The New Challenges posed to Statistical Physics by AI

Marc Mézard

Bocconi University, Milan

March 16, 2026
Galileo Galilei Institute: 20th anniversary
Firenze

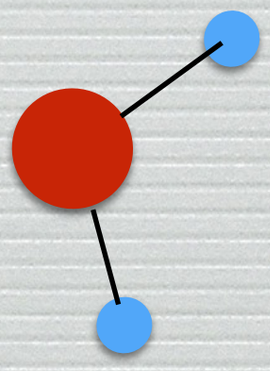
Statistical physics: a long-term perspective

- Maxwell, Boltzmann, etc., 150 years ago, create statistical physics

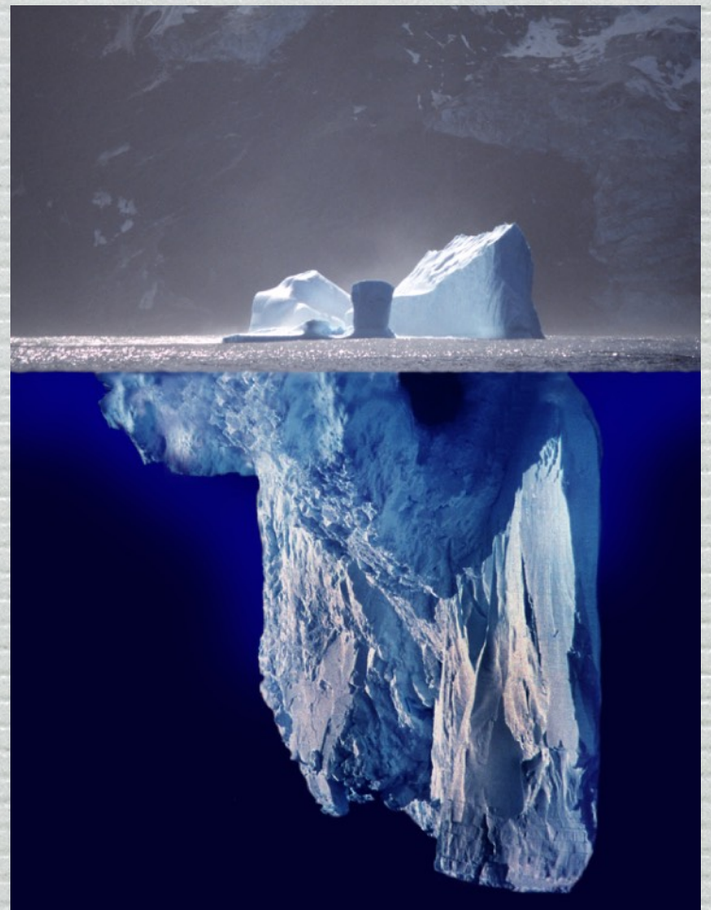
  Give up deterministic description
  Probabilistic approach

- Phase transitions and emergence

From the same microscopic elements: very different macroscopic behavior
Emergence
« More is Different » Anderson 1971



Onsager's solution of 2d Ising model 1944



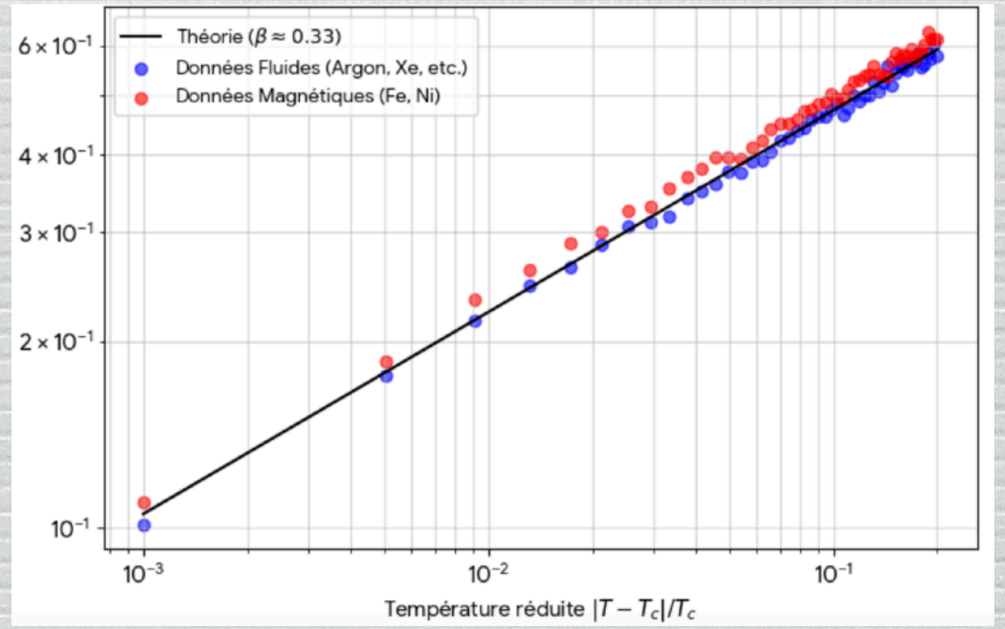
Statistical physics: a long-term perspective

- Zooming on phase transitions: universality and renormalization group

Magnetic systems: magnetization close to T_c : $M \propto (T_c - T)^\beta$

Fluid density close to T_c : $\rho_{liq} - \rho_{gas} \propto (T_c - T)^\beta$

Universal behavior which depends on the microscopic ingredients only through the dimension, the type and symmetry of the order parameter and of interactions
-> Deep connexions to field theory



Kadanoff 1966, Wilson 1971, etc.

Statistical physics: a long-term perspective

50 years ago, creation of a new branch of statistical physics, strongly disordered systems, posing several formidable challenges

Spin glasses. Major developments in the last four decades, starting with Parisi's replica solution of the SK model in 1979.

Statistical physics: a long-term perspective

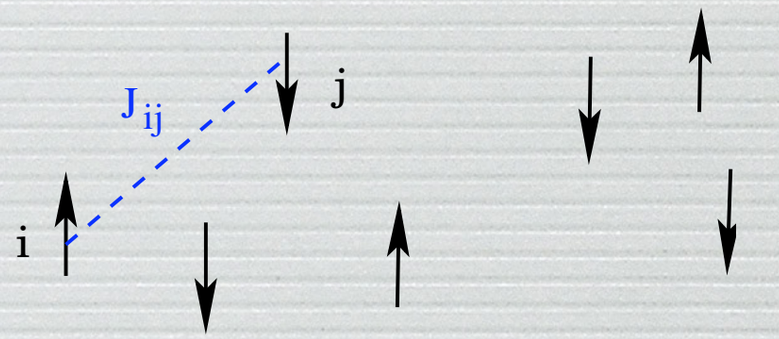
50 years ago, creation of a new branch of statistical physics, strongly disordered systems, posing several formidable challenges

Spin glasses. Major developments in the last four decades, starting with Parisi's replica solution of the SK model in 1979.

Many challenges -> new branch of statistical physics

Challenge 1: ensembles of samples

One sample of a spin glass= set of couplings J between $N \gg 1$ spins. Boltzmann probability measure on the spins $P_J(S)$



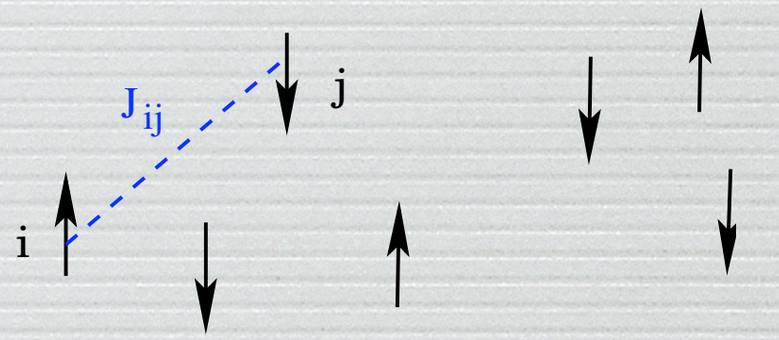
$$s_i = \pm 1$$

$$J = \{J_{ij}\}$$

$$E_J(S) = -\sum_{(i,j)} J_{ij} s_i s_j$$

$$P_J(S) = \frac{1}{Z_J} e^{-\beta E_J(S)}$$

Challenge 1: ensembles of samples



One sample of a spin glass= set of couplings J between $N \gg 1$ spins. Boltzmann probability measure on the spins $P_J(S)$

Ensemble of samples = probability distribution on the set of couplings $\mathcal{P}(J)$

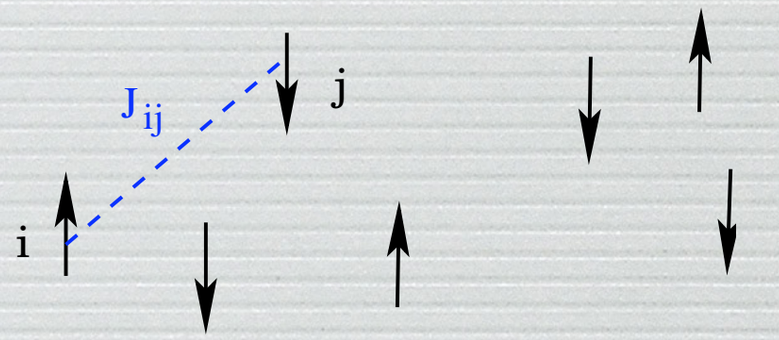
$$s_i = \pm 1$$

$$J = \{J_{ij}\}$$

$$E_J(S) = -\sum_{(i,j)} J_{ij} s_i s_j$$

$$P_J(S) = \frac{1}{Z_J} e^{-\beta E_J(S)}$$

Challenge 1: ensembles of samples



One sample of a spin glass= set of couplings J between $N \gg 1$ spins. Boltzmann probability measure on the spins $P_J(S)$

Ensemble of samples = probability distribution on the set of couplings $\mathcal{P}(J)$

Generate a sample with probability $\mathcal{P}(J)$
What are the properties of spin configurations sampled from $P_J(S)$?

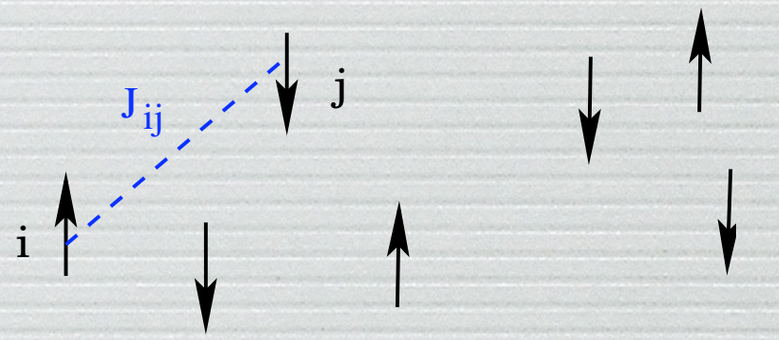
$$s_i = \pm 1$$

$$J = \{J_{ij}\}$$

$$E_J(S) = -\sum_{(i,j)} J_{ij} s_i s_j$$

$$P_J(S) = \frac{1}{Z_J} e^{-\beta E_J(S)}$$

Challenge 1: ensembles of samples



One sample of a spin glass= set of couplings J between $N \gg 1$ spins.
Boltzmann probability measure on the spins $P_J(S)$

Ensemble of samples = probability distribution on the set of couplings $\mathcal{P}(J)$

$$s_i = \pm 1$$

$$J = \{J_{ij}\}$$

$$E_J(S) = -\sum_{(i,j)} J_{ij} s_i s_j$$

$$P_J(S) = \frac{1}{Z_J} e^{-\beta E_J(S)}$$

Generate a sample with probability $\mathcal{P}(J)$
What are the properties of spin configurations sampled from $P_J(S)$?

Quenched disorder: each sample is different.
Thermal disorder: in a given sample, spins fluctuate.

