I. Mysteries of deep learning

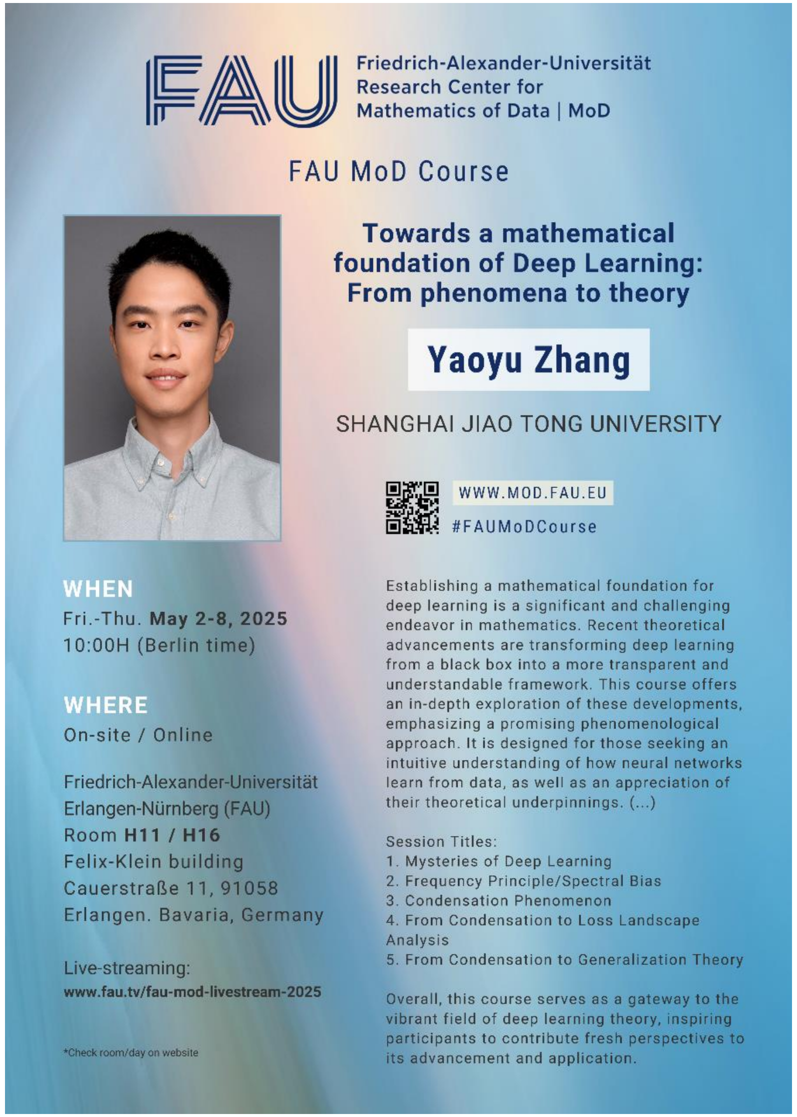
Yaoyu Zhang

Institute of Natural Sciences & School of Mathematical Sciences

Shanghai Jiao Tong University

FAU MoD Course

Deep learning is no longer a black-box



Towards a Mathematical Foundation of Deep Learning: From Phenomena to Theory

Date

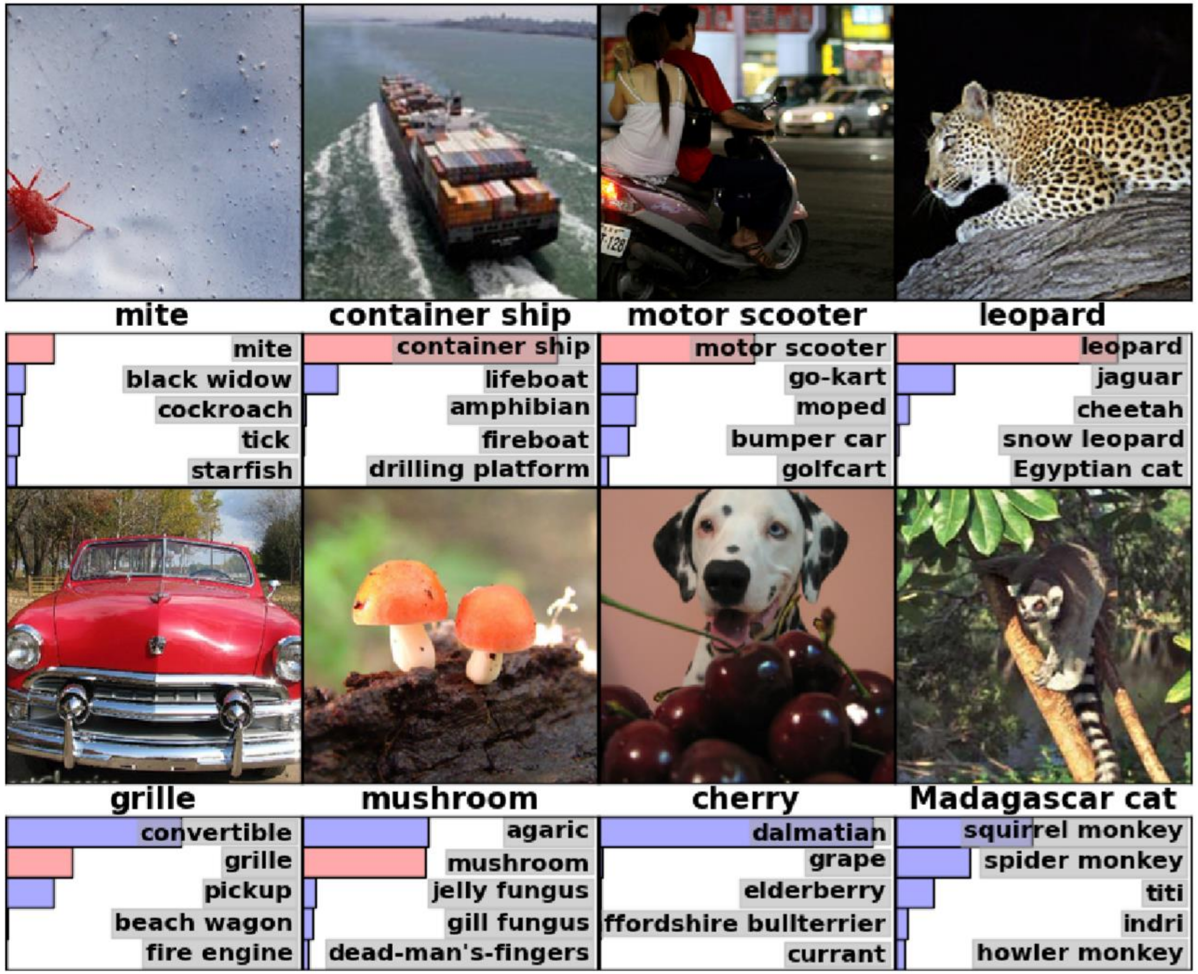
Fri. – Thu. May 2 – 8, 2025

Session Titles

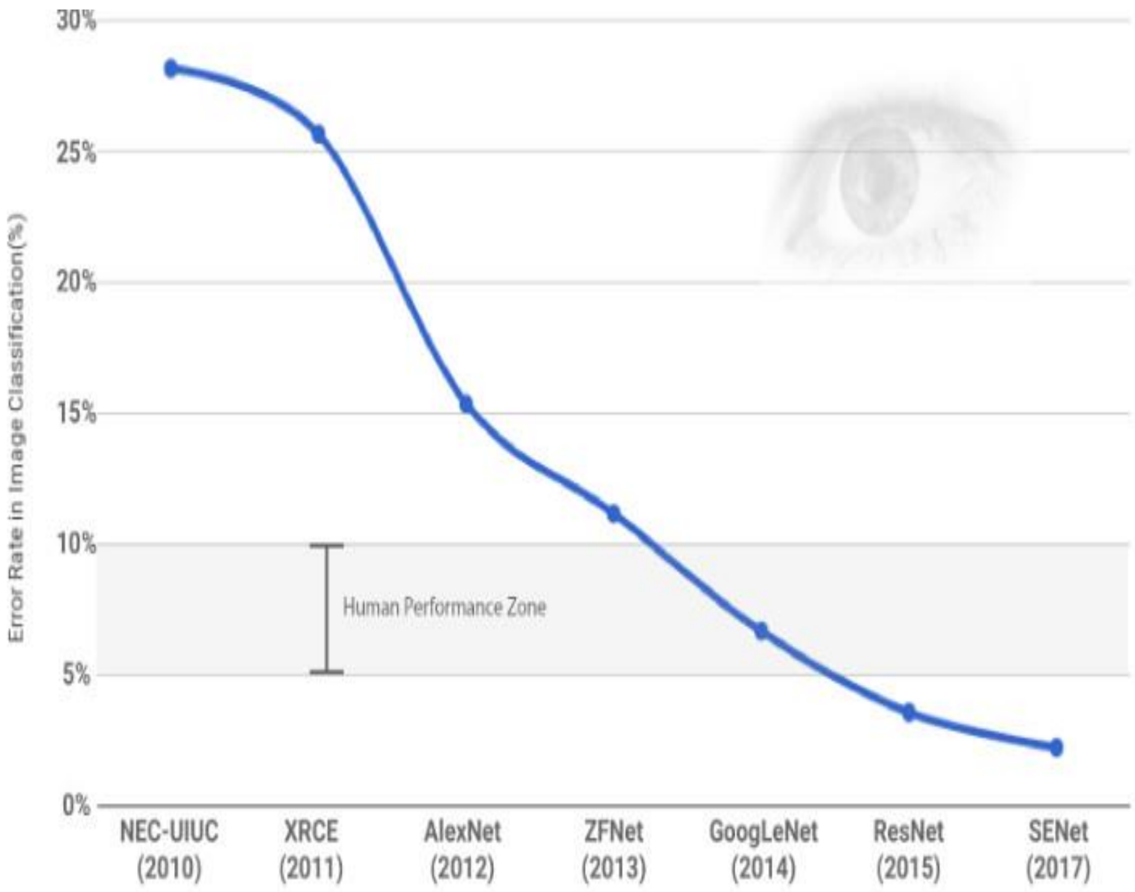
1. Mysteries of Deep Learning
2. Frequency Principle/Spectral Bias
3. Condensation Phenomenon
4. From Condensation to Loss Landscape Analysis
5. From Condensation to Generalization Theory

