CS280: Graduate Computer Vision

Spring 2024

Lecture: MW 12:30-2pm

1102 Berkeley Way West, UC Berkeley

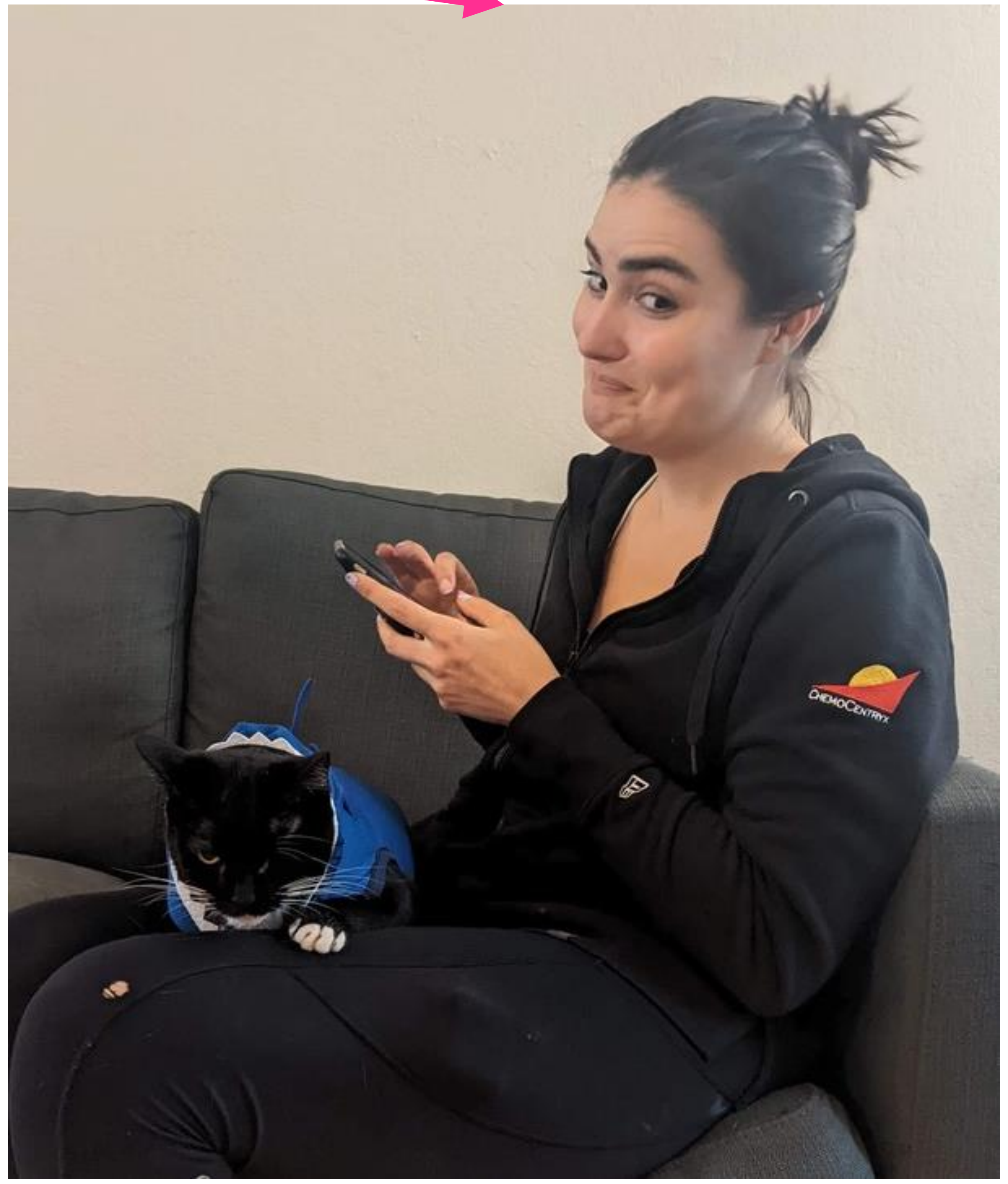
Meet your **AMAZING** course staff

Prof. Alexei (Alyosha) Efros

- loves gelato & bets
- thinks everything is nearest neighbors
- Prefers pixels to words

Suzie Petryk

- has never seen a moose
- thinks Grimes should give a guest lecture
- caretaker of BAIR class pet: DALL-E the stuffed sheep



Lisa Dunlap

- can be found painting nails at work
- trying and failing to start a prank war in BAIR
- uses the diagnostic manual on mental disorders as a monitor stand

Boring administrative things

Prereqs: solid command of Linear Algebra, programming, and Deep Learning

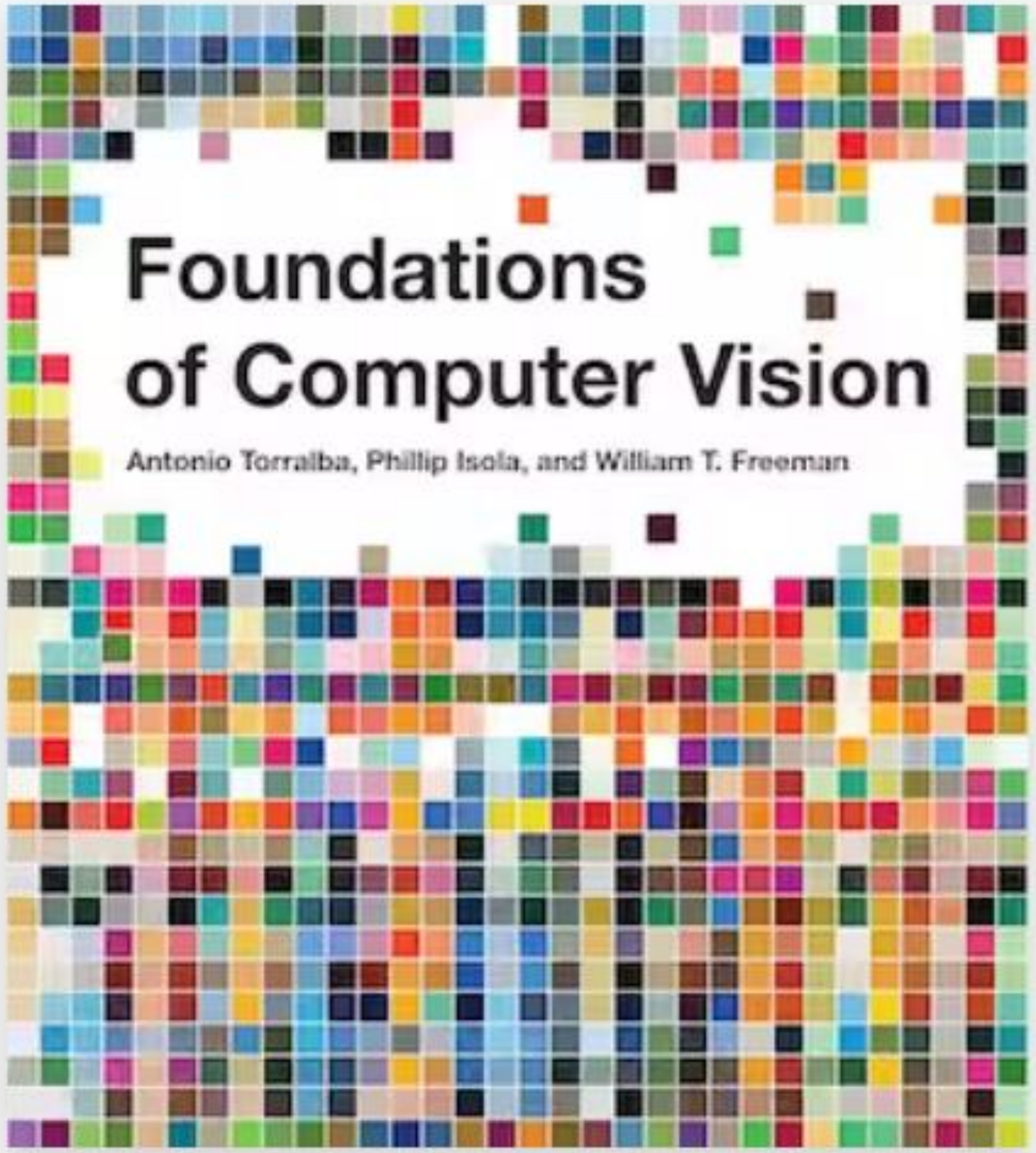
Grade Breakdown:

35% homework:	~4 assignments, due every 2-3 weeks
35% exams:	1 exam plus in-class pop quizzes
30% final project:	Presentations in the first week of May

Website: **https://cs280-berkeley.github.io/**

For those on the waitlist: you should have gotten an email about whether you are likely to get off the waitlist (the waitlist is long so temper expectations)

Textbook



From: Adaptive Computation and Machine Learning series

Foundations of Computer Vision

By Antonio Torralba, Phillip Isola and William T. Freeman

840 pp., 8 x 9 in, 317 color illus., 158 b&w illus.
Hardcover
ISBN: 9780262048972
Published: April 16, 2024
Publisher: The MIT Press

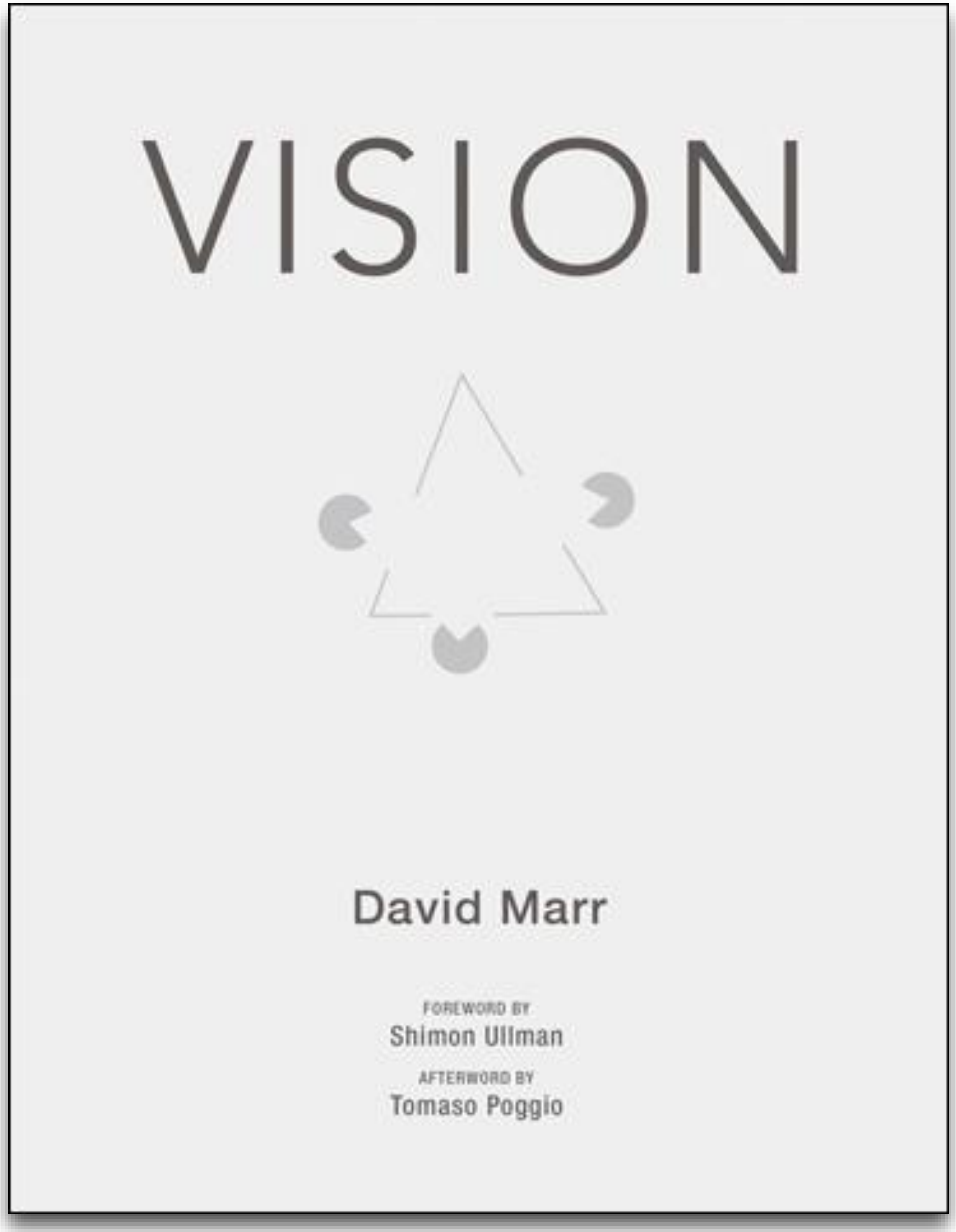
https://mitpress.mit.edu/9780262048972/foundations-of-computer-vision/

Lecture 1: Intro to Computer Vision

To see

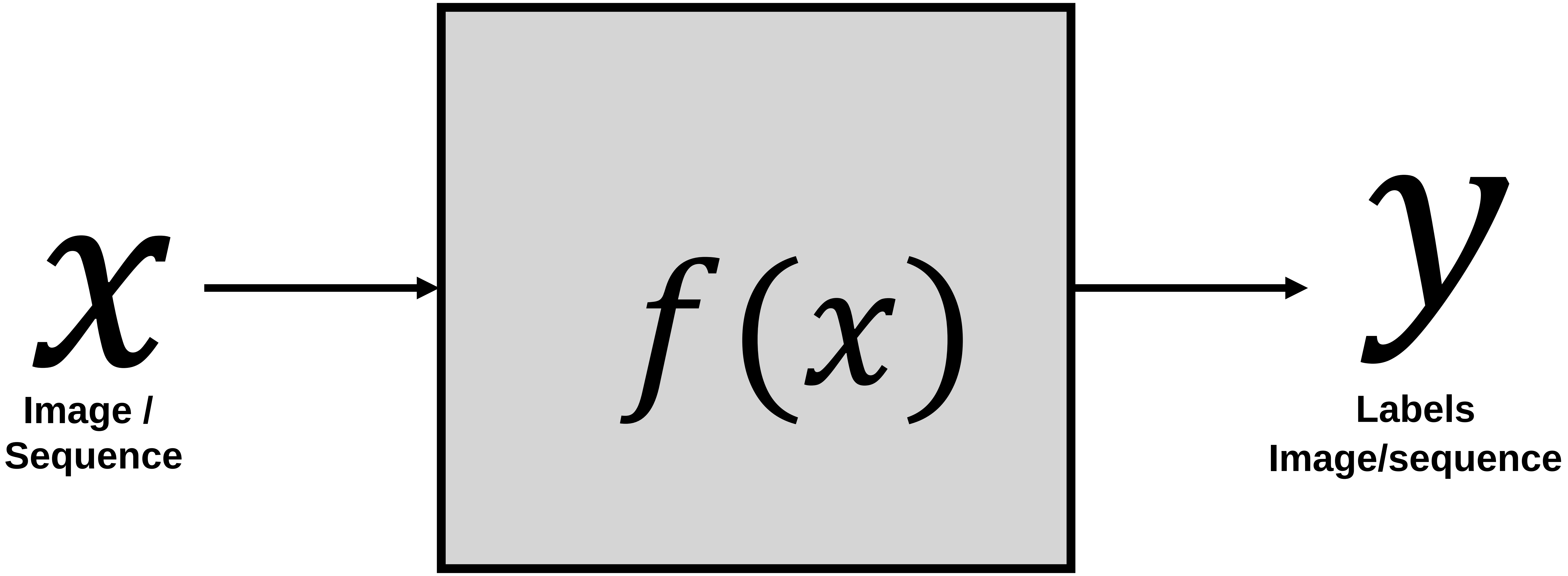
"What does it mean, to see? The plain man's answer (and Aristotle's, too). would be, to know what is where by looking."

