Causality, computation and learning in the study of the mind

@kordinglab

The standard reductionist dream in neuroscience

- Take brain and behavior
- Describe it in terms of what brain regions do
- Describe those in terms of what microcircuits do
- Describe those in terms of what neurons do
- Describe those in terms of what molecules do

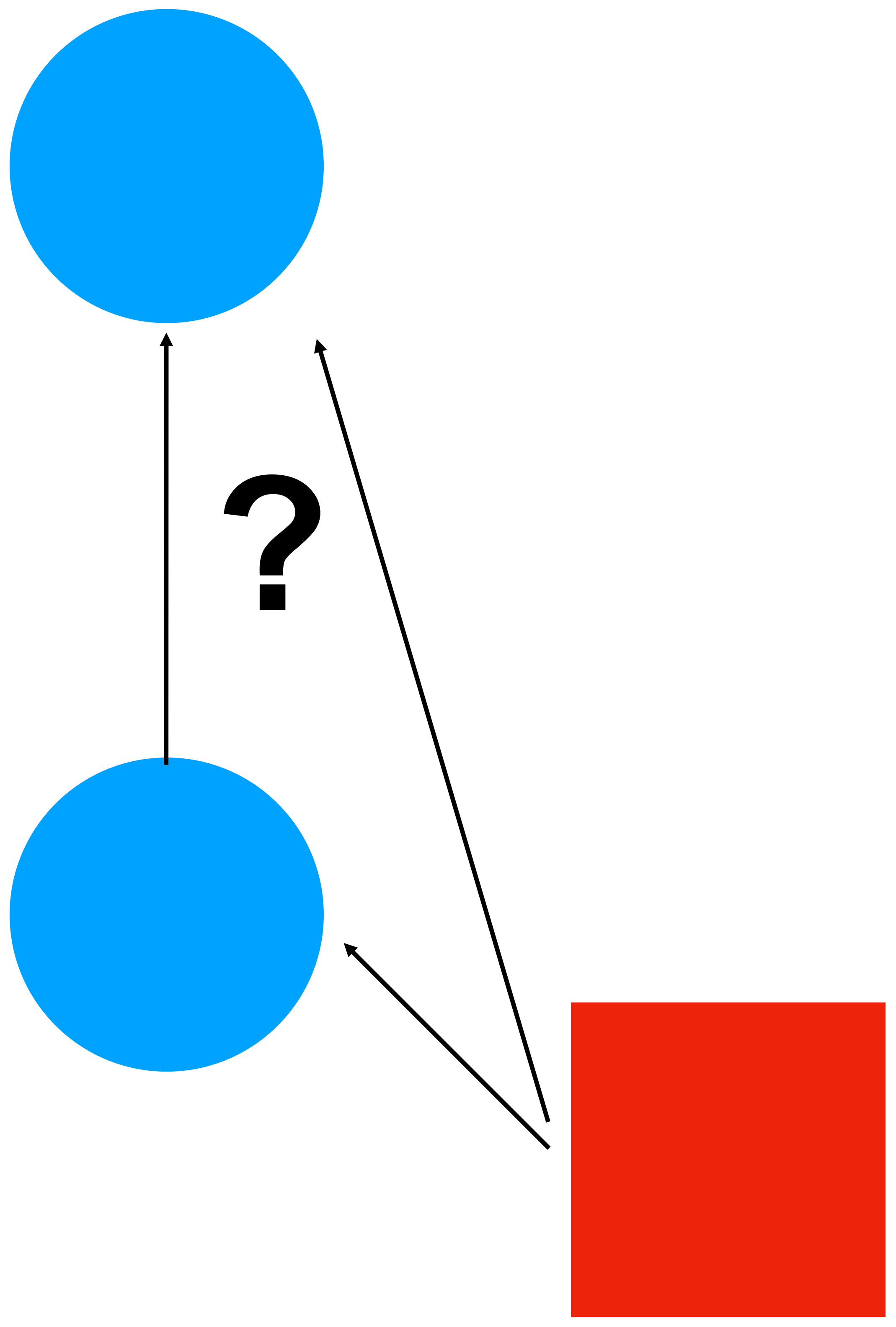
What is in each of these goals?

- A causal question:
  - How do parts of something make something happen at the bigger level
  - For this to be useful these parts must somehow be meaningfully independent
  - But also, information flow
- With the assumption:
  - The relevant causal inference/ discovery is a solvable problem
  - At that level of description there is a notion of simplicity

Why are causal answers so hard?

- Our perturbation methods are low-dimensional
- Our observational approaches are hopelessly confounded
  - Even if they were not we would be underpowered

Why causality is hard: Confounding



A continuum of confounding

- No confounders: e.g. atari, imagenet, go, chess
- Few confounders: starcraft
- Countless confounders: medicine
- 10^11 confounders: brain understanding
- **Let us focus on intuition (Neuromatch Academy)**

Simulate a trivial causal system

$$\vec{x}_{t+1} = \sigma(A\vec{x}_t + \epsilon_t)$$

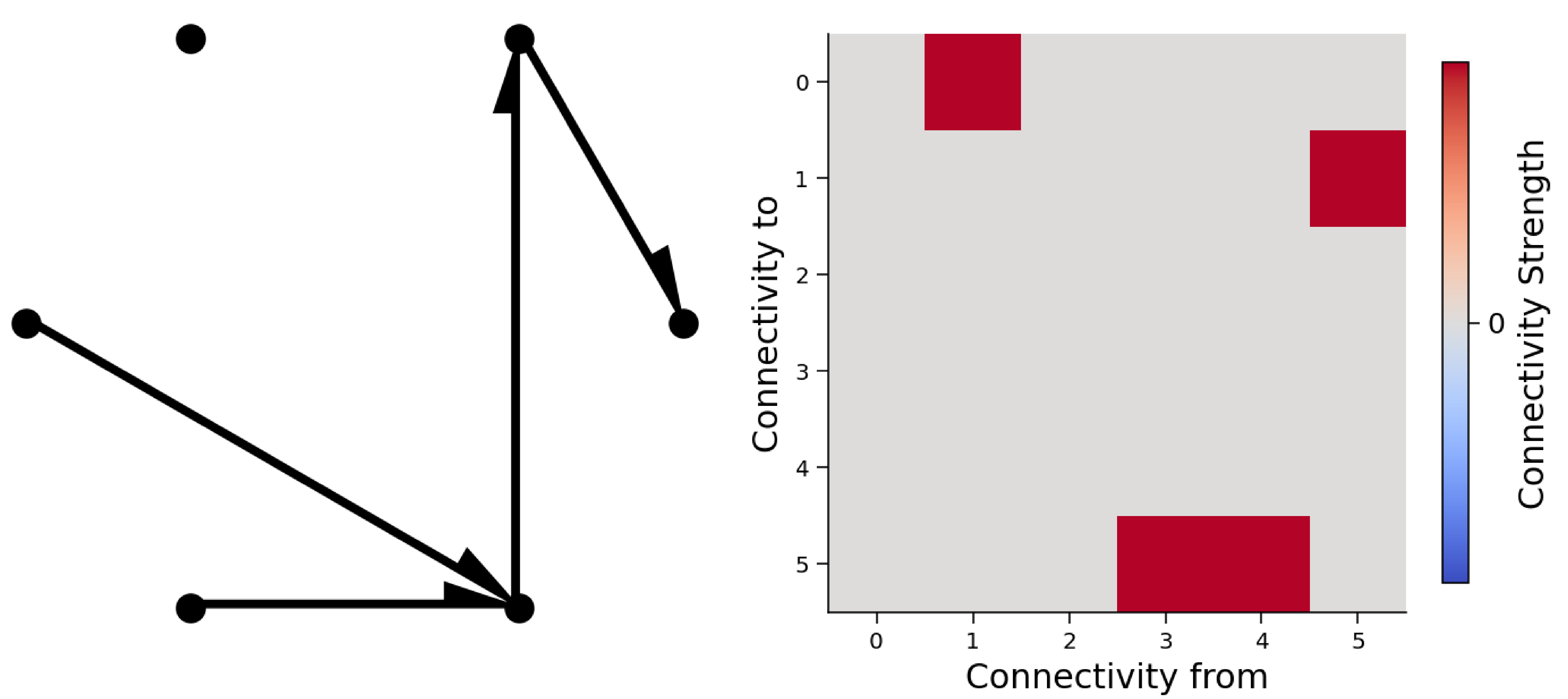
- $\vec{x}_t$ is an n-dimensional vector representing our n-neuron system at timestep t
- σ is a sigmoid nonlinearity
- A is our $n \times n$ *causal ground truth connectivity matrix*
- ϵ_t is random noise: $\epsilon_t \sim N(\vec{0}, I_n)$
- $\vec{x}_0$ is initialized to $\vec{0}$

Is correlation (delay =1) ~ causation?

