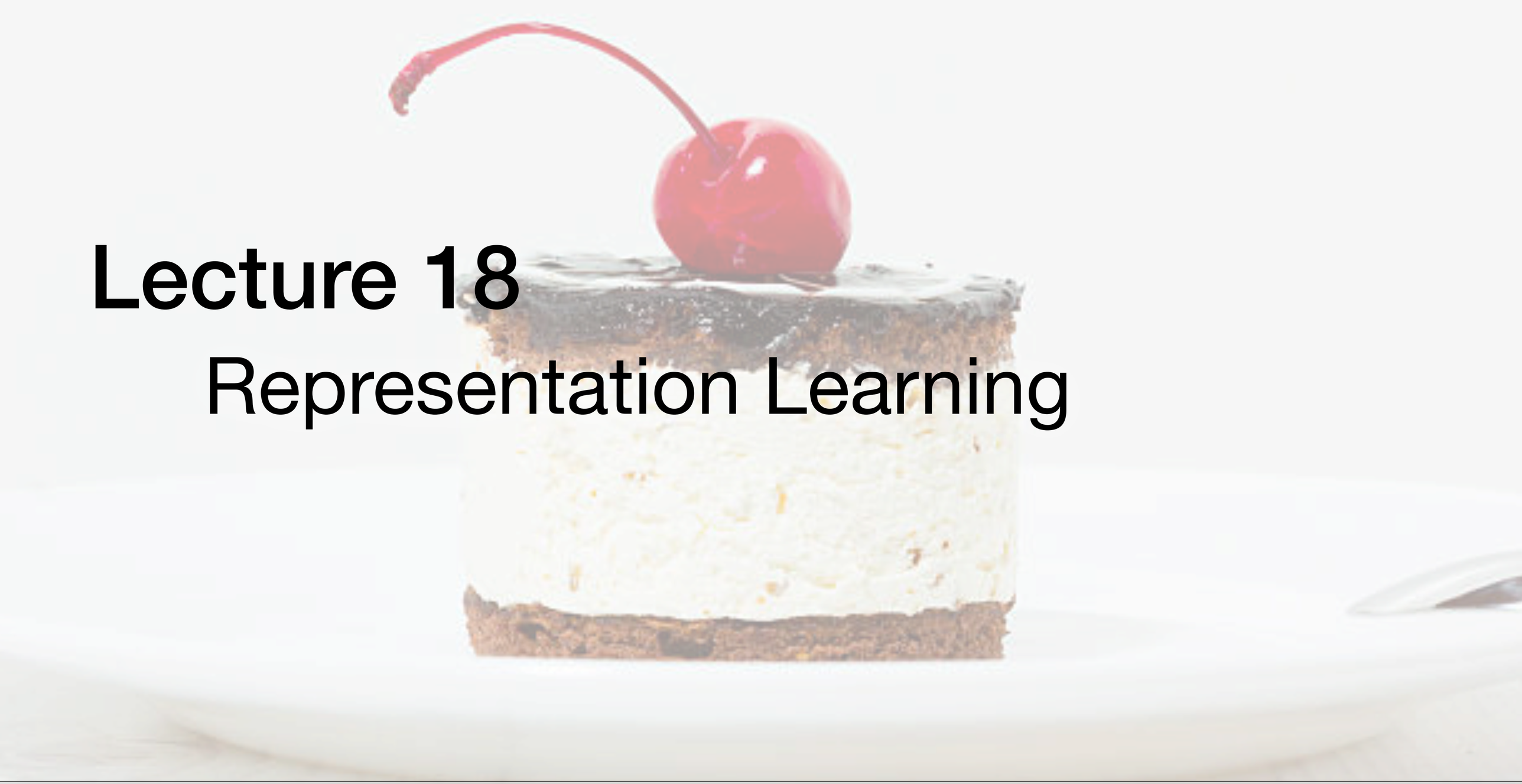


Lecture 18

Representation Learning

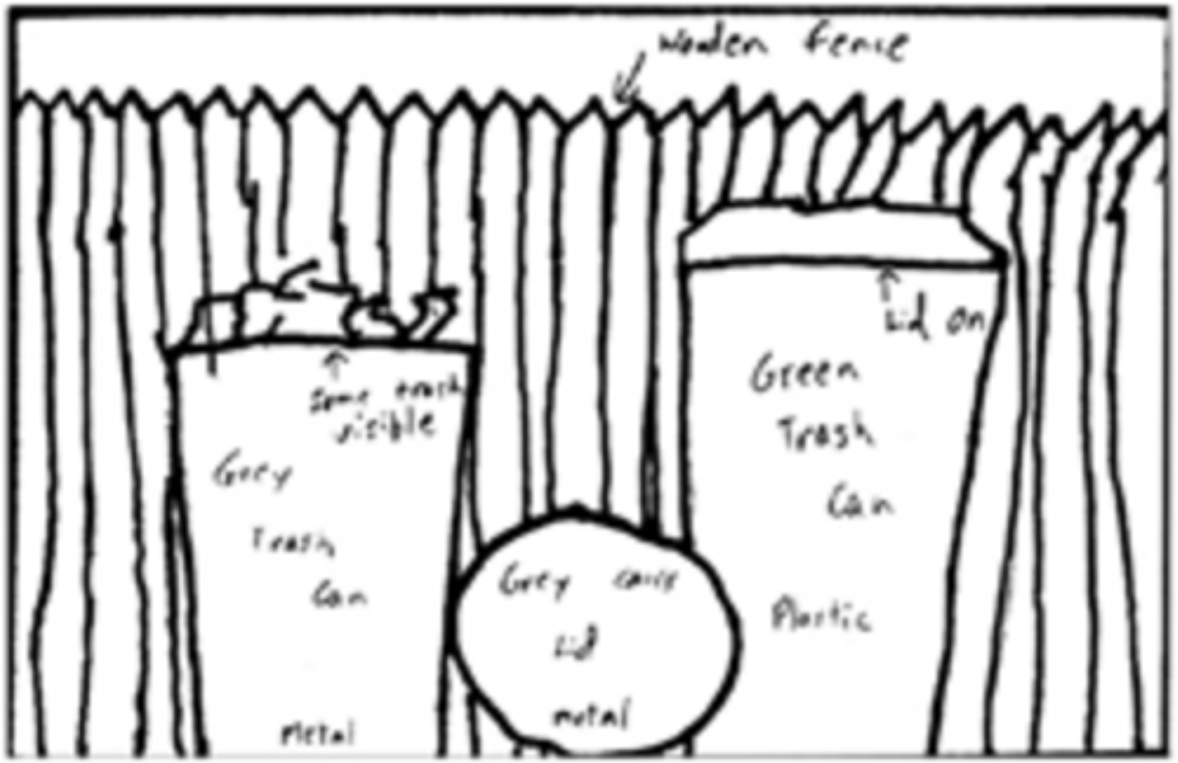
Bill Freeman, Antonio Torralba, Phillip Isola

18. Representation Learning

- Representations in the brain
- What is learned by a deep net?
- Transfer learning and finetuning
- Unsupervised and self-supervised learning

Observed image

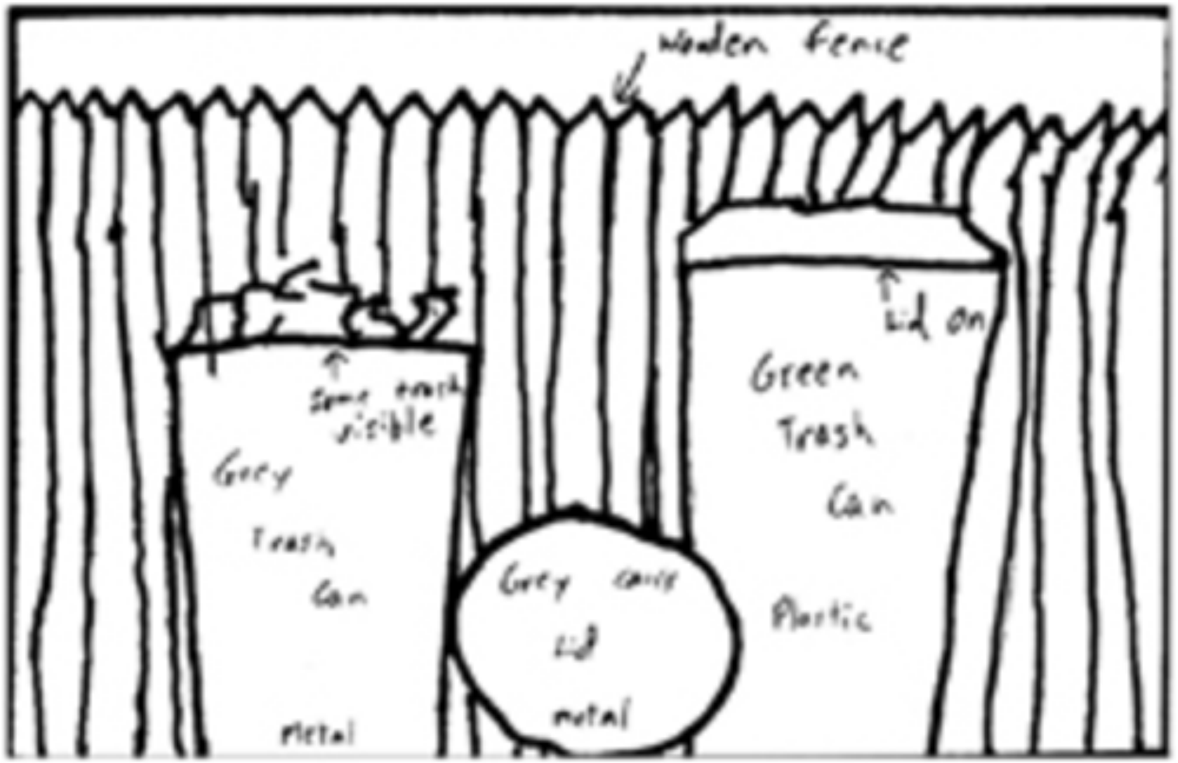
Drawn from memory



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

Observed image

Drawn from memory



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

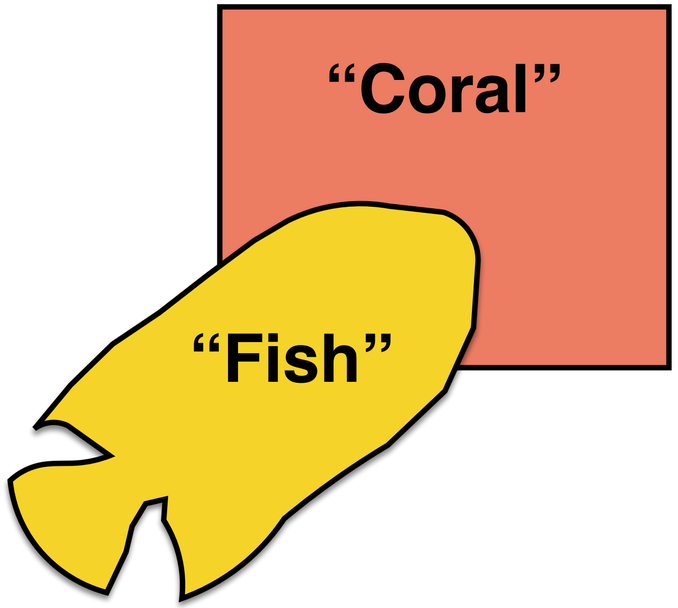
"I stand at the window and see a house, trees, sky. Theoretically I might say there were 327 brightnesses and nuances of colour. Do I have "327"? No. I have sky, house, and trees."

— Max Wertheimer, 1923

Representation learning

$\mathbf{x}$

Image

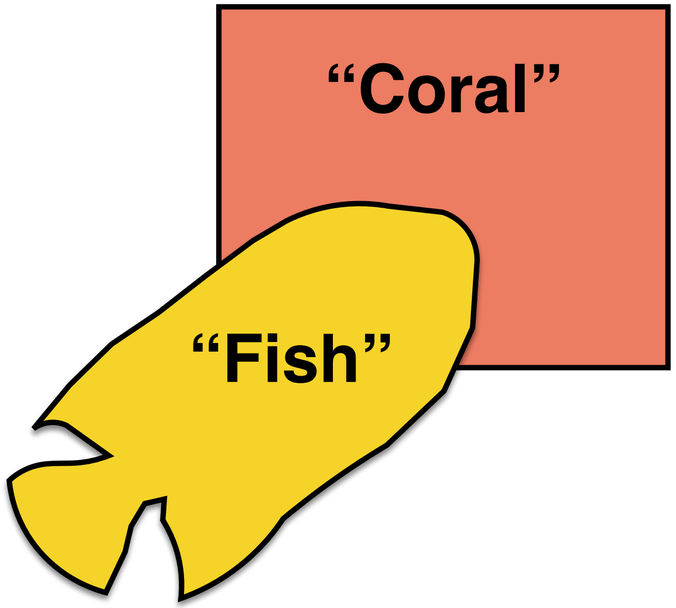


Compact mental representation

Representation learning

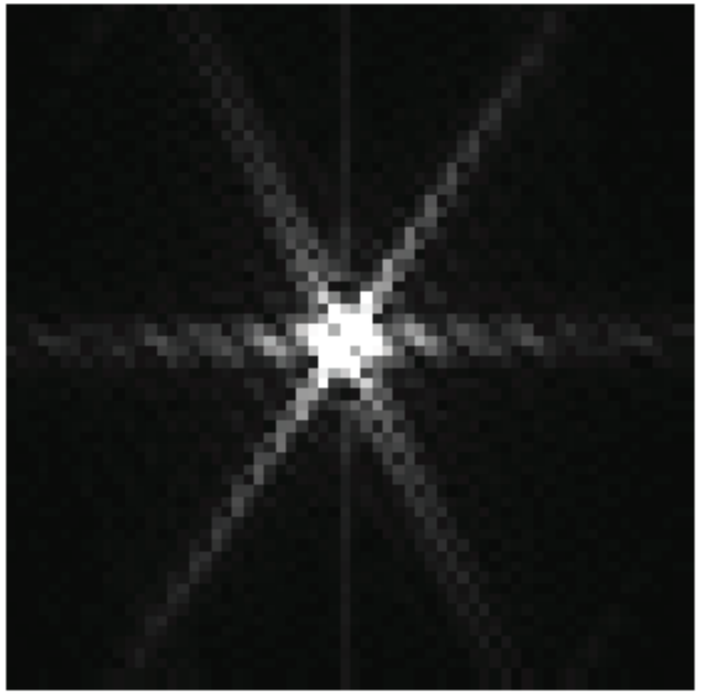
Good representations are:

1. Compact (*minimal*)
2. Explanatory (*sufficient*)
3. Disentangled (*independent factors*)
4. Interpretable
5. *Make subsequent problem solving easy*

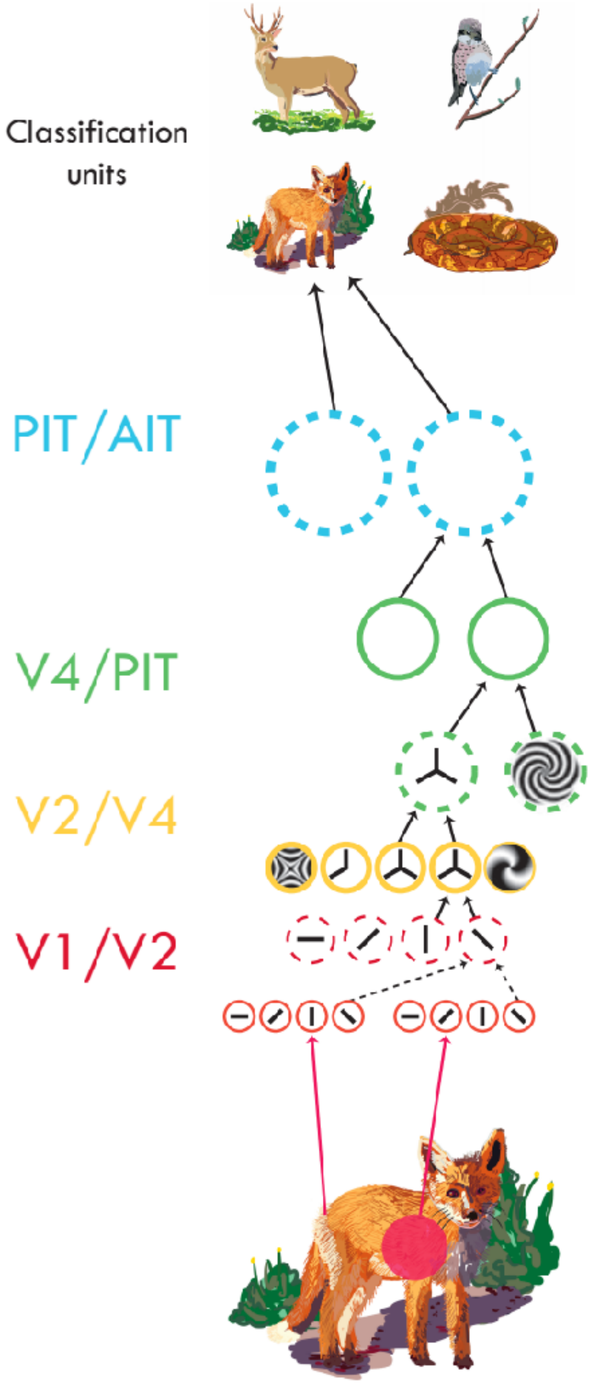


[See "Representation Learning", Bengio 2013, for more commentary]

Representation learning

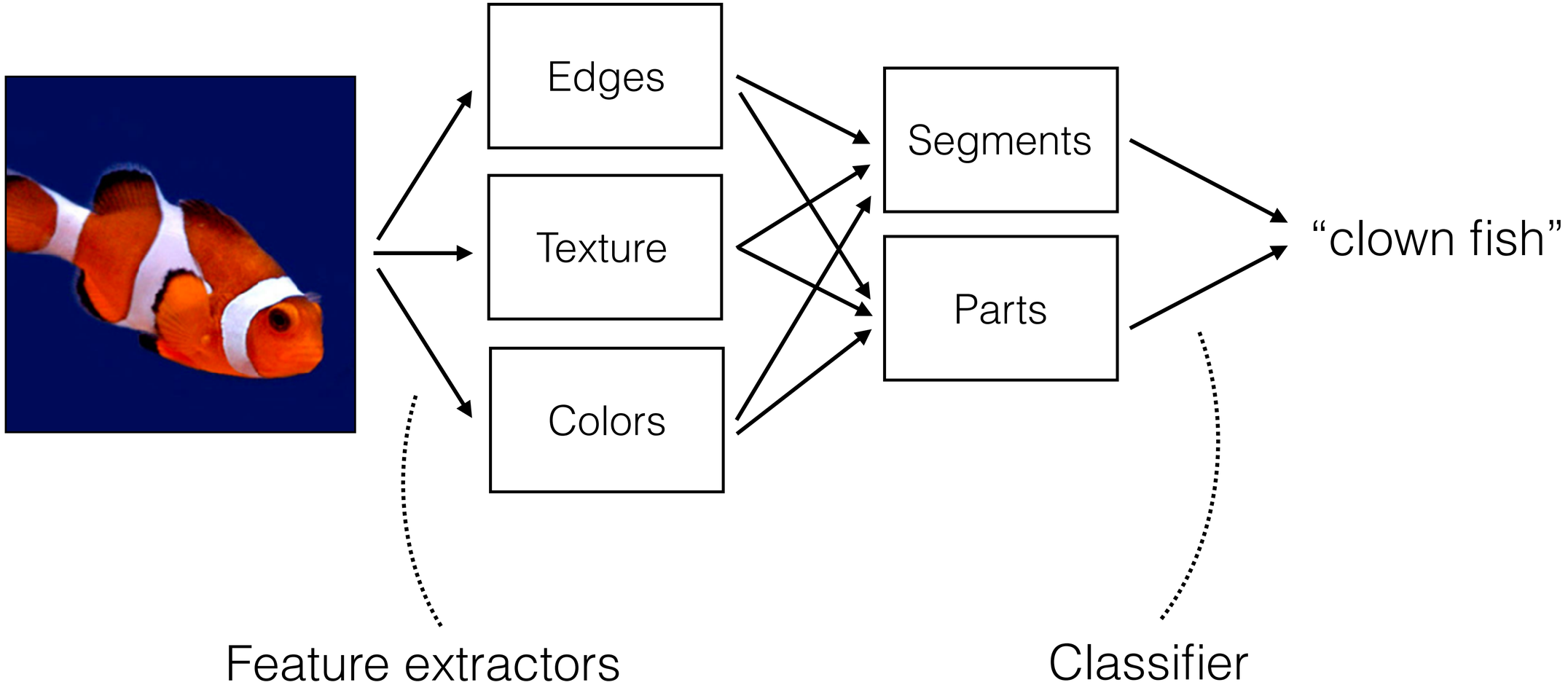


Convolution is pointwise multiplication in the frequency domain.

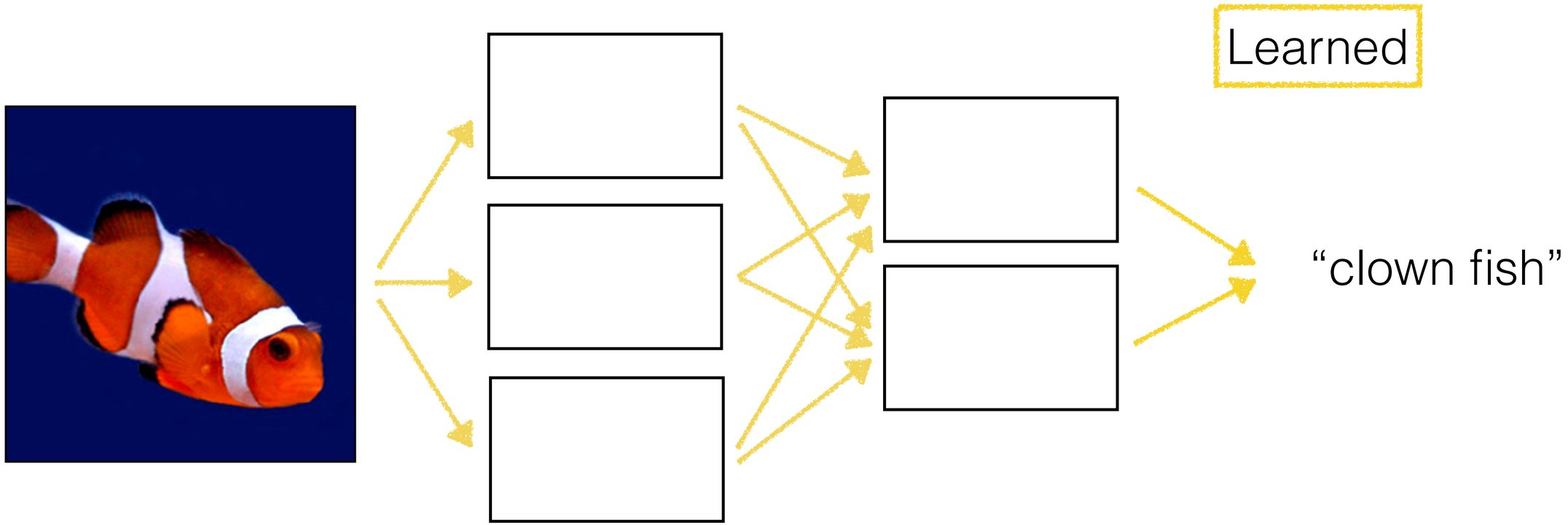


[Serre, 2014]

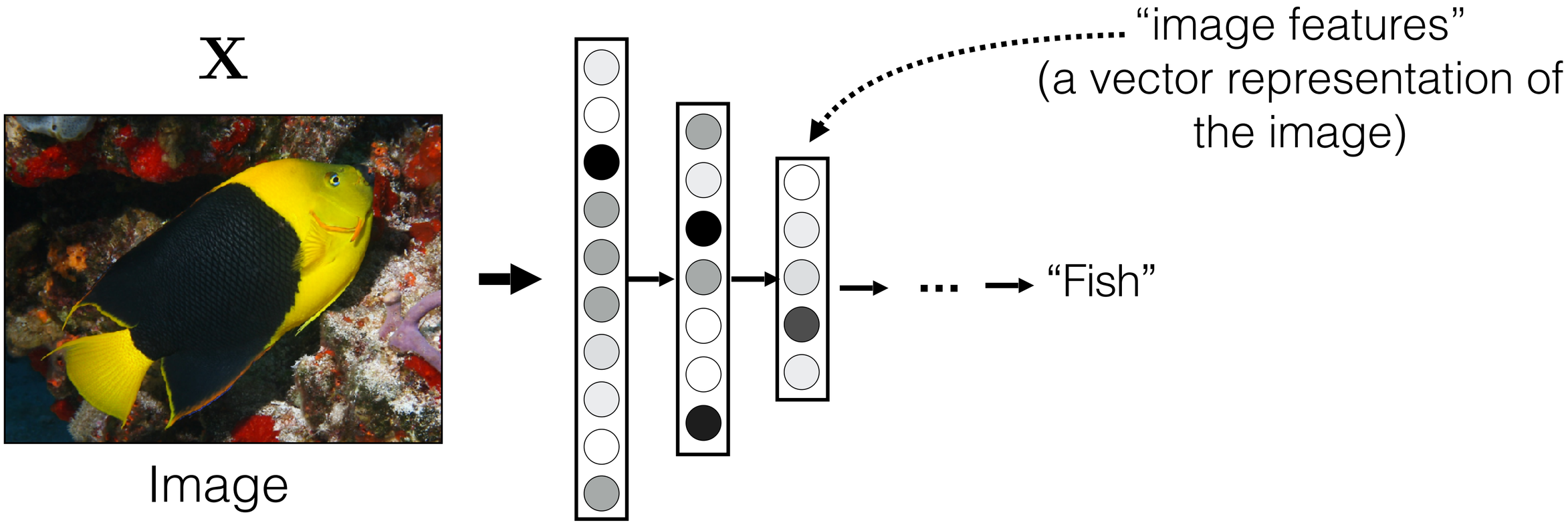
Classical object recognition



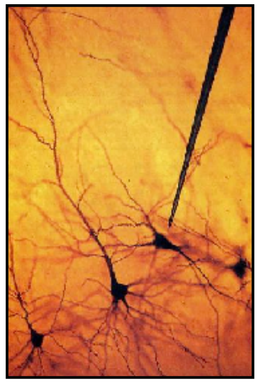
Deep learning

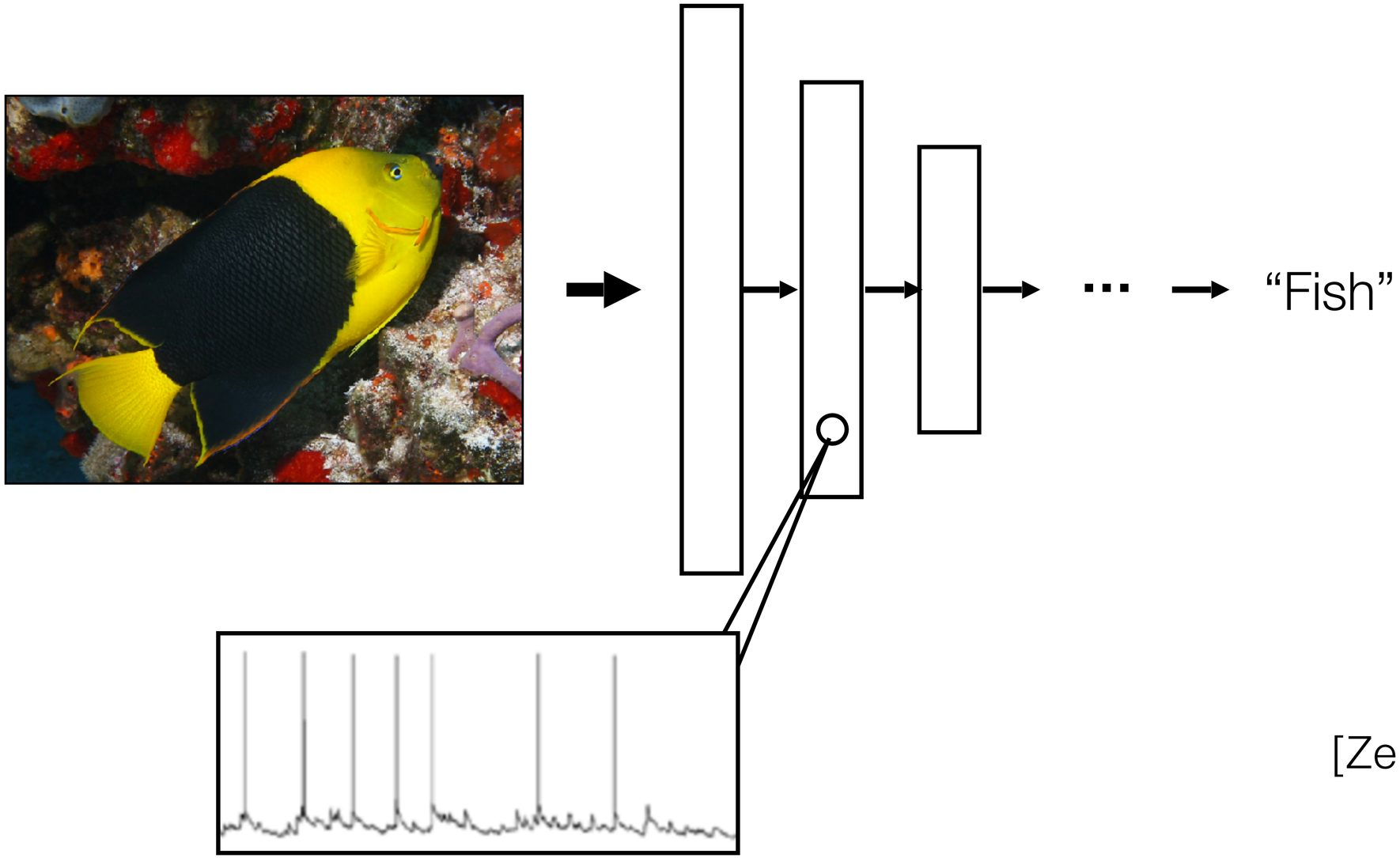


What do deep nets internally learn?



