What can the working mathematician expect from deep learning?

Geordie Williamson, University of Sydney Mathematical Research Institute

University of Sydney Colloquium, November 2022

Theorem: *There are infinitely many prime numbers.*

Proof:

1) *Assume there are finitely many:* $p_1, p_2, \ldots, p_n$.

2) *Consider* $p_1 p_2 \ldots p_n + 1$.

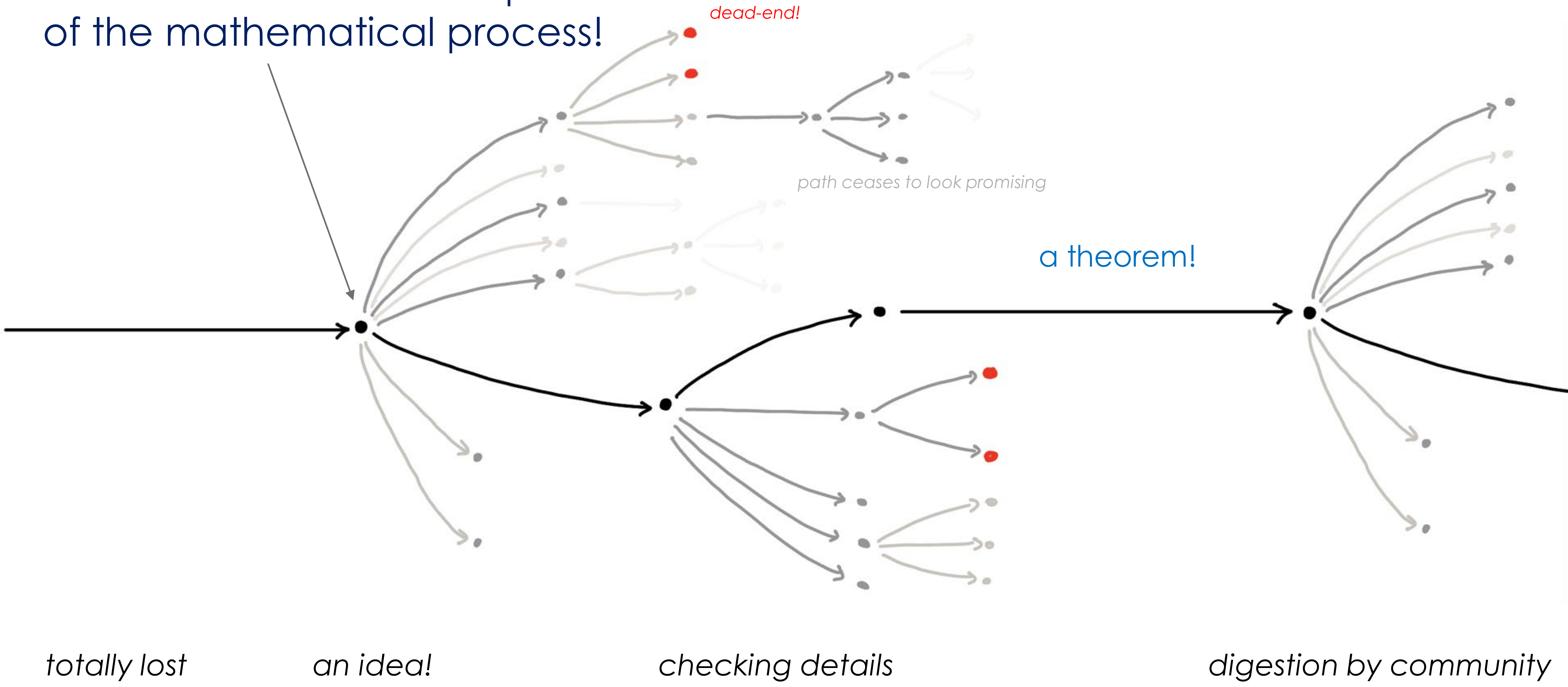
3) *You know the rest…*

"There is the monastic, introverted period, where we are just contemplating the ocean of our ignorance; but then suddenly something happens…the monk becomes busy and excited, in a hurry to look more closely at the details."

—Claire Voisin, *How to make a portrait of a bird.*

The Development of Ideas

We are interested in this part of the mathematical process!



Plan:

1) *Crash course in deep learning*

2) *Simple examples in mathematics*

3) *Myths, advice and scale*

4) *Some examples in (pure) mathematics research.*

Disclaimer:

1) I am a pure mathematician, interested (mostly) in representation theory, algebraic geometry… I won't discuss deep learning in applied math or mathematical questions raised by deep learning.

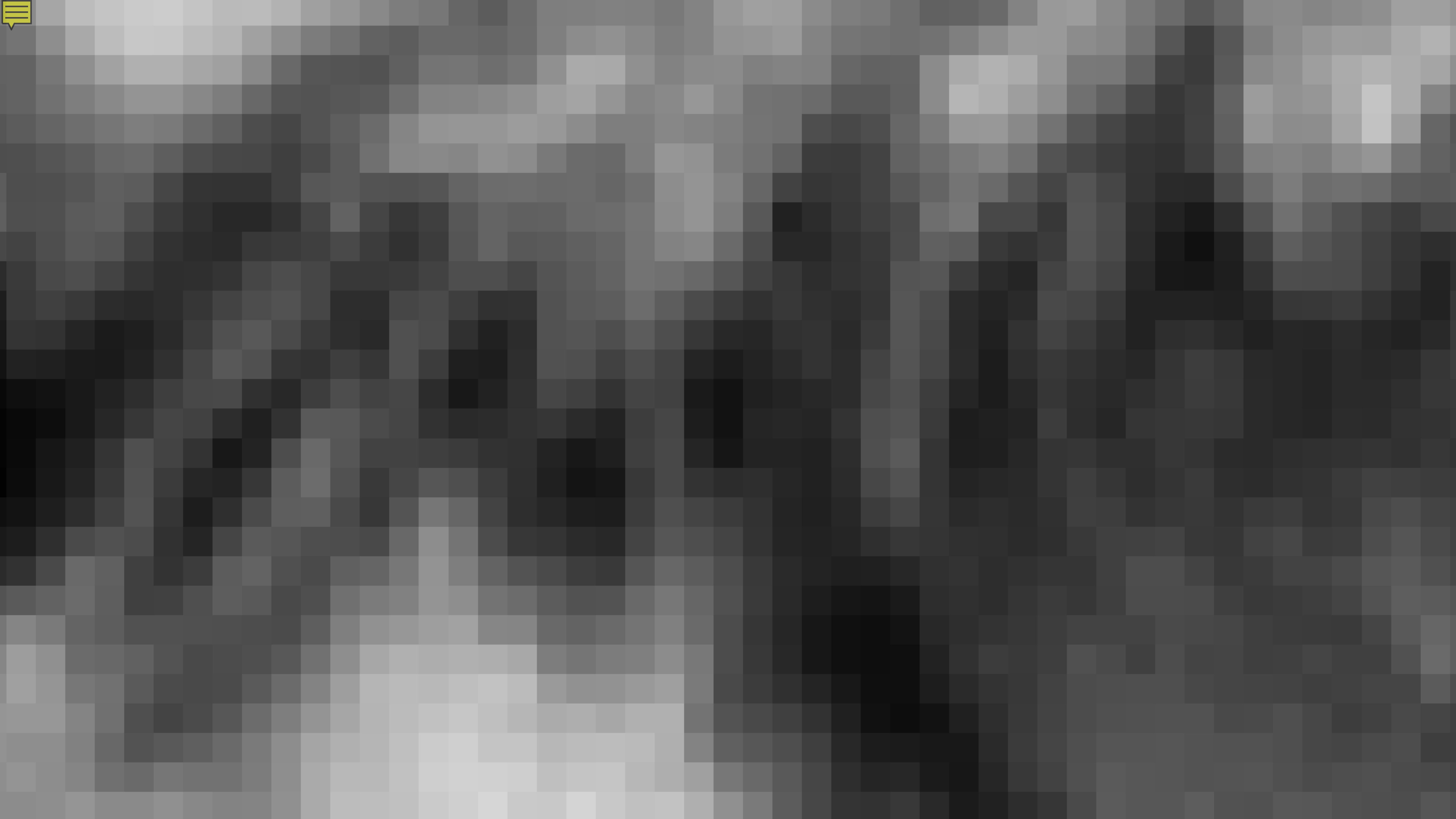
2) I have been working with DeepMind (of AlphaGo fame) tor two years. We are interested in potential interactions of AI and mathematics. This is a two-way bridge.

3) I have spent two years engaging with machine learning. I know the basics but am far from an expert.

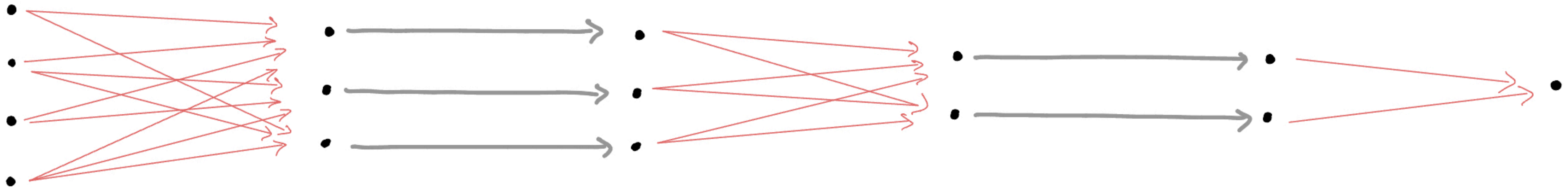
4) All opinions are my own.

A crash course in deep learning.