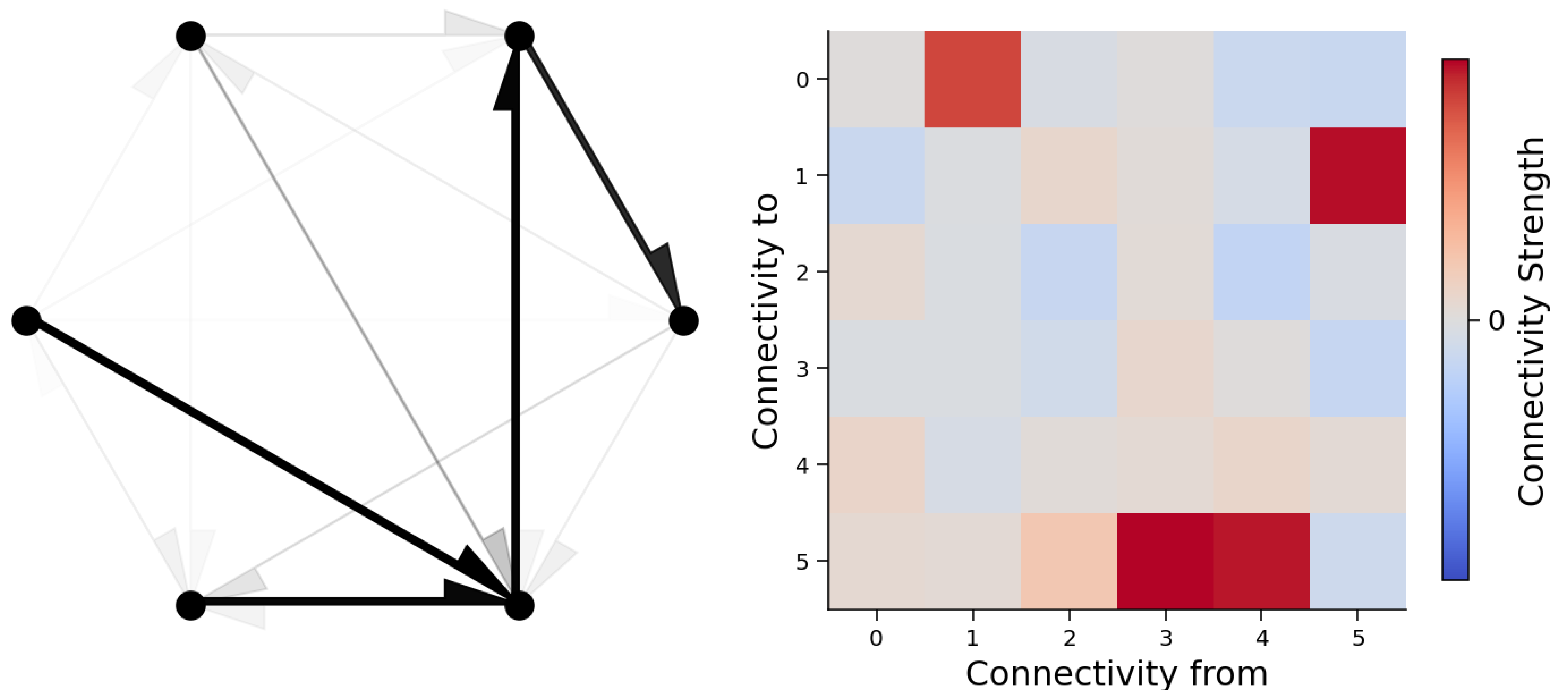
With Tony Liu, Ari Benjamin, Mike X Cohen, Madineh Sarvestani
Yoni Friedman, Ella Batty, Michael Waskom — D15 compneuro.neuromatch.io

