The Multidimensional Wisdom of Crowds

Welinder P., Branson S., Belongie S., Perona, P

Experiment Presentation
[CS395T] Visual Recognition Fall 2012

Presented by: Niveda Krishnamoorthy

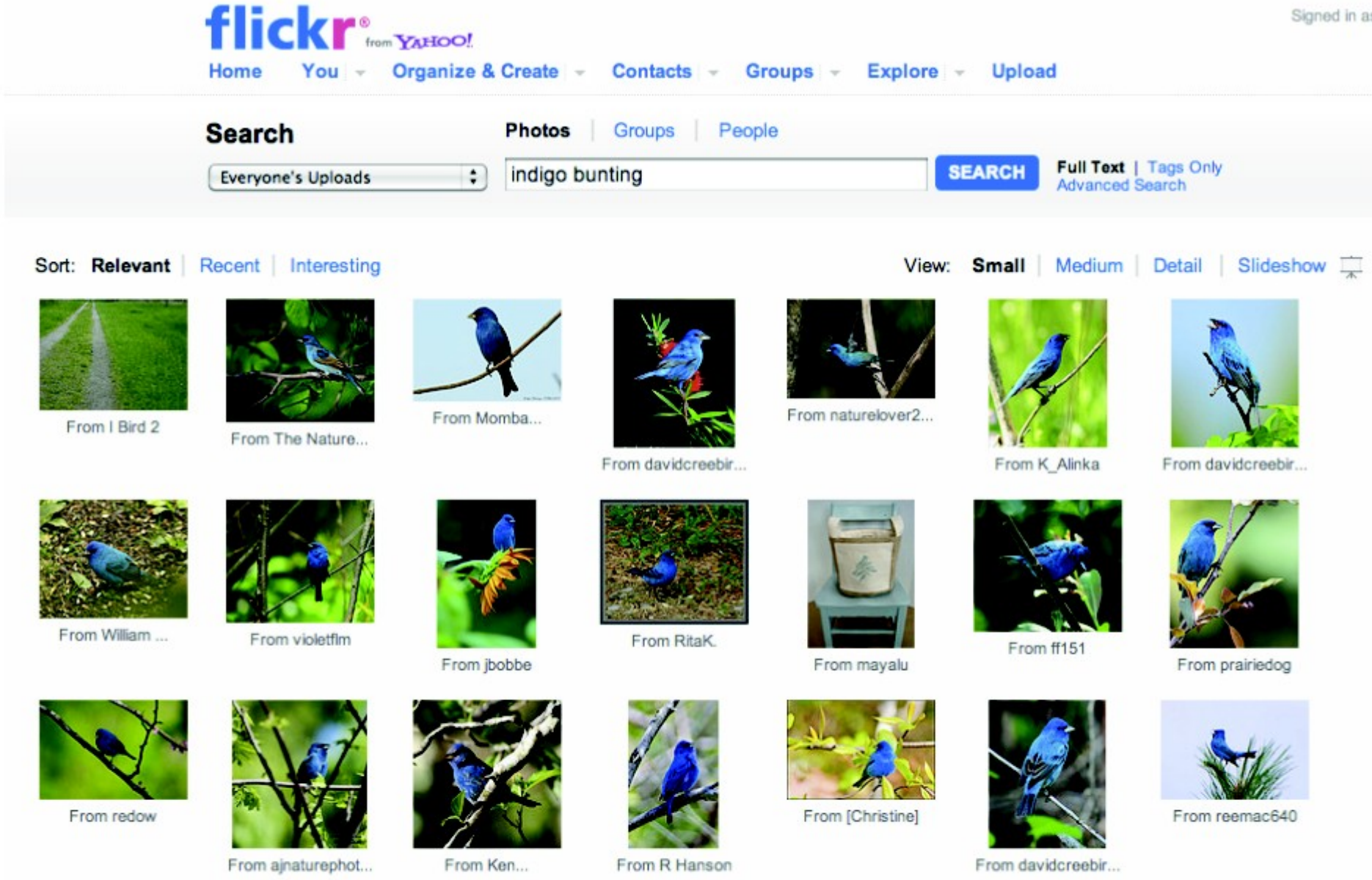
Problem Overview

Indigo Bunting






Slides from http://videolectures.net/nips2010_welinder_mwc/



5,926 results

Slides from http://videolectures.net/nips2010_welinder_mwc/

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

Slides from http://videolectures.net/nips2010_welinder_mwc/

Find the Indigo Bunting

A

A

A

A

A

A

Slides from http://videolectures.net/nips2010_welinder_mwc/

Find the Indigo Bunting

A B C D

A B C D

A B C D

A B C D

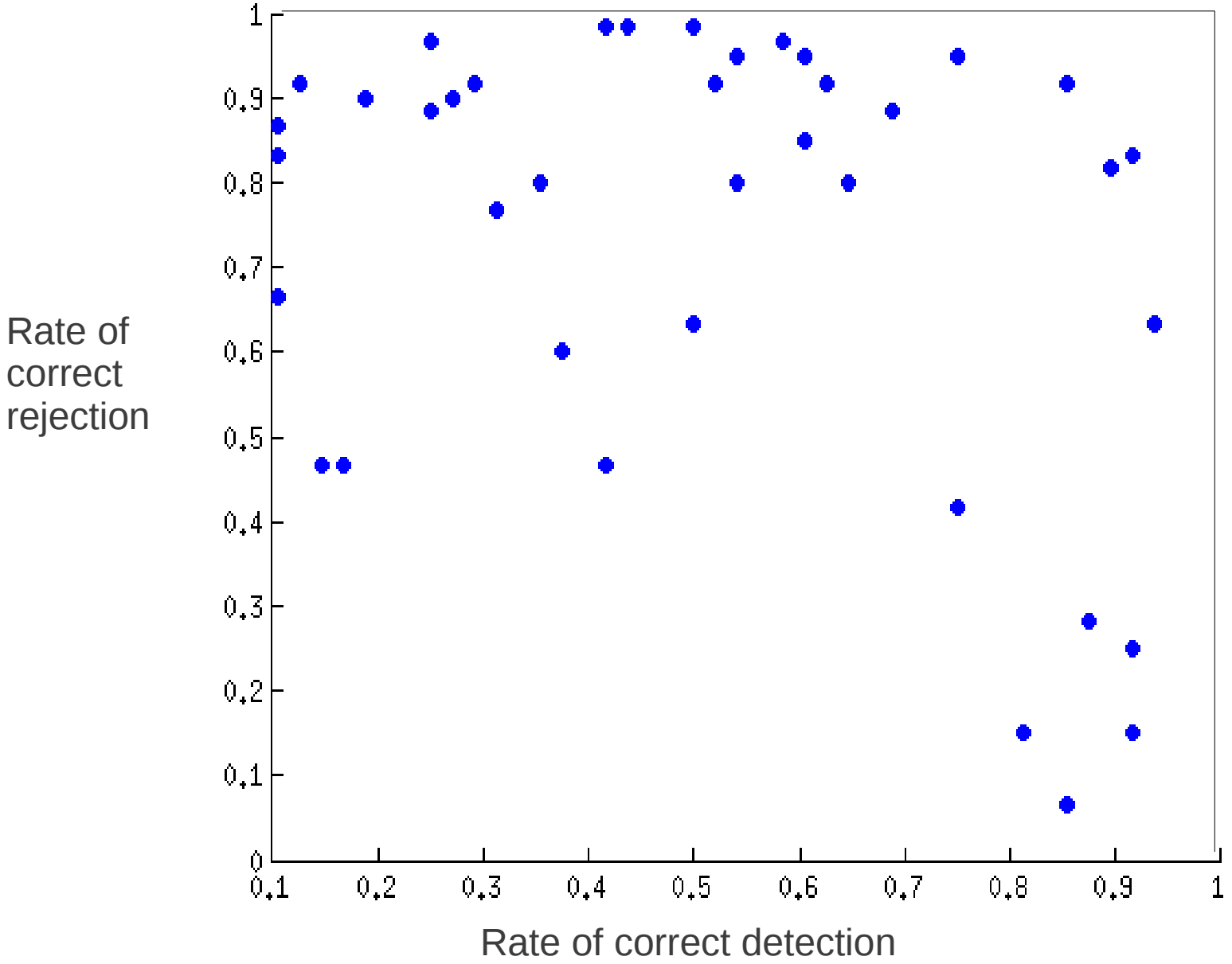
A B C D

A B C D

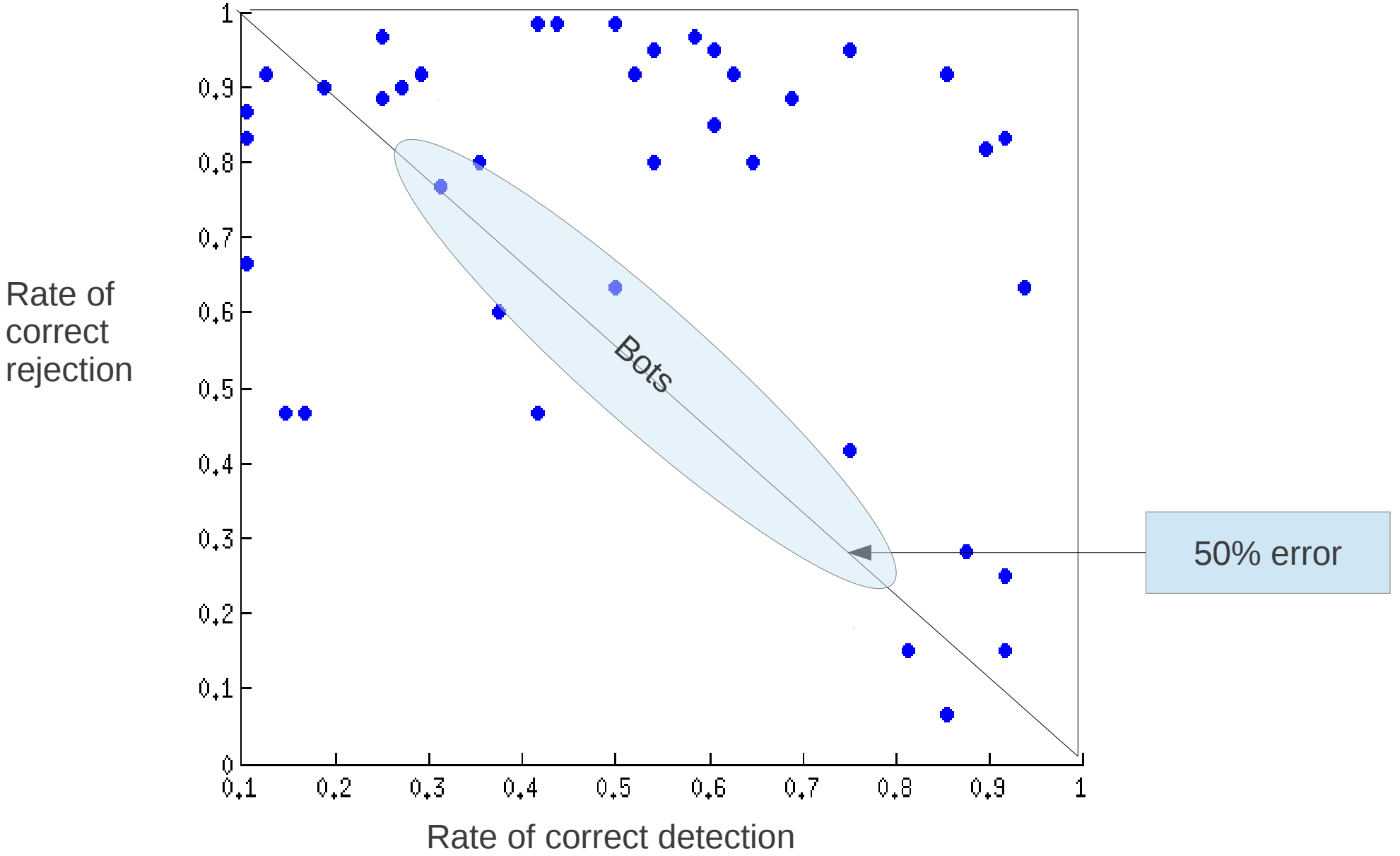
Slides from http://videolectures.net/nips2010_welinder_mwc/

Motivation

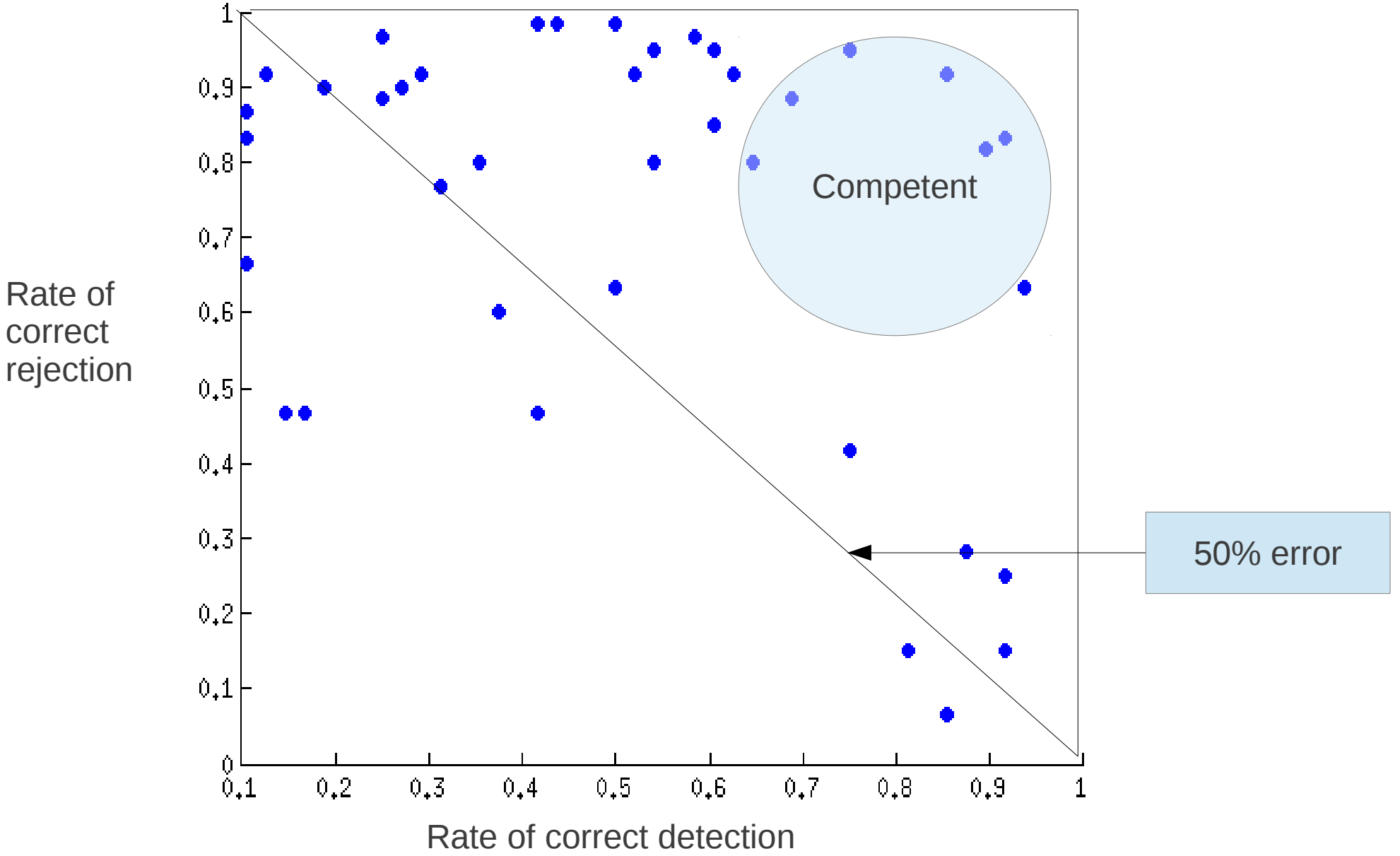
Distribution of Human Expertise – Task: Finding bluebirds Experiment #1



