Data Sets and Crowdsourcing

Or: My grad students are starting to hate me, but it looks like we need more training data.

Computer Vision

James Hays

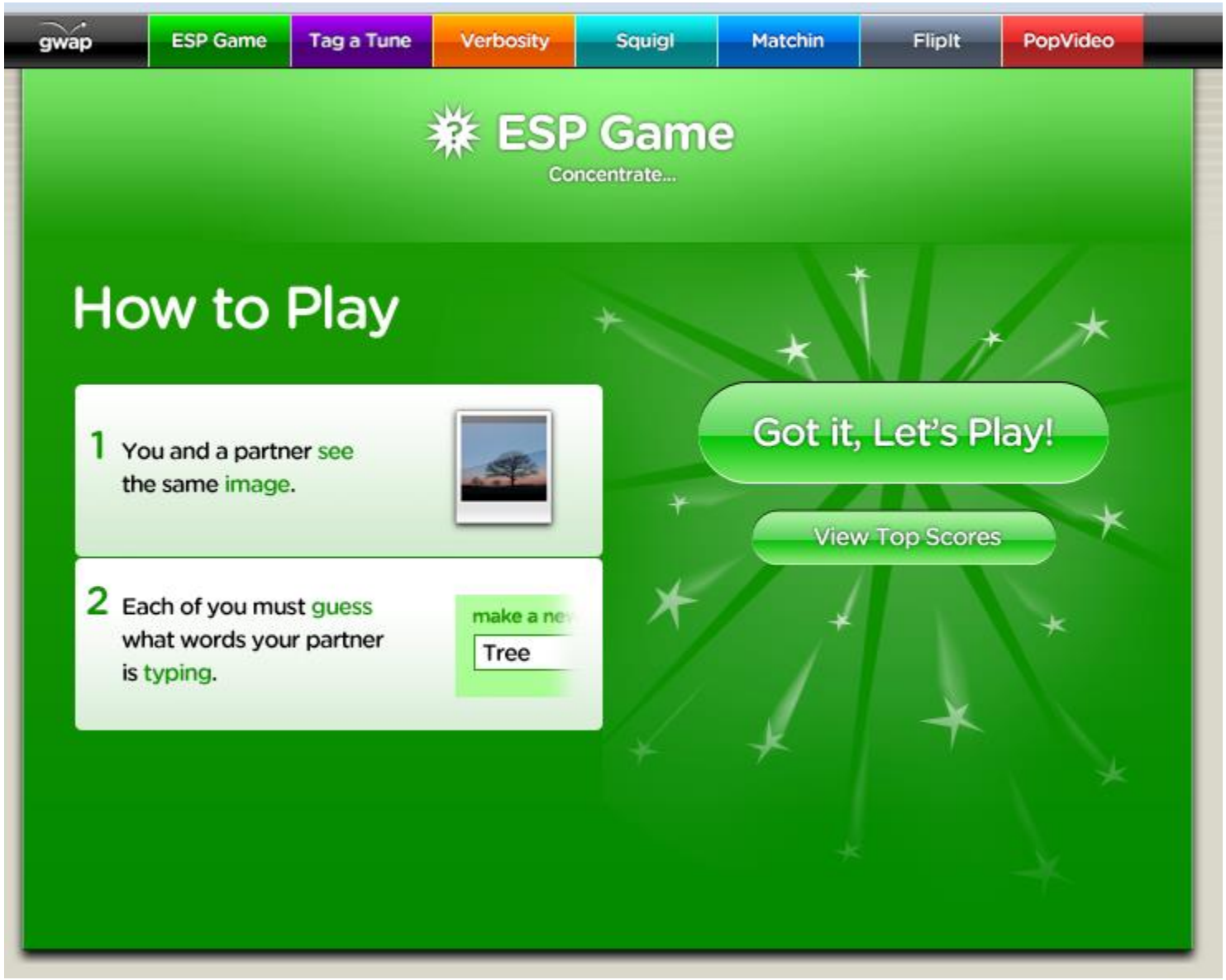
Outline

- Data collection with experts – PASCAL VOC
- Annotation with non-experts
  - LabelMe – no incentive (altruism, perhaps)
  - ESP Game – fun incentive (not fun enough?)
  - Mechanical Turk – financial incentive
- Human-in-the-loop Recognition
  - Visipedia

LabelMe

- http://labelme.csail.mit.edu
- “Open world” database annotated by the community*

- **Notes on Image Annotation**, Barriuso and Torralba 2012. http://arxiv.org/abs/1210.3448



Luis von Ahn and Laura Dabbish. Labeling Images with a Computer Game. ACM Conf. on Human Factors in Computing Systems, CHI 2004

ESP Game | Tag a Tune | Verbosity | Squigl | Matchin | logged i

score
0

ESP Game
Concentrate...

time
2:56

What do you see?

taboo words
student

guesses

submit | pass

Play Anonymously

Screencast-O-Matic.com

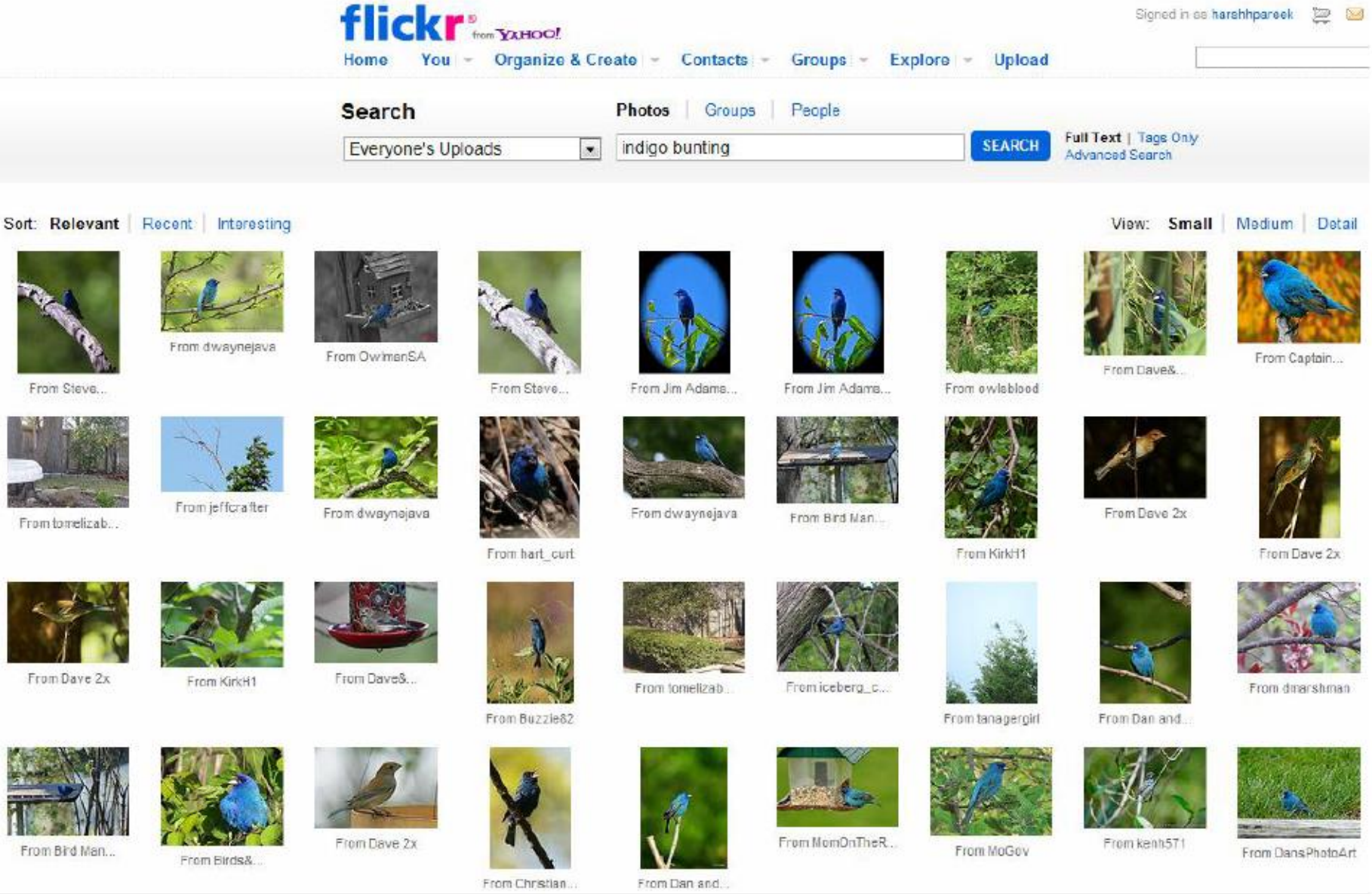
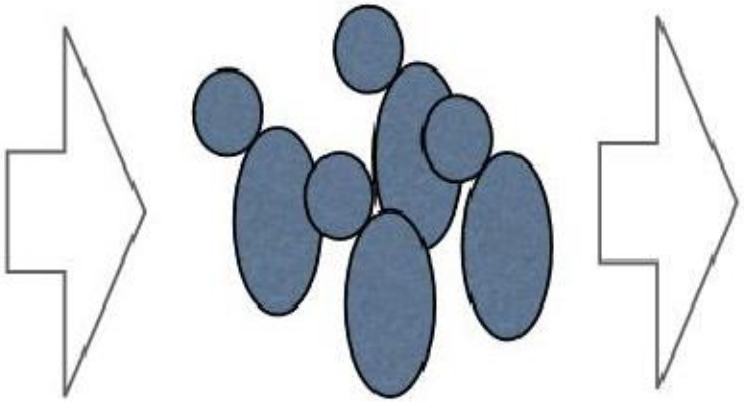


