A look toward the future of object recognition

Deva Ramanan

UC Irvine

Caveat

I'm no fortune teller!

I'll try to frame the discussion around what I might like to hear if I were in your shoes.

Most of these thoughts are due to discussion with students, colleagues,...

A quote...

"Good researchers know how to solve problems;
great researchers know what problems are worth solving"

-A senior colleague

What's the killer app for computer vision?

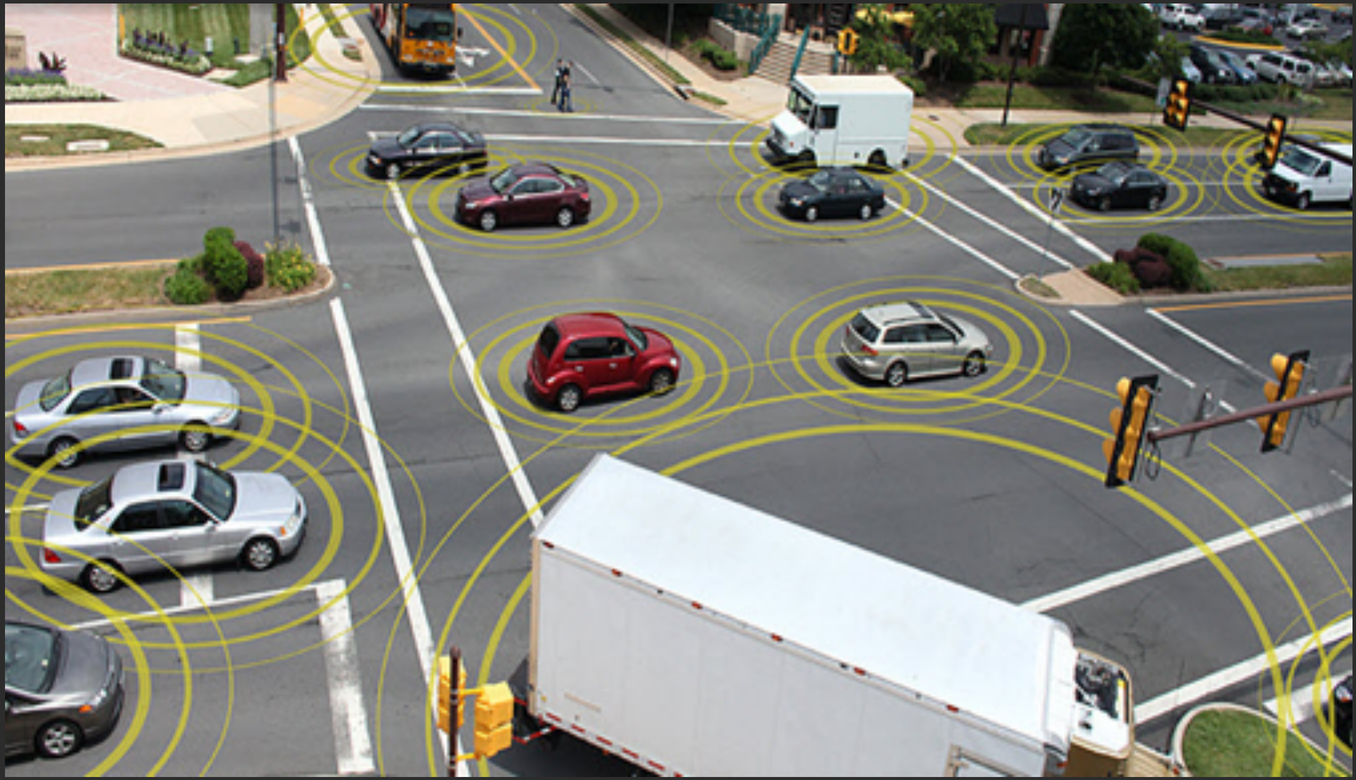
...its worth revisiting the tasks we're considering

"Good researchers know how to solve problems;
great researchers know what problems are worth solving"

-A senior colleague

Some proposals:

Visual perception for self-driving cars



Some proposals:

Reconstruction of 4D world

Some proposals:

Surveillance (while ensuring privacy?)

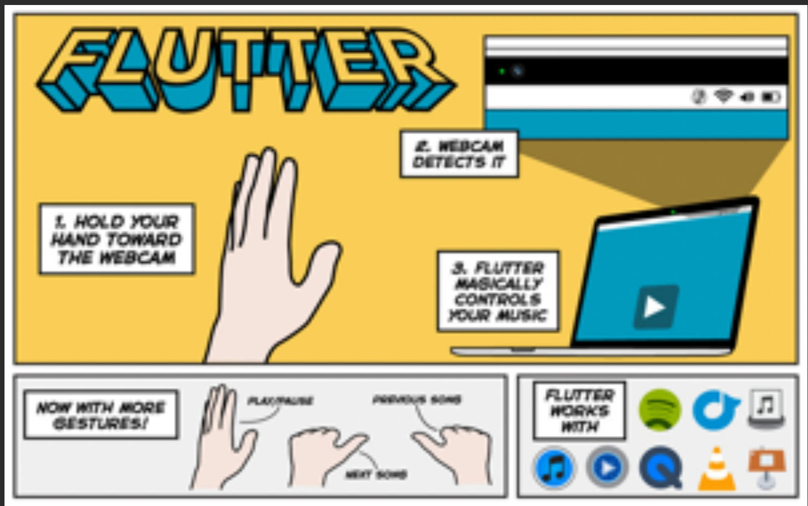
“The work was painstaking and mind-numbing: One agent watched the same segment of video 400 times.The goal was to construct a timeline of images, following possible suspects as they moved along the sidewalks, building a narrative out of a random jumble of pictures from thousands of different phones and cameras. It took a couple of days, but analysts began to focus on two men in baseball caps who had brought heavy black bags into the crowd near the marathon’s finish line but left without those bags.”

Washington Post

Some proposals:

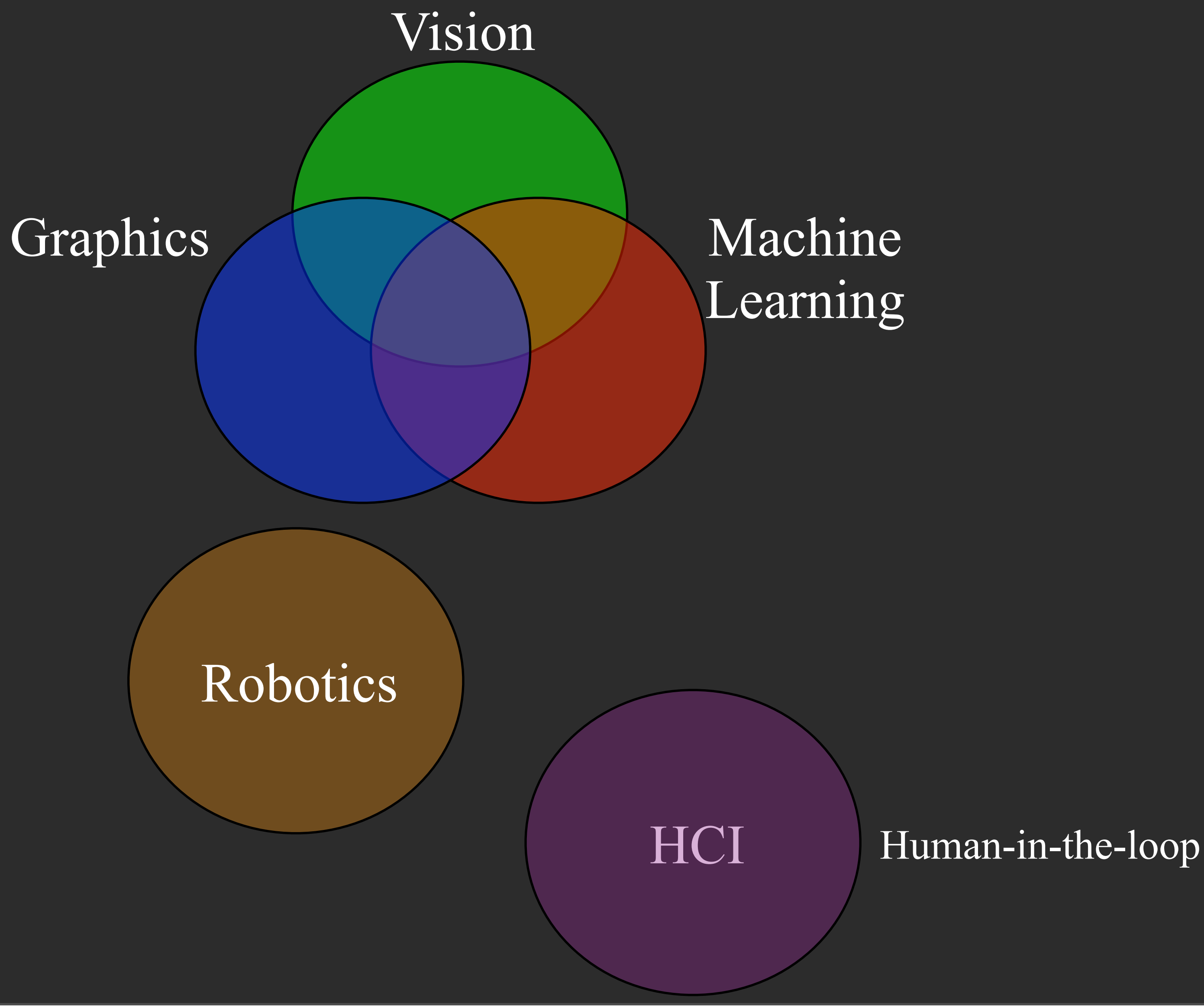
Assistive/medical technology

Entrepreneurial vision



Finding the right app will probably make you some $

Bird’s eye view



But we're scientists (not engineers), right?