Challenge 1: ensembles of samples

Disorder: each sample is different. Study sample **ensembles**. Find « self-averaging » quantities, which are identical in almost all samples. Understand differences (between samples)

eg Sherrington Kirkpatrick model

$$J_{ij} \sim \mathcal{N}(0, 1/N)$$

$$E_J(S) = -\sum_{(i,j)} J_{ij} s_i s_j$$

$$P_J(S) = \frac{1}{Z_J} e^{-\beta E_J(S)}$$

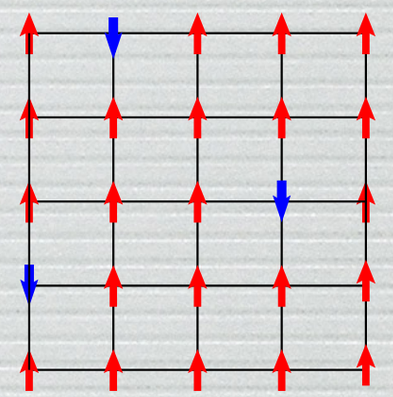
Self-averaging:

$$N \to \infty \qquad \frac{1}{N}\sum_i \langle s_i \rangle \qquad \frac{1}{N}\langle E_J(S) \rangle$$

Sample dependent: details of the landscape, ground state.

Challenge 2: inhomogeneity

Every spin is in a different environment.
Different magnetizations.
No « representative agent ».

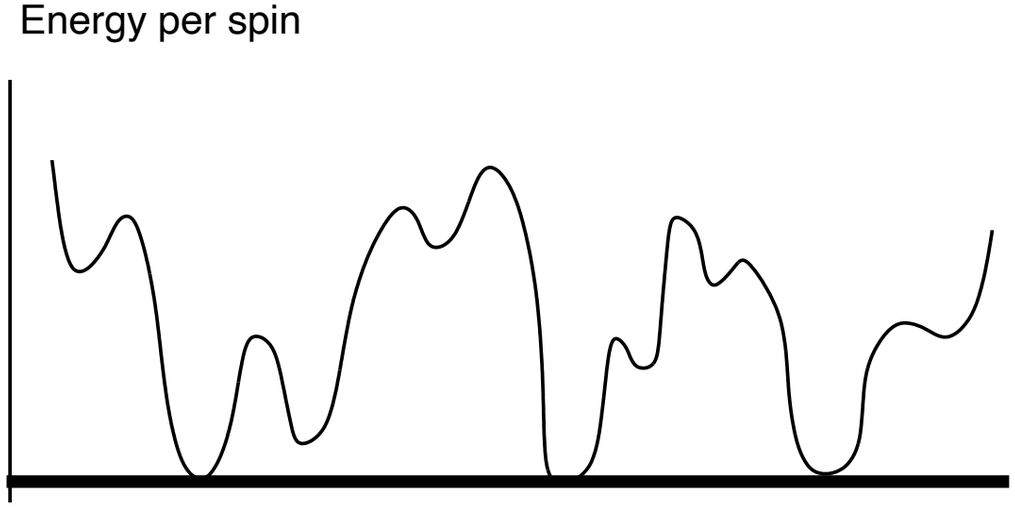


Mean-field equations= N coupled equations for the local magnetizations (Thouless, Anderson, Palmer 1976)
Major simplification from a probability over 2^N configurations

Statistical description of the magnetizations, the local fields: cavity method (M, Parisi, Virasoro 1986). NB: Difficult problem already at the level of existence and nature of the phases

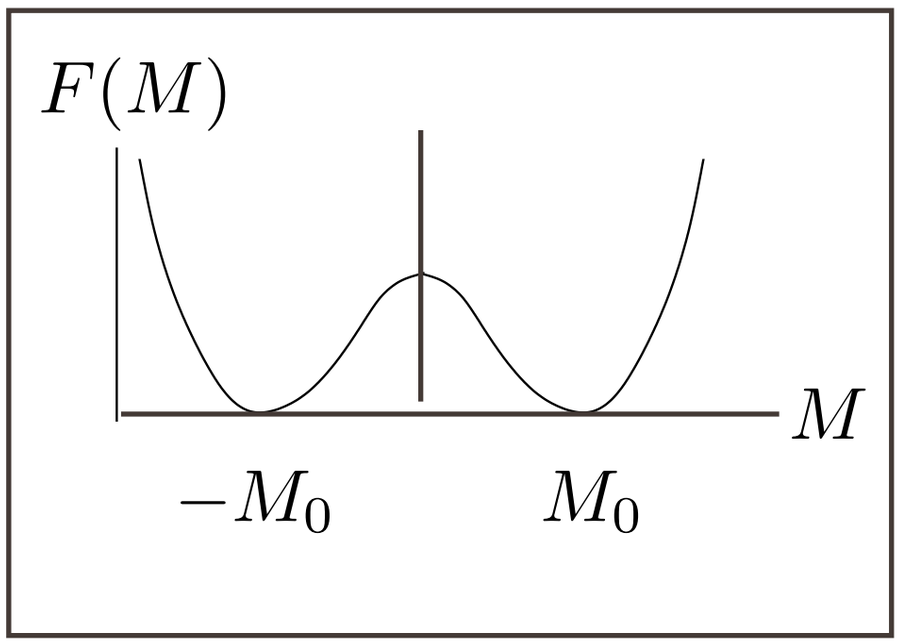
Challenge 3: rough landscape

Complicated landscape, many states in which the spin system can freeze. In SK: hierarchical (ultrametric) structure (MPSTV 85)



(sketch in a N-dimensional space)

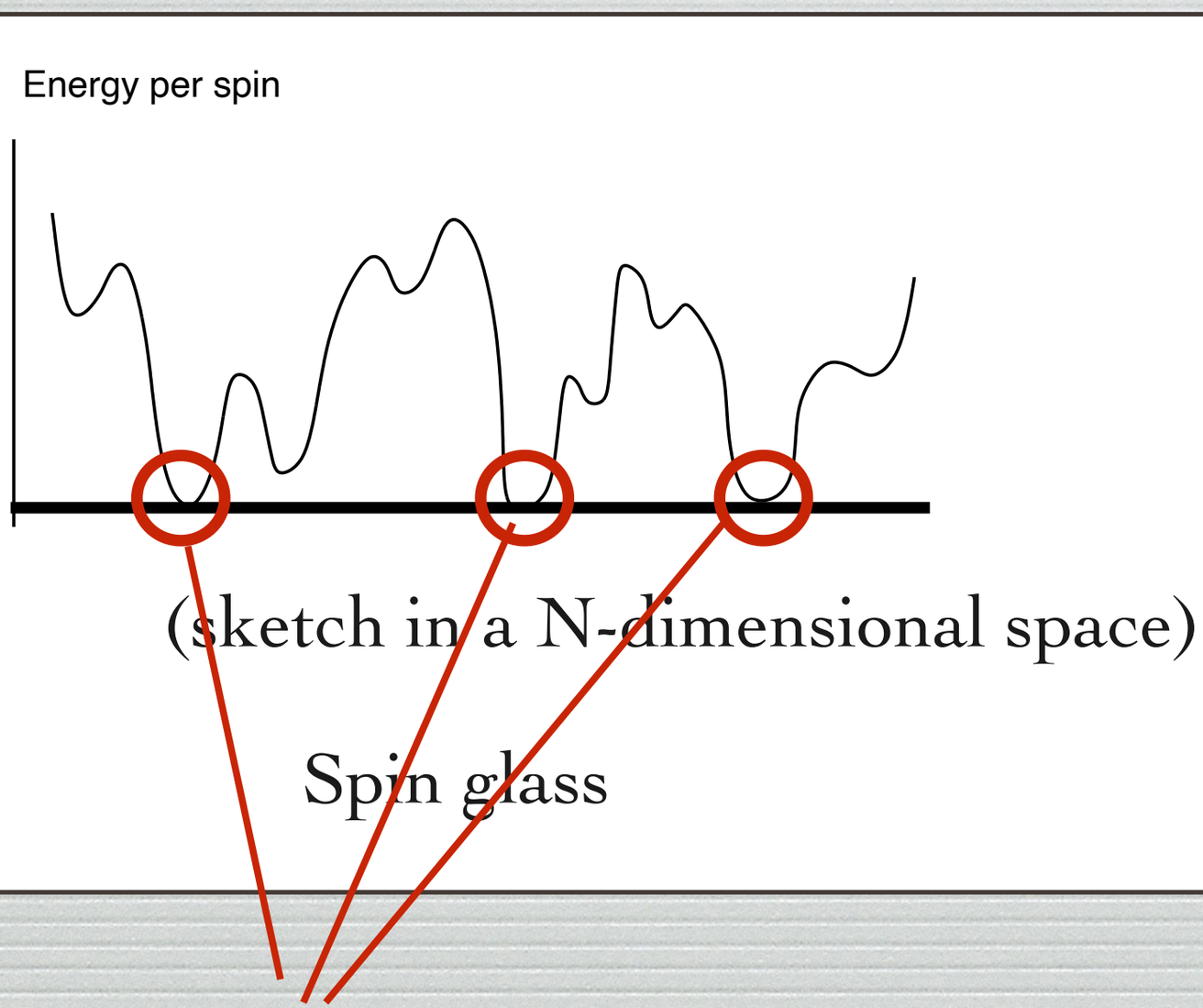
Spin glass



Ferromagnet

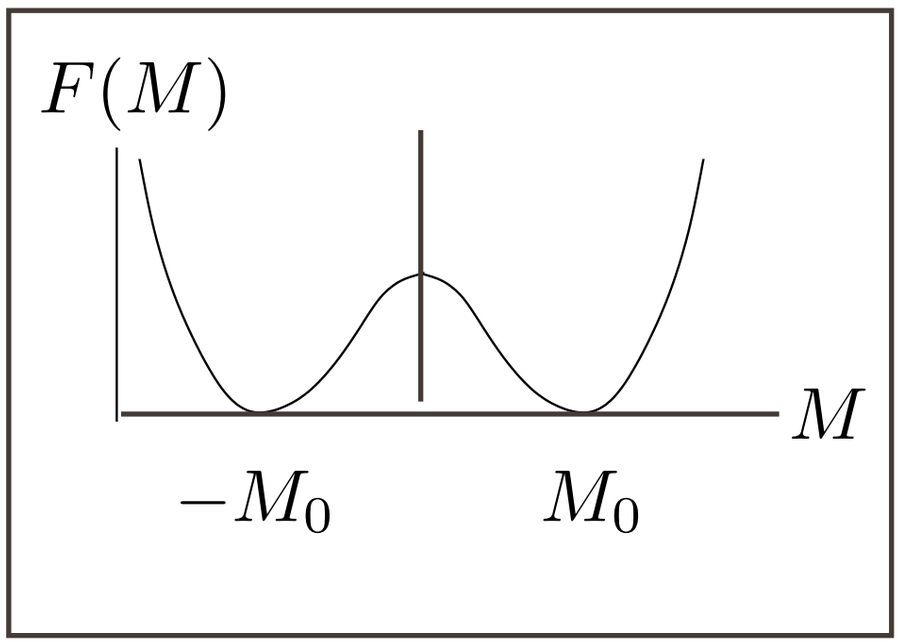
Challenge 3: rough landscape

Complicated landscape, many states in which the spin system can freeze. In SK: hierarchical (ultrametric) structure (MPSTV 85)



(sketch in a N-dimensional space)

Spin glass



Ferromagnet

Details of the landscape depend on the sample !

Landscape and **order parameters**

Ferromagnet: $$M^{\pm} = \lim_{B \to 0^{\pm}} \langle s_i \rangle_B$$

Spin glass: $$M_i^{\alpha} = \lim_{B_i \to 0^{\pm(\alpha)}} \langle s_i \rangle_B$$

Spontaneous symmetry breaking into an unknown, disordered state: unwieldy!

Landscape and **order parameters**

Ferromagnet: $$M^{\pm} = \lim_{B \to 0^{\pm}} \langle s_i \rangle_B$$

Spin glass: $$M_i^{\alpha} = \lim_{B_i \to 0^{\pm(\alpha)}} \langle s_i \rangle_B$$

Spontaneous symmetry breaking into an unknown, disordered state: unwieldy!

