Statistical physics and statistical inference

Marc Mézard

Ecole normale supérieure
PSL University

Aix en Provence, Le 5 juillet 2019
Festival de théorie
Conférence René Pellat

What is inference?

Statistics

Infer a hidden rule, or hidden variables, from data.

Restricted sense : find parameters of a probability distribution

Urn with 10.000 balls. Draw 100, find 70 white balls and 30 black

Best guess for the composition of the urn? How reliable? Probability that it has 6000 white- 4000 black?

If only black and white balls , with fraction x of white, probability to pick-up 70 white balls is $\binom{100}{70} x^{70}(1-x)^{30}$

Log likelihood of x : $L(x) = 70 \log x + 30 \log(1-x)$

Maximum at $x^* = .7$ Probability of .6 : $e^{L(.6)-L(.7)}$

Bayesian inference

Unknown parameters	x	Prior	$P(x)$
Measurements	y	Likelihood	$P(y\|x)$

Posterior

$$P(x|y) = \frac{P(y|x)P(x)}{P(y)}$$

Bayesian inference

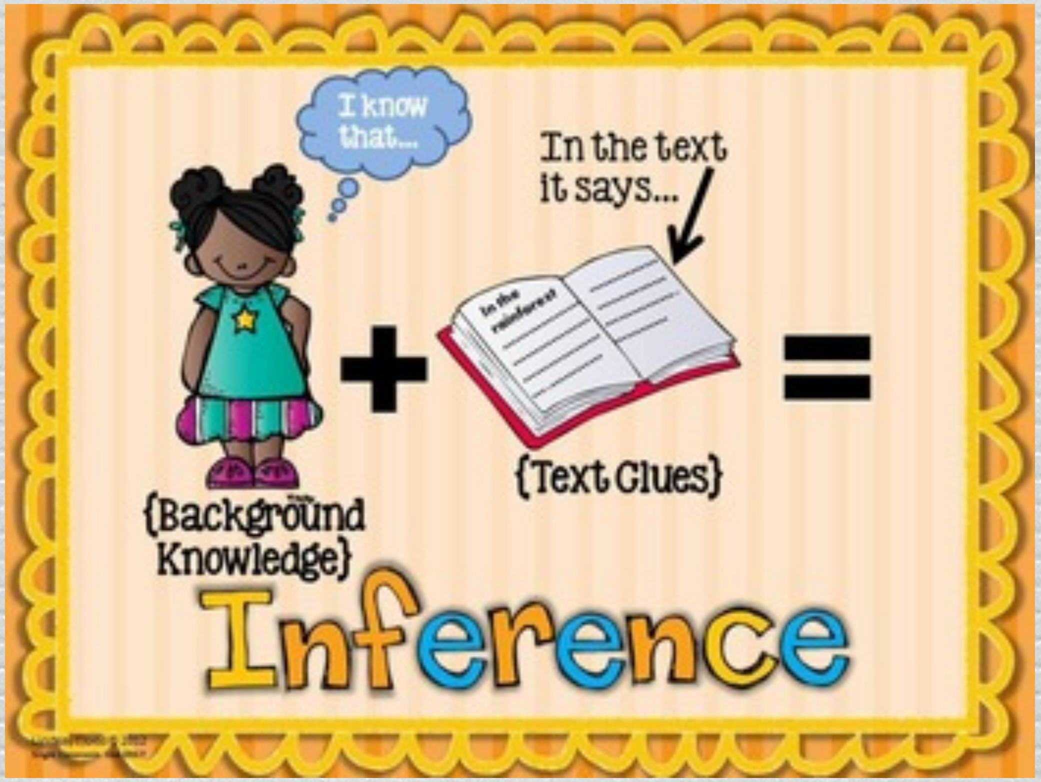
Unknown parameters	x	Prior	$P(x)$
Measurements	y	Likelihood	$P(y\|x)$

Posterior $$P(x|y) = \frac{P(y|x)P(x)}{P(y)}$$

Bayesian inference

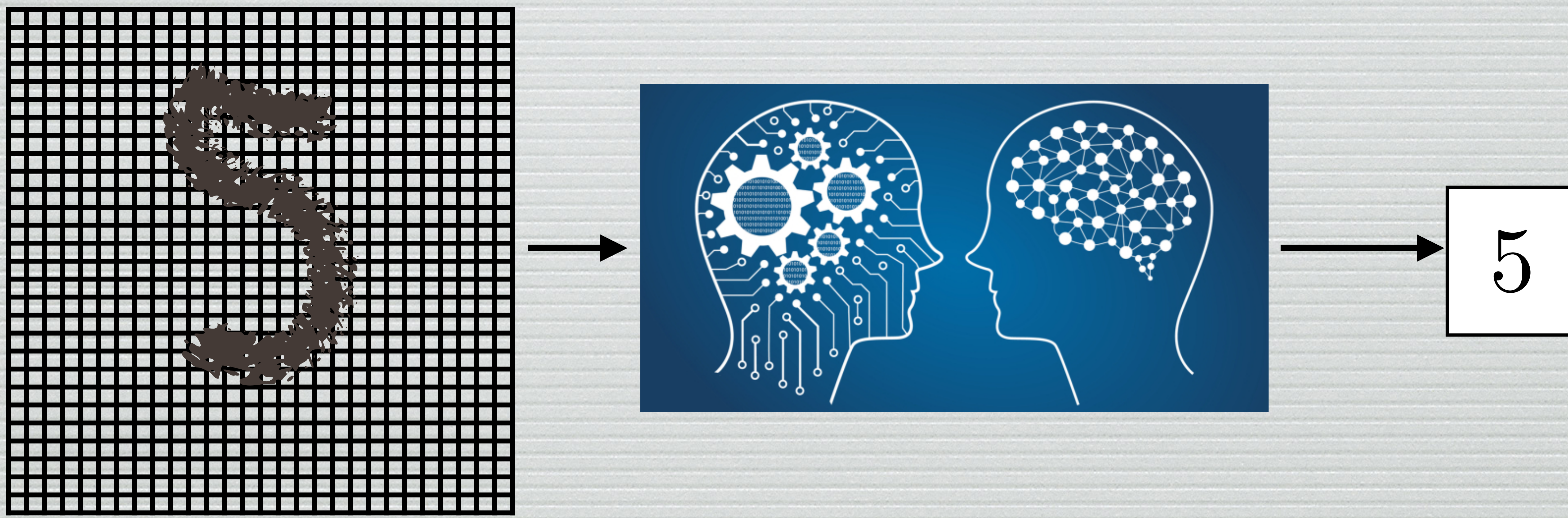
Unknown parameters	x	Prior	$P(x)$
Measurements	y	Likelihood	$P(y\|x)$

Posterior $$P(x|y) = \frac{P(y|x)P(x)}{P(y)}$$



What is inference?

Artificial intelligence,
machine learning



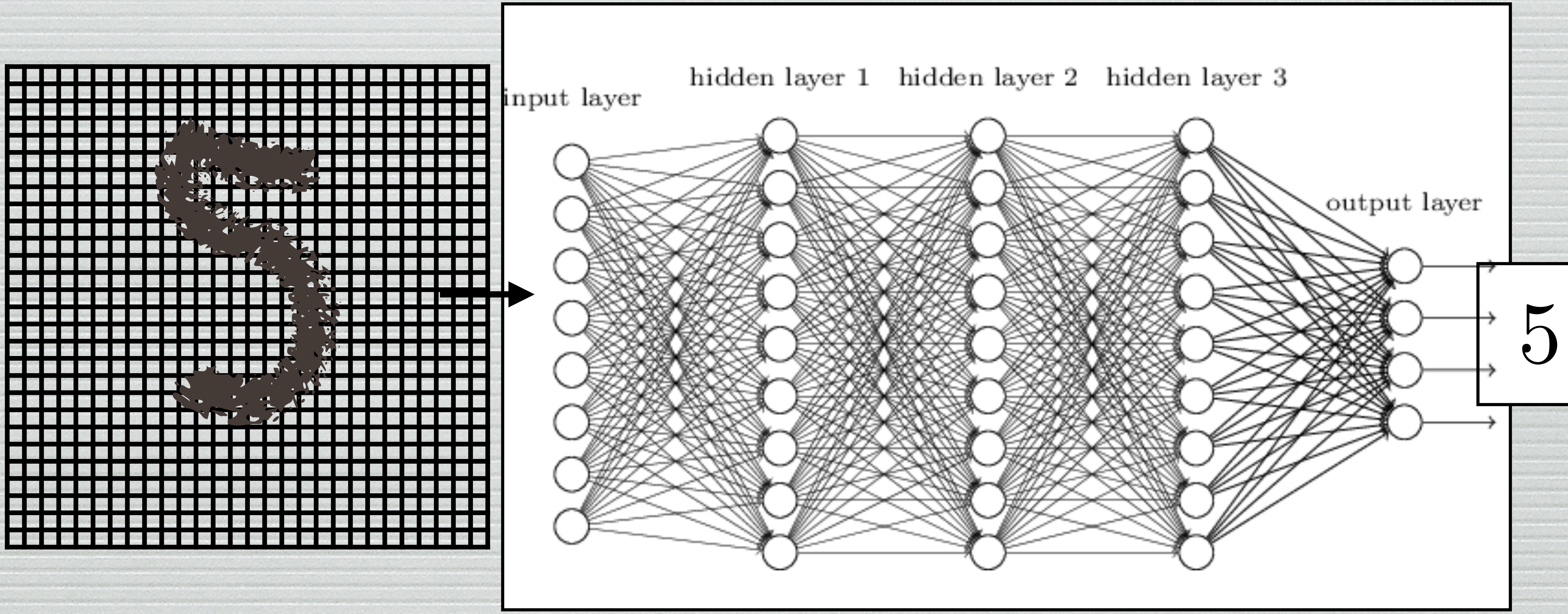
Find a machine that reads handwritten digits…

…inferring its parameters from examples

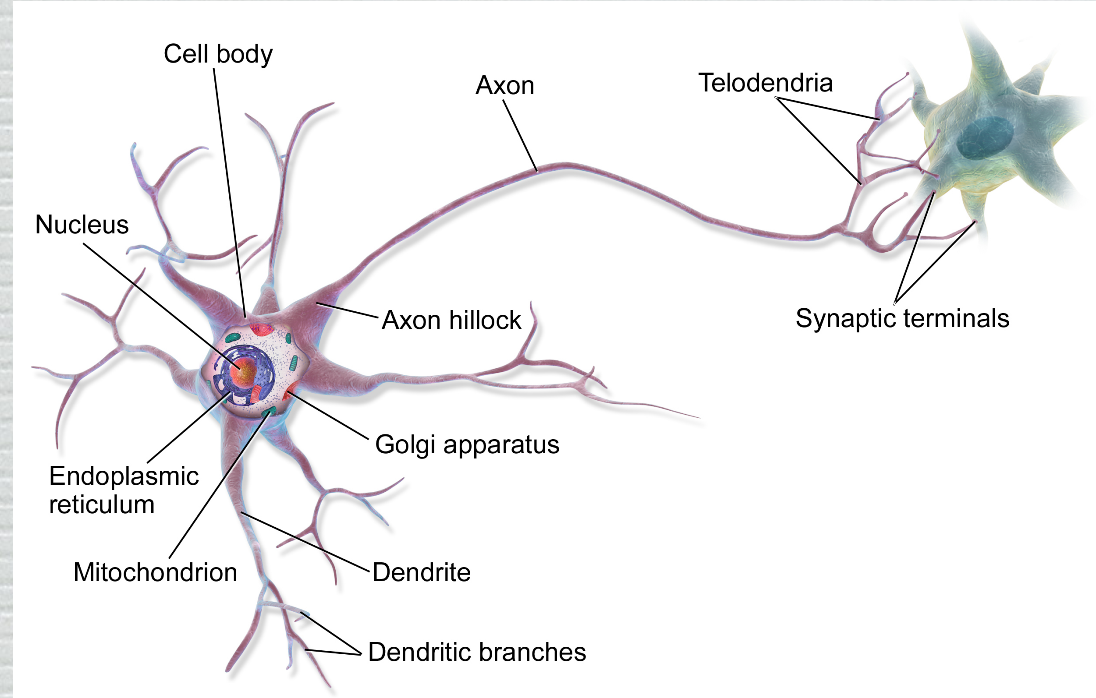
MNIST database : 70,000 images of digits, segmented, 28 $\times$ 28 pixels each, greyscale. Known output (supervised learning)