Romantic notions of AI

Replicate human visual system

See-ing robot

What should a vision system report?

Object/scene/action category labels

Segmentations

Attributes

....

What is the relevant perceptual output here?

Learning to predict the future

In general, temporal analysis still seems to be a second-class citizen in the world of recognition

Relatively in its infancy compared to static-image recognition

Direction 1: integration of video into recognition

Using video for learning

8 years worth of video is uploaded to YouTube... each day

Humans arguably use motion

Biological motivation

Hubel and Weisel's iconic experiments on simple vs complex "pooling"cells

Complex cells are tuned to movement

"Clicks" are action potentials generated by instrumented cortical neuron

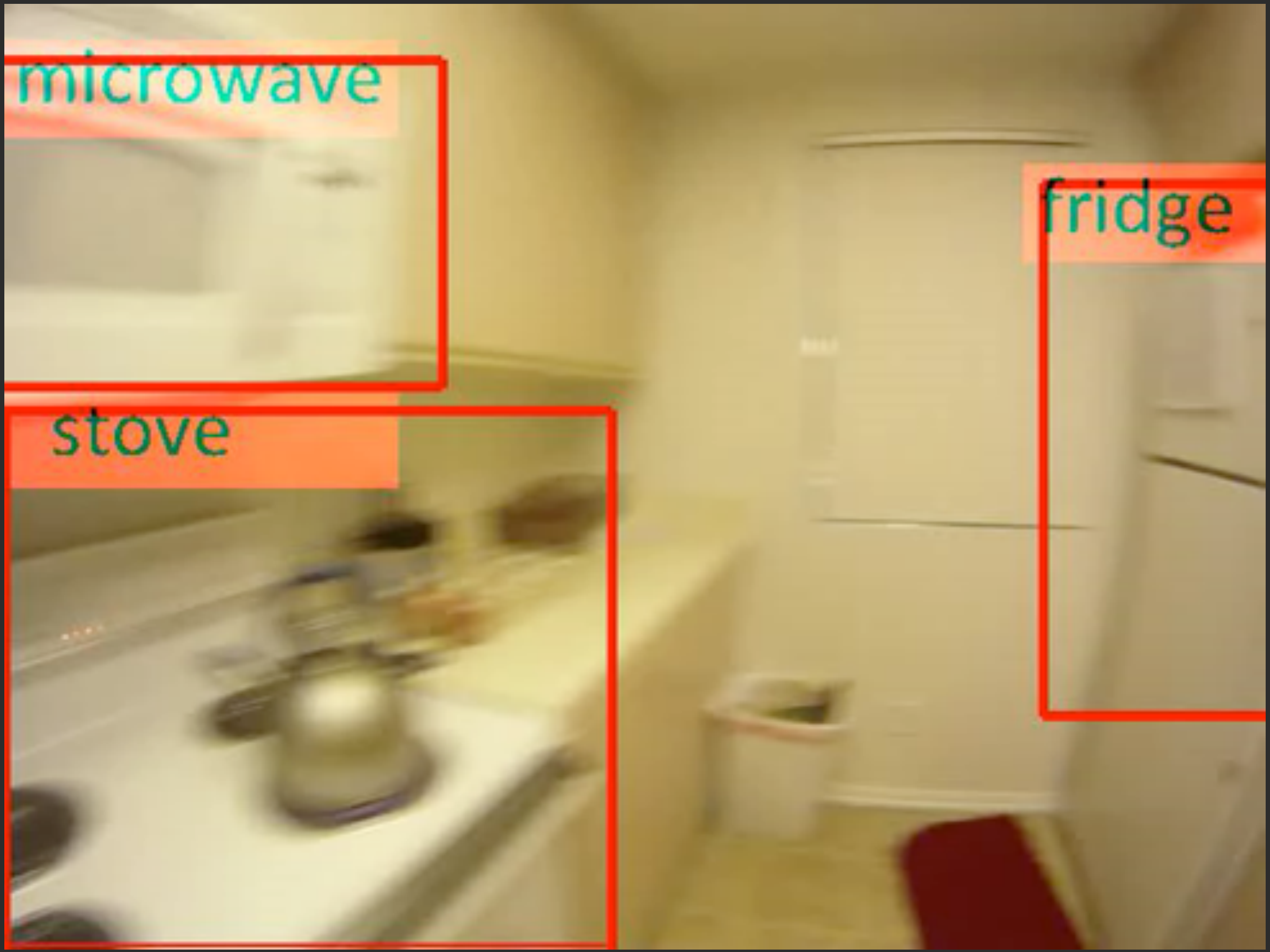
Online never-ending learning

Tom Mitchell's Never Ending Language Learning (NELL)

We should be processing a never-ending stream of input (temporal) data

Lots of untapped formulations for non-iid online learning (experts, bandits, etc.)

Egocentric vision



A killer app?

Functional prediction

If you know what can be done with a ... object, what it can be used for, you can call it whatever you please”

J. J. Gibson. *The Ecological Approach to Visual Perception*




“sittable” affordance label implies someone can sit in the future

Direction 2: Scalability

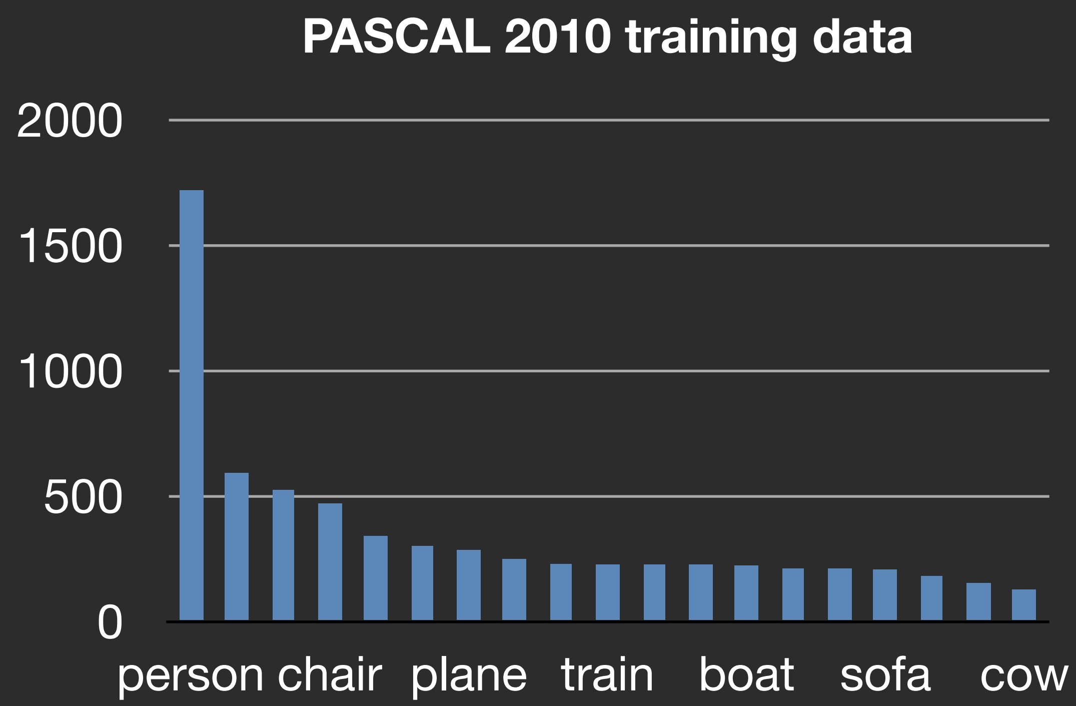
Direction 2: Scalability

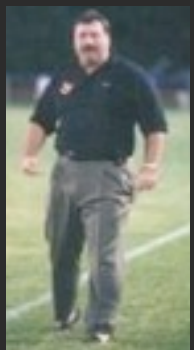
Approach 1: Built thousands of models and compress them

