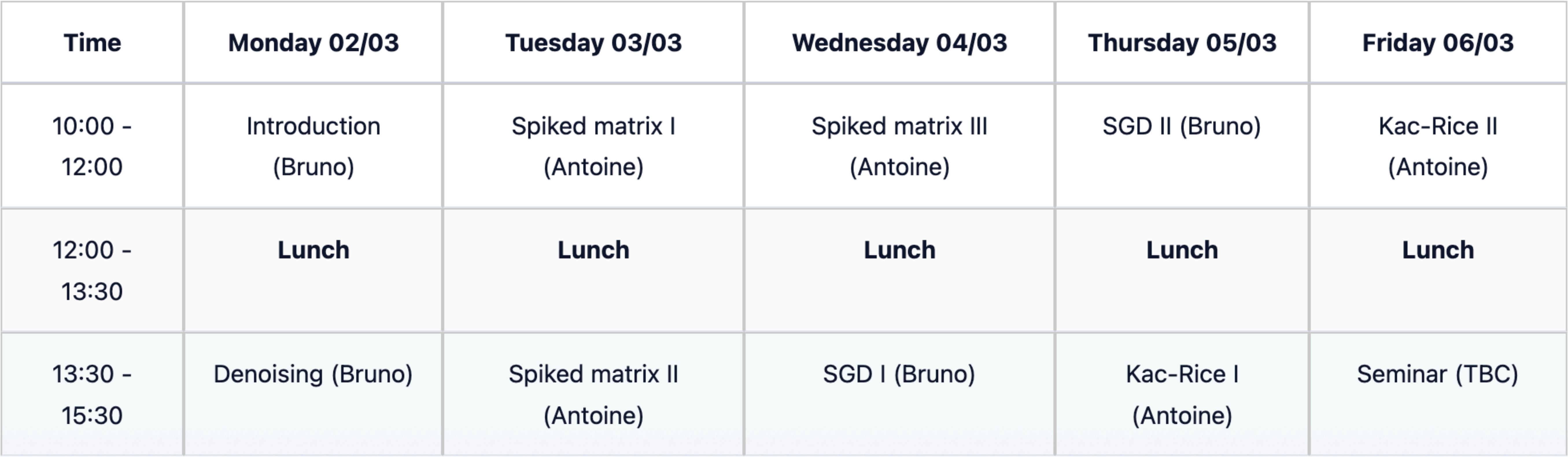
PSL Week
“Statistical physics of learning”

Bruno Loureiro (CNRS & DIENS)
Antoine Maillard (INRIA & DIENS)

PSL Week
02.03-06.03.2026

Plan of the week



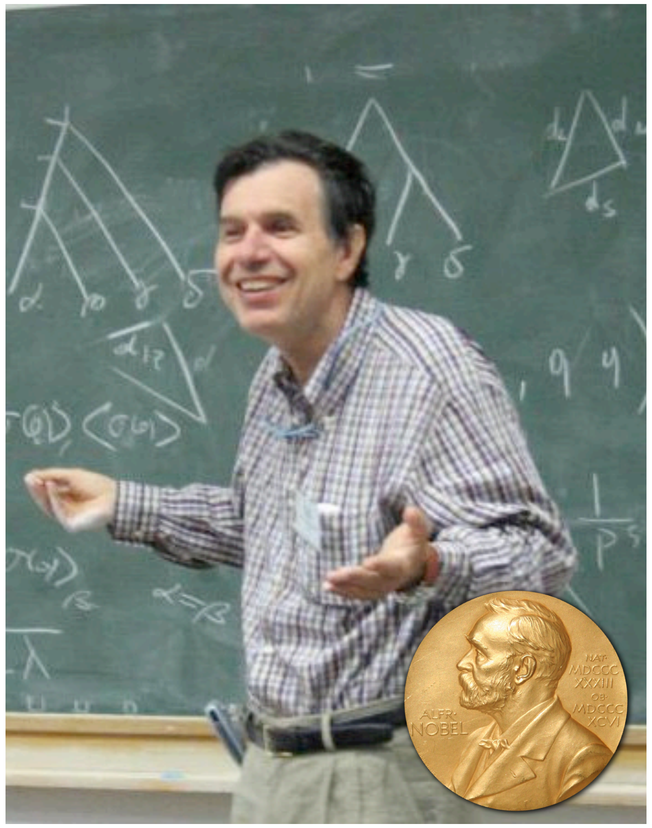
Time	Monday 02/03	Tuesday 03/03	Wednesday 04/03	Thursday 05/03	Friday 06/03
10:00 - 12:00	Introduction (Bruno)	Spiked matrix I (Antoine)	Spiked matrix III (Antoine)	SGD II (Bruno)	Kac-Rice II (Antoine)
12:00 - 13:30	**Lunch**	**Lunch**	**Lunch**	**Lunch**	**Lunch**
13:30 - 15:30	Denoising (Bruno)	Spiked matrix II (Antoine)	SGD I (Bruno)	Kac-Rice I (Antoine)	Seminar (TBC)

- **Evaluation:** Assiduity + Paper review
- **Ressources:** - Antoine’s lecture notes on high-d in inference

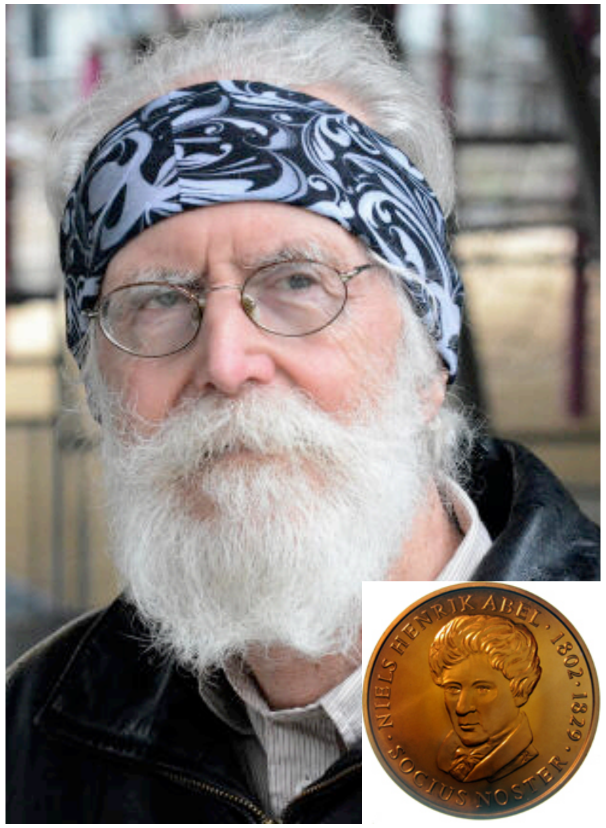
  - Bruno’s lecture notes on statistical physics of learning.

  - Bruno’s lecture notes on scaling limits of SGD

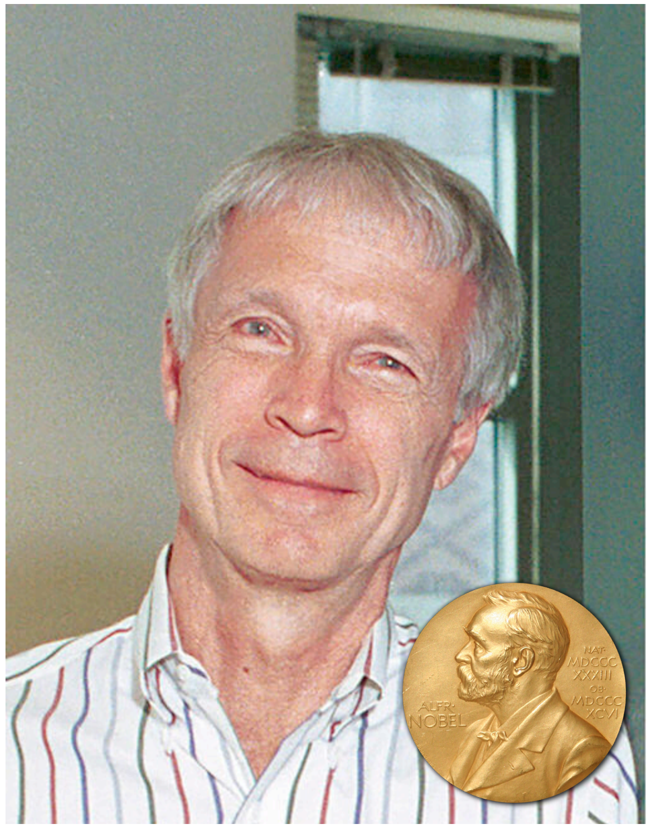
G. Parisi
(2021)



M. Talagrand
(2024)



J. Hopfield
(2024)



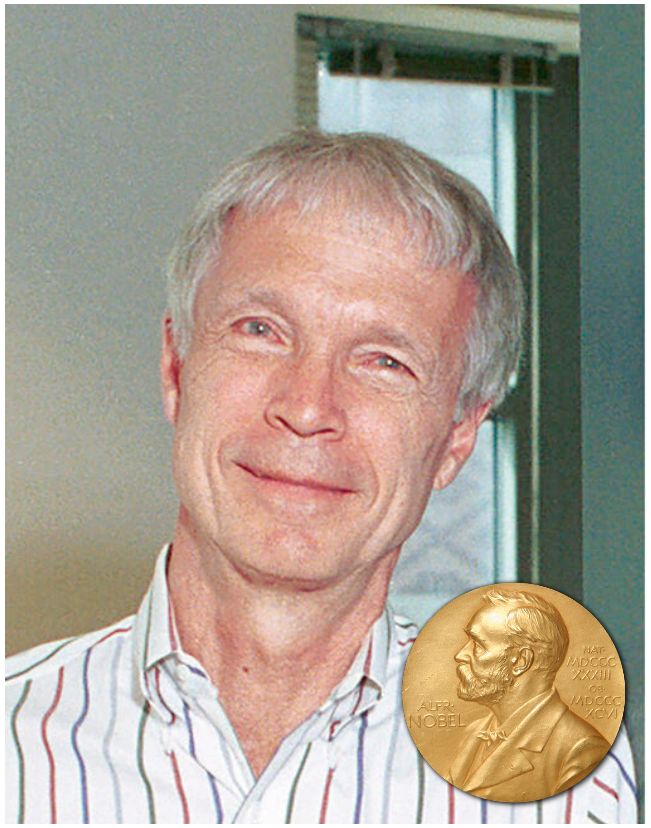
G. Hinton
(2024)

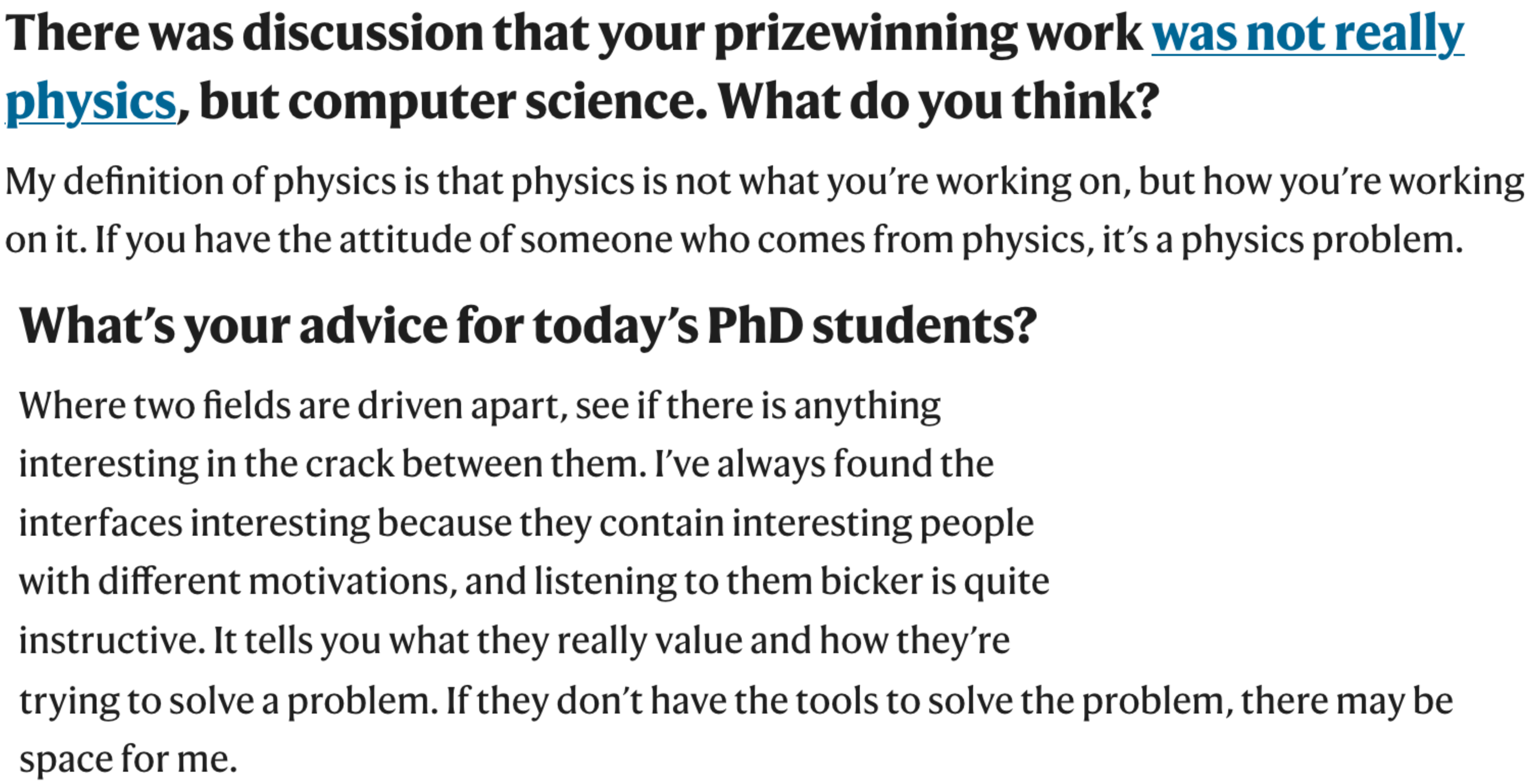
nature

NEWS Q&A | 28 October 2024

By Elizabeth Gibney

J. Hopfield (2024)





There was discussion that your prizewinning work was not really physics, but computer science. What do you think?

My definition of physics is that physics is not what you're working on, but how you're working on it. If you have the attitude of someone who comes from physics, it's a physics problem.

What's your advice for today's PhD students?

Where two fields are driven apart, see if there is anything interesting in the crack between them. I've always found the interfaces interesting because they contain interesting people with different motivations, and listening to them bicker is quite instructive. It tells you what they really value and how they're trying to solve a problem. If they don't have the tools to solve the problem, there may be space for me.

Statistical physics view of a *"theory of machine learning"*

Theory of machine learning?

Theory can mean different things.

fridge
Theory of ~~machine learning~~?

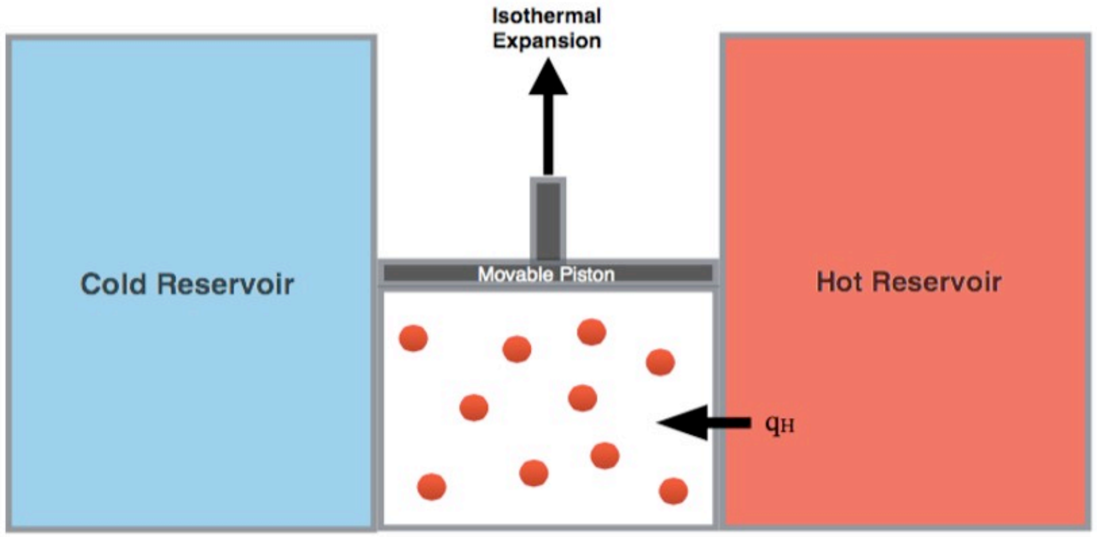
Theory can mean different things.

fridge
Theory of ~~machine learning~~?

Theory can mean different things.

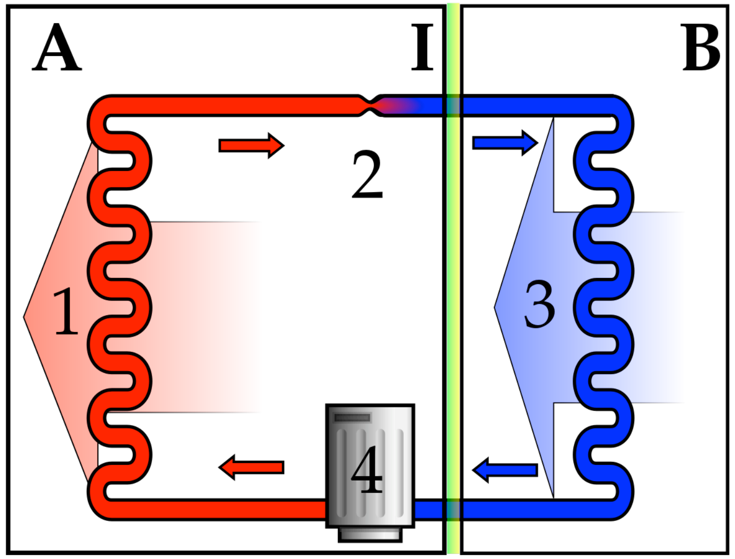
Physics

Fundamental laws that govern behaviour of the fridge



Engineering

How do I build a good fridge?

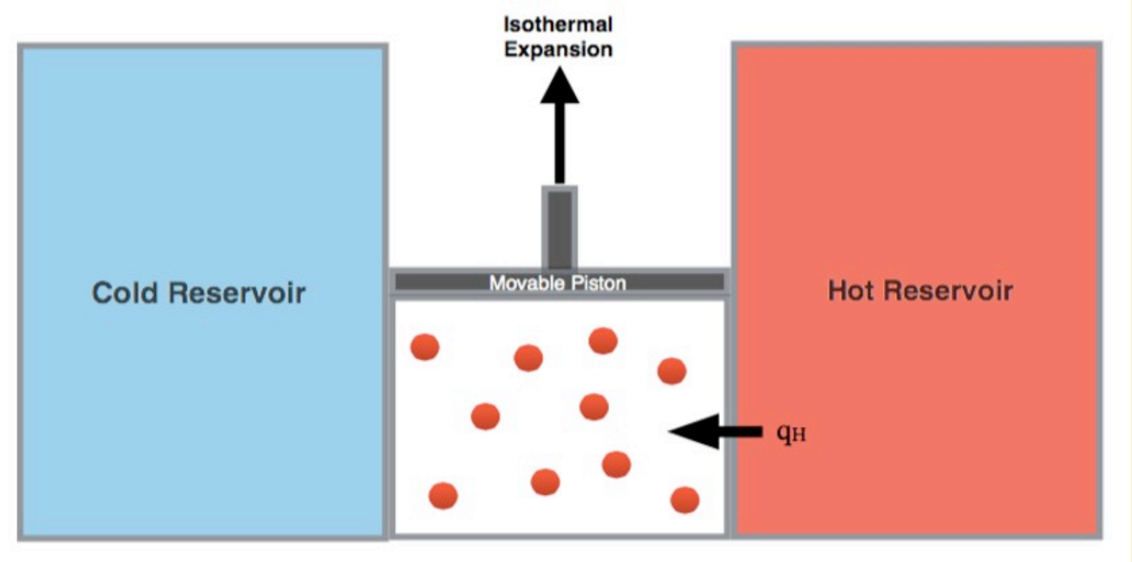


Theory of ~~machine learning~~ fridge?

Theory can mean different things.

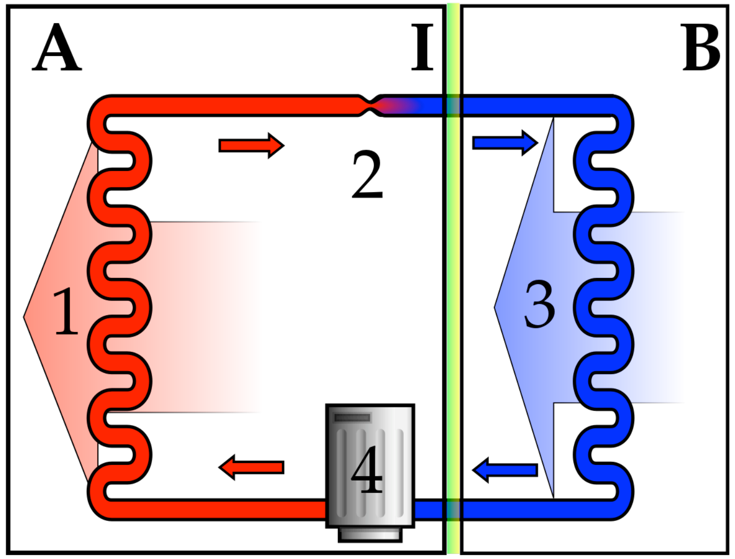
Physics

Fundamental laws that govern behaviour of the fridge



Engineering

How do I build a good fridge?

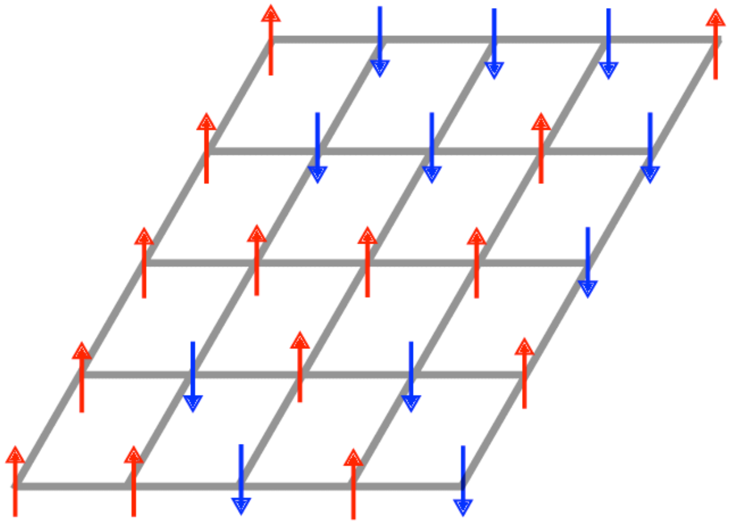


Theory of magnetism

a.k.a. the *Ising Model*

$$H_{J,h}(s) = -J\sum_{(ij)\in E} s_i s_j + h\sum_{i\in V} s_i$$

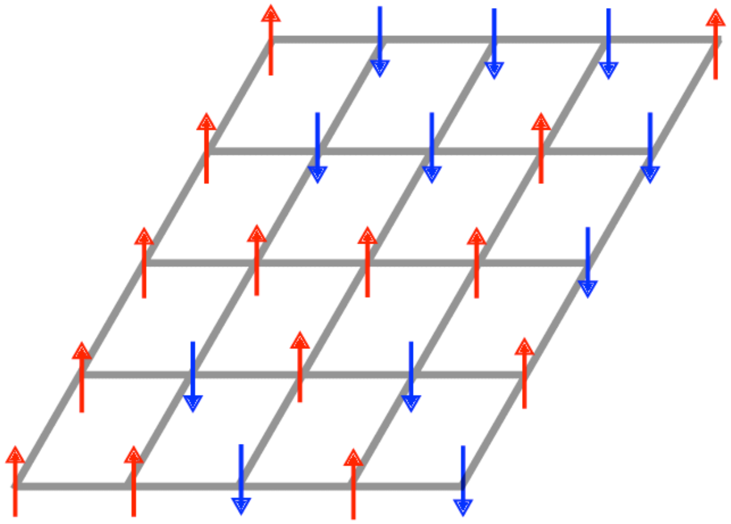
$$\mu_\beta(s) = \frac{1}{Z_{\beta,J,h}} e^{-\beta H_{J,h}(s)} \qquad s \in \{-1,+1\}^N$$



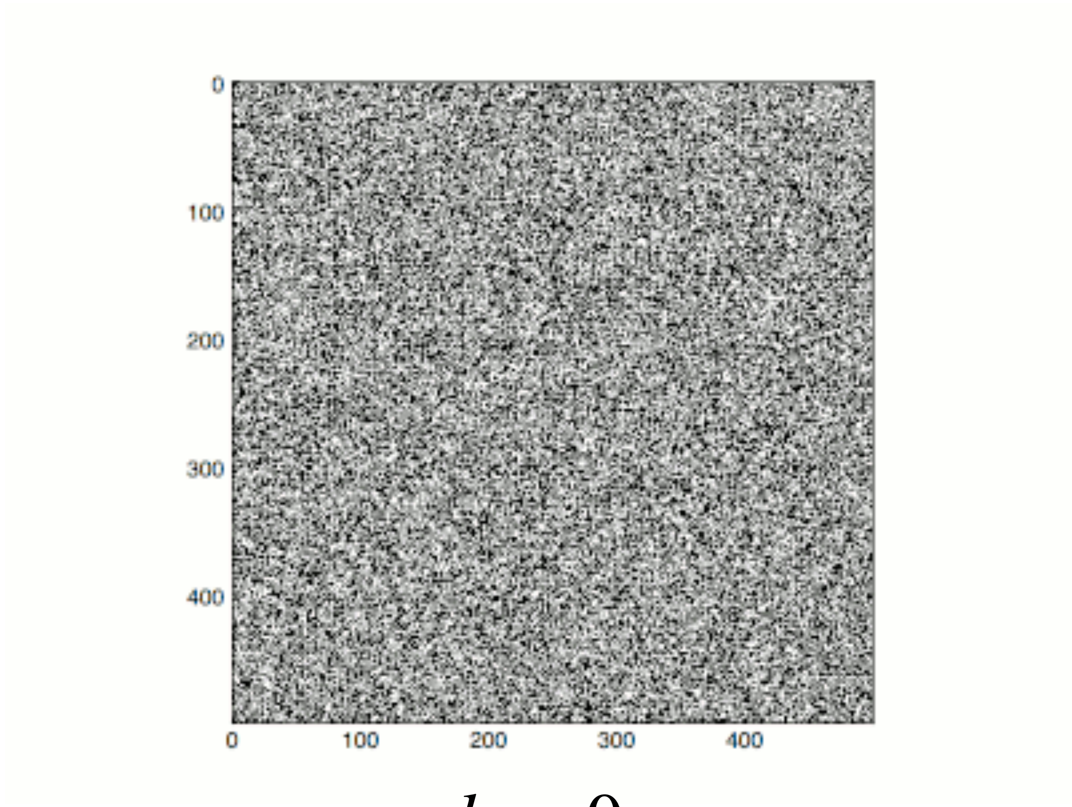
Theory of magnetism

a.k.a. the *Ising Model*

$$H_{J,h}(s) = -J \sum_{(ij)\in E} s_i s_j + h \sum_{i\in V} s_i$$

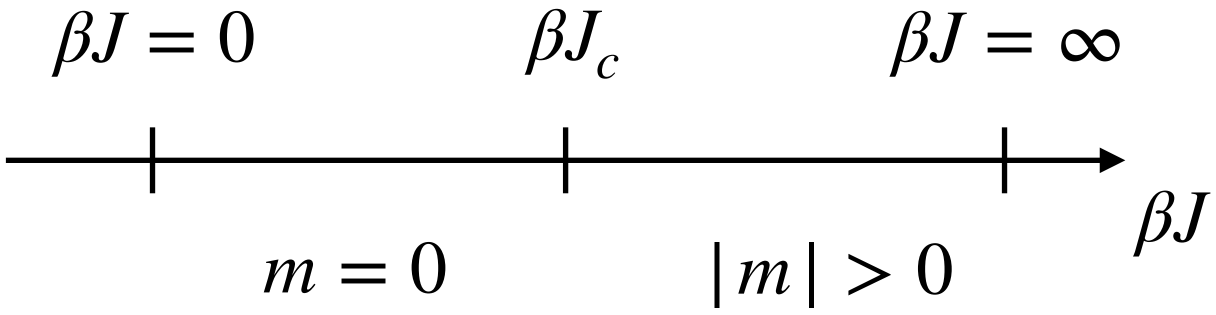


$$\mu_\beta(s) = \frac{1}{Z_{\beta,J,h}} e^{-\beta H_{J,h}(s)} \qquad s \in \{-1,+1\}^N$$



$h = 0$

Order parameter: $$m = \frac{1}{|V|} \sum_{i\in V} s_i$$



$\beta J = 0$ βJ_c $\beta J = \infty$

βJ

$m = 0$ $|m| > 0$

[Ising 1925; Onsager 1944]

Theory of machine learning?

Theory can mean different things.

