Targeted perturbations reveal brain-like local coding axes in robustified, but not standard, ANN-based brain models

Nikolas McNeal

Vision
Computation
Cognition
murtylab.com

Artificial neural networks as models of the brain

What do we want from a model of the brain?

- ***Precise, quantitative predictions*** *of neural data under novel conditions*

- *Implementation of* ***similar information-processing*** *strategies to the brain*

Artificial neural networks as models of the brain

What do we want from a model of the brain?

- ***Precise, quantitative predictions*** *of neural data under novel conditions*

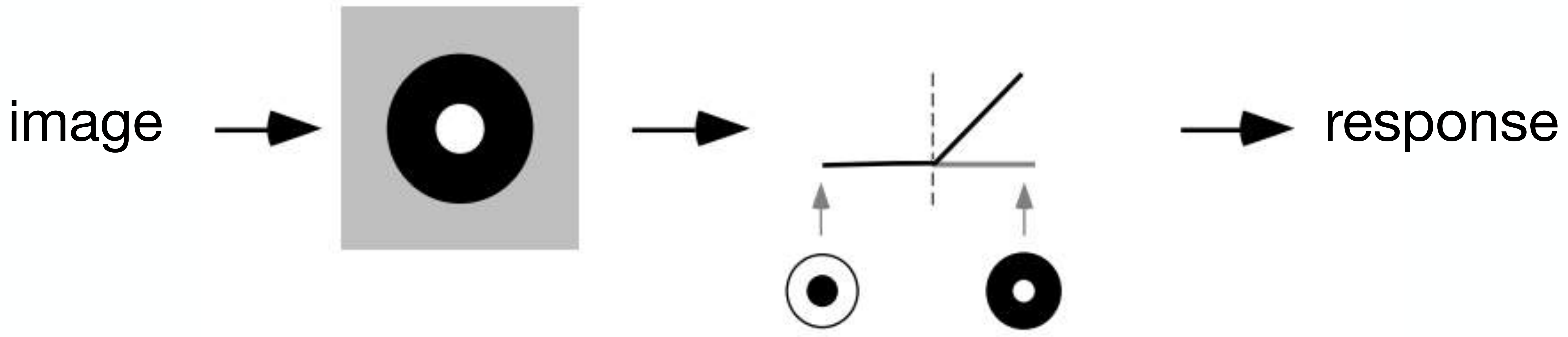
    "Encoding model" — input arbitrary sensory stimuli and predict exact brain responses

- *Implementation of* ***similar information-processing*** *strategies to the brain*

    Models with stable **brain-like** internal representations

How do we build predictive models?

- Early work used handcrafted features to predict brain responses
    - Image is convolved with feature and passed through a nonlinearity



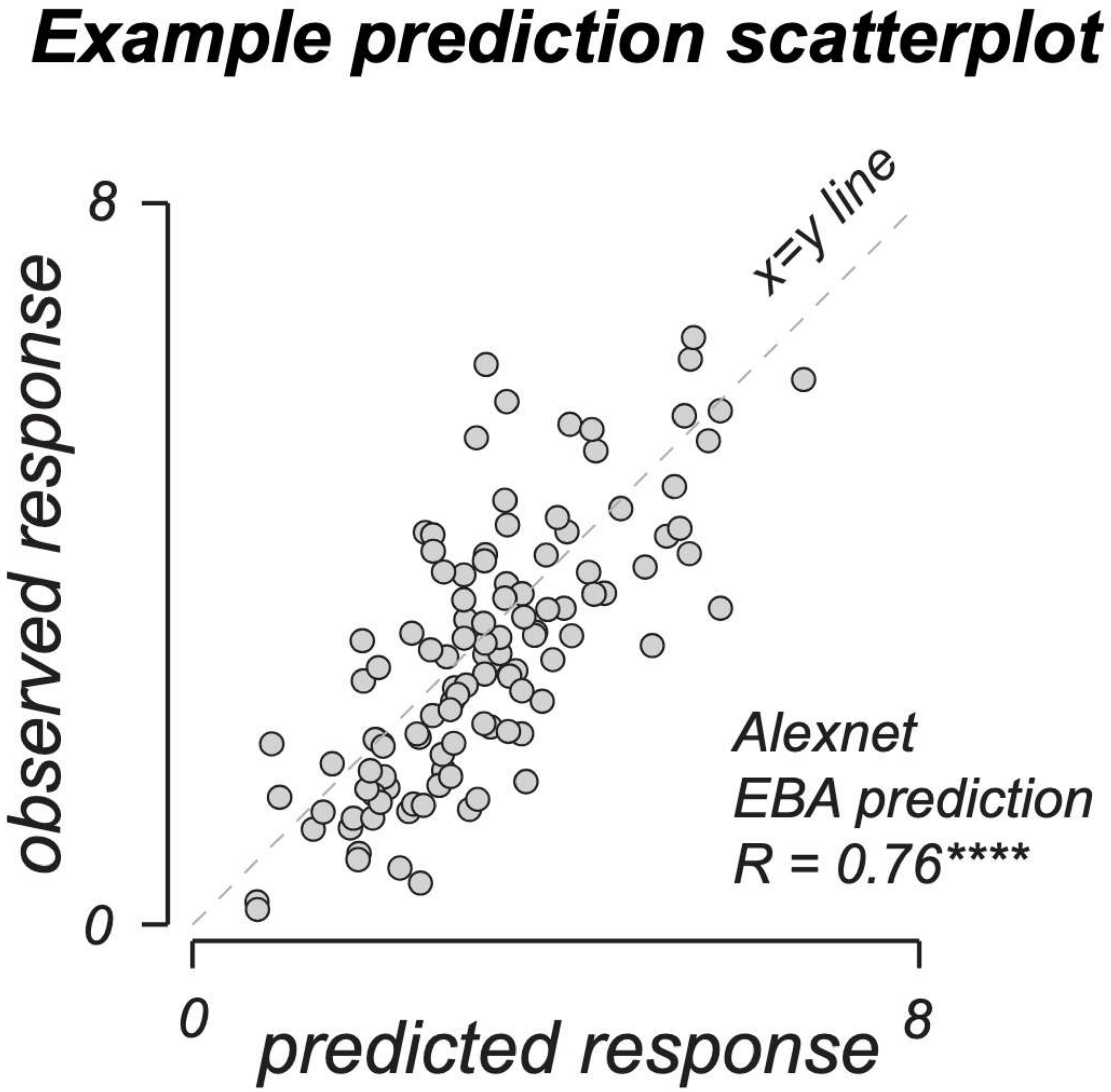
- Outputs are directly compared with neural firing rates (high correlation implies neuron is selective to this feature)

Artificial Neural Networks (ANNs) as models of the brain

- In more recent years, ANNs have emerged as leading models of the brain

- Features *learned by* ANNs are the most predictive

- We now have unprecedented predictive precision!

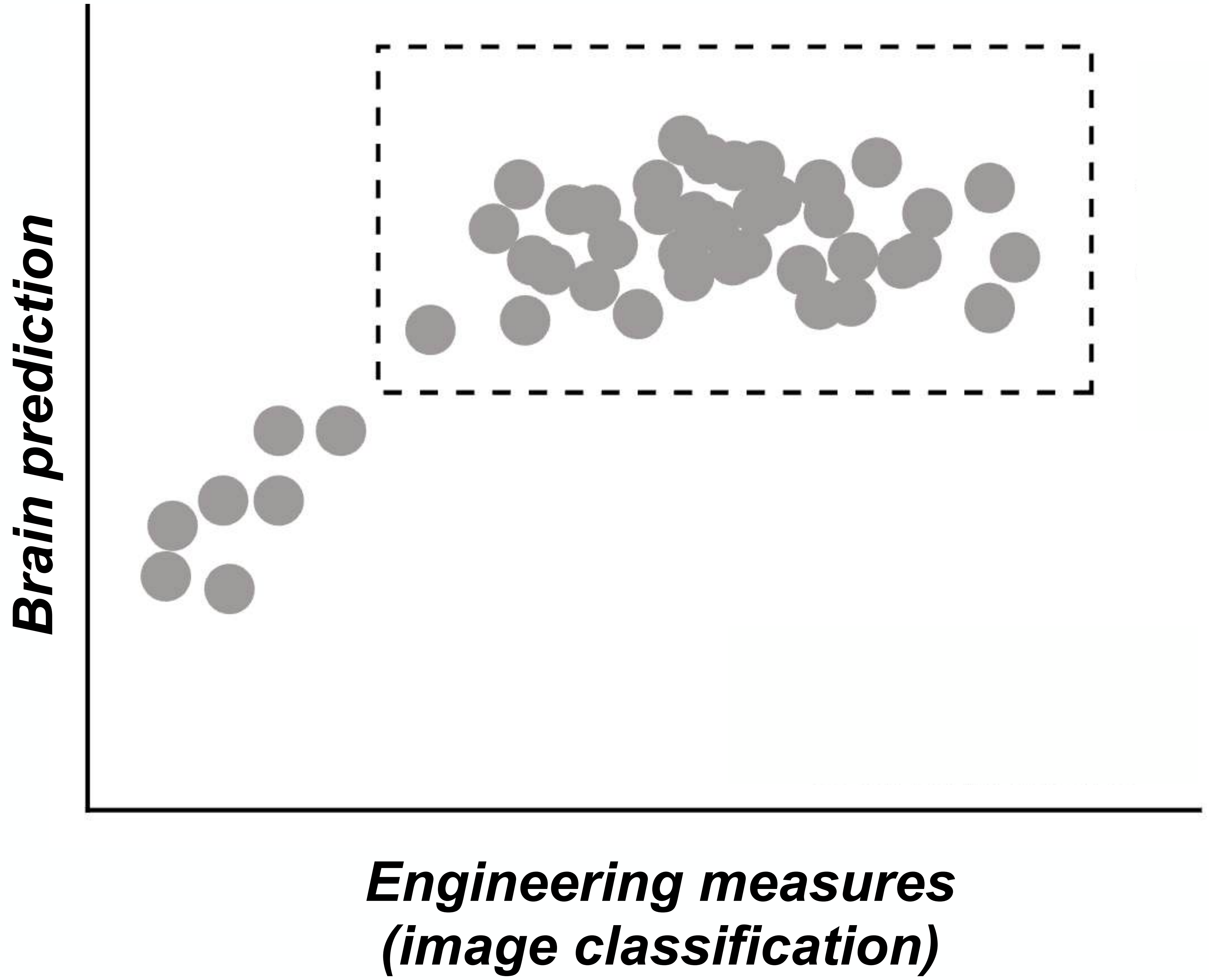


Artificial Neural Networks (ANNs) as models of the brain

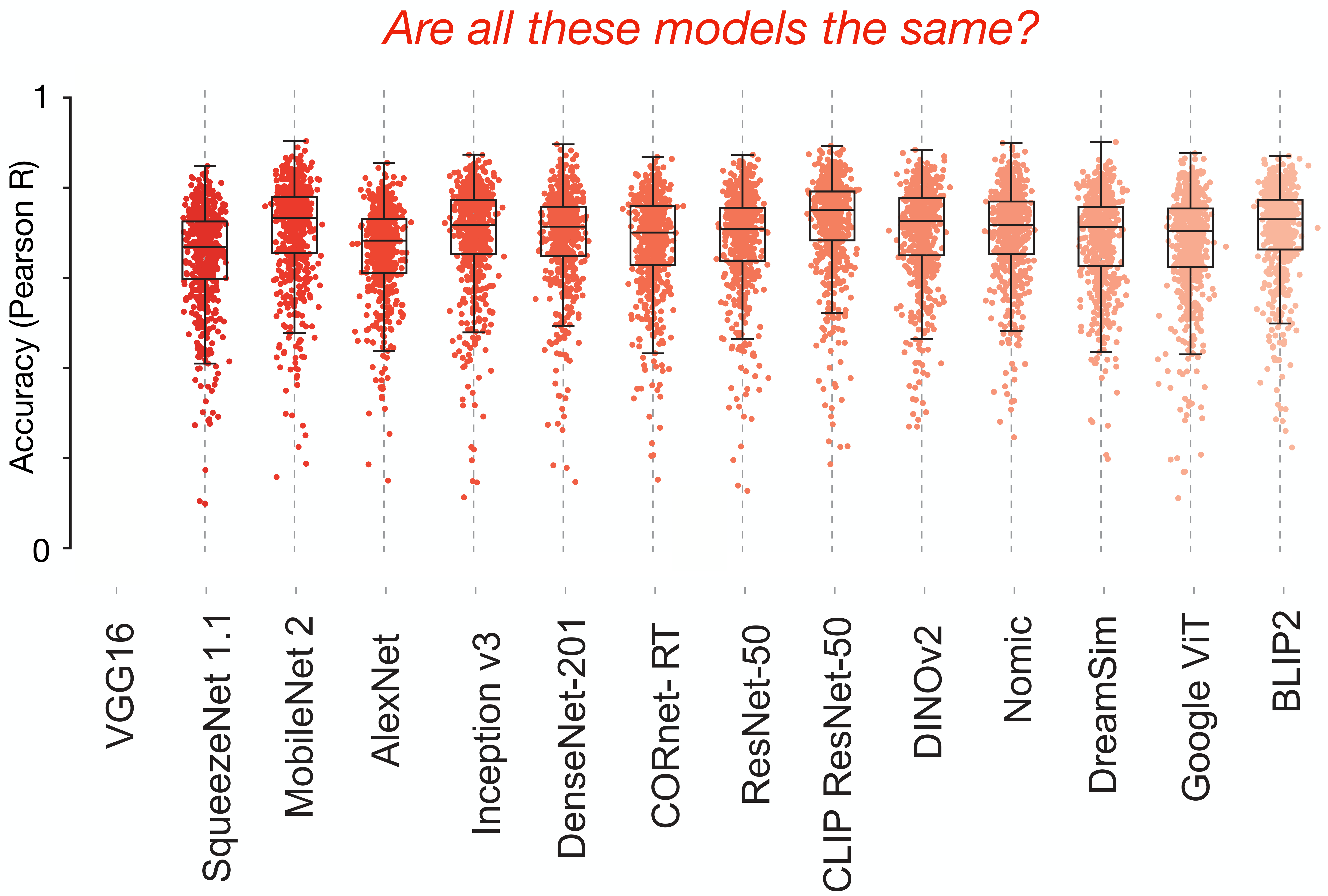
- As ANNs have become better on **engineering measures,** they typically have improved on **brain prediction**

This relationship has *plateaued!*

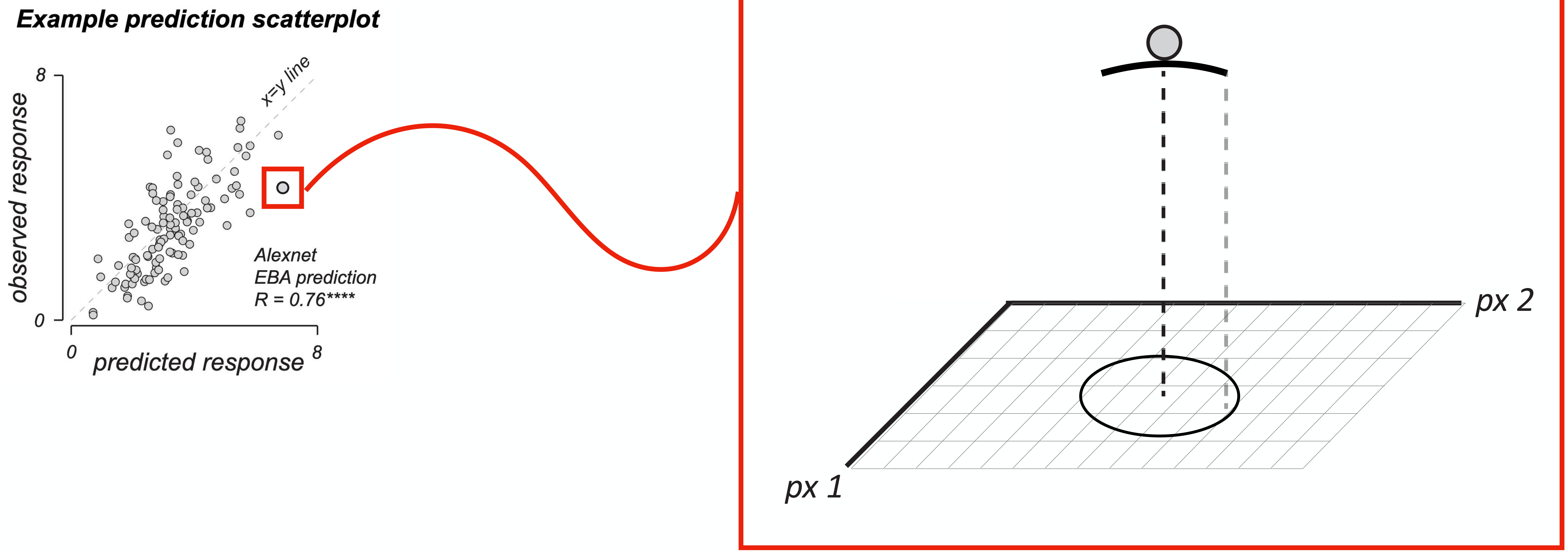
Do we need to move beyond prediction scores?



