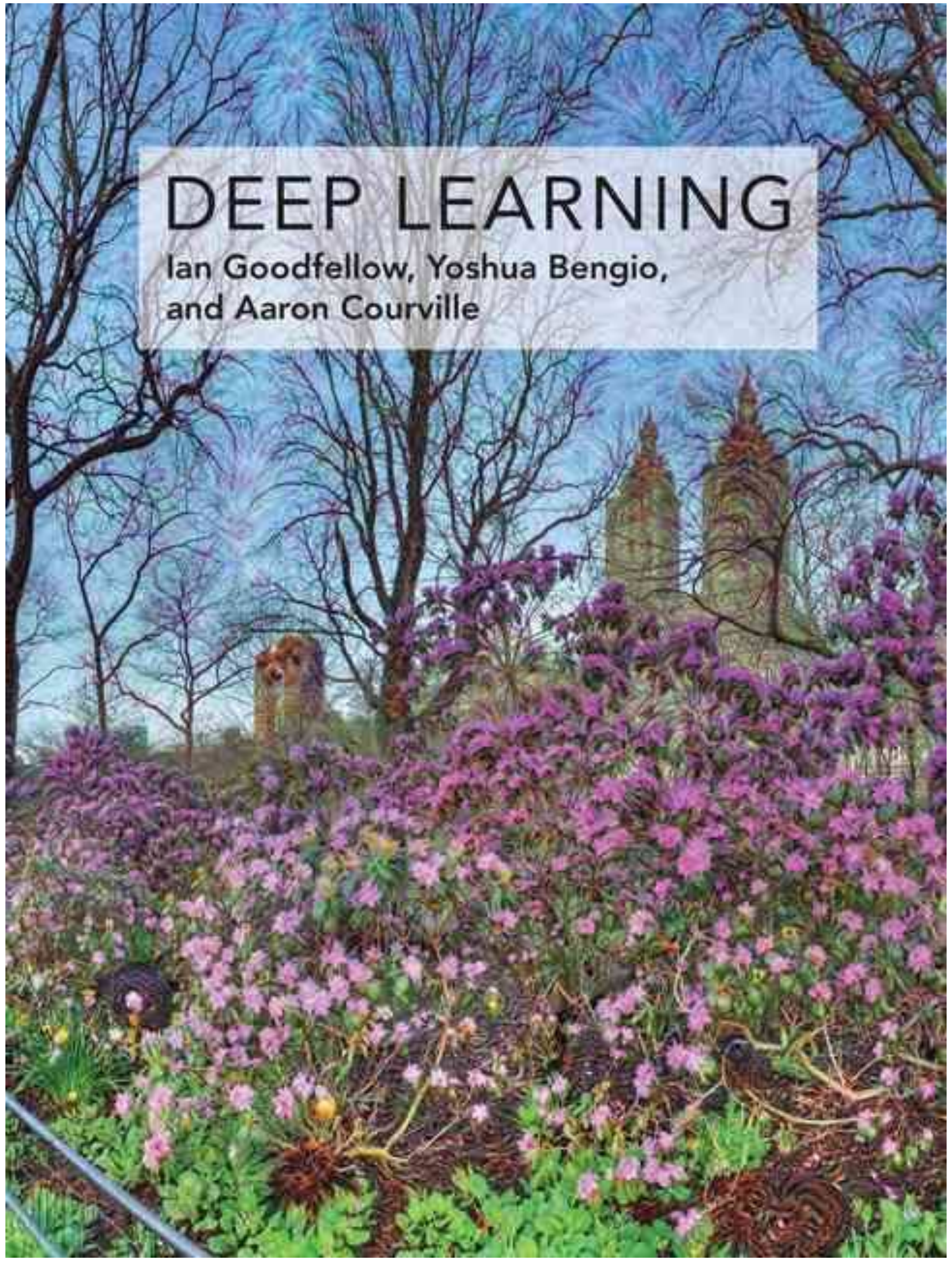
Visual representation learning

Bill Freeman, Antonio Torralba, Phillip Isola
6.819 / 6.869



http://www.deeplearningbook.org/

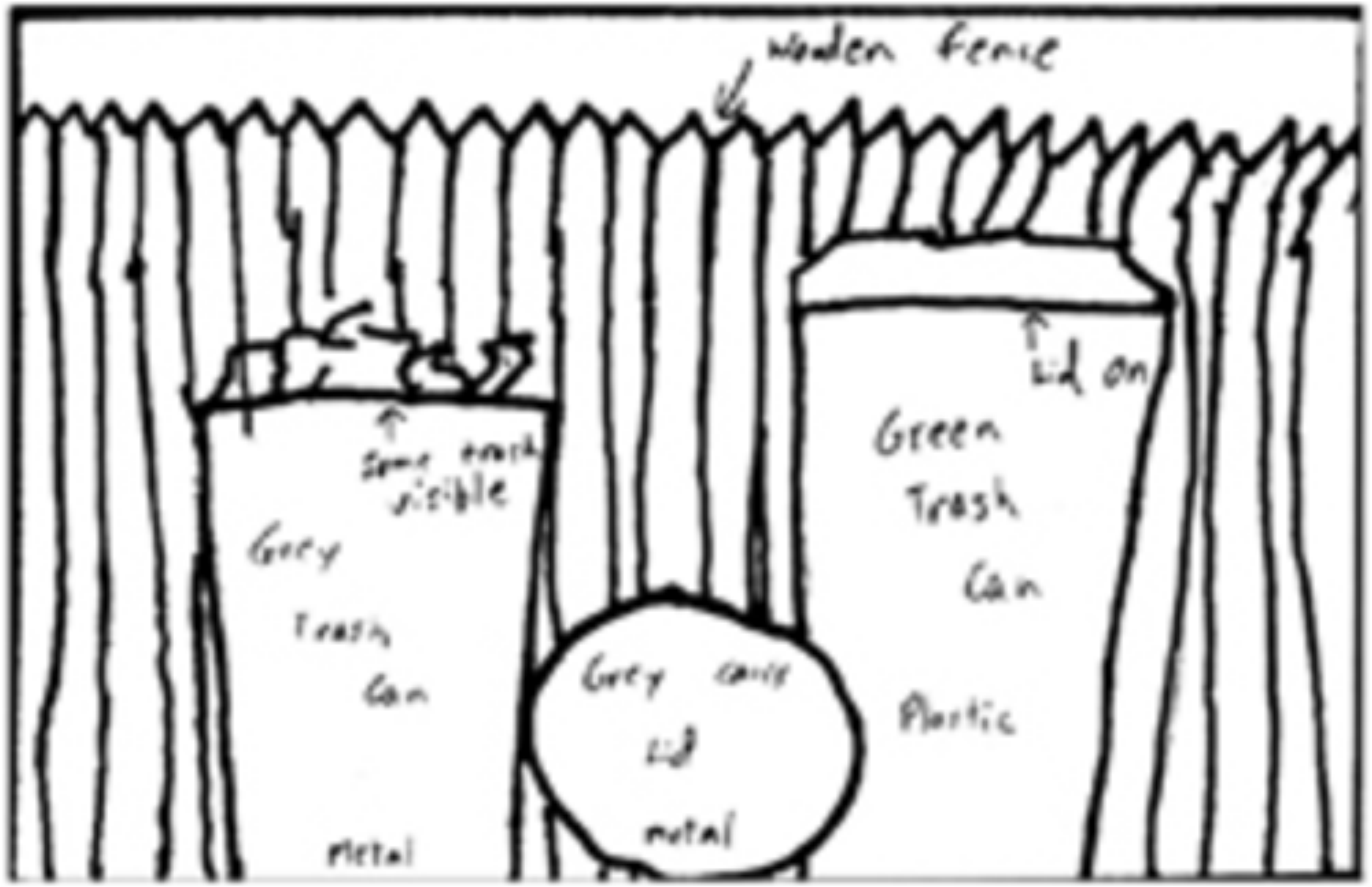
By Ian Goodfellow, Yoshua Bengio and Aaron Courville

November 2016

Today: parts of chapters 14 and 15 (but this lecture is mostly a departure from the book)

Observed image

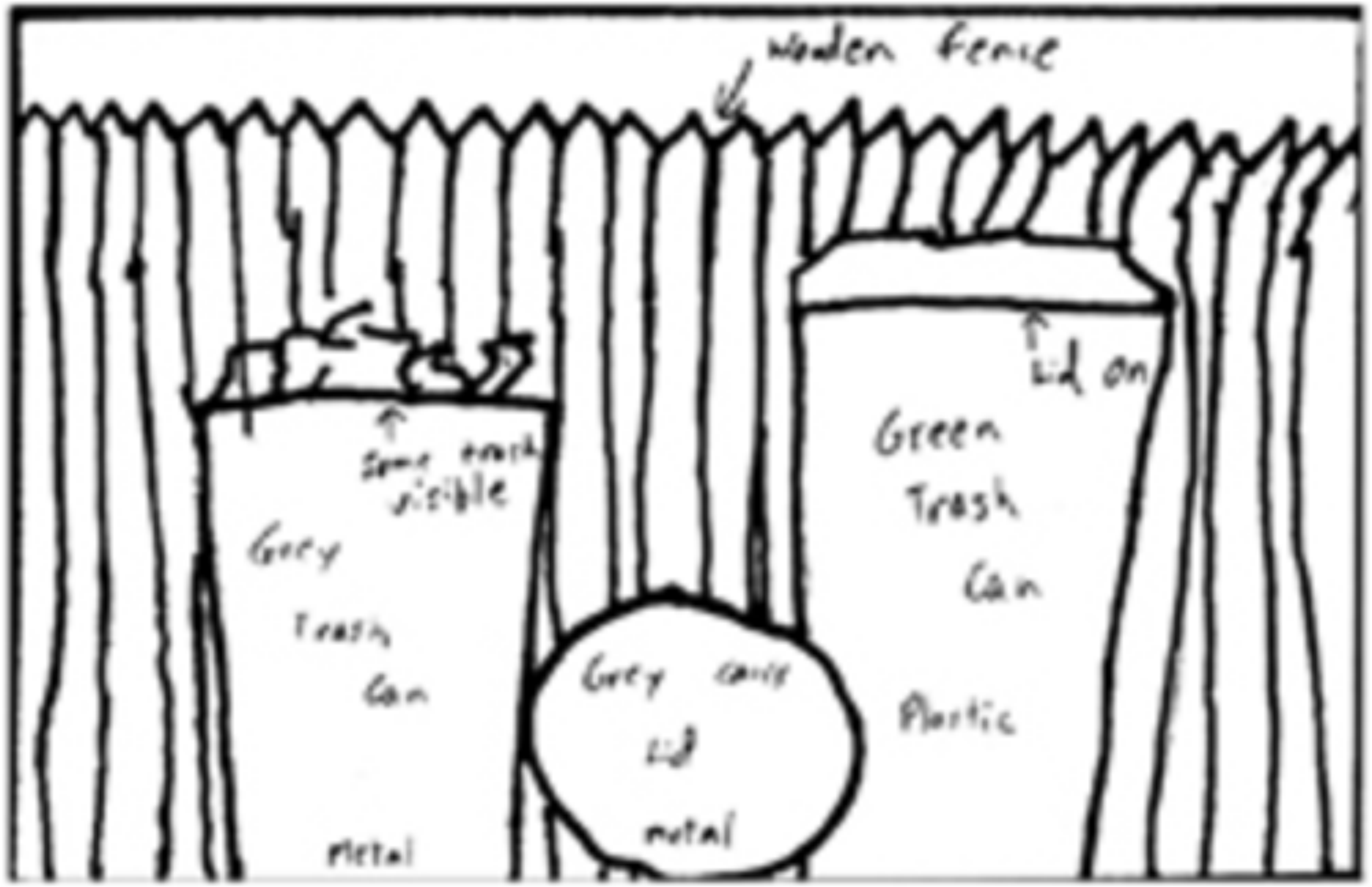
Drawn from memory



[Bartlett, 1932]
[Intraub & Richardson, 1989]

Observed image

Drawn from memory



[Bartlett, 1932]
[Intraub & Richardson, 1989]

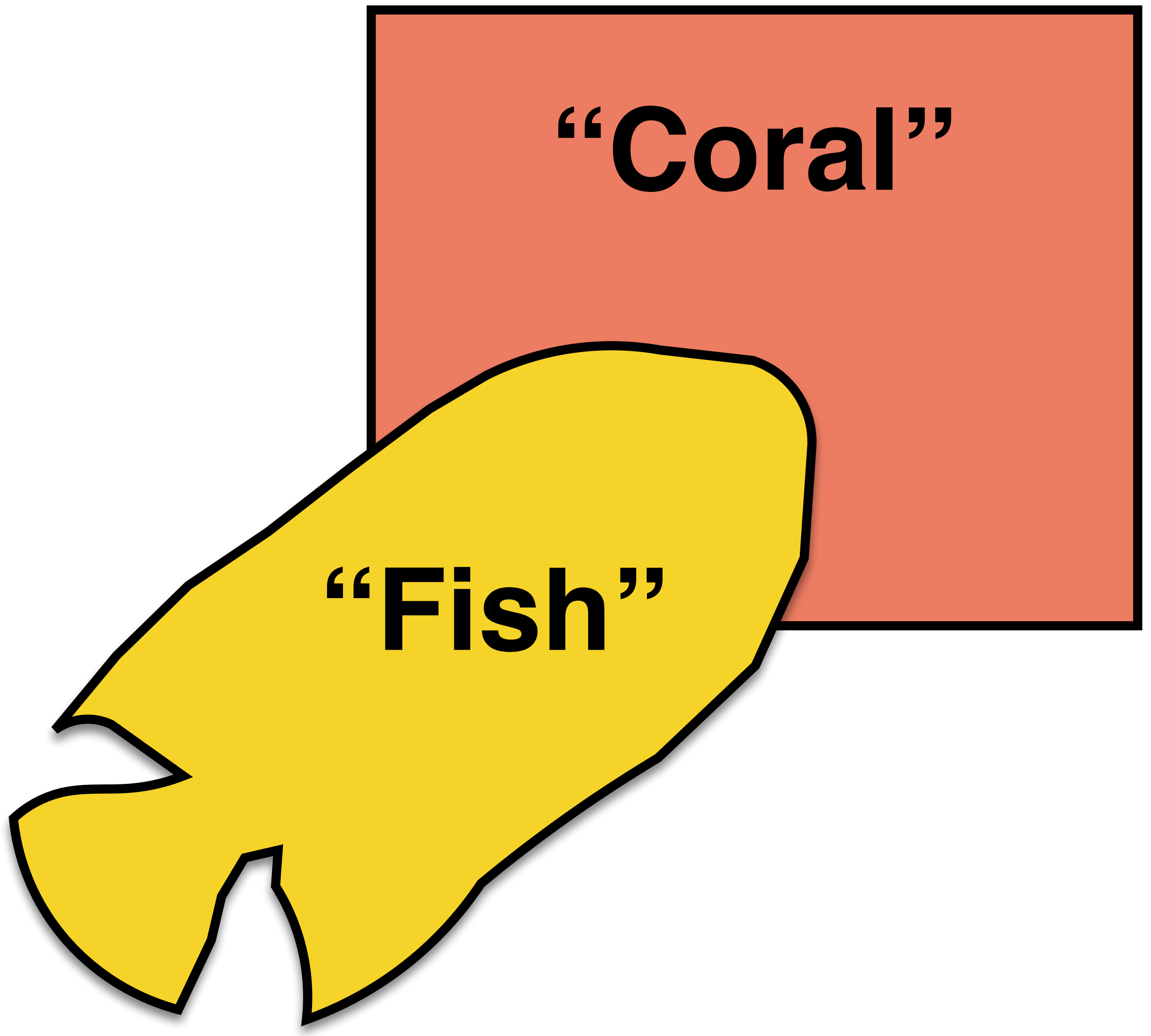
"I stand at the window and see a house, trees, sky. Theoretically I might say there were 327 brightnesses and nuances of colour. Do I have "327"? No. I have sky, house, and trees."

— Max Wertheimer, 1923

Representation learning

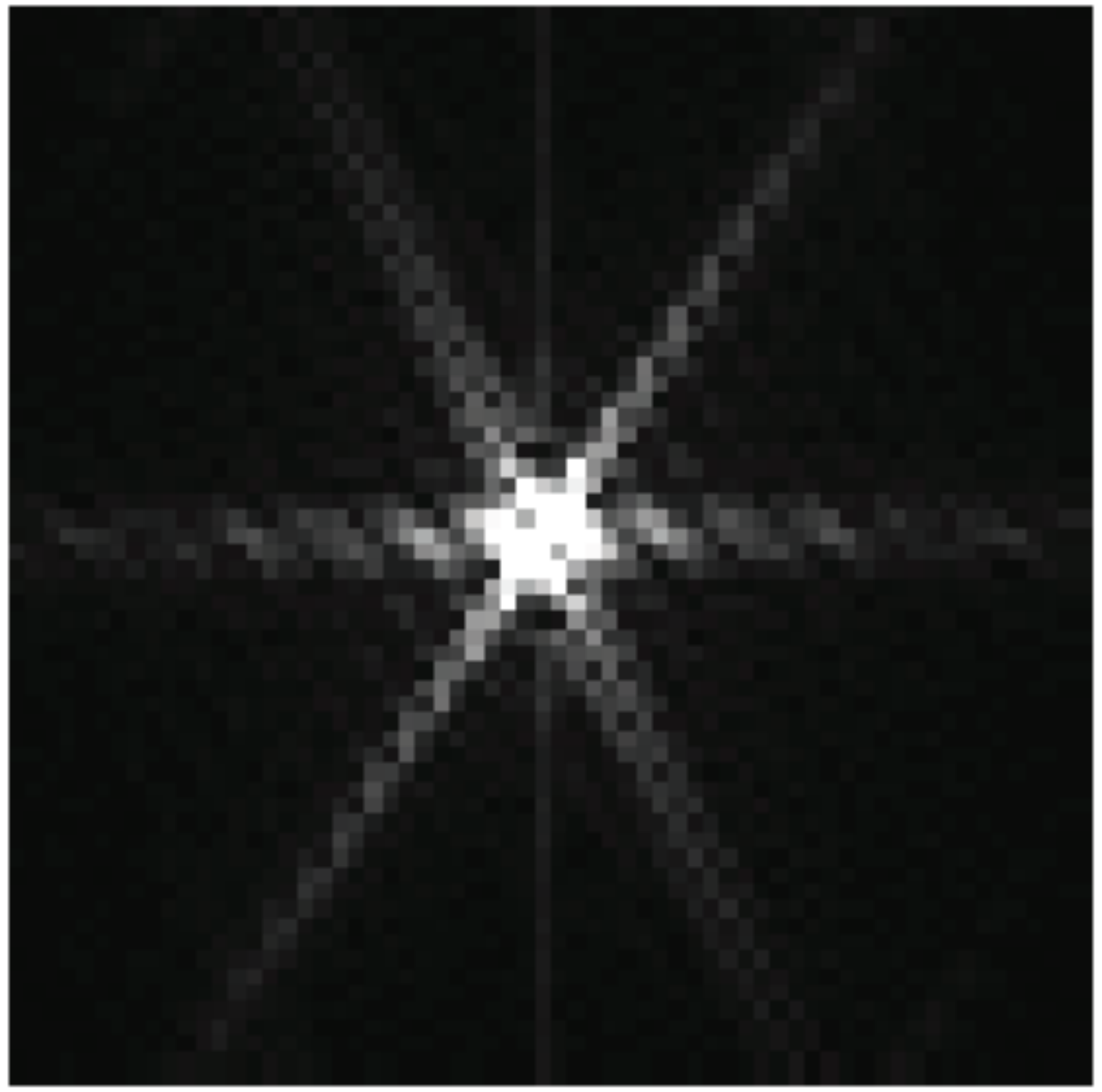
$\mathbf{x}$

Image



Compact mental representation

Representation learning

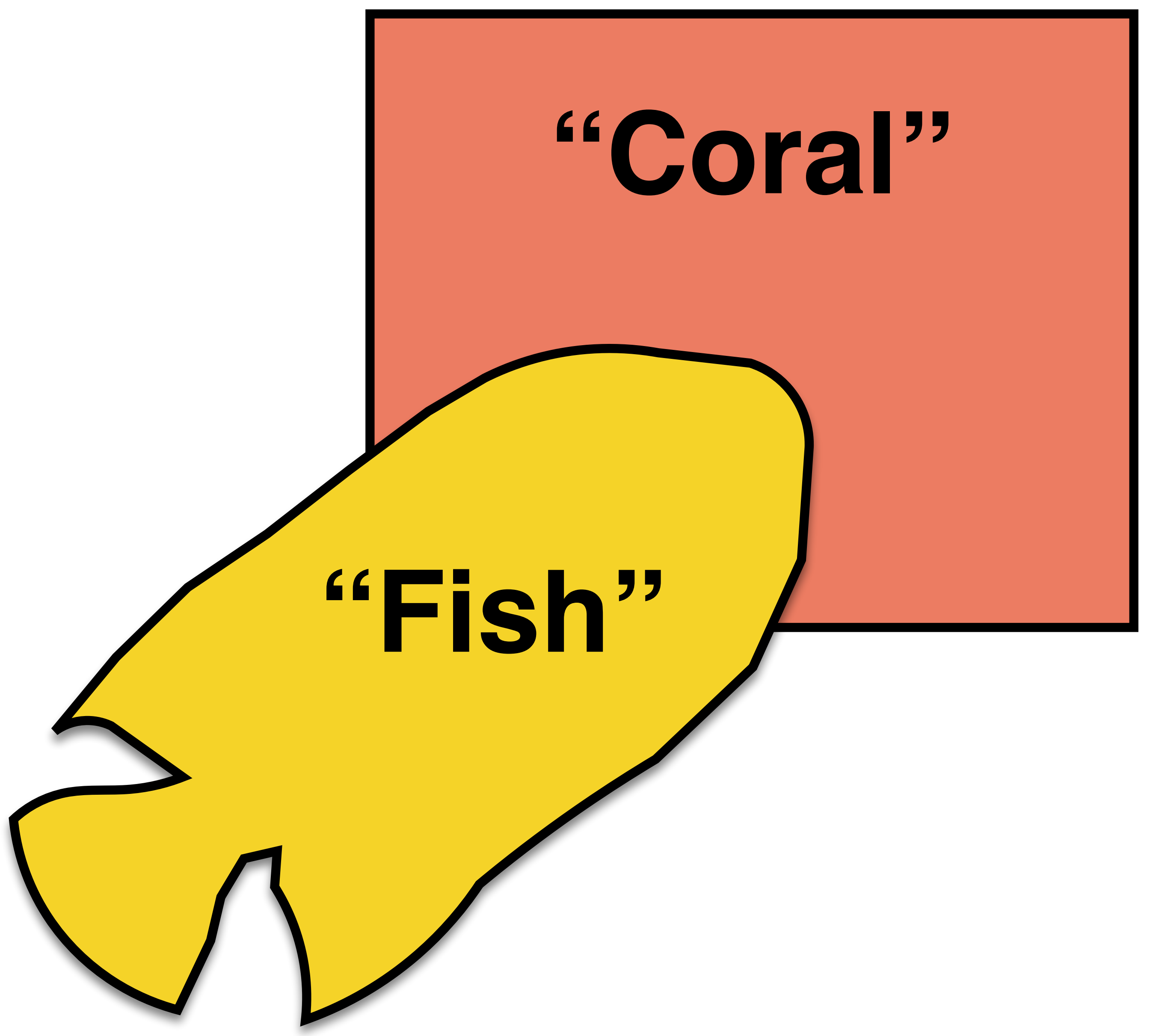


Convolution is pointwise multiplication in the frequency domain.

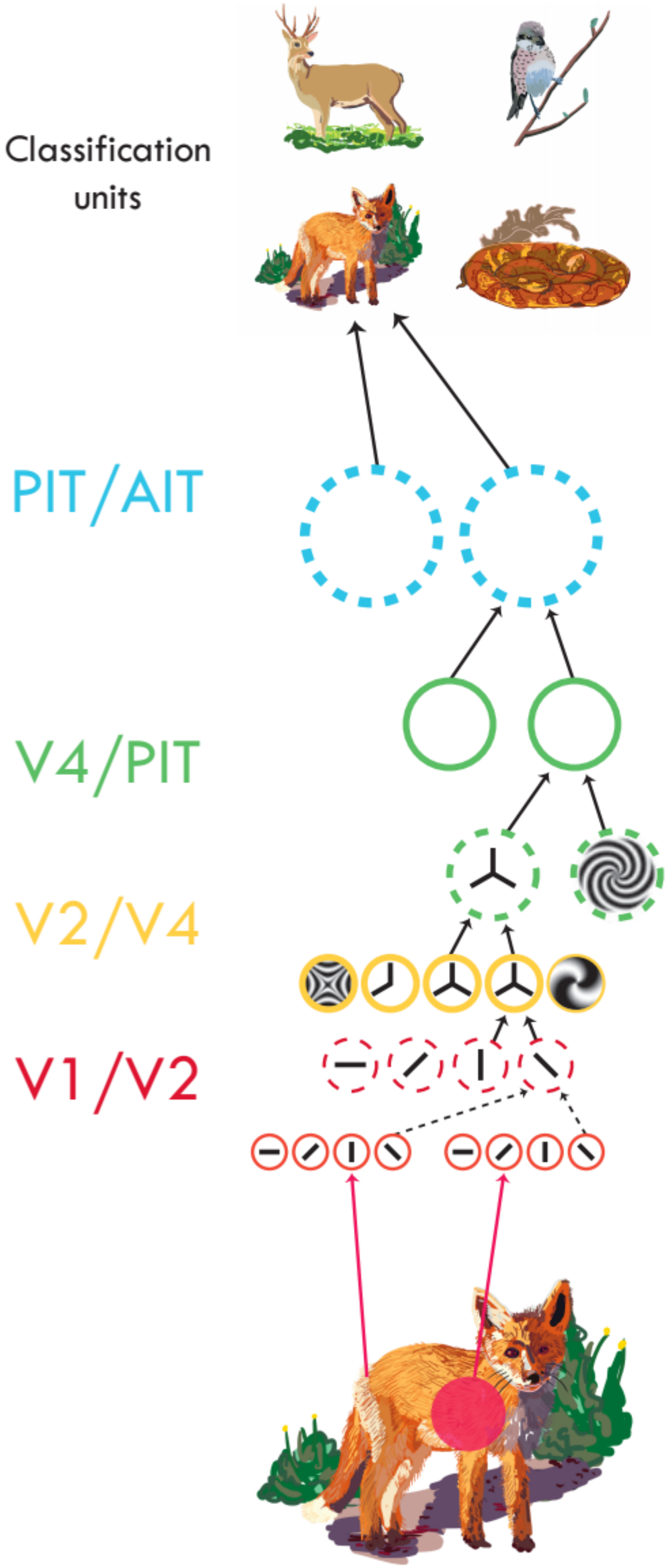
Representation learning

Good representations are:

1. Compact
2. Explanatory
3. Disentangled
4. Interpretable

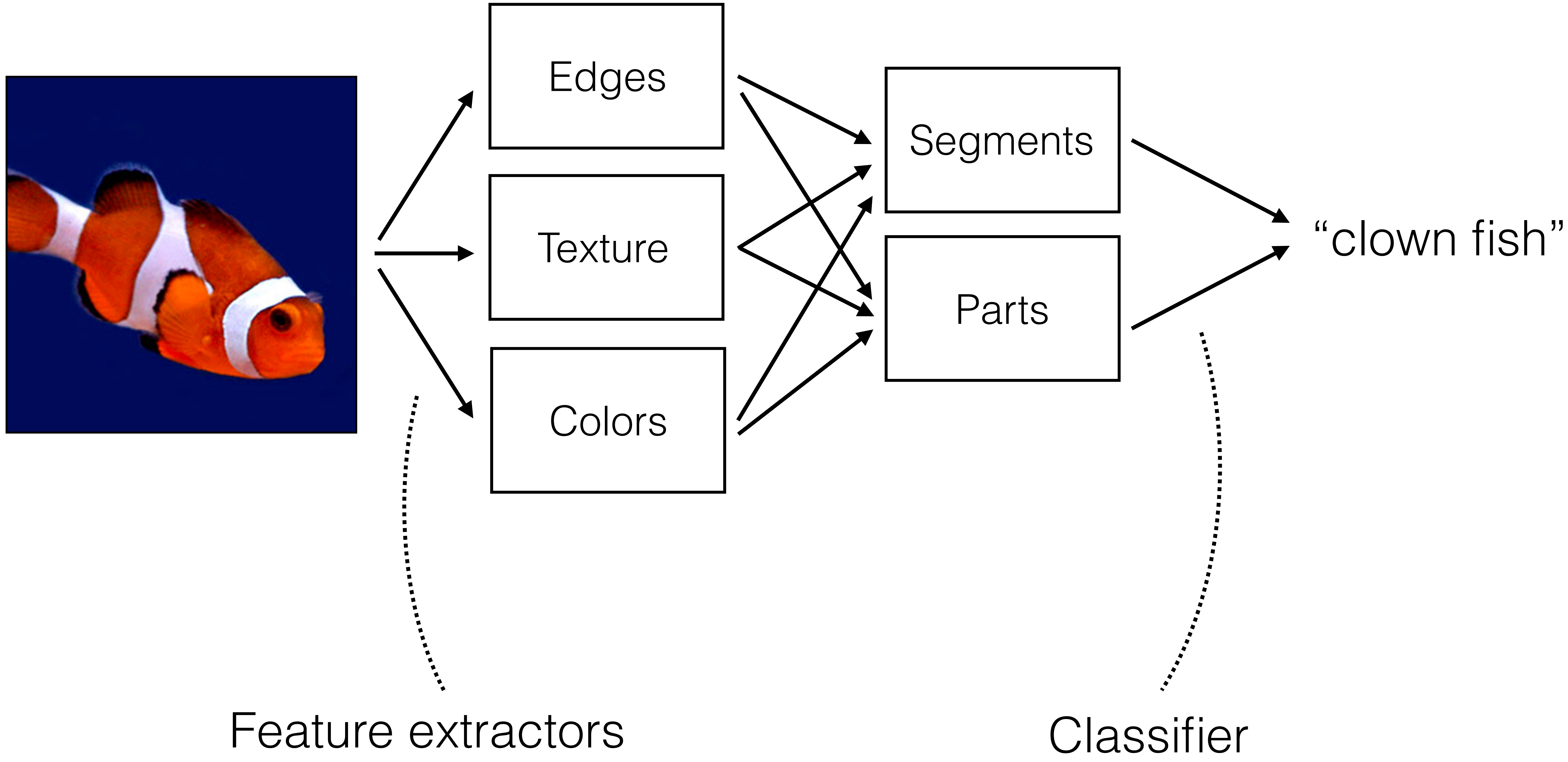


[See "Representation Learning", Bengio 2013, for more commentary]

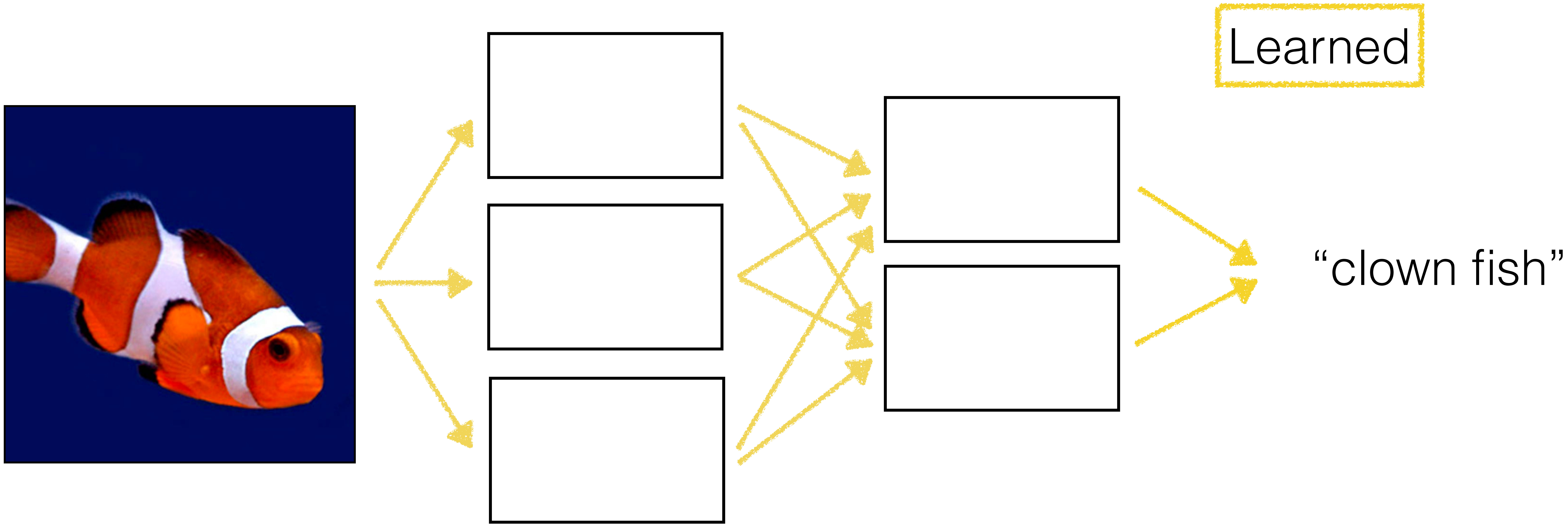


[Serre, 2014]

Classical object recognition



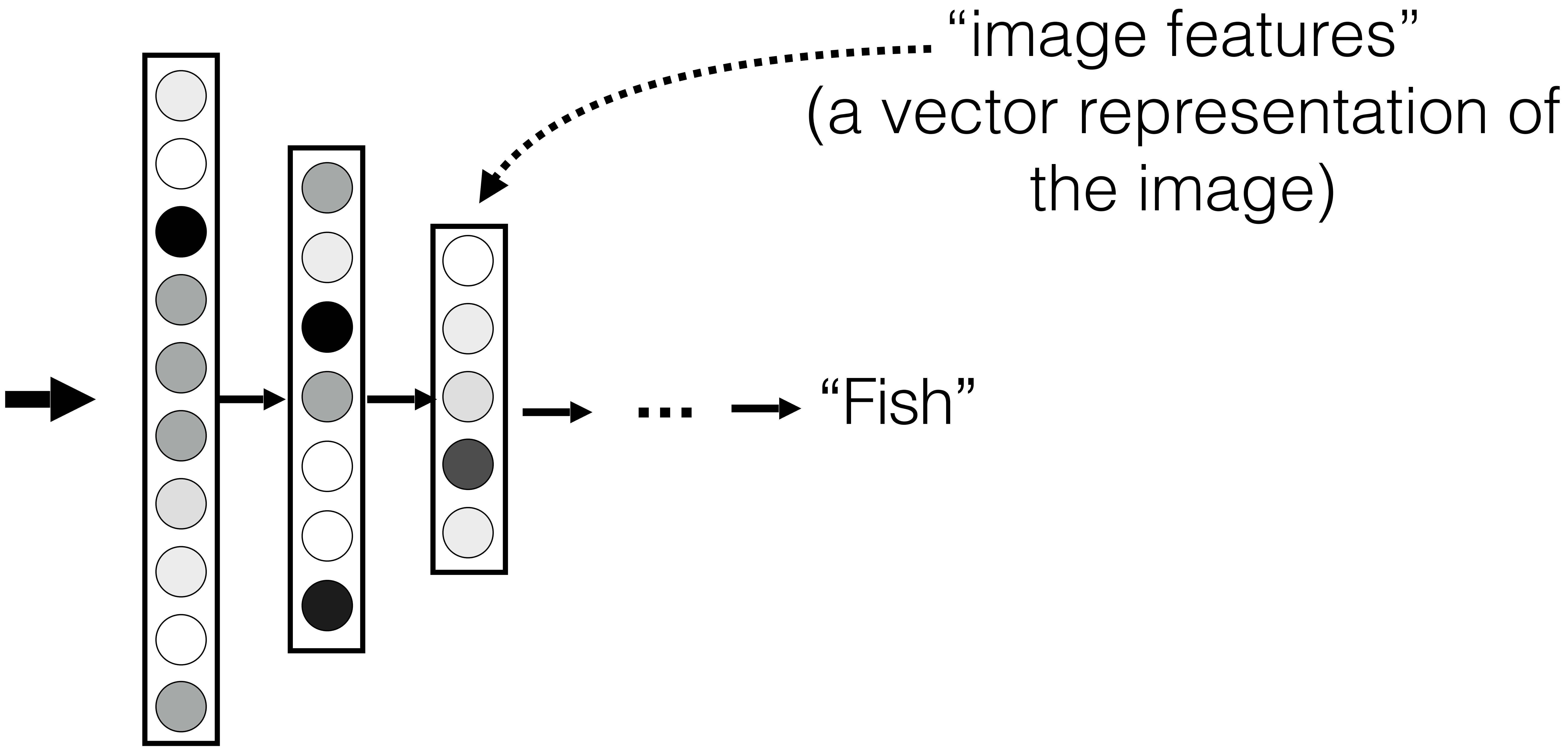
Deep learning



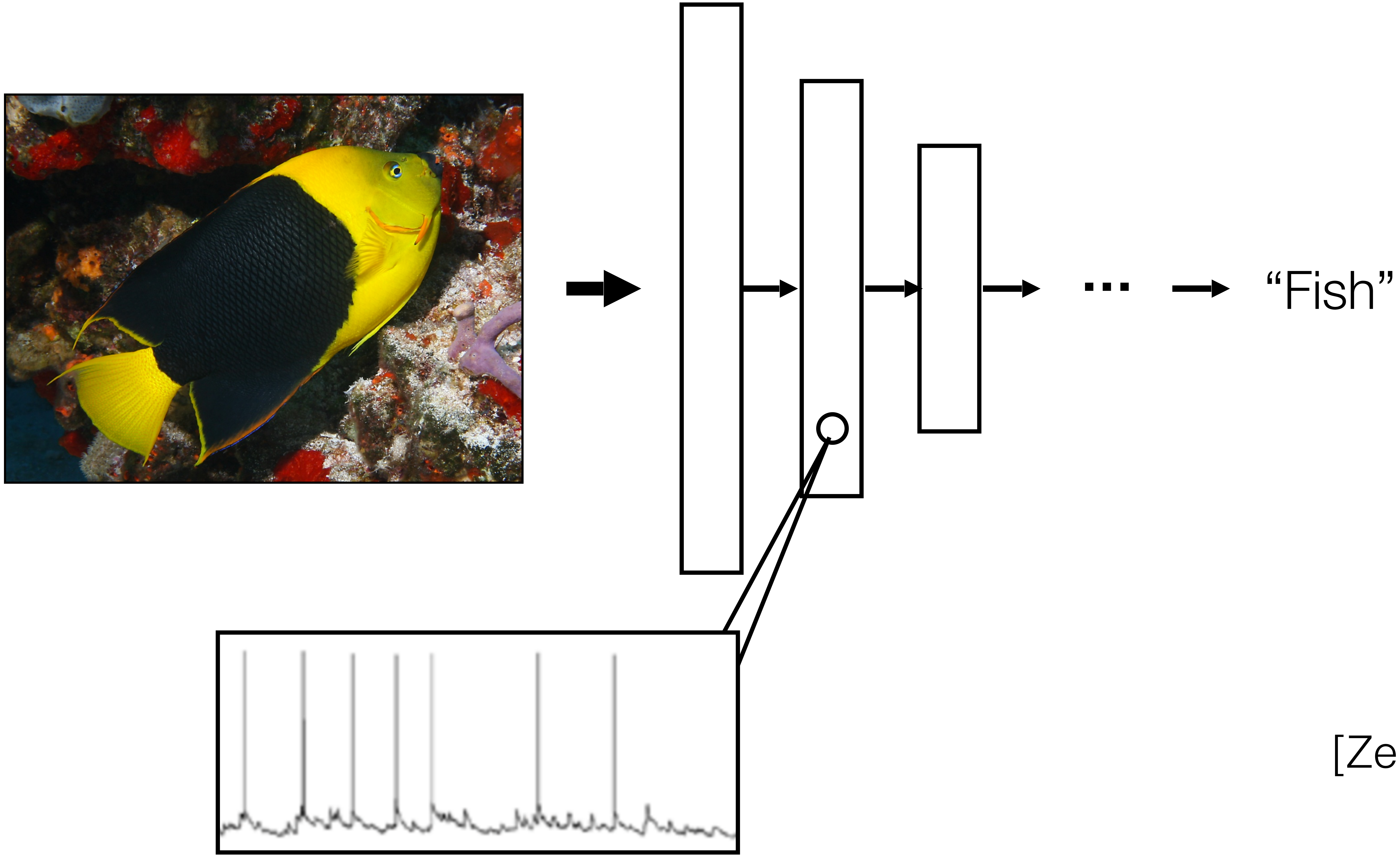
What do deep nets internally learn?

$\mathbf{x}$

Image



Deep Net “Electrophysiology”



[Zeiler & Fergus, ECCV 2014]

[Zhou et al., ICLR 2015]

Visualizing and Understanding CNNs

[Zeiler and Fergus, 2014]

Gabor-like filters learned by **layer 1**

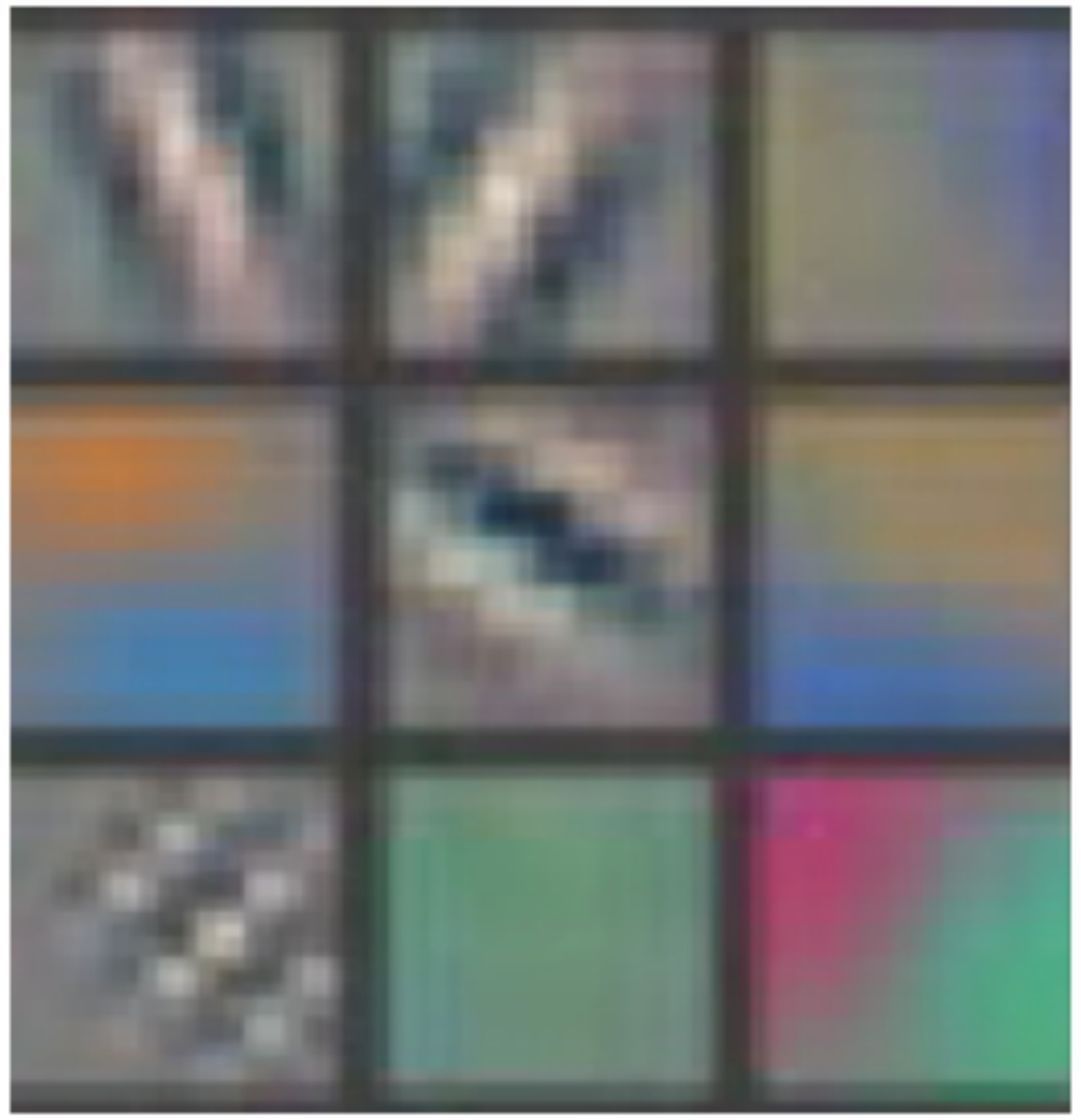
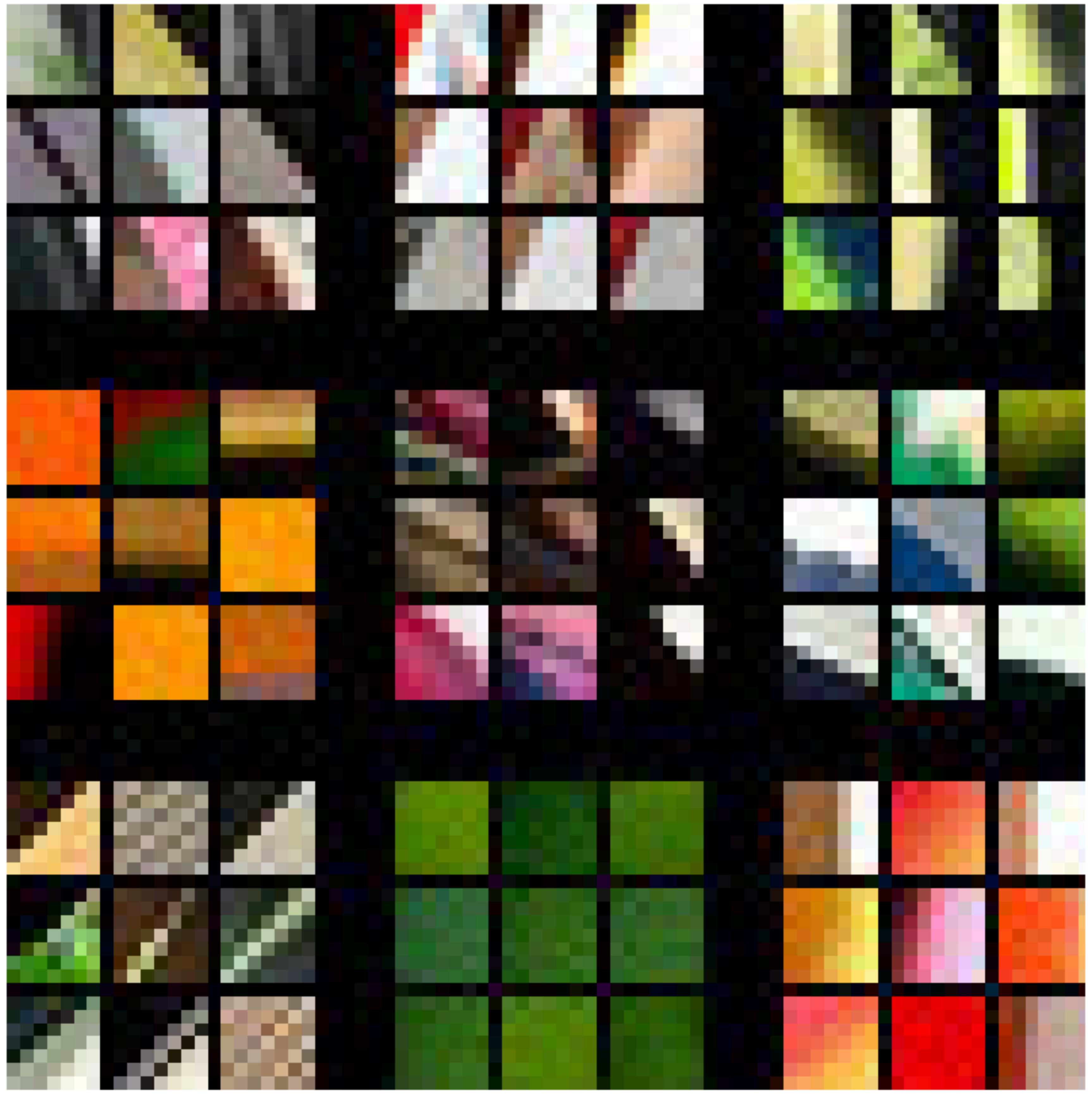
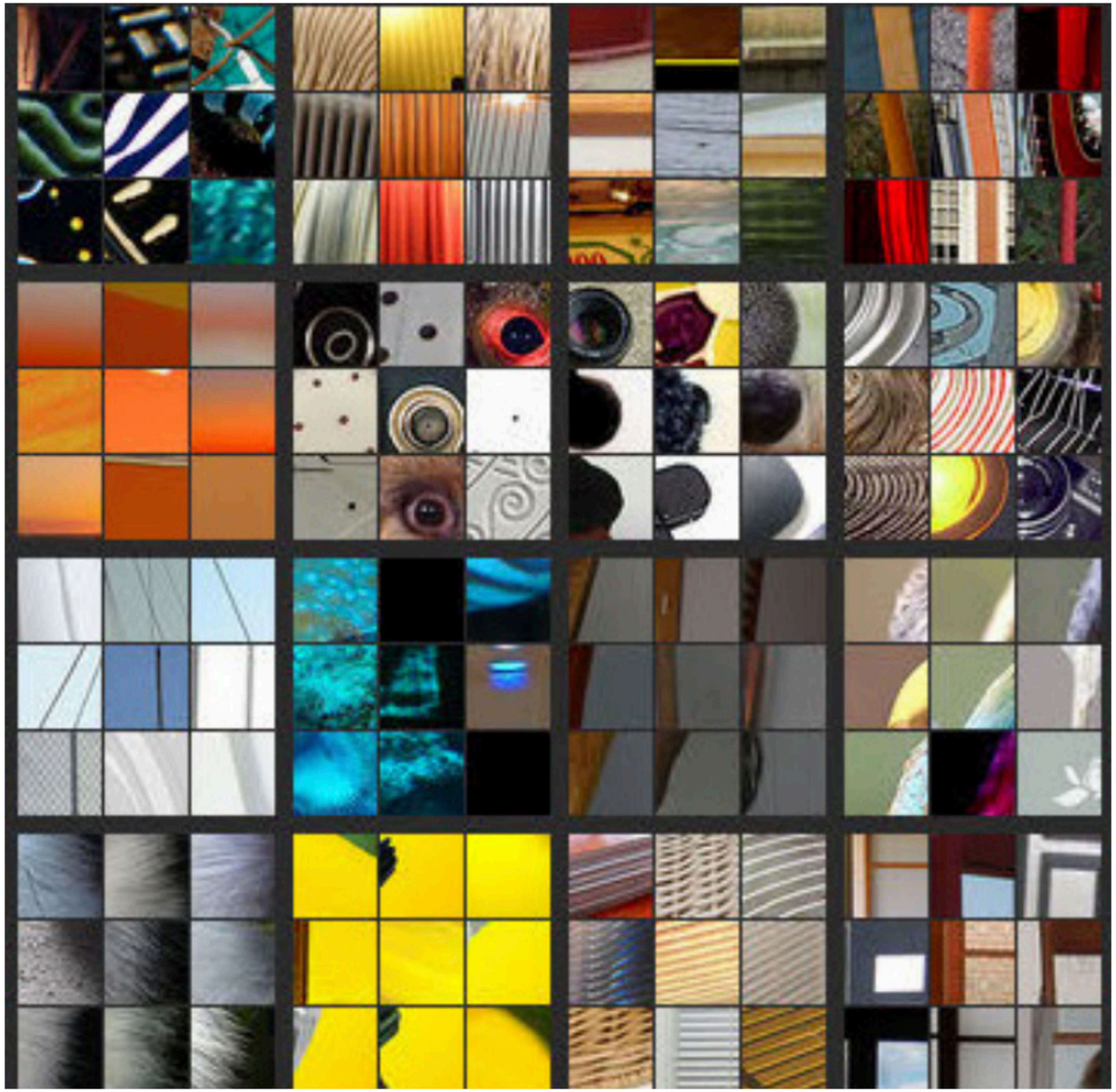


Image patches that activate each of the **layer 1** filters most strongly



[Zeiler and Fergus, 2014]

Image patches that activate each of the **layer 2** neurons most strongly



[Zeiler and Fergus, 2014]

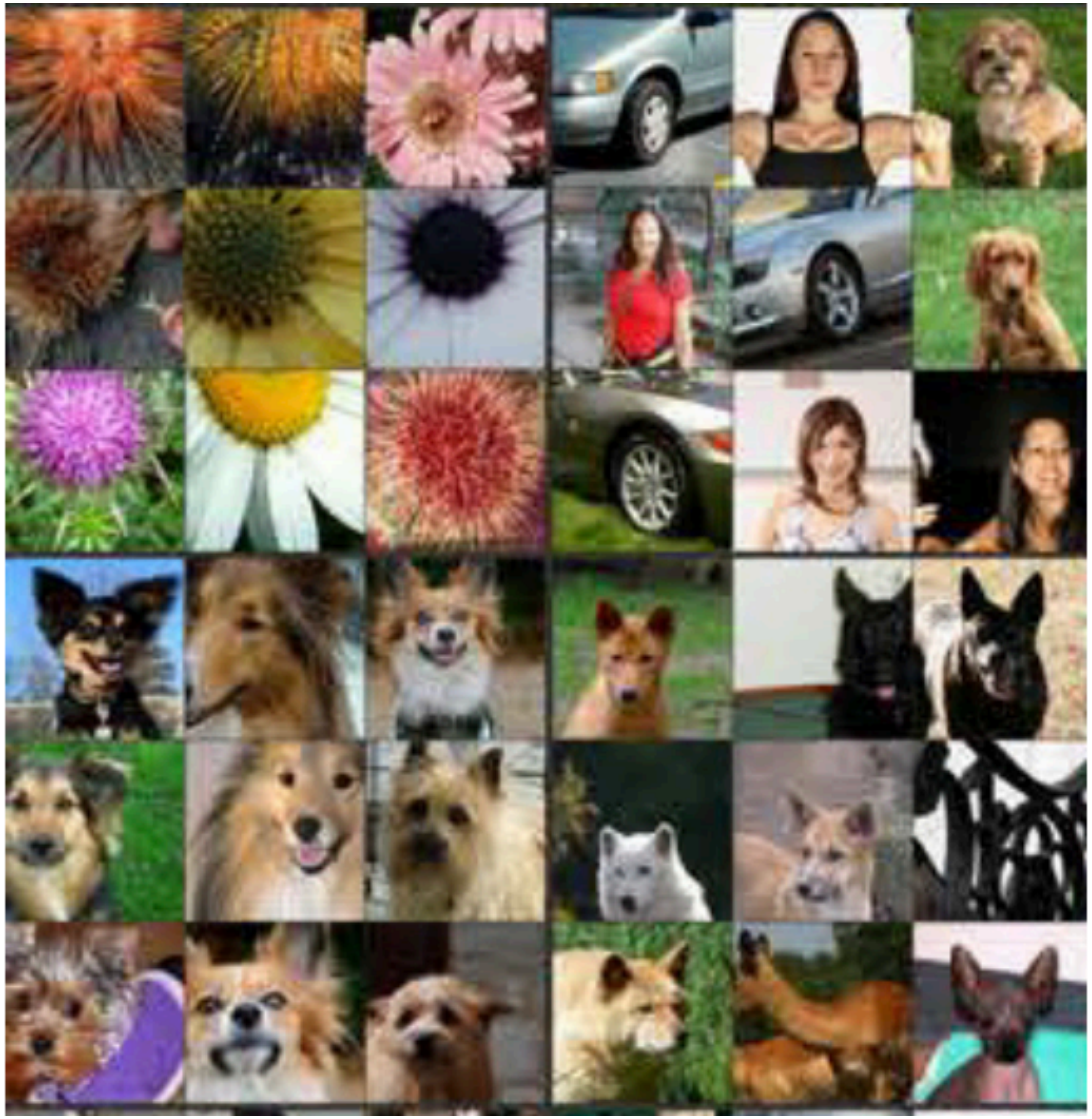
Image patches that activate each of the **layer 3** neurons most strongly

[Zeiler and Fergus, 2014]

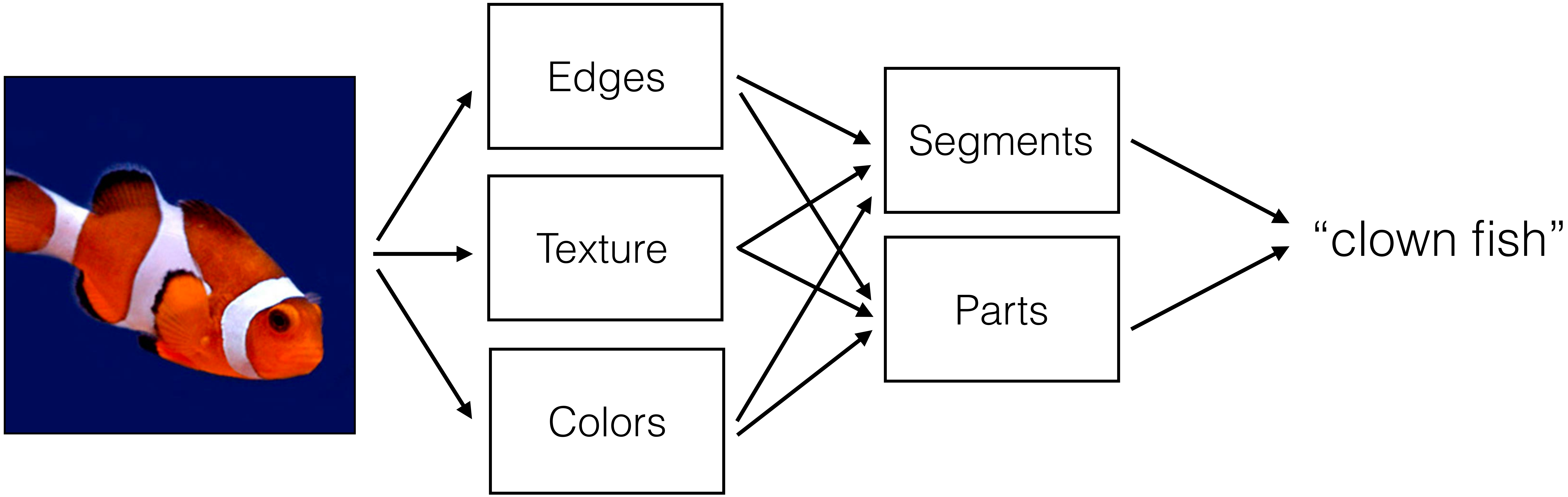
Image patches that activate each of the **layer 4** neurons most strongly

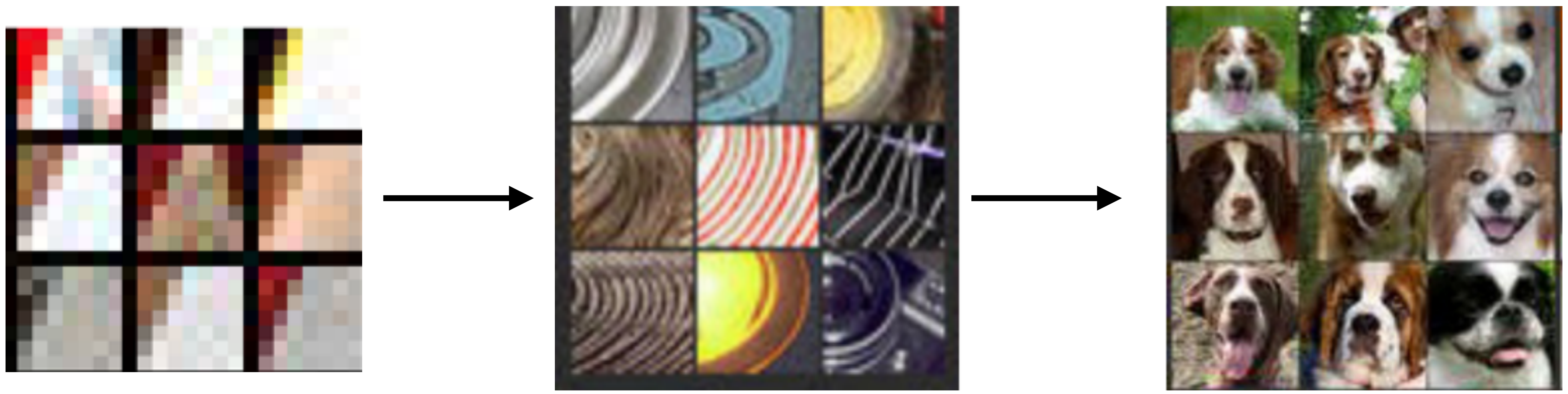
[Zeiler and Fergus, 2014]

Image patches that activate each of the **layer 5** neurons most strongly



CNNs *learned* the classical visual recognition pipeline!





