What is a model?

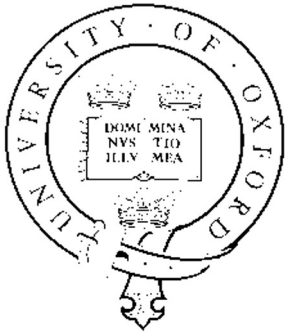
Christopher Summerfield
University of Oxford

BAMB! Summer School 2023

funding

European Research Council
Executive Agency

Human Brain Project



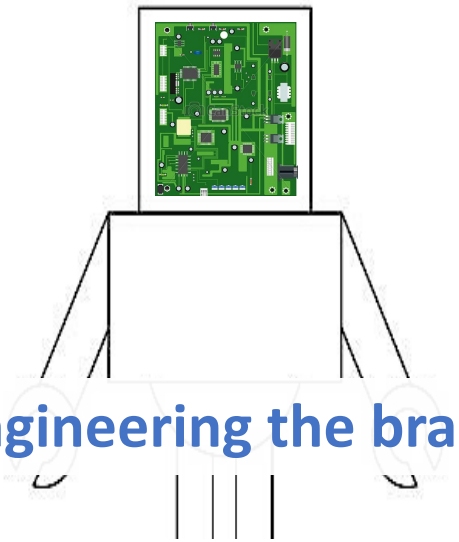
Models

UNIVERSITY OF OXFORD

Models are simulations (in computer code) of natural phenomena

Global finance
Language (LLMs)
Weather Forecasting
Structural Biology (e.g. Protein Folding)
Chess Computers
Ideal Gas Laws
etc

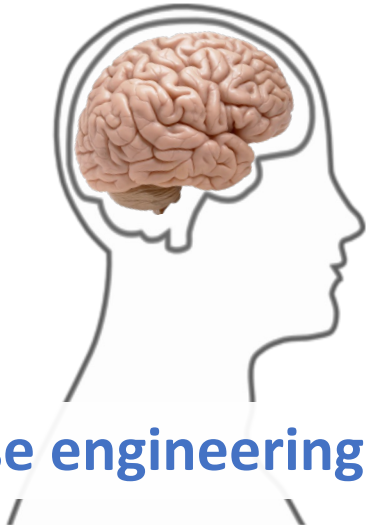
Outline



Engineering the brain

Artificial Intelligence: to build intelligent information processing systems *in silico*

Machine Learning: to use statistical principles to optimize information processing systems



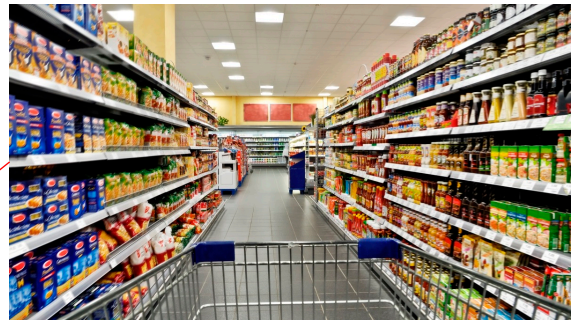
Reverse engineering the brain

Psychology: to understand the organisation of behavior and its foundations in cognition

Neuroscience: to understand neural coding and computation, and localise brain function

Why are we here?

predator-prey interactions



consumer behavior

mind / brain

social group dynamics

education

mental health

global finance

A problem in theory

nature human behaviour

PERSPECTIVE

https://doi.org/10.1038/s41562-018-0522-1

A problem in theory

Michael Muthukrishna [1*] and Joseph Henrich [2,3]

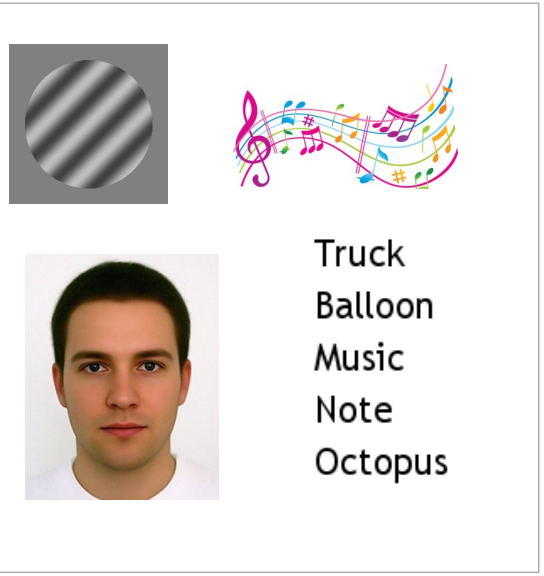
The replication crisis facing the psychological sciences is widely regarded as rooted in methodological or statistical shortcomings. We argue that a large part of the problem is the lack of a cumulative theoretical framework or frameworks. Without an overarching theoretical framework that generates hypotheses across diverse domains, empirical programs spawn and grow from personal intuitions and culturally biased folk theories. By providing ways to develop clear predictions, including through the use of formal modelling, theoretical frameworks set expectations that determine whether a new finding is confirmatory, nicely integrating with existing lines of research, or surprising, and therefore requiring further replication and scrutiny. Such frameworks also prioritize certain research foci, motivate the use diverse empirical approaches and, often, provide a natural means to integrate across the sciences. Thus, overarching theoretical frameworks pave the way toward a more general theory of human behaviour. We illustrate one such a theoretical framework: dual inheritance theory.

The psychological and behavioural sciences have a problem. By some accounts, half the literature doesn't replicate[1] and we don't know if the other half replicates for the 88% of our species who don't live in Western educated industrialized rich democratic (WEIRD) societies[2]. Although a few researchers insist that theory. If we discover fossil rabbits which appear to have originated in the Precambrian era, we would suspect something was wrong, because it conflicts with a cumulative understanding of how species evolved that has nothing to do with previous Precambrian finds per se but rather with a broad understanding of evolutionary change

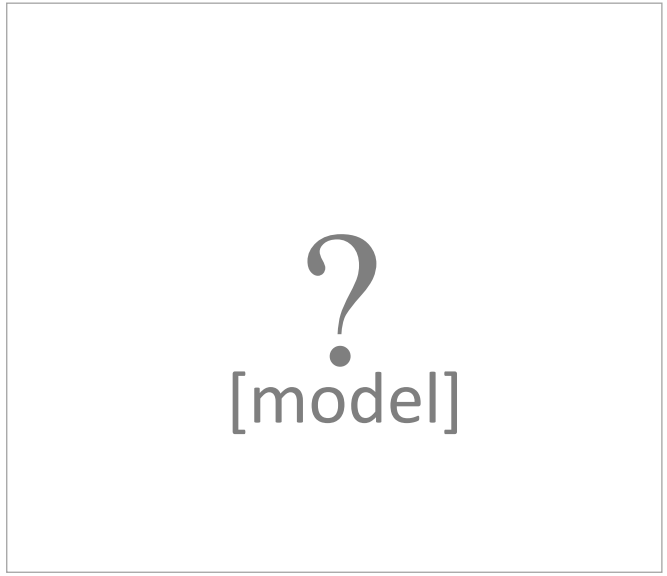
"Many subfields within psychology…lack any overarching, integrative general theoretical framework that would allow researchers to derive specific predictions from more general premises"

"Rather than building up principles that flow from overarching theoretical frameworks, psychology textbooks are largely a potpourri of disconnected empirical findings on topics that have been popular at some point in the discipline's history, and clustered based on largely American and European folk categories"