Artificial neural networks as models of the brain

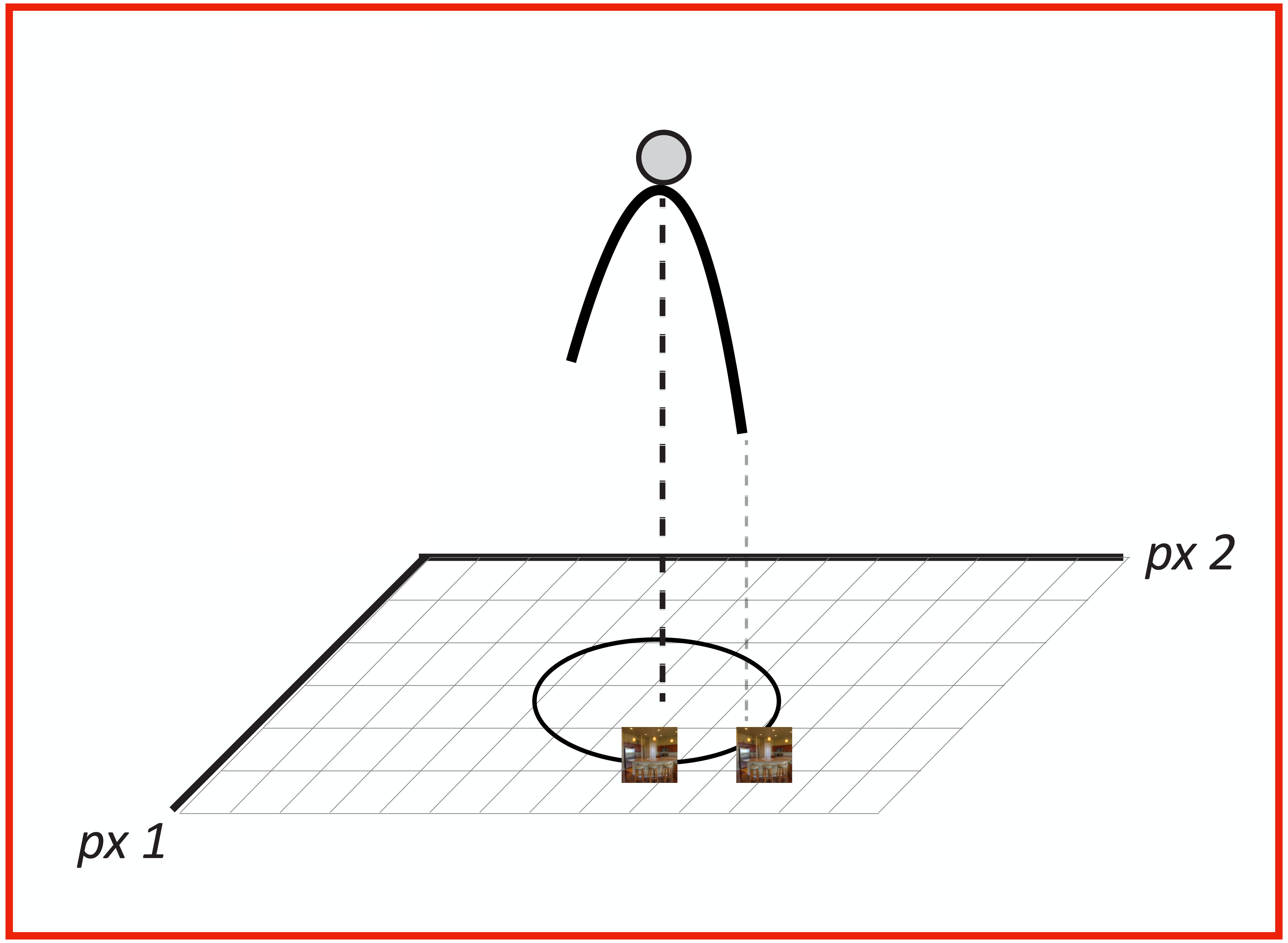


We should expect responses to be **stable** to small changes in the image



Can you tell the difference
between these two images?

If models have **failure modes,** this makes it a bad model!



Predictivity is an important metric, but a **predictive+stable** model is better than an **unstable+predictive** one

- How stable are model predictions?

- Do models share the same failure modes?

- Can we use stability to find better models of the brain?

- Can we use stable+predictive models to generate hypotheses about the brain?

- How stable are model predictions?
- Do models share the same failure modes?

Small perturbations

- Can we use stability to find better models of the brain?
- Can we use stable+predictive models to generate hypotheses about the brain?

Bigger perturbations

- How stable are model predictions?

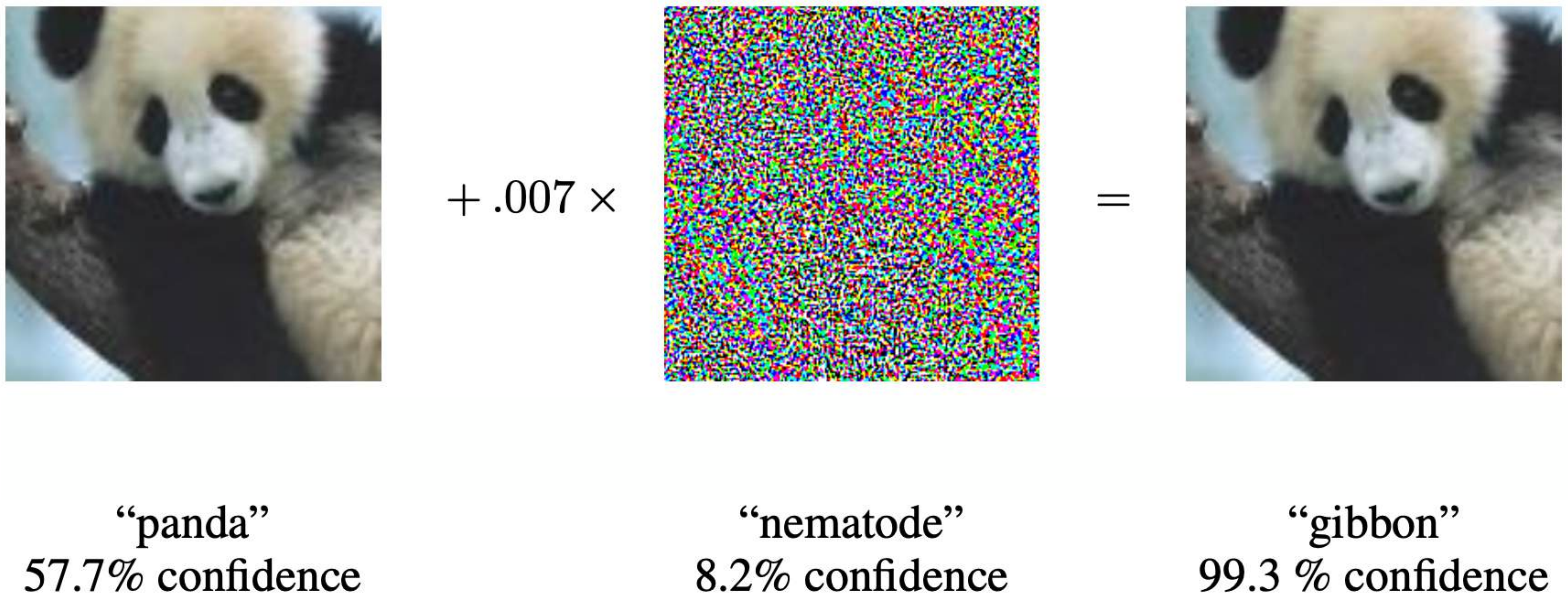
- Do models share the same failure modes?

- Can we use stability to find better models of the brain?

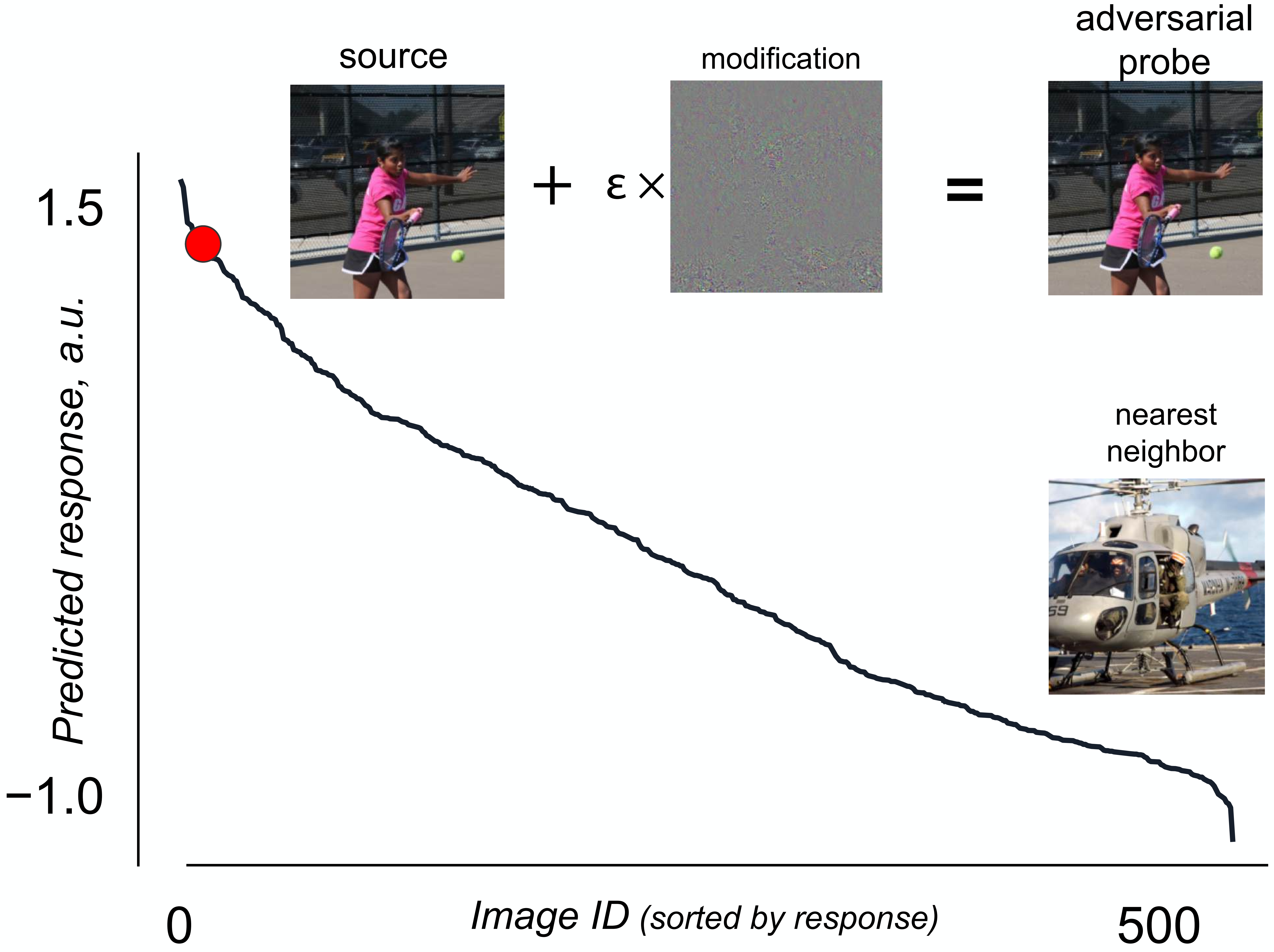
- Can we use stable+predictive models to generate hypotheses about the brain?

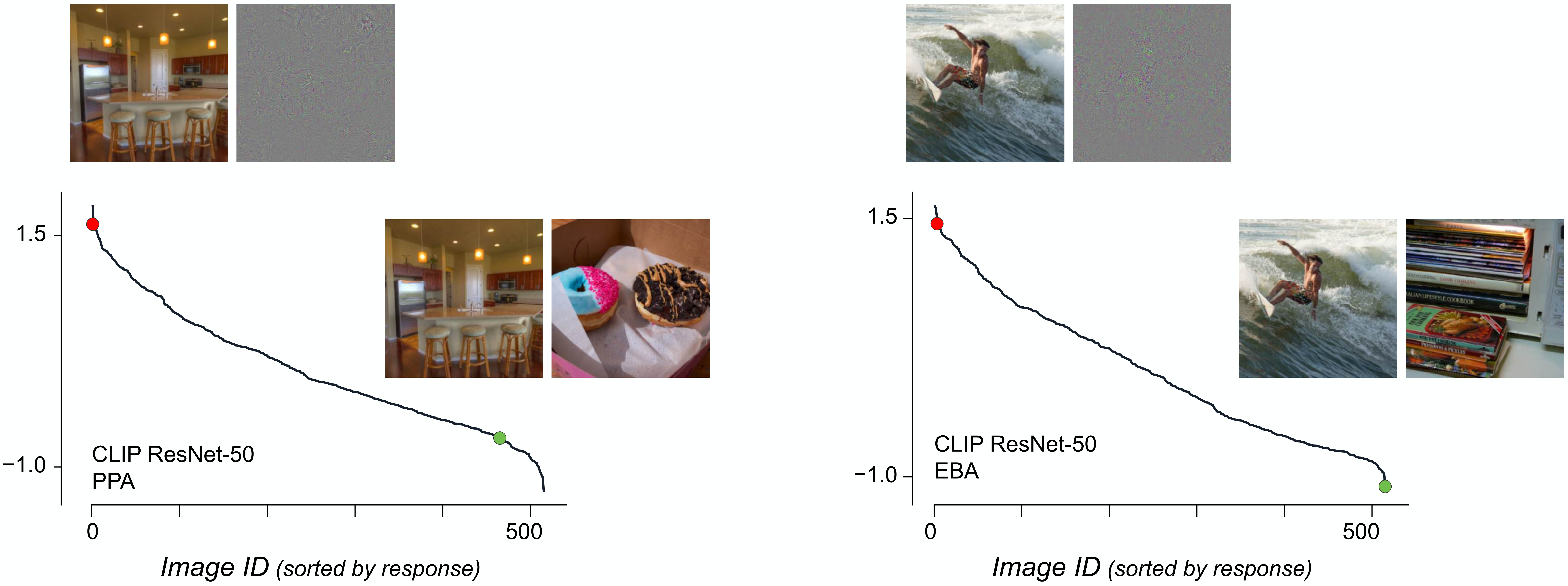
How do we find the worst-case change to an image?

- Machine learning work: ***adversarial attacks***
    - Formalizes “worst-case” perturbation under a certain pixel budget



Are brain models sensitive to small-scale noise?



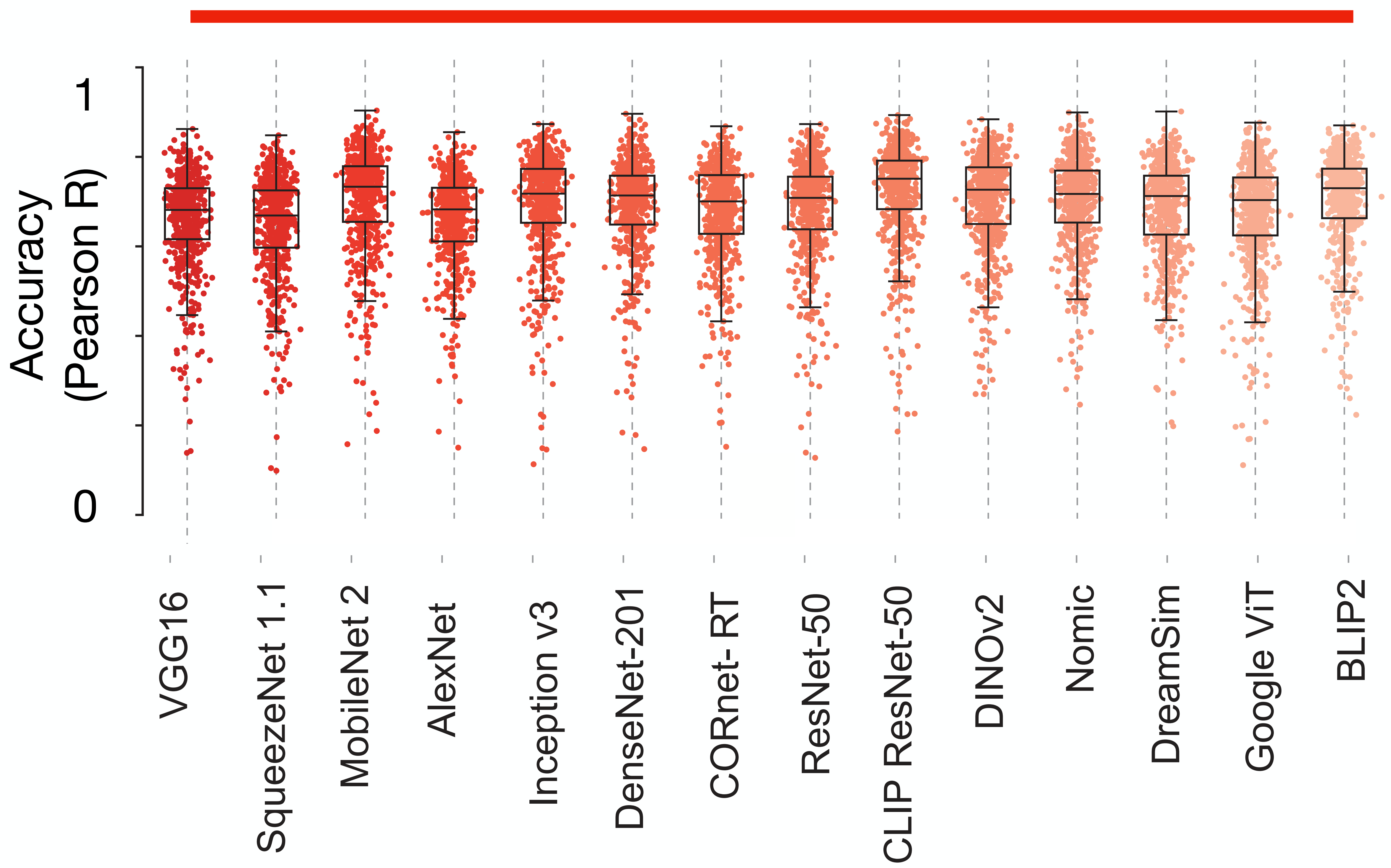

1.5
−1.0
CLIP ResNet-50
PPA
0
500
Image ID (sorted by response)
1.5
−1.0
CLIP ResNet-50
EBA
0
500
Image ID (sorted by response)

