How much do deep neural networks understand about the images they recognize?

Jeff Clune

Assistant Professor, Computer Science, University of Wyoming
Director, Evolving AI Lab

EVOLVING ARTIFICIAL INTELLIGENCE LABORATORY

UNIVERSITY OF WYOMING

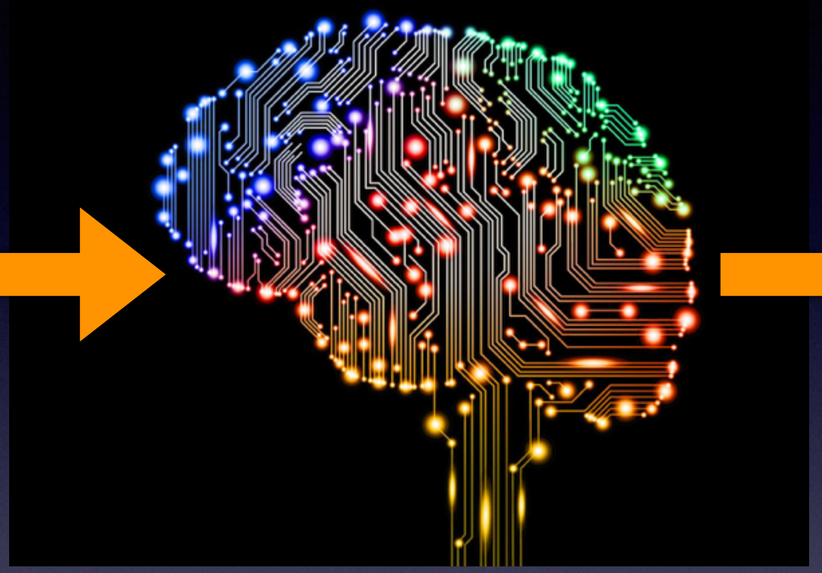
EVOLVING ARTIFICIAL INTELLIGENCE LABORATORY

UNIVERSITY OF WYOMING

Robotics & AI

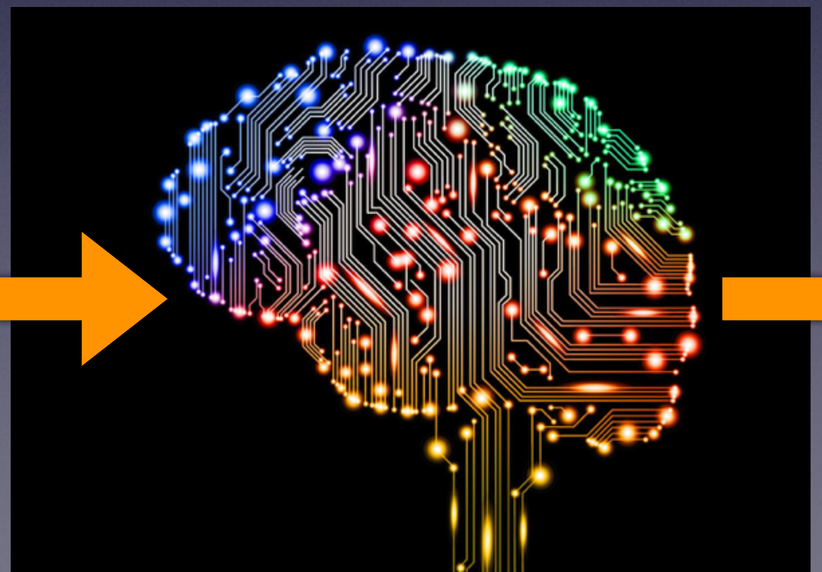
- **Robots that recover from damage**
- **Deep Learning**
- **Evolving neural network robot controllers**
- **Creative robots**
- **Computational biology**
- **etc.**

Deep Neural Networks/Deep Learning



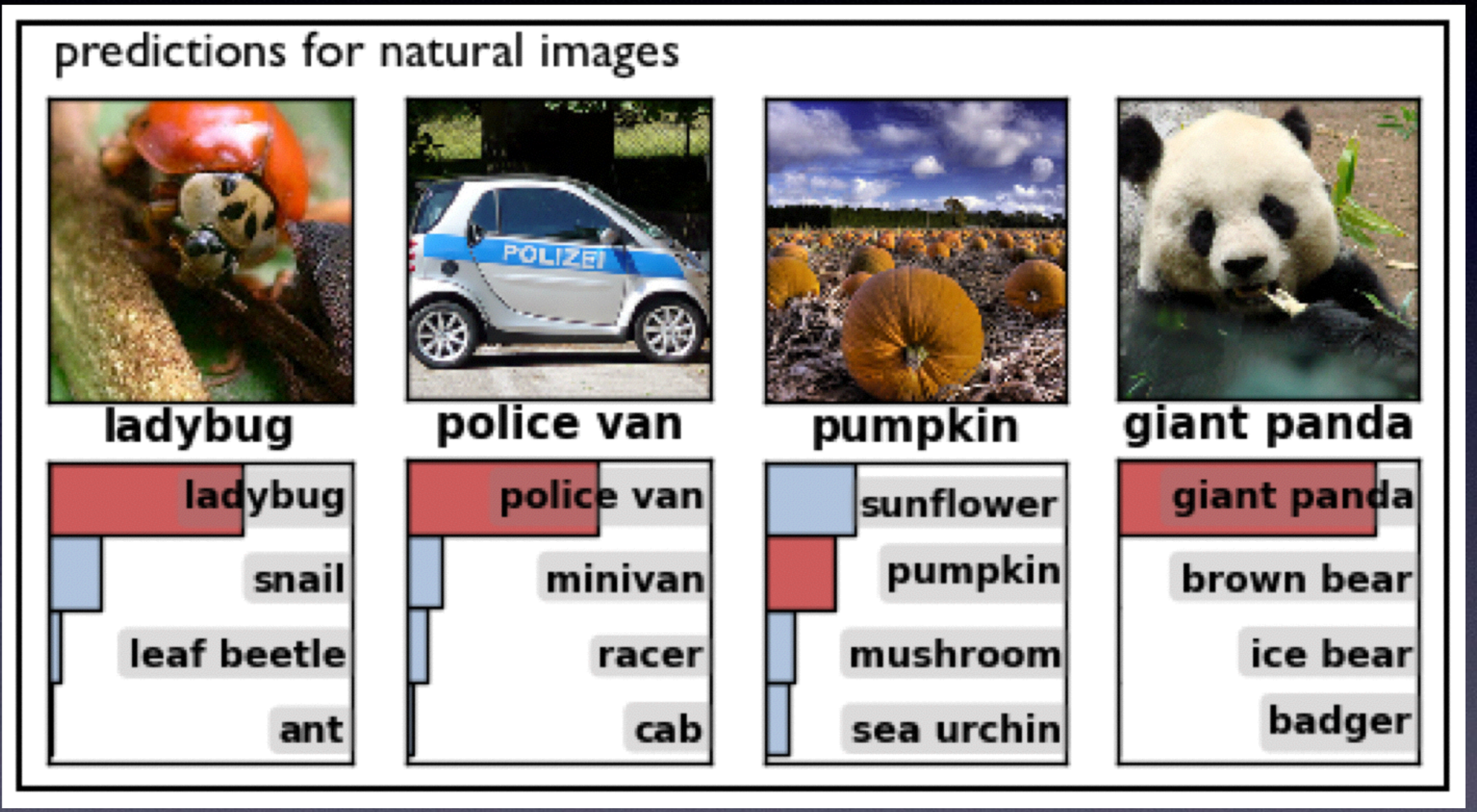
“Lion”

“Cogito, ergo sum”



“I think, therefore I am”

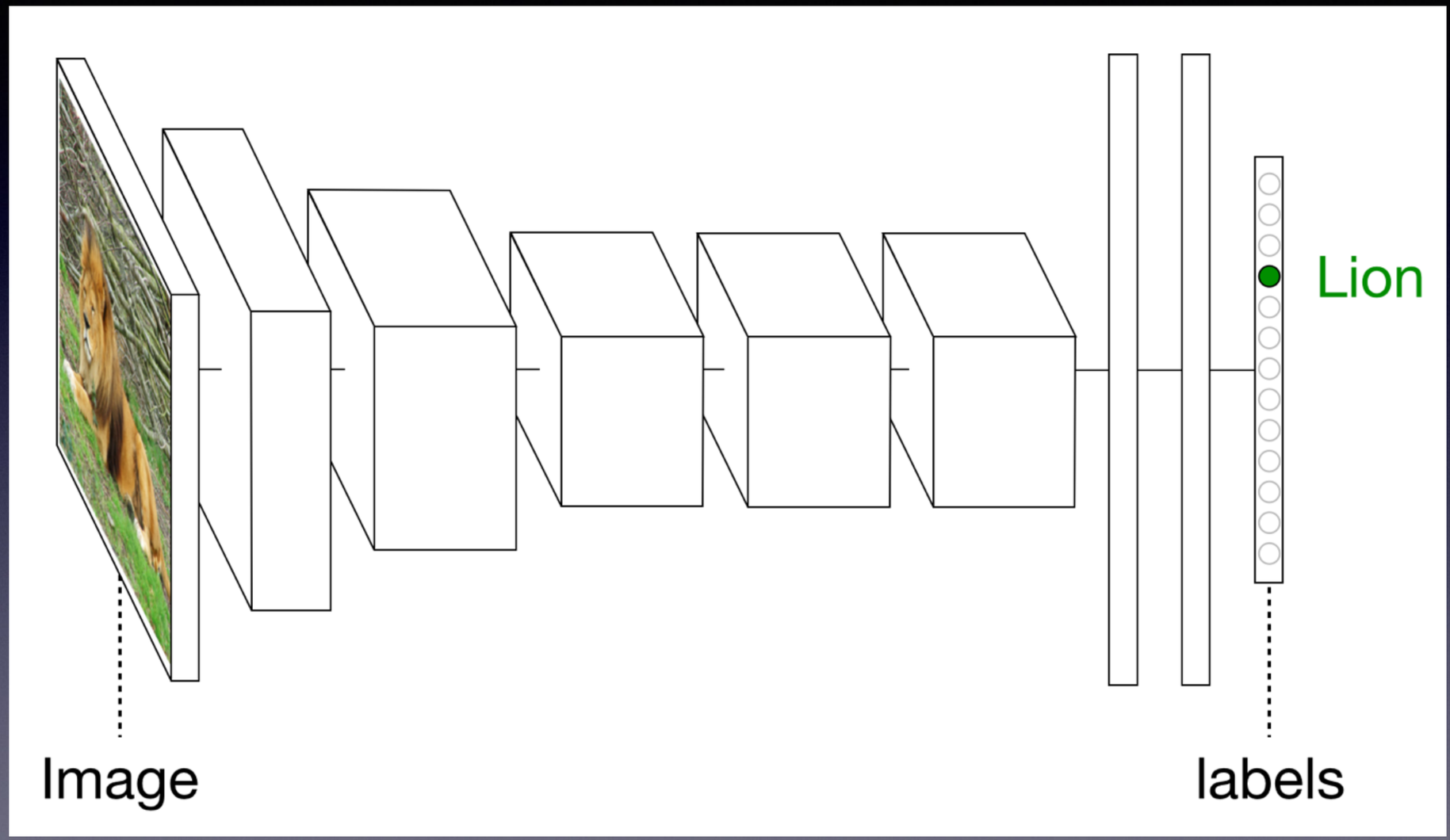
DNNs as good as humans at image recognition



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

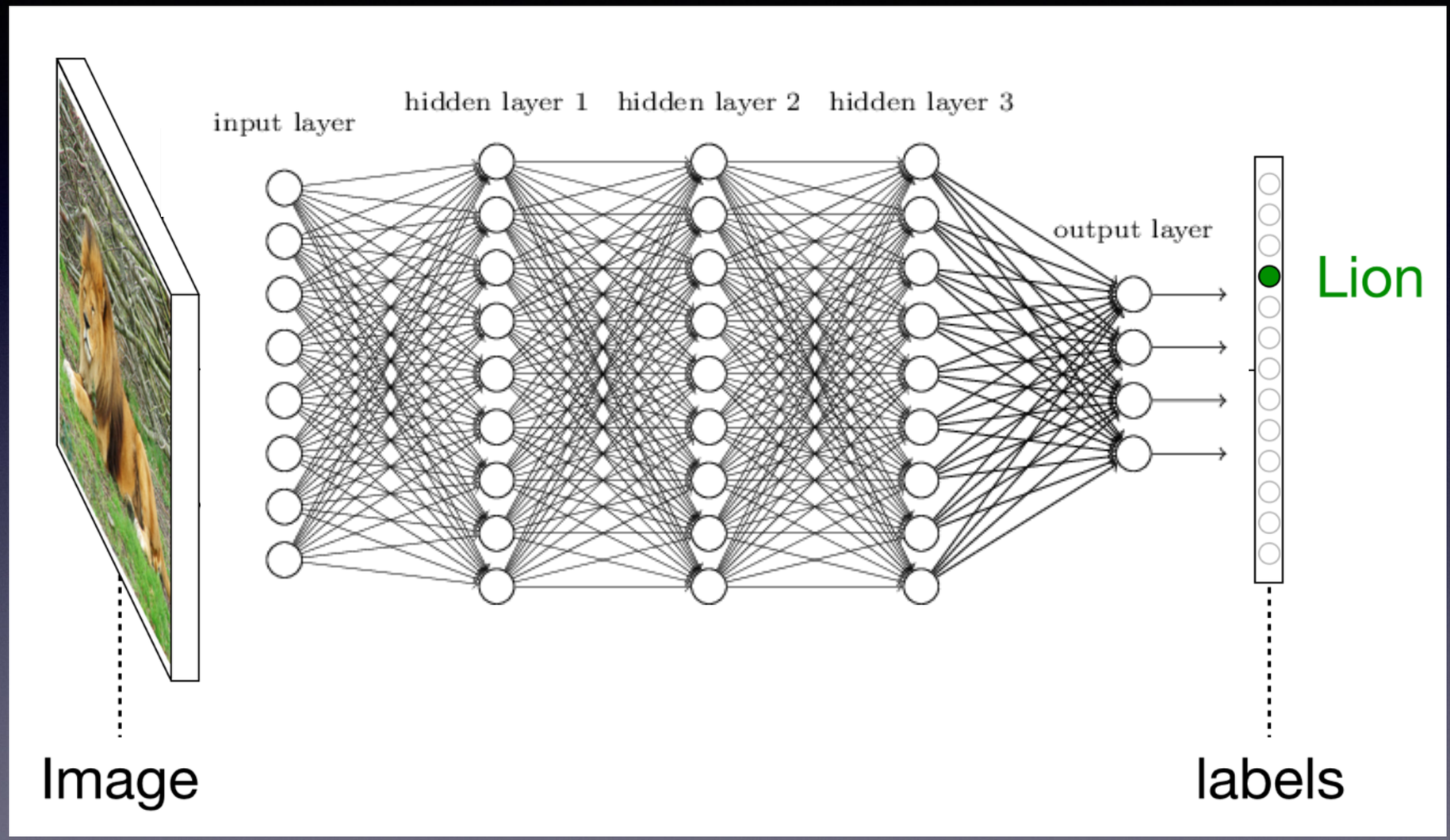
Alex Krizhevsky et al. Nips 2012

Deep Neural Networks/Deep Learning



~1M neurons
~100M weights

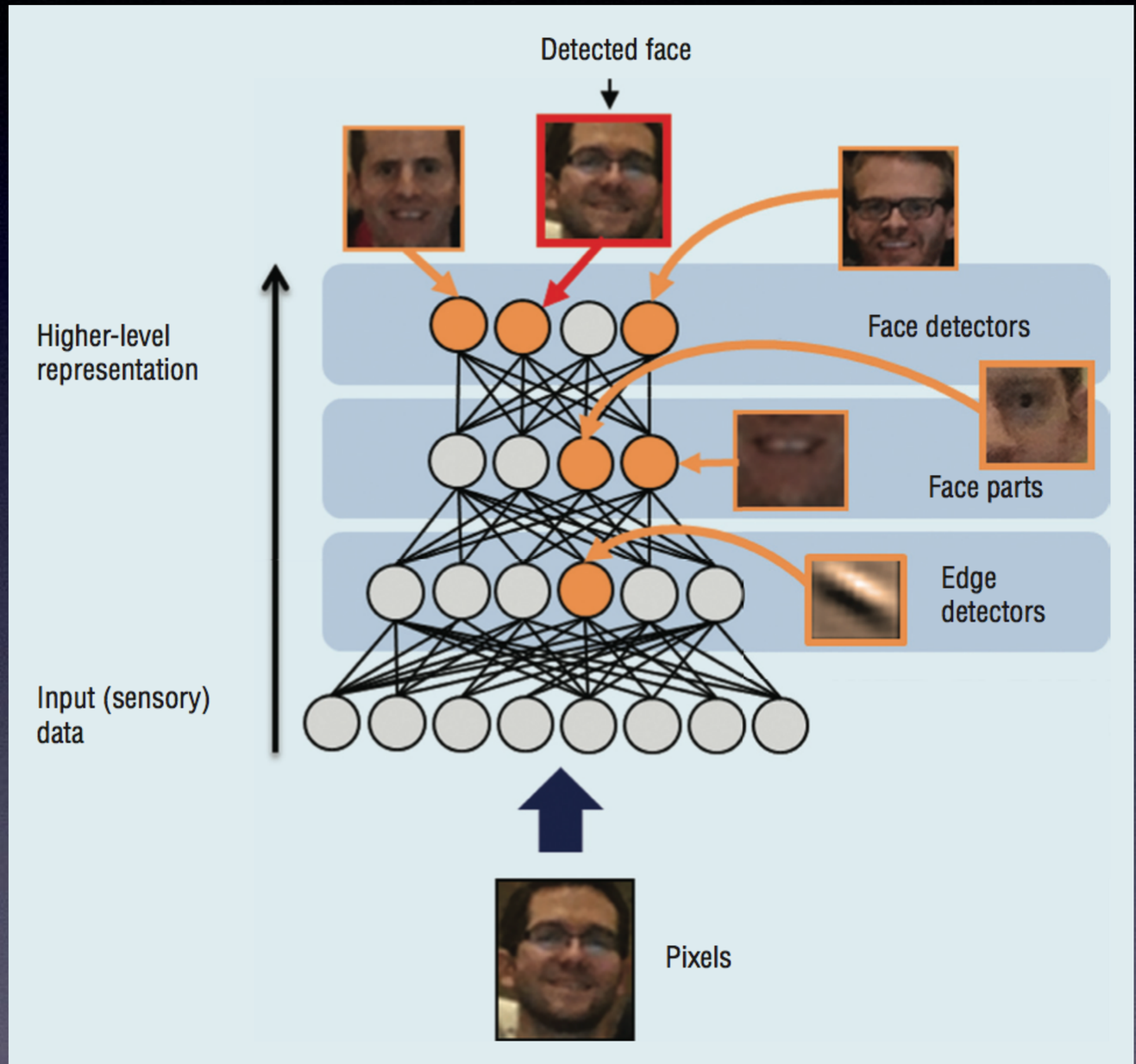
Deep Neural Networks/Deep Learning



DNNs: The Hope

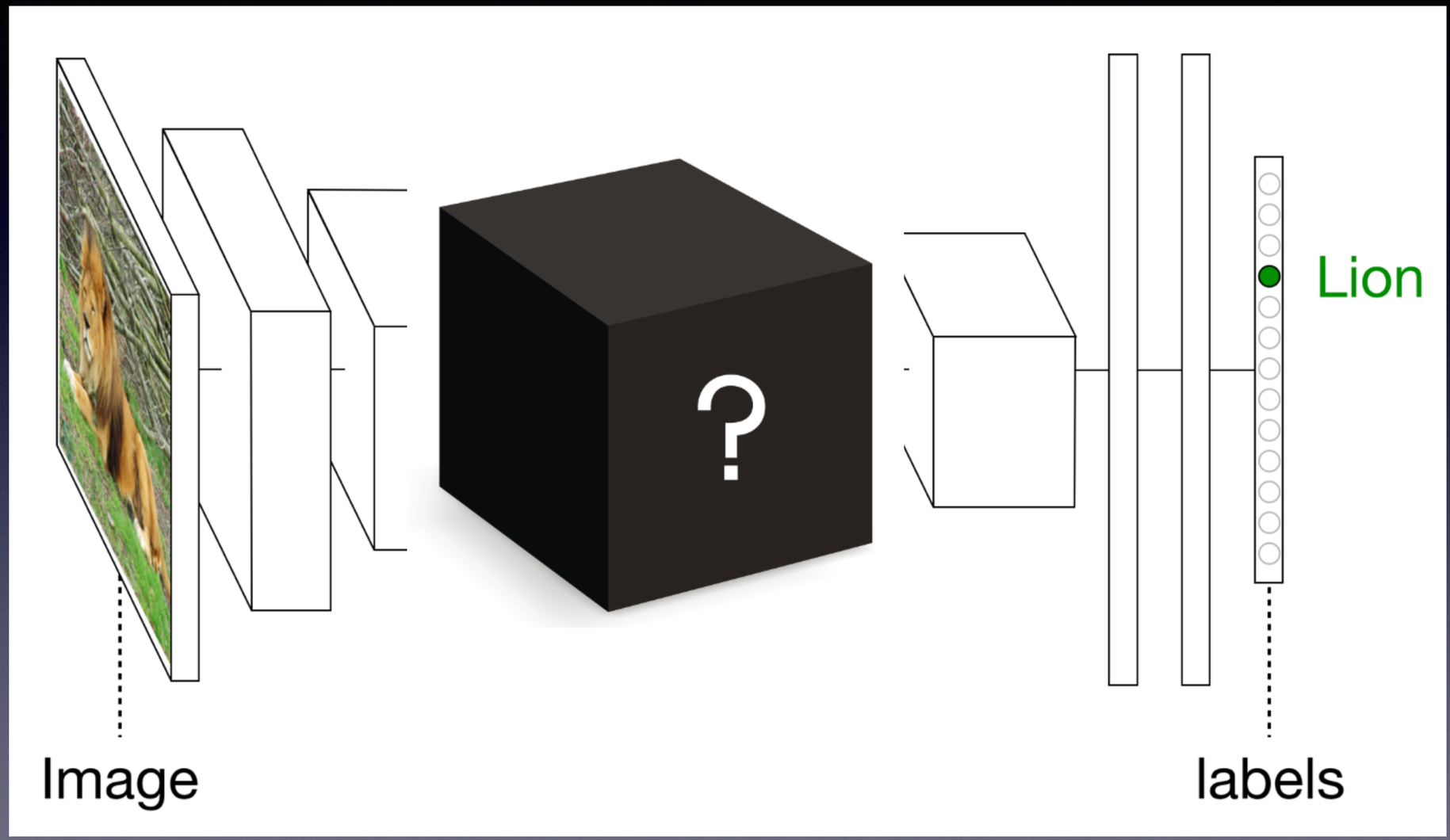
- Each layer transforms data
- Recognizes increasingly abstract “features”

