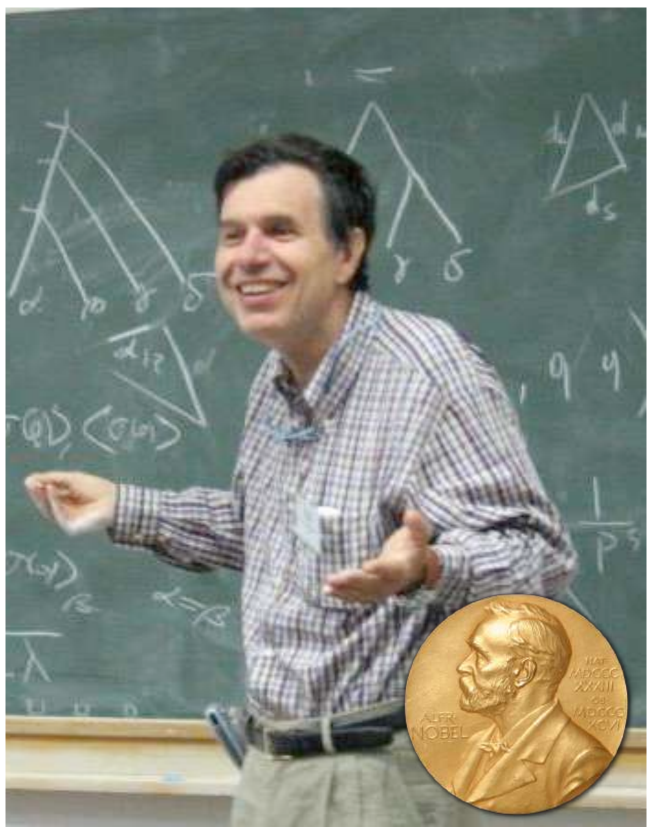
A Statistical Physics Perspective on the Theory of Machine Learning

Bruno Loureiro
Département d'Informatique
École Normale Supérieure & CNRS

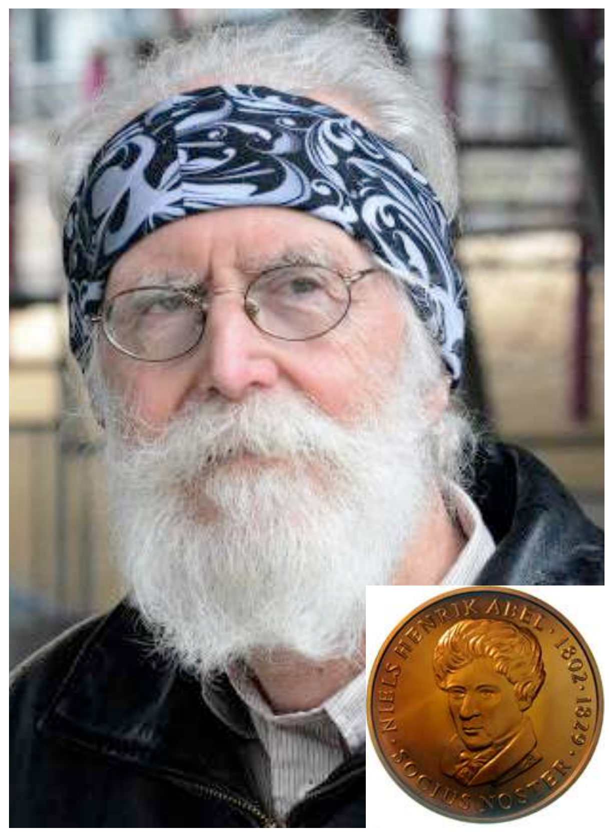
bruno.loureiro@di.ens.fr

INFORMS Applied Probability Society Summer School
GeorgiaTech, 25-29.06.2025

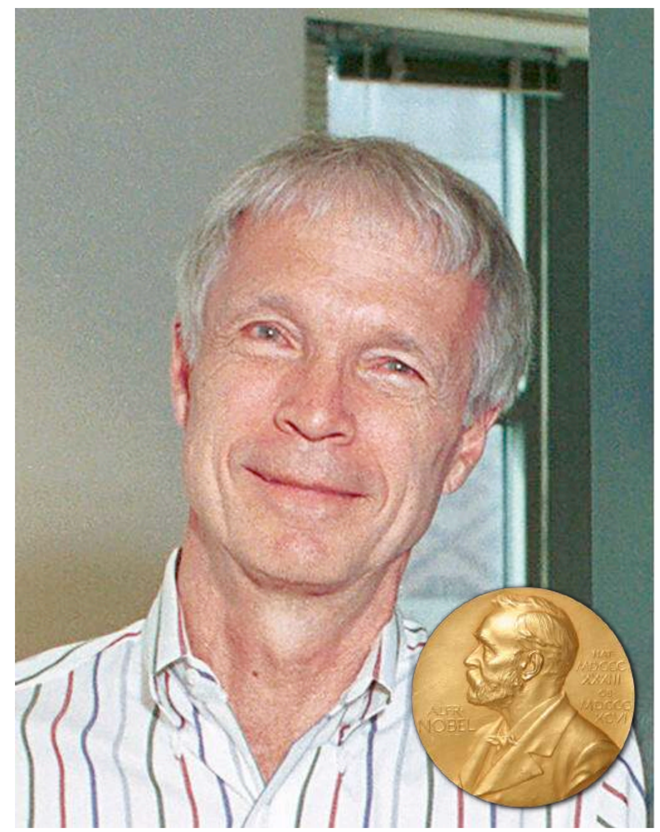
G. Parisi
(2021)



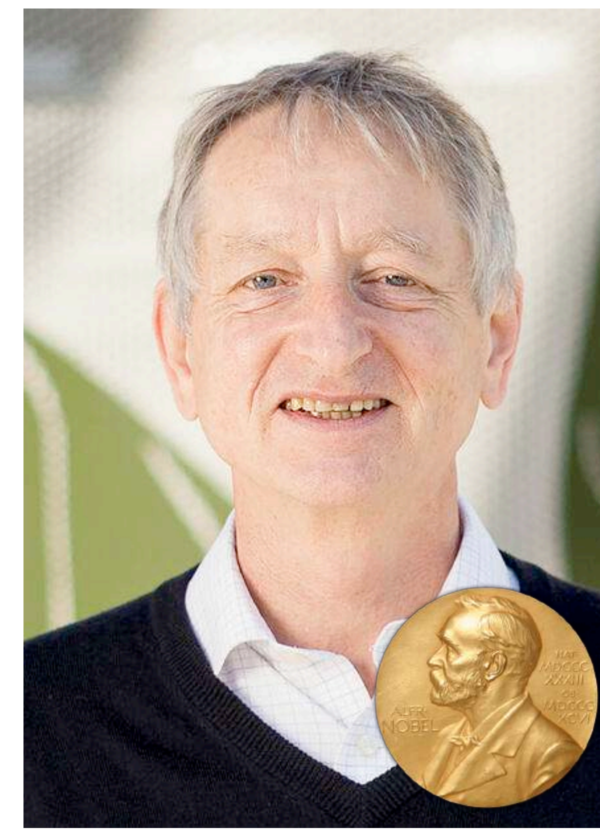
M. Talagrand
(2024)