Image credit: Flickr.com

Building datasets

6000 images from flickr.com

Annotators



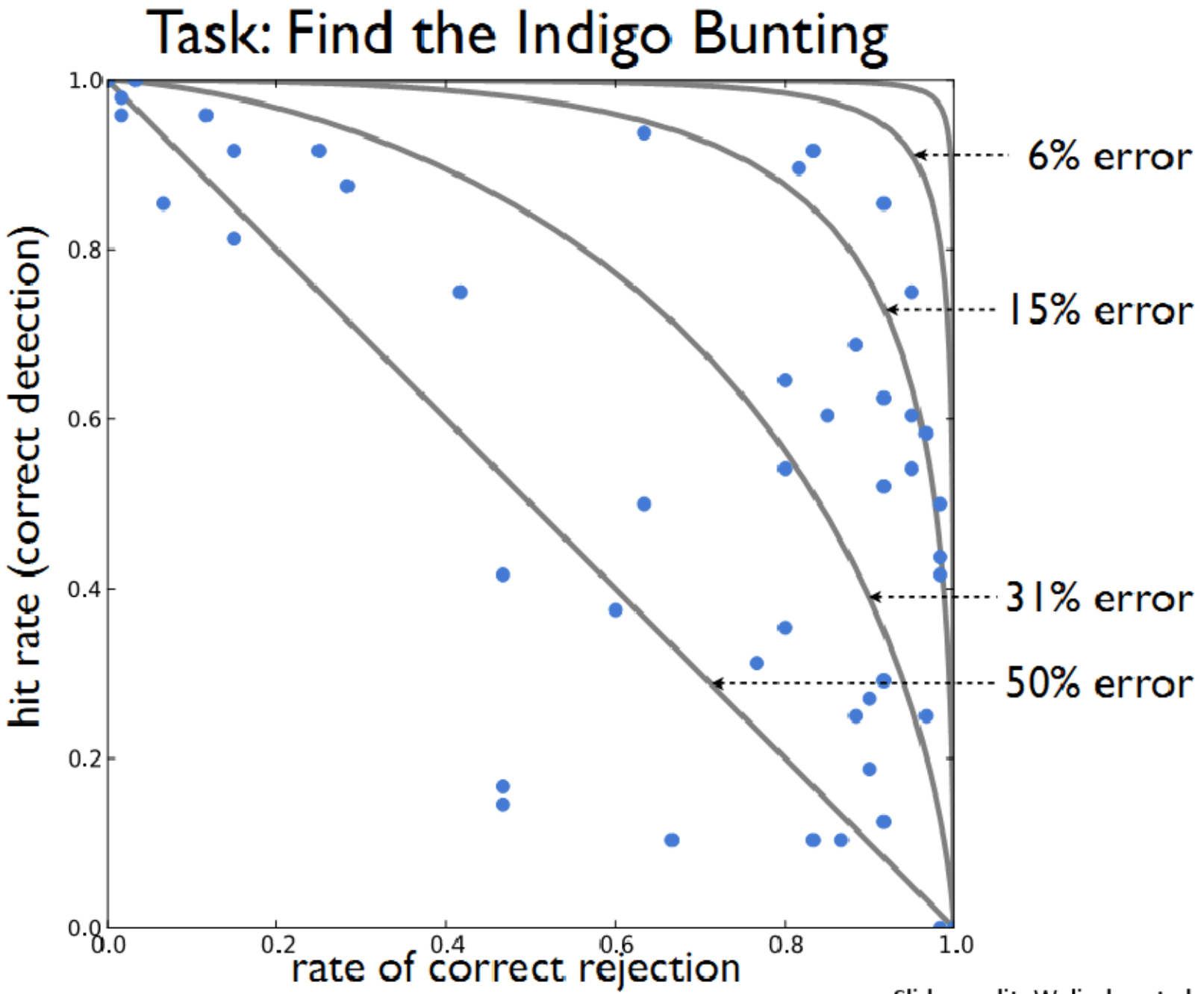
amazonmechanical turk
Artificial Artificial Intelligence

Is there an Indigo bunting in the image?