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**

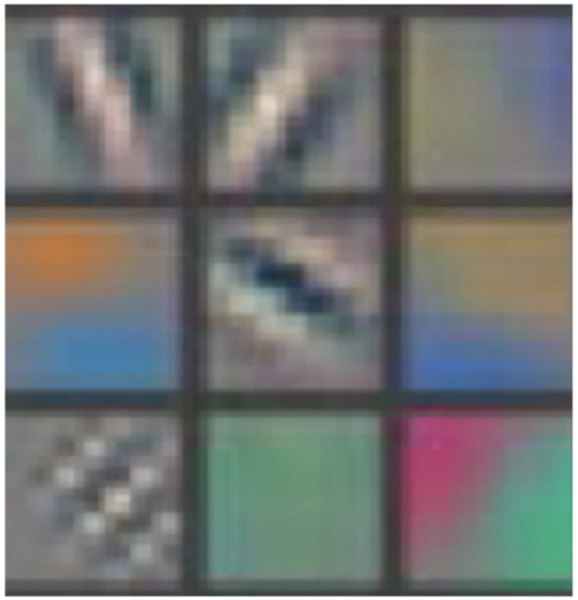
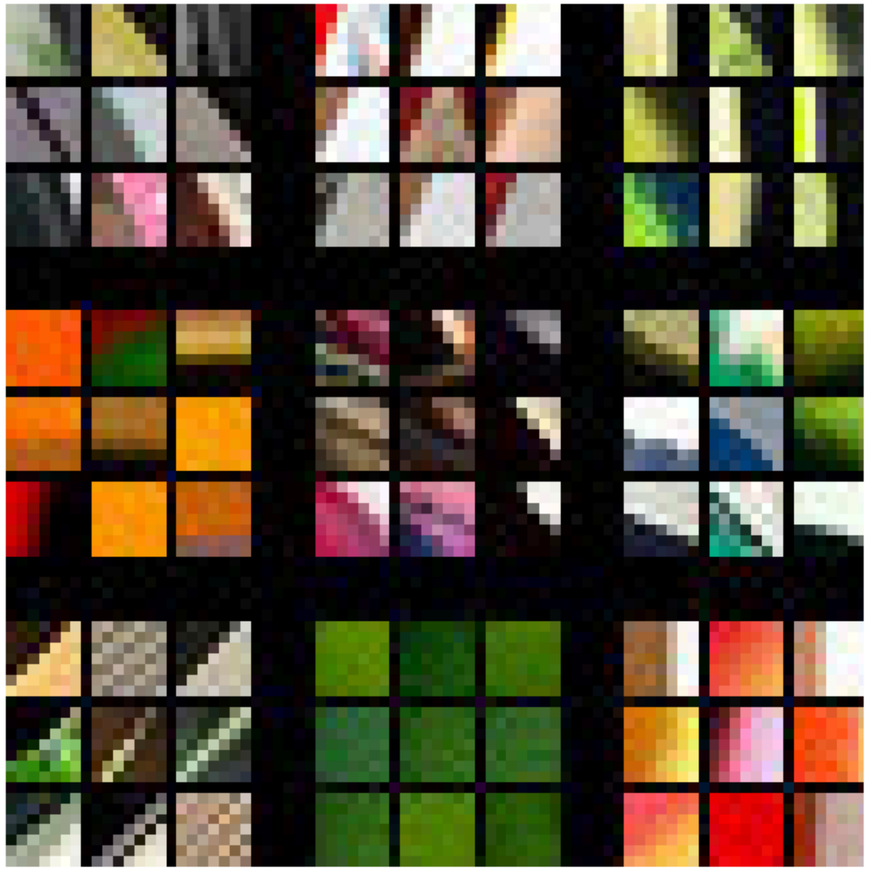
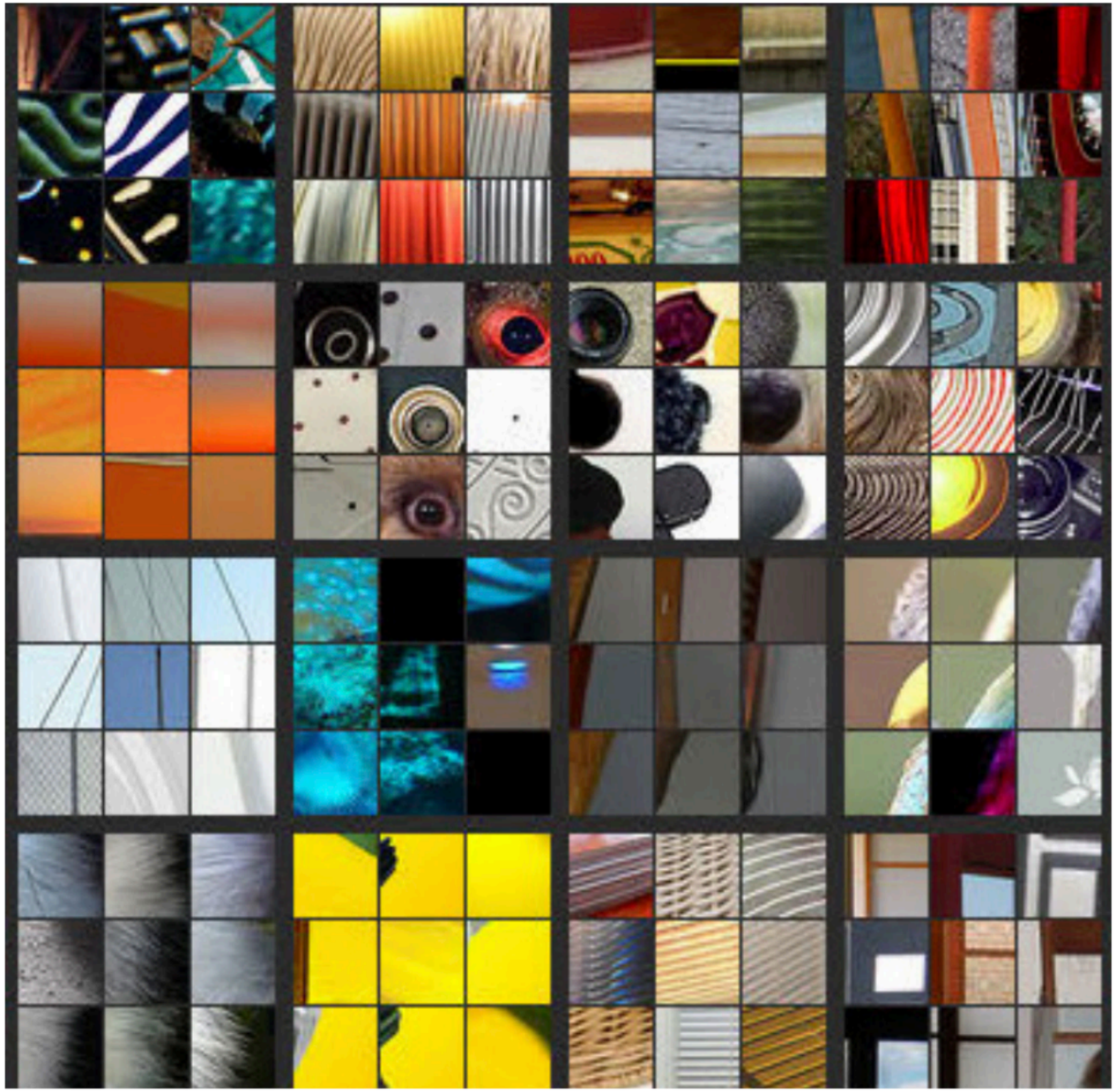


Image patches that activate each of the **layer 1** filters most strongly



[Zeiler and Fergus, 2014]

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



[Zeiler and Fergus, 2014]

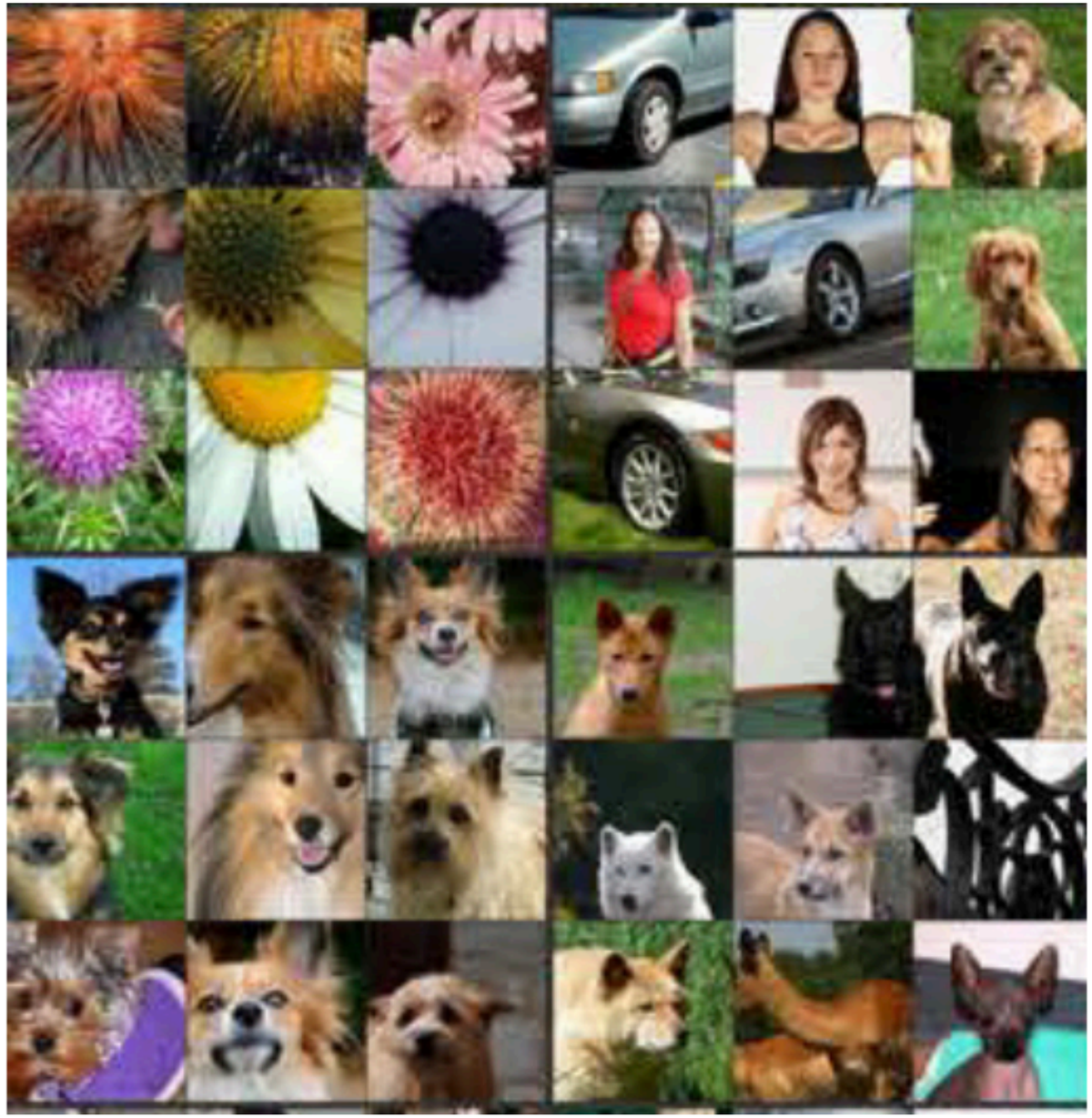
Image patches that activate each of the **layer 3** neurons most strongly

[Zeiler and Fergus, 2014]

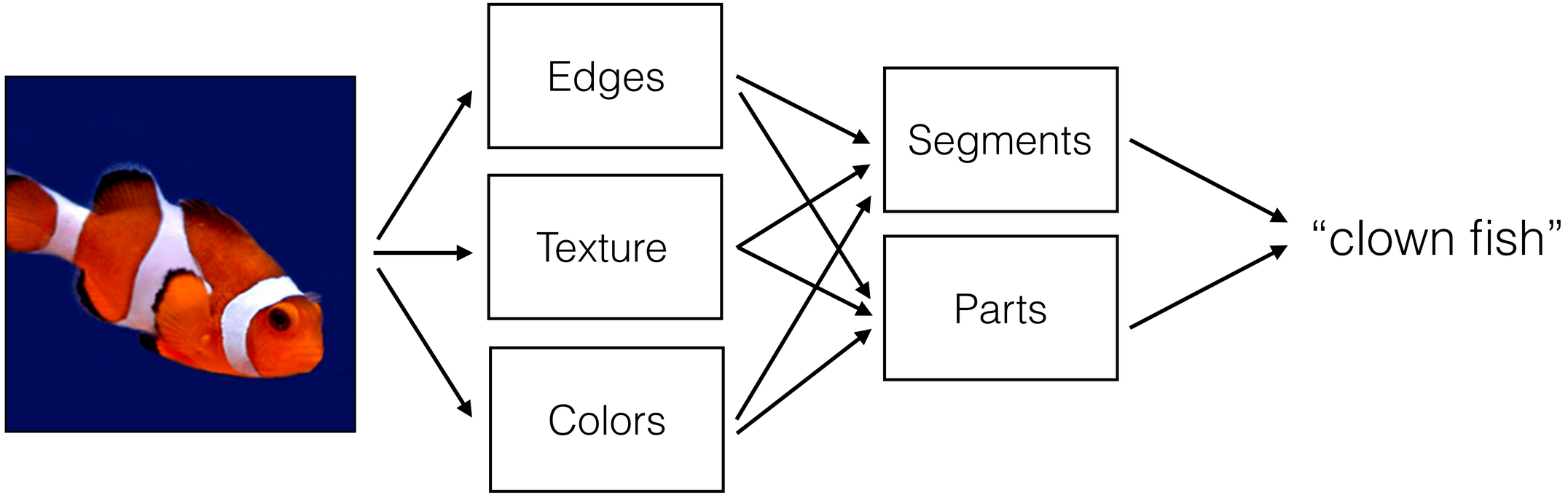
Image patches that activate each of the **layer 4** neurons most strongly

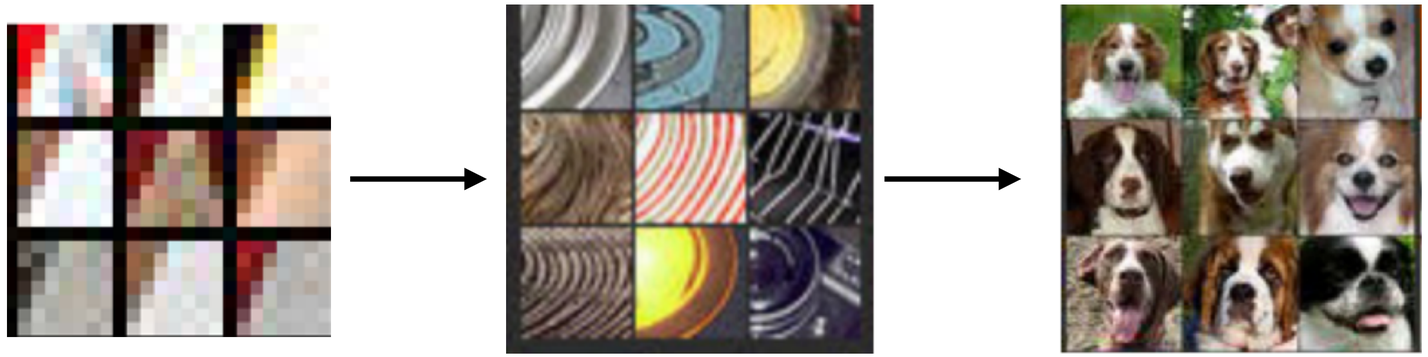
[Zeiler and Fergus, 2014]

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

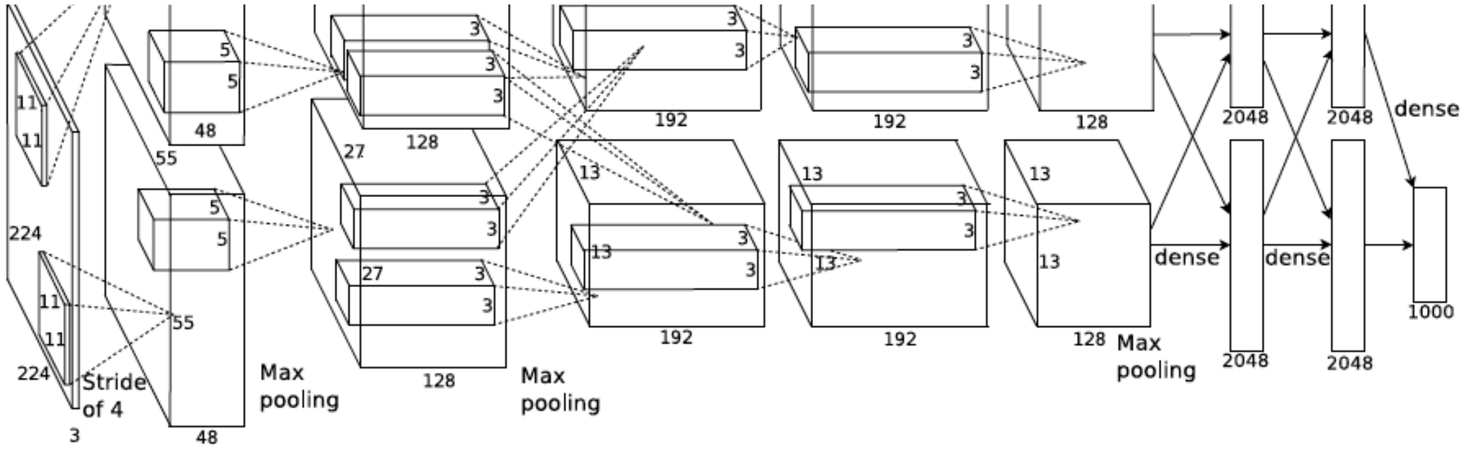


CNNs *learned* the classical visual recognition pipeline!





Alexnet — [Krizhevsky et al. NIPS 2012]



[227x227x3] INPUT
[55x55x96] CONV1: 96 11x11 filters at stride 4, pad 0
[27x27x96] MAX POOL1: 3x3 filters at stride 2
[27x27x96] NORM1: Normalization layer
[27x27x256] CONV2: 256 5x5 filters at stride 1, pad 2
[13x13x256] MAX POOL2: 3x3 filters at stride 2
[13x13x256] NORM2: Normalization layer
[13x13x384] CONV3: 384 3x3 filters at stride 1, pad 1
[13x13x384] CONV4: 384 3x3 filters at stride 1, pad 1
[13x13x256] CONV5: 256 3x3 filters at stride 1, pad 1
[6x6x256] MAX POOL3: 3x3 filters at stride 2
[4096] FC6: 4096 neurons
[4096] FC7: 4096 neurons
[1000] FC8: 1000 neurons (class scores)

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]

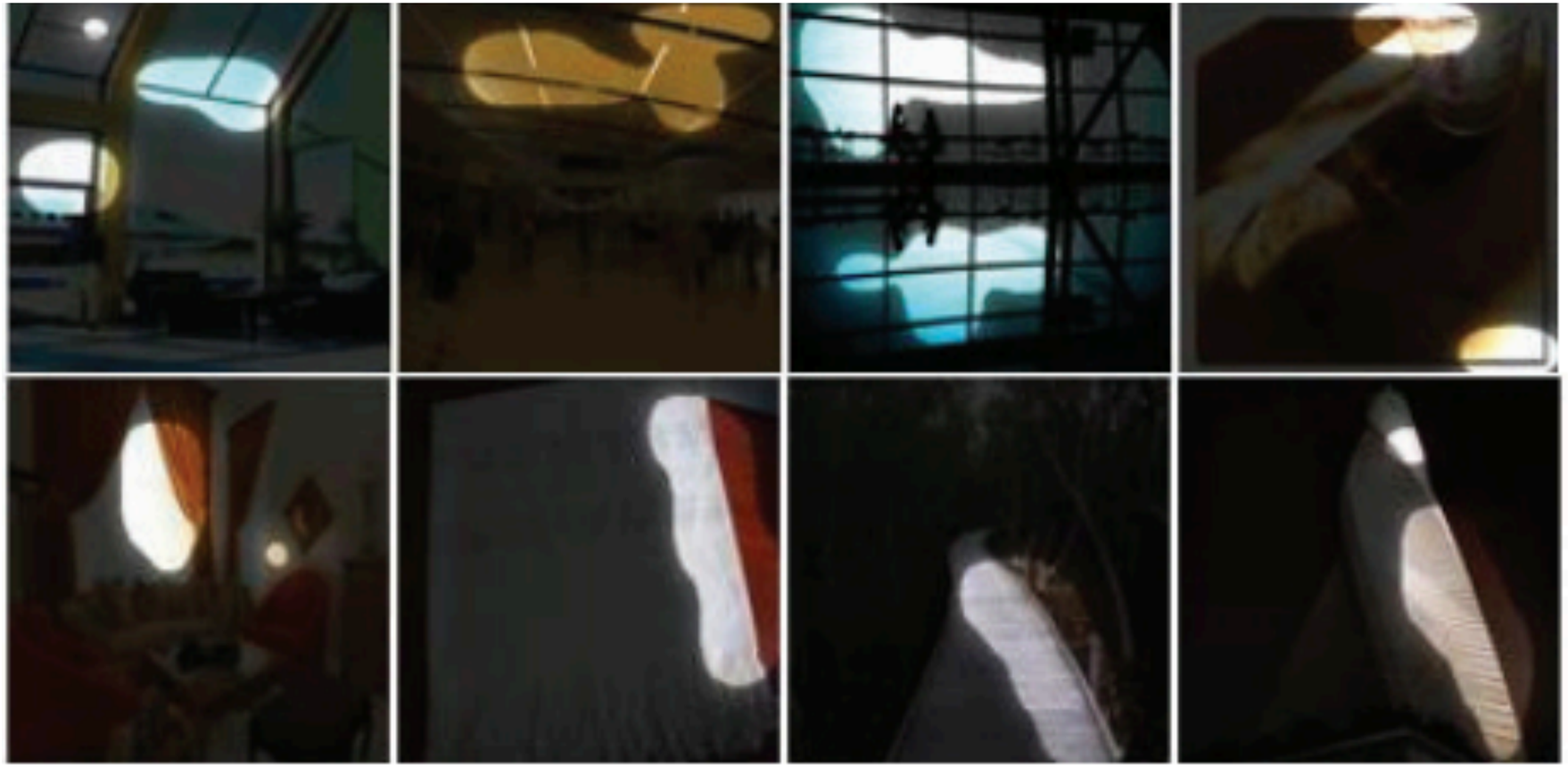
pool 1

[http://people.csail.mit.edu/torralba/research/drawCNN/drawNet.html]