Use the system itself as a conjugate field: **replicas**

Landscape and **order parameters**

Ferromagnet: $$M^{\pm} = \lim_{B \to 0^{\pm}} \langle s_i \rangle_B$$

Spin glass: $$M_i^{\alpha} = \lim_{B_i \to 0^{\pm(\alpha)}} \langle s_i \rangle_B$$

Spontaneous symmetry breaking into an unknown, disordered state: unwieldy!

Use the system itself as a conjugate field: **replicas**

Overlap between two equilibrium configurations

$$q = \frac{1}{N} \sum_i s_i^1 s_i^2$$

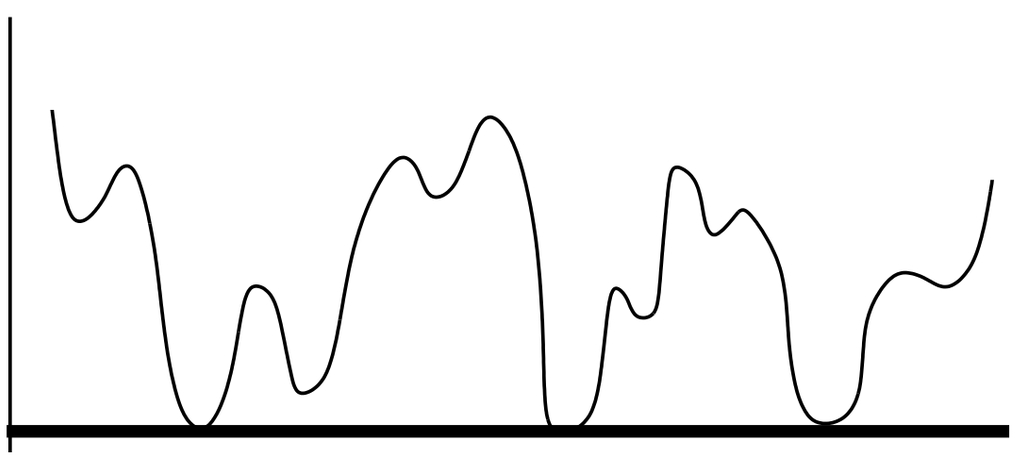
Order parameter = Probability of overlap q : $P_J(q)$

Parisi 82

This order parameter depends on the sample: study its distribution over an ensemble of samples

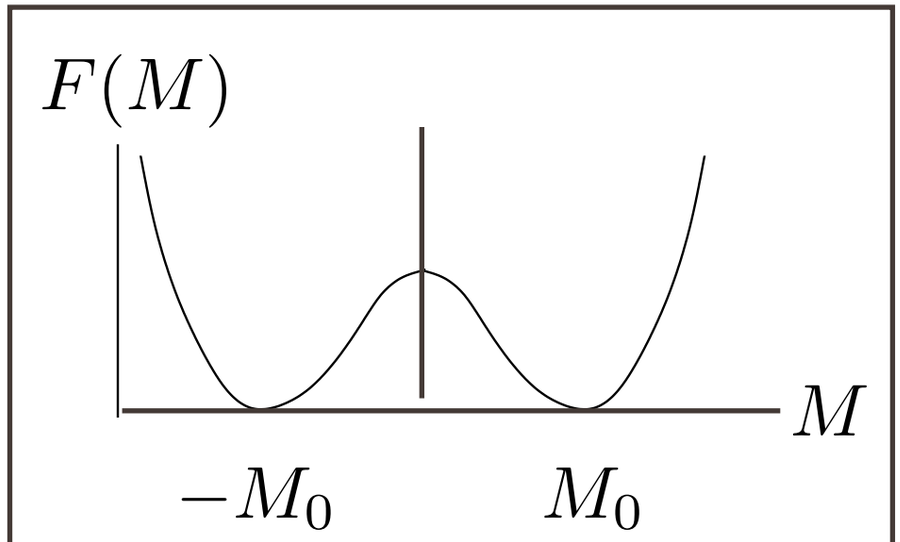
Challenge 4: non equilibrium

Slow relaxation, aging. Non-equilibrium effects crucial



(sketch in a N-dimensional space)

Spin glass



Ferromagnet

Relate equilibrium to non-equilibrium (landscape, fluctuation-dissipation relations)

A new branch of statistical physics

- Study ensembles of problems
- Each spin 'sees' a different local field

  Spins freeze in random directions
- Rough landscape: difficult to find min. of E
- Strong out of equilibrium dynamical effects

NB : beyond the simple mean field theory of the « representative agent »:
Statistics of agents. **Replicas, cavity**…

A new branch of statistical physics

- Study ensembles of problems
- Each spin 'sees' a different local field

  Spins freeze in random directions
- Rough landscape: difficult to find min. of E
- Strong out of equilibrium dynamical effects

NB : beyond the simple mean field theory of the « representative agent »:
Statistics of agents. **Replicas, cavity**…

Useless, but « cornucopia »…

SK= Generic model of binary variables interacting by pairs

Physics of glasses : spin, structural, quantum, interfaces, polymers, random lasers,…

Neural networks: brain, capacity, learning…

Mathematics of glasses and constraint satisfaction problems

Inference, statistics, machine learning, proteins, gene expression networks

Spin glass as a cornucopia

Economy and finance: portfolio, agent-based models, minority games, order books, risk

Information theory, codes, signal processing, compressed sensing

Evolution : biological, prebiotic, chemical, self-organization

Heteropolymers
Protein folding

Optimization and computer science : simulated annealing, quantum annealing, assignment, TSP, K-Sat, BP, SP…

Physics of glasses : spin, structural, quantum, interfaces, polymers, random lasers,…

Neural networks: brain, capacity, learning…

Mathematics of glasses and constraint satisfaction problems

Inference, statistics, **machine learning**, proteins, gene expression networks

Spin glass as a cornucopia

Economy and finance: portfolio, agent-based models, minority games, order books, risk

Information theory, codes, signal processing, compressed sensing

Evolution : biological, prebiotic, chemical, self-organization

Heteropolymers
Protein folding

Optimization and computer science : simulated annealing, quantum annealing, assignment, TSP, K-Sat, BP, SP…

Machine Learning and Large Dimensional Inference

Machine learning going deep: a decade of technological revolution

1- Image understanding.

In the last ten years, detection, segmentation and recognition of objects and regions in images. Image generation.

2- Language analysis: topic classification, question answering, language translation. Language generation

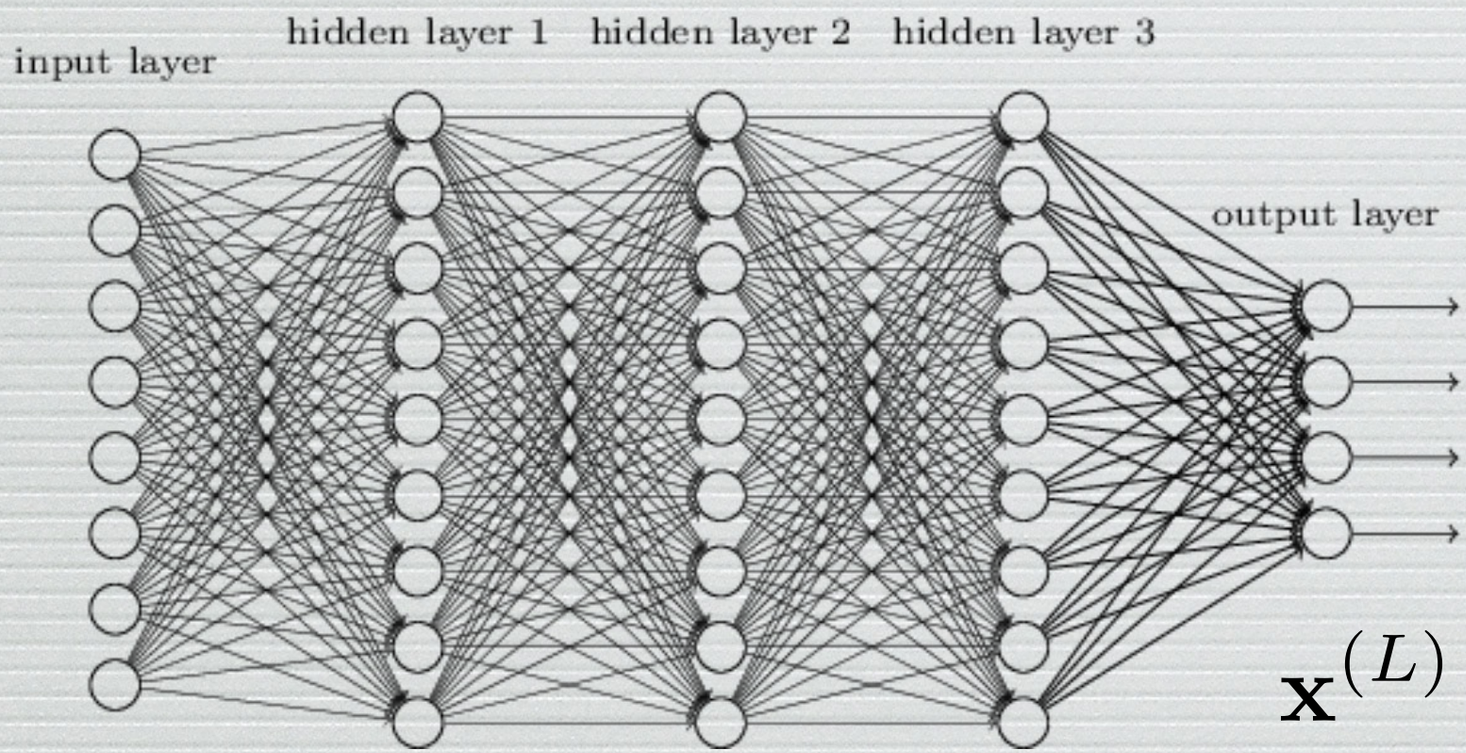
3- Science. Protein Folding. Predicting the activity of potential drug molecules. Meteorology and climate, Algorithmic speedup, feature detection in massive data analysis,…

4- Playing games (chess, go, poker, video-games,…)

etc.

waiting for a general theoretical framework

The tool:
Deep neural network



$\mathbf{x}^{(1)}$ $\mathbf{x}^{(2)}$

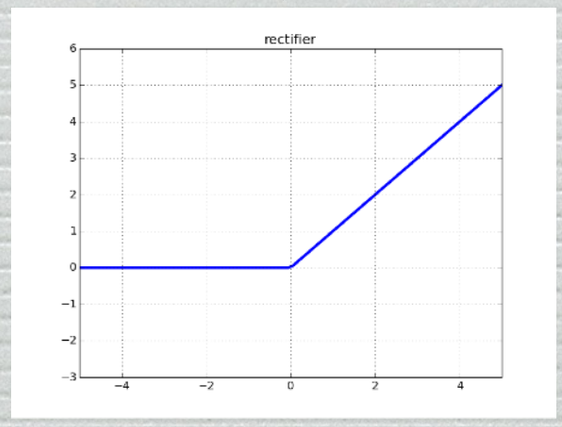
Parameters to be learnt: weights $\mathbb{W}$

$$\mathbf{x}^{(n+1)} = f\left(\mathbb{W}^{(n)}\mathbf{x}^{(n)}\right)$$

$$f = \text{Sign}, \text{Relu}, \tanh \ldots$$

Artificial neuron

$$x_i^{(n+1)} = f\left(\sum_j \mathbb{W}_{ij}^{(n)}\ x_j^{(n)}\right)$$



NB : component-wise nonlinearity

Machine learning

$$\xi \in \mathbb{R}^N \rightarrow y = f(W,\xi) \begin{cases} \in \mathbb{R} & \text{or} \\ \in \{0,1,...q\} \end{cases}$$

Database = P examples of input-output $\quad (\xi_\mu, y_\mu) \quad \mu = 1,...,P$

Machine learning

$$\xi \in \mathbb{R}^N \rightarrow y = f(W,\xi) \begin{cases} \in \mathbb{R} & \text{or} \\ \in \{0,1,...q\} \end{cases}$$

Database = P examples of input-output $(\xi_\mu, y_\mu) \quad \mu = 1,...,P$

Desired label (« supervised learning »)

Machine learning

$$\xi \in \mathbb{R}^N \rightarrow y = f(W, \xi) \begin{cases} \in \mathbb{R} & \text{or} \\ \in \{0, 1, ...q\} \end{cases}$$

Database = P examples of input-output $(\xi_\mu, y_\mu) \quad \mu = 1, ..., P$

Desired label (« supervised learning »)

Learning = Optimization

Find W^* that minimizes the training error: $\sum_{\mu=1}^{P} [f(W, \xi_\mu) - y_\mu]^2$

(or other « loss function »)

Example stochastic gradient descent Very large dimensional landscape.

