EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Inductive bias transfer between brains and machines

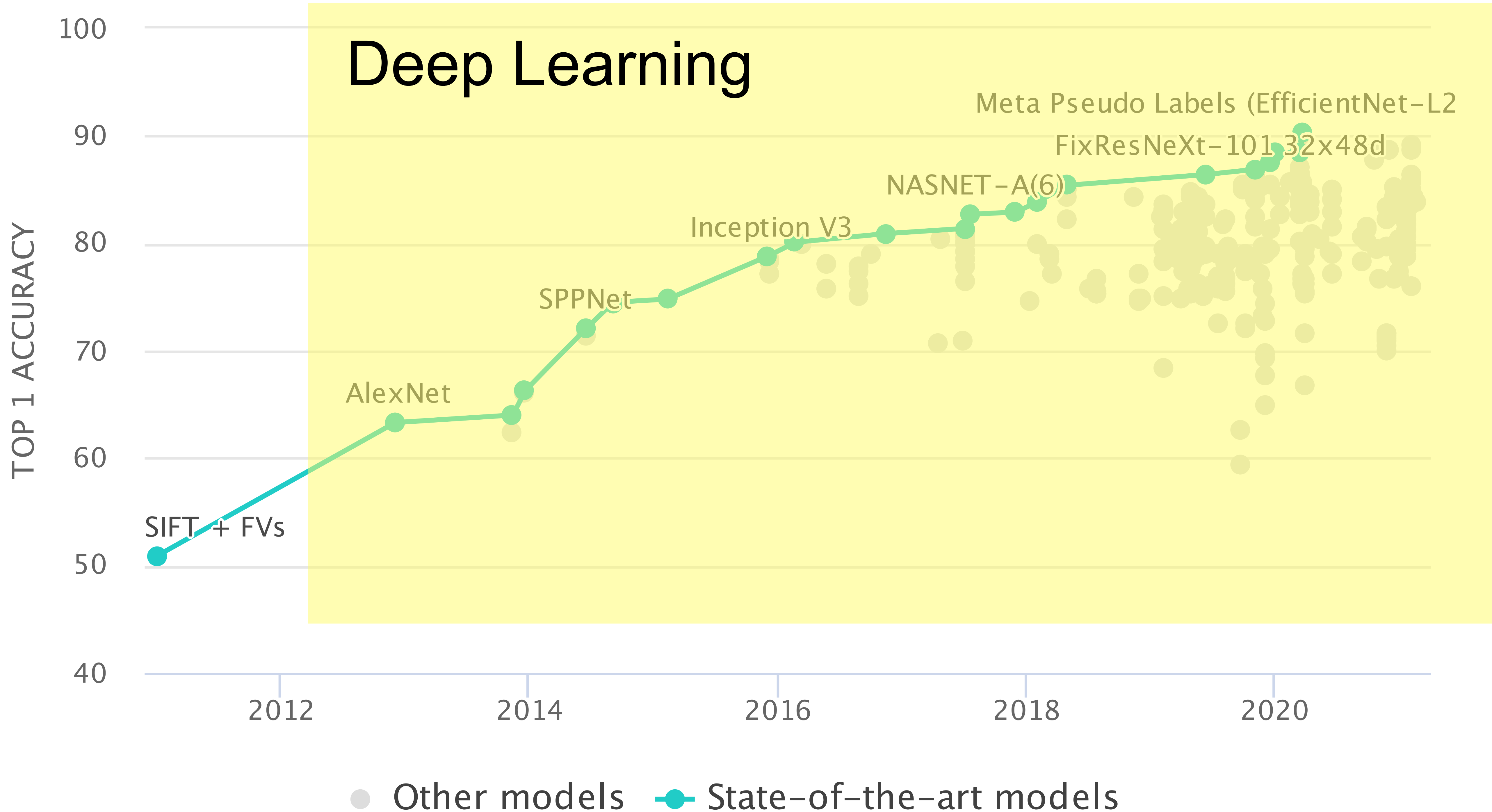
Fabian Sinz

Neural Intelligence Group, Uni Tübingen
soon Uni Göttingen

@sinzlab

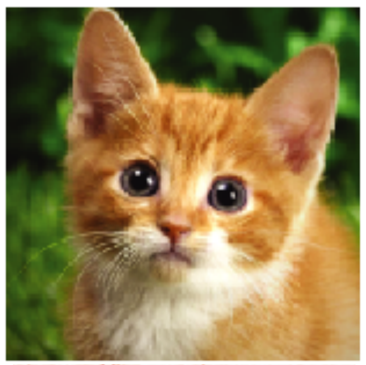
NICE 2021

Success of deep learning



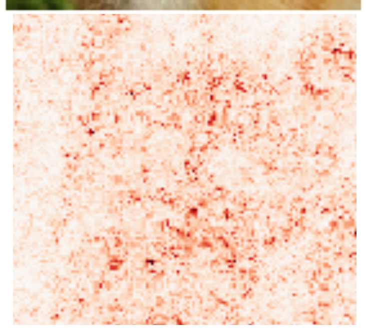
https://paperswithcode.com/sota/image-classification-on-imagenet

Current state-of-the-art is brittle



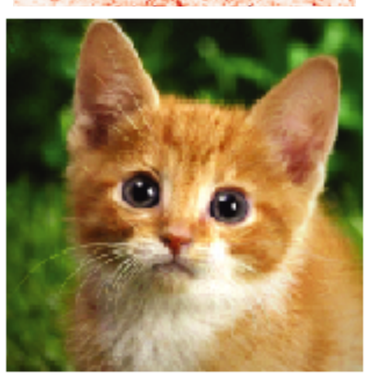
“cat”

+



adversarial perturbation

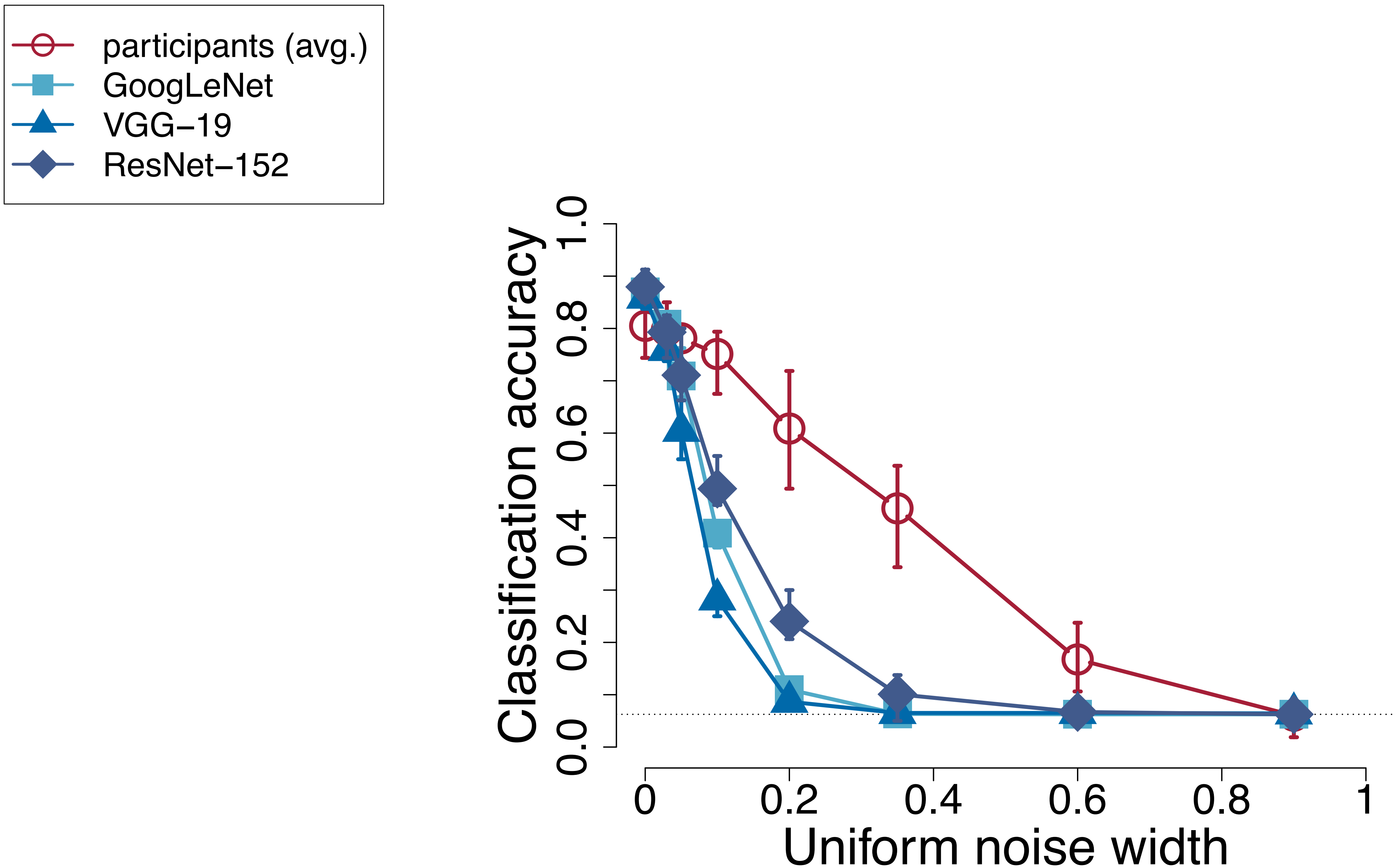
=



“moped”