Amélie Rec/Pexels



092	101	110	118	126	122	107	096	077	070	083	094	107	133	148	150	150	156	152	140	126	126	135	131	126	116	113	078	072	078	103	123	151	152	137	115	087	070	057	053	065	096	150
079	085	092	103	106	094	078	072	071	063	074	090	112	135	137	134	145	140	121	107	105	125	150	146	123	113	098	060	059	067	098	109	122	129	114	083	071	061	052	049	062	089	127
078	079	085	095	092	070	052	051	051	063	087	090	083	082	085	100	108	109	106	107	102	112	141	148	130	102	066	054	057	067	087	100	116	106	085	069	084	070	051	042	041	075	107
079	080	091	094	077	059	048	040	040	050	069	098	069	055	063	087	100	096	097	106	108	117	139	148	124	088	033	044	055	064	074	110	119	072	072	055	086	075	045	034	025	046	084
067	078	089	085	065	050	044	042	056	059	063	073	060	049	060	085	100	088	090	096	101	116	130	134	106	076	040	040	050	057	075	096	092	057	051	059	085	074	043	026	016	032	064
058	073	078	067	053	046	047	052	070	077	070	056	056	058	066	078	078	069	084	092	099	115	112	104	086	066	055	045	051	055	084	086	060	043	037	067	079	068	038	024	023	030	051
057	064	057	044	041	045	054	067	077	082	070	045	045	070	074	062	049	050	082	090	093	108	093	075	060	049	056	048	045	053	086	076	043	036	042	073	069	056	034	034	034	032	038
052	050	037	027	031	042	060	080	088	074	058	046	041	080	075	048	033	044	082	085	078	092	078	055	040	038	049	051	042	058	085	073	042	032	050	067	058	047	034	046	048	041	033
034	032	027	026	034	048	069	088	080	053	050	062	054	080	060	031	029	046	081	079	065	077	067	037	028	035	043	051	044	067	084	070	043	028	046	056	050	043	039	052	061	055	041
015	018	025	034	046	062	072	073	048	040	061	077	066	070	044	024	034	048	070	064	051	068	065	026	018	032	036	040	044	071	080	061	037	028	039	054	046	041	047	053	060	056	042
009	013	025	039	055	069	065	045	032	050	089	088	075	069	050	042	044	050	054	043	035	061	070	029	018	037	039	037	051	078	083	058	029	032	035	060	045	038	053	055	056	054	042
010	022	032	052	079	067	051	024	035	077	105	086	066	066	063	059	050	048	035	024	031	063	072	037	022	035	035	033	045	079	090	052	030	031	039	052	054	047	047	055	052	043	042
012	024	036	056	081	059	038	035	055	092	108	081	057	070	086	079	060	047	032	020	023	056	072	051	044	046	038	033	043	071	082	052	036	039	041	052	062	054	046	052	058	046	044
019	030	045	064	083	053	029	047	074	094	095	073	055	084	118	102	067	046	039	030	029	060	079	068	059	050	036	032	039	060	070	053	042	045	042	048	063	061	051	055	057	051	058
029	047	074	081	076	049	037	069	090	086	078	077	073	104	141	116	076	055	052	043	040	068	086	078	068	051	037	034	037	047	055	052	048	047	043	047	057	064	065	067	056	054	066
050	073	105	096	064	051	060	092	099	080	074	096	101	119	147	126	099	083	074	061	057	085	103	092	077	057	042	035	031	032	039	048	051	045	049	053	053	064	077	077	067	056	059
090	098	107	094	064	065	079	093	089	073	079	112	123	129	148	142	115	108	092	082	083	105	119	102	080	058	041	029	022	020	026	045	048	045	053	059	055	063	076	076	075	063	057
133	122	100	093	081	081	080	079	077	075	094	127	145	151	158	162	134	133	105	099	106	116	128	106	076	054	038	023	016	014	019	040	046	051	055	061	067	067	067	074	072	070	063
157	144	107	104	097	083	073	076	079	088	113	142	167	176	173	176	174	168	125	117	124	127	140	120	076	056	040	023	016	014	015	034	049	060	057	062	080	074	064	077	068	069	063
160	149	119	113	090	075	073	075	090	107	131	158	181	186	184	190	194	187	154	145	146	153	159	122	073	060	048	036	024	015	015	029	042	051	056	066	076	078	079	079	072	069	068
151	151	132	112	079	068	075	087	108	126	148	166	180	187	193	198	191	182	171	174	167	176	176	115	083	073	064	050	032	017	013	018	032	046	057	067	076	080	081	082	082	073	074
142	142	129	104	083	089	095	106	122	141	165	176	181	191	203	207	196	196	183	187	188	186	175	131	096	087	078	063	040	028	027	024	024	042	059	069	079	085	085	086	083	082	082
148	141	133	112	106	118	114	114	124	142	170	184	184	193	206	209	208	206	191	196	197	183	174	160	119	105	089	070	046	037	039	027	023	038	056	066	080	090	087	086	083	084	085



$$\phi : \mathbb{R}^{10^4} \xrightarrow[\text{linear}]{} \mathbb{R}^{10^3} \xrightarrow[\text{ReLU}]{} \mathbb{R}^{10^3} \xrightarrow[\text{linear}]{} \mathbb{R}^{10^2} \xrightarrow[\text{ReLU}]{} \mathbb{R}^{10^2} \xrightarrow[\text{linear}]{} \mathbb{R}$$

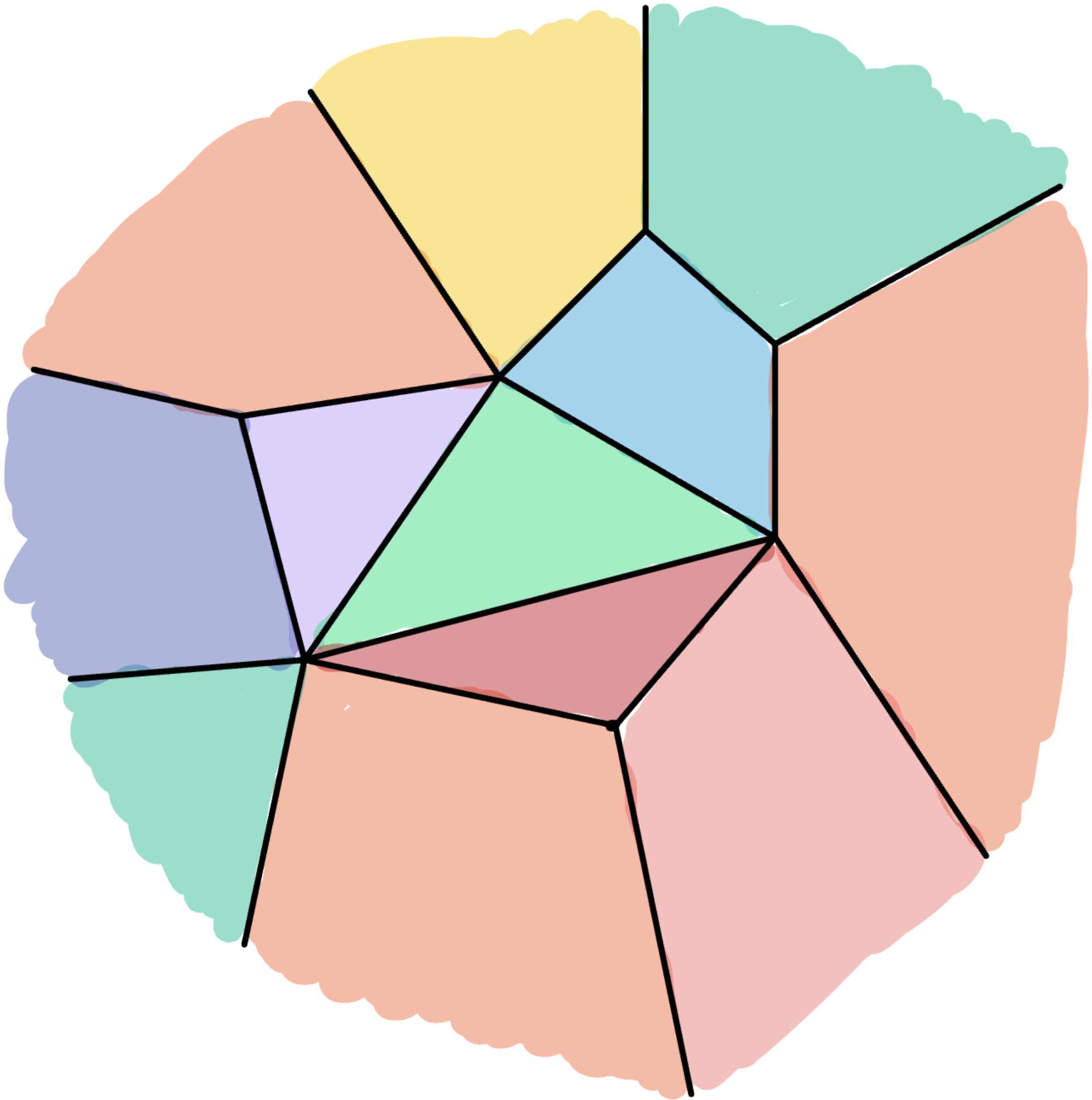
ReLU = coordinatewise max with 0 ("non-linearity")

Trained to approximate a target function $\tilde{\phi}$ via gradient descent on

$$\text{Loss} = \sum \ell(\tilde{\phi}(x), \phi(x)).$$

e.g. mean squared error
or cross entropy

Simple examples in mathematics

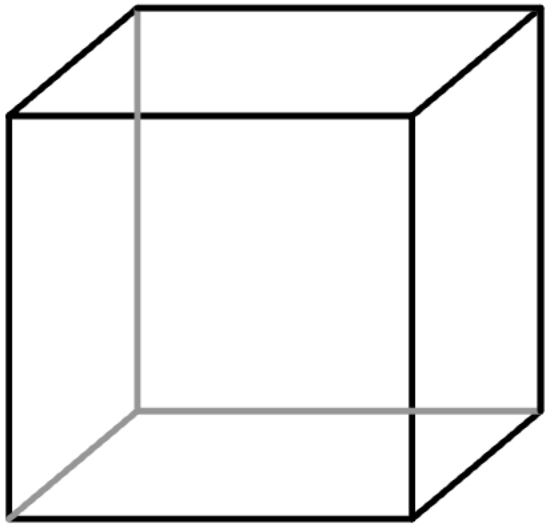


Geometric picture of training

(But try to imagine this happening in 10000 dimensions, rather than 2!)