Unimaginable achievements of AI

Image recognition



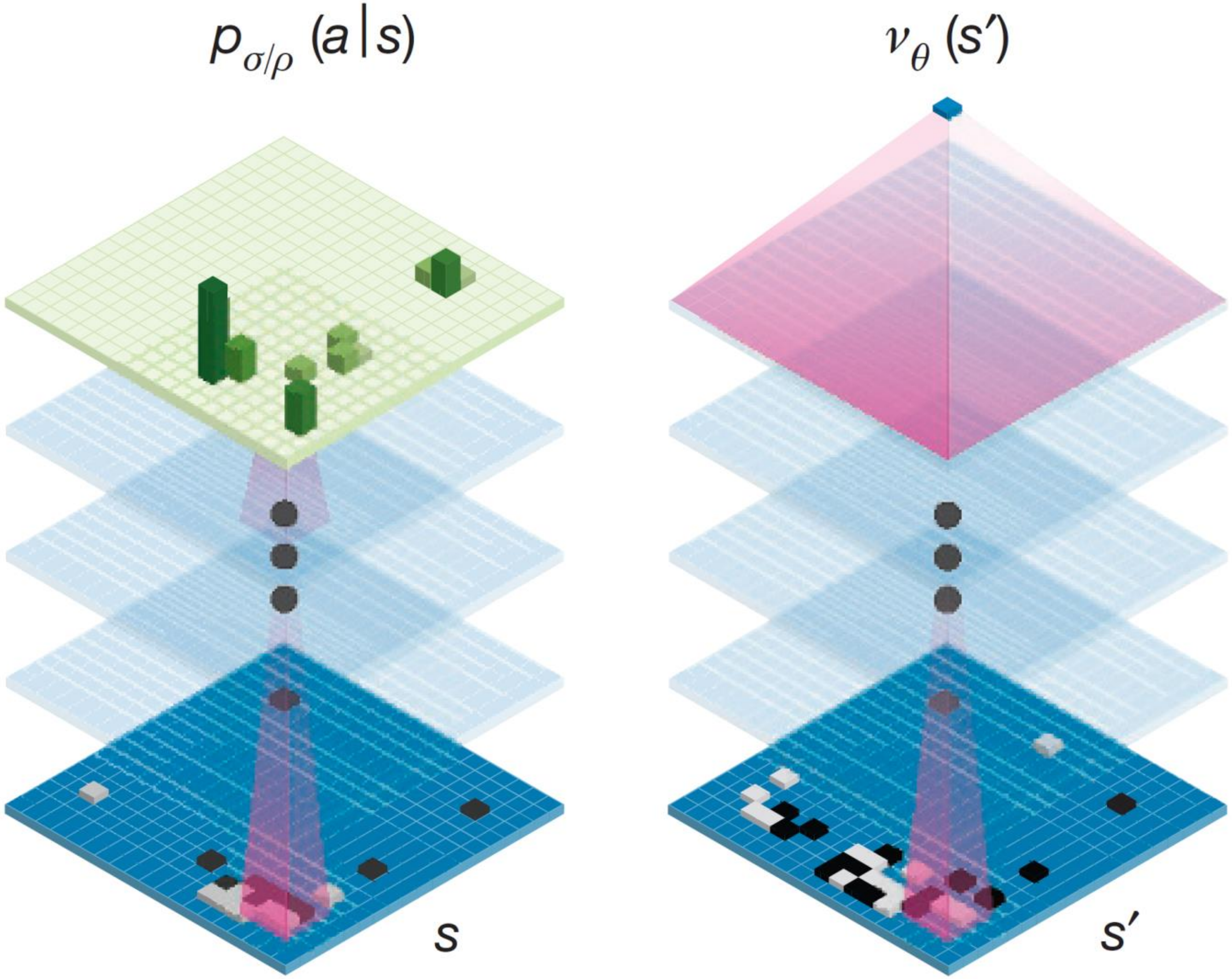
Krizhevsky, et al, 2012



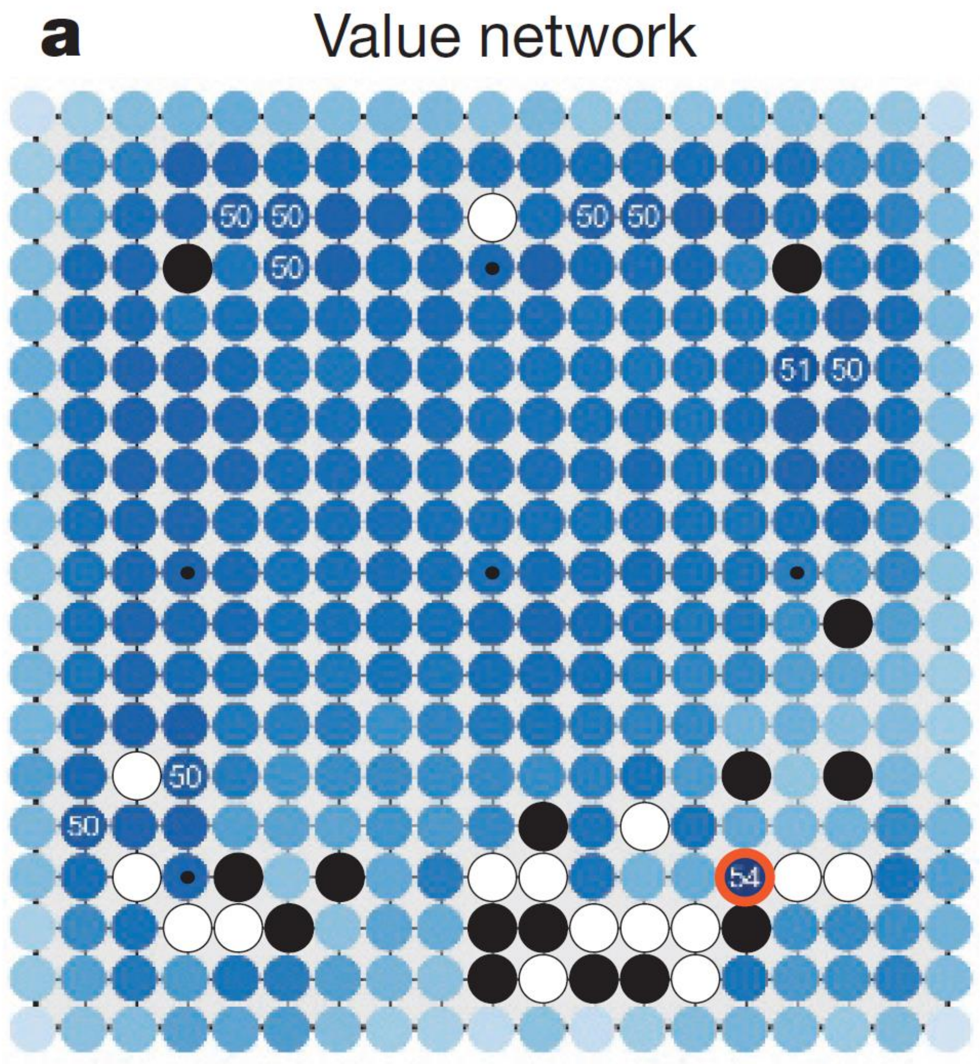
https://www.linkedin.com/pulse/must-read-path-breaking-papers-image-classification-muktabh-mayank

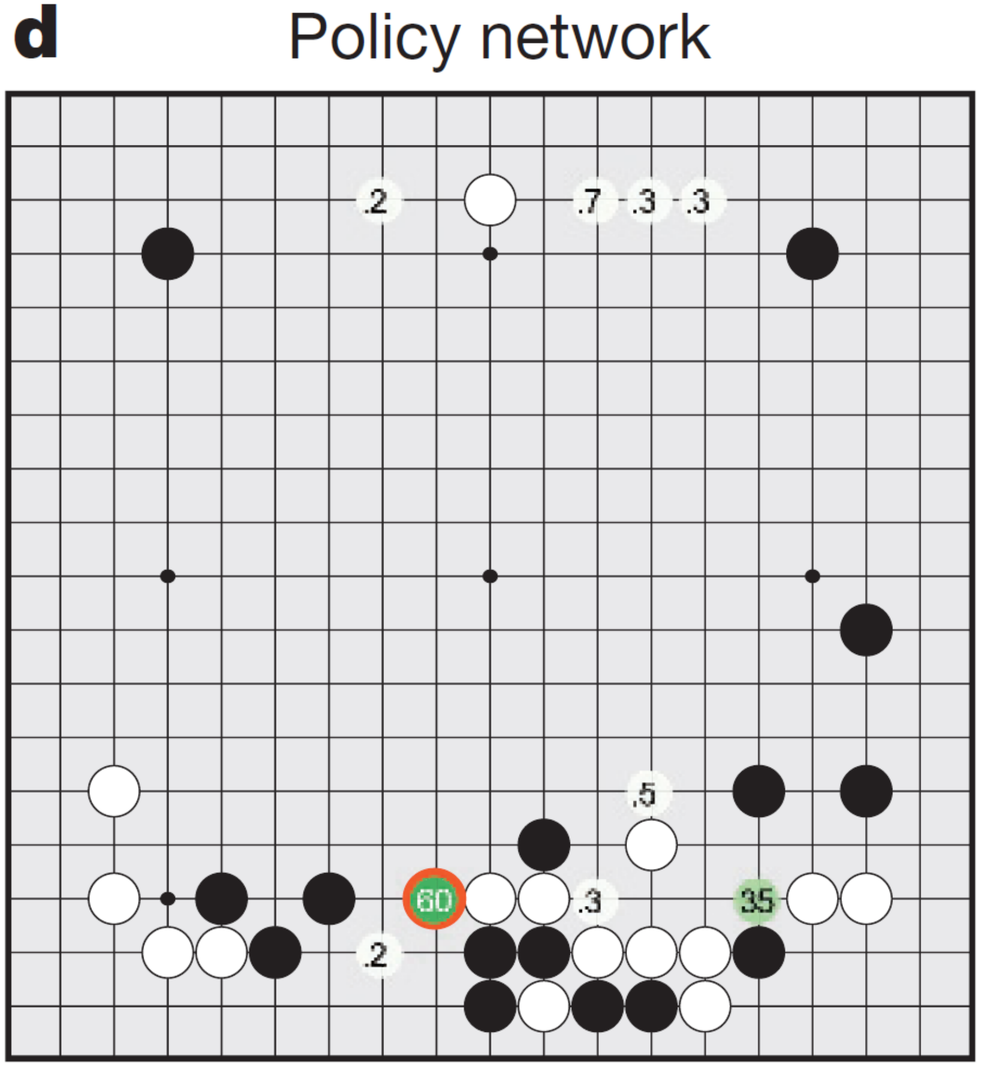
Go playing

Silver, et al, 2017



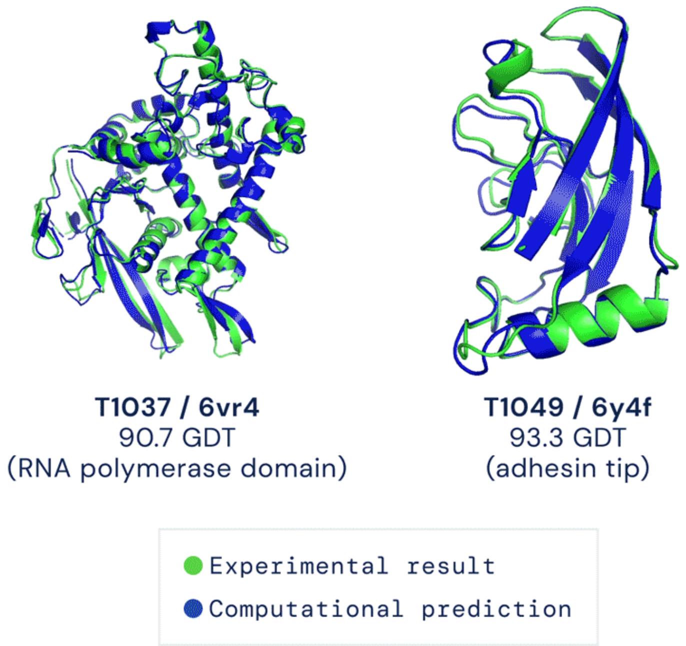
Go playing





Silver, et al, 2017

Protein structure prediction



The Nobel Prize in Chemistry 2024

David Baker

"for computational protein design"

Demis Hassabis

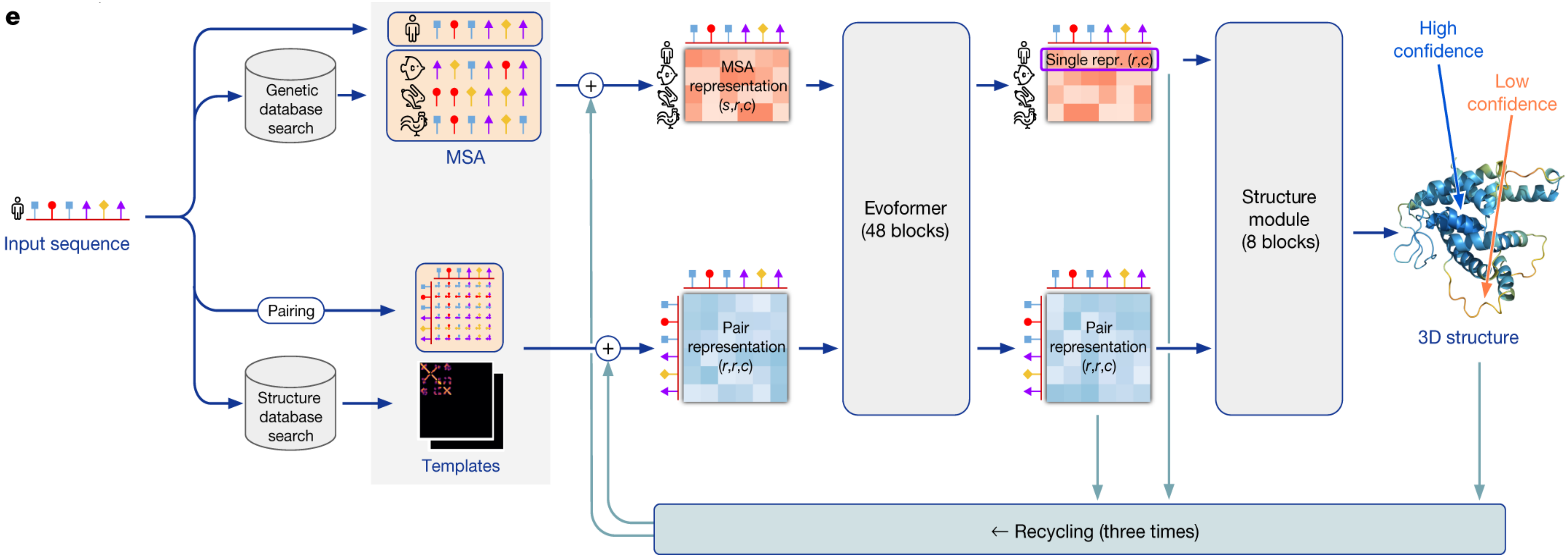
"for protein structure prediction"

John Jumper

"for protein structure prediction"

https://deepmind.com/blog/article/alphafold-a-solution-to-a-50-year-old-grand-challenge-in-biology