The information processing approach



input
(known)

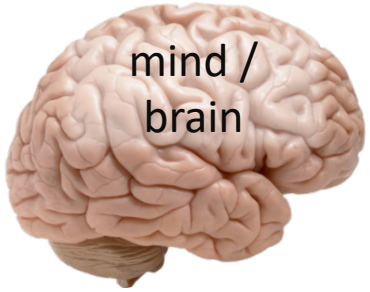


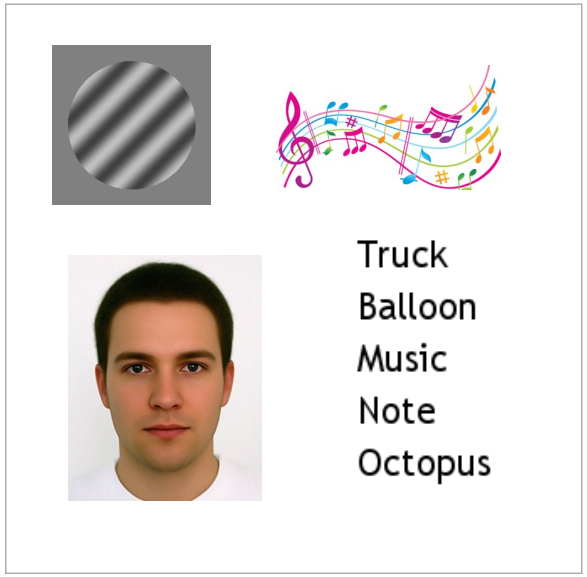
computation
(unobservable)

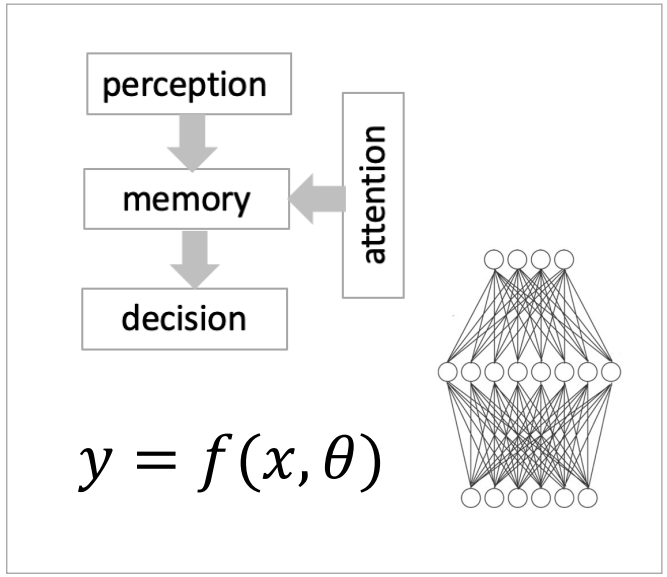
output
(measured)

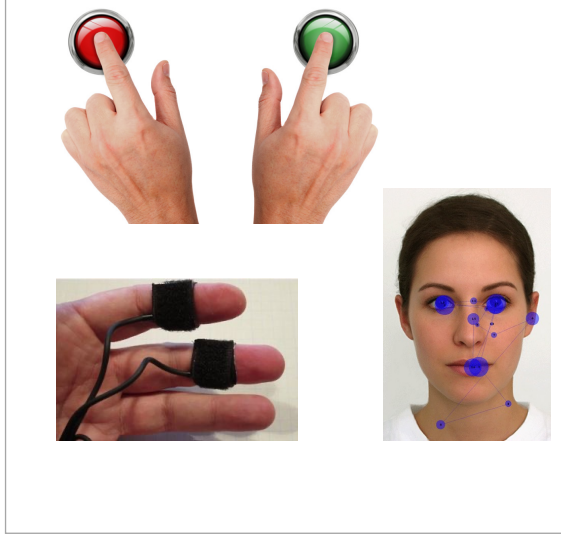
Interventions

The information processing approach









input
(known)

model
(inferred)

output
(measured)

What is a model?

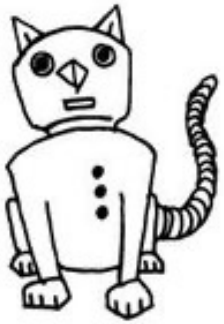
ground truth

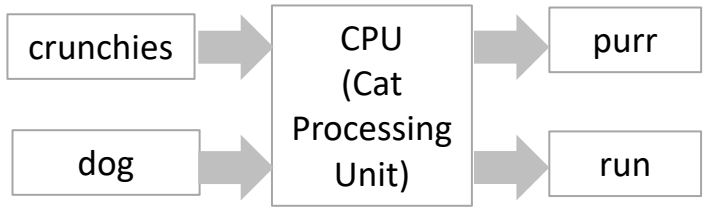
simulated (comp. neuro/cog. sci)

$$p(purr) = f(crunchies, hunger)$$

simulated
(same cat)

simulated
(robotics)





simulated (psychology)

"The best model of a cat is another cat, or preferably the same cat"

