

CS231A

Computer Vision: From 3D Reconstruction to Recognition



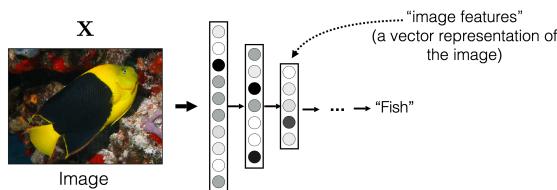
Representation & Representation Learning

How to reach me?

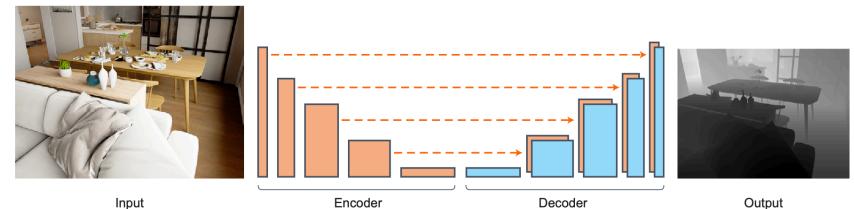
- Jeannette Bohg, Assistant Professor in Robotics
- Office hours, Fridays 1pm, zoom
- By appointment

Learning Goals for Upcoming Lectures

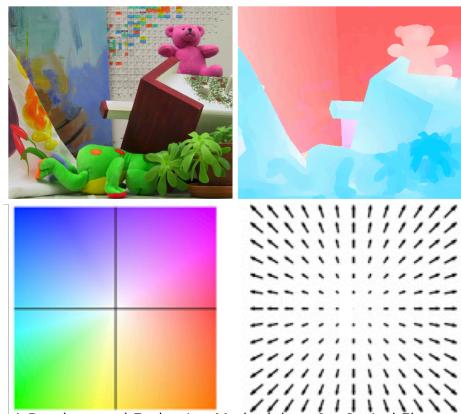
Representations & Representation Learning



Monocular Depth Estimation

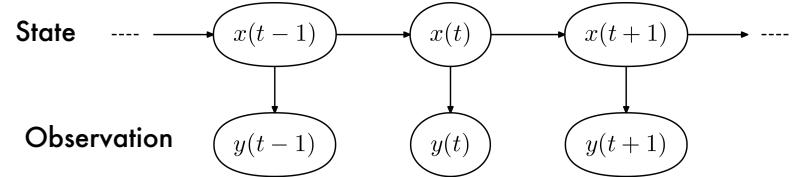


Optical & Scene Flow



A Database and Evaluation Methodology for Optical Flow.
Baker et al. IJCV. 2011

Optimal Estimation



Exercise

- Use the manipulation objects you brought
- What information do you need to solve the task, i.e., to make decision?
- How do you get this information?

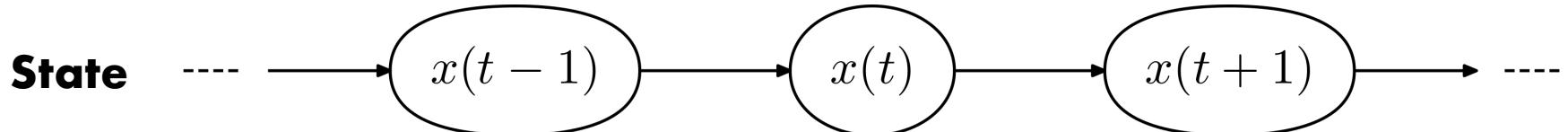
Polleverywhere

- <https://pollev.com/jeannetteboh707>

Outline of this lecture

- What is a state? What is a representation?
- What are the different kinds of representations?
- How can we extract state from raw sensory data?
- What kind of data can we process?

What is a state? What is a representation?



Markov Model



Sparse Cartpole



Acrobot Swingup



Hopper Hop



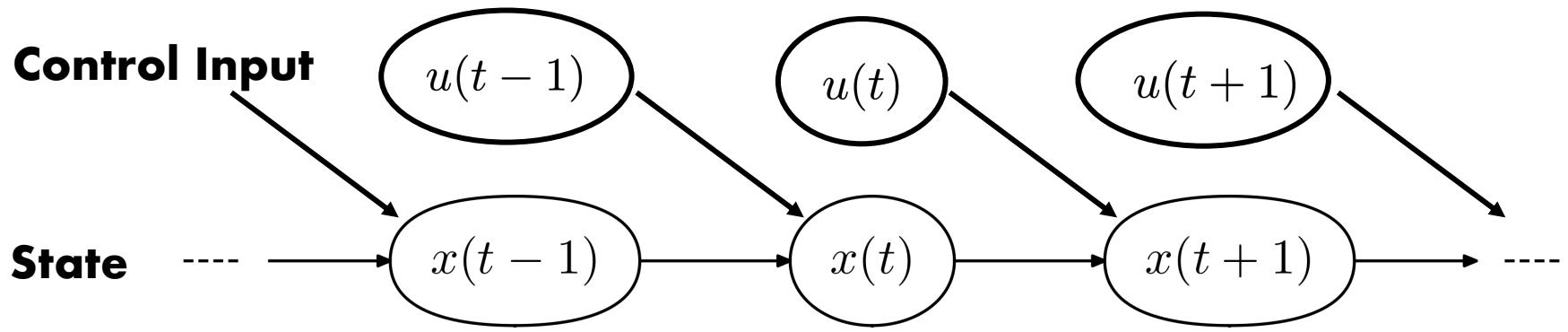
Walker Run



Quadruped Run

DeepMind Control Suite. Tassa et al. 2018

What is a state? What is a representation?



Markov Decision Process



Sparse Cartpole



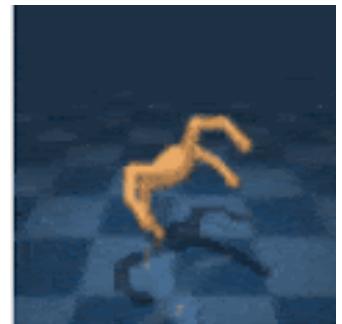
Acrobot Swingup



Hopper Hop



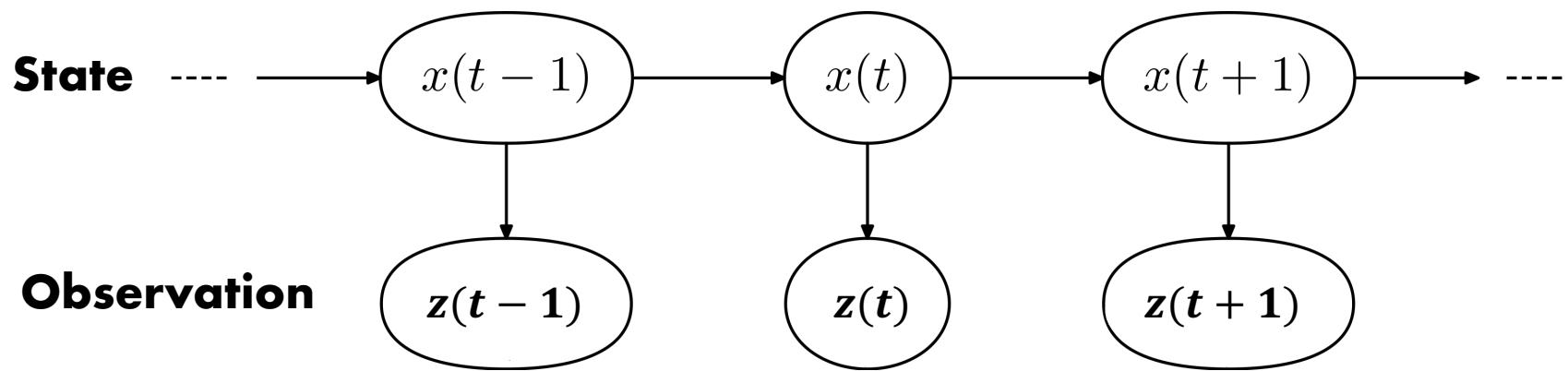
Walker Run



Quadruped Run

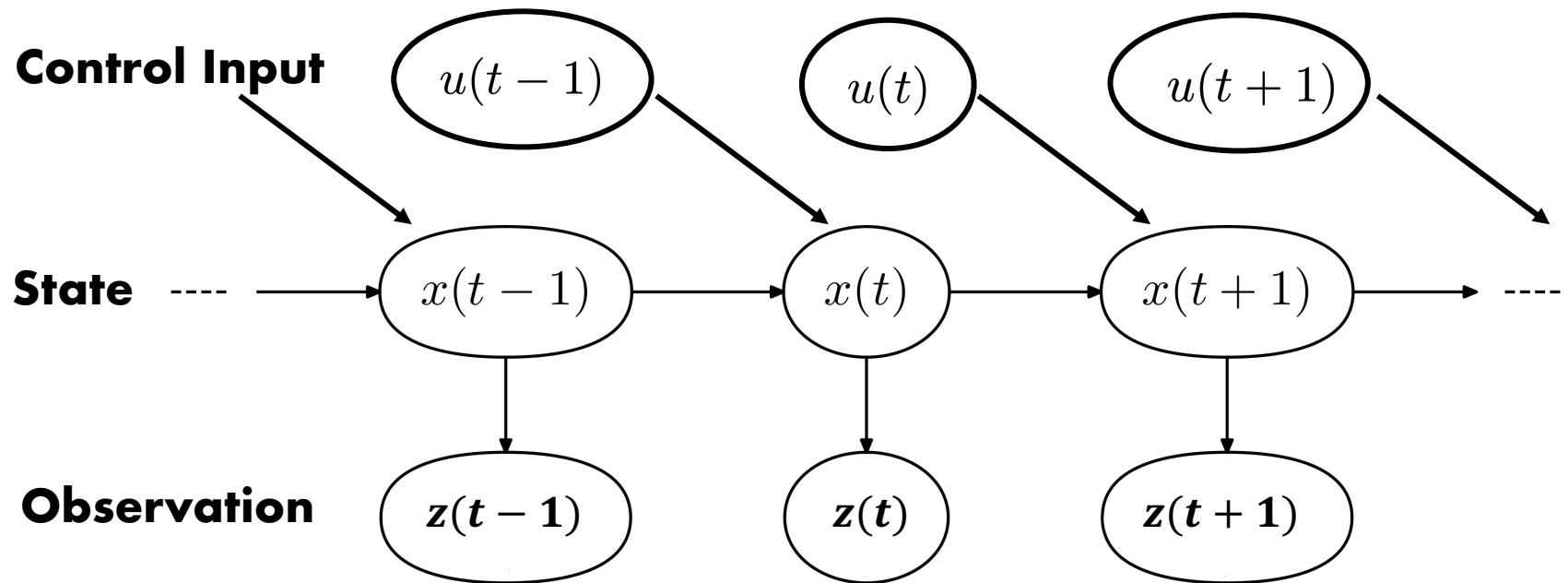
DeepMind Control Suite. Tassa et al. 2018

What is a state? What is a representation?



Hidden Markov Model

What is a state? What is a representation?