Great in small system

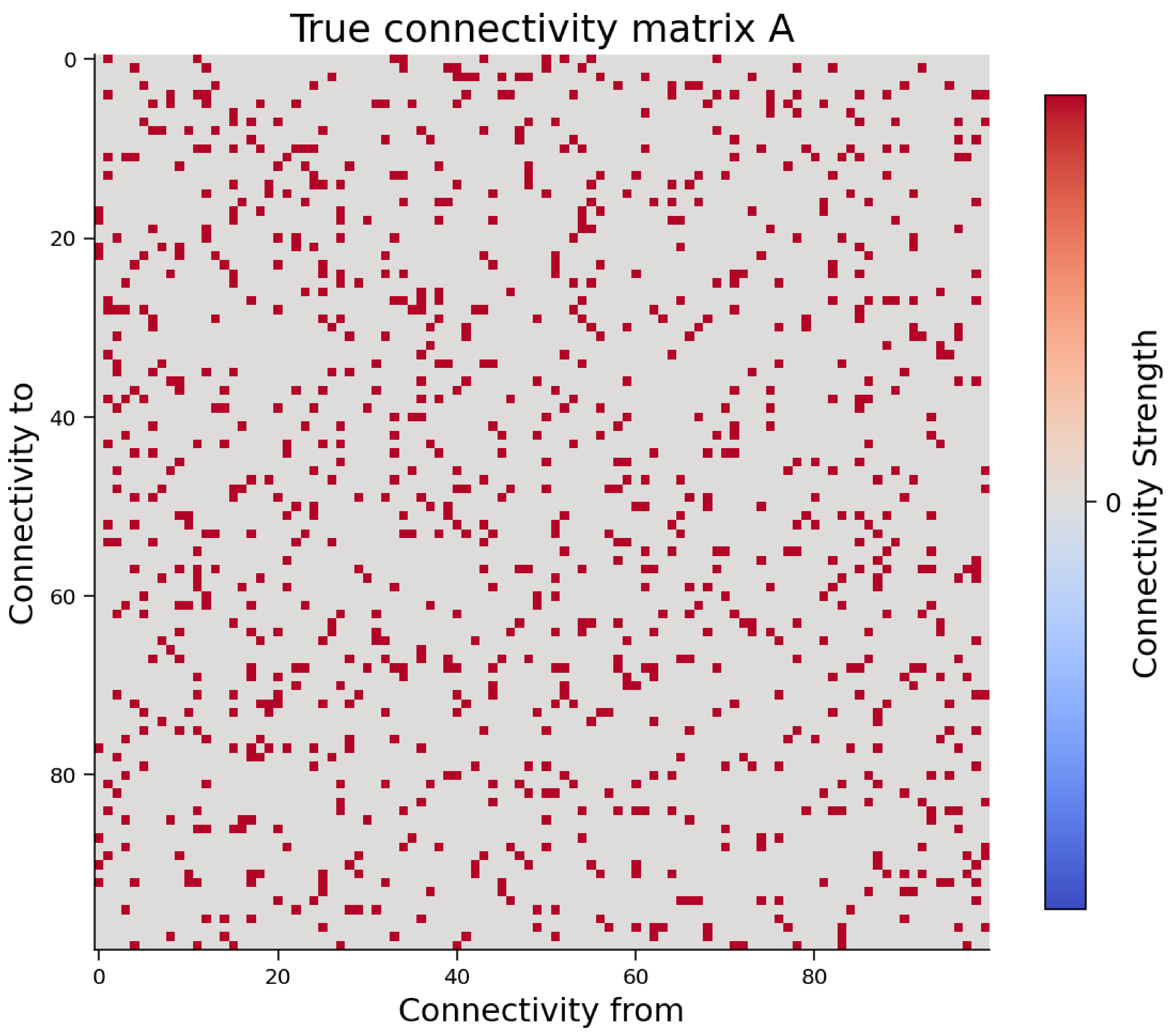
True connectivity matrix A

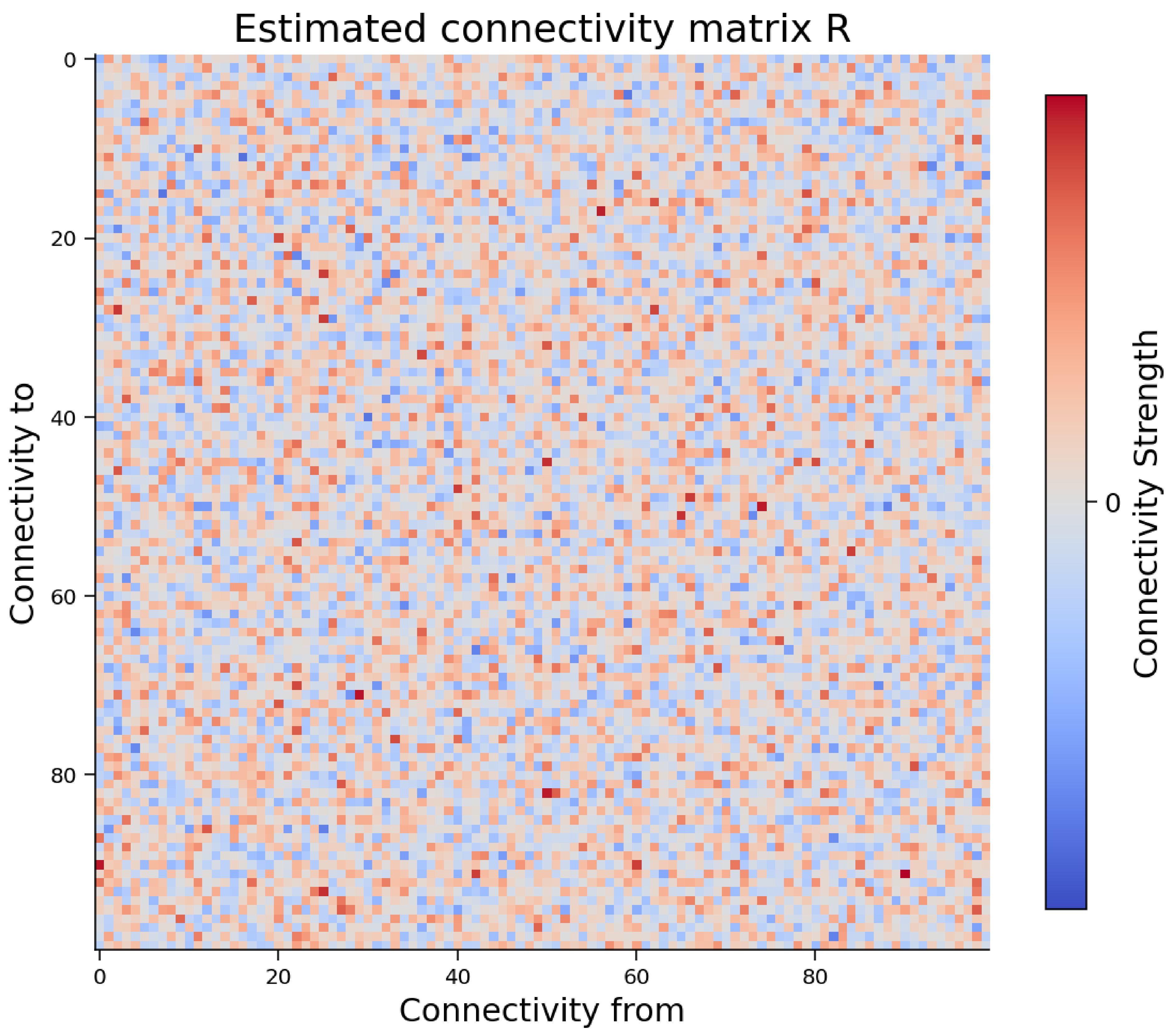


Estimated connectivity matrix R

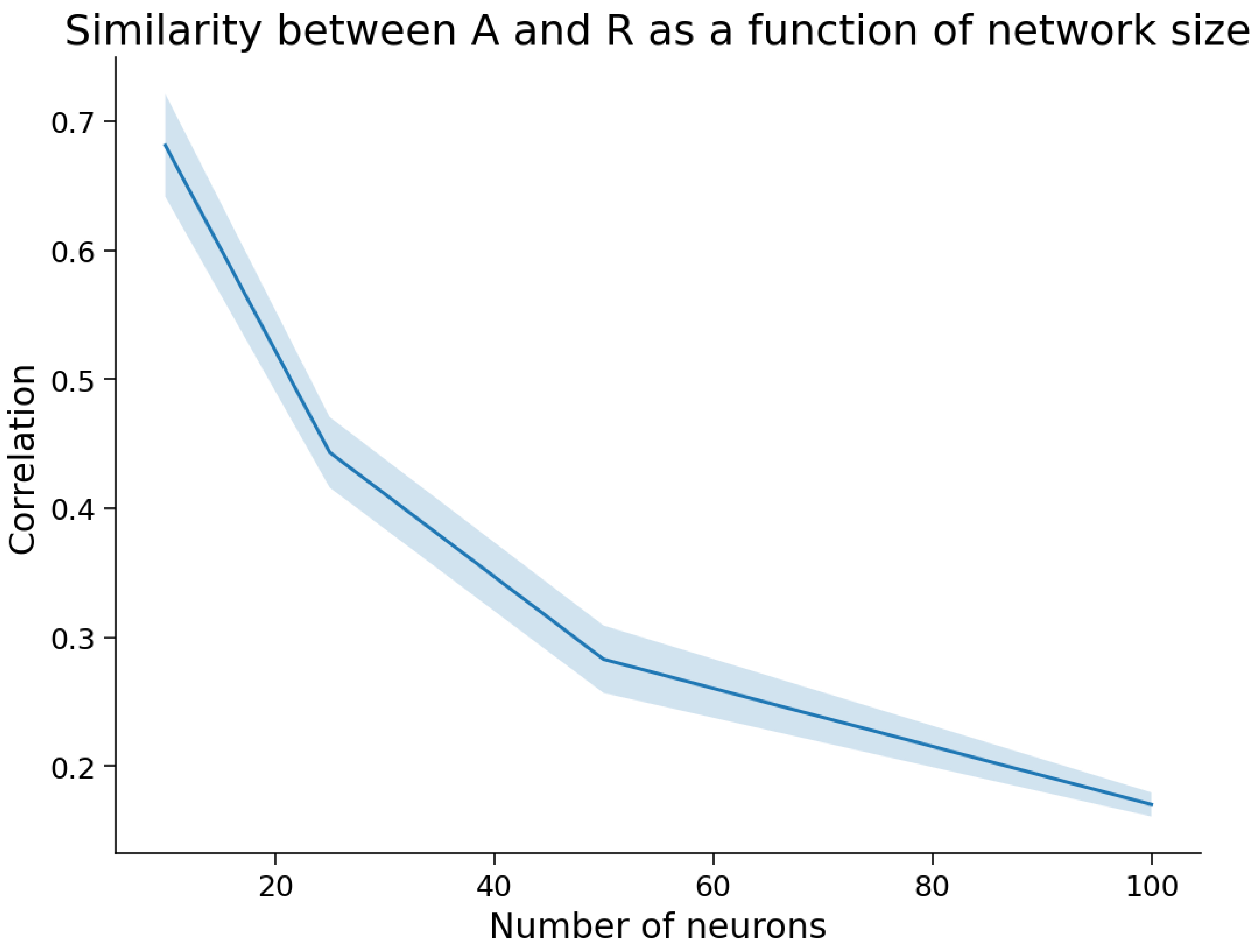


Bad in a big system





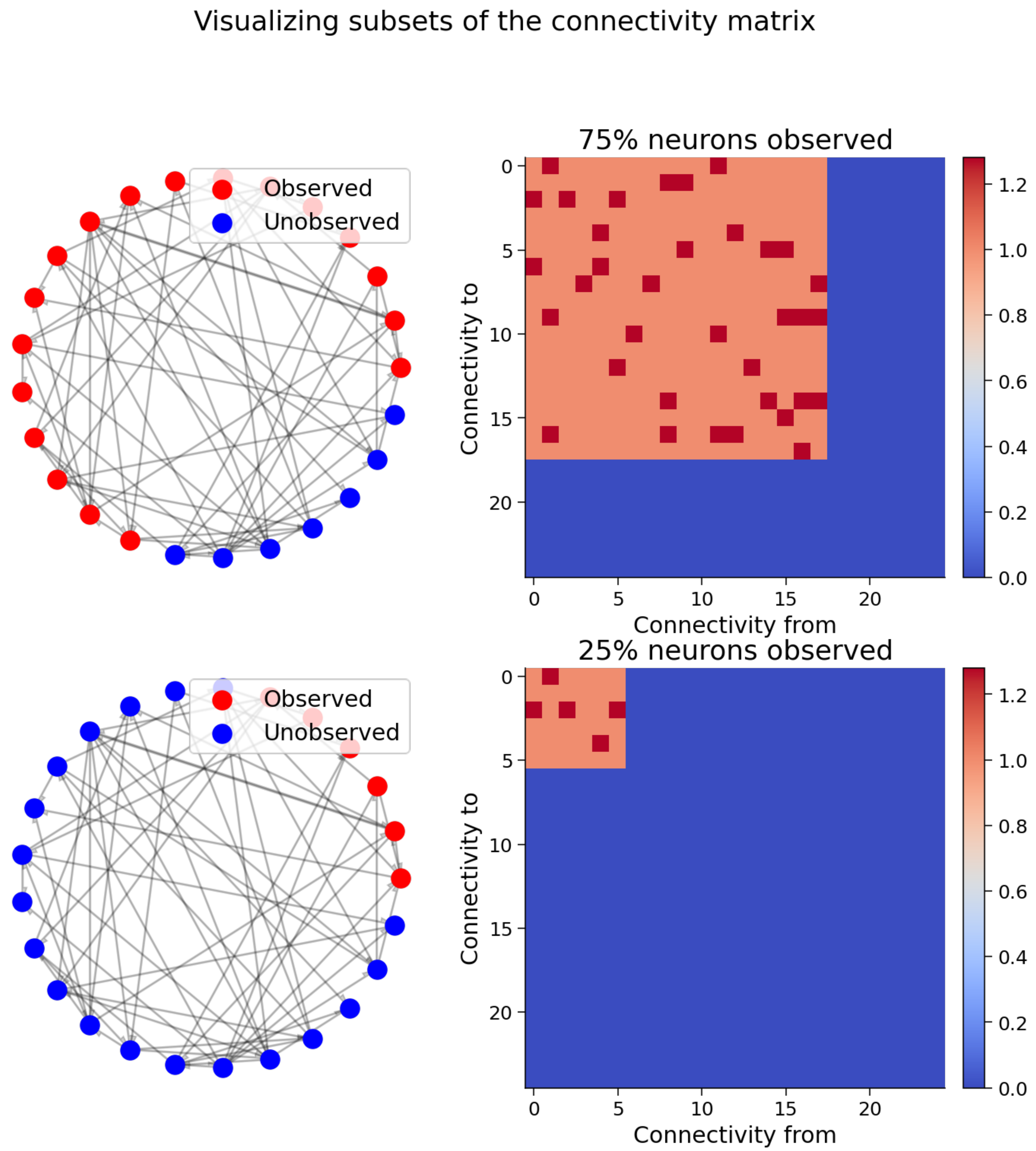
Delayed Correlation vs Causation



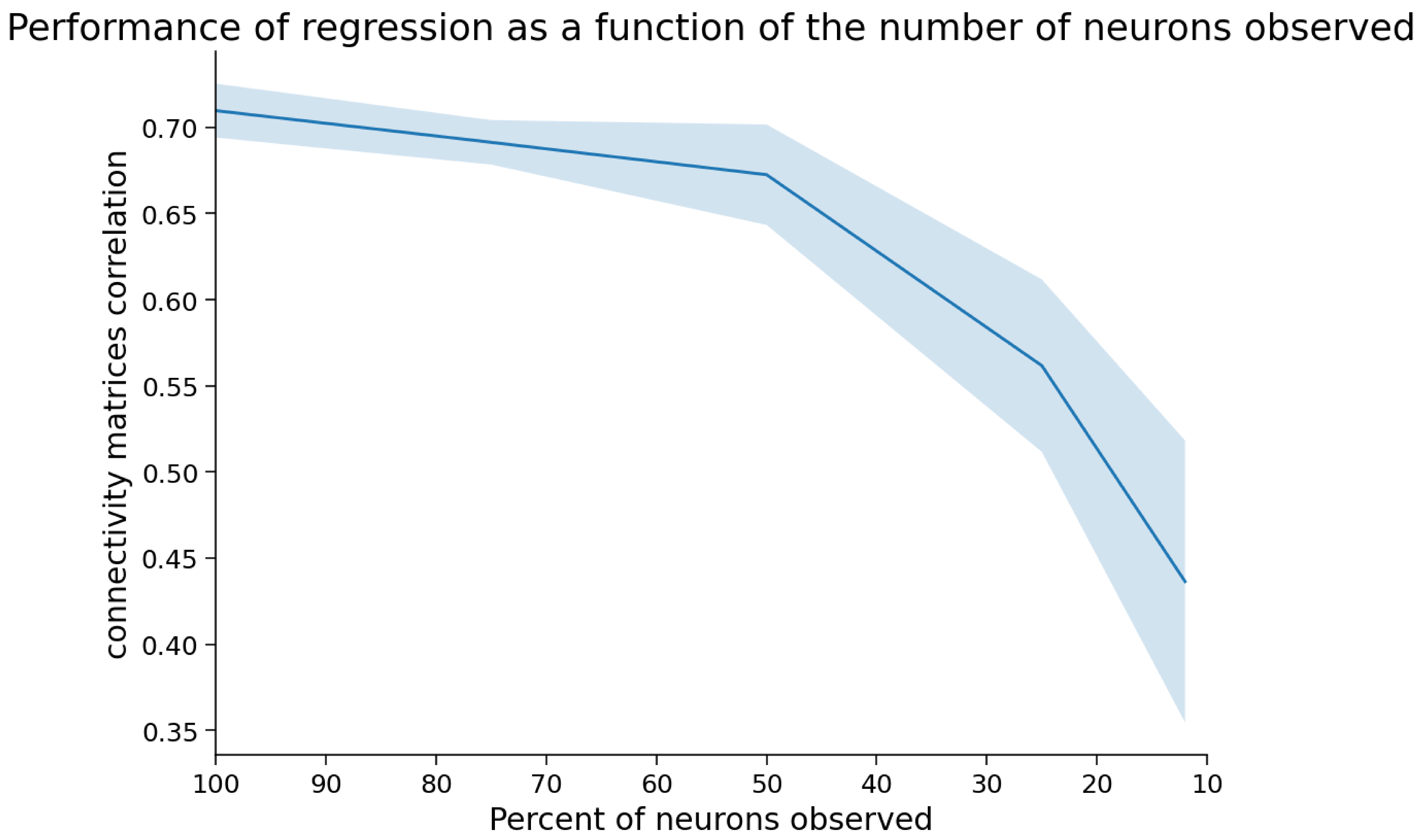
Fixable?

