The applicability of Deep Neural Networks in Biomedical Science

Alastair Droop

RSE Leeds AI and machine learning workshop, 19/12/2018

Deep Learning

Deep Neural Networks (DNNs) are neural networks with multiple layers between the input and output layers

Deep Learning

- A set of techniques in Artificial Intelligence to *learn patterns in complex data*
- Appropriate where *we do not know the structure of the data* beforehand

A good conceptual fit for biological data

- We need to infer higher-order structural information from messy low-level data
- DNNs *seem to fit* with the kinds of data and problems we have

	Image data	-omics' data
Pixels per image	~10k pixels	~50k 'pixels'
Higher-order features	edges & objects	pathways & modules
Information per pixel	–	+++
Number of samples	+++	–

Map -omics data onto standard DNNs

- Each image is a biological sample, and each pixel a transcript expression value
- We still want to extract higher-level information

– We have orders of magnitude fewer data points
– We have (many) orders of magnitude more complexity to extract
+ We have a lot more meaning for each individual 'pixel'

Problems with DNNs #1

Biological systems are not well constrained

Well Constrained problems

AlphaGO

+ GO now a ‘solved’ problem, much like chess or connect 4

CASP13

+ AlphaFold a clear winner in the recent protein folding contest

Not well constrained problems

IBM Watson Cancer Diagnosis

– Much hype; now seems to have been an expensive failure

Uber, Tesla *etc*…

– Not looking good yet.

Chess, GO, protein folding are *highly constrained* problems

Cancer diagnosis, and driving *are not*

Biological systems are extremely complex (and complicated)

- Simplistic representation are usually inappropriate
- We don't know the features and behaviours we're looking for
- We have many more variables than we have samples (always overfitting!)

We need

- More data
- Appropriate abstractions to reduce the state space

Problems with DNNs #2

We do not control what features of the data the DNN decides to learn

STOP	100%	<3%	40%
Speed limit 45	0%	86%	27%

Eykholt *et al* (2018) “Robust Physical-World Attacks on Deep Learning Visual Classification” in *IEEE CVPR 2018*

