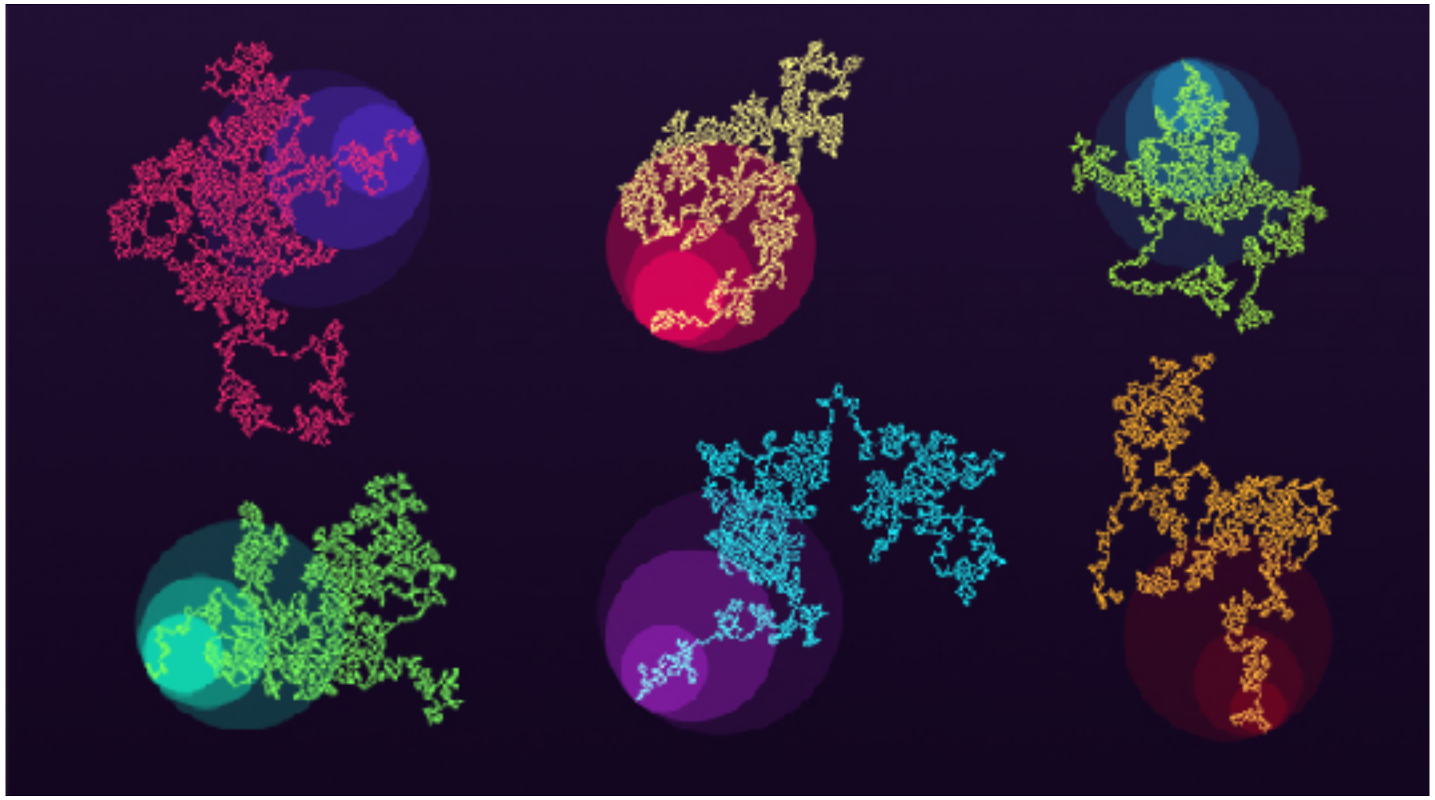
Deep learning and explainable ML

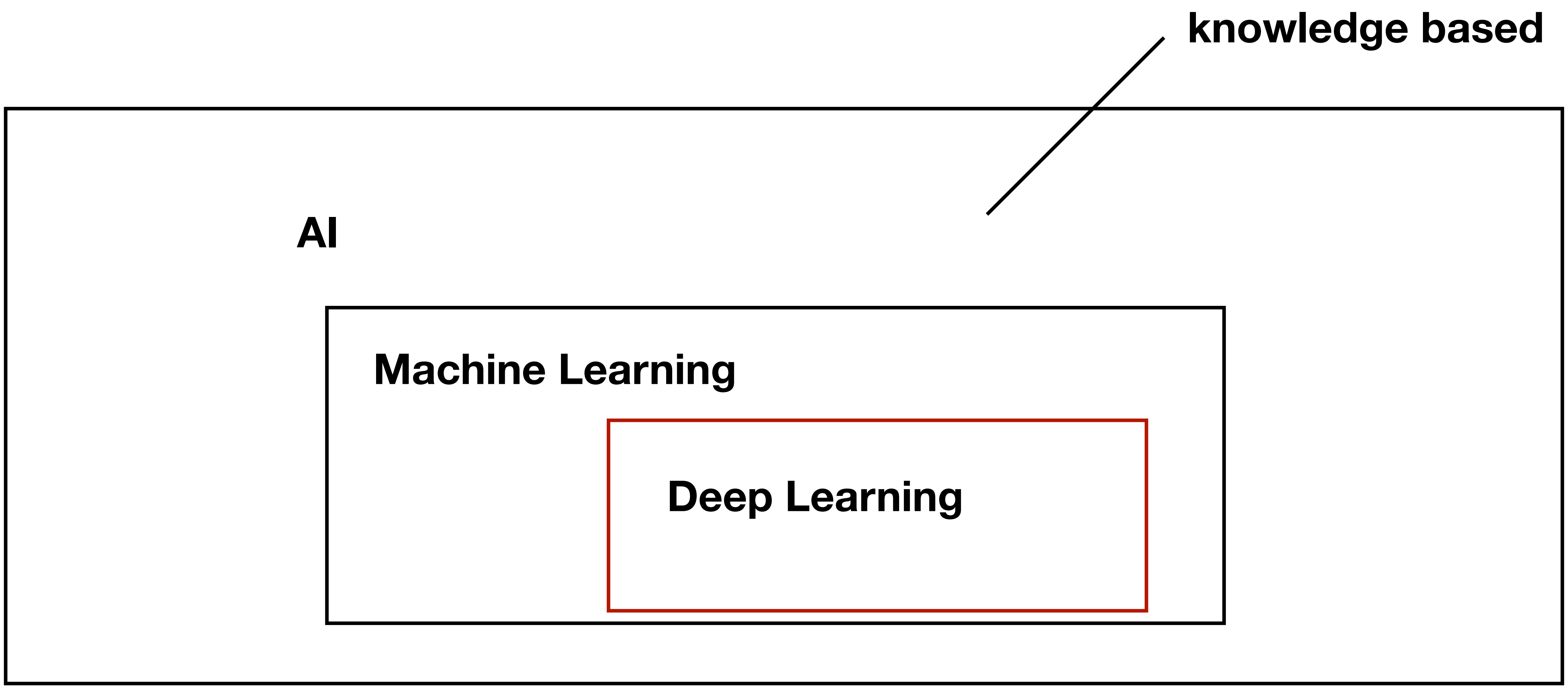
Anders Karlsson
Université de Genève and Uppsala University

SLMath, Berkeley, August 30, 2023



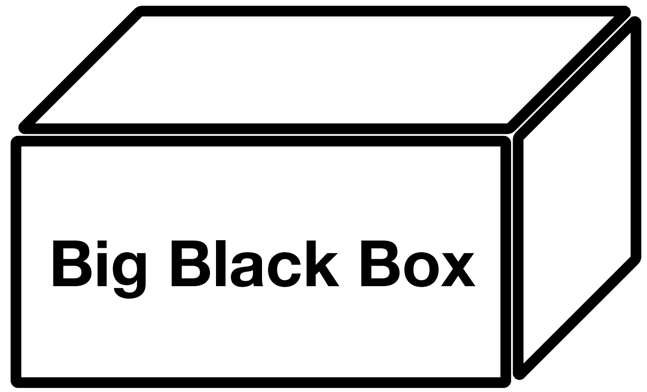
1. Introduction

The rise of Artificial Intelligence (AI)



The essential component is *neural networks* often described as software, but is just a type of mathematical function. In a one sentence description: piecewise linear maps.

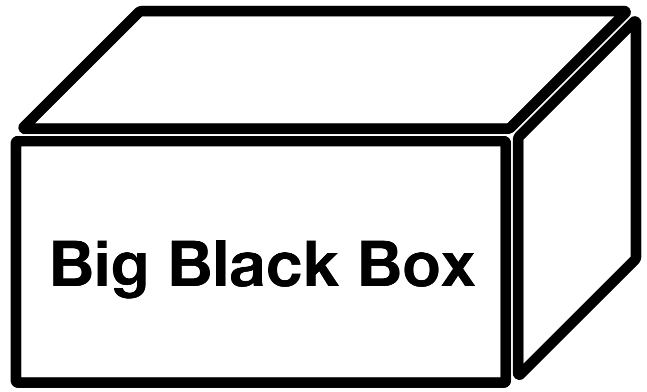
Some current or future problems with AI



contributing to:

- reliability problem
- fairness and bias problem
- alignment problem
- size and speed problem
- copyright and privacy problem
- extracting knowledge problem