- Problem occurs from ignoring confounders
- It may be possible to improve by fitting full models instead of correlation

Alternative, record many neurons, fit jointly



Partial recording makes advantage of multiple regression go away



Why is regression a problem?

$$\widehat{\beta} = (X'X)^{-1}X'Y$$

Omitted Variable Bias Equation

$$y_i = x_i\beta + z_i\delta + u_i$$

$$\widehat{\beta} = (X'X)^{-1}X'(X\beta + Z\delta + U)$$

$$E[\widehat{\beta} \mid X] = \beta + (X'X)^{-1}E[X'Z \mid X]\delta$$

The bias should be arbitrarily big relative to the signal
This problem does not go away with more data

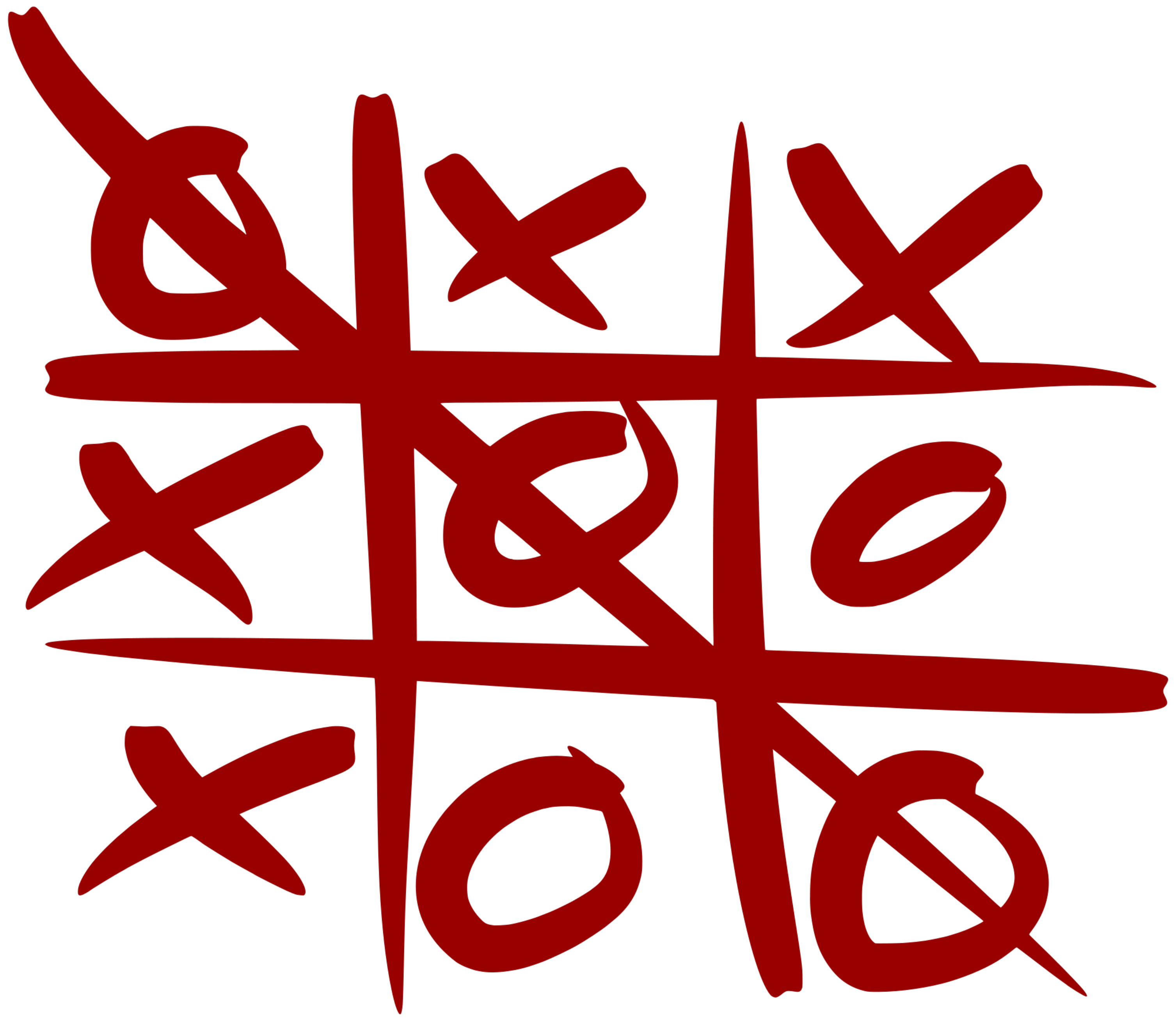
Measuring and interpreting neuronal correlations

Marlene R. Cohen[1] and Adam Kohn[2]

Flavors of understanding and link to complexity

- Know some truths about it
- Predict it
- Fix it
- Simulate it
- Understand how it works

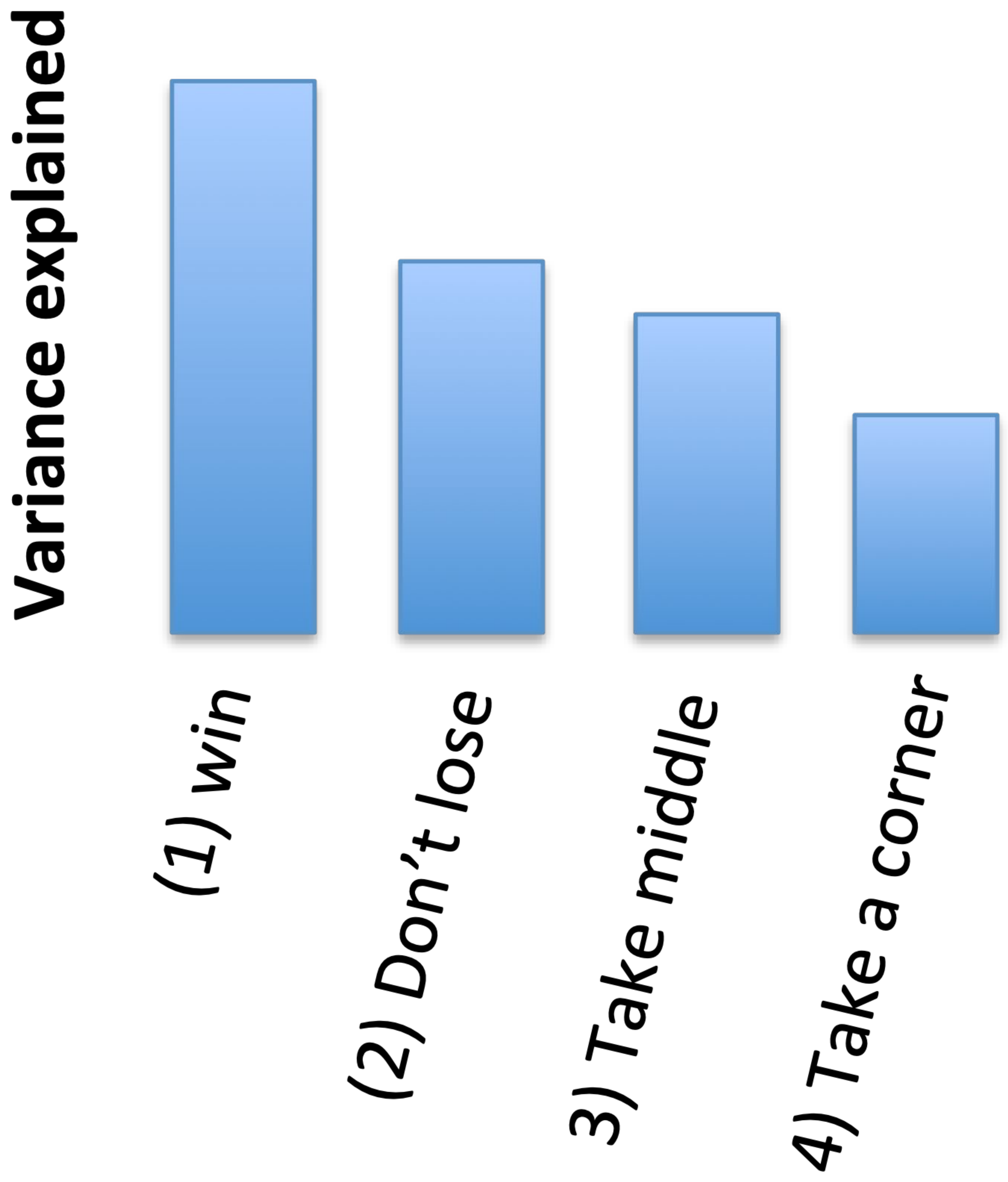
Think about complexity: Tic Tac Toe



255,168 distinct games!