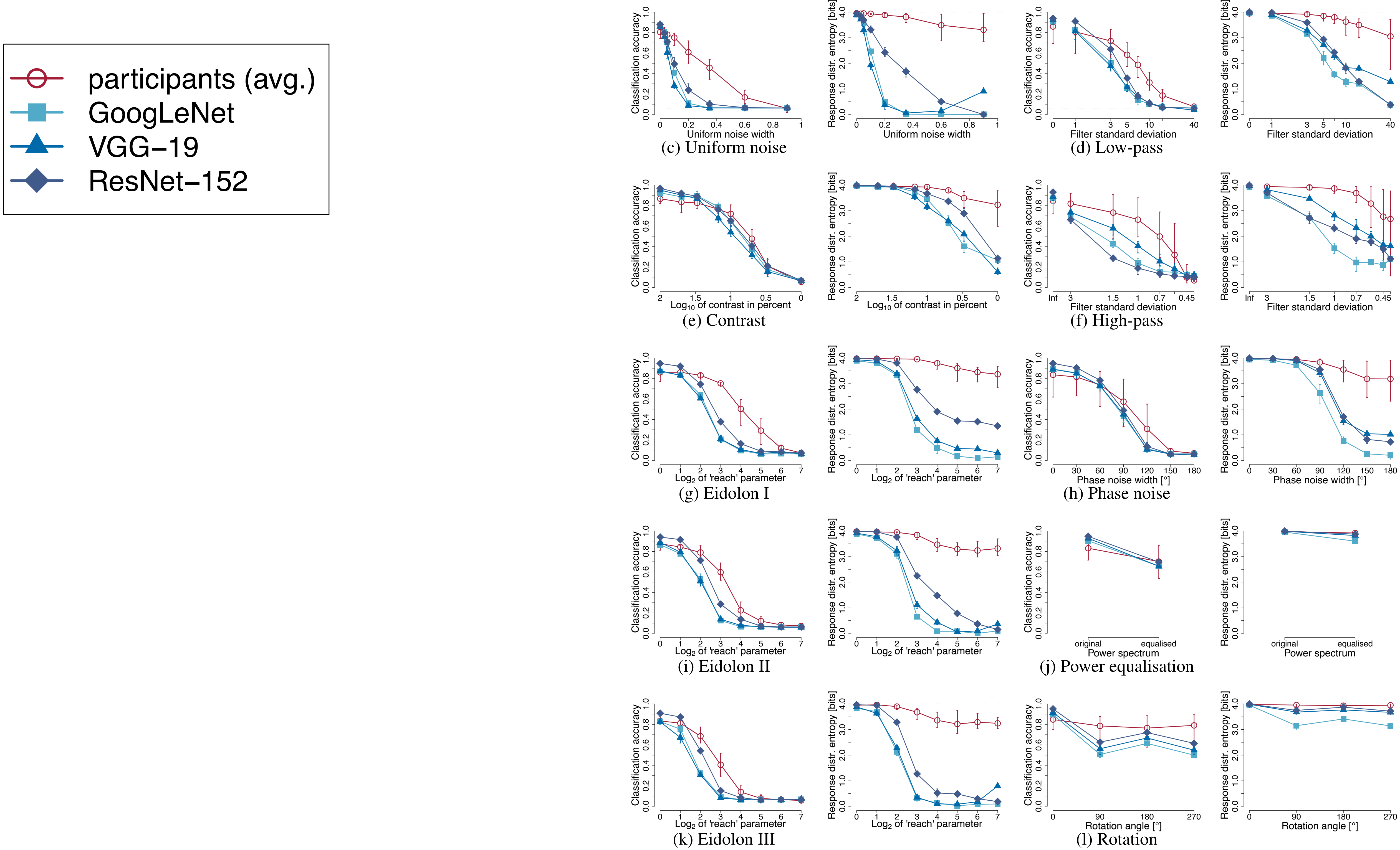
Geirhos et al. 2018. “Generalisation in Humans and Deep Neural Networks.”
Szegedy et al. 2013. “Intriguing Properties of Neural Networks.”
Sinz et al 2019. “Engineering a Less Artificial Intelligence.” *Neuron*.

Human visual system is robust



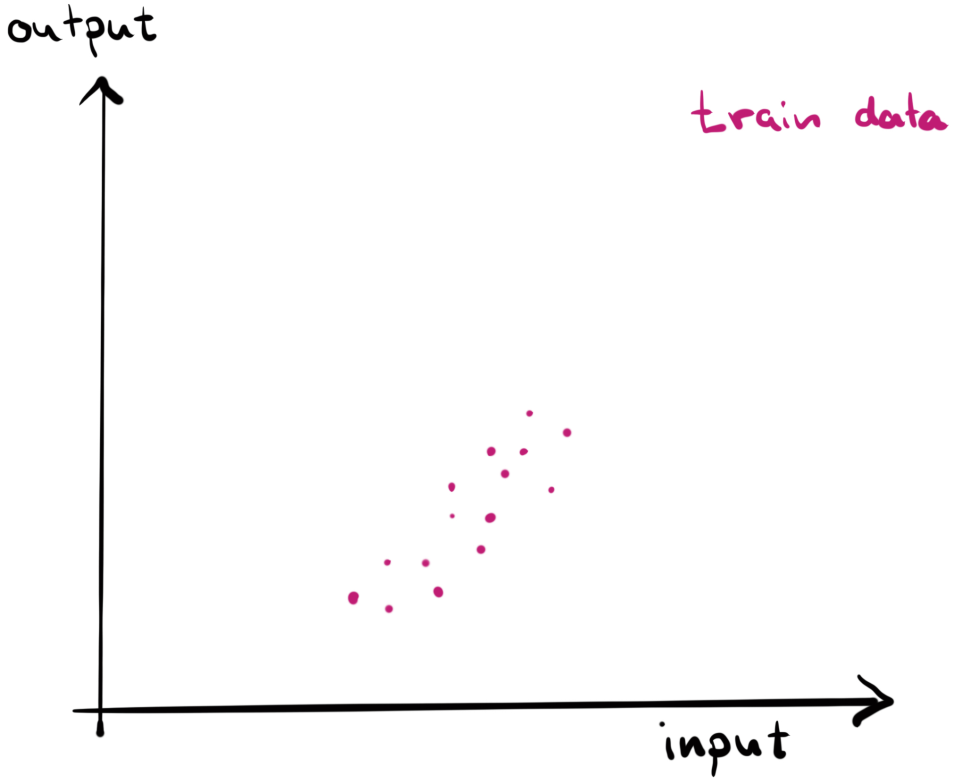
Geirhos et al. 2018. “Generalisation in Humans and Deep Neural Networks.”

Human visual system is robust

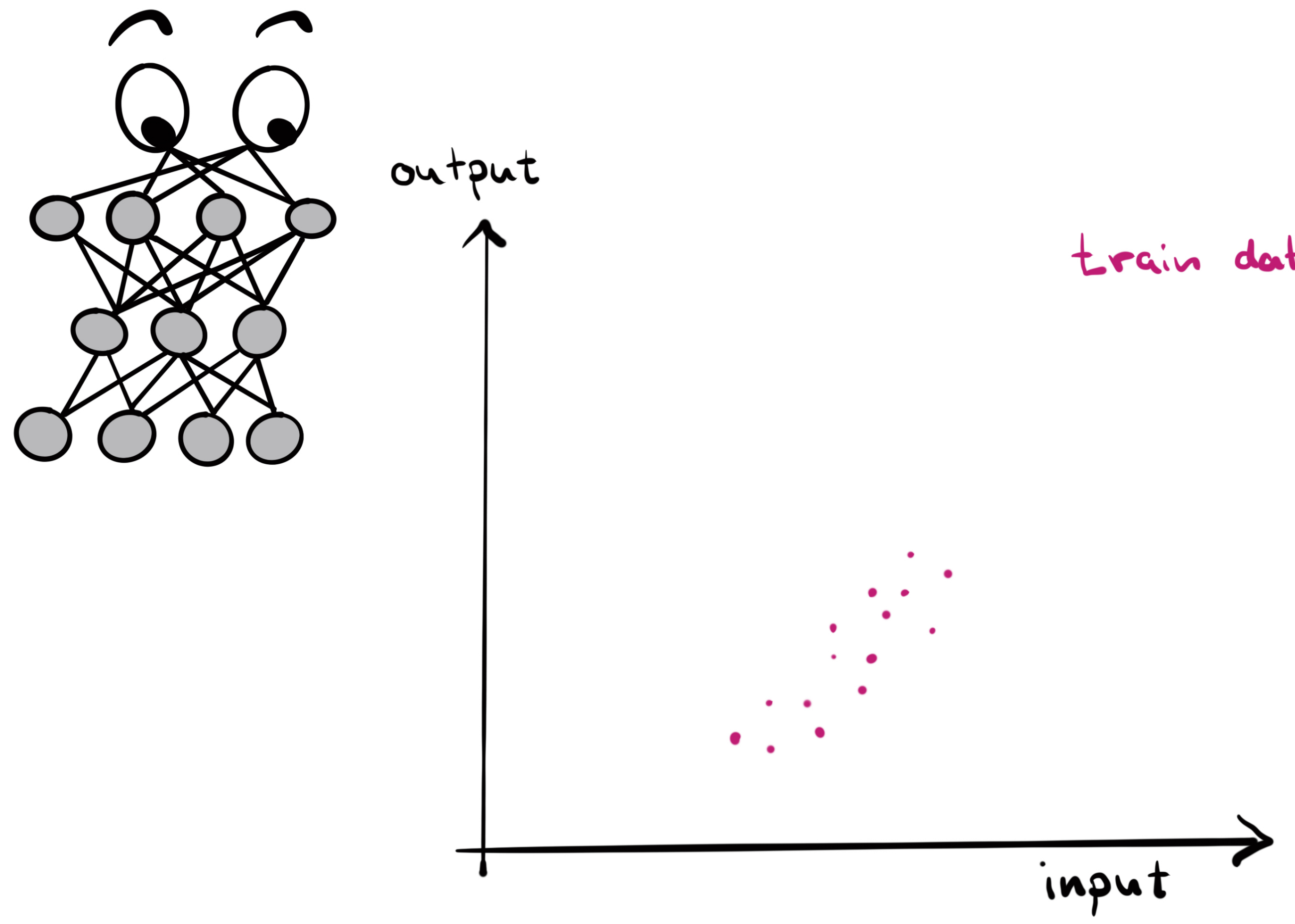


Geirhos et al. 2018. “Generalisation in Humans and Deep Neural Networks.”