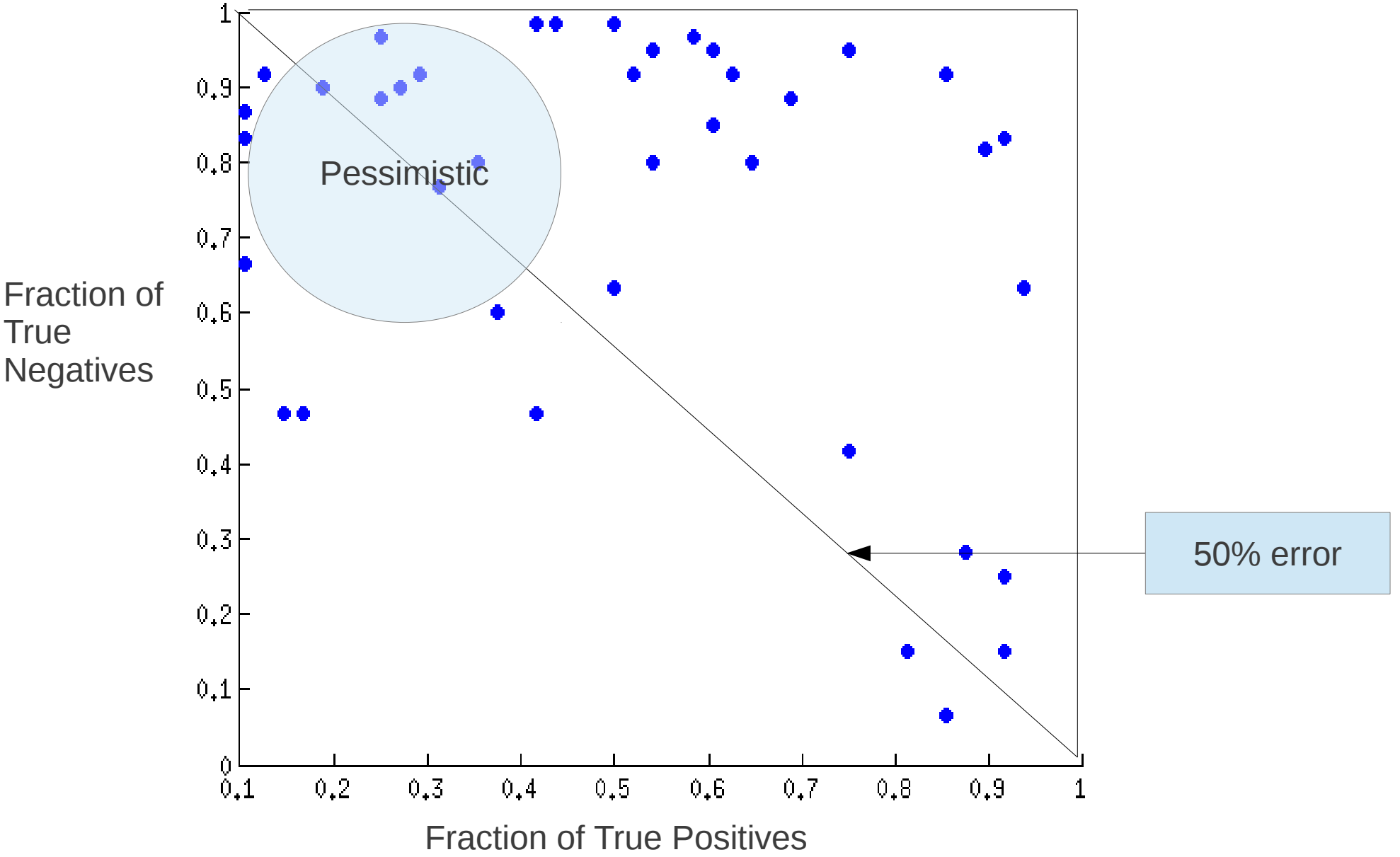
Distribution of Human Expertise – Task: Finding bluebirds
Experiment #1



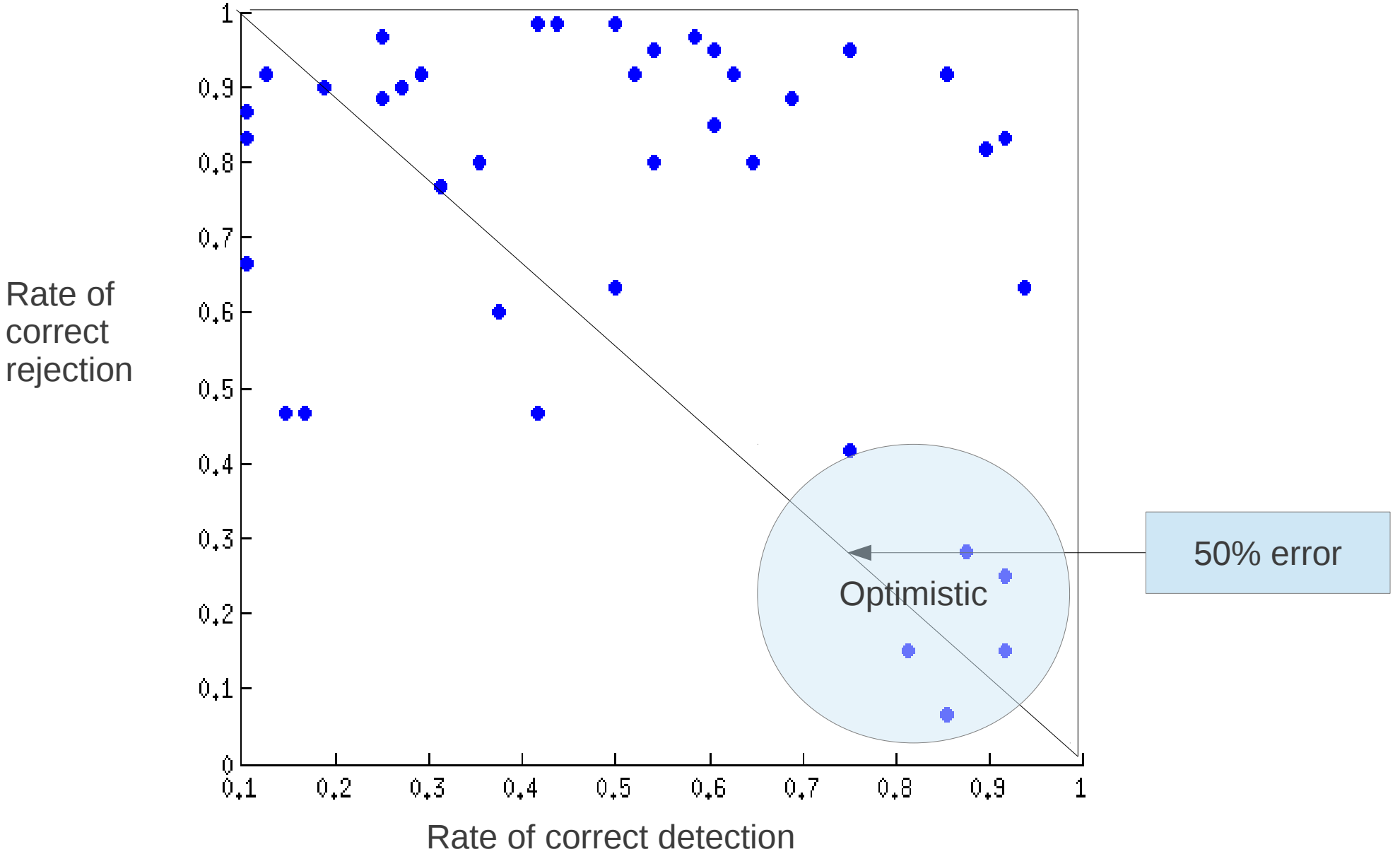
Distribution of Human Expertise – Task: Finding bluebirds
Experiment #1



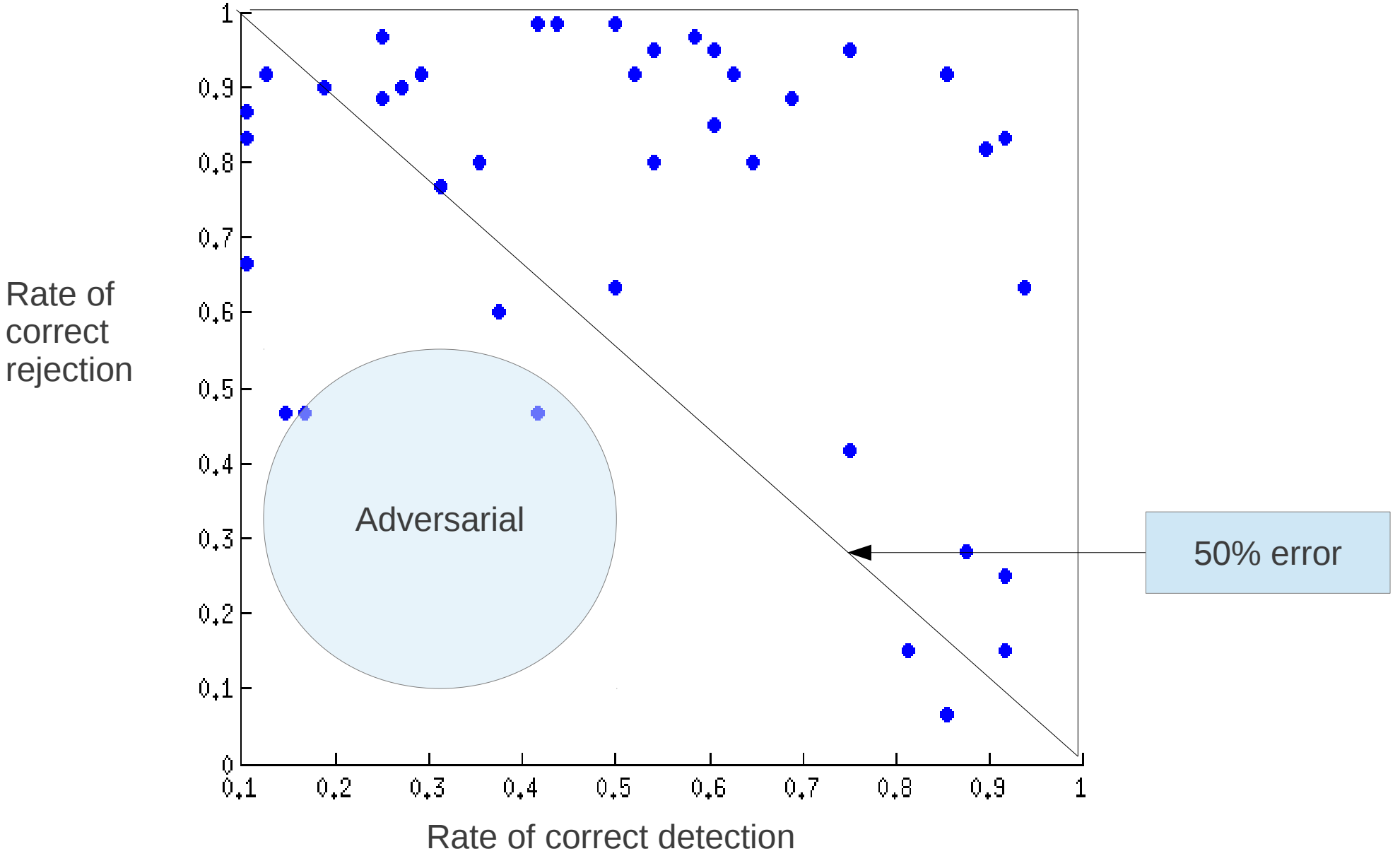
Distribution of Human Expertise – Task: Finding bluebirds
Experiment #1