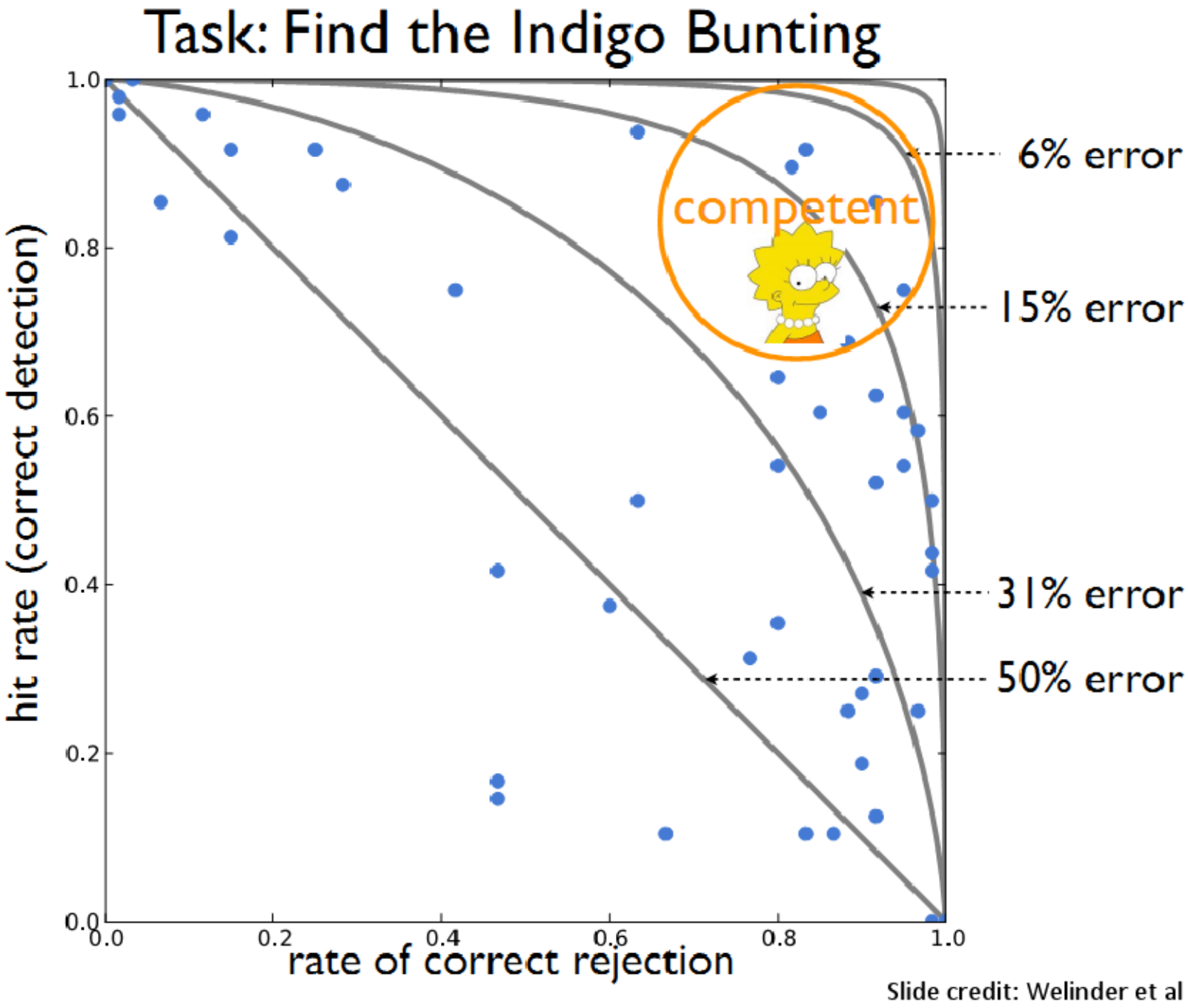
100s of training images

Slide credit: Welinder et al

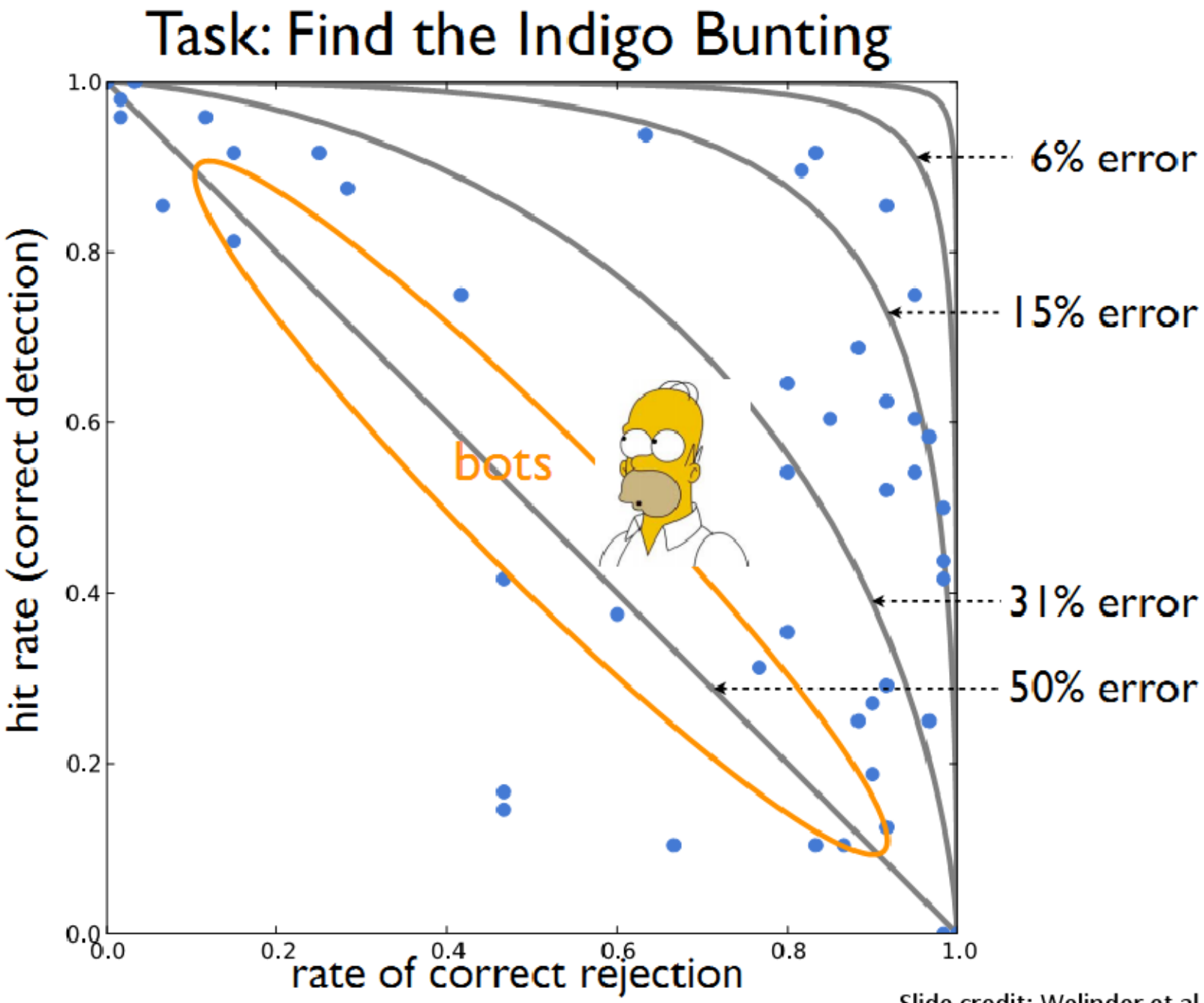
Slide credit: Welinder et al



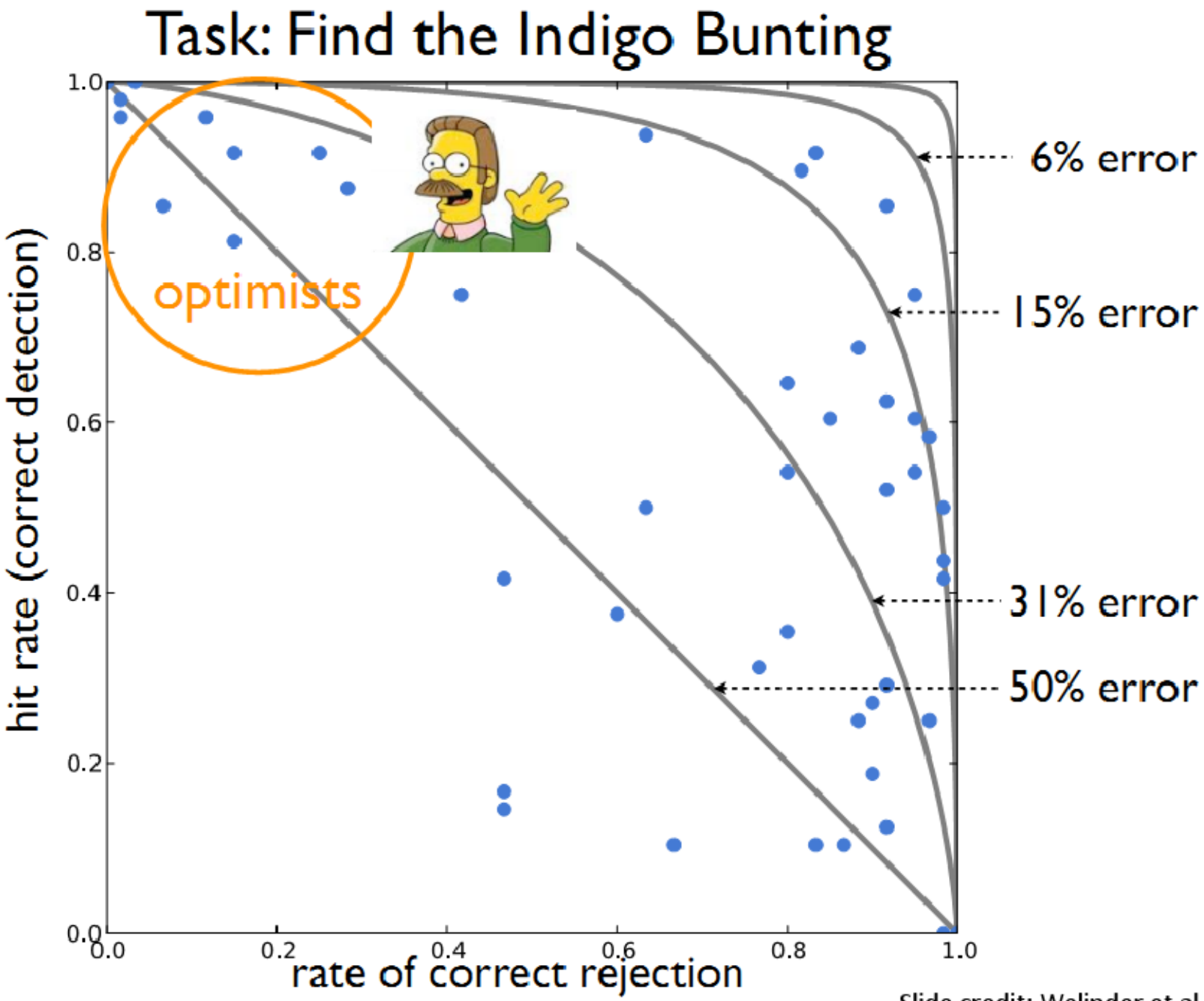
Slide credit: Welinder et al



Slide credit: Welinder et al

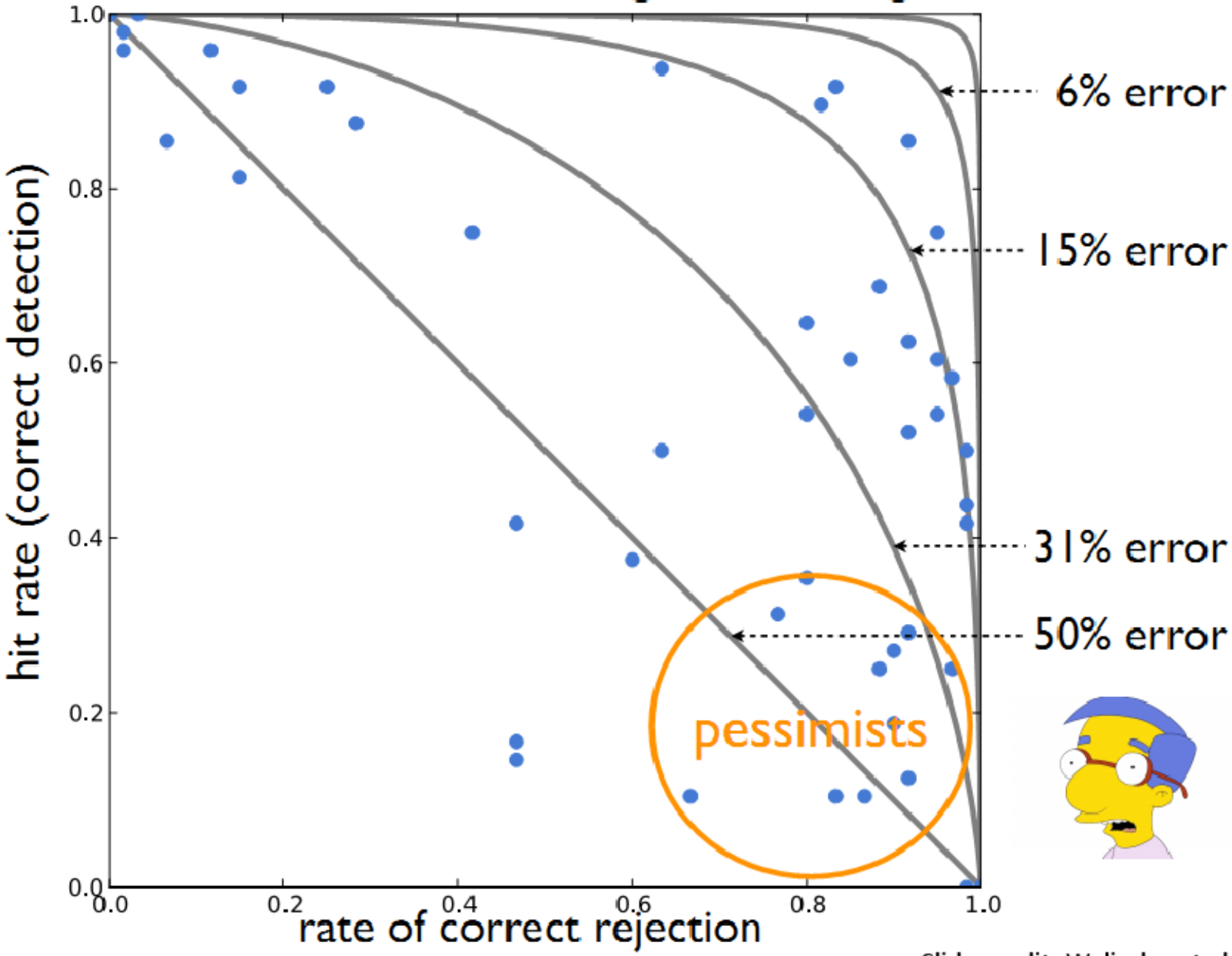


Slide credit: Welinder et al

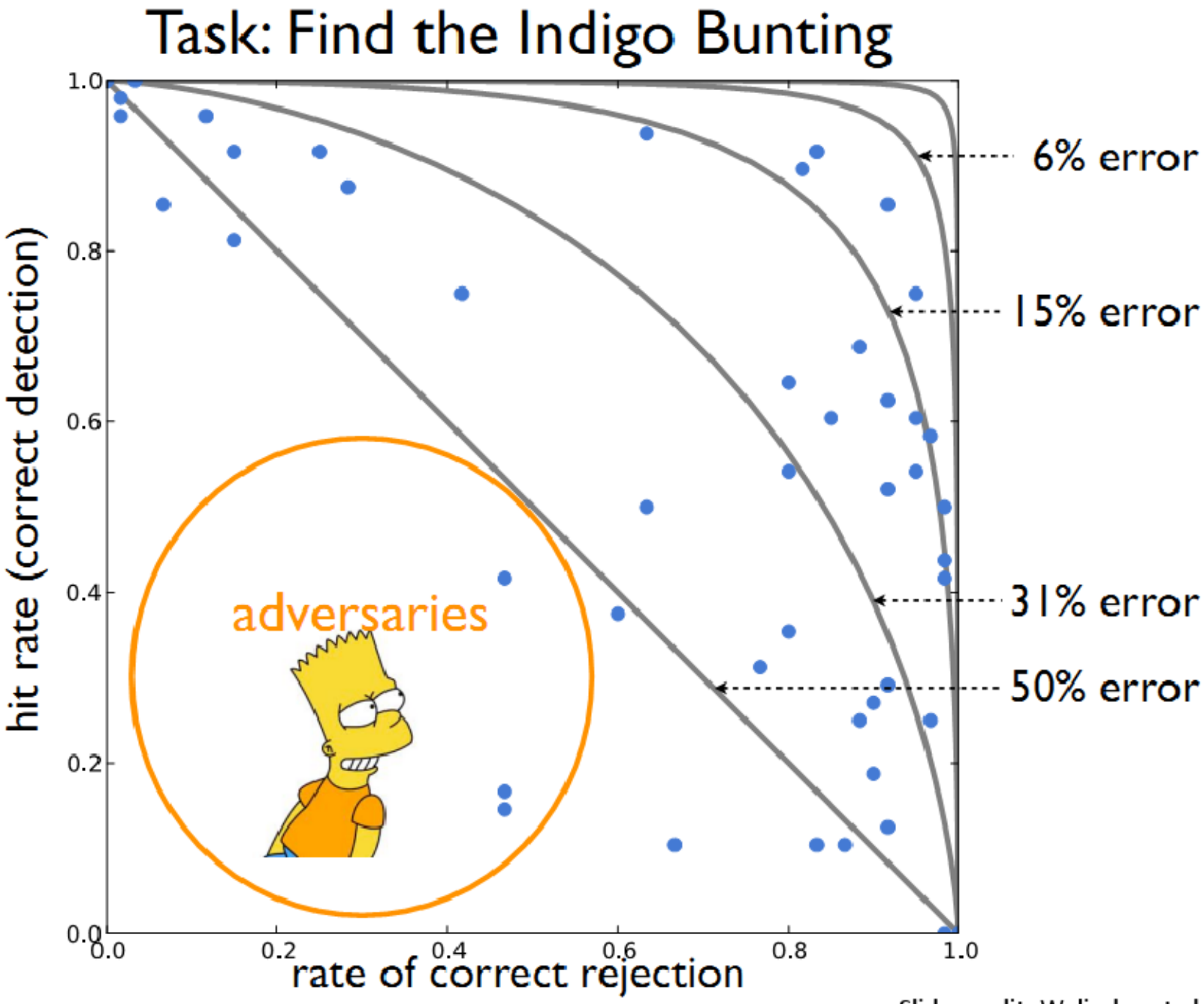


Slide credit: Welinder et al

Task: Find the Indigo Bunting



Slide credit: Welinder et al