Hierarchically composed feature representations



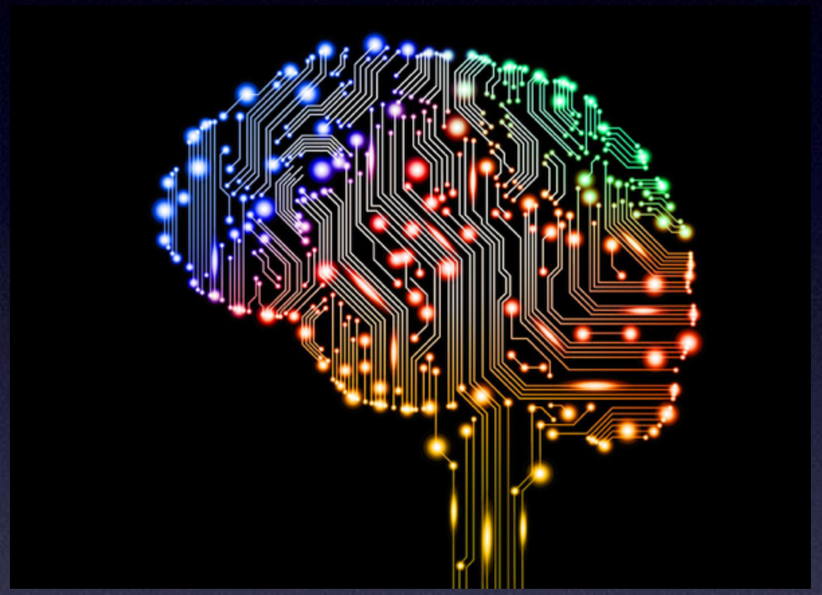
Lehman, Clune, Risi 2014

Deep Neural Networks/Deep Learning



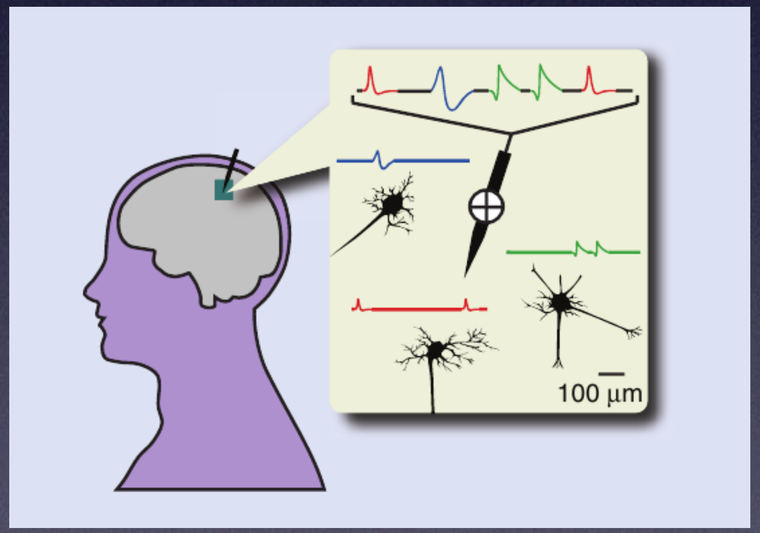
~1M neurons
~100M weights

Thus begins the field of “AI Neuroscience”



One way do neuroscientists do it (of many)

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

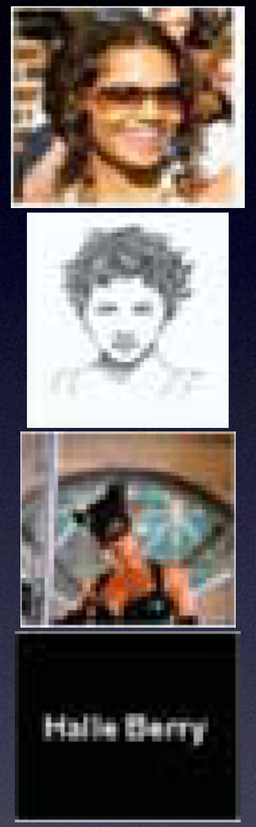


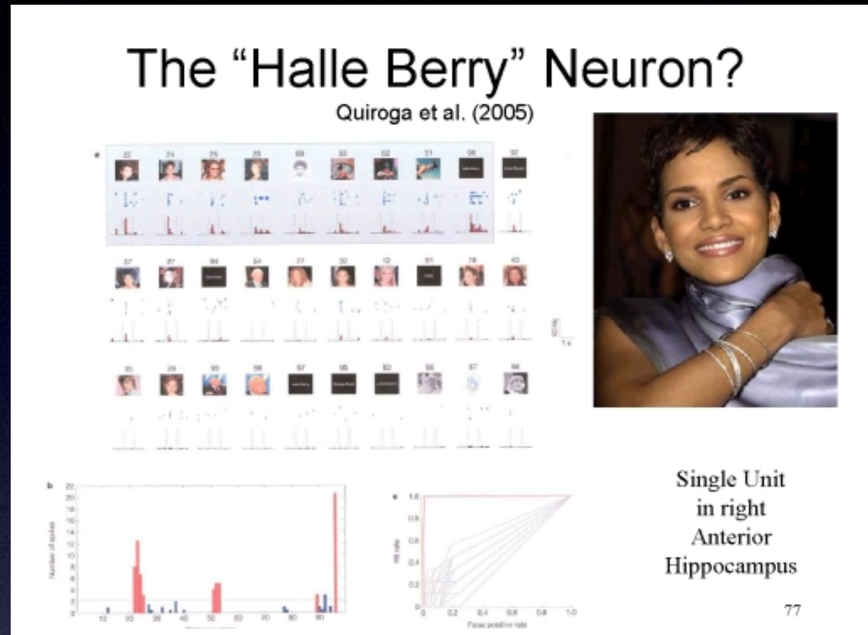
Quiroga et al. Nature 2005

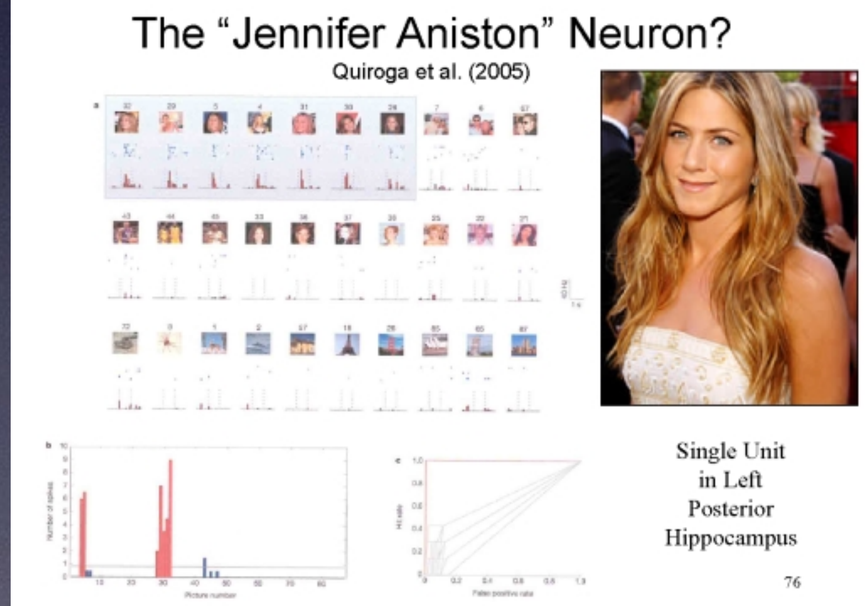
etc….

One way do neuroscientists do it (of many)

- “Halle Berry” Neuron
  - responded to
    - her picture (but not other pictures)
    - a line drawing of her
    - her picture in her Catwoman costume
    - her name (typed out)
- different neurons for
    - Jennifer Aniston
    - Bill Clinton
    - etc.



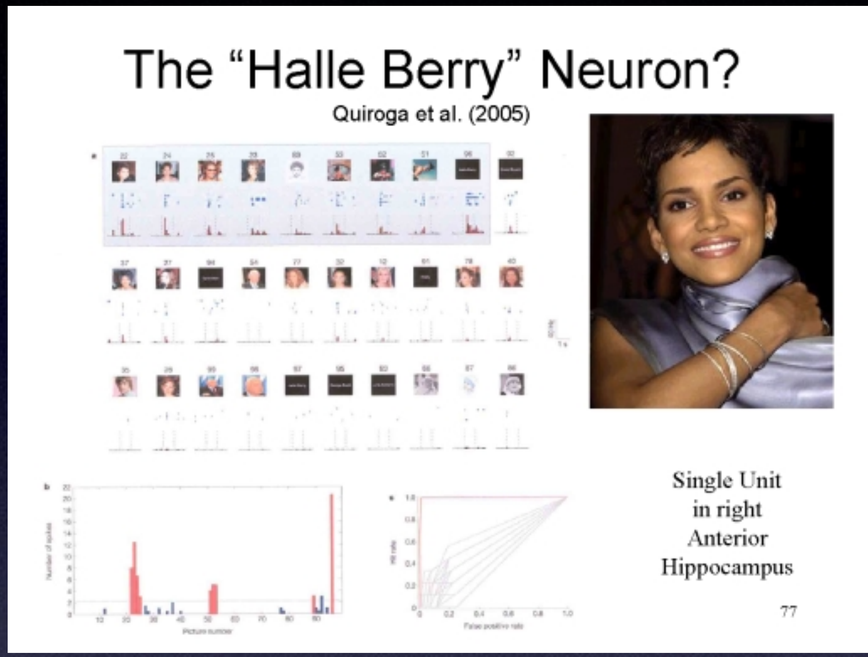




Quiroga et al. Nature 2005

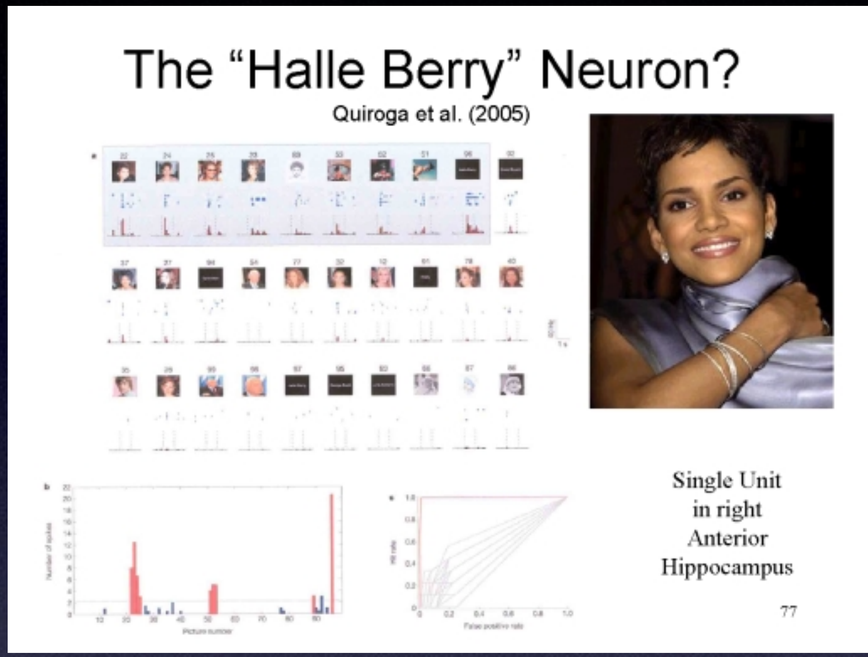
Open Questions

- Neuroscience
  - Is it really a Halle Berry neuron?
    - or an african-american actress neuron?
    - or?



Open Questions

- Neuroscience
  - Is it really a Halle Berry neuron?
    - or an african-american actress neuron?
    - or?



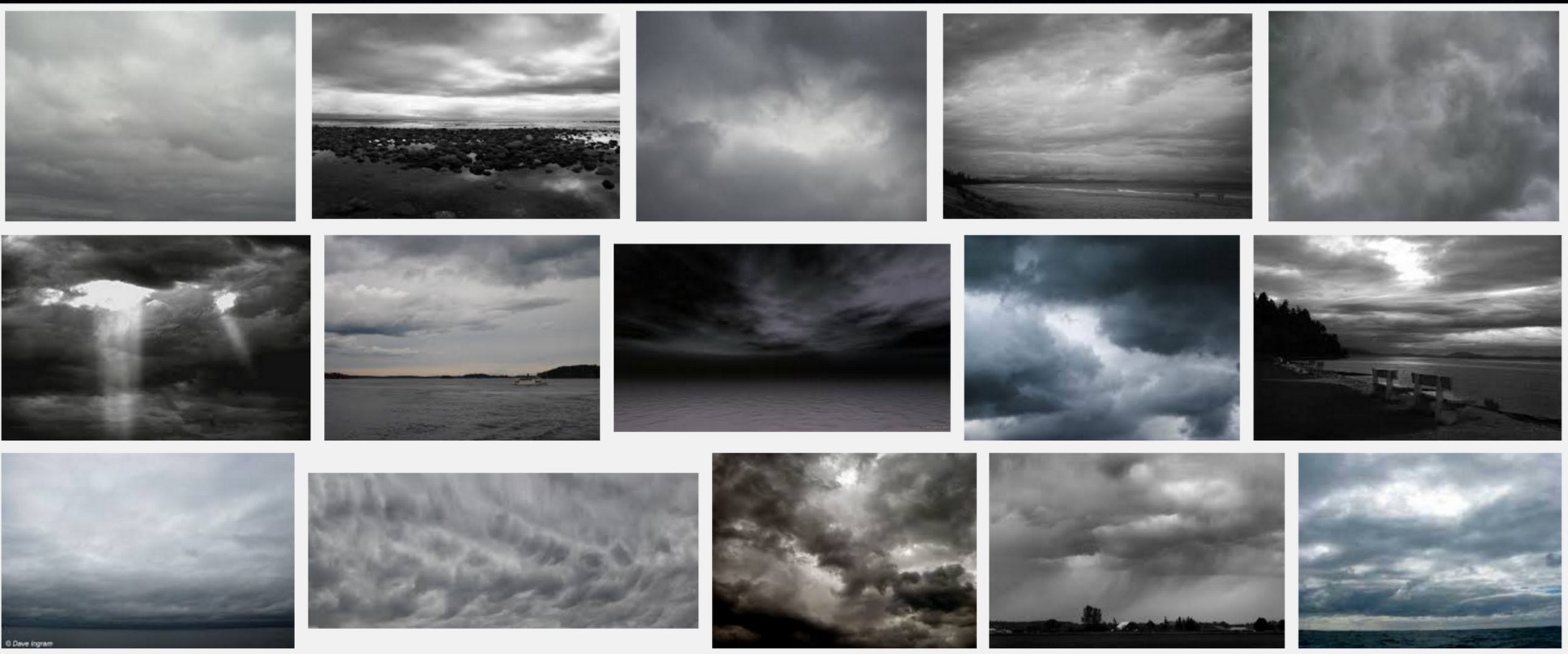
Ideal Test: Synthesize Preferred Inputs

African-american actress neuron

Ideal Test: Synthesize Preferred Inputs

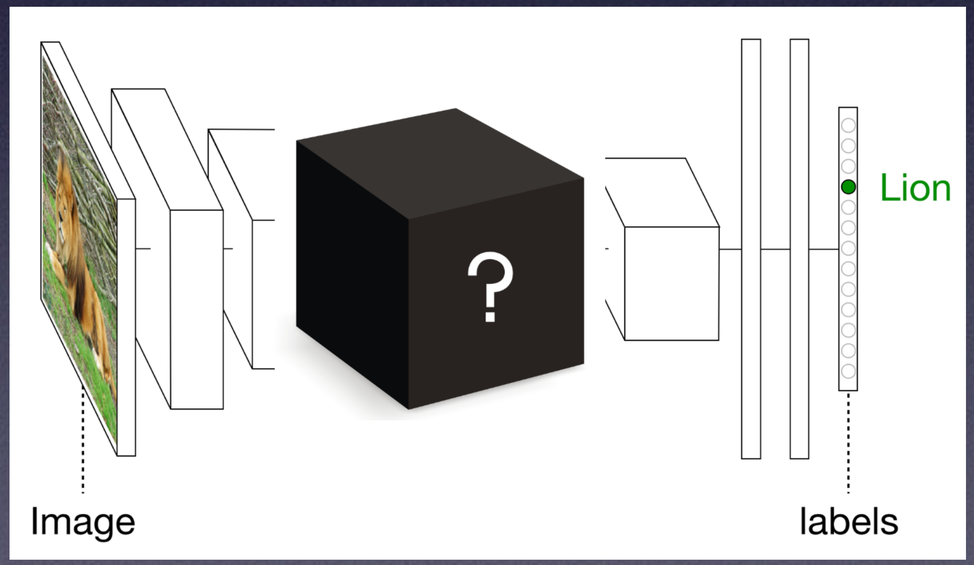
Halle Berry Neuron

Ideal Test: Synthesize Preferred Inputs

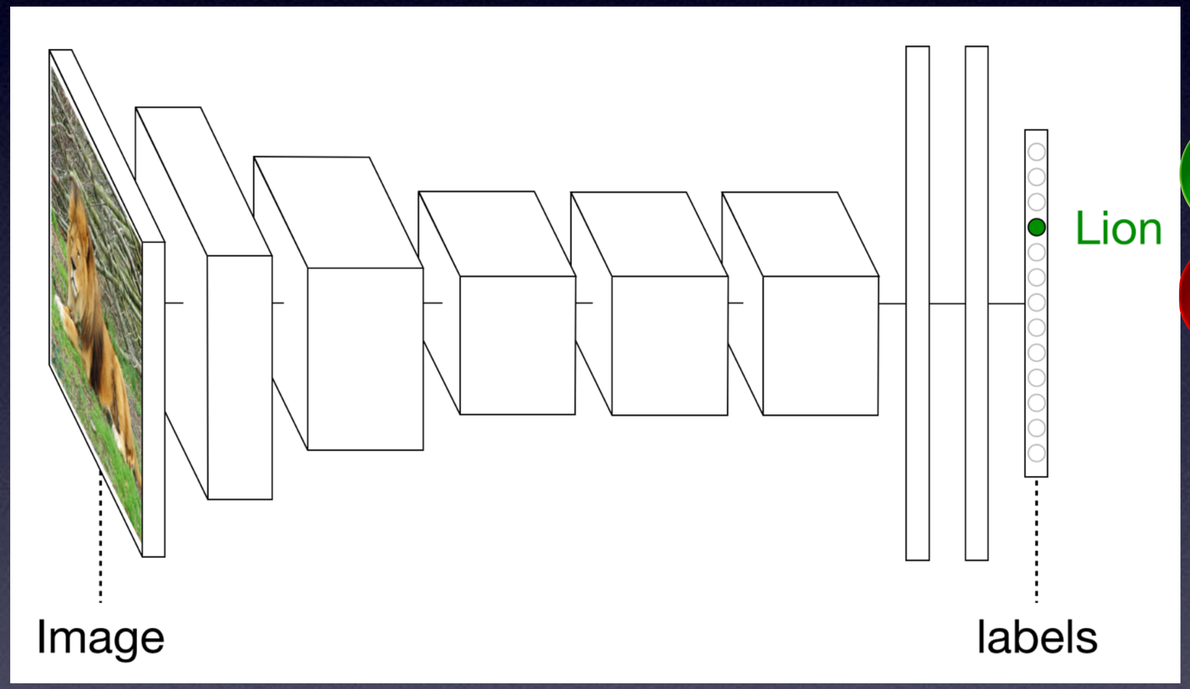


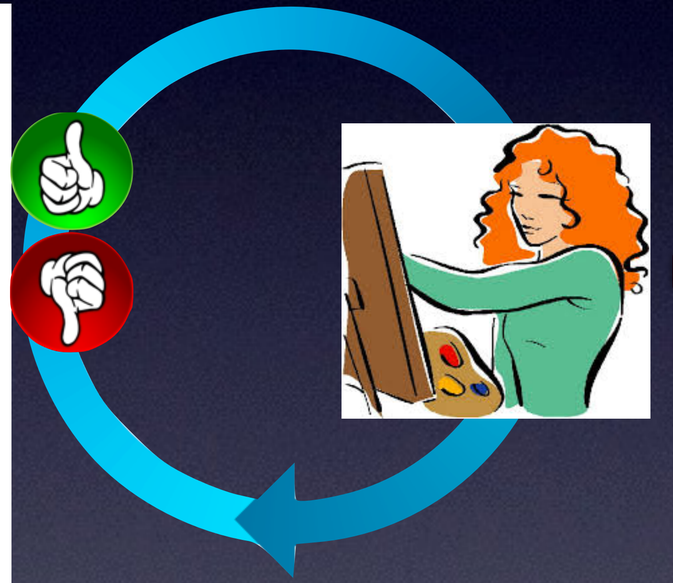
Possible with Artificial Neural Networks!

