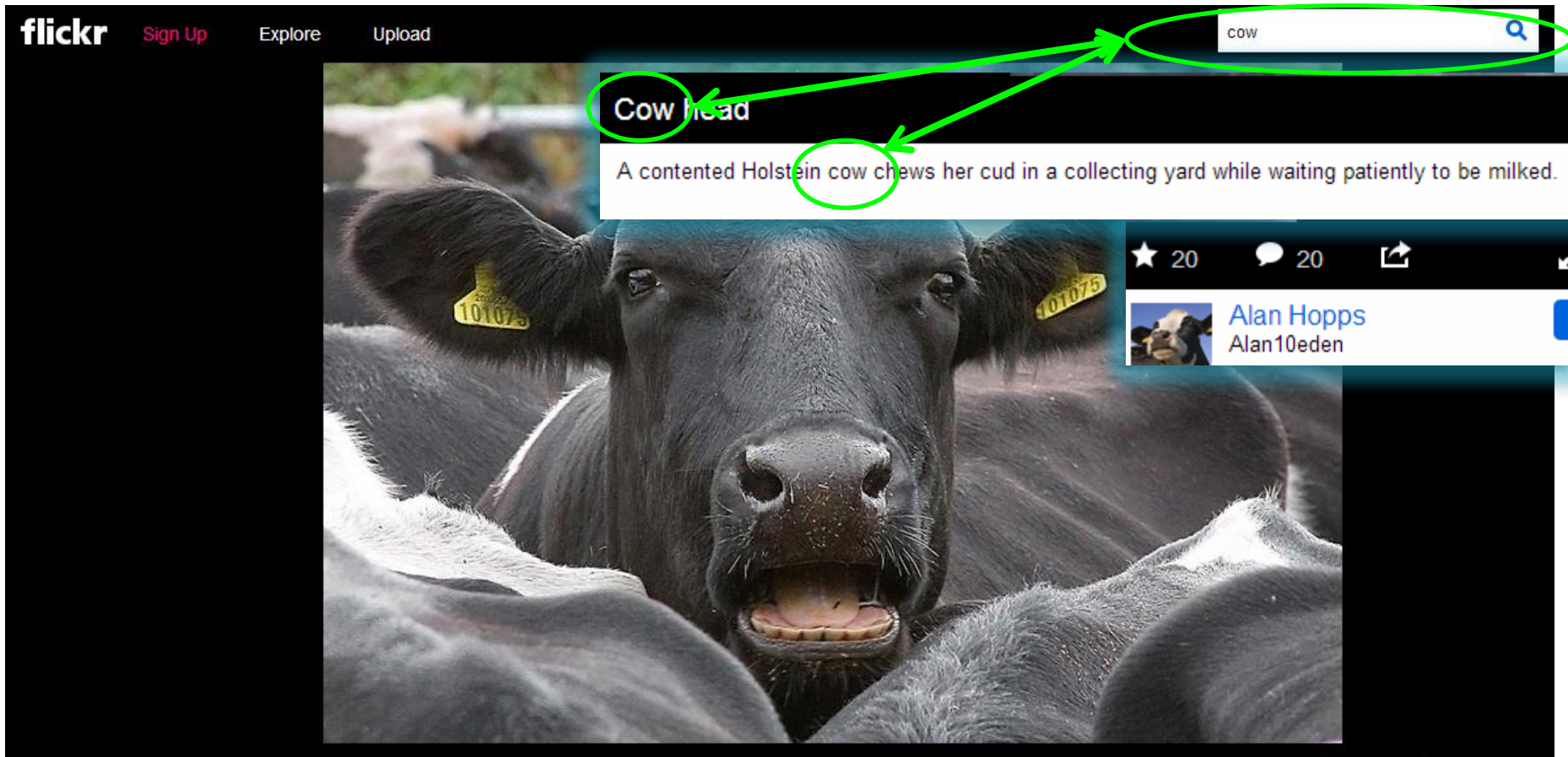
Kristen Grauman

University of Texas at Austin

With Adriana Kovashka, Devi Parikh, and
Sung Ju Hwang

“Visual” search 1.0

- Associate images by keywords and meta-data



Visual search 2.0

- Auto-annotate images with relevant keywords: objects, attributes, scenes, visual concepts...



outdoors

furry

black

[Kumar et al. 2008, Snoek et al. 2006, Naphade et al. 2006, Chang et al. 2006, Vaquero et al. 2009, Berg et al. 2010, and many others...]

Problem

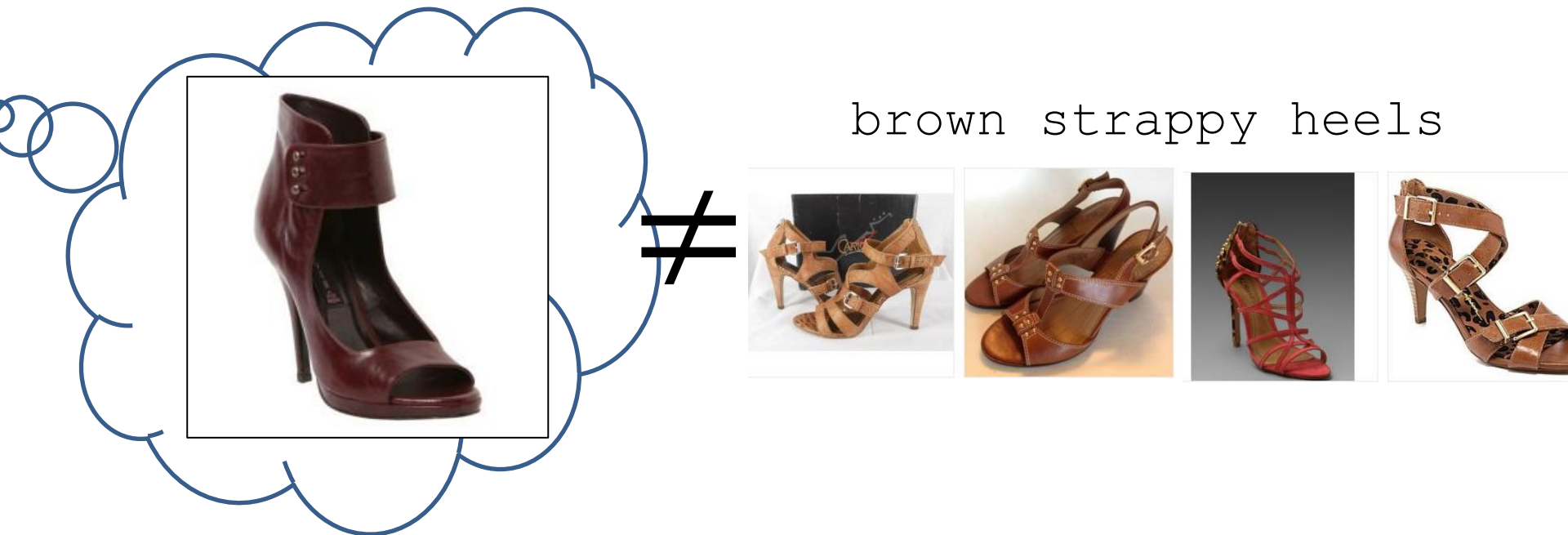
- Fine-grained visual differences *beyond keyword composition* influence image search relevance.



Similar object distributions, yet are they equally relevant?

Problem

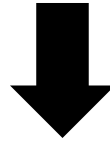
- Fine-grained visual differences *beyond keyword composition* influence image search relevance.



How to capture target with a single description?

Goal

- Fine-grained visual differences *beyond keyword composition* influence image search relevance.



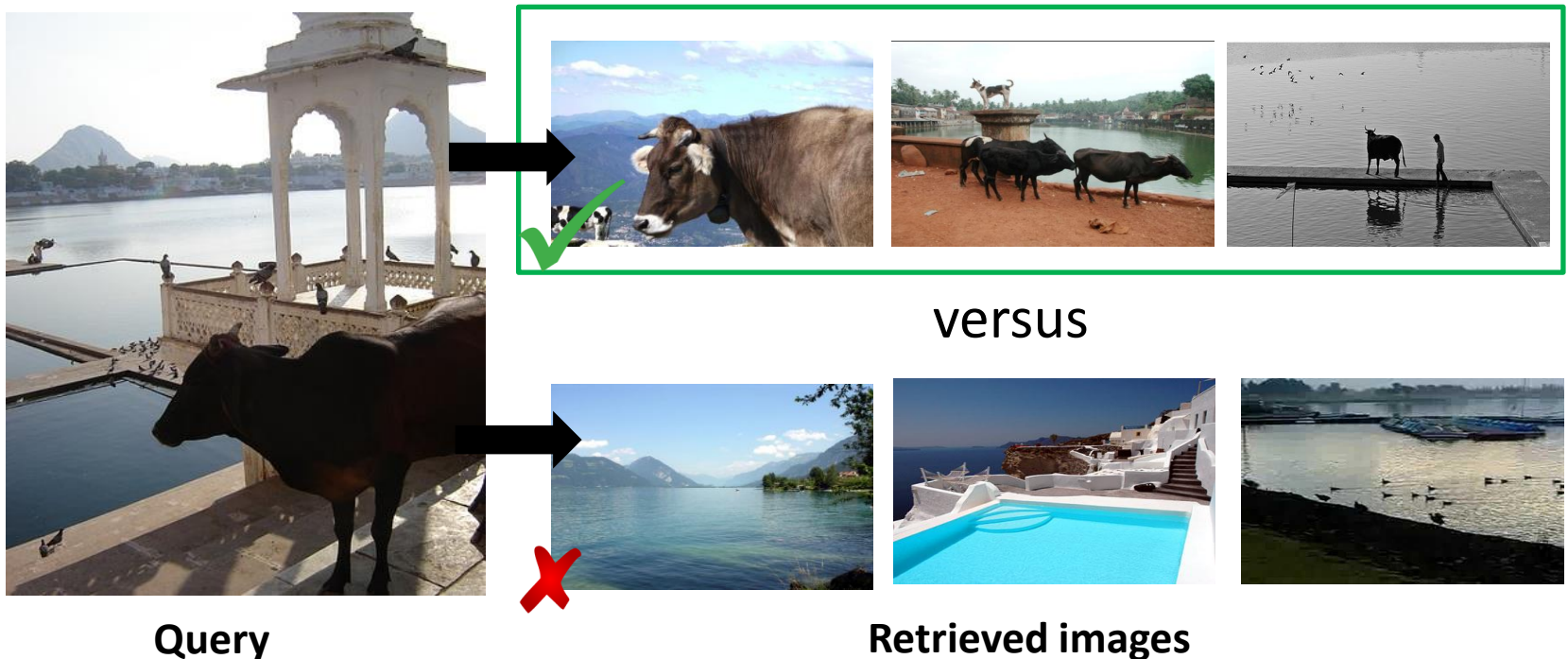
- **Goal:** Account for subtleties in visual relevance
 - **Implicit importance:**
Infer which objects most define the scene
 - **Explicit importance:**
Comparative feedback about which properties are (ir)relevant

