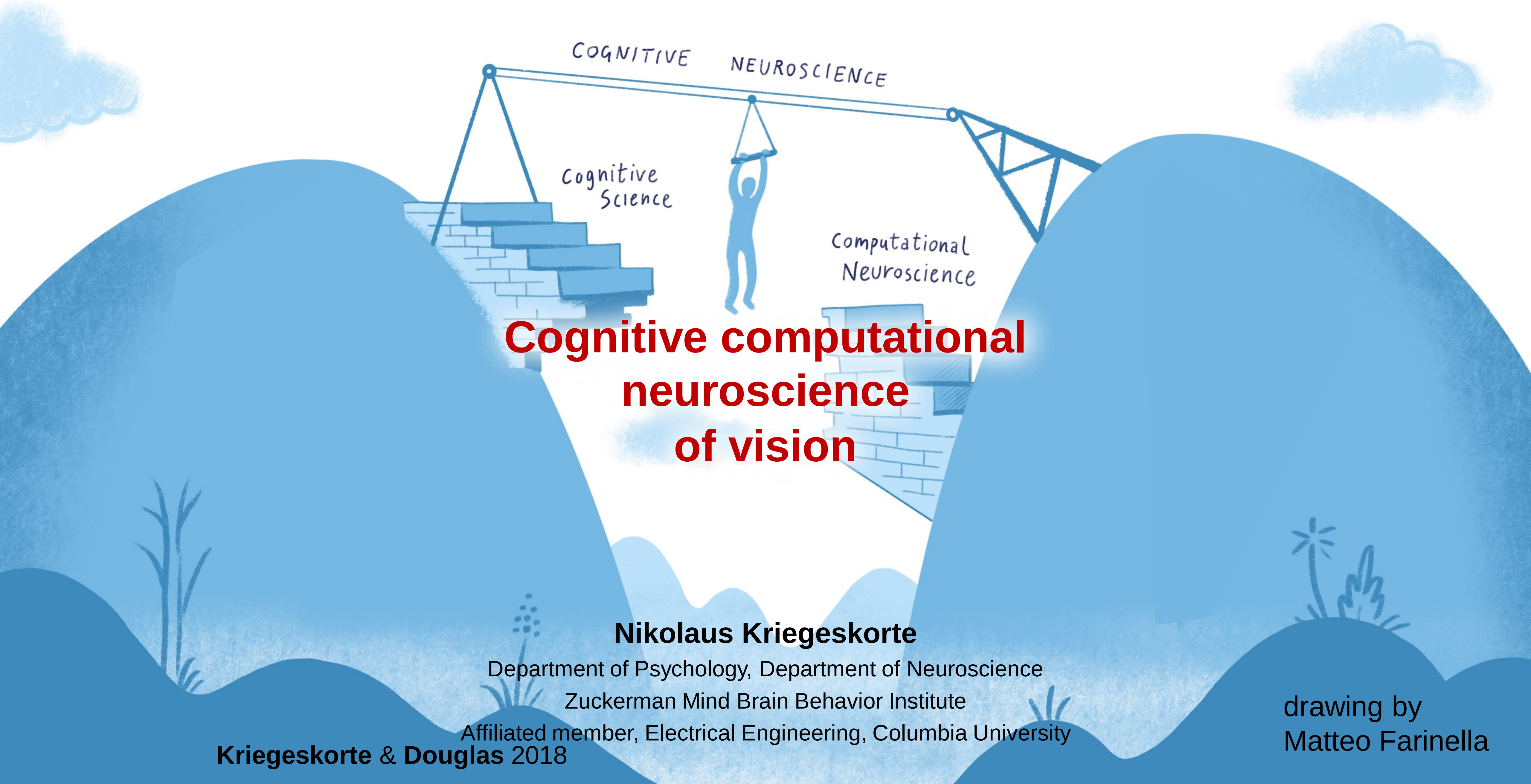
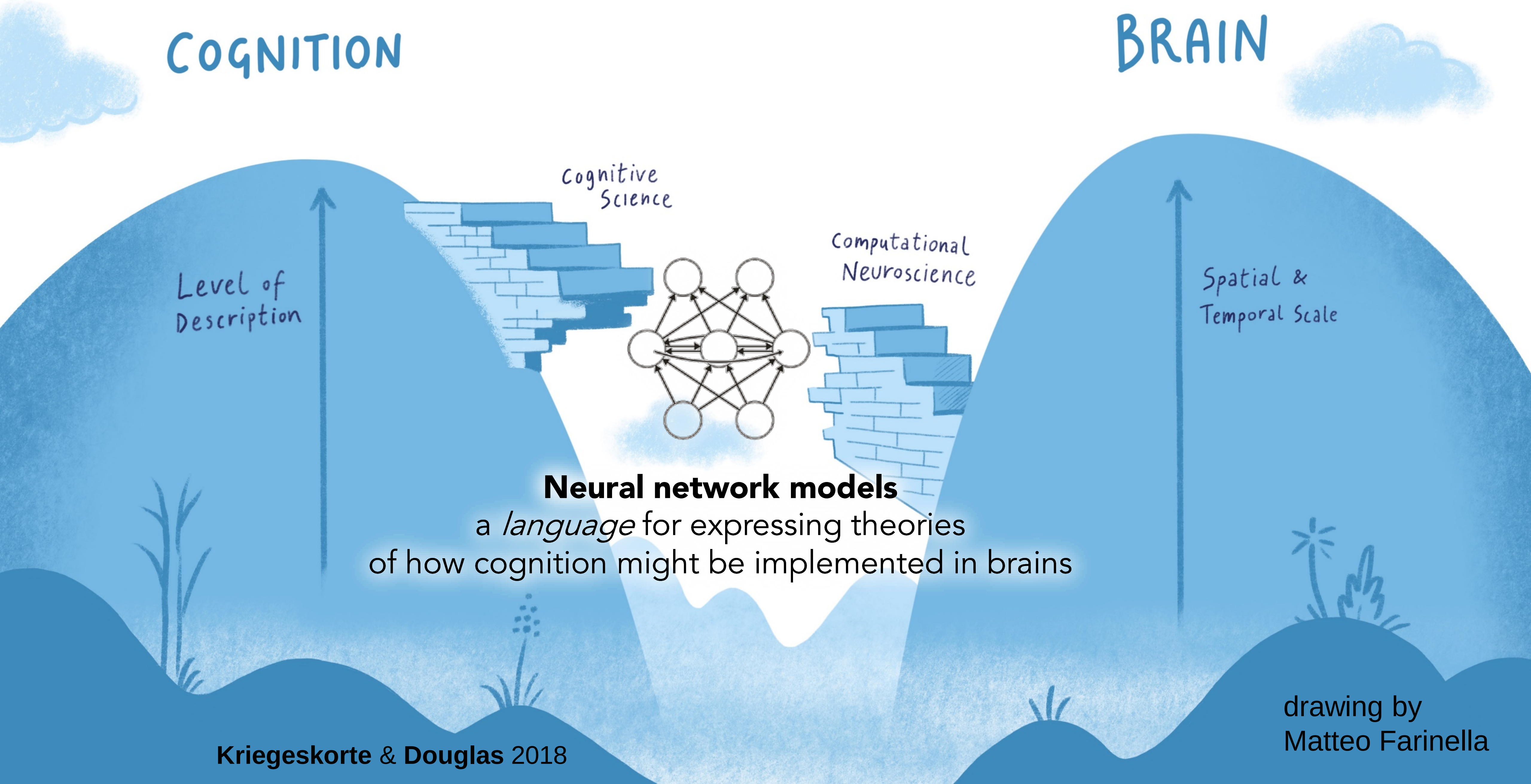


Cognitive computational neuroscience of vision

Nikolaus Kriegeskorte
Department of Psychology, Department of Neuroscience
Zuckerman Mind Brain Behavior Institute
Affiliated member, Electrical Engineering, Columbia University

Kriegeskorte & **Douglas** 2018

drawing by
Matteo Farinella



Neural network models
a *language* for expressing theories
of how cognition might be implemented in brains

Kriegeskorte & **Douglas** 2018

drawing by
Matteo Farinella

cognitive science
finding new roots
in neuroscience

glue

cognitive science

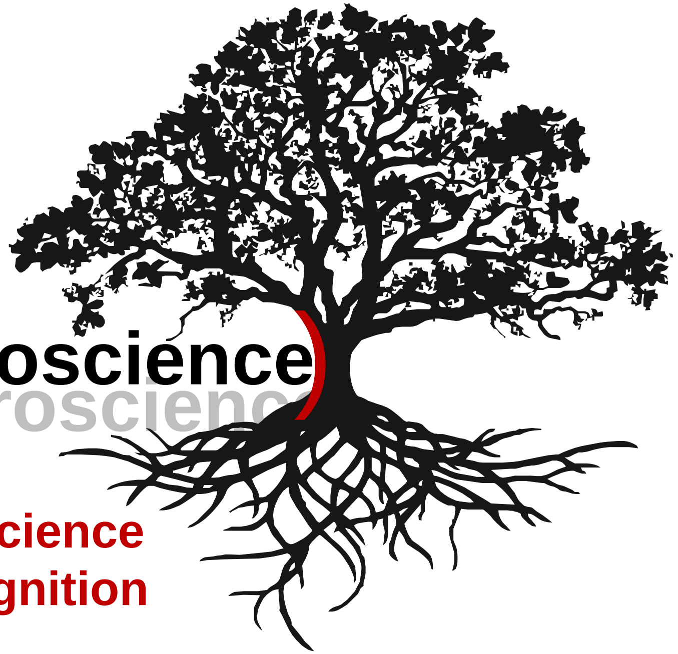
cognitive(computational neuroscience)

computational neuroscience

computational neuroscience
growing up toward cognition

history

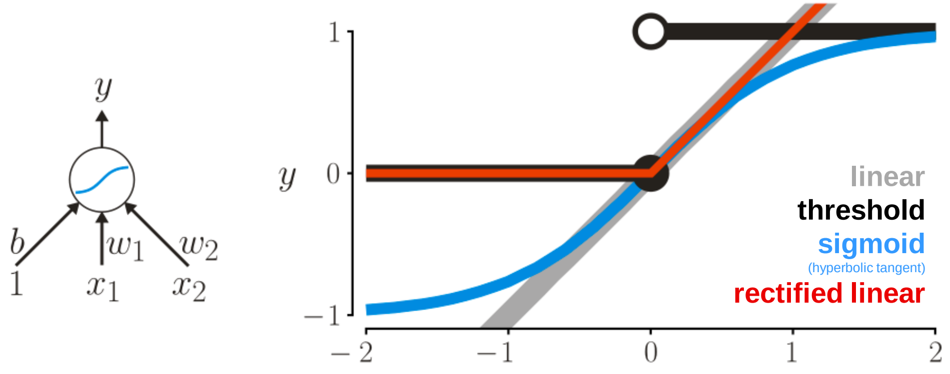
top-down



Kriegeskorte & **Douglas** 2018

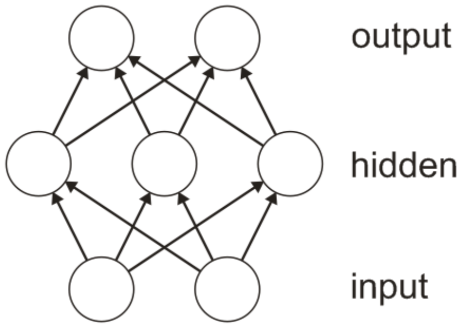
Neural network models

Neural network models

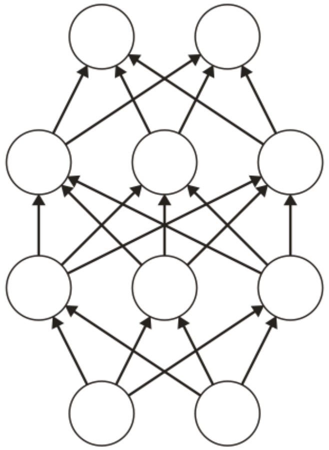


$$z = b + \sum_i x_i w_i$$

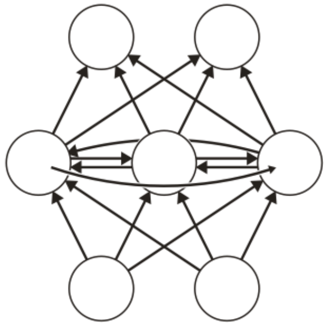
shallow feedforward
(1 hidden layer)



deep feedforward
(>1 hidden layer)

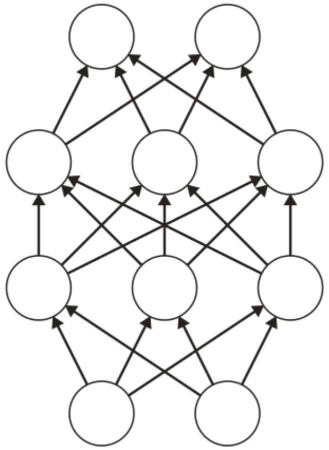


recurrent



Kriegeskorte 2015

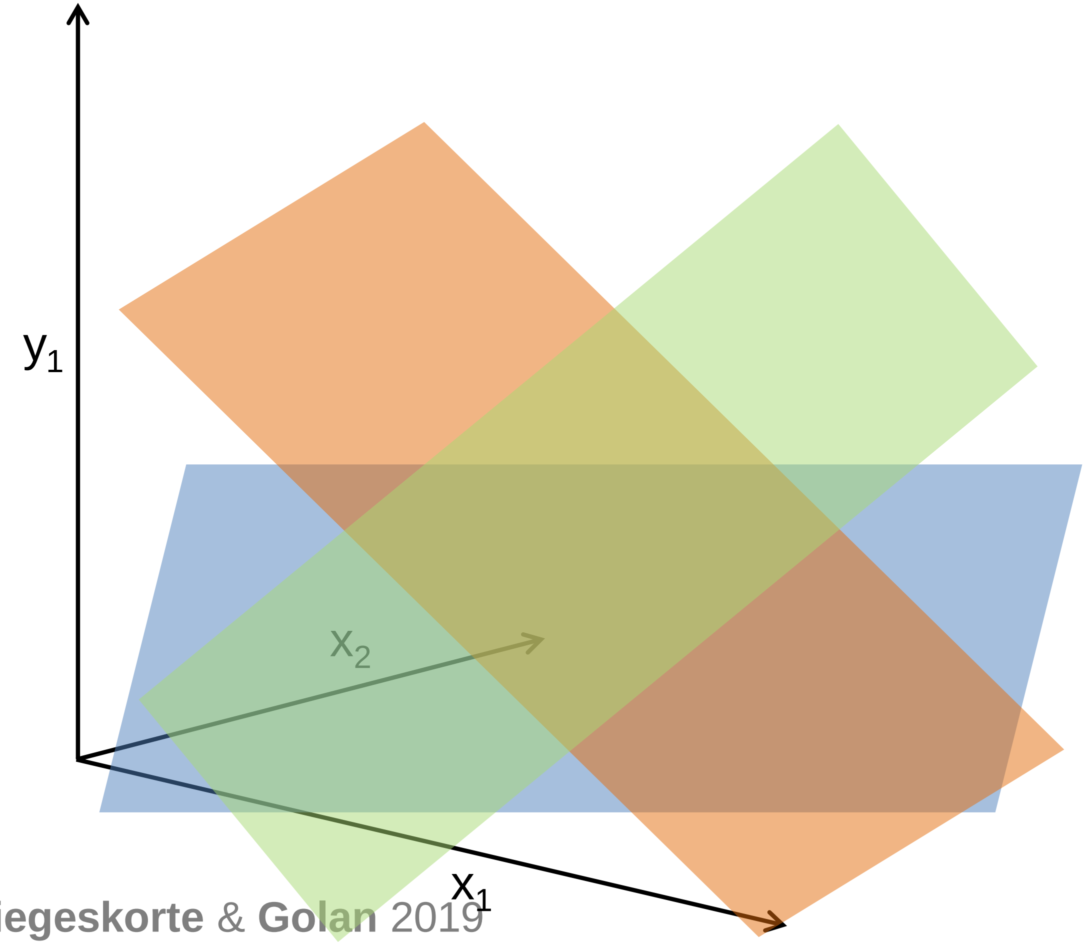
Neural network models

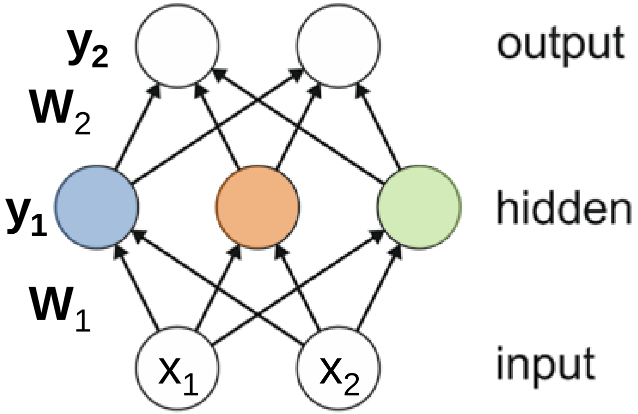


Kriegeskorte 2015

Nonlinear activation function needed to make a hidden layer useful

linear activation functions



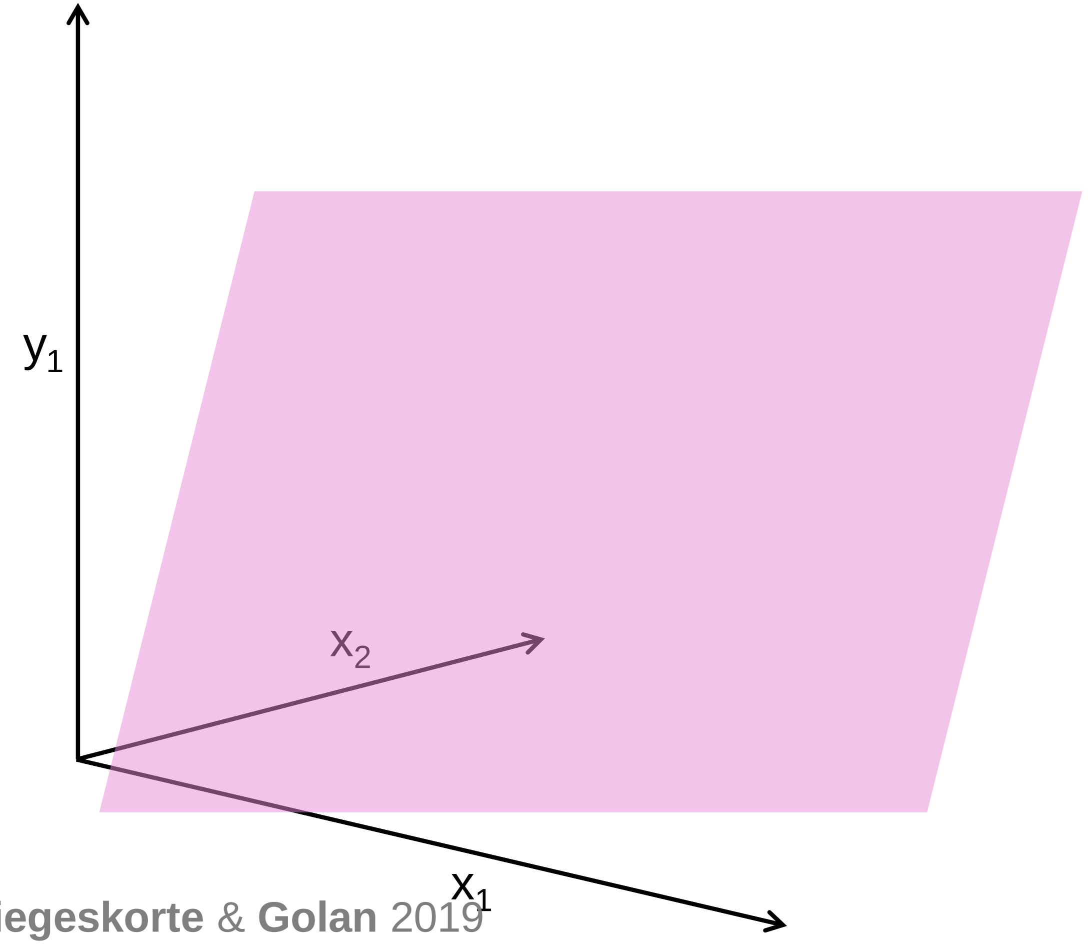


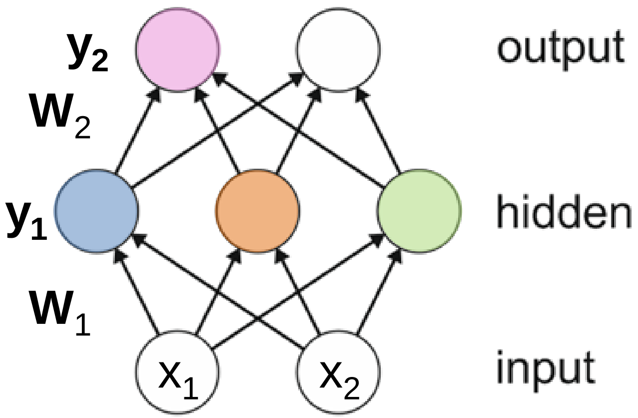
$\mathbf{y}_2 = \mathbf{x}\ \mathbf{W}_1\mathbf{W}_2$

$\mathbf{y}_2 = \mathbf{x}\ \mathbf{W'}$

Nonlinear activation function needed to make a hidden layer useful

linear activation functions





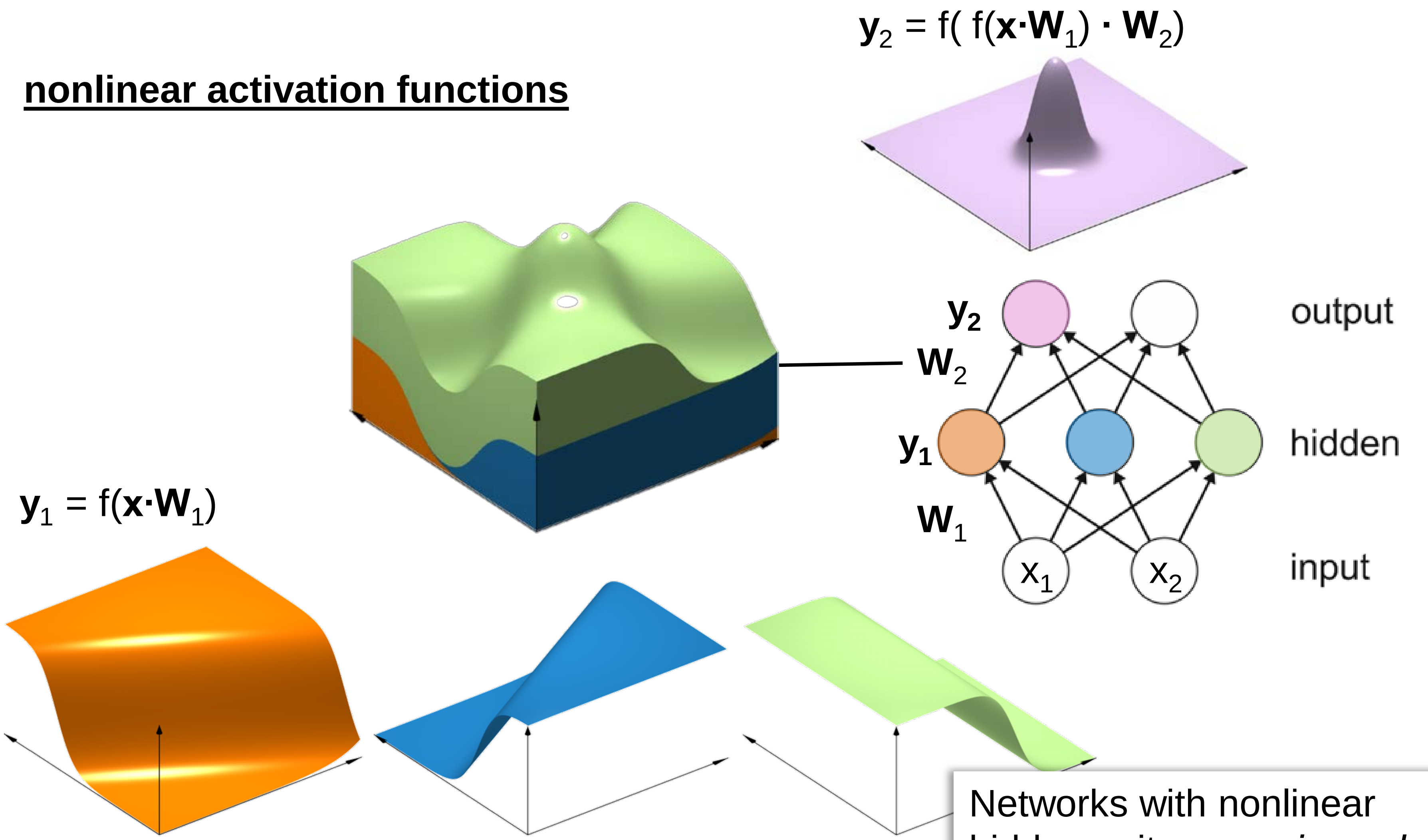
$\mathbf{y}_2 = \mathbf{x}\ \mathbf{W}_1\mathbf{W}_2$

$\mathbf{y}_2 = \mathbf{x}\ \mathbf{W'}$

Kriegeskorte & **Golan** 2019

Nonlinear activation function needed to make a hidden layer useful

nonlinear activation functions

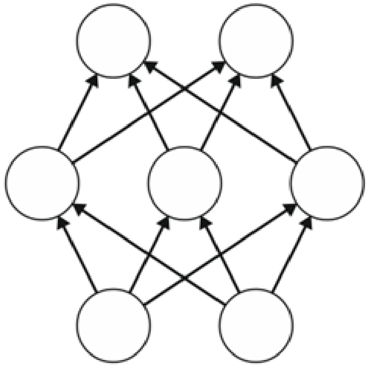


Networks with nonlinear hidden units are *universal* function approximators.