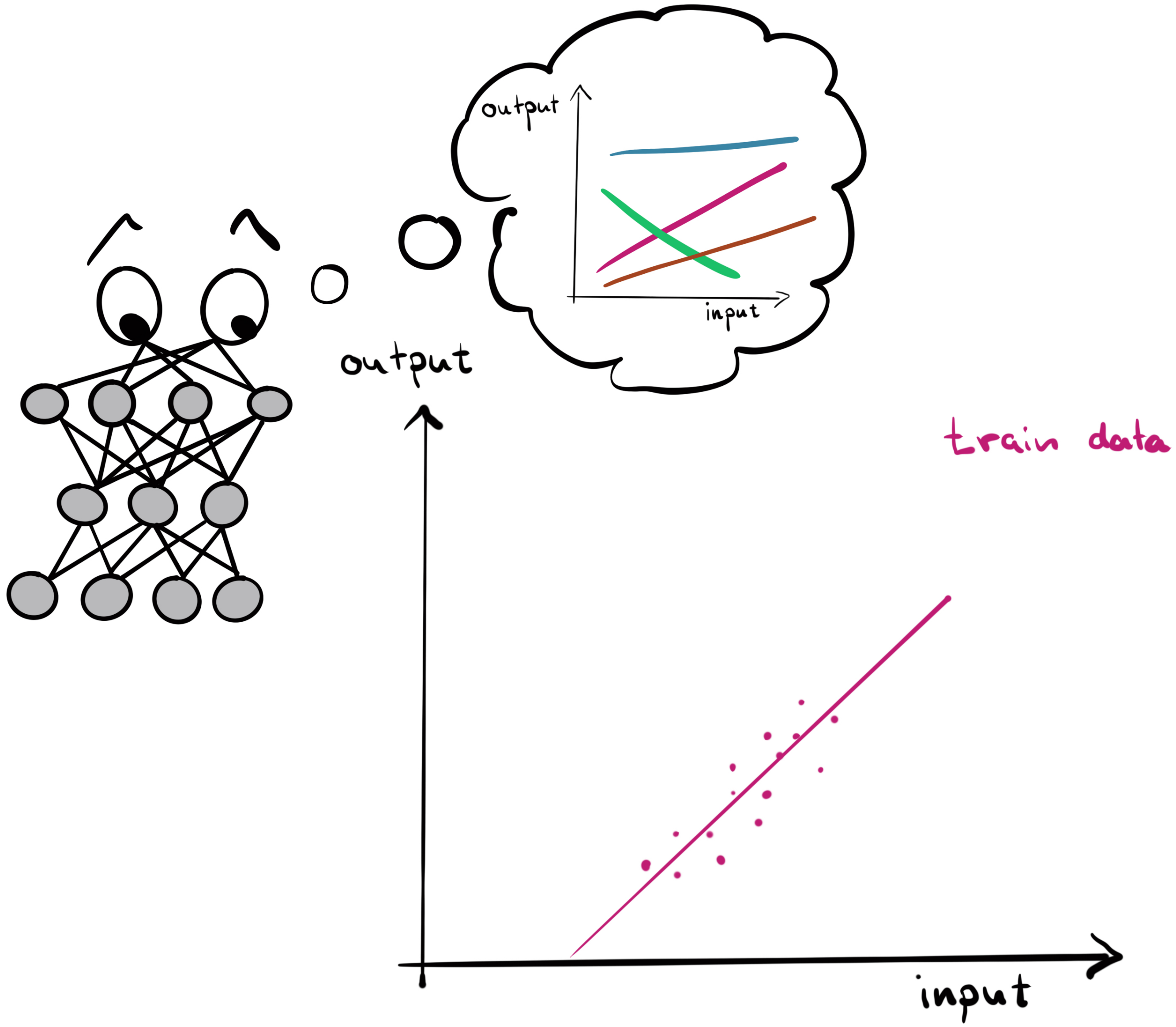
Inductive Bias - Implicit Assumptions



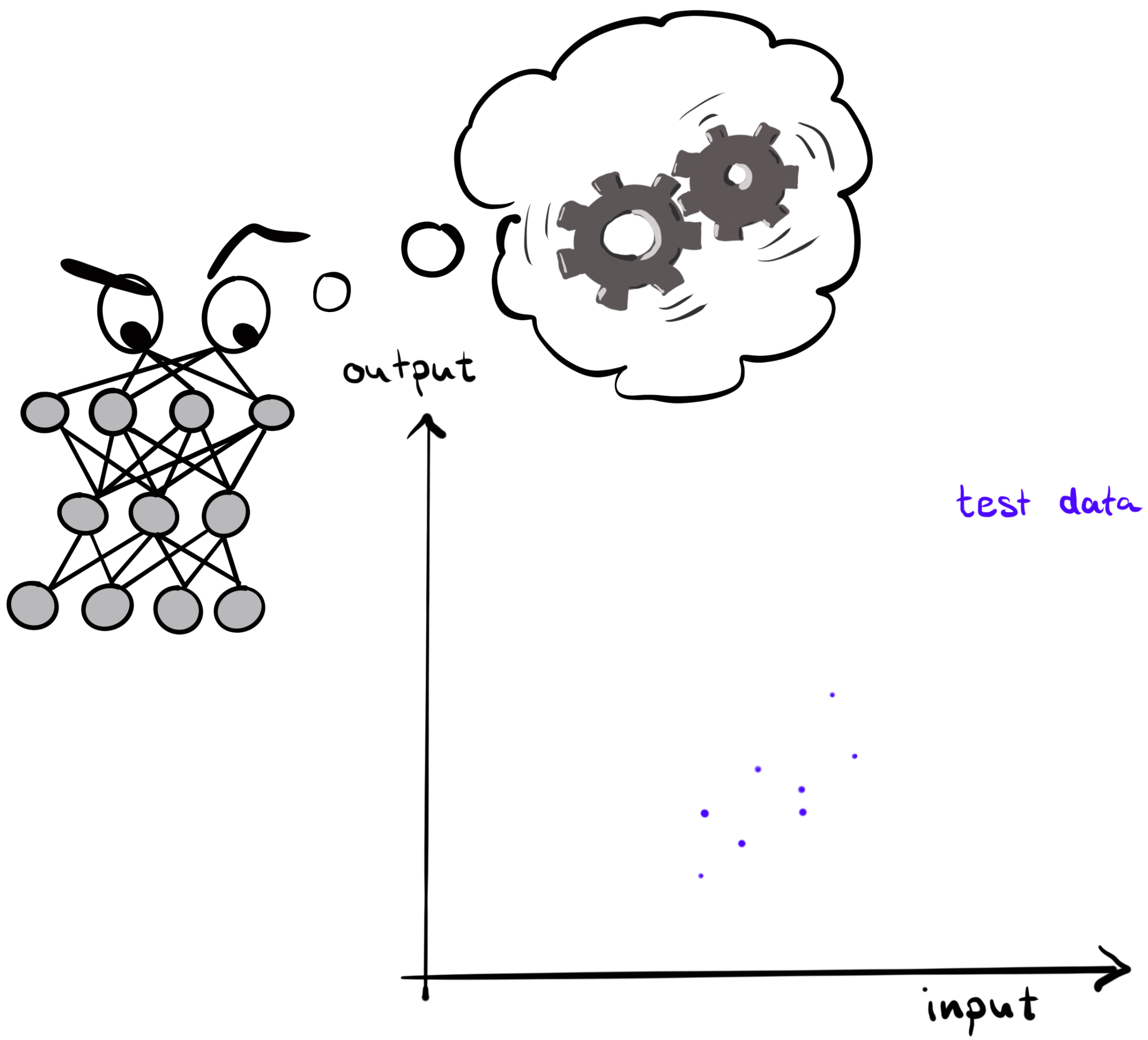
Inductive Bias - Implicit Assumptions



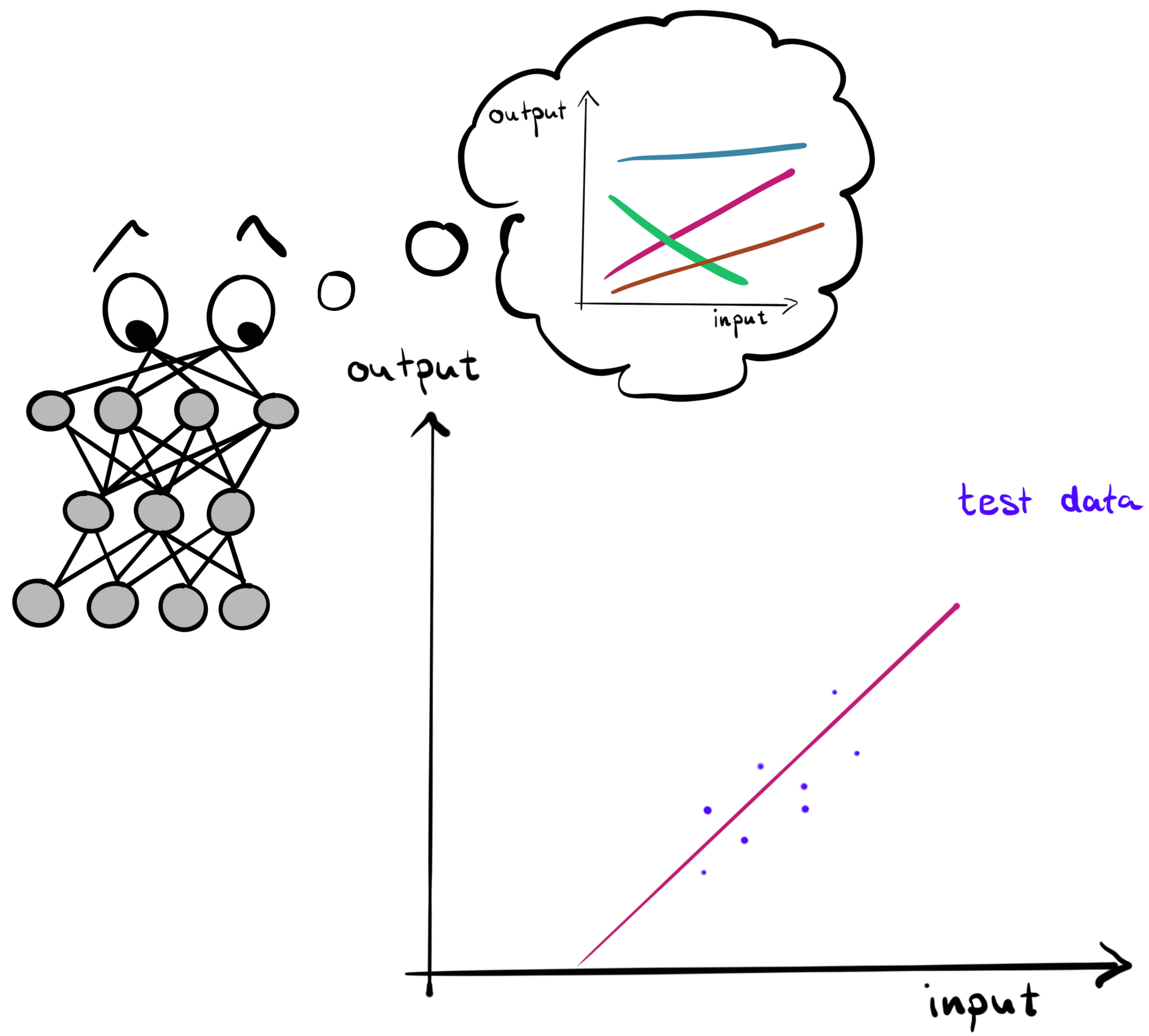
Inductive Bias - Implicit Assumptions



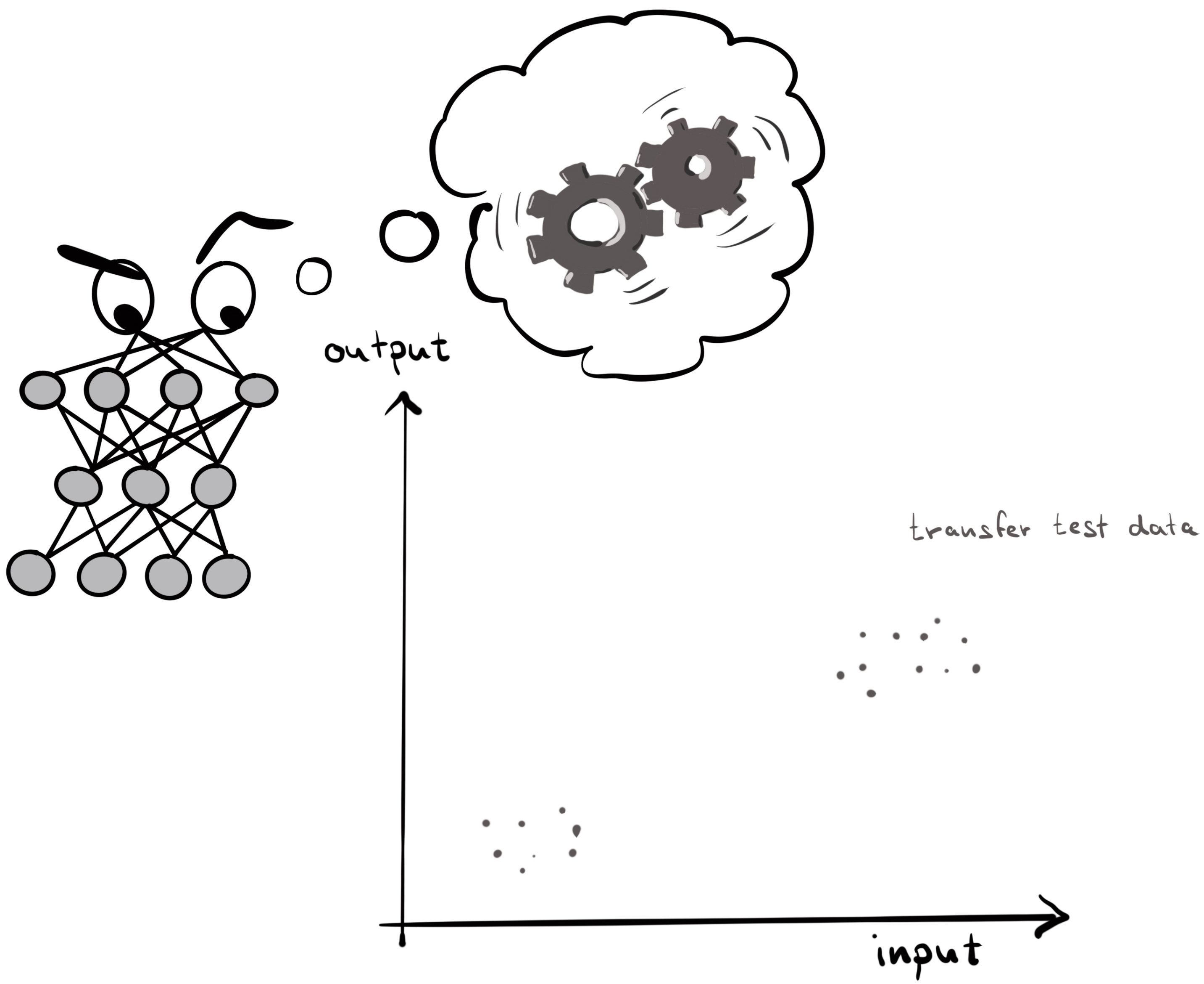
Inductive Bias - Implicit Assumptions



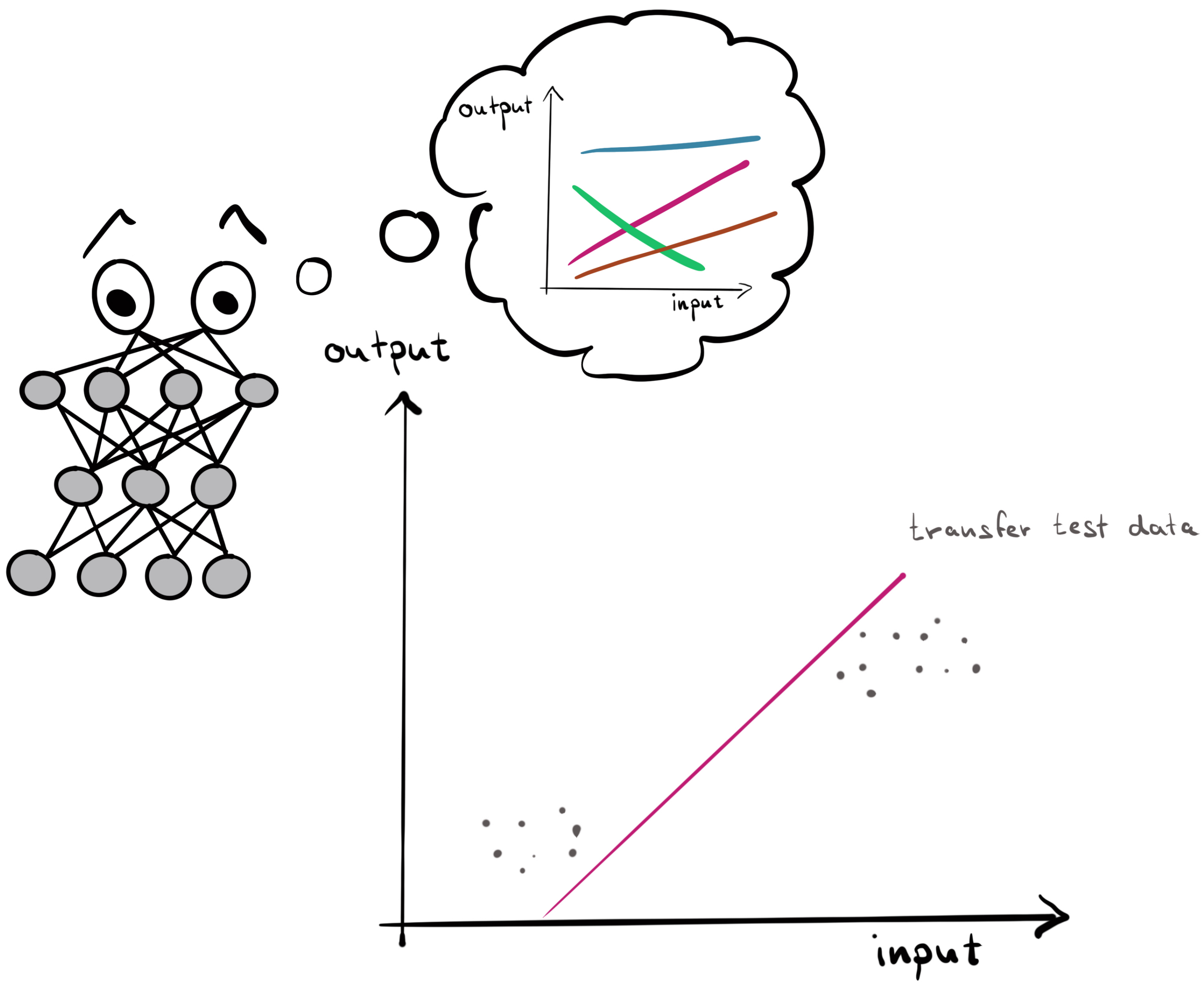