Protein structure prediction



Jumper et al., 2021

Image&video generation

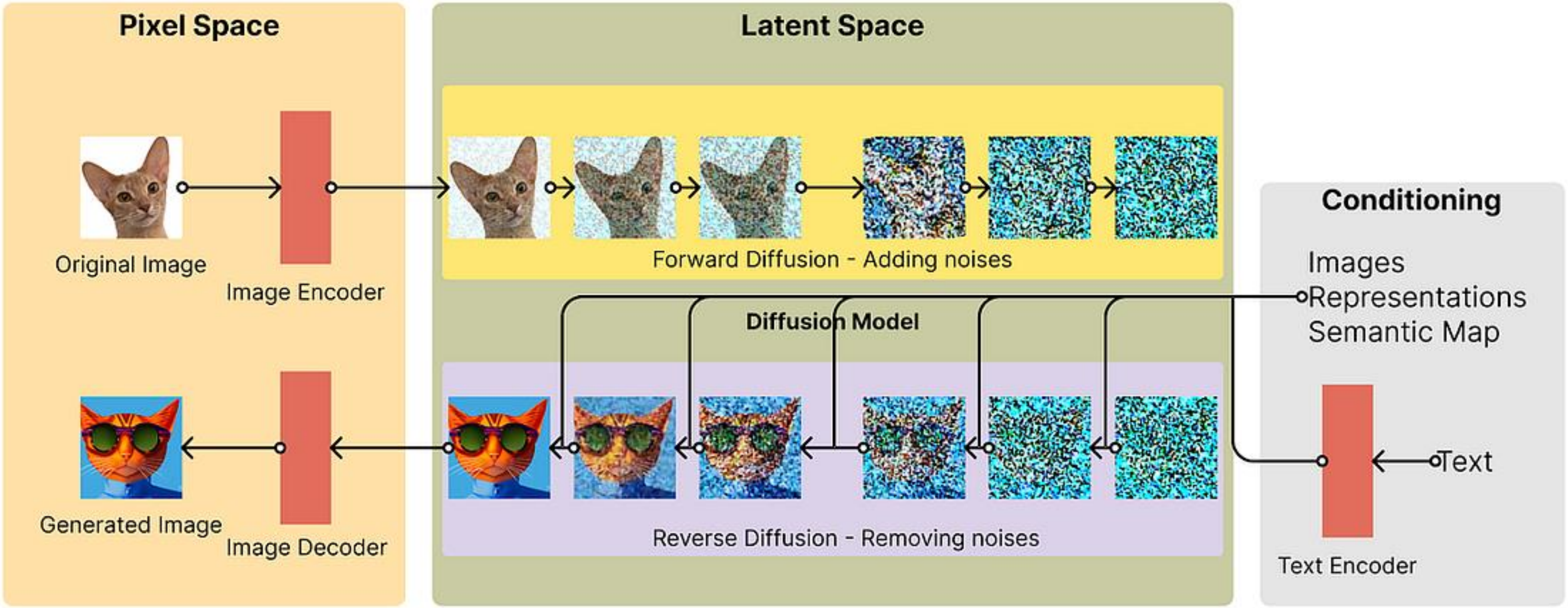
/v5_upscale 17 hrs ago
Woman with a Cactus Hat artwork by Edmund Dulac and Christian Schloe
yaros89 ...

/v5_upscale 17 hrs ago
Transparent Flask Water Bottle
alexzz ...

https://www.midjourney.com/

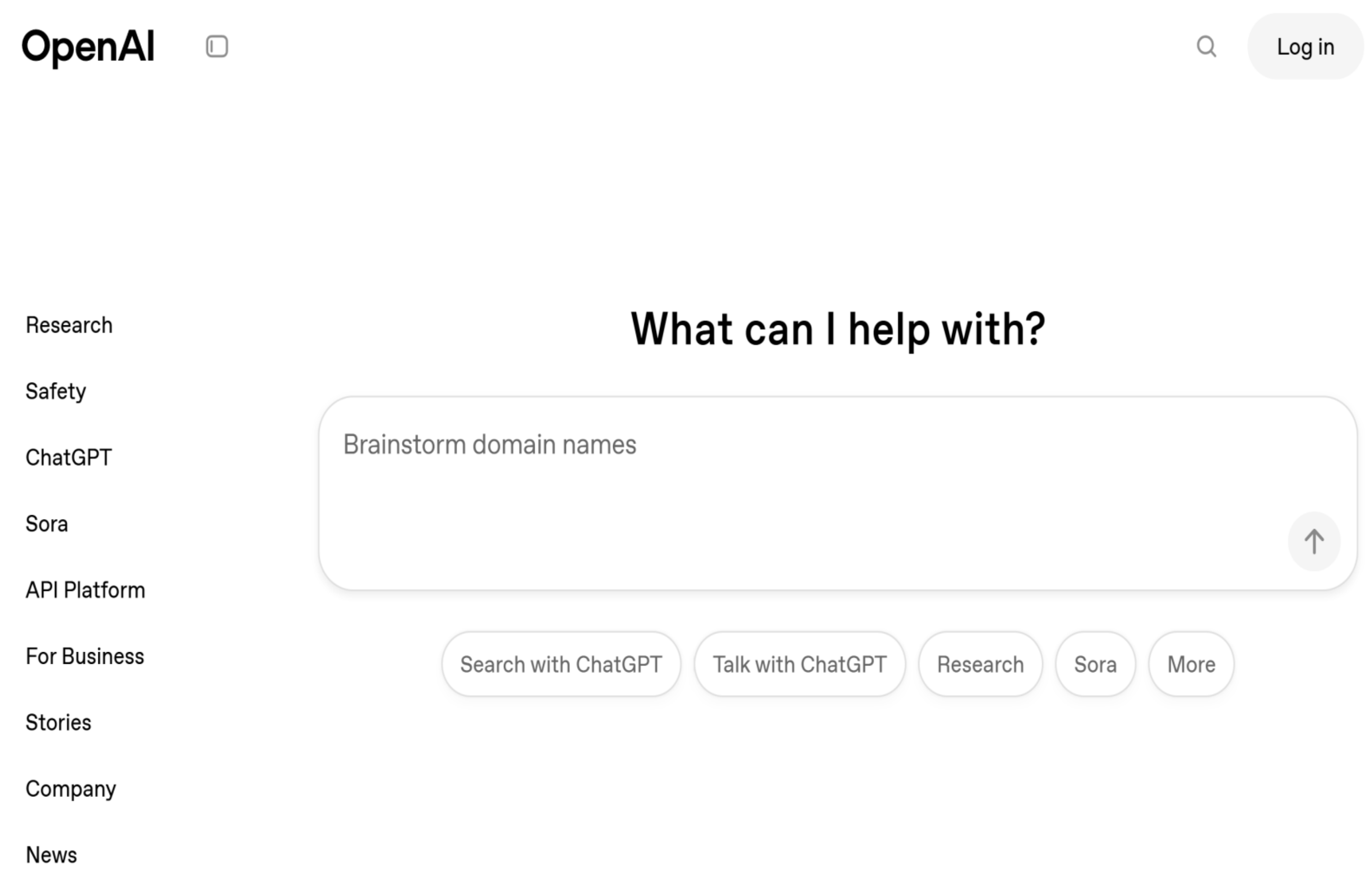
https://openai.com/index/sora/

Diffusion model

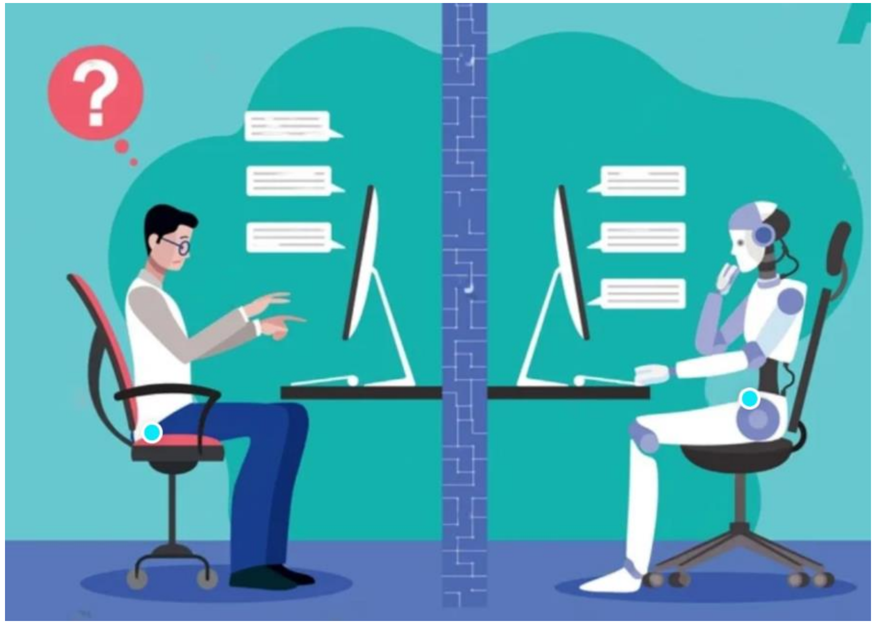


https://bootcamp.uxdesign.cc/how-stable-diffusion-works-explained-for-non-technical-people-be6aa674fa1d

Large language model—milestone towards AGI



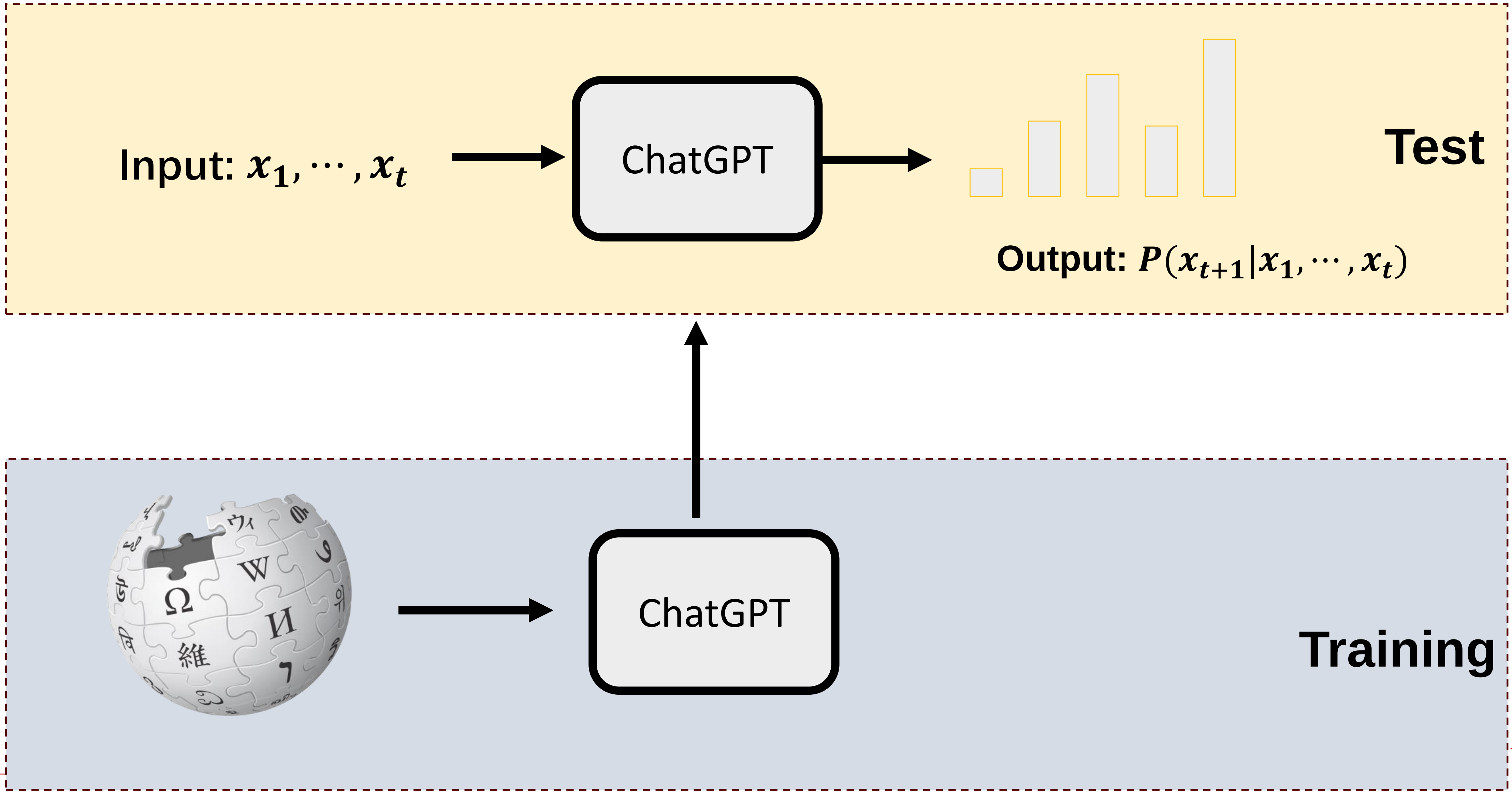
Turing test?



https://openai.com/

https://cshub.in/what-is-turing-test/

Next token prediction



The engine of AI: deep learning

Empirical risk: $R_S(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} l(f(\boldsymbol{x}_i, \boldsymbol{\theta}), \boldsymbol{y}_i)$

Model: $f(\boldsymbol{x}, \boldsymbol{\theta})$

Data: $S = \{(x_i, y_i)\}_{i=1}^{n}$

Just this?