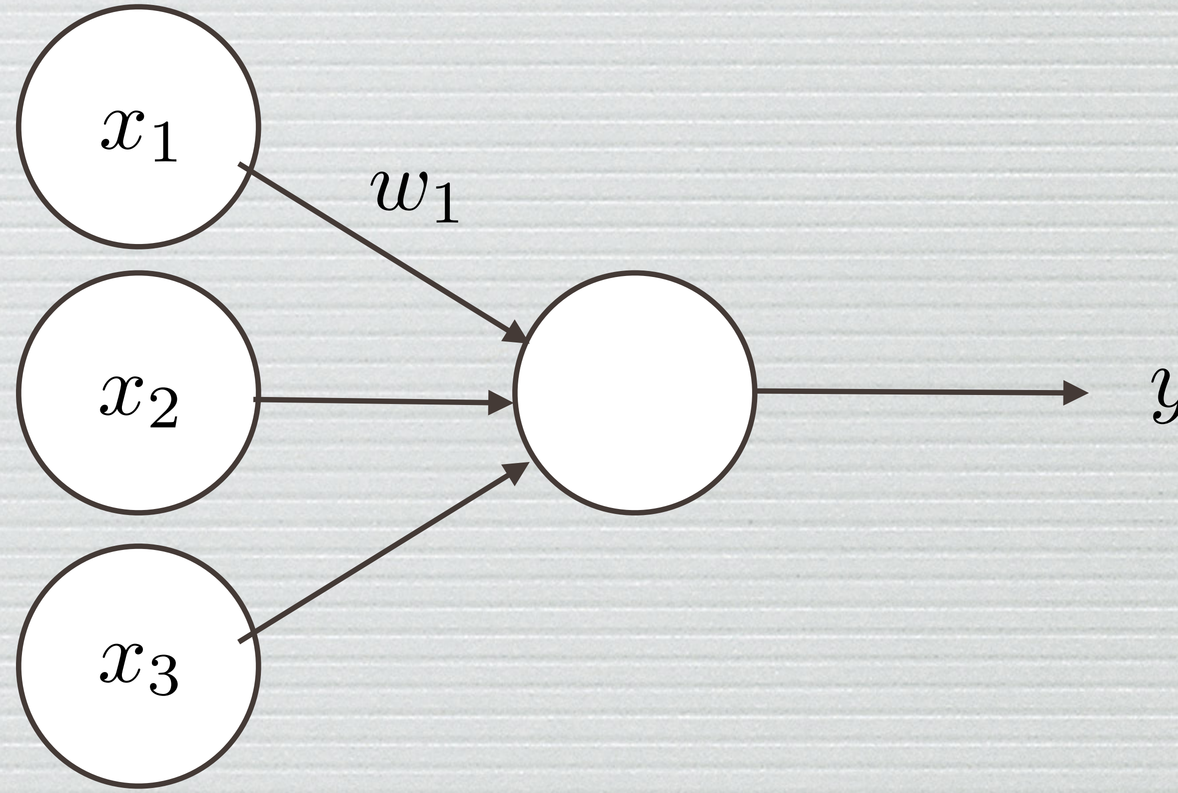
What is inference?

Artificial intelligence,
machine learning



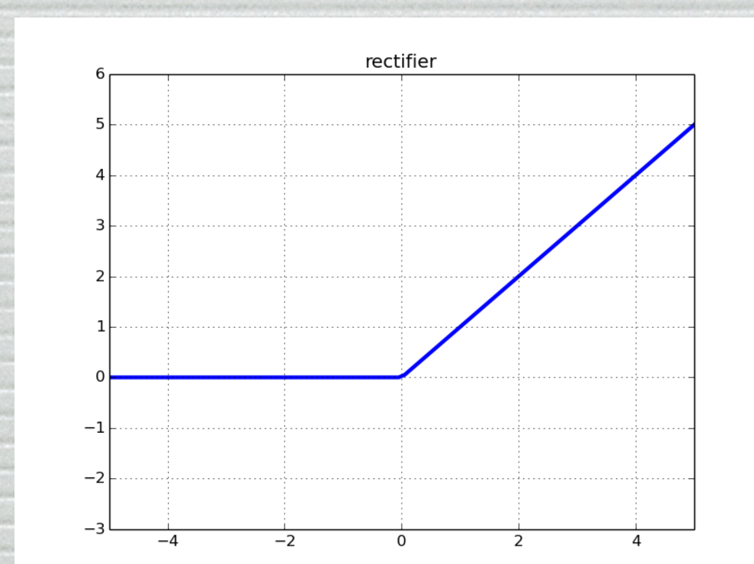
« Neural network » : artificial neurons

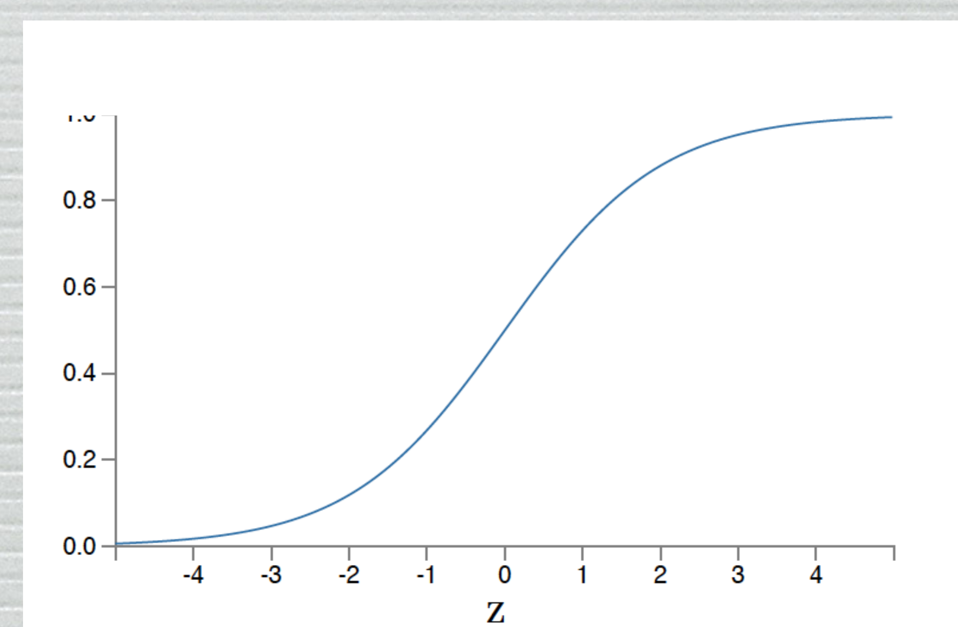




$$y = f(w_0 + w_1 x_1 + w_2 x_2 + w_3 x_3)$$

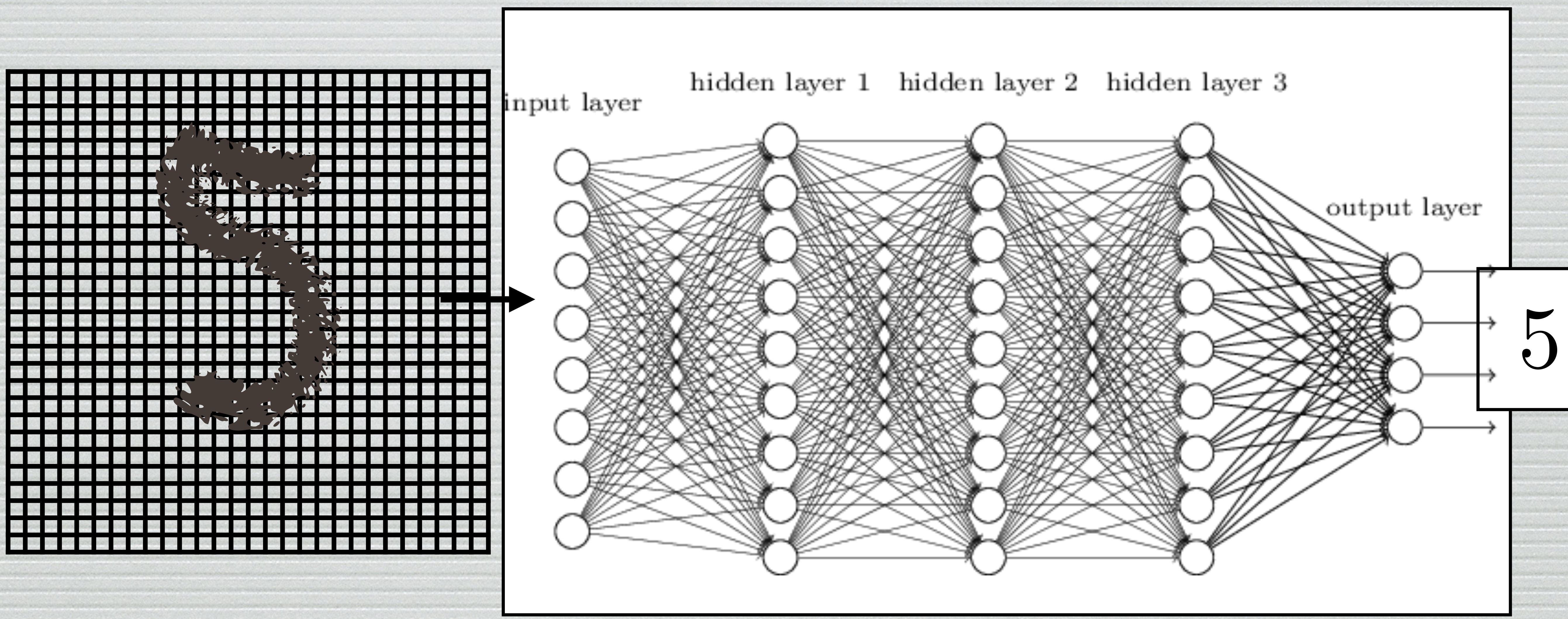
Formal neural network





What is inference?

Artificial intelligence,
machine learning



Machine with hundreds of thousands of parameters, trained on very large data base: infer the parameters from data (supervised learning)

Statistical inference

Challenge = rules with **many hidden parameters.** eg : machine learning with large machine and big data, decoding in commonication,…

$x = (x_1, \ldots, x_N) \qquad N \gg 1$

Many measurements $\qquad y = (y_1, \ldots, y_M) \qquad M \gg 1$

Measure of the amount of data $\qquad \alpha = M/N$

- **Algorithms**
- **Prediction on the quality of inference**, on the performance of the algorithms, on the type of situations where they can be applied

Bayesian inference with many unknown and many measurements

Unknown parameters $x = (x_1, \ldots, x_N)$ Prior $P^0(x)$

Measurements $y = (y_1, \ldots, y_M)$ $P(y|x)$

Bayesian inference $P(x|y) \propto P(y|x) P^0(x)$

Often (but not necessarily):

Independent measurements $P(y|x) = \prod_\mu P_\mu(y_\mu|x)$

Factorized prior $P^0(x) = \prod_i P_i^0(x_i)$

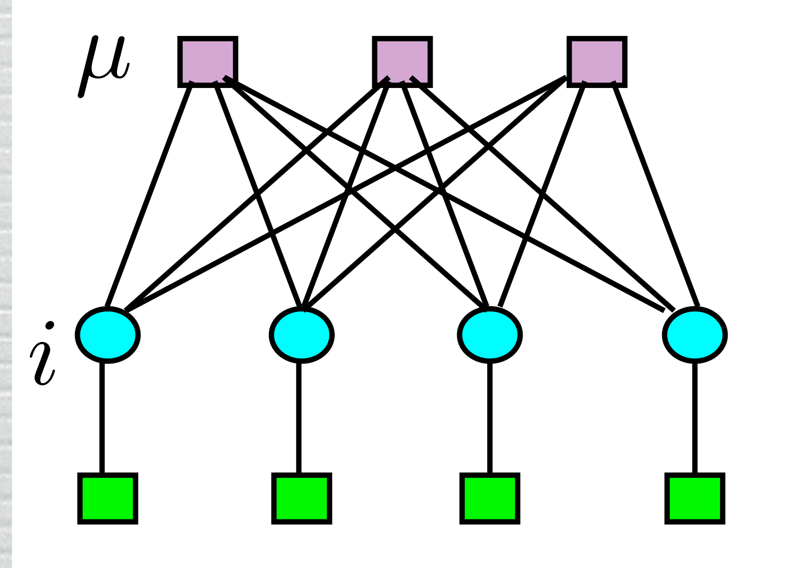
Posterior $P(x) = \frac{1}{Z(y)} \left(\prod_i P_i^0(x_i) \right) \exp\left[-\sum_\mu E_\mu(x, y_\mu) \right]$

$$E_\mu(x, y_\mu) = -\log P_\mu(y_\mu|x)$$

Bayesian inference with many unknown and many measurements

$$P(x) = \frac{1}{Z(y)} \left(\prod_i P_i^0(x_i) \right) \exp\left[-\sum_\mu E_\mu(x, y_\mu) \right]$$

$$E_\mu(x, y_\mu) = -\log P_\mu(y_\mu | x)$$



Statistical mechanics.

