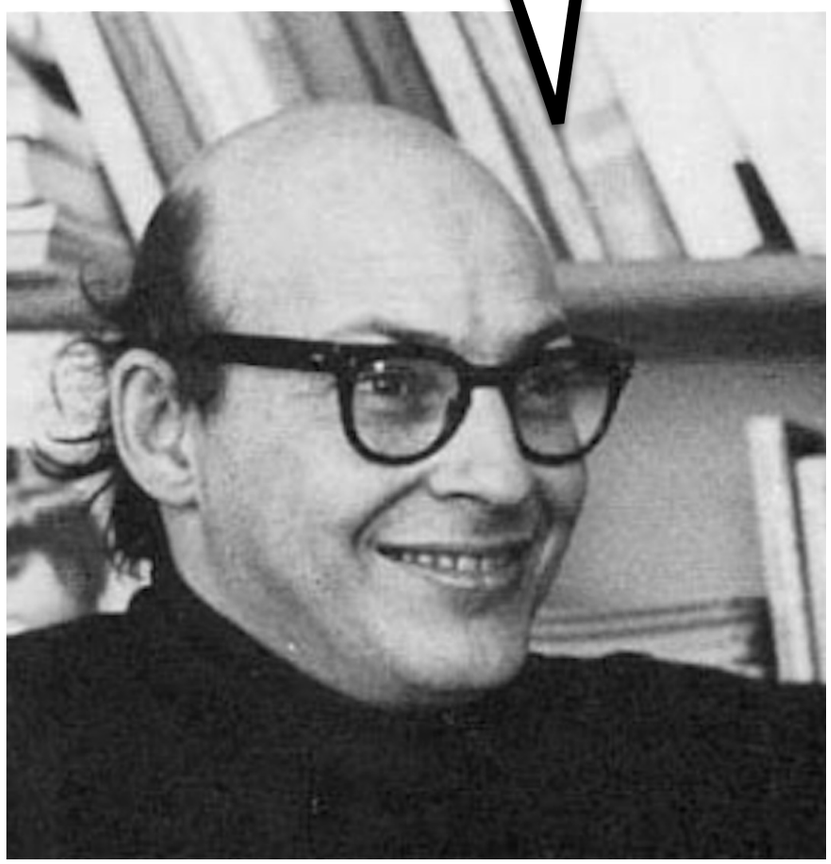
How Crowdsourcing Enabled Computer Vision

Crowdsourcing and Human Computation

Instructor: Chris Callison-Burch

Website: crowdsourcing-class.org

"Connect a television camera to a computer and **get the machine to describe what it sees.**"



MASSACHUSETTS INSTITUTE OF TECHNOLOGY

PROJECT MAC

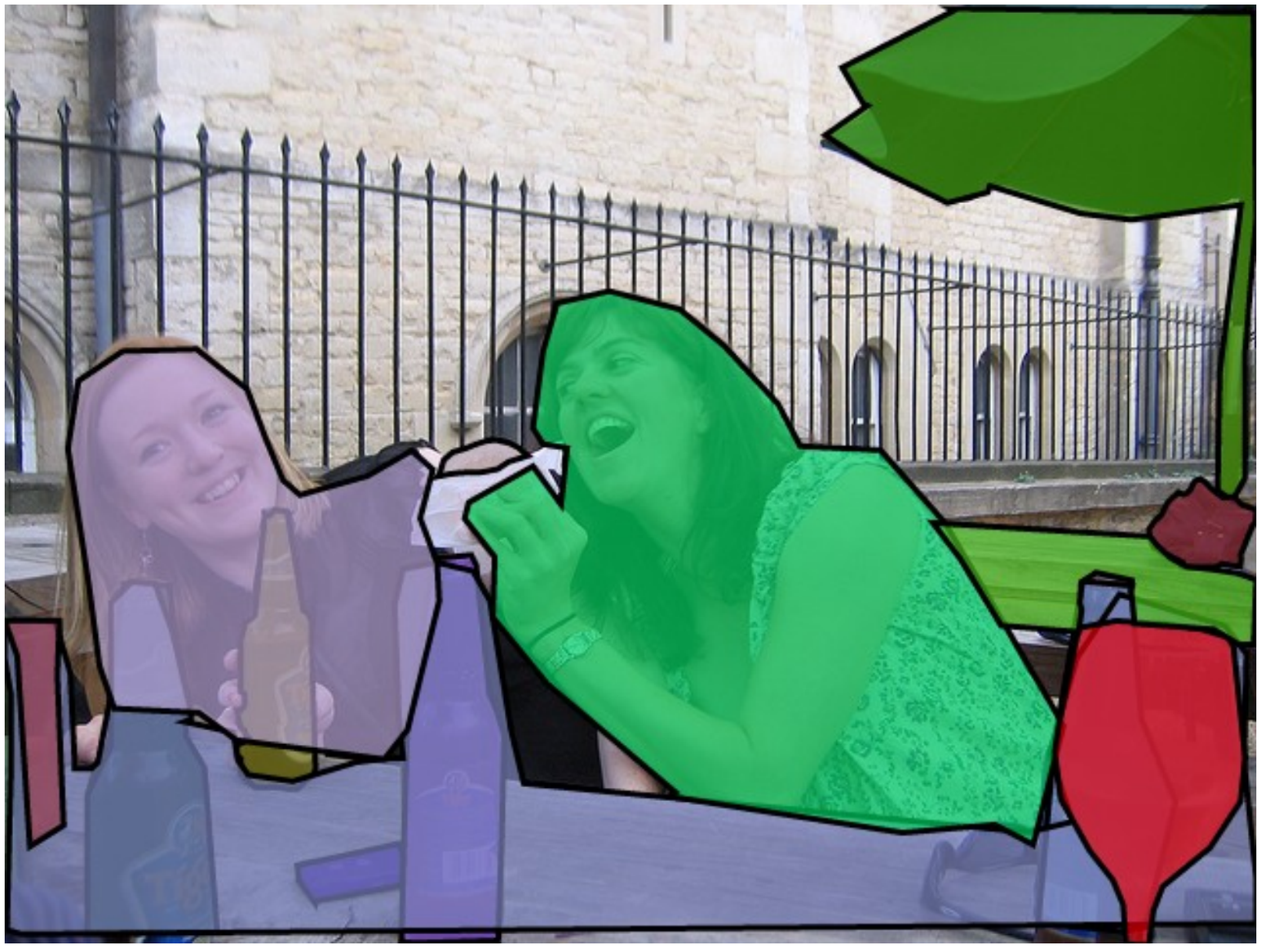
Artificial Intelligence Group — July 7, 1966
Vision Memo. No. 100.

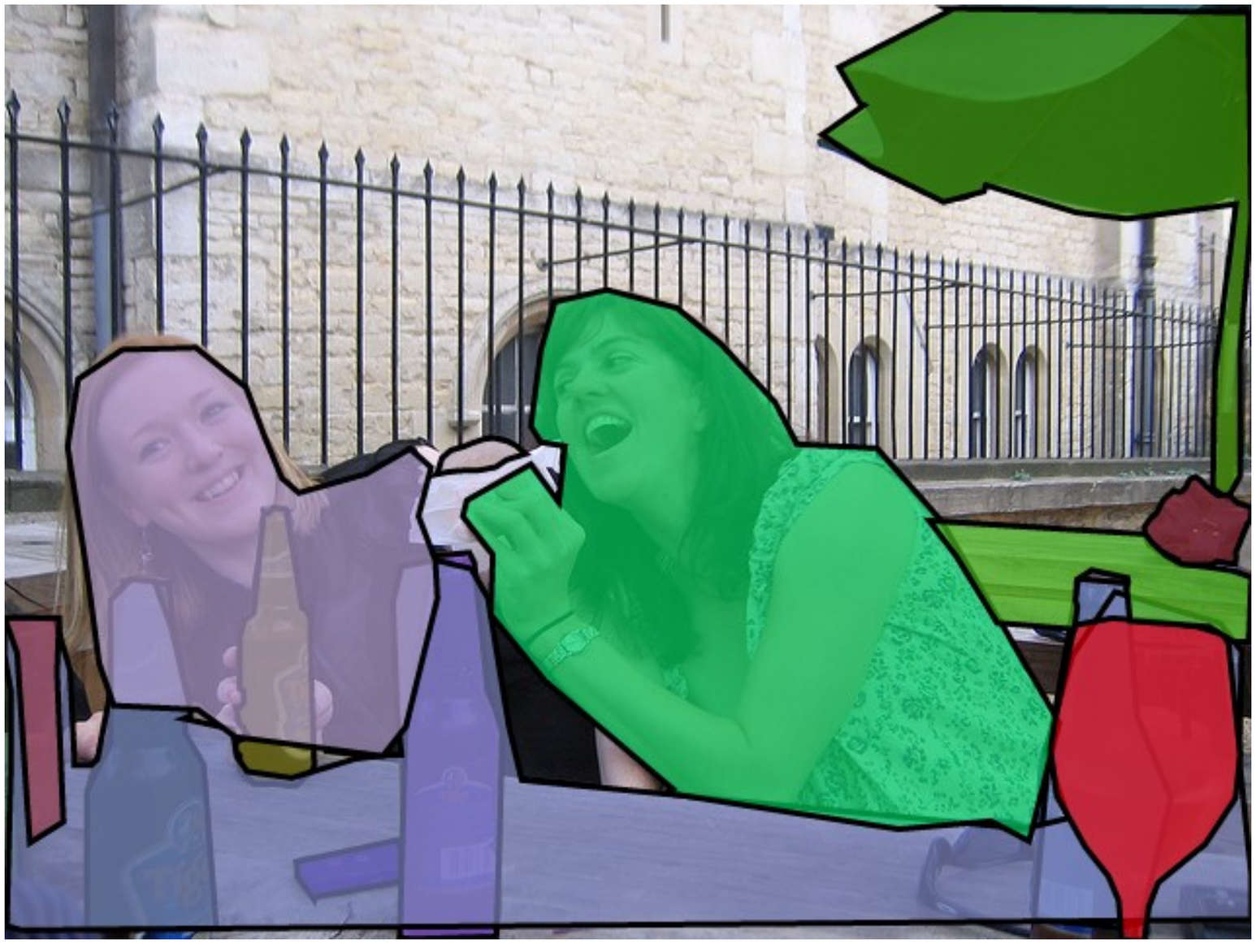
THE SUMMER VISION PROJECT

Seymour Papert

The summer vision project is an attempt to use our summer workers effectively in the construction of a significant part of a visual system. The particular task was chosen partly because it can be segmented into sub-problems which will allow individuals to work independently and yet participate in the construction of a system complex enough to be a real landmark in the development of "pattern recognition".


Tiger
Tiger
Tiger
Tiger

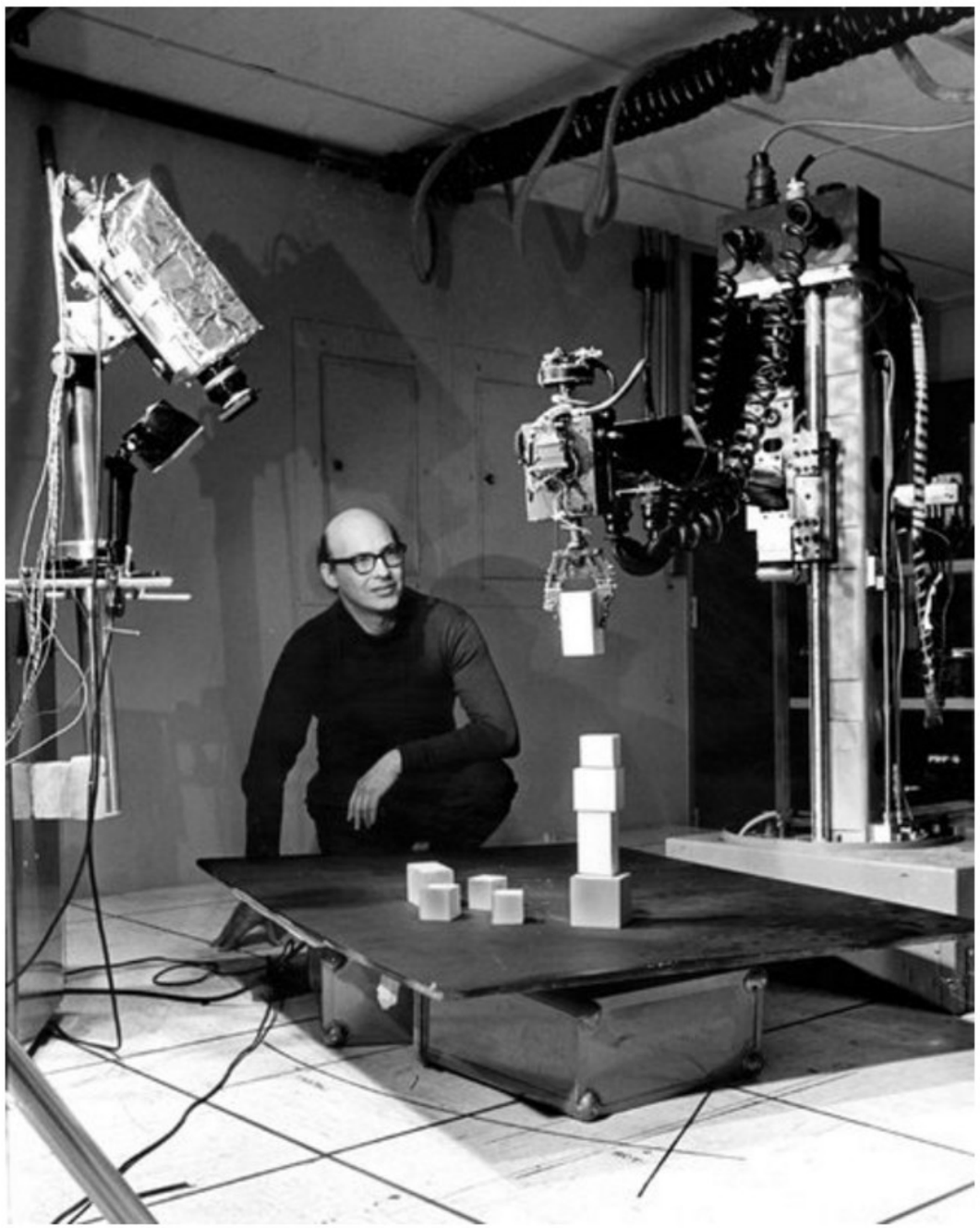




two girls sitting at a table smiling and eating and drinking.
a woman is eating a doughnut and drinking beer.
there are two woman drinking beers and eating food
a woman leaning into another woman as she holds a sandwich towards her.
two ladies are enjoying beer and treats at the table.