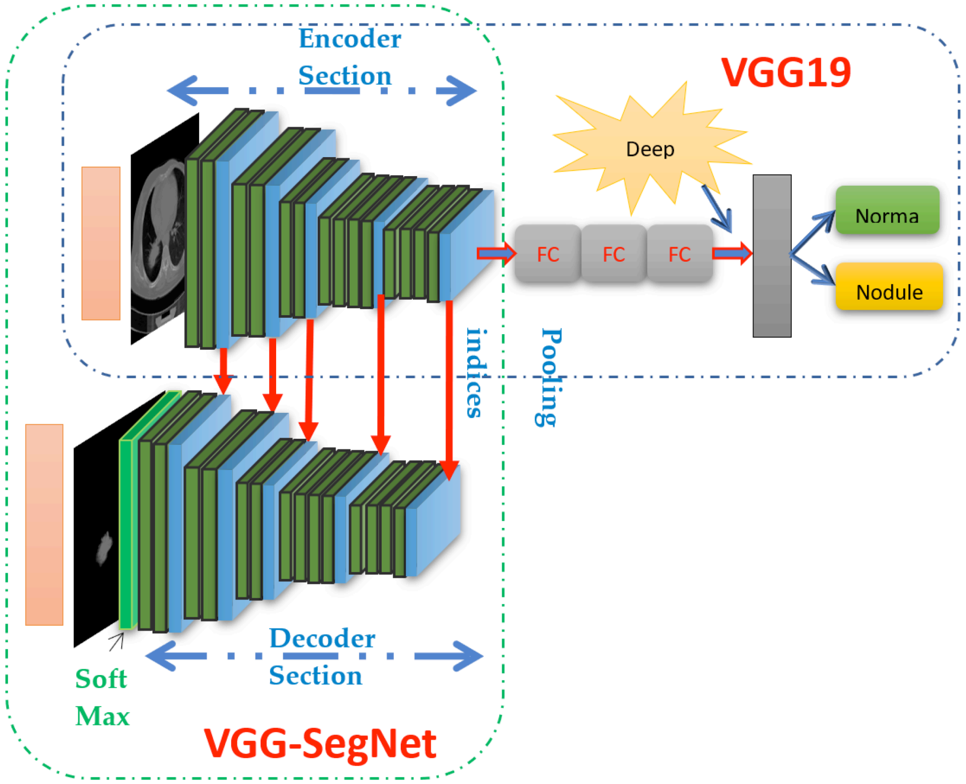
<u>Theory</u>

Fundamental principles
that govern learning

🤔

<u>Engineering</u>

How do I build and train a
state-of-the-art neural net?



Supervised Learning

Let $\mathcal{D} = \{(x_i, y_i)_{i\in[n]} \in \mathbb{R}^d \times \mathbb{R} : i \in [n]\}$ ind. sampled from ρ .

Supervised Learning

Let $\mathscr{D} = \{(x_i, y_i)_{i \in [n]} \in \mathbb{R}^d \times \mathbb{R} : i \in [n]\}$ ind. sampled from ρ .

Want: Learn $f : \mathbb{R}^d \to \mathbb{R}$ from data $\mathscr{D}$

Supervised Learning

Let $\mathscr{D} = \{(x_i, y_i)_{i \in [n]} \in \mathbb{R}^d \times \mathbb{R} : i \in [n]\}$ ind. sampled from ρ .

Want: Learn $f : \mathbb{R}^d \to \mathbb{R}$ from data $\mathscr{D}$

🤔 $f(x) = \begin{cases} y_i & \text{if} \quad x \in \mathscr{D} \\ 0 & \text{otherwise} \end{cases}$

Supervised Learning

Let $\mathscr{D} = \{(x_i, y_i)_{i\in[n]} \in \mathbb{R}^d \times \mathbb{R} : i \in [n]\}$ ind. sampled from ρ .

Want: Learn $f : \mathbb{R}^d \to \mathbb{R}$ from data $\mathscr{D}$

🤔 $$f(x) = \begin{cases} y_i & \text{if} \quad x \in \mathscr{D} \\ 0 & \text{otherwise} \end{cases}$$

Memorisation,
not learning!

Supervised Learning

Let $\mathscr{D} = \{(x_i, y_i)_{i\in[n]} \in \mathbb{R}^d \times \mathbb{R} : i \in [n]\}$ ind. sampled from ρ .

Want: Learn $f : \mathbb{R}^d \to \mathbb{R}$ from data $\mathscr{D}$

🤔 $$f(x) = \begin{cases} y_i & \text{if} \quad x \in \mathscr{D} \\ 0 & \text{otherwise} \end{cases}$$

Memorisation, not learning!

💡 Introduce a “cost function” $\ell(y, f(x)) \geq 0$

minimise $\mathscr{R}(f) = \mathbb{E}_{(x,y)\sim\rho}[\ell(y, f(x))]$

Population Risk

Supervised Learning

Let $\mathscr{D} = \{(x_i, y_i)_{i \in [n]} \in \mathbb{R}^d \times \mathbb{R} : i \in [n]\}$ ind. sampled from ρ .

Want: Learn $f : \mathbb{R}^d \to \mathbb{R}$ from data $\mathscr{D}$

🤔 $$f(x) = \begin{cases} y_i & \text{if} \quad x \in \mathscr{D} \\ 0 & \text{otherwise} \end{cases}$$

Memorisation, not learning!

💡 Introduce a "cost function" $\ell(y, f(x)) \geq 0$

minimise $\mathscr{R}(f) = \mathbb{E}_{(x,y)\sim\rho}[\ell(y, f(x))]$

Population Risk

☹ Problems:

- In practice, does't know ρ, only $\mathscr{D}$
- How to minimise over $\{f : \mathbb{R}^d \to \mathbb{R}\}$?

Supervised Learning

Let $\mathscr{D} = \{(x_i, y_i)_{i\in[n]} \in \mathbb{R}^d \times \mathbb{R} : i \in [n]\}$ ind. sampled from ρ .

Want: Learn $f : \mathbb{R}^d \to \mathbb{R}$ from data $\mathscr{D}$

minimise $\mathscr{R}(f) = \mathbb{E}_{(x,y)\sim\rho}[\ell(y, f(x))]$

Population Risk

minimise $\hat{\mathscr{R}}_n(f) = \dfrac{1}{n} \sum_{\nu\in[n]} [\ell(y_i, f(x_i))]$

Empirical Risk

☹ Problems:

- In practice, does't know ρ, only $\mathscr{D}$ ✓
- How to minimise over $\{f : \mathbb{R}^d \to \mathbb{R}\}$?

Supervised Learning

Let $\mathscr{D} = \{(x_i, y_i)_{i\in[n]} \in \mathbb{R}^d \times \mathbb{R} : i \in [n]\}$ ind. sampled from ρ .

Want: Learn $f_\Theta : \mathbb{R}^d \to \mathbb{R}$ from data $\mathscr{D}$

minimise $\mathscr{R}(\Theta) = \mathbb{E}_{(x,y)\sim\rho}[\ell(y, f_\Theta(x))]$

Population Risk

minimise $\hat{\mathscr{R}}_n(\Theta) = \dfrac{1}{n} \sum_{\nu\in[n]} [\ell(y_i, f(x_i))]$

Empirical Risk

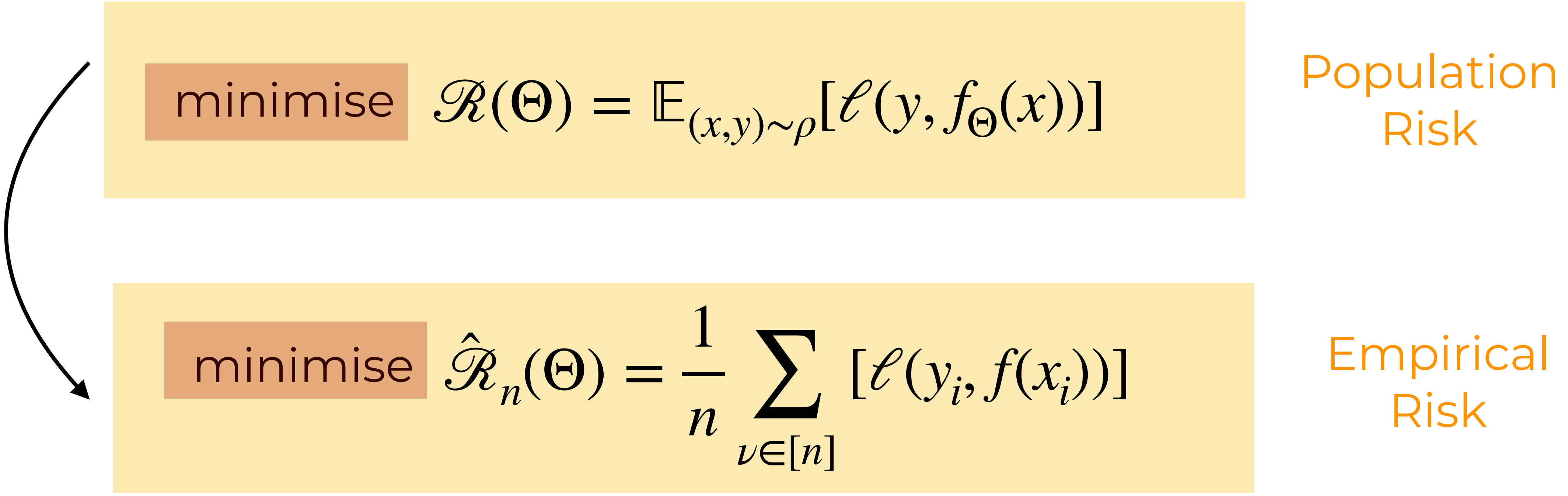
🙂 Problems:

- In practice, does't know ρ, only $\mathscr{D}$ ✓
- How to minimise over $\{f : \mathbb{R}^d \to \mathbb{R}\}$? ✓

Supervised Learning

Let $\mathscr{D} = \{(x_i, y_i)_{i\in[n]} \in \mathbb{R}^d \times \mathbb{R} : i \in [n]\}$ ind. sampled from ρ .

Want: Learn $f_\Theta : \mathbb{R}^d \to \mathbb{R}$ from data $\mathscr{D}$



minimise $\mathscr{R}(\Theta) = \mathbb{E}_{(x,y)\sim\rho}[\ell(y, f_\Theta(x))]$

Population Risk

minimise $\hat{\mathscr{R}}_n(\Theta) = \frac{1}{n}\sum_{\nu\in[n]}[\ell(y_i, f(x_i))]$

Empirical Risk

🙂 Problems:

- In practice, does't know ρ, only $\mathscr{D}$ ✓
- How to minimise over $\{f : \mathbb{R}^d \to \mathbb{R}\}$? ✓

Stat. Learning Theory

Supervised binary classification $(x_i, y_i) \in \mathbb{R}^d \times \{-1,1\}, \quad i = 1,\cdots,n$

Stat. Learning Theory

Supervised binary classification $(x_i, y_i) \in \mathbb{R}^d \times \{-1,1\}, \quad i = 1,\cdots,n$

Theorem (Uniform convergence): with probability at least $1-\delta$

$$\forall f_\Theta \in \mathscr{H} \qquad \mathscr{R}(\Theta) - \hat{\mathscr{R}}_n(\Theta) \leq \mathrm{Rad}(\mathscr{H}) + \sqrt{\frac{\log(1/\delta)}{n}}$$

Where $$\mathrm{Rad}(\mathscr{H}) = \frac{1}{n}\mathbb{E}\left[\sup_{f_\Theta \in \mathscr{H}} \sum_{i\in[n]} y_i f_\Theta(x_i)\right]$$

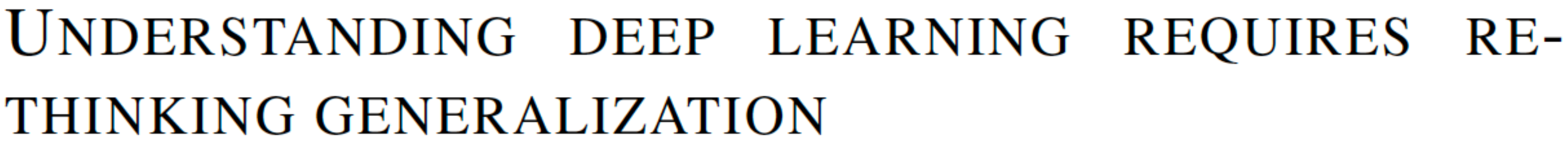
[Bartlett, Mendelson '03]

Stat. Learning Theory

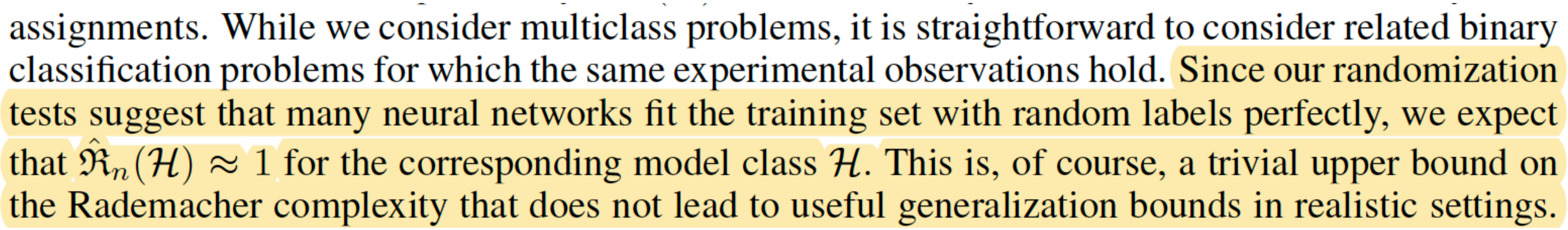
Supervised binary classification $(x_i, y_i) \in \mathbb{R}^d \times \{-1,1\}, \quad i = 1,\cdots,n$

Theorem (Uniform convergence): with probability at least $1-\delta$

$$\forall f_\Theta \in \mathscr{H} \qquad \mathscr{R}(\Theta) - \hat{\mathscr{R}}_n(\Theta) \leq \mathrm{Rad}(\mathscr{H}) + \sqrt{\frac{\log(1/\delta)}{n}}$$



UNDERSTANDING DEEP LEARNING REQUIRES RETHINKING GENERALIZATION

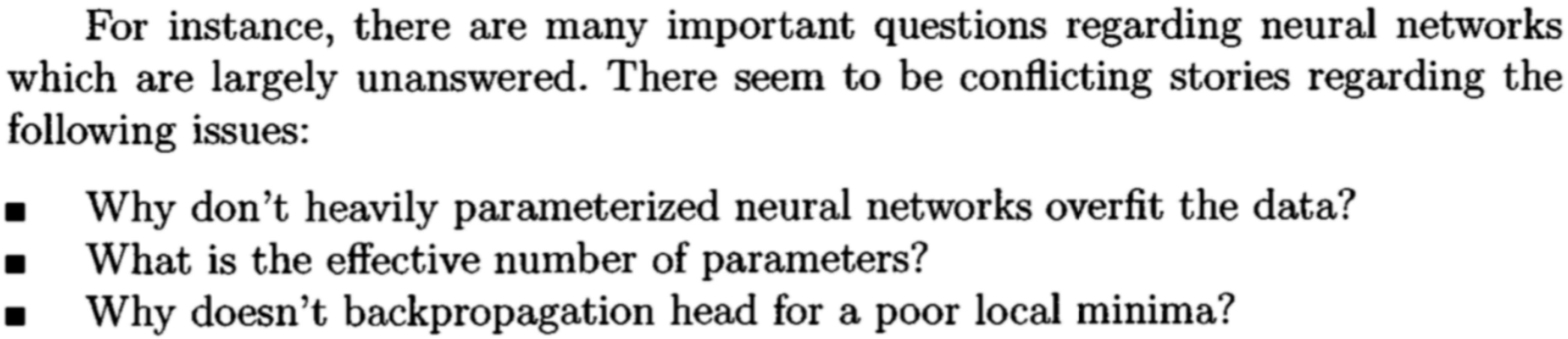


assignments. While we consider multiclass problems, it is straightforward to consider related binary classification problems for which the same experimental observations hold. Since our randomization tests suggest that many neural networks fit the training set with random labels perfectly, we expect that $\hat{\mathfrak{R}}_n(\mathcal{H}) \approx 1$ for the corresponding model class $\mathcal{H}$. This is, of course, a trivial upper bound on the Rademacher complexity that does not lead to useful generalization bounds in realistic settings.

[Zhang, Bengio, Hardt, Recht, Vinyals 17']

Many questions, few answers

Despite the amazing progress on the engineering side,
theory falls short.

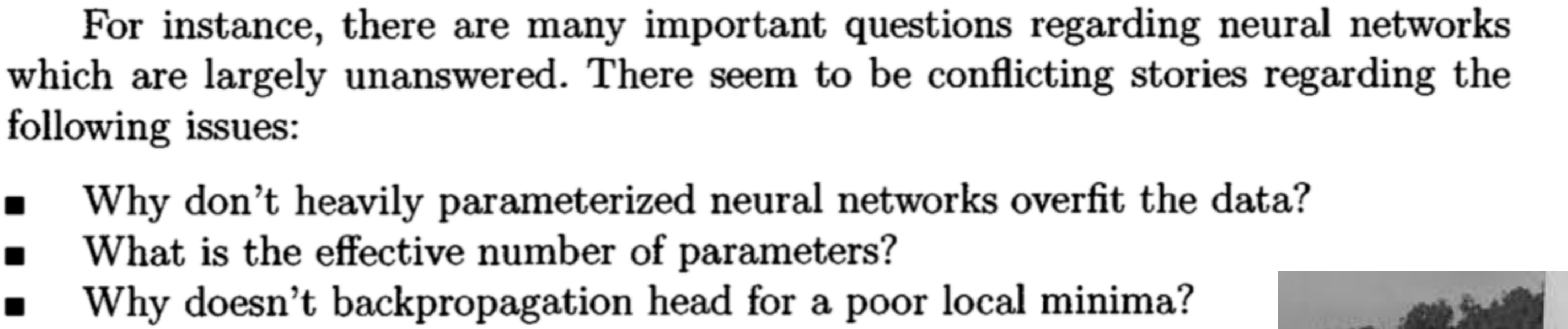


For instance, there are many important questions regarding neural networks which are largely unanswered. There seem to be conflicting stories regarding the following issues:

- 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?

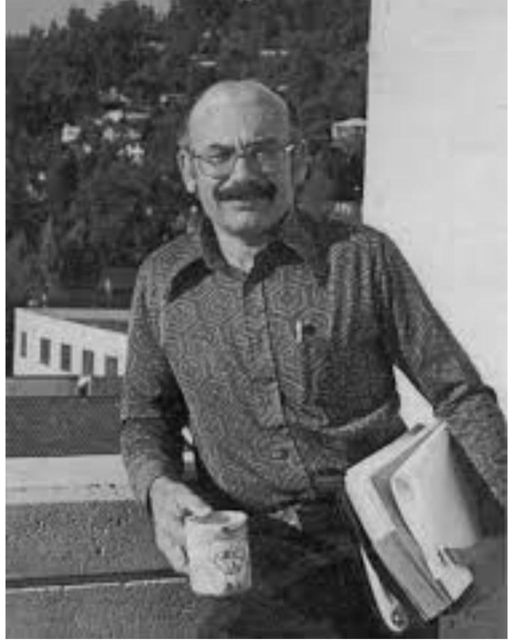
Many questions, few answers

Despite the amazing progress on the engineering side,
theory falls short.



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

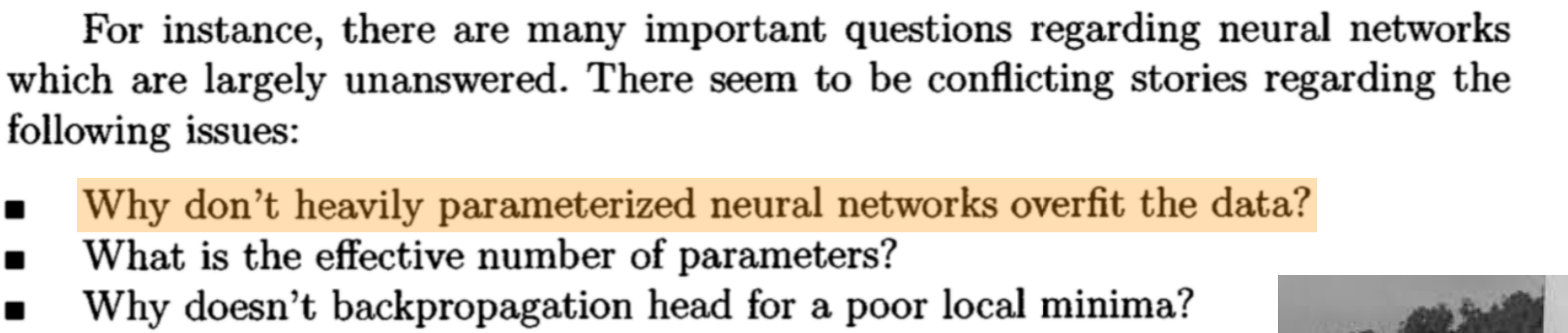
- Why don't heavily parameterized neural networks overfit the data?
- What is the effective number of parameters?
- Why doesn't backpropagation head for a poor local minima?



"Reflections after refereeing papers for NIPS",
Leo Breiman, **1995**

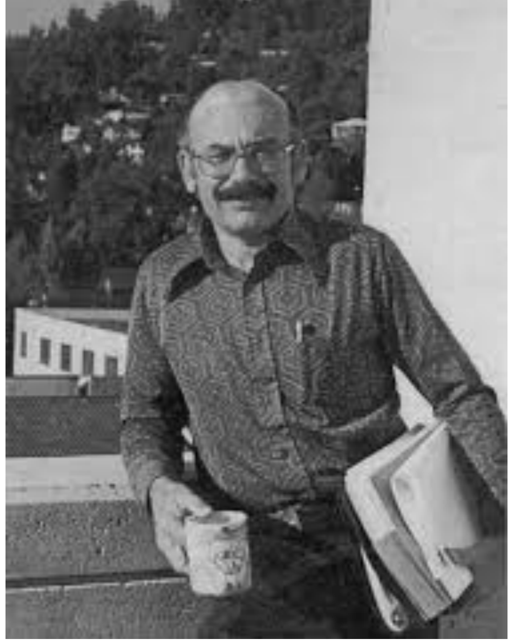
Many questions, few answers

Despite the amazing progress on the engineering side,
theory falls short.



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?



"Reflections after refereeing papers for NIPS",
Leo Breiman, **1995**

Bias-Variance decomposition

For $\ell(y, f_\Theta(x)) = (y - f_\Theta(x))^2$:

$$f_\star(x) = \underset{f}{\operatorname{argmin}}\ \mathscr{R}(f) = \mathbb{E}[y \mid x]$$

Bias-Variance decomposition

For $\ell(y, f_\Theta(x)) = (y - f_\Theta(x))^2$:

$$f_\star(x) = \operatorname*{argmin}_f \mathscr{R}(f) = \mathbb{E}[y \,|\, x]$$

Hence, for $\hat{\Theta} = \hat{\Theta}(X, y)$ the excess risk is given by:

$$\mathscr{R}(\hat{\Theta}) - \mathscr{R}(f_\star) = \mathbb{E}[(f_\star(x) - f(x; \Theta))^2]$$

Bias-Variance decomposition

For $\ell(y, f_{\Theta}(x)) = (y - f_{\Theta}(x))^2$:

$$f_{\star}(x) = \underset{f}{\operatorname{argmin}}\ \mathscr{R}(f) = \mathbb{E}[y \,|\, x]$$

Hence, for $\hat{\Theta} = \hat{\Theta}(X, y)$ the excess risk is given by:

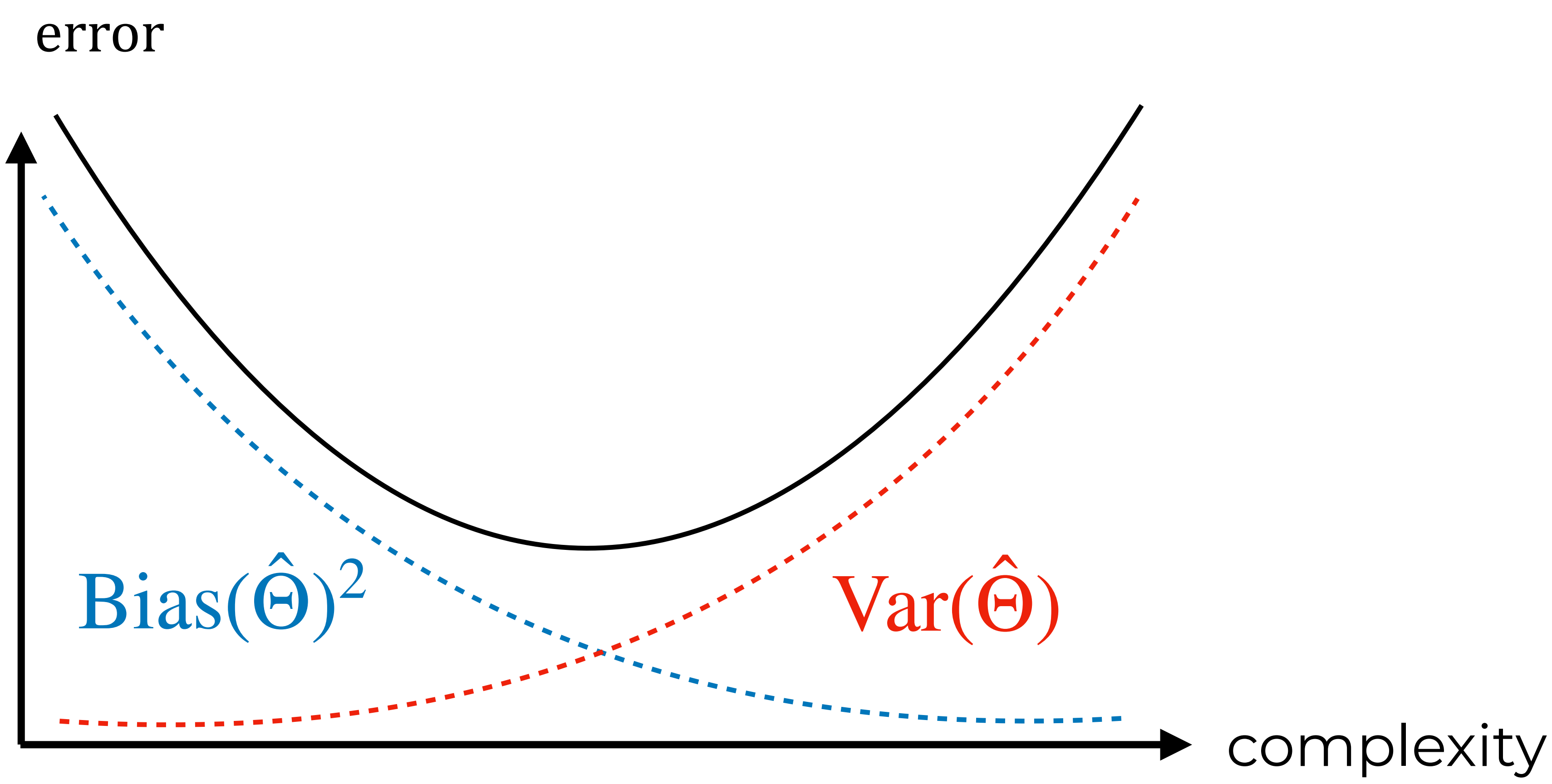
$$\begin{aligned} \mathscr{R}(\hat{\Theta}) - \mathscr{R}(f_{\star}) &= \mathbb{E}[(f_{\star}(x) - f(x;\Theta))^2] \\ &= \mathbb{E}_X[\mathrm{Bias}(\hat{\Theta})^2] + \mathbb{E}_X[\mathrm{Var}(\hat{\Theta})] \end{aligned}$$

Where:

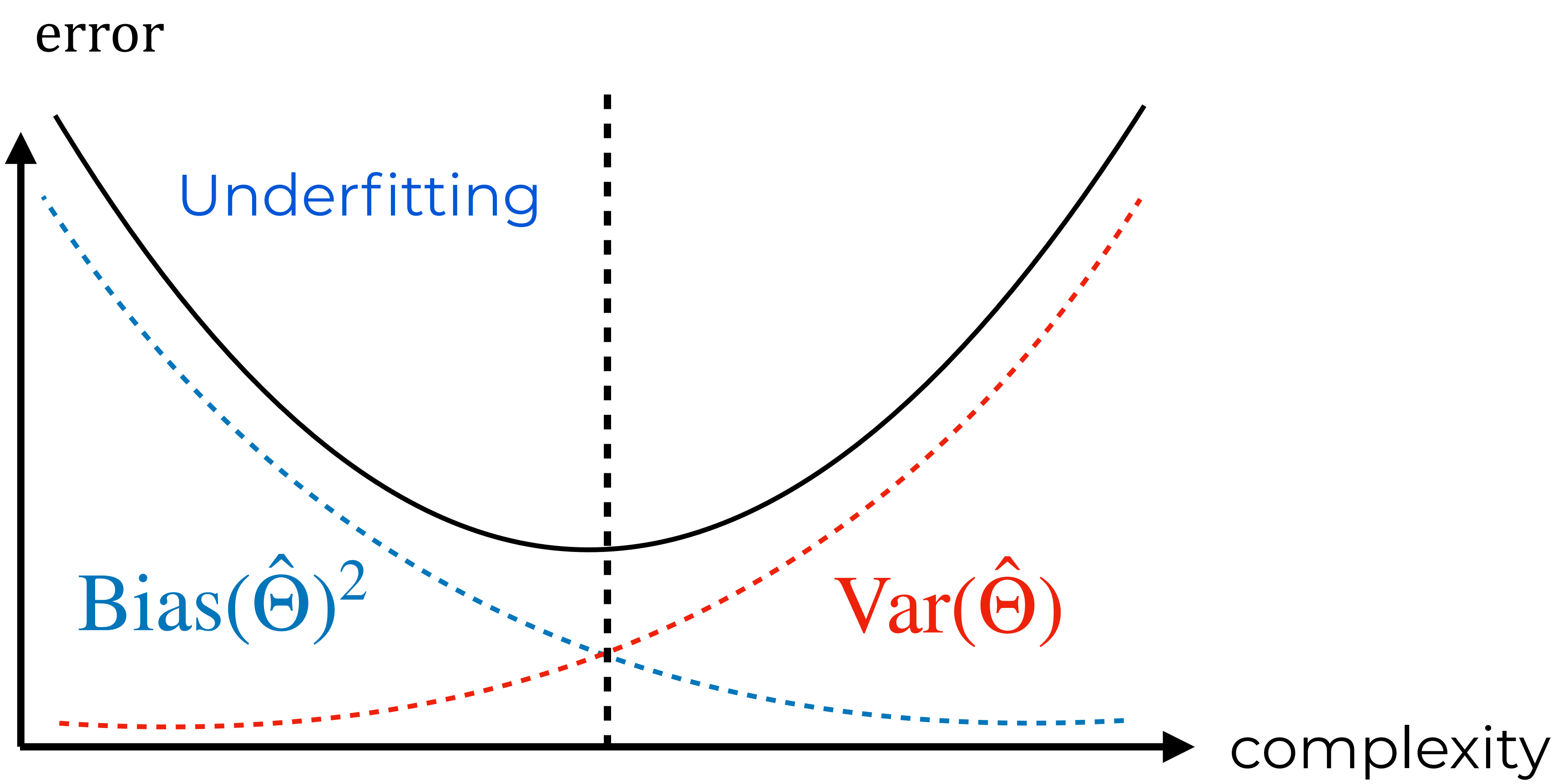
$$\mathrm{Bias}(\hat{\Theta})^2 = \mathbb{E}_x\left[\left(f_{\star}(x) - \mathbb{E}_y\left[f(x;\hat{\Theta})\right]\right)^2\right]$$

$$\mathrm{Var}(\hat{\Theta}) = \mathbb{E}_{x,y}\left[\left(f(x;\hat{\Theta}) - \mathbb{E}_y\left[f(x;\hat{\Theta})\right]\right)^2\right]$$