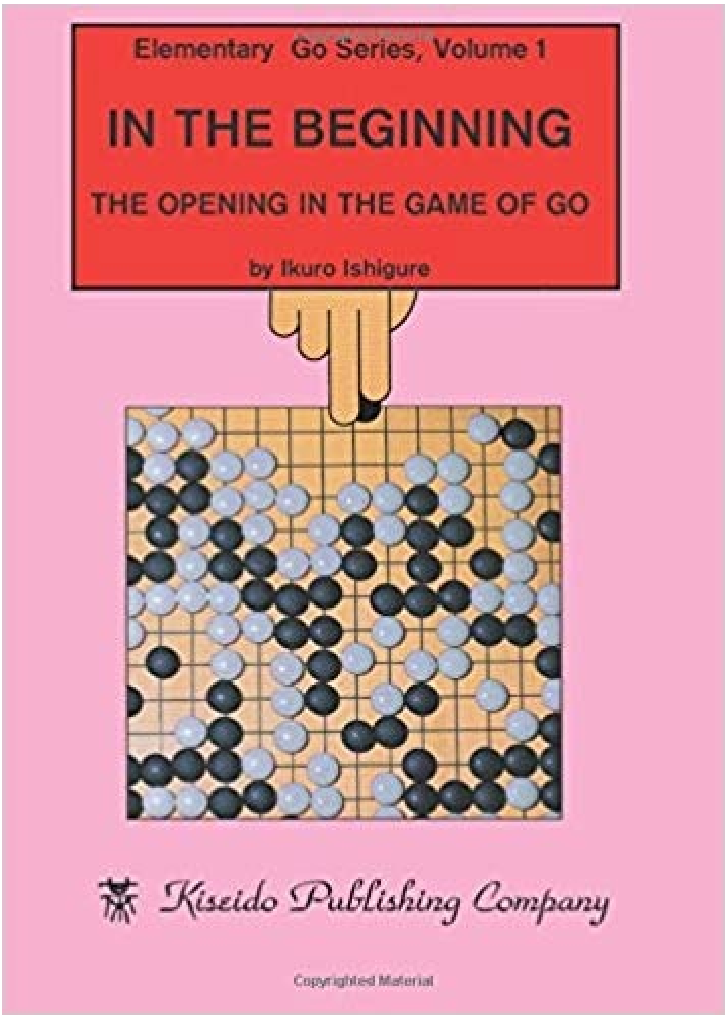
Lillicrap and Kording

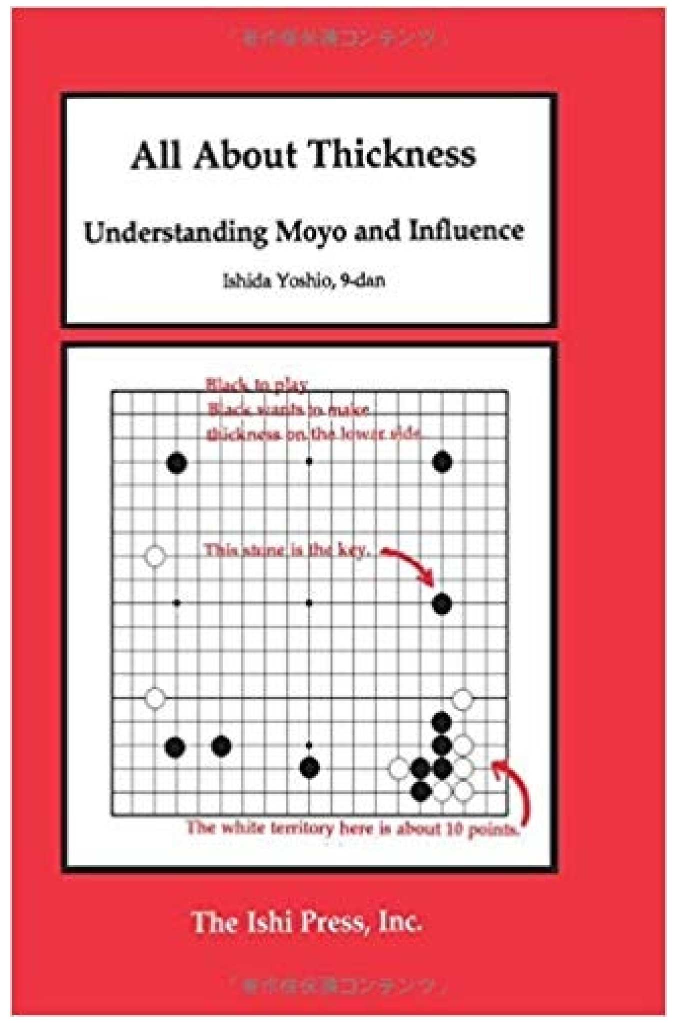
Compressible

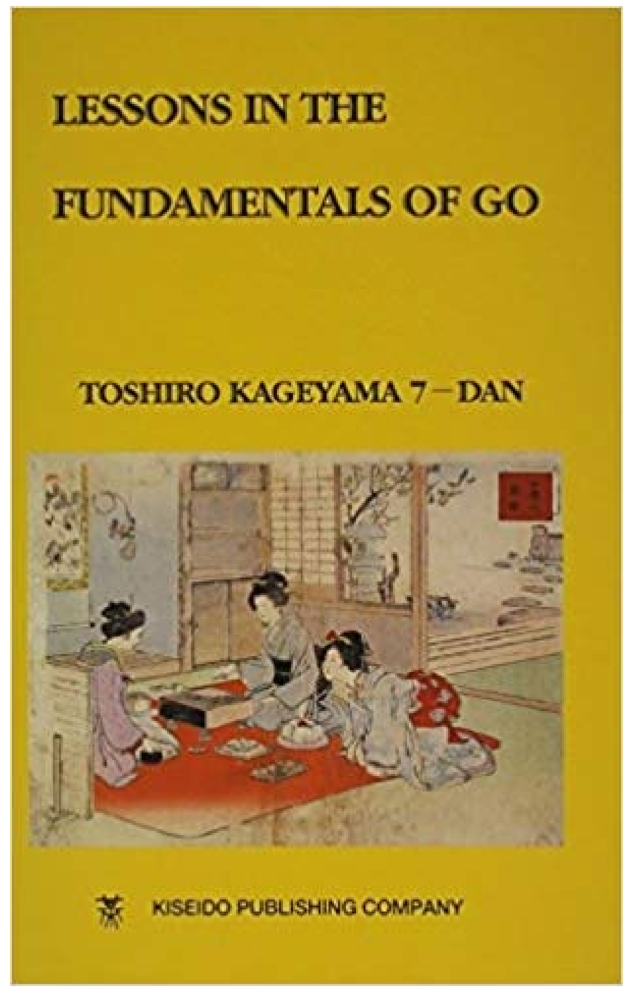


Go

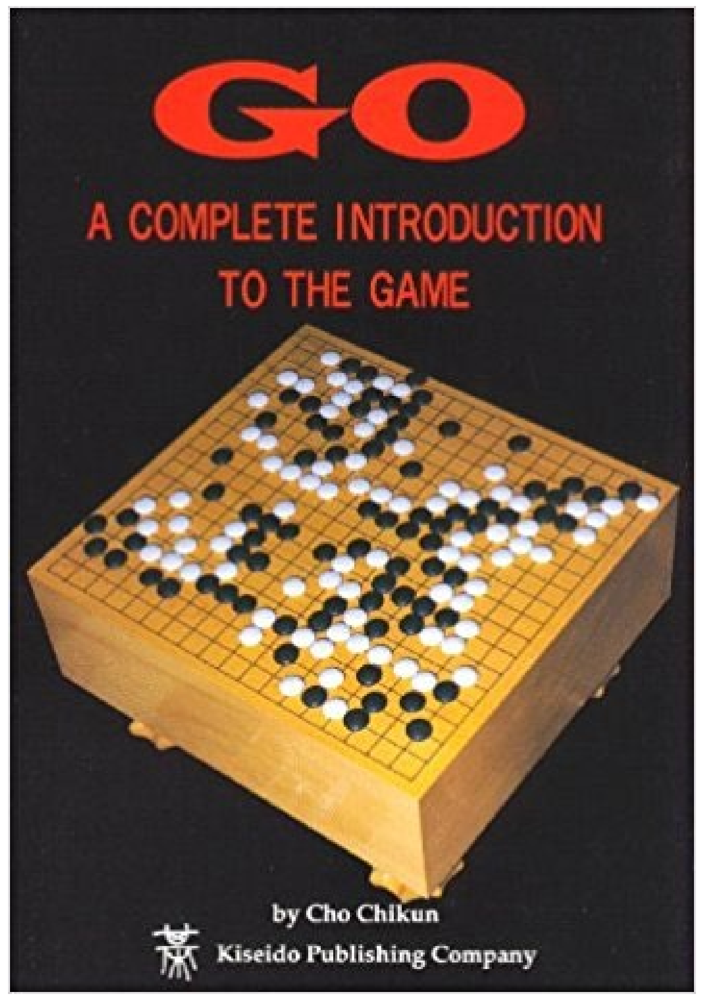
Probably no way to compactly describe it

