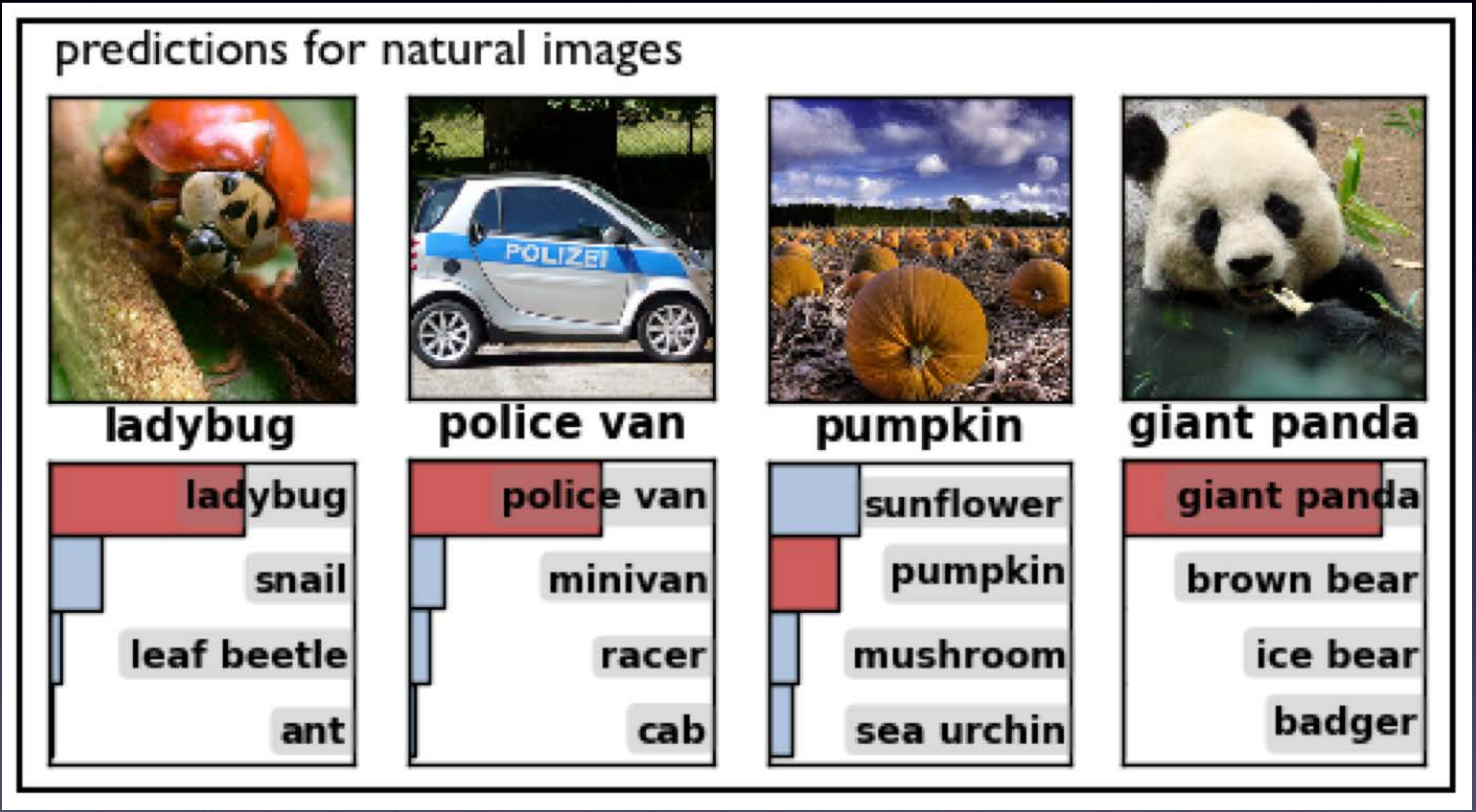
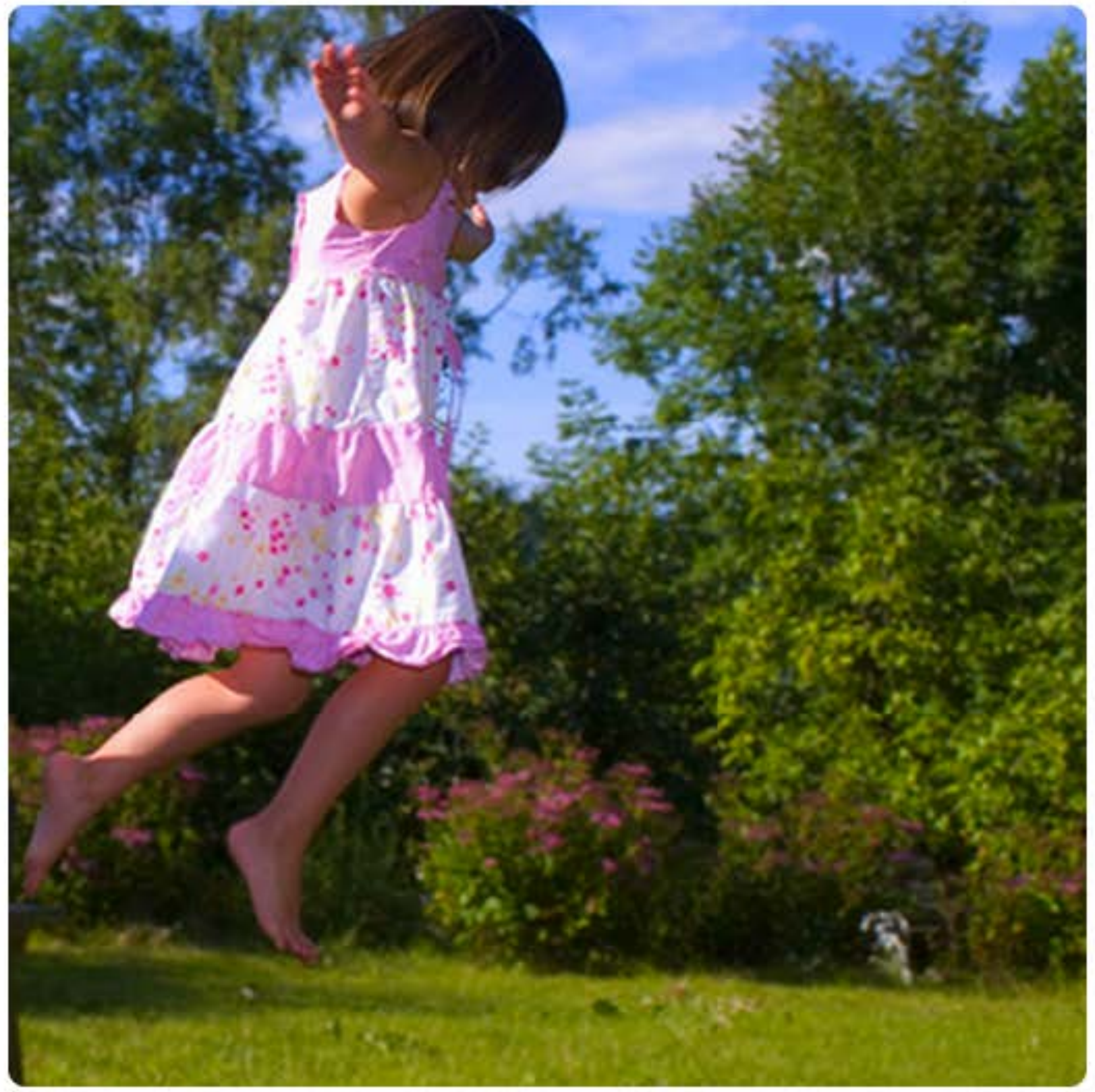
DNNs better than humans at image recognition!!



Alex Krizhevsky et al. Nips 2012

ImageNet
1,000 Categories
1.3 M Images
Human error: 5%
DNN: 3%

Deep Neural Networks/Deep Learning



"girl in pink dress is jumping in air."

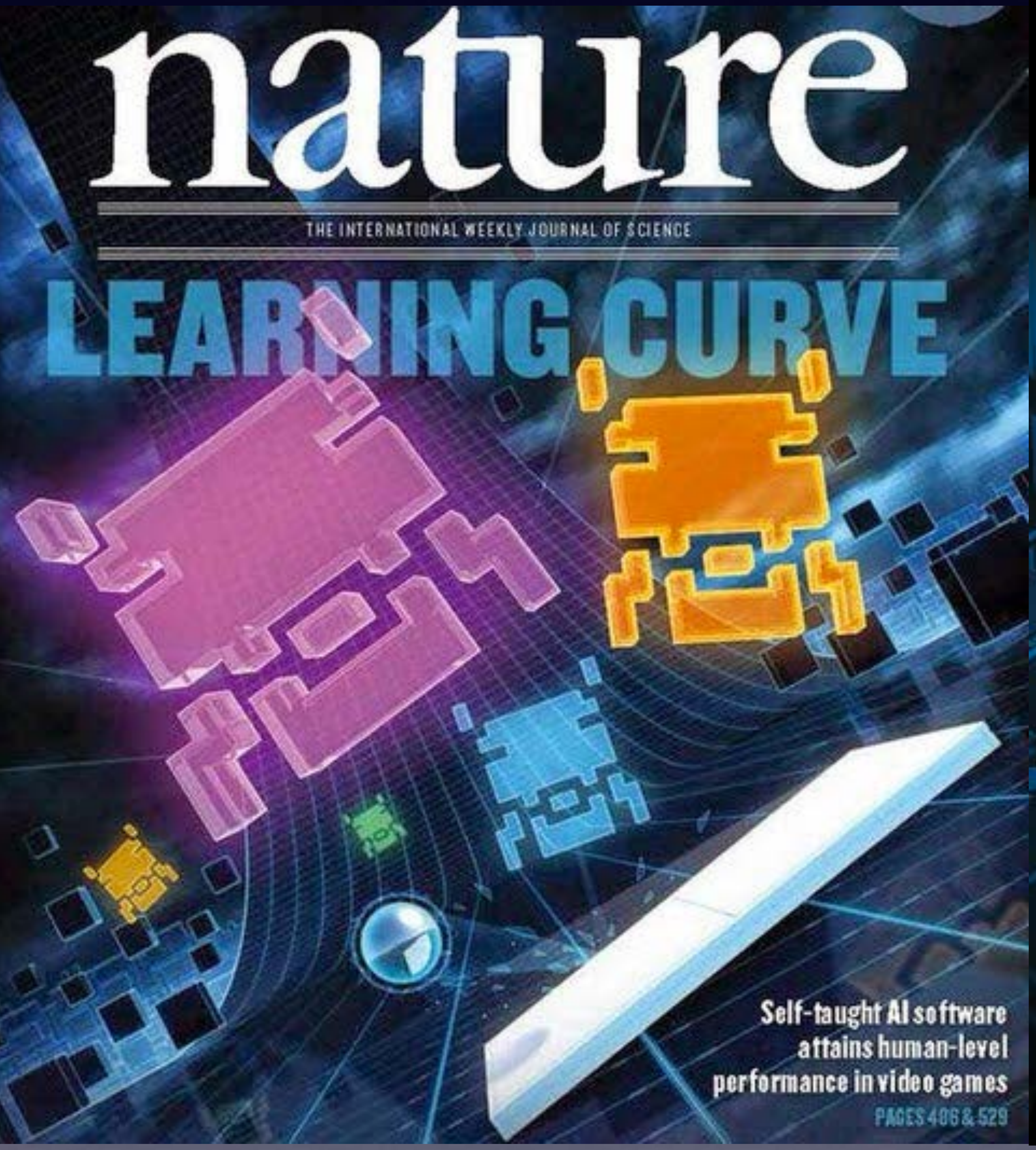
"black and white dog jumps over bar."

Understanding images

Describing them

Karpathy & Fei-Fei 2015

Deep Reinforcement Learning

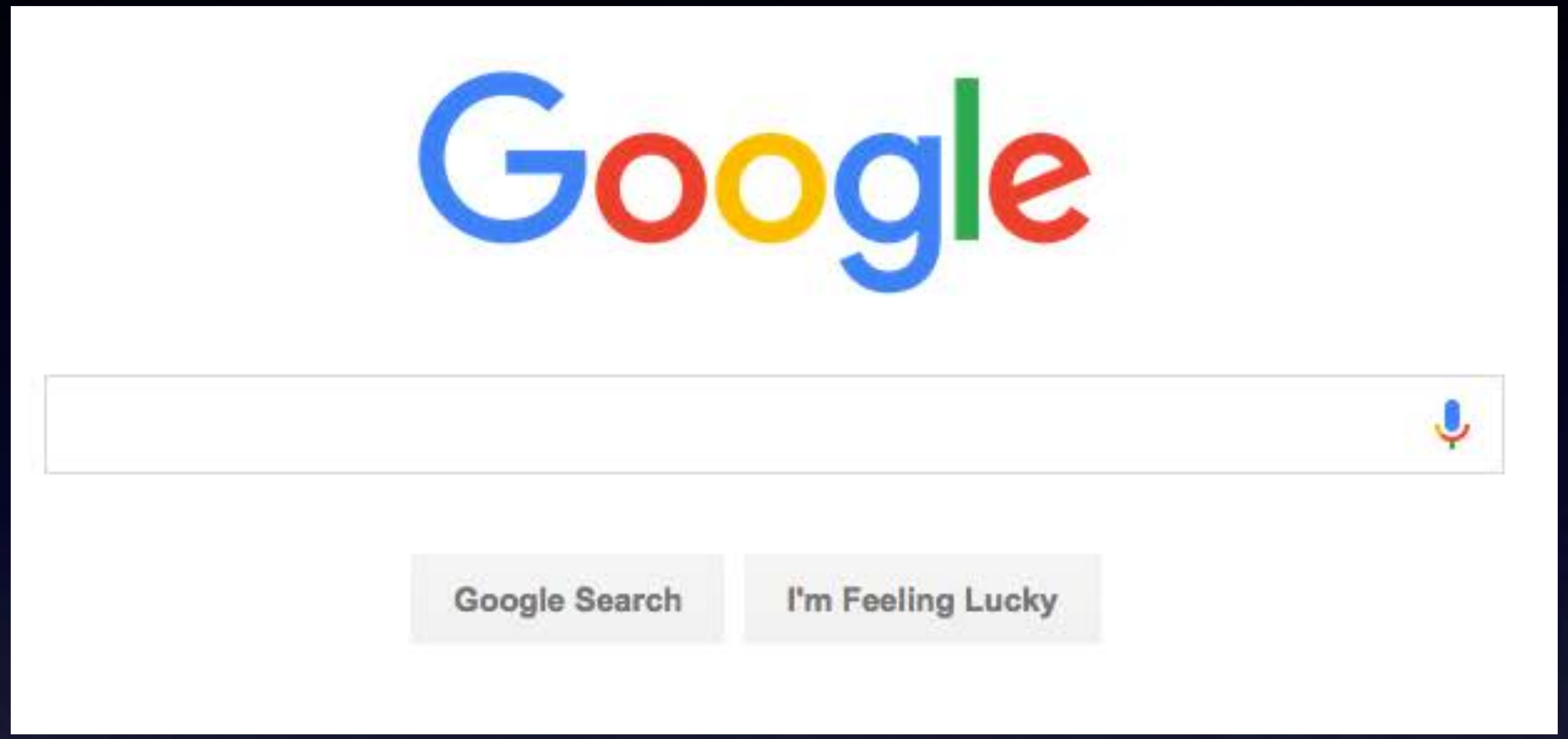


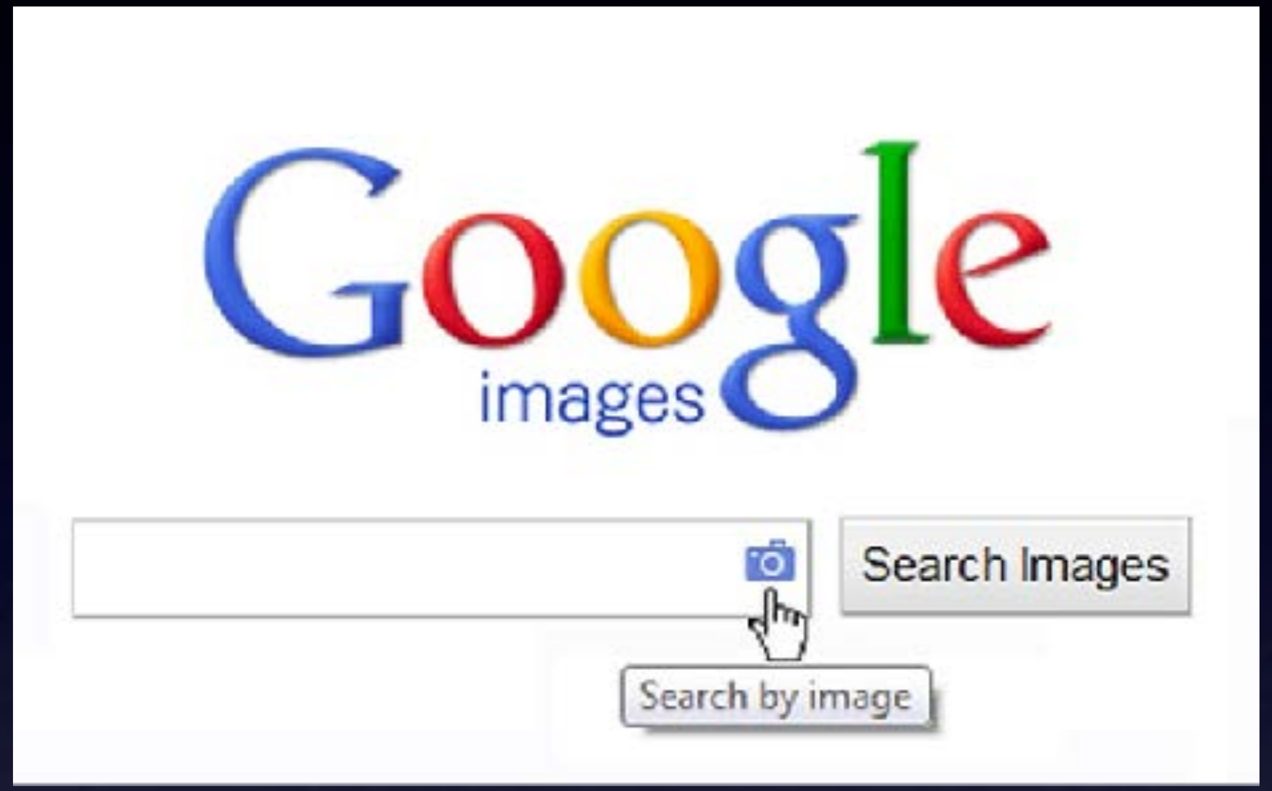

DeepMind


OpenAI

Just within Google

- Search
- Search by image
- Driverless cars
- Youtube recommendations
  - videos
  - thumbnails
- Maps
  - reading street addresses
- Etc.





facebook

- Tagging
- Determining close friends?

Every Major Company

Live translation

Deep Visualization Toolbox

yosinski.com/deepvis

#deepvis

Jason Yosinski

Jeff Clune

Anh Nguyen

Thomas Fuchs

Hod Lipson

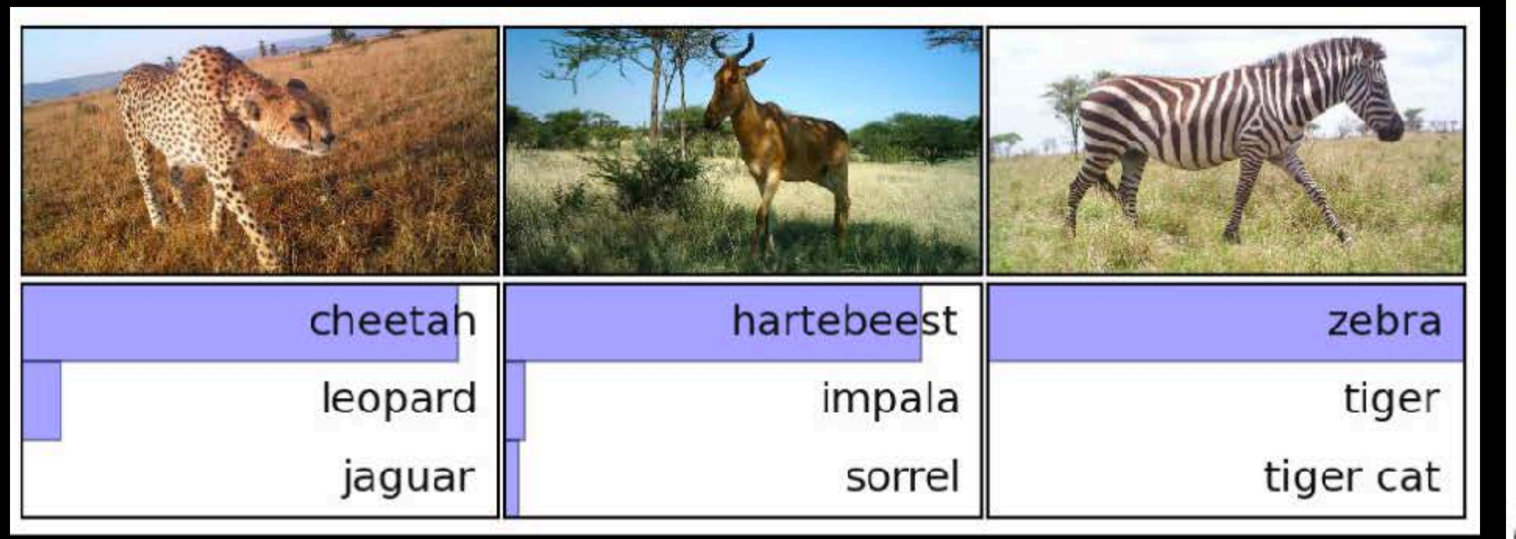
Cornell University

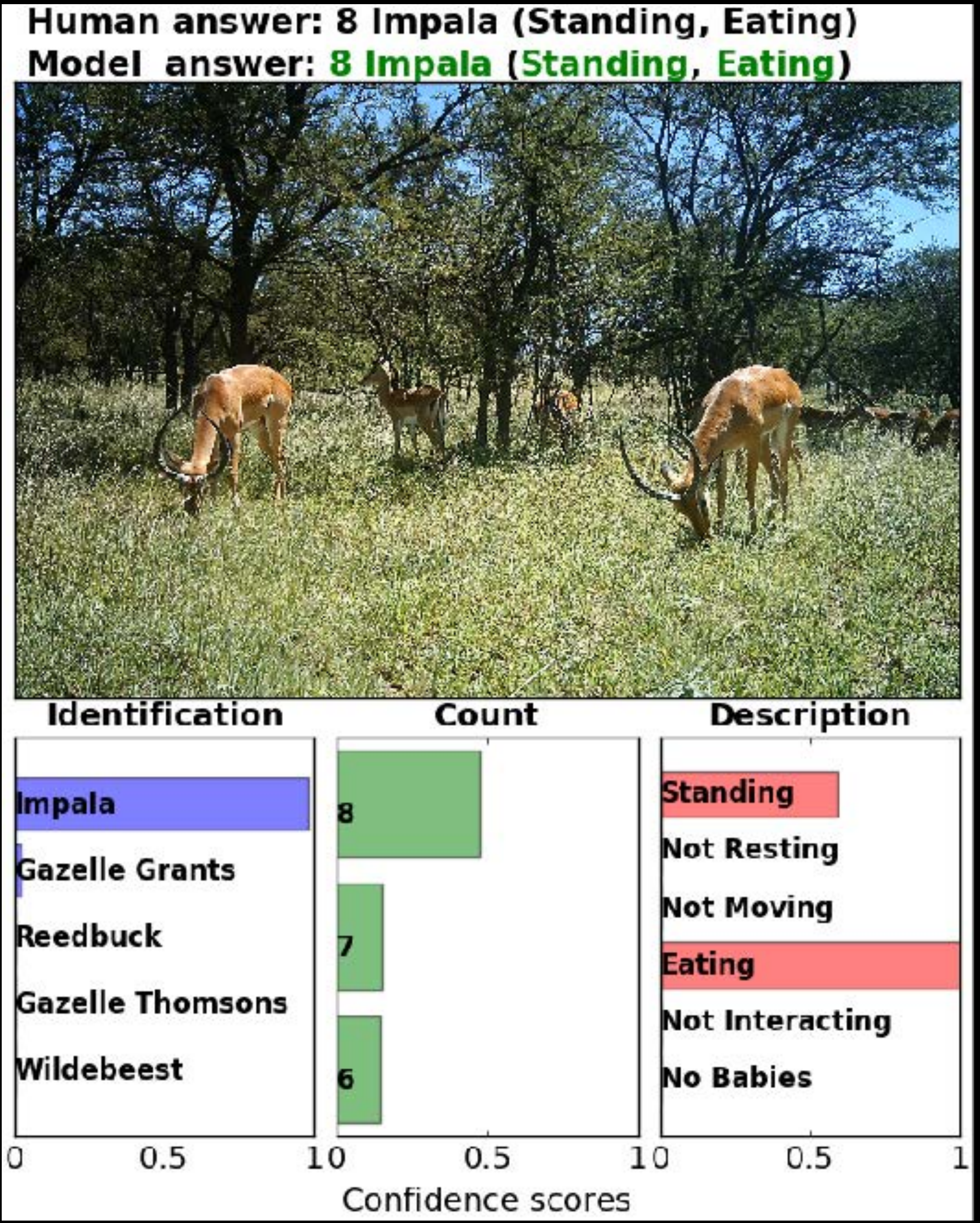
University of Wyoming

Jet Propulsion Laboratory
California Institute of Technology

Automated Ecological Understanding

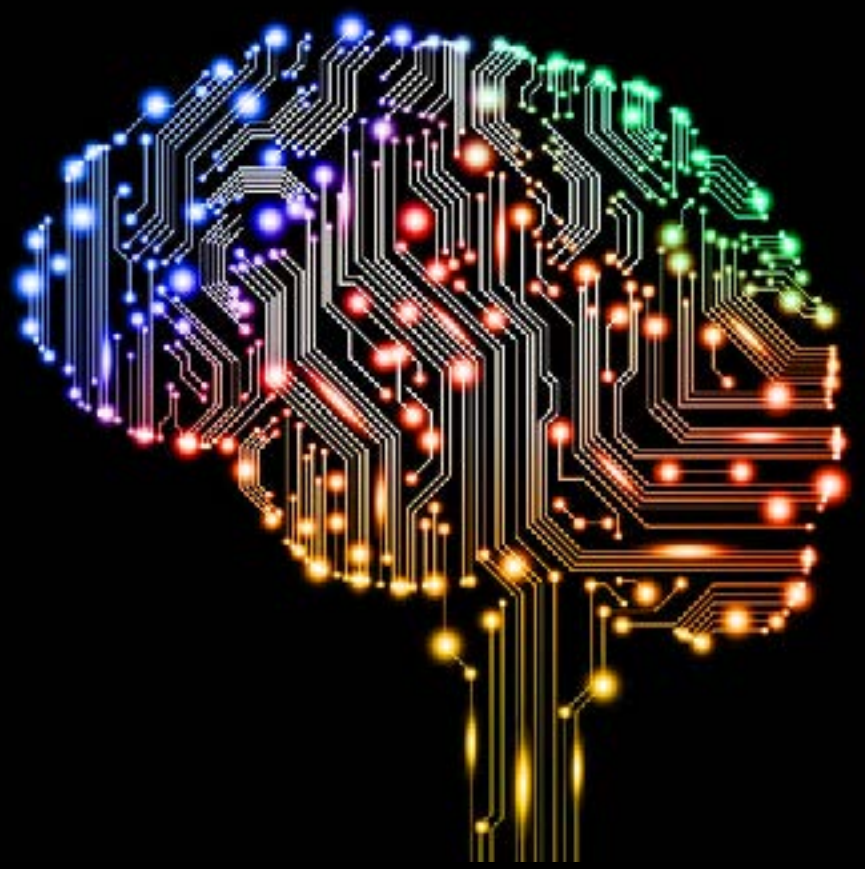
- 17,000 human hours to label. 3.2 million images. We automated 99.3% with human-level accuracy with deep neural networks
- Stop poaching, protect endangered species, transform ecology





Project 1:

“AI Neuroscience”: How much do deep neural networks understand about the images they classify?