To discover from images what is present in the world, where things are, what actions are taking place, to predict and anticipate events in the world.



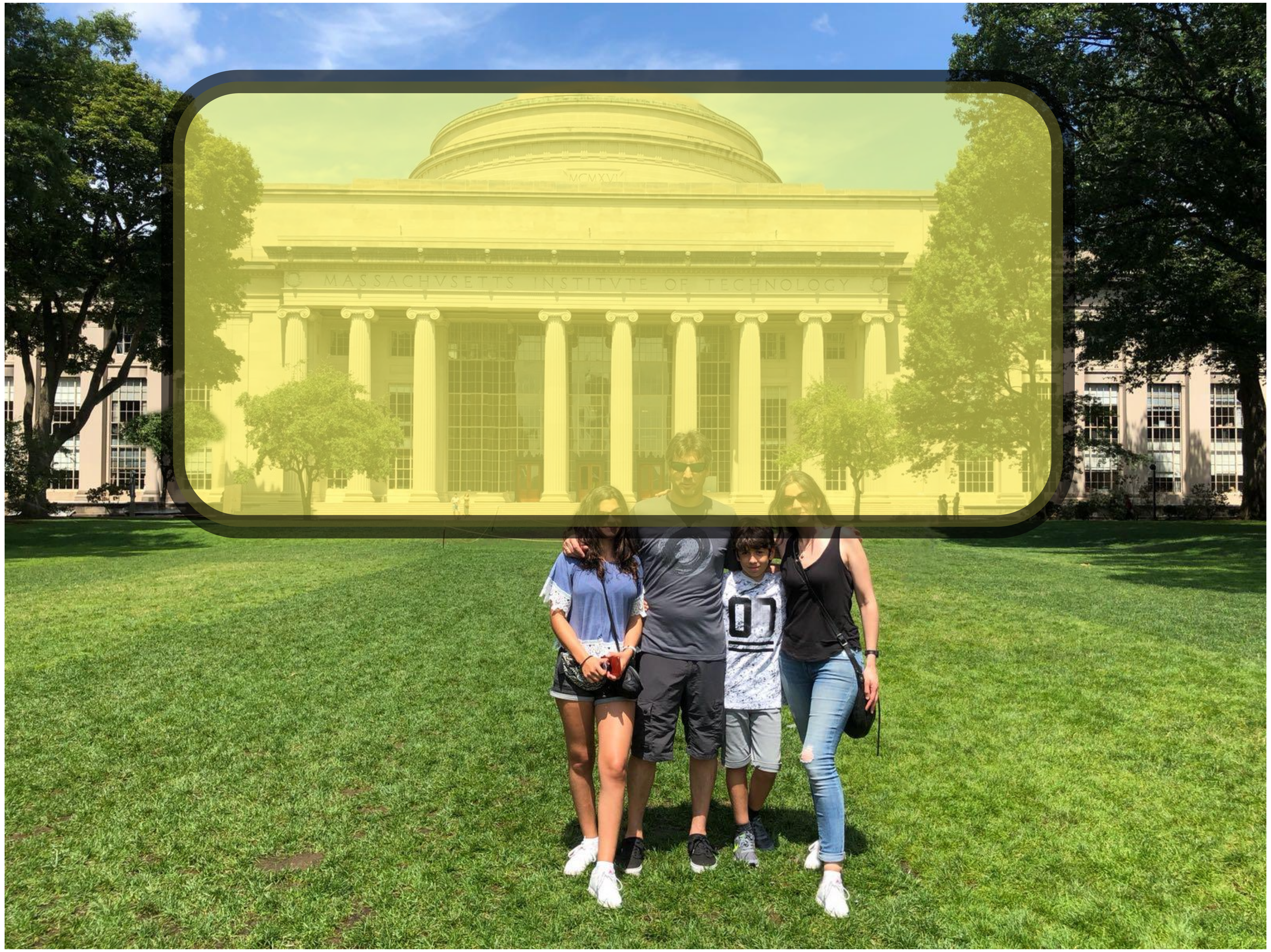
Some slides from Torralba, Freeman, Isola

Tasks: generic formulation



Tasks: what humans care about

Tasks: what humans care about



Verification: is this a building?

Recognition: which building is this?

Tasks: what humans care about

Image classification: list all the objects present in the image

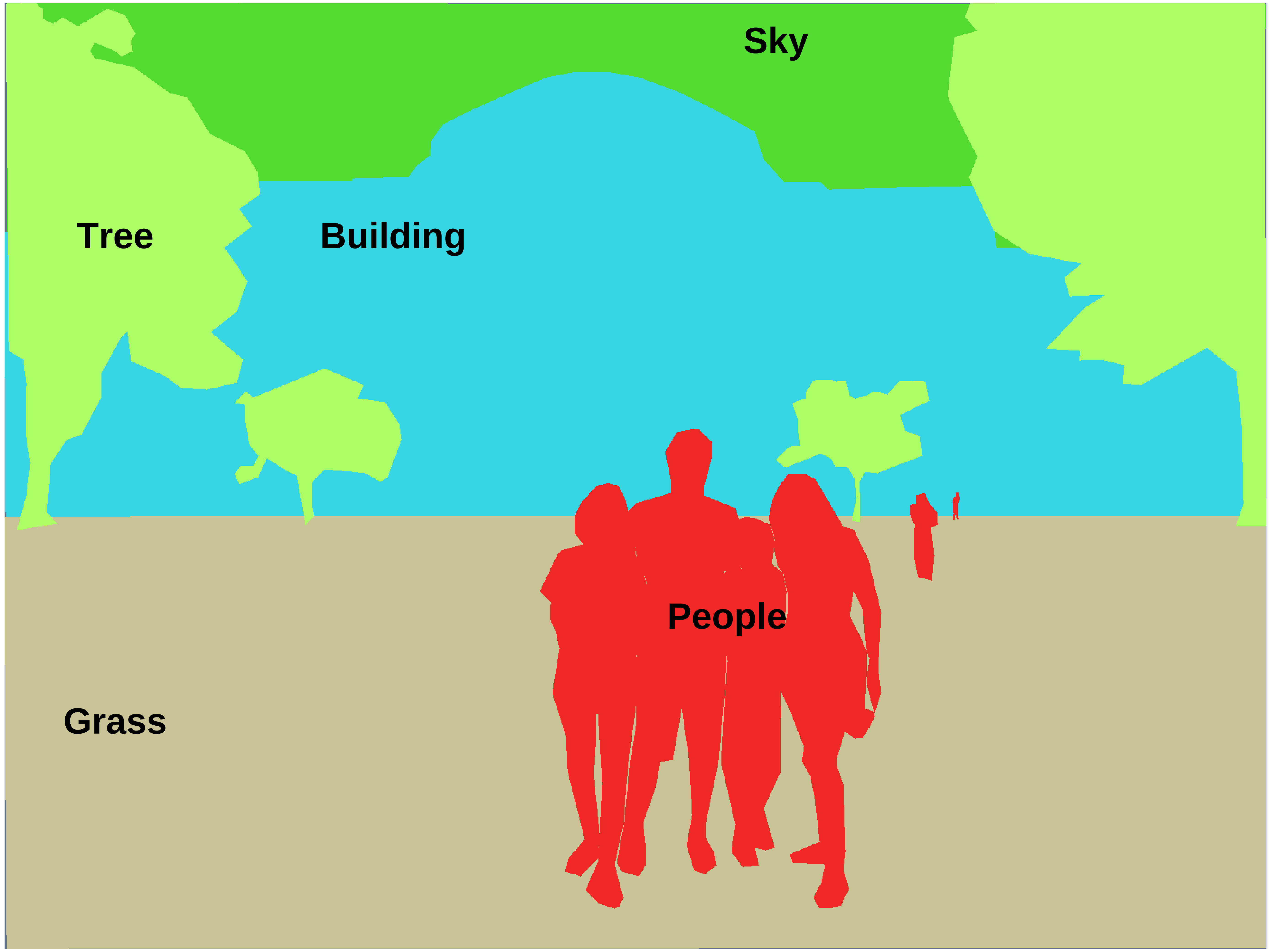
- Building
- Grass
- People
- Trees
- Sky
- Columns
- …