✦Discrete or continuous variables x_i

✦Interactions through $e^{-E_\mu(x, y_\mu)}$ can be

- pairwise : $E_\mu = J_\mu x_{i(\mu)} x_{j(\mu)}$
- multibody

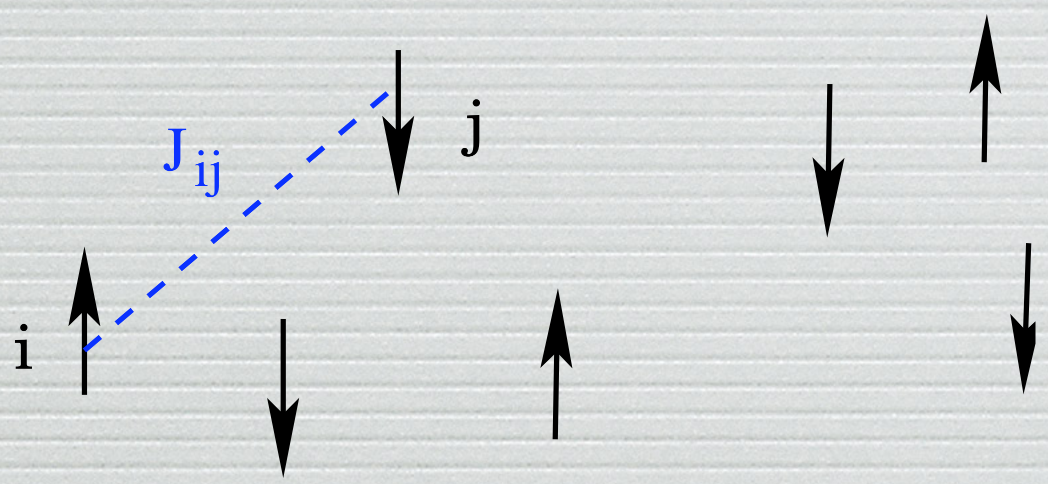
✦Disordered system, ensemble

✦Thermodynamic limit, phase transitions

« Spin glass »

Spin glasses

- Disordered magnetic systems   e.g.: CuMn



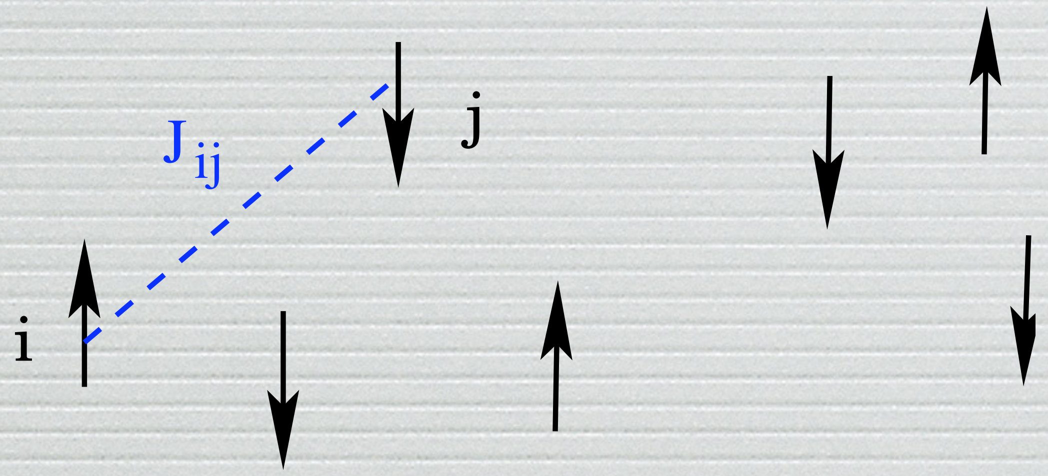
$$s_i = \pm 1$$

$$E = -\sum_{i,j} J_{ij} s_i s_j$$

$$P(s_1, \ldots, s_N) = \frac{1}{Z} e^{-E/T}$$

Spin glasses

- Disordered magnetic systems e.g.: CuMn



$$s_i = \pm 1$$

$$E = -\sum_{i,j} J_{ij} s_i s_j$$

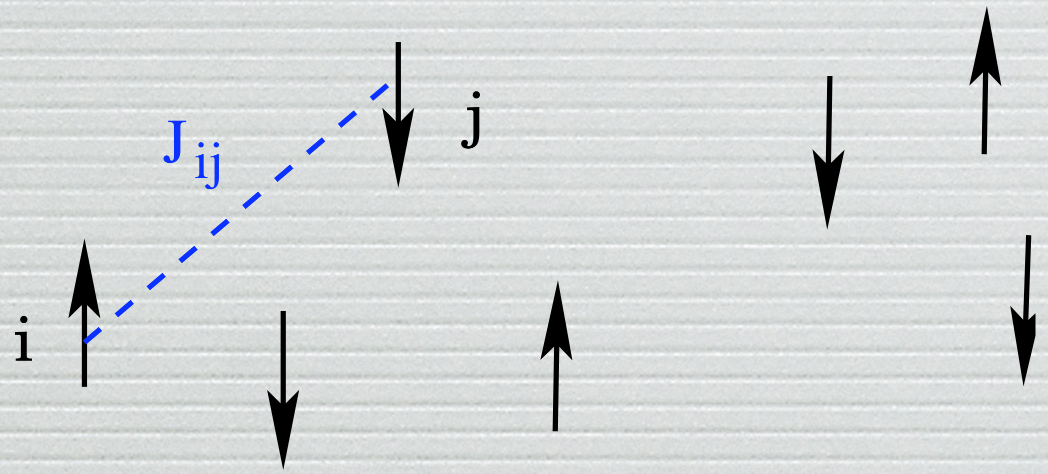
$$P(s_1, \ldots, s_N) = \frac{1}{Z} e^{-E/T}$$

➡ Each spin 'sees' a different local field

Phase transition with many states: spin glasses

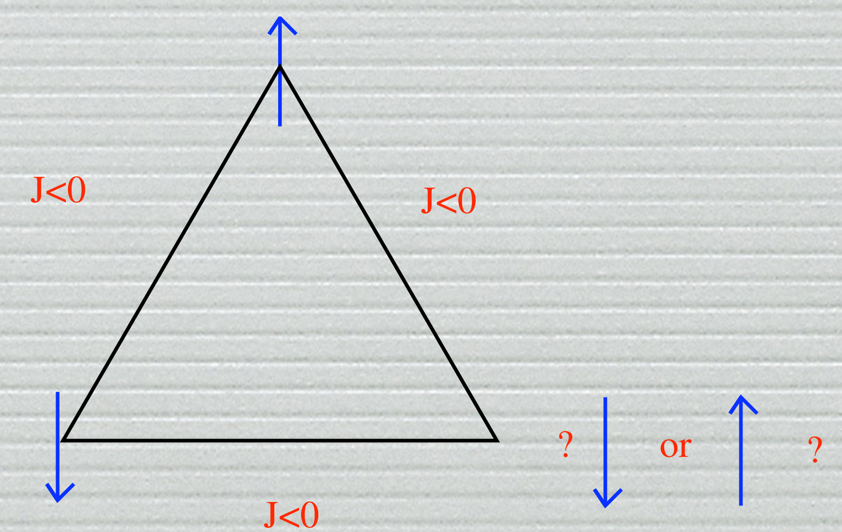
- Many atoms, microscopic interactions are known, "disordered systems"

e.g.: CuMn



➡ Each spin 'sees' a different local field

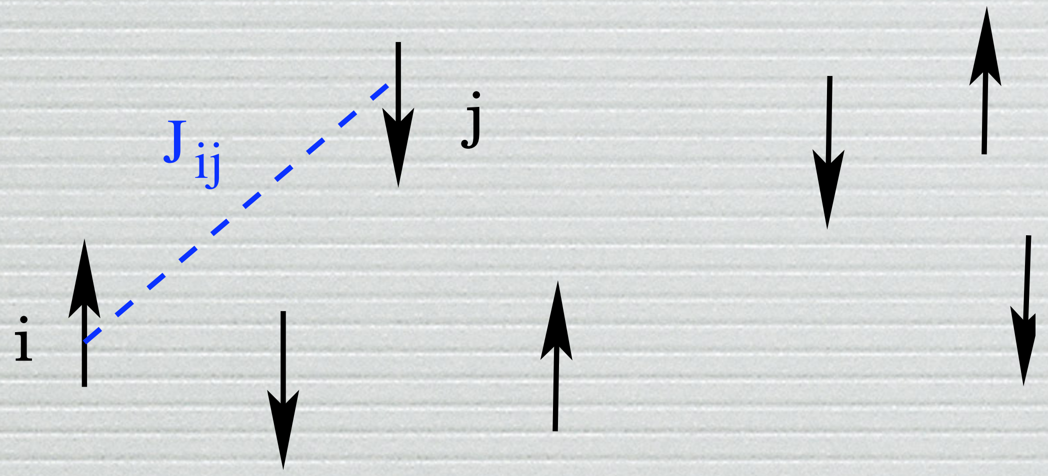
➡ Low temperature: frustration



Phase transition with many states: spin glasses

- Many atoms, microscopic interactions are known, "disordered systems"