Processing ResearchOne Data

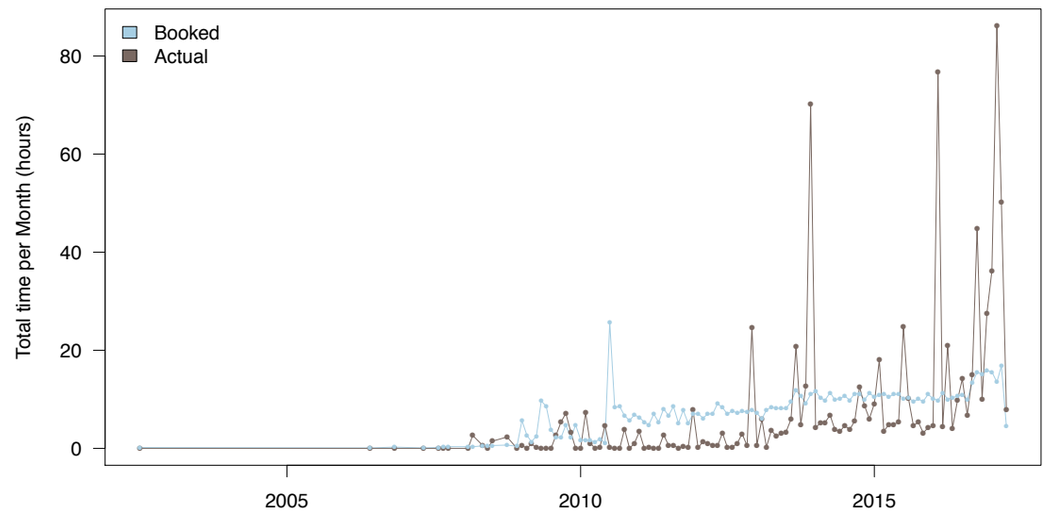
Technical errors & data artefacts

- Youngest patient is –12.5 years (oldest is 111.08)
- Illegal characters

Developing computational tools to identify and isolate artefacts

Data preprocessing to reduce unnecessary noise

file data	text representation
30 30 30 30 31 7C 4B 67 2F 6D FD 7C 32 7C 32 32 2E 35 7C 2D 31	00001|Kg/mý|2|22.5|-1
35 39 39 39 39 7C 4B 67 2F 6D FD 7C 32 7C 32 30 2E 30 7C 32 35	59999|Kg/mý|2|20.0|25
7C 32 36 2E 31 7C 4B 67 2F 6D FD 7C 32 7C 32 32 2E 35 7C 30 2E	|26.1|Kg/mý|2|22.5|0.
7C 32 36 2E 31 7C 4B 67 2F 6D FD 7C 32 7C 32 32 2E 35 7C 30 2E	|26.1|Kg/mý|2|22.5|0.
7C 32 33 2E 39 7C 4B 67 2F 6D FD 7C 32 7C 32 32 2E 35 7C 30 2E	|23.9|Kg/mý|2|22.5|0.
36 39 33 38 37 7C 4B 67 2F 6D FD 7C 32 7C 7C 7C 0D 0A 35 33 36	69387|Kg/mý|2||\r\n536
36 39 33 38 37 7C 4B 67 2F 6D FD 7C 32 7C 7C 7C 0D 0A 35 33 36	69387|Kg/mý|2||\r\n536
7C 33 30 2E 39 7C 4B 67 2F 6D FD 7C 32 7C 32 32 2E 35 7C 30 2E	|30.9|Kg/mý|2|22.5|0.