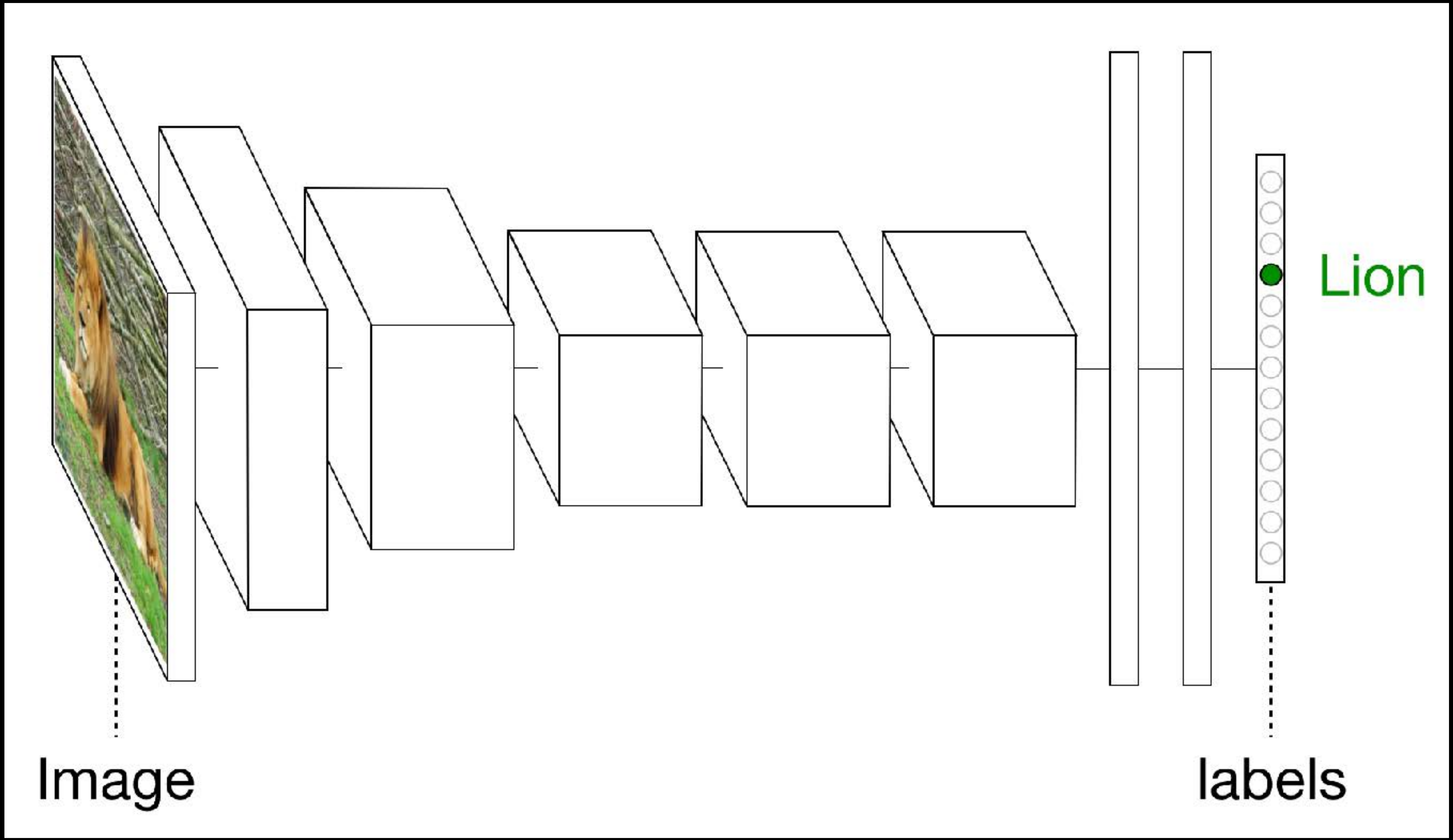
Main collaborators:

Anh Nguyen

Jason Yosinski

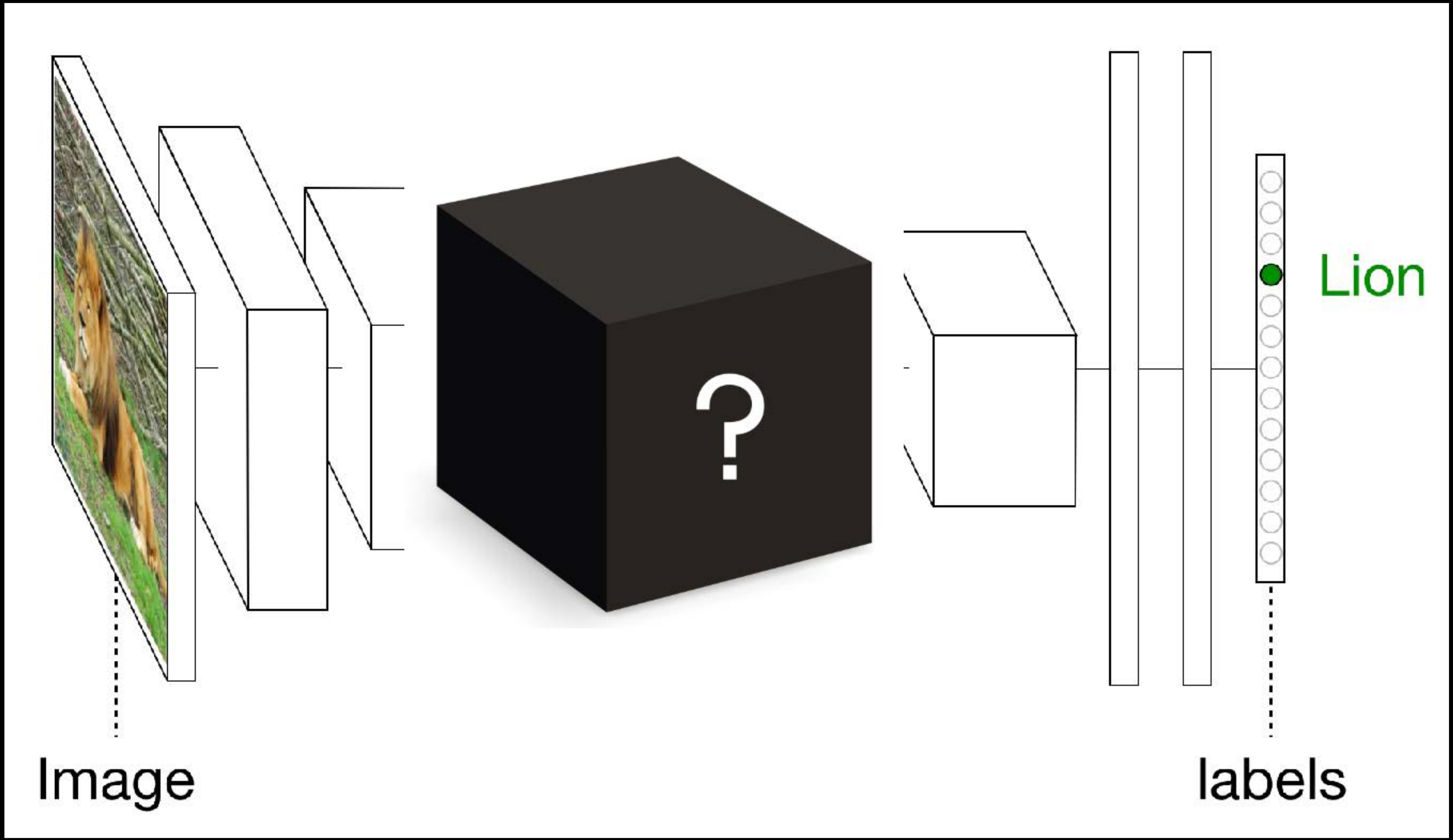
Alexey Dosovitskiy

Deep Neural Networks/Deep Learning



~1M neurons
~100M weights

Deep Neural Networks/Deep Learning

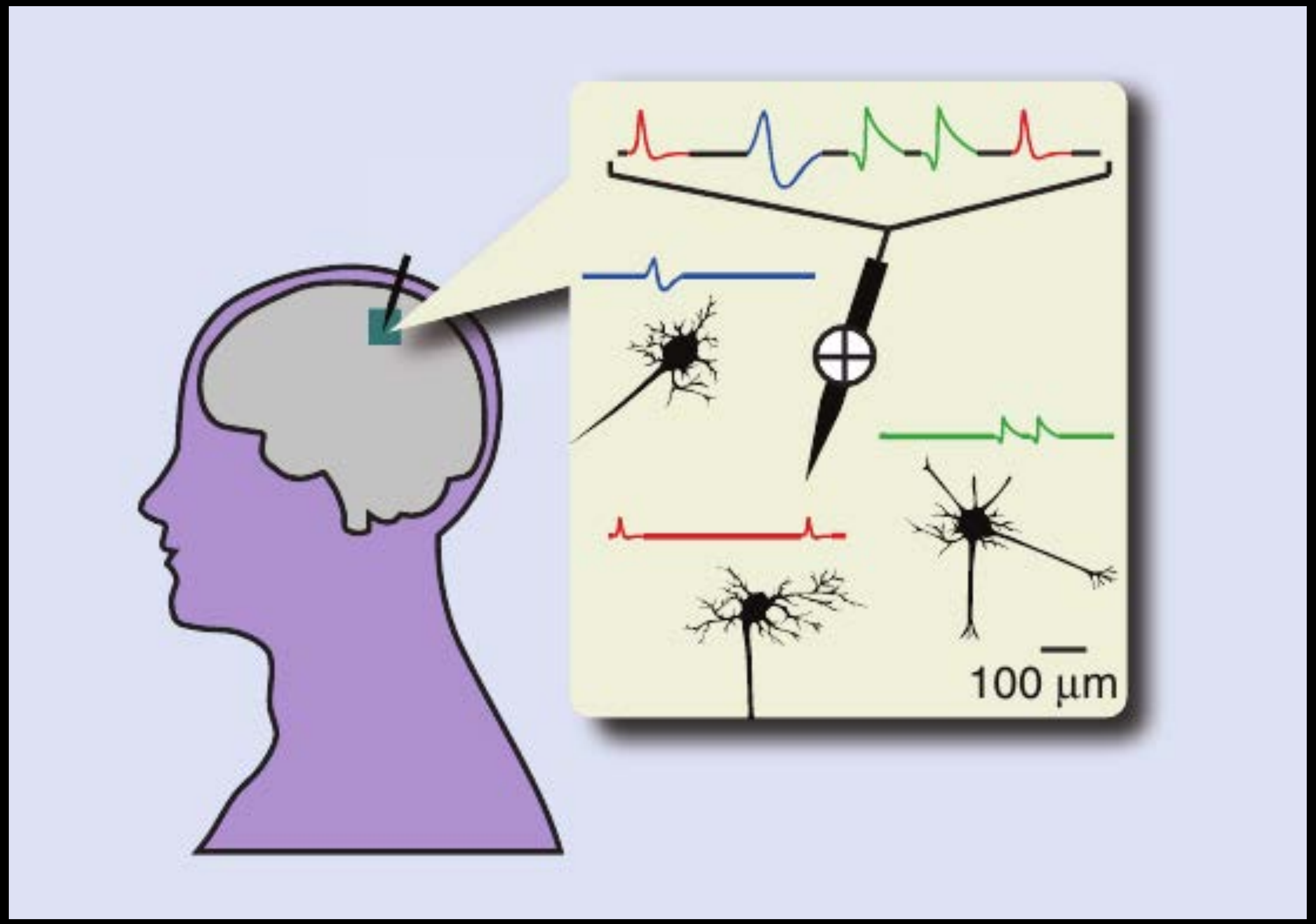


~1M neurons
~100M weights

One neuroscientist method: investigate function of individual neurons

- Record a single neuron
- Show it pictures
- See what it responds to



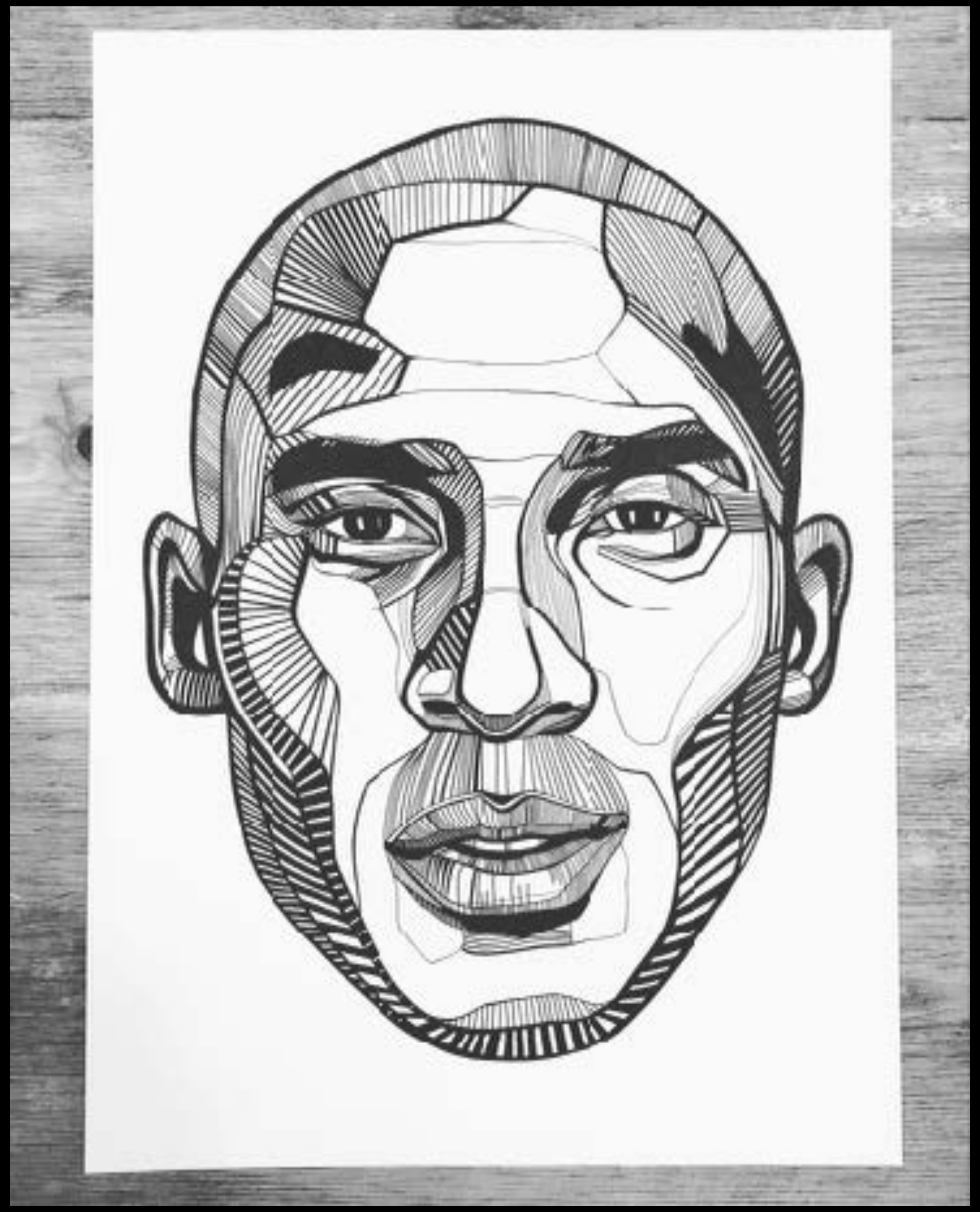




etc….

Quiroga et al. Nature 2005

"Kobe Bryant Neuron"

Quiroga et al. Nature 2005



Kobe Bryant

Multifaceted

Open Questions

- Is it really a Kobe Bryant neuron?
  - or a basketball player neuron?
  - or an LA laker neuron?
- Can’t show all possible images