Inductive Bias - Implicit Assumptions



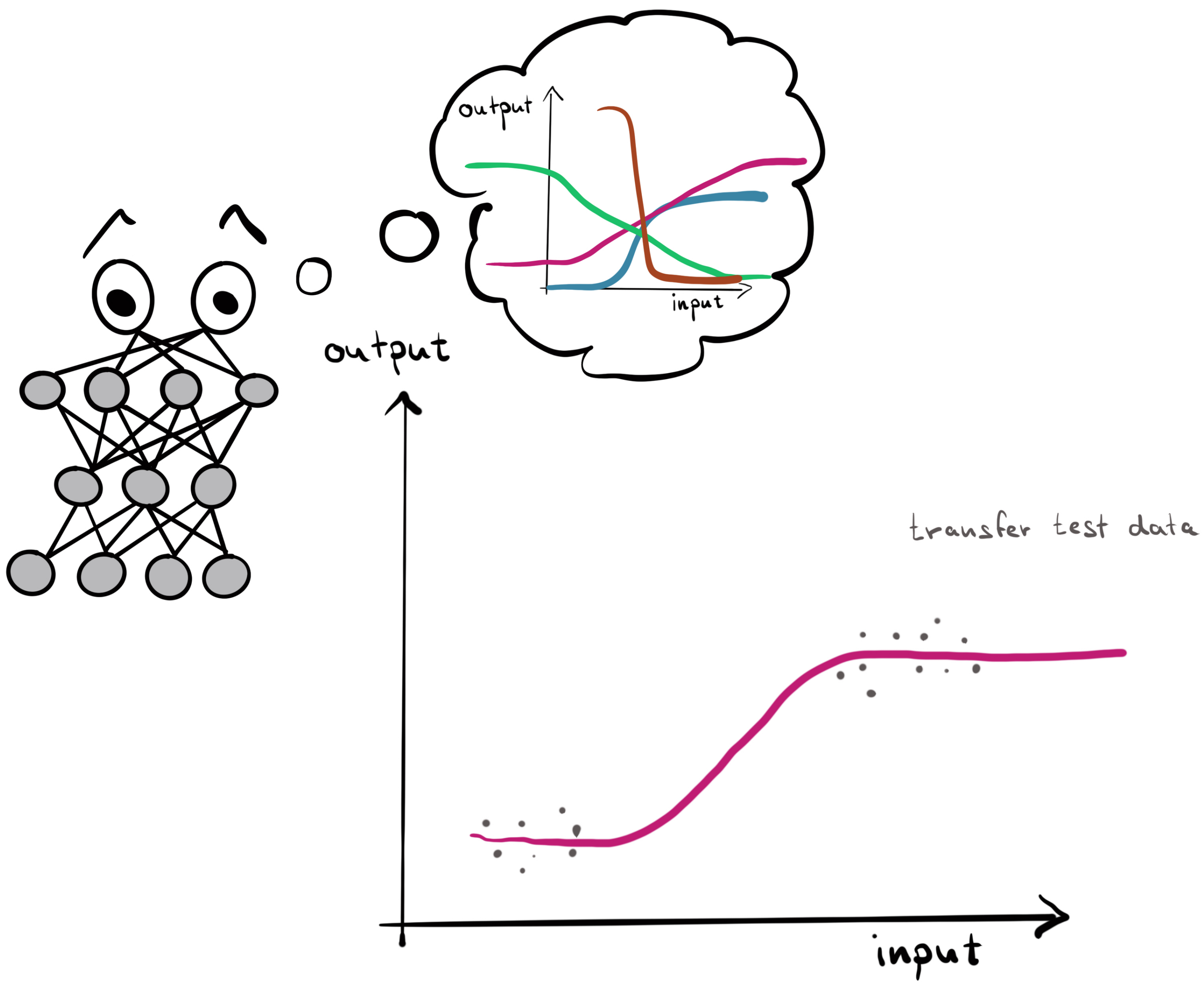
Inductive Bias - Implicit Assumptions



Inductive Bias - Implicit Assumptions

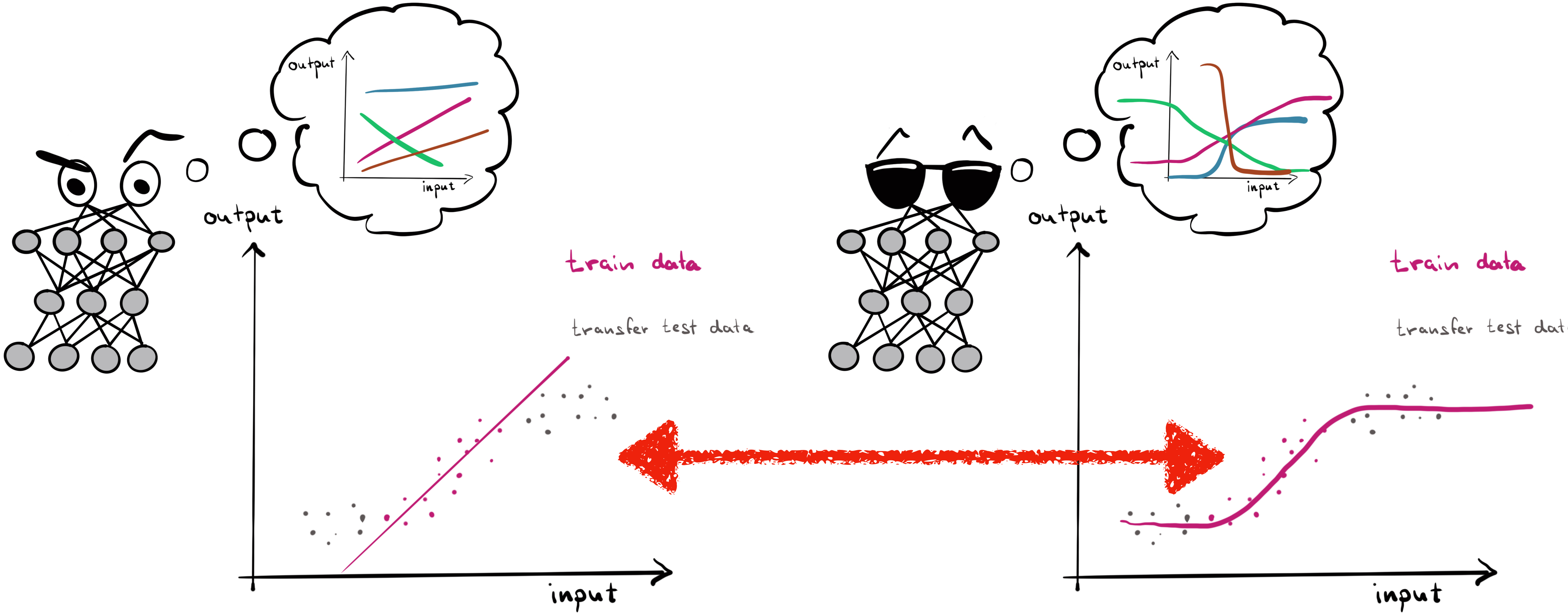


Inductive Bias - Implicit Assumptions



Inductive Bias - Implicit Assumptions

Differences in extrapolation between two algorithms given the **same** training data.