- How much do DNNs understand about the objects they recognize?
- Do they have increasingly abstract feature detectors?
  - Are they “multifaceted”
- Are representations “local” or “distributed”?
- What sorts of errors are they likely to make?

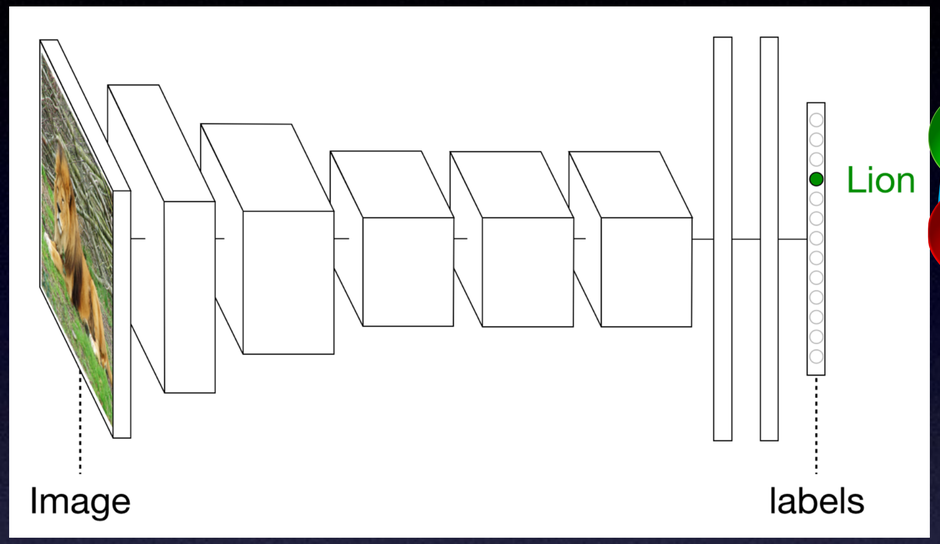


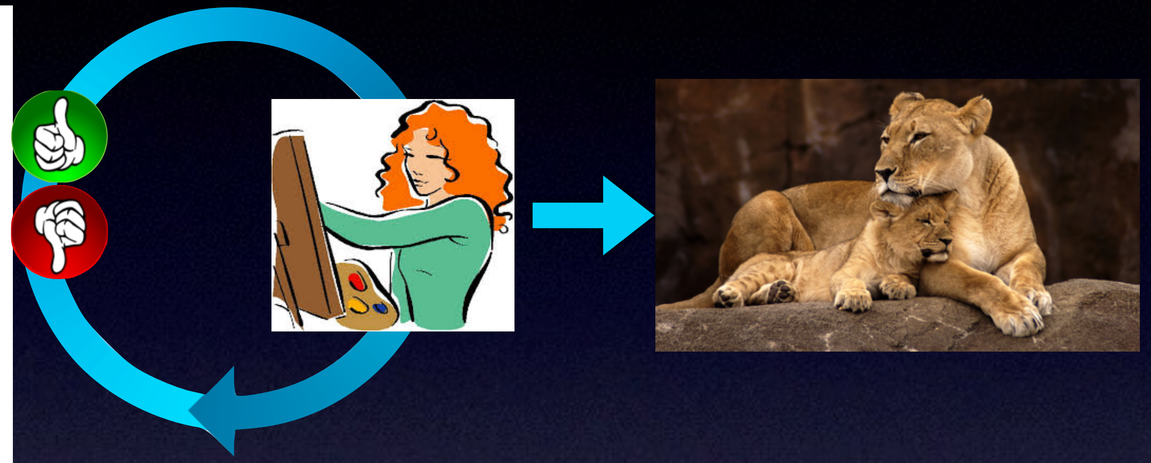
Investigating What Each Neuron Does

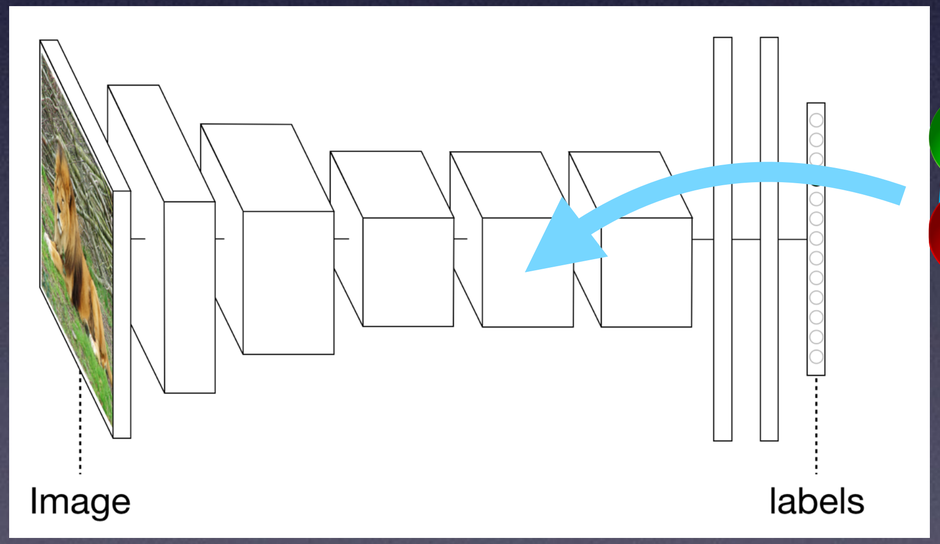


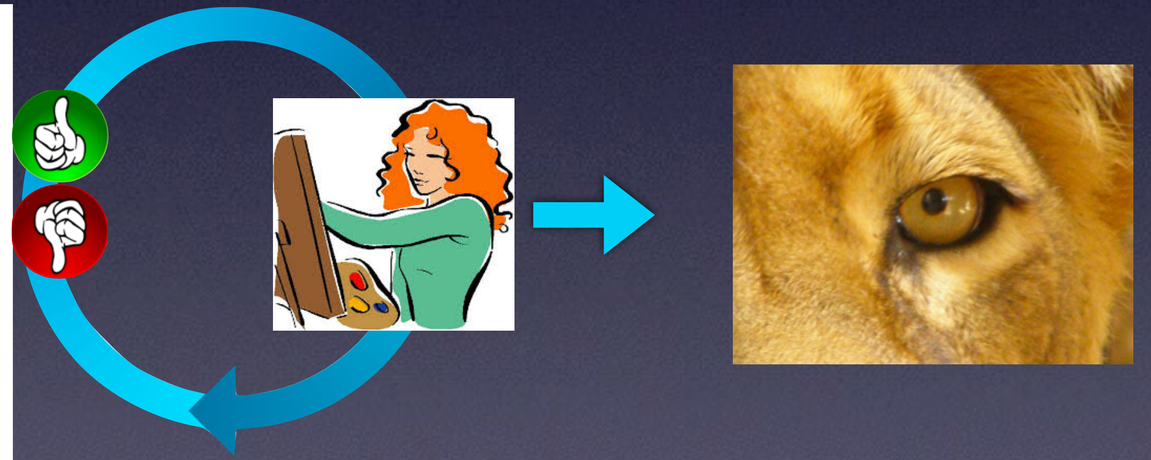


Investigating What Each Neuron Does

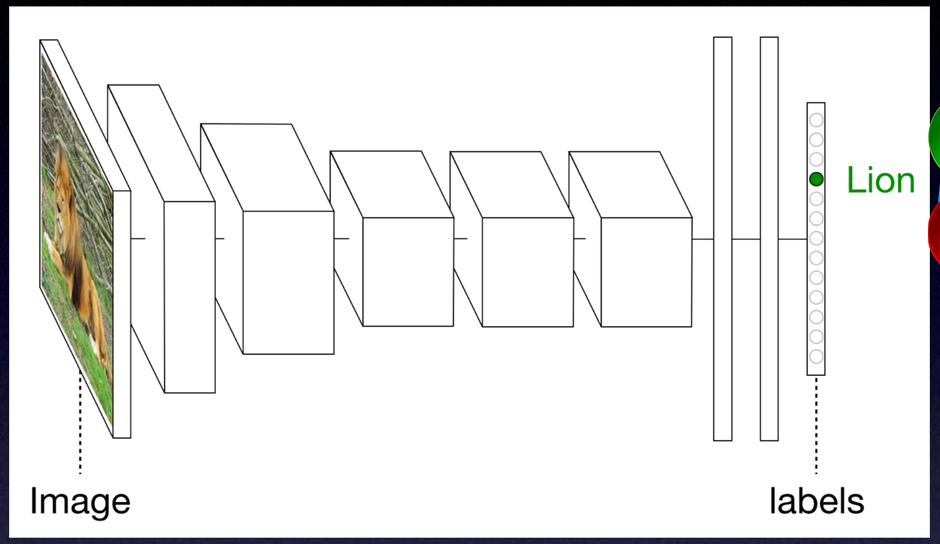


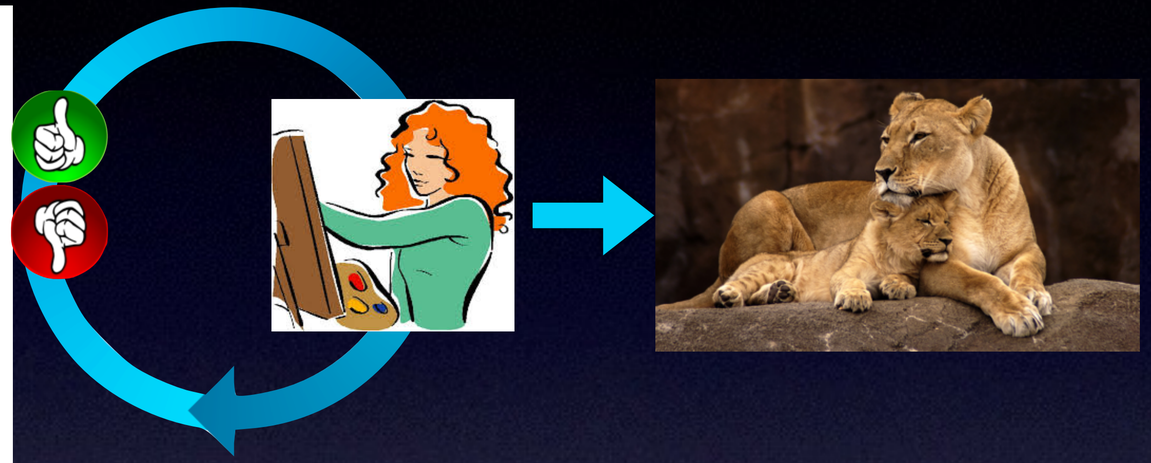


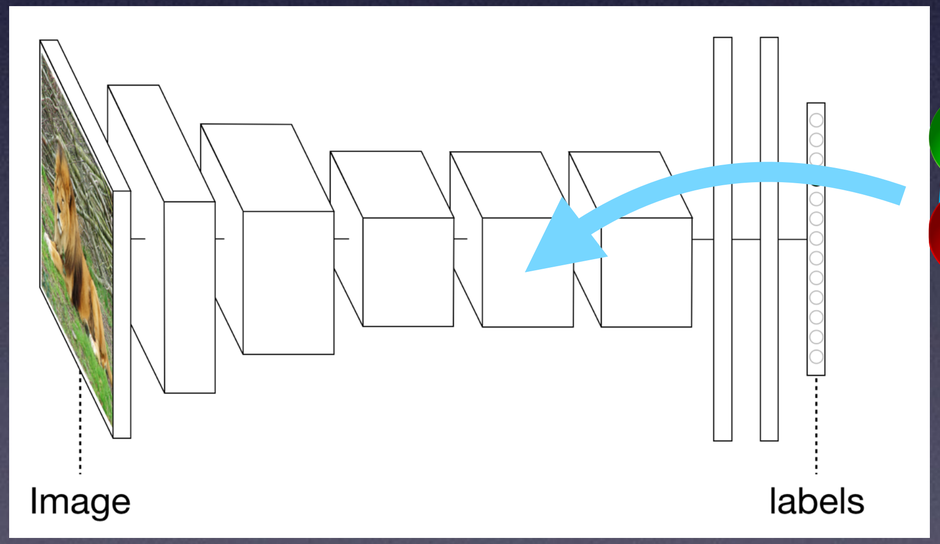


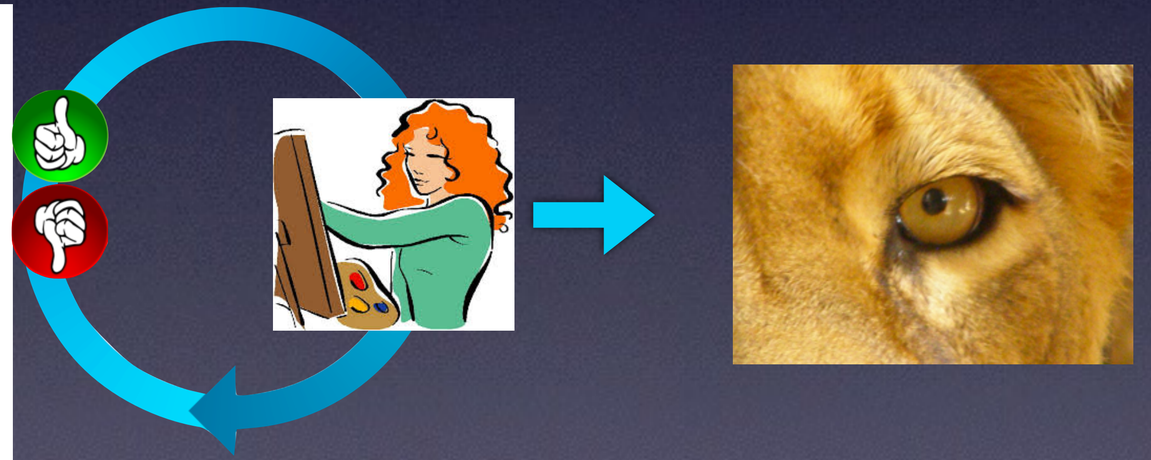


“Deep Visualization”