J. Hopfield
(2024)

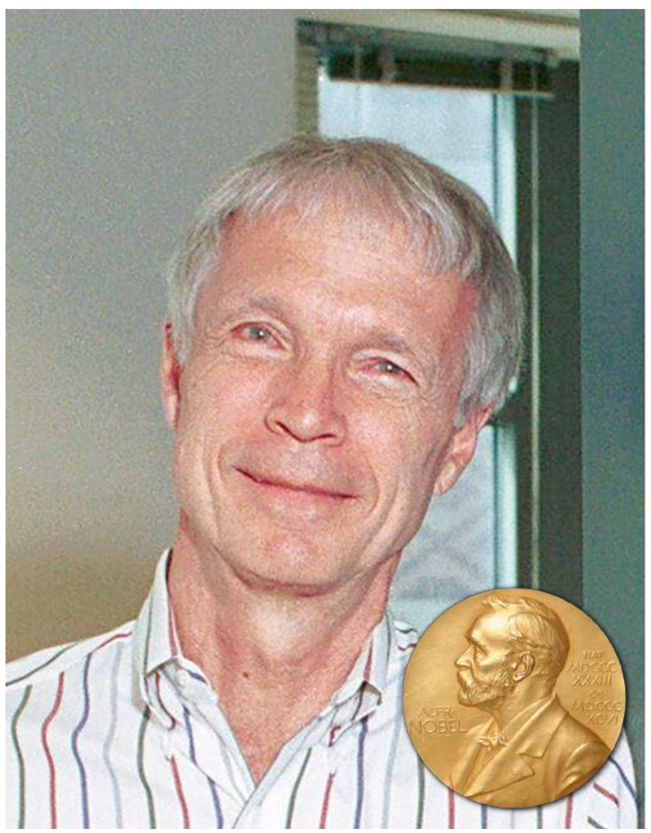


G. Hinton
(2024)



nature NEWS Q&A | 28 October 2024 By Elizabeth Gibney

J. Hopfield
(2024)

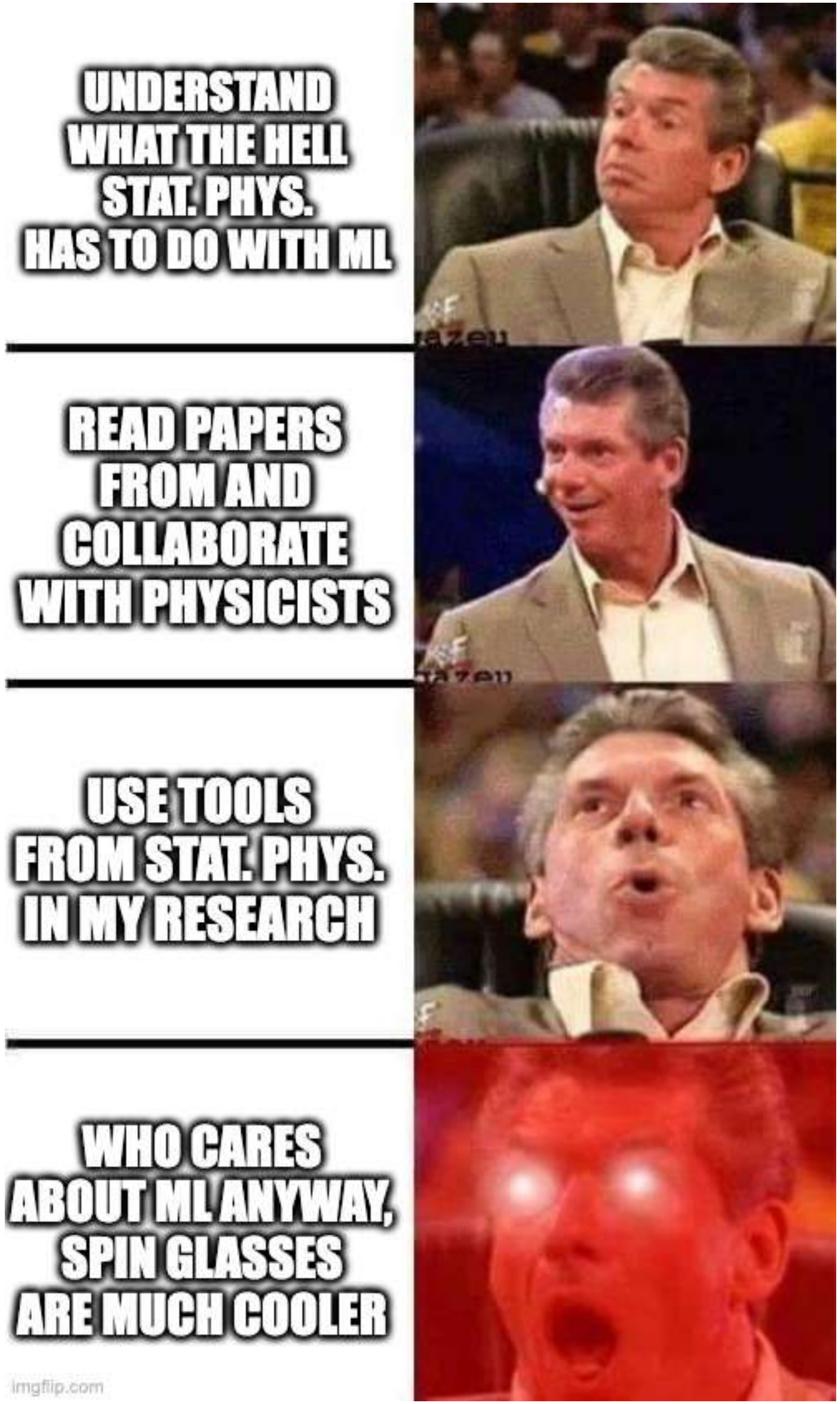


There was discussion that your prizewinning work was not really physics, but computer science. What do you think?

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.

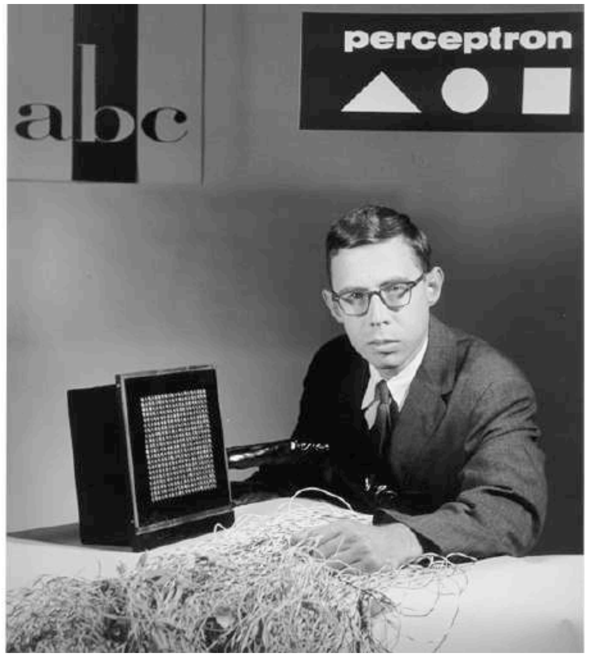
What's your advice for today's PhD students?

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.

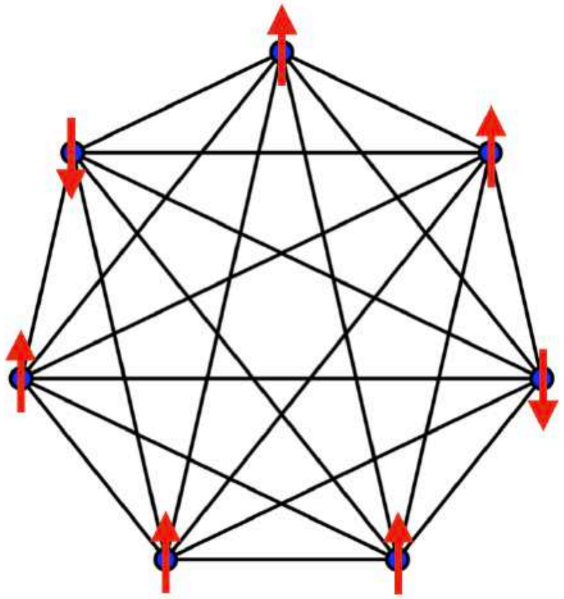

UNDERSTAND WHAT THE HELL STAT. PHYS. HAS TO DO WITH ML
READ PAPERS FROM AND COLLABORATE WITH PHYSICISTS
USE TOOLS FROM STAT. PHYS. IN MY RESEARCH
WHO CARES ABOUT ML ANYWAY, SPIN GLASSES ARE MUCH COOLER
imgflip.com