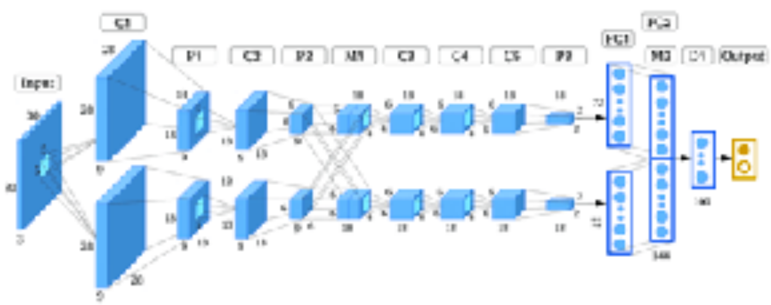
Inductive Bias is Essential for Generalization

No inductive bias

No free lunch

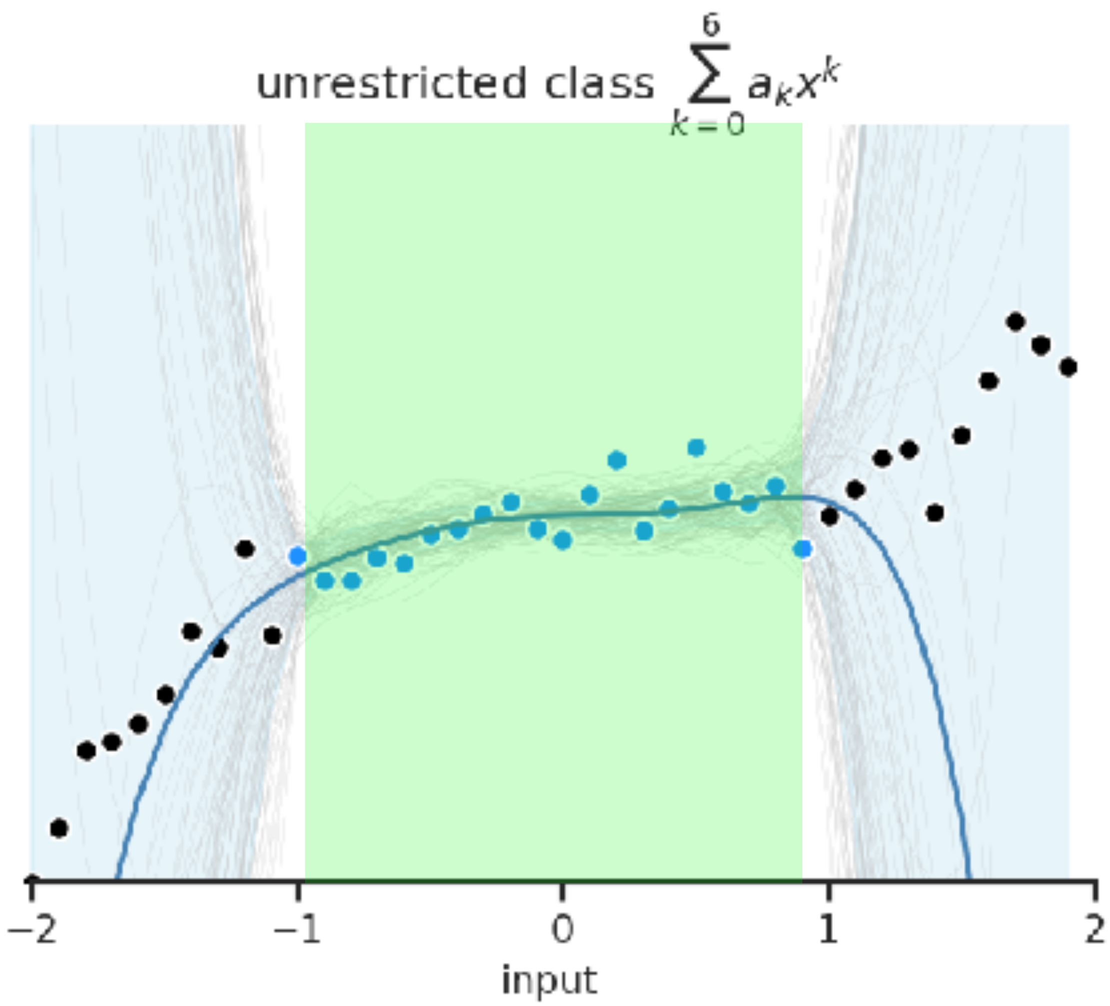
No generalization

Inductive Bias



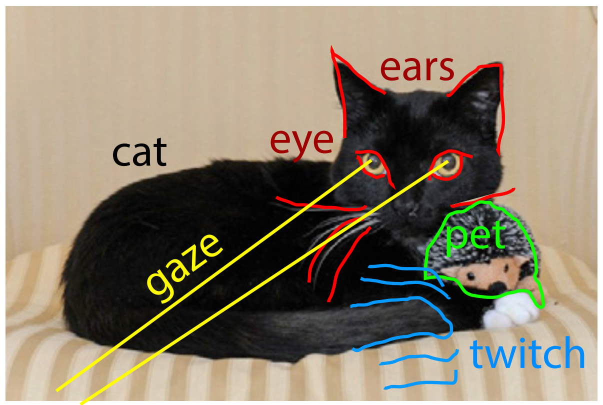
less ⟷ **more**

bias



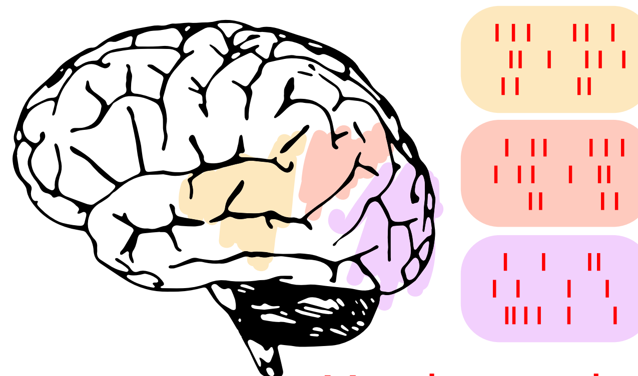
Levels of Inductive Bias Transfer

Computational level



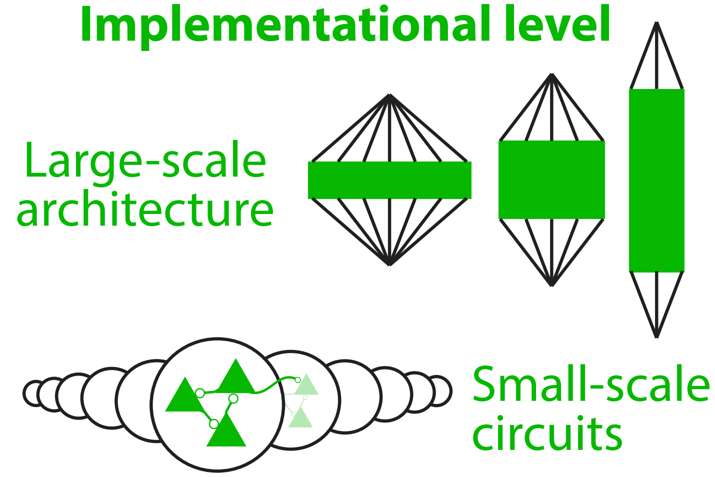
Multi-task training:
act on causal variables

Representational level



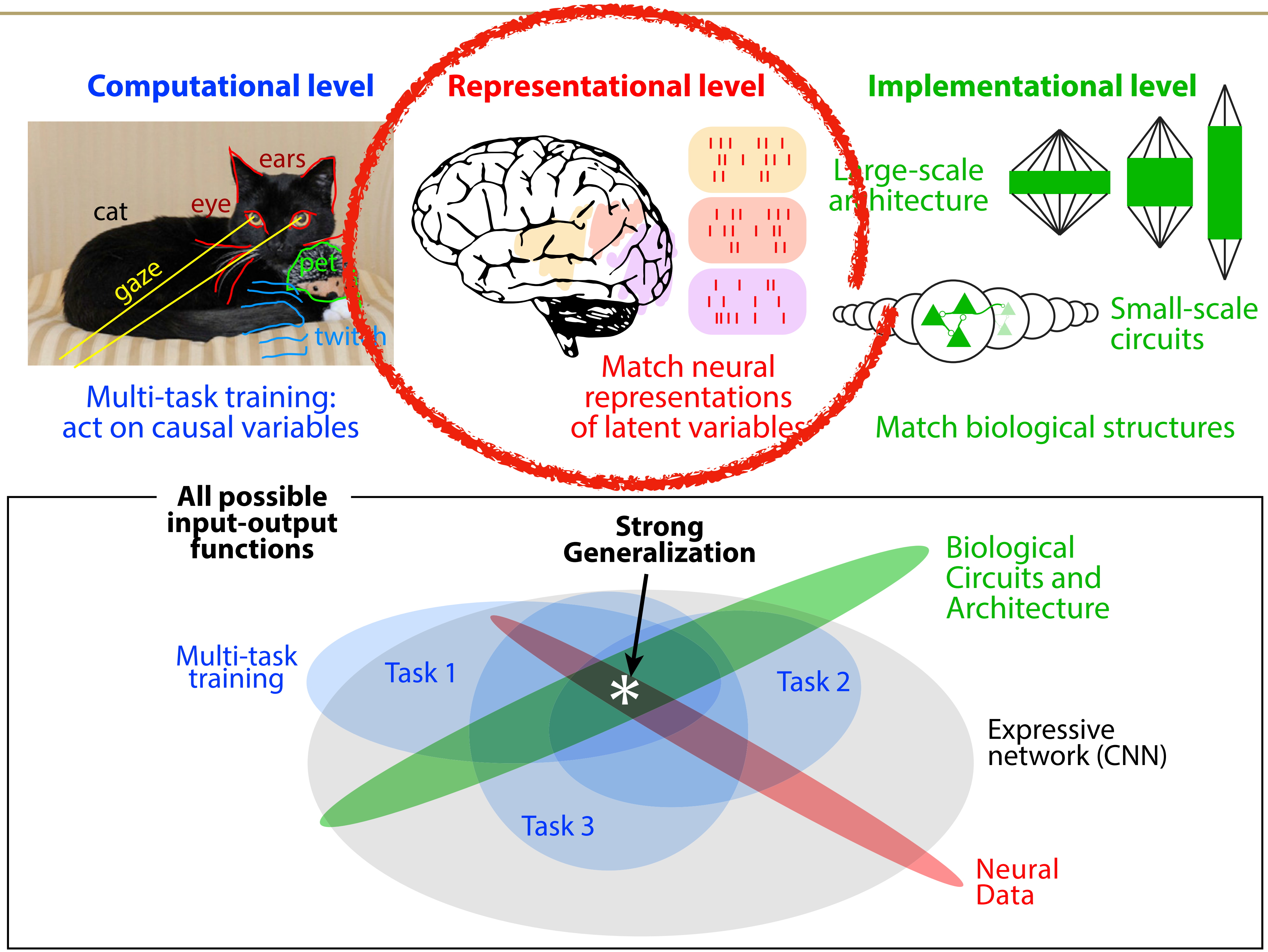
Match neural
representations
of latent variables

Implementational level



Match biological structures

How can we transfer good inductive biases?