Slide credit: Welinder et al

Utility data annotation via Amazon Mechanical Turk

X 100 000 = $5000

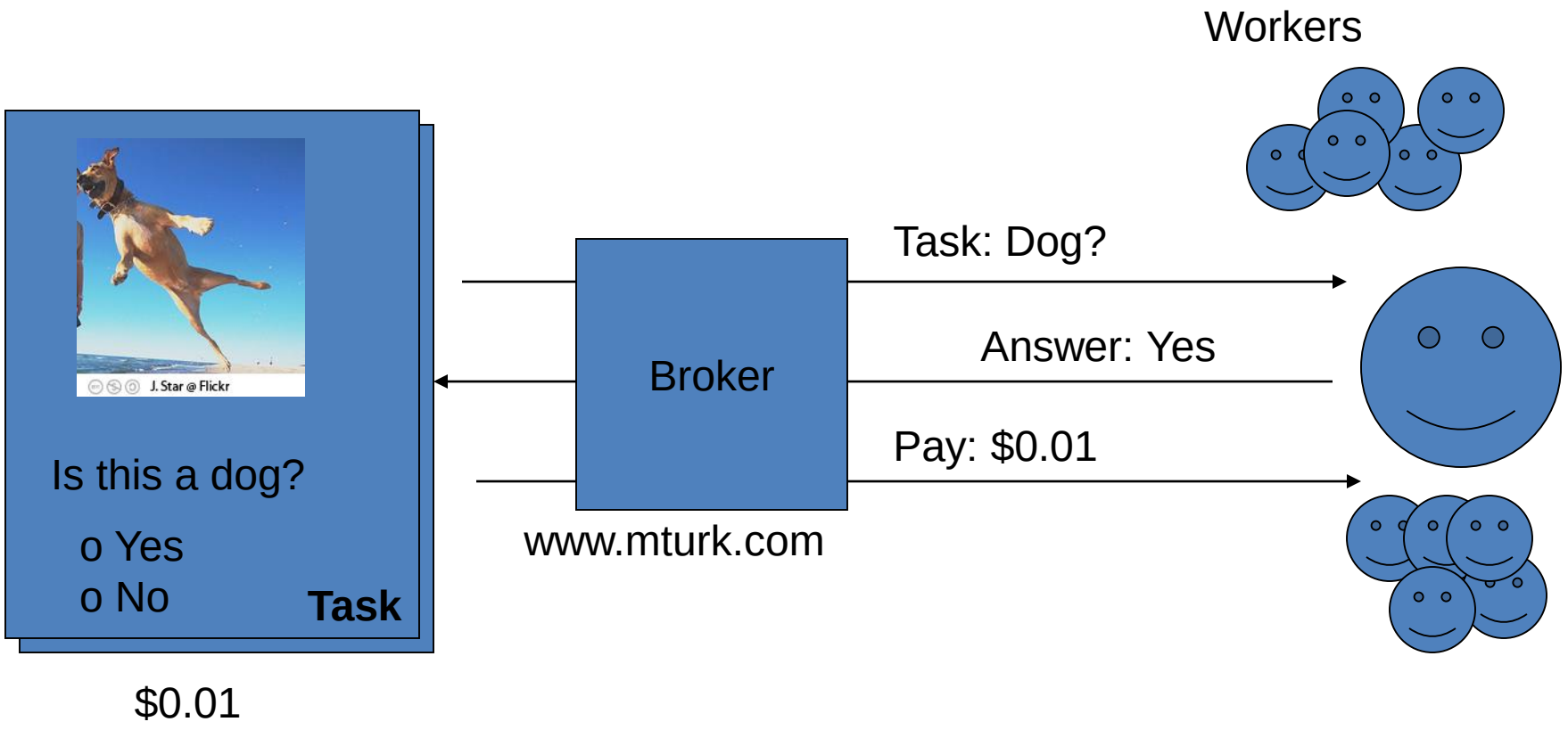
Alexander Sorokin

David Forsyth

CVPR Workshops 2008

Slides by Alexander Sorokin

Amazon Mechanical Turk



Annotation protocols

- Type keywords
- Select relevant images
- Click on landmarks
- Outline something
- Detect features

…….. anything else ……

Type keywords

Mechanical Turk Project

If you're using the turk, Be sure to copy the text back into the HIT page so that you can be credited.