Kriegeskorte & **Golan** 2019 (primer for biologists)

Why deep?

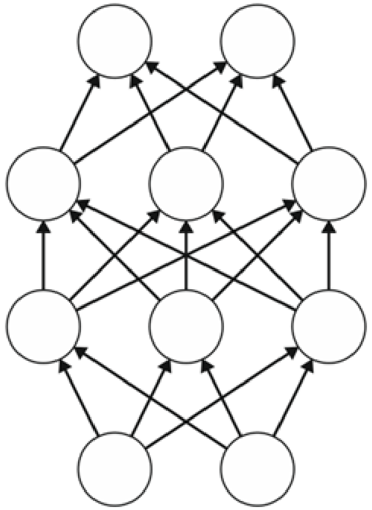
Why recurrent?



shallow

1 hidden layer

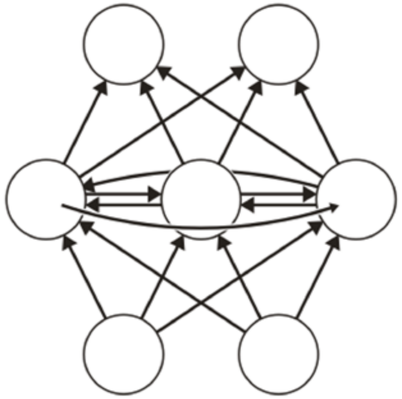
Networks with nonlinear hidden units are *universal function approximators*.



deep

>1 hidden layer

Deep nets can
- *reuse features* downstream
- represent many complex functions more concisely (fewer units and weights).



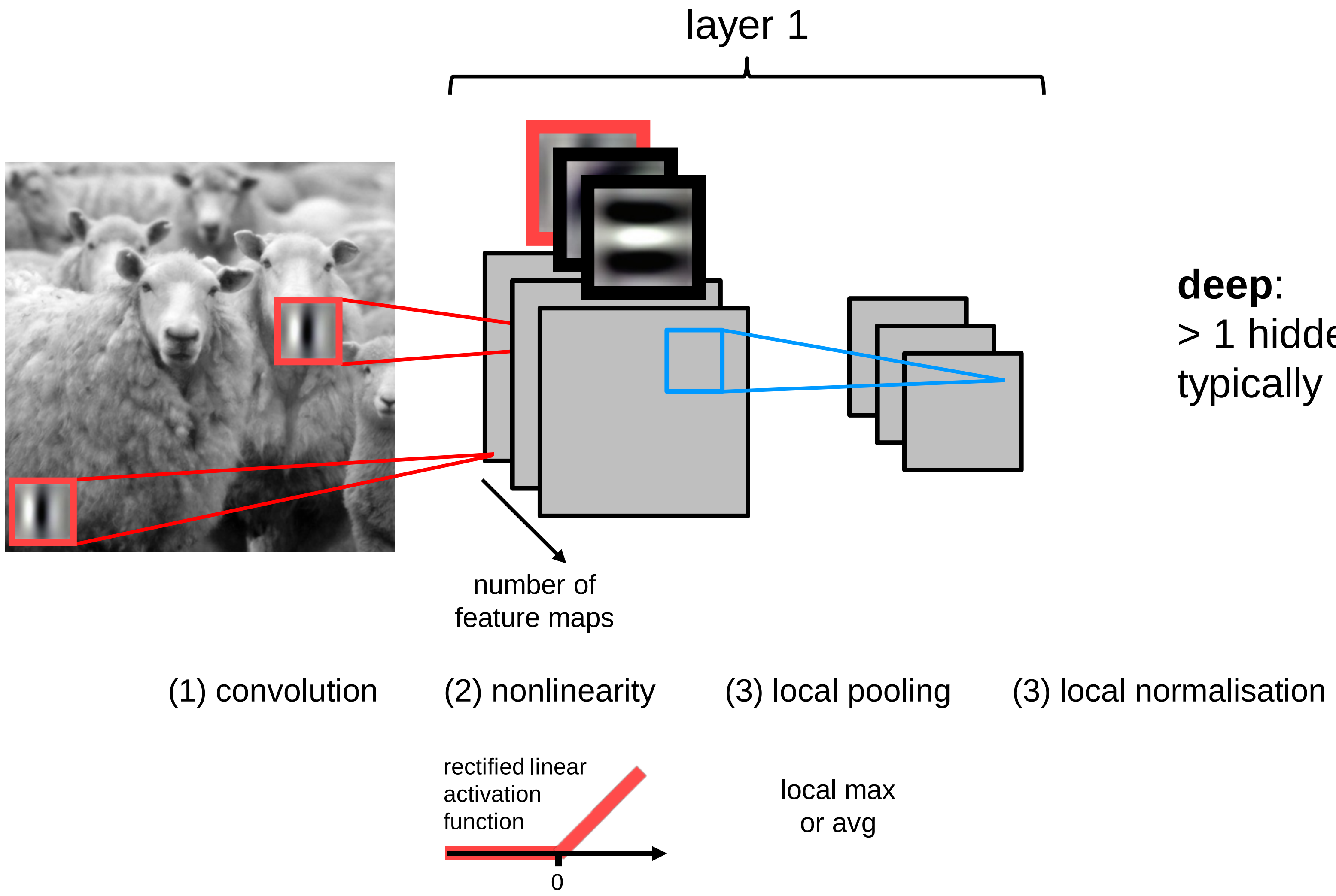
recurrent

Recurrent networks
- can *recycle weights and units* over time
- are *universal approximators of dynamical systems*.

hidden units are *universal* function approximators.

Kriegeskorte & **Golan** 2019 (primer for biologists)

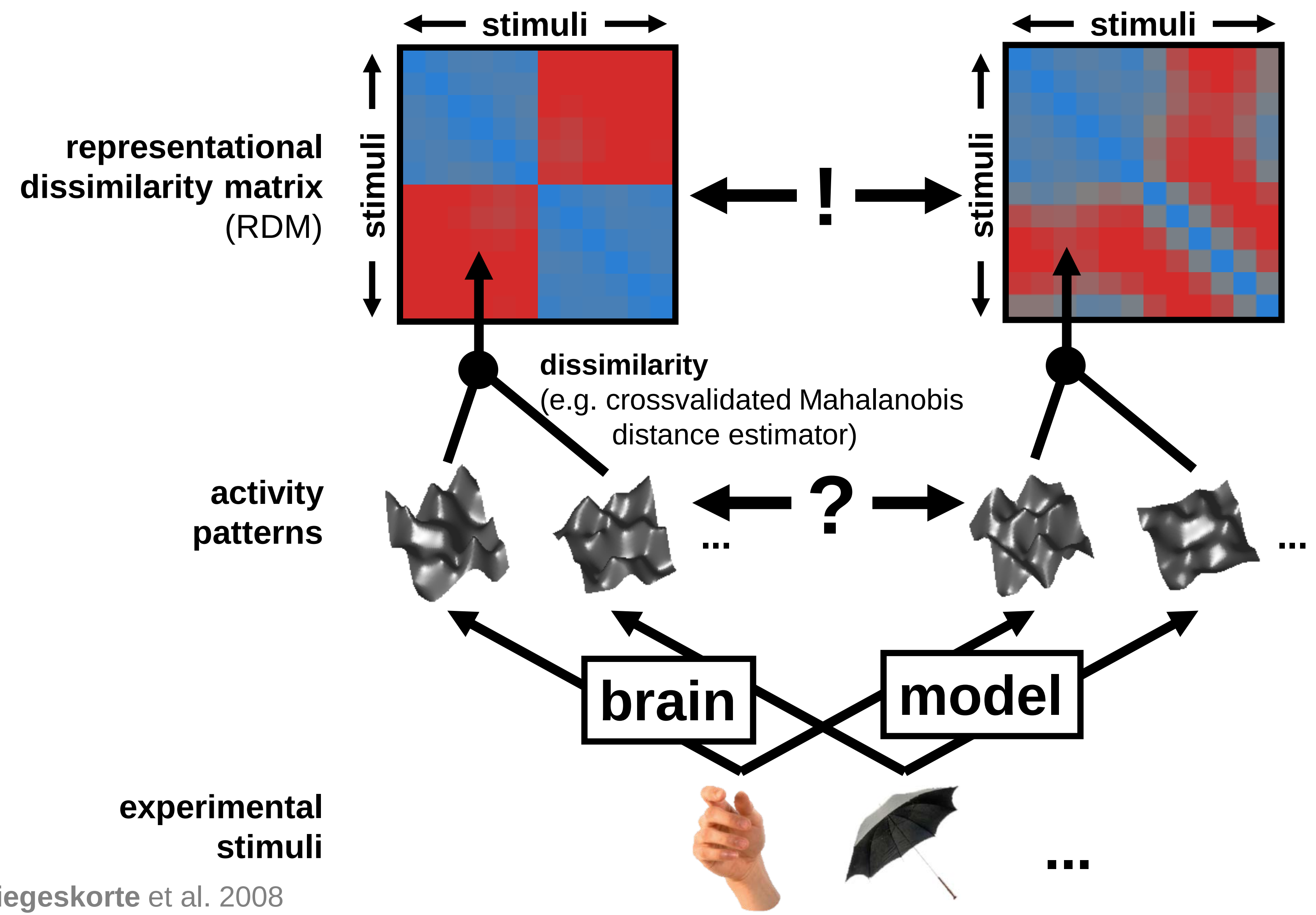
Deep *convolutional* feedforward neural networks



deep:
> 1 hidden layer,
typically many layers

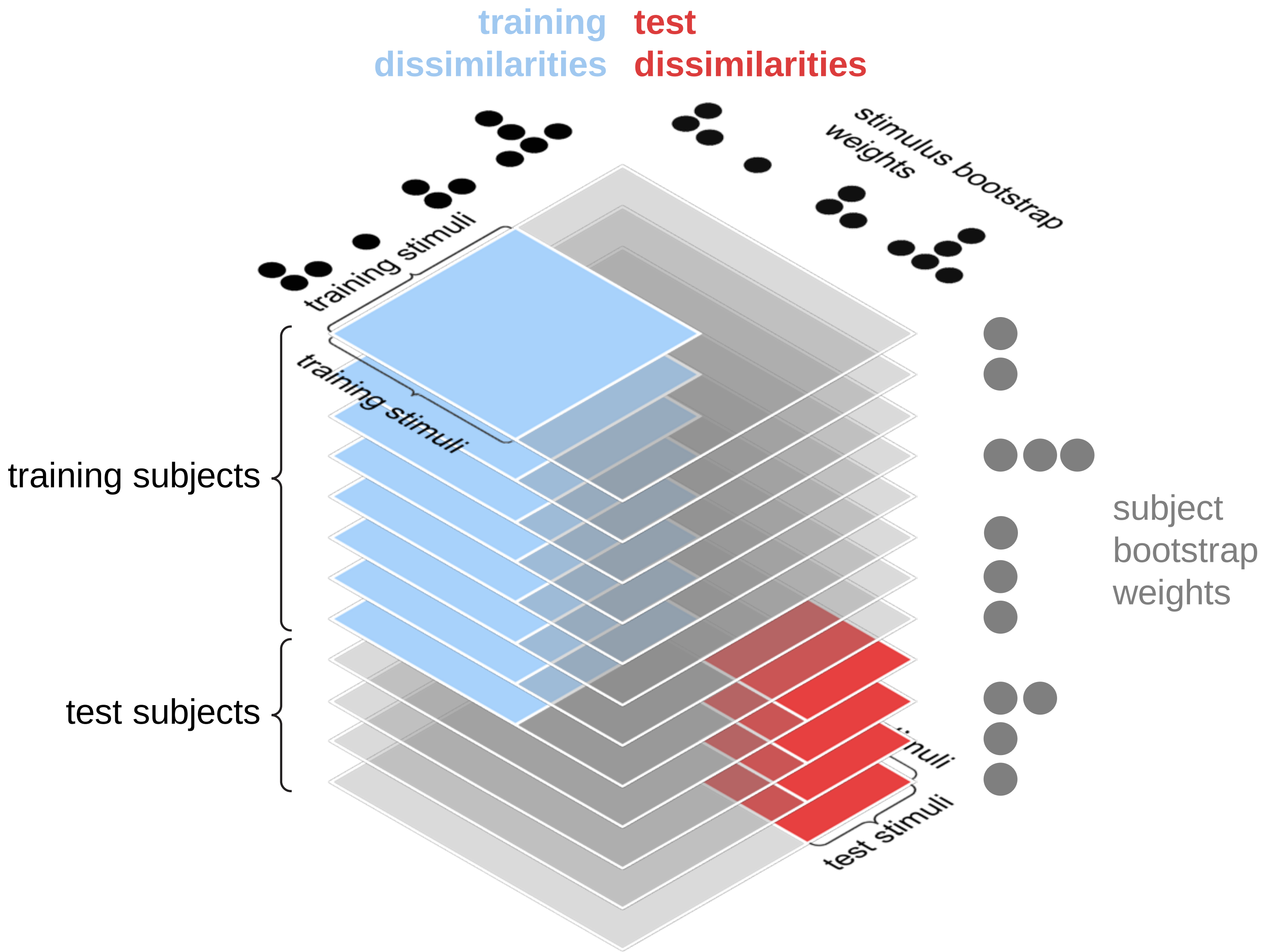
Testing neural network models with brain-activity data