Approach 2: Built representation that scales sublinearly with # of categories

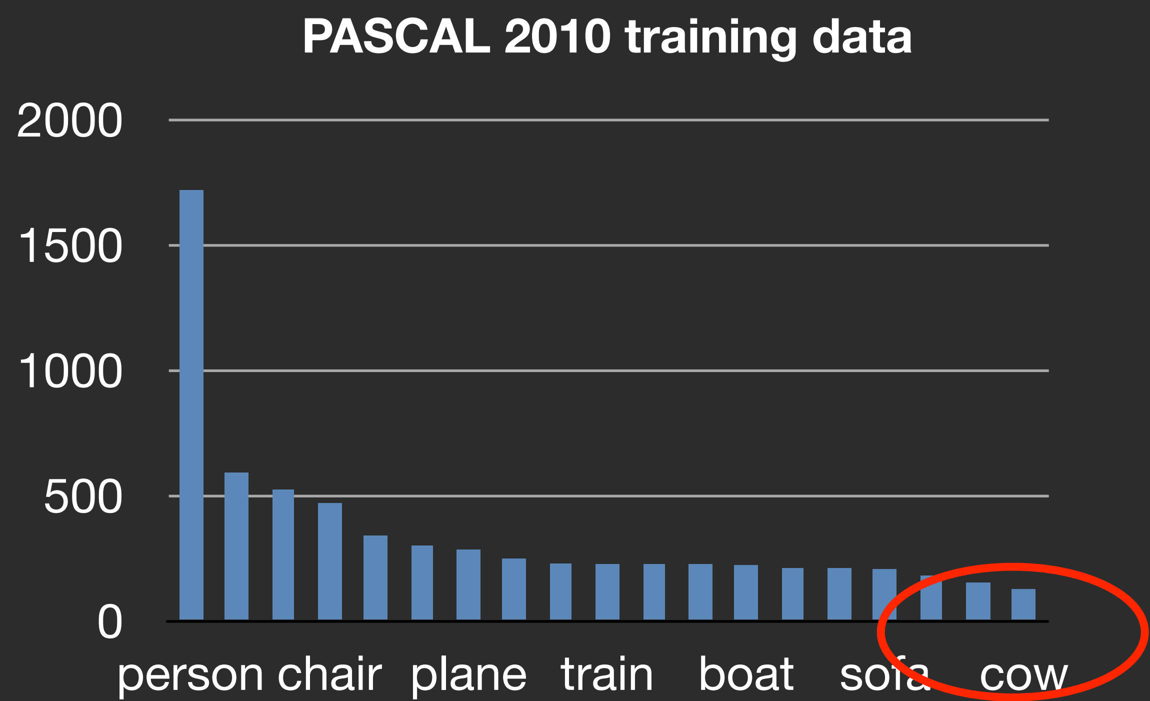
(c.f. compositional models)