- Photo should be rotated 90 degrees left (counter-clockwise)
- Photo should be rotated 90 degrees right (clockwise)
- Photo should be turned upside down
- Photo is oriented properly

Please describe the picture in the box using 10 words or more:

shells

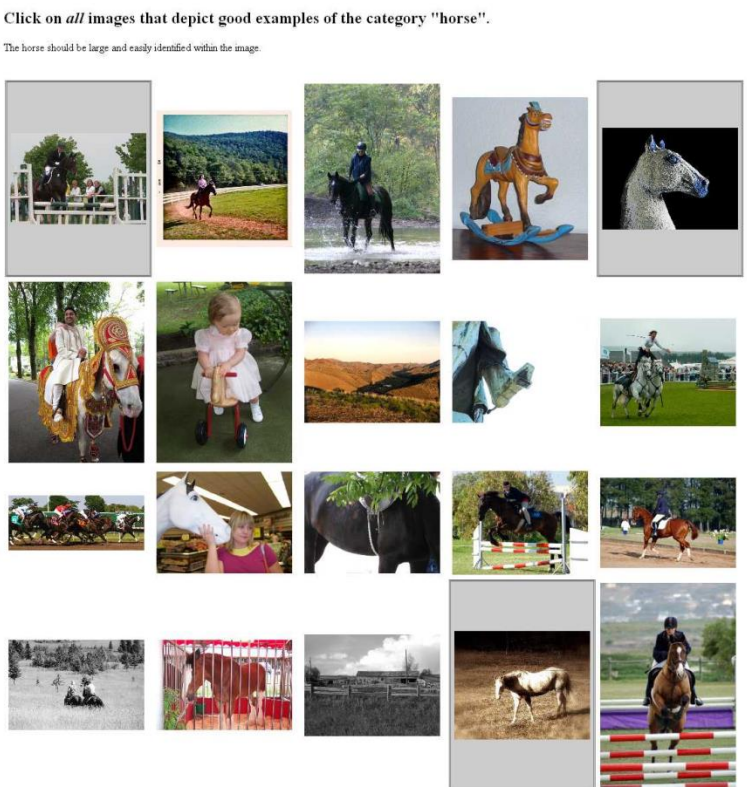
Submit Turk | Skip / Load a different photo

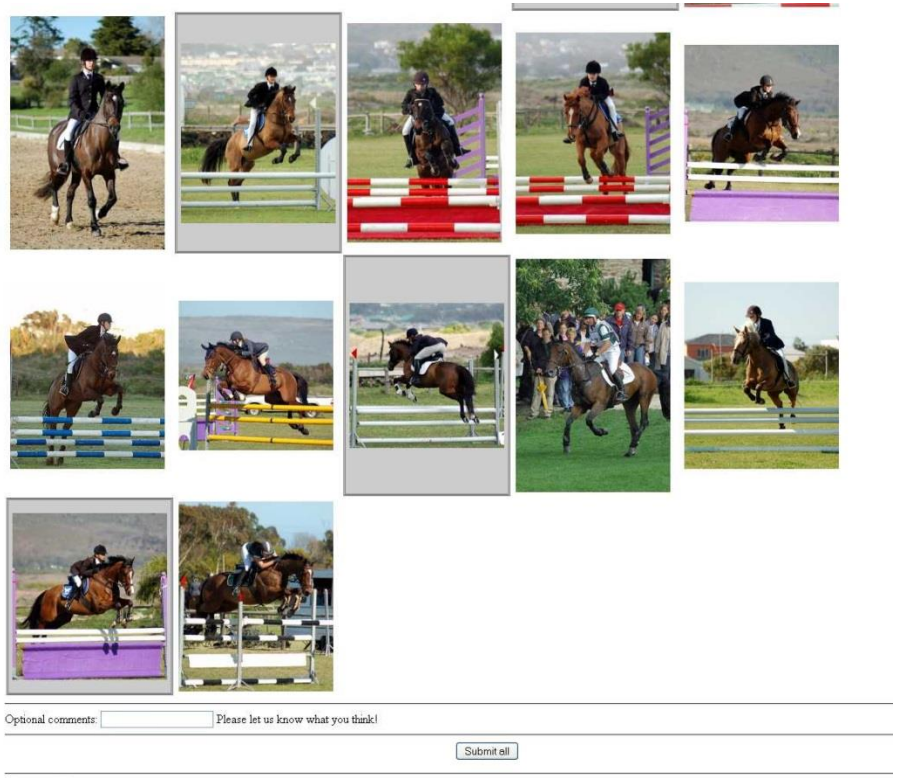
The submit button MUST be clicked!

$0.01

http://austinsmoke.com/turk/.