Tasks: what humans care about

Scene categorization

- Outdoor
- Campus
- Garden
- Clear sky
- Spring
- Group picture
- …

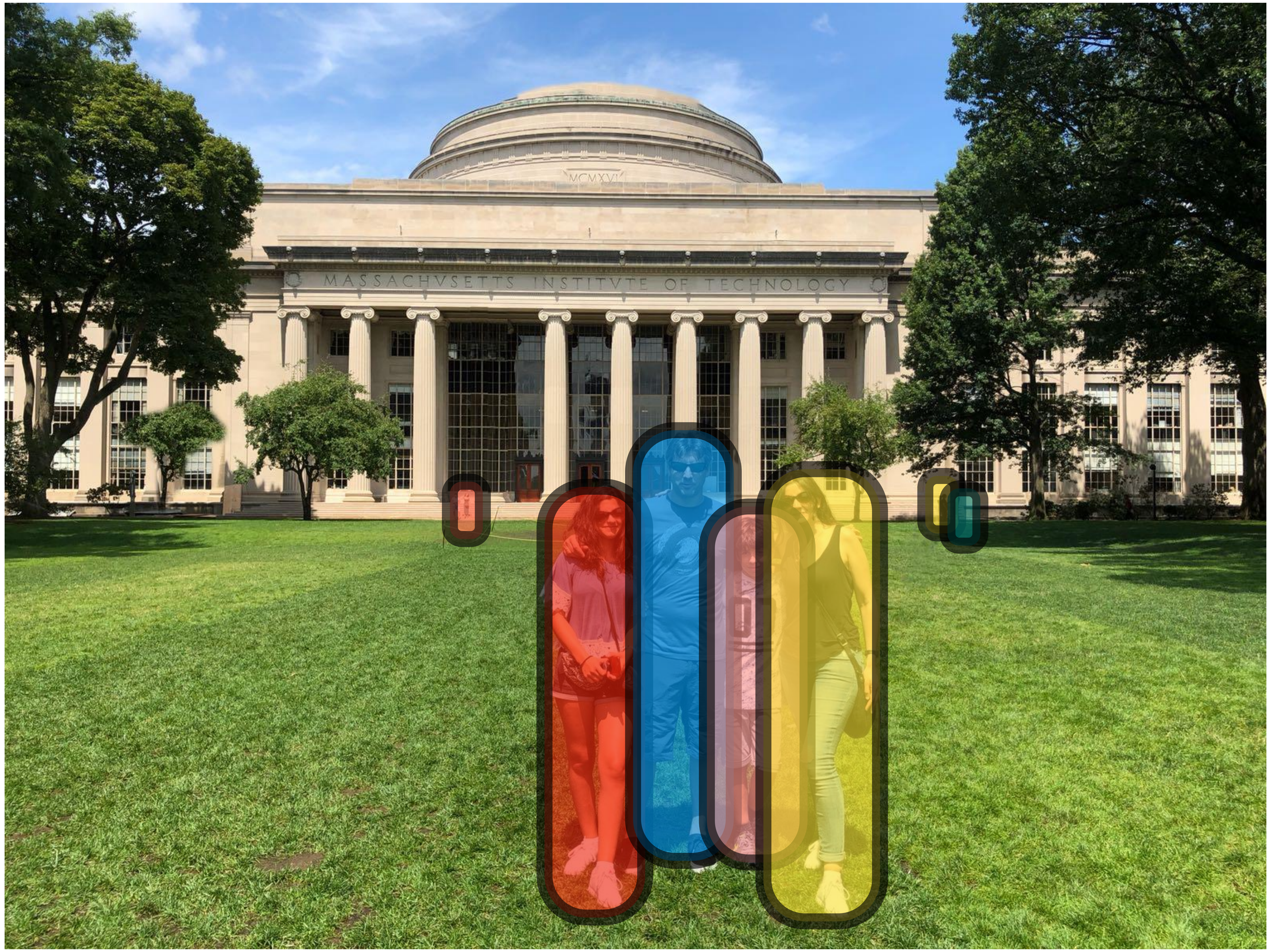
Tasks: what humans care about



Semantic segmentation:
Assign labels to all the pixels in the image

Related tasks:
- **Semantic segmentation**
- **Object categorization**

Tasks: what humans care about

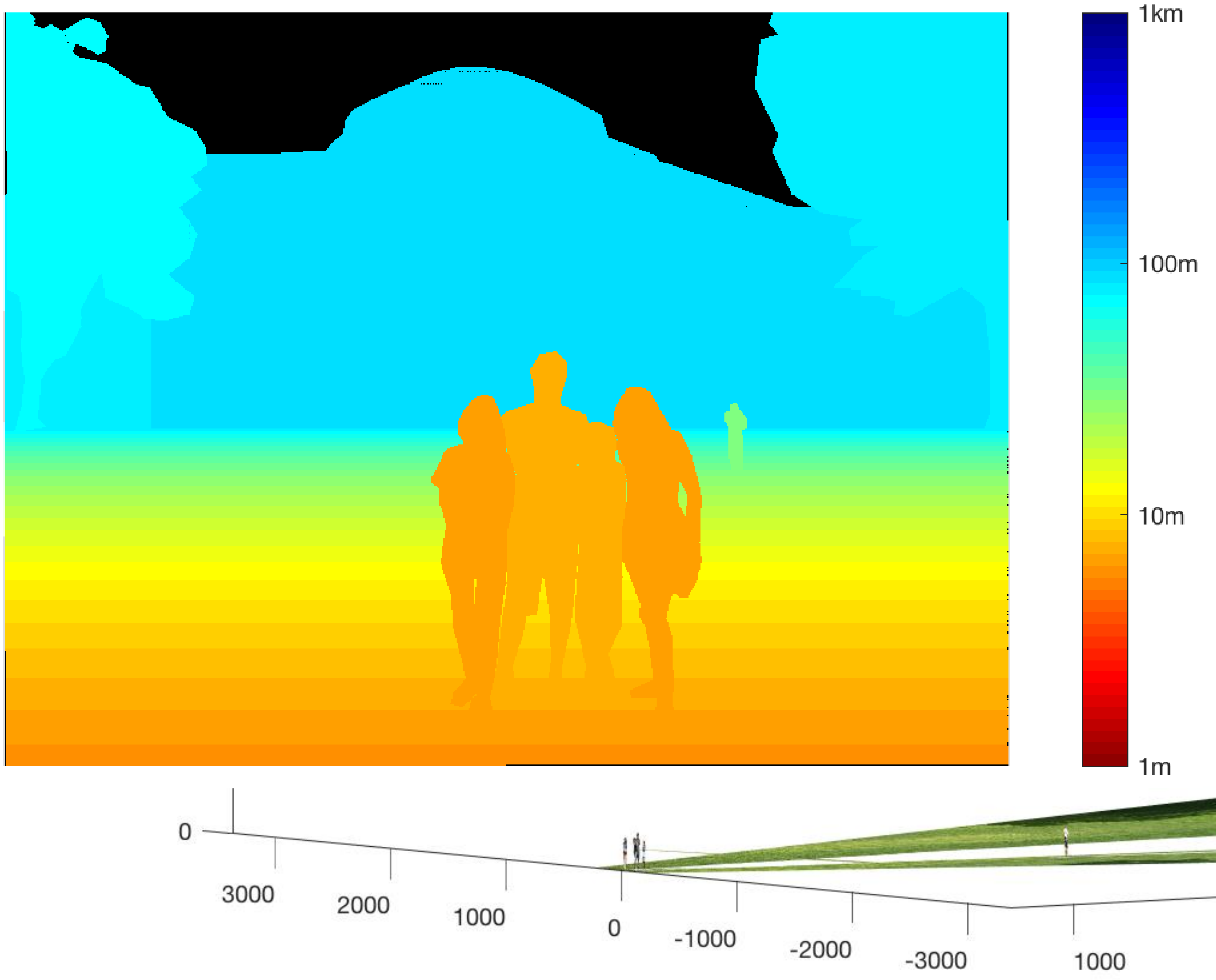


Detection: Locate all the people in this image

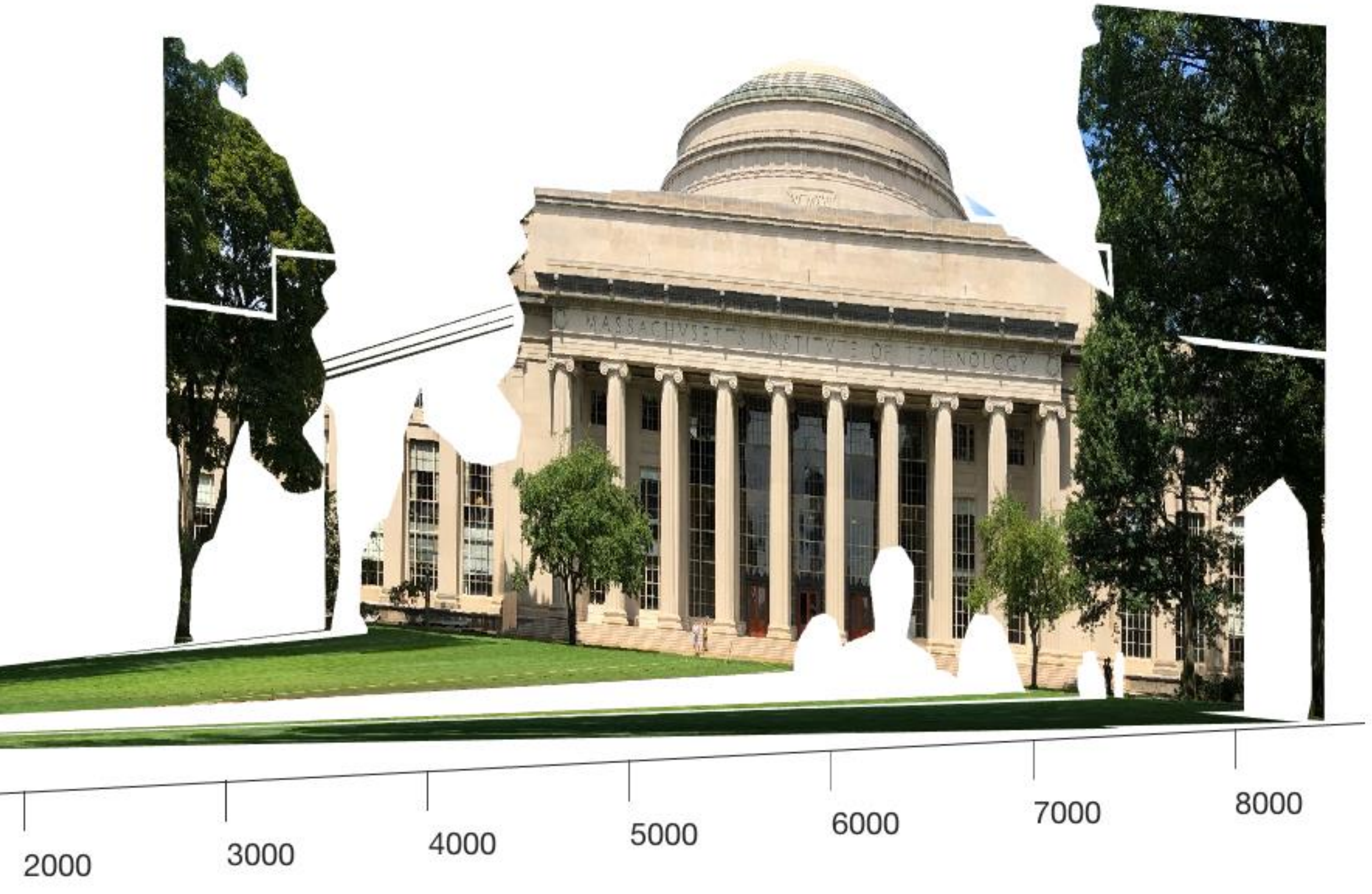
Tasks: what humans care about

Recognition: who is this person?