e.g.: CuMn

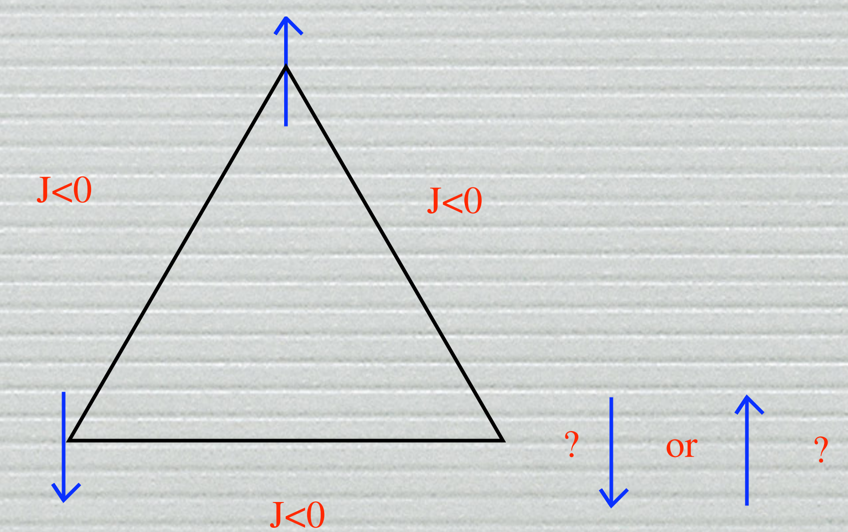


➡ Each spin 'sees' a different local field

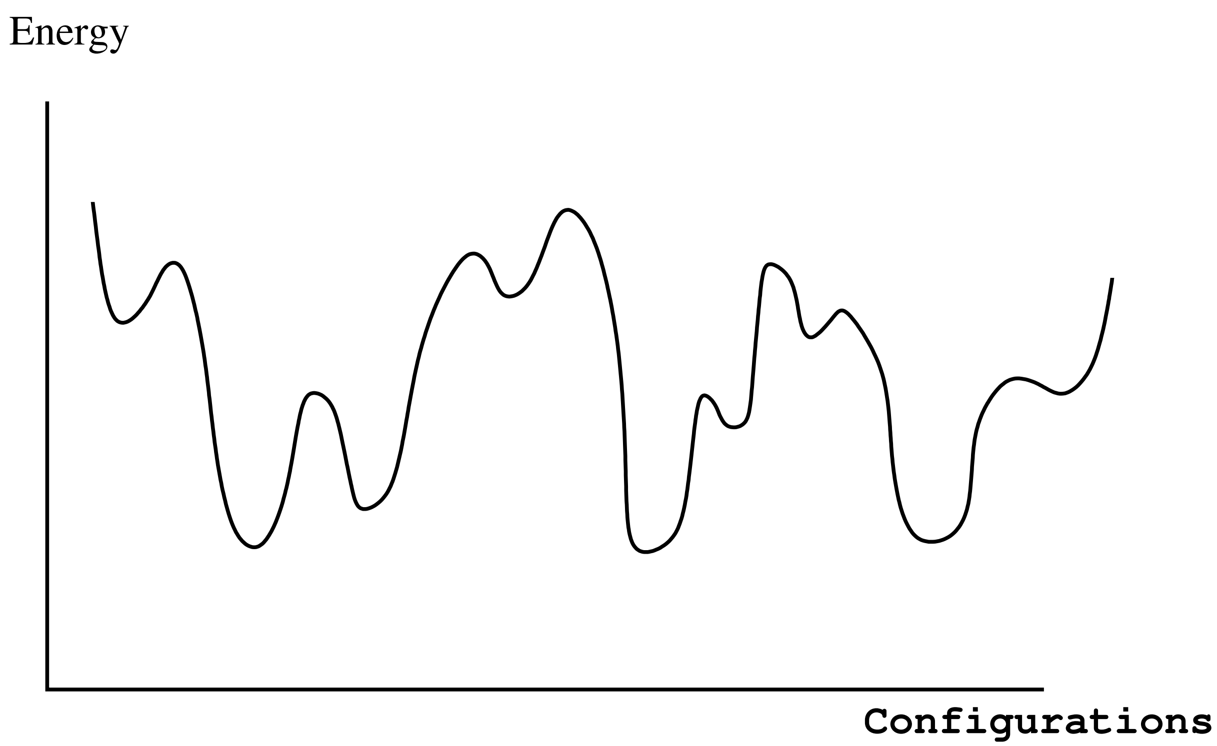
➡ Low temperature: frustration

➡ Spins freeze in random directions

➡ Difficult to find min. of E



Phase transition with many states: spin glasses

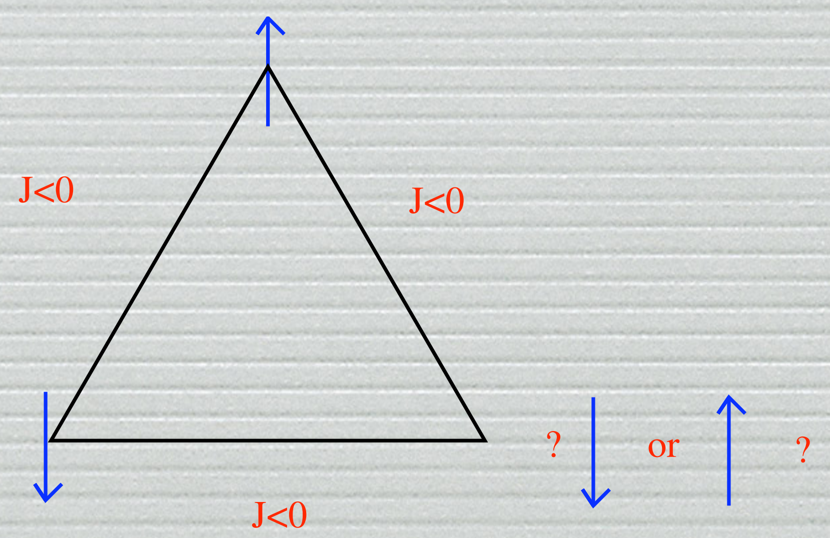


Many quasi-ground states unrelated by symmetries, many metastable states

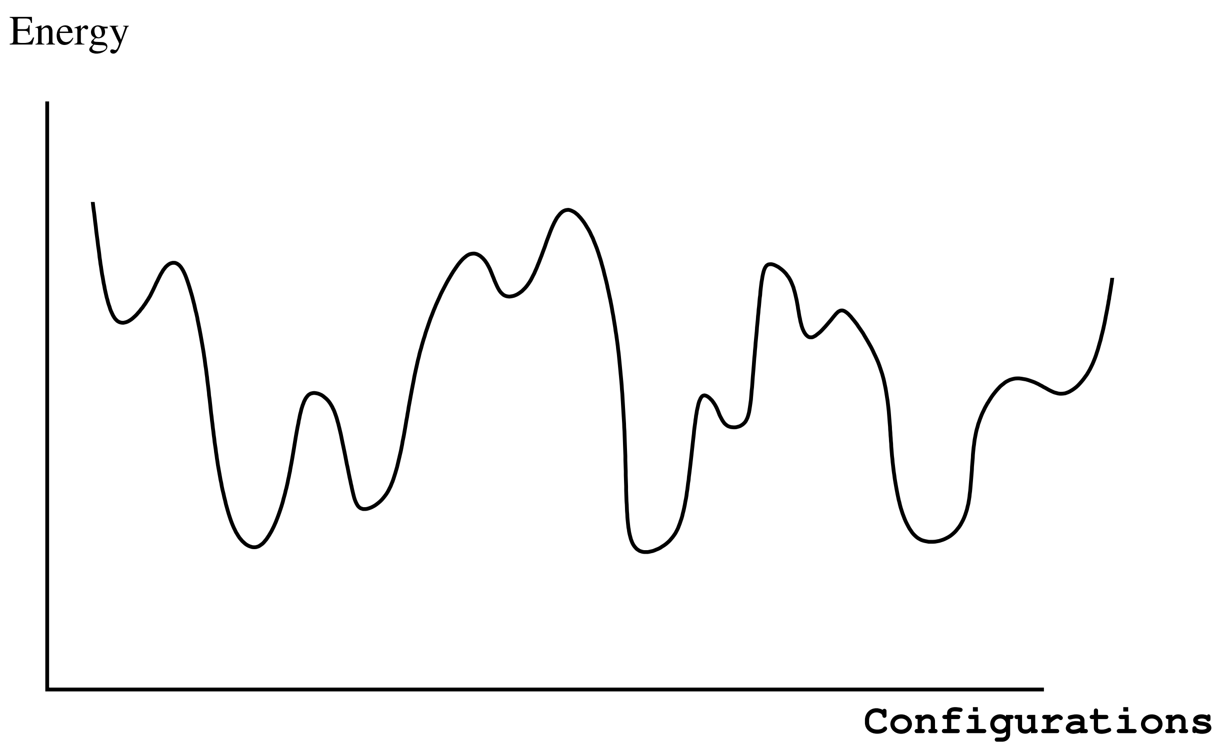
Slow dynamics, aging

Spin glass

- Each spin 'sees' a different local field
- Low temperature: frustration
- Spins freeze in random directions
- Difficult to find min. of E



Phase transition with many states: spin glasses

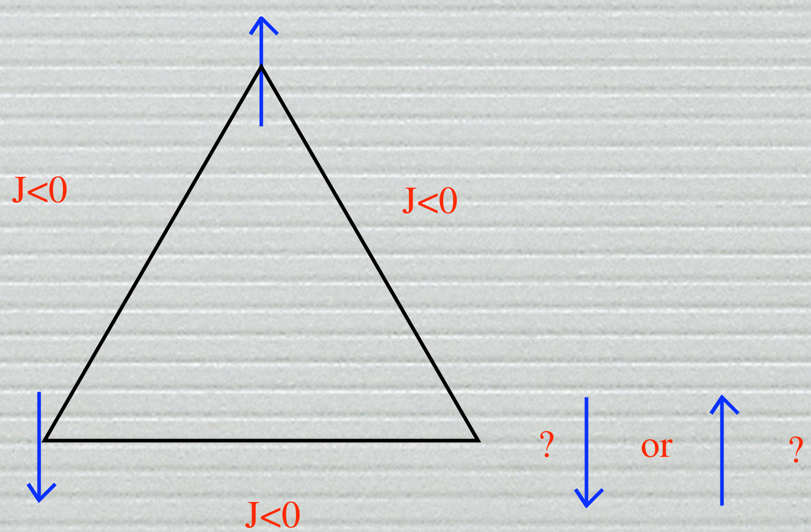


Spin glass

Many quasi-ground states unrelated by symmetries, many metastable states

Slow dynamics, aging

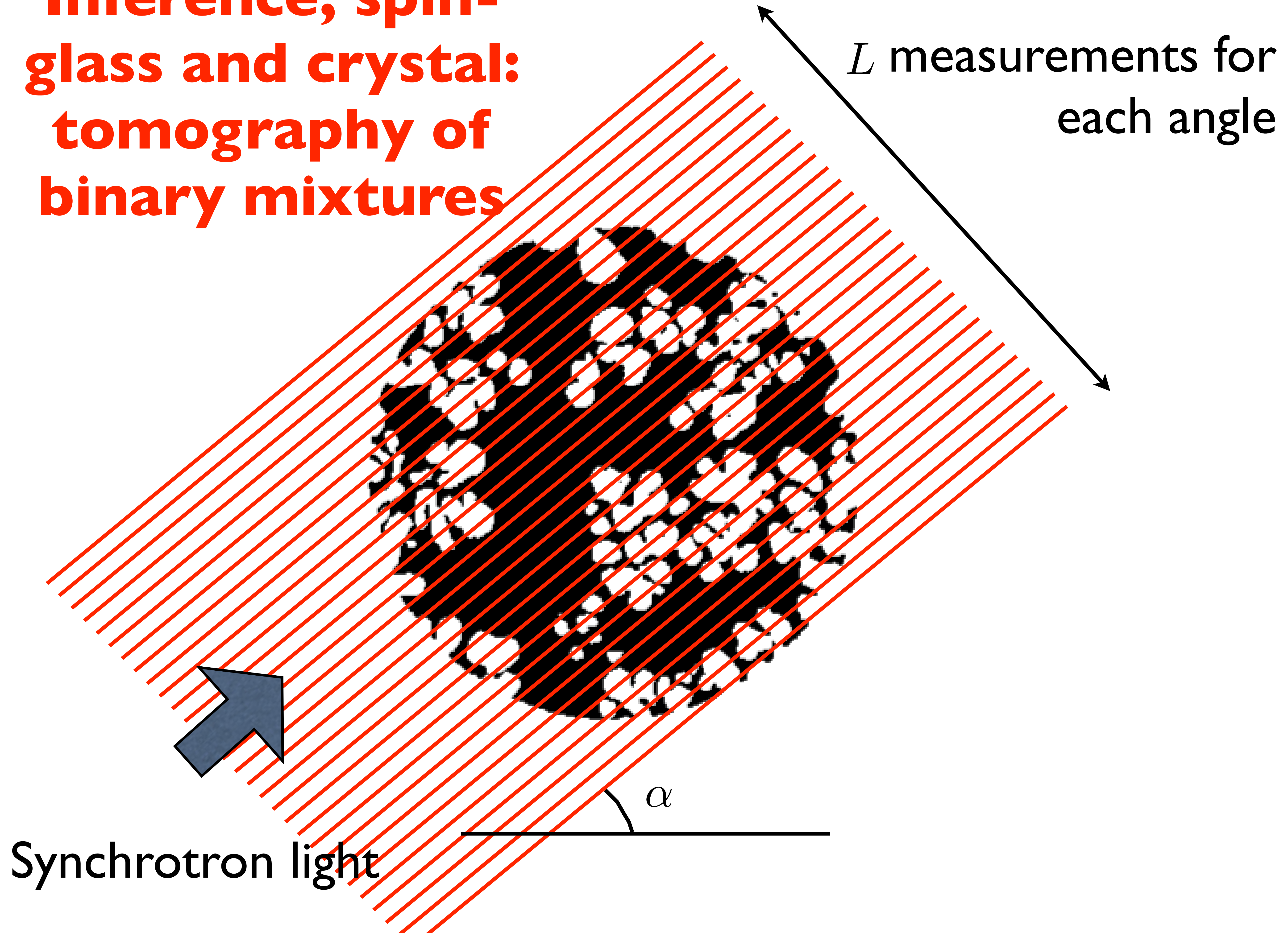
- Each spin 'sees' a different local field
- Low temperature: frustration
- Spins freeze in random directions
- Difficult to find min. of E



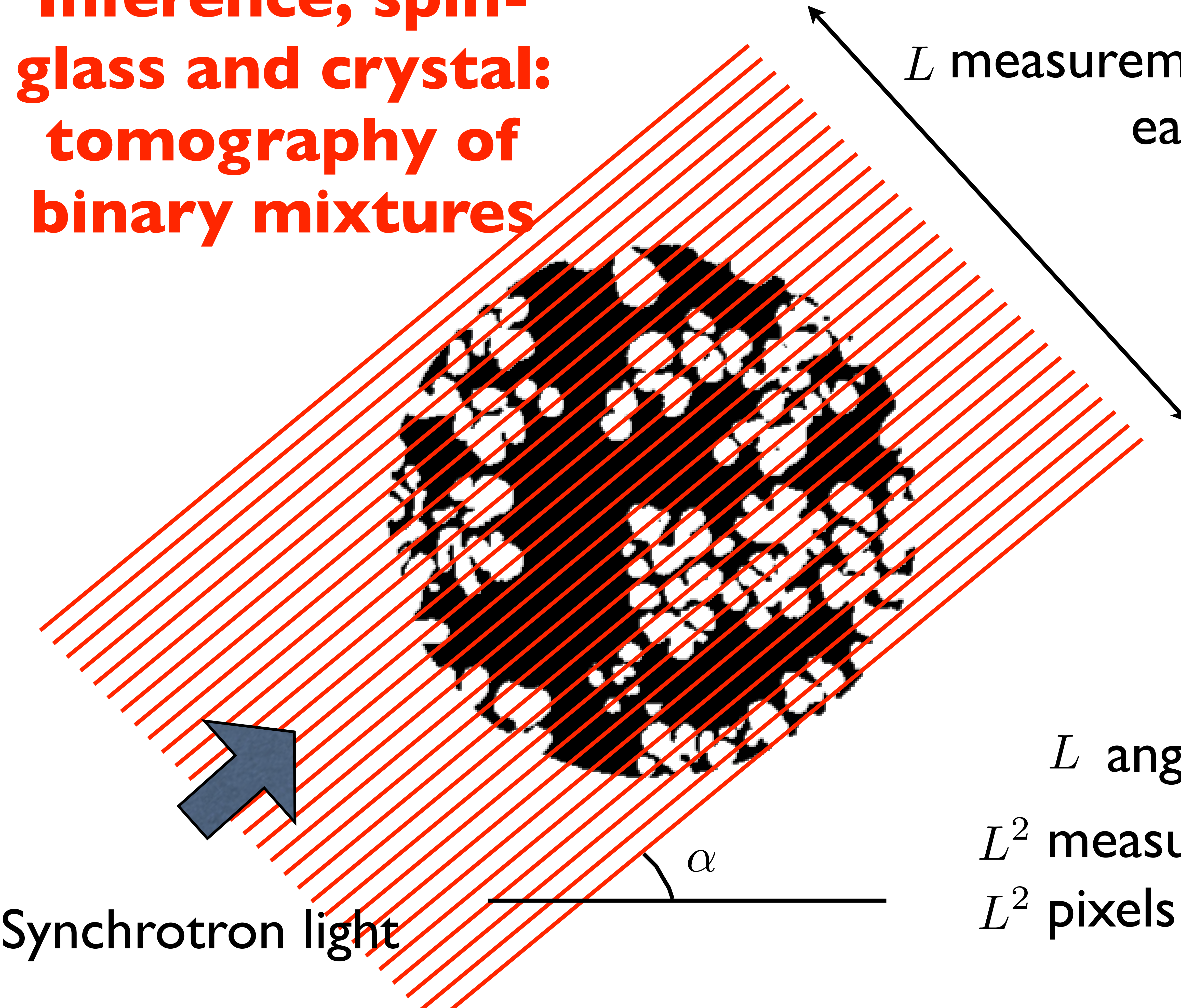
Useless, but thousands of papers...