Select examples

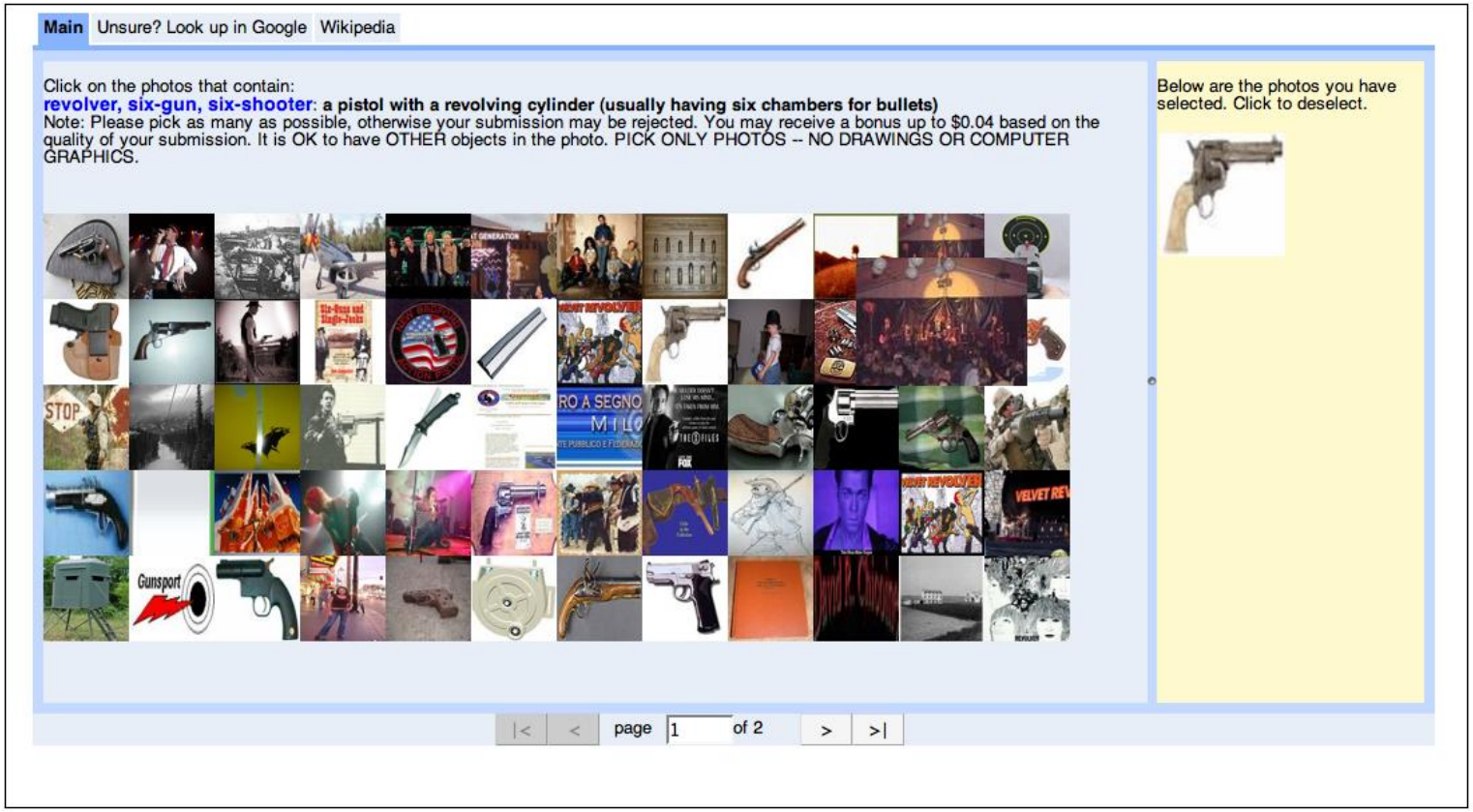




Joint work with Tamara and Alex Berg

http://visionpc.cs.uiuc.edu/~largescale/data/simpleevaluation/html/horse.html

Select examples



$0.02

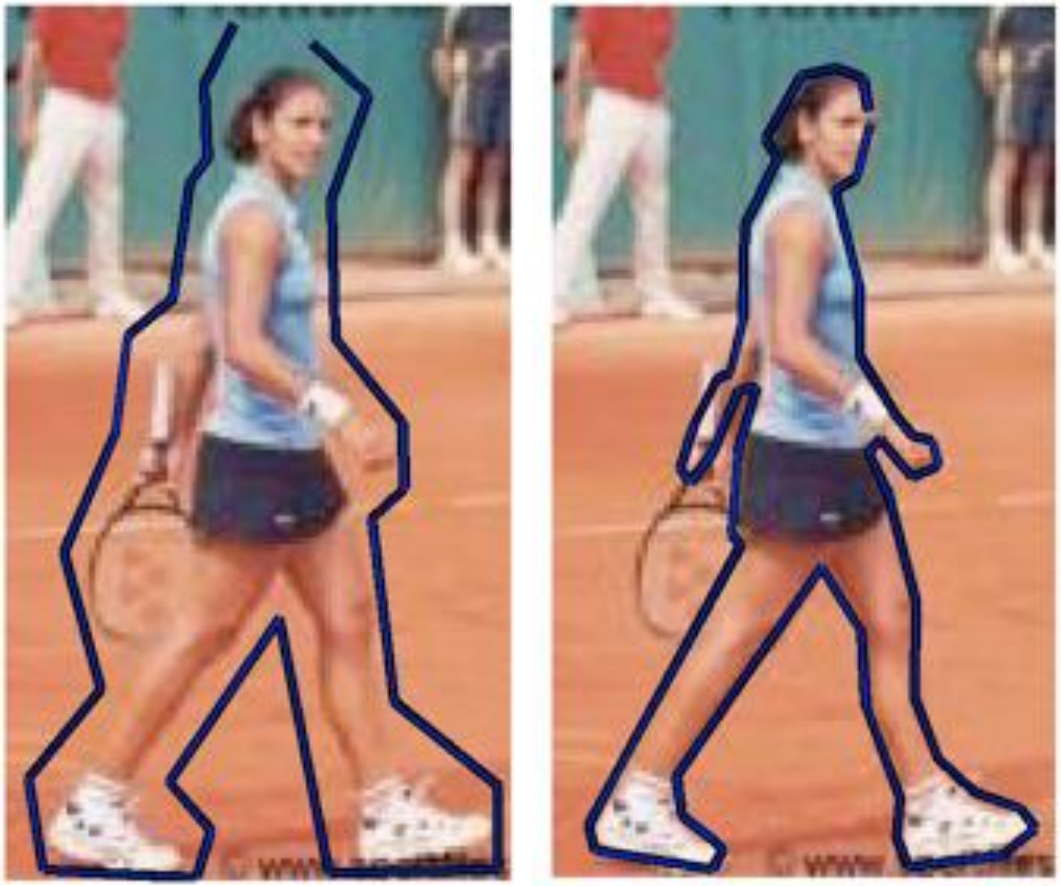
requester mtlabel

Click on landmarks

$0.01

http://vision-app1.cs.uiuc.edu/mt/results/people14-batch11/p7/

Outline something



$0.01

http://visionpc.cs.uiuc.edu/~largescale/results/production-3-2/results_page_013.html

Data from Ramanan NIPS06

Motivation

X 100 000 = $5000

Custom annotations

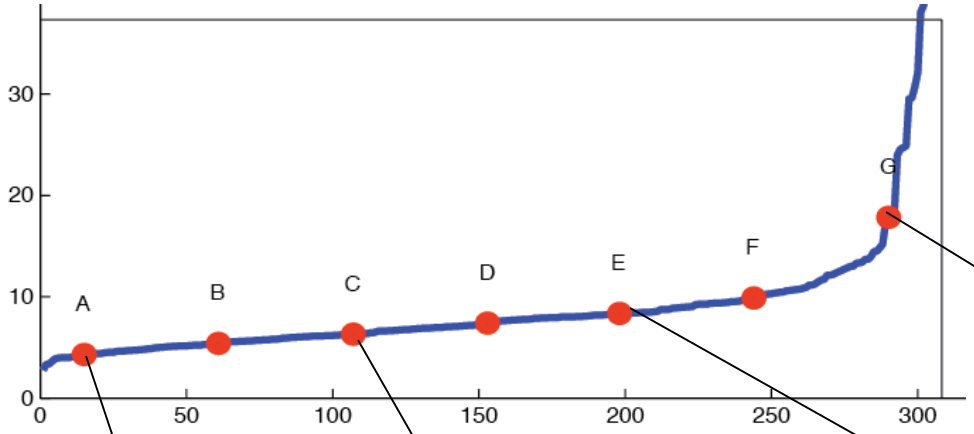
Large scale

Low price

Issues

- Quality?
  - How good is it?
  - How to be sure?
- Price?
  - How to price it?