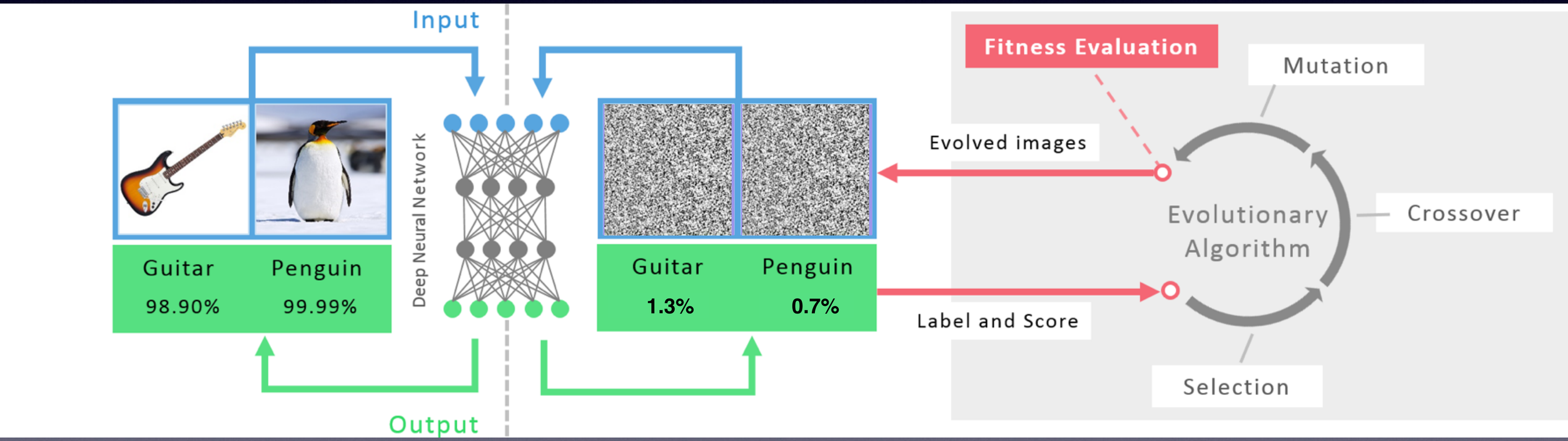






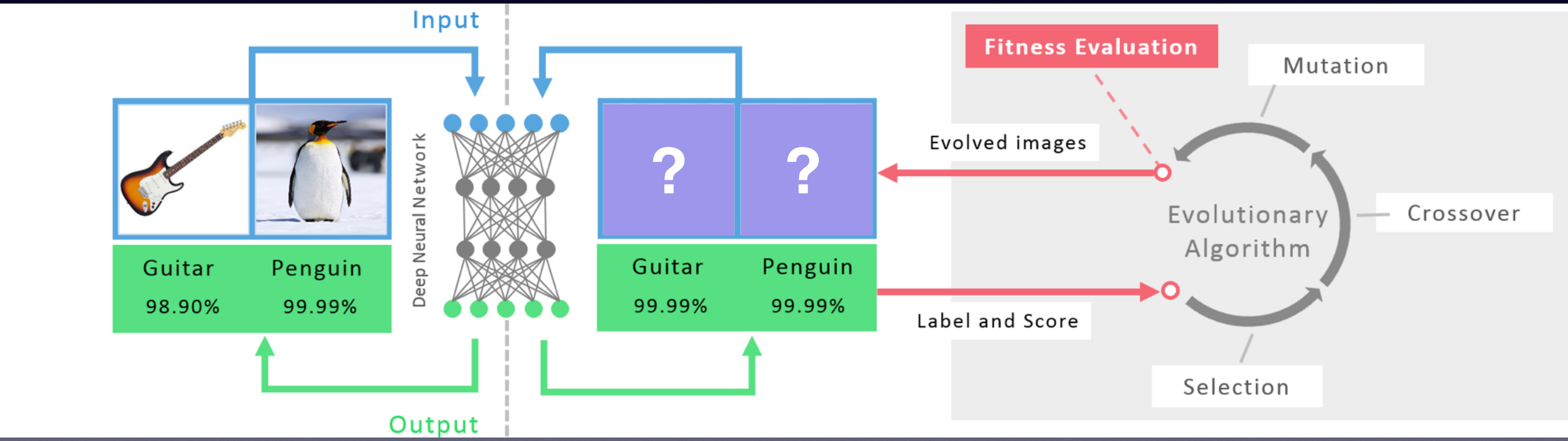
Deep Visualization Take 1

Nguyen, Yosinski, Clune, 2015, CVPR



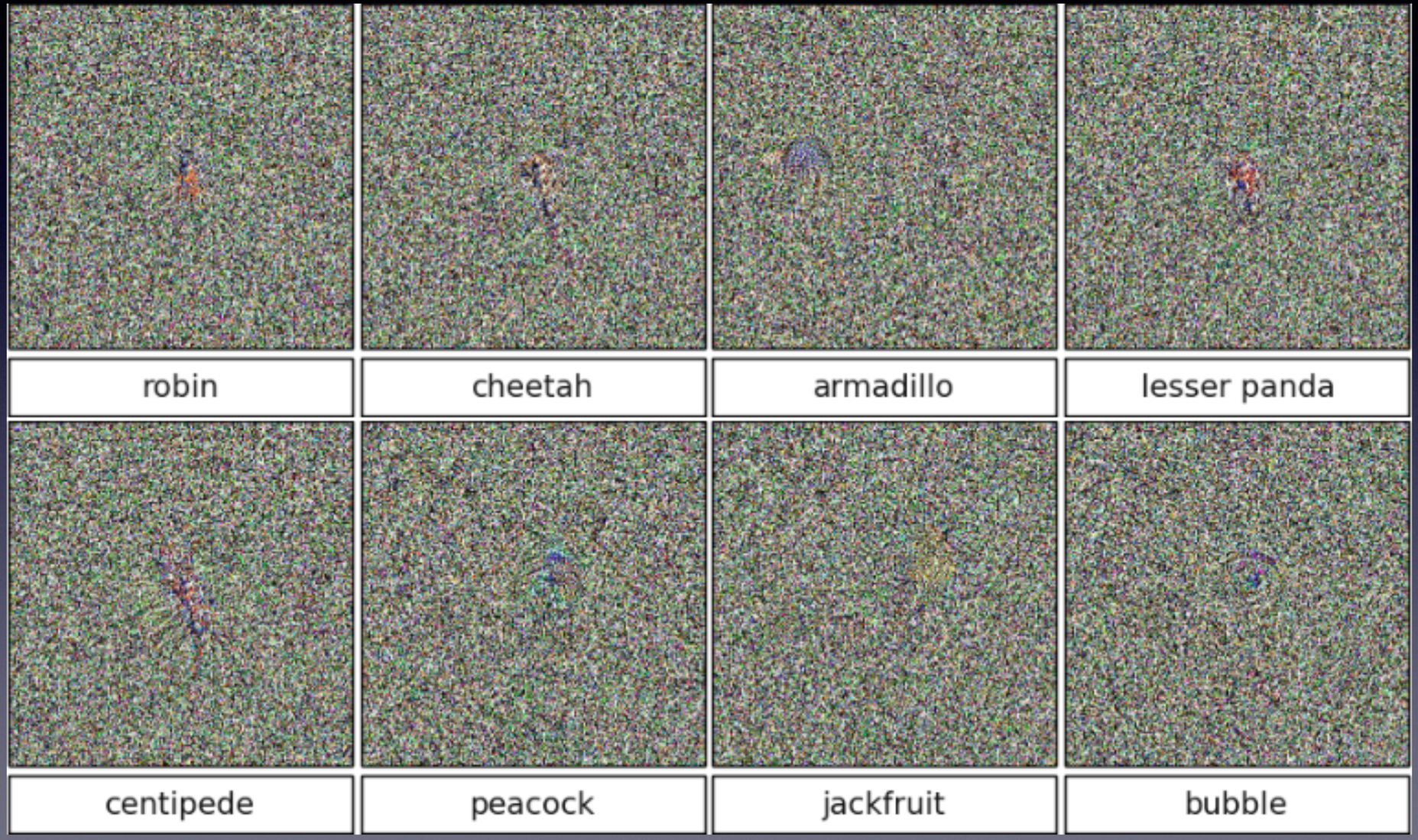
Deep Visualization Take 1

Nguyen, Yosinski, Clune, 2015, CVPR



Deep Visualization Take 1

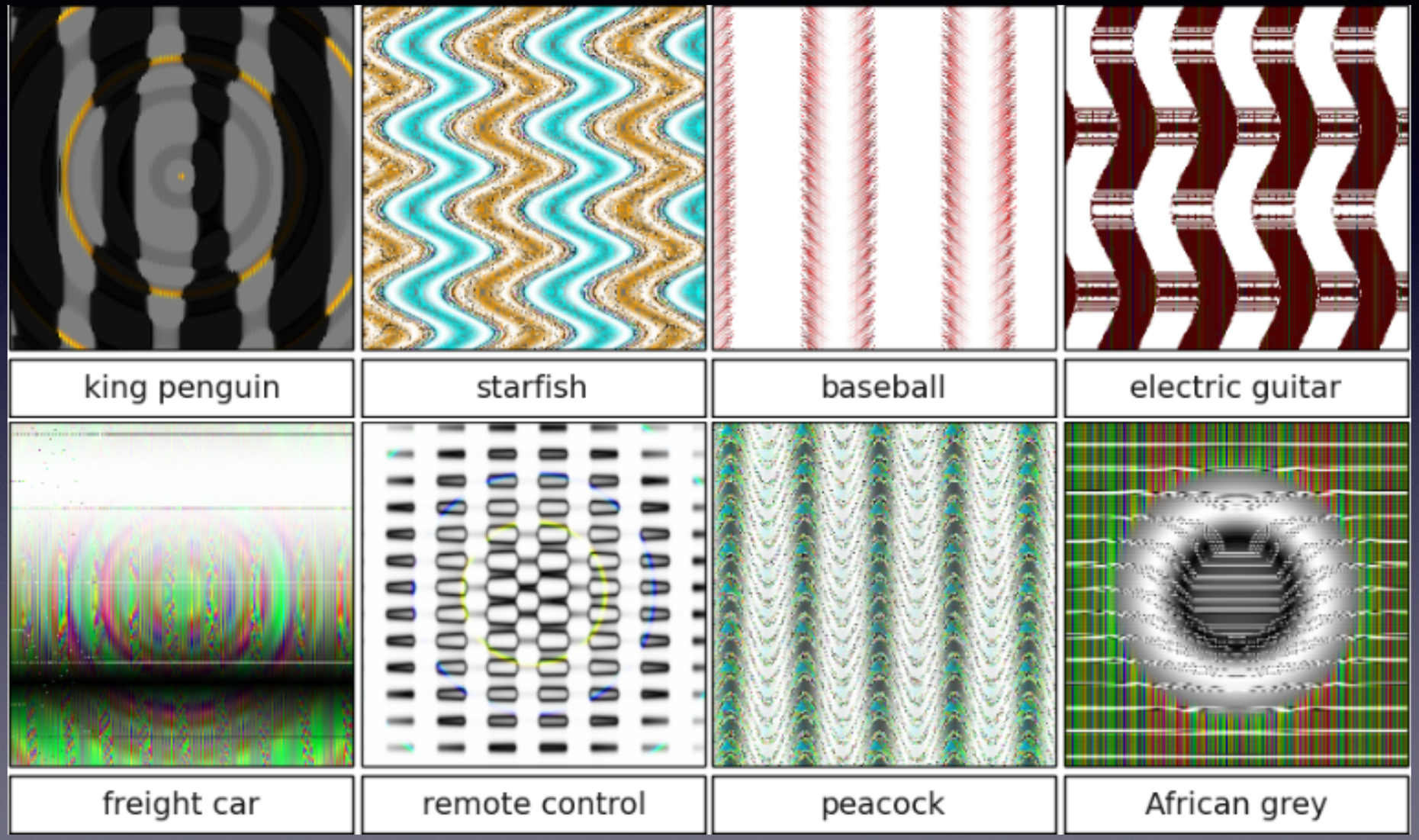
Nguyen, Yosinski, Clune, 2015, CVPR



DNN Confidence: > 99.6 % for all

Deep Visualization Take 1

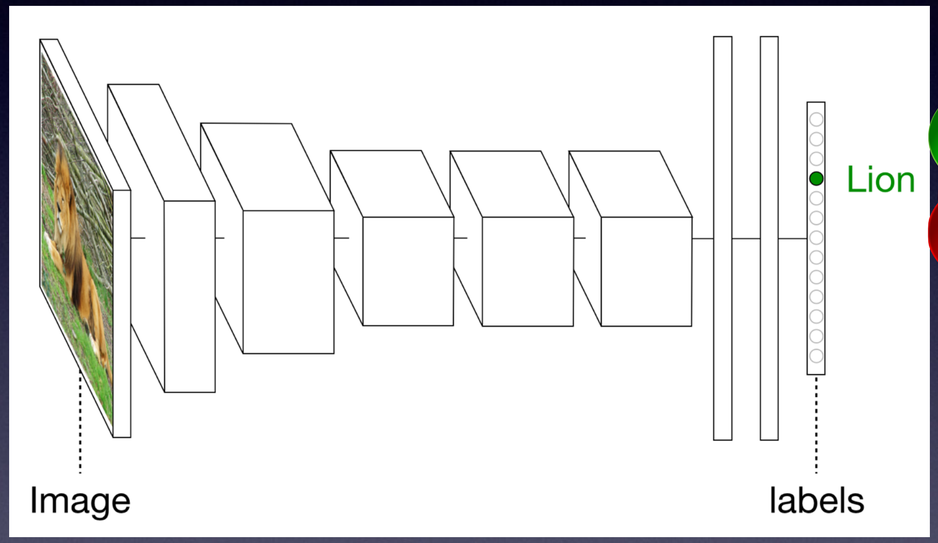
Nguyen, Yosinski, Clune, 2015, CVPR

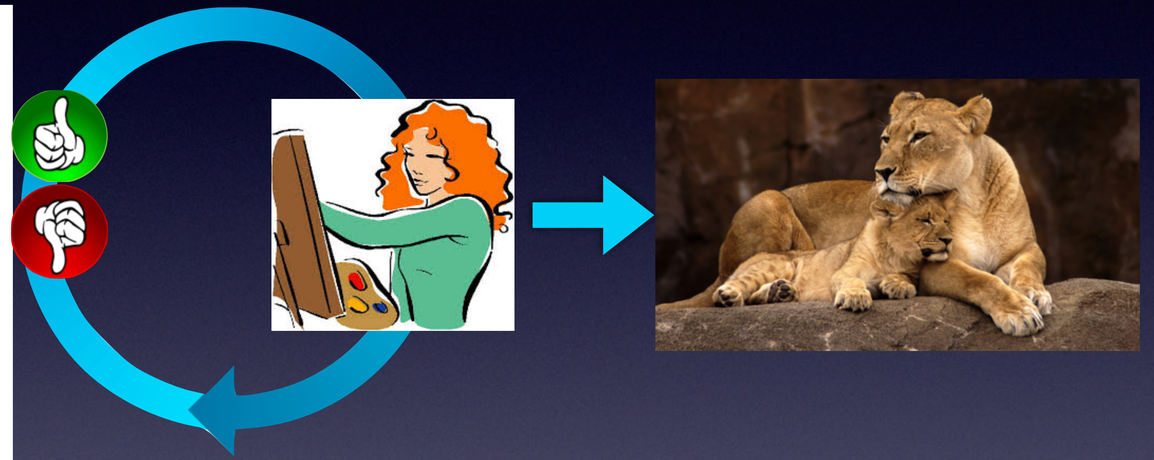


DNN Confidence: > 99.6 % for all

Deep Visualization Take 1

Nguyen, Yosinski, Clune, 2015, CVPR

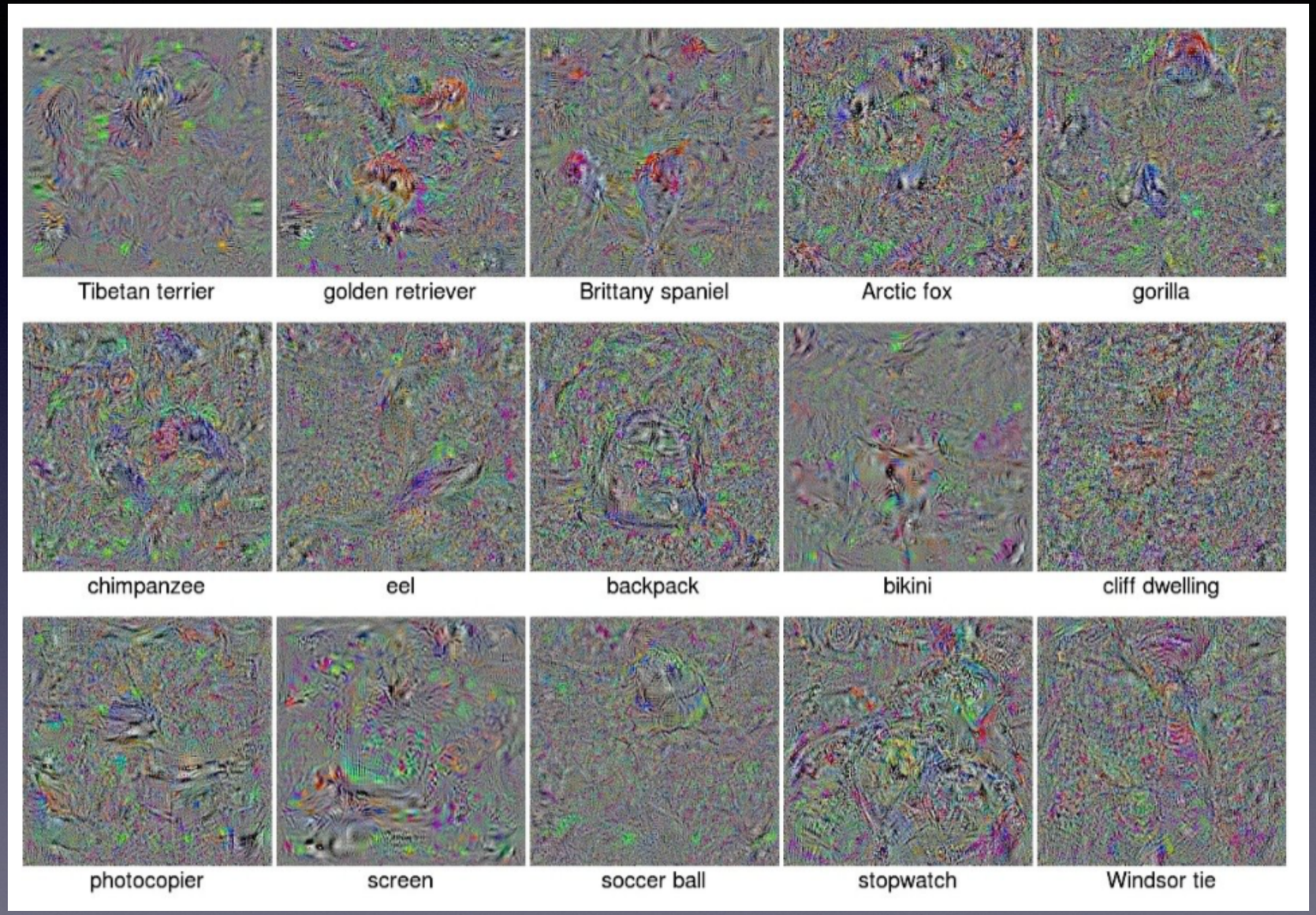




Backprop to
Pixel Changes

Deep Visualization Take 1

Nguyen, Yosinski, Clune, 2015, CVPR



DNN Confidence: > 99.9 % for all

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.

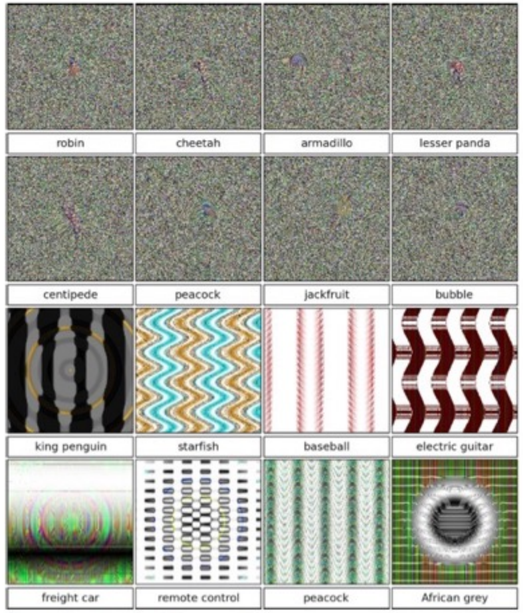


Figure 1. Evolved images that are unrecognizable to humans, but that state-of-the-art DNNs trained on ImageNet believe with $\geq$ 99.6% certainty to be a familiar object. This result highlights differences between how DNNs and humans recognize objects. Images are either directly (*top*) or indirectly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

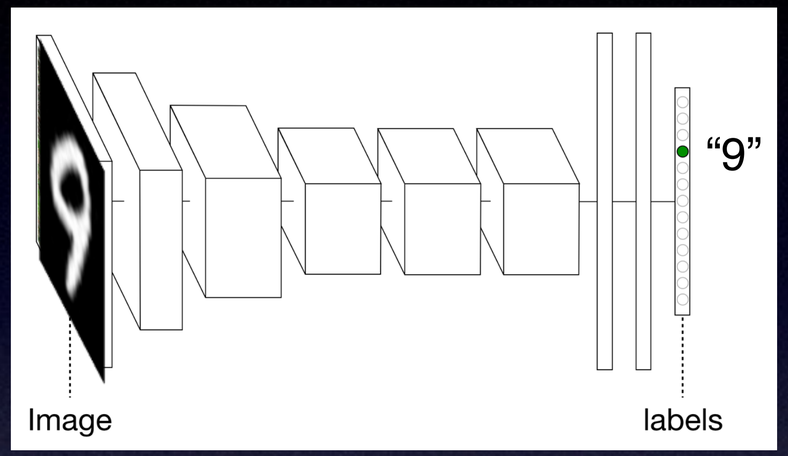
A recent study revealed a major difference between DNN and human vision [30]. Changing an image, originally correctly classified (e.g. as a lion), in a way imperceptible to human eyes, can cause a DNN to label the image as something else entirely (e.g. mislabeling a lion a library).

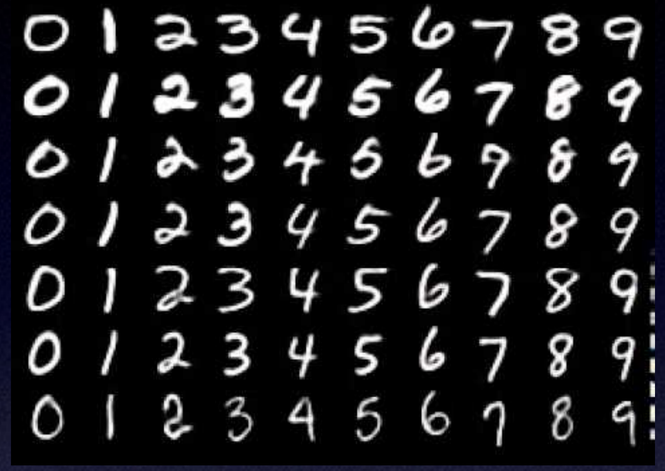
In this paper, we show another way that DNN and human vision differ: It is easy to produce images that are completely unrecognizable to humans (Fig. 1), but that state-of-the-art DNNs believe to be recognizable objects with over 99% confidence (e.g. labeling with certainty that TV static