Difficulties: long tails



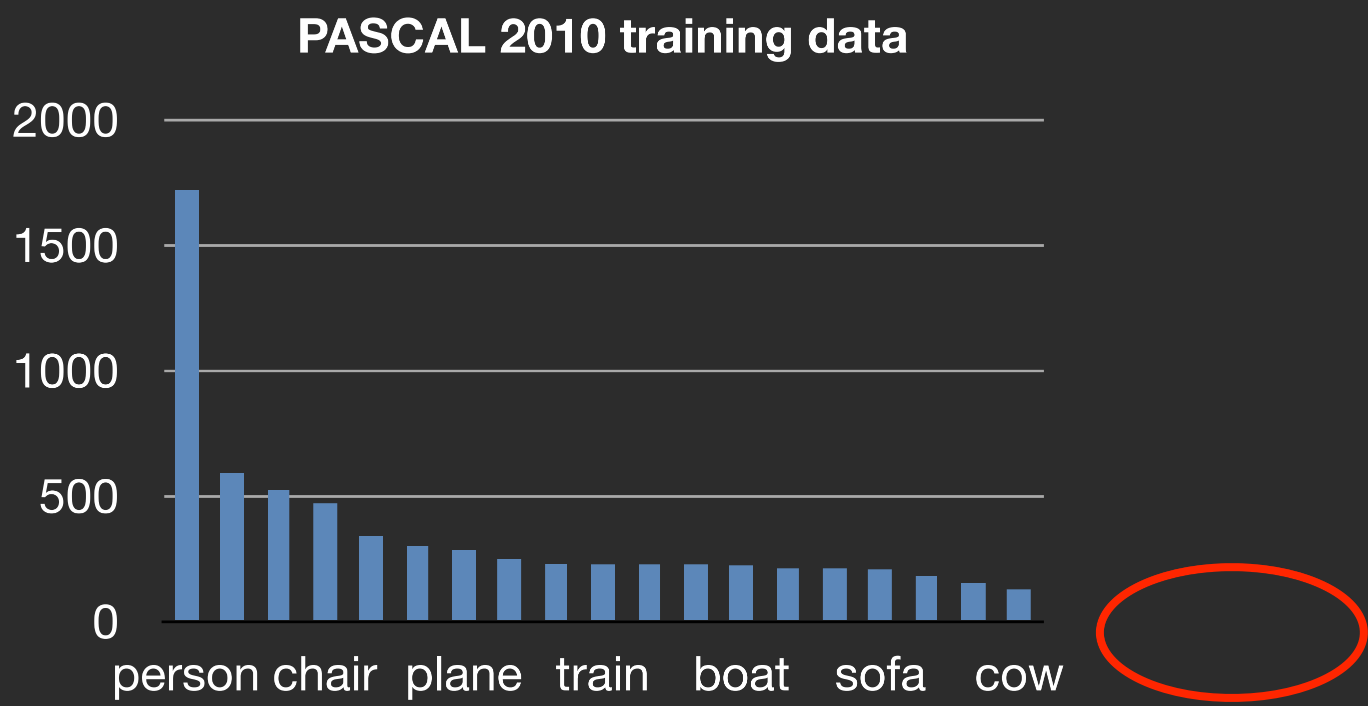


Difficulties: long tails

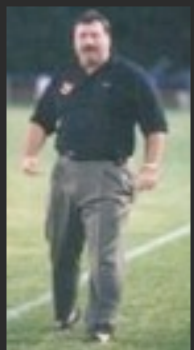


“One-shot learning”: sharing

Difficulties: long tails



“Zero-shot” learning: synthesis



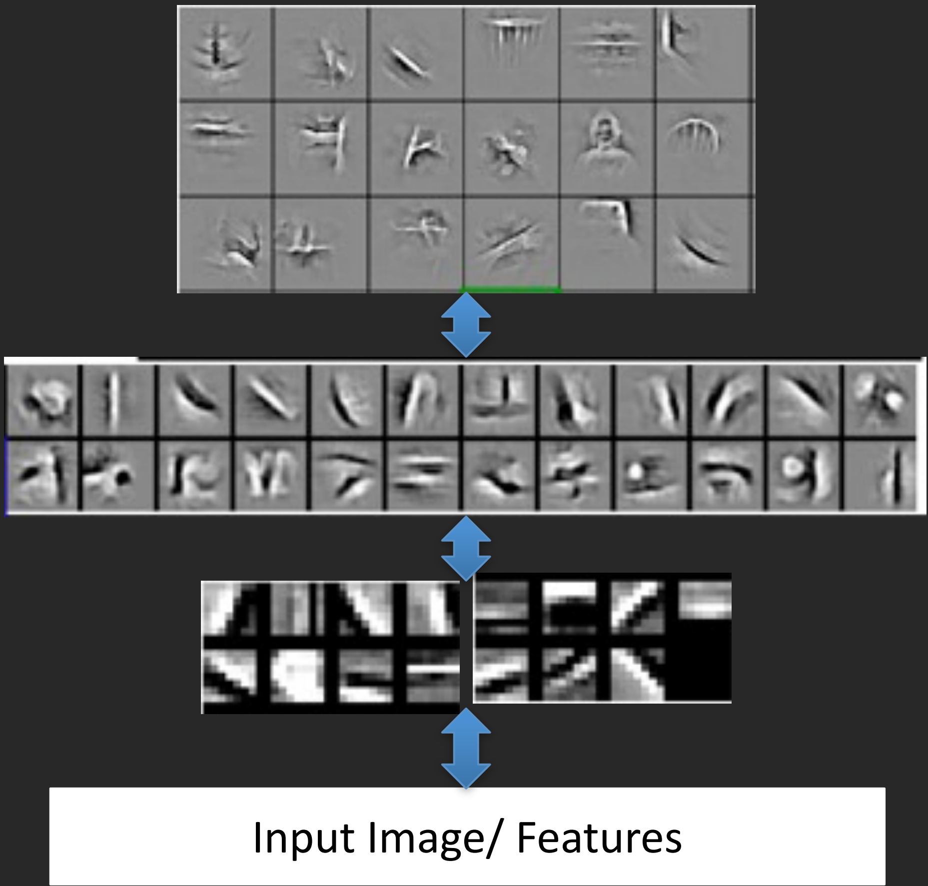
Explicit synthesis

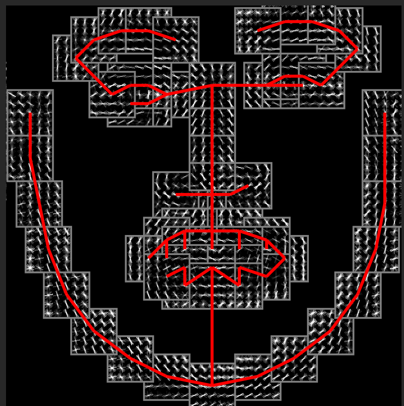
Kinect pose estimation

Subordinate categories

How to sub-linearly encode fine-scale differences between object categories?

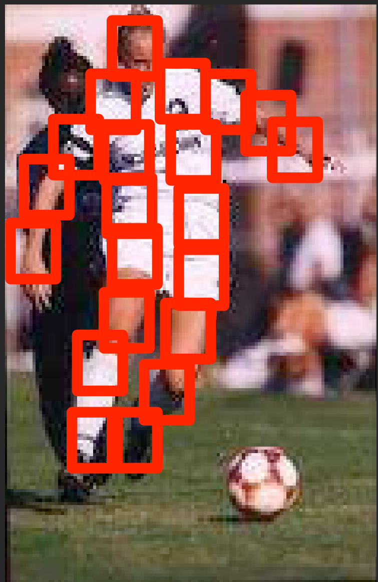
Comparison to deep networks



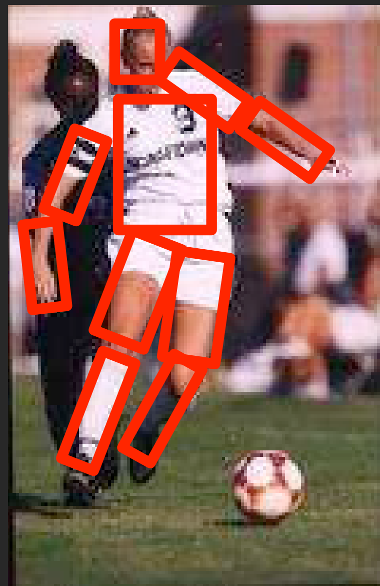


Deep models	Part models
Naturally shares parameters	Difficult to share across categories
Hierarchical	Trees / grammars
Learned representation	Engineered representation
Difficult to train (need lots of data)	Easier to train (100’s of examples)