Ideal Test: Synthesize Preferred Inputs

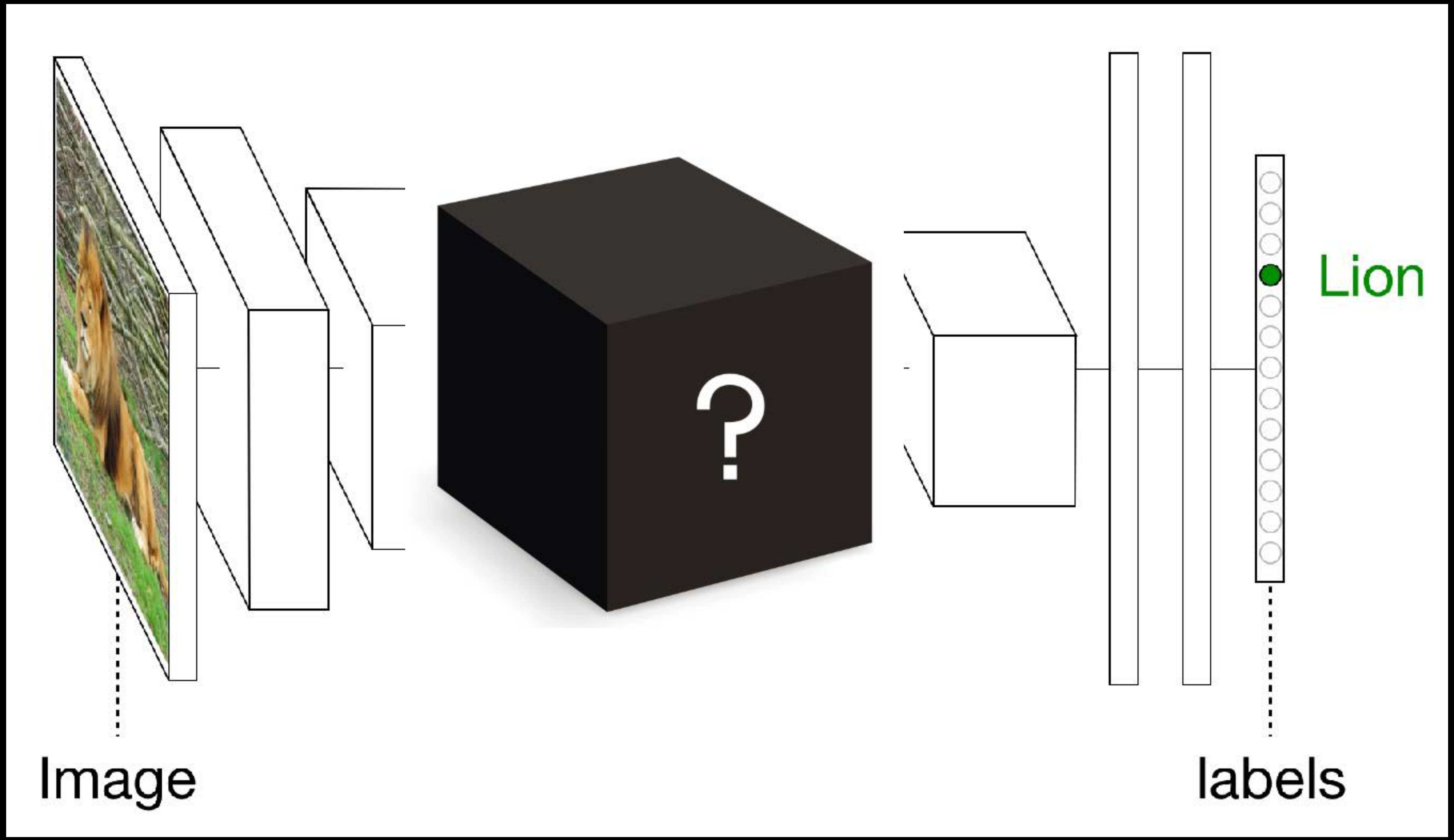
Ideal Test: Synthesize Preferred Inputs

LA Laker neuron

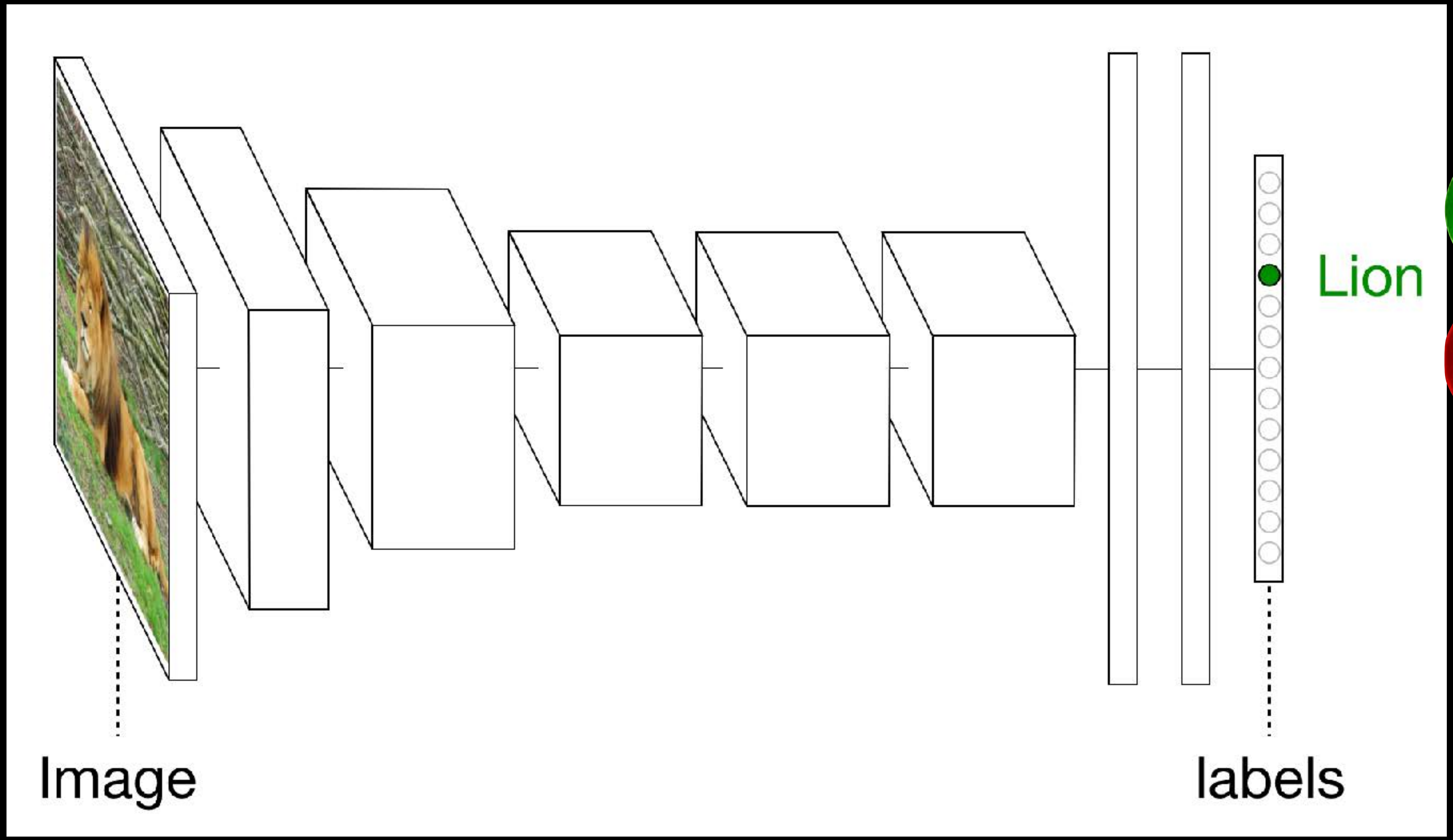
Ideal Test: Synthesize Preferred Inputs

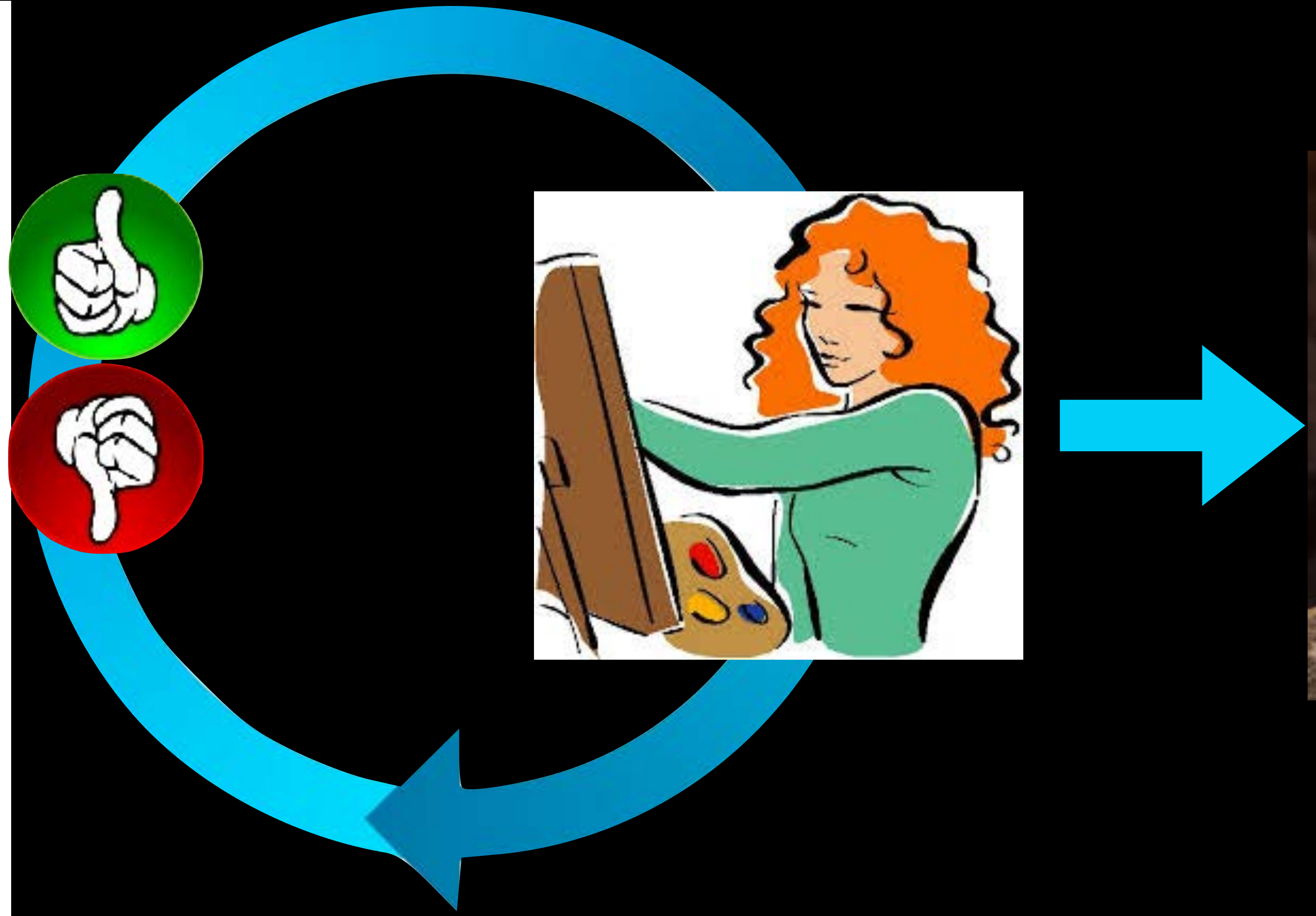
Kobe Bryant neuron

Possible with Artificial Neural Networks



Investigating What Each Neuron Does

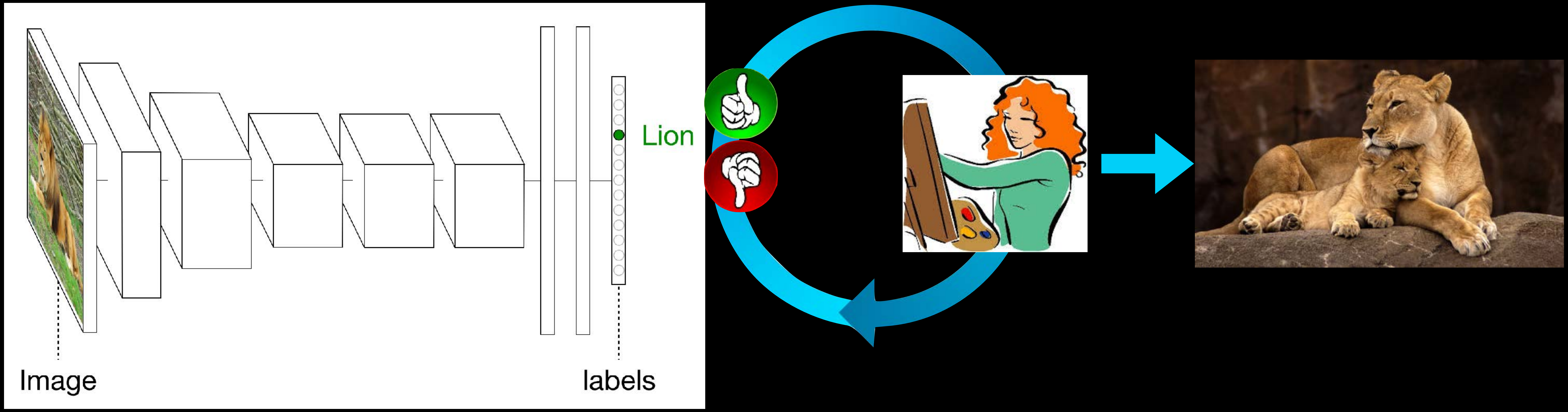


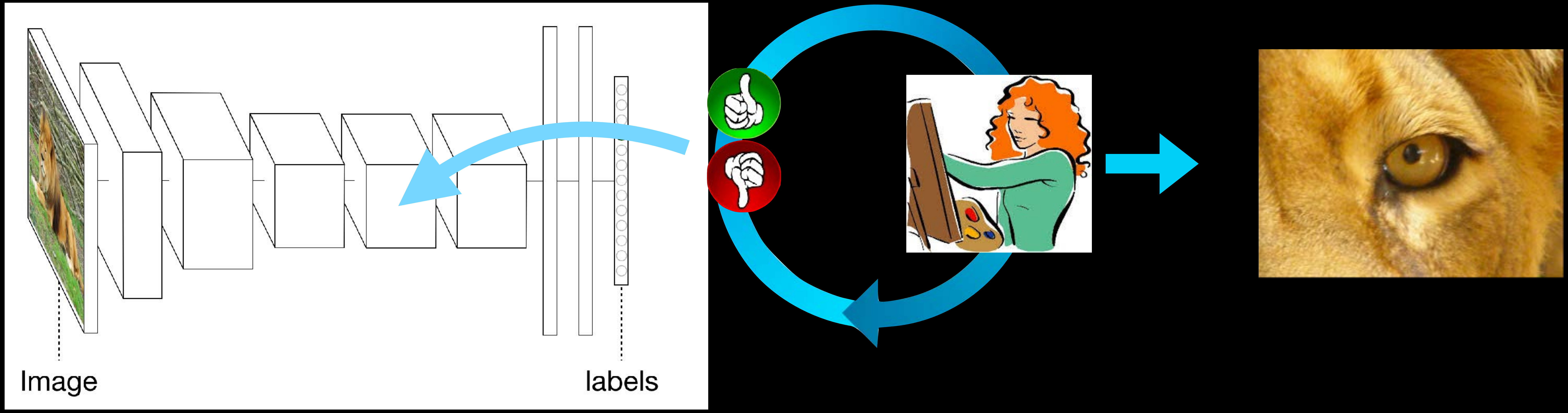


Pretrained, Fixed DNN

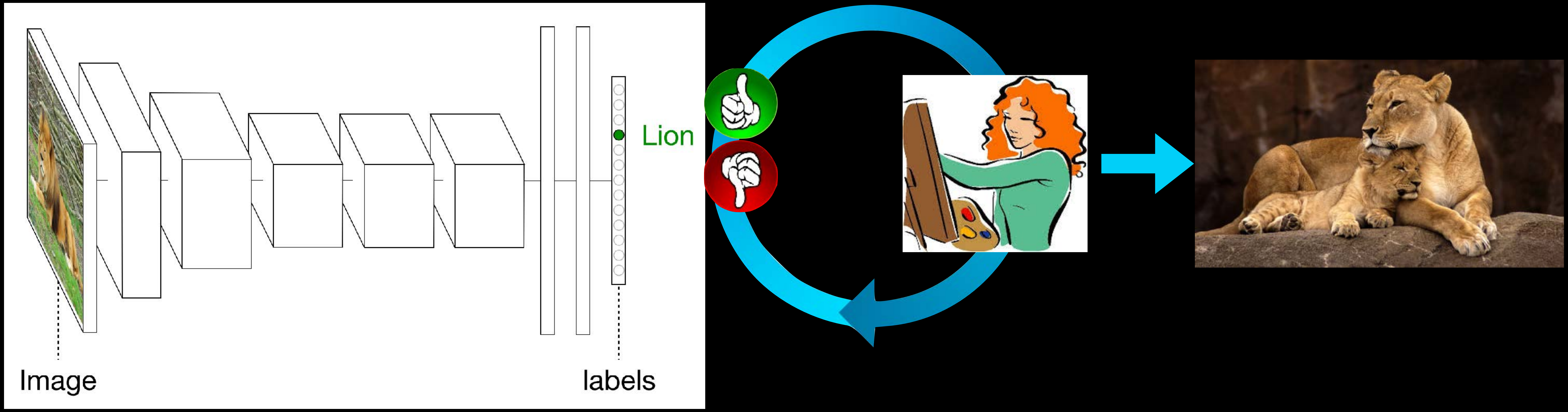
Optimize Pixels
e.g. via Backprop

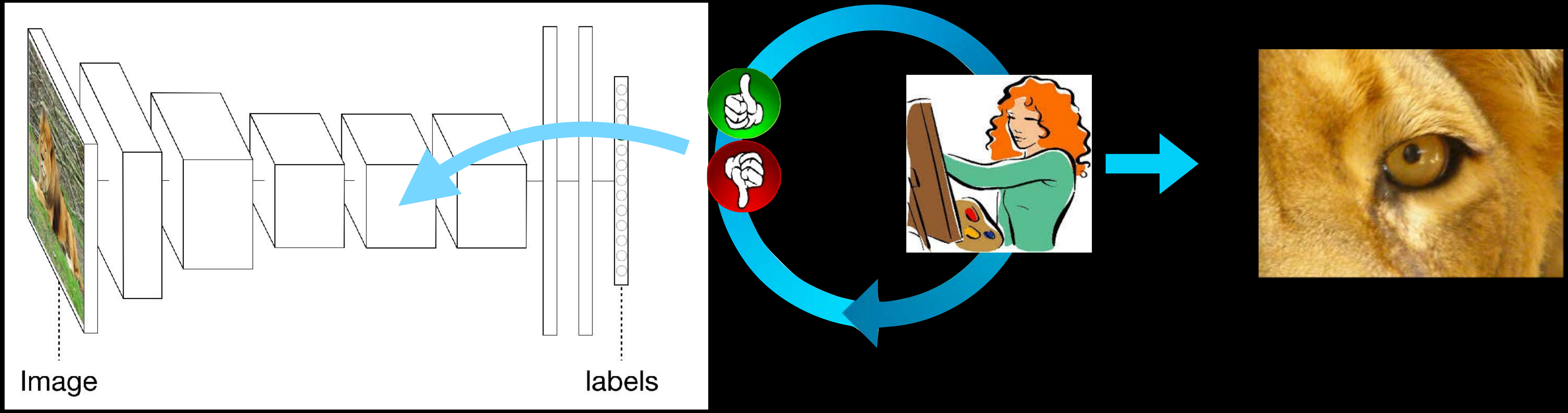
Investigating What Each Neuron Does





“Deep Visualization”