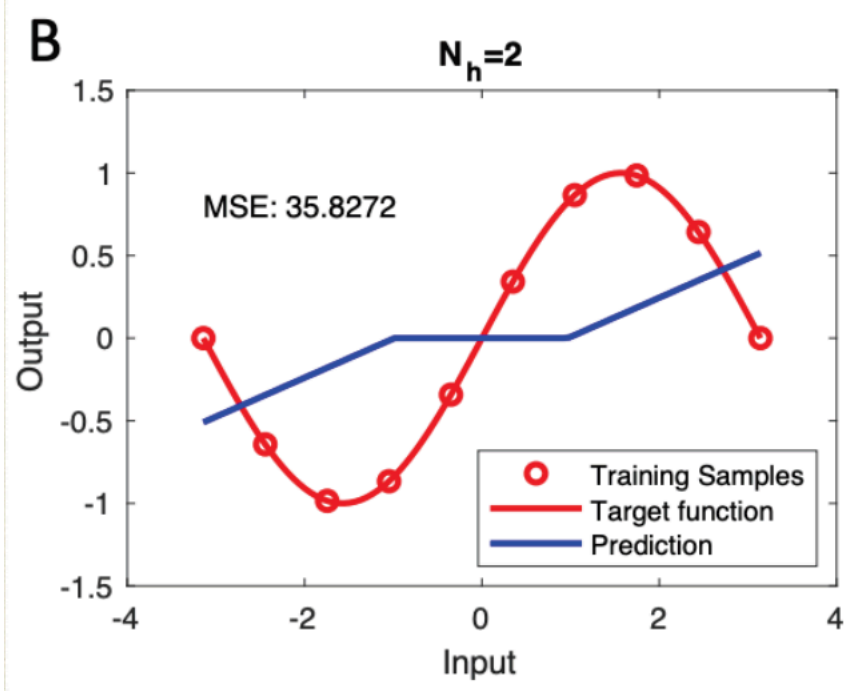
Bias-variance trade-off



Bias-variance trade-off

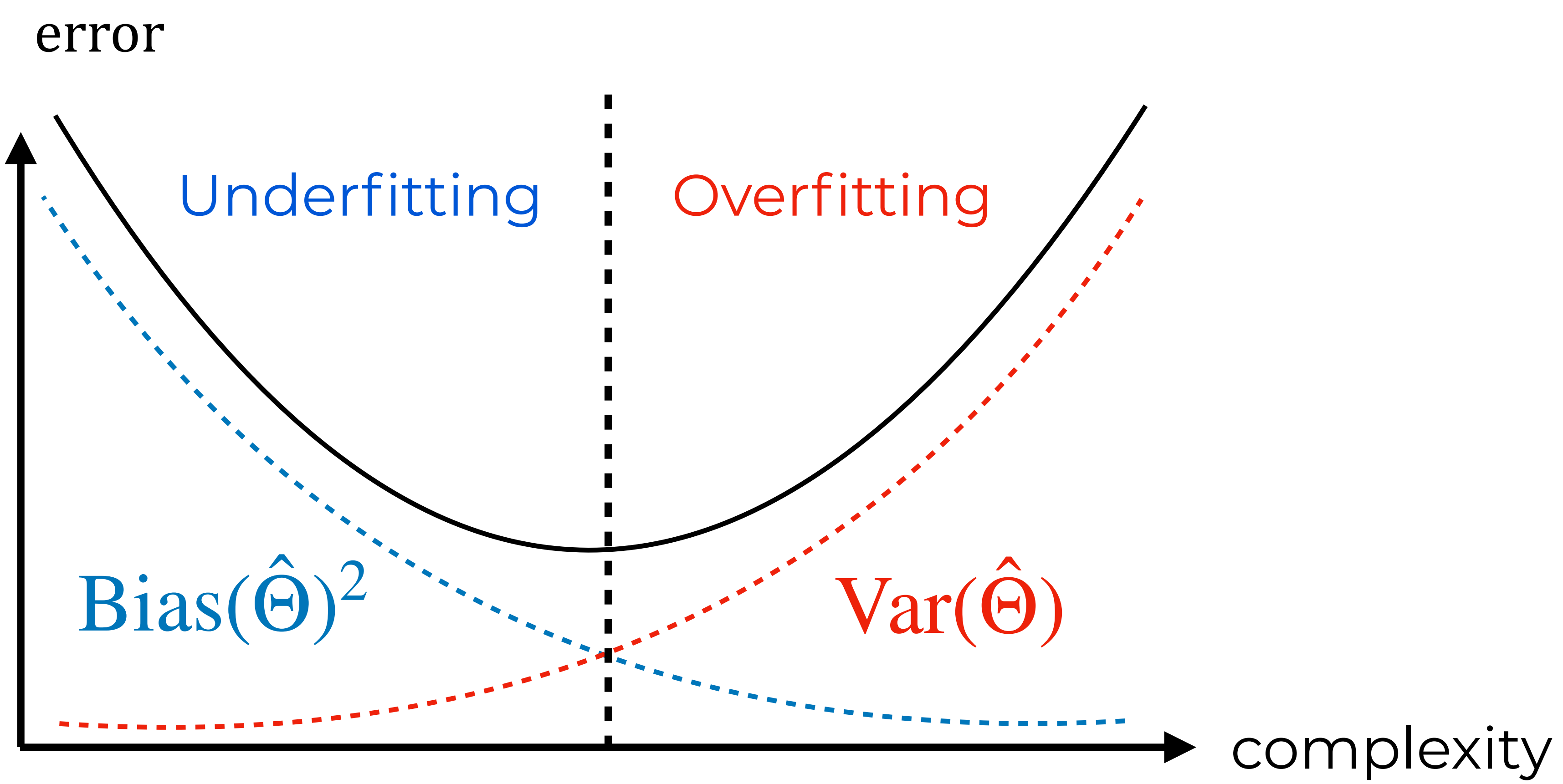


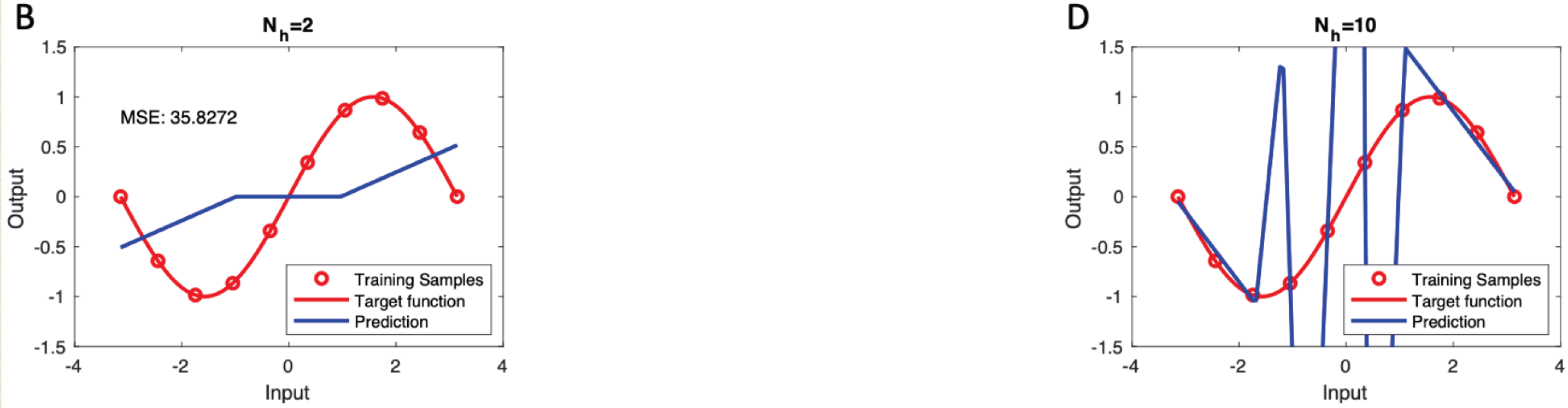
Underfitting



From [Advani, Saxe 17’]

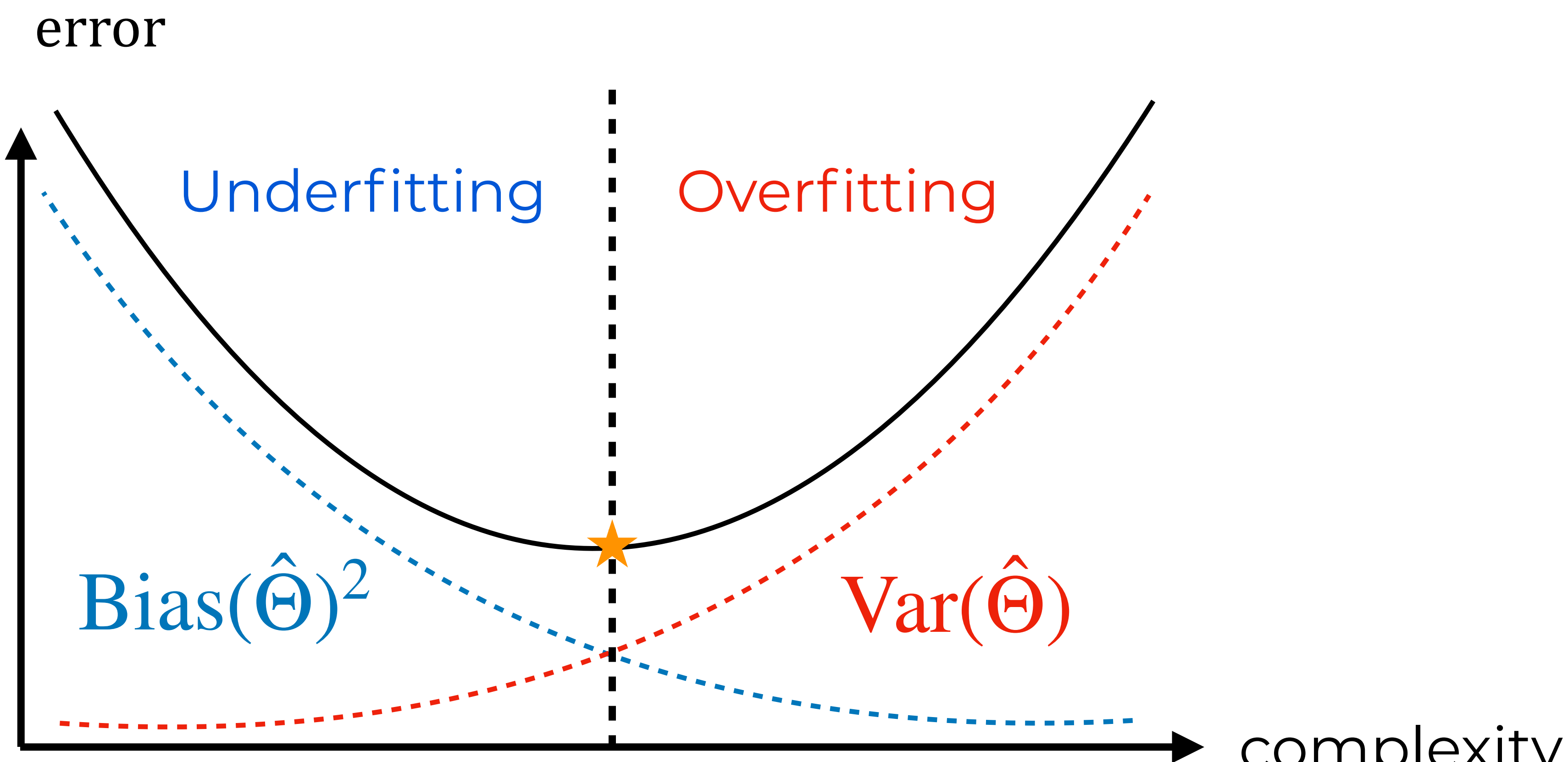
Bias-variance trade-off

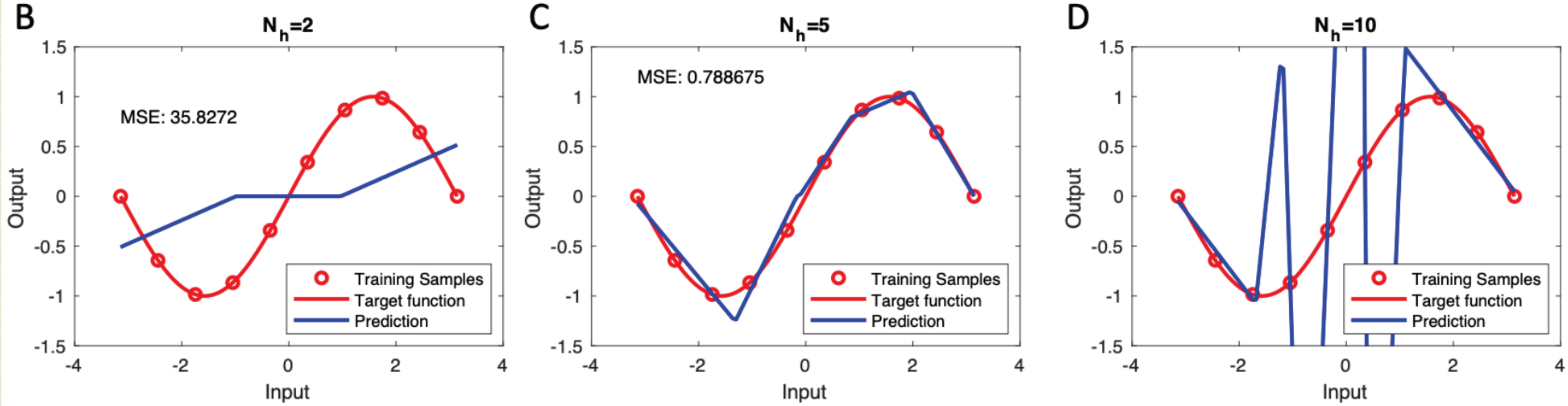




From [Advani, Saxe 17']

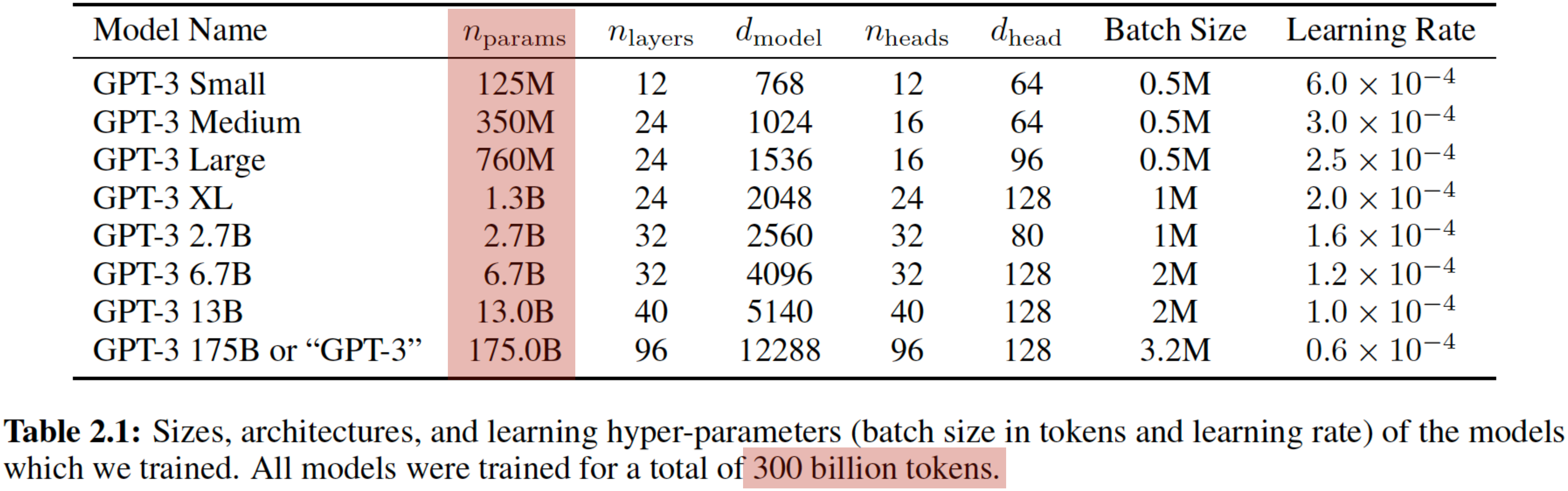
Bias-variance trade-off





From [Advani, Saxe 17']

Bias-variance trade-off



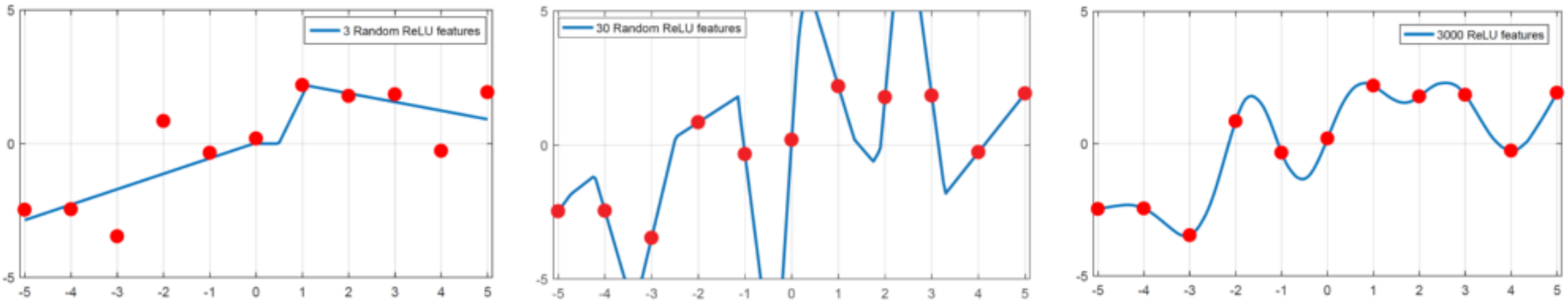
Model Name	n_{params}	n_{layers}	d_{model}	n_{heads}	d_{head}	Batch Size	Learning Rate
GPT-3 Small	125M	12	768	12	64	0.5M	6.0×10^{-4}
GPT-3 Medium	350M	24	1024	16	64	0.5M	3.0×10^{-4}
GPT-3 Large	760M	24	1536	16	96	0.5M	2.5×10^{-4}
GPT-3 XL	1.3B	24	2048	24	128	1M	2.0×10^{-4}
GPT-3 2.7B	2.7B	32	2560	32	80	1M	1.6×10^{-4}
GPT-3 6.7B	6.7B	32	4096	32	128	2M	1.2×10^{-4}
GPT-3 13B	13.0B	40	5140	40	128	2M	1.0×10^{-4}
GPT-3 175B or "GPT-3"	175.0B	96	12288	96	128	3.2M	0.6×10^{-4}

Table 2.1: Sizes, architectures, and learning hyper-parameters (batch size in tokens and learning rate) of the models which we trained. All models were trained for a total of 300 billion tokens.

From [Brown et al 2020]

“Double descent”

[Belkin ’18]



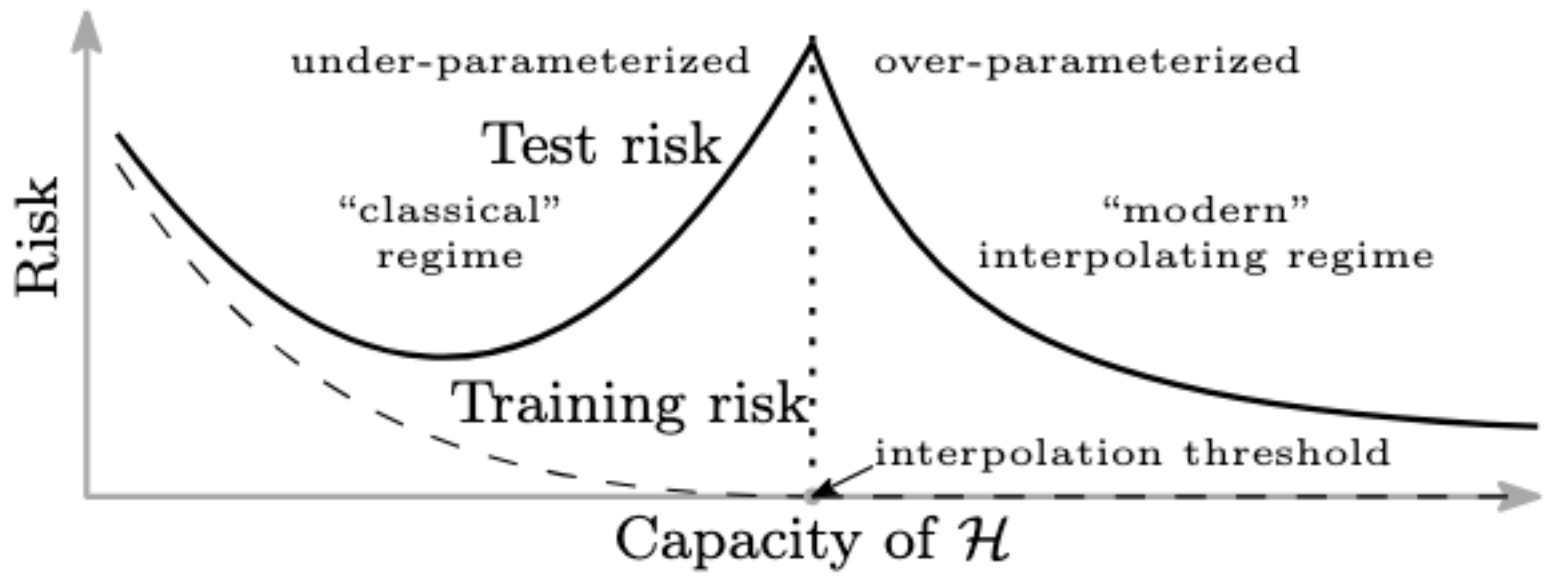
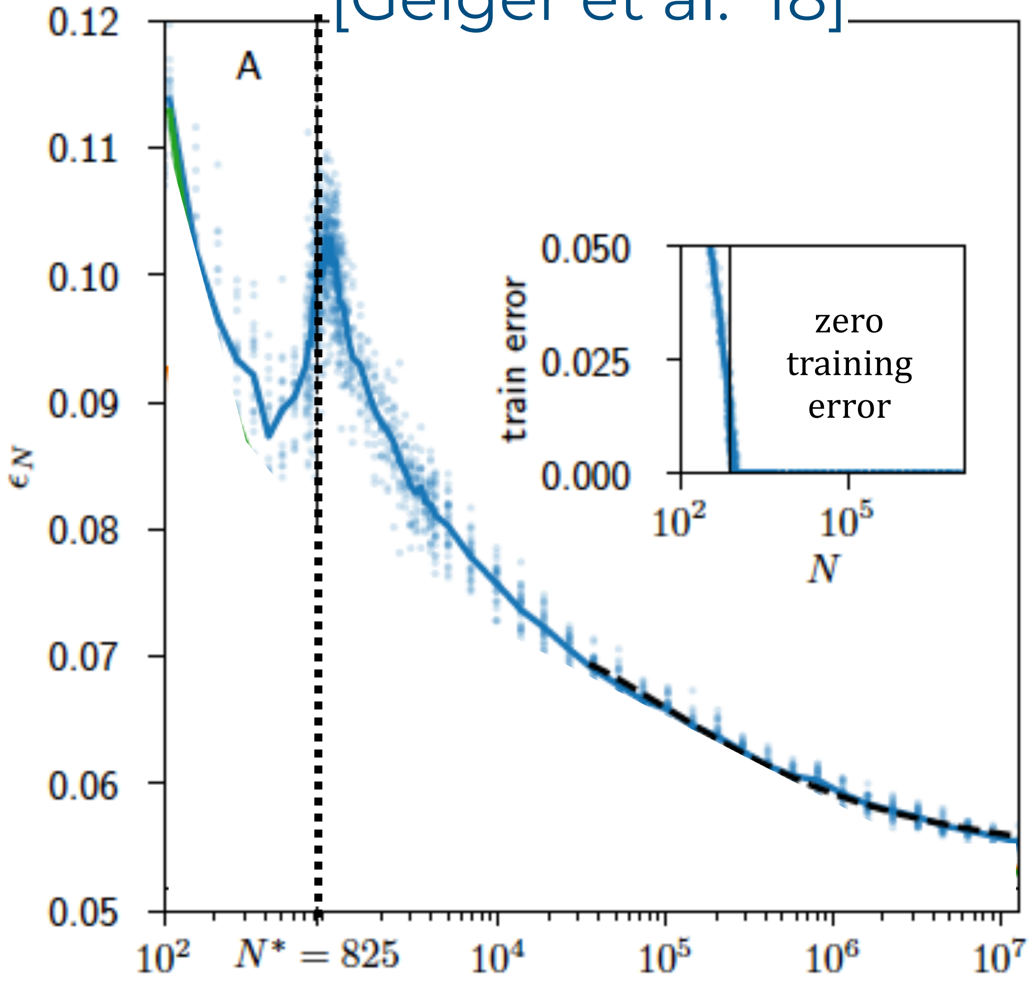


Figure from [Belkin 21’]

“Double descent”

[Belkin ’18]

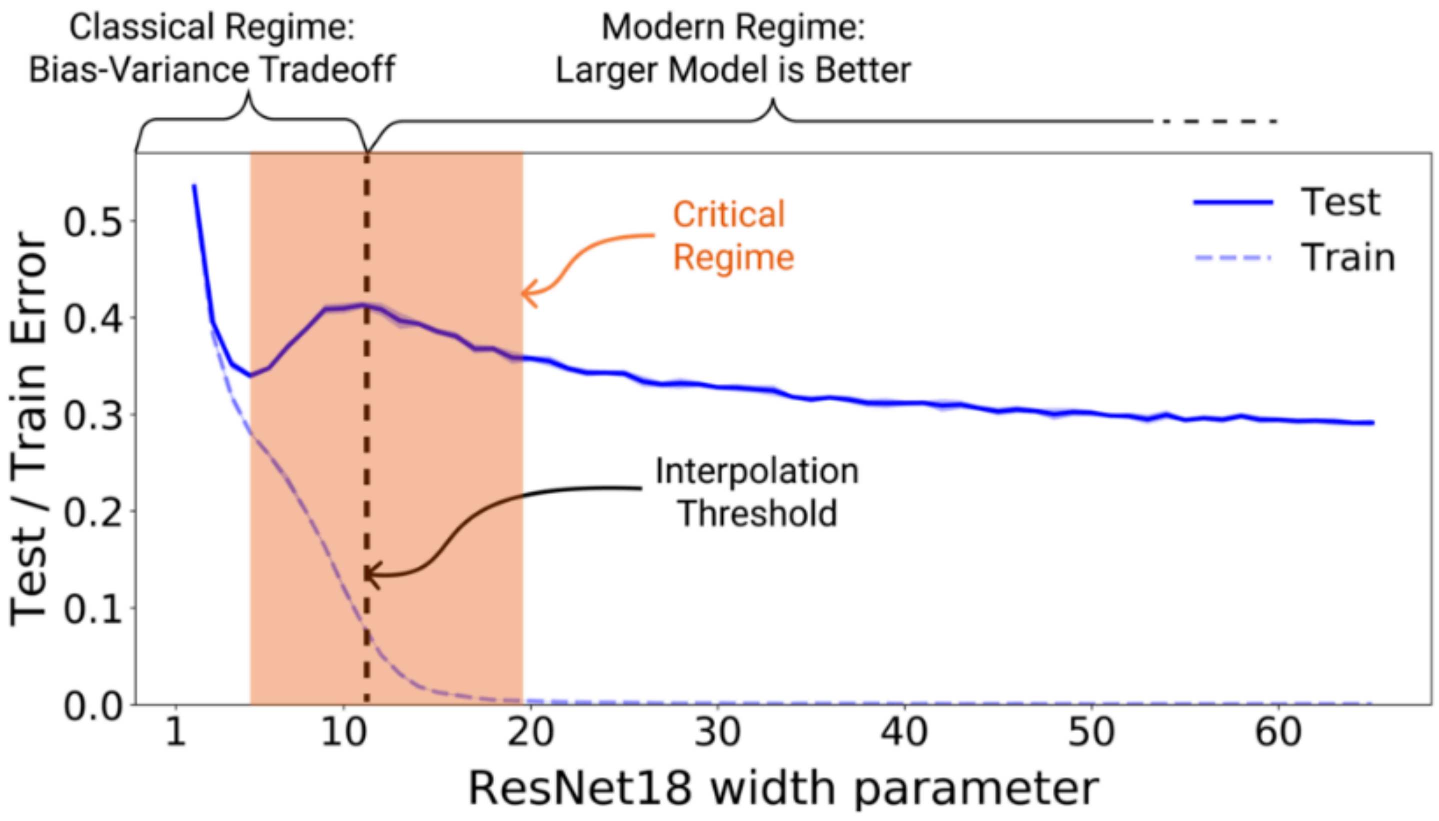
[Geiger et al. ’18]



Number of parameters

Parity-MNIST, 5 layers, fully-connected, no regularisation

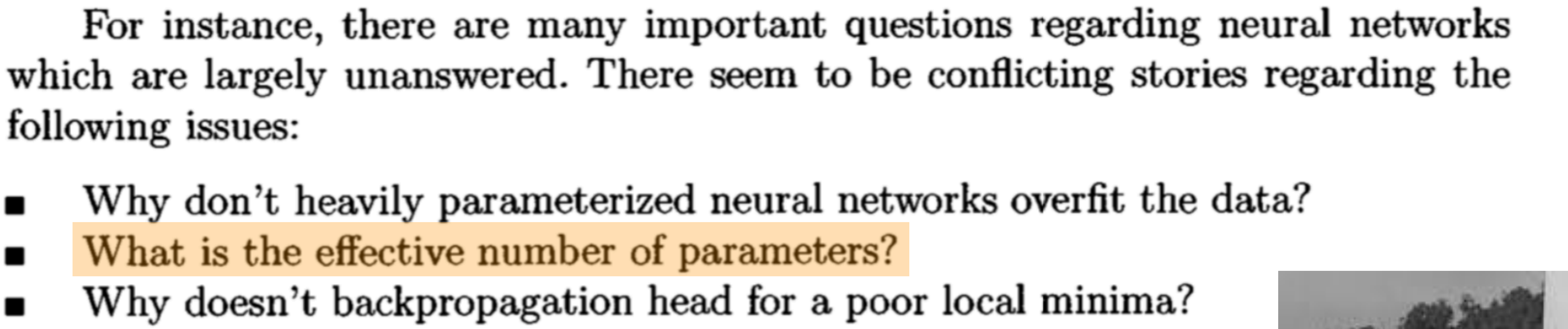
[Nakkiran et al. ’19]



CIFAR10, no regularisation

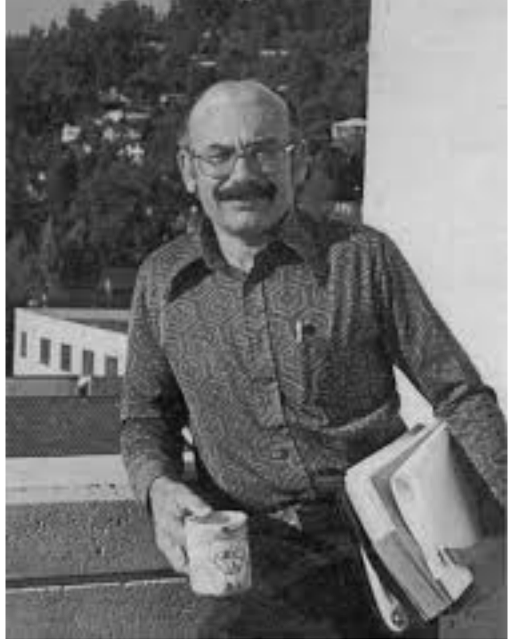
Many questions, few answers

Despite the amazing progress on the engineering side,
theory falls short.