Tasks: what humans care about



Rough 3D layout, depth ordering

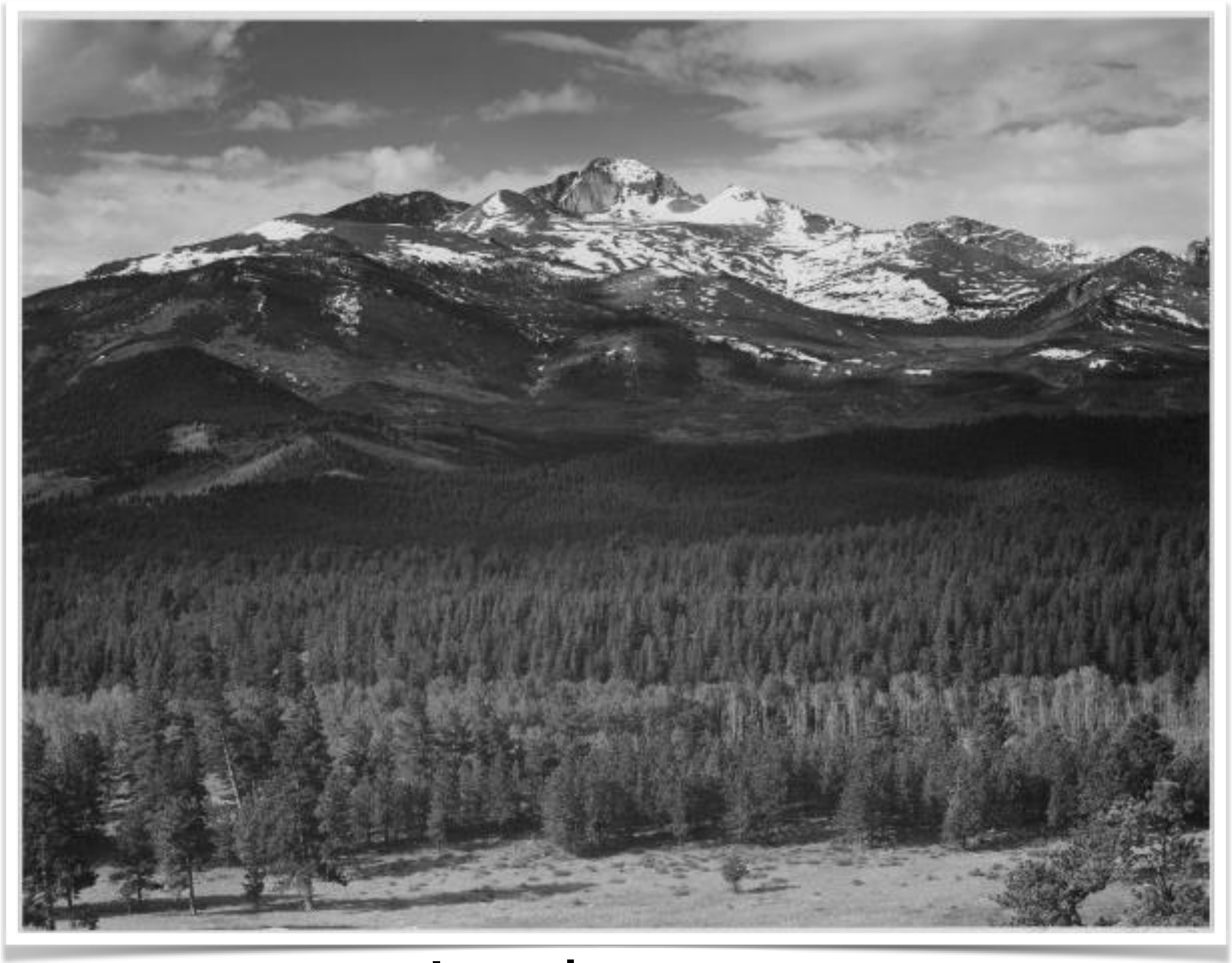


Tasks: what humans care about

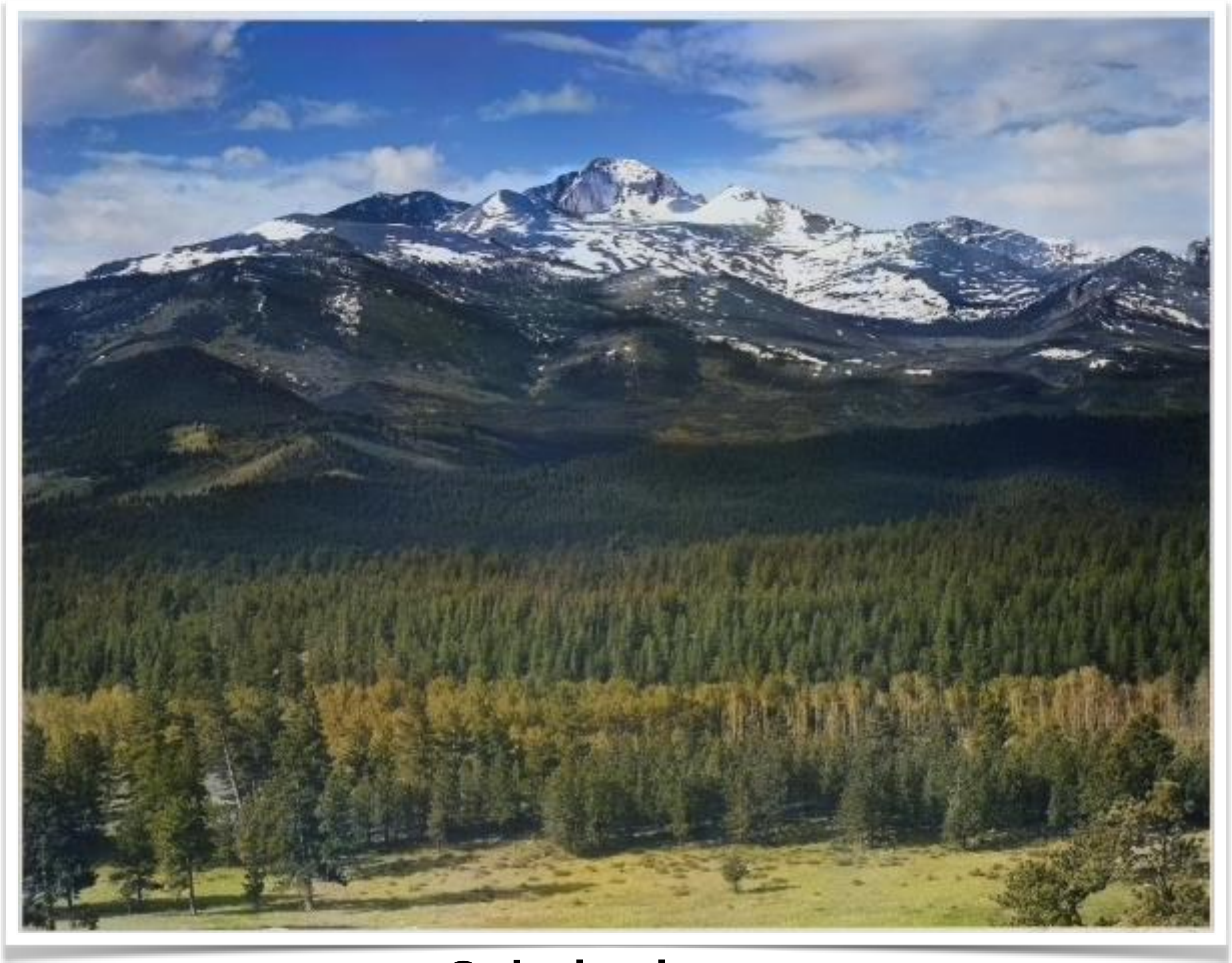
Making new images

Tasks: what humans care about

Adding missing content



Input image



Colorized output

Tasks: what humans care about

Predicting future events

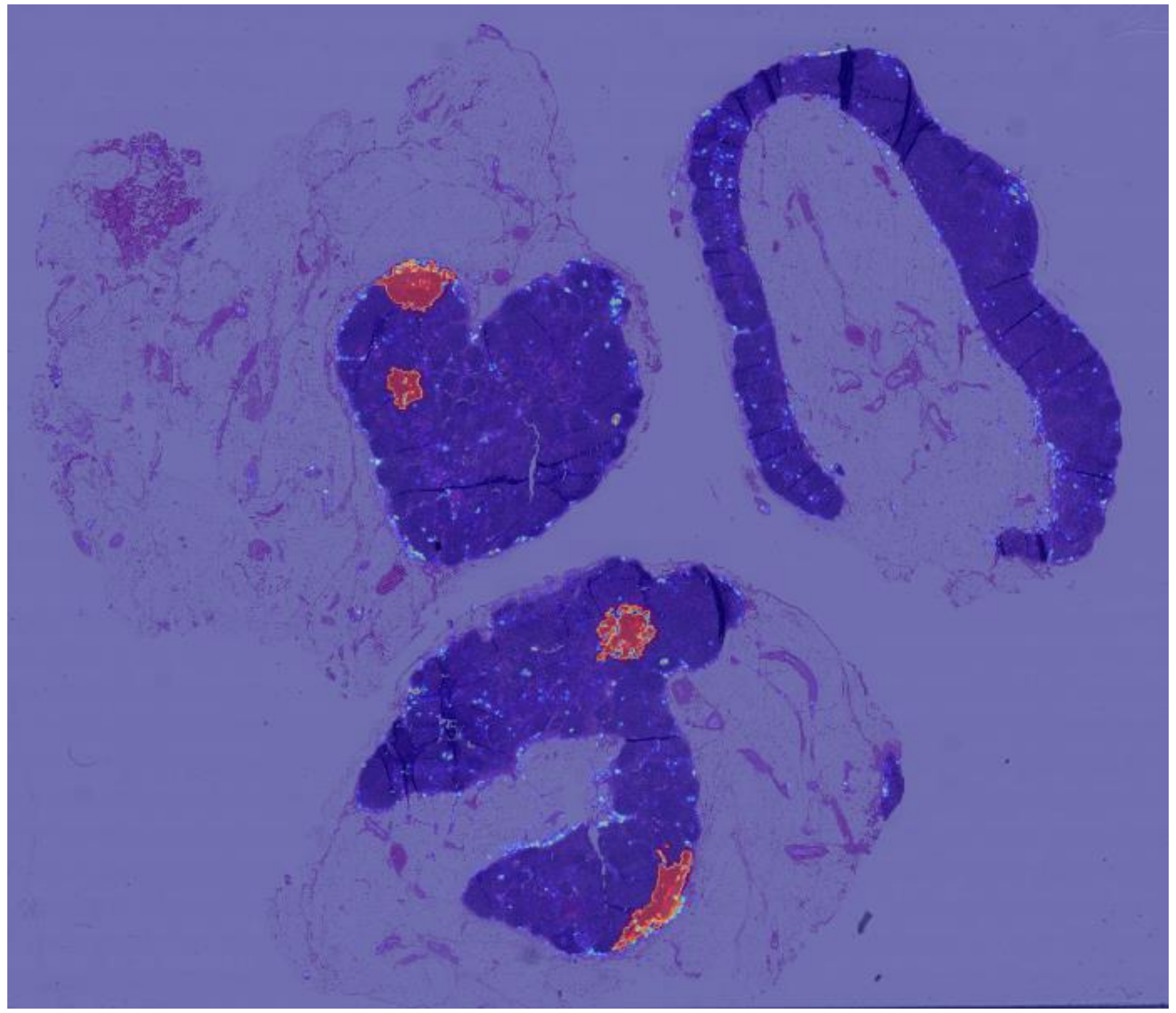
What is going to happen?

Exciting times in computer vision

Robotics

Driving

Medical applications



Gaming

Accessibility

Exciting times in computer vision!

"A cup of coffee"

"A cat"

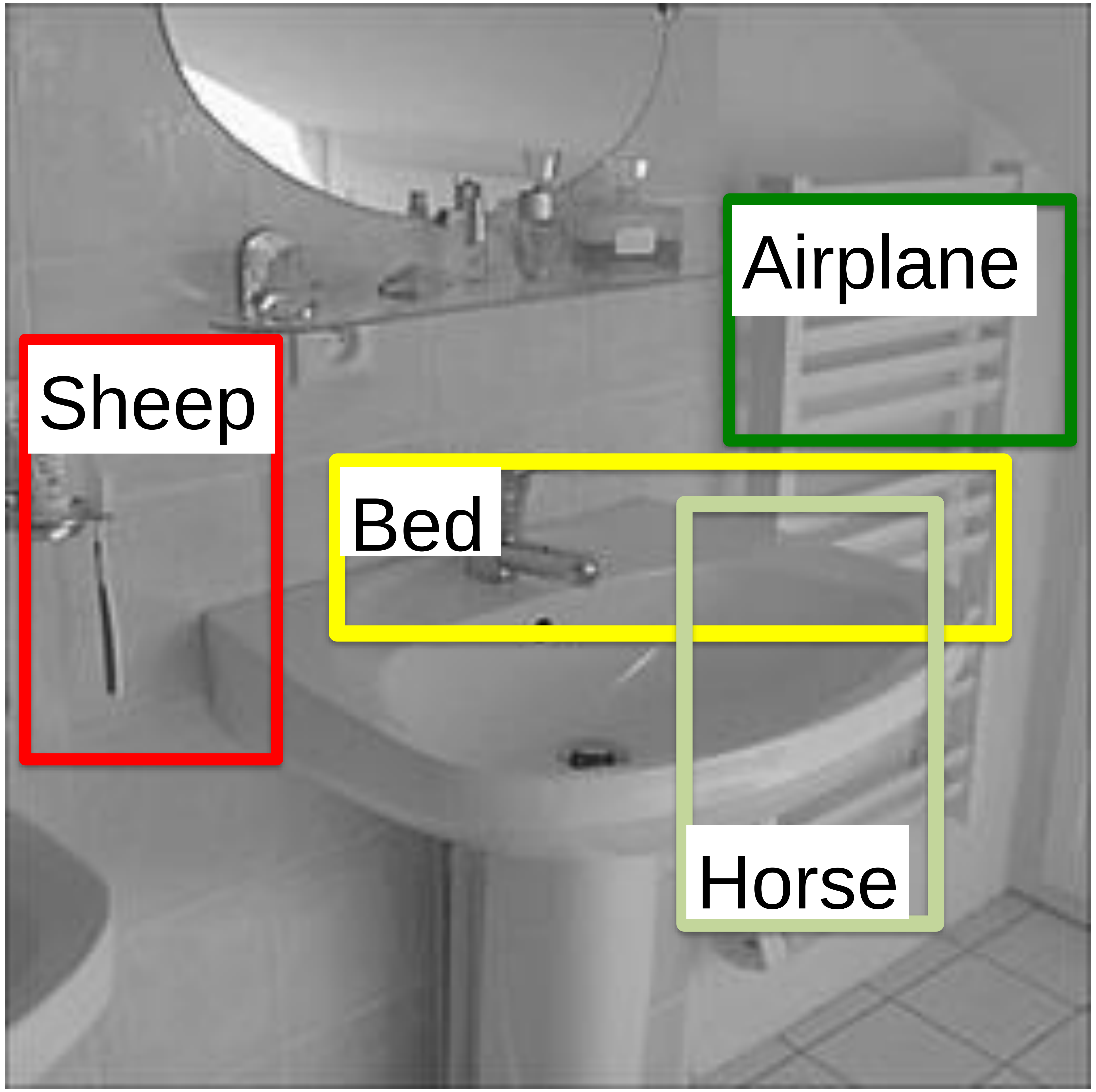
"A cup of cat"

DALL-E 2 (Open AI)

https://www.reddit.com/r/dalle2/comments/y4mygn/a_cup_of_cat/

Slide credit: Shuang Li

When I started…



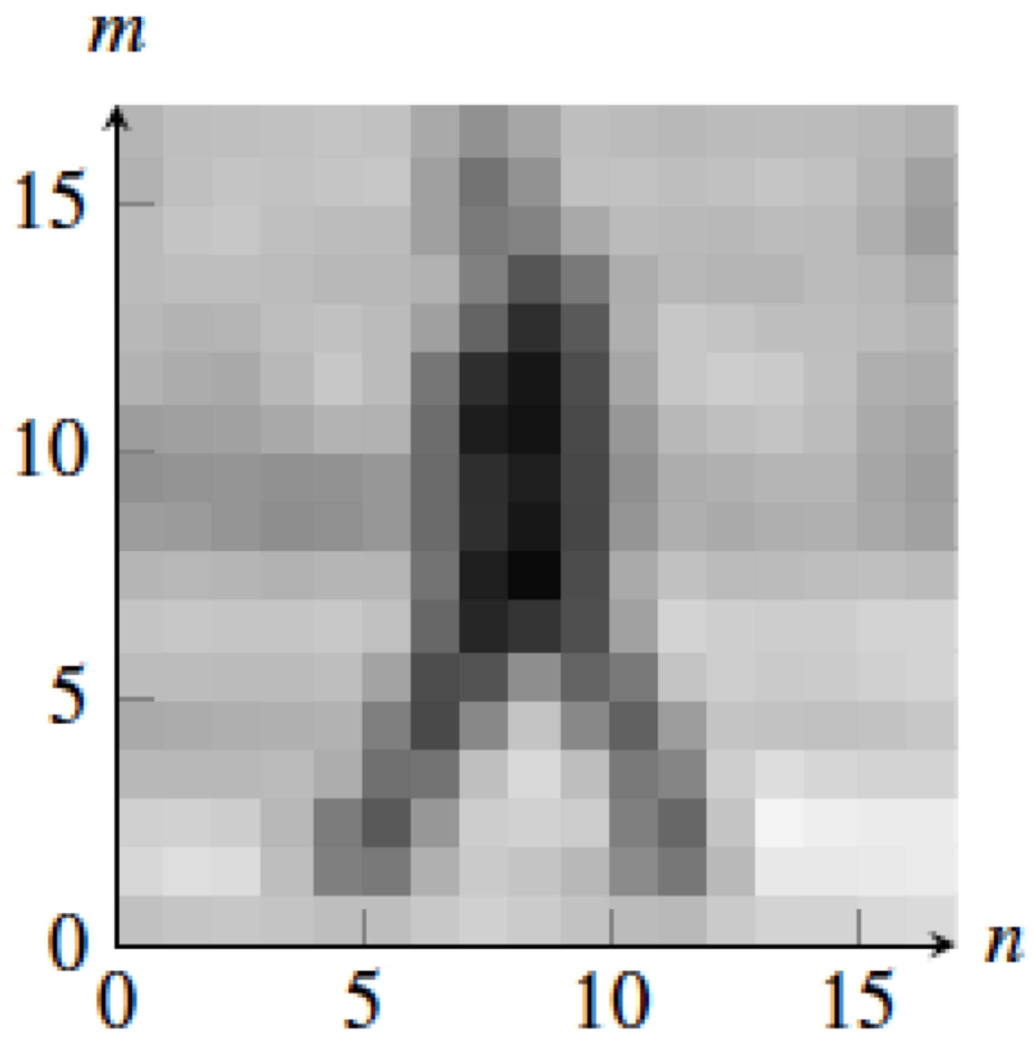
Vision is Hard

What the machine gets

$$\mathbf{I} = \begin{bmatrix}
160 & 175 & 171 & 168 & 168 & 172 & 164 & 158 & 167 & 173 & 167 & 163 & 162 & 164 & 160 & 159 & 163 & 162 \\
149 & 164 & 172 & 175 & 178 & 179 & 176 & 118 & 97 & 168 & 175 & 171 & 169 & 175 & 176 & 177 & 165 & 152 \\
161 & 166 & 182 & 171 & 170 & 177 & 175 & 116 & 109 & 169 & 177 & 173 & 168 & 175 & 175 & 159 & 153 & 123 \\
171 & 174 & 177 & 175 & 167 & 161 & 157 & 138 & 103 & 112 & 157 & 164 & 159 & 160 & 165 & 169 & 148 & 144 \\
163 & 163 & 162 & 165 & 167 & 164 & 178 & 167 & 77 & 55 & 134 & 170 & 167 & 162 & 164 & 175 & 168 & 160 \\
173 & 164 & 158 & 165 & 180 & 180 & 150 & 89 & 61 & 34 & 137 & 186 & 186 & 182 & 175 & 165 & 160 & 164 \\
152 & 155 & 146 & 147 & 169 & 180 & 163 & 51 & 24 & 32 & 119 & 163 & 175 & 182 & 181 & 162 & 148 & 153 \\
134 & 135 & 147 & 149 & 150 & 147 & 148 & 62 & 36 & 46 & 114 & 157 & 163 & 167 & 169 & 163 & 146 & 147 \\
135 & 132 & 131 & 125 & 115 & 129 & 132 & 74 & 54 & 41 & 104 & 156 & 152 & 156 & 164 & 156 & 141 & 144 \\
151 & 155 & 151 & 145 & 144 & 149 & 143 & 71 & 31 & 29 & 129 & 164 & 157 & 155 & 159 & 158 & 156 & 148 \\
172 & 174 & 178 & 177 & 177 & 181 & 174 & 54 & 21 & 29 & 136 & 190 & 180 & 179 & 176 & 184 & 187 & 182 \\
177 & 178 & 176 & 173 & 174 & 180 & 150 & 27 & 101 & 94 & 74 & 189 & 188 & 186 & 183 & 186 & 188 & 187 \\
160 & 160 & 163 & 163 & 161 & 167 & 100 & 45 & 169 & 166 & 59 & 136 & 184 & 176 & 175 & 177 & 185 & 186 \\
147 & 150 & 153 & 155 & 160 & 155 & 56 & 111 & 182 & 180 & 104 & 84 & 168 & 172 & 171 & 164 & 168 & 167 \\
184 & 182 & 178 & 175 & 179 & 133 & 86 & 191 & 201 & 204 & 191 & 79 & 172 & 220 & 217 & 205 & 209 & 200 \\
184 & 187 & 192 & 182 & 124 & 32 & 109 & 168 & 171 & 167 & 163 & 51 & 105 & 203 & 209 & 203 & 210 & 205 \\
191 & 198 & 203 & 197 & 175 & 149 & 169 & 189 & 190 & 173 & 160 & 145 & 156 & 202 & 199 & 201 & 205 & 202 \\
153 & 149 & 153 & 155 & 173 & 182 & 179 & 177 & 182 & 177 & 182 & 185 & 179 & 177 & 167 & 176 & 182 & 180
\end{bmatrix}$$