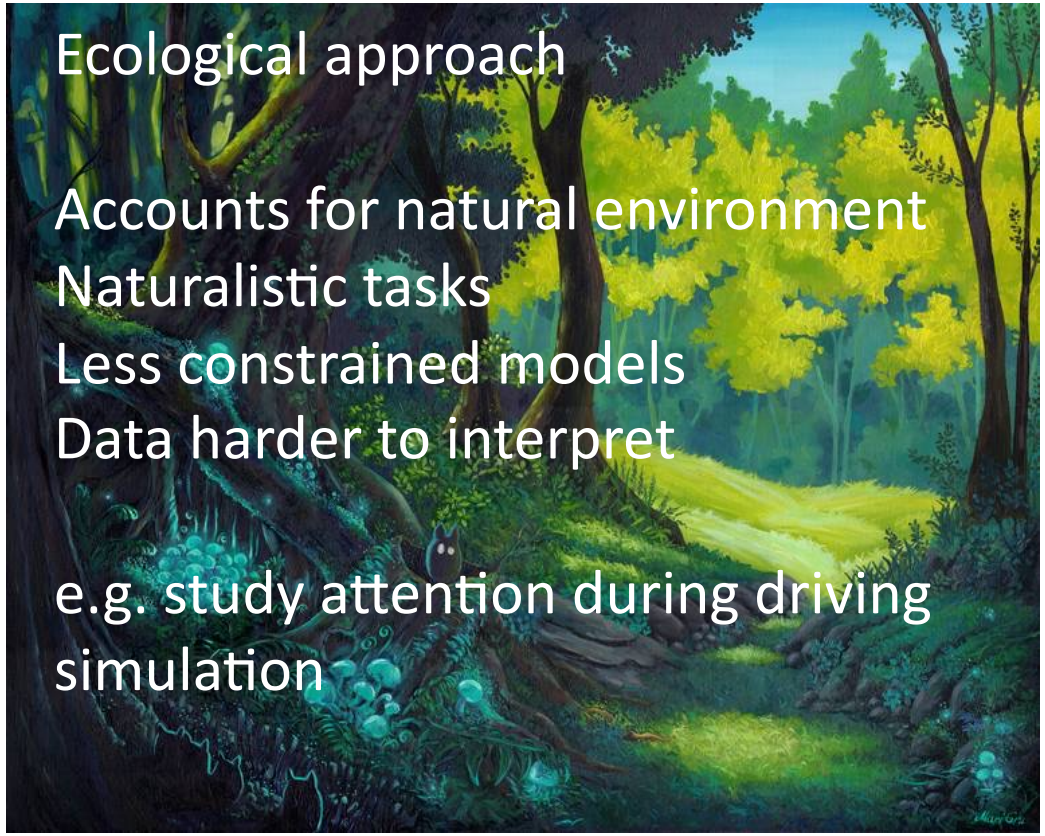
Norbert Weiner, Founder of Cybernetics

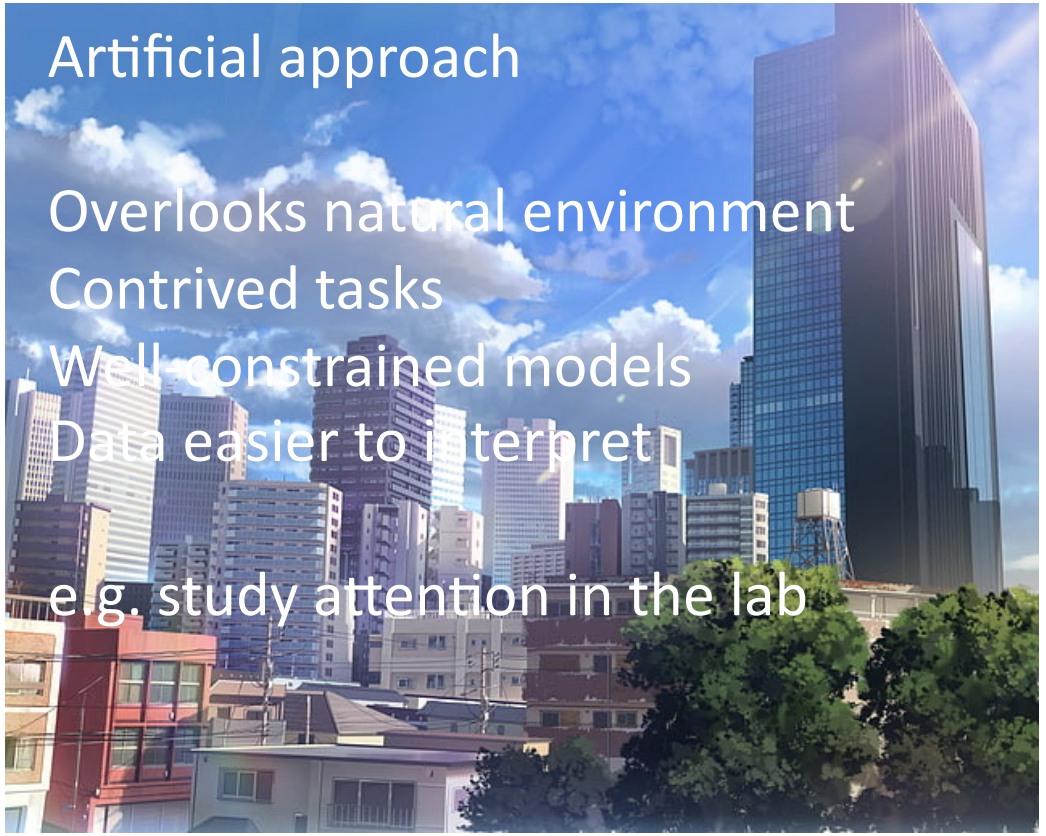
Levels of description

Computational theory	Representation and algorithm	Hardware implementation
What is the goal of the computation, why is it appropriate, and what is the logic of the strategy by which it can be carried out?	How can this computational theory be implemented? In particular, what is the representation for the input and output, and what is the algorithm for the transformation?	How can the representation and algorithm be realized physically?

Marr

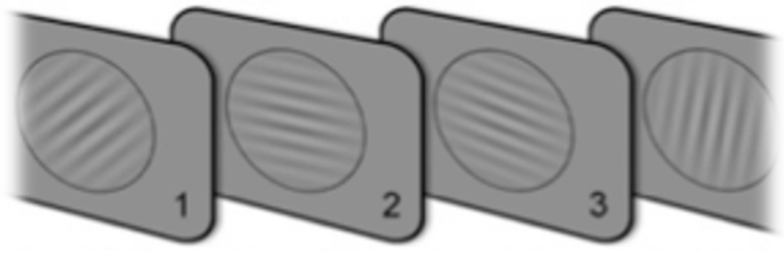
Approaches


Ecological approach
Accounts for natural environment
Naturalistic tasks
Less constrained models
Data harder to interpret
e.g. study attention during driving simulation


Artificial approach
Overlooks natural environment
Contrived tasks
Well-constrained models
Data easier to interpret
e.g. study attention in the lab

Model fitting: the basics

participants



stimuli

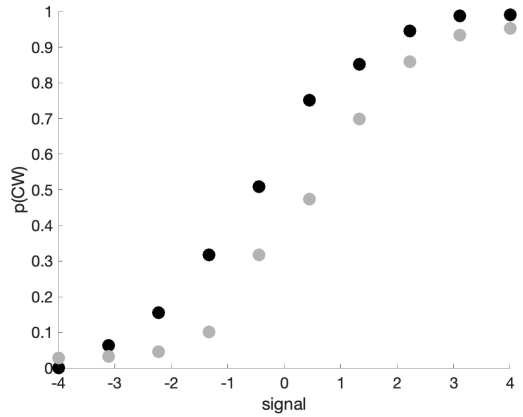
signals

$[0.2, -0.2, 0.4, 0.8, 0.8, -0.4]$

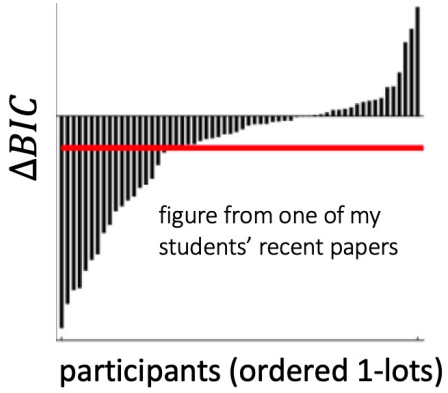
choices

$[0,1,1,0,0,0,1,1,0,0,0,1]$

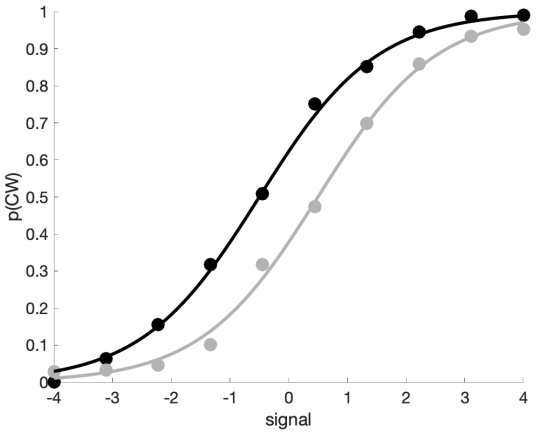
raw data



summaries



model fits



simulations

Model fitting: the basics

signals

$[0.2, -0.2, 0.4, 0.8, 0.8, -0.4]$

choices

$[0,1,1,0,0,0,1,1,0,0,0,1]$

raw data

model

signals

$[0.2, -0.2, 0.4, 0.8, 0.8, -0.4]$

choice probabilities

$[0.2, 0.9, 0.6, 0.3, 0.1, 0.2, 0.9]$

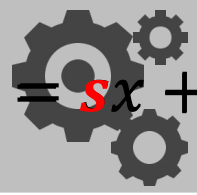
predicted data

Model fitting: the basics

$$\hat{y} = p(choice) = \frac{1}{1 + e^{x}}$$

$$x'$$

$$x' = sx + b$$



$$signal = x$$

Model fitting: the basics

$$\hat{y} = \frac{1}{1 + e^{x'}}$$

$$x'$$

$$x' = \boldsymbol{s}x + \boldsymbol{b}$$

$$signal = x$$

Ground truth choices (y)