Object Detectors Emergence in Deep Scene CNNs

[Zhou et al., ICLR 2015]

AlexNet

- For each unit (neuron) in network, find which images it is most selective for (cause it to have highest activation)
- Find which pixels in these images are responsible by occluding regions and seeing which pixels, when occluded, cause activation to change the most

Object Detectors Emergence in Deep Scene CNNs

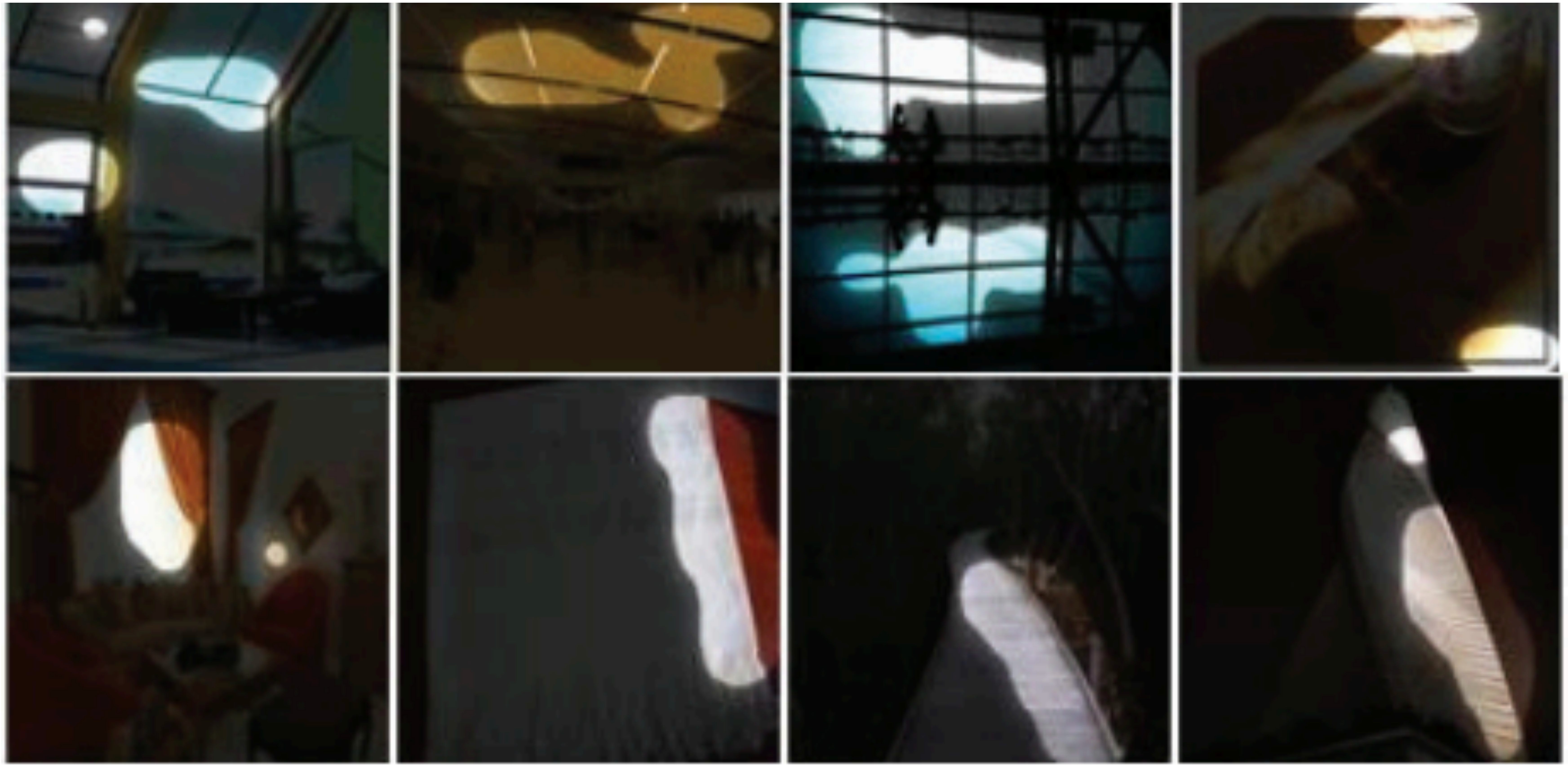
[Zhou et al., ICLR 2015]

pool 1

Object Detectors Emergence in Deep Scene CNNs

[Zhou et al., ICLR 2015]

pool 2



Object Detectors Emergence in Deep Scene CNNs

[Zhou et al., ICLR 2015]

conv 4

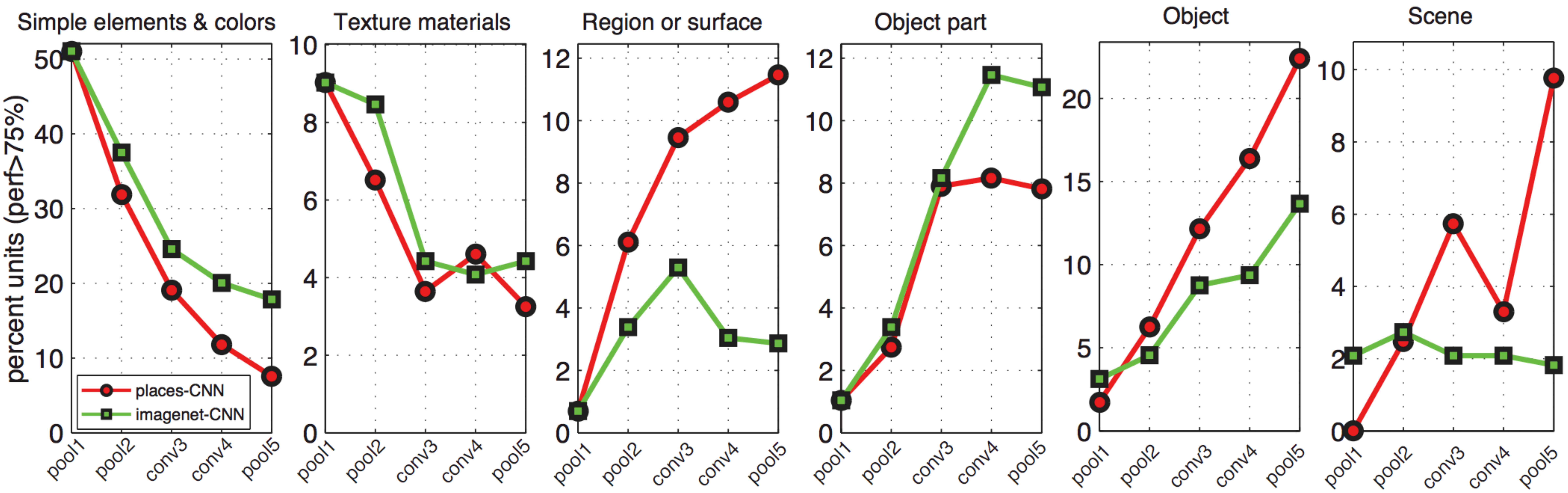
Object Detectors Emergence in Deep Scene CNNs

[Zhou et al., ICLR 2015]

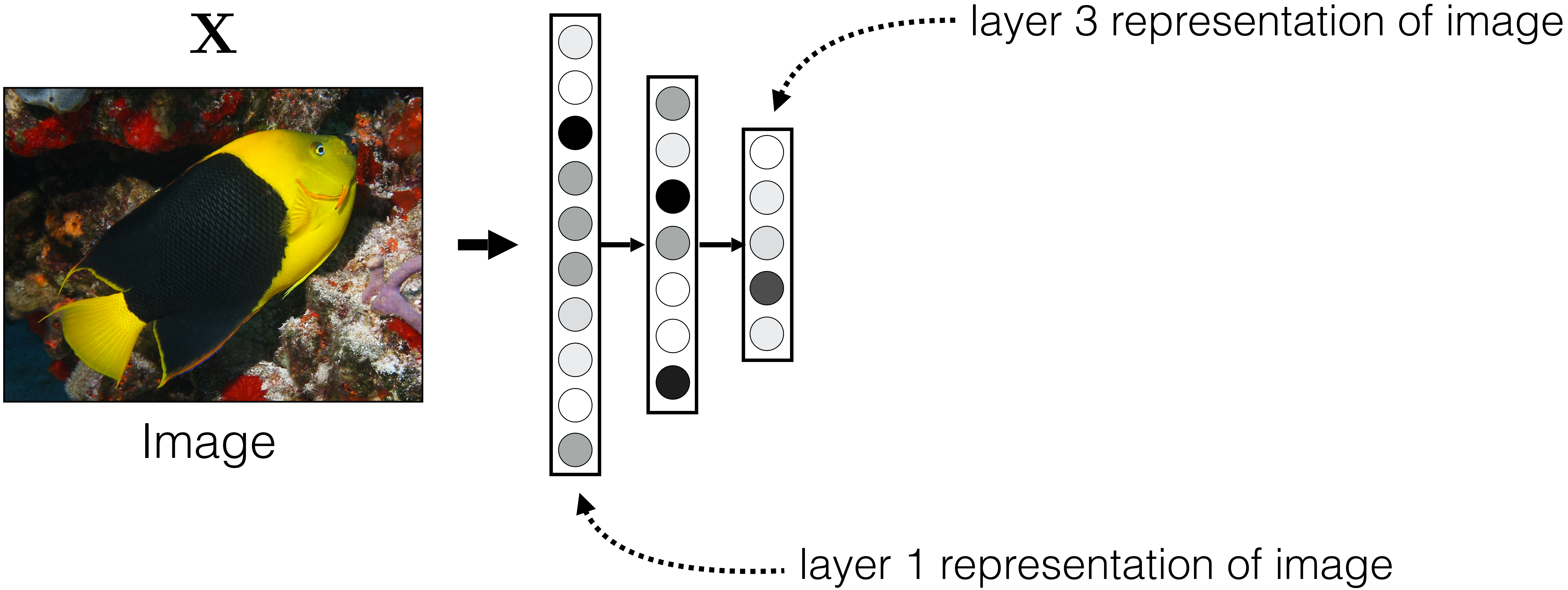
pool 5

Object Detectors Emergence in Deep Scene CNNs

[Zhou et al., ICLR 2015]



im2vec

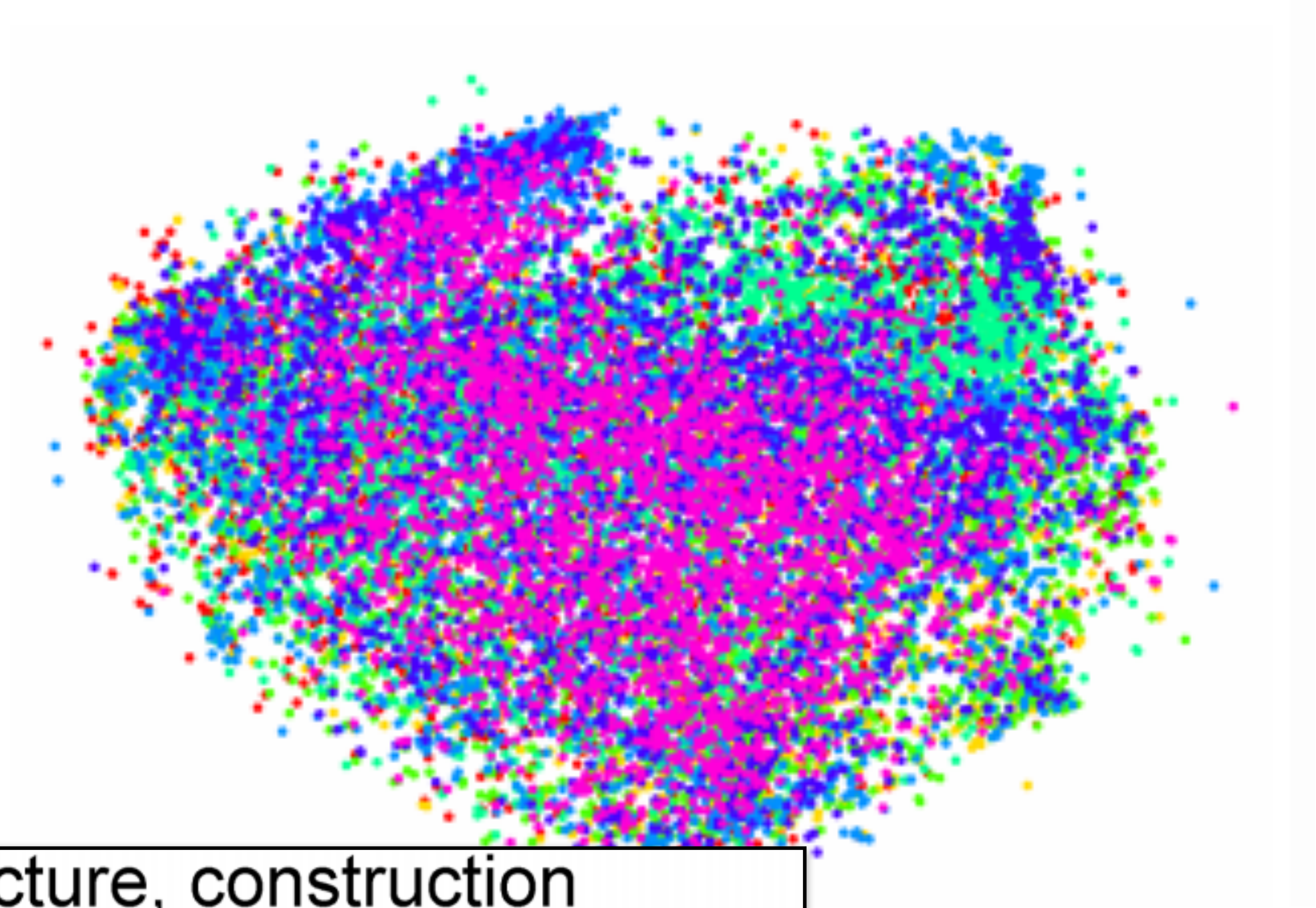


Represent image as a vector of neural activations

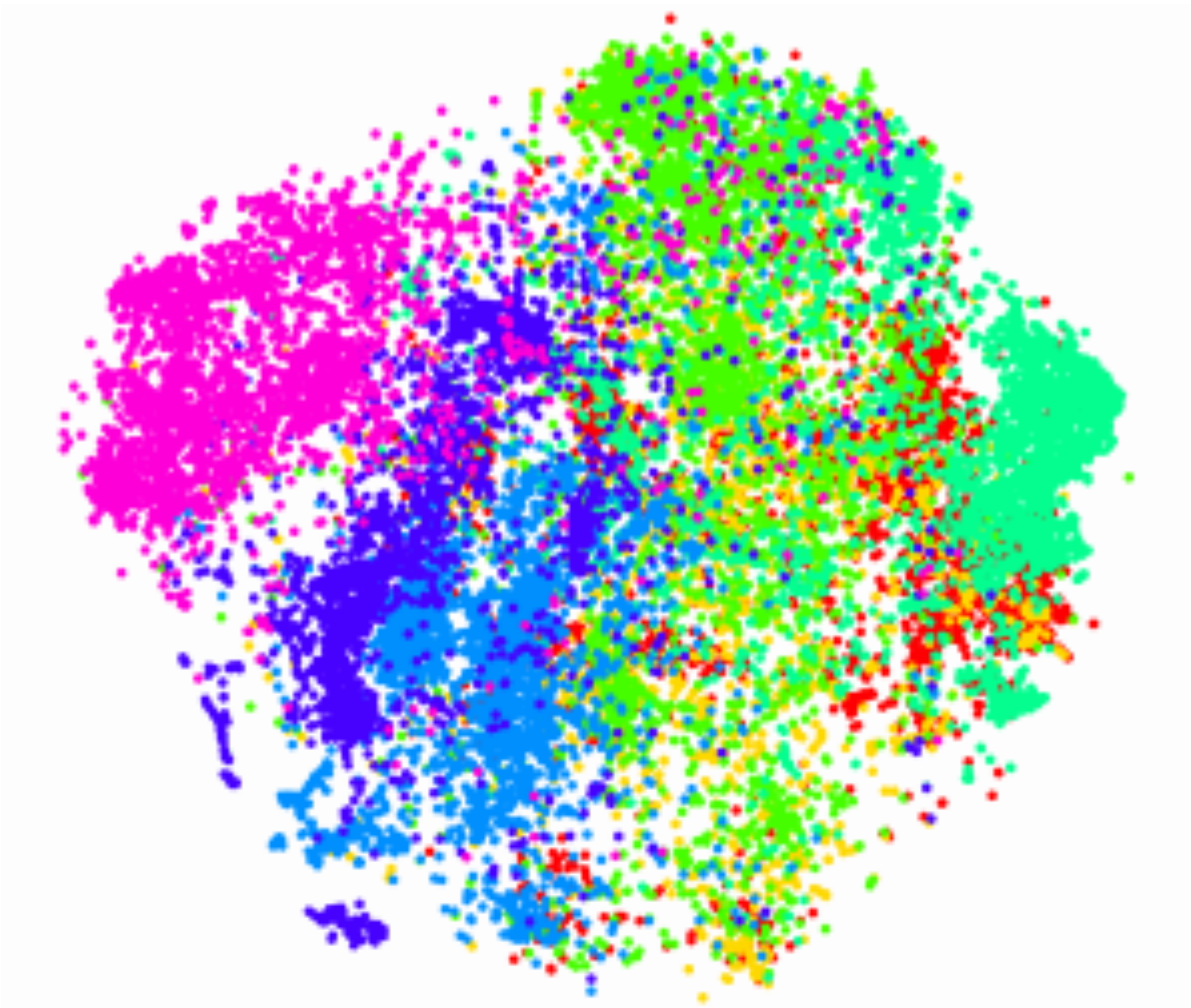
(perhaps representing a vector of detected texture patterns or object parts)

Disentangling

Layer 1 representation



Layer 6 representation



- structure, construction
- covering
- commodity, trade good, good
- conveyance, transport
- invertebrate
- bird
- hunting dog

[DeCAF, Donahue, Jia, et al. 2013]

[Visualization technique : t-sne, van der Maaten & Hinton, 2008]

Investigating a representation via similarity analysis

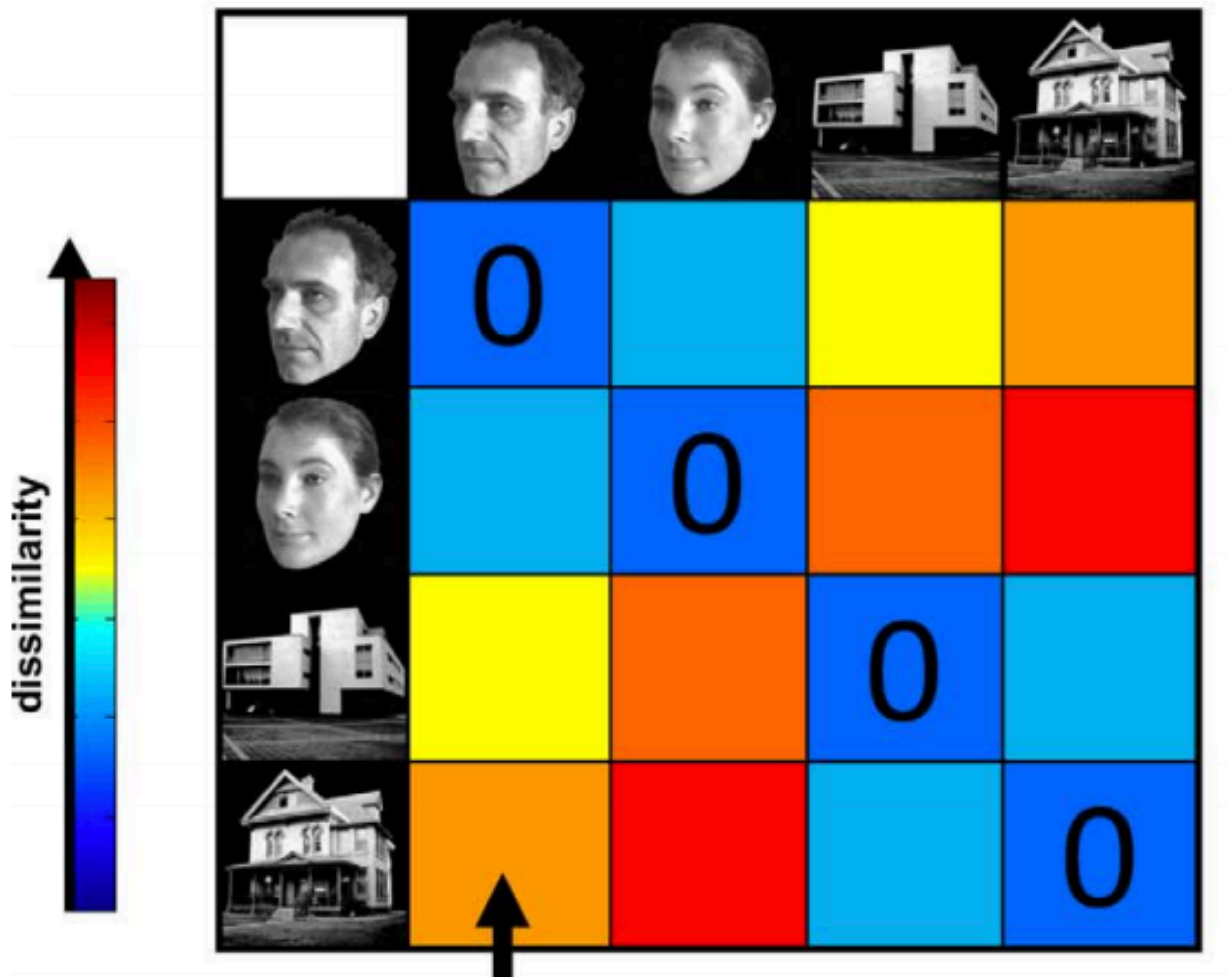
How similar are these two images?

How about these two?

