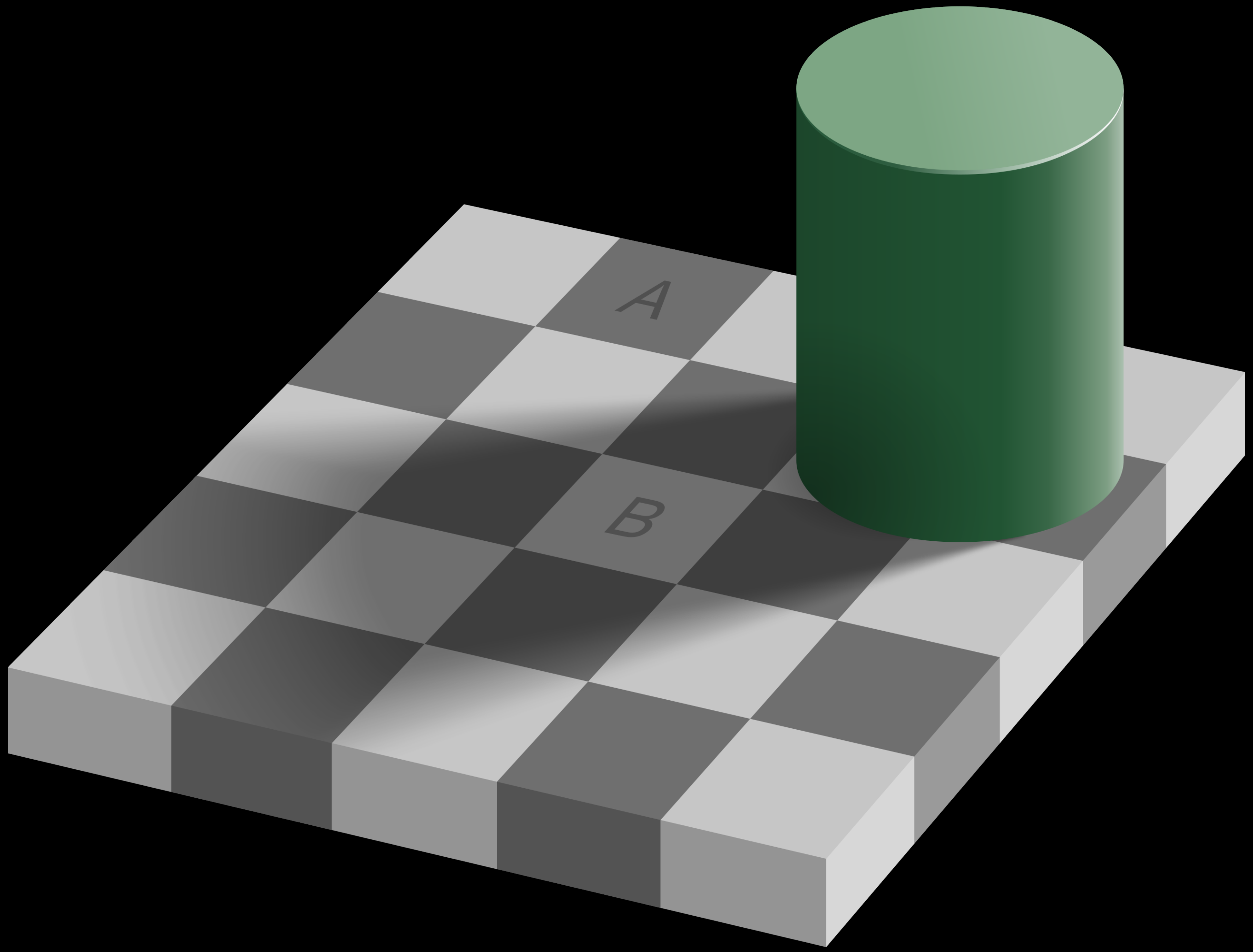
Advanced Computer Vision: Overview

MLMI17

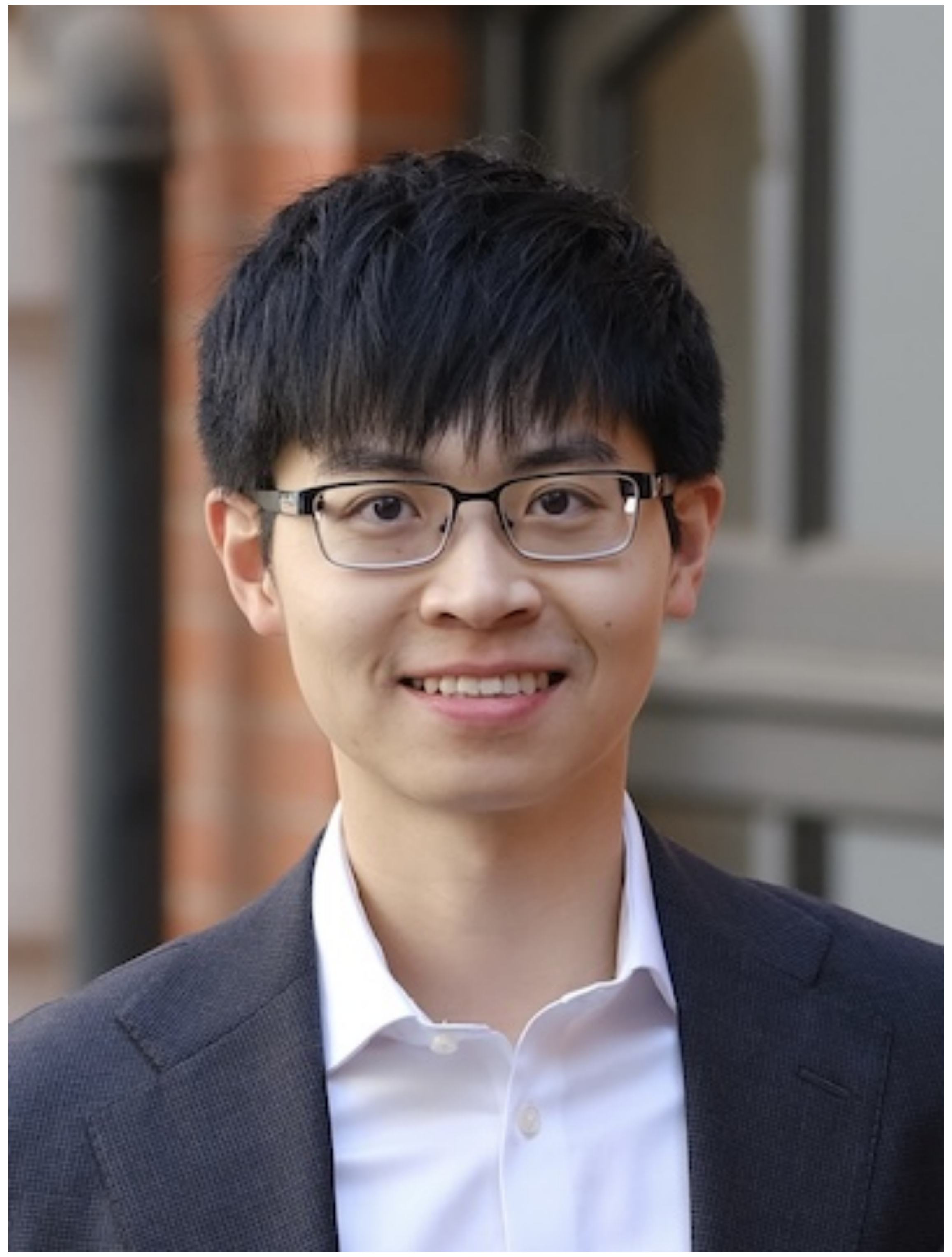
Ayush Tewari



Some slides adapted from https://visionbook.mit.edu/

UNIVERSITY OF CAMBRIDGE

Instructors



Elliott Wu

Ayush Tewari

Goal and Logistics

- Goal: Build background and intuition to understand and critique latest developments in computer vision.
- 4F12 is not a requirement
- Assessment
  - Two courseworks (individual): Deadline - 23rd Feb and 24th March
  - More on coursework later today
- Website: https://cambridgecvcourses.github.io/mlmi17/
- Office Hours: check website later this week
- Discussions: Use Moodle

