Perceptual Annotation: Measuring Human Vision to Improve Computer Vision

Walter J. Scheirer

Department of Molecular and Cellular Biology, School of Engineering and Applied Sciences, Center for Brain Science
Harvard University

Motivation

Persistent gap between state-of-the-art computer vision systems and human performance.

ILSVRC2014 Best Classification Result: 6.66% Error Rate

ILSVRC2014 Best Detection Result: 43.9 mean AP

ImageNet Classification

Krizhevsky et al. 2012

ImageNet Detection

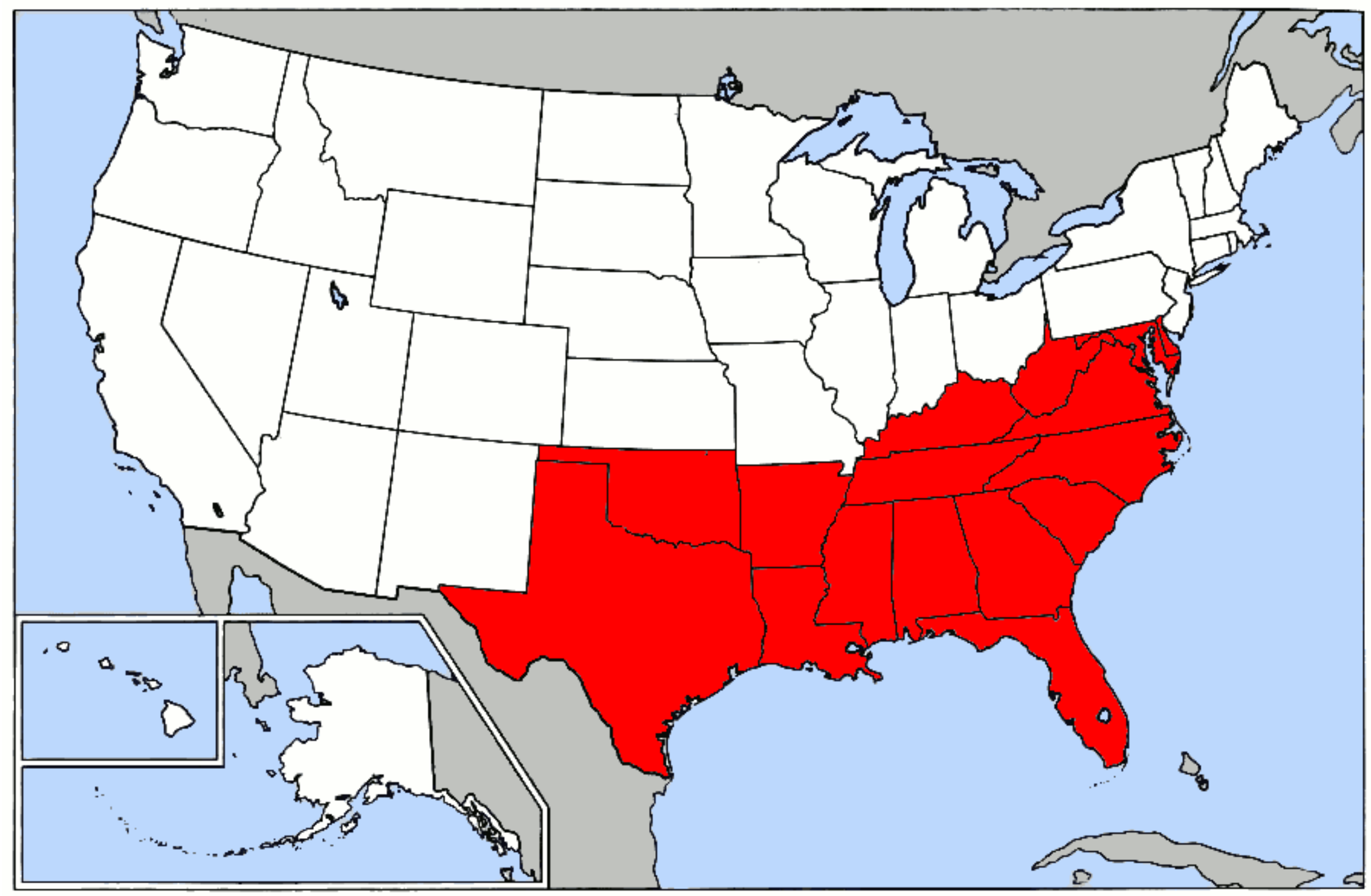
Girshick et al. 2014

There is concern that such methods will asymptote well below the level of human performance.

Learnability

Imagine a newly arrived foreigner in the US…

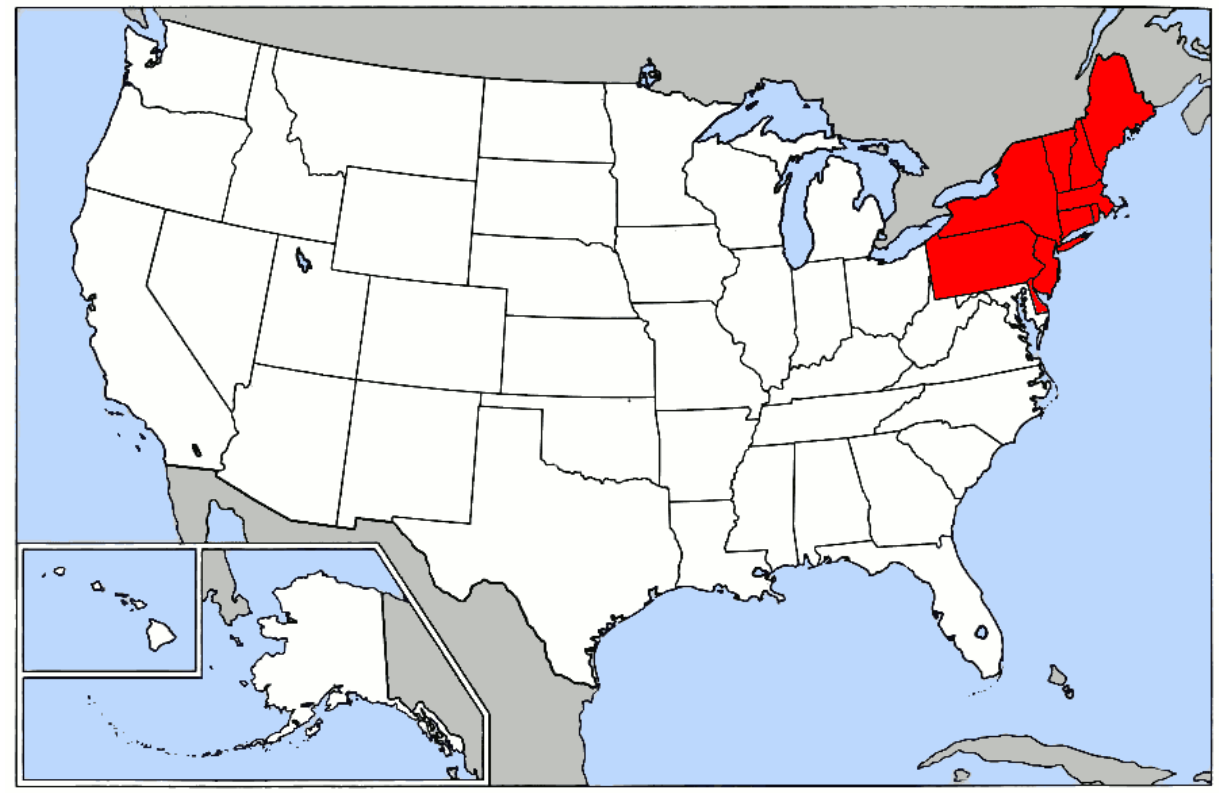
Could they recognize a person's origin based on their speech?



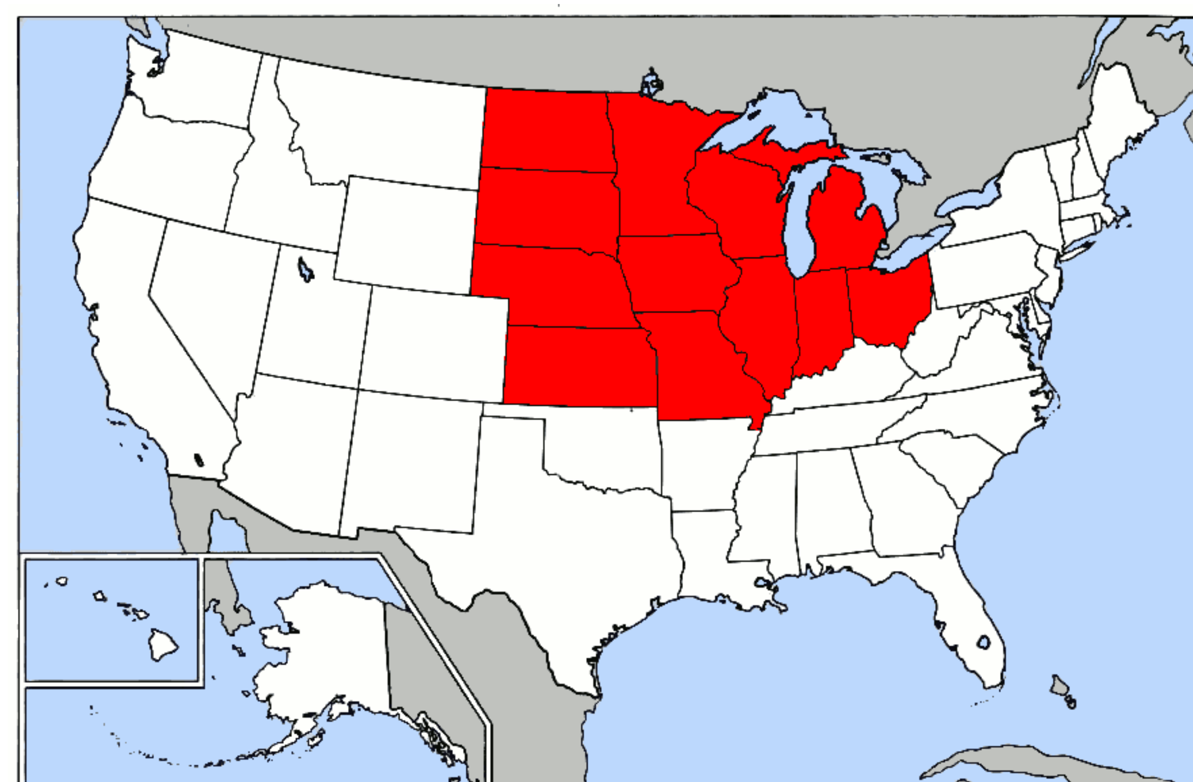
Map of USA Highlighting South

Learnability

What about the distinction between the Northeastern and the Mid-Western accents?



Map of USA Highlighting Northeast



Map of USA Highlighting Midwest

Learnability

Or the distinction between the people who originated from different parts of Brooklyn?

Brooklyn neighborhoods map ⓒ BY-SA 3.0 Peter Fitzgerald

The Practice of Teaching

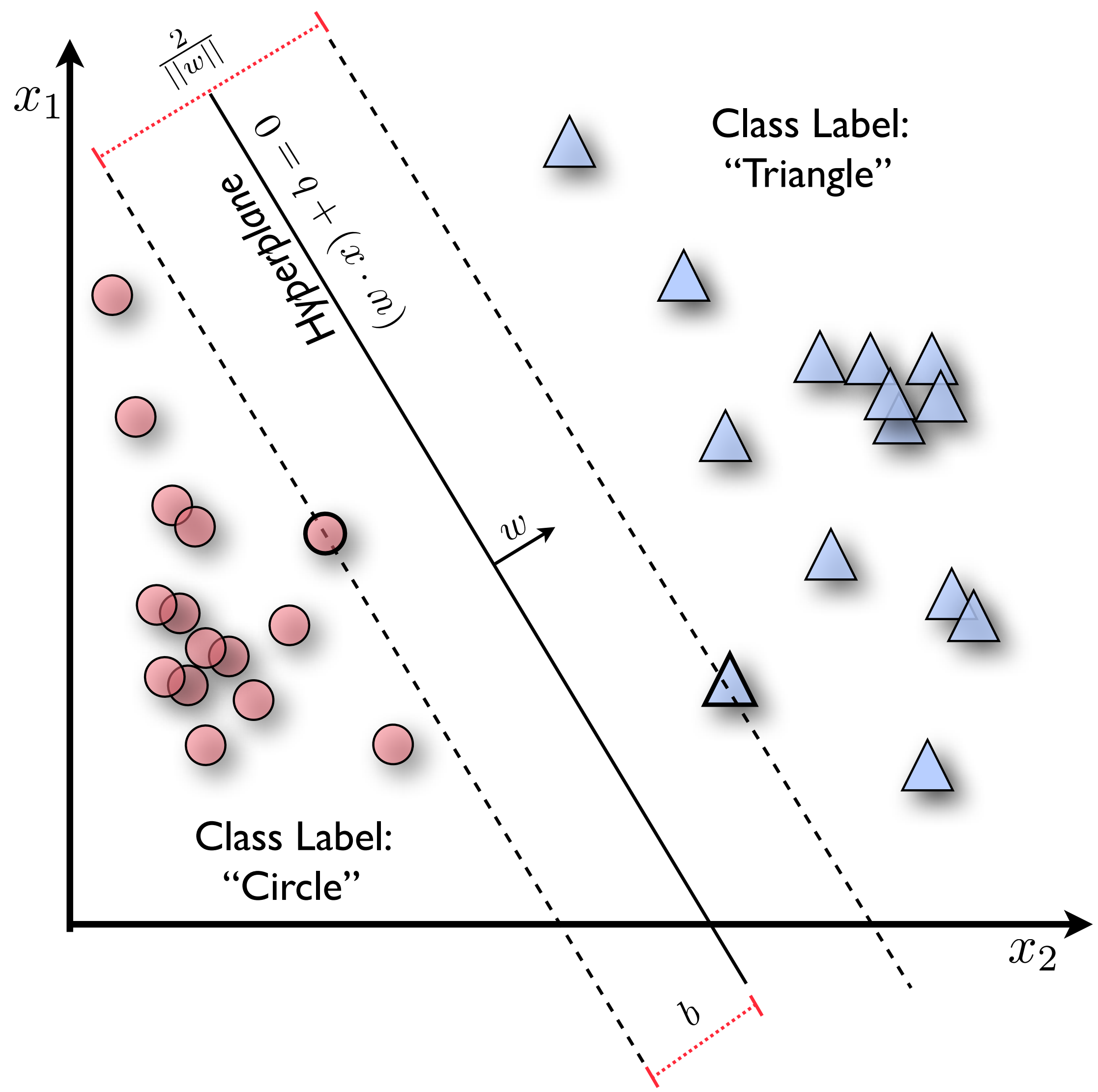
How would we teach a new arrival to identify accents?