$[0,1,1,0,0,0,1,1,0,0,0,1]$

Choice probabilities ($\hat{y}$)

$[0.2,0.9,0.6,0.3,0.1,0.2,0.9]$

$$-\left[\sum_{j=1}^{N} y_j \log(\hat{y}_j) + (1 - y_j)\log(1 - \hat{y}_j)\right]$$

Binary cross-entropy loss
(equivalent to $-\log(p(data|model))$

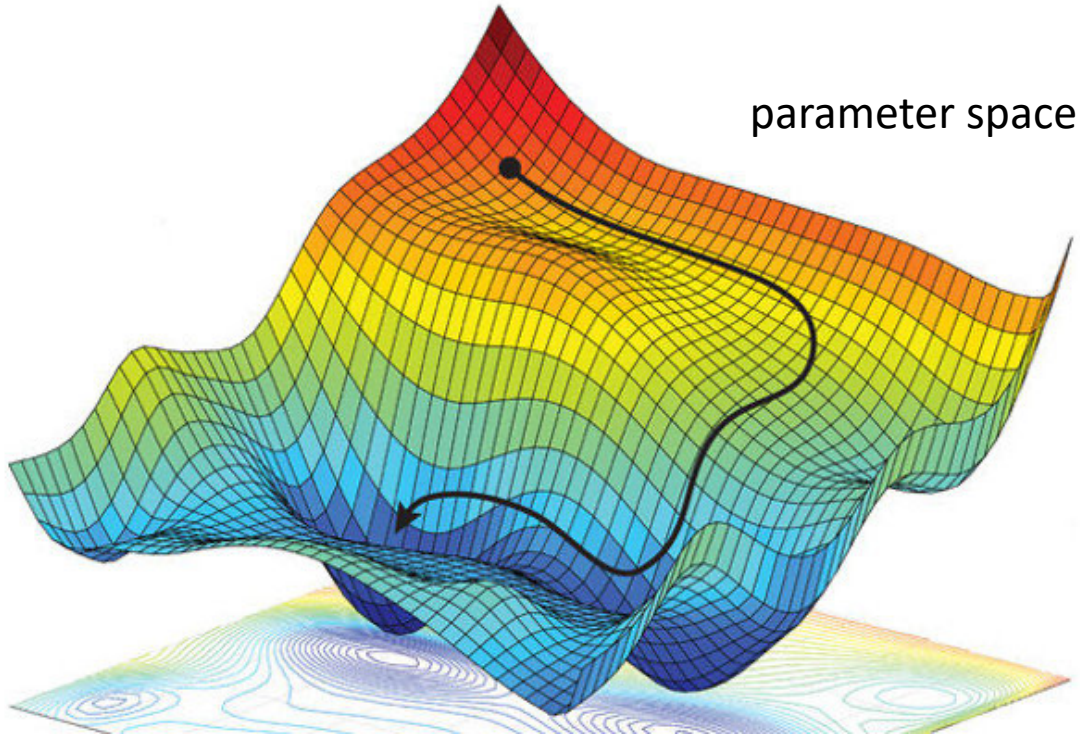
Model fitting: the basics

$$\hat{y} = \frac{1}{1 + e^{x'}}$$

$$x'$$

$$x' = \boldsymbol{s}x + \boldsymbol{b}$$

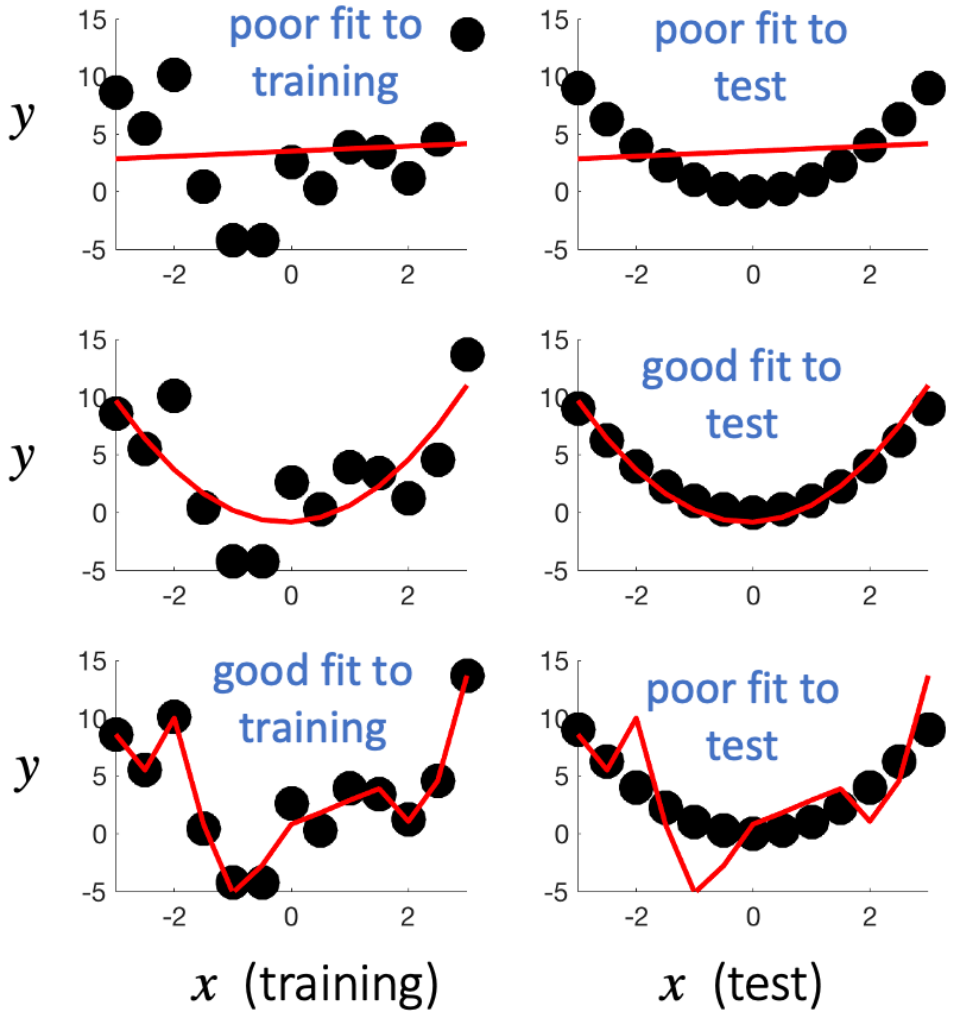
$$signal = x$$



bias b

slope s

Over and underfitting



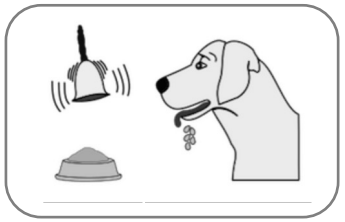
This model is **underfit** (high bias). It can't capture the training or test data.