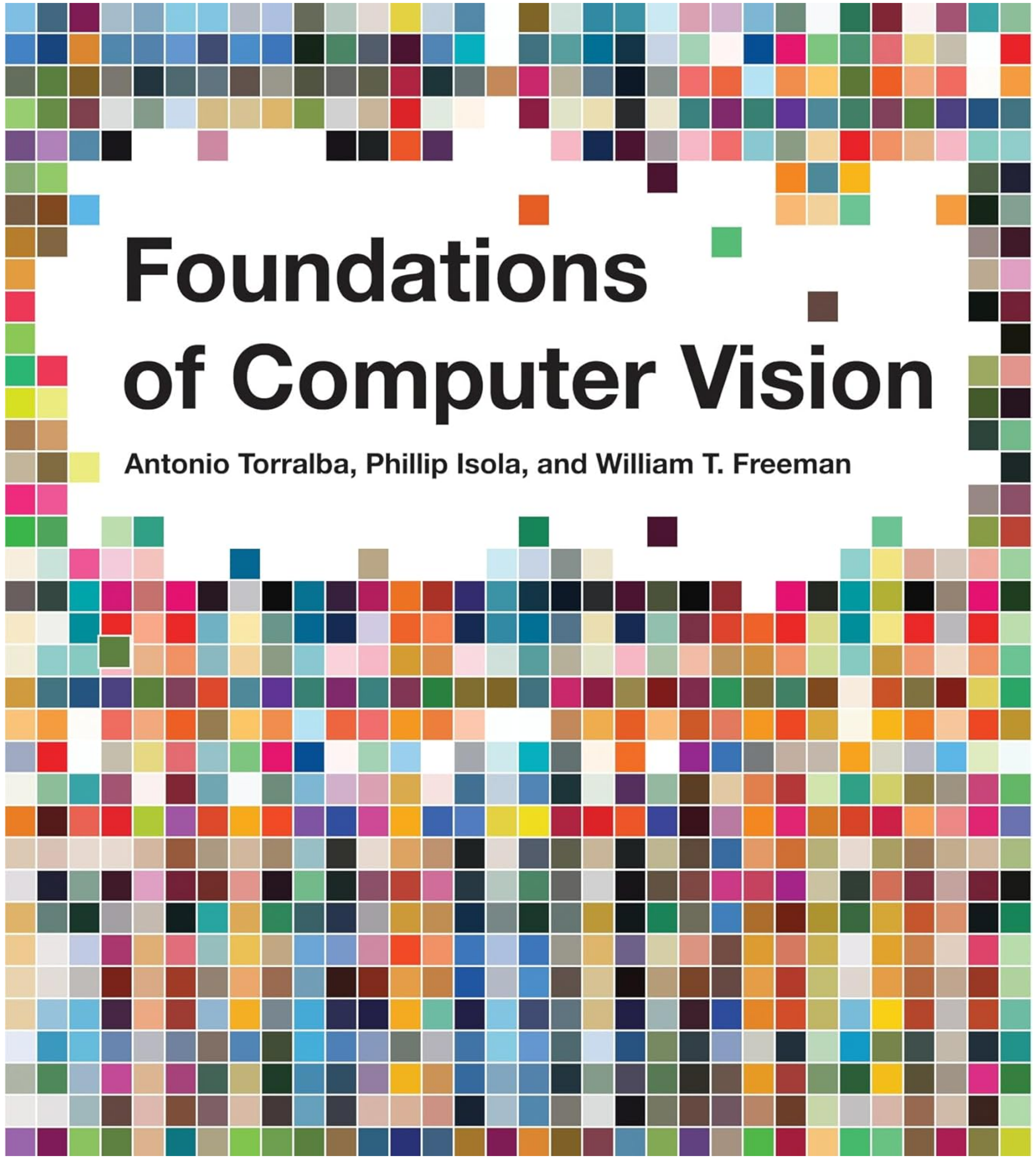
What is Computer Vision?

Graphics

Vision

"what is present in the world, and where it is"-David Marr

What is Computer Vision?


Foundations of Computer Vision
Antonio Torralba, Phillip Isola, and William T. Freeman

“The goal of a vision scientist is twofold: to understand how perception works and to build systems that can interpret the world around them using images…”

Computer Vision is Interdisciplinary

Graphics

Vision

"what is present in the world, and where it is"

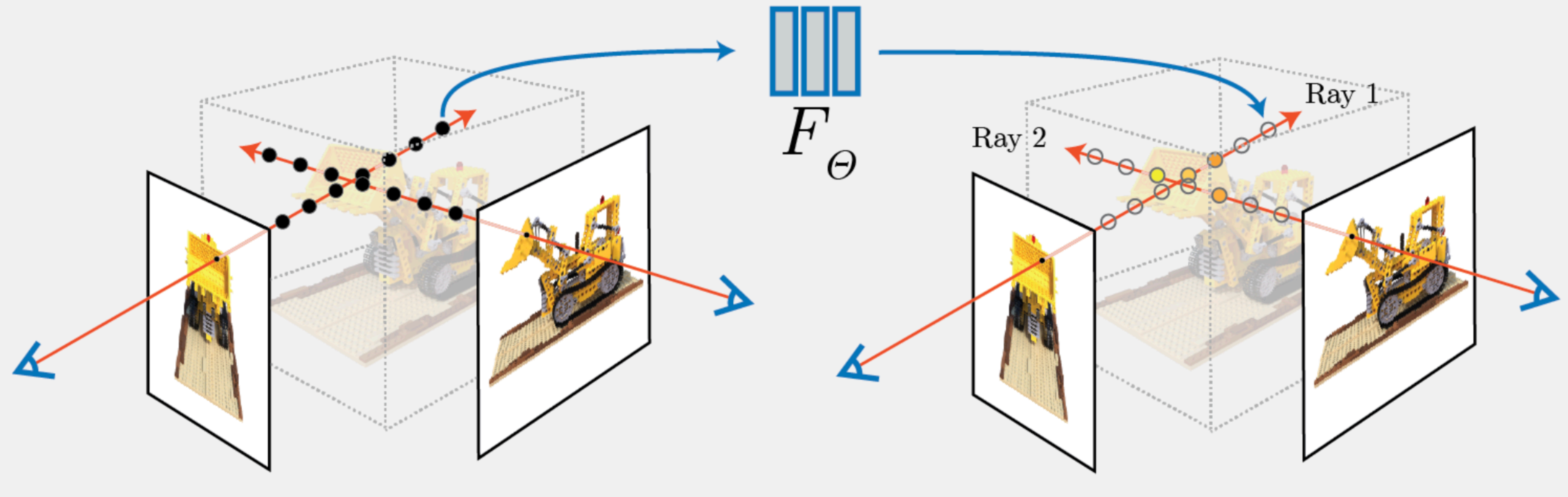
Computer Vision is Interdisciplinary - Graphics

Graphics

Vision

"what is present in the world, and where it is"

Computer Vision is Interdisciplinary



NeRF [Mildenhall et al., ECCV 2020 Best Paper]

Computer Vision is Interdisciplinary

Graphics

Vision

“what is present in the world, and where it is”

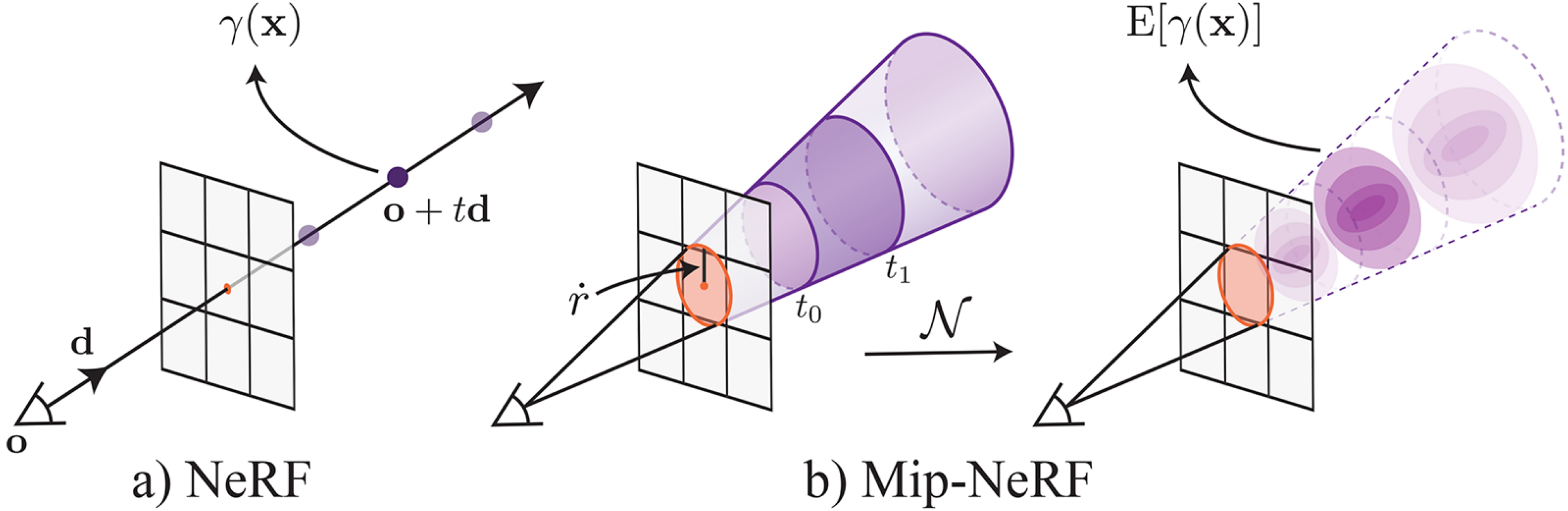
Computer Vision is Interdisciplinary - Signal Processing

Graphics

Vision

"what is present in the world, and where it is"-David Marr

Computer Vision is Interdisciplinary - Signal Processing



Mip-NeRF [Barron et al., ICCV 2021 Best Paper Honorable Mention]

Computer Vision is Interdisciplinary

Graphics

Vision

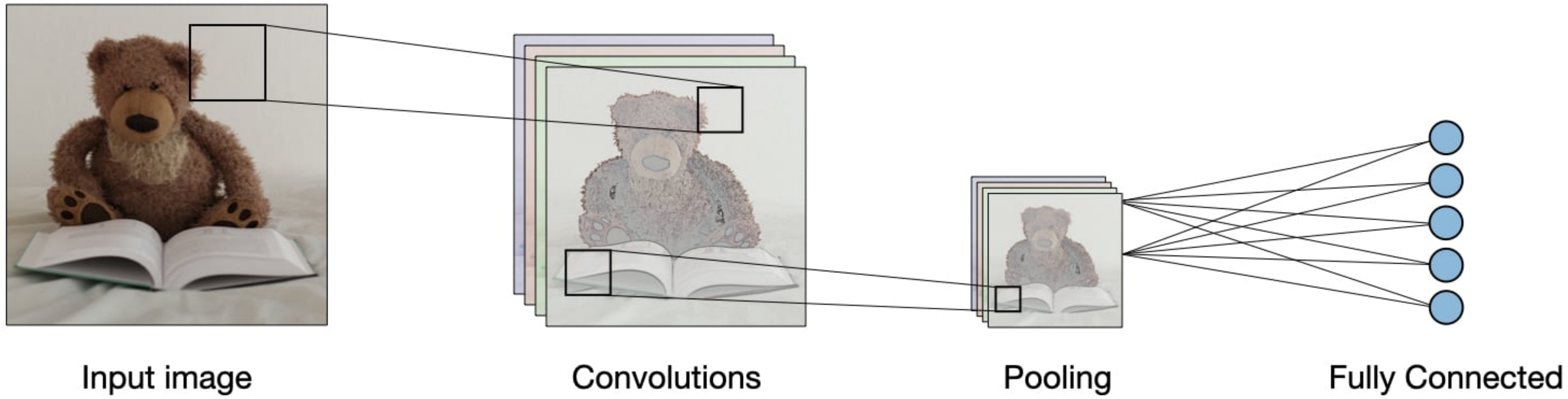
“what is present in the world, and where it is”