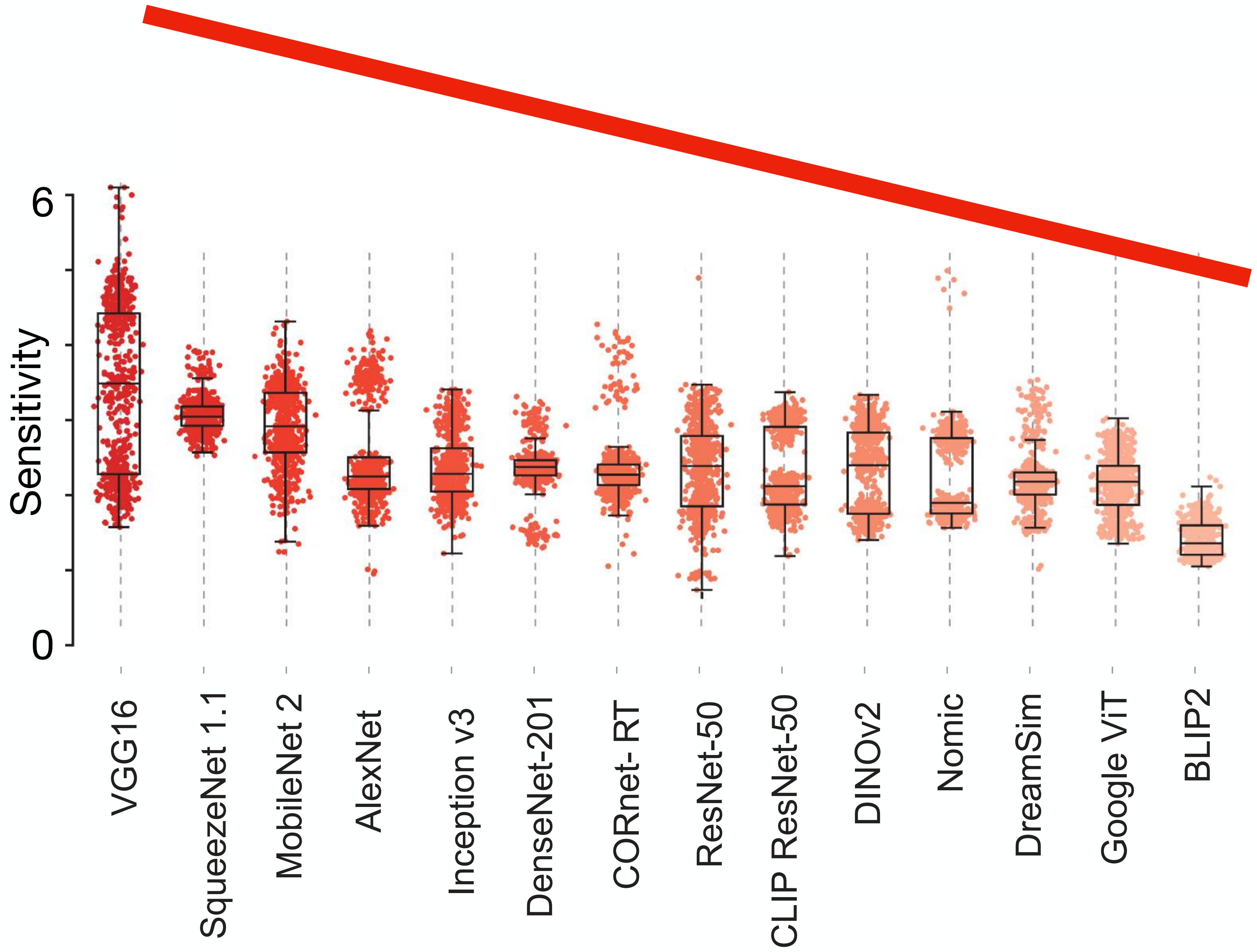
- We perform this experiment on **category-selective** regions (EBA, FFA, PPA)

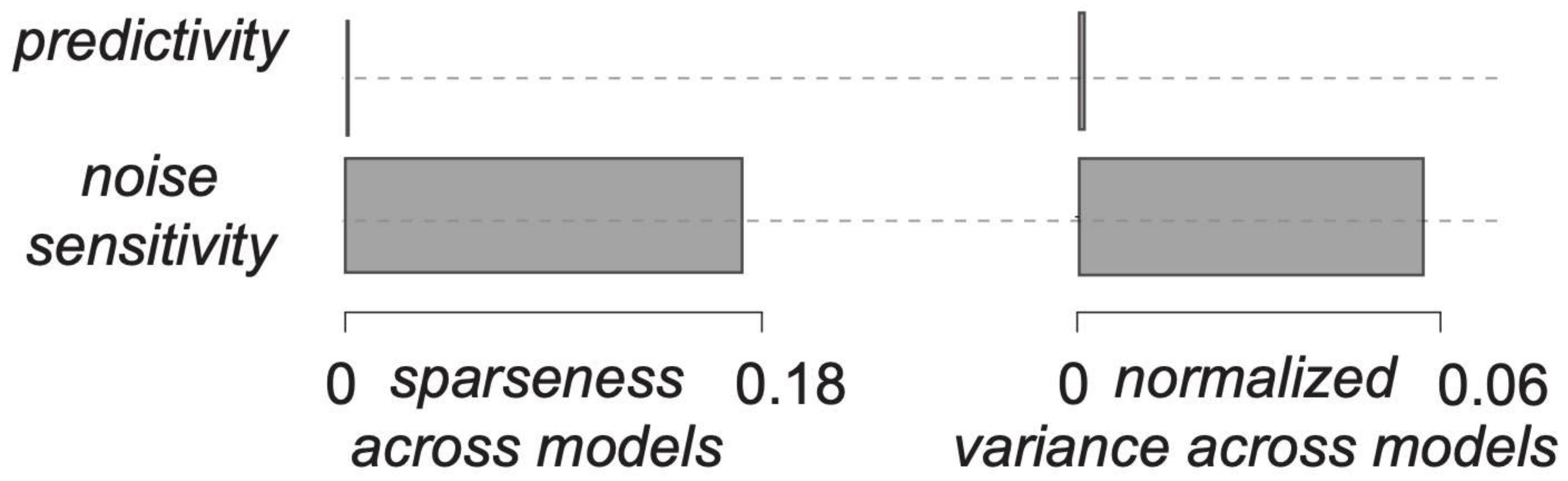
- Natural Scenes Dataset (NSD) for all analyses (515 shared stimuli, eight subjects)

- The readout of 14 pre-trained neural networks is is fit via linear regression to predict brain responses

Does adversarial sensitivity discriminate between equally well-predicting brain models?

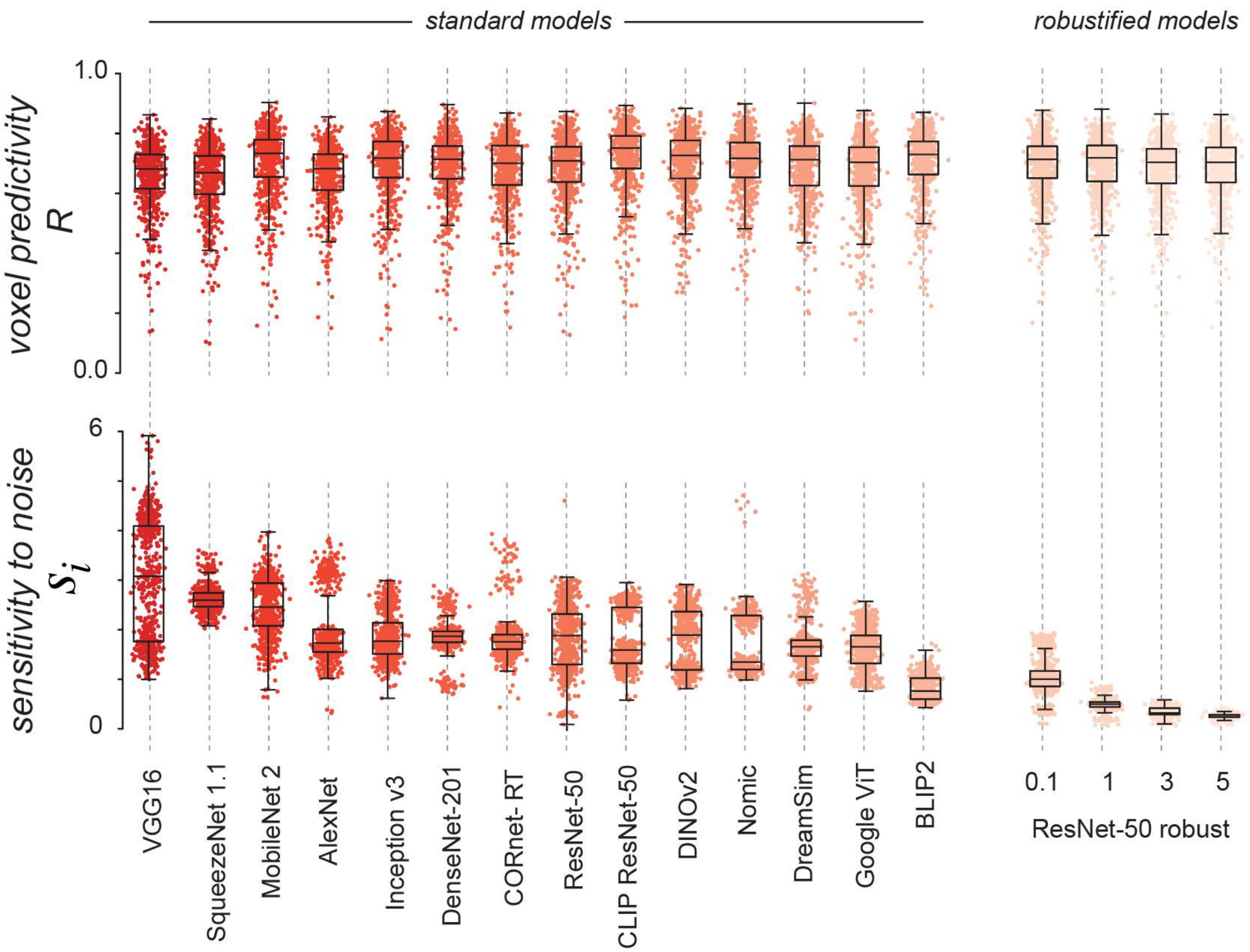






How do we find predictive models with ***low sensitivity***?

- *Robustified neural networks*
    - Models trained using an *adversarial loss* function
    - Trained to correctly classify adversarial images during training
    - Do *robustified ResNet-50* models solve our problem?



standard models
robustified models
voxel predictivity R
1.0
0.0
sensitivity to noise s_i
6
0
VGG16
SqueezeNet 1.1
MobileNet 2
AlexNet
Inception v3
DenseNet-201
CORnet- RT
ResNet-50
CLIP ResNet-50
DINOv2
Nomic
DreamSim
Google ViT
BLIP2
0.1
1
3
5
ResNet-50 robust

- How sensitive are brain models to adversarial attacks?

  Models are **very** sensitive to adversarial attacks

- Do models share the same failure modes?

- Can we use stability to find better models of the brain?

- Can we use stable+predictive models to generate hypotheses about the brain?

- How sensitive are brain models to adversarial attacks?

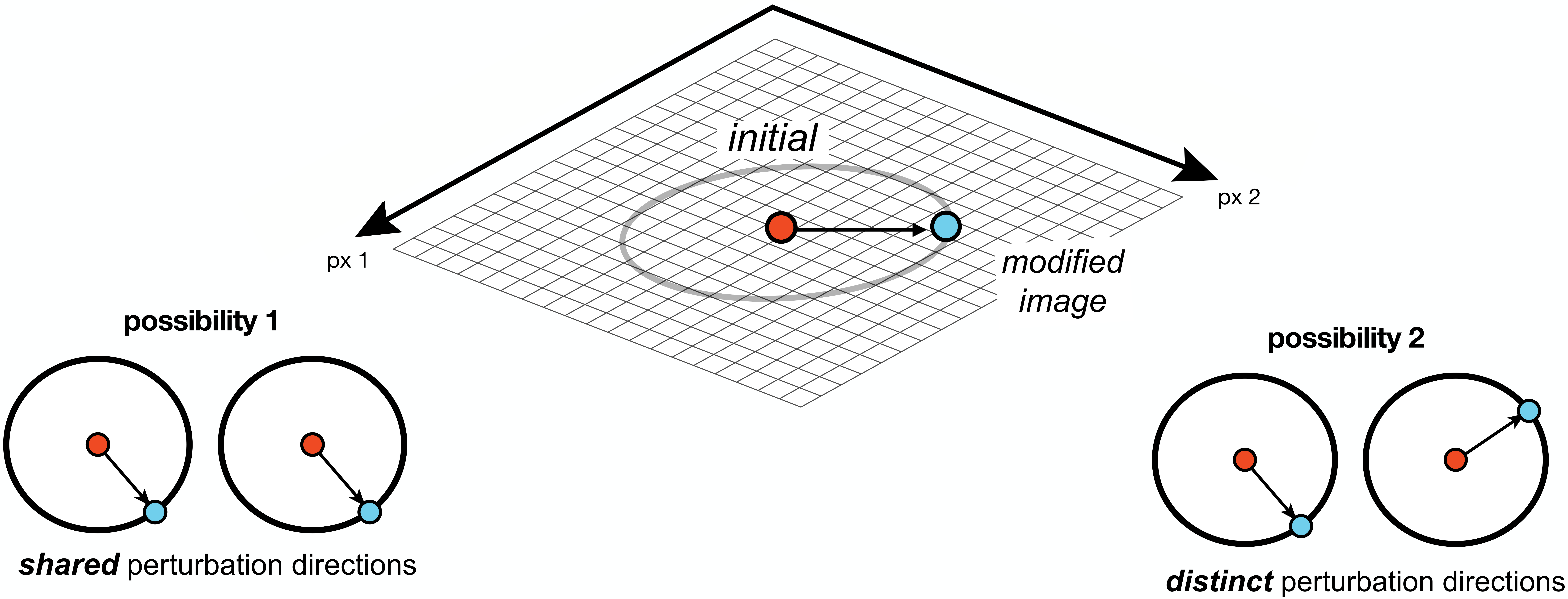
  Models are **very** sensitive to adversarial attacks

- Do models share the same failure modes?

- Can we use stability to find better models of the brain?

- Can we use stable+predictive models to generate hypotheses about the brain?

Do models share the same failure modes?



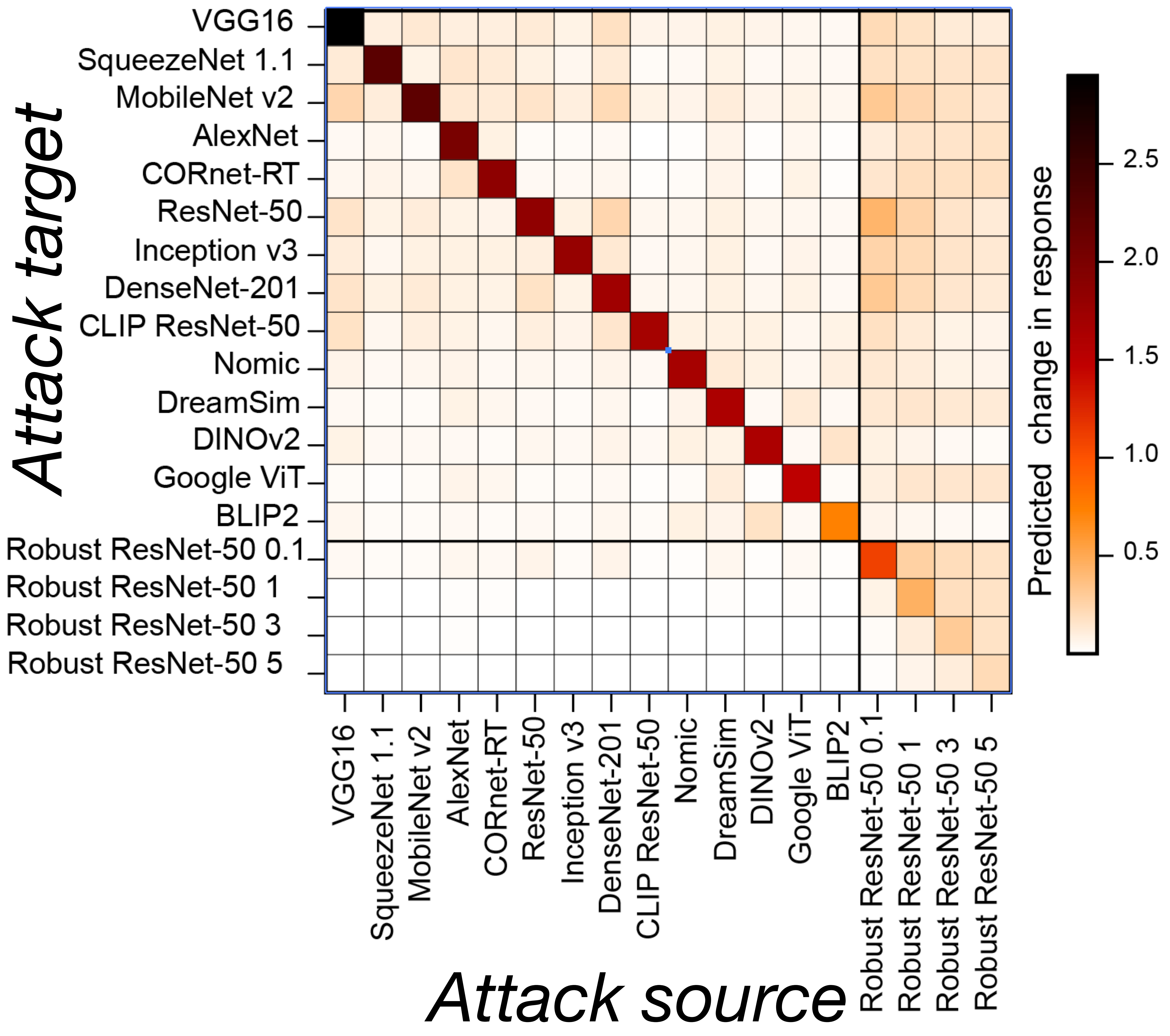
Do models share the same failure modes?