Marvin Minsky, Pioneer in Artificial Intelligence, Dies at 88

By GLENN RIFKIN JAN. 25, 2016



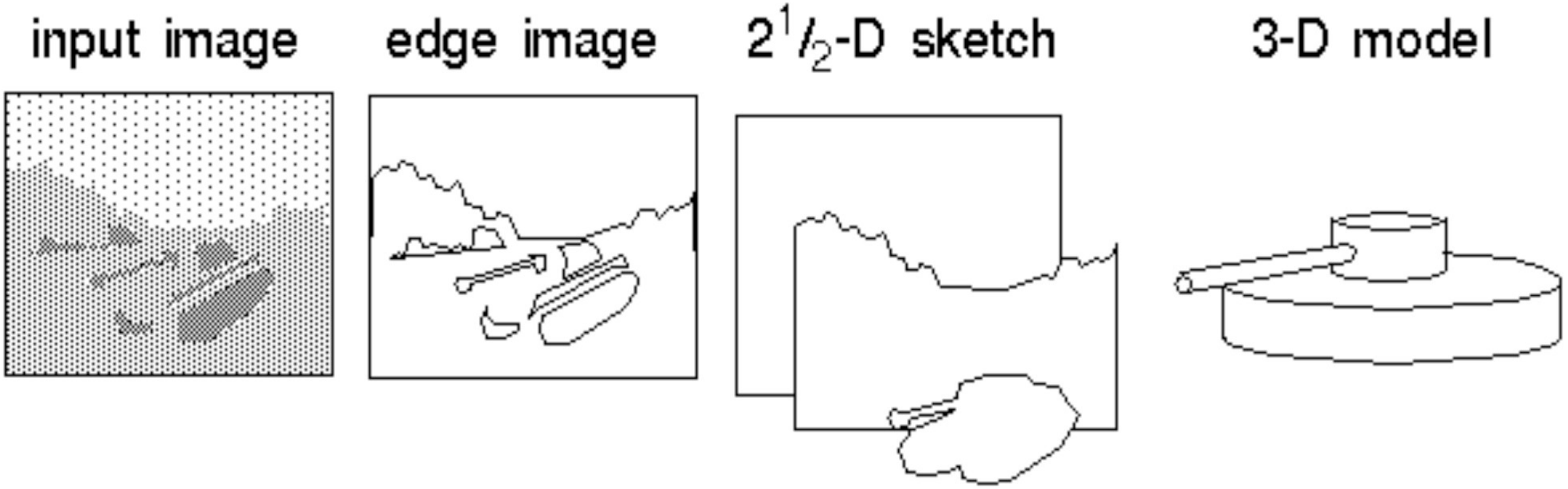
Marvin Minsky in a lab at M.I.T. in 1968. M.I.T.

Marvin Minsky, who combined a scientist's thirst for knowledge with a philosopher's quest for truth as a pioneering explorer of artificial intelligence, work that helped inspire the creation of the personal computer and the Internet, died on Sunday night in Boston. He was 88.

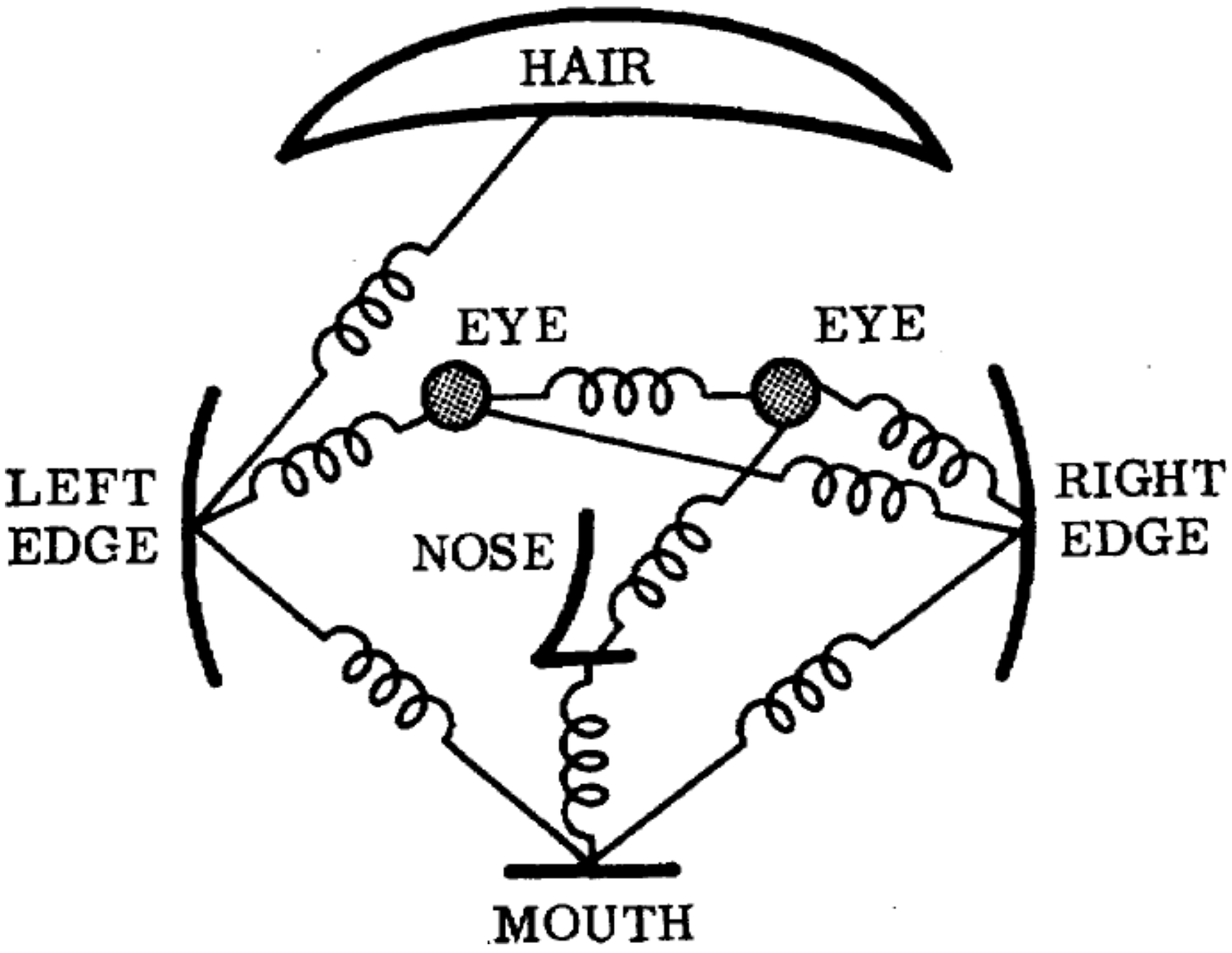
His family said the cause was a cerebral hemorrhage.

Well before the advent of the microprocessor and the supercomputer, Professor Minsky, a revered computer science educator at M.I.T., laid the foundation for the field of artificial intelligence by demonstrating the possibilities of imparting common-sense reasoning to computers.

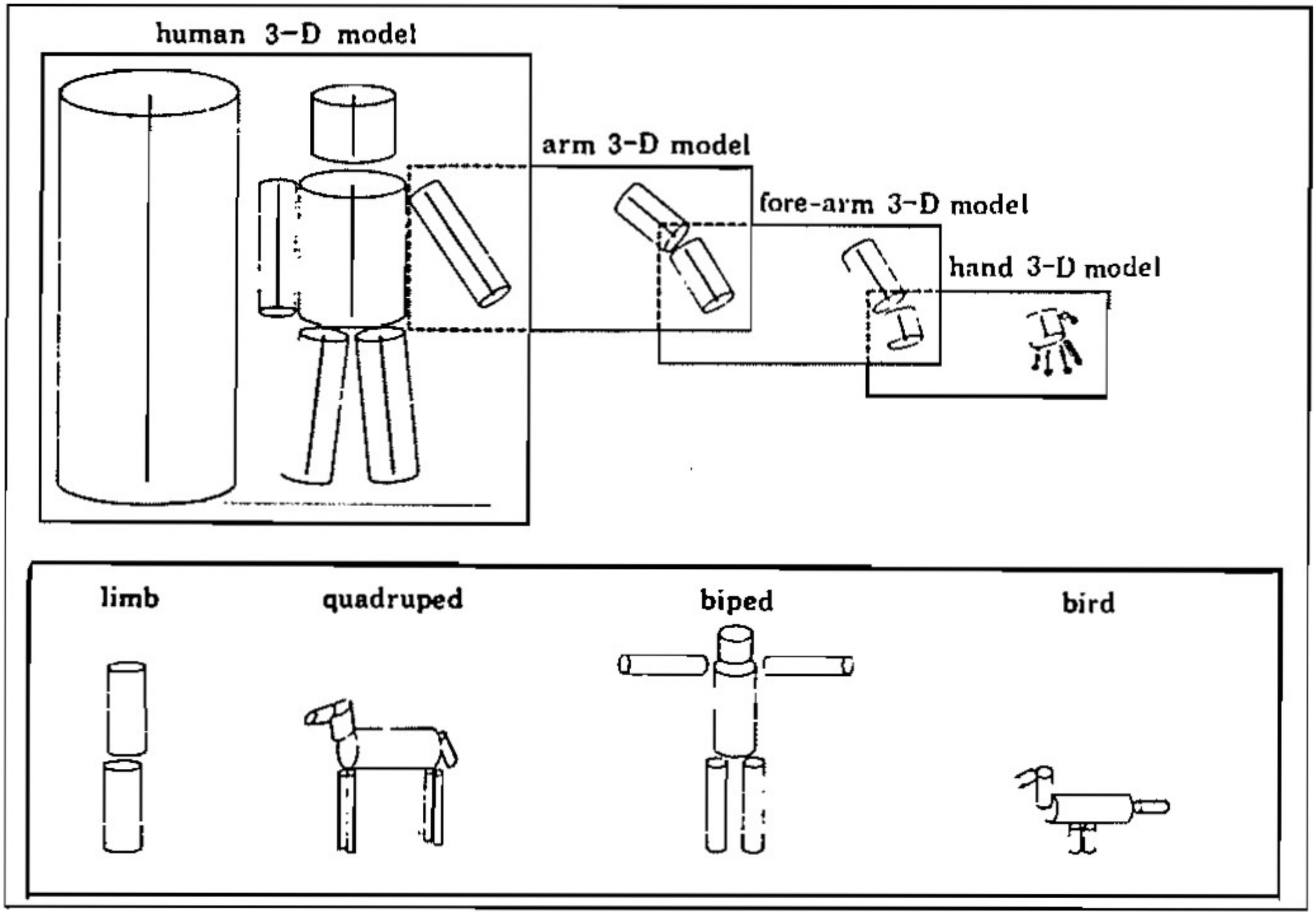
"Marvin was one of the very few people in computing whose visions and perspectives liberated the computer



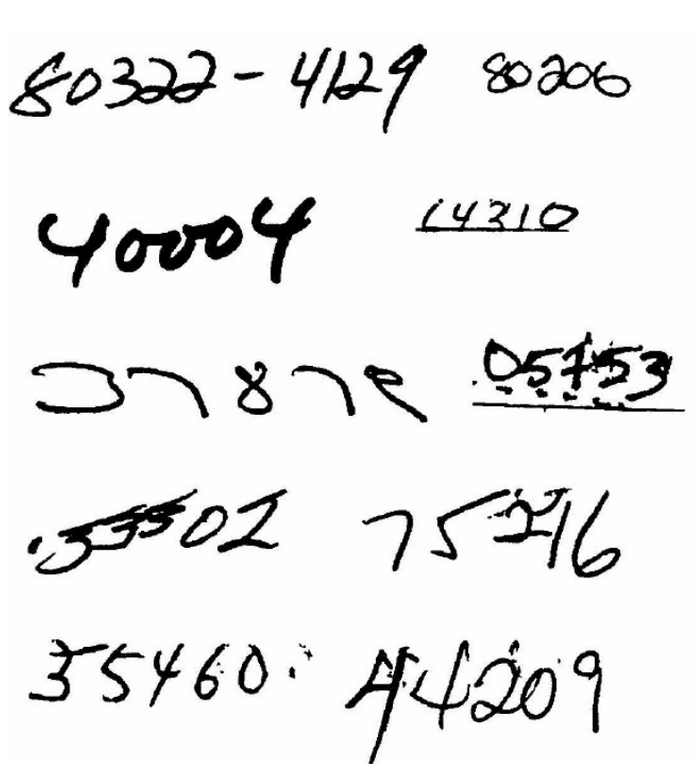
Stages of Visual Representation, David Marr, 1970

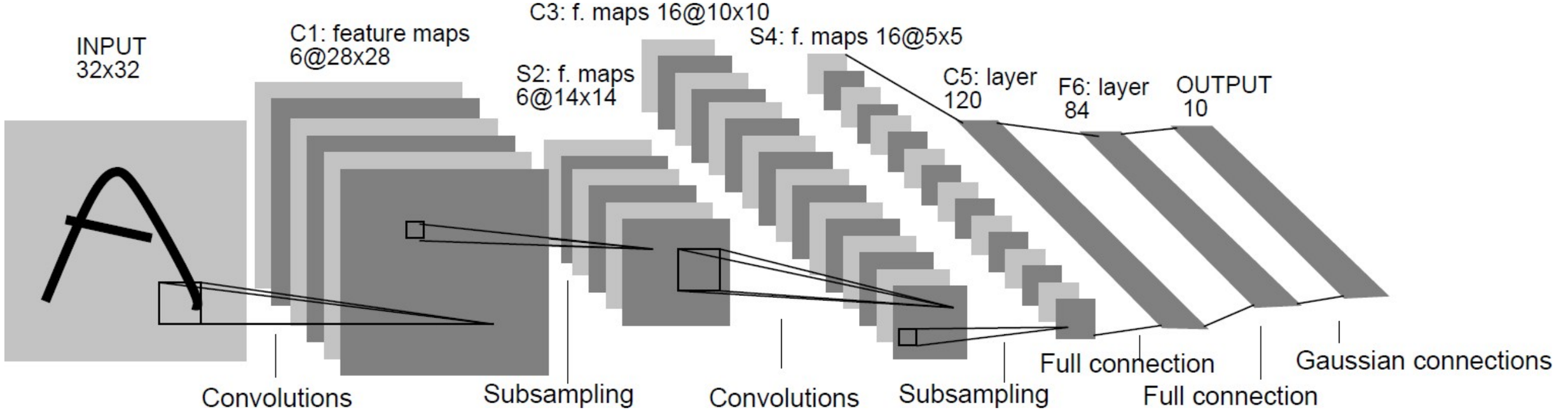


The representation and matching of pictorial structures,
Fischler and Elschlager, 1973