Distribution of Human Expertise – Task: Finding bluebirds
Experiment #1

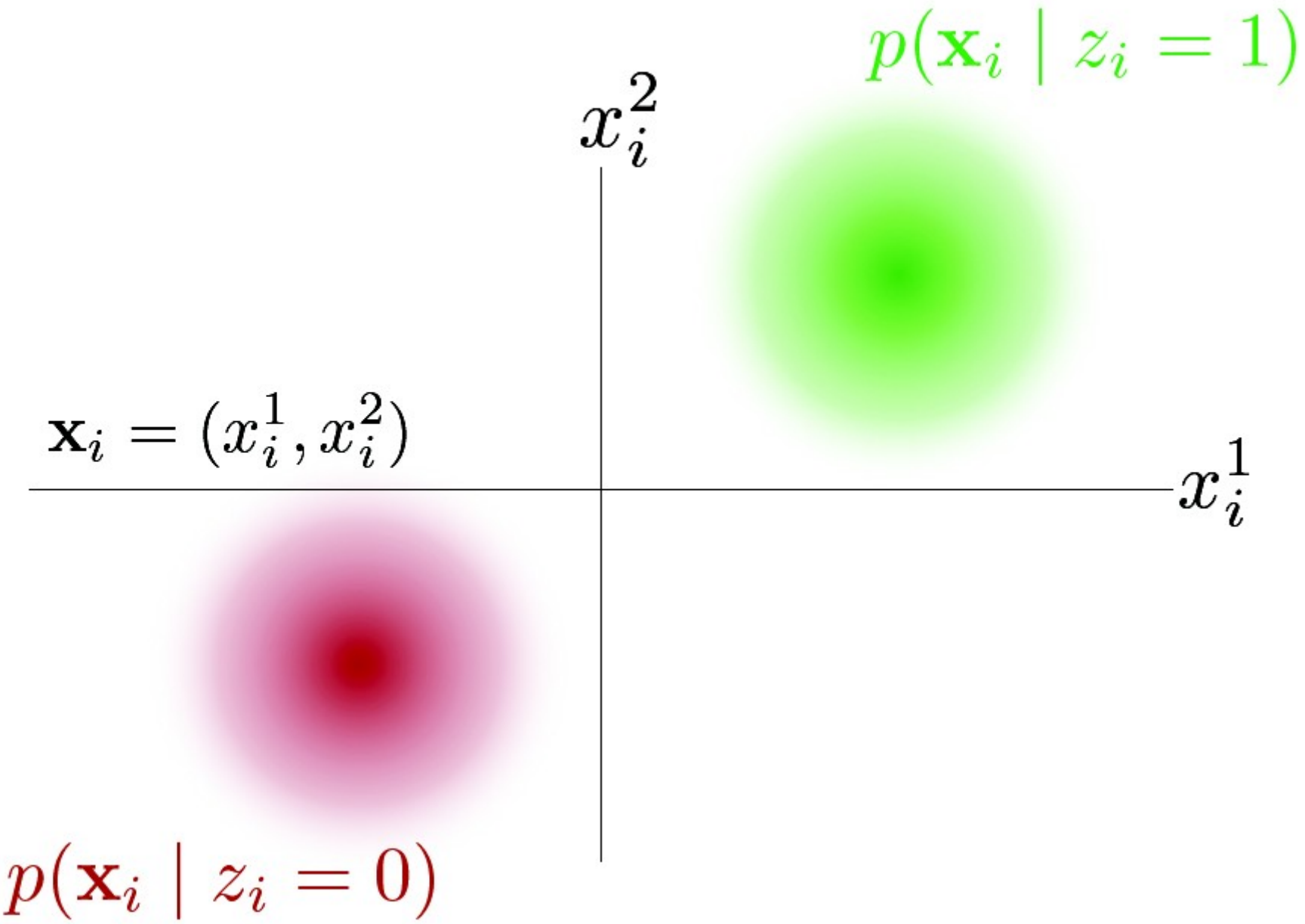


Distribution of Human Expertise – Task: Finding bluebirds Experiment #1



The Idea

Multidimensional signals and annotators



Slides from http://videolectures.net/nips2010_welinder_mwc/

Multidimensional signals and annotators

x_i^2

$p(\mathbf{x}_i \mid z_i = 1)$

$\mathbf{x}_i = (x_i^1, x_i^2)$

x_i^1

$p(\mathbf{x}_i \mid z_i = 0)$

Multidimensional signals and annotators

$p(\mathbf{x}_i \mid z_i = 1)$

x_i^2

$\mathbf{x}_i = (x_i^1, x_i^2)$

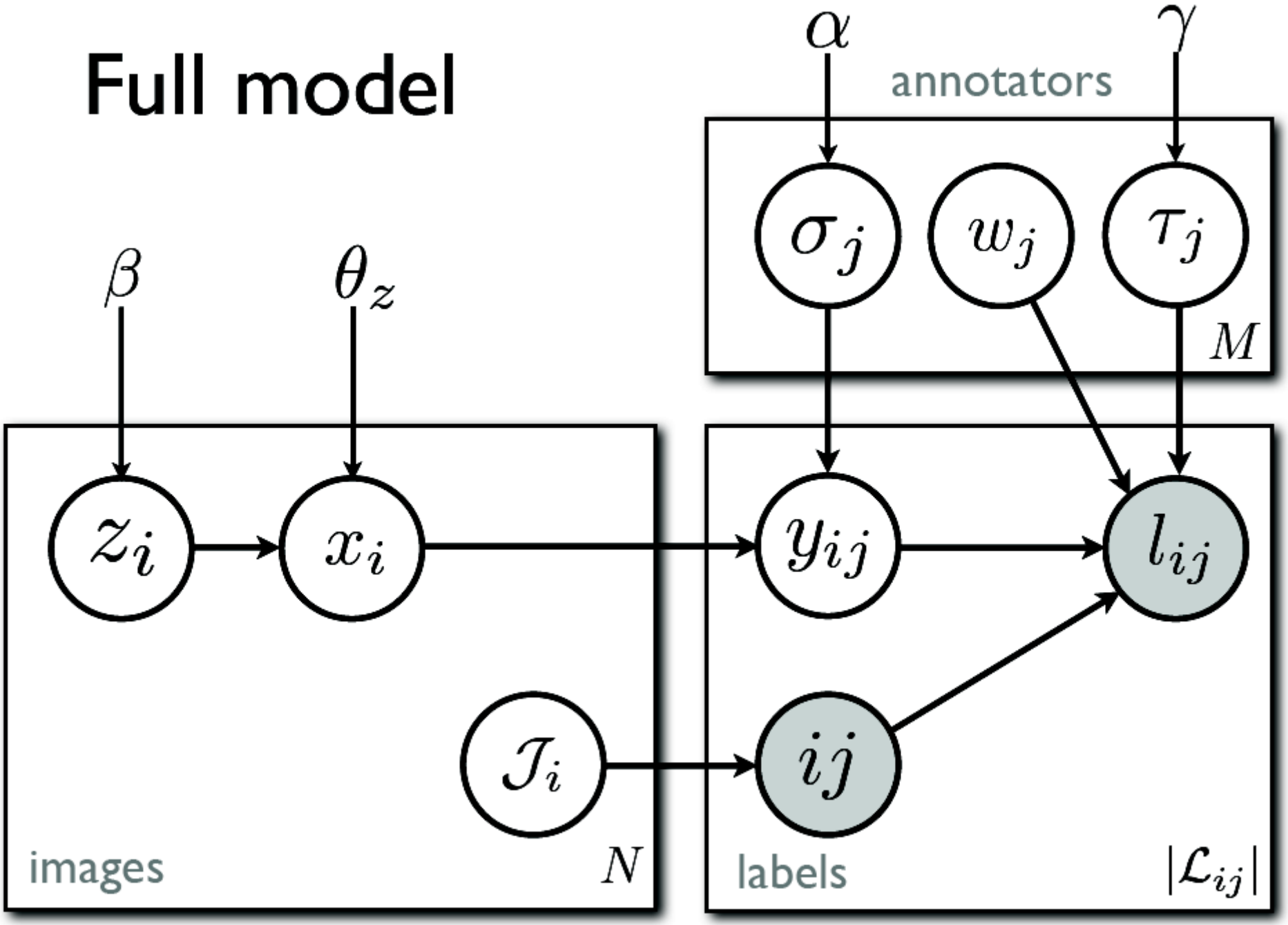
x_i^1

τ_j

$\mathbf{w}_j = (w_j^1, w_j^2)$

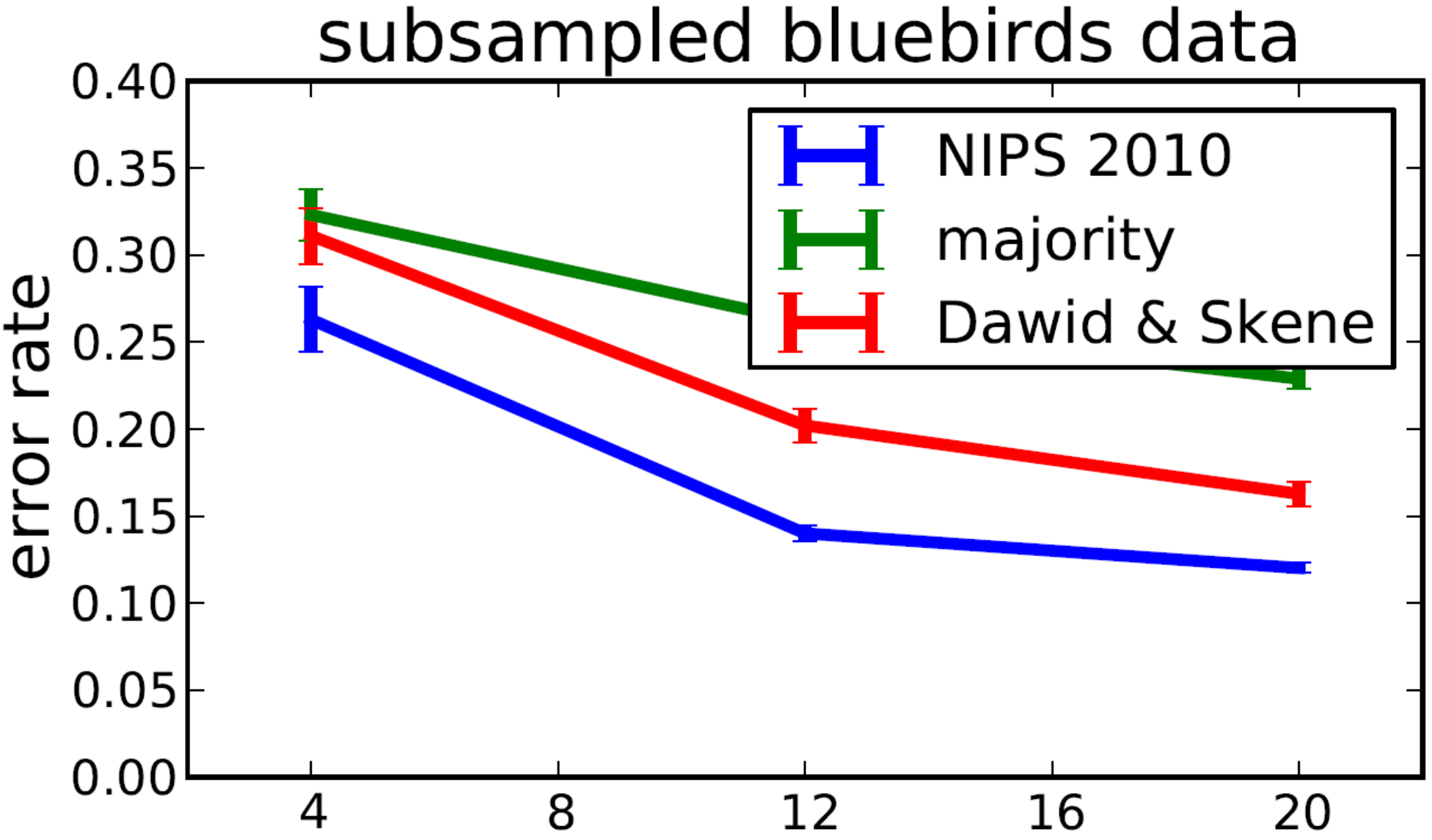
$p(\mathbf{x}_i \mid z_i = 0)$

Full model

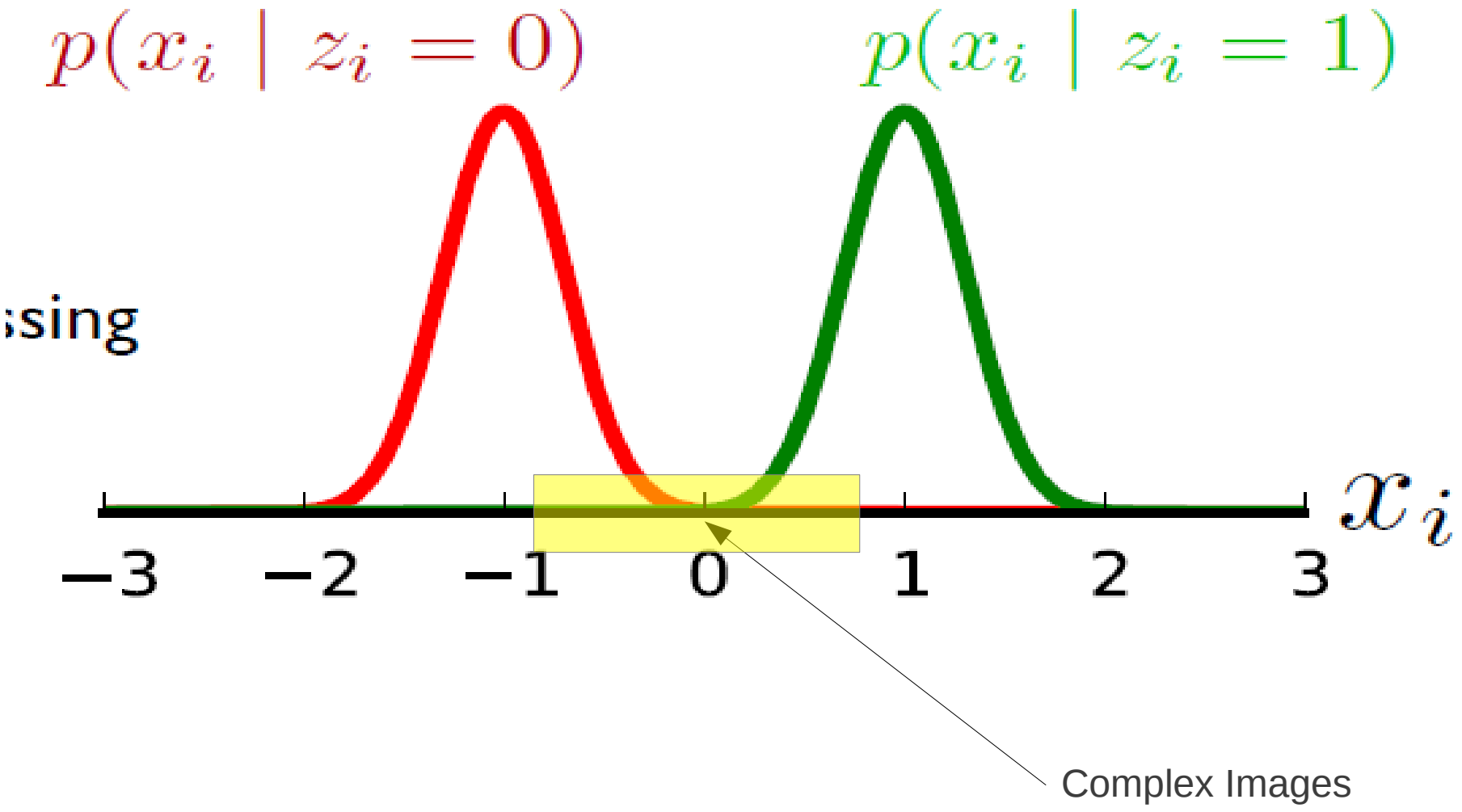


Slides from http://videolectures.net/nips2010_welinder_mwc/

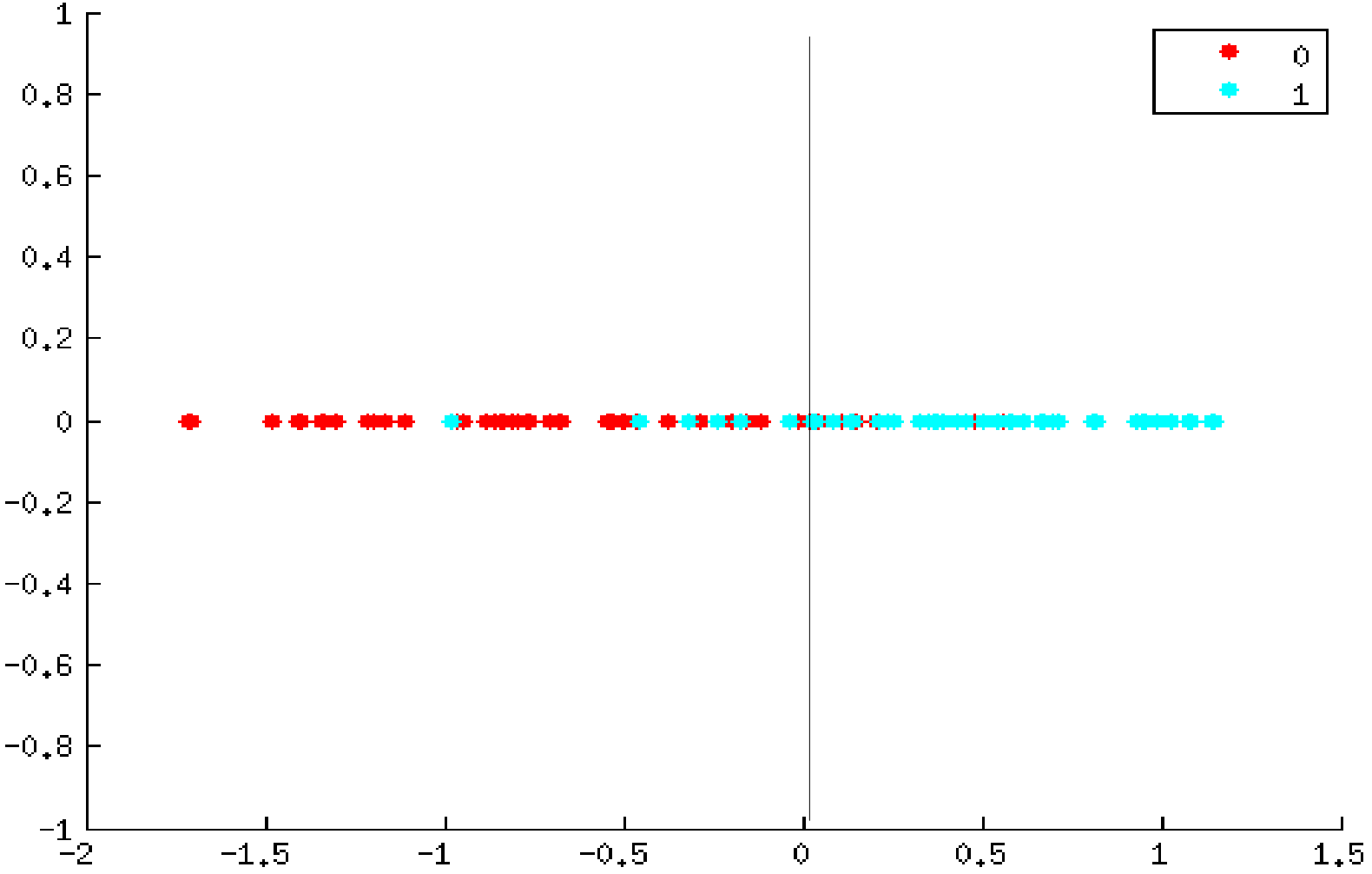