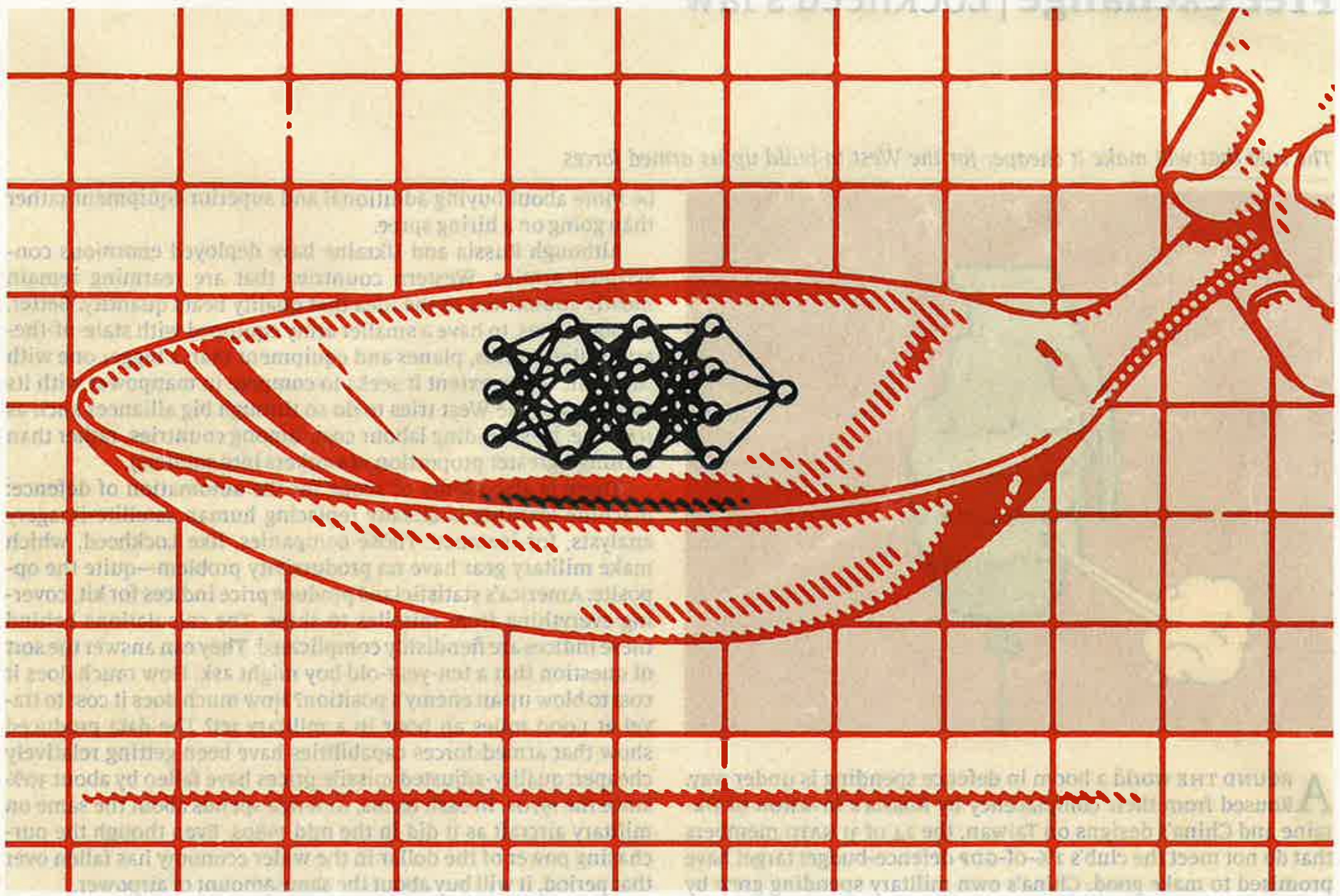
Some current or future problems with AI



contributing to:
- reliability problem
- fairness and bias problem
- alignment problem
- size and speed problem
- copyright and privacy problem
- extracting knowledge problem

- Taking-over-the-world problem
- Killing-us-all problem

The Economist, June 24, 2023



Artificial intelligence

Time for a diet

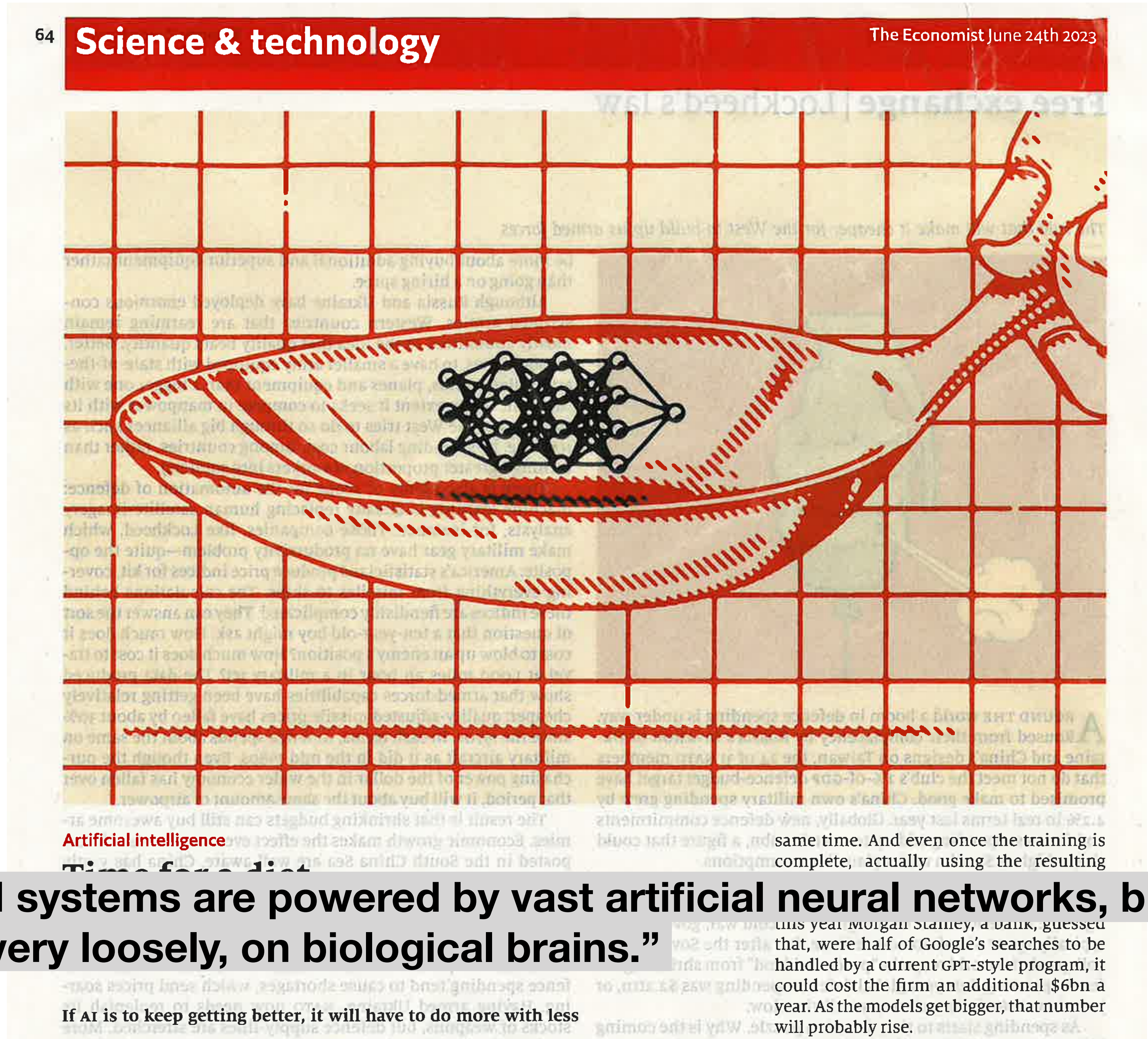
If AI is to keep getting better, it will have to do more with less

WHEN IT COMES to "large language ... times as many as its predecessor Sam Alt- ...

same time. And even once the training is complete, actually using the resulting model can be expensive as well. The bigger the model, the more it costs to run. Earlier this year Morgan Stanley, a bank, guessed that, were half of Google's searches to be handled by a current GPT-style program, it could cost the firm an additional $6bn a year. As the models get bigger, that number will probably rise.

Many in the field therefore think the

The Economist, June 24, 2023



Artificial intelligence

Time for a diet

If AI is to keep getting better, it will have to do more with less

WHEN IT COMES to "large language ... times as many as its predecessor. Sam Alt-

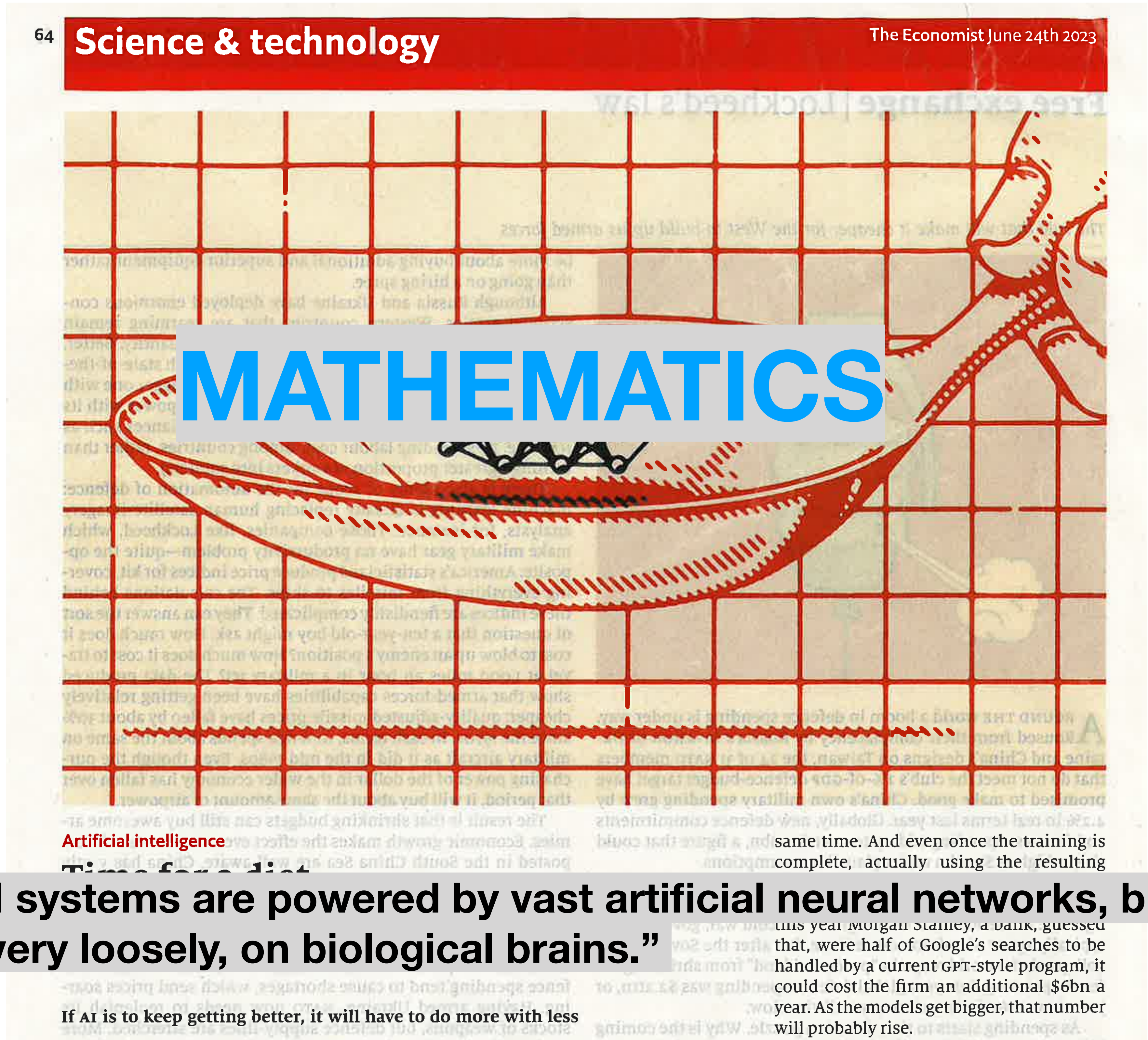
same time. And even once the training is complete, actually using the resulting ...

this year Morgan Stanley, a bank, guessed that, were half of Google's searches to be handled by a current GPT-style program, it could cost the firm an additional $6bn a year. As the models get bigger, that number will probably rise.

Many in the field therefore think the

"Modern AI systems are powered by vast artificial neural networks, bits of software modelled, very loosely, on biological brains."

