Self-supervised Representation Learning

Computer Vision
Fall 2018
Columbia University

Project

- Presentation schedule posted on Piazza.
- Review it ASAP and let us know of any problems by Wednesday
- For those presenting on December 5: OK to have some experiments in progress
- Final reports due December 10 midnight — no extensions!

GPU Credits

- If you have not requested GPU credits, do so immediately.
- We are starting to give them away…

Homeworks

- HW3 is back: median is 100% !
- HW4 grades soon
- HW5 due today

Final Grades

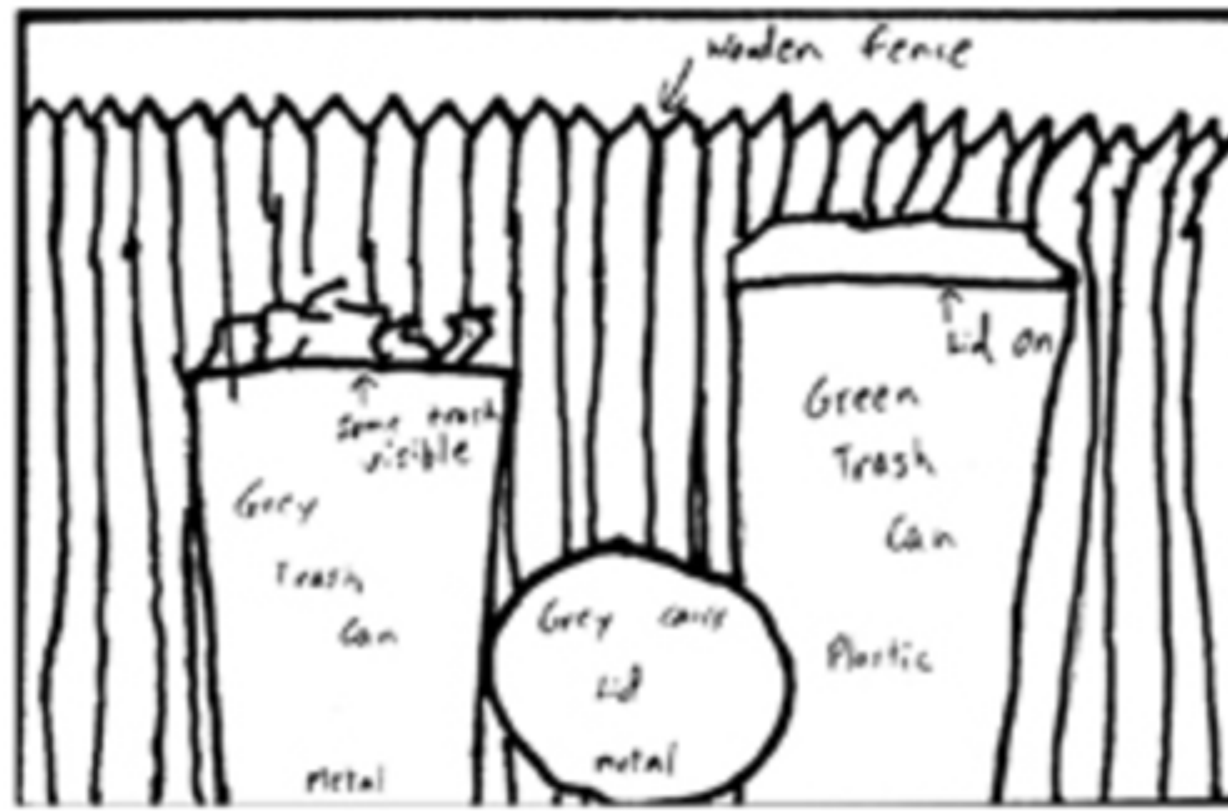
- We will likely curve down, but we will guarantee:
  - 90% is at least A
  - 80% is at least B
  - 70% is at least C

Next Semester

- E6998 Advanced Computer Vision, offered Spring 2019
- Focuses on research frontier of computer vision and applied machine learning
- Make sure to fill out survey to get off wait list

Observed image

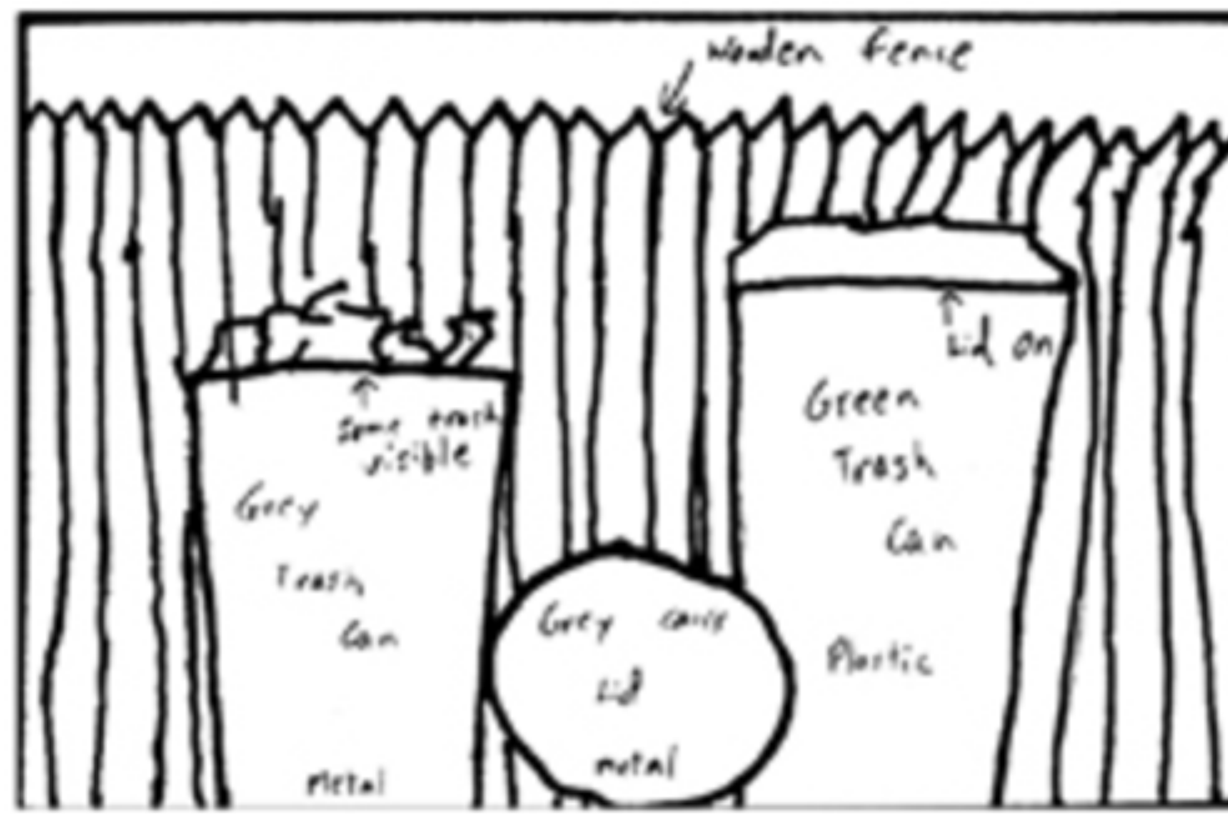
Drawn from memory



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

Observed image

Drawn from memory



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

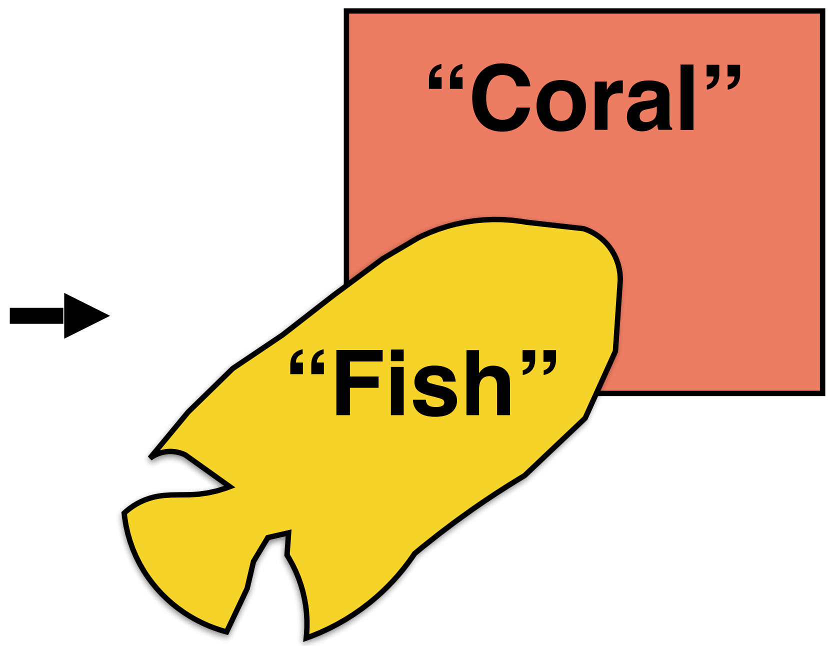
"I stand at the window and see a house, trees, sky. Theoretically I might say there were 327 brightnesses and nuances of colour. Do I have "327"? No. I have sky, house, and trees.”

— Max Wertheimer, 1923

Representation learning

$\mathbf{x}$

Image



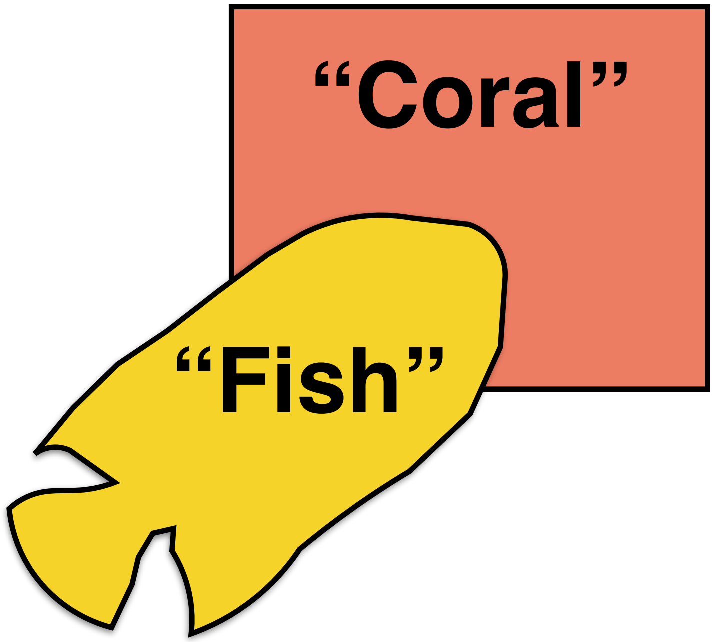
Compact mental representation

Slide credit: Phillip Isola

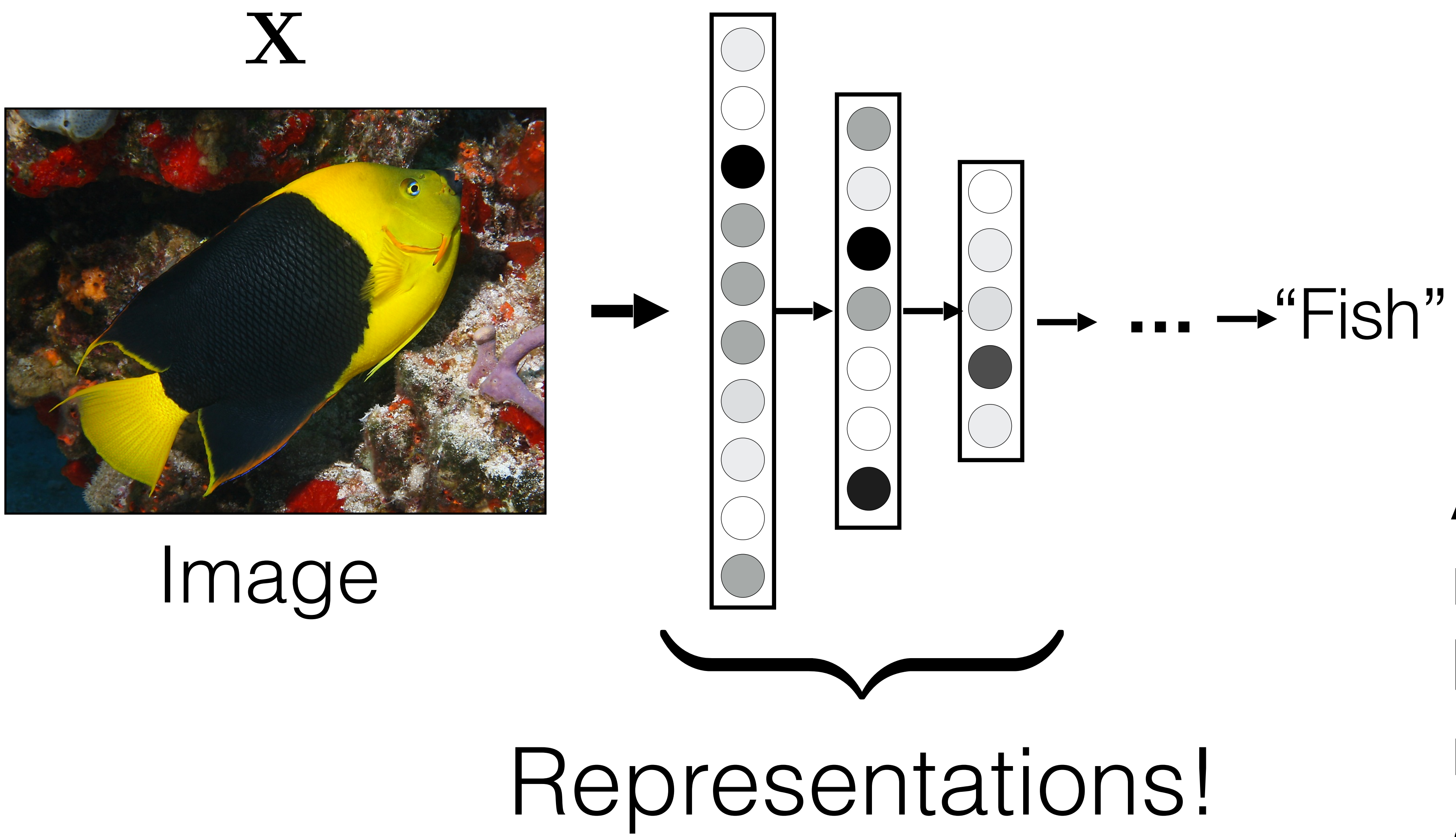
Representation learning

Good representations are:

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



Slide credit: Phillip Isola

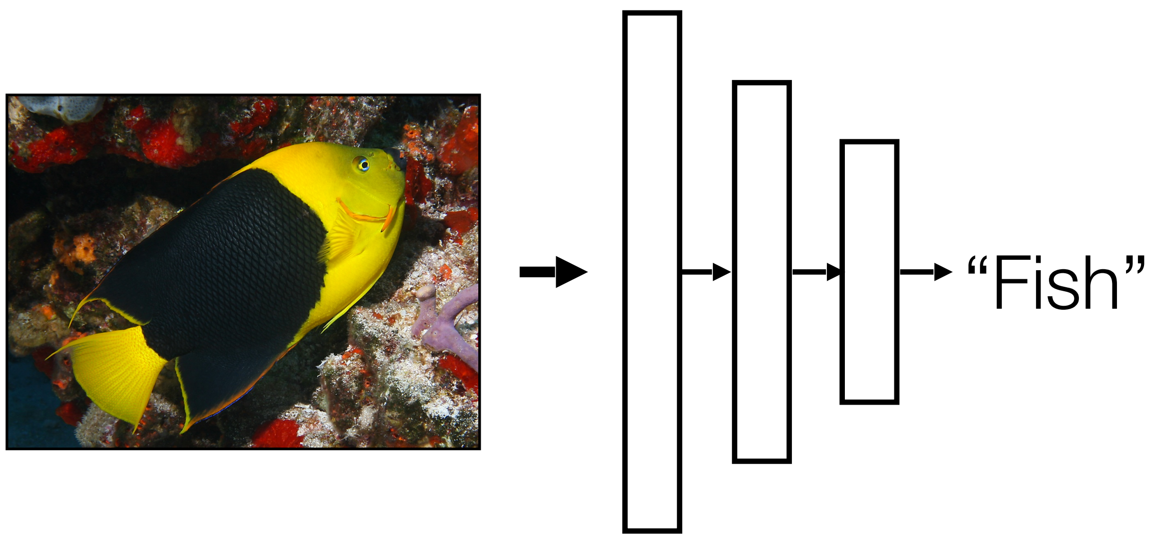


A CNN is a multiscale, hierarchical representation of data

Slide credit: Phillip Isola

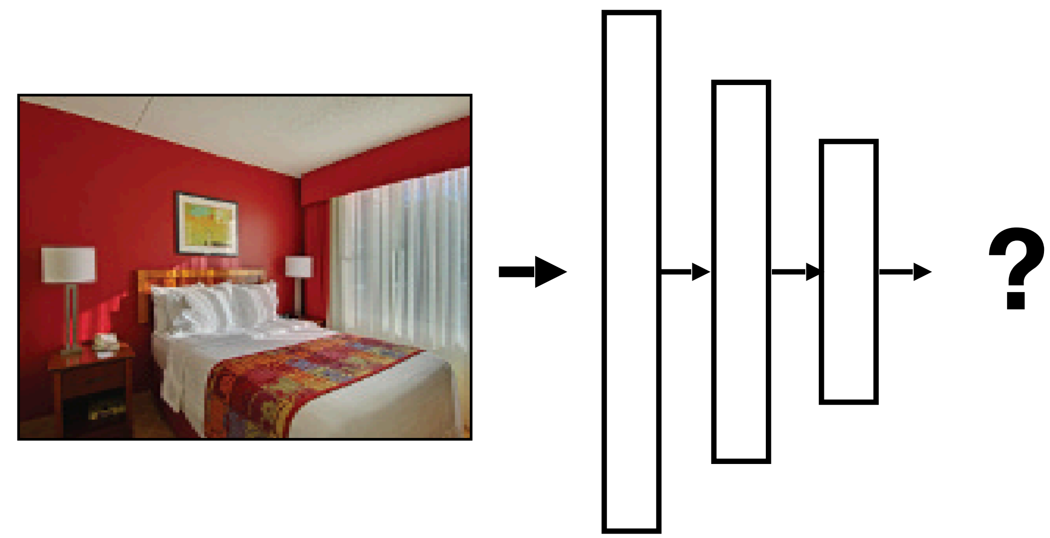
Training

Object recognition



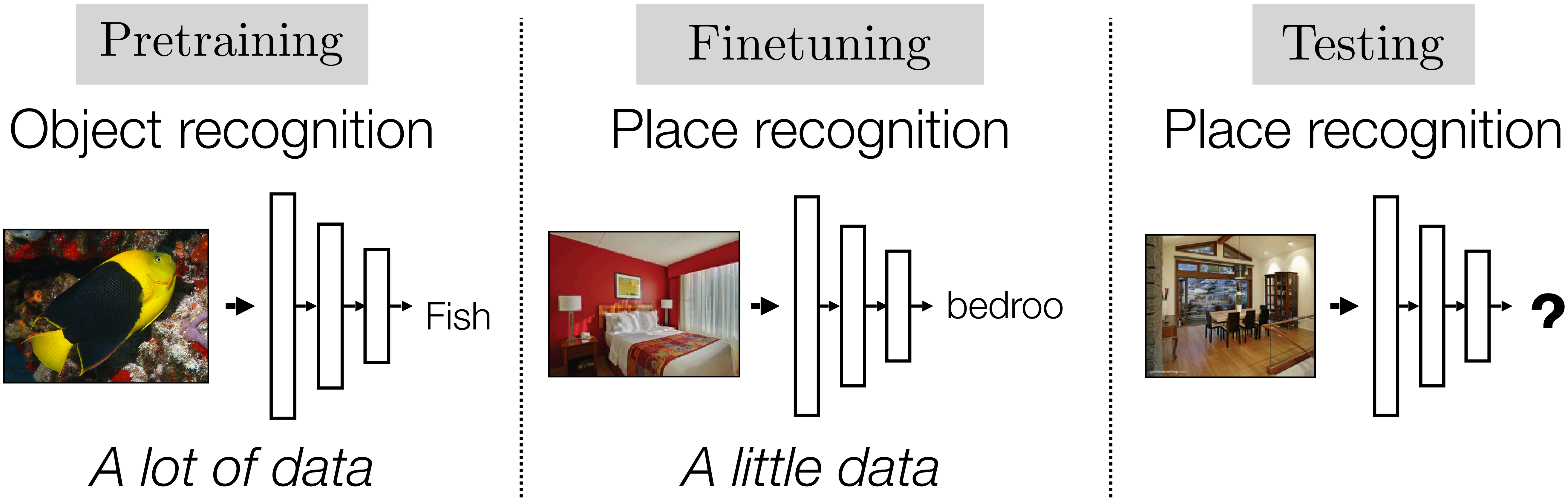
Testing

Place recognition



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

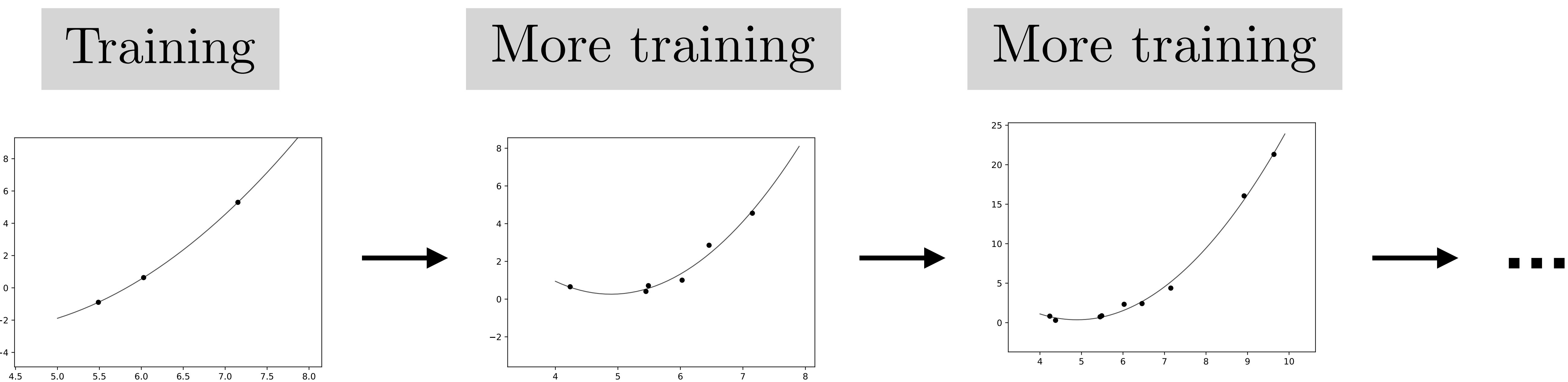
Slide credit: Phillip Isola



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

Slide credit: Phillip Isola

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.