Common Models:

Linear models: polynomial models, random feature models, $\cdots$

Neural networks: fully-connected, convolutional, ResNet, Transformer, $\cdots$

Common loss function:

Mean-squared error (l2) loss: $l(y, y') = \|y - y'\|_2^2$,

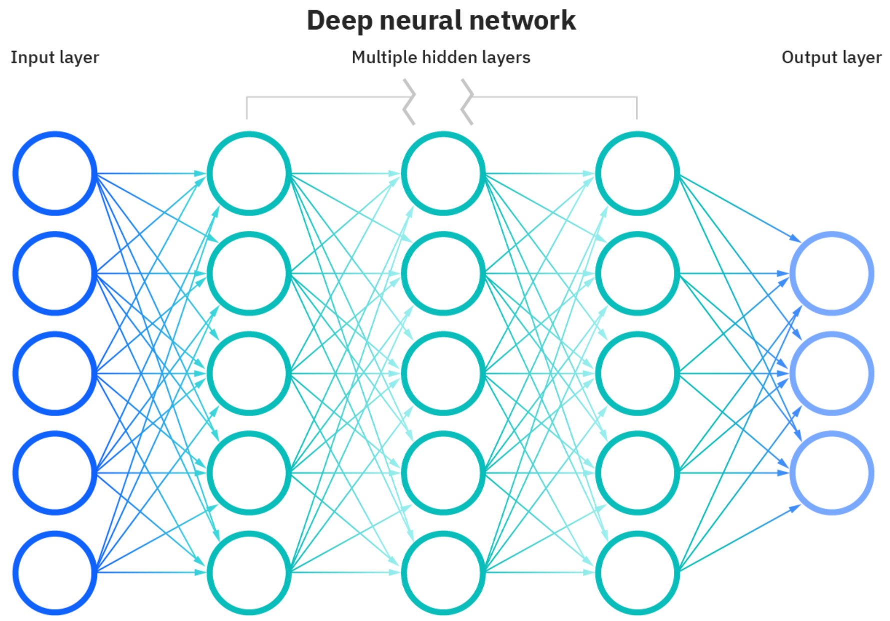
Cross entropy, Hinge loss, …

Common training algorithm:

Gradient decent (GD): $\theta^{t+1} = \theta^{t} - \eta \nabla R_S(\theta^{t})$,

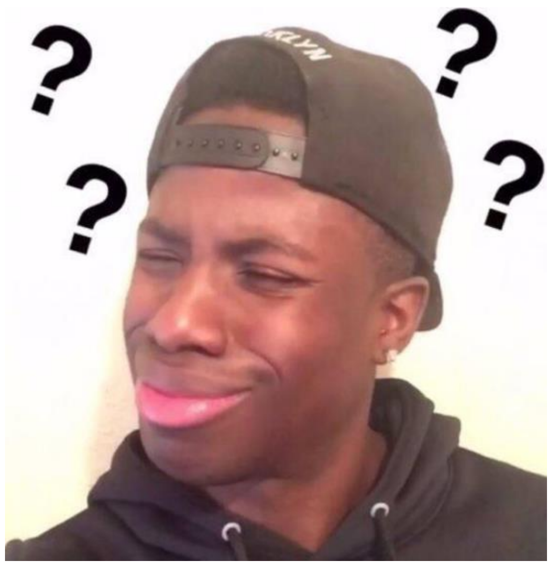
Stochastic gradient descent (SGD), Adam, …

Deep learning remains a “black” technology



$$\theta := \left(W^{[1]}, b^{[1]}, \ldots, W^{[L]}, b^{[L]}\right)$$

$$f_\theta^{[l]}(x) := \sigma(W^{[l]} f_\theta^{[l-1]}(x) + b^{[l]})$$



AlphaGo,
AlphaFold,
ChatGPT,
SORA,
…

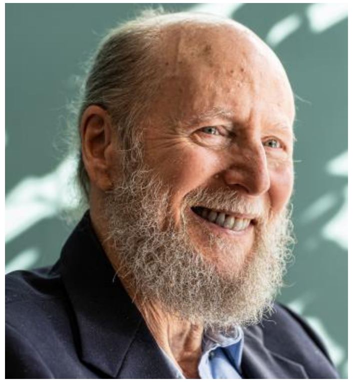
https://www.ibm.com/cloud/learn/neural-networks

Which are black technologies?

- Synthetic diamond
- Atomic bomb
- The Apollo Program
- ChatGPT
- Quantum computer
- 0.1 light-speed spaceship

Bitter lesson for deep learning theory

The Bitter Lesson for AI research



The Bitter Lesson

Rich Sutton

March 13, 2019

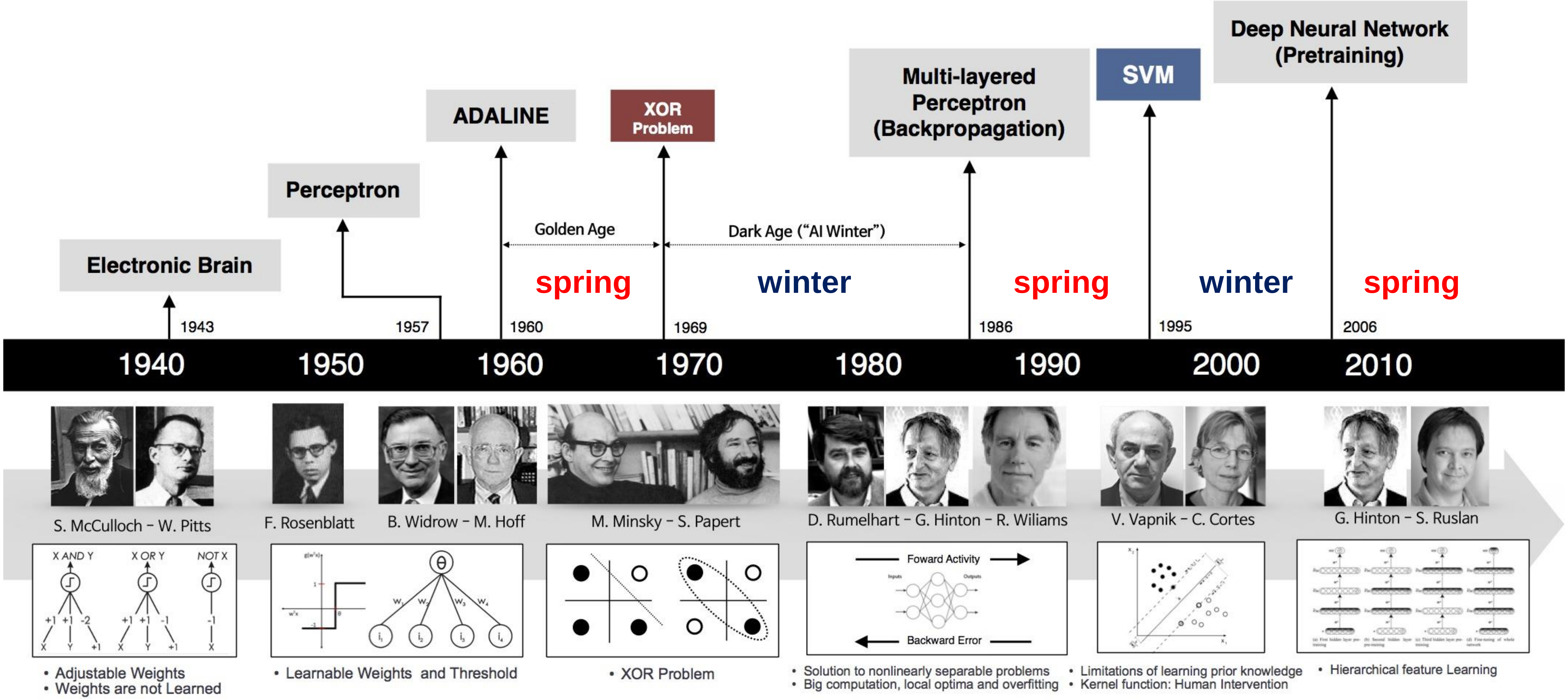
The biggest lesson that can be read from 70 years of AI research is that general methods that leverage computation are ultimately the most effective, and by a large margin. The ultimate reason for this is Moore's law, or rather its generalization of continued exponentially falling cost per unit of

- “The biggest lesson that can be read from **70 years of AI research** is that **general methods that leverage computation** are ultimately the most effective, and by a large margin.”

Leverage computation (learning) instead of human knowledge

http://www.incompleteideas.net/IncIdeas/BitterLesson.html

Timeline of neural network development



https://beamandrew.github.io/deeplearning/2017/02/23/deep_learning_101_part1.html

Timeline of deep learning theory

- **1969 book of *Perceptrons*** (lead to the first winter)
- **1986 Backpropagation** (emergence of modern deep learning)
- **1989 Universal approximation theorem**
- **1995 Generalization puzzle proposed** (not well solved till now)

  The Vapnik-Jackel Bet (witnessed by Yann Lecun)
- **2017 Generalization puzzle demonstrated in SOTA settings**
- **2018 Neural Tangent Kernel** (lead to a surge in DL theory research)

  Frequency principle/Spectral bias

Despite 40 years of effort, framework for its math foundation yet to emerge

The bet on deep learning theory

The Vapnik-Jackel Bet in 1995

1. Jackel bets (one fancy dinner) that by March 14, 2000 people will understand
quantitatively why big neural nets working on large databases are not so bad.
(Understanding means that there will be clear conditions and bounds)

Vapnik bets (one fancy dinner) that Jackel is wrong.

But .. If Vapnik figures out the bounds and conditions, Vapnik still wins the bet.

****

2. Vapnik bets (one fancy dinner) that by March 14, 2005, no one in his right mind will use neural nets that are essentially like those used in 1995.

Jackel bets (one fancy dinner) that Vapnik is wrong

V. Vapnik 3/14/95

L.. Jackel 3/14//95

Witnessed by Y. LeCun 3/14/95

From Lecun's talk

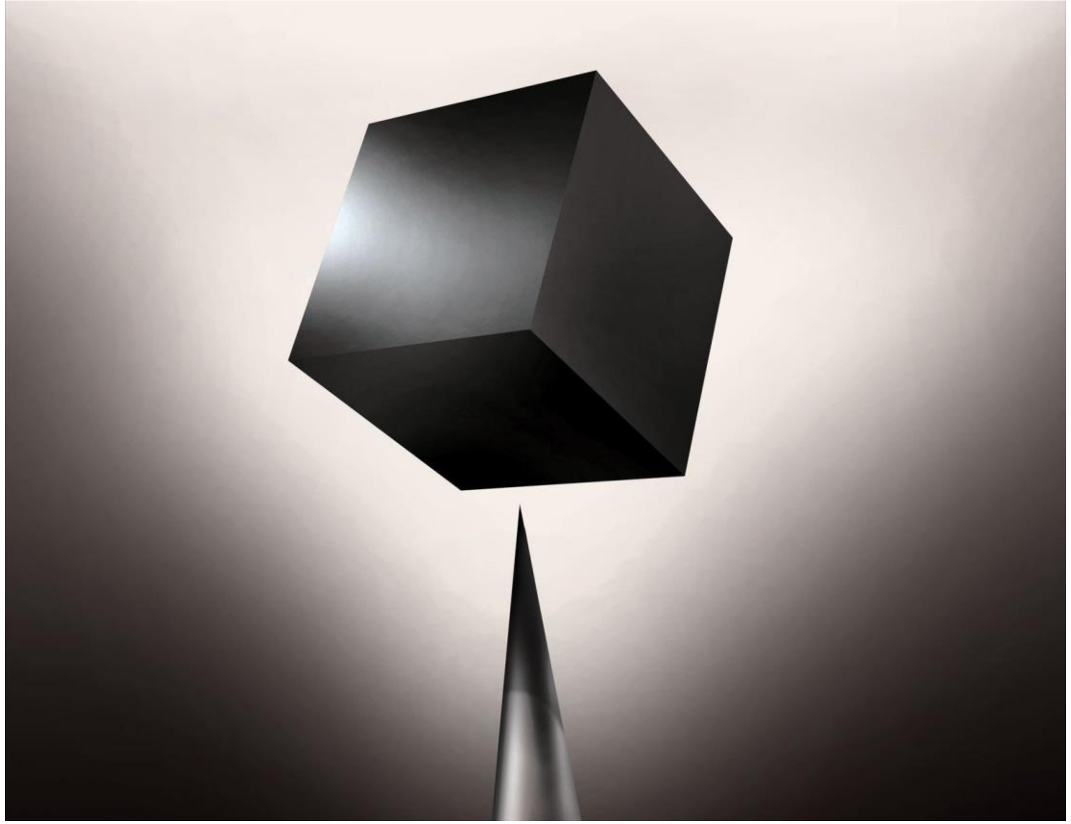
Intelligent Machines

The Dark Secret at the Heart of AI

No one really knows how the most advanced algorithms do what they do. That could be a problem.

by Will Knight April 11, 2017

Last year, a strange self-driving car was released onto the quiet roads of Monmouth County, New Jersey. The experimental vehicle, developed by researchers at the chip maker Nvidia, didn't look different from other autonomous cars, but it was unlike anything demonstrated by Google, Tesla, or General Motors, and it showed the rising power of artificial intelligence. The car didn't follow a single instruction provided by an engineer or programmer. Instead, it relied entirely on an algorithm that had taught itself to drive by watching a human do it.