The Economist, June 24, 2023

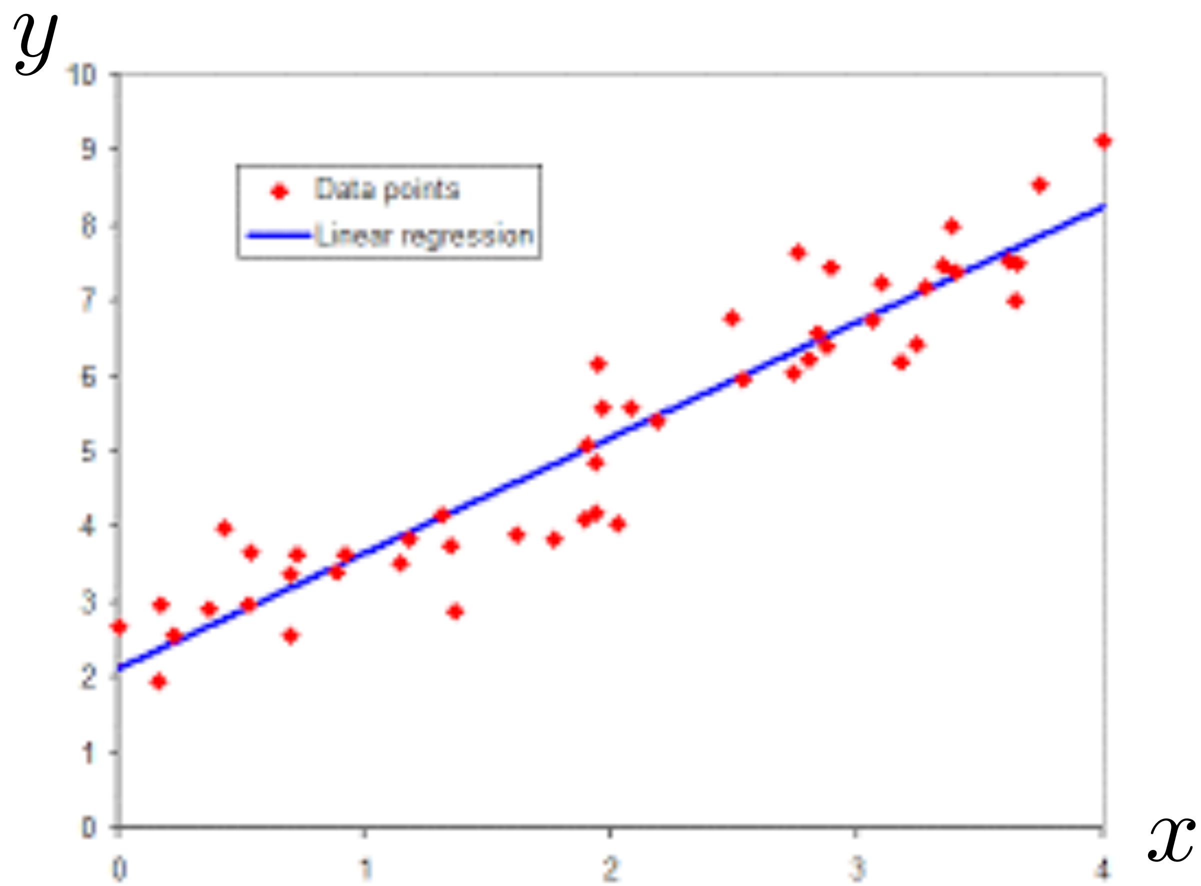


MATHEMATICS

"Modern AI systems are powered by vast artificial neural networks, bits of software modelled, very loosely, on biological brains."

2. A statistical problem — deep learning

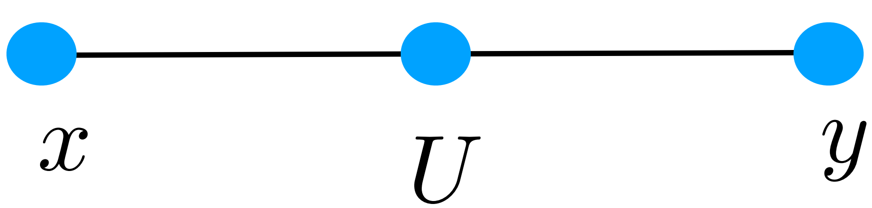
Linear regression



Find $y = ax + b$

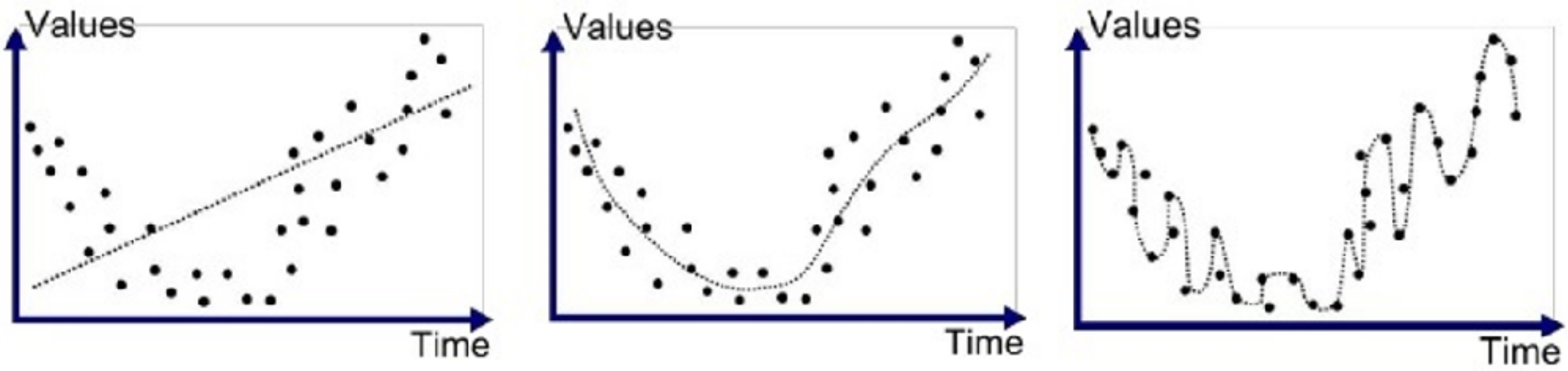
that minimizes the errors.

Smallest neural network



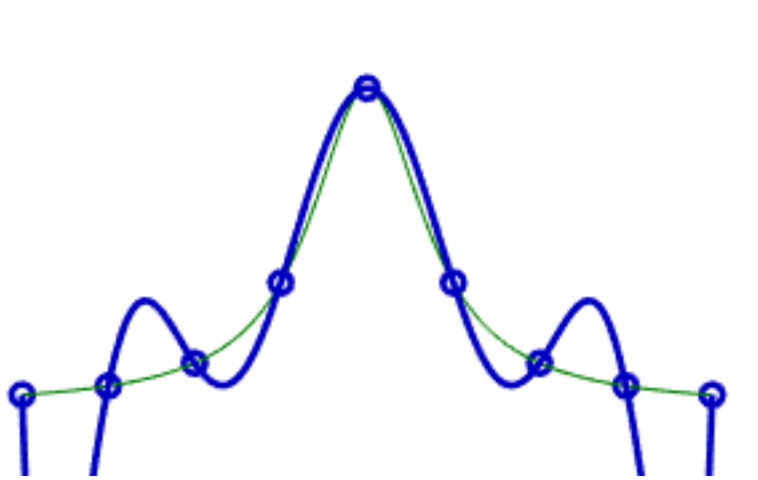
$$y = U(x)$$

A problem of statistics

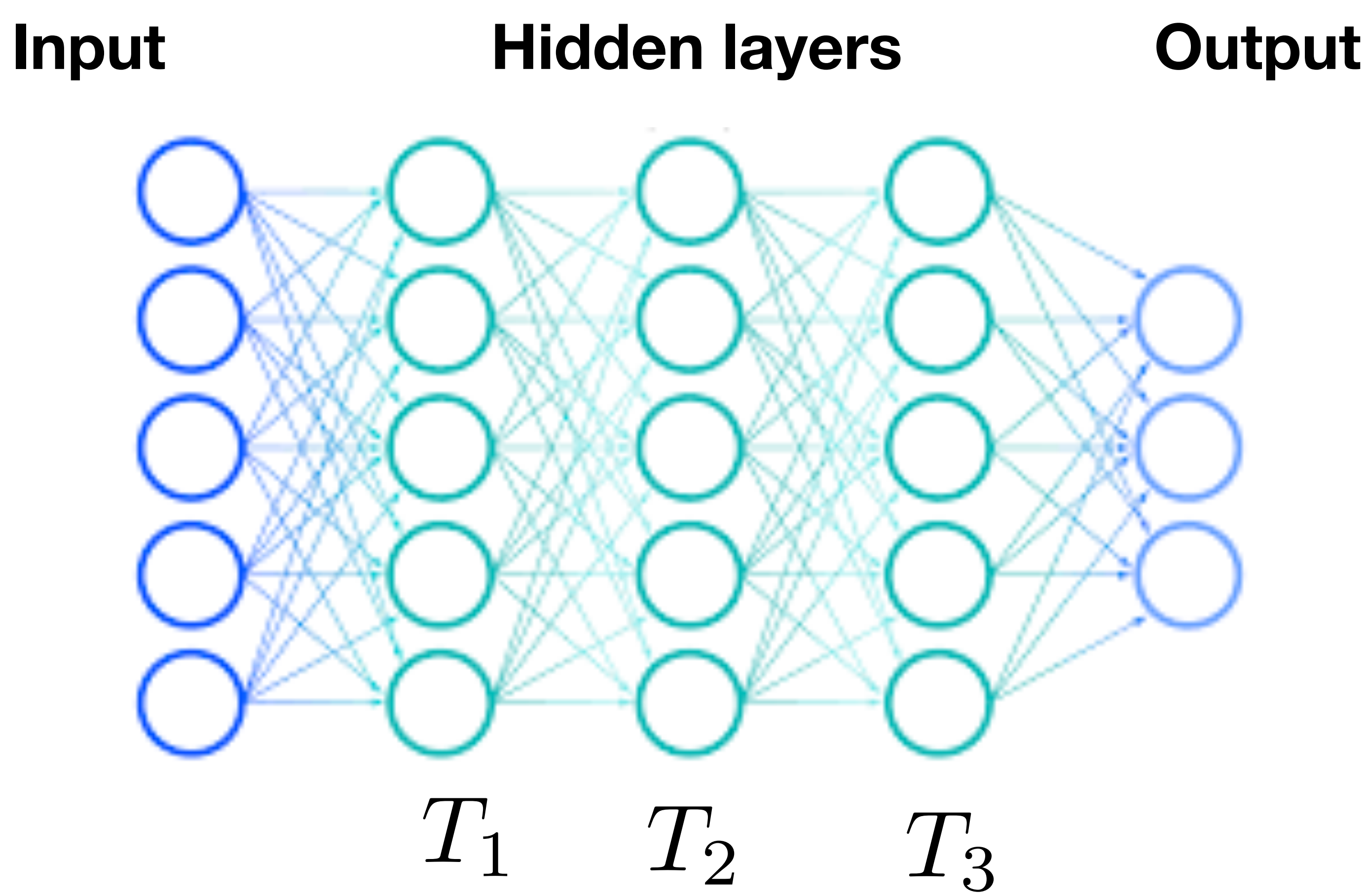


Underfitted Good Fit/Robust Overfitted

even worse with polynomials:



Neural networks



$$T_i(x) = \sigma(A_i x + b_i)$$

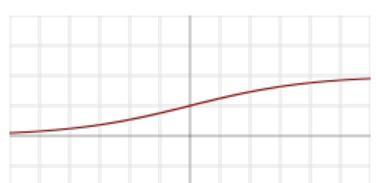
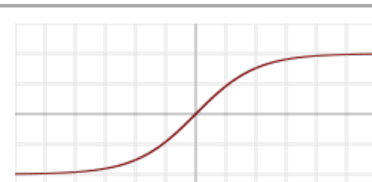
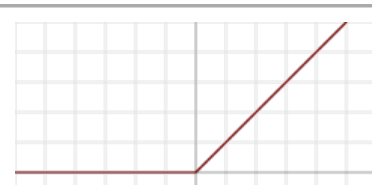
$$x_{out} = T_3 \circ T_2 \circ T_1(x_{in})$$

Neural networks

A layer is a transformation $T : x \mapsto \sigma(Ax + b)$

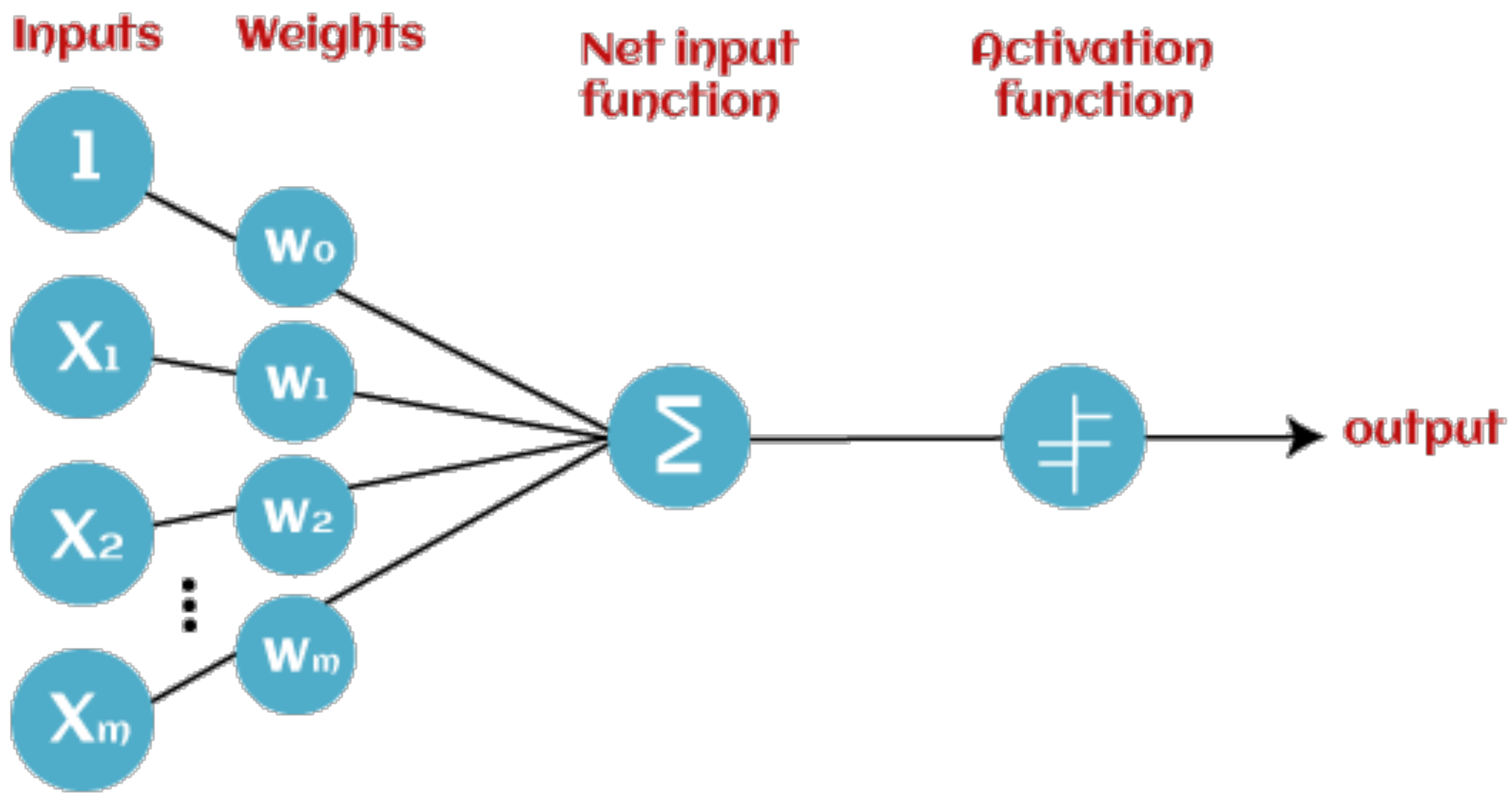
A a matrix and b a vector.

where $\sigma(t)$ is an activity function, a non-linear function applied coordinatewise.
For example, $\sigma(t) = \max 0, t$ (ReLU) or $\sigma(t) = \tanh(t)$ (TanH).

Logistic, sigmoid, or soft step		$\sigma(x) = \frac{1}{1 + e^{-x}}$
Hyperbolic tangent (tanh)		$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$
Rectified linear unit (ReLU)[9]		$\begin{cases} 0 & \text{if } x \leq 0 \\ x & \text{if } x > 0 \end{cases}$ $= \max\{0, x\} = x\mathbf{1}_{x>0}$

Nonlinearity!

Perceptron, or McCulloch-Pitts neuron, 1943



Frank Rosenblatt, 1950s, built "embryo" of computer from these, and claimed it learns by itself, and in future can do many things.

NEW NAVY DEVICE LEARNS BY DOING

Psychologist Shows Embryo of Computer Designed to Read and Grow Wiser

The New York Times, July 8, 1958

WASHINGTON, July 7 (UPI)—The Navy revealed the embryo of an electronic computer today that it expects will be able to walk, talk, see, write, reproduce itself and be conscious of its existence.

The embryo—the Weather Bureau's $2,000,000 "704" computer—learned to differentiate between right and left after fifty attempts in the Navy's demonstration for newsmen.

The service said it would use this principle to build the first of its Perceptron thinking machines that will be able to read and write. It is expected to be finished in about a year at a cost of $100,000.

Dr. Frank Rosenblatt, designer of the Perceptron, conducted the demonstration. He said the machine would be the first device to think as the human brain. As do human beings, Perceptron will make mistakes at first, but will grow wiser as it gains experience, he said.

Dr. Rosenblatt, a research psychologist at the Cornell Aeronautical Laboratory, Buffalo, said Perceptrons might be fired to the planets as mechanical space explorers.

Without Human Controls

The Navy said the perceptron would be the first non-living mechanism "capable of receiving, recognizing and identifying its surroundings without any human training or control."

The "brain" is designed to remember images and information it has perceived itself. Ordinary computers remember only what is fed into them on punch cards or magnetic tape.

Later Perceptrons will be able to recognize people and call out their names and instantly translate speech in one language to speech or writing in another language, it was predicted.

Mr. Rosenblatt said in principle it would be possible to build brains that could reproduce themselves on an assembly line and which would be conscious of their existence.

In today's demonstration, the "704" was fed two cards, one with squares marked on the left side and the other with squares on the right side.

Learns by Doing

In the first fifty trials, the machine made no distinction between them. It then started registering a "Q" for the left squares and "O" for the right squares.

Dr. Rosenblatt said he could explain why the machine learned only in highly technical terms. But he said the computer had undergone a "self-induced change in the wiring diagram."

The first Perceptron will have about 1,000 electronic "association cells" receiving electrical impulses from an eye-like scanning device with 400 photo-cells. The human brain has 10,000,000,000 responsive cells, including 100,000,000 connections with the eyes.

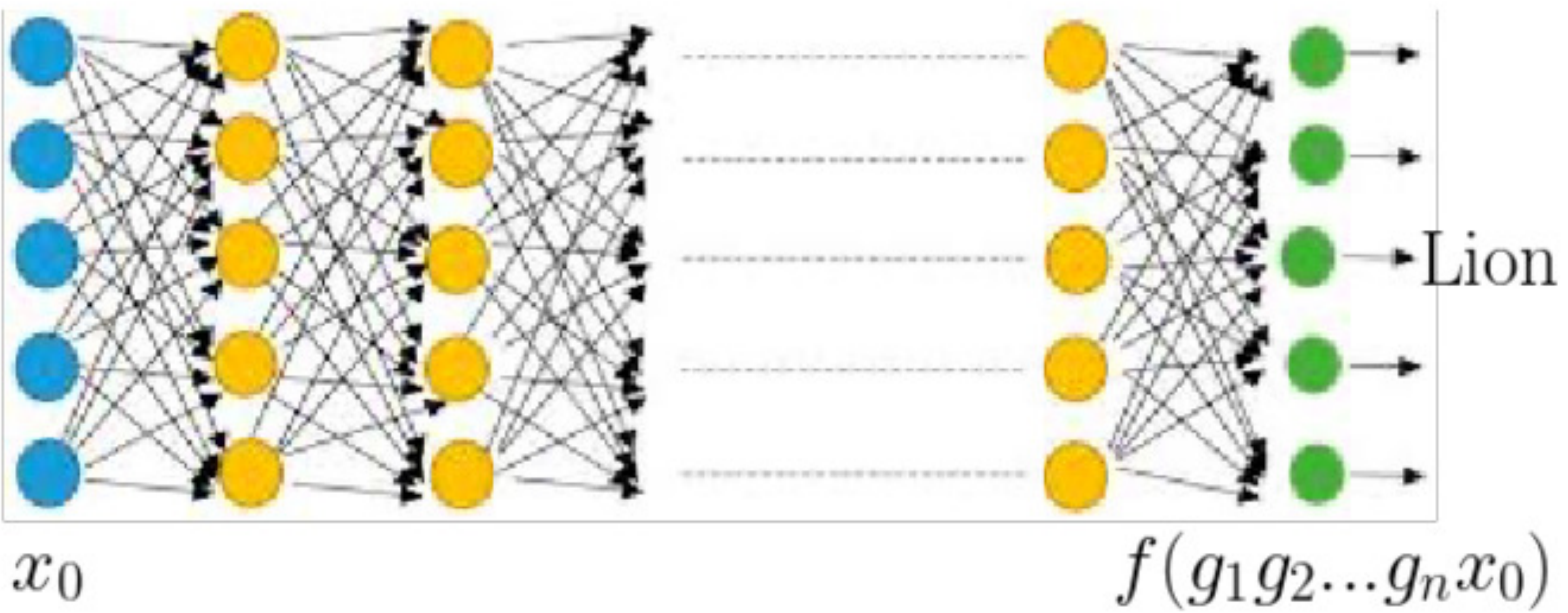
The "brain" is designed to remember images and information it has perceived itself. Ordinary computers remember only what is fed into them on punch cards or magnetic tape.

Later Perceptrons will be able to recognize people and call out their names and instantly translate speech in one language to speech or writing in another language, it was predicted.

