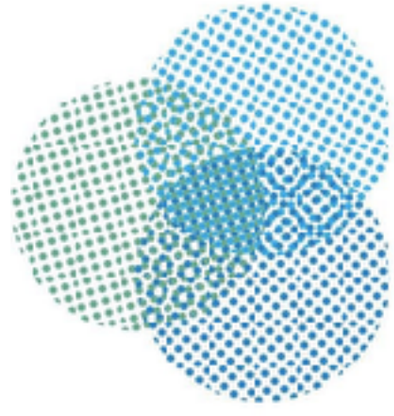
Making a Science from the Computer Vision Zoo

Xavier Boix

xboix@mit.edu



CENTER FOR
Brains
Minds+
Machines

Computer Vision Zoo

Brains

SLAM

Super Pixels

Human

Primates

ResNet

HMAX

Zebra Fish

CRFs

VGG

Fly

Rat

AlexNet

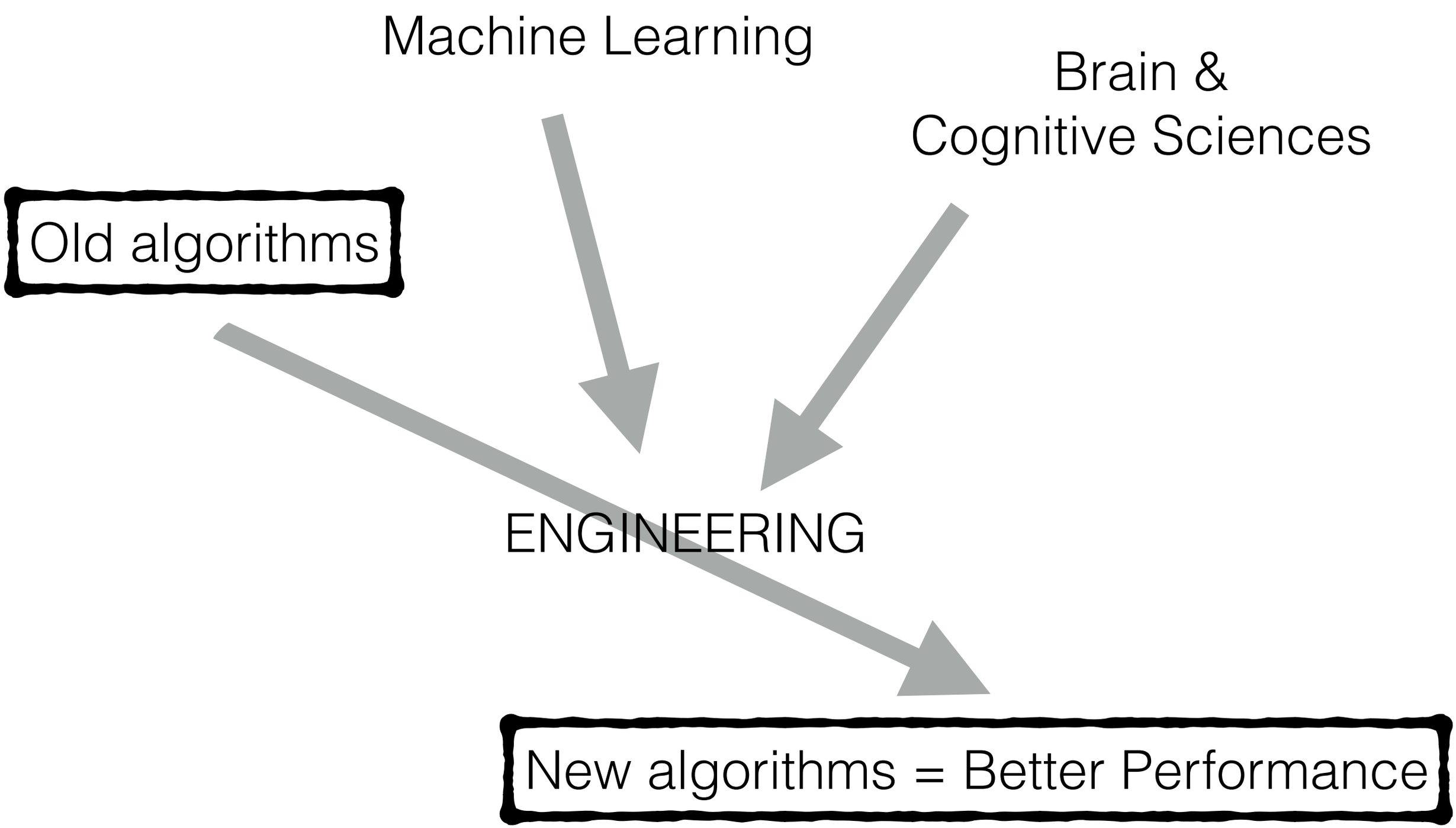
The Computer Vision Zoo

Engineering of algorithms

New algorithms = Better Performance

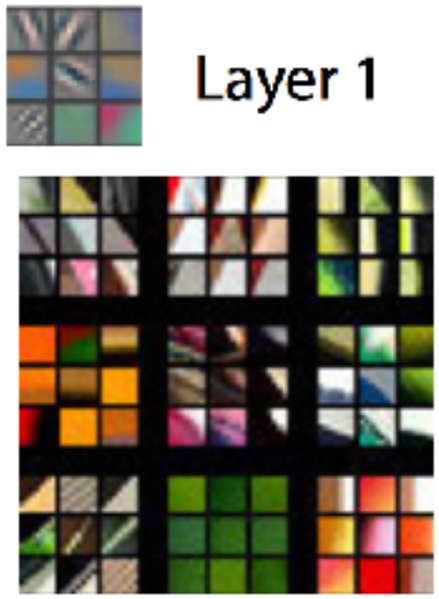
("Darwinian evolution")

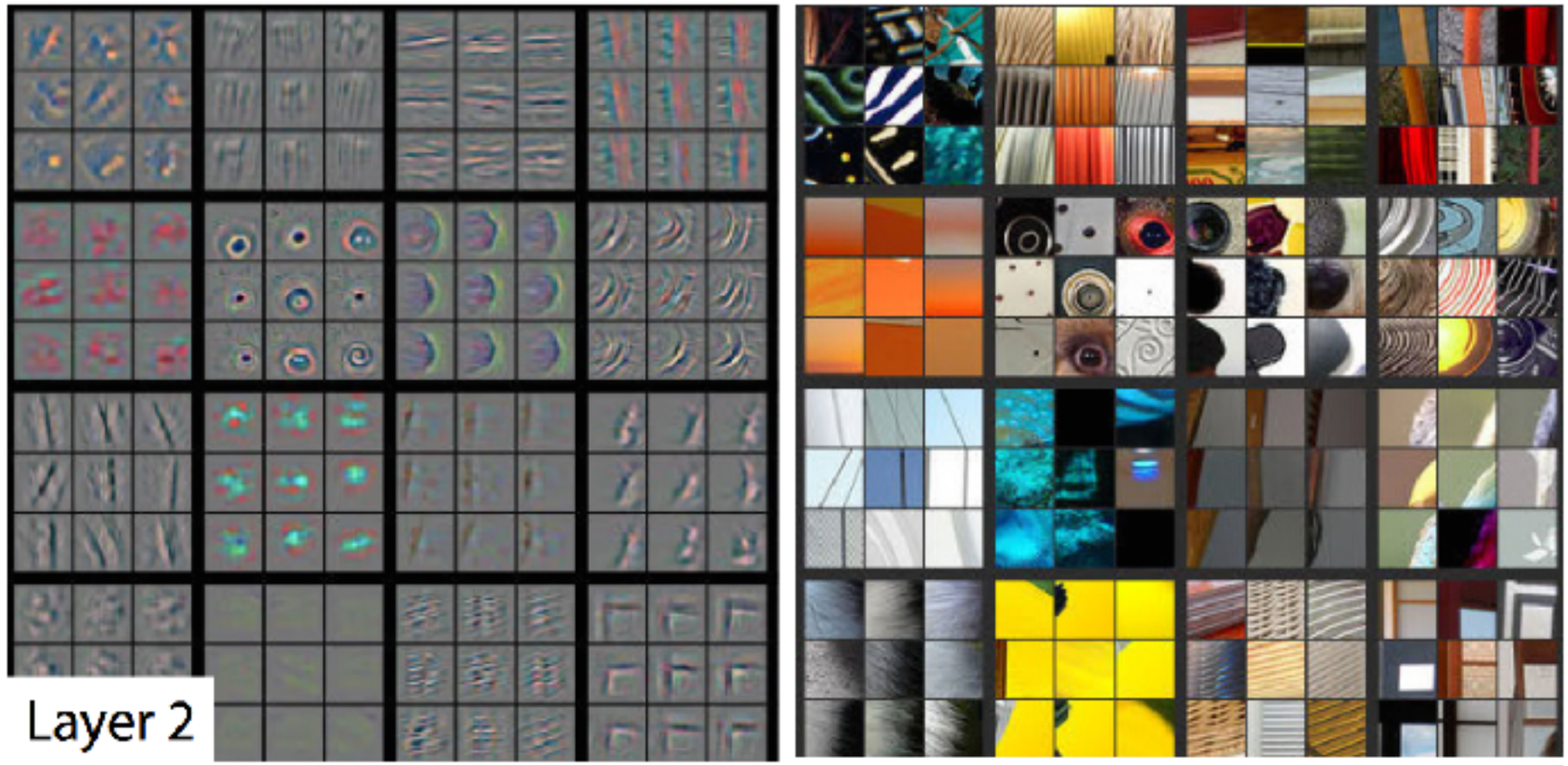
Evolution of the Computer Vision Zoo

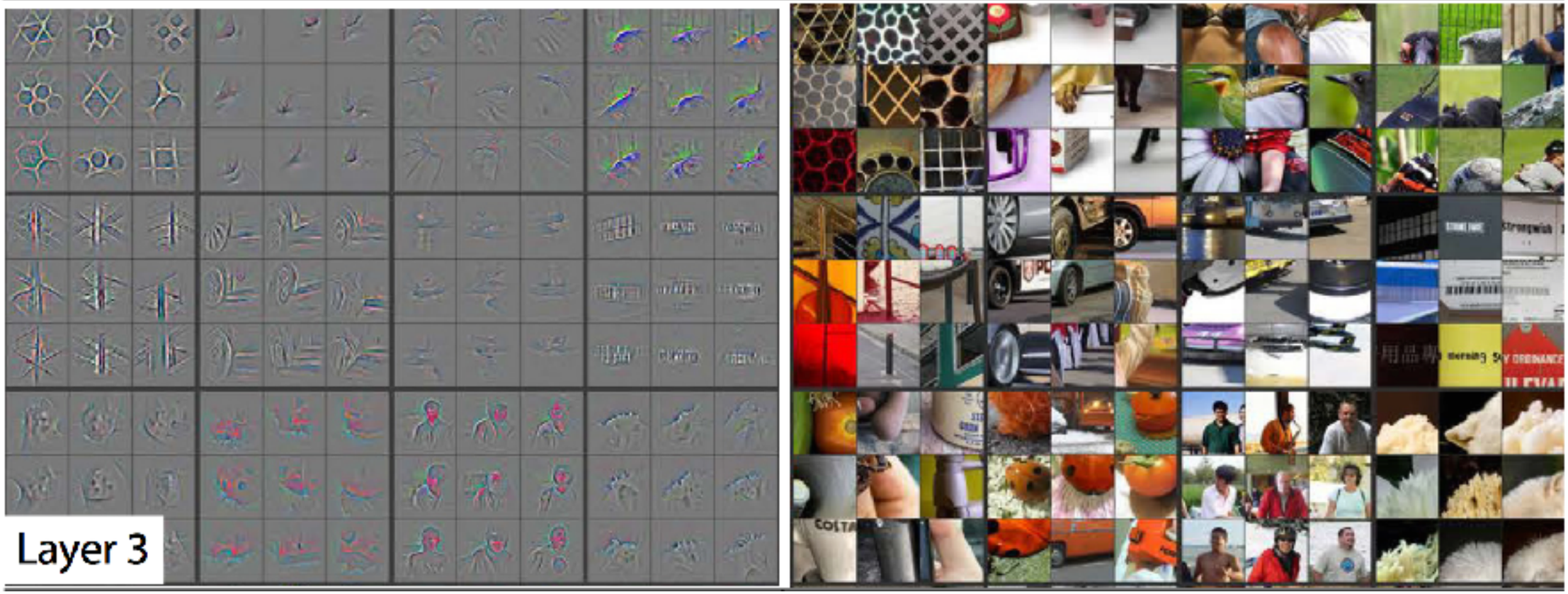


Are we going to understand intelligence by building AIs that pass the Turing test

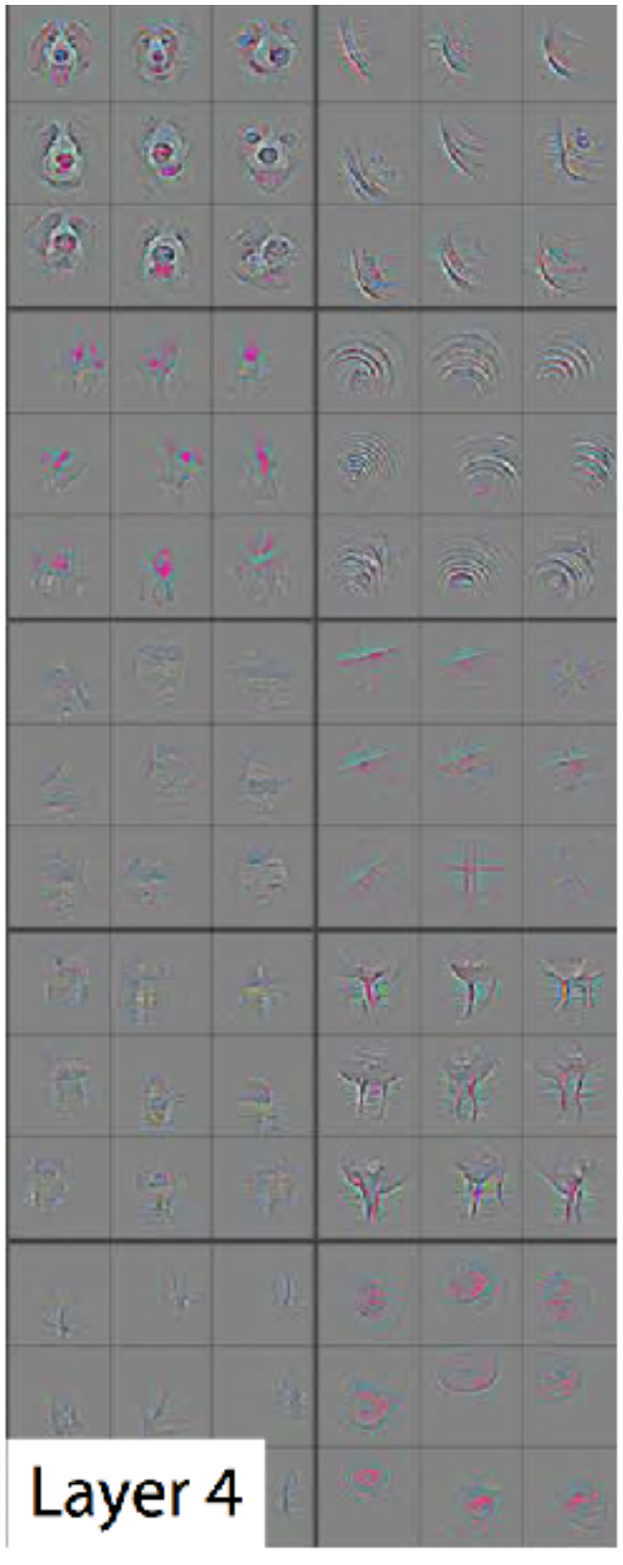
?