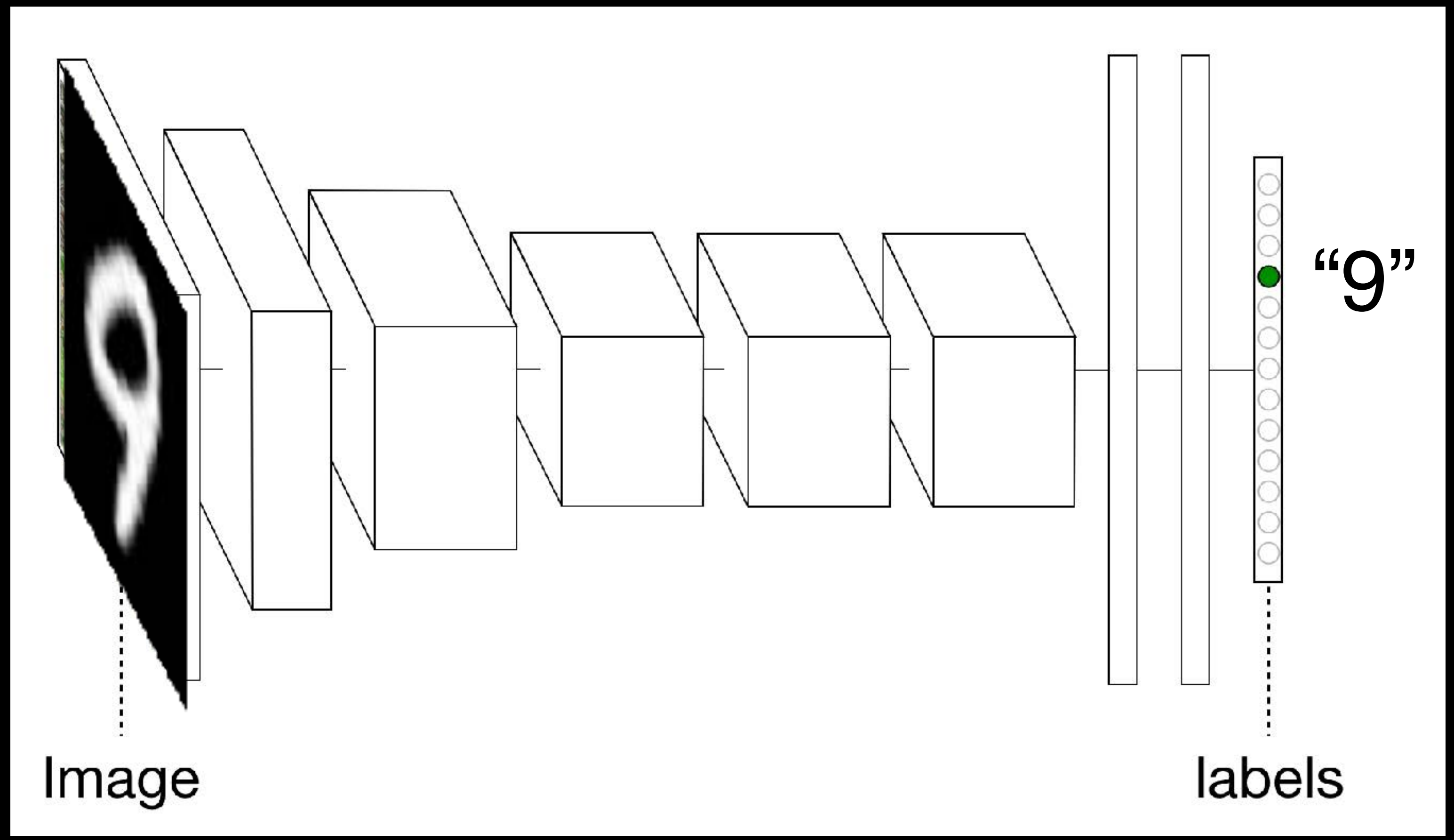


Deep Visualization Take 1

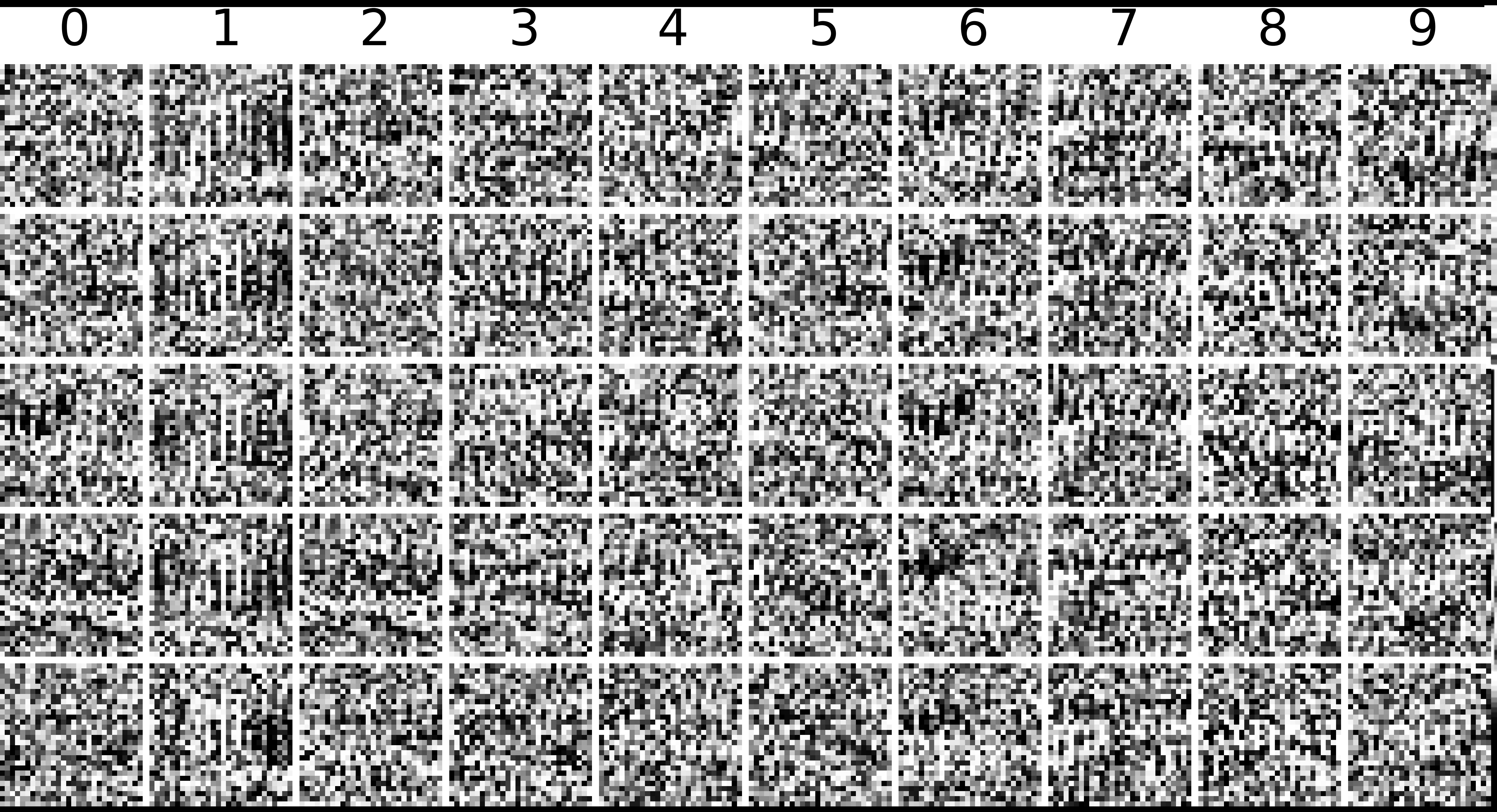
Nguyen, Yosinski, Clune, 2015, CVPR

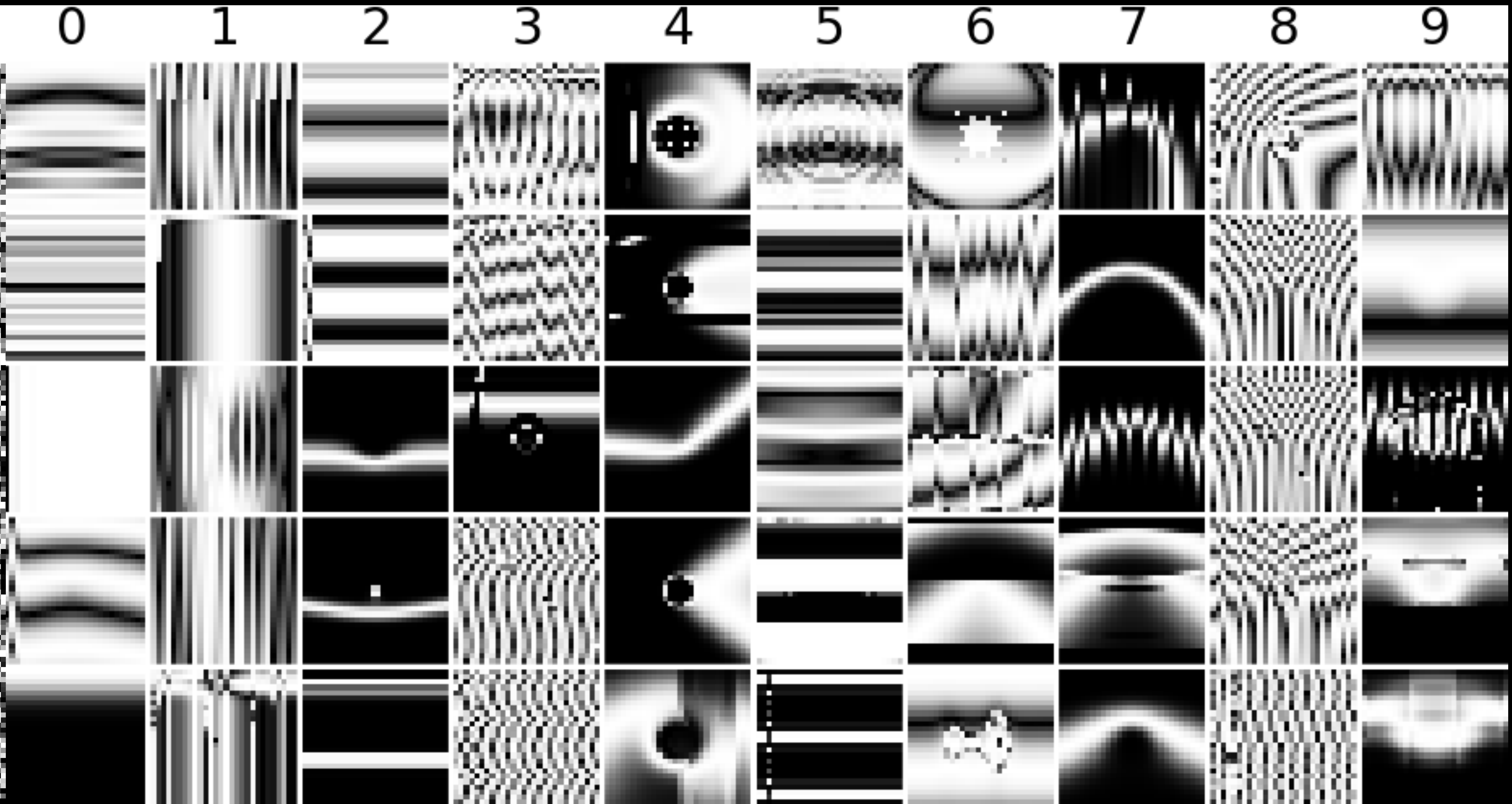
DNN Confidence: > 99.6 % for all

Digits

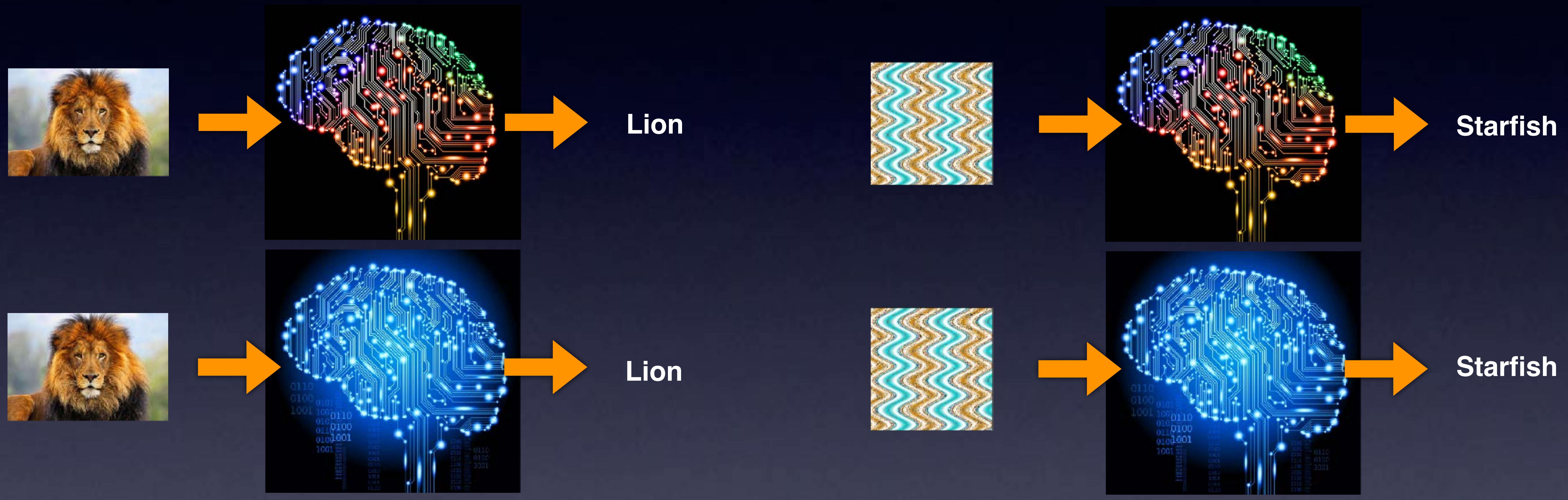


>99% accurate

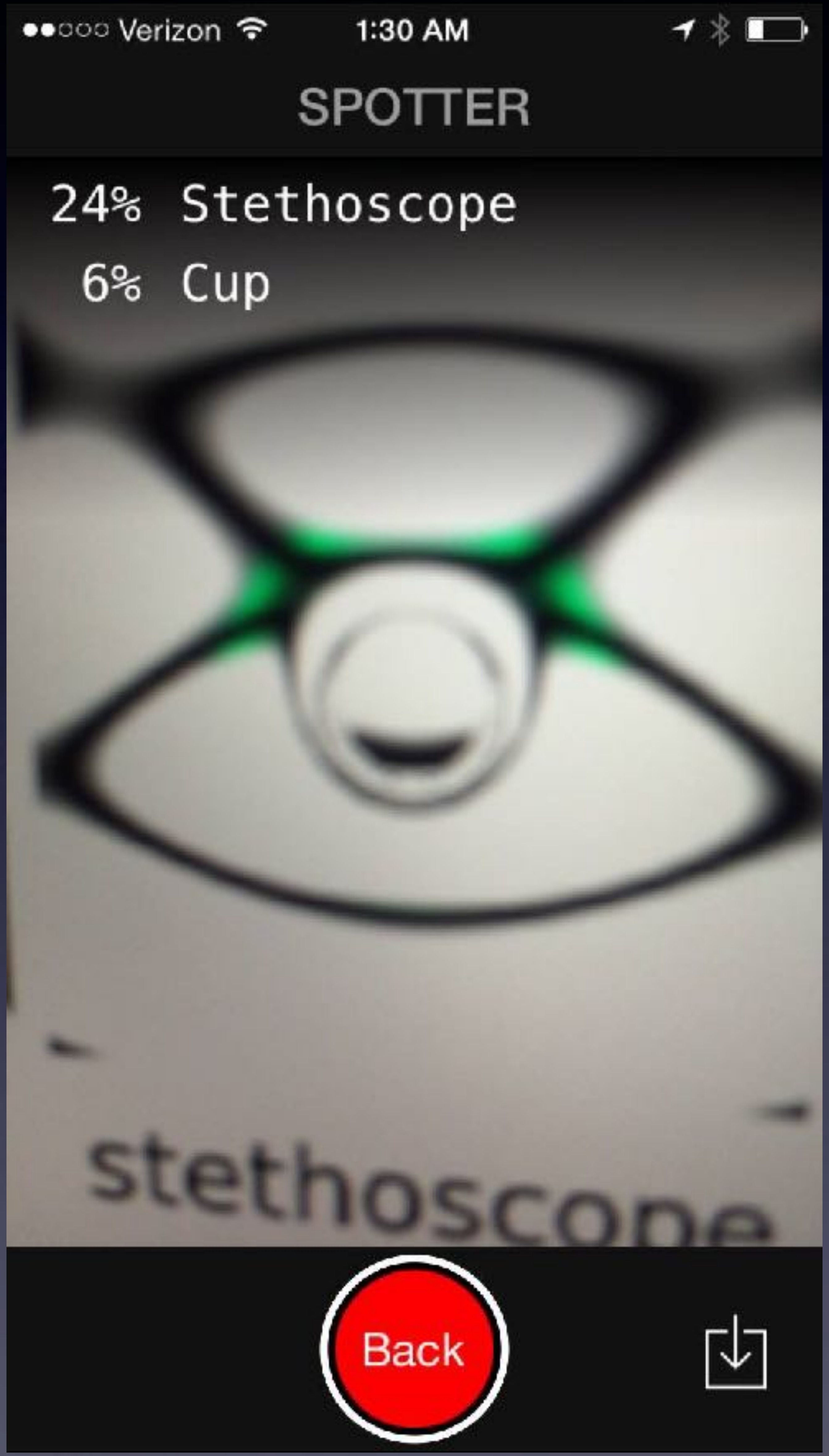


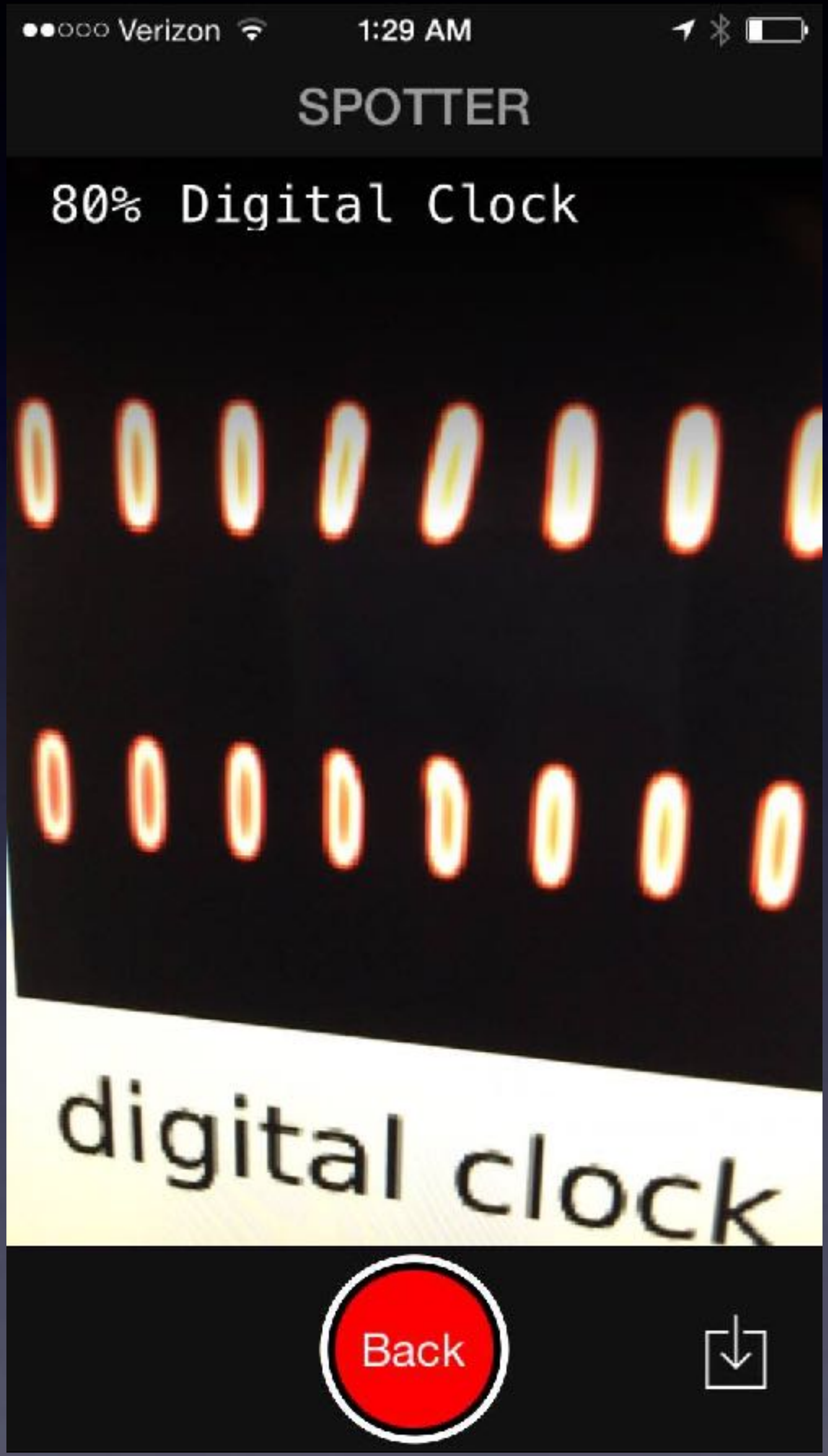


Images that fool one network fool others!



Images that fool one network fool others!





Courtesy: Dileep George, co-founder Vicarious

Huge reaction

TODO: ADD NATURE

63rd most talked about scientific paper worldwide in 2015 - Altmetric

The Economist

The Atlantic

WIRED

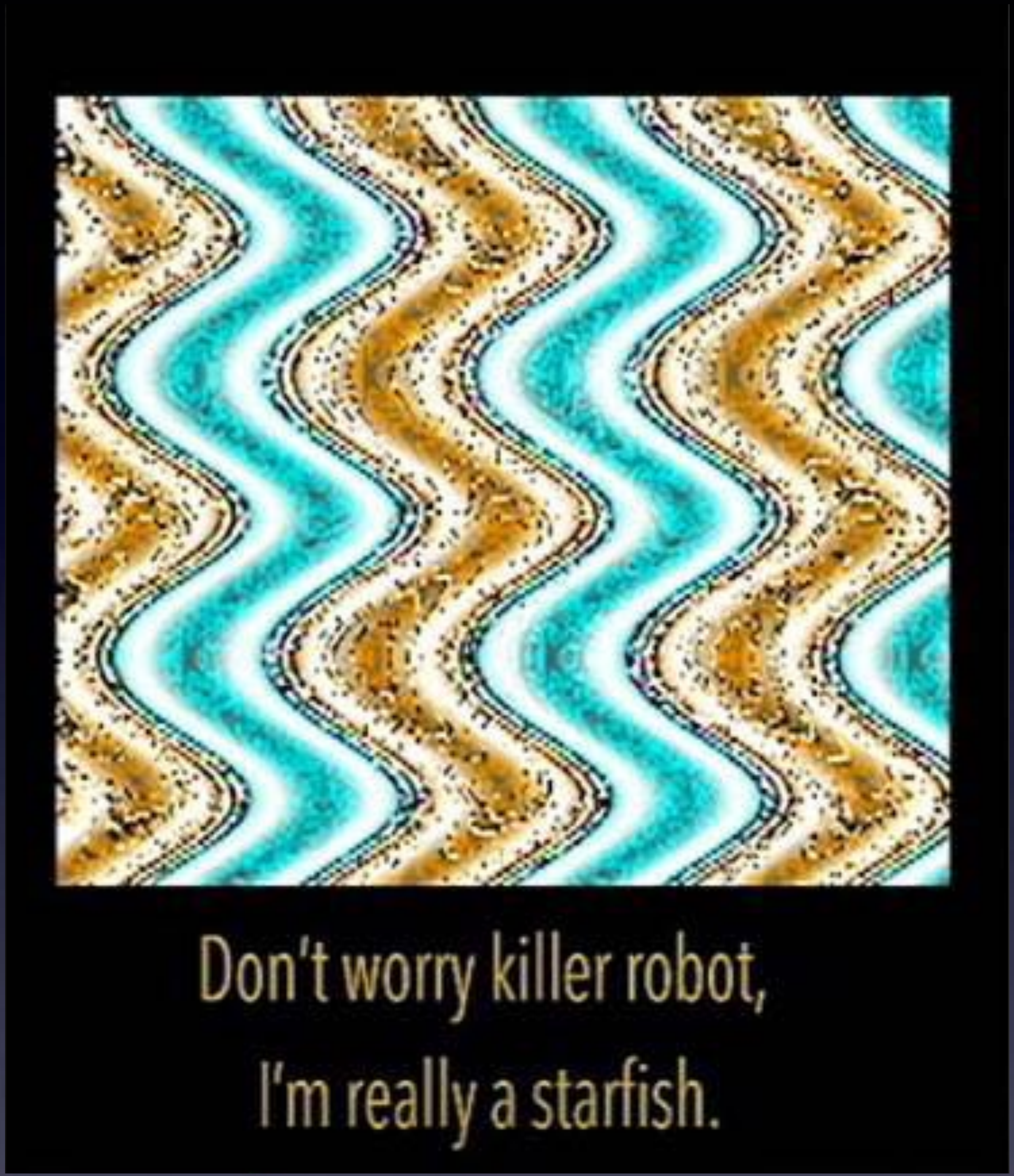
NewScientist

SCIENTIFIC AMERICAN MIND
BEHAVIOR • BRAIN SCIENCE • INSIGHTS

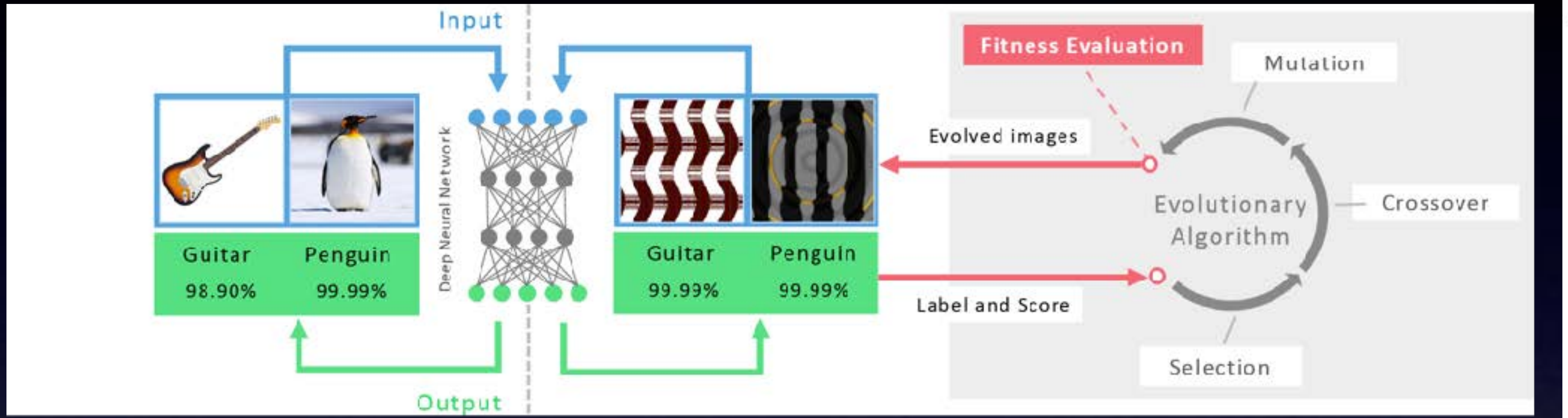
MIT Technology Review

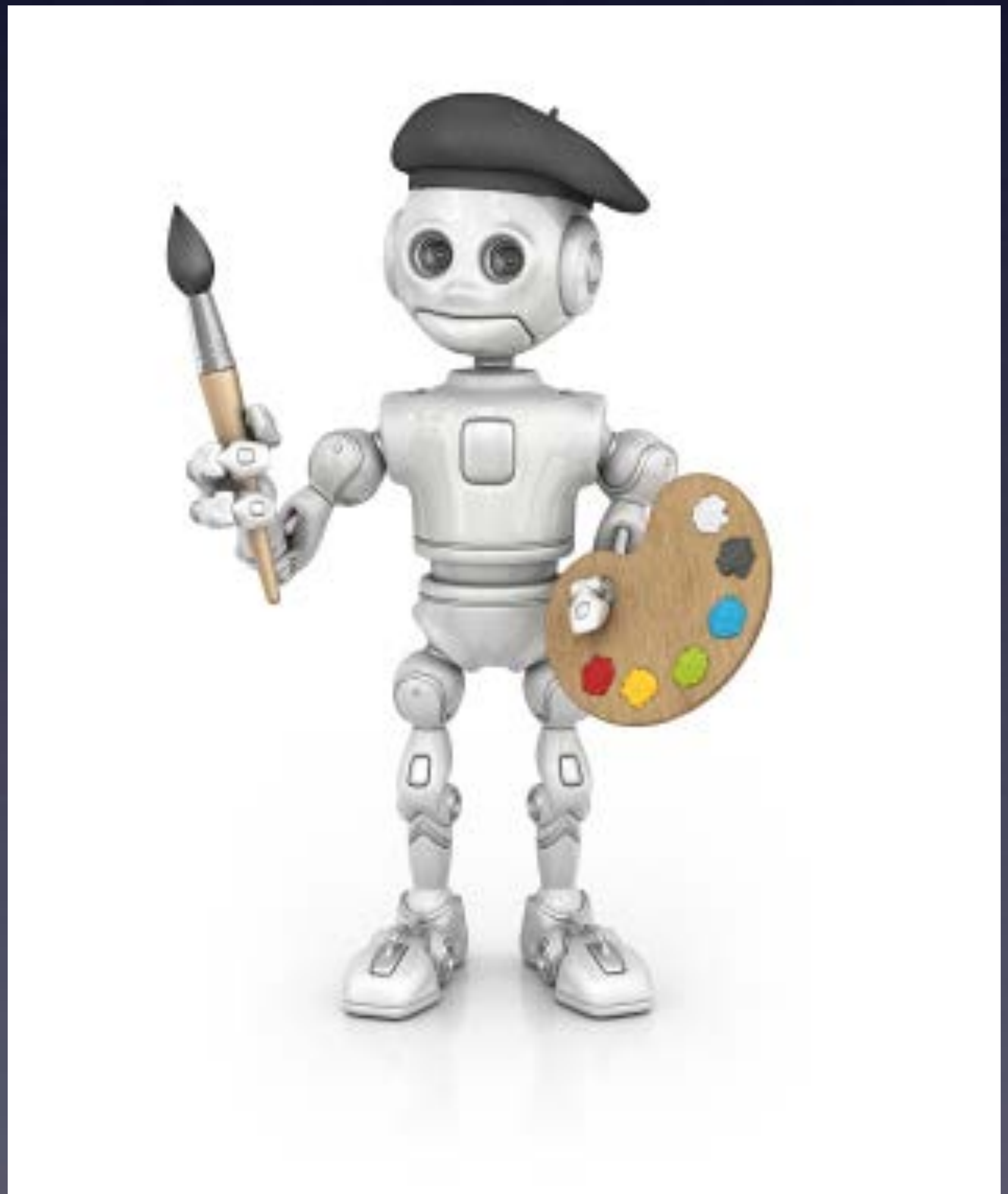
Larry Page,
Google co-Founder

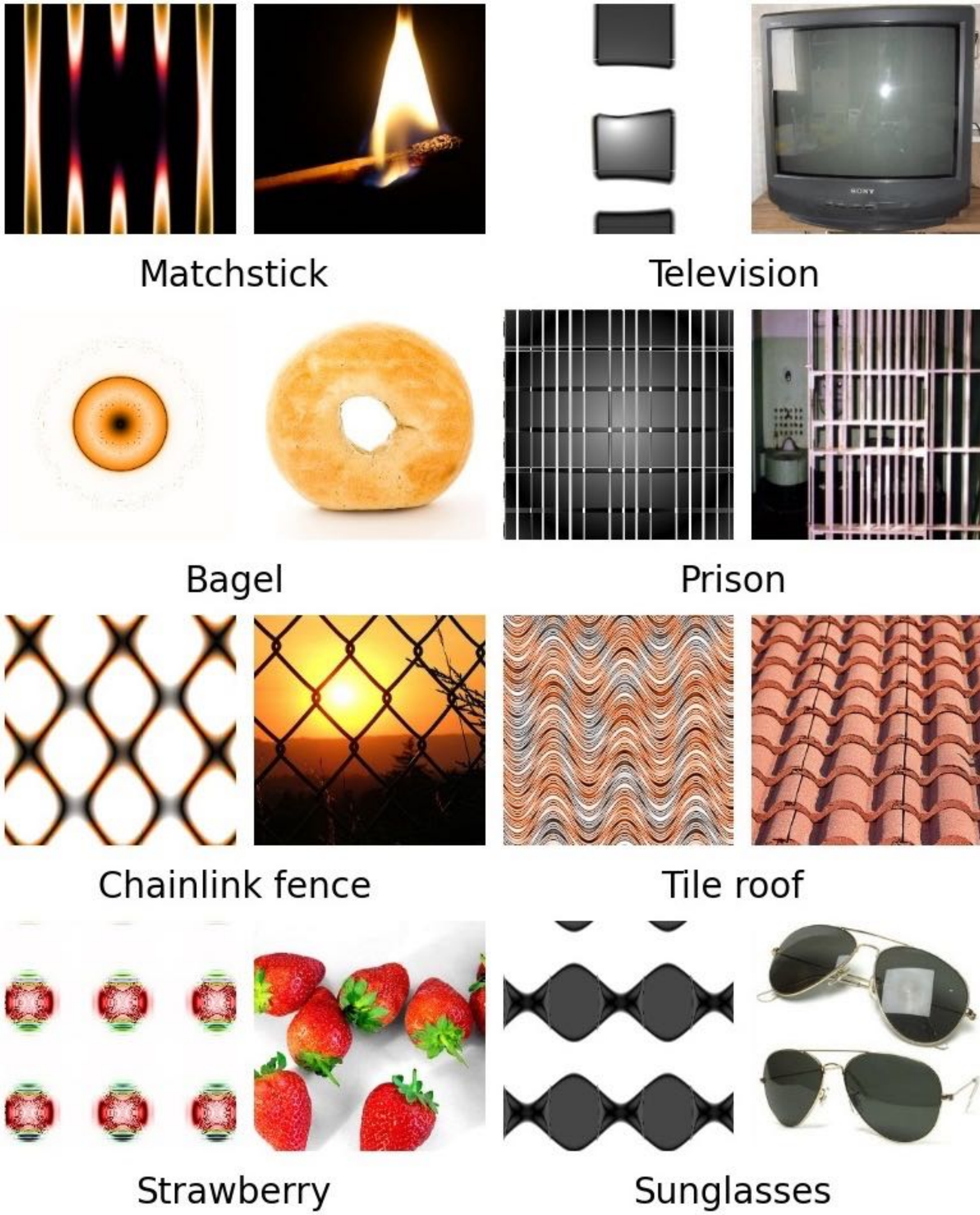
Gary Marcus,
NYU Prof & CEO

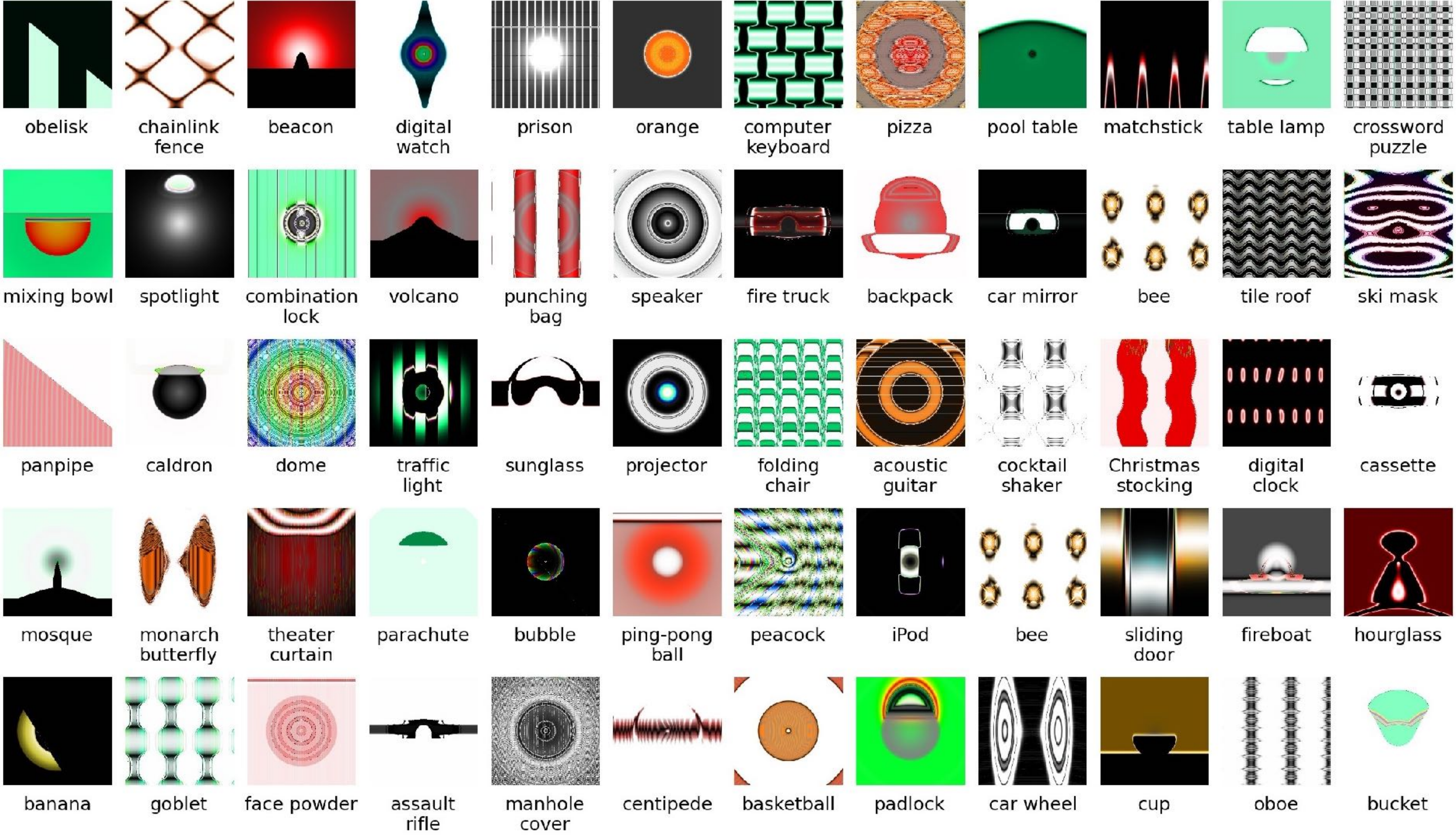


Automatic Art Generator







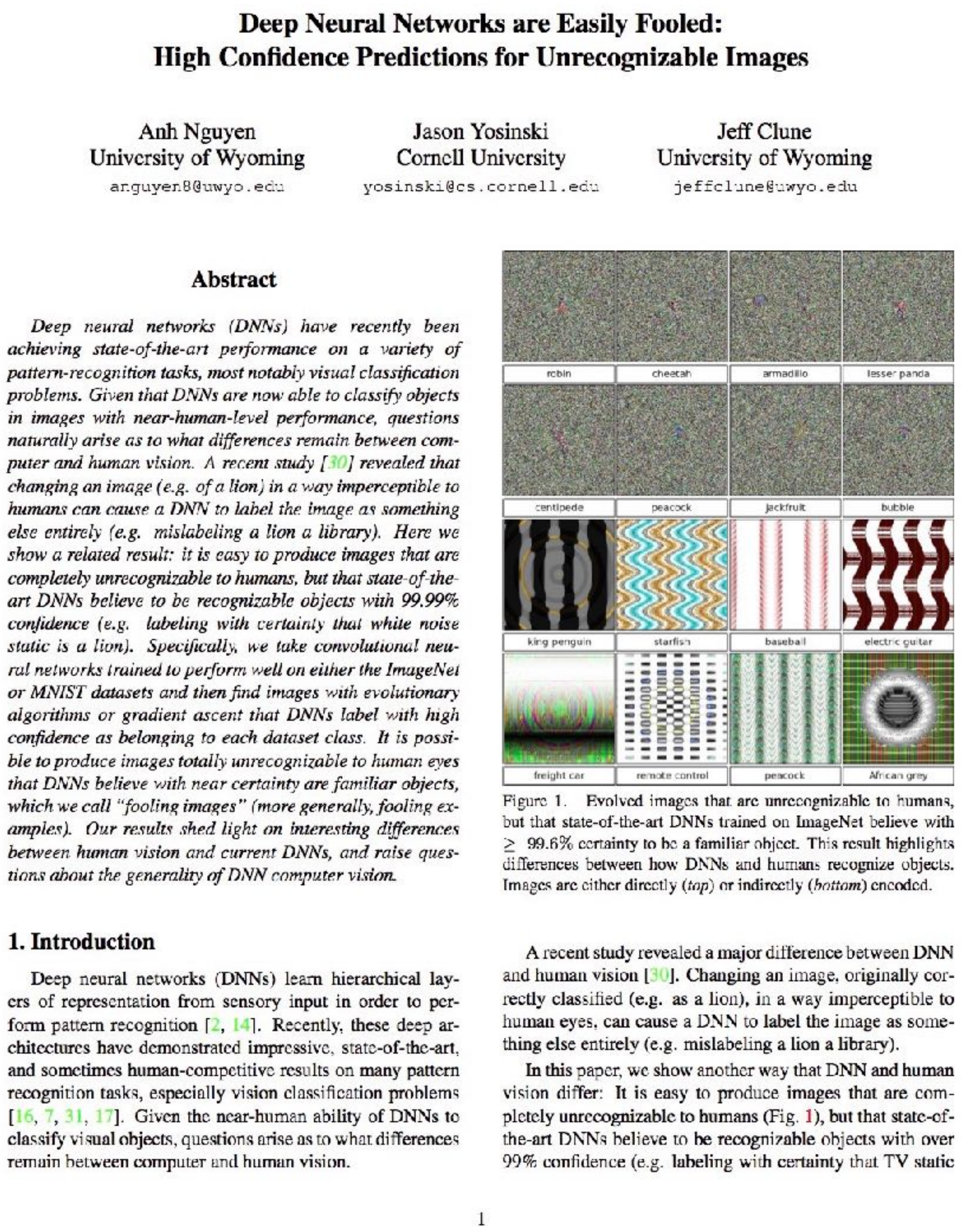

obelisk
chainlink fence
beacon
digital watch
prison
orange
computer keyboard
pizza
pool table
matchstick
table lamp
crossword puzzle
mixing bowl
spotlight
combination lock
volcano
punching bag
speaker
fire truck
backpack
car mirror
bee
tile roof
ski mask
panpipe
caldron
dome
traffic light
sunglass
projector
folding chair
acoustic guitar
cocktail shaker
Christmas stocking
digital clock
cassette
mosque
monarch butterfly
theater curtain
parachute
bubble
ping-pong ball
peacock
iPod
bee
sliding door
fireboat
hourglass
banana
goblet
face powder
assault rifle
manhole cover
centipede
basketball
padlock
car wheel
cup
oboe
bucket

- UW Museum Student Art Competition
- Judges did not know art was AI-generated (and not human artist)
- 35% acceptance rate, and an award

Innovation Engines

Nguyen, Yosinski, Clune, 2015, GECCO

- Automatically generate interesting, new solutions in any domain
  - art
  - robotics
  - engineering challenges
  - tests and informs biodiversity theories
- Interested in more?
  - ICML Tutorial: https://www.youtube.com/watch?v=g6HiuEnbwJE
  - CORL Keynote: https://www.youtube.com/watch?v=zpUD9rf5YaQ&t=15069s



Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images

Anh Nguyen
University of Wyoming
anguyen8@uwyo.edu

Jason Yosinski
Cornell University
yosinski@cs.cornell.edu

Jeff Clune
University of Wyoming
jeffclune@uwyo.edu

Abstract

Deep neural networks (DNNs) have recently been achieving state-of-the-art performance on a variety of pattern-recognition tasks, most notably visual classification problems. Given that DNNs are now able to classify objects in images with near-human-level performance, questions naturally arise as to what differences remain between computer and human vision. A recent study [30] revealed that changing an image (e.g. of a lion) in a way imperceptible to humans can cause a DNN to label the image as something else entirely (e.g. mislabeling a lion a library). Here we show a related result: it is easy to produce images that are completely unrecognizable to humans, but that state-of-the-art DNNs believe to be recognizable objects with 99.99% confidence (e.g. labeling with certainty that white noise static is a lion). Specifically, we take convolutional neural networks trained to perform well on either the ImageNet or MNIST datasets and then find images with evolutionary algorithms or gradient ascent that DNNs label with high confidence as belonging to each dataset class. It is possible to produce images totally unrecognizable to human eyes that DNNs believe with near certainty are familiar objects, which we call "fooling images" (more generally, fooling examples). Our results shed light on interesting differences between human vision and current DNNs, and raise questions about the generality of DNN computer vision.

robin | cheetah | armadillo | lesser panda
centipede | peacock | jackfruit | bubble
king penguin | starfish | baseball | electric guitar
freight car | remote control | peacock | African grey

Figure 1. Evolved images that are unrecognizable to humans, but that state-of-the-art DNNs trained on ImageNet believe with ≥ 99.6% certainty to be a familiar object. This result highlights differences between how DNNs and humans recognize objects. Images are either directly (*top*) or indirectly (*bottom*) encoded.