Annotation quality



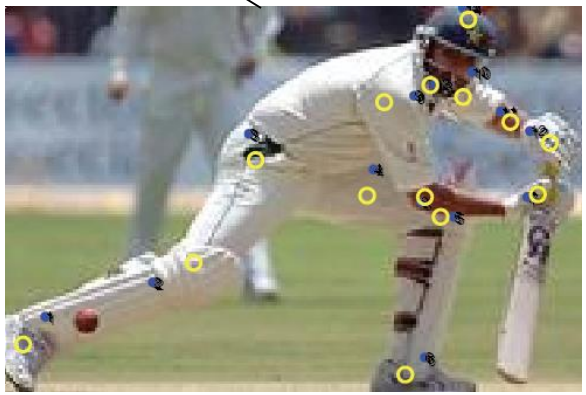
Agree within 5-10 pixels on 500x500 screen

There are bad ones.

A

C

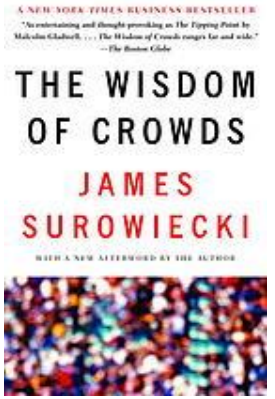
E



G

How do we get quality annotations?

Ensuring Annotation Quality



- Consensus / Multiple Annotation / "Wisdom of the Crowds"

  Not enough on its own, but widely used

- Gold Standard / Sentinel
  - Special case: qualification exam

  Widely used and most important. Find good annotators and keep them honest.
- Grading Tasks
  - A second tier of workers who grade others

  Not widely used

Pricing

- Trade off between throughput and cost
  - *NOT* as much of a trade off with quality
- Higher pay can actually attract scammers

Outline

- Data collection with experts – PASCAL VOC
- Annotation with non-experts
  - LabelMe
  - ESP Game
  - Mechanical Turk
- Human-in-the-loop Recognition
  - Visipedia

Visual Recognition with Humans in the Loop

Steve Branson, Catherine Wah, Florian Schroff, Boris Babenko, Peter Welinder, Pietro Perona, Serge Belongie

Part of the Visipedia project

Slides from Brian O'Neil

Introduction:

(A) Easy for Humans

Chair? Airplane? …

Computers starting to get good at this.

(B) Hard for Humans

Finch? Bunting?…

If it's hard for humans, it's probably too hard for computers.

(C) Easy for Humans

Yellow Belly? Blue Belly? …

Semantic feature extraction difficult for computers.

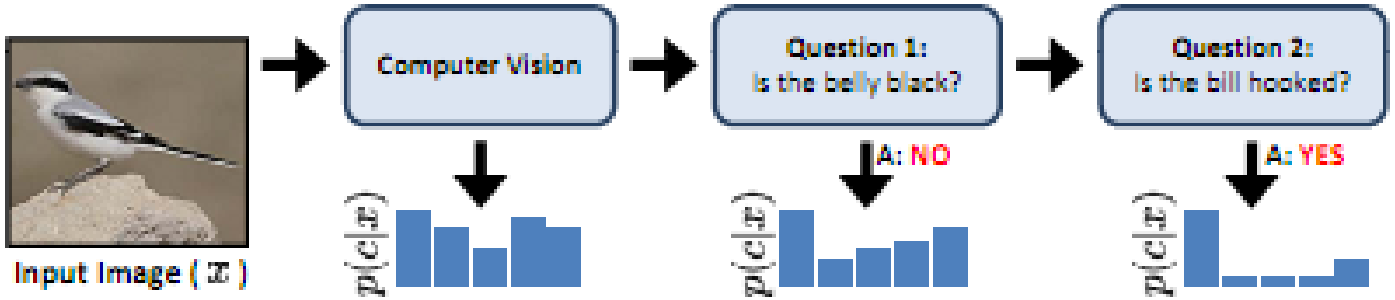
Combine strengths to solve this problem.

The Approach: What is progress?

- Supplement visual recognition with the human capacity for visual feature extraction to tackle difficult (fine-grained) recognition problems.
- Typical progress is viewed as increasing data difficulty while maintaining full autonomy
- Here, the authors view progress as reduction in human effort on difficult data.

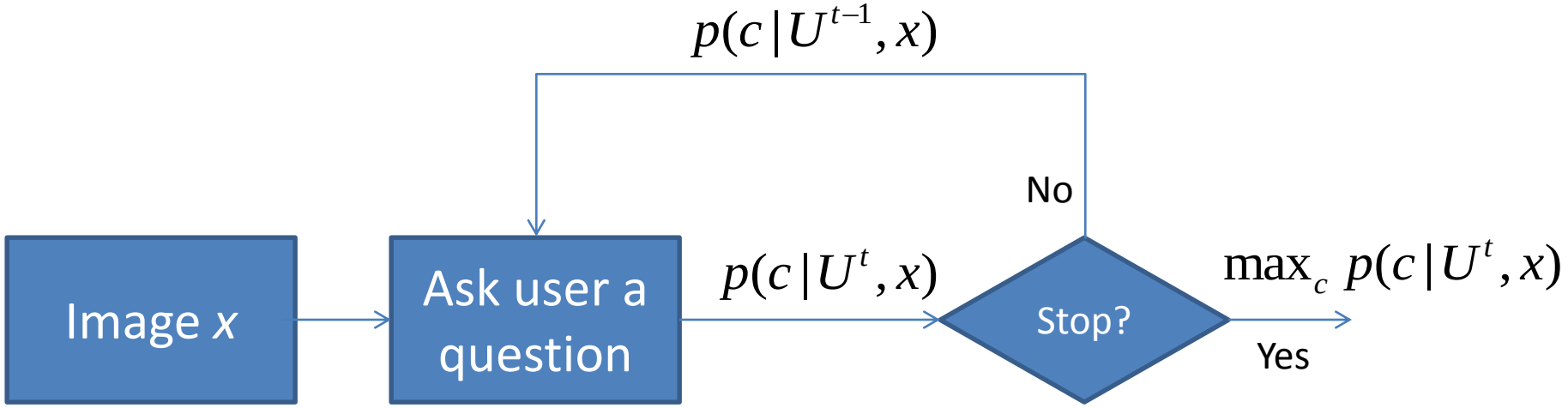
The Approach: 20 Questions

- Ask the user a series of discriminative visual questions to make the classification.



Which 20 questions?

- At each step, exploit the image itself and the user response history to select the most informative question to ask next.



Which question to ask?

- The question that will reduce entropy the most, taking into consideration the computer vision classifier confidences for each category.

The Dataset: Birds-200

- 6033 images of 200 species





Implementation

- Assembled 25 visual questions encompassing 288 visual attributes extracted from www.whatbird.com
- Mechanical Turk users asked to answer questions and provide confidence scores.

User Responses.