Error Rate for Bluebirds dataset



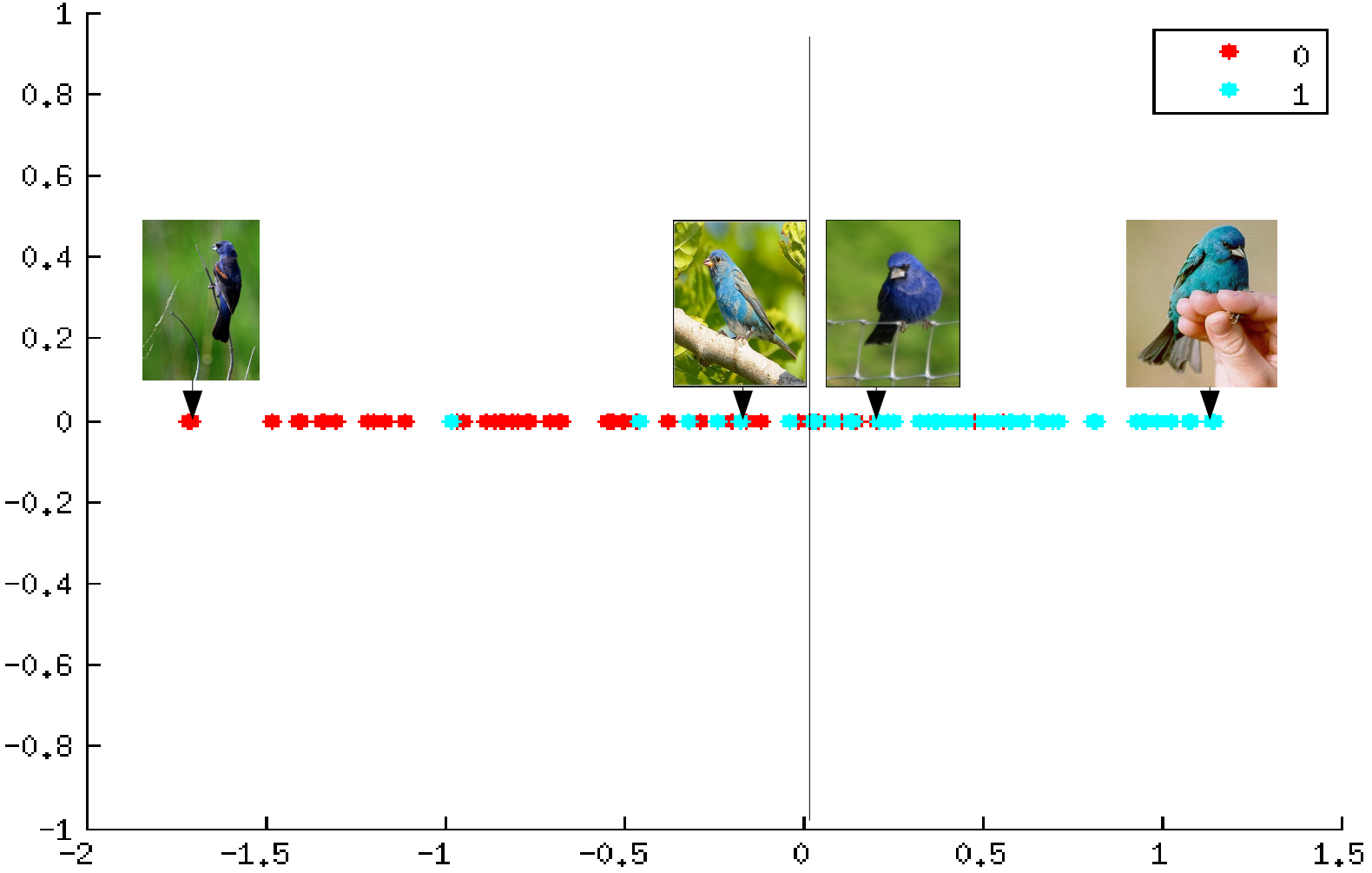
Estimating Image Difficulty



1D clusters from learned X_i values
Dataset: Bluebirds



1D clusters from learned X_i values
Dataset: Bluebirds



How do these learned image complexities compare with vision-based techniques?

Vision-based measure:
Predicted time* to label an image as a measure of image complexity

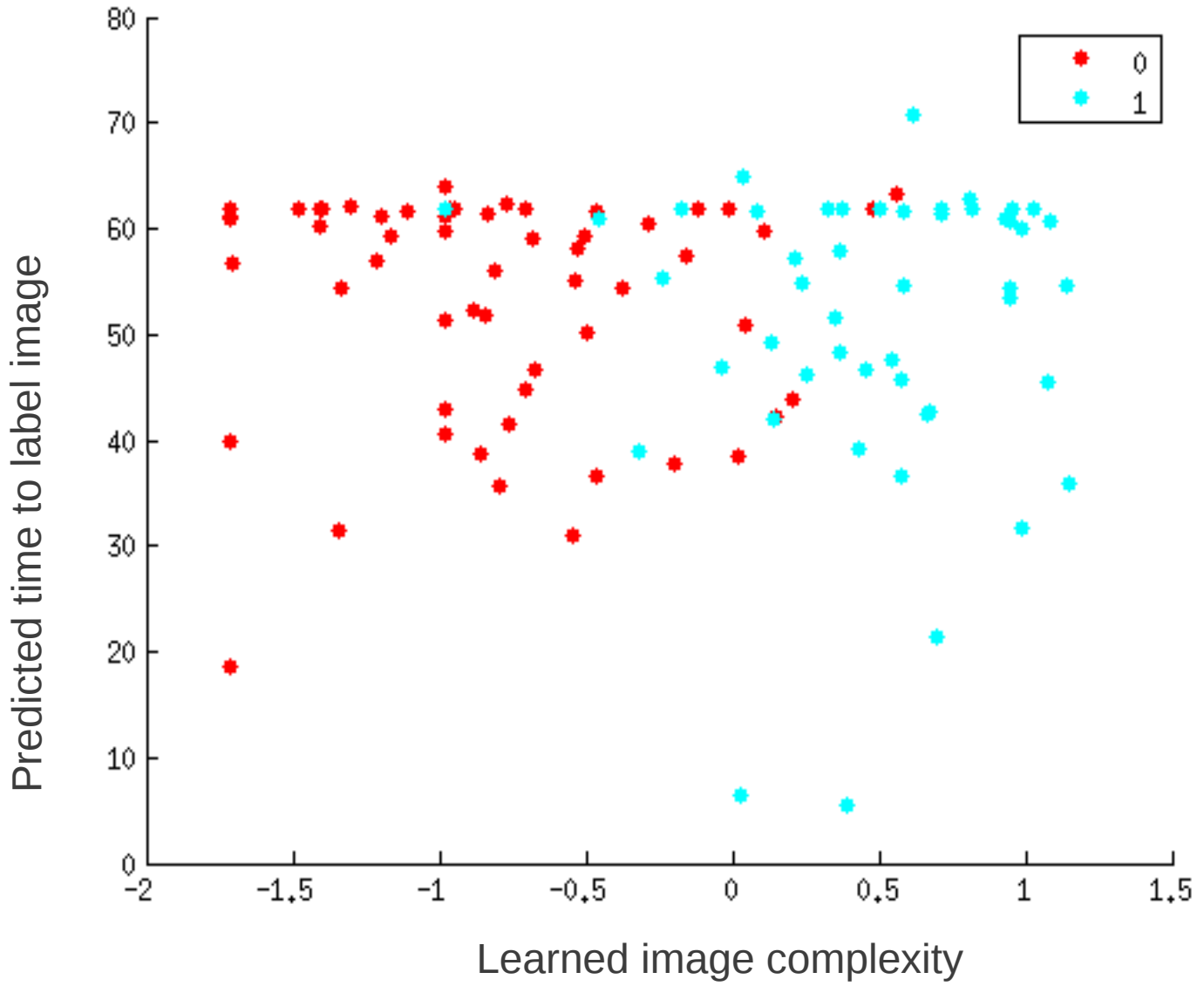
*What's It Going to Cost You? : Predicting Effort vs. Informativeness for Multi-Label Image Annotations. S. Vijayanarasimhan and K. Grauman. CVPR 2009

Approach:
Extract 2804-d feature vectors for MSRC dataset

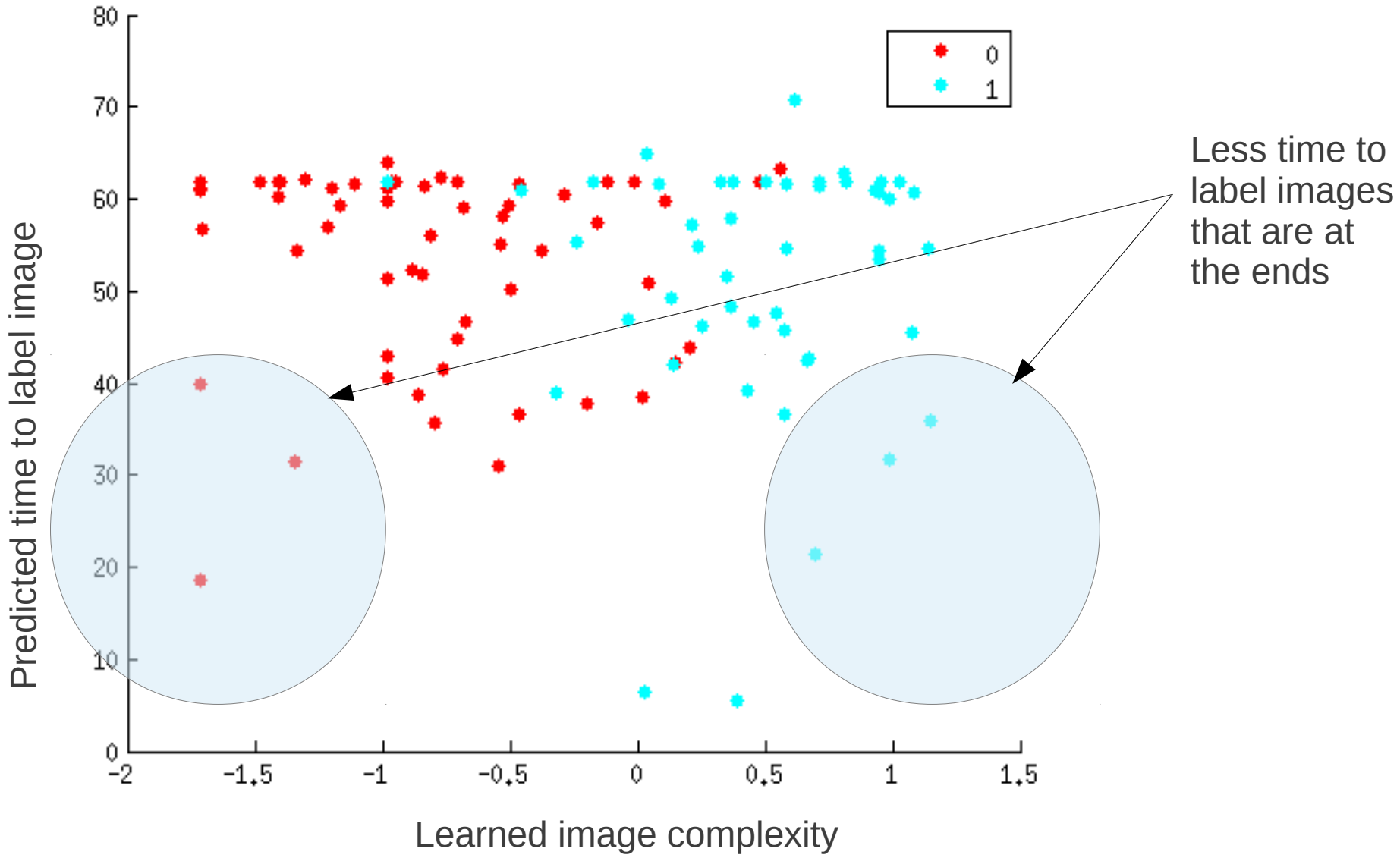
- Pyramid of HoG
- Color histogram
- Grayscale histogram
- Spatial pyramid of edge density (Canny edge)

Train a regressor on top 200 features selected using ReliefF
Predict time to label images for bluebirds dataset