Related work

- **Region-noun correspondence** [Duygulu et al. 2002, Barnard et al. 2003, Berg et al. 2004, Gupta & Davis 2008, Li et al. 2009, Hwang & Grauman 2010,...]
- **Dual-view image-text representations** [Monay & Gatica-Perez 2003, Hardoon & Shawe-Taylor 2003, Quattoni et al. 2007, Bekkerman & Jeon 2007, Quack et al. 2008, Blaschko & Lampert 2008, Qi et al. 2009,...]
- **Image description and memorability** [Spain & Perona 2008, Farhadi et al. 2010, Berg et al. 2011, Parikh & Grauman 2011, Isola et al. 2011, Berg et al. 2012]

Capturing relative importance

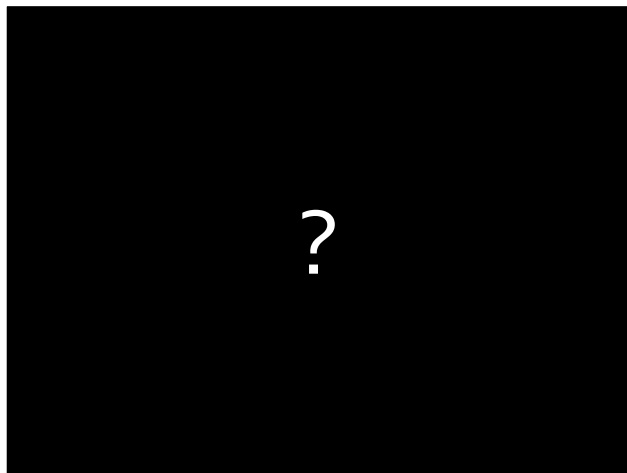
- Object presence $\neq$ importance



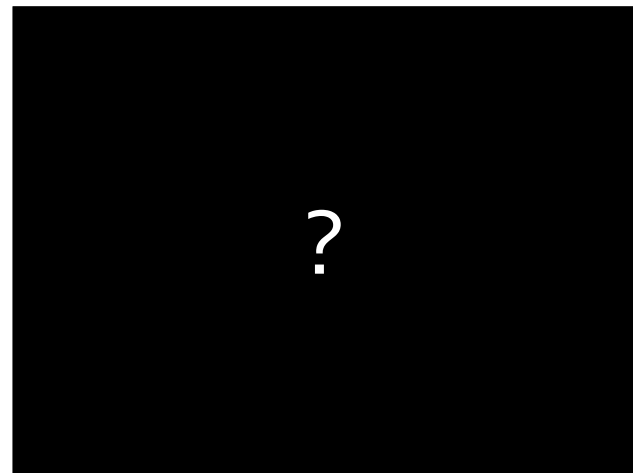
Can we infer what human viewers find most important?

Capturing relative importance

- Intuition: Human-provided tags give useful cues beyond just which objects are present.



Mug
Key
Keyboard
Toothbrush
Pen
Photo
Post-it

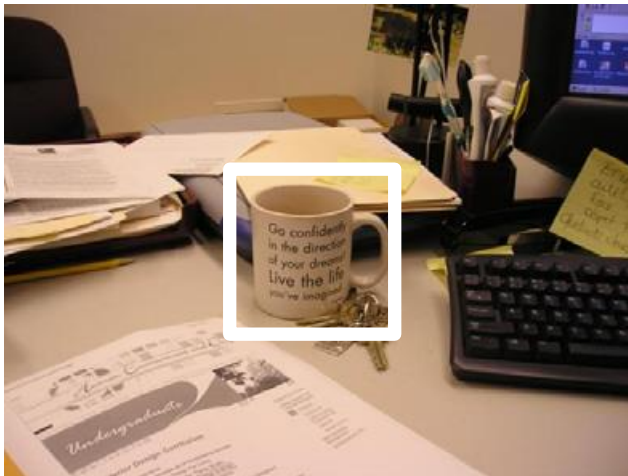


Computer
Poster
Desk
Bookshelf
Screen
Keyboard
Screen
Mug
Poster

Based on tags alone, what can you say about the mug in each image?

Capturing relative importance

- Intuition: Human-provided tags give useful cues beyond just which objects are present.



Mug
Key
Keyboard
Toothbrush
Pen
Photo
Post-it



Computer
Poster
Desk
Bookshelf
Screen
Keyboard
Screen
Mug
Poster

Based on tags alone, what can you say about the mug in each image?

Our idea: Learning implicit importance

- Learn cross-modal representation that accounts for “what to mention” using implicit cues from text

Visual:

- Texture
- Scene
- Color...



TAGS:

Cow

Birds

Architecture

Water

Sky

Textual:

- Frequency
- Relative order
- Mutual proximity

Training: human-given descriptions

Our idea: Learning implicit importance

- Learn cross-modal representation that accounts for “what to mention” using implicit cues from text

Visual

- Text
- Scene
- Color

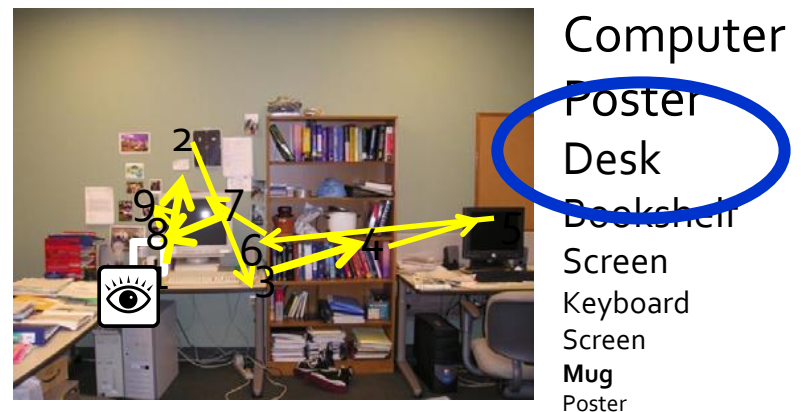
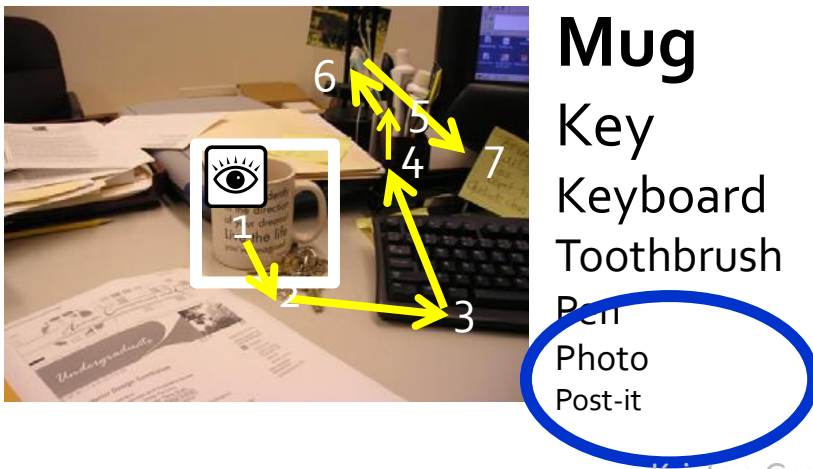
Importance = how likely an object is named early on by a human describing an image.