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

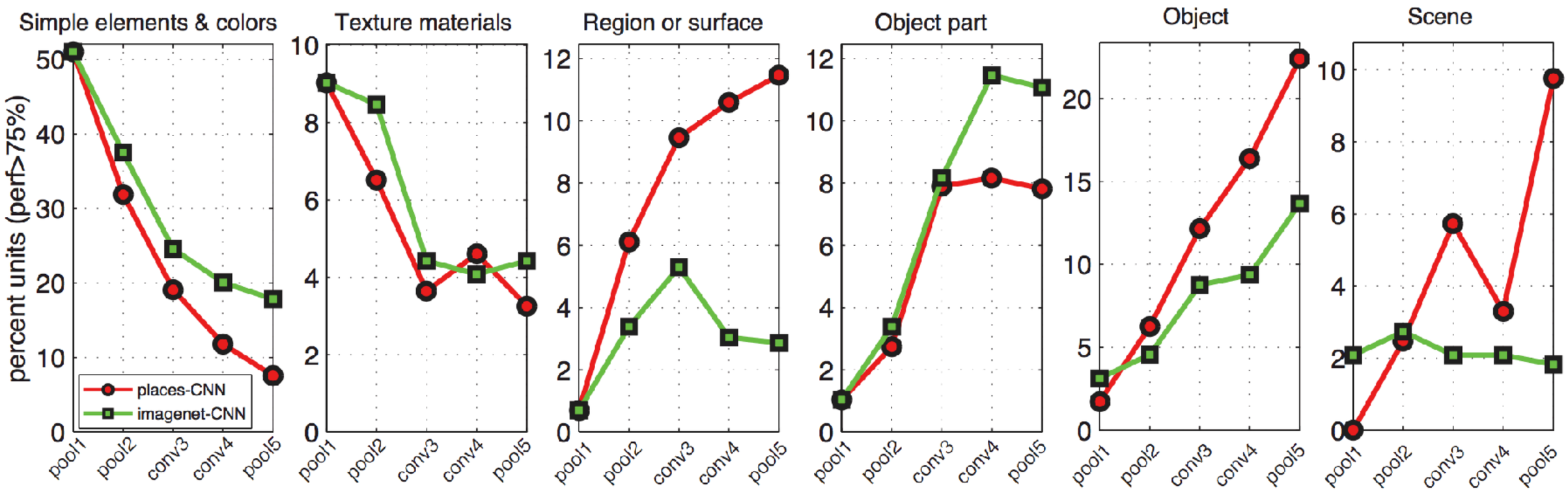
Object Detectors Emergence in Deep Scene CNNs

[Zhou et al., ICLR 2015]

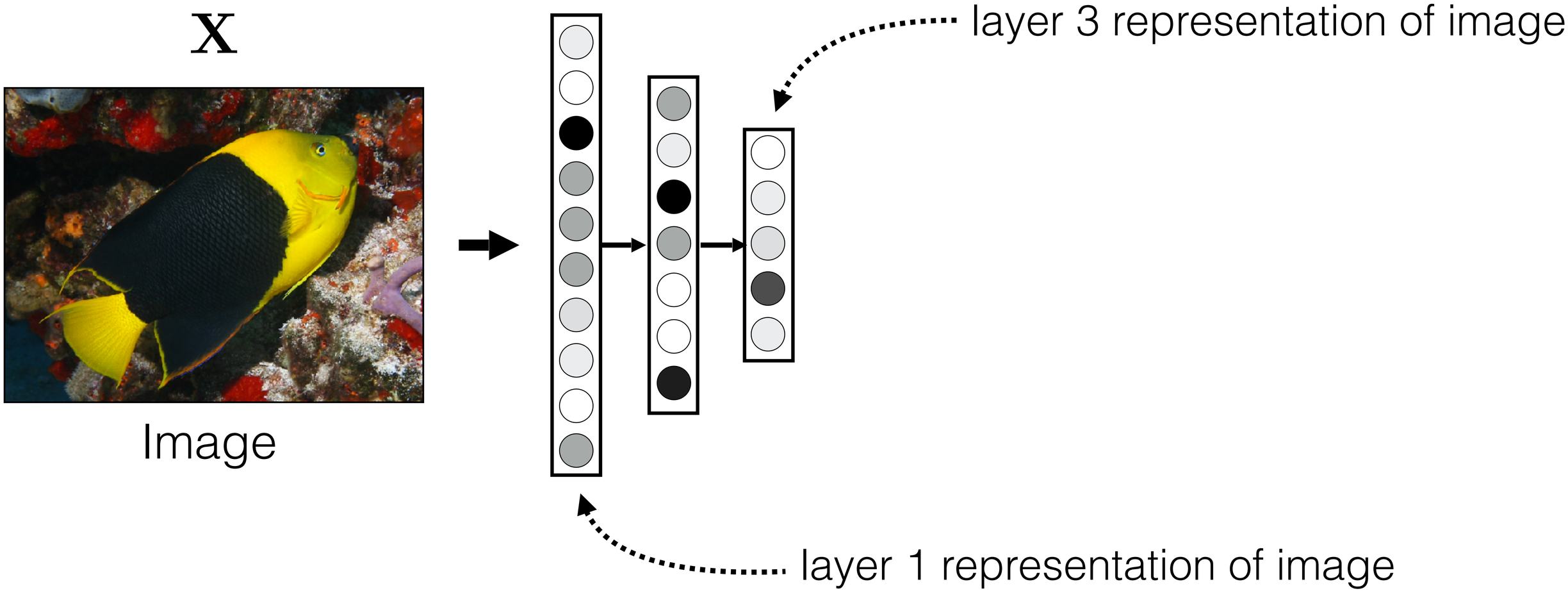
pool 5

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)

Investigating a representation via similarity analysis

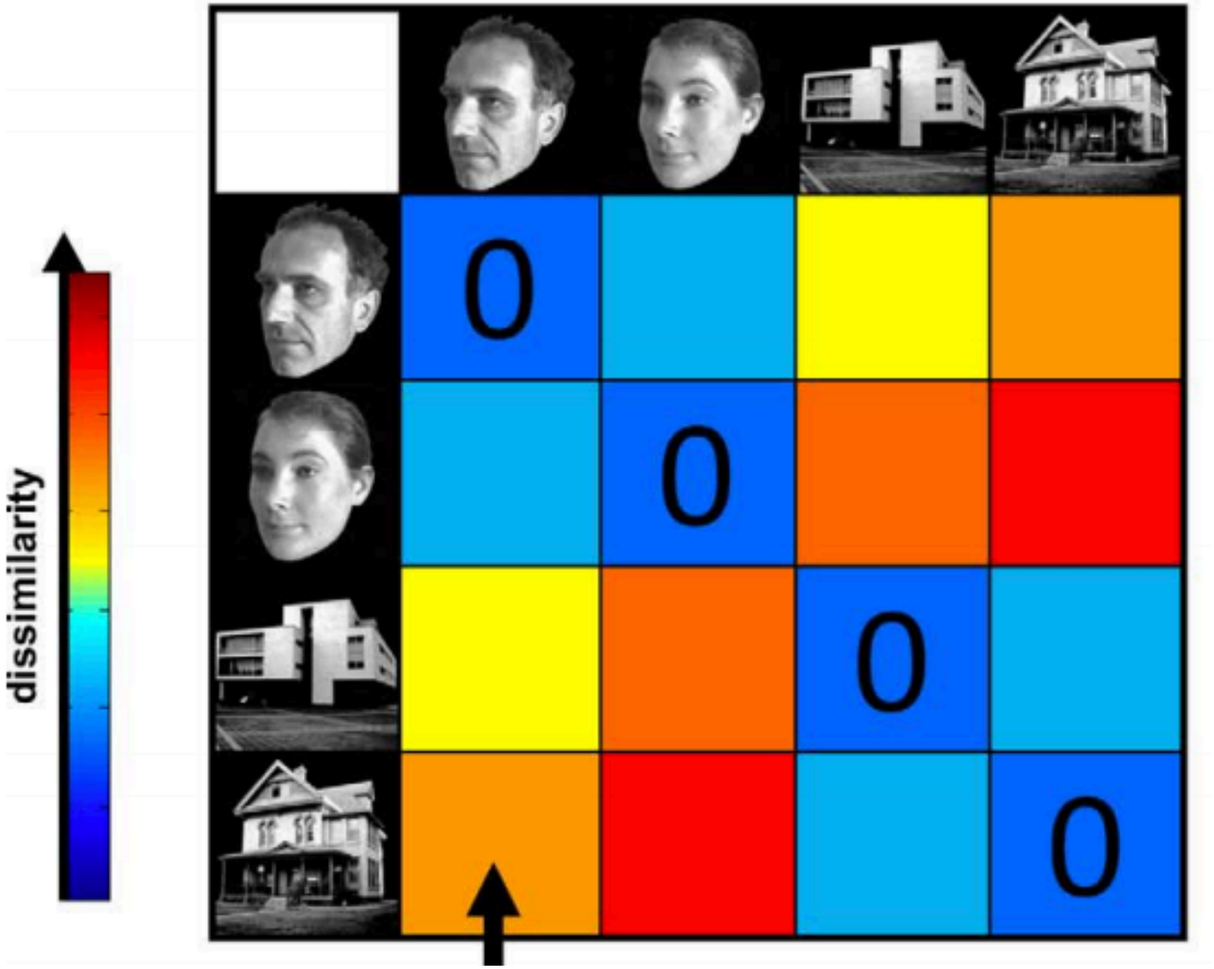
How similar are these two images?

How about these two?

[Kriegeskorte et al. 2008]

Investigating a representation via similarity analysis

Representational Dissimilarity Matrix



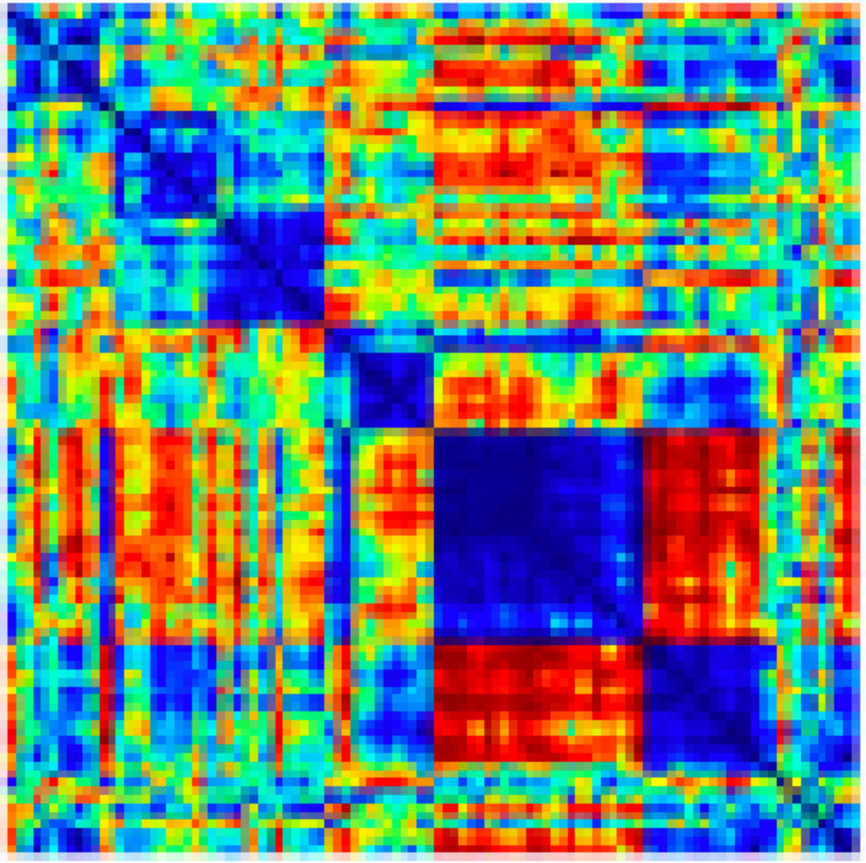
$$\|\mathbf{h}_i - \mathbf{h}_j\|$$

Neural activation vector

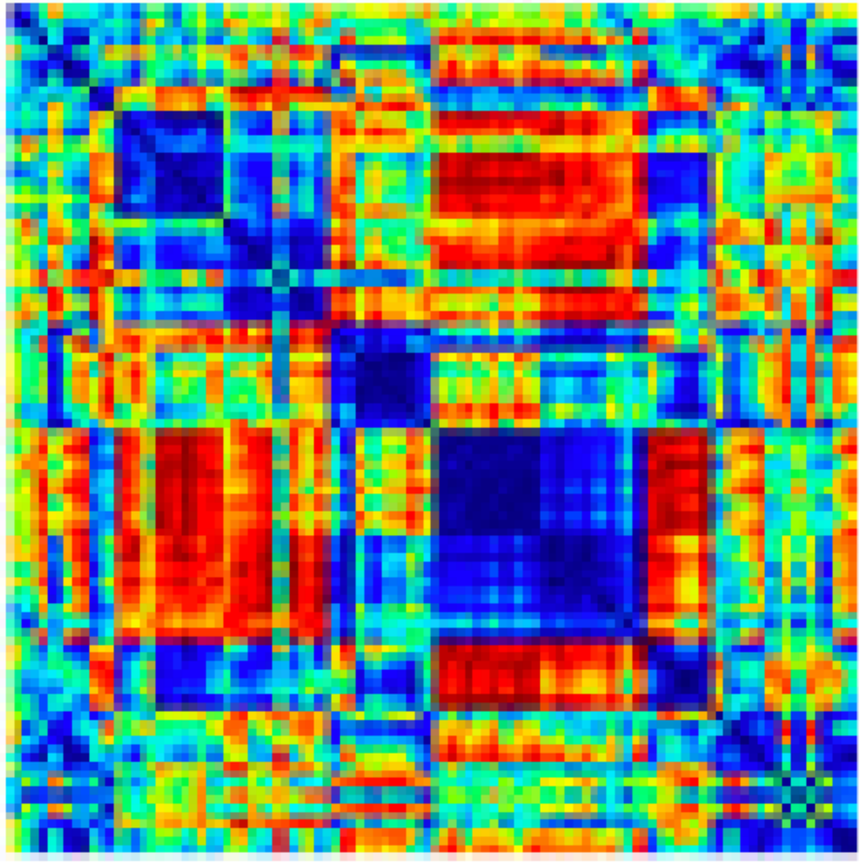
[Kriegeskorte et al. 2008]

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.

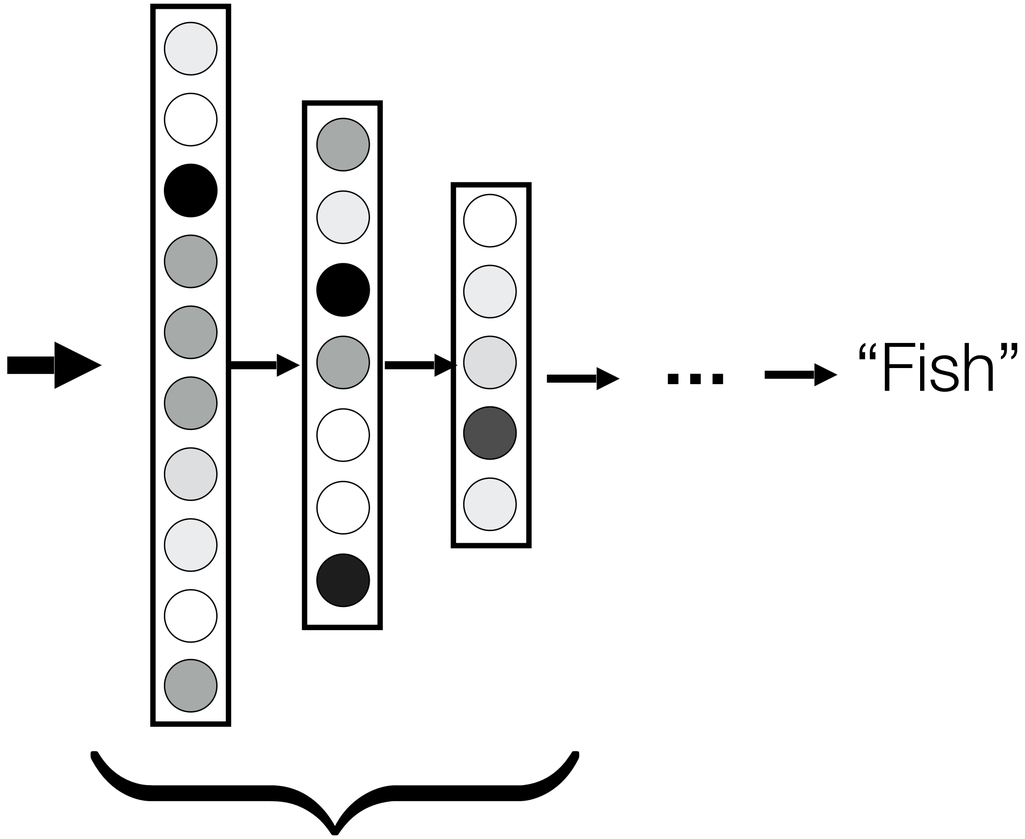
Deep nets organize visual information similarly to how our brains do!

[Yamins et al., PNAS 2014]

What do deep nets internally learn?

$\mathbf{x}$

Image

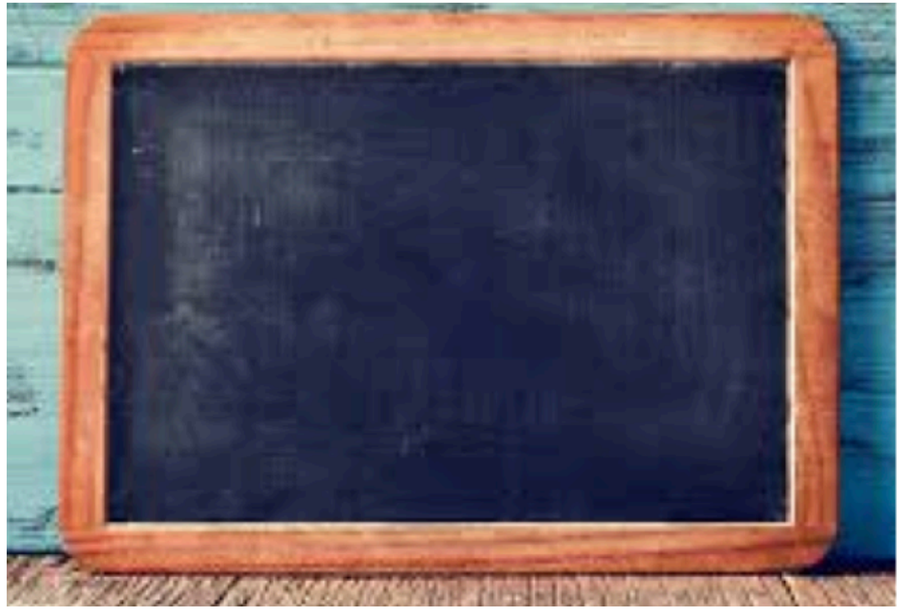


Representations!

A CNN is a multiscale, hierarchical representation of data

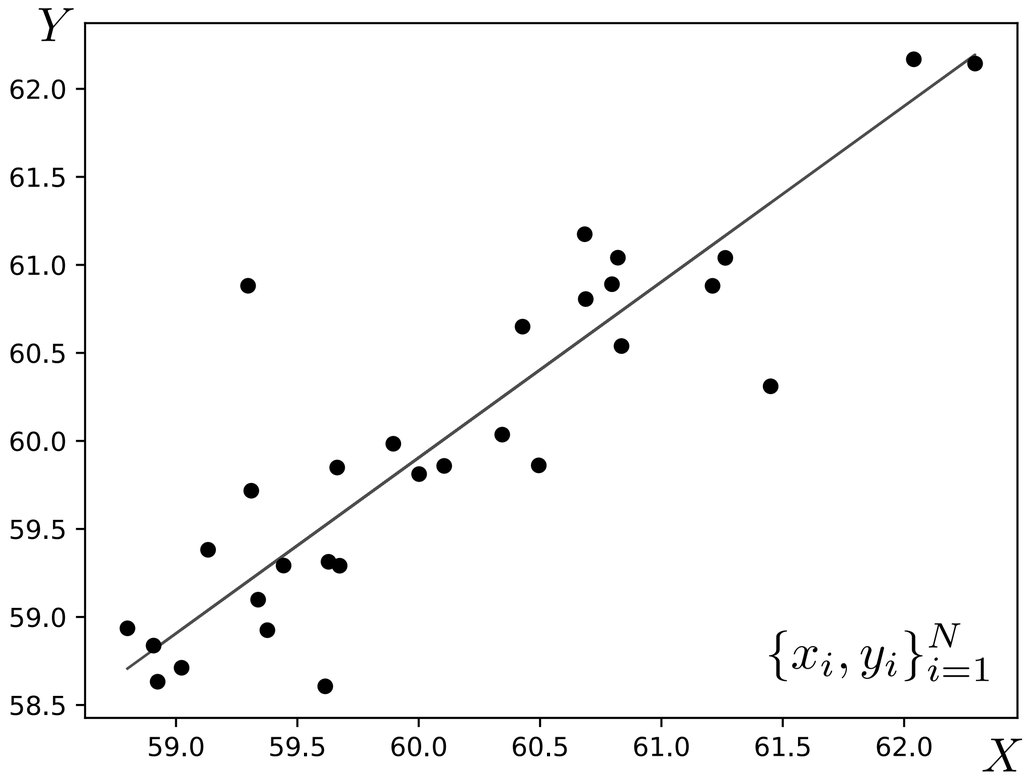
Transfer learning

“Generally speaking, a good representation is one that makes a subsequent learning task easier.” — *Deep Learning*, Goodfellow et al. 2016

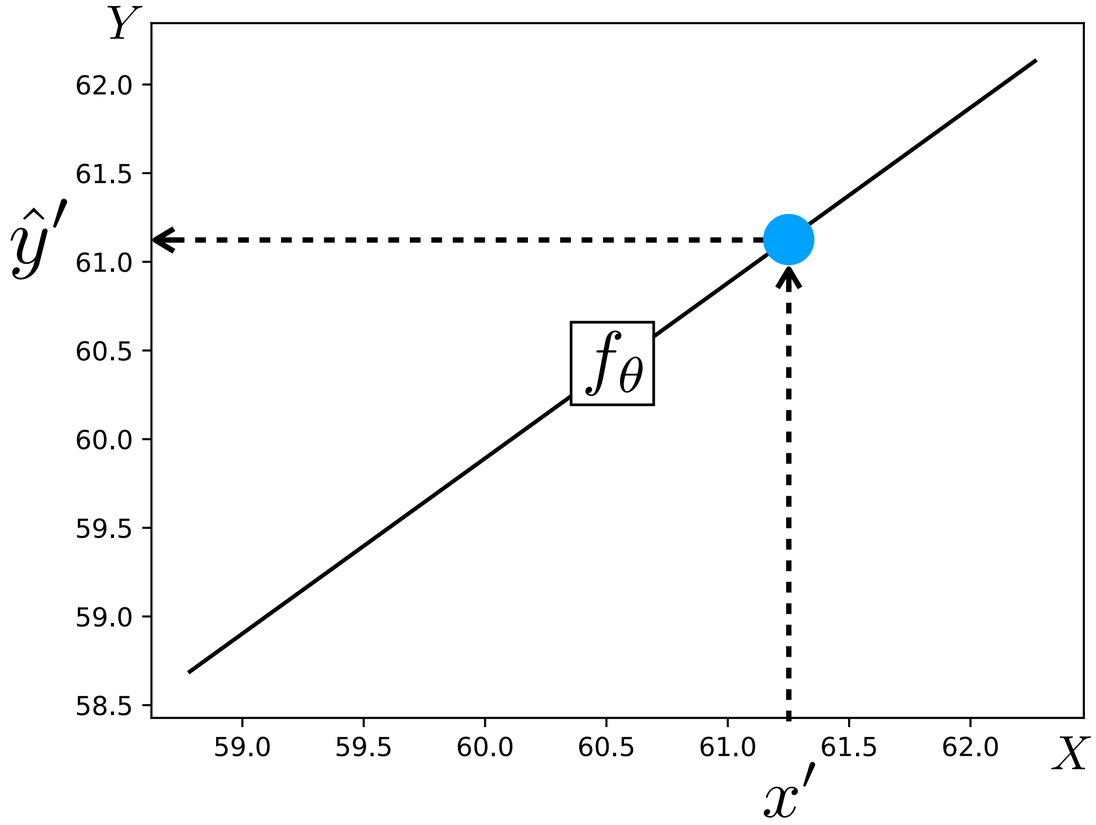


?

Training

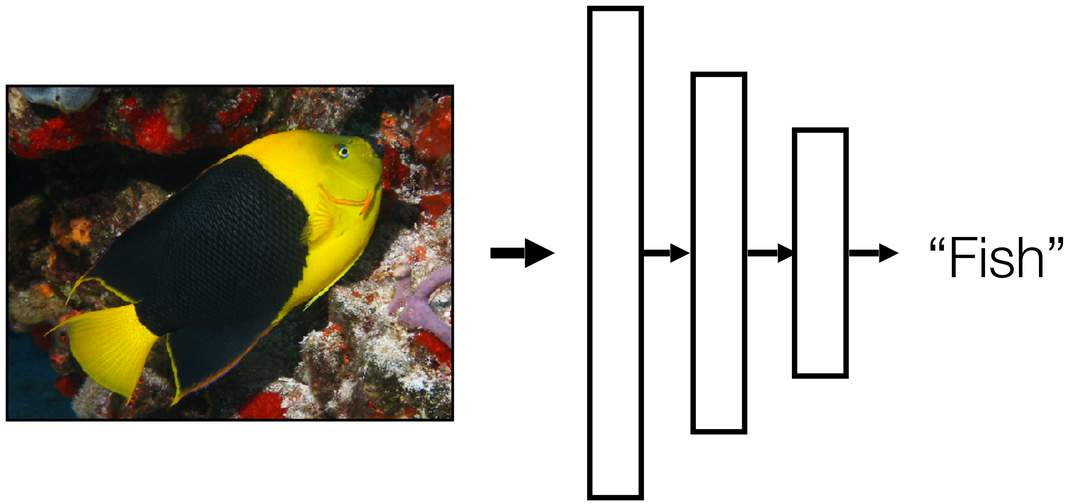


Testing



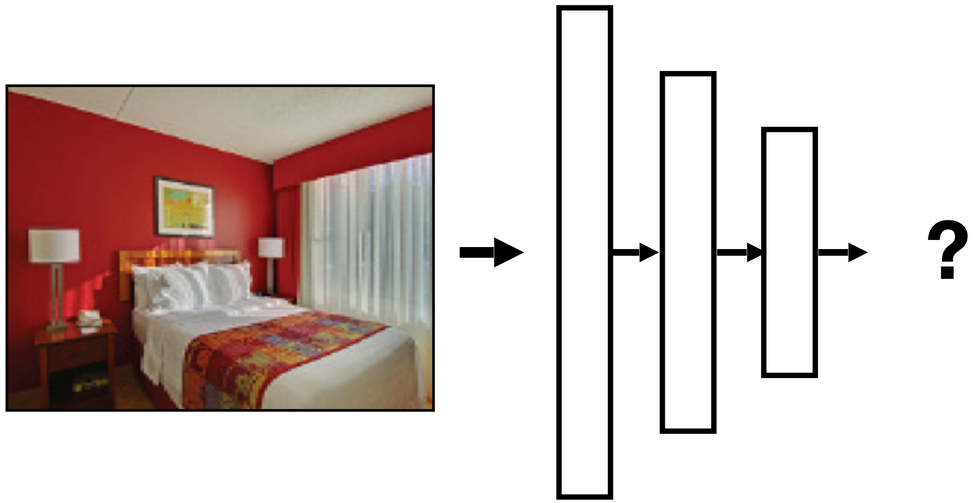
Training

Object recognition

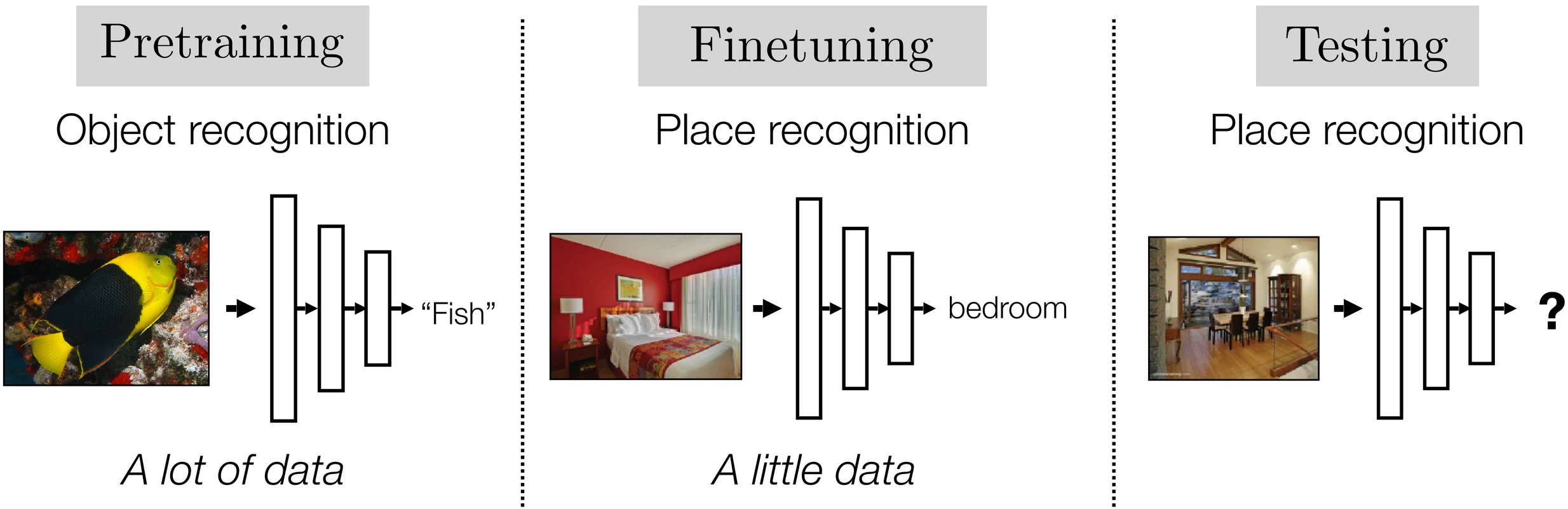


Testing

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.