- If a model is sensitive to an **adversarial direction**, will other models also be sensitive to this direction?

- Darker colors represent **more transfer**

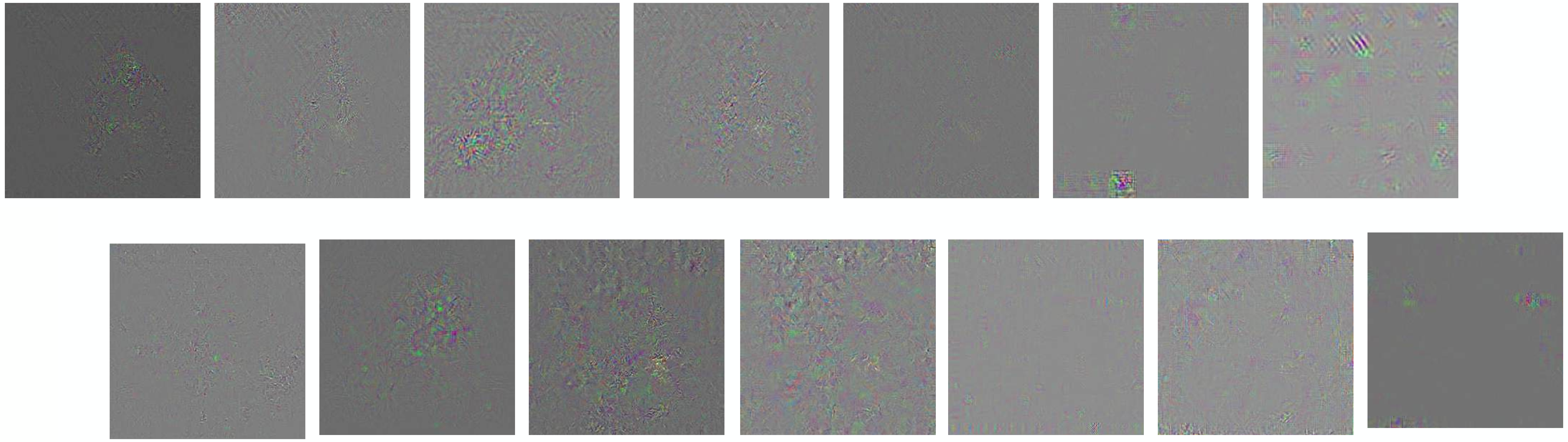
- Overall **limited transfer** between models

- Robust models tend to **transfer** a bit more



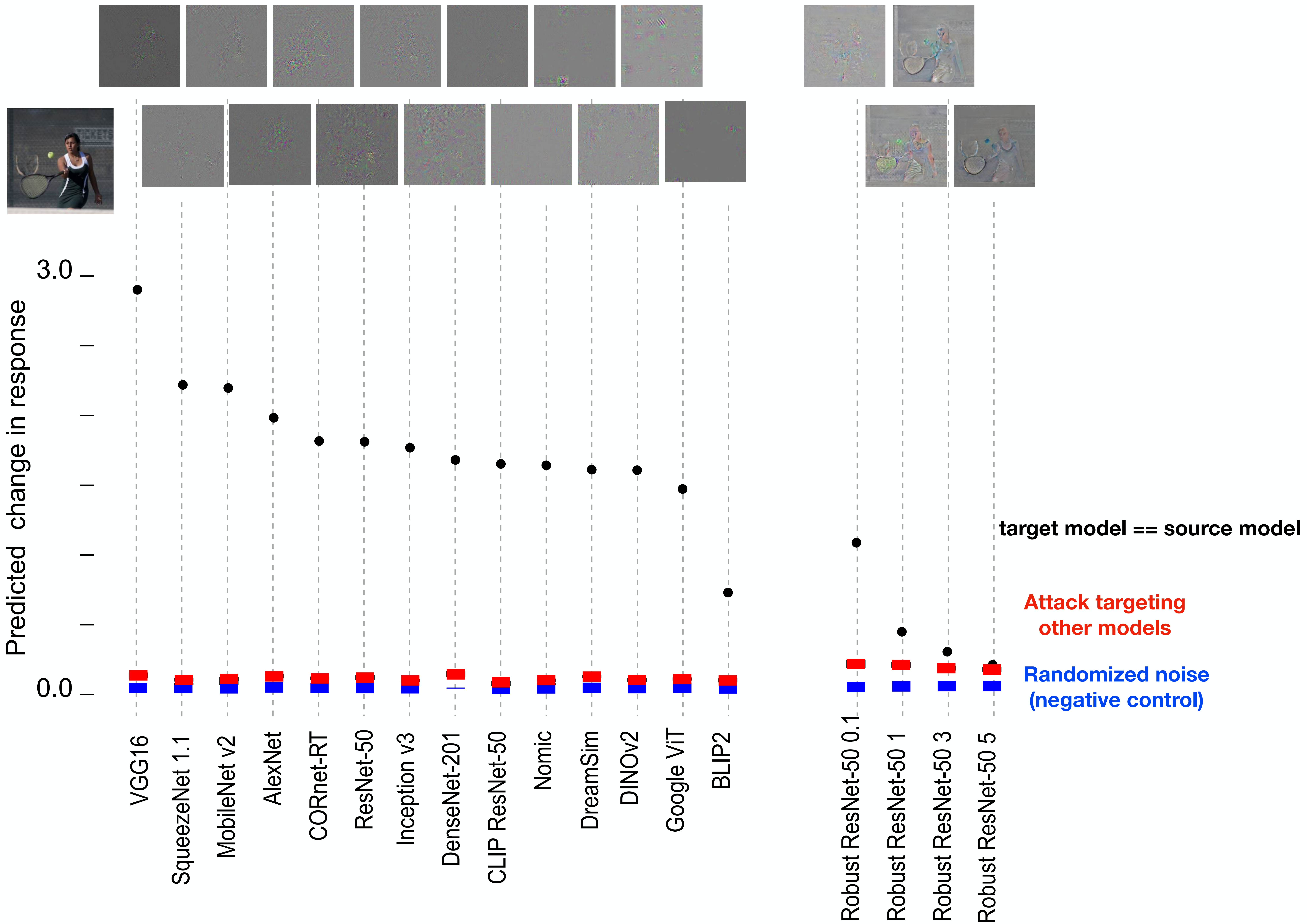
- Another way of thinking about this…
    - Consider this image
      And the associated noise patterns

Standard models



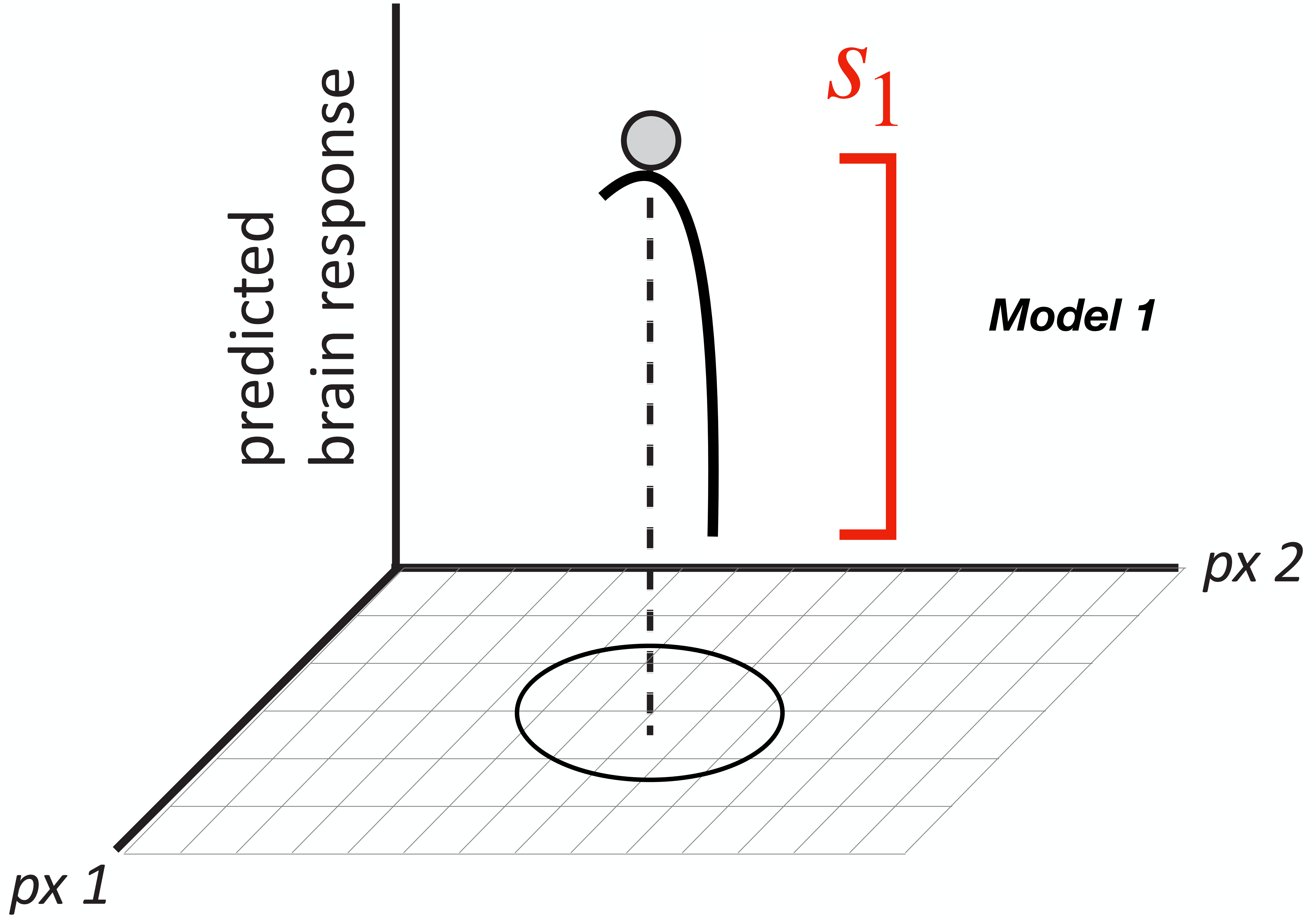
Robust models

Which of these sensitive directions **transfer** to other models?



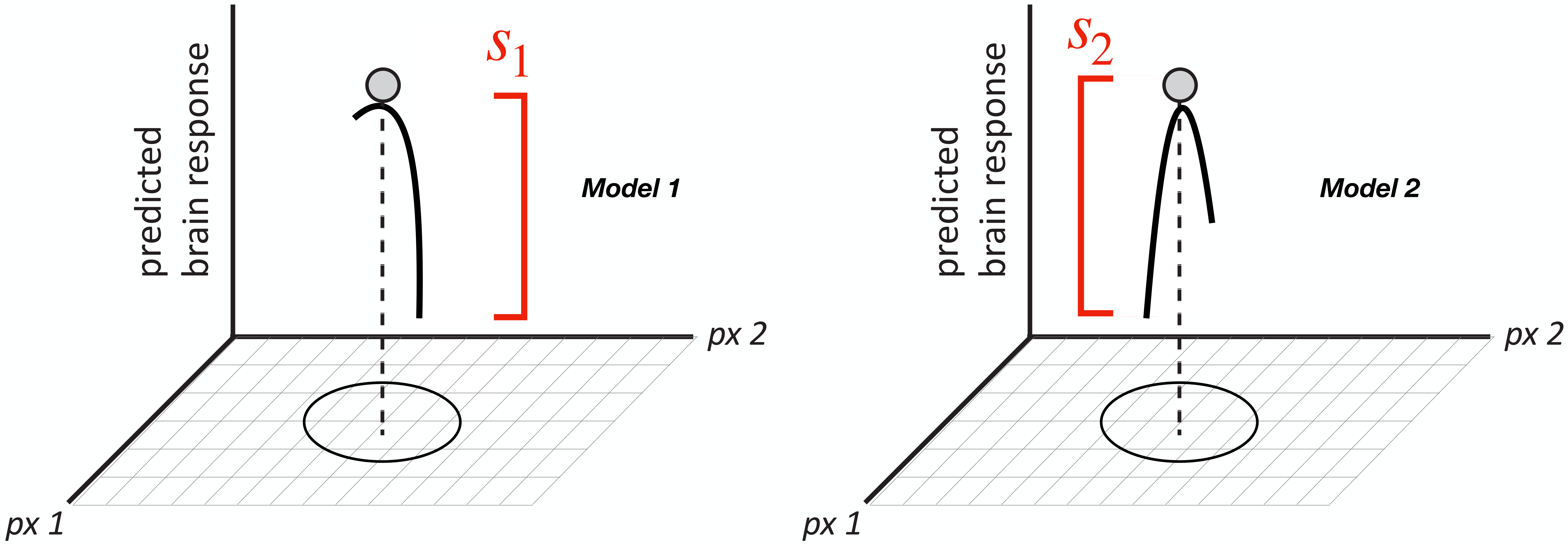
Predicted change in response
3.0
0.0
VGG16
SqueezeNet 1.1
MobileNet v2
AlexNet
CORnet-RT
ResNet-50
Inception v3
DenseNet-201
CLIP ResNet-50
Nomic
DreamSim
DINOv2
Google ViT
BLIP2
Robust ResNet-50 0.1
Robust ResNet-50 1
Robust ResNet-50 3
Robust ResNet-50 5
target model == source model
Attack targeting other models
Randomized noise (negative control)

So far…



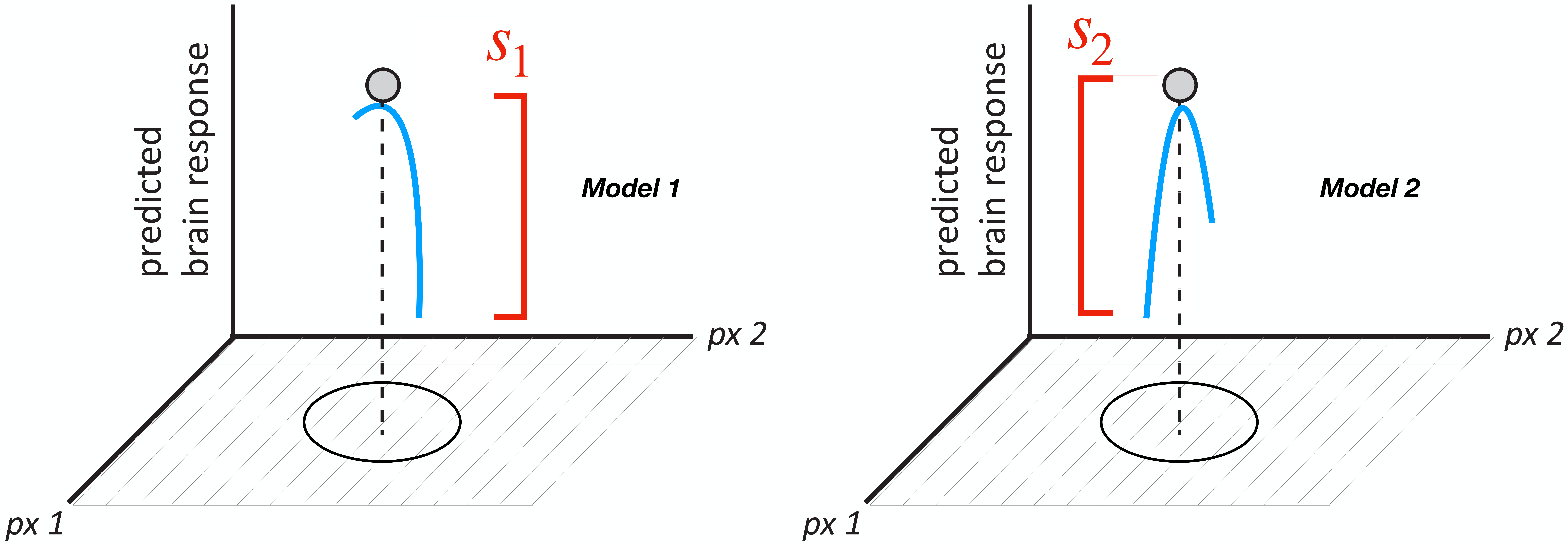
We've estimated the **adversarial sensitivity (s)** of each model

So far…



We've estimated the **adversarial sensitivity (s)** of each model

$$s = r(x) - r(x+\delta), \quad \delta = \underset{\|\delta\| \leq \epsilon}{\arg\max} \left[|r(x) - r(x+\delta)| \right]$$