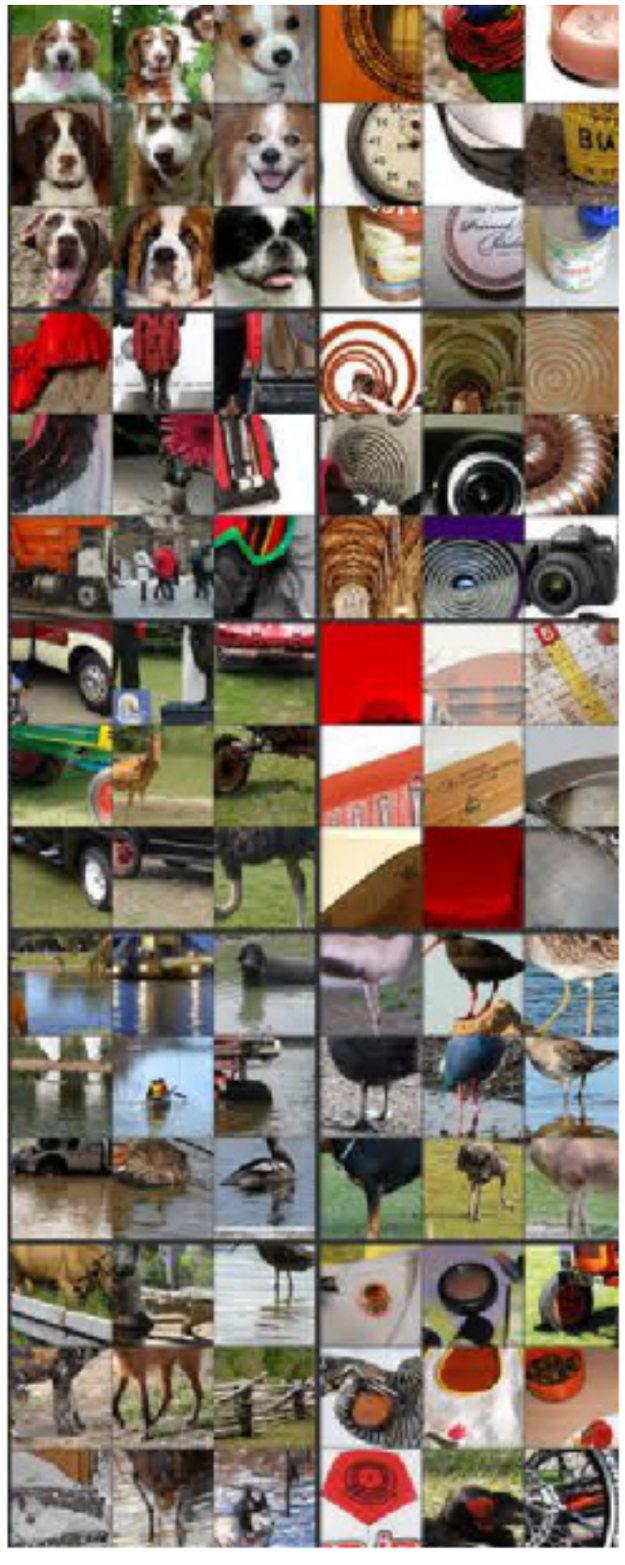


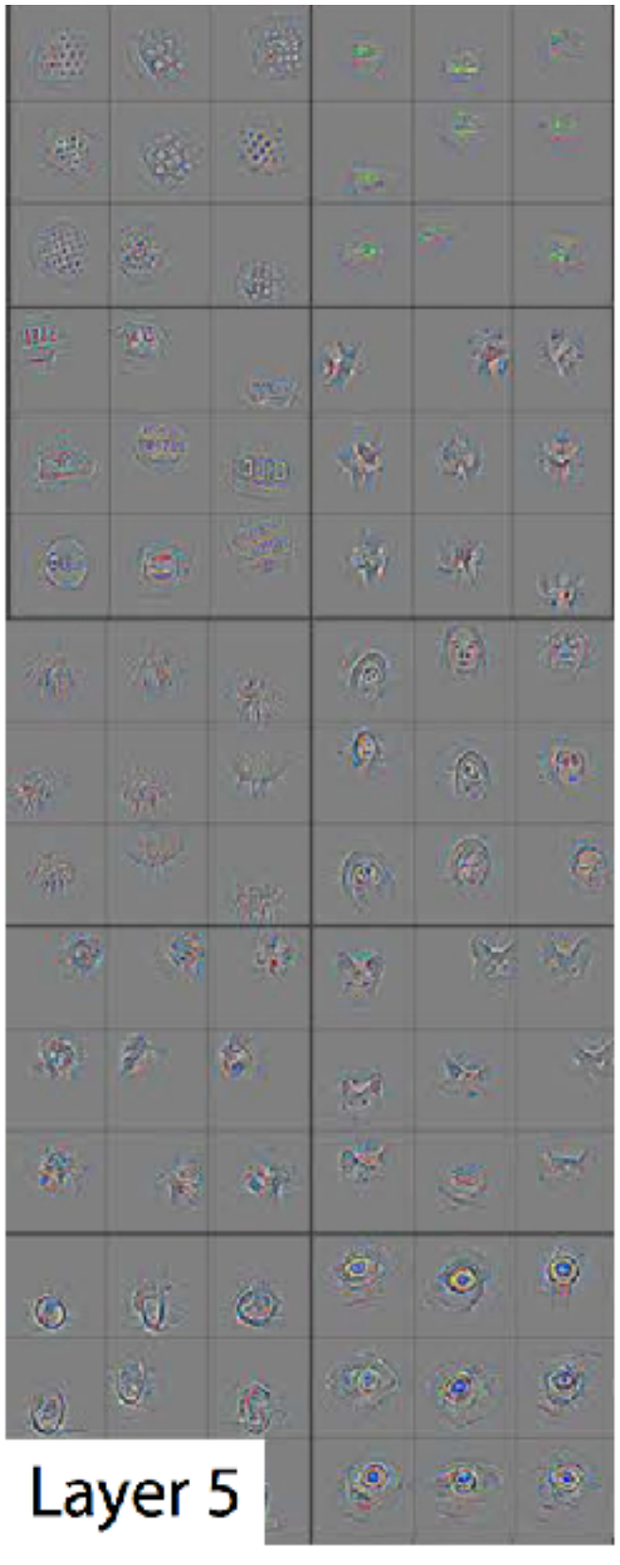


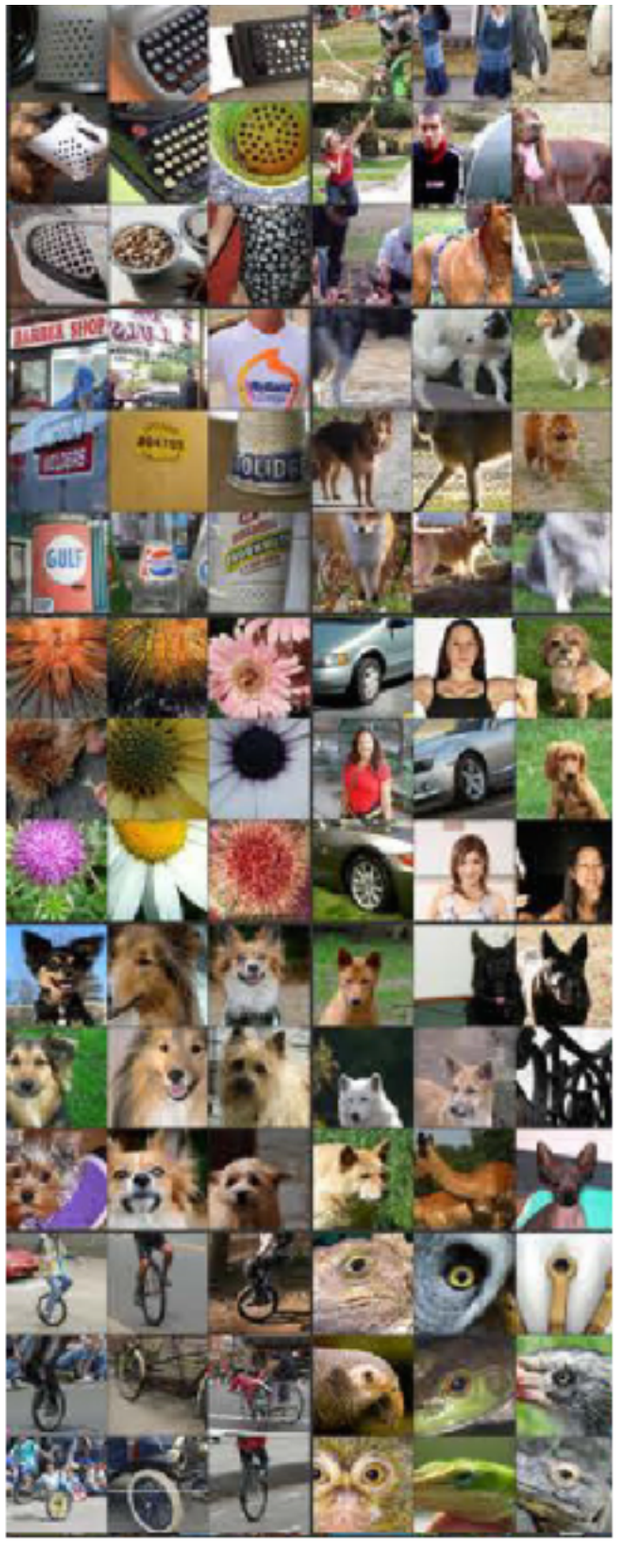


Zeiler and Fergus 13

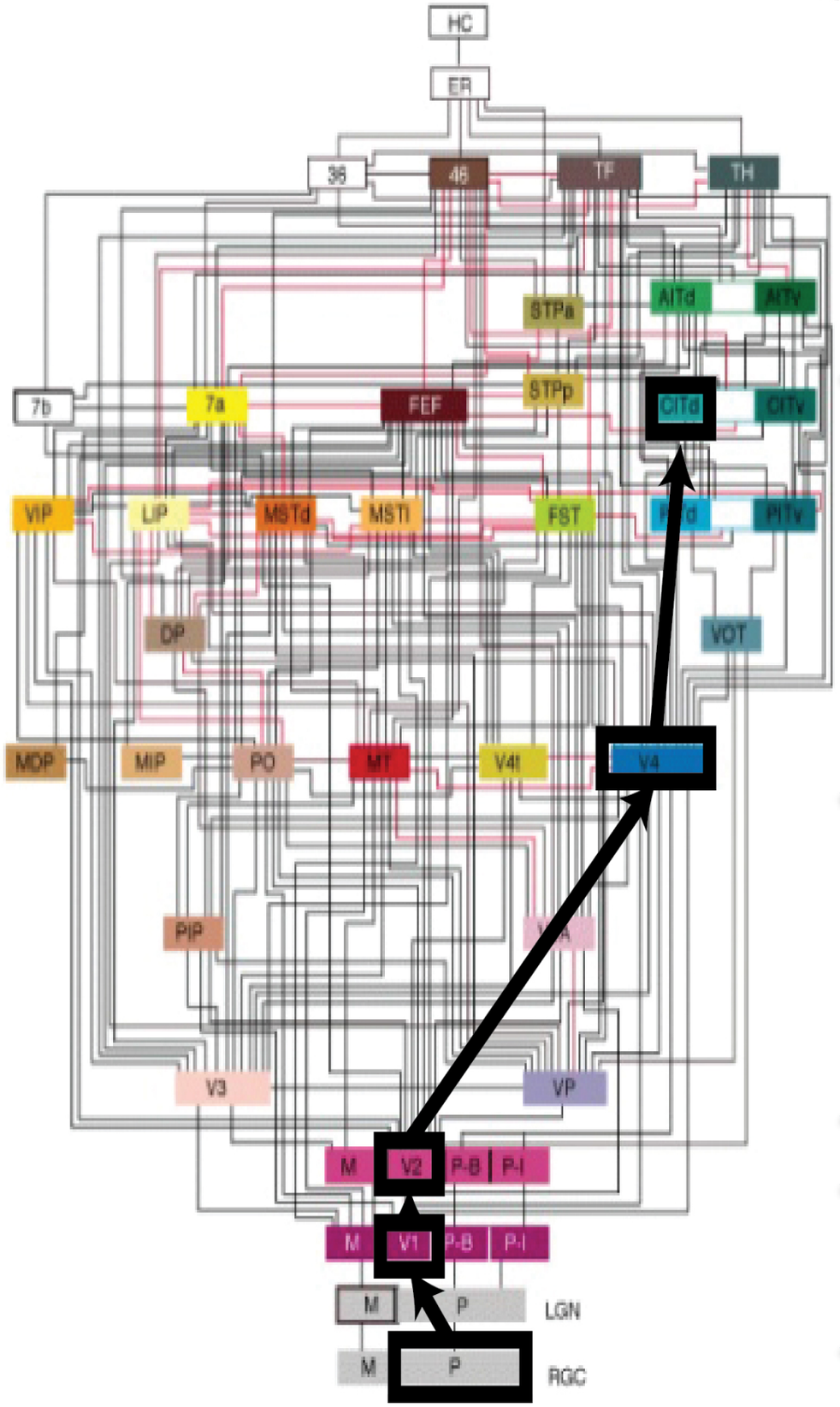


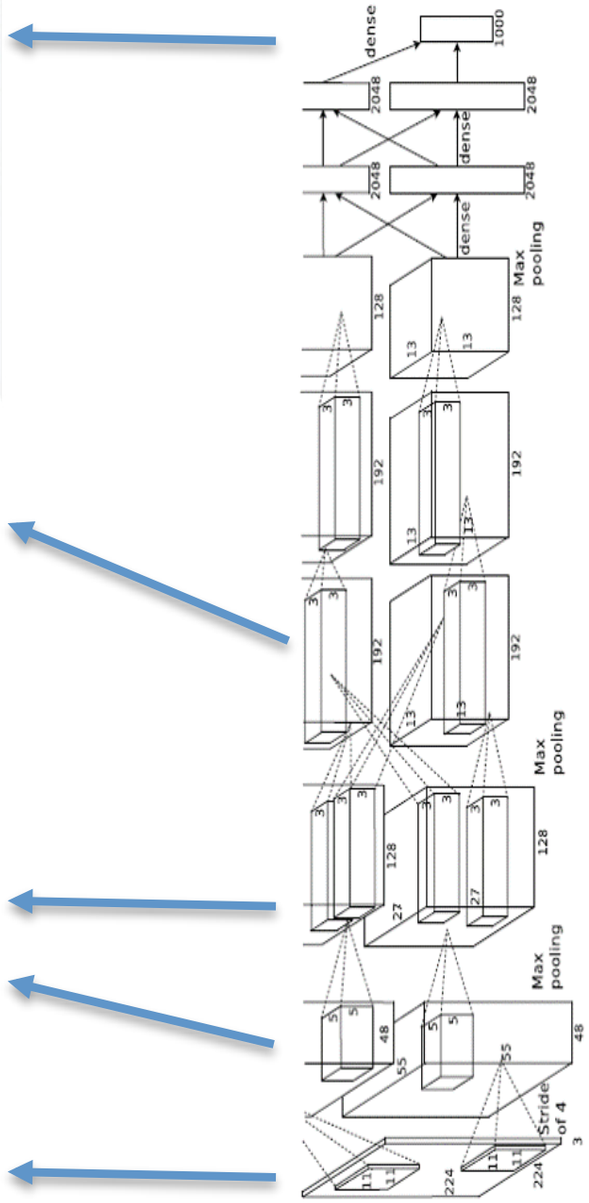






Zeiler and Fergus 13





Felleman and Van Essen 1991

Krizhevsky et al 2012

Marr Levels of Understanding

1) Computational

Computer Vision
(Build better algorithms)

2) Algorithmic

3) Implementation

Science of Computer Vision

Make a science from the Zoo

computational principles among algorithms

- Develop testable hypothesis

- Test and challenge the hypothesis

At the Computational Level

Develop scientific hypothesis…

-Meta-algorithms, unification of algorithms

-Visualization analysis

-Mathematical theories

etc.

…to predict properties of the algorithms before testing them

Computer Vision Zoo

Brains

SLAM

Super Pixels

Human

Primates

ResNet

HMAX

Models

Zebra Fish

CRFs

VGG

Rat

Fly

AlexNet

Computational hypothesis

1) Computational

Same principles should explain
artificial and biological intelligence

(eg. Why the brain is how it is?
Why an algorithm works better than others?)

---

2) Algorithmic

Computer Vision
(Build better algorithms)

Modelling the Brain
(How does the brain work?)