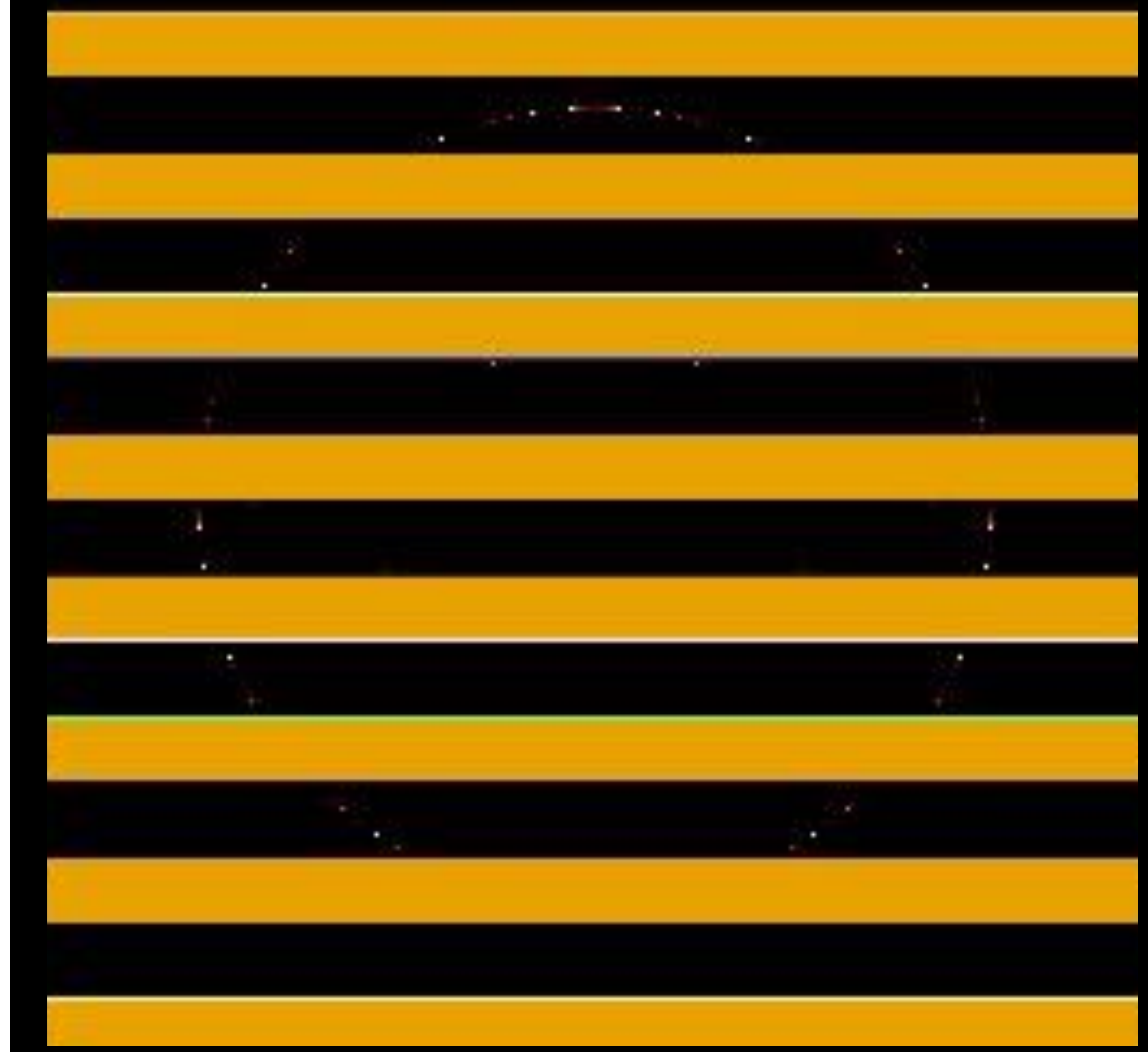
1. Introduction

Deep neural networks (DNNs) learn hierarchical layers of representation from sensory input in order to perform pattern recognition [2, 14]. Recently, these deep architectures have demonstrated impressive, state-of-the-art, and sometimes human-competitive results on many pattern recognition tasks, especially vision classification problems [16, 7, 31, 17]. Given the near-human ability of DNNs to classify visual objects, questions arise as to what differences remain between computer and human vision.

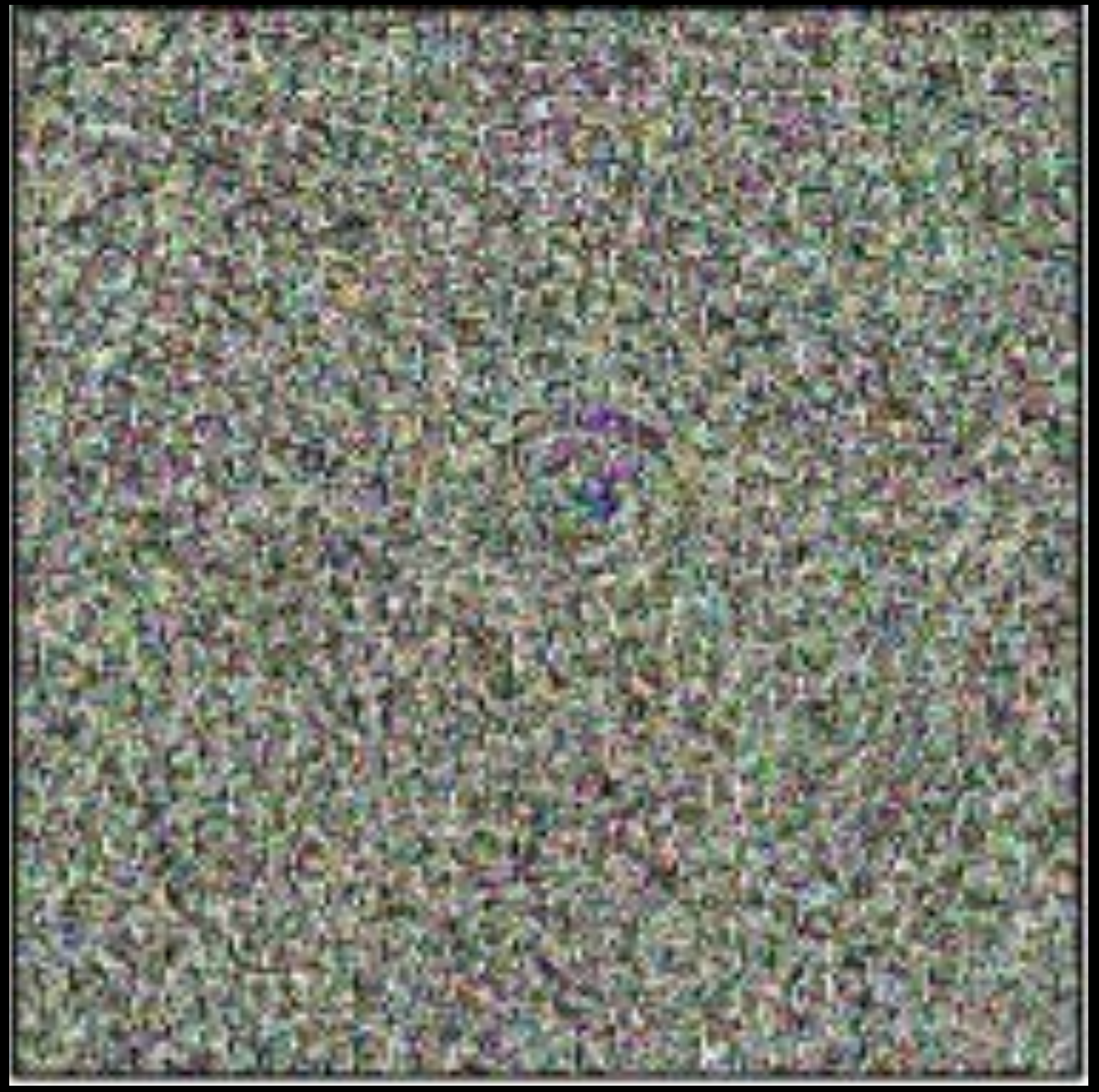
A recent study revealed a major difference between DNN and human vision [30]. Changing an image, originally correctly classified (e.g. as a lion), in a way imperceptible to human eyes, can cause a DNN to label the image as something else entirely (e.g. mislabeling a lion a library).

In this paper, we show another way that DNN and human vision differ: It is easy to produce images that are completely unrecognizable to humans (Fig. 1), but that state-of-the-art DNNs believe to be recognizable objects with over 99% confidence (e.g. labeling with certainty that TV static

- May not understand much
- Huge security concern
- Helped launch avalanche of work into "adversarial & fooling examples"
  - with Szegedy et al. 2013



School bus



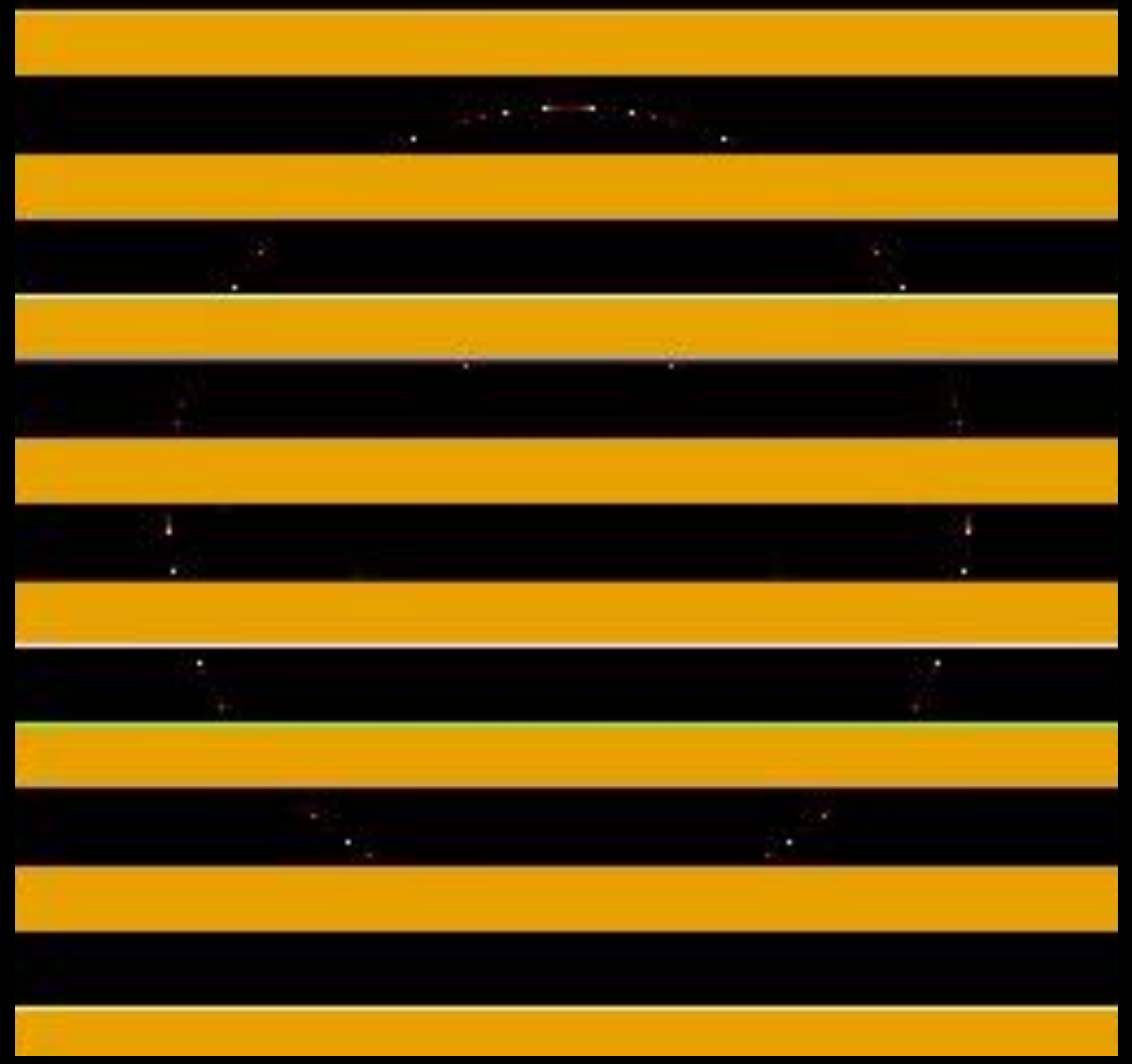
Open road!

Why are networks easily fooled?

Hypothesis 1: DNNs do understand, test is bad

Prediction: With constraints to stay in the space of natural images, we WOULD get recognizable objects.

Hypothesis 2: Only learns distinguishing features



School Bus

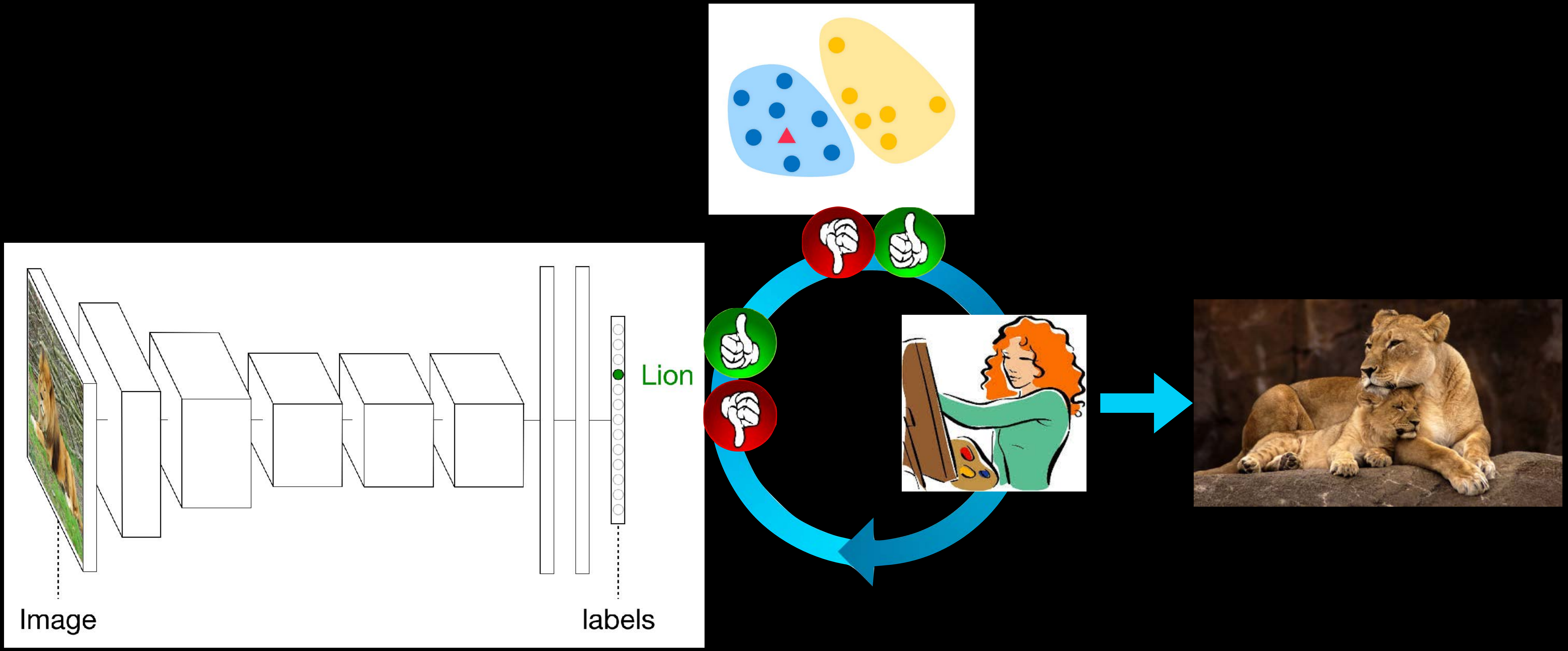
Starfish

Prediction: With constraints to stay in the space of natural images, we WOULD NOT get recognizable objects.

Our "fooling" work suggests the "DNNs don't understand" hypothesis is more likely