cy
order
proximity

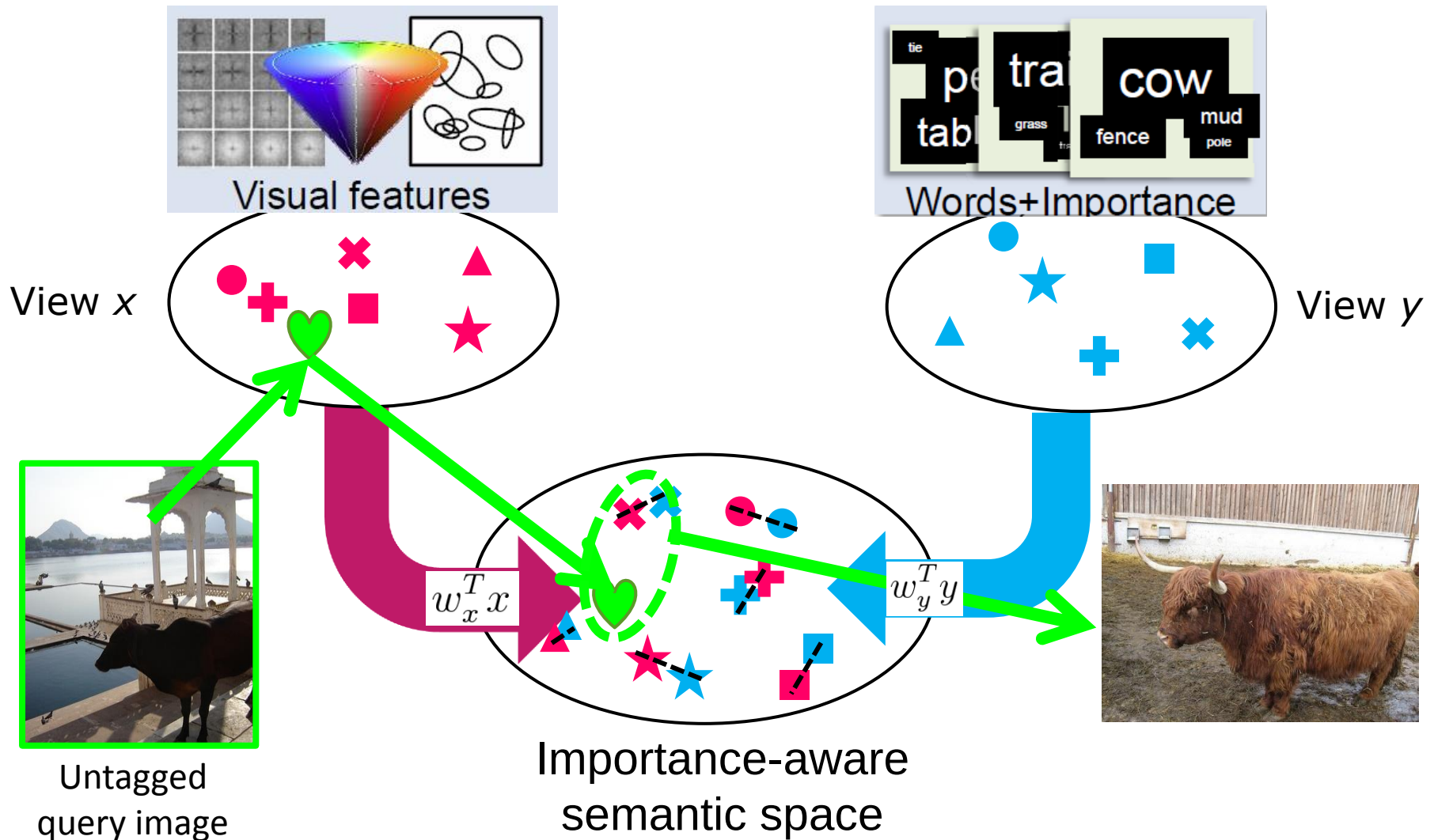
Training: human-given descriptions

Implicit tag features

- Presence or absence of other objects affects the scene layout → record bag-of-words **frequency**.
- People tag the “important” objects earlier → record **rank** of each tag compared to its typical rank.
- People tend to move eyes to nearby objects after first fixation → record **proximity** of all tag pairs.



Learning an importance-aware semantic space



Learning an importance-aware semantic space

Linear CCA: Given paired data $\{(x_1, y_1), \dots, (x_N, y_N)\}$,

Select projection bases:

$$w_x^*, w_y^* = \arg \max_{w_x, w_y} \frac{\hat{E}[\langle x, w_x \rangle \langle y, w_y \rangle]}{\sqrt{\hat{E}[\langle x, w_x \rangle^2] \hat{E}[\langle y, w_y \rangle^2]}}$$

Kernel CCA: Given pair of kernel functions

$$k_x(x_i, x_j) = \phi_x(x_i)^T \phi_x(x_j),$$

$$k_y(y_i, y_j) = \phi_y(y_i)^T \phi_y(y_j),$$

Same objective, but projections in kernel space:

$$w_x = \sum_i \alpha_i \phi_x(x_i), \quad w_y = \sum_i \beta_i \phi_y(y_i),$$

Assumptions

1. People tend to agree about which objects most define a scene.
2. Significance of those objects in turn influences the order in which they are mentioned.

Evidence from previous studies that these hold:

[von Ahn & Dabbish 2004, Tatler et al. 2005, Spain & Perona 2008, Einhauser et al. 2008, Elazary & Itti 2008, Berg et al. 2012]

Image+text datasets

- PASCAL VOC 2007 with tags ~10K images
- LabelMe images with tags ~4K images
- PASCAL VOC 2007 with sentences ~500 images

Text data collected on MTurk (~750 unique workers)



Tags

person, person,
son, bicycle,
bag, tree,
board



Sentences

A motorcyclist getting ready for his journey his baggages are tied in the motorcycle.

Results: Accounting for importance in image search

Visual only



Words + Visual



Query
Image

Our method



Results: Accounting for importance in image search

Visual only



Words + Visual



Our method



Query
Image



Results: Accounting for importance in image search

Visual only



Words + Visual



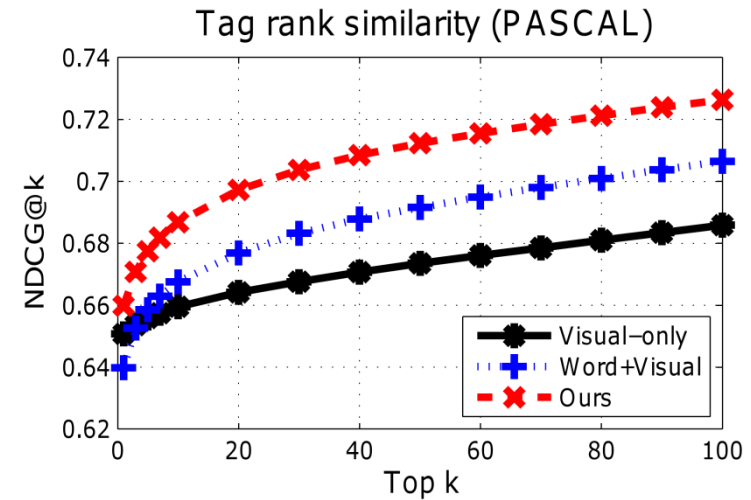
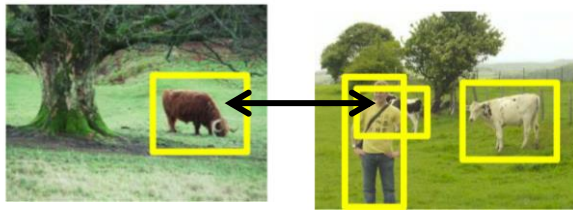
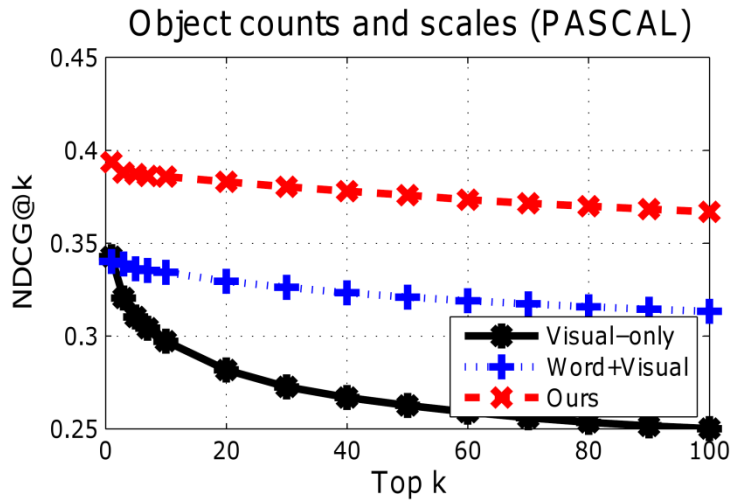
Our method



Query
Image



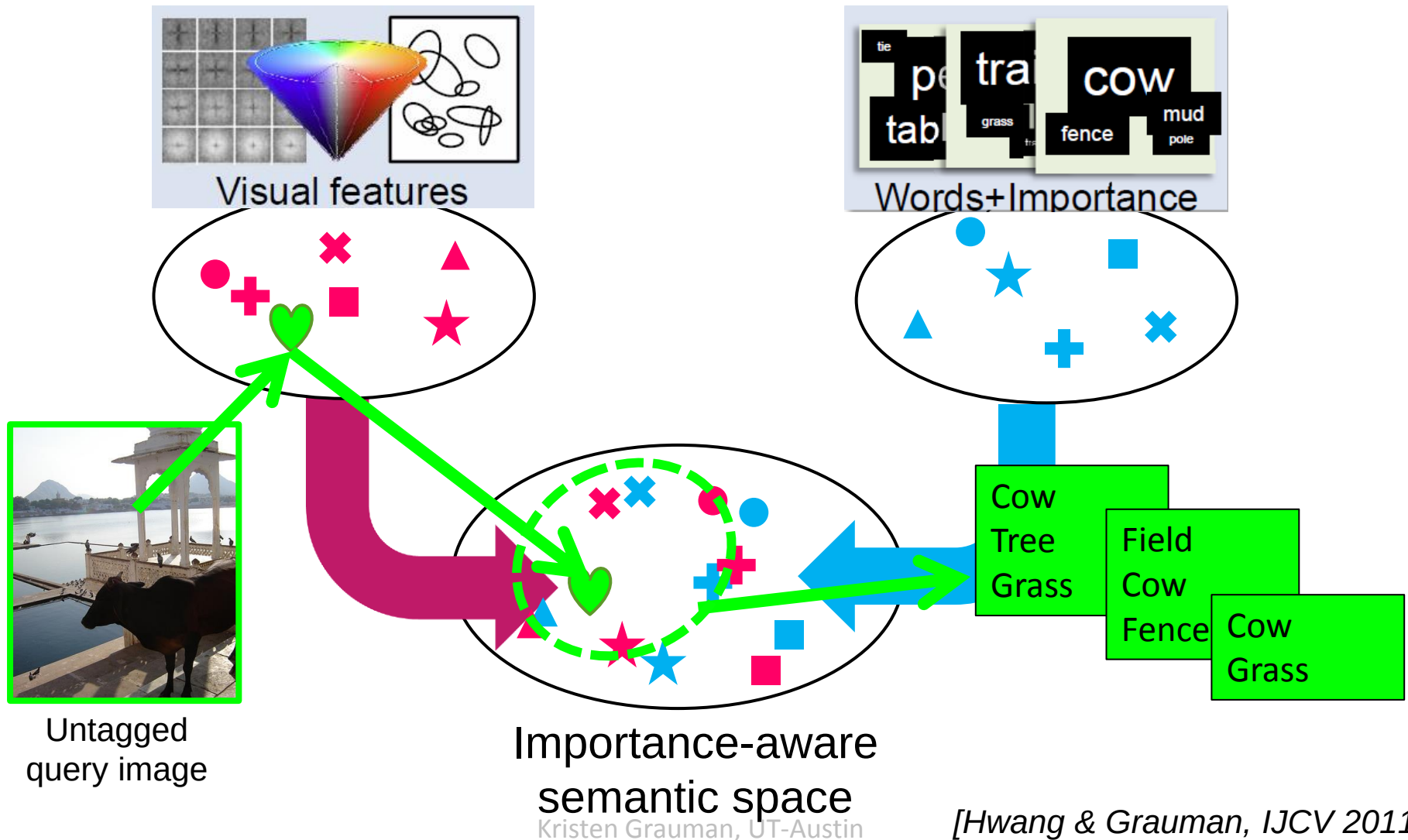
Results: Accounting for importance in image search