3) Implementation

Machine Learning

Brain &
Cognitive Sciences

Old algorithms

ENGINEERING

**Science of
Computer Vision**

New algorithms = Better Performance

3 Snapshots

1) Failures of DNNs

2) Beyond Object Recognition

3) DNNs to explain the Brain

Minimal Images in Deep Neural Networks

with

Sanjana Srivastava and Guy Ben-Yosef

Adversarial examples

Original | Adversarial | Adversarial (exaggerated to show perturbation)

(Ullman et al., 2016)

(Ullman et al., 2016)

(Ullman et al., 2016)

(Ullman et al., 2016)

Minimal images in humans

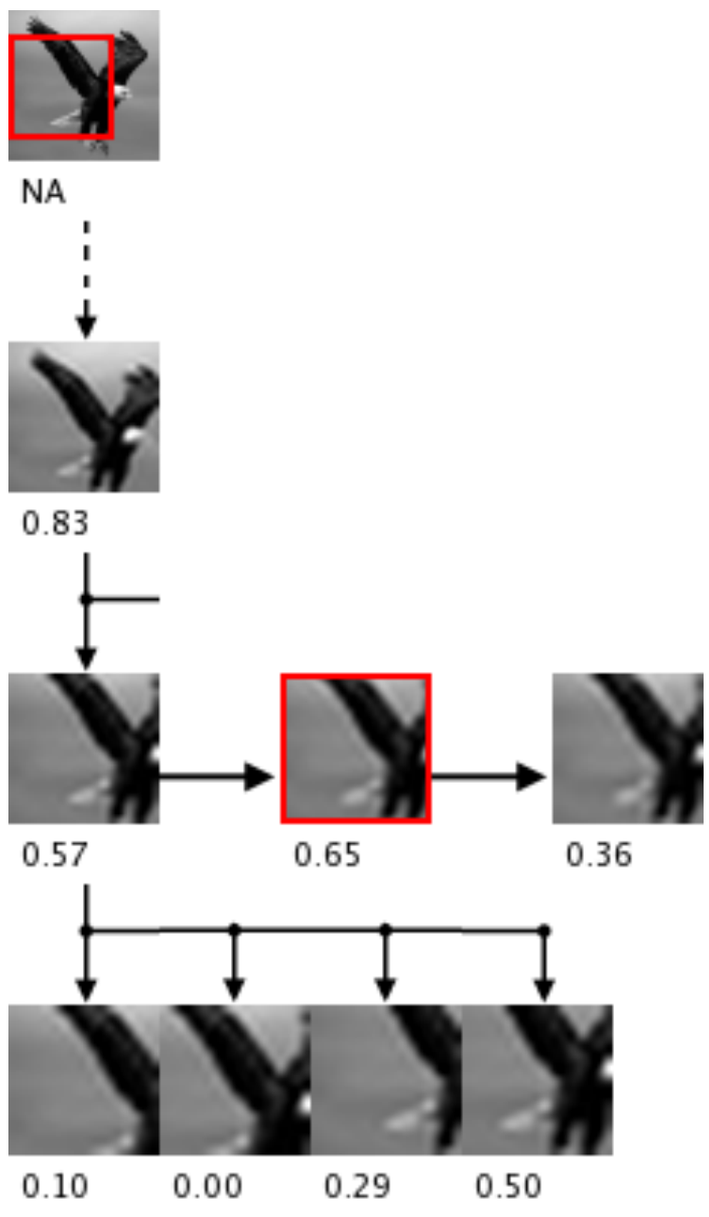
- Smallest region of an image that is still recognizable to humans (Ullman et al., 2016)

- DNNs unable to recognise human minimal images (Ullman et al., 2016; Ben-Yosef et al., 2018)

Unrecognizable to humans

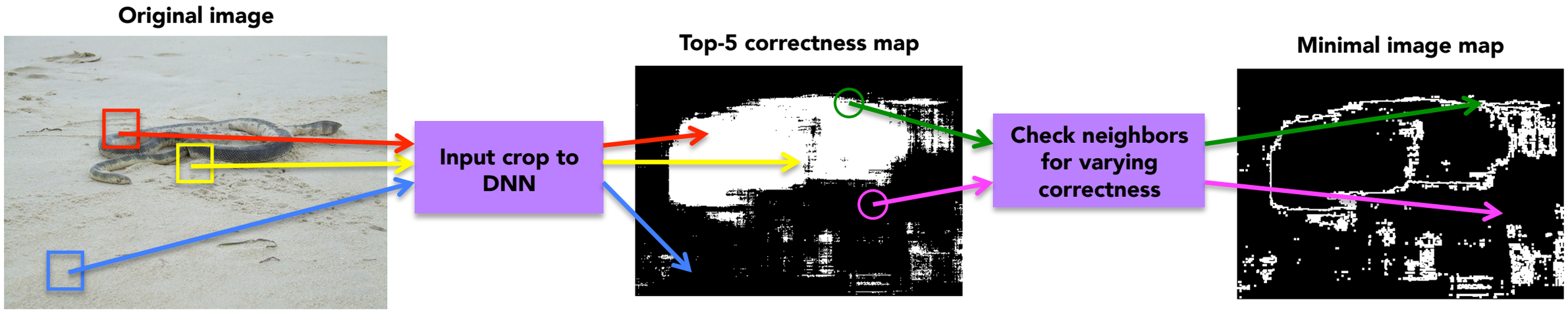
Recognizable to humans

Extracting Human Minimal Images

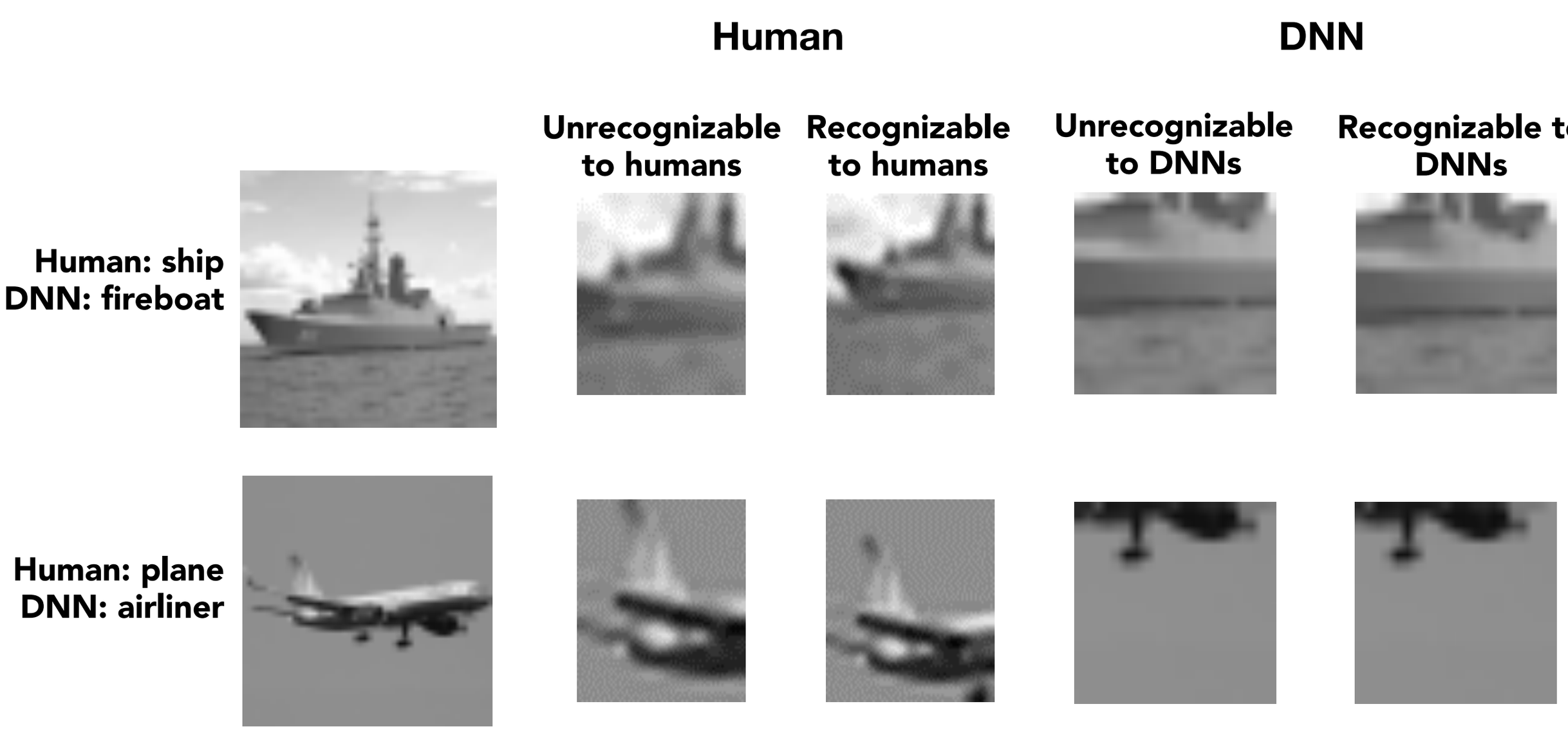


(Ullman et al., 2016)

Extracting DNNs Minimal Images



Do DNNs have their own set of minimal images?



Fragile Recognition Image

FRI: *Image region that is correctly categorised but a small shift of the region location or shrink of the region scale produces misclassification.*

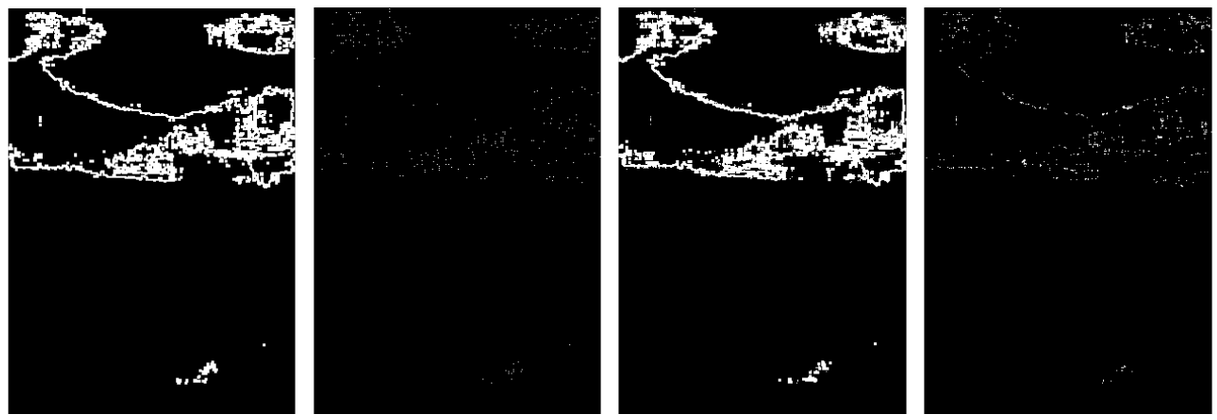
Minimal Images $\subset$ FRIs

Different ways to evaluate fragile object recognition

Original image

Label: coffee

Loose shift | **Strict shift** | **Loose scale** | **Strict scale**

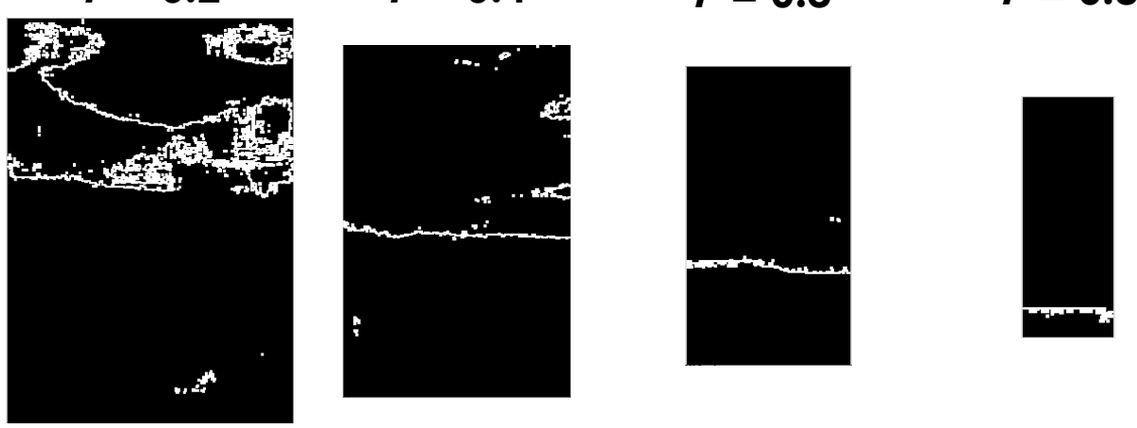


Loose vs. strict

Scale vs. shift

***P* = crop side length / image lesser dimension**

***P* = 0.2** | ***P* = 0.4** | ***P* = 0.6** | ***P* = 0.8**



FRIs occur at boundaries of useful regions

Loose FRI: there exists a slight change that will cause failure

Strict FRI: any slight change will cause a failure

Experimental Setup

VGG, ResNet, Inception: time to process 1 image in 8 K80 GPU is ~5 minutes

500 images randomly chosen from the validation set

DNNs fail on smaller changes

ImageNet class: Electric guitar
Human option: Instrument

ResNet score: ~0.8
Human success
rate: 65%

ResNet score: ~0.4
Human success
rate: 55%

ImageNet class: Broccoli
Human option: Vegetable

ResNet score: ~0.9
Human success
rate: 60%

ResNet score: ~0.5
Human success
rate: 80%

ImageNet class: Sturgeon
Human option: Fish

ResNet score: ~0.7
Human success
rate: 5%

ResNet score: ~0.3
Human success
rate:: 15%

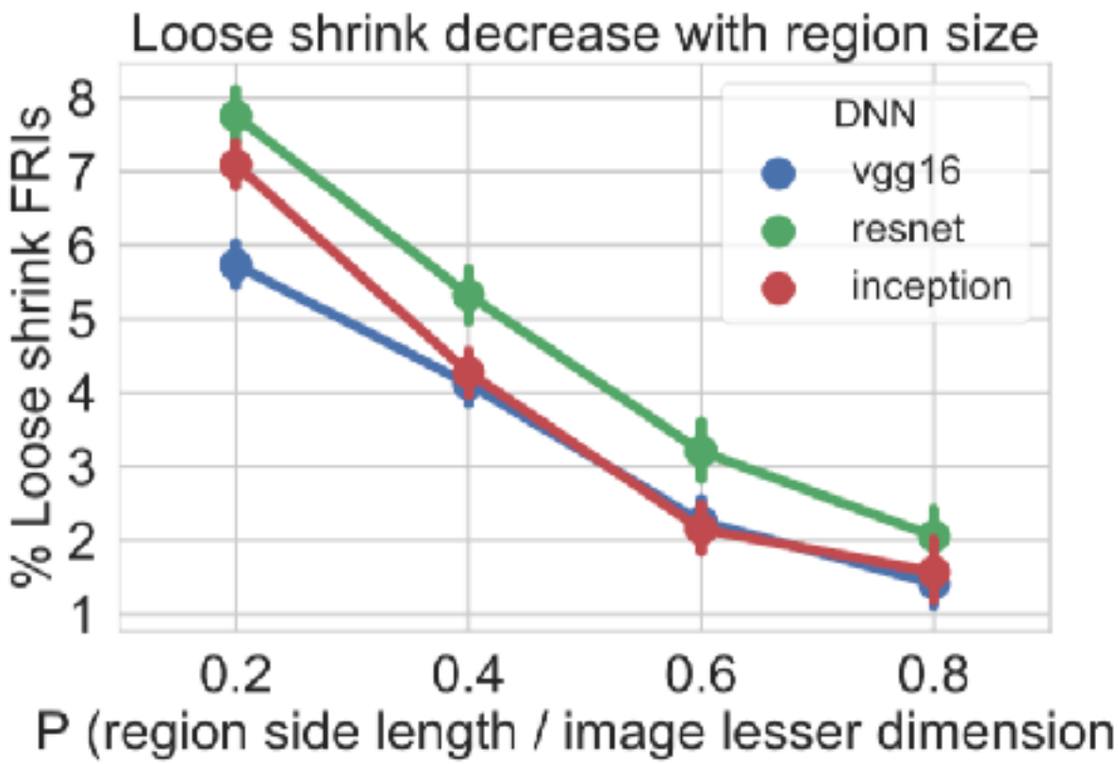
ImageNet class: Artichoke
Human option: Vegetable

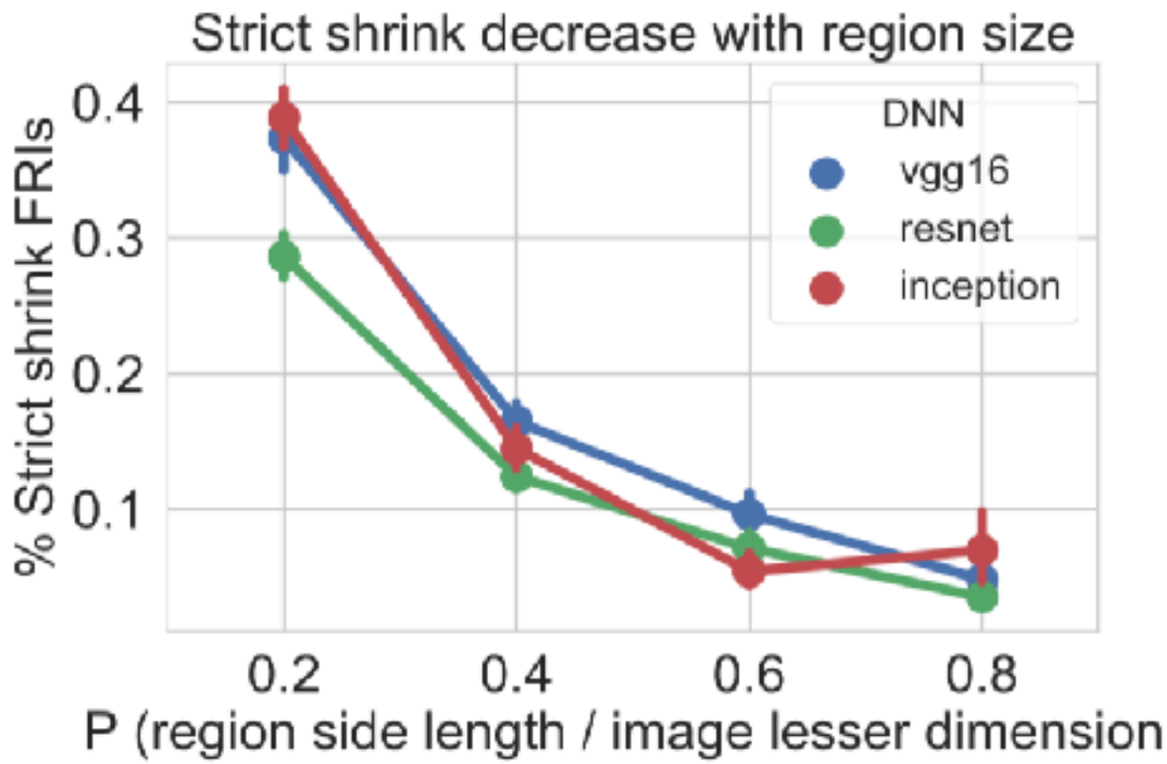
ResNet score: ~0.8
Human success
rate: 65%

ResNet score: ~0.4
Human success
rate: 75%

Difference from humans (1): DNNs can be affected by a single-pixel shift or shrink of the visible image region

DNNs have large FRIs





Difference with humans (2): DNNs are generally affected by visible image regions that are larger than human minimal images

Is this a new type of adversarial example without synthetic patterns?