Slide credit: Phillip Isola

Supervised object recognition

→ **Learner** →

"Fish"

image X

label Y

Slide credit: Phillip Isola

Supervised object recognition

→ **Learner** →

"Fish"

image X

label Y

Slide credit: Phillip Isola

Supervised object recognition

→ **Learner** →

"Fish"

image X

label Y

Slide credit: Phillip Isola

Supervised object recognition

image X → **Learner** → "Fish"

⋮

image X

label Y

Slide credit: Phillip Isola

Kitten Carousel

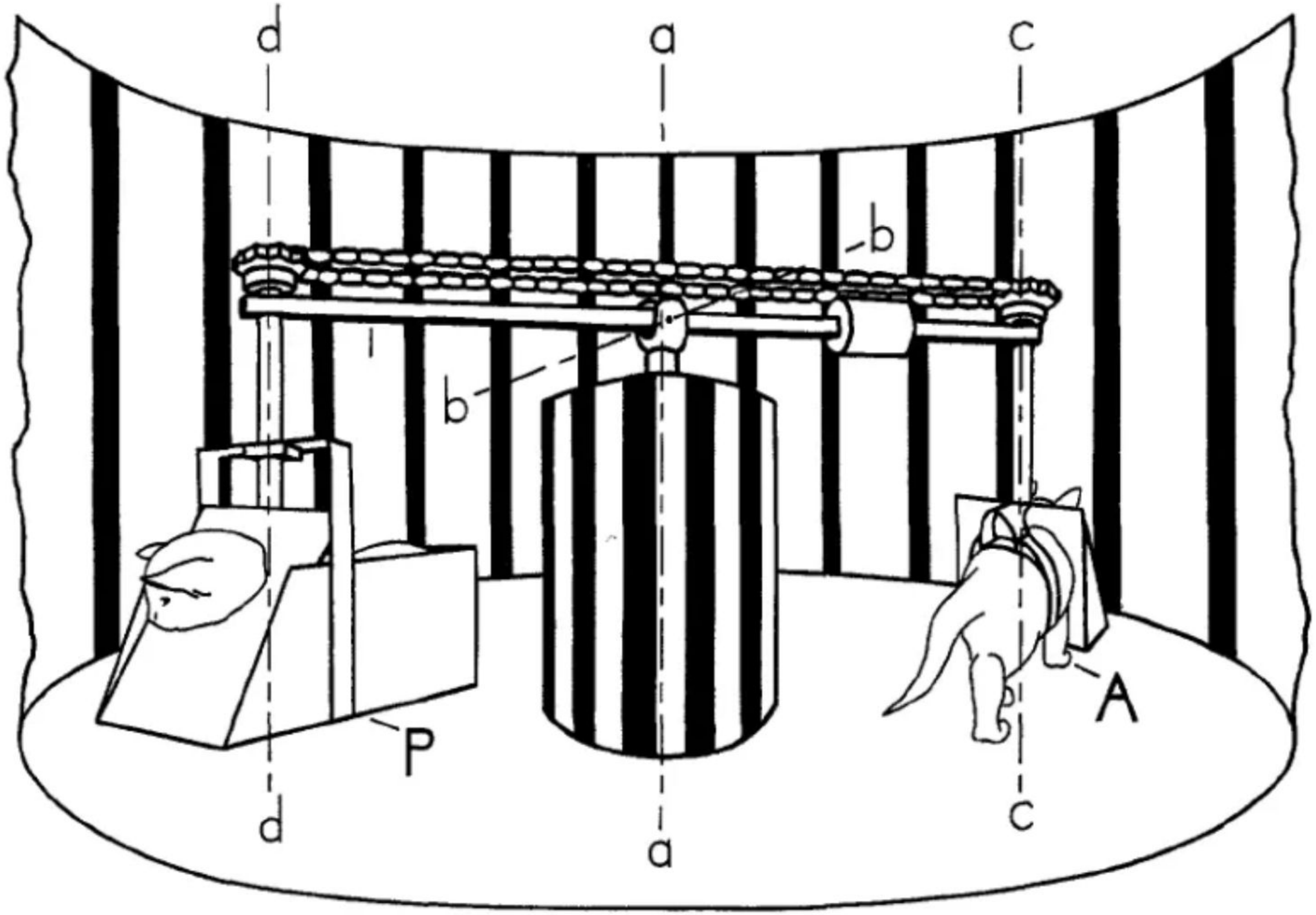
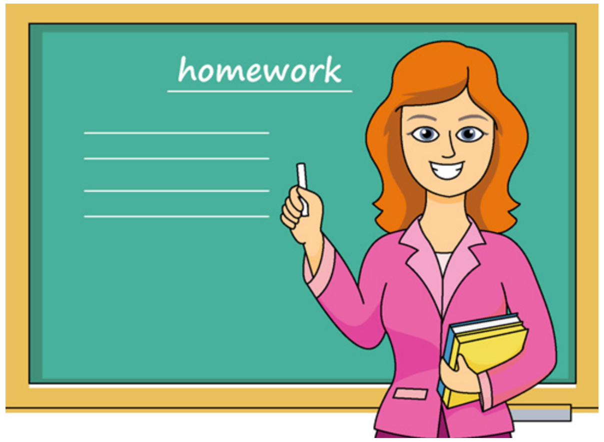


FIG. 1. Apparatus for equating motion and consequent visual feedback for an actively moving (A) and a passively moved (P) *S*.

Held and Hein, 1963

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

Slide credit: Phillip Isola

Learning from examples

(aka **supervised learning**)

Training data

$\{x_1, y_1\}$

$\{x_2, y_2\}$ $\rightarrow$ Learner $\rightarrow$ $f : X \rightarrow Y$

$\{x_3, y_3\}$

$\dots$

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

Slide credit: Phillip Isola

Learning without examples

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

Data

$\{x_1\}$

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

$\{x_3\}$

$\ldots$

Slide credit: Phillip Isola

Unsupervised Representation Learning

Data

$\{x_1\}$

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

$\{x_3\}$

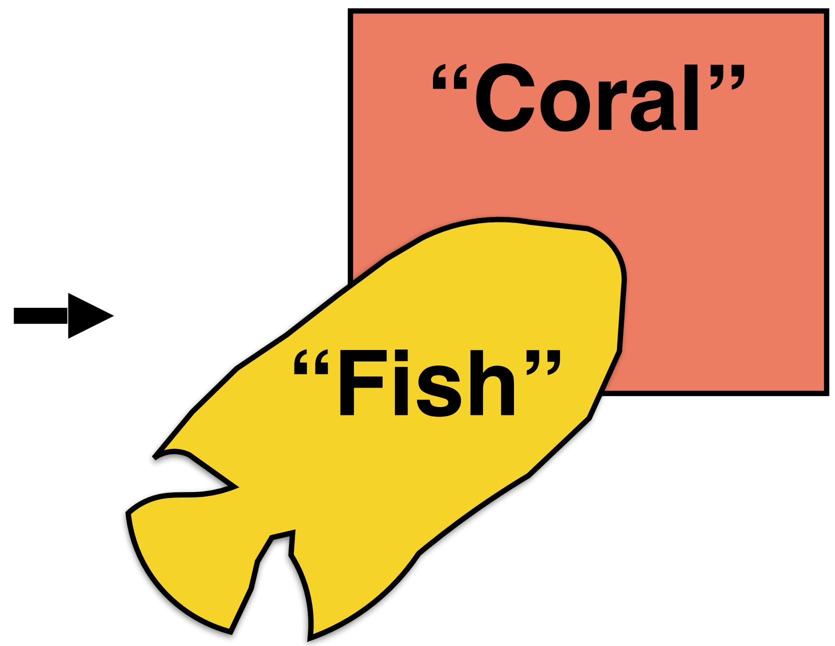
$\dots$

Slide credit: Phillip Isola

Unsupervised Representation Learning

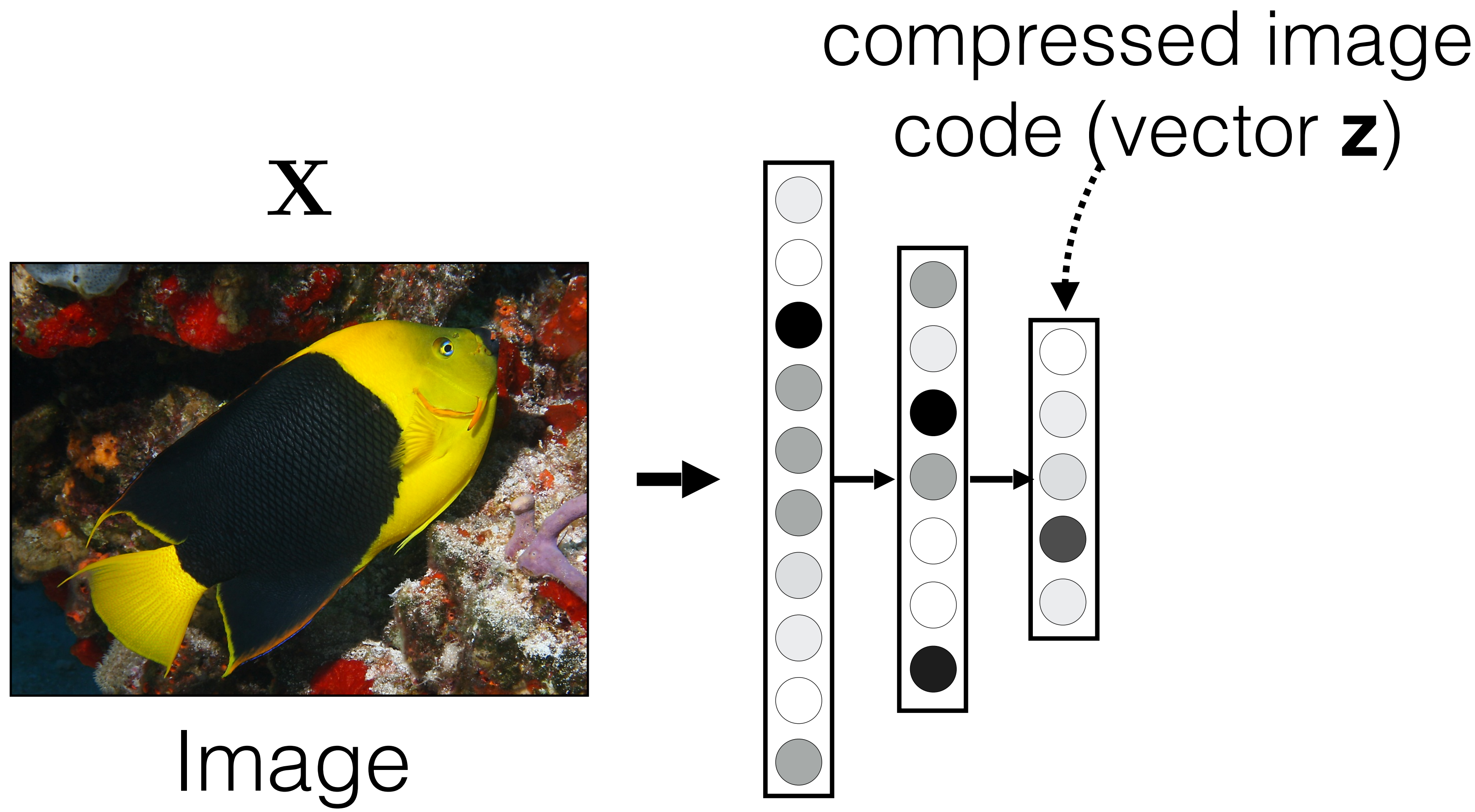
$\mathbf{x}$

Image



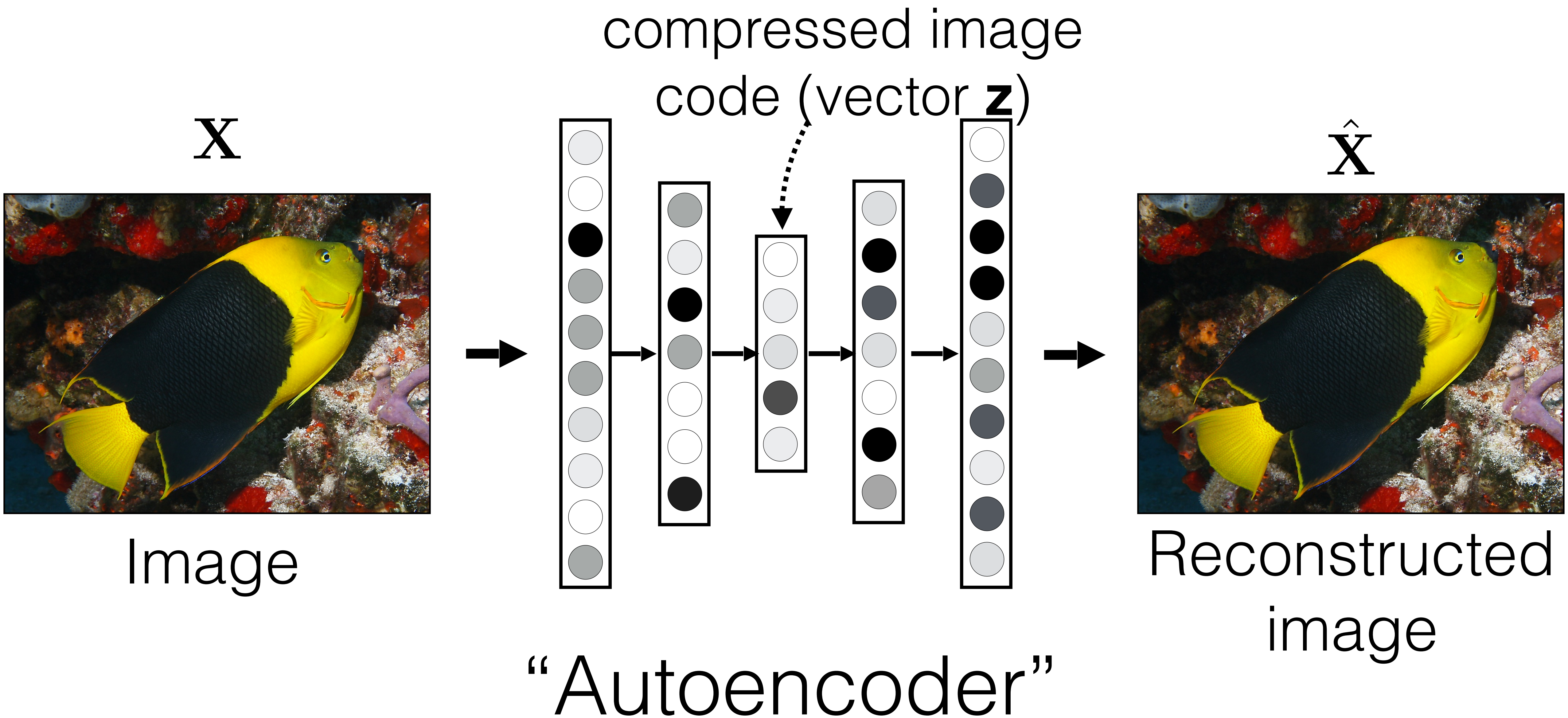
Compact mental representation

Unsupervised Representation Learning



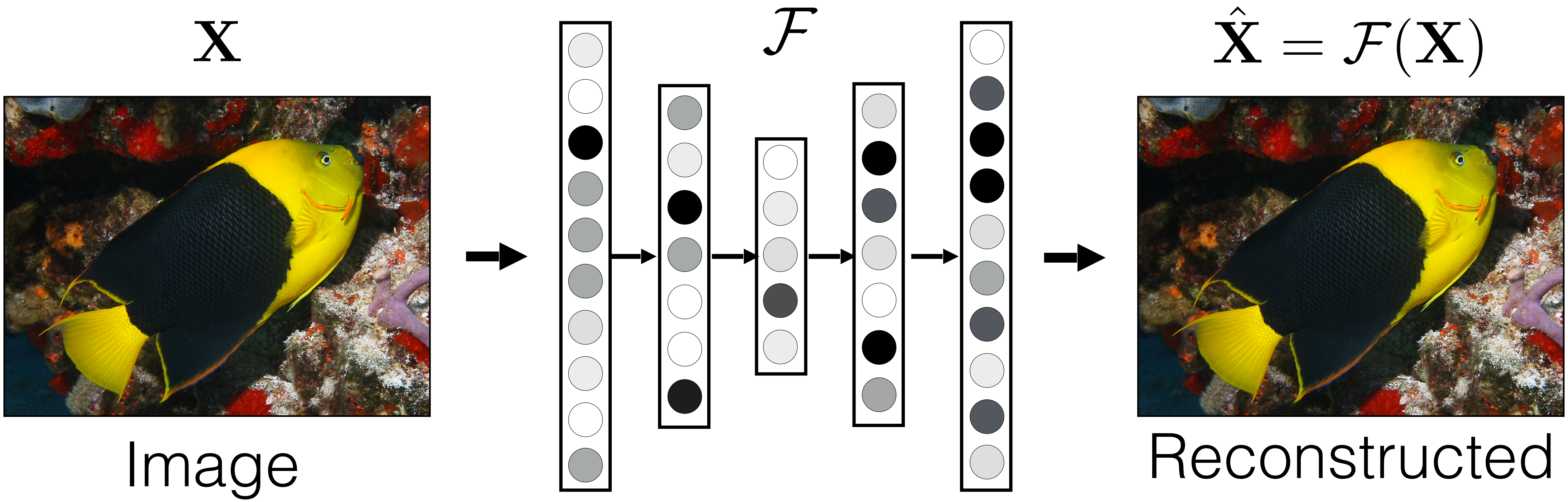
Slide credit: Phillip Isola

Unsupervised Representation Learning



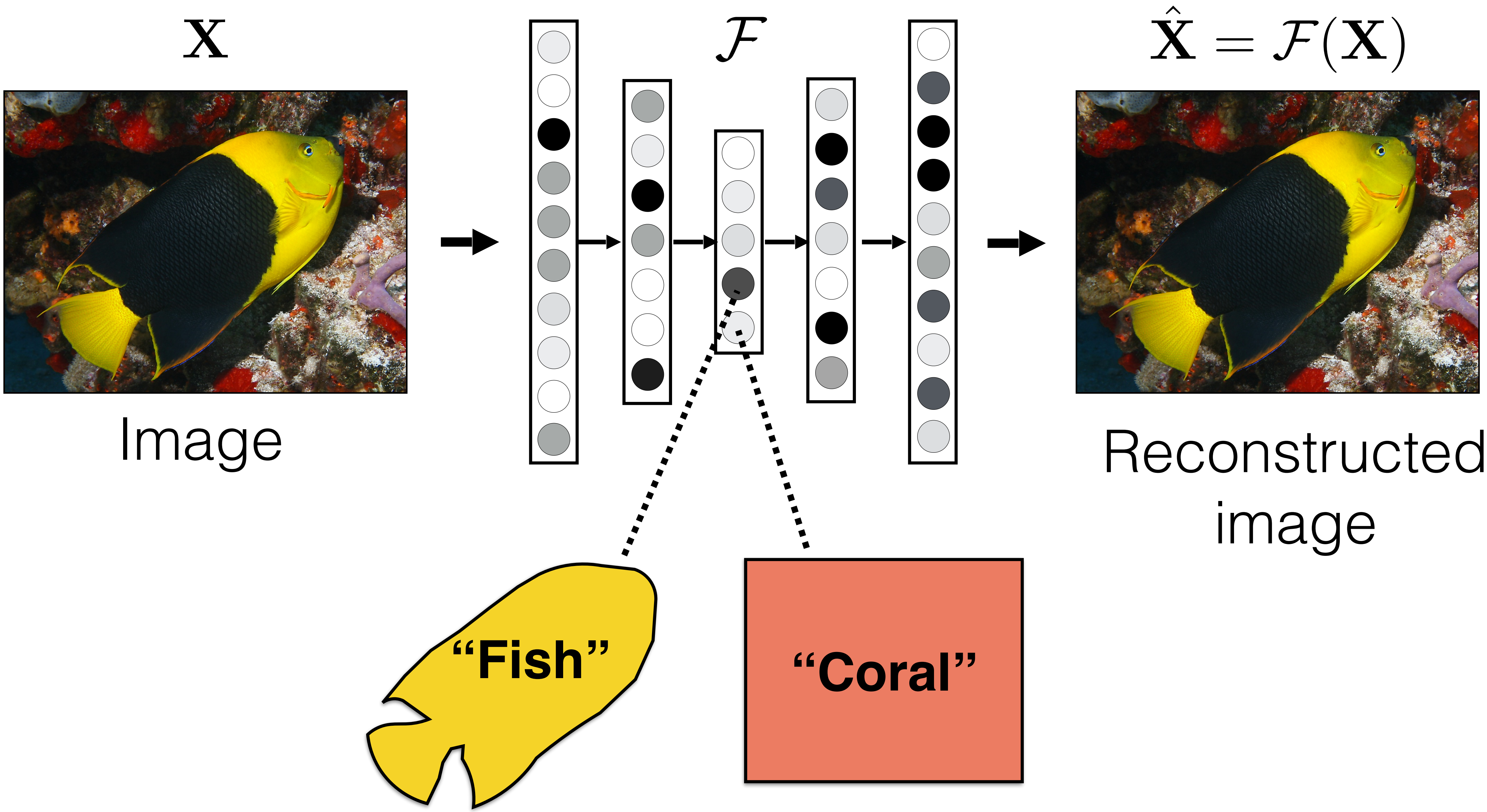
[e.g., Hinton & Salakhutdinov, Science 2006]

Autoencoder



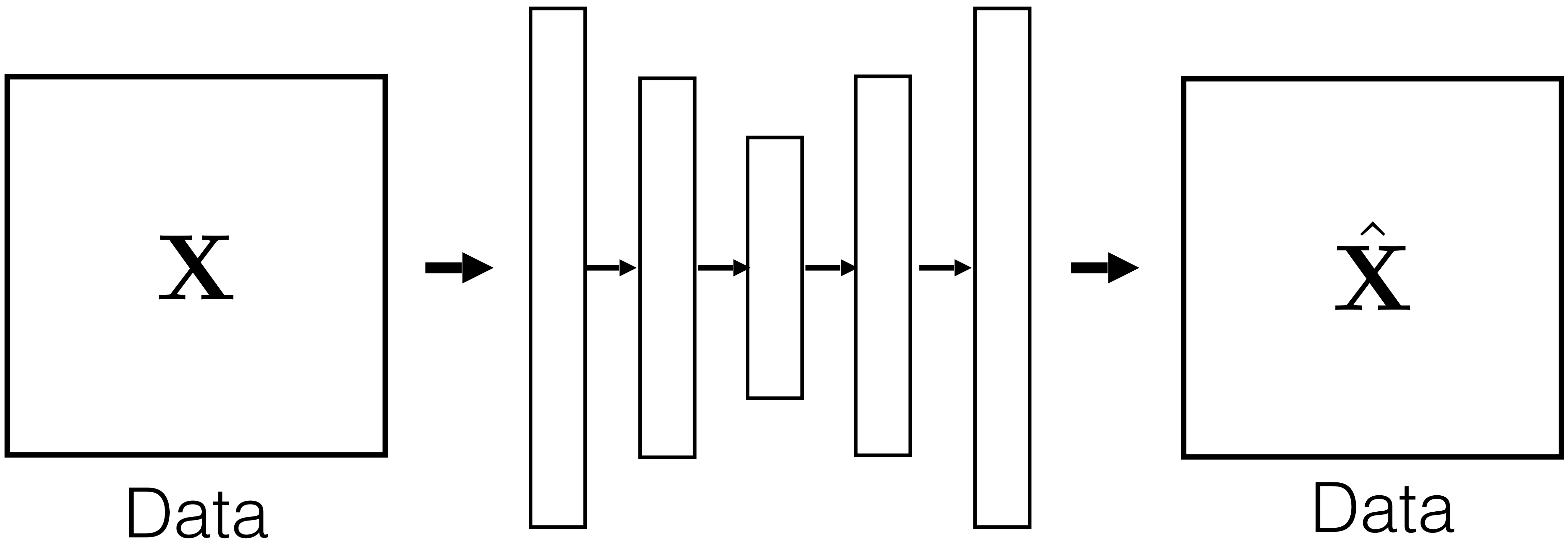
$$\arg\min_{\mathcal{F}} \mathbb{E}_{\mathbf{X}}[||\mathcal{F}(\mathbf{X}) - \mathbf{X}||]$$

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



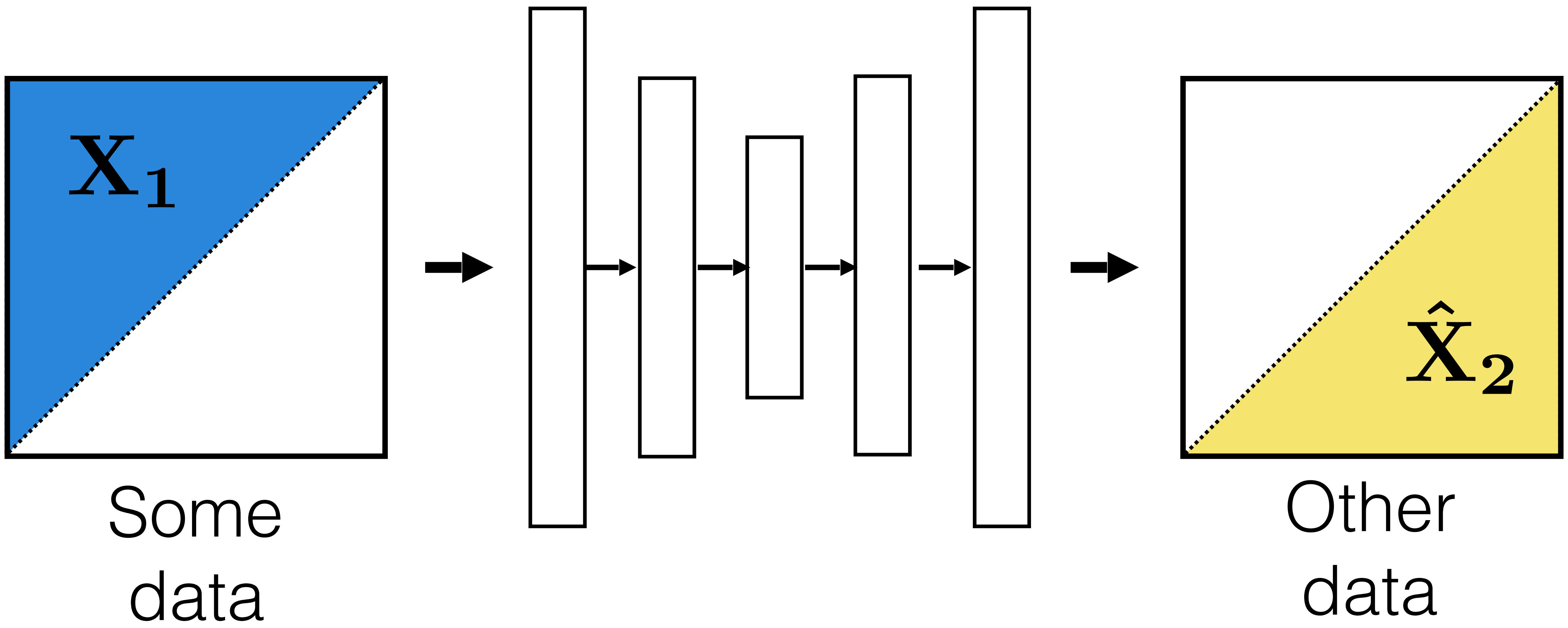
[e.g., Hinton & Salakhutdinov, Science 2006]