Computer Vision is Interdisciplinary - Machine Learning

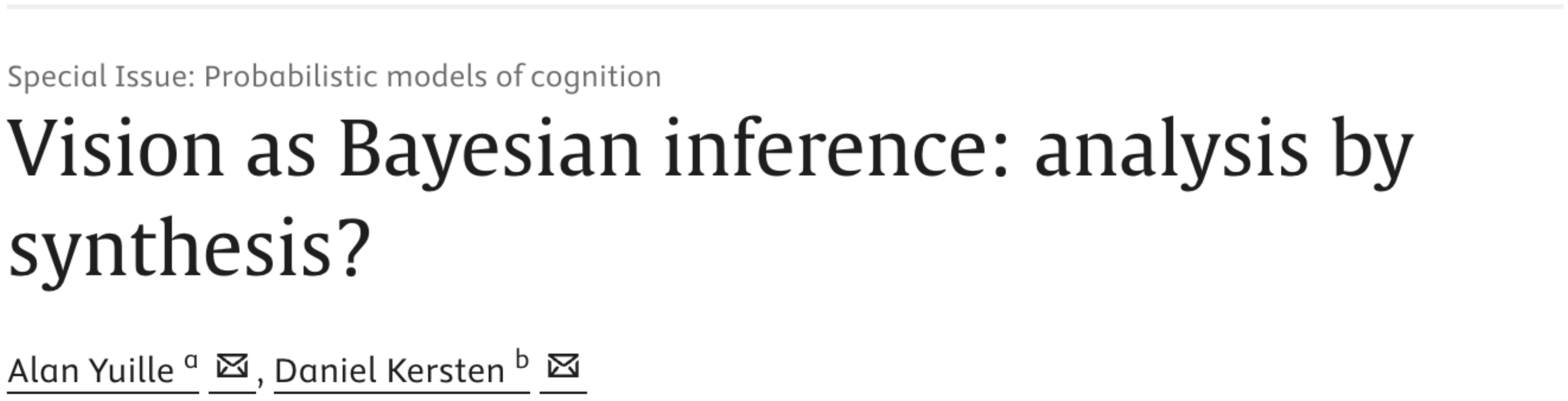
Vision

"what is present in the world, and where it is"

Computer Vision is Interdisciplinary - Machine Learning

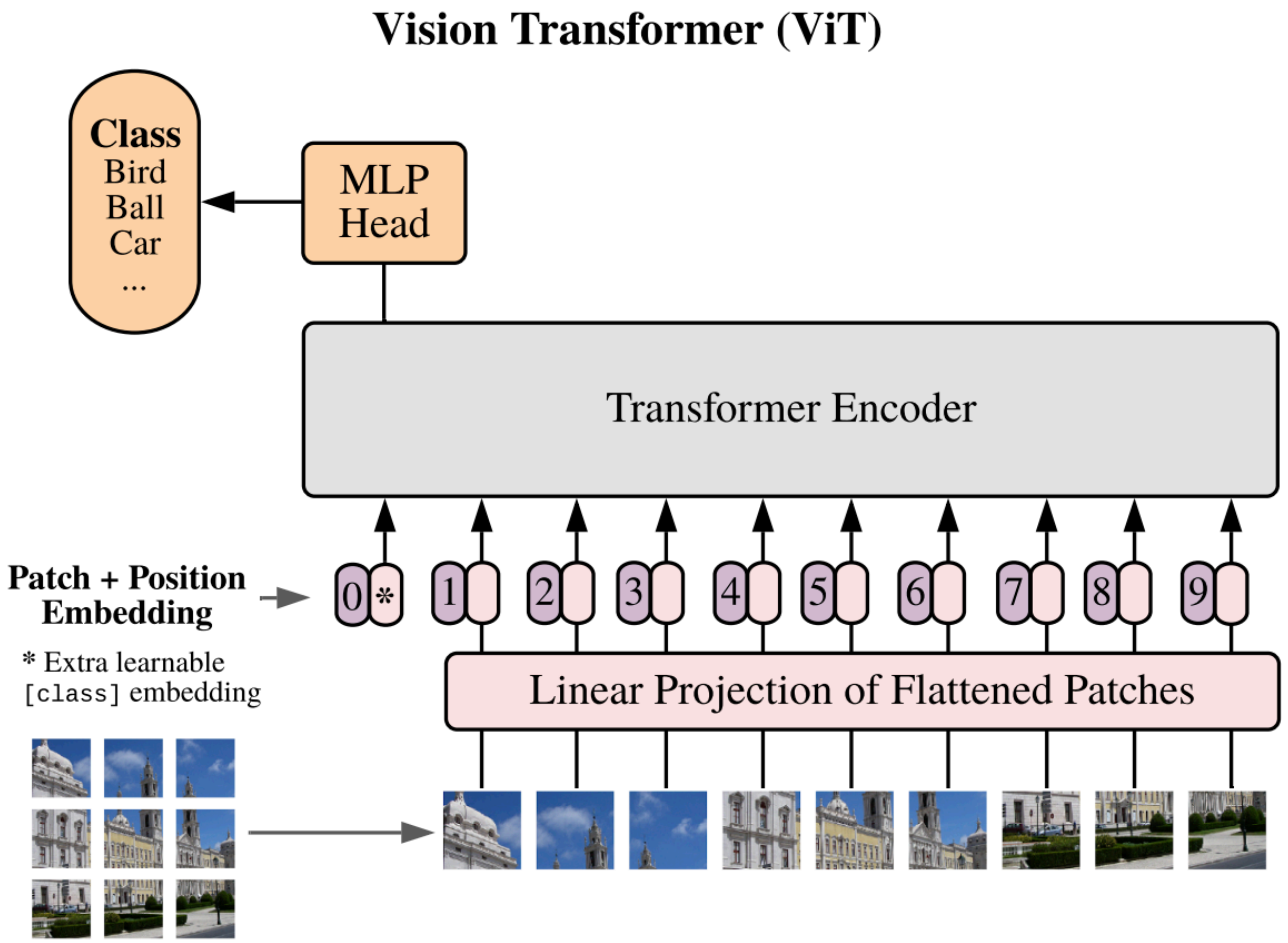


CNNs (https://stanford.edu/~shervine/teaching/cs-230/)


Special Issue: Probabilistic models of cognition

Vision as Bayesian inference: analysis by synthesis?

Alan Yuille [a] ✉, Daniel Kersten [b] ✉


Vision Transformer (ViT)
Class
Bird
Ball
Car
...
MLP Head
Transformer Encoder
Patch + Position Embedding
* Extra learnable [class] embedding
0
*
1
2
3
4
5
6
7
8
9
Linear Projection of Flattened Patches

ViT [Dosovitskiy et al., ICLR 2021]

Computer Vision is Interdisciplinary

Graphics

Vision

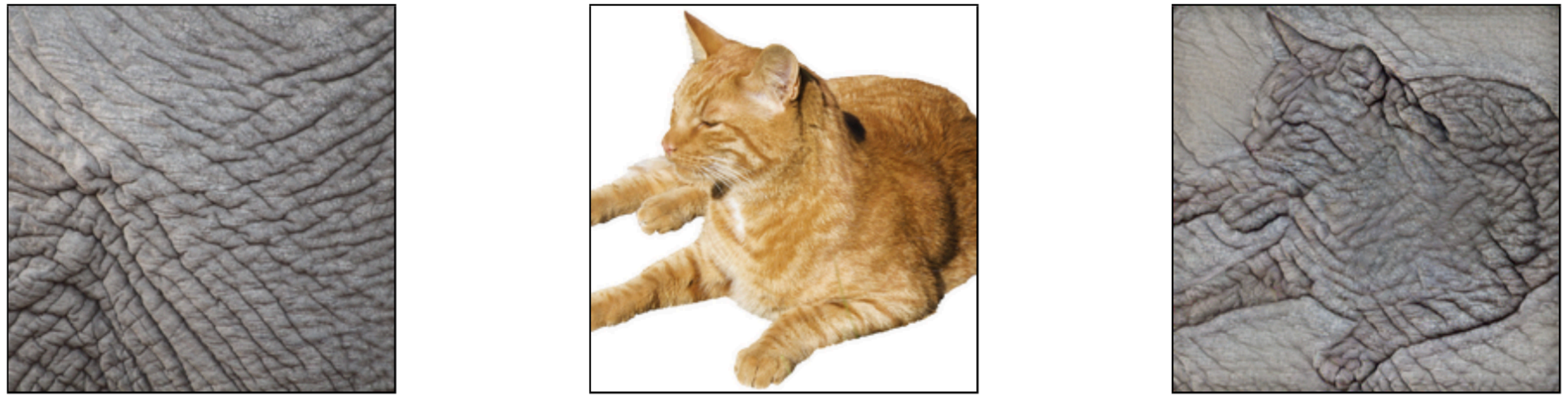
"what is present in the world, and where it is"

Computer Vision is Interdisciplinary - Cognitive Science / Neuroscience

Graphics
Vision

"what is present in the world, and where it is"-David Marr

Computer Vision is Interdisciplinary - Cognitive Science / Neuroscience



(a) Texture image

81.4%	**Indian elephant**
10.3%	indri
8.2%	black swan

(b) Content image

71.1%	**tabby cat**
17.3%	grey fox
3.3%	Siamese cat

(c) Texture-shape cue conflict

63.9%	**Indian elephant**
26.4%	indri
9.6%	black swan

[Geirhos et al., ICLR 2019]

Why study vision?