Representational similarity analysis

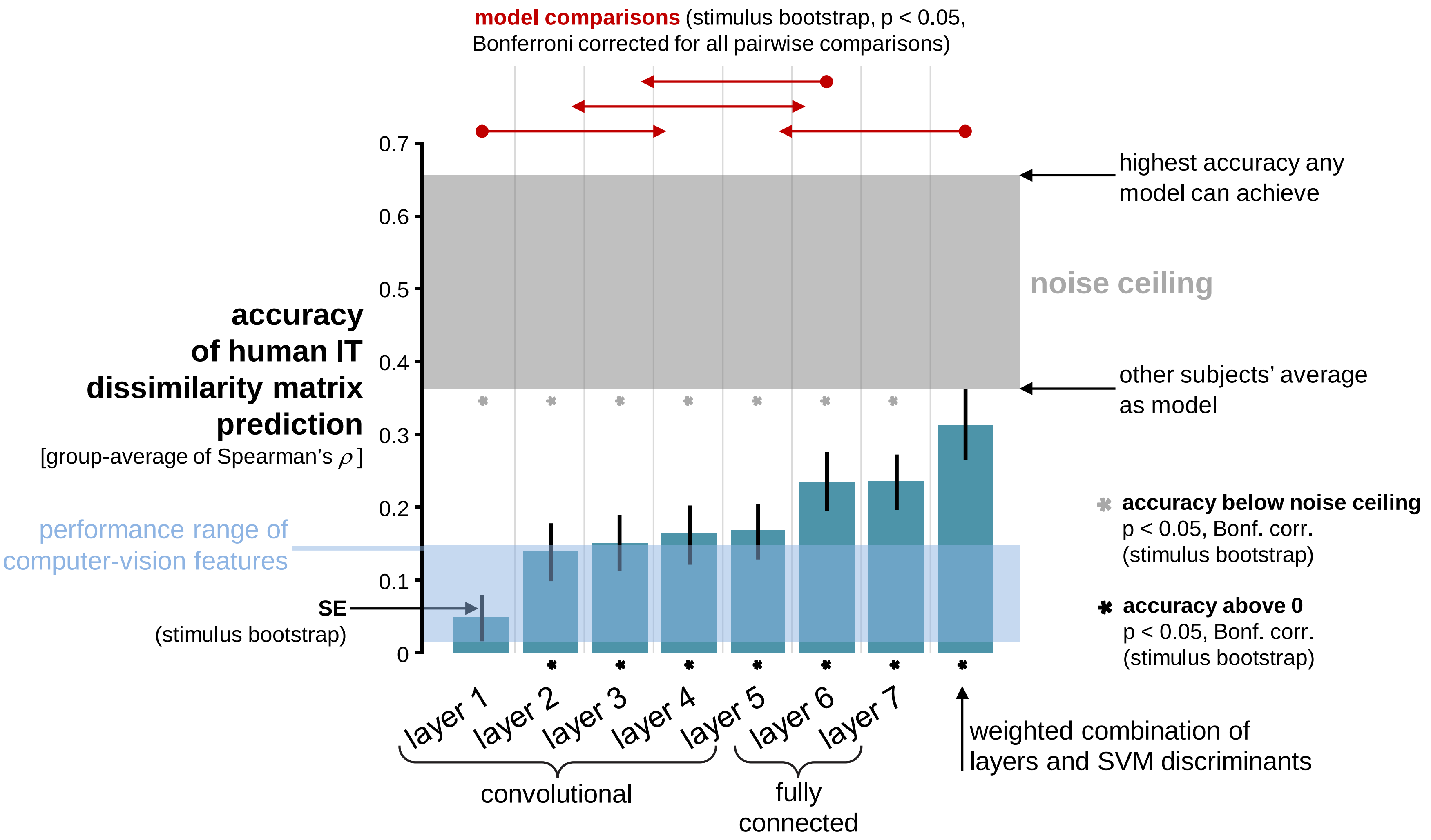


Kriegeskorte et al. 2008

Training and testing in crossvalidation

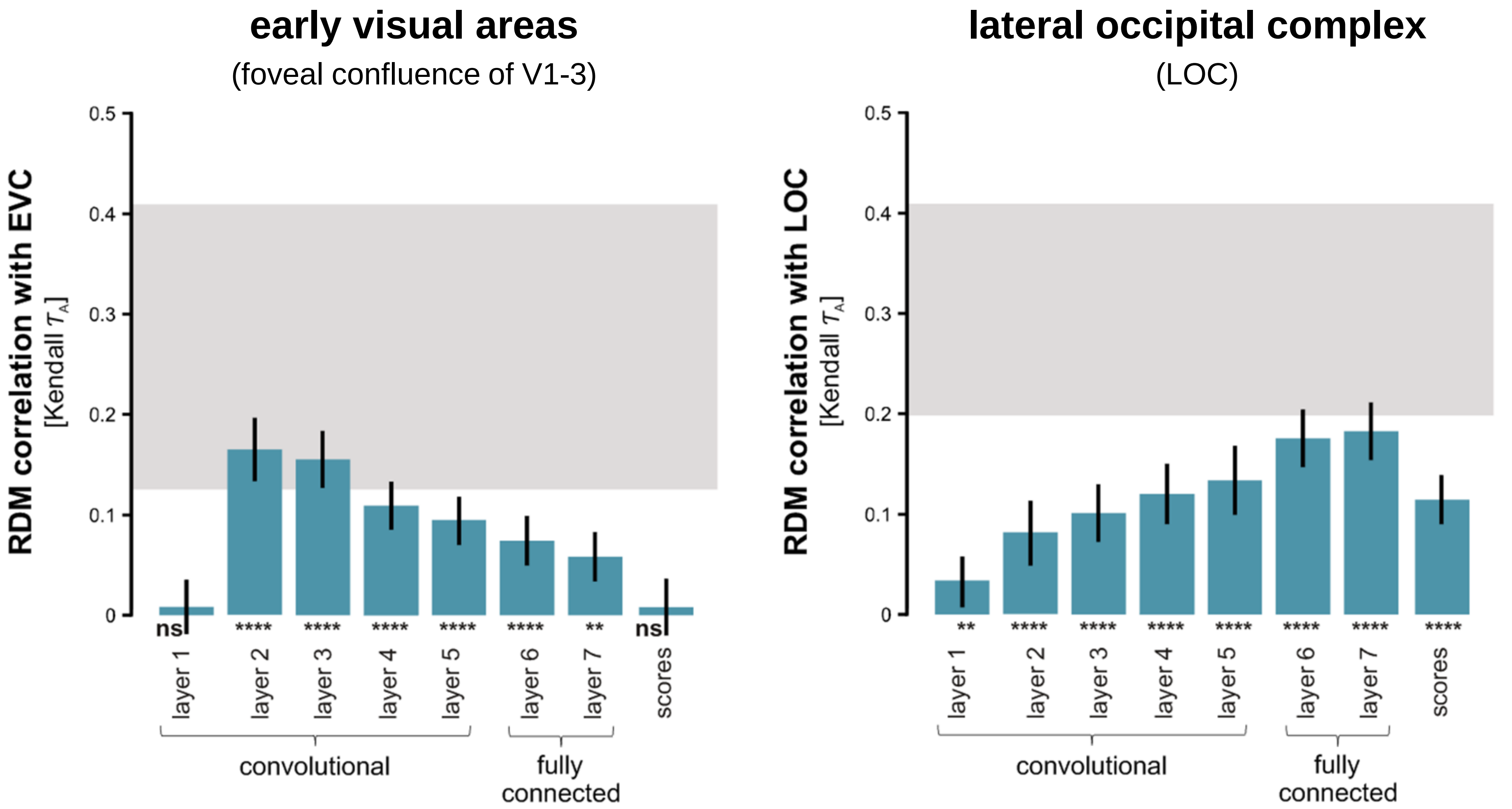


Deep convolutional feedforward networks predict IT representational geometry



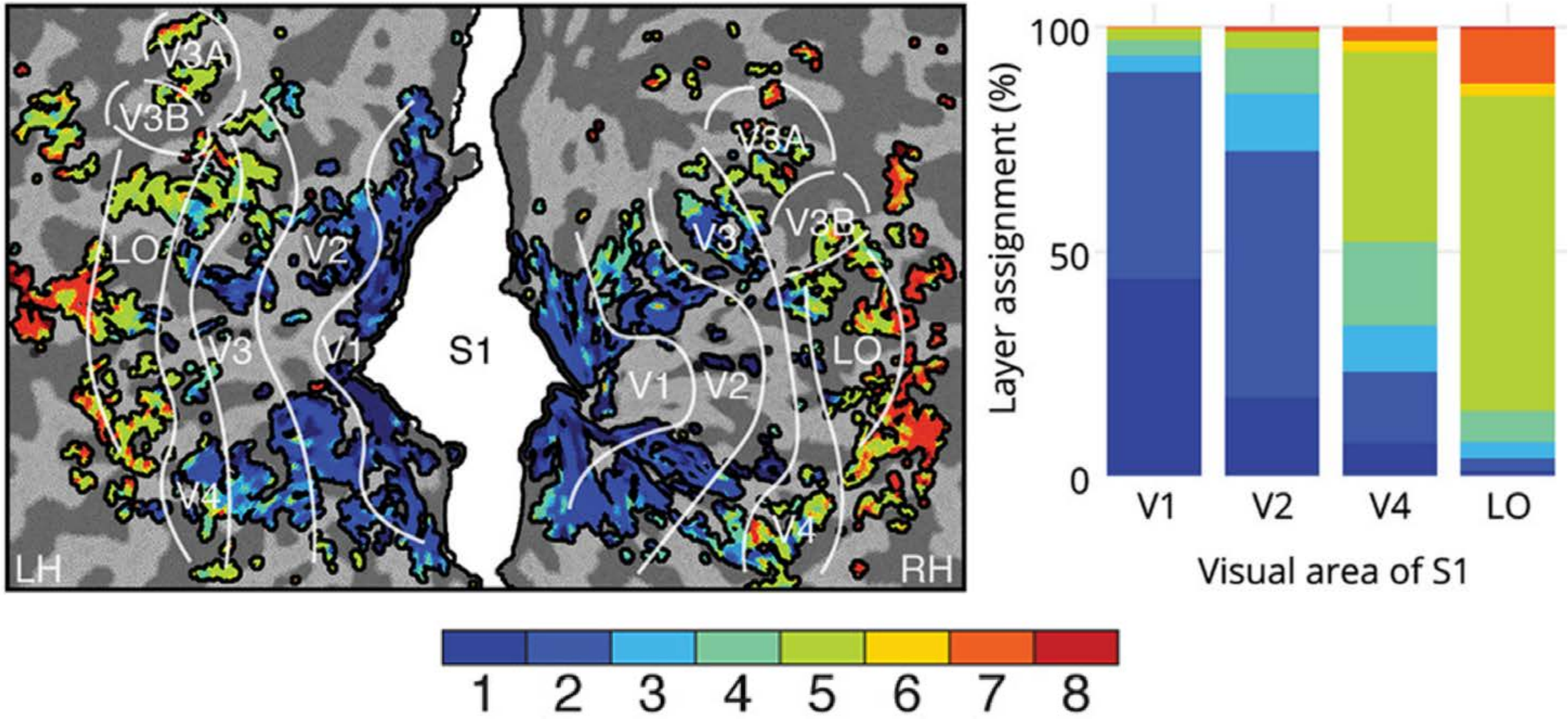
Khaligh-Razavi & Kriegeskorte 2014, Nili et al. 2014 (RSA Toolbox), Storrs et al. (in prep.)

Deep-net layers correspond to stages of the ventral visual stream



Khaligh-Razavi & **Kriegeskorte** 2014, **Nili** et al. 2014 (RSA Toolbox)

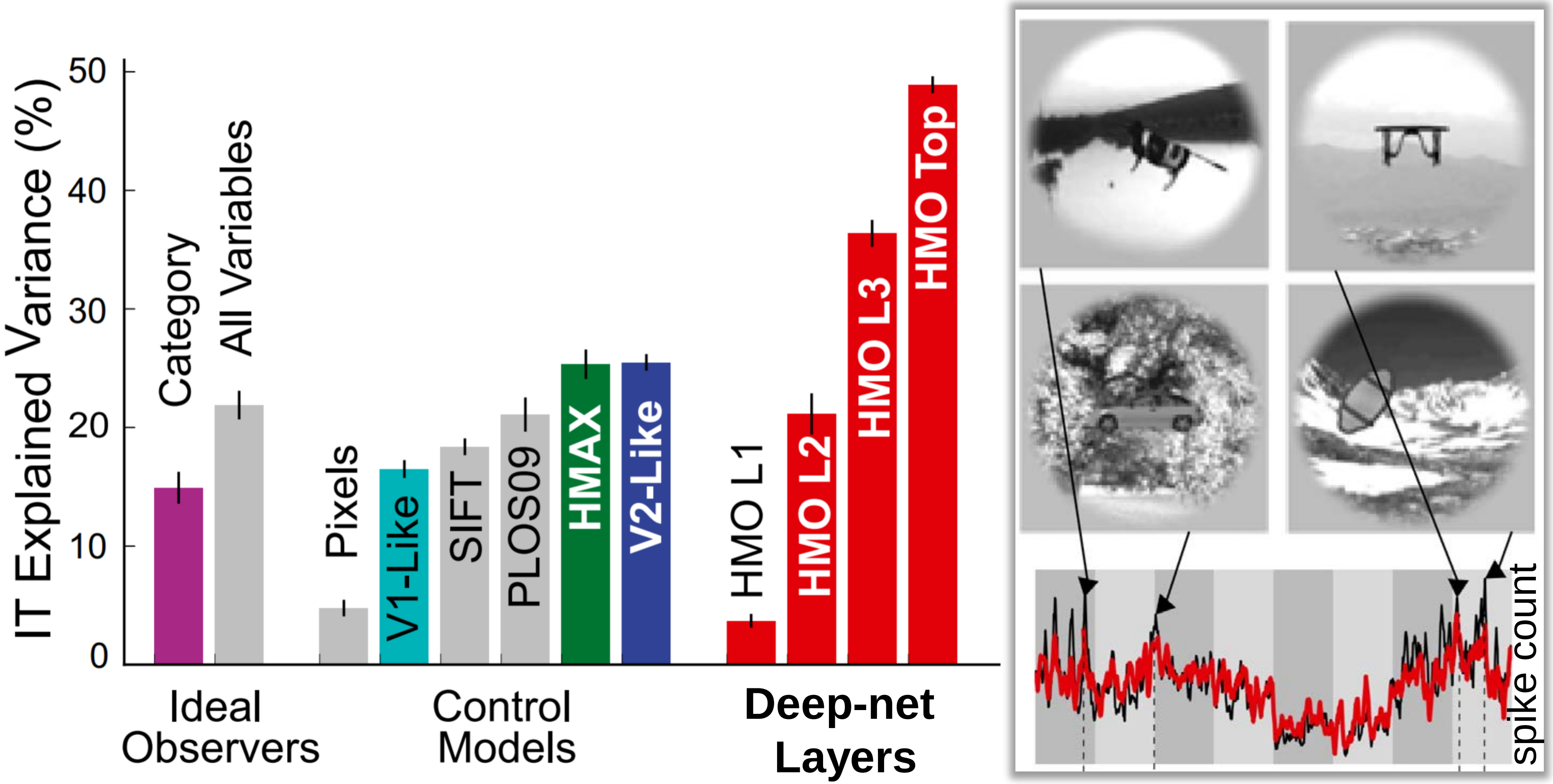
Deep-net layers correspond to stages of the ventral visual stream



deep-net layer that best explains each voxel

Güçlü & **van Gerven** 2015, data from **Kay** et al. 2008

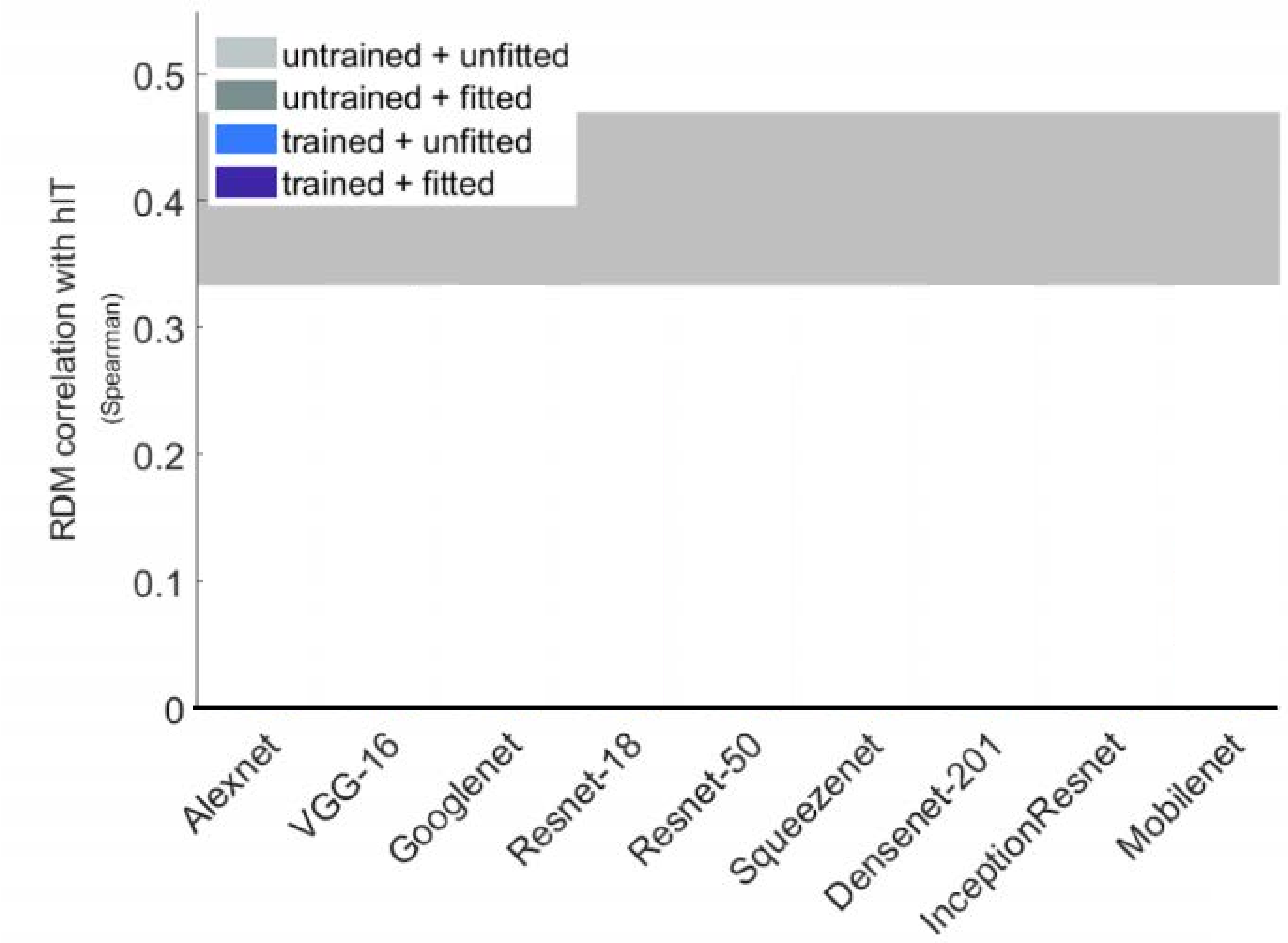
High explained variance for IT neuronal recordings



— single neuronal recording site
— deep-net predicted response for novel image

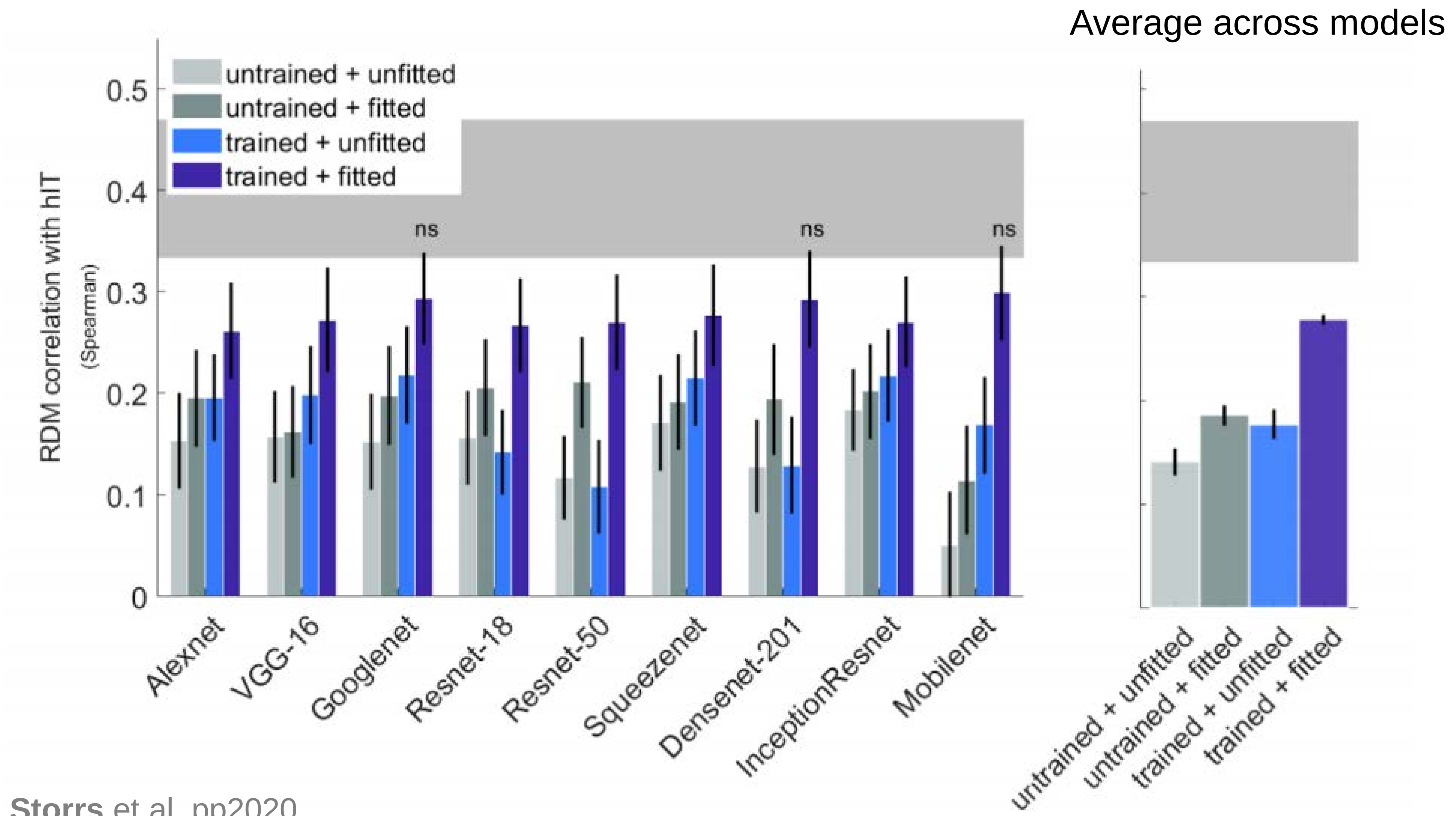
Yamins et al. 2014

Diverse deep feedforward neural networks predict IT, after *task-training* and *IT-fitting*



Storrs et al. pp2020

Diverse deep feedforward neural networks predict IT, after *task-training* and *IT-fitting*



Storrs et al. pp2020

The converging feedforward story…

- Deep convolutional feedforward neural networks explain how the initial sweep through the primate visual hierarchy enables recognition at a glance.
- They predict representations of novel images better than any alternative current models.
- Both the *architecture* of the model and the *task training* contribute substantially to these successes.

However, we need to build models whose architecture more closely resembles the visual hierarchy.

A major feature of biological neural networks is recurrent signal flow.

Overview

1. Recurrent neural network models

2. Controversial stimuli

Overview

1. Recurrent neural network models

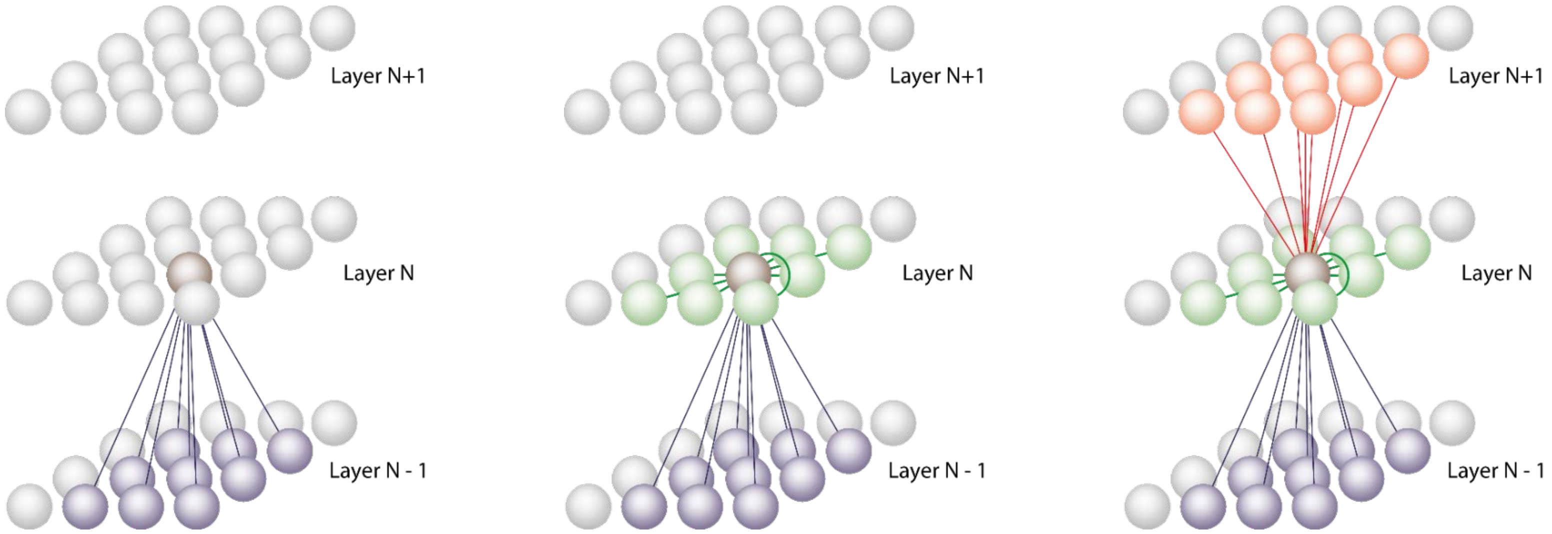
2. Controversial stimuli

Do *recurrent* convolutional neural networks provide better models of vision?