Partially Observable Markov Decision Process

Representations for Autonomous Driving

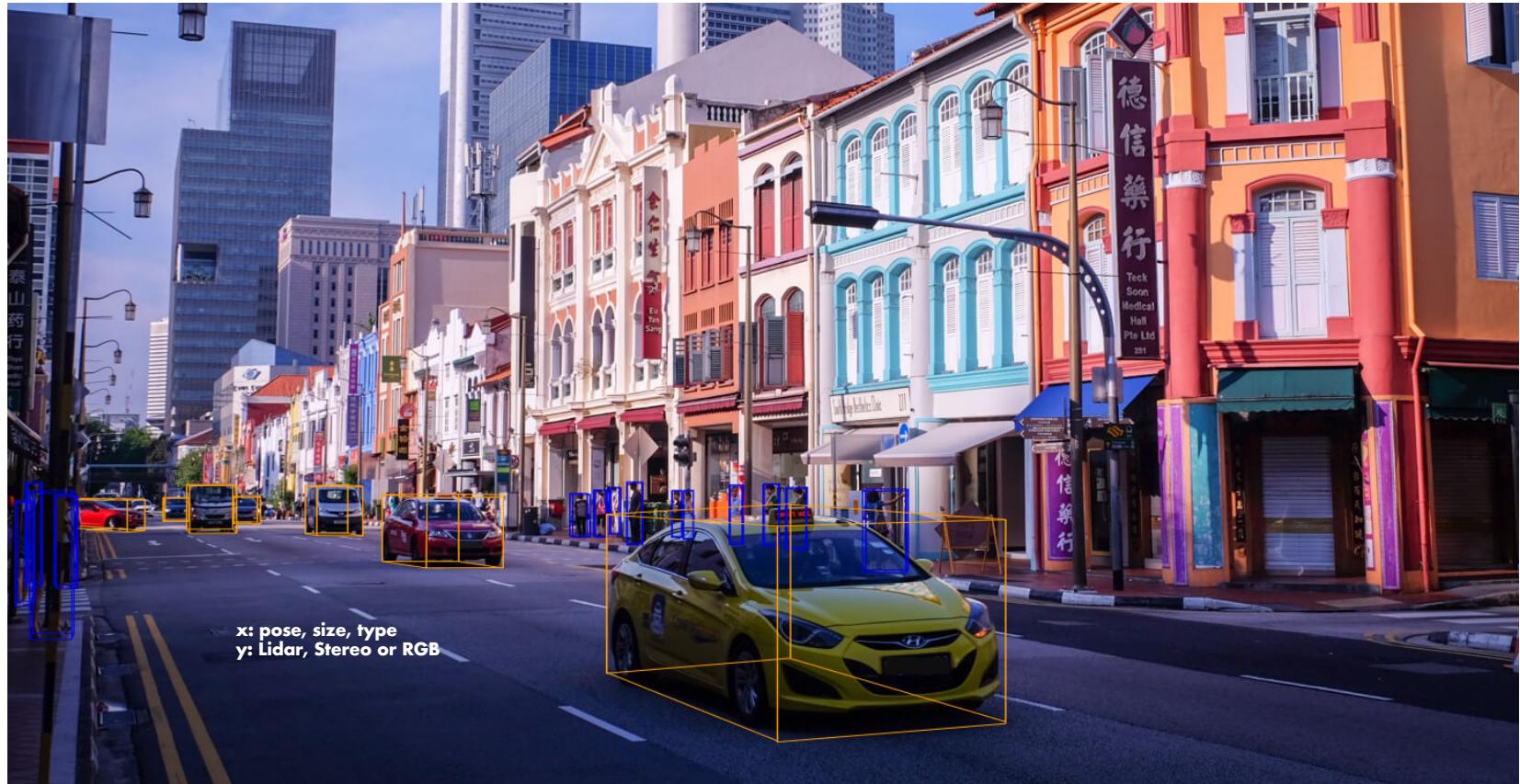
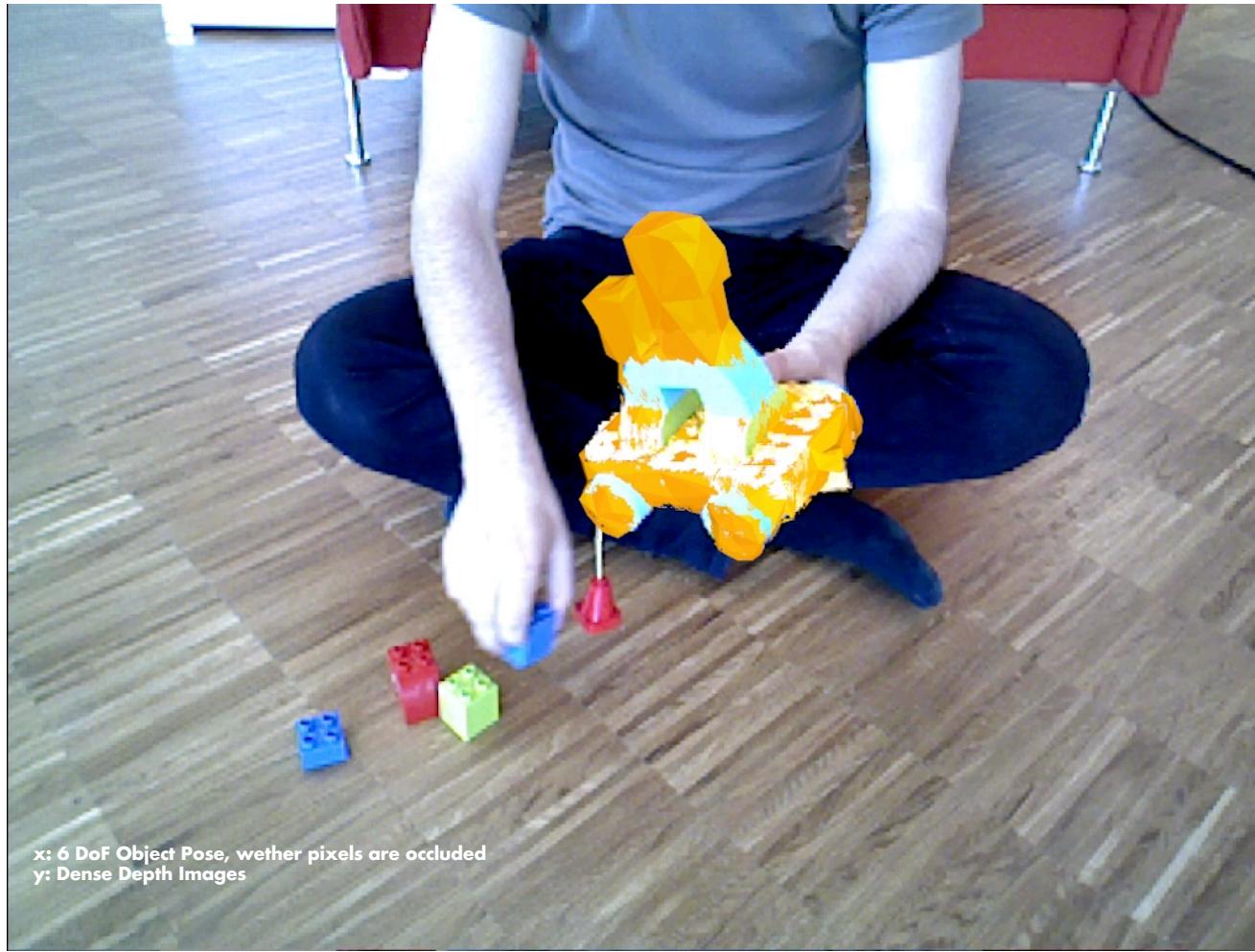


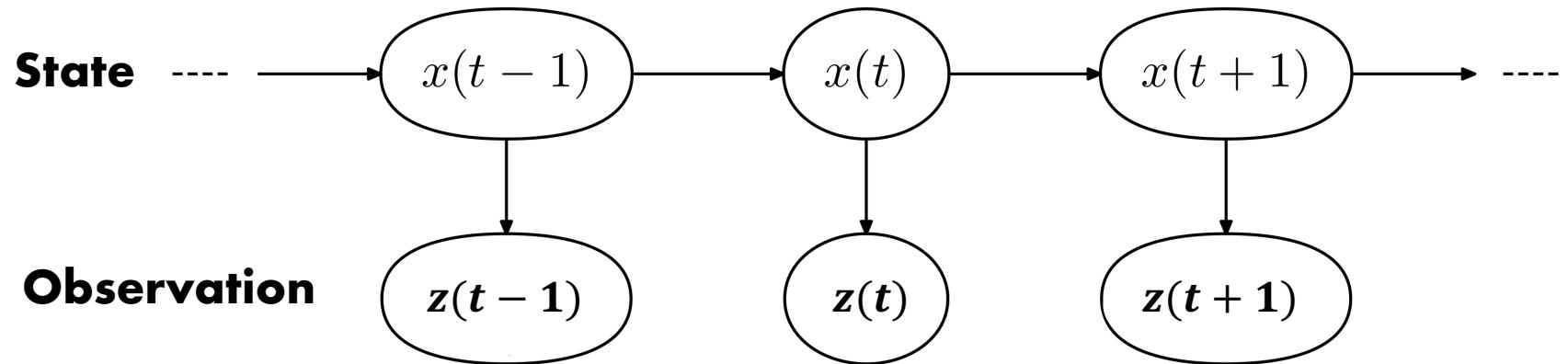
Image adapted from NuScenes by Motional. nuscenes.org

Representations for Manipulation



Manuel Wüthrich et al. "Probabilistic Object Tracking using a Depth Camera", IROS 2013

Examples in Robotics: Generative Observation Model



State x



Point cloud y overplayed with estimated object pose x .



Generative Model

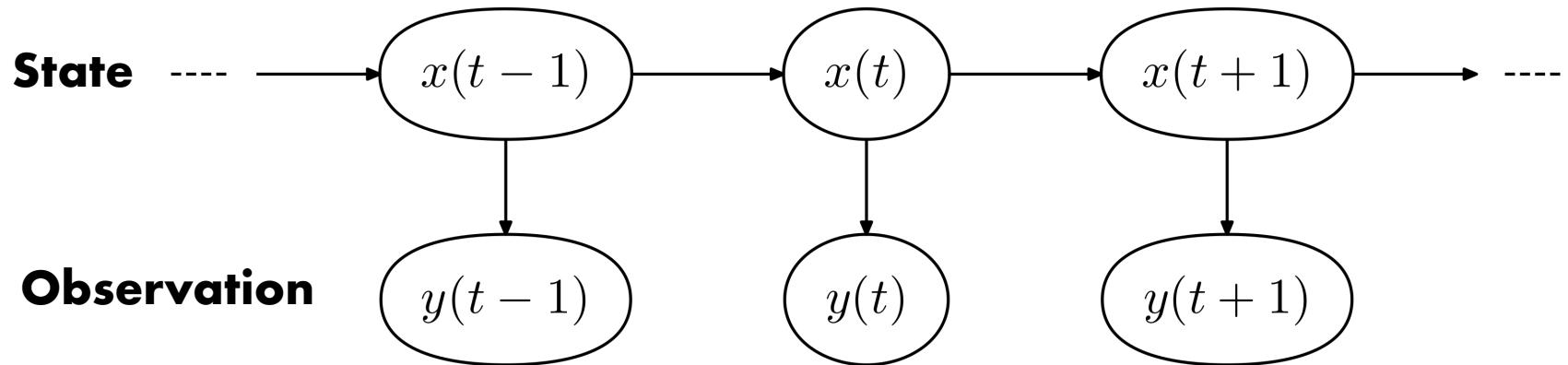
$$z = h(x)$$

Likelihood

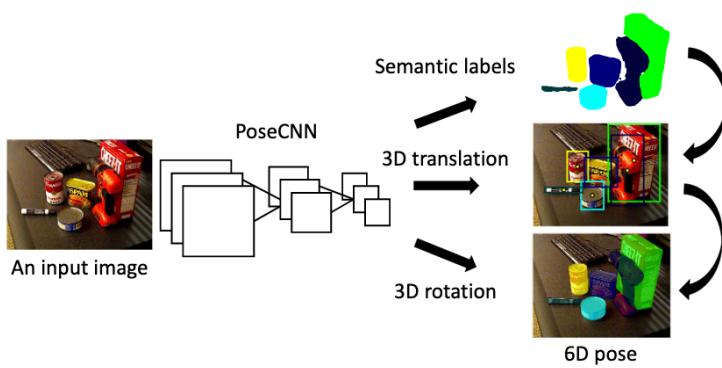
$$P(z|x)$$

Hidden Markov Model

Examples in Robotics: ‘Discriminative’ Observation Model



Discriminative Model



$$z = g(I)$$

Generative Model

$$z = x = f(x)$$

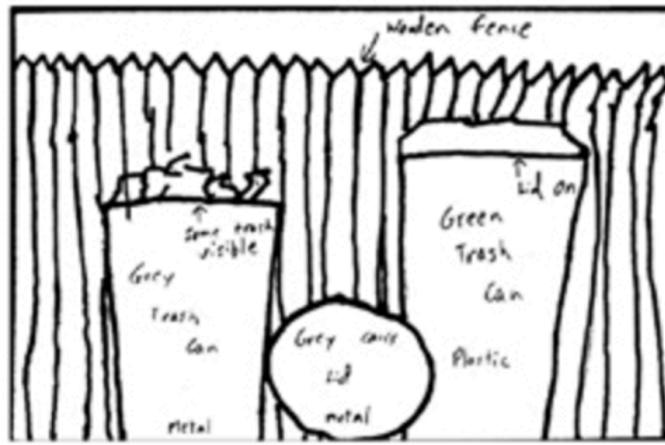
Yu Xiang et al. PoseCNN: A Convolutional Neural Network for 6D Object Pose Estimation in Cluttered Scenes. RSS 2018.