Our method better retrieves images that share the query's important objects

Auto-tagging

We can also predict **descriptions** for novel images



Results: Accounting for importance in auto-tagging

We can also predict **descriptions** for novel images



Boat
Person
Water
Sky
Rock



Tree
Boat
Grass
Water
Person



Bottle
Knife
Napkin
Light
Fork



Person
Tree
Car
Chair
Window

What do human judges think?

relevantimage

Requester: Chad Barnes

Qualifications Required: HIT approval rate (%) is greater than 25

: 60 minutes

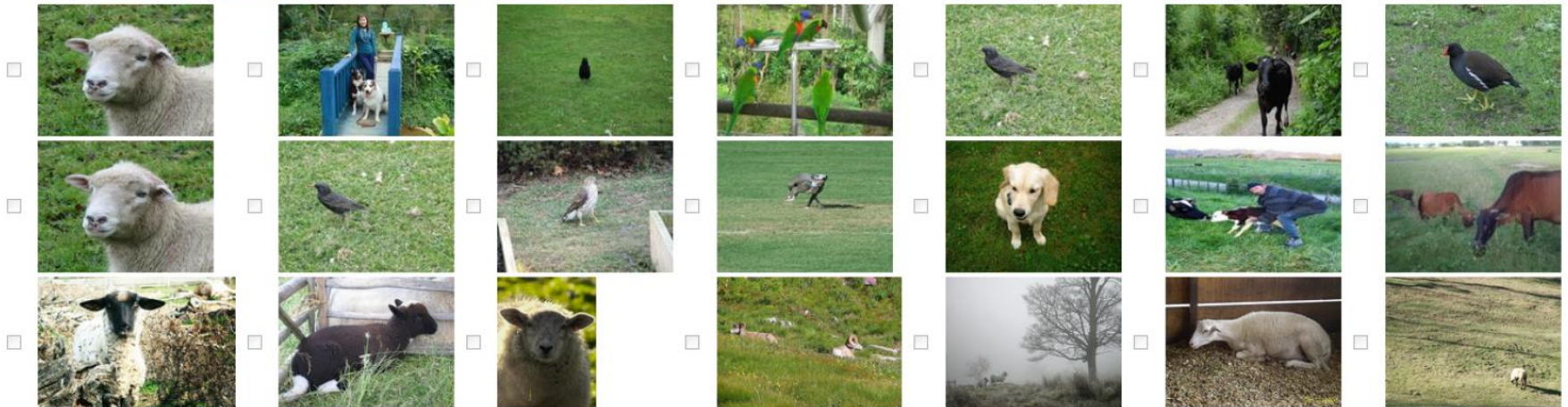
Before starting to work, please read the instruction carefully

First look at the topmost Query image. Then select those images below that

When you have selected all relevant images, press the Submit button at the bottom of the page. If you are poorly done.

Select those images below that contain the “most important” objects seen in the query.

Query Image

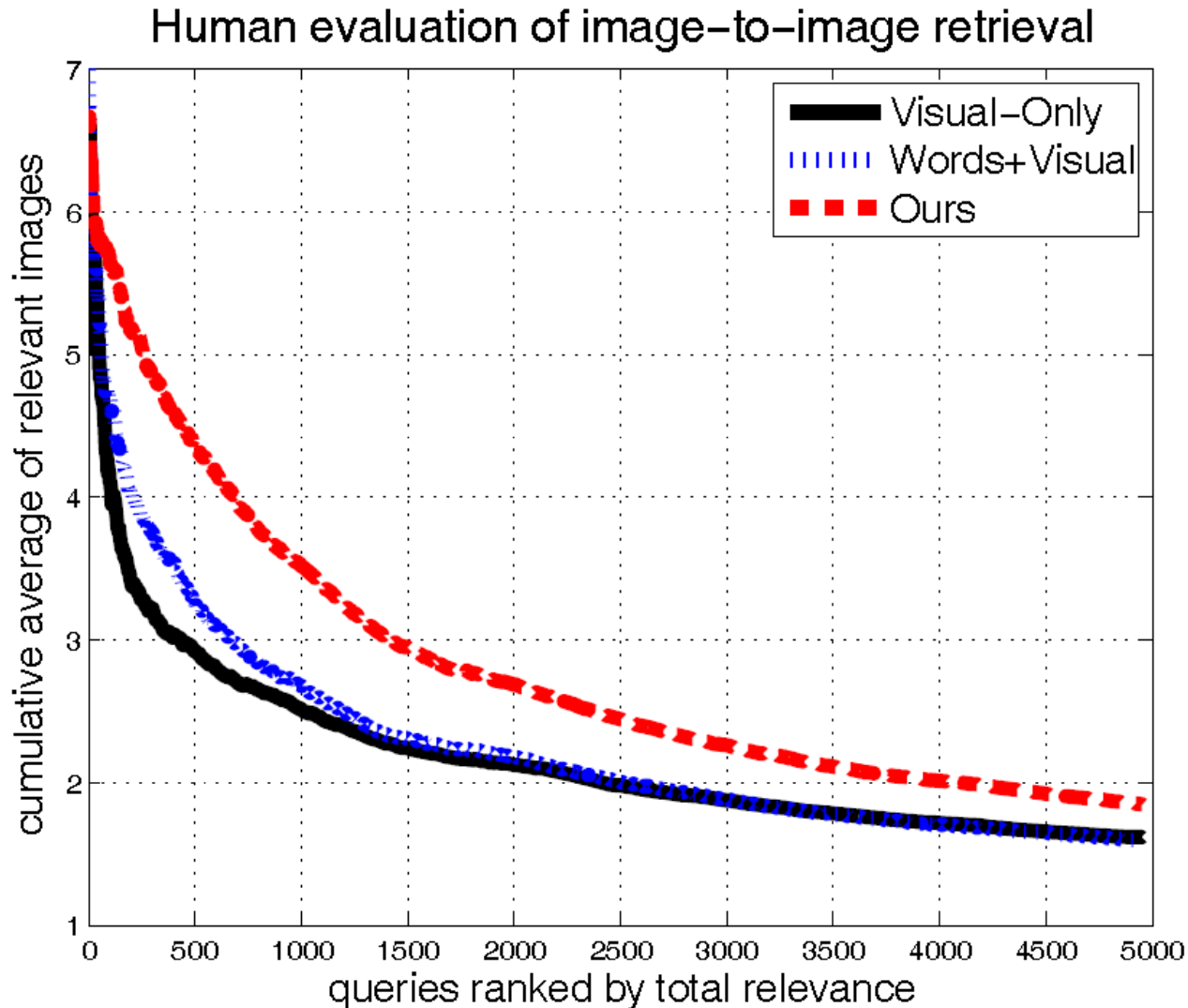


You must ACCEPT the HIT first.

Want to work on this HIT?

Kristen Grainger, UT-Austin

What do human judges think?

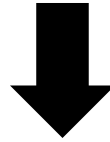


Subjects are
323 MTurk
workers

Require
unanimous
vote among 5
for image to be
considered
relevant

Goal

- Fine-grained visual differences *beyond keyword composition* influence image search relevance.



- **Goal:** Account for subtleties in visual relevance

- **Implicit importance:**

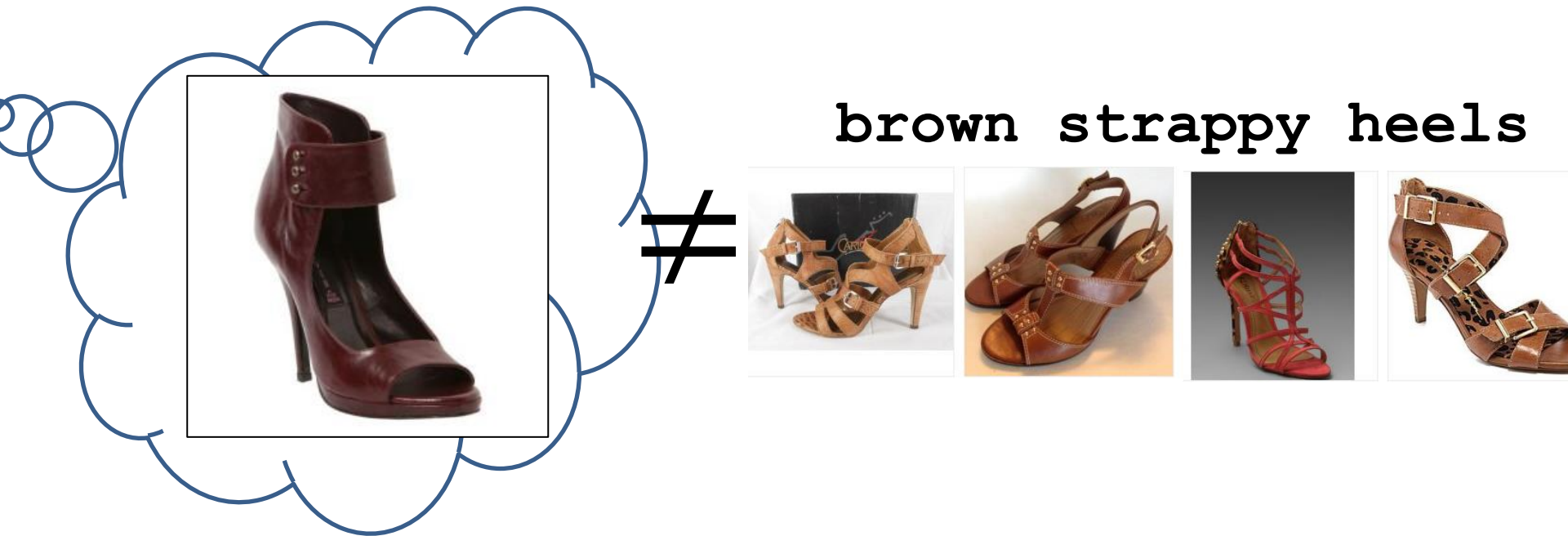
- Infer which objects most define the scene