Data compression

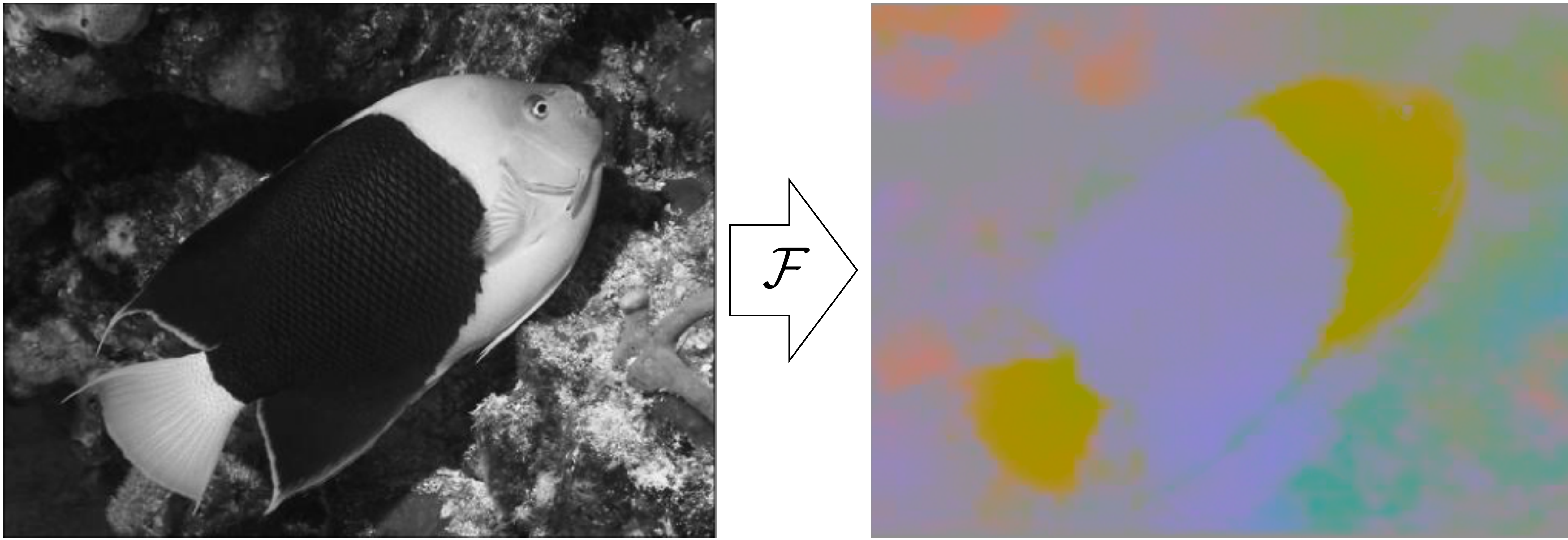


[e.g., Hinton & Salakhutdinov, Science 2006]

Data prediction

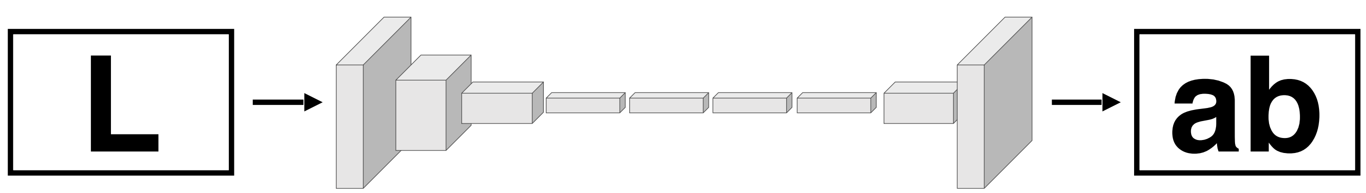


Slide credit: Phillip Isola

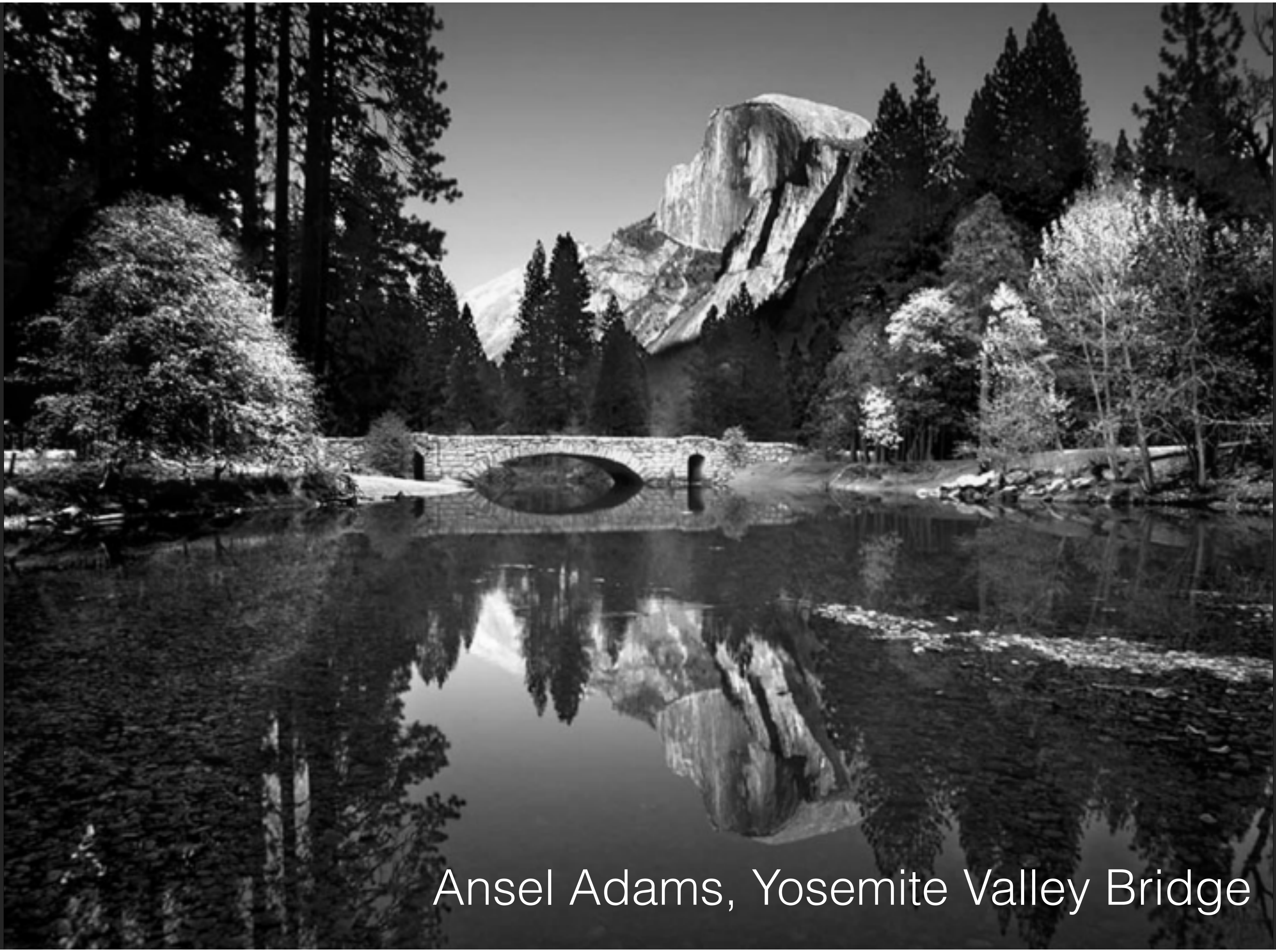


Grayscale image: L channel
$\mathbf{X} \in \mathbb{R}^{H \times W \times 1}$

Color information: ab channels
$\widehat{\mathbf{Y}} \in \mathbb{R}^{H \times W \times 2}$



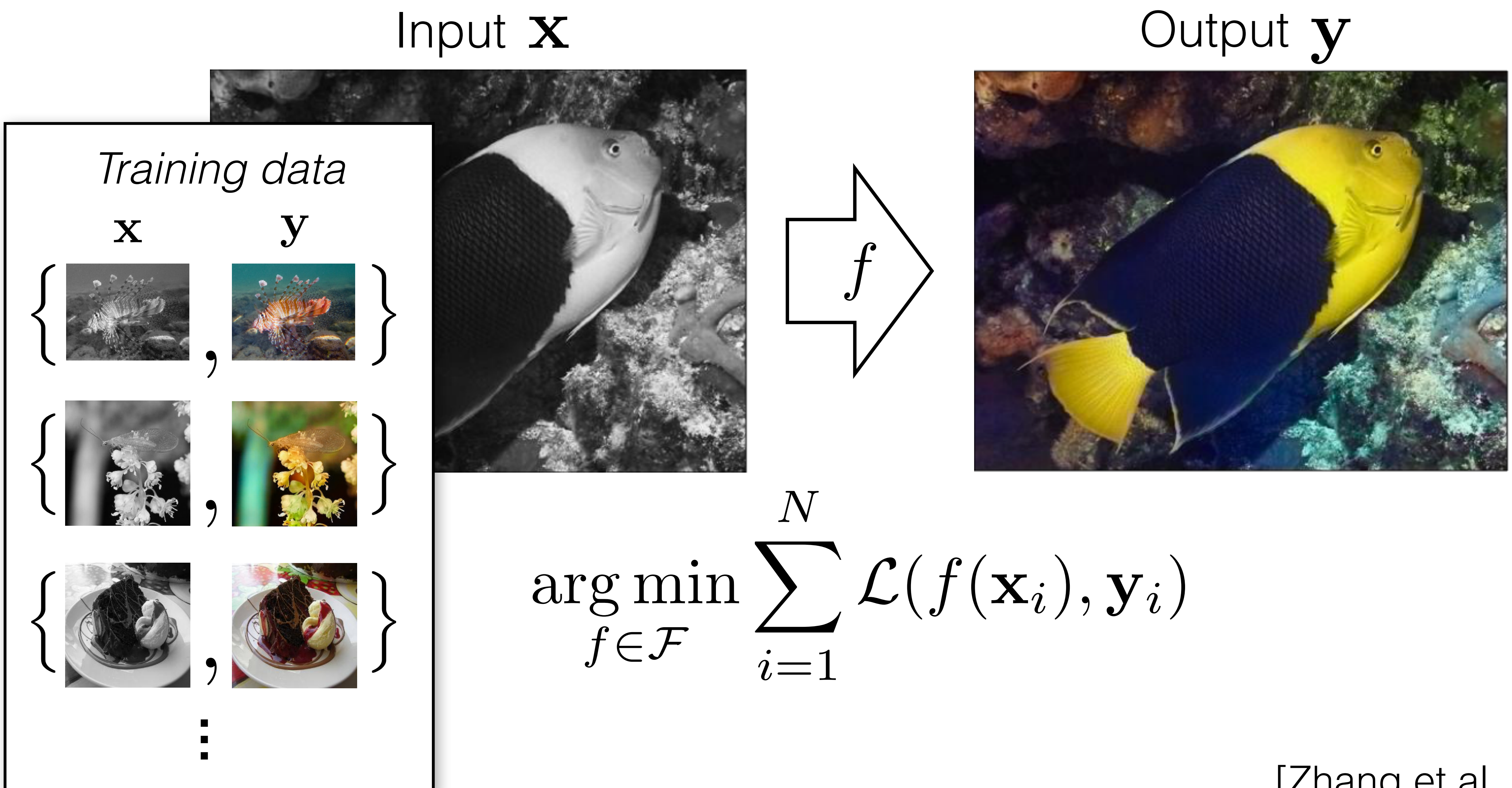
[Zhang, Isola, Efros, ECCV 2016]



Ansel Adams, Yosemite Valley Bridge


Result of [Zhang et al., ECCV 2016]

Image colorization



$$\arg\min_{f \in \mathcal{F}} \sum_{i=1}^{N} \mathcal{L}(f(\mathbf{x}_i), \mathbf{y}_i)$$

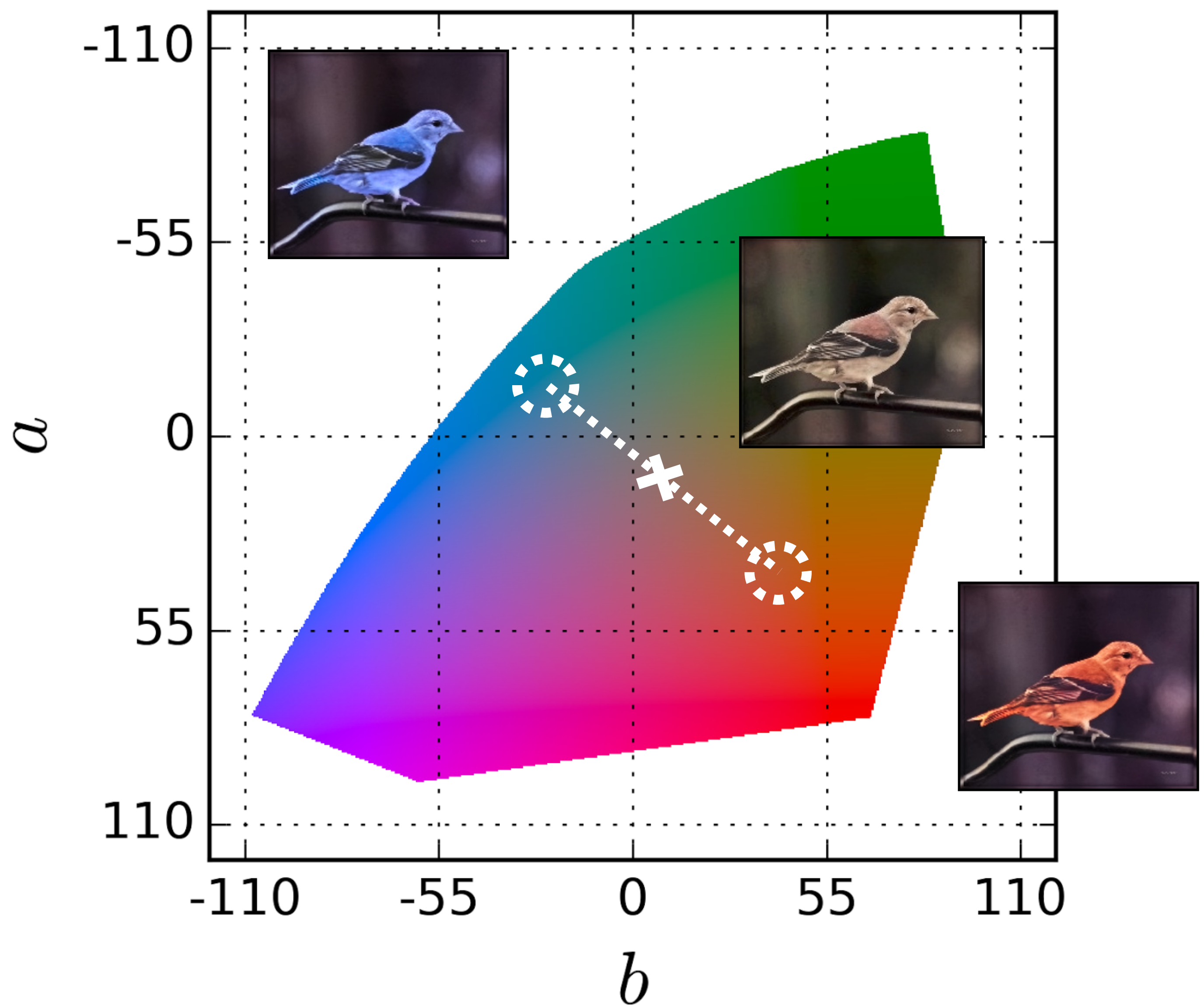
[Zhang et al., ECCV 2016]

Choosing loss and representation

Input Output Ground truth

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

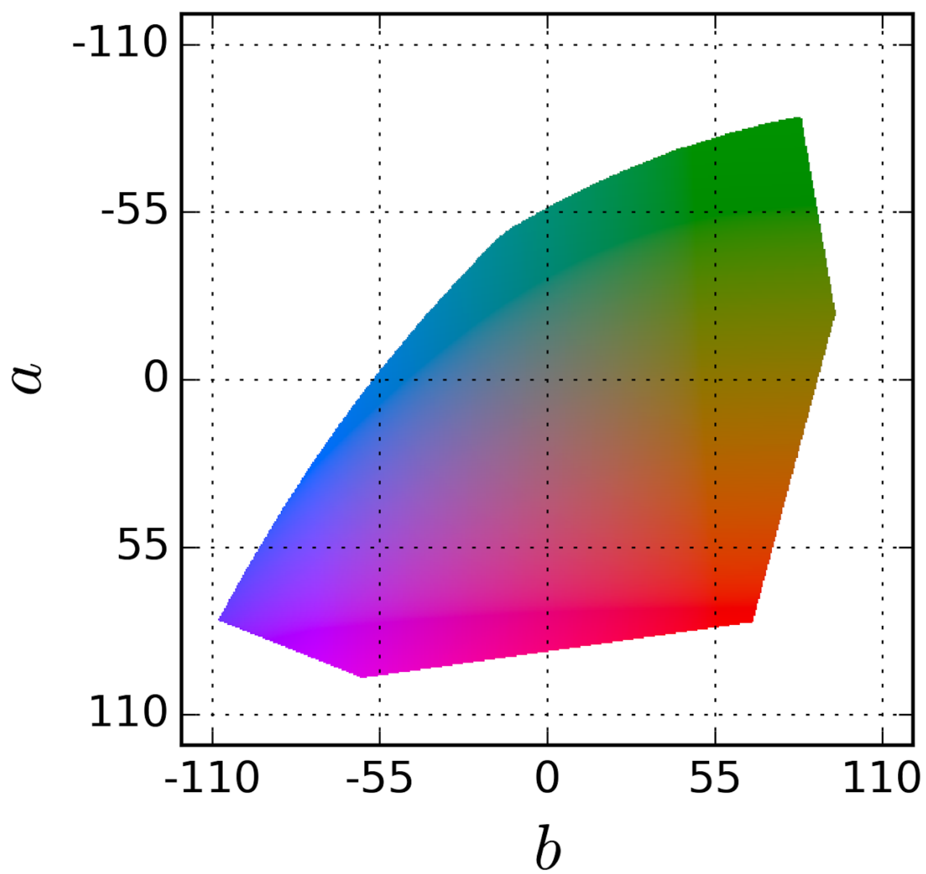
Slide credit: Phillip Isola



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

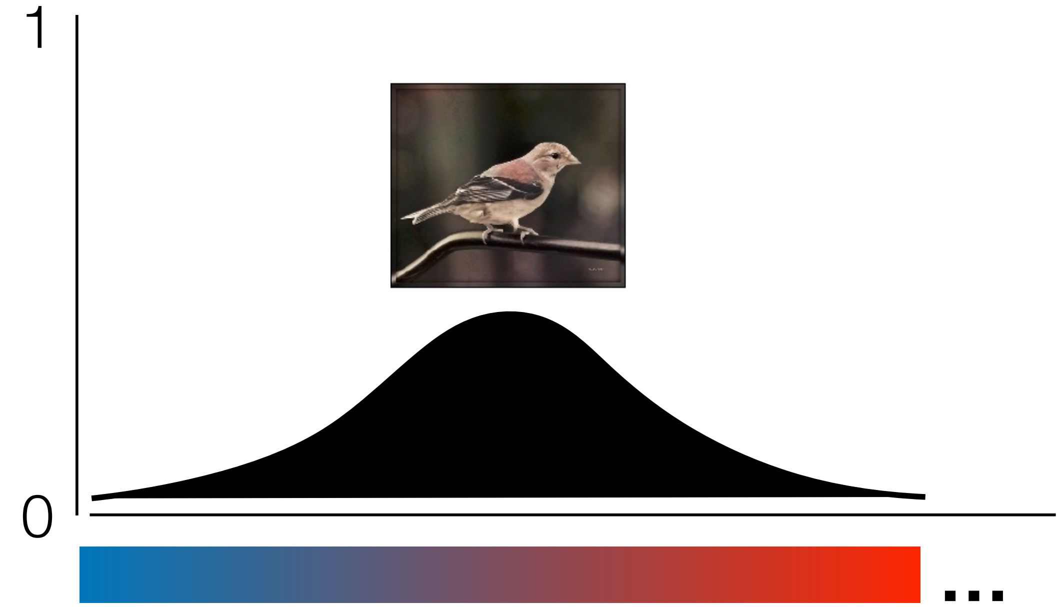
Slide credit: Phillip Isola

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



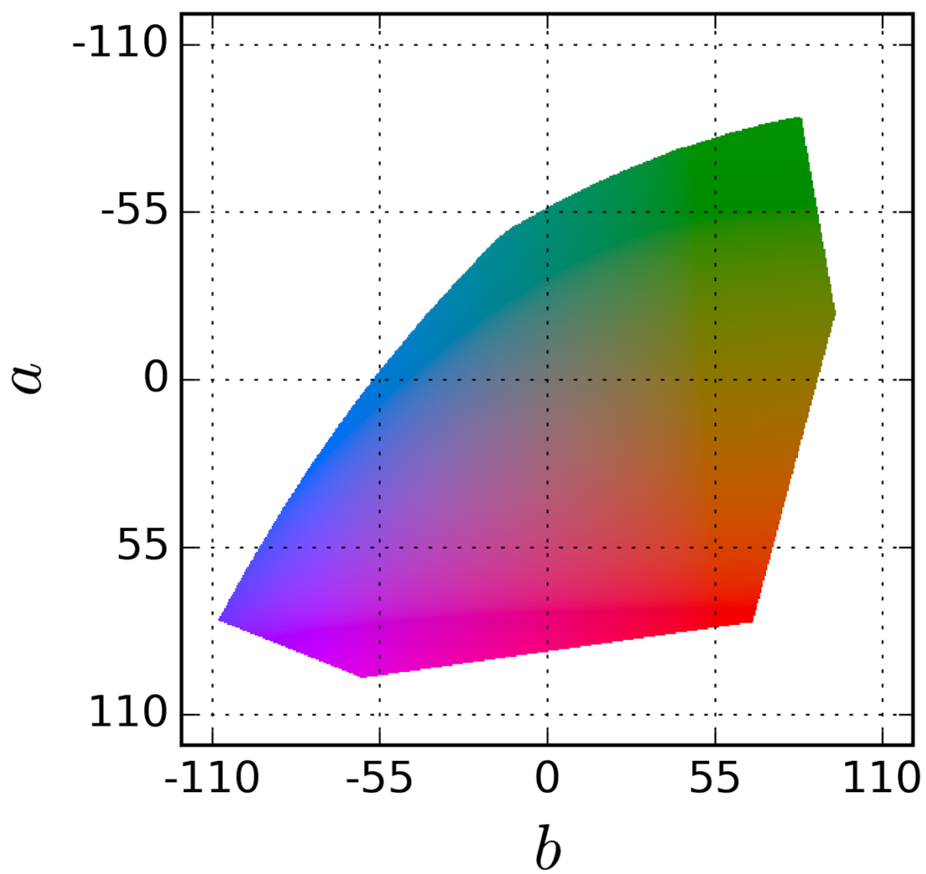
$$\mathcal{L}(f(\mathbf{x}), \mathbf{y}) = \|f(\mathbf{x}) - \mathbf{y}\|_2^2$$