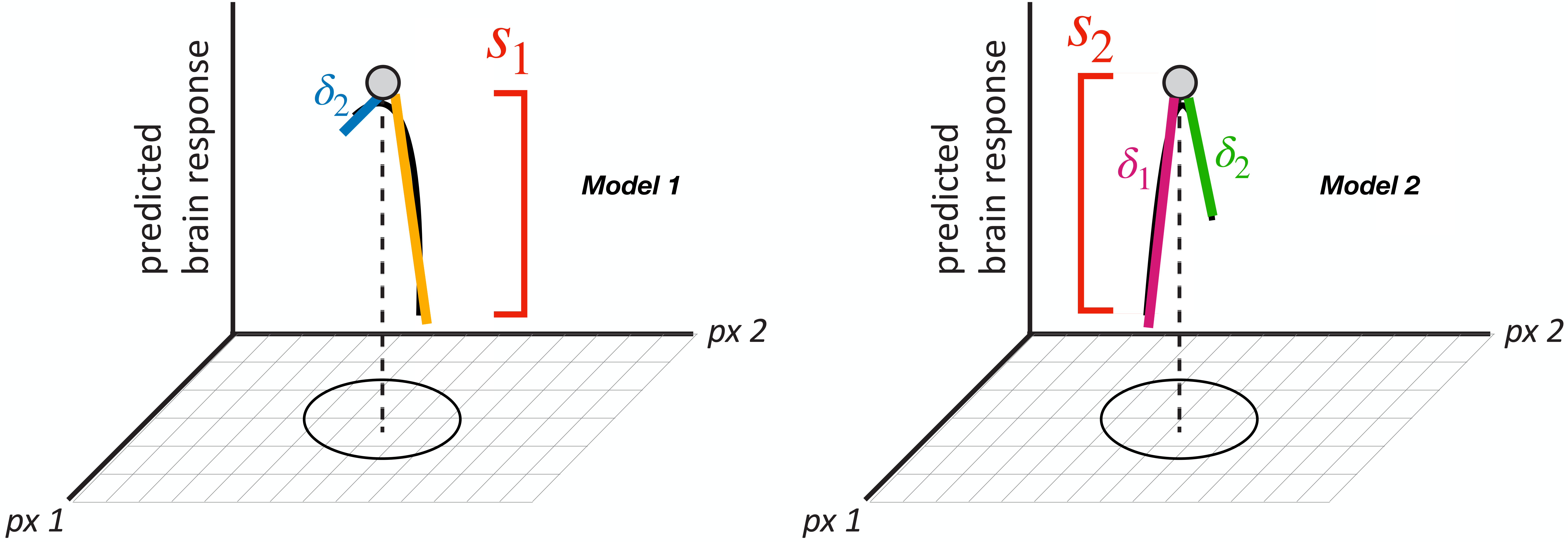


We've estimated the **adversarial sensitivity (s)** of each model

Two models with the same ***sensitivity*** might have different ***representational geometries***

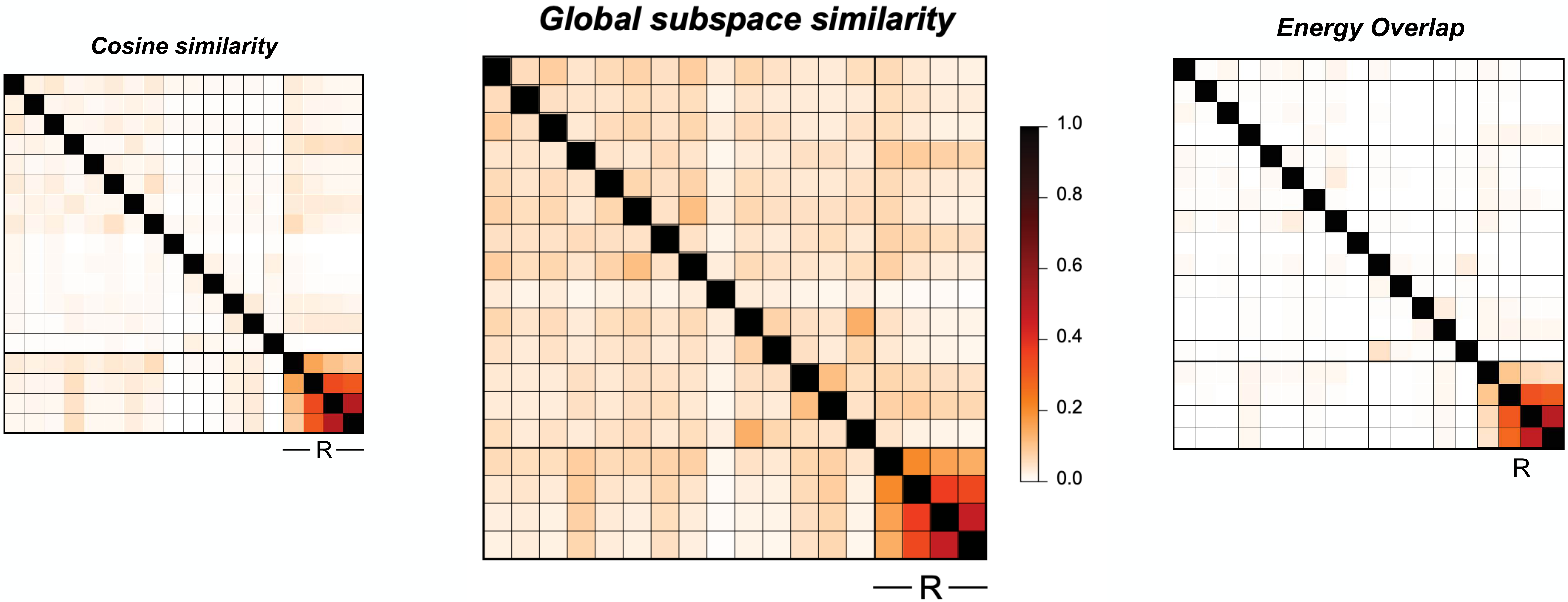
How do we characterize differences in geometry?

By looking at the *principle directions*



Stack the directions in a matrix,
forming a **perturbation subspace**

How similar are perturbation subspaces across models?



We wish to maximize the distance between $r(x)$ and $r(x+\delta)$

$$r(x) = r(x)$$

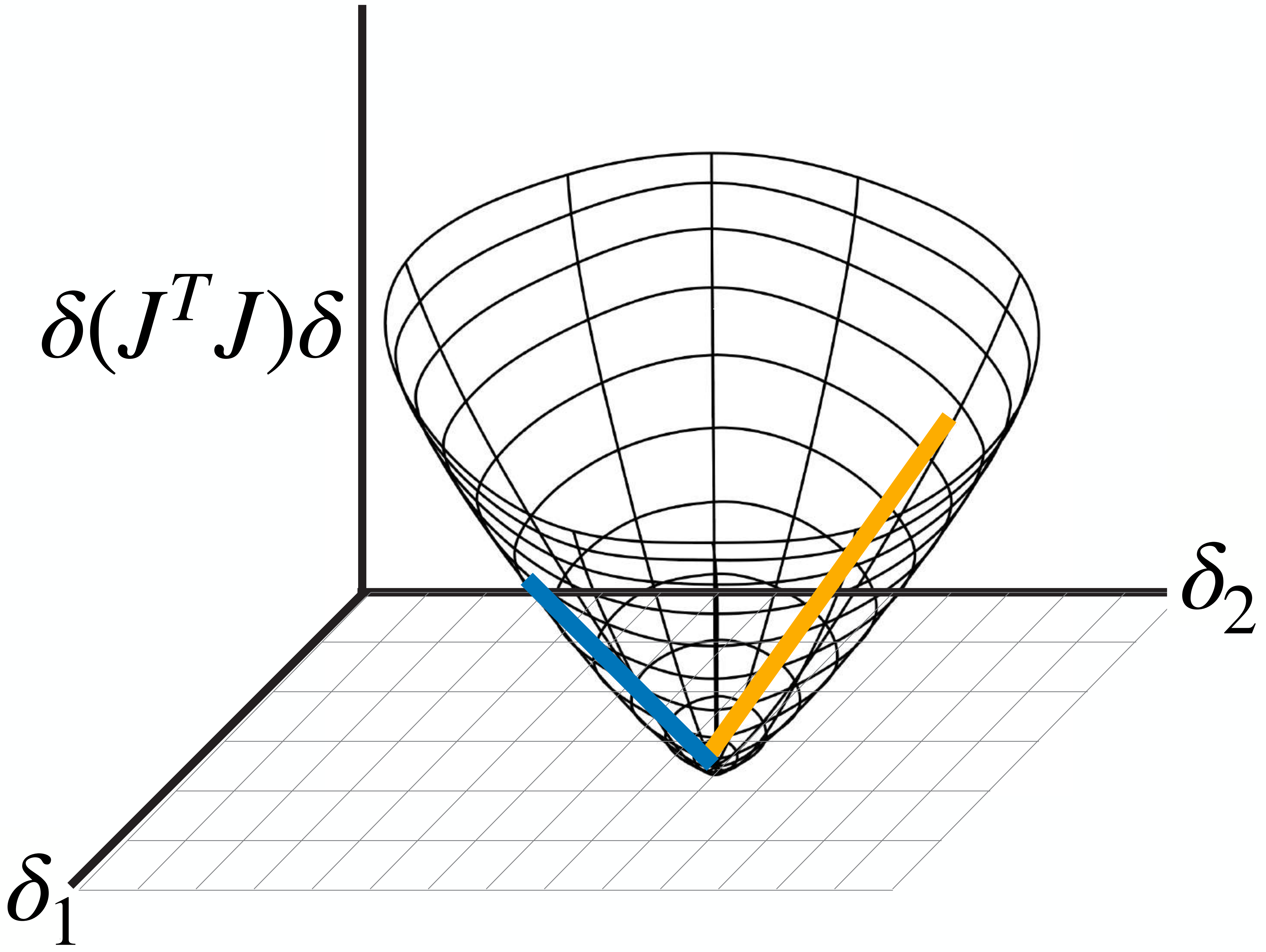
$$r(x+\delta) = r(x) + J\delta + \cancel{\frac{1}{2}\delta^T H(x+\theta\delta)\delta}$$

(first-order)
Taylor expansion

$$\Delta r \equiv r(x) - r(x+\delta) \approx J\delta$$

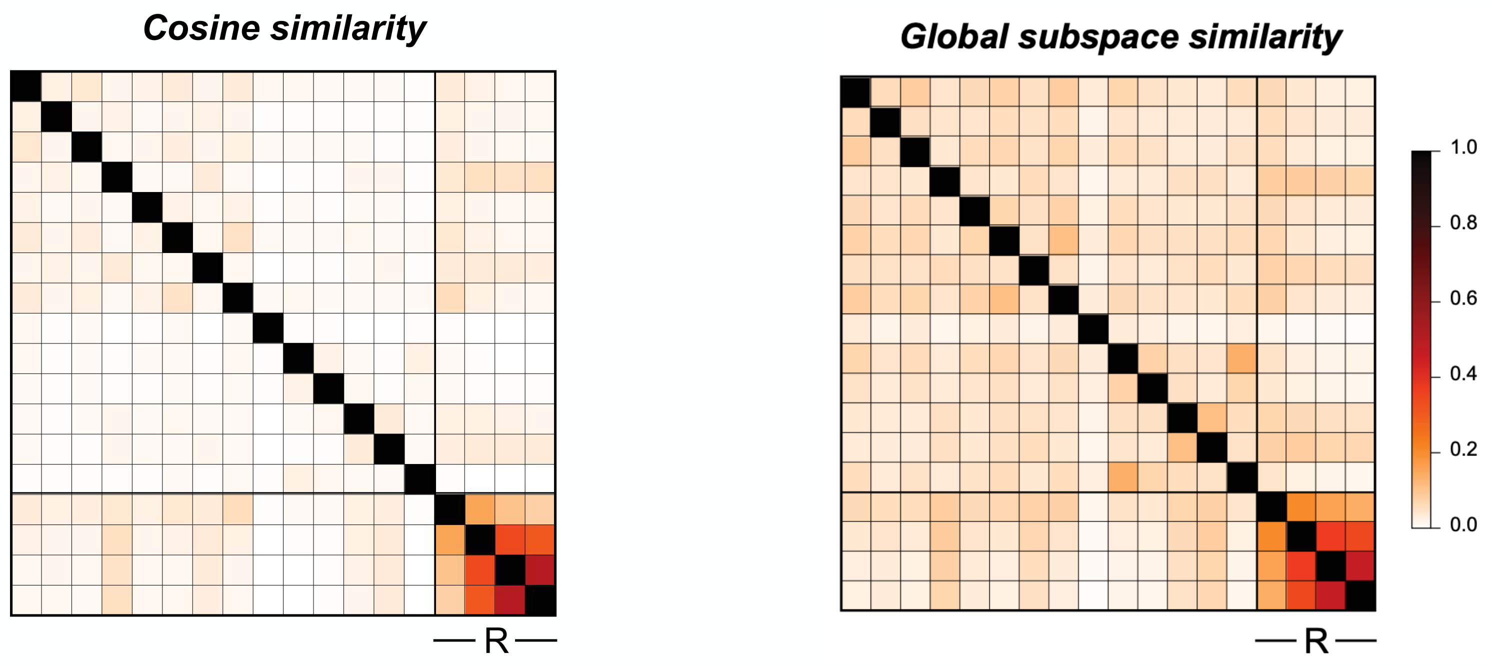
$$||\Delta r||_2^2 = \delta^T (J^T J)\delta$$

We want to find directions δ which **maximize** $\delta(J^T J)\delta$
(the eigenvectors!)

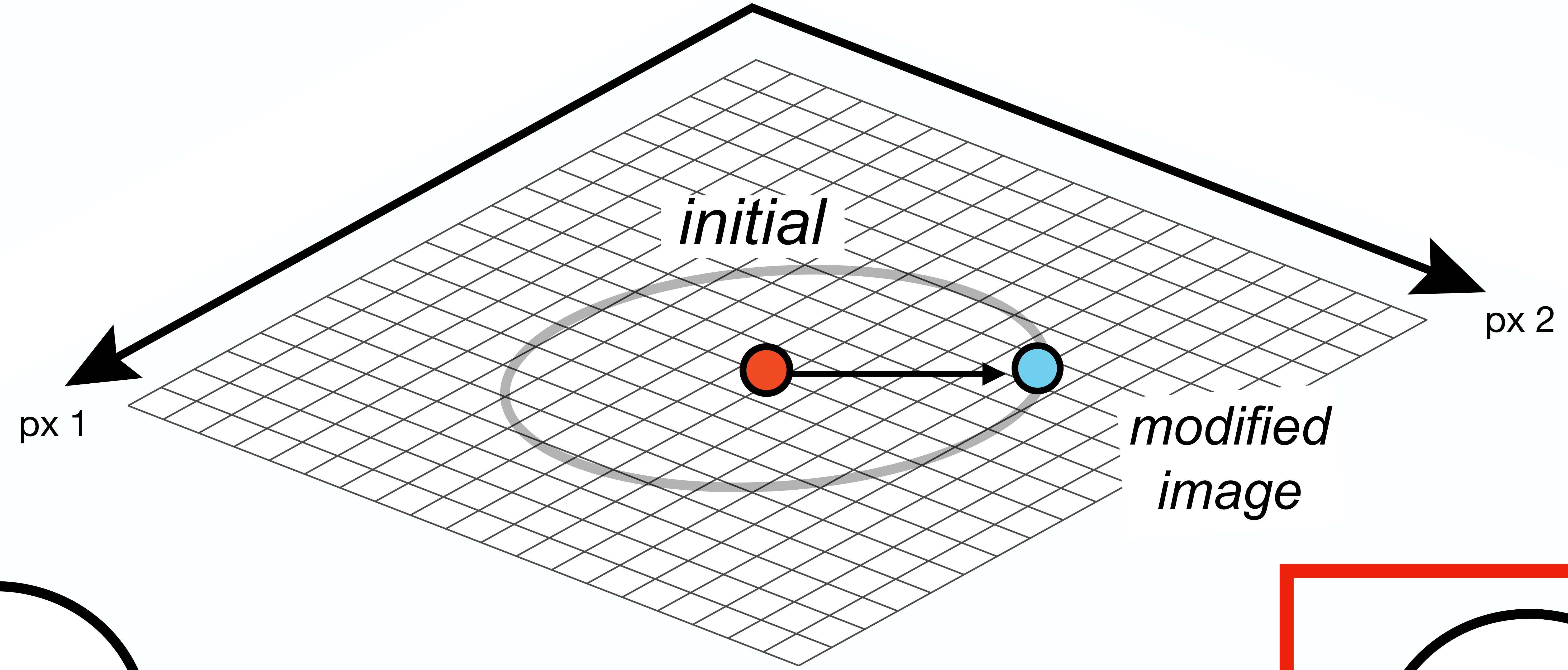


Top-k eigenvectors stacked in perturbation matrices P_i, P_j

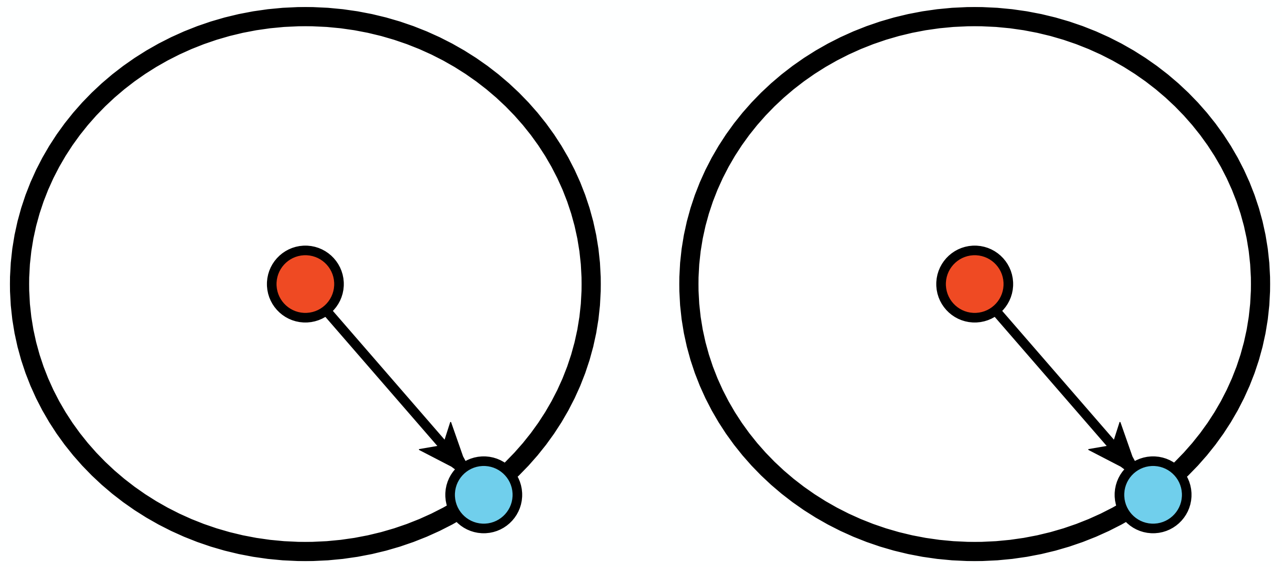
Similarity measurement $Sim(P_i, P_j)$



Do models share the same failure modes?

