

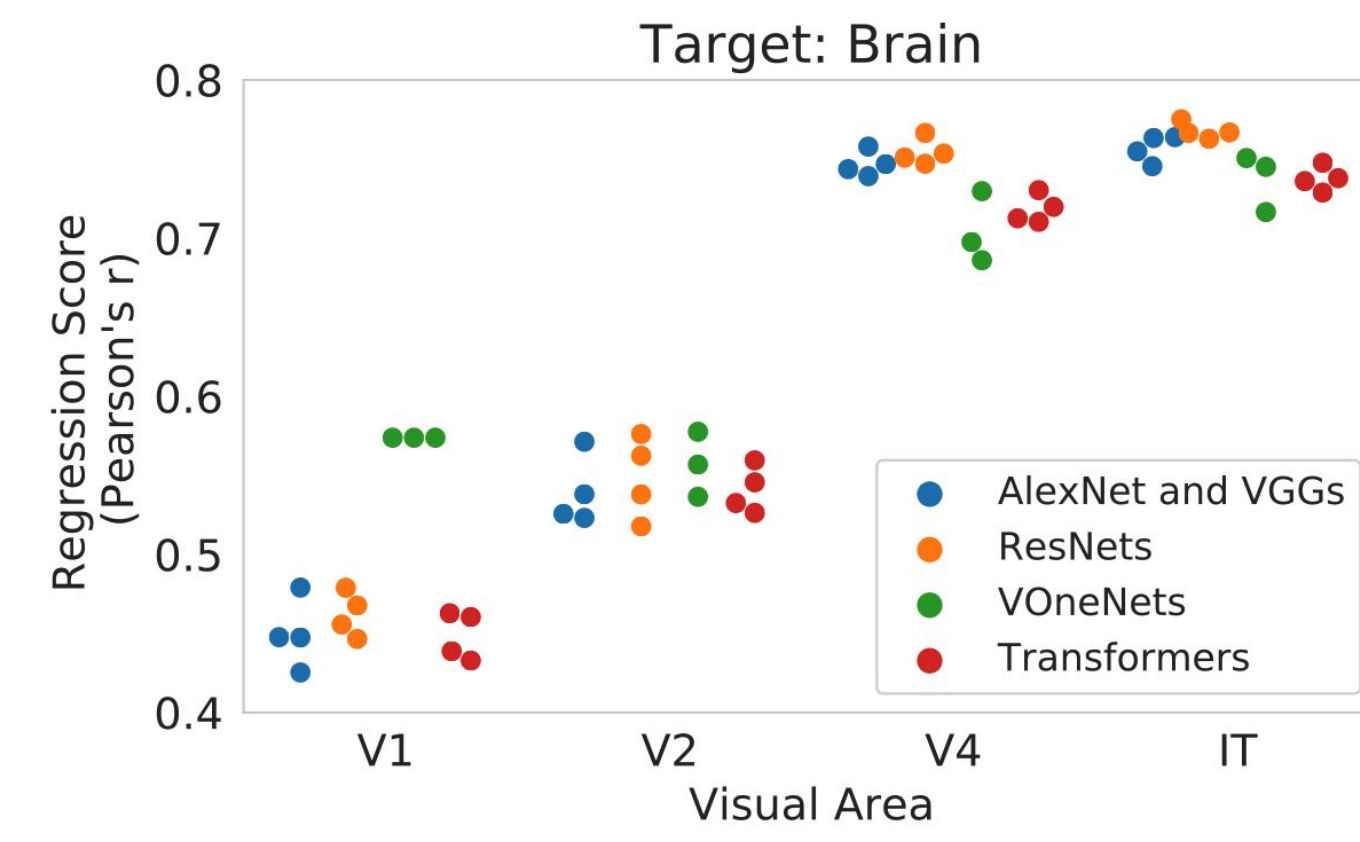
Making things go wrong...for educational purposes

Brian Cheung, Erin Grant, Helen Yang, Yena Han, Tomaso Poggio, Boris Katz
cheungb@mit.edu
<https://briancheung.github.io/>