- AI "sees" the world differently
- Clever Hans?
- Huge security concern

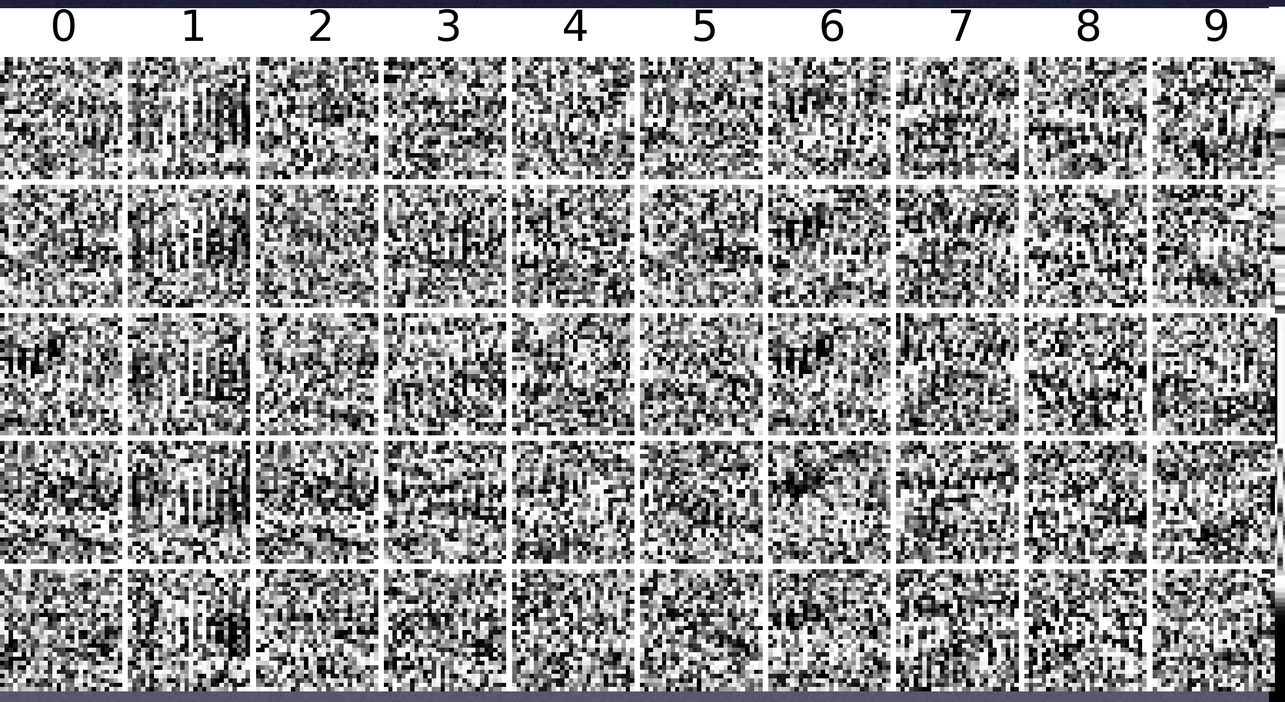
School bus

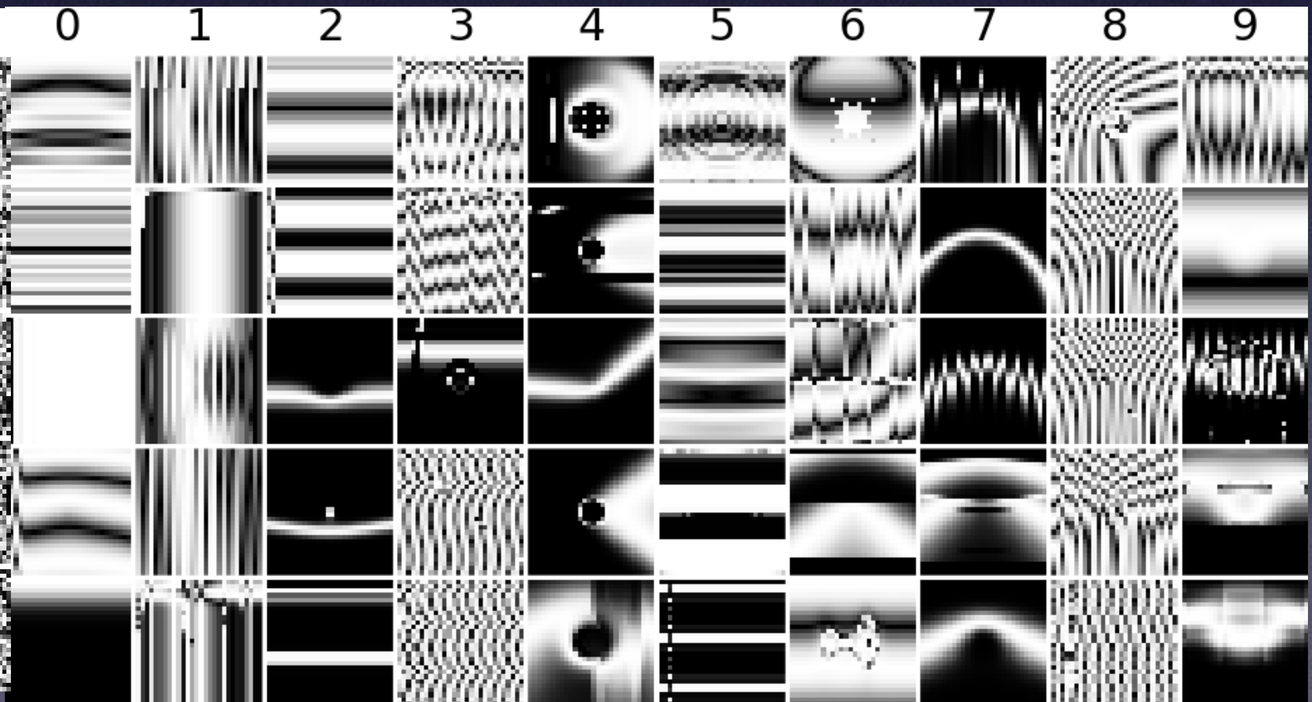
Digits



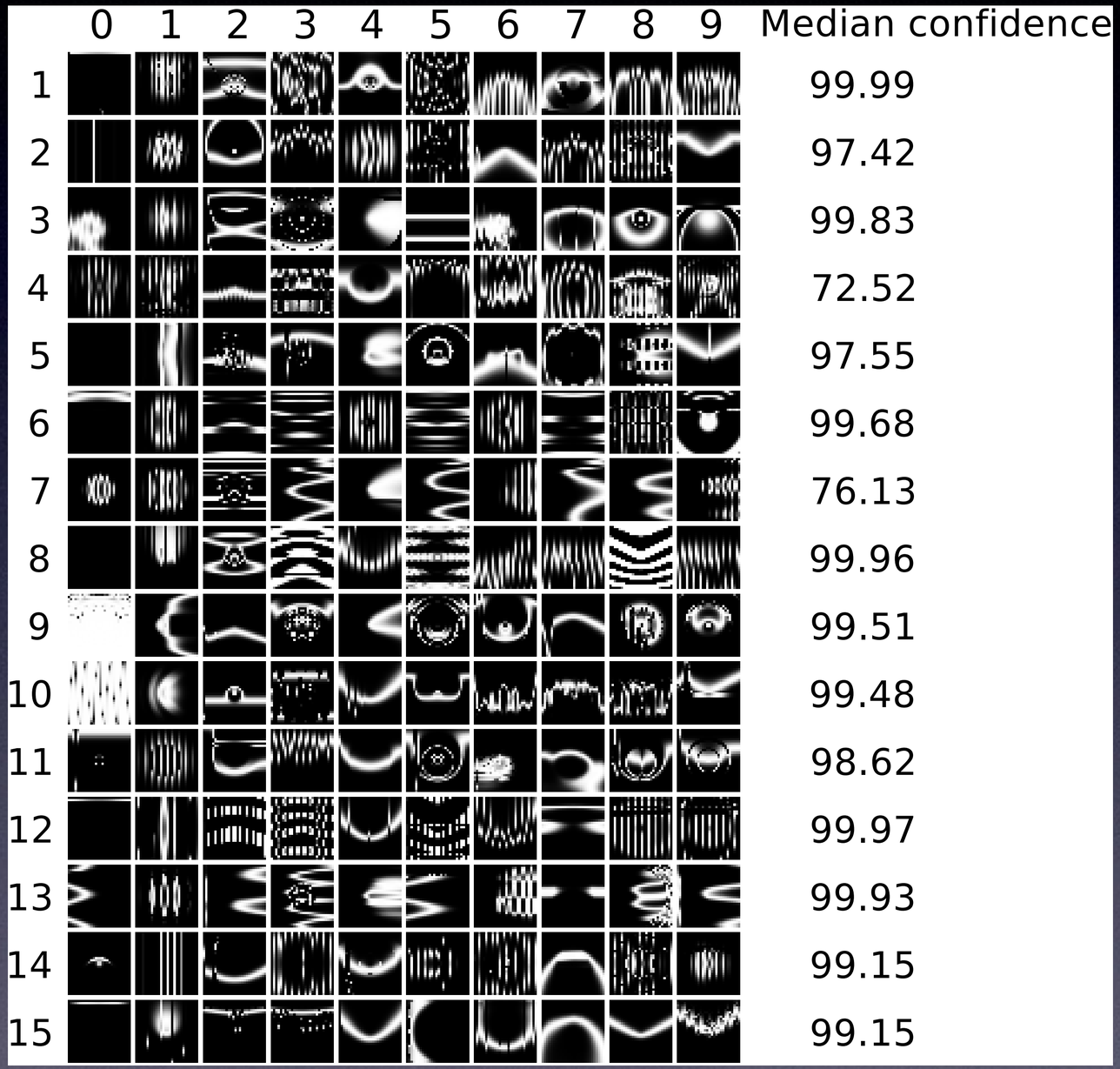


>99% accurate

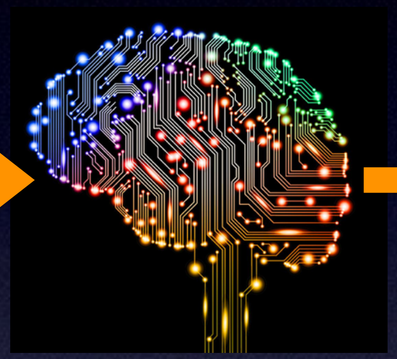




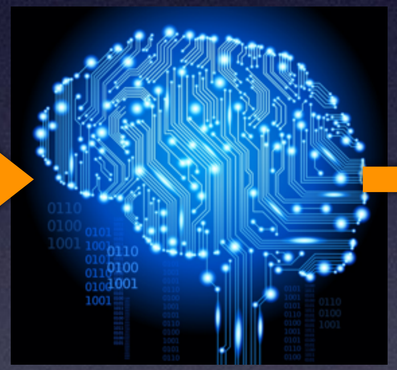
Training on these examples does not help



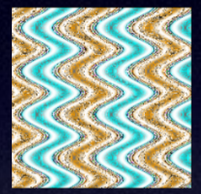
Images that fool one network fool others!

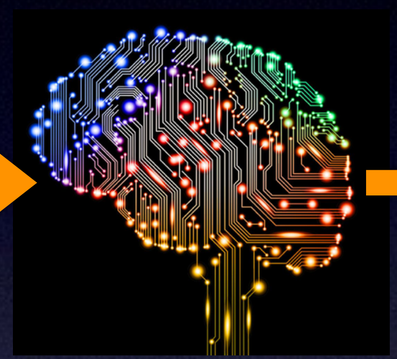


Lion

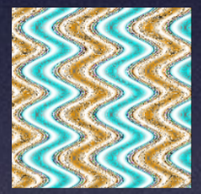


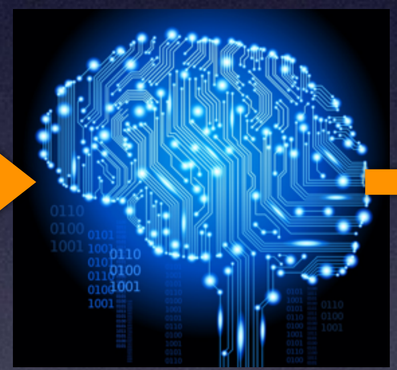
Lion





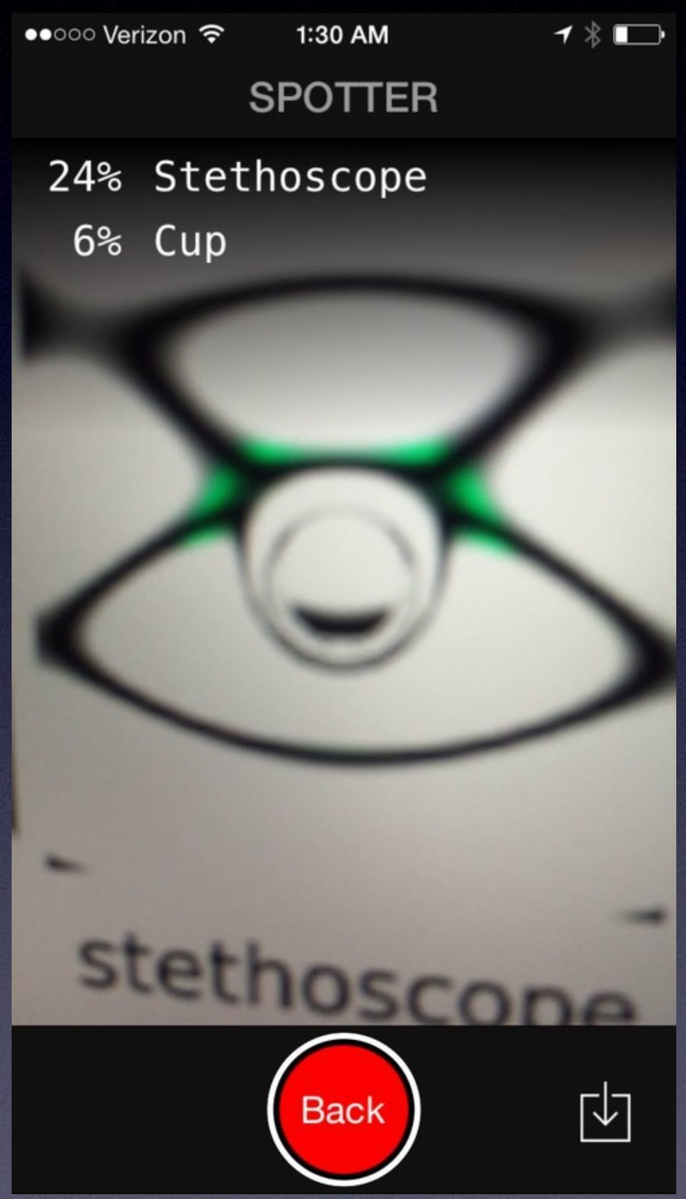
Starfish

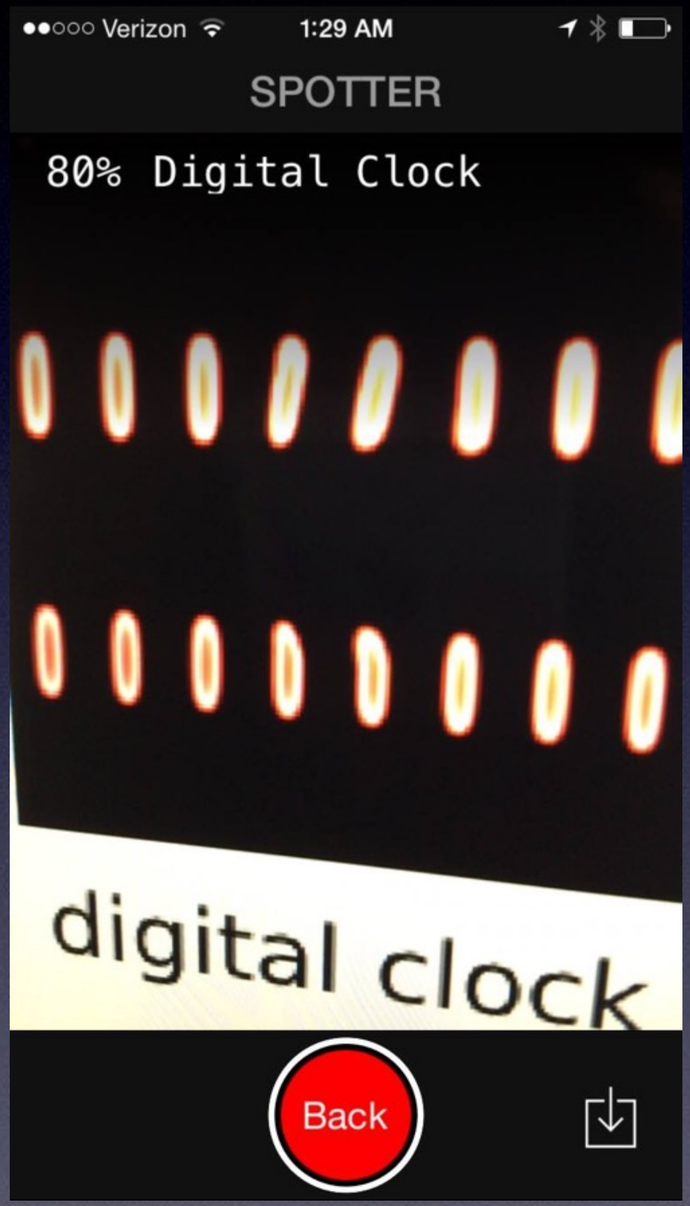




Starfish

Images that fool one network fool others!





Courtesy: Dileep George, co-founder Vicarious

Huge reaction

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

The Economist

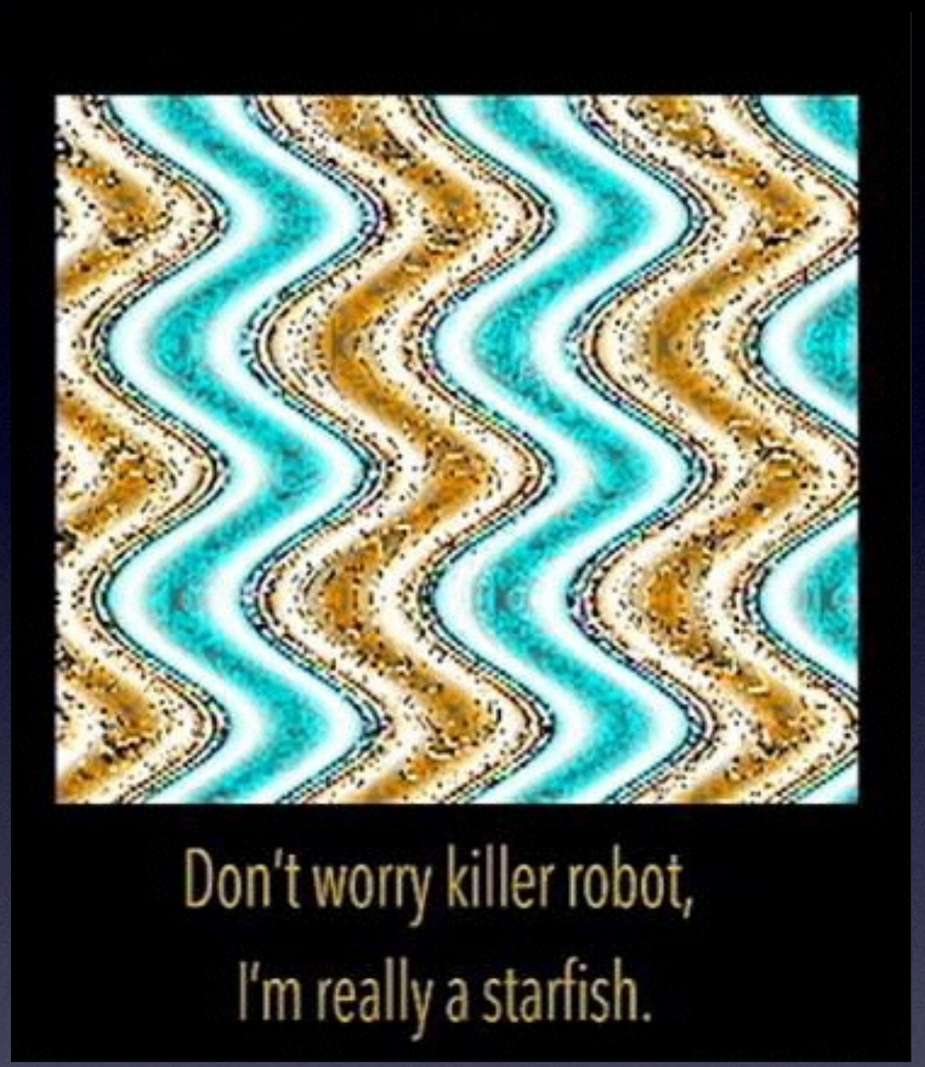
The Atlantic

WIRED

NewScientist

SCIENTIFIC AMERICAN MIND
BEHAVIOR • BRAIN SCIENCE • INSIGHTS

MIT Technology Review



Larry Page,
Google co-Founder

Gary Marcus,
NYU Prof & CEO

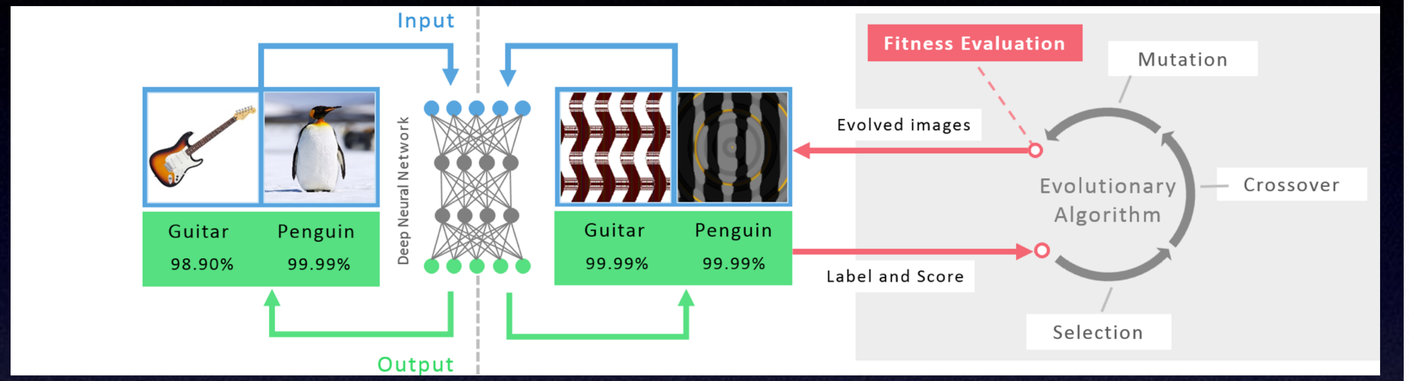
Well-received Academically

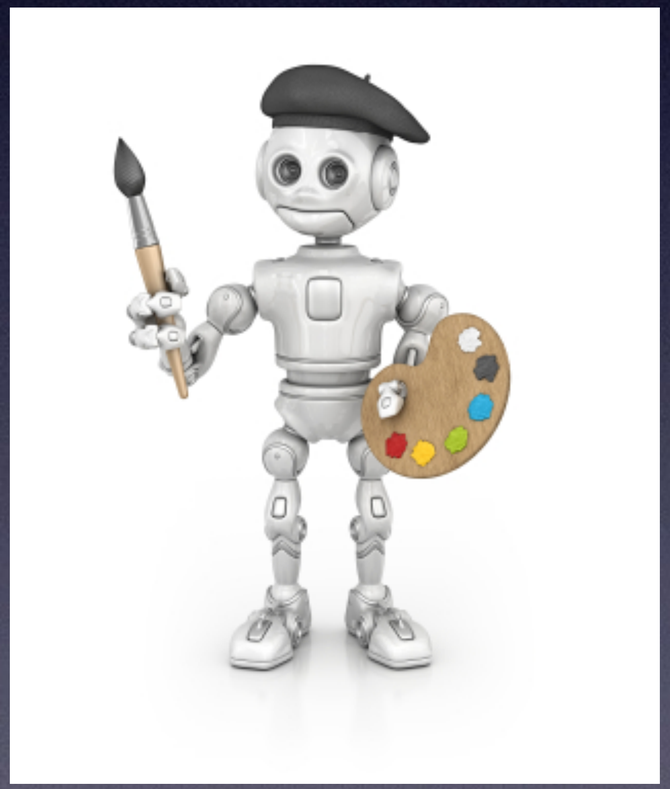
- Computer Vision and Pattern Recognition (CVPR)
  - Oral presentation (3% acceptance rate)
- Winner: Community Top Paper Award (3rd place)

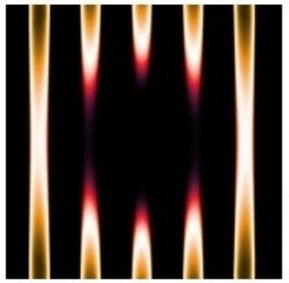
What does the huge reaction mean?