Courtney Spoerer

Recurrent convolutional neural networks

B (FF-CNN) | BL (RCNN) | BLT (RCNN)

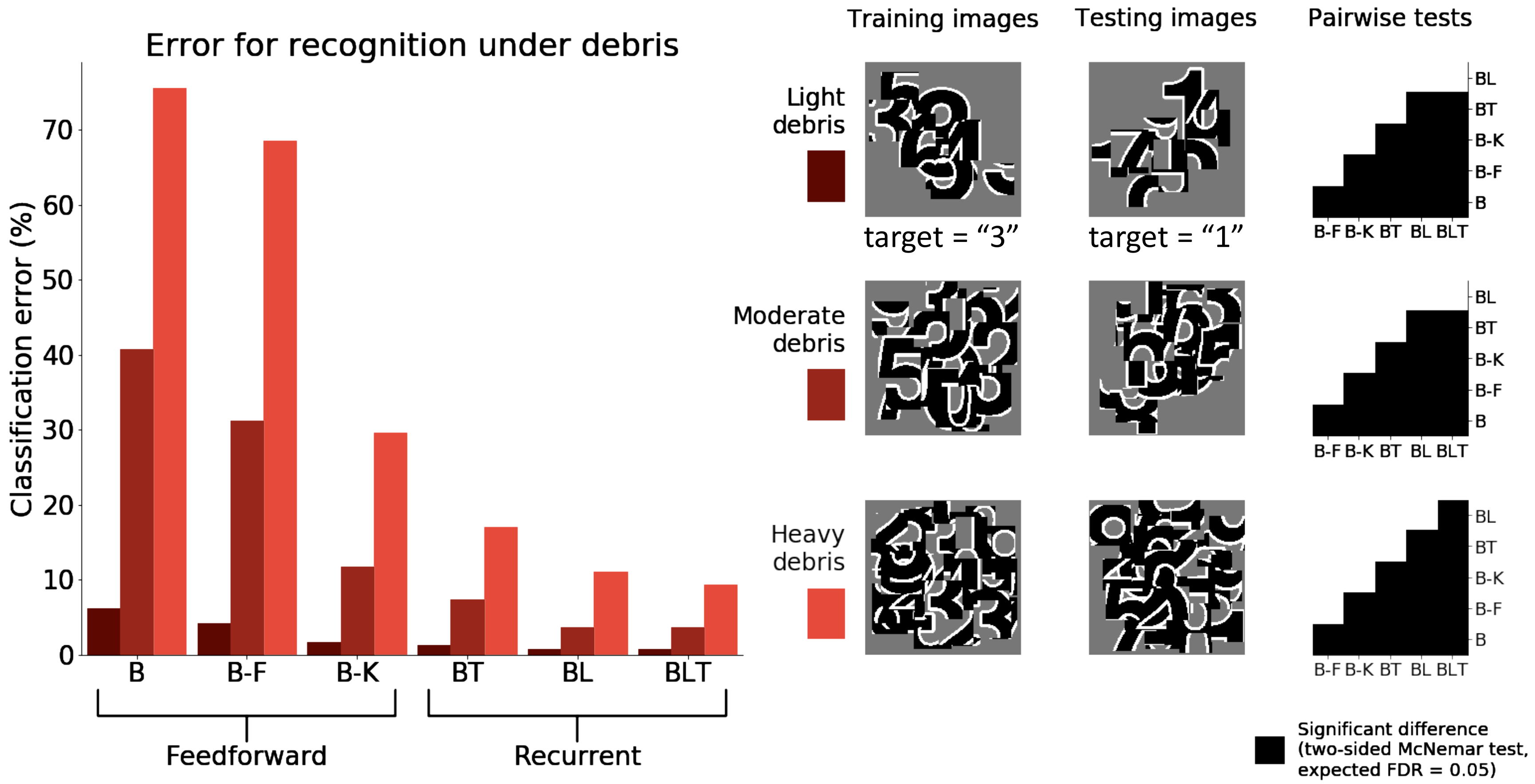


$\sigma(h_B)$ | $\sigma(h_B + h_L)$ | $\sigma(h_B + h_L + h_T)$

σ - non-linearity
h_x - convolution

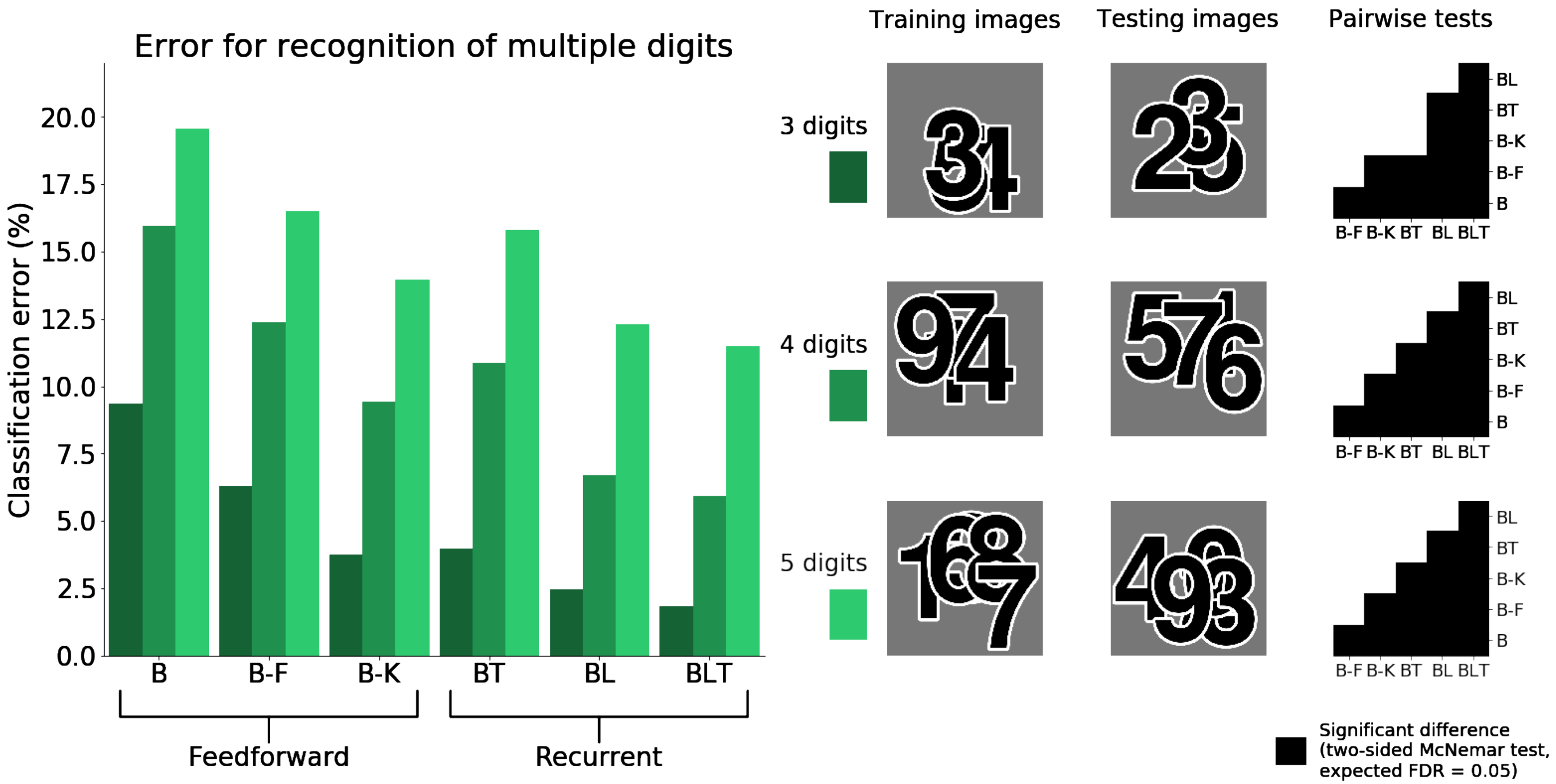
Liang & **Hu** 2015, **Spoerer** et al. 2017

***Digit debris*:**
recognition under occlusion



Spoerer et al. 2017

Digit clutter: Multiple digit recognition

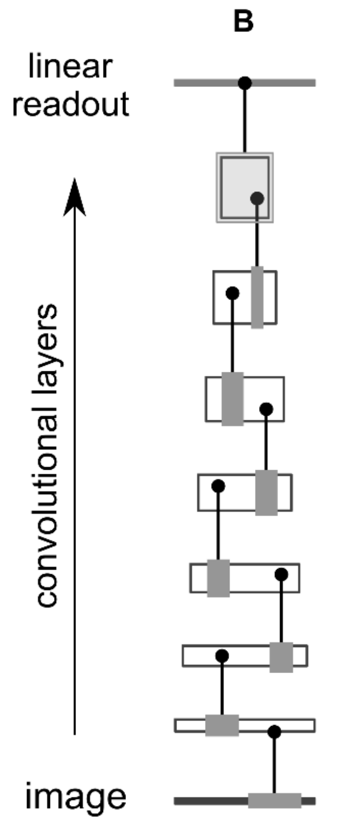


Spoerer et al. 2017

Can recurrent convolutional networks be scaled up to process natural images?

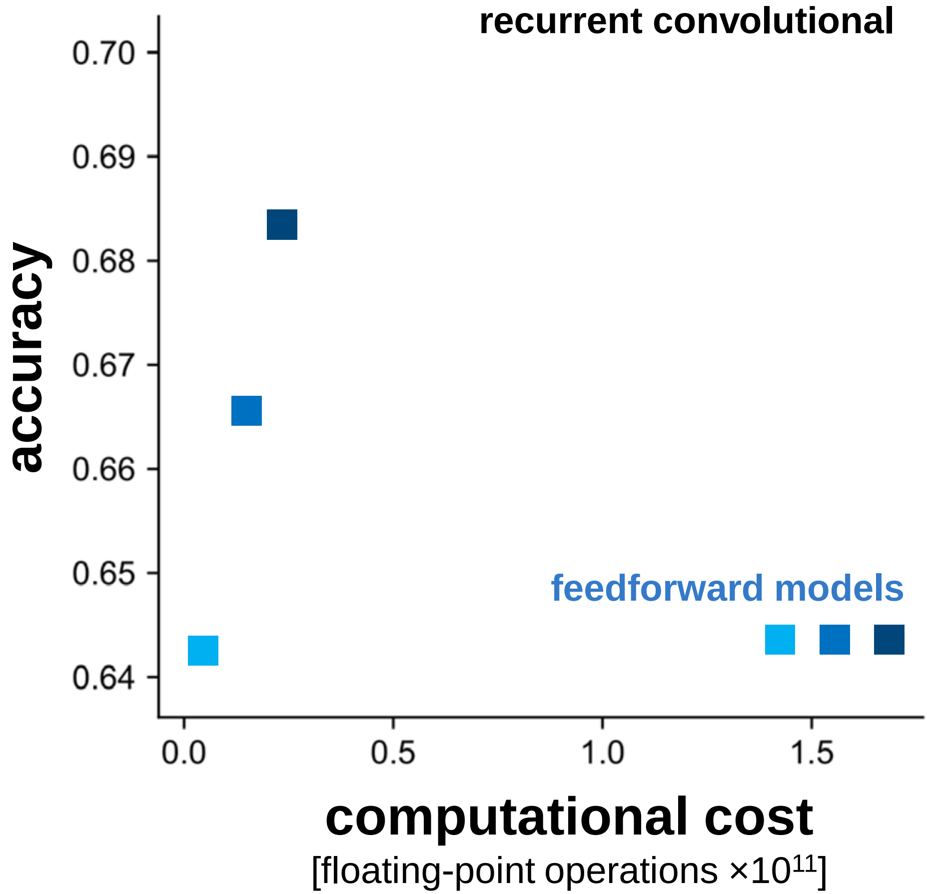
Recurrent convolutional networks trained to recognize natural images

feedforward



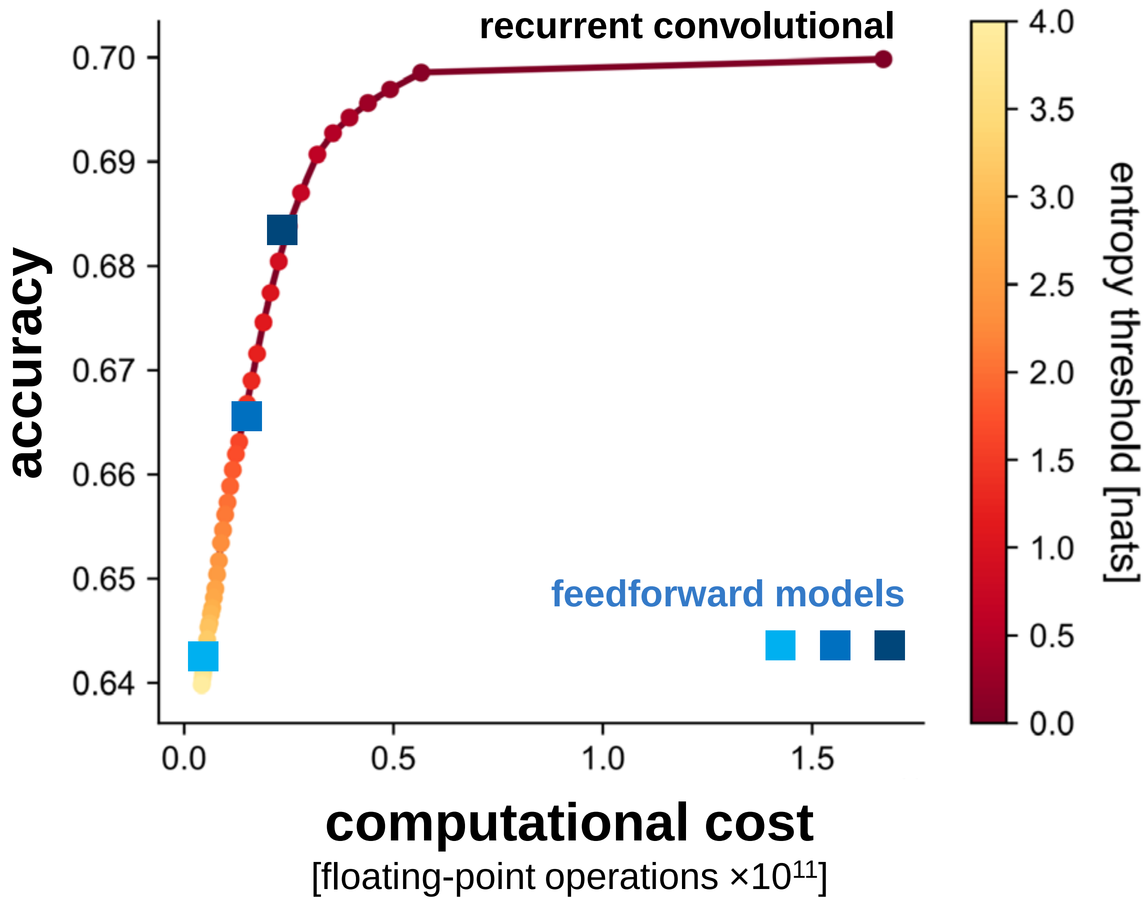
Spoerer et al. 2020

Recurrent models can trade off speed of computation for accuracy



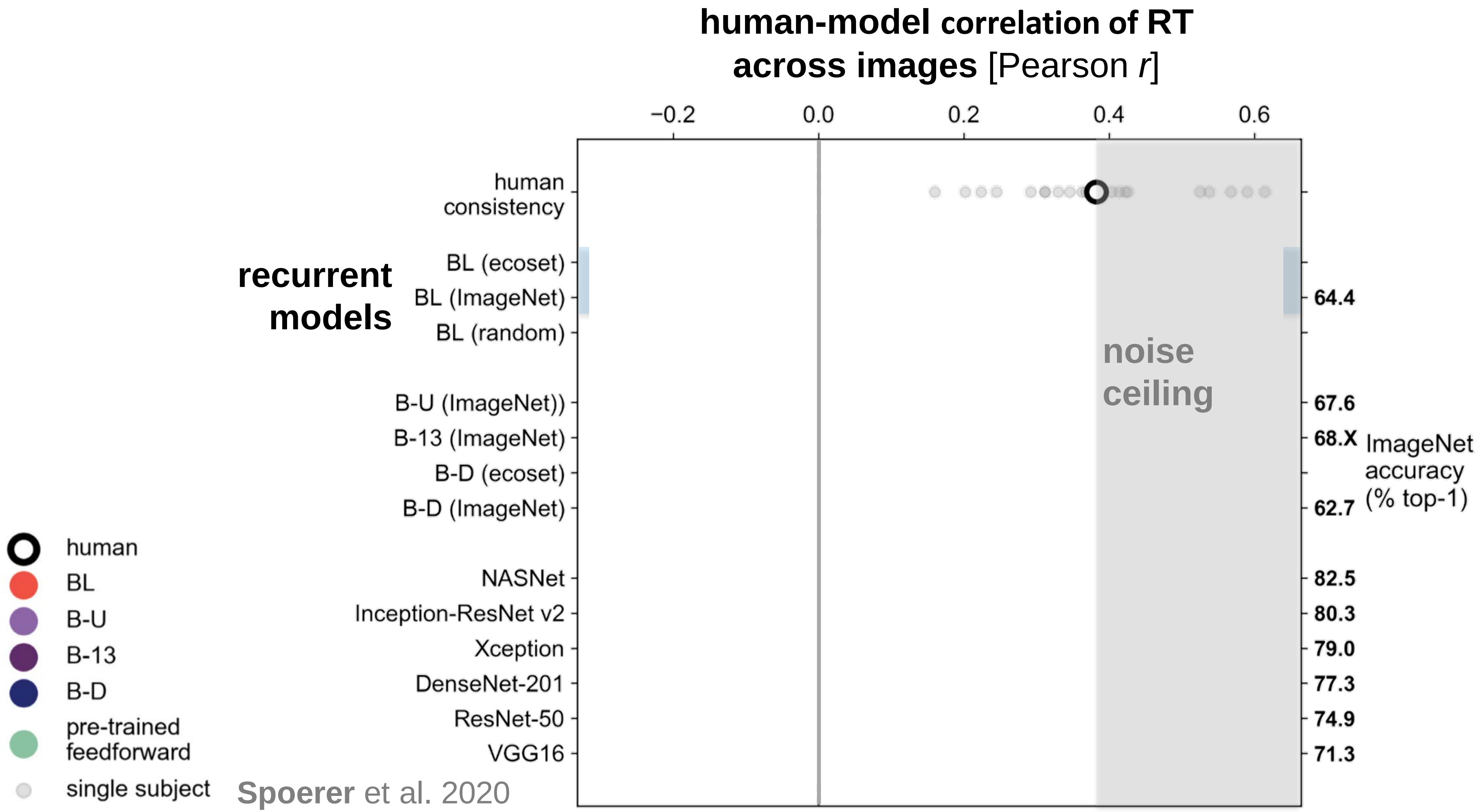
Spoerer et al. 2020

Recurrent models can trade off speed of computation for accuracy



Spoerer et al. 2020

RCNNs predict human reaction times



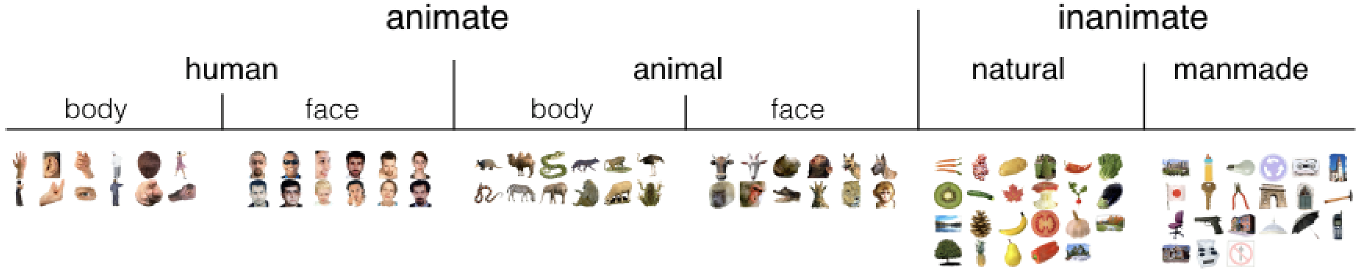
Spoerer et al. 2020

Tim Kietzmann

Can recurrent neural network models capture the representational dynamics in the human ventral stream?