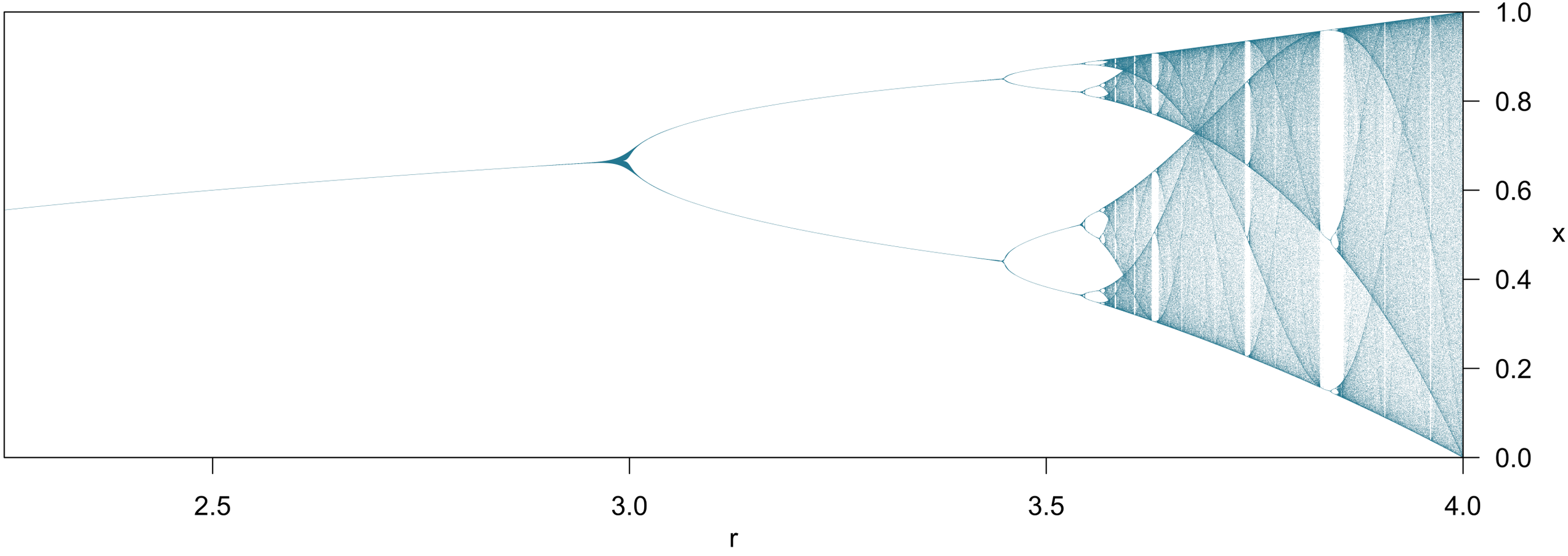


The Logistic Map

Very simple iterated equation yielding complex behaviour

$$x_{n+1} = r x_n (1 - x_n)$$

For each value of *r* in the range (1,4) count the number of stable values of *x*



Many biological systems exhibit complex, dynamical behaviour

- Information processing in metabolic networks
- Cell communication
- Cell cycle control feedback

Many biological phenotypes are emergent

Can DNNs capture these complex behaviours?