Prediction for a single pixel i,j



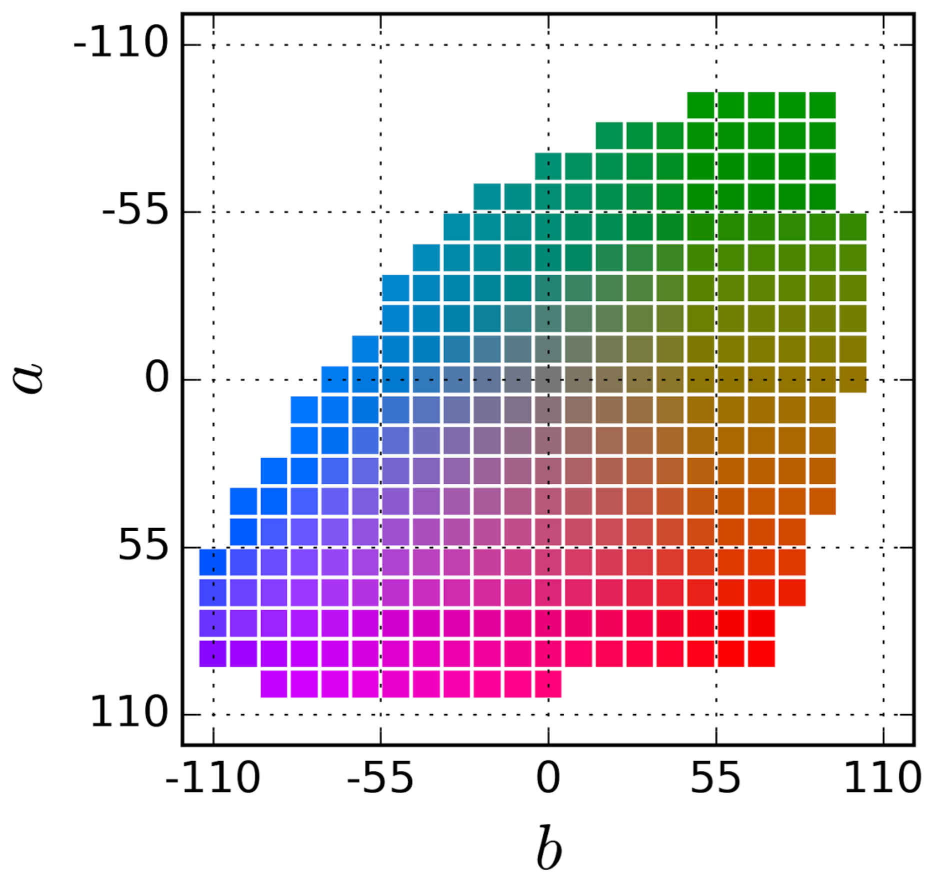
Slide credit: Phillip Isola

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



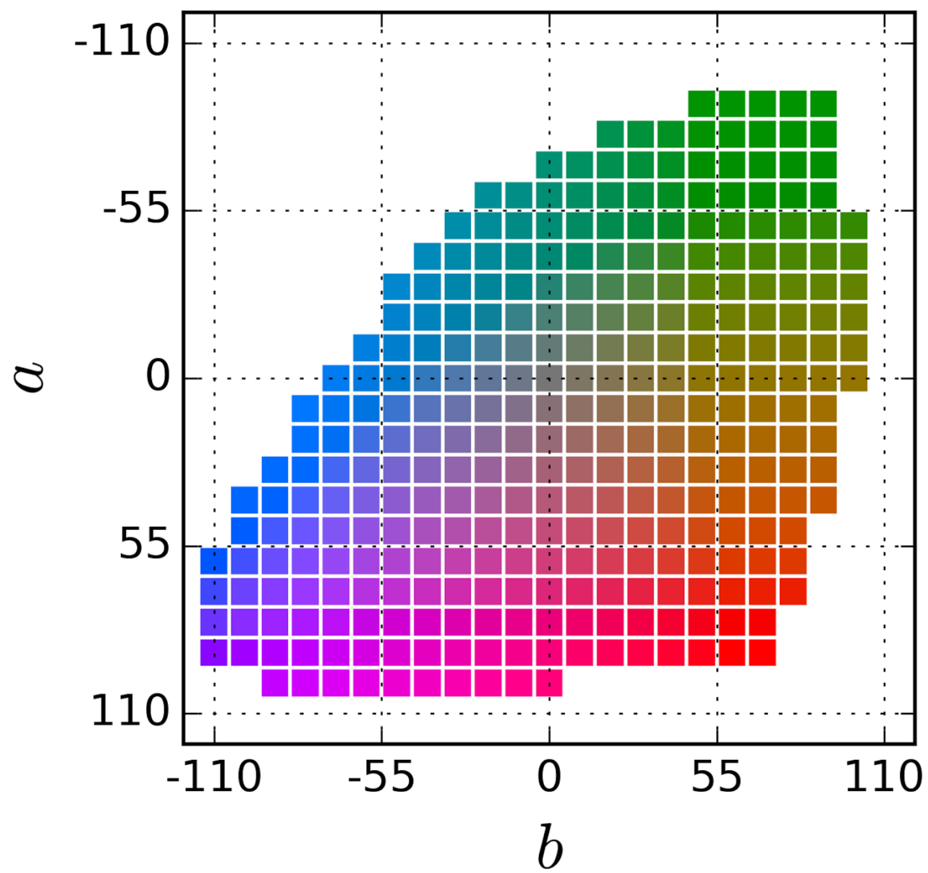
one-hot representation of K discrete classes

$$\mathbf{y} \in \mathbb{R}^{H \times W \times K}$$

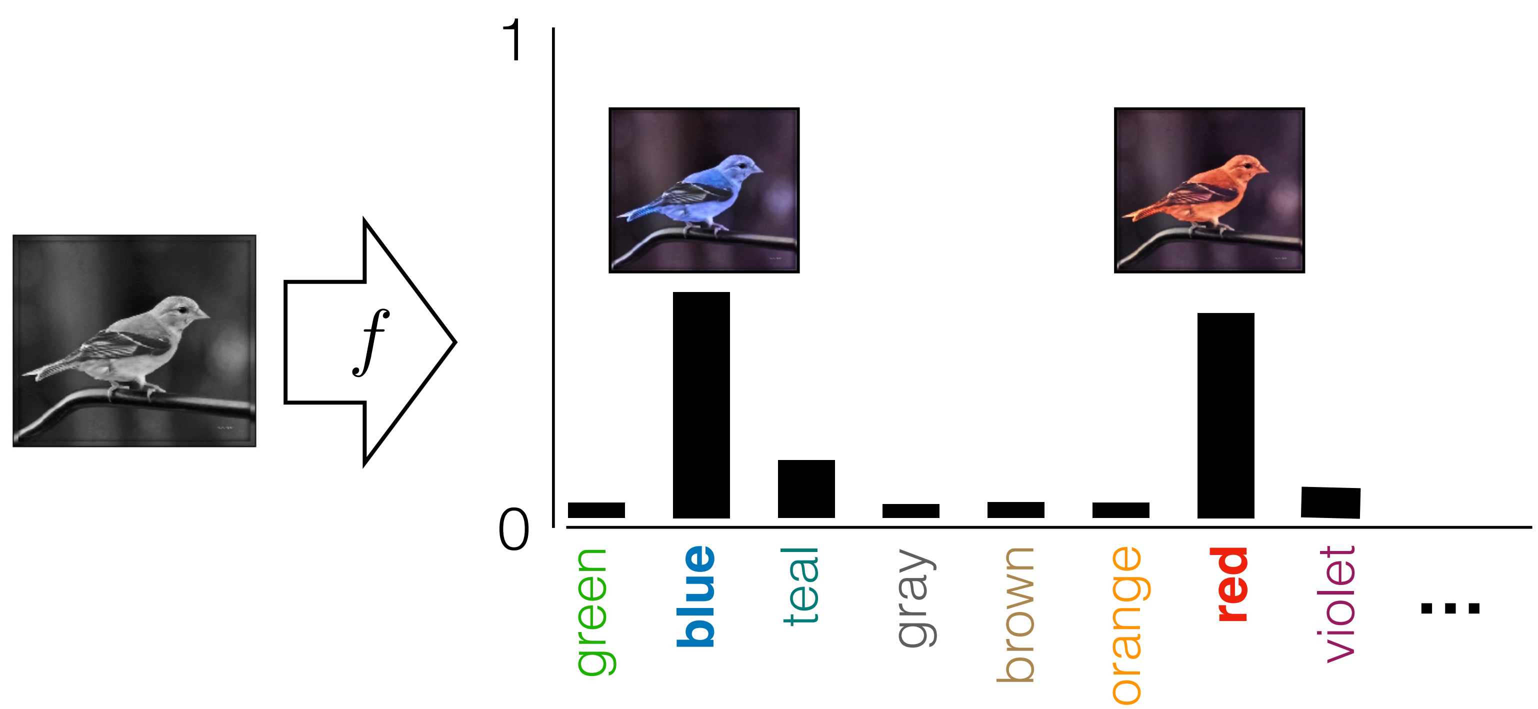


Slide credit: Phillip Isola

$$\mathbf{y} \in \mathbb{R}^{H \times W \times K}$$



Prediction for a single pixel i,j



$$\mathcal{L}(\mathbf{y}, f_\theta(\mathbf{x})) = H(\mathbf{y}, \mathtt{softmax}(f_\theta(\mathbf{x})))$$

Slide credit: Phillip Isola

Input

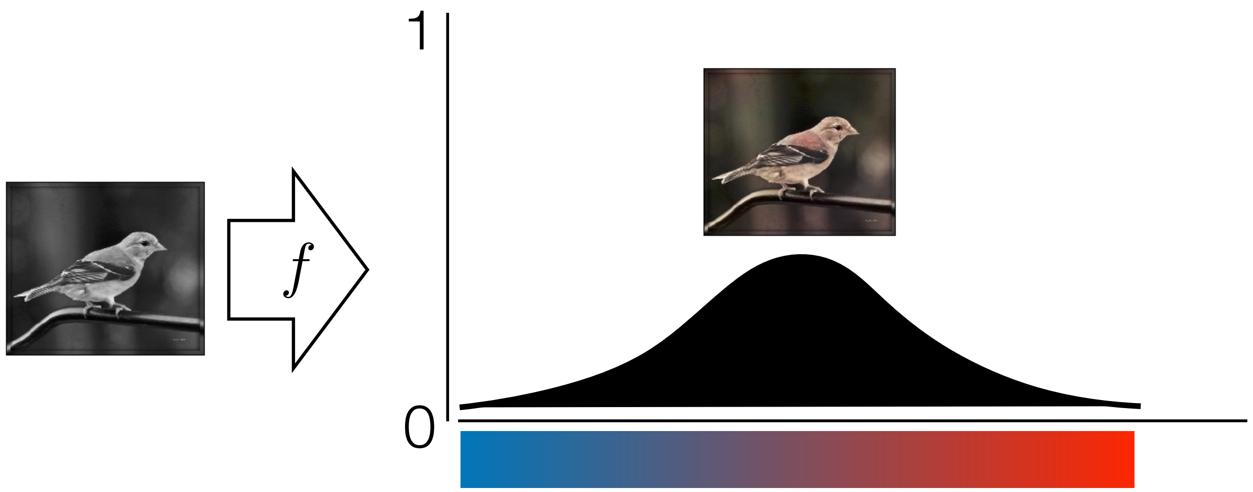
Zhang et al. 2016

Ground truth

$$\mathcal{L}(\mathbf{x}, \mathbf{y}) = H(\mathbf{y}, \texttt{softmax}(f_\theta(\mathbf{x})))$$

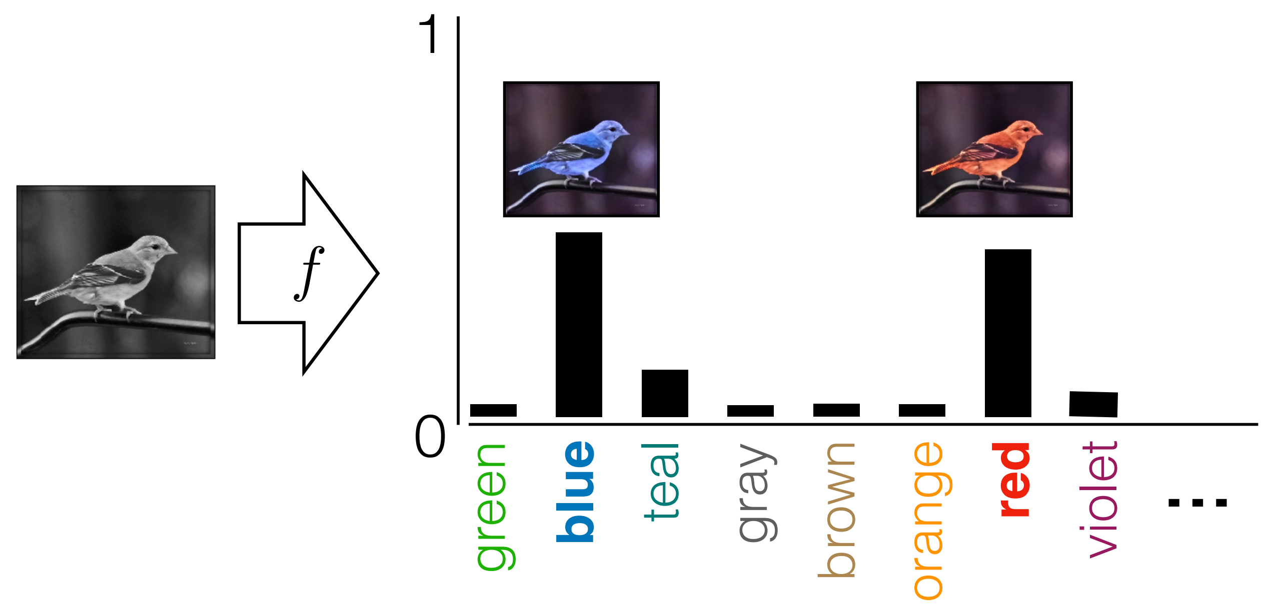
Slide credit: Phillip Isola

“Regression”



- Continuous-valued prediction
- (Usually) models unimodal distribution

“Classification”

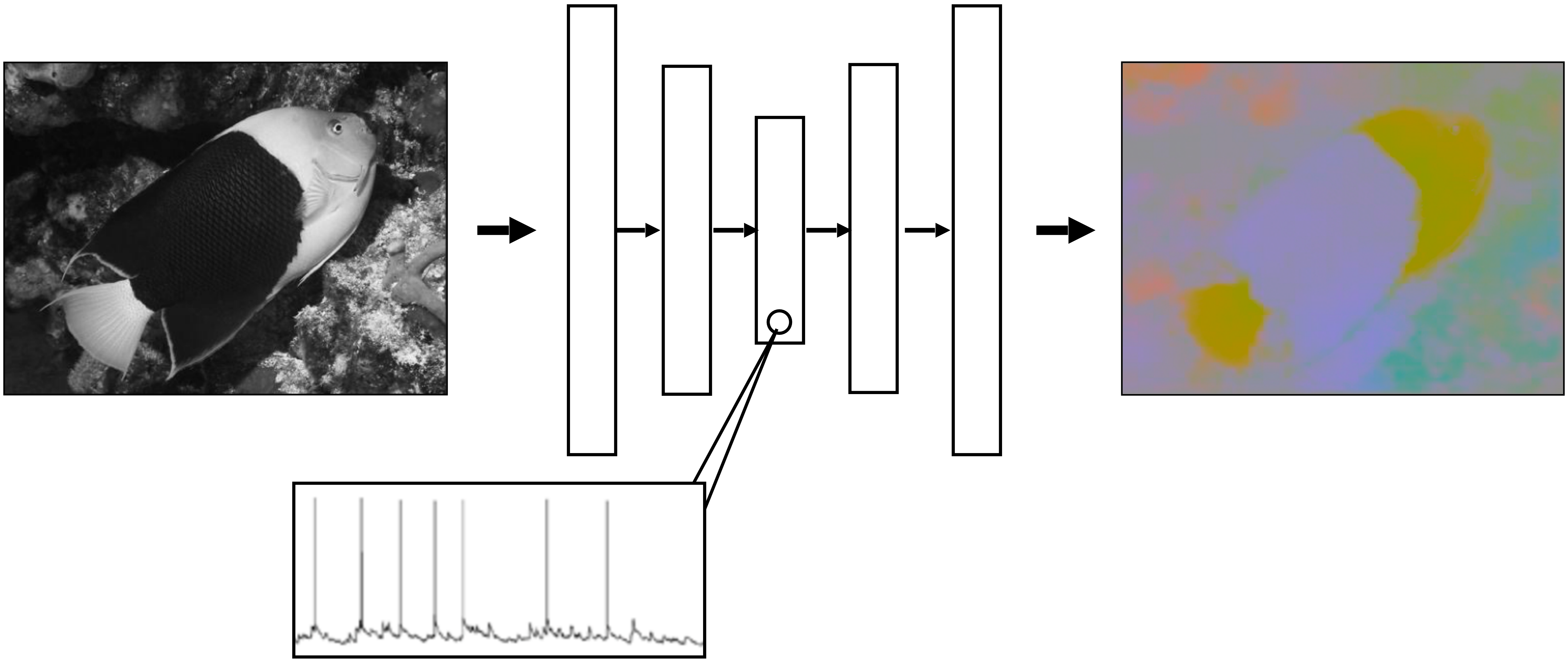


- Discrete-valued prediction
- Models multimodal distribution

Instructive failure

Instructive failure

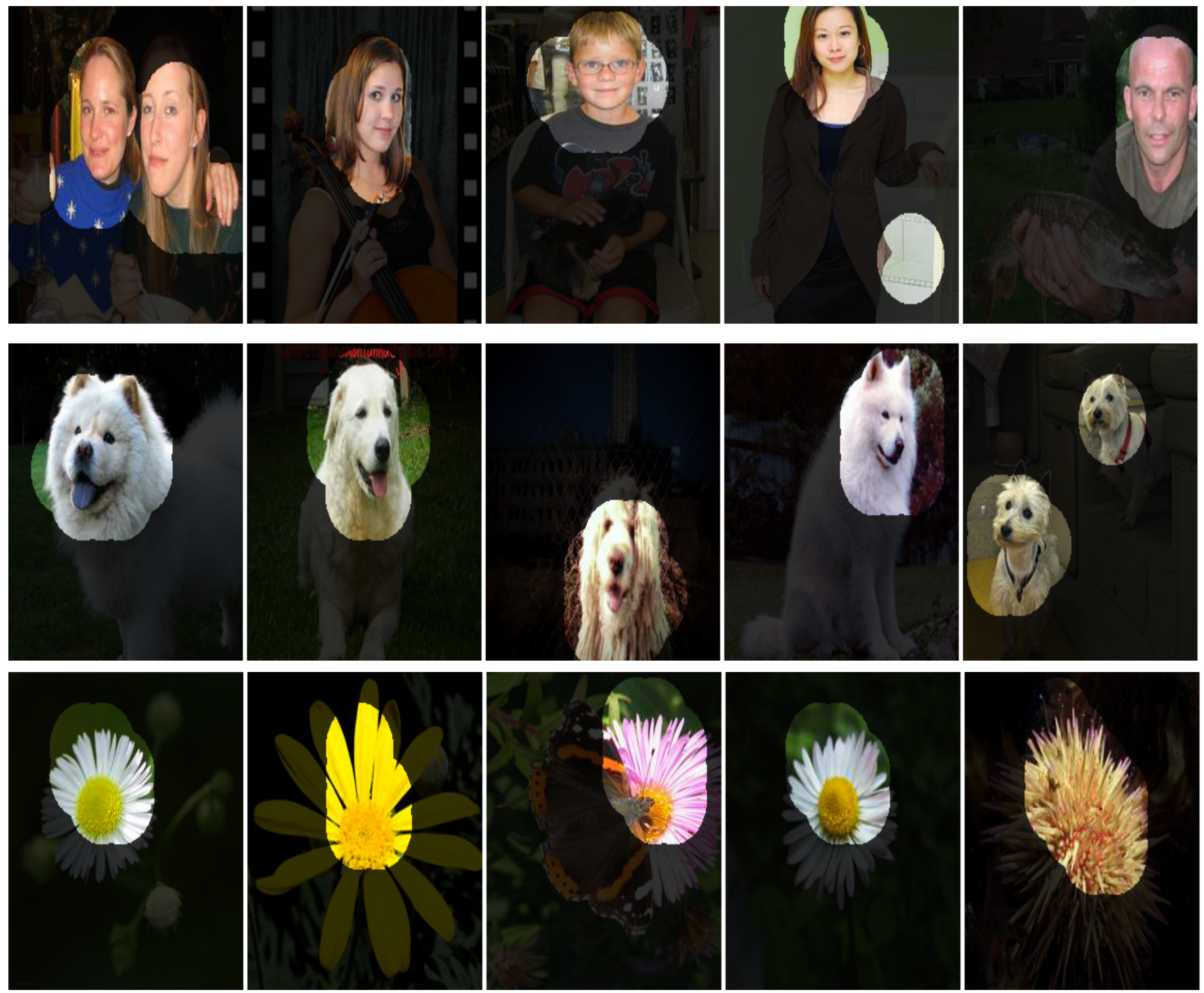
Deep Net "Electrophysiology"

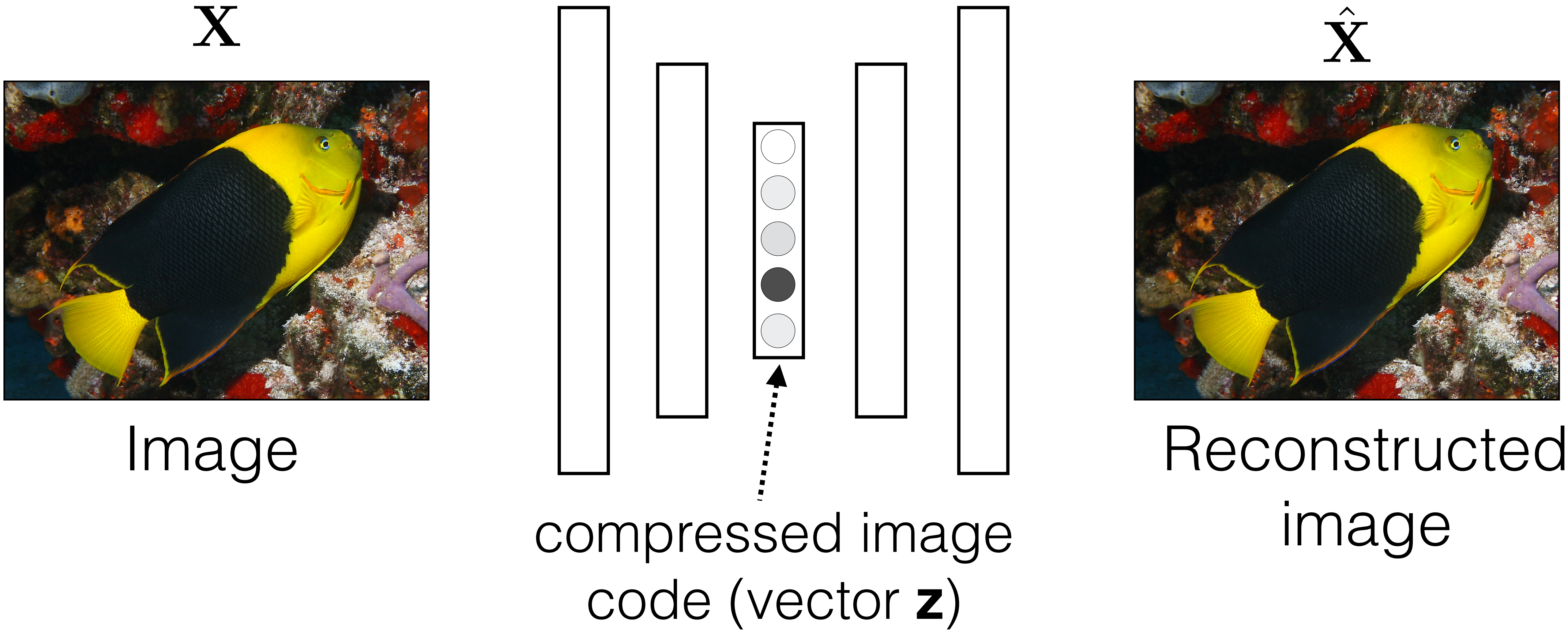


[Zeiler & Fergus, ECCV 2014]

[Zhou et al., ICLR 2015]

Stimuli that drive selected neurons (conv5 layer)