Mr. Rosenblatt said in principle it would be possible to build brains that could reproduce themselves on an assembly line and which would be conscious of their existence.

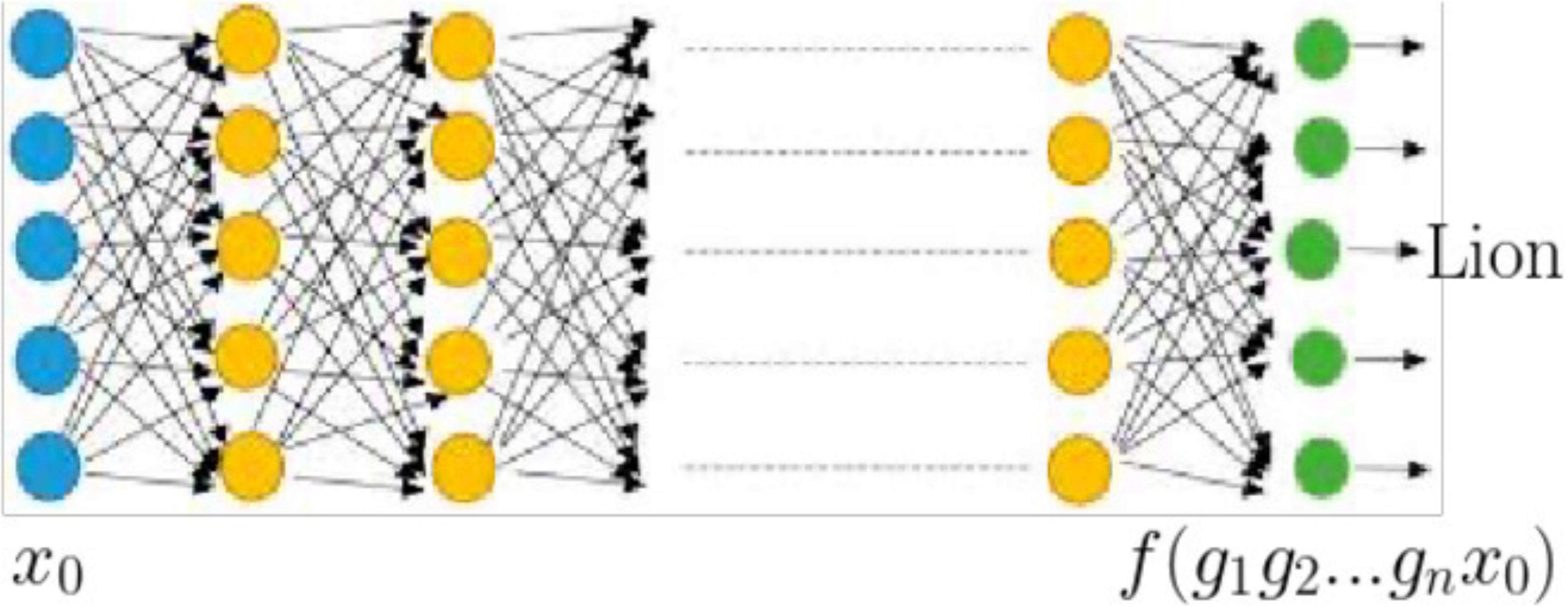
Deep learning

Find parameters A_i, b_i such that with $g_i(x) = \sigma(A_i x + b_i)$



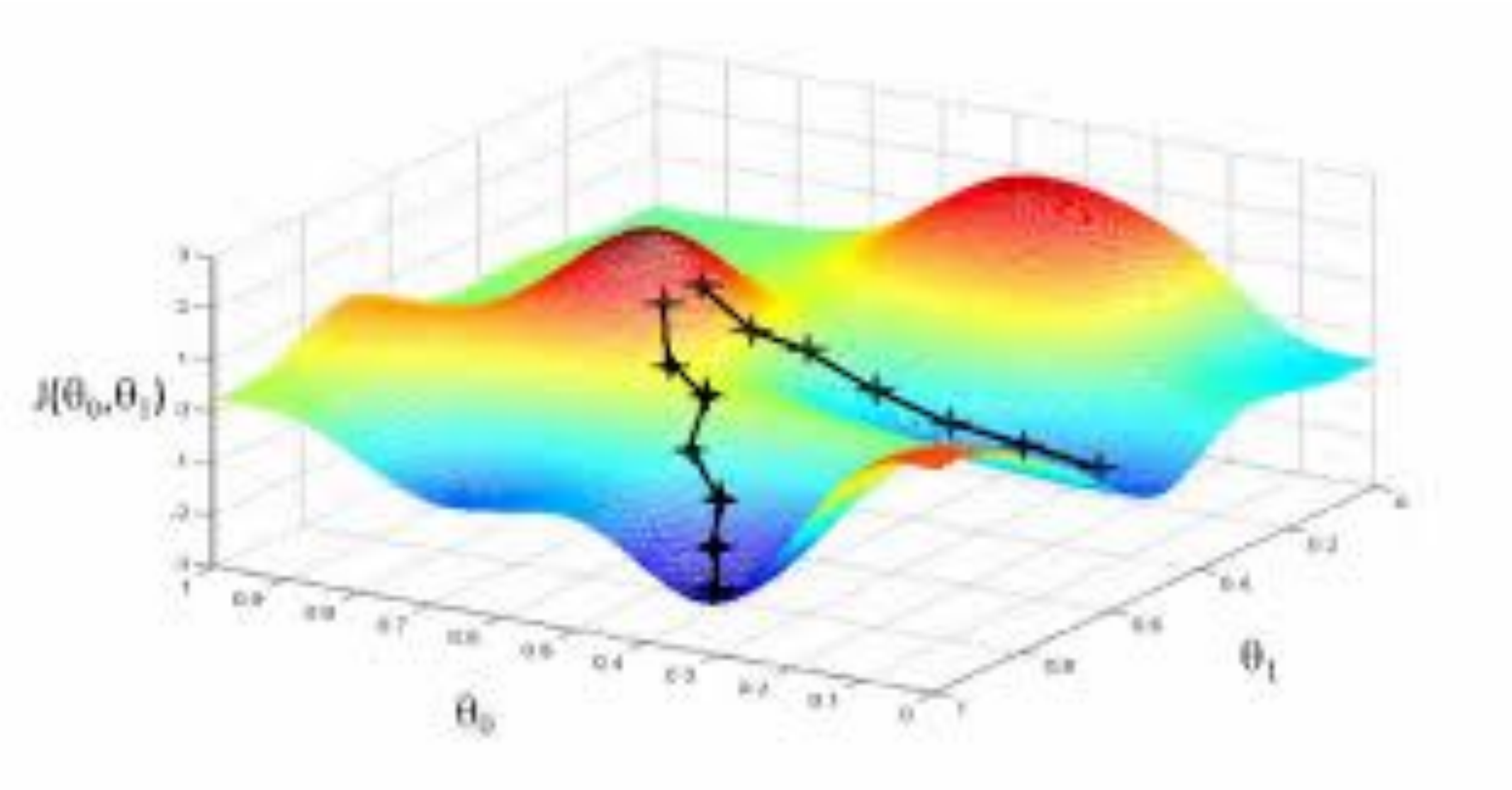
Deep learning

Find parameters A_i, b_i such that with $g_i(x) = \sigma(A_i x + b_i)$



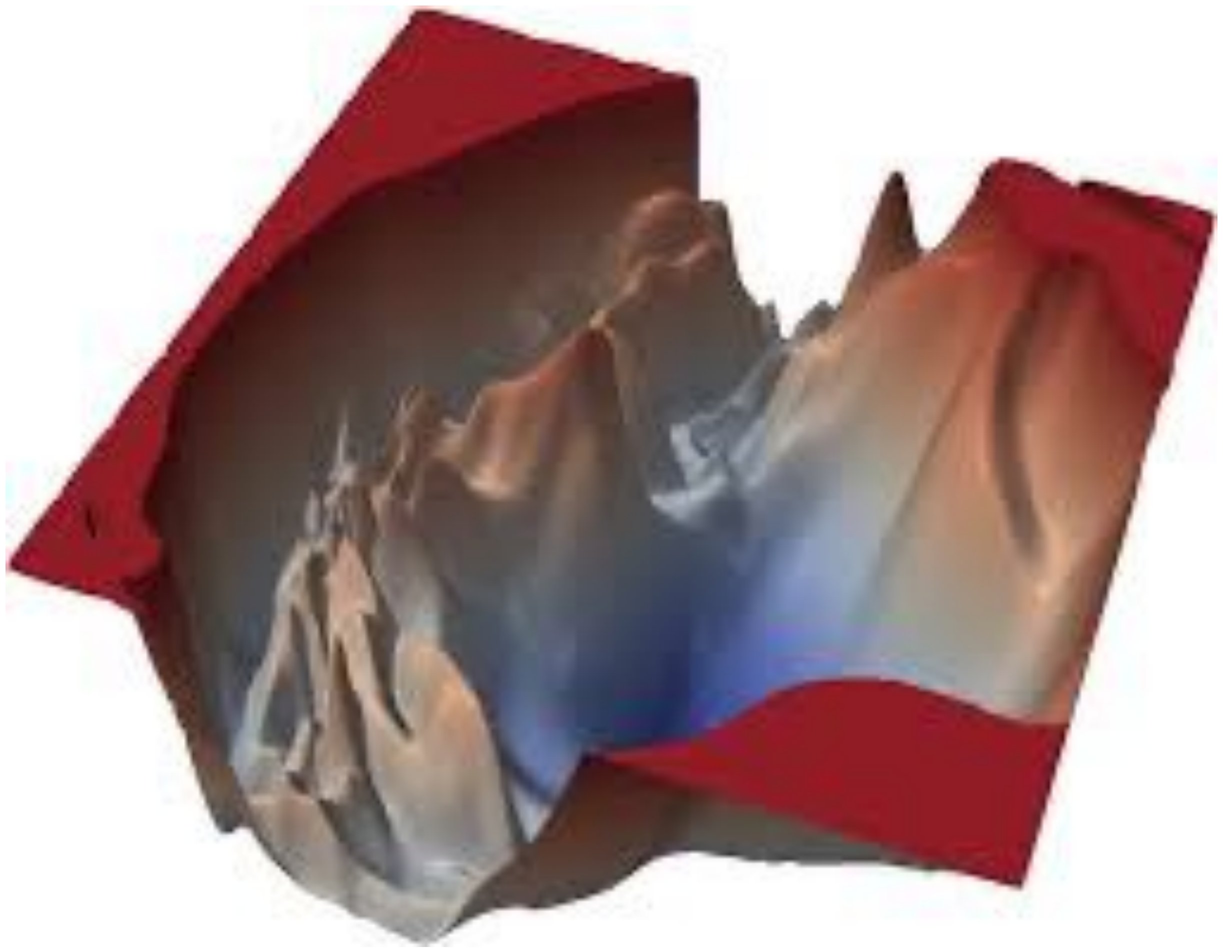
How to find these many parameters? Possible not to overfit !?