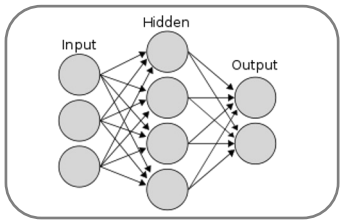
This model **fits just right**. It has some training error, but fits test well

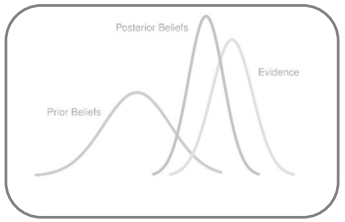
This model is **overfit** (high variance). It captures the training data perfectly, but fails at test

Simple example of overfitting with ground truth $y = x^2 + \mathcal{N}(0, \sigma)$

The search for a unified theory







1900 1920 1940 1960 1980 2000 2020

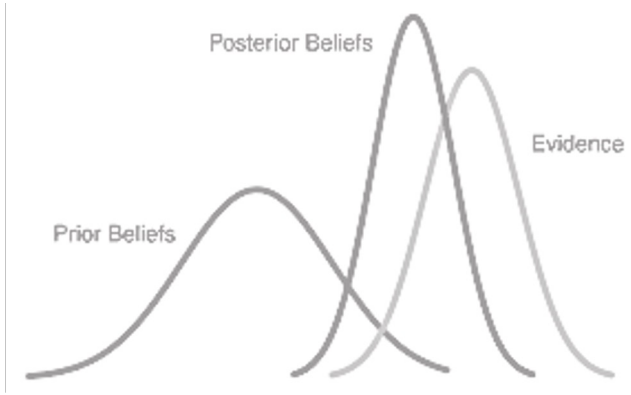
Pavlov Watson Skinner behaviourism

connectionism Hinton, McClelland DiCarlo

Bayesianism Tenebaum, Friston

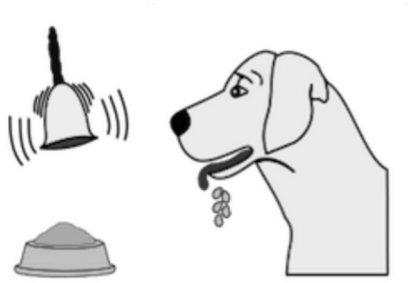
cognitivism Marr, Chomsky Posner, Shallice

Model classes

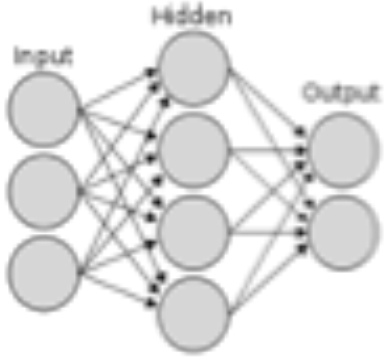


Bayesianism

cognitivism



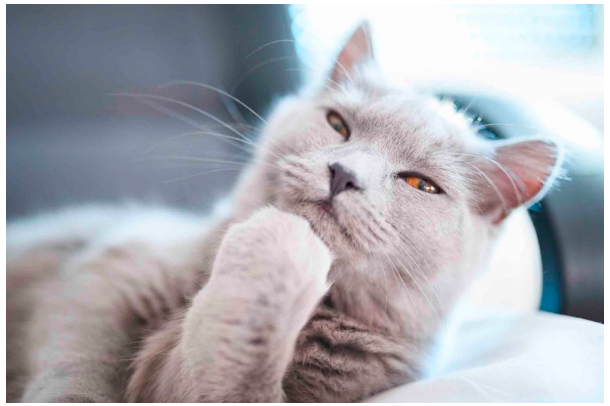
behaviourism



connectionism

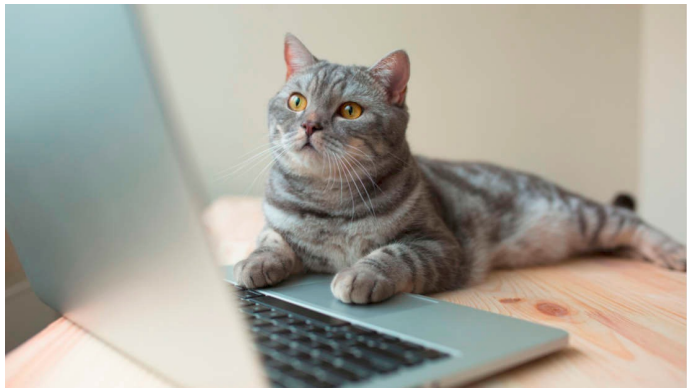
What nobody teaches you

You can read thousands of papers about how to compute model evidence and compare models (and this course will help!).



vague thoughts about cognition/computation

what actually happens here?



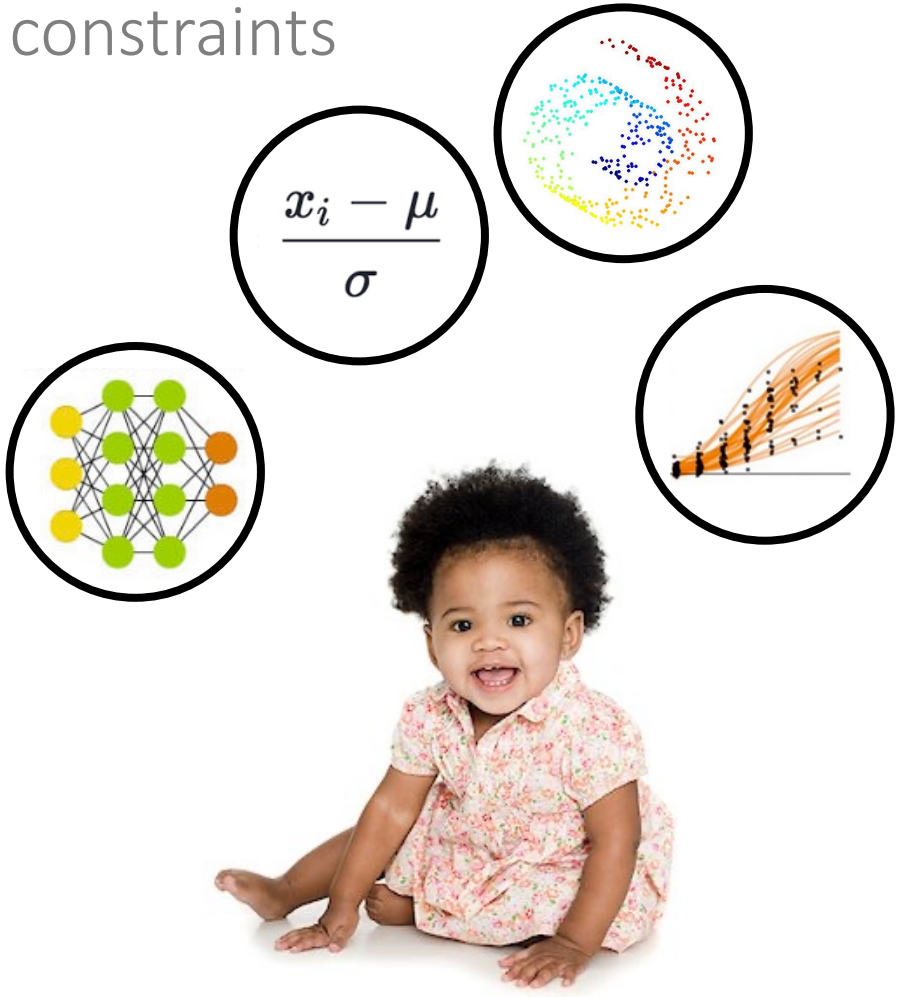
sit down to write model

But nobody ever tells you how to identify the space of models that you should consider in the first place.