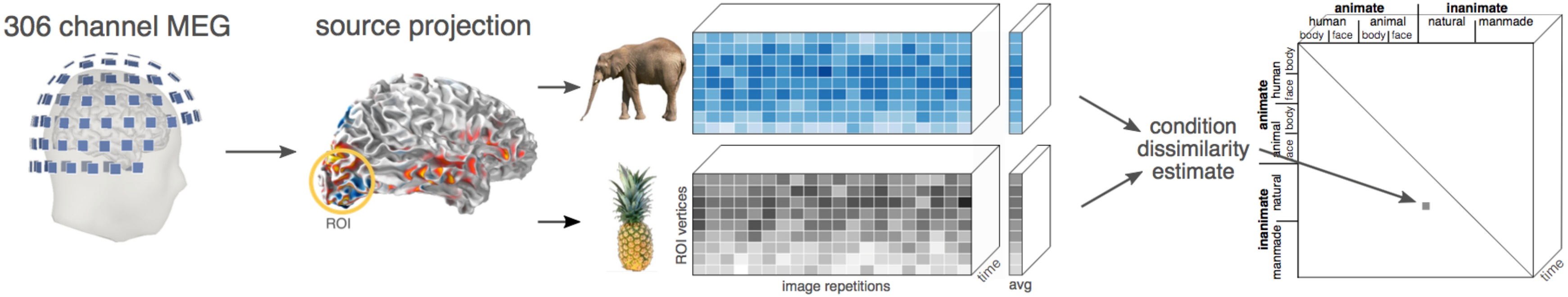
Representational dynamics

stimuli



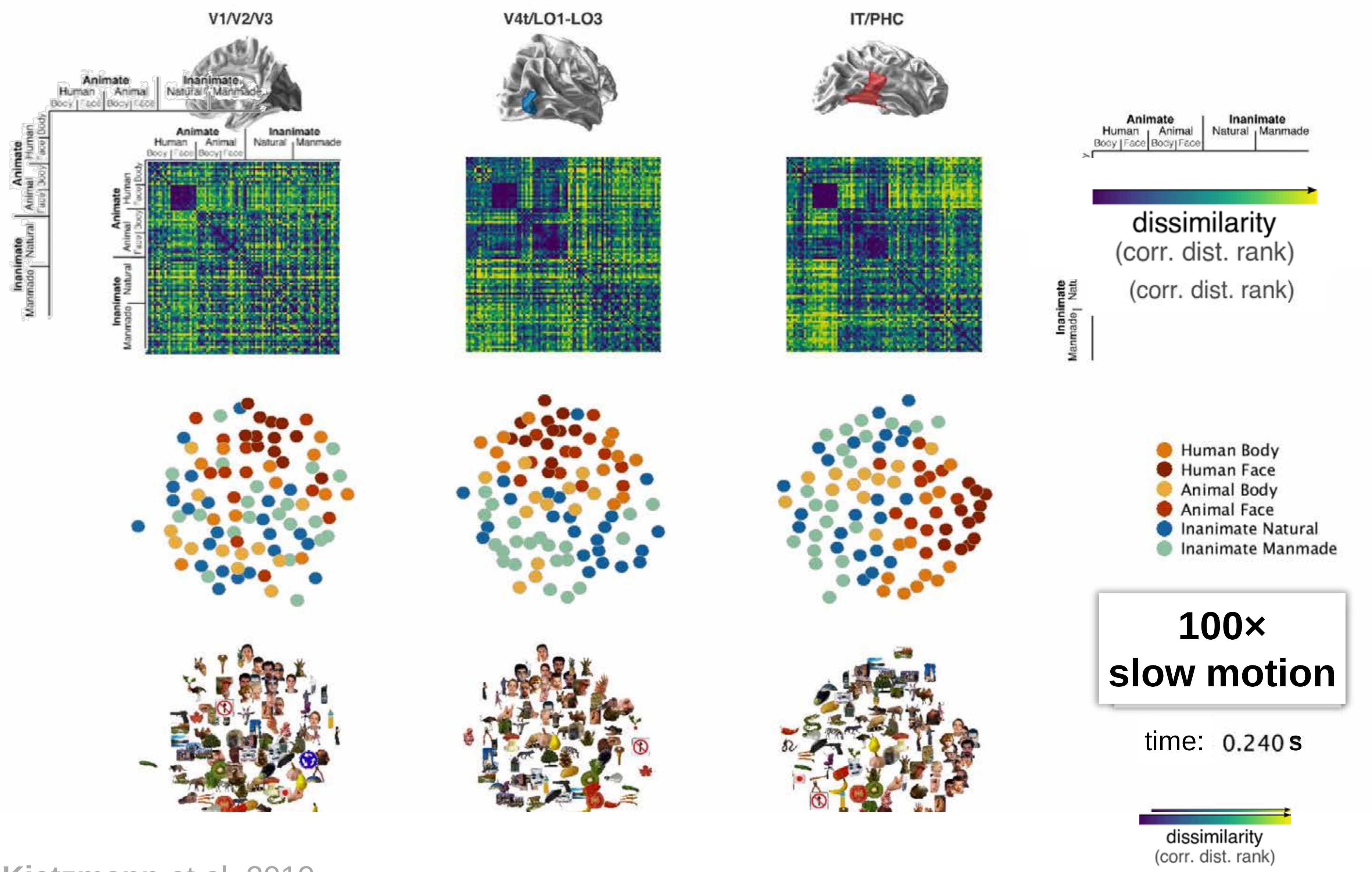
brain regions

human magnetoencephalography



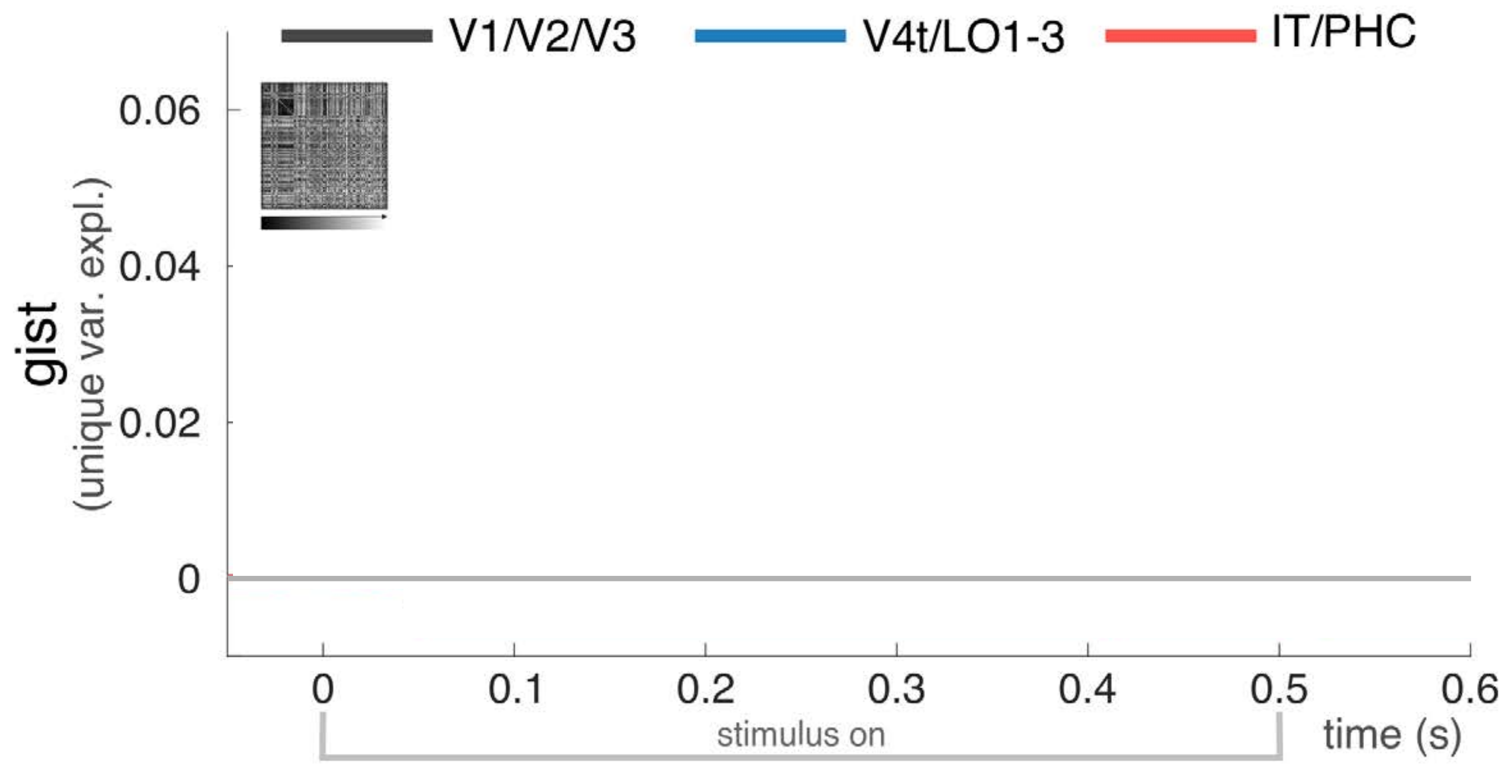
Kietzmann et al. 2019, **Cichy** et al. 2015

Movie time



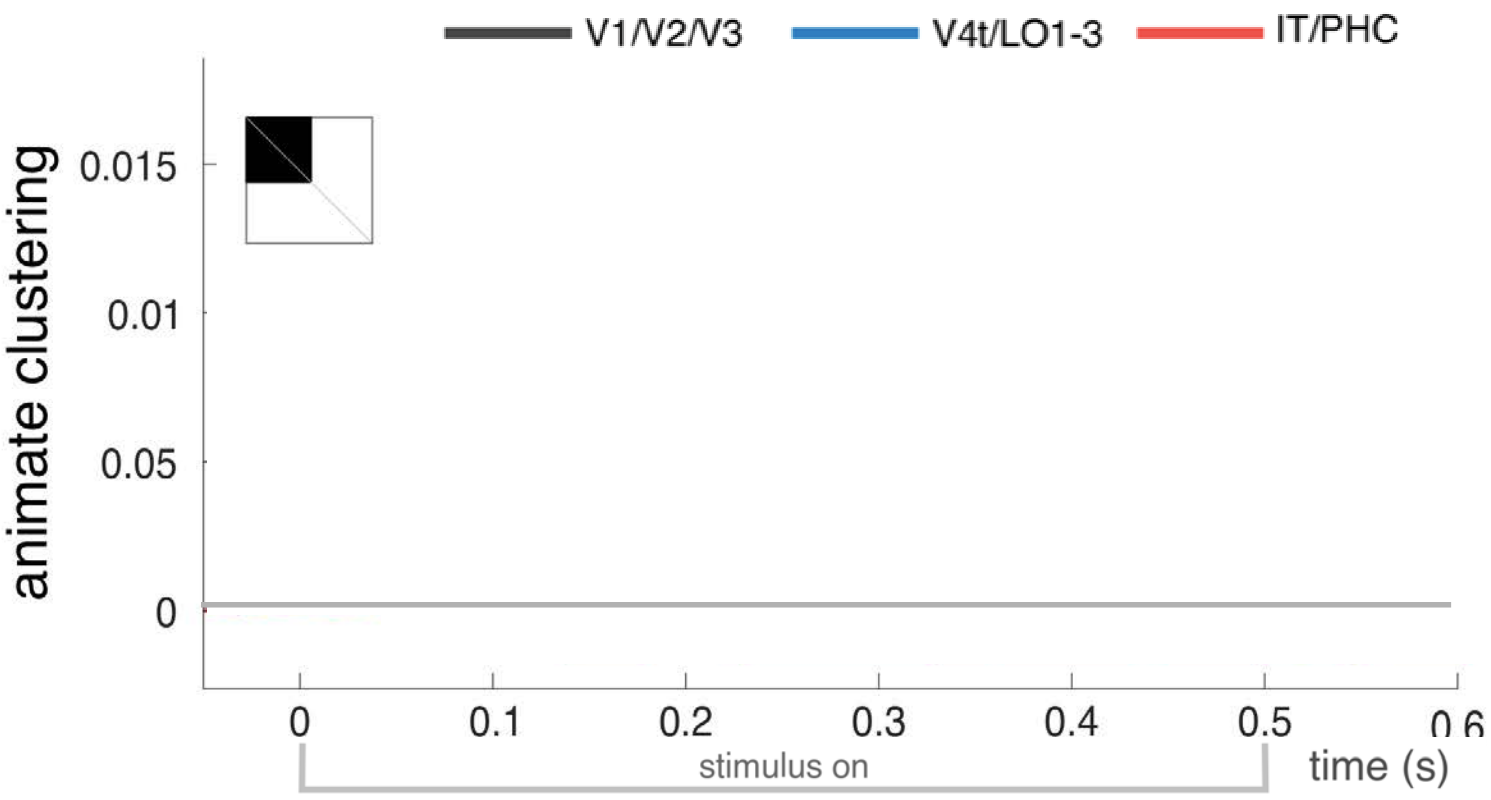
Kietzmann et al. 2019

Low-level features: gist model



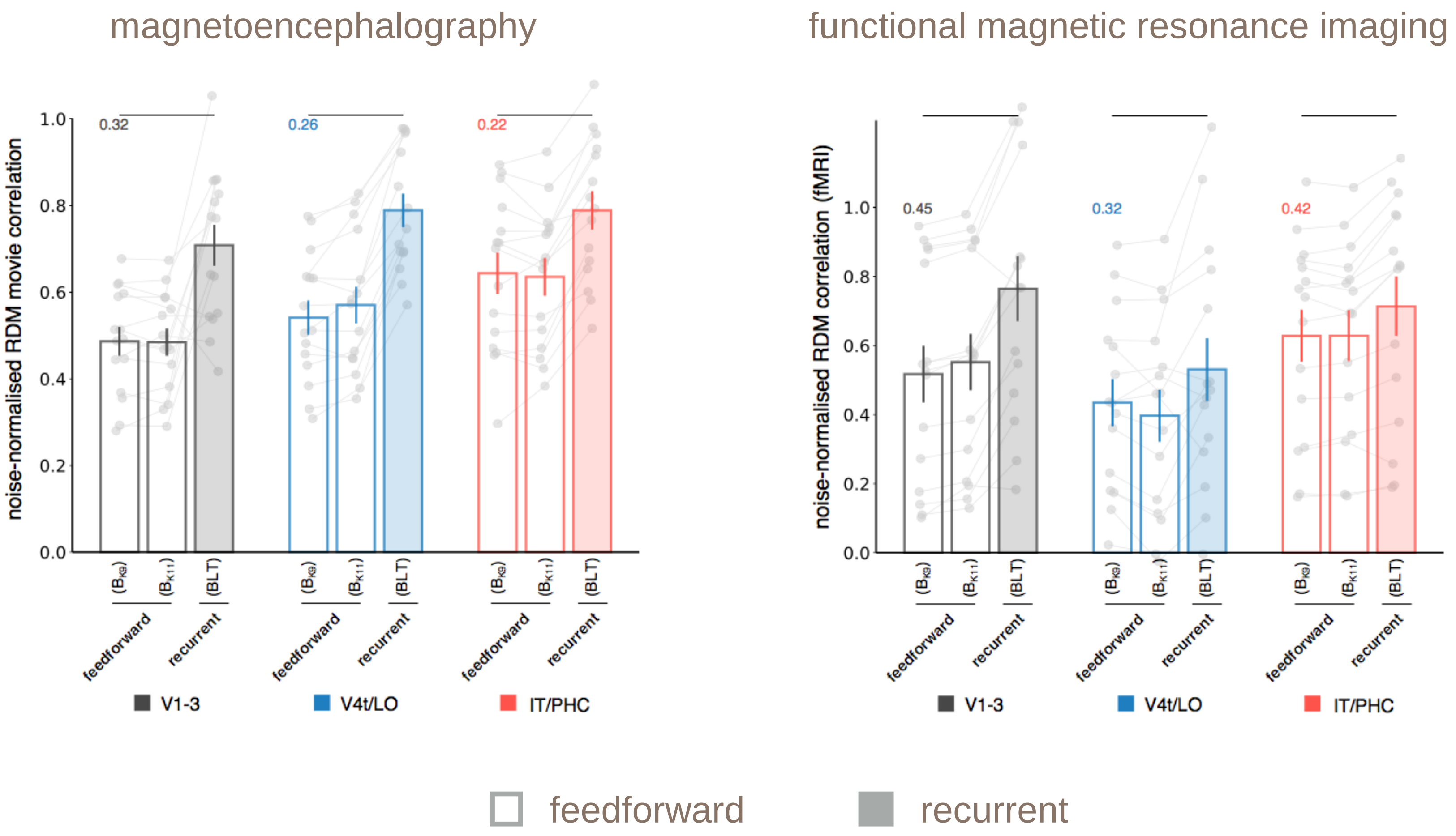
Gist-like geometries first emerge in early visual areas,
where they remain stronger throughout.

Categorical clustering: animacy



Animacy emerges first in IT/PHC, and only later in V4t/LO1-3.

Recurrent models better explain representations and their dynamics



Kietzmann et al. 2019

The emerging recurrent story…

- Recurrent neural networks provide a **more neurobiologically realistic and computationally powerful** modeling framework.
- Recurrent processing can enable a network to
  - **recycle its computational resources**,
  - perform **more robust inferences**, and
  - **flexibly trade off speed and accuracy**.
- Recurrent models also **better explain the representational dynamics** of the human ventral stream.

Controversial stimuli:
pitting neural networks
against each other
as models of
human recognition

Golan et al. 2020

Controversial stimuli

Golan et al. 2020

Controversial stimuli: motivation

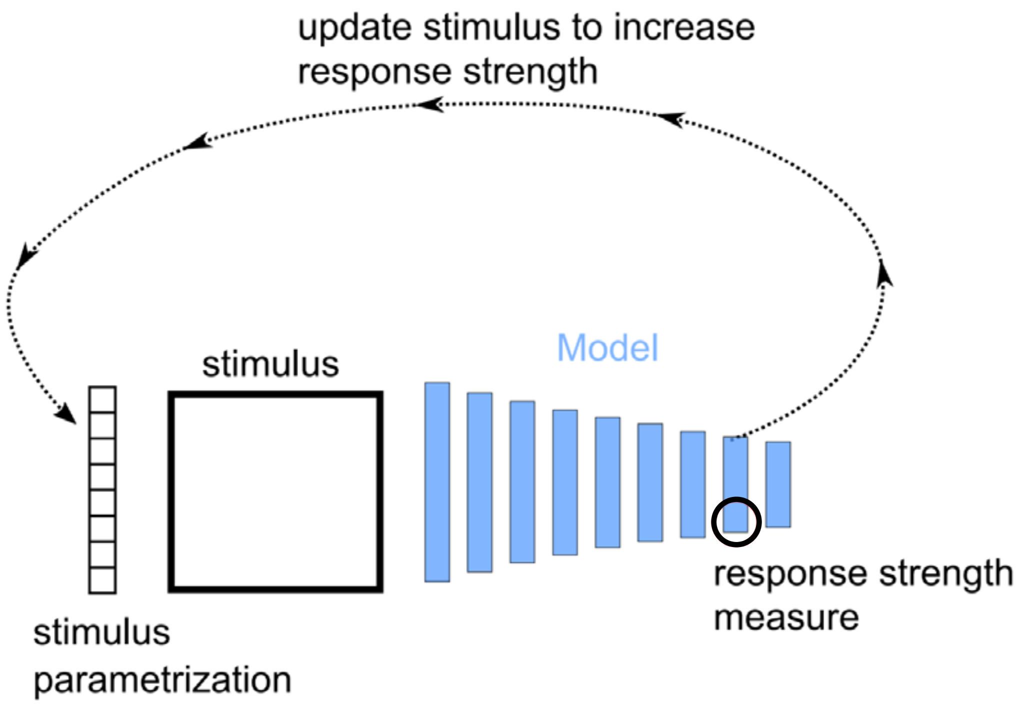
- Theoretical progress depends on experiments for which competing theories make distinct predictions.
- We can implement competing theories in testable NN models.
- However, NN models have many parameters, and theoretically distinct models often make similar predictions for natural stimuli.

Insight 1: To elicit models' distinct *inductive biases* we can test models on a population of stimuli not used in training (*out of distribution*).
  - natural stimuli drawn from a different stimulus population
  - synthetic stimuli (optimized to elicit bolder predictions, e.g. superstimuli, adversarial stimuli, and metamers)

Insight 2: Since our goal is to adjudicate among models, we can create synthetic stimuli optimized to elicit distinct predictions from different models: stimuli that are ***controversial*** among the models.

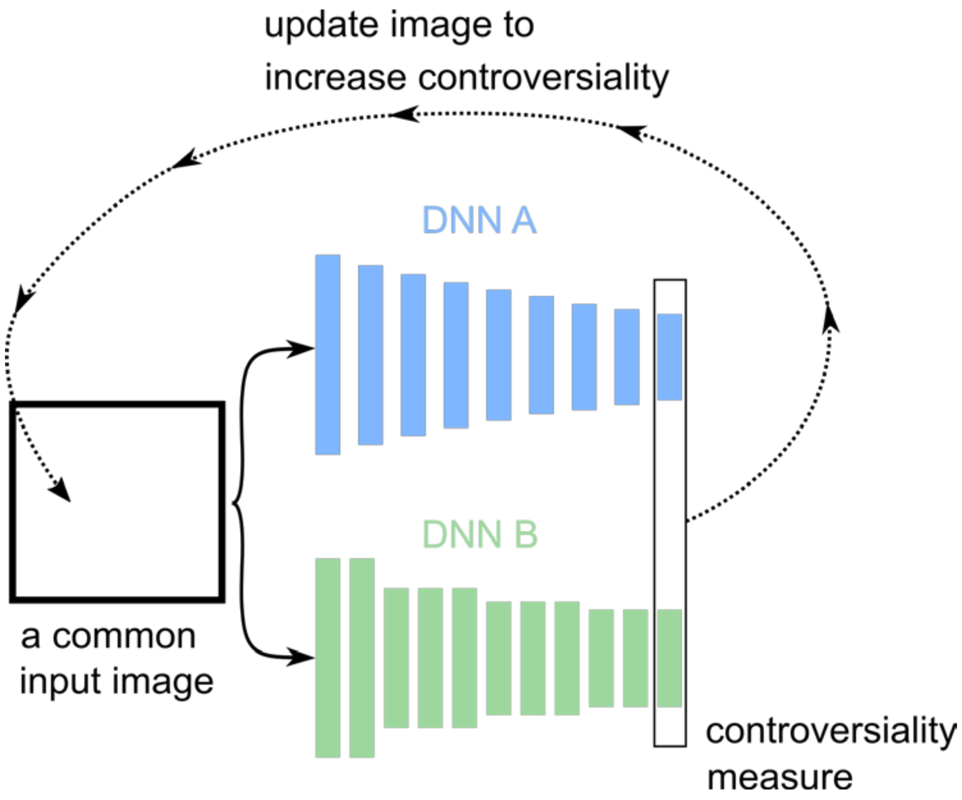
Golan et al. 2020, **Wang** & **Simoncelli** 2008

Superstimulus



Abbasi-Asl et al. 2018, **Malakhova** 2018, **Ponce** et al. 2019, **Bashivan** et al. 2019, **Walker** et al. 2019