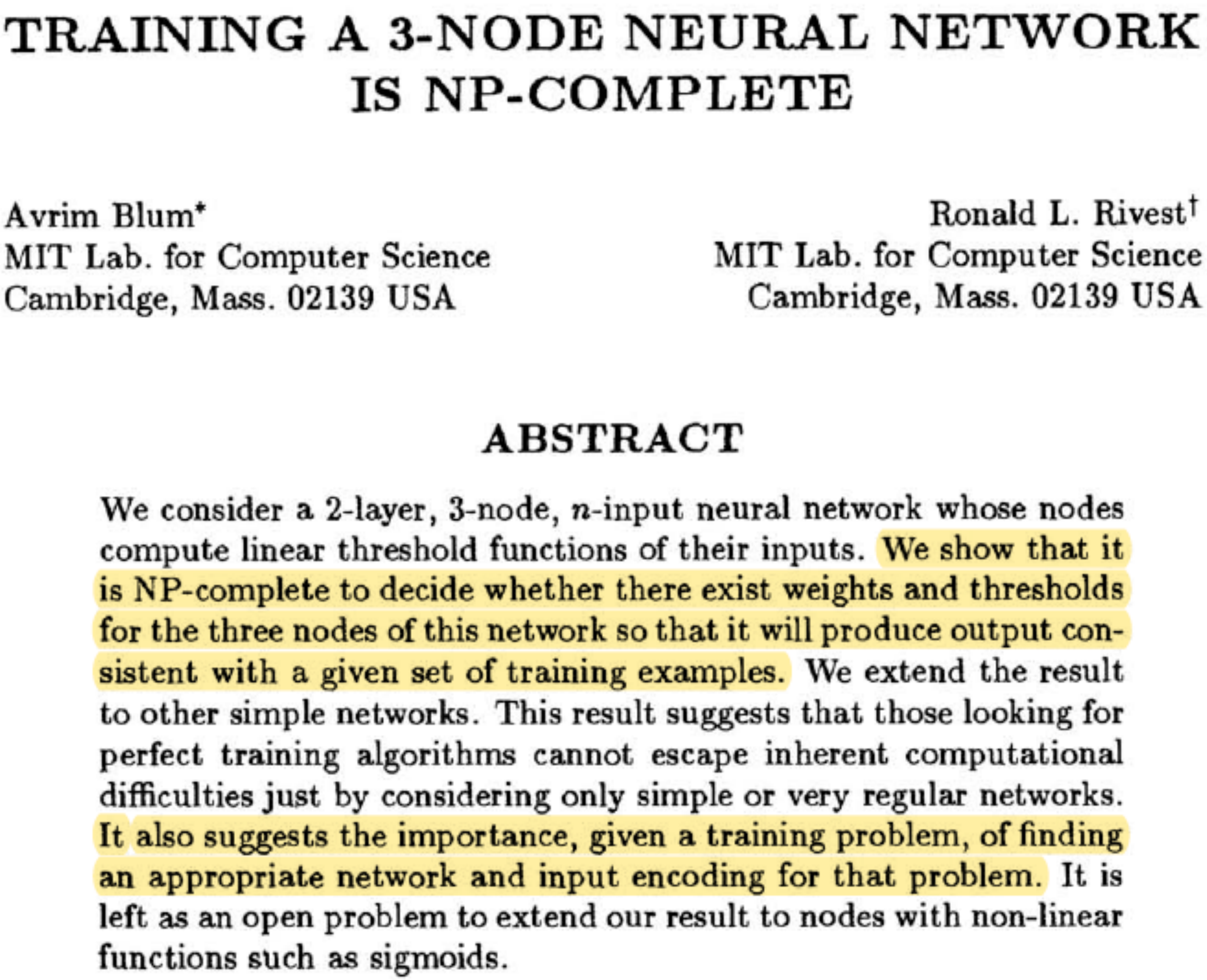
For instance, there are many important questions regarding neural networks which are largely unanswered. There seem to be conflicting stories regarding the following issues:

- 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?



"Reflections after refereeing papers for NIPS",
Leo Breiman, **1995**

Worst case can be hard



TRAINING A 3-NODE NEURAL NETWORK IS NP-COMPLETE

Avrim Blum*
MIT Lab. for Computer Science
Cambridge, Mass. 02139 USA

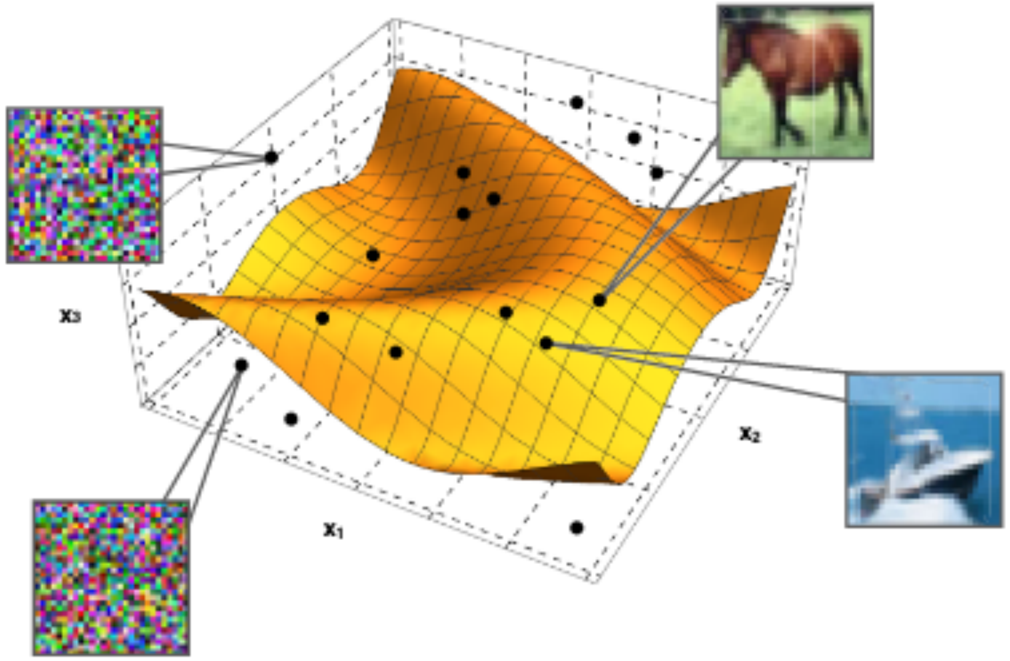
Ronald L. Rivest†
MIT Lab. for Computer Science
Cambridge, Mass. 02139 USA

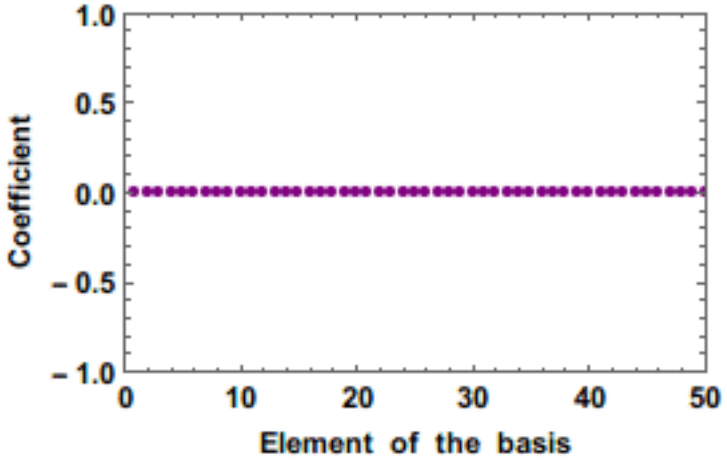
ABSTRACT

We consider a 2-layer, 3-node, n-input neural network whose nodes compute linear threshold functions of their inputs. We show that it is NP-complete to decide whether there exist weights and thresholds for the three nodes of this network so that it will produce output consistent with a given set of training examples. We extend the result to other simple networks. This result suggests that those looking for perfect training algorithms cannot escape inherent computational difficulties just by considering only simple or very regular networks. It also suggests the importance, given a training problem, of finding an appropriate network and input encoding for that problem. It is left as an open problem to extend our result to nodes with non-linear functions such as sigmoids.

Effective dimension?

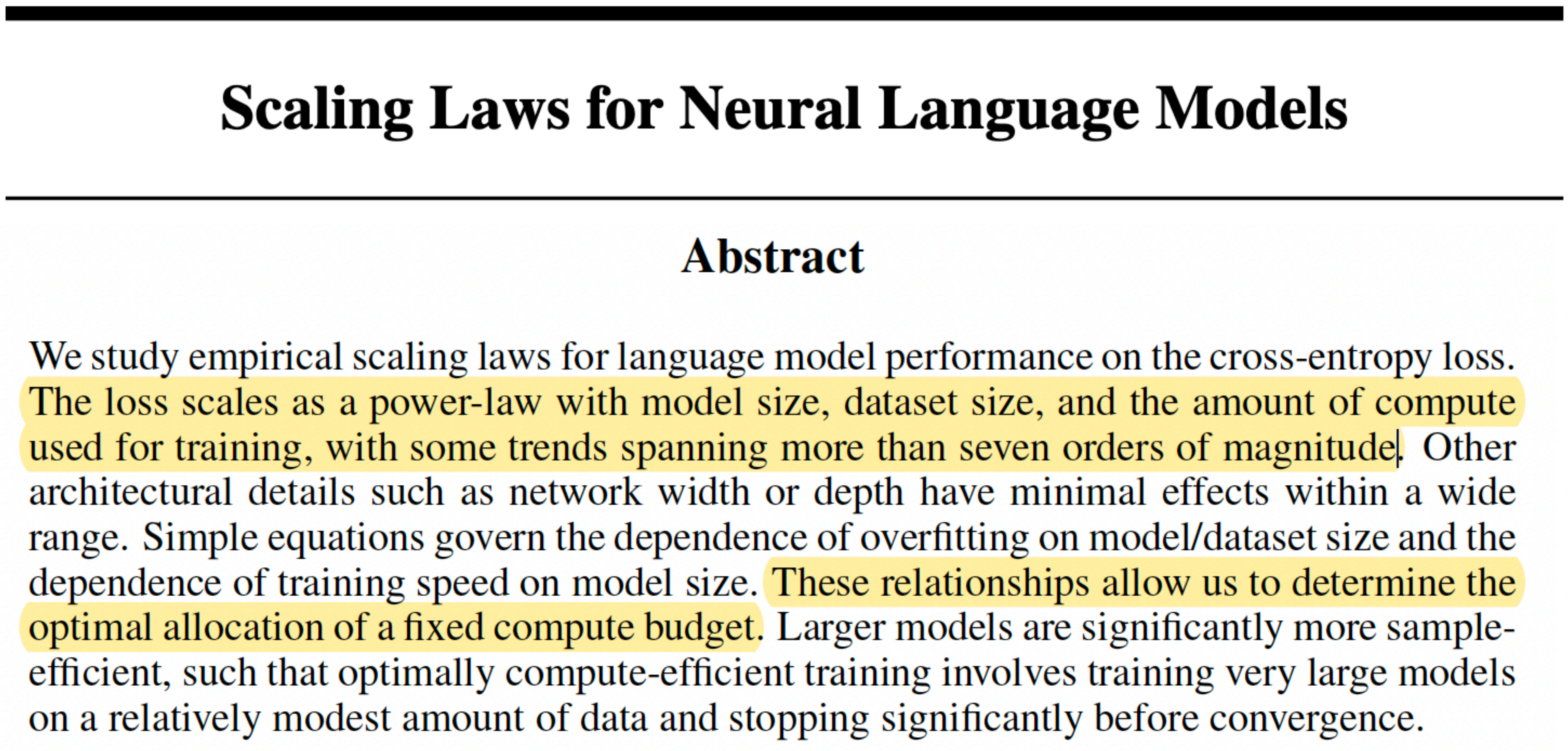
How many features / samples needed to correctly learn?





Neural scaling laws

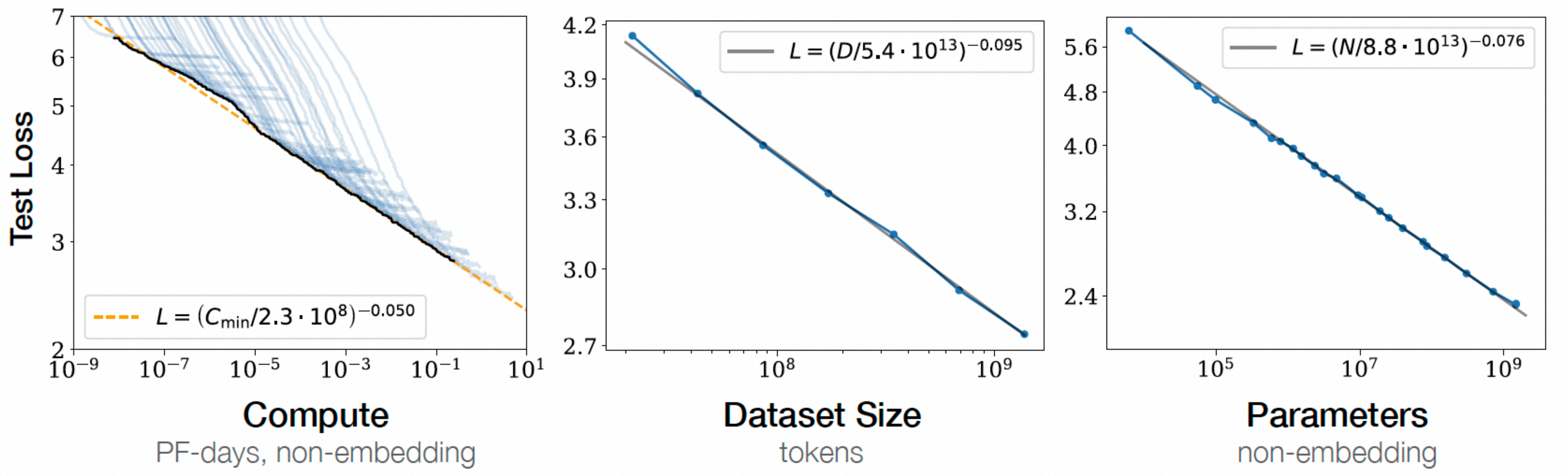
[Kaplan et al. '20]



Scaling Laws for Neural Language Models

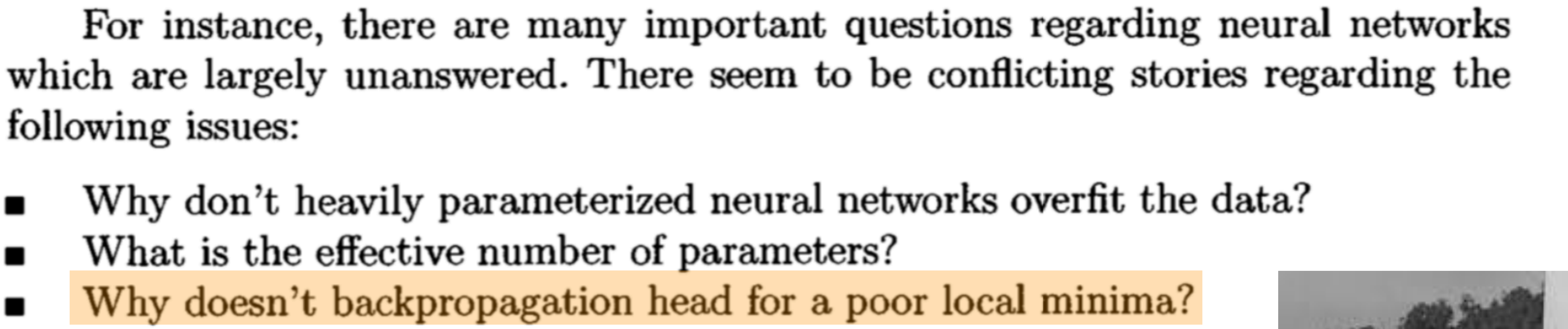
Abstract

We study empirical scaling laws for language model performance on the cross-entropy loss. The loss scales as a power-law with model size, dataset size, and the amount of compute used for training, with some trends spanning more than seven orders of magnitude. Other architectural details such as network width or depth have minimal effects within a wide range. Simple equations govern the dependence of overfitting on model/dataset size and the dependence of training speed on model size. These relationships allow us to determine the optimal allocation of a fixed compute budget. Larger models are significantly more sample-efficient, such that optimally compute-efficient training involves training very large models on a relatively modest amount of data and stopping significantly before convergence.



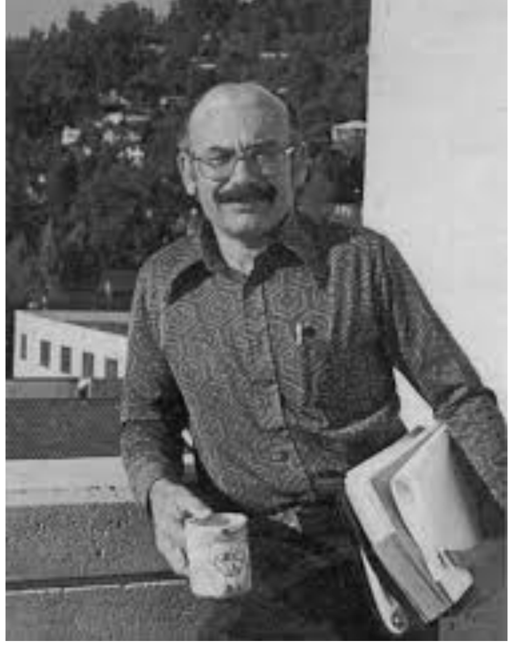
Many questions, few answers

Despite the amazing progress on the engineering side,
theory falls short.



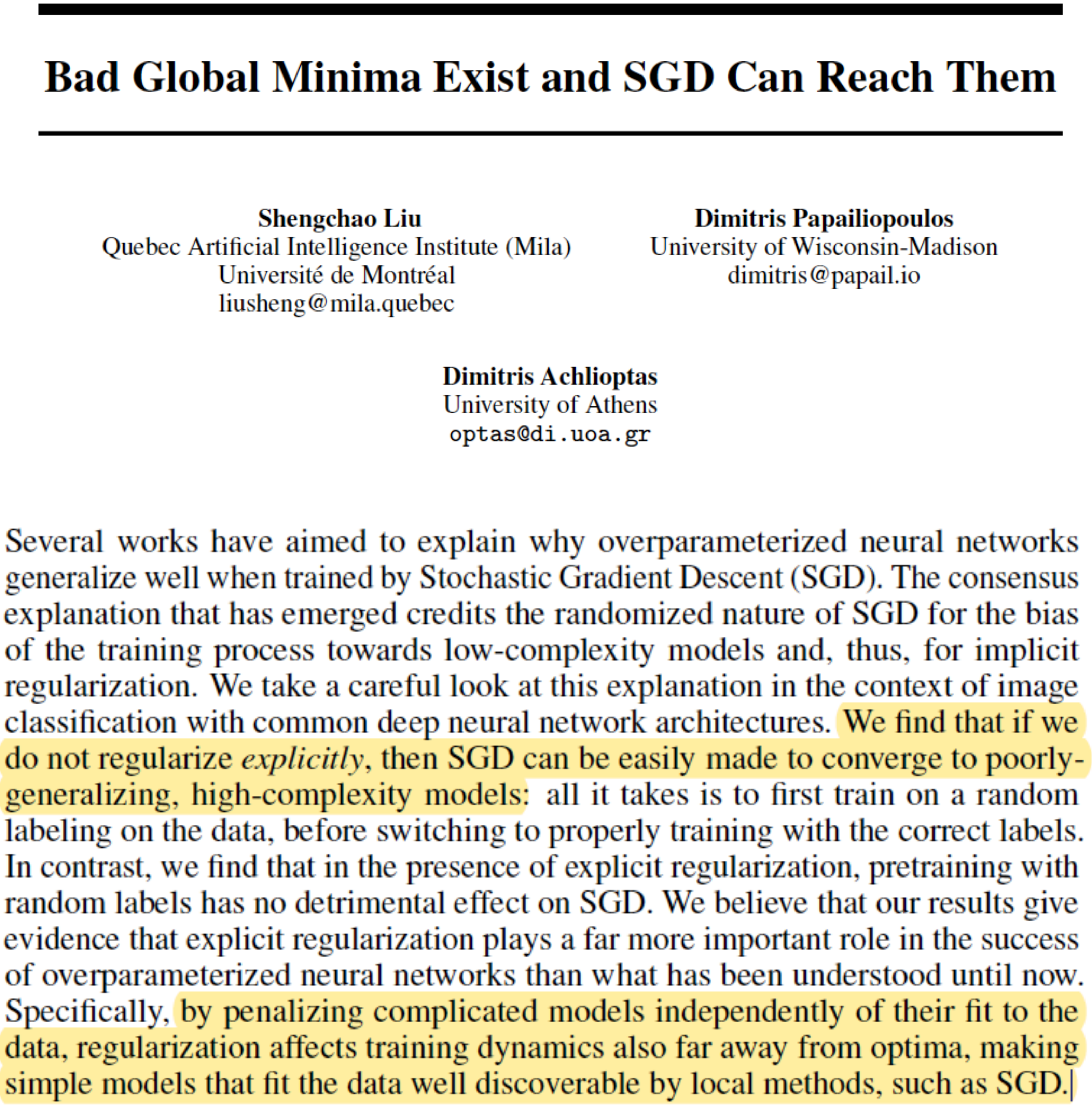
For instance, there are many important questions regarding neural networks which are largely unanswered. There seem to be conflicting stories regarding the following issues:

- 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?



"Reflections after refereeing papers for NIPS",
Leo Breiman, **1995**

Bad minima exist



Bad Global Minima Exist and SGD Can Reach Them

Shengchao Liu
Quebec Artificial Intelligence Institute (Mila)
Université de Montréal
liusheng@mila.quebec

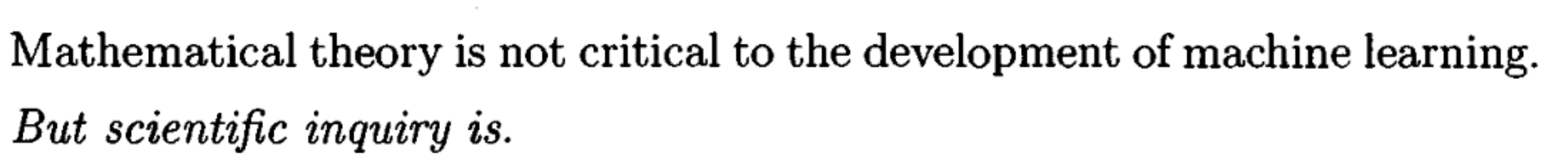
Dimitris Papailiopoulos
University of Wisconsin-Madison
dimitris@papail.io

Dimitris Achlioptas
University of Athens
optas@di.uoa.gr

Several works have aimed to explain why overparameterized neural networks generalize well when trained by Stochastic Gradient Descent (SGD). The consensus explanation that has emerged credits the randomized nature of SGD for the bias of the training process towards low-complexity models and, thus, for implicit regularization. We take a careful look at this explanation in the context of image classification with common deep neural network architectures. We find that if we do not regularize *explicitly*, then SGD can be easily made to converge to poorly-generalizing, high-complexity models: all it takes is to first train on a random labeling on the data, before switching to properly training with the correct labels. In contrast, we find that in the presence of explicit regularization, pretraining with random labels has no detrimental effect on SGD. We believe that our results give evidence that explicit regularization plays a far more important role in the success of overparameterized neural networks than what has been understood until now. Specifically, by penalizing complicated models independently of their fit to the data, regularization affects training dynamics also far away from optima, making simple models that fit the data well discoverable by local methods, such as SGD.

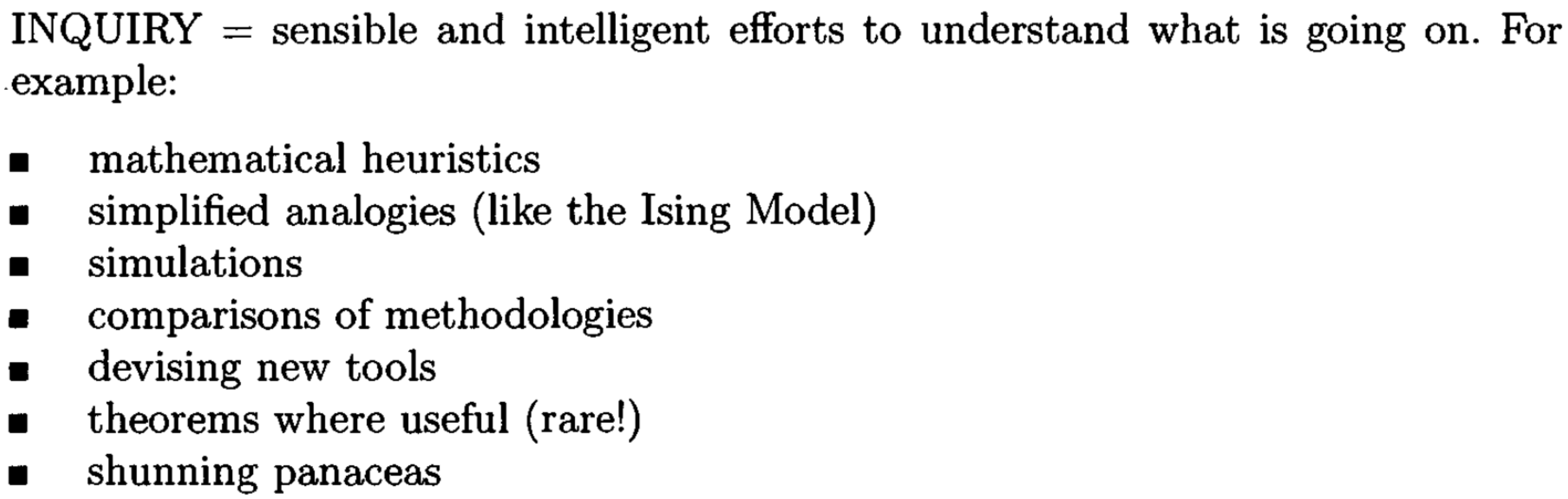
Breiman’s suggestions

“Reflections after refereeing papers for NIPS”, Leo Breiman, **1995**



Mathematical theory is not critical to the development of machine learning.

But scientific inquiry is.