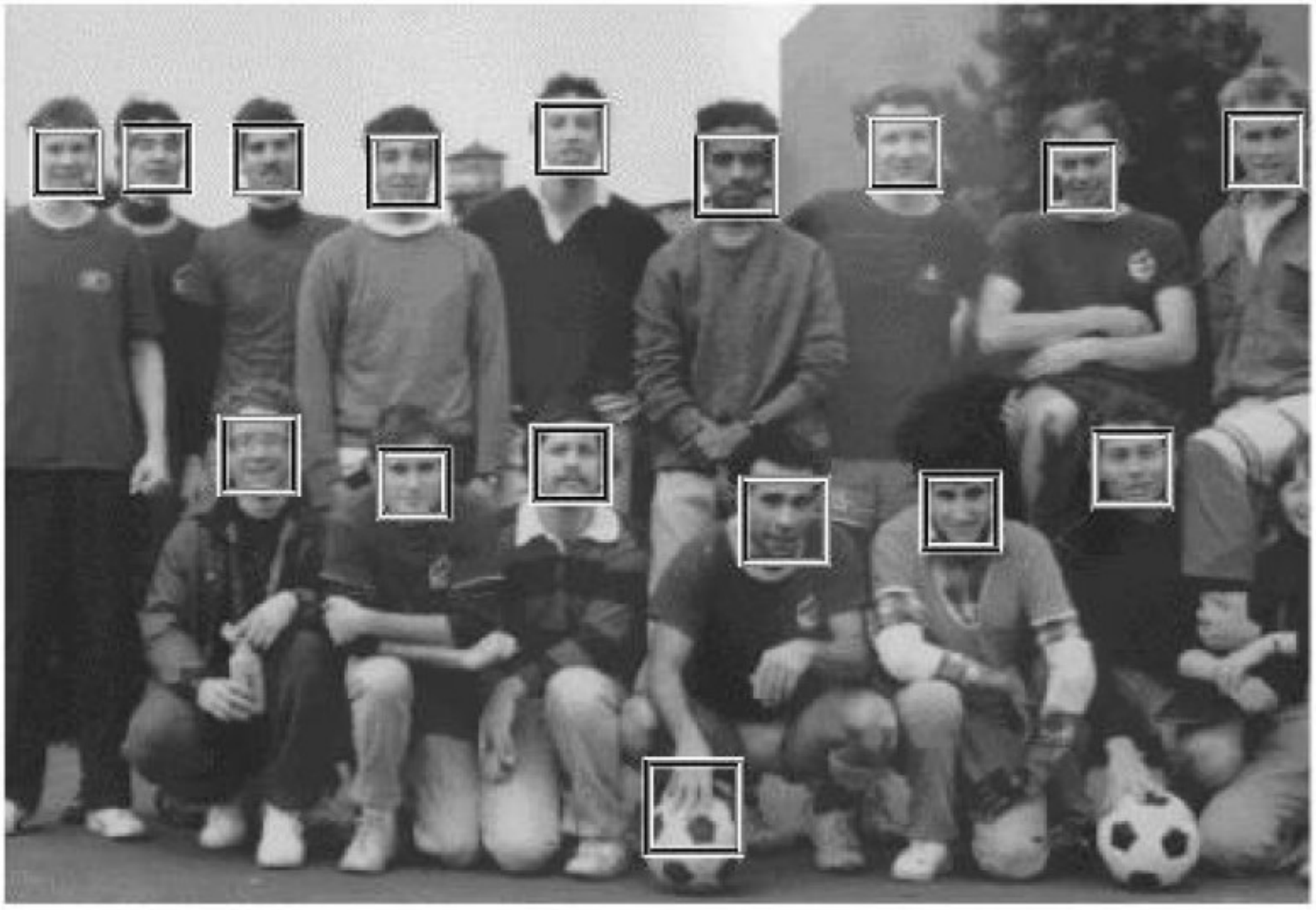


Perceptual organization and the representation of natural form
Alex Pentland, 1986

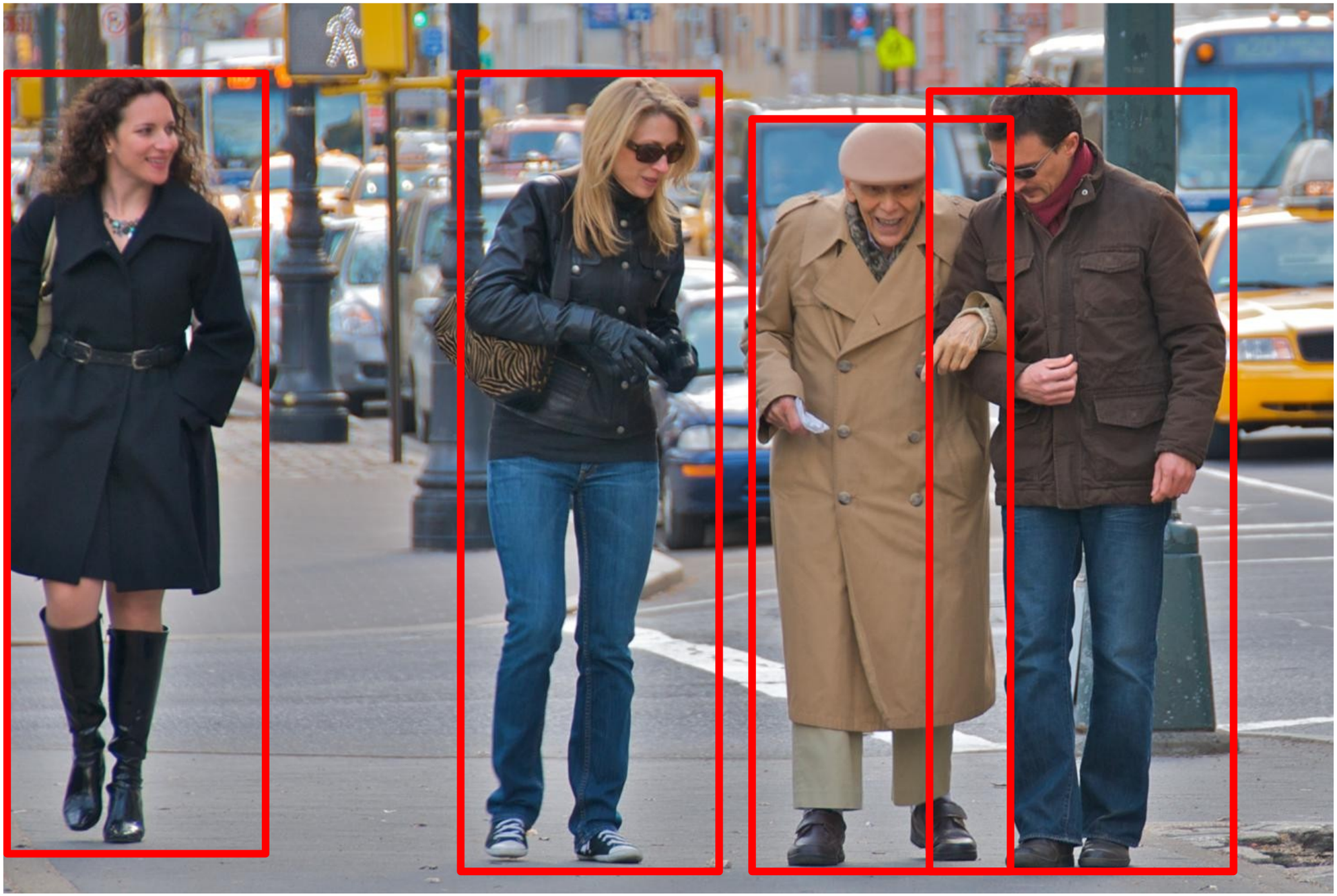




Backpropagation applied to handwritten zip code recognition,
Lecun et al., 1989



Rapid Object Detection using a Boosted Cascade of Simple Features,
Viola and Jones, CVPR 2001



Histograms of oriented gradients for human detection, Dalal and Triggs, CVPR 2005.

Datasets and computer vision

MNIST digits (1998-10)
Y LeCun & C. Cortes

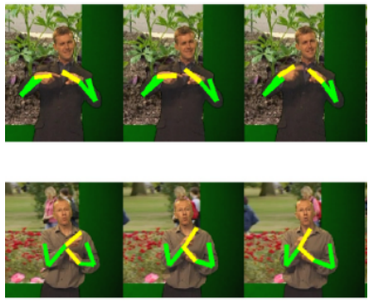
CMU/VASC Faces (1998)
H. Rowley, S. Baluja, T. Kanade

FERET Faces (1998)
P. Phillips, H. Wechsler, J. Huang, P. Raus

COIL Objects (1996)
S. Nene, S. Nayar, H. Murase

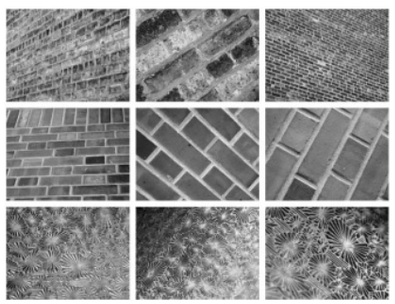
UIUC Cars (2004)
S. Agarwal, A. Awan, D. Roth

KTH human action (2004)
I. Leptev & B. Caputo

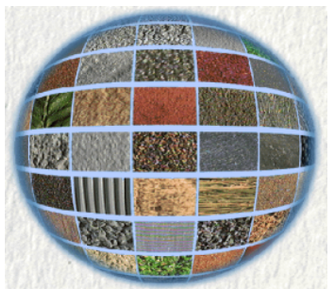


Sign Language (2008)
P. Buehler, M. Everingham, A. Zisserman

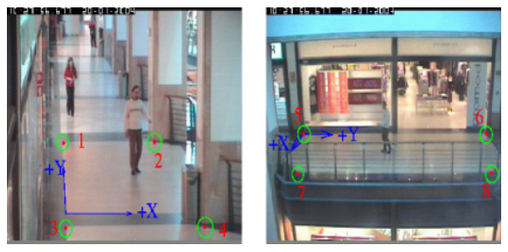
Segmentation (2001)
D. Martin, C. Fowlkes, D. Tal, J. Malik.



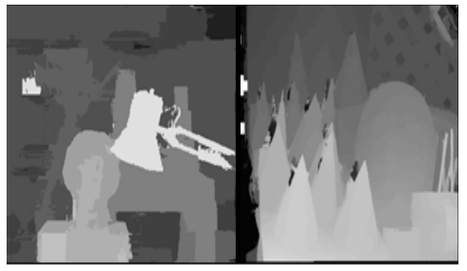
3D Textures (2005)
S. Lazebnik, C. Schmid, J. Ponce



CuRRET Textures (1999)
K. Dana B. Van Ginneken S. Nayar J. Koenderink



CAVIAR Tracking (2005)
R. Fisher, J. Santos-Victor J. Crowley



Middlebury Stereo (2002)
D. Scharstein R. Szeliski

In 2006 Fei-Fei Li was a new CS professor at UIUC. Everyone was trying to develop better algorithms that would make better decisions, regardless of the data.

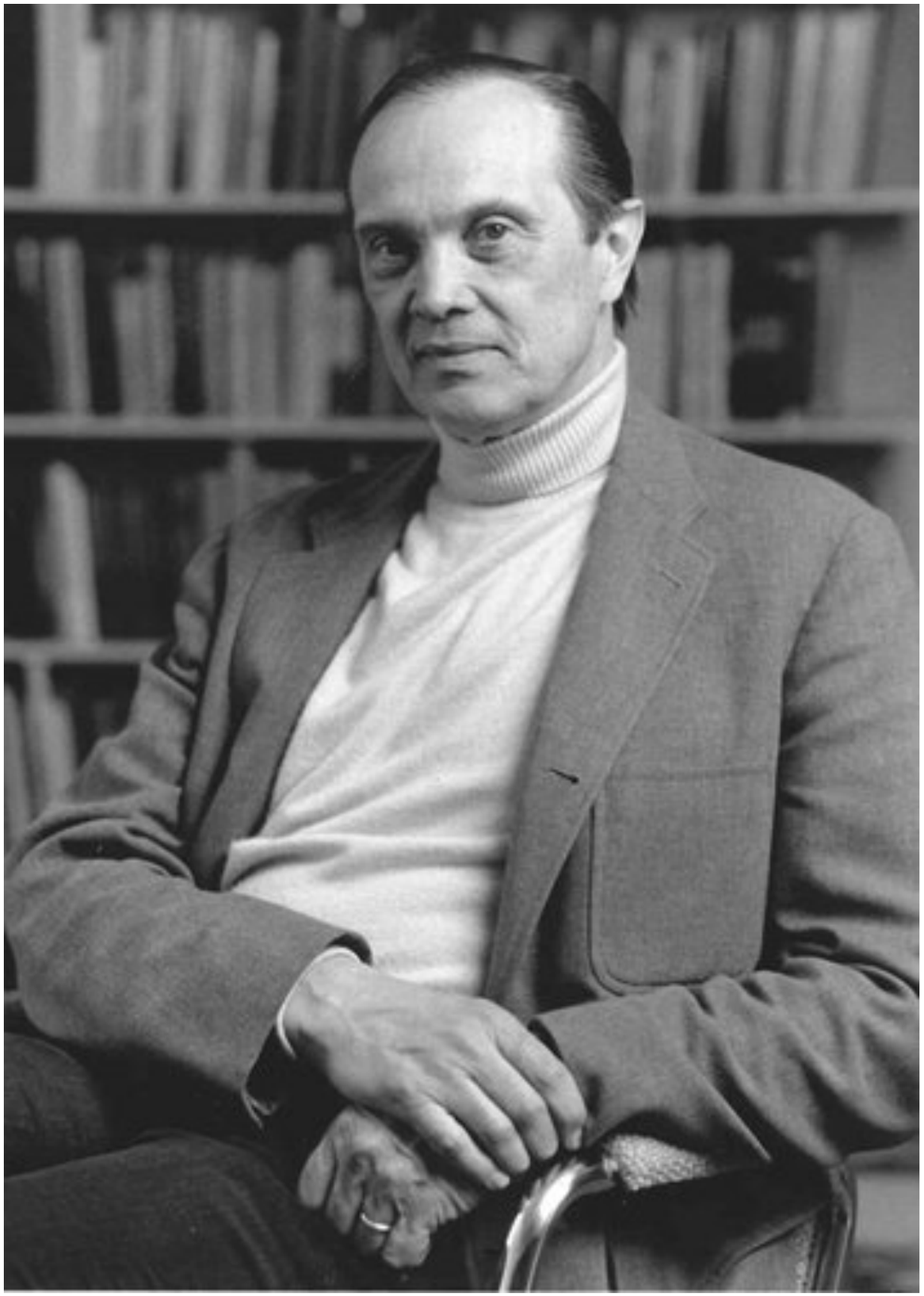
But she realized a limitation to this approach—the best algorithm wouldn’t work well if the data it learned from didn’t reflect the real world.

Her solution: build a better dataset.

“We decided we wanted to do something that was completely historically unprecedented. We’re going to map out the entire world of objects.” The resulting dataset was called ImageNet

What is IMAGENET

In the late 1980s, Princeton psychologist George Miller started a project called WordNet, with the aim of building a hierarchal structure for the English language. For example, dog is-a canine is-a mammal. It helped to organize language into a machine-readable logic, indexed more than 155,000 words.