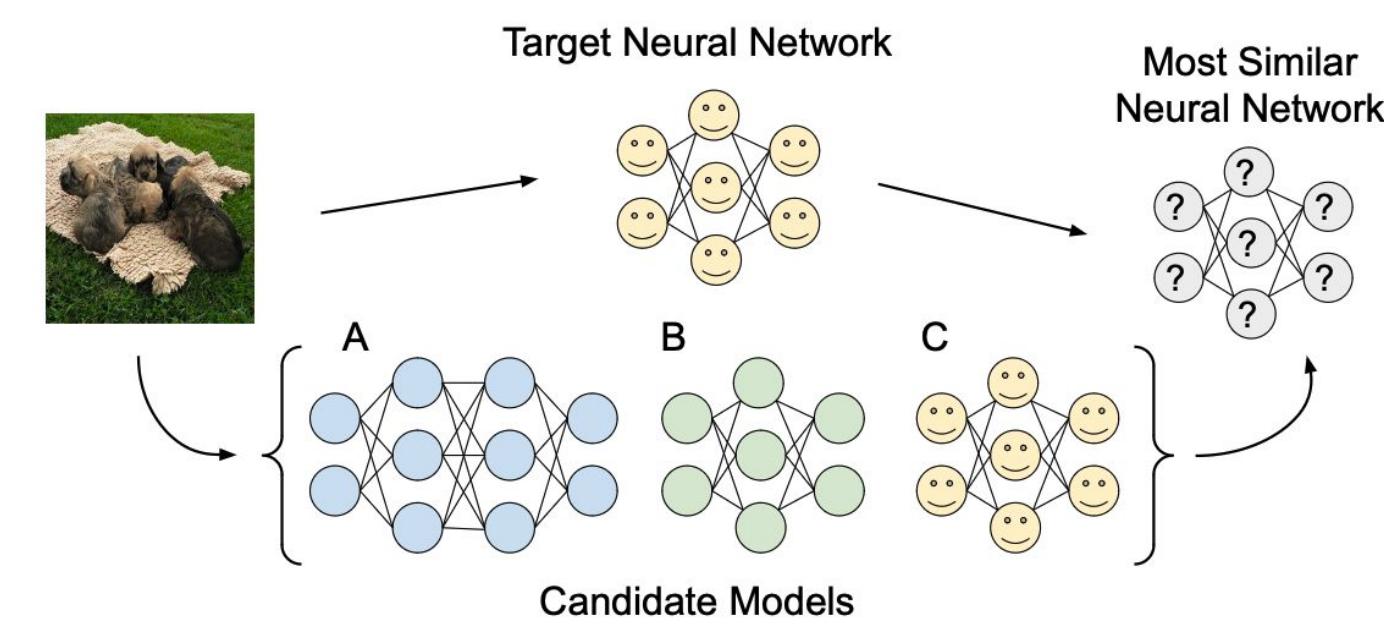
Overview

- Artificial intelligence has led to the development of vision systems that resemble primate visual systems at the level of representations and behavior.
- Identifying individual differences among high-performing AI systems and comparing them with primate visual systems is challenging.
- Leveraging disagreement amongst AI models can maximally distinguish models.
- Systematically exploring the space of identifiable stimuli can break the noise barrier.
- Our approach aims to provide a method for comparisons **at scale**:
 - A live repository of the strongest AI models available today
 - Requires no supervision
 - Grows alongside AI progress in the future