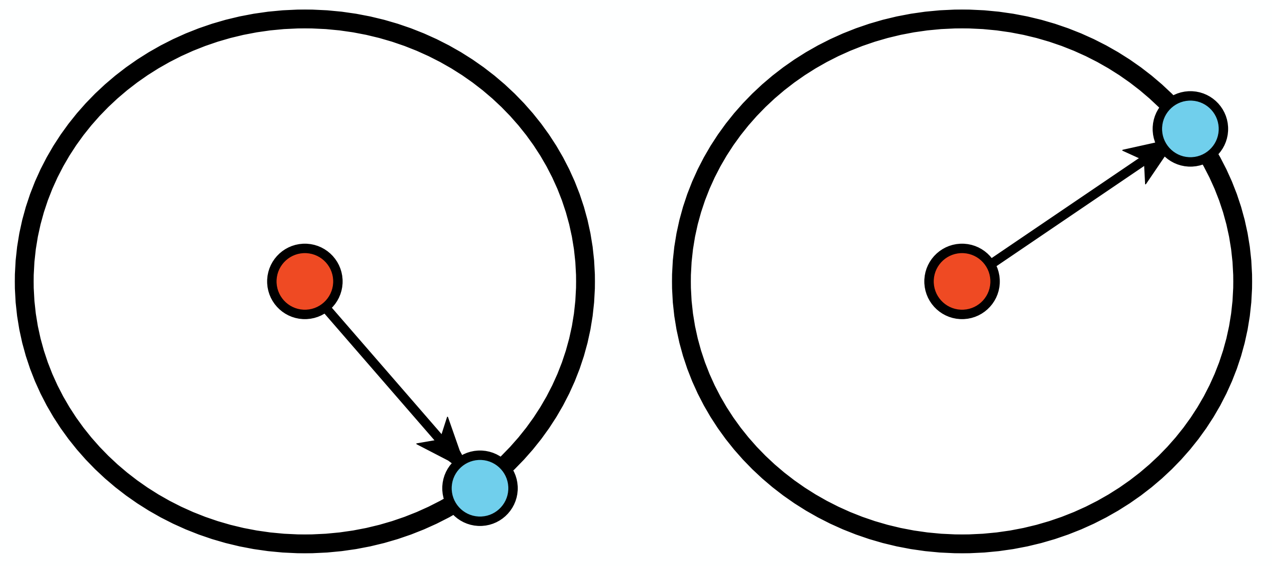


possibility 1



shared perturbation directions

possibility 2



distinct perturbation directions

- How sensitive are brain models to adversarial attacks?

  Models are **very** sensitive to adversarial attacks

- Do models share the same failure modes?

  Standard models generally have **distinct** failure modes; robust models have **shared** directions

- Can we use stability to find better models of the brain?

- Can we use stable+predictive models to generate hypotheses about the brain?

- How sensitive are brain models to adversarial attacks?

  Models are **very** sensitive to adversarial attacks

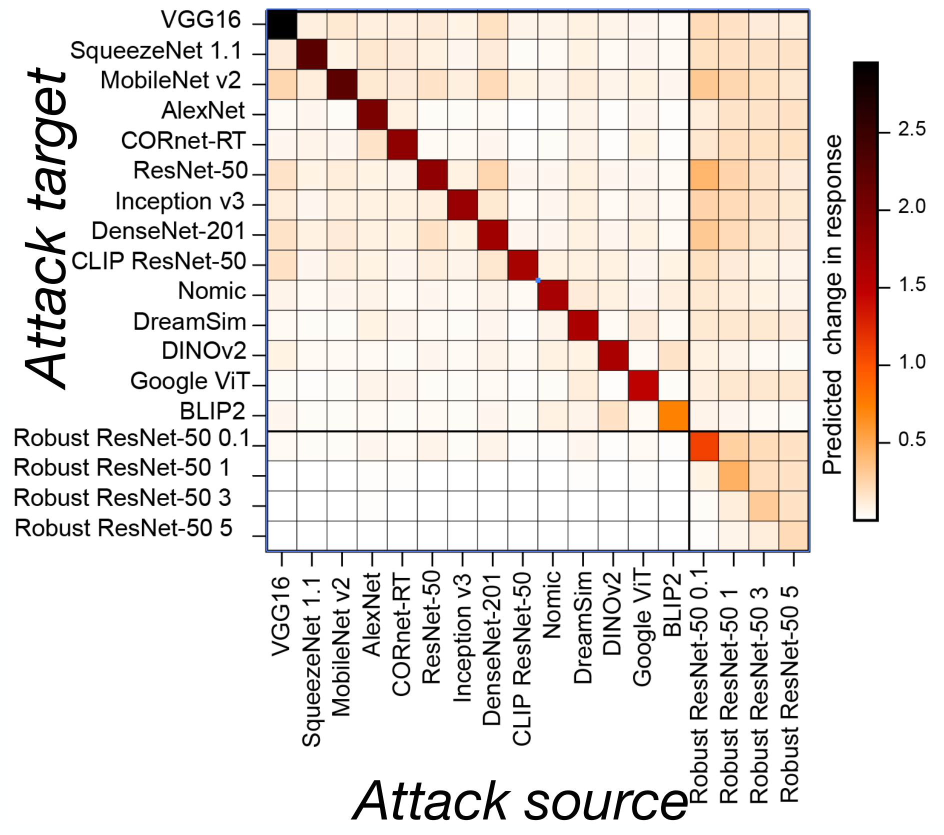
- Do models share the same failure modes?

  Standard models generally have **distinct** failure modes; robust models have **shared** directions

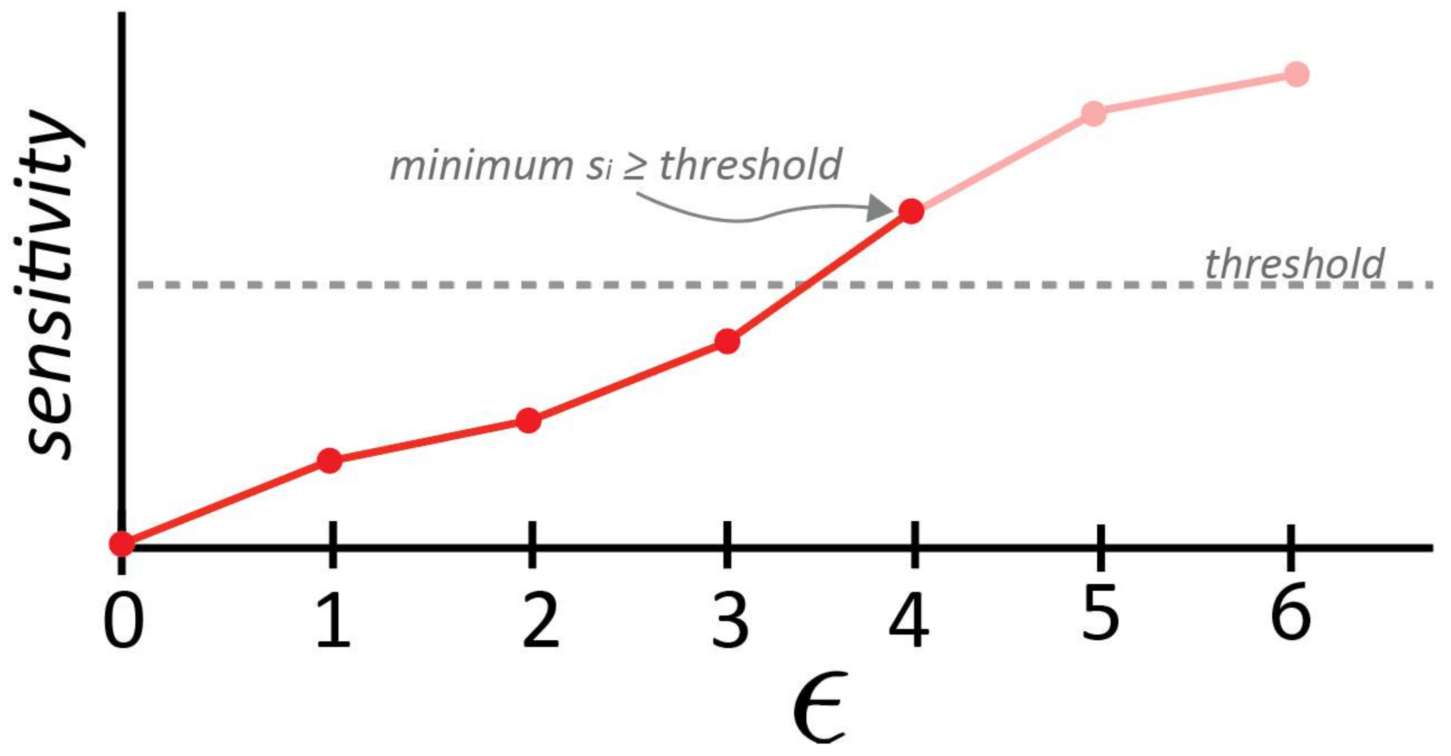
- Can we use stability to find better models of the brain?

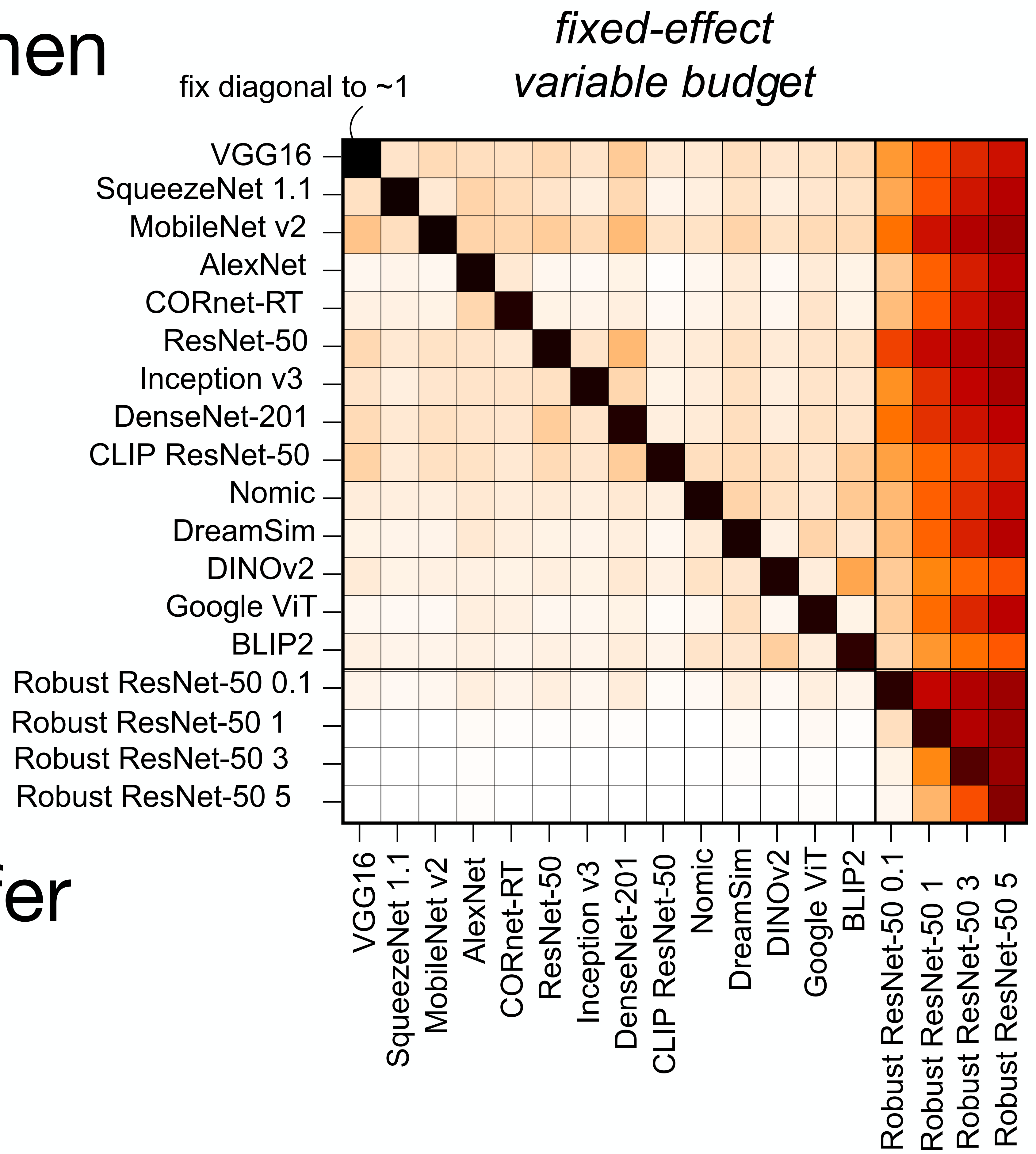
- Can we use stable+predictive models to generate hypotheses about the brain?

- Earlier, we fixed the **perturbation size** and evaluated how well attacks transferred between models



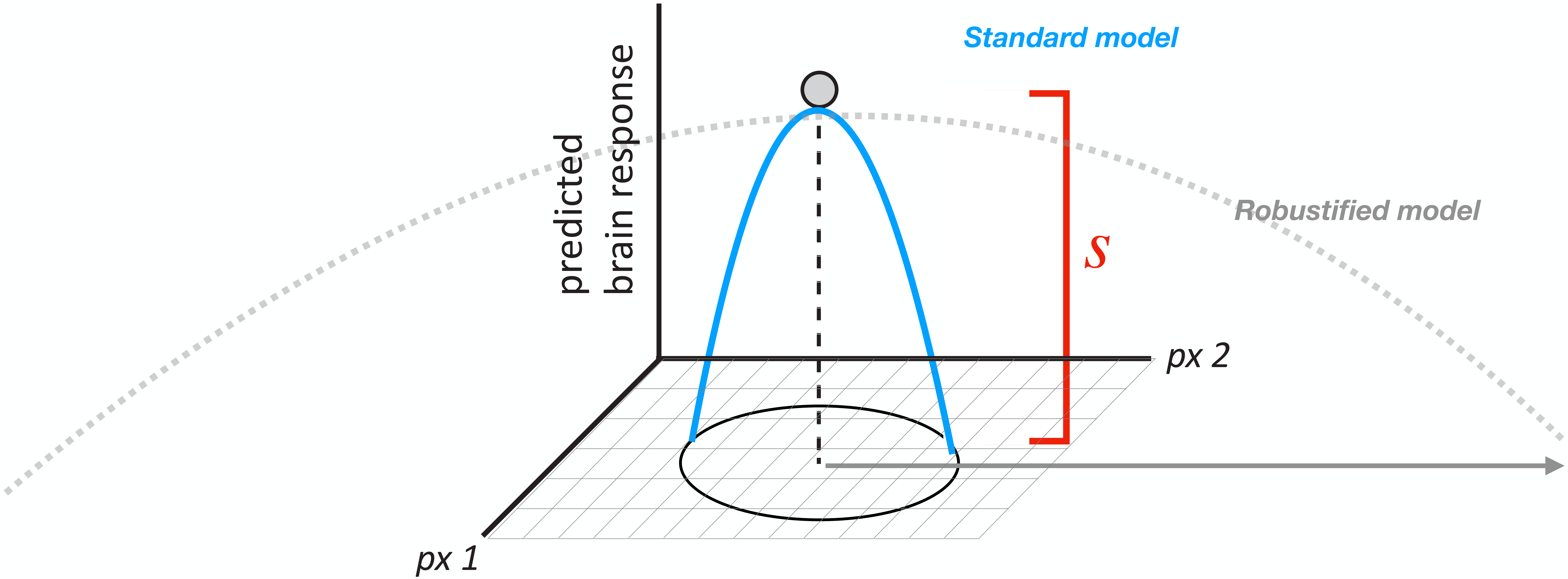
- What if we fix the **transfer effect first,** and then evaluate how well attacks transfer between models?