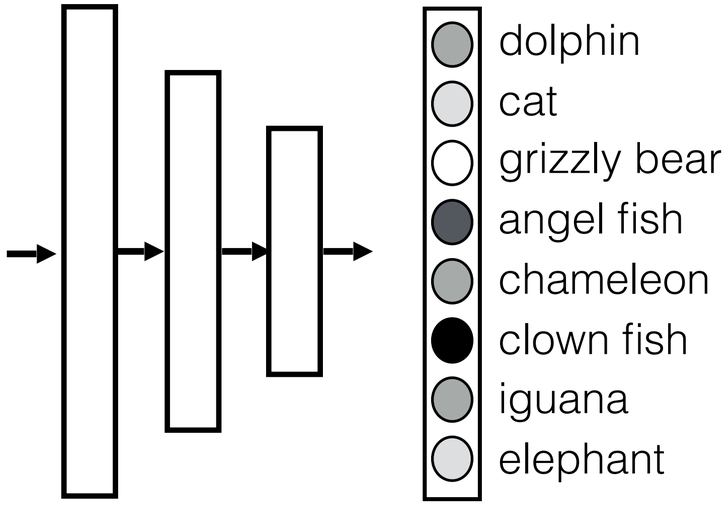
Finetuning in practice

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

Finetuning in practice

Pretraining

Object recognition



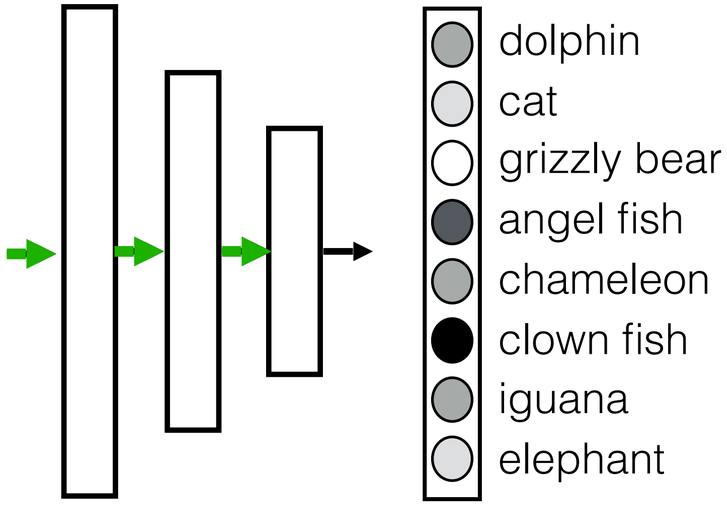
Finetuning

Place recognition

Finetuning in practice

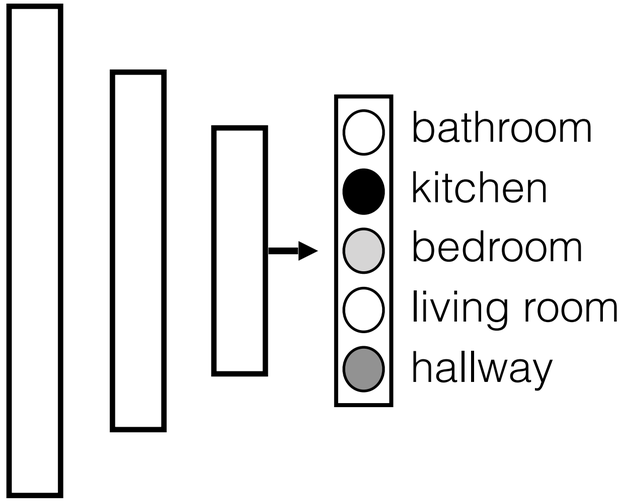
Pretraining

Object recognition



Finetuning

Place recognition



The “learned representation” is just the weights and biases, so that’s what we transfer

Supervised object recognition

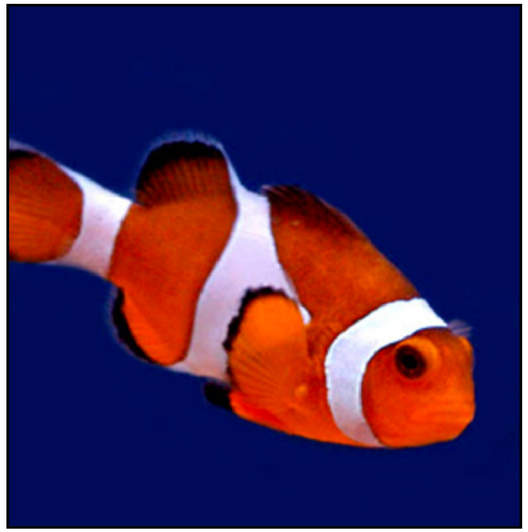


image X

Learner

"Fish"

label Y

Supervised object recognition

image X

Learner

"Fish"

label Y

Supervised object recognition

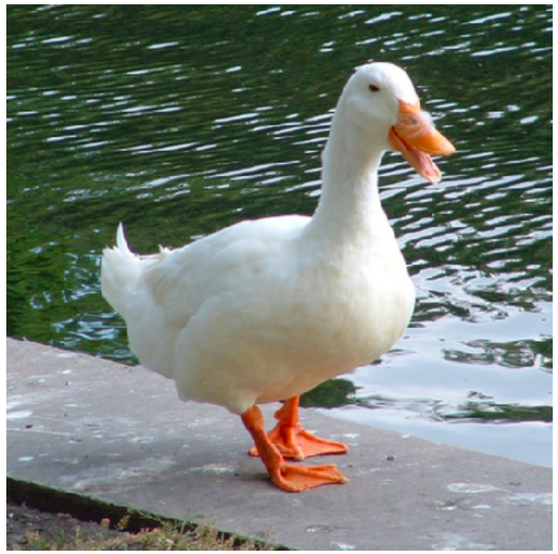
image X

Learner

"Fish"

label Y

Supervised object recognition



⋮

image X

Learner

"Duck"

label Y

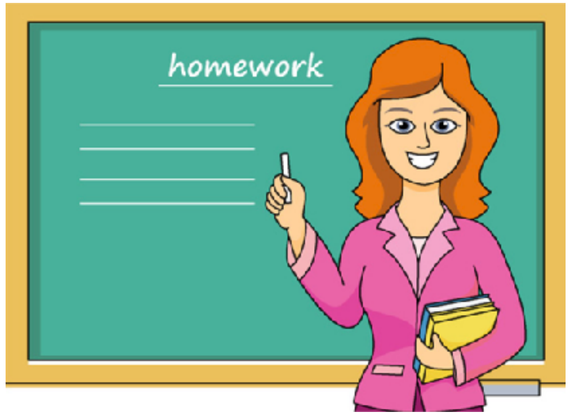
Supervised computer vision

Hand-curated training data

+ Informative

- Expensive

- Limited to teacher's knowledge



Vision in nature

Raw unlabeled training data

+ Cheap

- Noisy

- Harder to interpret

Learning from examples

(aka **supervised learning**)

Training data

$$\{x_1, y_1\}$$
$$\{x_2, y_2\}$$
$$\{x_3, y_3\}$$
$$\ldots$$

$$\rightarrow \quad \boxed{\text{Learner}} \quad \rightarrow \quad f : X \rightarrow Y$$

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

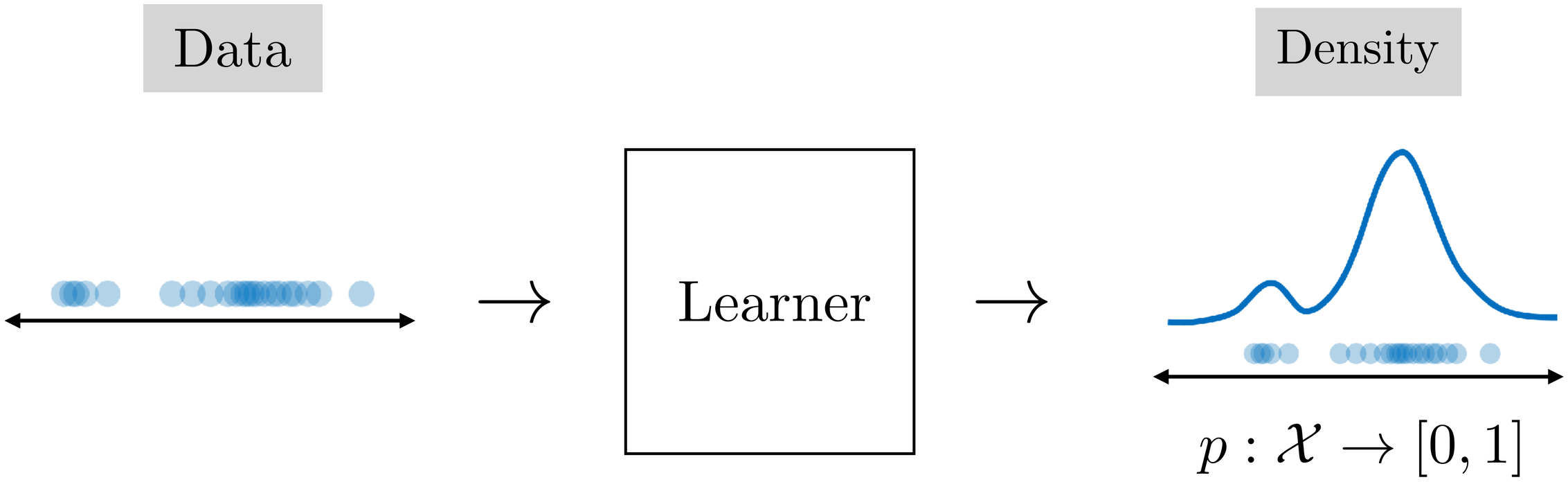
Learning without examples

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

Data

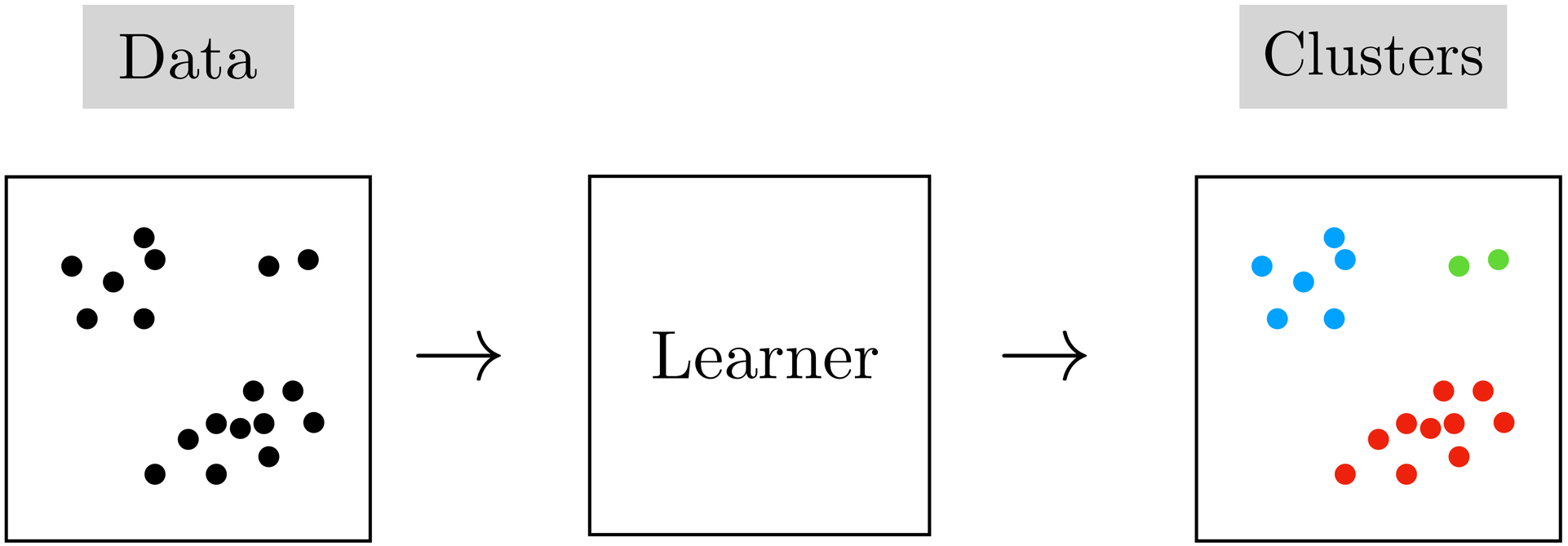
$$\{x_1\} \quad \{x_2\} \quad \{x_3\} \quad \ldots \quad \rightarrow \quad \boxed{\text{Learner}} \quad \rightarrow \quad ?$$

Density modeling



[figs modified from: http://introtodeeplearning.com/materials/2019_6S191_L4.pdf]

Clustering



$$f : \mathcal{X} \to \{1, \ldots, k\}$$

Representation Learning

Data

$\{x_1\}$

$\{x_2\}$ $\rightarrow$ $\boxed{\text{Learner}}$ $\rightarrow$ Representations

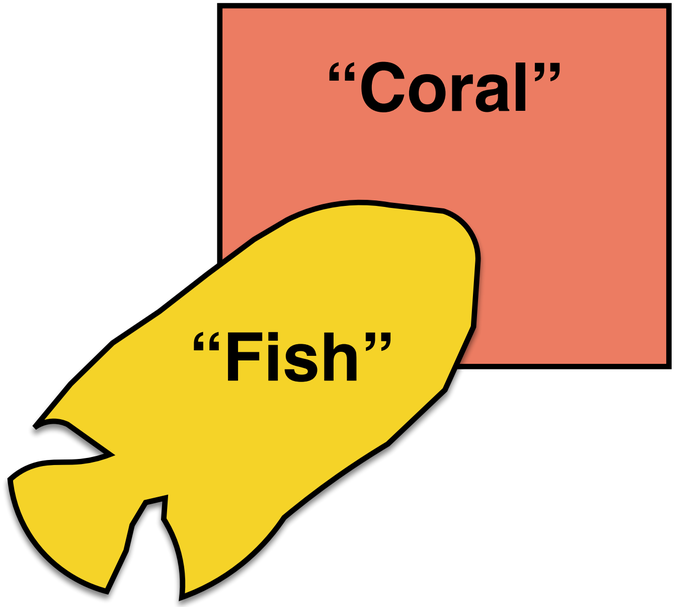
$\{x_3\}$

$\ldots$

Unsupervised Representation Learning

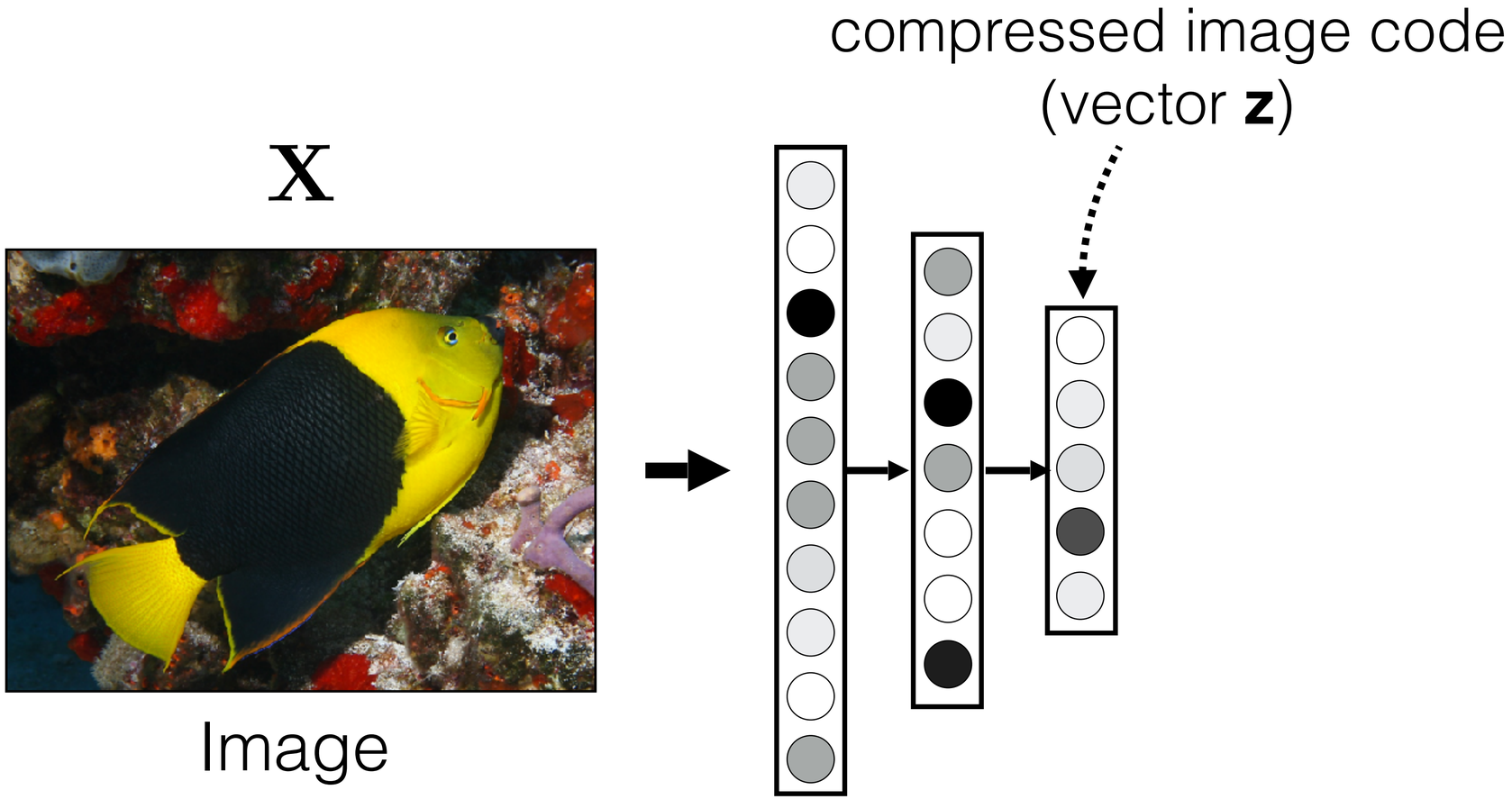
$\mathbf{x}$

Image

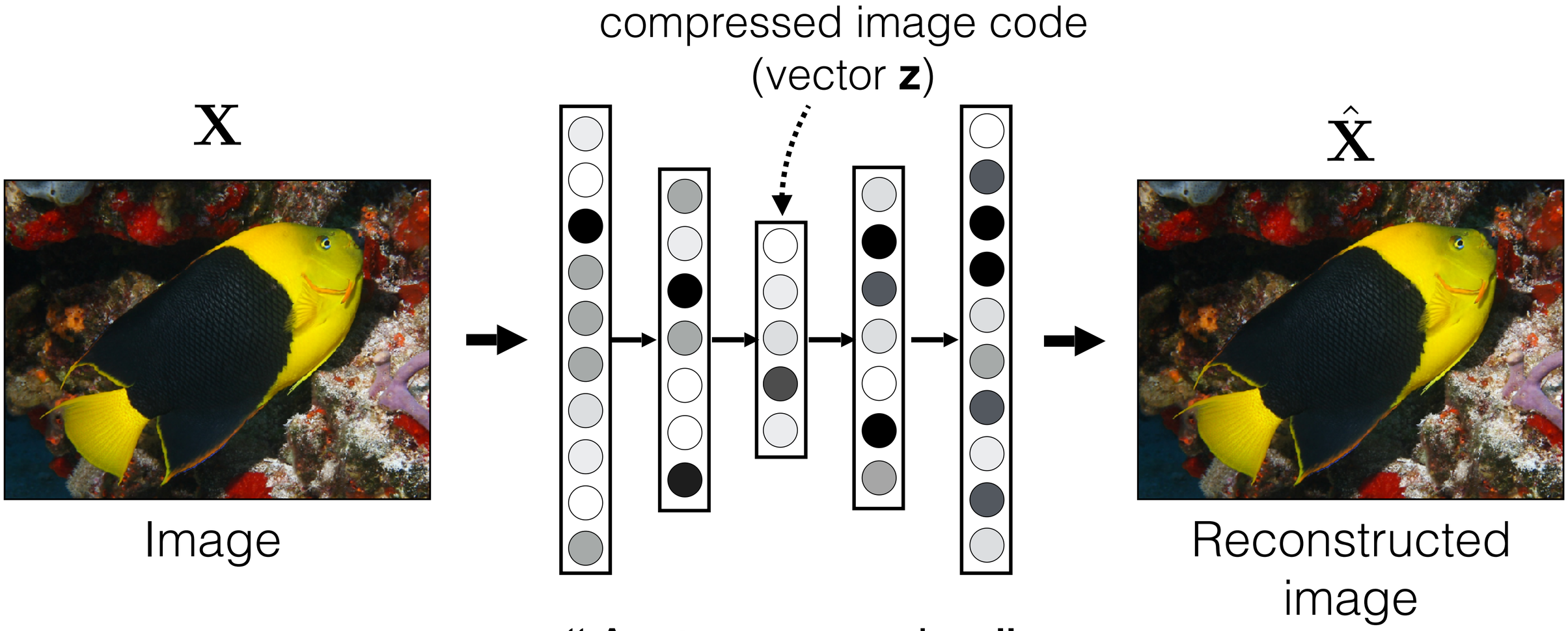


Compact mental representation

Unsupervised Representation Learning



Unsupervised Representation Learning



"Autoencoder"

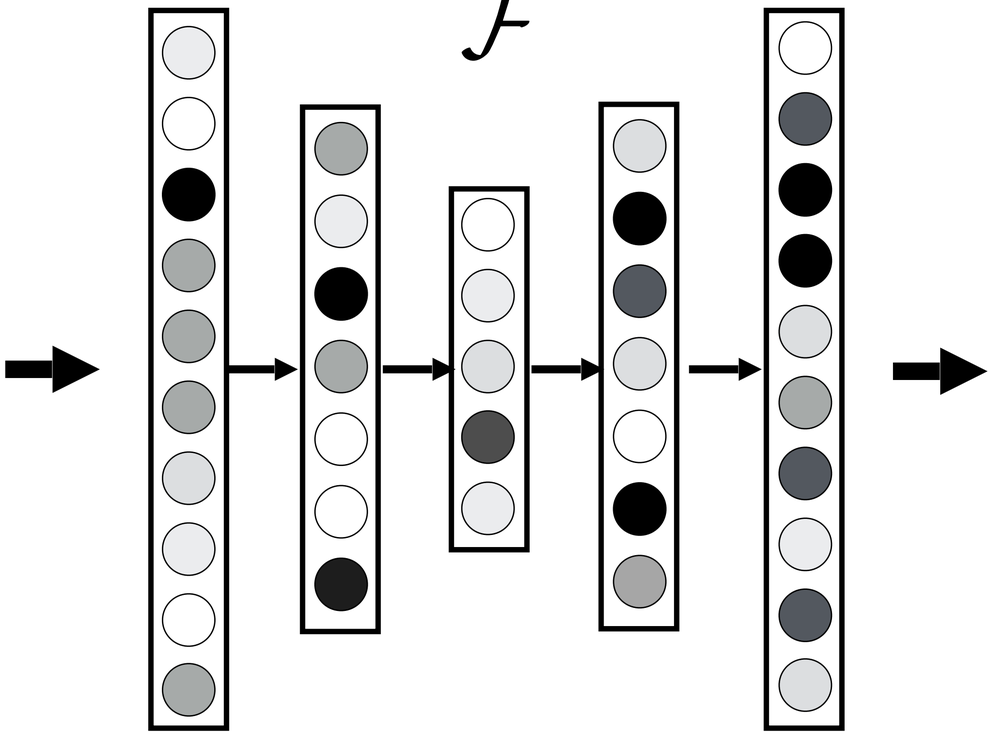
[e.g., Hinton & Salakhutdinov, Science 2006]