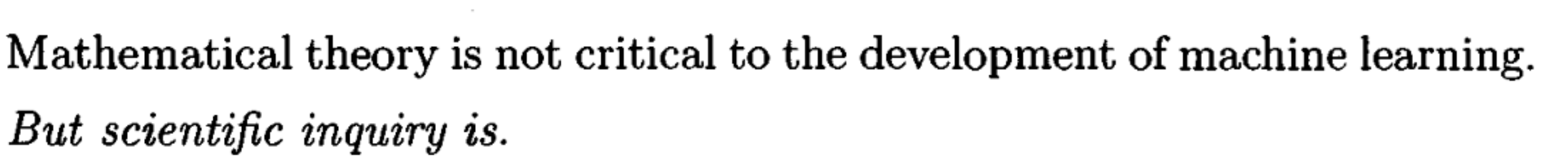


INQUIRY = sensible and intelligent efforts to understand what is going on. For example:

- mathematical heuristics
- simplified analogies (like the Ising Model)
- simulations
- comparisons of methodologies
- devising new tools
- theorems where useful (rare!)
- shunning panaceas

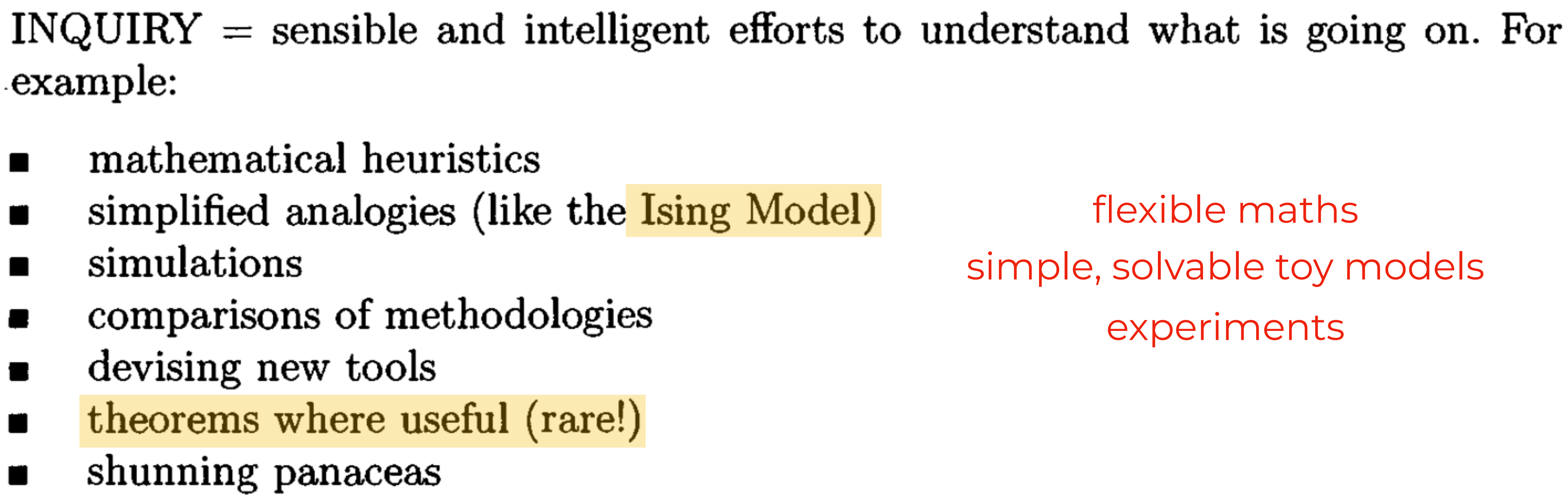
Breiman's suggestions

"Reflections after refereeing papers for NIPS", Leo Breiman, **1995**



Mathematical theory is not critical to the development of machine learning.

But scientific inquiry is.



INQUIRY = sensible and intelligent efforts to understand what is going on. For example:

- mathematical heuristics
- simplified analogies (like the Ising Model)
- simulations
- comparisons of methodologies
- devising new tools
- theorems where useful (rare!)
- shunning panaceas

flexible maths

simple, solvable toy models

experiments

Smells of… physics.

Partial summary

Physics provide a conceptual framework to think about the challenges in ML.

Partial summary

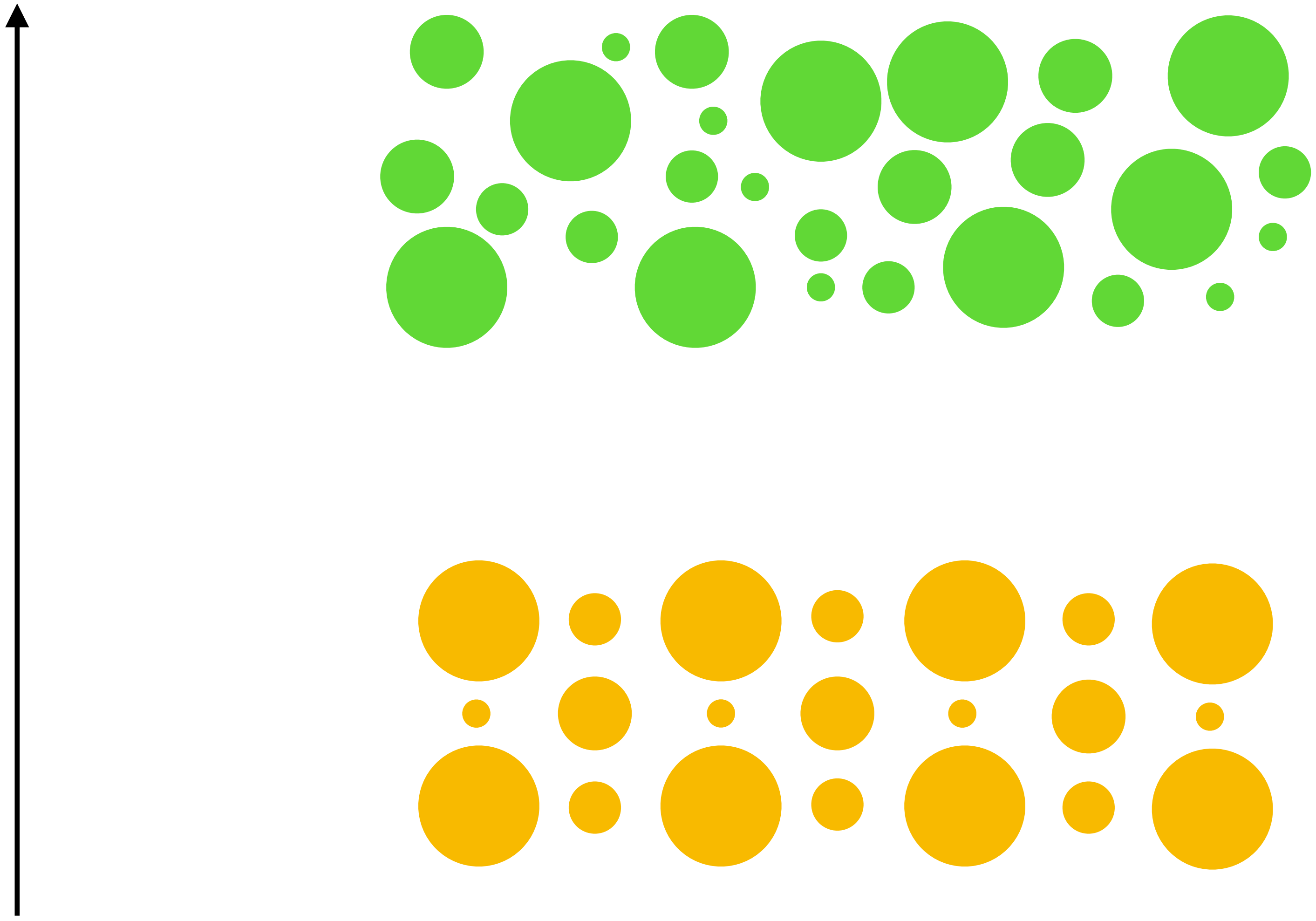
Physics provide a conceptual framework to think about the challenges in ML.

What about tools?

Physics of glasses

Temperature

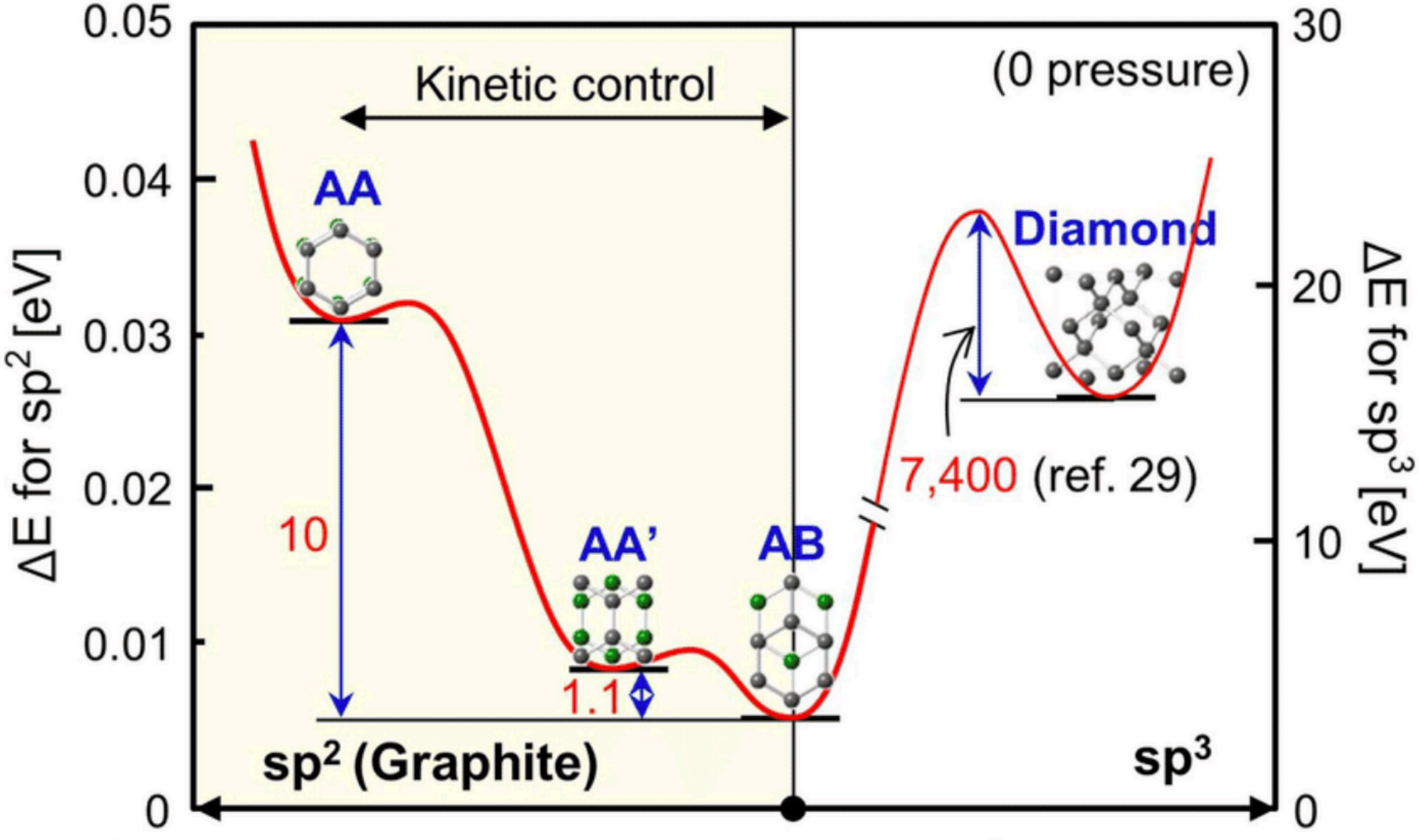
“Liquid”



“Solid”

“Melting transition”

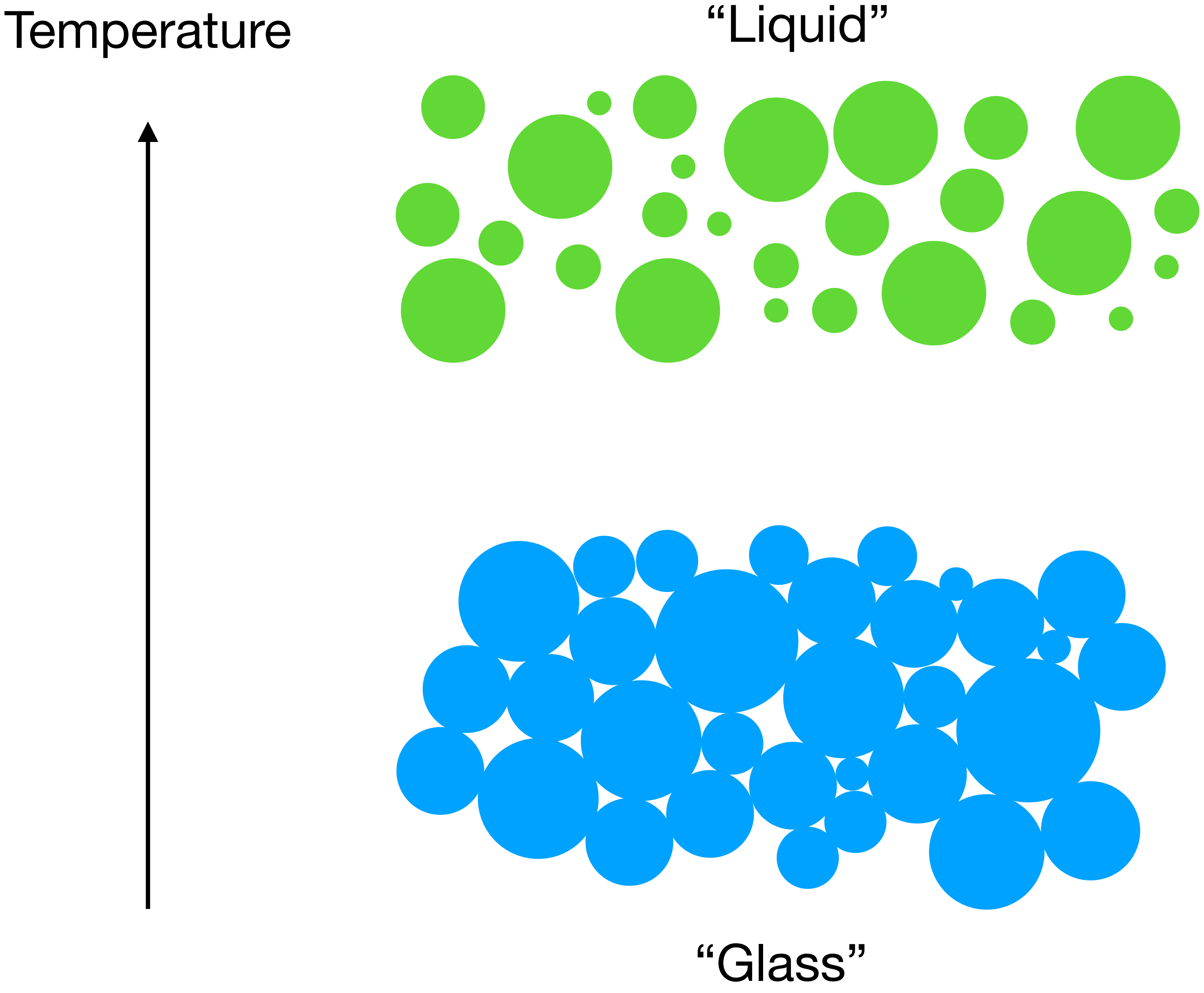
Physics of glasses



$$\tau \propto e^{\Delta E}$$

“Arrhenius law”

Physics of glasses

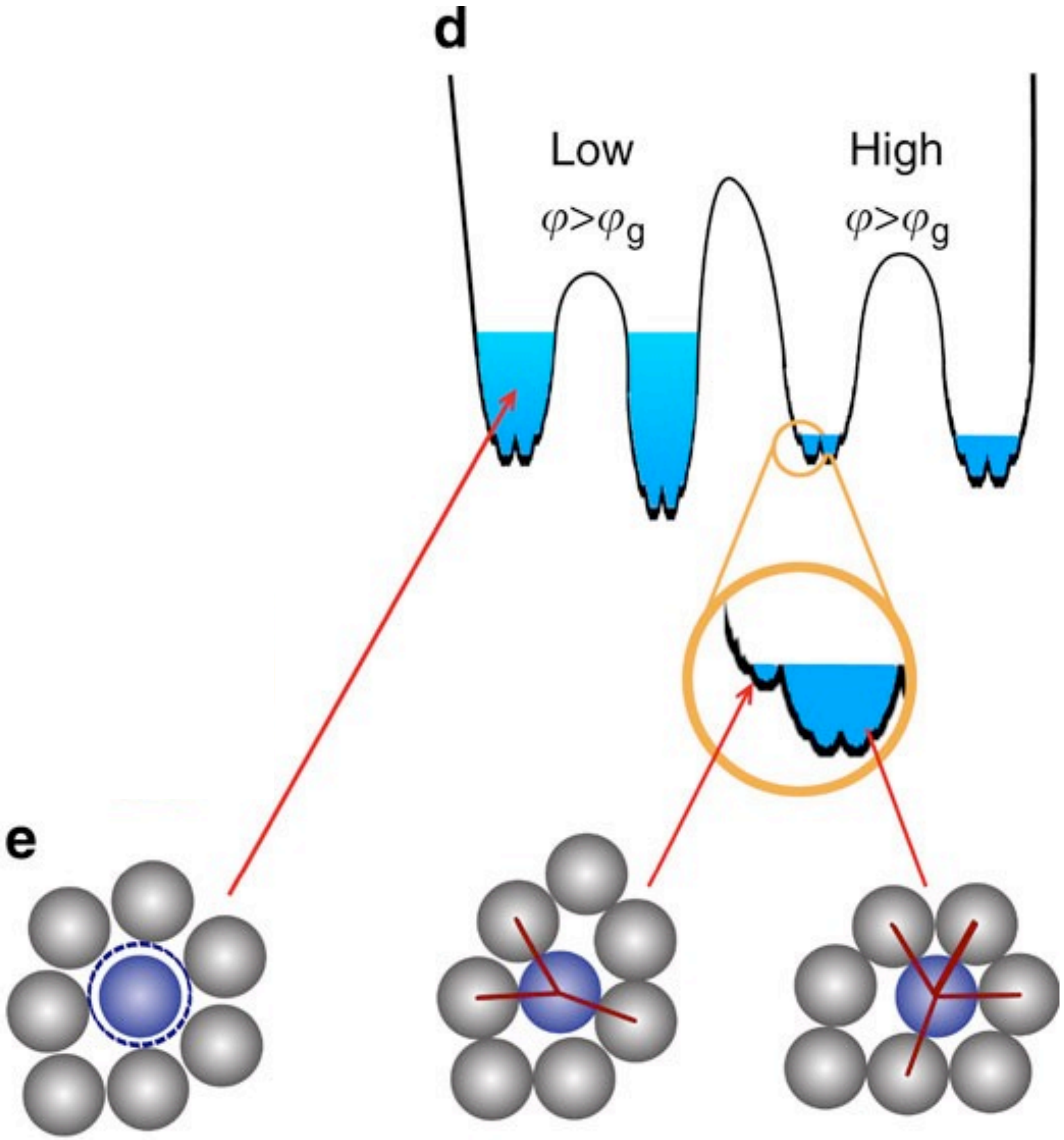


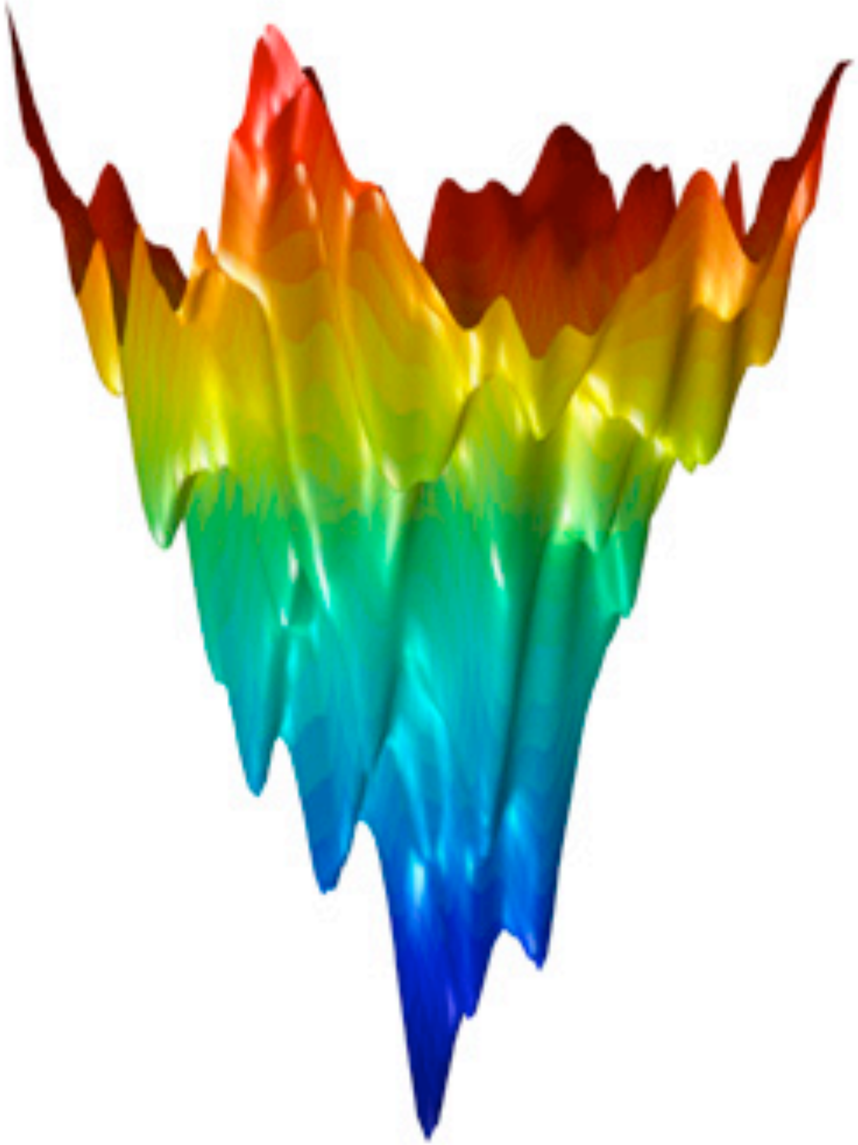
Temperature

“Liquid”

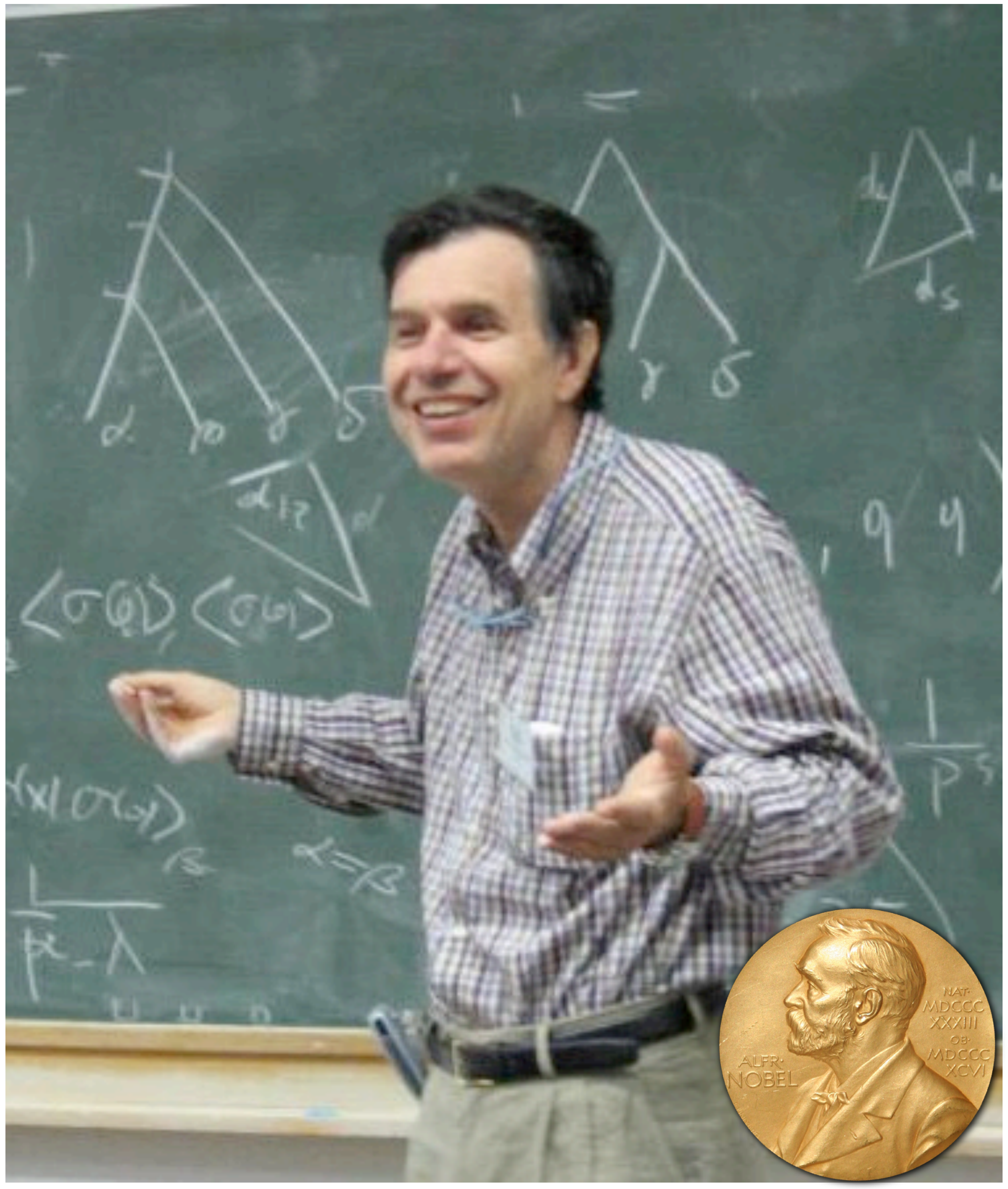
“Glass”

Physics of glasses



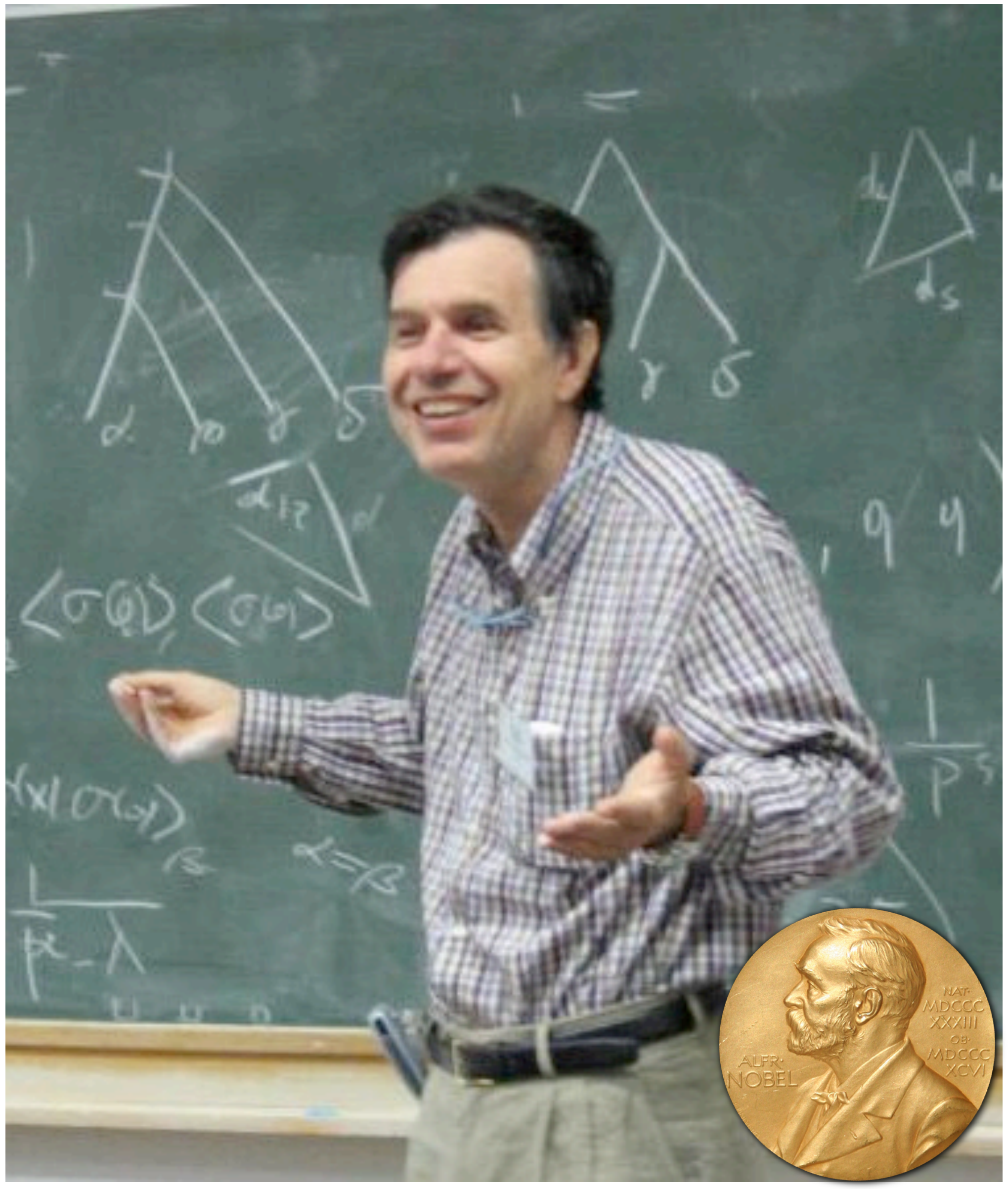


Giorgio Parisi



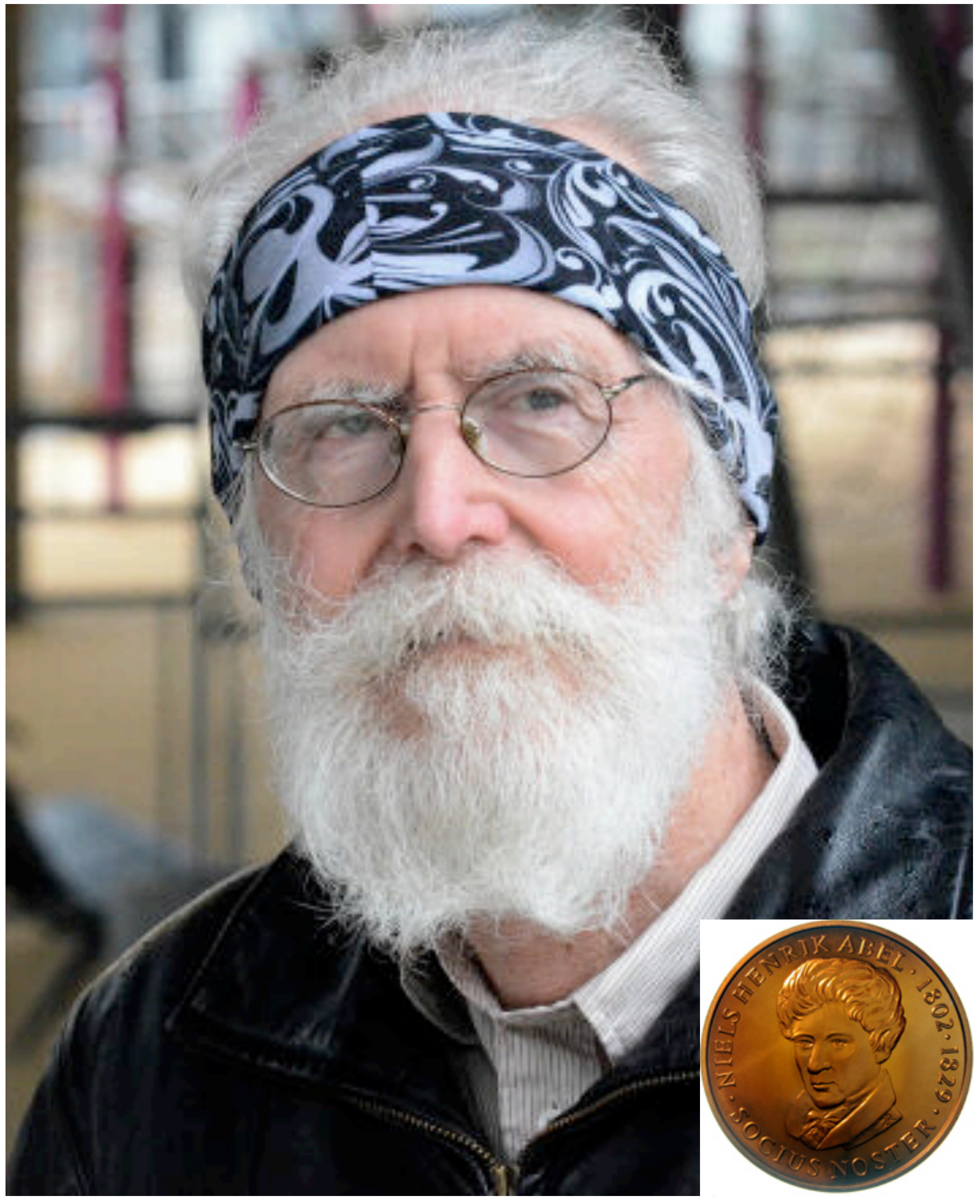
"They make it possible to understand and describe many different and apparently entirely random materials and phenomena, not only in physics but also in other, very different areas, such as mathematics, biology, neuroscience and machine learning."