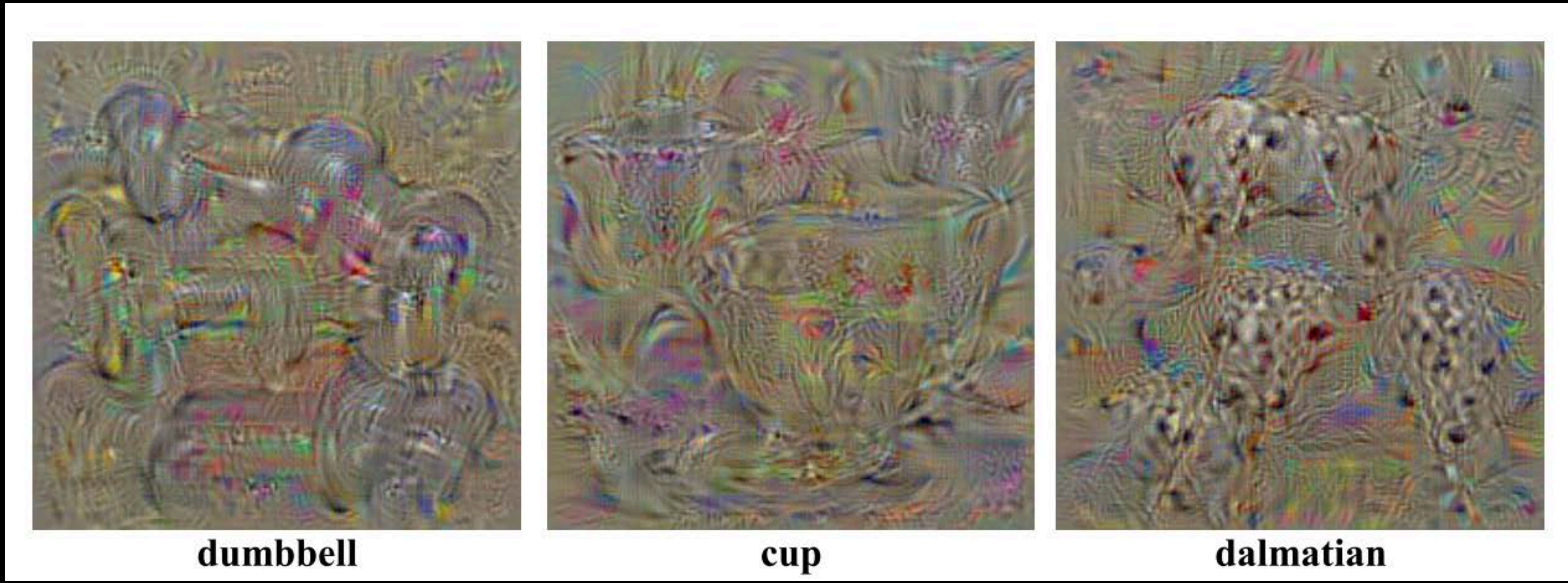
Deep Visualization Take 2

Manually Engineered Natural Image Priors



Manual Priors

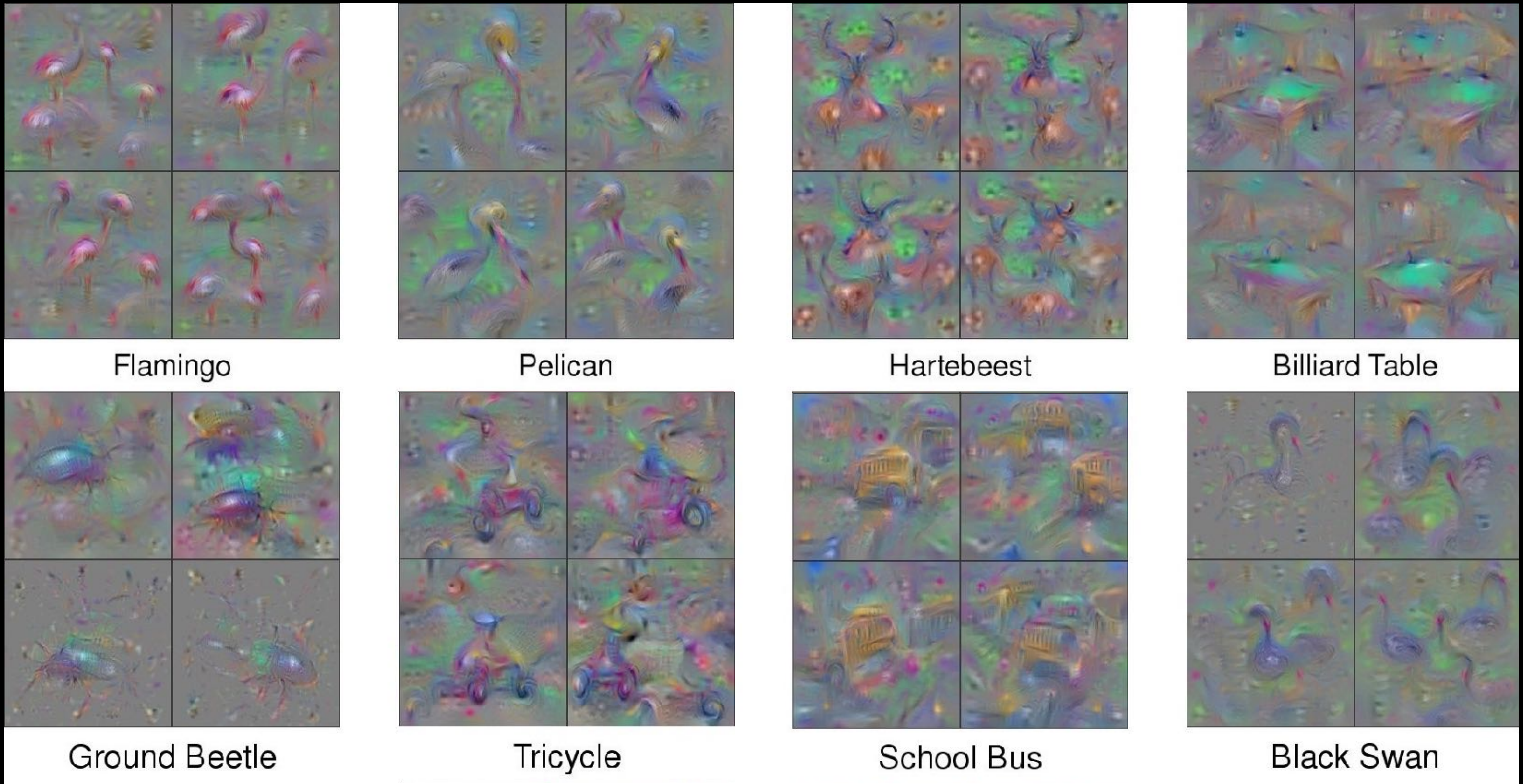
L2 loss from mean image



Simonyan, Vedaldi, & Zisserman 2013

Deep Visualization Take 2

Yosinski, Clune, Nguyen, Lipson, 2015, ICML Deep Learning Workshop



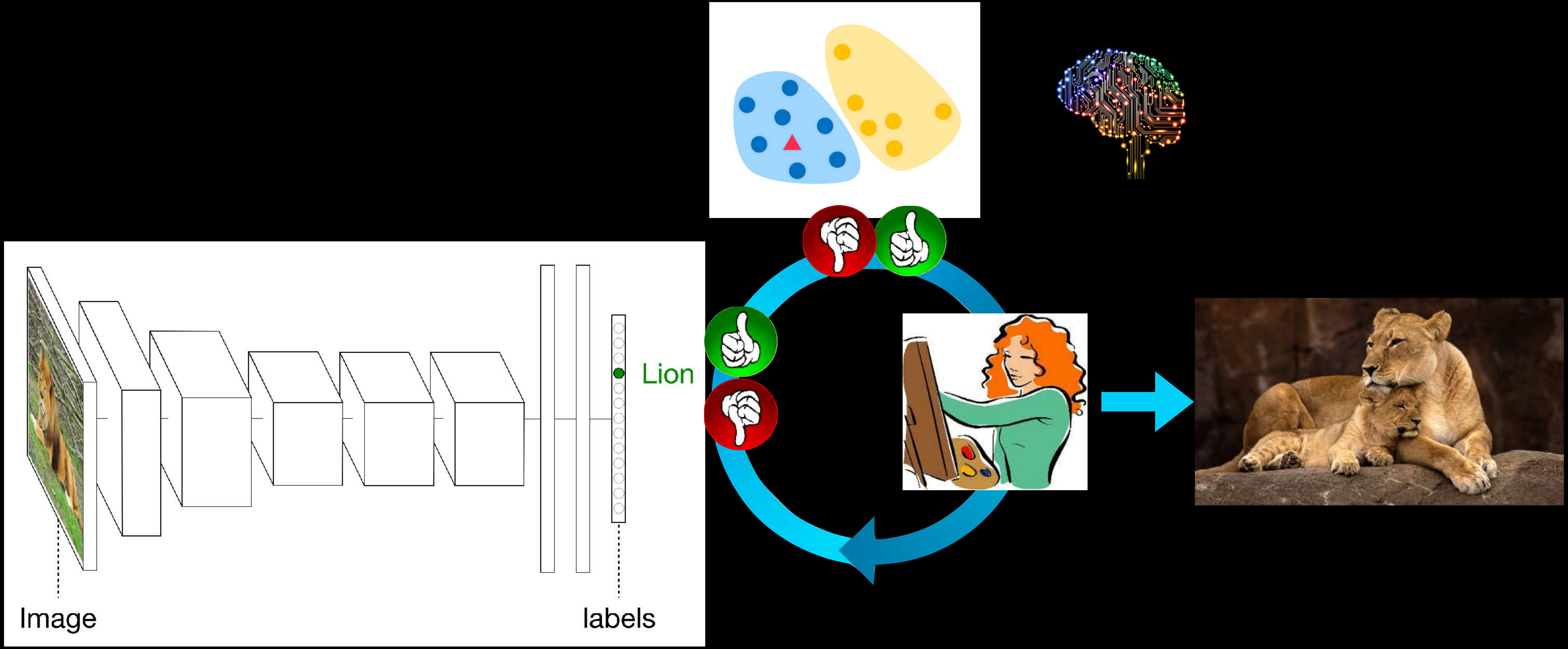
Deep Visualization Take 3

Multifaceted Feature Visualization. Nguyen, Yosinski, Clune 2016, ICML Workshop

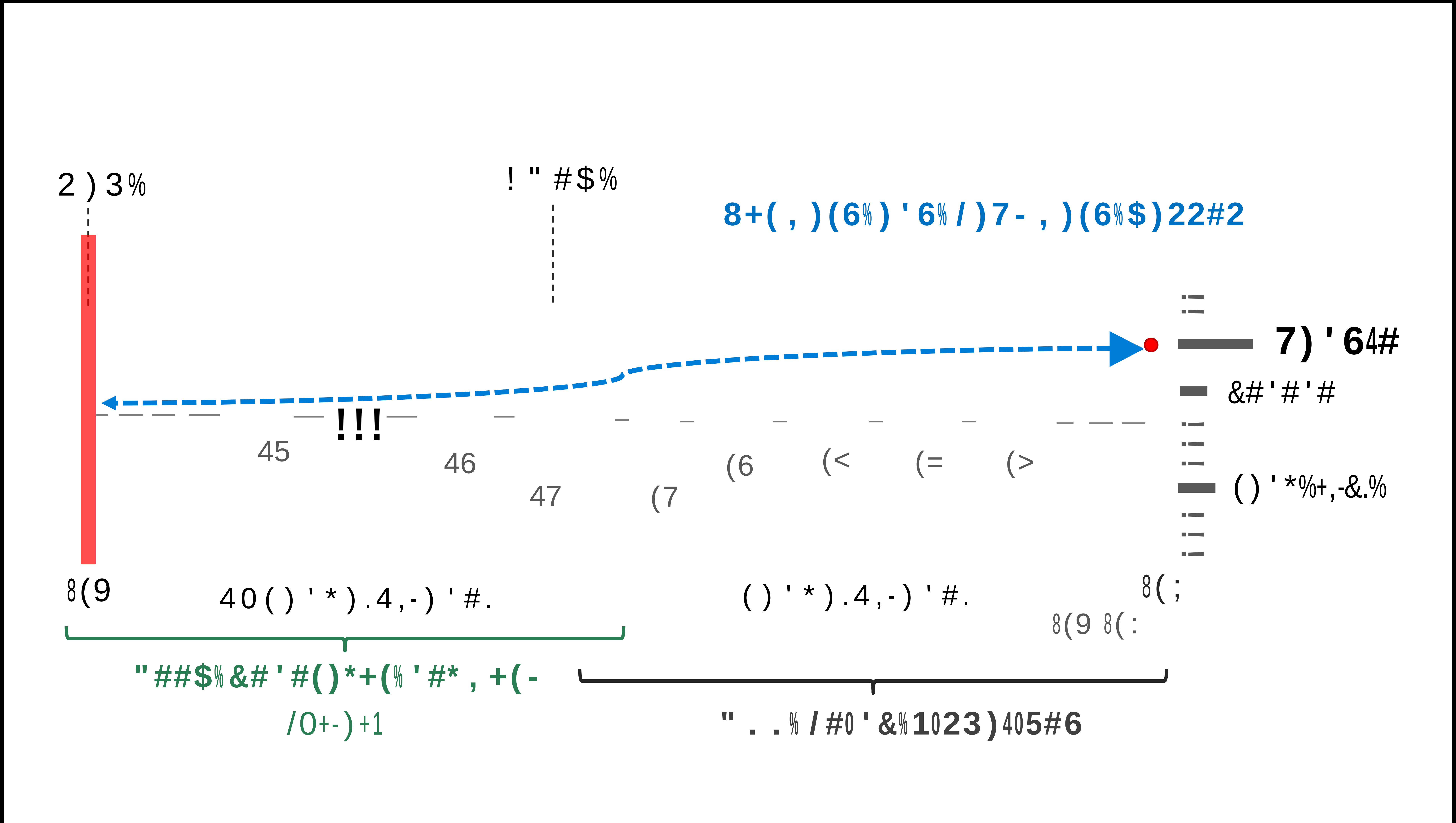
Deep Visualization Take 4

Nguyen, Dosovitskiy, Yosinski, Brox, Clune. NeurIPS. 2016

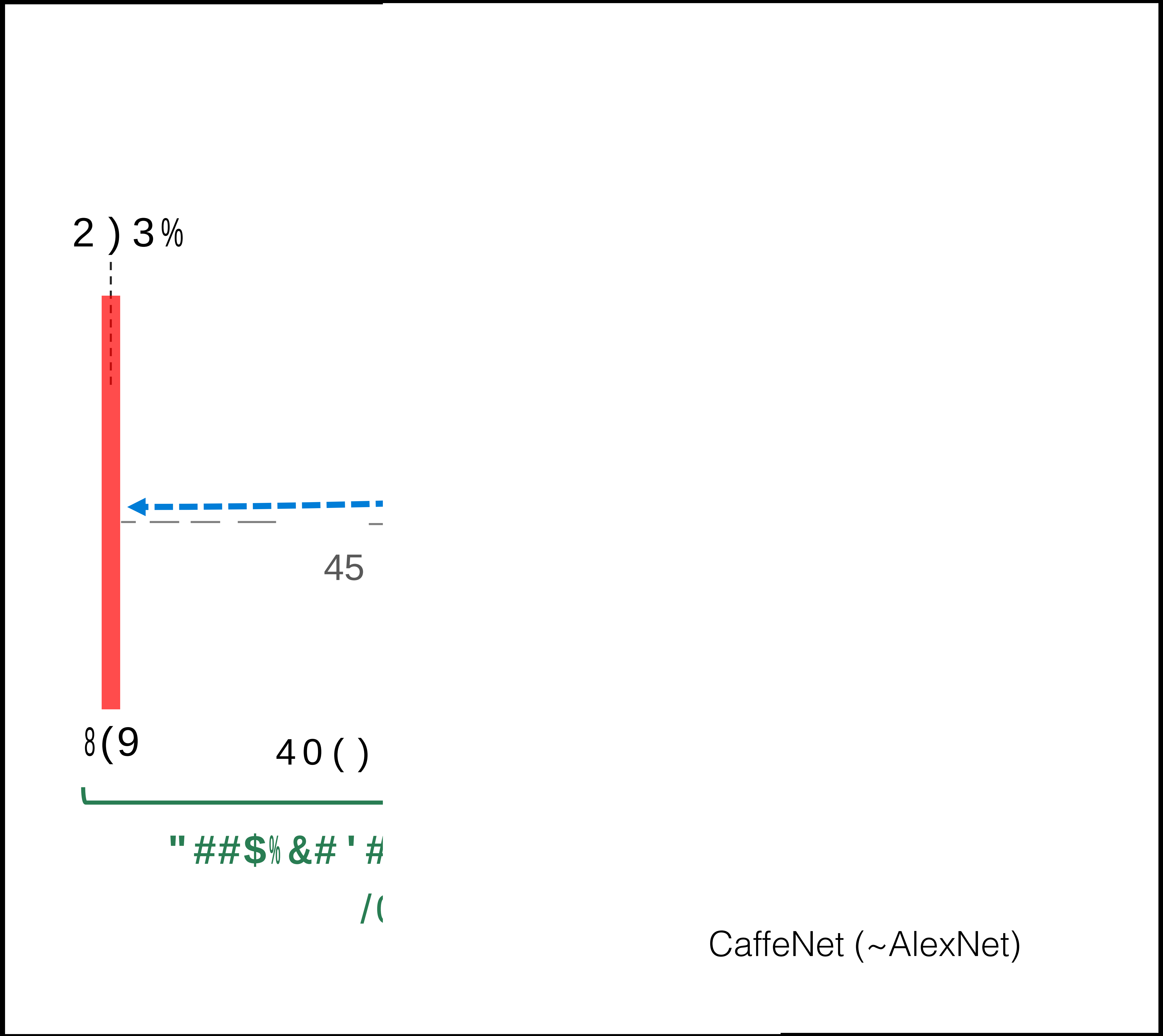
Learned Natural Image Priors



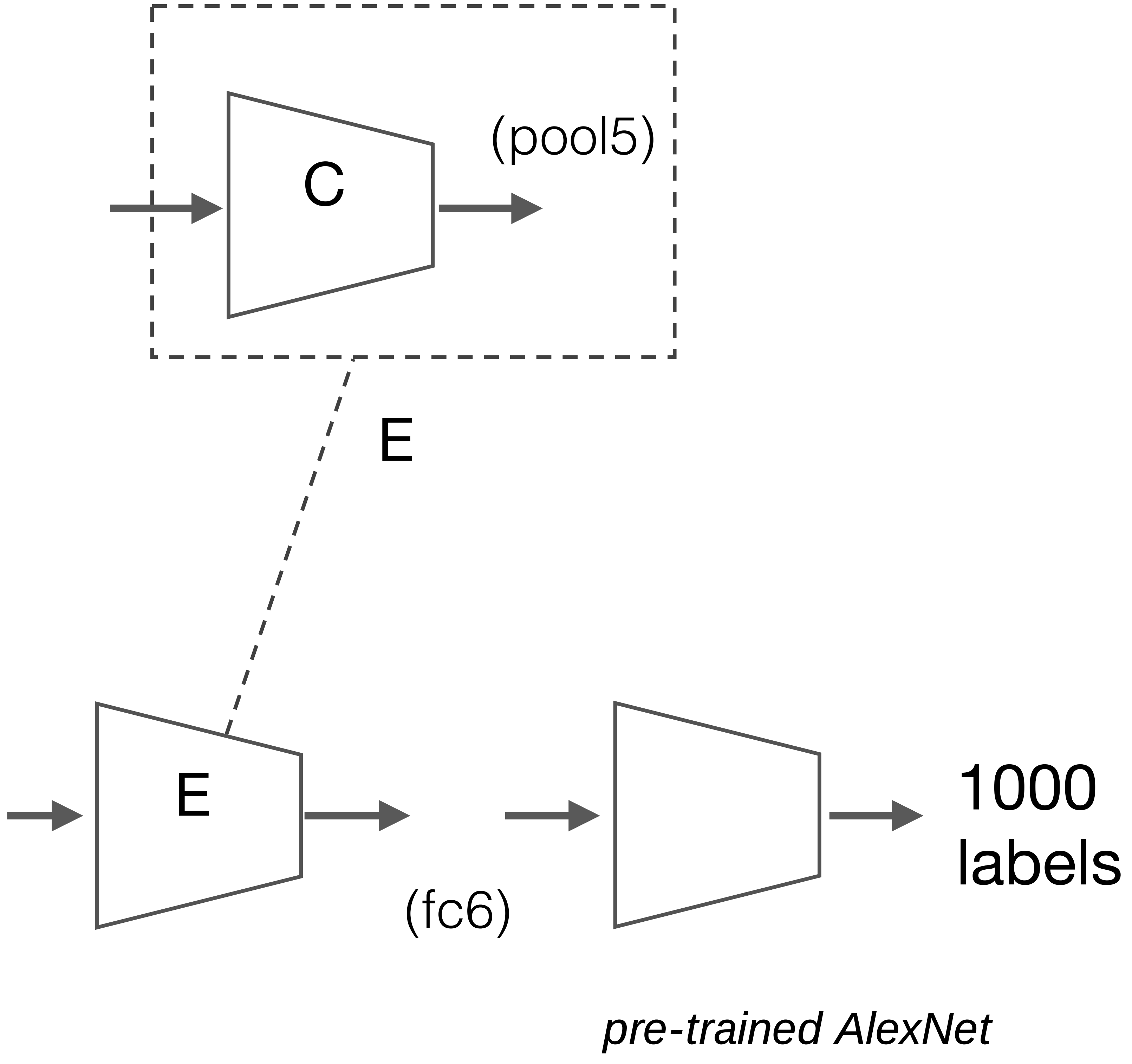
Deep Generator Network based Activation Maximization (DGN-AM)



Training the Deep Generator Network (DGN)

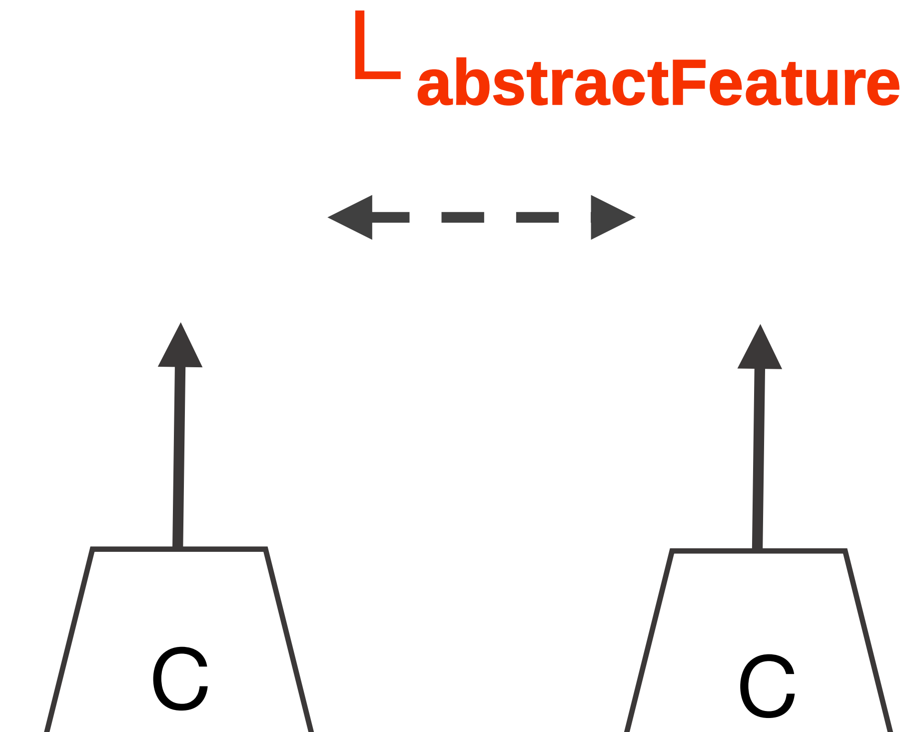


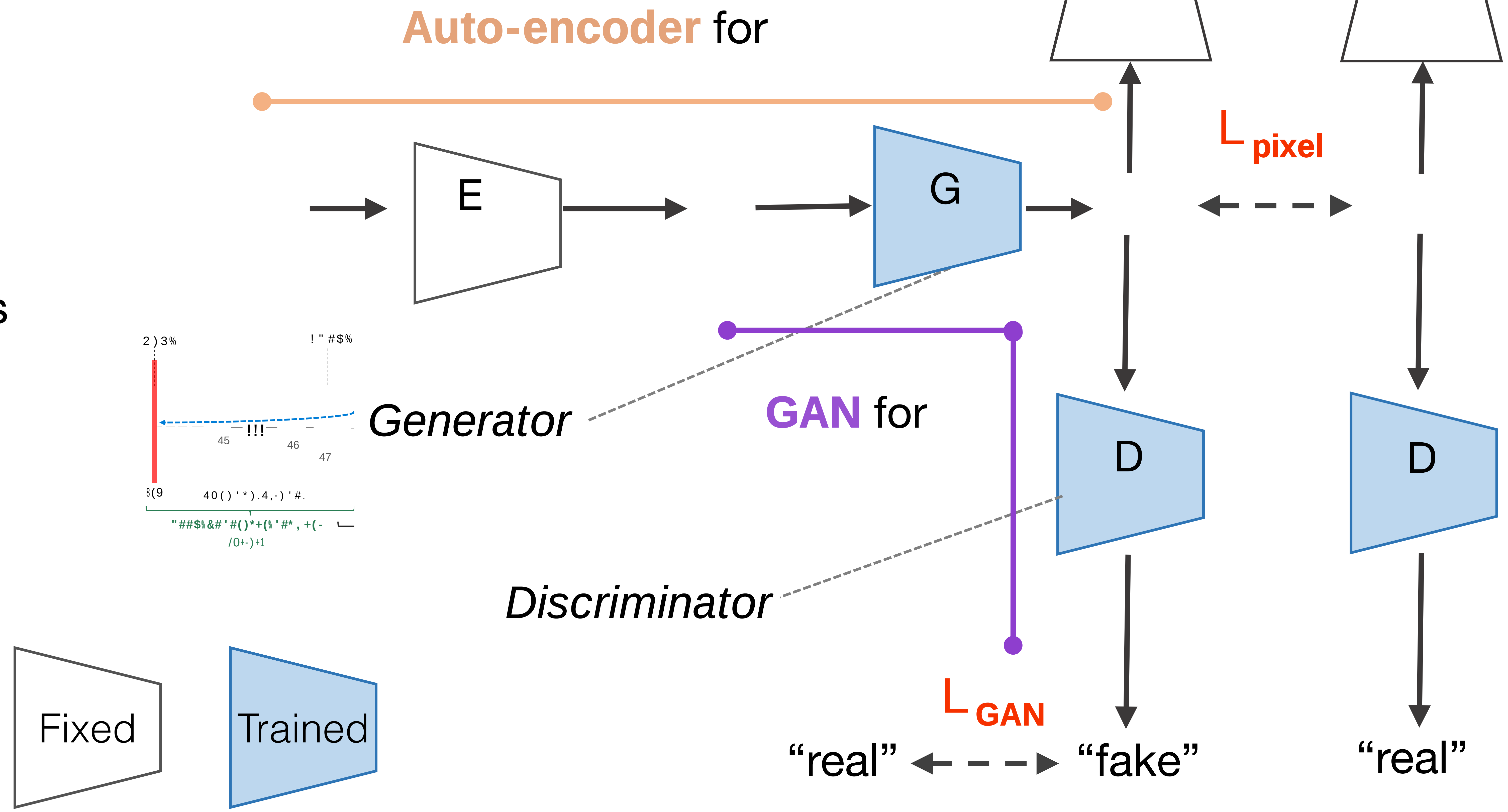
Training the Deep Generator Network (DGN)



Train w 3 Losses

- Pixelwise (L2)
- Semantics Dosovitskiy 2015
- GAN Goodfellow et al. 2014





Deep Visualization Take 4

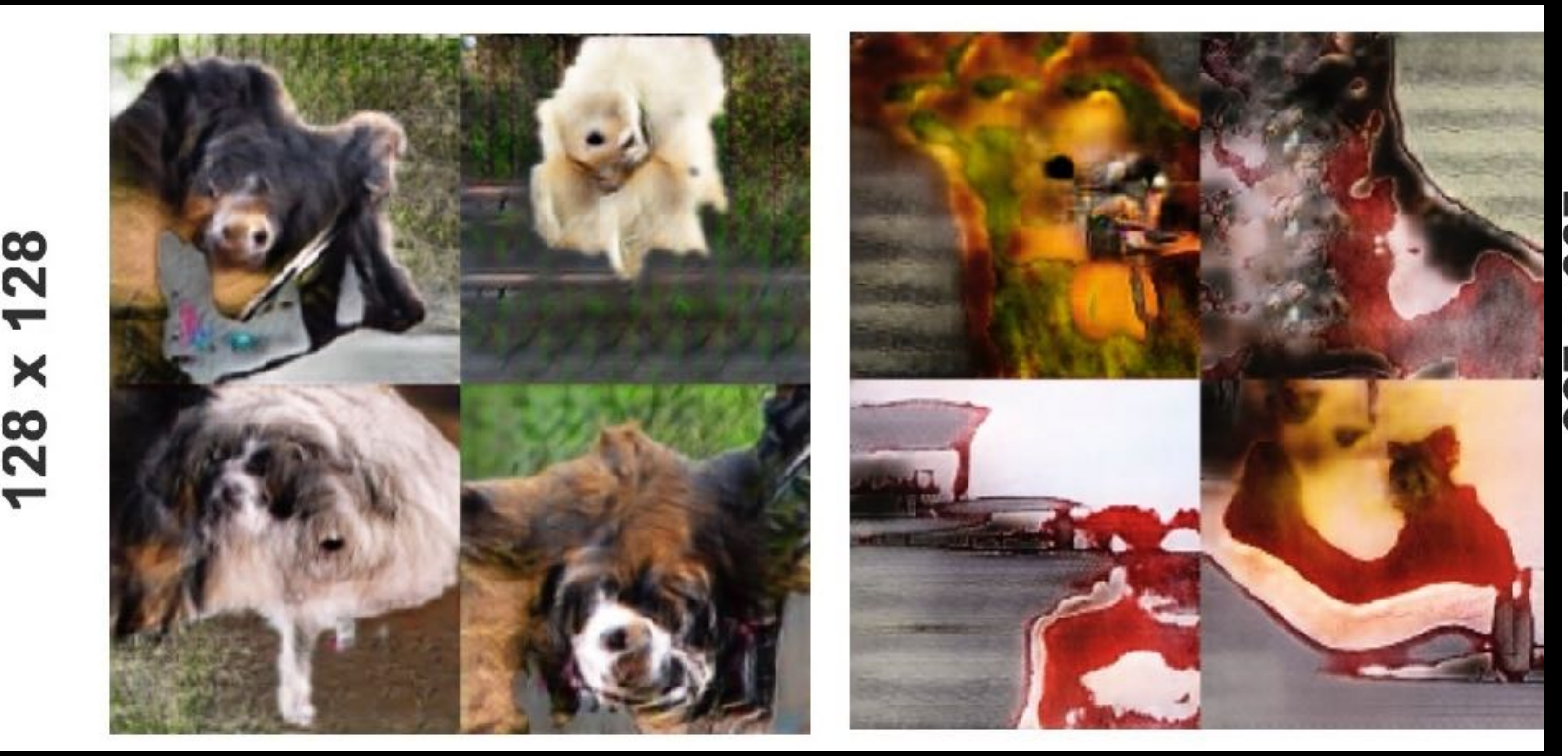
Nguyen, Dosovitskiy, Yosinski, Brox, Clune. 2016. NeurIPS

Deep Visualization Take 4

Nguyen, Dosovitskiy, Yosinski, Brox, Clune. 2016. NeurIPS


Real
Synthetic
Real
Synthetic

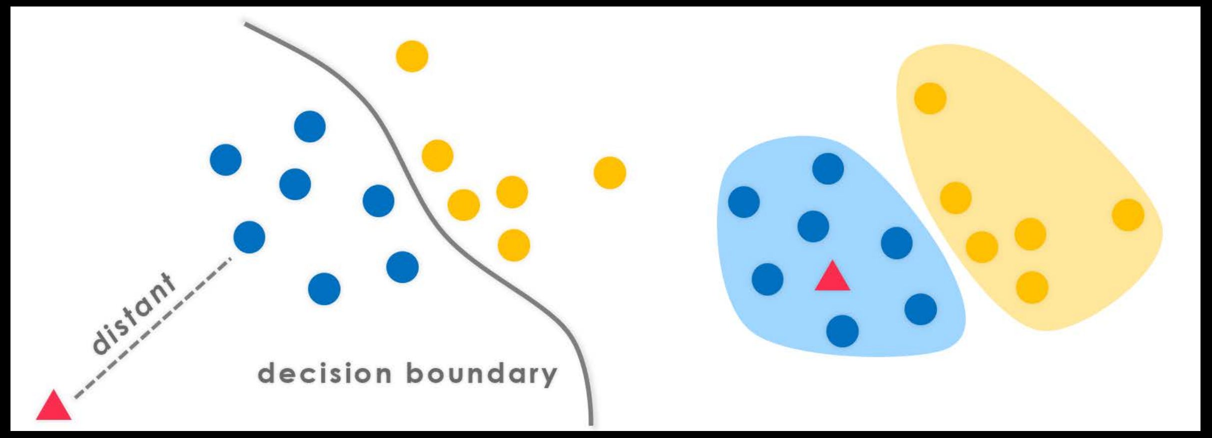
State of the Art Generative Model (at the time)