Representations in Computer Vision

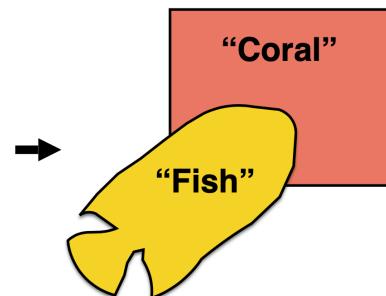
Observed image



Drawn from memory



X



Image

Compact mental
representation

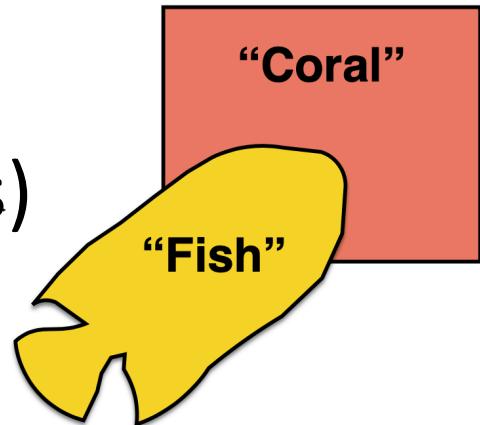
[Bartlett, 1932]
[Intraub & Richardson, 1989]

Input/Output/Intermediate Representations

Example from Advances in Computer Vision – MIT – 6.869/6.819

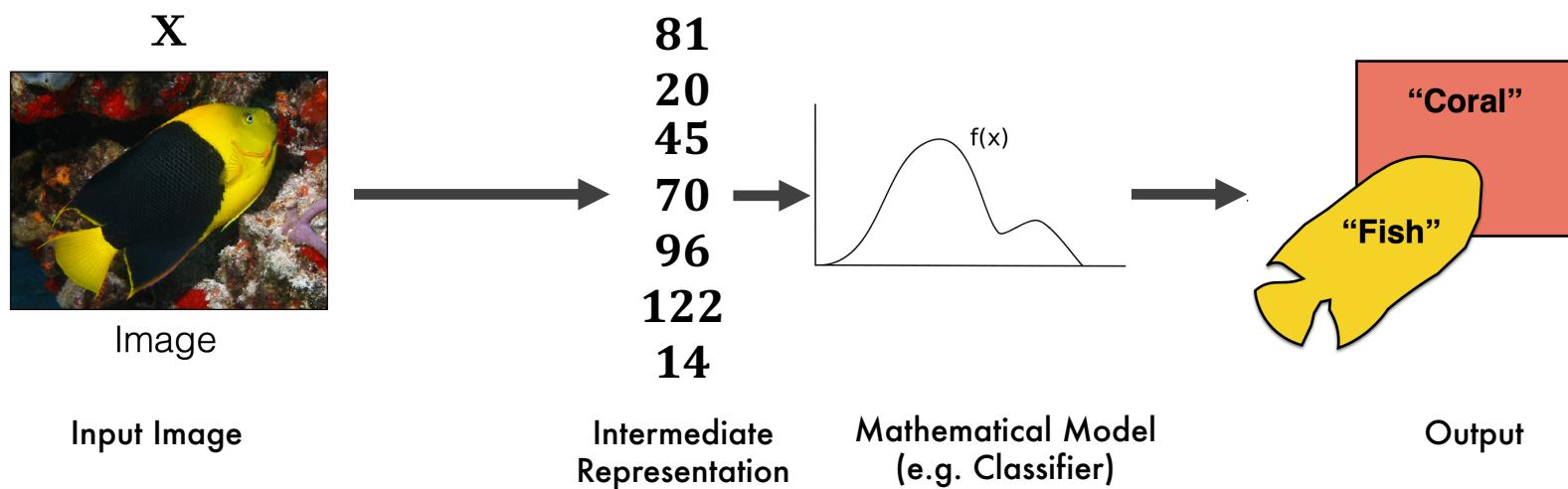
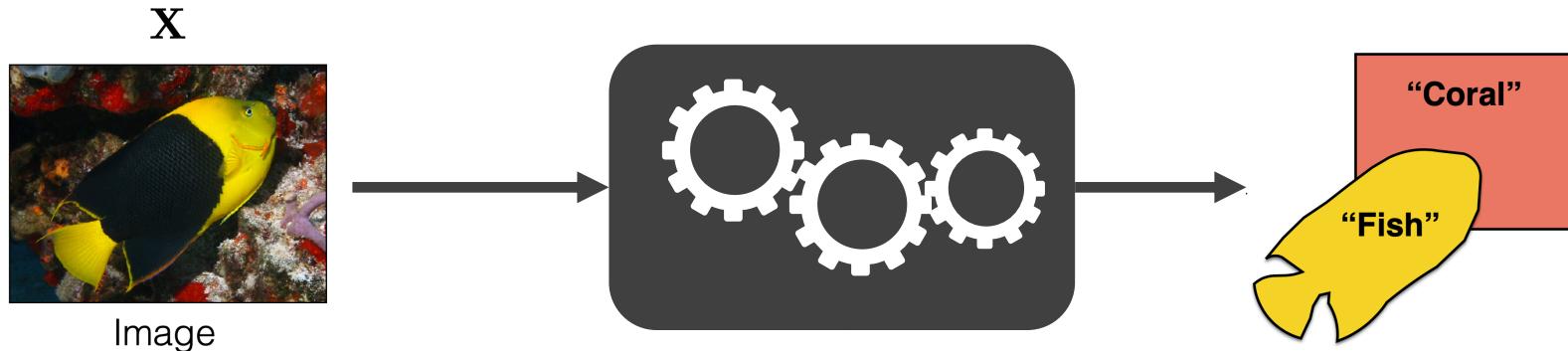
Requirements for Good Representations

- Compact (minimal)
- Explanatory (sufficient)
- Disentangled (independent factors)
- Interpretable
- Makes subsequent problem easier

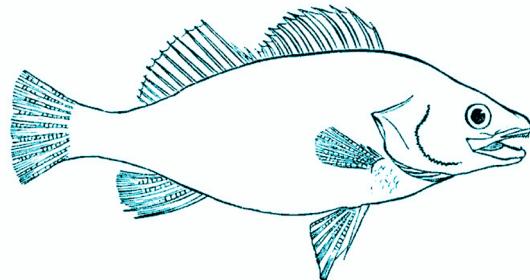


[See “Representation Learning”, Bengio 2013, for more commentary]

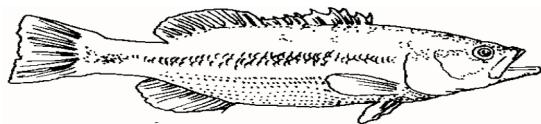
Typical CV Pipeline



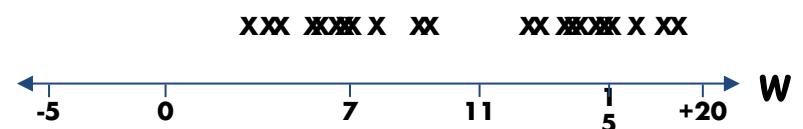
Example



~12 lbs

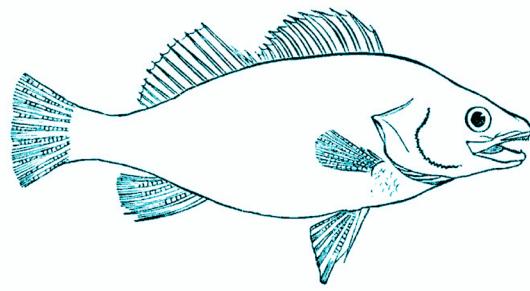


~8 lbs

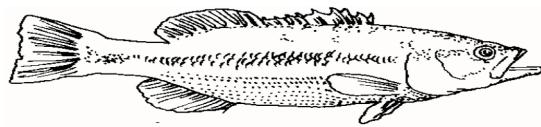


Example from CS331B: Representation Learning in Computer Vision

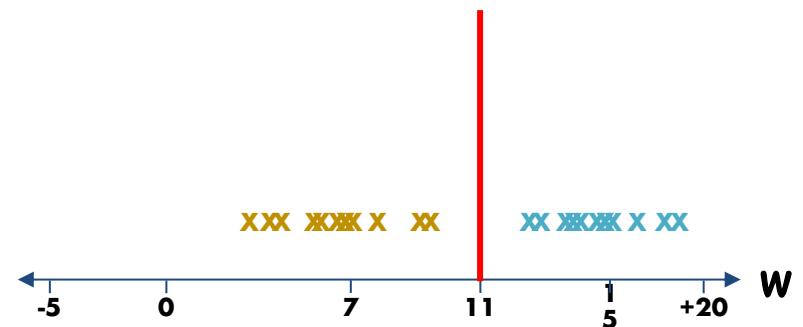
Example



~12 lbs

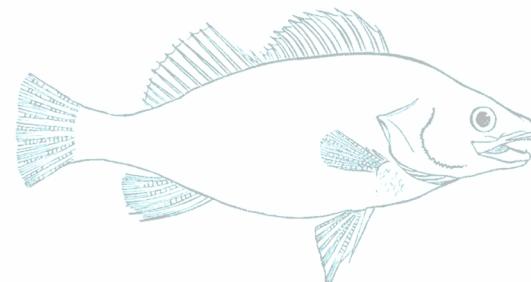


~8 lbs



Example from CS331B: Representation Learning in Computer Vision

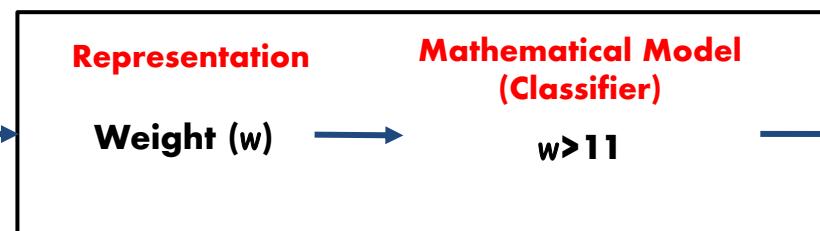
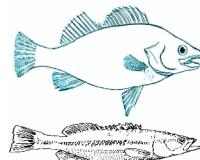
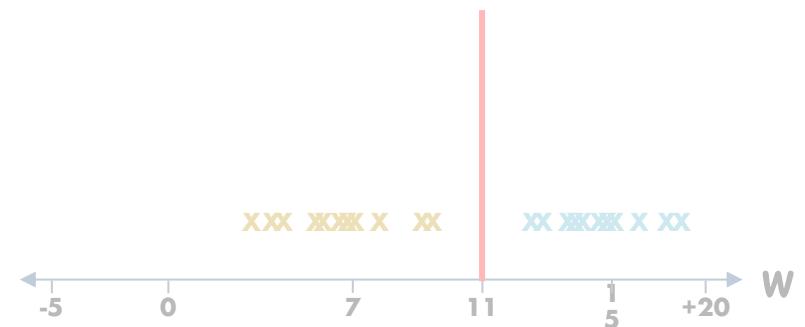
Example