In synthetic images:

Luo et al. 2016

Original | Adversarial | Adversarial (exaggerated to show perturbation)

Targeted synthetic changes that cause DNN failure

Generally imperceptible to humans

In **natural images**:

Various translations of a single image

Translations can cause DNNs to fail (Engstrom et al. 2016) (Azulay, Weiss 2018)

Location Invariance and Minimal Images

Experiment:

- **CIFAR10**

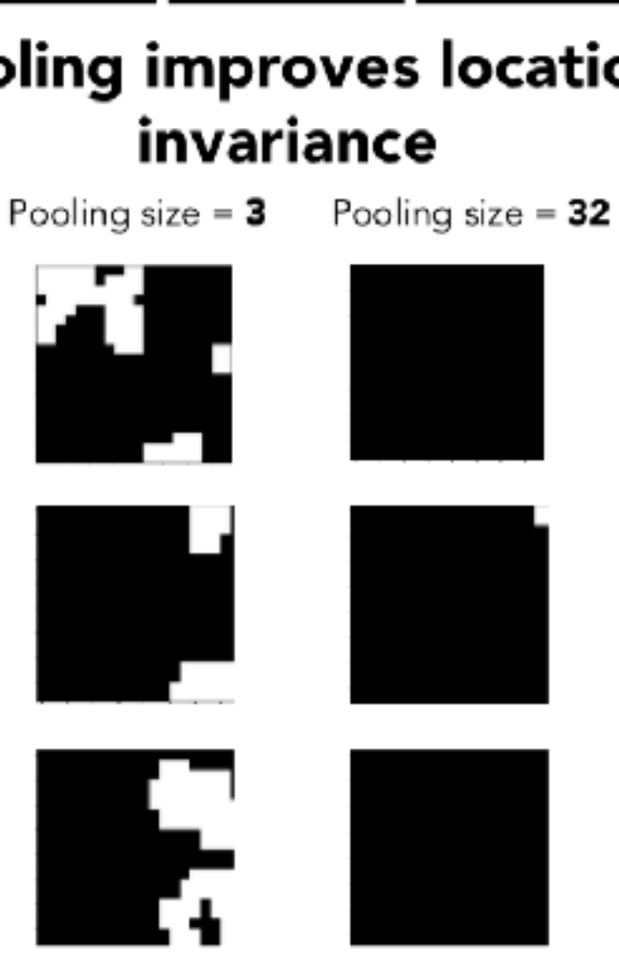
- **Network: 2 convolutions + 2 fully connected (standard, ~75% accuracy)**

- **We vary the *pooling region size:***

  Larger pooling sizes lead to more position invariance.

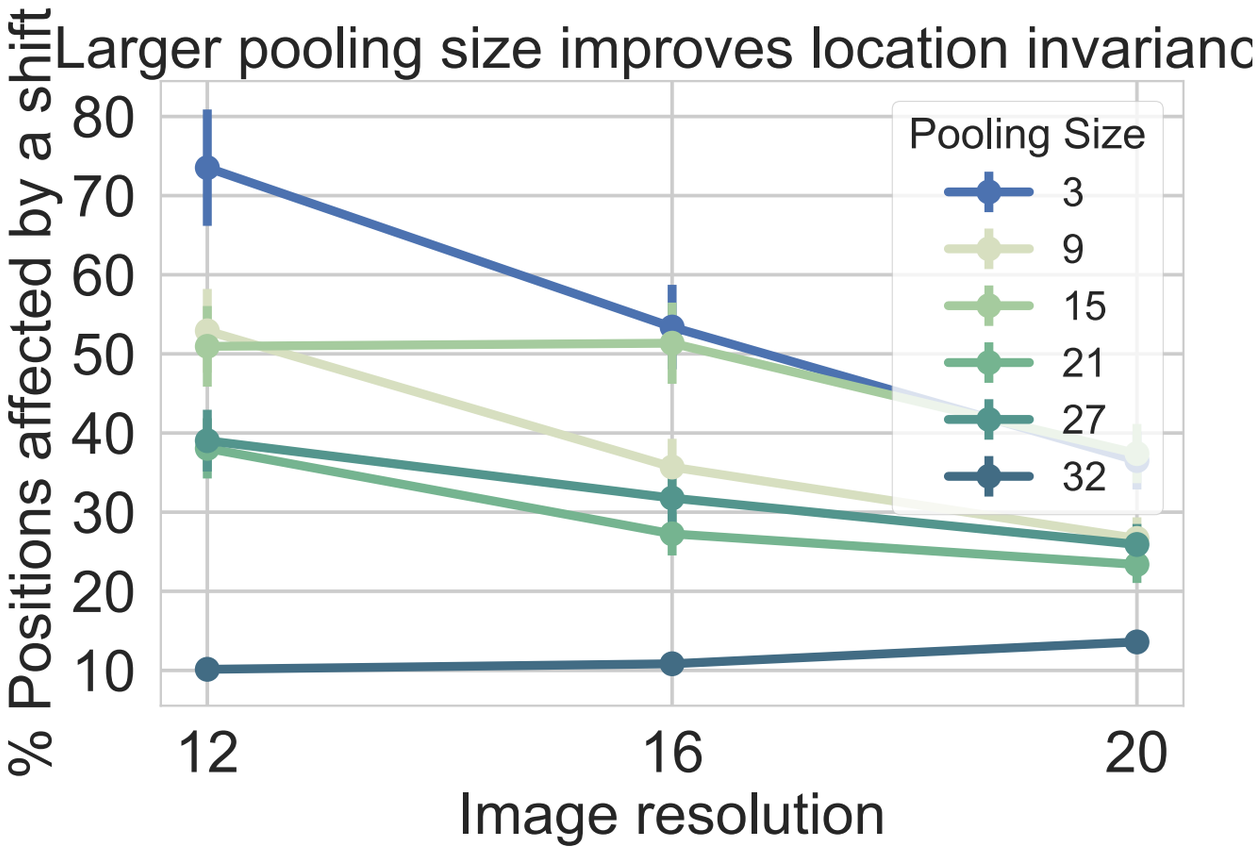
  Pooling size = 32: entire image pooled (maximum invariance, similar accuracy ~72%)

Pooling improves location invariance



White pixels indicate positions for which a shift of 1 pixel will change network accuracy

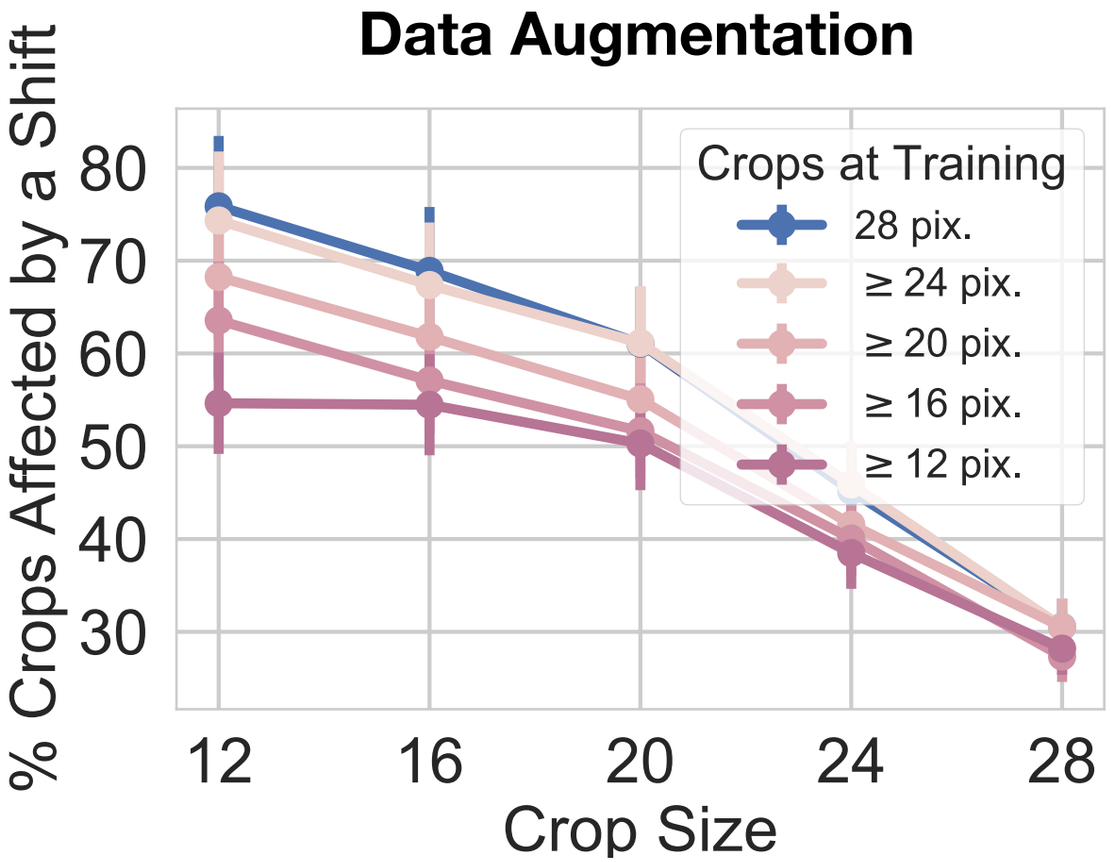
Pooling and Fragile Recognition

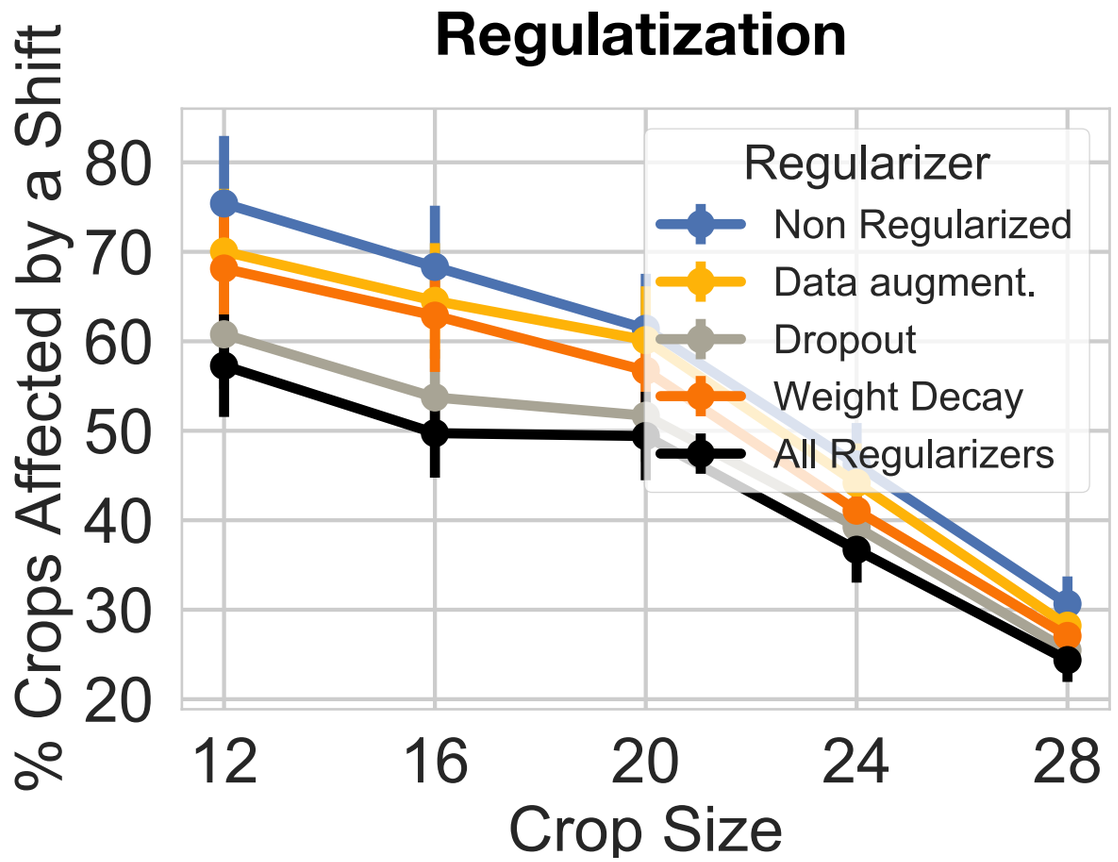


Increasing pooling size significantly improves translation invariance at small image sizes, but only somewhat improves minimal images.

FRIs are related to location invariance but also depend on other factors

Can Fragile Recognition be Alleviated?





Data augmentation helps generalization to limited information; regularizers also help

Neither eliminates the problem

VGG16, Resnet, Inception already have data augmentation of relatively large crops

3 Snapshots

1) Failures of DNNs

Minimal images are a common phenomenon among humans and DNNs

How can we make DNNs robust to minimal images as humans?

2) Beyond Object Recognition

3) DNNs to explain the Brain

3 Snapshots

1) Failures of DNNs

Minimal images are a common phenomenon among humans and DNNs

How can we make DNNs robust to minimal images as humans?

2) Beyond Object Recognition

Insideness is solvable but not learnable by state-of-the-art DNNs

What learning mechanisms could lead to the general solution?

3) DNNs to explain the Brain

Single units in a DNN correspond with neurons in the brain

with

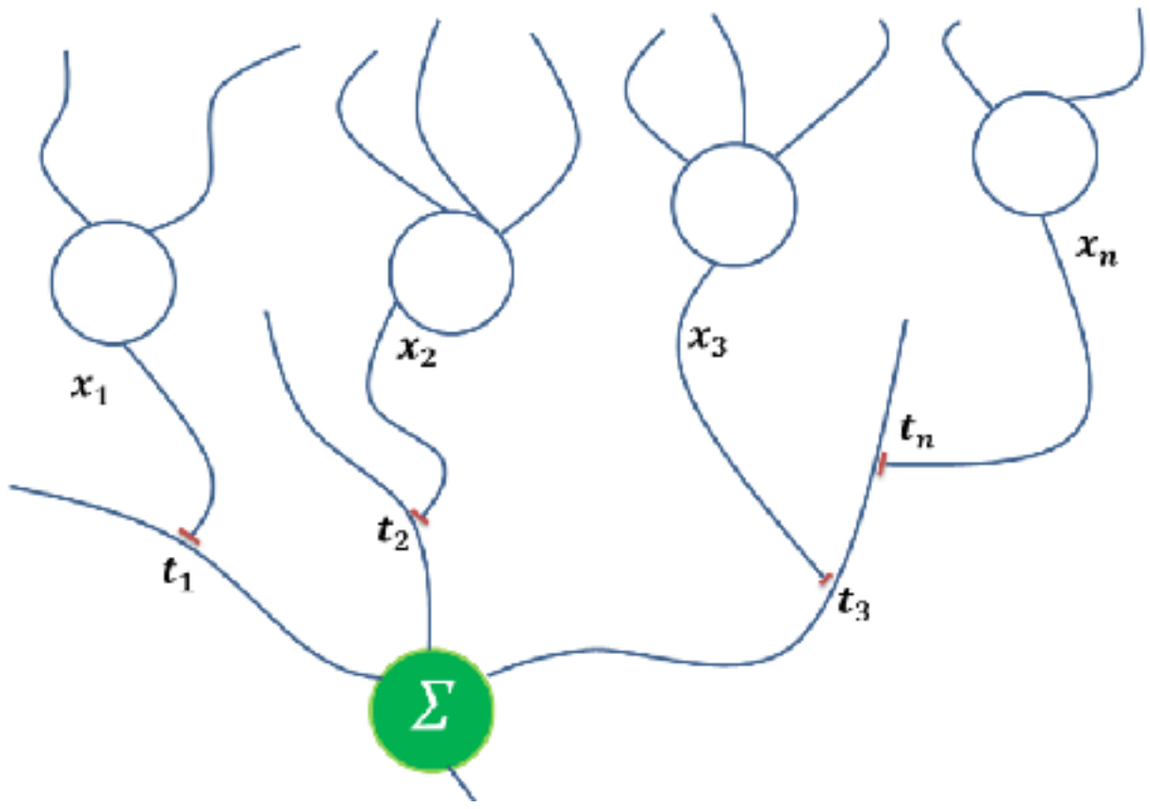
Luke Arend, Yena Han, Martin Schrimpf, Pouya Bashivan, Kohitij Kar, Tomaso Poggio, James DiCarlo