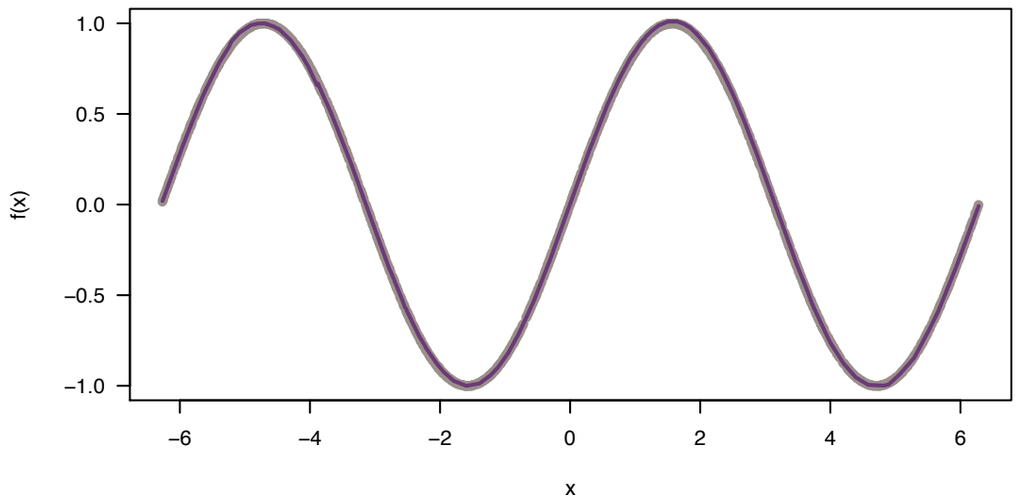
Health Data Research UK

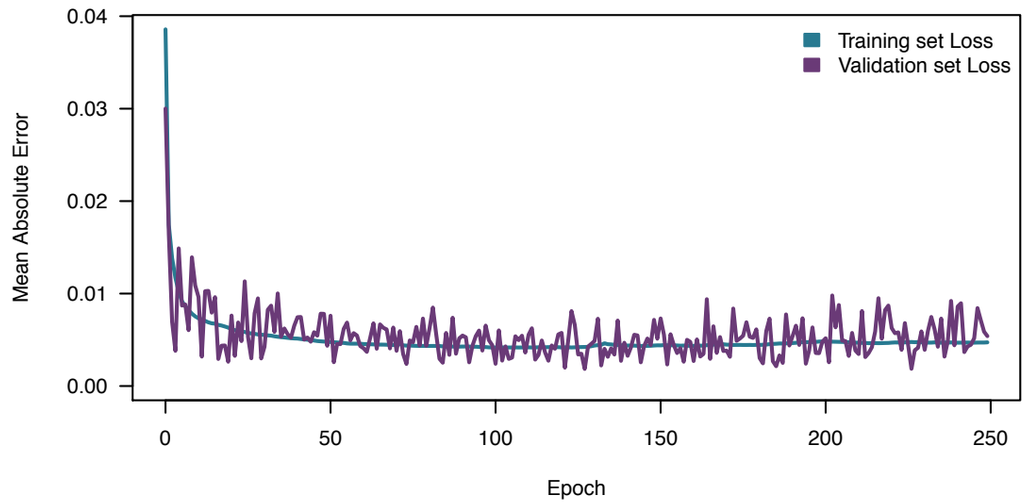
UNIVERSITY OF LEEDS

Model Predictions

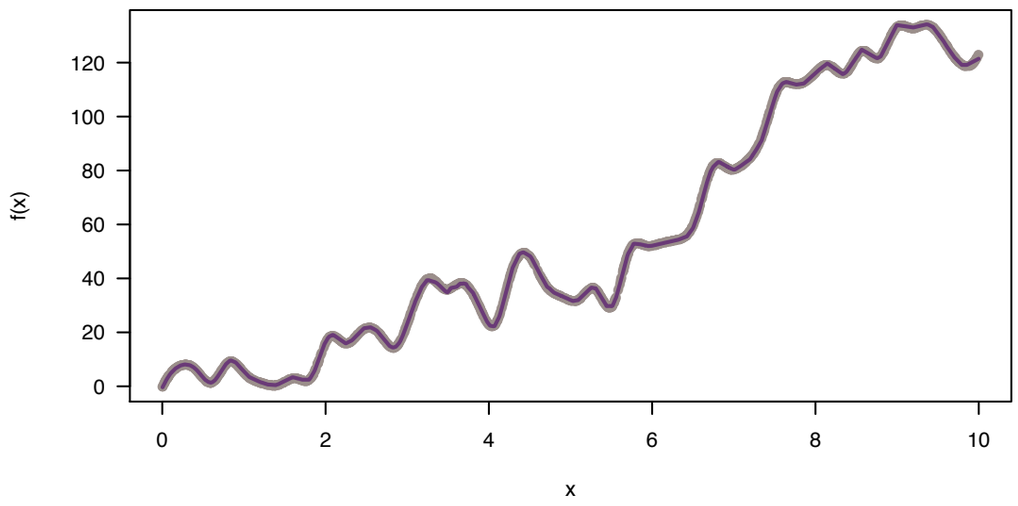
