Learning perception independent of goals: self-supervised learning

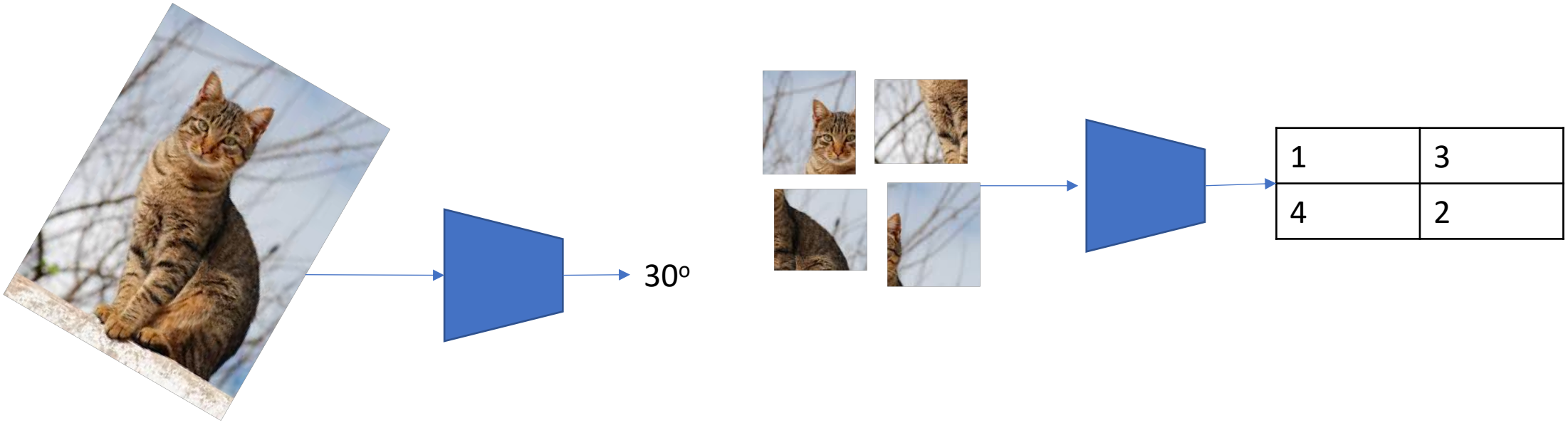
- An embodied agent can’t depend on labels
- Reinforcement learning goals are inevitably tied to particular goals / tasks
- Need another way to build good feature representation.

Pretext task

- Labels are unavailable
- Idea: create your own labels from data
- “Pretext” task
- Hope: Solving the task leads to good feature representations

Pretext tasks

- Transform input, task network with predicting transformation

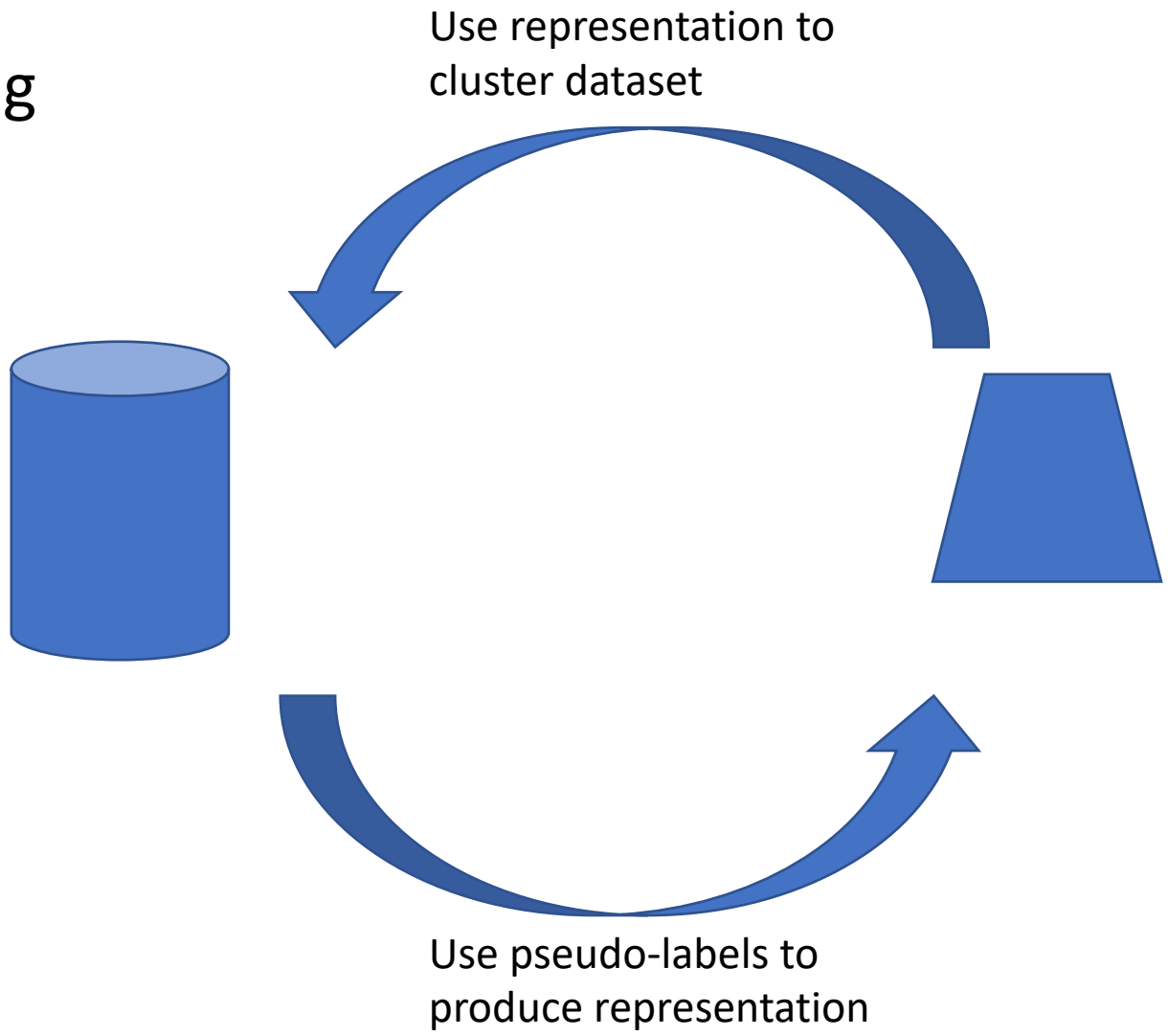


Pretext tasks

- Remove data, then task network with predicting it

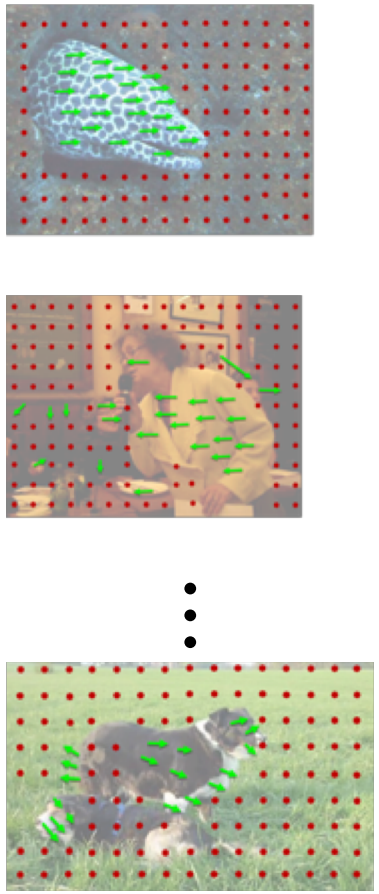
Pretext tasks

- Clustering

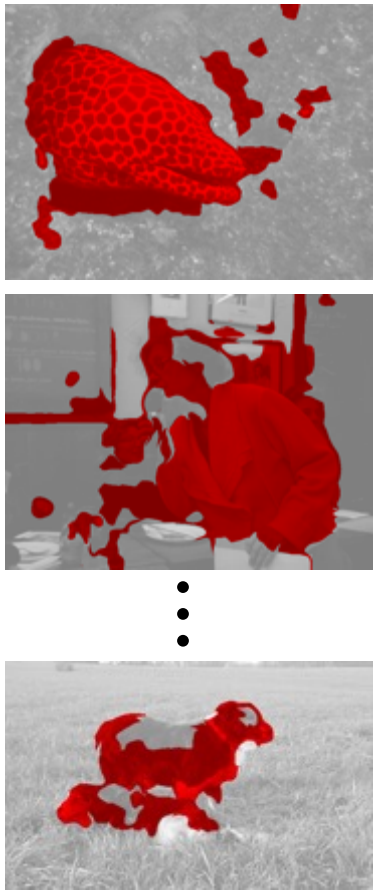


Pretext tasks

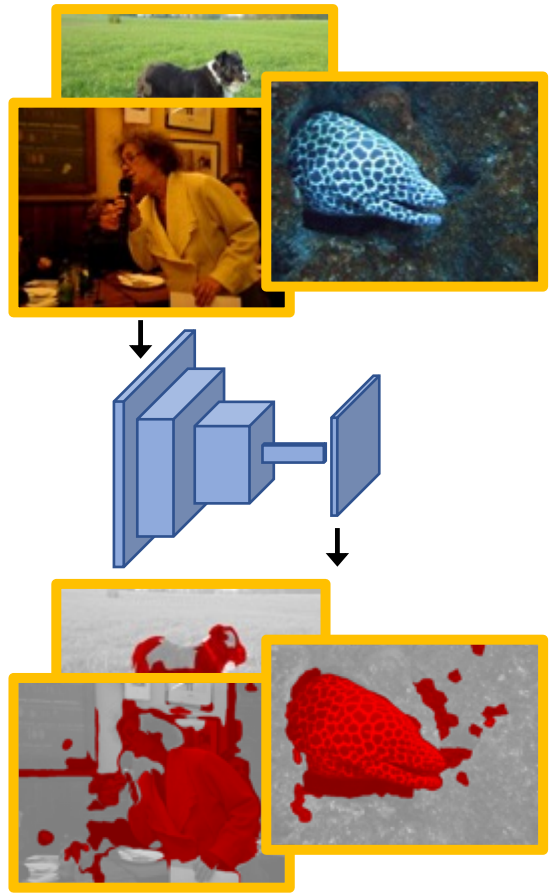
- Use some source with additional data
- E.g. videos