Machine learning

$$\xi \in \mathbb{R}^N \rightarrow y = f(W, \xi) \begin{cases} \in \mathbb{R} & \text{or} \\ \in \{0, 1, ...q\} \end{cases}$$

Database = P examples of input-output $(\xi_\mu, y_\mu) \quad \mu = 1, ..., P$

Desired label (« supervised learning »)

Learning = Optimization

Find W^* that minimizes the training error: $\sum_{\mu=1}^{P} [f(W, \xi_\mu) - y_\mu]^2$

(or other « loss function »)

Example stochastic gradient descent Very large dimensional landscape.

The big Challenge: Generalization

Use the optimal W^*, test the machine on new data

Machine learning: learning phase

$$\xi \in \mathbb{R}^N \rightarrow y = f(W,\xi) \begin{cases} \in \mathbb{R} & \text{or} \\ \in \{0,1,...q\} \end{cases}$$

Database = P examples of input-output (ξ_μ, y_μ) $\mu = 1,...,P$

Desired label (« supervised learning »)

Machine learning: learning phase

$$\xi \in \mathbb{R}^N \rightarrow y = f(W,\xi) \begin{cases} \in \mathbb{R} & \text{or} \\ \in \{0,1,...q\} \end{cases}$$

Database = P examples of input-output (ξ_μ, y_μ) $\mu = 1,...,P$

Desired label (« supervised learning »)

Bayesian learning:

$$P(W|\{\xi_\mu, y_\mu\}) = \frac{1}{Z}\, P^0(W)\, \exp\left(-\beta \sum_\mu [f(W,\xi_\mu) - y_\mu)]^2\right)$$

Unknown Data Prior Loss

Effective inverse temperature allows to tune the importance of data wrt prior (annealing)

Machine learning: learning phase

Disordered system. Database = sample= disorder. For each database, study the properties of the probability measure on the weights

- Specific database, MNIST, CIFAR, etc
- Statistical ensemble of database. Generative models

Bayesian learning:

$$P(W|\{\xi_\mu, y_\mu\}) = \frac{1}{Z}\, P^0(W)\, \exp\left(-\beta \sum_\mu [f(W,\xi_\mu) - y_\mu)]^2\right)$$

Unknown Data Prior Loss

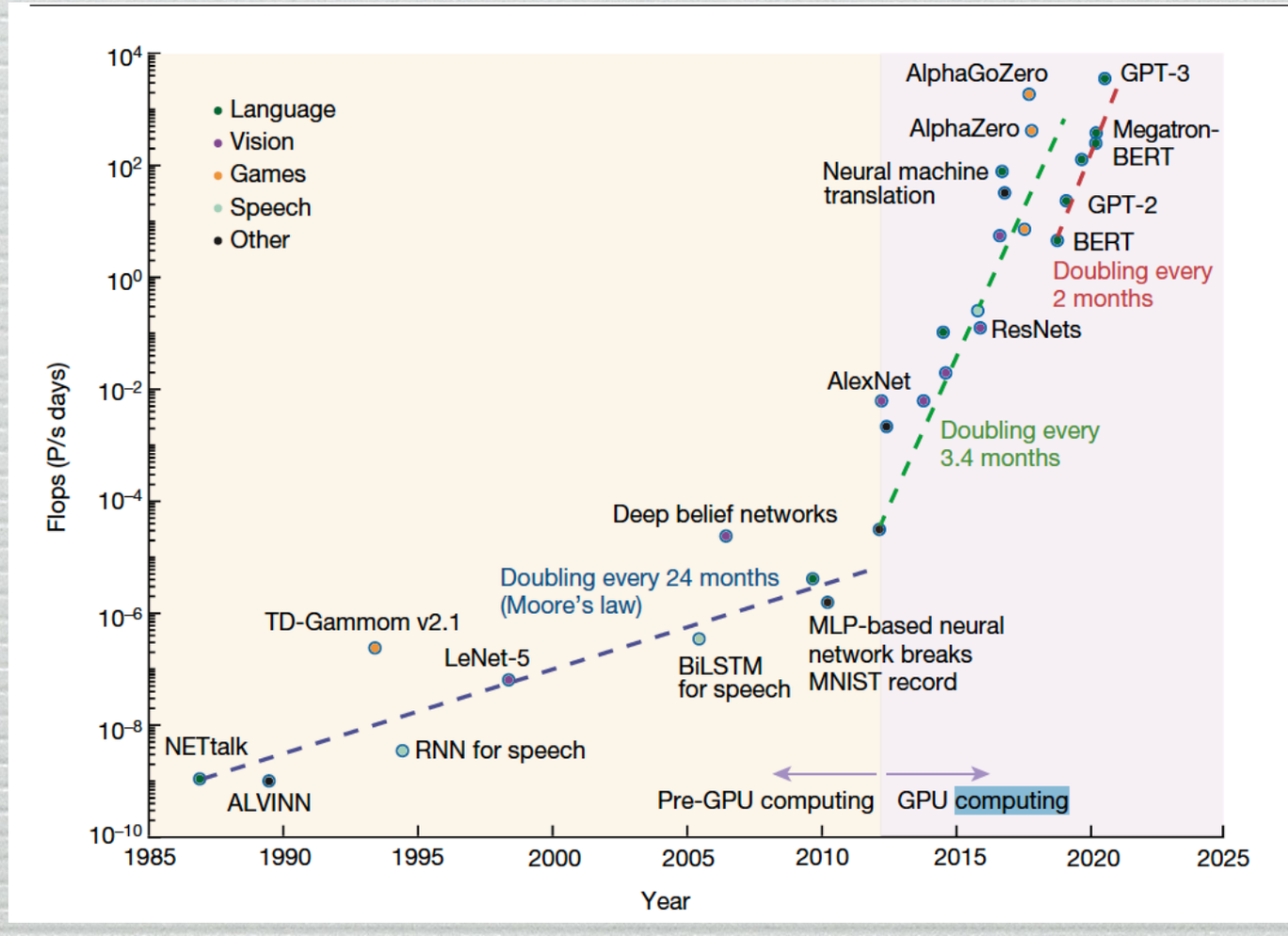
Effective inverse temperature allows to tune the importance of data wrt prior (annealing)

The (old) ingredients

- Feedforward neural networks
- Trained with gradient descent learning, implemented with gradient back propagation

What is new in practice since the 80's ?

- Availability of very large data bases
- Much larger computing power
- Much deeper networks
- Generative models

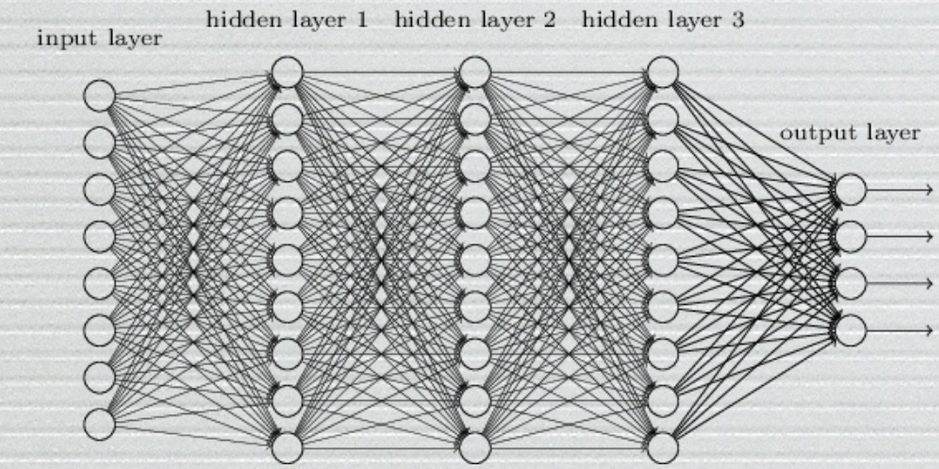


Some surprises and questions

Training

Some surprises and questions

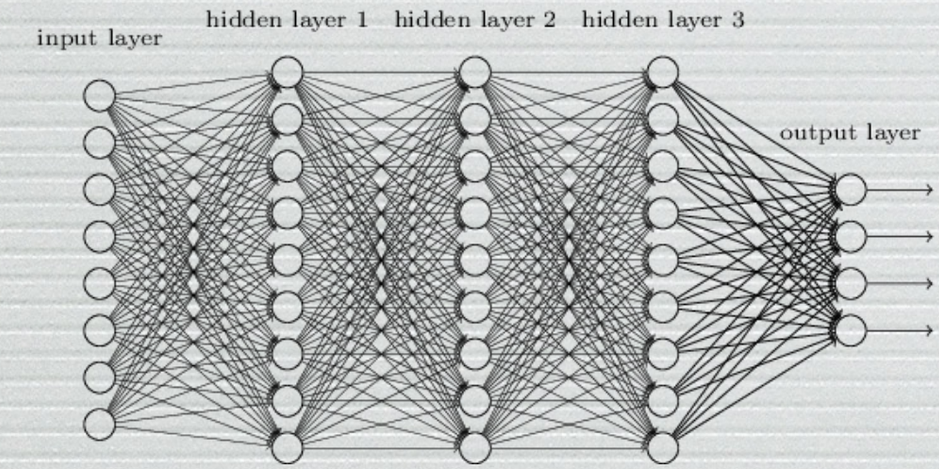
Training



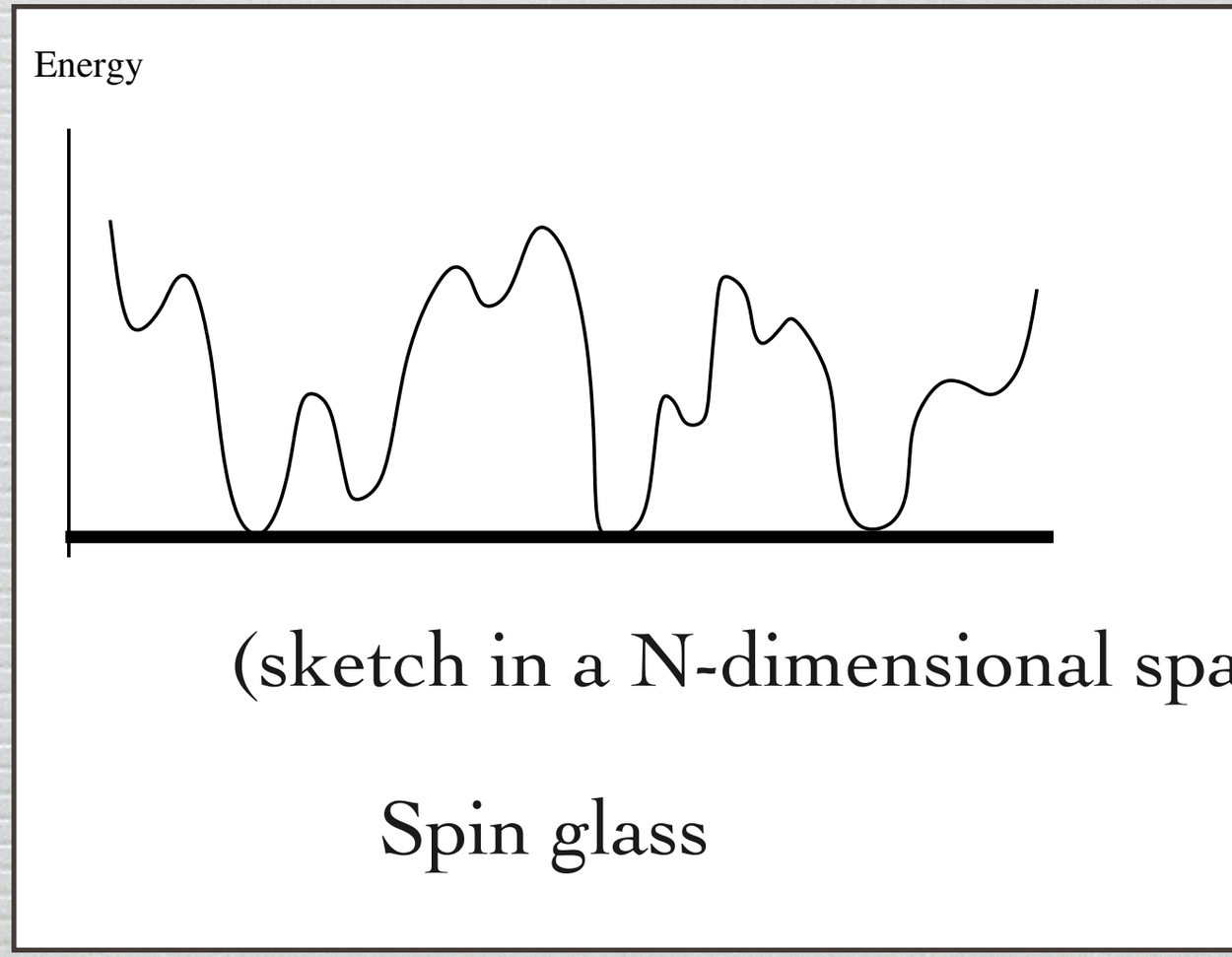
Training = optimization of a disordered system in a large dimensional space