Giorgio Parisi



“They make it possible to understand and describe many different and apparently entirely random materials and phenomena, not only in physics but also in other, very different areas, such as mathematics, biology, neuroscience and machine learning.”

Michel Talagrand

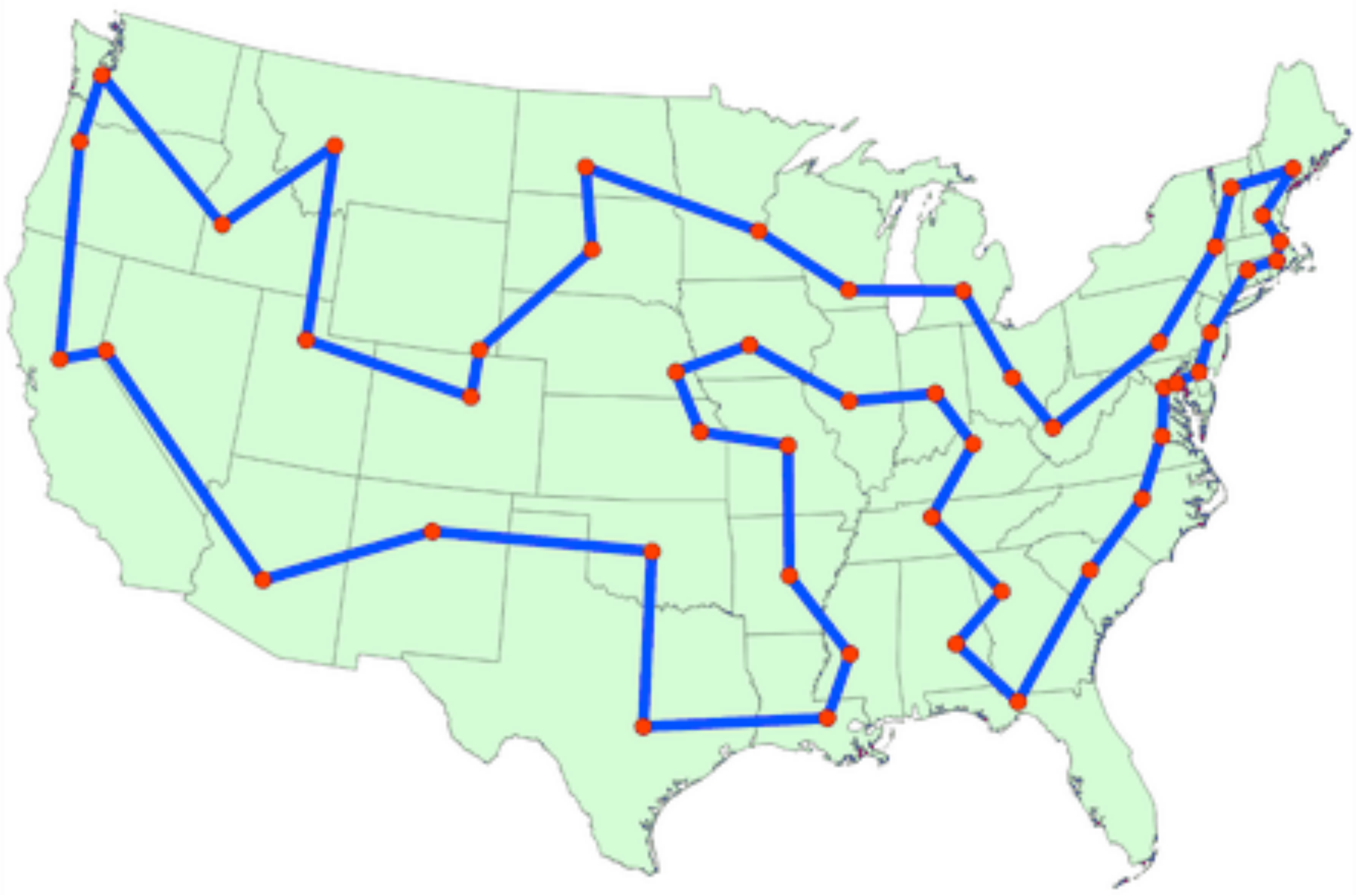


“Talagrand used his knowledge of statistics and probability to prove limits on how spin glass matter can behave, and thereby completed the proof of Giorgio Parisi’s Nobel Prize winning work (2021).”

Other type of “glasses”

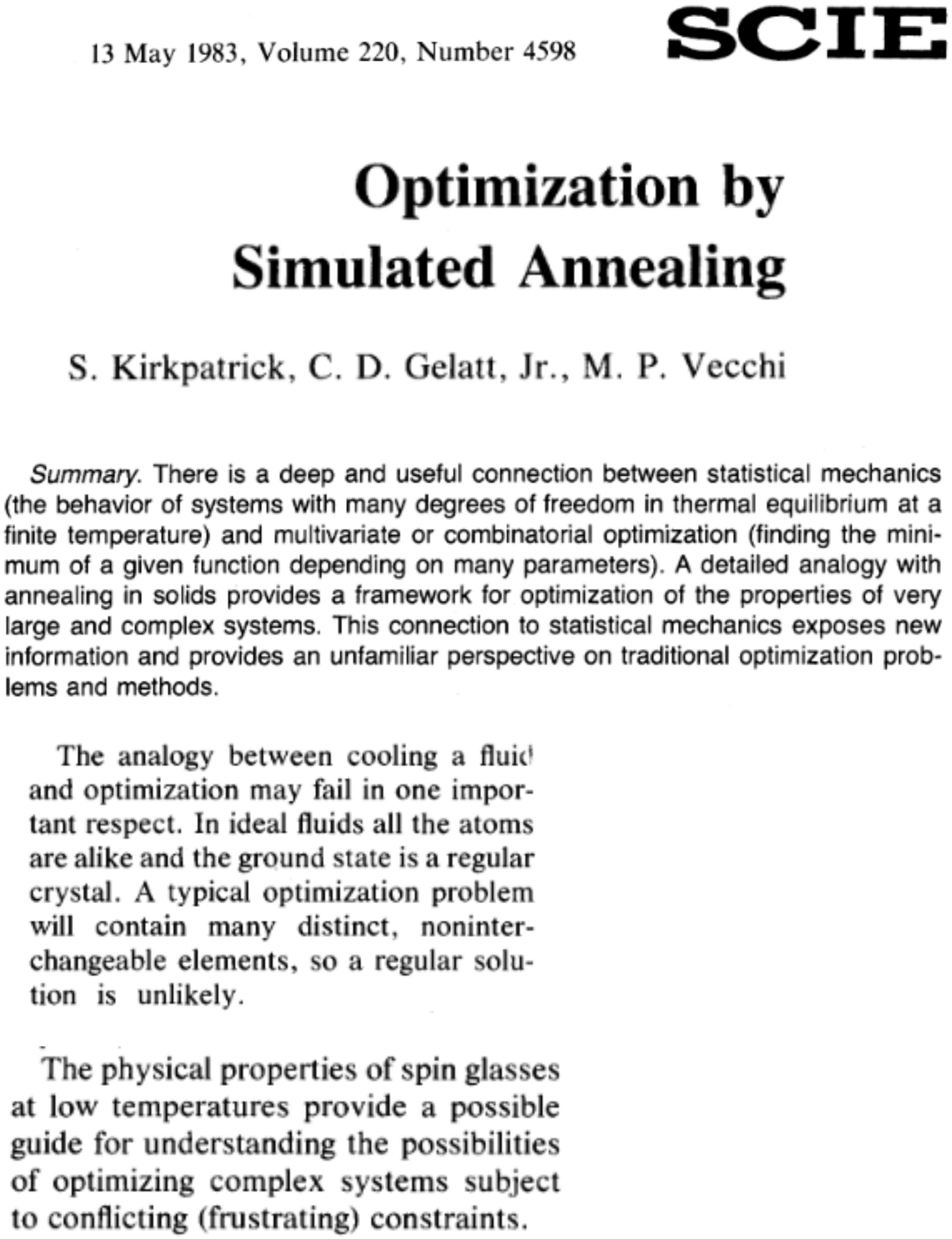
Traveling salesman problem:

“Given a list of cities and the distances between each pair of cities, what is the shortest possible route that visits each city exactly once and returns to the origin city?”



Simulated annealing

1983



13 May 1983, Volume 220, Number 4598 **SCIENCE**

Optimization by Simulated Annealing

S. Kirkpatrick, C. D. Gelatt, Jr., M. P. Vecchi

Summary. There is a deep and useful connection between statistical mechanics (the behavior of systems with many degrees of freedom in thermal equilibrium at a finite temperature) and multivariate or combinatorial optimization (finding the minimum of a given function depending on many parameters). A detailed analogy with annealing in solids provides a framework for optimization of the properties of very large and complex systems. This connection to statistical mechanics exposes new information and provides an unfamiliar perspective on traditional optimization problems and methods.

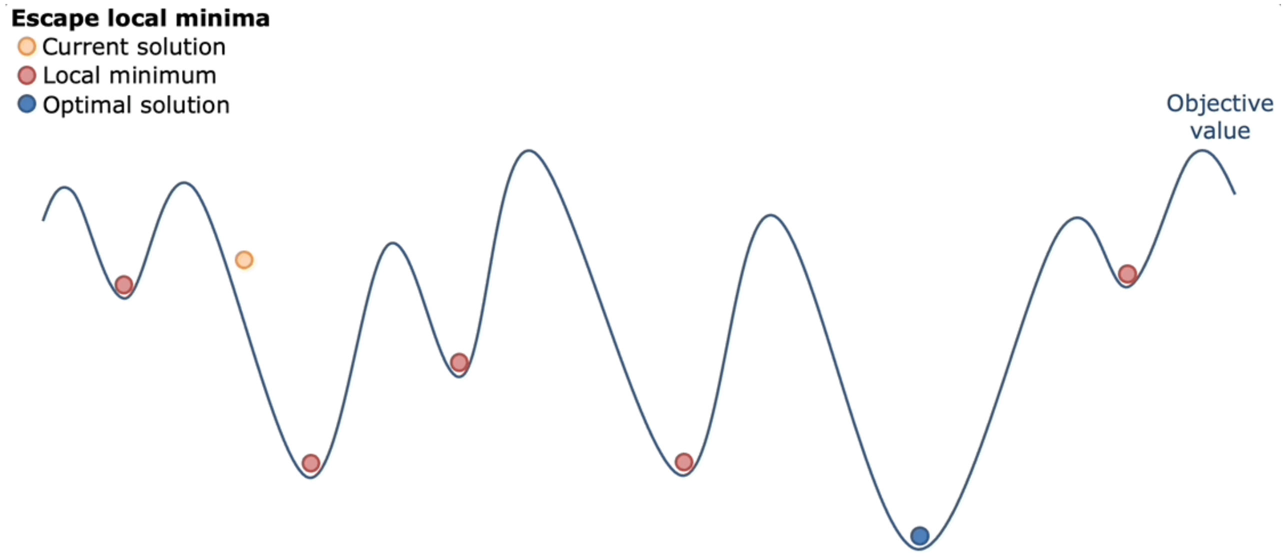
The analogy between cooling a fluid and optimization may fail in one important respect. In ideal fluids all the atoms are alike and the ground state is a regular crystal. A typical optimization problem will contain many distinct, noninterchangeable elements, so a regular solution is unlikely.

The physical properties of spin glasses at low temperatures provide a possible guide for understanding the possibilities of optimizing complex systems subject to conflicting (frustrating) constraints.

S. Kirkpatrick

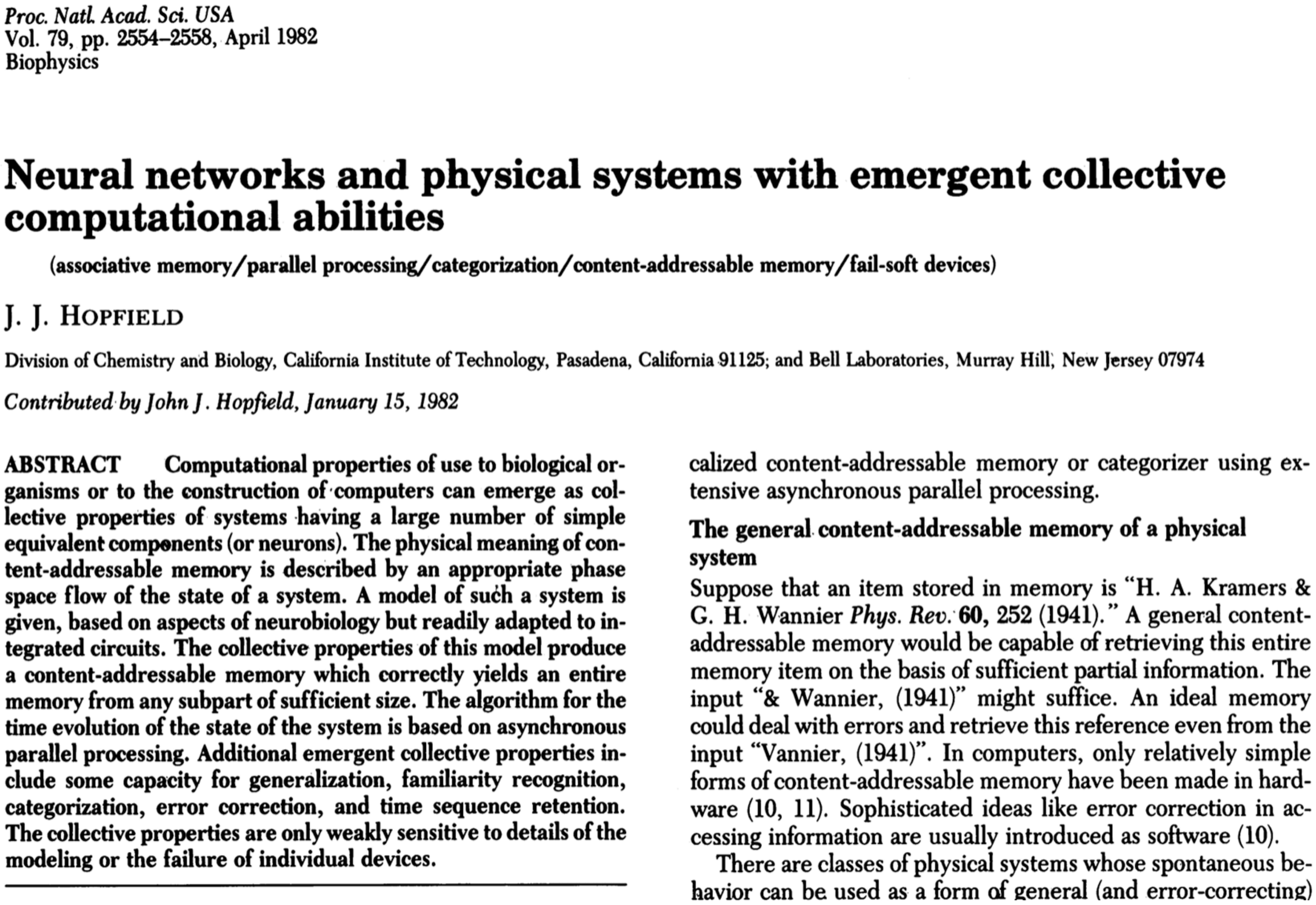
C.D. Gelatt

M.P. Vecchi



The Hopfield Model

1982



Proc. Natl. Acad. Sci. USA
Vol. 79, pp. 2554–2558, April 1982
Biophysics

Neural networks and physical systems with emergent collective computational abilities

(associative memory/parallel processing/categorization/content-addressable memory/fail-soft devices)

J. J. Hopfield

Division of Chemistry and Biology, California Institute of Technology, Pasadena, California 91125; and Bell Laboratories, Murray Hill, New Jersey 07974

Contributed by John J. Hopfield, January 15, 1982

ABSTRACT Computational properties of use to biological organisms or to the construction of computers can emerge as collective properties of systems having a large number of simple equivalent components (or neurons). The physical meaning of content-addressable memory is described by an appropriate phase space flow of the state of a system. A model of such a system is given, based on aspects of neurobiology but readily adapted to integrated circuits. The collective properties of this model produce a content-addressable memory which correctly yields an entire memory from any subpart of sufficient size. The algorithm for the time evolution of the state of the system is based on asynchronous parallel processing. Additional emergent collective properties include some capacity for generalization, familiarity recognition, categorization, error correction, and time sequence retention. The collective properties are only weakly sensitive to details of the modeling or the failure of individual devices.

calized content-addressable memory or categorizer using extensive asynchronous parallel processing.

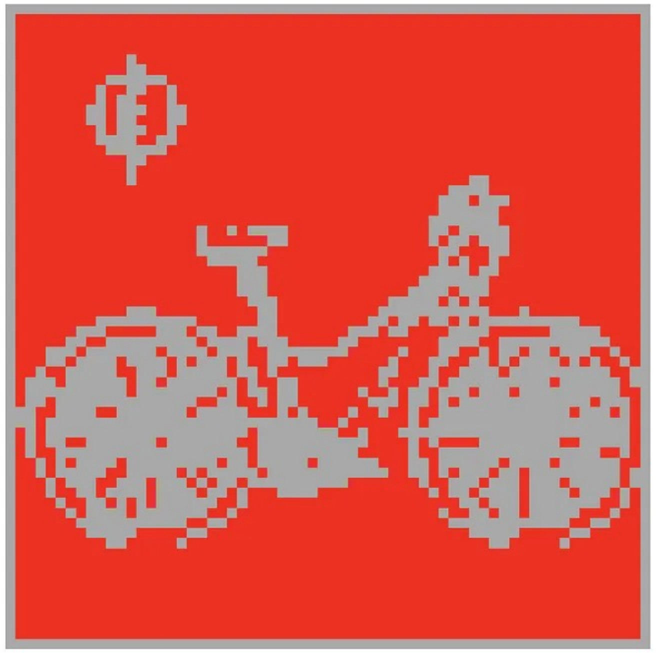
The general content-addressable memory of a physical system

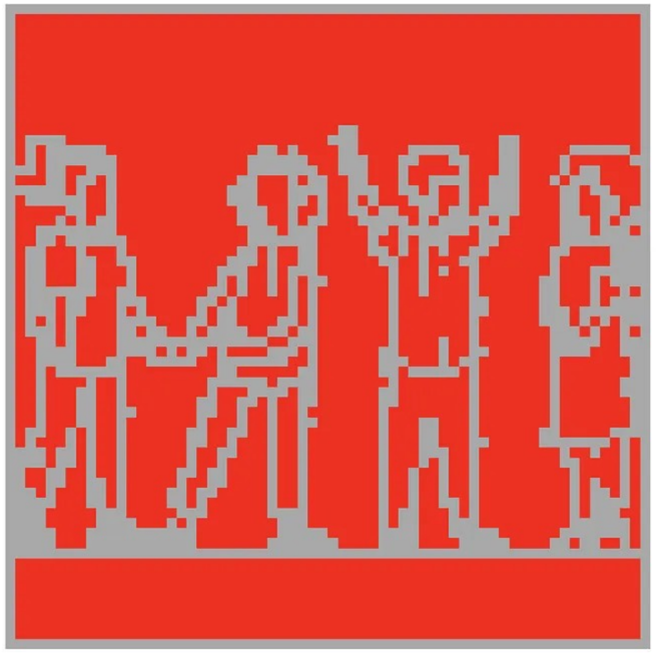
Suppose that an item stored in memory is "H. A. Kramers & G. H. Wannier *Phys. Rev.* **60,** 252 (1941)." A general content-addressable memory would be capable of retrieving this entire memory item on the basis of sufficient partial information. The input "& Wannier, (1941)" might suffice. An ideal memory could deal with errors and retrieve this reference even from the input "Vannier, (1941)". In computers, only relatively simple forms of content-addressable memory have been made in hardware (10, 11). Sophisticated ideas like error correction in accessing information are usually introduced as software (10).

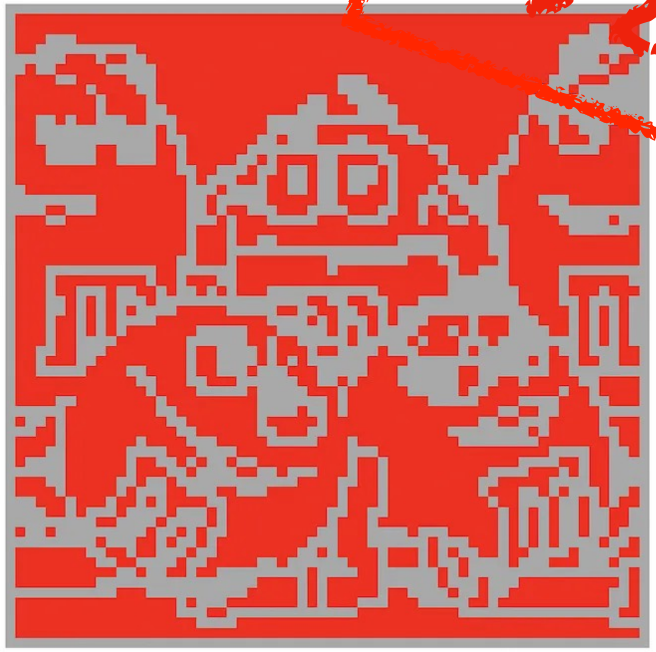
There are classes of physical systems whose spontaneous behavior can be used as a form of general (and error-correcting)

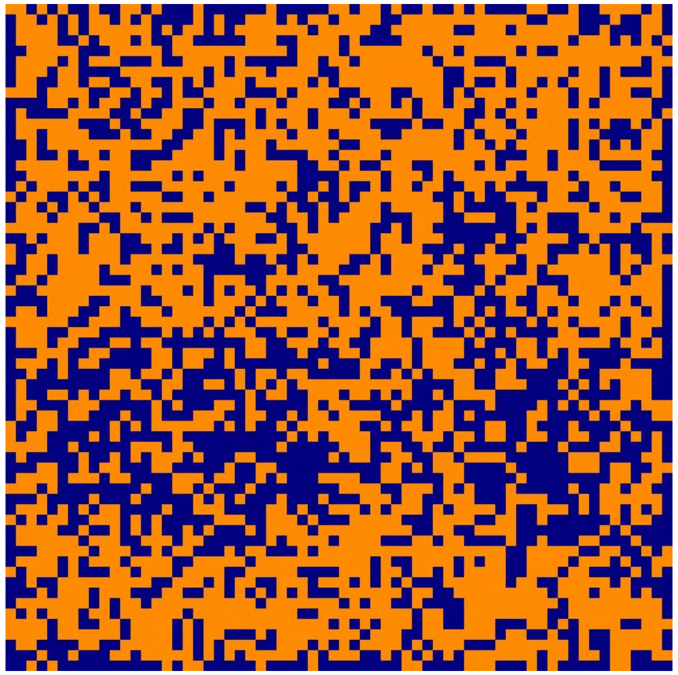
The Hopfield Model

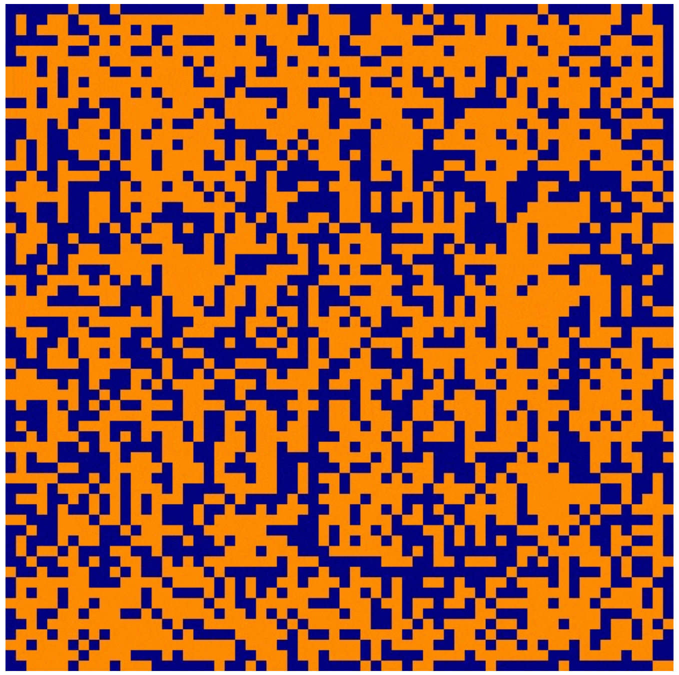
1982

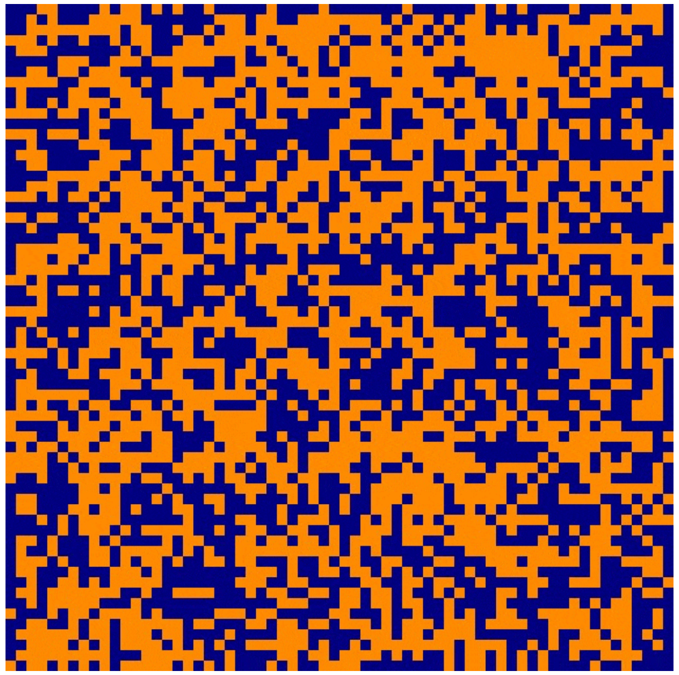






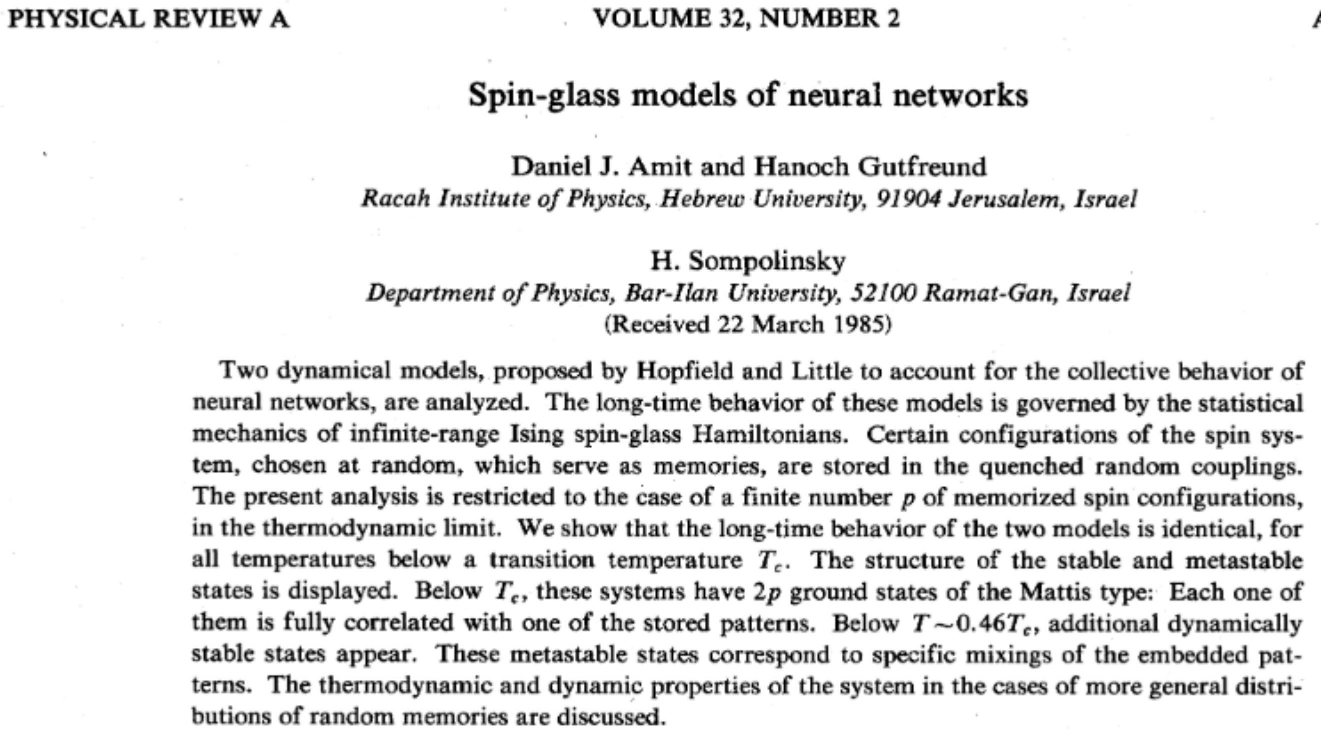






The Hopfield Model

1985



PHYSICAL REVIEW A VOLUME 32, NUMBER 2 AUGUST 1985

Spin-glass models of neural networks

Daniel J. Amit and Hanoch Gutfreund
Racah Institute of Physics, Hebrew University, 91904 Jerusalem, Israel

H. Sompolinsky
Department of Physics, Bar-Ilan University, 52100 Ramat-Gan, Israel

(Received 22 March 1985)

Two dynamical models, proposed by Hopfield and Little to account for the collective behavior of neural networks, are analyzed. The long-time behavior of these models is governed by the statistical mechanics of infinite-range Ising spin-glass Hamiltonians. Certain configurations of the spin system, chosen at random, which serve as memories, are stored in the quenched random couplings. The present analysis is restricted to the case of a finite number p of memorized spin configurations, in the thermodynamic limit. We show that the long-time behavior of the two models is identical, for all temperatures below a transition temperature T_c. The structure of the stable and metastable states is displayed. Below T_c, these systems have $2p$ ground states of the Mattis type: Each one of them is fully correlated with one of the stored patterns. Below $T \sim 0.46T_c$, additional dynamically stable states appear. These metastable states correspond to specific mixings of the embedded patterns. The thermodynamic and dynamic properties of the system in the cases of more general distributions of random memories are discussed.