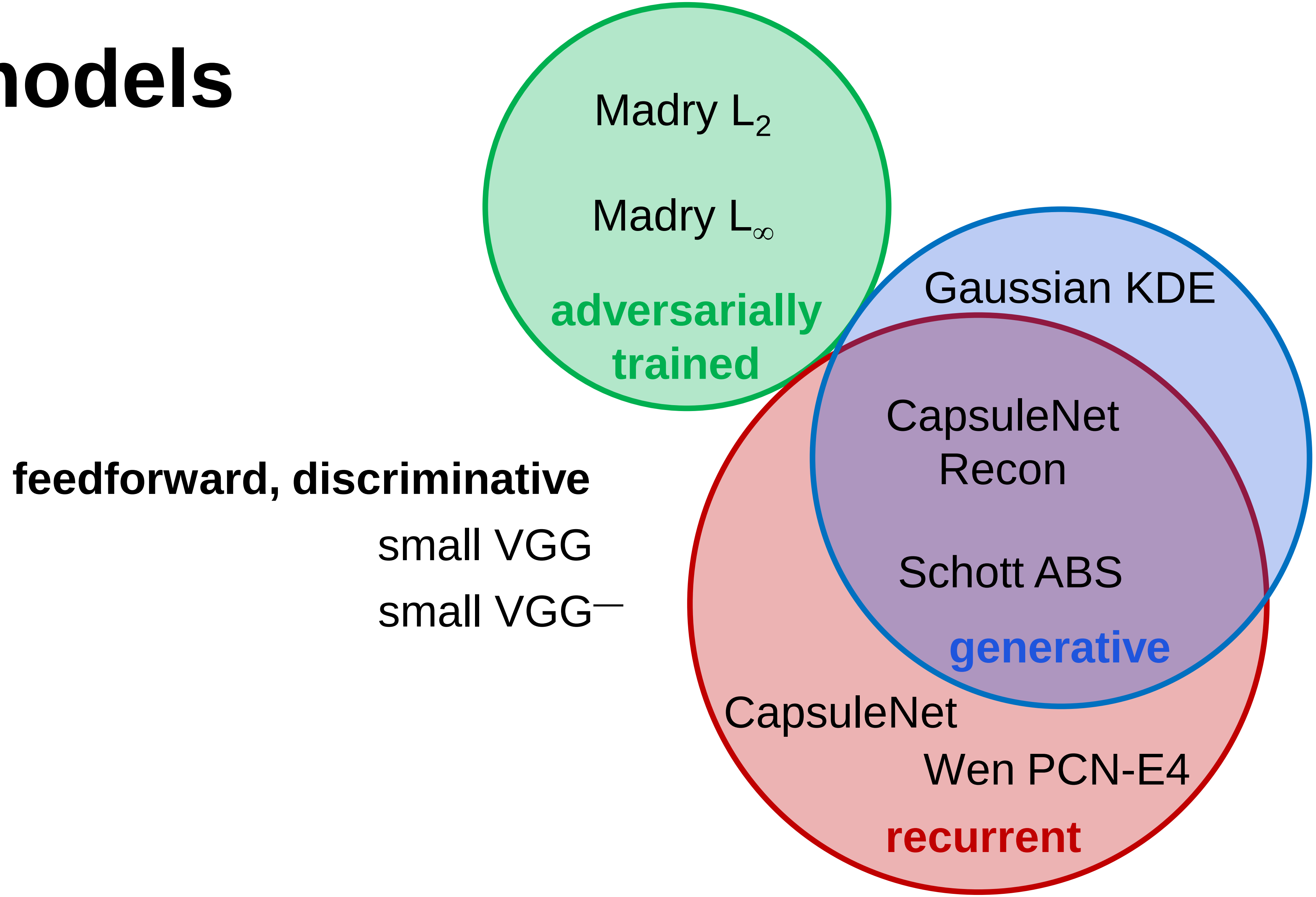
Controversial stimulus



Golan et al. 2020

MNIST

Tested models



Controversiality index

digit pair

$$c^{d_a,d_b}_{A,B}(x)$$

model pair

image

model **A** detects digit **a** **and** not digit **b**

and model **B** detects digit **b** **and** not digit **a**

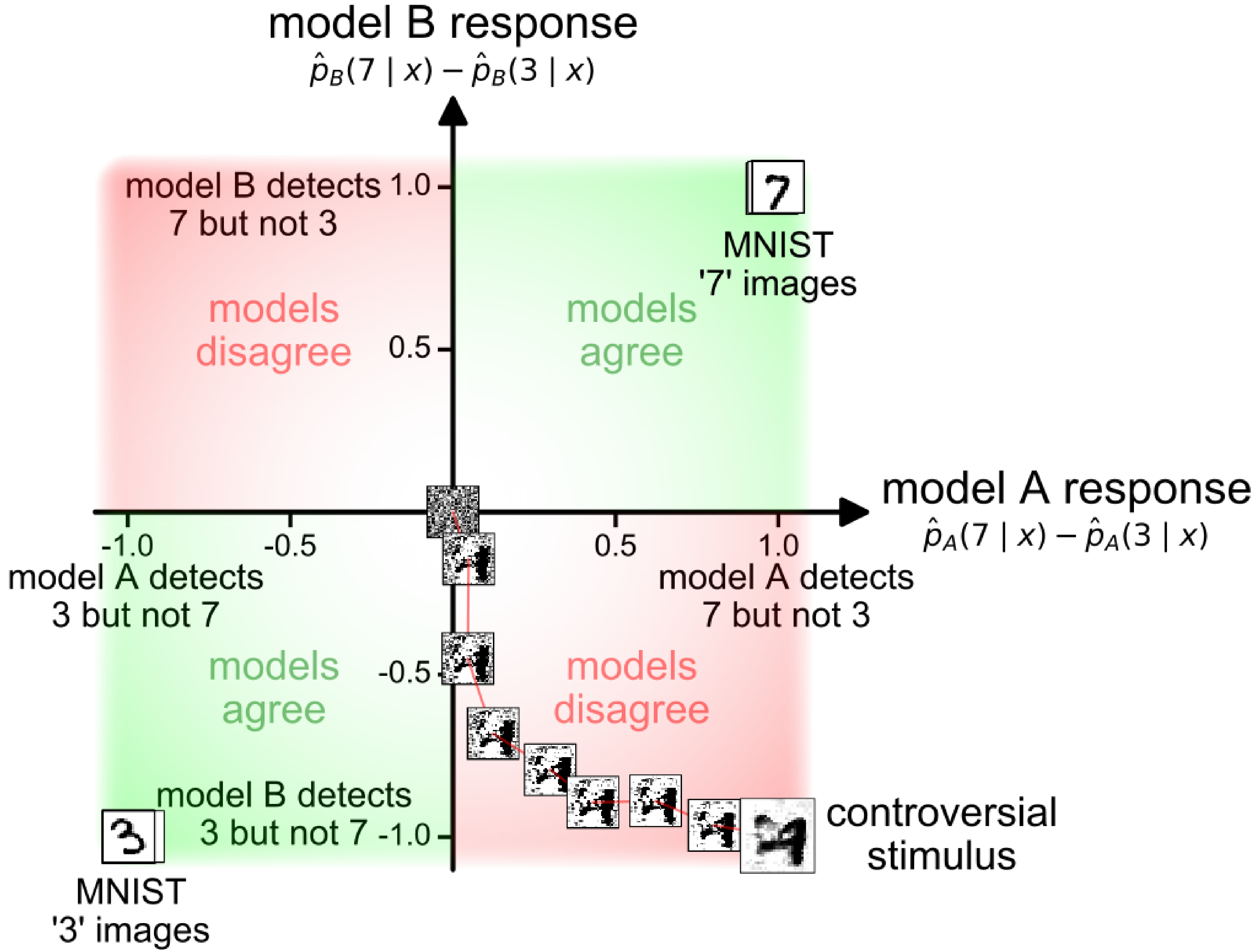
Golan et al. 2020

Controversiality index

$$c_{\underbrace{A,B}_{\text{model pair}}}^{\overbrace{d_a,d_b}^{\text{digit pair}}}(\underbrace{x}_{\text{image}}) = \min\Big\{\overbrace{\hat{p}_A(d_a \mid x)}^{\text{model }\mathbf{A}\text{ detects digit }\mathbf{a}}, \overbrace{1 - \hat{p}_A(d_b \mid x)}^{\mathbf{and}\text{ not digit }\mathbf{b}},$$

$$\underbrace{\hat{p}_B(d_b \mid x)}_{\mathbf{and}\text{ model }\mathbf{B}\text{ detects digit }\mathbf{b}}, \underbrace{1 - \hat{p}_B(d_a \mid x)}_{\mathbf{and}\text{ not digit }\mathbf{a}}\Big\},$$

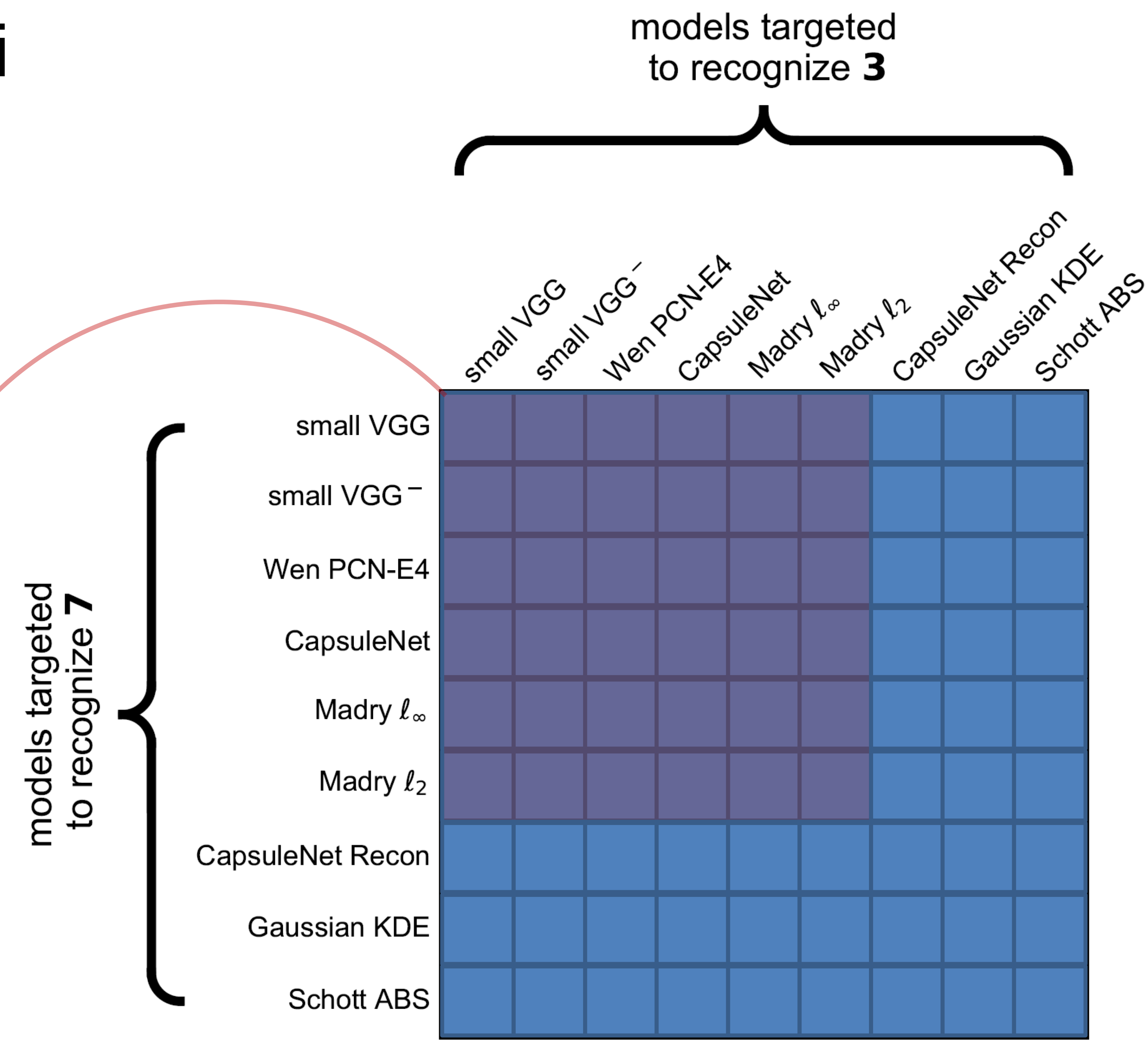
Golan et al. 2020


model B response
$\hat{p}_B(7 \mid x) - \hat{p}_B(3 \mid x)$
model B detects
7 but not 3
1.0
0.5
-0.5
-1.0
models
disagree
models
agree
MNIST
'7' images
model A response
$\hat{p}_A(7 \mid x) - \hat{p}_A(3 \mid x)$
-1.0
-0.5
0.5
1.0
model A detects
3 but not 7
model A detects
7 but not 3
models
agree
models
disagree
model B detects
3 but not 7
controversial
stimulus
MNIST
'3' images

Controversial stimuli

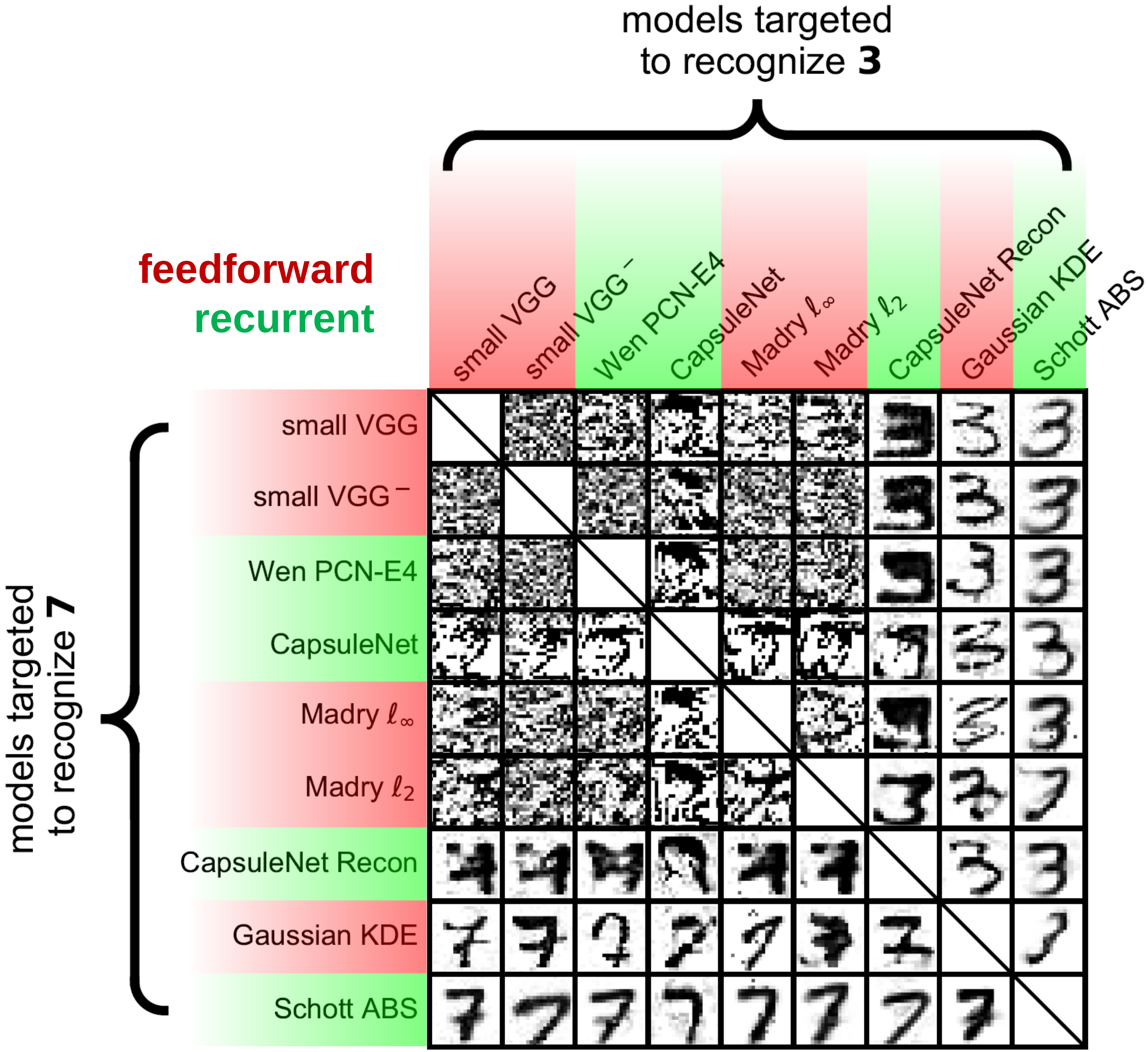
(optimized by gradient descent)

adversarial example
a stimulus that is controversial
between a model and ground truth