Training Performance

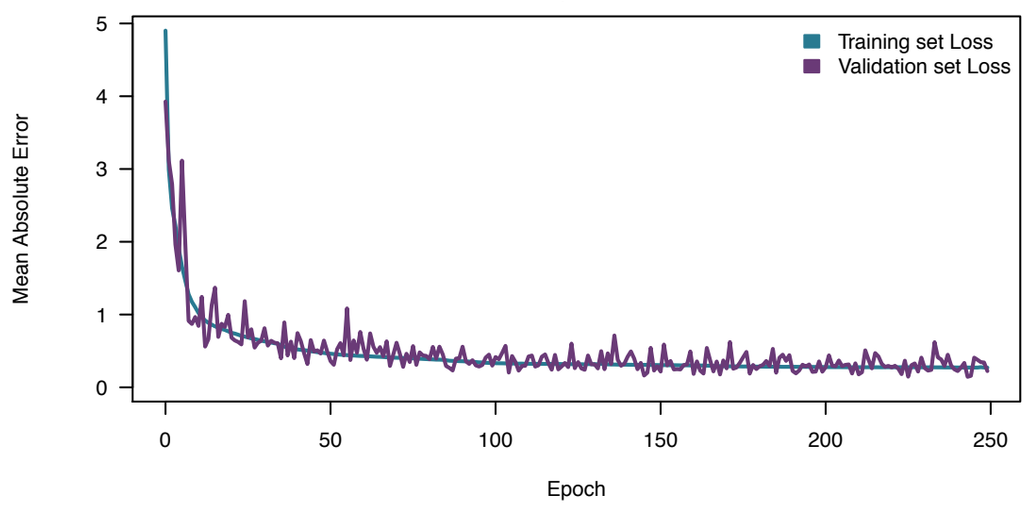
Simple



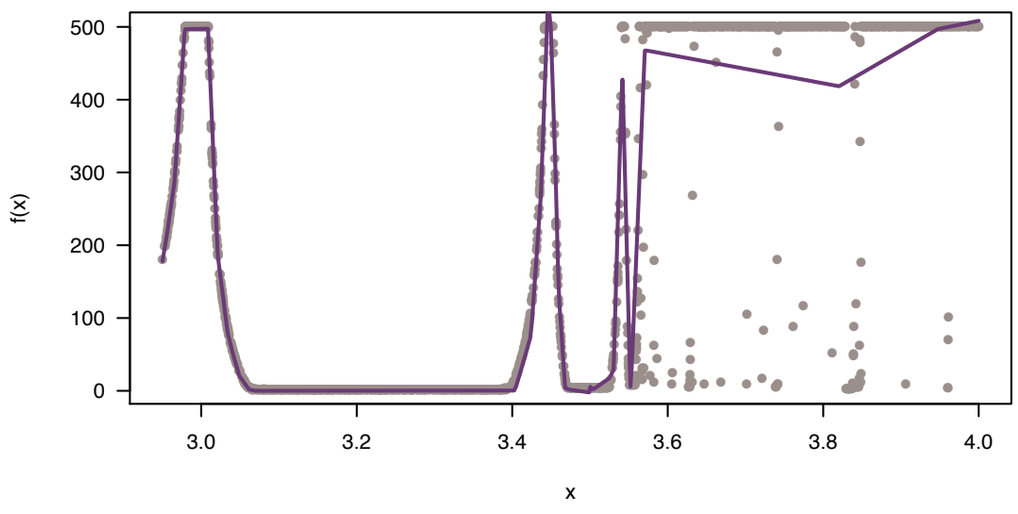


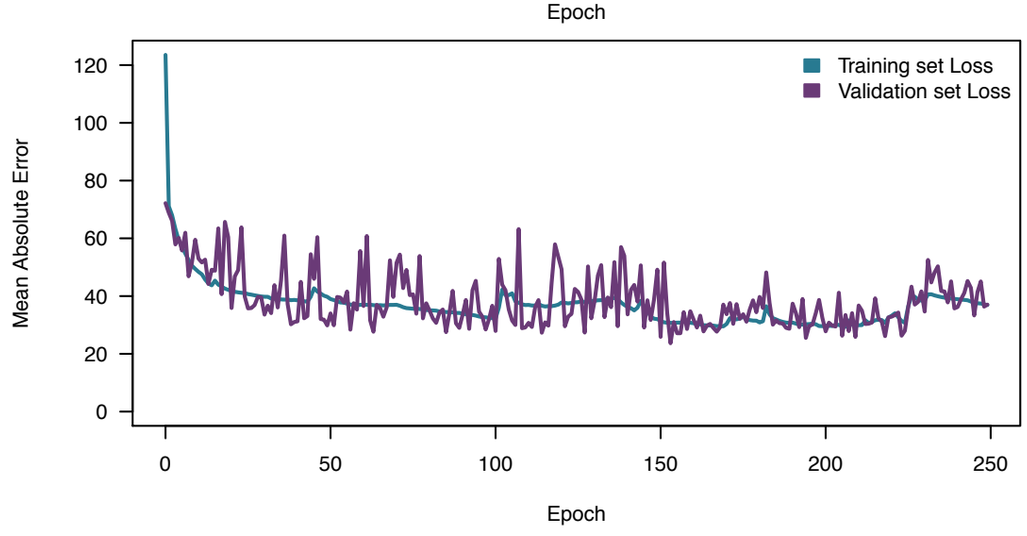