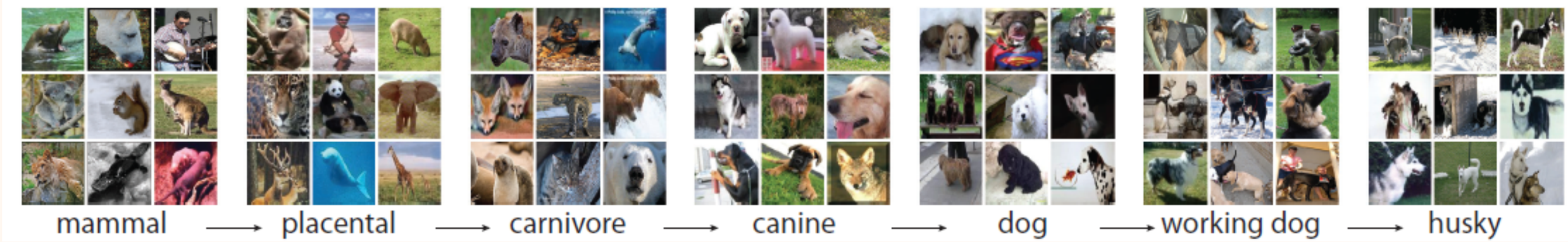


S: (n) **dog**, domestic dog, Canis familiaris (a member of the genus Canis (probably descended from the common wolf) that has been domesticated by man since prehistoric times; occurs in many breeds) *"the dog barked all night"*

- *direct hyponym* / *full hyponym*
- *part meronym*
- *member holonym*
- *direct hypernym* / ***inherited hypernym*** / *sister term*
  - S: (n) canine, canid (any of various fissiped mammals with nonretractile claws and typically long muzzles)
    - S: (n) carnivore (a terrestrial or aquatic flesh-eating mammal) *"terrestrial carnivores have four or five clawed digits on each limb"*
      - S: (n) placental, placental mammal, eutherian, eutherian mammal (mammals having a placenta; all mammals except monotremes and marsupials)
        - S: (n) mammal, mammalian (any warm-blooded vertebrate having the skin more or less covered with hair; young are born alive except for the small subclass of monotremes and nourished with milk)
          - S: (n) vertebrate, craniate (animals having a bony or cartilaginous skeleton with a segmented spinal column and a large brain enclosed in a skull or cranium)
            - S: (n) chordate (any animal of the phylum Chordata having a notochord or spinal column)
              - S: (n) animal, animate being, beast, brute, creature, fauna (a living organism characterized by voluntary movement)
                - S: (n) organism, being (a living thing that has (or can develop) the ability to act or function independently)
                  - S: (n) living thing, animate thing (a living (or once living) entity)

Christiane Fellbaum

IMAGENETontology



- S: (n) Eskimo dog, **husky** (breed of heavy-coated Arctic sled dog)
  - *direct hypernym* / ***inherited hypernym*** / *sister term*
    - S: (n) working dog (any of several breeds of usually large powerful dogs bred to work as draft animals and guard and guide dogs)
      - S: (n) dog, domestic dog, Canis familiaris (a member of the genus Canis (probably descended from the common wolf) that has been domesticated by man since prehistoric times; occurs in many breeds) *"the dog barked all night"*
        - S: (n) canine, canid (any of various fissiped mammals with nonretractile claws and typically long muzzles)
          - S: (n) carnivore (a terrestrial or aquatic flesh-eating mammal) *"terrestrial carnivores have four or five clawed digits on each limb"*

Constructing IMAGENET

Step 1:
Collect candidate images via the Internet

→

Step 2:
Clean up the candidate Images by humans

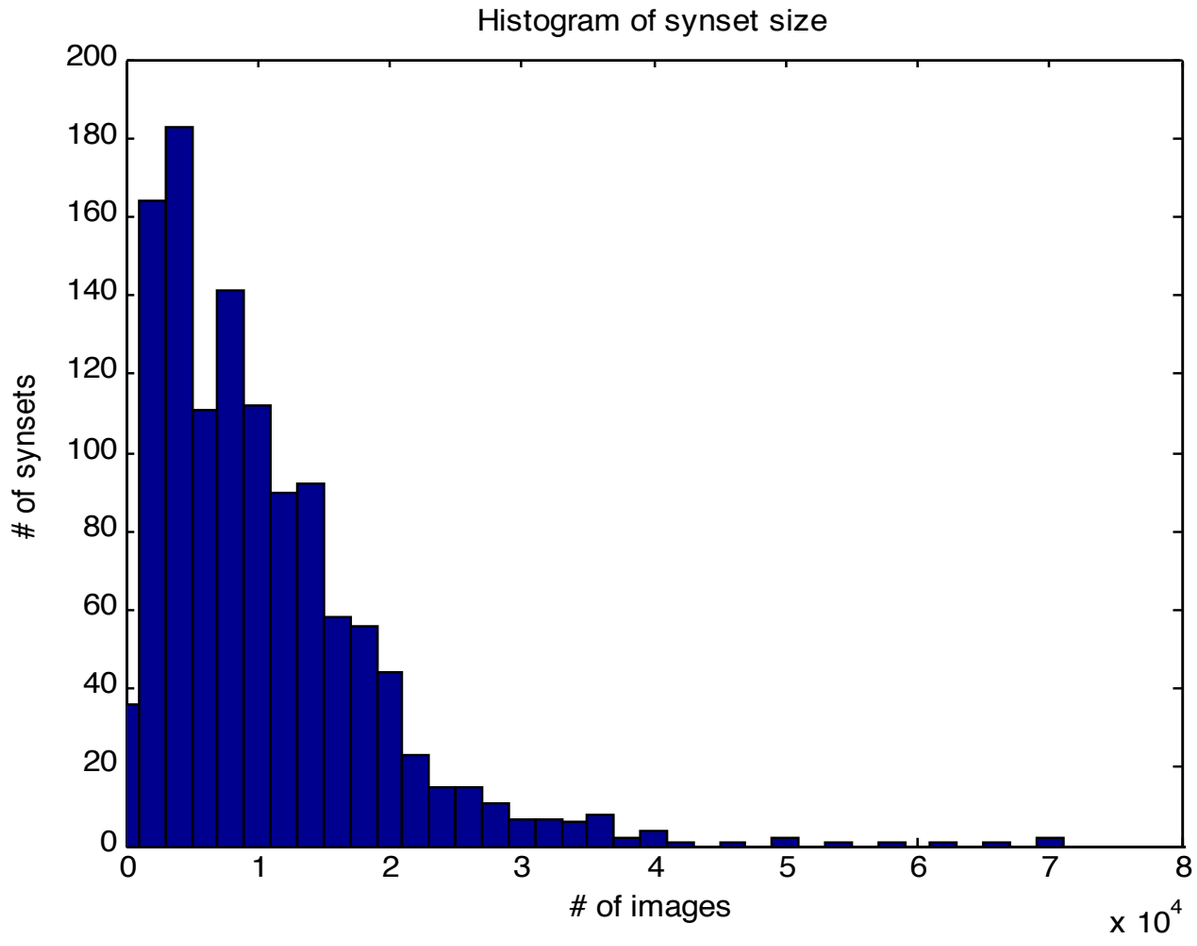
Step 1: Collect Candidate Images from the Internet

- Query expansion
  - Synonyms: *German shepherd, German police dog, German shepherd dog, Alsatian*
  - Appending words from ancestors: *sheepdog, dog*

- Collect images from multiple internet search engines

Step 1: Collect Candidate Images from the Internet

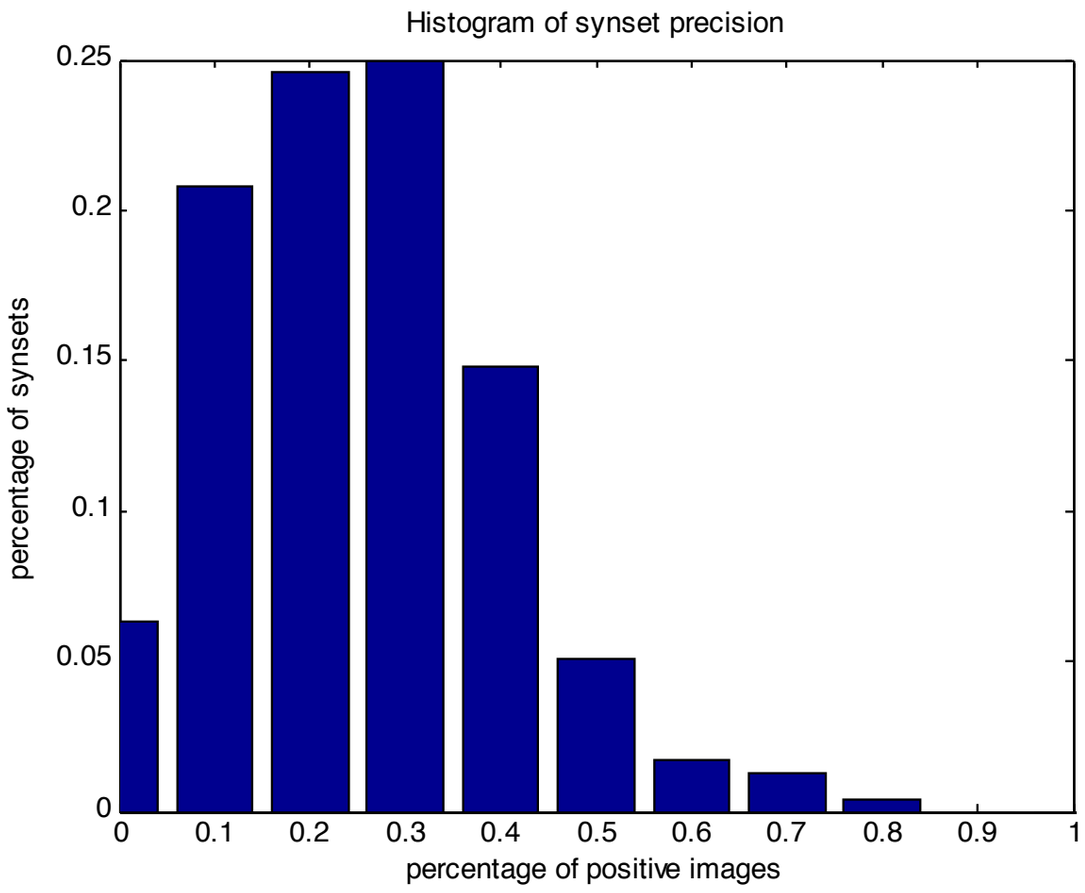
- "Mammal" subtree (1180 synsets)
  - Average # of images per synset: 10.5K



Most populated	Least populated
Humankind (118.5k)	Algeripithecus minutus (90)
Kitty, kitty-cat (69k)	Striped muishond (107)
Cattle, cows (65k)	Mylodonitid (127)
Pooch, doggie (62k)	Greater pichiciego (128)
Cougar, puma (57k)	Damaraland mole rat (188)
Frog, toad (53k)	Western pipistrel (196)
Hack, jade, nag (50k)	Muishond (215)

Step 1: Collect Candidate Images from the Internet

- "Mammal" subtree (1180 synsets)
  - Average accuracy per synset: 26%



Most accurate	Least accurate
Bottlenose dolpin (80%)	Fanaloka (1%)
Meerkat (74%)	Pallid bat (3%)
Burmese cat (74%)	Vaquita (3%)
Humpback whale (69%)	Fisher cat (3%)
African elephant (63%)	Walrus (4%)
Squirrel (60%)	Grison (4%)
Domestic cat (59%)	Pika, Mouse hare (4%)

Constructing IMAGENET

Step 1:
Collect candidate images via the Internet

→

Step 2:
Clean up the candidate Images by humans

How long will it take?

Li's first idea was to hire undergraduate students for $10 an hour to manually find images and add them to the dataset. But back-of-the-napkin math quickly made Li realize that at the undergrads' rate of collecting images it would take too long to complete.

$$40{,}000 \times 10{,}000 \times 3 / 2 = 600{,}000{,}000\,\text{sec} \approx 19\,\text{years}$$

synsets images people .4 seconds per image

After the undergrad task force was disbanded, Li and the team went back to the drawing board. What if computer-vision algorithms could pick the photos from the internet, and humans would then just curate the images? But the team decided the technique wasn't sustainable either—future algorithms would be constricted to only judging what algorithms were capable of recognizing at the time the dataset was compiled.

Undergrads were time-consuming, algorithms were flawed, and the team didn't have money—Li said the project failed to win any of the federal grants she applied for, receiving comments on proposals that it was shameful Princeton would research this topic, and that the only strength of proposal was that Li was a woman.

A solution finally surfaced in a chance hallway conversation with a graduate student who asked Li whether she had heard of Amazon Mechanical Turk, a service where hordes of humans sitting at computers around the world would complete small online tasks for pennies.

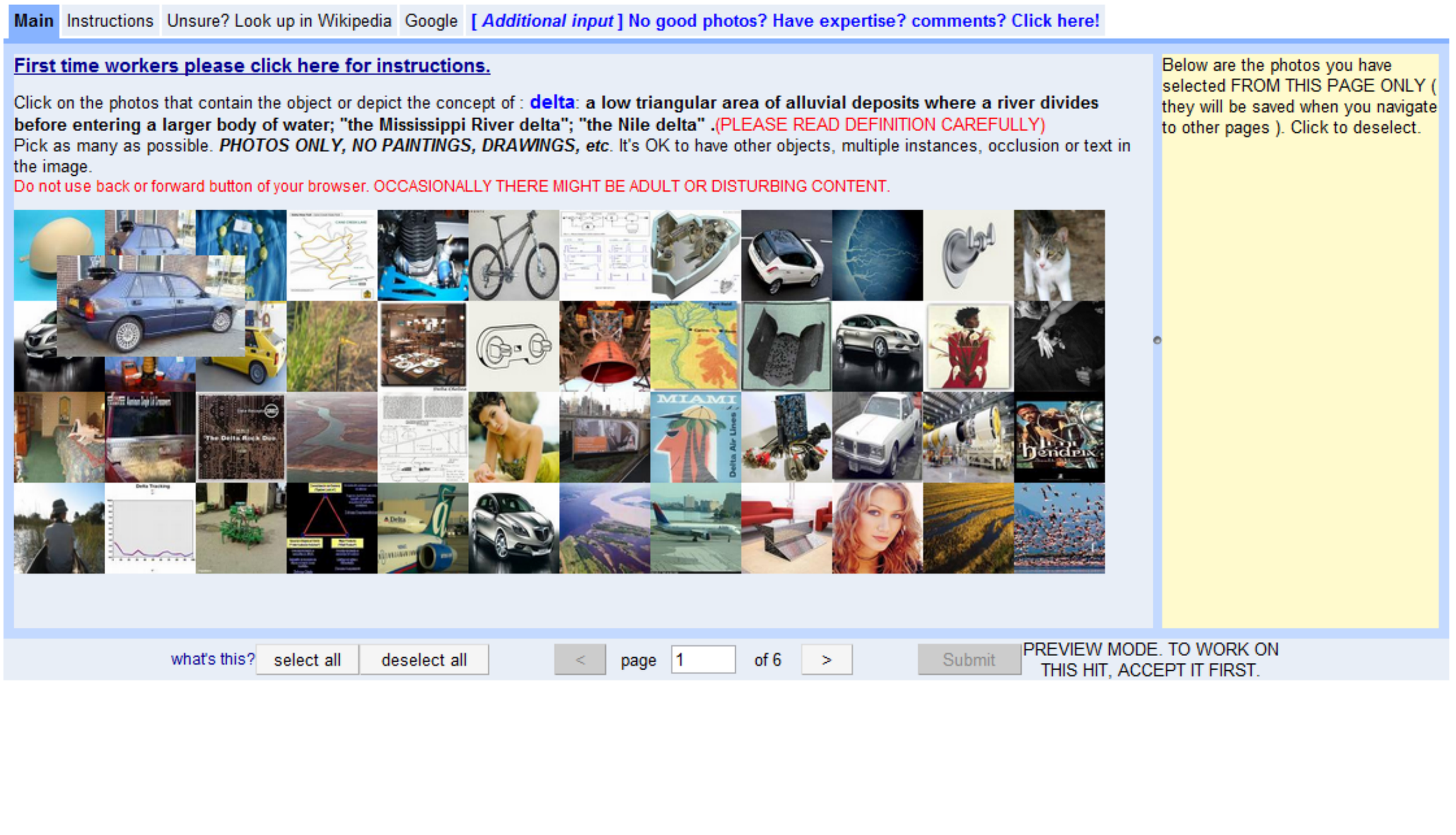
The data that transformed AI research—and possibly the world

By Dave Gershgorn
July 26, 2017

"He showed me the website, and I can tell you literally that day I knew the ImageNet project was going to happen," she said. "Suddenly we found a tool that could scale, that we could not possibly dream of by hiring Princeton undergrads."