1. Start with the easiest distinctions
2. Proceed with finer distinctions

We would never suggest that a novice learn all distinctions at the same time.

Supervised Learning

A “sink or swim” approach



No effort to tailor the learning to the human ability to learn from particular images.

Perceptual Annotation

Much information about human capacities can be of direct value for machine learning:

Some images are learnable, and some are not.

Learnability varies with experience.

Some things are easily learned, other things take more time.

Such detailed information reflecting human capacity is what we call a perceptual annotation.

Prior Work

Active Learning: B. Settles. Active Learning. Morgan & Claypool, 2012.

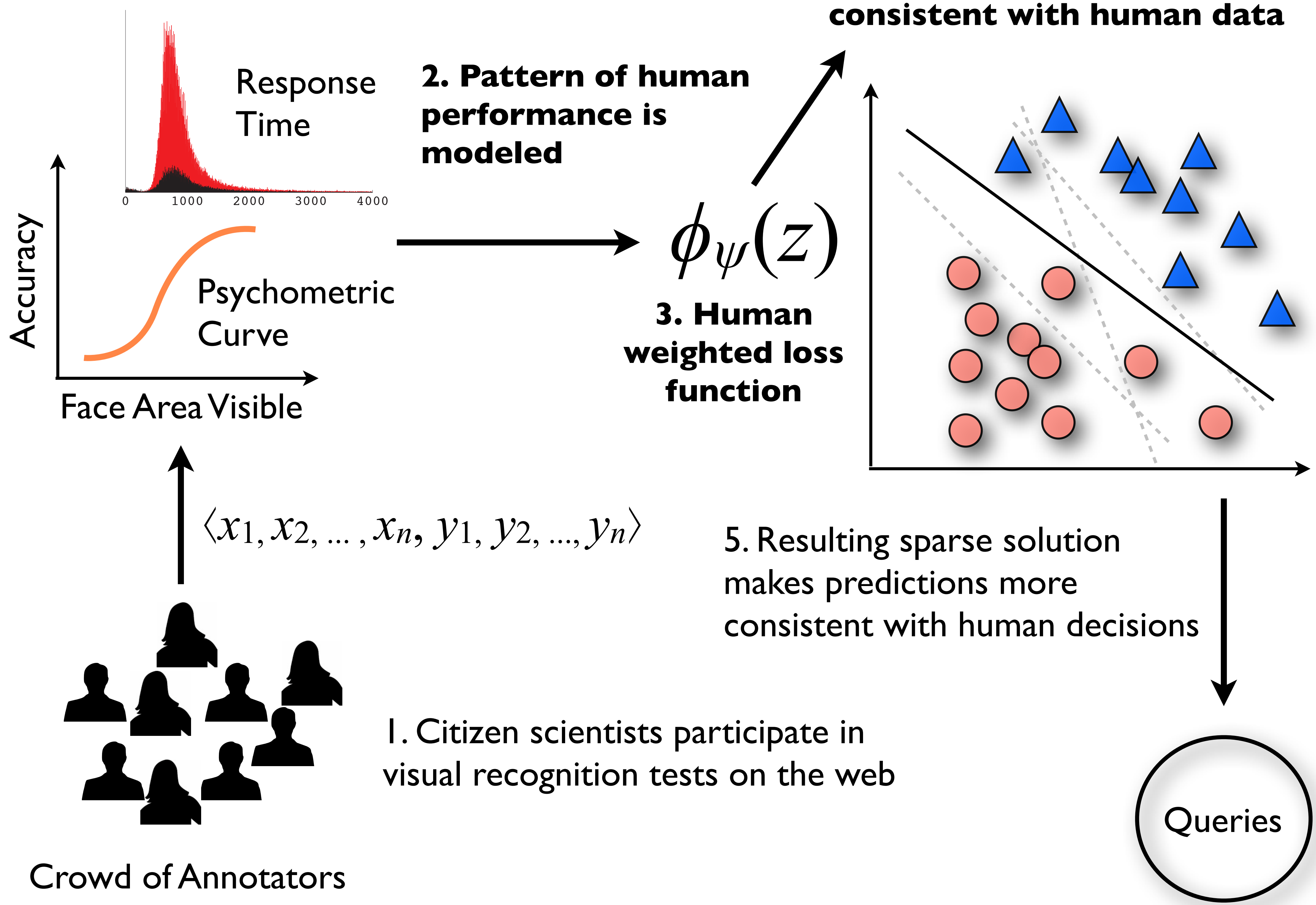
Structured Domain Knowledge: G. Kunapuli, R. Maclin, and J. Shavlik, “Advice refinement for knowledge-based support vector machines,” NIPS, 2011.

Human Annotation Process Modeling: P. Welinder, S. Branson, S. Belongie, and P. Perona, “The Multidimensional Wisdom of Crowds,” NIPS, 2010.

Fine-Grained Crowdsourcing: J. Deng, J. Krause, and L. Fei-Fei, “Fine-Grained Crowdsourcing for Fine-Grained Recognition,” IEEE CVPR 2013.

Semantic Retention: C. Xu, R.F. Doell, S.J. Hanson, C. Hanson, and J.J. Corso, “A Study of Actor and Action Semantic Retention in Video Supervoxel Segmentation,” Int. J. Semantic Computing, vol. 7, no. 4, Dec. 2013.

Perceptual Annotation



W. Scheirer, S. Anthony, K. Nakayama, and D.D. Cox, "Perceptual Annotation: Measuring Human Vision to Improve Computer Vision," IEEE T-PAMI, vol. 36, no. 8, August 2014.

Visual Psychophysics Using TestMyBrain.org

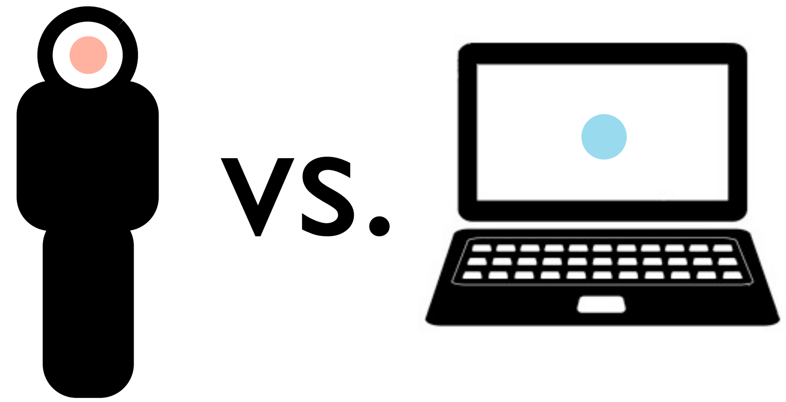
Visual Psychophysics

Probe psychological and perceptual thresholds through controlled manipulation of stimuli.

Careful management of stimulus construction, ordering and presentation allows for precise determination of perceptual thresholds.

Canonical Early Example*: minimum threshold for stimulation of an individual photoreceptor.

S. Hecht, S. Shlaer, M. Pirenne, "Energy, quanta and vision," J. Gen. Physiol., 25 (1942), pp. 819–840.



Face Detection: Identical face stimuli shown to humans and computer algorithms.

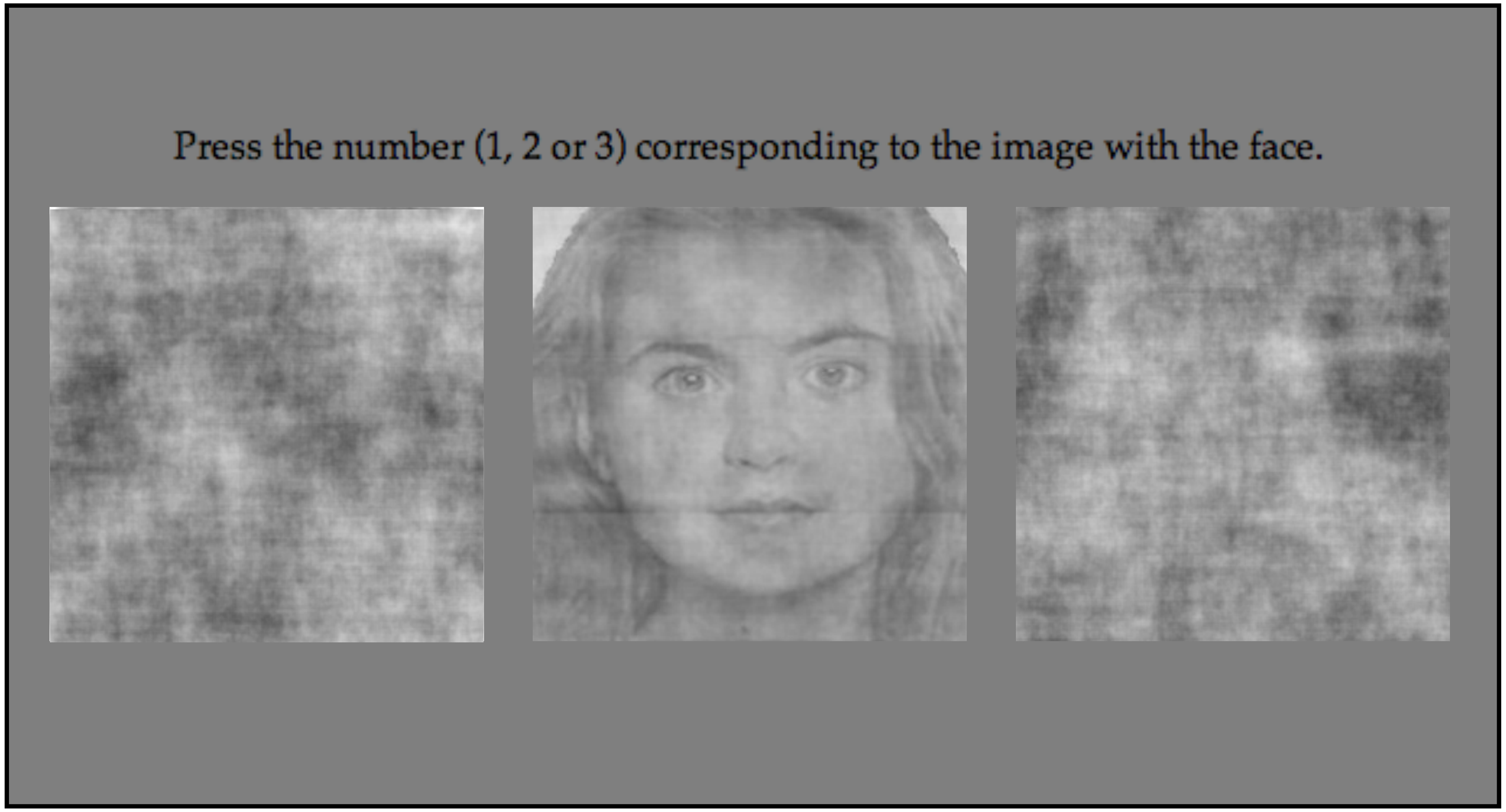
A selection of common algorithms, including state-of-the-art commercial algorithms from Google and face.com (now part of Facebook).

Large-scale web samples captured on the TestMyBrain platform.

L. Germine, K. Nakayama, B. Duchaine, C. Chabris, G. Chatterjee, and J. Wilmer, "Is the web as good as the lab? Comparable performance from web and lab in cognitive/perceptual experiments," Psychonomic Bulletin & Review, vol. 19, pp. 847–857, 2012.