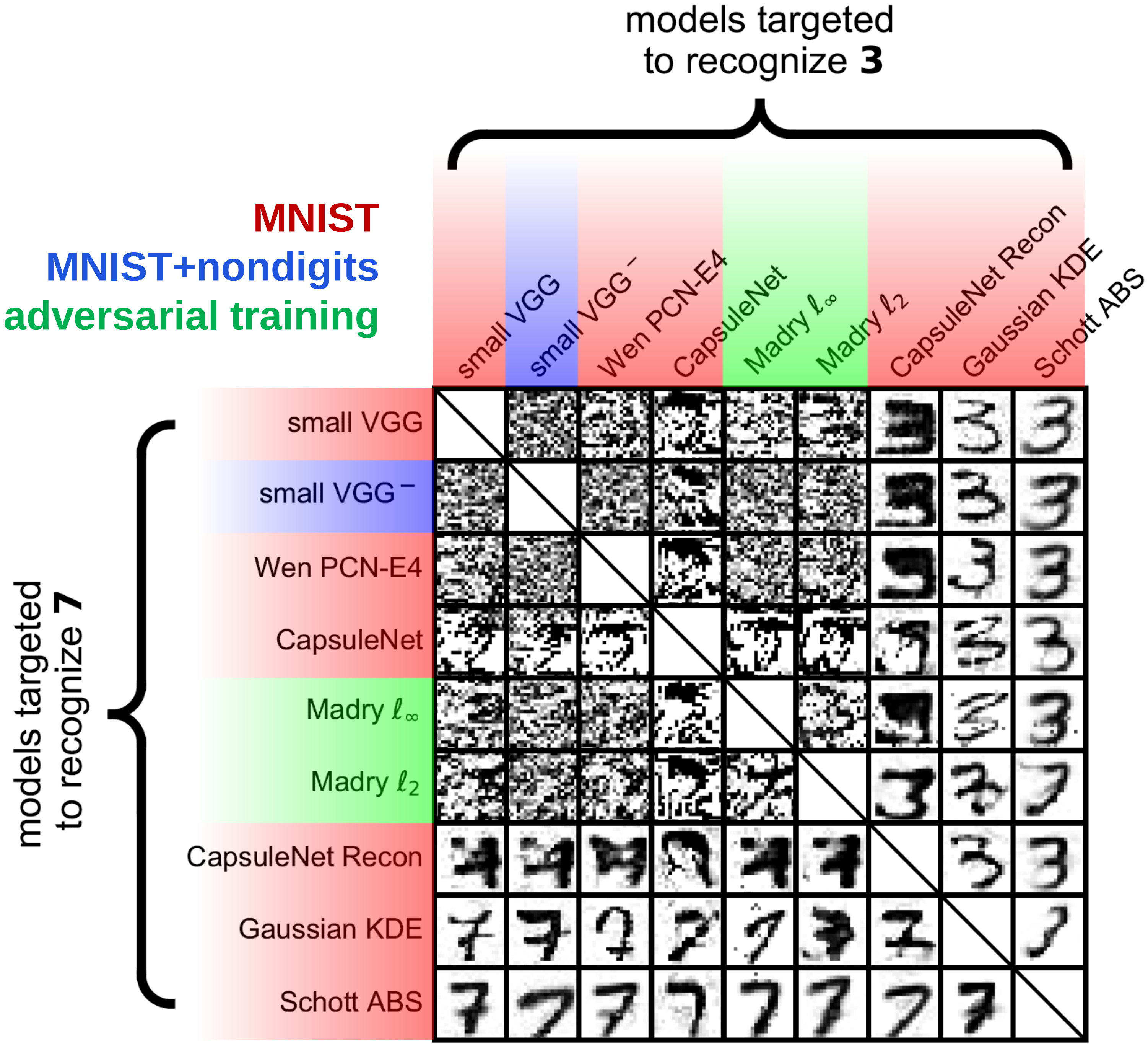
Golan et al. 2020

Controversial stimuli



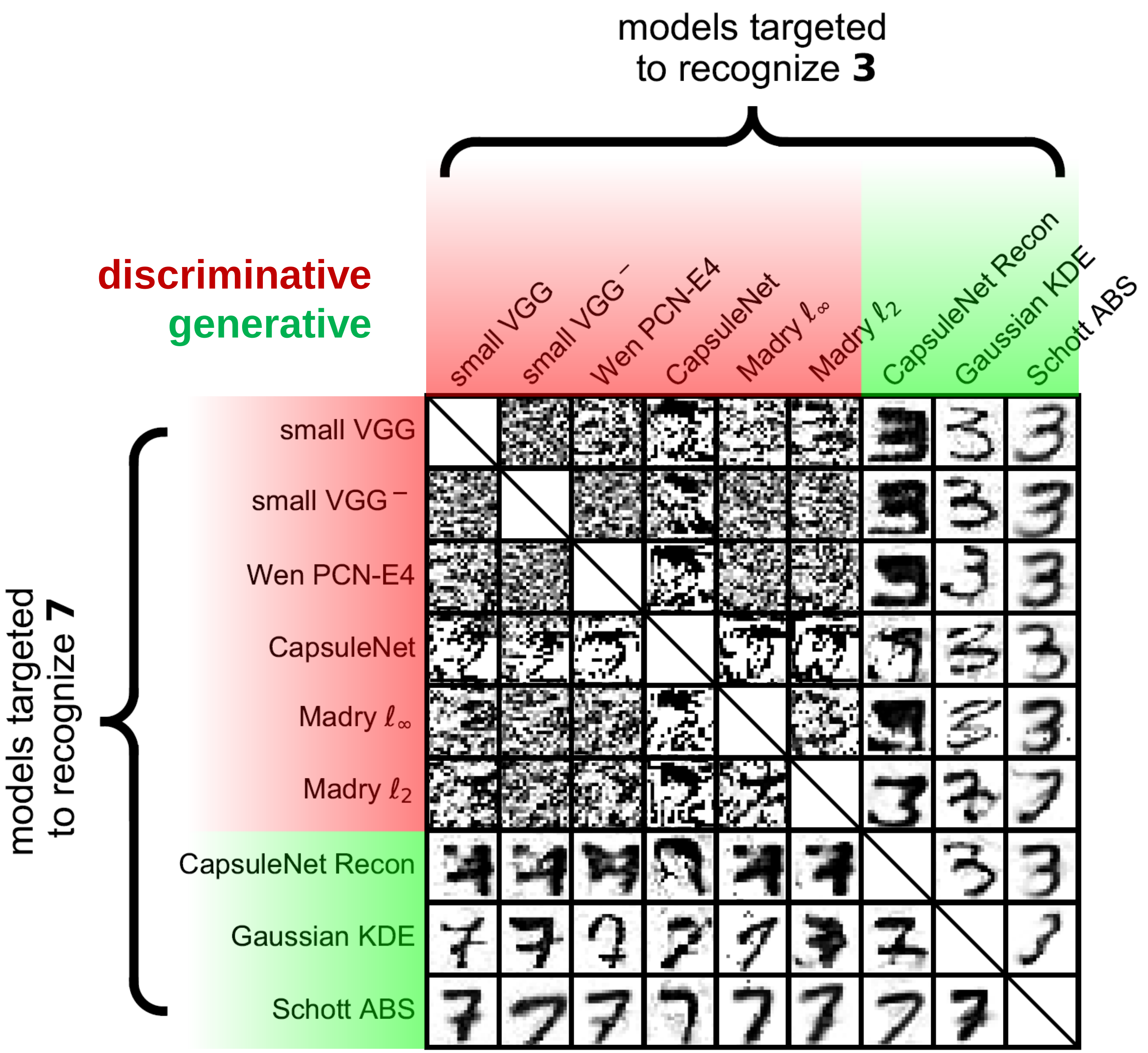
Golan et al. 2020

Controversial stimuli

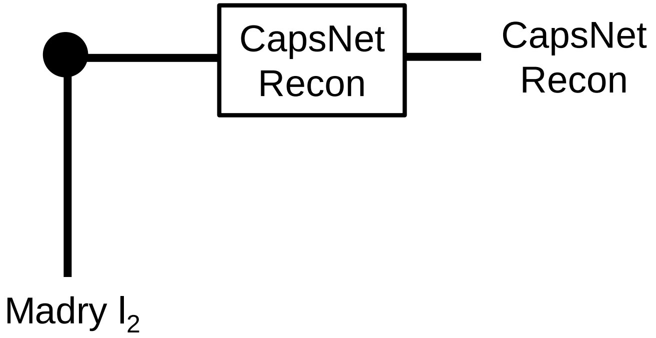


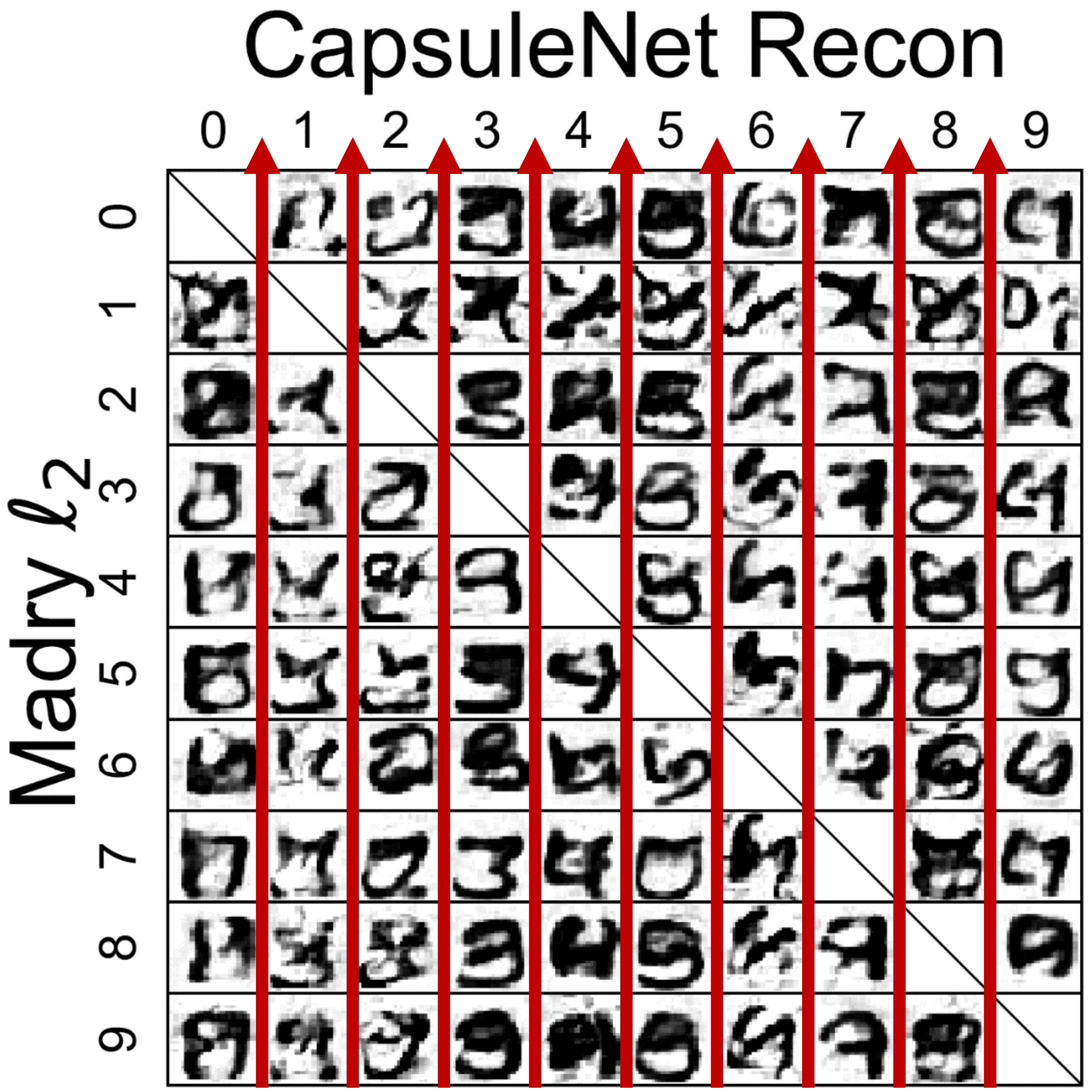
Golan et al. 2020

Controversial stimuli

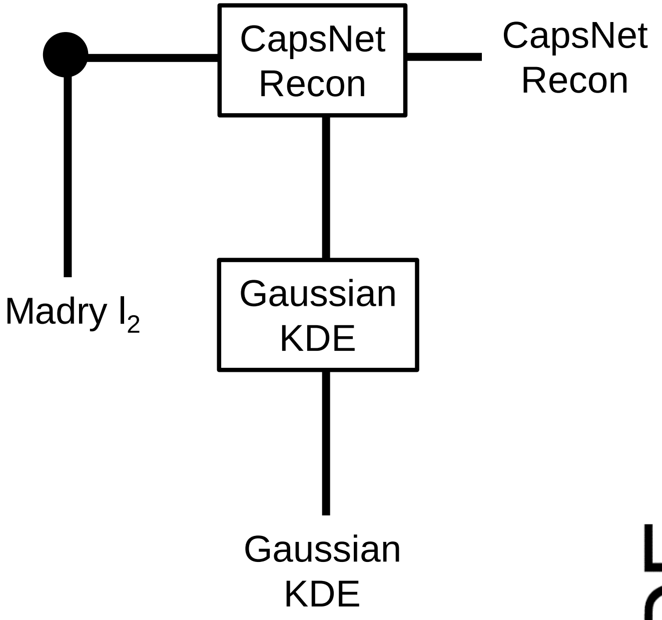


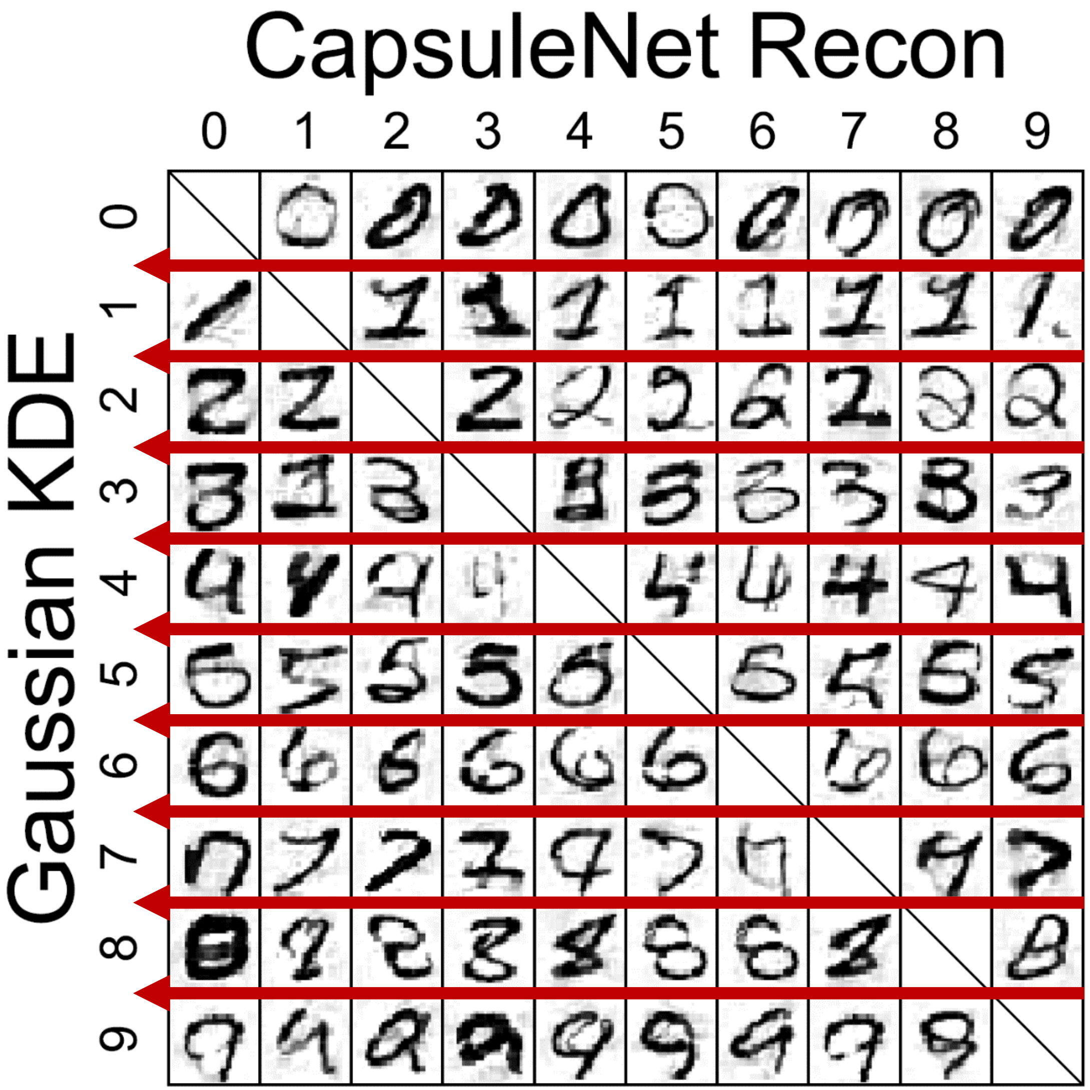
Golan et al. 2020



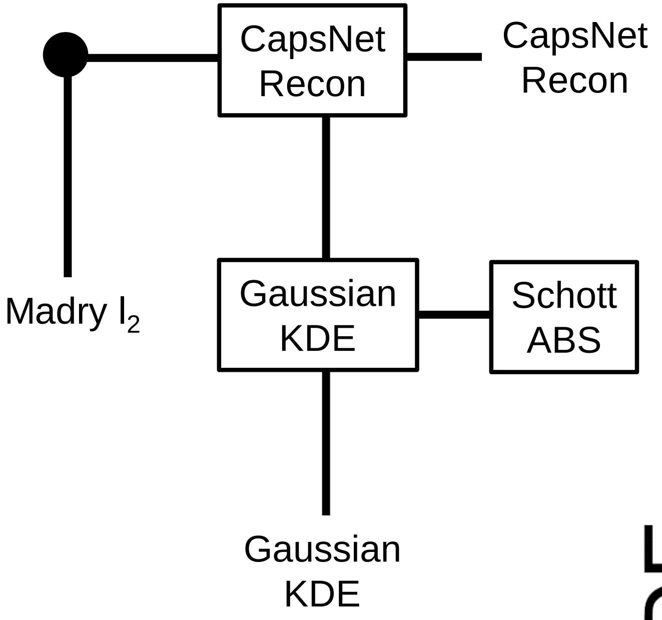


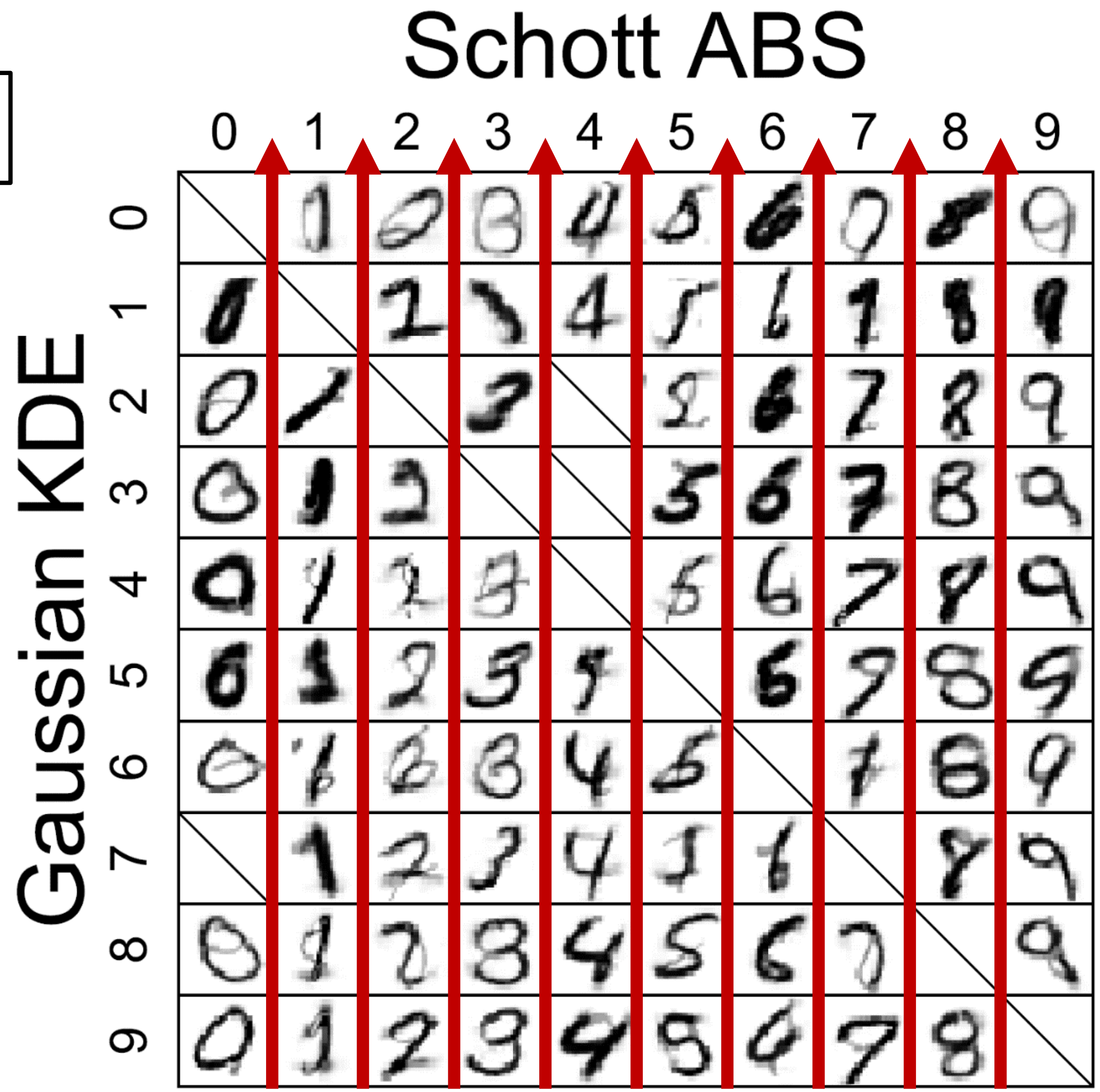
Golan et al. 2020



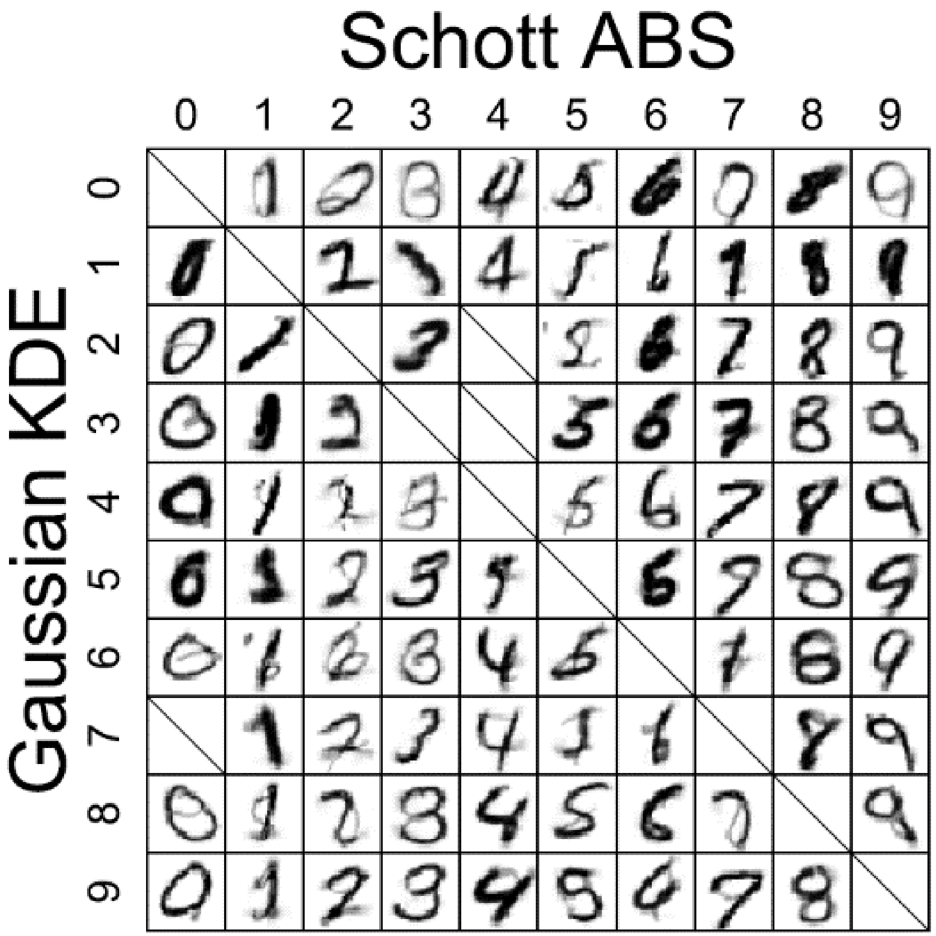


Golan et al. 2020



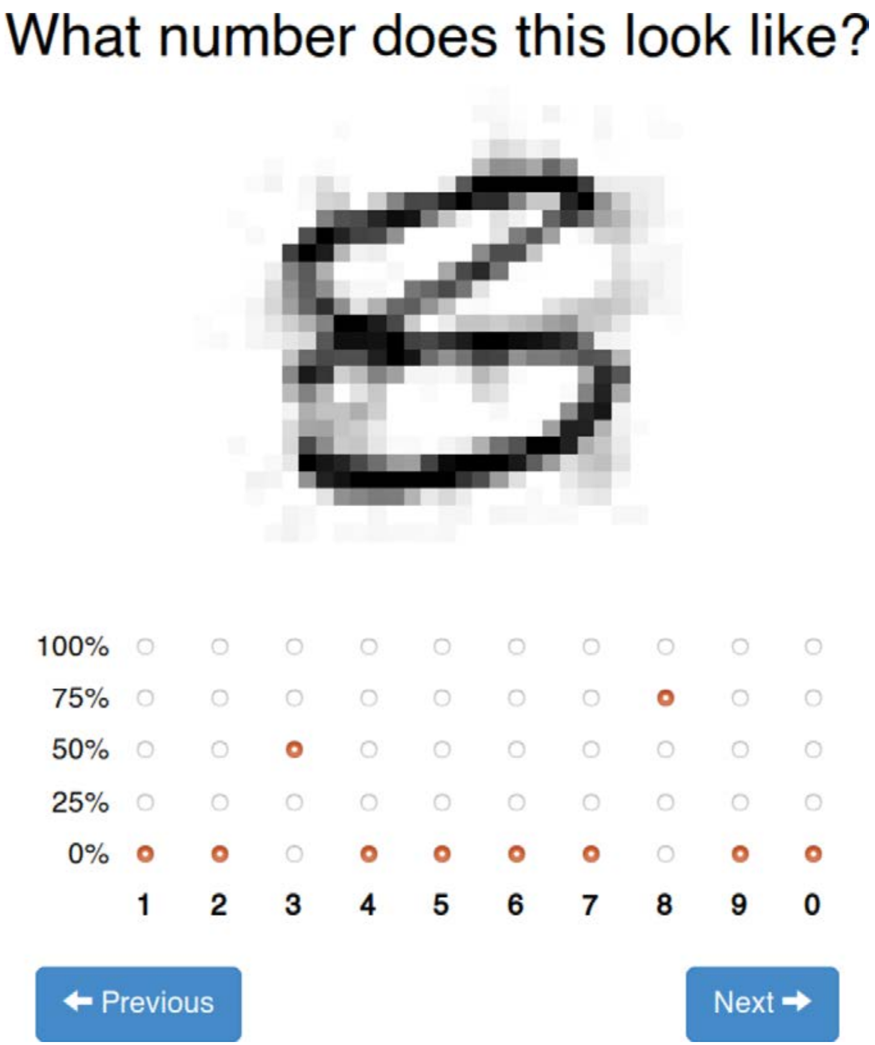


Golan et al. 2020



Golan et al. 2020

Behavioral experiment

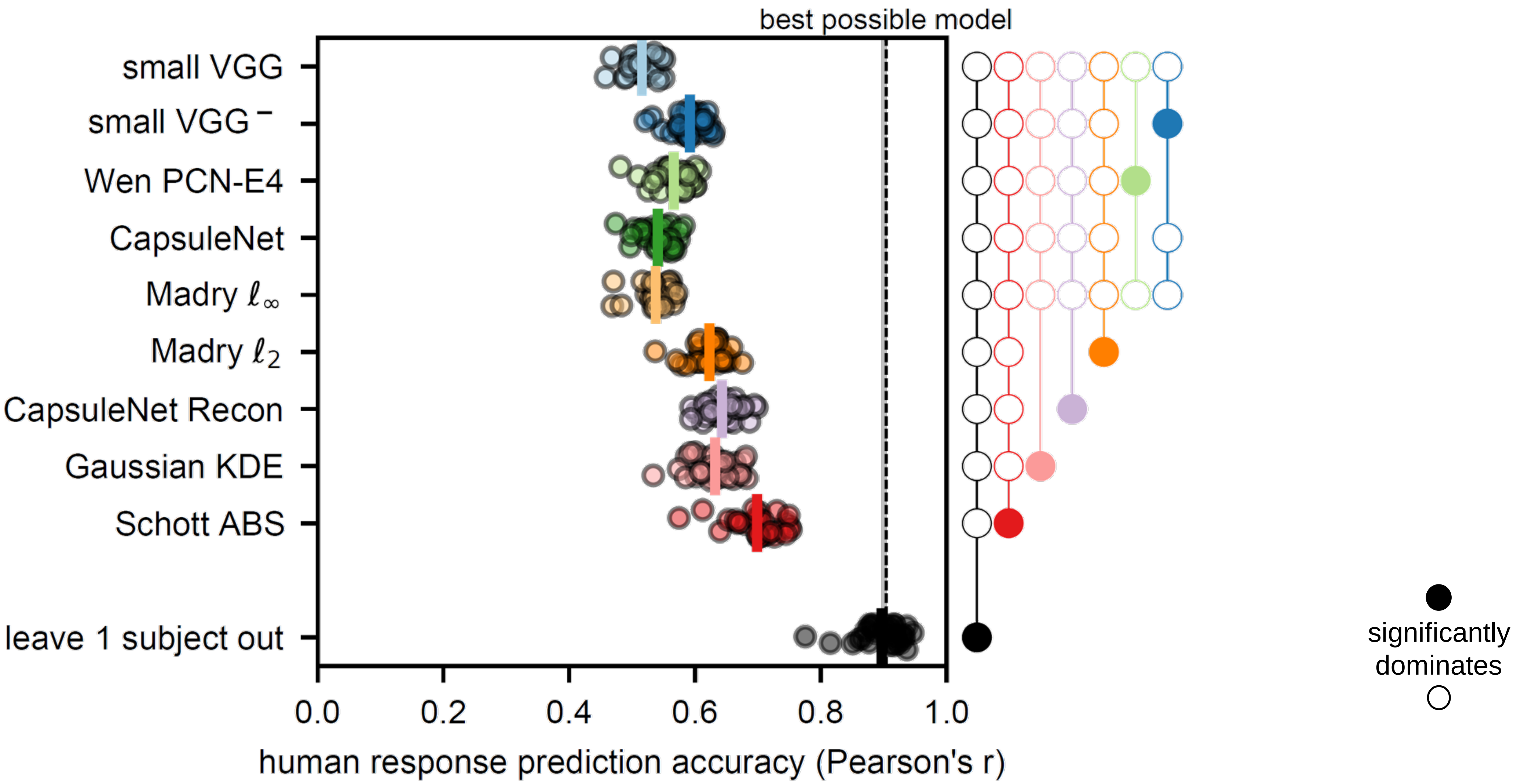


Golan et al. 2020

Behavioral experiment

- 30 subjects (tested via *Prolific*)
- stimuli included 20 controversial stimuli per model pair
  (36*20) + 100 MNIST images = 820 images per subject
- stimuli presented in a randomized order
- 820 stimuli x 10 scales x 30 subjects
  (246,000 data points)