IMAGENET Basic User Interface

Click on the good images.

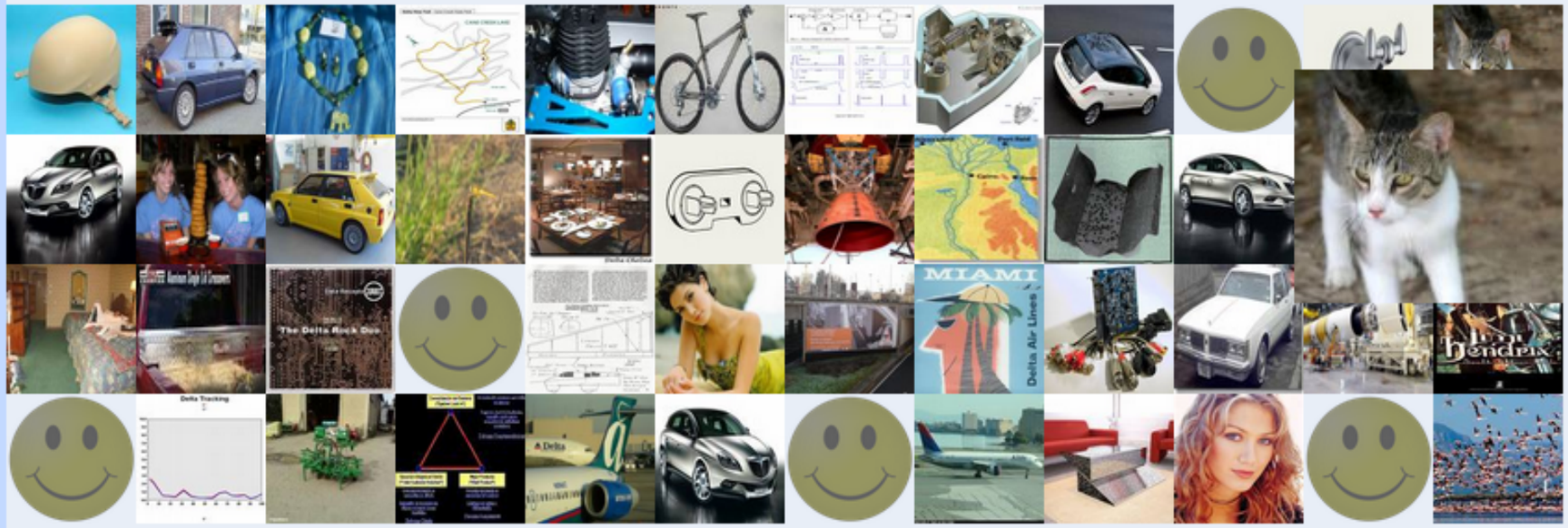


IMAGENET Basic User Interface

Main | Instructions | Unsure? Look up in Wikipedia | Google | [*Additional input*] No good photos? Have expertise? comments? Click here!

First time workers please click here for instructions.

Click on the photos that contain the object or depict the concept of : **delta**: **a low triangular area of alluvial deposits where a river divides before entering a larger body of water; "the Mississippi River delta"; "the Nile delta" .**(PLEASE READ DEFINITION CAREFULLY)
Pick as many as possible. ***PHOTOS ONLY, NO PAINTINGS, DRAWINGS, etc***. It's OK to have other objects, multiple instances, occlusion or text in the image.
Do not use back or forward button of your browser. OCCASIONALLY THERE MIGHT BE ADULT OR DISTURBING CONTENT.



Below are the photos you have selected FROM THIS PAGE ONLY (they will be saved when you navigate to other pages). Click to deselect.

what's this? select all | deselect all | < | page 1 | of 6 | > | Submit | PREVIEW MODE. TO WORK ON THIS HIT, ACCEPT IT FIRST.

Mechanical Turk brought its own slew of hurdles, with much of the work fielded by two of Li's Ph.D students, Jia Deng and Olga Russakovsky . For example, how many Turkers needed to look at each image? Maybe two people could determine that a cat was a cat, but an image of a miniature husky might require 10 rounds of validation. What if some Turkers tried to game or cheat the system? Li's team ended up creating a batch of statistical models for Turker's behaviors to help ensure the dataset only included correct images.

Even after finding Mechanical Turk, the dataset took two and a half years to complete. It consisted of 3.2 million labelled images, separated into 5,247 categories, sorted into 12 subtrees like "mammal," "vehicle," and "furniture."

Enhancement 1

- Provide wiki and google links

Enhancement 2

- Make sure workers read the definition.
  - Words are ambiguous. E.g.
    - Box: *any one of several designated areas on a ball field where the batter or catcher or coaches are positioned*
    - Keyboard: *holder consisting of an arrangement of hooks on which keys or locks can be hung*
  - These synsets are hard to get right
  - Some workers do not read or understand the definition.

Definition quiz

This HIT is about 'delta'.

Definition: a low triangular area of alluvial deposits where a river divides before entering a larger body of water; "the Mississippi River delta"; "the Nile delta"

Please read the above definition carefully. 'delta' might mean something different from what you think.

I HAVE READ IT

Definition quiz

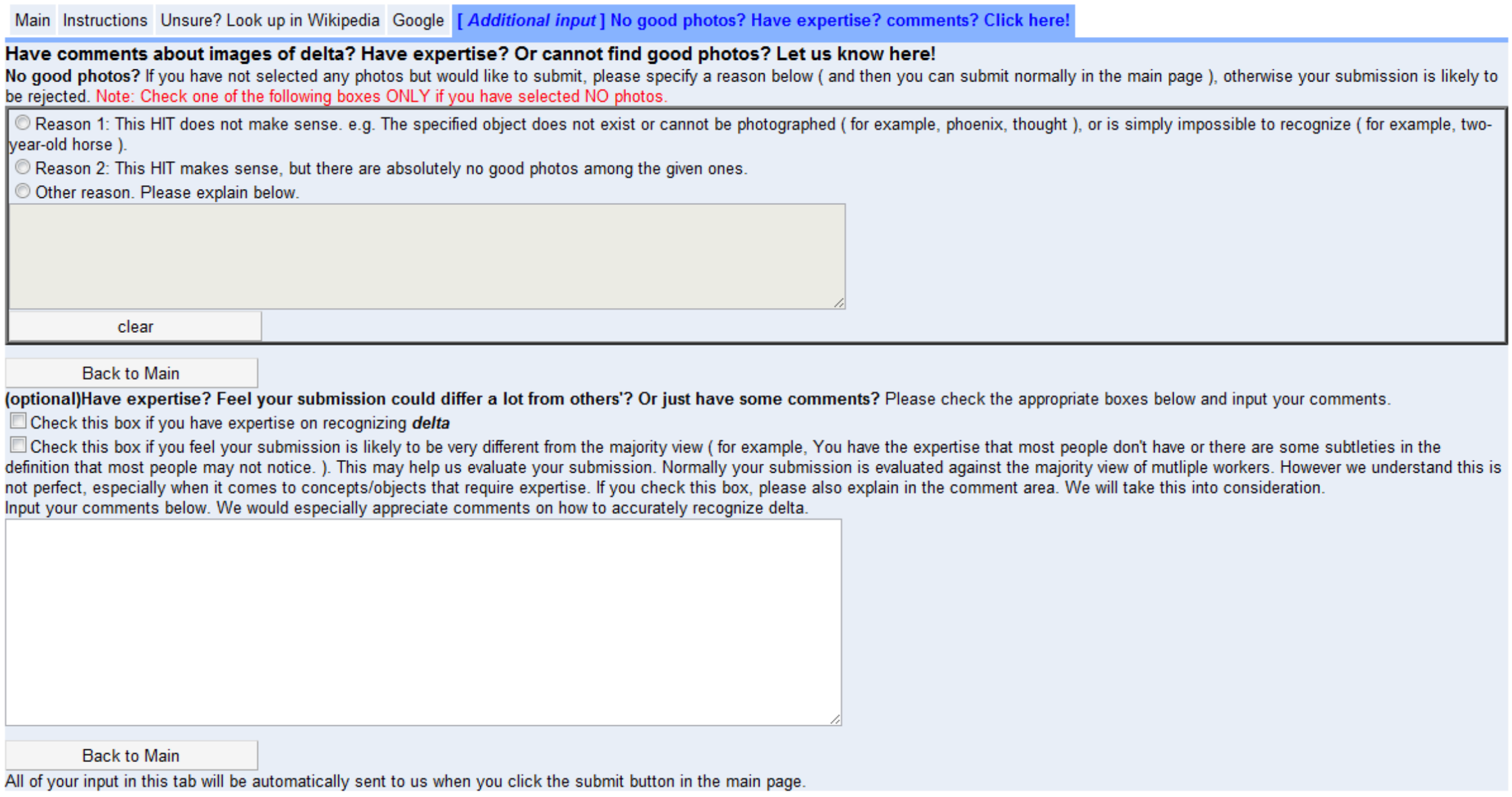
Please answer: what is the meaning of 'delta' in this HIT?

Go back and read the definition again.

- the normal brainwave in the encephalogram of a person in deep dreamless sleep; occurs with high voltage and low frequency (1 to 4 hertz)
- the 4th letter of the Greek alphabet
- a low triangular area of alluvial deposits where a river divides before entering a larger body of water; "the Mississippi River delta"; "the Nile delta"
- an airplane with wings that give it the appearance of an isosceles triangle
- an object shaped like an equilateral triangle

Enhancement 3

- Allow more feedback. E.g. “unimagable synsets” expert opinion


Main | Instructions | Unsure? Look up in Wikipedia | Google | [*Additional input*] No good photos? Have expertise? comments? Click here!

Have comments about images of delta? Have expertise? Or cannot find good photos? Let us know here!

No good photos? If you have not selected any photos but would like to submit, please specify a reason below (and then you can submit normally in the main page), otherwise your submission is likely to be rejected. Note: Check one of the following boxes ONLY if you have selected NO photos.

- Reason 1: This HIT does not make sense. e.g. The specified object does not exist or cannot be photographed (for example, phoenix, thought), or is simply impossible to recognize (for example, two-year-old horse).
- Reason 2: This HIT makes sense, but there are absolutely no good photos among the given ones.
- Other reason. Please explain below.

clear

Back to Main

(optional)Have expertise? Feel your submission could differ a lot from others'? Or just have some comments? Please check the appropriate boxes below and input your comments.

- Check this box if you have expertise on recognizing ***delta***
- Check this box if you feel your submission is likely to be very different from the majority view (for example, You have the expertise that most people don't have or there are some subtleties in the definition that most people may not notice.). This may help us evaluate your submission. Normally your submission is evaluated against the majority view of mutliple workers. However we understand this is not perfect, especially when it comes to concepts/objects that require expertise. If you check this box, please also explain in the comment area. We will take this into consideration.

Input your comments below. We would especially appreciate comments on how to accurately recognize delta.

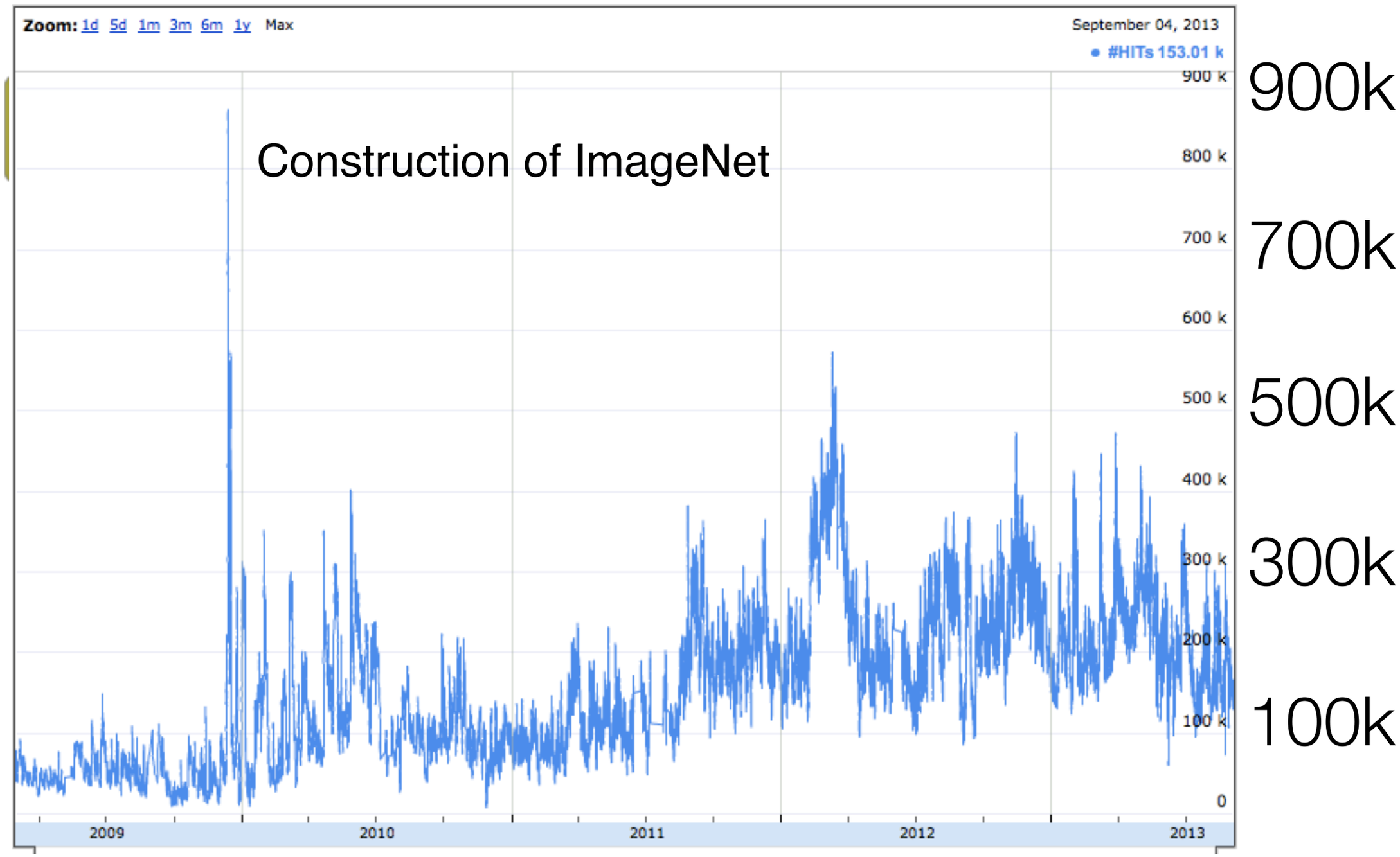
Back to Main

All of your input in this tab will be automatically sent to us when you click the submit button in the main page.