1. Collect videos



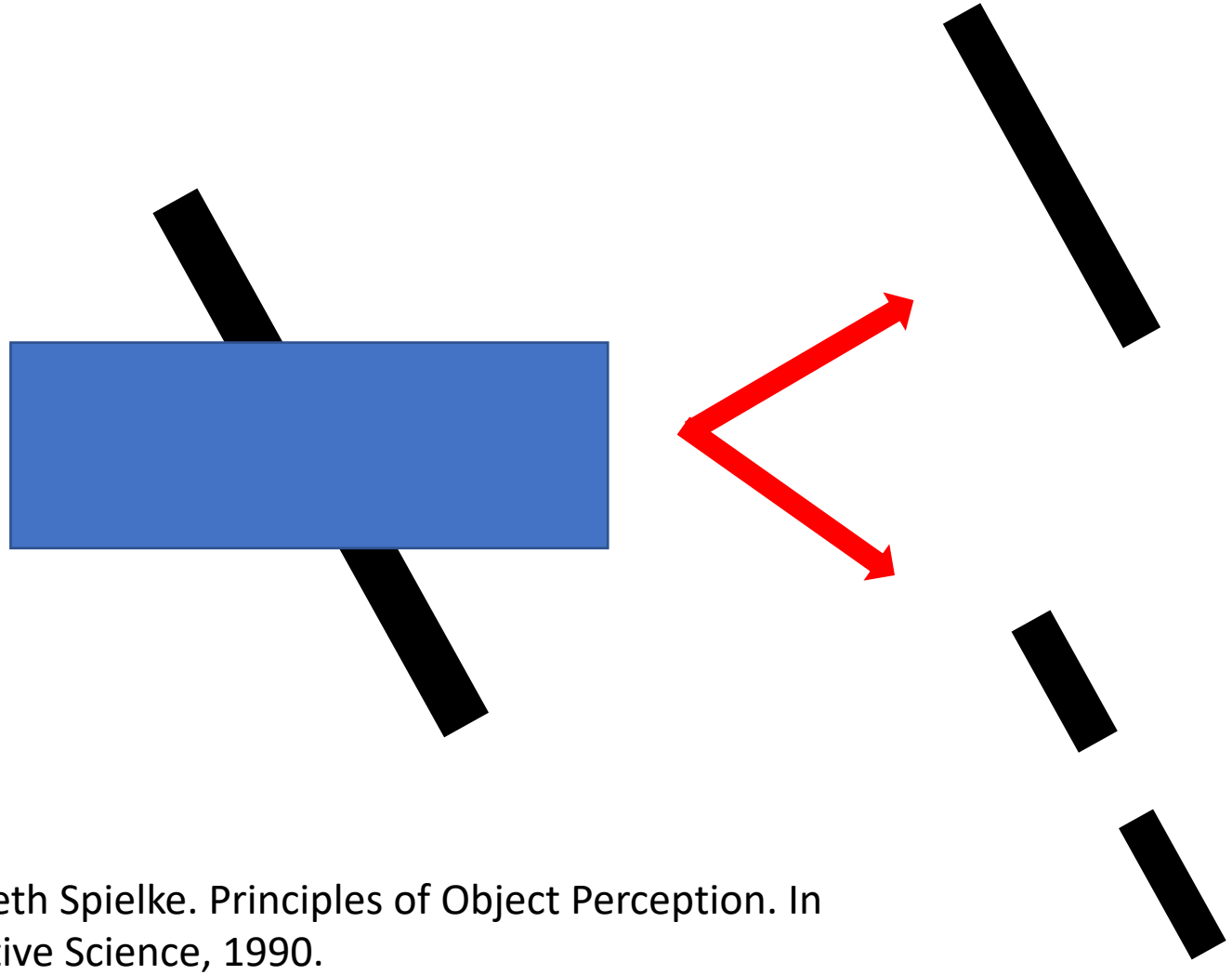
2. Segment using motion



3. Train ConvNet

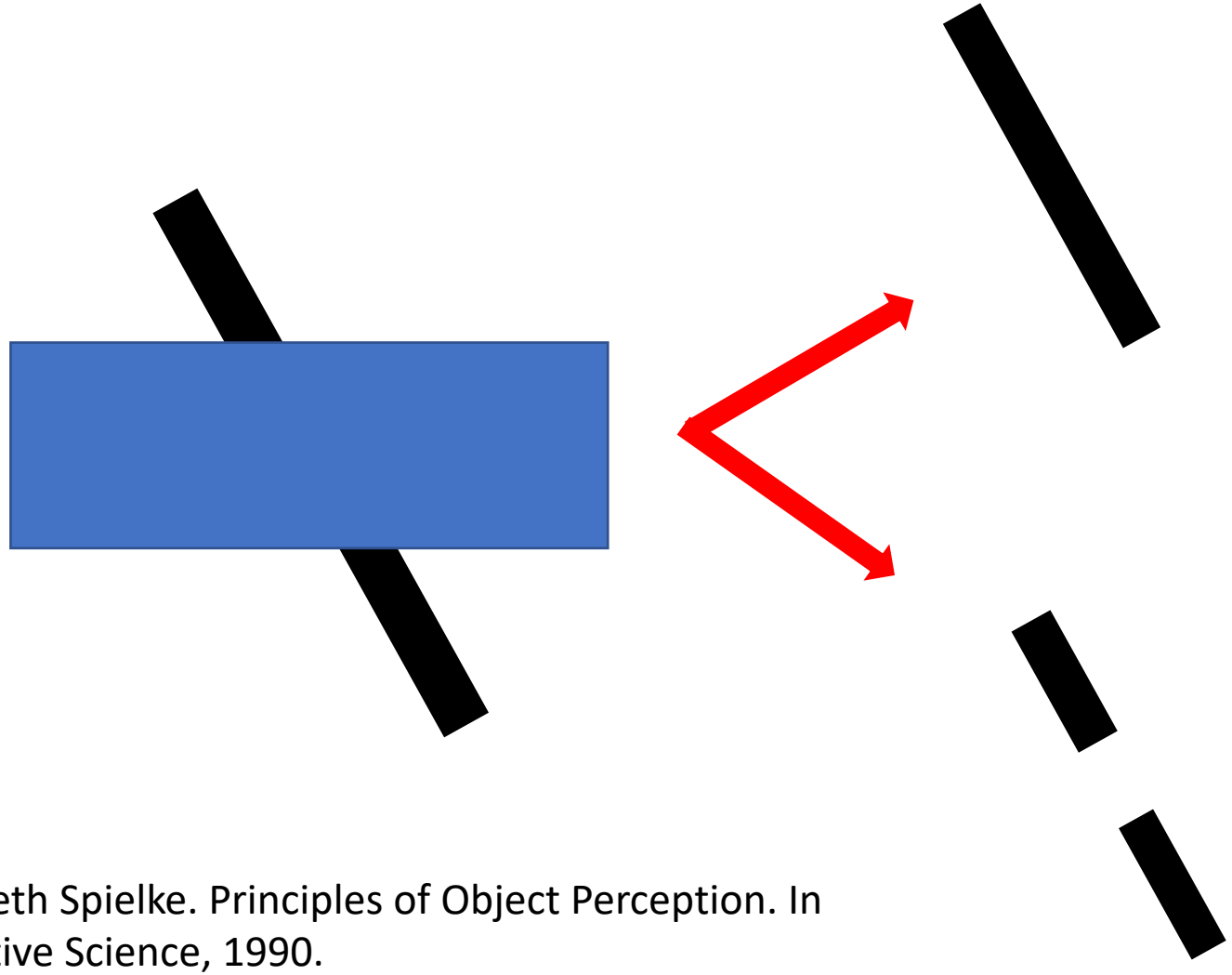
Pathak, Deepak, et al. "Learning Features by Watching Objects Move." *CVPR*. Vol. 1. No. 2. 2017.

What do humans do?



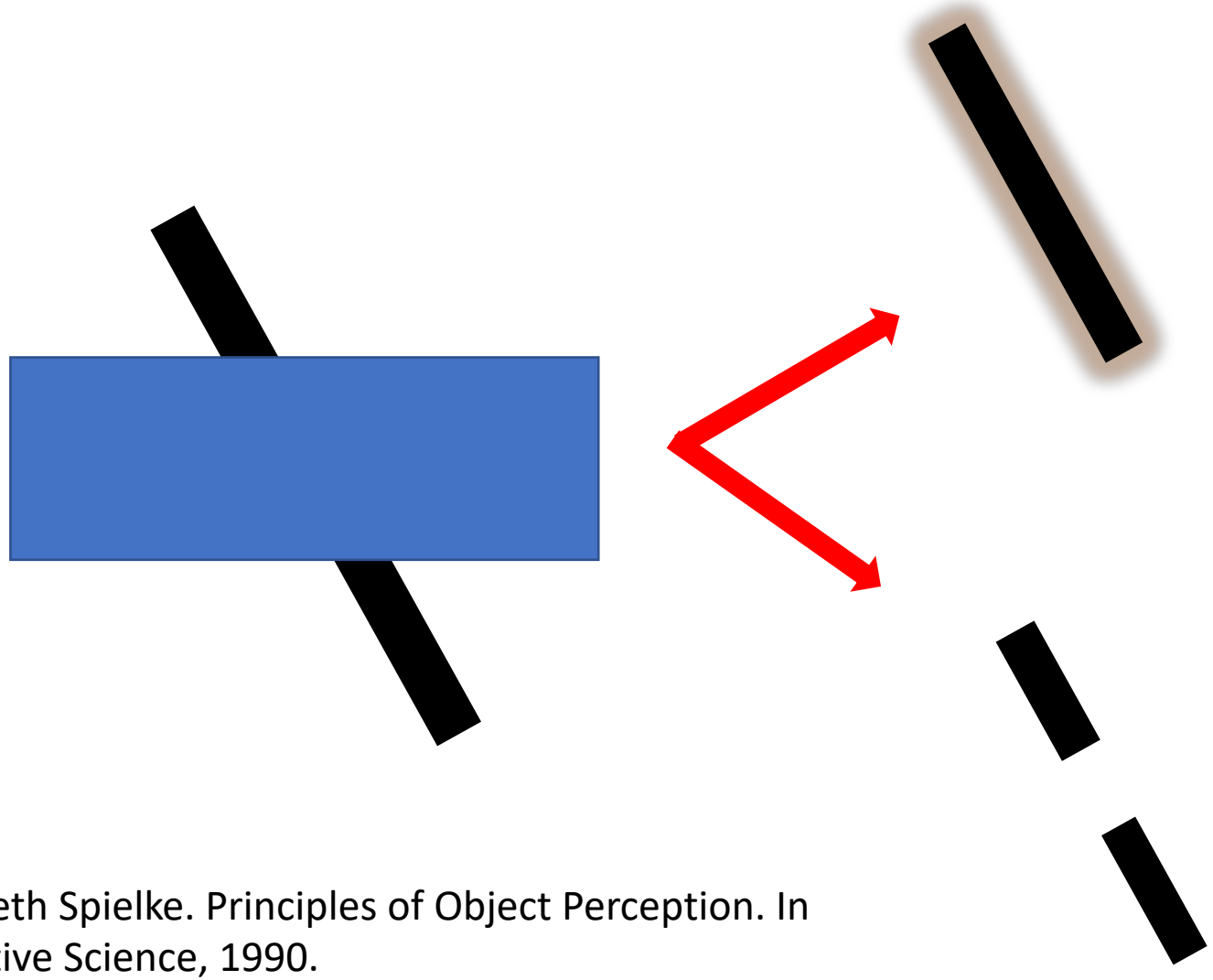
Elizabeth Spielke. Principles of Object Perception. In Cognitive Science, 1990.

What do humans do?