[Kriegeskorte et al. 2008]

Investigating a representation via similarity analysis

Representational Dissimilarity Matrix



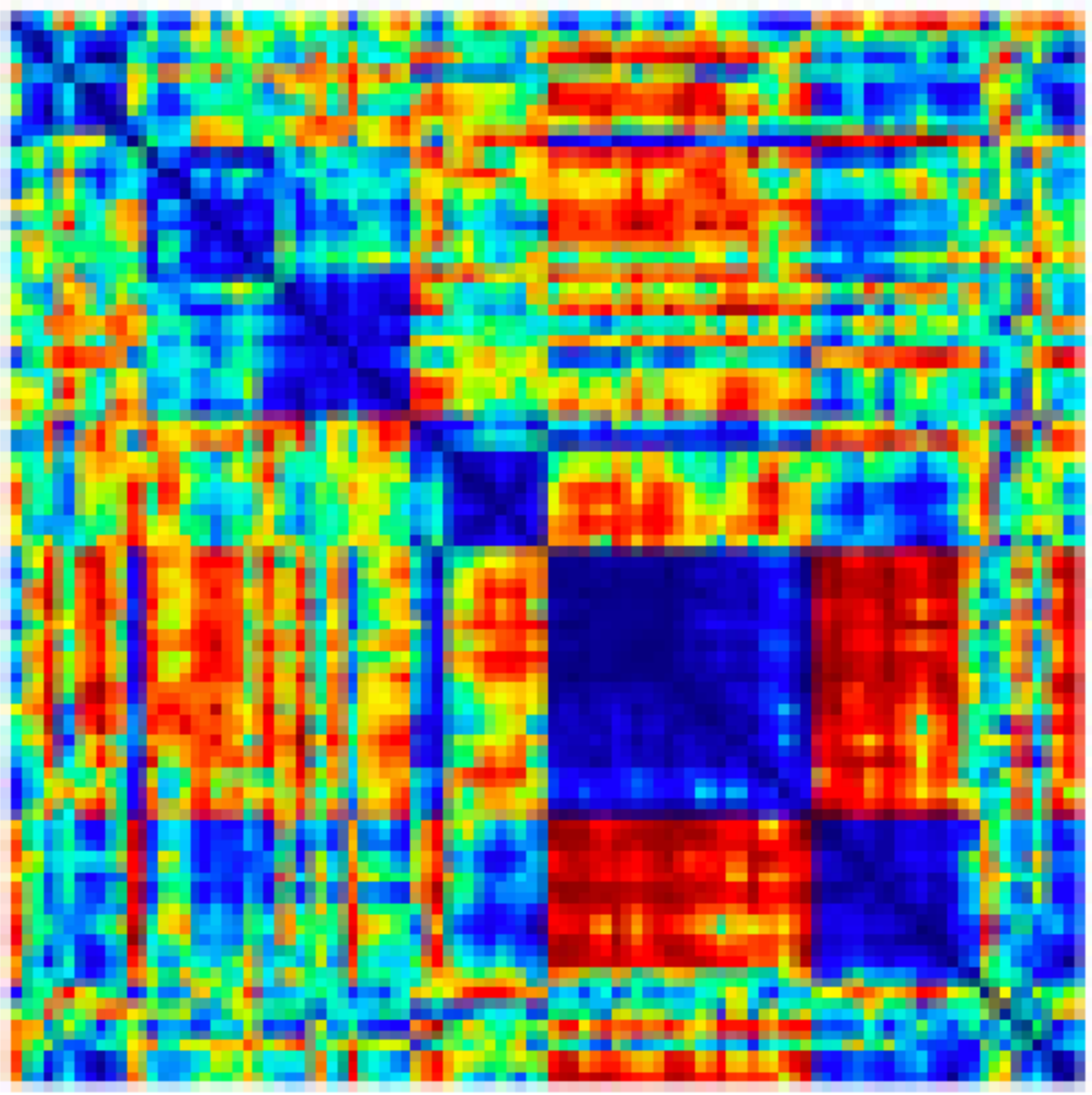
$$\|\mathbf{h}_i - \mathbf{h}_j\|$$

Neural activation vector

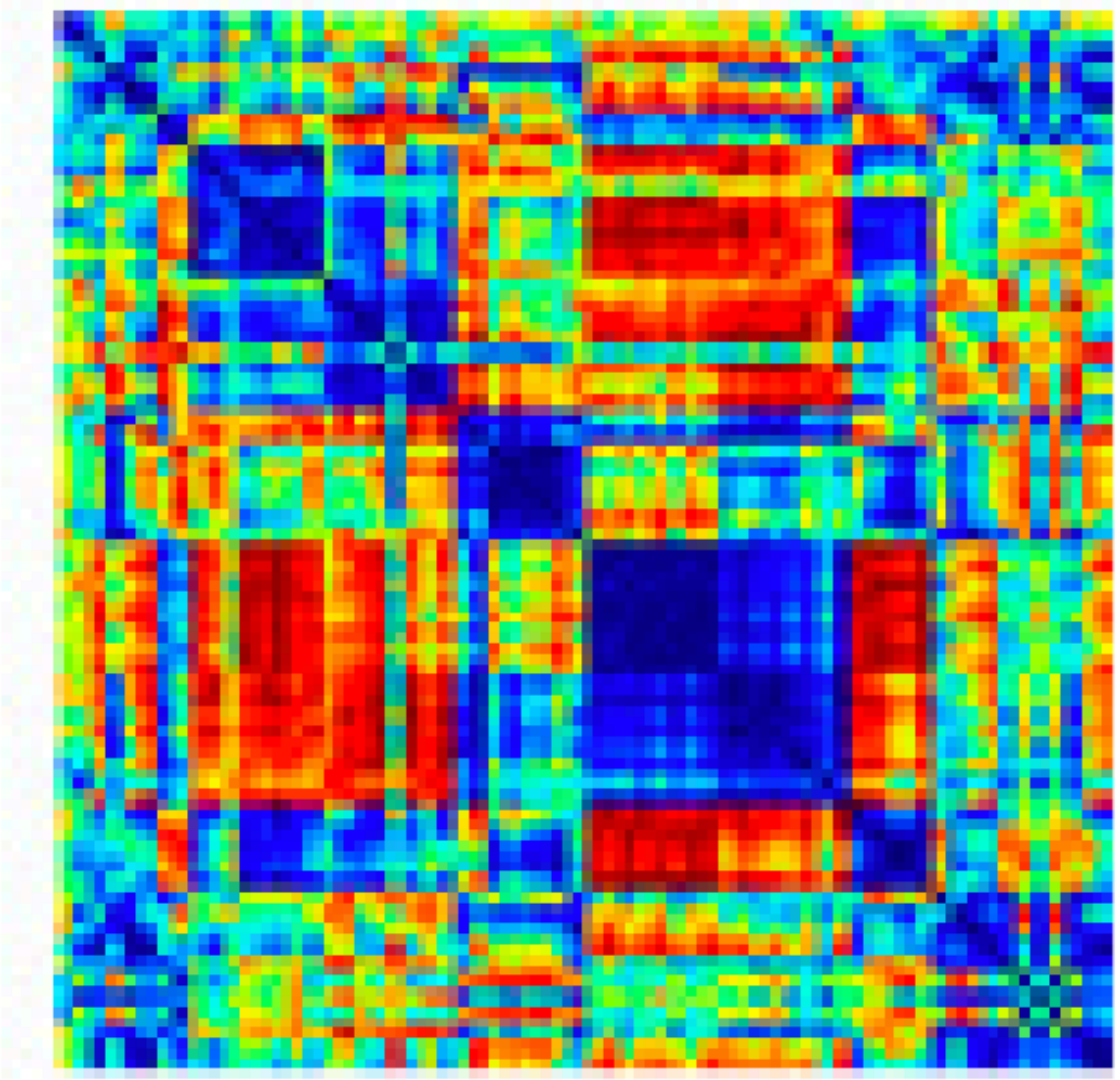
[Kriegeskorte et al. 2008]

Investigating a representation via similarity analysis

IT Neuronal Units



Deep net (in paricular, HMO)



[Yamins et al., PNAS 2014]

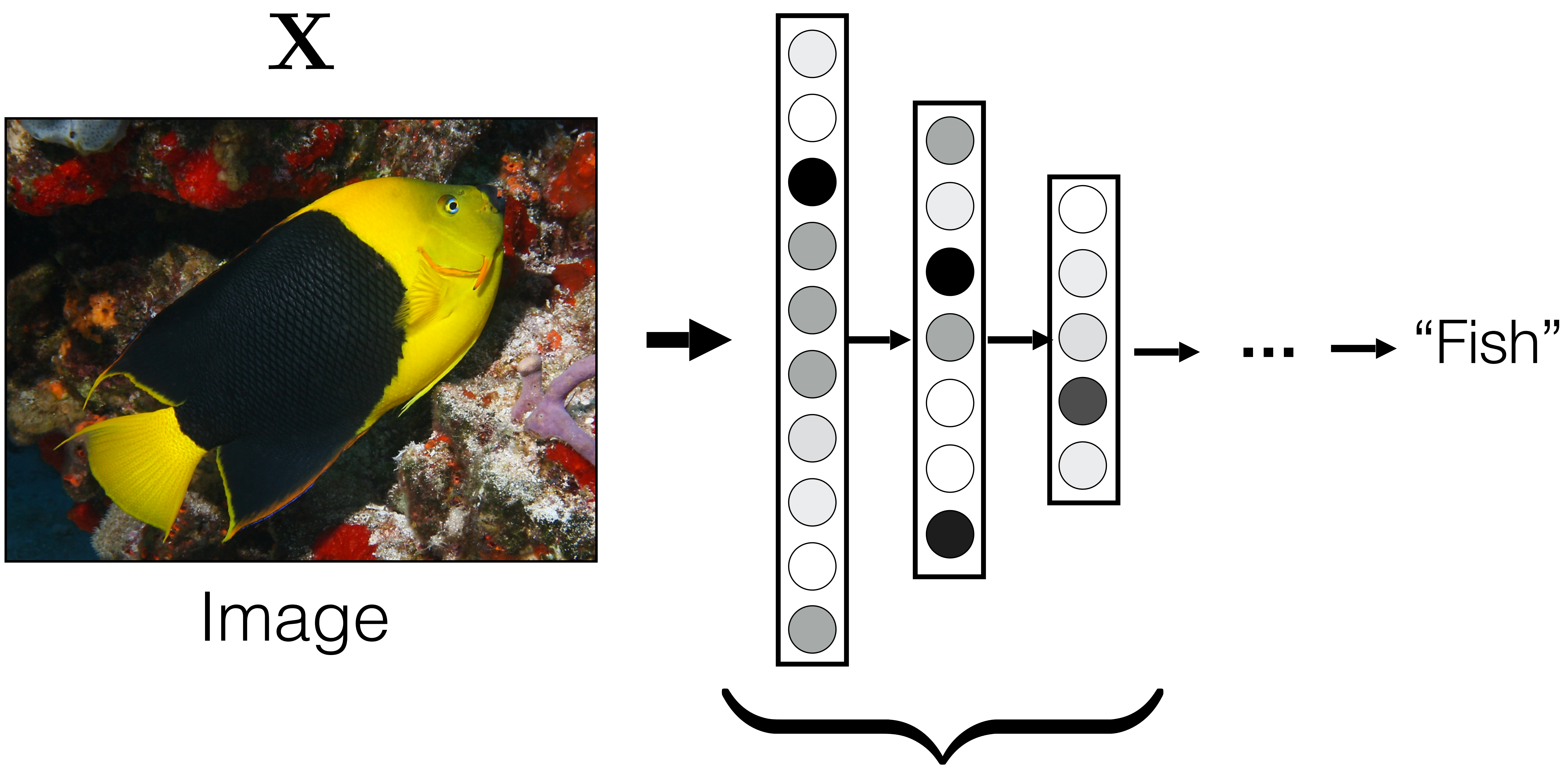
Investigating a representation via similarity analysis

Deep nets and the primate brain both learn similar metric spaces.

Deep nets organize visual information similarly to how our brains do!

[Yamins et al., PNAS 2014]

What do deep nets internally learn?



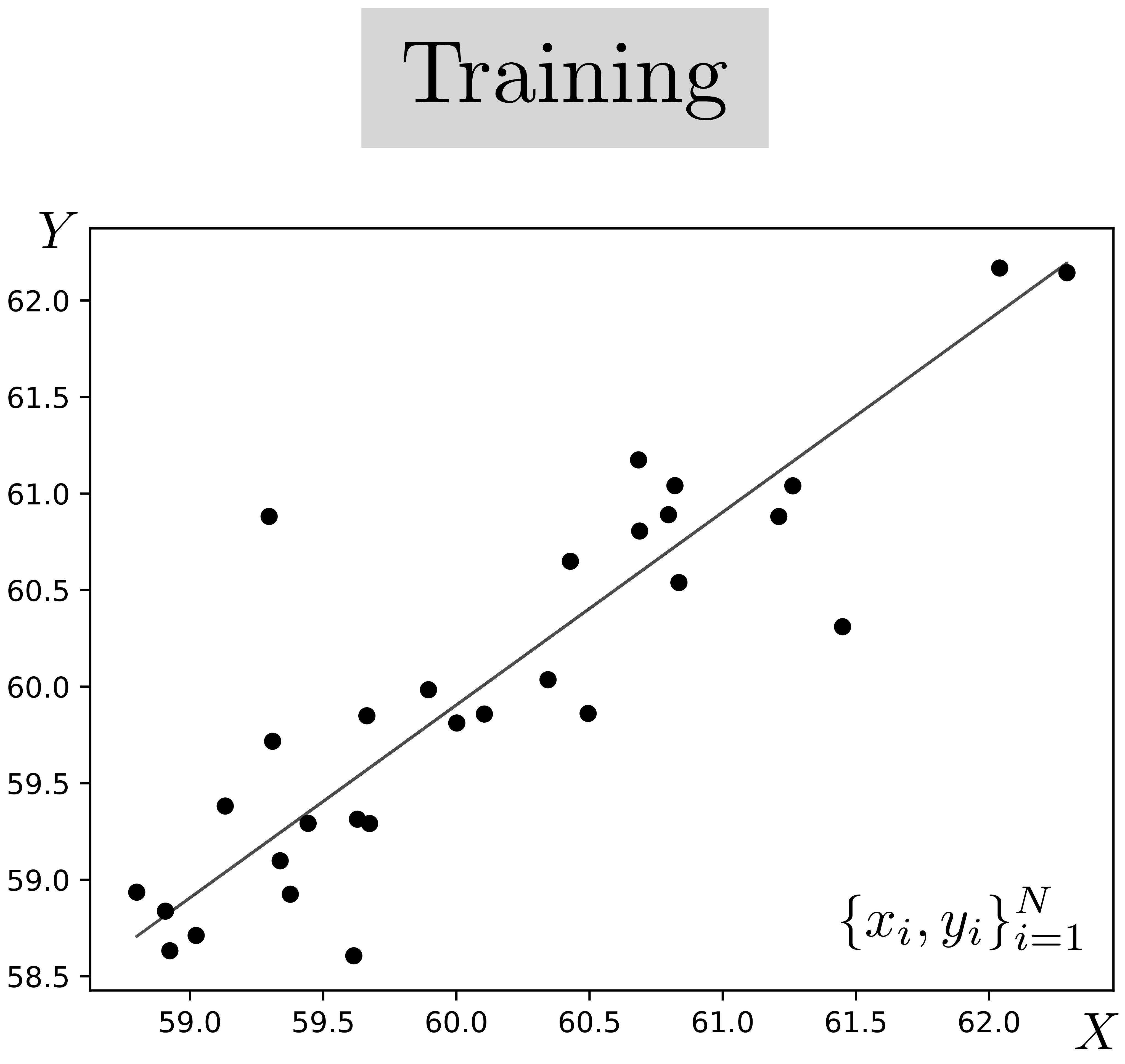
Representations!

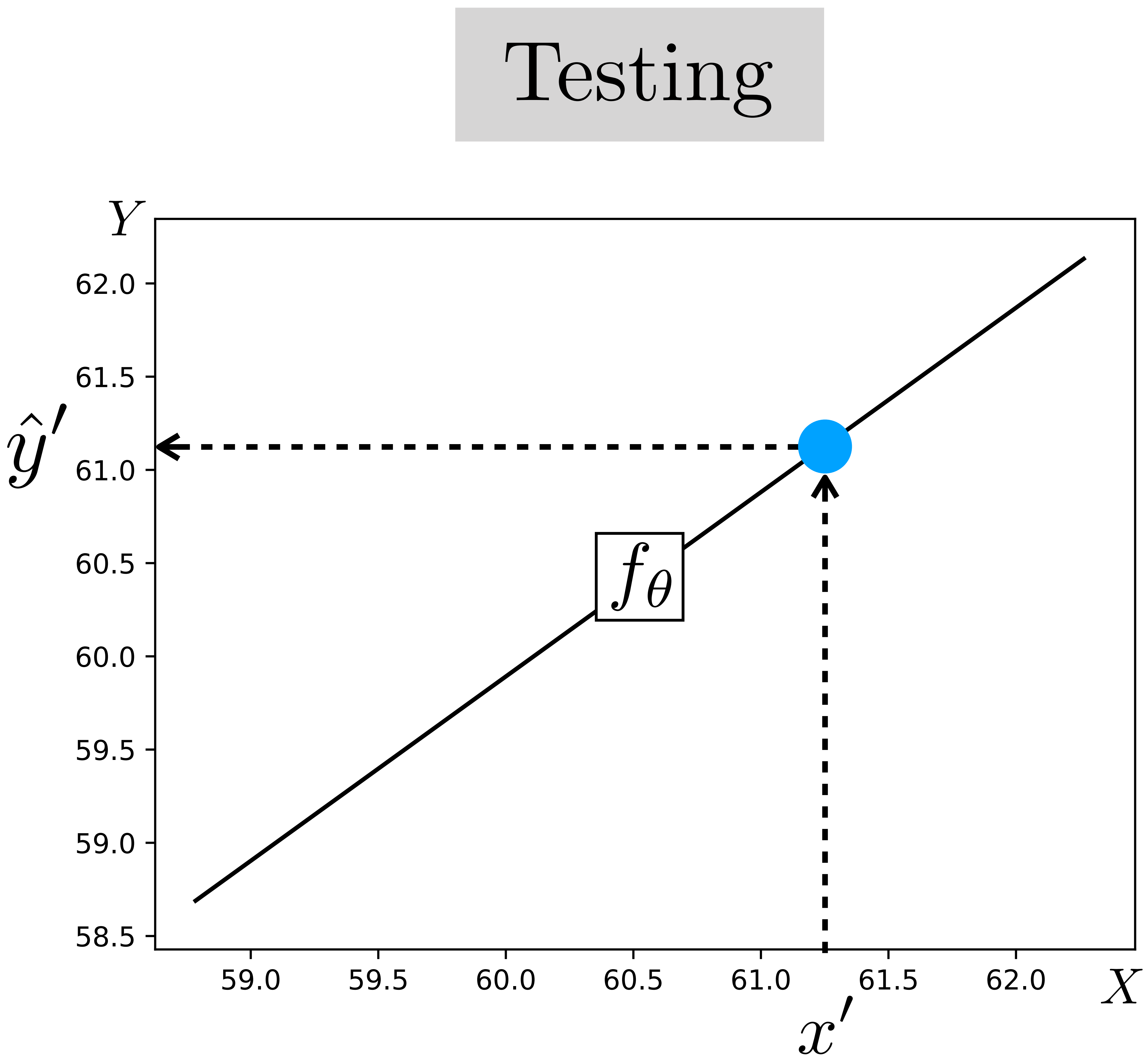
A CNN is a multiscale, hierarchical representation of data

Transfer learning

"Generally speaking, a good representation is one that makes a subsequent learning task easier." — Deep Learning textbook

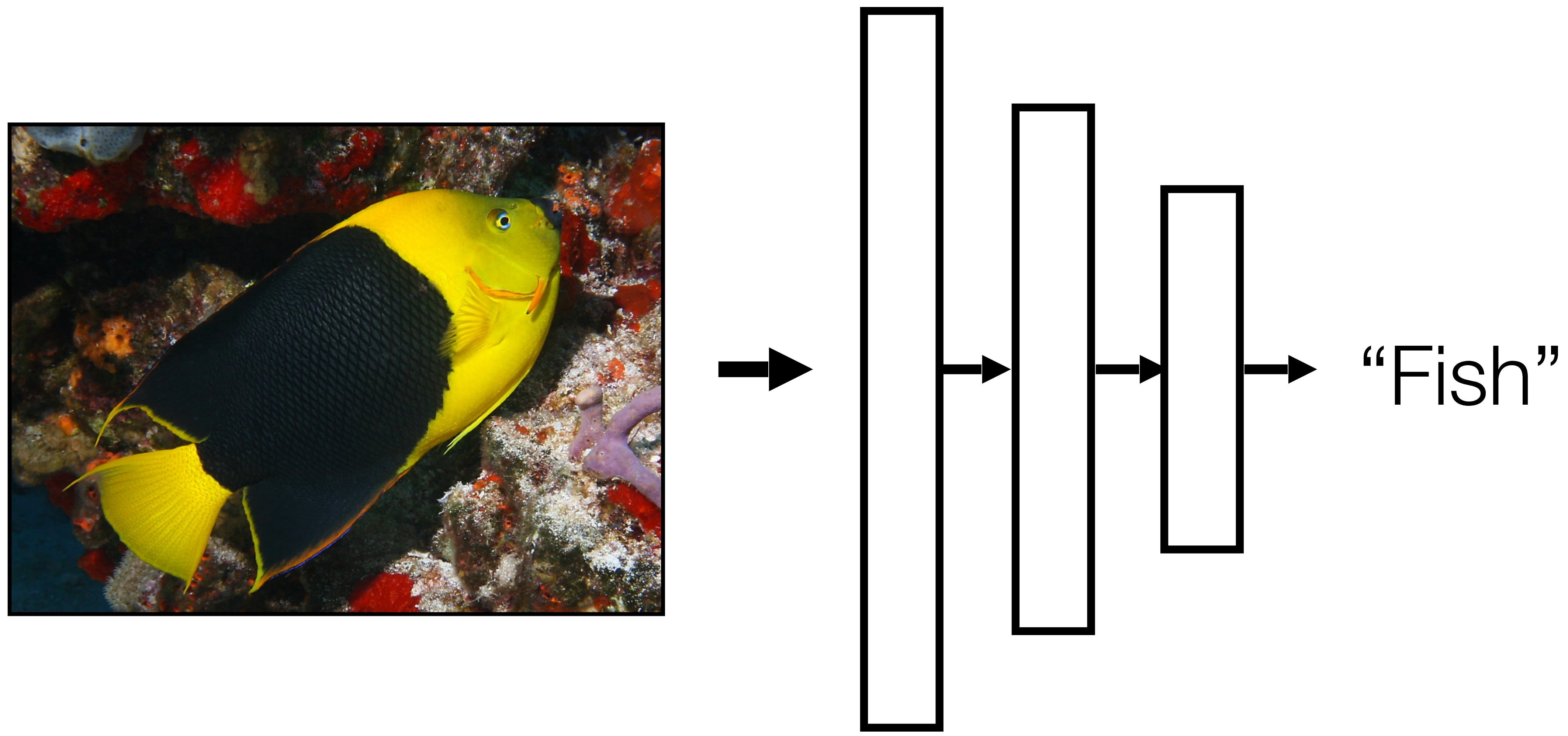
?


Training
Y
X
62.0
61.5
61.0
60.5
60.0
59.5
59.0
58.5
59.0
59.5
60.0
60.5
61.0
61.5
62.0
$\{x_i, y_i\}_{i=1}^N$


Testing
Y
X
$\hat{y}'$
x'
f_θ
62.0
61.5
61.0
60.5
60.0
59.5
59.0
58.5
59.0
59.5
60.0
60.5
61.0
61.5
62.0

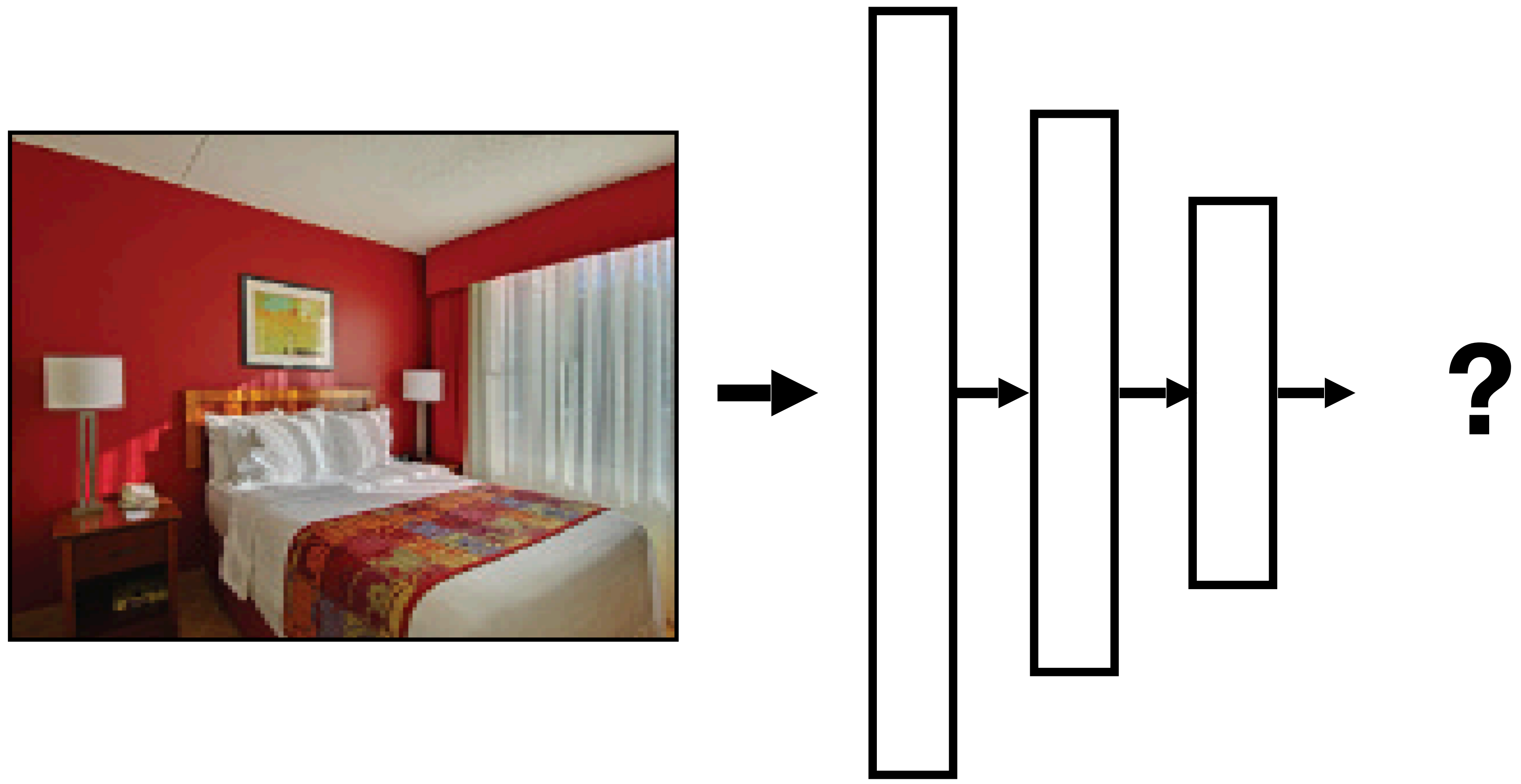
Training

Object recognition

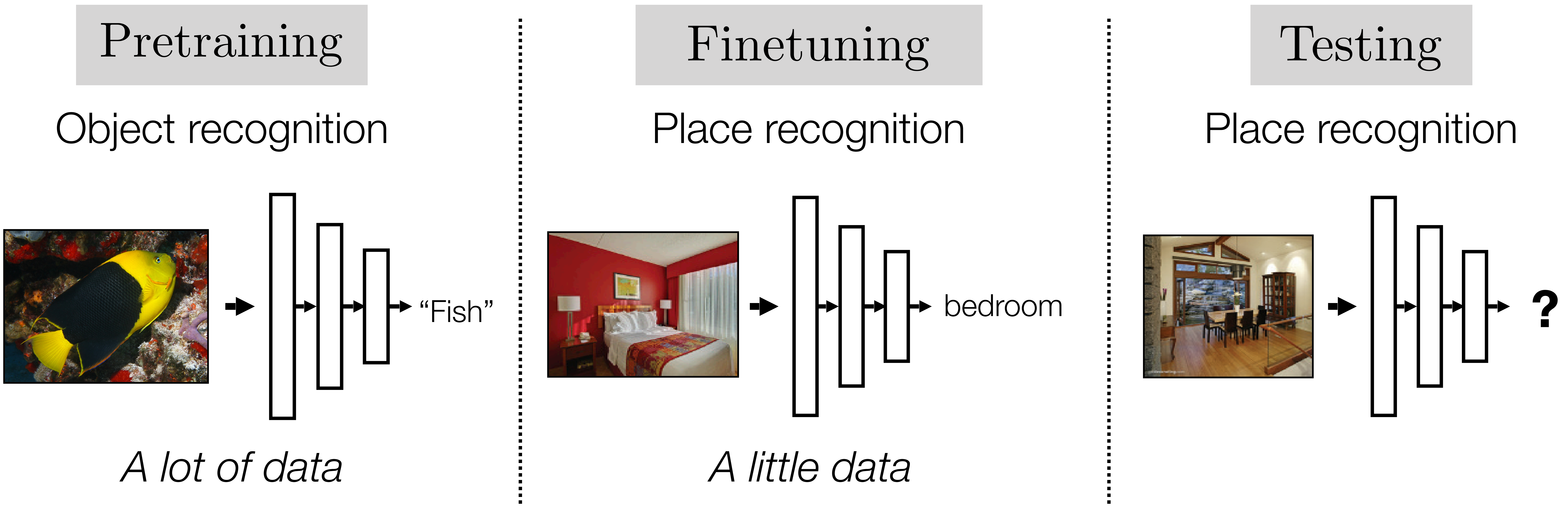


Testing

Place recognition



Often, what we will be “tested” on is to learn to do a new thing.



Finetuning starts with the representation learned on a previous task, and adapts it to perform well on a new task.