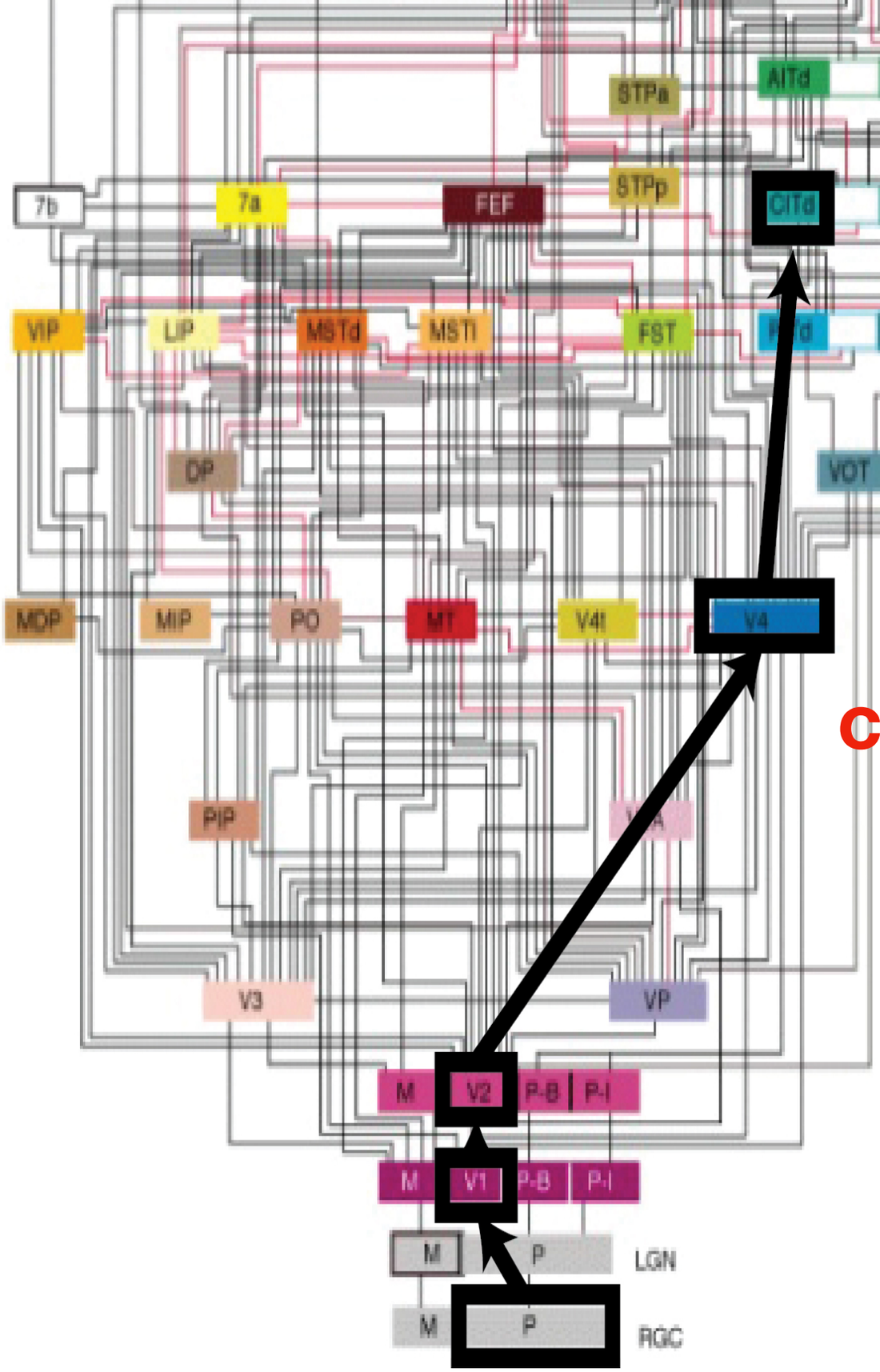
Model of a neuron

Simplified neuroscience: a neuron computes a dot product between its inputs and the synaptic weights

$< x, t >$ ⟷

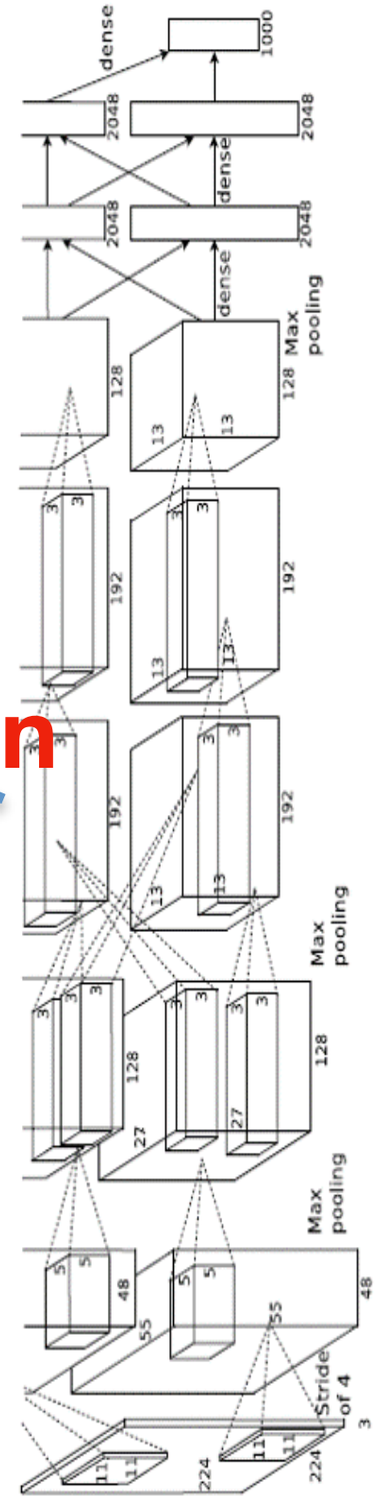
Neuroscience definition of dot product!





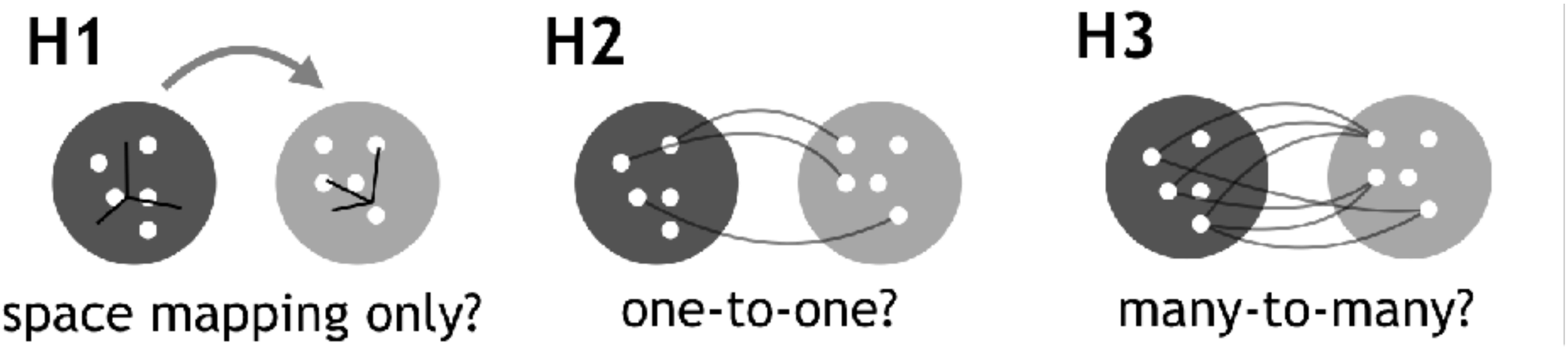
Felleman and Van Essen 1991

Linear combination



Krizhevsky et al 2012

Can we map individual DNN units to individual neurons in the brain?



If yes,

1) **DNN are better models of the brain than previously thought**

2) **the unit to neuron mapping is more interpretable than a “linear combination of units” to neuron**

1) is there a one-to-one mapping?

2) which cells in the model match the neural recordings?

3) which cells in the model don’t match the neural recordings?

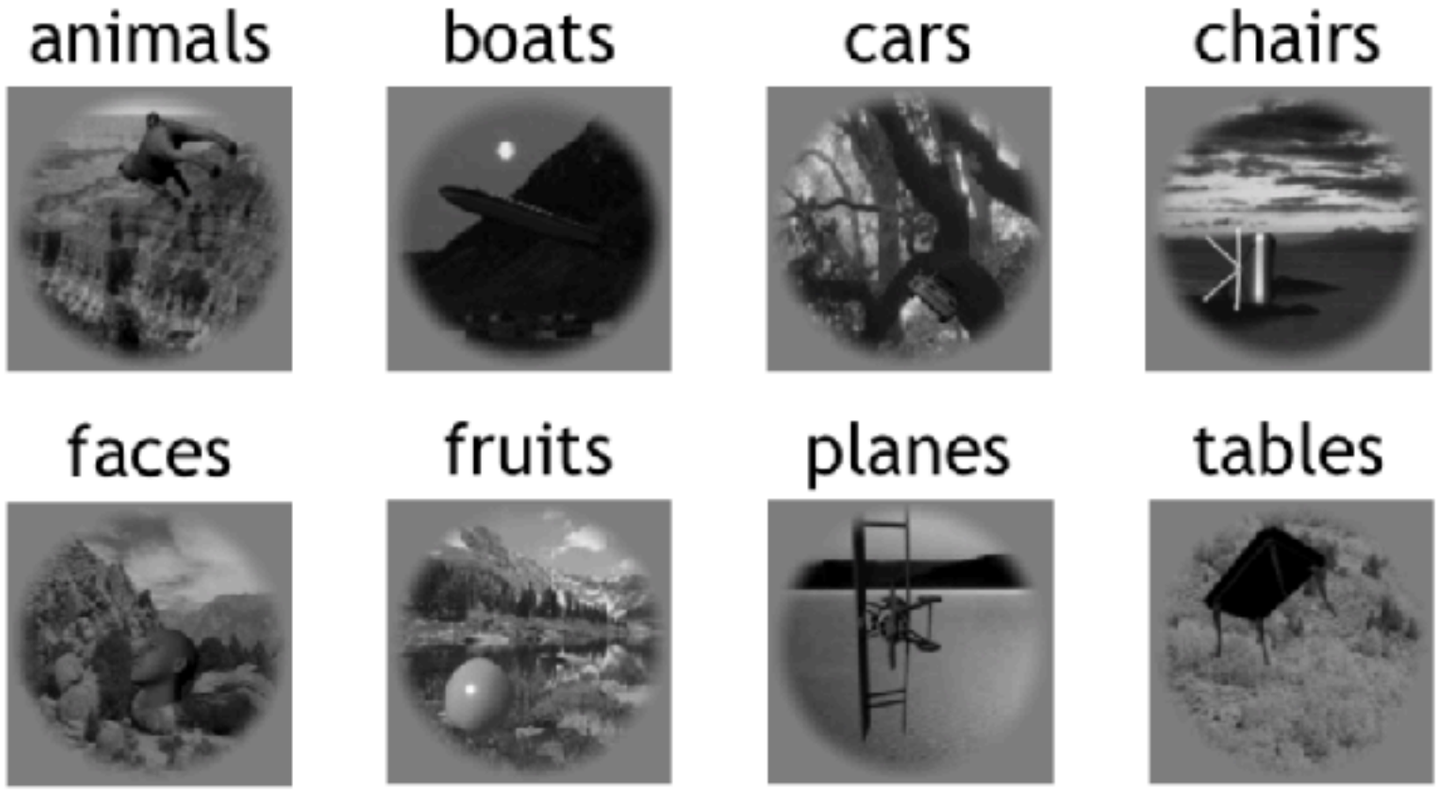
1) is there a one-to-one mapping?

2) which cells in the model match the neural recordings?

3) which cells in the model don't match the neural recordings?

stimuli

- 2560 images
- 8 categories
- high variation in location, pose and size



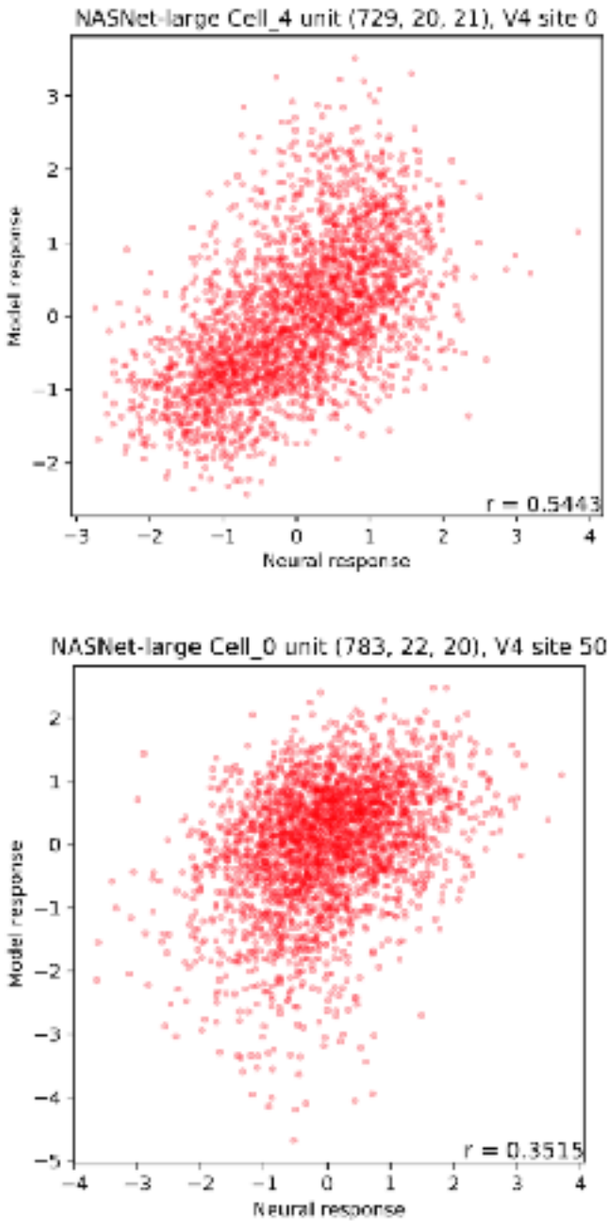
neural recordings

- 88 V4 sites, 168 IT sites
- previously published in [1, 2, 3]

models

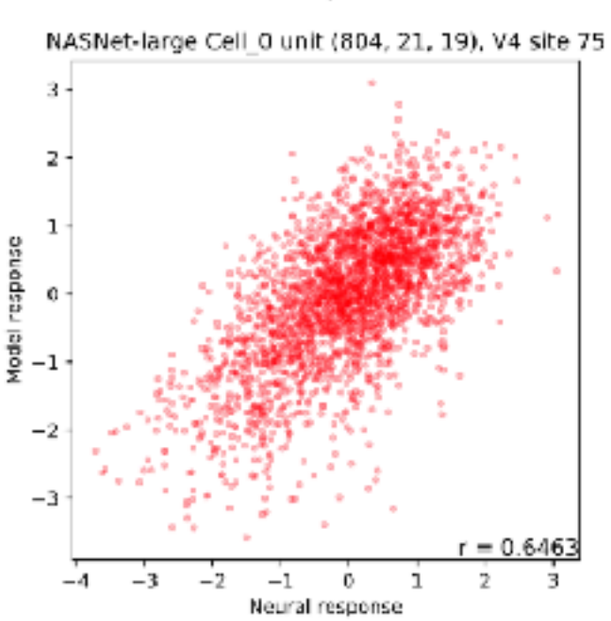
- 14 state-of-the-art deep neural networks
- optimized for object recognition

V4

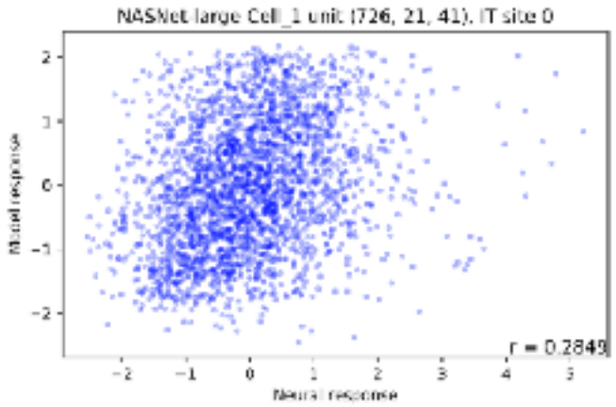

NASNet-large Cell_4 unit (729, 20, 21), V4 site 0
r = 0.5443
Model response
Neural response

NASNet-large Cell_2 unit (828, 25, 21), V4 site 25
r = 0.5935
Model response
Neural response