Some surprises and questions

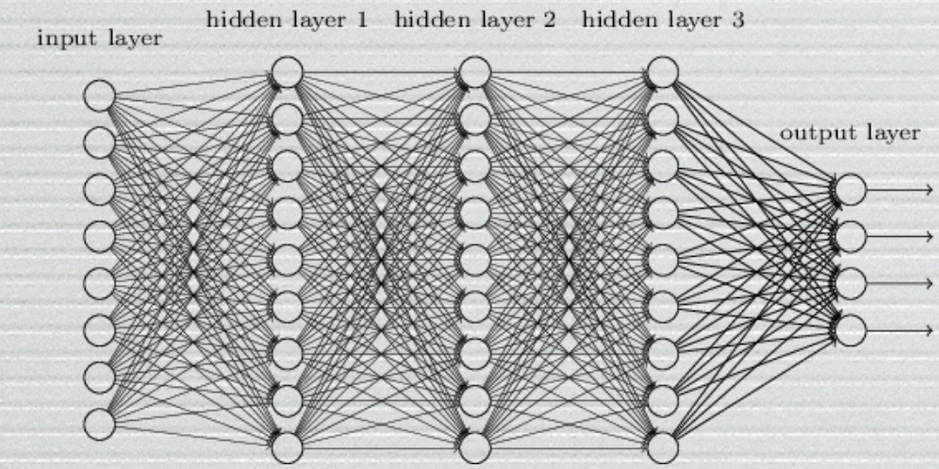
Training



Training = optimization of a disordered system in a large dimensional space

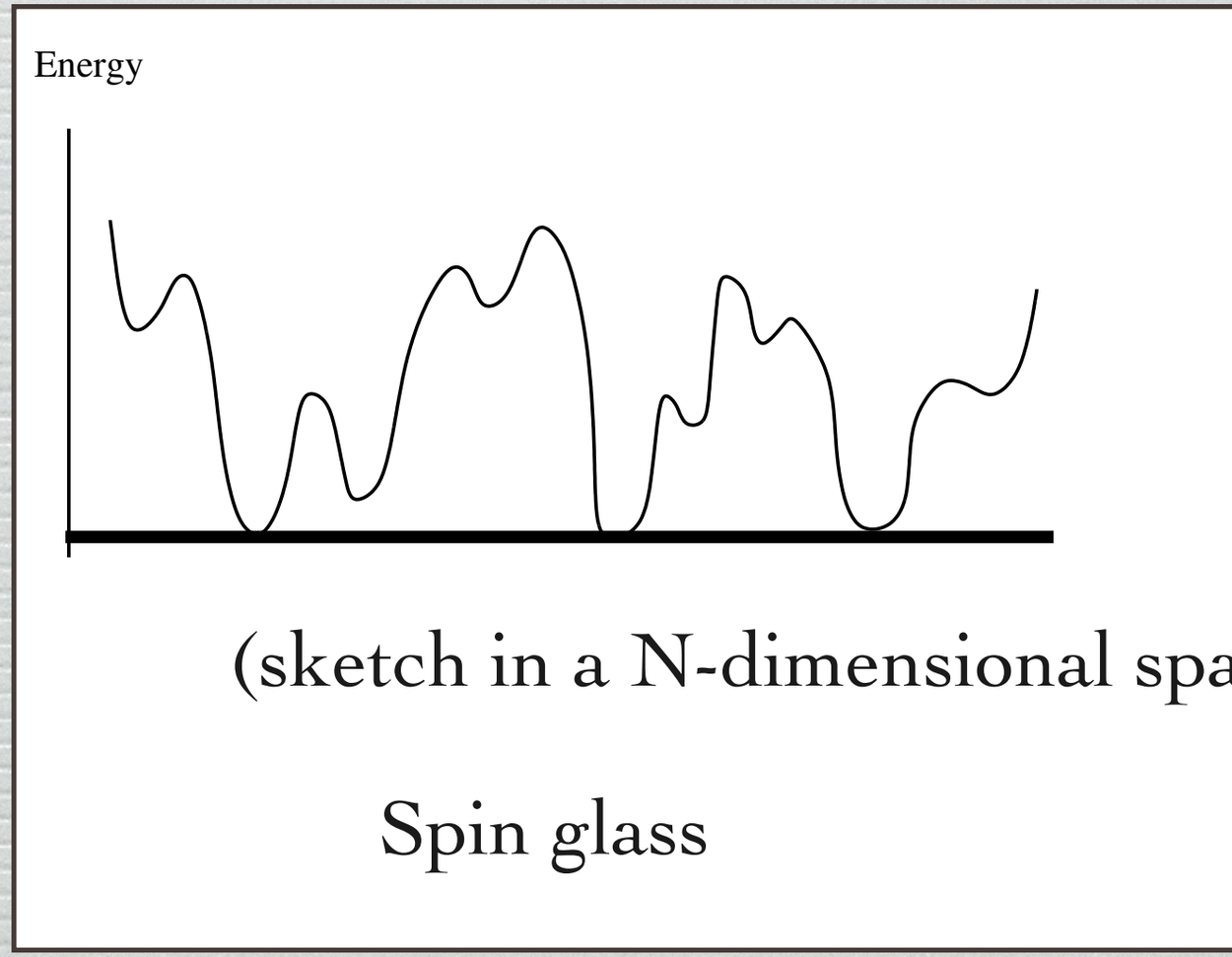


Some surprises and questions Training



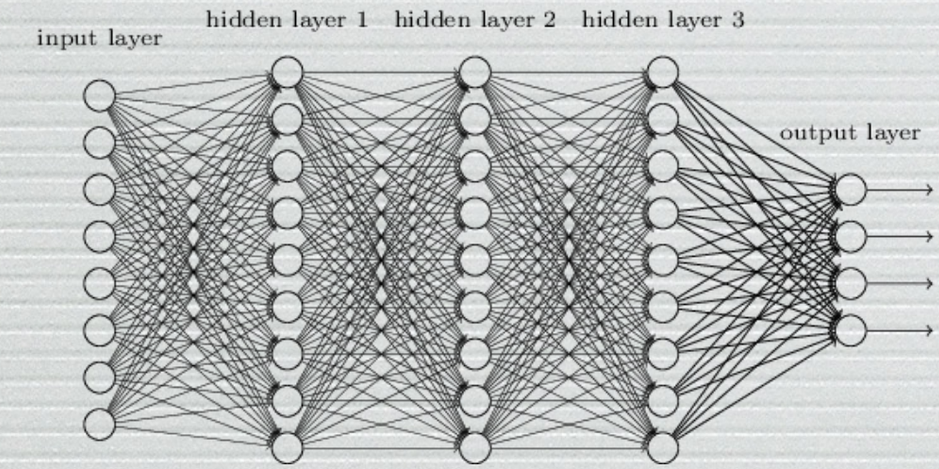
Training = optimization of a disordered system in a large dimensional space

Experimentally: one can reach zero training error, using simple stochastic gradient descent, in the neighborhood of any random starting point provided the network is deep enough



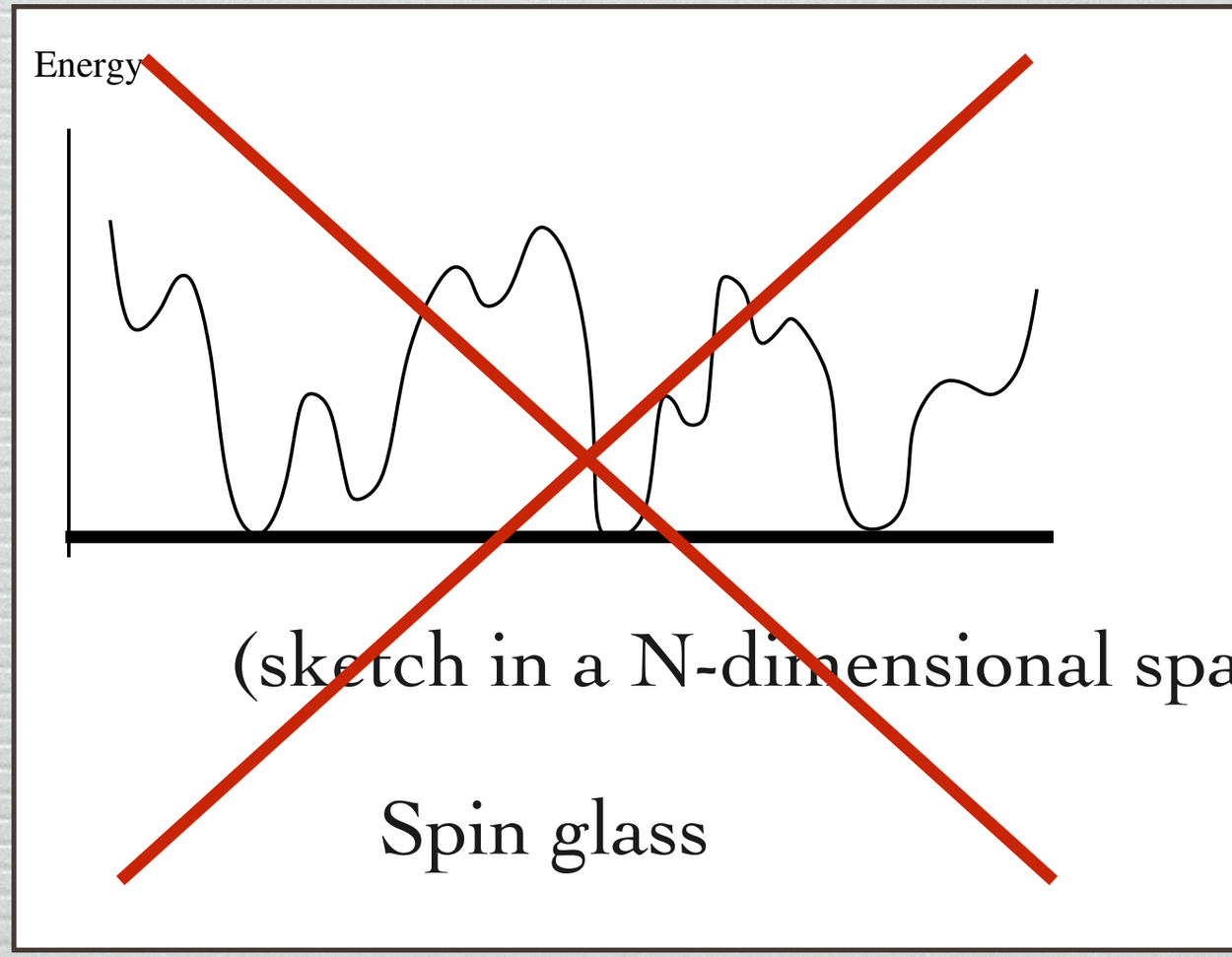
Some surprises and questions

Training



Training = optimization of a disordered system in a large dimensional space

Experimentally: one can reach zero training error, using simple stochastic gradient descent, in the neighborhood of any random starting point provided the network is deep enough