Vision is Hard

What we see



What the machine gets

$$
\mathbf{I} = \begin{bmatrix}
160 & 175 & 171 & 168 & 168 & 172 & 164 & 158 & 167 & 173 & 167 & 163 & 162 & 164 & 160 & 159 & 163 & 162 \\
149 & 164 & 172 & 175 & 178 & 179 & 176 & 118 & 97 & 168 & 175 & 171 & 169 & 175 & 176 & 177 & 165 & 152 \\
161 & 166 & 182 & 171 & 170 & 177 & 175 & 116 & 109 & 169 & 177 & 173 & 168 & 175 & 175 & 159 & 153 & 123 \\
171 & 174 & 177 & 175 & 167 & 161 & 157 & 138 & 103 & 112 & 157 & 164 & 159 & 160 & 165 & 169 & 148 & 144 \\
163 & 163 & 162 & 165 & 167 & 164 & 178 & 167 & 77 & 55 & 134 & 170 & 167 & 162 & 164 & 175 & 168 & 160 \\
173 & 164 & 158 & 165 & 180 & 180 & 150 & 89 & 61 & 34 & 137 & 186 & 186 & 182 & 175 & 165 & 160 & 164 \\
152 & 155 & 146 & 147 & 169 & 180 & 163 & 51 & 24 & 32 & 119 & 163 & 175 & 182 & 181 & 162 & 148 & 153 \\
134 & 135 & 147 & 149 & 150 & 147 & 148 & 62 & 36 & 46 & 114 & 157 & 163 & 167 & 169 & 163 & 146 & 147 \\
135 & 132 & 131 & 125 & 115 & 129 & 132 & 74 & 54 & 41 & 104 & 156 & 152 & 156 & 164 & 156 & 141 & 144 \\
151 & 155 & 151 & 145 & 144 & 149 & 143 & 71 & 31 & 29 & 129 & 164 & 157 & 155 & 159 & 158 & 156 & 148 \\
172 & 174 & 178 & 177 & 177 & 181 & 174 & 54 & 21 & 29 & 136 & 190 & 180 & 179 & 176 & 184 & 187 & 182 \\
177 & 178 & 176 & 173 & 174 & 180 & 150 & 27 & 101 & 94 & 74 & 189 & 188 & 186 & 183 & 186 & 188 & 187 \\
160 & 160 & 163 & 163 & 161 & 167 & 100 & 45 & 169 & 166 & 59 & 136 & 184 & 176 & 175 & 177 & 185 & 186 \\
147 & 150 & 153 & 155 & 160 & 155 & 56 & 111 & 182 & 180 & 104 & 84 & 168 & 172 & 171 & 164 & 168 & 167 \\
184 & 182 & 178 & 175 & 179 & 133 & 86 & 191 & 201 & 204 & 191 & 79 & 172 & 220 & 217 & 205 & 209 & 200 \\
184 & 187 & 192 & 182 & 124 & 32 & 109 & 168 & 171 & 167 & 163 & 51 & 105 & 203 & 209 & 203 & 210 & 205 \\
191 & 198 & 203 & 197 & 175 & 149 & 169 & 189 & 190 & 173 & 160 & 145 & 156 & 202 & 199 & 201 & 205 & 202 \\
153 & 149 & 153 & 155 & 173 & 182 & 179 & 177 & 182 & 177 & 182 & 185 & 179 & 177 & 167 & 176 & 182 & 180
\end{bmatrix}
$$

The camera is a measurement device, not a vision system

Let’s Imagine how Computer Thinks

Pablo Picasso
The Guitar Player (1911)

Why is vision hard?
Why is it getting easer now?

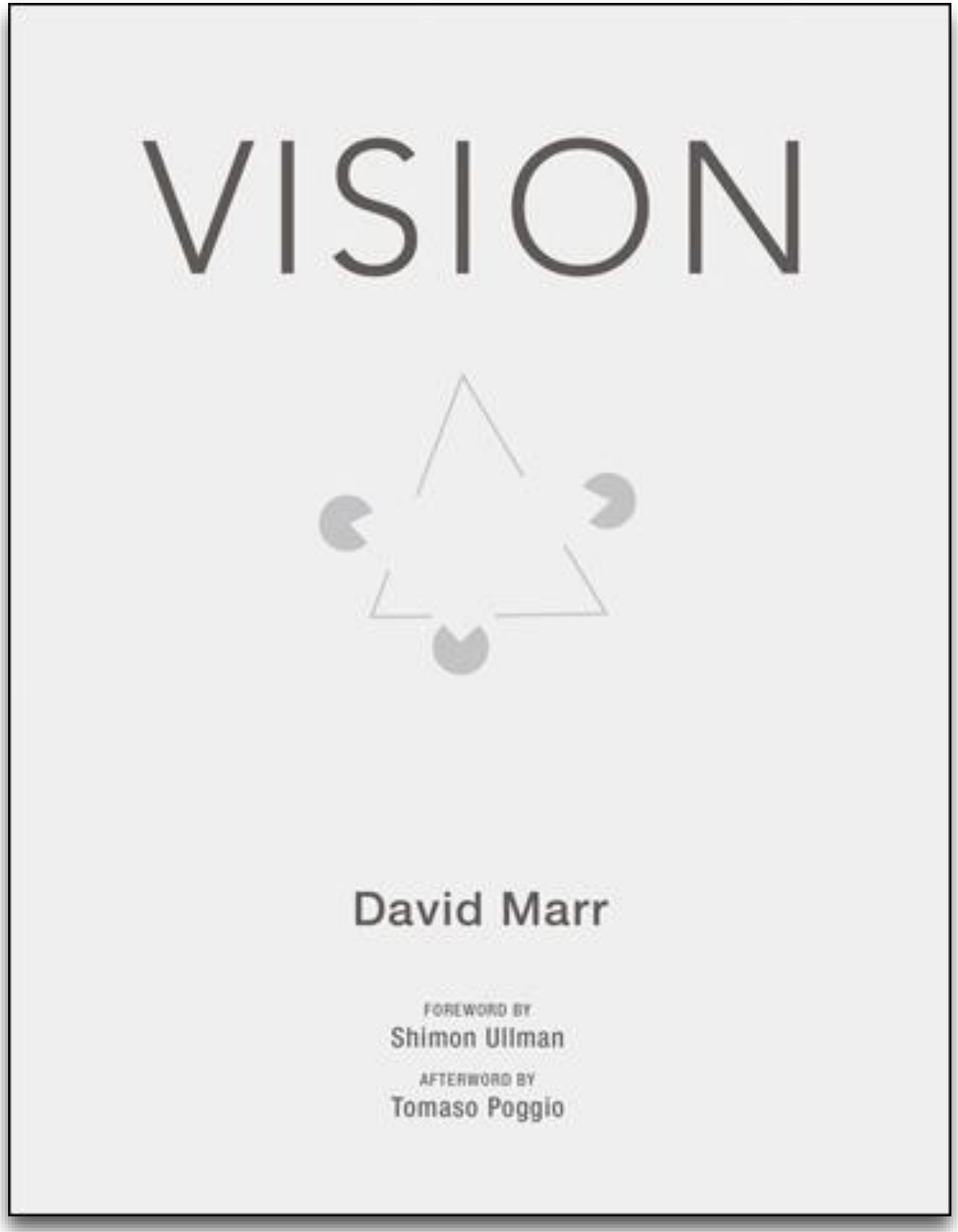
We don’t quite know. Many “axis of confusion”:

- Measurement vs. Understanding
- Given Pixels vs. Past Experience (priors)
- Algorithms vs. Data
- top-down Supervision vs. bottom-up Emergence
- Discriminative vs. Generative
- Vision is special vs. just another type of data

The Vision Story confused from the beginning…

"What does it mean, to see? The plain man's answer (and Aristotle's, too). would be, to know what is where by looking… "

"In other words, vision is the process of discovering from images what is present in the world, and where it is."



Computer Vision: a split personality

…as measurement

…as understanding

Goals: **Objective** (depth, distance, etc)

Represented by: meters, angles, 3D meshes, etc.

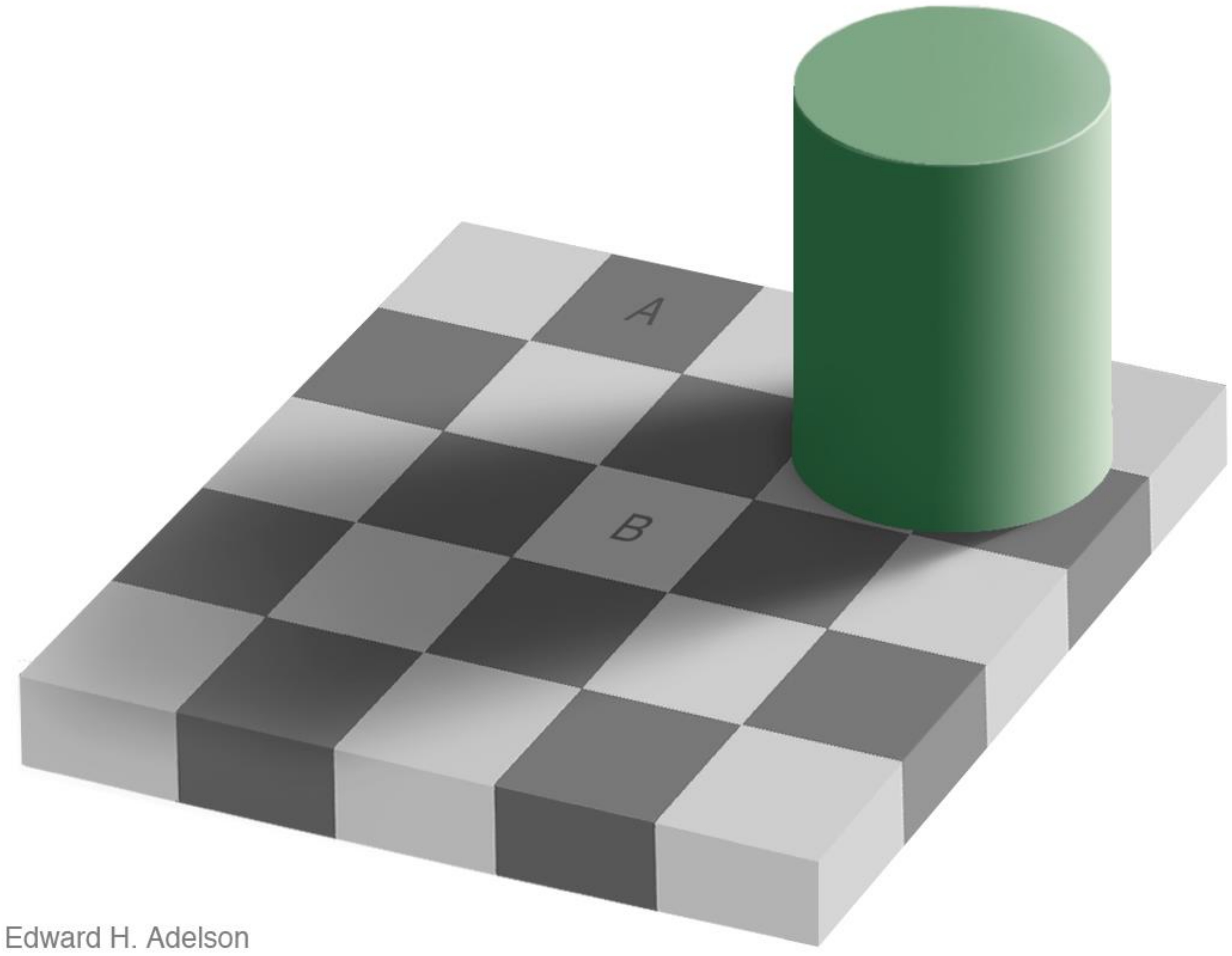
Related fields: mathematics, optics, physics, etc.

Goals: **Subjective** (objects, parts, affordances)

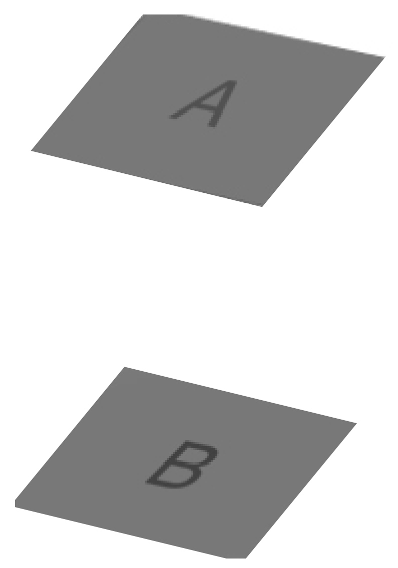
Represented by: words, human annotations, etc.

Related fields: statistics, learning, psychology, epistomology, etc.