Training the network



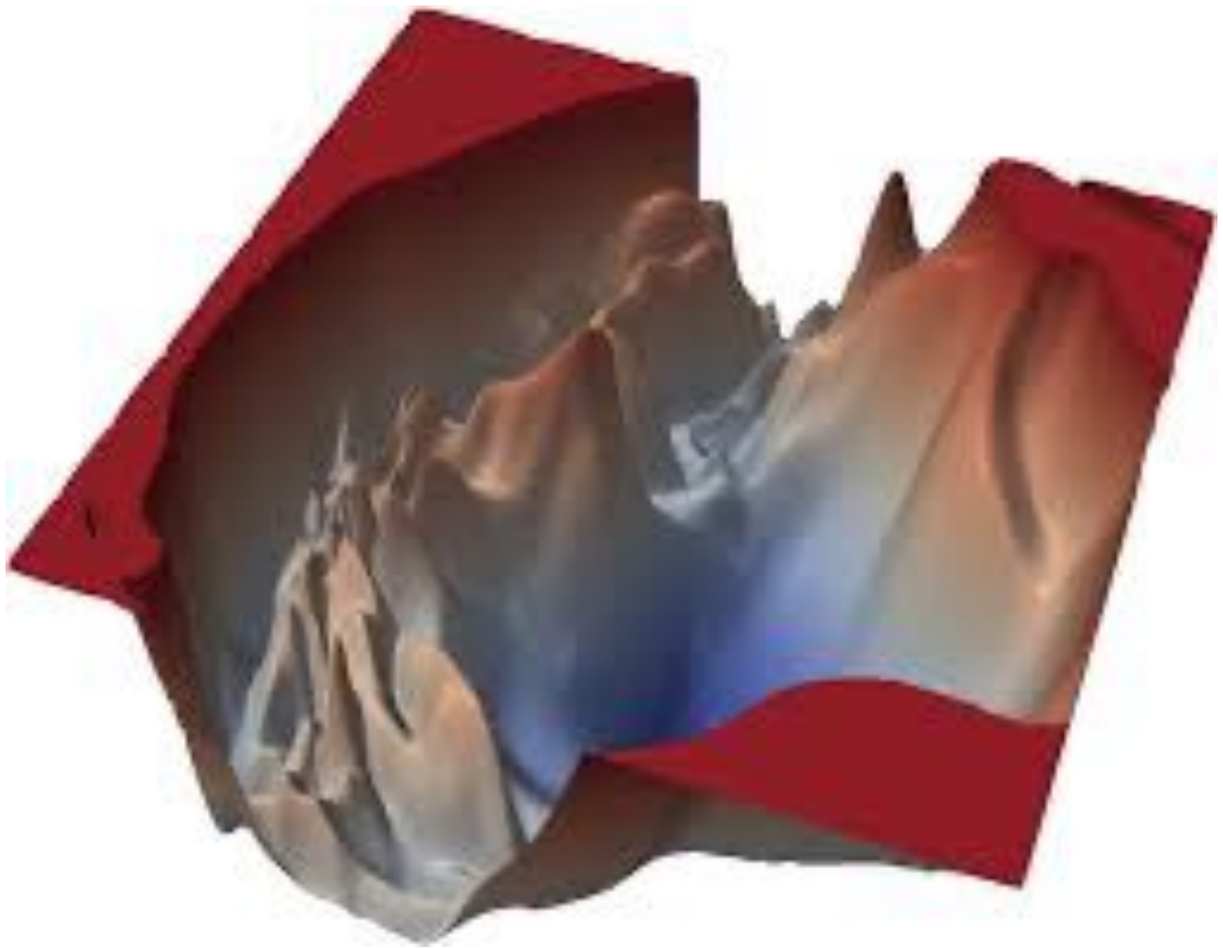
Finding the global minimum of the error function.
Where to start? Random initialization.
Then stochastic gradient descent to local minimum.
Regularization by drop-out procedure.

Training the network



Finding the global minimum of the error function.
Where to start? Random initialization.
Then stochastic gradient descent to local minimum.
Regularization by drop-out procedure.

Training the network



Finding the global minimum of the error function.
Where to start? Random initialization.
Then stochastic gradient descent to local minimum.
Regularization by drop-out procedure.

Involve a random product of noncommuting operations.

3. An ergodic theorem for the composition of noncommuting operations

Limit law for noncommutative operations

The Law of Large Numbers asserts that for i.i.d $X_1, X_2, X_3, \ldots$

$$\frac{1}{n}(X_1 + X_2 + \ldots + X_n) \to E[X_1].$$

Limit law for noncommutative operations

The Law of Large Numbers asserts that for i.i.d $X_1, X_2, X_3, \ldots$

$$\frac{1}{n}(X_1 + X_2 + \ldots + X_n) \to E[X_1].$$

Is there a similar law for

$$X_1 \cdot X_2 \cdot \ldots \cdot X_n?$$

Where X_i are noncommuting operations, for example elements of an arbitrary group.

Limit law for noncommutative operations

The Law of Large Numbers asserts that for i.i.d $X_1, X_2, X_3, ...$

$$\frac{1}{n}(X_1 + X_2 + ... + X_n) \to E[X_1].$$

Is there a similar law for

$$X_1 \cdot X_2 \cdot ... \cdot X_n?$$

Where X_i are noncommuting operations, for example elements of an arbitrary group.

Duke Math. J. 1954

LIMIT THEOREMS FOR NON-COMMUTATIVE OPERATIONS. I.

By Richard Bellman

1. **Introduction.** In this paper a start is made in the construction of a general theory involving the limiting behavior of systems subjected to non-commutative effects.

The classical central limit theorem states that under certain assumptions

The metric category

Let X be a metric space. $f : X \to X$ is *nonexpansive* if

$$d(f(x), f(y)) \leq d(x, y)$$

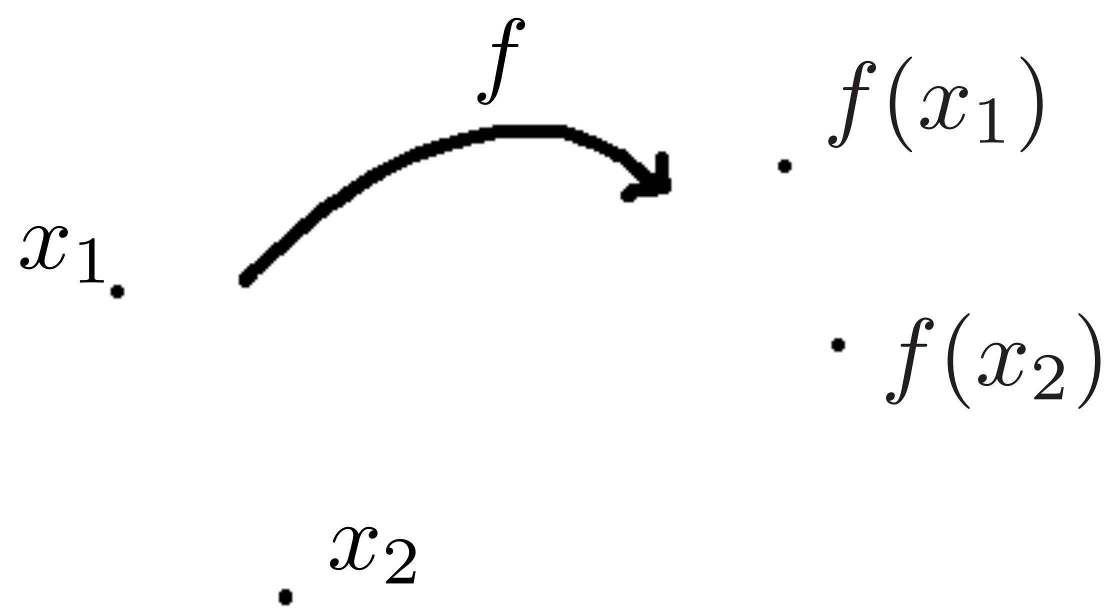
for all $x, y \in X$.

The metric category

Let X be a metric space. $f : X \to X$ is *nonexpansive* if

$$d(f(x), f(y)) \leq d(x, y)$$

for all $x, y \in X$.



Ex: Compositions; ISOMETRIES.

Nonexpanding maps appear in many contexts

Geometry: Riemannian geometry, Banach spaces, etc

Linear algebra / Lie groups, operator theory, diffeomorphisms

Complex analysis, group theory, cone maps, …

Nonexpanding maps appear in many contexts

Geometry: Riemannian geometry, Banach spaces, etc

Linear algebra / Lie groups, operator theory, diffeomorphisms

Complex analysis, group theory, cone maps, …

and certain **neural networks.**

Nonexpanding maps appear in many contexts

Geometry: Riemannian geometry, Banach spaces, etc

Linear algebra / Lie groups, operator theory, diffeomorphisms

Complex analysis, group theory, cone maps, …

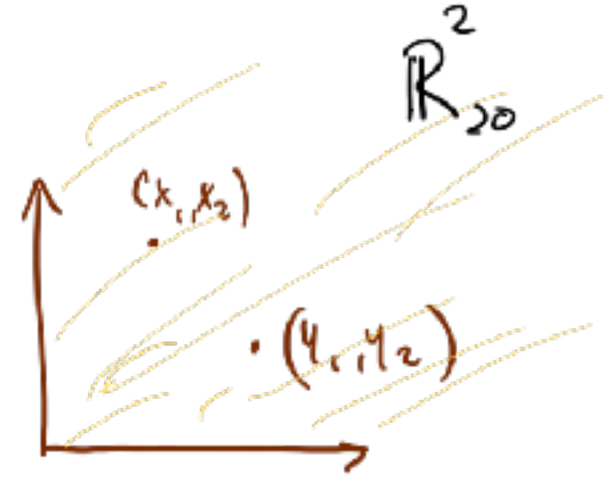
and certain **neural networks.**

A.K. *From liner to metric functional analysis,* PNAS 2021

Metrics

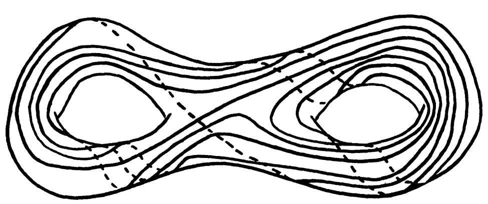
Thompson metric

$$d(x,y) = \max\{\log \max_i \frac{x_i}{y_i}, \log \max_i \frac{y_i}{x_i}\}$$



The Thurston asymmetric distance on Teichmüller space

$$L(x,y) = \log \sup_{\alpha \in \mathcal{S}} \frac{l_y(\alpha)}{l_x(\alpha)}.$$



x,y represent metrics on a surface,
and homeomorphisms are isometries.

An example

An observation in D. Blackwell, *Discounted Dynamic Programming*, 1965:

Let S be a set and $B(S)$ the space of functions on S, equipped with sup-norm.

Let $T : B(S) \to B(S)$ such that

- $f \leq g$ implies $Tf \leq Tg$
- $T(f + C) = Tf + \beta C$ certain $\beta \in (0, 1]$ all constants C

Then $||Tf - Tg|| \leq \beta ||f - g||$ for all f, g.

An example

An observation in D. Blackwell, *Discounted Dynamic Programming*, 1965:

Let S be a set and $B(S)$ the space of functions on S, equipped with sup-norm.