Inference, spin-glass and crystal: tomography of binary mixtures

Inference, spin-glass and crystal: tomography of binary mixtures



Inference, spin-glass and crystal: tomography of binary mixtures



L angles:

L^2 measurements

L^2 pixels

Tomography of binary mixtures

L angles:

L^2 measurements

L^2 pixels

sample size L

pixel size 1

domain size ξ

If the size of domains is $\gg$ pixel: possible to reconstruct with $\ll L^2$ measurements

$\xi \gg 1$

Tomography of binary mixtures

sample size L

pixel size 1

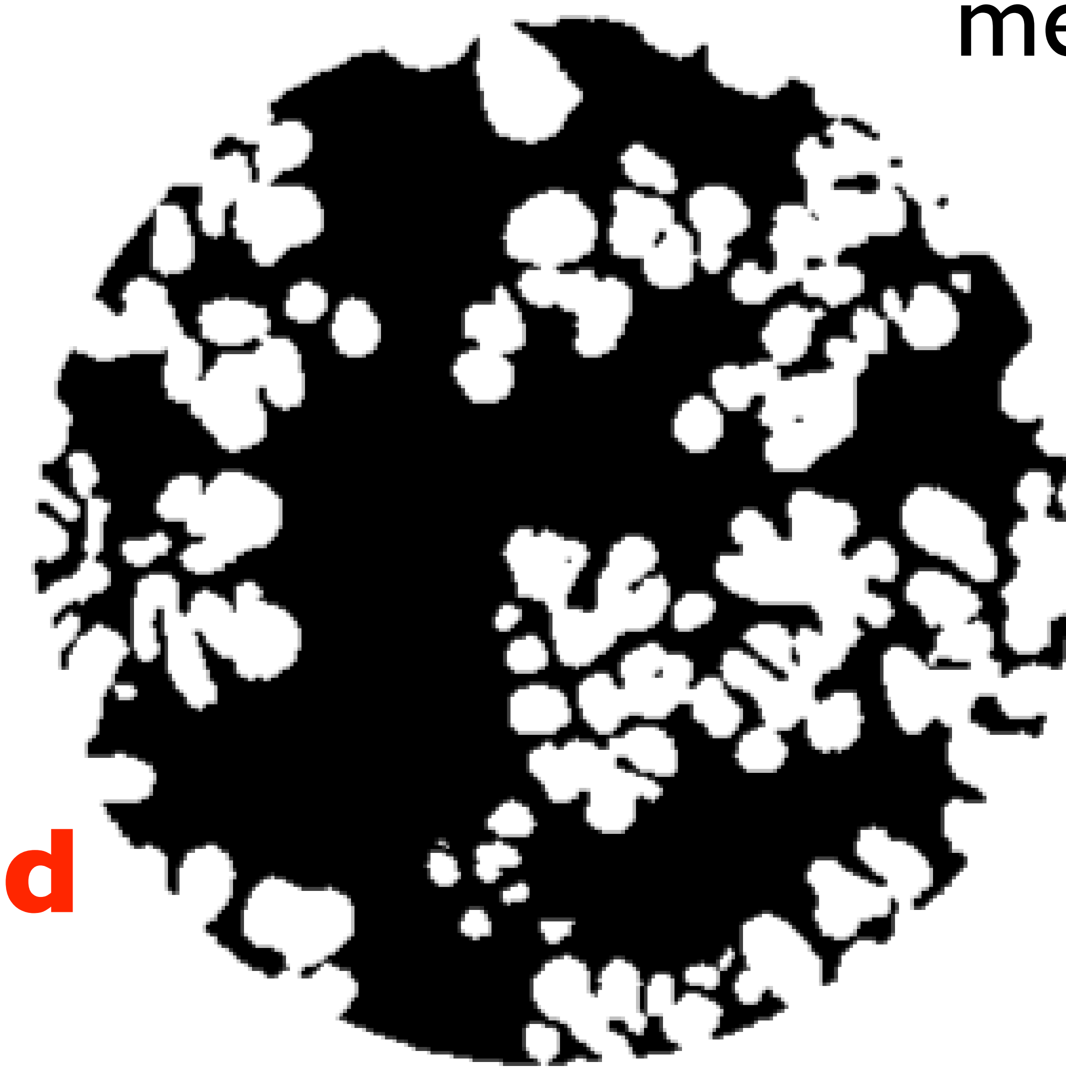
domain size ξ

If the size of domains is $\gg$ pixel: possible to reconstruct with $\ll L^2$ measurements

$\xi \gg 1$

Tomography of binary mixtures

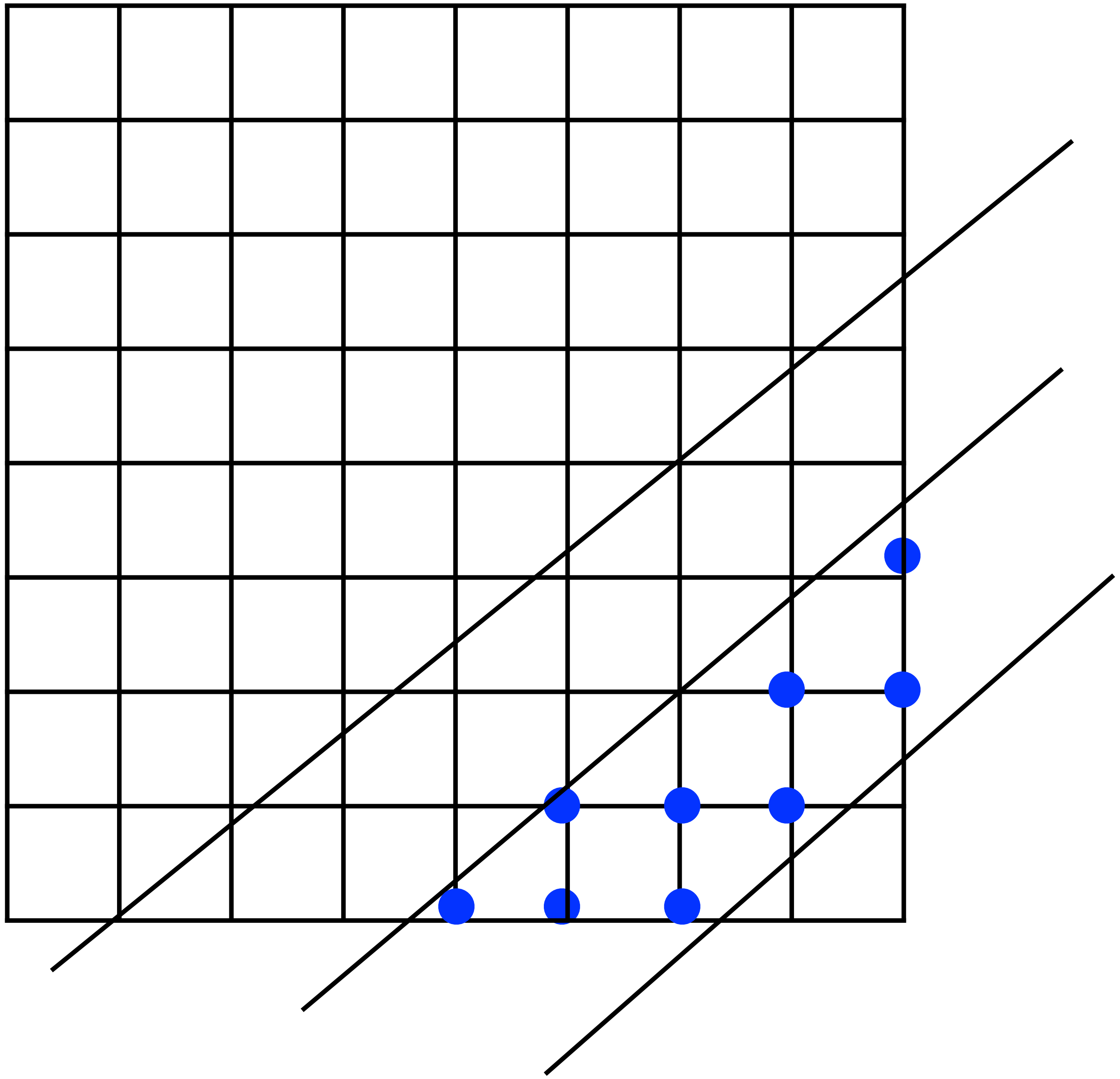
This picture, digitalized on 1000×1000 grid, can be reconstructed fom measurements with 16 angles



Gouillart et al., Inverse problems 2013

Compressed sensing

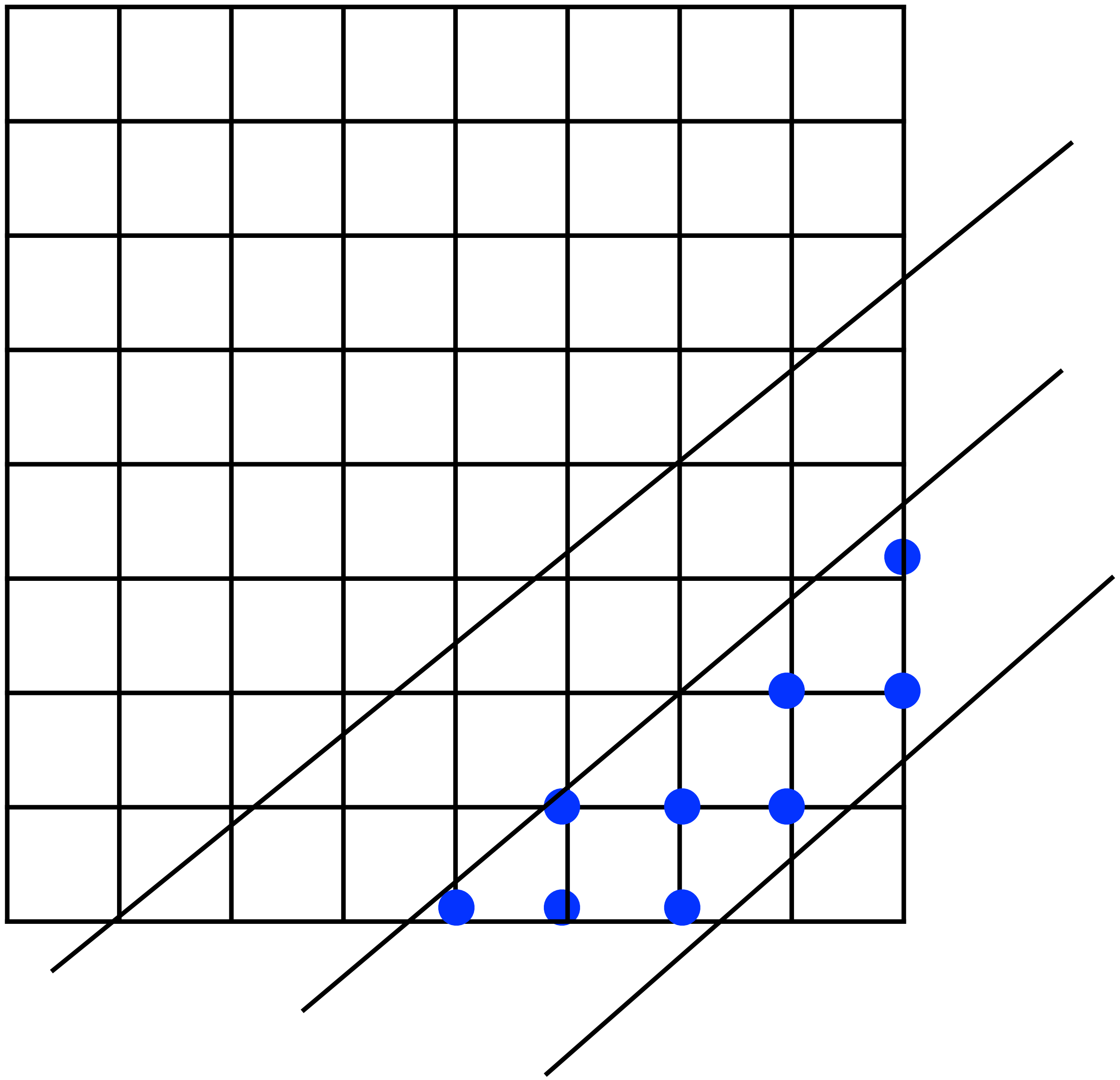
If the size of domains is $\gg$ pixel: possible to reconstruct with $\ll L^2$ measurements



μ

$$y_\mu = \sum_{i \in \partial\mu} s_i$$

Prior knowledge on $\{s_i\}$: neighboring pixels more likely to be equal



$$\mu \qquad y_\mu = \sum_{i \in \partial\mu} s_i$$

Prior knowledge on $\{s_i\}$: neighboring pixels more likely to be equal

$$P(S) = \prod_{ij \in \text{grid}} e^{J s_i s_j} \prod_{\mu} \delta\left(y_\mu, \sum_{i \in \partial\mu} s_i \right)$$