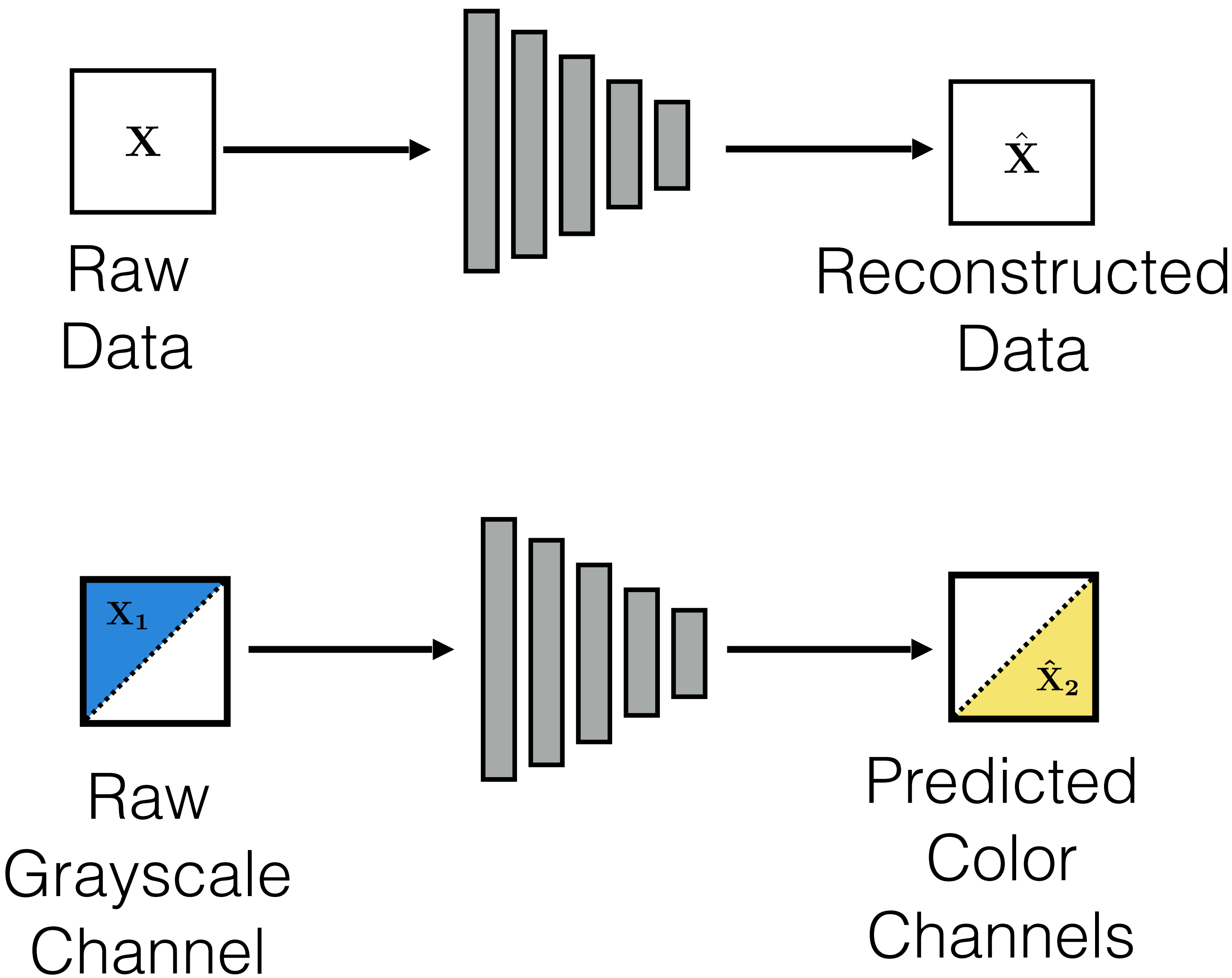


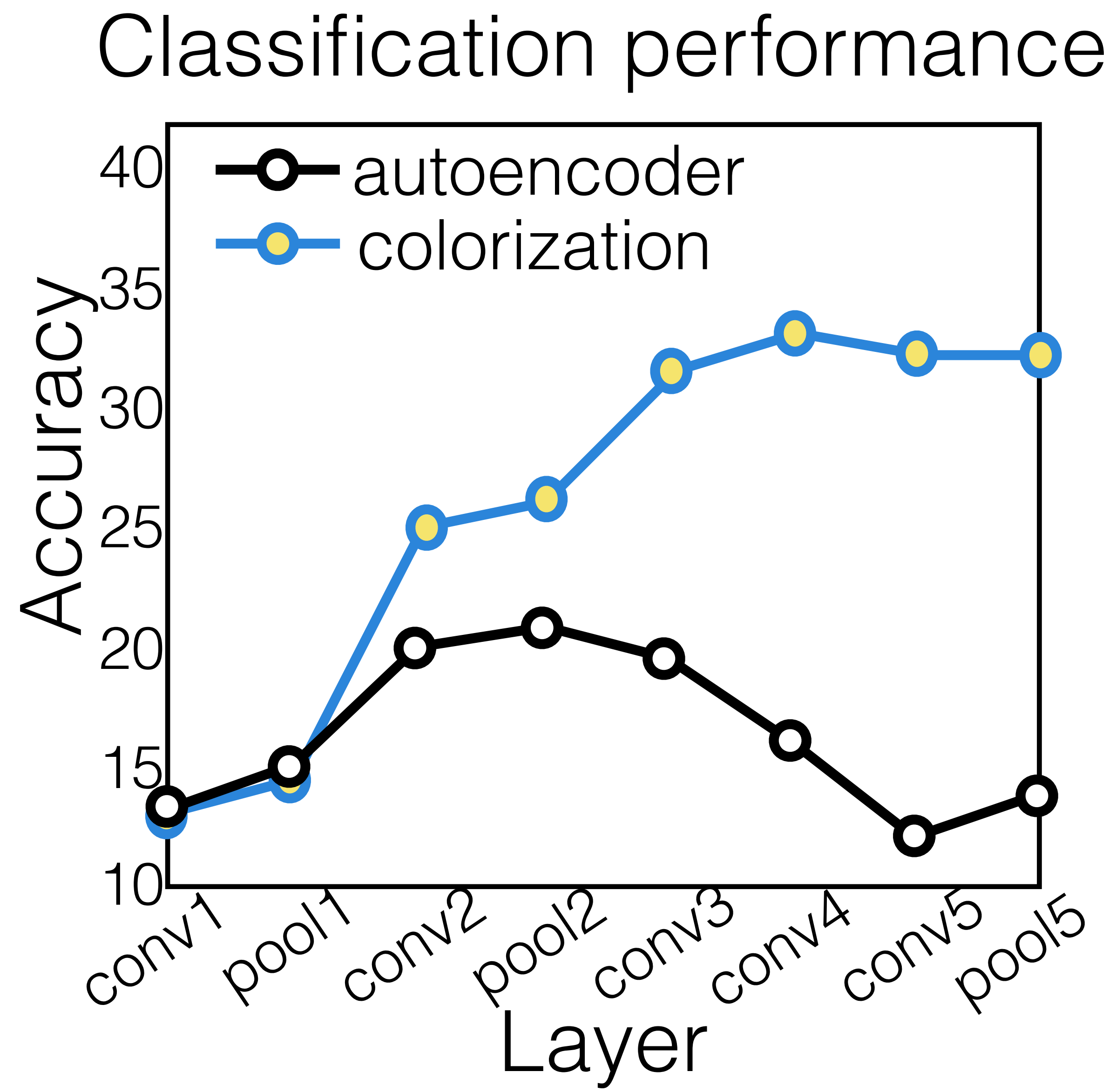


Is the code informative about object class y?

Logistic regression:

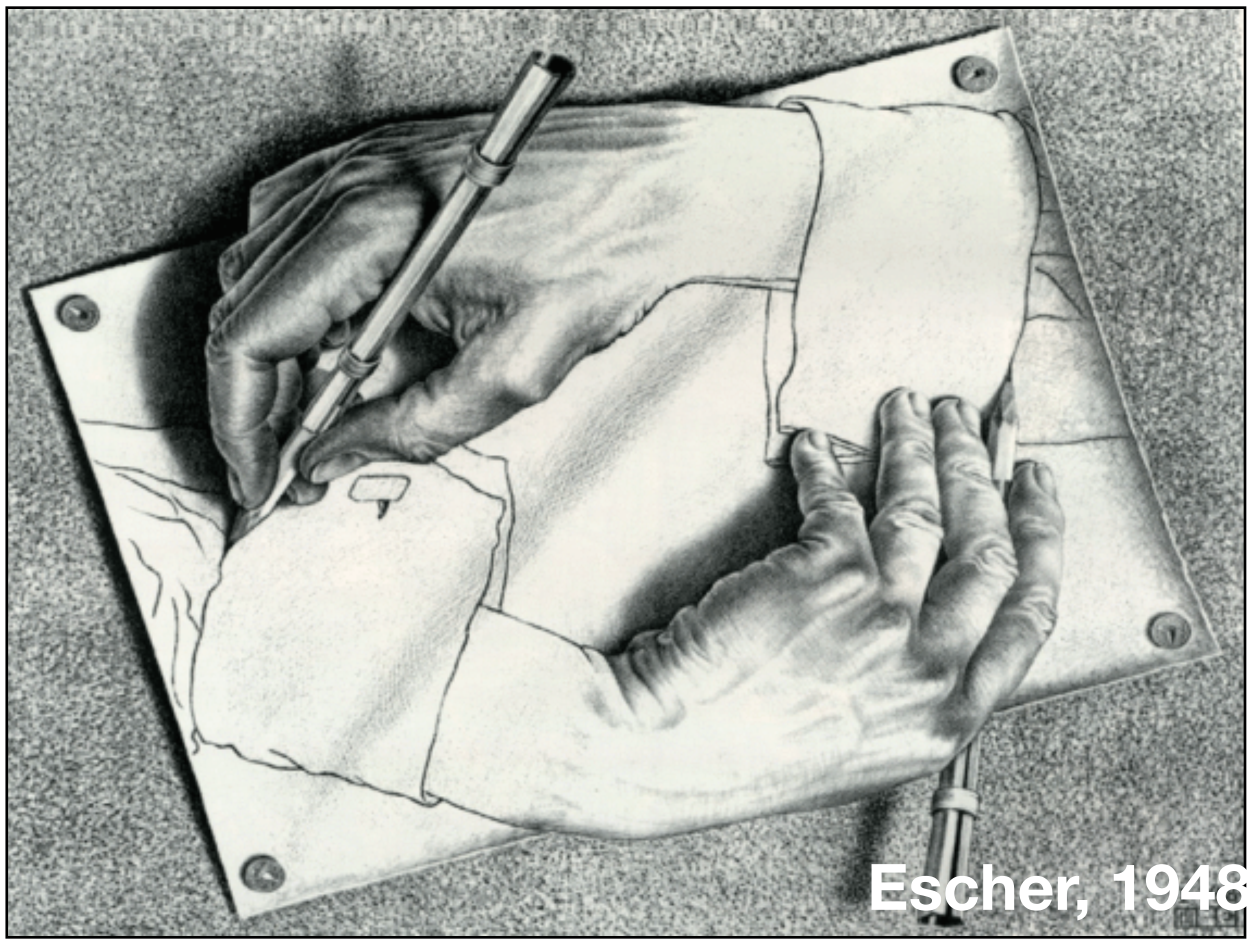
$$y = \sigma(\mathbf{W}\mathbf{z} + \mathbf{b})$$





Task from [Russakovsky et al. 2015]

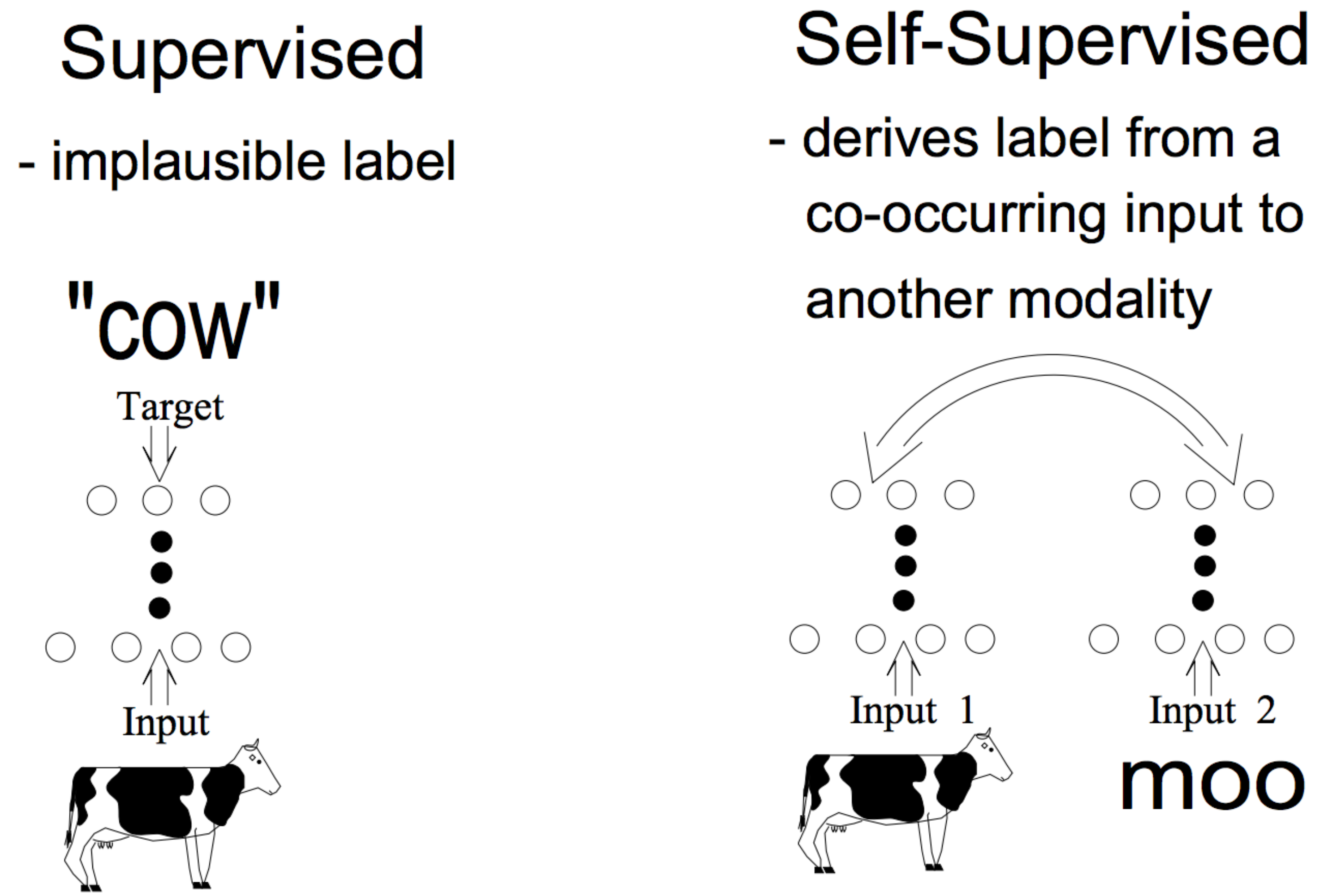
Self-supervised learning



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



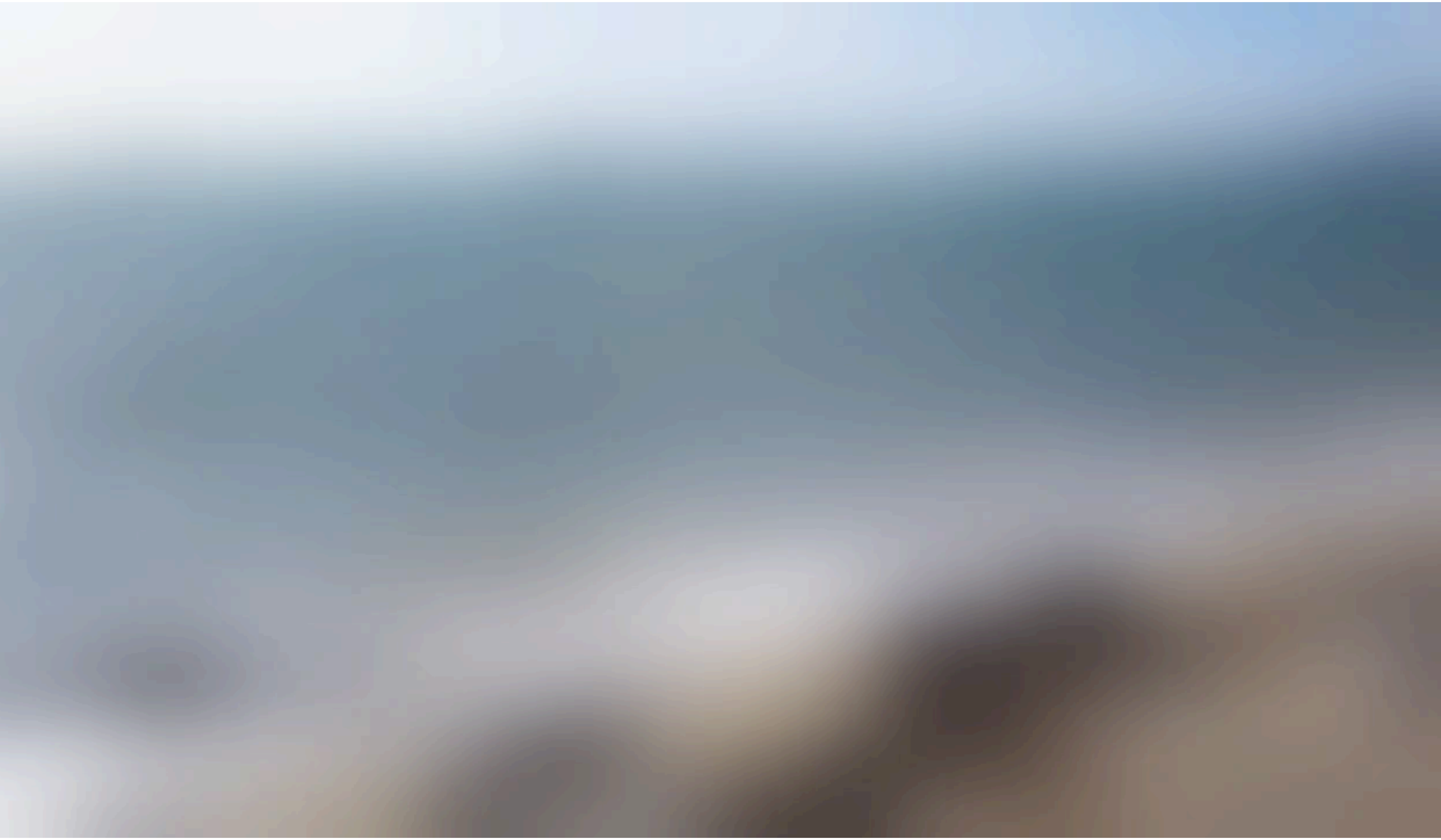
Virginia de Sa. *Learning Classification with Unlabeled Data*. NIPS 1994.

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

Ambient Sound Provides Supervision for Visual Learning

Andrew Owens Jiajun Wu Josh McDermott
William Freeman Antonio Torralba

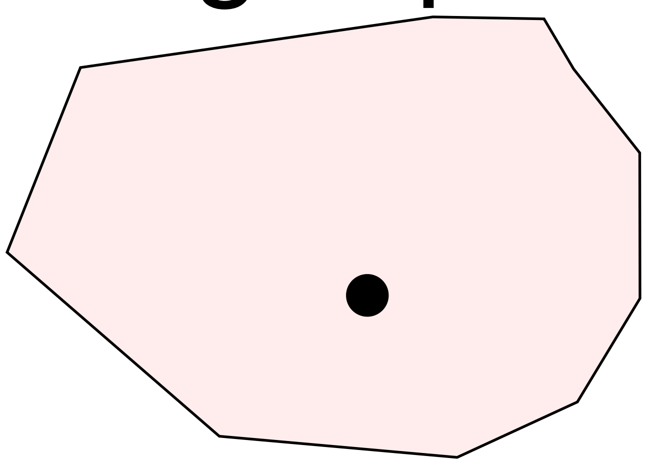
MIT, Google Research



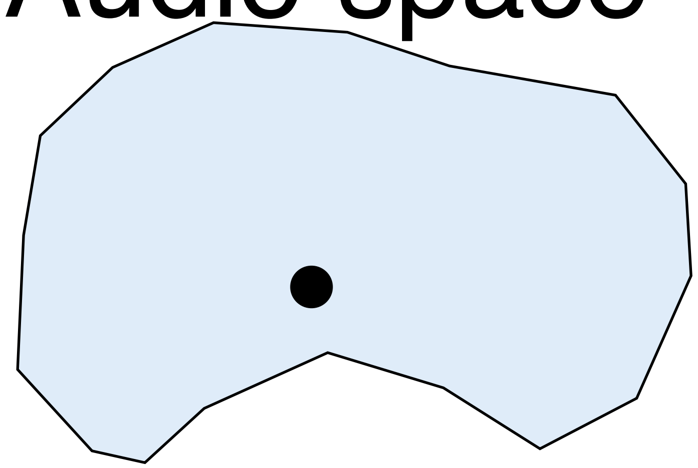
Slide credit: Andrew Owens

Audio is invariant to many visual transformations

Image space



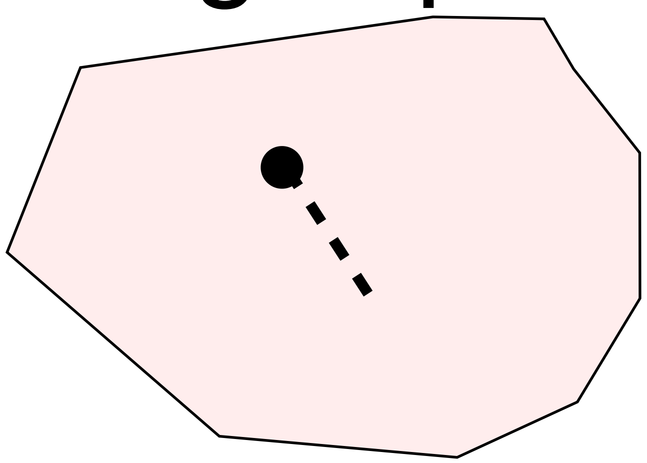
Audio space



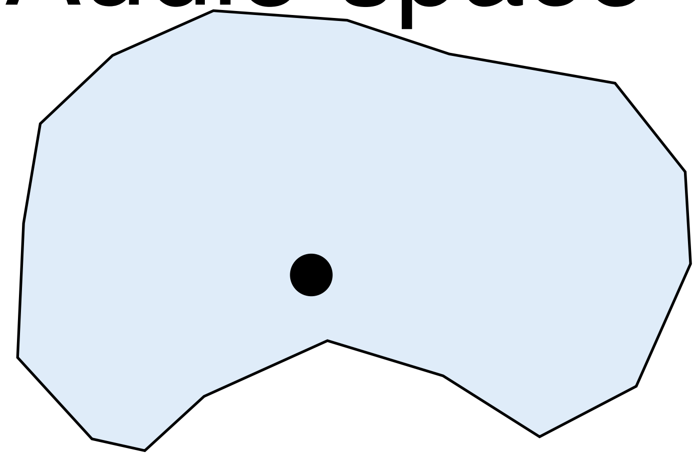
Slide credit: Andrew Owens

Audio is invariant to many visual transformations

Image space

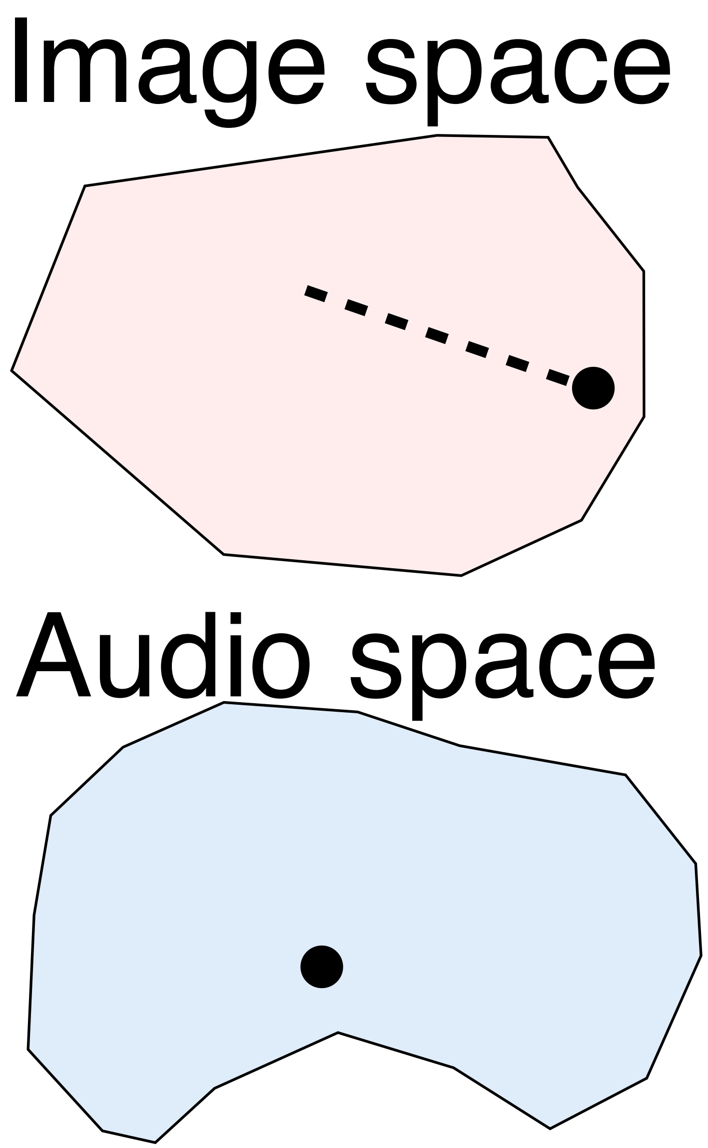


Audio space



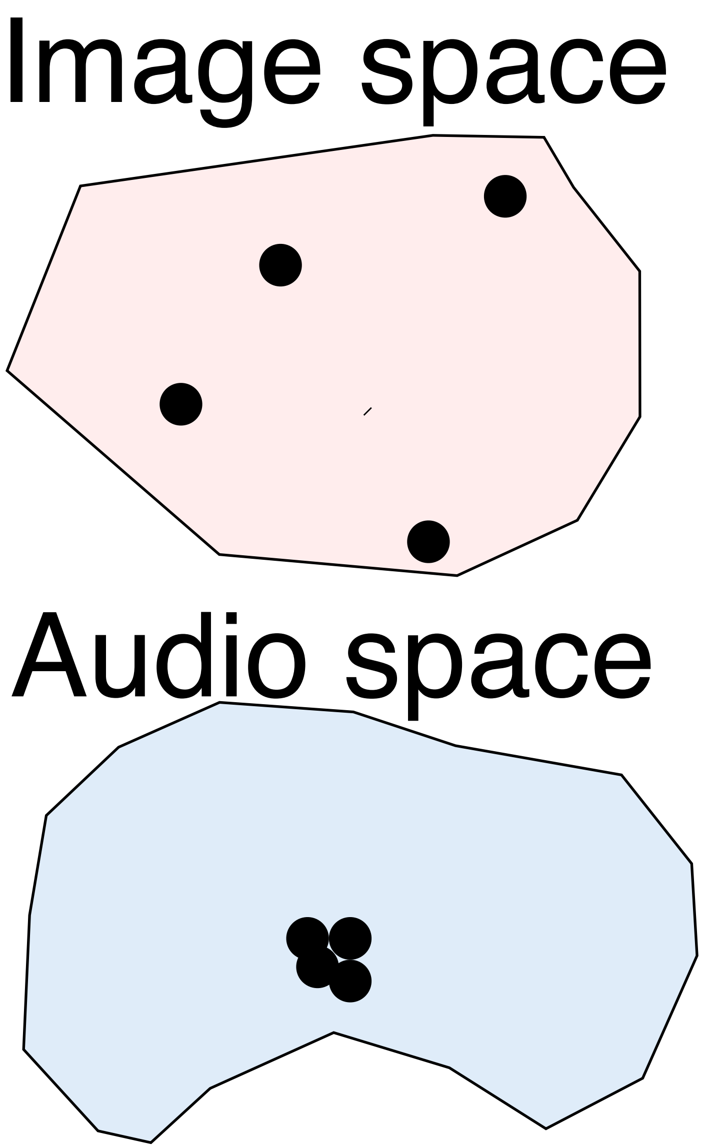
Slide credit: Andrew Owens

Audio is invariant to many visual transformations



Slide credit: Andrew Owens

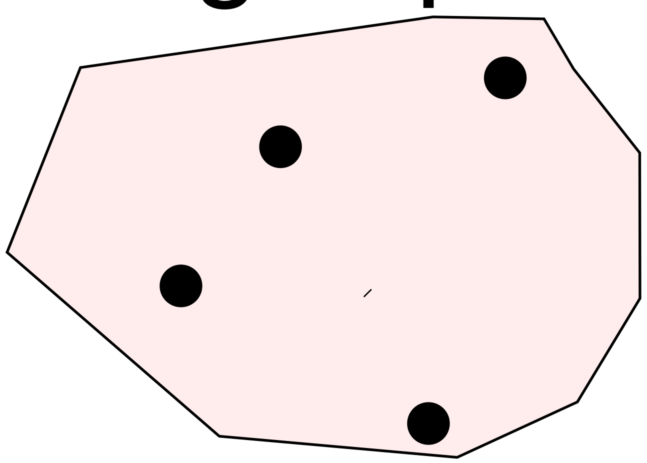
Audio is invariant to many visual transformations