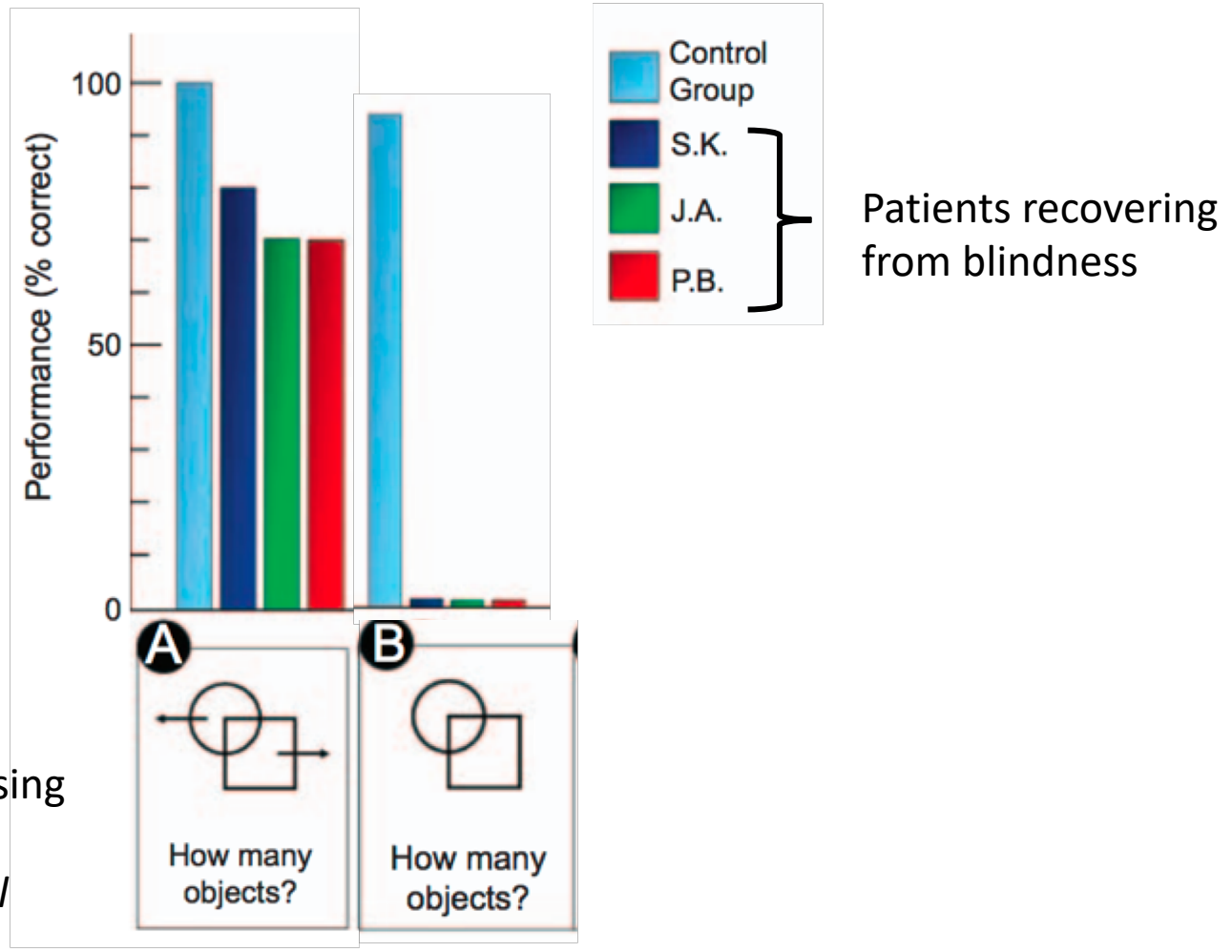
Elizabeth Spielke. Principles of Object Perception. In Cognitive Science, 1990.

What do humans do?



Elizabeth Spielke. Principles of Object Perception. In Cognitive Science, 1990.

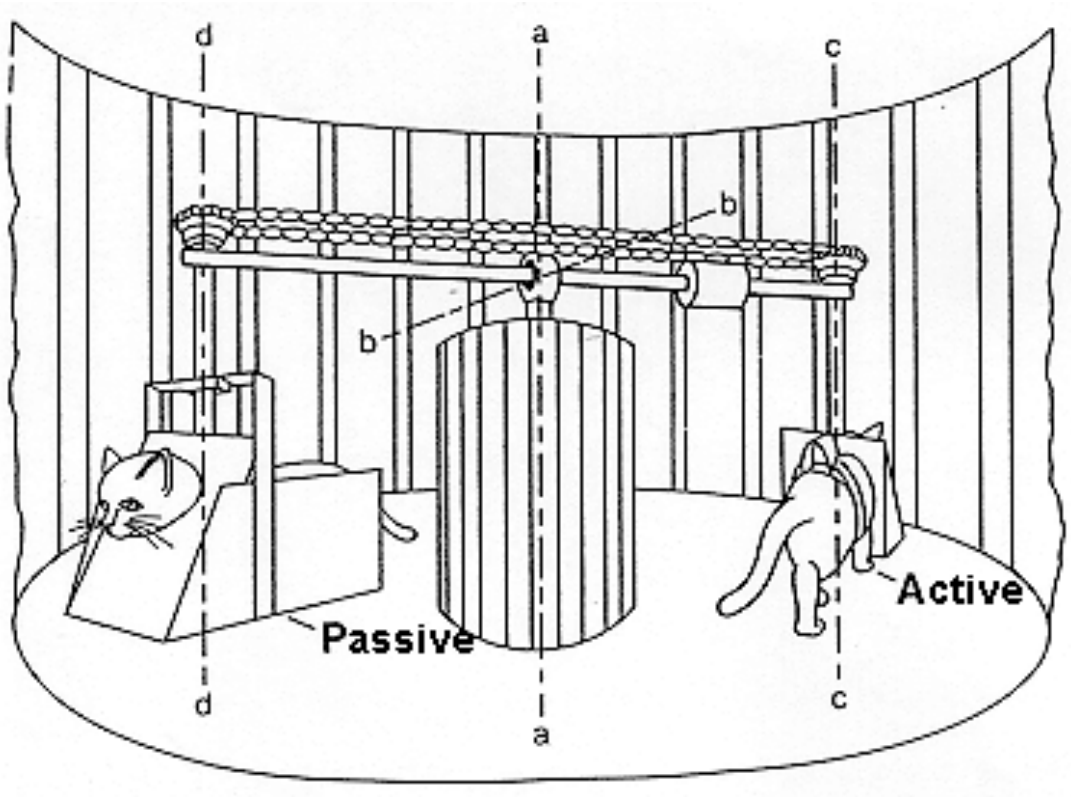
What do humans do?



Ostrovsky et al. Visual Parsing After Recovery From Blindness. In *Psychological Science*, 2009.

The kitten carousel

- Both kittens see same visual input
- Active kitten learns well, passive kitten does not. Why?
  - Knowledge of motion?
  - Actively choosing action?
  - Paid more attention?



Held, R. and Hein A. (1963). Movement-produced stimulation in the development of visually guided behavior. Jouranal of Comparative and Physiological Psychology
56(5): 872-876.

Walk, Richard D., Jane D. Shepherd, and David R. Miller. "Attention and the depth perception of kittens." *Bulletin of the Psychonomic Society* 26.3 (1988): 248-251.

Ego-motion ↔ vision: view prediction

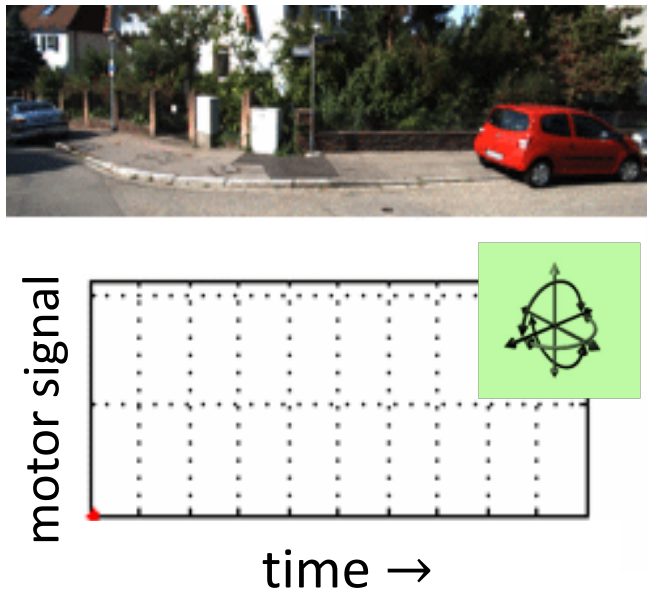
After moving:

Slide credit: Dinesh Jayaraman

Approach idea: Ego-motion equivariance

Training data

Unlabeled video + motor signals



Learn

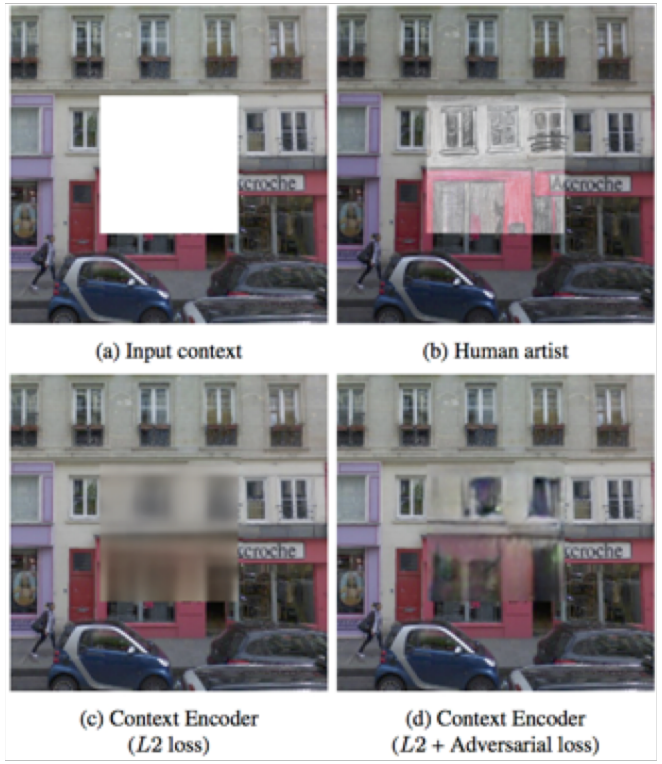
Equivariant embedding

organized by ego-motions

Pairs of frames related by similar ego-motion should be related by same feature transformation