- **Explicit importance:**

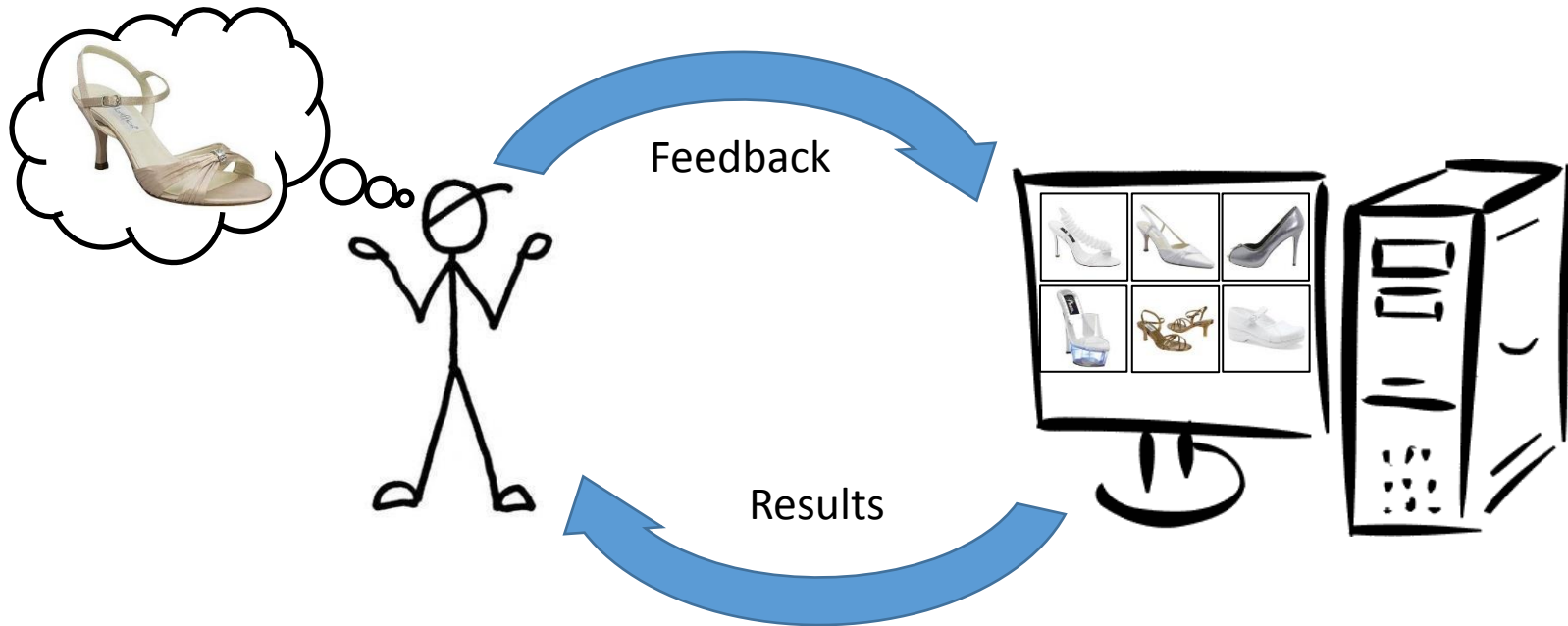
- Comparative feedback about which properties are (ir)relevant

Problem with one-shot visual search

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

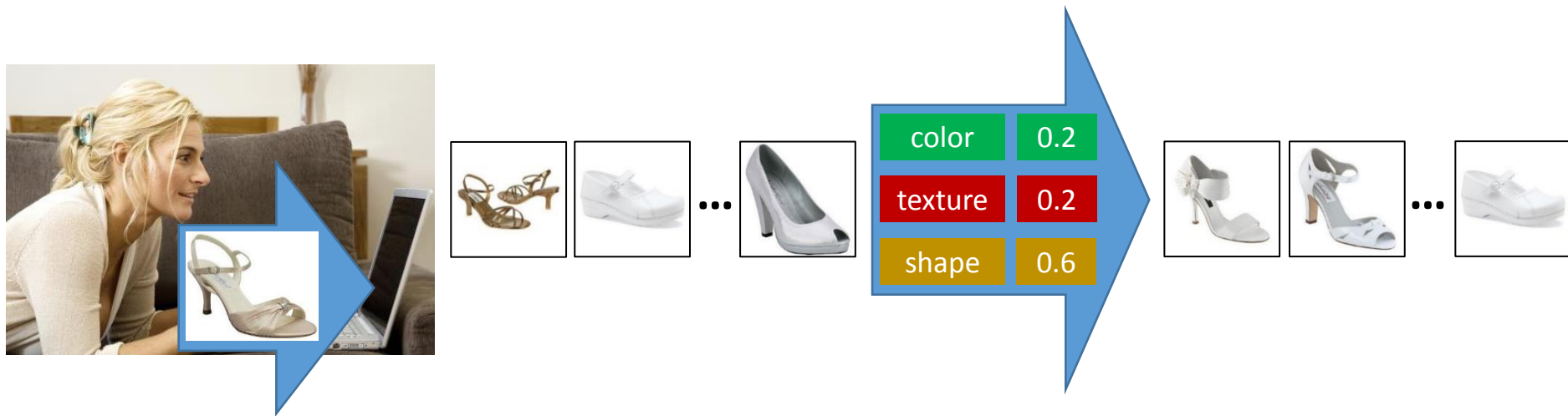


Interactive visual search



- **Iteratively refine** the set of retrieved images based on user feedback on results so far
- Potential to communicate more precisely the desired visual content

Limitations of traditional interactive methods



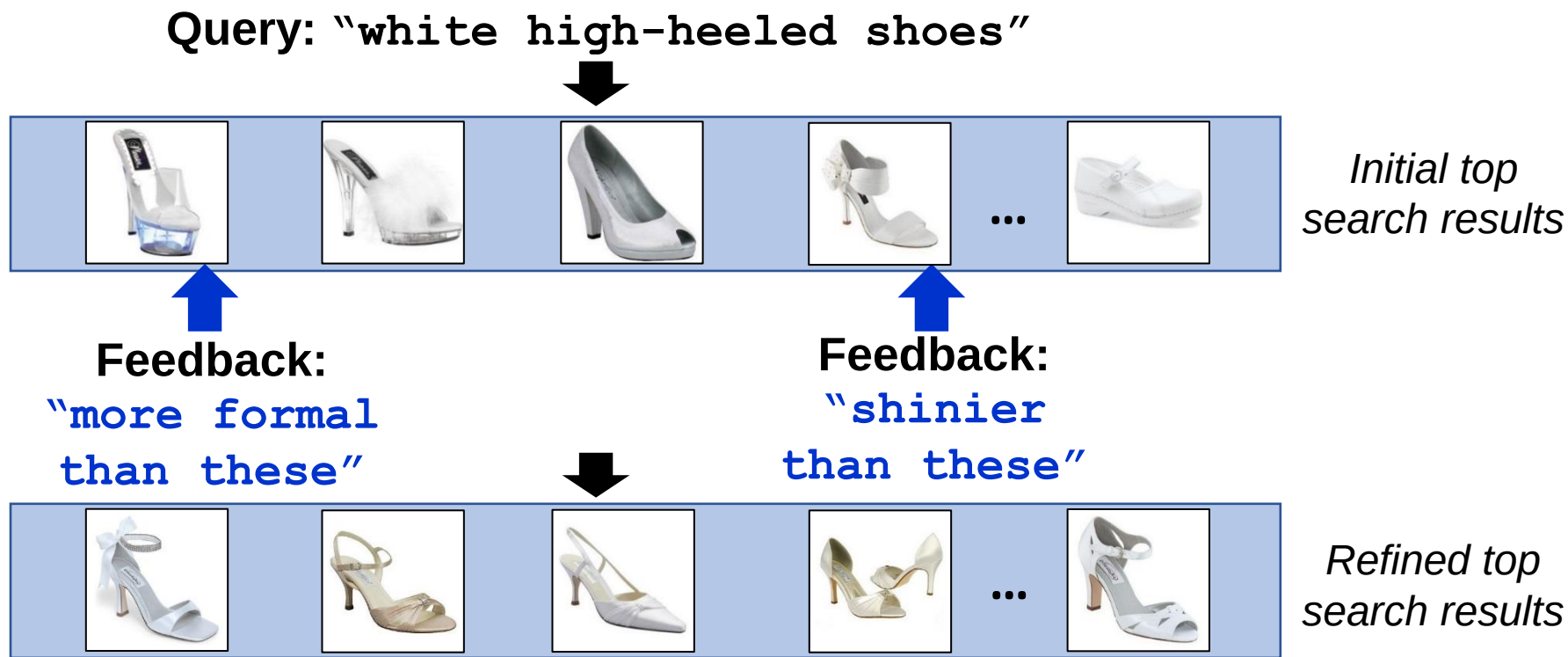
- Tuning system parameters difficult for user
[Flickner et al. 1995, Ma & Manjunath 1997, Iqbal & Aggarwal 2002]