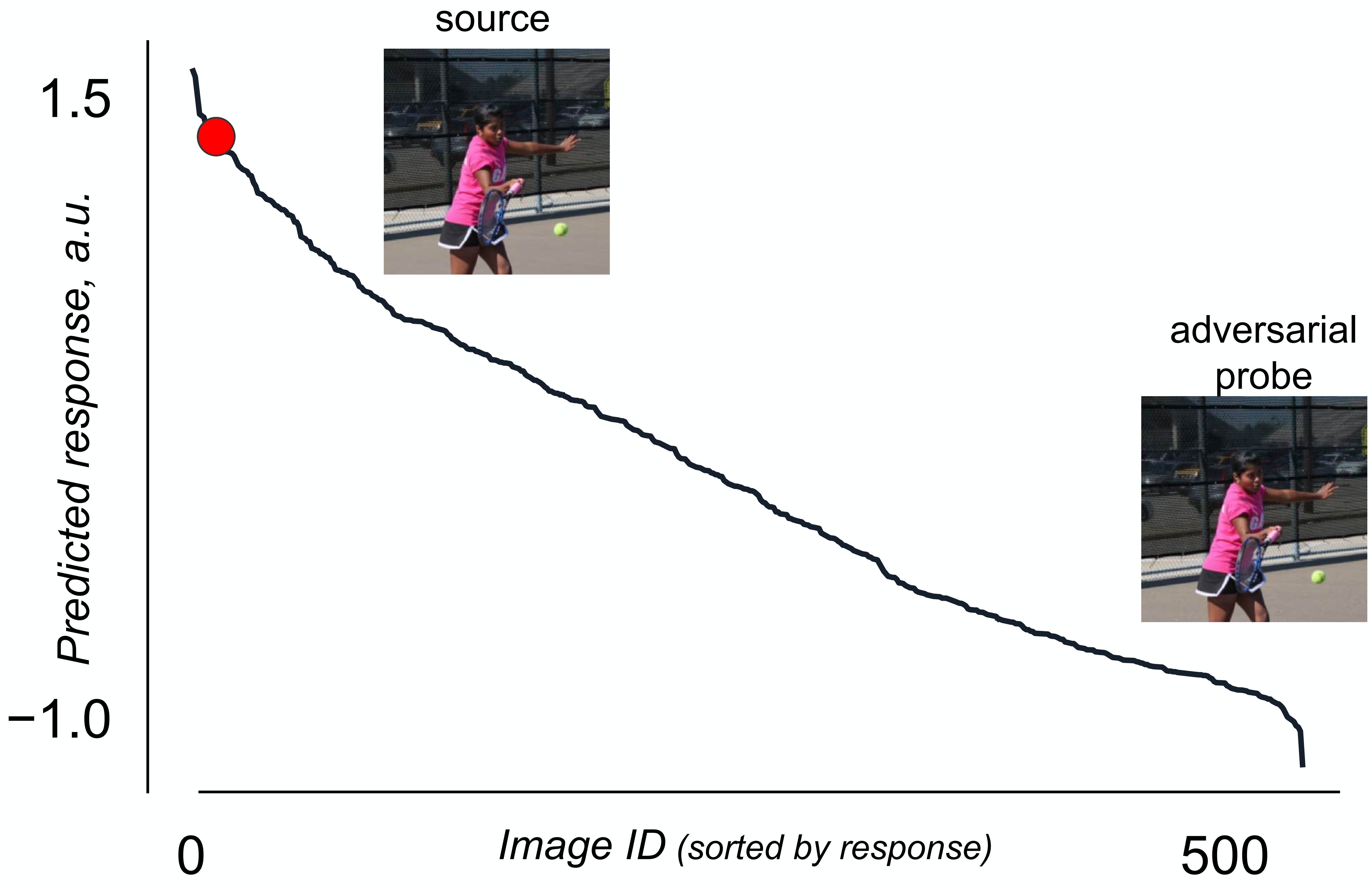
- Perturbations from robust models now transfer to all models! Possibly the **brain-like coding axis?**

Another way to think about this…



To achieve the same sensitivity, we need a much larger perturbation

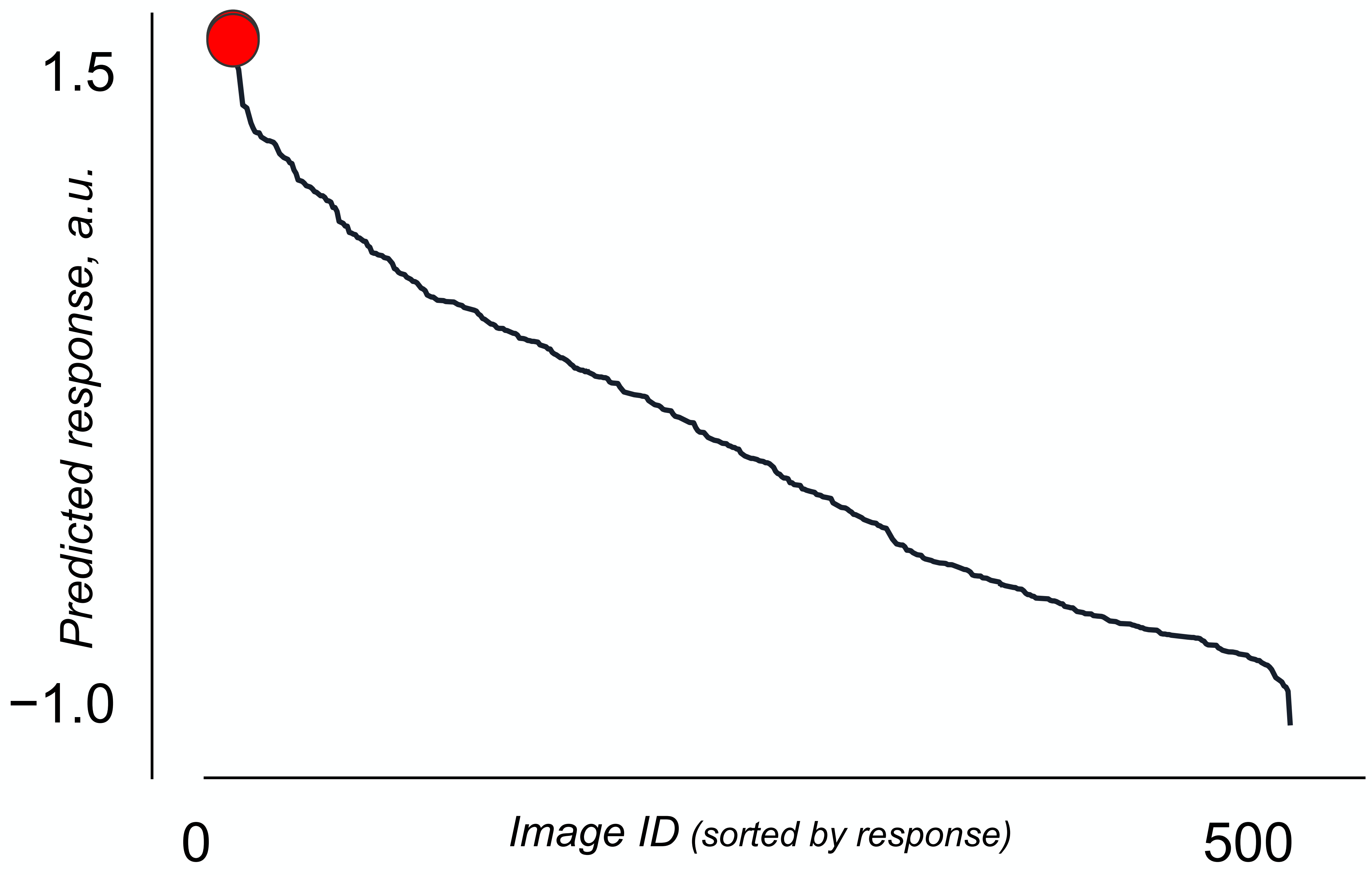
- Earlier, we saw how **imperceptible** noise patterns can drastically alter standard model predictions



Since *robustified* models need a larger perturbation size, what does the adversarial probe look like?

Consider an image with a **face** in it with a high **predicted FFA response**

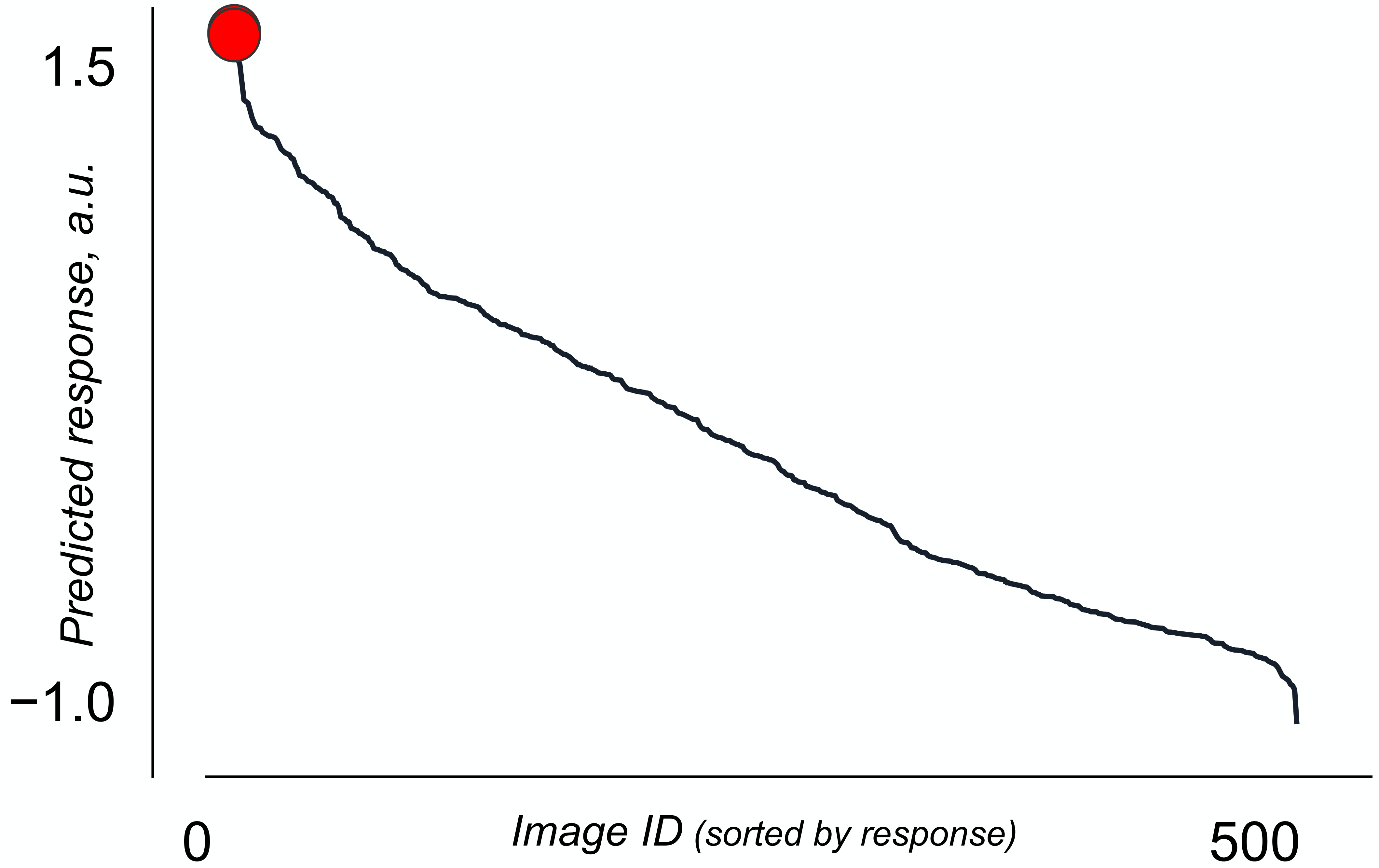
Let's use a **robustified** model to find a perturbation **which minimizes the predicted FFA response**



(video)

Another example…

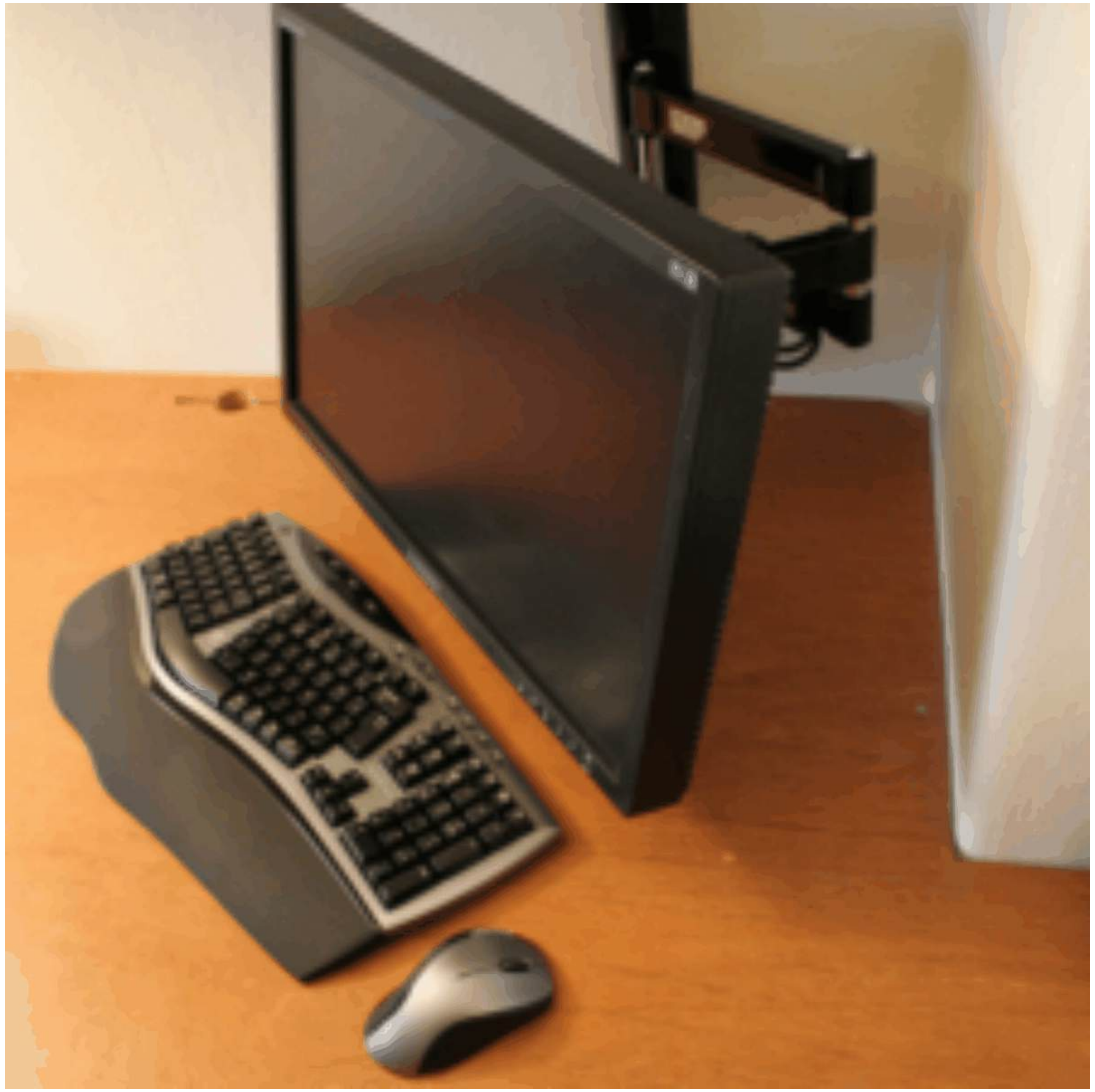
Let’s use a **robustified** model to find a perturbation **which minimizes the predicted FFA response**



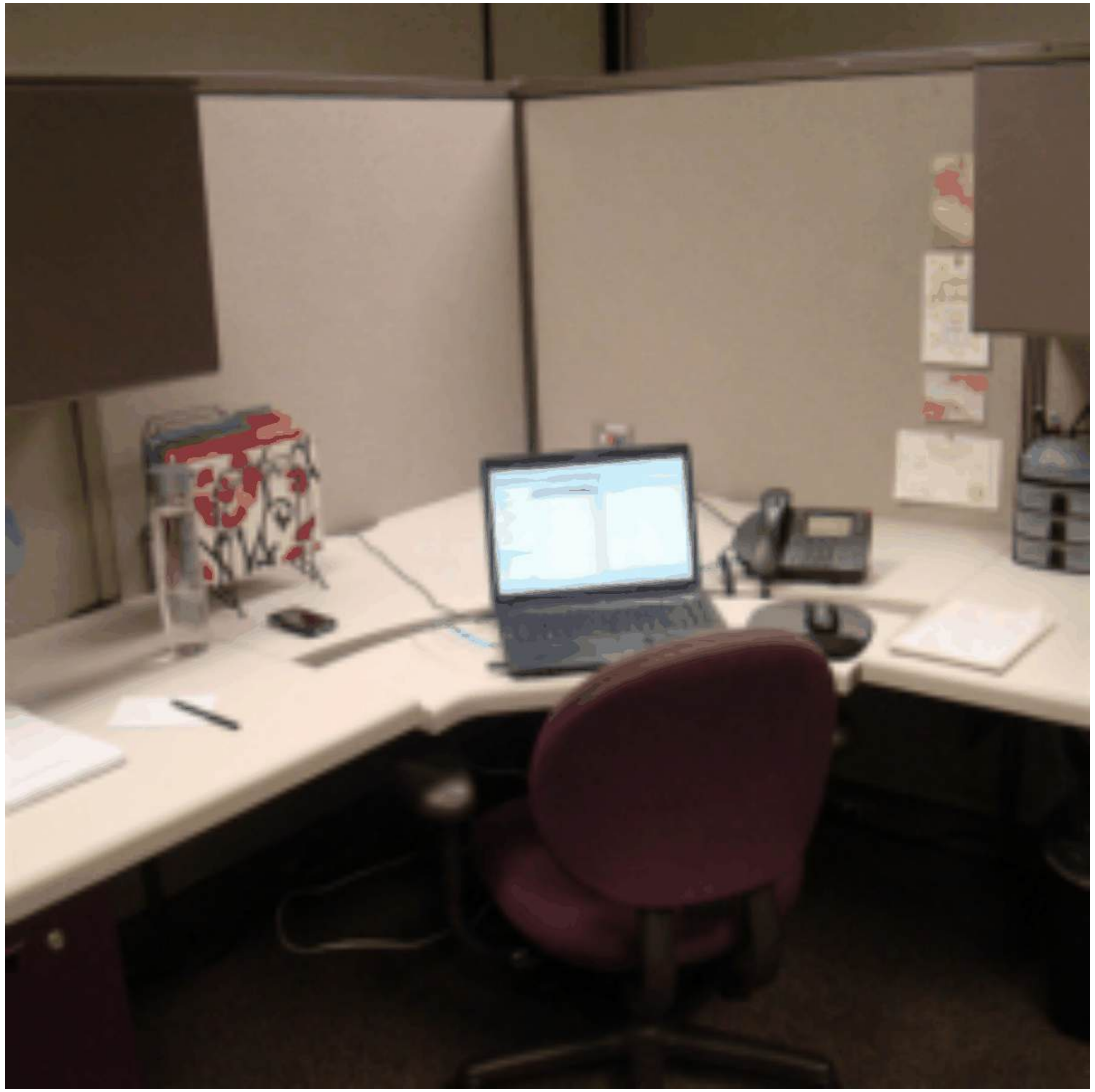
(video)

A challenging example… maximize **EBA**?