Measurement vs. perception

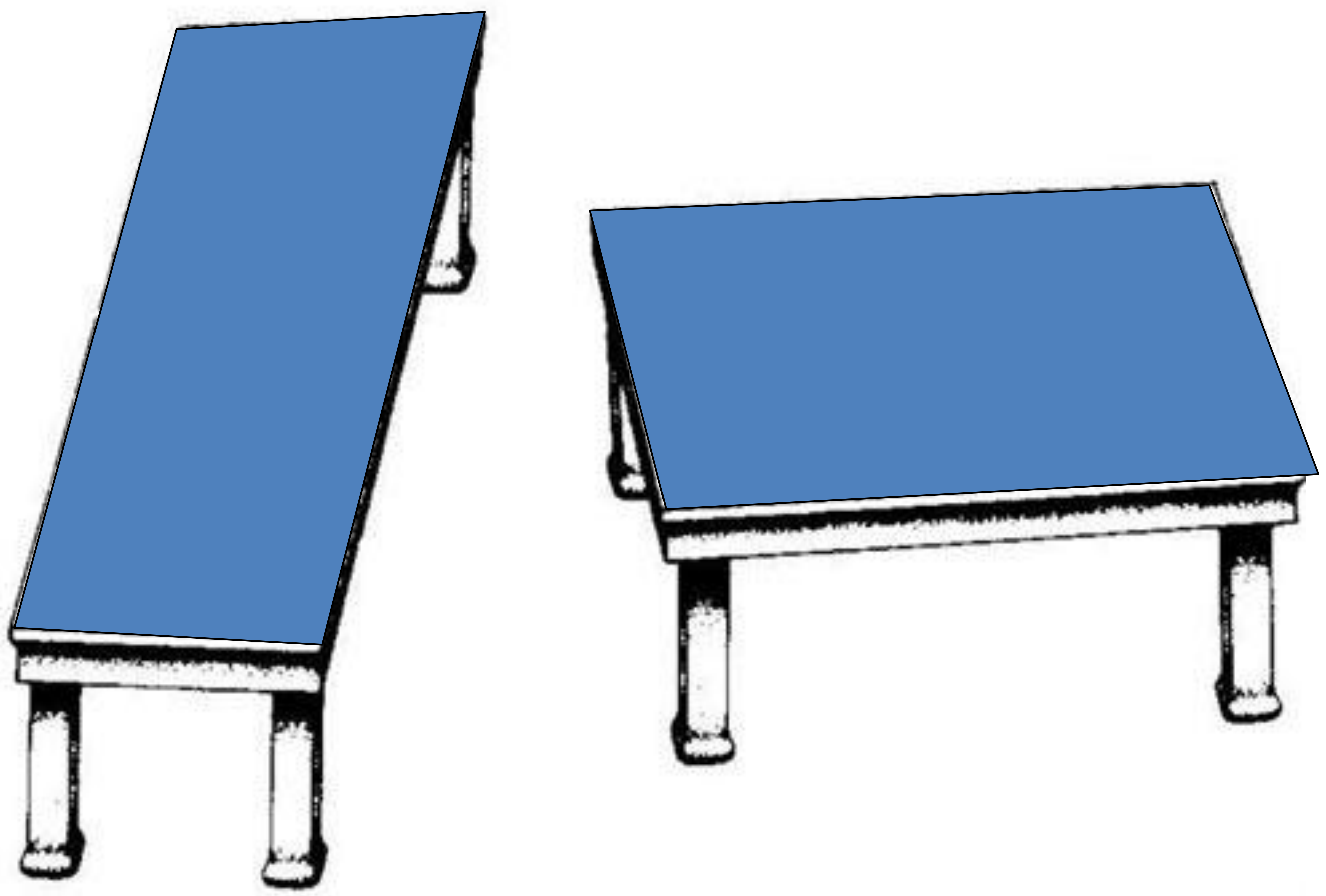


Measurement vs. perception



Measurement vs. perception

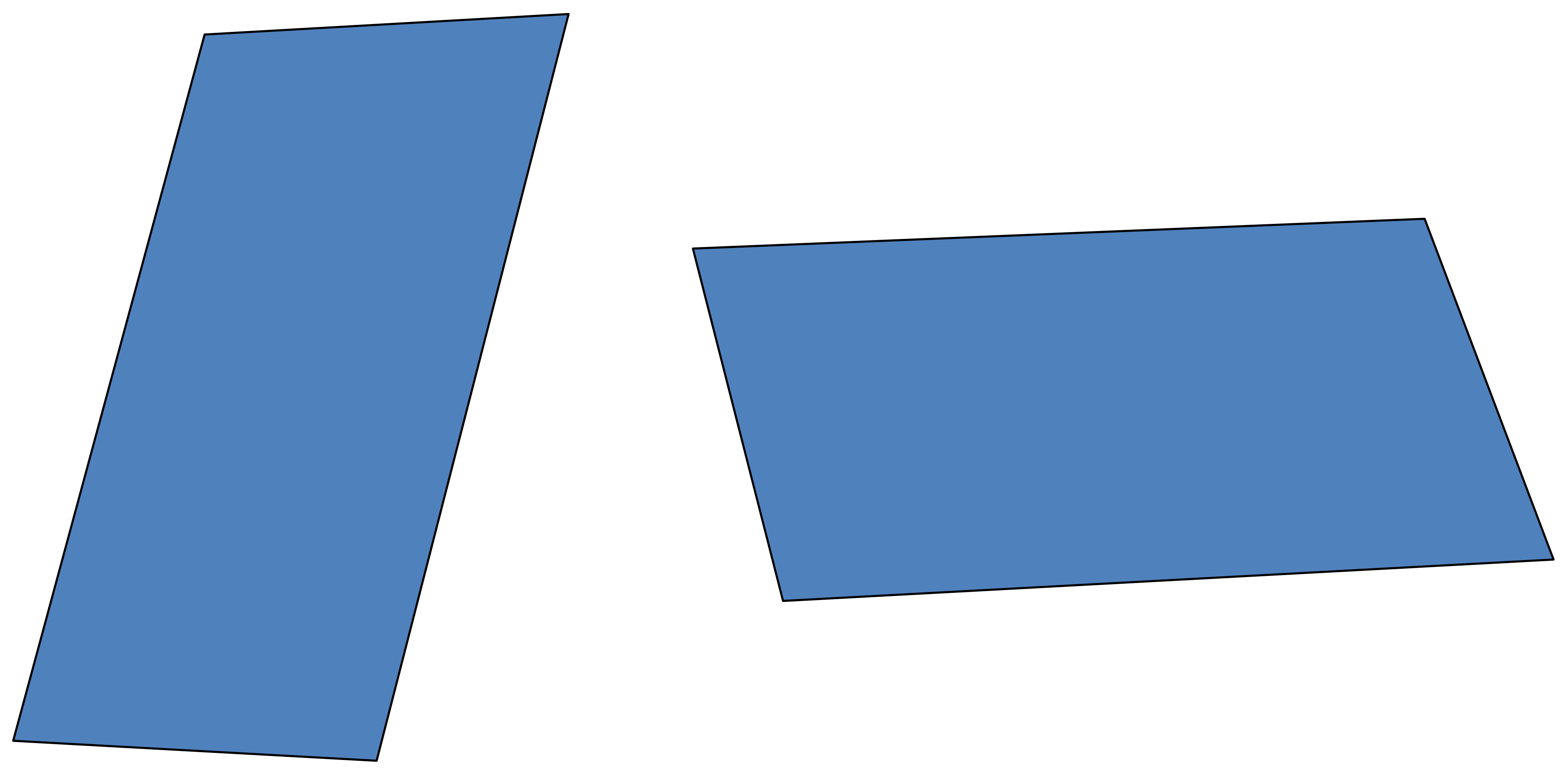
Depth processing is automatic, and we can not shut it down…



by Roger Shepard (”Turning the Tables”)

Measurement vs. perception

Depth processing is automatic, and we can not shut it down…



by Roger Shepard (”Turning the Tables”)

Measurement vs. perception

https://en.wikipedia.org/wiki/The_dress

Given Pixels vs. Past Experience

Claude Monet
Gare St.Lazare
Paris, 1877

There is almost nothing inside!

Importance of Past Experience

Claude Monet
Gare St.Lazare
Paris, 1877

Seeing less than you think…

Seeing less than you think…

Need to think “outside the box”

Seeing more than the pixels

Video by Antonio Torralba (starring Rob Fergus)

But actually…

Video by Antonio Torralba (starring Rob Fergus)

"Our perception relies
on memory as much as it does
on incoming information, which
blurs the border between
perception and cognition."

-- Moshe Bar

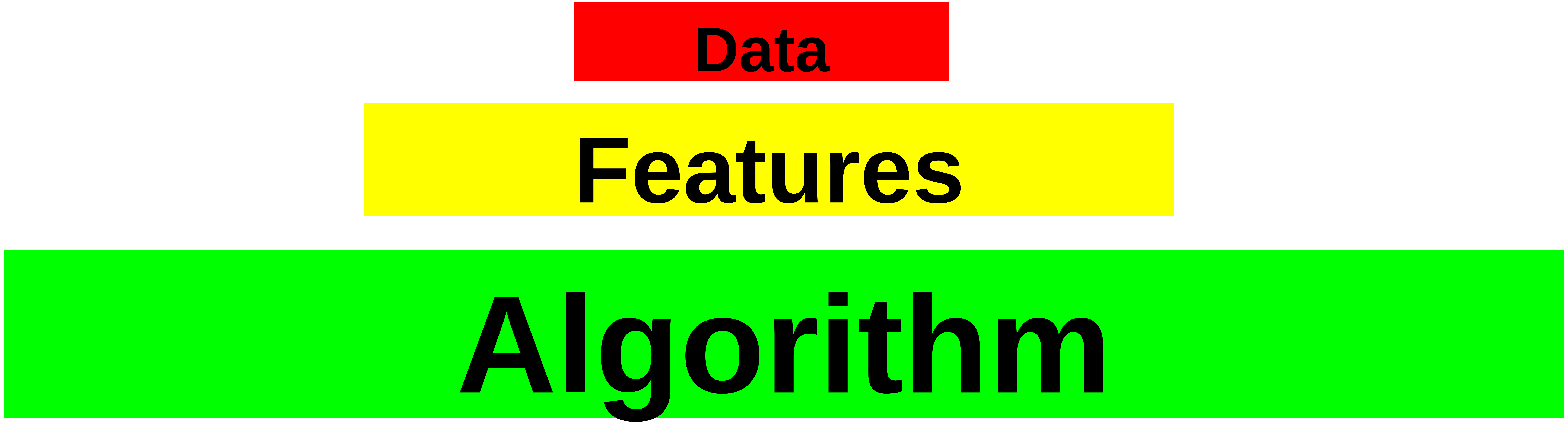
“Our perception relies on memory as much as it does on incoming information, which blurs the border between perception and cognition.”

-- Moshe Bar

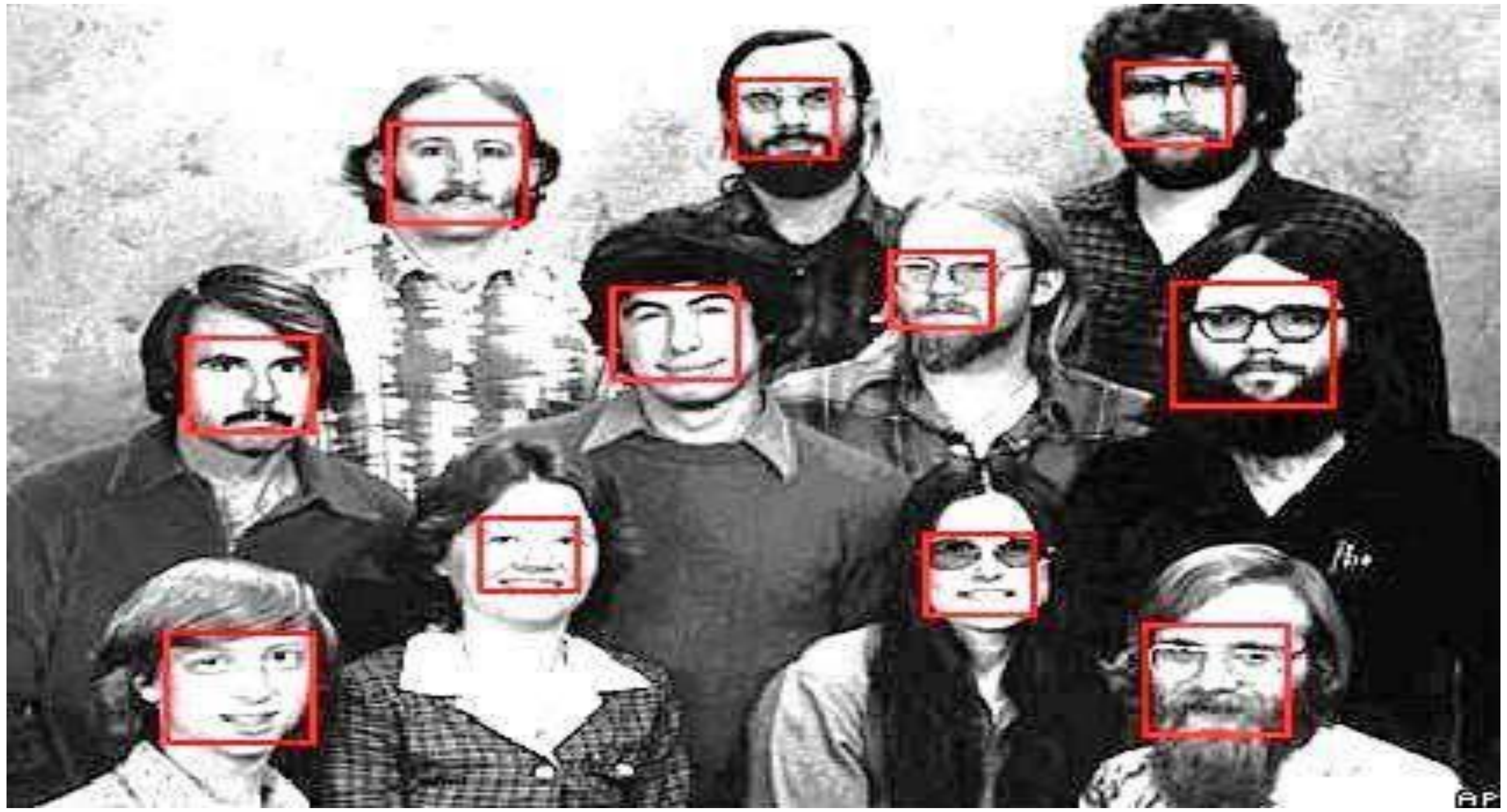
“Mind" is largely an emergent property of "data."

-- Lance Williams

Algorithms vs. Data



Vignette 1: Face Detection (late 1990s)



- Rowley, Baluja, and Kanade, 1998
  - features: **pixels**, algorithm**: neural network**
- Schniderman & Kanade, 1999
  - features: **pairs of wavelet coeff**., algorithm**: naïve Bayes**
- Viola & Jones, 2001
  - features: **haar**, algorithm: **boosted cascade**

Our Scientific Narcissism

All things being equal, we prefer to credit our own cleverness

Vignette 2: Geolocation (late 2000s)

Query Photograph

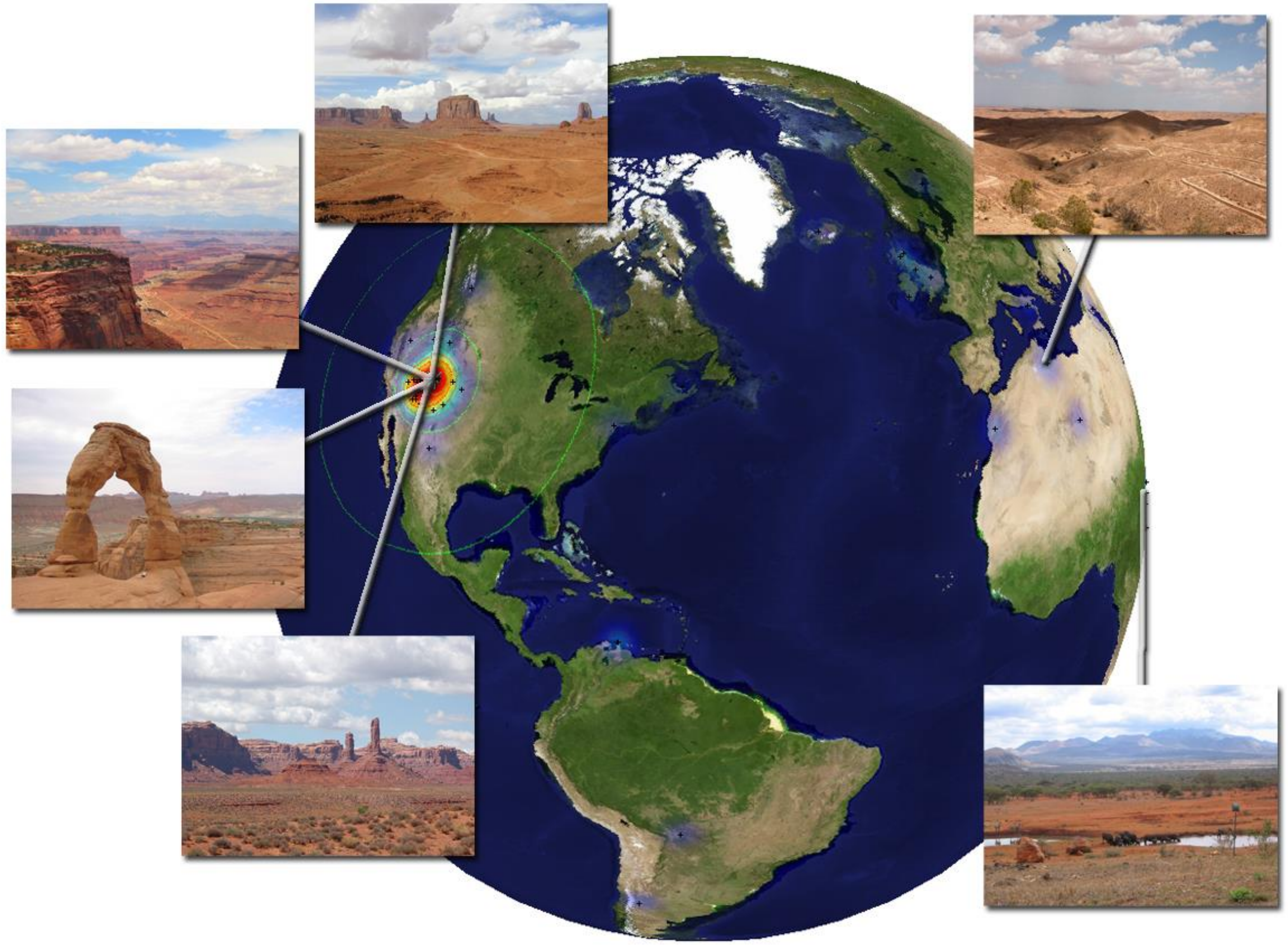
Im2gps [Hays & Efros, CVPR'08]