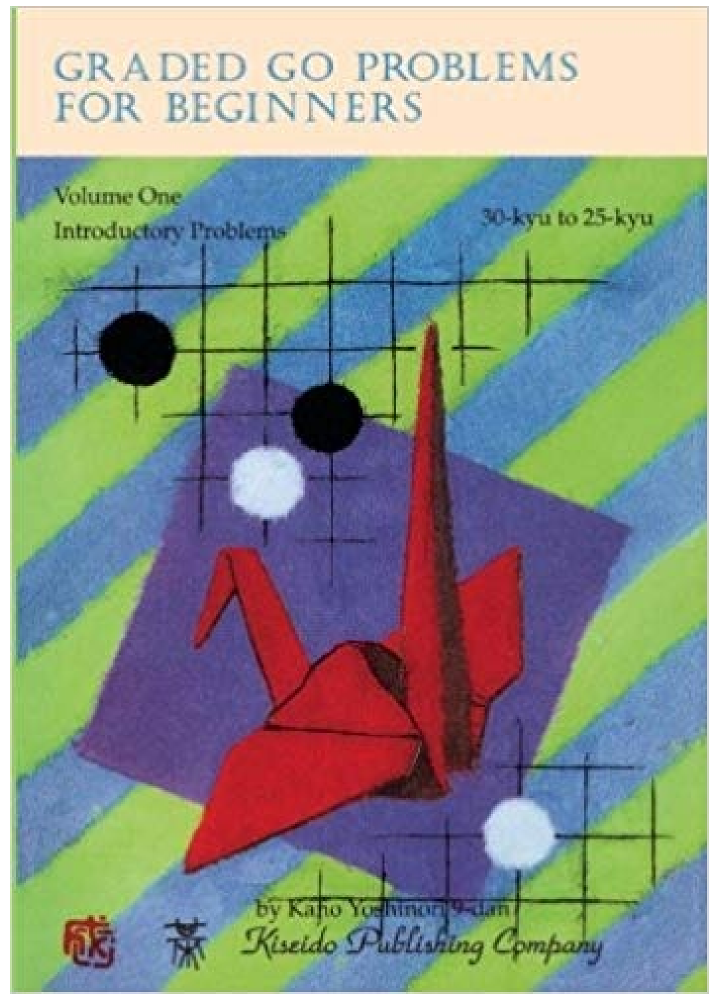


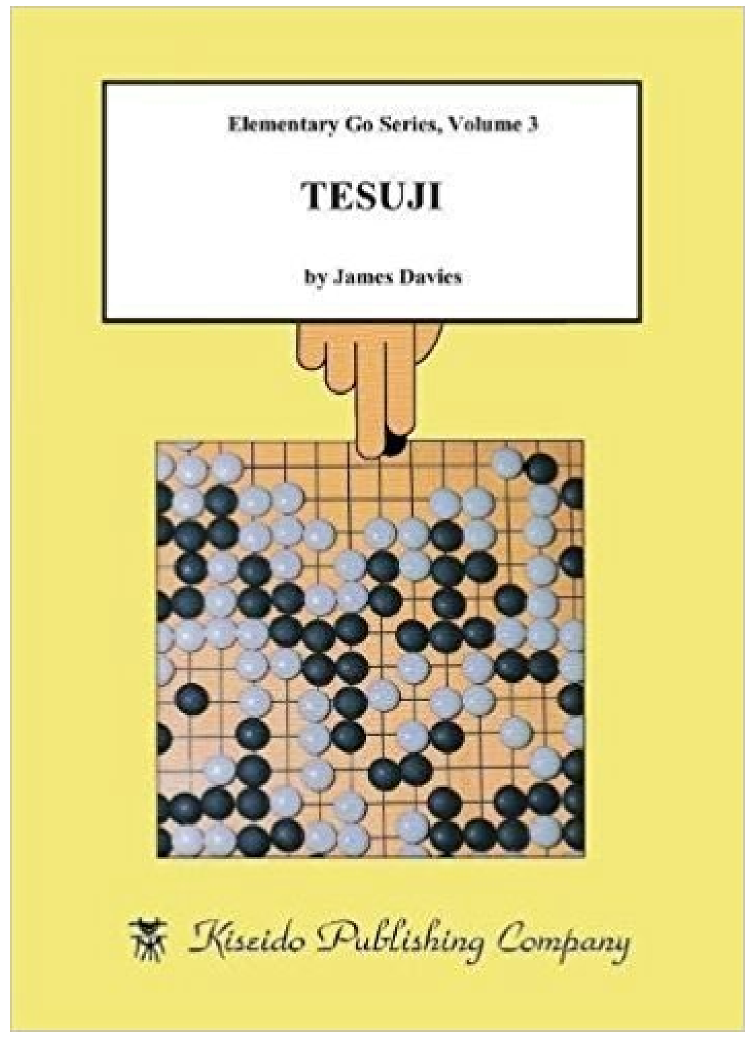


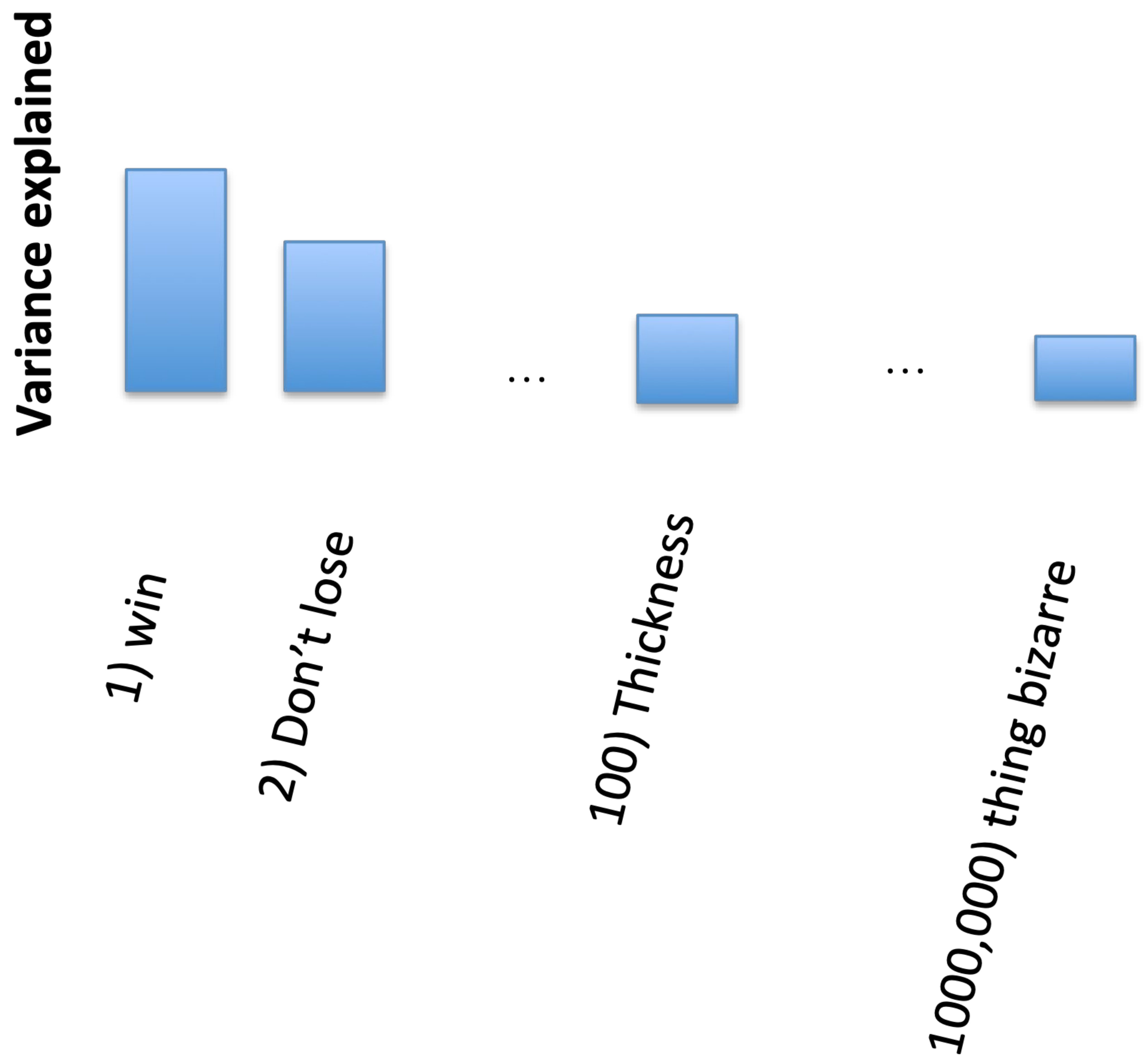


...









They are all real. Replicable from Go grand master to Go grandmaster.

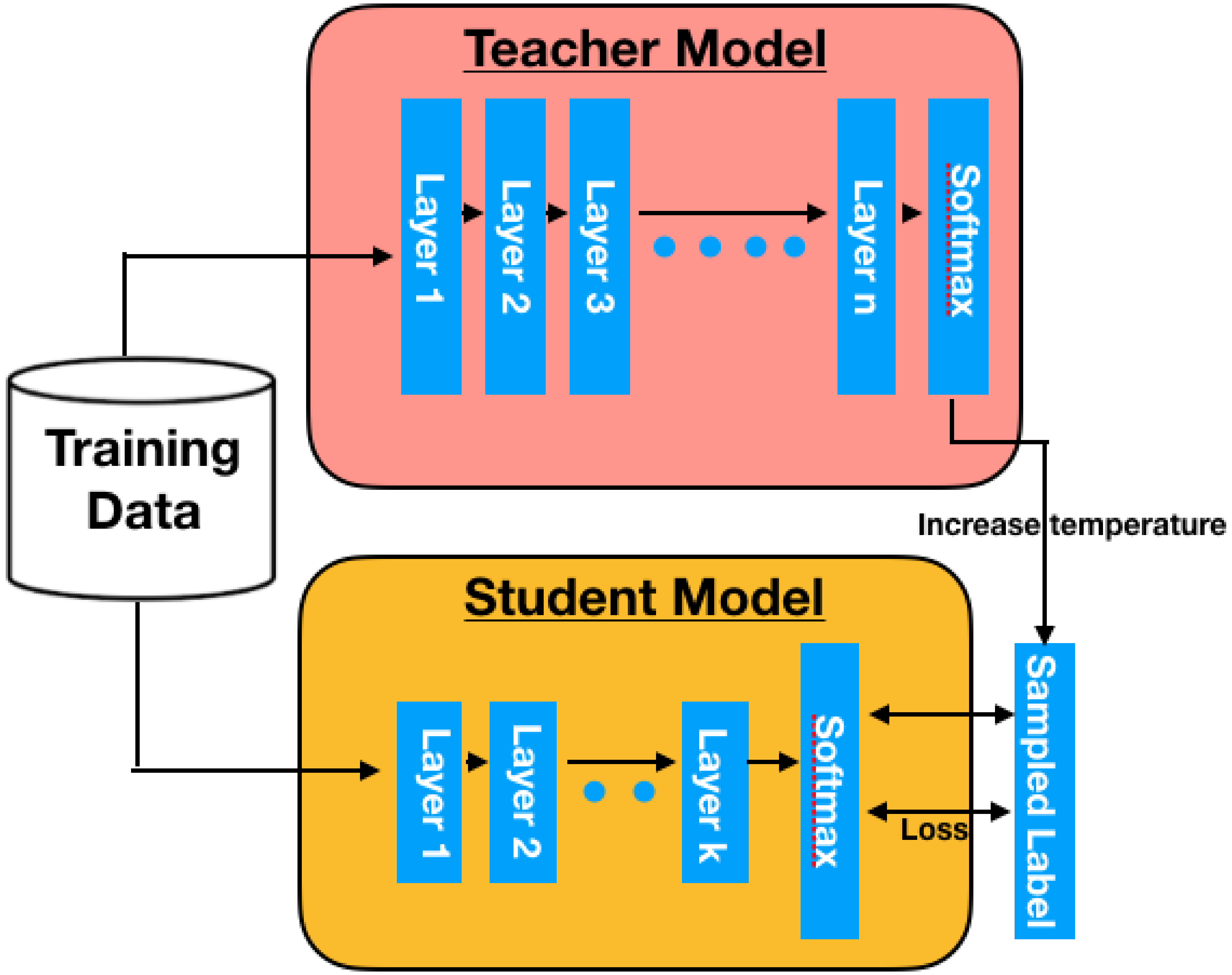
Understand a neural network

- **Pytorch code**
- Vs the **Weight dump**

How much information about the world does an intelligence have?