Theory of deep learning?

Donoho’s PPT, Stats 385 Stanford

Deep learning as a magic mirror

Figure : Every theorist who looks at it see what they wish

Donoho's PPT, Stats 385 Stanford

Is there a bitter lesson we can learn?

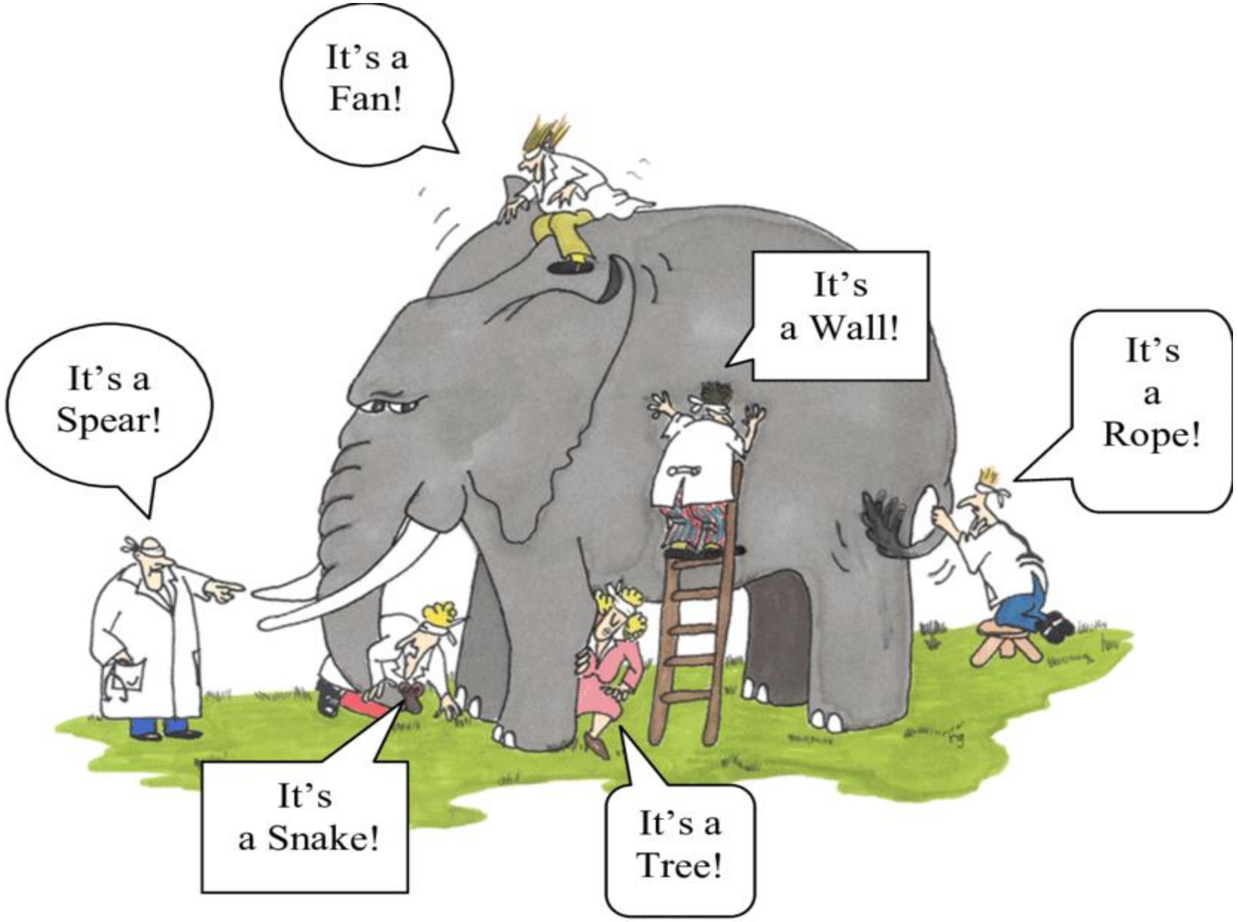
A (personal) bitter lesson:

All previously existing frameworks, irrespective of their origin or demonstrated success, are ineffective for understanding deep learning.

Existing frameworks:

statistical learning theory, numerical analysis, statistical physics, statistics, optimization, neuroscience, psychology, …

Existing frameworks often mislead

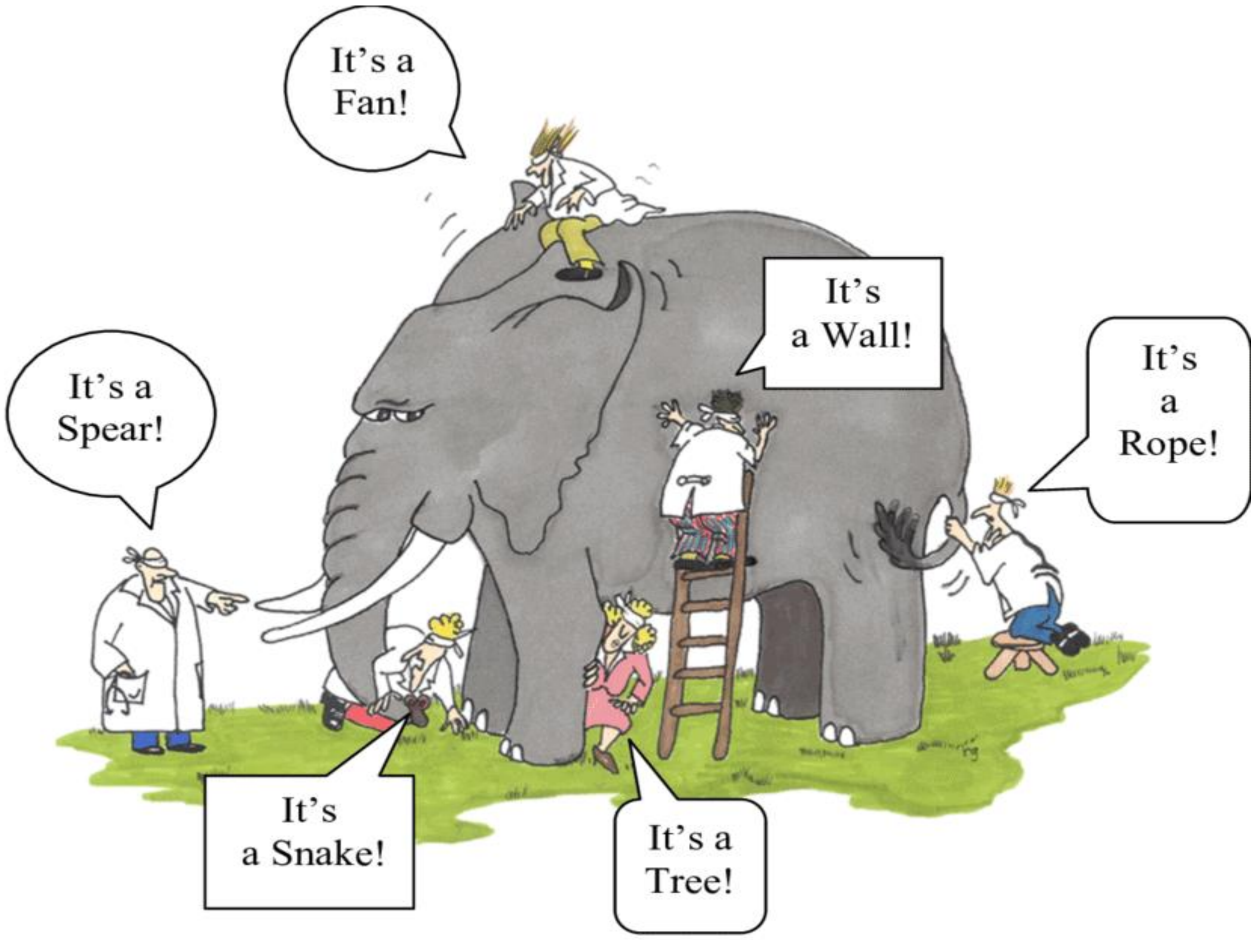


In face of deep learning, all of us are blind men.

https://www.sloww.co/blind-men-elephant/

A phenomenological methodology for blind men

1. **Suspension:** Suspend the prior and belief one may hold and focus on the facts about the object.
2. **Cumulation:** Discover and cumulate all possible facts about the object. Prioritize the more informative ones.
3. **Emergence:** A new framework shall emerge once enough pieces are uncovered.



https://www.sloww.co/blind-men-elephant/

Our phenomenological methodology--illustration

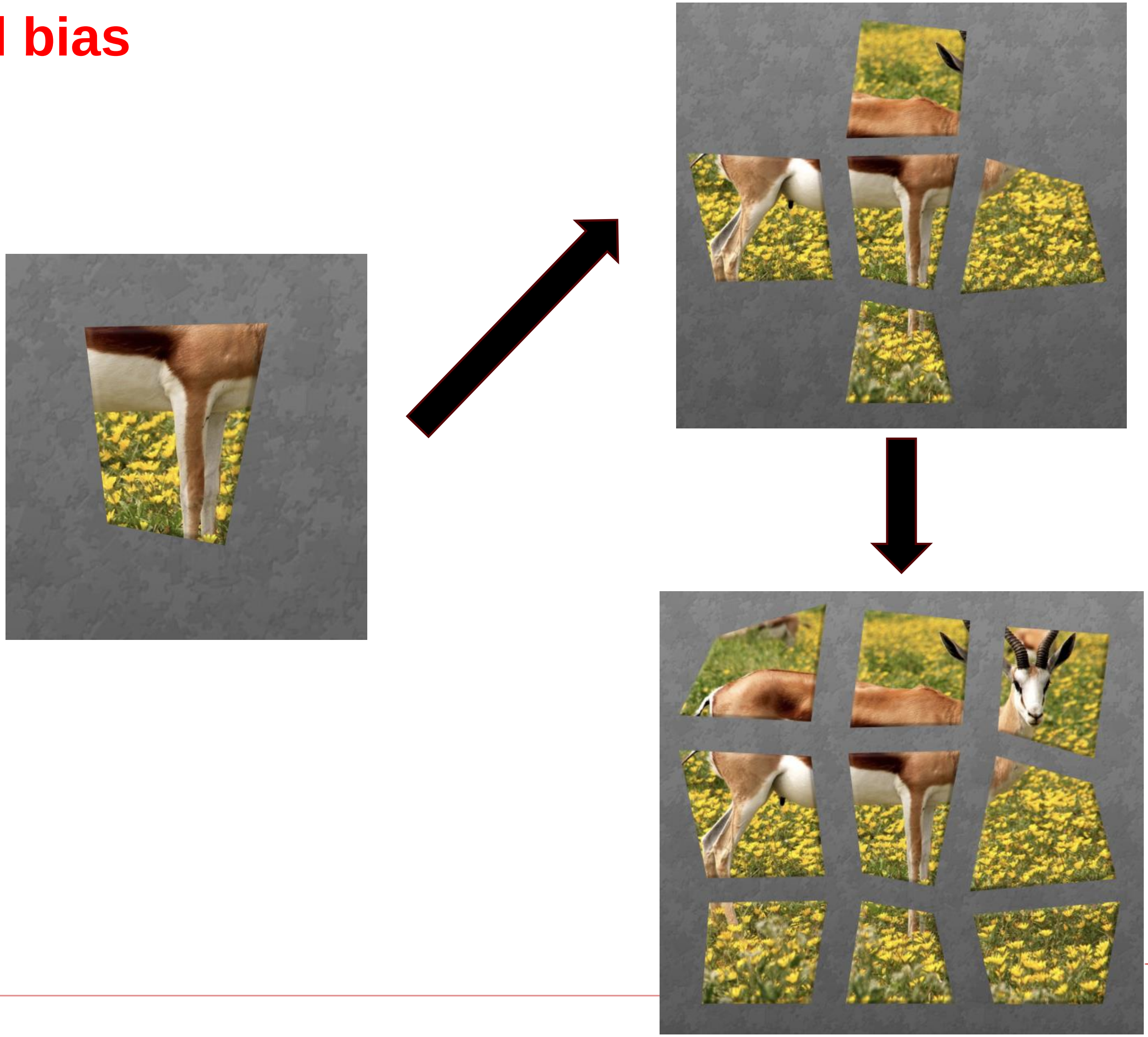
Suspension

Cumulation

Emergence

Phenomenon as a key family of facts to uncover

- **Frequency principle/spectral bias**
- **Condensation**
- **Double descent**
- **Edge of stability**
- **Lottery ticket**
- **Neural collapse**
- **Grokking**
- **……**