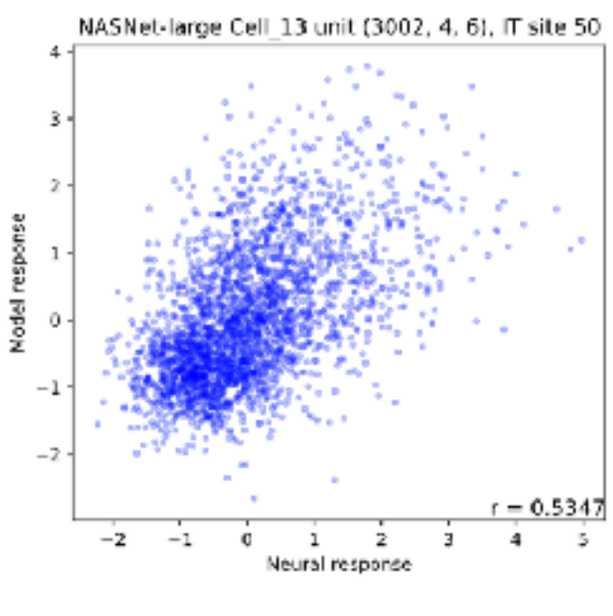
NASNet-large Cell_0 unit (783, 22, 20), V4 site 50
r = 0.3515
Model response
Neural response

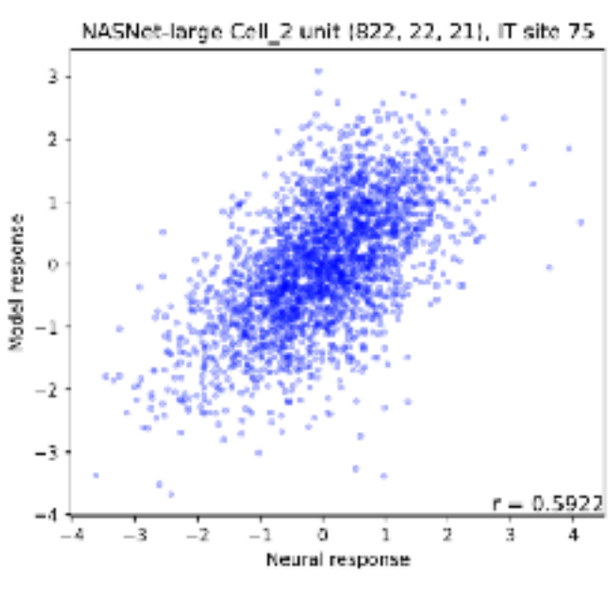

NASNet-large Cell_0 unit (804, 21, 19), V4 site 75
r = 0.6463
Model response
Neural response

IT

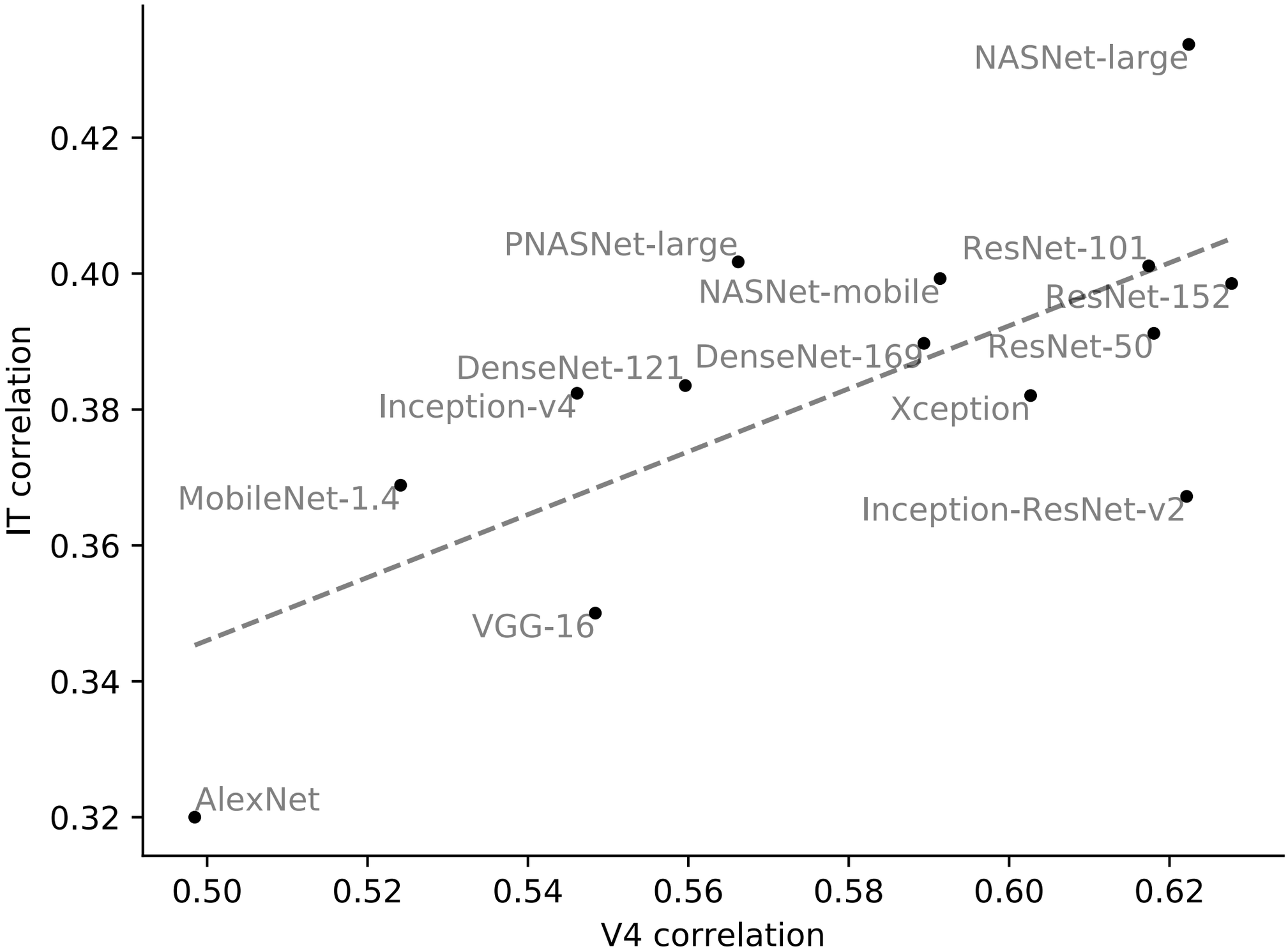

NASNet-large Cell_1 unit (726, 21, 41), IT site 0
r = 0.2849
Model response
Neural response

NASNet-large Cell_13 unit (3002, 4, 6), IT site 25
r = 0.4833
Model response
Neural response

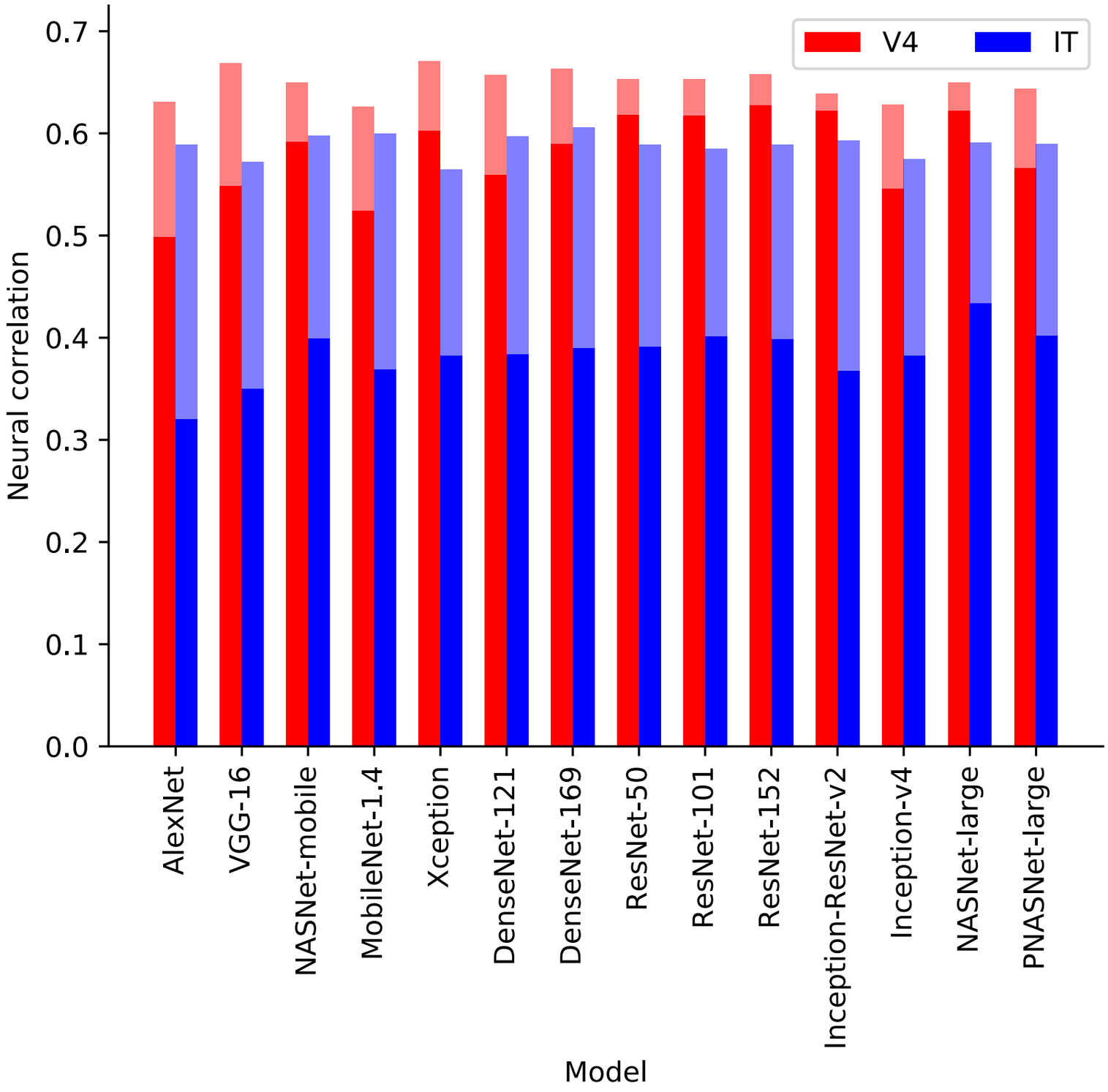

NASNet-large Cell_13 unit (3002, 4, 6), IT site 50
r = 0.5347
Model response
Neural response

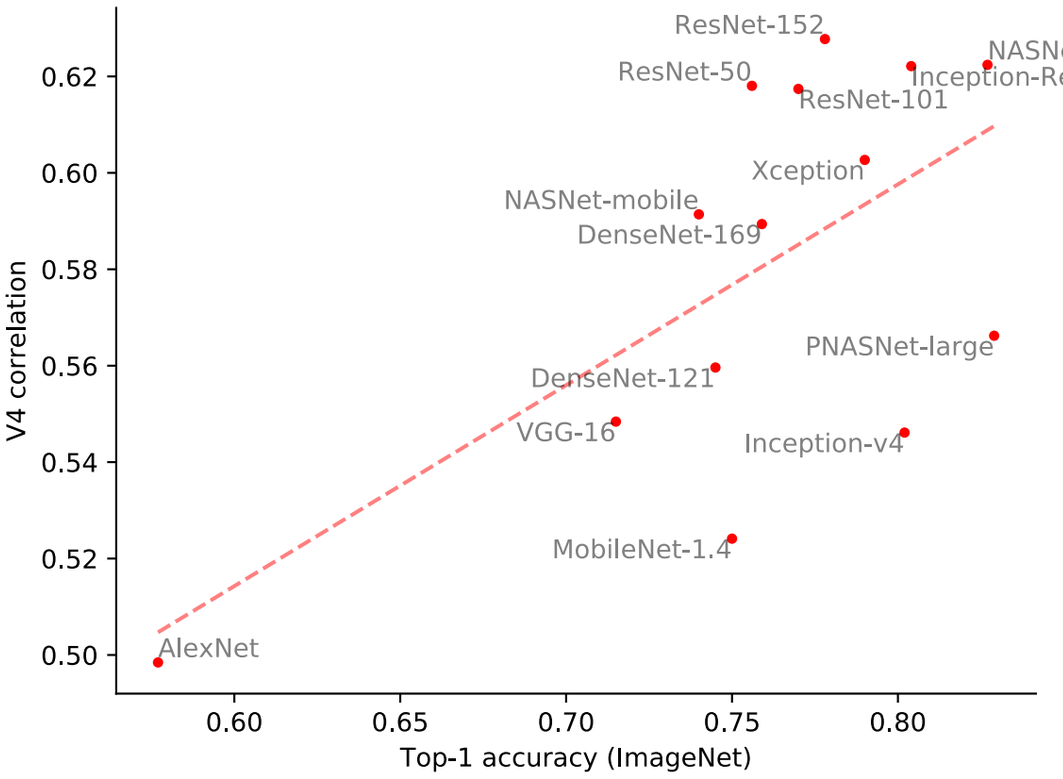

NASNet-large Cell_2 unit (822, 22, 21), IT site 75
r = 0.5922
Model response
Neural response

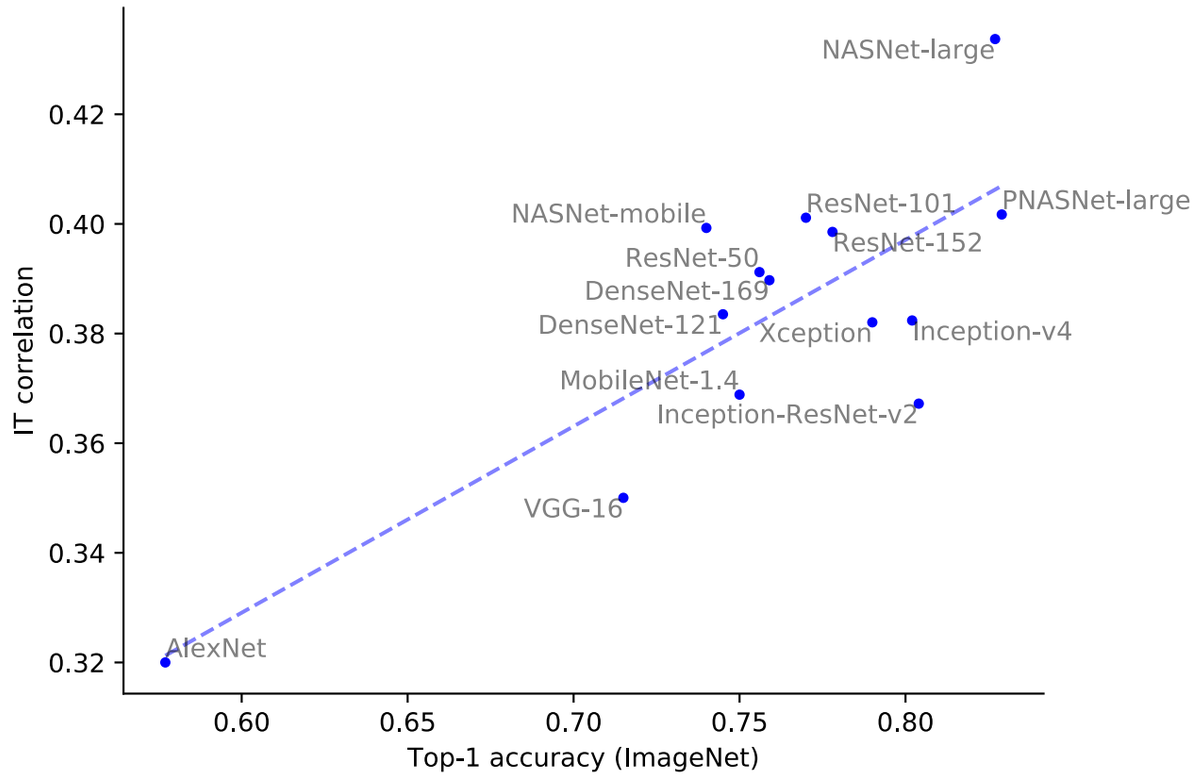
median across population of correlation for best-matched units