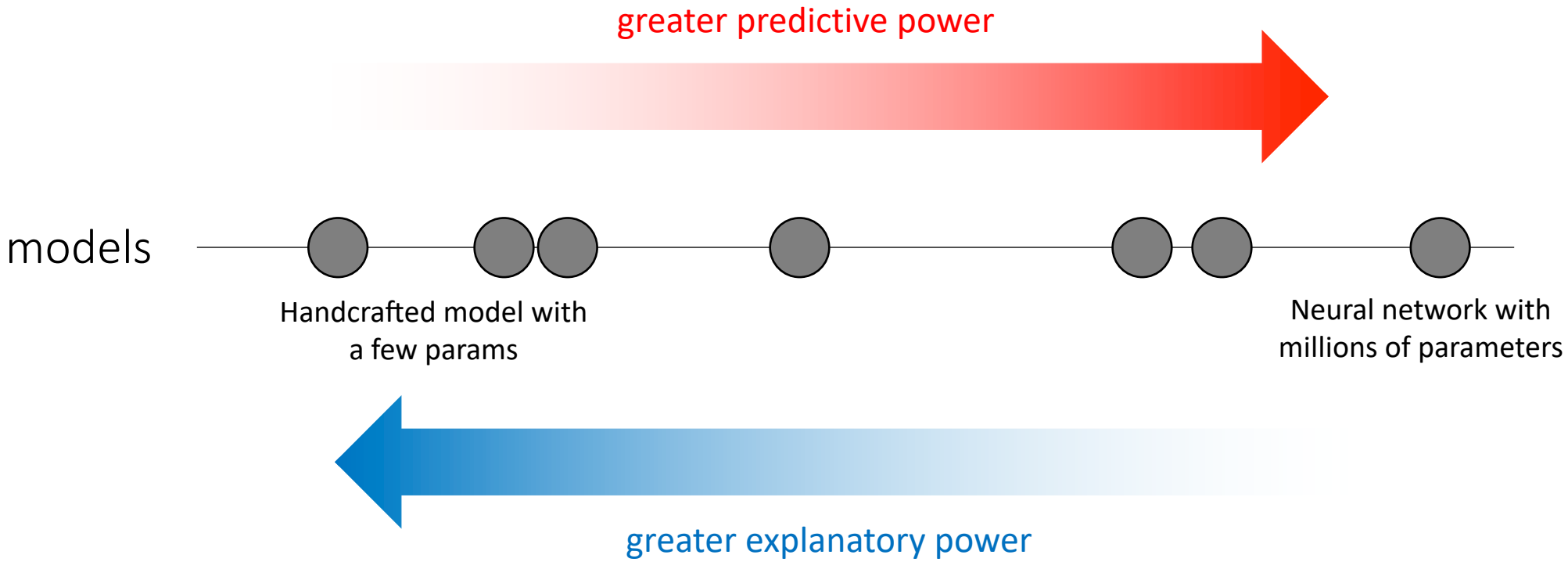
Inductive biases

prior beliefs

computational constraints

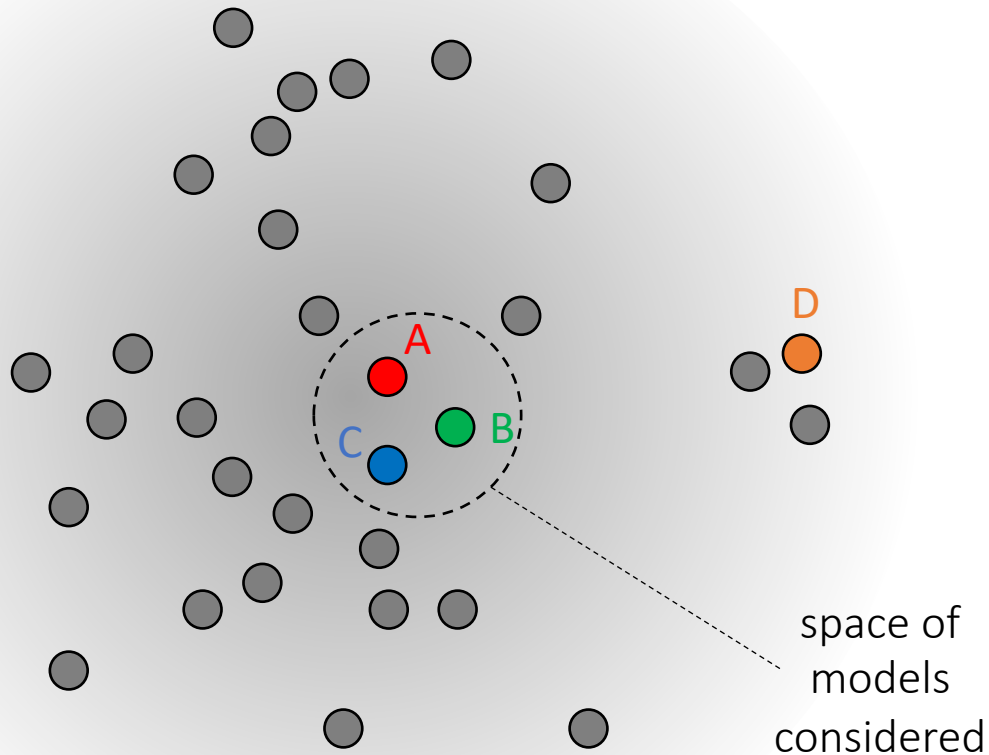


Predicting and explaining



Models are useful for predicting and explaining, and these two virtues typically trade off with complexity