Basics of deep learning theory

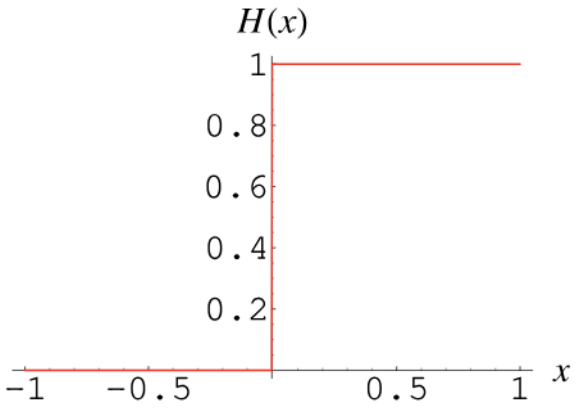
One neuron and its behavior

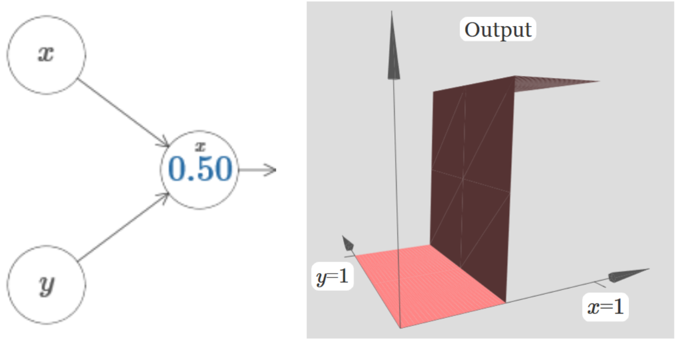
Single artificial neuron:

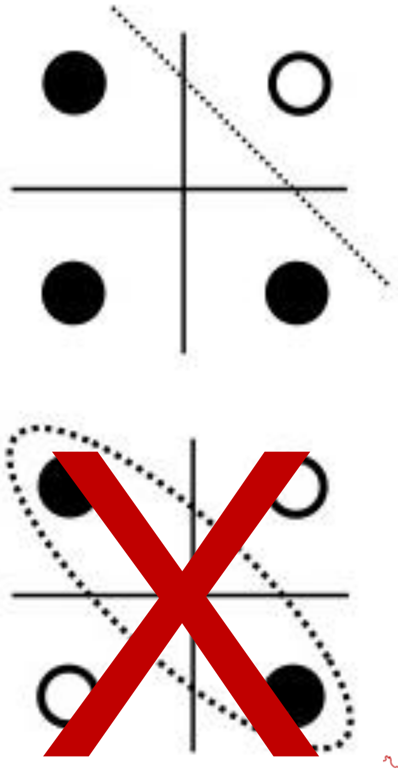
$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \sigma(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b)$$

Parameters (weights): $\boldsymbol{\theta} = (\boldsymbol{w}, \boldsymbol{b})$, activation function: $\sigma(\cdot): \mathbb{R} \to \mathbb{R}$

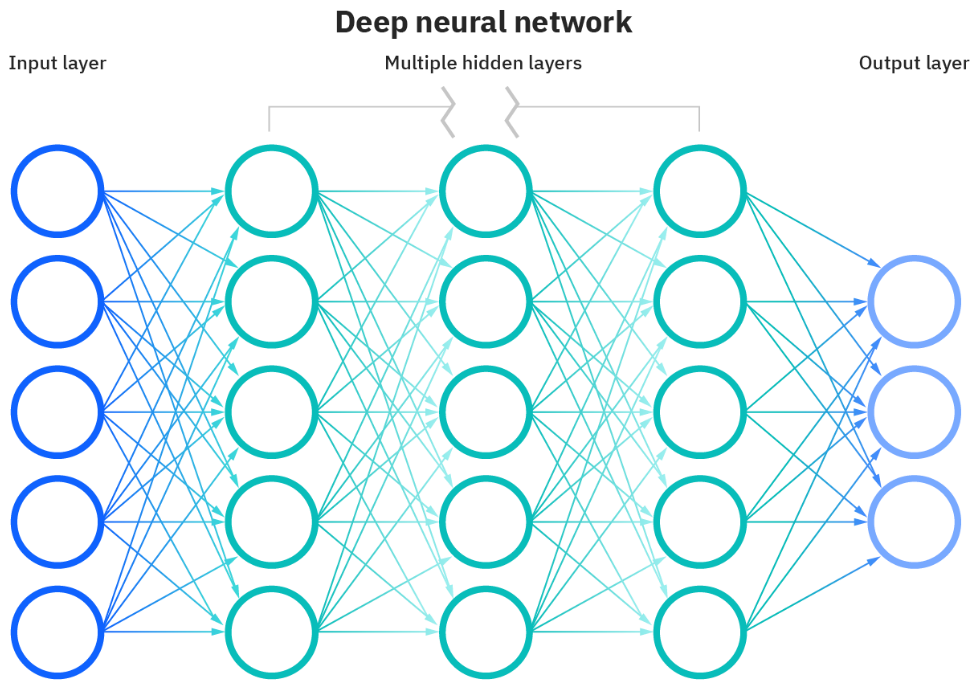
Illustration:







Deep neural networks:



$$\boldsymbol{\theta} := \left(\boldsymbol{W}^{[1]}, \boldsymbol{b}^{[1]}, \ldots, \boldsymbol{W}^{[L]}, \boldsymbol{b}^{[L]}\right)$$

$$\boldsymbol{f}_{\boldsymbol{\theta}}^{[l]}(\boldsymbol{x}) := \sigma(\boldsymbol{W}^{[l]} \boldsymbol{f}_{\boldsymbol{\theta}}^{[l-1]}(\boldsymbol{x}) + \boldsymbol{b}^{[l]})$$

https://www.ibm.com/cloud/learn/neural-networks

Universal Approximation Theorem

Neural networks with a single hidden layer can be used to approximate any continuous function to any desired precision.

Cybenko 89, Hornik 89, Hornik 91, Barron 93

Requirement for transfer function:

$\sigma(z)$ is well-defined as $z \to -\infty$ and $z \to \infty$

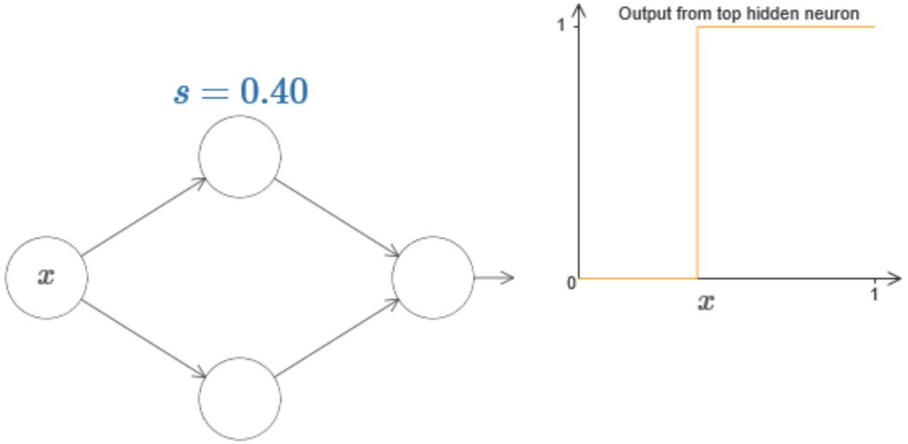
$$\boxed{\left| f(x) - \sum_j k_j \sigma\left(w_{ij} x_i + b_j\right) \right| < \epsilon}$$

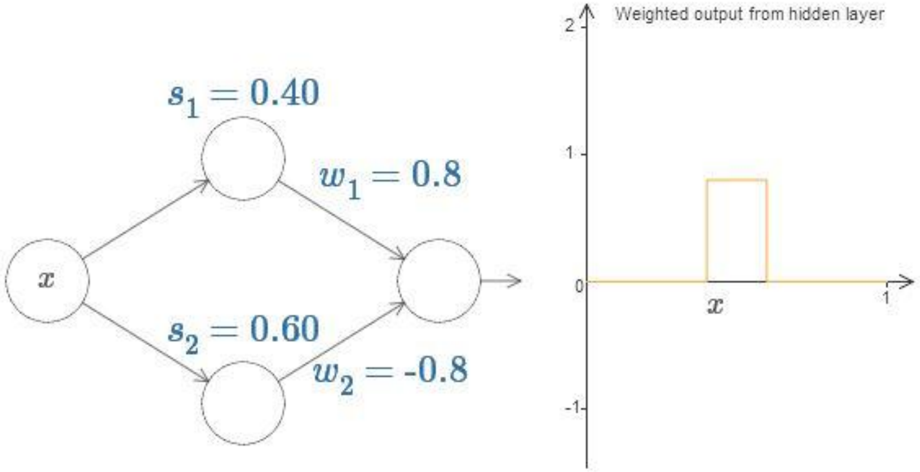
Sketch of a constructive proof:

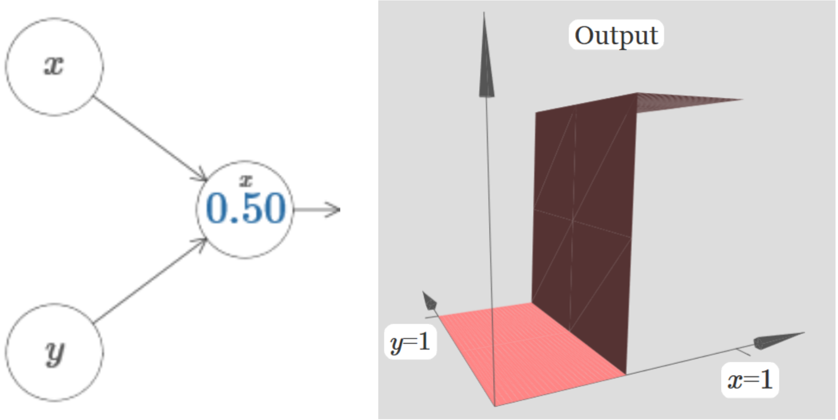
1. Construct Heaviside function from the given transfer function
2. Construct “bump” function (1-d) or “tower” function (2-d)
3. Approximate the target continuous function with “bump” or “tower” functions

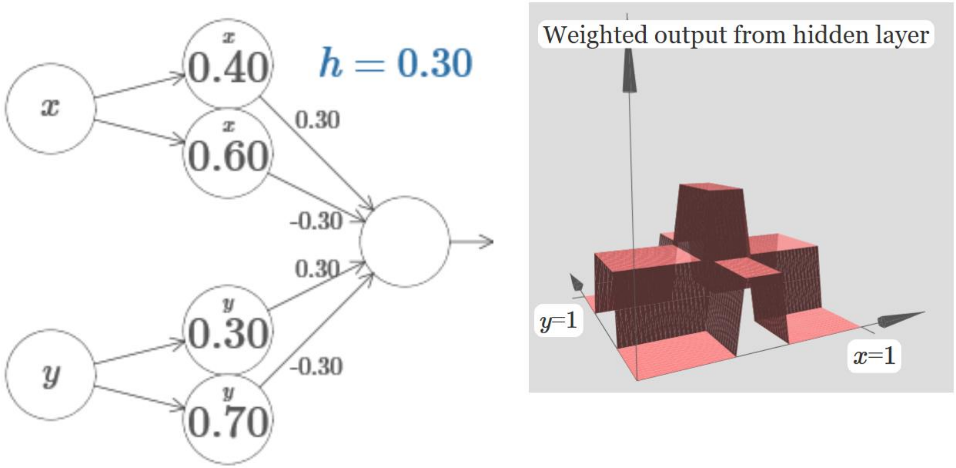
Ref: http://neuralnetworksanddeeplearning.com/chap4.html

Illustration of constructive proof (three layer)



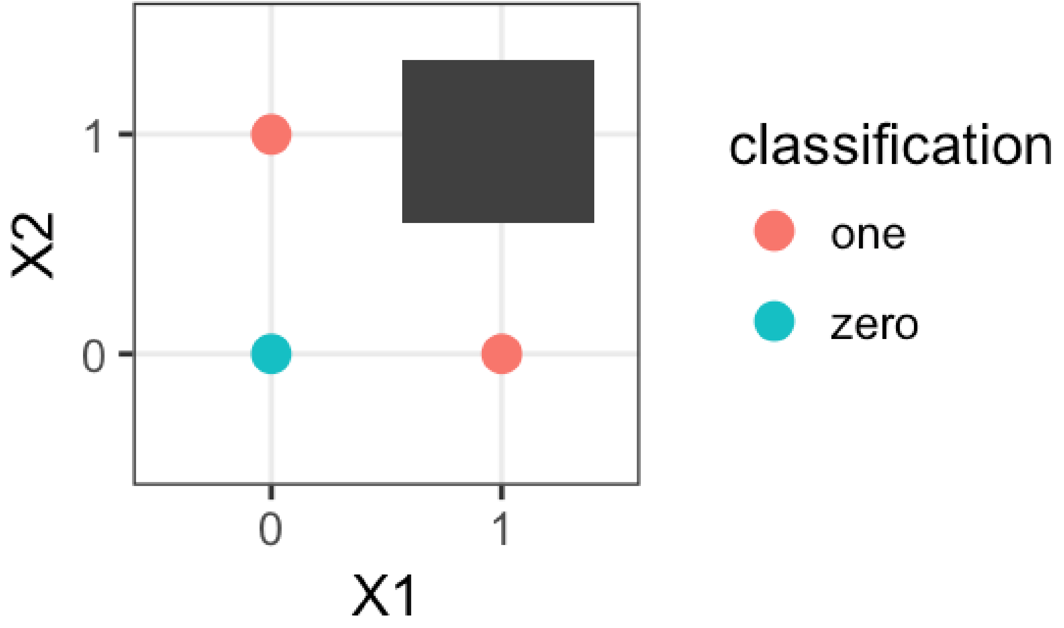






No Free Lunch Theorem (Wolpert and Macready)

Theorem—Given a finite set V and a finite set S of real numbers, assume that $f : V \to S$ is chosen at random according to uniform distribution on the set S^V of all possible functions from V to S. For the problem of optimizing f over the set V, then no algorithm performs better than blind search.