On the menu

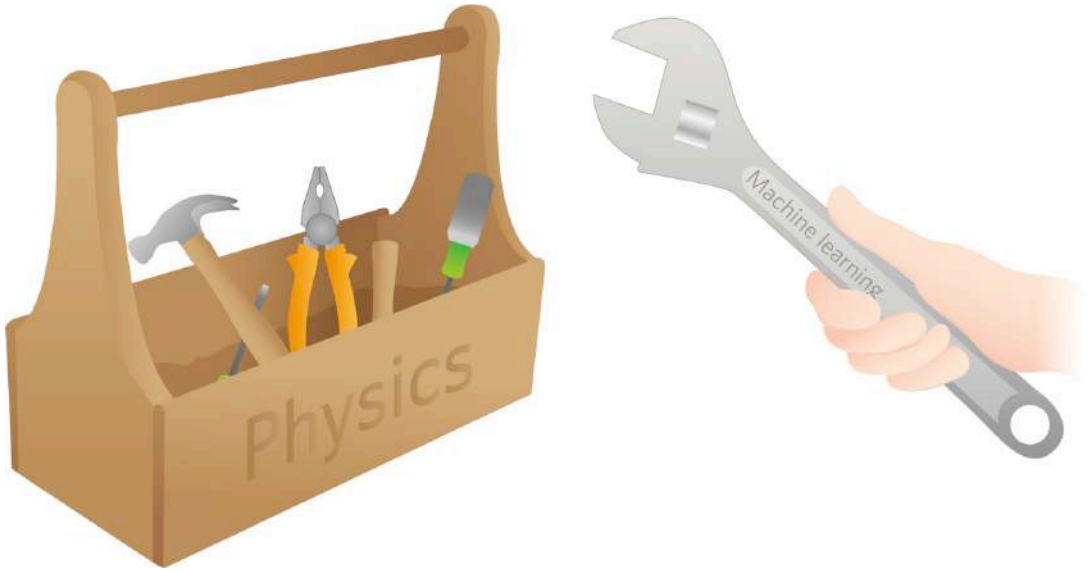
Historical context



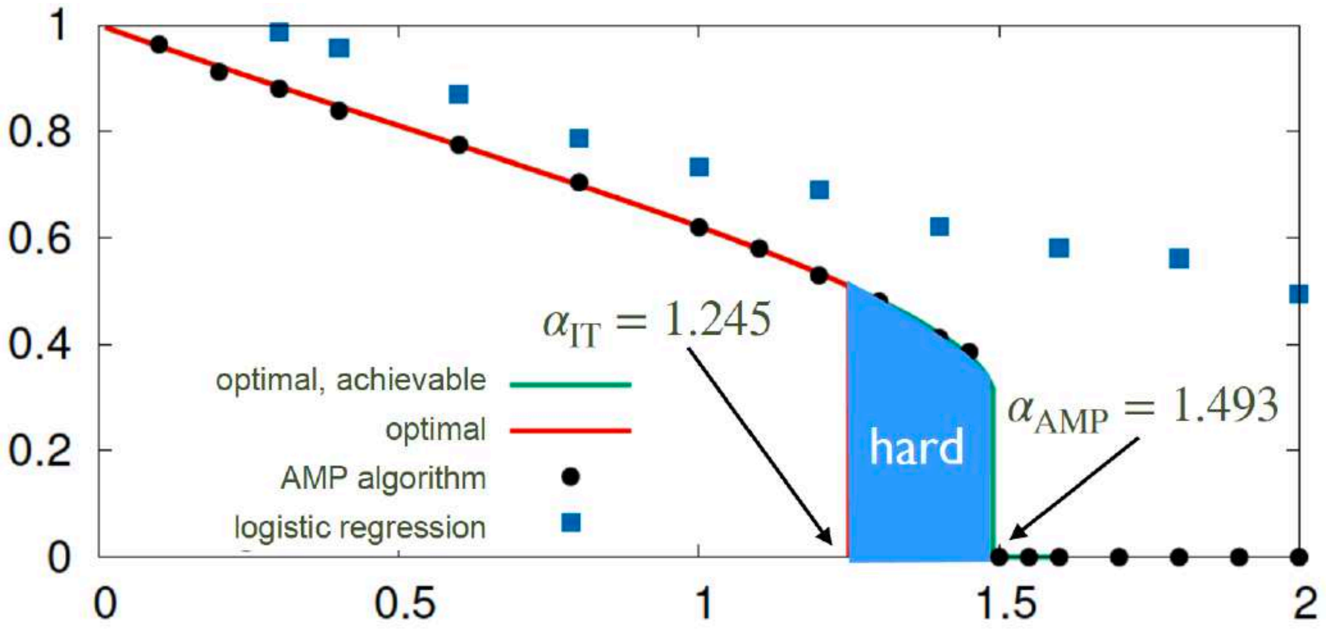
Curie-Weiss model

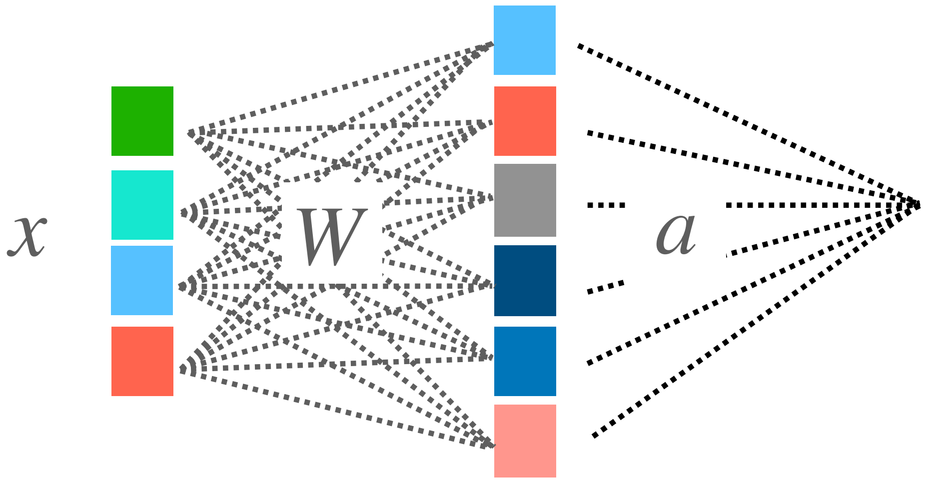


The toolbox



Lessons from simple models

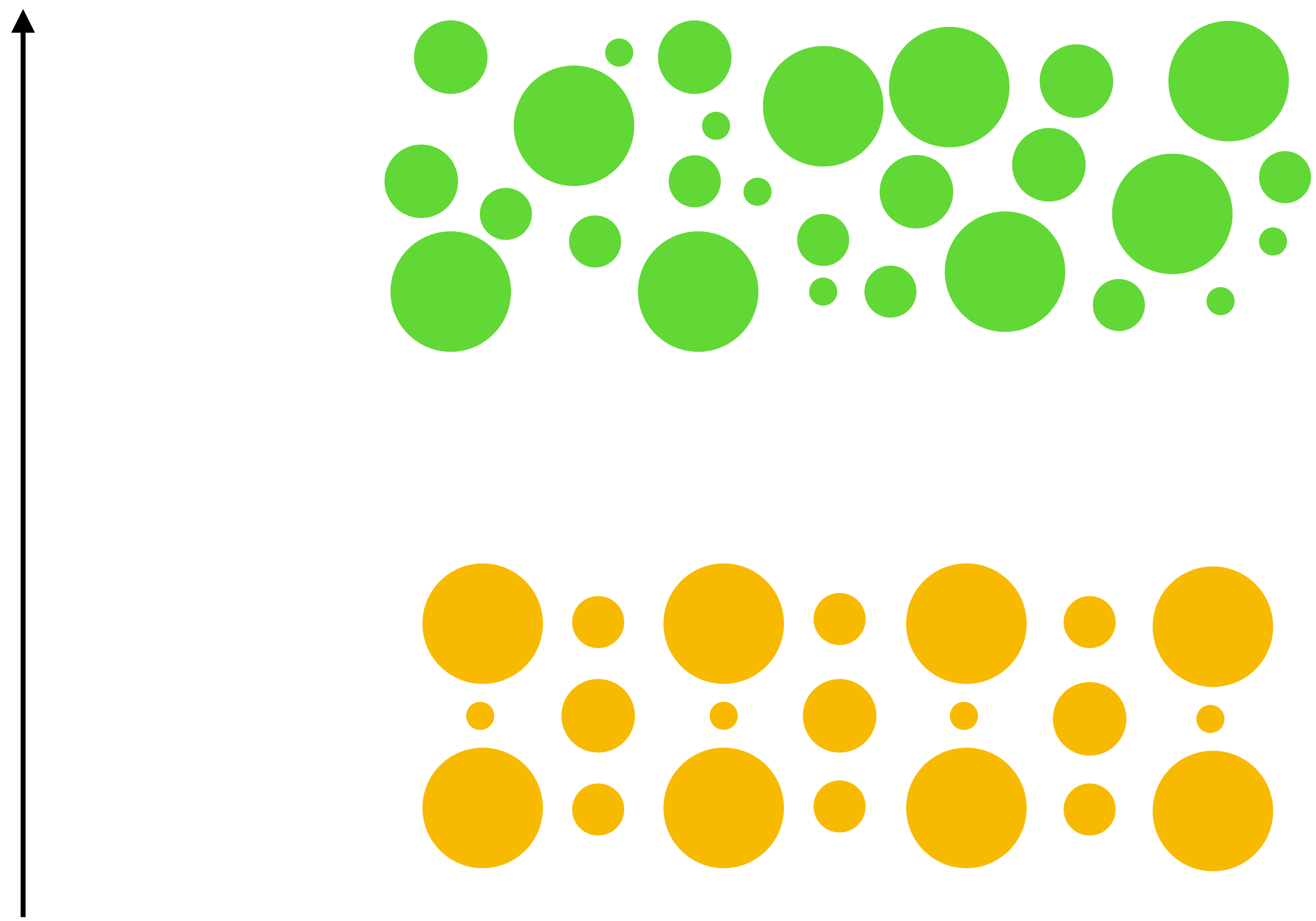




Physics of glasses

Temperature

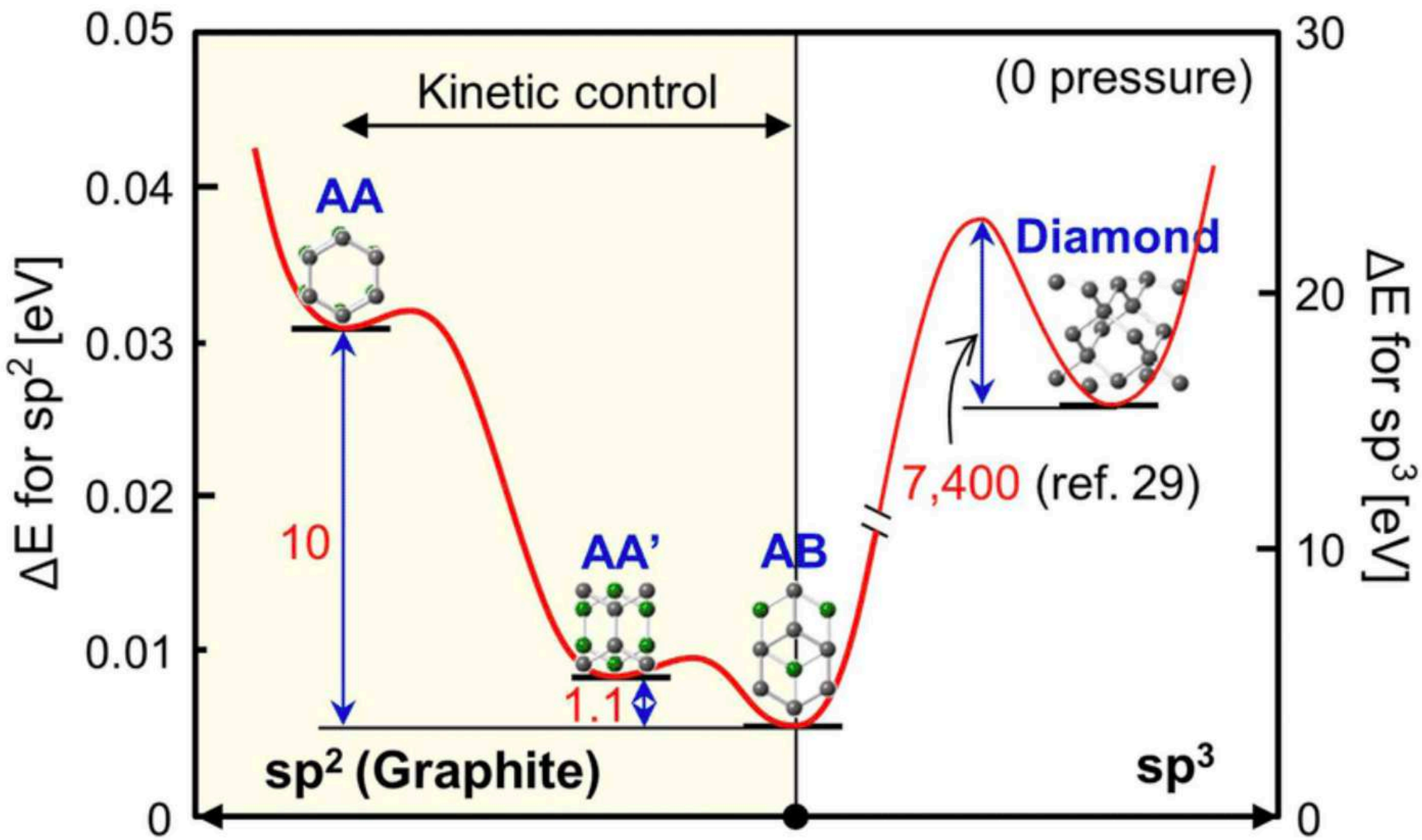
“Liquid”



“Solid”

“Melting transition”