IMAGENET is built by crowdsourcing

- July 2008: 0 images
- Dec 2008: 3 million images, 6000+ synsets
- April 2010: 11 million images, 15,000+ synsets

MTurk Tracker



900k

700k

500k

300k

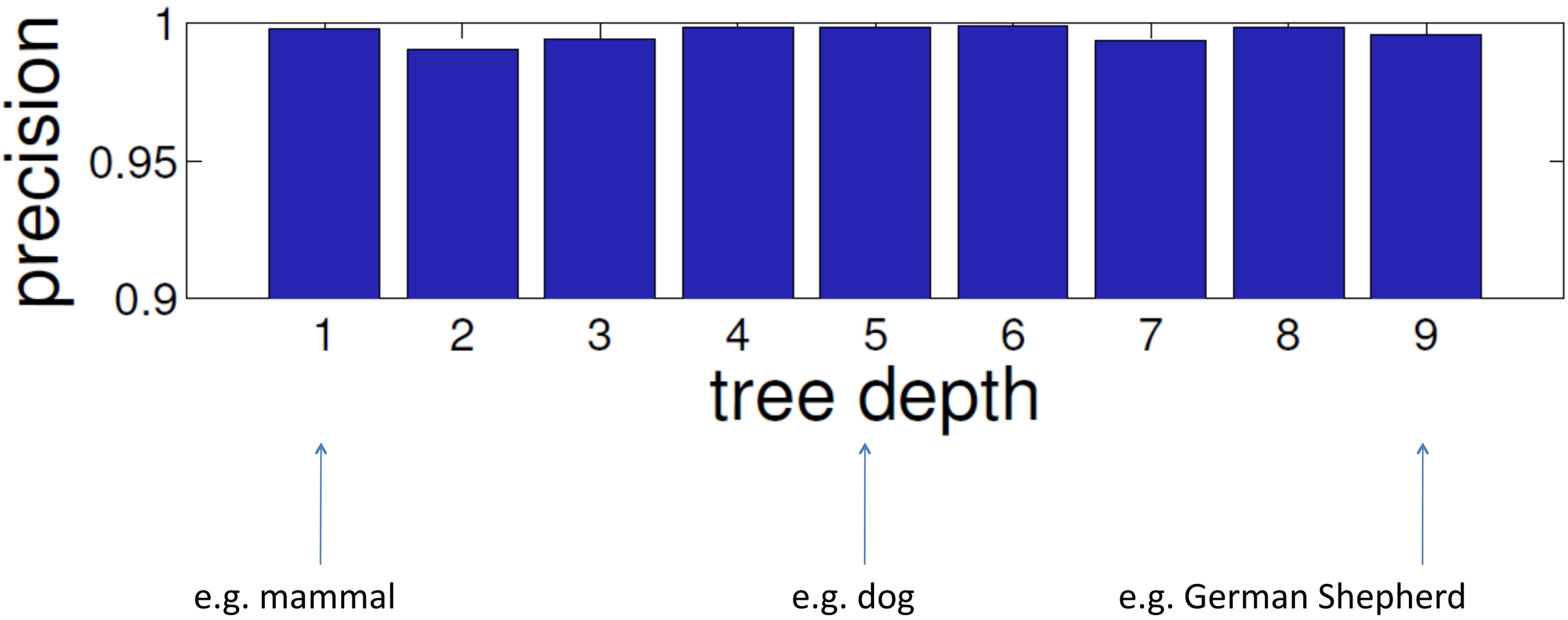
100k

2009 2010 2011 2012 2013

U.S. economy 2008 - 2010

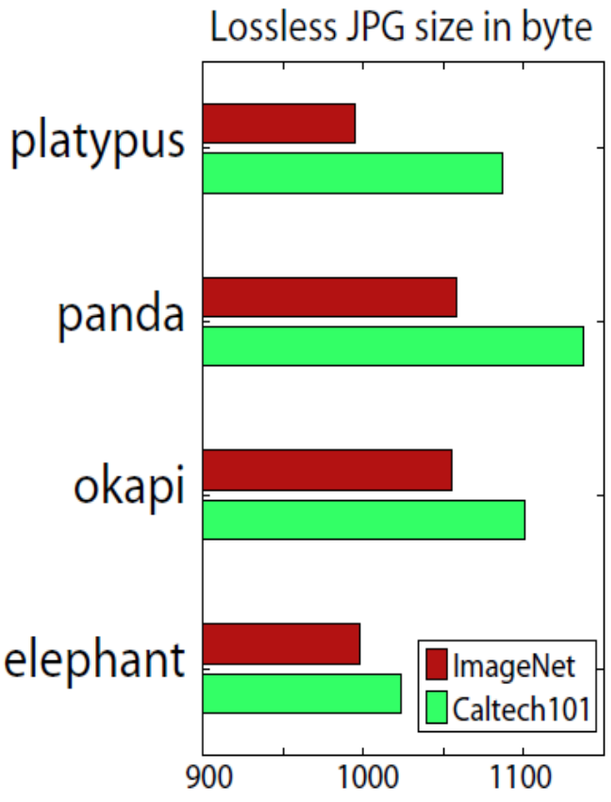
IMAGENET hired more than 25,000 AMT workers in this period of time!!