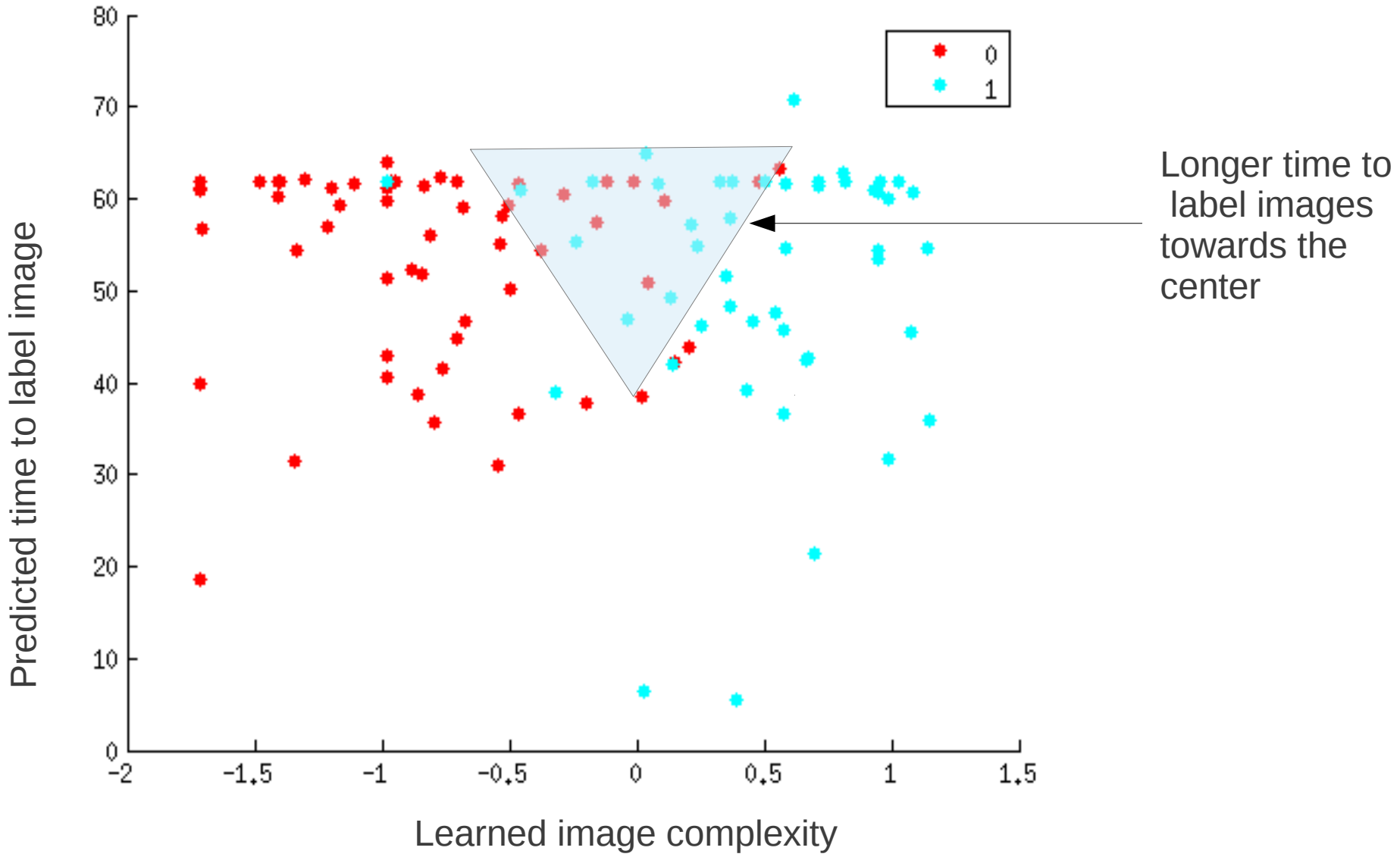
Vision-based complexity (vs) Learned Image Complexity



Vision-based complexity (vs) Learned Image Complexity



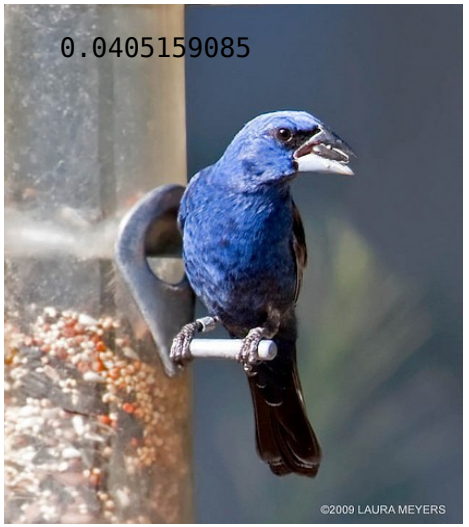
Vision-based complexity (vs) Learned Image Complexity



Qualitative Comparison

Complex Images – Examples
False negatives

Complex Images – Examples
False positive



Easy Images – Examples
True negatives

Easy Images – Examples
True positives

Task: Finding ducks

Mallard

American Black Duck

Canada Goose

Red Necked Grebe

Non-bird

Slides from http://videolectures.net/nips2010_welinder_mwc/

DUCKS

Mallard

American Black Duck

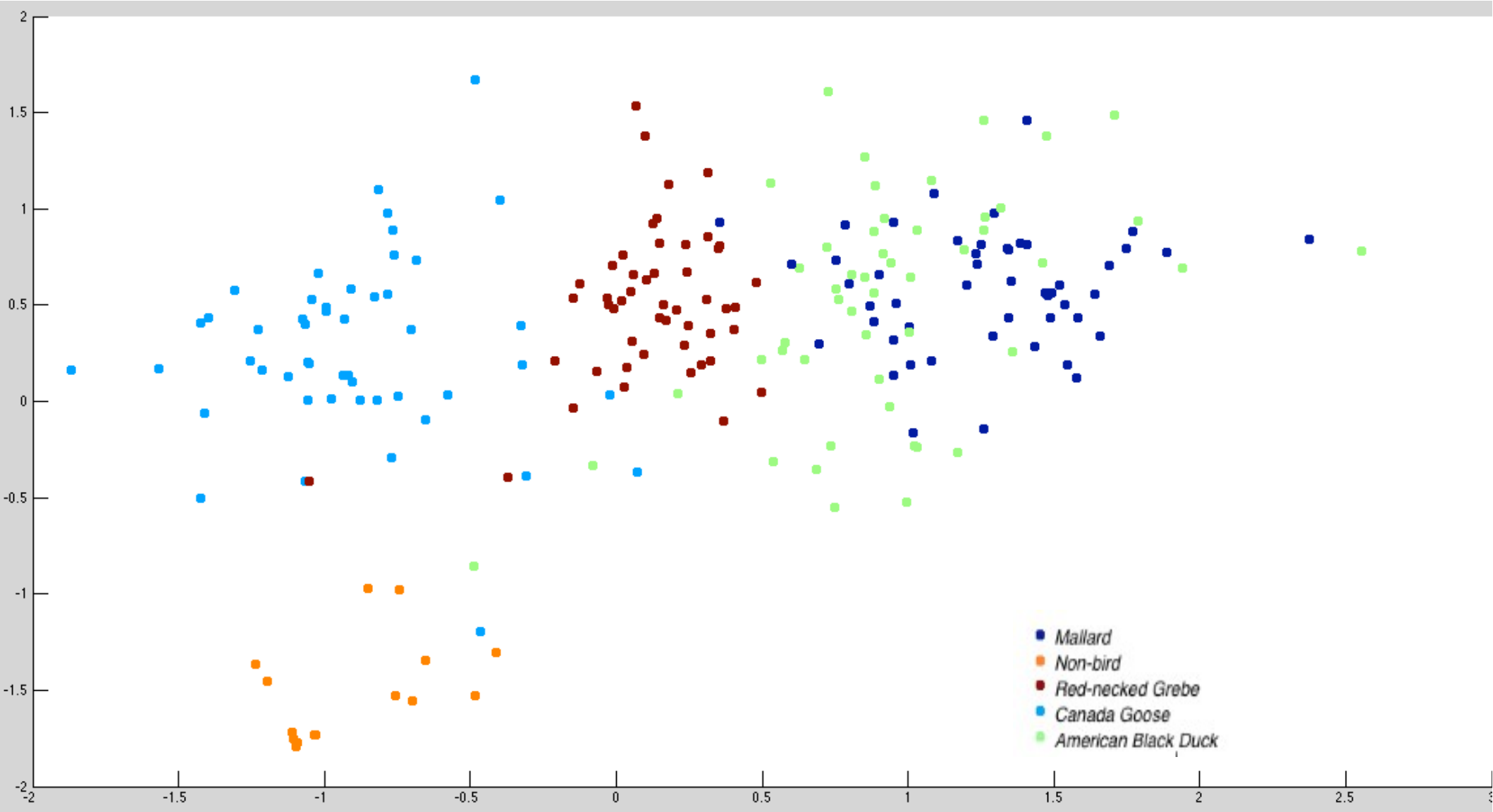
Canada Goose

Red Necked Grebe

Non-bird

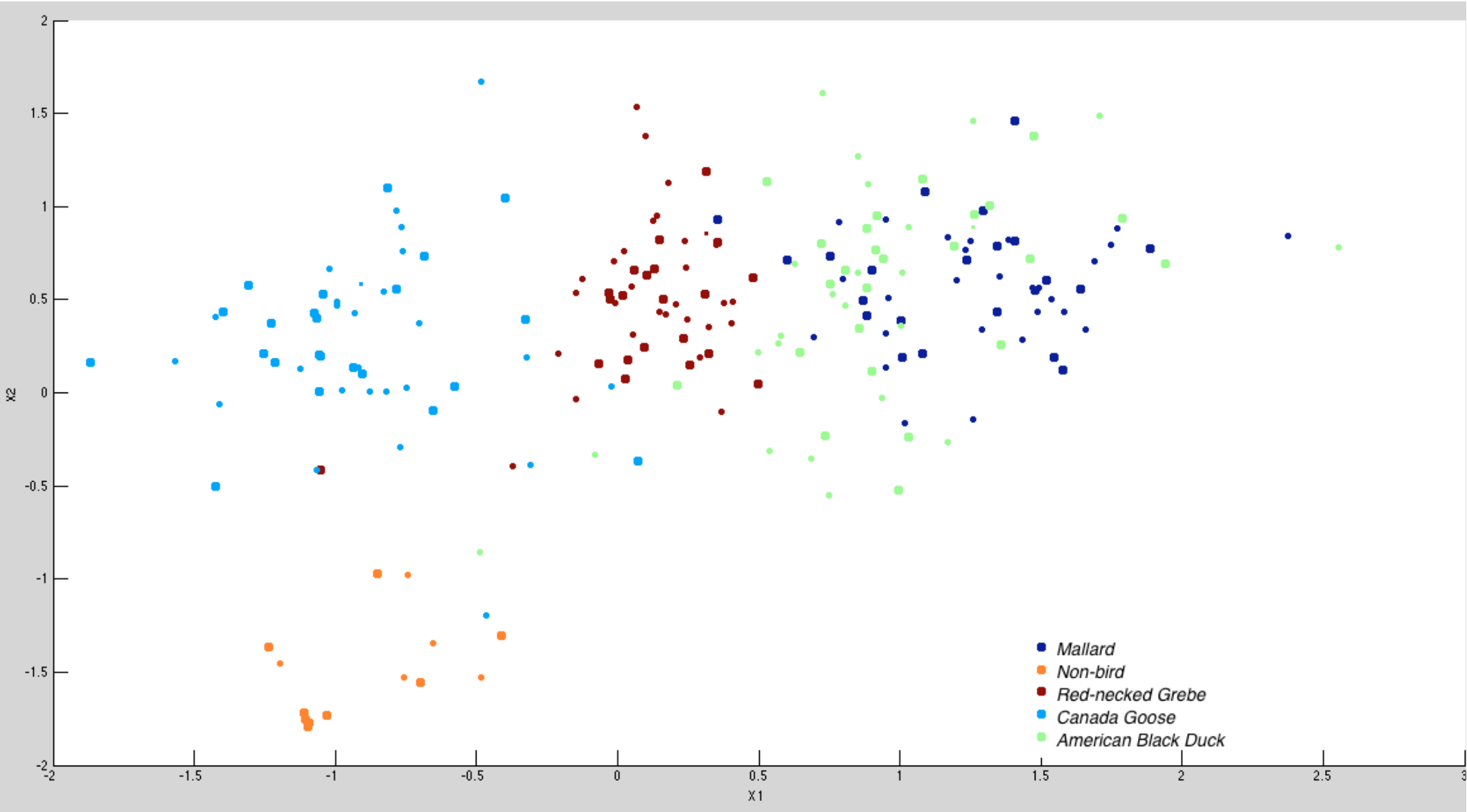
Slides from http://videolectures.net/nips2010_welinder_mwc/