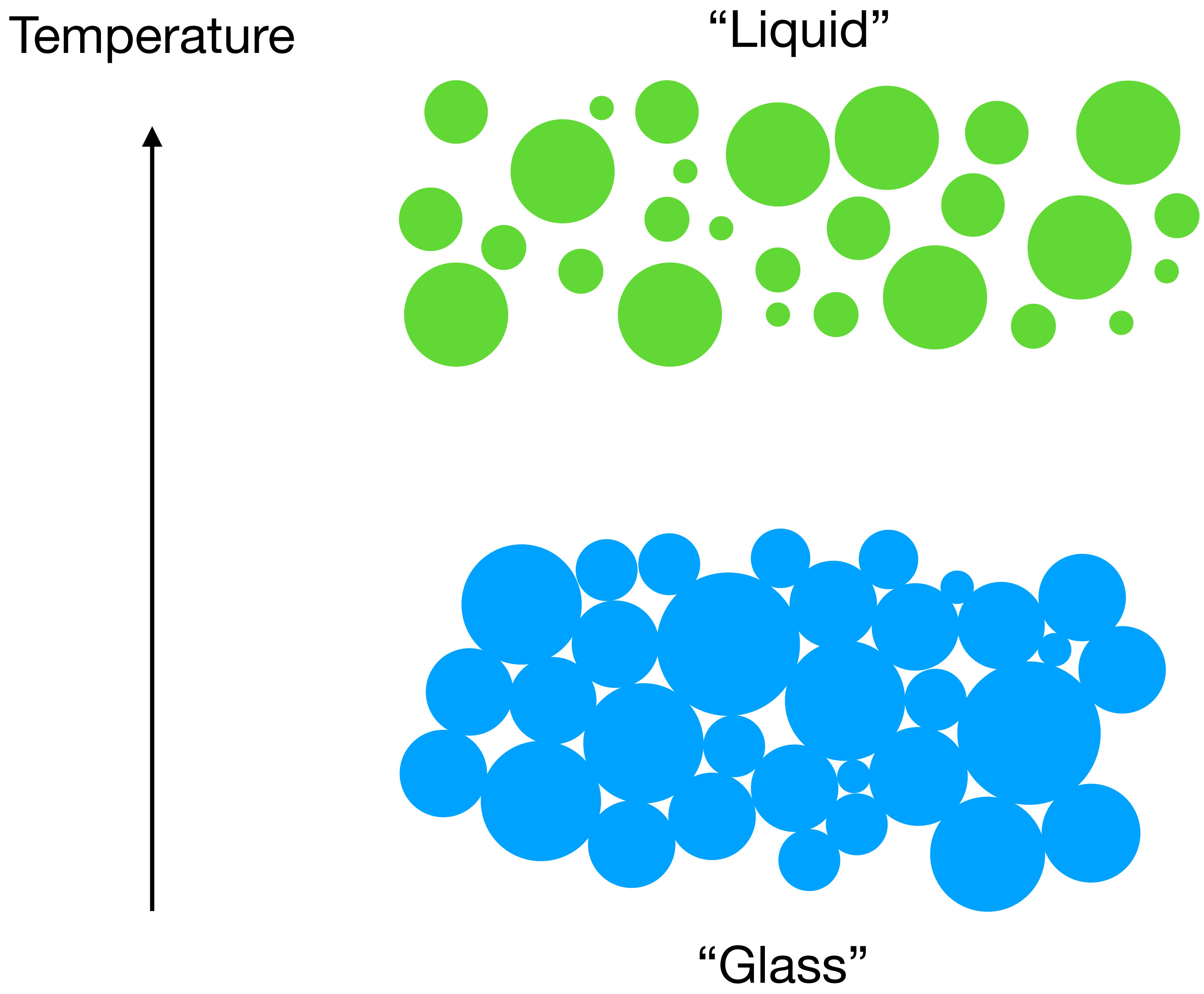
Physics of glasses



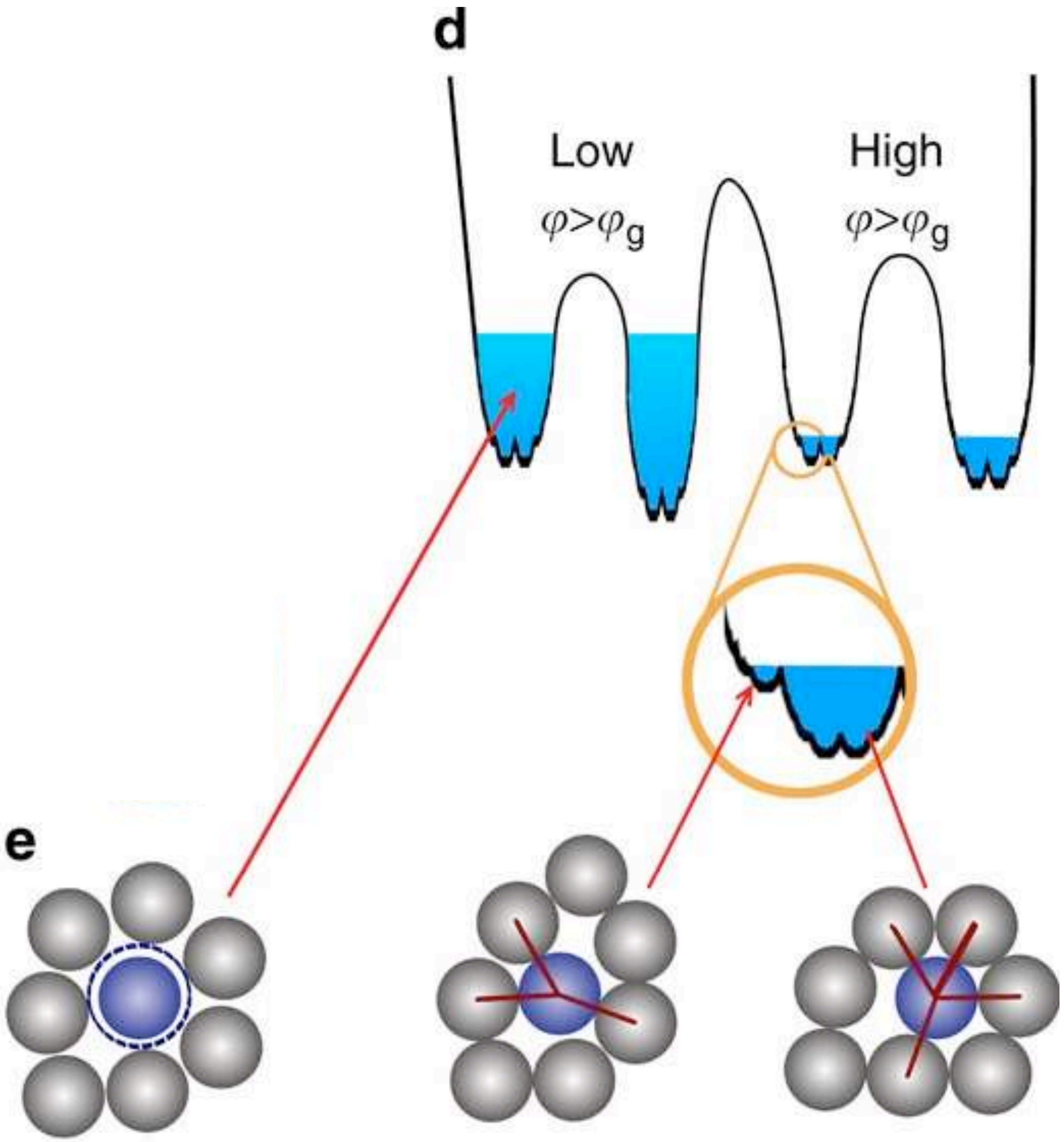
$$\tau \propto e^{\Delta E}$$

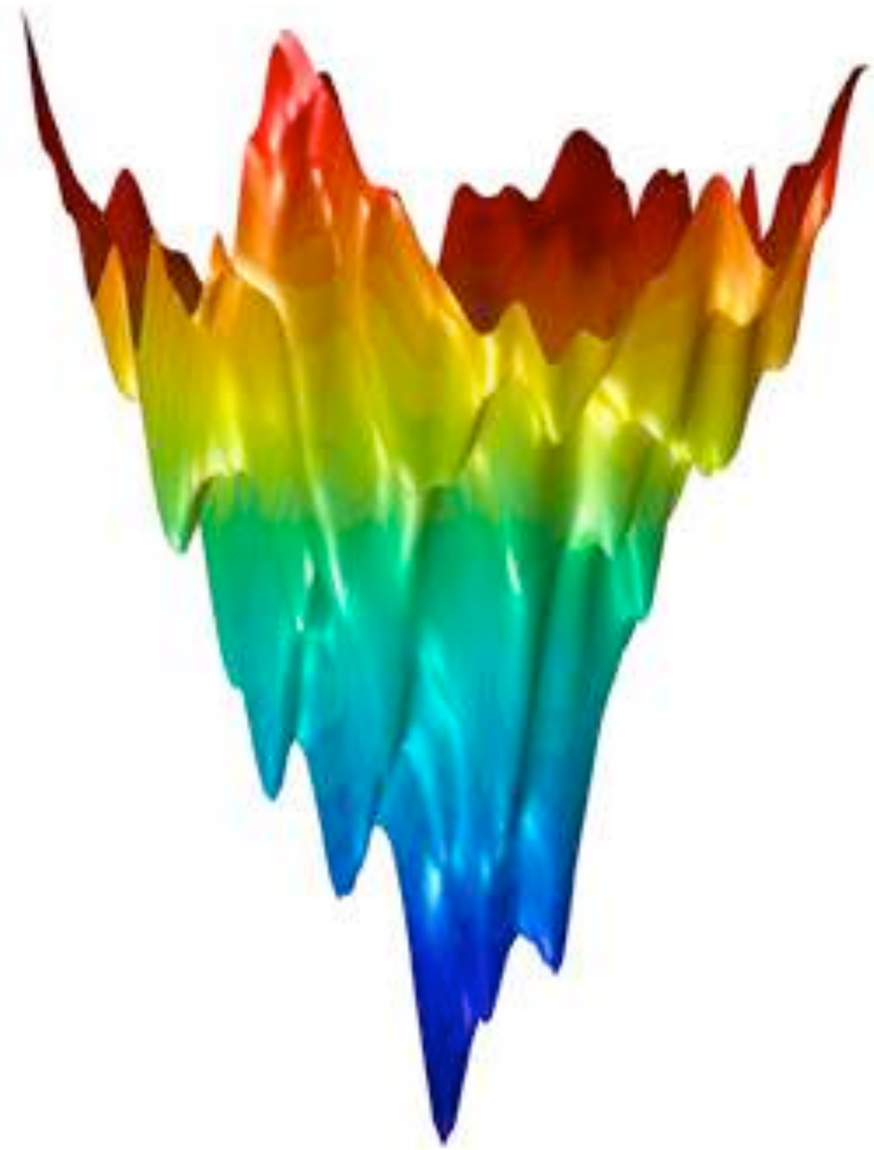
"Arrhenius law"