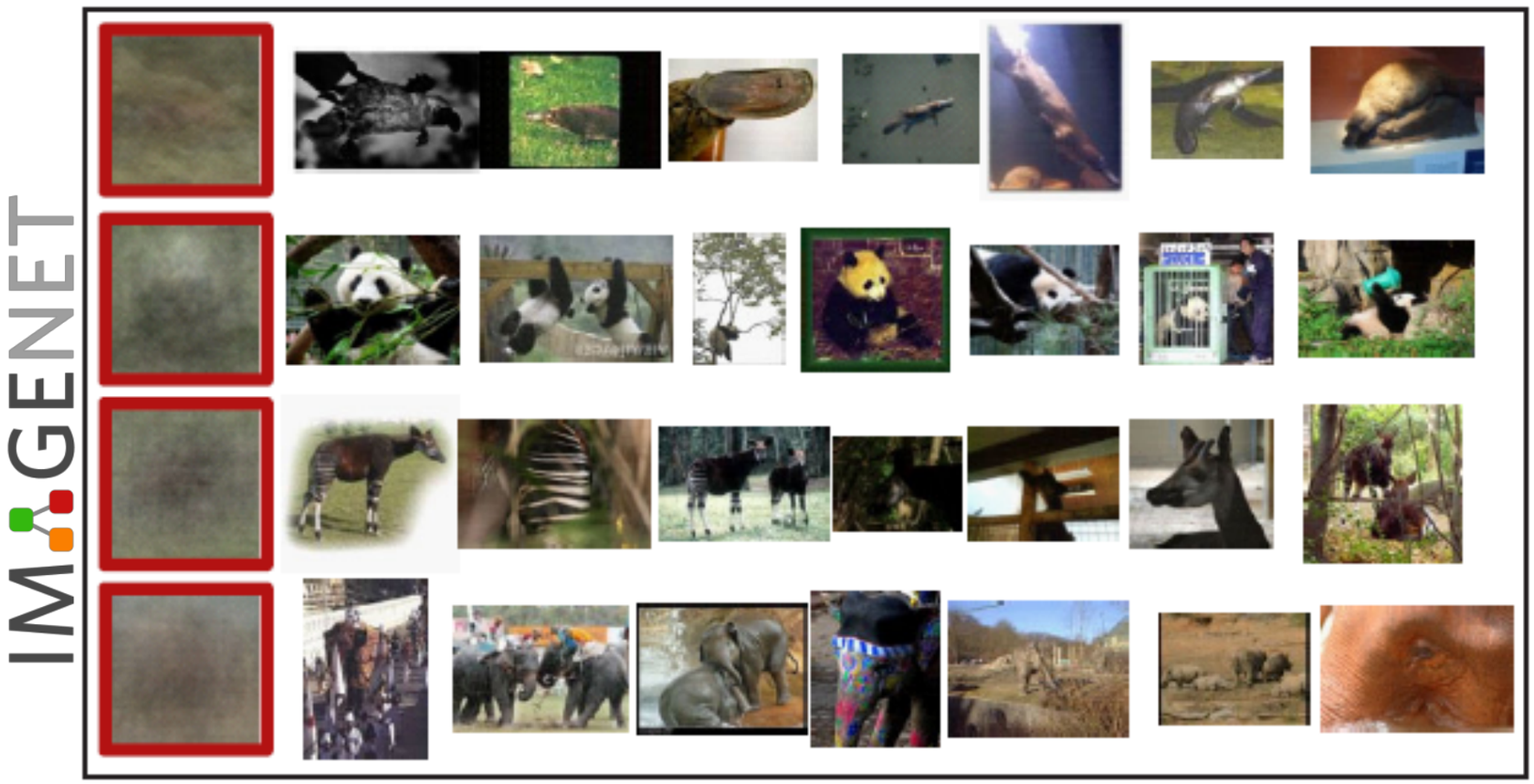
Accuracy

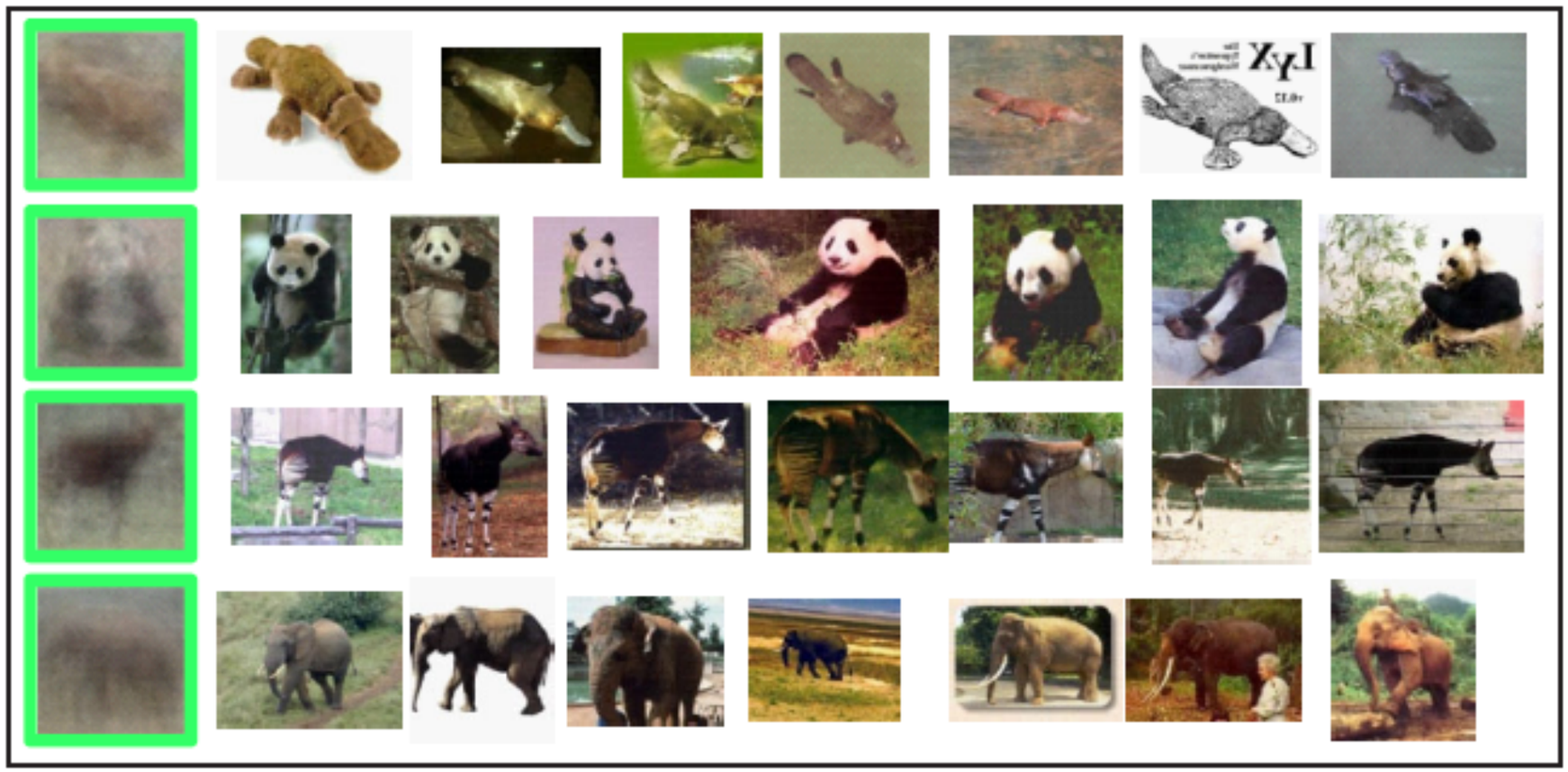


Deng, Dong, Socher, Li, Li, & Fei-Fei, *CVPR*, 2009

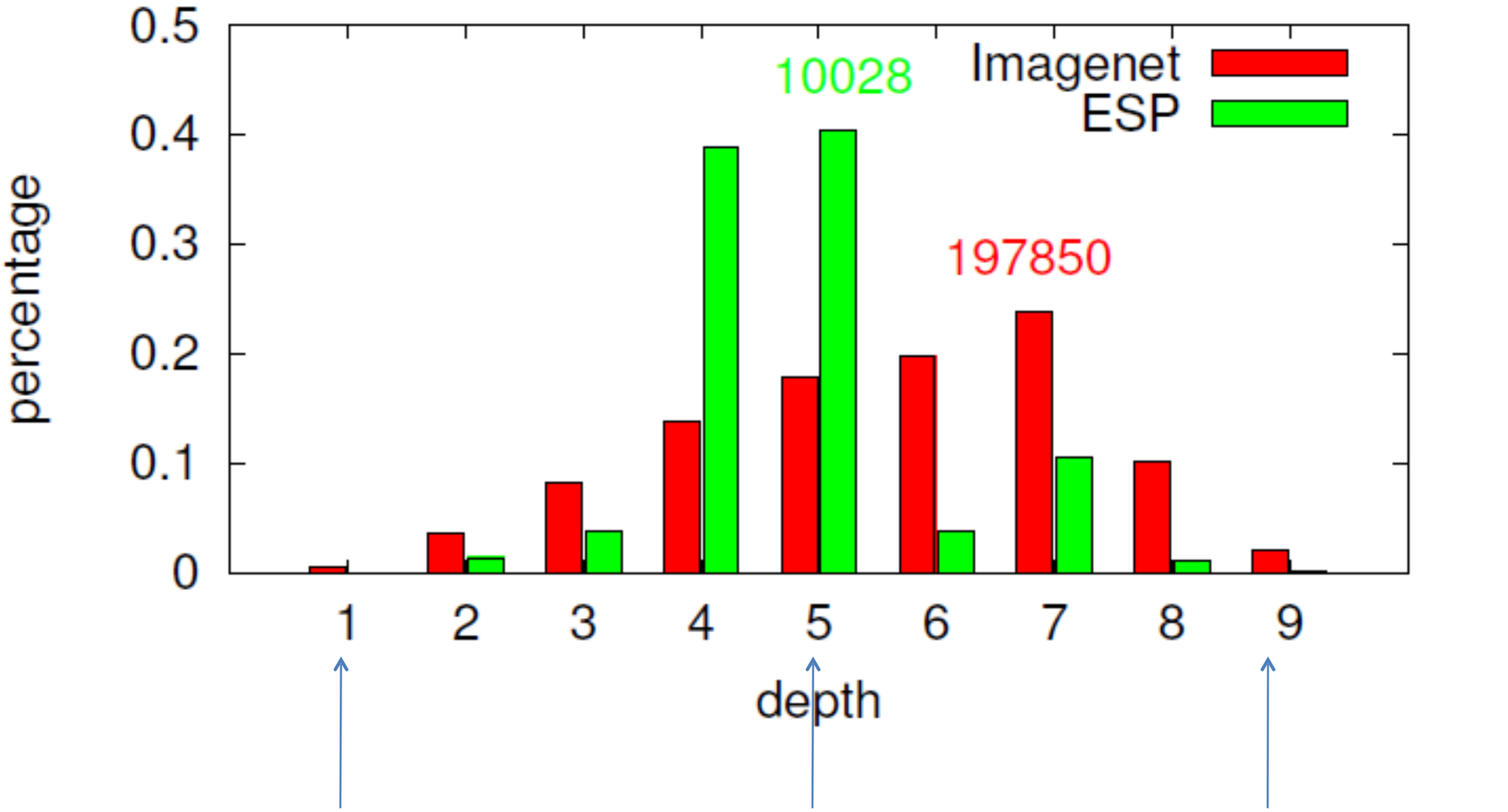
Diversity


Lossless JPG size in byte
platypus
panda
okapi
elephant
ImageNet
Caltech101
900
1000
1100


IMAGENET


Caltech101

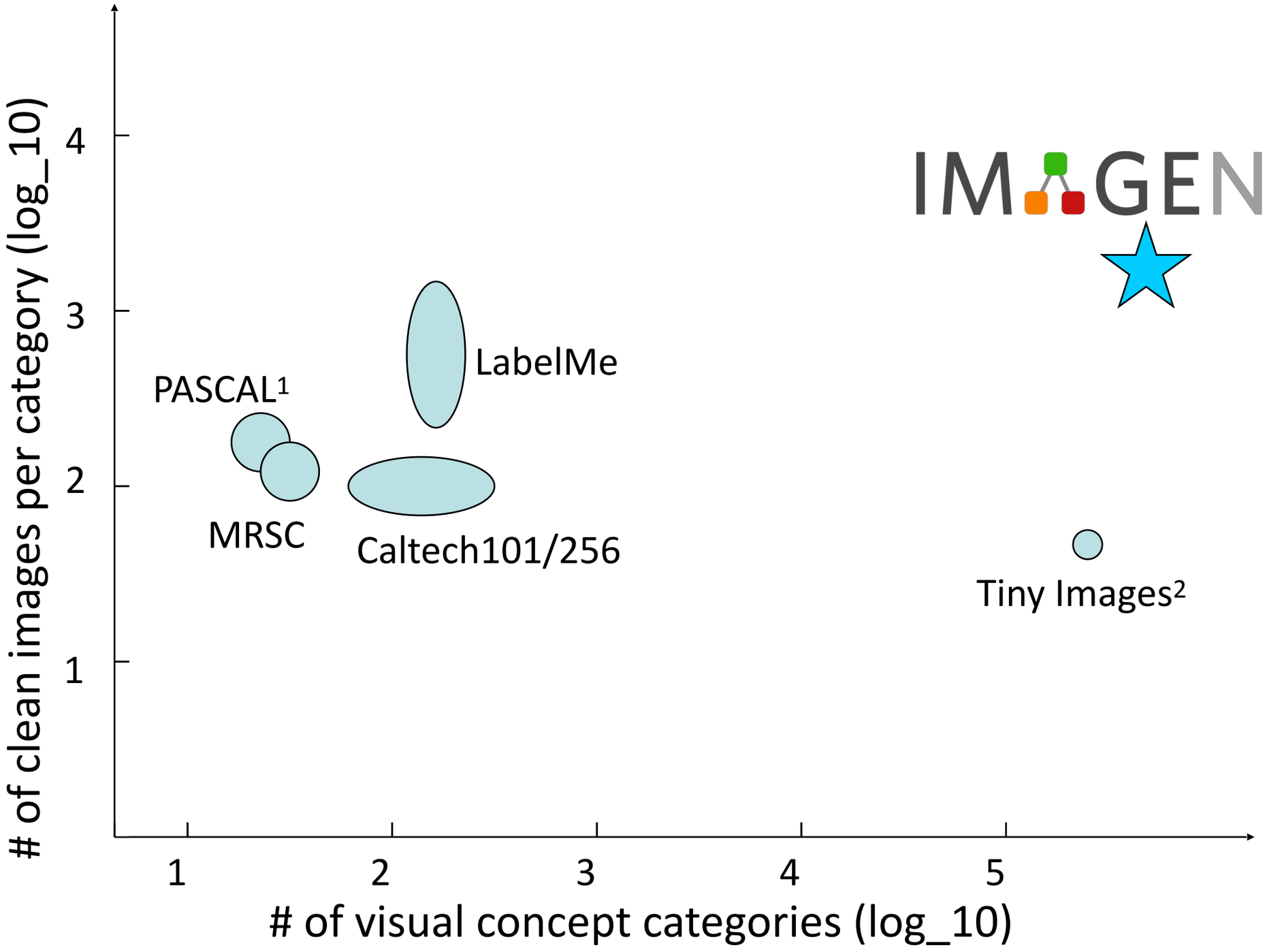
Diversity



ESP: Ahn et al. 2006

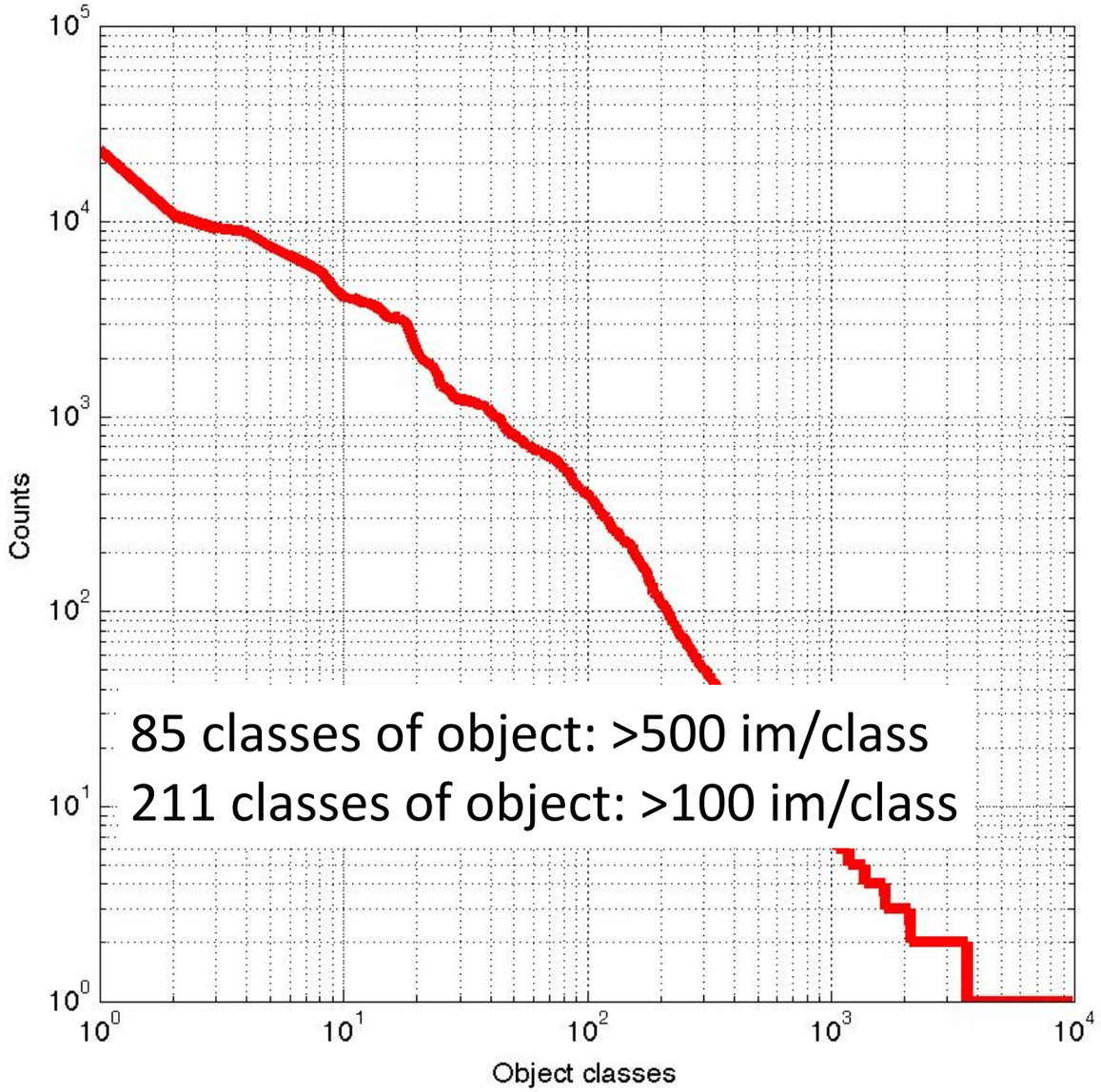
Deng, Dong, Socher, Li, Li, & Fei-Fei, *CVPR*, 2009

Comparison among free datasets



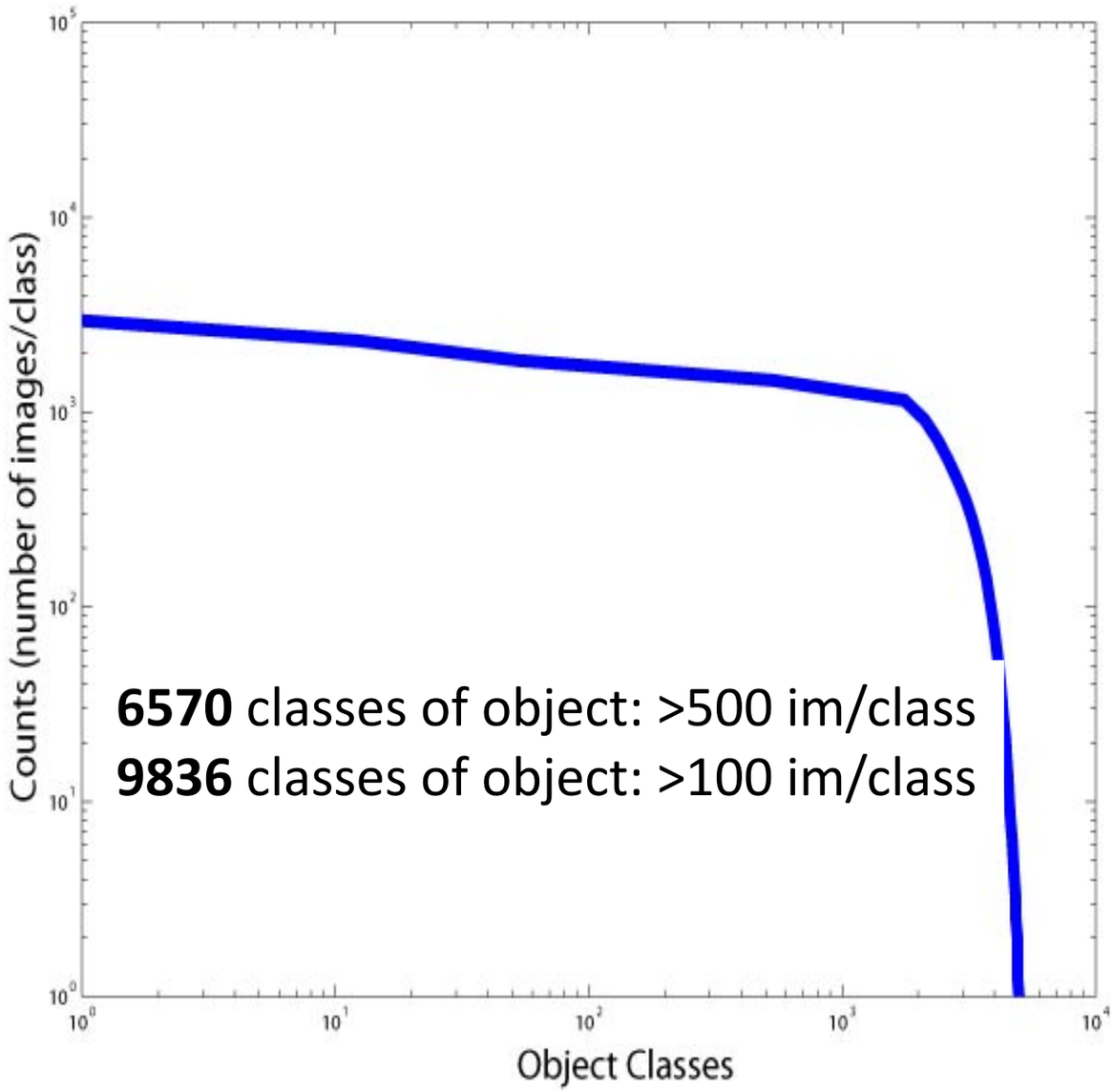
1. Excluding the Caltech101 datasets from PASCAL
2. No image in this dataset is human annotated. The # of clean images per category is a rough estimation

Scale



LabelMe

Russell et al. 2005;
statistics obtained in 2009



IMAGENET

What does classifying more than 10,000 image categories tell us?