Deep learning works best when:

1) *Input dimension is high*
2) *Function is on unit cube*
3) *Coordinates have low symbolic content*

$\phi : \mathbb{R}^{10^4} \supset$  $\longrightarrow \mathbb{R}$

In some settings, overcomes "curse of dimensionality".

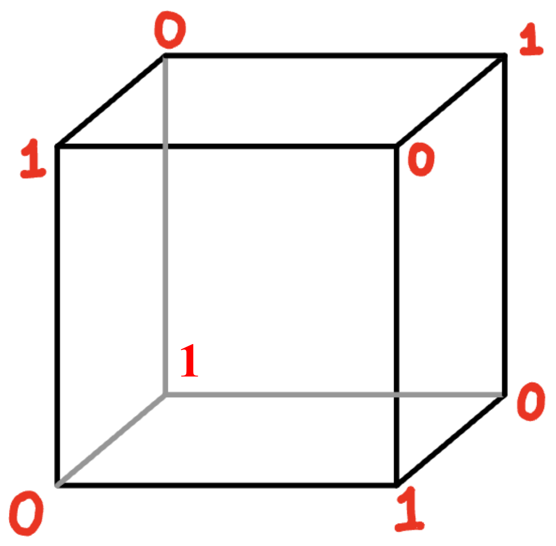
Example: "parity bit" $\{0,1\}^{1000} \longrightarrow \{0,1\}$

$$(x_0, \dots, x_{999}) \longmapsto \sum x_i \mod 2$$

Very noise sensitive.

Difficult to learn!

Many examples in number theory are like this one.



Example: (due to Joel Gibson)

$x = (x_1, \dots, x_n)$ a permutation of n.

$\mathcal{R}(x) = \{ i \mid x_i > x_{i+1} \}$ "right descent set"

$\mathcal{L}(x) = \{ i \mid i \text{ occurs to the right of } i+1 \text{ in } (x_1, \dots, x_n) \}$

"left descent set"

$\mathcal{R}(x) = \mathcal{L}(x^{-1}) \Rightarrow$ symmetrical concepts.

$x = (x_1, \dots, x_n)$ a permutation of n.

$\mathcal{R}(x) = \{ i \mid x_i > x_{i+1} \}$ "right descent set"

$\mathcal{L}(x) = \{ i \mid i \text{ occurs to the right of } i+1 \text{ in } (x_1, \dots, x_n) \}$

"left descent set"

Input vector x ($n = 50$):

```
Right descent set:
Epoch  299: Train loss 0.01, Test loss 0.01,  4907 out of  5000 correct (98.14%).
Left descent set:
Epoch  299: Train loss 0.68, Test loss 0.70,     0 out of  5000 correct (0%).
```

Input permutation matrices ($n=20$):

```
Right descent set:
Epoch 64: Train loss 0.00, Test loss 0.01, 4975 out of 5000 correct (99.50%).
Left descent set:
Epoch 64: Train loss 0.00, Test loss 0.01, 4977 out of 5000 correct (99.54%).
```

How input sits in space (the "representation") really matters.

Myths, advice and scale

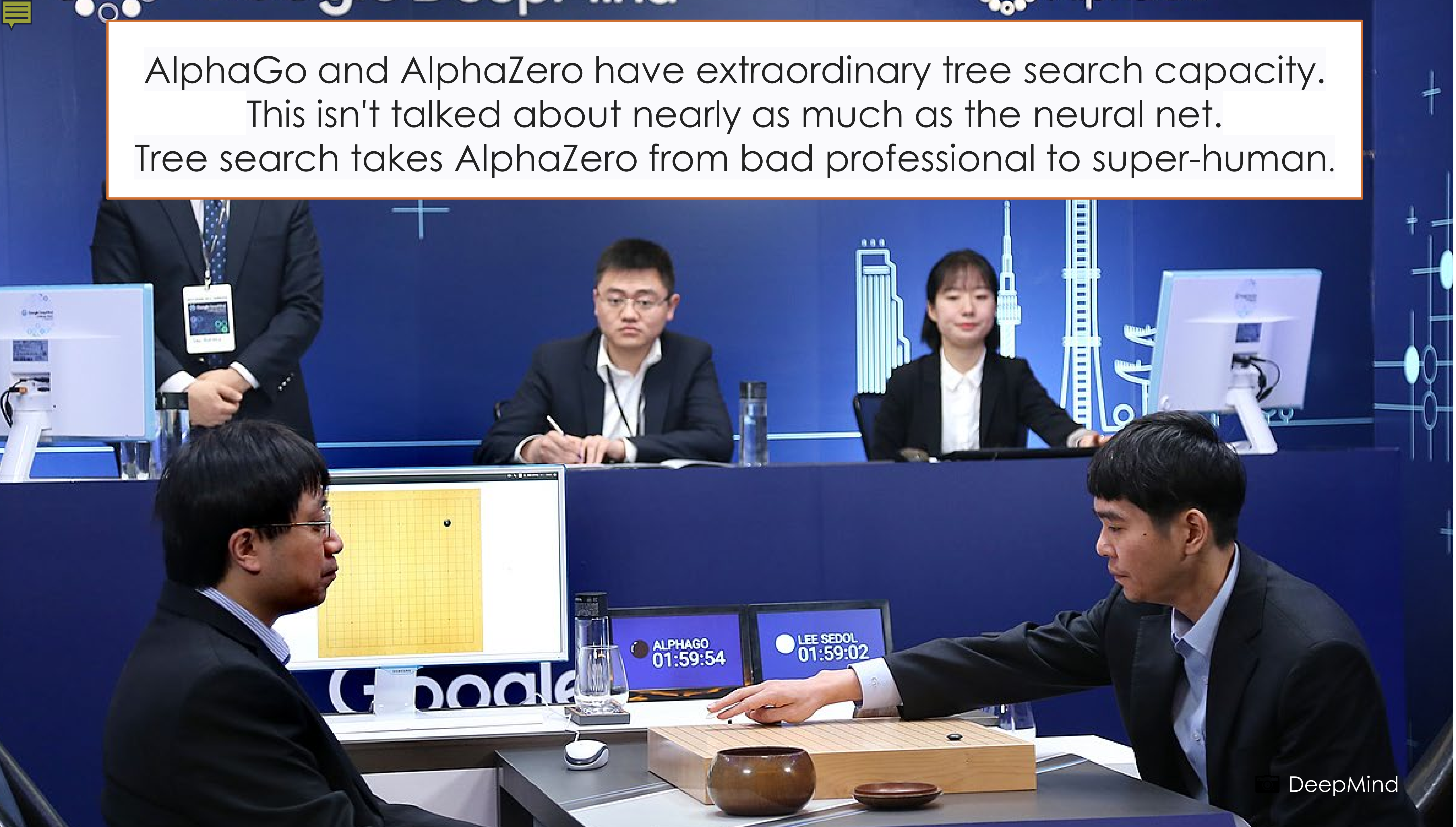
TheDigitalArtist

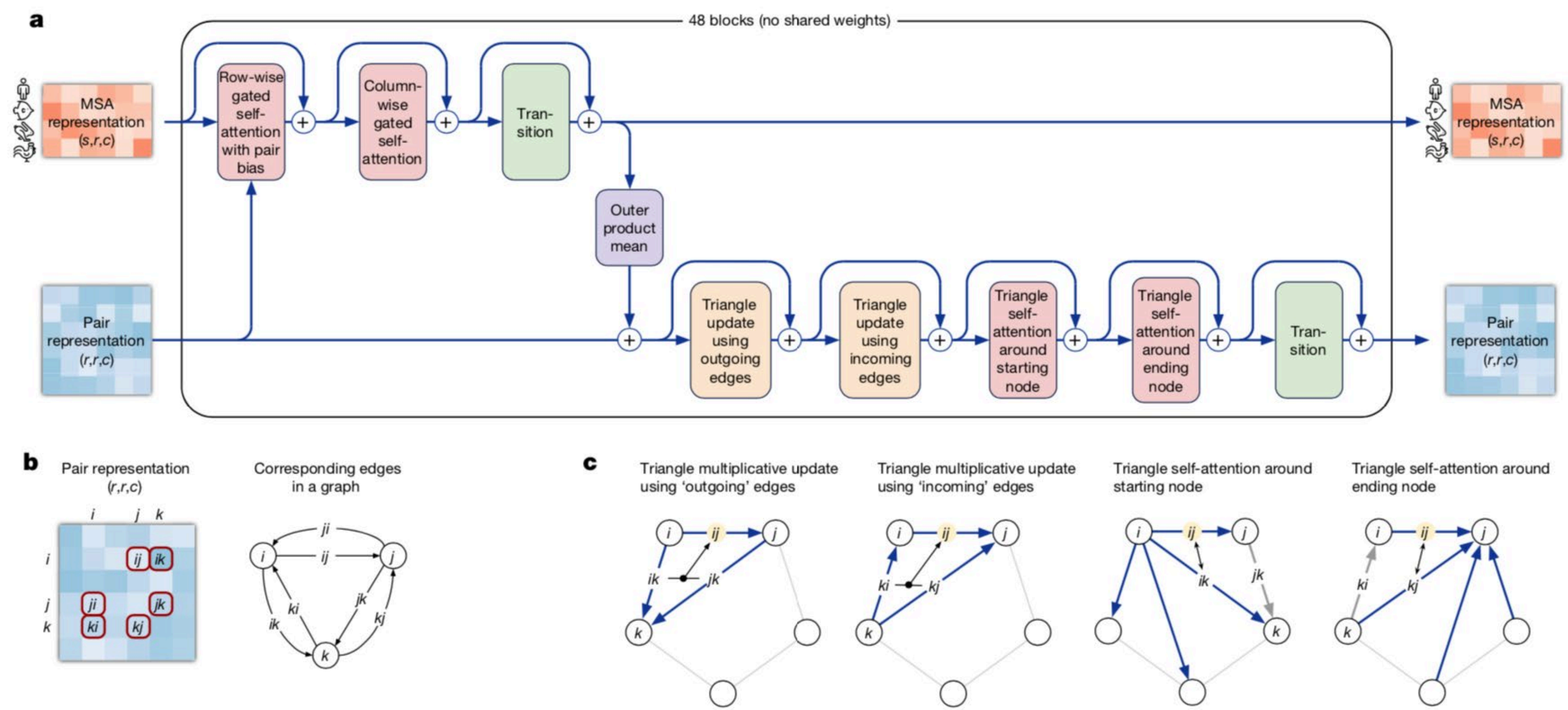
Hype: There is one "vanilla" structure (a "fully connected deep neural net") which is versatile on many problems.