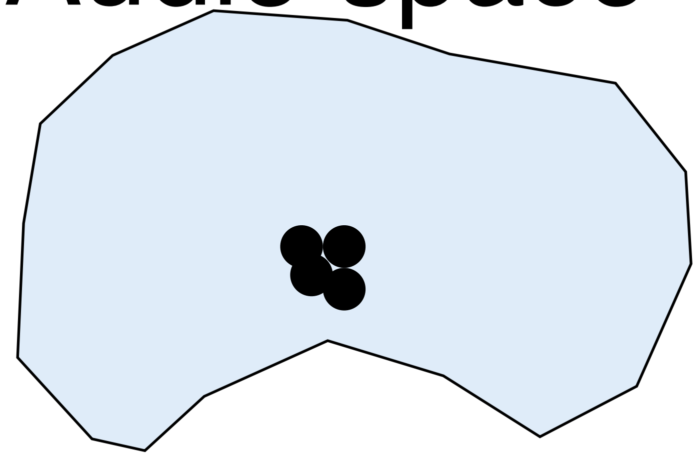
Slide credit: Andrew Owens

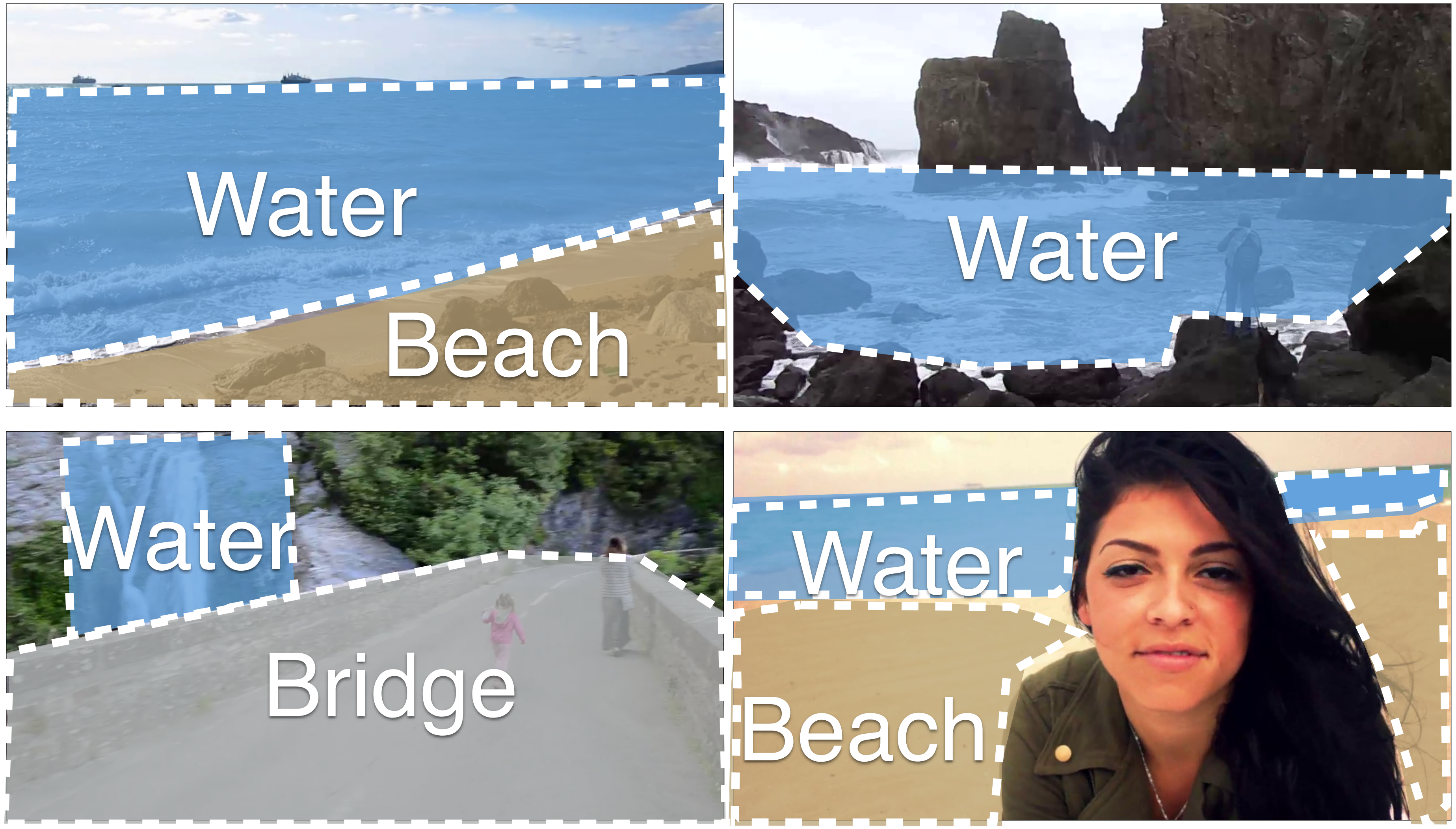
Audio is invariant to many visual transformations

Image space



Audio space



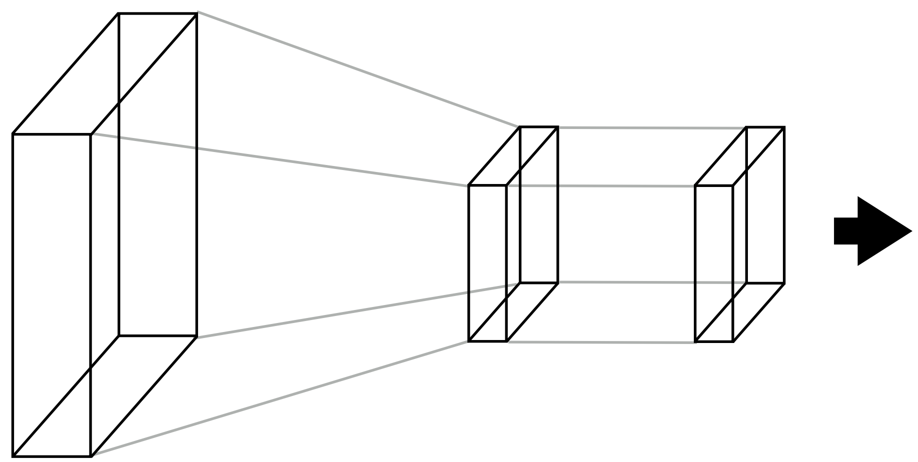


Slide credit: Andrew Owens

Predicting sound

- Flickr video dataset.
- 180K videos, 10 random frames from each.
- Trained from scratch

Video frame



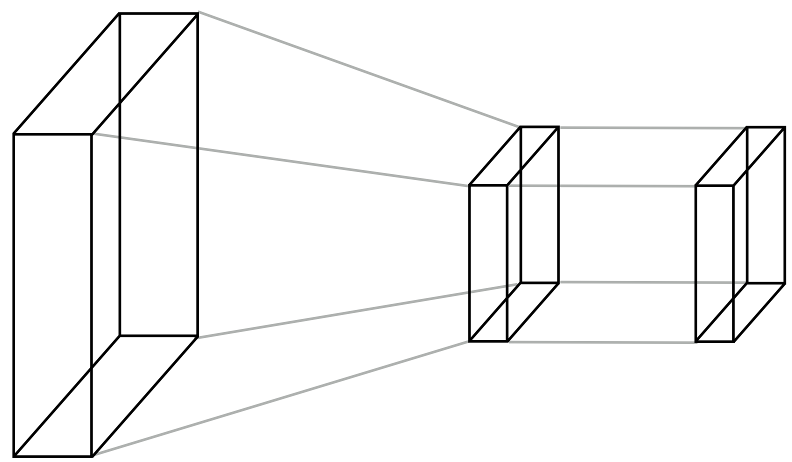
ConvNet

Sound texture
[McDermott & Simoncelli 2011]

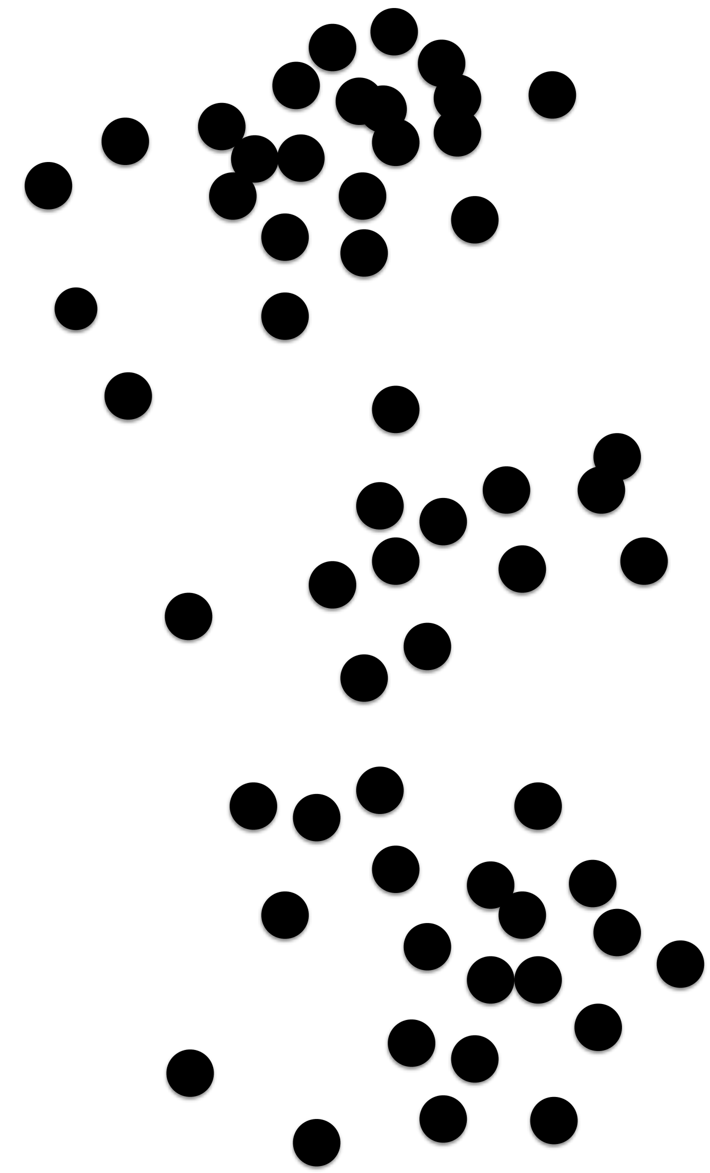
Slide credit: Andrew Owens

Predicting sound

Video frame



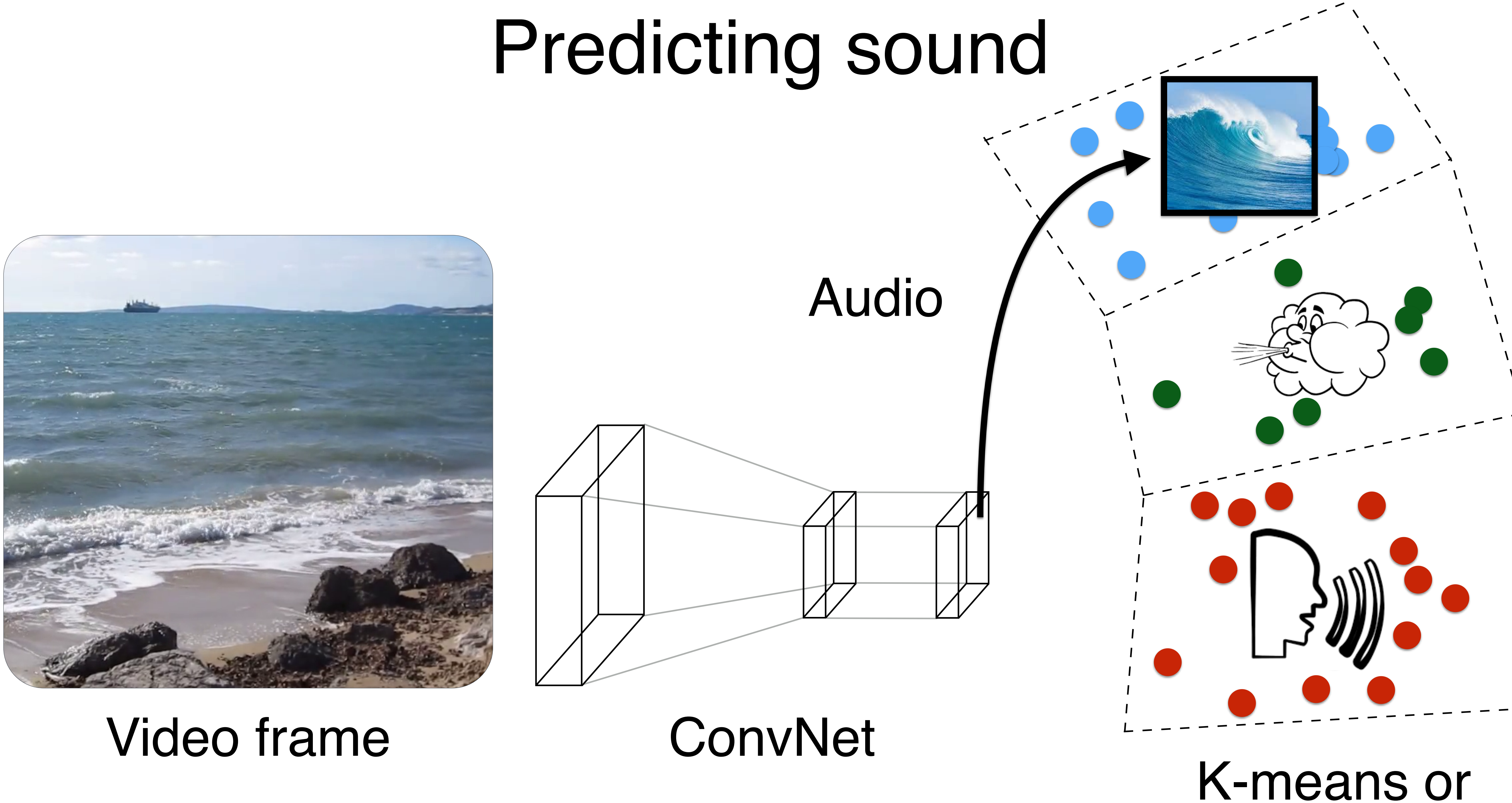
ConvNet



Sound feature

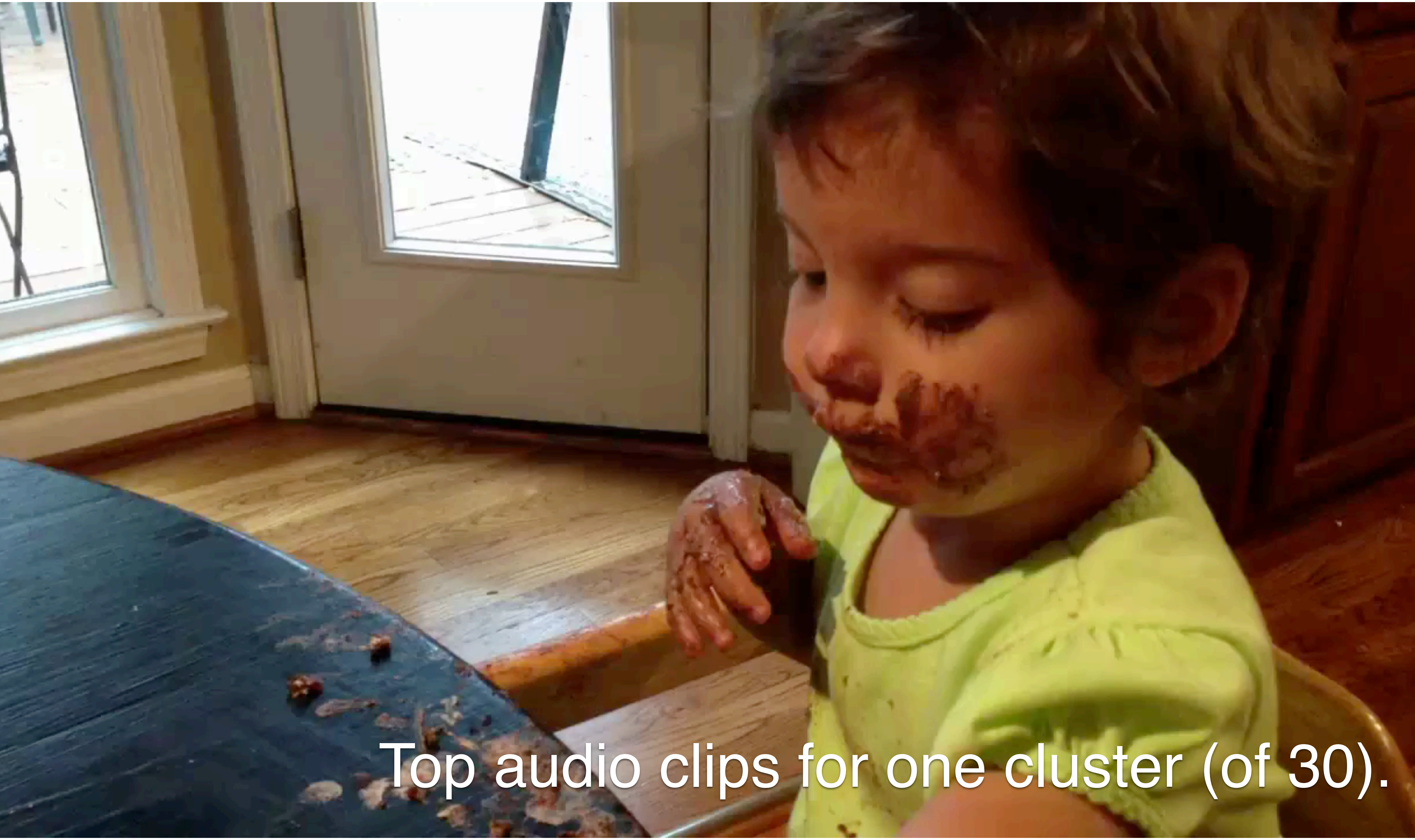
Slide credit: Andrew Owens

Predicting sound

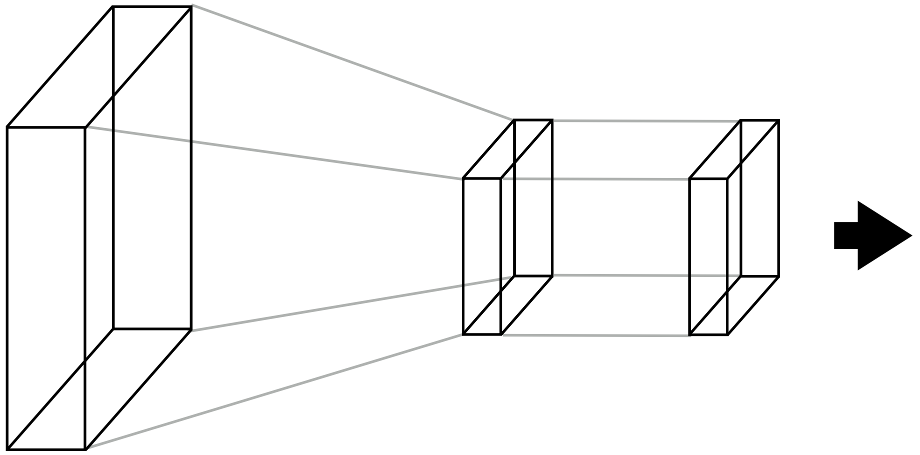


Slide credit: Andrew Owens


Top audio clips for one cluster (of 30).


Top audio clips for one cluster (of 30).

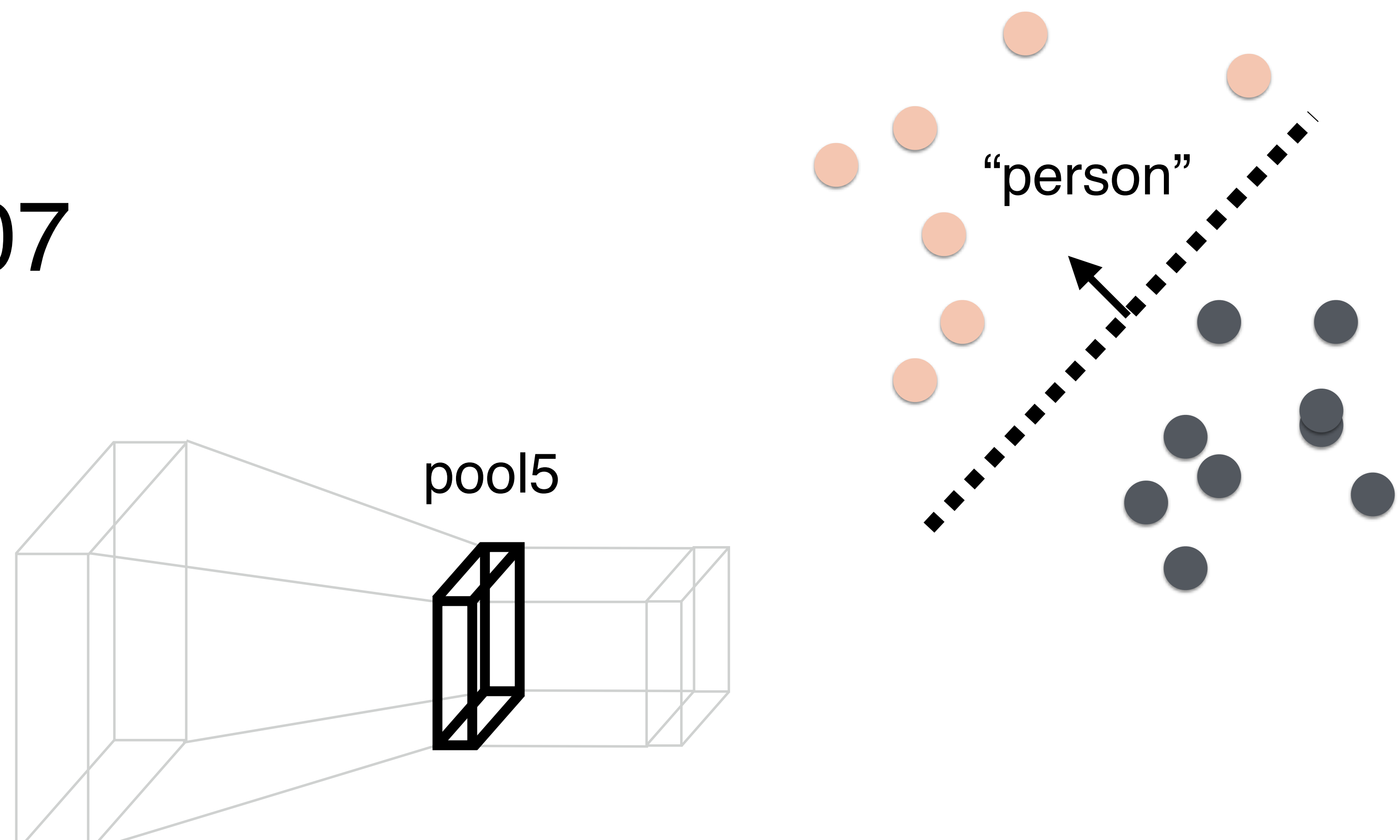
What did the network learn?