How to infer the missing spot?

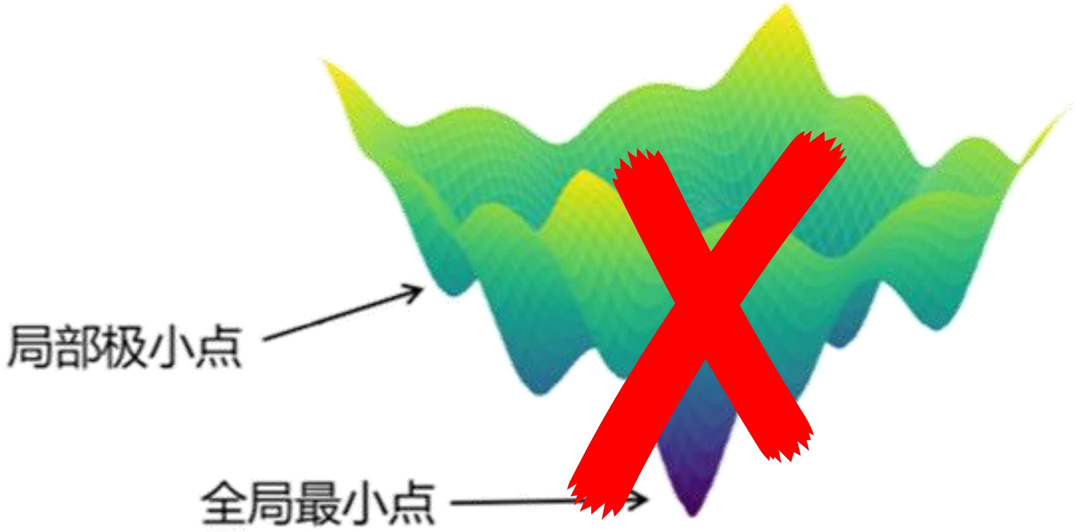
https://en.wikipedia.org/wiki/No_free_lunch_theorem

Aspects and problems in deep learning theory

- Generalization
- Optimization
- Approximation
- Robustness
- Interpretability
- …

Phenomenon

Despite strongly nonconvex loss landscape, gradient-based training of large DNNs often find global minima.



Problem

What is the geometry of loss landscape?

Approximation

Phenomenon

Some architectures are more parameter efficient than others regarding particular class of tasks.

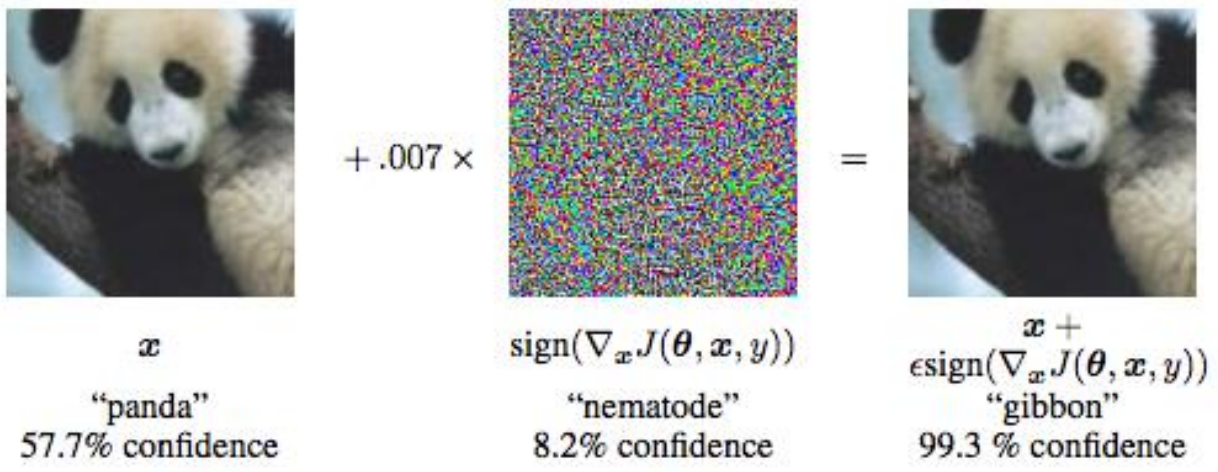
Ex: CNN vs. FNN for image, Transformer vs. LSTM for language

Problem

How to quantify the difference in parameter efficiency between architectures?

Phenomenon

Output of well-trained DNNs are often susceptible to tiny adversarial perturbation.



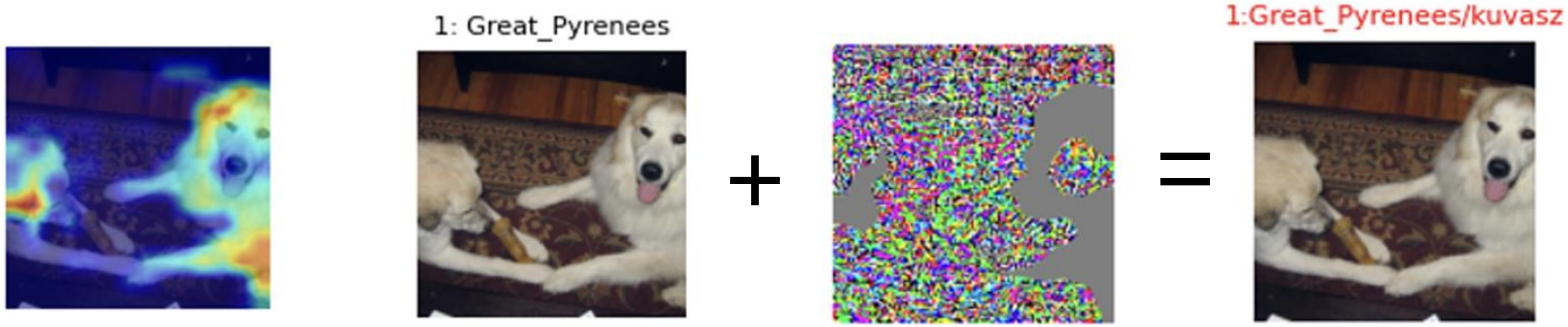
Goodfellow et al.

Problem

Why is that? How to improve robustness?

Phenomenon

One can hardly obtain an explanation with prediction power.

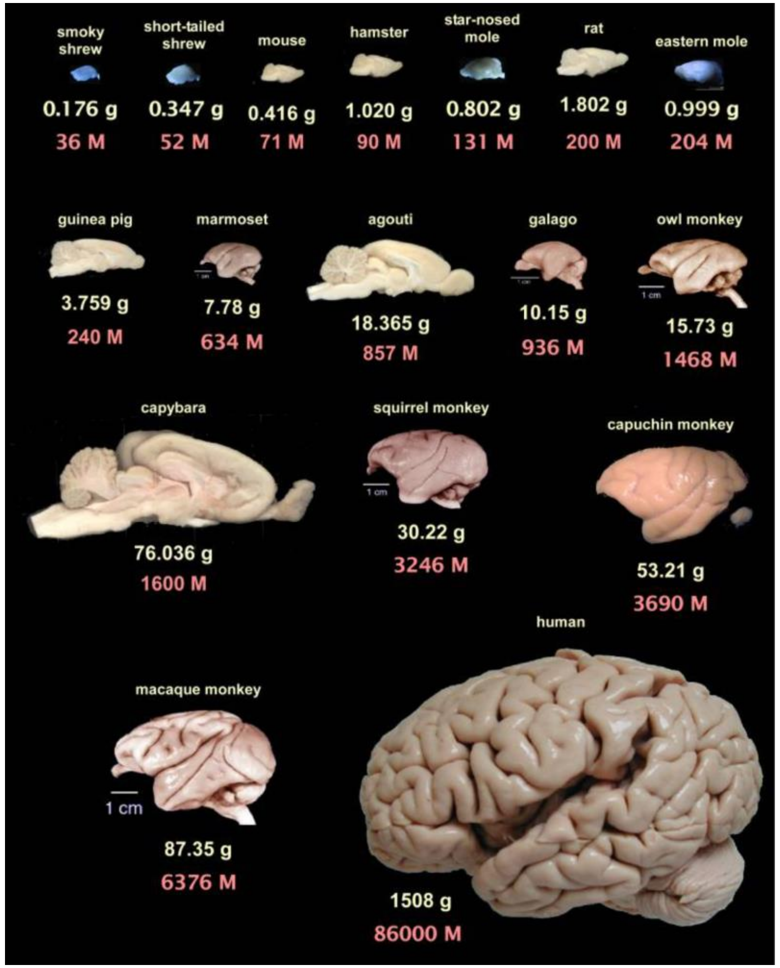


Problem

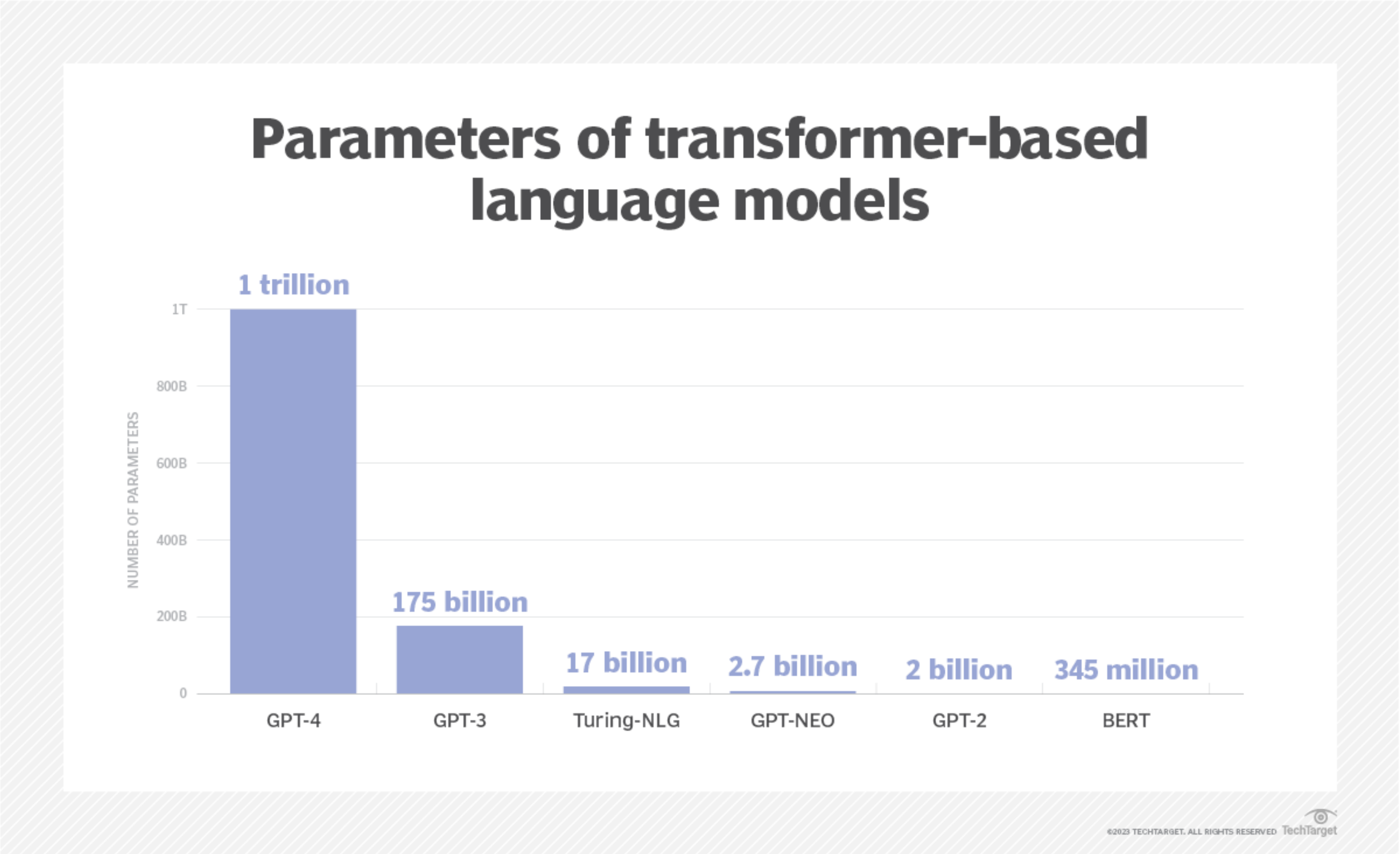
When is it possible to obtain explanations with prediction power?