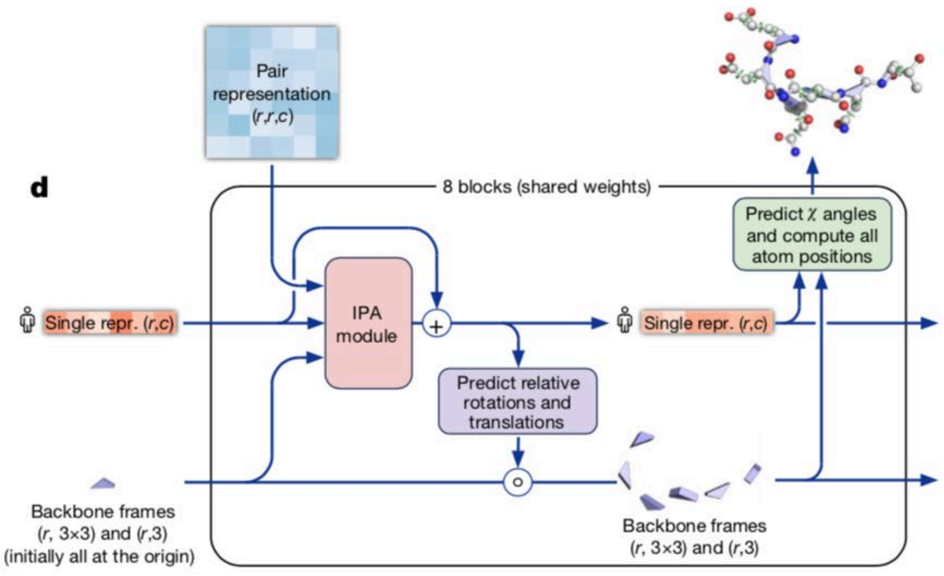
Reality: Vanilla architectures usually only work out of the box on simple problems.

Architecture matters.
Deep learning is often a *piece* of the puzzle.

AlphaGo and AlphaZero have extraordinary tree search capacity.
This isn't talked about nearly as much as the neural net.
Tree search takes AlphaZero from bad professional to super-human.







AlphaFold is often referred to as a "neural net".
This is not accurate.
It is a remarkable piece of software which incorporates neural nets in an essential way. It also incorporates several other components distilled from human understanding of the key features of an extremely difficult problem.

Advice for the interested mathematician:

1) Expect to spend considerable time experimenting with details like learning rate, model selection etc.

2) Try to work with someone who has background in machine learning.

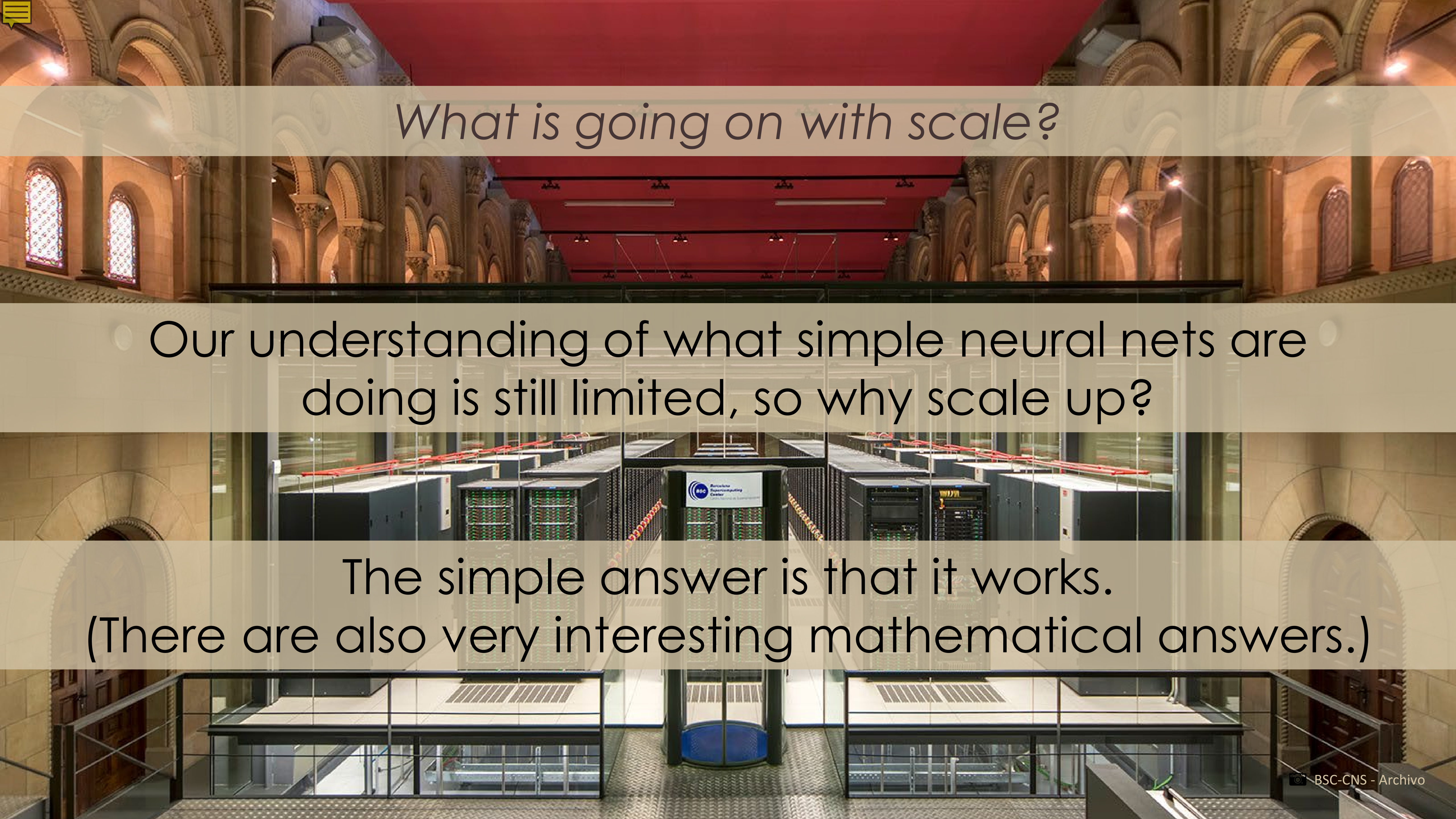
3) Try to push *either* mathematics or machine learning, but not both! (Remember that AlphaGo began as a supervised learning task.)

4) Have a precise idea of what you want machine learning to achieve. (We are not yet at the stage where we can "throw AlphaZero at the problem".)