Some surprises and questions

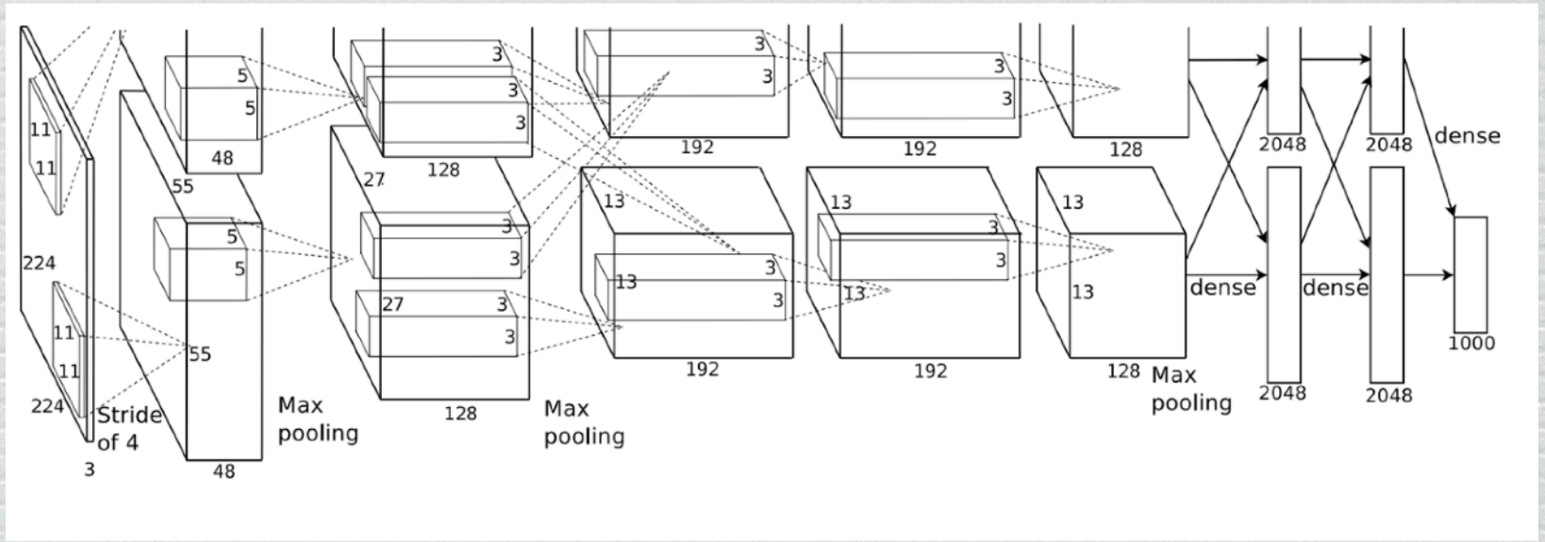
Generalization :

We train with billions of parameters.

Why no overfitting?

Some surprises and questions

Mechanism



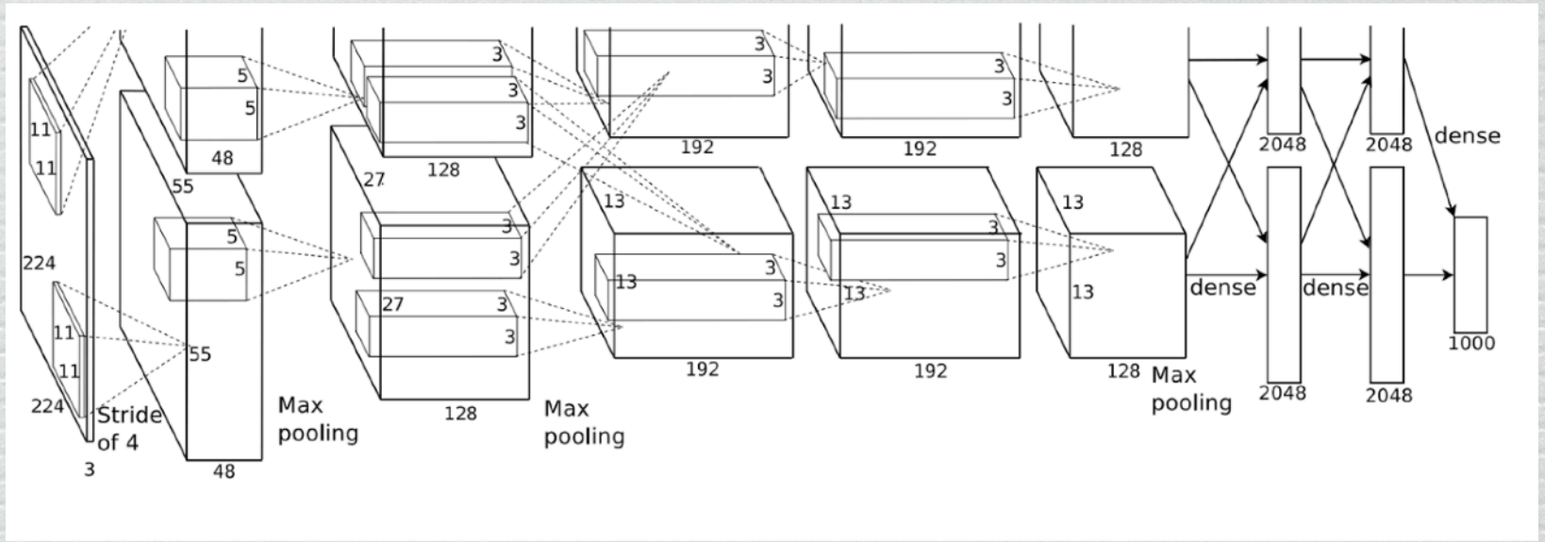
We know everything of the trained network

(neuroscientist's dream)

We do not understand much. Emergent phenomenon

Some surprises and questions

Mechanism



We know everything of the trained network

(neuroscientist's dream)

We do not understand much. Emergent phenomenon

No guarantee

No explanation

Ingredients of deep networks

Architecture

Art. Go deep, use convolutions in first layers, use pooling, etc…

Learning algorithms

Art. The (nearly) most naive algorithm, stochastic gradient descent initialized with small weights, works well (with optimized hardware

« Simple » Data structure

Maybe the tasks that machine learning addresses are easier than expected because **data has a lot more structure** than our theories (worst case, or typical case with iid data) used so far

The Challenge of Data Structure

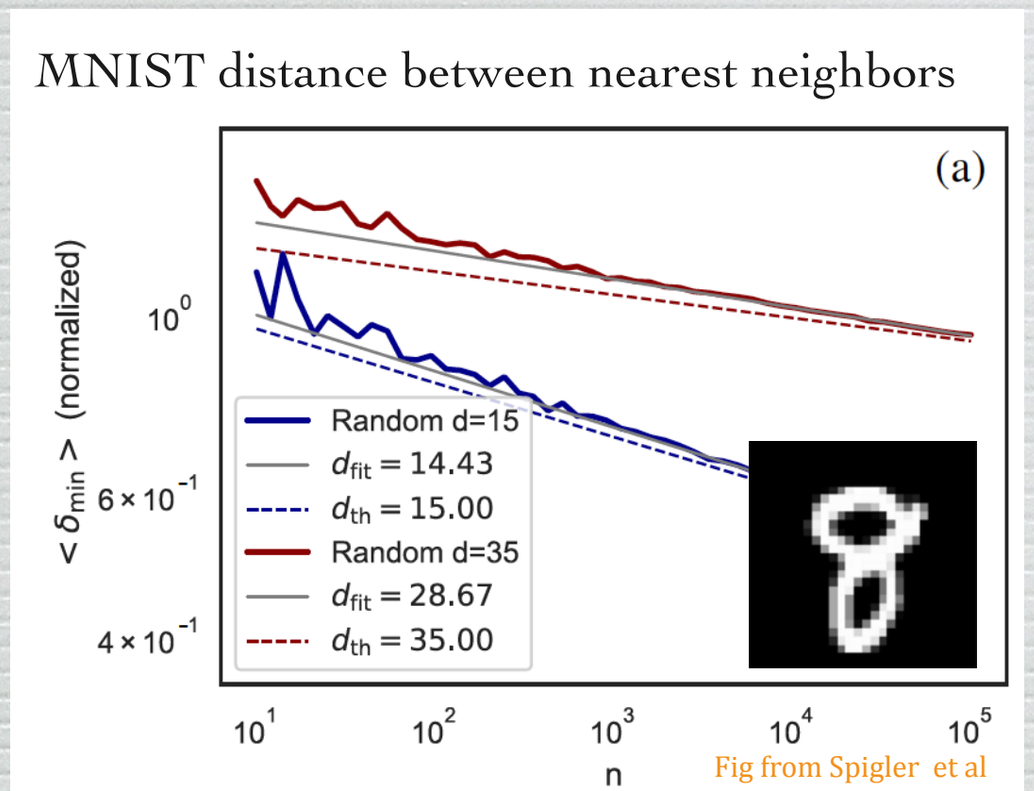


Fig from Spigler et al

'Low' effective dimension

Compositional/Combinatorial

Large range of scales

Multimodal

Structure is crucial for Machine Learning

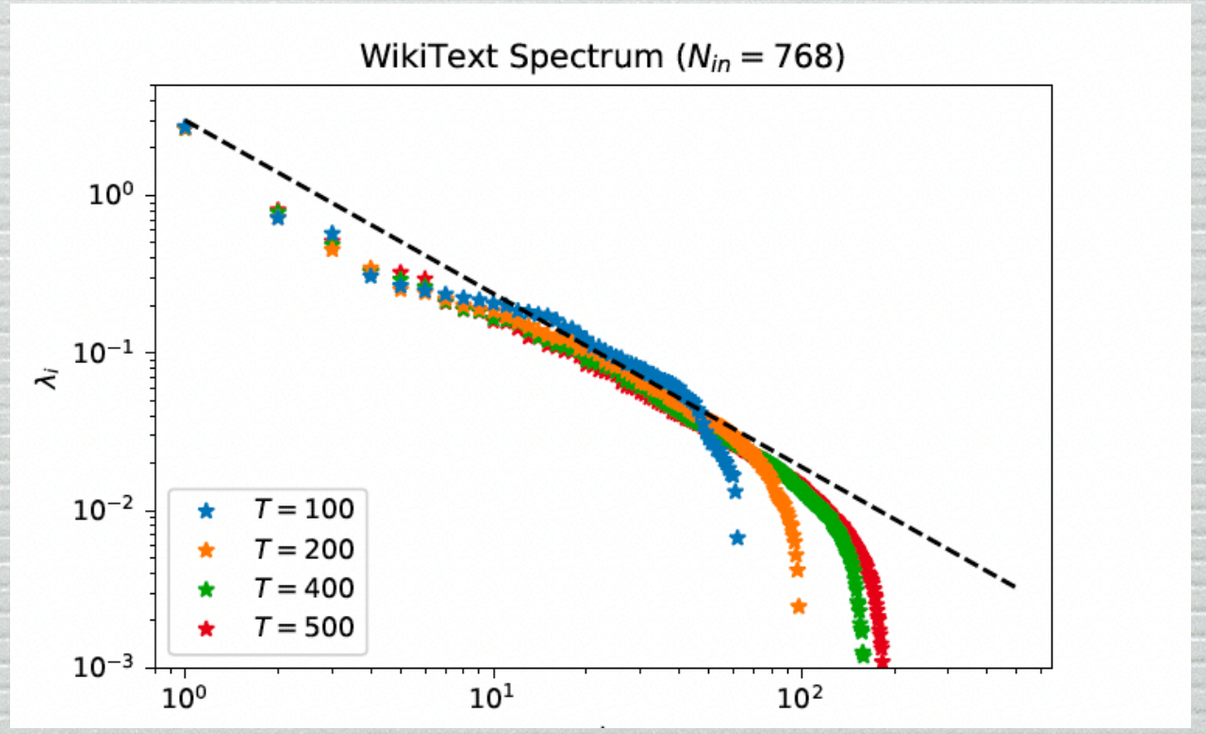


Fig from Maloney et al

A window to study Data Structure: Generative Models

A general scientific objective: Given many examples in $\mathbb{R}^d$ sampled from an unknown probability distribution P_0 , generate new samples.