Limitations of traditional interactive methods



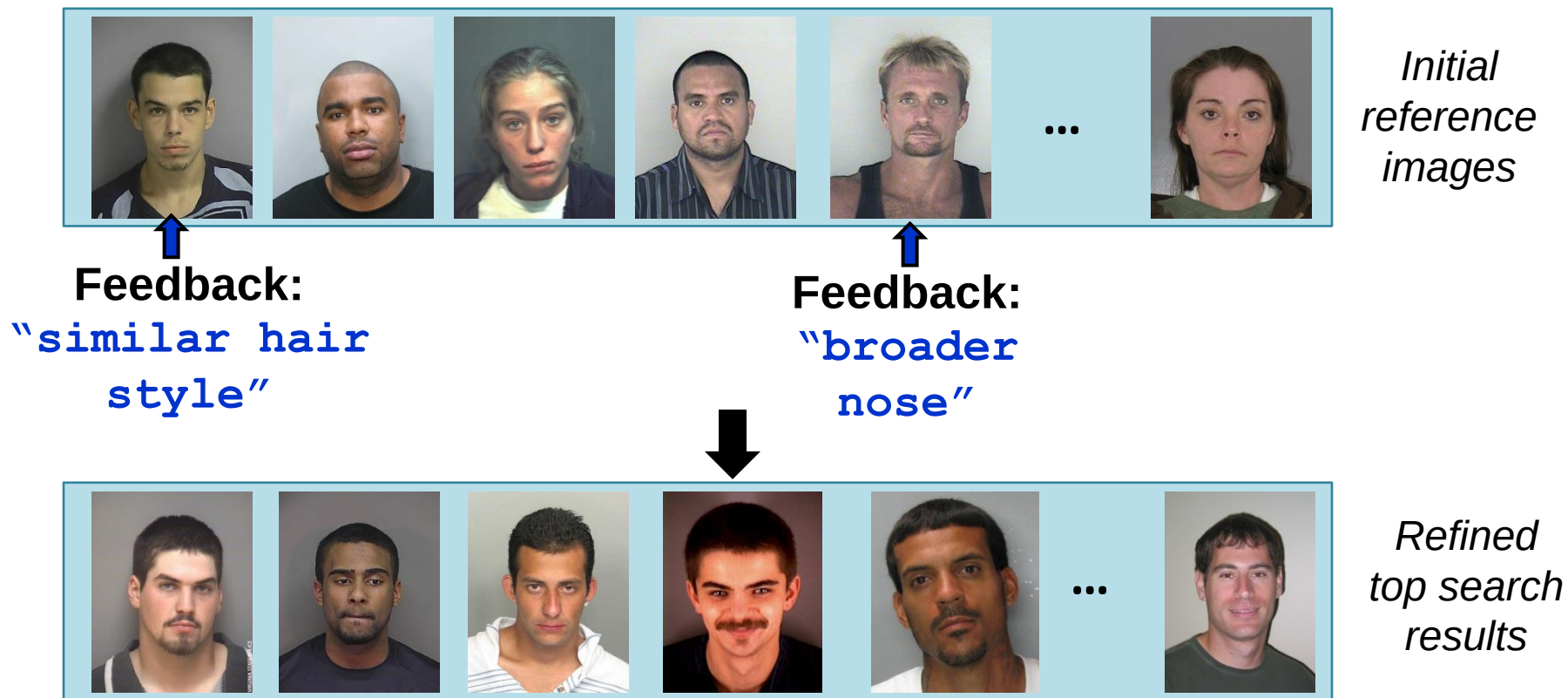
- Tuning system parameters difficult for user
[Flickner et al. 1995, Ma & Manjunath 1997, Iqbal & Aggarwal 2002]
- Traditional binary feedback imprecise
[Rui et al. 1998, Zhou et al. 2003, ...]

WhittleSearch: Relative attribute feedback



Whittle away irrelevant images via precise semantic feedback

WhittleSearch: Relative attribute feedback



Whittle away irrelevant images via precise semantic feedback

Visual attributes

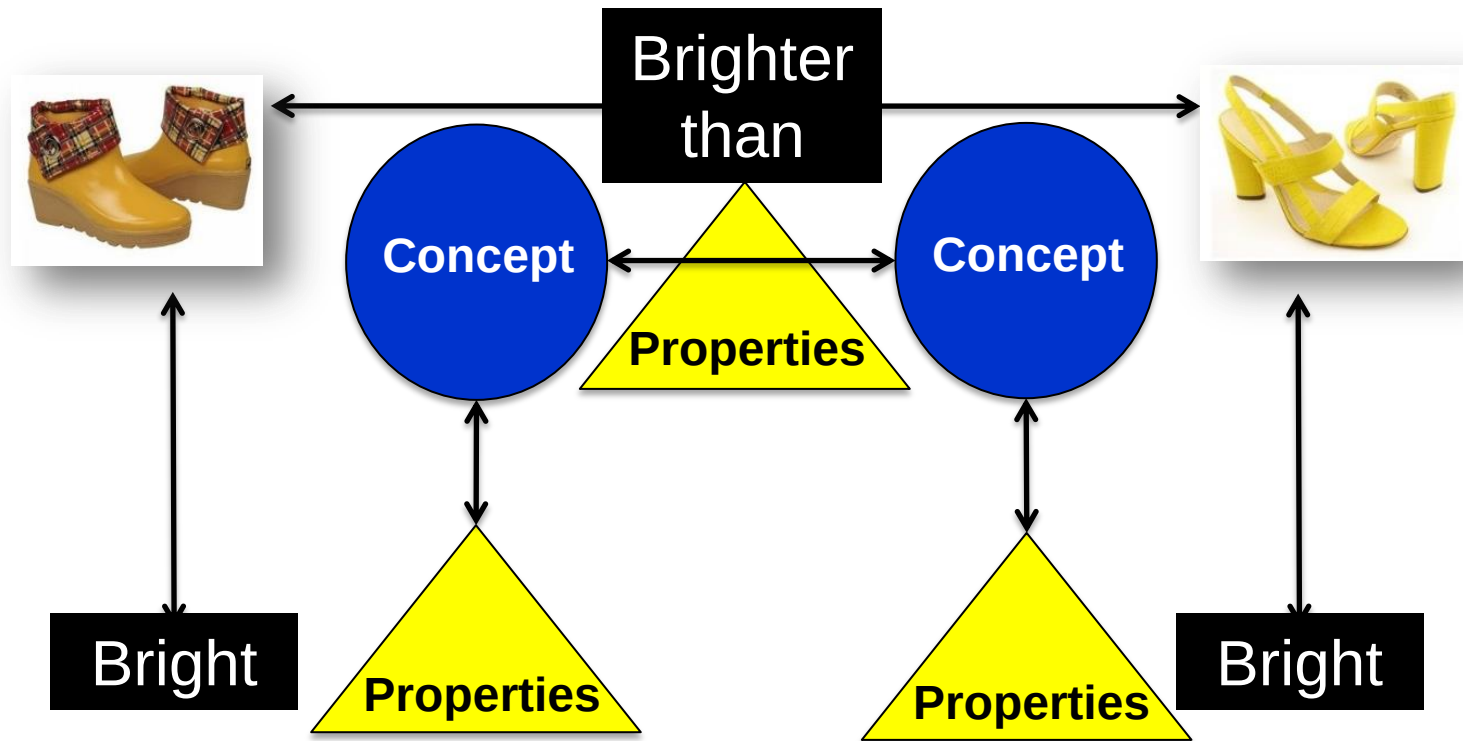
- High-level semantic properties shared by objects
- Human-understandable and machine-detectable



[Farhadi et al. 2009, Lampert et al. 2009, Kumar et al. 2009, Endres et al. 2010, Wang & Mori 2010, Berg et al. 2010, Branson et al. 2010, Parikh & Grauman 2011, ...]

Relative attributes

- Represent *comparative relationships* between classes, images, and their properties.



Learning relative attributes

- We want to learn a spectrum (ranking model) for an attribute, e.g. “brightness”.
- Supervision consists of:



Learning relative attributes

Learn a ranking function

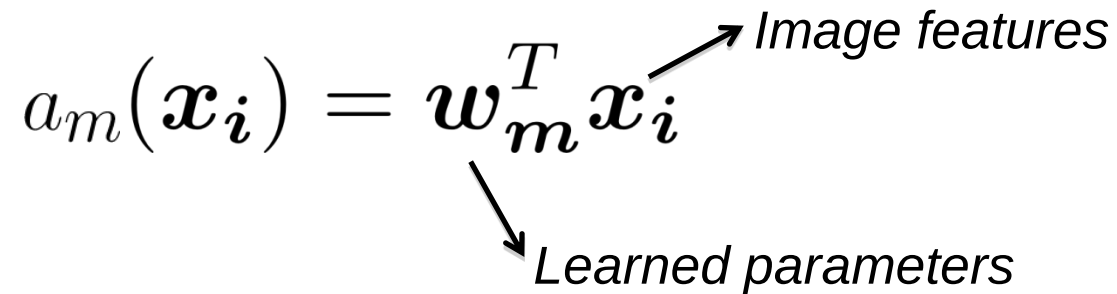
$$a_m(\mathbf{x}_i) = \mathbf{w}_m^T \mathbf{x}_i$$


Image features

Learned parameters

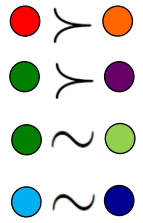
that best satisfies the constraints:

$$\forall (i, j) \in O_m : \mathbf{w}_m^T \mathbf{x}_i > \mathbf{w}_m^T \mathbf{x}_j$$

$$\forall (i, j) \in E_m : \mathbf{w}_m^T \mathbf{x}_i = \mathbf{w}_m^T \mathbf{x}_j$$

Learning relative attributes

Max-margin learning to rank formulation



$$\min \left(\frac{1}{2} \|\mathbf{w}_m^T\|_2^2 + C \left(\sum \xi_{ij}^2 + \sum \gamma_{ij}^2 \right) \right)$$

$$s.t. \quad \mathbf{w}_m^T (\mathbf{x}_i - \mathbf{x}_j) \geq 1 - \xi_{ij}$$

$$|\mathbf{w}_m^T (\mathbf{x}_i - \mathbf{x}_j)| \leq \gamma_{ij}$$

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

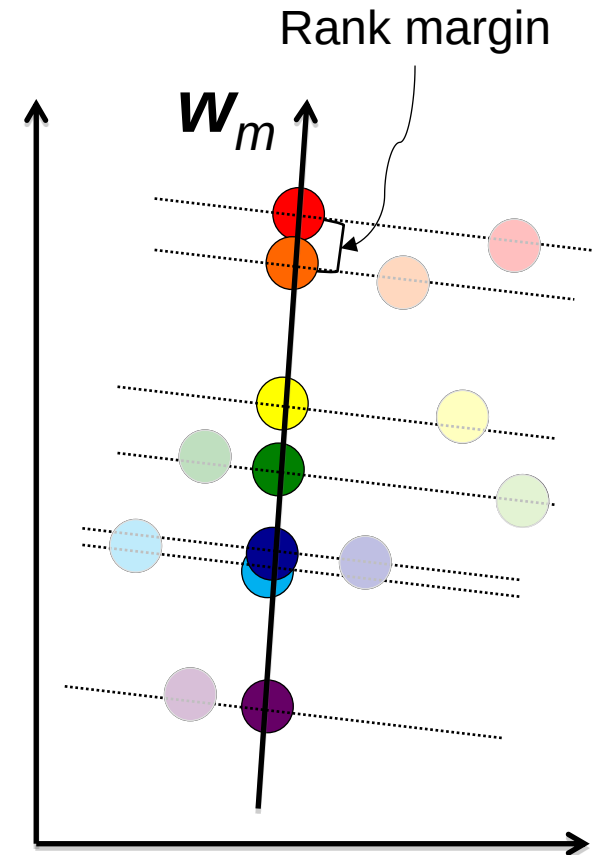


Image $\rightarrow$ Relative attribute score

Relating images

bright →



formal →



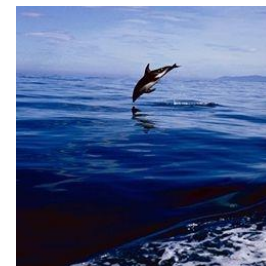
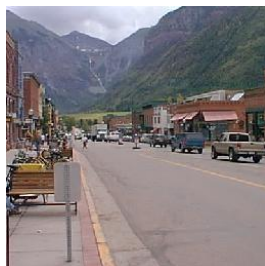
natural →