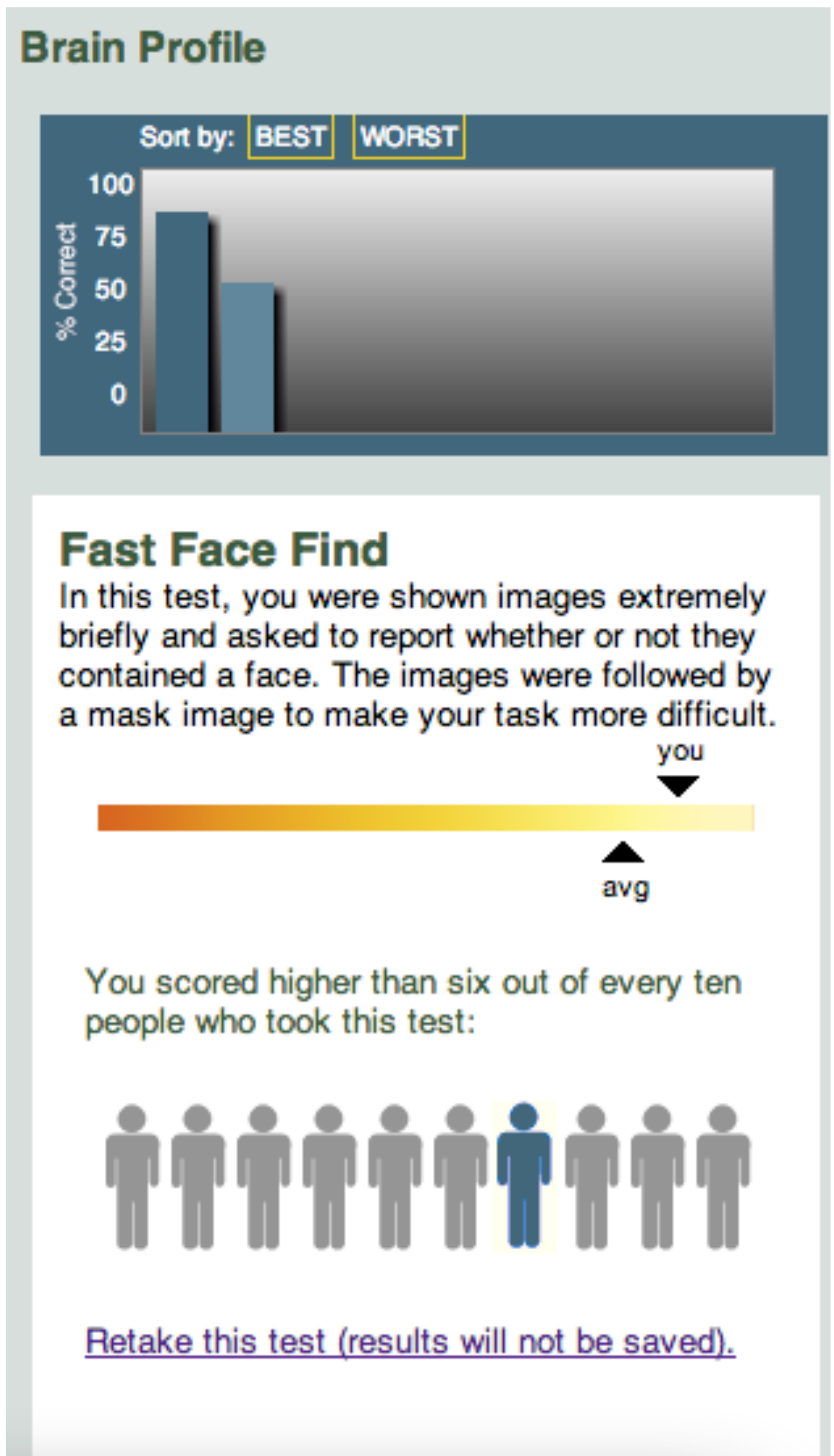
Behavioral Task

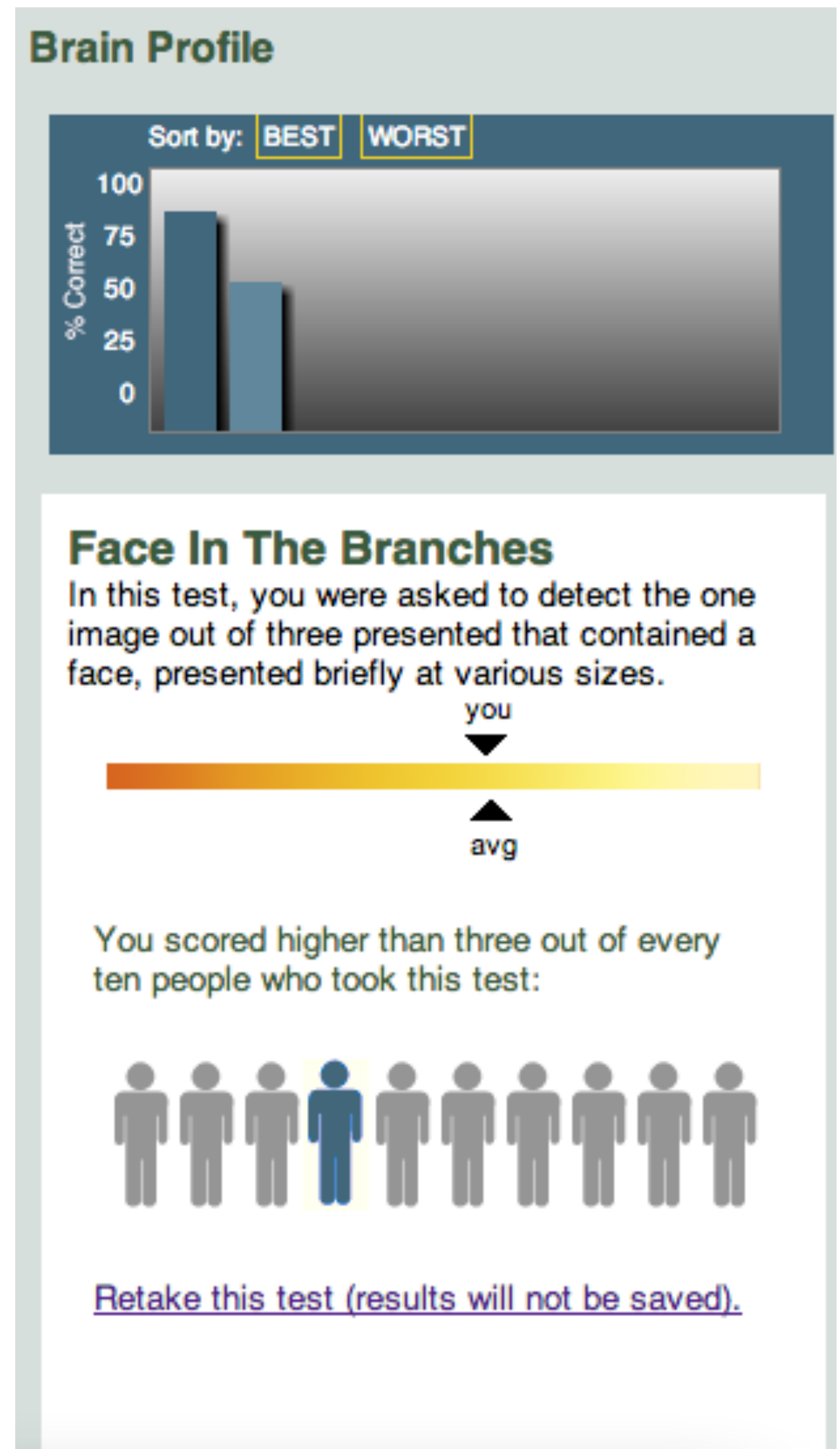
3 Alternative Forced Choice

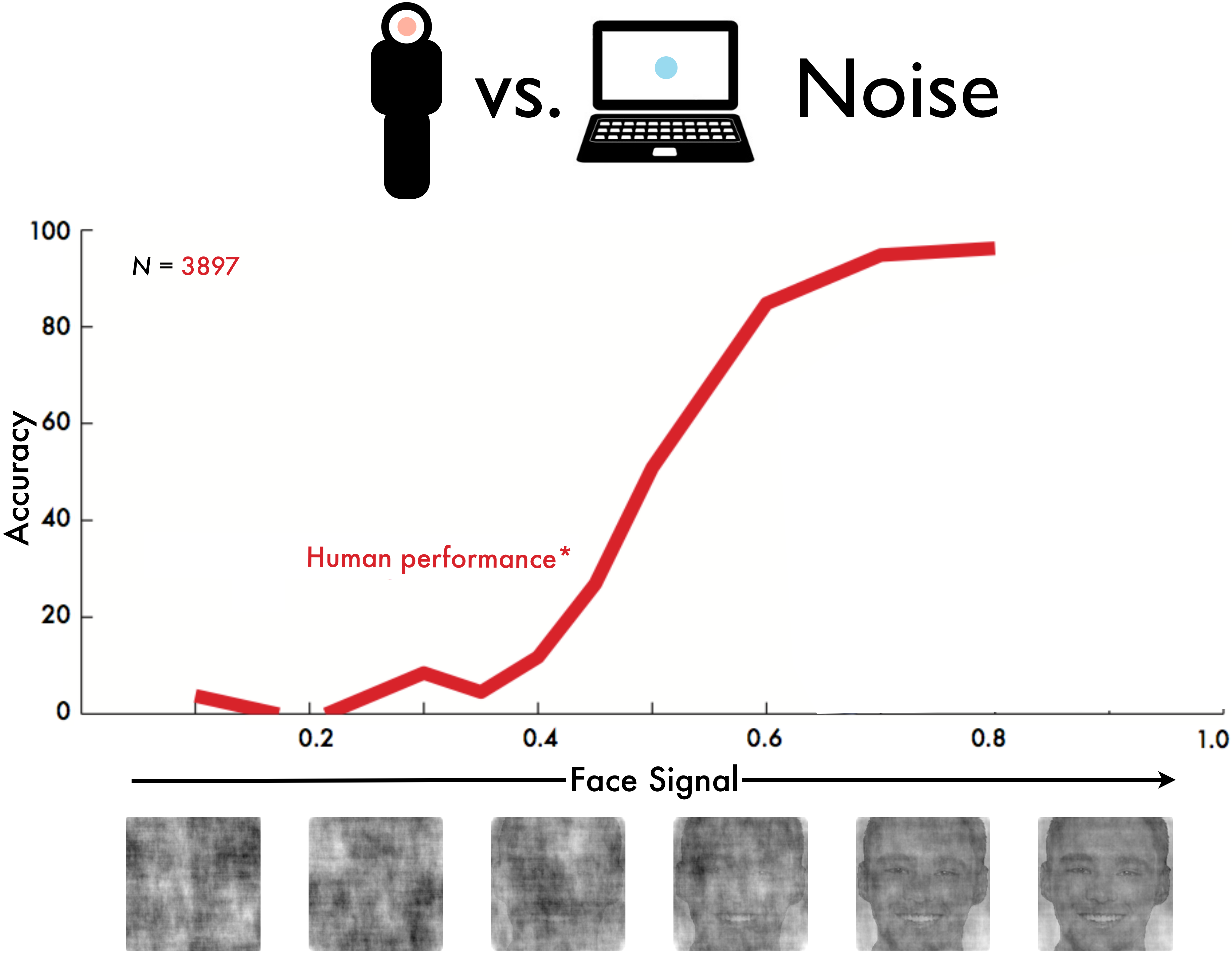


TESTMYBRAIN

Behavioral Task

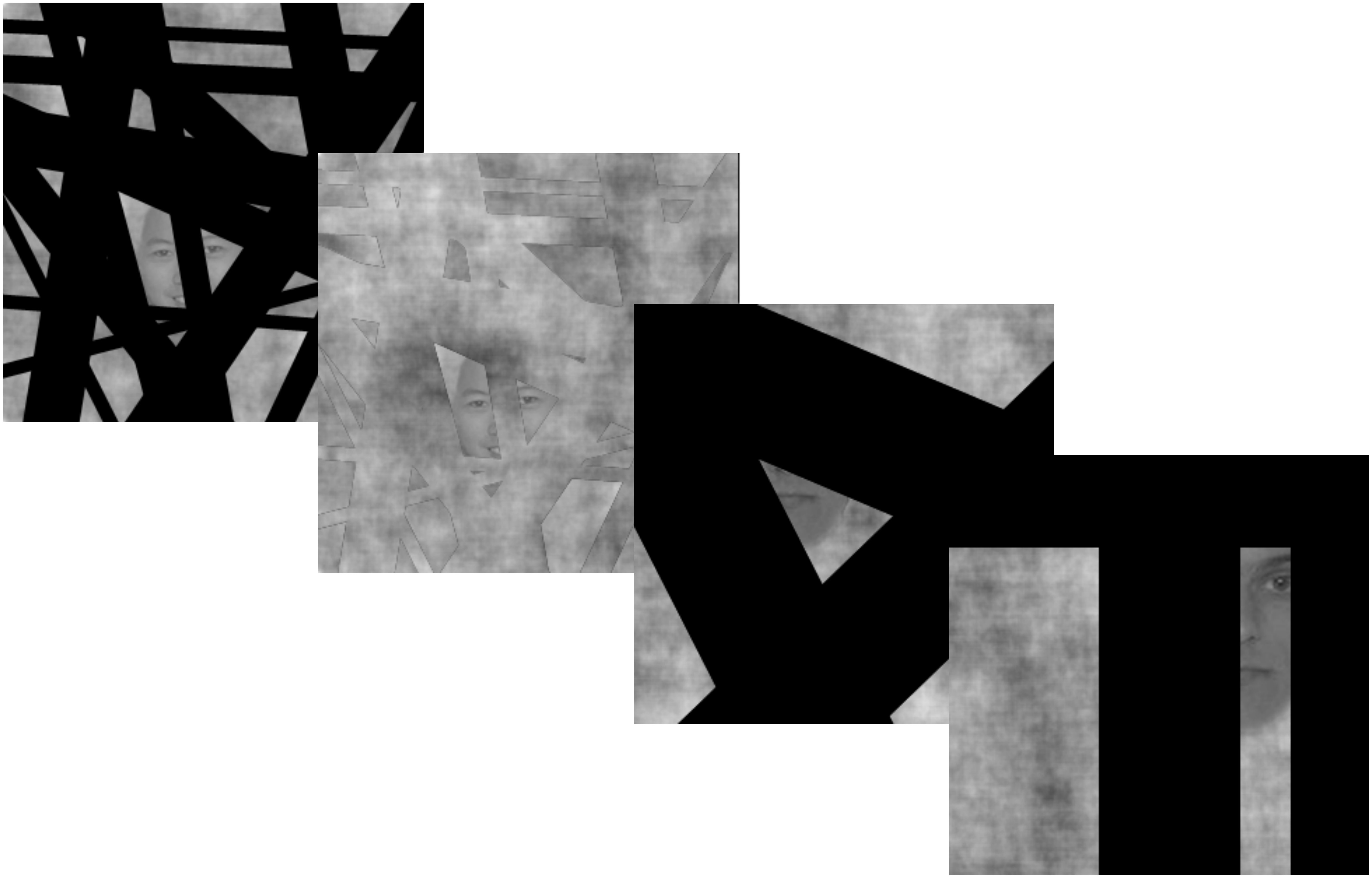


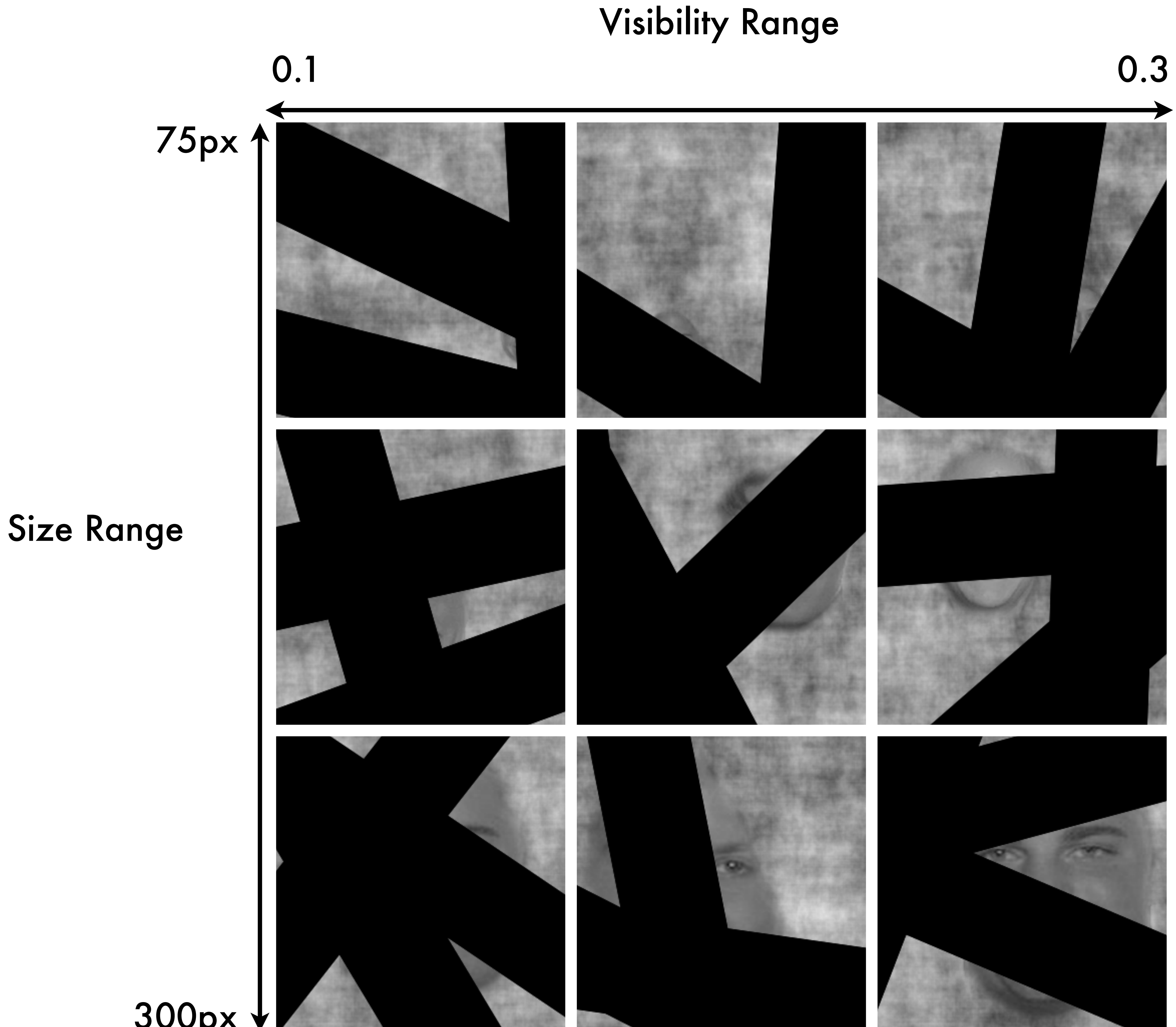


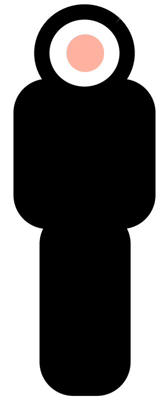


* normalized so chance is zero

Occlusion

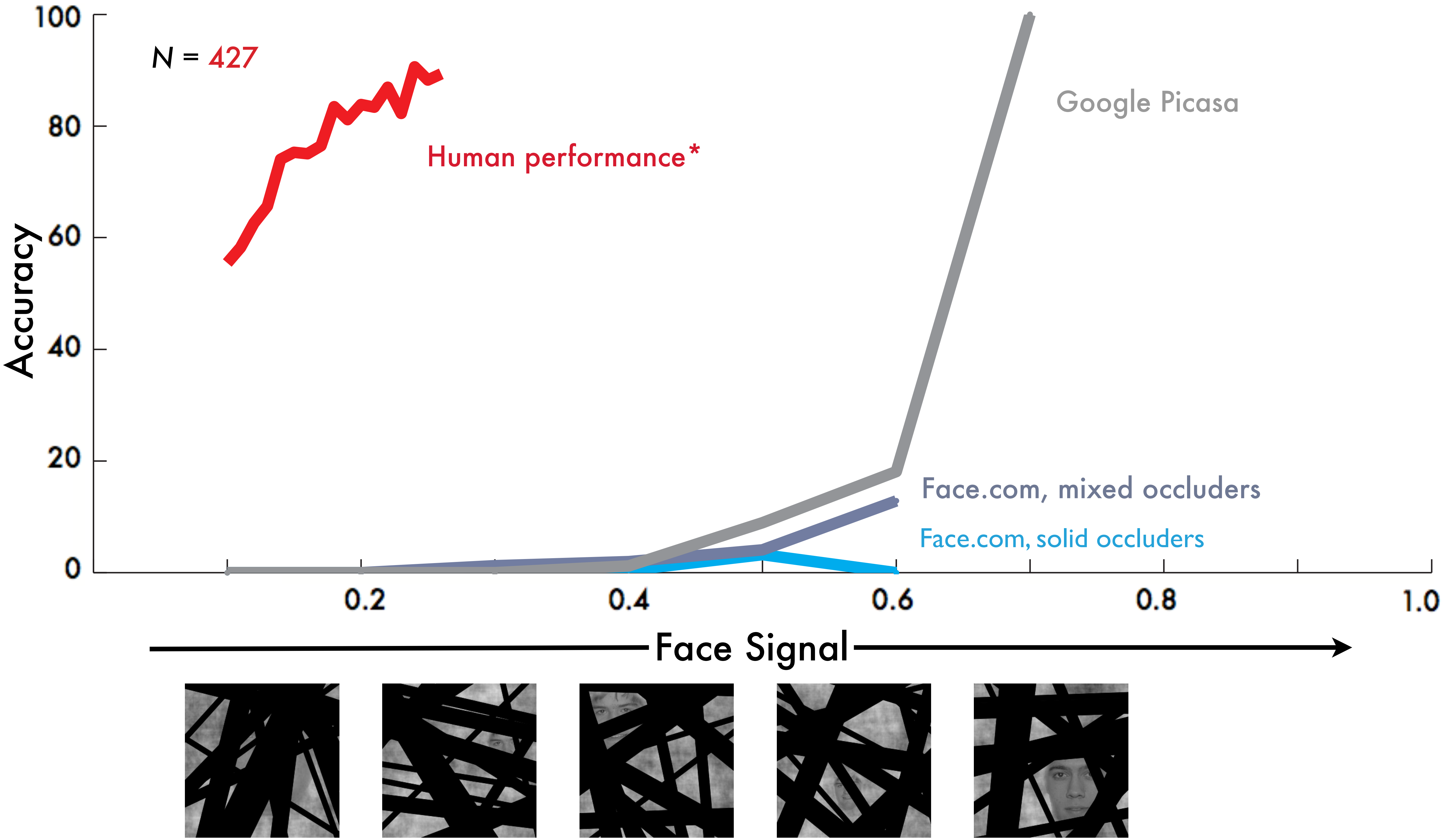



Visibility Range
0.1
0.3
75px
Size Range
300px



vs.

Occlusion

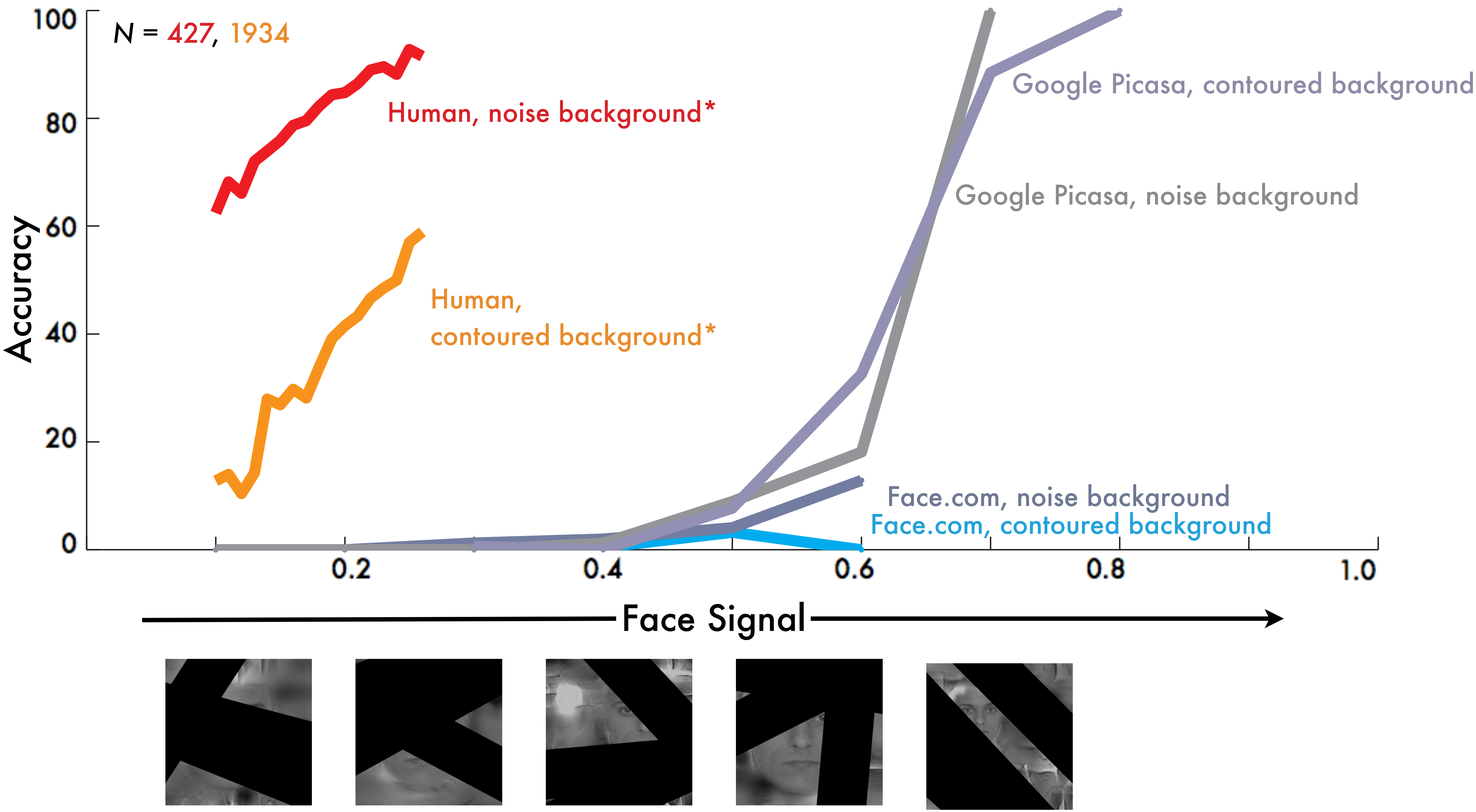


* normalized

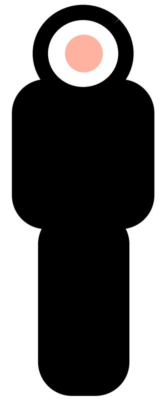
Black occluders with Portilla-Simoncelli Backgrounds

J. Portilla and E. Simoncelli, "A parametric texture model based on joint statistics of complex wavelet coefficients," IJCV, vol. 40, no. 1, 2000.

vs. Occlusion



* normalized

 vs.  Summary

Humans beat even the best algorithms.

Algorithms have enormous problems with degradations like occlusion that people find trivial.

Contoured image backgrounds reduce human performance; people are still much better.

Perceptual Annotations

What information are we recording from a psychophysics experiment for machine learning training?

1. Per Image Avg. Accuracy

2. Per Image Avg. Reaction Time

Perceptual Annotation for SVM

Classification Risk

$$\underset{f}{\operatorname{argmin}}\left\{R_{\mathcal{I}}(f) := \int_{\mathbb{R}^d \times \mathbb{N}} L(x, y, f(x)) P(x, y)\right\}$$

Ideal Risk   Loss Function   Joint Distribution