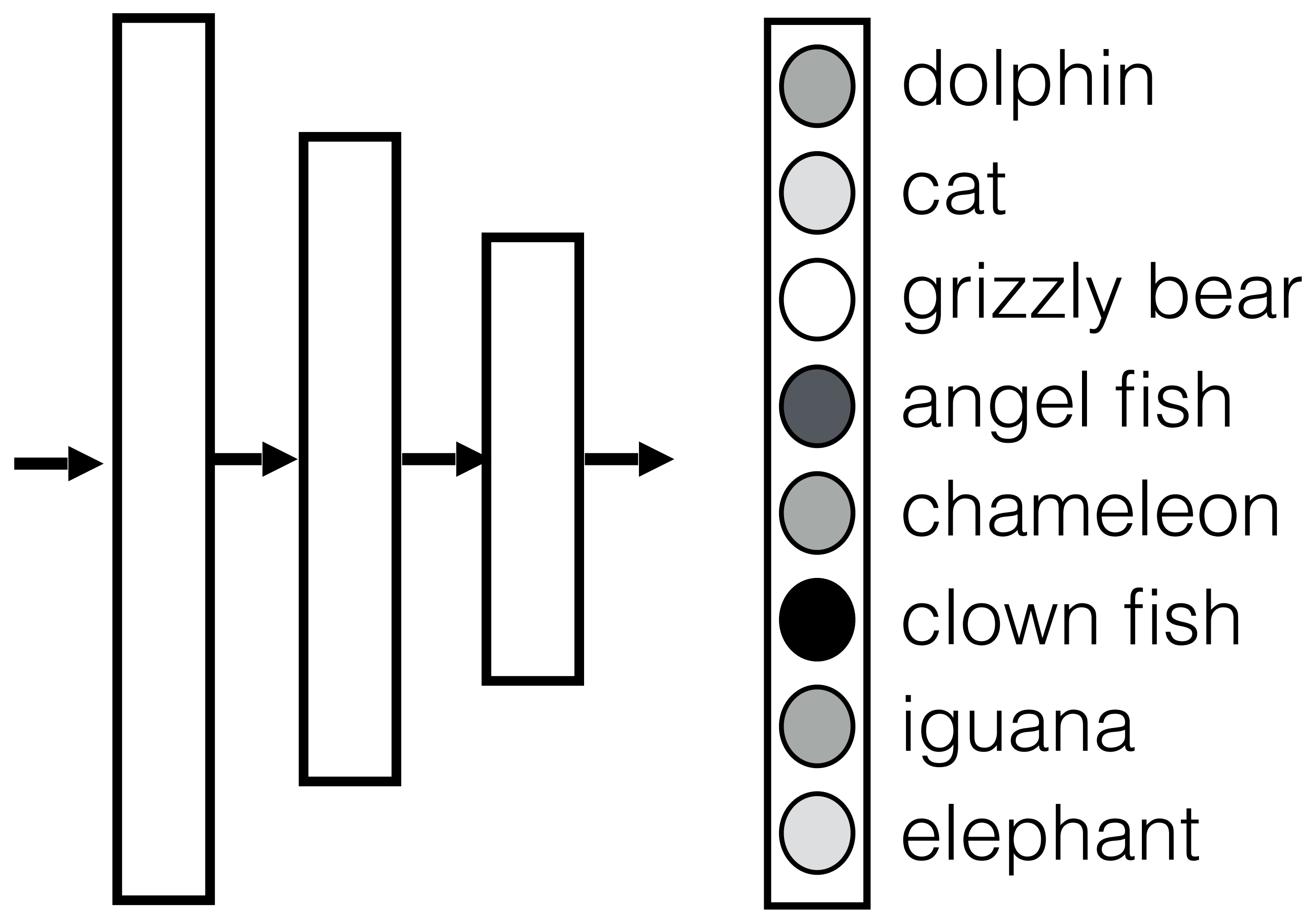
Finetuning in practice

- Pretrain a network on task A (often object recognition), resulting in parameters **W** and **b**
- Initialize a second network with some all of **W** and **b**
- Train the second network on task B, resulting in parameters **W’** and **b’**

Finetuning in practice

Pretraining

Object recognition



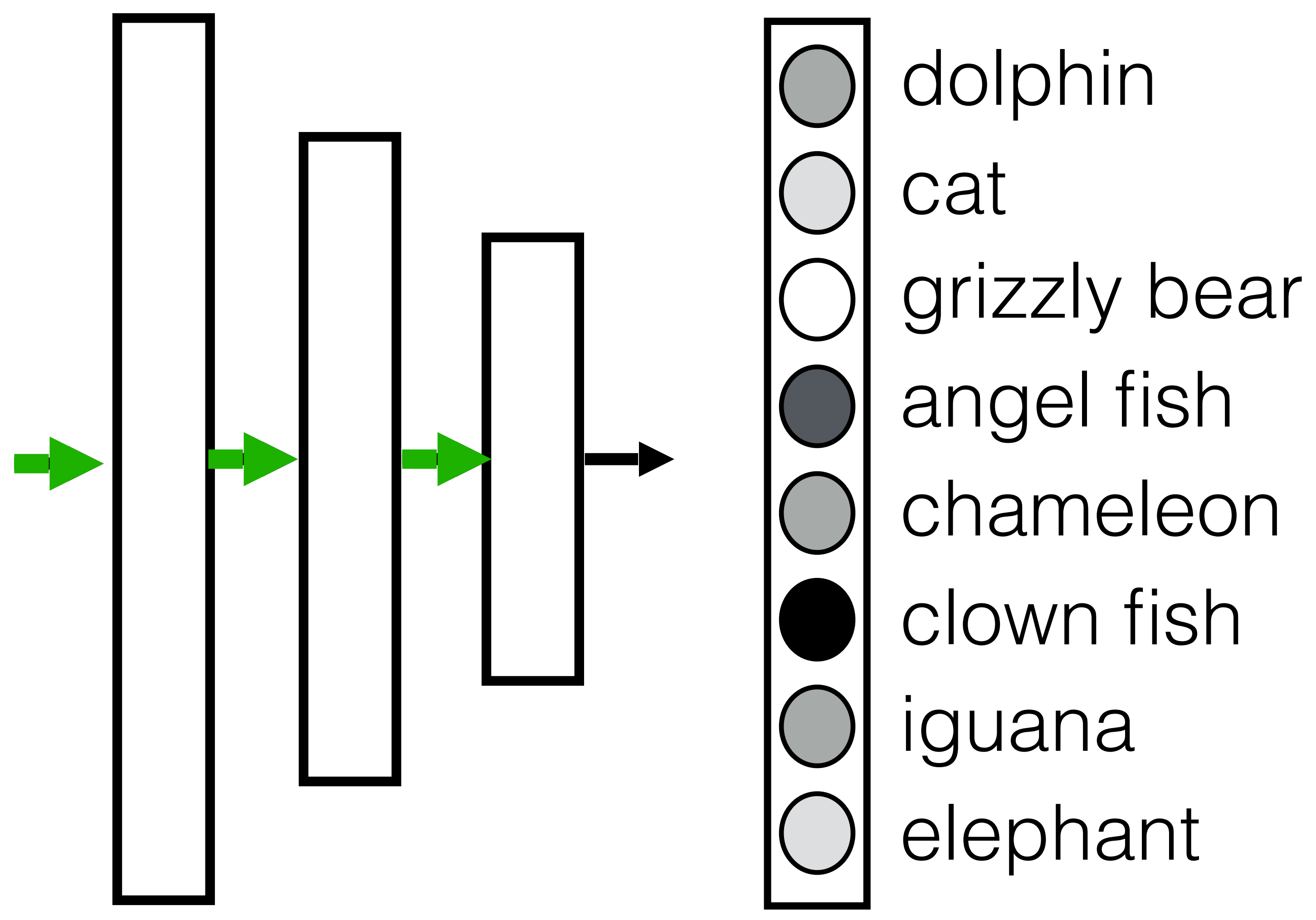
Finetuning

Place recognition

Finetuning in practice

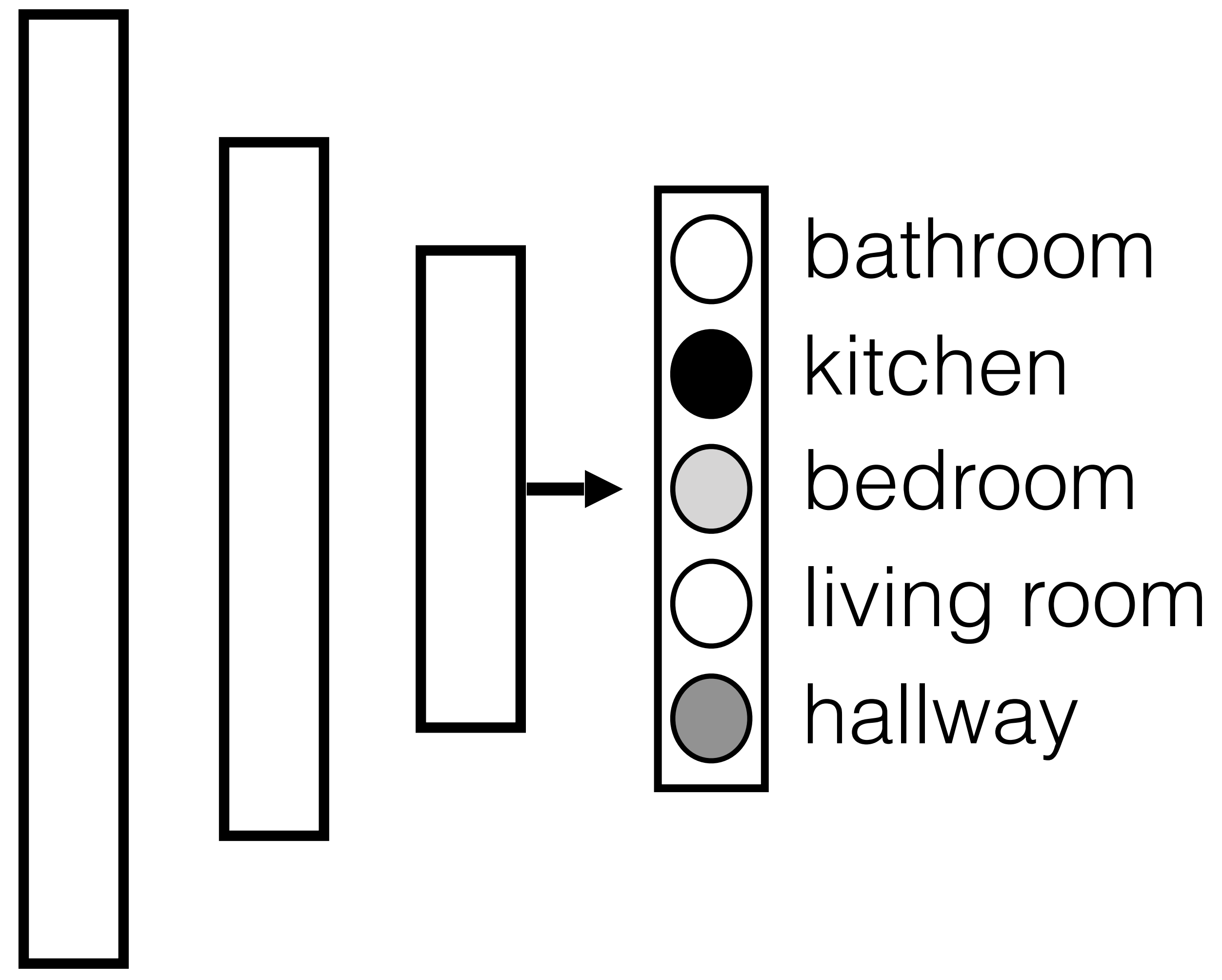
Pretraining

Object recognition



Finetuning

Place recognition



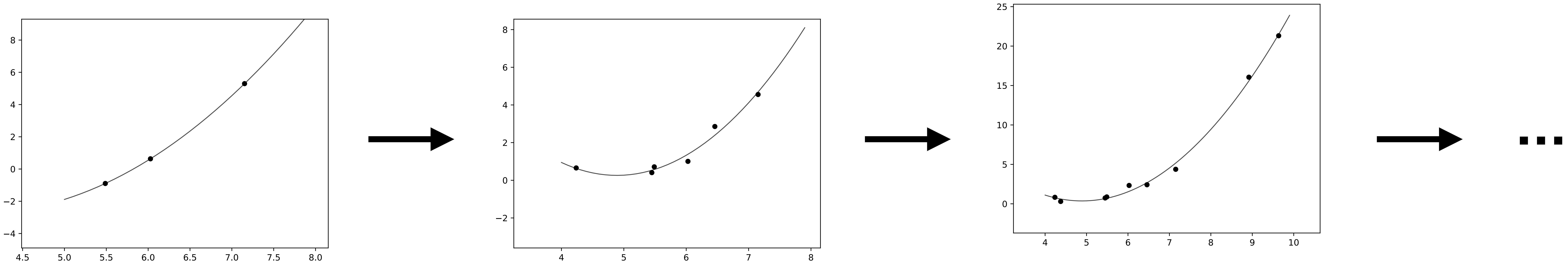
The "learned representation" is just the weights and biases, so that's what we transfer

If we keep on finetuning for every new datapoint or task that comes our way, we get **online learning**. Humans seem to do this, we never stop learning.

Training

More training

More training



Supervised object recognition

→ **Learner** →

"Fish"

image X

label Y

Supervised object recognition

Learner

"Fish"

image X

label Y

Supervised object recognition

Learner

"Fish"

image X

label Y

Supervised object recognition

Learner

"Fish"

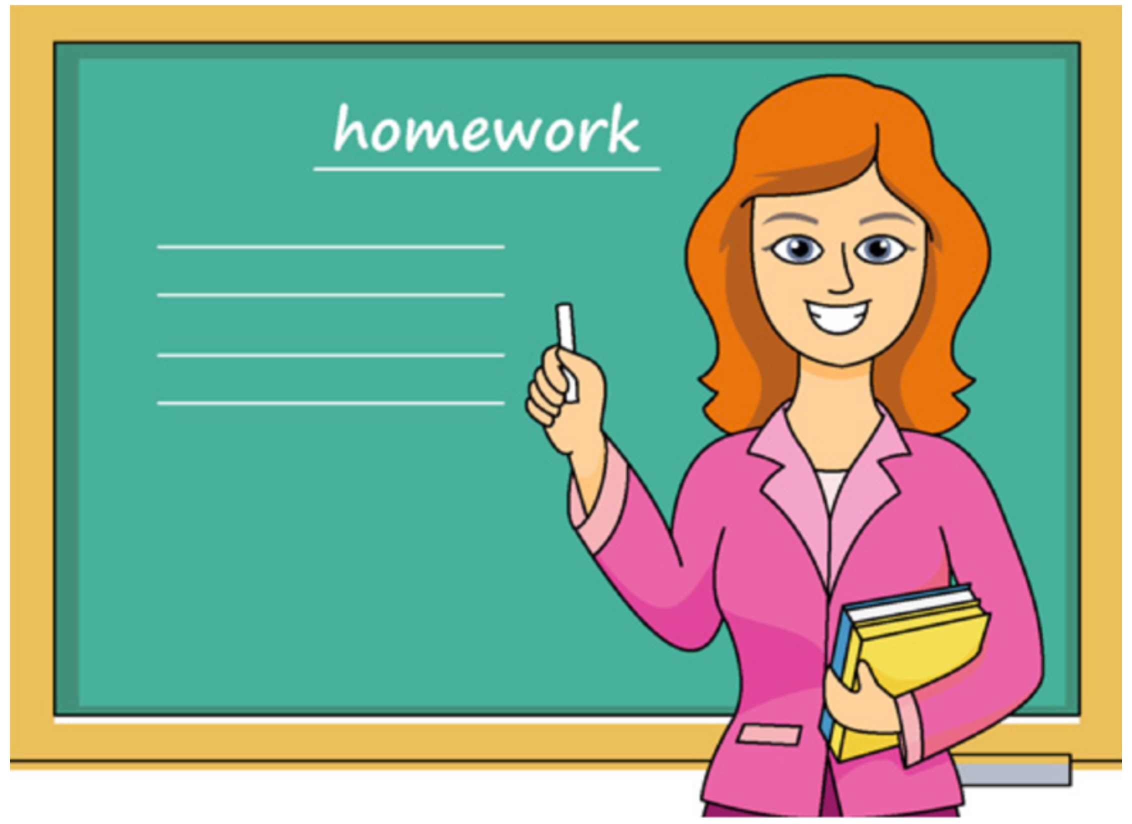
image X

label Y

Supervised computer vision

Hand-curated training data

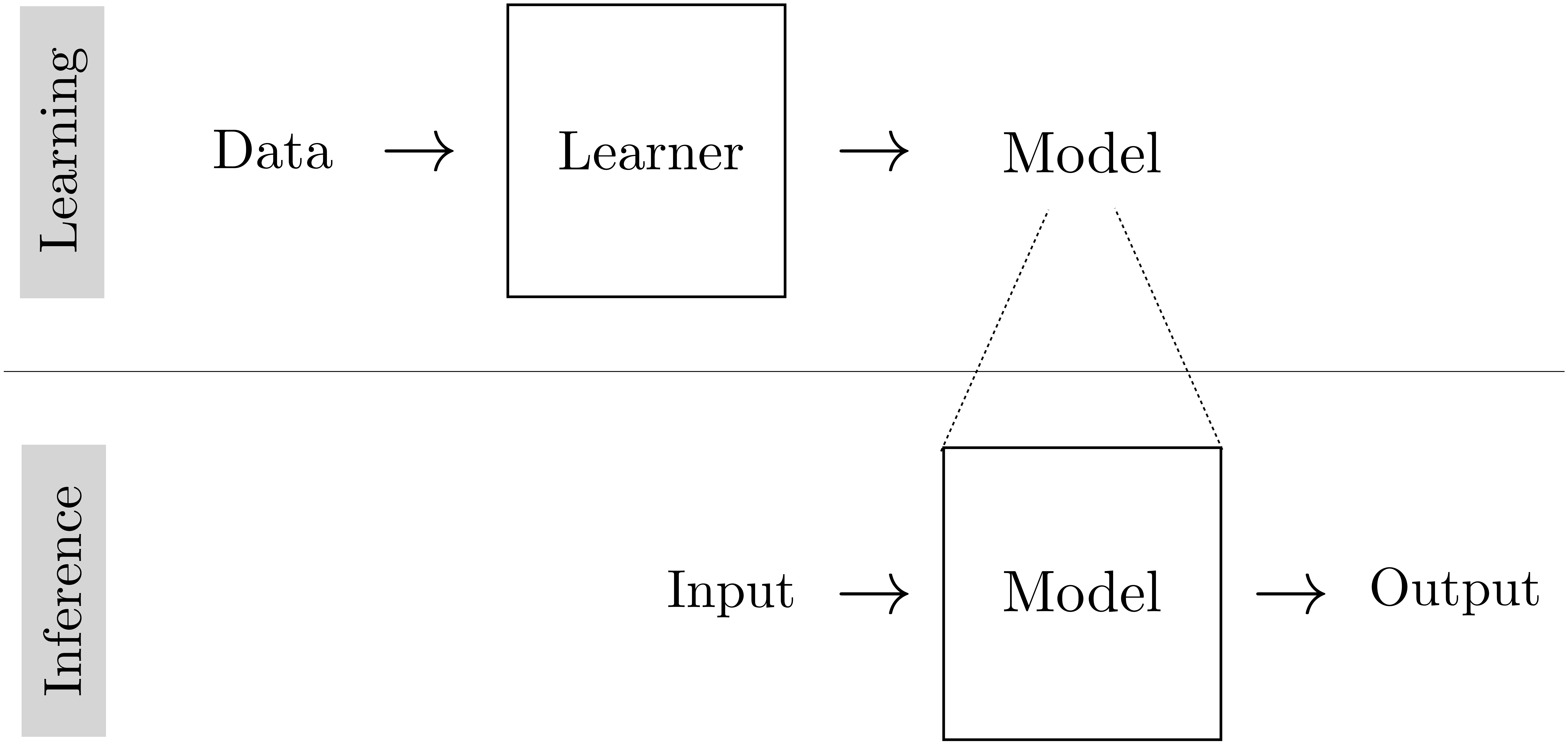
+ Informative
- Expensive
- Limited to teacher's knowledge



Vision in nature

Raw unlabeled training data

+ Cheap
- Noisy
- Harder to interpret


Learning
Data
Learner
Model
Inference
Input
Model
Output

Learning from examples

(aka **supervised learning**)

Training data

$$\{x_1, y_1\}$$
$$\{x_2, y_2\} \quad \rightarrow \quad \boxed{\text{Learner}} \quad \rightarrow \quad f : X \rightarrow Y$$
$$\{x_3, y_3\}$$
$$\dots$$

$$f^* = \arg\min_{f \in \mathcal{F}} \sum_{i=1}^{N} \mathcal{L}(f(x_i), y_i)$$

Learning without examples

(includes **unsupervised learning** and **reinforcement learning**)

Data

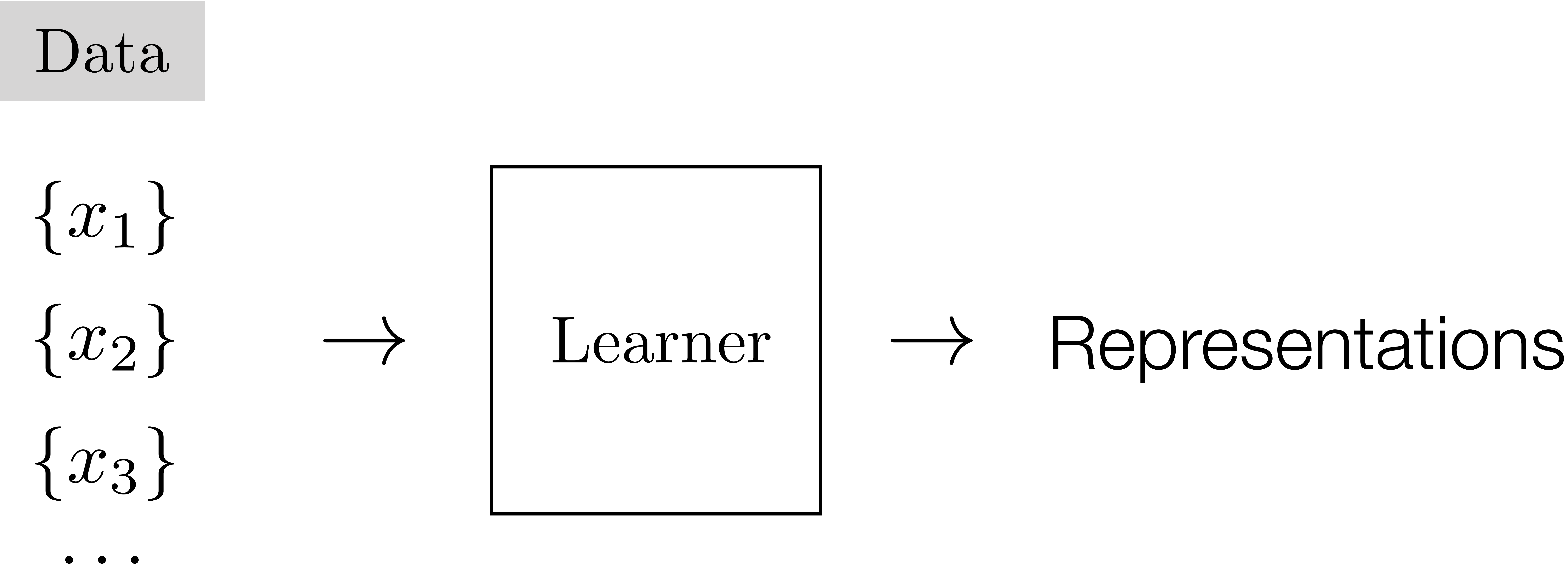
$\{x_1\}$

$\{x_2\}$ $\rightarrow$ Learner $\rightarrow$?