Complicated





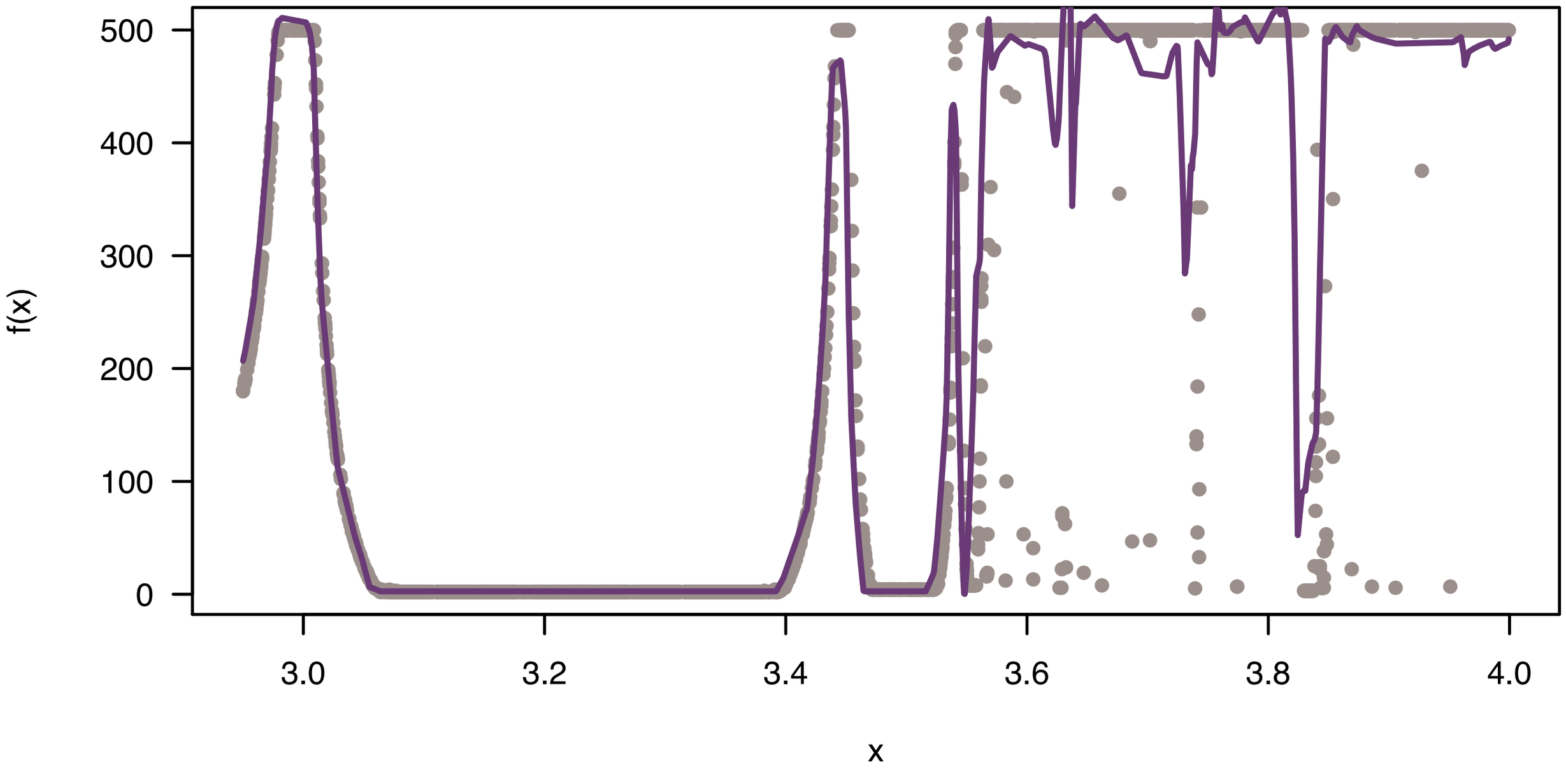
Complex





Data from fully-connected DNN with 2 hidden layers (each with 32 neurones) using 10^6 training points