Let $T : B(S) \to B(S)$ such that

- $f \leq g$ implies $Tf \leq Tg$
- $T(f + C) = Tf + \beta C$ certain $\beta \in (0, 1]$ all constants C

Then $||Tf - Tg|| \leq \beta||f - g||$ for all f, g.

Example:

$$Tf(s) = \max_a\{r_a(s) + \sum_t p_{st}(a)f(t)\}$$

A noncommutative ergodic theorem

Let (X, d) be a weak metric space, i.e. $d(x, x) = 0$ and

$$d(x, y) \leq d(x, z) + d(z, y).$$

Let g_i be i.i.d. selected nonexpansive maps $X \to X$.

Let

$$u(n, \omega) := g_1 \circ g_2 \circ g_3 \circ \ldots \circ g_n.$$

Assume everything measurable and $\mathbb{E}[d(x, g(x))] < \infty$.

A noncommutative ergodic theorem

Let

$$u(n,\omega) = g_1 g_2 g_3 \ldots g_n$$

be an integrable ergodic cocycle of nonexpansive maps of X.

Theorem (K.-Ledrappier, Ann Prob '06 ; Gouëzel-K., JEMS '20)
For a.e. ω there exists a metric functional $h = h^{\omega}$ s.t.

$$\lim_{n\to\infty} -\frac{1}{n} h(u(n,\omega)x) = \lim_{n\to\infty} \frac{1}{n} d(x, u(n,\omega)x).$$

A noncommutative ergodic theorem

Let

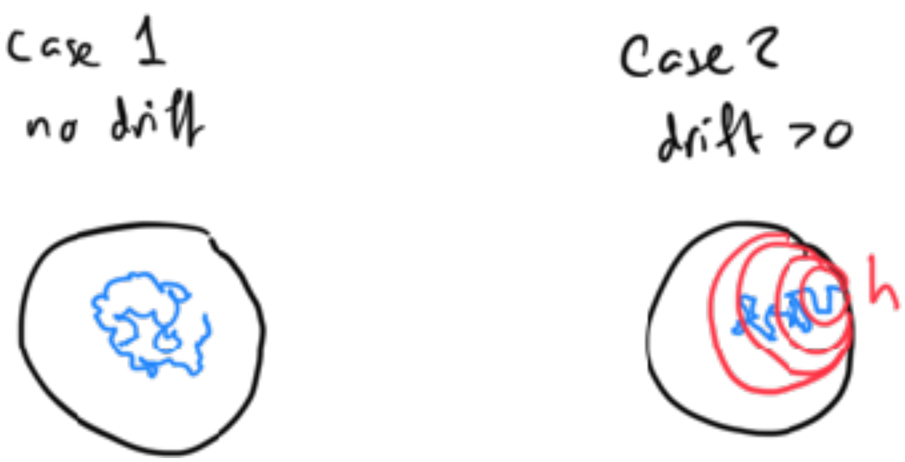
$$u(n,\omega) = g_1 g_2 g_3 \ldots g_n$$

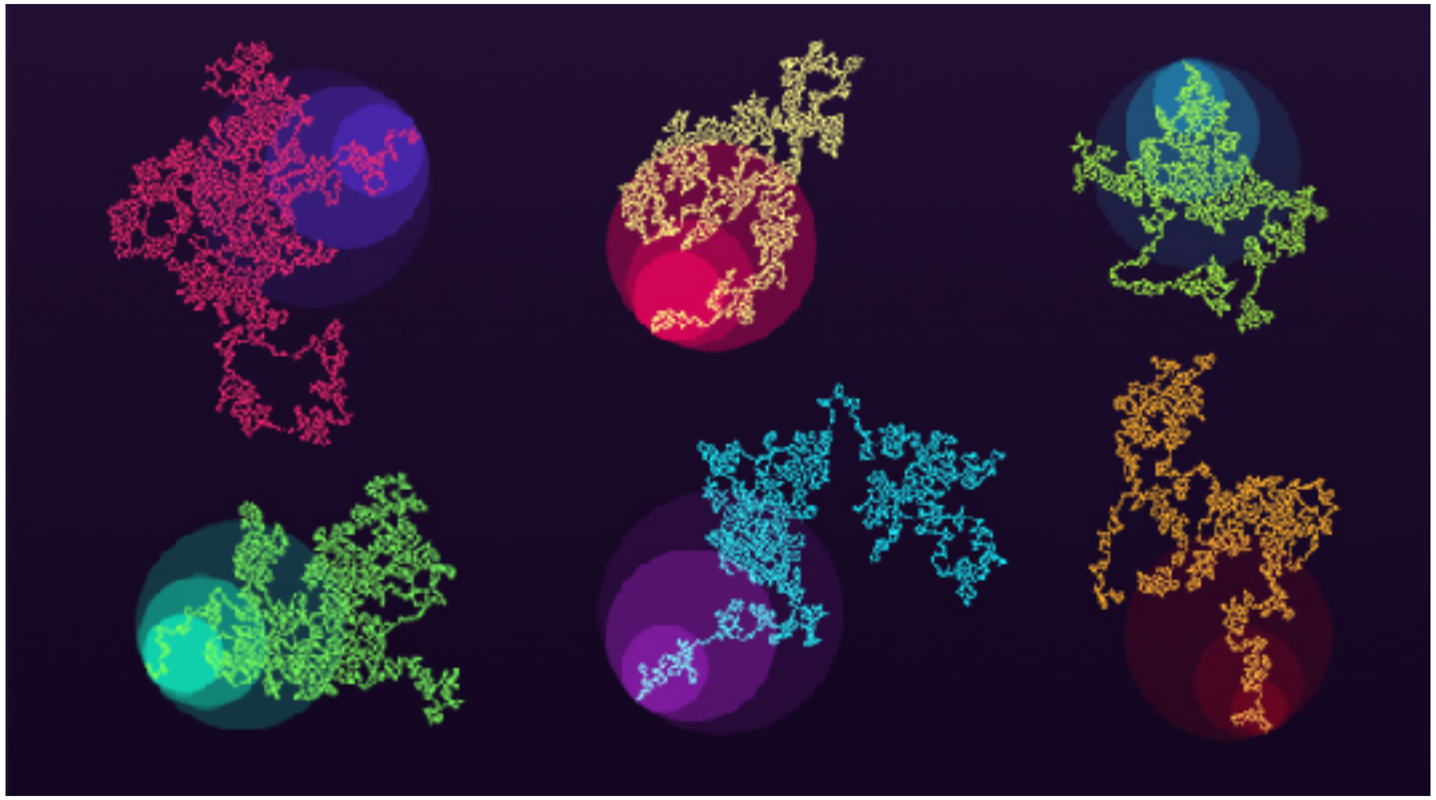
be an integrable ergodic cocycle of nonexpansive maps of X.

Theorem (K.-Ledrappier, Ann Prob '06 ; Gouëzel-K., JEMS '20)
For a.e. ω there exists a metric functional $h = h^\omega$ s.t.

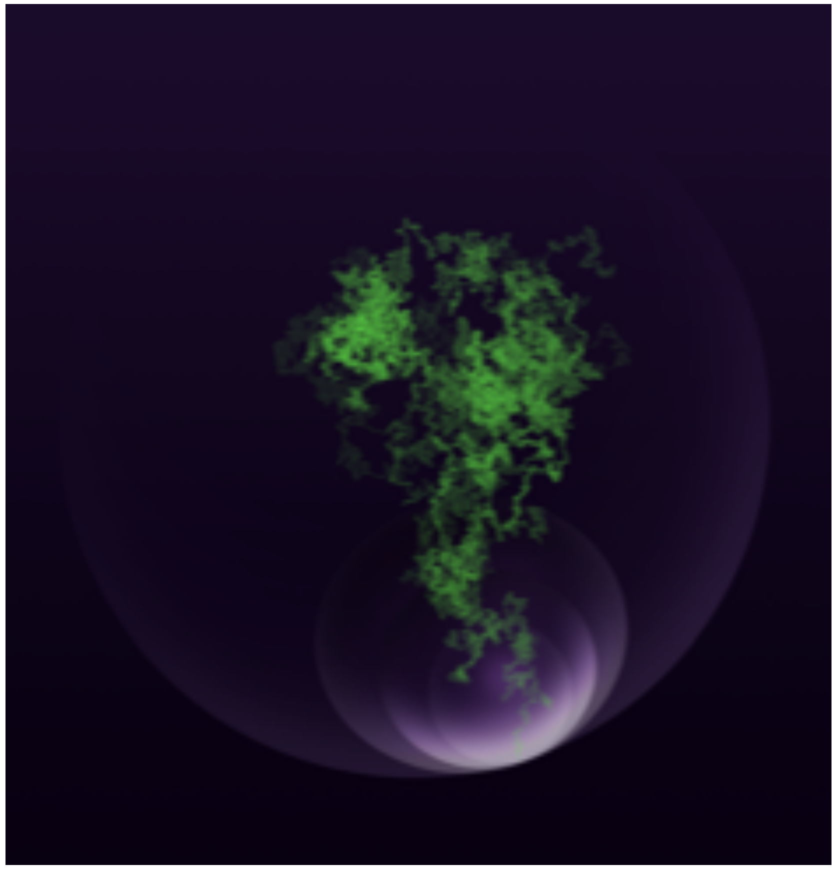
$$\lim_{n\to\infty} -\frac{1}{n} h(u(n,\omega)x) = \lim_{n\to\infty} \frac{1}{n} d(x, u(n,\omega)x).$$