- A lot of interest in whether Deep Learning is “real” intelligence
  - or a “Clever Hans”

Automatic Art Generator

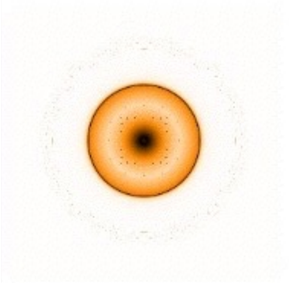






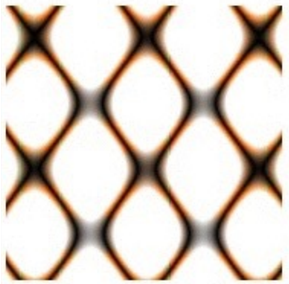

Matchstick


Television



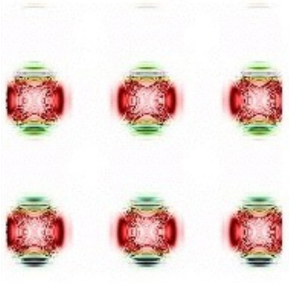

Bagel


Prison

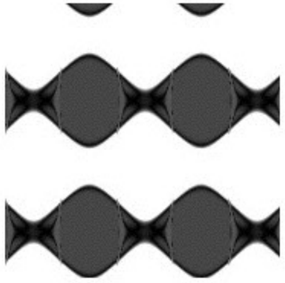



Chainlink fence


Tile roof




Strawberry

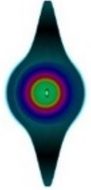



Sunglasses

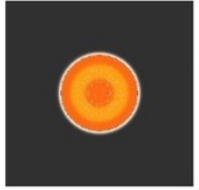

obelisk

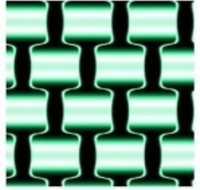

chainlink fence


beacon


digital watch


prison


orange


computer keyboard


pizza


pool table

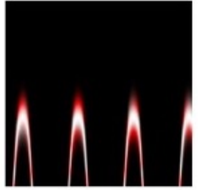
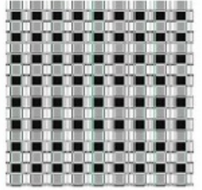

matchstick

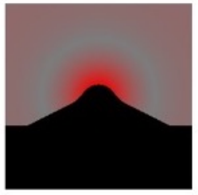

table lamp

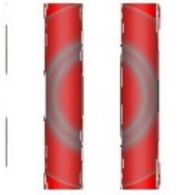

crossword puzzle


mixing bowl


spotlight


combination lock


volcano


punching bag


speaker


fire truck


backpack


car mirror


bee


tile roof


ski mask


panpipe


caldron


dome

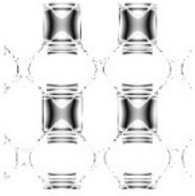

traffic light


sunglass

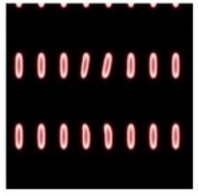

projector


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

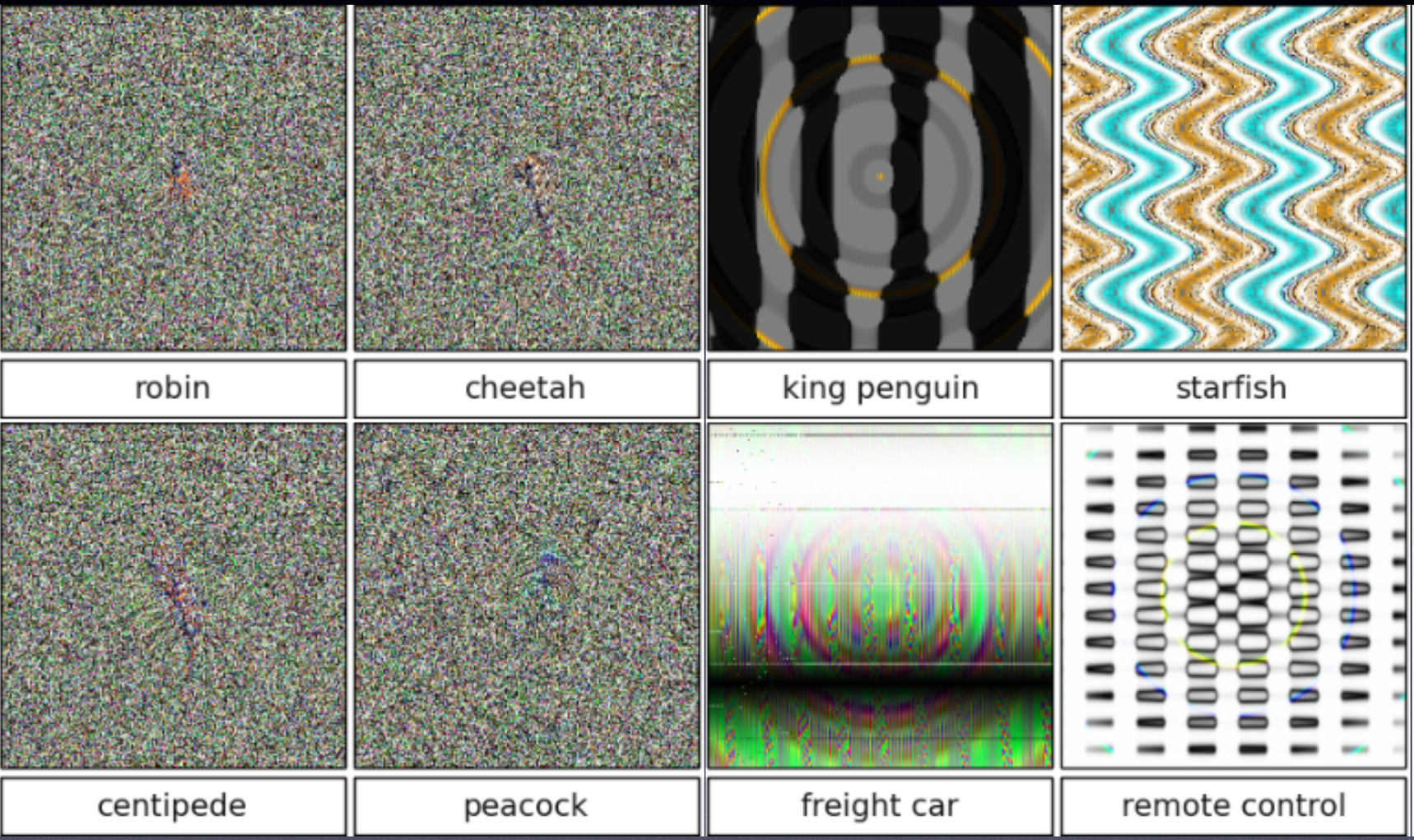
cup

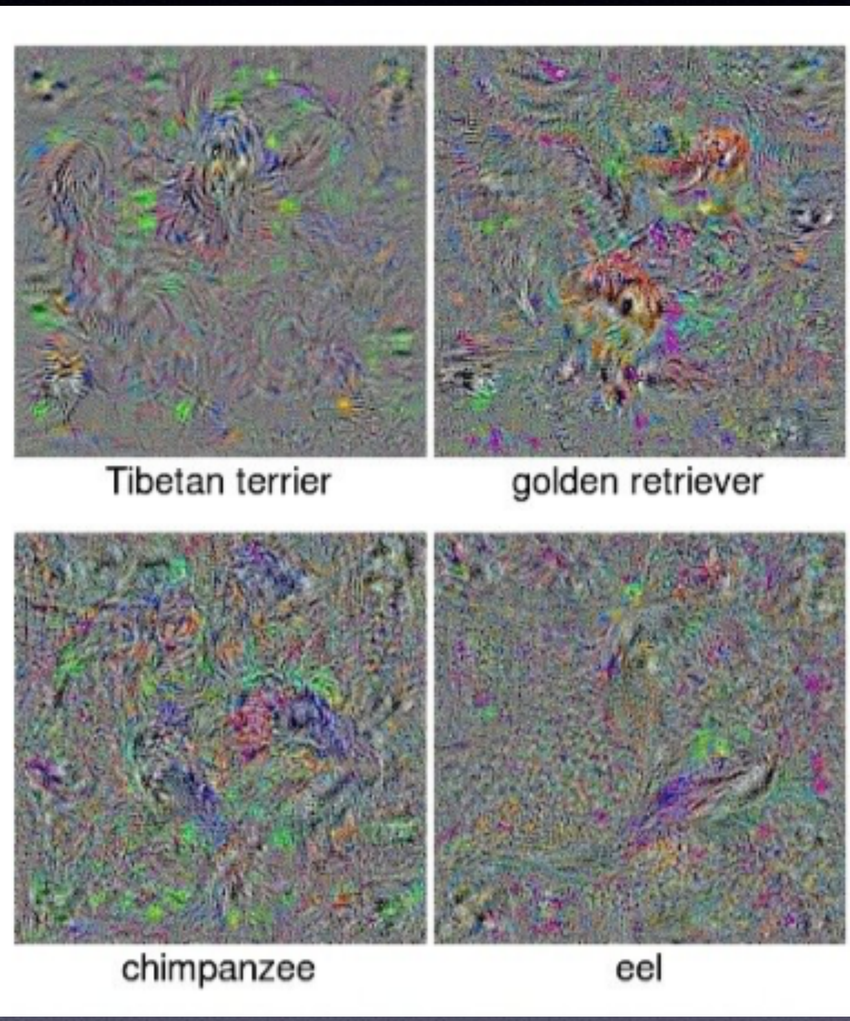
oboe

bucket

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

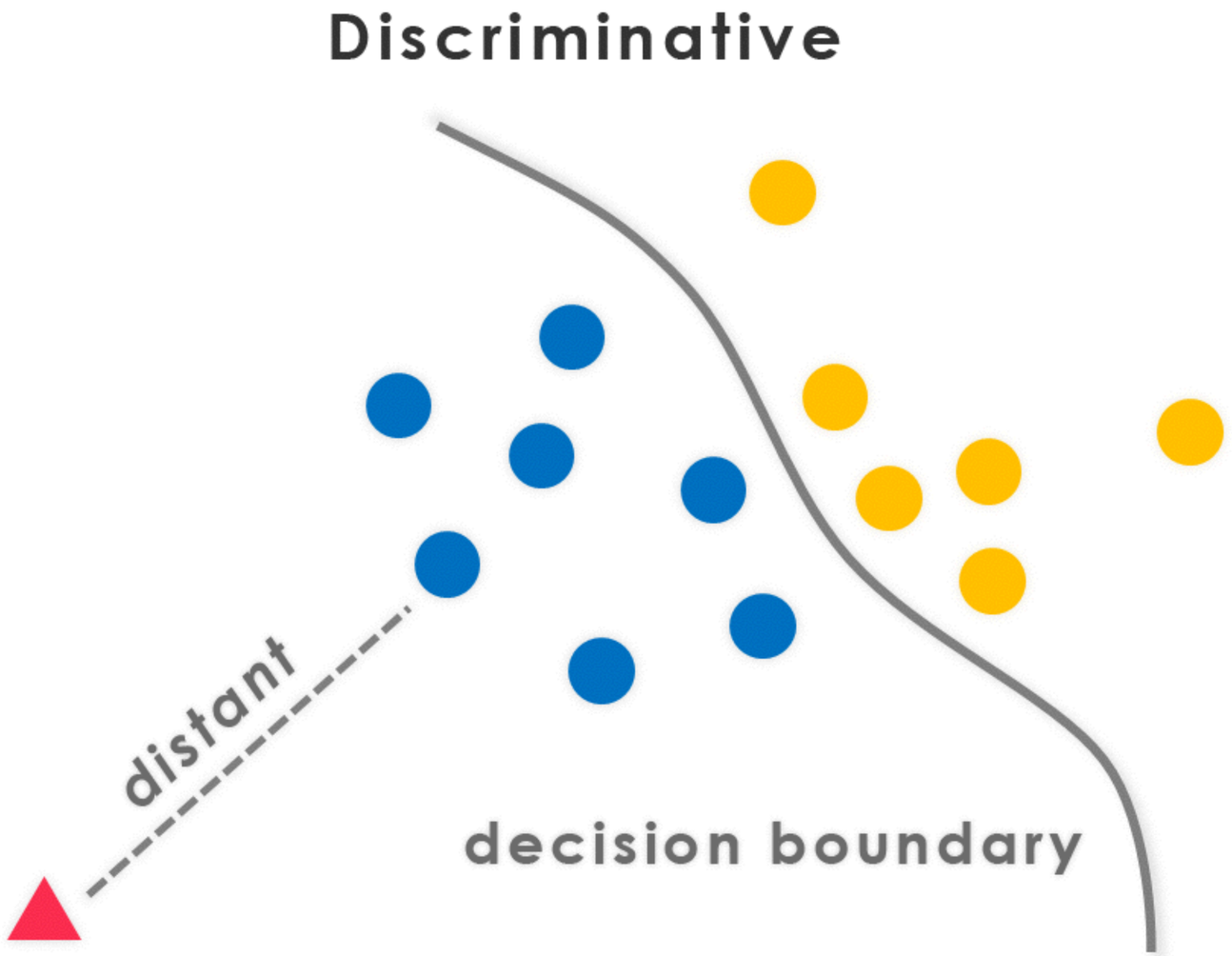
Why are networks easily fooled?

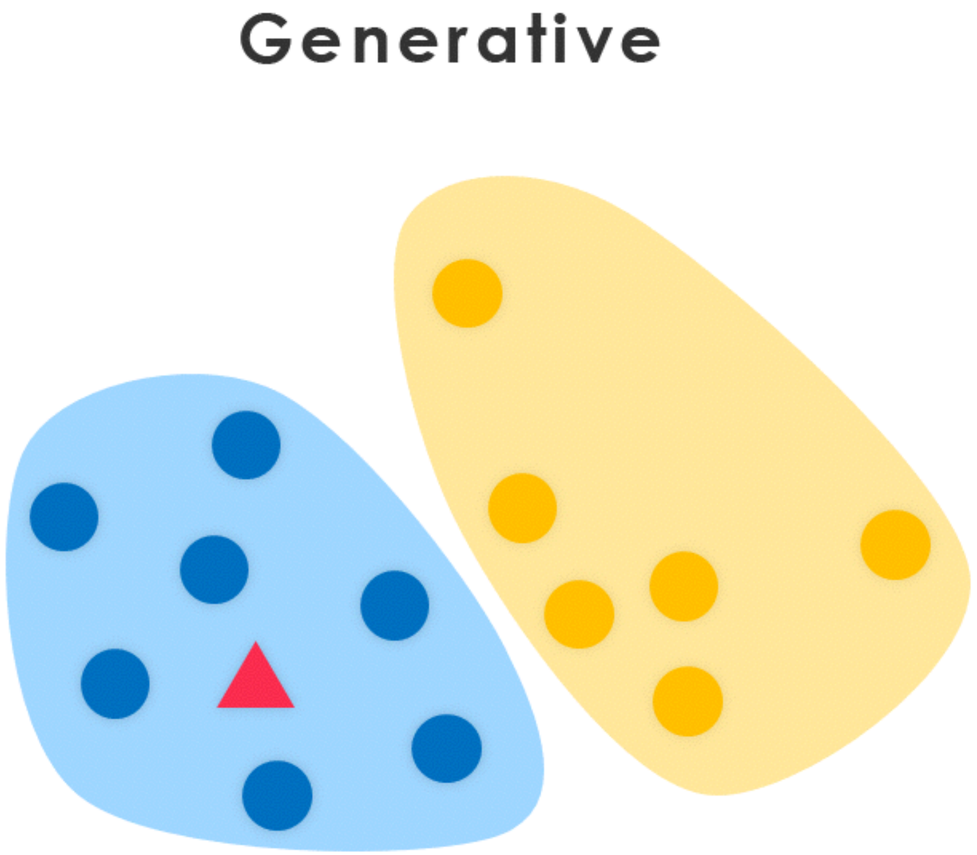




DNN Confidence: > 99.6 % for all

Hypothesis 1: DNNs do understand, test is bad



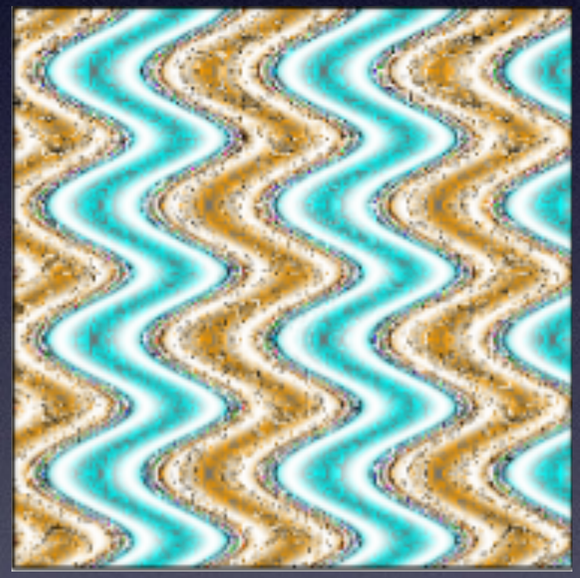


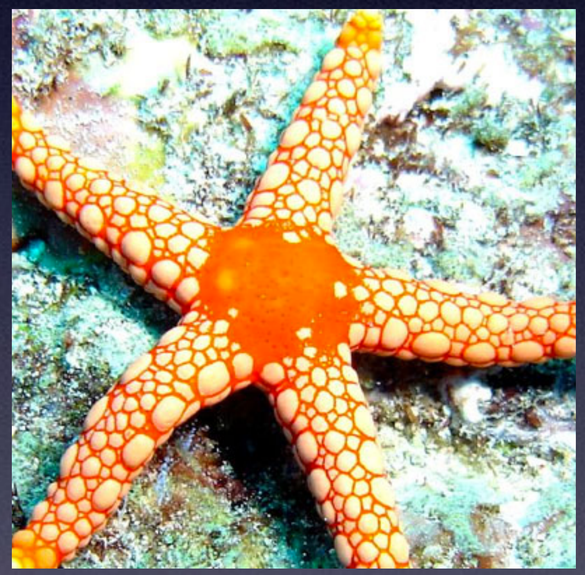
Prediction: With constraints to stay in the space of natural images, we WOULD get recognizable objects.