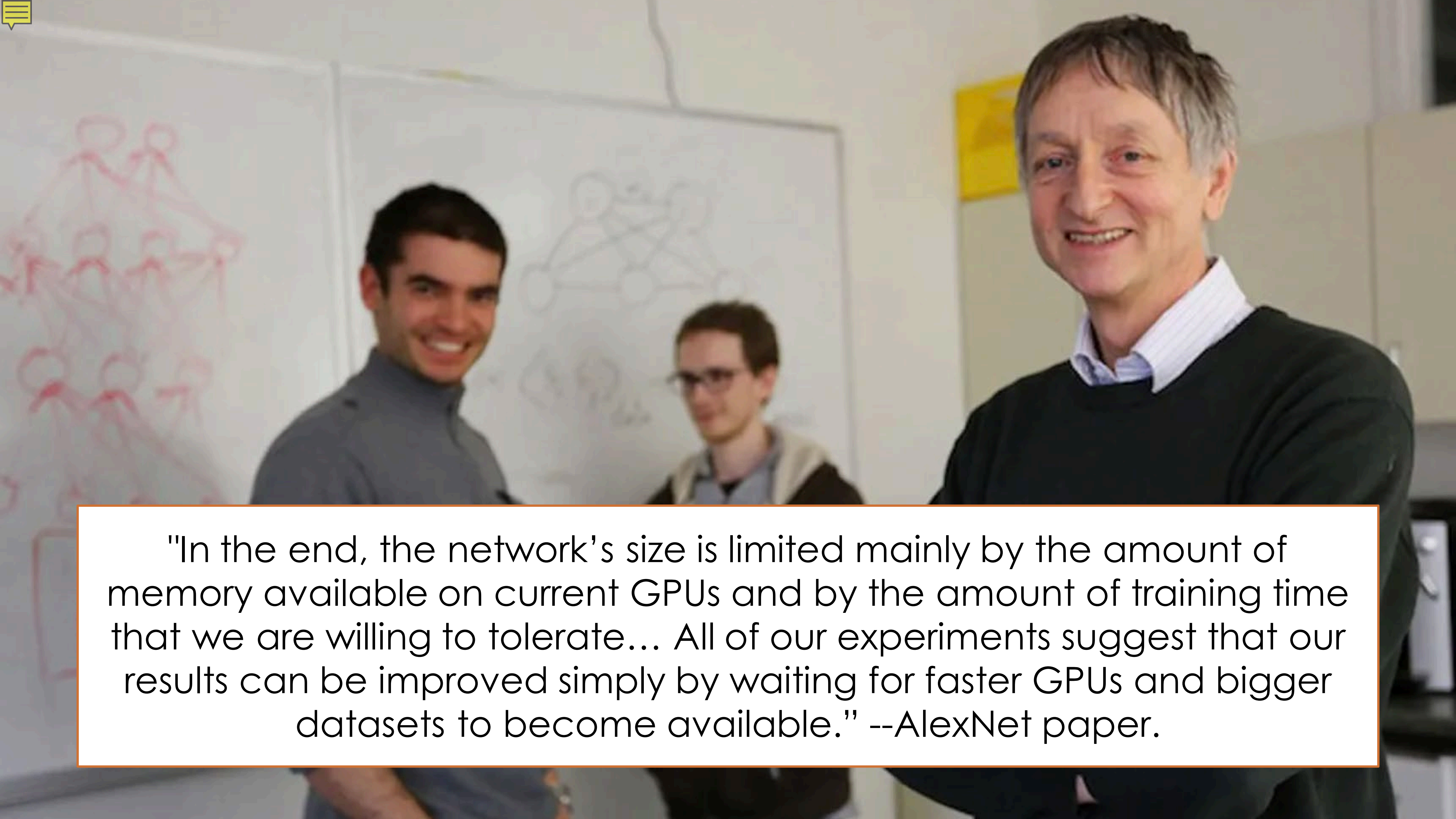
What is going on with scale?

Our understanding of what simple neural nets are doing is still limited, so why scale up?

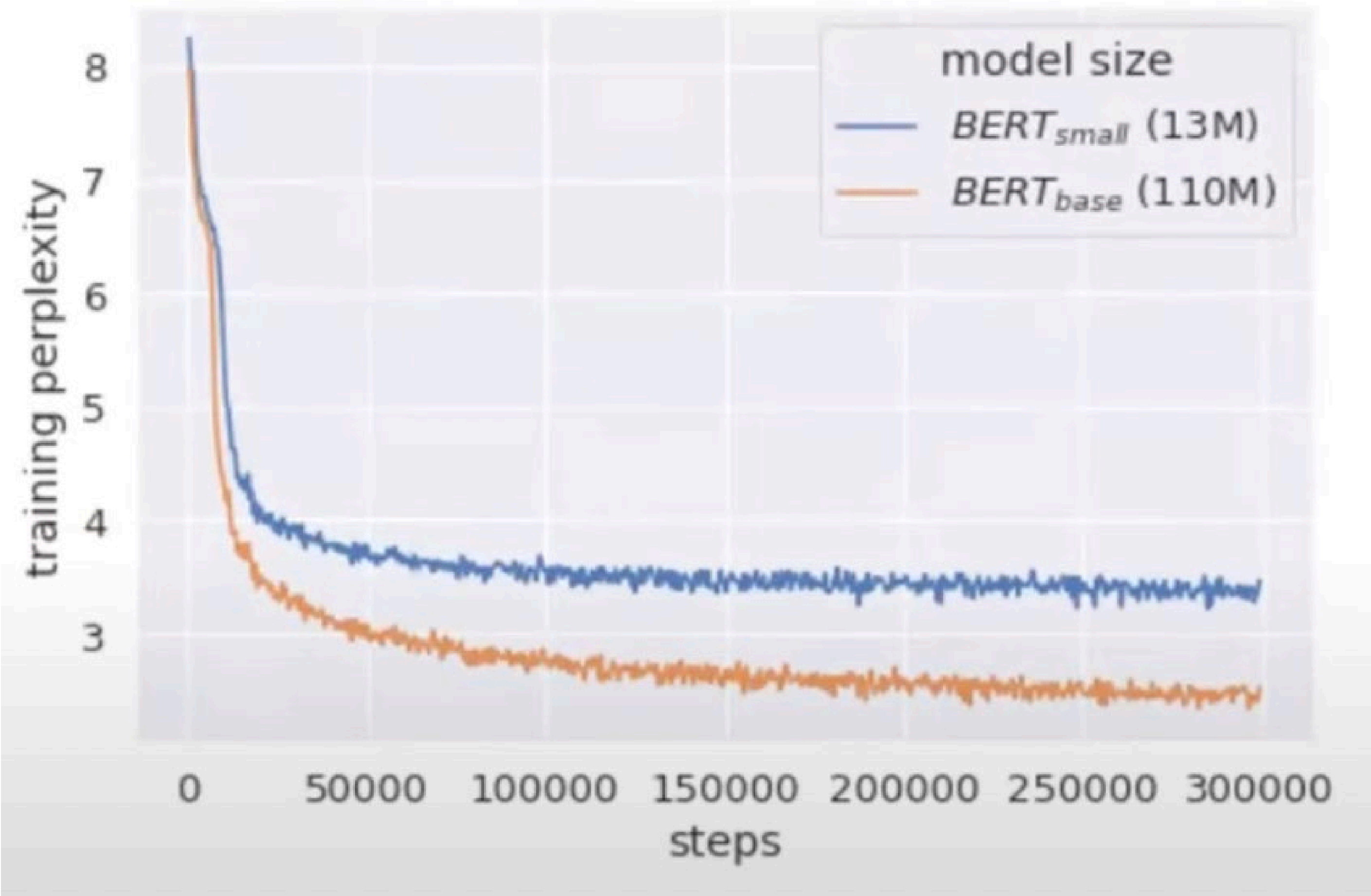
The simple answer is that it works.
(There are also very interesting mathematical answers.)



BSC-CNS - Archivo



"In the end, the network's size is limited mainly by the amount of memory available on current GPUs and by the amount of training time that we are willing to tolerate… All of our experiments suggest that our results can be improved simply by waiting for faster GPUs and bigger datasets to become available." --AlexNet paper.



Greg Yang, https://www.youtube.com/watch?v=XpU3mDKJOak&ab_channel=AutoMLSeminars

Increasing network size appears to monotonically increase performance. (After getting numerous details right!)

Created with DALL·E, an AI system by OpenAI

“An epic fight between a laptop, a lone tiger and a compass, oil painting ”

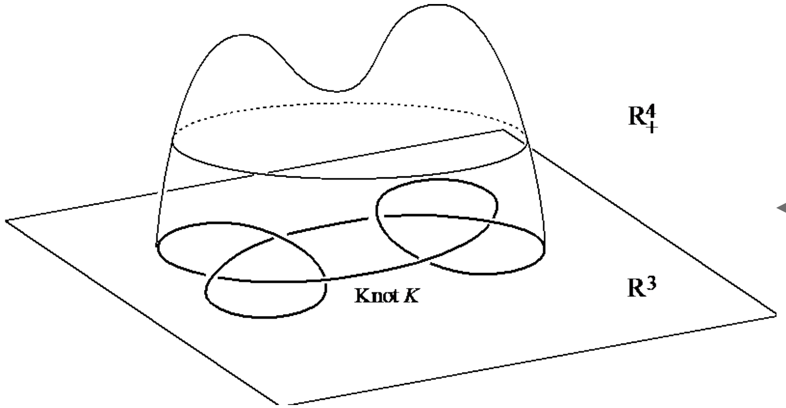
Three Examples

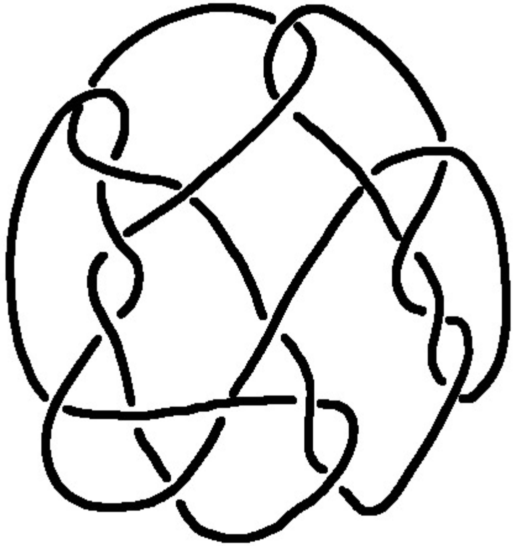
Machine learning in use in knot theory, representation theory and graph theory