- Distillation
- Complexity calculations
- Back of the envelope calculations

Distillation



from mc.ai

Factor 10-100 on MNIST, imagenet

e.g. Ba and Caruana, Zhu et al 2018

Can we compress NNs?

- MNIST -> soft decision trees
    - BAD
- imagenet

Back of the envelope human

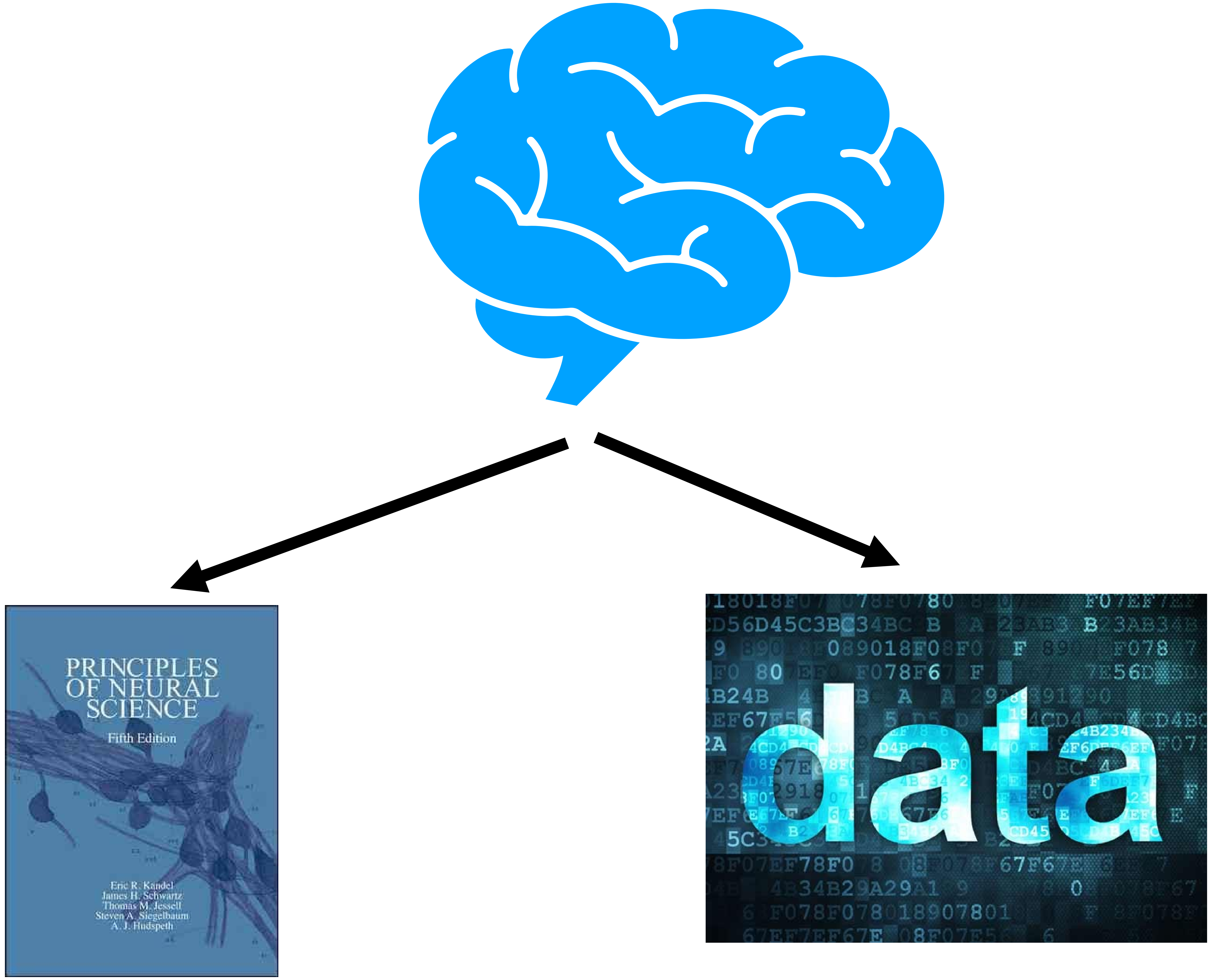
- 10 bits/s
- pi*10^7 seconds/a
- 30 years

- 10^10 bits
- 10^6 bits/book -> 10^4 books

H(DNA)<<H(World)

- DNA: 2*3*10^9 nucleotides
  - mostly non-nervous system
  - of nervous system possibly much non-computational
  - very non-compressed
- Nurture >> Nature

Ok. So what if the brain is not compressible?

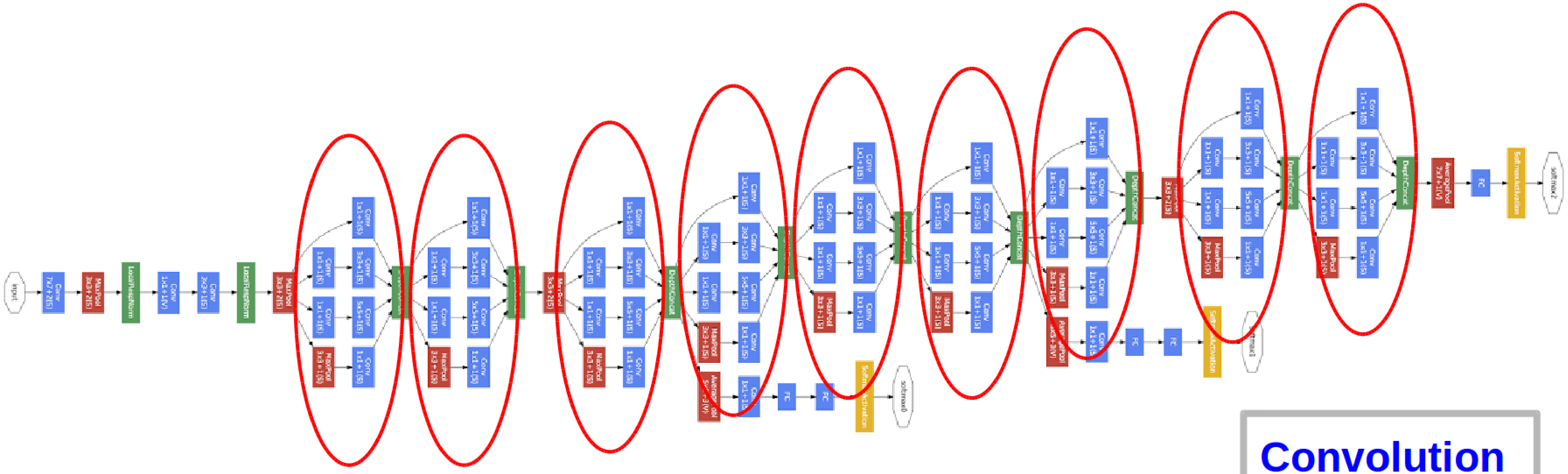


Argument in a nutshell

- Pytorch code to make an ANN is easy for us to understand
- Resulting network is (probably) impossible to understand
- So lets do the analogue of the first in neuroscience

- **Don’t study the mind. Study the process that makes the mind.**

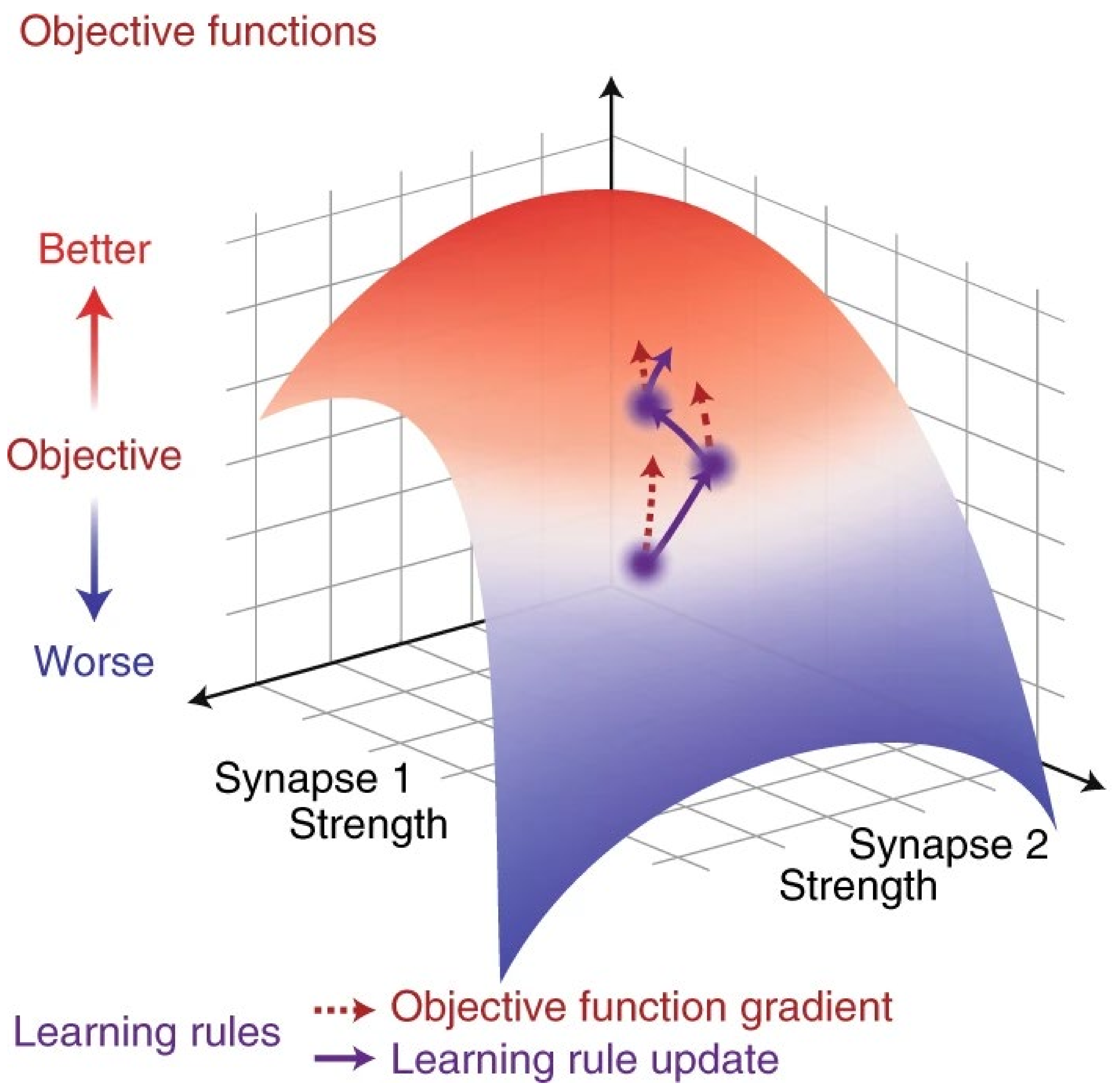
Architecture (Googlenet)