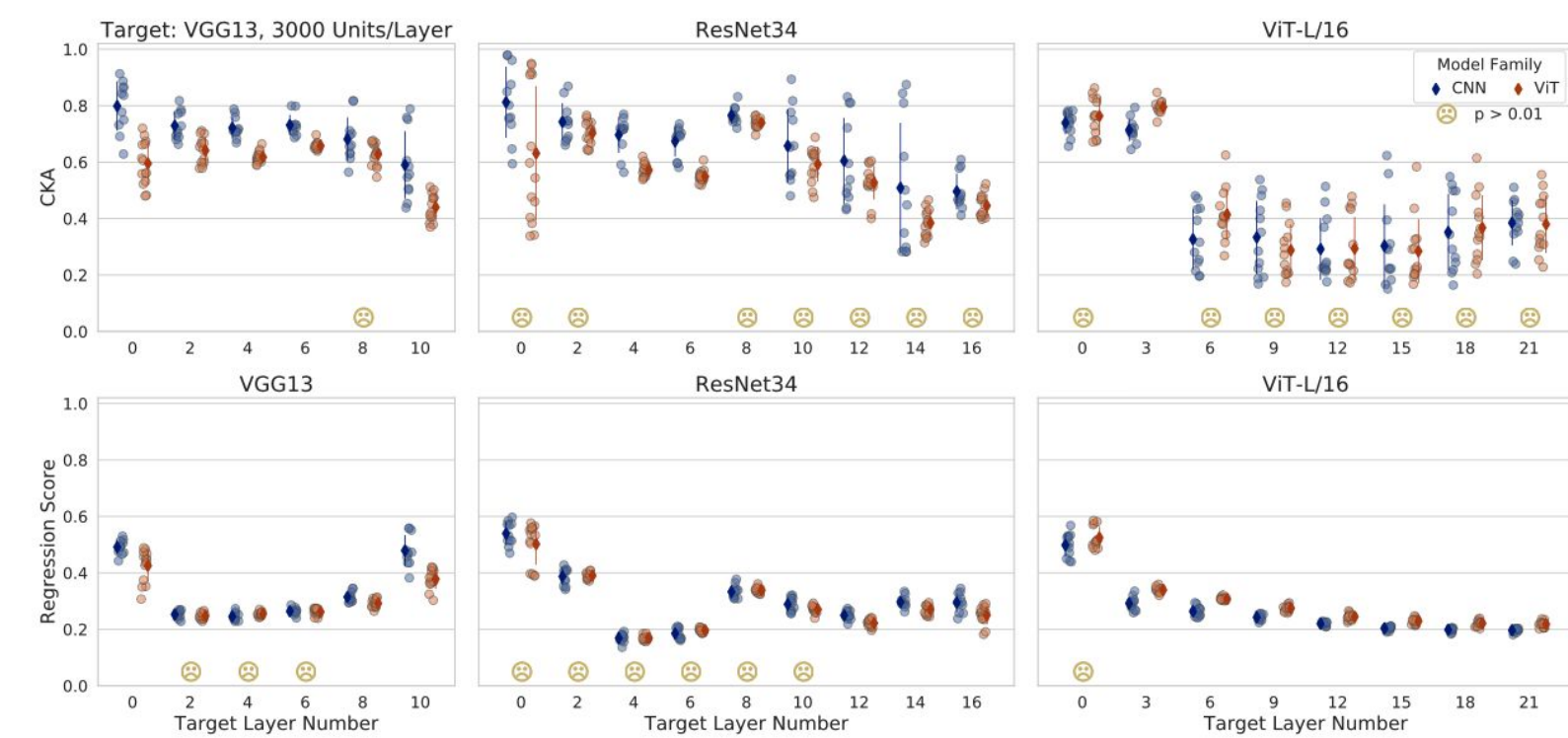
Neuroscience vs AI



System Identification as a Turing Test



The Result



Making Models Disagree

Why Disagree?