2D clusters from learned X_i values
Dataset: Ducks



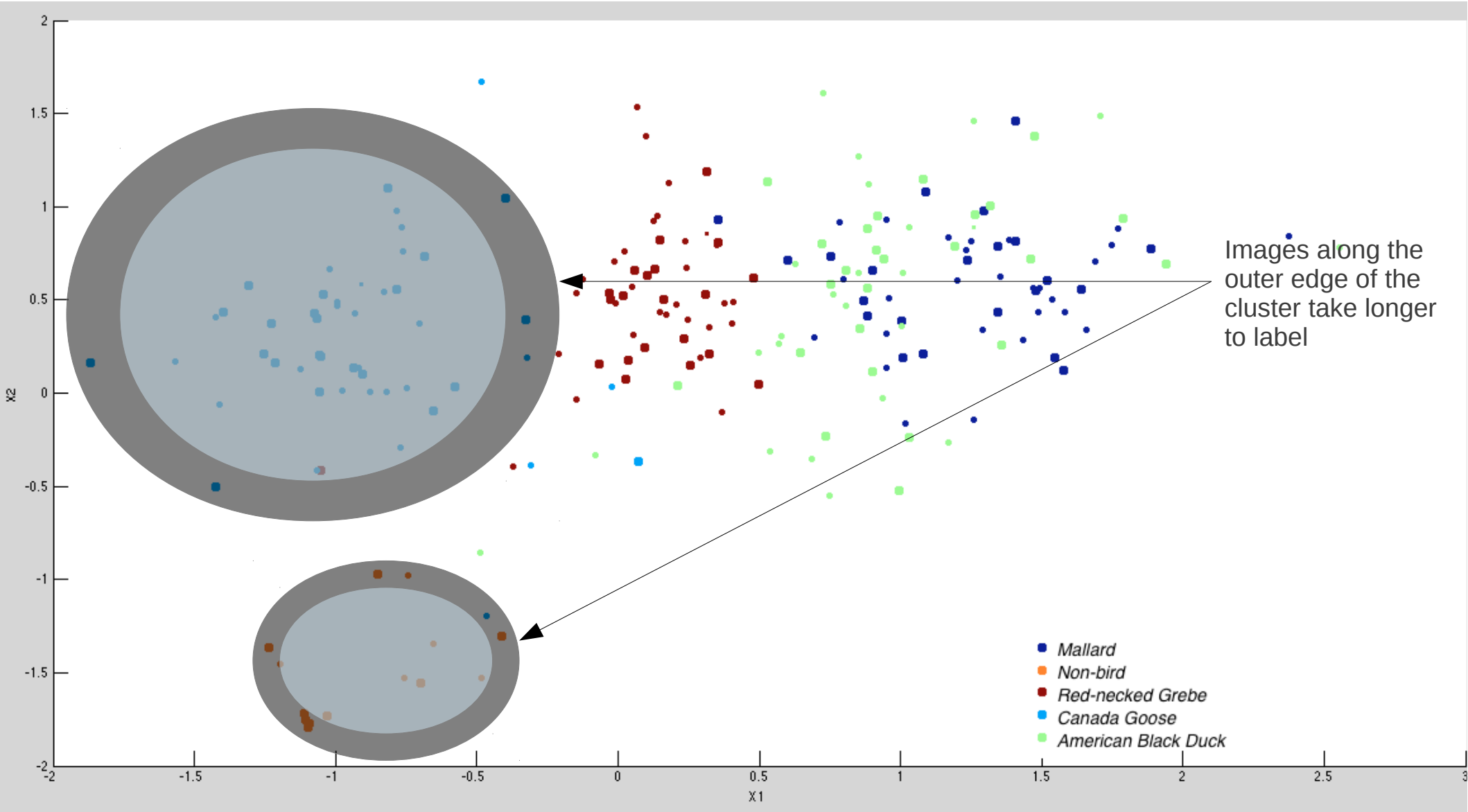
Recreated in MATLAB

Vision-based complexity (vs) Learned Image Complexity



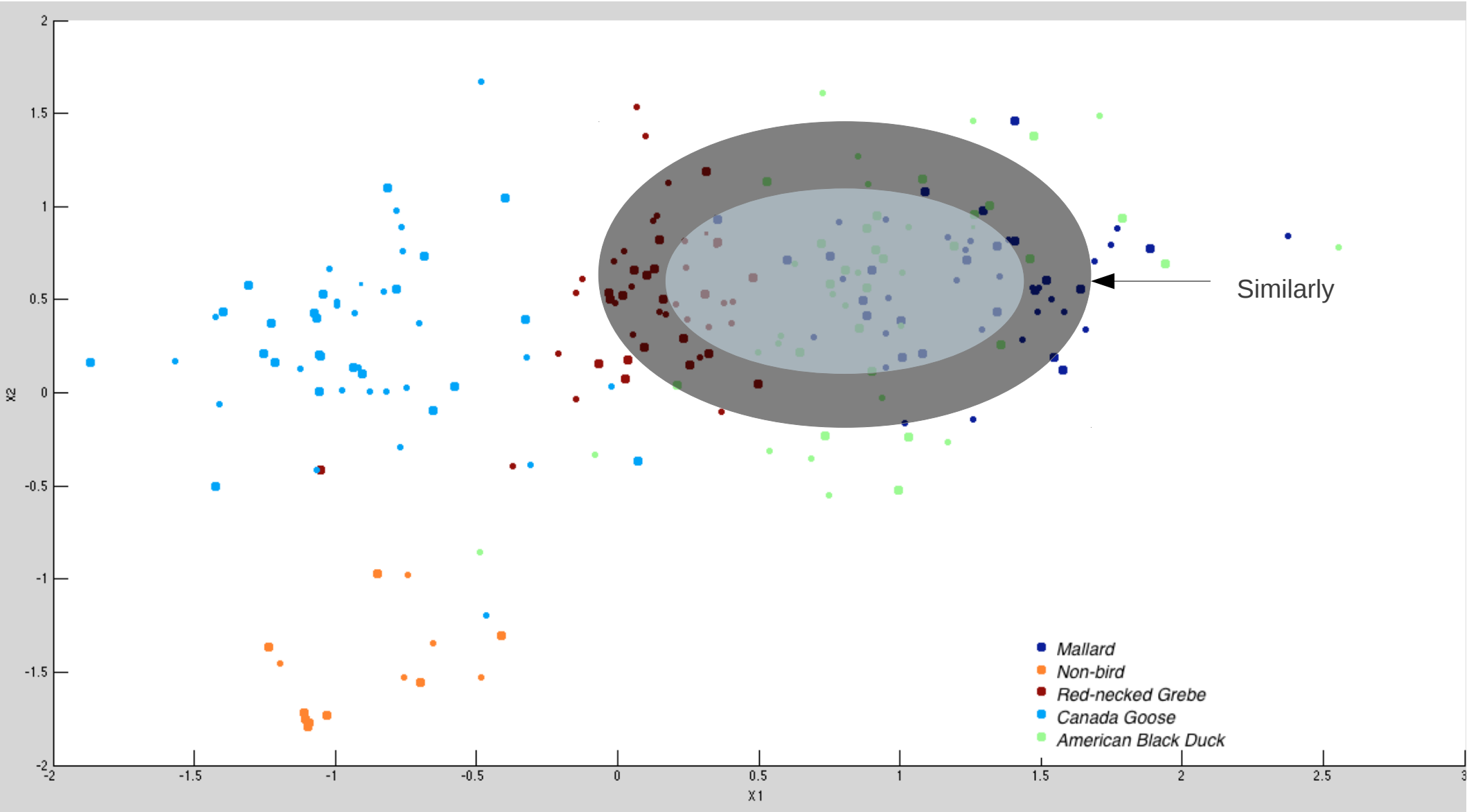
Recreated in MATLAB: **Size of point is proportional to the predicted time needed to label it**

Vision-based complexity (vs) Learned Image Complexity



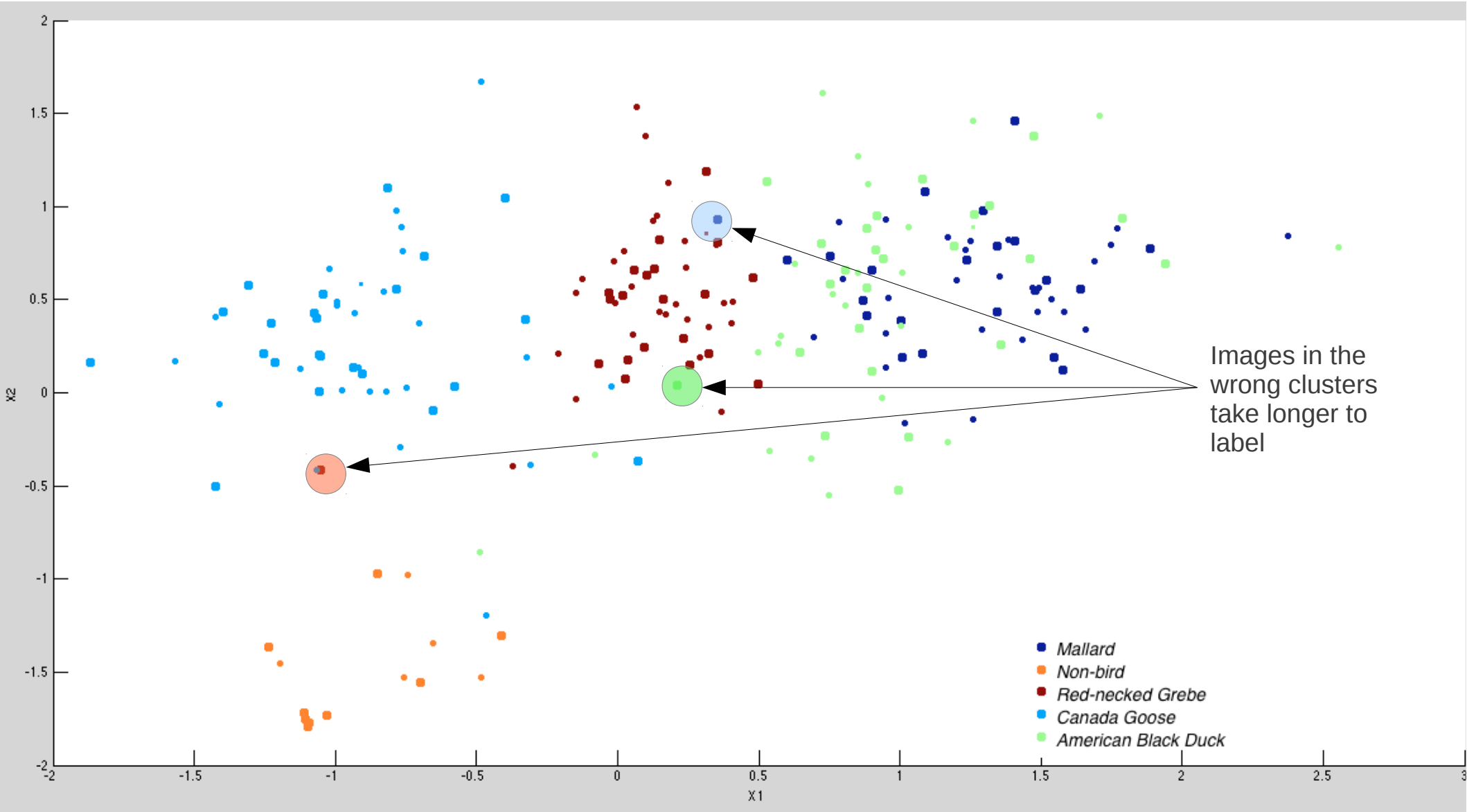
Recreated in MATLAB: **Size of point is proportional to the predicted time needed to label it**

Vision-based complexity (vs) Learned Image Complexity



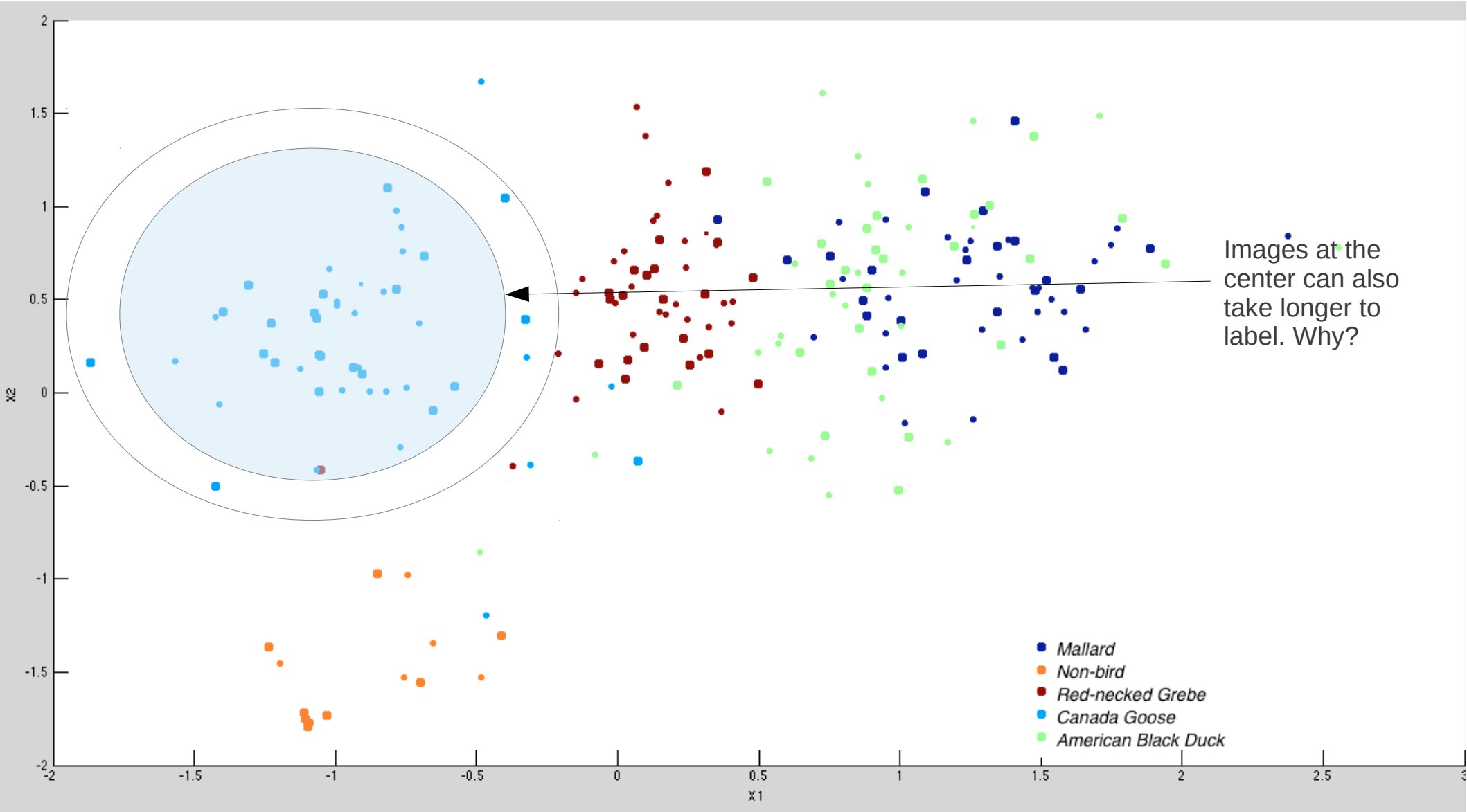
Recreated in MATLAB: **Size of point is proportional to the predicted time needed to label it**

Vision-based complexity (vs) Learned Image Complexity



Recreated in MATLAB: **Size of point is proportional to the predicted time needed to label it**

Vision-based complexity (vs) Learned Image Complexity



Recreated in MATLAB: **Size of point is proportional to the predicted time needed to label it**

Discussion

Is vision-based image complexity a good indicator of difficulty in labeling an image?

What are the other factors?

Discussion

Is vision-based image complexity a good indicator of difficulty in labeling an image?

What are the other factors?

Bird pose
Occlusions
Lighting

Discussion

1. The authors experiment only with a **2-dimensional** model of human expertise

   How would this model perform by increasing the number of intrinsic dimensions?

Extending this approach
to a video dataset
YouTube corpus

Example YouTube video with descriptions

http://youtu.be/FYyqIJ36dSU

A french bulldog is **playing** with a big ball
A small dog **chases** a big ball.
A French bulldog is **running** fast and **playing** with a blue yoga ball all by himself in a field.
The little dog **pushed** a big blue ball.
A dog is **playing** with a very large ball.
A dog **chases** a giant rubber ball around
A dog is **playing** with ball

Approach

YouTube corpus is not cut out for this task.