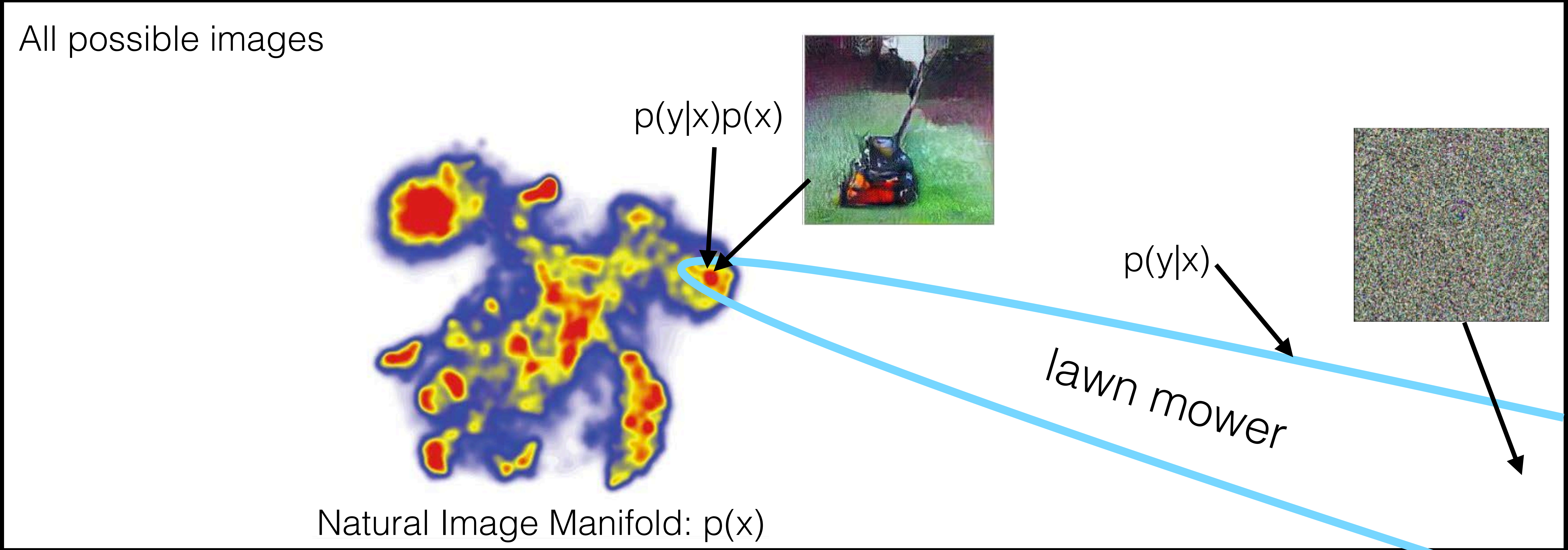
Improved GAN: Salimans et al. 2016

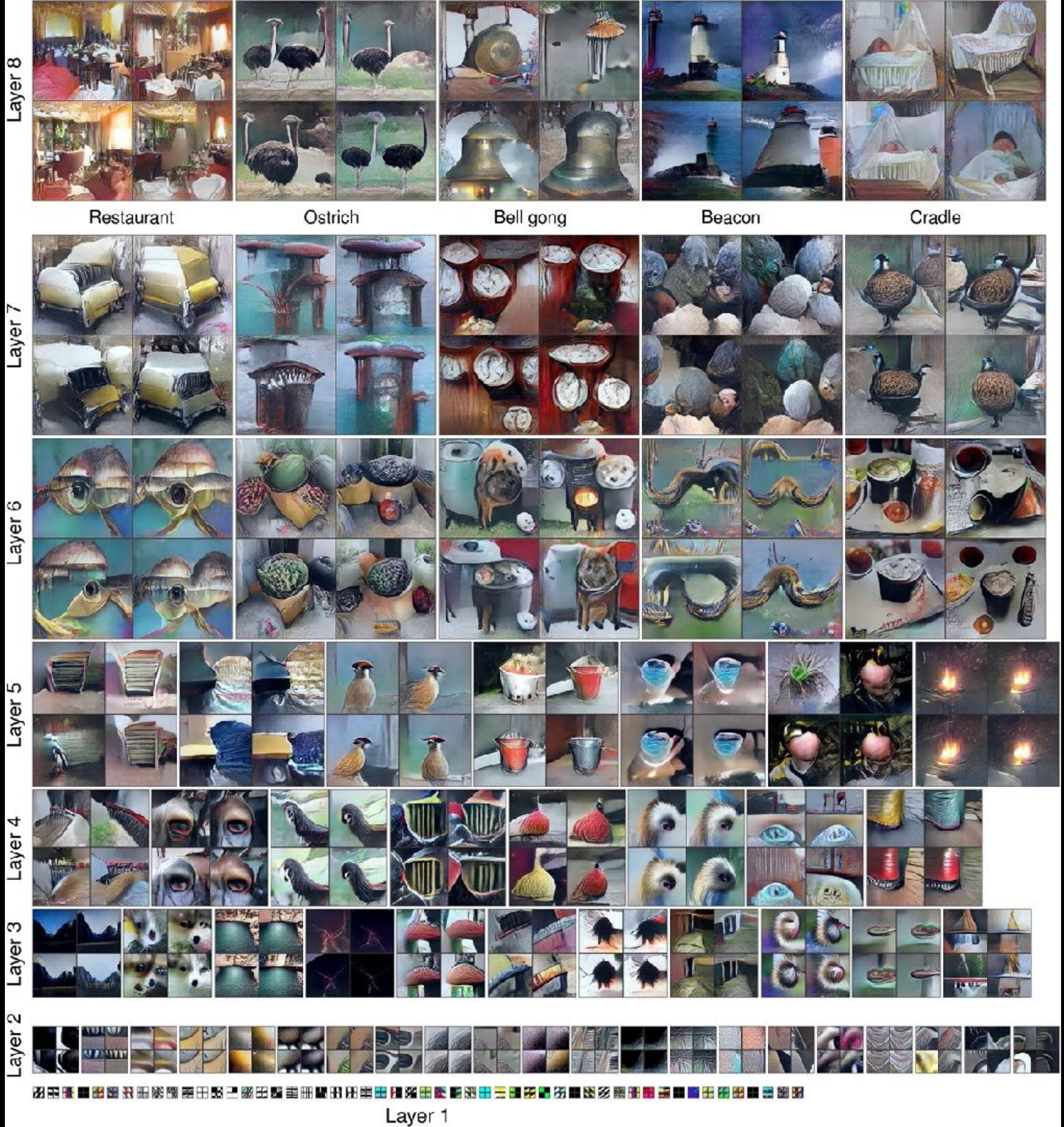
DGN-AM: Nguyen et al. 2016

Discussion



- Are they easily fooled, or do they understand?
  - Both!




Layer 8
Restaurant
Ostrich
Bell gong
Beacon
Cradle
Layer 7
Layer 6
Layer 5
Layer 4
Layer 3
Layer 2
Layer 1


Layer 3
Layer 2
Layer 1

Layer 5
Layer 4

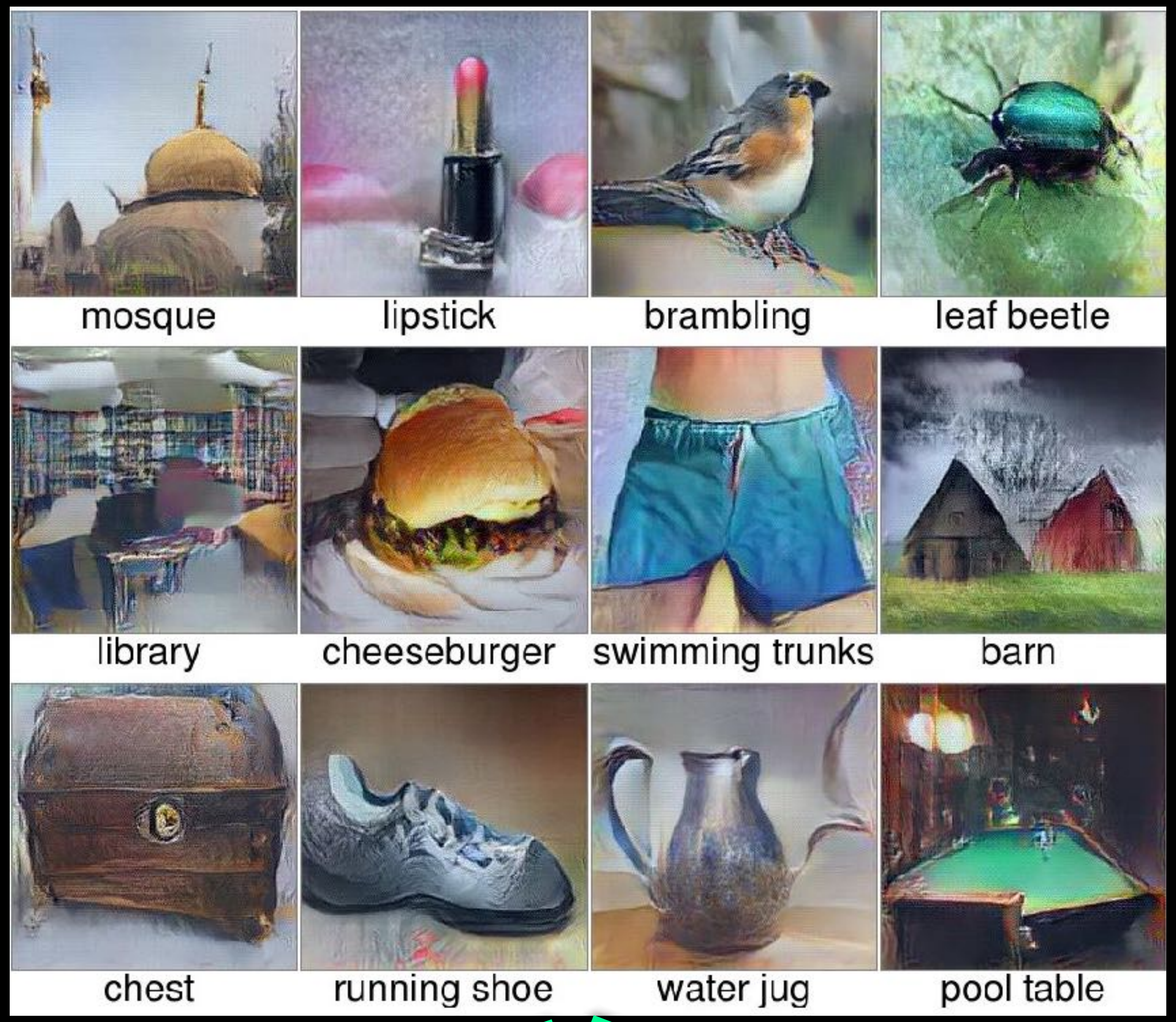
Layer 8

Restaurant

Ostrich

One drawback to DGN-AM

Deep Generator Network (DGN) + More Diversity



Better
generative model

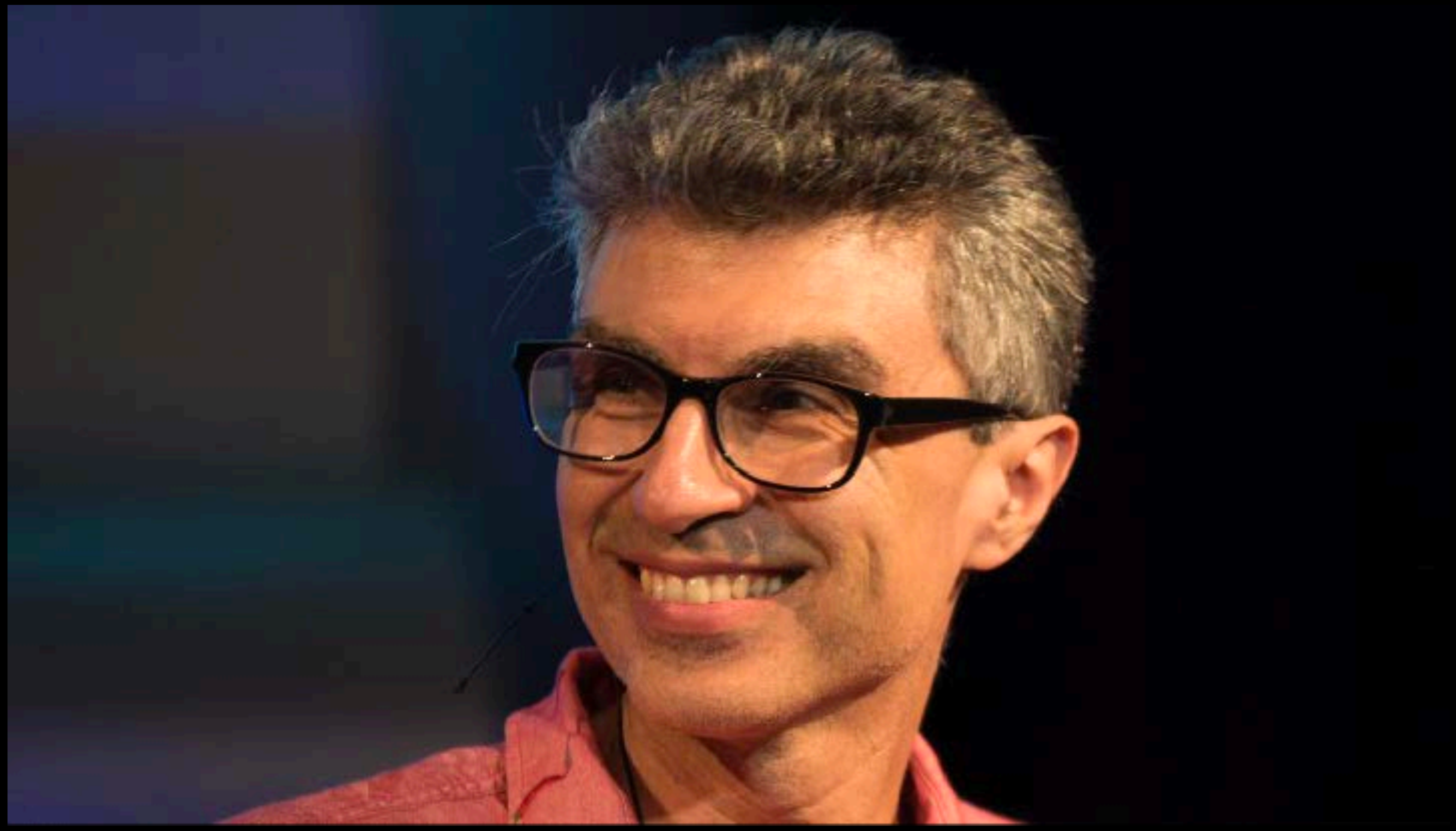
Improved multifaceted
feature visualization

Plug & Play Generative Networks (PPGNs)

Nguyen, Clune, Dosovitskiy, Bengio, Yosinski. CVPR 2017

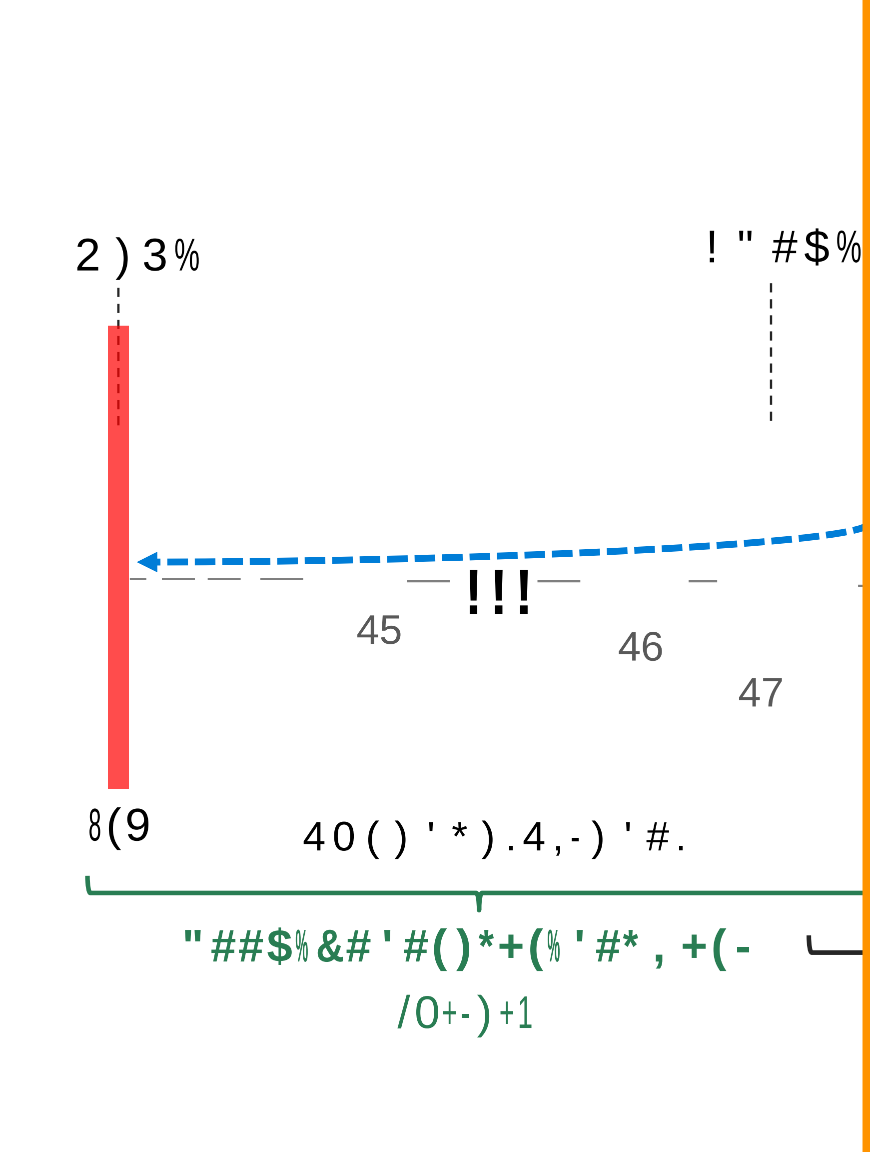
Take 5

+ Yoshua Bengio



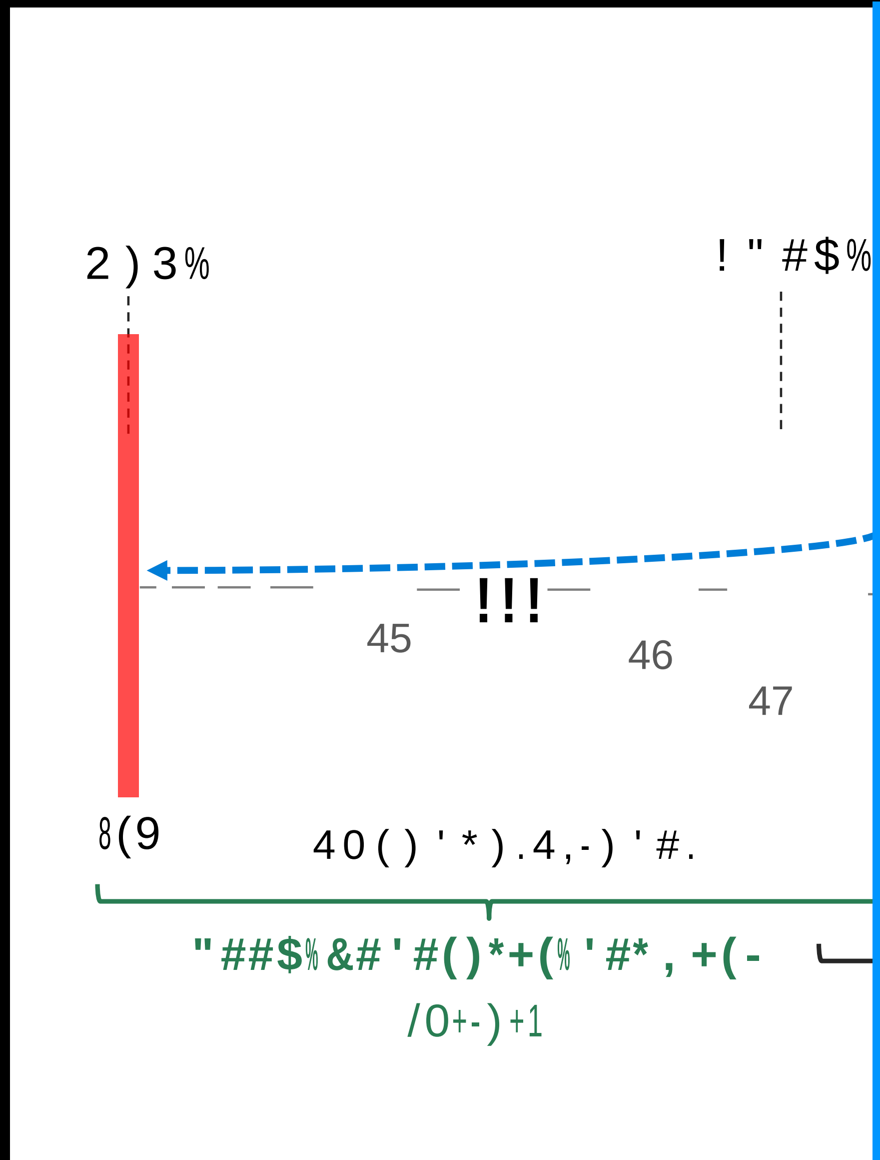
$p(x,y) = p(x)p(y|x)$

“Plug & Play Generative Networks”



ImageNet

=



MIT Places, captioner regressor etc.

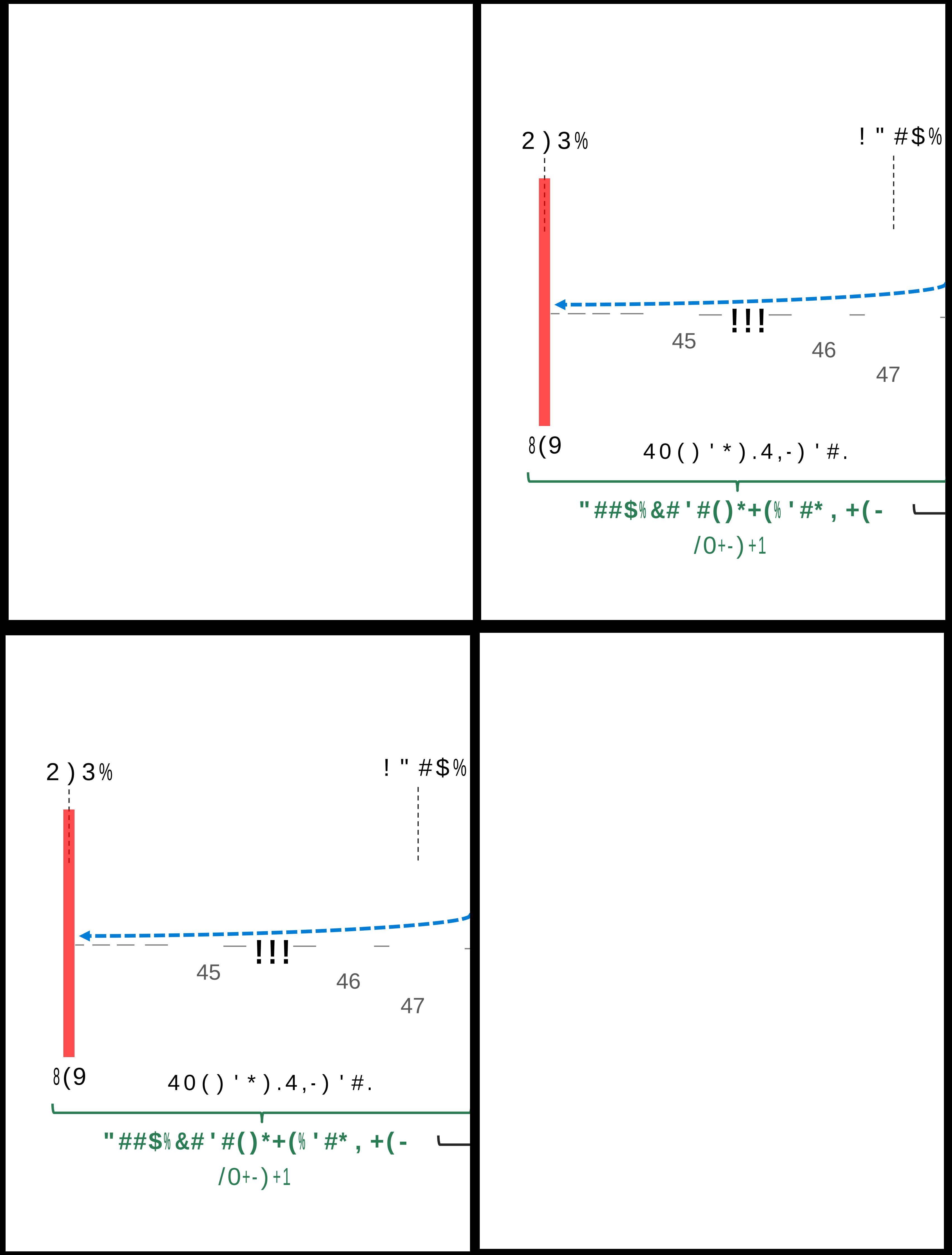
=

PPGNs: DGN-AM with Better Sampling

- Denoising auto-encoders model the data density: you can get the derivative of log p(x) easily

Alain & Bengio, 2014

- We create a code (h) auto-encoder
  - input current code h
  - get 'more real' output code h'
  - move input code in that direction



PPGNs: DGN-AM with Better Sampling

Denoising auto-encoders model the data density & provide the derivative of log p(x)