$\{x_3\}$

$\ldots$

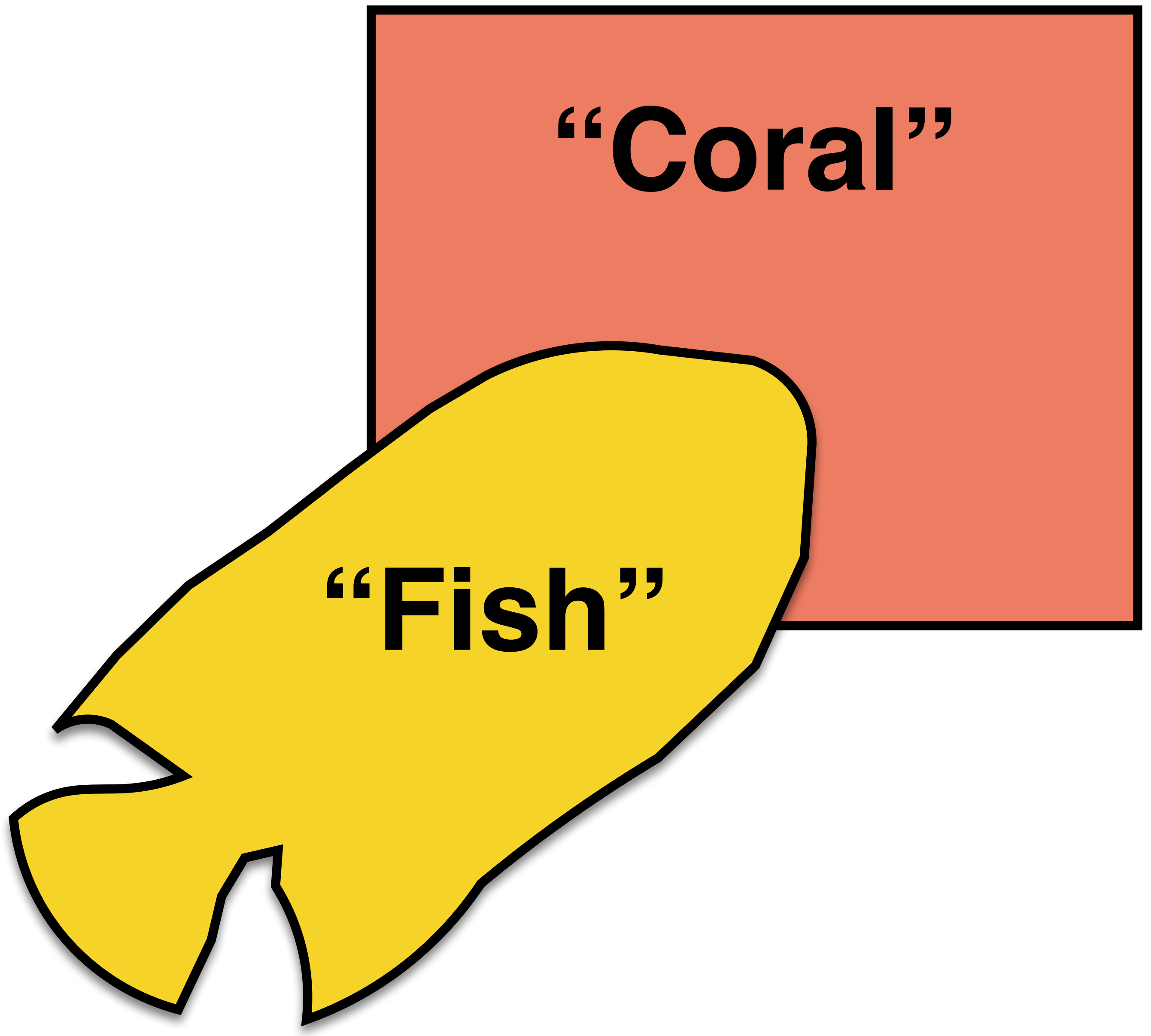
Unsupervised Representation Learning



Unsupervised Representation Learning

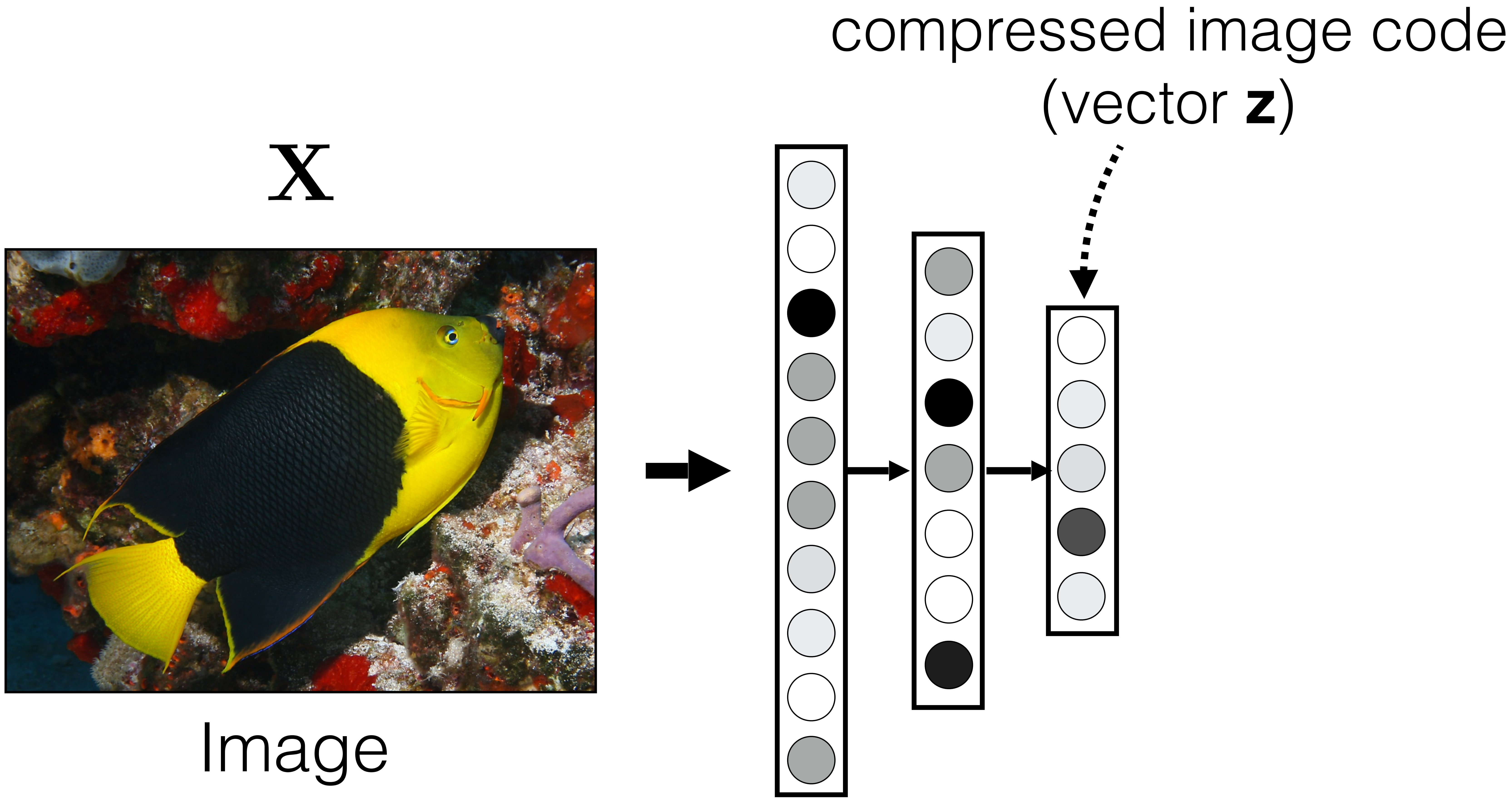
$\mathbf{x}$

Image



Compact mental representation

Unsupervised Representation Learning

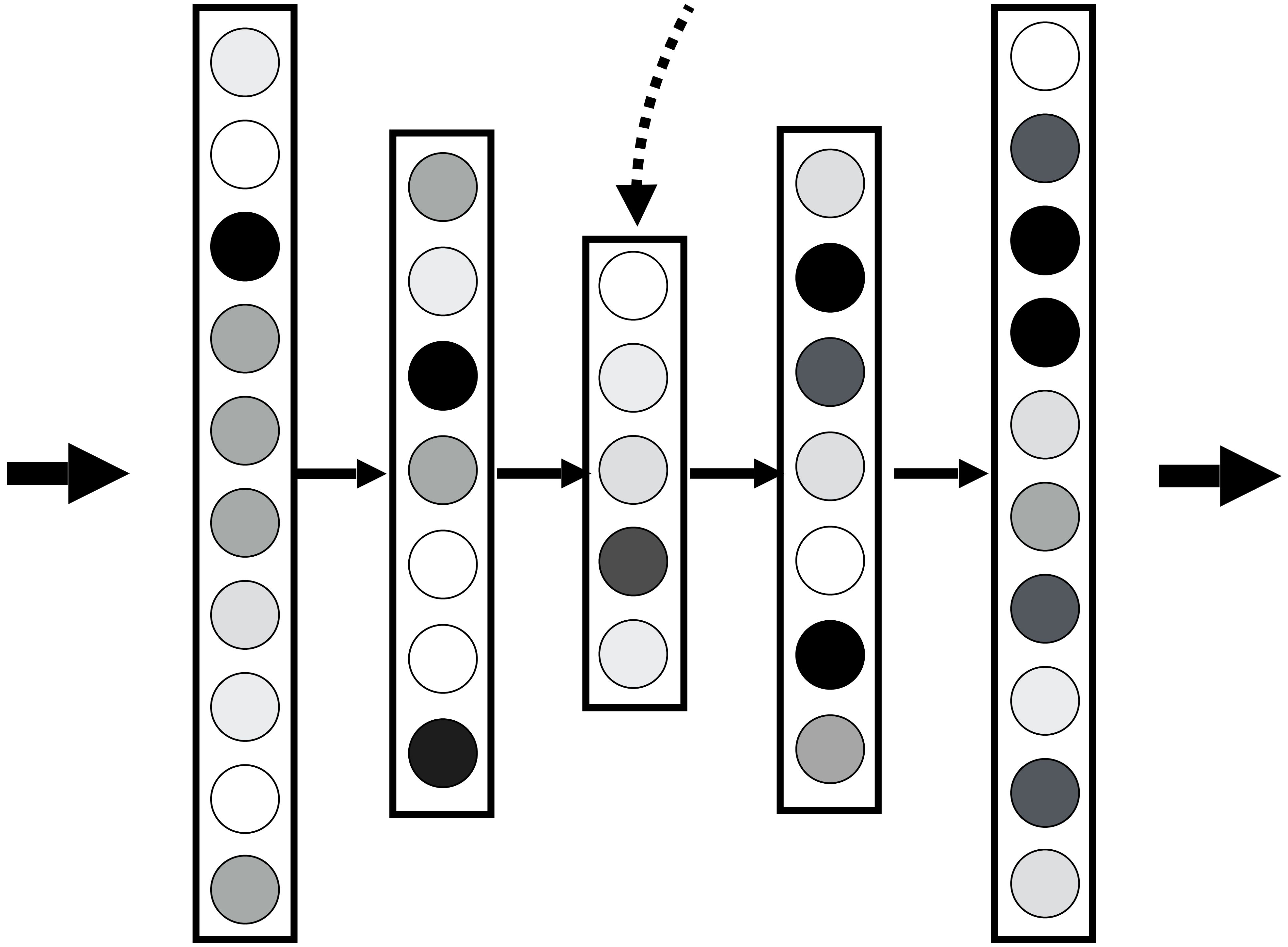


Unsupervised Representation Learning

$\mathbf{x}$

Image

compressed image code
(vector **z**)



“Autoencoder”

$\hat{\mathbf{x}}$

Reconstructed image

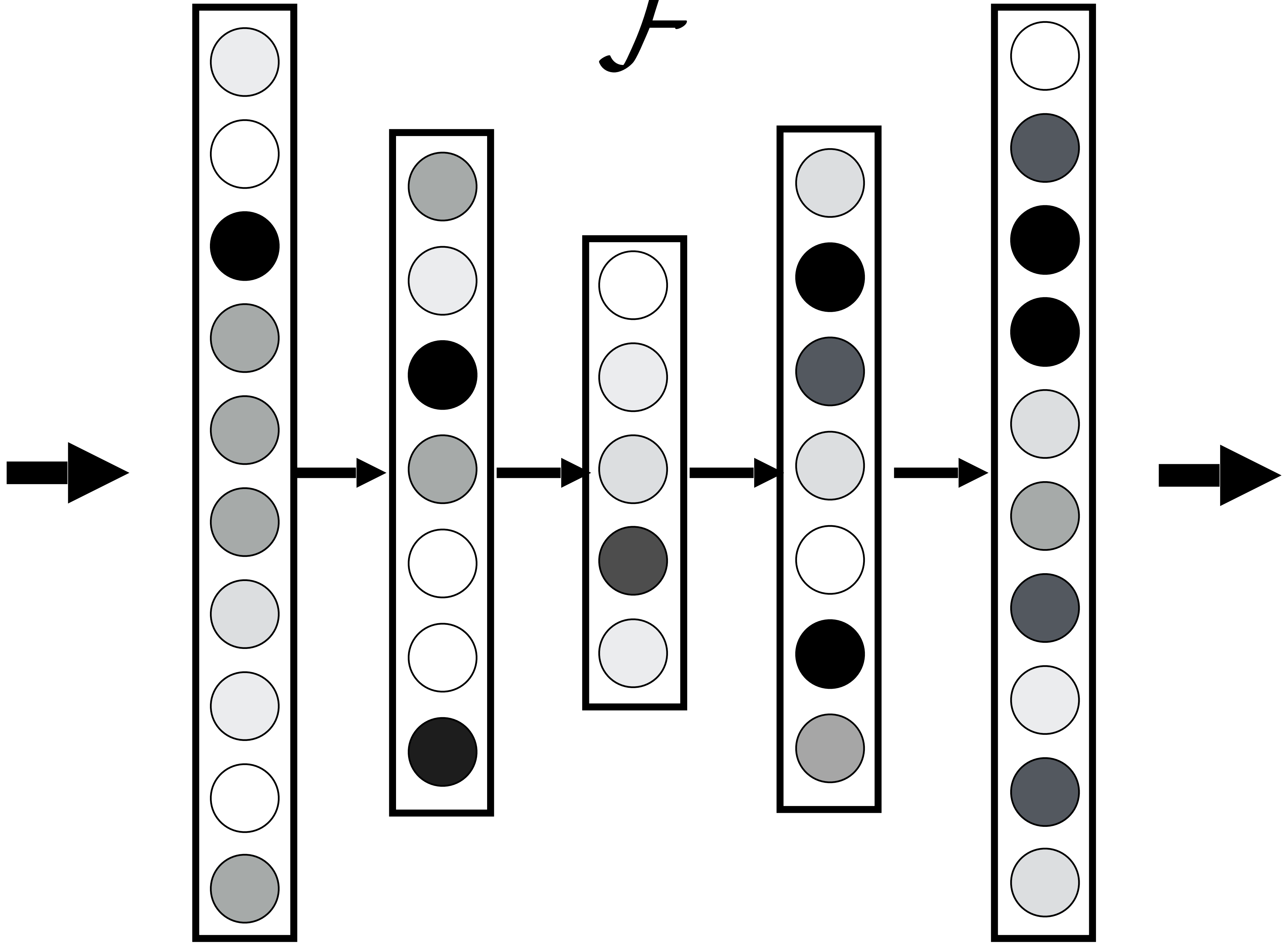
[e.g., Hinton & Salakhutdinov, Science 2006]

Autoencoder

$\mathbf{X}$

Image

$\mathcal{F}$

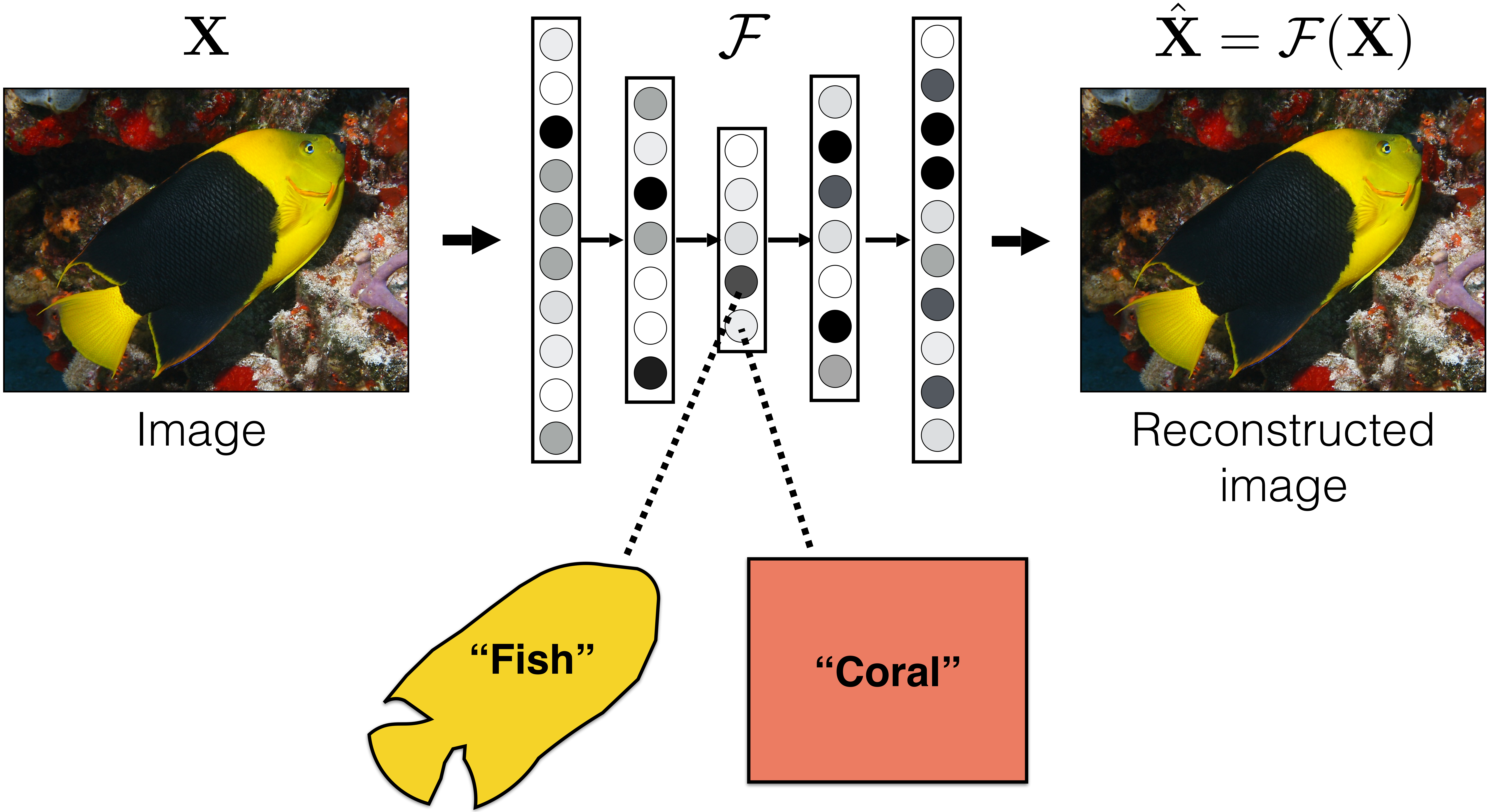


$$\arg\min_{\mathcal{F}} \mathbb{E}_{\mathbf{X}}[||\mathcal{F}(\mathbf{X}) - \mathbf{X}||]$$

$\hat{\mathbf{X}} = \mathcal{F}(\mathbf{X})$

Reconstructed image

[e.g., Hinton & Salakhutdinov, Science 2006]



[e.g., Hinton & Salakhutdinov, Science 2006]

Autoencoder

Data

$\{\mathbf{x}_i\}_{i=1}^N \quad \rightarrow$

Learner

Objective

$$\mathcal{L}(f(\mathbf{x}), \mathbf{x}) = \|f(\mathbf{x}) - \mathbf{x}\|_2^2$$

Hypothesis space

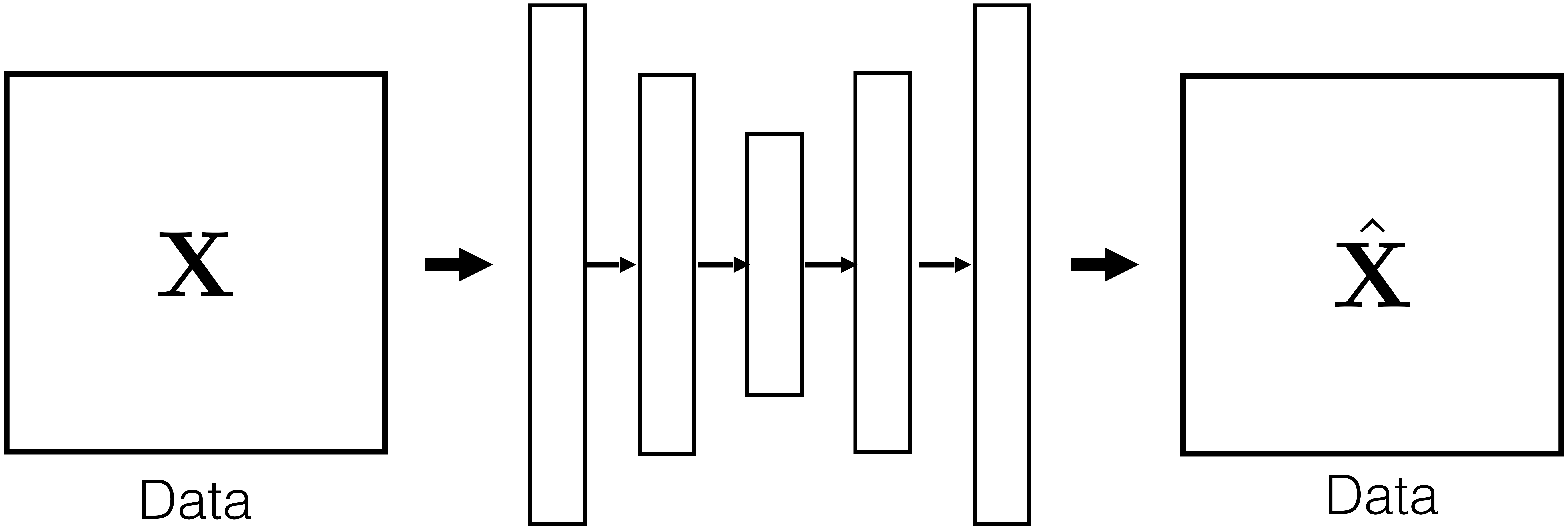
CNN

Optimizer

SGD

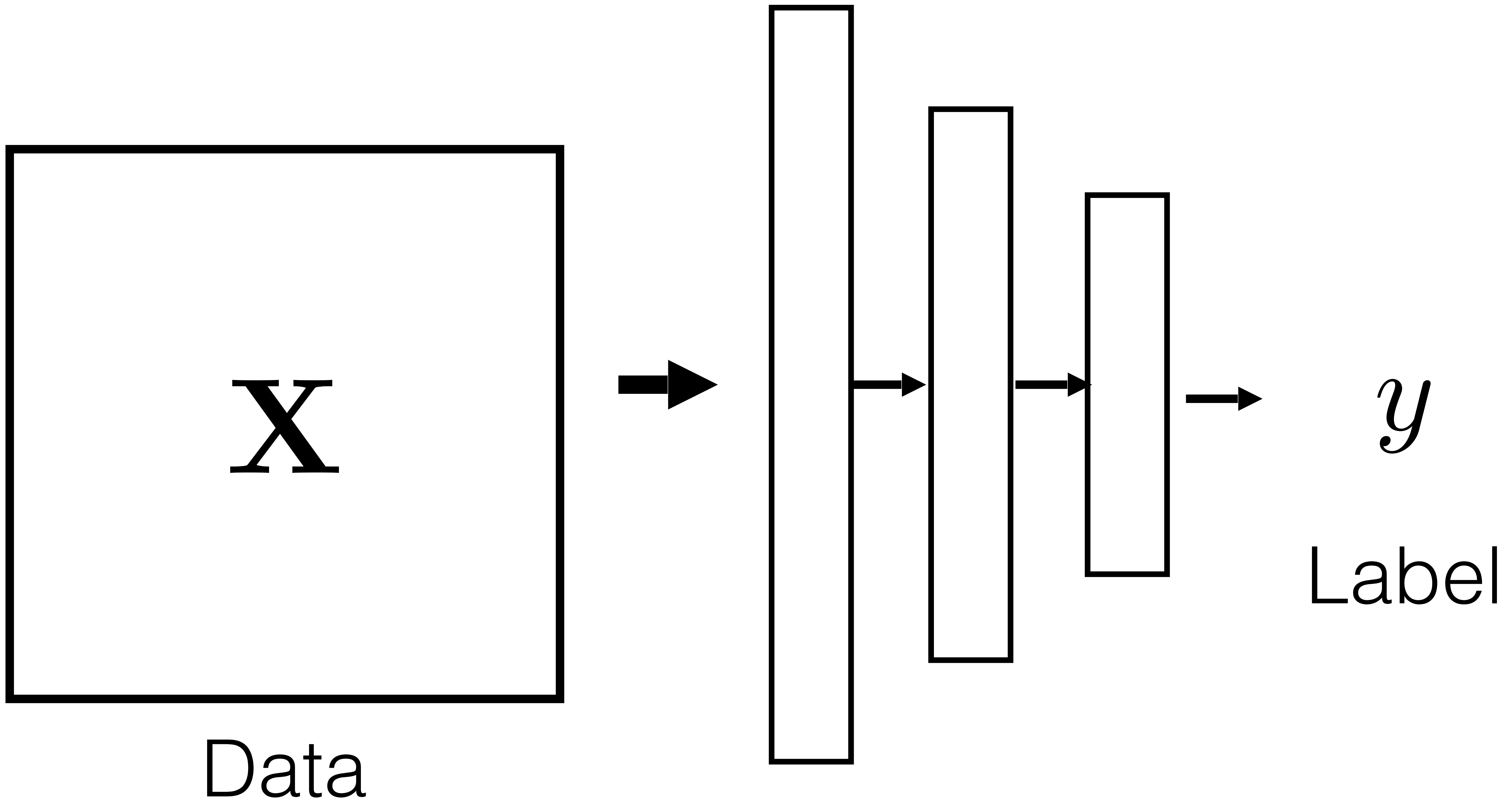
$\rightarrow f$

Data compression



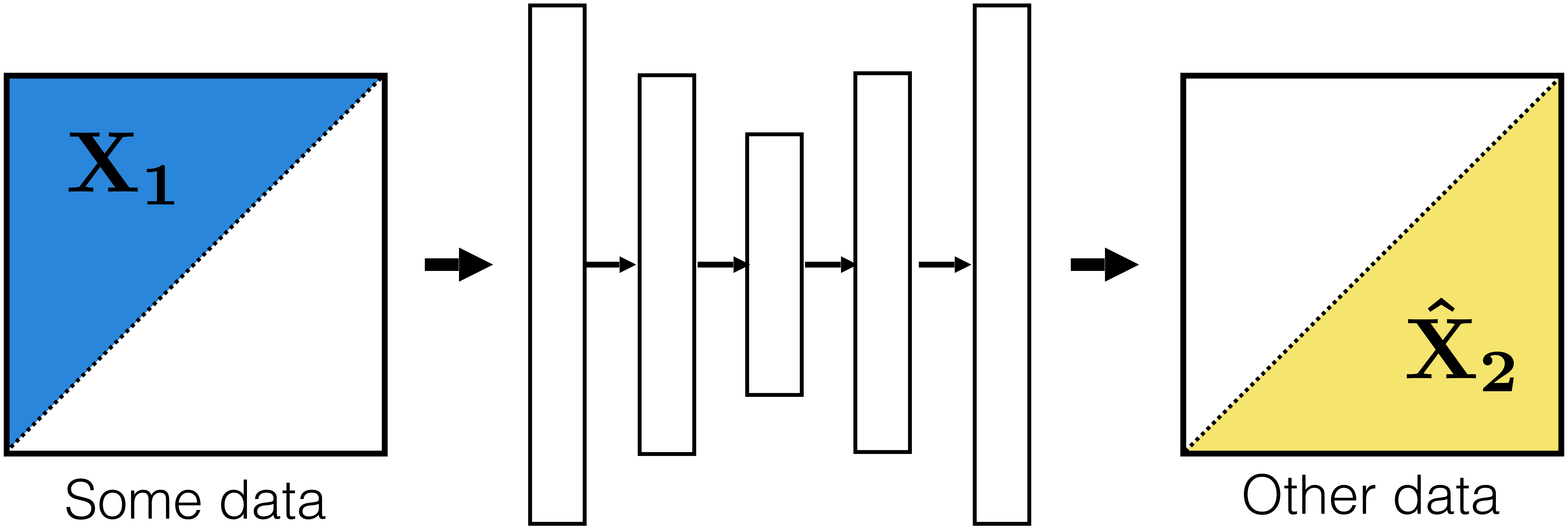
[Hinton & Salakhutdinov, Science 2009]

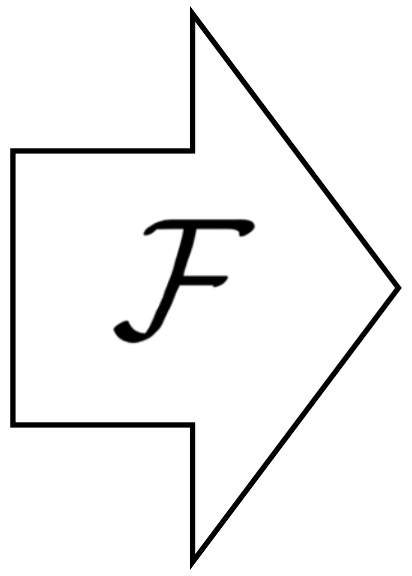
Label prediction



e.g., image classification

Data prediction



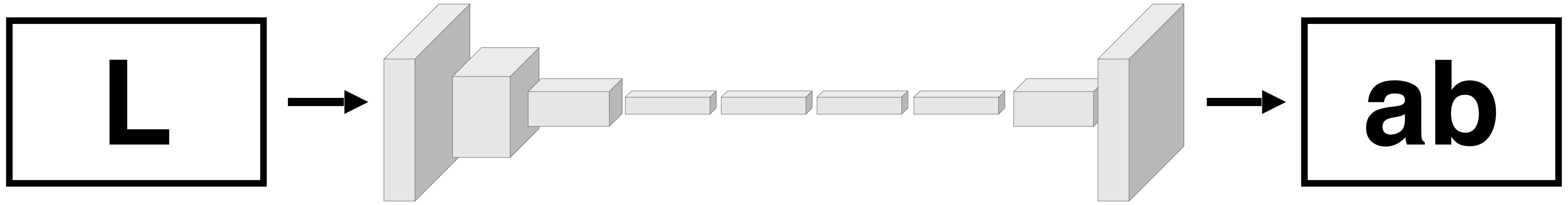


Grayscale image: L channel

$\mathbf{X} \in \mathbb{R}^{H \times W \times 1}$

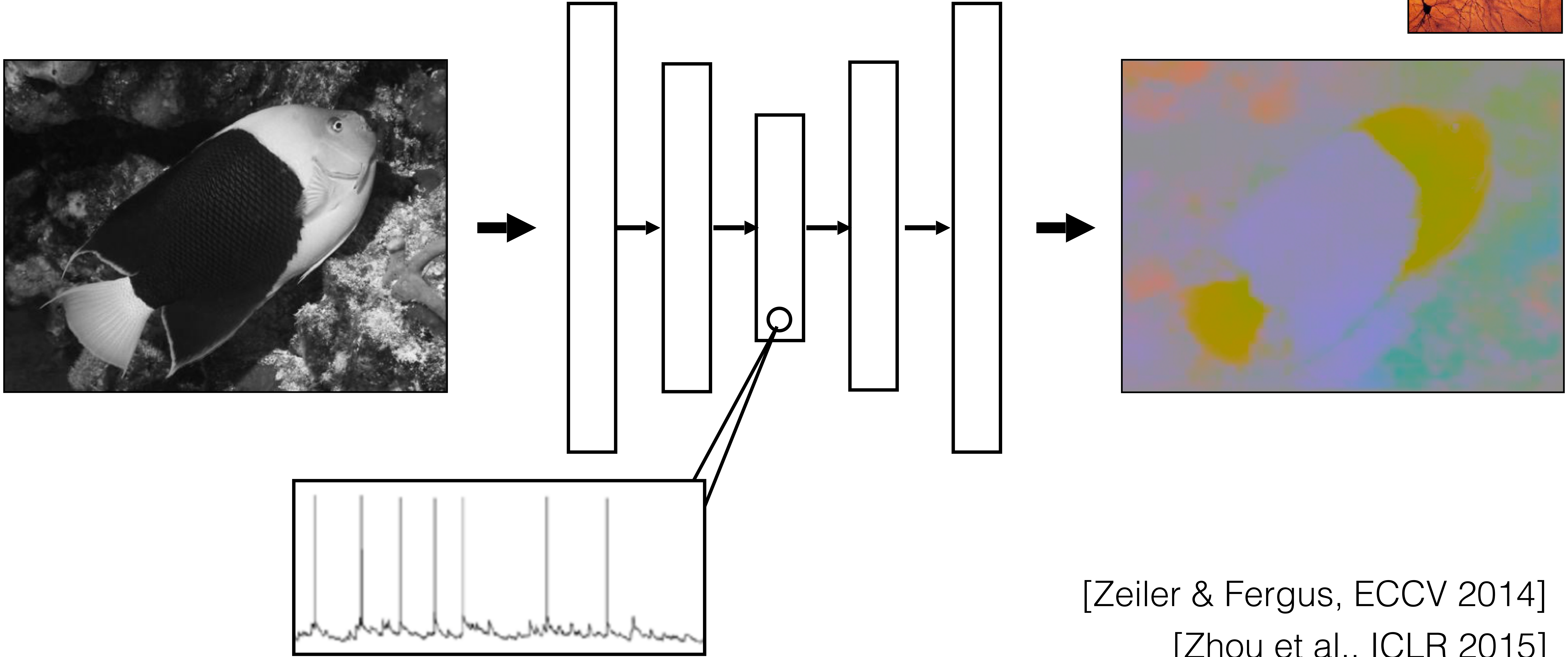
Color information: ab channels

$\widehat{\mathbf{Y}} \in \mathbb{R}^{H \times W \times 2}$



[Zhang, Isola, Efros, ECCV 2016]