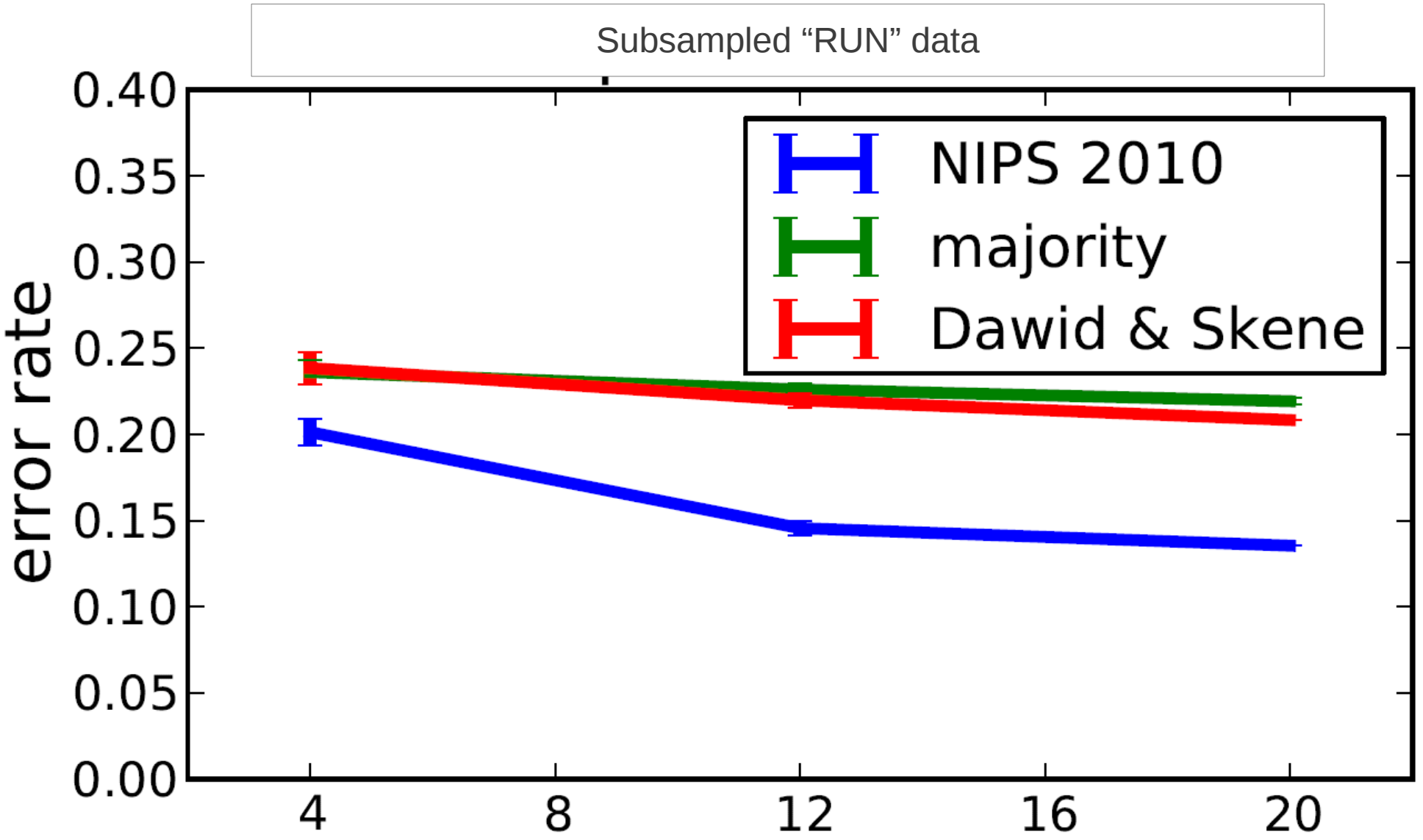
Consider predicting the presence of the activity **“run”**

1. Selected **50 videos where “run”** was the predicted activity using majority voting
2. Selected **30 videos where “play”** was the predicted activity using majority voting
3. Selected **20 videos where “walk”** was the predicted activity using majority voting

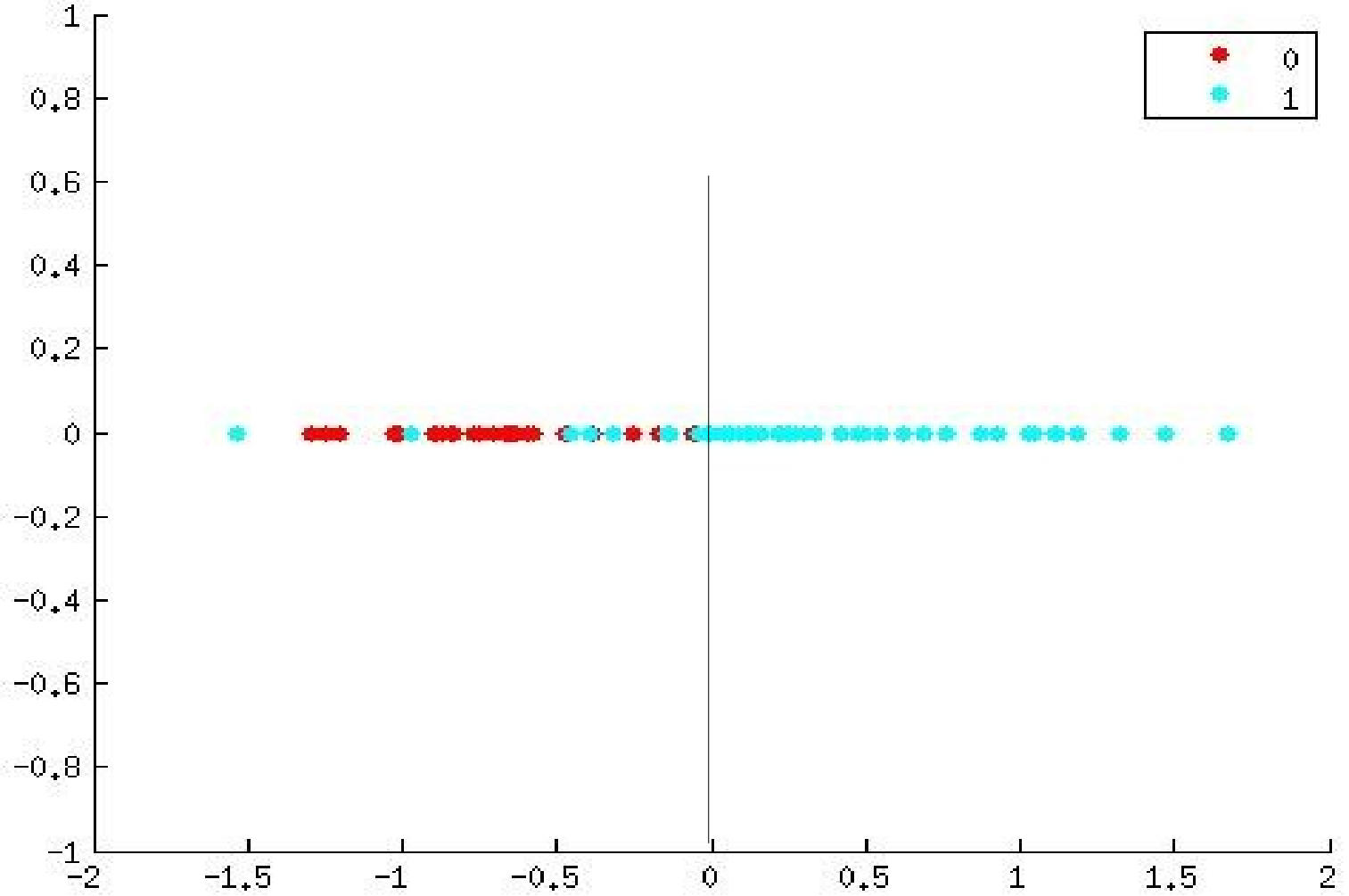
Ground Truth Labels were assigned accordingly

Each video had variable number of annotators. Picked the **20 most frequent annotators**.

Results


Subsampled “RUN” data
NIPS 2010
majority
Dawid & Skene
error rate
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
4
8
12
16
20

1D clusters from learned X_i values
Dataset: YouTube videos

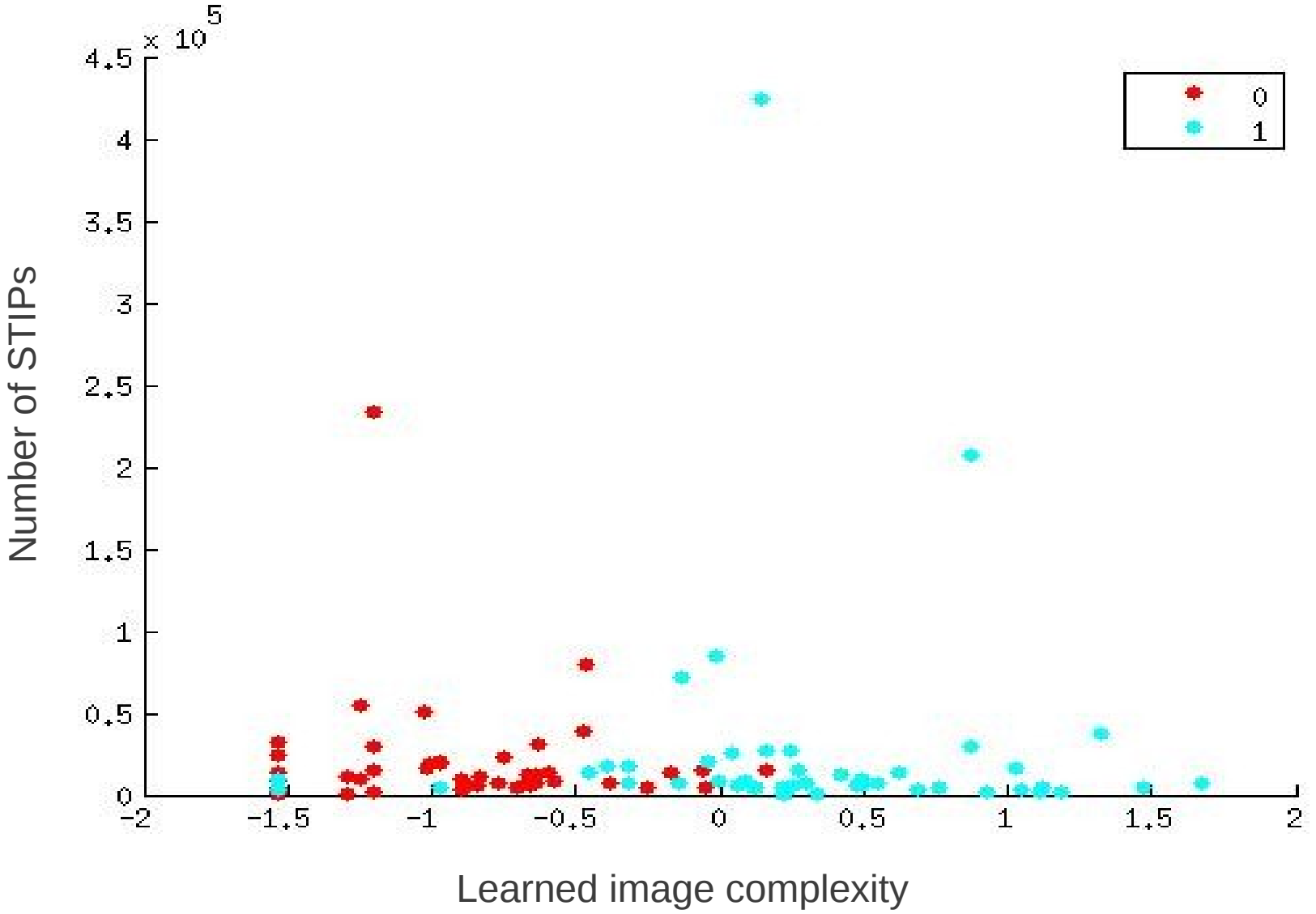


How do these learned video complexities compare with vision-based techniques?

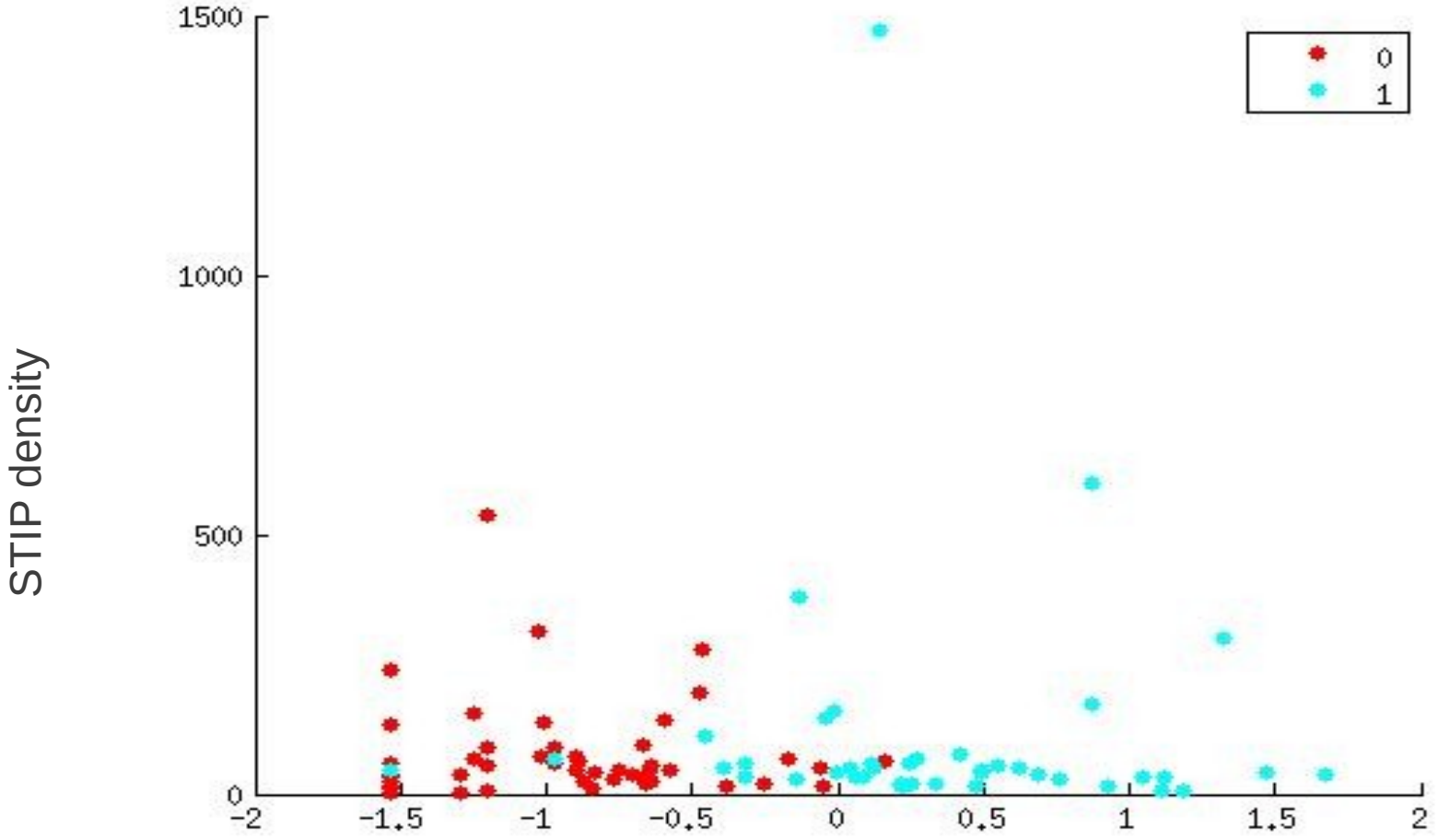
Vision-based measure:
Number of STIPS in the video
STIP density

*Learning Realistic Human Actions from Movies. I. Laptev, M. Marszałek, C. Schmid and B. Rozenfeld. CVPR 2008.