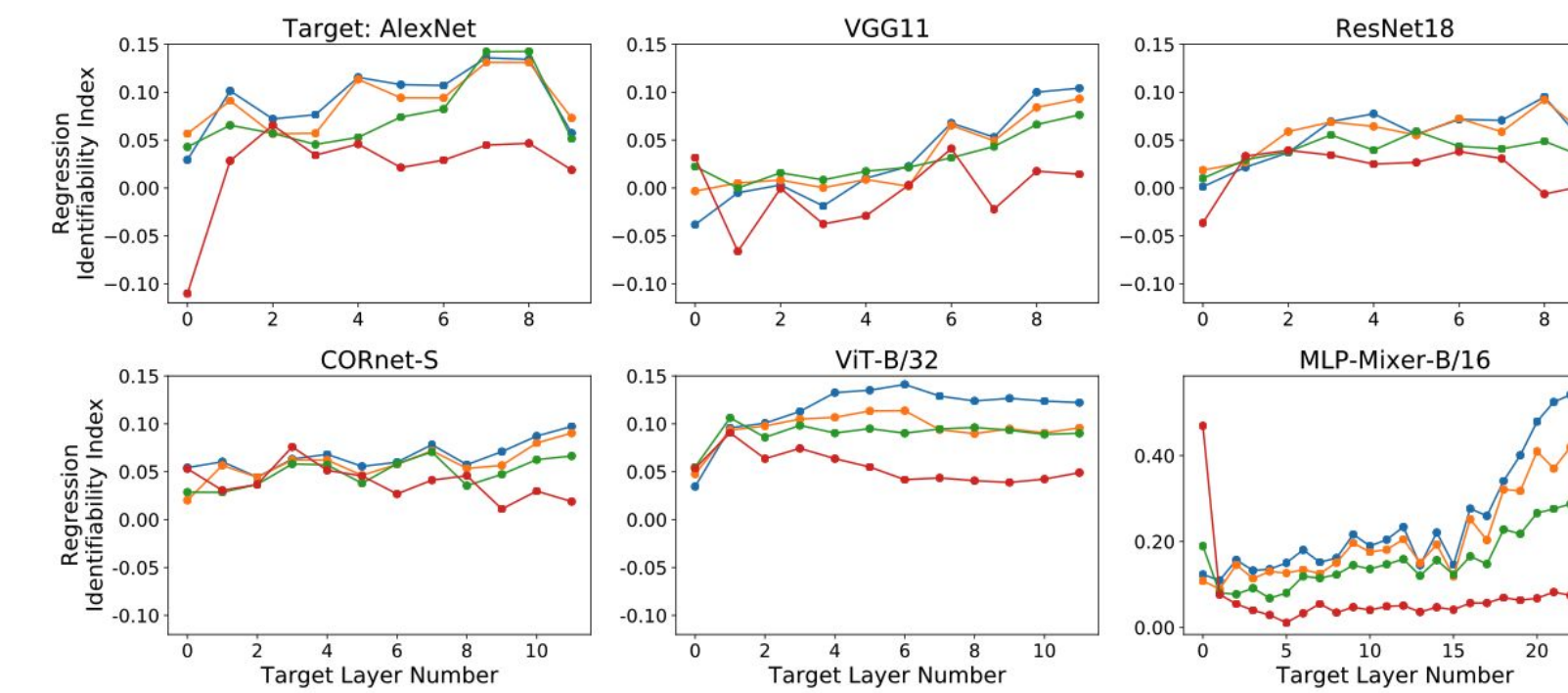
- Guarantees at least one model is wrong
- No supervision required
- Many ways to be wrong, fewer to be right



The Dress (2015)

Do models disagree about the same things?