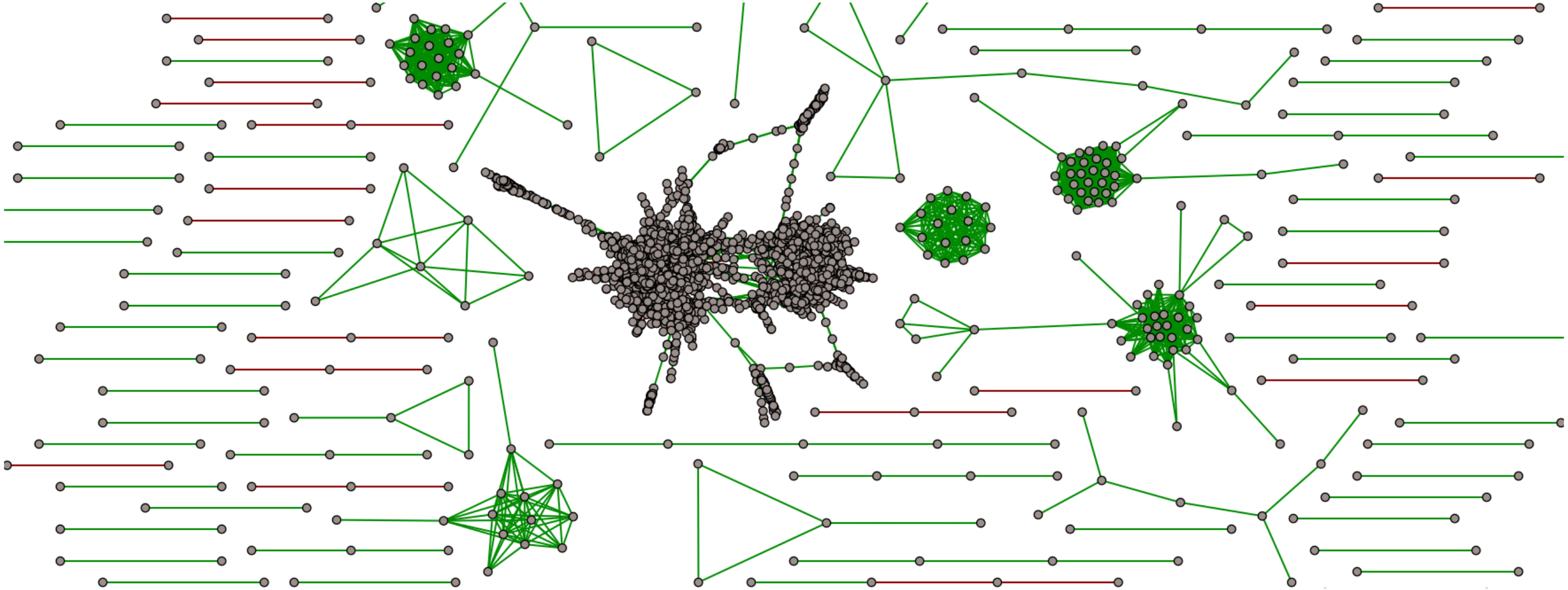
Chaotic behaviour seems to degrade entire model learnability



Appropriate DNN Topologies

Build DNN hidden layers using biological topologies

Derived from transcriptomic data correlation analyses



Summary

DNNs are not a magic bullet

DNNs work best on well-constrained problems

- More data; appropriate simplification of the system

Artefacts in the data can throw off the learning

- We are developing methods to tidy ResearchOne data

DNNs not good at learning dynamical biological phenomena

- We are developing methods to allow learning on transcriptomic data

Acknowledgements

Lucy Stead

- Transcriptomic network data & analyses

Klaus Witte & Mike Drozd

- SystmOne data & access

Prof David Westhead's group

Addendum: What's Wrong with AI?

"It's more complicated than that"

Biology is essentially unconstrained

Current AI fails badly in these cases (fails to *recognise* this as a sofa)

Maximally accurate | Maximally specific

big cat	3.79080
feline	3.16167
jaguar	2.50780
leopard	1.57470
carnivore	0.87198

CNN took 0.383 seconds.