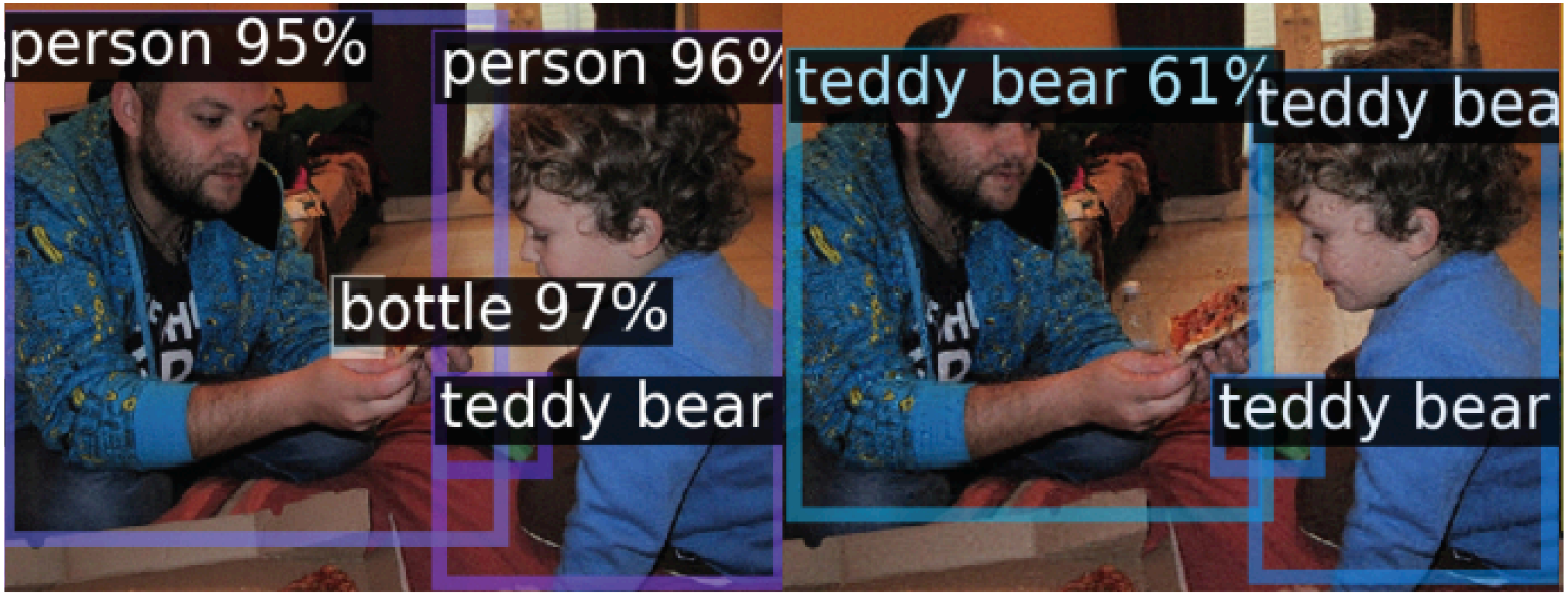
- It is a requirement
- It is cool

Course Structure

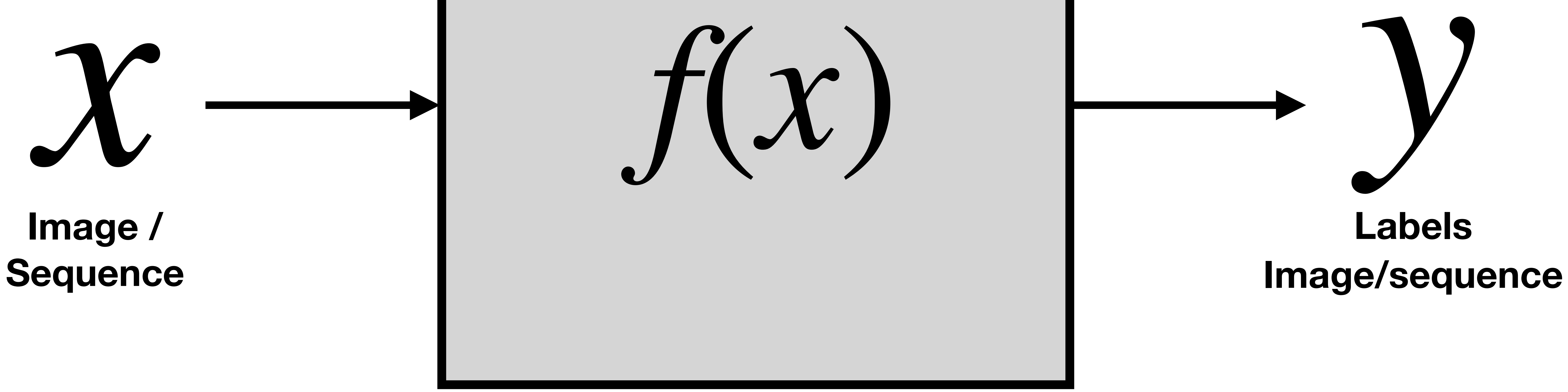
Syllabus

Session	Reading Material	Slides
Part I: Representation Learning and Generative Models (Ayush Tewari)		
Overview of Modern Computer Vision Supervised Learning	Foundations of Computer Vision (Chapter 1, 9, 24)	[TBD]
Unsupervised and Self-supervised Learning	[TBD]	[TBD]
Generative Models in Computer Vision	[TBD]	[TBD]
Part II: 3D Vision & Foundation Models (Elliott Wu)		
3D Computer Vision	[TBD]	[TBD]
Vision Foundation Models	[TBD]	[TBD]
Applications of Computer Vision	[TBD]	[TBD]

Machine Learning for Computer Vision: Supervised Learning



Generic formulation of a computer vision task



x

Image /
Sequence

$f(x)$

y

Labels
Image/sequence

What would make this problem supervised?

Some tasks that humans care about..

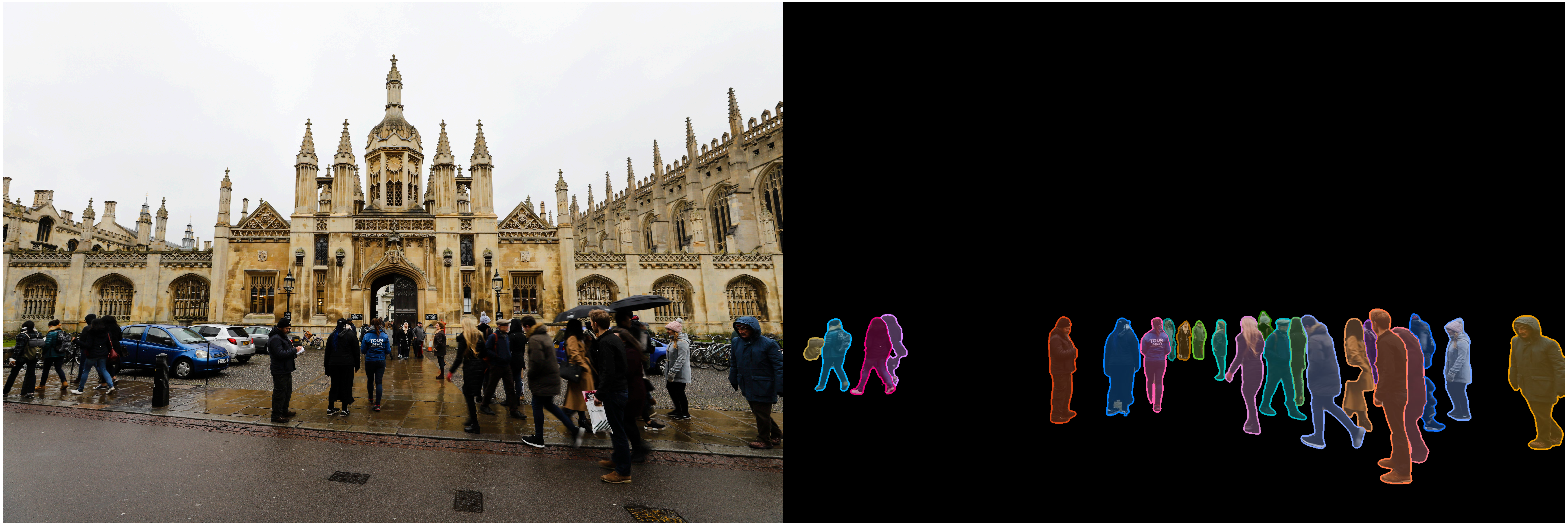
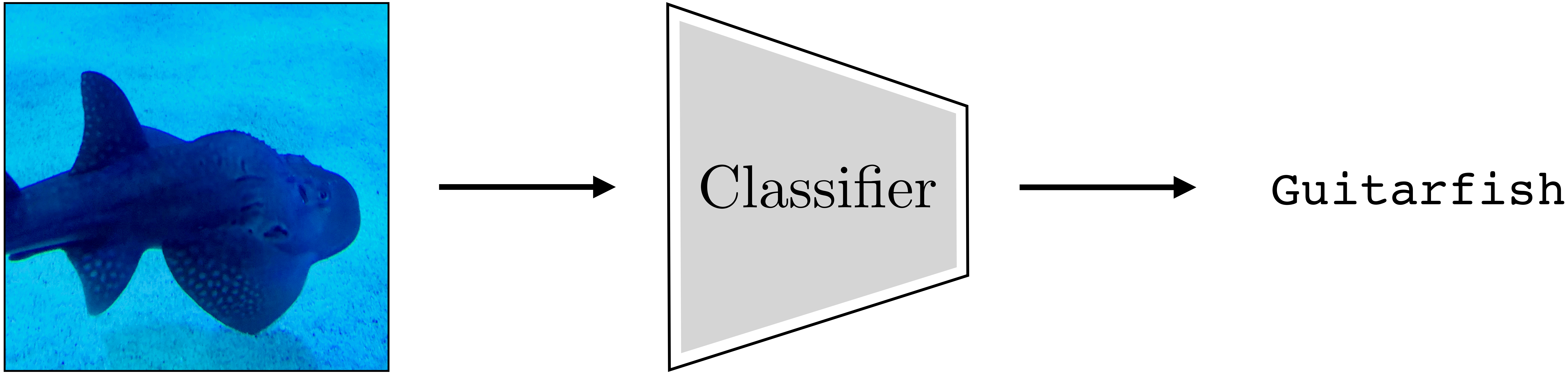