Physics of glasses

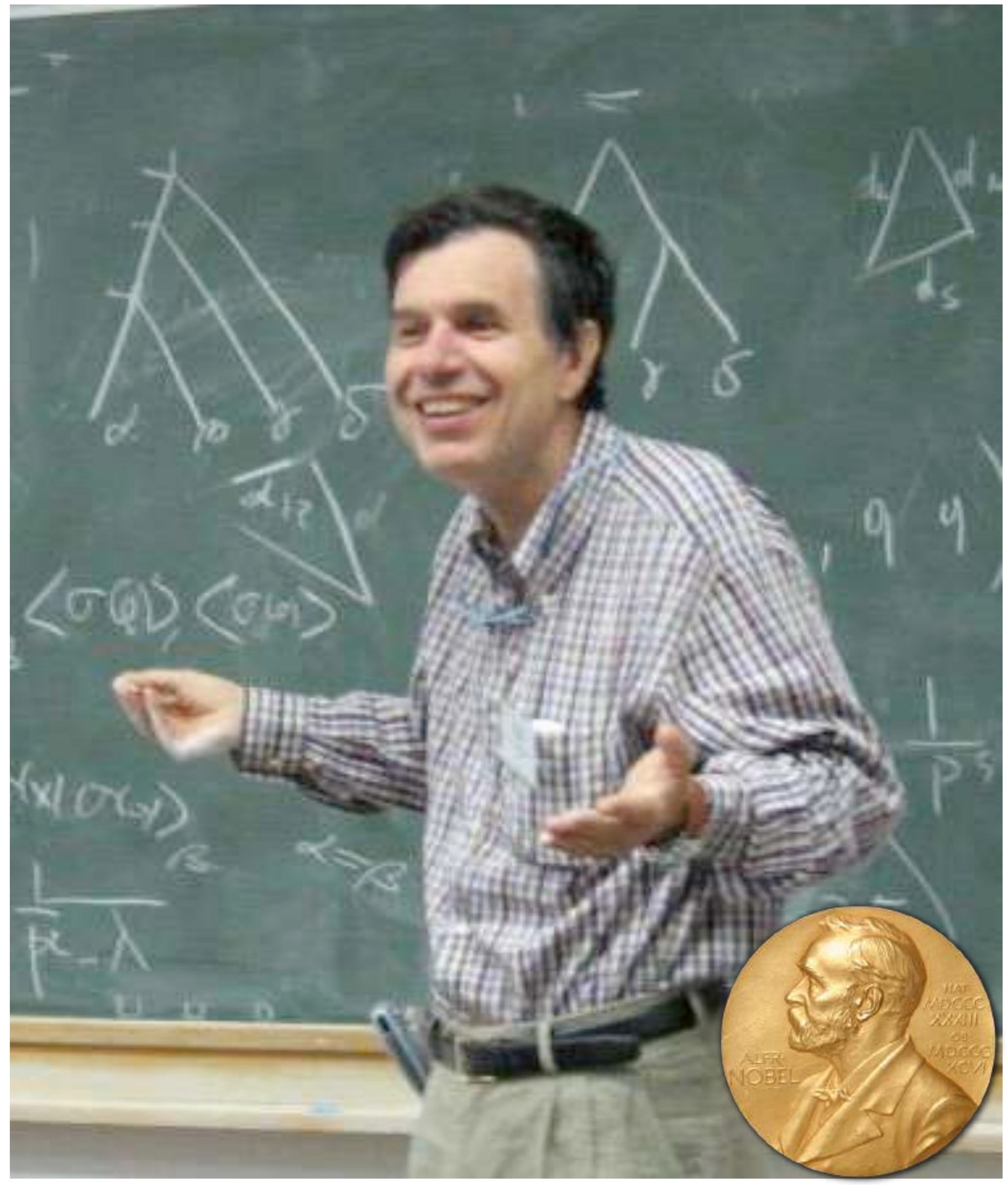


Physics of glasses



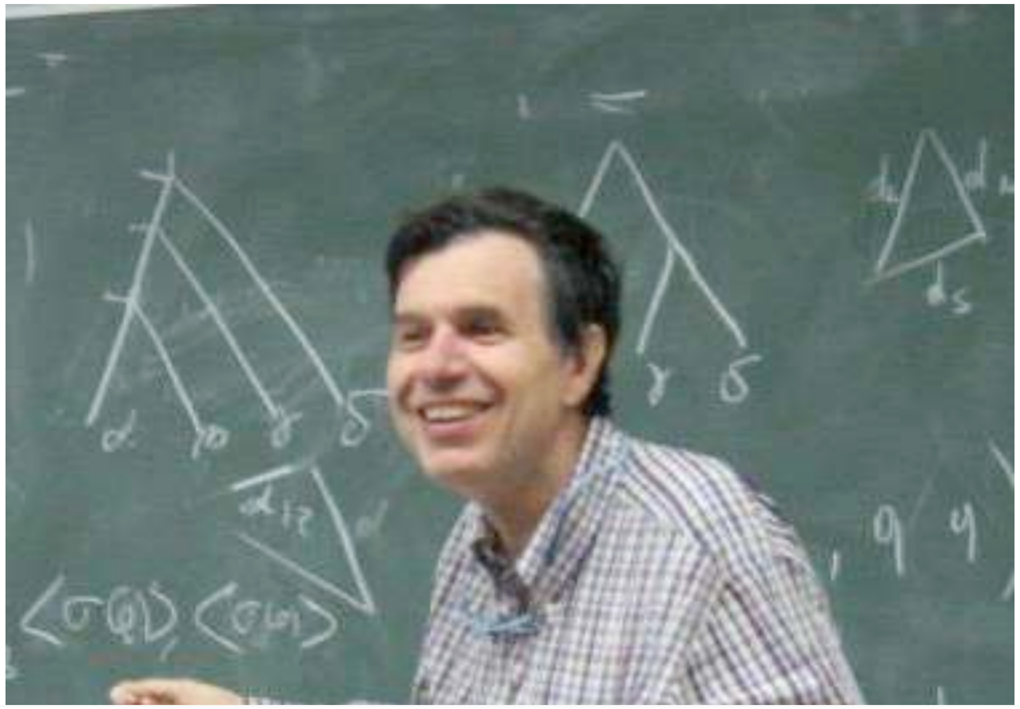


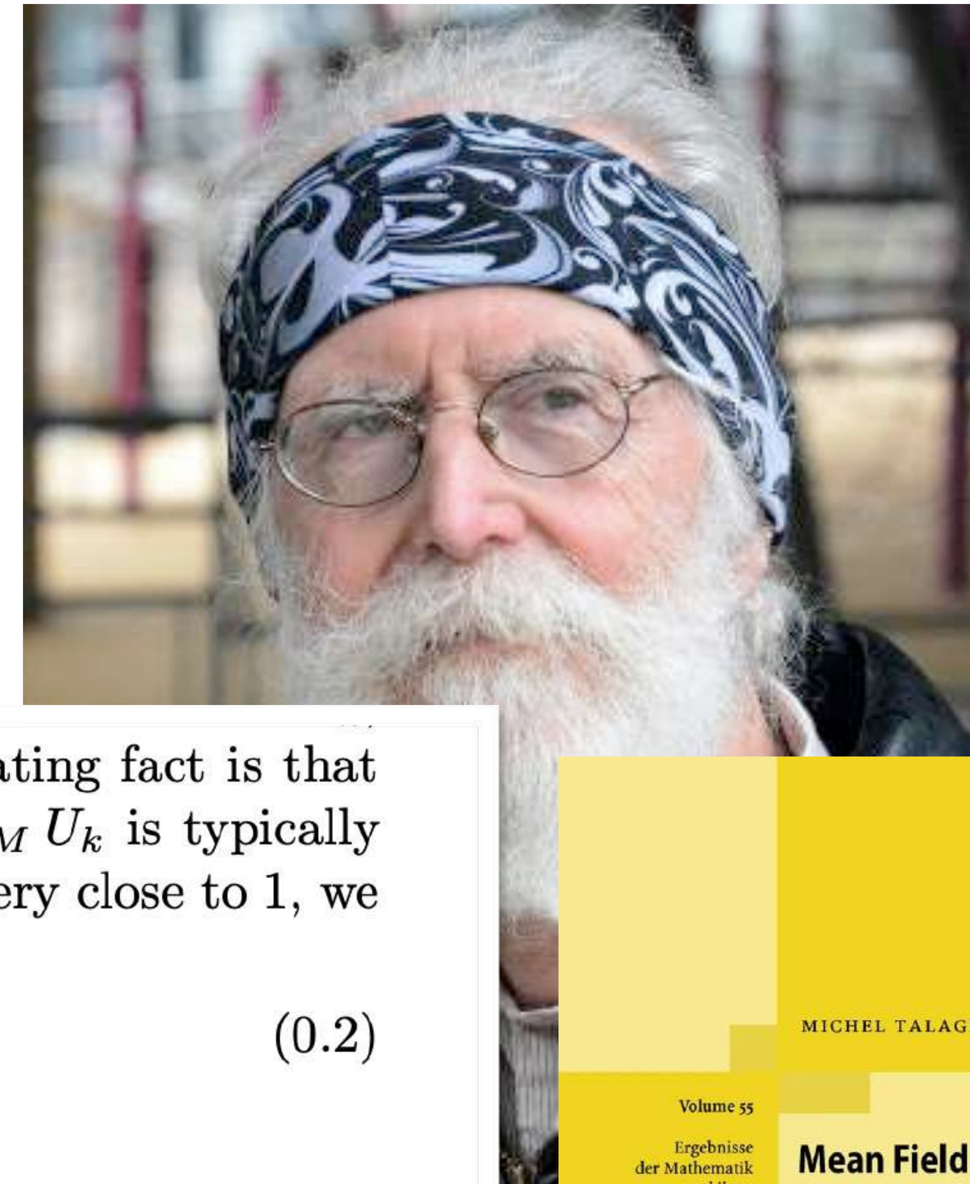
Giorgio Parisi



“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 Michel Talagrand





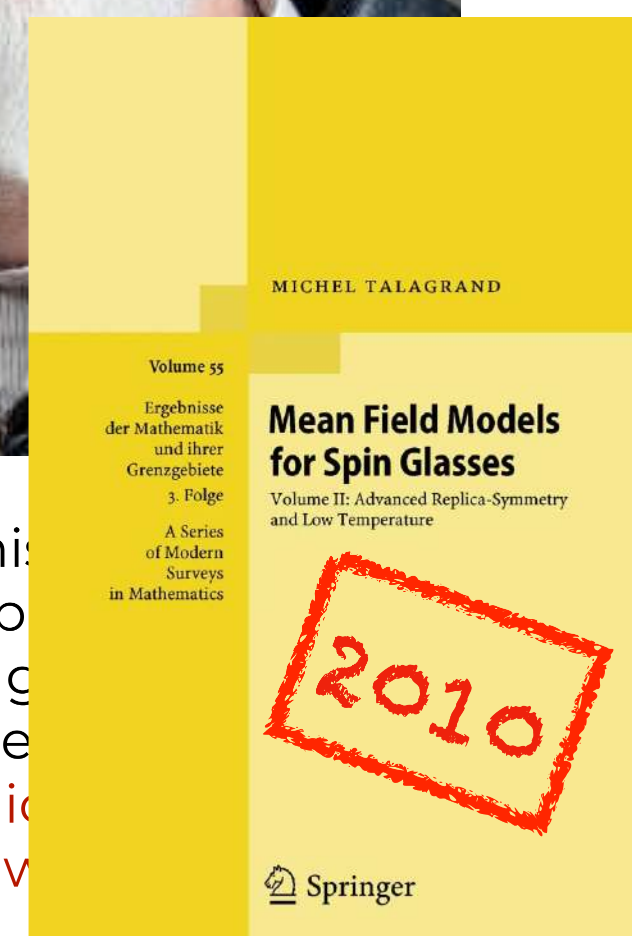
$k \leq M$. This however is not really interesting. The fascinating fact is that when N is large and $M/N \simeq \alpha$, if $\alpha > 2$ the set $\mathbb{S}_N \cap_{k \leq M} U_k$ is typically empty (a classical result), while if $\alpha < 2$, with probability very close to 1, we have

$$\frac{1}{N} \log \mu_N \Big(\mathbb{S}_N \bigcap_{k \leq M} U_k \Big) \simeq \mathrm{RS}(\alpha)\ . \tag{0.2}$$

Here,

$$\mathrm{RS}(\alpha) = \min_{0<q<1} \left(\alpha \, \mathsf{E} \log \mathcal{N} \left(\frac{z\sqrt{q}}{\sqrt{1-q}} \right) + \frac{1}{2} \frac{q}{1-q} + \frac{1}{2} \log(1-q) \right) ,$$