6 Million Flickr Images

im2GPS

(using 6 million GPS-tagged Flickr images)

Query Photograph



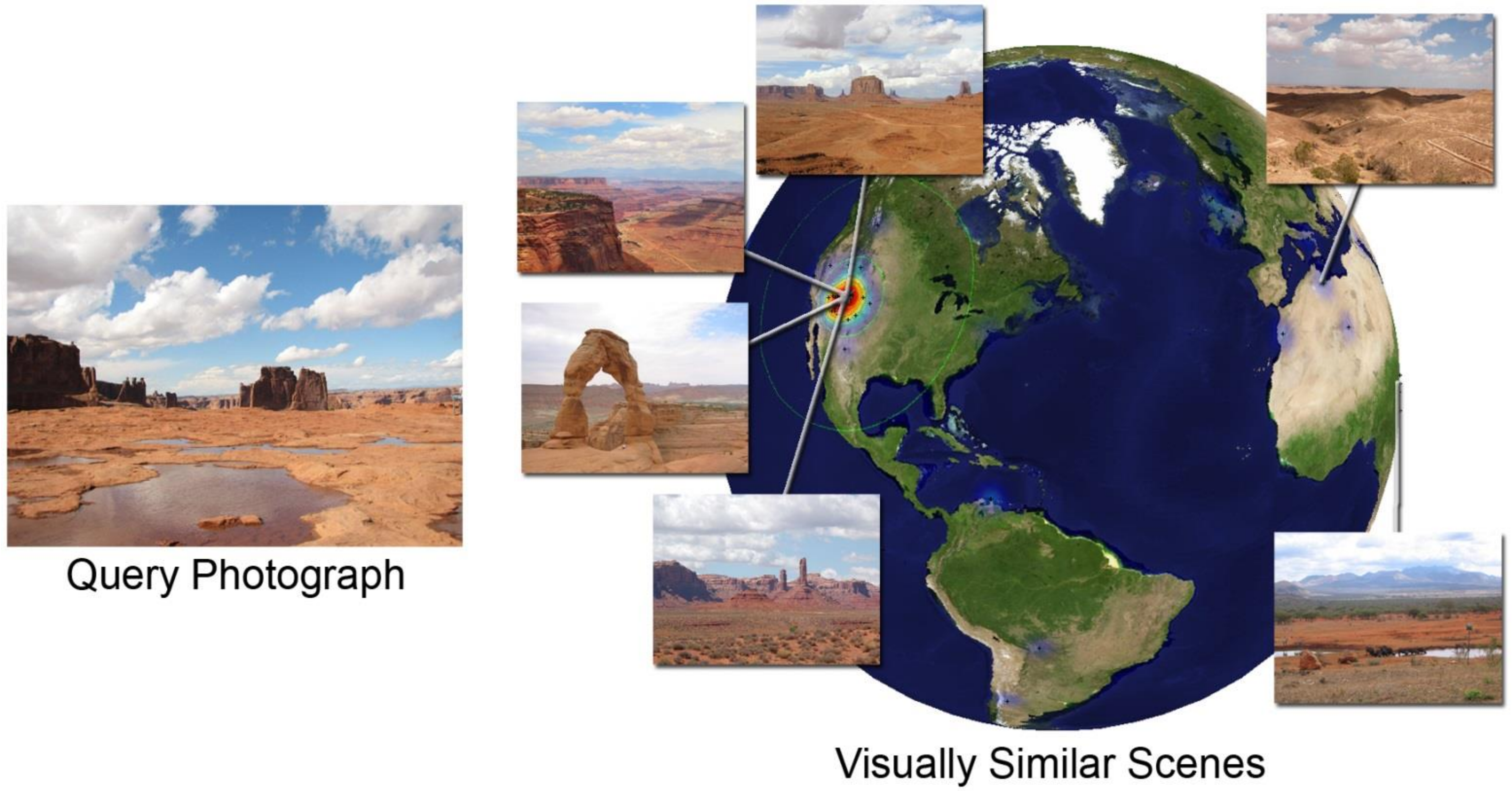
Visually Similar Scenes

Im2gps [Hays & Efros, CVPR'08]

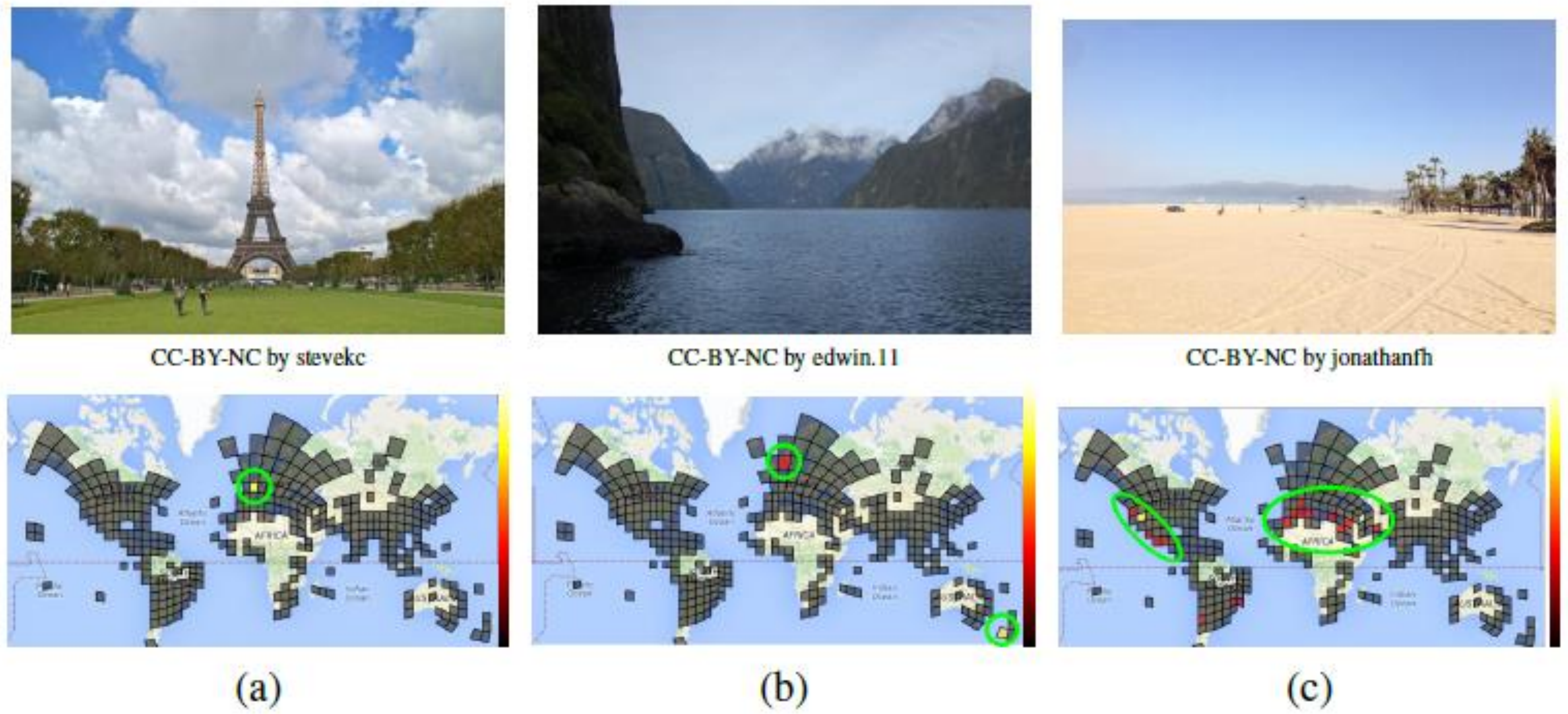
15 years later…

Algorithm vs. Data

im2gps, 2008



PlaNet, 2016



- Nearest Neighbors
- 6 million images

- Deep Net
- 91 million images

Algorithm vs. Data

Method	Street 1 km	City 25 km	Region 200 km	Country 750 km	Continent 2500 km
Im2GPS (orig) [19]		12.0%	15.0%	23.0%	47.0%
Im2GPS (new) [20]	2.5%	21.9%	32.1%	35.4%	51.9%
PlaNet (900k)	0.4%	3.8%	7.6%	21.6%	43.5%
PlaNet (6.2M)	6.3%	18.1%	30.0%	45.6%	65.8%
PlaNet (91M)	**8.4%**	**24.5%**	**37.6%**	**53.6%**	**71.3%**

Weyand et al, PlaNet, ECCV 2016

Vignette 3: Image Generation (2023)