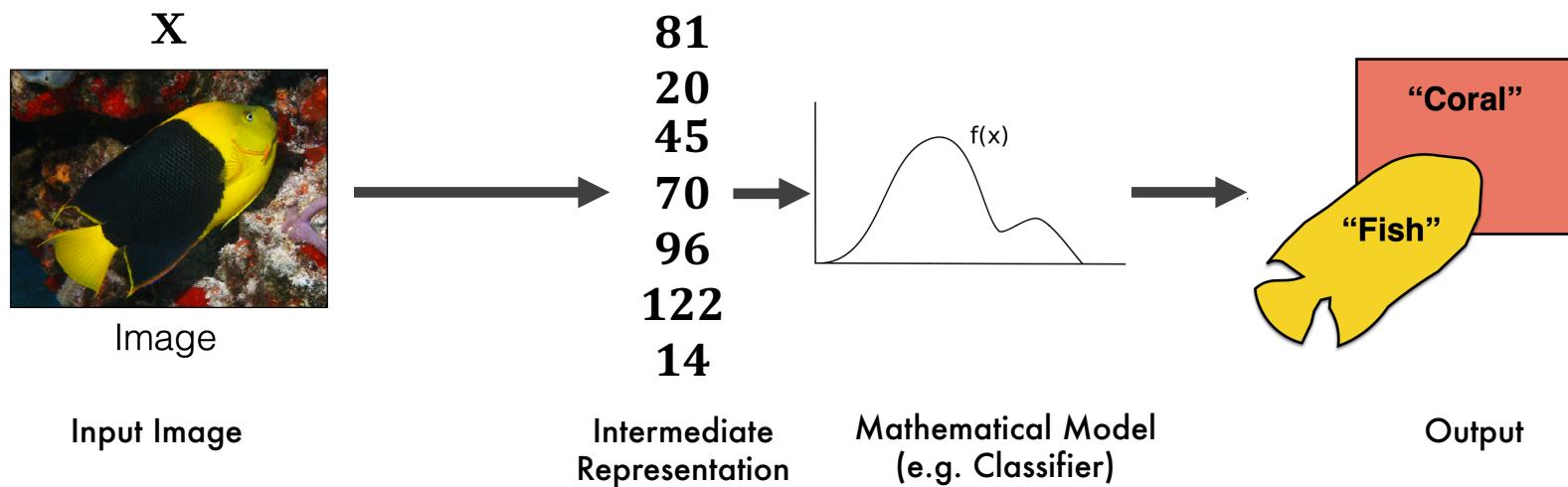
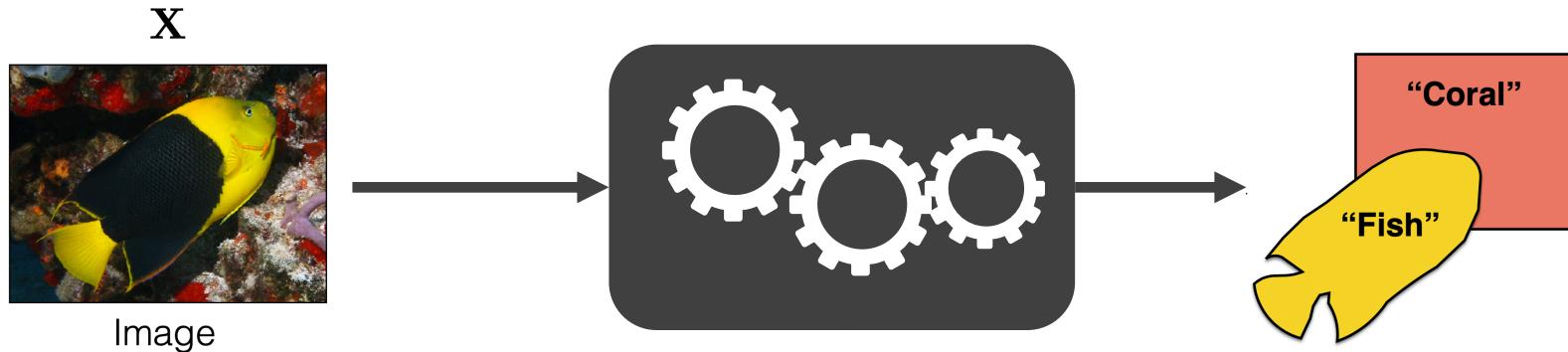
~12 lbs



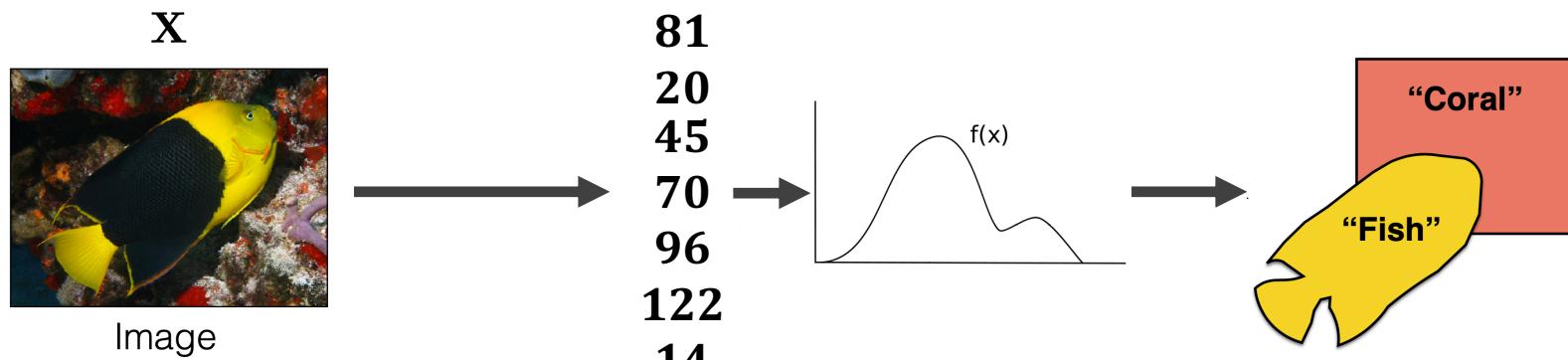
~8 lbs



Typical CV Pipeline



Traditional CV Pipeline

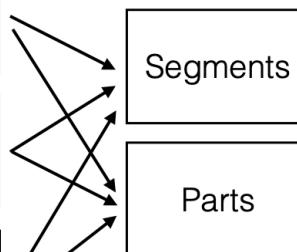
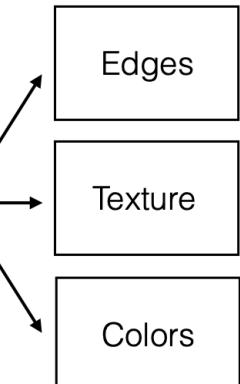


Input Image

Intermediate Representation

Mathematical Model
(e.g. Classifier)

Output

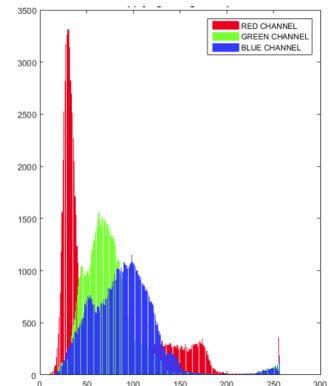


"clown fish"

Classifier

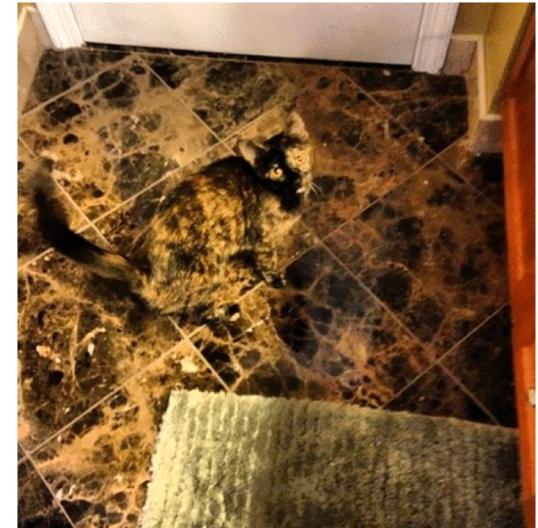
Example from Advances in Computer
Vision – MIT – 6.869/6.819

Represent these cats with a cat detector!



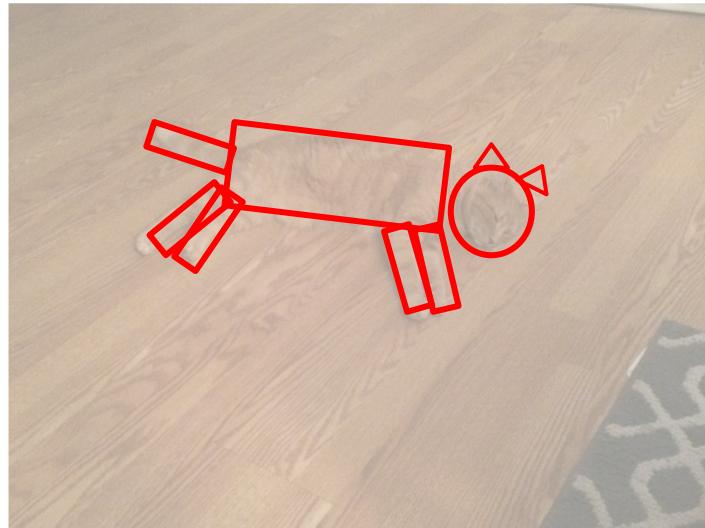
Example from CS331B: Representation Learning in Computer Vision

Represent these cats with a cat detector! (II)



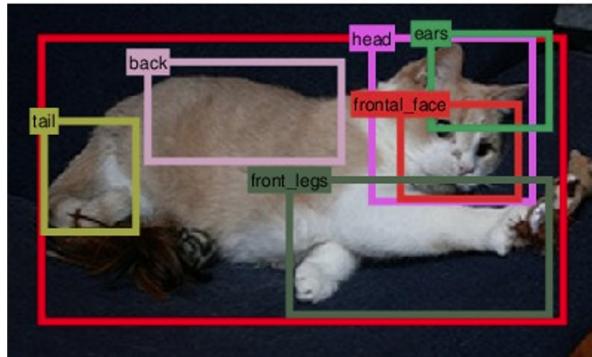
Example from CS331B: Representation Learning in Computer Vision

Represent these cats with a cat detector! (II)