A. Smola, "Learning with Kernels," Ph.D. dissertation, Technische Universitat Berlin, Berlin, Germany, 1998.

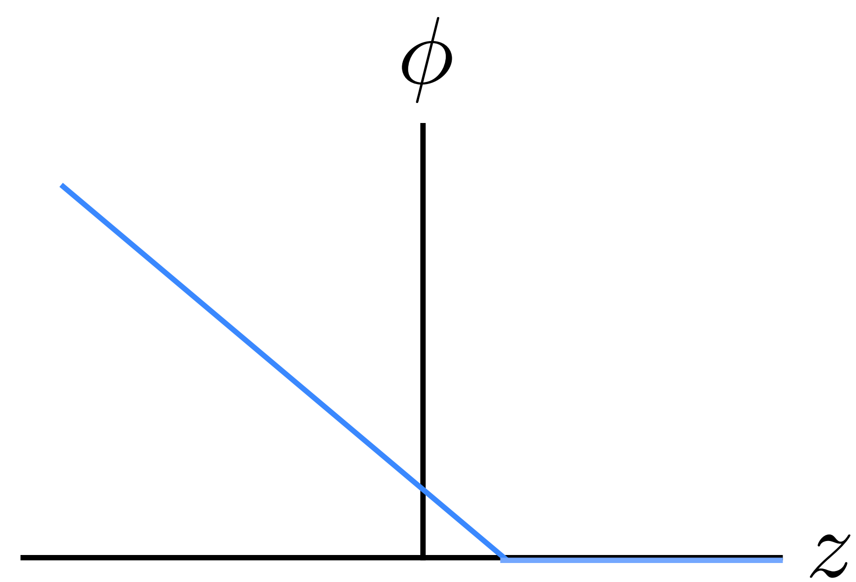
Loss Functions

A prediction during training is calculated as the output of a classifier multiplied by its label:

$$z = yf(x)$$

Typical Loss Function: Hinge Loss

$$\phi(z) = max(0, 1 - z)$$



Non-linear nature of psychometric curves for visual recognition tasks suggests a much different model.

Human Weighted Loss

Besides data x and labels y, assume we also have a cost c for each training sample:

$$\phi_\psi(x, z) = \max(0, (1 - z) + \mathrm{M}(x, z))$$

where

$$\mathrm{M}(x, z) = \begin{cases} c_x & \text{if } z < 1 \\ 0, & \text{otherwise} \end{cases}$$

Human Weighted Loss

c can take on one of two types of values:

A static penalty (*e.g.* 0 if a sample doesn't have a perceptual annotation)

A point on the psychometric curve (*e.g.* accuracy or reaction time)

*All training samples do not require an associated perceptual annotation.

Optimization Problem

For the linear binary case, solve the following optimization problem:

$$min\frac{1}{2}||w||^2 + C\sum_{l=1}^{L}\phi_{\psi}(x_l, y_l f(x_l))$$

Perceptual Annotations

Who is afraid of non-convex loss functions?

Human weighted loss is non-convex.

(simple reason: the same x can take on multiple c values)

Bengio and LeCun: biological systems contain many layers of adaptive non-linear components.

Perceptual annotations are the measurable output of such machinery.

No expectation for convex formulation.