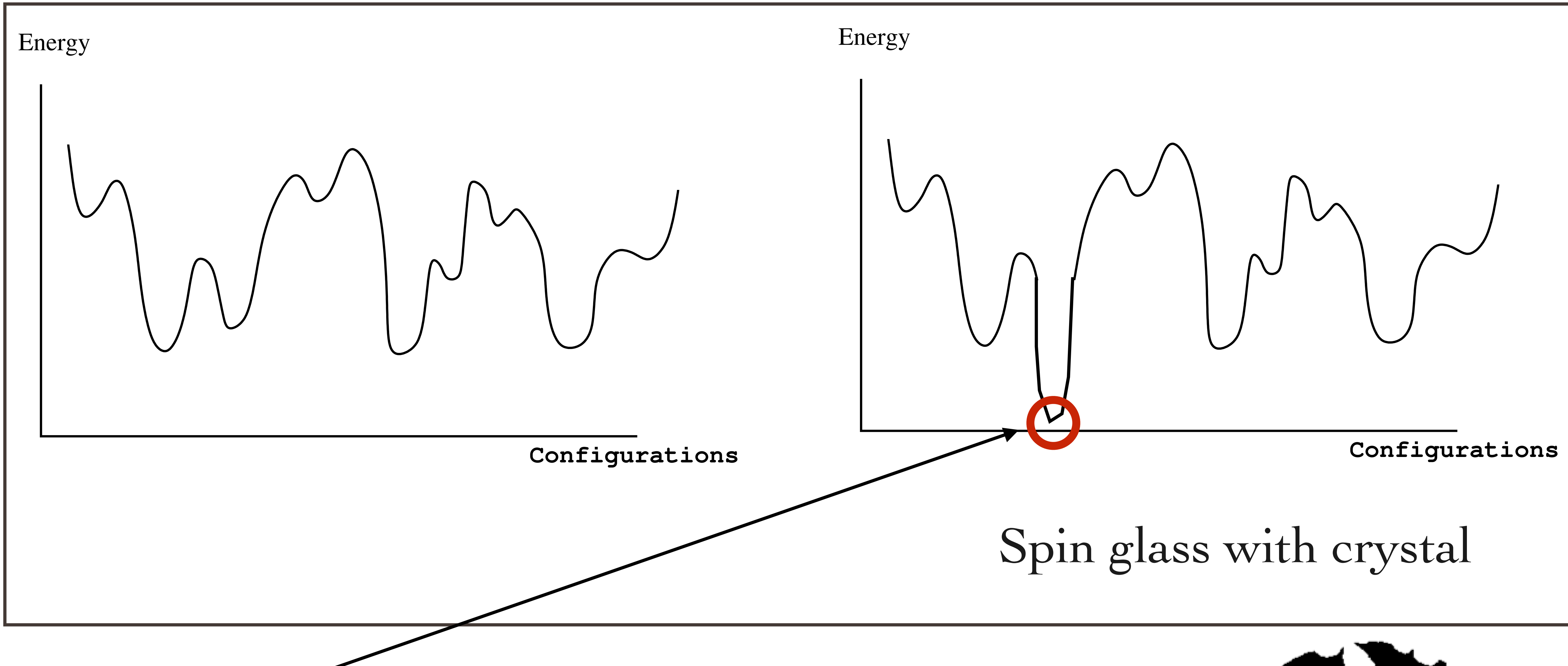
prior measurement

Studied with mean-field

$$P(S) = \prod_{ij \in \mathrm{grid}} e^{J s_i s_j} \prod_{\mu} \delta\left(y_\mu, \sum_{i \in \partial\mu} s_i \right)$$

If enough measurements: The most probable S (the ground state) gives the perfect composition of the sample.

« Crystal » : much more probable



« Crystal » : much more probable

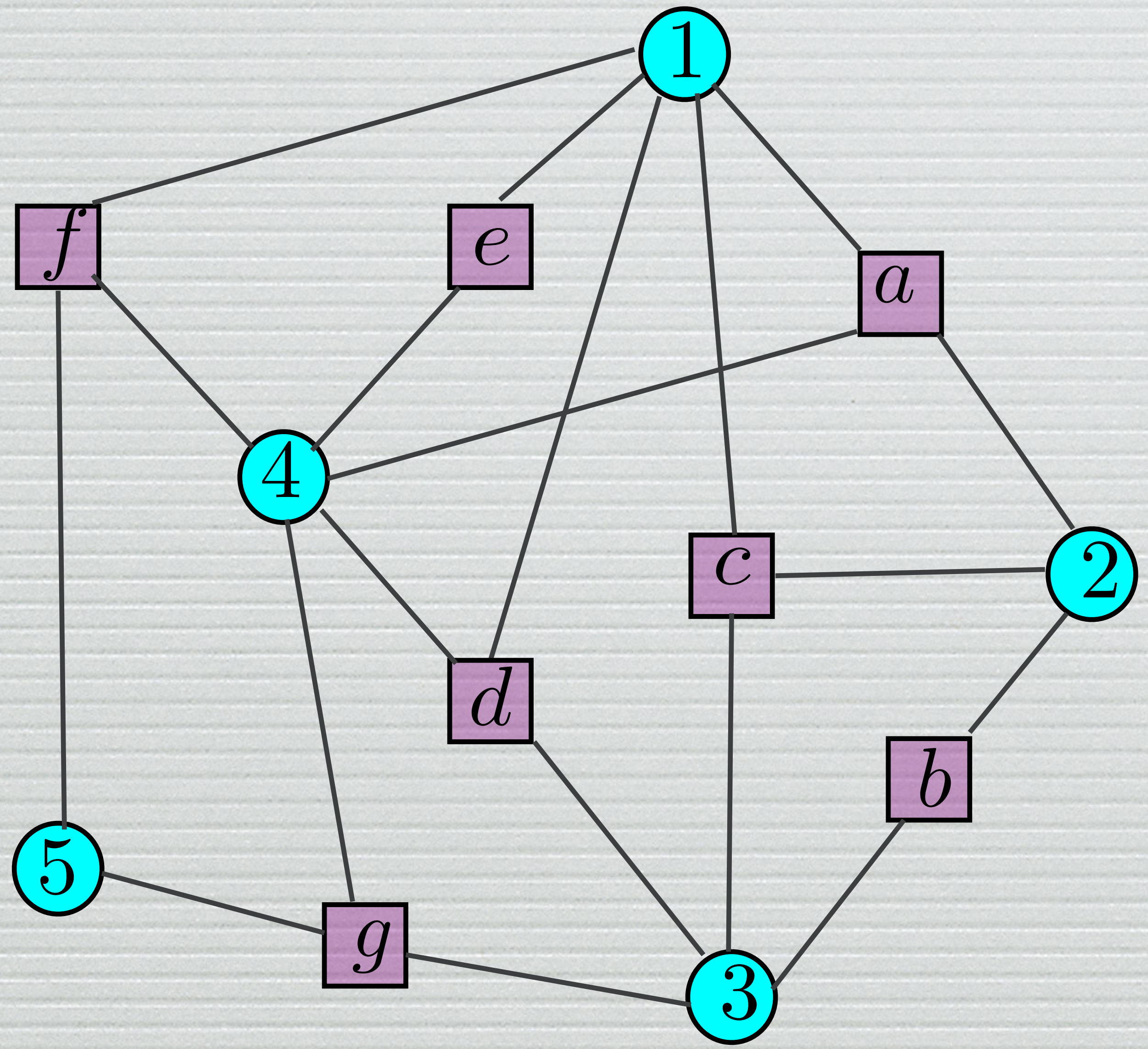
But in some cases « crystal hunting » may be computationally very hard !

Inference with many unknowns : « crystal hunting » with mean-field based algorithms

Historical development of mean field equations :

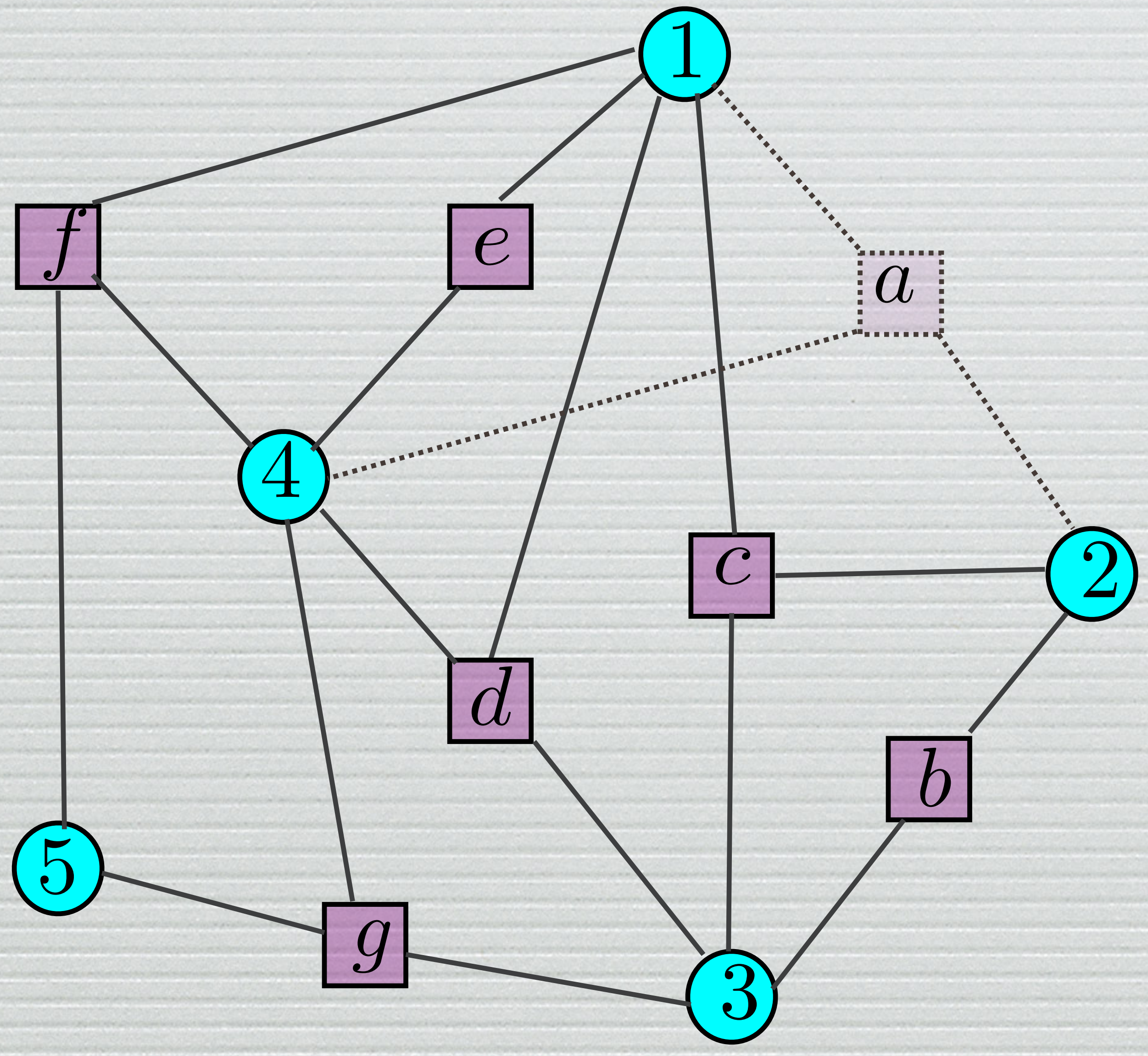
- In homogeneous ferromagnets:
  - Weiss (infinite range, 1907)
  - Bethe Peierls (finite connectivity, 1935)
- In glassy systems:
  - Thouless Anderson Palmer 1977,
  - MM Parisi Virasoro 1986 (infinite range)
  - MM Parisi 2001 (finite connectivity)
- As an algorithm:
  - Gallager 1963
  - Pearl 1986
  - MM Parisi Zecchina 2002
  - Kabashima 2003, 2008
  - Donoho Bayati Montanari 2009
  - Rangan 2010
  - Krzakala MM Zdeborova 2012 .