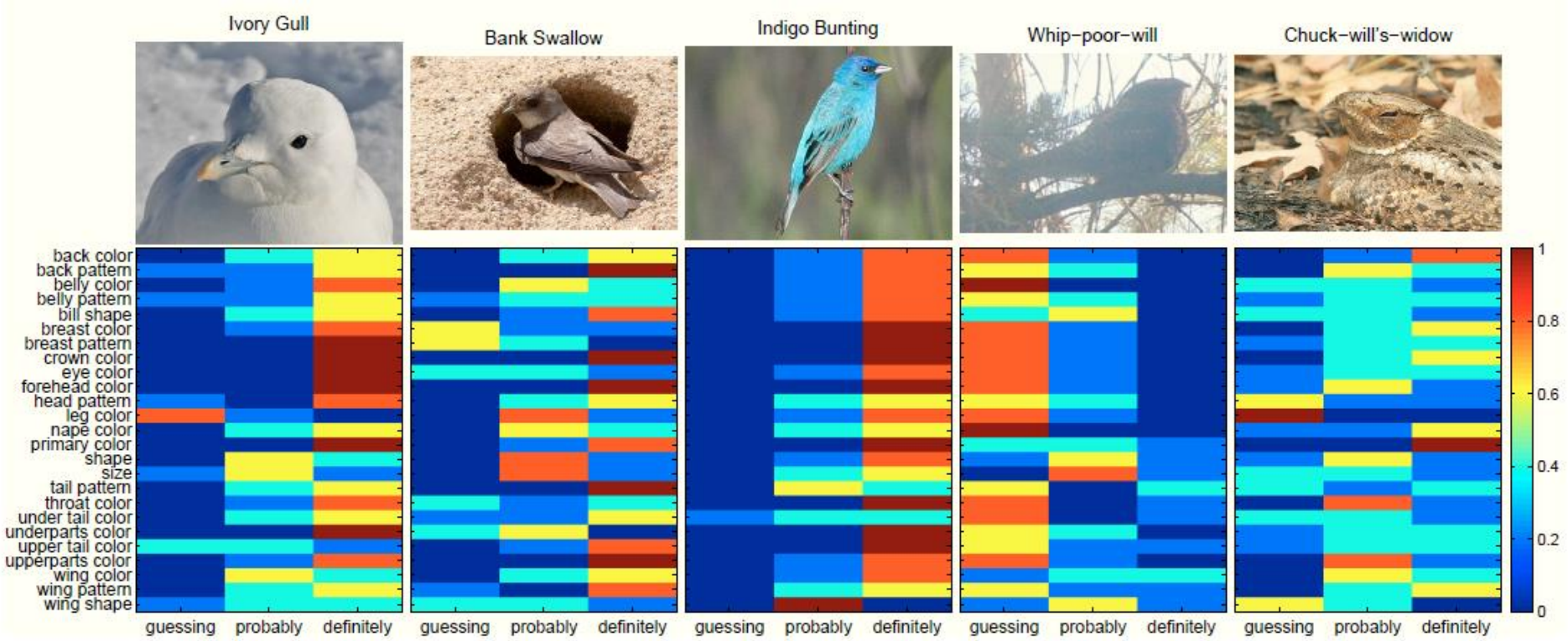
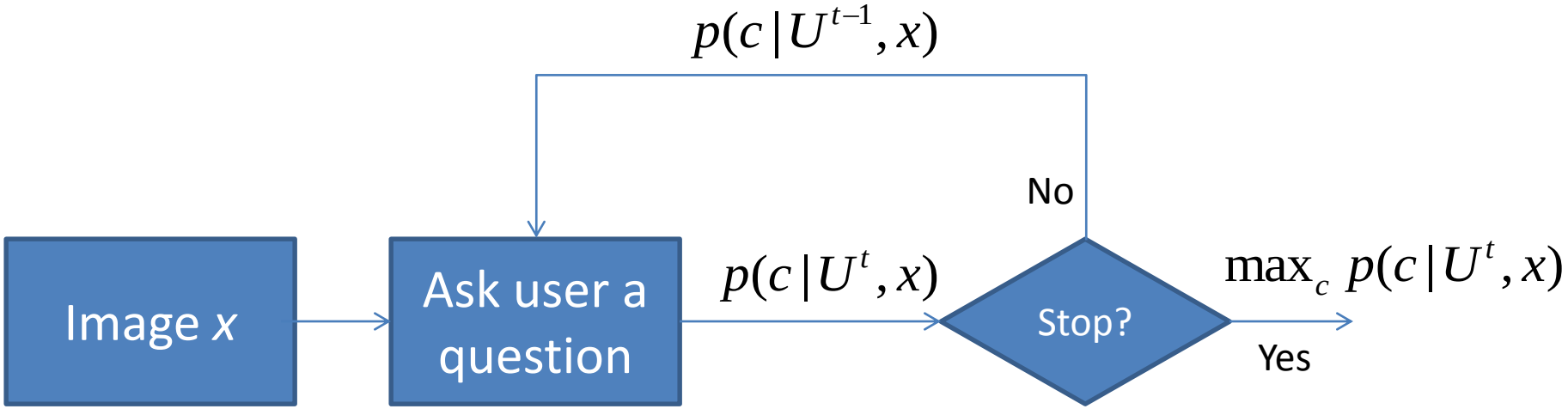


Fig. 4. Examples of user responses for each of the 25 attributes. The distribution over {*Guessing, Probably, Definitely*} is color coded with blue denoting 0% and red denoting 100% of the five answers per image attribute pair.

Visual recognition

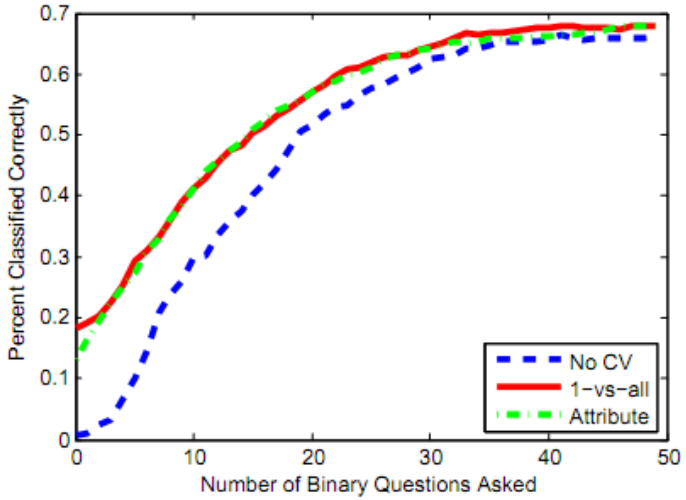
- Any vision system that can output a probability distribution across classes will work.
- Authors used Andrea Vedaldis's code.
  - Color/gray SIFT
  - VQ geometric blur
  - 1 v All SVM
- Authors added full image color histograms and VQ color histograms

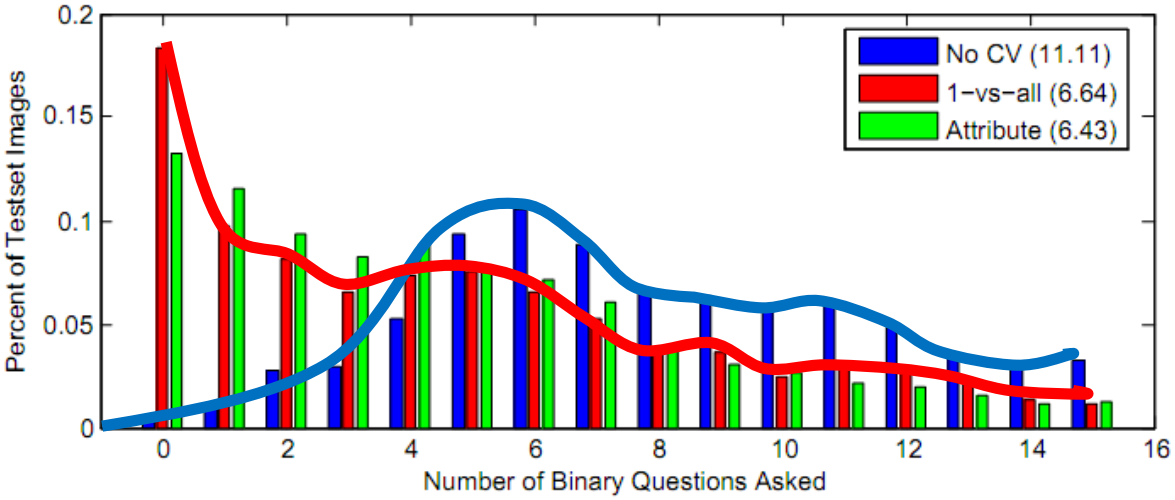
Experiments



- 2 Stop criteria:
  - Fixed number of questions – evaluate accuacy
  - User stops when bird identified – measure number of questions required.

Results





- Average number of questions to make ID reduced from 11.11 to 6.43
- Method allows CV to handle the easy cases, consulting with users only on the more difficult cases.

Key Observations

- Visual recognition reduces labor over a pure “20 Q” approach.
- Visual recognition improves performance over a pure “20 Q” approach. (69% vs 66%)
- User input dramatically improves recognition results. (66% vs 19%)

Strengths and weaknesses

- Handles very difficult data and yields excellent results.
- Plug-and-play with many recognition algorithms.
- Requires significant user assistance
- Reported results assume humans are perfect verifiers
- Is the reduction from 11 questions to 6 really that significant?

Next lecture(s)

- Human-in-the-loop
- Attributes
- More crowdsourcing (ImageNet, MS COCO)