Y. Bengio and Y. LeCun, Large Scale Kernel Machines. MIT Press, 2007, ch. Scaling Learning Algorithms Towards AI.

Case Study: Face Detection

TestMyBrain Data Collections

Collection #1: "Fast Face Finder"

7.5 weeks

3,250 research subjects

337,932 annotations for 4,255 AFLW images

Collection #2: "Faces in the Branches"

2 weeks

410 research subjects

41,650 annotations for 2,448 Portilla-Simoncelli textures

FDDB: Face Detection Dataset and Benchmark

- 2,845 images with a total of 5,171 faces
- A wide range of challenges including occlusions, difficult poses, and low resolution and out-of-focus faces
- The specification of face regions as elliptical regions
- Both grayscale and color images
- 10-fold cross-validation style testing

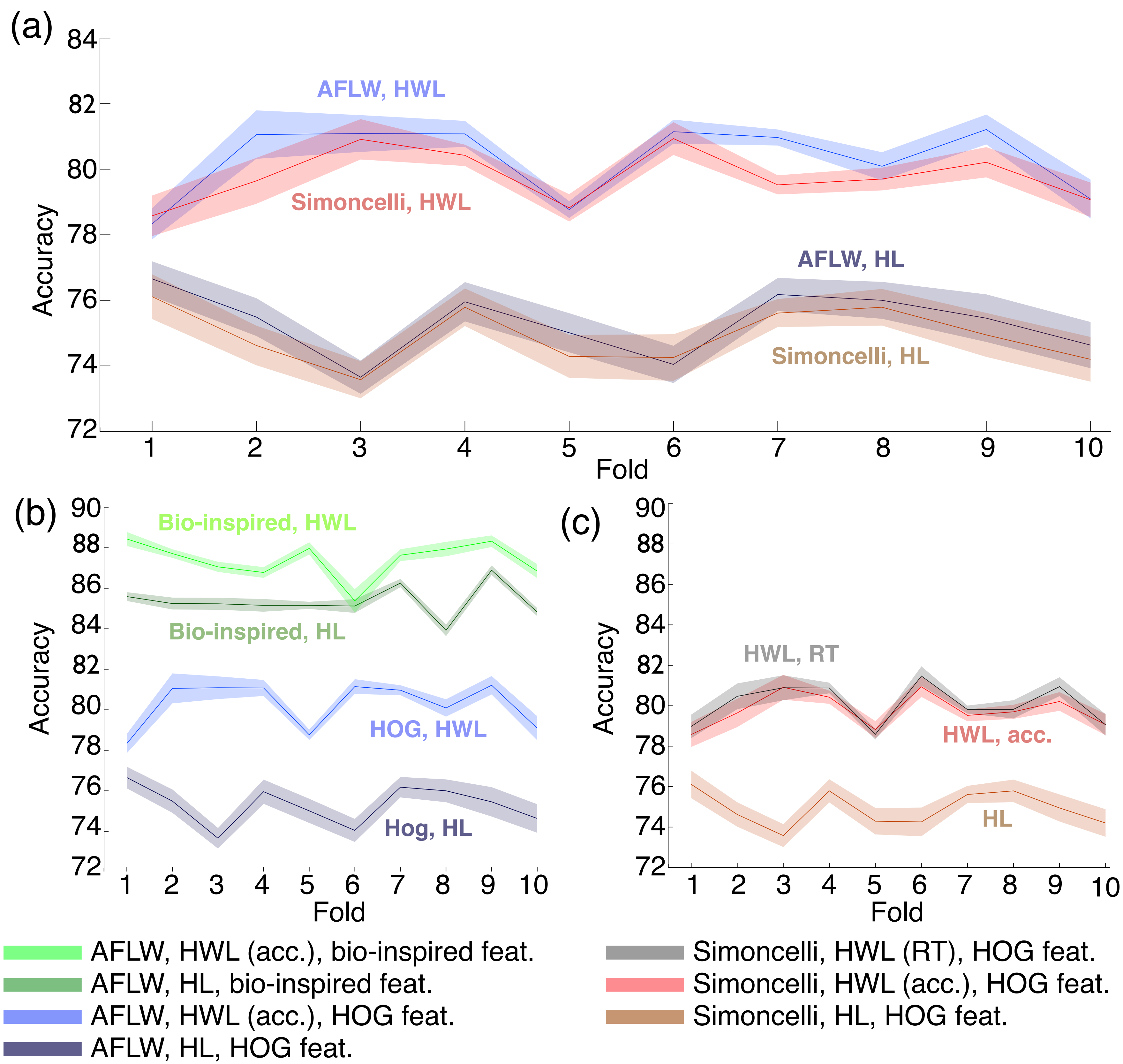
V. Jain and E. Learned-Miller, “FDDB: A Benchmark for Face Detection in Unconstrained Settings,” UMass Amherst Tech Report UM-CS-2010-009, 2010

Experiment #1: Is there an effect?

10-fold cross-validation classification task

- HOG and Bio-Inspired CNN Features
- 500 +/- patches from each fold
- Features sampled from 30x30 chips
- 100 +/- training images from each fold
- 50 +/- perceptually annotated training images
- 9 tests per fold

Effect: HL Replaced by HWL



Experiment #2: FDDB Benchmark

A simple detector

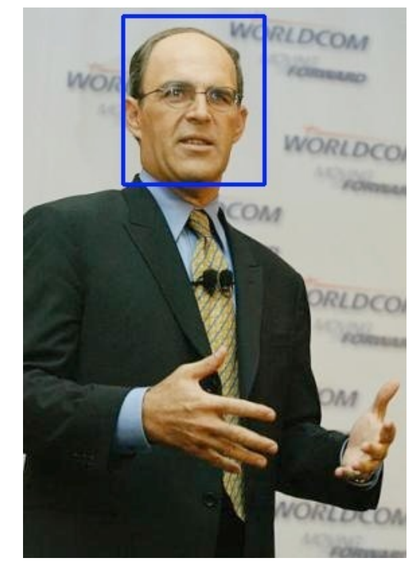
1. Viola-Jones at numerous scales

SVM w/ HWL

2. Filter face candidates

Scoring Strategy

3. Compute scores for ROC calculation. Select best rectangle from a neighborhood.



4. Final result

Bio-inspired features:
900 +/- images from fold
300 +/- annotated AFLW images
30x30 patches

HOG features:
1800 +/- images from fold
200 +/- annotated Portilla-Simoncelli images
40x40 patches