BP = Bethe-Peierls = Belief Propagation



$$P(x_1, \cdots, x_5) = \psi_a(x_1, x_2, x_4)\psi_b(x_2, x_3) \cdots$$

BP equations

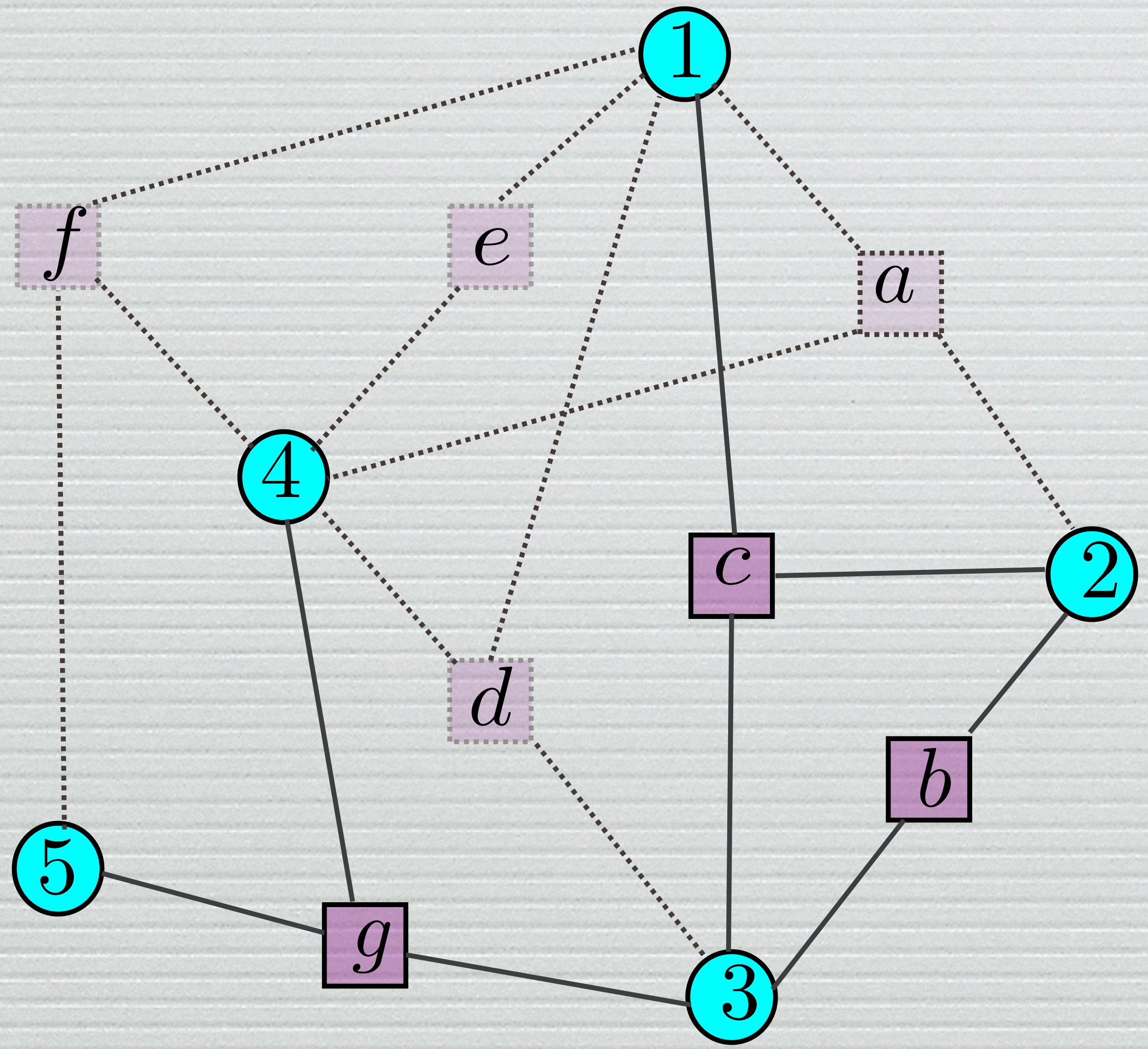


First type of messages:

Probability of x_1 in the absence of a:

$$m_{1\to a}(x_1)$$

BP equations

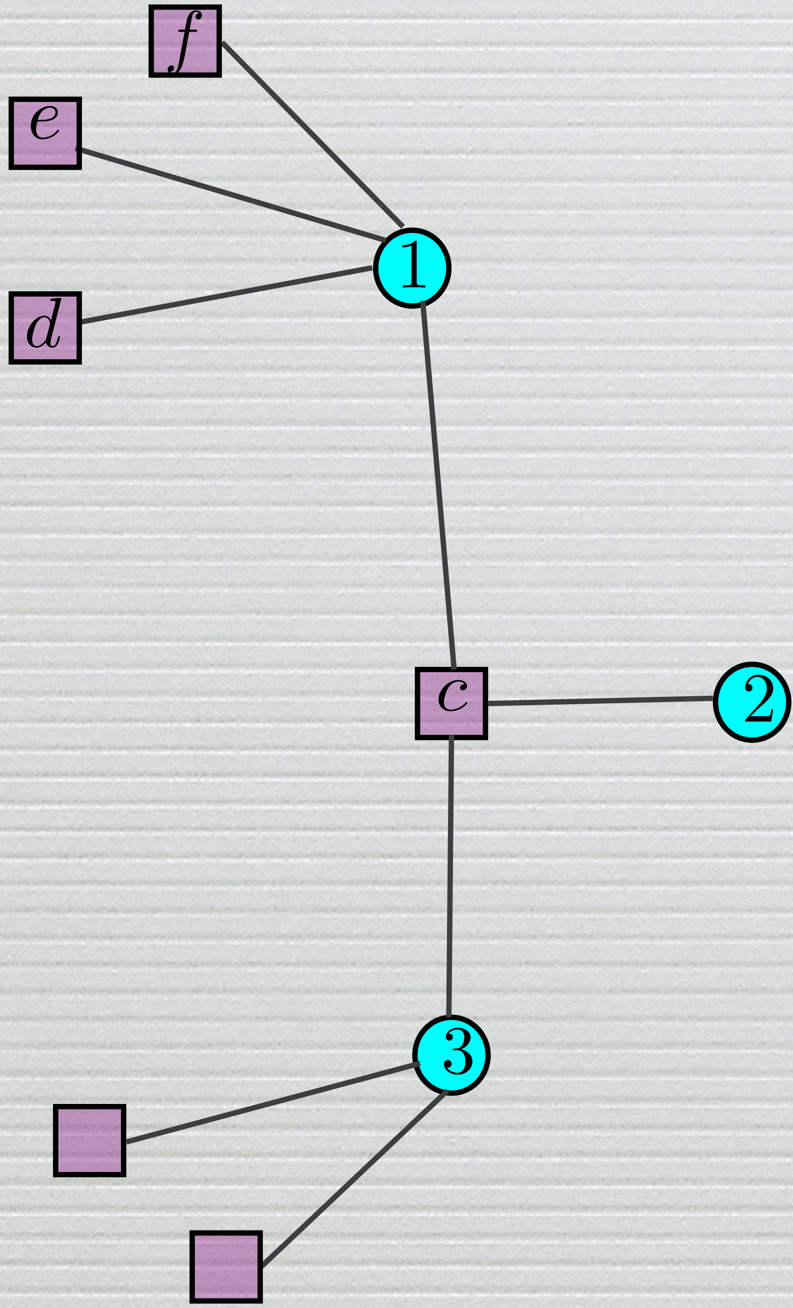


Second type of messages:

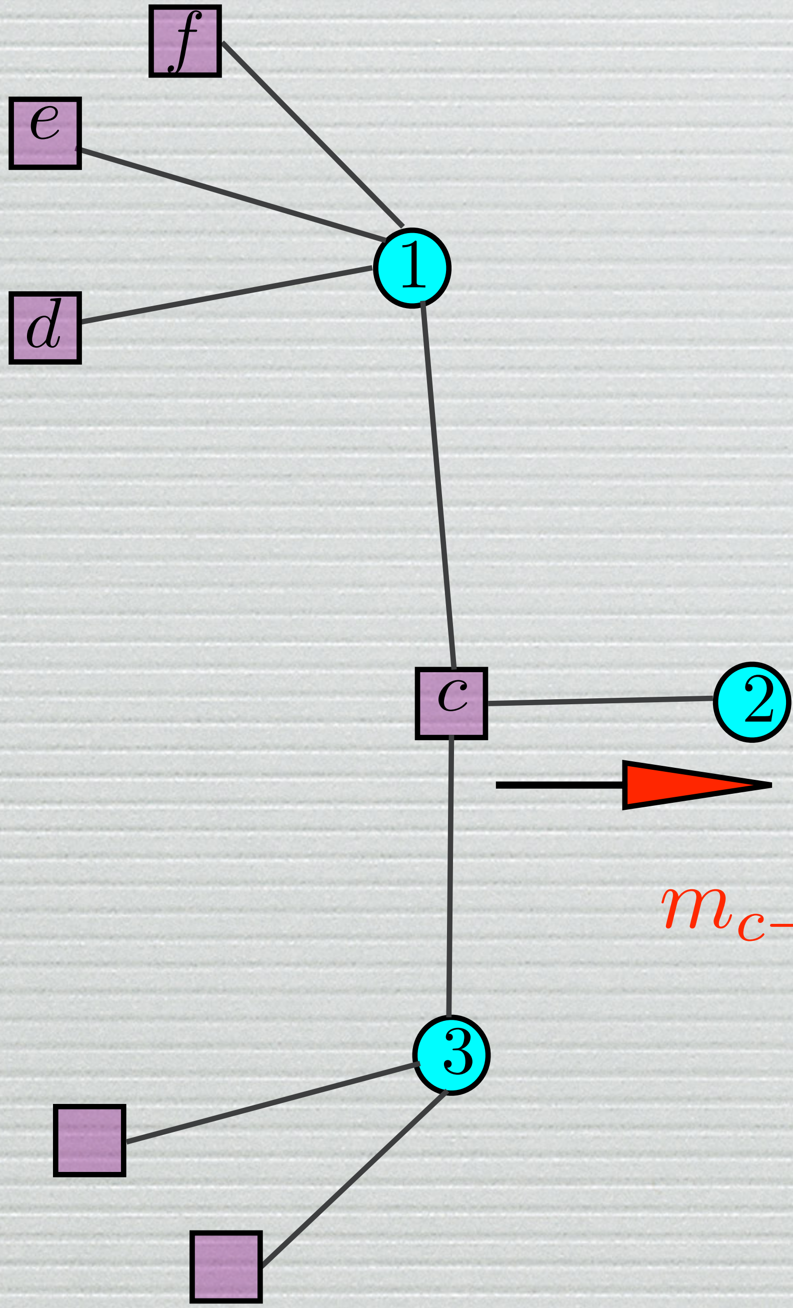
Probability of x_1 when it is connected only to c :

$$m_{c\to 1}(x_1)$$

BP equations

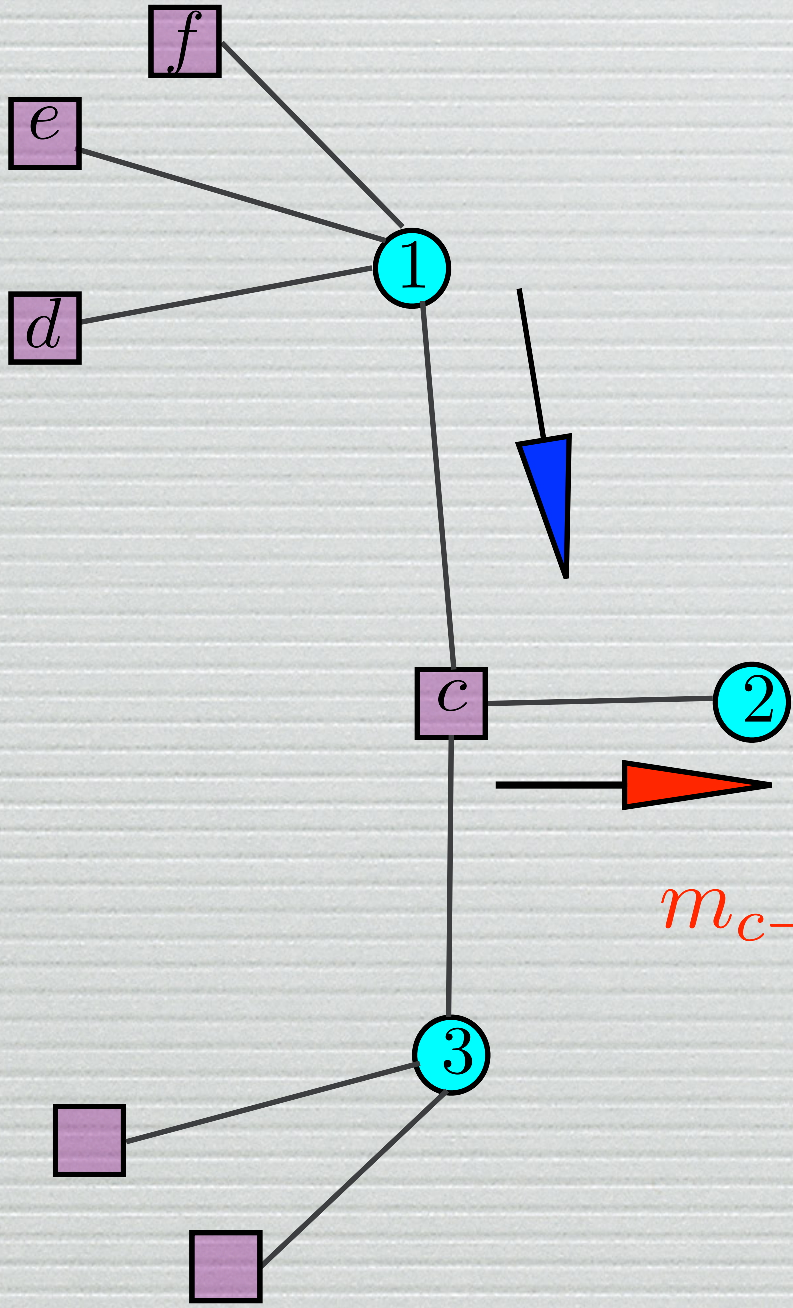


BP equations



$$m_{c\to 2}(x_2) = \sum_{x_1, x_3} \psi_c(x_1, x_2, x_3) m_{1\to c}(x_1) m_{3\to c}(x_3)$$