Controversial stimuli



Golan et al. 2020

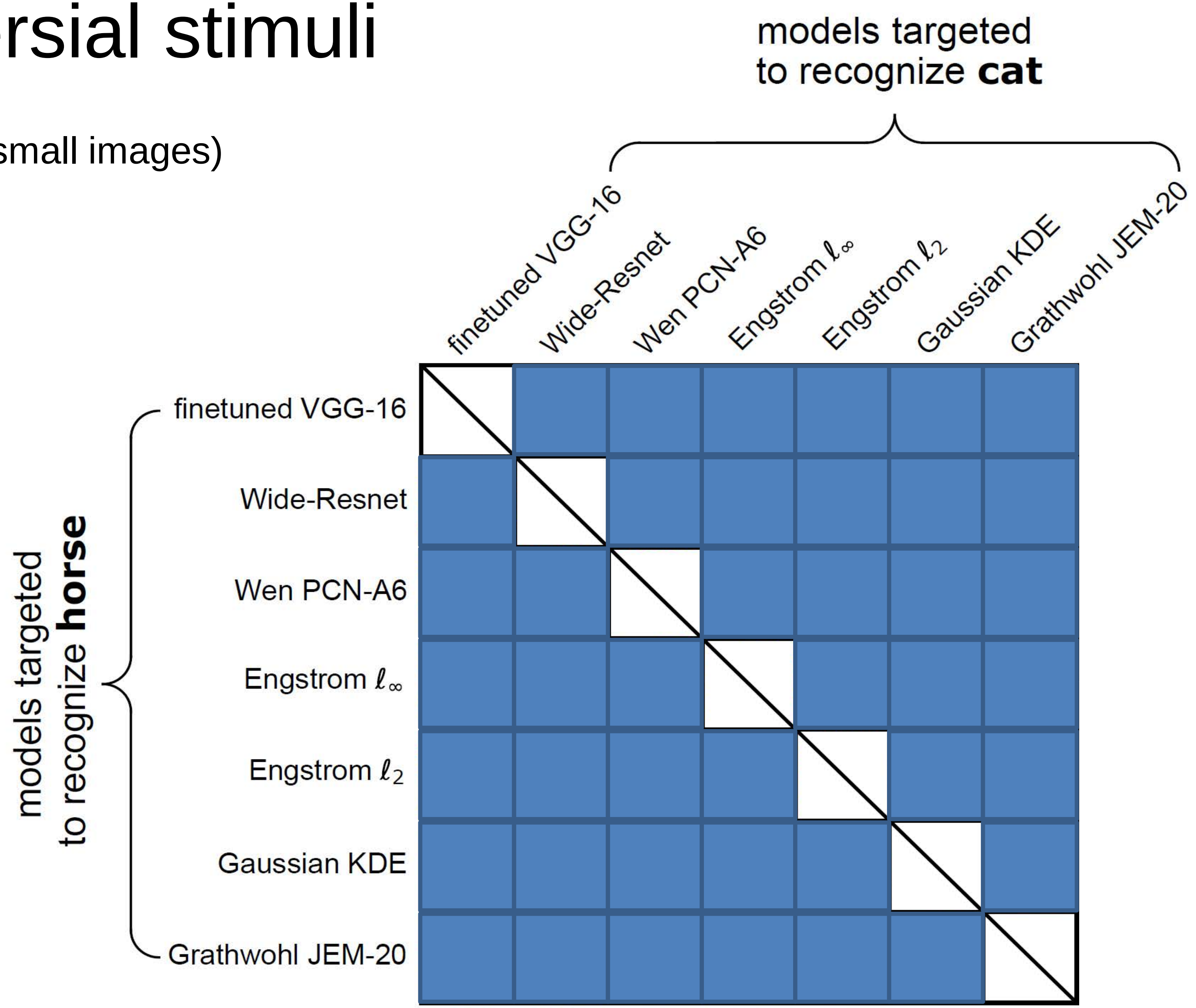
Controversial stimuli

natural images
(CIFAR-10 set of small images)

Golan et al. 2020

Controversial stimuli

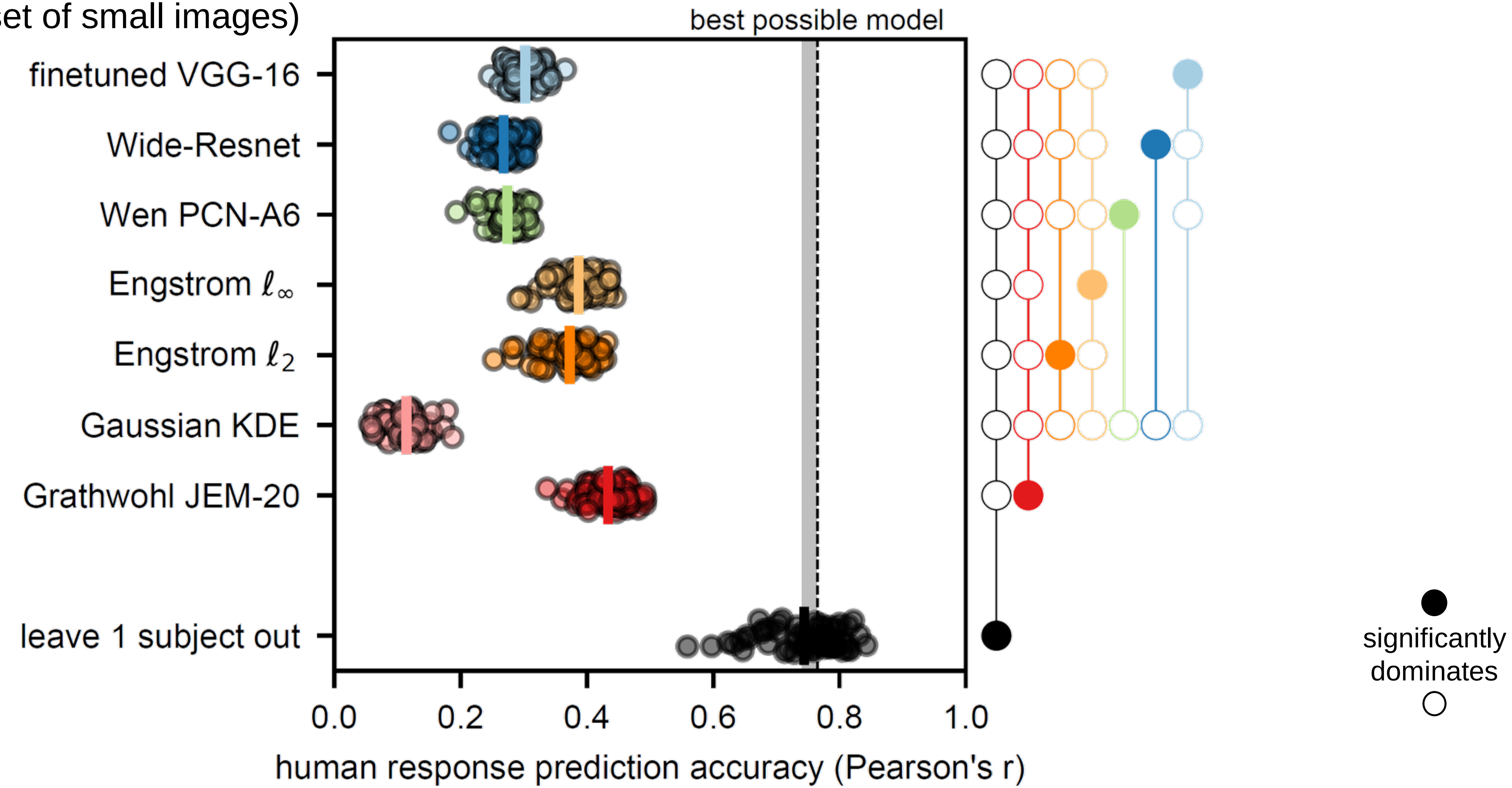
natural images
(CIFAR-10 set of small images)



Golan et al. 2020

Controversial stimuli

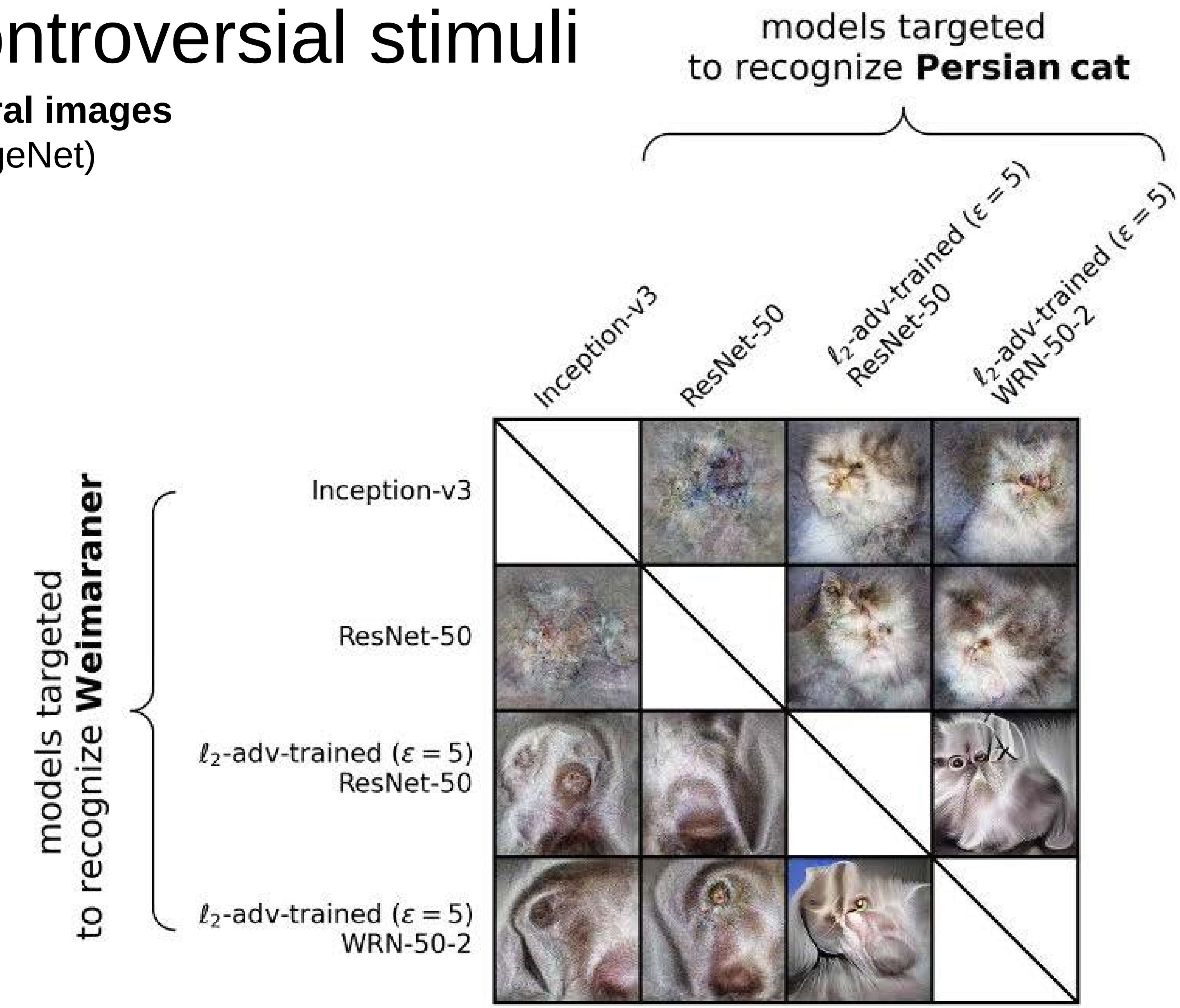
natural images
(CIFAR-10 set of small images)



Golan et al. 2020

Controversial stimuli

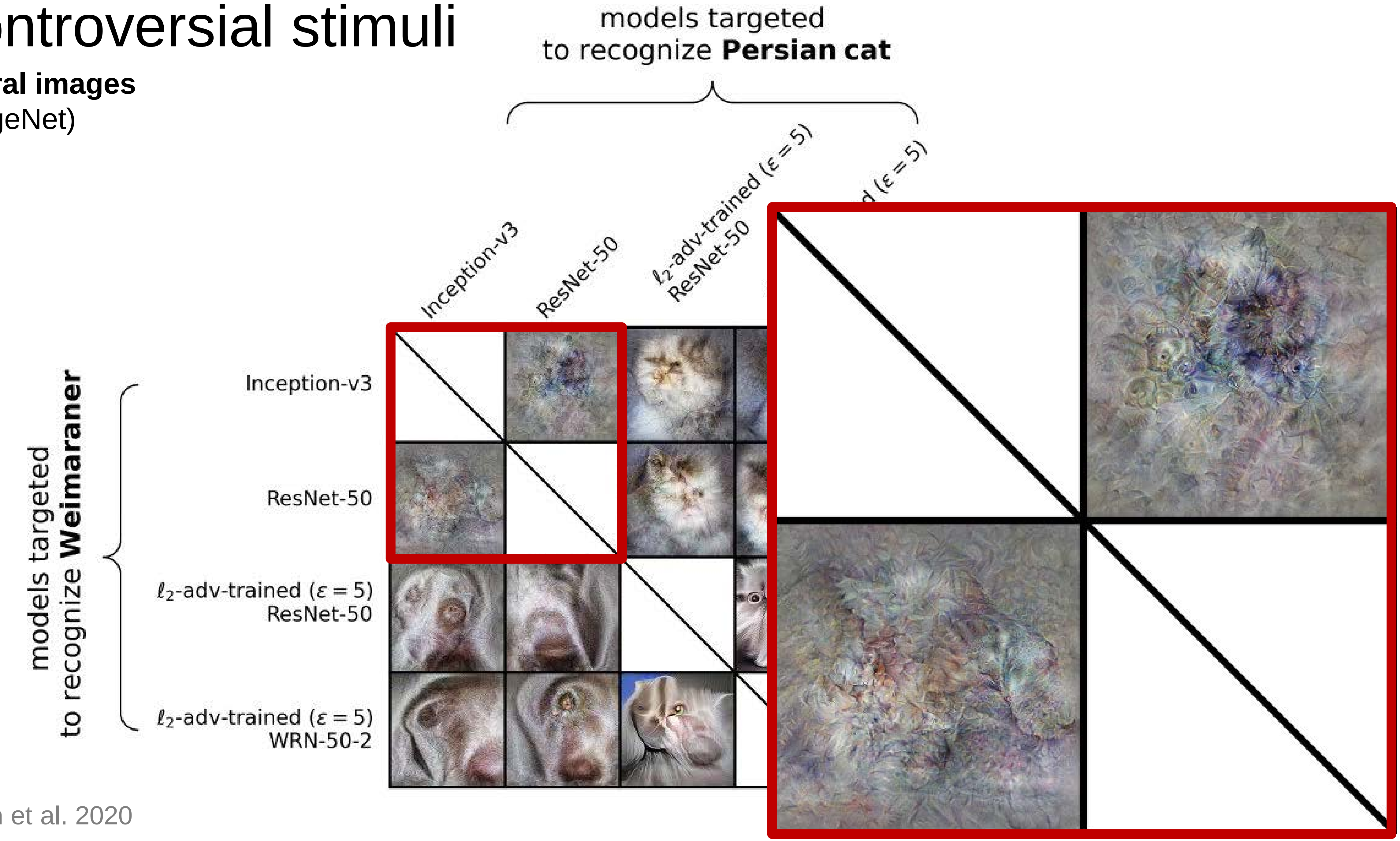
natural images
(ImageNet)



Golan et al. 2020

Controversial stimuli

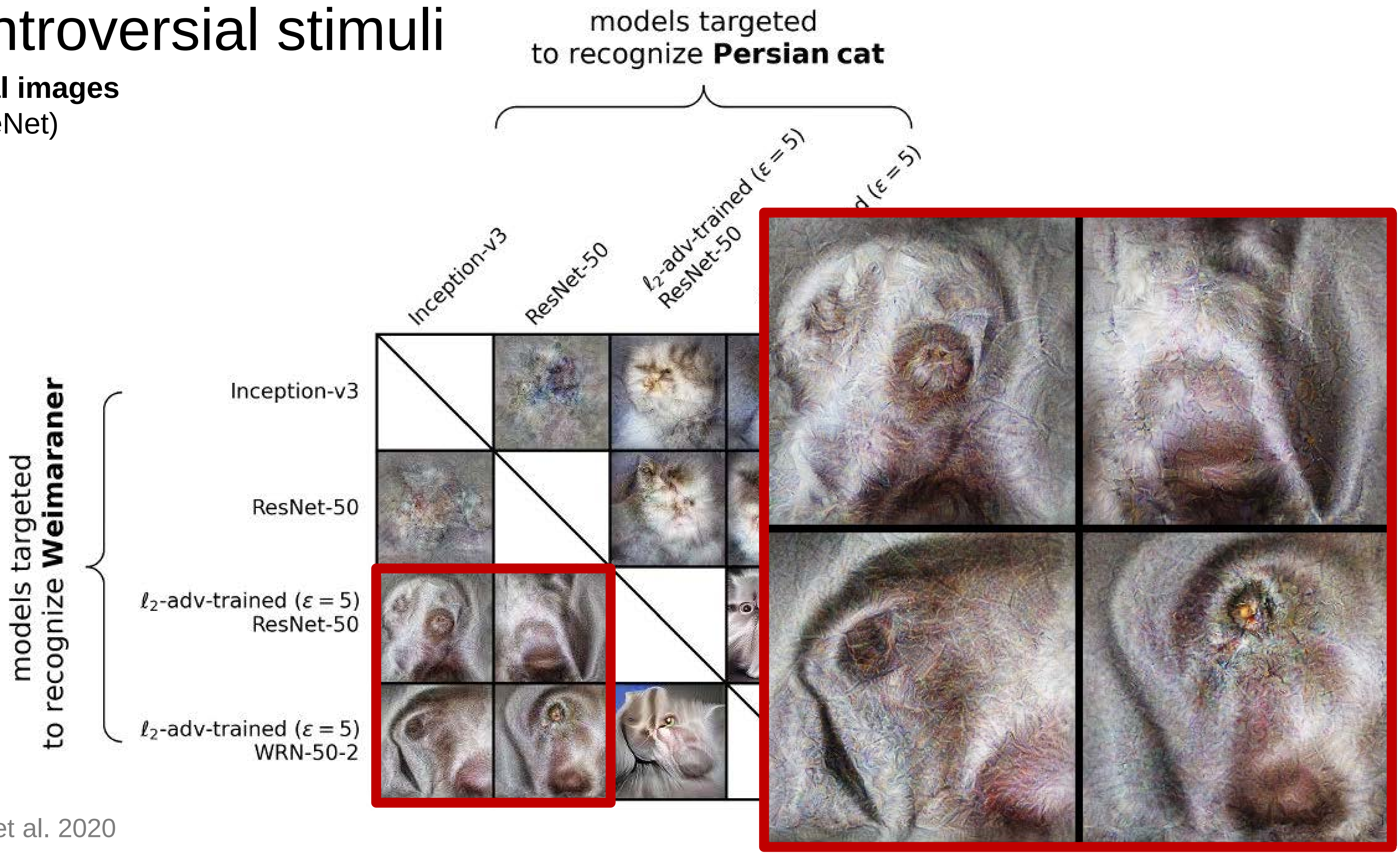
natural images
(ImageNet)



Golan et al. 2020

Controversial stimuli

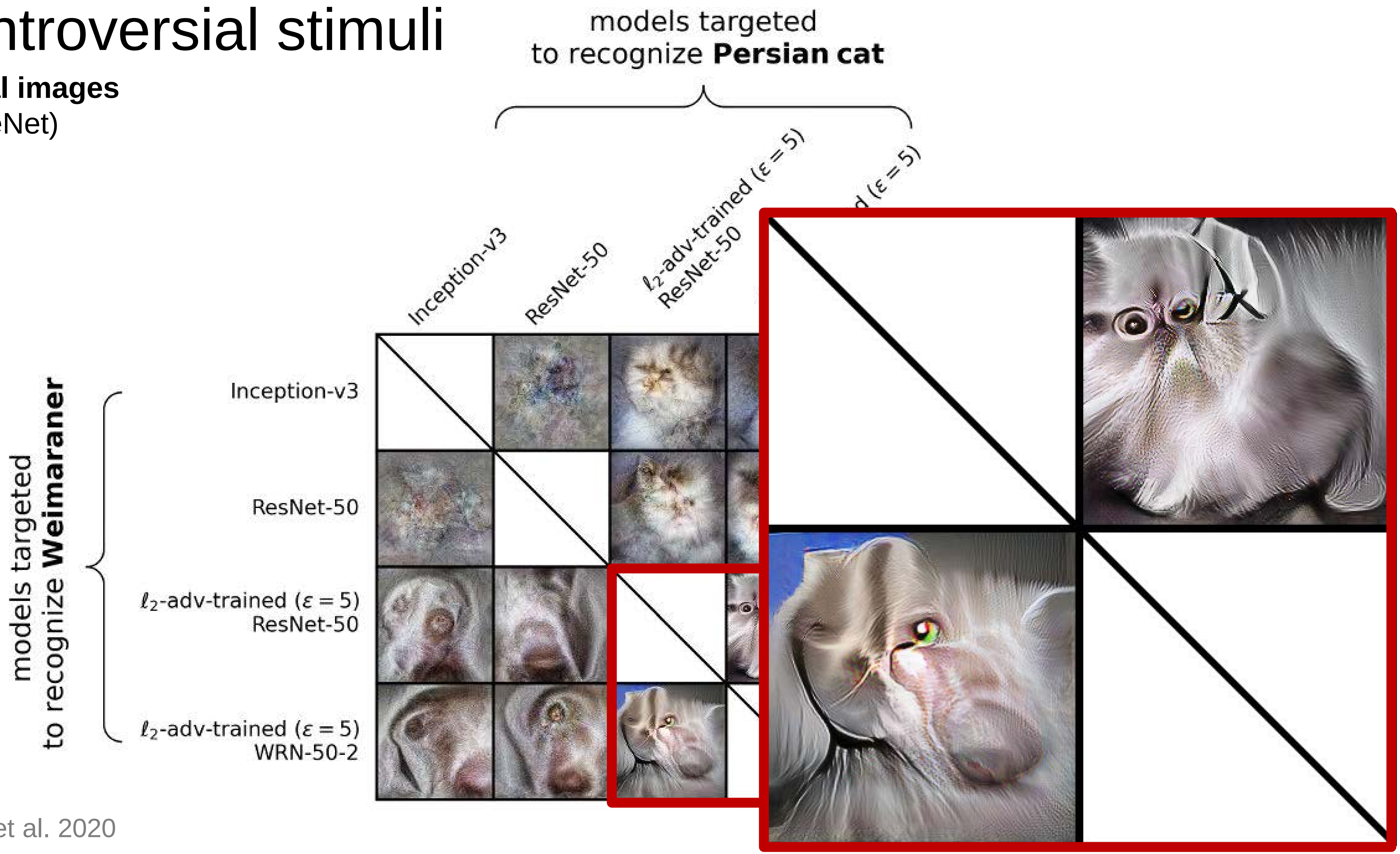
natural images
(ImageNet)



Golan et al. 2020

Controversial stimuli

natural images
(ImageNet)



Golan et al. 2020

Overall conclusions

1. We can adjudicate among task-performing deep net models by inferentially comparing their representations to brain representations.
   Nili et al. 2014, **Kriegeskorte** & **Diedrichsen** 2019

2. Recurrent convolutional vision models better predict human ventral stream representational dynamics and reaction times
   Kietzmann et al. 2019, **Spoerer** et al. 2020

3. Controversial stimuli enable us to elicit differences in the inductive biases of deep net model.
   Golan et al. 2020

4. Human vision may rely on a computational mechanism that combines elements of discriminative and generative inference.
   Golan et al. 2020