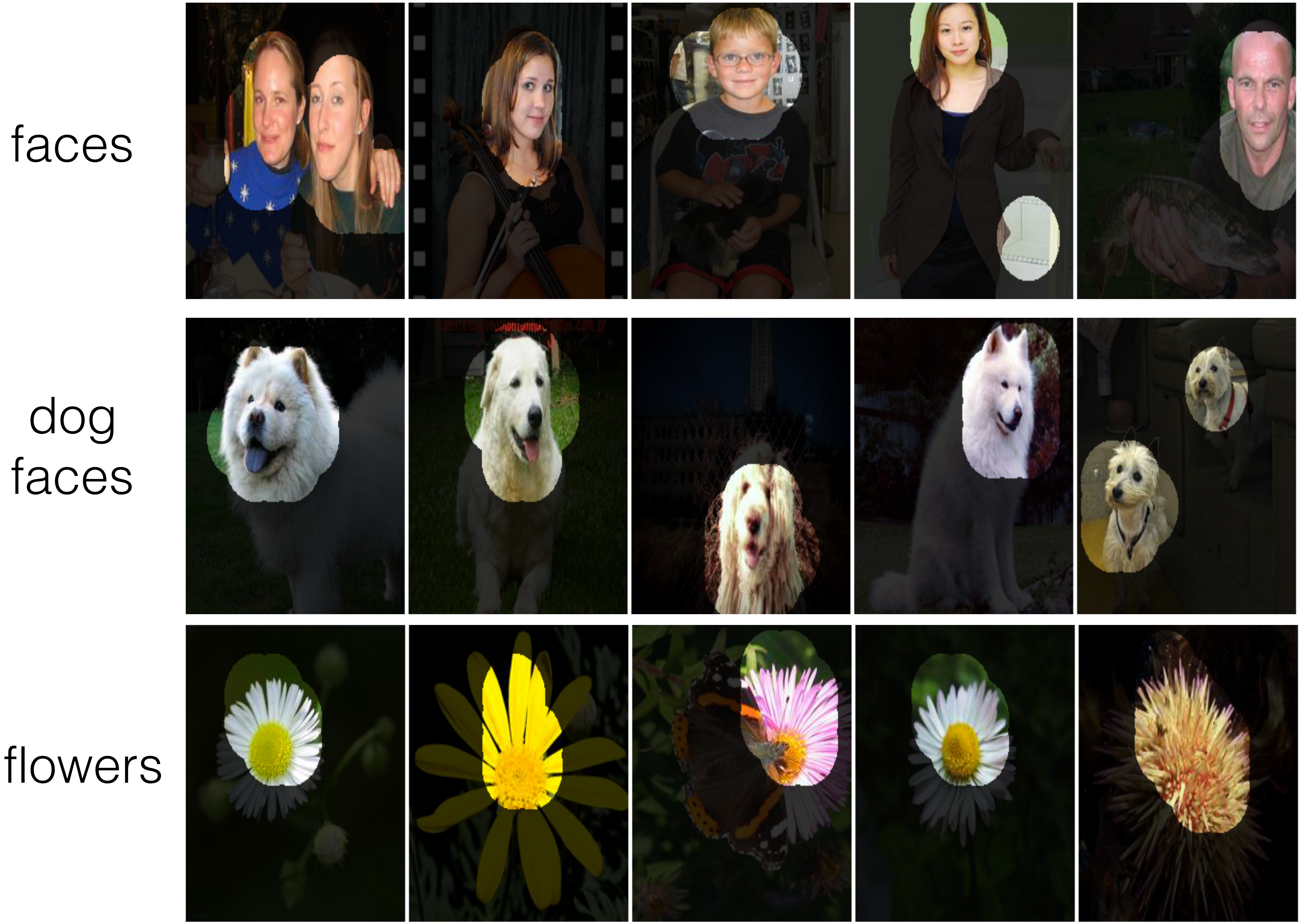
Deep Net "Electrophysiology"

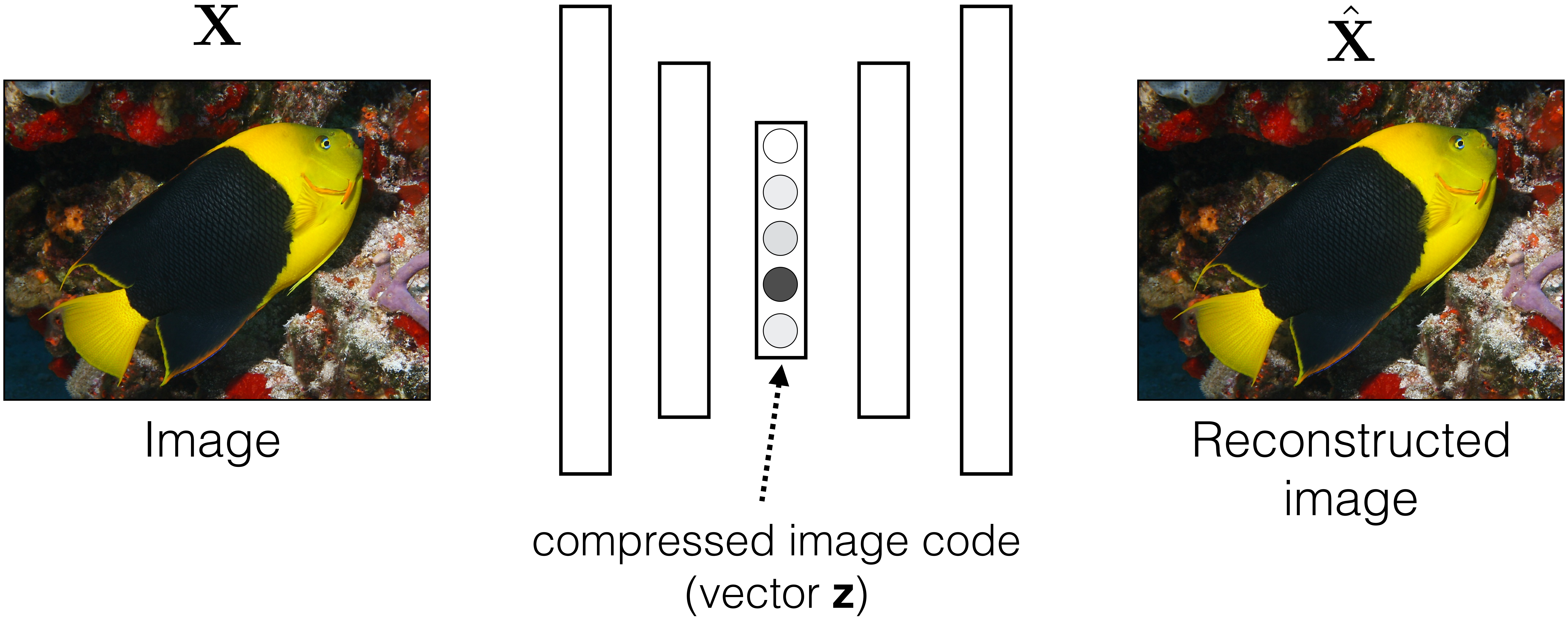


[Zeiler & Fergus, ECCV 2014]

[Zhou et al., ICLR 2015]

Stimuli that drive selected neurons (conv5 layer)

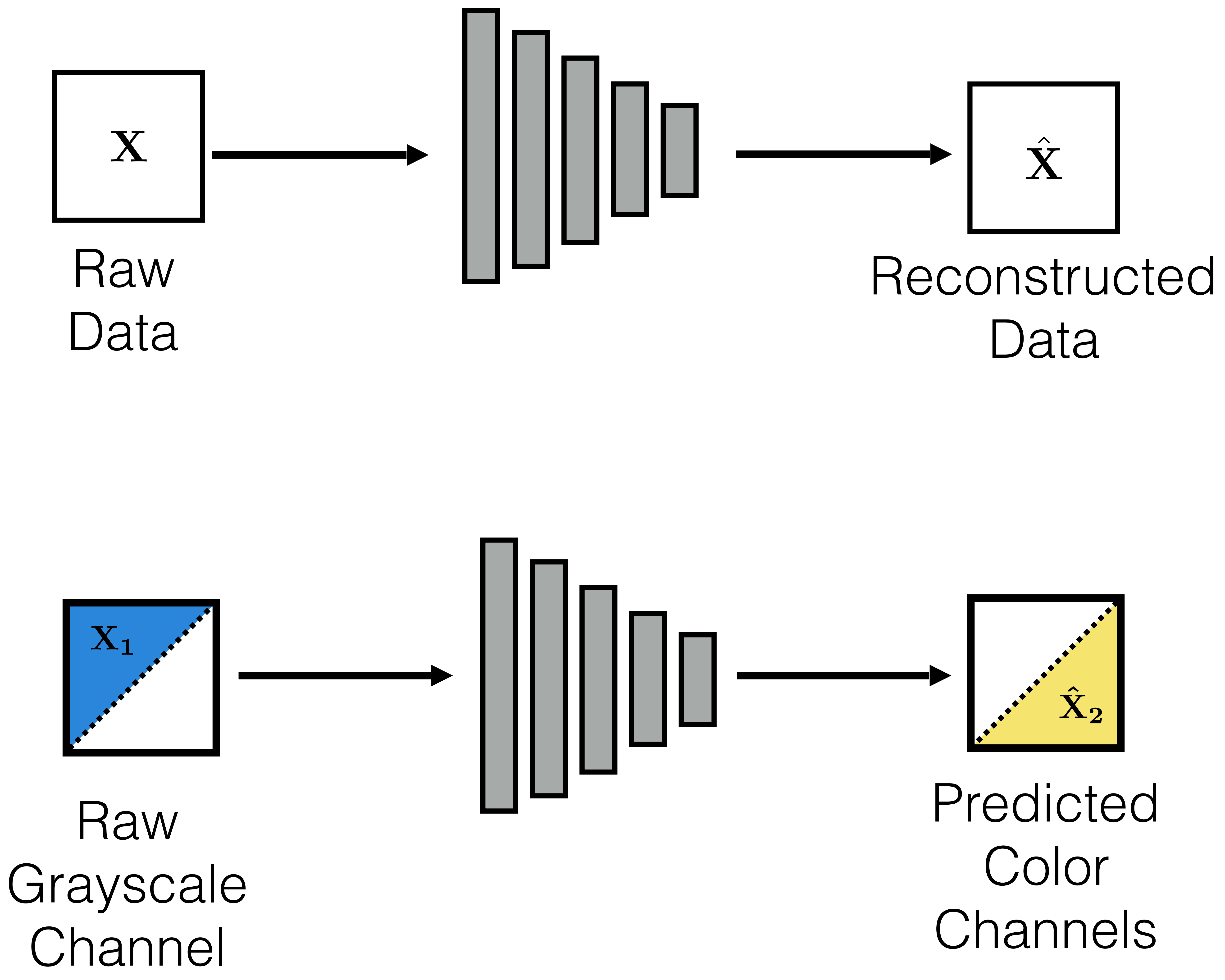


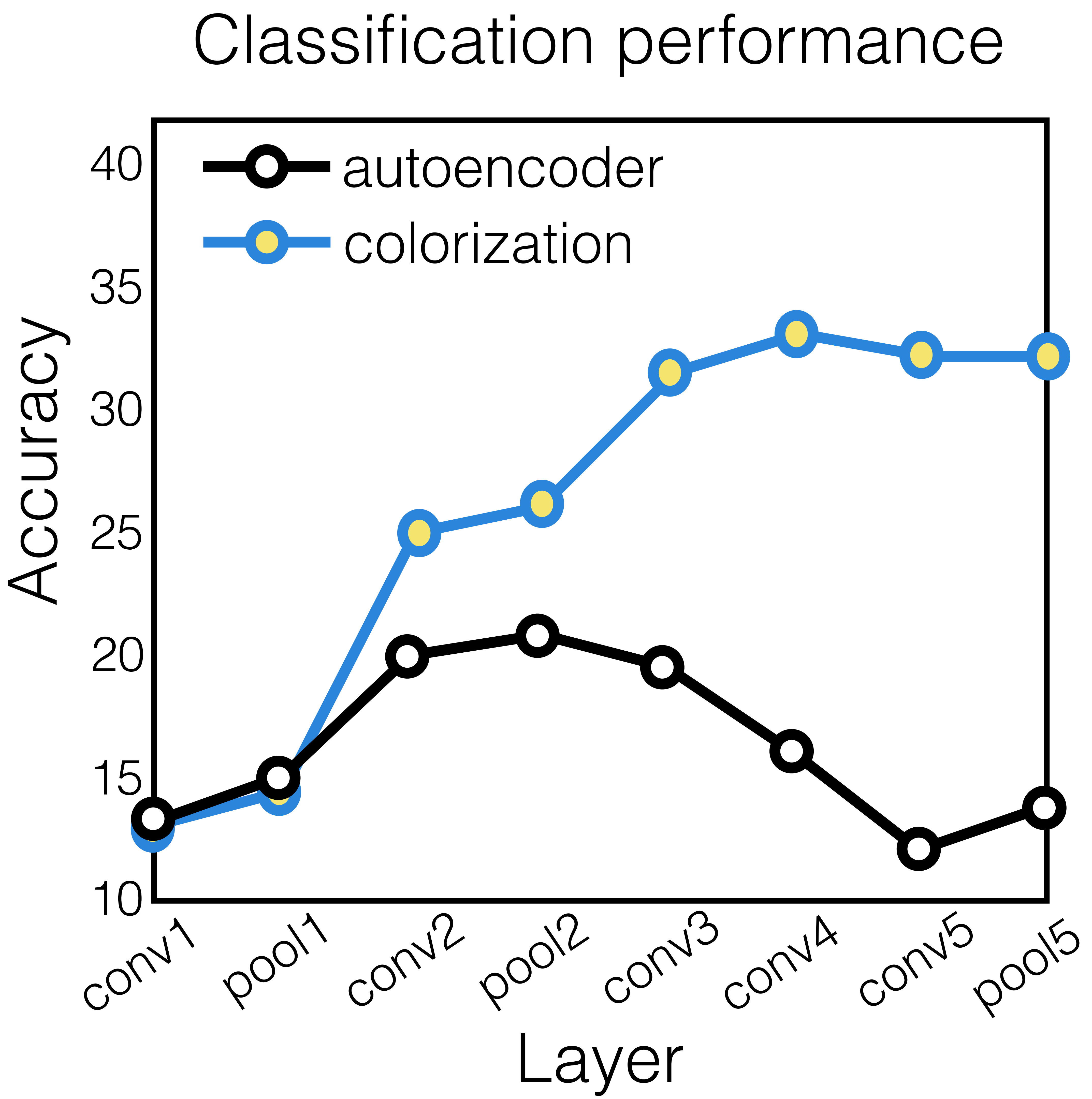


Is the code informative about object class y?

Logistic regression:

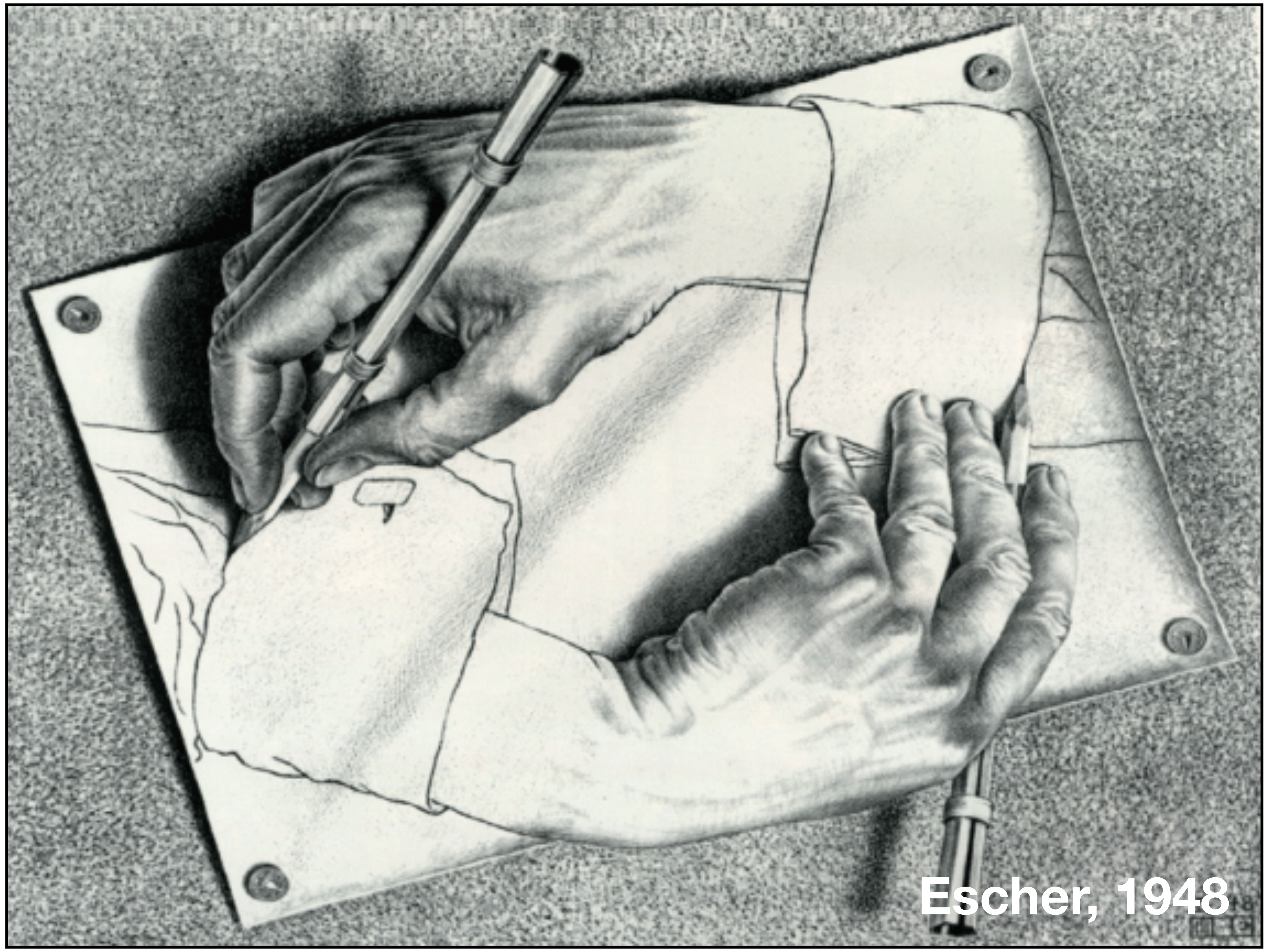
$$y = \sigma(\mathbf{Wz} + \mathbf{b})$$





Task from [Russakovsky et al. 2015]

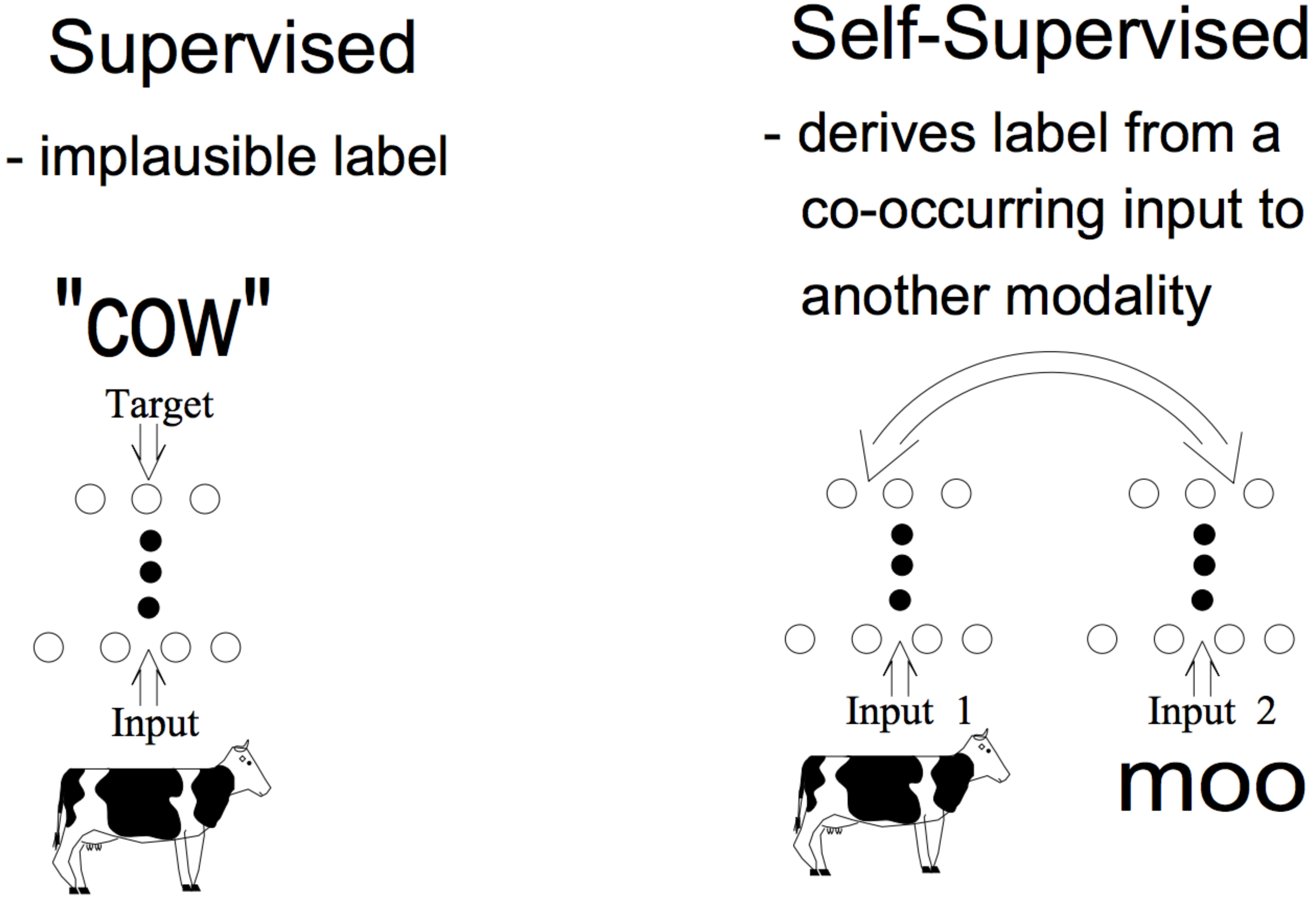
Self-supervised learning



Common trick:

- Convert “unsupervised” problem into “supervised” empirical risk minimization
- Do so by cooking up “labels” (prediction targets) from the raw data itself

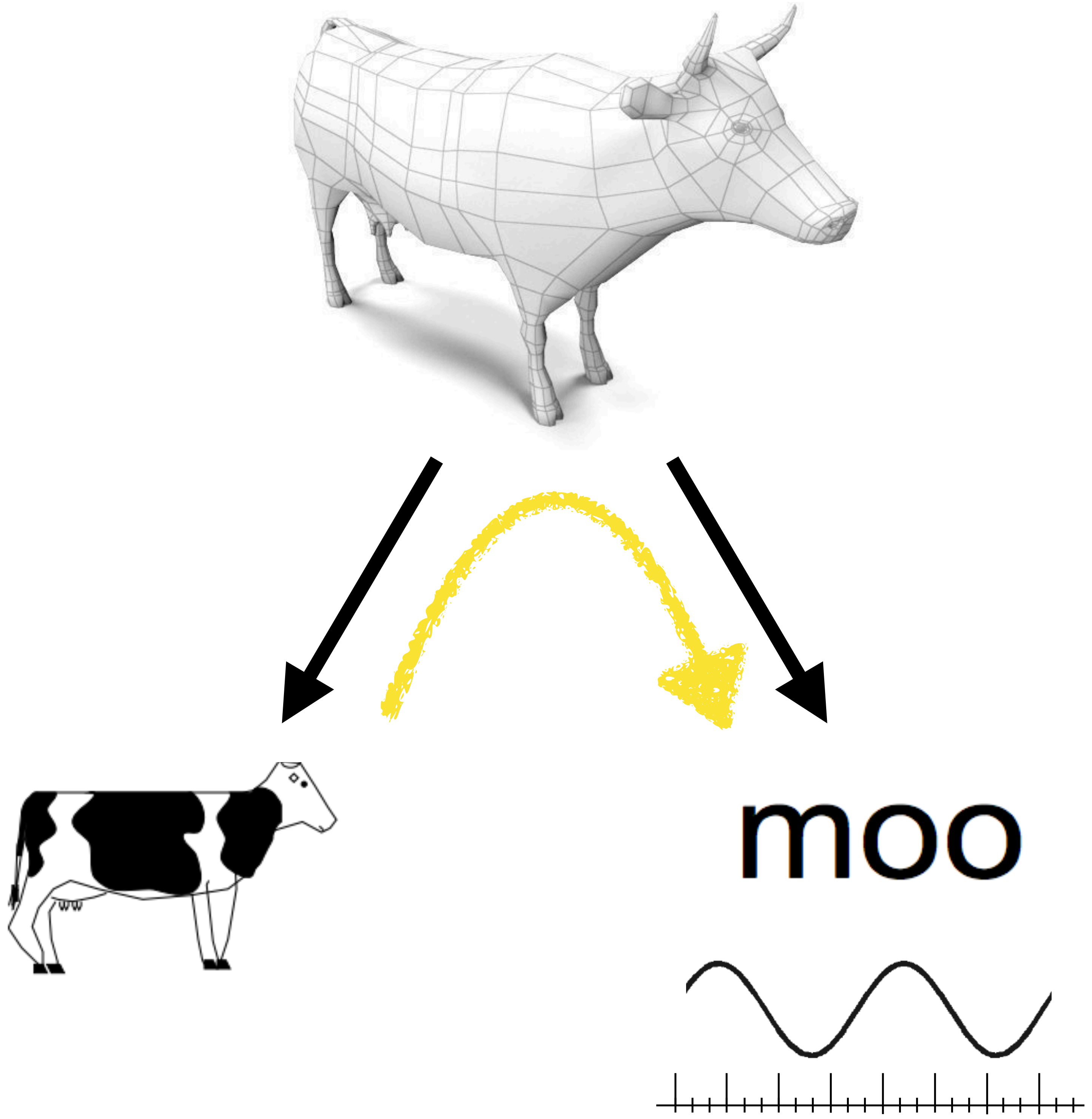
Multisensory self-supervision



Virginia de Sa. *Learning Classification with Unlabeled Data*. NIPS 1994.