Slide credit: Dinesh Jayaraman

Self supervision by reconstructing hidden data

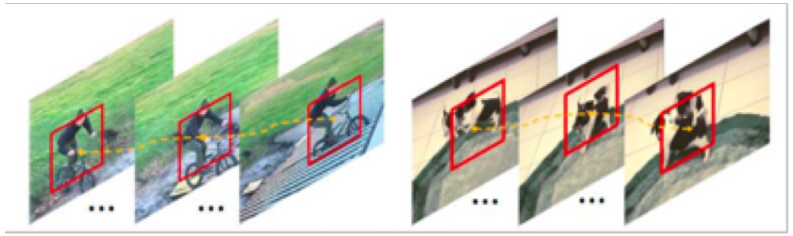


Pathak et al, CVPR 2016

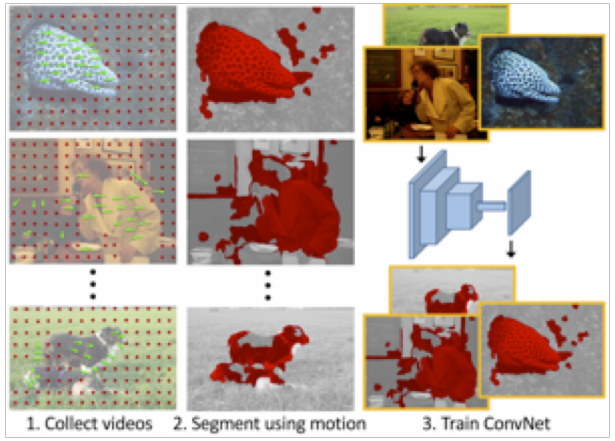
Doersch et al, ICCV 2015

Zhang et al, ECCV 2016

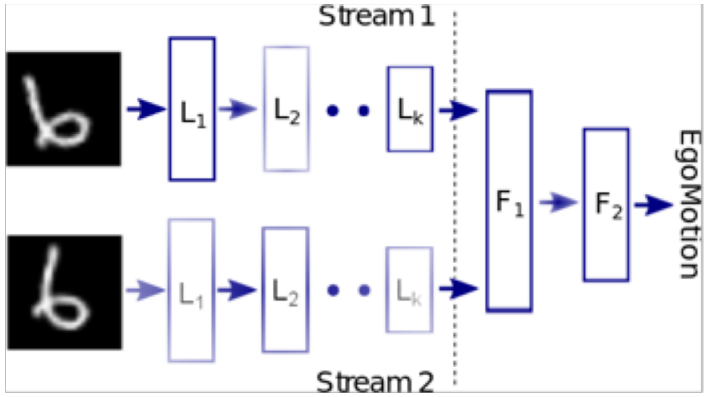
Self supervision by using external information



Wang et al, ICCV 2015



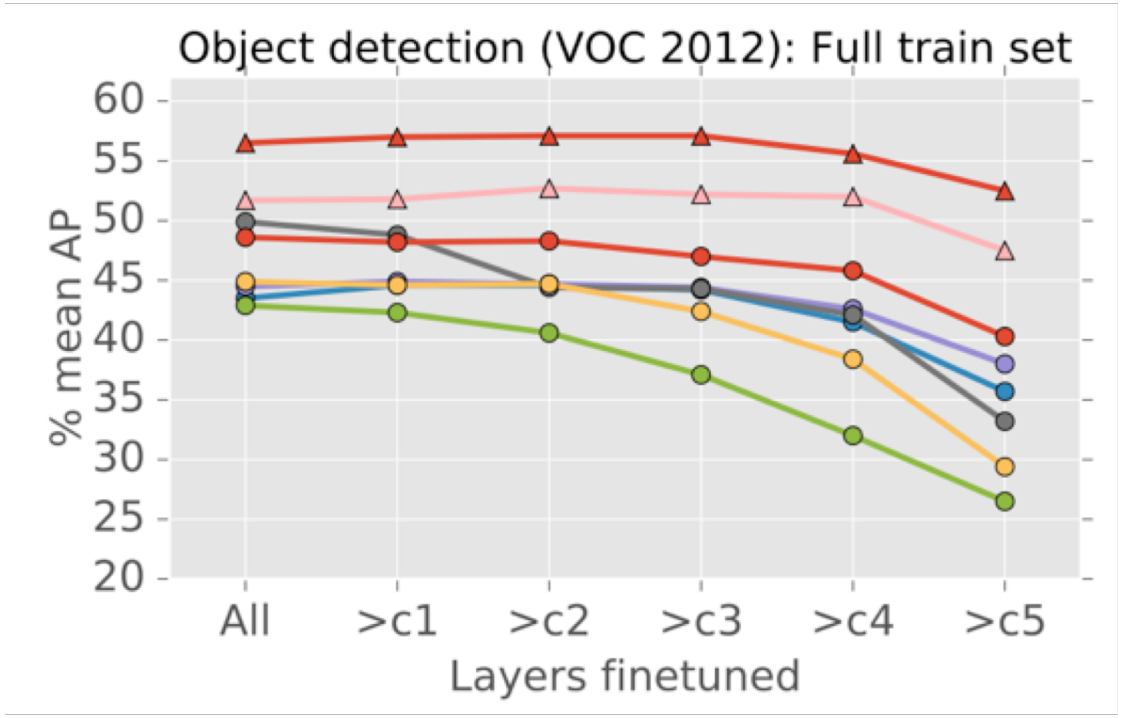
Pathak et al, CVPR 2017

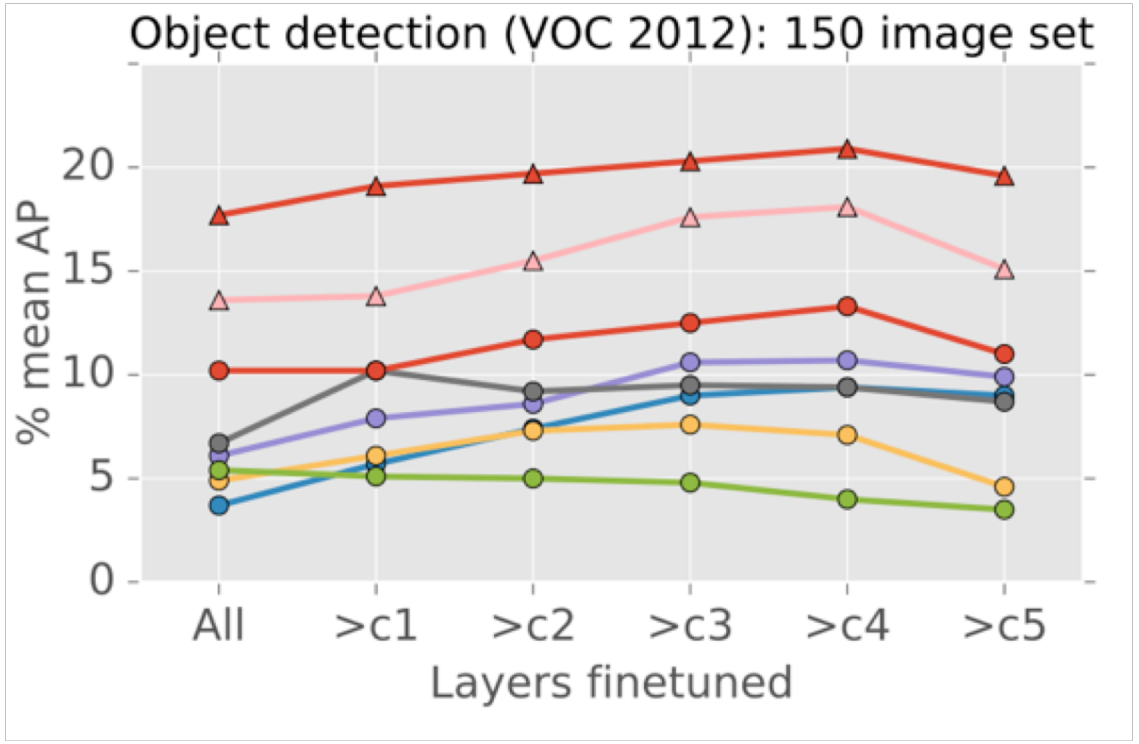


Agrawal et al, ICCV 2015.
Jayaraman and Grauman, ICCV 2015.