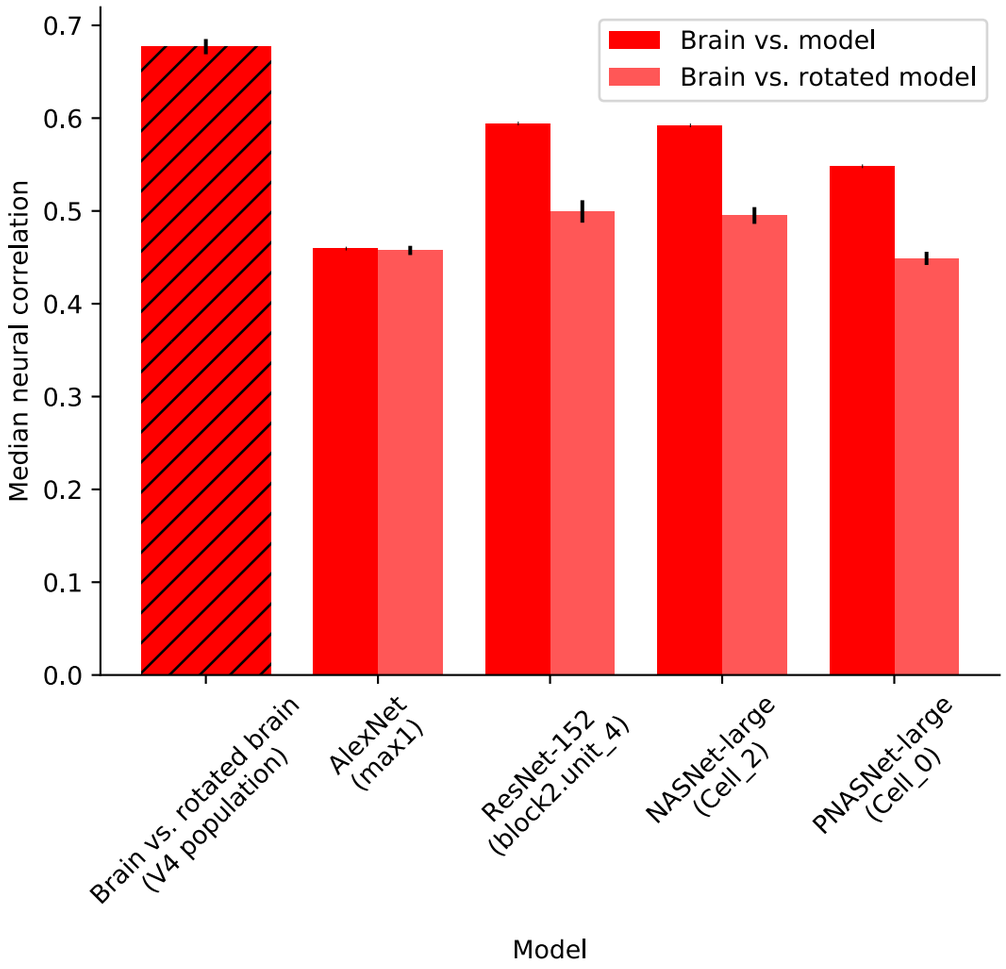
Linear Combination vs Individual Units



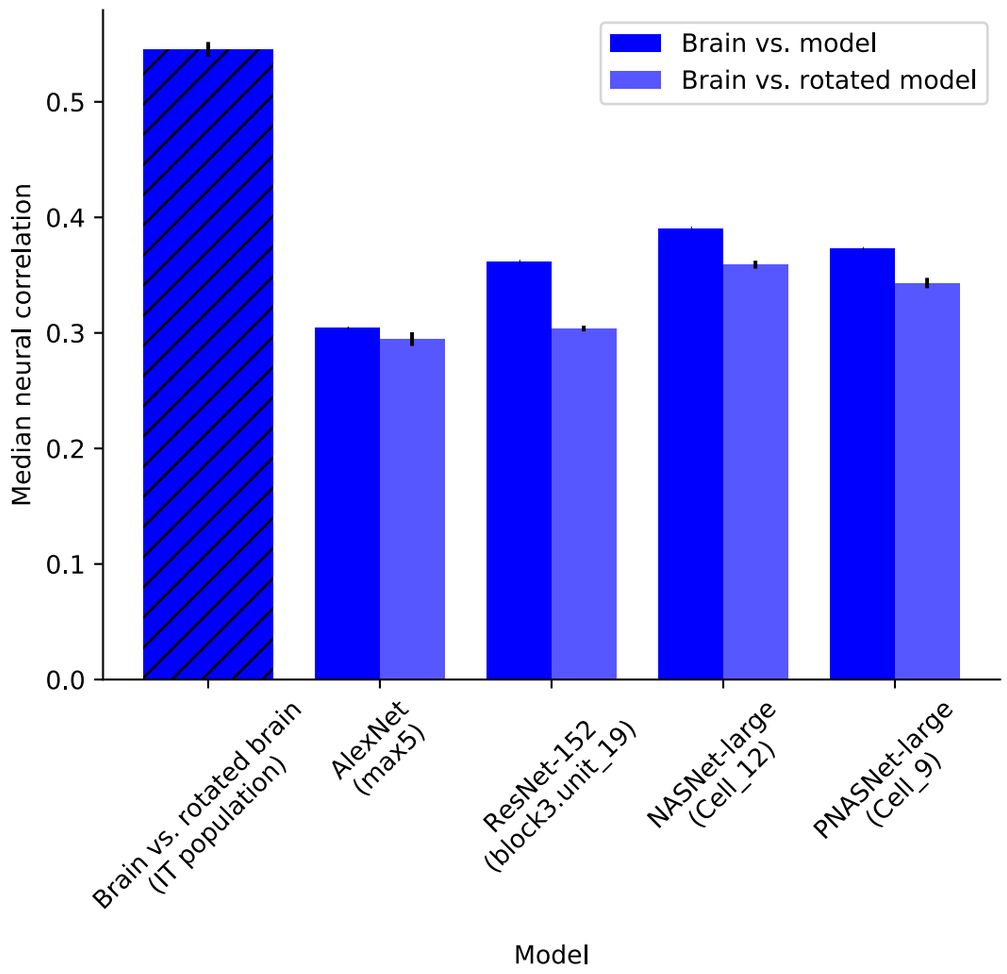

V4
ResNet-152
ResNet-50
NASNet-large
Inception-ResNet-v2
ResNet-101
Xception
NASNet-mobile
DenseNet-169
PNASNet-large
DenseNet-121
VGG-16
Inception-v4
MobileNet-1.4
AlexNet
V4 correlation
0.62
0.60
0.58
0.56
0.54
0.52
0.50
0.60
0.65
0.70
0.75
0.80
Top-1 accuracy (ImageNet)


IT
NASNet-large
NASNet-mobile
ResNet-101
PNASNet-large
ResNet-50
ResNet-152
DenseNet-169
DenseNet-121
Xception
Inception-v4
MobileNet-1.4
Inception-ResNet-v2
VGG-16
AlexNet
IT correlation
0.42
0.40
0.38
0.36
0.34
0.32
0.60
0.65
0.70
0.75
0.80
Top-1 accuracy (ImageNet)

V4


Brain vs. model
Brain vs. rotated model
Median neural correlation
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Brain vs. rotated brain (V4 population)
AlexNet (max1)
ResNet-152 (block2.unit_4)
NASNet-large (Cell_2)
PNASNet-large (Cell_0)
Model

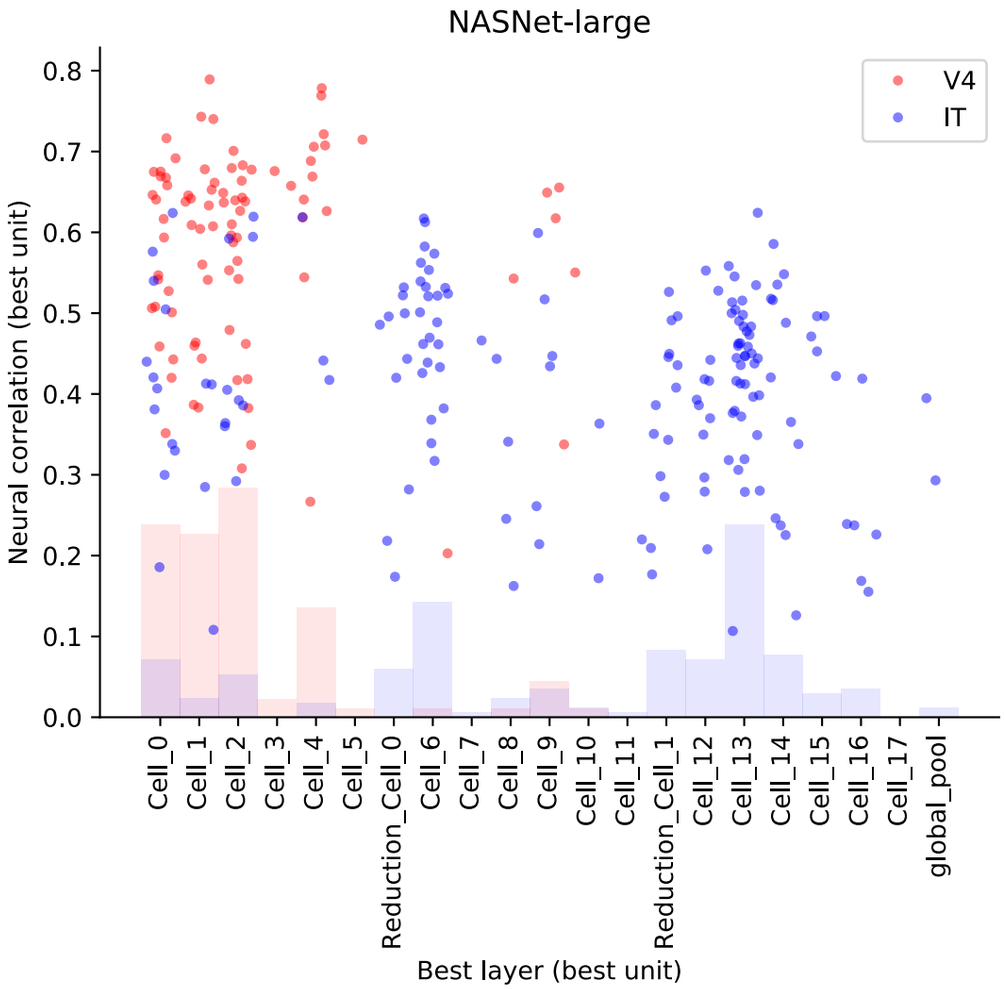
IT

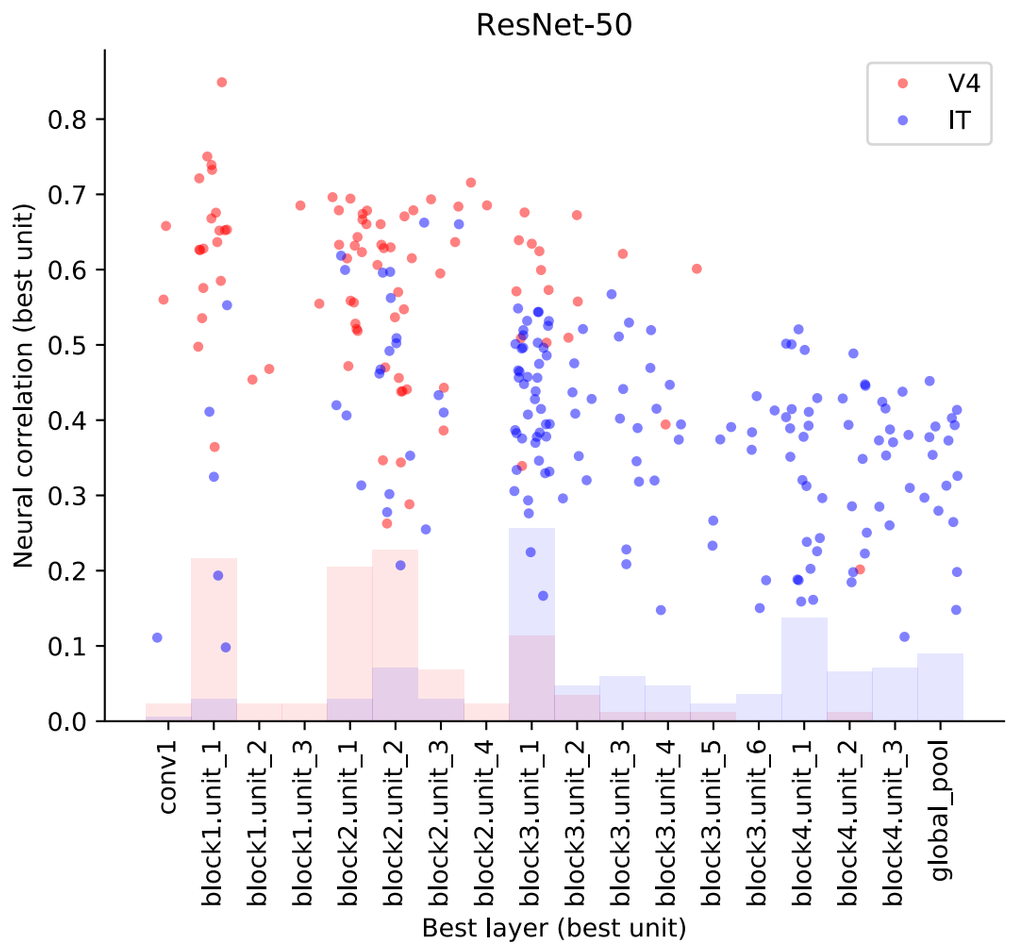

Brain vs. model
Brain vs. rotated model
Median neural correlation
0.0
0.1
0.2
0.3
0.4
0.5
Brain vs. rotated brain (IT population)
AlexNet (max5)
ResNet-152 (block3.unit_19)
NASNet-large (Cell_12)
PNASNet-large (Cell_9)
Model

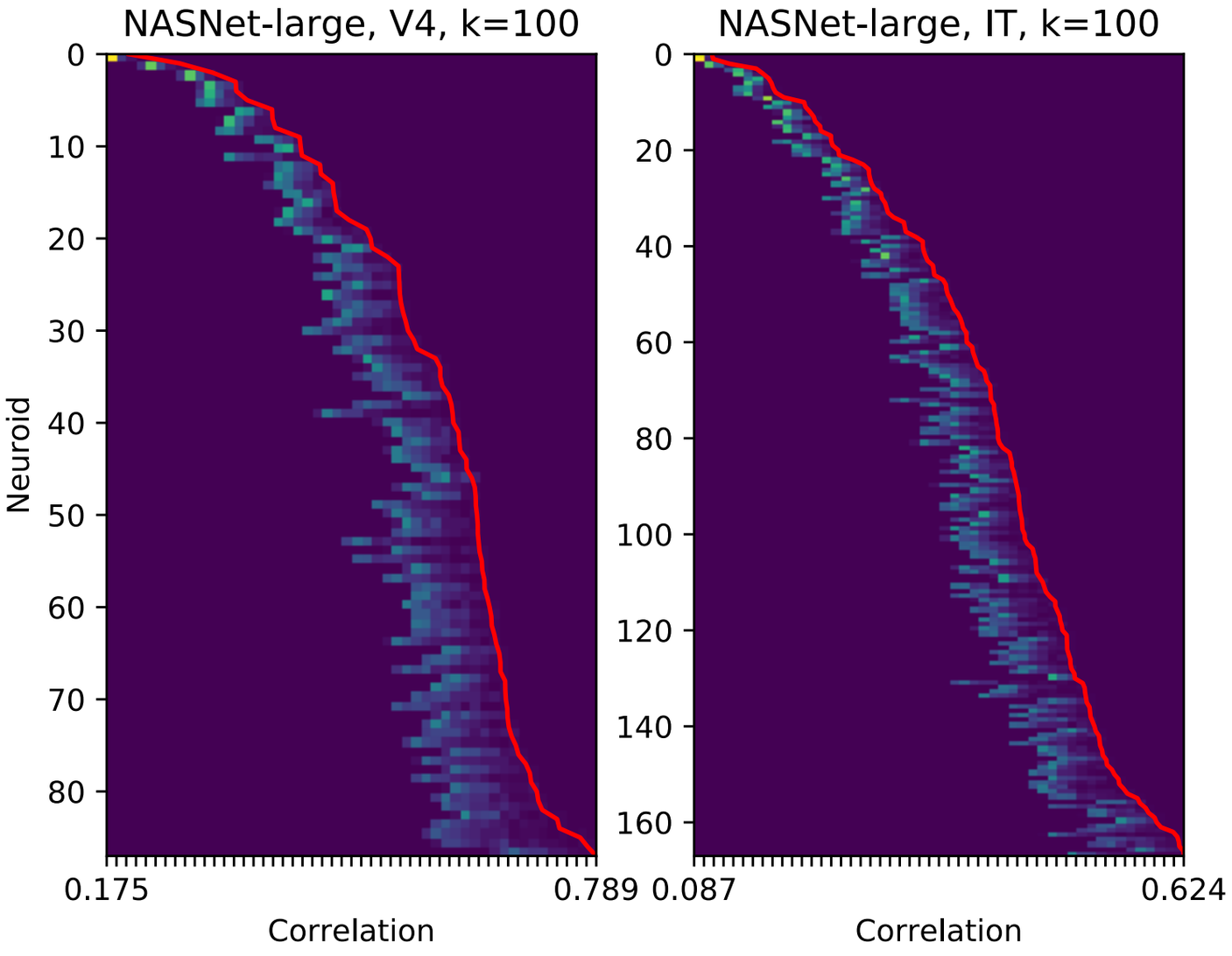
1) is there a one-to-one mapping?