Limitless models



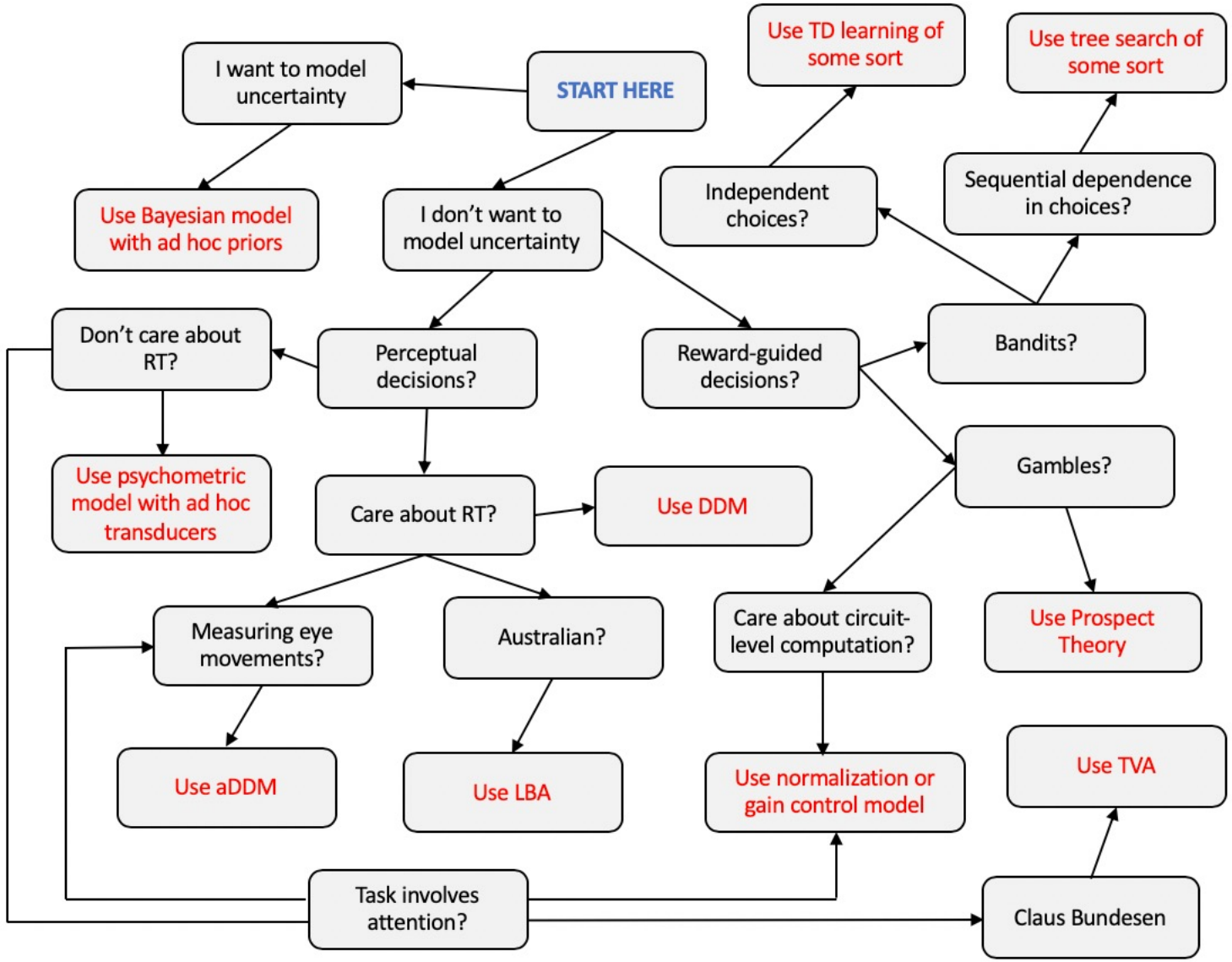
universe of models

The universe of models is theoretically limitless.

How do I know if my model space is sensible or not?

So, how do I know where to start?

Solution 1: use past models as theories

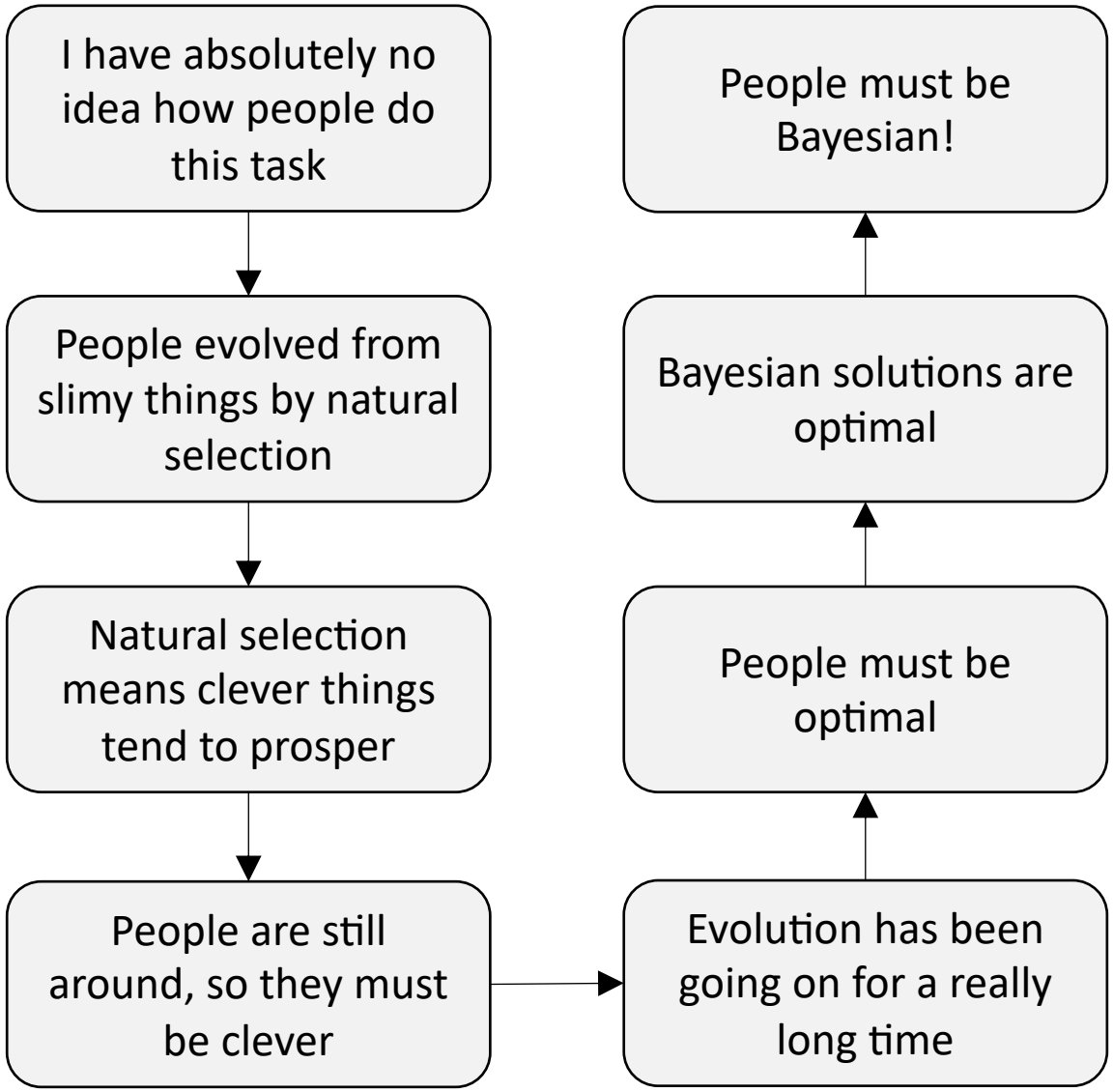


Science is an incremental endeavour.

So, many people use a sort of mental modelling cheat sheet that involves copying what others have done.

BUT sometimes this is done without thinking very hard about why (more on this later).

Solution 2: take a normative stance

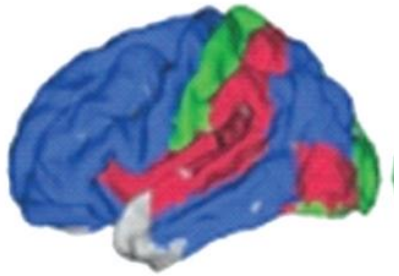


I built the task, so I know the generative model. I just need to assume that people optimally invert this model for inference.

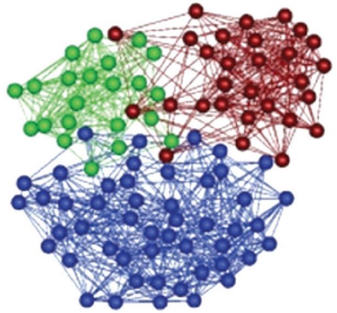
I don't even need any free parameters!