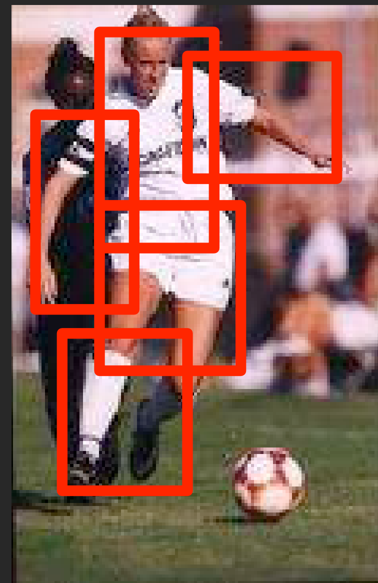
Various representations



Patches
2011



Skeleton
1970's

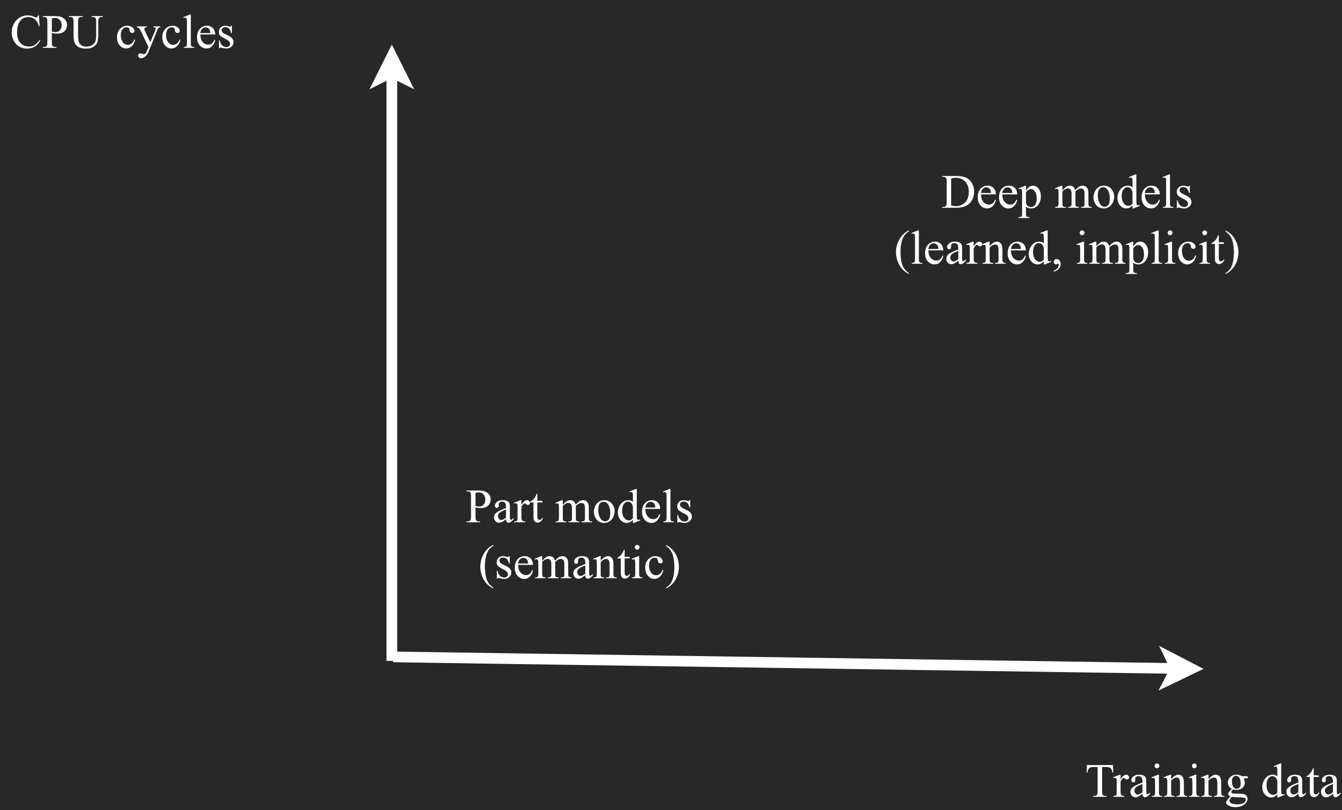


Poselets
2009

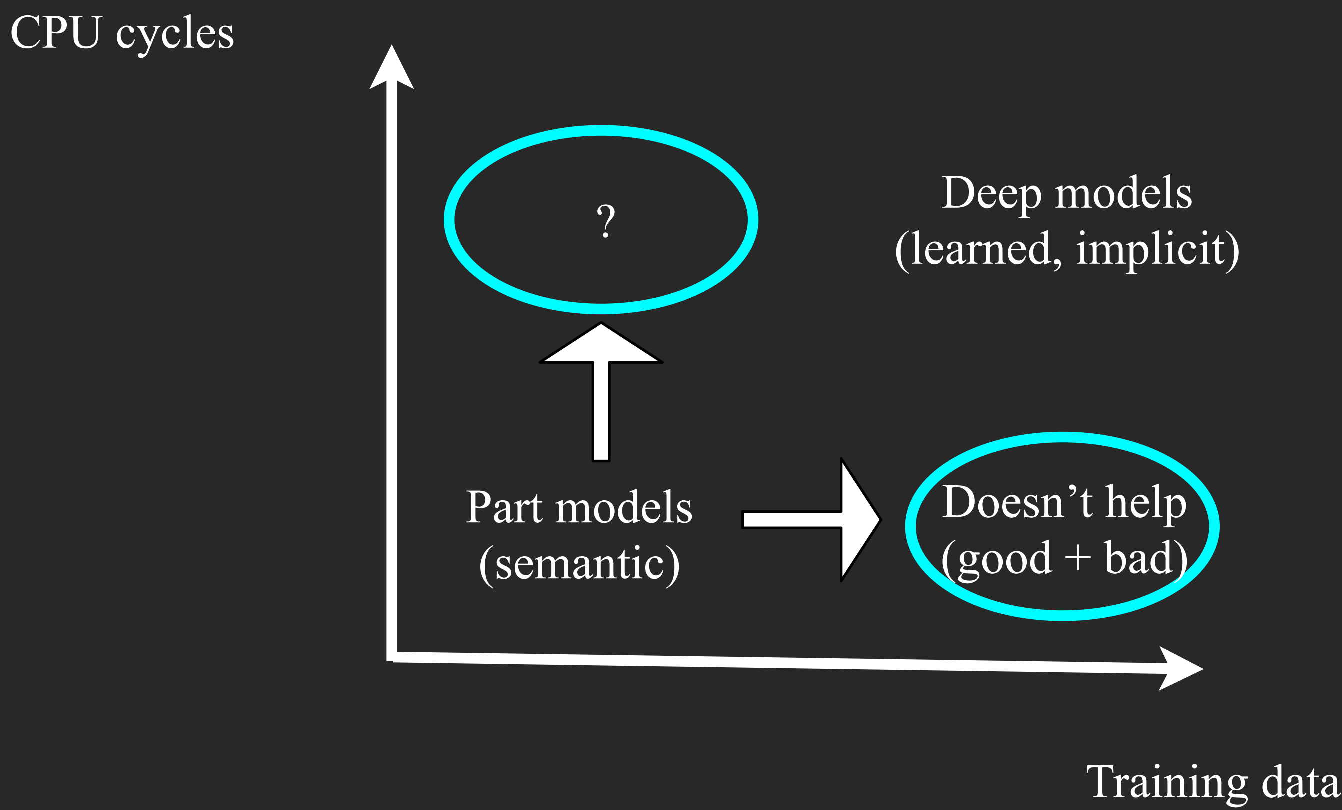
As a field, we perform a human-in-the-loop search over representations, at the time-scale of years or decades

We must be able to do better!

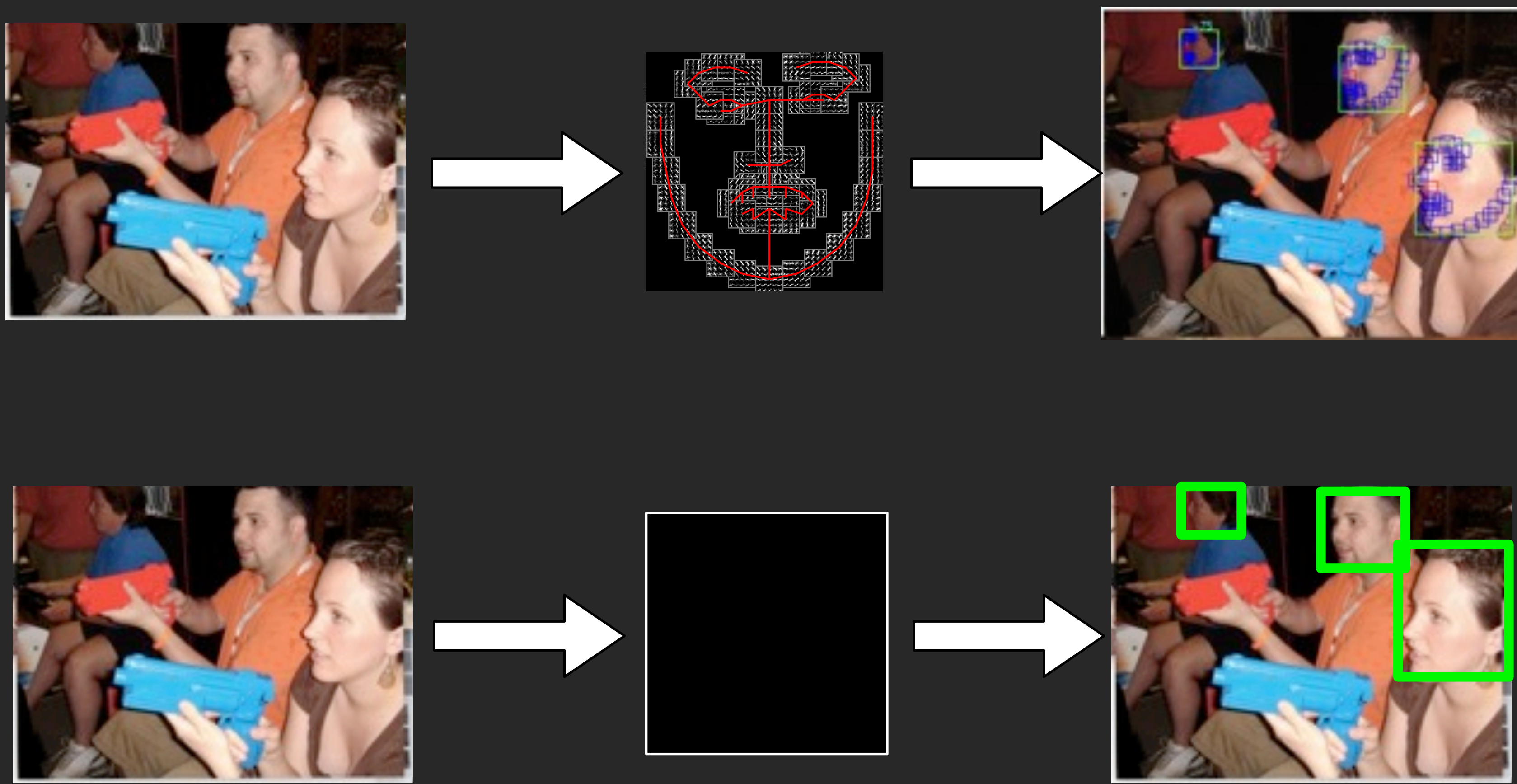
Thought experiment



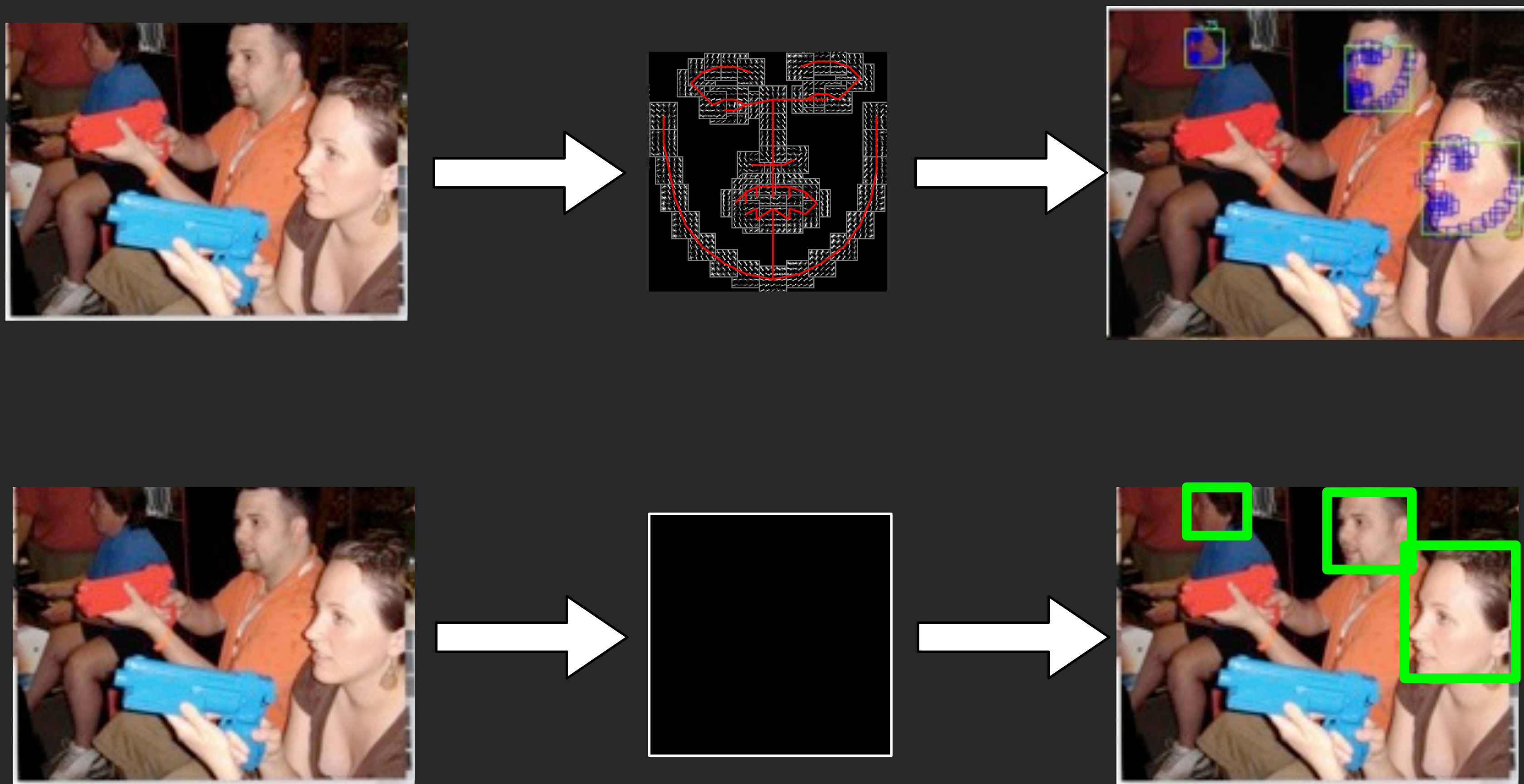
Thought experiment



Why should representations be interpretable?