Slide credit: Andrew Owens

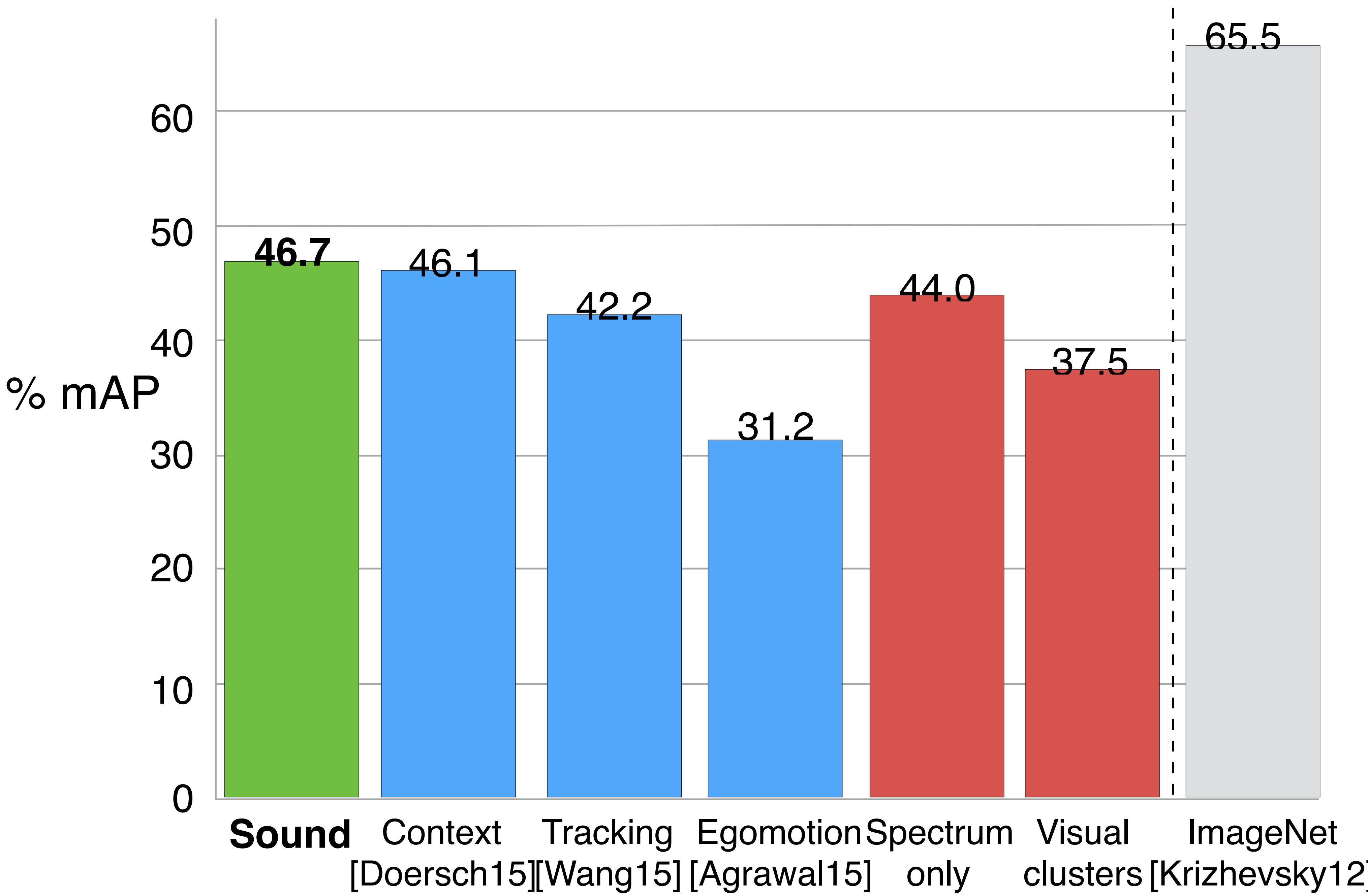
What did the network learn?

PASCAL VOC 2007



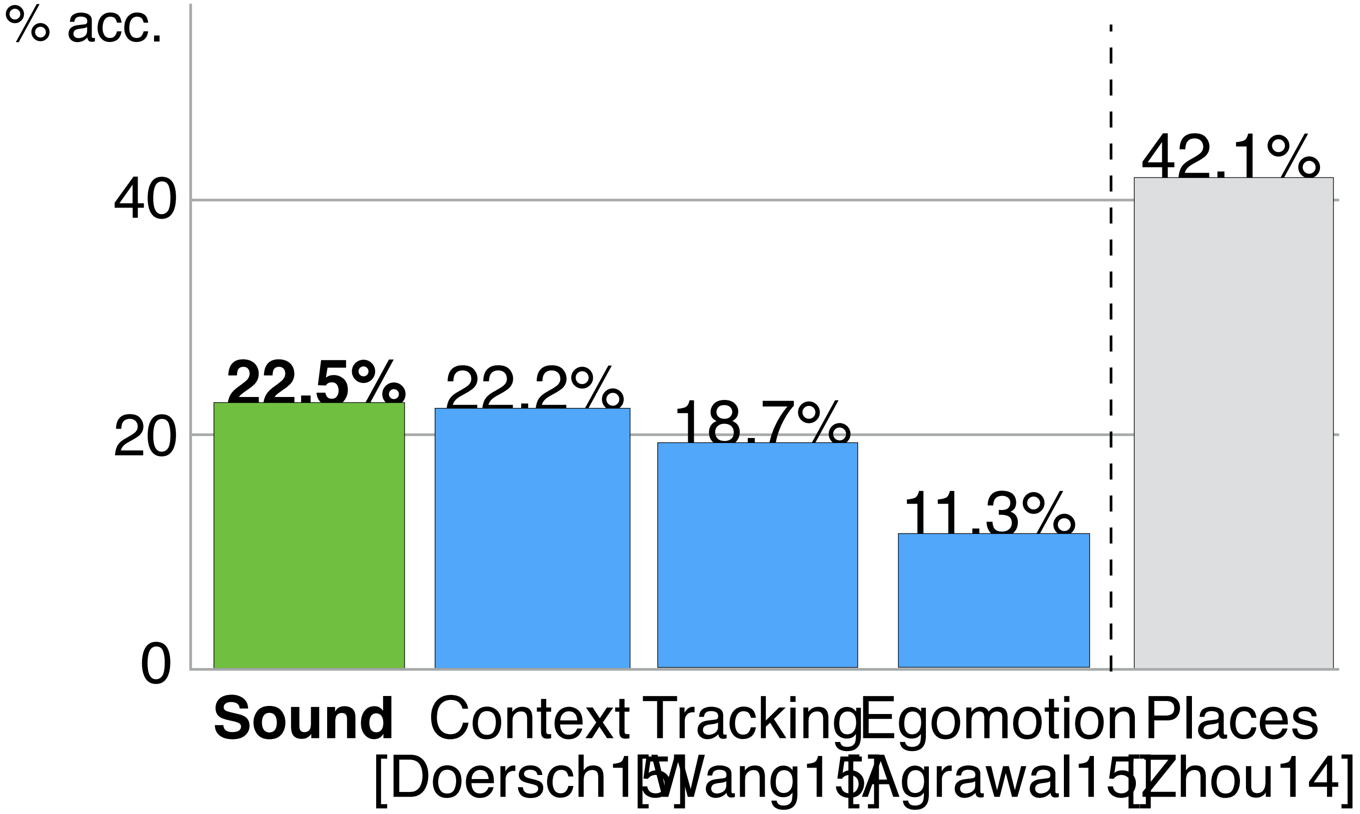
Slide credit: Andrew Owens

PASCAL VOC Classification



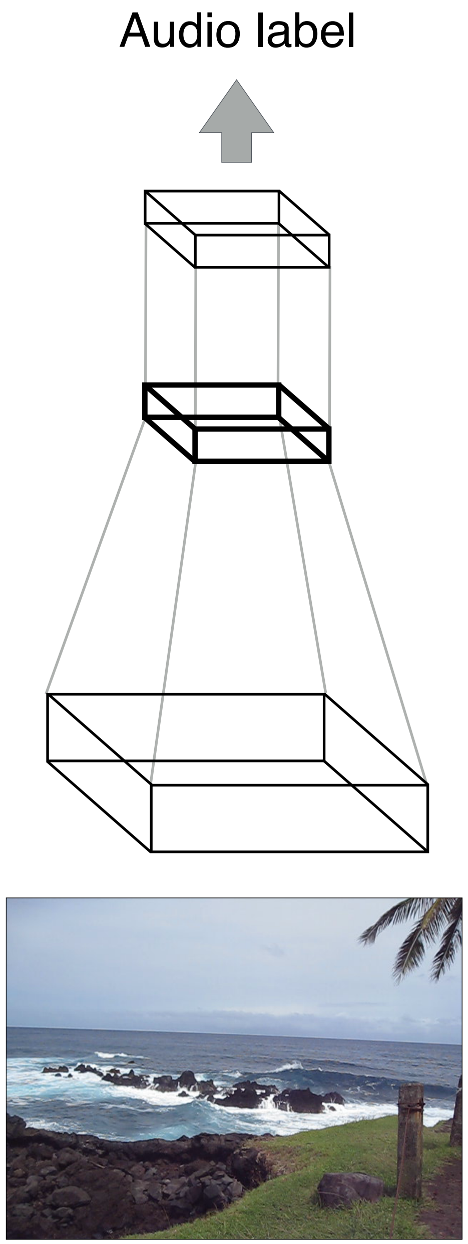
Slide credit: Andrew Owens

SUN397 Scene Recognition



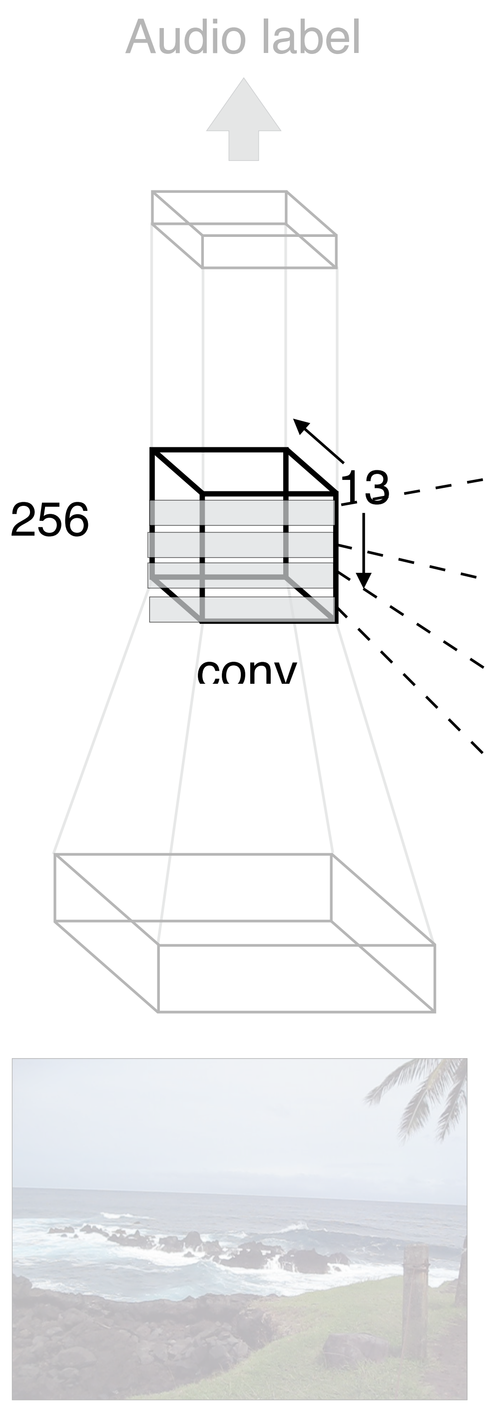
Slide credit: Andrew Owens

What did the network learn?