Autoencoder

$\mathbf{X}$

Image

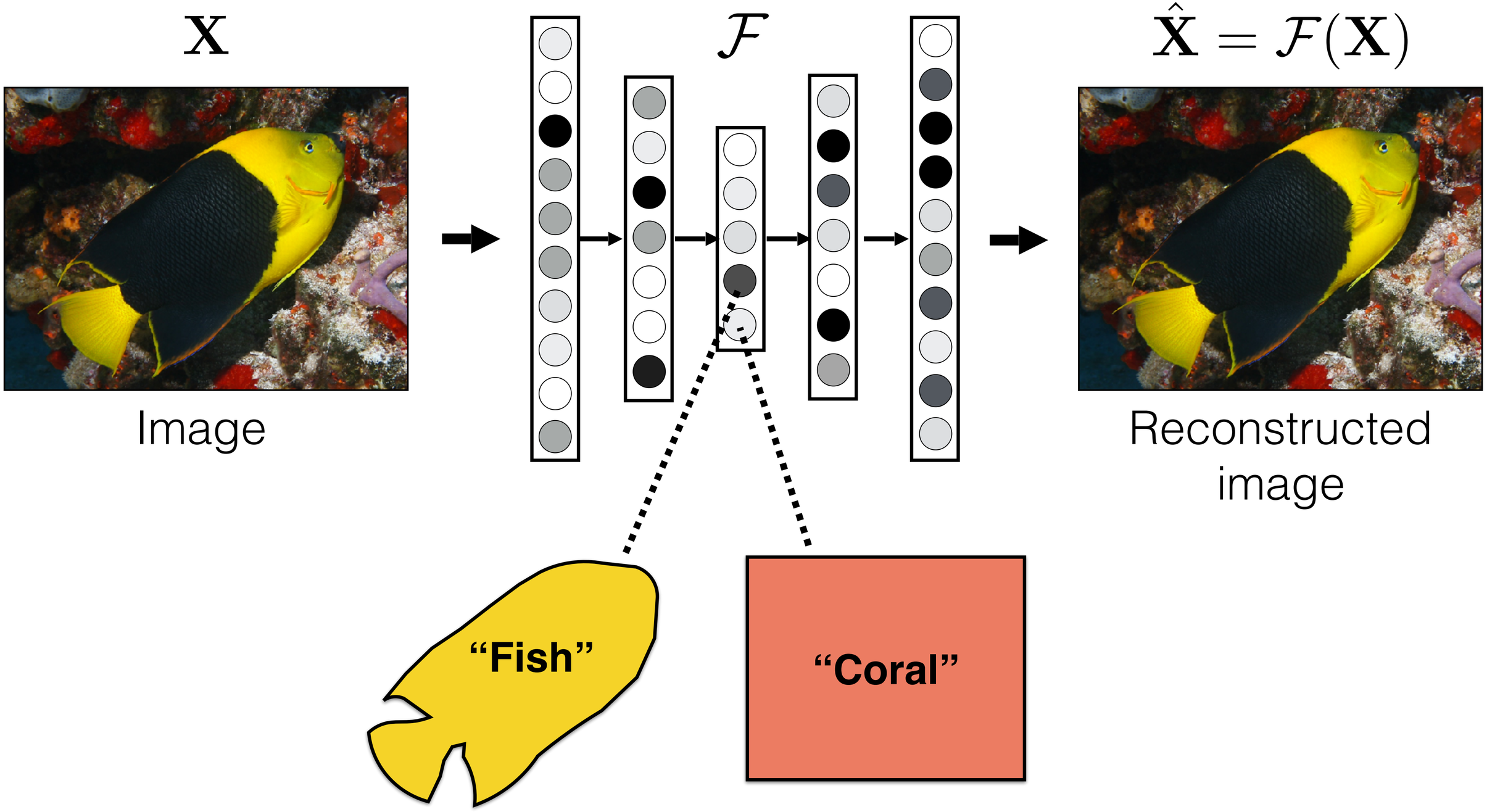
$\mathcal{F}$



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

Reconstructed image

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


$\mathbf{X}$
$\mathcal{F}$
$\hat{\mathbf{X}} = \mathcal{F}(\mathbf{X})$
Image
Reconstructed image
"Fish"
"Coral"

Autoencoder

Data

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

Learner

Objective

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

Hypothesis space

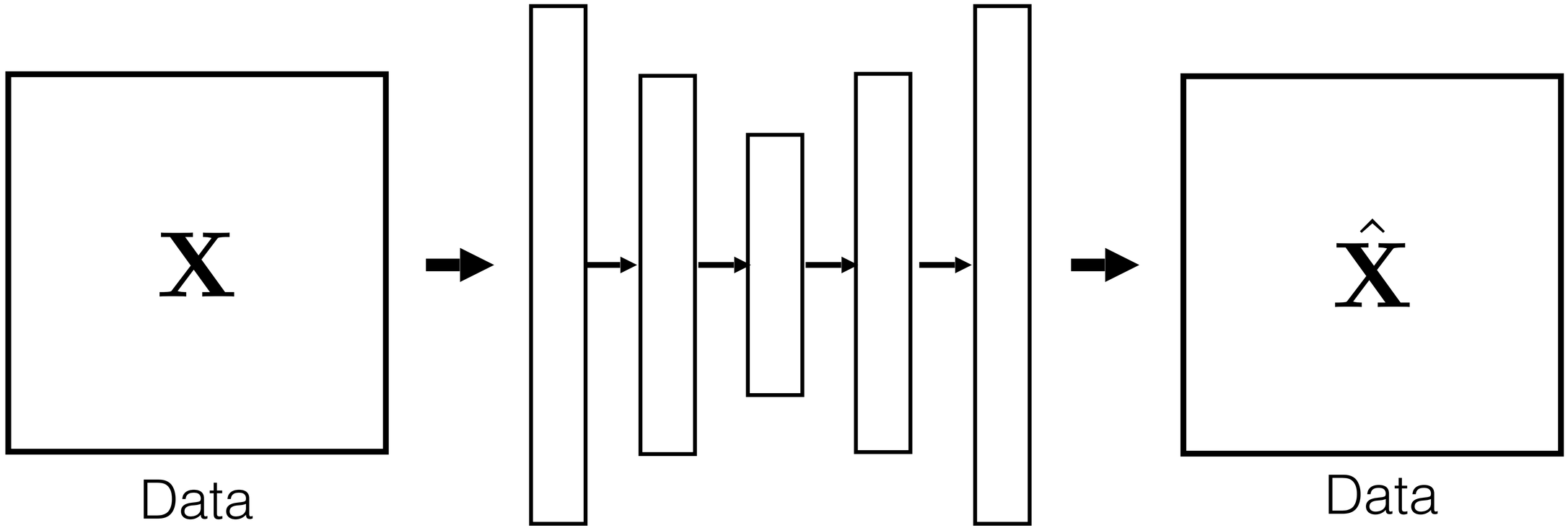
Neural net with a bottleneck

Optimizer

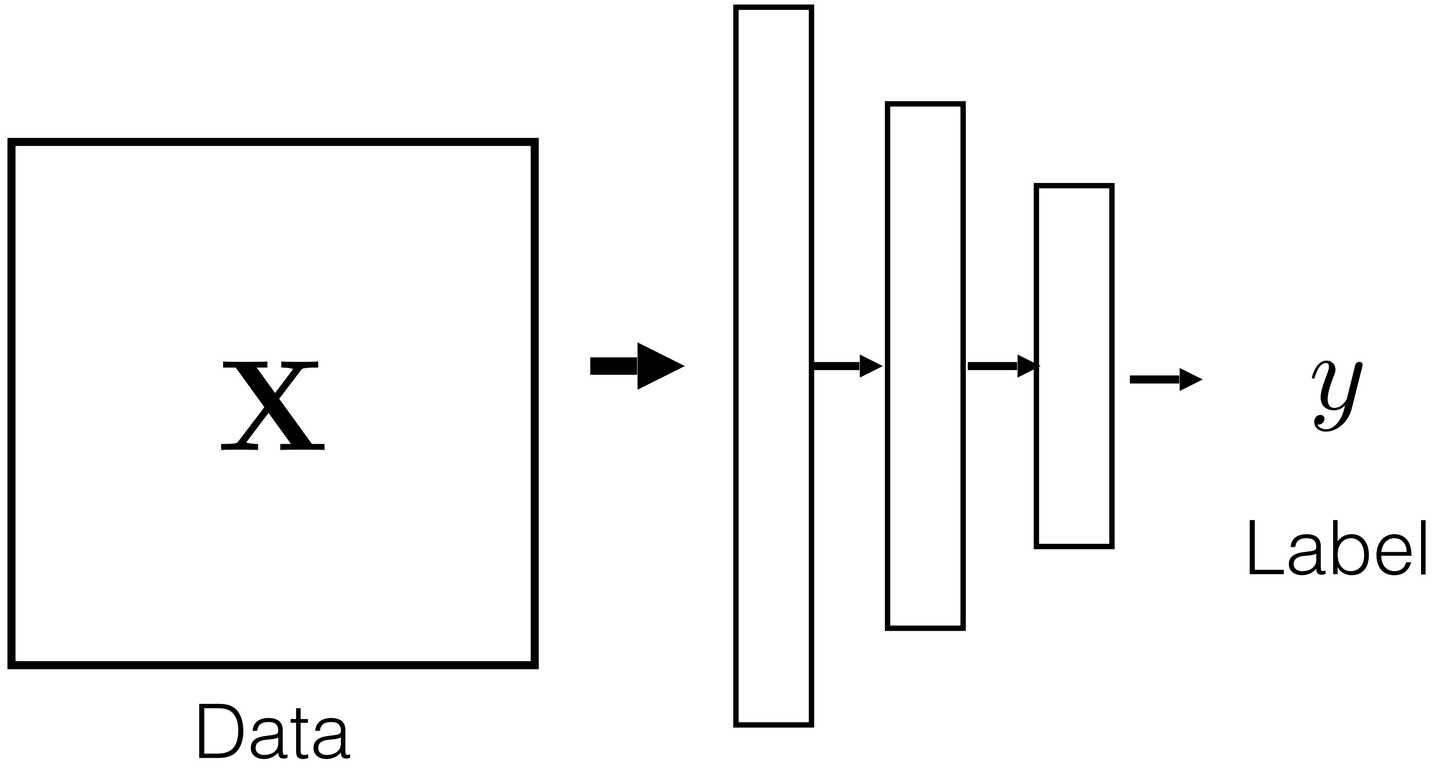
SGD

$\rightarrow f$

Data compression



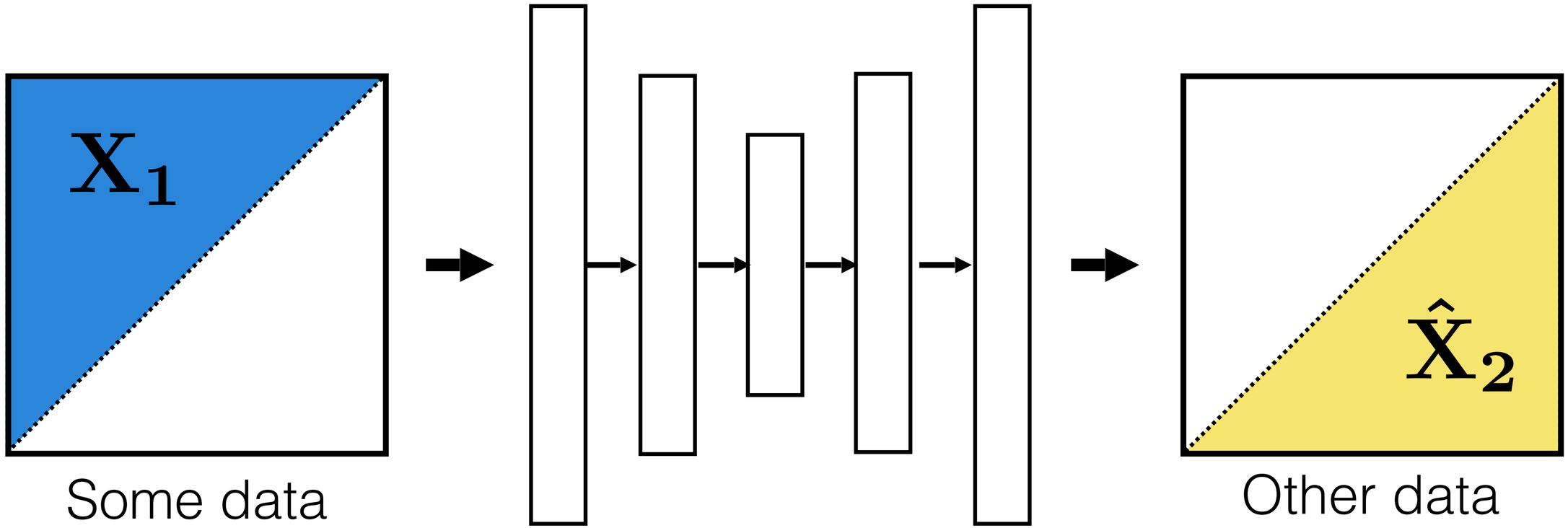
Label prediction

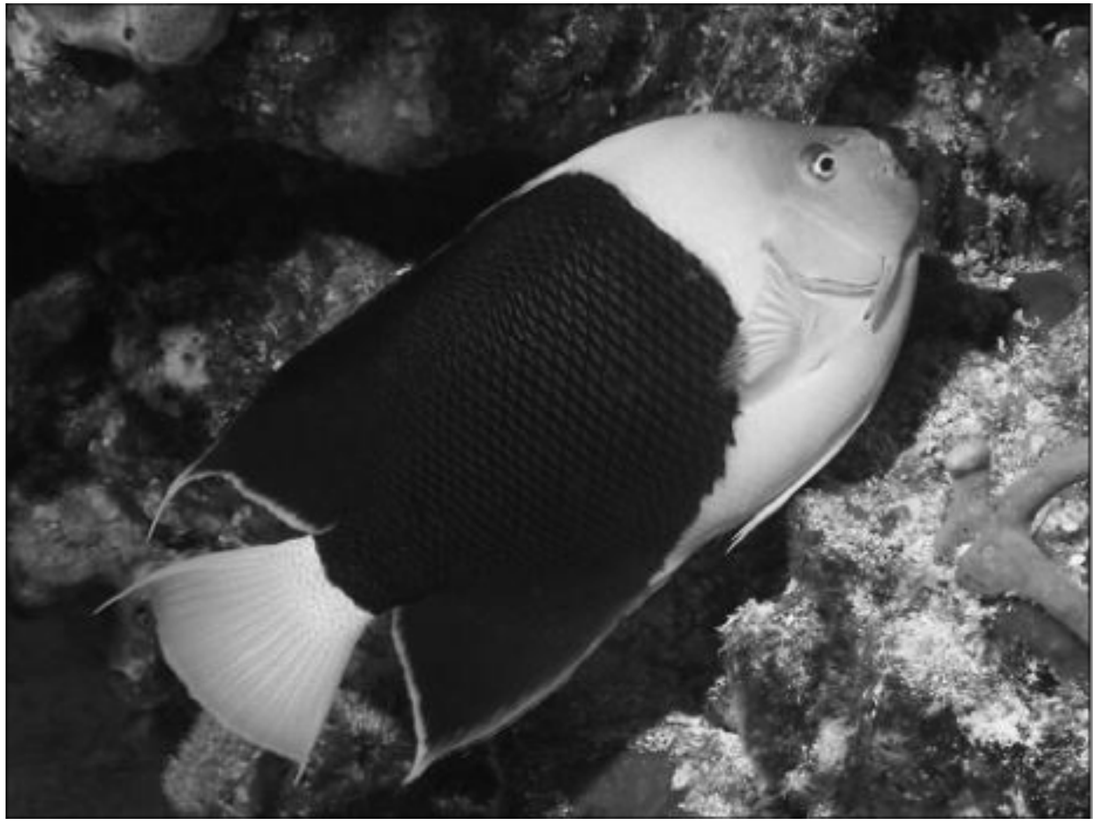


e.g., image classification

Data prediction

aka “self-supervised learning”





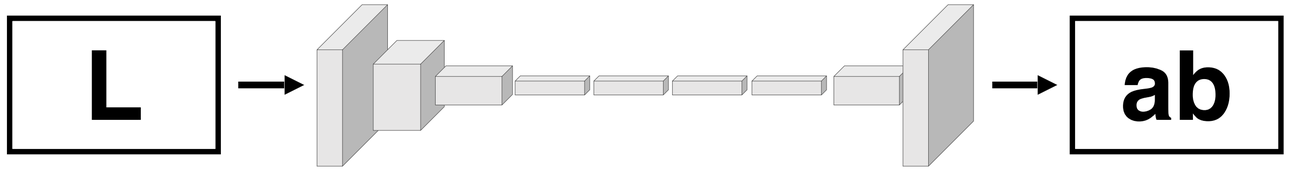
$\mathcal{F}$

Grayscale image: L channel

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

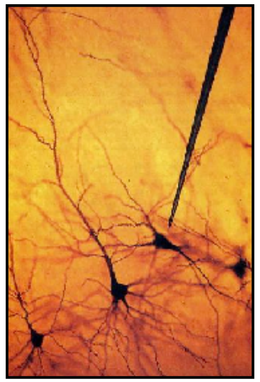
Color information: ab channels

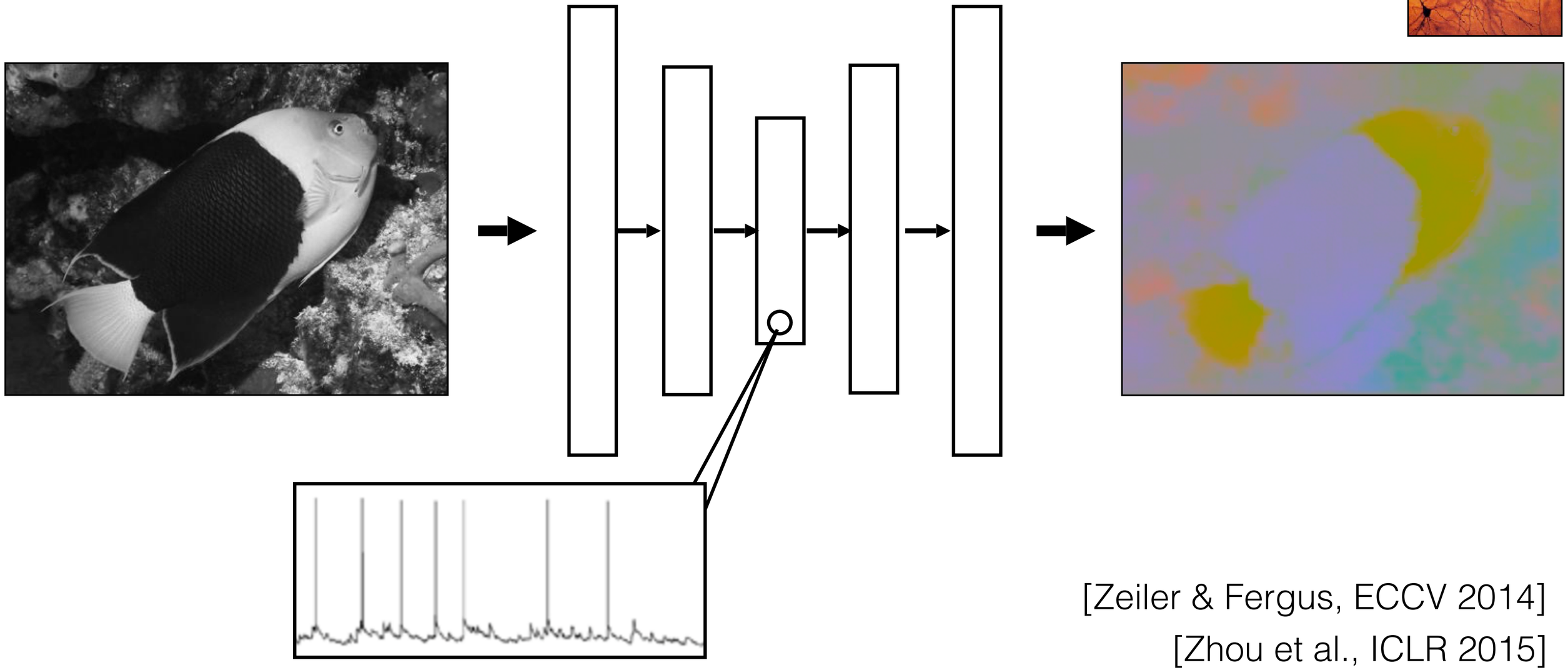
$\hat{\mathbf{Y}} \in \mathbb{R}^{H \times W \times 2}$



[Zhang, Isola, Efros, ECCV 2016]

Deep Net “Electrophysiology”





[Zeiler & Fergus, ECCV 2014]

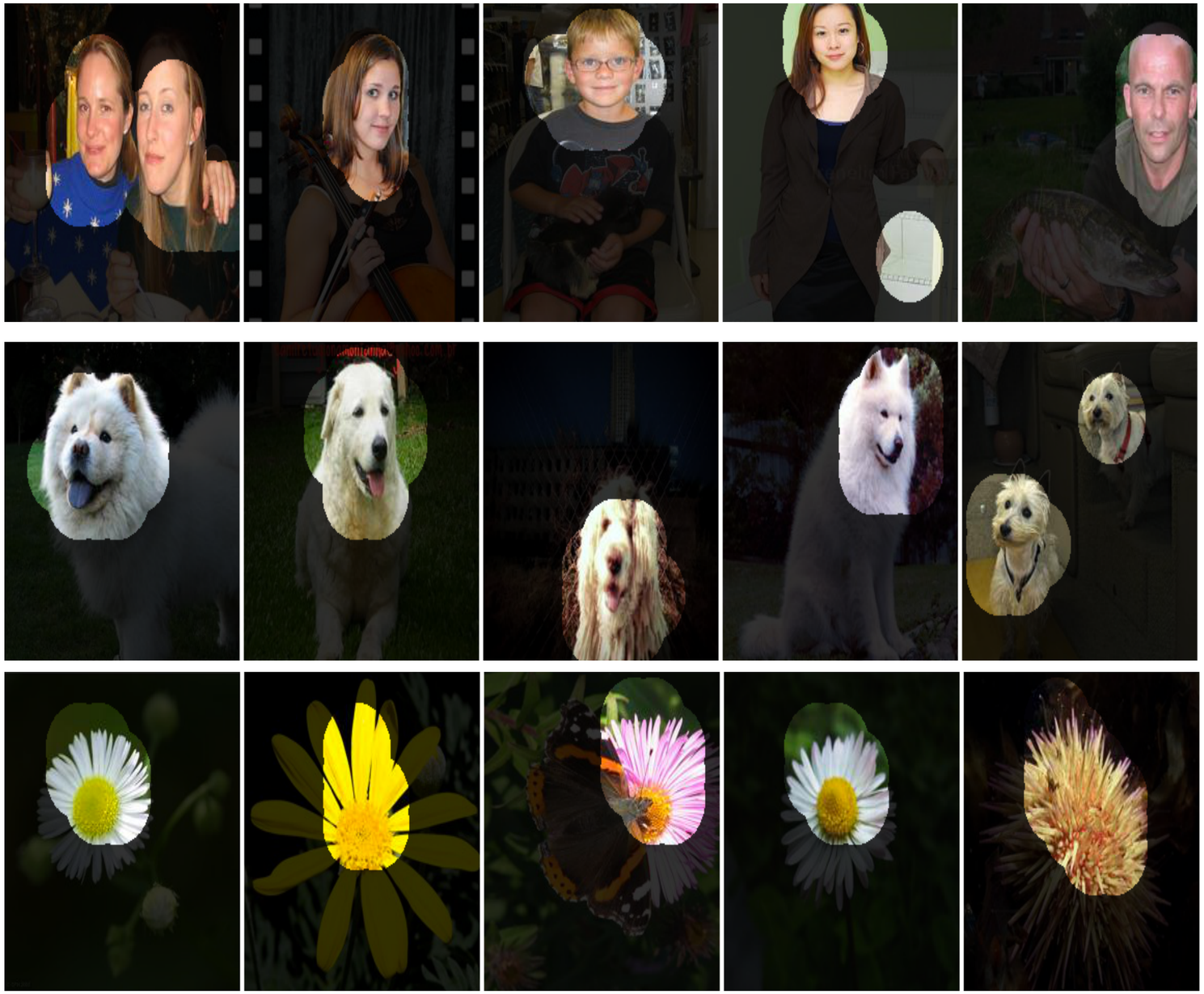
[Zhou et al., ICLR 2015]

Stimuli that drive selected neurons (conv5 layer)

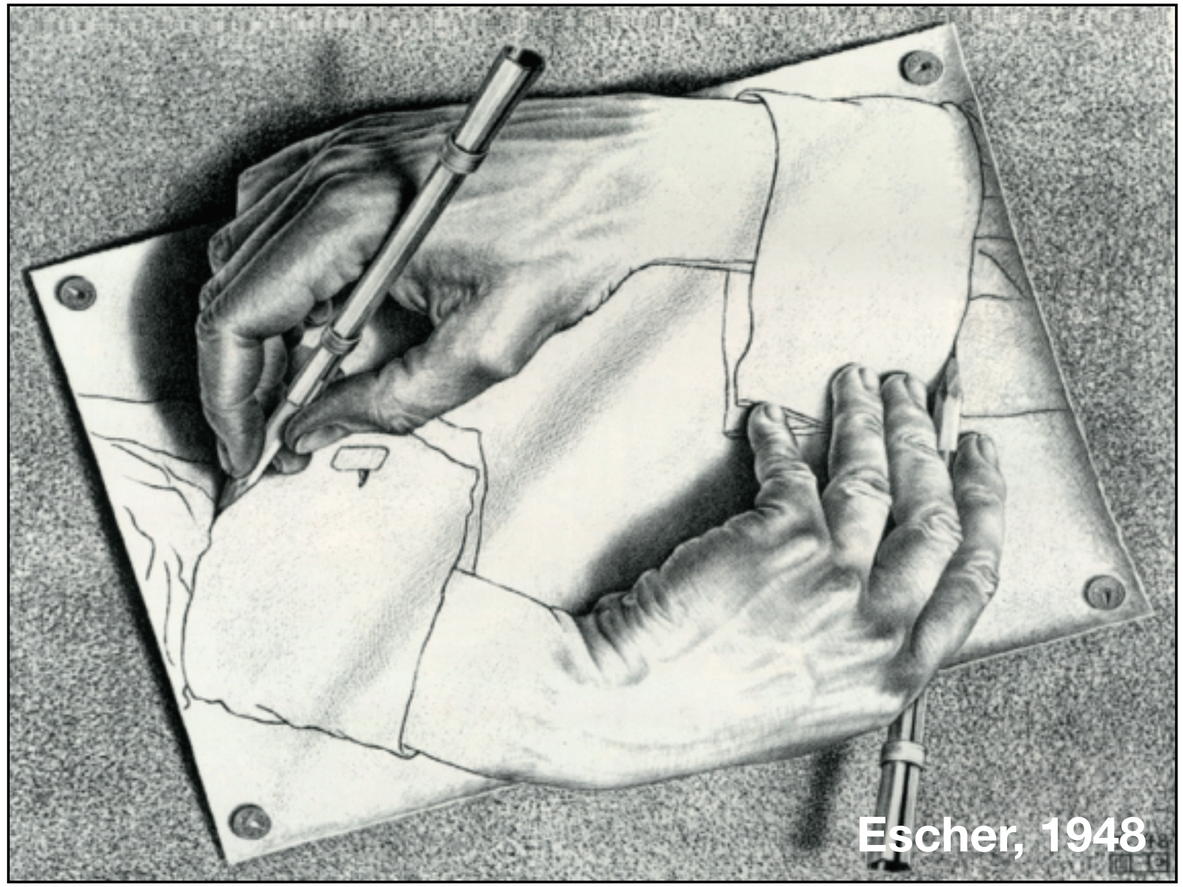
faces

dog faces

flowers



Self-supervised learning



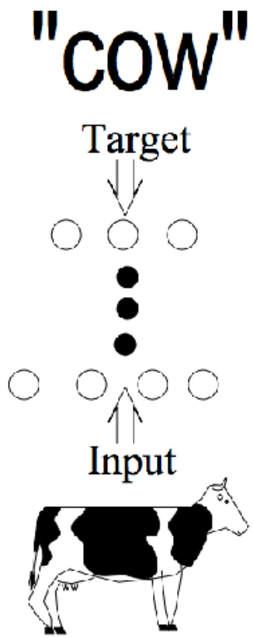
Common trick:

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

Multisensory self-supervision

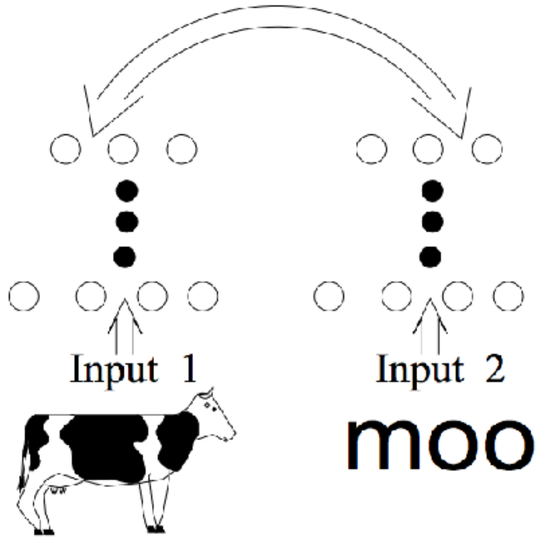
Supervised

- implausible label



Self-Supervised

- derives label from a co-occurring input to another modality



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

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

“Multiview” self-supervised learning

Data

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

Learner

Objective

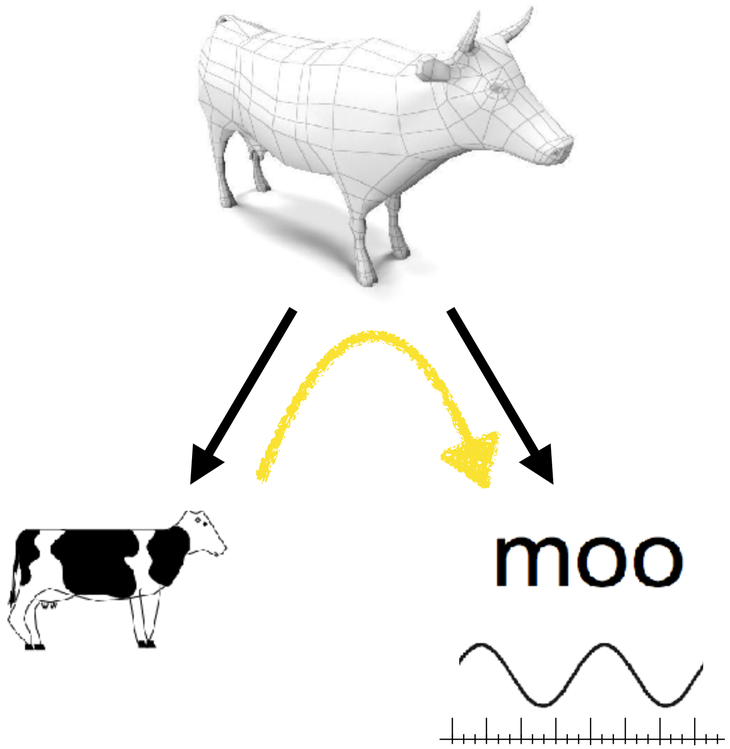
$$\arg\min_f \sum_i D(f(g(\mathbf{x}_i)) - h(\mathbf{x}_i))$$

$\rightarrow f$

Distance function

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

The allegory of the cave



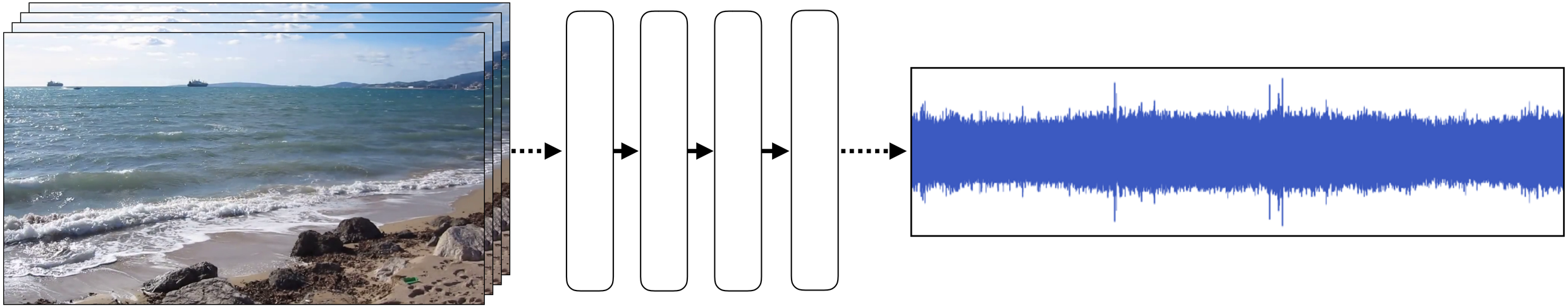

Predicted sound
X

Ambient Sound Provides Supervision for Visual Learning

Andrew Owens Jiajun Wu Josh McDermott
William Freeman Antonio Torralba

[Slide credit: Andrew Owens]

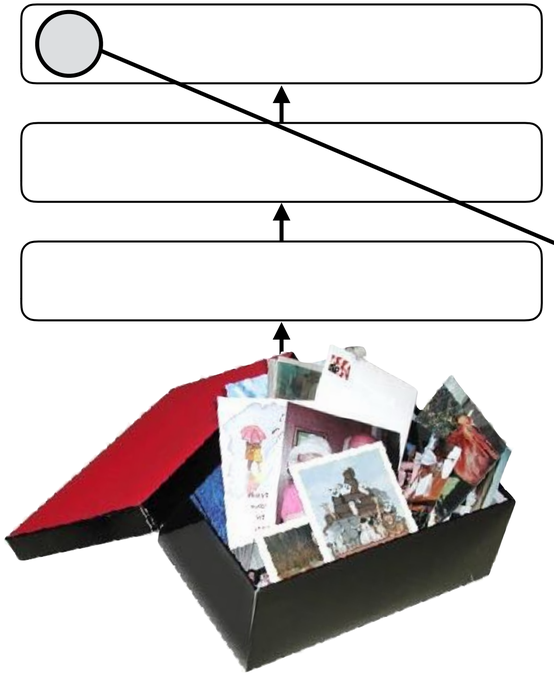
Predicting ambient sound



[Slide credit: Andrew Owens]

What did the model learn?

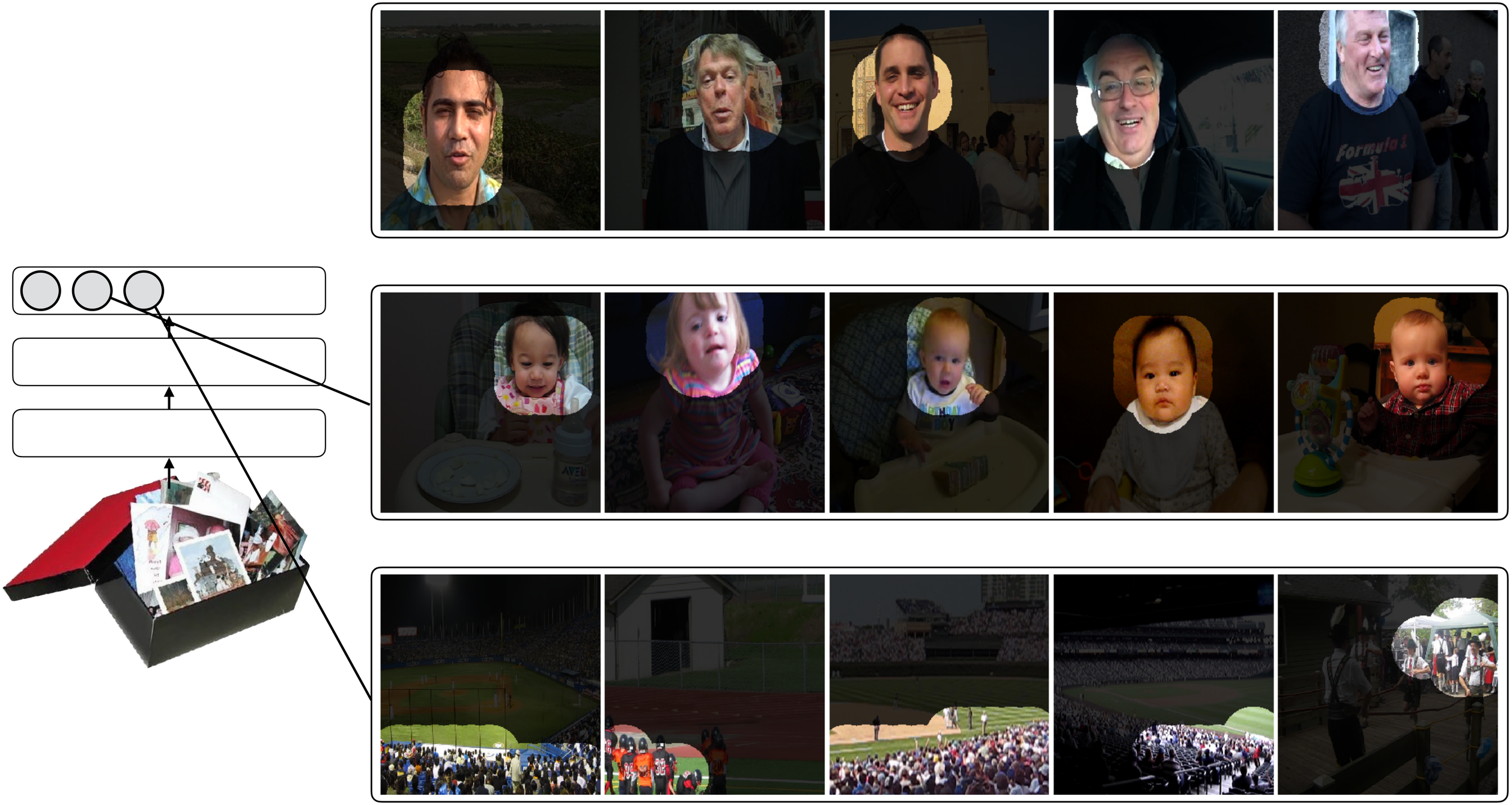
Unit #90 of 256



Strongest responses in dataset

Visualization method from (Zhou 2015)

[Slide credit: Andrew Owens]

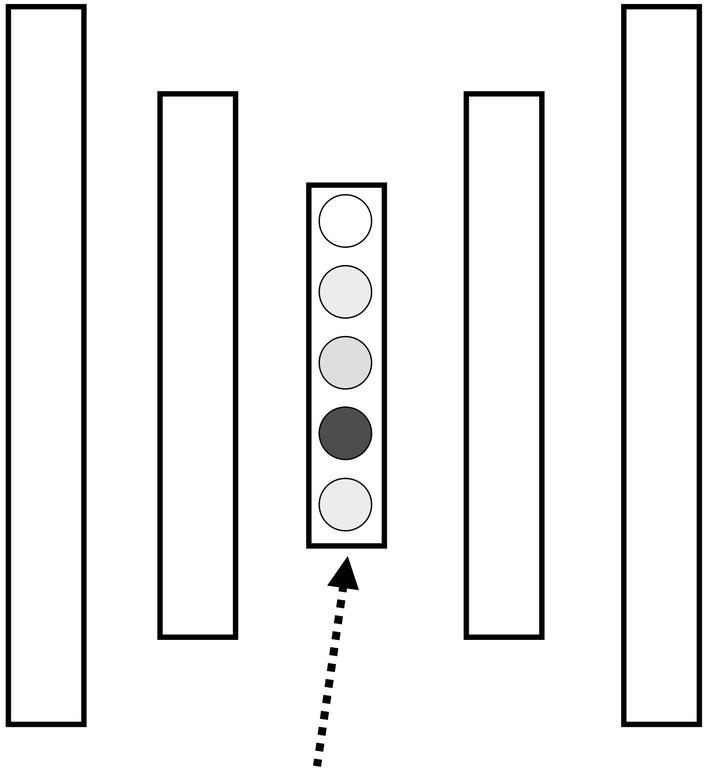


[Slide credit: Andrew Owens]

[Slide credit: Andrew Owens]

$\mathbf{x}$

Image



compressed image code
(vector **z**)

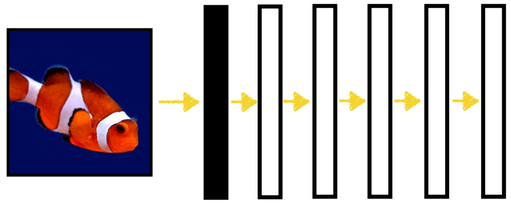
$\hat{\mathbf{x}}$

Reconstructed image

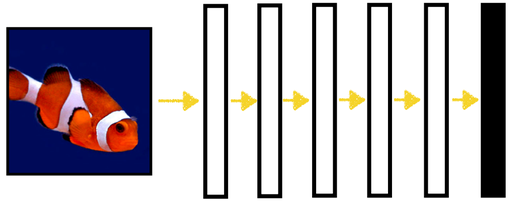
Is the code informative about object class y?

Logistic regression:

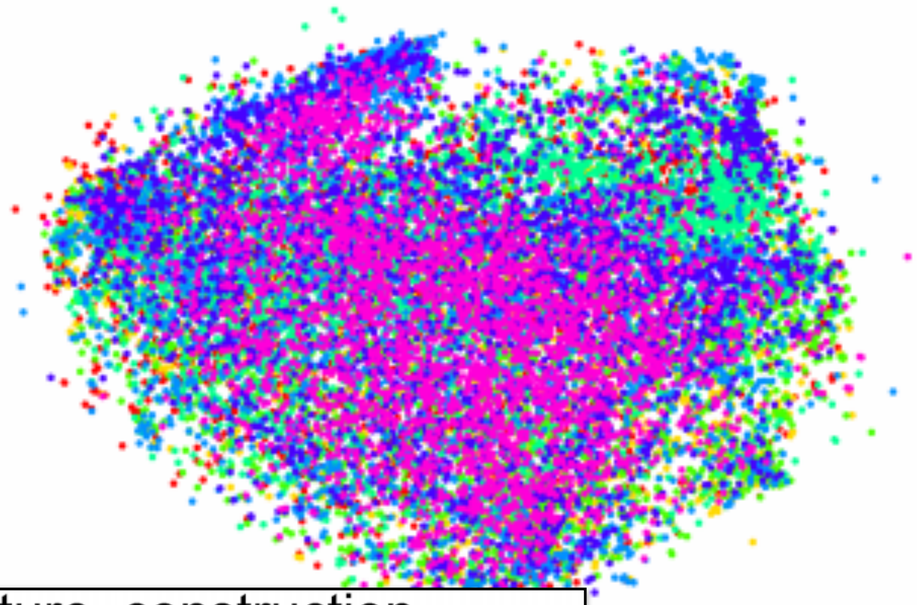
$$y = \sigma(\mathbf{W}\mathbf{z} + \mathbf{b})$$

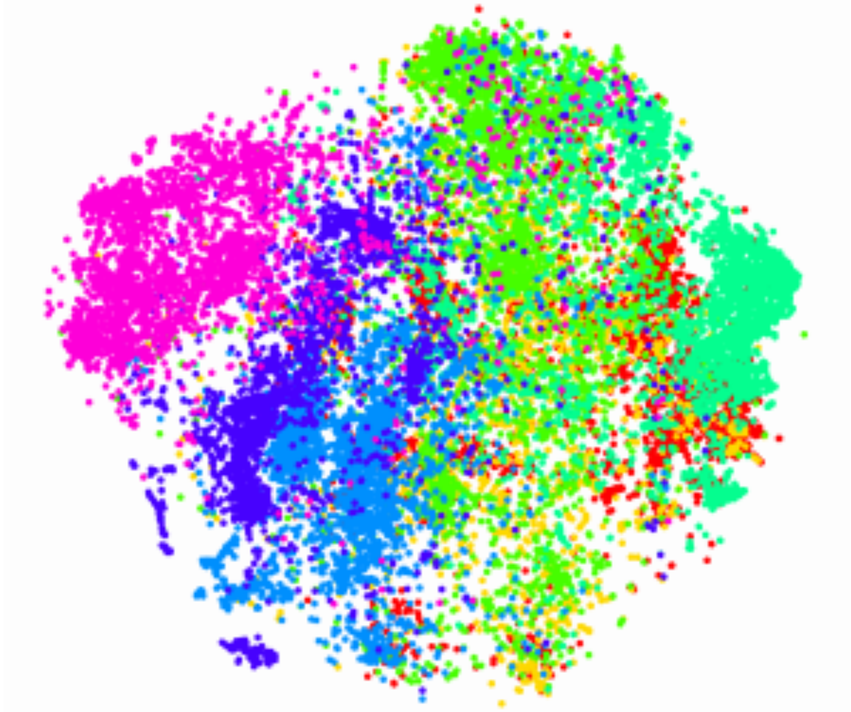


Layer 1 representation



Layer 6 representation

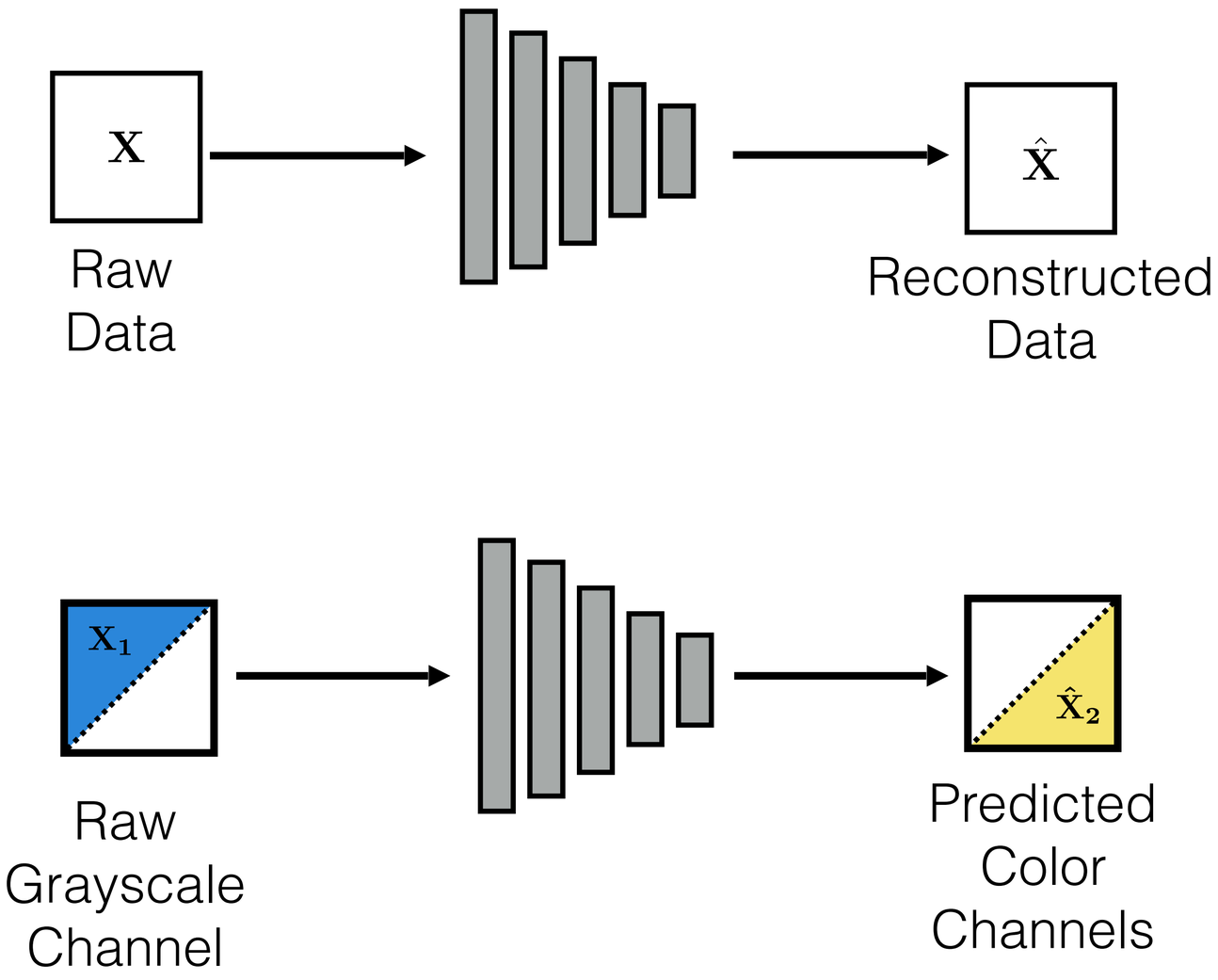


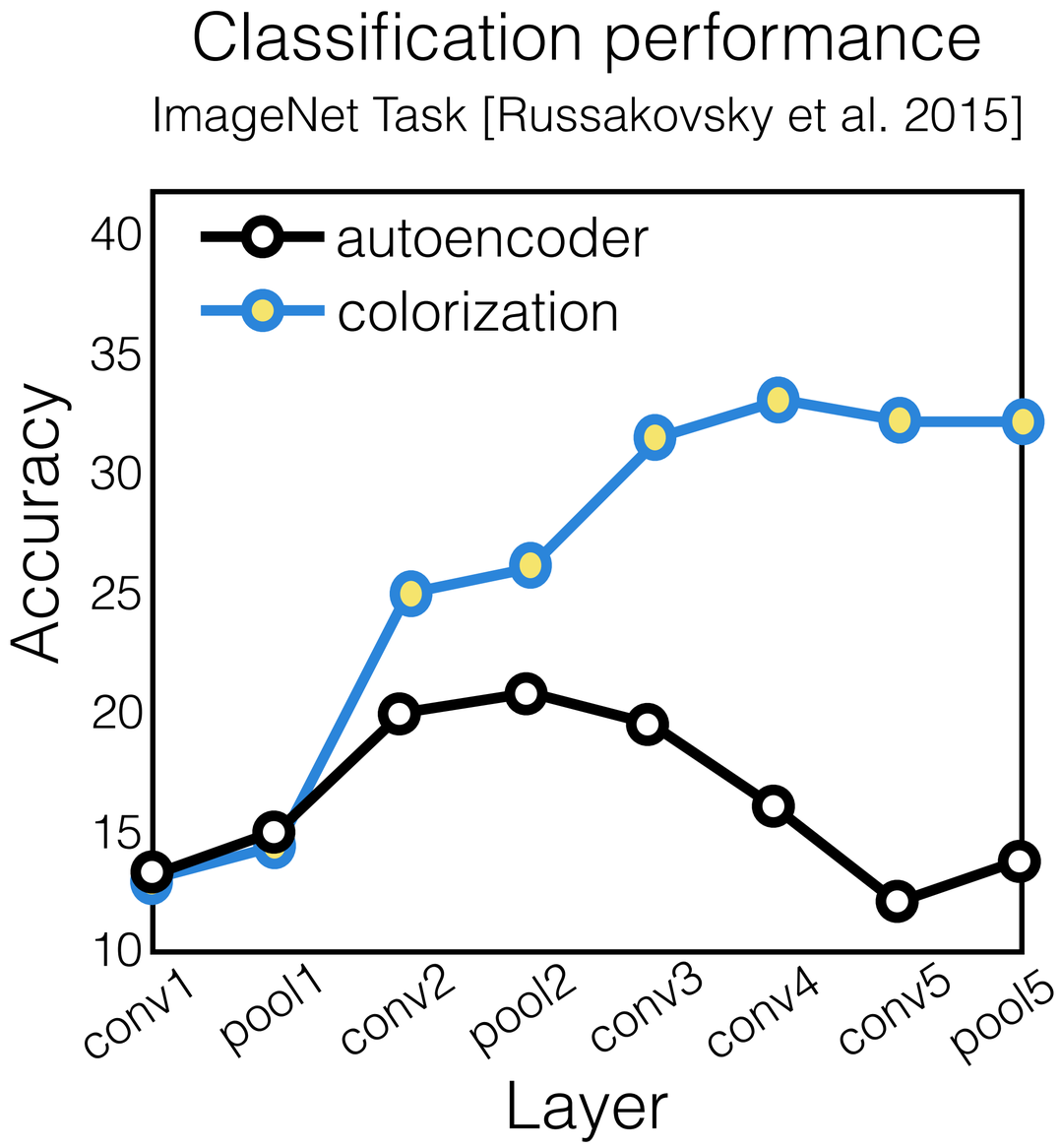


- structure, construction
- covering
- commodity, trade good, good
- conveyance, transport
- invertebrate
- bird
- hunting dog

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

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

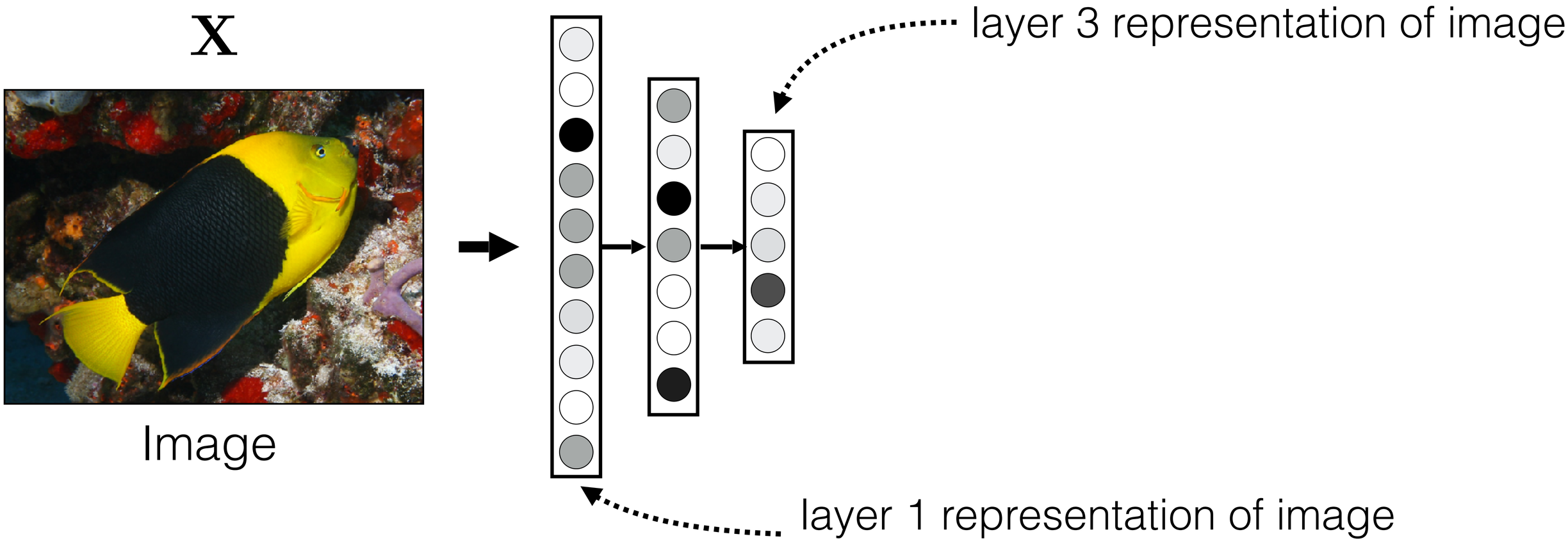

$\mathbf{x}$
Raw Data
$\hat{\mathbf{x}}$
Reconstructed Data
$\mathbf{x}_1$
Raw Grayscale Channel
$\hat{\mathbf{x}}_2$
Predicted Color Channels