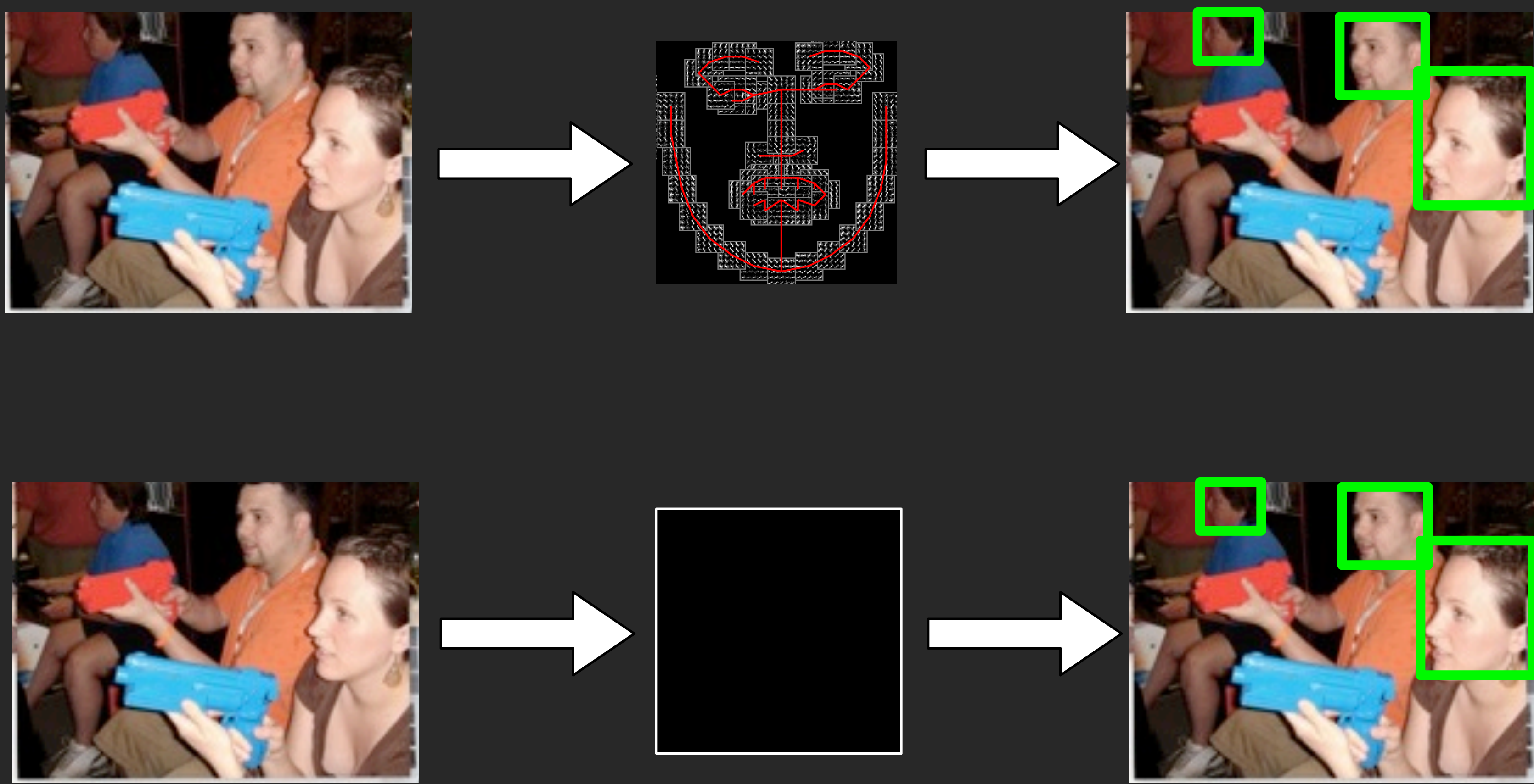


Why should representations be interpretable?



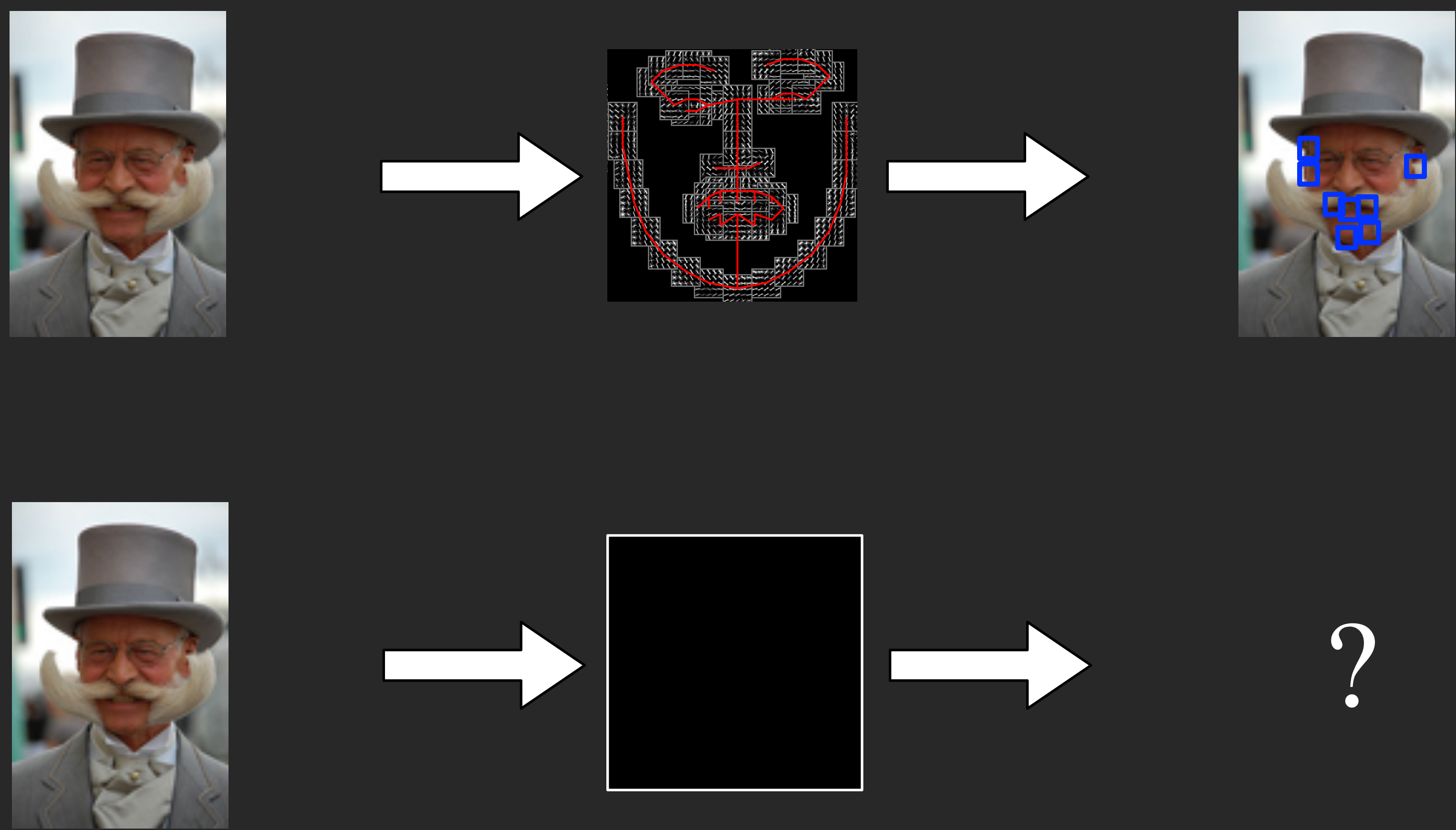
Detailed outputs (pose, landmarks) seem to "force" the black box to internally represent 3D shape

Why should representations be interpretable?