FDDB Protocols

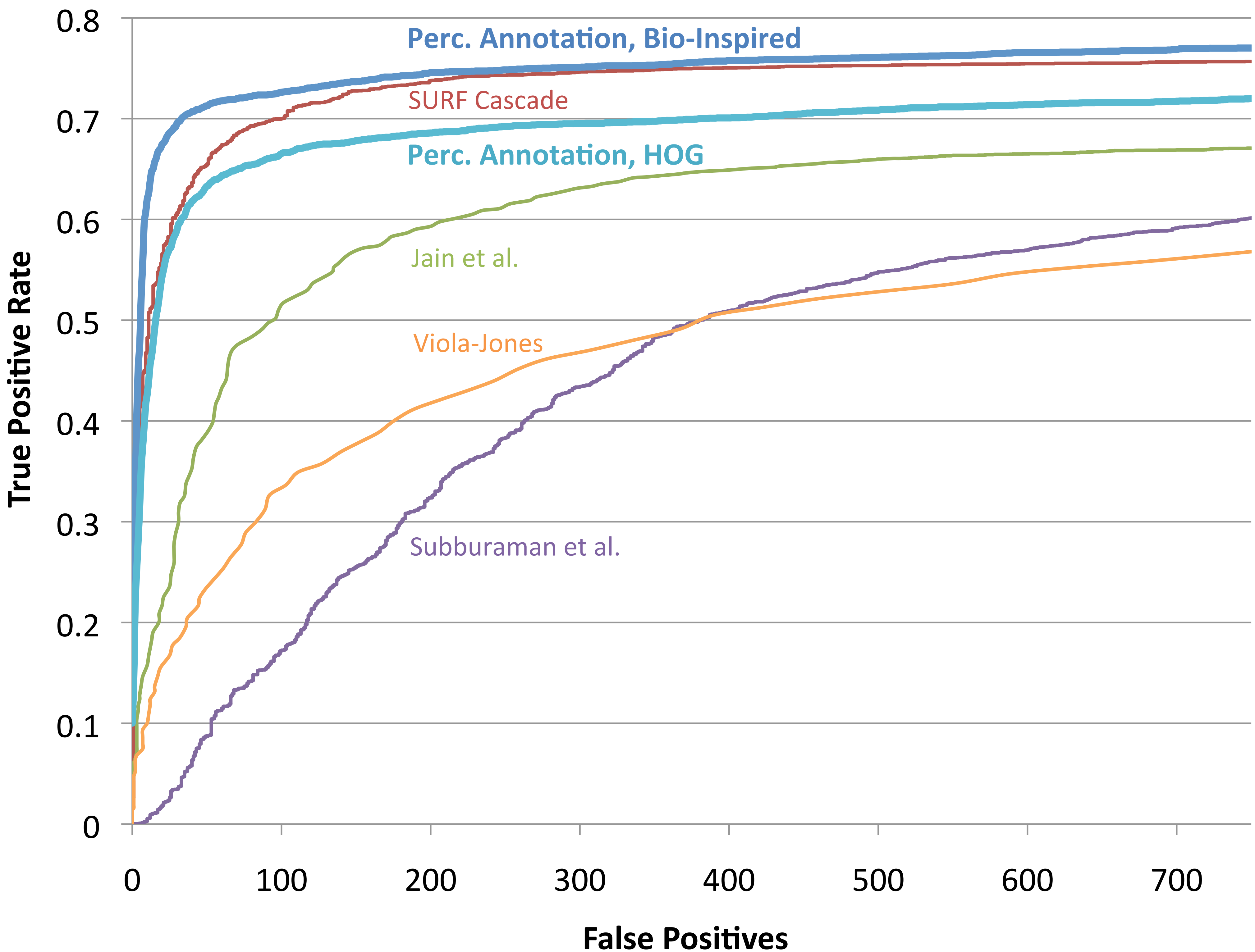
Ground-Truth

True Positive

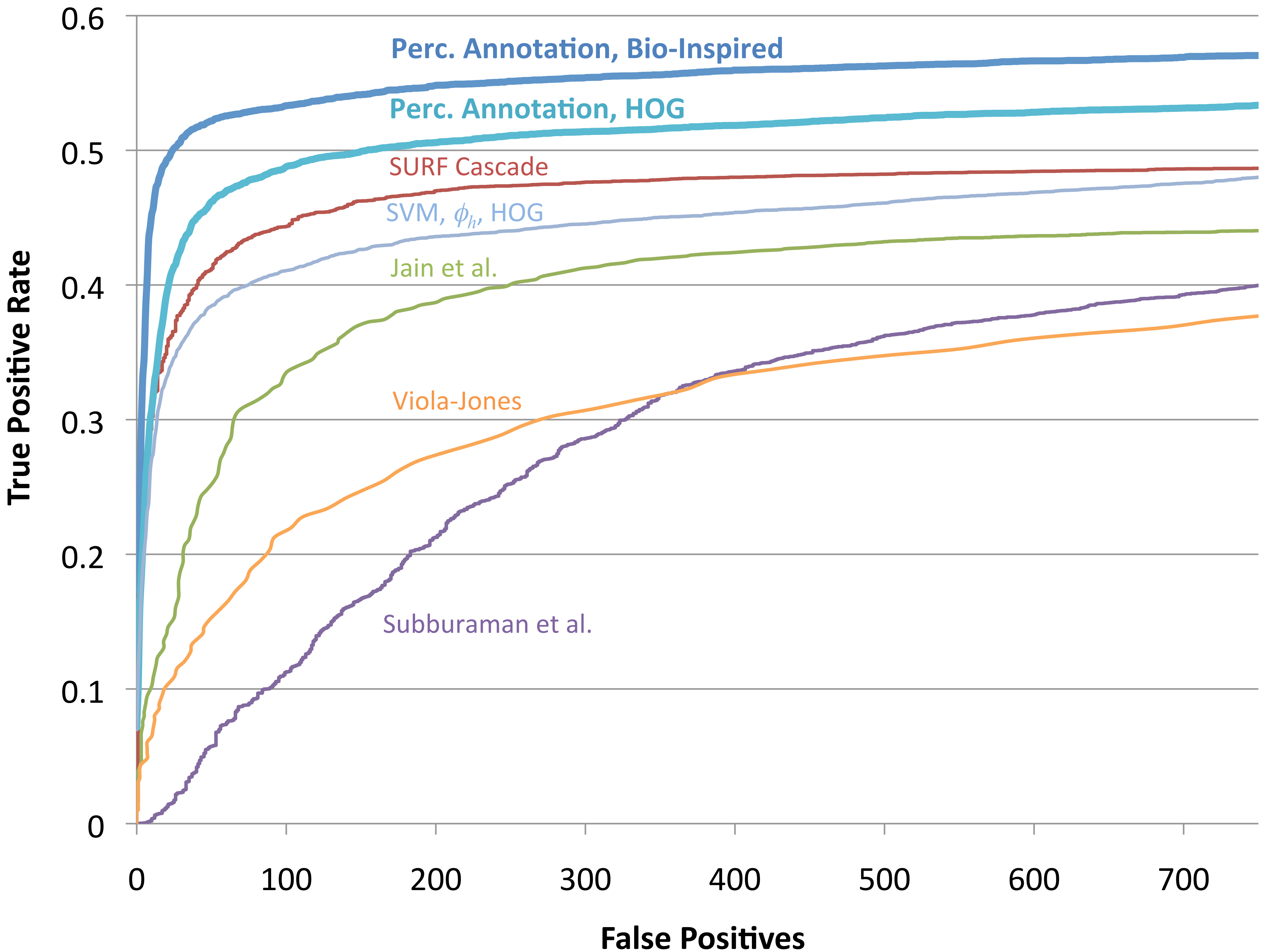
False Positive

Discrete Metric: presence of detection
Continuous Metric: quality of detection

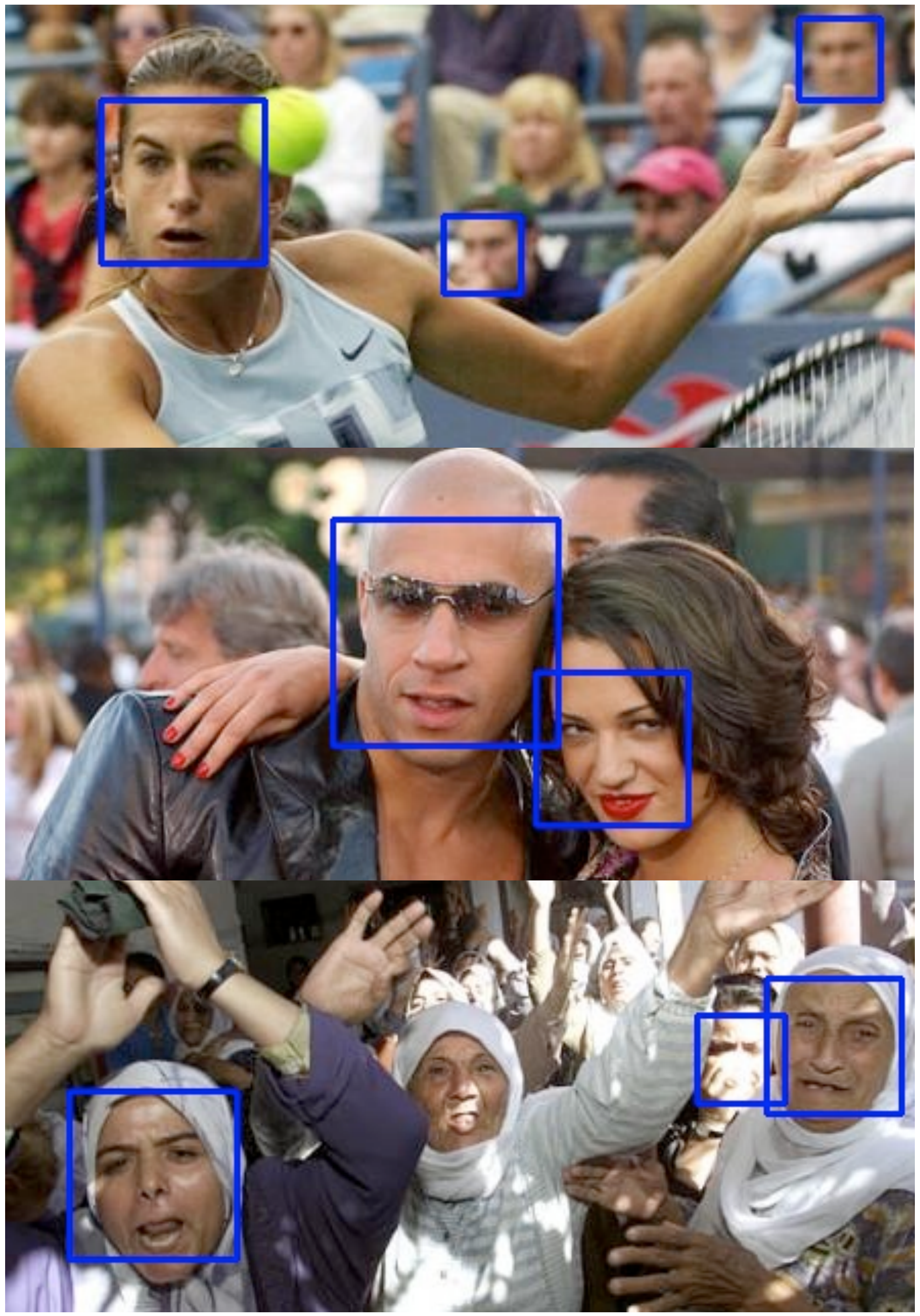
FDDB Discrete Score Metric



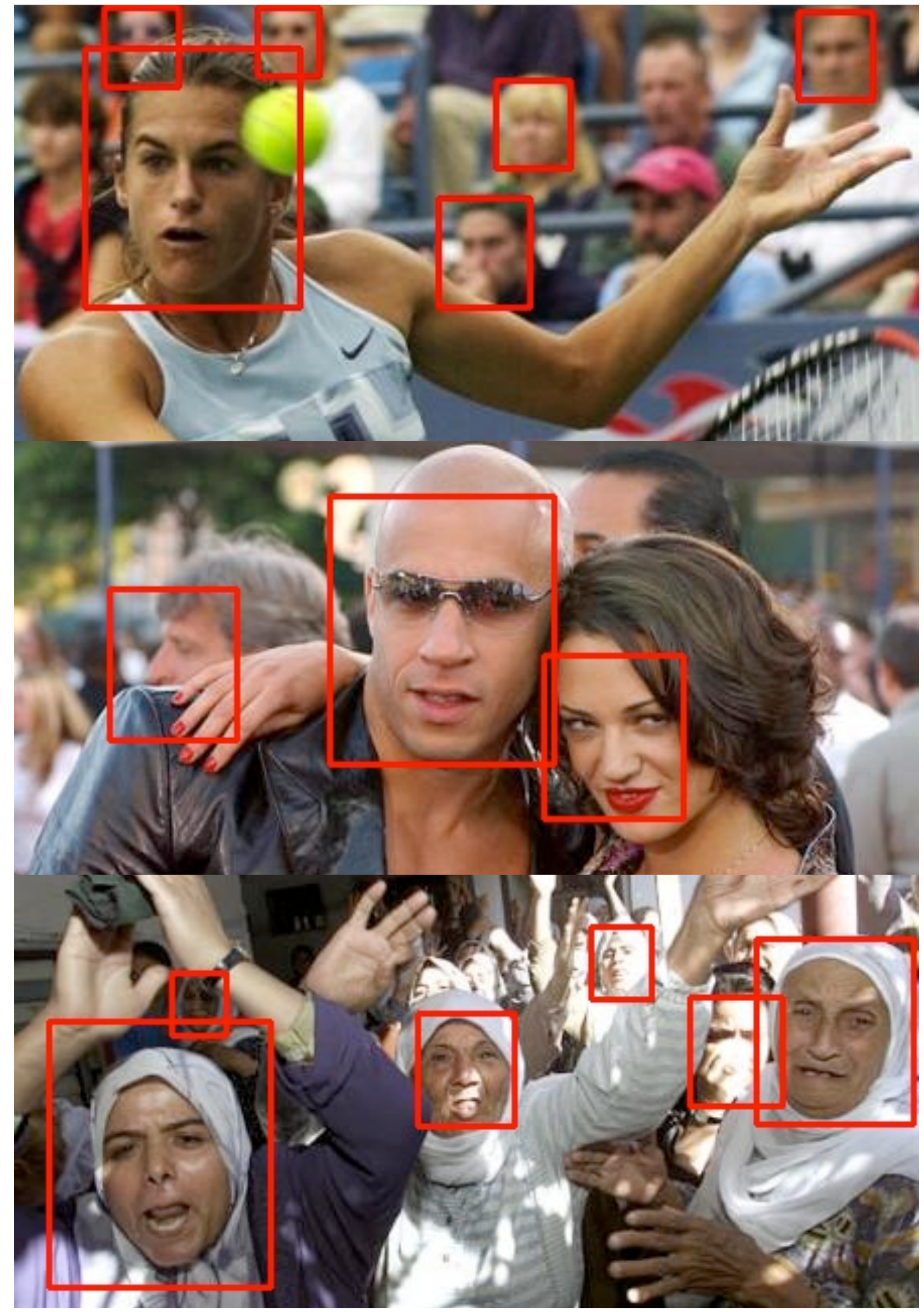
FDDB Continuous Score Metric



Example Detections



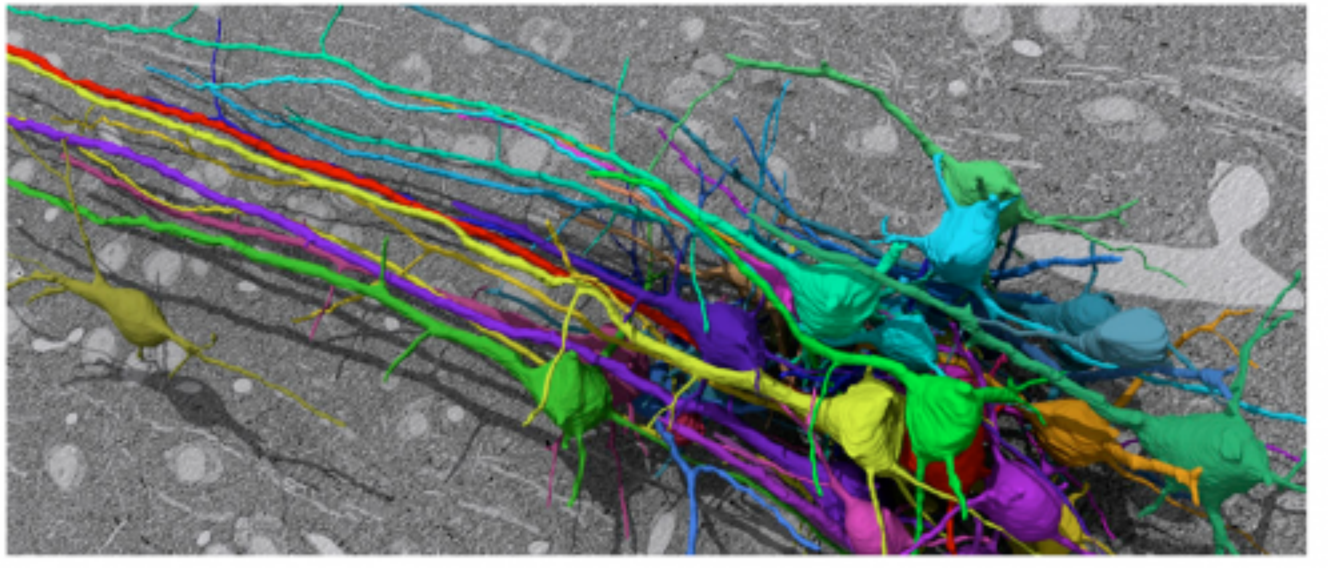
Viola-Jones



Perceptual Annotation

New Directions:
Health and Education

Connectomics



EM reconstruction of mouse brain cortex.
Lichtman Lab @ Harvard

Map all of the connections in a brain, neuron by neuron, synapse by synapse

Connectomics

Understand the elements of neural computation

- Vision, Motor Control, Language, Learning

Understand abstract aspects of the mind

- Memory, Intelligence, Personality, Identity

New therapies for mental illnesses that present without an obvious pathology

- Autism and Schizophrenia

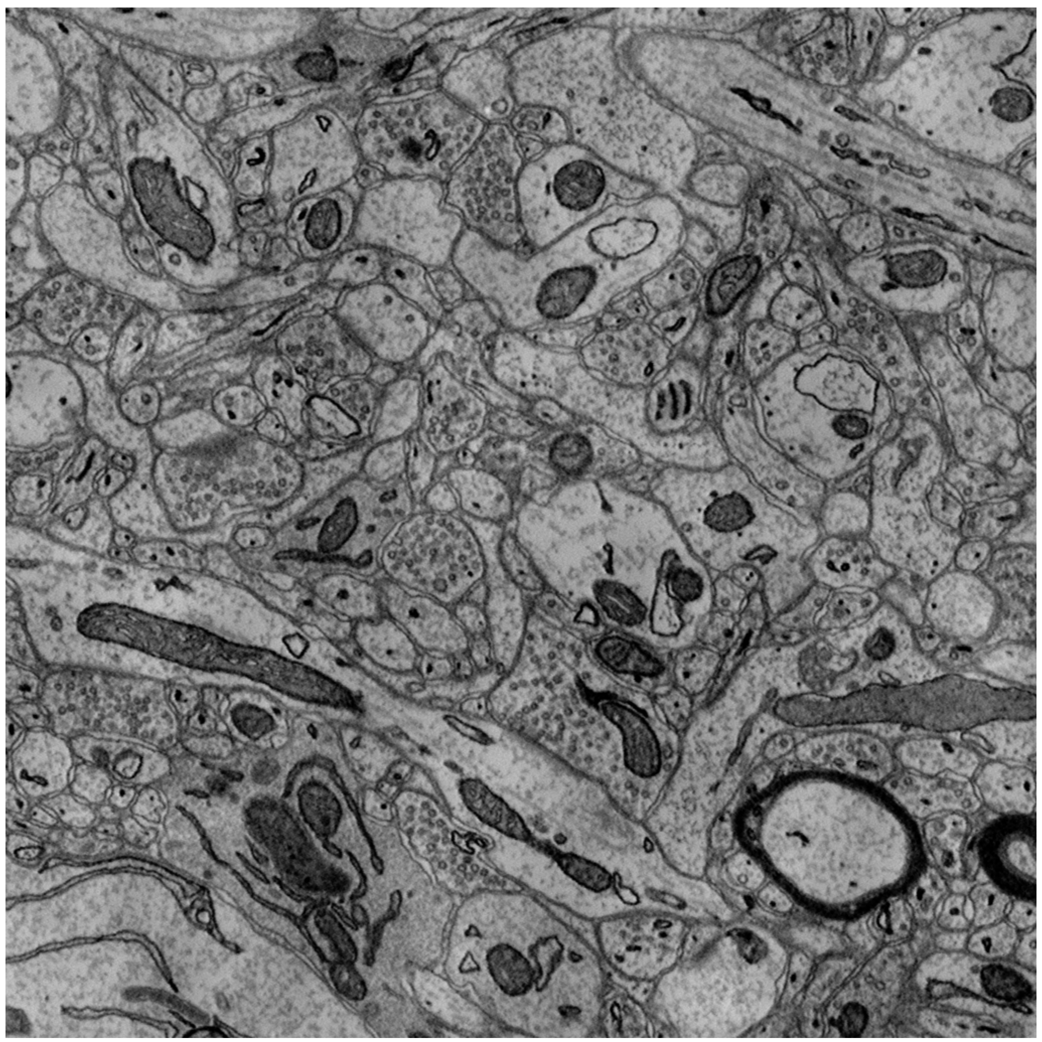
Rat Connectome

Tens of millions of neurons and billions of connections between them