Sinz et al 2019. “Engineering a Less Artificial Intelligence.” *Neuron*.

Neural co-training on monkey V1

Shahd Safarani

Arne Nix

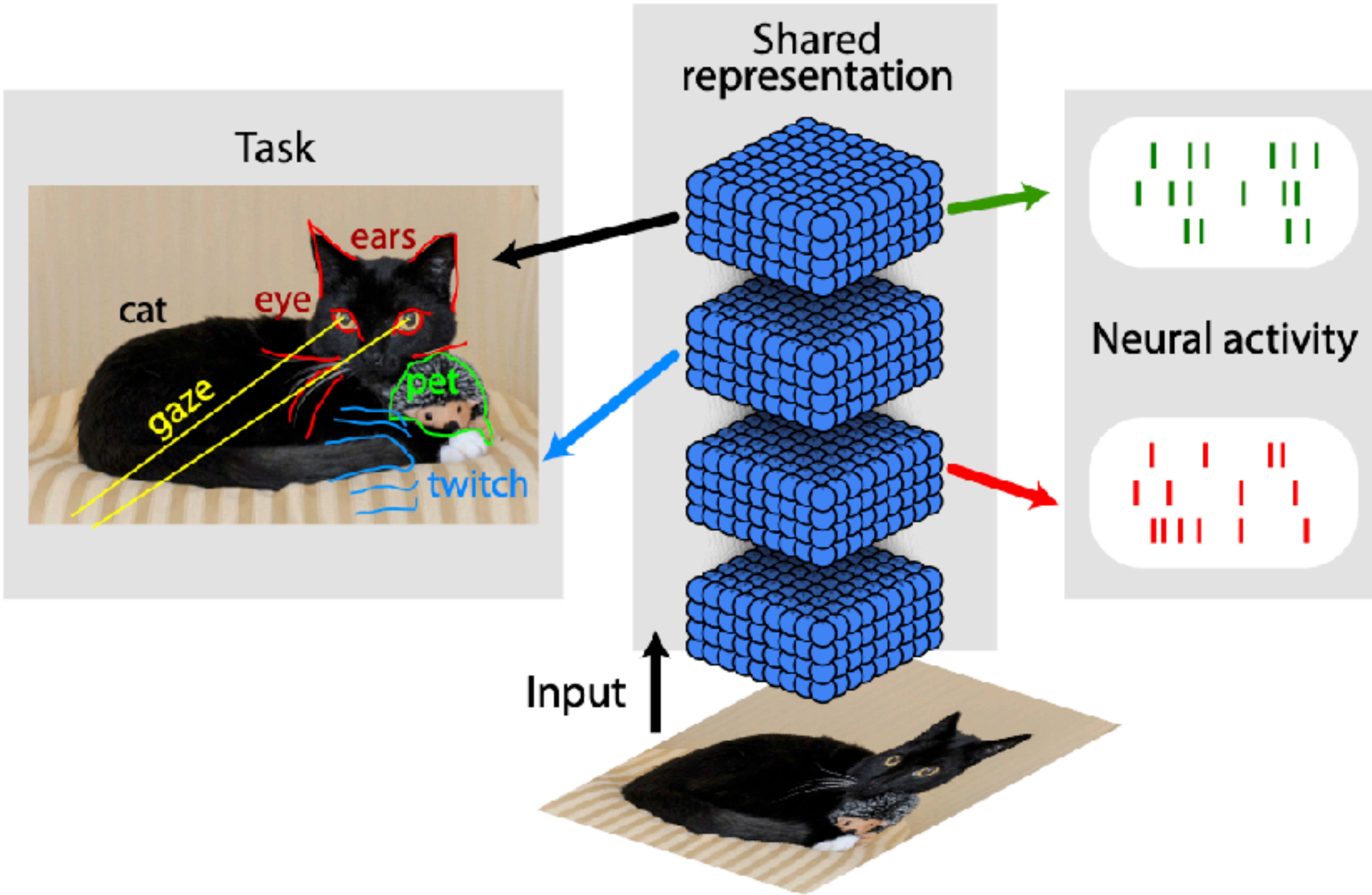
Konstantin Willeke

In collaboration with:

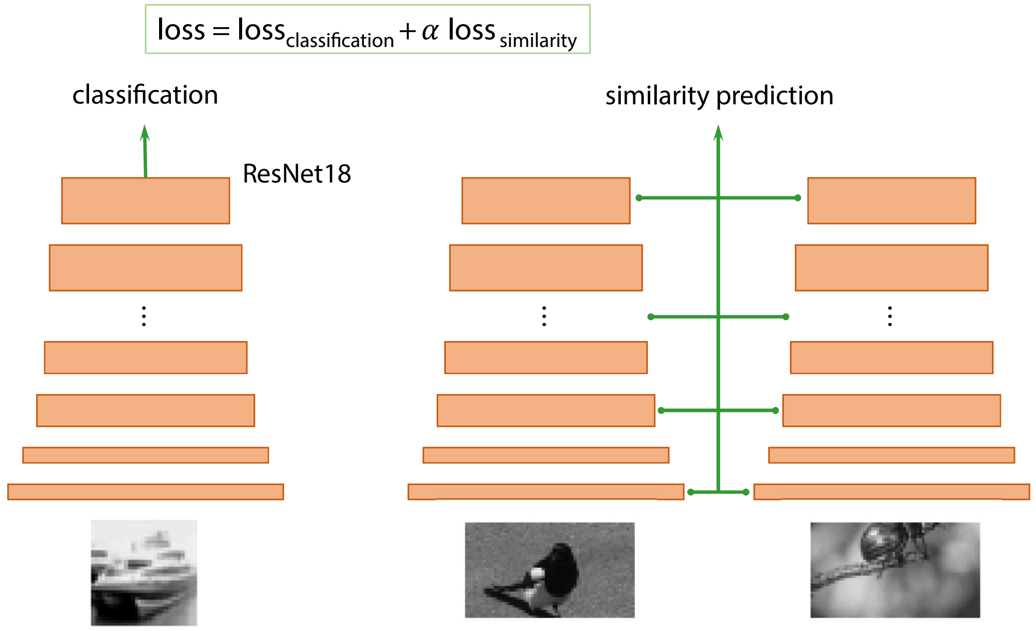
Baylor College of Medicine

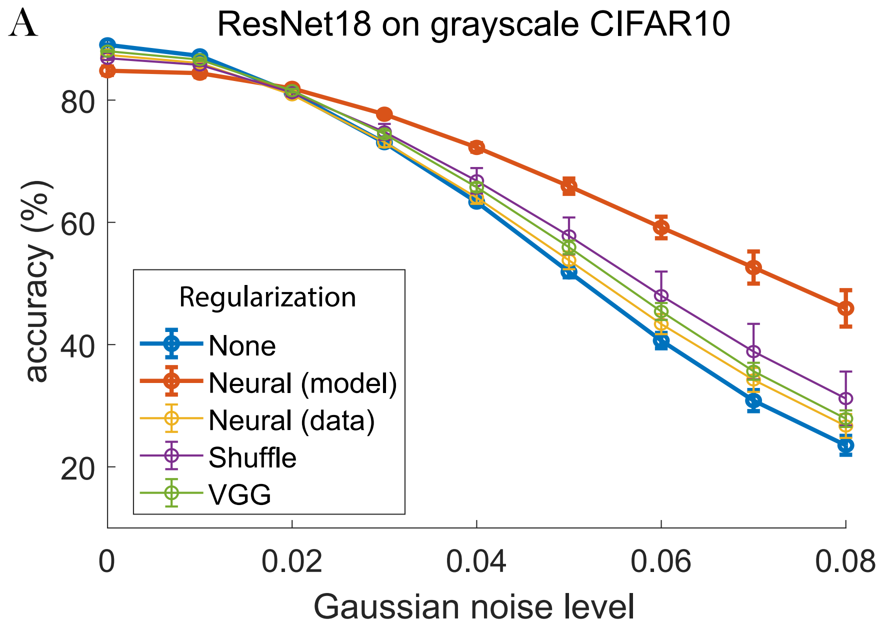
Andreas Tolias, PhD

Neural co-training hypothesis



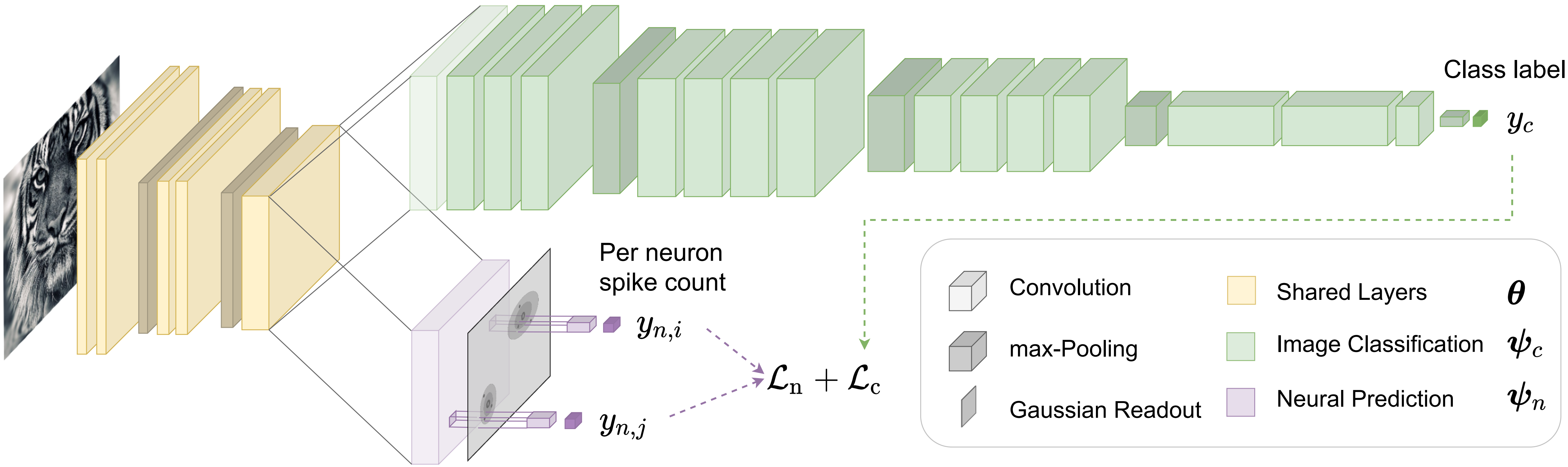
Do we expect it to work?