[see also "Six lessons from babies", Smith and Gasser 2005]

The allegory of the cave



“Multiview” self-supervised learning

Data

$\{\mathbf{x}_i\}_{i=1}^N \quad \rightarrow$

Learner

Objective

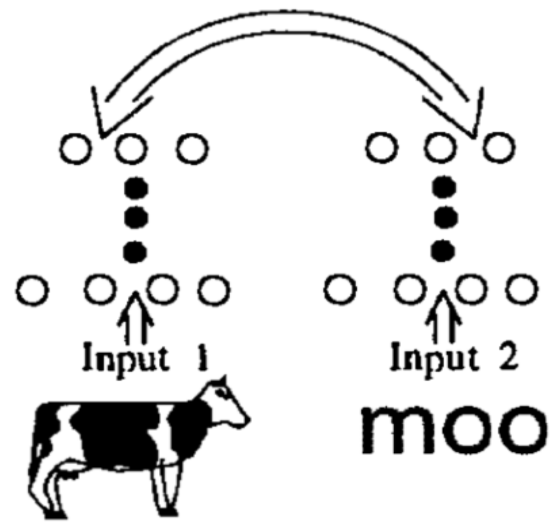
$$\arg\min_f \sum_i D(f(g(\mathbf{x}_i)) - h(\mathbf{x}_i))$$

$\rightarrow f$

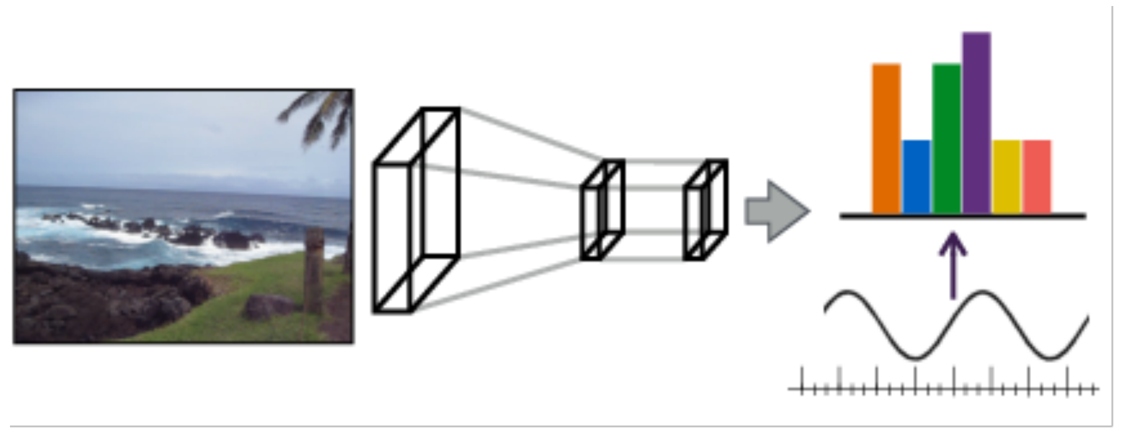
Distance function

g and h are two “views” of the data x,
e.g., two different image channels

Audio

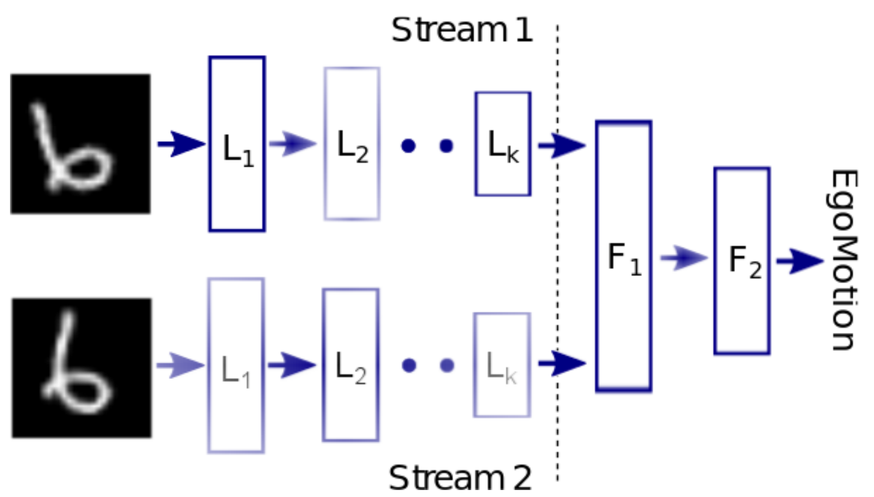


de Sa. NIPS 1994.



Owens et al. ECCV 2016.

Egomotion

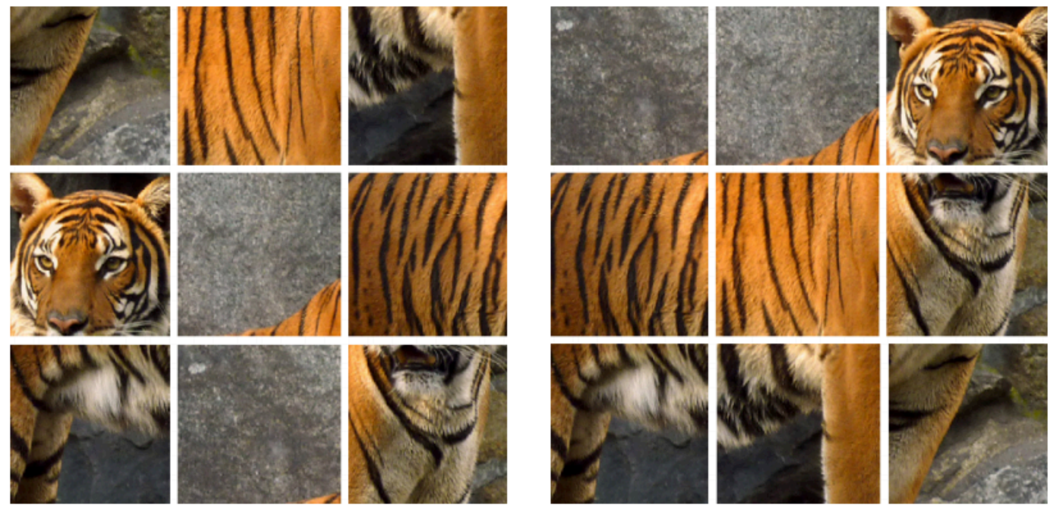


Agrawal et al. ICCV 2015.

Jayaraman et al. ICCV 2015.

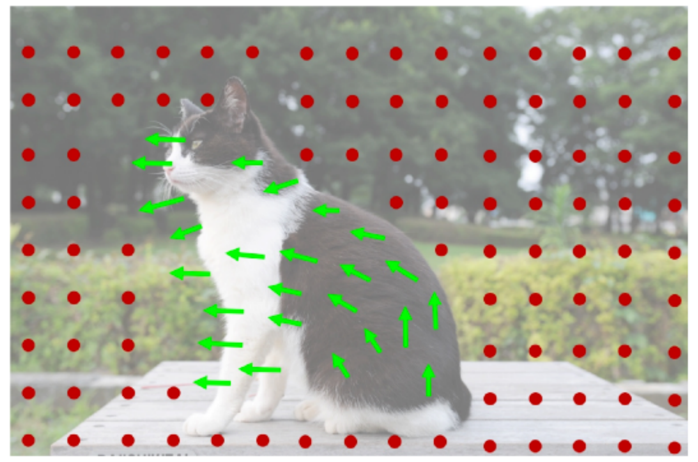
Context

Pathak et al. CVPR 2016.



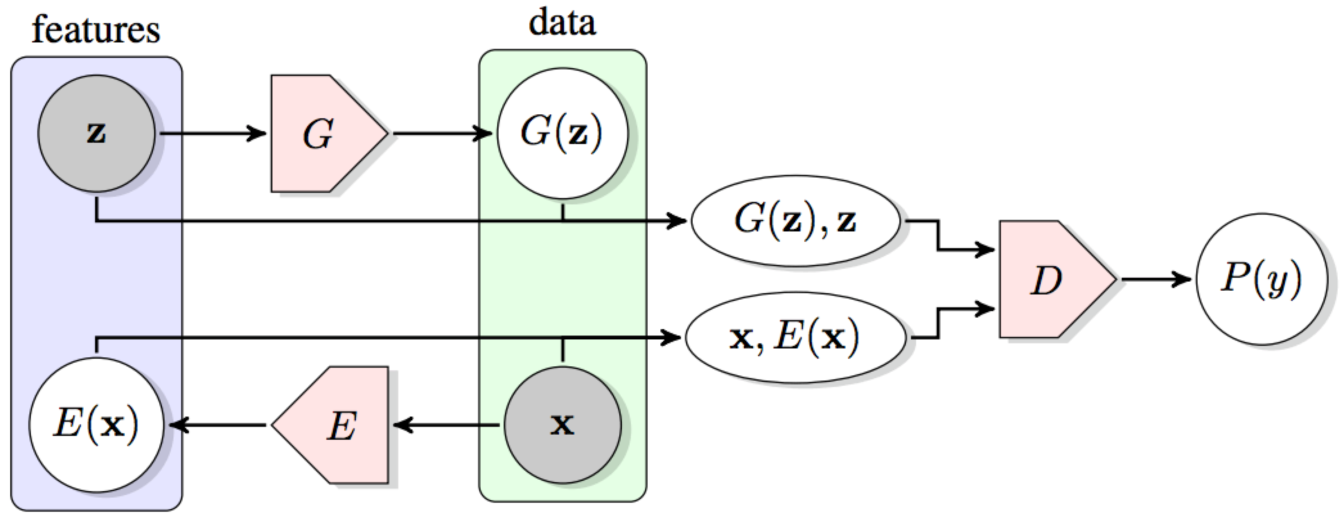
Noroozi and Favaro. ECCV 2016.
Doersch et al. ICCV 2015.

Video



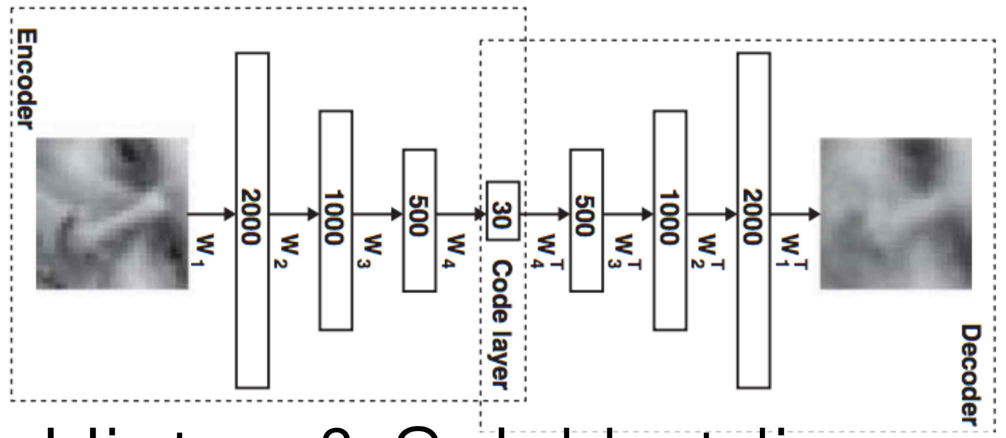
Wang et al. ICCV 2015. Pathak et al. CVPR 2017.
Misra et al. ECCV 2016.

Generative Modeling



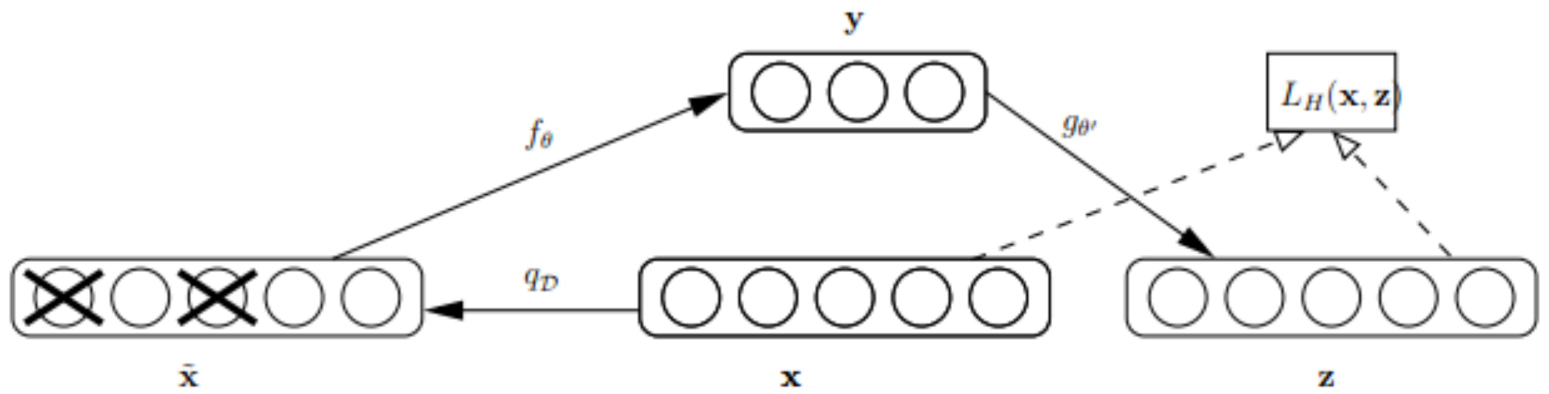
Donahue et al. Dumoulin et al. ICLR 2017.

Autoencoders



Hinton & Salakhutdinov. Science 2006.

Denoising Autoencoders



Vincent et al. ICML 2008.

Goal: Set up a pre-training scheme to induce a “useful” representation

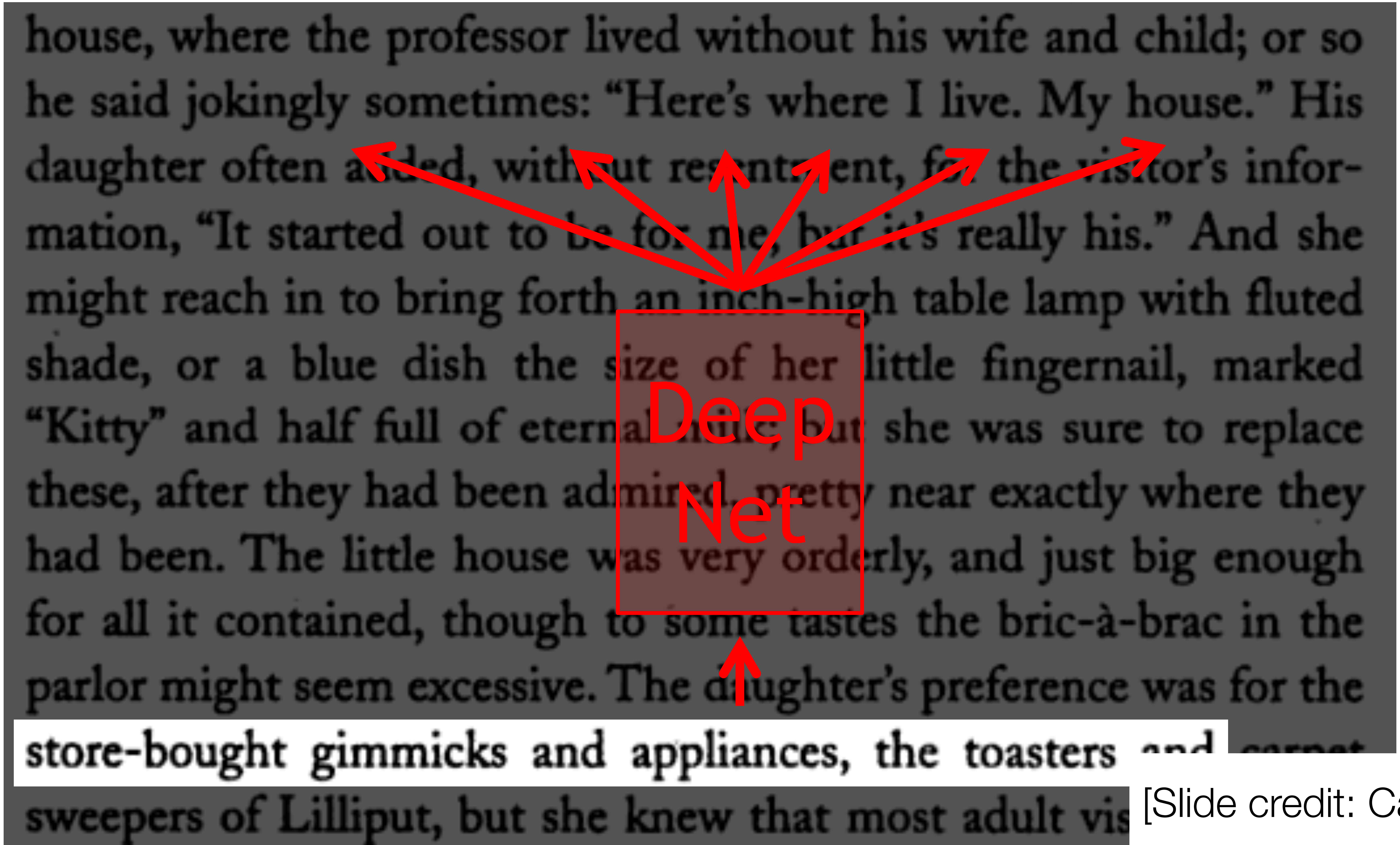
[Slide credit: Richard Zhang]

Unsupervised visual representation learning by context prediction

[Carl Doersch, Abhinav Gupta, Alexei A. Efros, ICCV 2015]

Context as Supervision

[Collobert & Weston 2008; Mikolov et al. 2013]

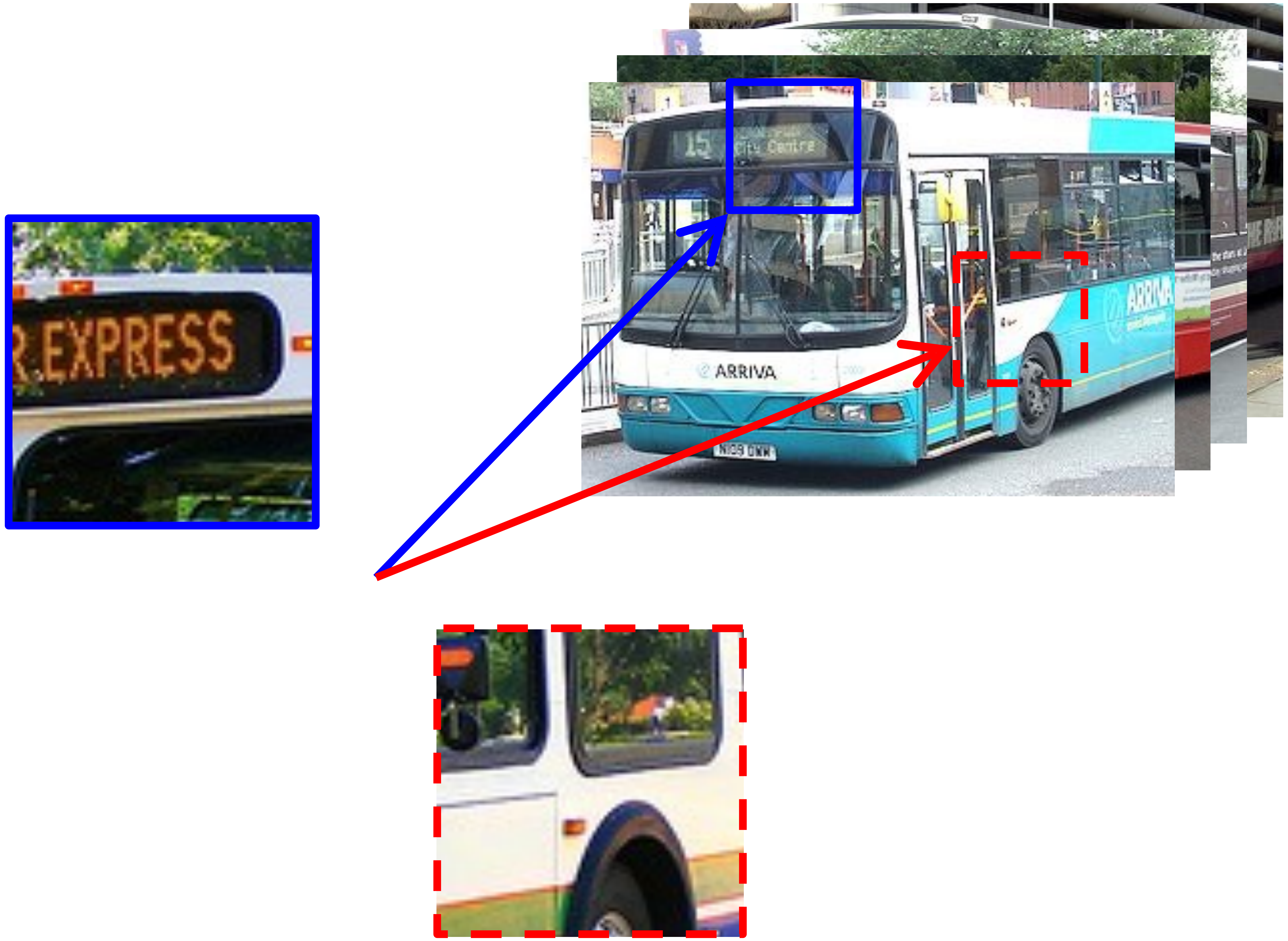


[Slide credit: Carl Doersch]

Context Prediction for Images

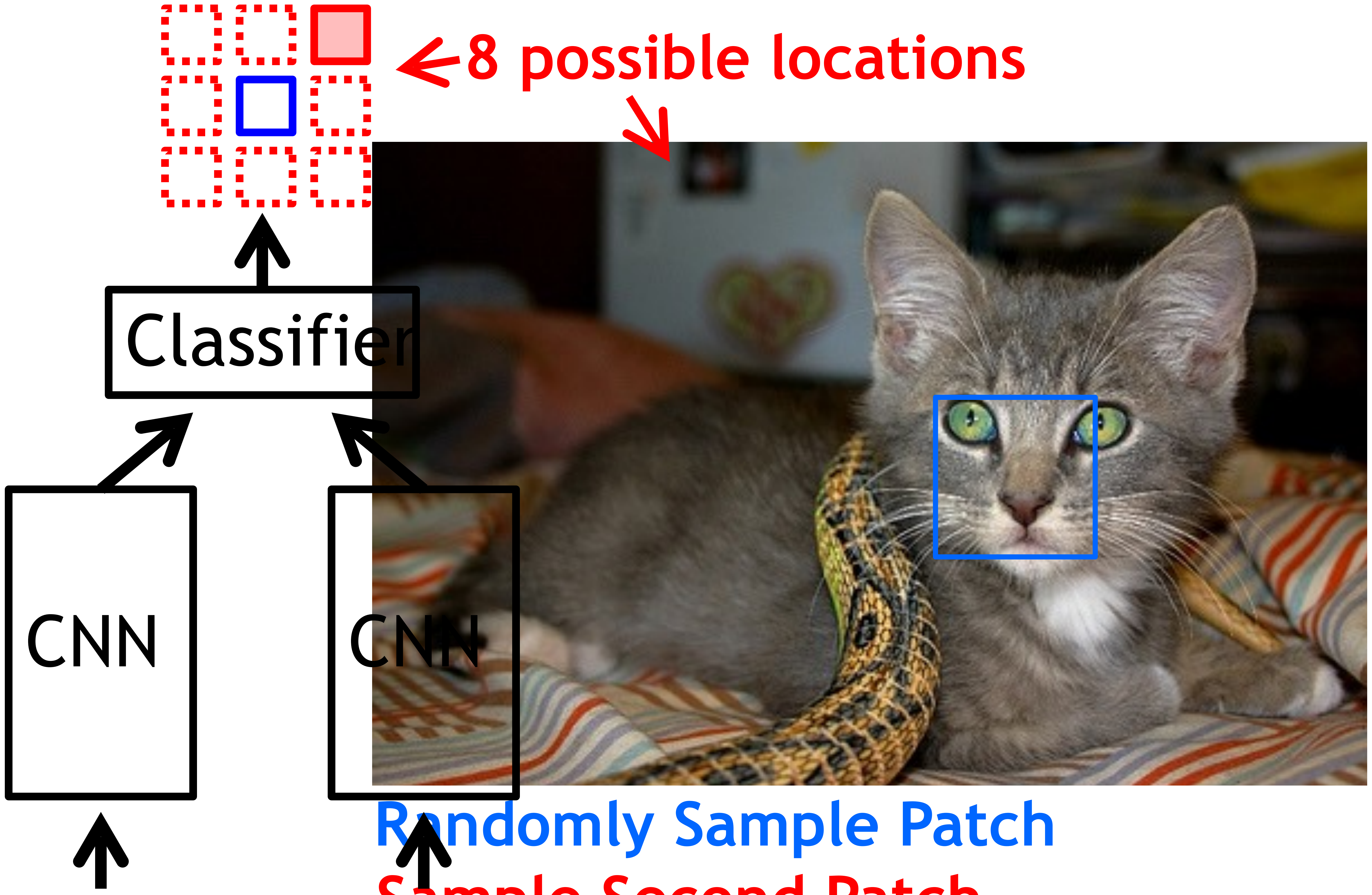
[Slide credit: Carl Doersch]

Semantics from a non-semantic task

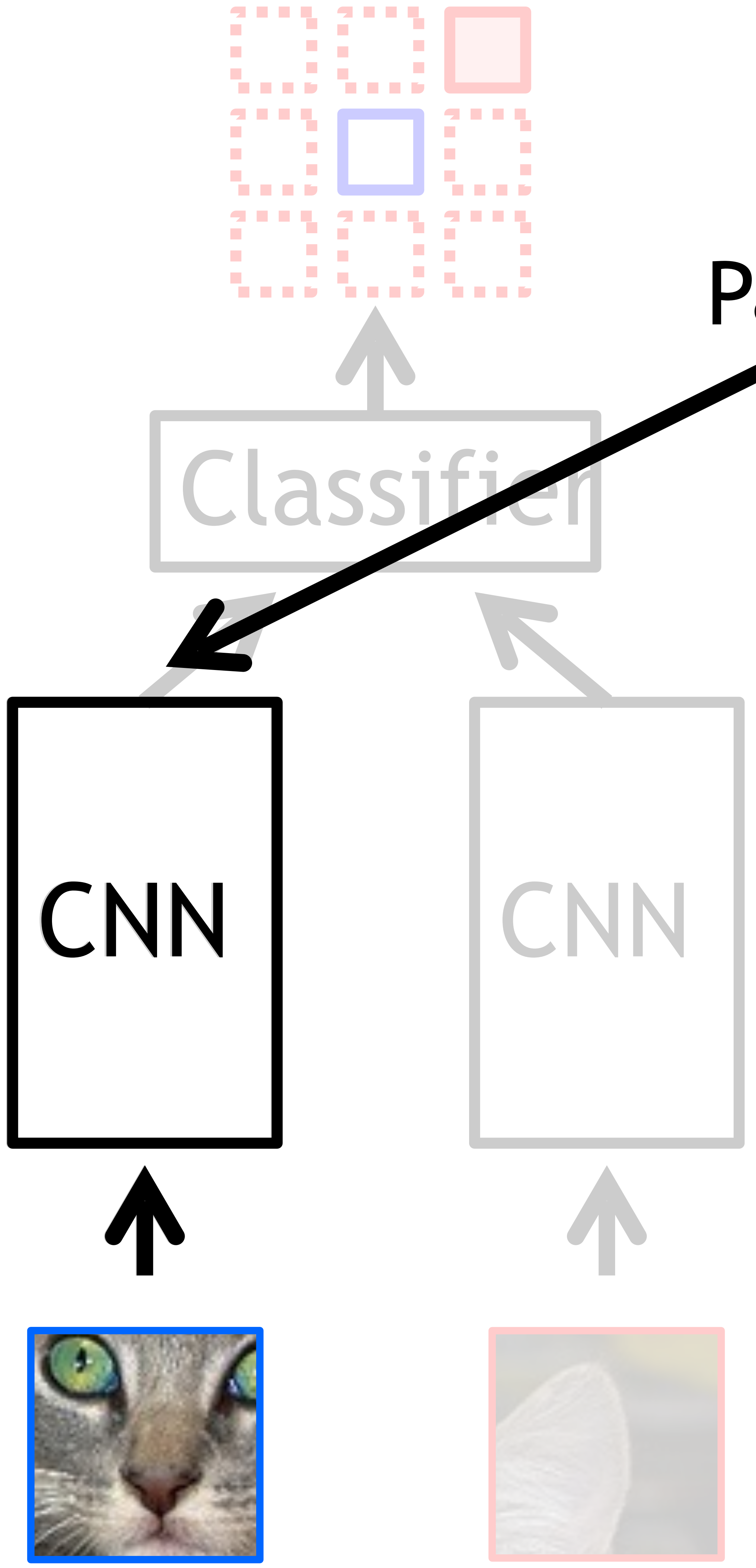


[Slide credit: Carl Doersch]

Relative Position Task



[Slide credit: Carl Doersch]



Patch Embedding (representation)

Input | Nearest Neighbors

Note: connects ***across*** instances!

[Slide credit: Carl Doersch]

Prediction hypothesis

Henri Cartier-Bresson

1. To survive, biological agents are constantly trying to anticipate, to predict sensations

2. This trains up representations useful for prediction — surfaces, objects, events!

Summary

1. Deep nets learn *representations*, just like our brains do

2. This is useful because representations transfer — they act as prior knowledge that enables quick learning on new tasks

3. Representations can also be learned without labels, which is great since labels are expensive and limiting

4. Without labels there are many ways to learn representations. We saw:

    1. representations as compressed codes

    2. representations that are shared across sensory modalities

    3. representations that are predictive of the future

Yann LeCun’s cake:

1. Cake is unsupervised representation learning

2. Frosting is supervised transfer learning

3. Cherry on top is reinforcement learning (model-based RL)