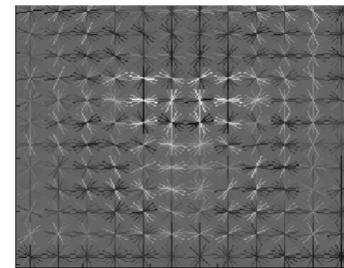
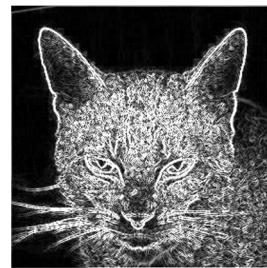
Example from CS331B: Representation Learning in Computer Vision

Represent these cats with a cat detector! (III)



Example from CS331B: Representation Learning in Computer Vision

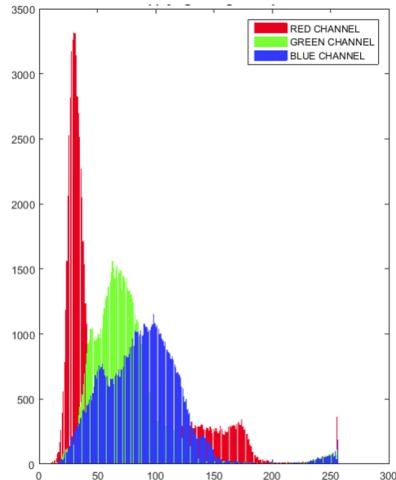
Represent these cats with a cat detector! (IV)



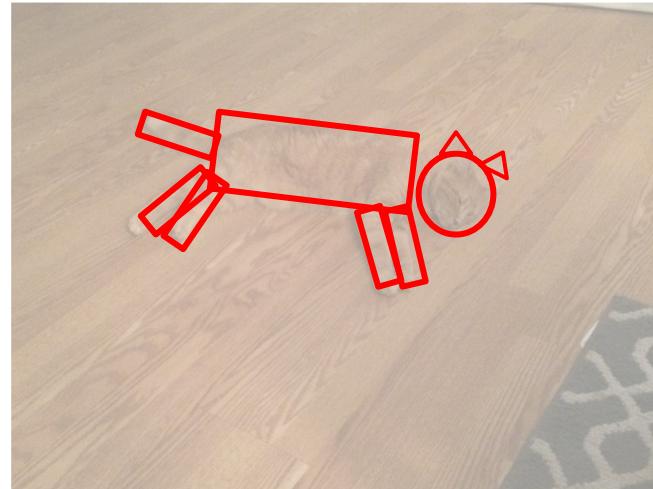
Example from CS331B: Representation Learning in Computer Vision

Summary of Traditional Components

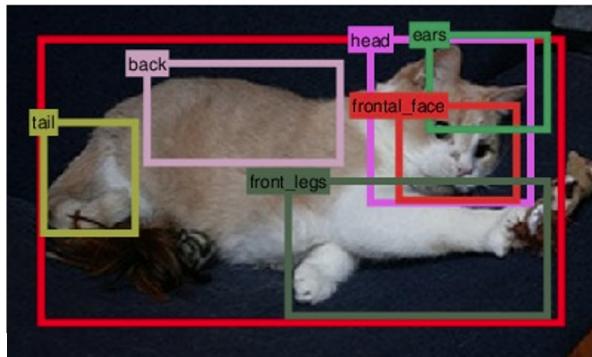
Color
Histograms



Model
based
Shapes

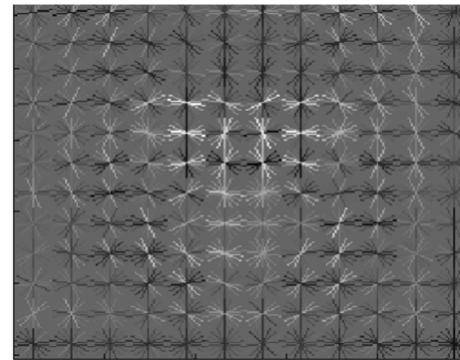


Deformable
Part based
Models
(DPM)



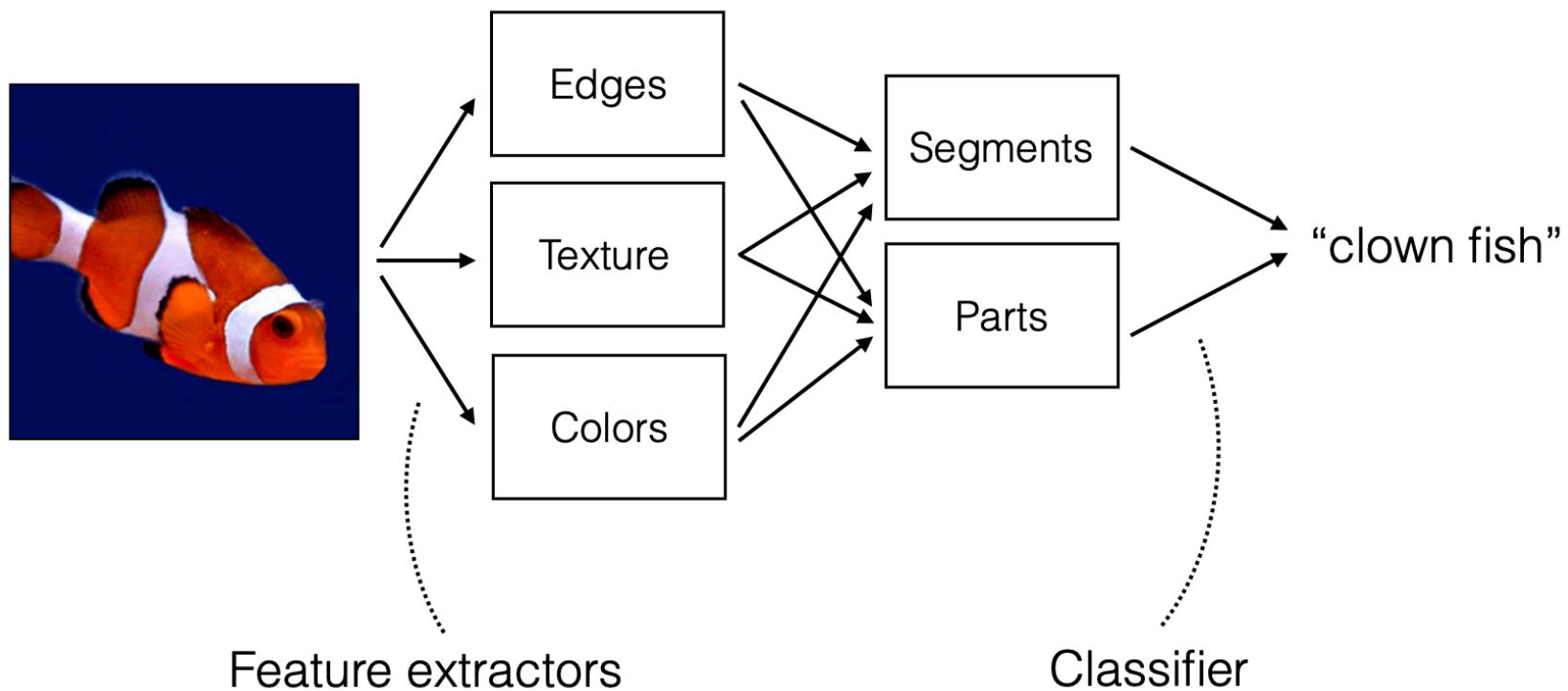
Felzenszwalb et al. 2010.
Dalal and Triggs, 2005.
Beis and Lowe, 1997.

Histogram of
Gradients
(HOG)



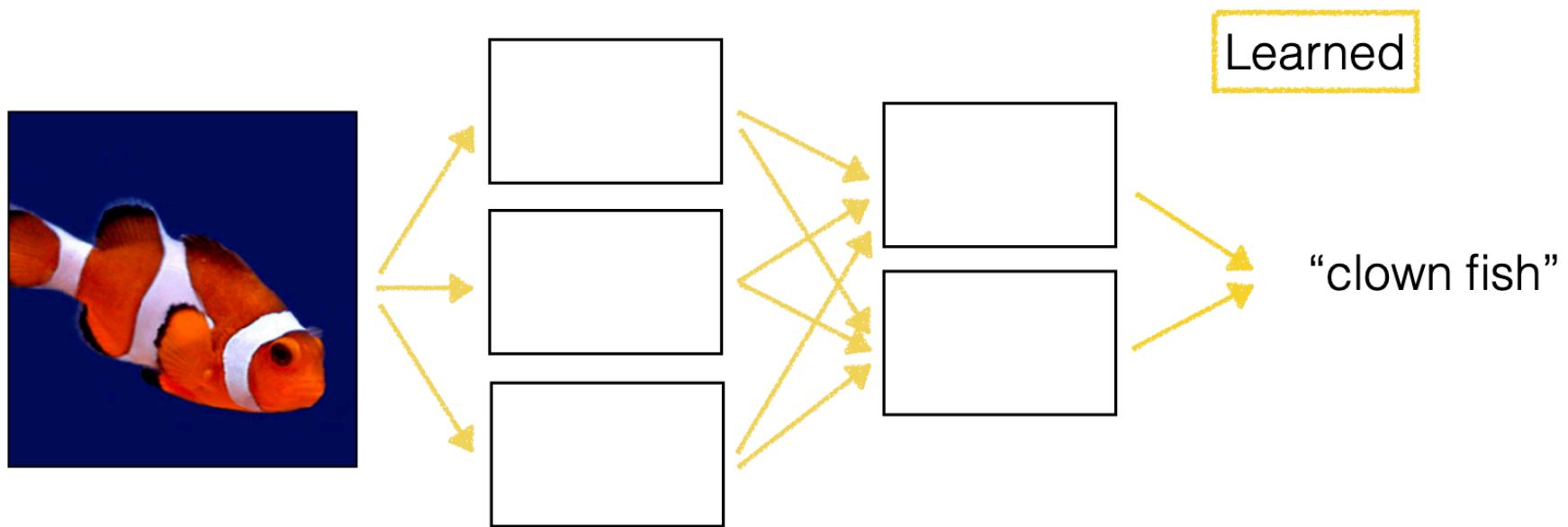
Example from CS331B: Representation Learning in Computer Vision

Traditional CV Pipeline



Example from Advances in Computer Vision – MIT – 6.869/6.819

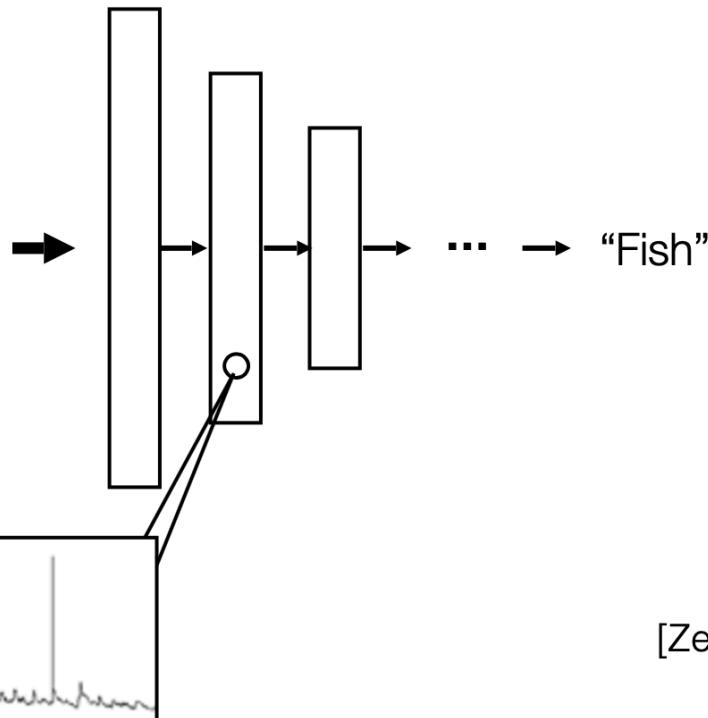
Traditional CV Pipeline



Example from Advances in Computer Vision – MIT – 6.869/6.819

How do you interpret what the network has learned?