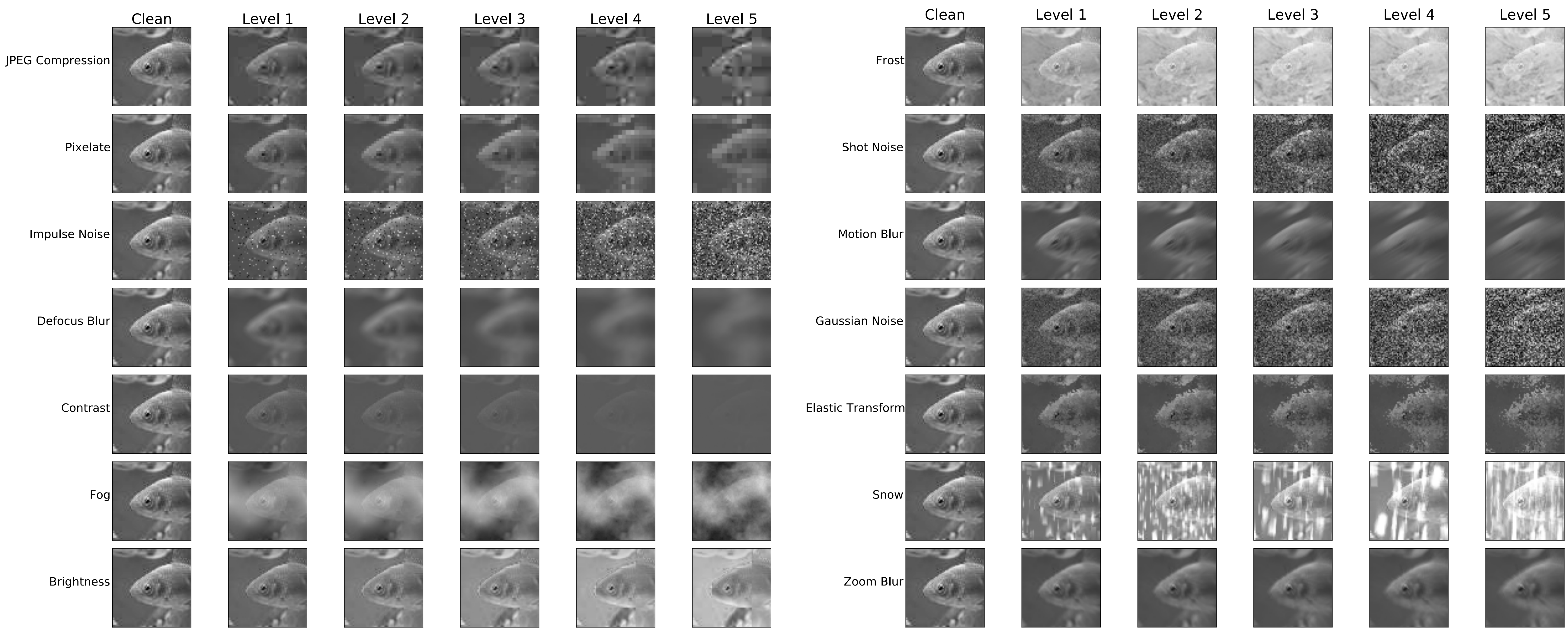


Li et al. 2019. “Learning From Brains How to Regularize Machines.” NeurIPS

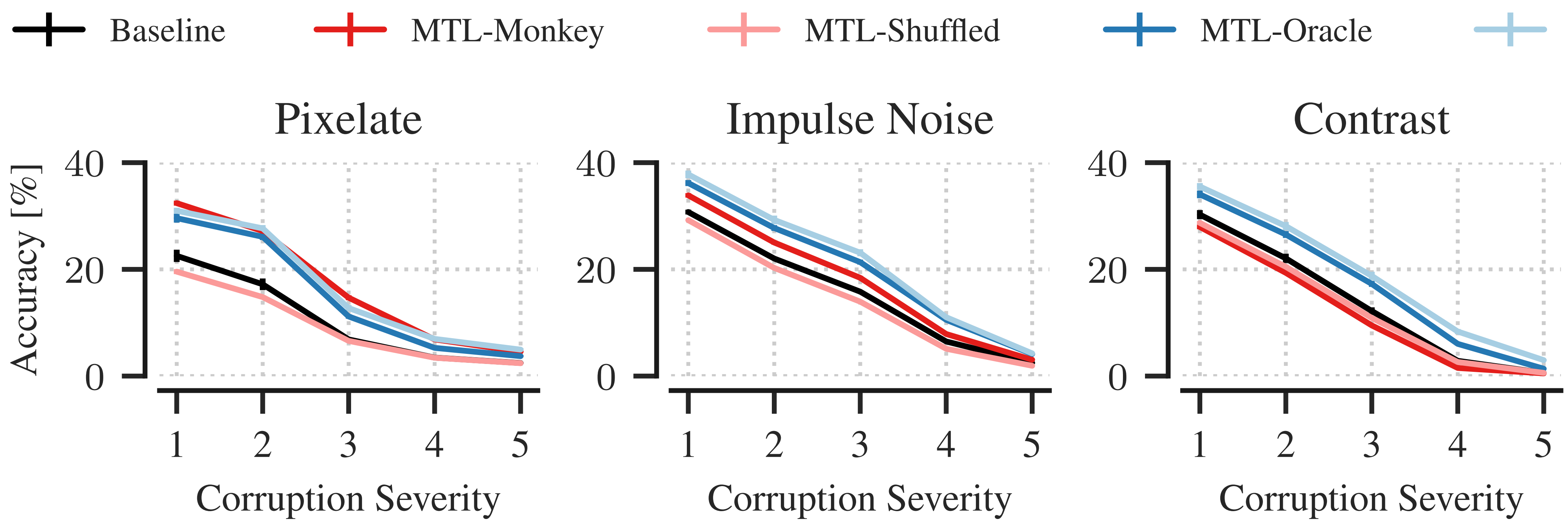
Multi-Task-Learning with monkey V1



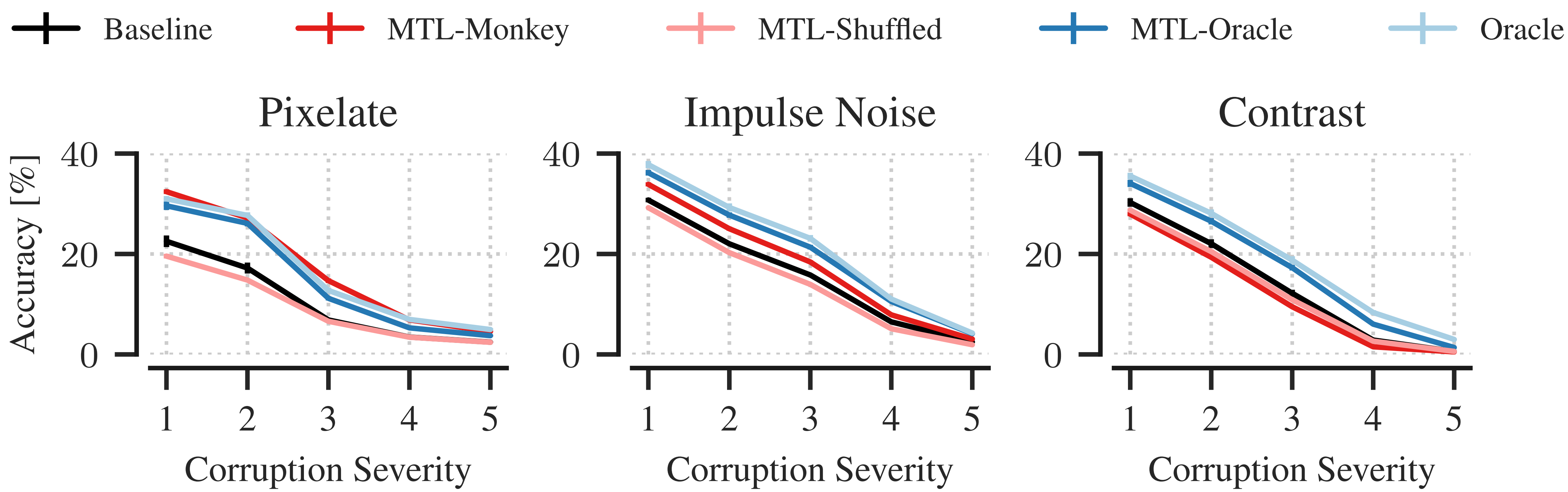
How do we test inductive bias?