Classification performance
ImageNet Task [Russakovsky et al. 2015]
autoencoder
colorization
Accuracy
40
35
30
25
20
15
10
conv1
pool1
conv2
pool2
conv3
conv4
conv5
pool5
Layer

How to represent words as numbers

“Fish”	→	[**1**,0,0,0,0,0,0,…]
“Shark”	→	[0,**1**,0,0,0,0,0,…]
“Whale”	→	[0,0,**1**,0,0,0,0,…]
“Water”	→	[0,0,0,**1**,0,0,0,…]
“Cat”	→	[0,0,0,0,**1**,0,0,…]
“Couch”	→	[0,0,0,0,0,**1**,0,…]
“Sun”	→	[0,0,0,0,0,0,**1**,…]

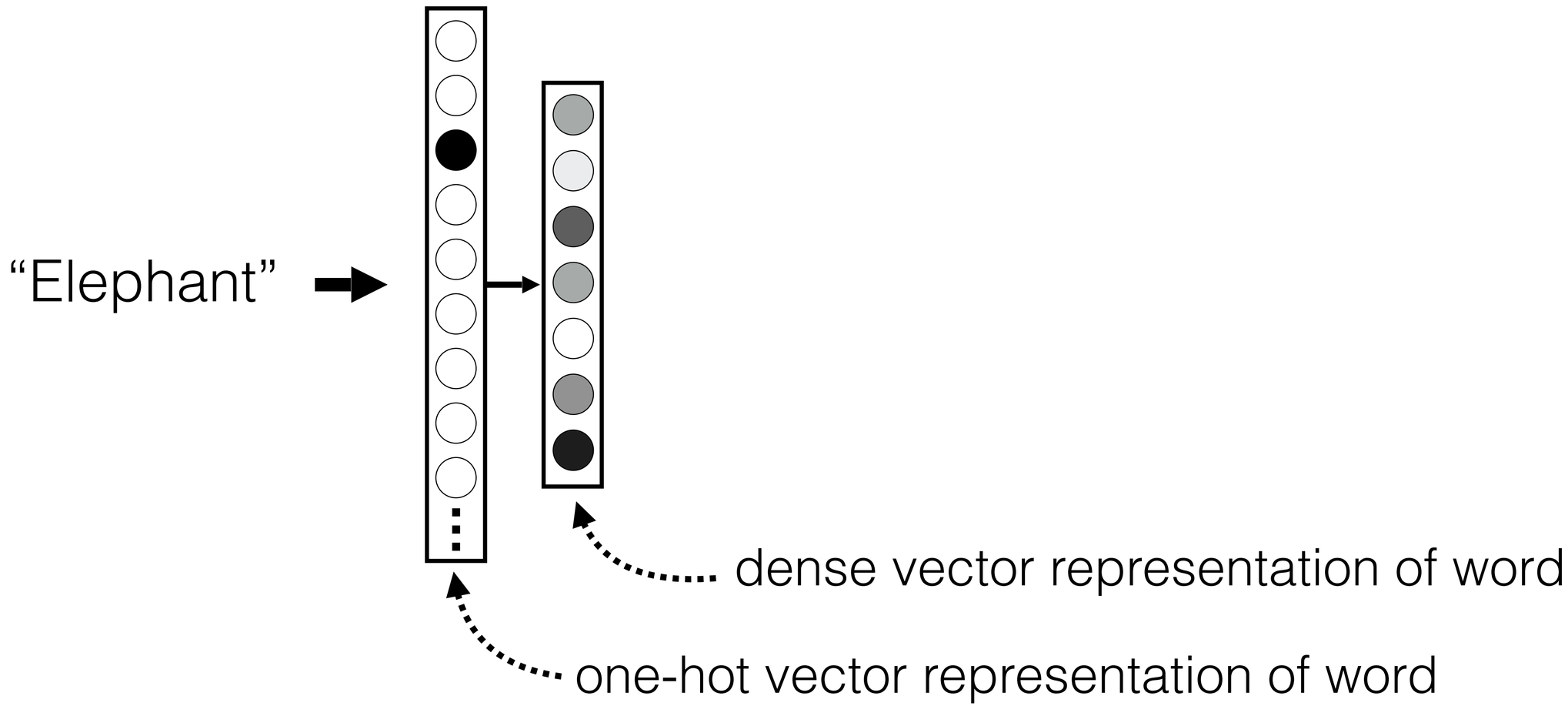
im2vec



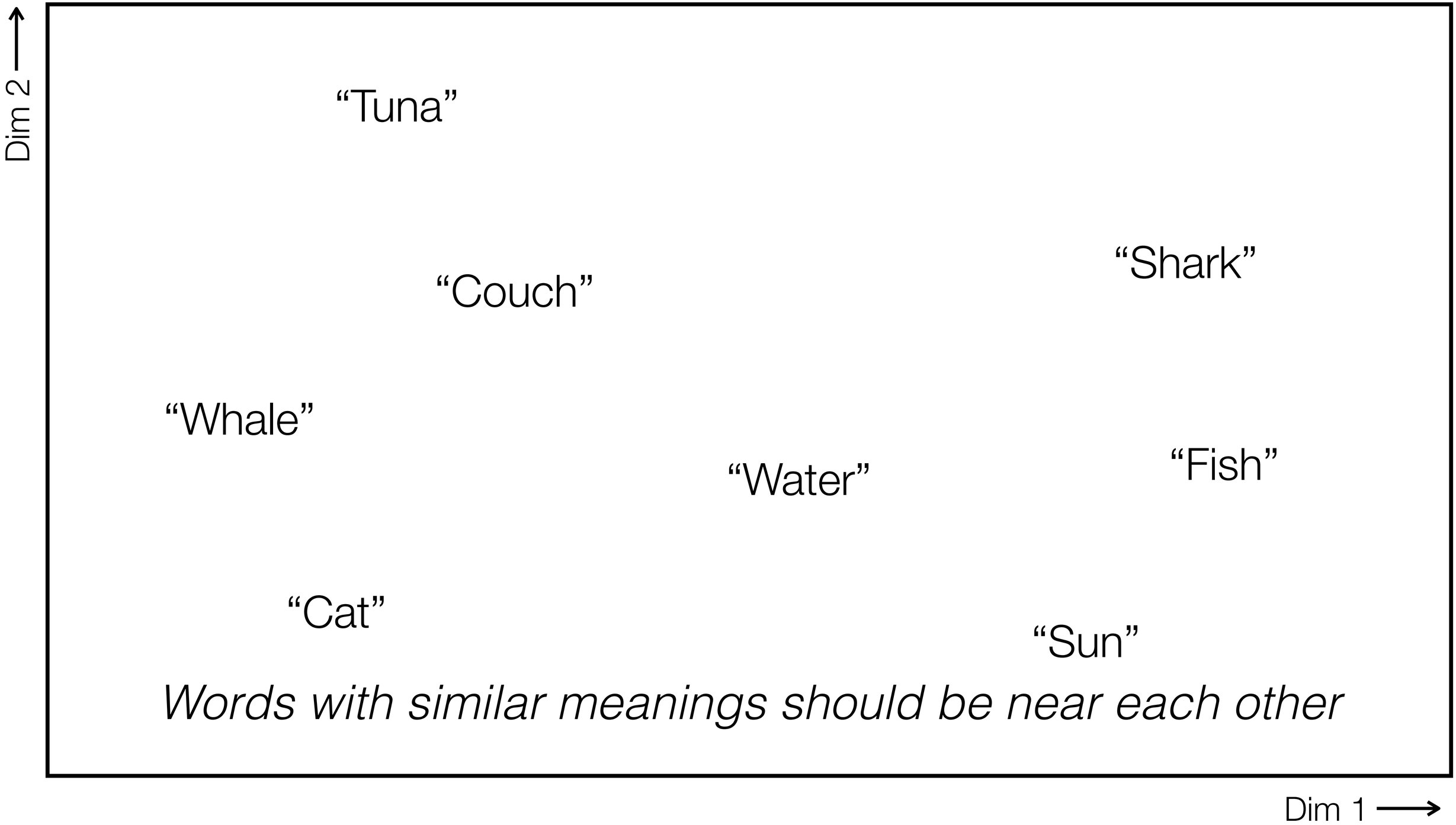
Represent image as a vector of neural activations

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

word2vec



X2vec methods are also called embeddings of X, e.g., a **word embedding**


Dim 2
"Tuna"
"Shark"
"Couch"
"Whale"
"Fish"
"Water"
"Cat"
"Sun"
Words with similar meanings should be near each other
Dim 1

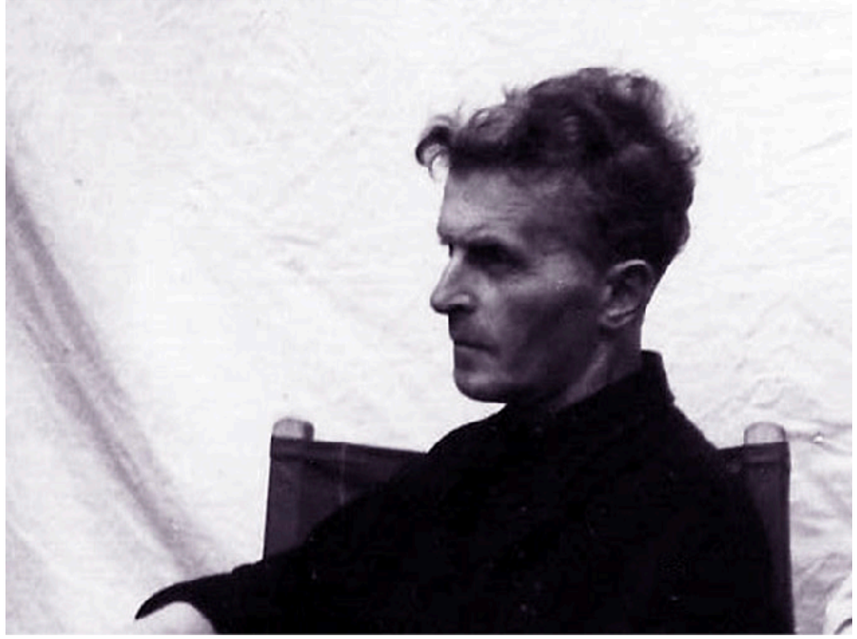
word2vec

Words with similar meanings should be near each other

Proxy: words that are used in the same context tend to have similar meanings

words with similar contexts should be near each other

“Meaning is use” — Wittgenstein



Next to the 'sofa' is a desk, and a 'person' is sitting behind it.

'armchair'	'man'
'bench'	'woman'
'chair'	'child'
'deck chair'	'teenager'
'ottoman'	'girl'
'seat'	'boy'
'stool'	'baby'
'swivel chair'	'daughter'
'loveseat'	'son'
...	...

word2vec

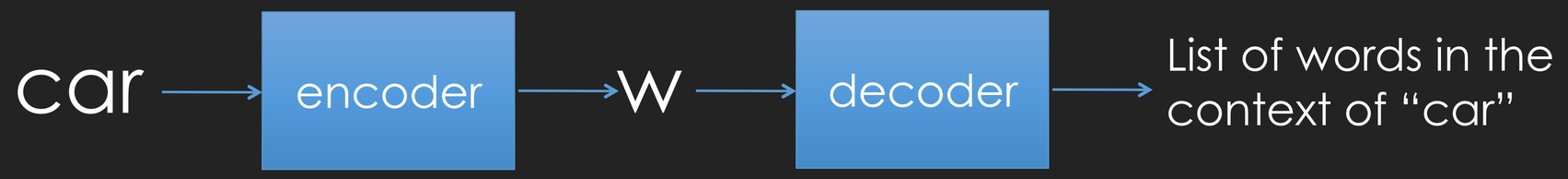
I parked the car in a nearby street. It is a red car with two doors, …

I parked the vehicle in a nearby street…

T. Mikolov, K. Chen, G. Corrado, J. Dean. Efficient Estimation of Word Representations in Vector Space. arXiv:1301.3781, 2013

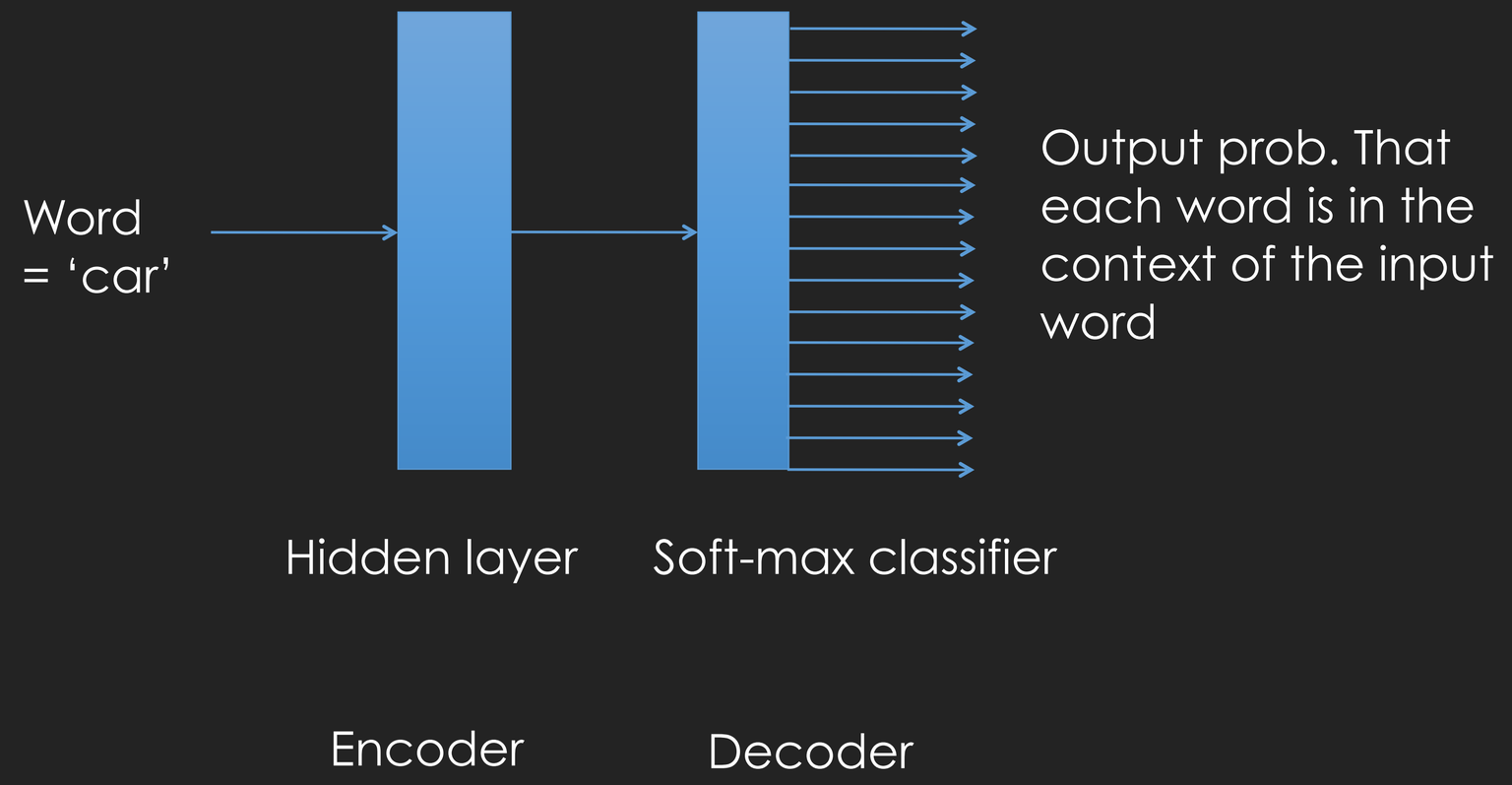
word2vec

I parked the car in a nearby street. It is a red car with two doors, …



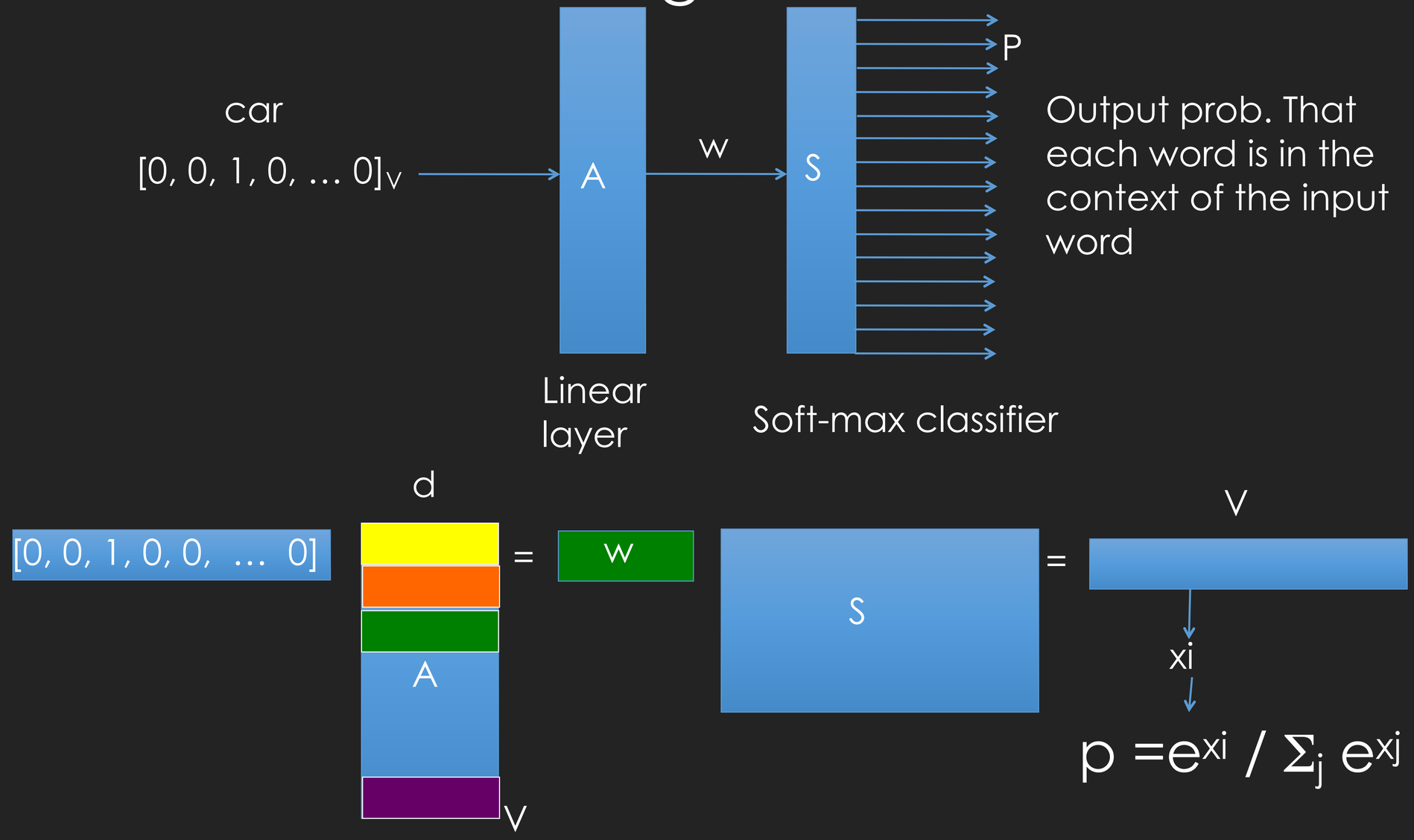
T. Mikolov, K. Chen, G. Corrado, J. Dean. Efficient Estimation of Word Representations in Vector Space. arXiv:1301.3781, 2013

word2vec



T. Mikolov, K. Chen, G. Corrado, J. Dean. Efficient Estimation of Word Representations in Vector Space. arXiv:1301.3781, 2013

word2vec, training



T. Mikolov, K. Chen, G. Corrado, J. Dean. Efficient Estimation of Word Representations in Vector Space. arXiv:1301.3781, 2013

word2vec, training

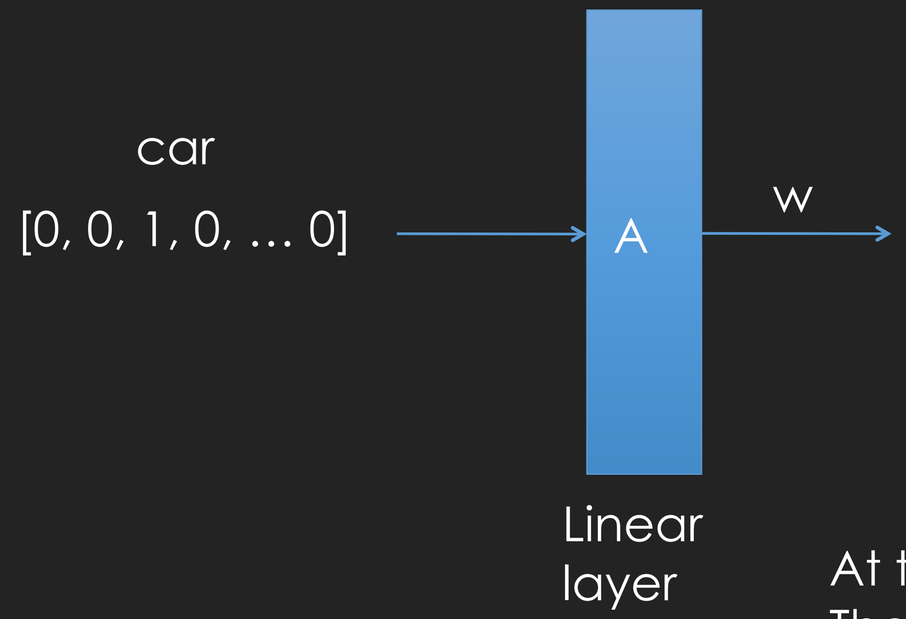
- In training maximize log-likelihood over the training set:

$$\sum_{t=1}^{T} \sum_{i=-c}^{c} \log p(w_{t+i} | w_t)$$

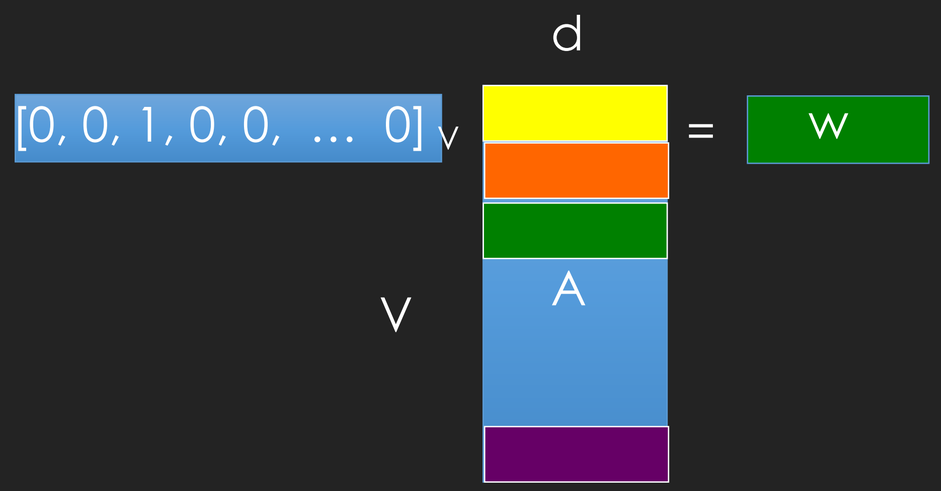
T … training set size
c … context window size

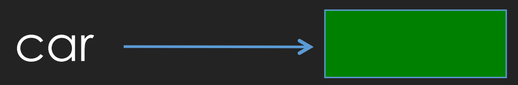
T. Mikolov, K. Chen, G. Corrado, J. Dean. Efficient Estimation of Word Representations in Vector Space. arXiv:1301.3781, 2013

word2vec, test time



At test time, w is our word embedding.
The encoding is just a look up table.