Deep Net “Electrophysiology”



[Zeiler & Fergus, ECCV 2014]
[Zhou et al., ICLR 2015]

Example from Advances in Computer Vision – MIT – 6.869/6.819

Visualizing and Understanding CNNs

[Zeiler and Fergus, 2014]

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

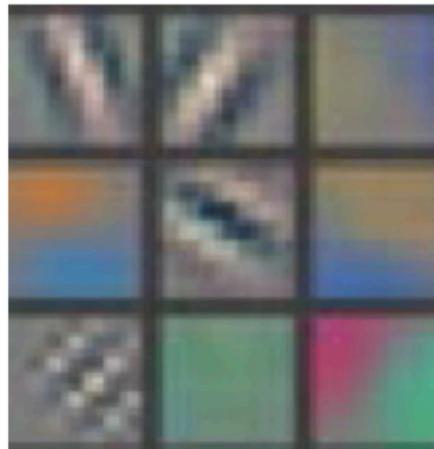
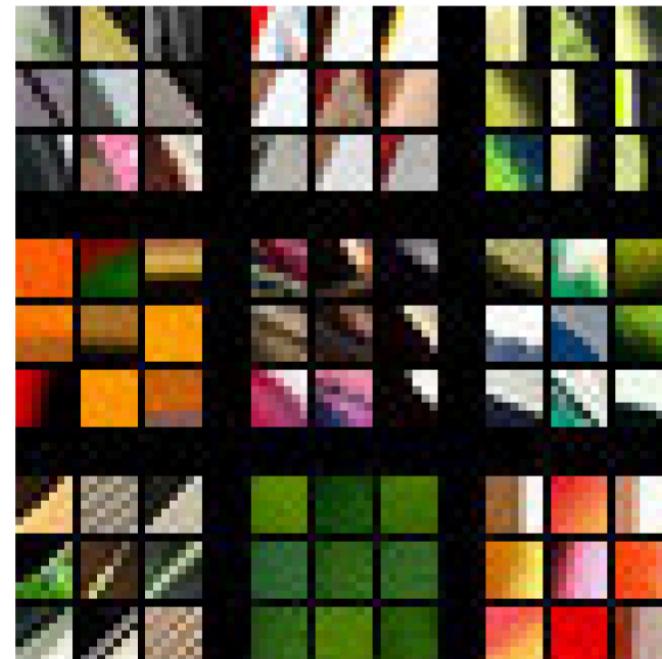


Image patches that activate each of the
layer 1 filters most strongly

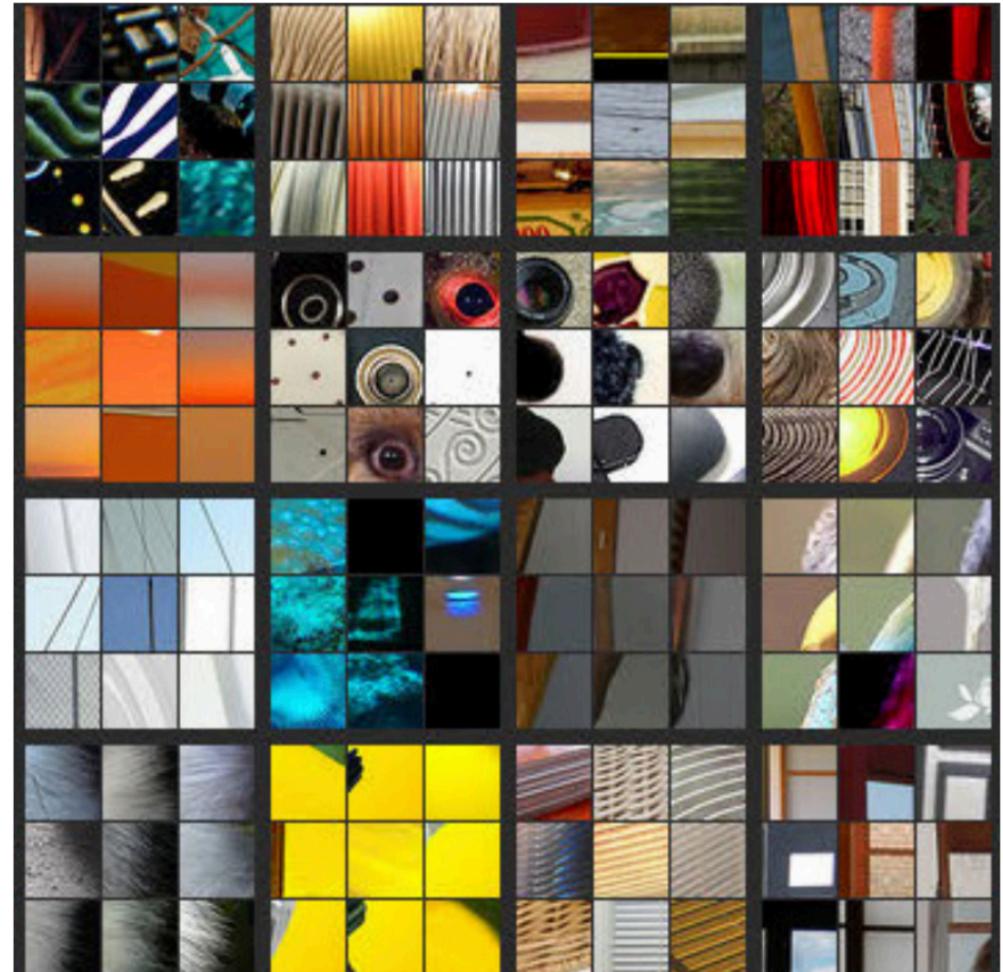


Example from Advances in Computer Vision – MIT – 6.869/6.819

Visualizing and Understanding CNNs

[Zeiler and Fergus, 2014]

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



Example from Advances in Computer Vision – MIT – 6.869/6.819

Visualizing and Understanding CNNs

[Zeiler and Fergus, 2014]



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

Example from Advances in Computer Vision – MIT – 6.869/6.819

Visualizing and Understanding CNNs

[Zeiler and Fergus, 2014]

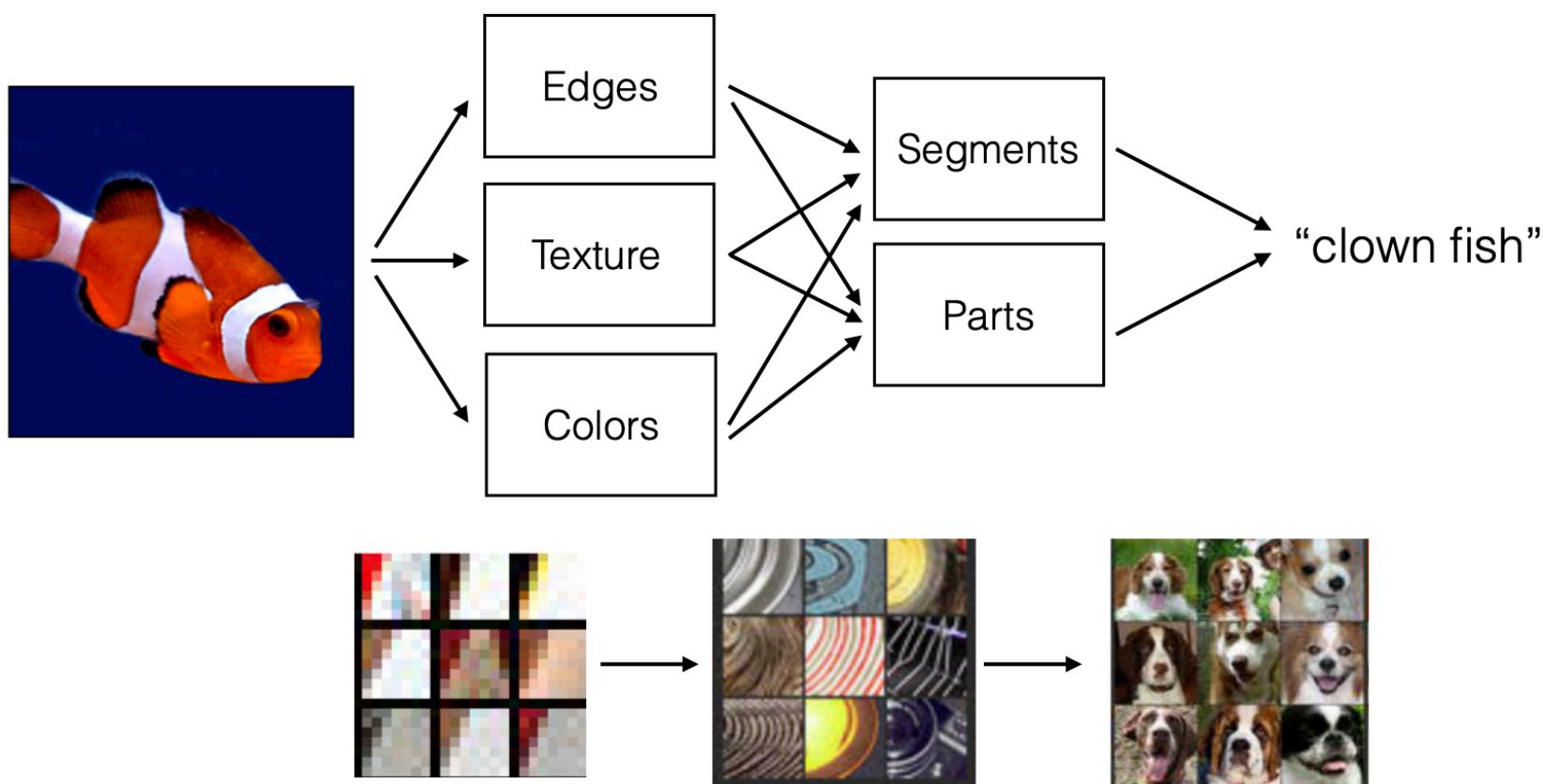


Image patches that activate
each of the **layer 5** neurons
most strongly

Example from Advances in Computer Vision – MIT – 6.869/6.819

Visualizing and Understanding CNNs

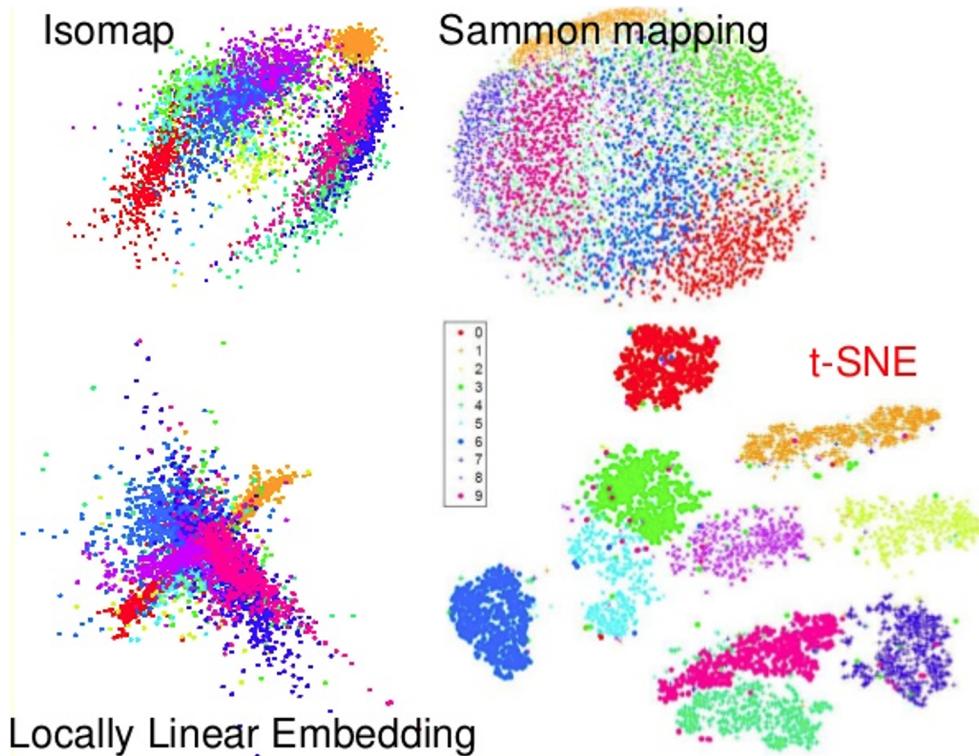
CNNs *learned* the classical visual recognition pipeline!