(video)



(video)

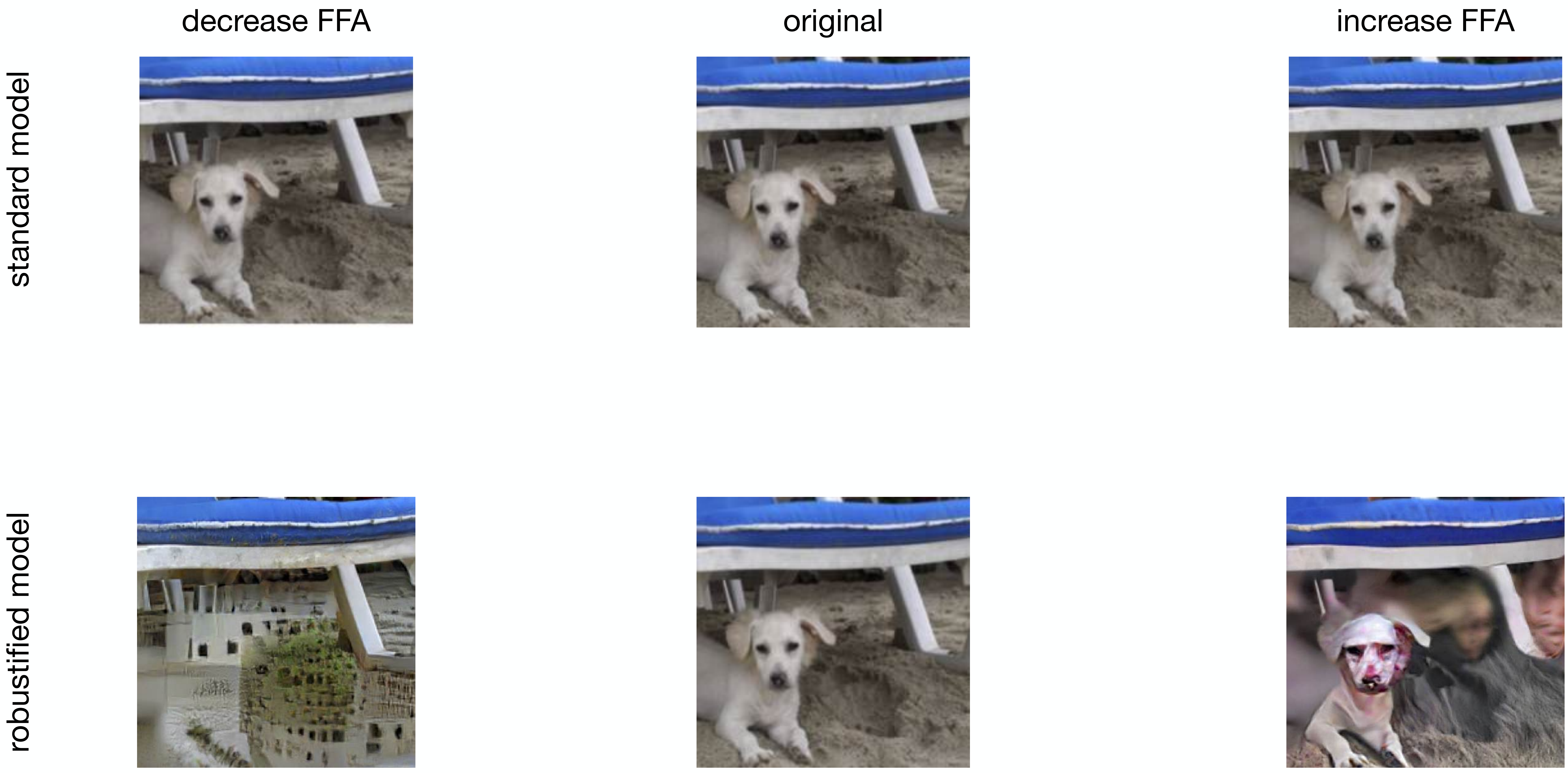


minimize EBA

(video)

maximize PPA

(video)


decrease FFA
original
increase FFA
standard model
robustified model

- How sensitive are brain models to adversarial attacks?

  Models are **very** sensitive to adversarial attacks

- Do models share the same failure modes?

  Standard models generally have **distinct** failure modes; robust models have **shared** directions

- Can we use stability to find better models of the brain?

  Yes! Robustified models have *interpretable, semantically meaningful* features, whereas standard models are **unstable** and **brittle**

- Can we use stable+predictive models to generate hypotheses about the brain?

- How sensitive are brain models to adversarial attacks?

  Models are **very** sensitive to adversarial attacks

- Do models share the same failure modes?

  Standard models generally have **distinct** failure modes; robust models have **shared** directions

- Can we use stability to find better models of the brain?

  Yes! Robustified models have *interpretable, semantically meaningful* features, whereas standard models are **unstable** and **brittle**

- Can we use stable+predictive models to generate hypotheses about the brain?

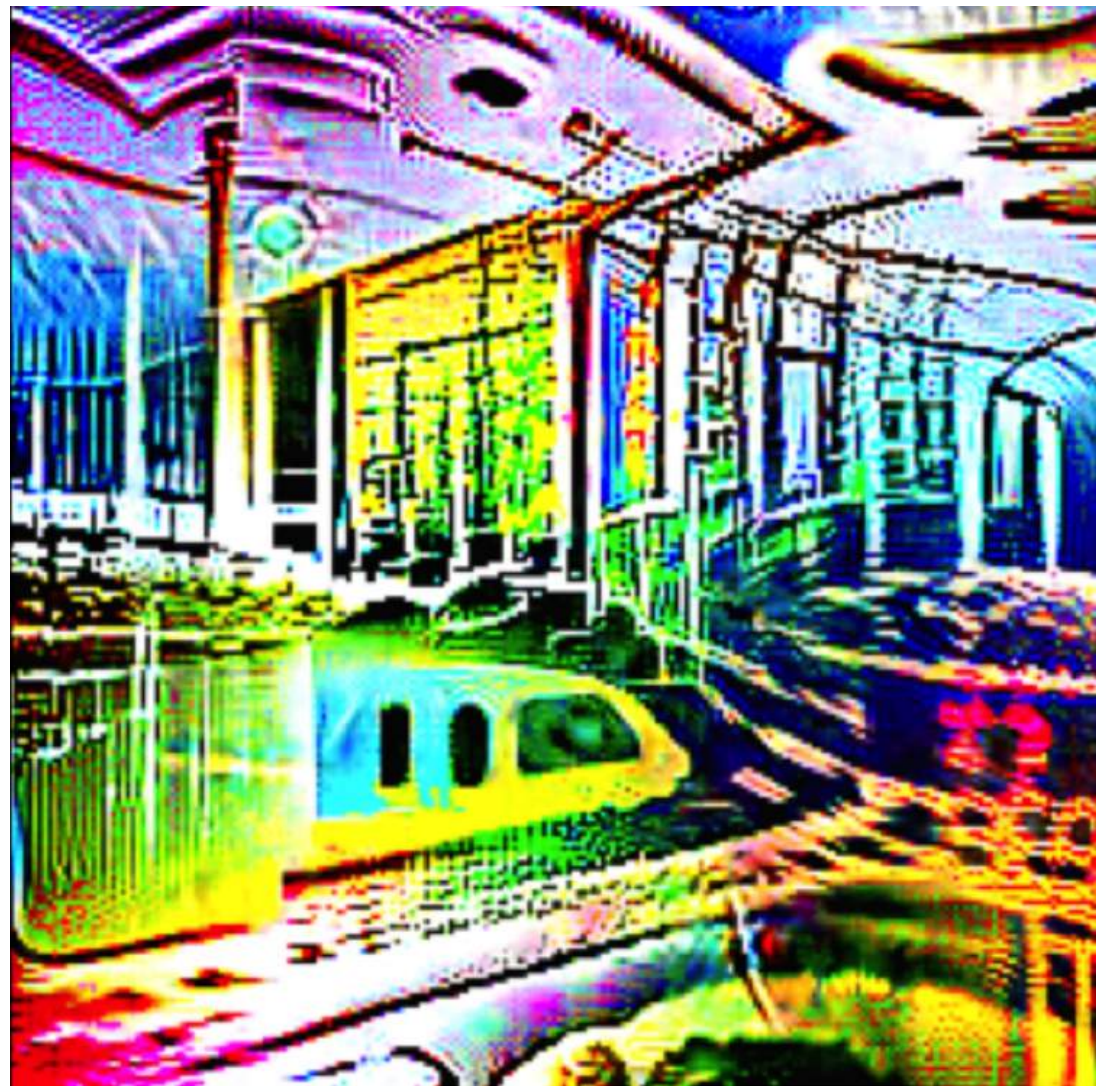
Hypothesis generation with robust encoding models

- In an unconstrained problem, we can continue maximizing the predicted response of a region to obtain a maximally exciting image

FFA

This may reveal the features encoded by a certain brain region

PPA



Many methods for identifying to what features a brain region is selective

BrainACTIV: Identifying visuo-semantic properties driving cortical selectivity using diffusion-based image manipulation

Diego García Cerdas[1,*], Christina Sartzetaki[1], Magnus Petersen[2], Gemma Roig[2], Pascal Mettes[1] and Iris Groen[1]

[1]Informa
[2]Departm
Corresp

Computational models of category-selective brain regions enable high-throughput tests of selectivity

N. Apurva Ratan Murty[1,2,3,5], Pouya Bashivan[4],
Nancy Kanwisher[1,2,3]

Energy Guided Diffusion for Generating Neurally Exciting Images

Brain Diffusion for Visual Exploration: Cortical Discovery using Large Scale Generative Models

Andrew F. Luo
Carnegie Mellon University
afluo@cmu.edu

Margaret M. Henderson
Carnegie Mellon University
mmhender@cmu.edu

Leila Wehbe*
Carnegie Mellon University
lwehbe@cmu.edu

Michael J. Tarr*
Carnegie Mellon University
michaeltarr@cmu.edu

mized image synthesis for discovery

Meenakshi Khosla[a], Emily J. Allen[c,d], Yihan
selaris[c,d], Kendrick Kay[c], Mert R. Sabuncu[a], Amy

ineering, Cornell University, Ithaca, New York, USA
Medicine, New York, New York, USA
rch(CMRR), Department of Radiology, University of Minnesota, Minneapolis, Minnesota, USA
[d]Department of Neuroscience, University of Minnesota, Minneapolis, Minnesota, USA

F. Nix[1-2], Pavithra Elumalai[2],
abrielle Rodriguez[3-4],
as[3-5], Fabian H. Sinz[1-4]
n University, Tübingen, Germany
e, University of Göttingen, Germany
icine, Houston, TX, USA
ge of Medicine, Houston, TX, USA
University, Houston, TX, USA
.de

Which method is most accurate in driving brain responses?
And under what constraints?

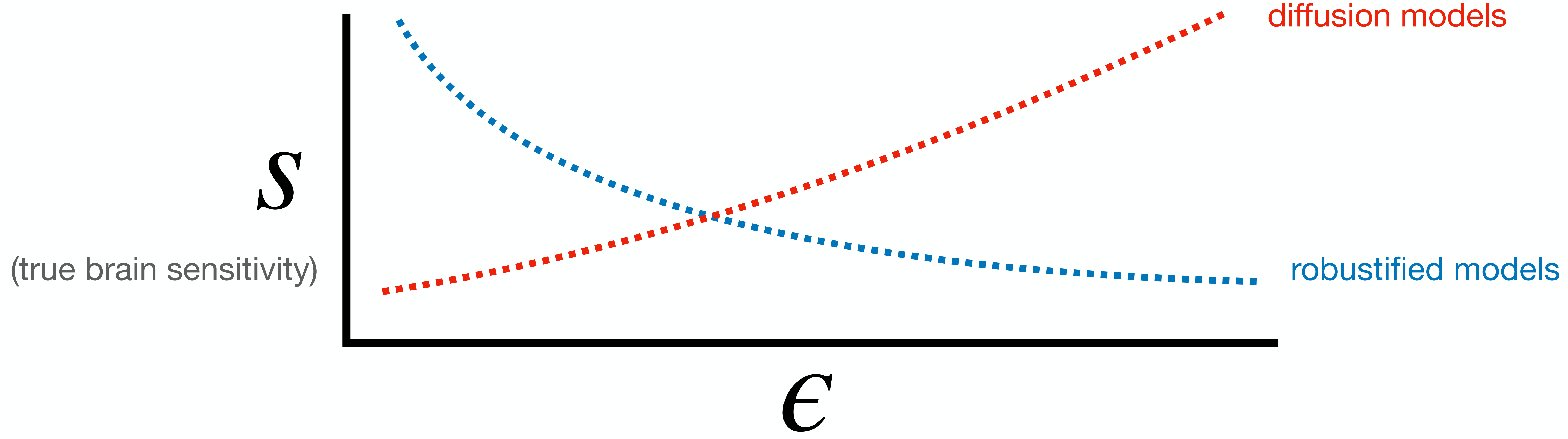
- Many of these methods use *priors* from generative models to guide the sampling

- This may be better for generating *realistic* images (high-norm perturbations), but worse for controlling neural responses to low-norm perturbations:

$$\delta = \arg\max_{||\delta|| \leq \epsilon} \left[|r(x) - r(x+\delta)| \right]$$

- Can we compare all these methods? We need *experimental tests* to validate these methods

An example hypothesis:



We don't know yet!

- How sensitive are brain models to adversarial attacks?

  Models are **very** sensitive to adversarial attacks

- Do models share the same failure modes?

  Standard models generally have **distinct** failure modes; robust models have **shared** directions

- Can we use stability to find better models of the brain?

  Yes! Robustified models have *interpretable, semantically meaningful* features, whereas standard models are **unstable** and **brittle**

- Can we use stable+predictive models to generate hypotheses about the brain?

- How sensitive are brain models to adversarial attacks?

  Models are **very** sensitive to adversarial attacks

- Do models share the same failure modes?

  Standard models generally have **distinct** failure modes; robust models have **shared** directions

- Can we use stability to find better models of the brain?

  Yes! Robustified models have *interpretable, semantically meaningful* features, whereas standard models are **unstable** and **brittle**

- Can we use stable+predictive models to generate hypotheses about the brain?

  Yes, more soon!

Acknowledgements

Murty Lab

Ruolin Wang
Flo Addiego

Alish Dipani
Mayukh Deb
Haider Al Tahan
Junxia Wang
Mainak Deb
Nikolas McNeal

- How sensitive are brain models to adversarial attacks?

  Models are **very** sensitive to adversarial attacks

- Do models share the same failure modes?

  Standard models generally have **distinct** failure modes; robust models have **shared** directions

- Can we use stability to find better models of the brain?

  Yes! Robustified models have *interpretable, semantically meaningful* features, whereas standard models are **unstable** and **brittle**

- Can we use stable+predictive models to generate hypotheses about the brain?

  Yes, more soon!

TARGETED PERTURBATIONS REVEAL BRAIN-LIKE LOCAL CODING AXES IN ROBUSTIFIED, BUT NOT STANDARD, ANN-BASED BRAIN MODELS

Nikolas McNeal[1,2] N. Apurva Ratan Murty[1,3]
[1]Center for Excellence in Computational Cognition, Georgia Tech
[2]School of Mathematics, Georgia Tech
[3]School of Psychology, Georgia Tech
{nikolas, ratan}@gatech.edu

arXiv preprint (2025)

Small-scale adversarial perturbations expose differences between predictive encoding models of human fMRI responses

Nikolas McNeal[1,2,*], Mainak Deb[3,*], and N. Apurva Ratan Murty[4,5]
[1]Machine Learning, Georgia Tech
[2]School of Mathematics, Georgia Tech
[3]Independent contributor
[4]CoE in Computational Cognition, Georgia Tech
[5]Cognition and Brain Science, Georgia Tech

NeurIPS workshop (UniReps) 2024