- Rank images according to attribute presence

WhittleSearch with relative attribute feedback

natural →



scores = scores + 1

“I want something
less **natural** than this.”

scores = scores + 0

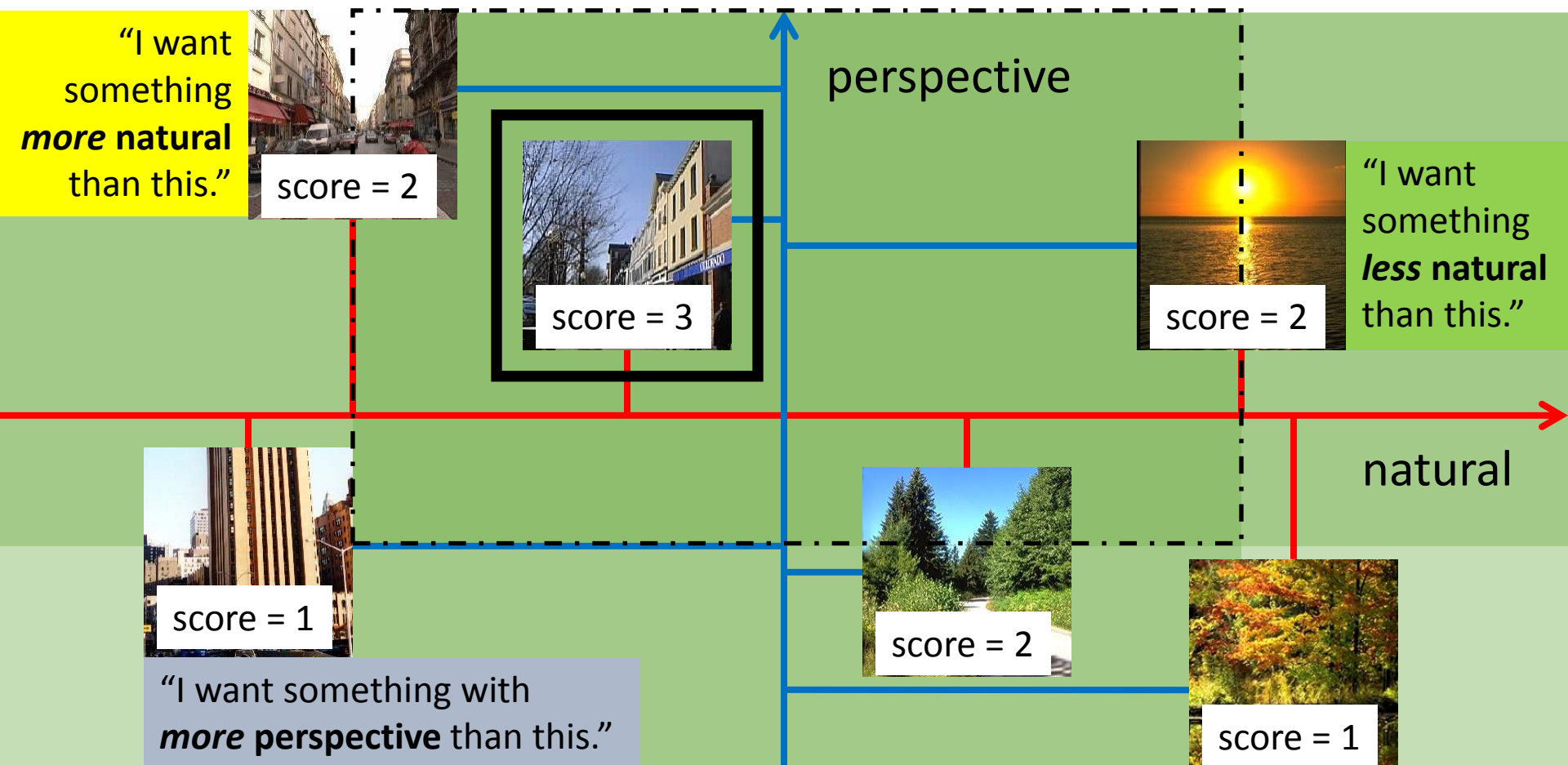
Offline:

We learn a spectrum for each attribute

During search:

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

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.]

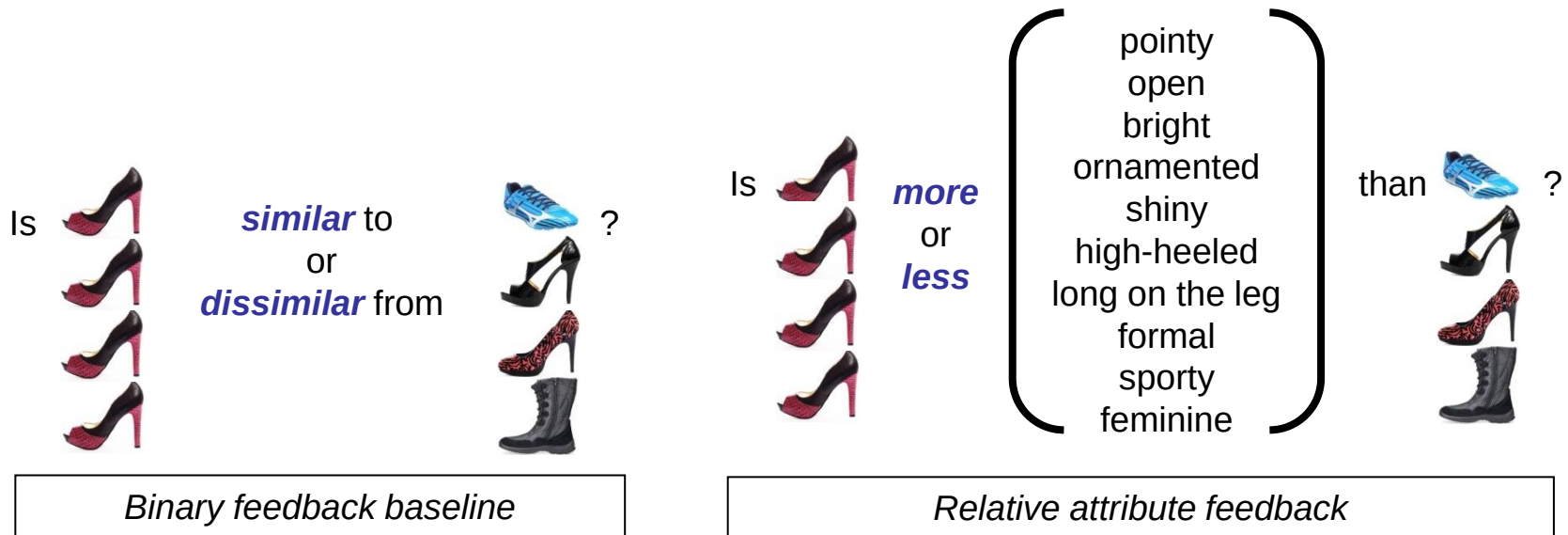
772 face images;

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



WhittleSearch Results



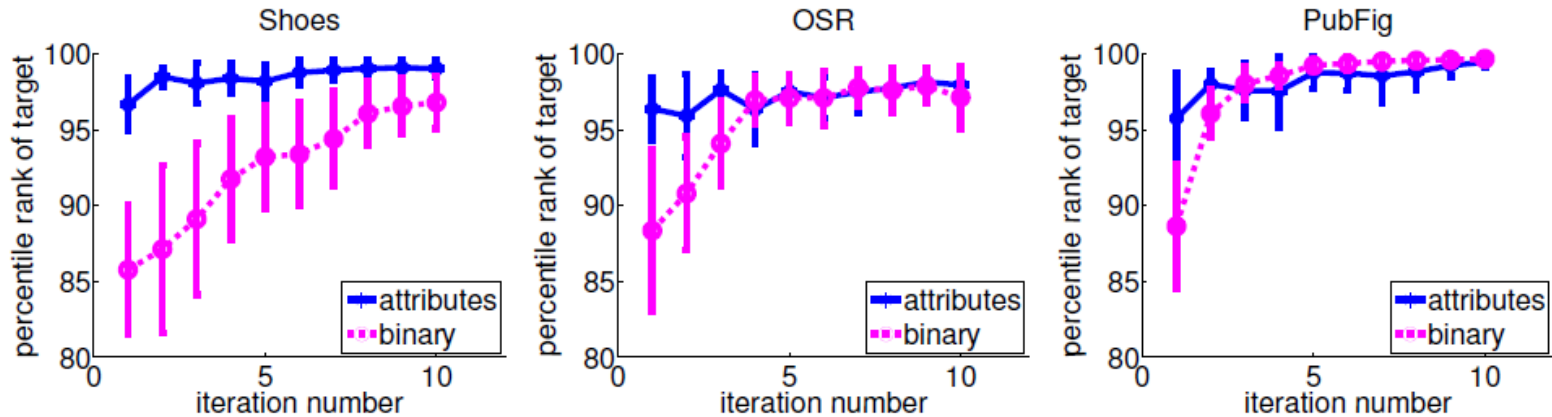
Binary relevance feedback

vs.