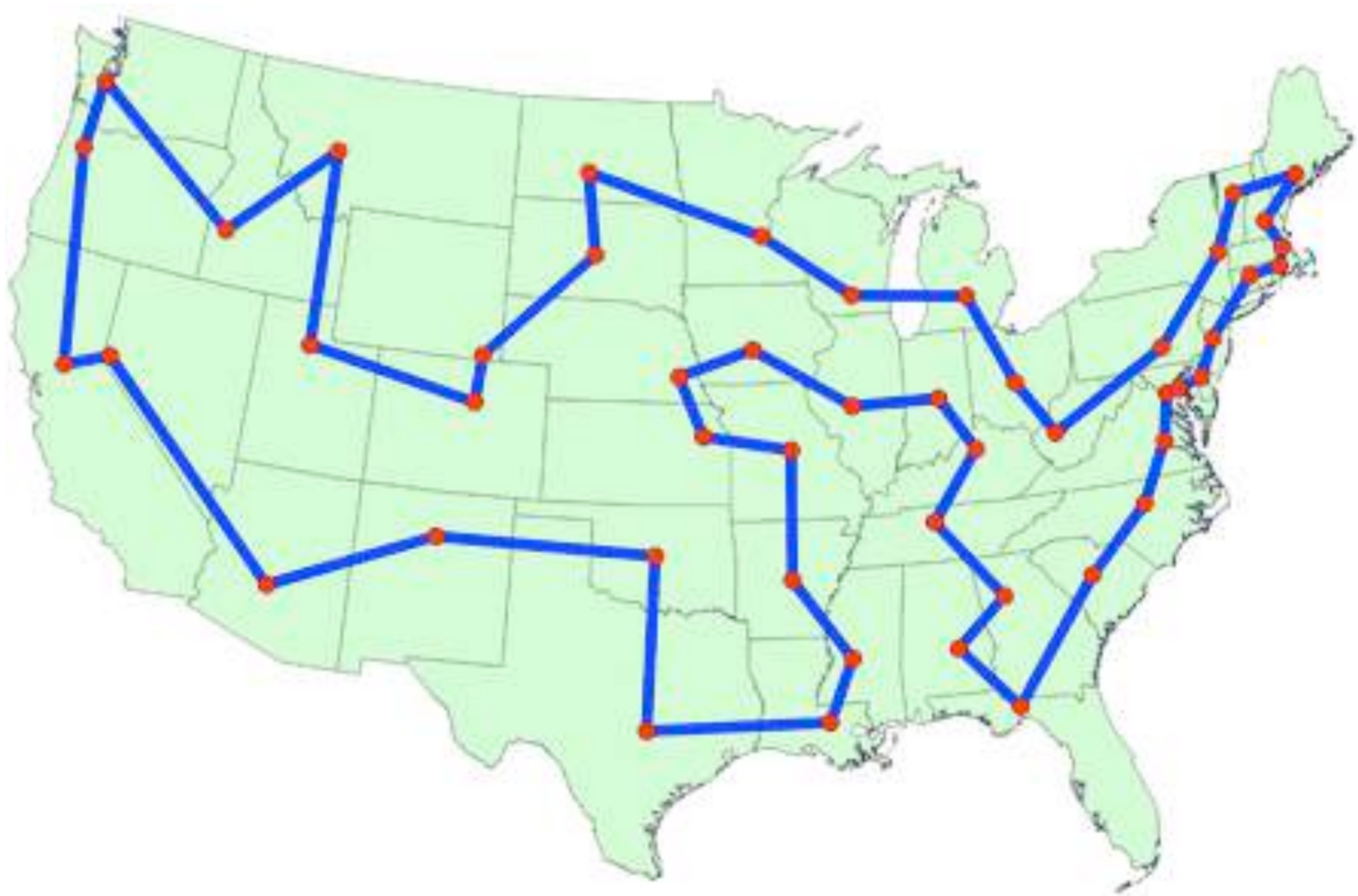
where $\mathcal{N}(x)$ denotes the probability that a standard Gaussian r.v. g is $\geq x$, and where $\log x$ denotes (as everywhere through the book) the natural logarithm of x. Of course you should rush to require medical attention if this formula seems transparent to you. We simply give it now to demonstrate



Other type of "glasses"

<u>Traveling salesman problem:</u>

"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.

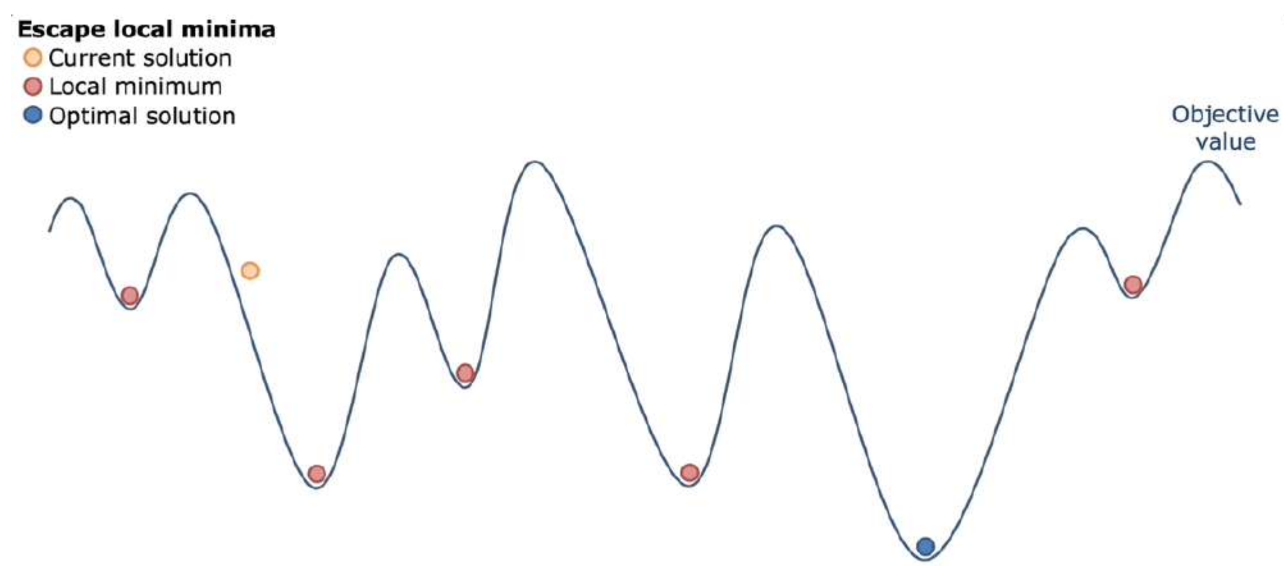

S. Kirkpatrick


C.D. Gelatt


M.P. Vecchi

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.



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.

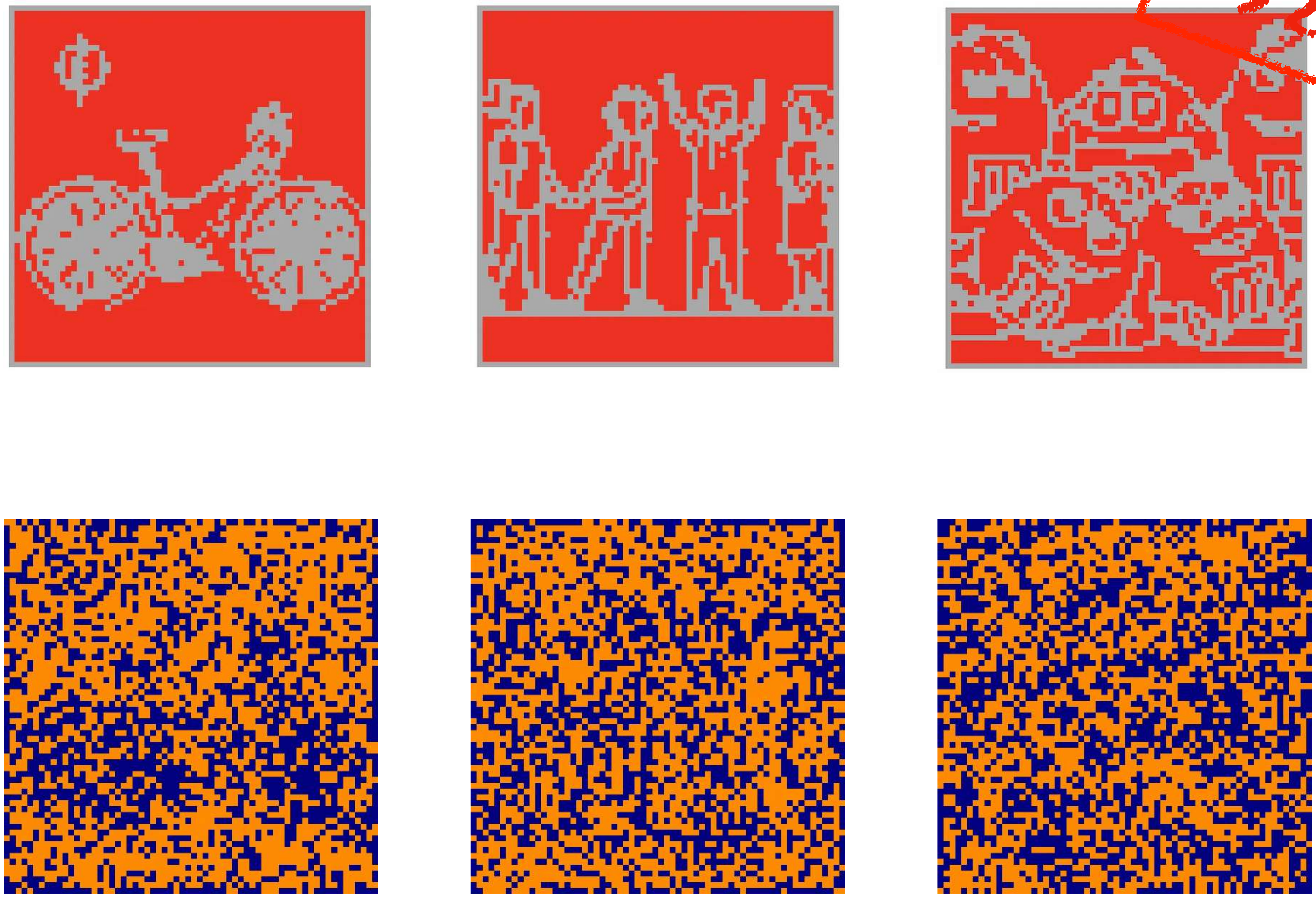
The general content-addressable memory of a physical system

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 $T \sim 0.46 T_c$, 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

The Hopfield Model

1985

PHYSICAL REVIEW A VOLUME 32, NUMBER 2 AUGUST 1985

Spin-glass models of neural networks

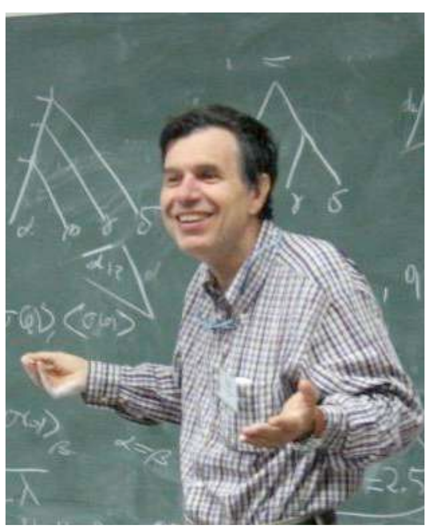
Daniel J. Amit and Hanoch Gutfreund
Racah Institute of Physics, Hebrew University, 91904 Jerusalem, Israel