Image Classification



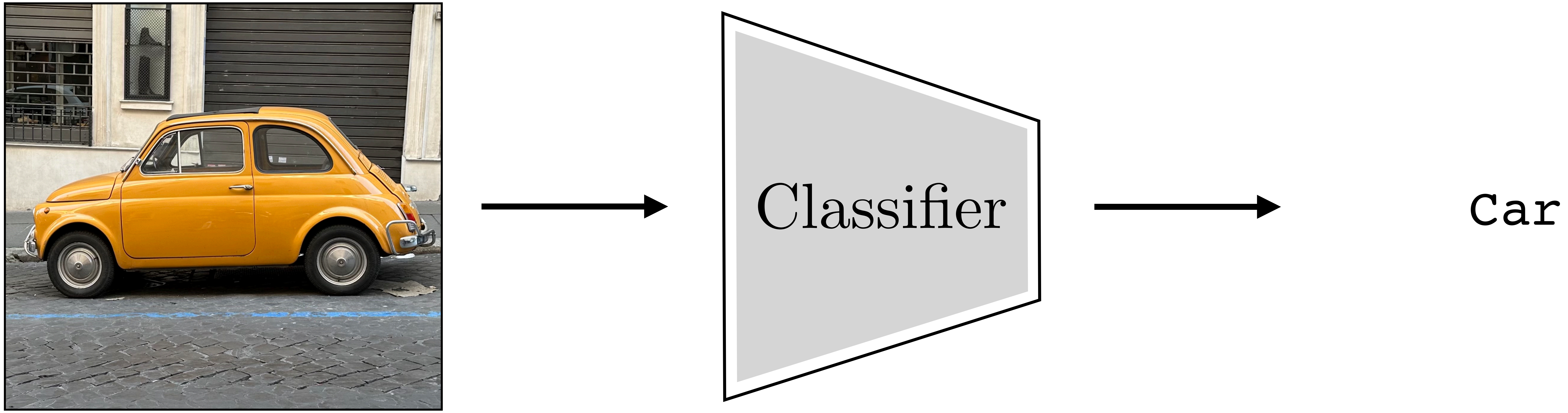
Classifier

`Guitarfish`

image **x**

label **y**

Image Classification



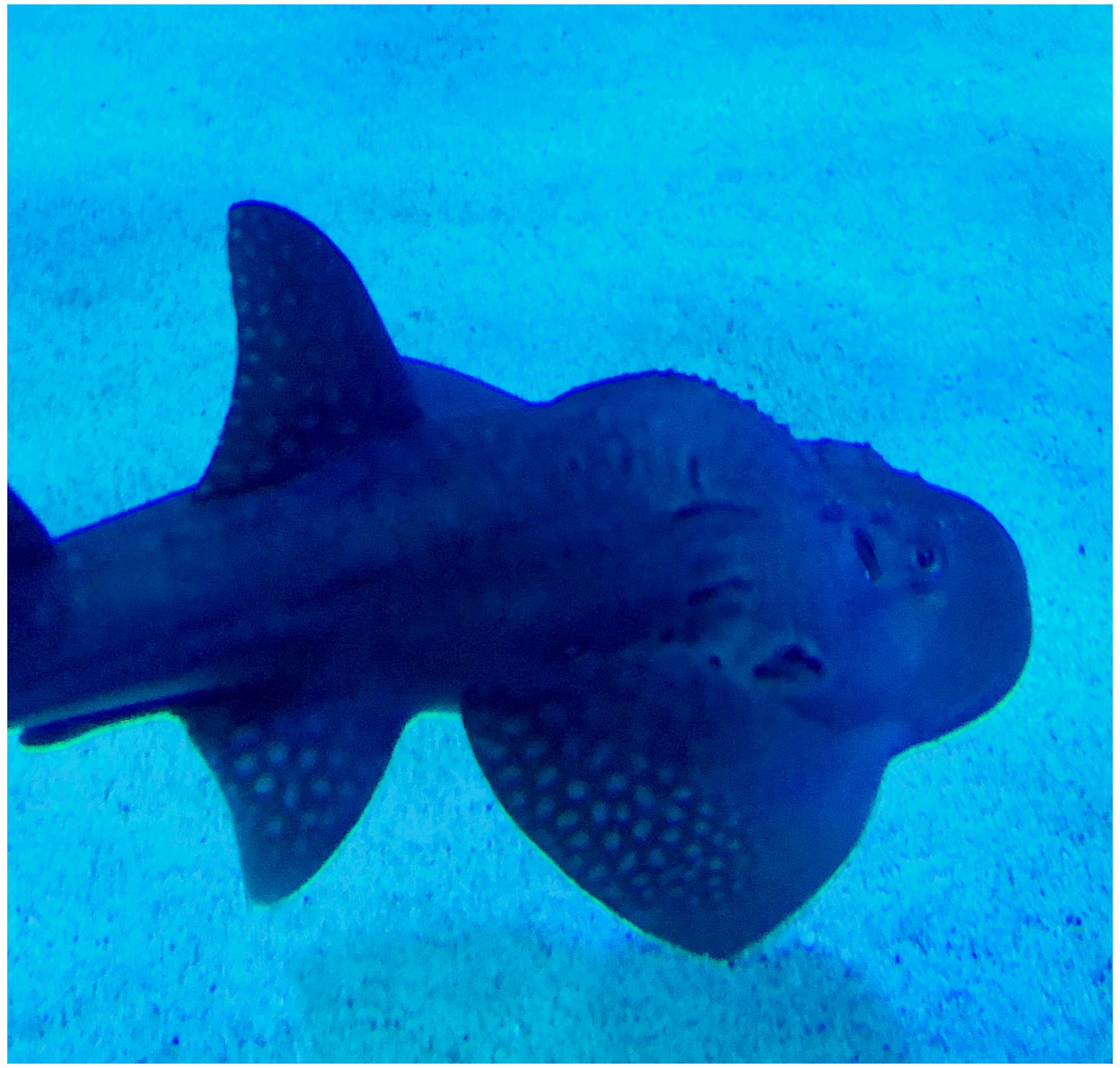
Classifier

car

image $\mathbf{x}$

label $\mathbf{y}$

$\mathbf{x}$



$\xrightarrow{f}$

$\mathbf{y}$

`Guitarfish`

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

Important Questions

1. What data do you train on?
2. How do you represent the input and output?
3. What is the objective?
4. What is the hypothesis space? (e.g., linear, polynomial, neural net?)
5. How do you optimise? (e.g., gradient descent, Newton's method?)

Important Questions

1. **What data do you train on?**

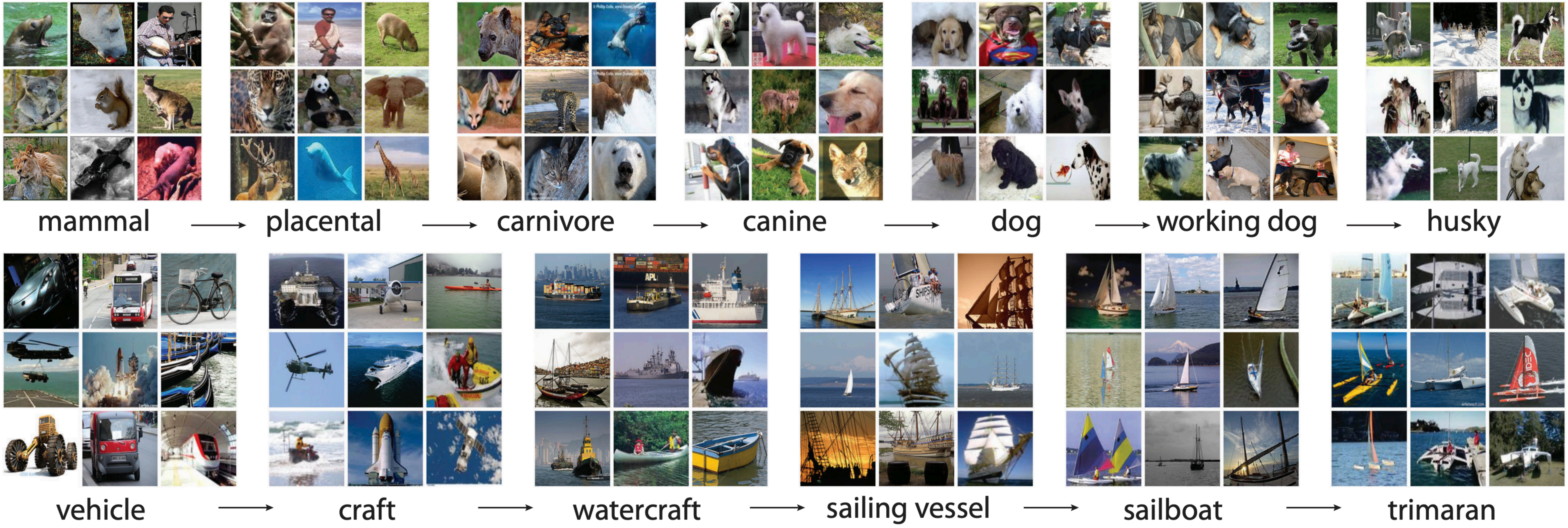
2.How do you represent the input and output?