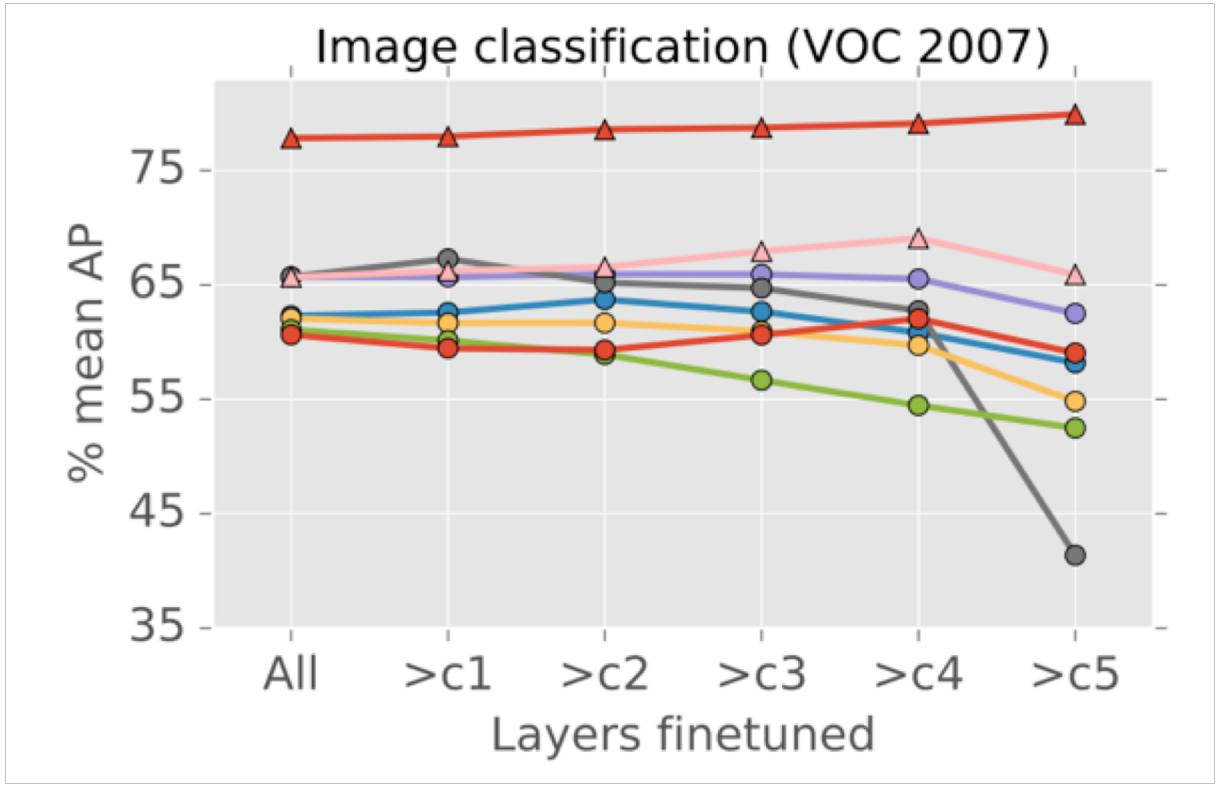
Owens et al, CVPR 2016

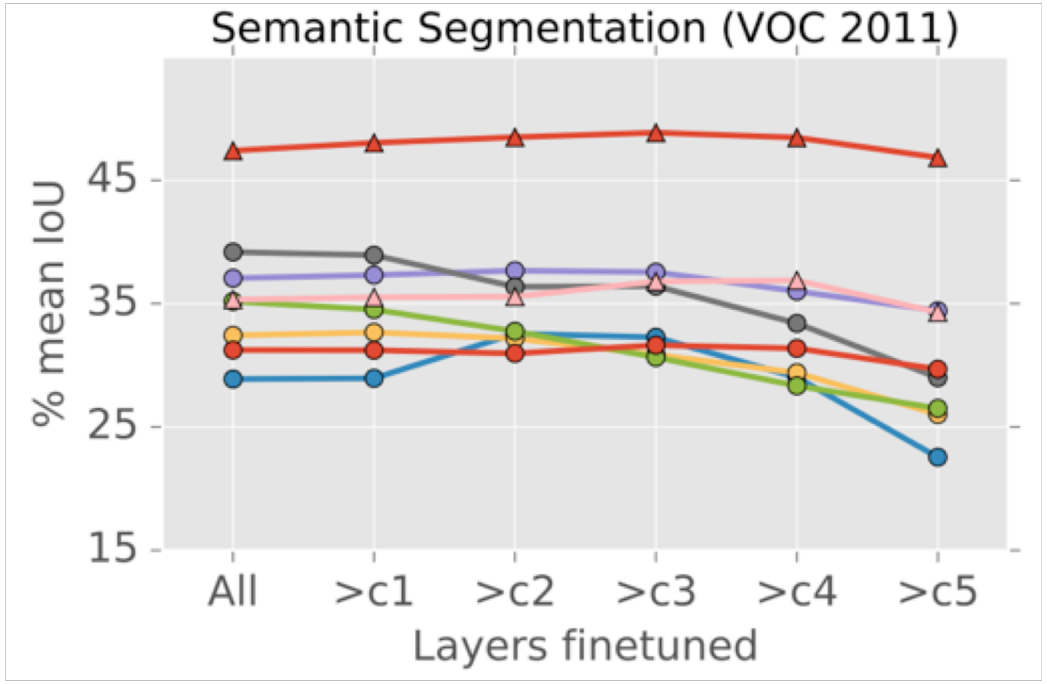
The unreasonable effectiveness of ImageNet





The unreasonable effectiveness of ImageNet





ImageNet [21]
Colorization [48]
BiGAN [8]
Sup. Masks (Ours)
Tracking-gray [43]
Context [6]
Sounds [30]
Motion Masks (Ours)

Why is ImageNet so good?

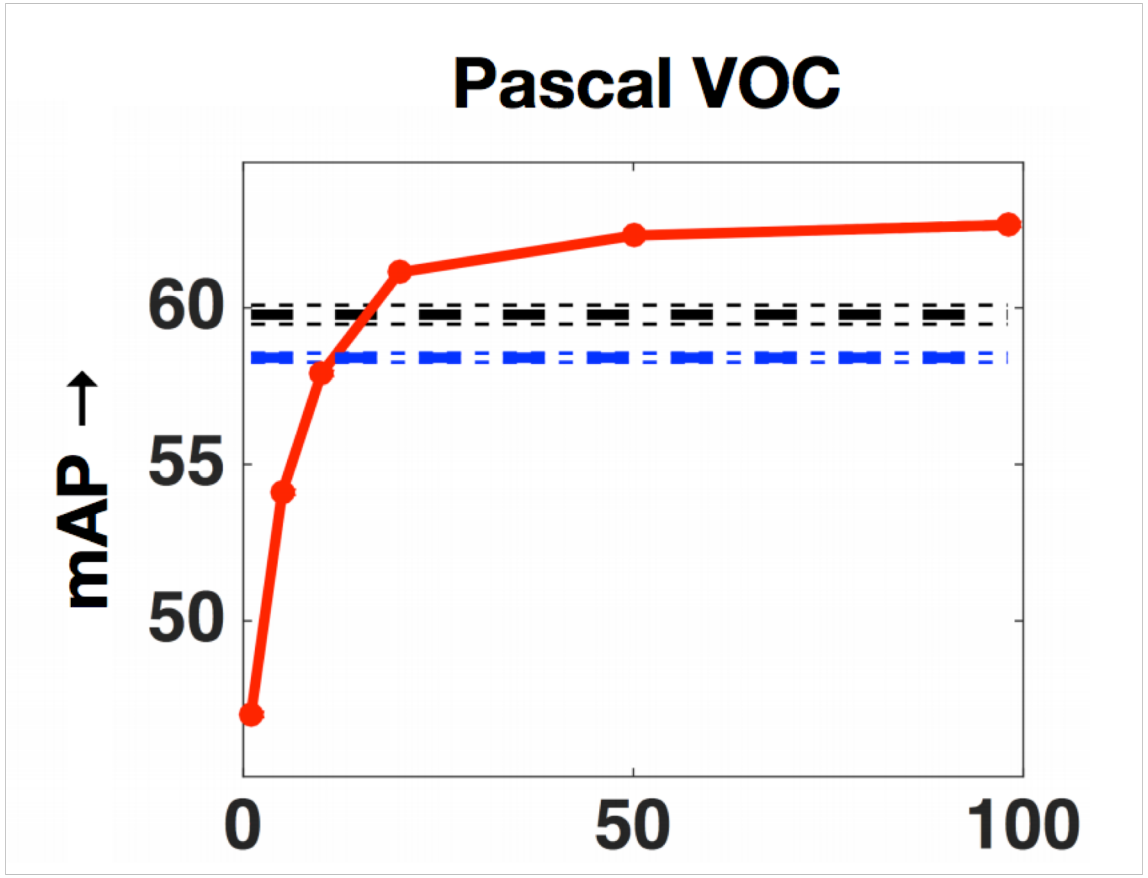
- Classification?
- Curated images?
- Many diverse classes?
- Related to target tasks?

Lessons from human cognition

- Babies experience of the world is profoundly **multi-modal**
- Babies develop **incrementally**, and they are not smart at the start.
- Babies live in a **physical world**, full of rich regularities that organize perception, action, and ultimately thought.
- Babies **explore** – they move and act in highly variable and playful ways that are not goal-oriented and are seemingly random.
- Babies act and learn in a **social world** in which more mature partners guide learning and add supporting structures to that learning
- Babies learn a **language**, a shared communicative system that is symbolic.

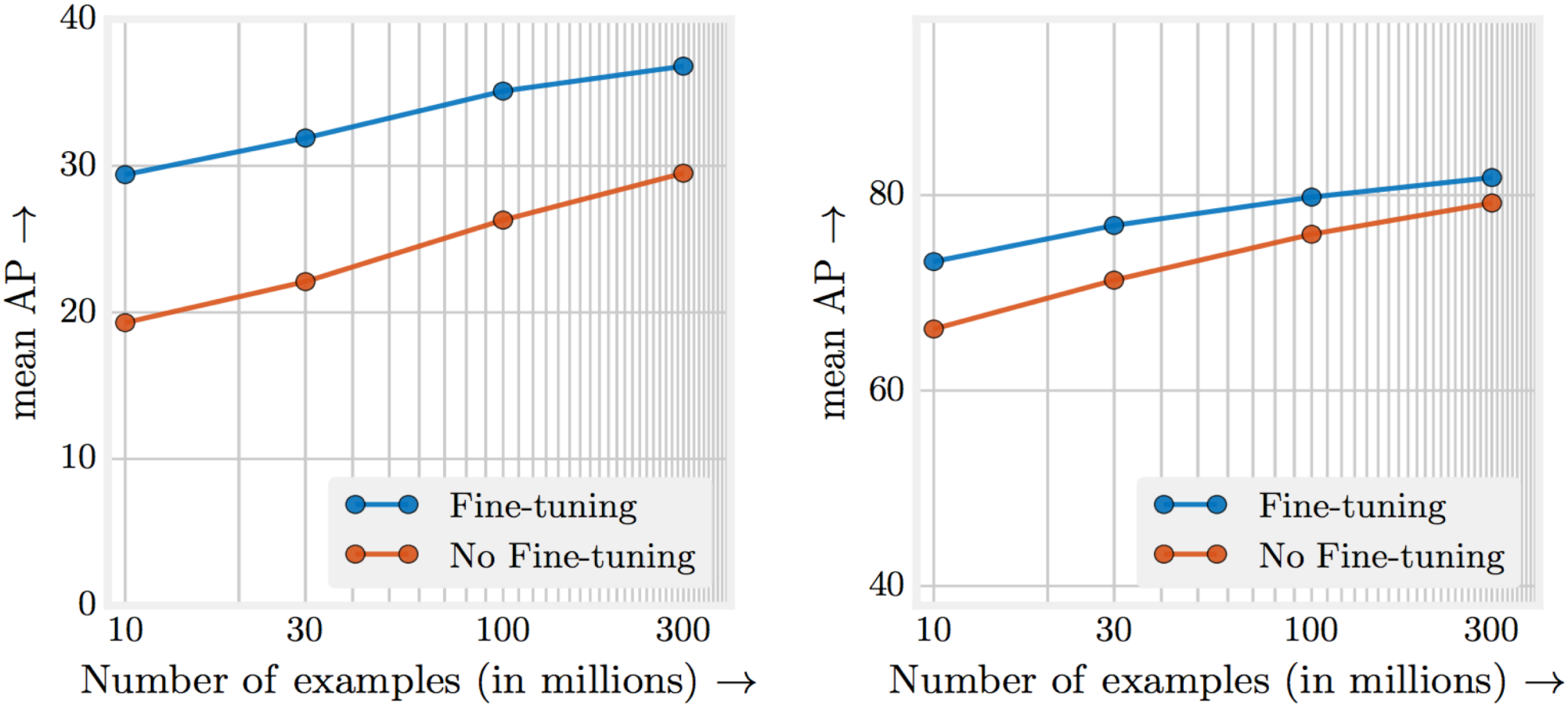
Smith, Linda, and Michael Gasser. "The development of embodied cognition: Six lessons from babies." *Artificial life* 11.1-2 (2005): 13-29.

Weak supervision



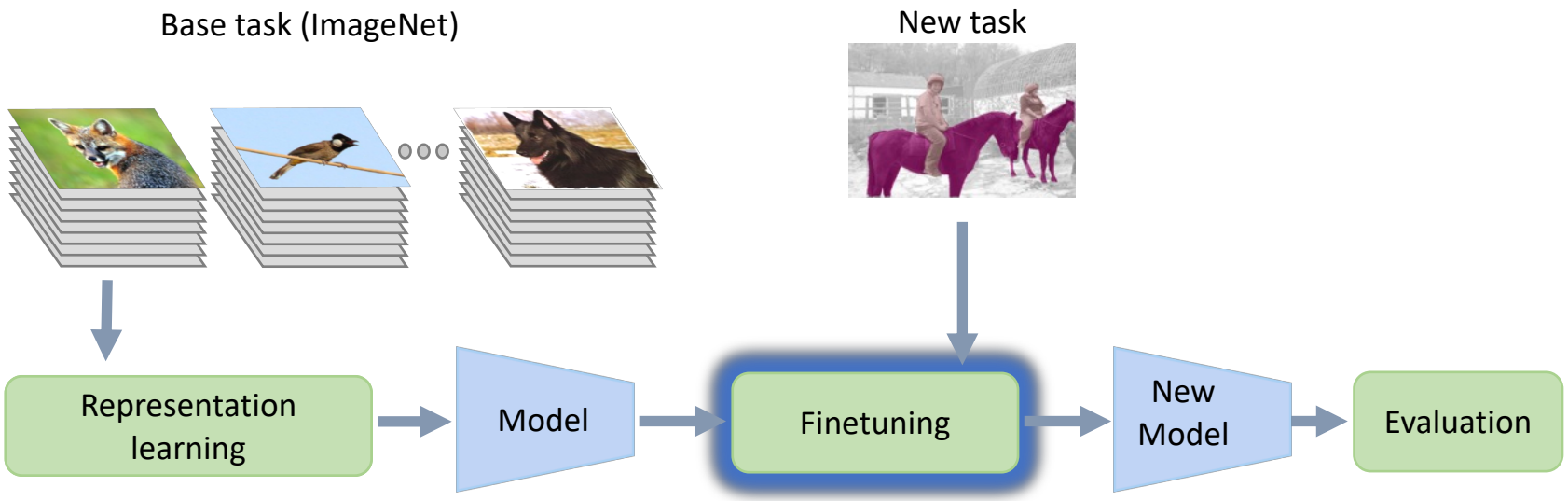
A. Joulin*, L.J.P. van der Maaten*, A. Jabri, and N. Vasilache (*both authors contributed equally). **Learning Visual Features from Large Weakly Supervised Data**. In *ECCV, 2016.*

Beating supervised learning with tons and tons of data



Revisiting Unreasonable Effectiveness of Data in Deep Learning Era. Chen Sun, Abhinav Srivastava, Saurabh Singh, Abhinav Gupta. In *ICCV,* 2017.