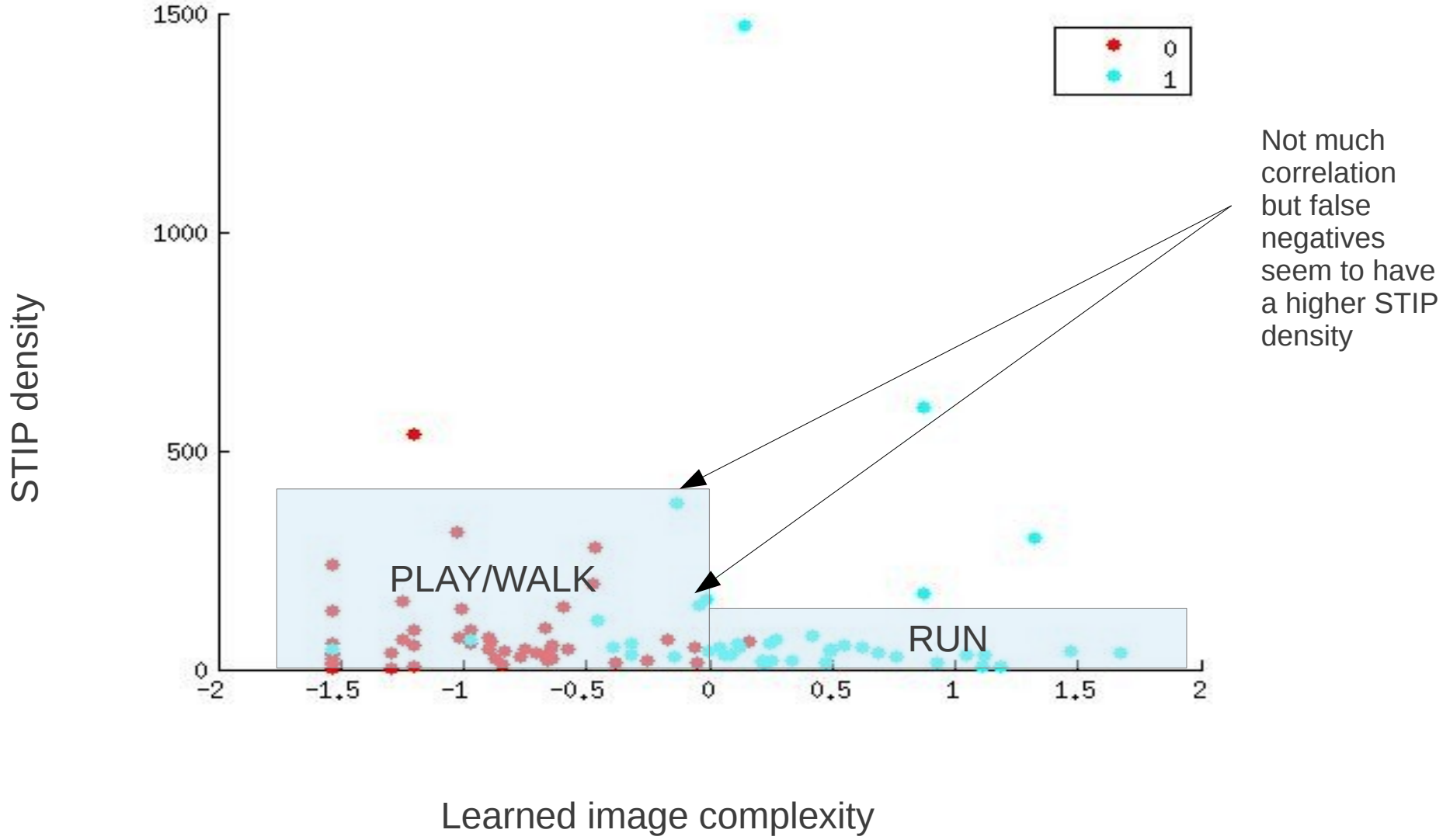
Vision-based complexity (vs) Learned Image Complexity



Vision-based complexity (vs) Learned Image Complexity



Vision-based complexity (vs) Learned Image Complexity



Discussion

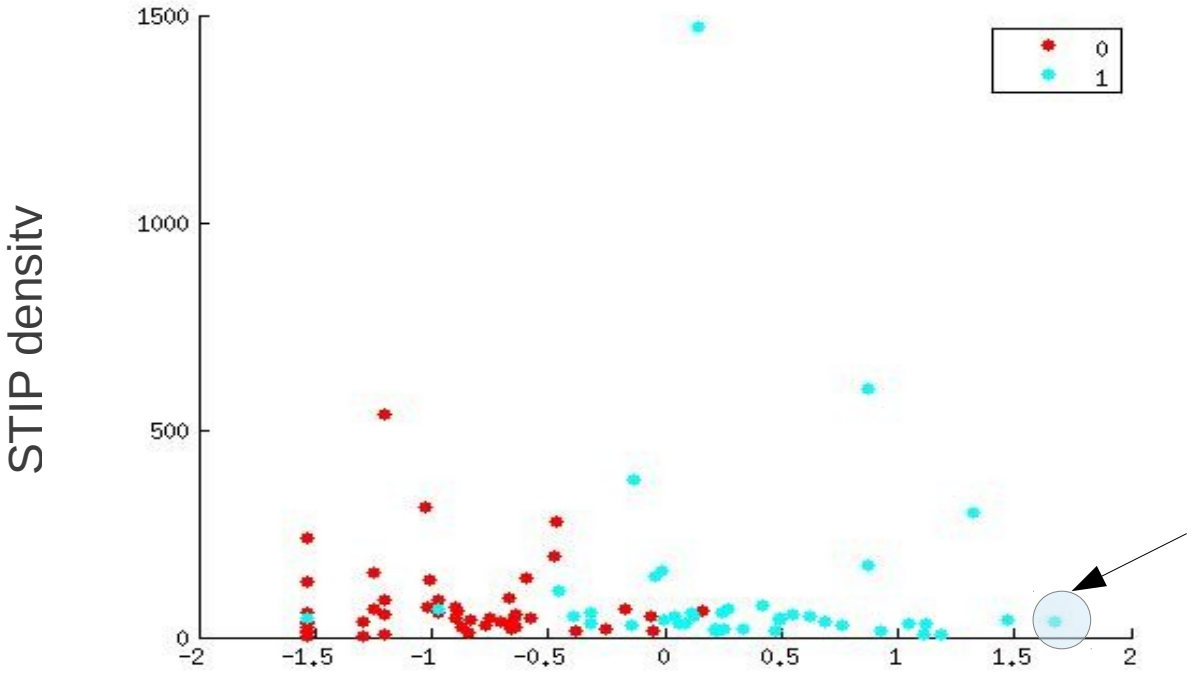
How can we quantify the complexity of a video?

STIP density?
Video length?
Variety in STIPS?
Confusion amongst multiple annotators?

How can we quantify the effort involved in labeling a video?

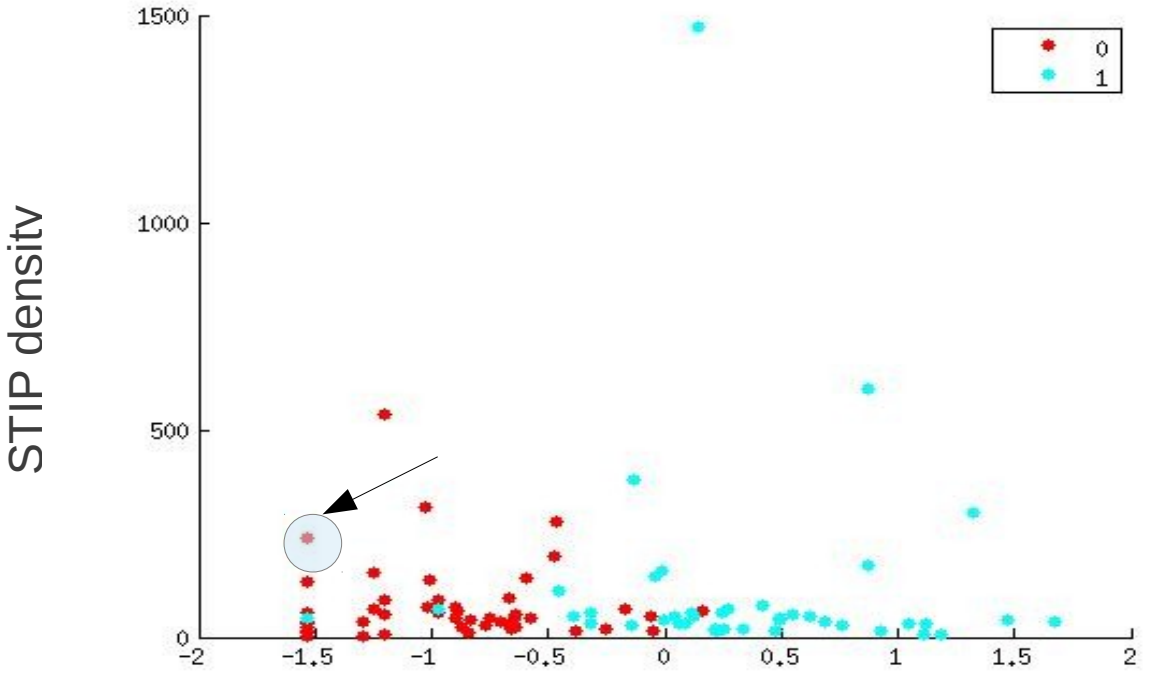
How do these relate to video ambiguity?

Qualitative Comparison – True positive



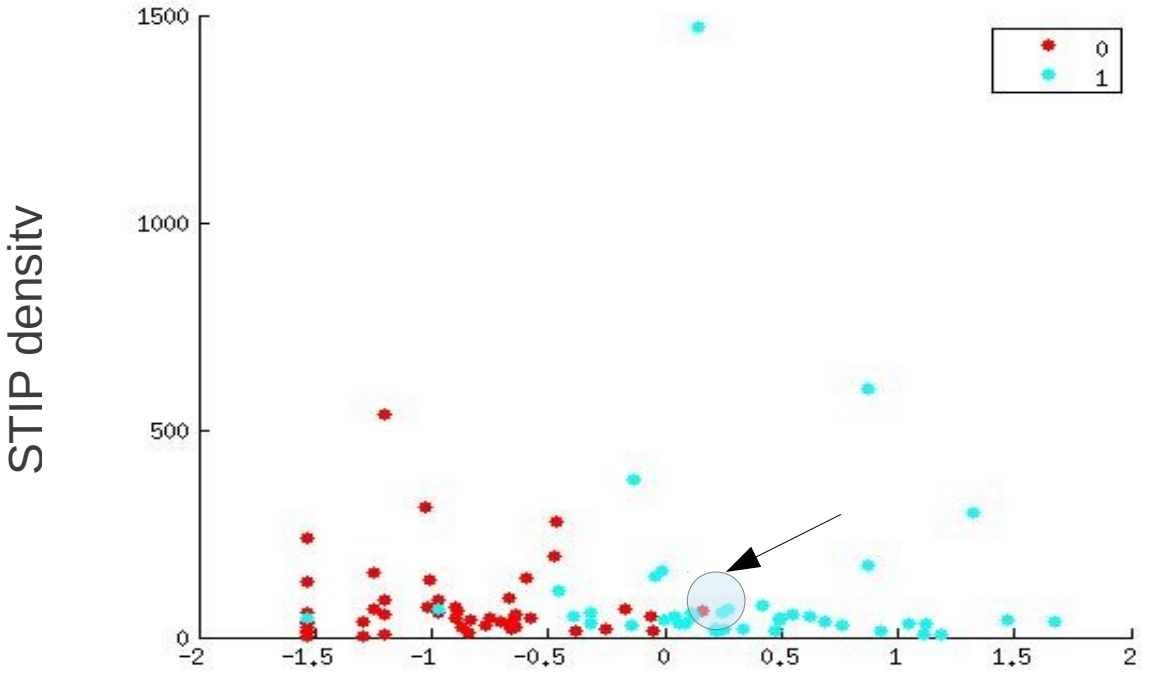
http://youtu.be/NKm8c_7mgx4

Qualitative Comparison – True negative



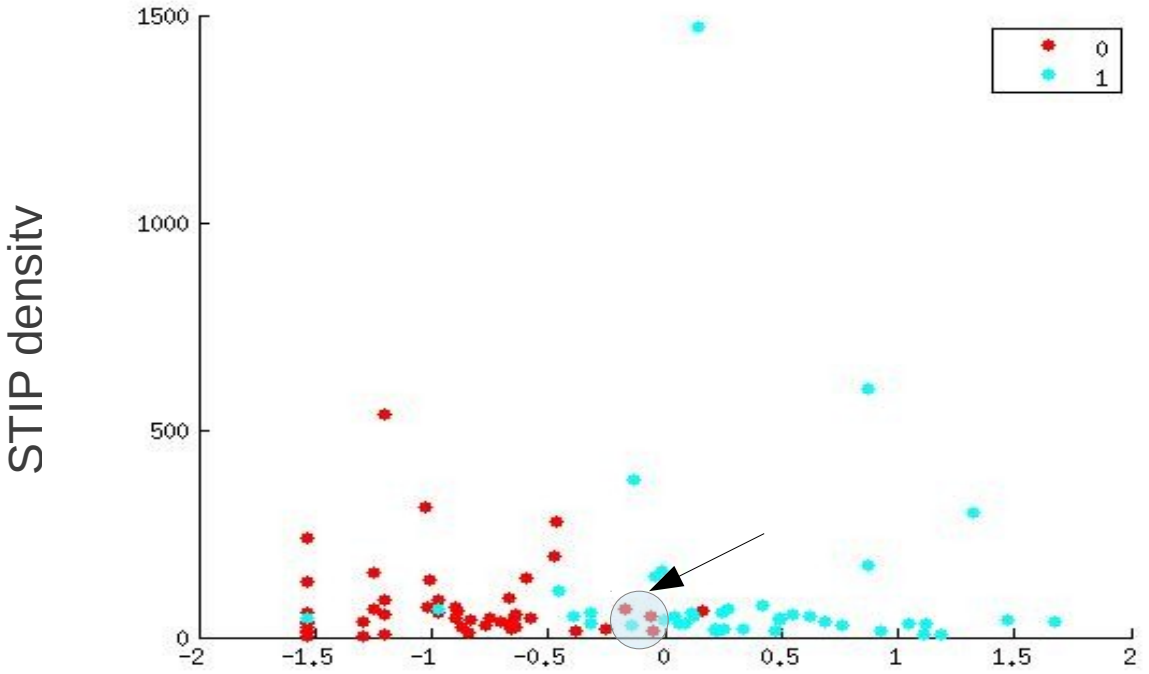
http://youtu.be/abiezv1p7SY

Qualitative Comparison – False positive



http://youtu.be/1l9Hx1kX_tQ

Qualitative Comparison – False negative



http://youtu.be/8miosT-Fs1k

Strengths

1. Each annotator is modeled as a multi-dimensional entity – competence, expertise, bias

2. Can be extended to any domain to estimate the ground truth with least error

3. Models image complexities without even seeing the image

4. The model discovers groups of annotators with varying skill sets.

Discussion

1. Image difficulties are learned from human annotations only, which is great!

   But would the model perform better if image difficulty was incorporated as a known parameter (using some vision-based technique) into the graphical model?