3. What is the objective?

4. What is the hypothesis space? (e.g., linear, polynomial, neural net?)

5. How do you optimise? (e.g., gradient descent, Newton's method?)

What data to use?



mammal → placental → carnivore → canine → dog → working dog → husky

vehicle → craft → watercraft → sailing vessel → sailboat → trimaran

ImageNet [Deng et al., CVPR 2009]

>14M images with human annotations!

Important Questions

1. What data do you train on?

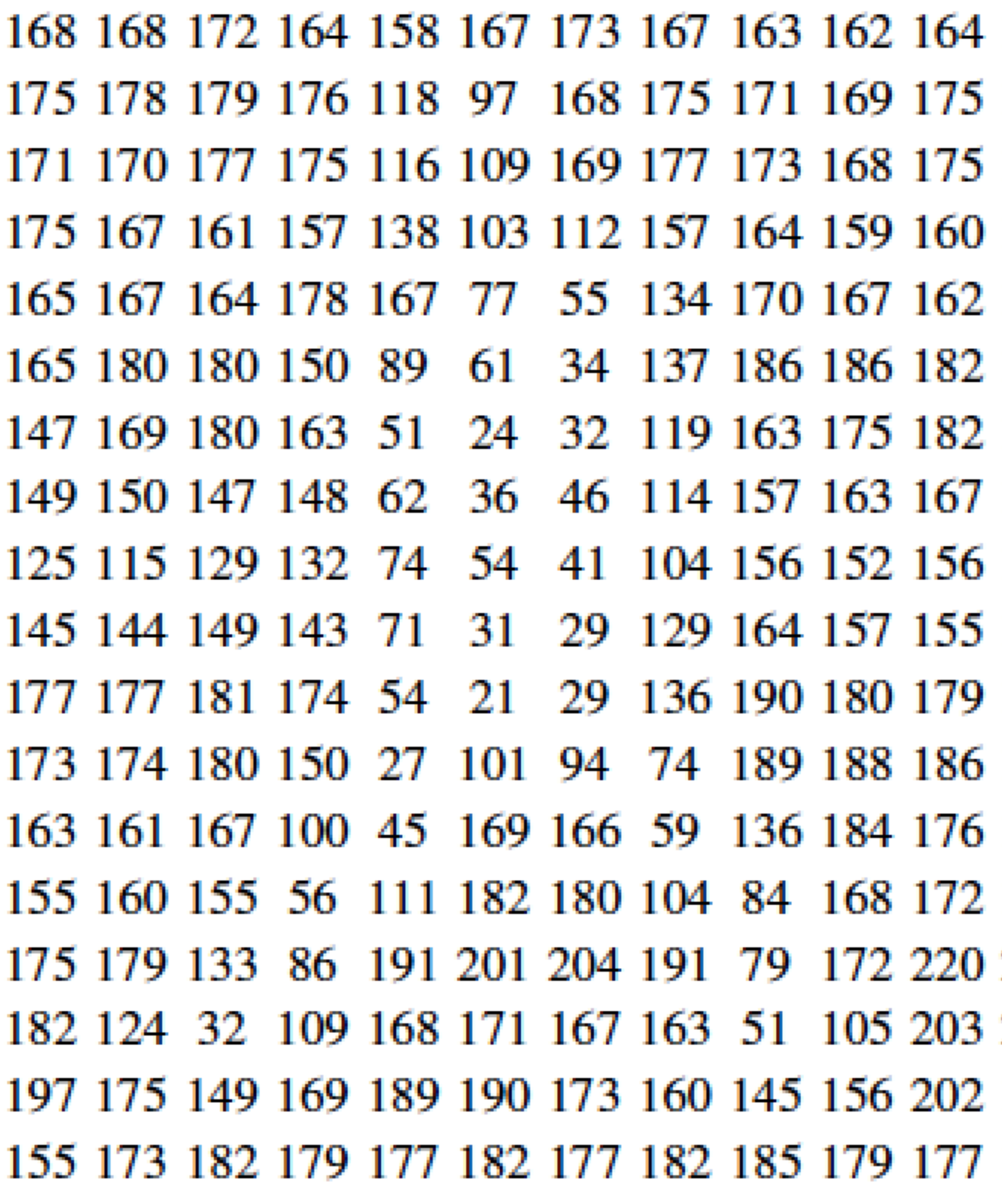
2. How do you represent the input and output?

3. What is the objective?

4. What is the hypothesis space? (e.g., linear, polynomial, neural net?)

5. How do you optimise? (e.g., gradient descent, Newton’s method?)

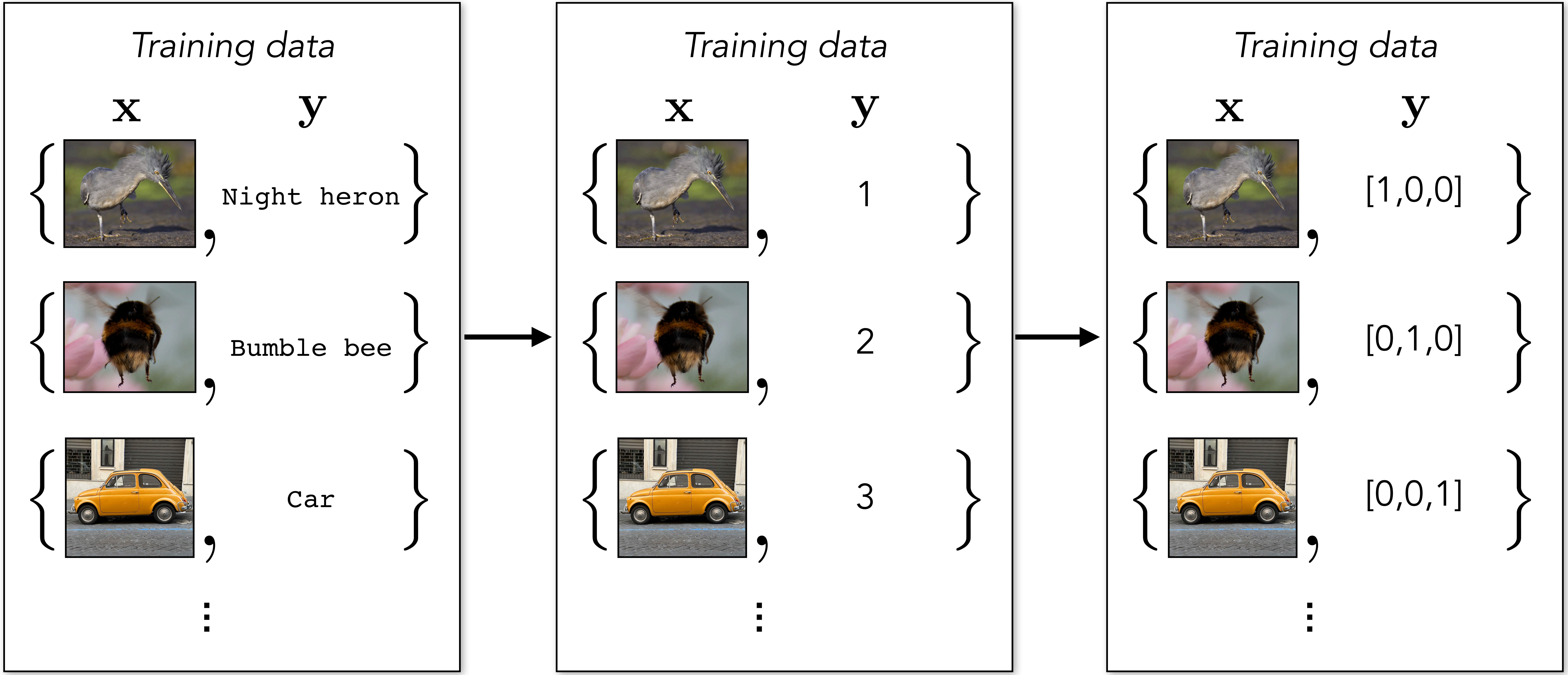
How to represent images?



168	168	172	164	158	167	173	167	163	162	164
175	178	179	176	118	97	168	175	171	169	175
171	170	177	175	116	109	169	177	173	168	175
175	167	161	157	138	103	112	157	164	159	160
165	167	164	178	167	77	55	134	170	167	162
165	180	180	150	89	61	34	137	186	186	182
147	169	180	163	51	24	32	119	163	175	182
149	150	147	148	62	36	46	114	157	163	167
125	115	129	132	74	54	41	104	156	152	156
145	144	149	143	71	31	29	129	164	157	155
177	177	181	174	54	21	29	136	190	180	179
173	174	180	150	27	101	94	74	189	188	186
163	161	167	100	45	169	166	59	136	184	176
155	160	155	56	111	182	180	104	84	168	172
175	179	133	86	191	201	204	191	79	172	220
182	124	32	109	168	171	167	163	51	105	203
197	175	149	169	189	190	173	160	145	156	202
155	173	182	179	177	182	177	182	185	179	177

How to represent class labels?