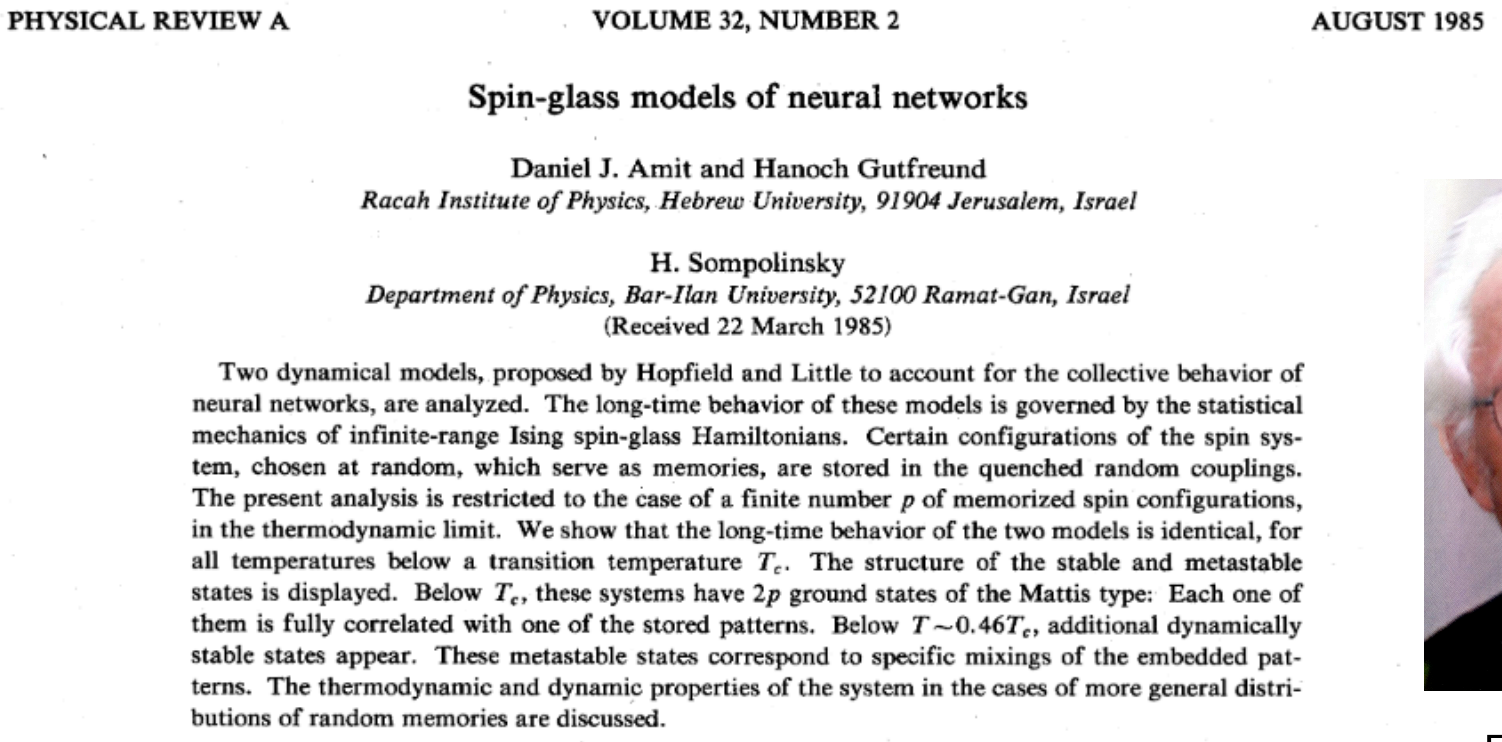
D. Amit

H. Gutfreund

H. Sompolinsky

The Hopfield Model

1985



PHYSICAL REVIEW A VOLUME 32, NUMBER 2 AUGUST 1985

Spin-glass models of neural networks

Daniel J. Amit and Hanoch Gutfreund
Racah Institute of Physics, Hebrew University, 91904 Jerusalem, Israel

H. Sompolinsky
Department of Physics, Bar-Ilan University, 52100 Ramat-Gan, Israel

(Received 22 March 1985)

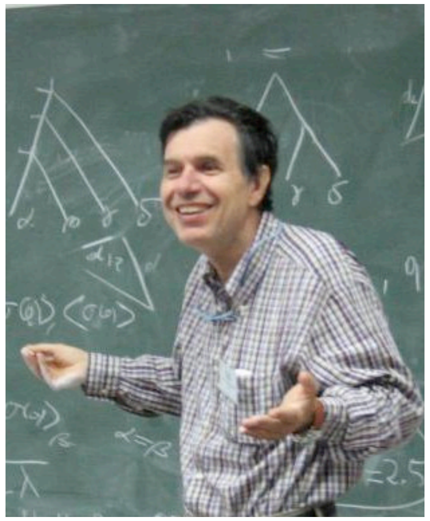
Two dynamical models, proposed by Hopfield and Little to account for the collective behavior of neural networks, are analyzed. The long-time behavior of these models is governed by the statistical mechanics of infinite-range Ising spin-glass Hamiltonians. Certain configurations of the spin system, chosen at random, which serve as memories, are stored in the quenched random couplings. The present analysis is restricted to the case of a finite number p of memorized spin configurations, in the thermodynamic limit. We show that the long-time behavior of the two models is identical, for all temperatures below a transition temperature T_c. The structure of the stable and metastable states is displayed. Below T_c, these systems have $2p$ ground states of the Mattis type: Each one of them is fully correlated with one of the stored patterns. Below $T \sim 0.46 T_c$, additional dynamically stable states appear. These metastable states correspond to specific mixings of the embedded patterns. The thermodynamic and dynamic properties of the system in the cases of more general distributions of random memories are discussed.

D. Amit

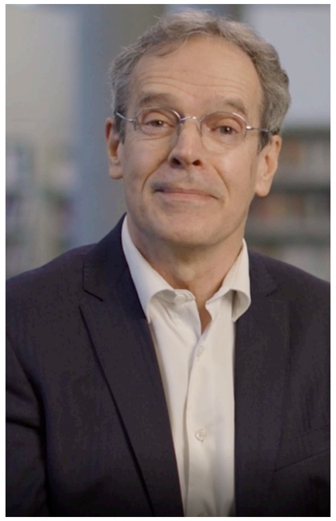
H. Gutfreund

H. Sompolinsky

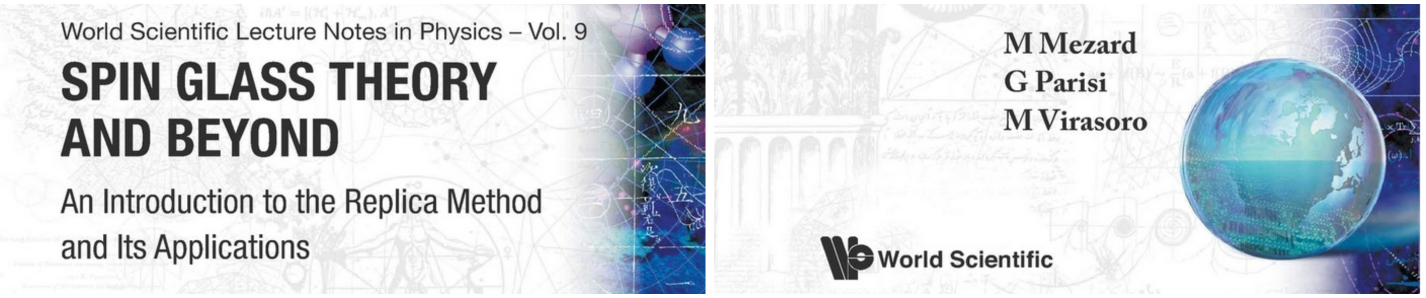
G. Parisi



M. Mézard



M. Virasoro



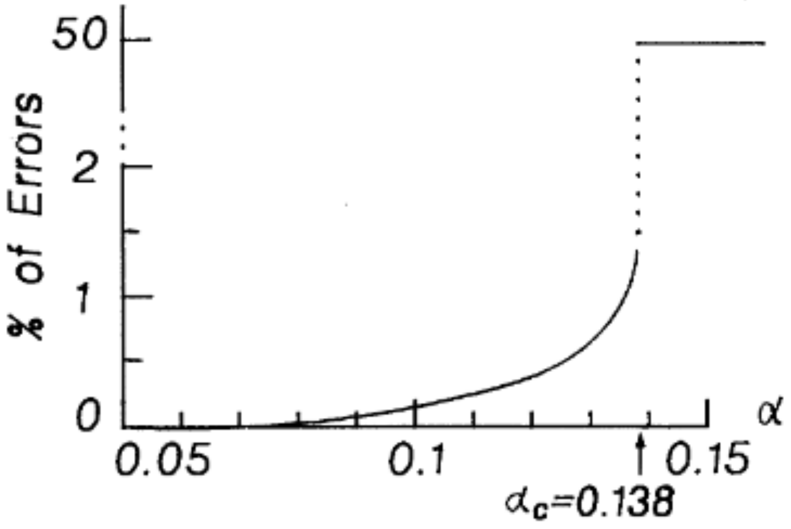
World Scientific Lecture Notes in Physics – Vol. 9

SPIN GLASS THEORY AND BEYOND

An Introduction to the Replica Method and Its Applications

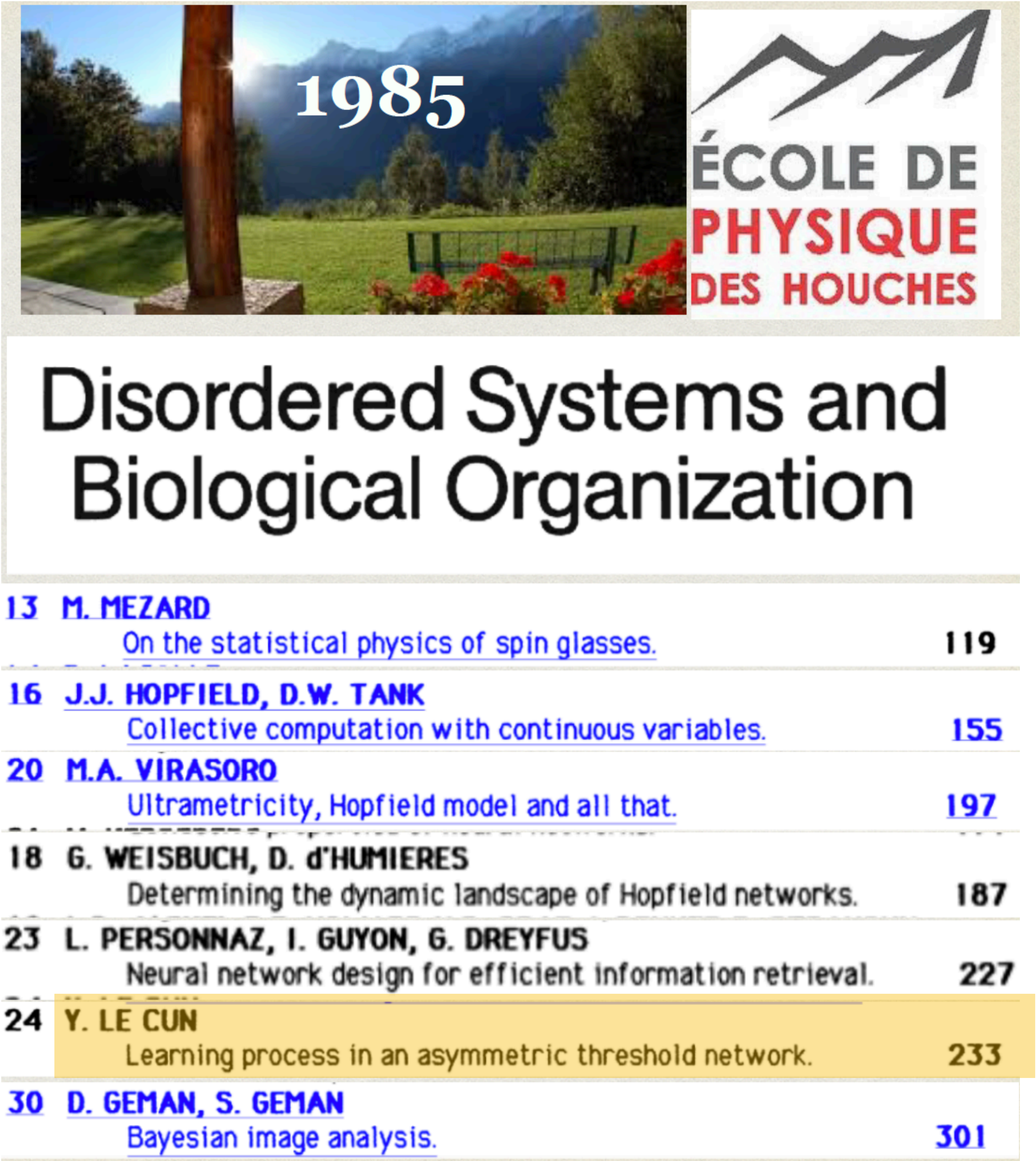
M Mezard
G Parisi
M Virasoro

World Scientific



And they were not alone...

1985



1985

ÉCOLE DE PHYSIQUE DES HOUCHES

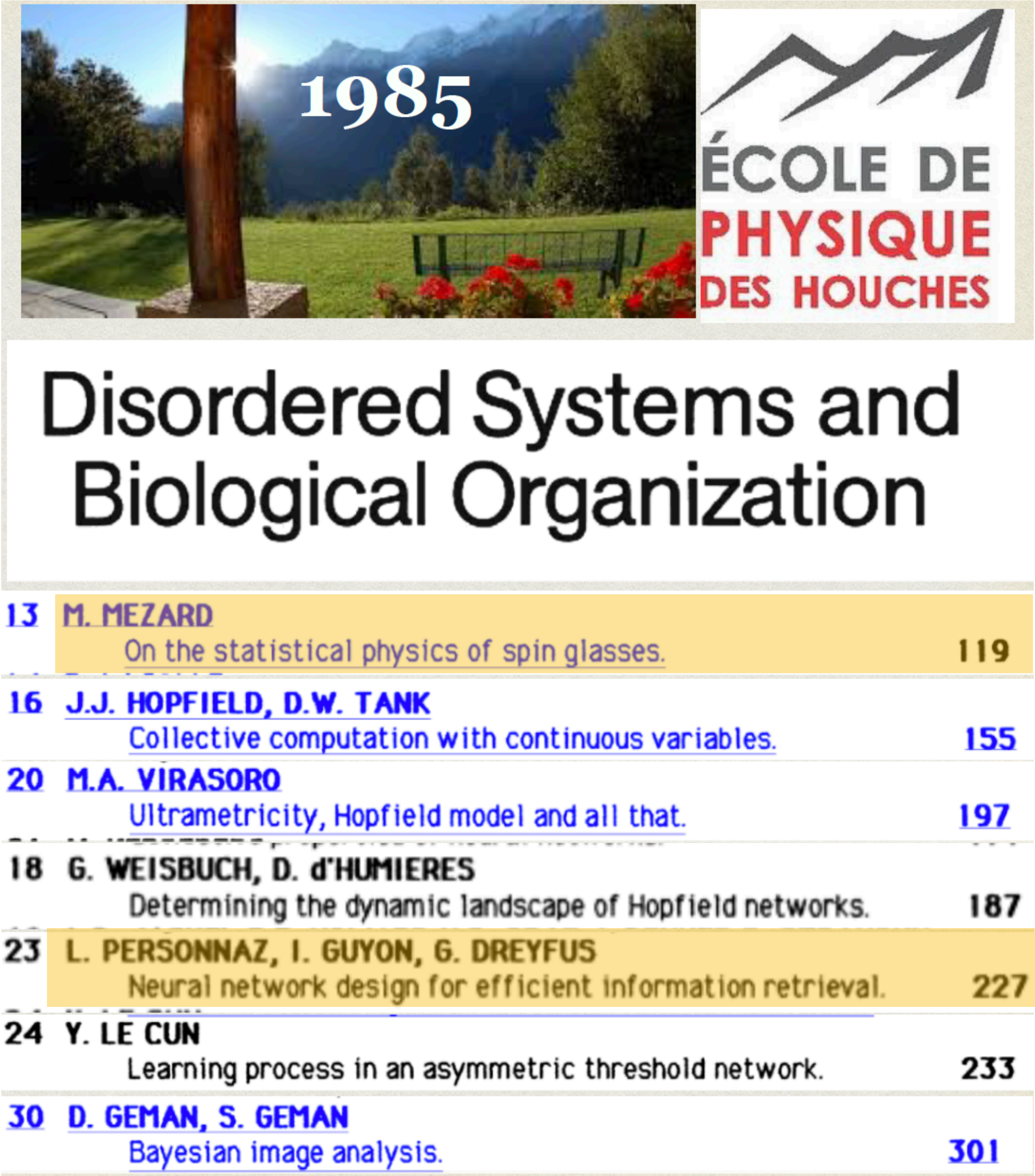
Disordered Systems and Biological Organization

"Only physicists were interested in neural networks at the time [...] My professional life truly shifted in February 1985 during a physics symposium in Les Houches, in the French Alps. There, I met the crème de la crème of international research interested in neural networks and gave my very first talk (in English!)."

From *"Quand la Machine Apprend"*

And they were not alone...

1985



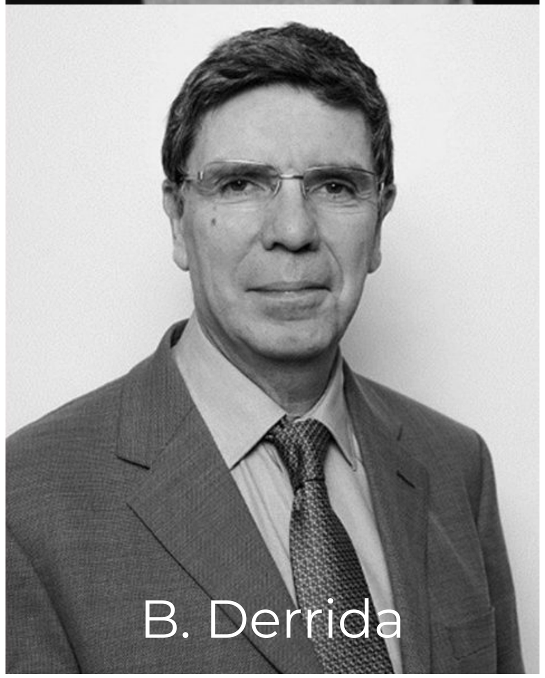
Disordered Systems and Biological Organization

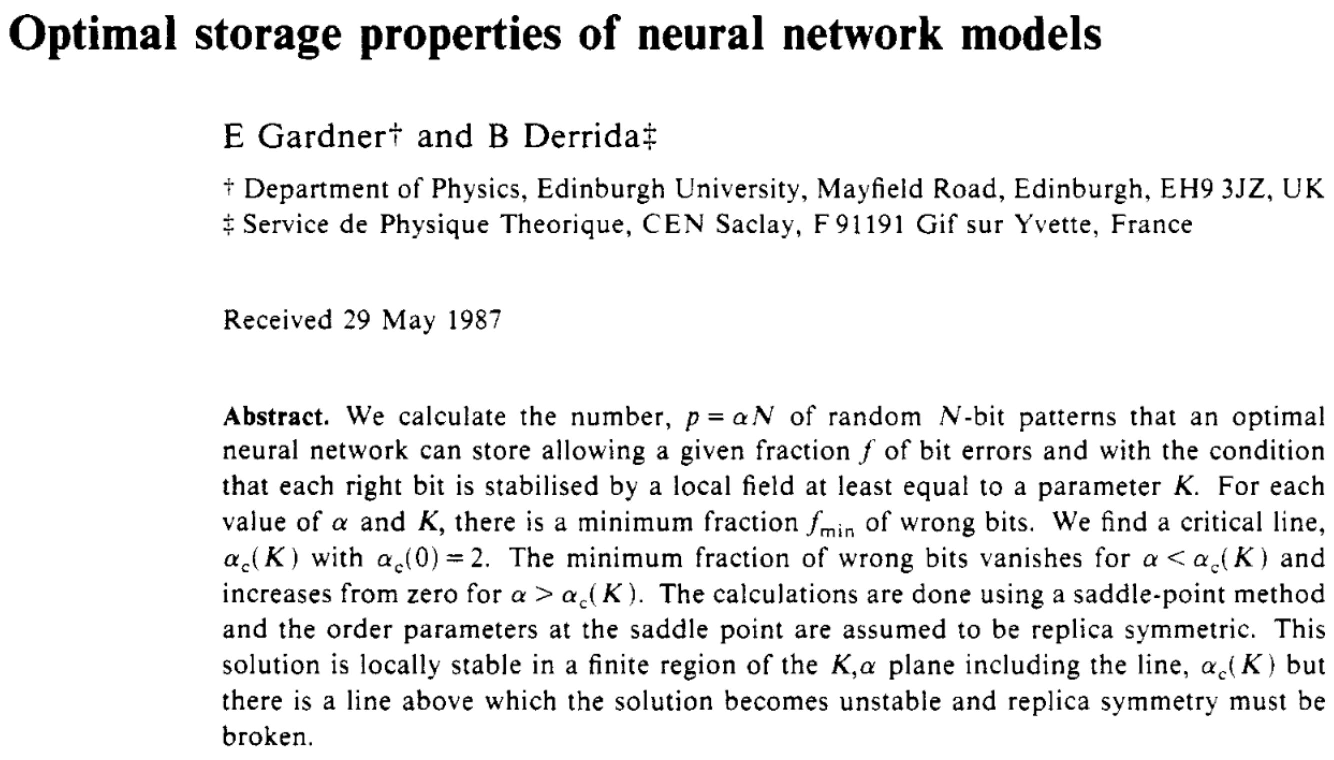
I benchmarked neural networks against kernel methods with my Ph.D advisors Gerard Dreyfus and Leon Personnaz. The same year, two physicists working close-by (Marc Mezard & Werner Krauth) published a paper on an optimal margin algorithm called 'minover,' which attracted my attention.... but it was not until I joined Bell Labs that I put things together and we created support vector machines.

From *"Data Mining History: The Invention of Support Vector Machines"*

The Perceptron

1987





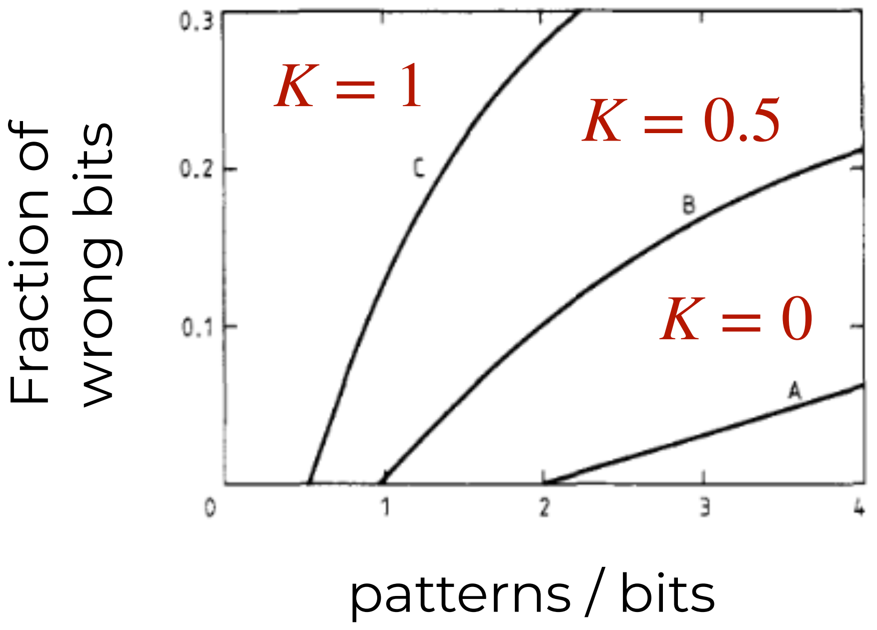
Optimal storage properties of neural network models

E Gardner† and B Derrida‡

† Department of Physics, Edinburgh University, Mayfield Road, Edinburgh, EH9 3JZ, UK
‡ Service de Physique Theorique, CEN Saclay, F 91191 Gif sur Yvette, France

Received 29 May 1987

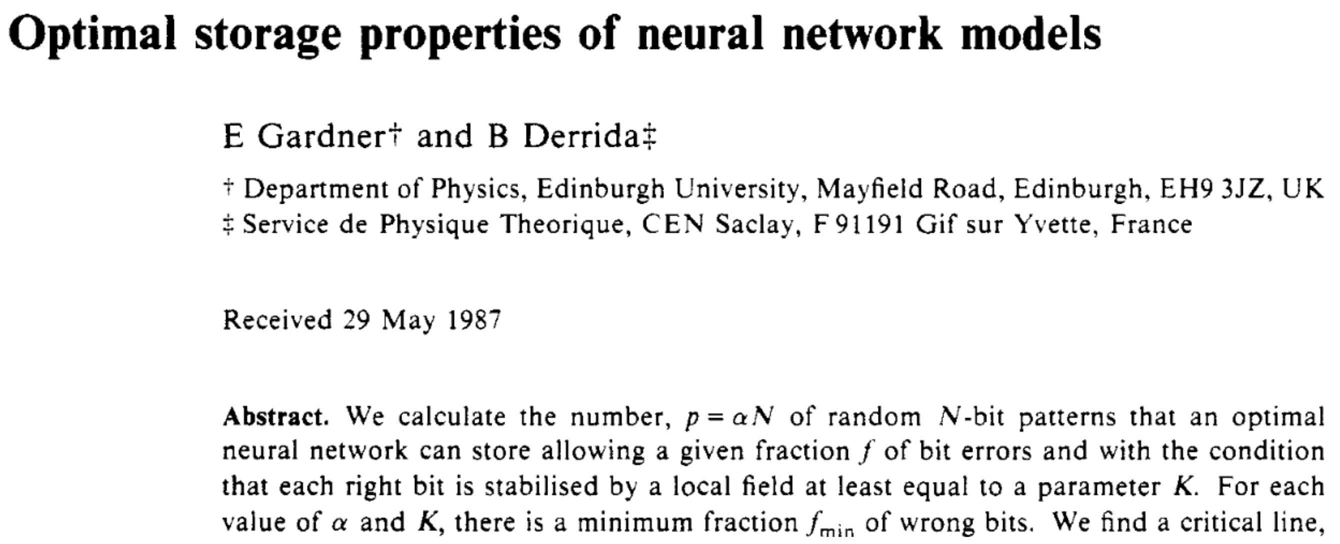
Abstract. We calculate the number, $p = \alpha N$ of random N-bit patterns that an optimal neural network can store allowing a given fraction f of bit errors and with the condition that each right bit is stabilised by a local field at least equal to a parameter K. For each value of α and K, there is a minimum fraction $f_{\min}$ of wrong bits. We find a critical line, $\alpha_c(K)$ with $\alpha_c(0) = 2$. The minimum fraction of wrong bits vanishes for $\alpha < \alpha_c(K)$ and increases from zero for $\alpha > \alpha_c(K)$. The calculations are done using a saddle-point method and the order parameters at the saddle point are assumed to be replica symmetric. This solution is locally stable in a finite region of the K,α plane including the line, $\alpha_c(K)$ but there is a line above which the solution becomes unstable and replica symmetry must be broken.