Generalization puzzle of deep learning

Learning systems with increasingly large size

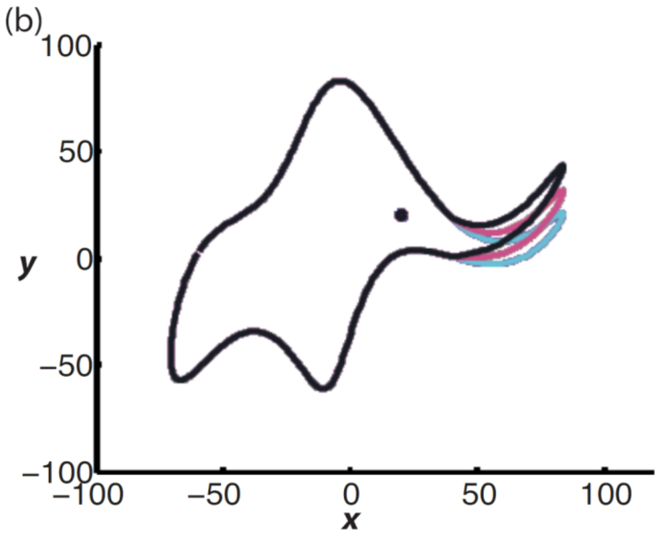


Suzana Herculano-Houzel, 2009

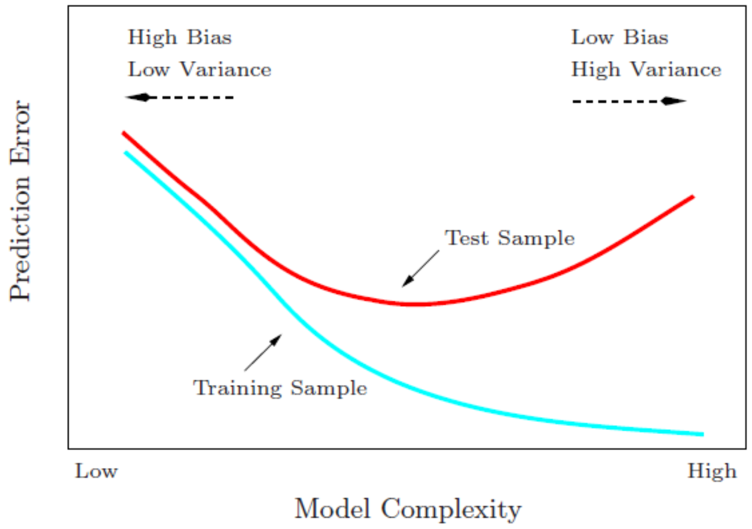


"With four parameters you can fit an elephant to a curve; with five you can make him wiggle his trunk.”

-- John von Neumann



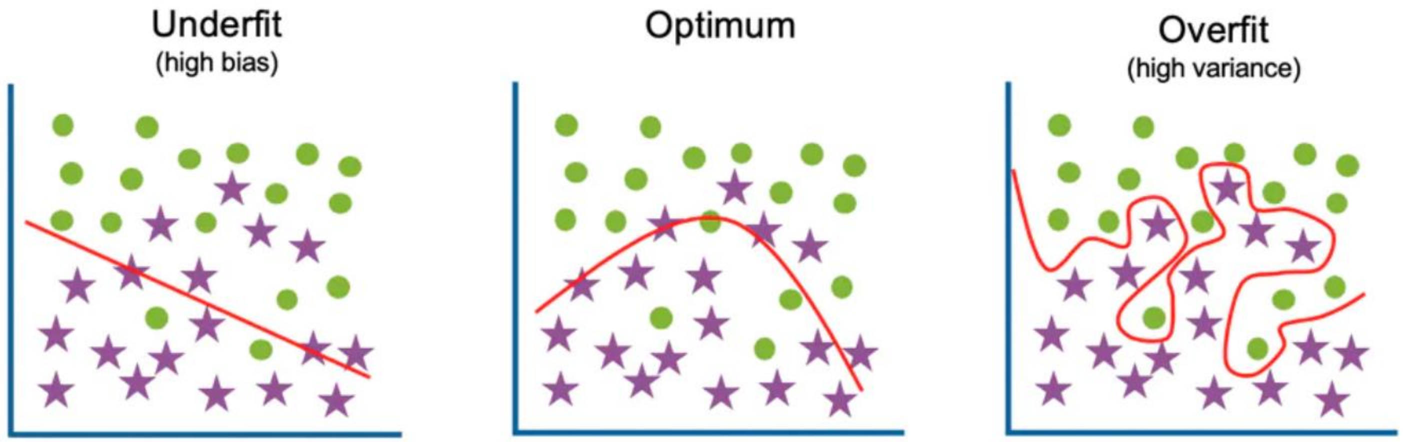
Mayer et al., 2010



Complex models easily overfit.

Failure of traditial wisdom

Large complexity → Large generalization gap

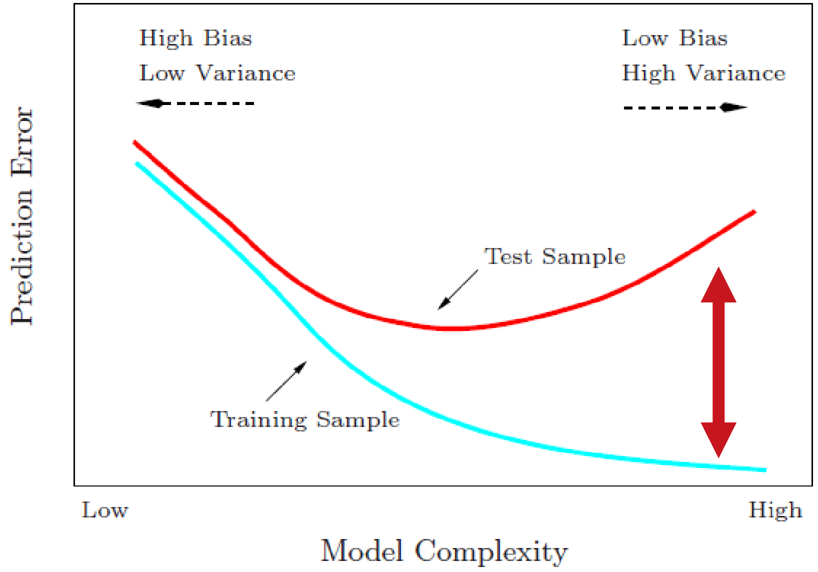


High training error
High test error

Low training error
Low test error

Low training error
High test error

Generalization Gap



Occam Razor: Entities should not be multiplied unnecessarily

Mystery: overparameterized NN often generalize well

1995

Leo Breiman
Statistics Department, University of California, Berkeley, CA 94305;
e-mail: leo@stat.berkeley.edu

Reflections After Refereeing Papers for NIPS

Our fields would be better off with far fewer theorems, less emphasis on faddish stuff, and much more scientific inquiry and engineering. But the latter requires real thinking.

For instance, there are many important questions regarding neural networks which are largely unanswered. There seem to be conflicting stories regarding the following issues:

- Why don't heavily parameterized neural networks overfit the data?
- What is the effective number of parameters?
- Why doesn't backpropagation head for a poor local minima?
- When should one stop the backpropagation and use the current parameters?

Modern verification of generalization mystery

UNDERSTANDING DEEP LEARNING REQUIRES RE-THINKING GENERALIZATION

Chiyuan Zhang*
Massachusetts Institute of Technology
chiyuan@mit.edu

Samy Bengio
Google Brain
bengio@google.com

Moritz Hardt
Google Brain
mrtz@google.com

Benjamin Recht†
University of California, Berkeley
brecht@berkeley.edu

Oriol Vinyals
Google DeepMind
vinyals@google.com

Cifar10: 60,000 training data

model	# params	random crop	weight decay	train accuracy	test accuracy
Inception	1,649,402	yes	yes	100.0	89.05
		yes	no	100.0	89.31
		no	yes	100.0	86.03
		no	no	100.0	85.75
(fitting random labels)		no	no	100.0	9.78

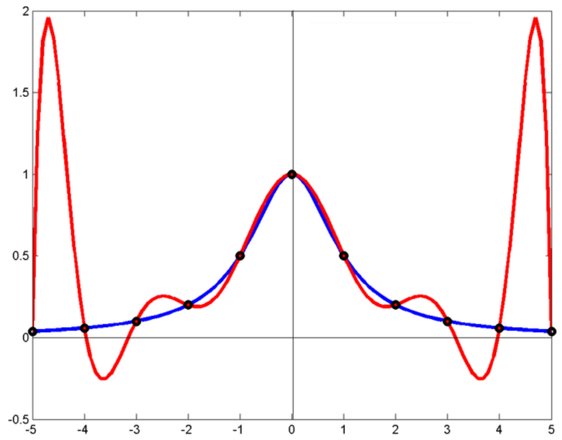
Zhang et al., 2017

Generalization mystery in 1-d interpolation

Find an interpolation of $\mathcal{D}: \{(x_i, y_i)\}_{i=1}^{n}$ in $\mathcal{H}: \{h(\cdot;\Theta) | \Theta \in \mathbb{R}^m\}$

Example:

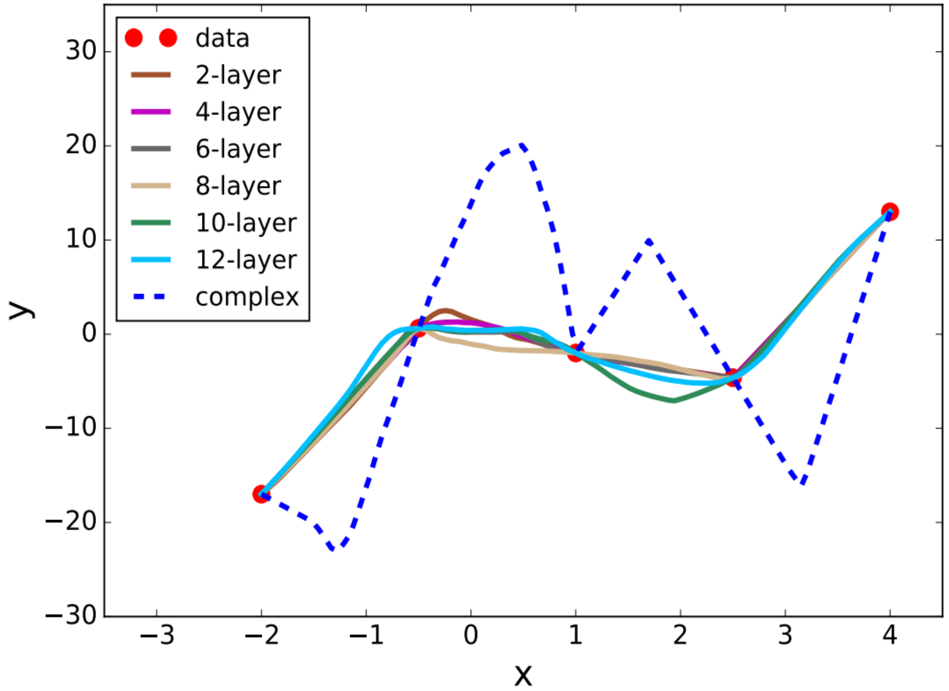
$h(x;\Theta) = \theta_1 + \theta_2 x + \cdots + \theta_M x^{m-1}$ with $m = n$



Traditional wisdom: $m < n$.

Modern wisdom?

Using neural network with $m \gg n$.



Lei Wu, Zhanxing Zhu, Weinan E, 2017

Thanks！

饮 水 思 源　爱 国 荣 校