Hypothesis 2: Clever Hans

- Only learns distinguishing features

School Bus

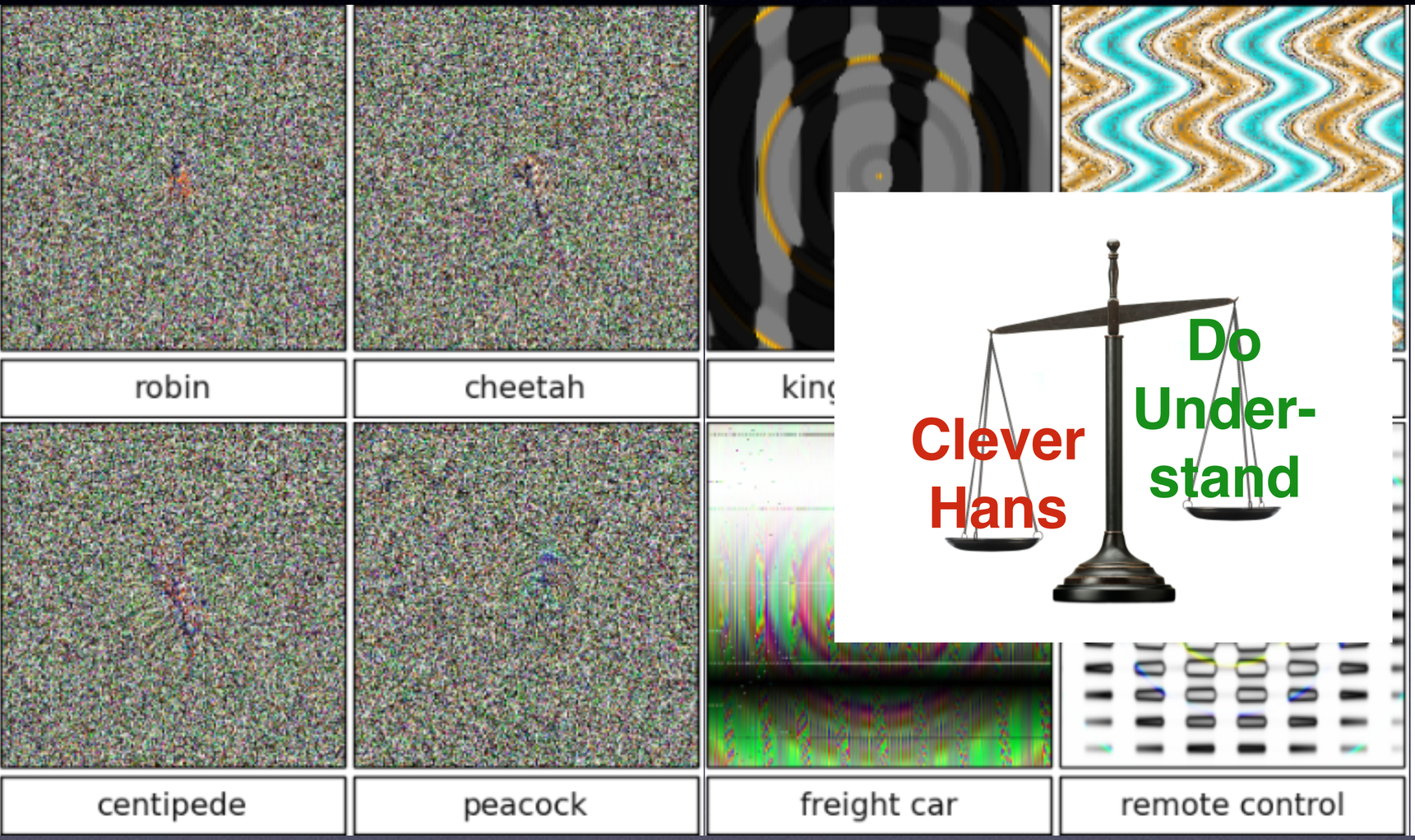


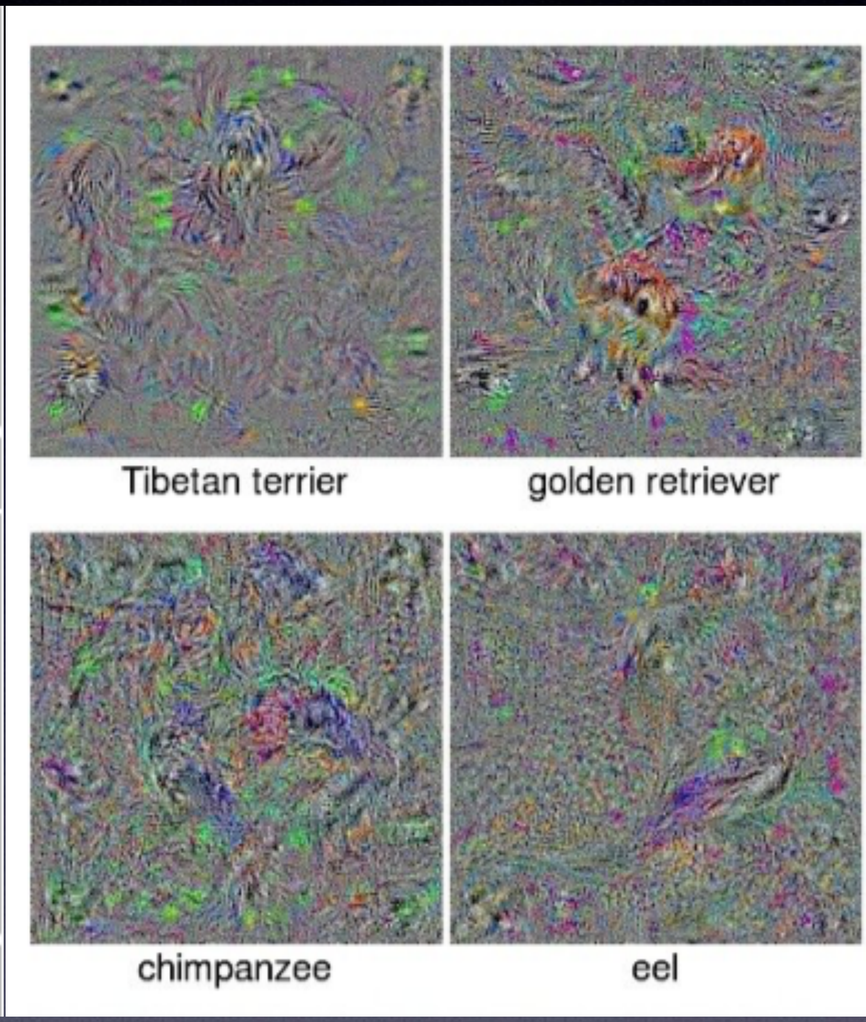


Starfish

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

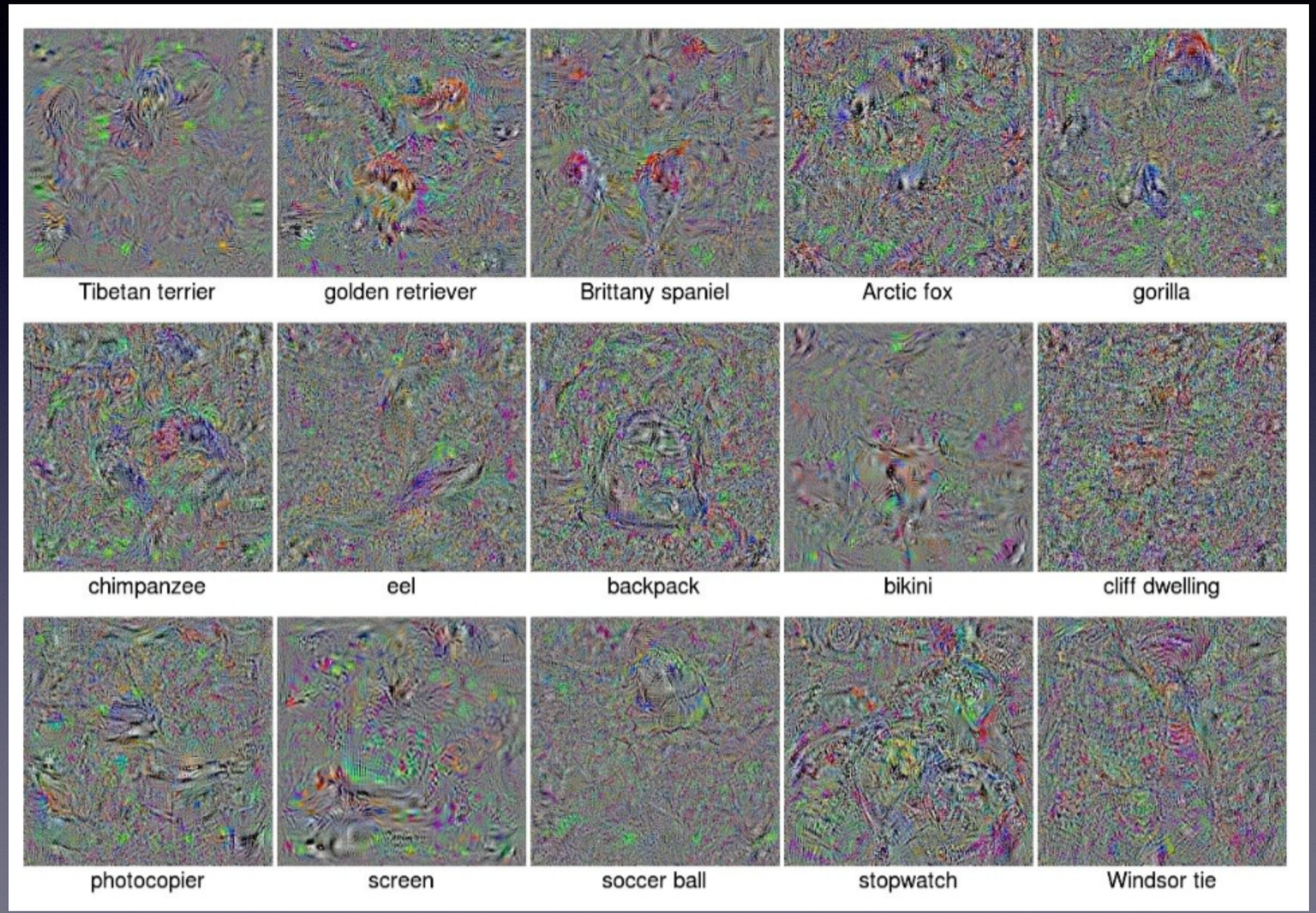
Clever Hans seems more likely…





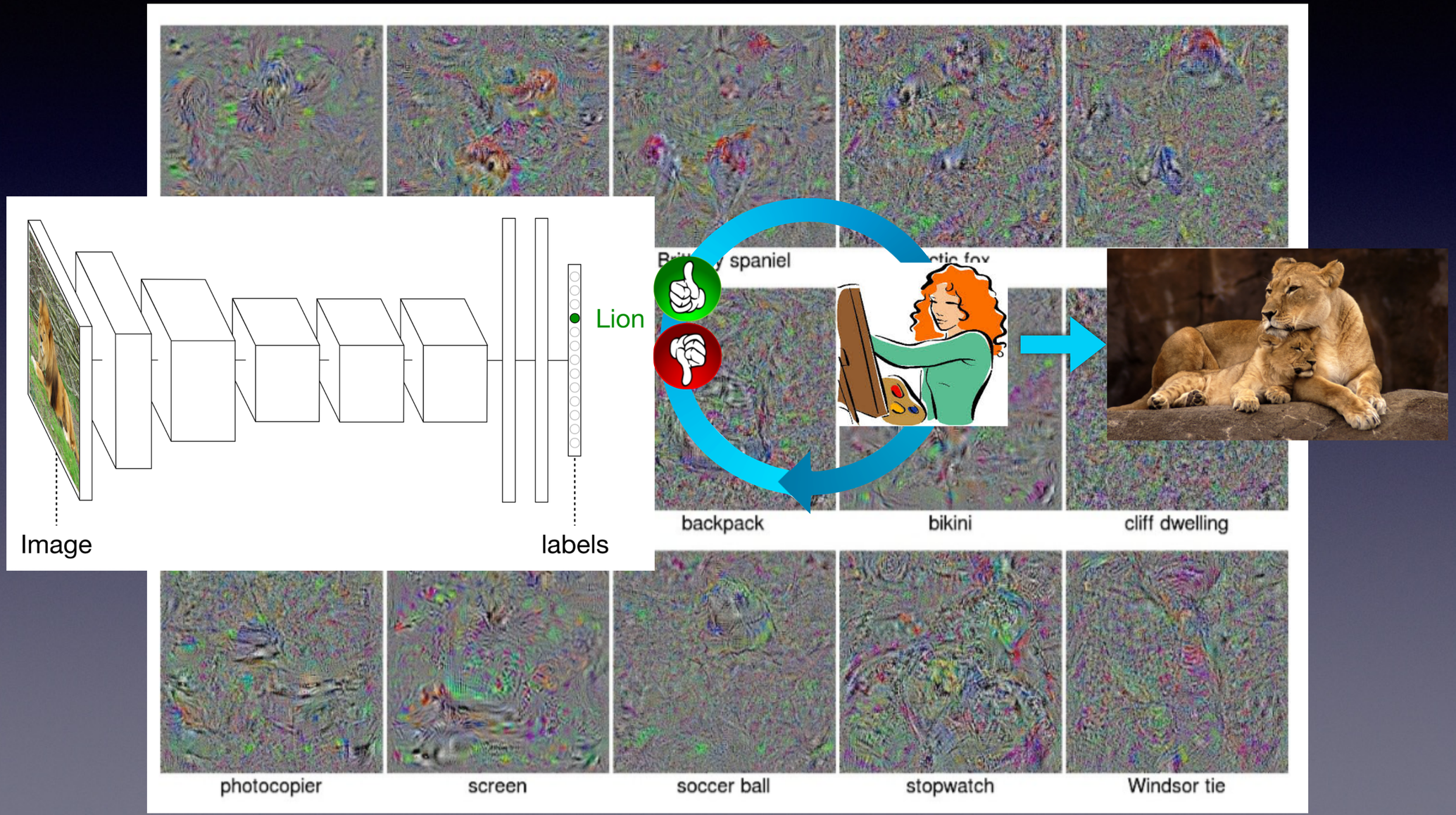
Deep Visualization Take 1

Nguyen, Yosinski, Clune, 2015, CVPR



Deep Visualization Take 1

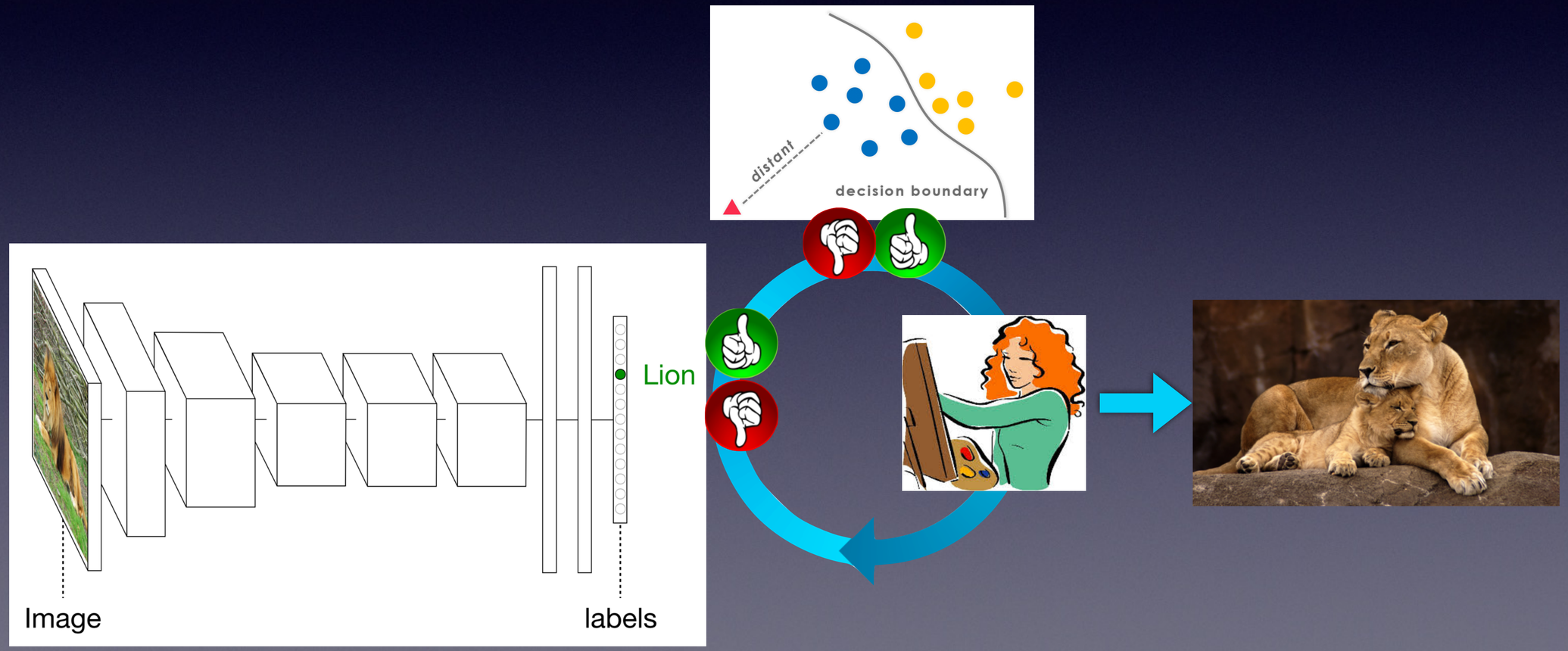
Nguyen, Yosinski, Clune, 2015, CVPR



Deep Visualization Take 2

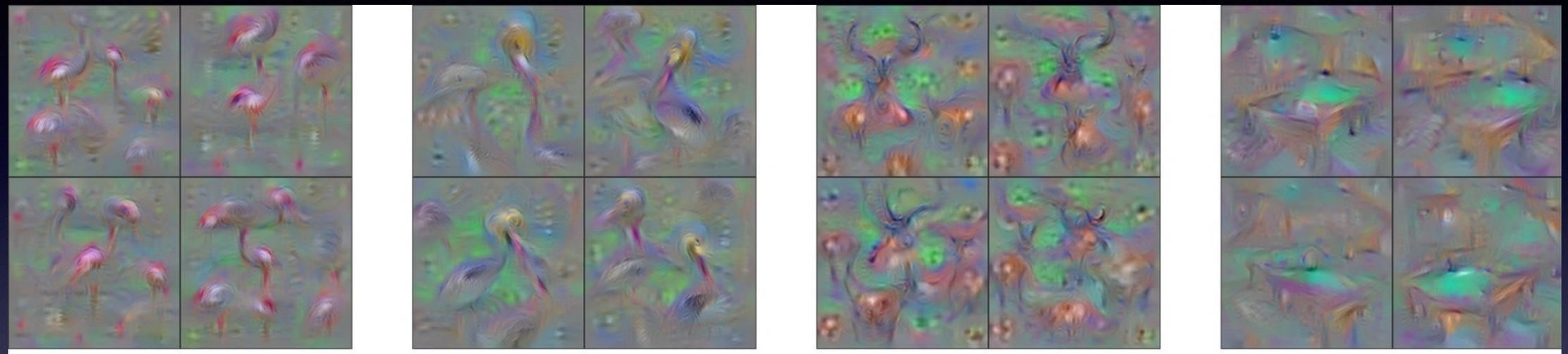
Yosinski, Clune, Nguyen, Lipson, 2015, ICML Deep Learning Workshop

Manually Engineered Natural Image Priors



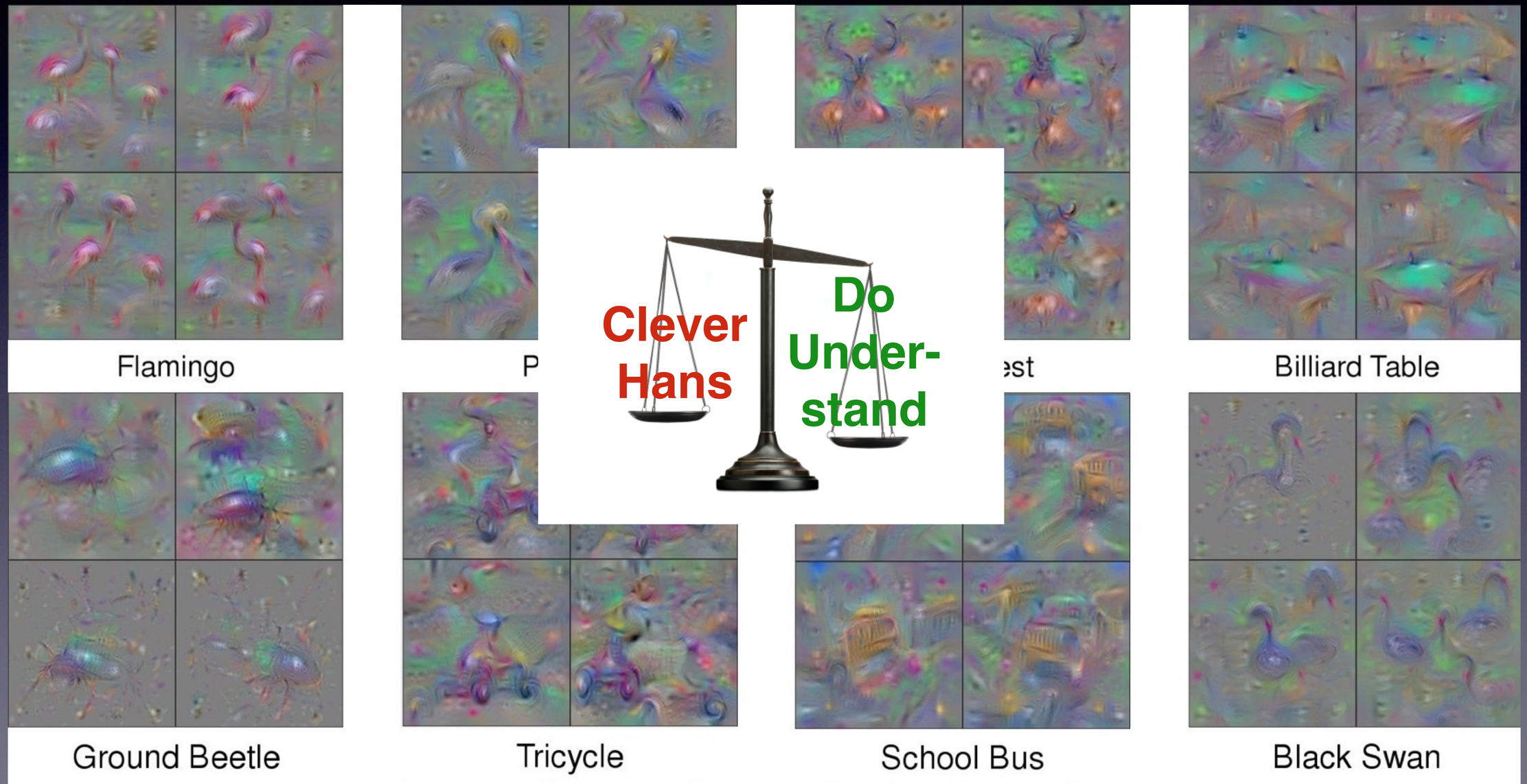
Deep Visualization Take 2

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



Deep Visualization Take 2

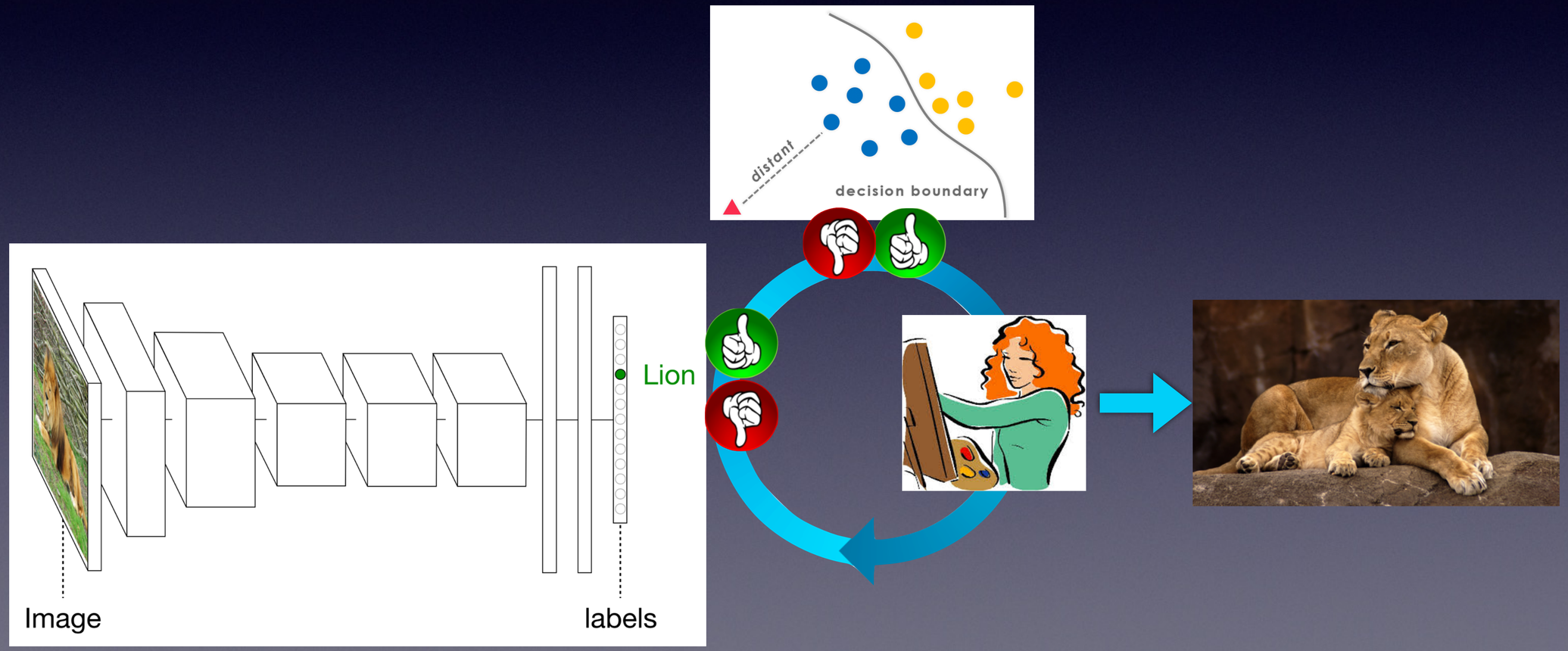
Yosinski, Clune, Nguyen, Lipson, 2015, ICML Deep Learning Workshop



Deep Visualization Take 3

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

Different Manually Engineered Natural Image Priors (+ mean seeds from data)