Kingman's subadditive ergodic theorem



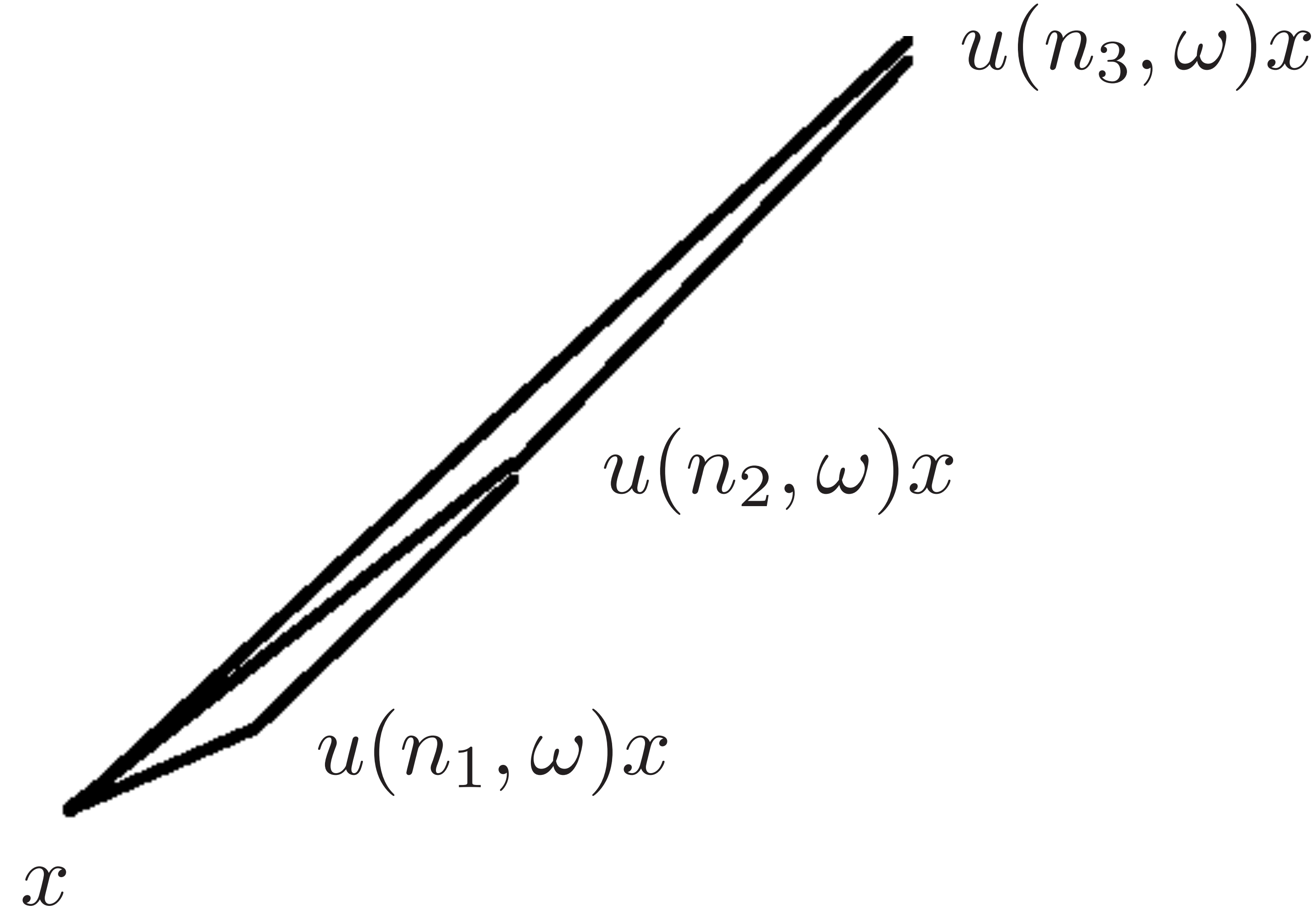


Proof based on substantial refinement of the subadditive ergodic theorem,



Rough idea of the proof

There exist good times n_i when the subadditive cocycle $d(x, u(n,\omega)x)$ is nearly additive.



Take a weak limit point of $h_{u(n_i,\omega)x}$ for these special orbit points. QED.

Special cases of the noncommutative ergodic theorem

1. Oseledets multiplicative ergodic theorem

When applied to $X = Pos$ and $G = GL(n, \mathbb{R})$

2. Random mean ergodic theorems (Ulam-von Neumann,…)

When applied to X =Hilbert space and $g_i(x) = U_i x + v$

3. Operator multiplicative ergodic theorems (Ruelle,…)

4. Muliplicative ergodic theorem for CAT(0)-spaces (K.-Margulis)

5. Random walks on groups and Brownian motion (with Ledrappier, 2007)

6. A Furstenberg-Khasminskii type formula (with Ledrappier, 2007)

3. Deep learning: metric frameworks

Providing a metric and dynamical framework

Avelin, B, Karlsson, A, Deep limits and a cut-off phenomenon for neural networks,
Journal of Machine Learning Research, 2022
NeurIPS 2022 presentation

In this paper we:

- **Display invariant metrics or associated metric spaces on which the layer maps act by nonexpansive maps.**
- **Apply the noncommutative ergodic theorem**
- **Found evidence for a cut-off phenomenon**

Instances of recent deep learning literature

**“The Principles of Deep Learning Theory:
An Effective Theory Approach to Understanding Neural Networks”
Daniel A. Roberts and Sho Yaida
based on research in collaboration with
Boris Hanin
Cambridge Univ. Press, 2022**

“Beyond illuminating the properties of networks at the start of training, the analysis of random neural networks can reveal a great deal about networks after training as well.” Boris Hanin, 2021

Benoit Dherin, Michael Munn, Mihaela Rosca, David G.T. Barrett, Why neural networks find simple solutions: the many regularizers of geometric complexity, NeurIPS (2022)

"Direct" metrics

Positive models

$$T(x) = \sigma(Ax + b)$$

where $A_{ij} \geq 0$, $b_i \geq 0$ and $\sigma =$ sigmoid or ReLU.

Thompson or Blackwell

“Direct” metrics

Positive models

$$T(x) = \sigma(Ax + b)$$

where $A_{ij} \geq 0$, $b_i \geq 0$ and $\sigma =$ sigmoid or ReLU.

Thompson or Blackwell

Residual neural networks, “ResNets”

$$T(x) = W^T \sigma(Wx + b)$$

where $||W|| \leq 1$ and $\sigma =$ one of the standard.

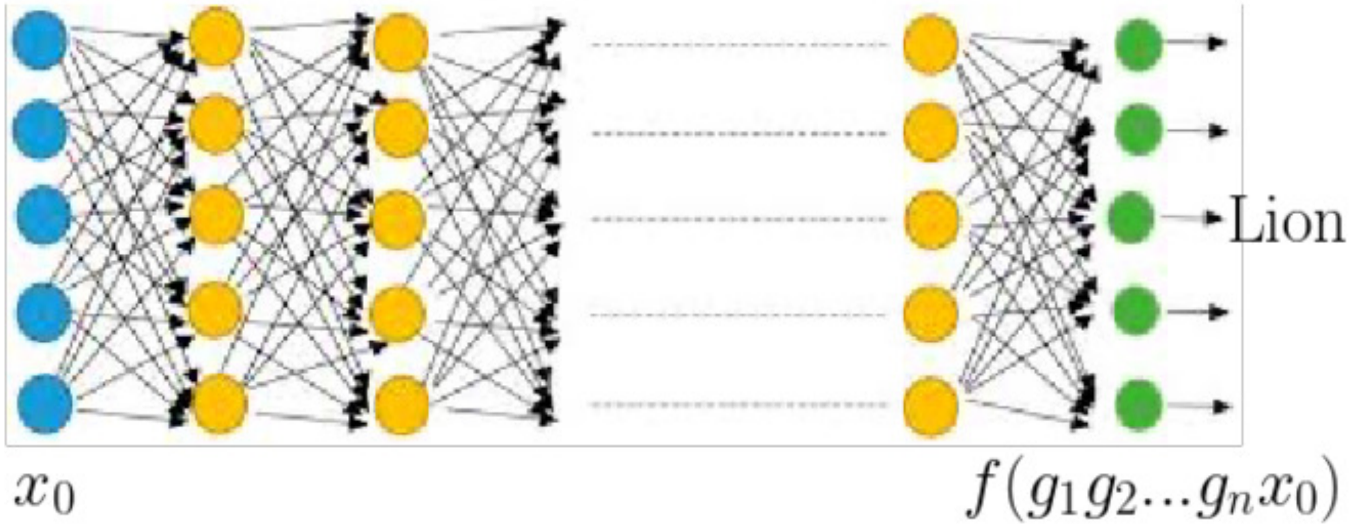
The norm

“Associated” metrics

Distances on space of metrics

Ex. $$D(d_1, d_2) = \log\left(\max\left\{\sup_{x\neq y}\frac{d_2(x,y)}{d_1(x,y)}, \sup_{x\neq y}\frac{d_1(x,y)}{d_2(x,y)}\right\}\right).$$

Distance on the set of decision functions

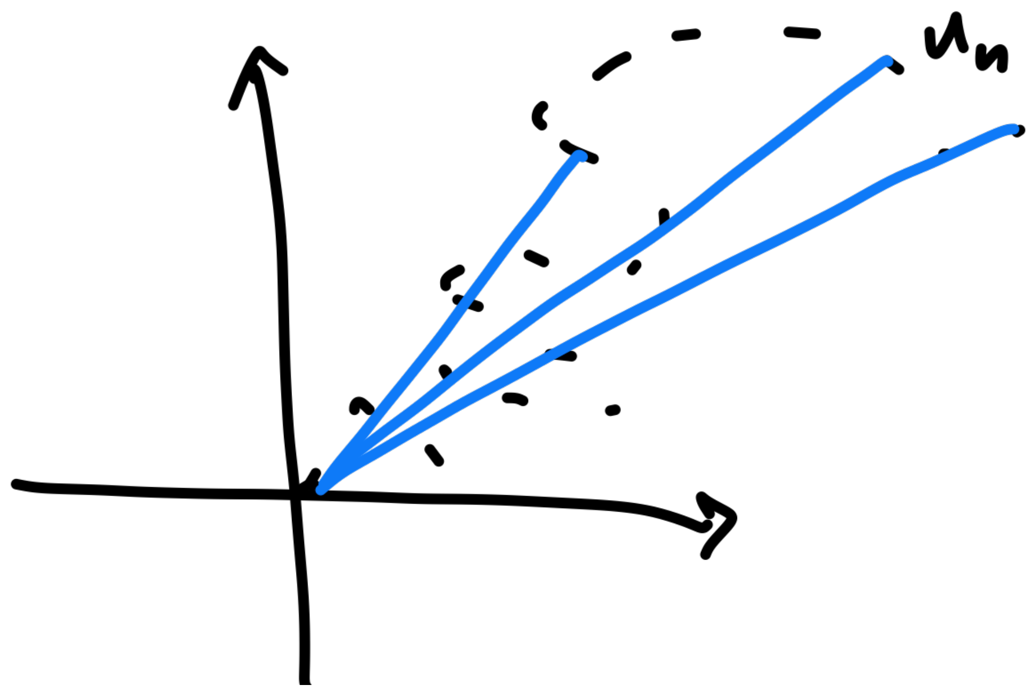


$$=: f_n(x_0)$$

Sample result, “ResNets”

Theorem(Avelin-K. ’20) Given layer maps $x \overset{U}{\mapsto} W\sigma(Ax+b)$ where W, A, b are selected iid (or stationary) with $||W||, ||A|| \leq 1$, and $\sigma(t) = \max\{0, t\}$. Then there is a random vector v such that

$$\frac{1}{n} U_1 U_2 \ldots U_n x_0 \to v. \qquad n \to \infty$$



Sample results

Take $X = [-1,1]^N$ and activation function $\sigma(t) = \tanh(t)$ and invertible weights A. As before $U(x) = \sigma(Ax + b)$. Select at random say with finite support.

Theorem (Avelin-K. 21) There is a well-defined maximal exponential rate separating two nearby points. And when it is strictly positive, there is moreover a random point x whose neighborhood is stretched with this maximal rate.

$$\lim_{n\to\infty} \left(\sup_{x\neq y} \frac{||u_n(x) - u_n(y)||}{||x-y||} \right)^{1/n} = e^{\lambda}$$

Summary and outlook

- **Main idea in deep learning, hence AI, is mathematical, simple to understand**
- **Lack of theoretical understanding contributes to the main problems of AI**
- **Products of noncommuting operations is a feature of deep learning, and several other scientific contexts, as are metrics**
- **There is a general noncommutative ergodic theorem**

Summary and outlook

- Main idea in deep learning, hence AI, is mathematical, simple to understand
- Lack of theoretical understanding contributes to the main problems of AI
- Products of noncommuting operations is a feature of deep learning, and several other scientific contexts, as are metrics
- There is a general noncommutative ergodic theorem

THANK YOU FOR YOUR ATTENTION!