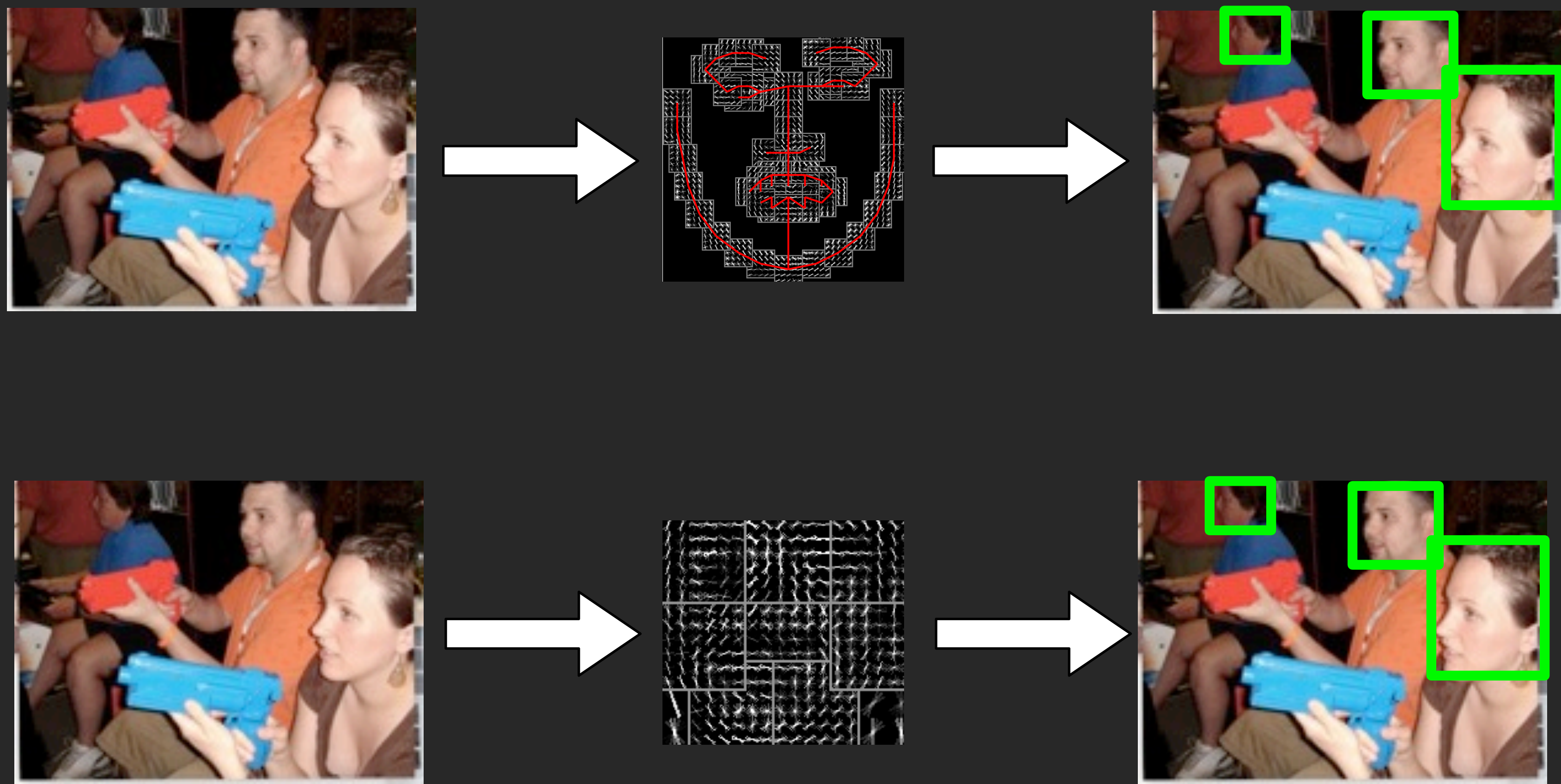
Ignoring that, why do we need explicit semantic representations?

Why should representations be interpretable?



Practical issue (dataset bias)

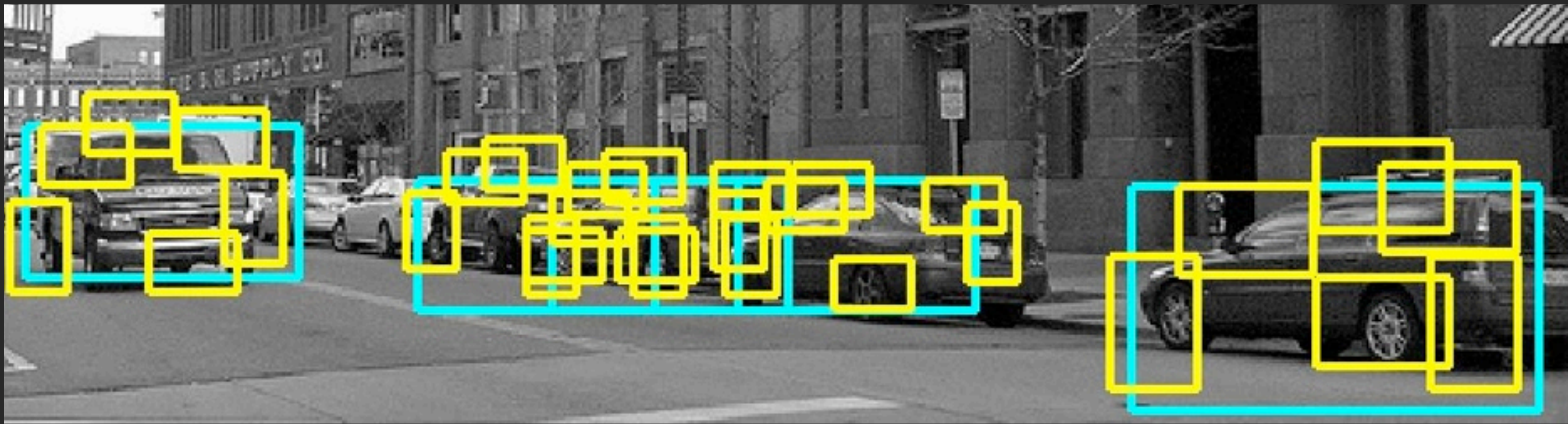
Post-hoc interpretation

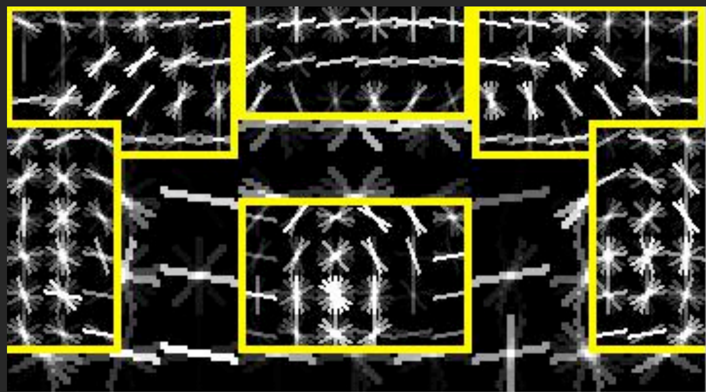


Perhaps in retrospect, we'll be able to visualize/interpret the black box

If so, do semantic constructs (eyes, mouths) play any role during learning?