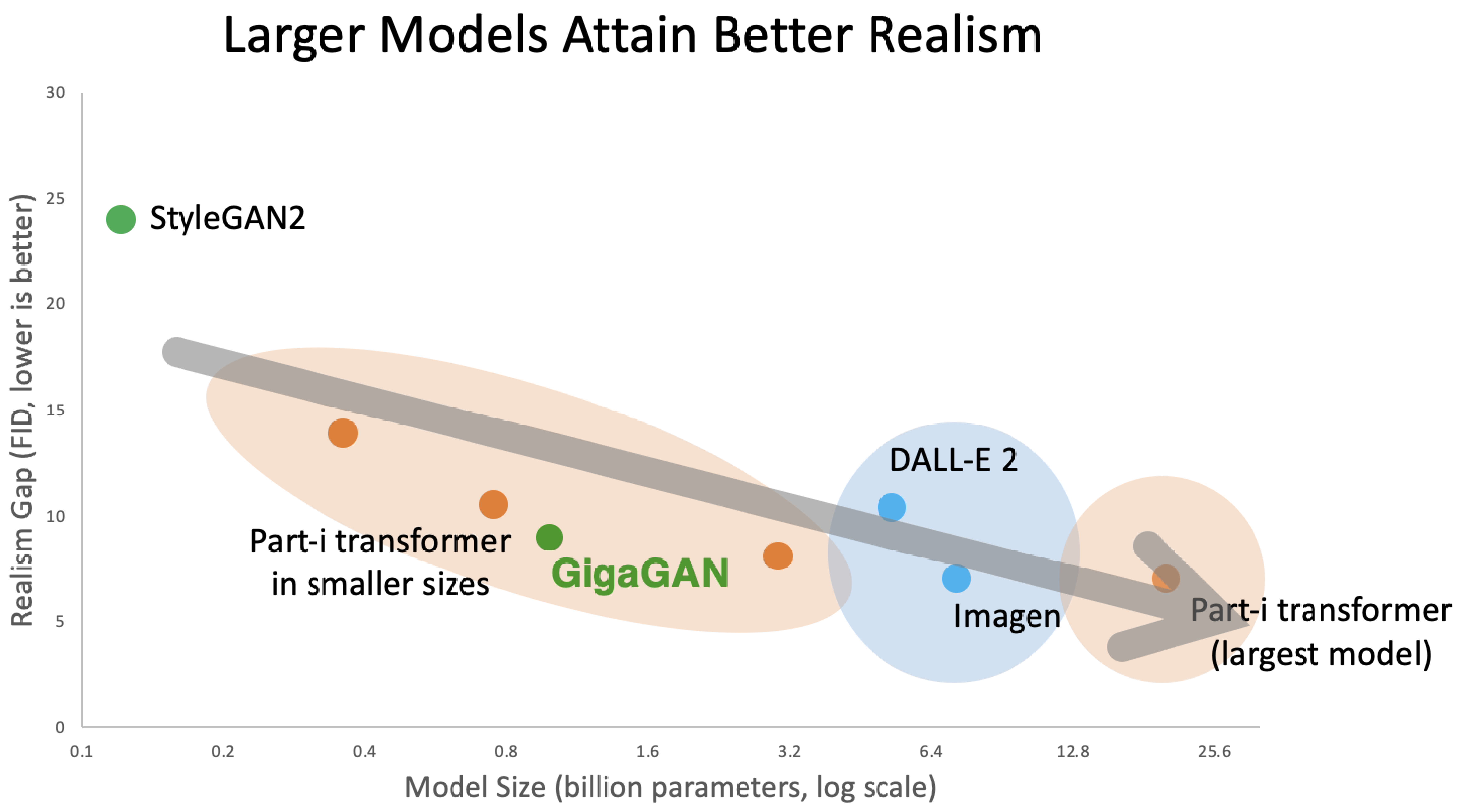
Diffusion-based

Auto-regressive

GAN-based

Prompt: *"squirrel reaching for a nut"*

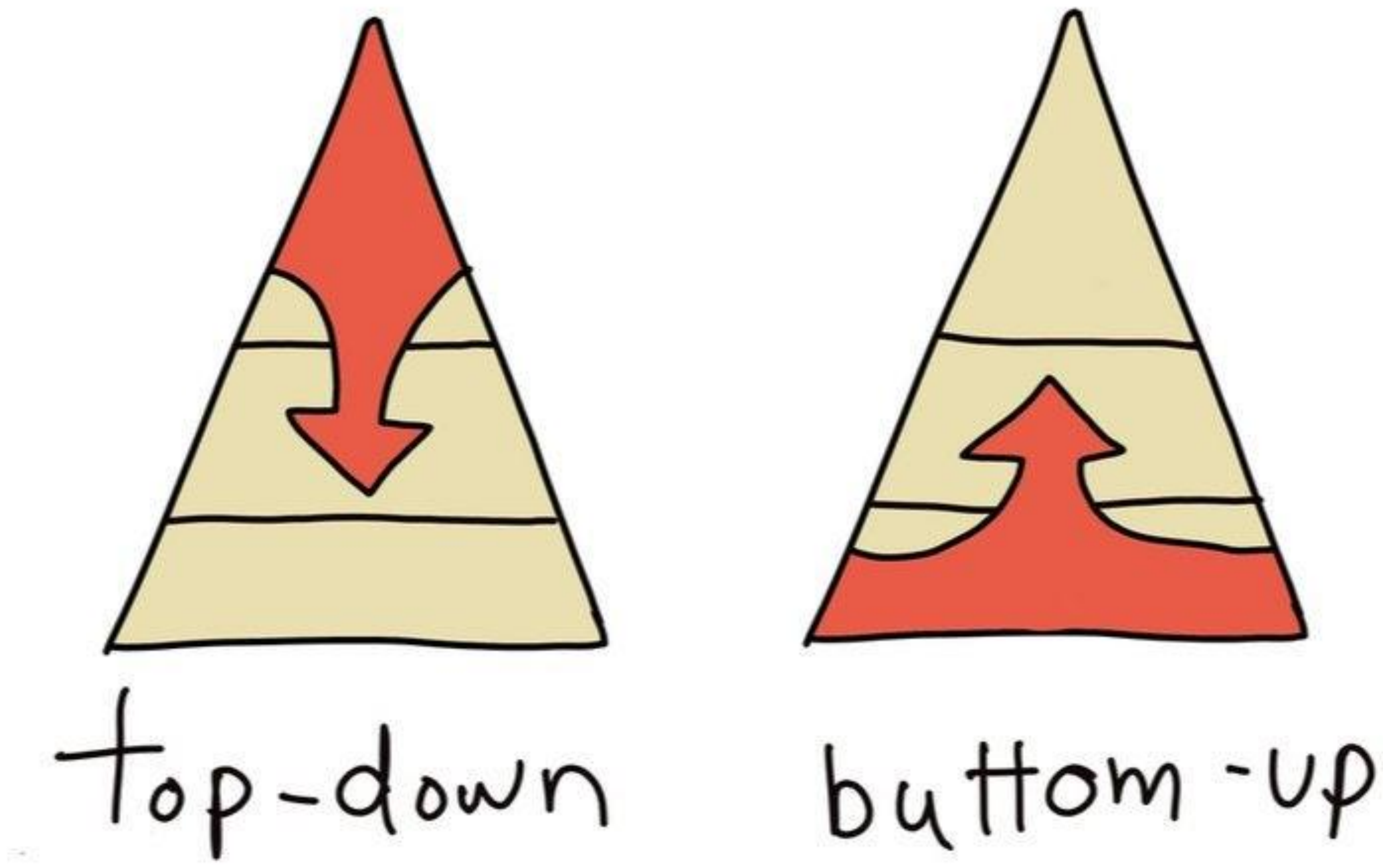
model data capacity vs. image quality



Slide by Taesung Park

Top-down Supervision vs. Bottom-up Emergence

Semantics, Language, Concepts



Pixels, sound, touch, torques, etc

Why do we have vision?

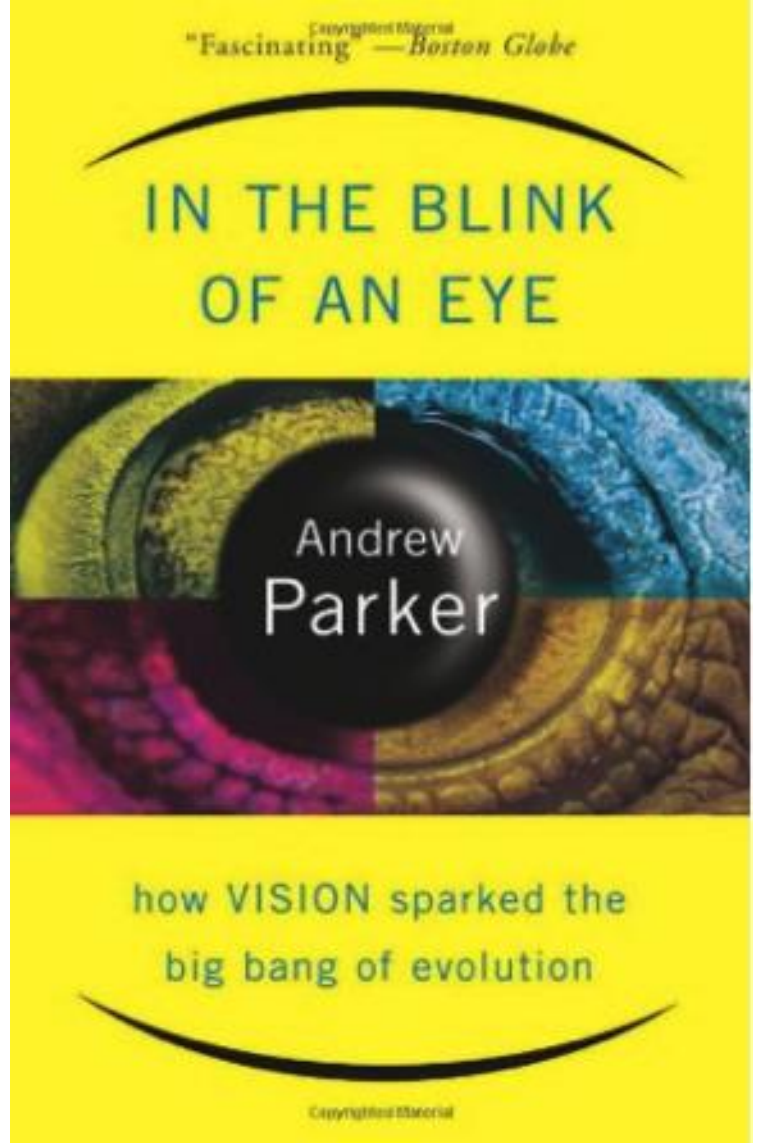
- “To see what is where by looking”
  - Aristotle, Marr, etc
- .
- .
- .
- .
- “To make babies who make babies, etc”
  - Darwin, Dawkins, etc.

Phylogeny of Intelligence

Cambrian Explosion
540 million years ago

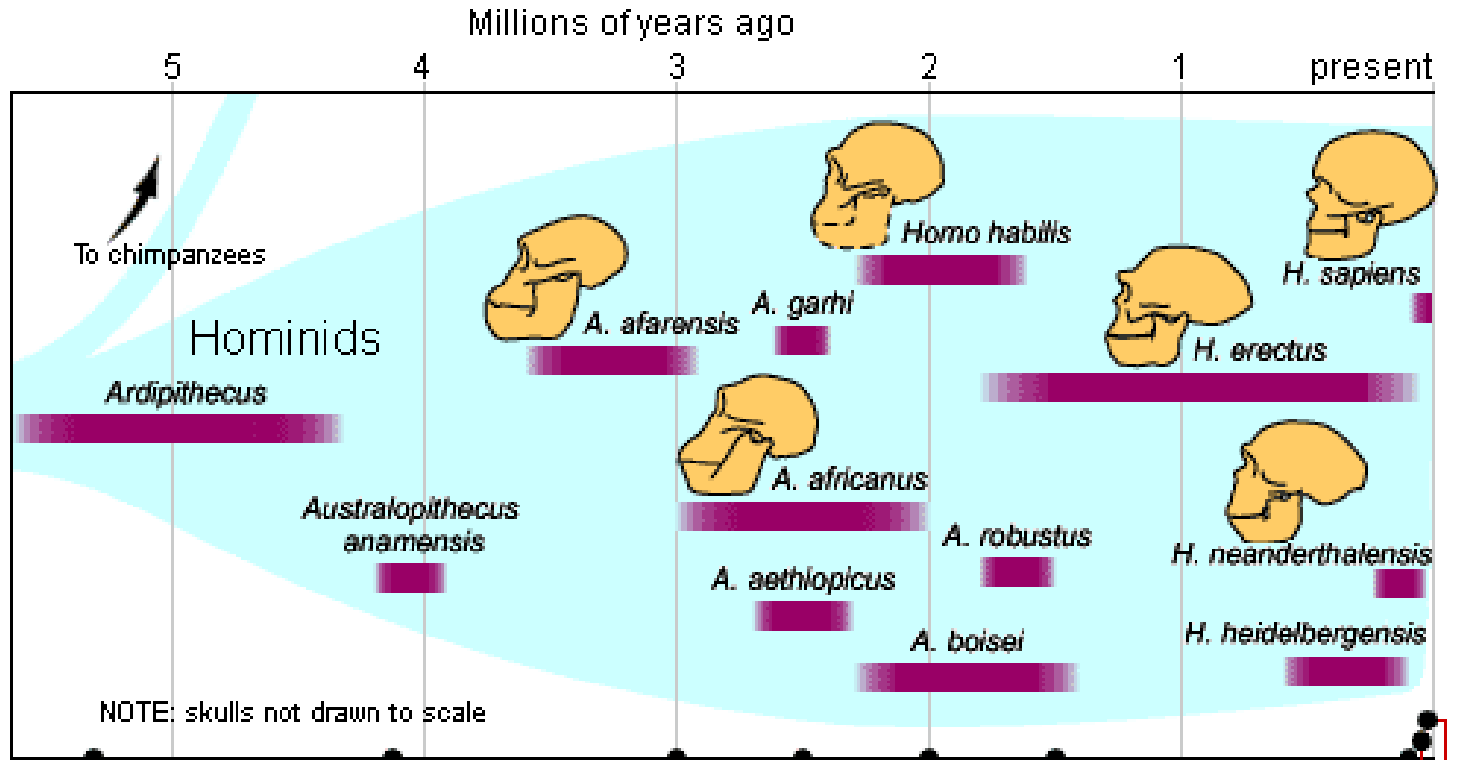
Variety of life forms, almost all phyla emerge

Animals that could see and move



Gibson: we see in order to move and we move in order to see

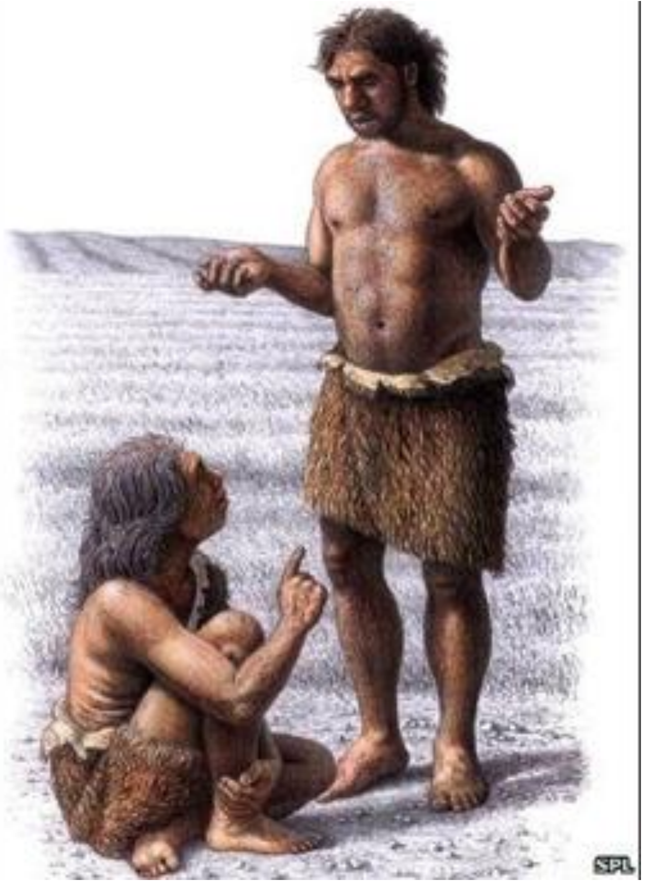
Hominid evolution, last 5 million years

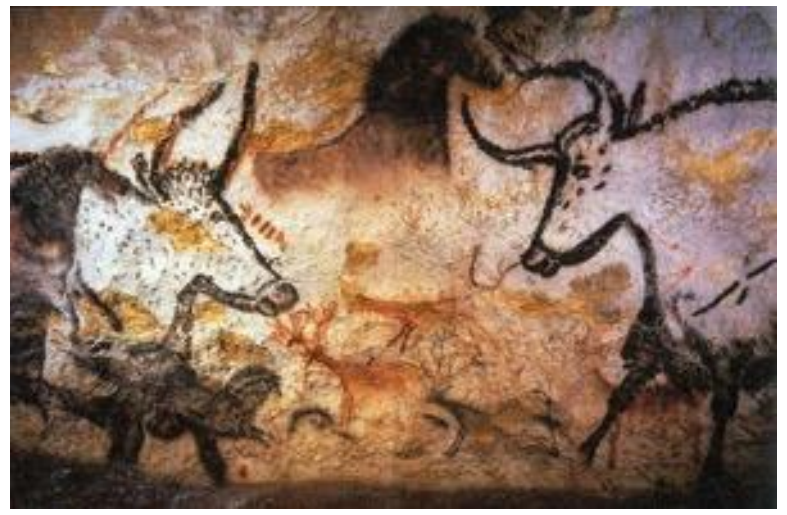


Bipedalism
Opposable thumb
Tool use

Anaxogaras: It is because of his being armed with hands that man is the most intelligent animal

Modern humans, last 50 K years





Language
Abstract thinking
Symbolic behavior

The evolutionary progression

- Vision and Locomotion
- Manipulation
- Language

Why do we have vision?

- "To see what is where by looking"
  - Aristotle, Marr, etc.
- .
- .
- "To make babies who make babies, etc"
  - Darwin, Dawkins, etc.

Why do we have vision?

- “To see what is where by looking”
  - Aristotle, Marr, etc.
- .
- “To predict the world”
  - Jakob Uexküll, Jan Koenderink, Moshe Bar, etc.
- .
- “To make babies who make babies, etc”
  - Darwin, Dawkins, etc.

Self-supervision: the world as supervision

Try to predict some aspect of the world that we interact with / have effect on:

– What’s gonna happen next?

– What’s to my left?

– What can I touch?

– What will make a sound?

– Etc.

Discriminative vs. Generative

Think of **“Zebra”**

Generative Models

Discriminative Models

Slide credit: Alyosha Efros + Lisa Dunlap + DALL-E

CS280 will (hopefully) make you think

- Measurement vs. Understanding
- Given Pixels vs. Past Experience (priors)
- Algorithms vs. Data
- top-down Supervision vs. bottom-up Emergence
- Discriminative vs. Generative
- Vision is special vs. just another type of data

POP QUIZ!

Full Credit for Participation!