Knot theory

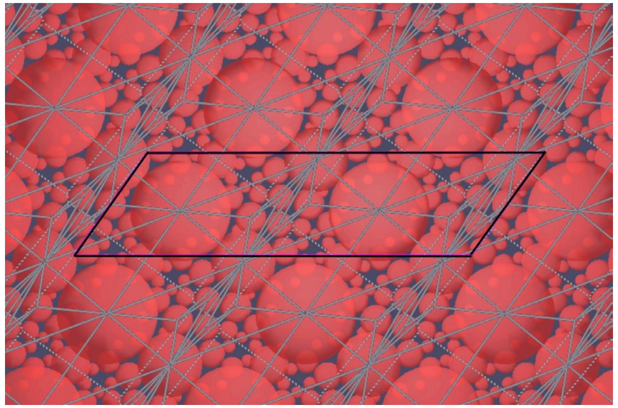
Knot Theory

3d and 4d topology





hyperbolic geometry



quantum topology, mathematical physics, …

Knot Theory

Signature: 2

Alexander polynomial: $t^3 - 5t^2 + 12t - 15 + 12t^{-1} - 5t^{-2} + t^{-3}$

Hyperbolic volume: 13.29

Jones: $-2q^6 - 5q^5 - 7q^4 + 9q^3 - 9q^2 + 8q - 6 + 4q^{-1} - q^{-2}$

HOMFLY-PT: $z^6a^{-2} + 3z^4a^{-4} - z^4a^{-4} - z^4 + 2z^2a^{-2} - z^2 - a^{-2} + 2a^{-4} - a^{-6} + 1$

A2: $-q^6 + 2q^4 + 1 + 2q^{-2} - 3q^{-4} + q^{-6} - 2q^{-8} + 2q^{-10} + 2q^{-12} + 2q^{-16} - 2q^{-18} - q^{-20}$

3-genus: 3

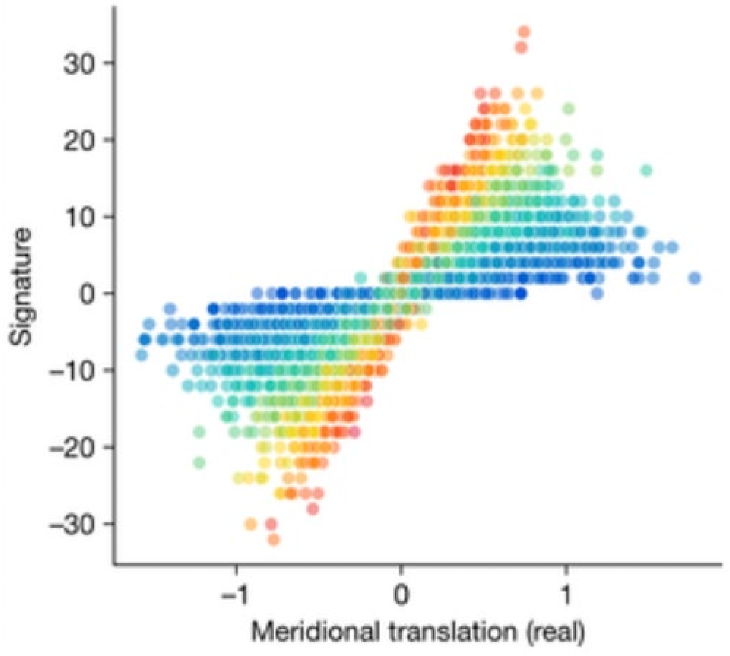
Topological 4-genus: 1

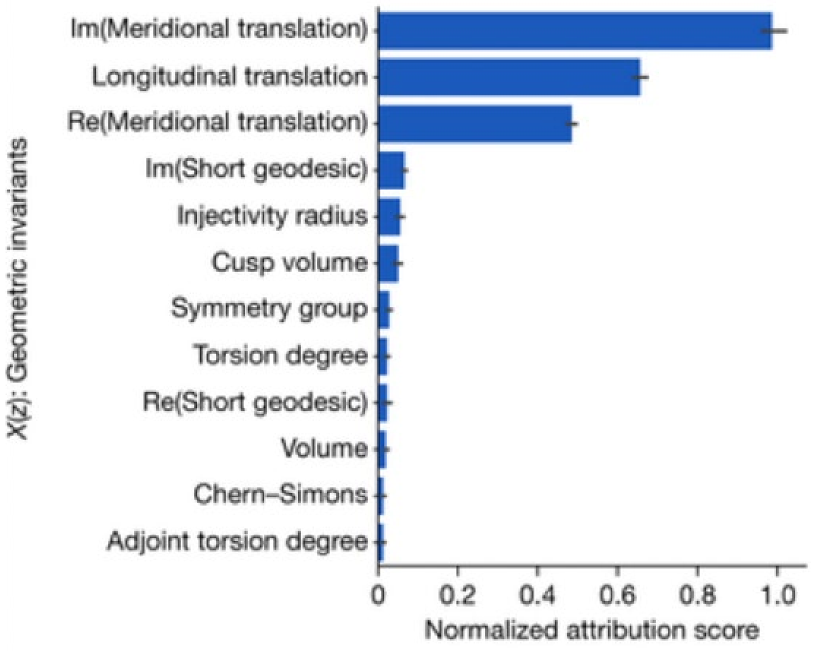
Do there exist unexpected relations between these invariants?

Smooth 4-genus: 1

Determinant: 51

Knot Theory





Davies, Juhász, Lackenby and Tomasev prove:

Theorem 1.1. *There exists a constant c_1 such that, for any hyperbolic knot K,*

$$|2\sigma(K) - \mathrm{slope}(K)| \leq c_1 \,\mathrm{vol}(K)\,\mathrm{inj}(K)^{-3}.$$

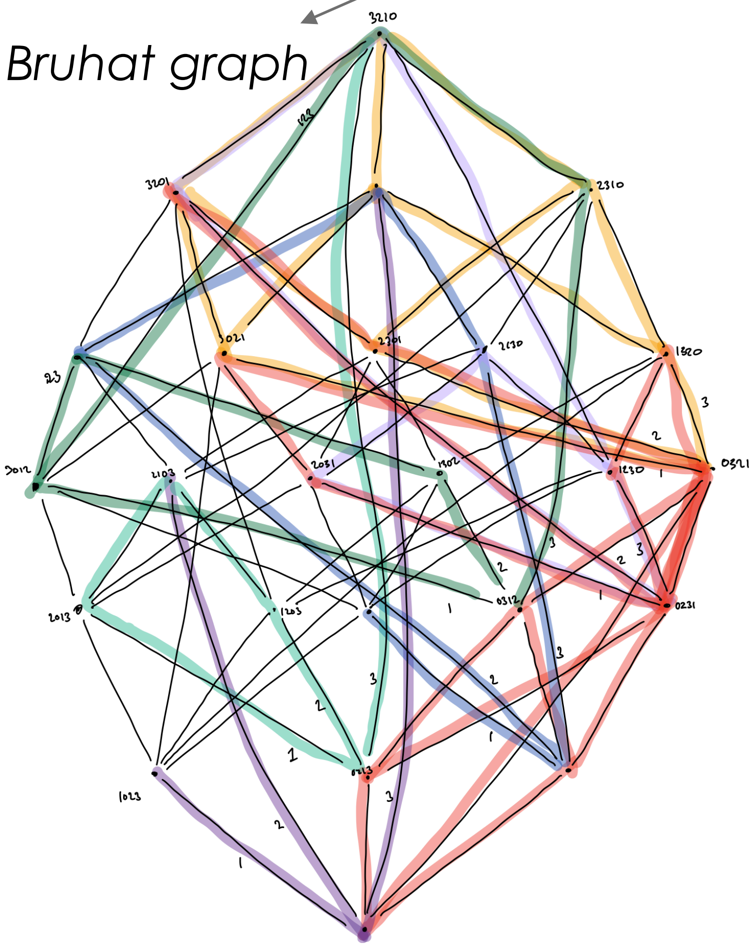
Representation theory

Representation Theory

Combinatorial invariance conjecture (Dyer, Lusztig 1980s)

x, y (pair of permutations)

Bruhat graph



Kazhdan-Lusztig polynomial

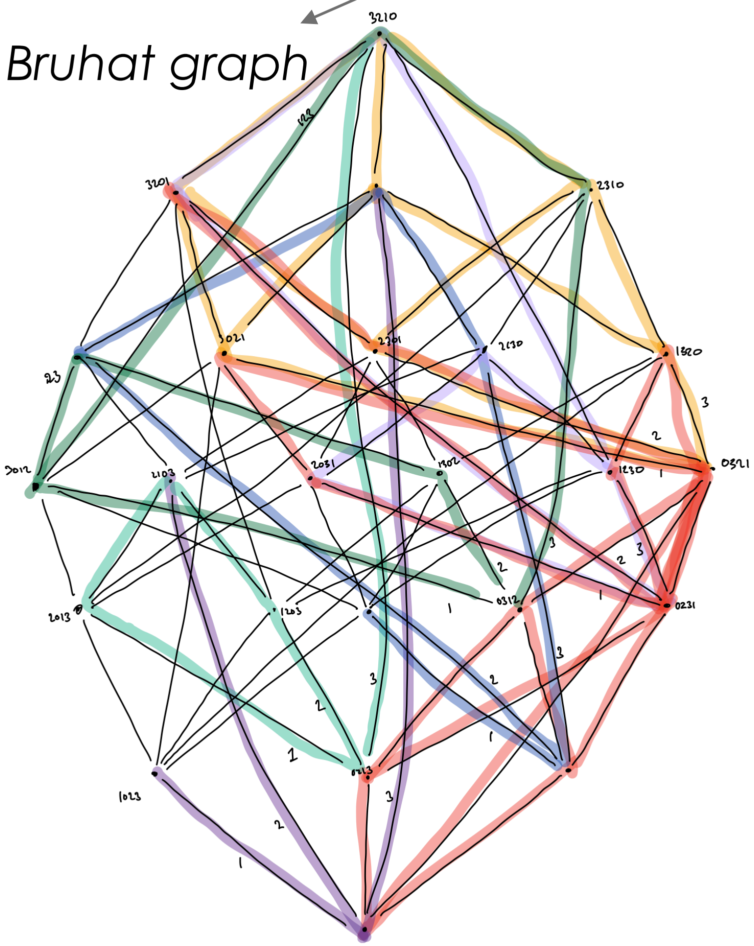
$$1+3q+q^2$$

Representation Theory

Combinatorial invariance conjecture (Dyer, Lusztig 1980s)

x, y (pair of permutations)