c.f. [Cover 1967]

The Perceptron

1987



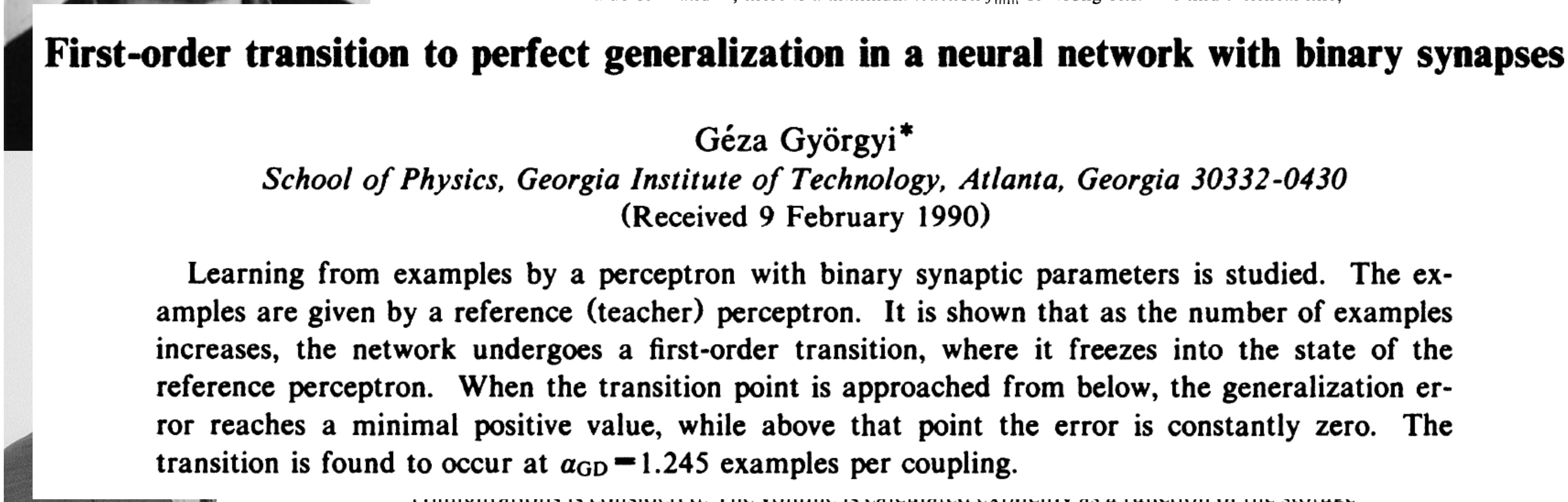
Optimal storage properties of neural network models

E Gardner† and B Derrida‡

† Department of Physics, Edinburgh University, Mayfield Road, Edinburgh, EH9 3JZ, UK
‡ Service de Physique Theorique, CEN Saclay, F 91191 Gif sur Yvette, France

Received 29 May 1987

Abstract. We calculate the number, $p = \alpha N$ of random N-bit patterns that an optimal neural network can store allowing a given fraction f of bit errors and with the condition that each right bit is stabilised by a local field at least equal to a parameter K. For each value of α and K, there is a minimum fraction $f_{\min}$ of wrong bits. We find a critical line,



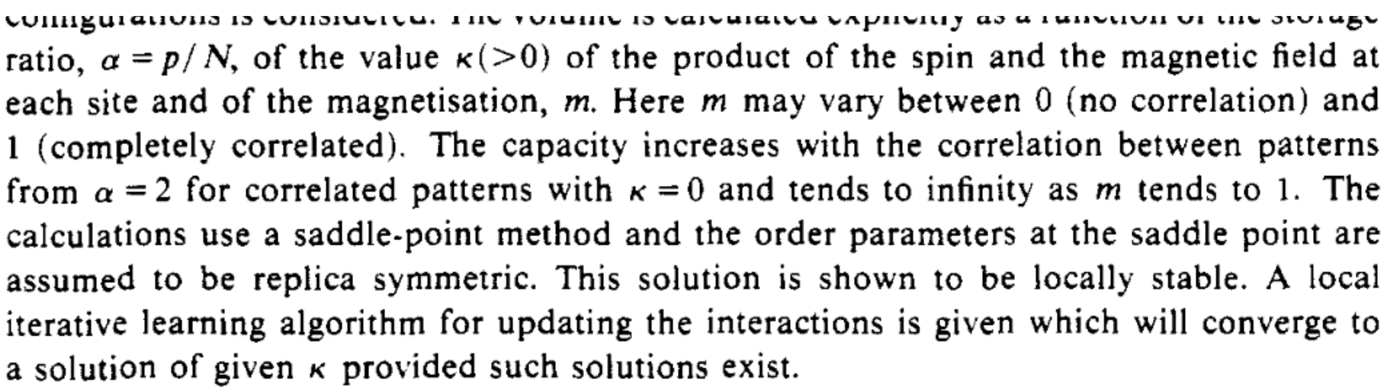
First-order transition to perfect generalization in a neural network with binary synapses

Géza Györgyi*

School of Physics, Georgia Institute of Technology, Atlanta, Georgia 30332-0430

(Received 9 February 1990)

Learning from examples by a perceptron with binary synaptic parameters is studied. The examples are given by a reference (teacher) perceptron. It is shown that as the number of examples increases, the network undergoes a first-order transition, where it freezes into the state of the reference perceptron. When the transition point is approached from below, the generalization error reaches a minimal positive value, while above that point the error is constantly zero. The transition is found to occur at $\alpha_{GD} = 1.245$ examples per coupling.

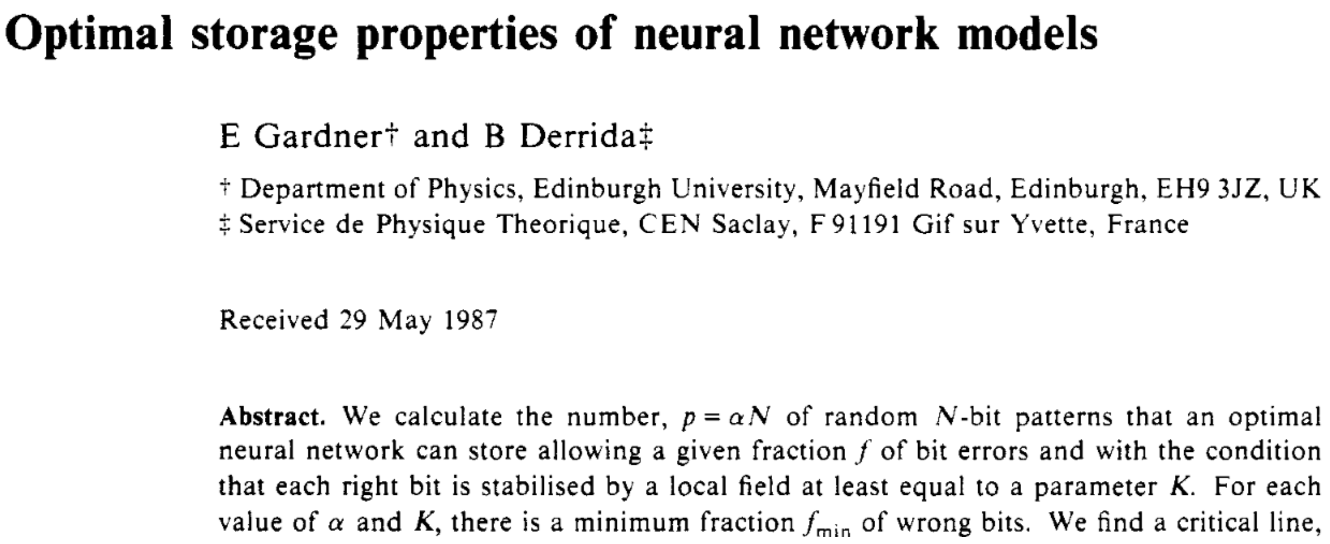


configurations is considered. The volume is calculated explicitly as a function of the storage ratio, $\alpha = p/N$, of the value $\kappa(>0)$ of the product of the spin and the magnetic field at each site and of the magnetisation, m. Here m may vary between 0 (no correlation) and 1 (completely correlated). The capacity increases with the correlation between patterns from $\alpha = 2$ for correlated patterns with $\kappa = 0$ and tends to infinity as m tends to 1. The calculations use a saddle-point method and the order parameters at the saddle point are assumed to be replica symmetric. This solution is shown to be locally stable. A local iterative learning algorithm for updating the interactions is given which will converge to a solution of given κ provided such solutions exist.

c.f. [Cover 1967]

The Perceptron

1987



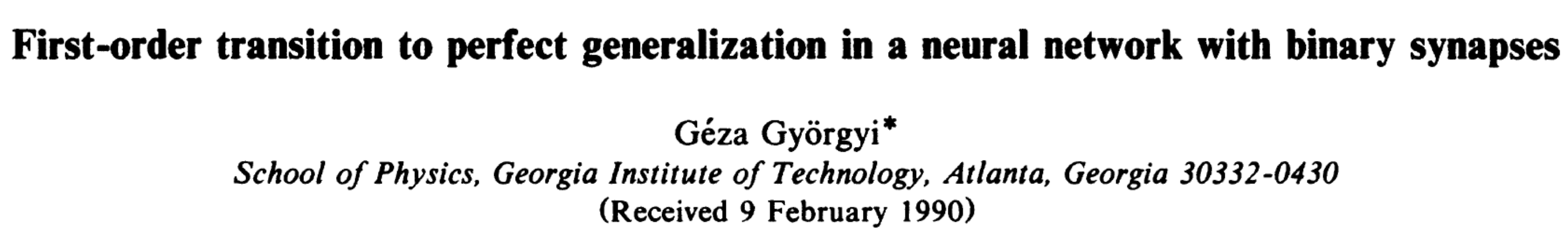
Optimal storage properties of neural network models

E Gardner† and B Derrida‡

† Department of Physics, Edinburgh University, Mayfield Road, Edinburgh, EH9 3JZ, UK
‡ Service de Physique Theorique, CEN Saclay, F 91191 Gif sur Yvette, France

Received 29 May 1987

Abstract. We calculate the number, $p = \alpha N$ of random N-bit patterns that an optimal neural network can store allowing a given fraction f of bit errors and with the condition that each right bit is stabilised by a local field at least equal to a parameter K. For each value of α and K, there is a minimum fraction $f_{\min}$ of wrong bits. We find a critical line,

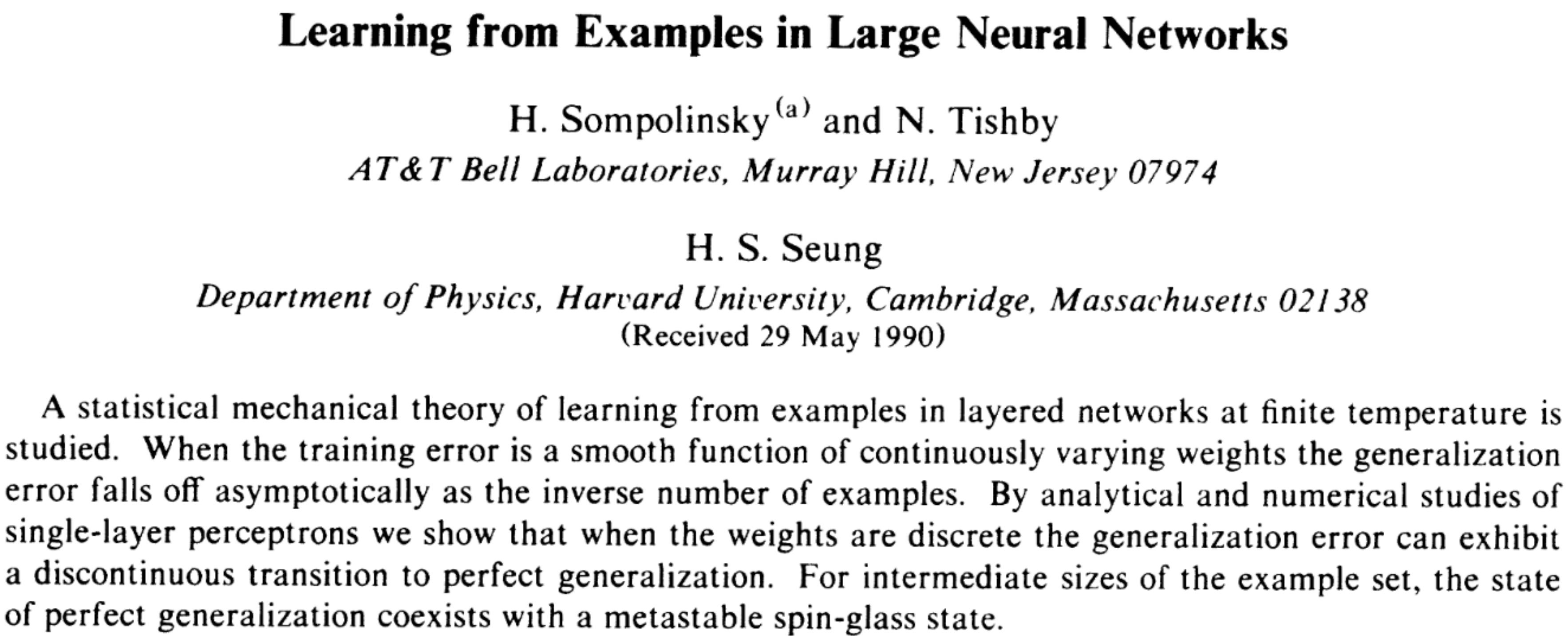


First-order transition to perfect generalization in a neural network with binary synapses

Géza Györgyi*
School of Physics, Georgia Institute of Technology, Atlanta, Georgia 30332-0430

(Received 9 February 1990)

Learning fr
amples are giv
increases, the
reference perc
ror reaches a
transition is fc



Learning from Examples in Large Neural Networks

H. Sompolinsky[(a)] and N. Tishby
AT&T Bell Laboratories, Murray Hill, New Jersey 07974

H. S. Seung
Department of Physics, Harvard University, Cambridge, Massachusetts 02138

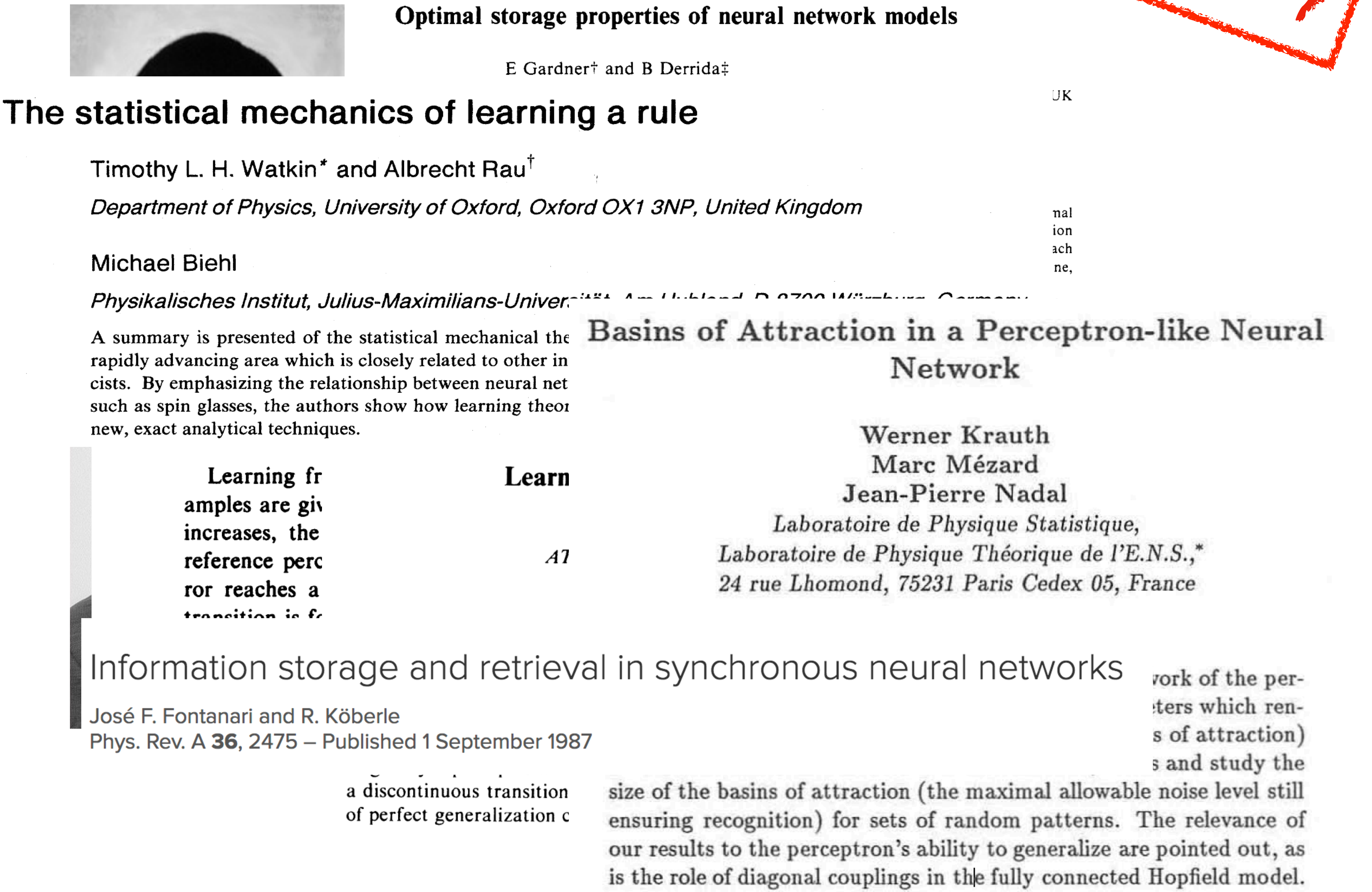
(Received 29 May 1990)

A statistical mechanical theory of learning from examples in layered networks at finite temperature is studied. When the training error is a smooth function of continuously varying weights the generalization error falls off asymptotically as the inverse number of examples. By analytical and numerical studies of single-layer perceptrons we show that when the weights are discrete the generalization error can exhibit a discontinuous transition to perfect generalization. For intermediate sizes of the example set, the state of perfect generalization coexists with a metastable spin-glass state.

B. Derr

The Perceptron

1987



Optimal storage properties of neural network models

E Gardner† and B Derrida‡

UK

nal
ion
ach
ne,

The statistical mechanics of learning a rule

Timothy L. H. Watkin* and Albrecht Rau†

Department of Physics, University of Oxford, Oxford OX1 3NP, United Kingdom

Michael Biehl

Physikalisches Institut, Julius-Maximilians-Univer

A summary is presented of the statistical mechanical the
rapidly advancing area which is closely related to other in
cists. By emphasizing the relationship between neural net
such as spin glasses, the authors show how learning theo
new, exact analytical techniques.

Learning fr
amples are gi
increases, the
reference perc
ror reaches a

Learn

A1

a discontinuous transition
of perfect generalization c

Basins of Attraction in a Perceptron-like Neural Network

Werner Krauth
Marc Mézard
Jean-Pierre Nadal
Laboratoire de Physique Statistique,
*Laboratoire de Physique Théorique de l'E.N.S.,**
24 rue Lhomond, 75231 Paris Cedex 05, France

ork of the per-
ters which ren-
s of attraction)
s and study the
size of the basins of attraction (the maximal allowable noise level still ensuring recognition) for sets of random patterns. The relevance of our results to the perceptron's ability to generalize are pointed out, as is the role of diagonal couplings in the fully connected Hopfield model.

Information storage and retrieval in synchronous neural networks

José F. Fontanari and R. Köberle
Phys. Rev. A **36**, 2475 – Published 1 September 1987

The Perceptron

1987

Optimal storage properties of neural network models

E Gardner† and B Derrida‡

UK

UK

The statistical mechanics of learning a rule

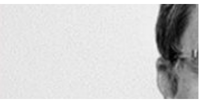
Timothy L. H. Watkin* and Albrecht Rau†

Department of Physics, University of Oxford, Oxford OX1 3NP, United Kingdom

Michael Biehl

Physikalisches Institut, Julius-Maximilians-Universität, Am Hubland, D-8700 Würzburg, Germany

A summary is presented of the statistical mechanical the
rapidly advancing area which is closely related to other in
cists. By emphasizing the relationship between neural net
such as spin glasses, the authors show how learning theo
new, exact analytical techniques.



Learning from examples by

Learning from Examples in Large

H. Sompolinsky[a] and N. T

AT&T Bell Laboratories, Murray Hill, N

Basins of Attraction in a Perceptron-like Neural Network

Werner Krauth
Marc Mézard
Jean-Pierre Nadal
Laboratoire de Physique Statistique,
*Laboratoire de Physique Théorique de l'E.N.S.,**
24 rue Lhomond, 75231 Paris Cedex 05, France

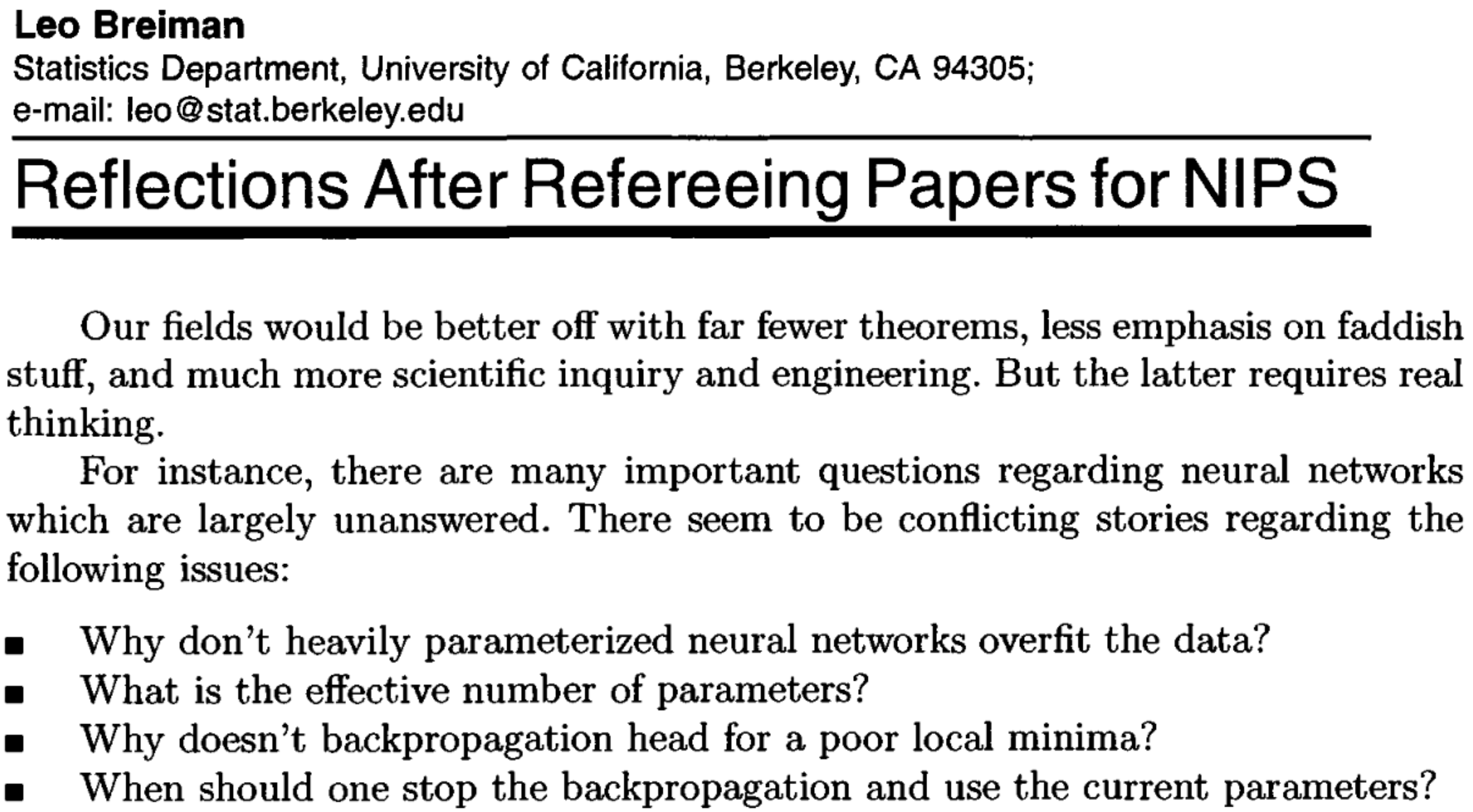
Information storage and retrieval in synchronous neural networks

José F. Fontanari and R. Köberle
Phys. Rev. A **36**, 2475 – Published 1 September 1987

A stat
studied.
error falls off asymptotically as the inverse number of examples.
single-layer perceptrons we show that when the weights are disc
a discontinuous transition to perfect generalization. For interme
of perfect generalization coexists with a metastable spin-glass sta

vork of the per-
ters which ren-
s of attraction)
s and study the
size of the basins of attraction (the maximal allowable noise level still ensuring recognition) for sets of random patterns. The relevance of our results to the perceptron's ability to generalize are pointed out, as is the role of diagonal couplings in the fully connected Hopfield model.

1995



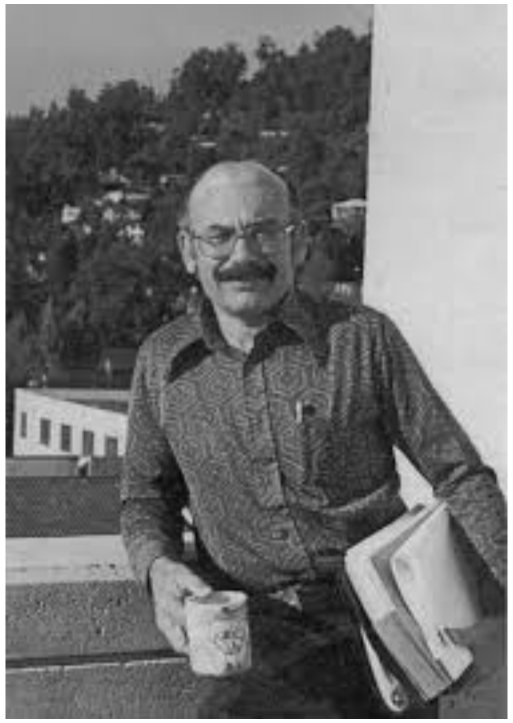
Leo Breiman
Statistics Department, University of California, Berkeley, CA 94305;
e-mail: leo@stat.berkeley.edu

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.

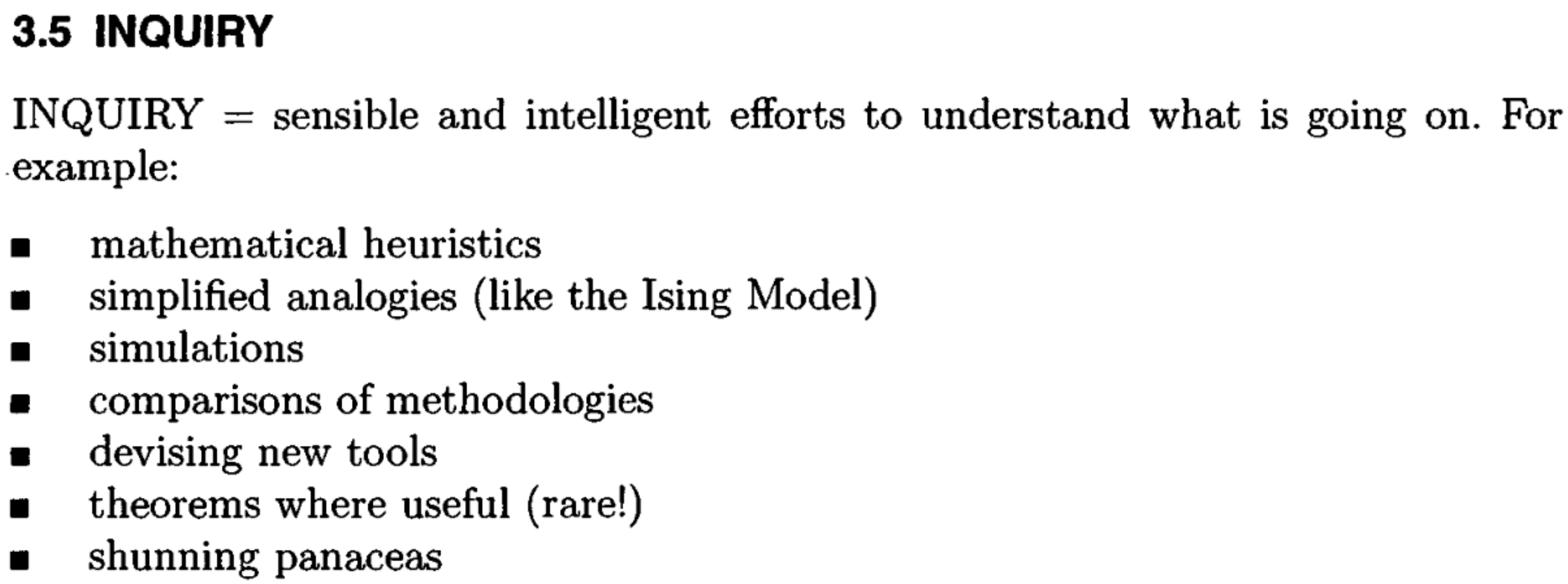
For instance, there are many important questions regarding neural networks which are largely unanswered. There seem to be conflicting stories regarding the following issues:

- 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?



Mathematical theory is not critical to the development of machine learning.

But scientific inquiry is.



3.5 INQUIRY

INQUIRY = sensible and intelligent efforts to understand what is going on. For example:

- mathematical heuristics
- simplified analogies (like the Ising Model)
- simulations
- comparisons of methodologies
- devising new tools
- theorems where useful (rare!)
- shunning panaceas

Final summary

Physics provide not only a conceptual framework to think about the challenges in ML but also a set of useful tools to solve them.

This will be subject of our next lectures.