Problem solved!

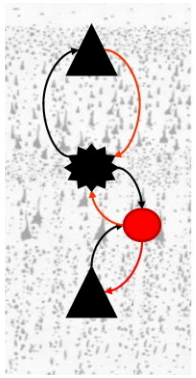
Solution 3: draw inspiration from neurobiology



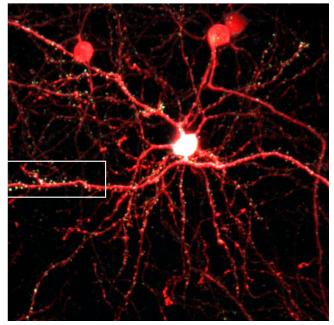
systems



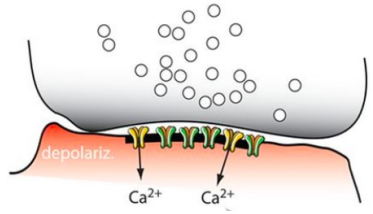
networks



circuits



cells

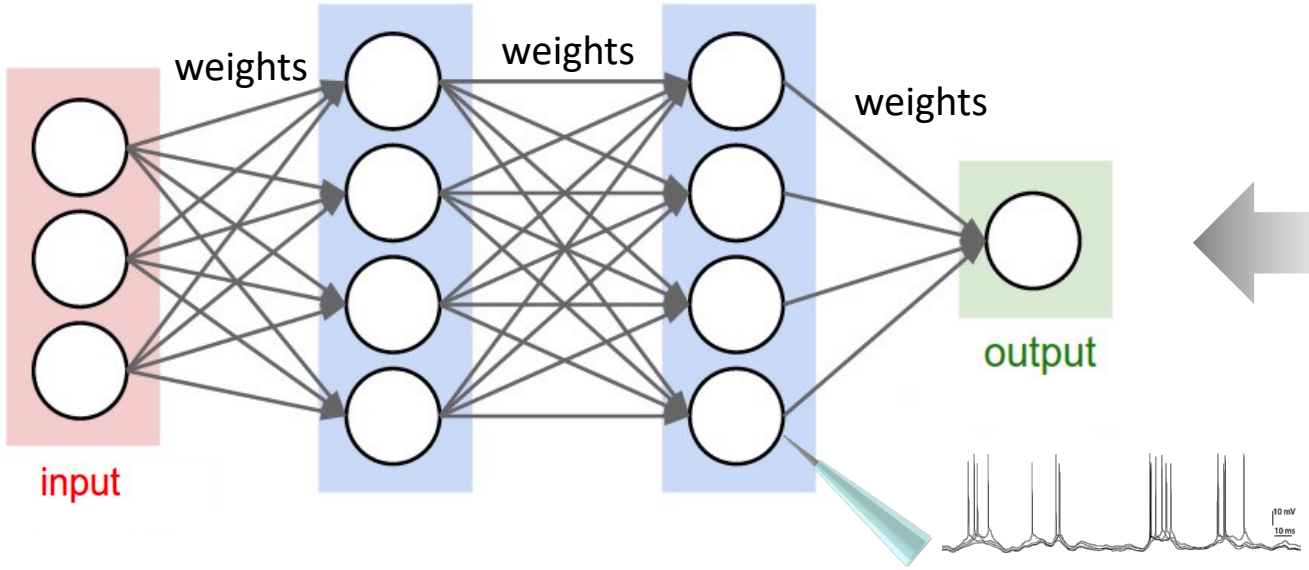


synapses

There are at least 50,000 neuroscientists out there. And they know some stuff (a little, at least).

We can draw inspiration from what we know about the brain to build our models. For example, tuning curves are Gaussian, neurons are mutually inhibitory, and exhibit recurrent excitation, coding is adaptive in time and space, etc, etc.

Solution 4: design a cost function instead



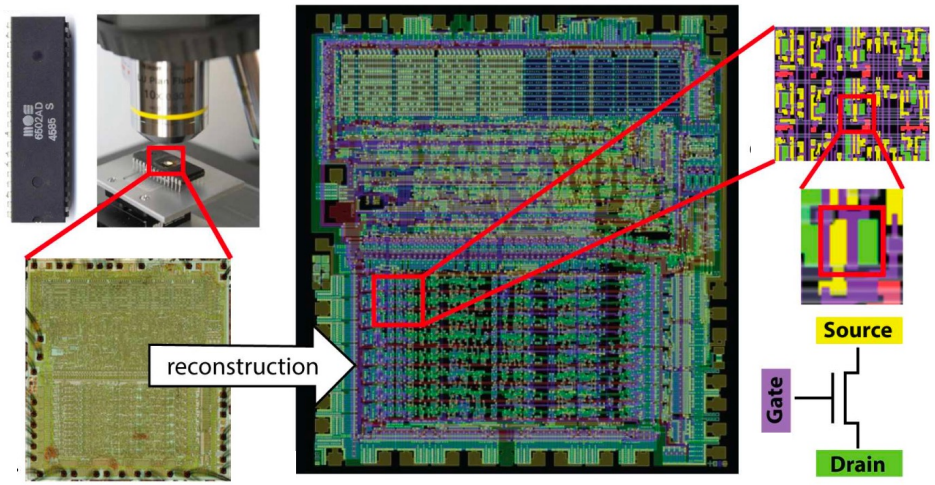
cost function:
minimise classification error
minimise reconstruction error
maximise external reward
+ intrinsic costs
etc

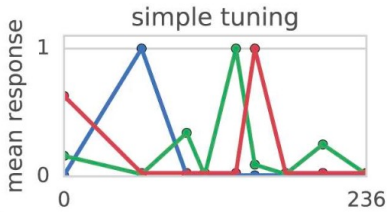
Why not forget hand-designing your model? Focus on the optimisation principle instead. If the network is powerful enough, it will learn the requisite set of computations through raw function approximation.

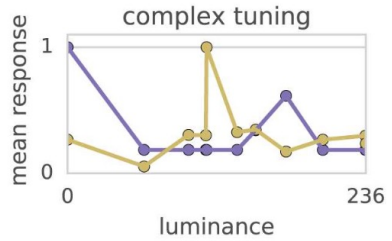
But….interpretability!

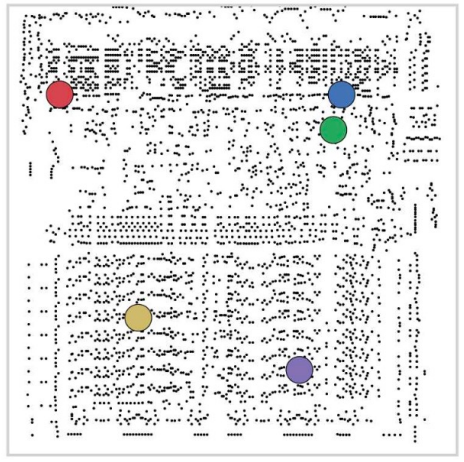
Marblestone et al

And we should be duly (but not unduly) sceptical…









Take a system that is fully understood because we built it (e.g. a microprocessor)

do cognitive neuroscience e.g. single cell recordings, brain imaging, lesion studies

Results reveals tuning curves, connectivity profiles, lesion-symptom maps and oscillatory activity just as in the real brain.

But we know that the interpretative logic applied to these phenomena is completely wrong!

Jonas & Kording 2017

Thanks