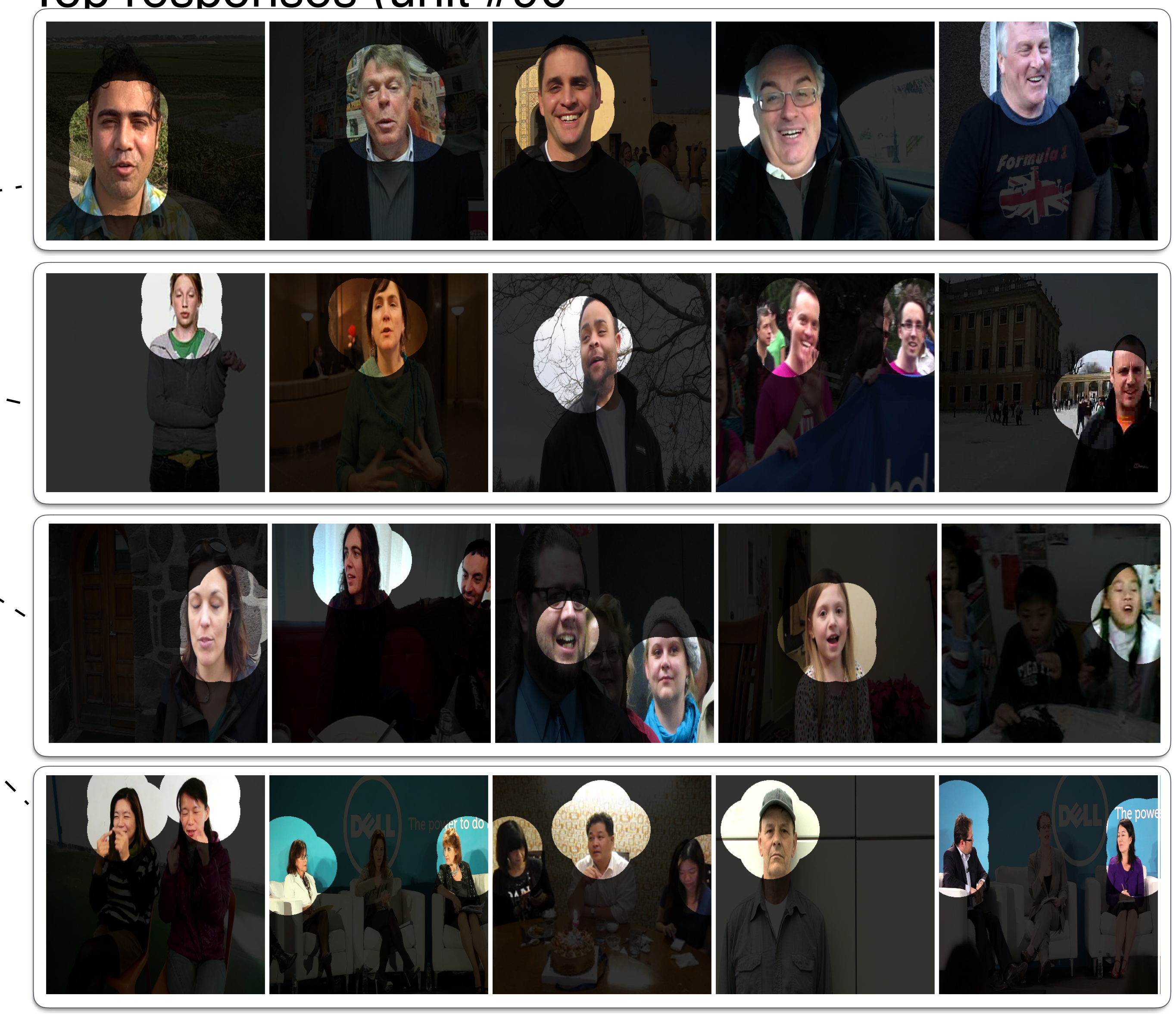


Slide credit: Andrew Owens

Unit visualizations

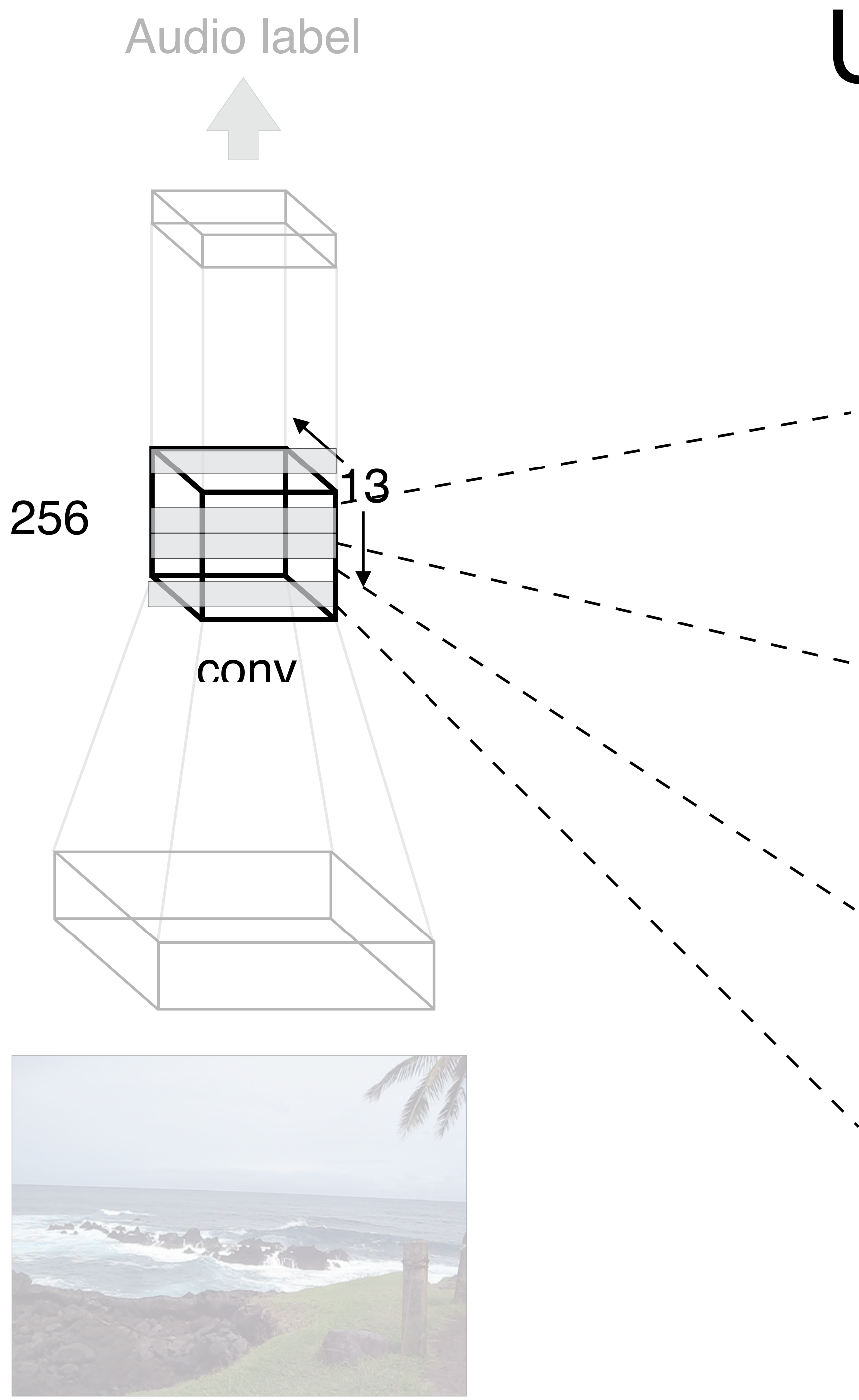


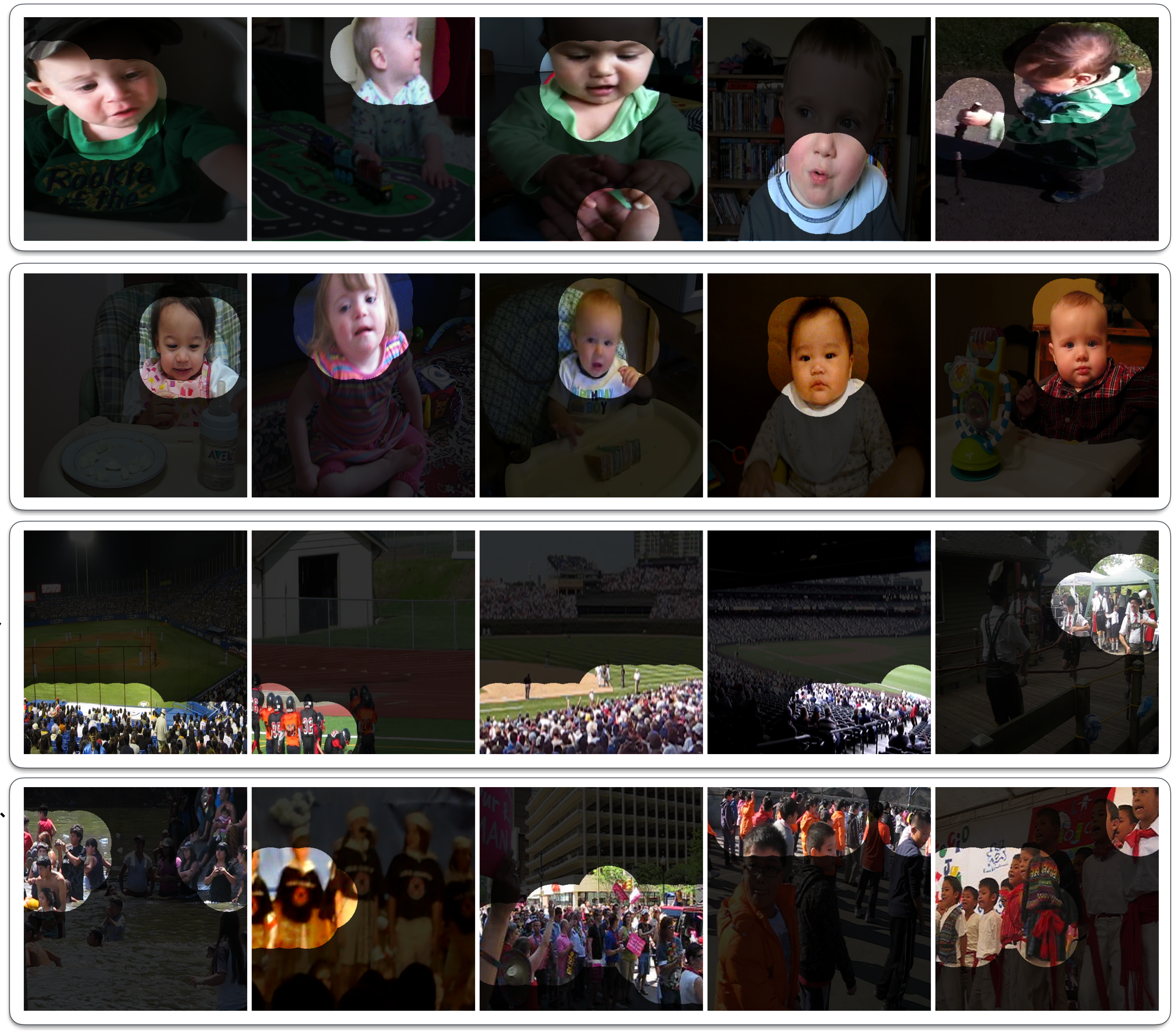
Top responses (unit #90



Slide credit: Andrew Owens

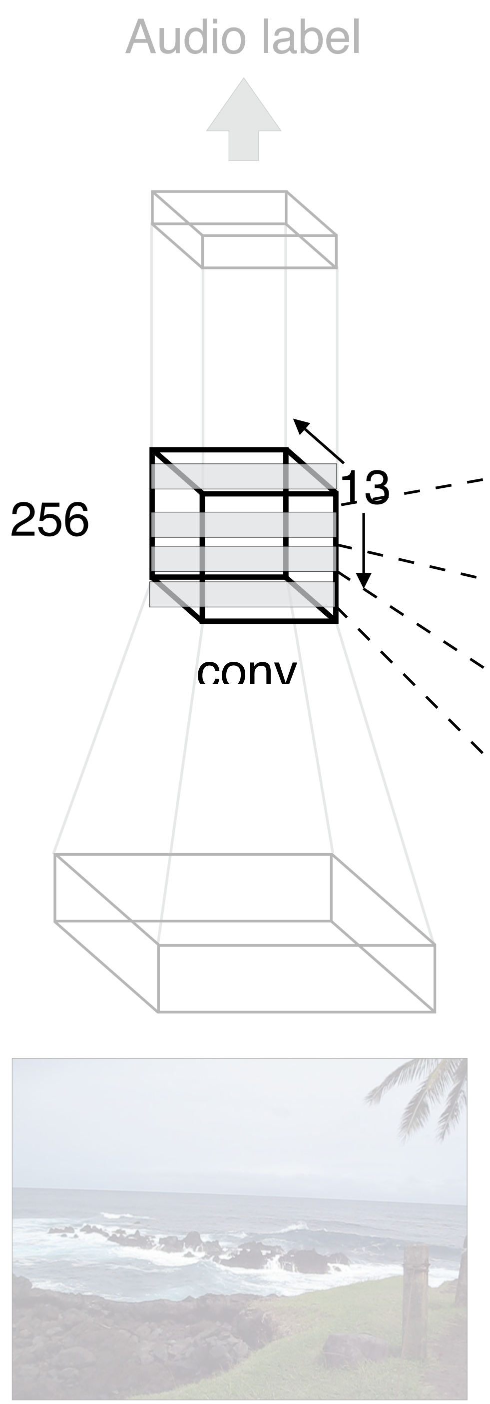
Unit visualizations

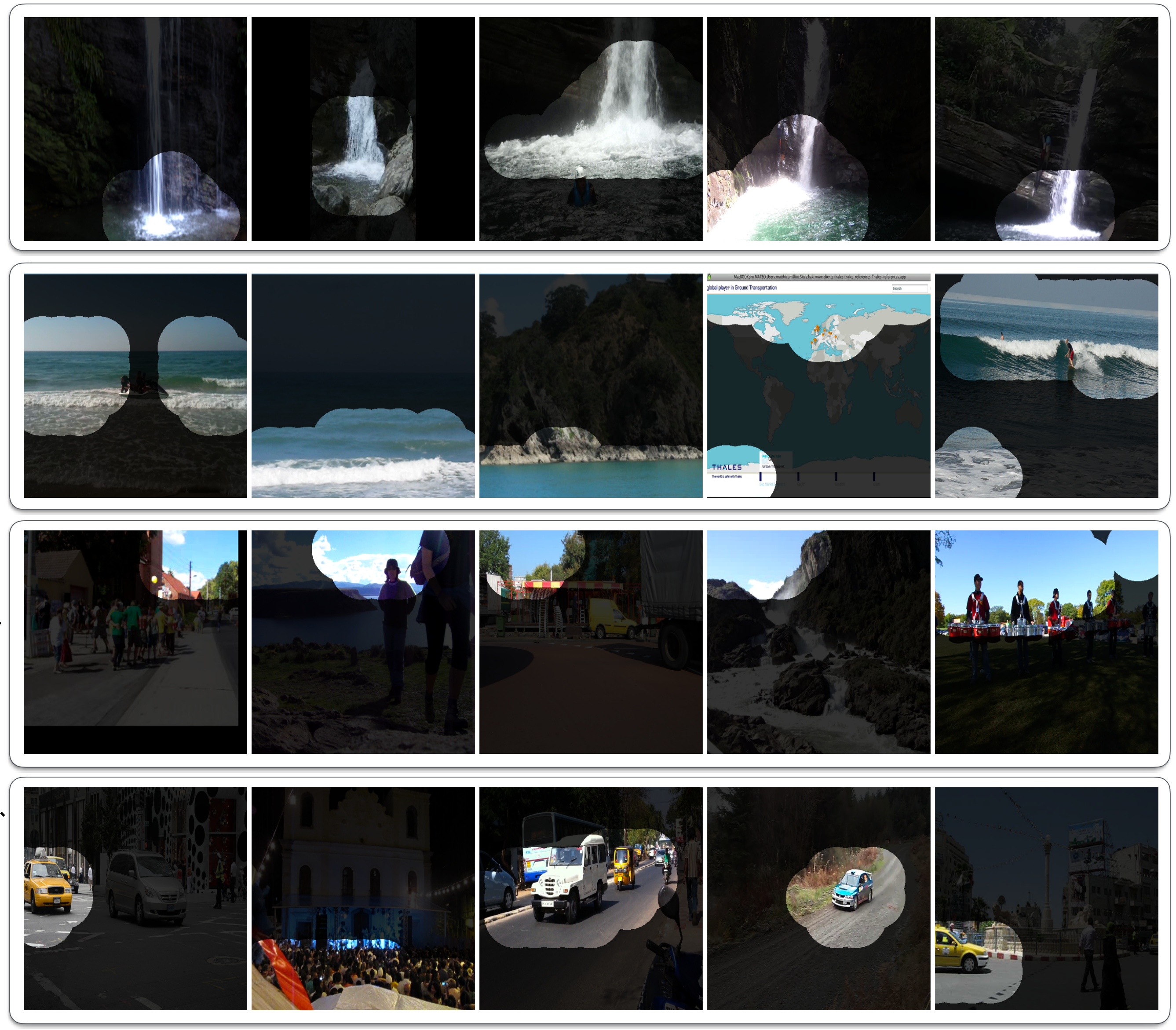




Slide credit: Andrew Owens

Unit visualizations





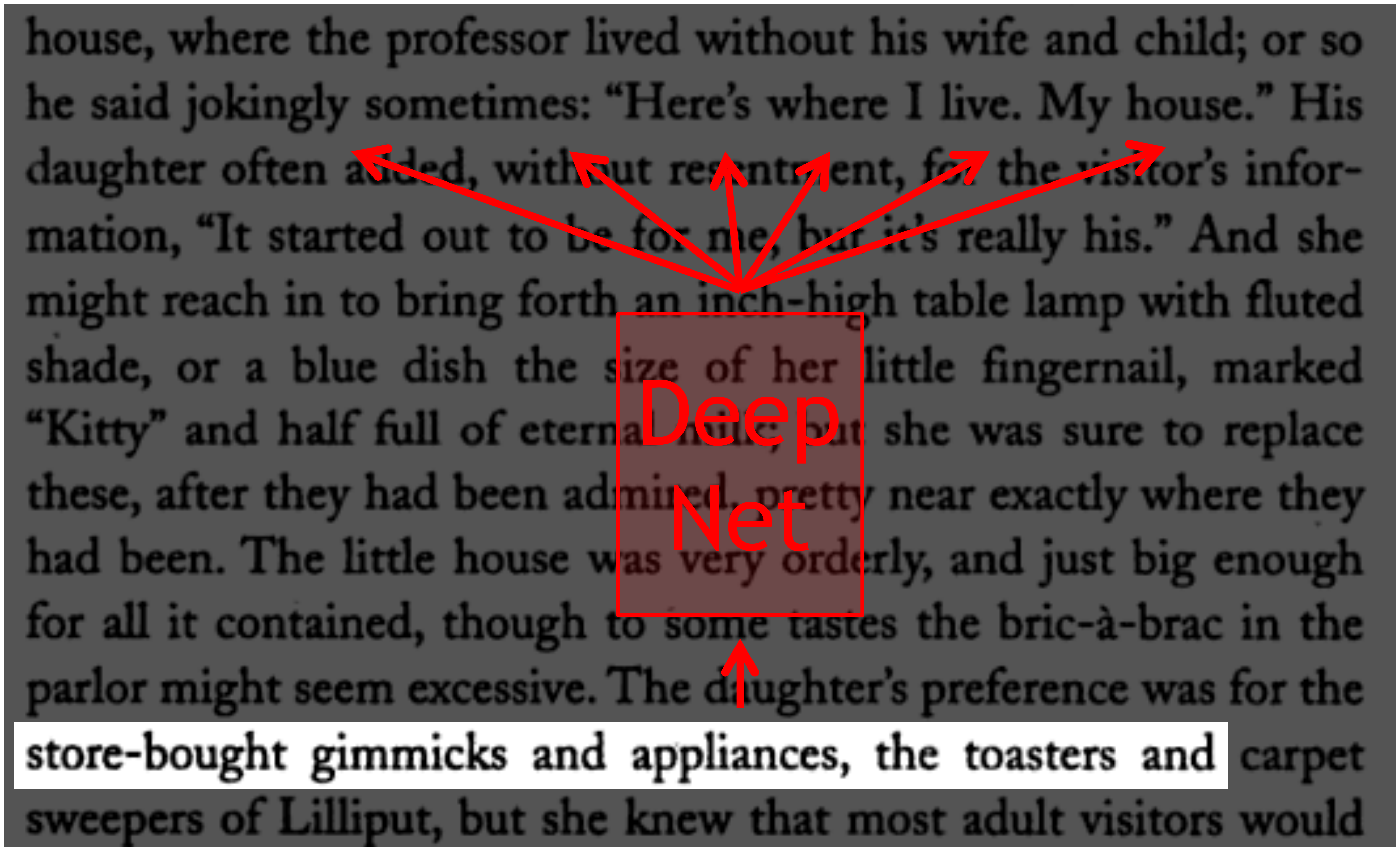
Slide credit: Andrew Owens

Unsupervised visual representation learning by context prediction

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

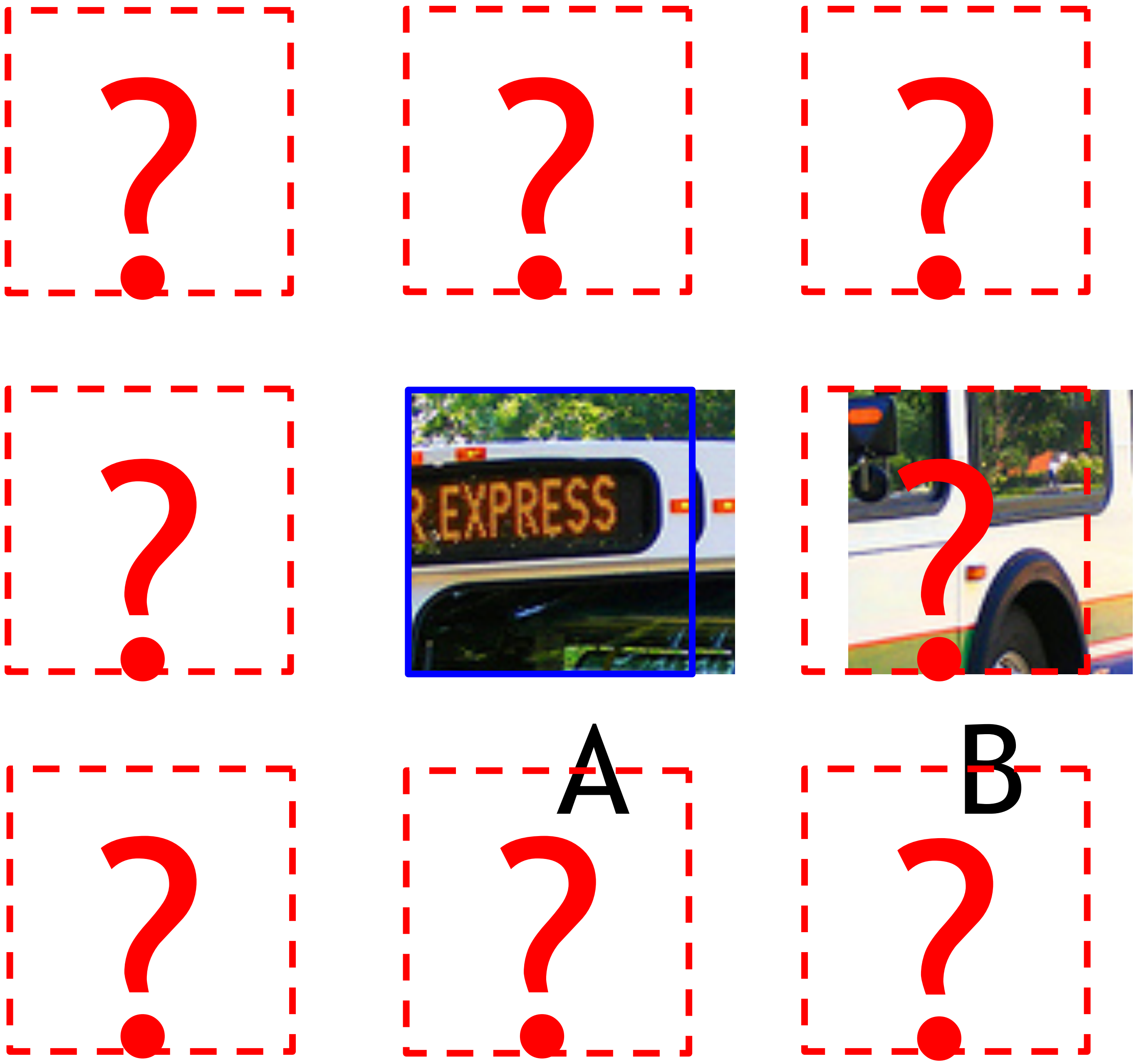
Context as Supervision

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



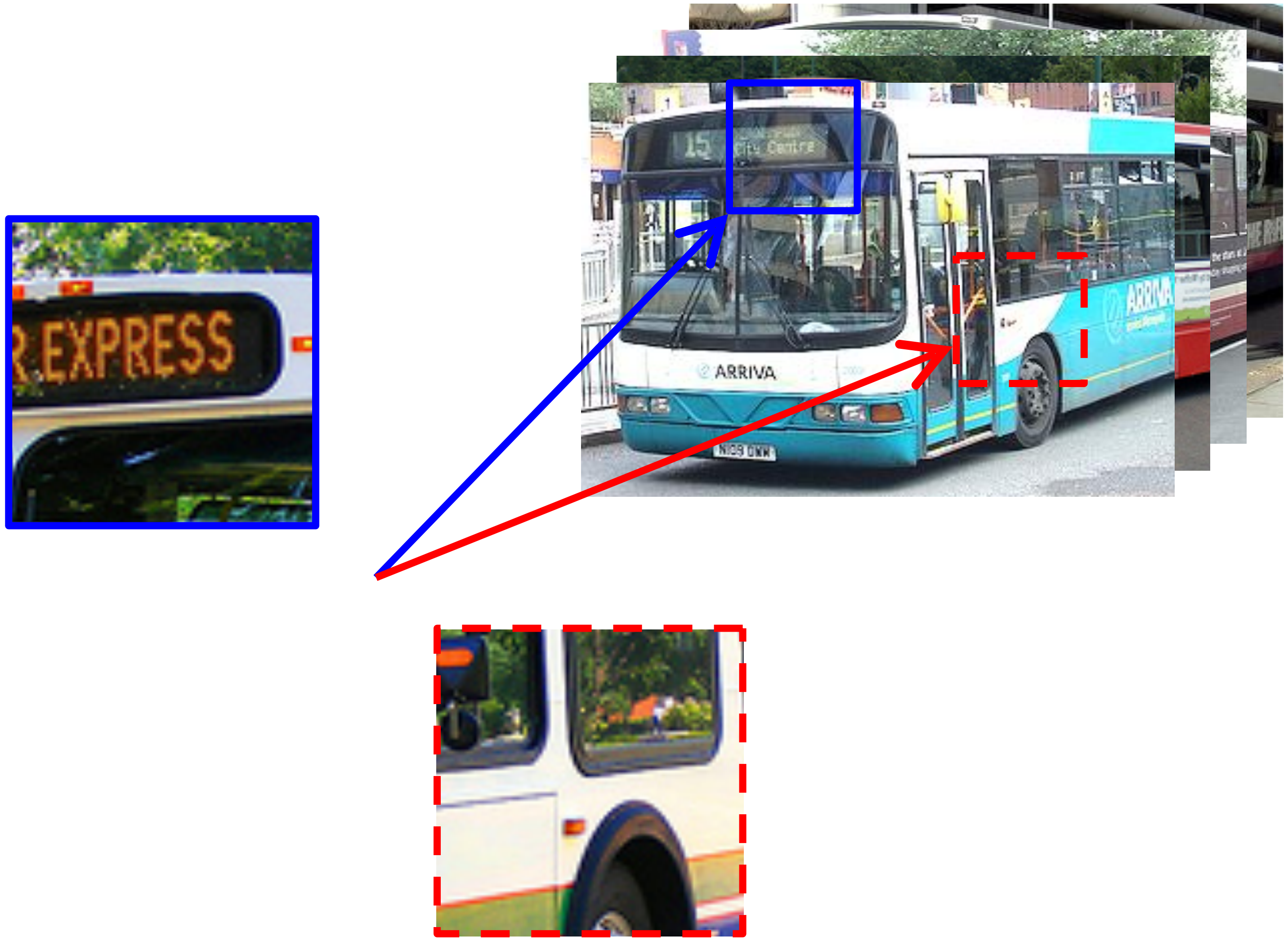
[Slide credit: Carl Doersch]

Context Prediction for Images



[Slide credit: Carl Doersch]

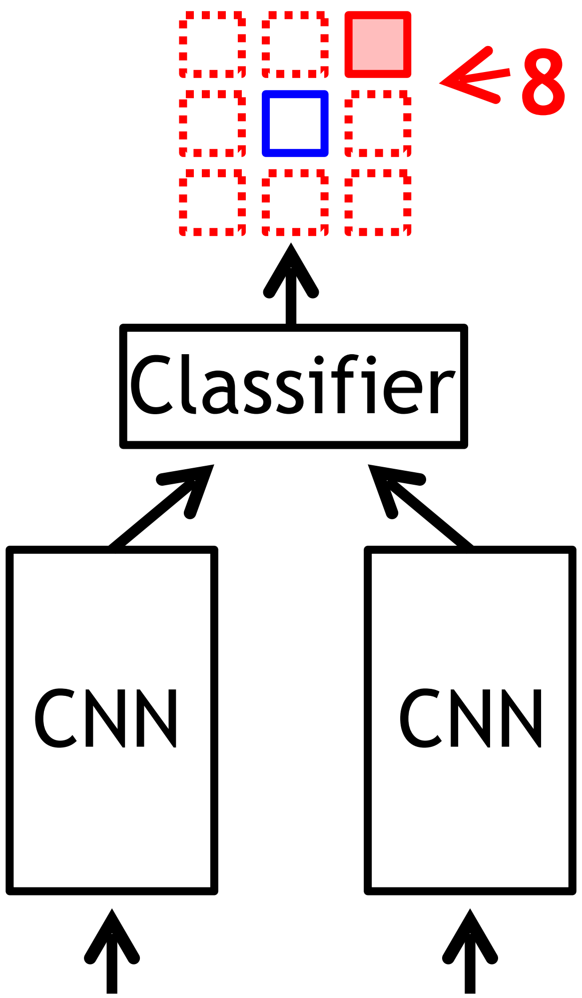
Semantics from a non-semantic task

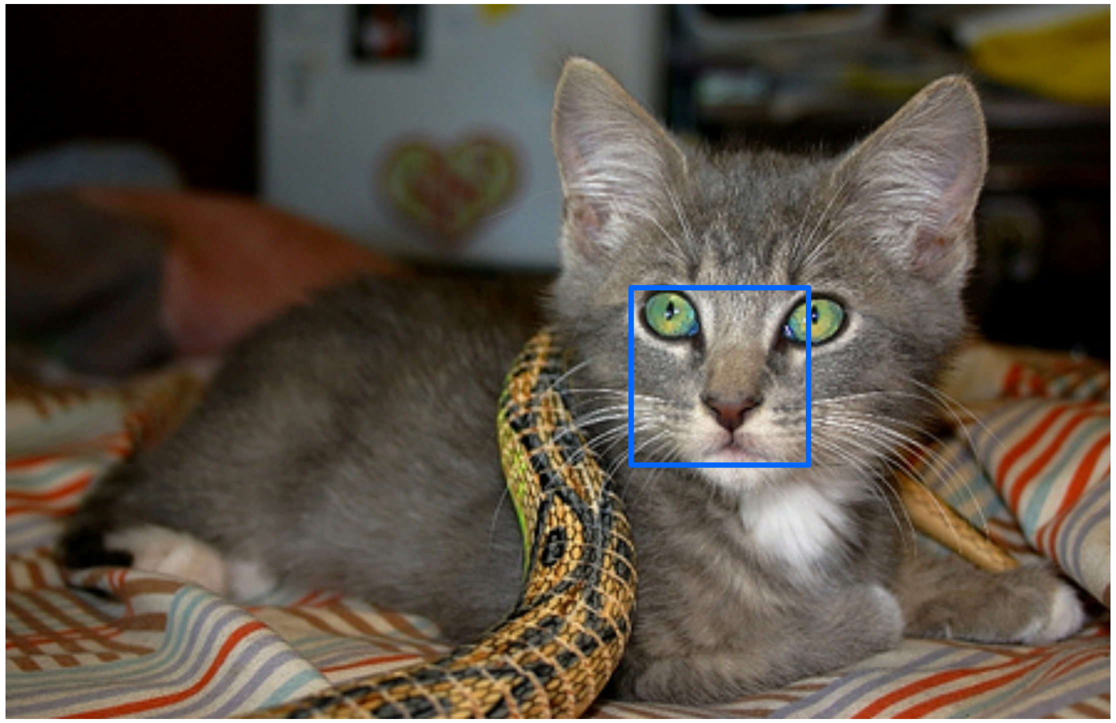


[Slide credit: Carl Doersch]

Relative Position Task

8 possible locations

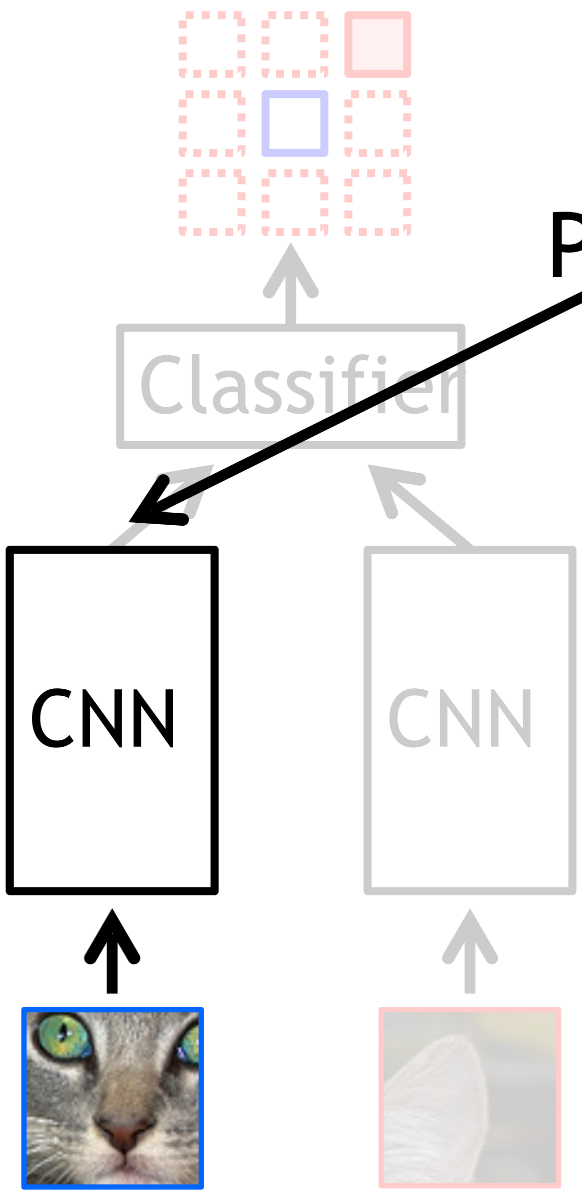




Randomly Sample Patch

Sample Second Patch

[Slide credit: Carl Doersch]



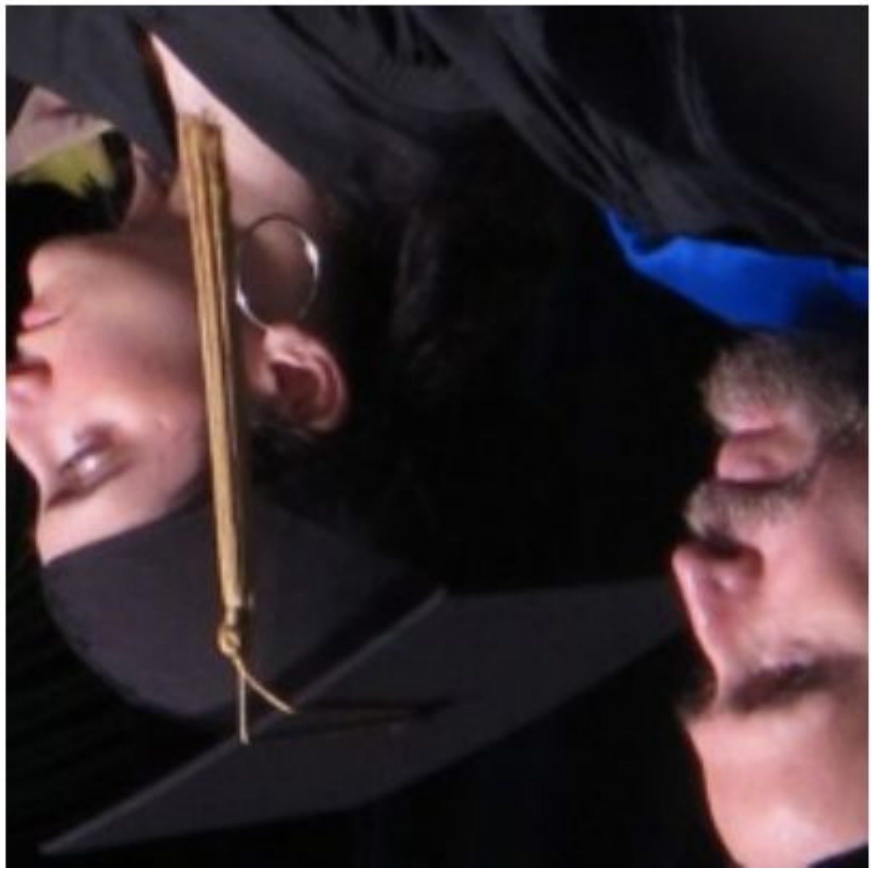
Patch Embedding (representation)

Input — Nearest Neighbors

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

[Slide credit: Carl Doersch]

Learning by Rotating



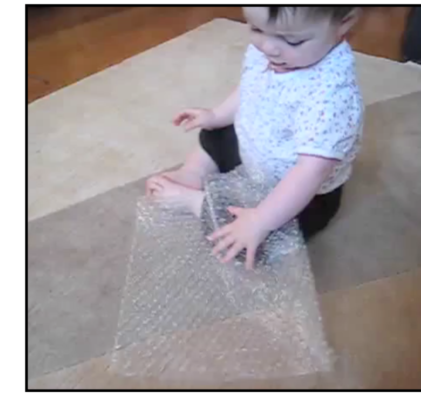
Unsupervised Representation Learning by Predicting Image Rotations
Spyros Gidaris, Praveer Singh, Nikos Komodakis

How are we doing?

	Classification	Detection	Segmentation
ImageNet	78.2%	56.8%	48.0%
Context	55.3%	46.6%	-
Jigsaw Puzzle	67.6%	53.2%	37.6%
Inpainting	56.5%	44.5%	30.0%
Colorization	61.5%	46.9%	35.6%
Tracking	58.7%	47.4%	-
Counting	67.7%	51.4%	36.6%
Rotation	72.9%	54.4%	39.1%

PASCAL VOC 2007

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!

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)