Two phases of supervision



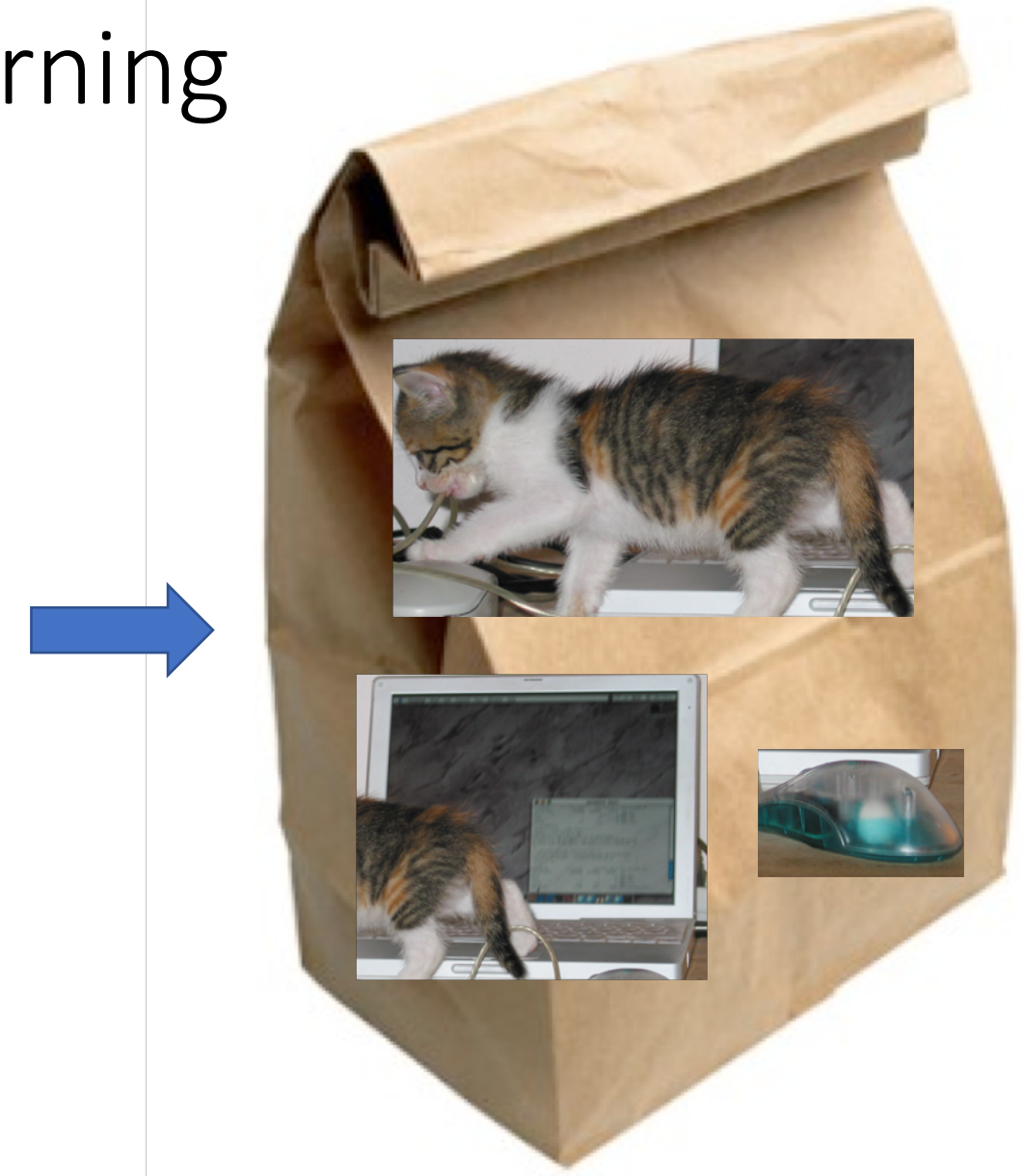
Different forms of reduced supervision

- Weakly supervised
  - Use less rich annotation / noisy annotation
- Semi supervised
  - Use a few labeled images and a bank of unlabeled images
- Few-shot
  - Use a few labeled images
- Zero-shot
  - Use no labeled images but side-information

Weak supervision for detection

- Can we learn object detection / semantic segmentation with only image level labels?
- Idea: image label = “cat” => somewhere in the image there is a cat

Multiple instance learning

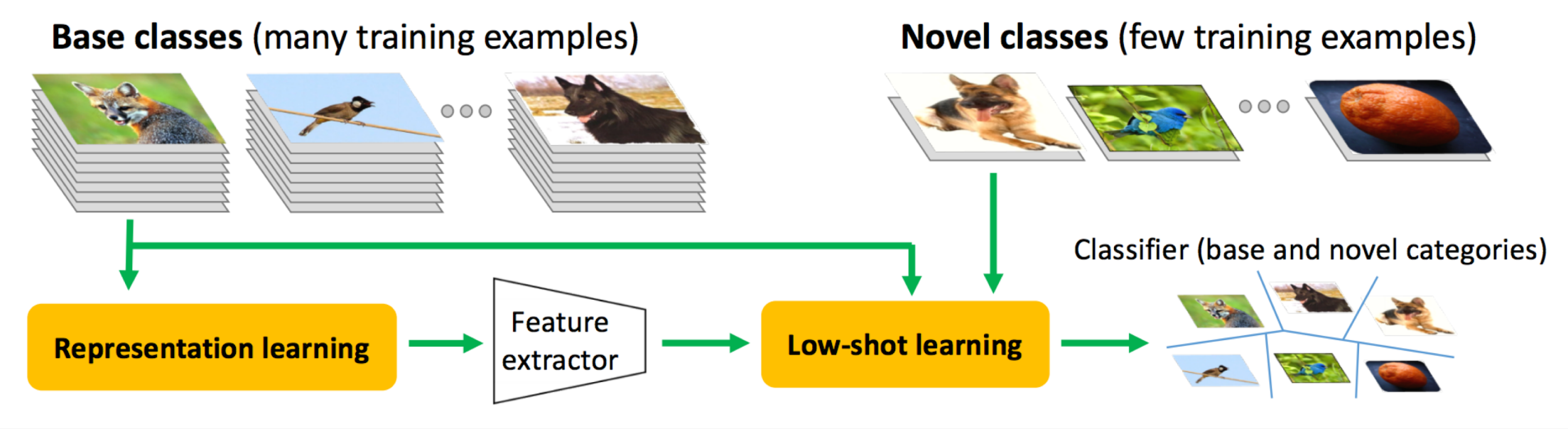


Multiple instance learning

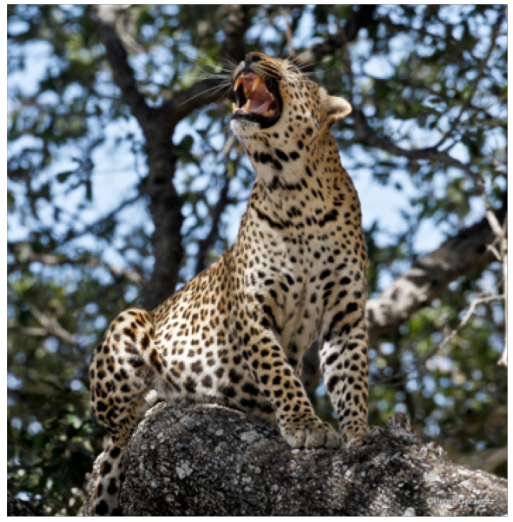
- Bag is negative if all instances are negative
- Bag is positive if one or more instances are positive

$$p_i = \max_j p_{ij} \qquad\qquad p_i = 1 - \prod_j (1 - p_{ij})$$

Few-shot learning



The challenge: Intra-class variation



“Train set”

Philippine Tarsier

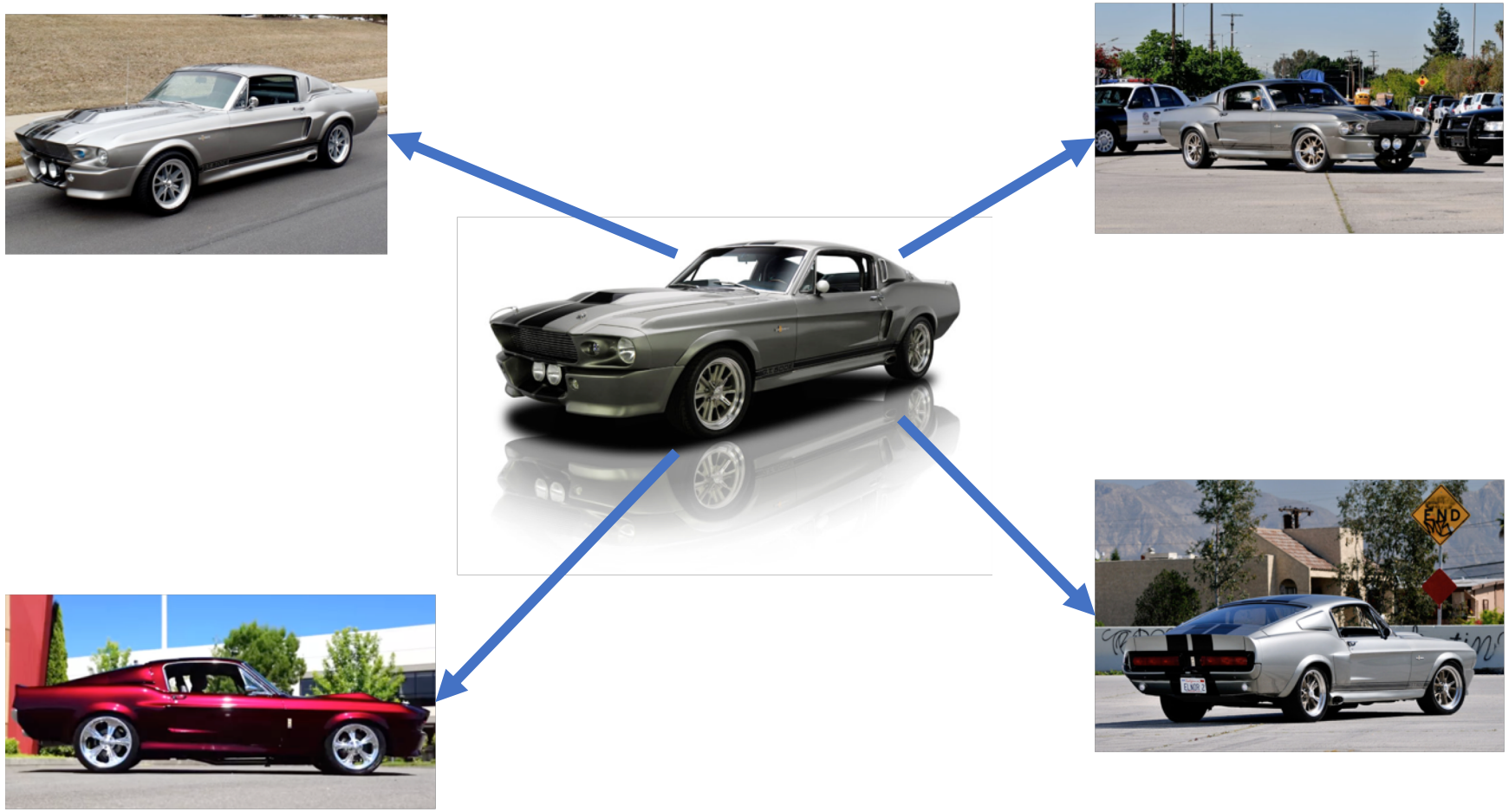
“Test set”