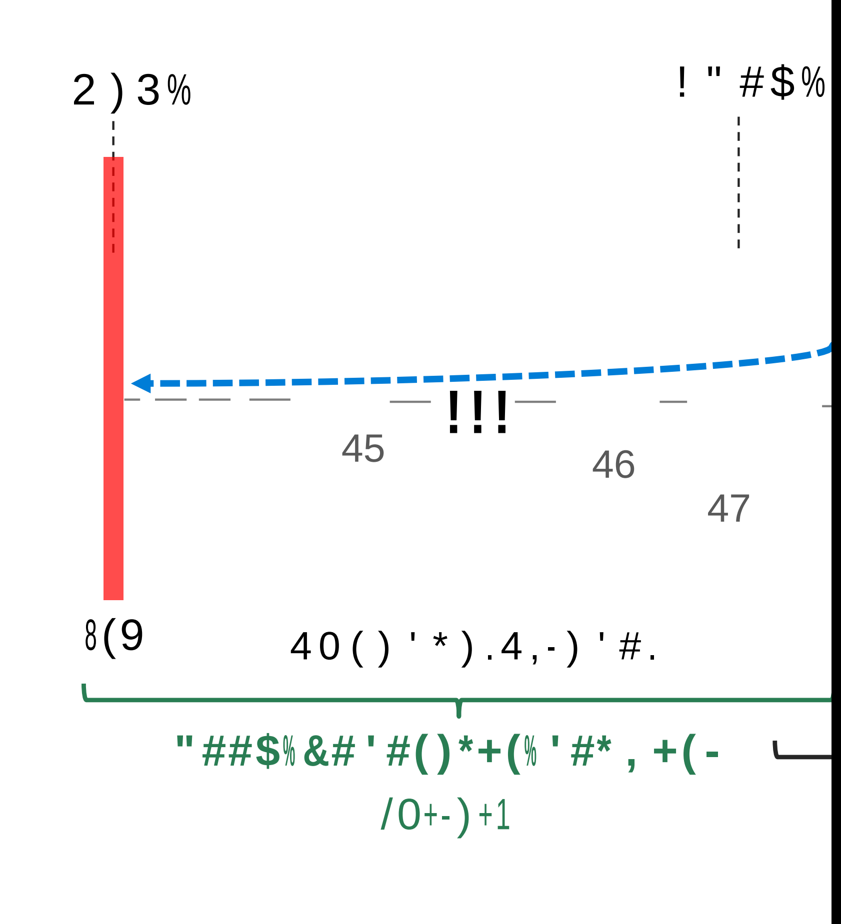
realism prior

new code

current code

Activate target neuron: DGN-AM v1

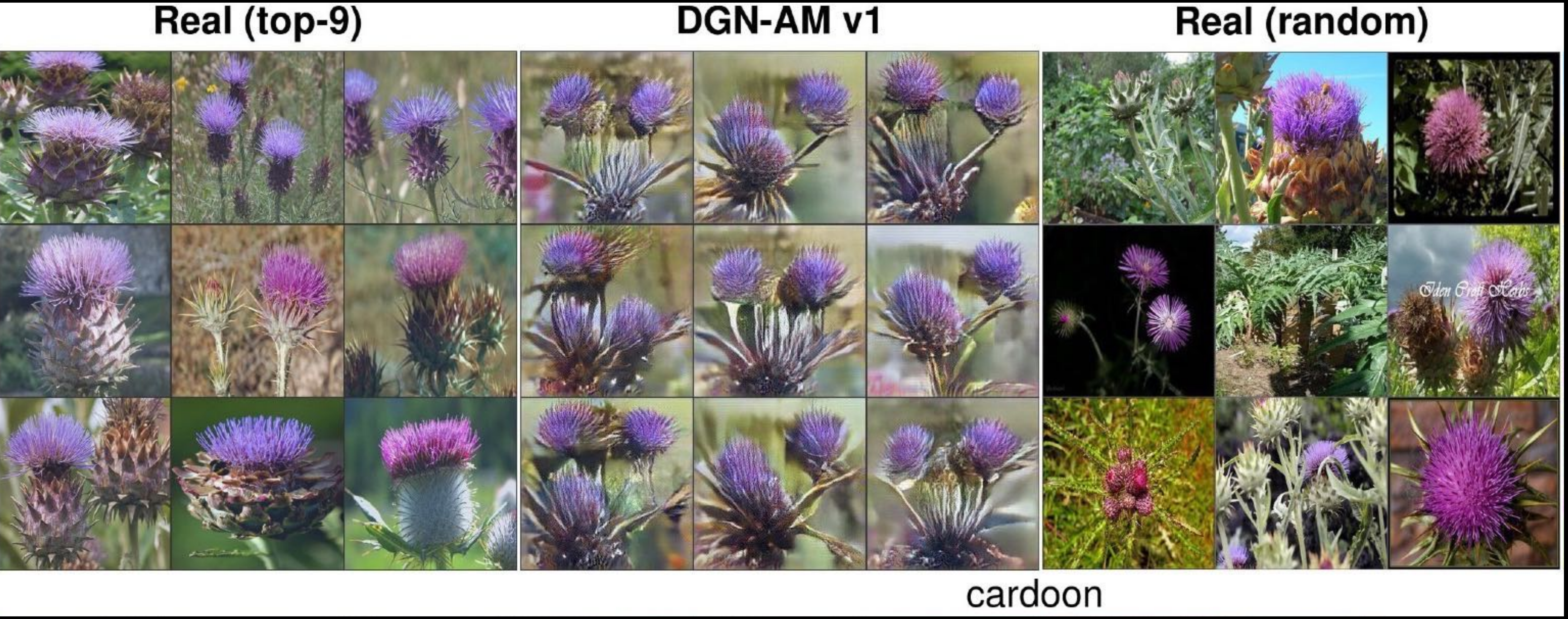
noise



softmax of neuron in target network

~Langevin sampler without the rejection step

PPGNs: Better MFV & Generative Model

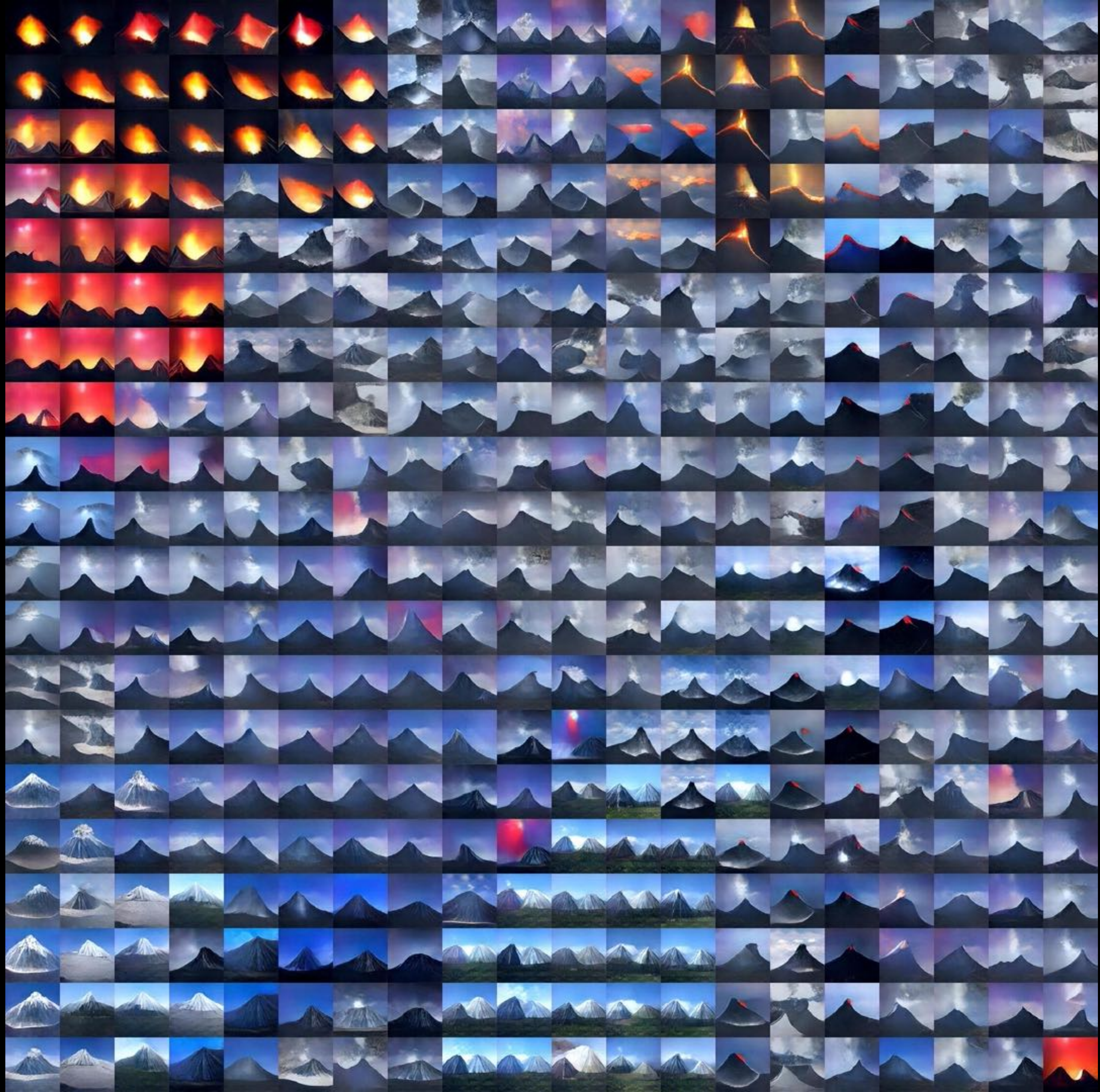



Real (top-9)
DGN-AM v1
Real (random)
PPGN
Iden Croft Herbs
cardoon

DGN-AM

Plug & Play Generative Networks

Improved diversity



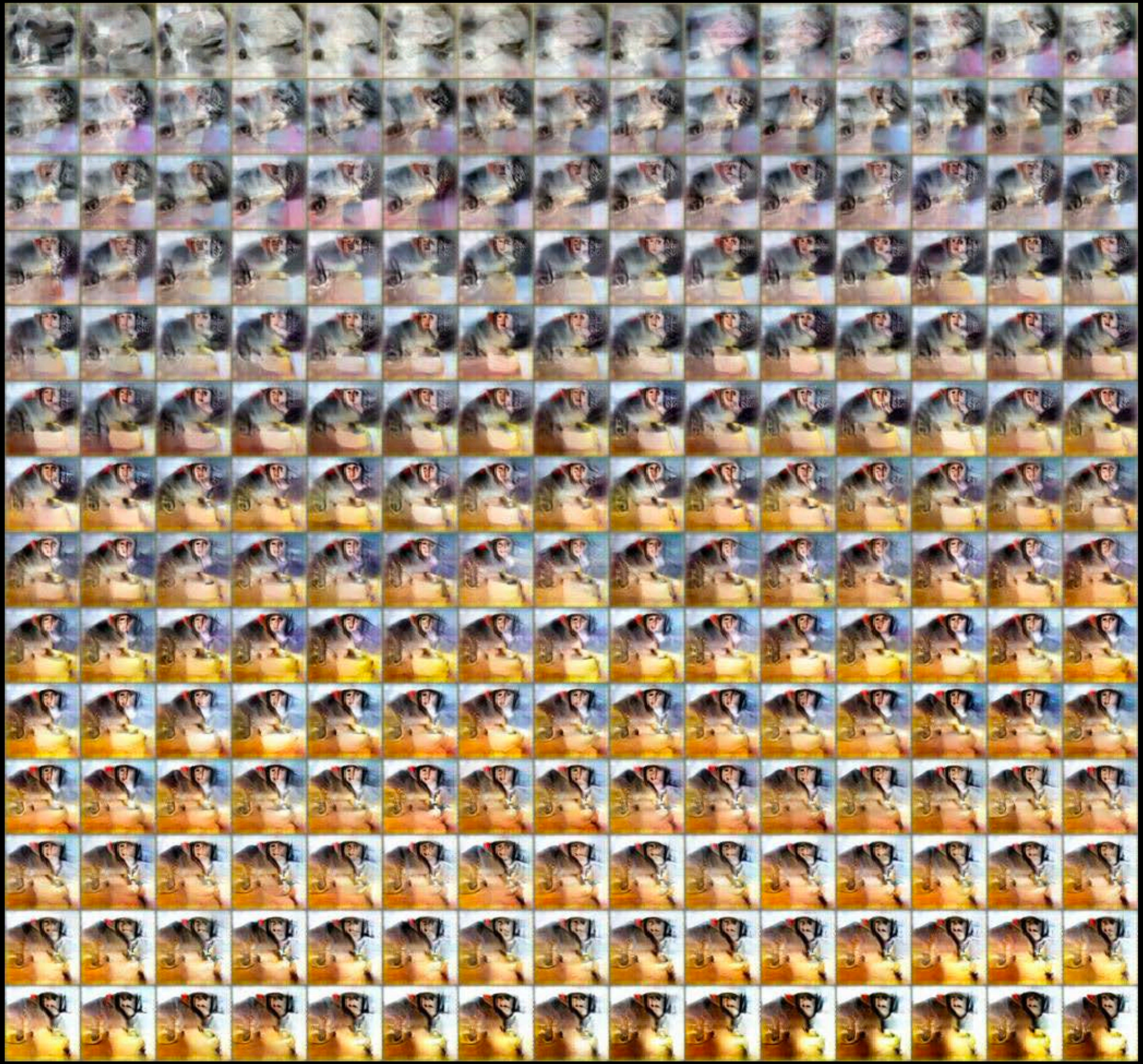
Conclusions: AI Neuroscience

- Despite our initial conclusions after the “fooling” work,
- DNNs do understand the objects they classify
  - their global structure, context, and multifaceted nature
- PPGNs: Generative model & multifaceted deep visualization tool

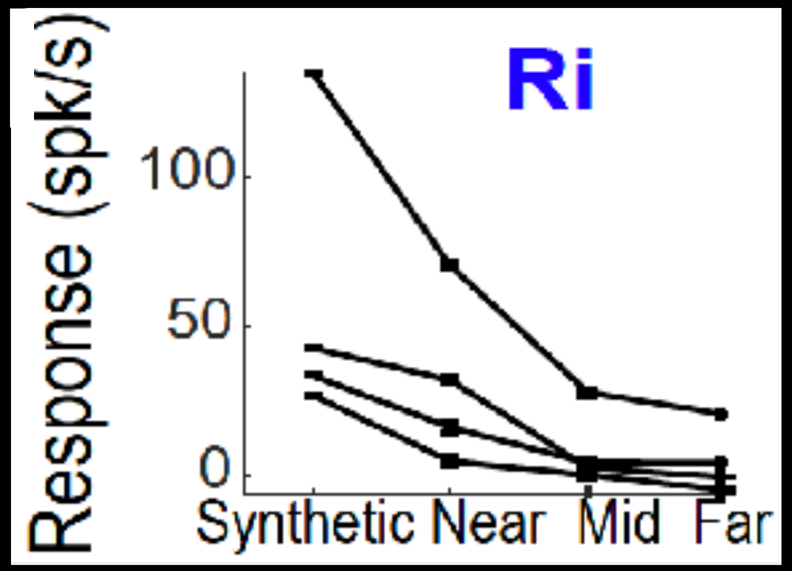
Future Work Ideas

- Generate videos, entire virtual worlds
- Other modalities
  - e.g. speech recognition, music classification
- Interpret deep RL networks
- Try with animal brains?

DGNs on real monkeys!

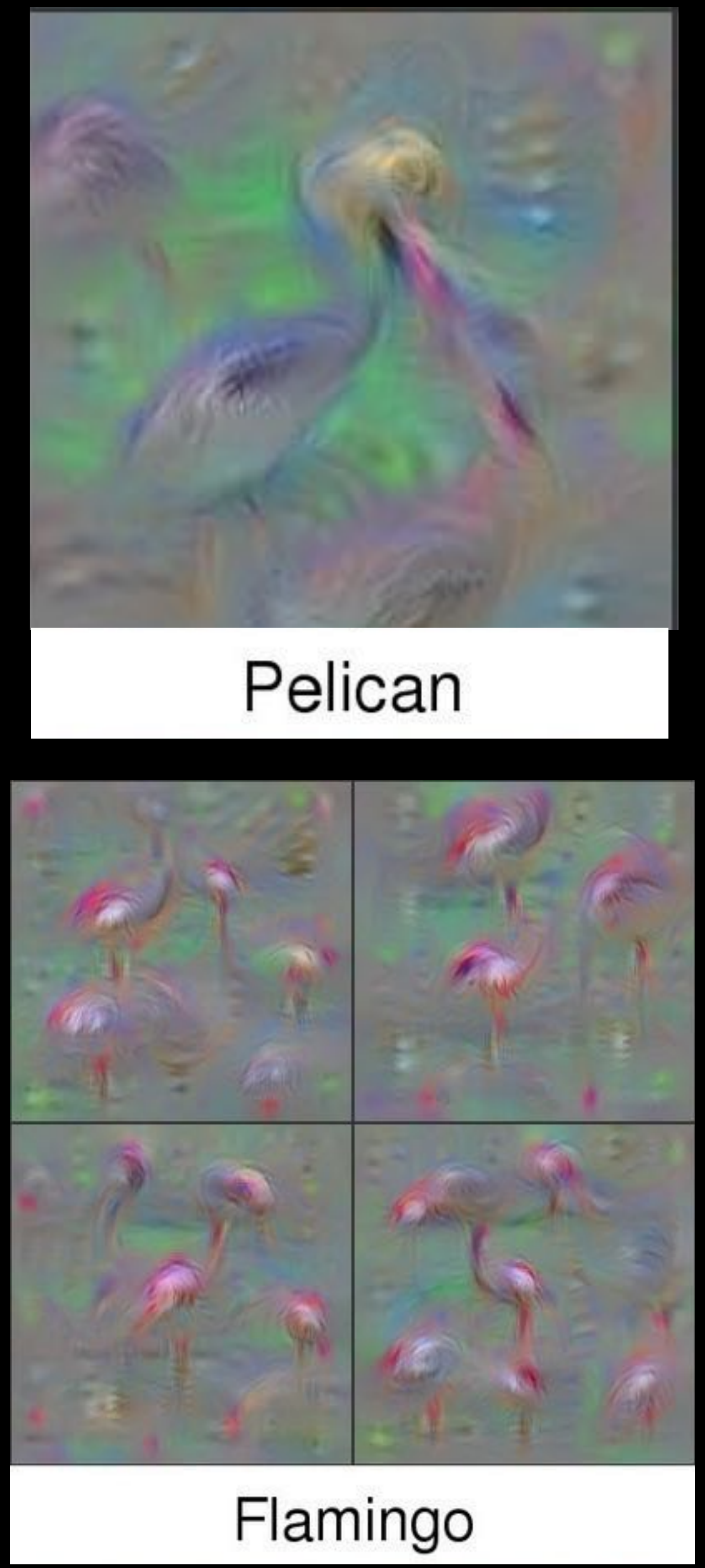


- finds both fooling and recognizable images
- predicts a neuron's function!



Ponce et al. Evolving super stimuli for real neurons using deep generative networks. Cell. 2019

Rapid Progress



2015

2015

2016

2017

2022 (today!)

DALL·E

Edit the detailed description

HELP

a green dragon breathing blue flame flying above a blood red ocean

DALL·E

Edit the detailed description

HELP

logo for a research lab on artificial intelligence

a tiger playing a violin

OpenAI @OpenAI · 15m

"A photo of an astronaut riding a horse" #dalle

2 7 69

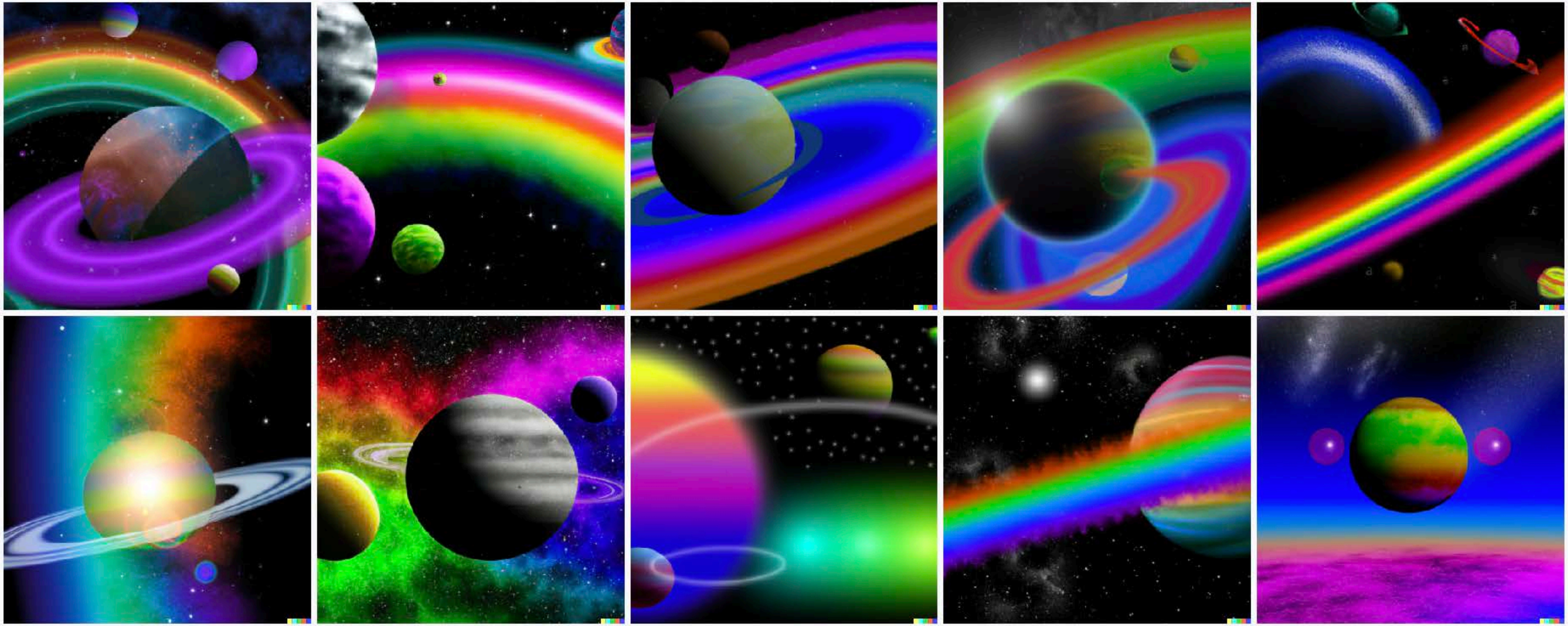
OpenAI @OpenAI · 13m

"A photo of a quaint flower shop storefront with a pastel green and clean white facade and open door and big window" #dalle

1 4 34

a rainbow in outer space between planets

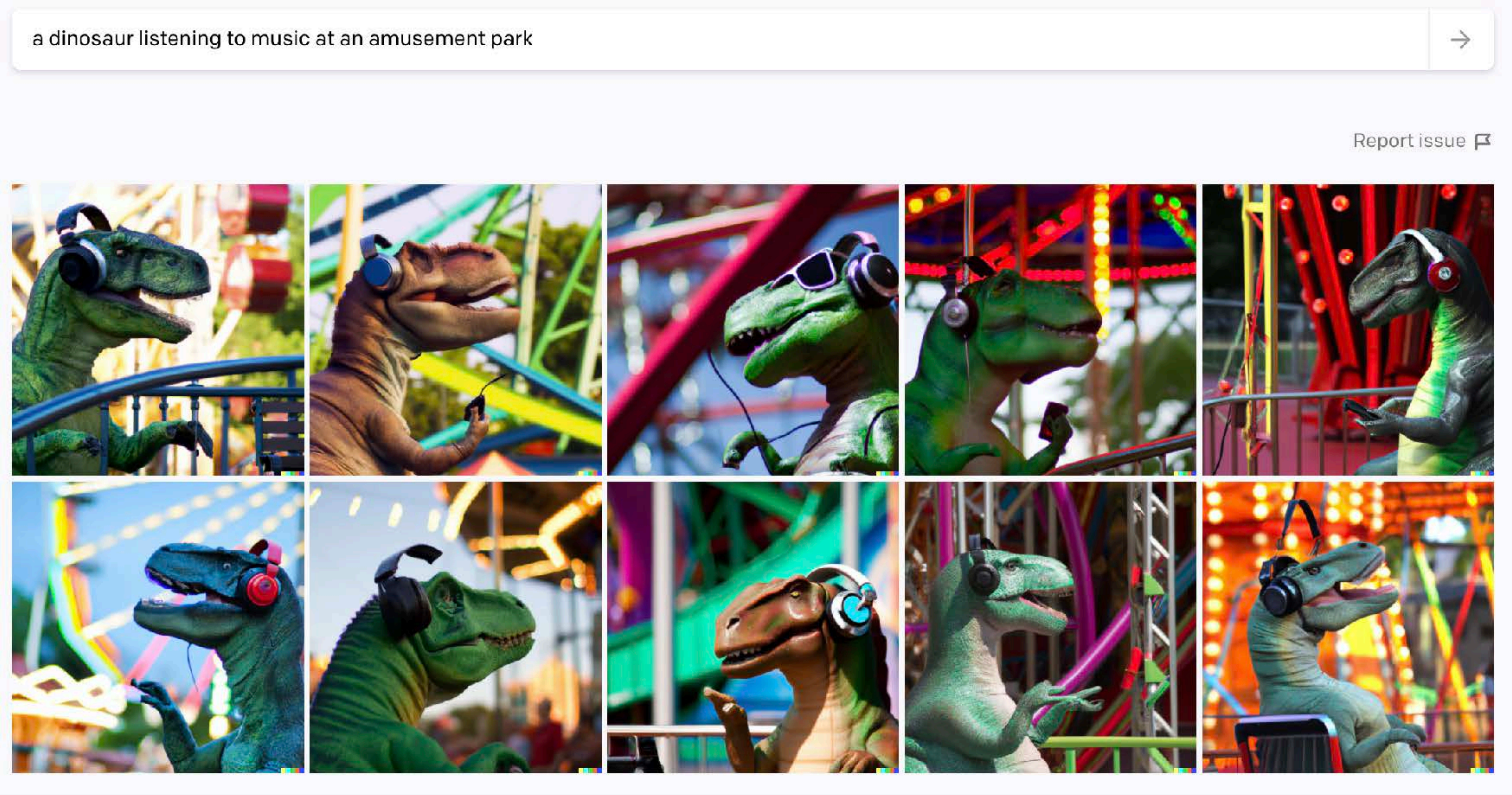
Report issue



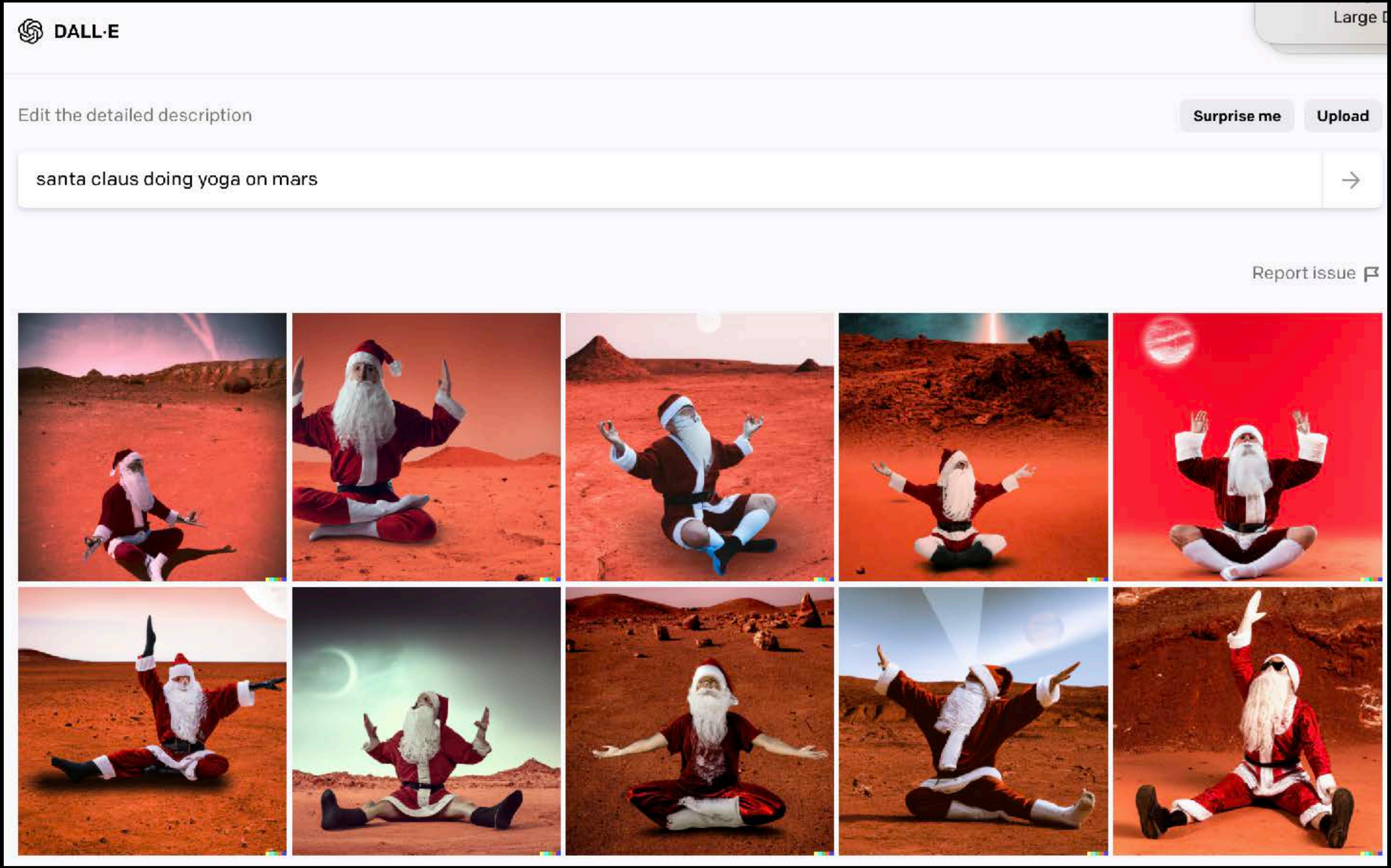
a carton three-layered cake with a unicorn walking on a rainbow on the top

Report issue

From the 12pm class



From the 2pm Class



DALL-E 2 results for “Teddy bears mixing sparkling chemicals as mad scientists, steampunk.” | OpenAI

Many more, done live: https://twitter.com/sama/status/1511724264629678084

Erase part of the image, then describe your desired new image

a bunch of red grapes blocking a man's face


ORIGINAL

We can try it!

https://labs.openai.com/e/mLqds5DwxGVud1QXYjg6LfMs a cartoon of a lecture hall celebrating the last day of class phd comic

Report issue

Thanks

Anything else you want to know?