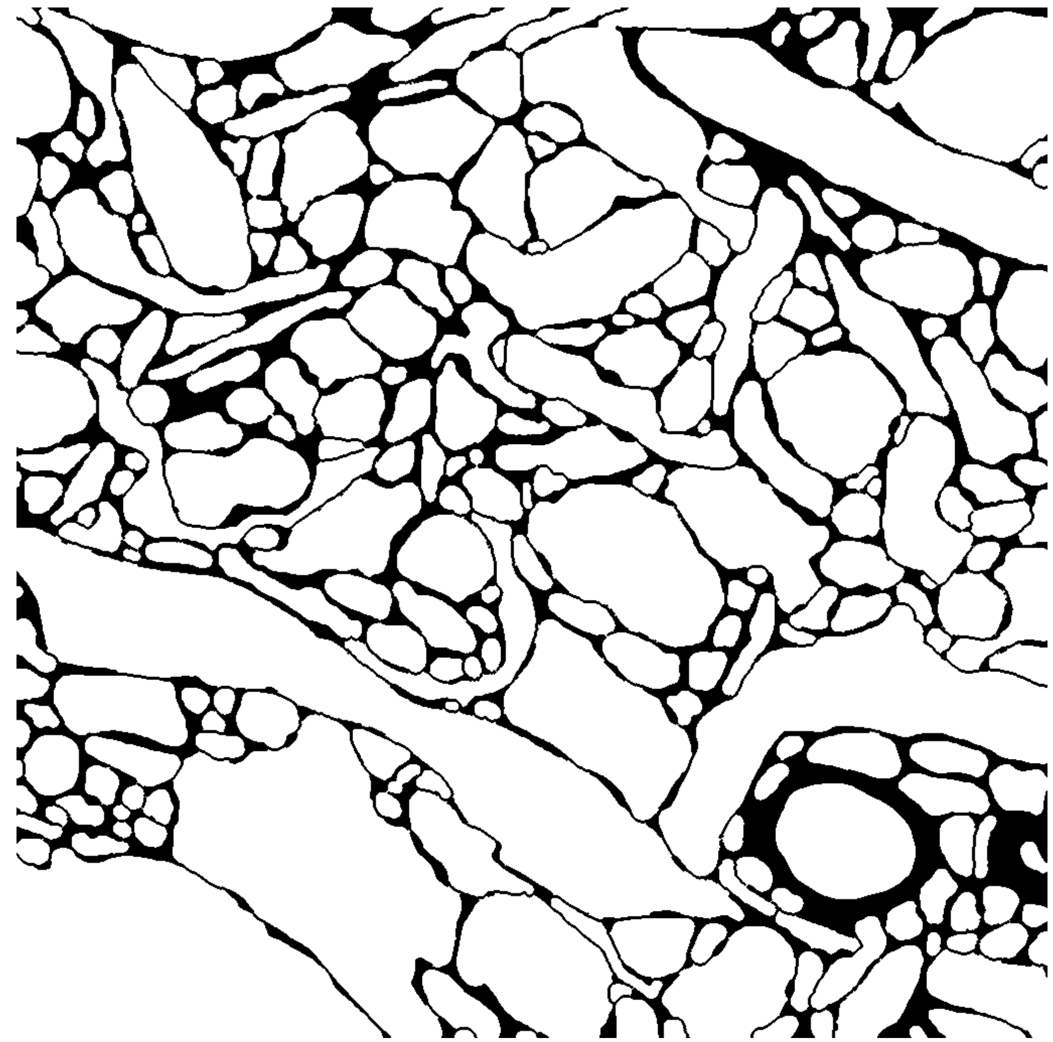
Petabytes of data

Cannot do this by hand: we need computer vision

2D Segmentation

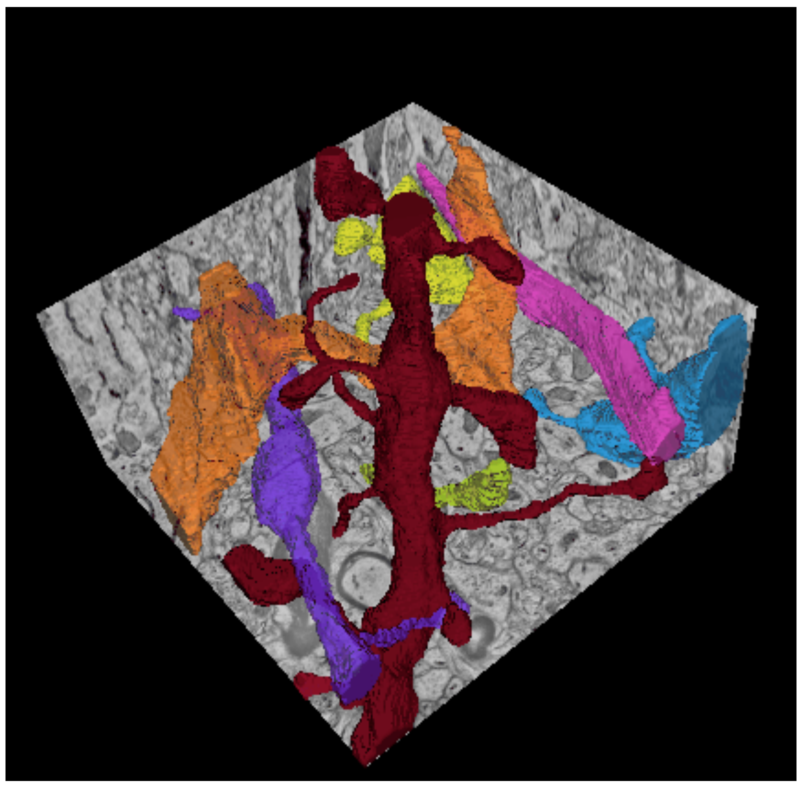


Slice of mouse cortex.
Lichtman Lab @ Harvard

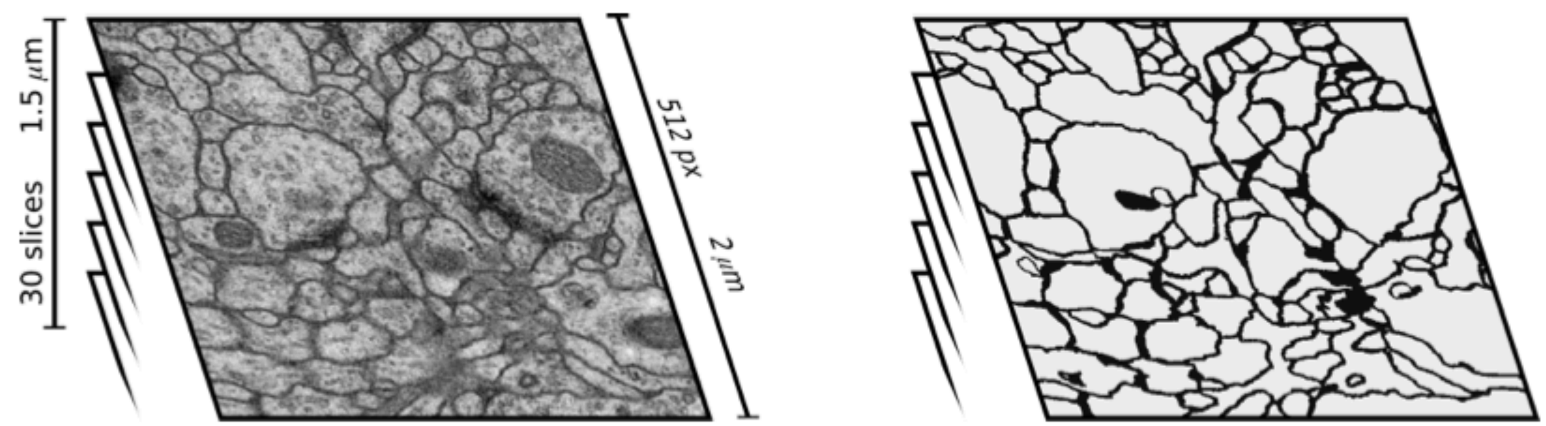


Goal: Automatically segment all neural structures

3D Reconstruction



Stack of mouse cortex.
Lichtman Lab @ Harvard



The x- and y-directions have a high resolution, whereas the z-direction has a low resolution

Goal: Connect corresponding segments across the volume

Progress has been slow…

From a computer vision perspective, why is this problem hard?

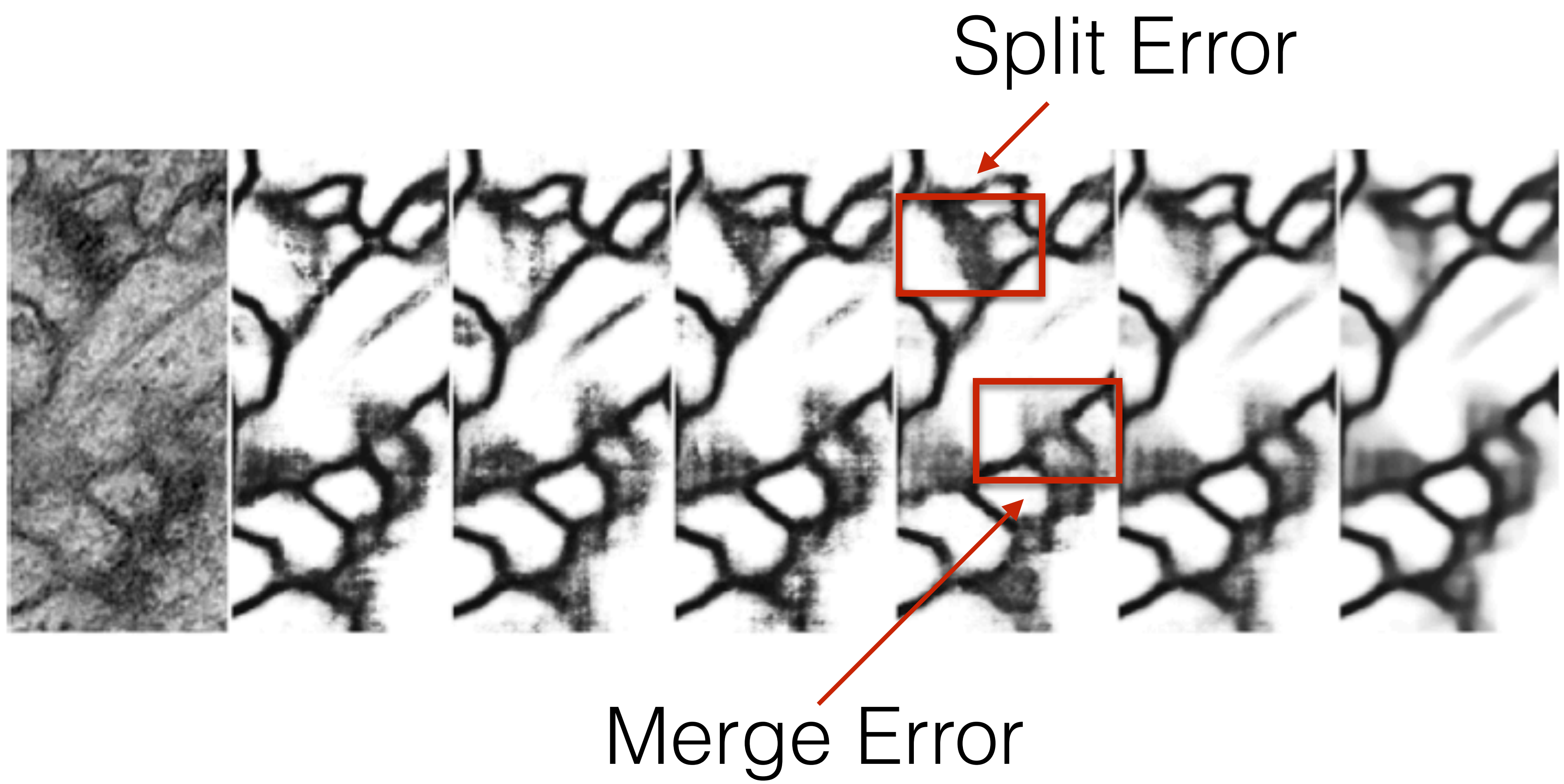
Trouble exploiting context

Lack of useful progress in recognition

Cycles required to handle sufficiently large amounts of training data

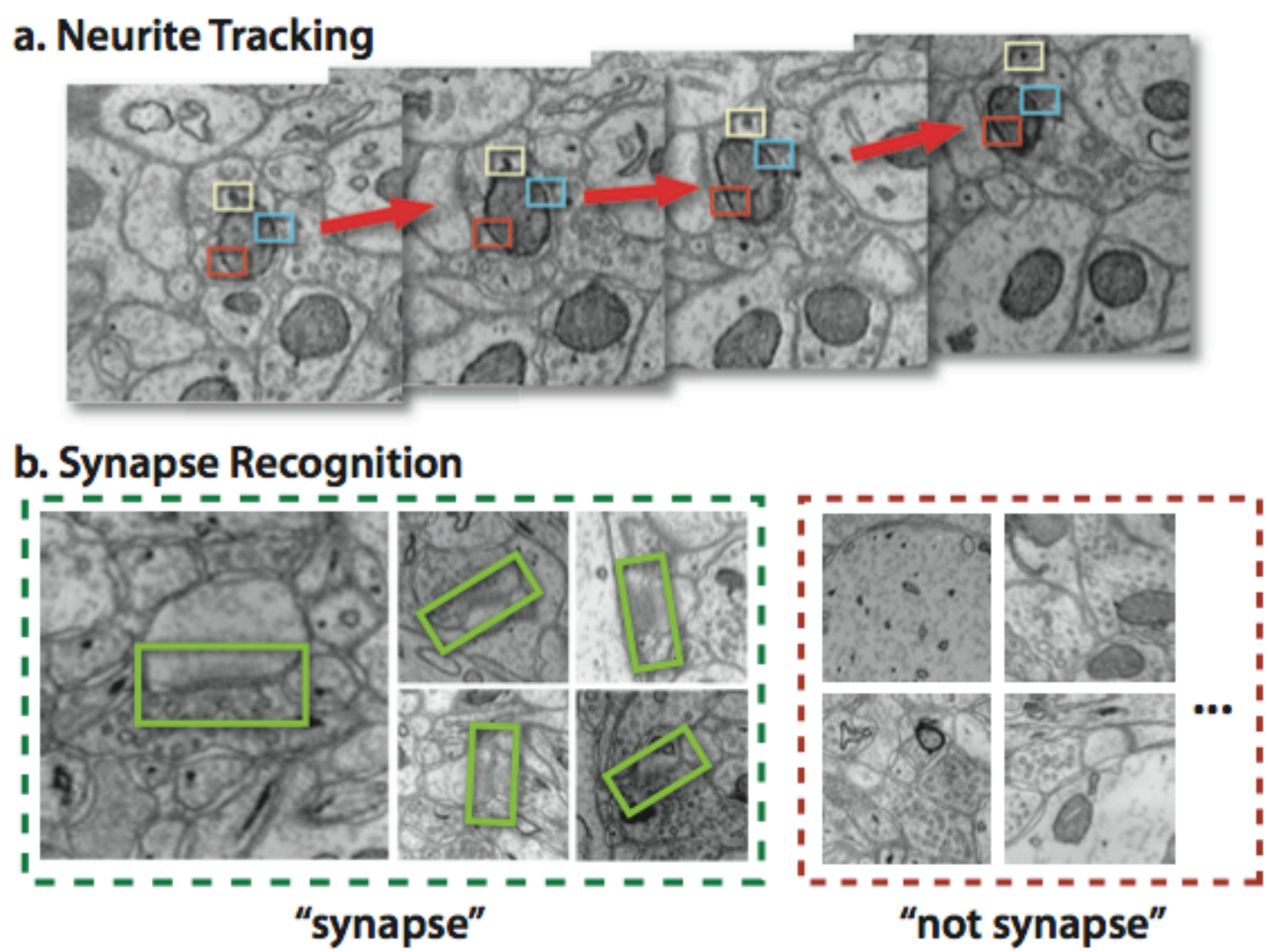
Limitations of existing algorithms

IDSIA Deep Neural Network Segmentation



D. Cirşan, A. Giusti, L. Gambardella, and J. Schmidhuber, "Deep Neural Networks Segment Neuronal Membranes in Electron Microscopy Images" NIPS 2012