Moped

Bicycle

Motorbike

Go-cart

Trail Bike

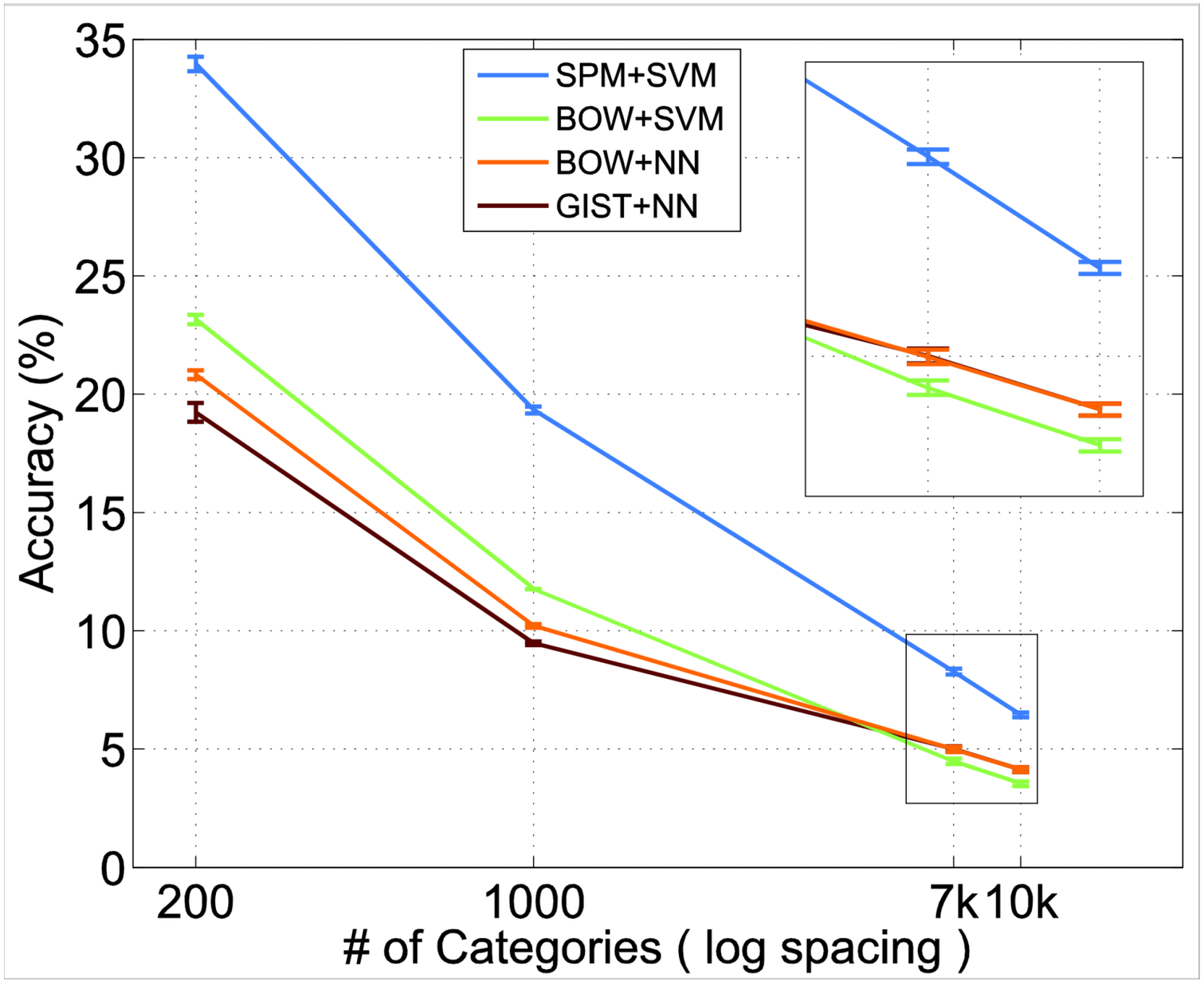
Car, auto

Helicopter

Background image courtesy: Antonio Torralba

Size matters

- 6.4% for 10K categories
- Better than we expected (instead of dropping at the rate of 10x; it's roughly at about 2x)
- An ordering switch between SVM and NN methods when the # of categories becomes large

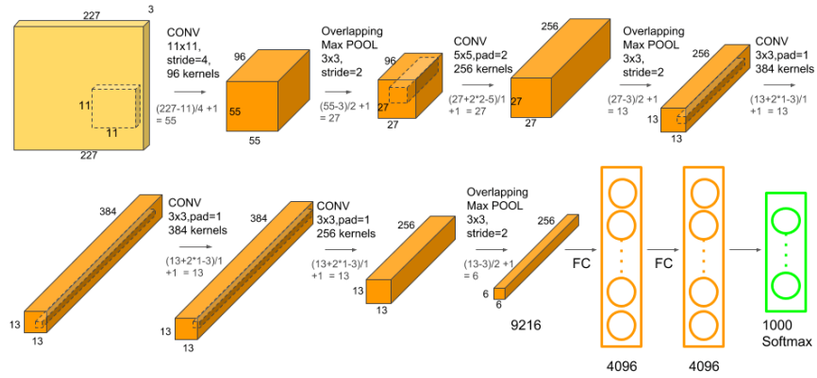


Deng, Berg, Li, & Fei-Fei, *ECCV2010*

Li, Berg, and Deng authored five papers together based on the dataset, exploring how algorithms would interpret such vast amounts of data. The first paper would become a benchmark for how an algorithm would react to thousands of classes of images, the predecessor to the ImageNet competition.

“We realized to democratize this idea we needed to reach out further,” Li said. She approached a well-known image recognition competition in Europe called PASCAL VOC, which agreed to collaborate and co-brand their competition with ImageNet. The PASCAL challenge was a well-respected competition and dataset, but representative of the previous method of thinking. It only had 20 classes, compared to ImageNet’s 1,000.

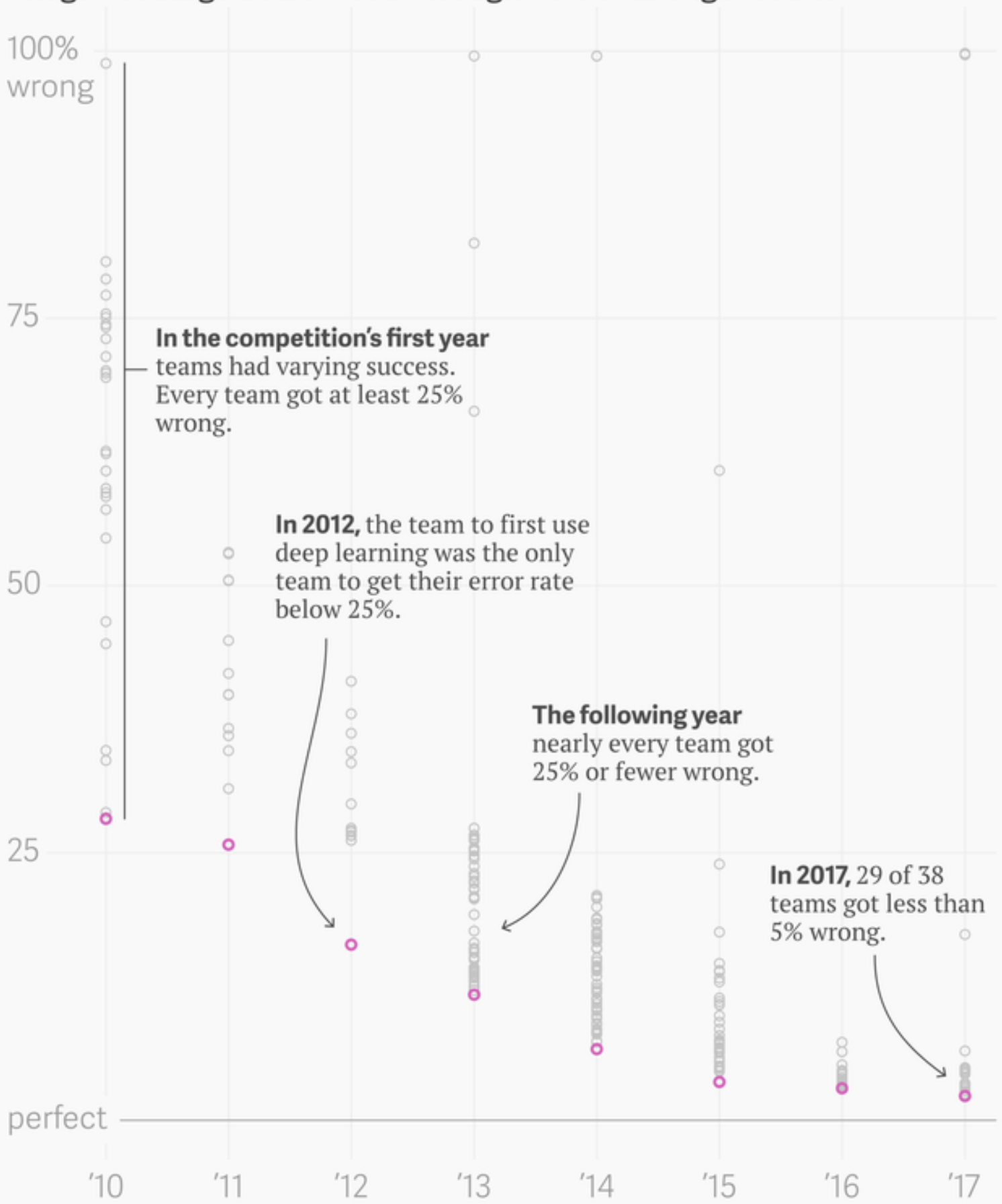
Two years after the first ImageNet competition, in 2012, something even bigger happened. Indeed, **if the artificial intelligence boom we see today could be attributed to a single event, it would be the announcement of the 2012 ImageNet challenge results.** Geoffrey Hinton, Ilya Sutskever, and Alex Krizhevsky submitted a deep convolutional neural network architecture called AlexNet which beat the field by a whopping 10.8 percentage point margin, which was 41% better than the next best.



The data that transformed AI research—and possibly the world

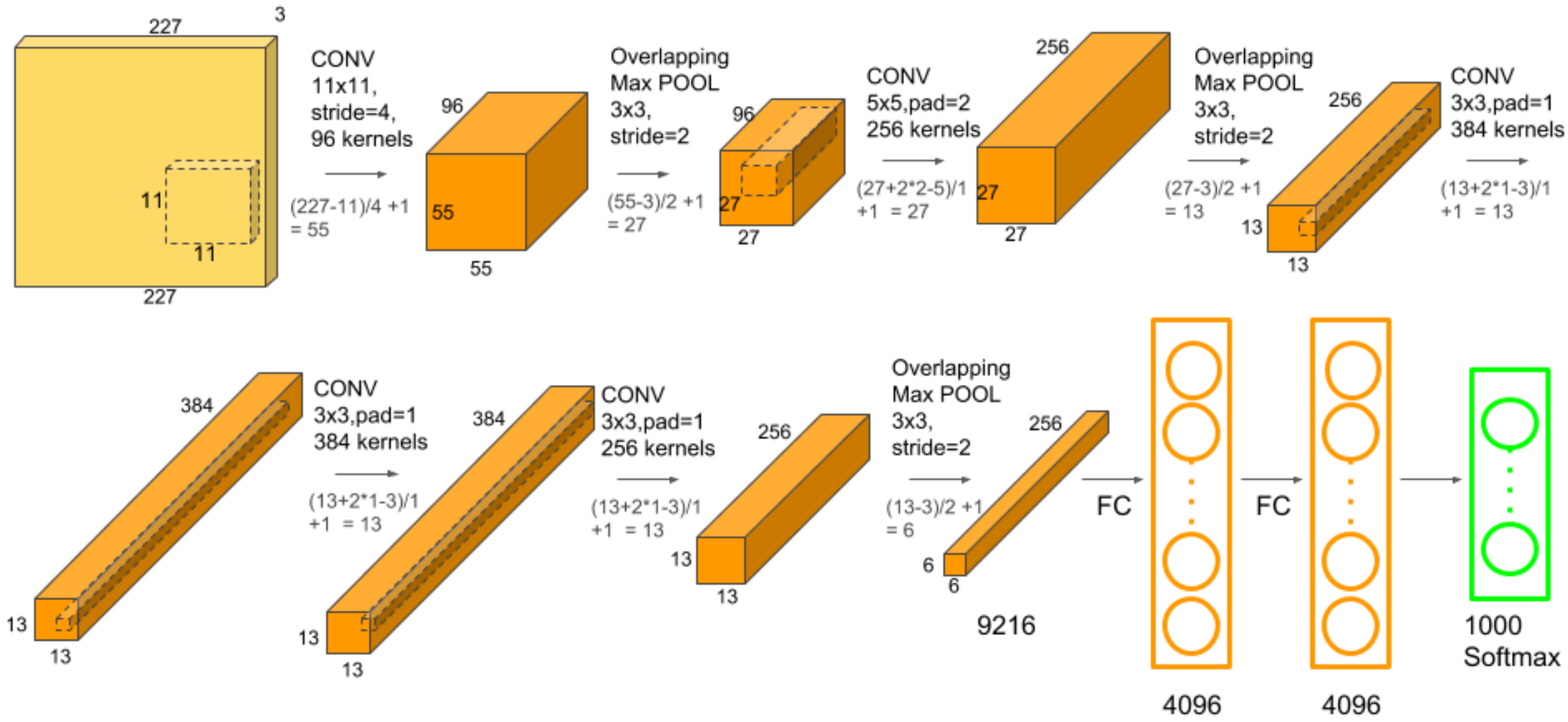
By Dave Gershgorn
July 26, 2017

ImageNet Large Scale Visual Recognition Challenge results



David Yanofsky | Quartz

Data: ImageNet



AlexNet is the name of a convolutional neural network which has had a large impact on the field of machine learning, specifically in the application of deep learning to machine vision. It famously won the 2012 ImageNet LSVRC-2012 competition by a large margin (15.3% VS 26.2% (second place) error rates). The network had a very similar architecture as LeNet by Yann LeCun et al but was deeper, with more filters per layer, and with stacked convolutional layers. It consisted of 11×11, 5×5,3×3, convolutions, max pooling, dropout, data augmentation, ReLU activations, SGD with momentum. It attached ReLU activations after every convolutional and fully-connected layer. AlexNet was trained for 6 days simultaneously on two Nvidia Geforce GTX 580 GPUs which is the reason for why their network is split into two pipelines.

After ImageNet

- 2017 was the final year of the ImageNet competition.
- Many people consider it to be "solved"– the error rate is now around 2%.
- ImageNet inspired many subsequent research efforts.