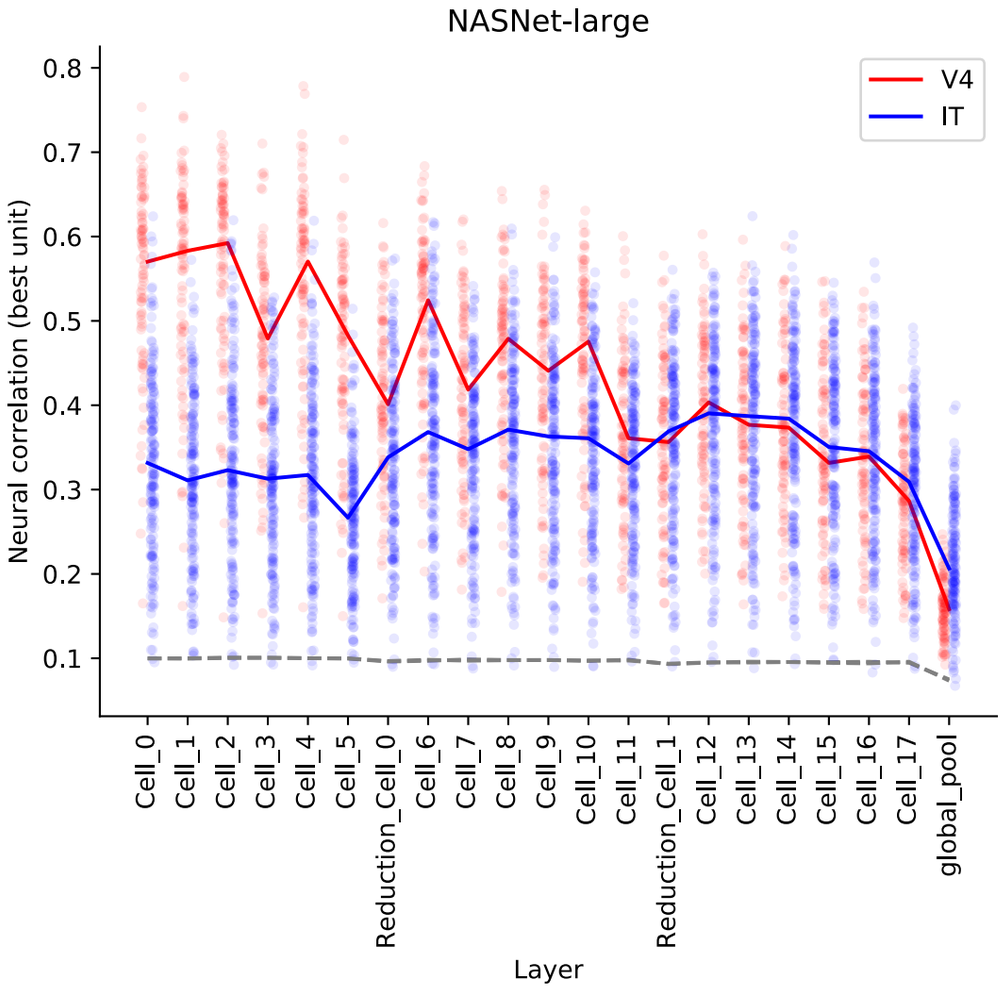
2) which cells in the model match the neural recordings?

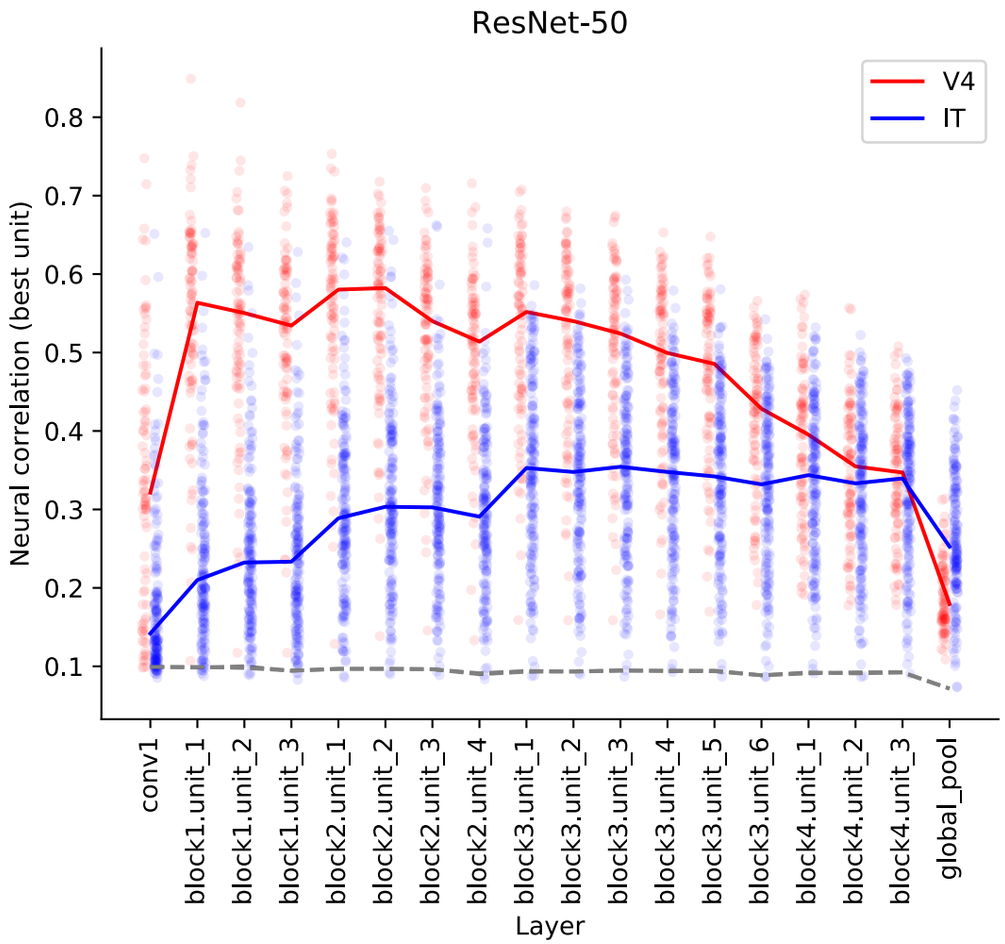
3) which cells in the model don't match the neural recordings?


NASNet-large
V4
IT
Neural correlation (best unit)
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
Cell_0
Cell_1
Cell_2
Cell_3
Cell_4
Cell_5
Reduction_Cell_0
Cell_6
Cell_7
Cell_8
Cell_9
Cell_10
Cell_11
Reduction_Cell_1
Cell_12
Cell_13
Cell_14
Cell_15
Cell_16
Cell_17
global_pool
Best layer (best unit)


ResNet-50
V4
IT
Neural correlation (best unit)
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
conv1
block1.unit_1
block1.unit_2
block1.unit_3
block2.unit_1
block2.unit_2
block2.unit_3
block2.unit_4
block3.unit_1
block3.unit_2
block3.unit_3
block3.unit_4
block3.unit_5
block3.unit_6
block4.unit_1
block4.unit_2
block4.unit_3
global_pool
Best layer (best unit)


NASNet-large, V4, k=100
NASNet-large, IT, k=100
Neuroid
0
10
20
30
40
50
60
70
80
0
20
40
60
80
100
120
140
160
0.175
0.789
0.087
0.624
Correlation
Correlation

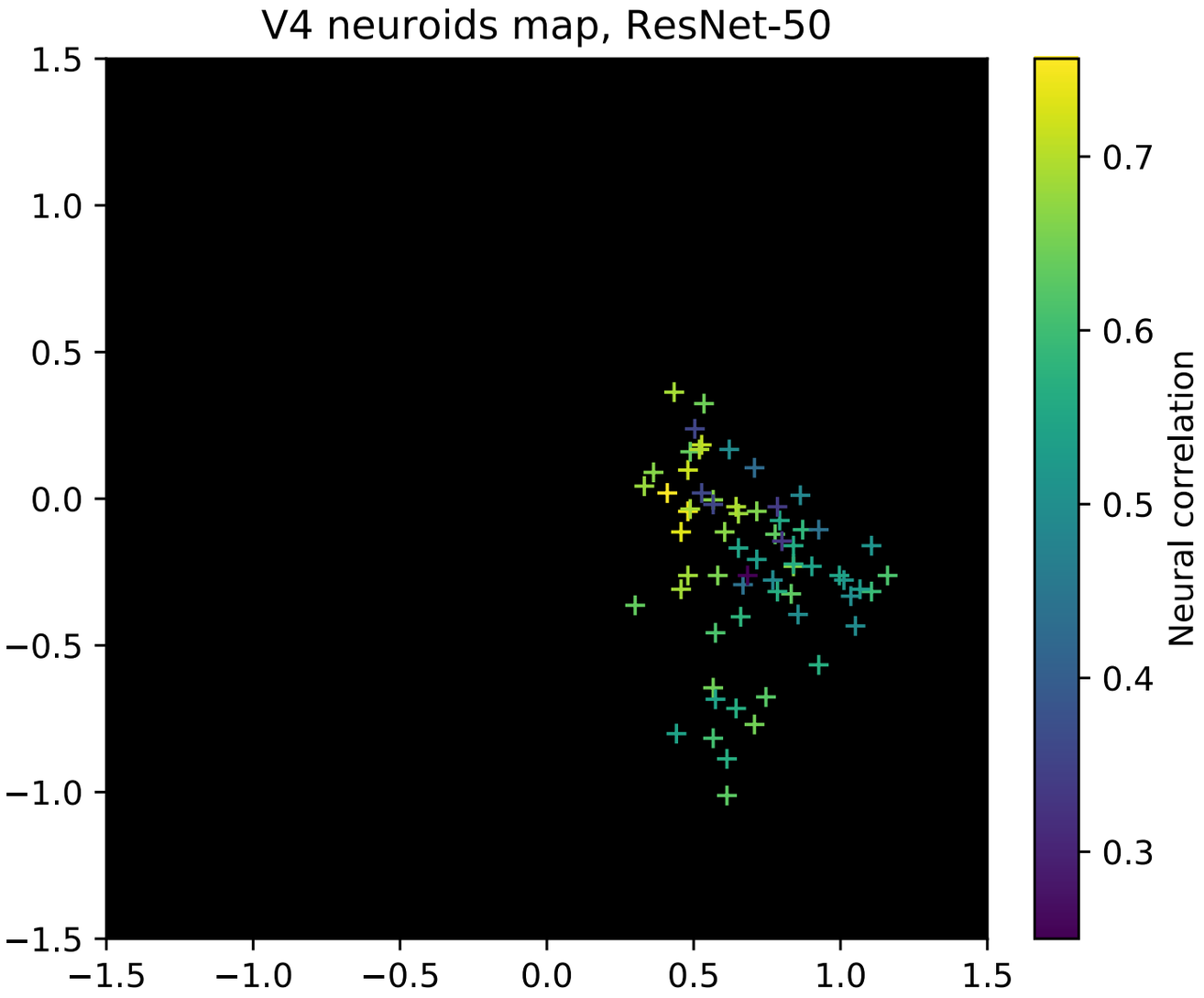

NASNet-large
V4
IT
Neural correlation (best unit)
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
Cell_0
Cell_1
Cell_2
Cell_3
Cell_4
Cell_5
Reduction_Cell_0
Cell_6
Cell_7
Cell_8
Cell_9
Cell_10
Cell_11
Reduction_Cell_1
Cell_12
Cell_13
Cell_14
Cell_15
Cell_16
Cell_17
global_pool
Layer

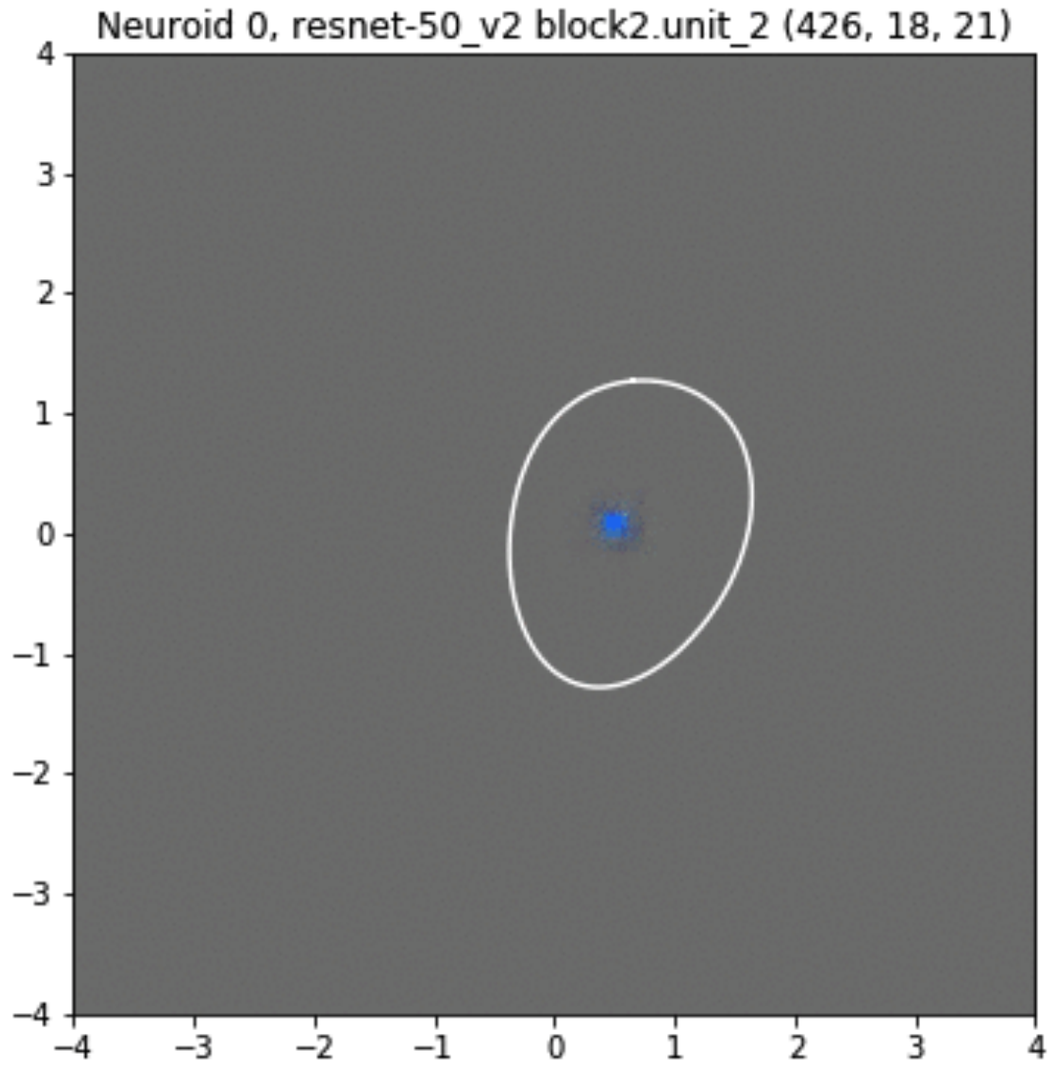

ResNet-50
V4
IT
Neural correlation (best unit)
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
conv1
block1.unit_1
block1.unit_2
block1.unit_3
block2.unit_1
block2.unit_2
block2.unit_3
block2.unit_4
block3.unit_1
block3.unit_2
block3.unit_3
block3.unit_4
block3.unit_5
block3.unit_6
block4.unit_1
block4.unit_2
block4.unit_3
global_pool
Layer

1) is there a one-to-one mapping?

2) which cells in the model match the neural recordings?

3) which cells in the model don't match the neural recordings?


V4 neuroids map, ResNet-50
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
0.7
0.6
0.5
0.4
0.3
Neural correlation


Neuroid 0, resnet-50_v2 block2.unit_2 (426, 18, 21)
4
3
2
1
0
−1
−2
−3
−4
−4
−3
−2
−1
0
1
2
3
4

3 Snapshots

1) Failures of DNNs

Minimal images are a common phenomenon among humans and DNNs

How can we make DNNs robust to minimal images as humans?

2) Beyond Object Recognition

Insideness is solvable by state-of-the-art DNNs for image segmentation

What learning mechanisms could lead to the general solution?

3) DNNs to explain the Brain

Single units in DNNs correspond to neurons in V4 and IT (to lesser degree)

What leads to this correspondence?

1) Computational

Same principles should explain
artificial and biological intelligence

(eg. Why the brain is how it is?
eg. Why an algorithm works better than others?)

---

2) Algorithmic

Computer Vision
(Build better algorithms)

Modelling the Brain
(How does the brain work?)

3) Implementation

CENTER FOR
Brains
Minds+
Machines

Making a Science from the Computer Vision Zoo

Xavier Boix xboix@mit.edu

Sanjana

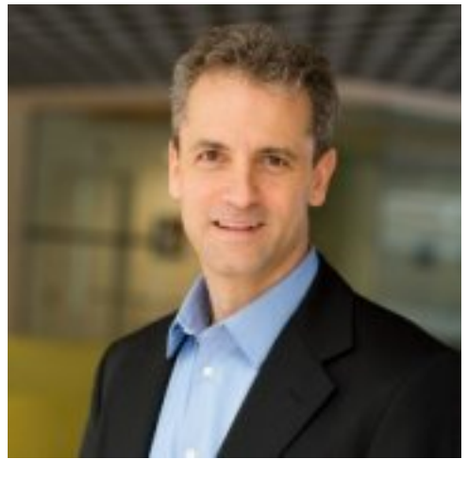
Kim

Luke

Guy

Tomotake

Fred



Prof. DiCarlo

Prof. Poggio