Bruhat graph



Kazhdan-Lusztig polynomial

conjecture

$$1+3q+q^2$$


3210
4
3201
2310
2
3021
2301
2130
1320
1
3012
3
2103
2031
1302
1230
0321
2013
1203
0312
0231
1023
0213
0

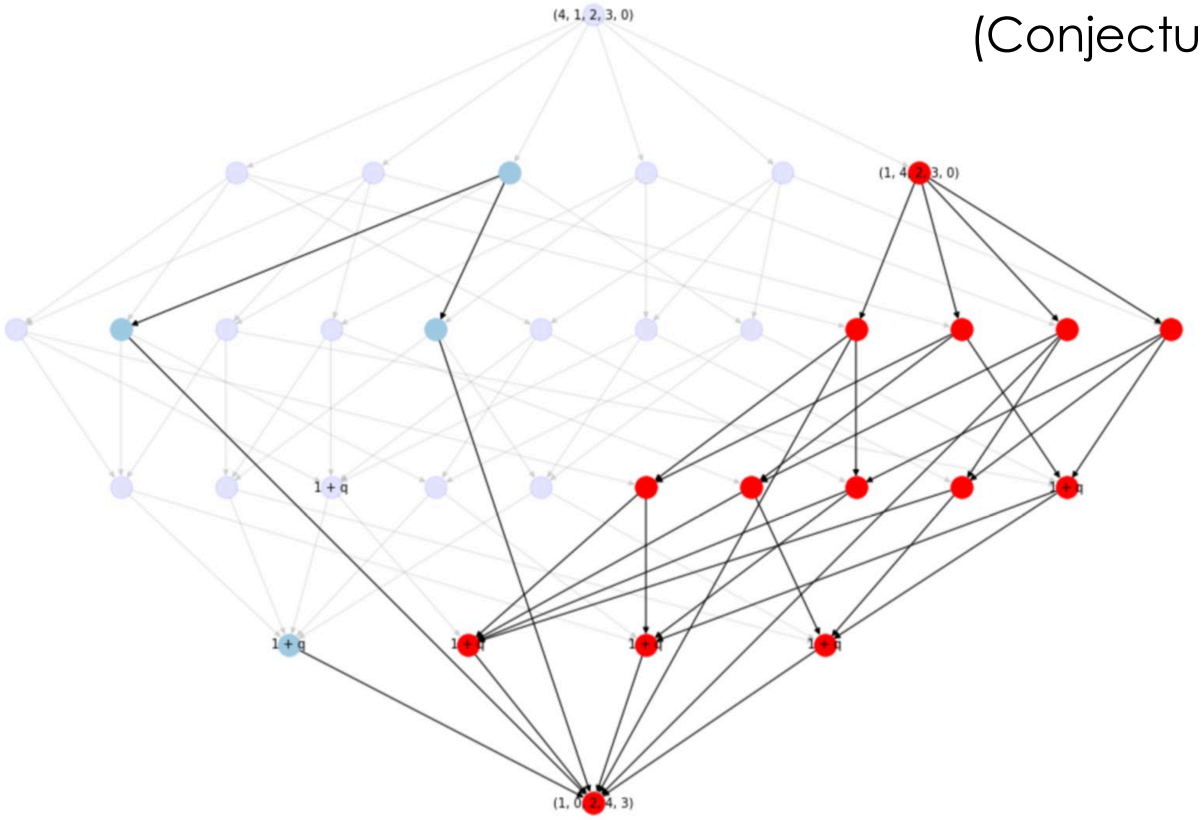
Representation Theory

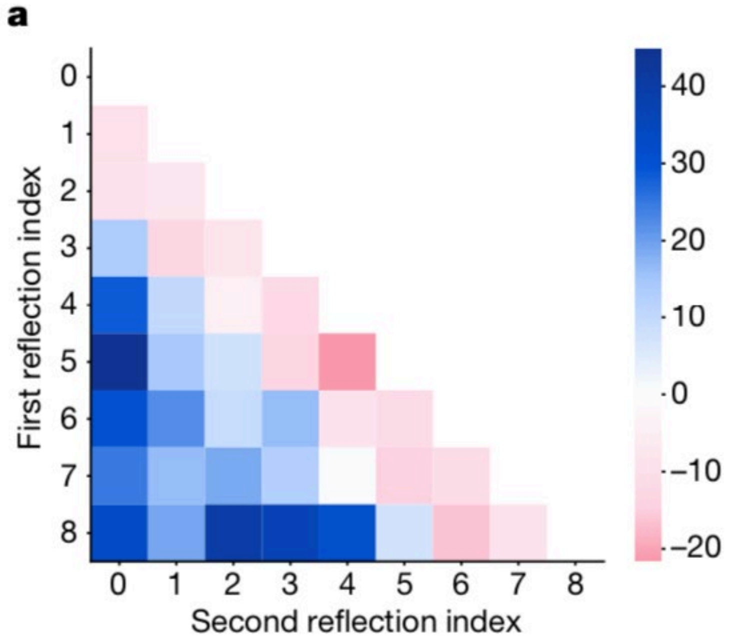
Blundell, Buesing, Davies, Veličković, Williamson:

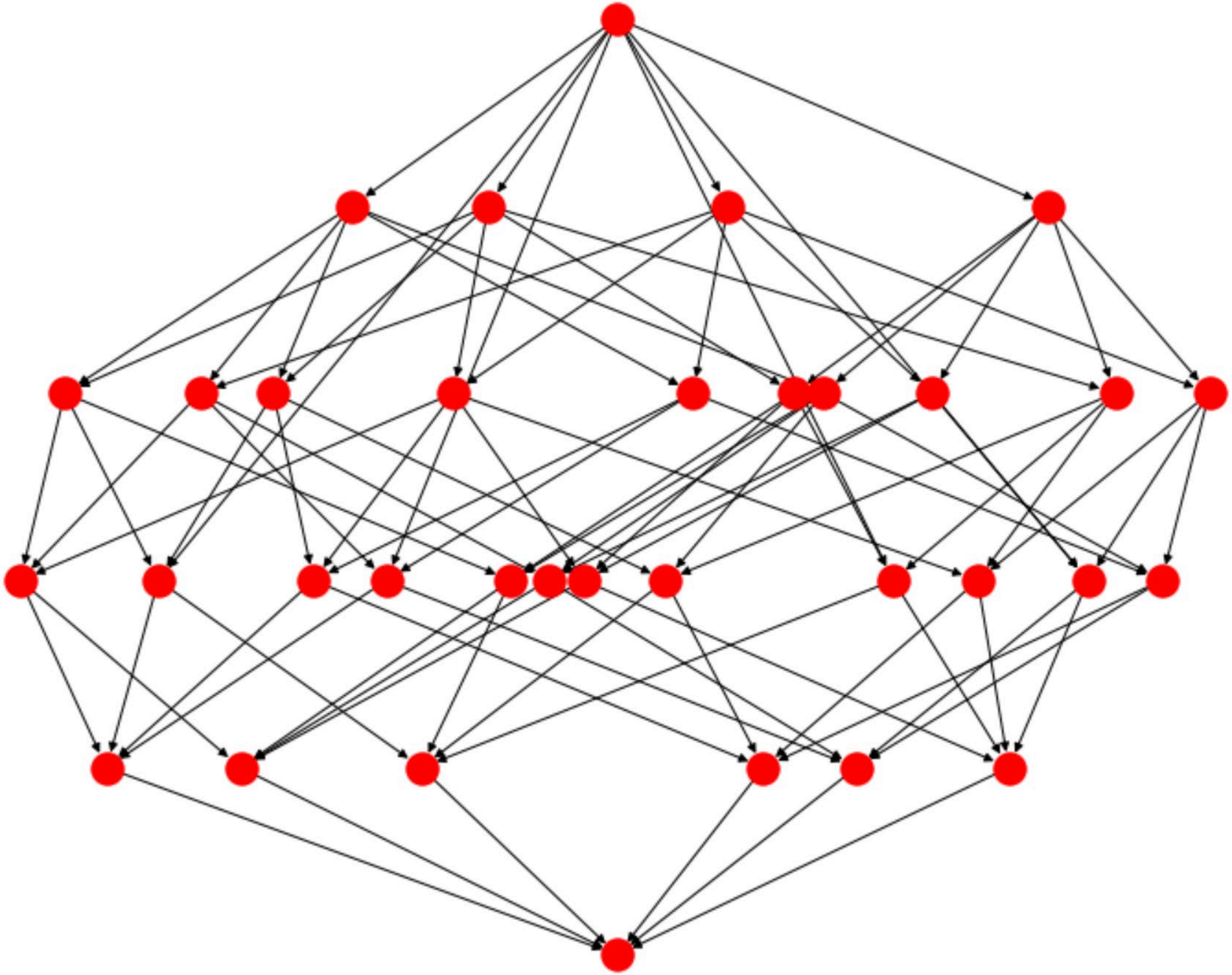
Conjecture 3.1. *For any hypercube decomposition we have*

$$P^{\partial}_{x,y} = q^{\ell(y)-\ell(x)-1} \sum_{\varnothing \neq I \subset E} (q^{-1}-1)^{|I|-1} P_{\theta(I),y}(q^{-1}) + \sum_{x \neq v \in J} \gamma_v \cdot P^{\partial}_{x,v}$$