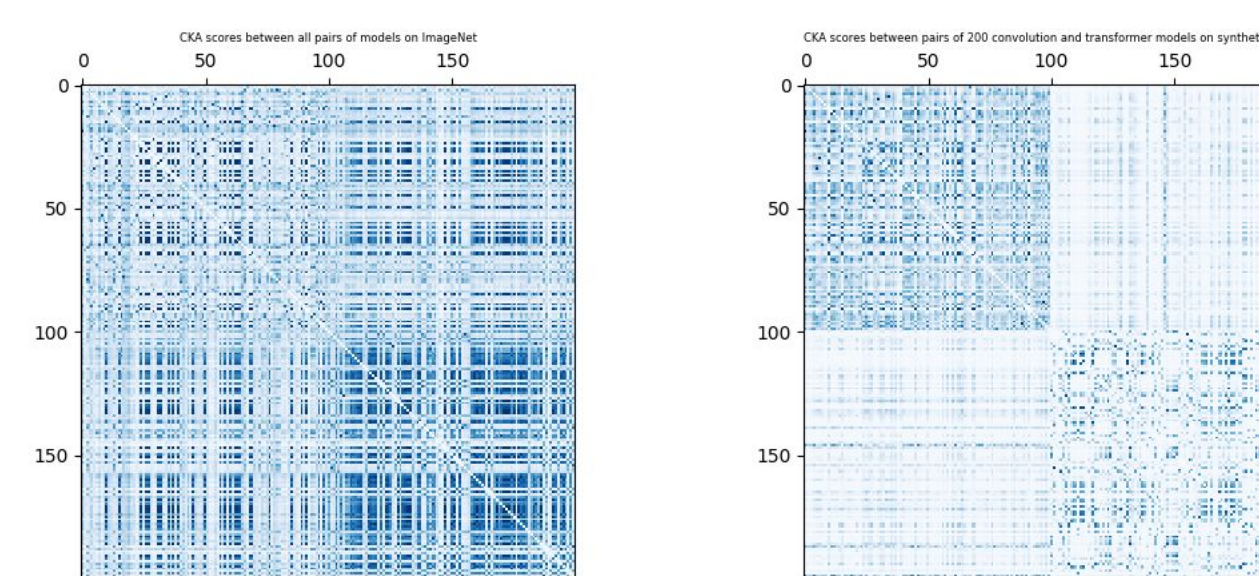
Data Matters



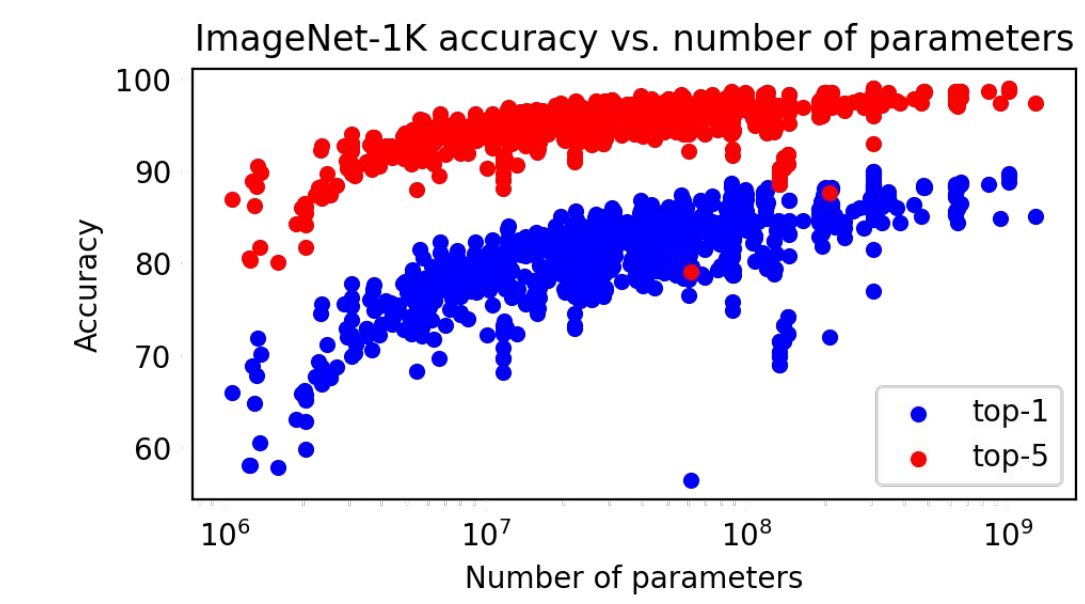
Disagreement Engineering

Centered Kernel Alignment (CKA):

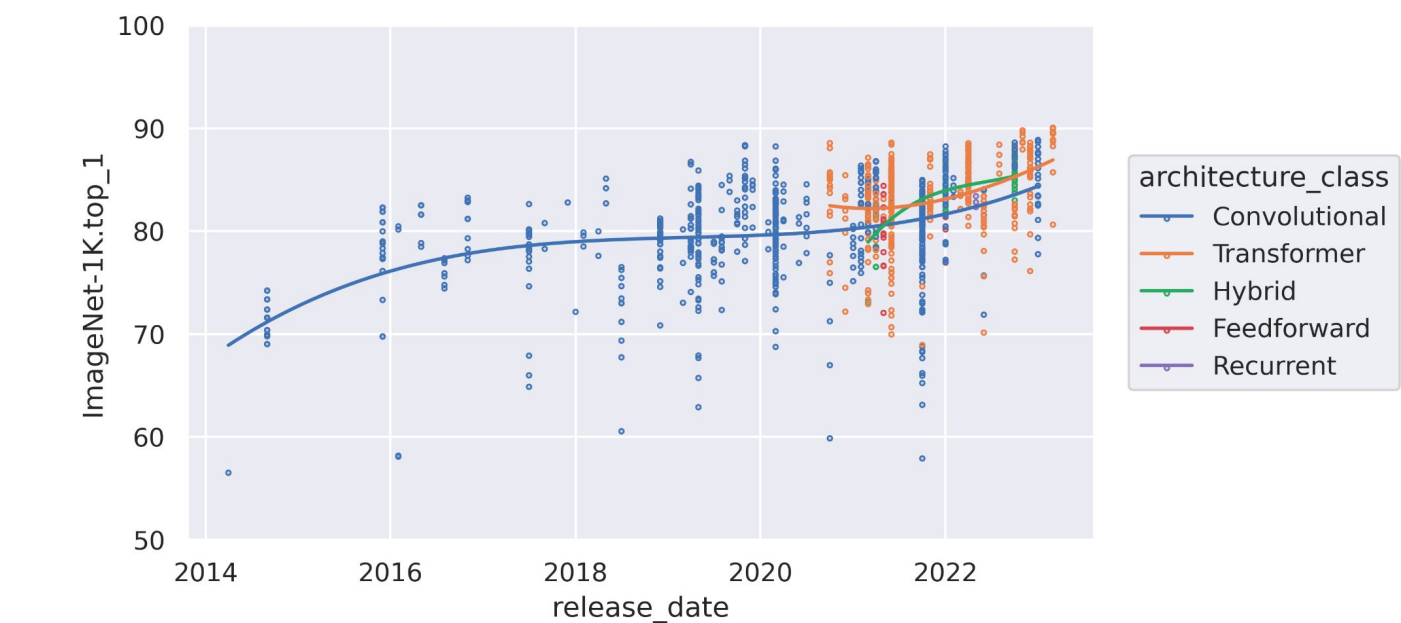
$$CKA(X, Y) = \frac{\|Y^T X\|_F^2}{\|X^T X\|_F \|Y^T Y\|_F}$$