Philippine Tarsier

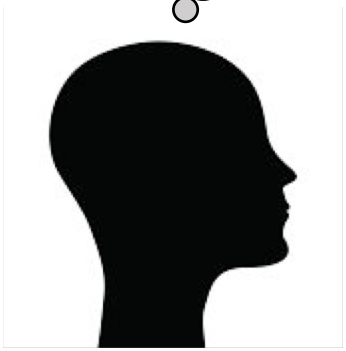
Mouse lemur

Beaver

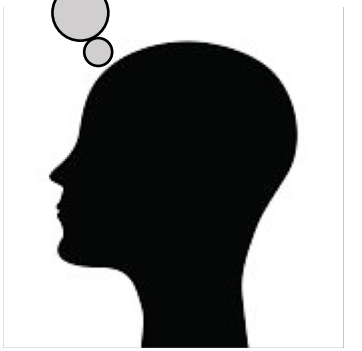
Key cue: shared modes of variation



How do humans do this?

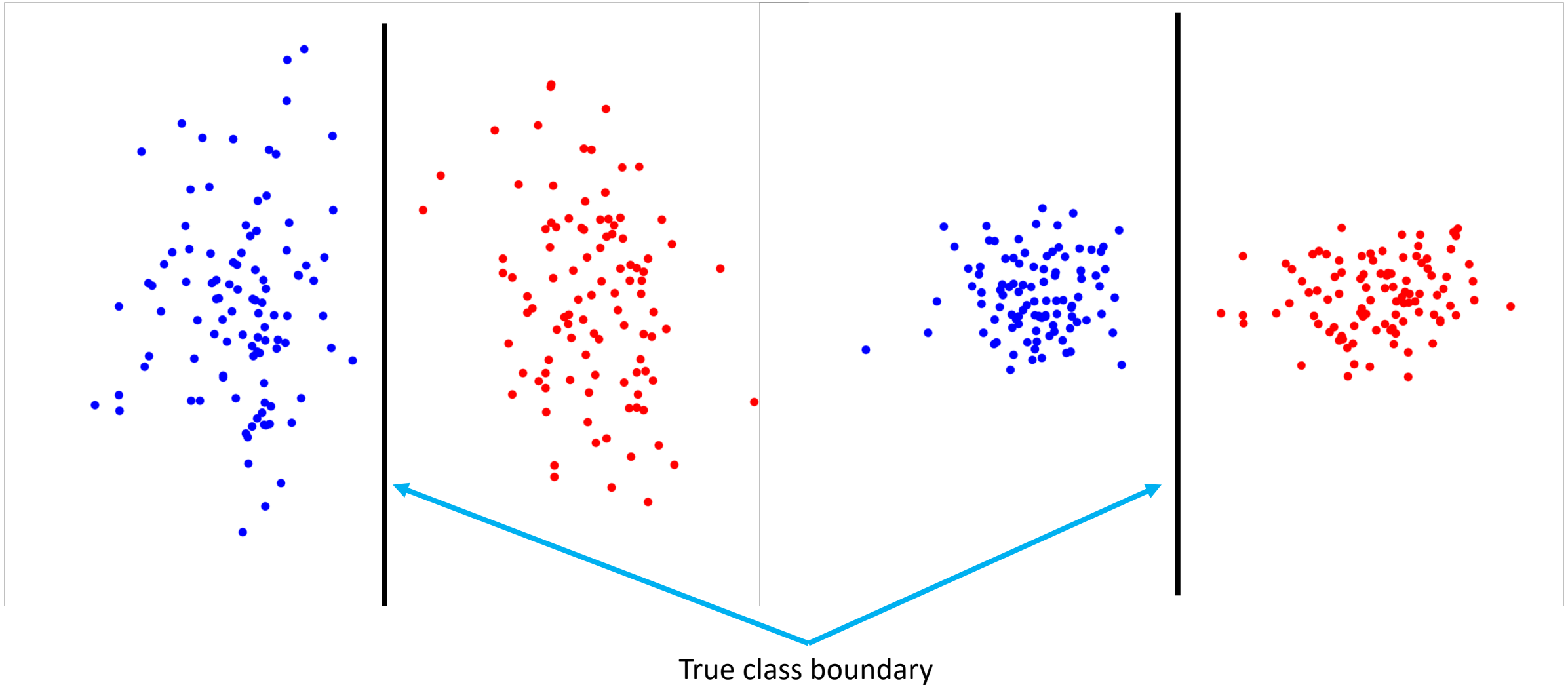


More invariant representations

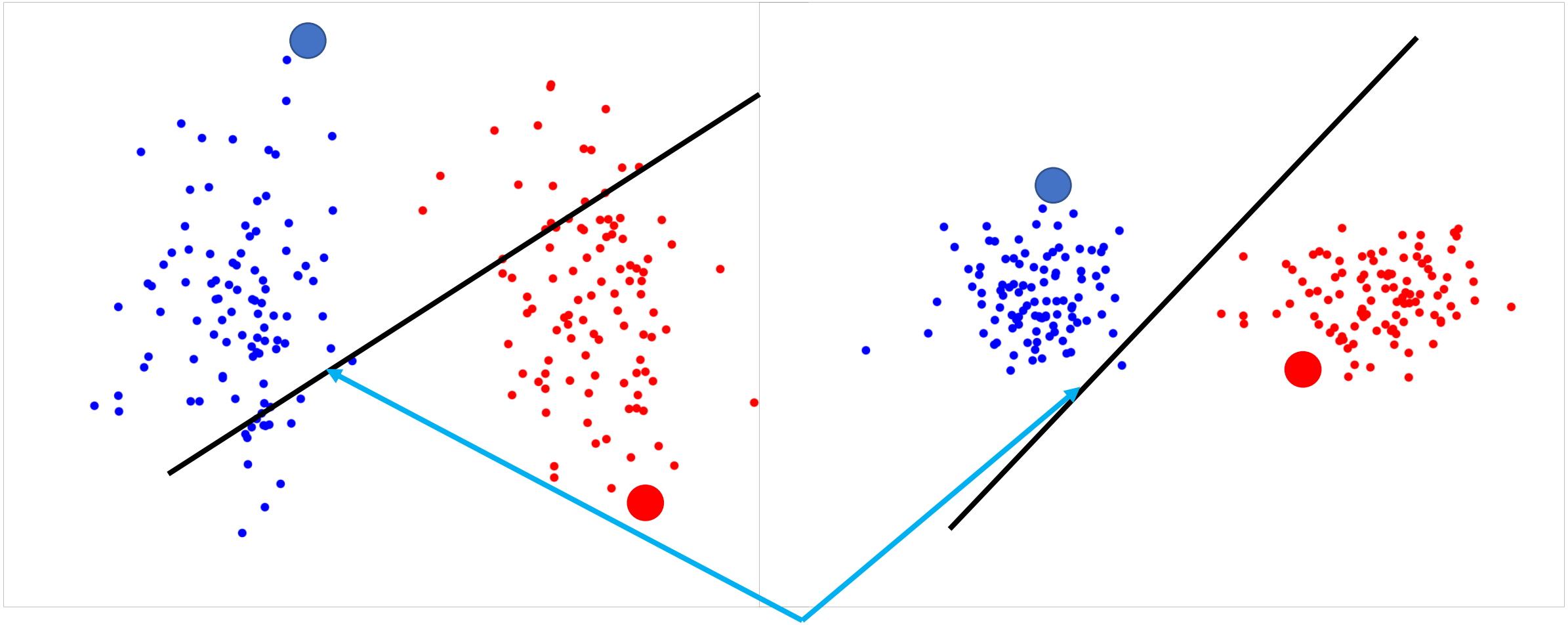


Inductive biases during learning

Better representations: metric learning



Better representations: metric learning



"One-shot" class boundary

Metric learning

- Pull same-class pairs closer and different-class pairs apart
- Contrastive loss (DrLIM)
  - $= d(x, x')^2 \ \ if \ y = y'$
  - $= \ \max\left(0, m \ - d(x, x')\right)^2 \ if \ y \neq y'$
- Triplet loss
  - $= \max(d(x, x_+) - d(x, x_-) + \gamma, 0)$

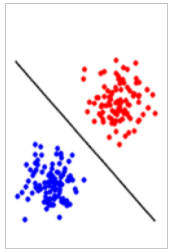
Dimensionality reduction by learning an invariant mapping. *Raia Hadsell, Sumit Chopra, Yann LeCun.* **Computer Vision and Pattern Recognition (CVPR), 2006**

Meta-learning

- Given:

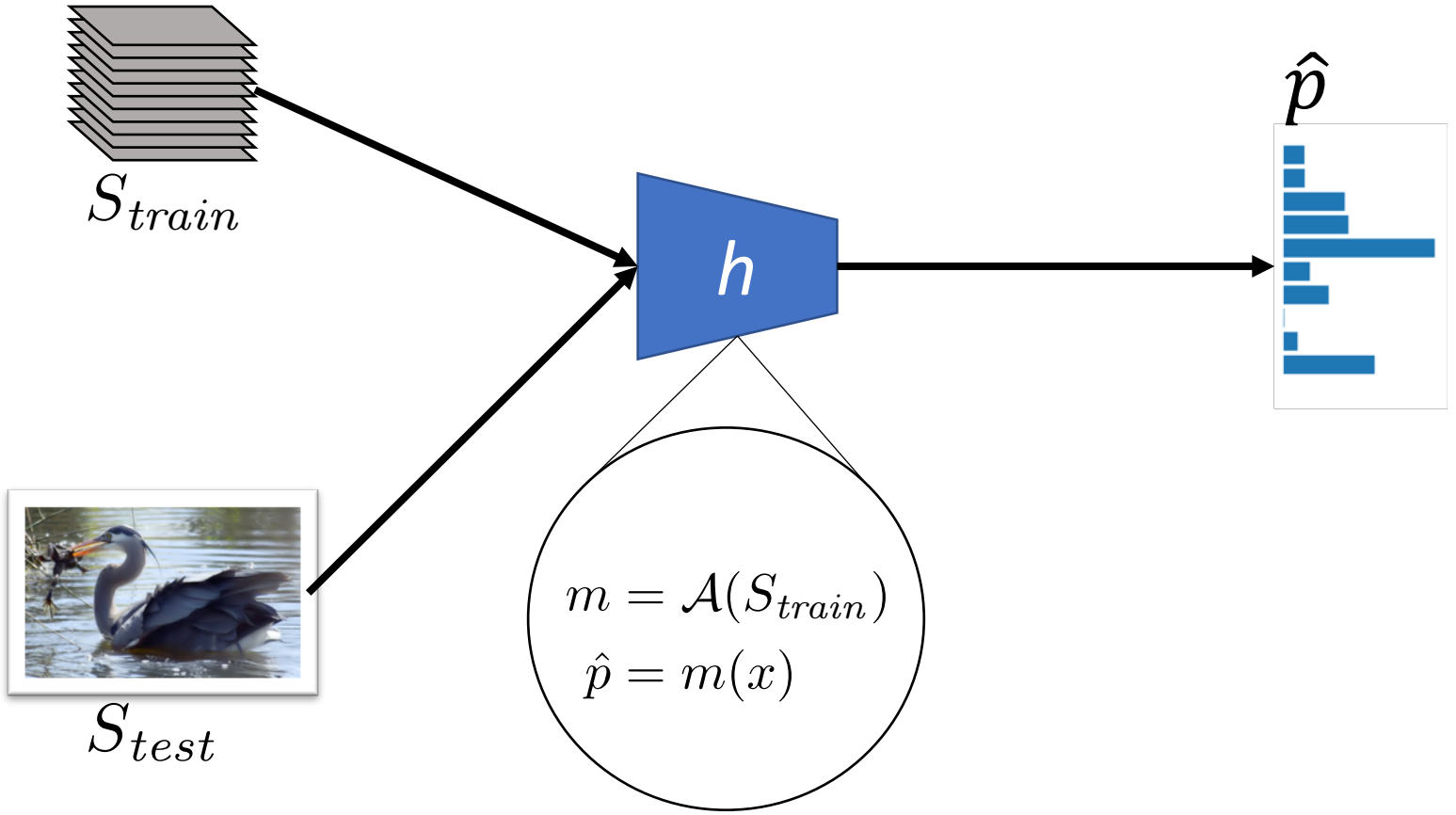
Small training set
(few training examples)

- Produce:

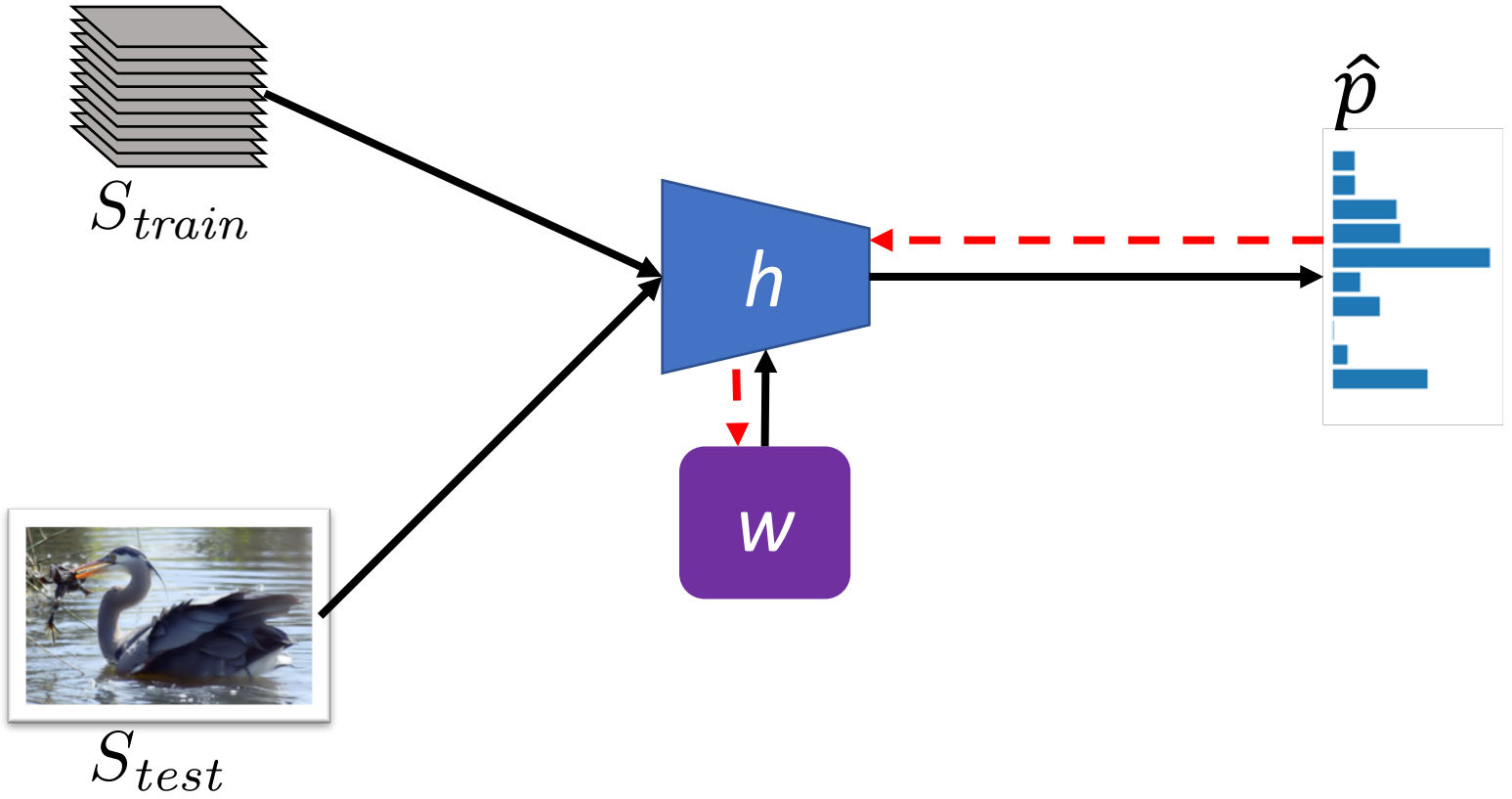


- Idea: Make this a learnable function!

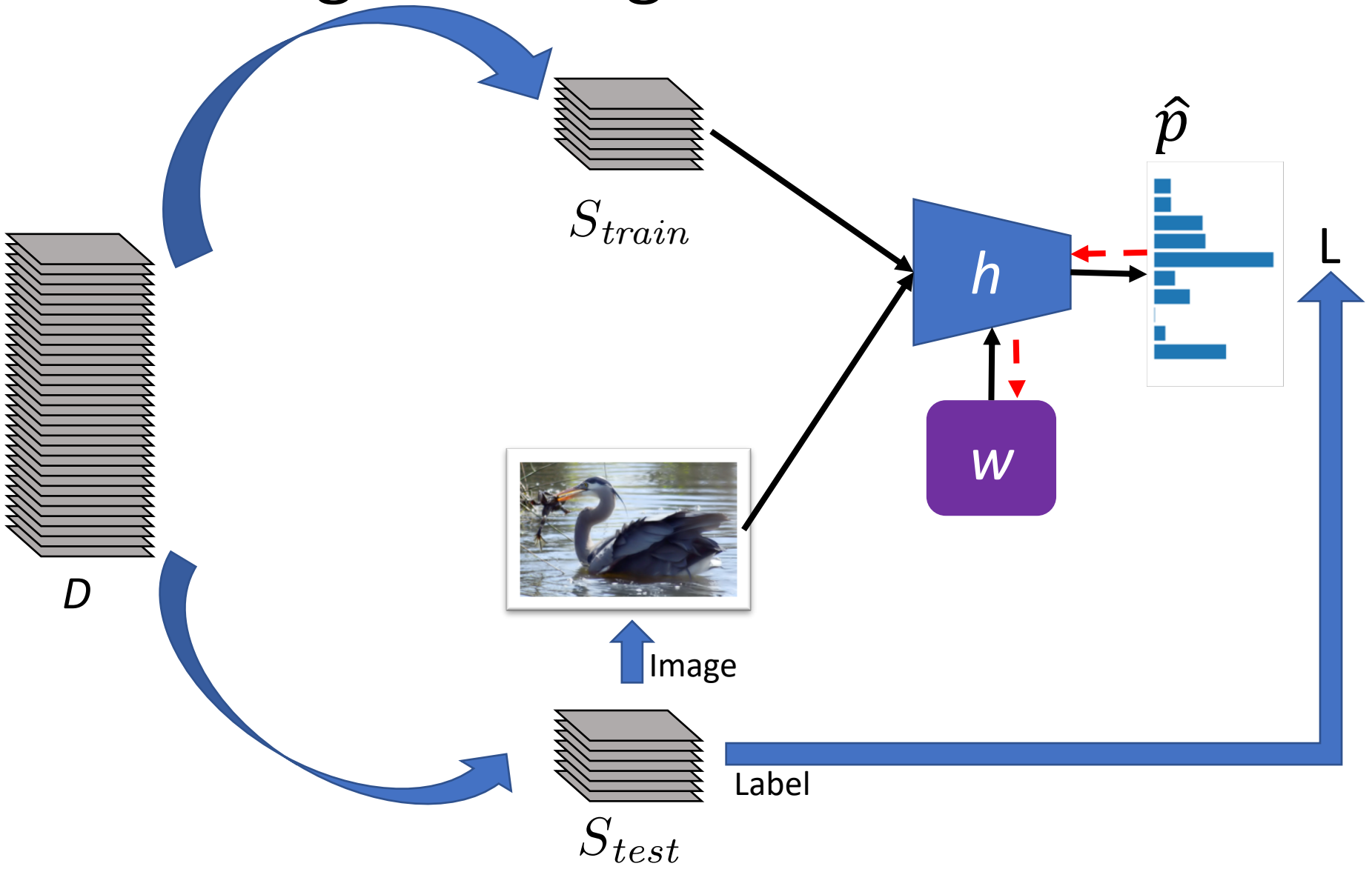
Meta-learning



Meta-learning



Meta-learning: training



An army of meta-learners

- Vinyals, Oriol, et al. "Matching networks for one shot learning." *NIPS*. 2016.
- Ravi, Sachin, and Hugo Larochelle. "Optimization as a model for few-shot learning." *ICLR,* 2017.
- Snell, Jake, Kevin Swersky, and Richard Zemel. "Prototypical networks for few-shot learning." *NIPS*. 2017.
- Finn, Chelsea, Pieter Abbeel, and Sergey Levine. "Model-agnostic meta-learning for fast adaptation of deep networks." *ICML.* 2017.

Meta-learning: Prototypical Networks