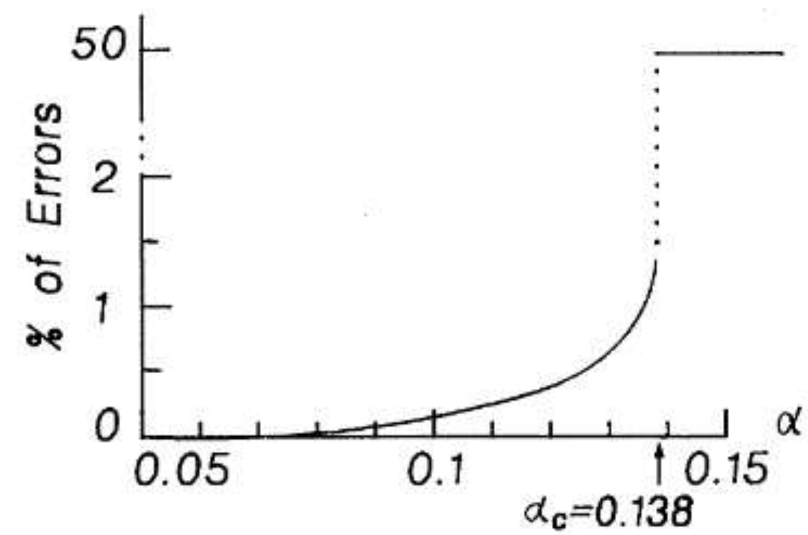
H. Sompolinsky
Department of Physics, Bar-Ilan University, 52100 Ramat-Gan, Israel
(Received 22 March 1985)

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.



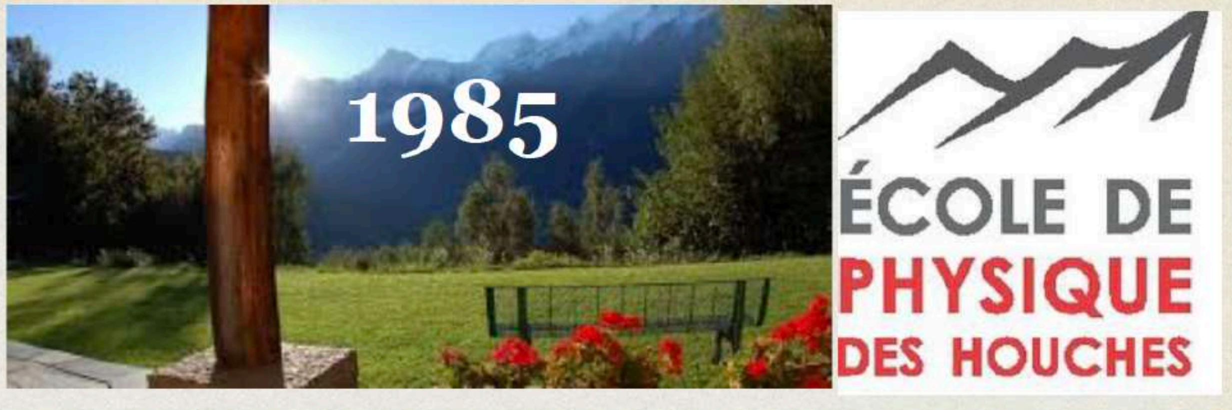
D. Amit H. Gutfreund H. Sompolinsky





And they were not alone...

1985



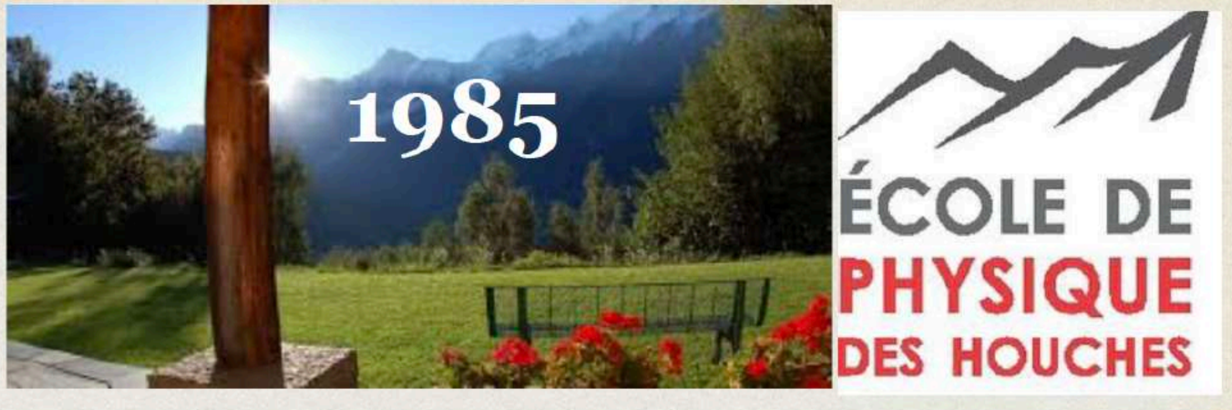
Disordered Systems and Biological Organization

"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



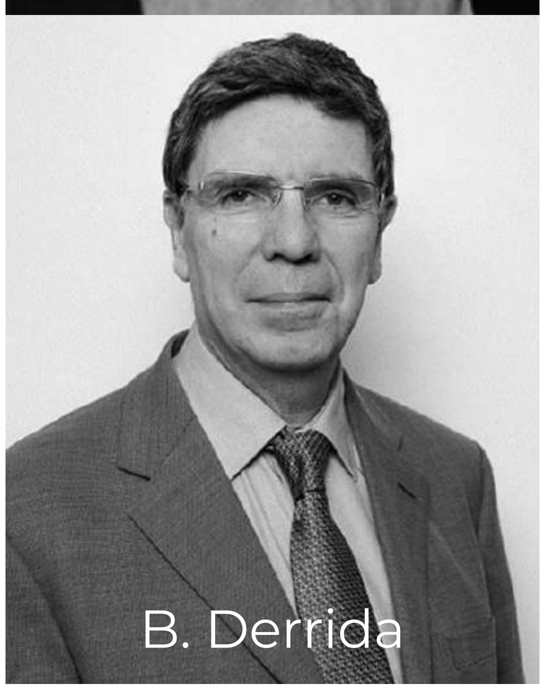
Disordered Systems and Biological Organization

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



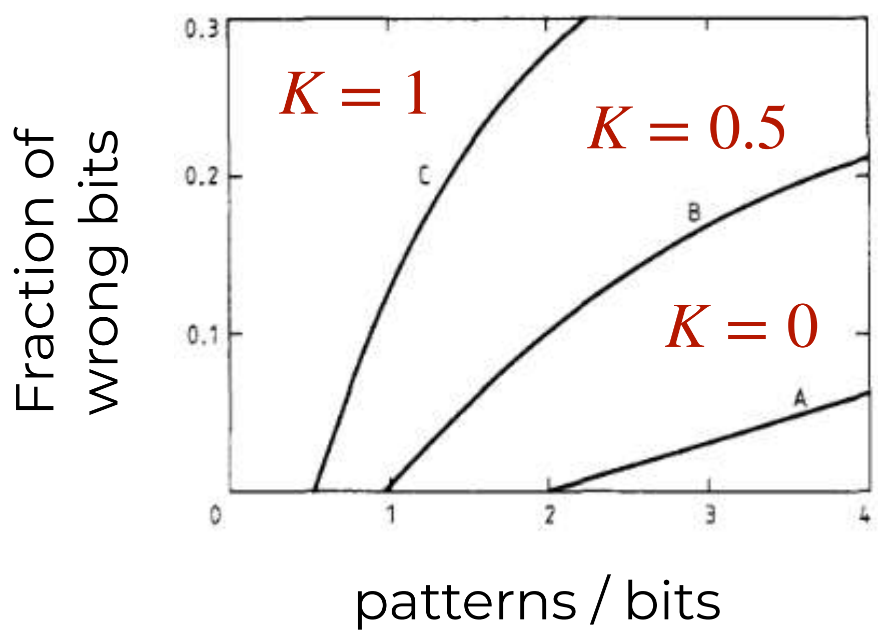
Optimal storage properties of neural network models

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

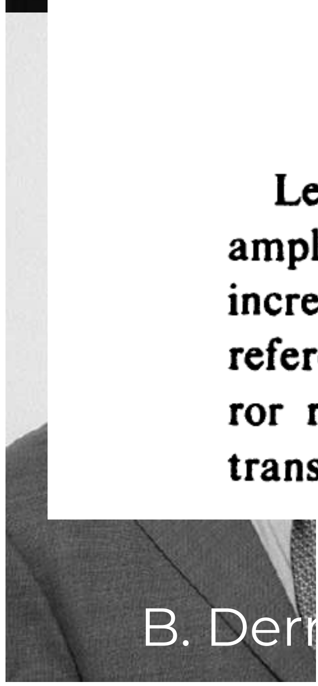
Received 29 May 1987

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.

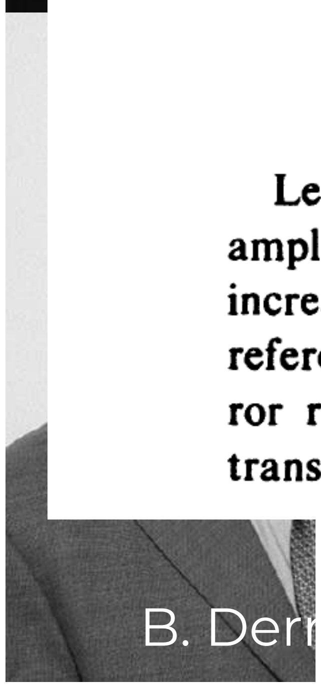


configurations is considered. The volume is calculated explicitly as a function of the storage ratio, $\alpha = p/N$, of the value $\kappa (>0)$ of the product of the spin and the magnetic field at each site and of the magnetisation, m. Here m may vary between 0 (no correlation) and 1 (completely correlated). 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 gi
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.

The Perceptron

1987

Optimal storage properties of neural network models

E Gardner† and B Derrida‡

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-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.

nal
ion
ach
ne,

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-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 from examples by

Learning from Examples in Large I

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

work of the per-
eters 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.