Leveraging a Model Zoo



- 1000+ models extracted from Hugging Face 🤗:
- Architecture type
 - Number of parameters
 - ImageNet Score
 - and more!



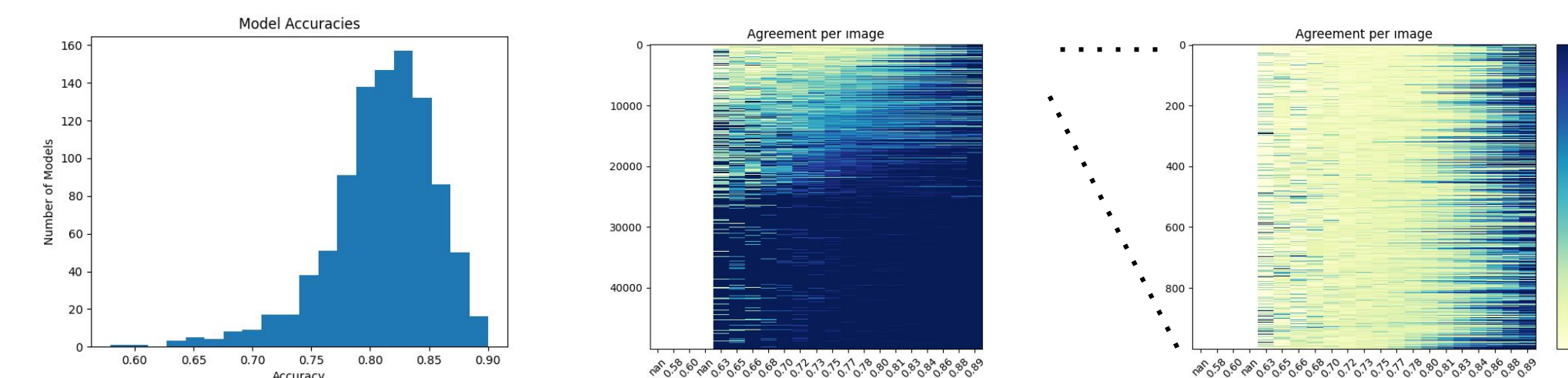
Measuring Disagreement

Fleiss' kappa 🌐 4 languages

Article [Talk](#) Read Edit View history Tools

From Wikipedia, the free encyclopedia

Fleiss' kappa (named after [Joseph L. Fleiss](#)) is a **statistical measure** for assessing the **reliability of agreement** between a fixed number of raters when assigning **categorical ratings** to a number of items or classifying items. This contrasts with other kappas such as **Cohen's kappa**, which only work when assessing the agreement between not more than two raters or the intra-rater reliability (for one appraiser versus themself). The measure calculates the degree of agreement in classification over that which would be expected by chance.



Disagreeable ImageNet



Relevant Work

System Identification of Neural Systems: If We Got It Right, Would We Know?

Yena Han¹ Tomaso Poggio¹ Brian Cheung¹

Controversial stimuli: Pitting neural networks against each other as models of human cognition

Tal Golan,^{a,1} Prashant C. Raju,^b and Nikolaus Kriegeskorte^{a,c,d,e,1}

Brain-Score: Which Artificial Neural Network for Object Recognition is most Brain-Like?

Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib J. Majaj, Rishi Rajalingham, Elias B. Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Kailyn Schmidt, Daniel L. K. Yamins, James J. DiCarlo