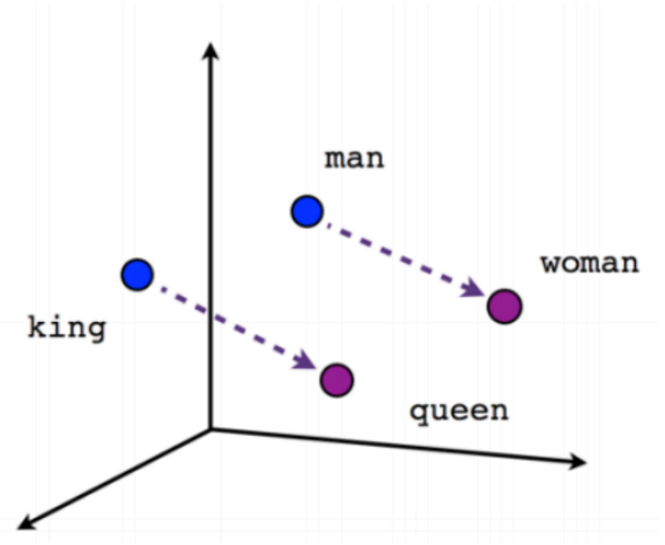


T. Mikolov, K. Chen, G. Corrado, J. Dean. Efficient Estimation of Word Representations in Vector Space. arXiv:1301.3781, 2013

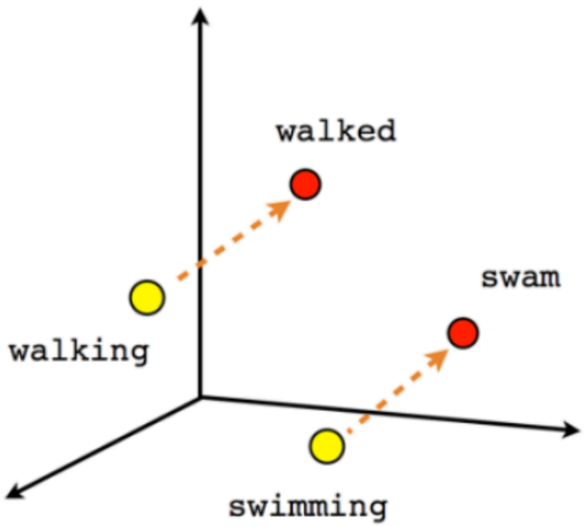
Algebraic operations with the vector representation of words

X = Vector(“Paris”) – vector(“France”) + vector(“Italy”)

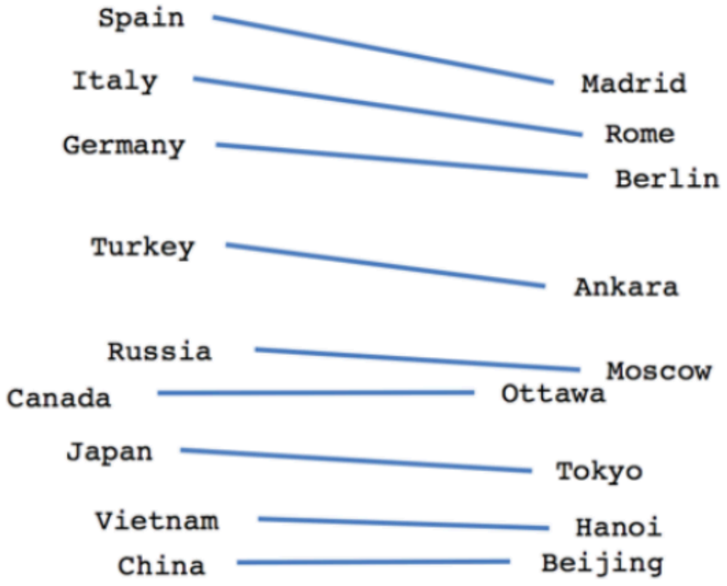
Closest nearest neighbor to X is vector(“Rome”)



Male-Female



Verb tense



Country-Capital

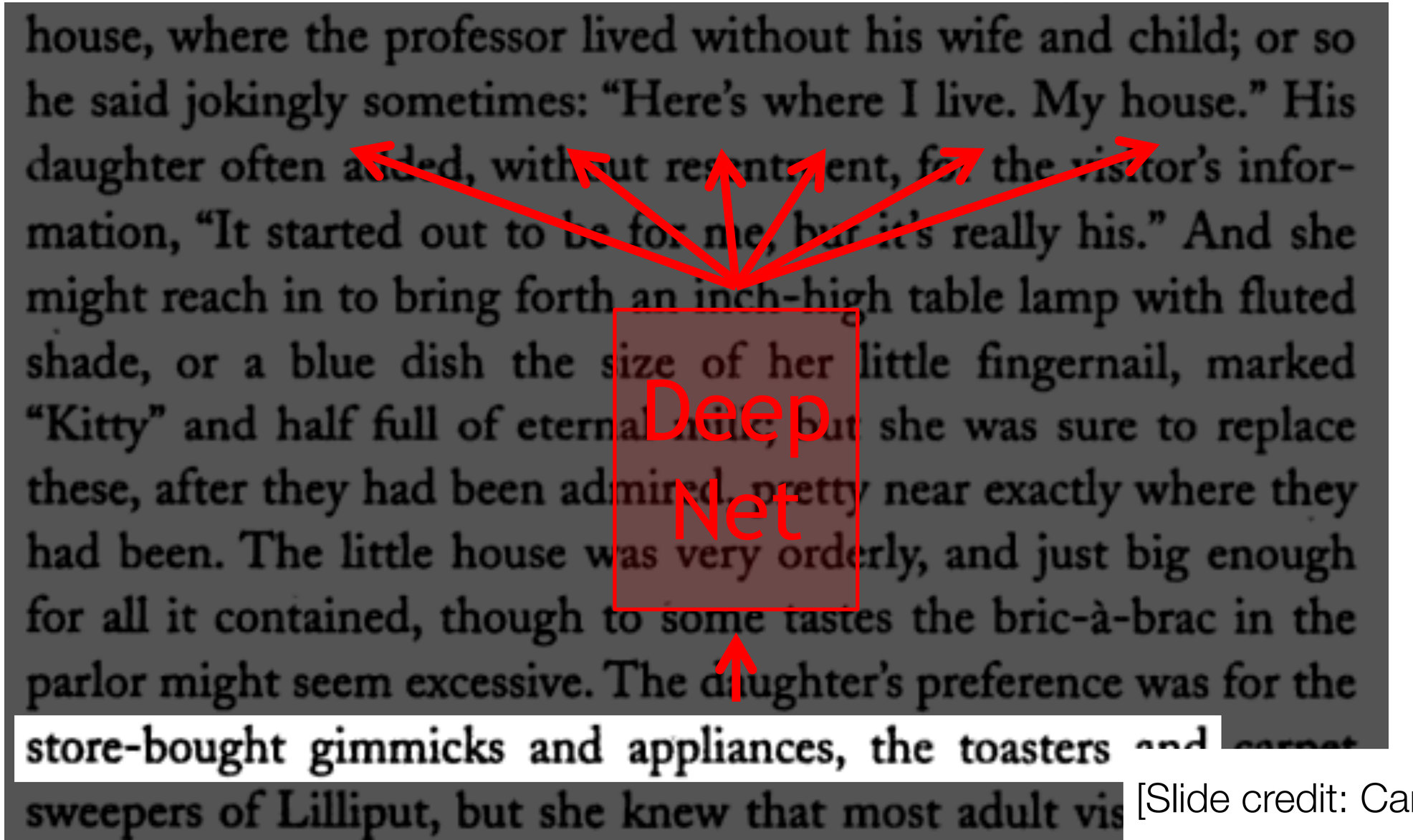
Examples from https://www.tensorflow.org/tutorials/representation/word2vec

Unsupervised visual representation learning by context prediction

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

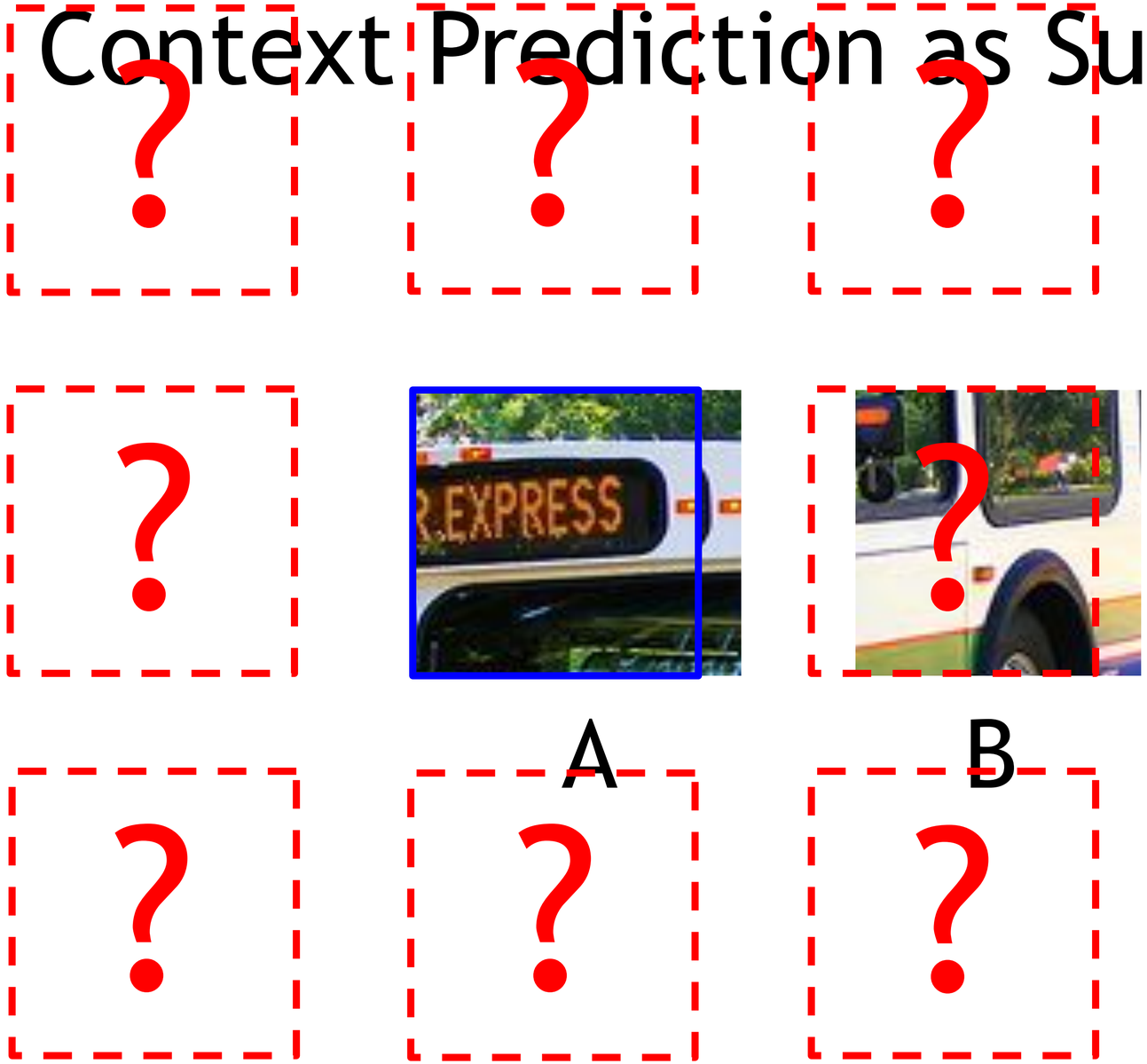
Context as Supervision

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



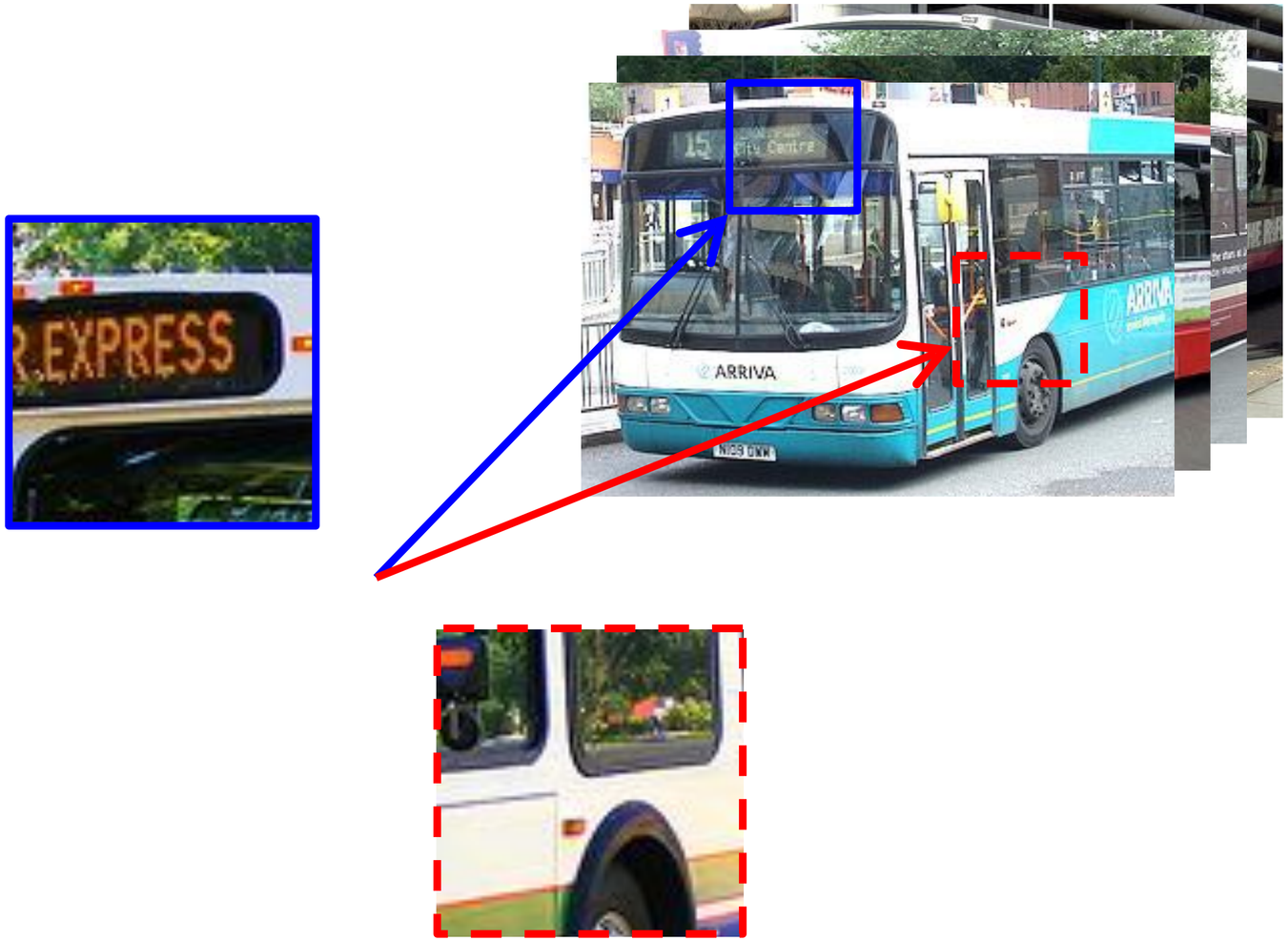
[Slide credit: Carl Doersch]

Context Prediction as Supervision



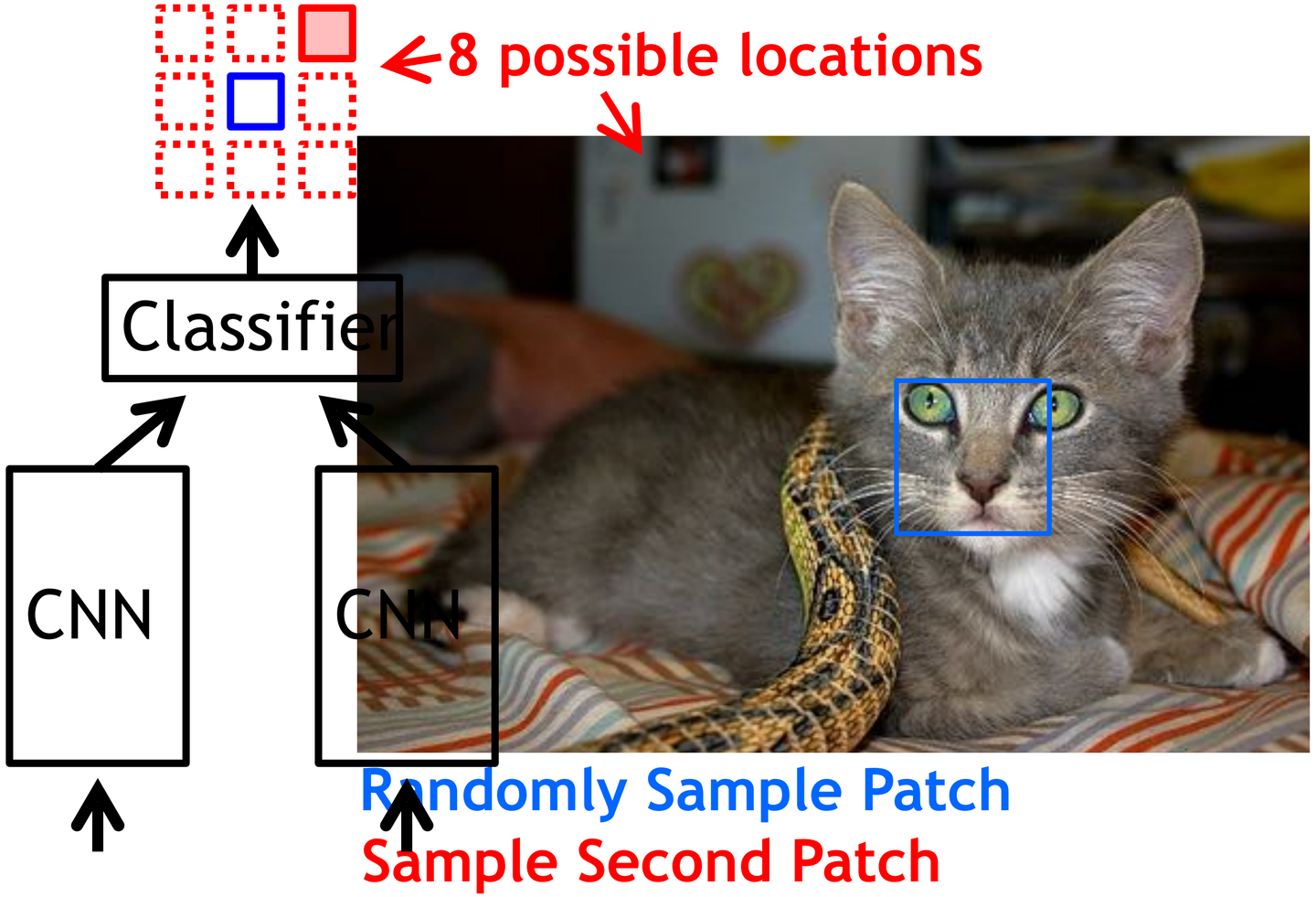
[Slide credit: Carl Doersch]

Semantics from a non-semantic task

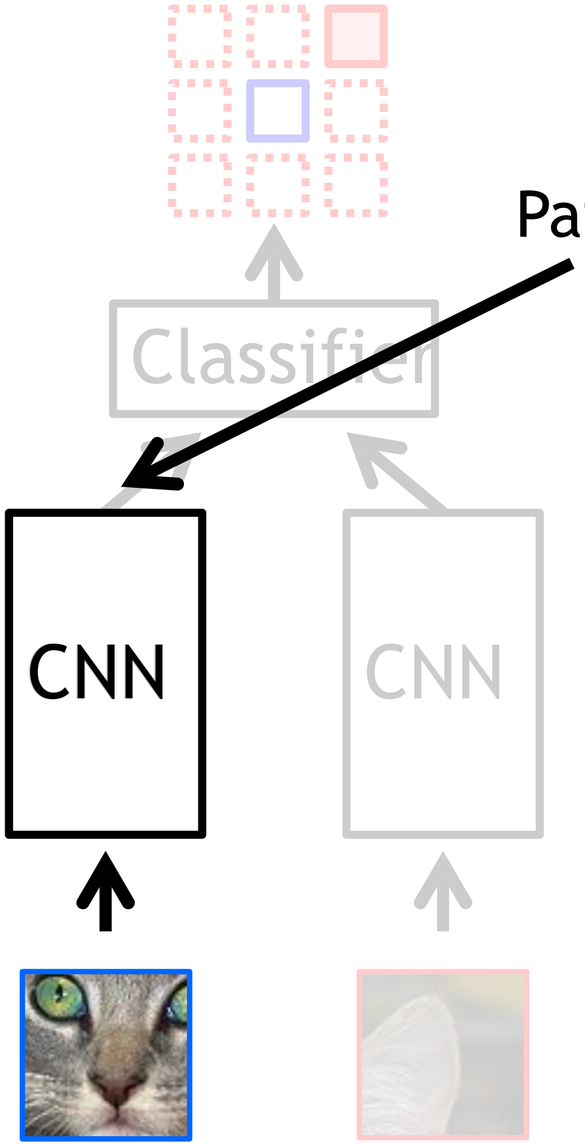


[Slide credit: Carl Doersch]

Relative Position Task



[Slide credit: Carl Doersch]



Patch Embedding (representation)

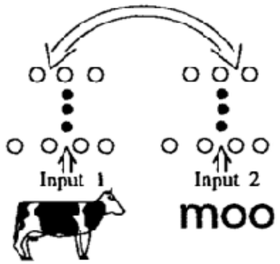
Input

Nearest Neighbors

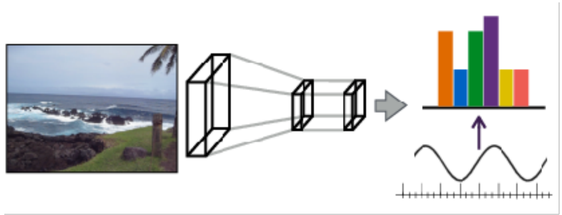
Note: connects ***across*** instances!

[Slide credit: Carl Doersch]

Audio

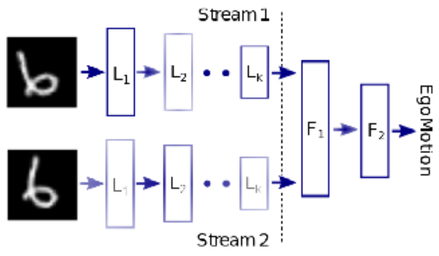


de Sa. NIPS 1994.



Owens et al. ECCV 2016.

Egomotion



Agrawal et al. ICCV 2015.

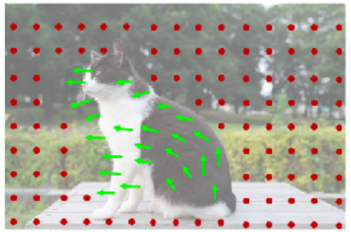
Jayaraman et al. ICCV 2015.

Context

Pathak et al. CVPR 2016.

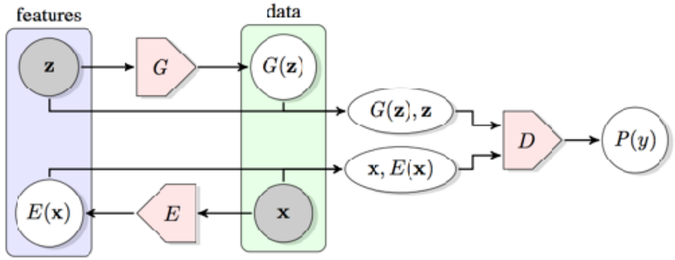
Noroozi and Favaro. ECCV 2016.
Doersch et al. ICCV 2015.

Video



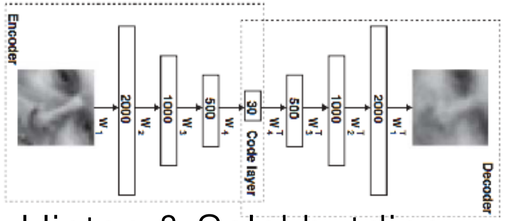
Wang et al. ICCV 2015. Pathak et al. CVPR 2017.
Misra et al. ECCV 2016.

Generative Modeling



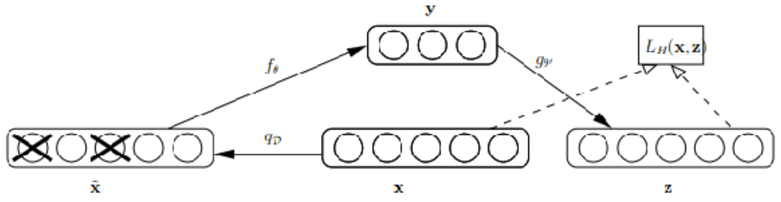
Donahue et al. Dumoulin et al. ICLR 2017.

Autoencoders



Hinton & Salakhutdinov.
Science 2006.

Denoising Autoencoders



Vincent et al. ICML 2008.

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

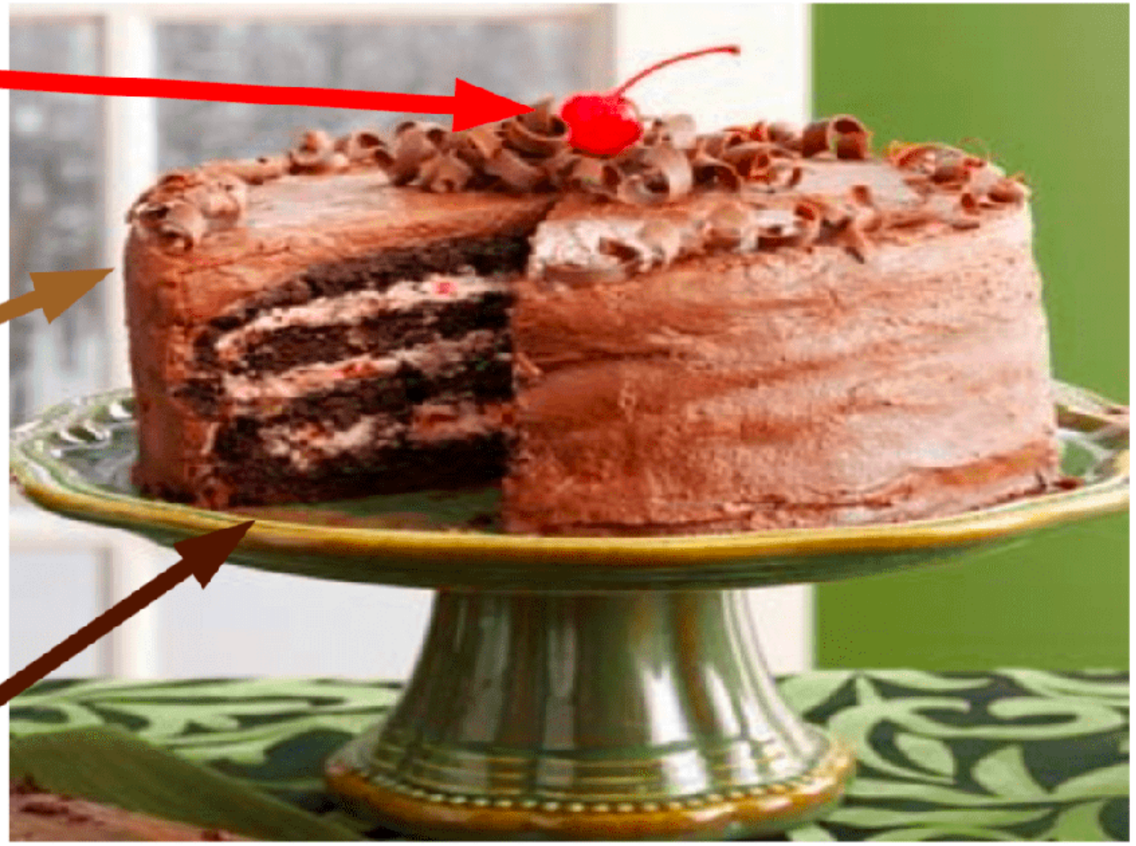
[Slide credit: Richard Zhang]

How Much Information is the Machine Given during Learning?

- **"Pure" Reinforcement Learning (cherry)**
  - The machine predicts a scalar reward given once in a while.
  - **A few bits for some samples**

- **Supervised Learning (icing)**
  - The machine predicts a category or a few numbers for each input
  - Predicting human-supplied data
  - **10→10,000 bits per sample**

- **Self-Supervised Learning (cake génoise)**
  - The machine predicts any part of its input for any observed part.
  - Predicts future frames in videos
  - **Millions of bits per sample**



[Slide Credit: Yann LeCun]

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 missing data