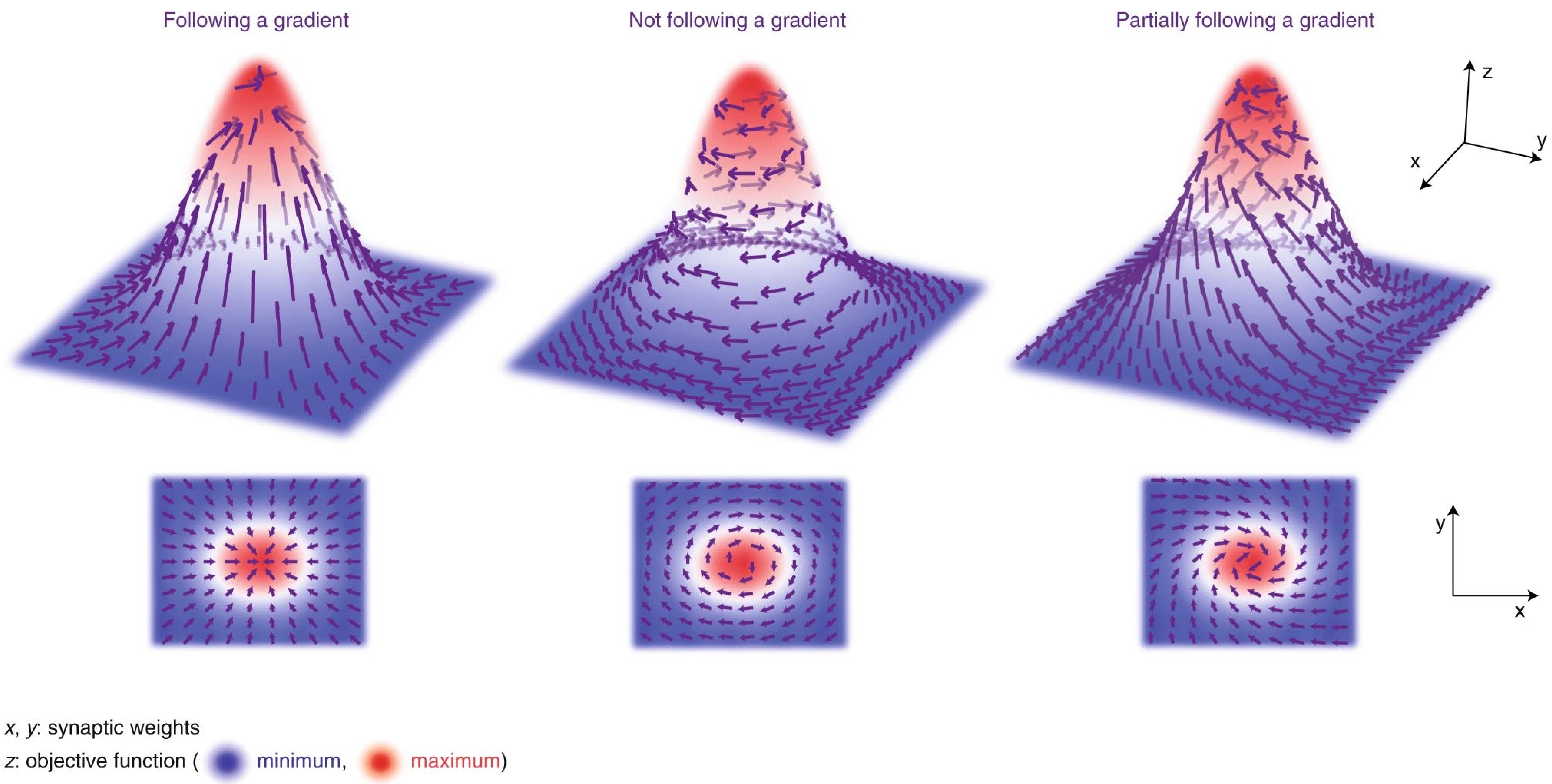
Convolution
Pooling
Softmax
Concat/Normalize

Richards et al review paper

Objective function (softmax)



Should brain follow a gradient?

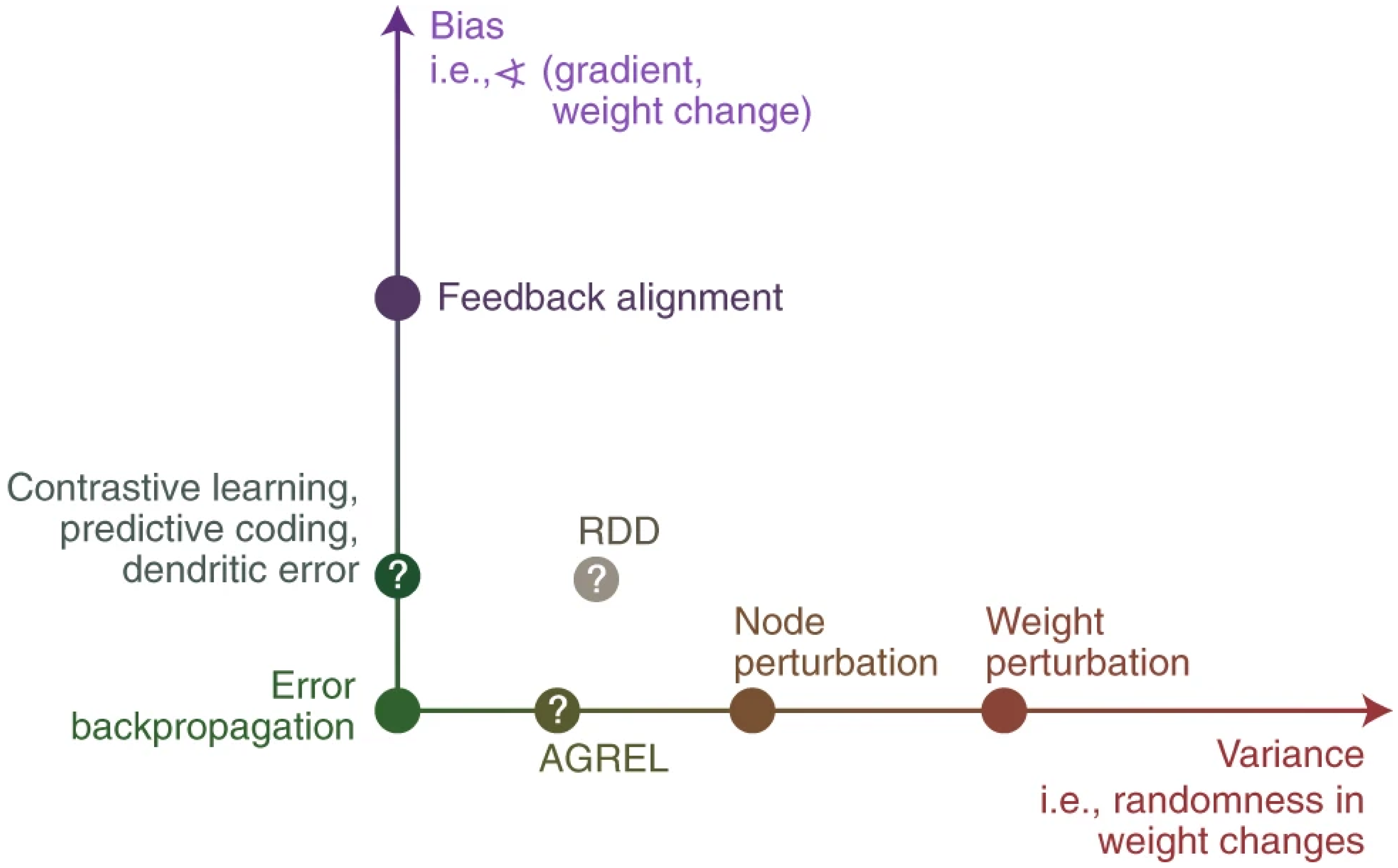


Optimizer (SGD)

$$Q(w) = \frac{1}{n}\sum_{i=1}^{n} Q_i(w)$$

$$w := w - \eta \nabla Q(w) = w - \eta \sum_{i=1}^{n} \nabla Q_i(w)/n$$

Efficiency is a real criterion



Data and embodiment

- Embodiment matters
  - Part of mechanism
  - Makes causality possible
  - Recasts problem of intelligence
  - Intelligence may be defined in relation to embodied cognition
- Curricula matter to make anything work
- A lot of aspects of data matter

Zador et al review paper

Learning is for the future

- Classical ML: Future distribution is like past distribution (i.i.d.)
- Multi-task: Future distribution is drawn from fixed distribution (i.i.d.)
- Etc
- But in reality: we want to be optimal in potential future world that relates to past worlds

For that reason we need

- Continual progressive learning
- Causal representations: causal structures are stable
- Curiosity: we want to learn what matters for future
- Constraints: we need to use our species’ past knowledge of evolutions

Three causal paths

- Bottom up: molecules -> spikes -> populations ->behavior
- Evolution: Ecological Niche -> Specification of Brains
  - There is something unique about the human niche
- Learning: Niche+Specification -> Actual Brains

Take home message

- Pytorch code to make an ANN is easy for us to understand
- Resulting network is (probably) impossible to understand
- So lets do the analogue of the first in neuroscience