One-hot vector



Training data

$\mathbf{x}$ $\mathbf{y}$

{ , Night heron }

{ , Bumble bee }

{ , Car }

⋮

Training data

$\mathbf{x}$ $\mathbf{y}$

{ , 1 }

{ , 2 }

{ , 3 }

⋮

Training data

$\mathbf{x}$ $\mathbf{y}$

{ , [1,0,0] }

{ , [0,1,0] }

{ , [0,0,1] }

⋮

Important Questions

1. What data do you train on?

2. How do you represent the input and output?

3. **What is the objective?**

4. What is the hypothesis space? (e.g., linear, polynomial, neural net?)

5. How do you optimise? (e.g., gradient descent, Newton's method?)

What should the loss be?

0-1 loss (number of misclassifications)

$$\mathcal{L}(\hat{\mathbf{y}}, \mathbf{y}) = \mathbb{1}(\hat{\mathbf{y}} \neq \mathbf{y})$$

← discrete, NP-hard to optimize!

Cross entropy

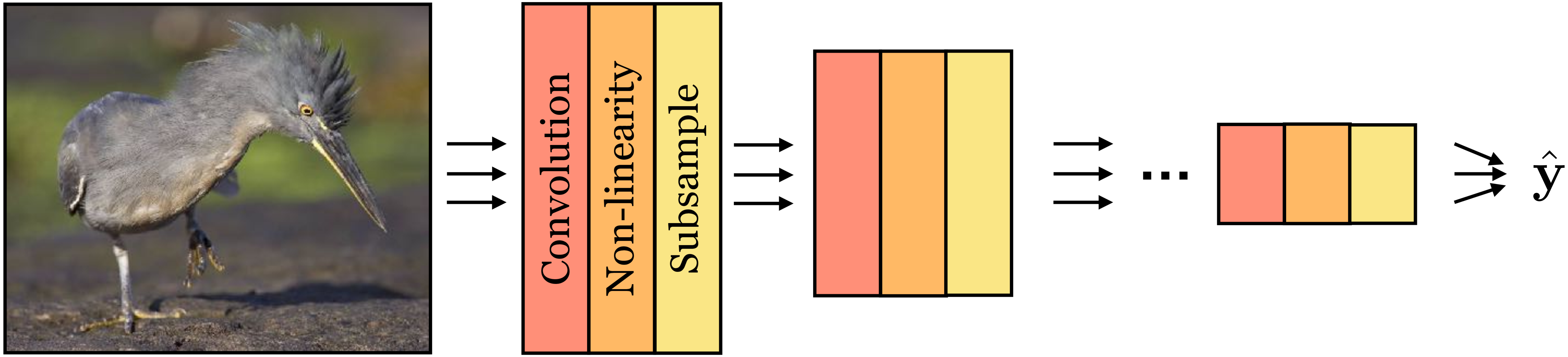
$$\mathcal{L}(\hat{\mathbf{y}}, \mathbf{y}) = H(\mathbf{y}, \hat{\mathbf{y}}) = -\sum_{k=1}^{K} y_k \log \hat{y}_k$$

← continuous,
differentiable,
convex

Important Questions

1. How do you represent the input and output?
2. What data do you train on?
3. What is the objective?
4. **What is the hypothesis space? (e.g., linear, polynomial, neural net?)**
5. How do you optimise? (e.g., gradient descent, Newton’s method?)

Neural Networks..



Convolution

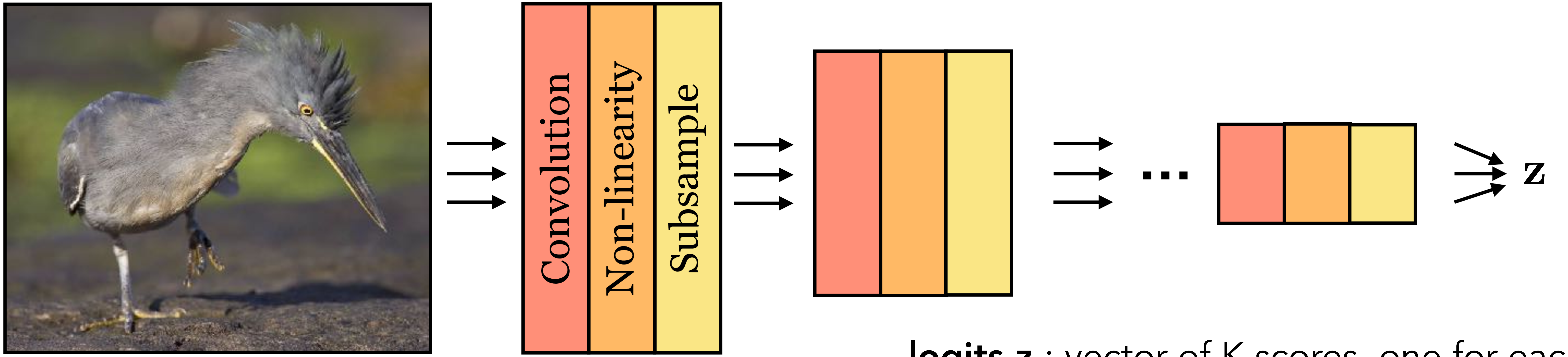
Non-linearity

Subsample

...

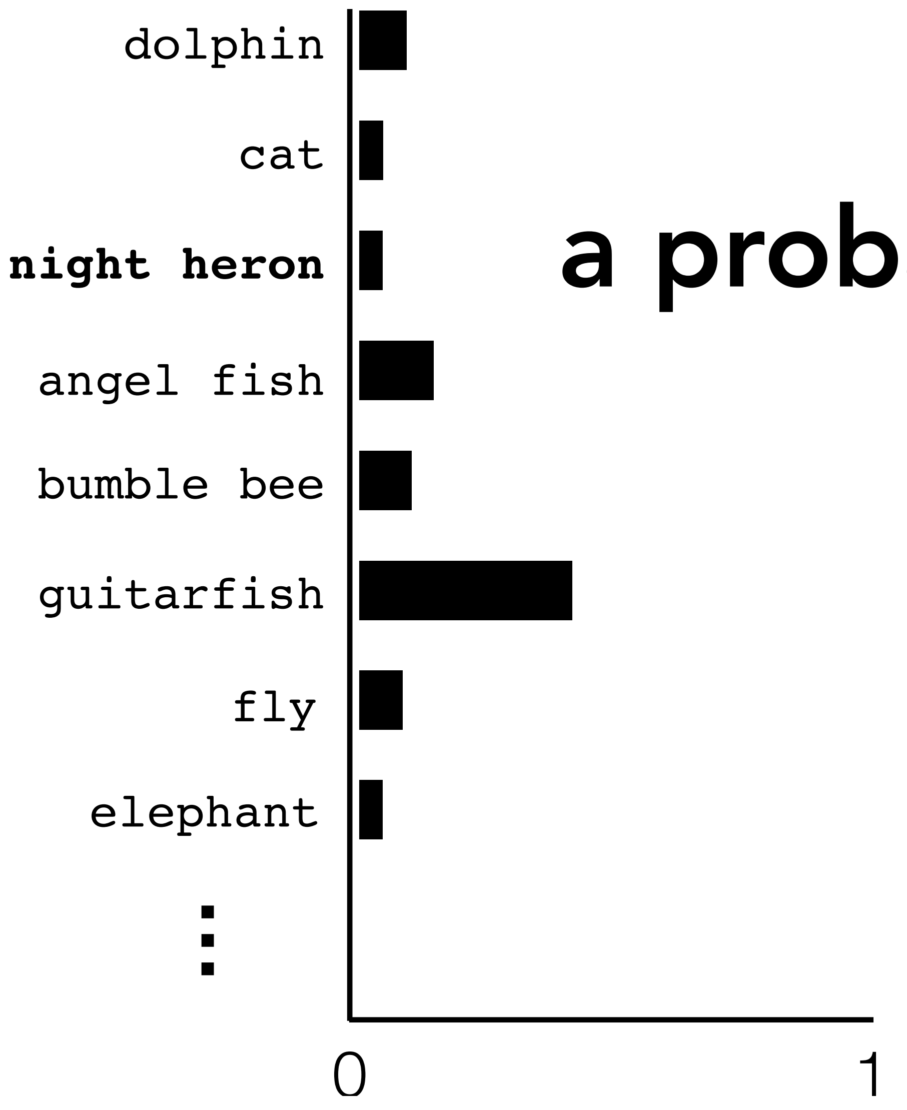
$\hat{\mathbf{y}}$

Neural Networks..



logits z : vector of K scores, one for each class

$$\hat{\mathbf{y}} =$$



a probability mass function!