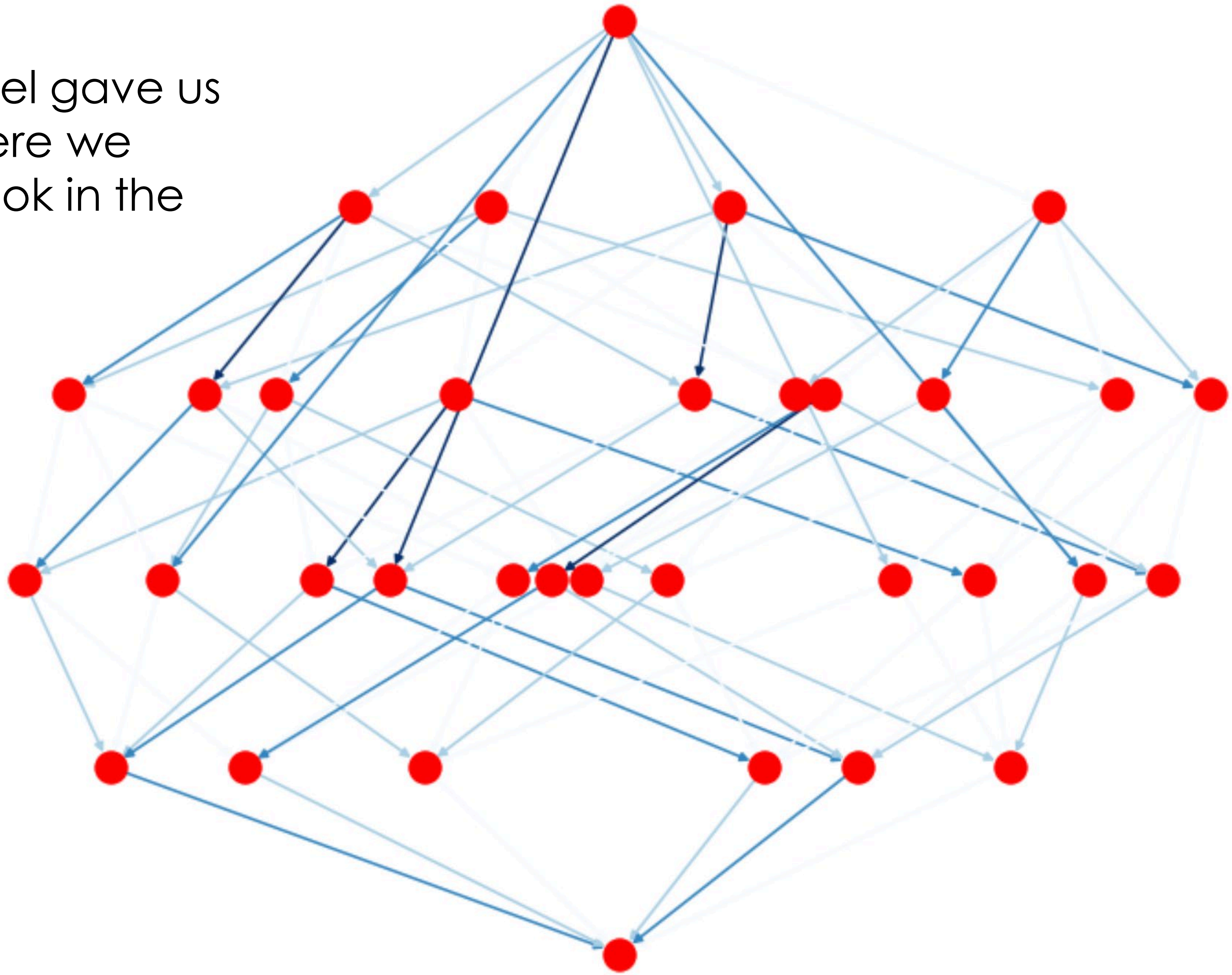
(Conjecture is proved in an important special case.)





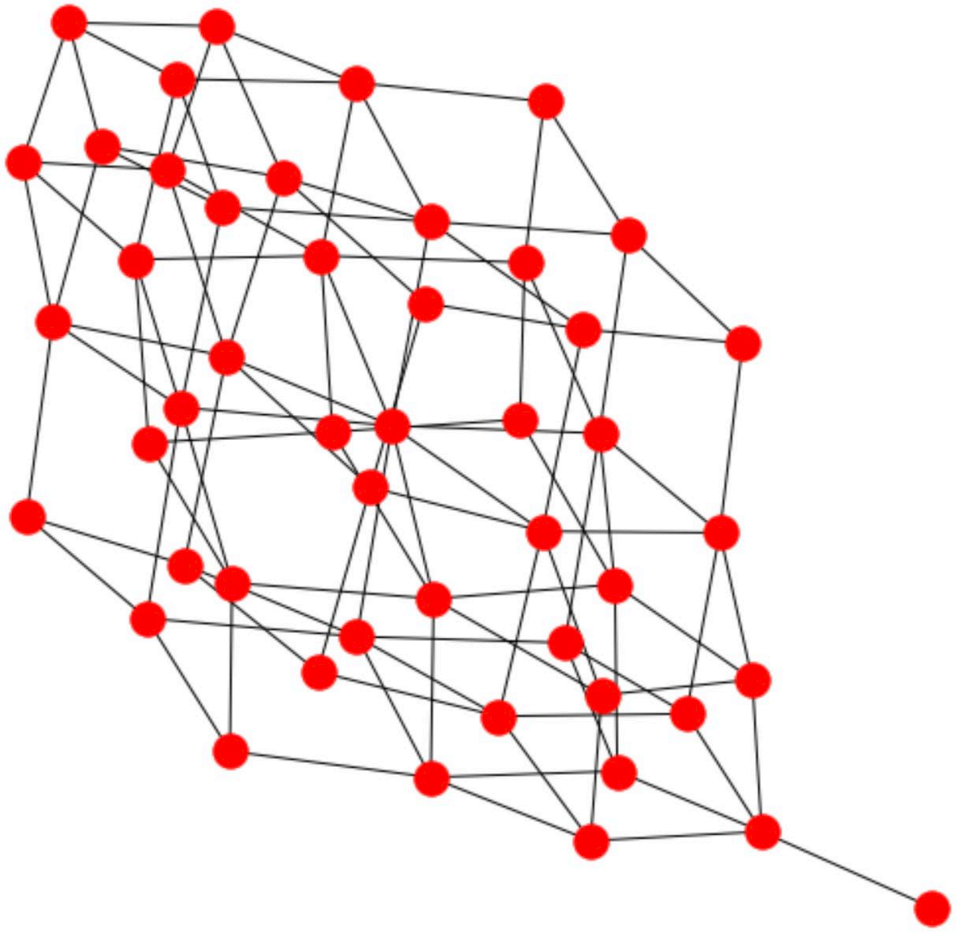


The model gave us hints where we should look in the graph.



Graph Theory

Graph Theory



Graph theory contains many conjectures.
Some are true. Some are simply false.

However, finding counter-examples is difficult!

Wagner: Finding a counter-example can be posed as a game, and computers can be trained to play the game via reinforcement learning.

Thus, the computer generates hundreds of examples at random, by accepting or rejecting an edge. Over multiple training rounds it learns patterns that result in graphs which are close to being counter-examples.

Conjecture 2.1 ([4]). *Let G be a connected graph on $n \geq 3$ vertices, with largest eigenvalue λ_1 and matching number μ. Then*

$$\lambda_1 + \mu \geq \sqrt{n-1} + 1.$$

Conjecture 2.3 (Auchiche–Hansen [6]). *Let G be a connected graph on $n \geq 4$ vertices with diameter D, proximity π and distance spectrum $\partial_1 \geq \ldots \geq \partial_n$. Then*

$$\pi + \partial_{\lfloor \frac{2D}{3} \rfloor} > 0.$$

Summary

Neural nets perform some tasks remarkably well. They are strongest on tasks like speech recognition and image classification that is simple and intuitive for us.

The functions that neural nets like to learn are rather different from the functions I usually think about.

Architecture matters, and most applications of neural nets to “difficult” problems incorporate them into more complicated architectures.

Neural nets can provide useful tools for conjecture generation and refutation.

I suspect that the next few years will see many more applications in pure mathematics, particularly organising calculation and guiding search.

I don’t yet see convincing evidence that neural nets are capable of replicating the “system 2” parts of the mathematical process.

“We can predict everything, except the future.”
-- A Sydney fortune cookie.

THE UNIVERSITY OF SYDNEY | Mathematical Research Institute

A philanthropically funded Institute in Mathematics and Statistics within the University of Sydney

www.sydney.edu.au/research/centres/mathematical-research-institute.html

Photography & artwork

DeepMind
Mare Nostrum/BSC-CNS
Christian Haugen/Flickr
Marc Chagall: Sandi Hemmerlein/avoidingregret.com

Simulation & knot measurements

A Neural Network Playground: TensorFlow on GitHub
bit.ly/network-playground

Benjamin Burton (Regina), Jessica Purcell

Papers

Davies et al., *Advancing mathematics by guiding human intuition with AI*: nature.com/articles/s41586-021-04086-x

Davies, Juhász, Lackenby, Tomasev, *The signature and cusp geometry of hyperbolic knots*: arXiv:2111.15323

Blundell, Buesing, Davies, Veličković, Williamson, *Towards combinatorial invariance for Kazhdan-Lusztig polynomials*: arXiv:2111.15161

Wagner, *Constructions in combinatorics via neural networks*: arXiv:2104.14516v1