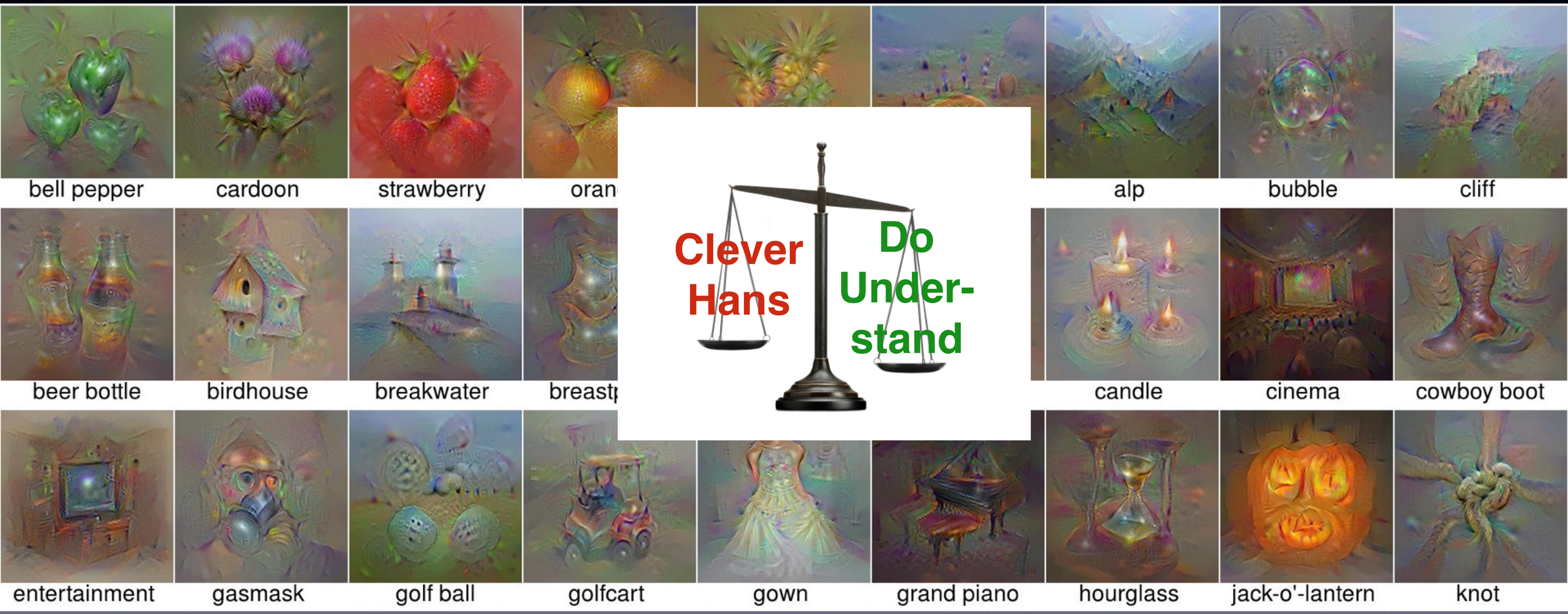


Deep Visualization Take 3

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

Deep Visualization Take 3

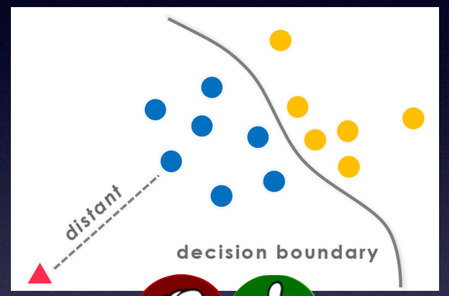
Multifaceted Feature Visualization. Nguyen, Yosinski, Clune 2016, ICML Workshop

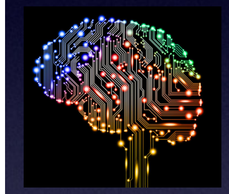


Deep Visualization Take 4

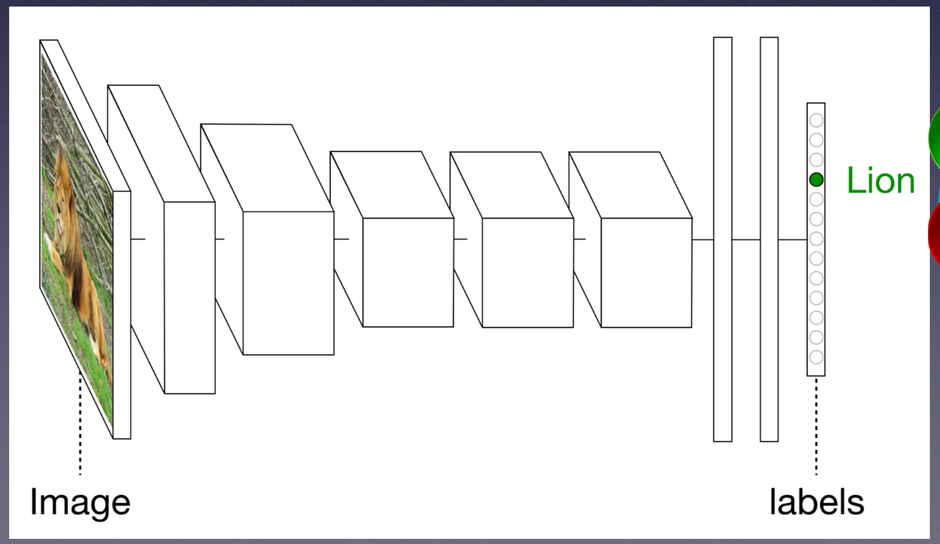
Nguyen, Dosovitskiy, Yosinski, Brox, Clune. 2016. Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. NIPS

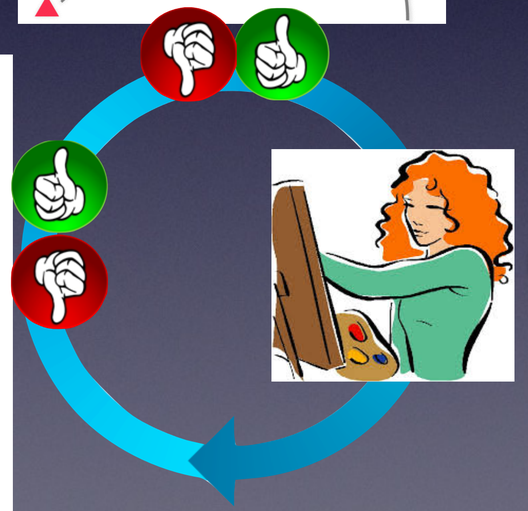
Learned Natural Image Priors





GANs, DeepSim, Upconv





Deep Visualization Take 4

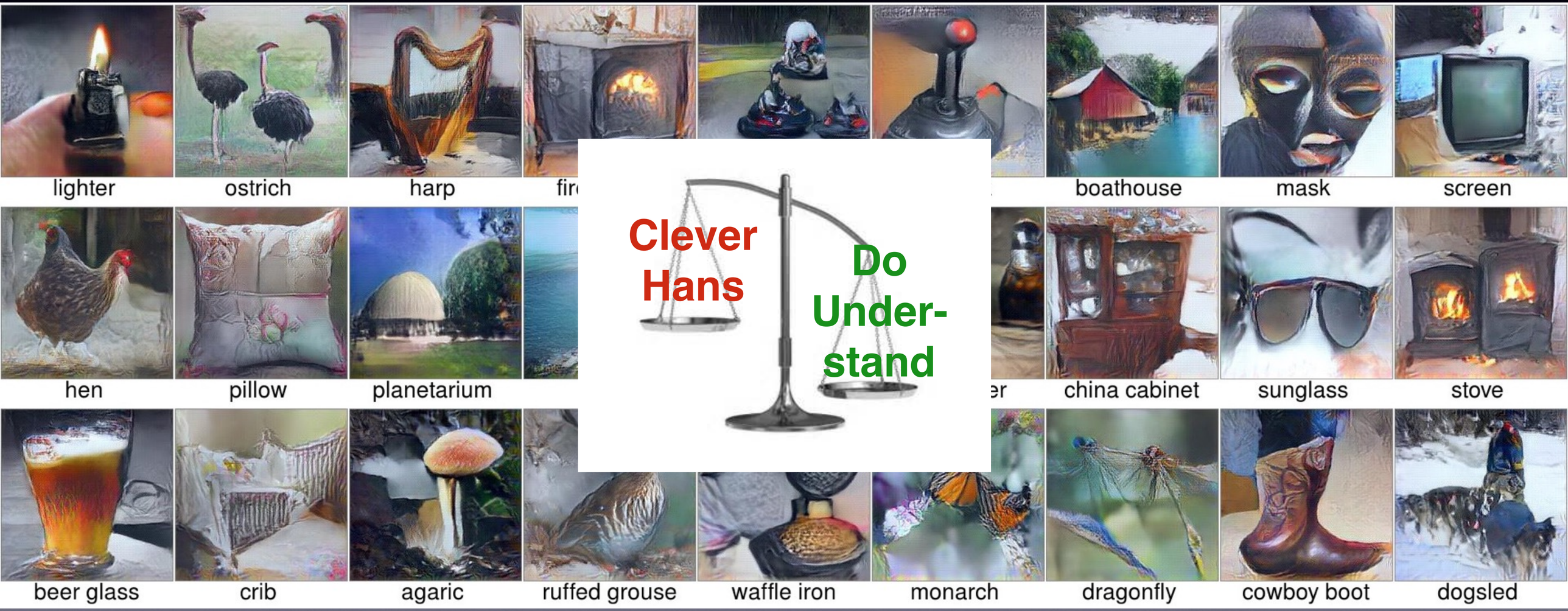
Nguyen, Dosovitskiy, Yosinski, Brox, Clune. 2016. NIPS

Deep Visualization Take 4

Nguyen, Dosovitskiy, Yosinski, Clune. Work in progress.

Deep Visualization Take 4

Nguyen, Dosovitskiy, Yosinski, Clune. Work in progress.



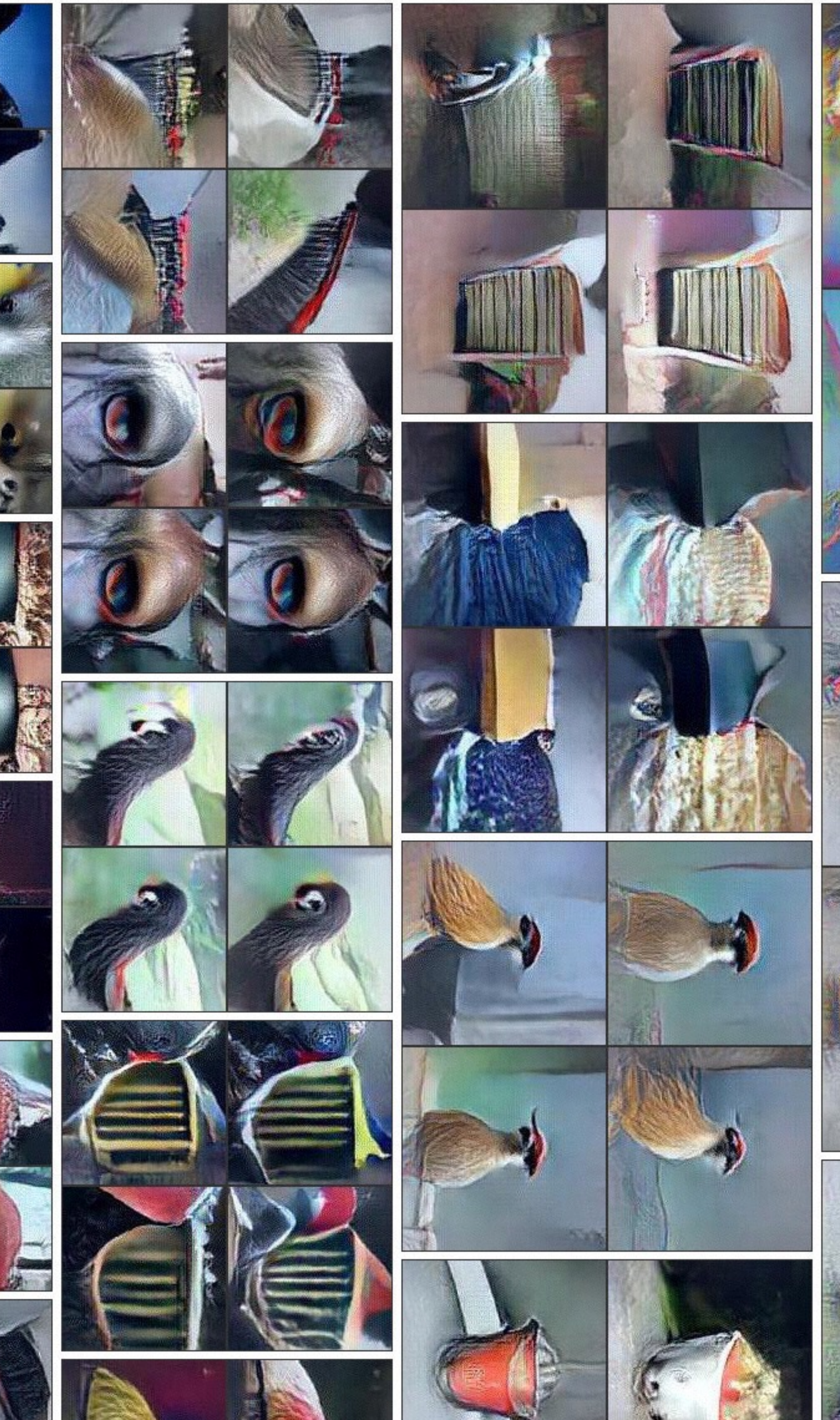

Layer 8
Restaurant
Ostrich
Bell gong
Beacon
Cradle
Layer 7
Layer 6
Layer 5
Layer 4
Layer 3
Layer 2
Layer 1

Layer 3

Layer 2

Layer 1

3 Layer 4 Layer 5



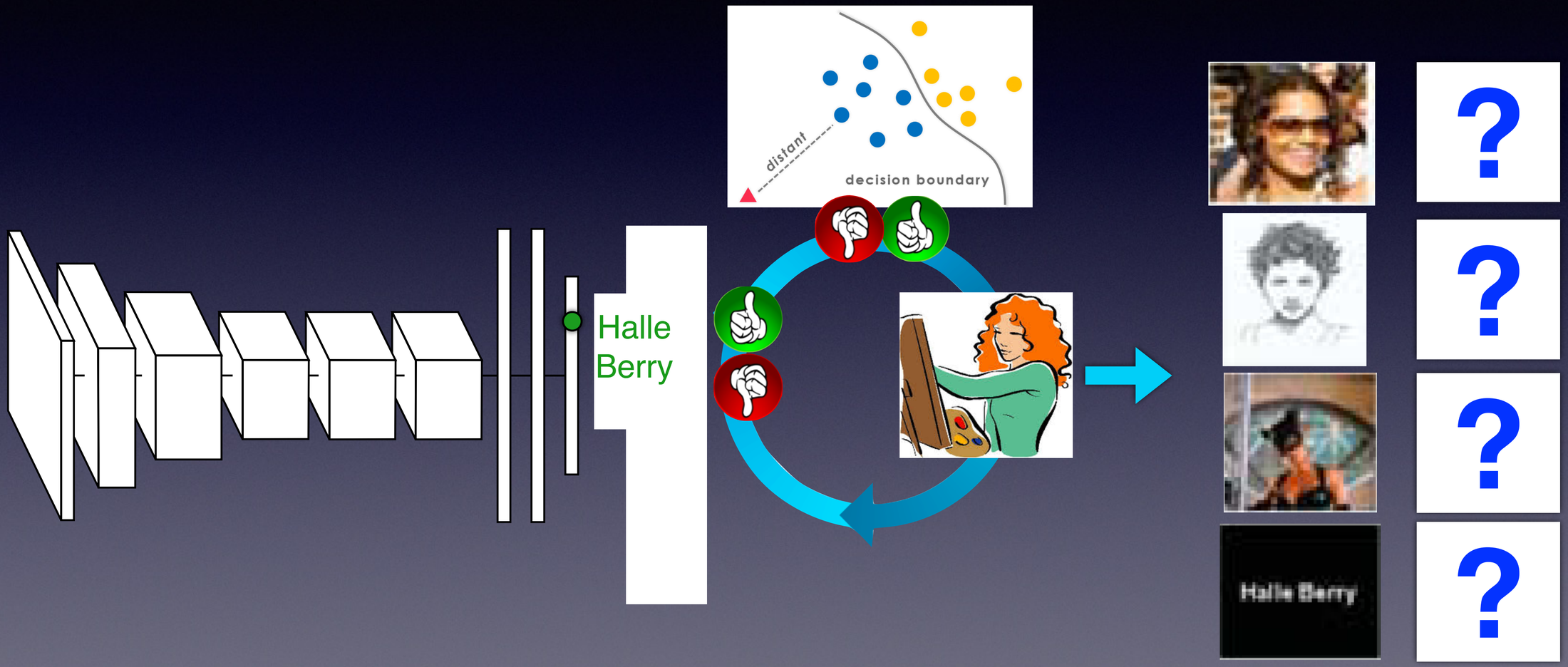
Layer 8

Restaurant

Ostrich

Multi-Faceted Feature Visualization

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




4
9
6
7
2
10
1
5
8
3


1
2
3
4
5
6
7
8
9
10
1
2
3
4
5
6
7
8
9
10


Star


PASADENA
Capitan


REGENT THEATRE
THEATRE

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

Are DNN representations local or distributed?

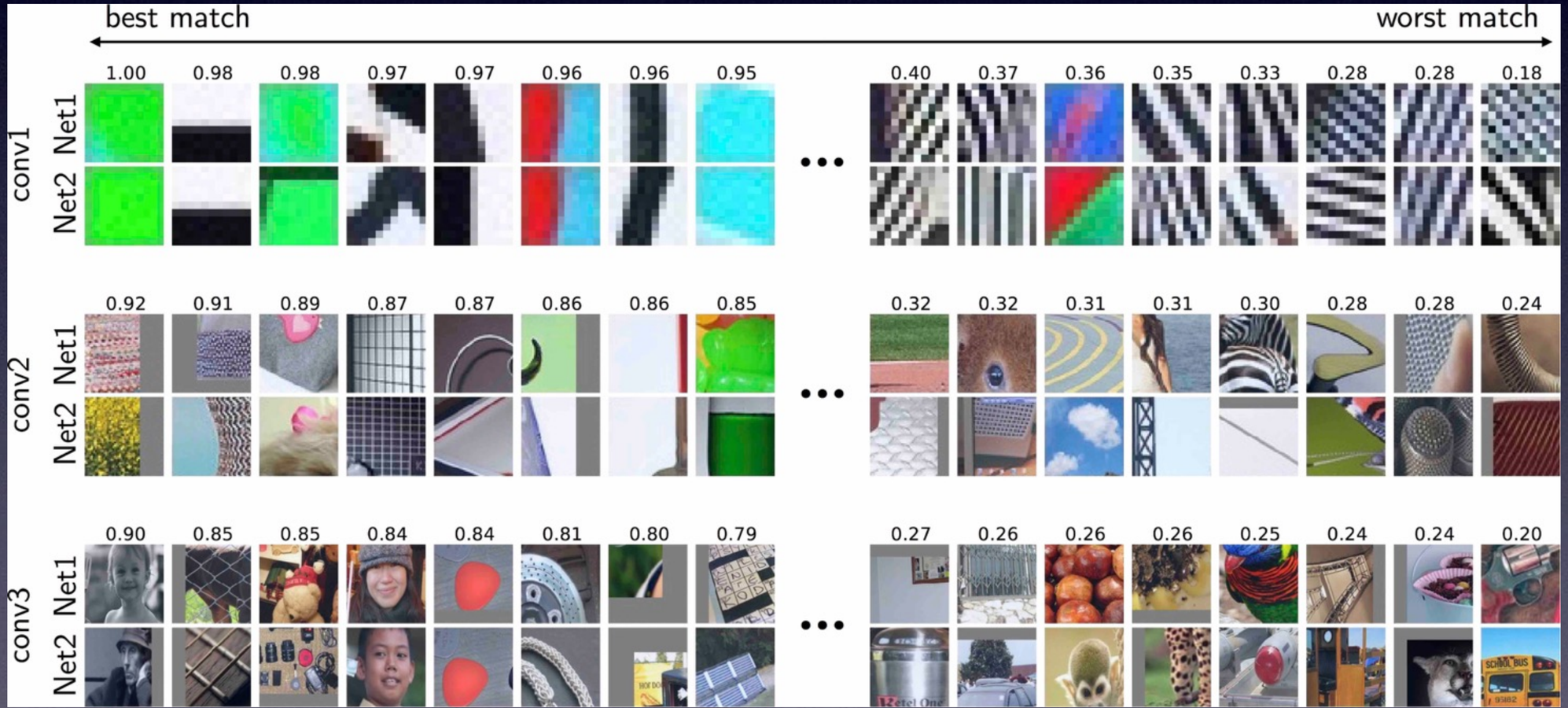
Li Y, Yosinski J, Clune J, Lipson H, Hopcroft J (2016) Convergent Learning: Do different neural networks learn the same representations? International Conference on Learning Representations (ICLR '16).

- Answer: “It’s complicated.”

Do different DNNs learn the same features?

Li Y, Yosinski J, Clune J, Lipson H, Hopcroft J (2016) Convergent Learning: Do different neural networks learn the same representations? International Conference on Learning Representations (ICLR '16).

Answer: Yes for most active features. No for least active ones.



Conclusions

- Initial “fooling” work suggested Deep Learning is a Clever Hans
- Work since reveals DNNs learn a surprising amount about the concepts they classify
  - Understand an entire object & its context
  - Understand the multifaceted nature of concepts

Conclusions

- Why fooling?
  - Asking for maximum confidence
  - Space of images is vast
  - We are not training for fool-proof images

Future Work Potential

- Demonstrate all of this in other modes
  - Speech recognition
  - Music classification
  - Including learning natural priors to prevent fooling
- Investigate with unsupervised learning
  - and reinforcement learning
- Can we generate such “preferred stimuli” for biological brains?

Thanks

Anh Nguyen

Jason Yosinski

Alexey Dosovitskiy