Example from Advances in Computer Vision – MIT – 6.869/6.819

Understanding representations through low-dimensional embeddings

- 6000 MNIST Digit
 - tSNE
 - Isomap
 - Sammon M
 - LLE



Van der Maaten & Hinton. 2008

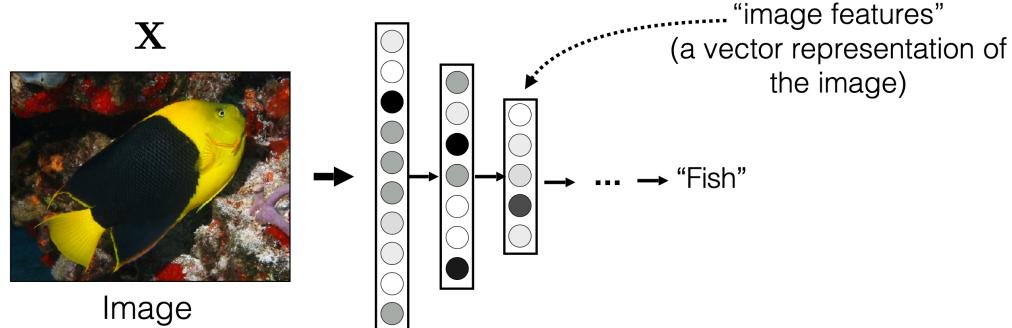
Understanding representations through low-dimensional embeddings

- tSNE

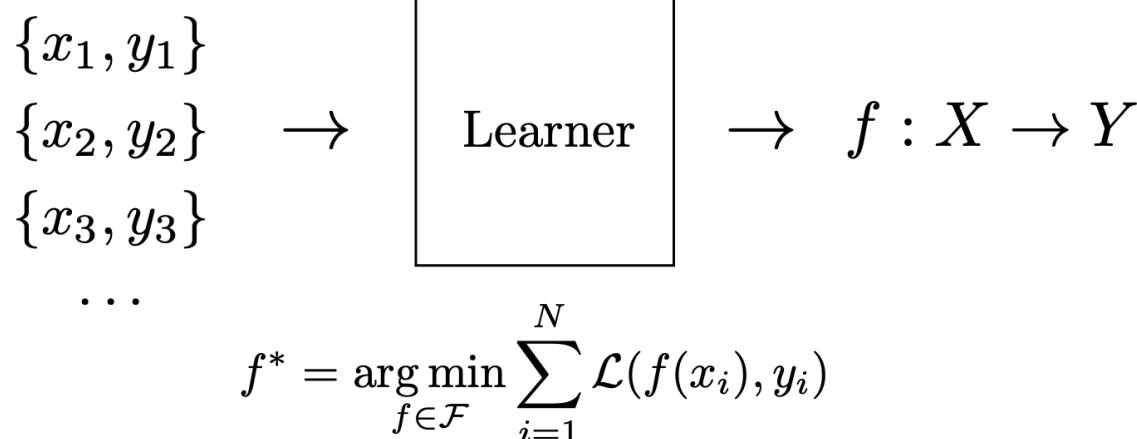


Van der Maaten & Hinton. 2008

How do you learn a representation?



Training data



Supervised Object Recognition

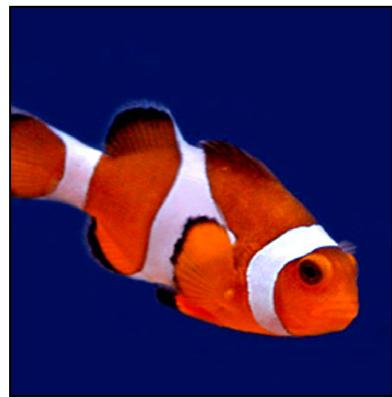
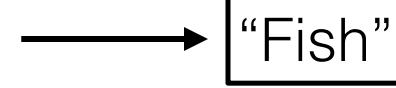


image X

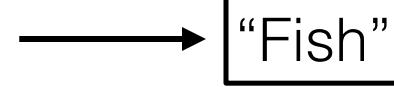


label Y

Supervised Object Recognition



image X

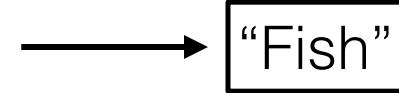


label Y

Supervised Object Recognition

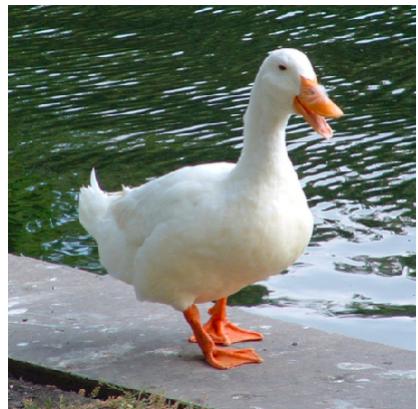


image X



label Y

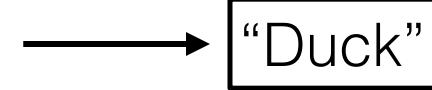
Supervised Object Recognition



:

A vertical ellipsis indicating multiple inputs.

image X



label Y

Learning in the wild



Time lapse of a baby playing with toys. Francis Vachon. YouTube

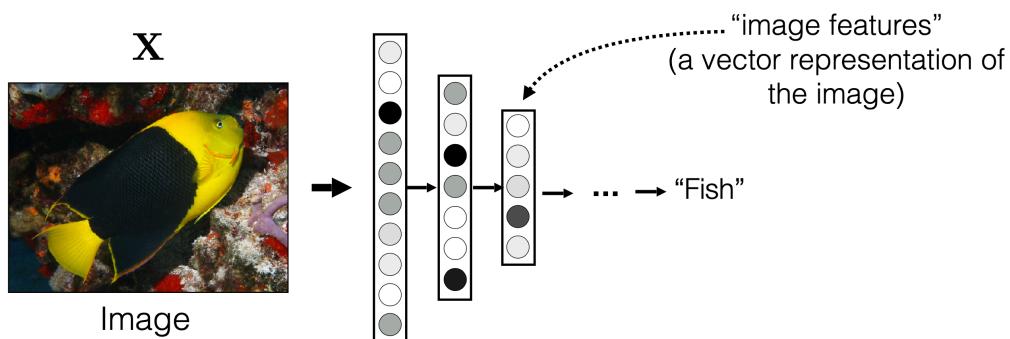
Supervised Computer Vision

- Informative
- Expensive
- Limited to teacher knowledge

Vision in Nature

- Cheap
- Noisy
- Harder to interpret

Learning without Labels



Training data

$$\{x_1, y_1\}$$

$$\{x_2, y_2\}$$

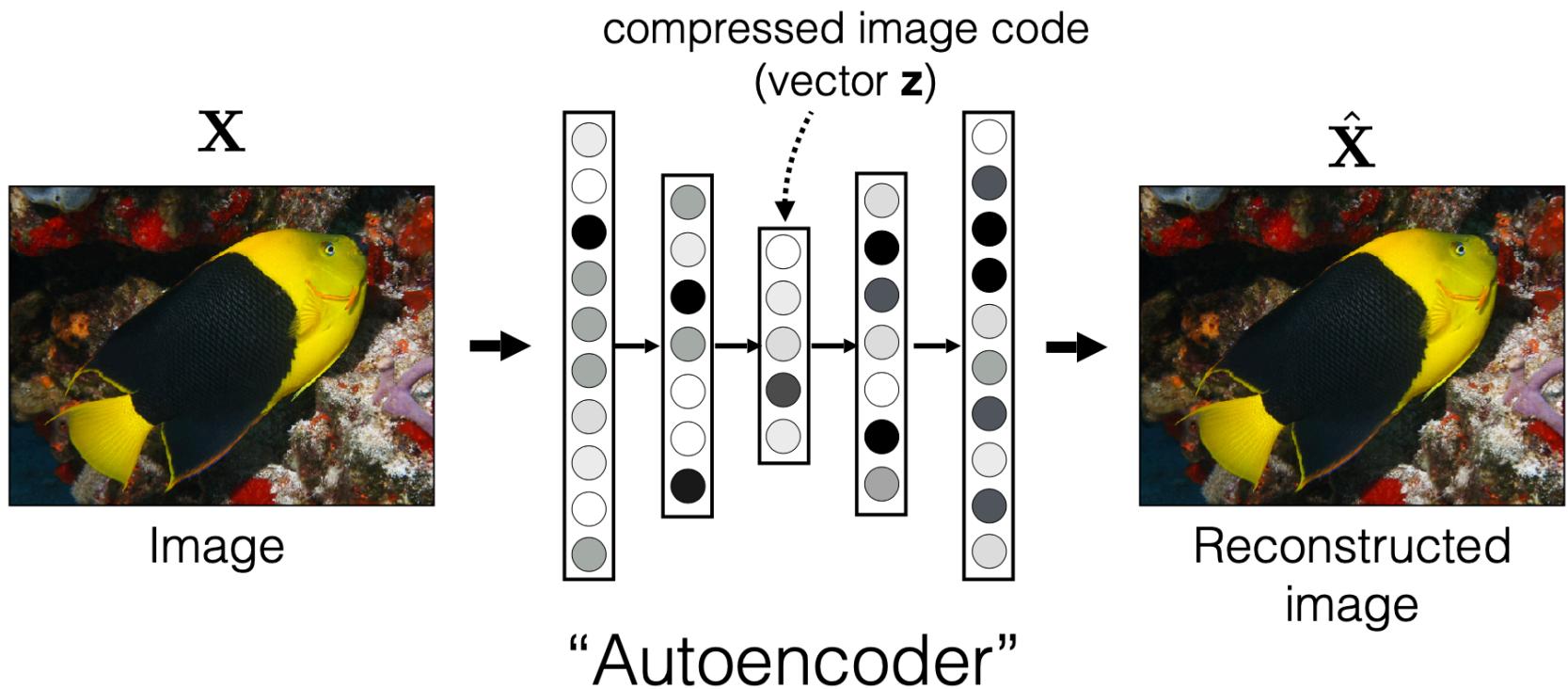
$$\{x_3, y_3\}$$

...