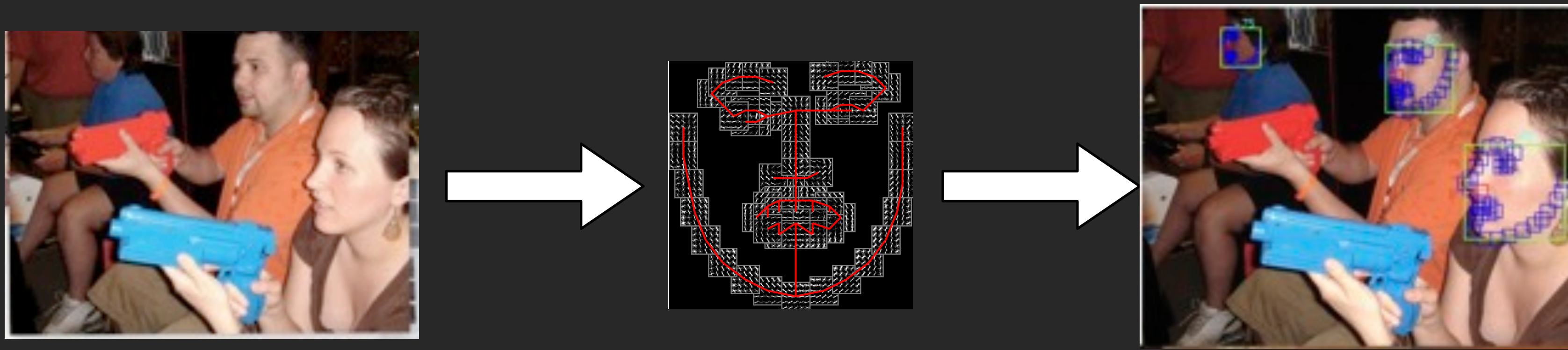
Post-hoc interpretation

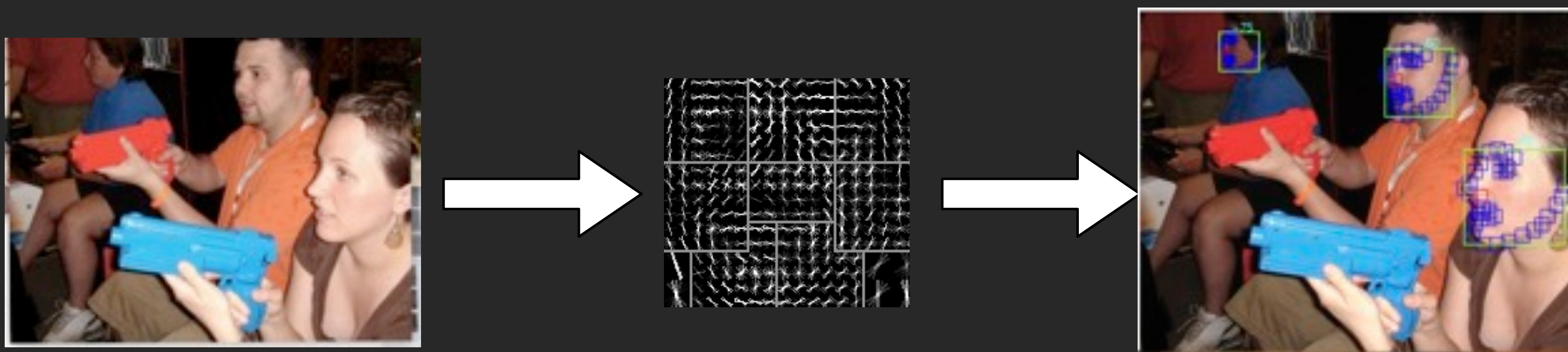




The three "wheel" parts sometimes fire on non-wheels.
We thought this meant that this was the wrong representation

Post-hoc interpretation



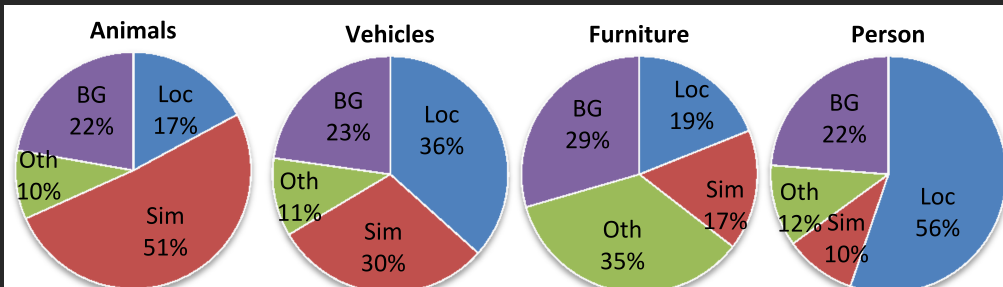


Perhaps in retrospect, we’ll be able to visualize/interpret the black box

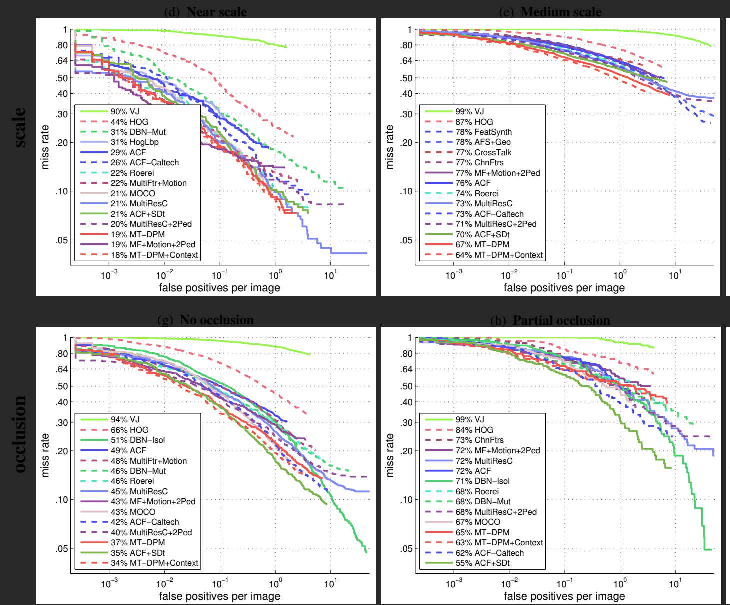
If so, what is the role of semantic tokens (eyes, mouths) when learning models?

Perhaps they are most crucial in defining the output...

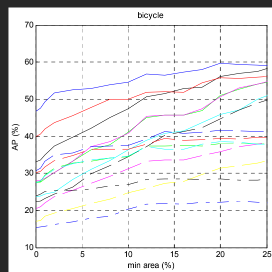
Direction 3: Diagnostic evaluation



Hoeim et al, ECCV12



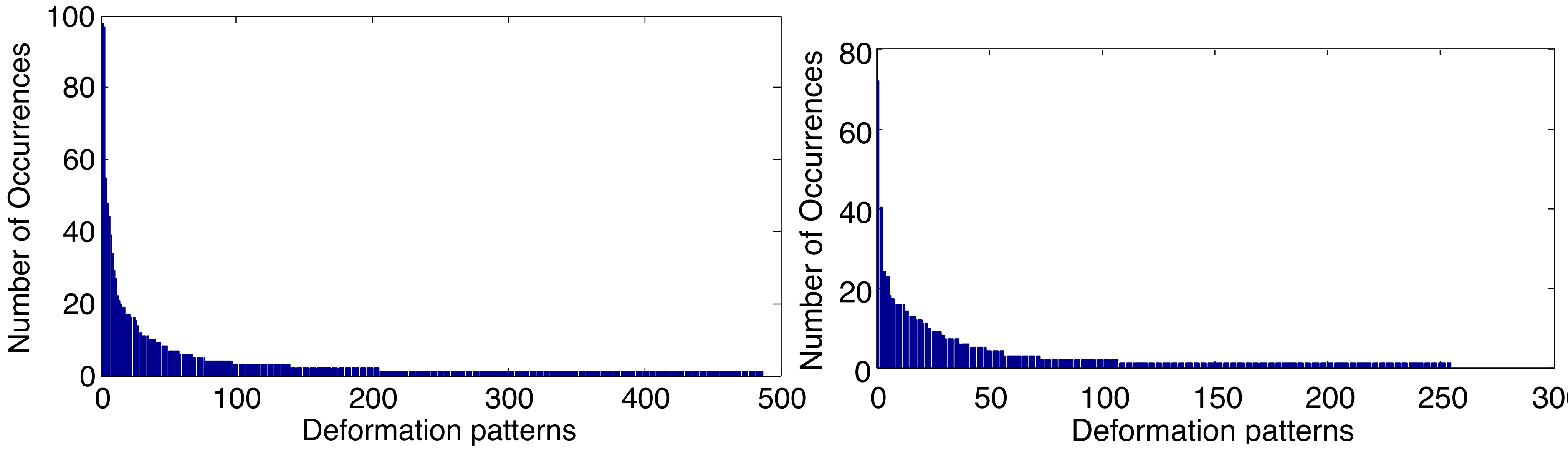
Dollar et al, 12



Everingham et al, IJCV10

Claim: diagnostic evaluation is just as important than dataset collection, but is even less appreciated

Long tails complicate evaluation



(a) Bus (b) Face

Less training instances mean unusual poses are harder to get right

Less test instances mean unusual poses impact performance less

A look back

Pick a good problem
(c.f. robotics, HCI)

Putting temporal reasoning back into recognition
(more training data, online learning, functional labels)

Scalable representations
(semantic vs learned vs interpretable)

Diagnostic evaluation
(systematic progress)