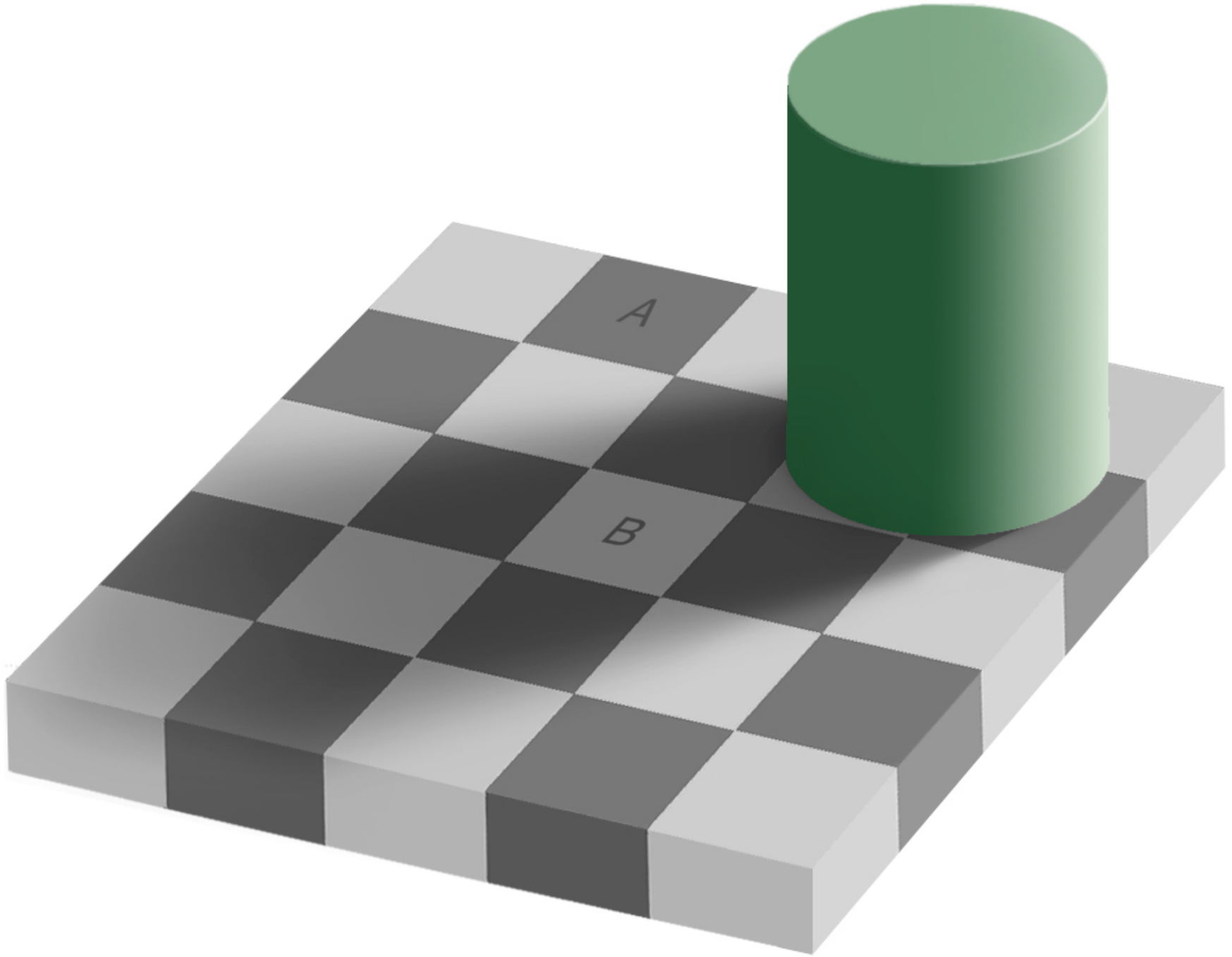
http://demo.caffe.berkeleyvision.org

Concept & sofa image from https://rocknrollnerd.github.io/ml/2015/05/27/leopard-sofa.html

The model is only as good as its data

In biology, we don’t have a model (yet), and by definition it is incomplete

It gets worse…



Edward H. Adelson (1995) Dept. of Brain and Cognitive Sciences MIT (http://persci.mit.edu/gallery/checkershadow)