Benign overfitting

1992

Neural Networks and the Bias/Variance Dilemma

Stuart Geman
Division of Applied Mathematics,
Brown University, Providence, RI 02912 USA

Elie Bienenstock
René Doursat
ESPCI, 10 rue Vauquelin,
75005 Paris, France

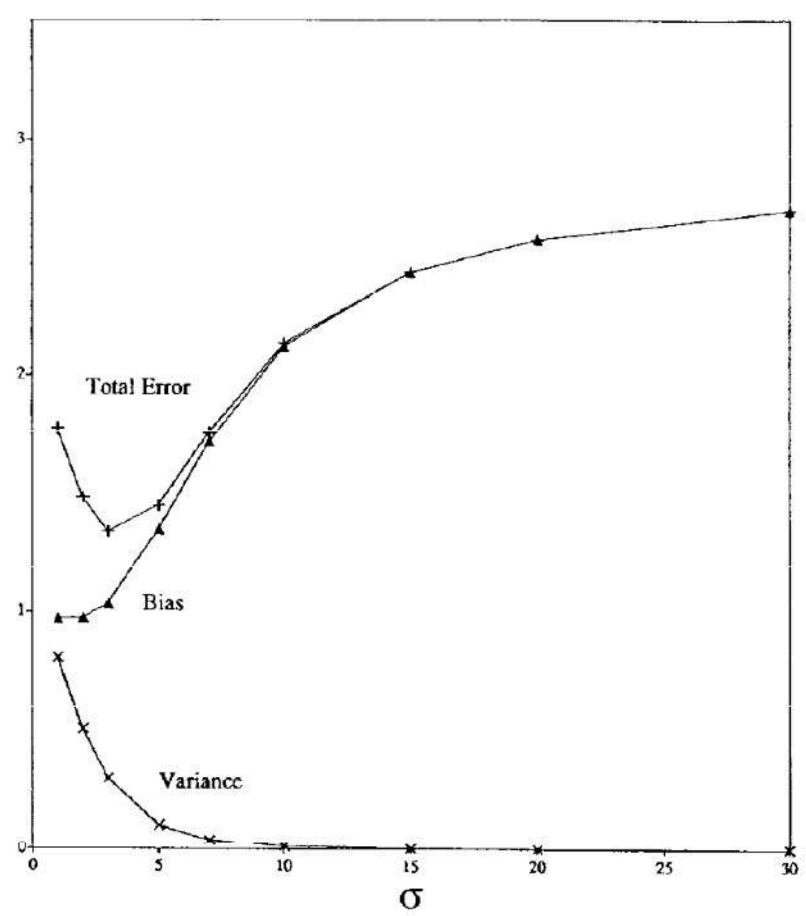


Figure 14: Kernel regression for handwritten numeral recognition. Bias, variance, and total error as a function of kernel bandwidth.

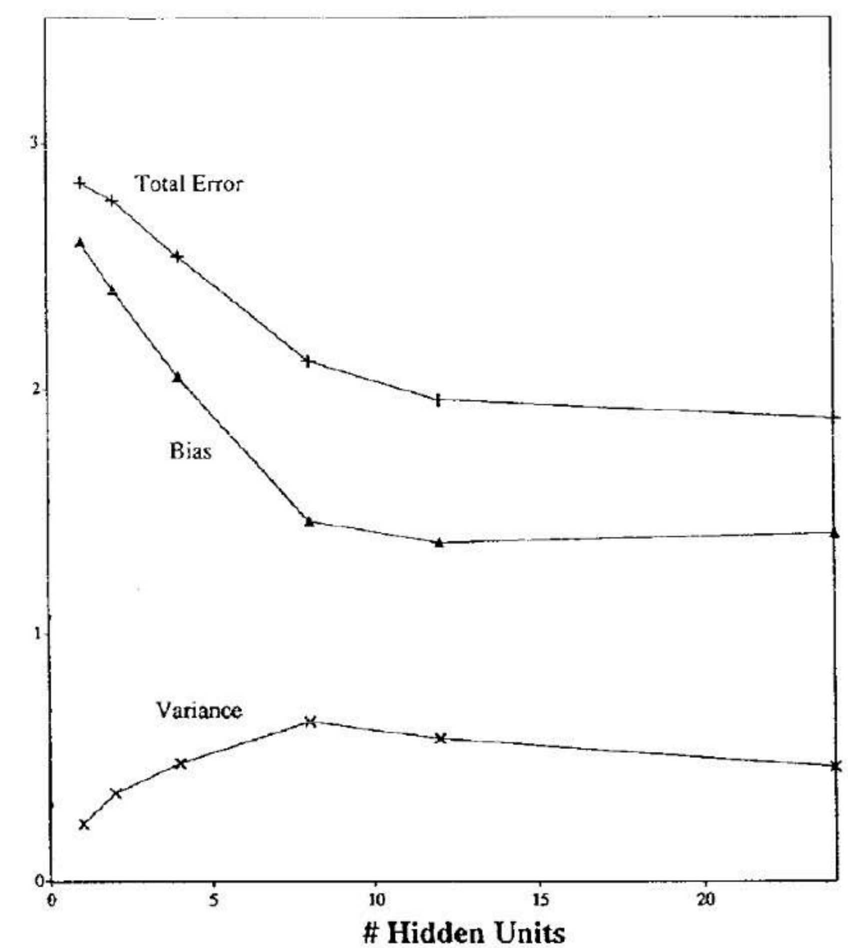


Figure 16: Total error, bias, and variance of feedforward neural network as a function of the number of hidden units. Training is by error backpropagation. For a fixed number of hidden units, the number of iterations of the backpropagation algorithm is chosen to minimize total error.

Double descent

1991

A Simple Weight Decay Can Improve Generalization

Part of Advances in Neural Information Processing Systems 4 (NIPS 1991)

Anders Krogh*
CONNECT, The Niels Bohr Institute
Blegdamsvej 17
DK-2100 Copenhagen, Denmark
krogh@cse.ucsc.edu

John A. Hertz
Nordita
Blegdamsvej 17
DK-2100 Copenhagen, Denmark
hertz@nordita.dk

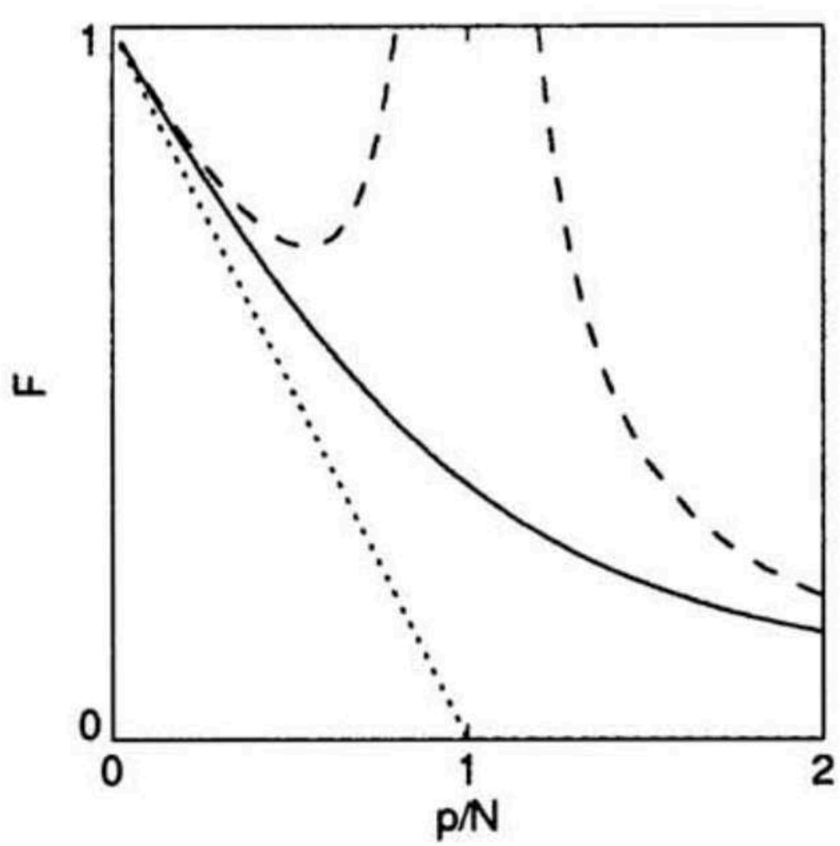


Figure 1: Generalization error as a function of $\alpha = p/N$. The full line is for $\lambda = \sigma^2 = 0.2$, and the dashed line for $\lambda = 0$. The dotted line is the generalization error with no noise and $\lambda = 0$.

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

Reflections After Refereeing Papers for NIPS

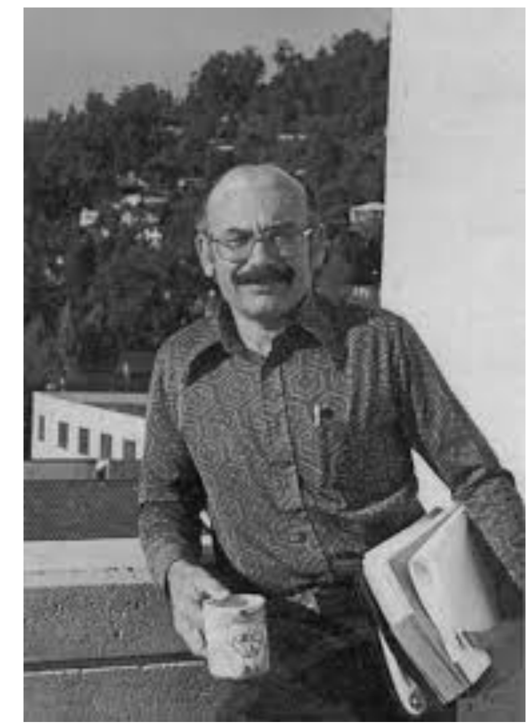
1995

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.



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

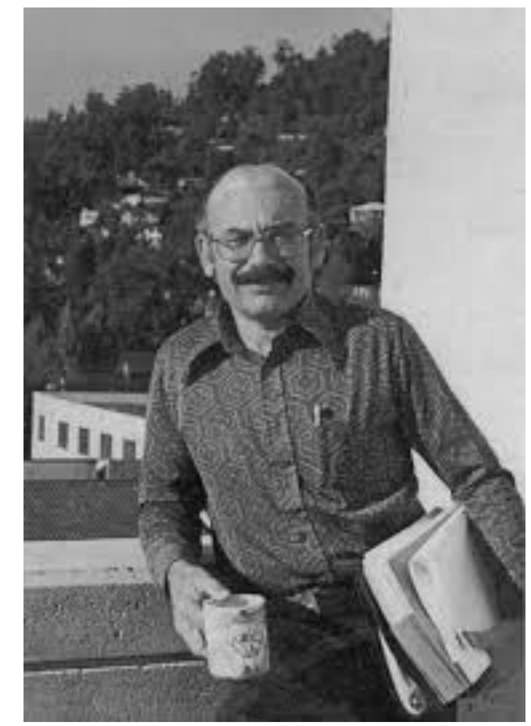
1995

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