$$\hat{\mathbf{y}} = \texttt{softmax}(\mathbf{z})$$

$$\hat{y}_j = \frac{e^{-z_j}}{\sum_{k=1}^{K} e^{-z_k}}$$

Important Questions

1. How do you represent the input and output?

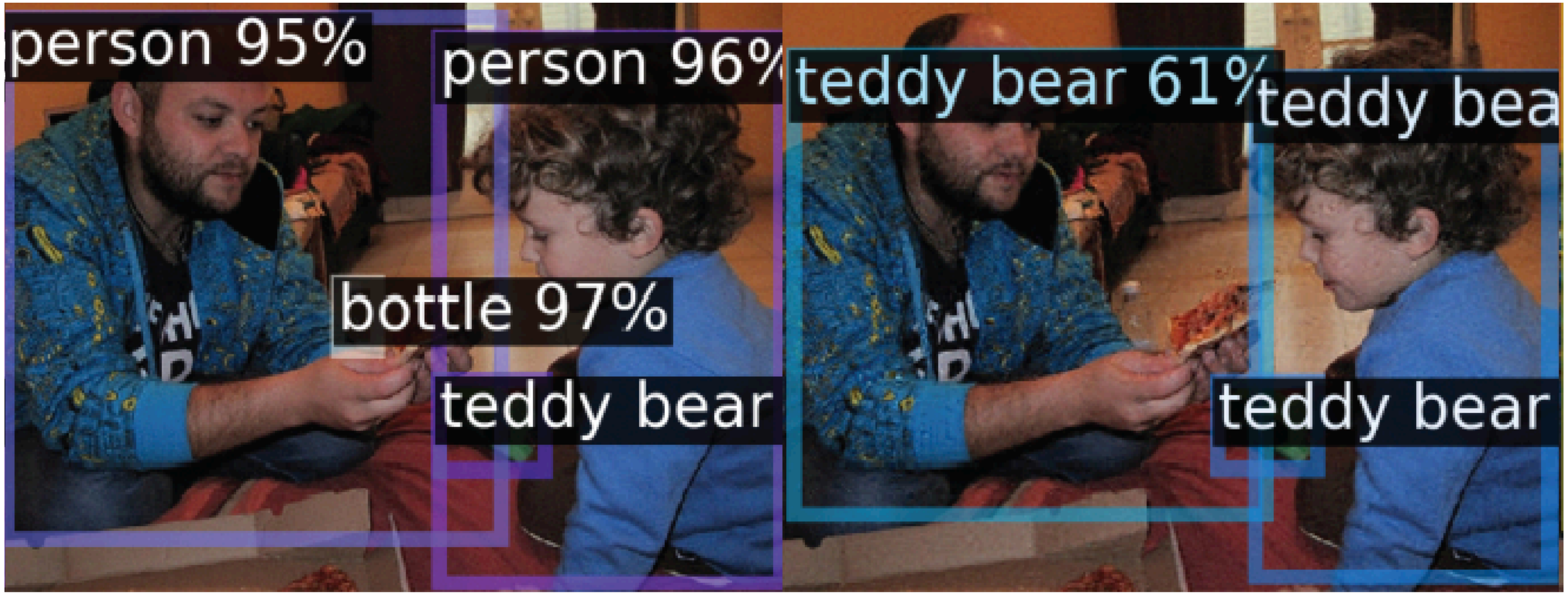
2. What data do you train on?

3. What is the objective?

4. What is the hypothesis space? (e.g., linear, polynomial, neural net?)

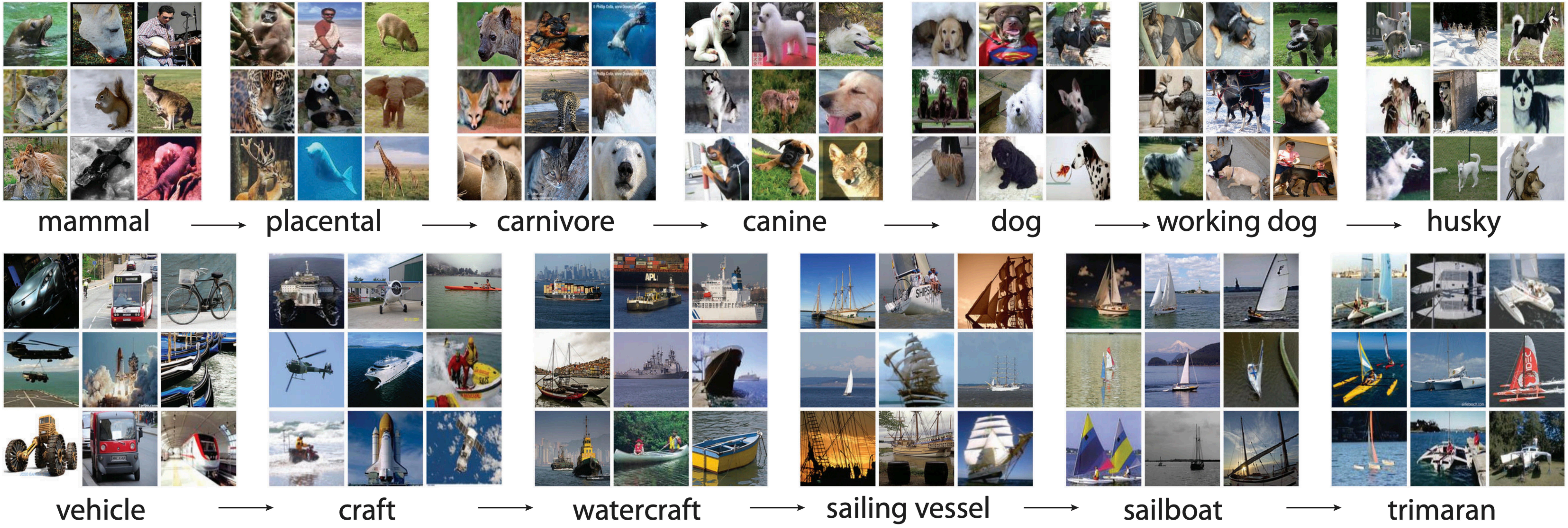
5. **How do you optimise? (e.g., gradient descent, Newton’s method?)**

Machine Learning for Computer Vision: Supervised Learning



[Harrington et al., ICLR 2024]

Why is this dataset necessary?



mammal → placental → carnivore → canine → dog → working dog → husky

vehicle → craft → watercraft → sailing vessel → sailboat → trimaran

ImageNet [Deng et al., CVPR 2009]

>14M images with human-verified annotations!

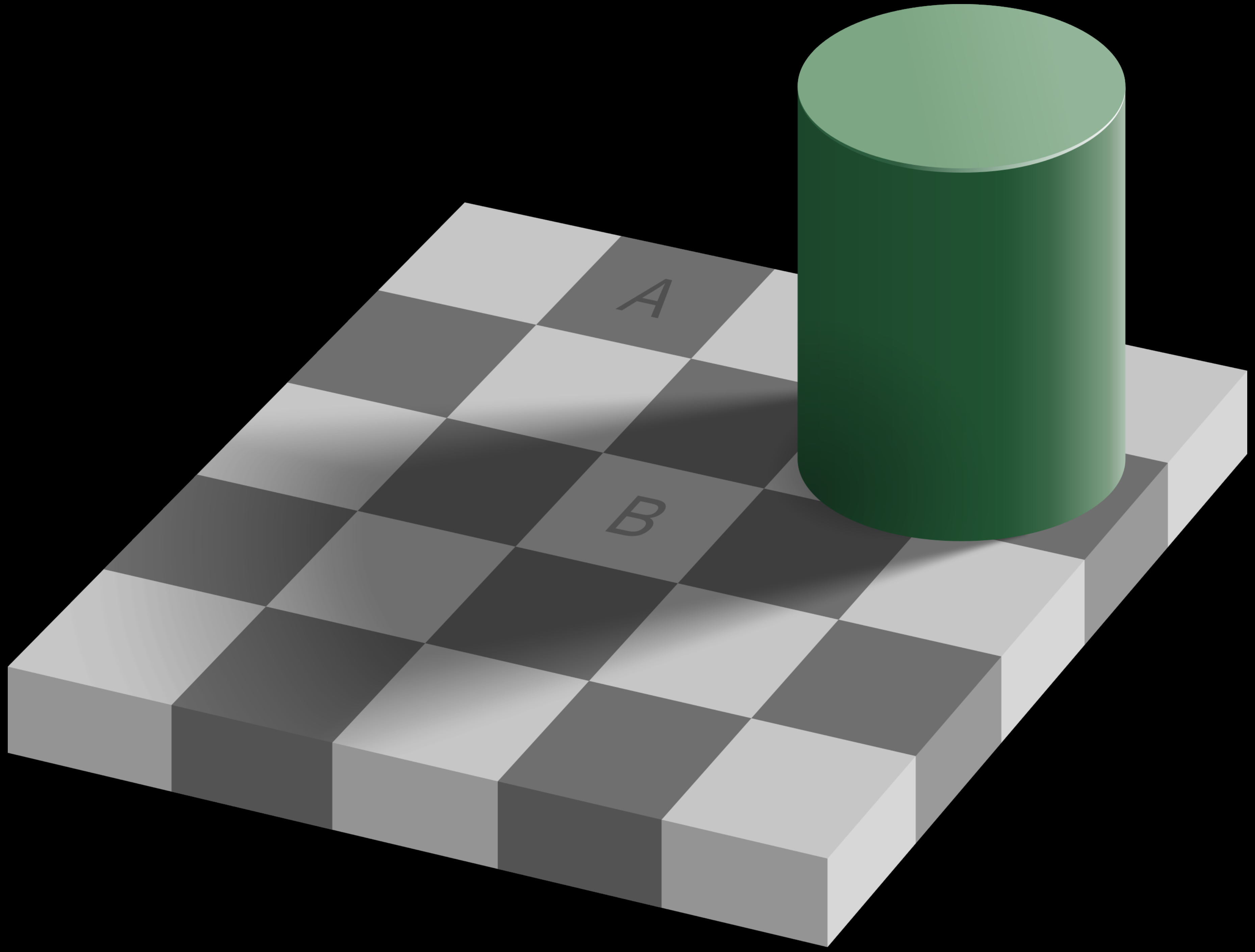
What will courseworks look like?

- No large-scale training
- Focus will be on analysis, possibly with some smaller-scale finetuning
- We will give you pretrained models, optimisation environment, etc.
- Similar (in spirit) to 4F13 courseworks

Advanced Computer Vision: Overview

MLMI17

Ayush Tewari



UNIVERSITY OF CAMBRIDGE