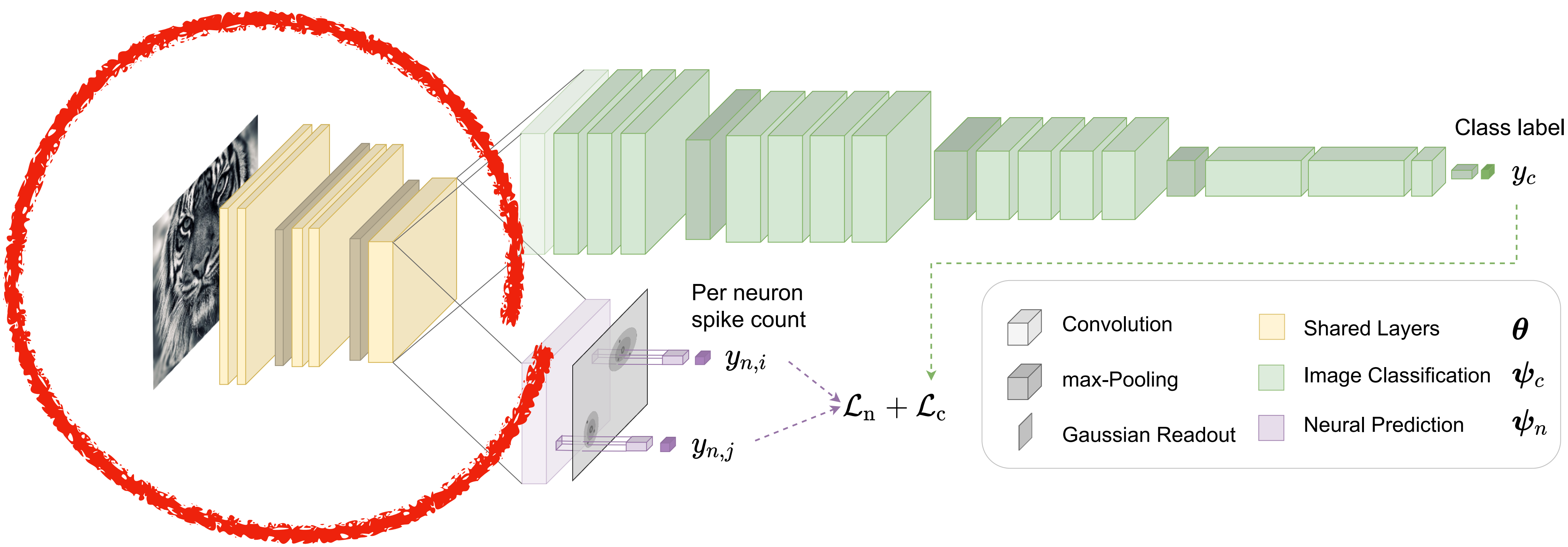


V1 co-training yields benefits

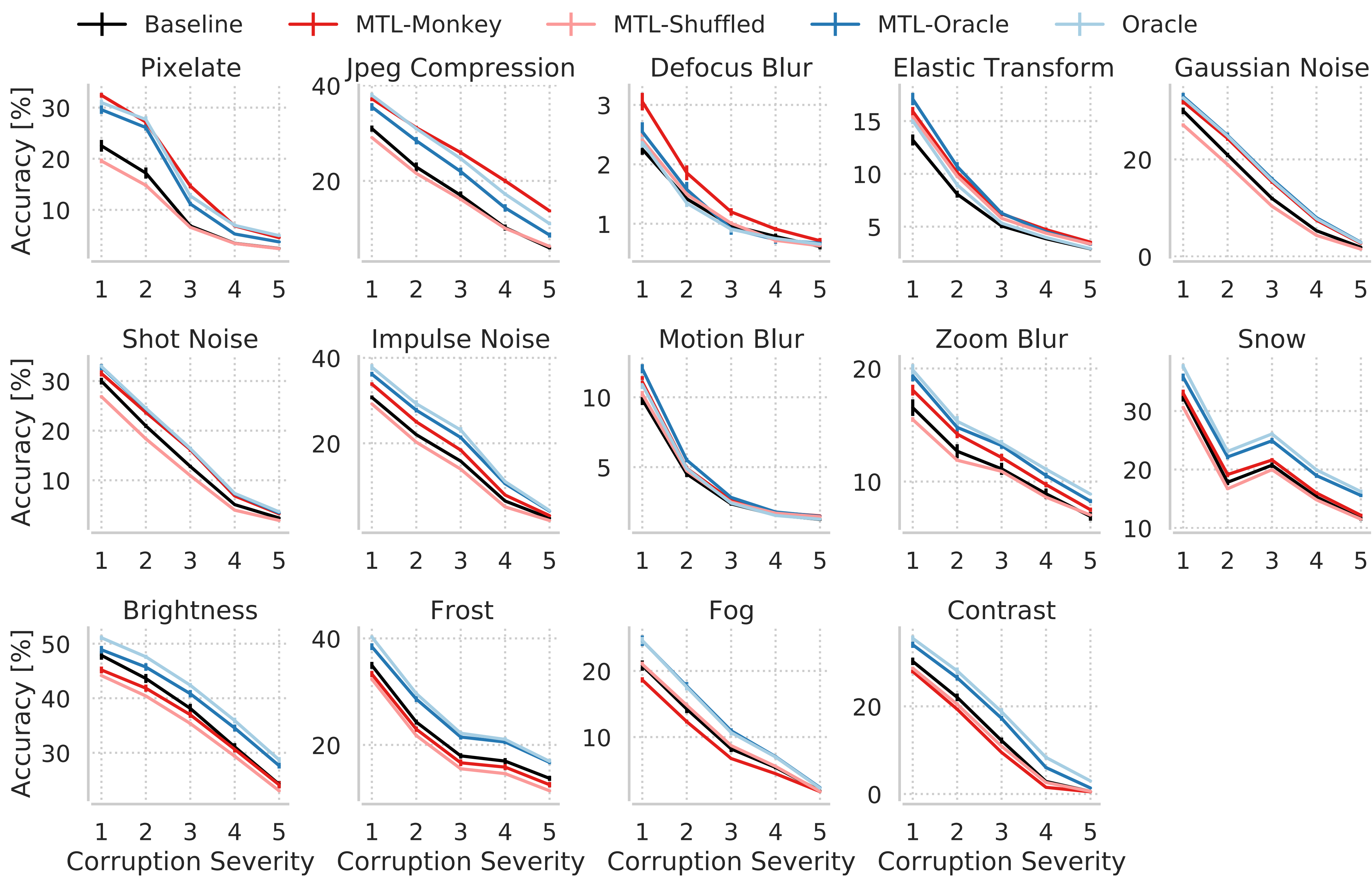


V1 co-training yields benefits

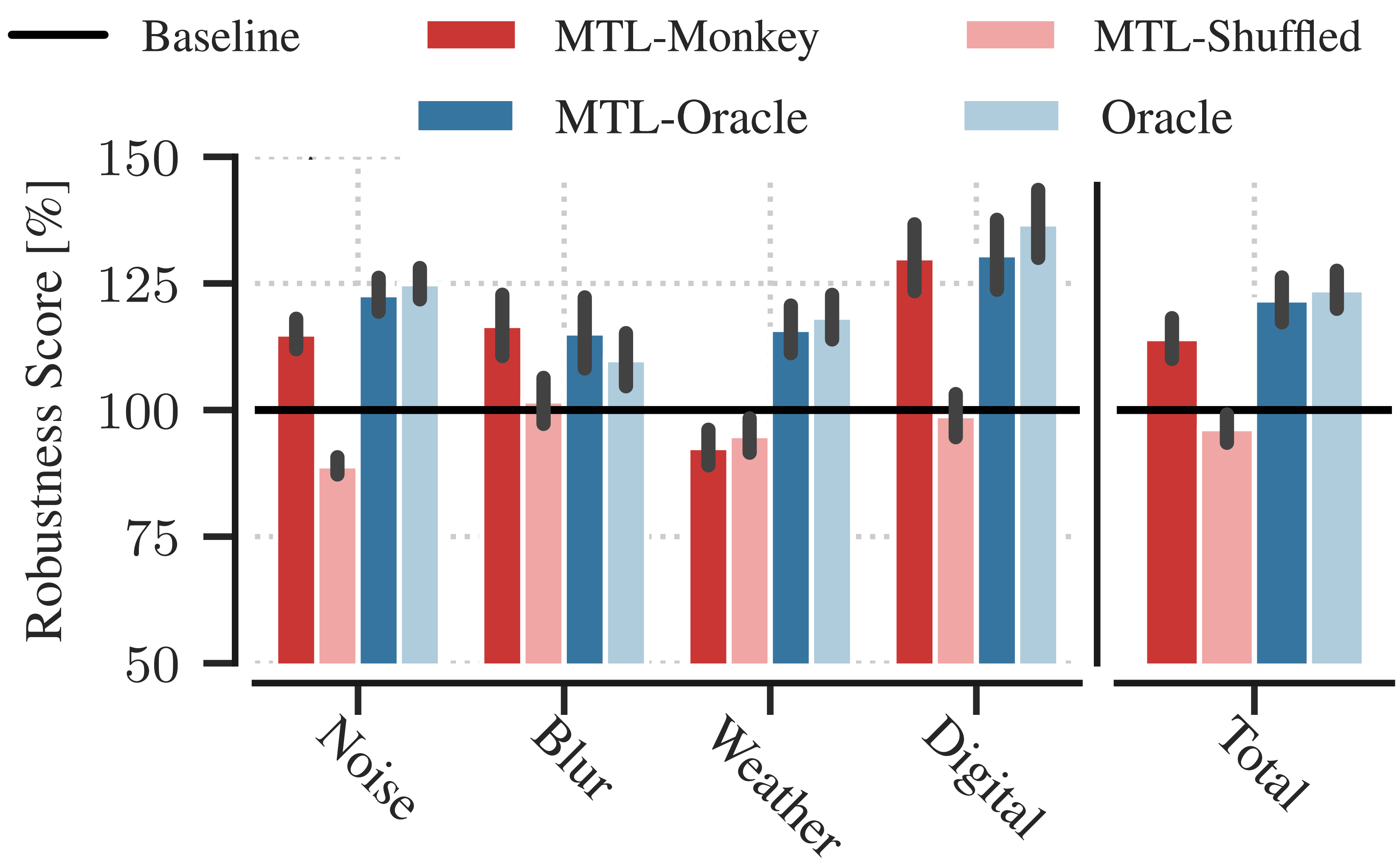




V1 co-training yields benefits



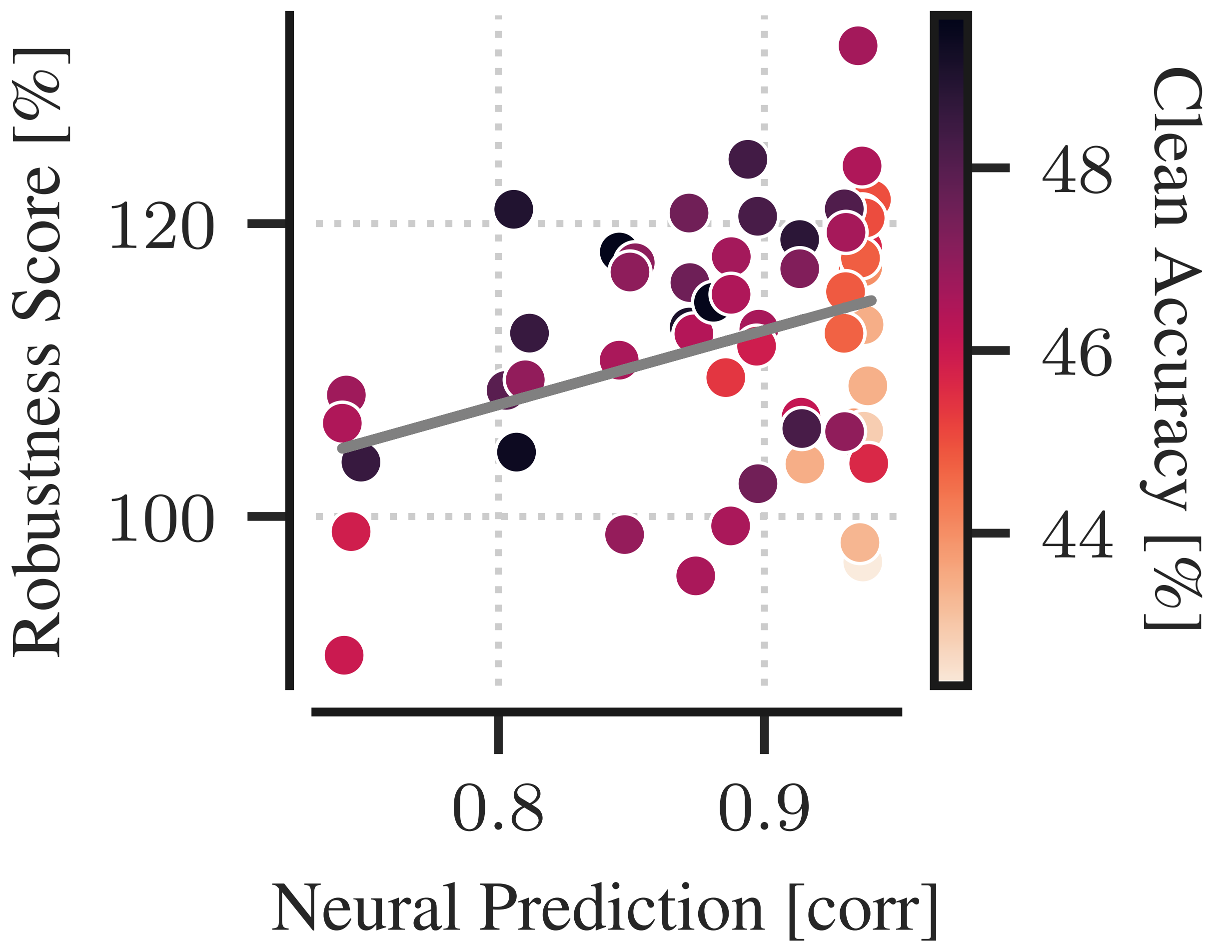
No all distortions show the same effect



$$\frac{1}{C}\sum_{c=1}^{C}\frac{A_c^{\textbf{robust}}}{A_c^{\textbf{baseline}}}$$

$$A_c = \frac{1}{5}\sum_{l,s=1}^{5} A_{l,c,s}$$

Robustness correlates with “brain-likeness”



Summary

- Mammalian visual systems have a better inductive bias than deep networks
- Multi task learning can be one avenue to improve inductive biases of models
- Co-training on monkey V1 yields improves robustness classification models
- Brain-likeness correlates with robustness

Funding

CyberValley

Carl Zeiss Stiftung

aws

Thanks for listening! Questions?

We are looking for PhD students!

Check out: https://sinzlab.org/openpositions.html
or scan code