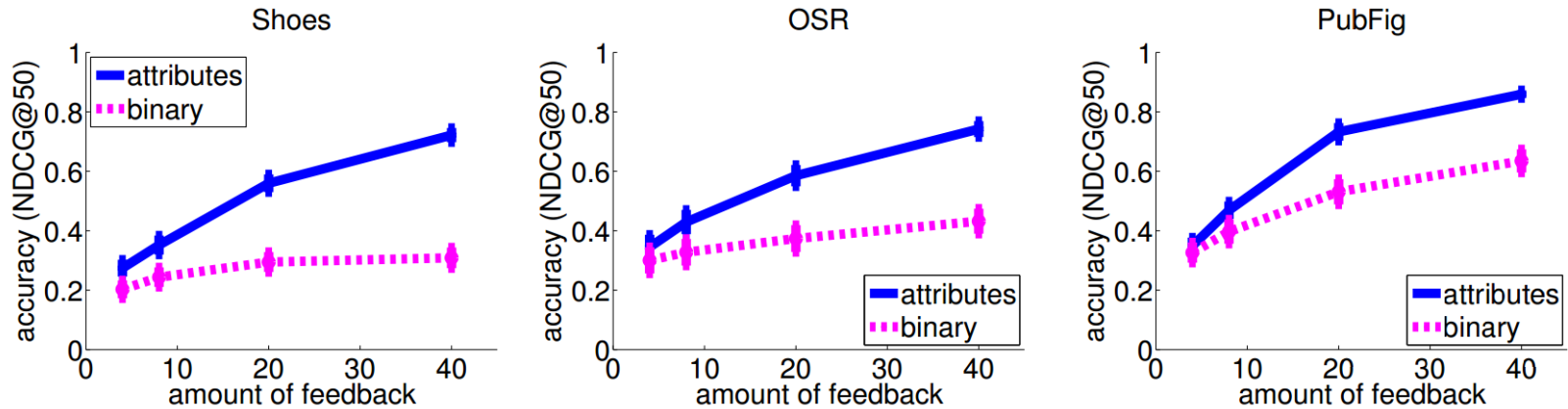


Relative attribute feedback

WhittleSearch Results



We more rapidly converge on the envisioned visual content.



Richer feedback → faster gains per unit of user effort.

Example WhittleSearch

Query: "I want a bright, open shoe that is short on the leg."

Round 1

More open than

Selected feedback

More bright in color than

Less ornaments than

Less high at the heel than



Round 2

Round 3

More formal than

More bright in color than

Higher at the heel than

More open than

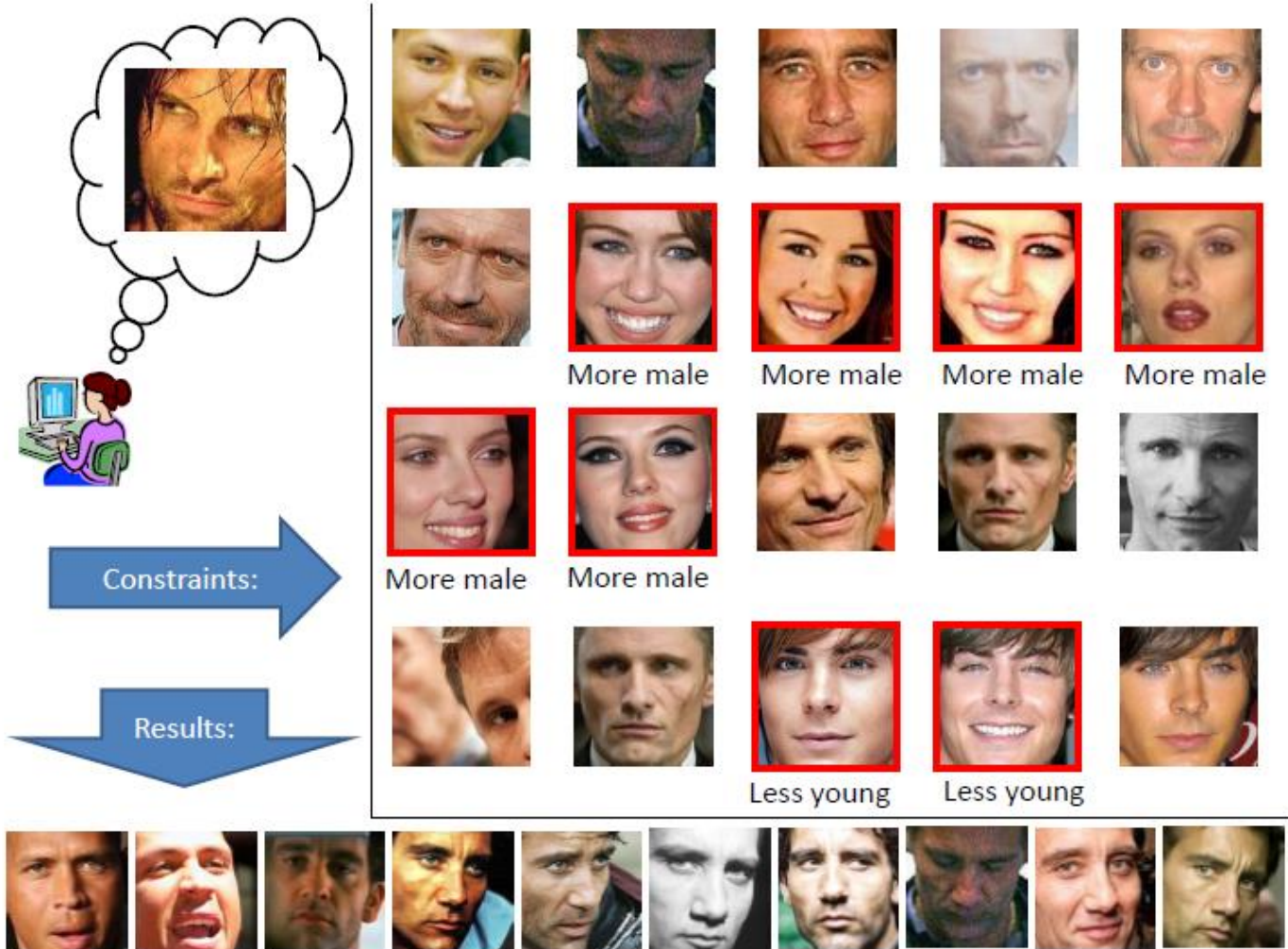
Match



45

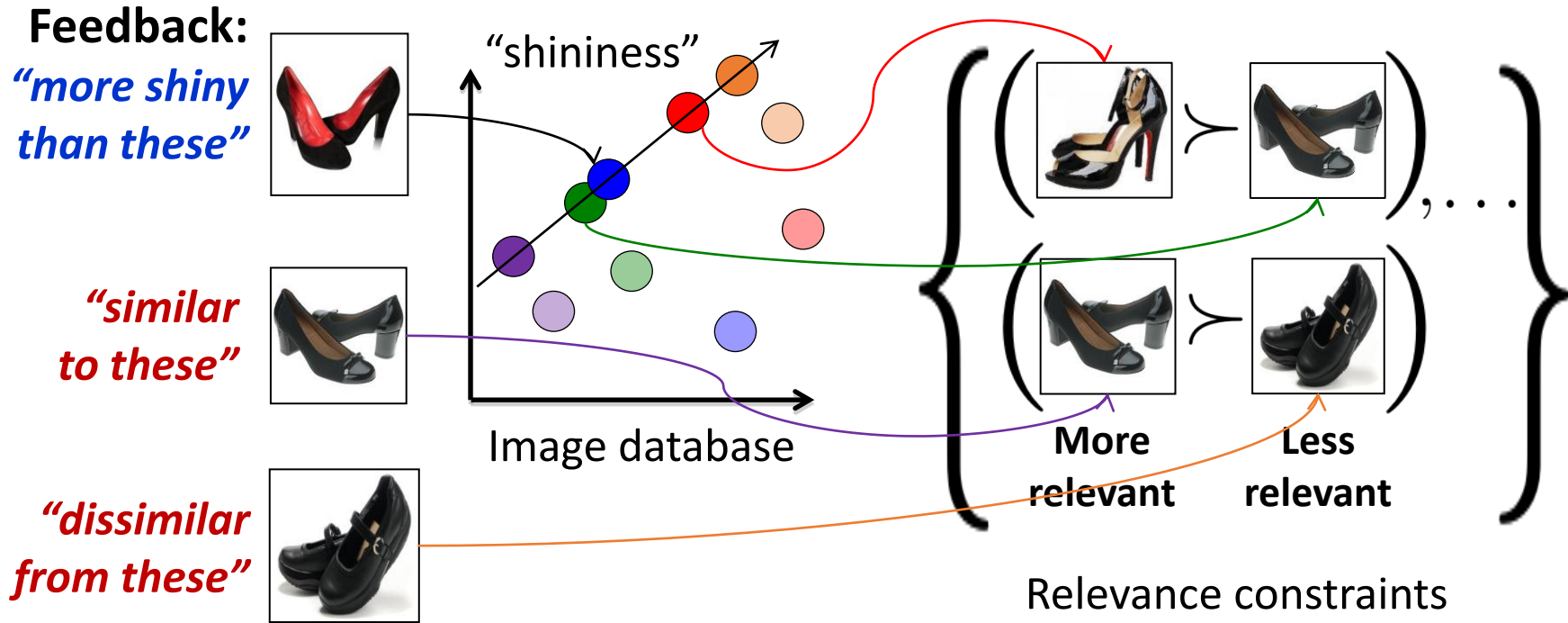


Failure case (?)



Is the user searching for a specific person (identity),
or a person meeting the description?

Hybrid relevance feedback



- We integrate relative attribute and binary feedback by learning a relevance ranking function.

Example hybrid WhittleSearch

Query: “I want a non-open shoe that is long on the leg and covered in ornaments.”



Round 1

Selected feedback
Dissimilar from



Similar to



More bright than



Less open than



Round 2

Match



Summary

- Fine-grained visual relevance is essential for next steps in image search
- Beyond tags when learning from text+images
→ model implied importance cues
- Beyond clicks as feedback
→ visual comparisons to refine search

Looking forward

- What is implied by natural language description beyond ordering? (tags vs. sentences)
- How to ensure that feedback user gives is useful (e.g., not redundant)?
- What attributes should be in the vocabulary?
- How to align user's attribute language with the visual attribute models?

Summary

- Fine-grained visual relevance is essential for next steps in image search
- Beyond tags when learning from text+images
→ model implied importance cues
- Beyond clicks as feedback
→ visual comparisons to refine search