New Perspectives

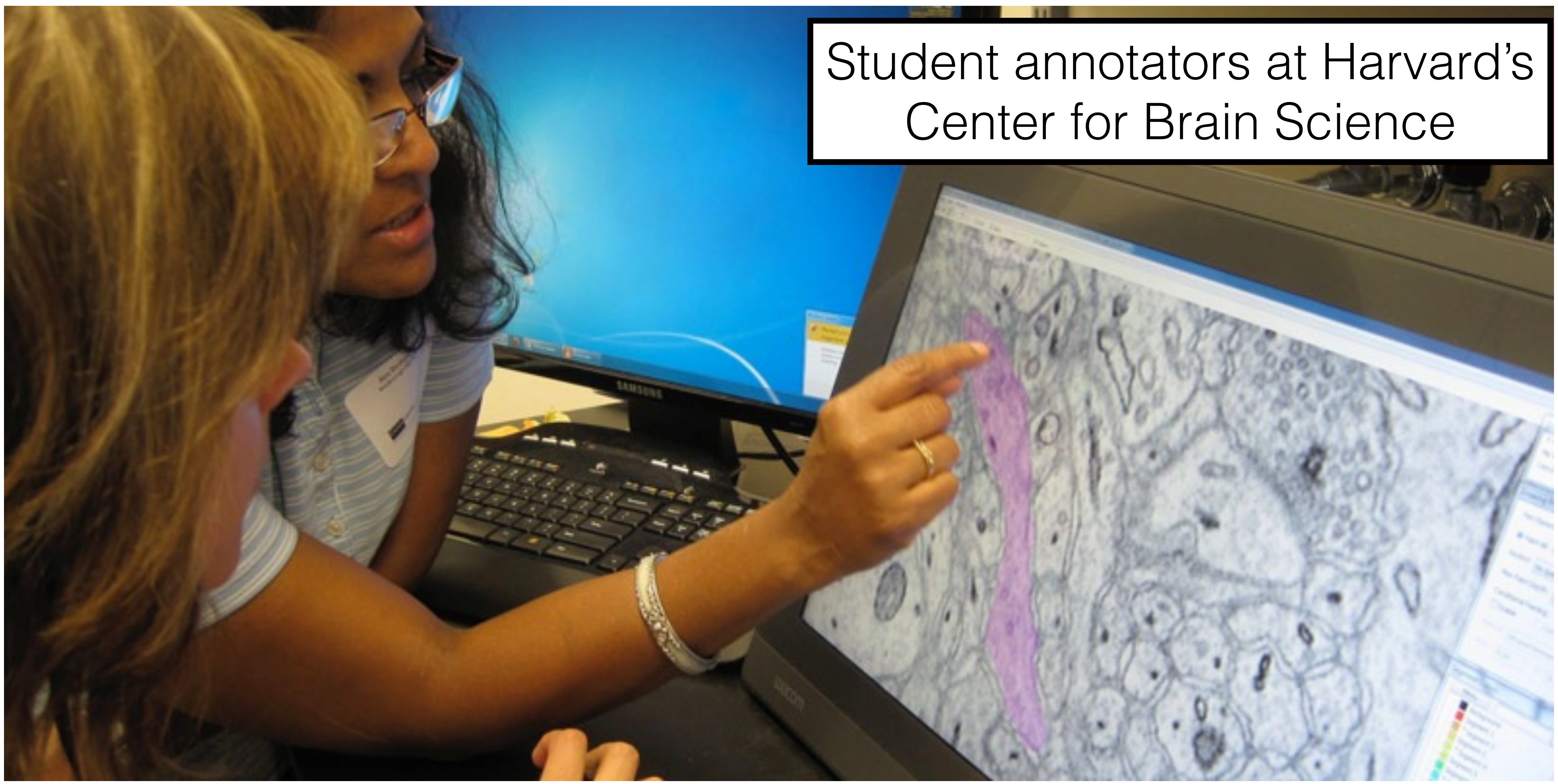


Perceptual Annotation

Looking beyond accuracy and reaction time:

- Stack flips during segmentation
- Trial and error patterns
- Eye movements

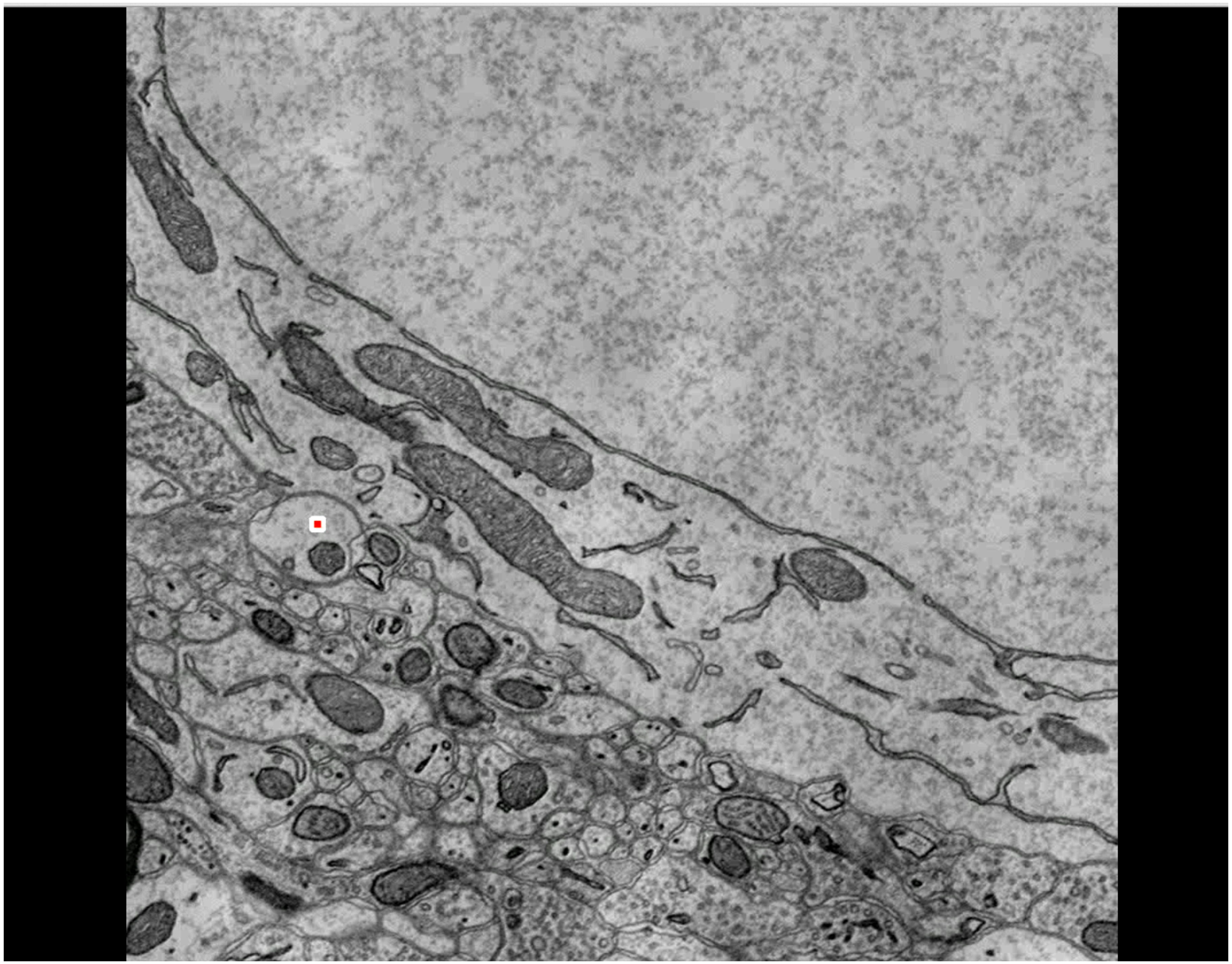
Learn how humans segment images



Student annotators at Harvard's Center for Brain Science

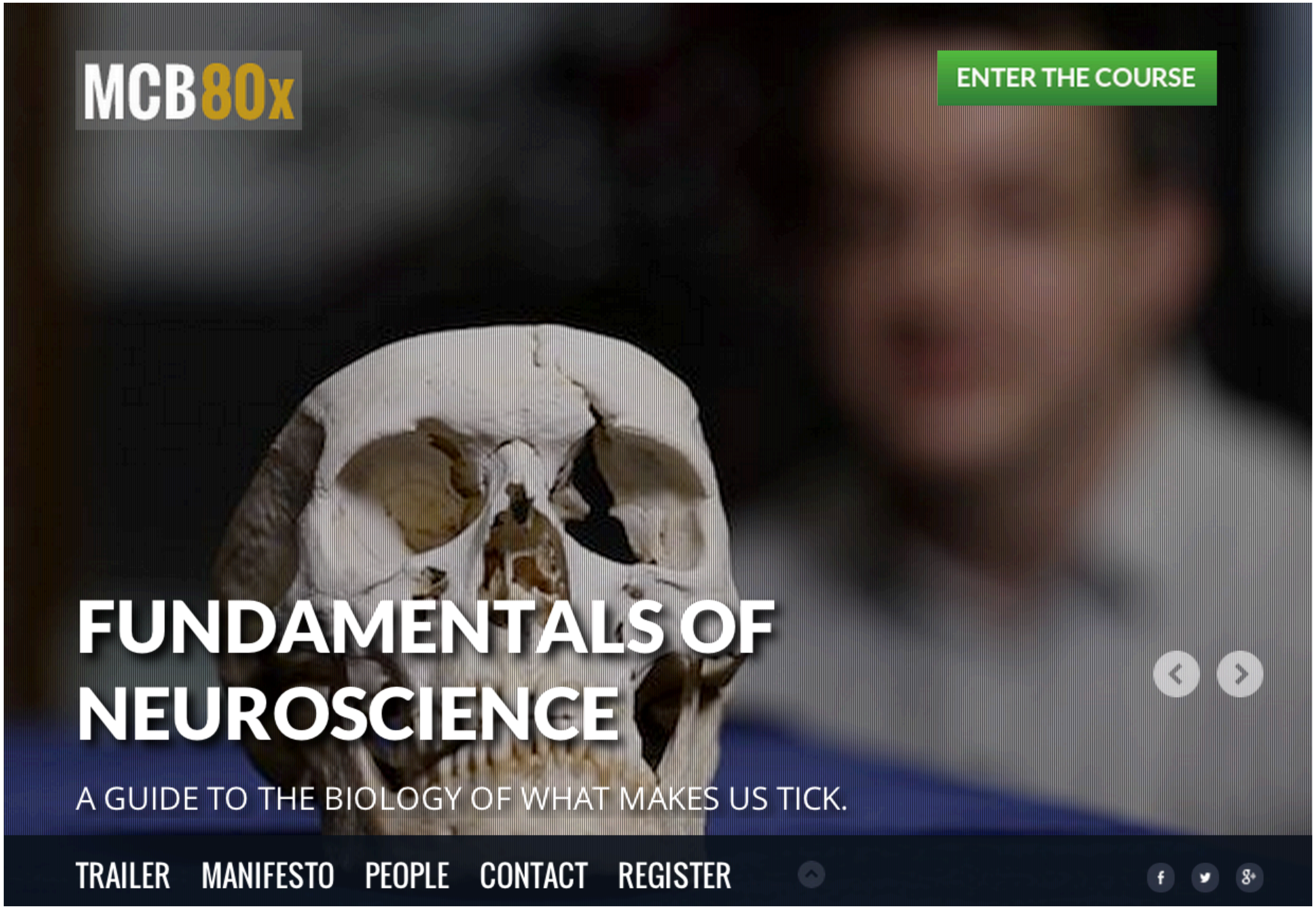
If a region is initially ambiguous to a human annotator, the computer should be given this information.

Eye Movement as a Perceptual Annotation



Online Learning

http://www.mcb80x.org



Let’s think about learning one more time…

How do we know when a student is struggling or understanding?

Skilled teachers rapidly glean clues from implicit cues students provide.

Can the statistics of behavior signals related to learning be captured with machine learning?

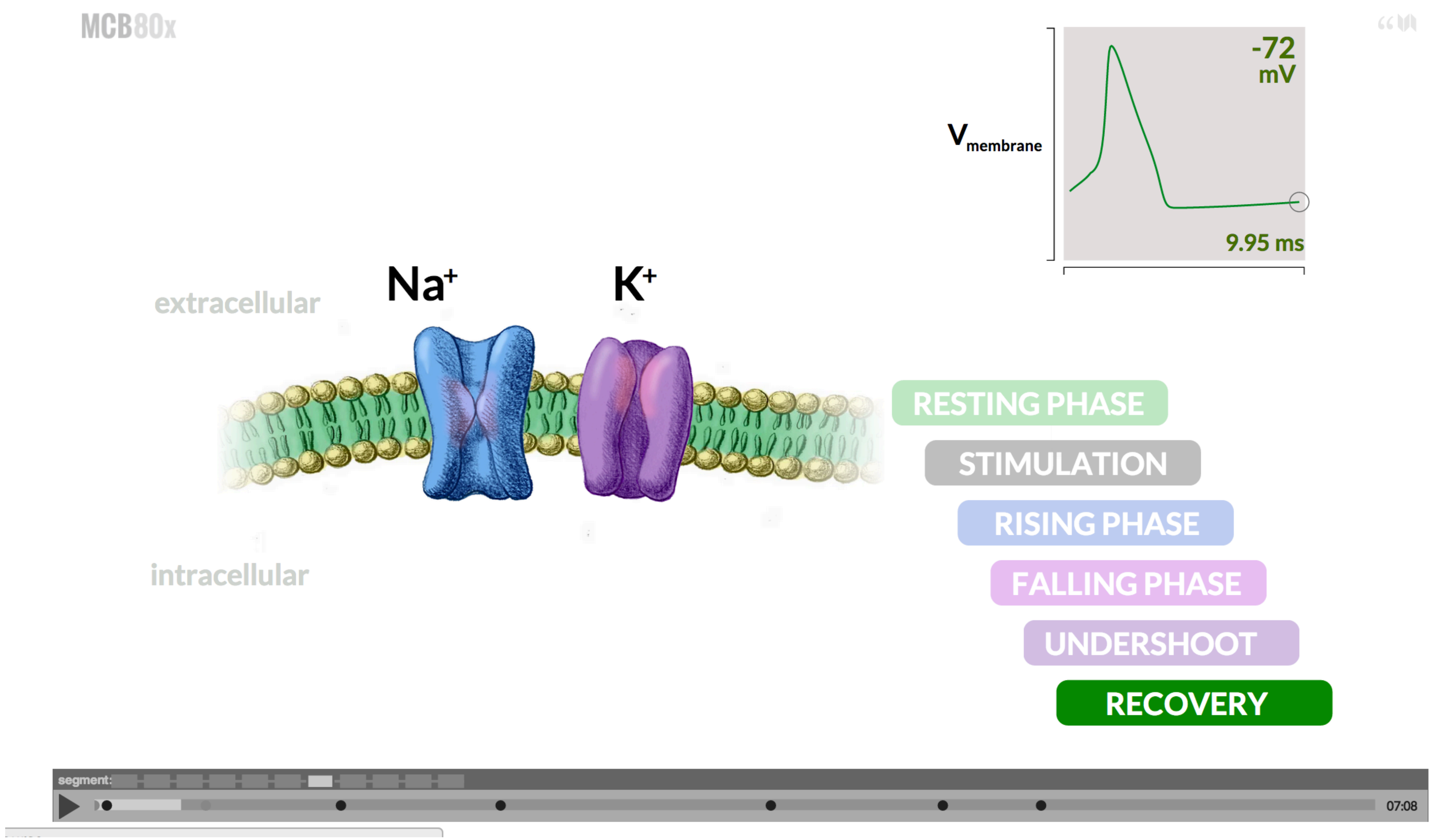
> If yes, then we can create MOOCs that automatically adapt to the needs of individual students.

Perceptual Annotation

Quantifiable patterns in:

- Actions
- Facial expressions
- Facial micro-movements
- Interaction with input devices
- Performance on course activities and tests

mcb80x interactive activities



Wrapping Up…

Collaborators @ Harvard

Sam Anthony

Ken Nakayama

David Cox

Resources

Code:

https://github.com/coxlab/perceptual-annotation

Data:

http://www.wjscheirer.com/datasets/perceptual_annotation/

Paper:

http://www.wjscheirer.com/papers/wjs_tpami2014_perceptual.pdf

TestMyBrain:

http://TestMyBrain.org

Questions?