A new technology: State of the art for generating images and videos.

- Principle

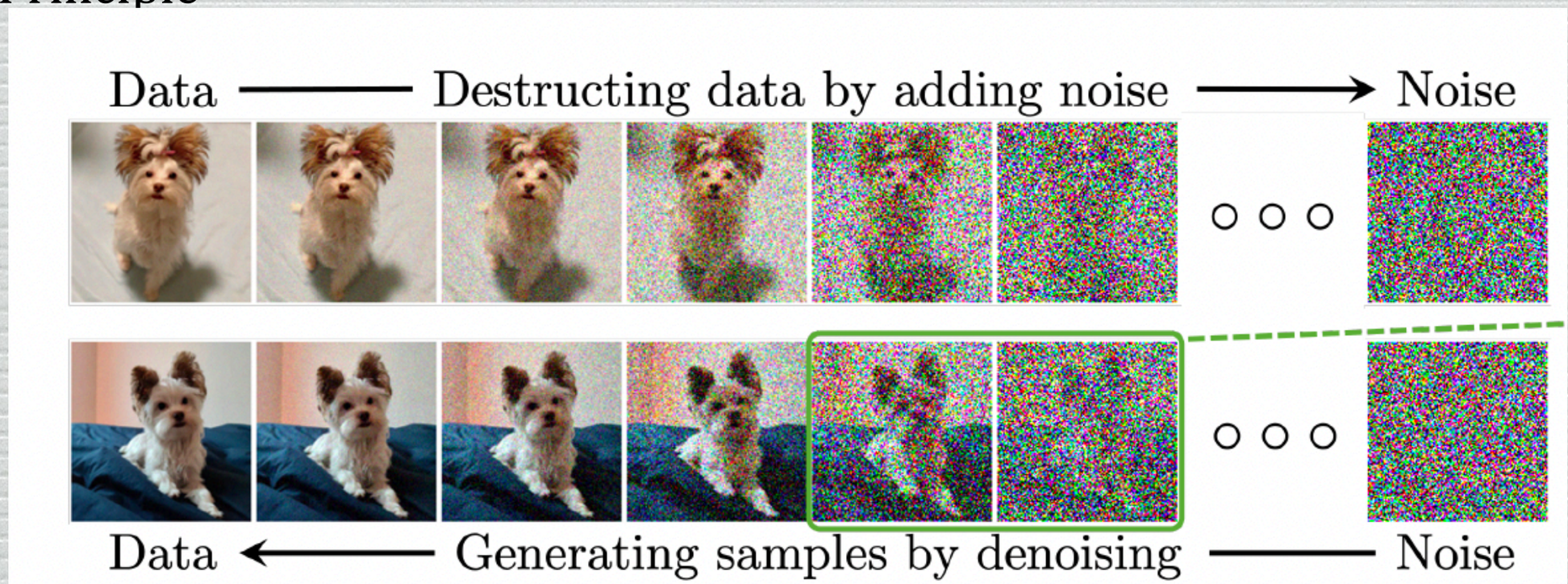


Fig from Yang et al.

Related to Flow-based generative models, stochastic interpolation (Albergo, Vanden-Eijnden), stochastic localization (Eldan, …),..

Fig from Maloney et al

Generative diffusion in a nutshell

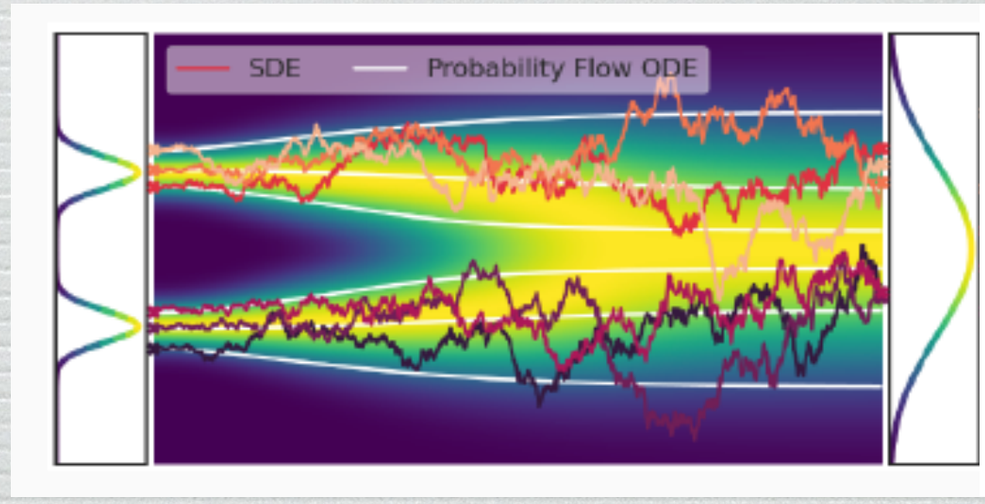
Initial distribution: $\vec{a} \in \mathbb{R}^N$, probability P_0

Start from $\vec{a}$ and add noise. Langevin equation for an Ornstein-Uhlenbeck process

$$\frac{d\vec{x}}{dt} = -\vec{x} + \vec{\eta}(t)$$

$$\langle \eta_i(t)\eta_j(t') \rangle = 2\delta_{ij}\delta(t-t')$$

P_0 —Forward→ P_{Gauss}



At long time $t = t_f \gg 1$

$$P_{t_f}(\vec{x}) \propto e^{-\vec{x}^2/2}$$

Generative diffusion in a nutshell

Initial distribution: $\vec{a} \in \mathbb{R}^N$, probability P_0

Start from $\vec{a}$ and add noise. Langevin equation for an Ornstein-Uhlenbeck process

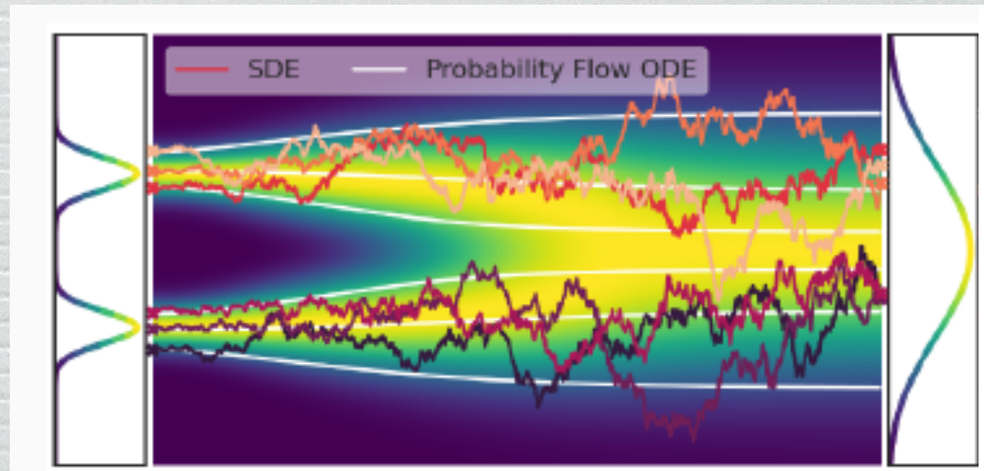
$$\frac{d\vec{x}}{dt} = -\vec{x} + \vec{\eta}(t)$$

$$\langle \eta_i(t)\eta_j(t') \rangle = 2\delta_{ij}\delta(t-t')$$

At long time $t = t_f \gg 1$

$$P_{t_f}(\vec{x}) \propto e^{-\vec{x}^2/2}$$

P_0 —Forward→ P_{Gauss}



P_0 ←Backward— P_{Gauss}

Generative diffusion in a nutshell

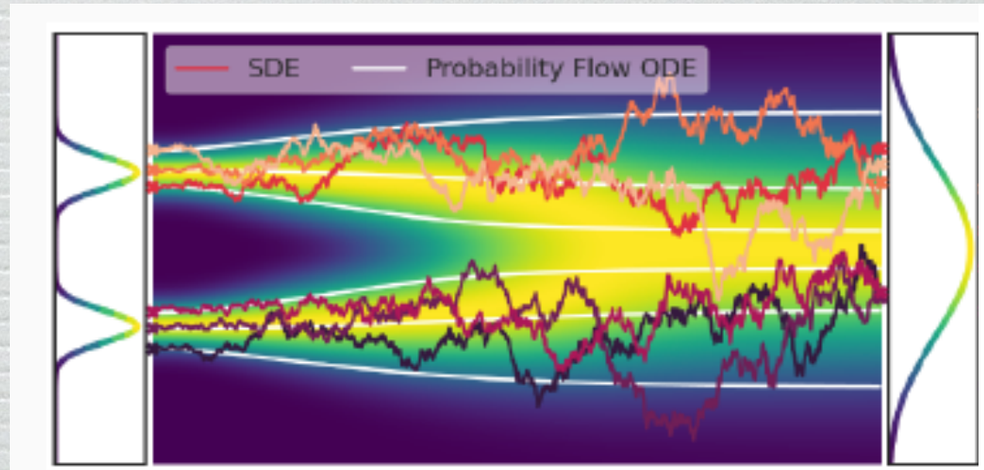
Initial distribution: $\vec{a} \in \mathbb{R}^N$, probability P_0

Start from $\vec{a}$ and add noise. Langevin equation for an Ornstein-Uhlenbeck process

$$\frac{d\vec{x}}{dt} = -\vec{x} + \vec{\eta}(t)$$

$$\langle \eta_i(t)\eta_j(t') \rangle = 2\delta_{ij}\delta(t-t')$$

P_0 —Forward→ P_{Gauss}



P_0 ←Backward— P_{Gauss}

At long time $t = t_f \gg 1$

$$P_{t_f}(\vec{x}) \propto e^{-\vec{x}^2/2}$$

Time-reversed Langevin equation transforms iid Gaussians in new data

Start from final time t_f and run backward, writing $t = t_f - t'$

$$\frac{d\vec{y}}{dt'} = \vec{y} + 2\vec{\mathcal{F}}(\vec{y}, t_f - t') + \vec{\eta}(t')$$

Score

$$\mathcal{F}_i(\vec{x}, t) = \frac{\partial \log P_t(\vec{x})}{\partial x_i}$$

The score function

Exact score

$$P_t(\vec{x}) = \int d\vec{a}\; P_0(\vec{a}) \frac{1}{\sqrt{2\pi\Delta_t}^N} \exp\left(-\frac{1}{2}\frac{(\vec{x}-\vec{a}e^{-t})^2}{\Delta_t}\right) = \int d\vec{a}\; P_t(\vec{a},\vec{x})$$

$$\mathcal{F}_i(\vec{x},t) = \frac{\partial \log P_t(\vec{x})}{\partial x_i}$$

$$\vec{\mathcal{F}}(\vec{x},t) = \frac{-\vec{x} + \langle \vec{a} \rangle_{|\vec{x}}\, e^{-t}}{\Delta_t}$$

Denoiser

The score function

Exact score

$$P_t(\vec{x}) = \int d\vec{a}\; P_0(\vec{a}) \frac{1}{\sqrt{2\pi\Delta_t}^N} \exp\left(-\frac{1}{2}\frac{(\vec{x}-\vec{a}e^{-t})^2}{\Delta_t}\right) = \int d\vec{a}\; P_t(\vec{a},\vec{x})$$

$$\mathcal{F}_i(\vec{x},t) = \frac{\partial \log P_t(\vec{x})}{\partial x_i}$$

$$\vec{\mathcal{F}}(\vec{x},t) = \frac{-\vec{x} + \langle \vec{a} \rangle_{|\vec{x}}\, e^{-t}}{\Delta_t}$$