$$\rightarrow f : X \rightarrow Y$$

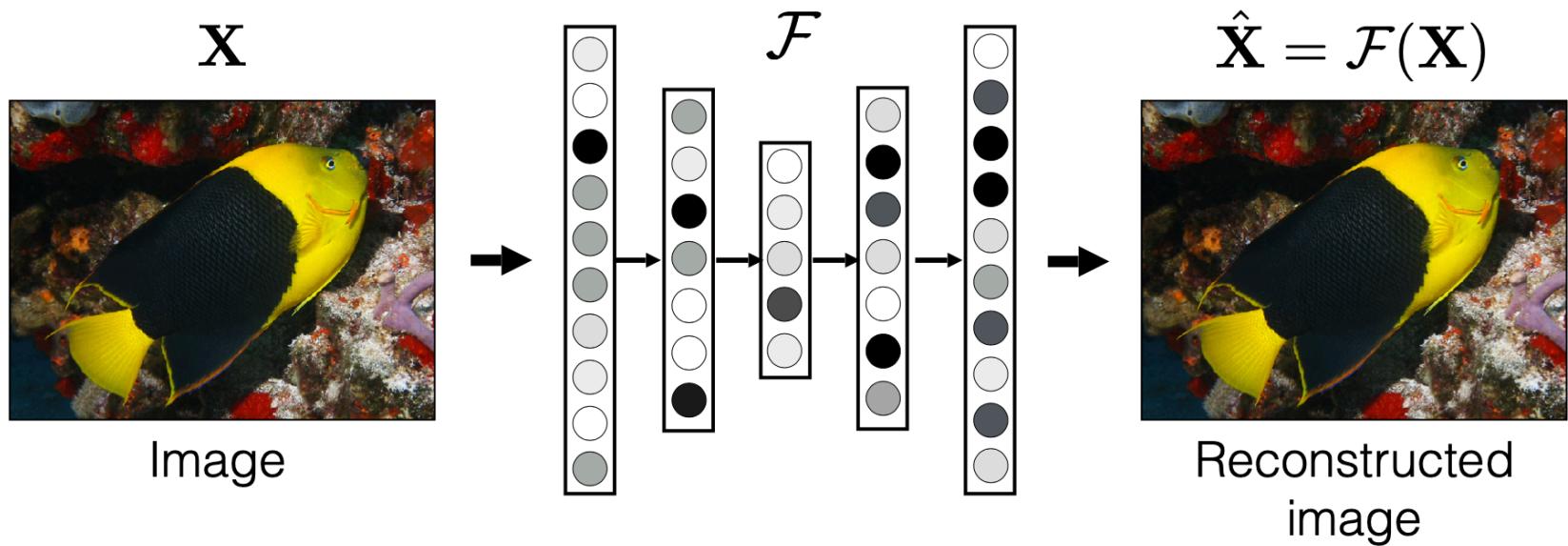
$$f^* = \arg \min_{f \in \mathcal{F}} \sum_{i=1}^n \mathcal{L}(f(x_i), y_i)$$

Unsupervised Representation Learning



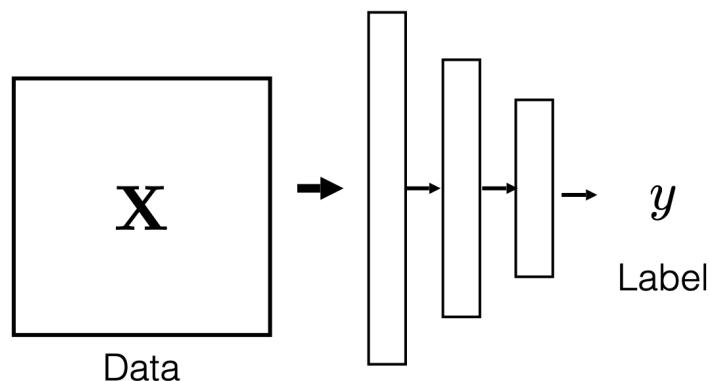
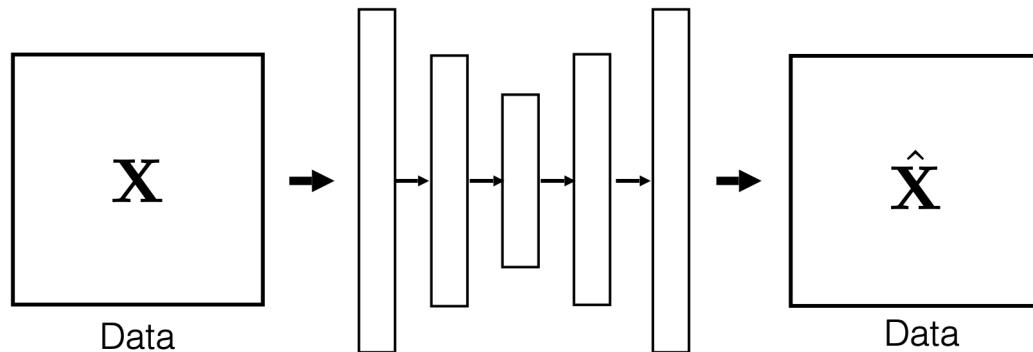
[e.g., Hinton & Salakhutdinov, Science 2006]

Autoencoder



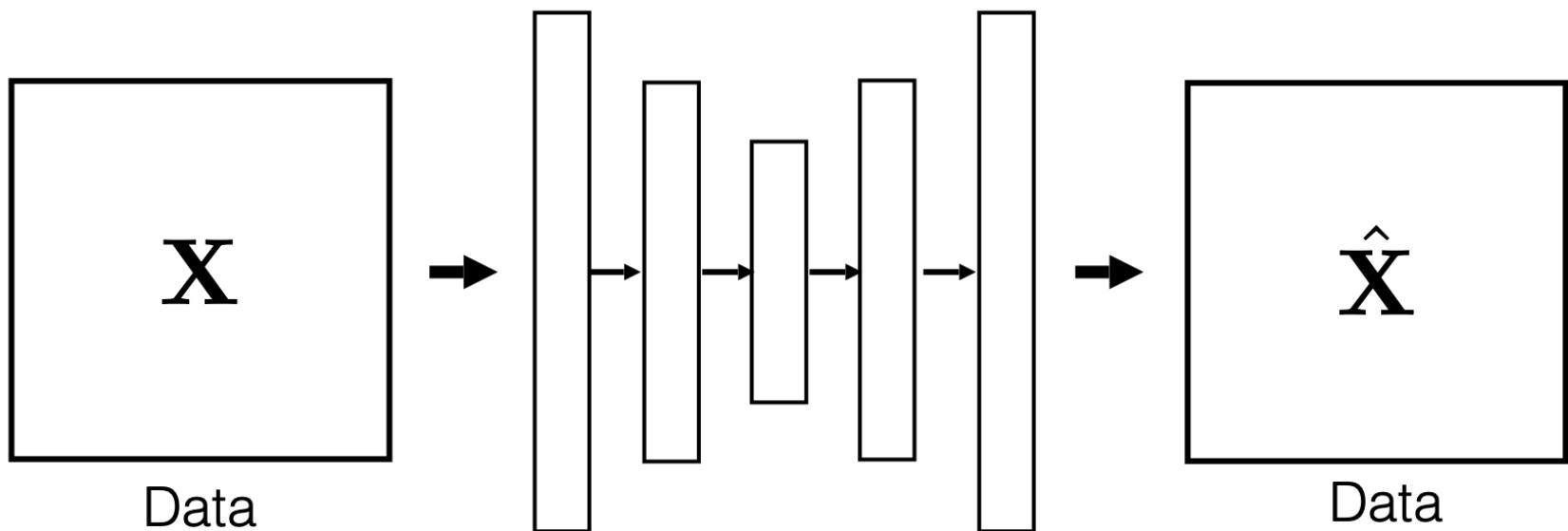
$$\arg \min_{\mathcal{F}} \mathbb{E}_{\mathbf{X}} [||\mathcal{F}(\mathbf{X}) - \mathbf{X}||]$$

Data Compression & Task Transfer



e.g., image classification

Self-Supervision



Representation Learning

Reinforcement Learning (Cherry)

Predicting a scalar reward given once in a while

A few bits for some samples

Supervised Learning (Chocolate Coat)

Predicting category or vector of scalars per input as provided by human labels.
10-10k bits per sample

Unsupervised / Self-Supervised Learning (Cake)

Predicting parts of observed input or predicting future observations or events
Millions of bits per sample



Visualisation Idea by Yann LeCun
Photo by [Kristina Paukshtite](#) from [Pexels](#)

Summary of what you learned today

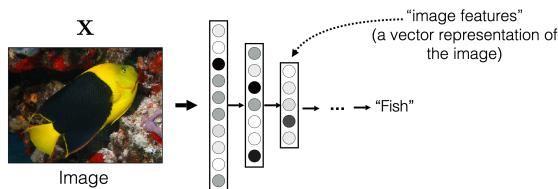
- **State:** Quantity that describes the most important aspect of a dynamical system at time t_i
- **Representation:** data format of input or output including a low-dimensional representation of sensor data

Summary of what you learned today

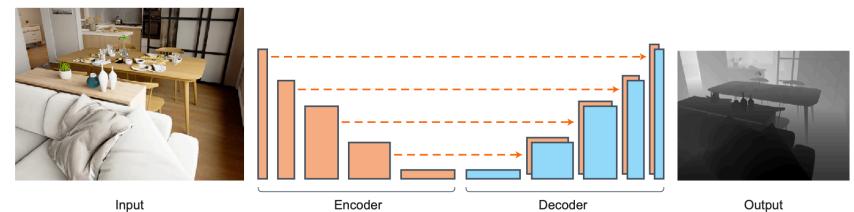
- Learned versus interpretable representations
- Visualize learned representations
- How to learn representations?
 - Supervised
 - Unsupervised
 - Self-supervised

What is in the next lectures?

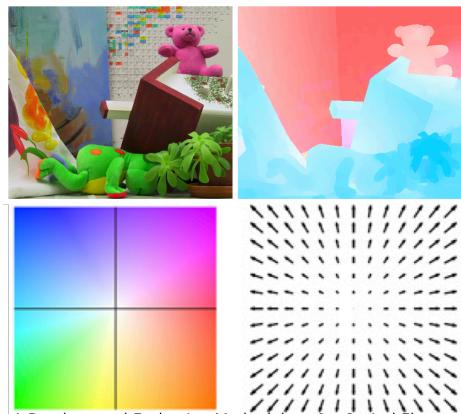
Representations & Representation Learning



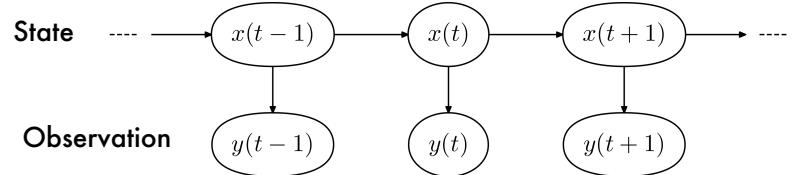
Monocular Depth Estimation



Optical & Scene Flow

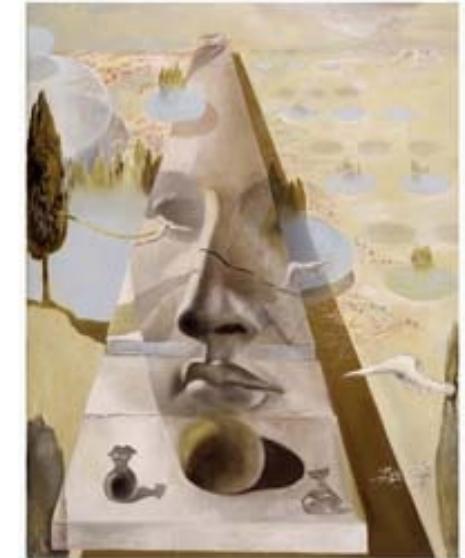


Optimal Estimation



CS231

Introduction to Computer Vision



Next lecture:

Monocular Depth Estimation
Low Level Tracking