Denoiser

Exact score from empirical distribution

$$P_t^{emp}(\vec{x},\vec{a}) = \frac{1}{P}\frac{1}{\sqrt{2\pi\Delta_t}^N}\sum_\mu \exp\left(-\frac{1}{2\Delta_t}|\vec{x}-\vec{a}^\mu e^{-t}|^2\right)$$

The score function

Exact score

$$P_t(\vec{x}) = \int d\vec{a}\; P_0(\vec{a}) \frac{1}{\sqrt{2\pi\Delta_t}^N} \exp\left(-\frac{1}{2}\frac{(\vec{x}-\vec{a}e^{-t})^2}{\Delta_t}\right) = \int d\vec{a}\; P_t(\vec{a},\vec{x})$$

$$\mathcal{F}_i(\vec{x},t) = \frac{\partial \log P_t(\vec{x})}{\partial x_i}$$

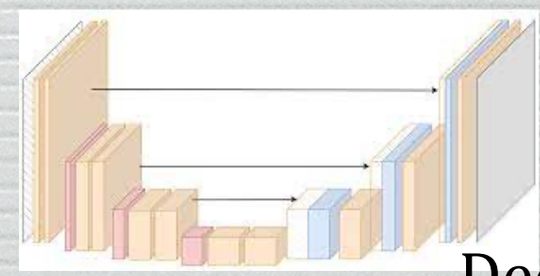
$$\vec{\mathcal{F}}(\vec{x},t) = \frac{-\vec{x} + \langle \vec{a} \rangle_{|\vec{x}}\, e^{-t}}{\Delta_t}$$

Denoiser

Exact score from empirical distribution

$$P_t^{emp}(\vec{x},\vec{a}) = \frac{1}{P}\frac{1}{\sqrt{2\pi\Delta_t}^N}\sum_{\mu} \exp\left(-\frac{1}{2\Delta_t}|\vec{x}-\vec{a}^{\mu}e^{-t}|^2\right)$$

Regularized score, parameters θ : $\vec{\mathcal{S}}^{\theta}(\vec{x})$



Deepnet (UNet)

The score function

Exact score

$$P_t(\vec{x}) = \int d\vec{a}\; P_0(\vec{a}) \frac{1}{\sqrt{2\pi\Delta_t}^N} \exp\left(-\frac{1}{2}\frac{(\vec{x}-\vec{a}e^{-t})^2}{\Delta_t}\right) = \int d\vec{a}\; P_t(\vec{a},\vec{x})$$

$$\mathcal{F}_i(\vec{x},t) = \frac{\partial \log P_t(\vec{x})}{\partial x_i}$$

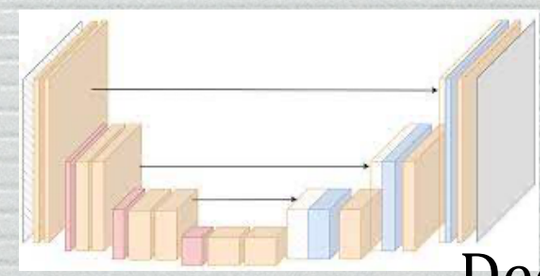
$$\vec{\mathcal{F}}(\vec{x},t) = \frac{-\vec{x} + \langle \vec{a} \rangle_{|\vec{x}}\, e^{-t}}{\Delta_t}$$

Denoiser

Exact score from empirical distribution

$$P_t^{emp}(\vec{x},\vec{a}) = \frac{1}{P}\frac{1}{\sqrt{2\pi\Delta_t}^N}\sum_\mu \exp\left(-\frac{1}{2\Delta_t}|\vec{x}-\vec{a}^\mu e^{-t}|^2\right)$$

Regularized score, parameters θ : $\vec{\mathcal{S}}^\theta(\vec{x})$



Deepnet (UNet)

Regression problem: find θ by minimizing a « loss »

$$\mathcal{L}(\theta) = \int d\vec{x} P_t(\vec{x}) \left\| \vec{\mathcal{S}}^\theta(\vec{x}) - \vec{\mathcal{F}}(\vec{x},t) \right\|^2$$

The score function

Exact score

$$P_t(\vec{x}) = \int d\vec{a}\; P_0(\vec{a}) \frac{1}{\sqrt{2\pi\Delta_t}^N} \exp\left(-\frac{1}{2}\frac{(\vec{x}-\vec{a}e^{-t})^2}{\Delta_t}\right) = \int d\vec{a}\; P_t(\vec{a},\vec{x})$$

$$\mathcal{F}_i(\vec{x},t) = \frac{\partial \log P_t(\vec{x})}{\partial x_i}$$

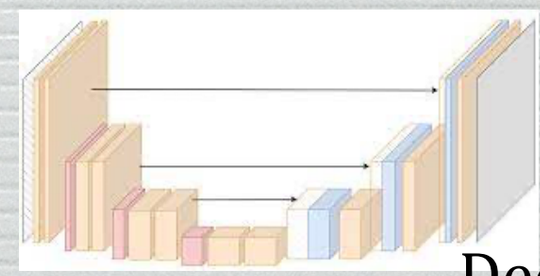
$$\vec{\mathcal{F}}(\vec{x},t) = \frac{-\vec{x} + \langle \vec{a} \rangle_{|\vec{x}}\, e^{-t}}{\Delta_t}$$

Denoiser

Exact score from empirical distribution

$$P_t^{emp}(\vec{x},\vec{a}) = \frac{1}{P}\frac{1}{\sqrt{2\pi\Delta_t}^N}\sum_\mu \exp\left(-\frac{1}{2\Delta_t}|\vec{x}-\vec{a}^\mu e^{-t}|^2\right)$$

Regularized score, parameters θ : $\vec{\mathcal{S}}^\theta(\vec{x})$



Deepnet (UNet)

Regression problem: find θ by minimizing a « loss »

$$\mathcal{L}(\theta) = \int d\vec{x} P_t(\vec{x}) \left\| \vec{\mathcal{S}}^\theta(\vec{x}) - \vec{\mathcal{F}}(\vec{x},t) \right\|^2$$

Rewrite after integration by parts (Tweedie):

$$\mathcal{L}(\theta) = \mathbb{E}_{x,a} \left\| \vec{\mathcal{S}}^\theta(\vec{x}) + \frac{\vec{x}-\vec{a}e^{-t}}{\Delta_t} \right\|^2 + C$$

$$\mathcal{L}(\theta(t)) = \sum_{x_\mu(t),a_\mu} \left\| \vec{S}^{\theta(t)}(\vec{x}_\mu(t)) + \frac{\vec{x}_\mu(t) - \vec{a}_\mu e^{-t}}{\Delta_t} \right\|^2$$

Questions

What happens when data is in large dimensions ? Curse of dimensionality?

Can generative diffusion work in presence of phase transitions? At what stage of the diffusion process does one split the phase space?

How many data does one need in order to get a good diffusion model ?

When does collapse take place?

What is the role of the approximation class ?

What is the role of learning dynamics ?

Statistical physics study

Analysis of diffusion models in the large dimension- large number of data limit:

Simple models: Curie Weiss, Gaussian Mixtures, solvable

→ **Identify General Mechanisms**

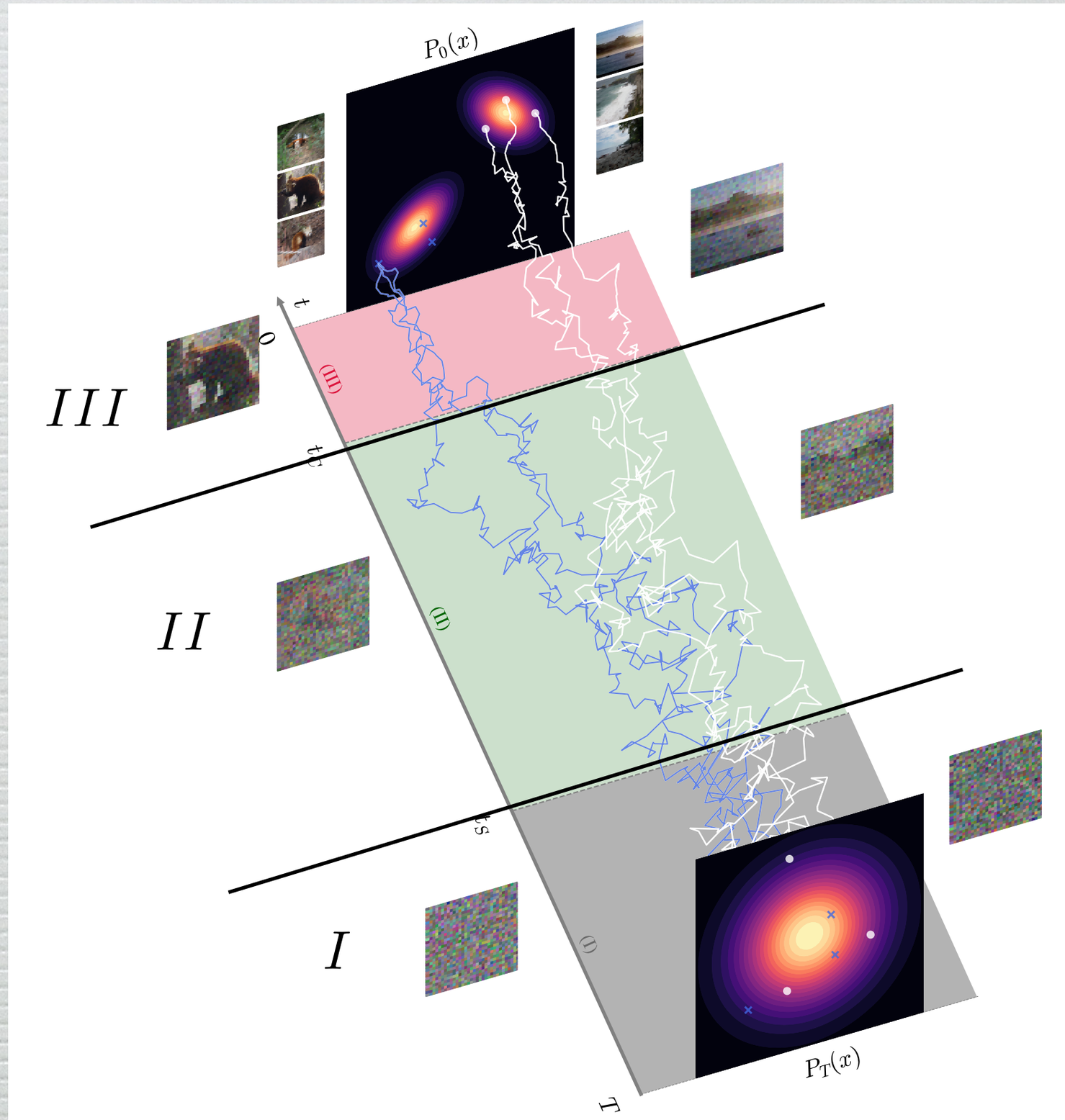
→ **Tests on realistic database**

Three time regimes and two dynamical transitions

- **Three different time-regimes, with dynamical transitions: speciations and collapse**

- **Analytical prediction for t_s (spectral estimate) and t_c (entropy estimate), computable on any database**

- **Collapse takes place early on if the score function is too good, and the number of data is not exponential**



Take home

Statistical physics of disordered systems has been a major evolution of statistical physics in the last 50 years, opening the way to applications in many other branches of science.

Machine learning in modern AI is a series of remarkable technical breakthroughs which still wait for a theory. A major question is that of emergence of macroscopic « order parameters » related to information processing. Statistical physics can play a role in this understanding.

When using statistical physics ideas in the study of machine learning, one must be able to face a new challenge, the one of highly structured data (manifolds, attention, etc.)

The End