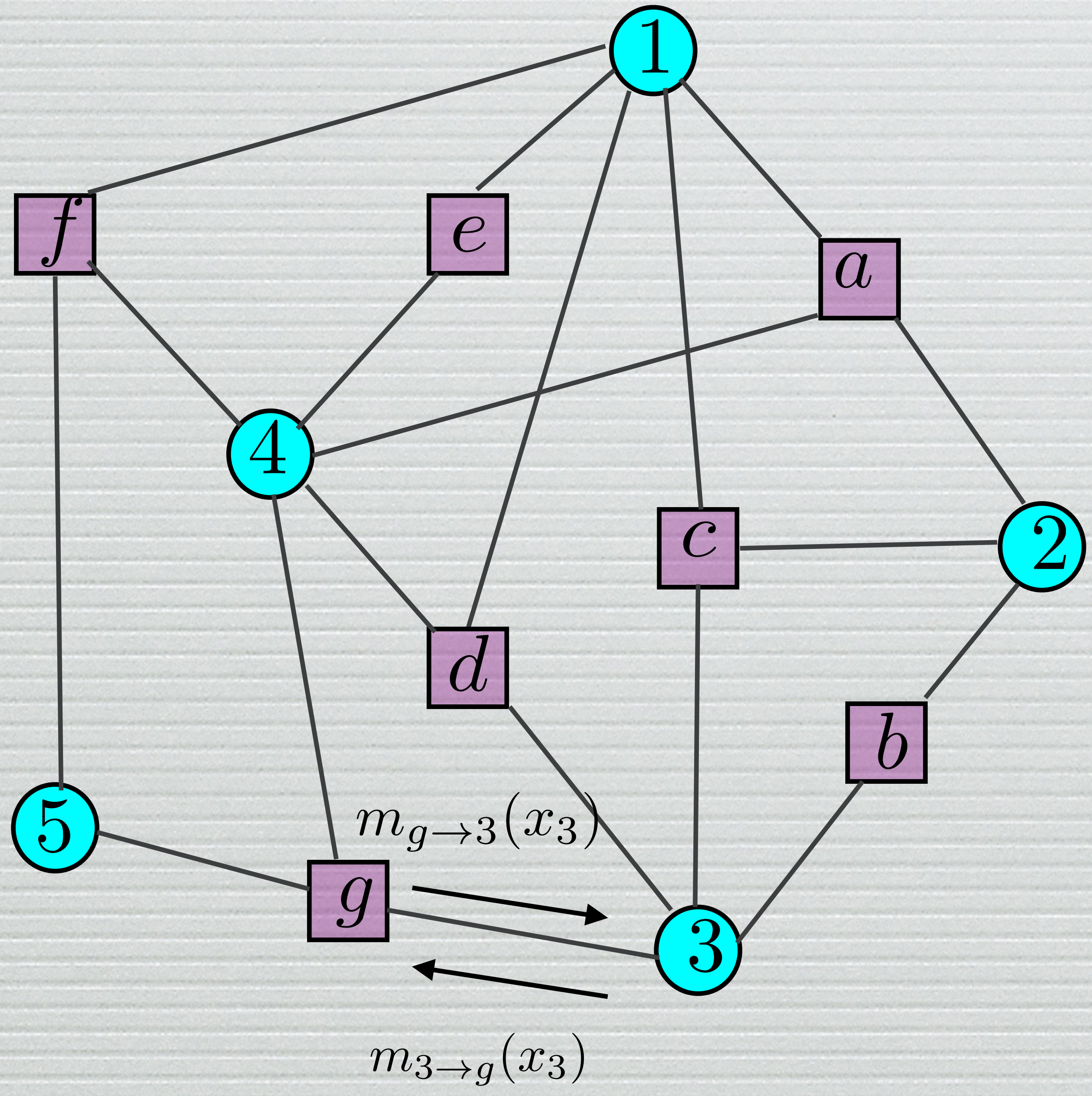
BP equations



$$m_{1\to c}(x_1) = C m_{d\to 1}(x_1) m_{e\to 1}(x_1) m_{f\to 1}(x_1)$$

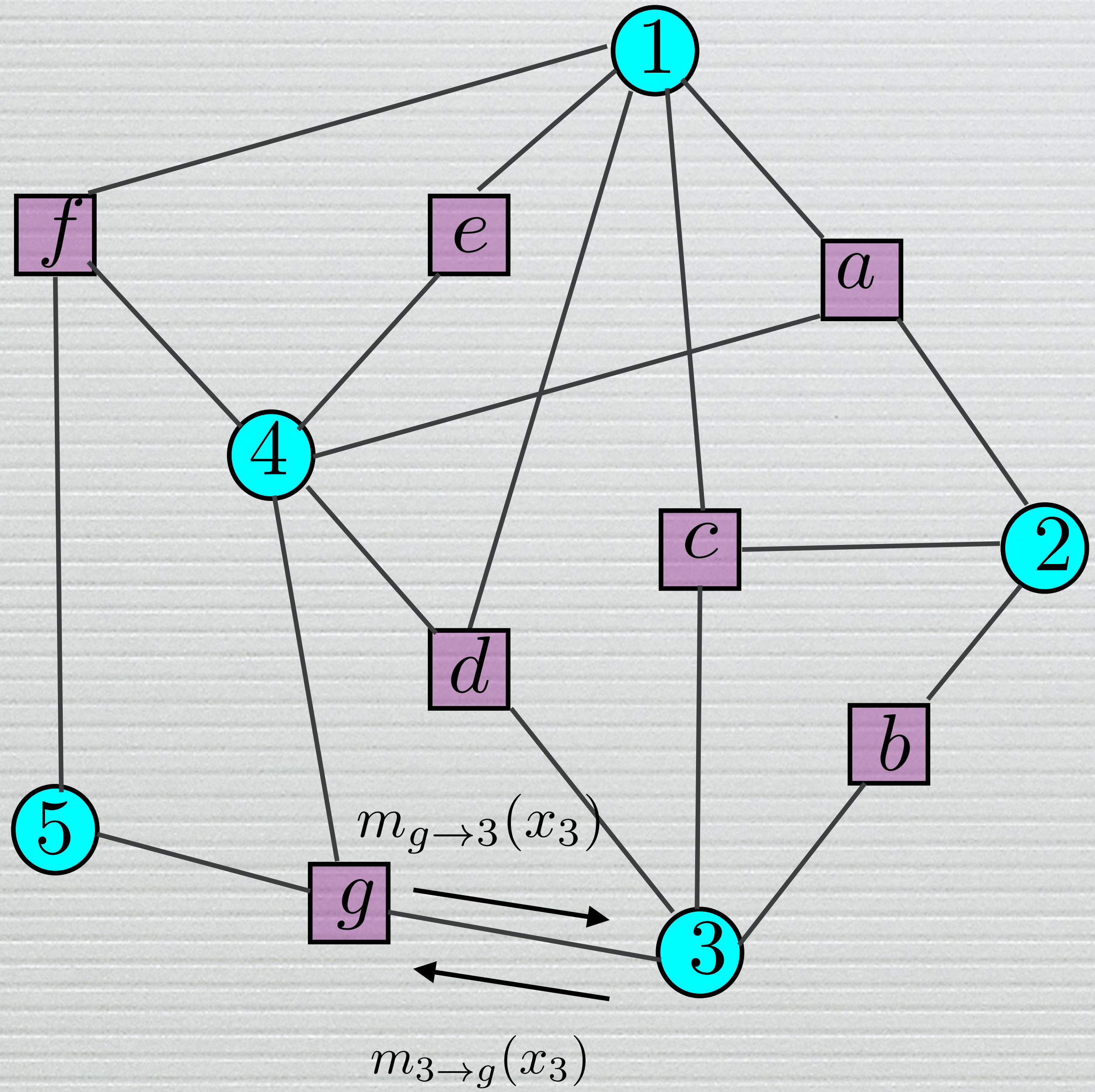
$$m_{c\to 2}(x_2) = \sum_{x_1, x_3} \psi_c(x_1, x_2, x_3) m_{1\to c}(x_1) m_{3\to c}(x_3)$$

BP equations



Propagate messages along the edges, update messages at vertices, using elementary local probabilistic rules

BP equations



Propagate messages along the edges, update messages at vertices, using elementary local probabilistic rules

Closed set of equations: two messages "propagate" on each edge of the factor graph.

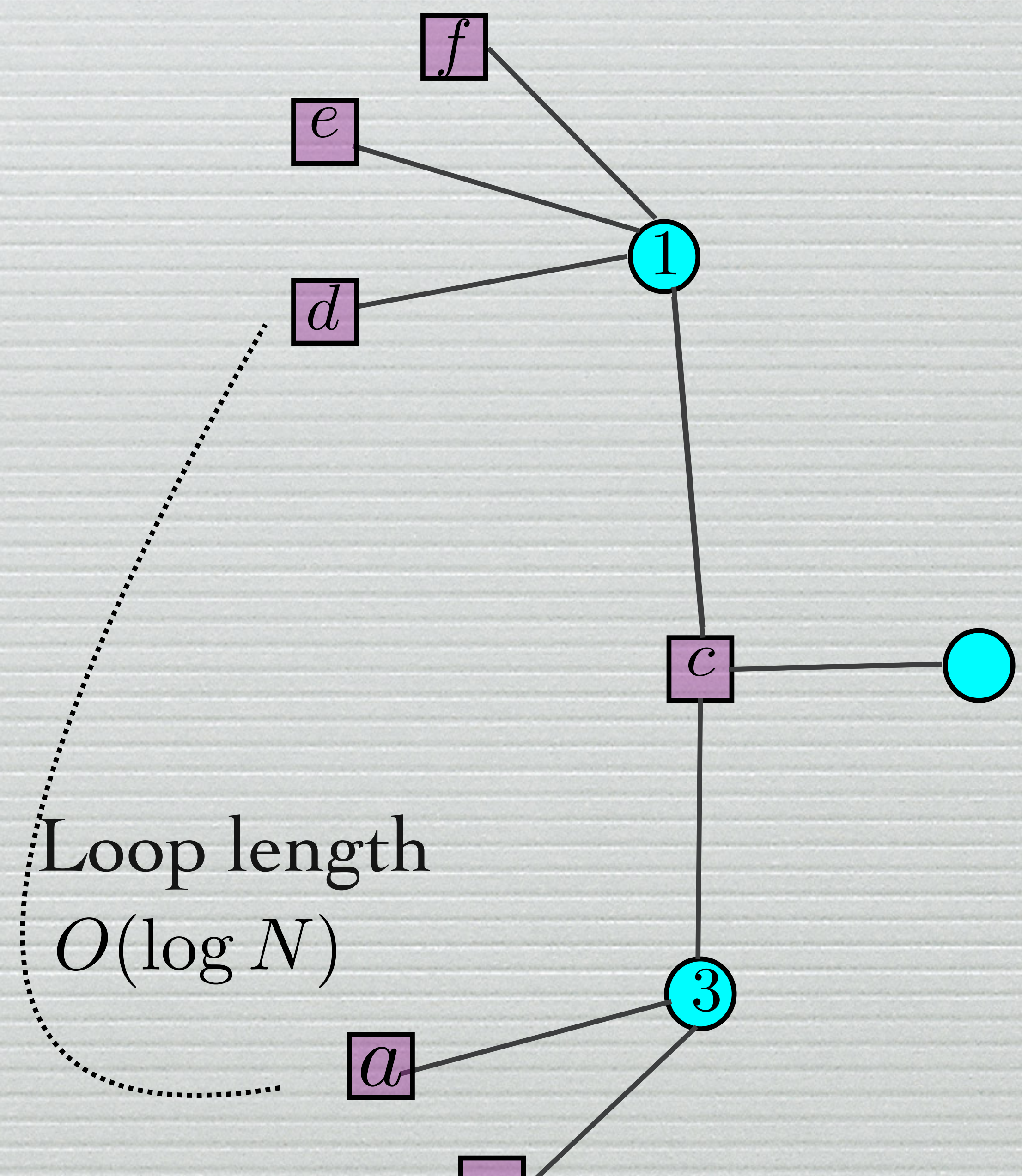
When is BP exact?

$$m_{1\to c}(x_1) = C m_{d\to 1}(x_1) m_{e\to 1}(x_1) m_{f\to 1}(x_1)$$

$$m_{c\to 2}(x_2) = \sum_{x_1,x_3} \psi_c(x_1,x_2,x_3) m_{1\to c}(x_1) m_{3\to c}(x_3)$$

Fluctuations are handled correctly, but beware of correlations

- Exact in one dimension (transfer matrix = dynamic programming)
- Exact on a tree (uncorrelated b.c)
- Exact on locally tree-like graphs (Erdös Renyi etc.) if correlations decay fast enough (single pure state) and uncorrelated disorder
- Exact in infinite range problems if correlations decay fast enough (single pure state) and uncorrelated disorder

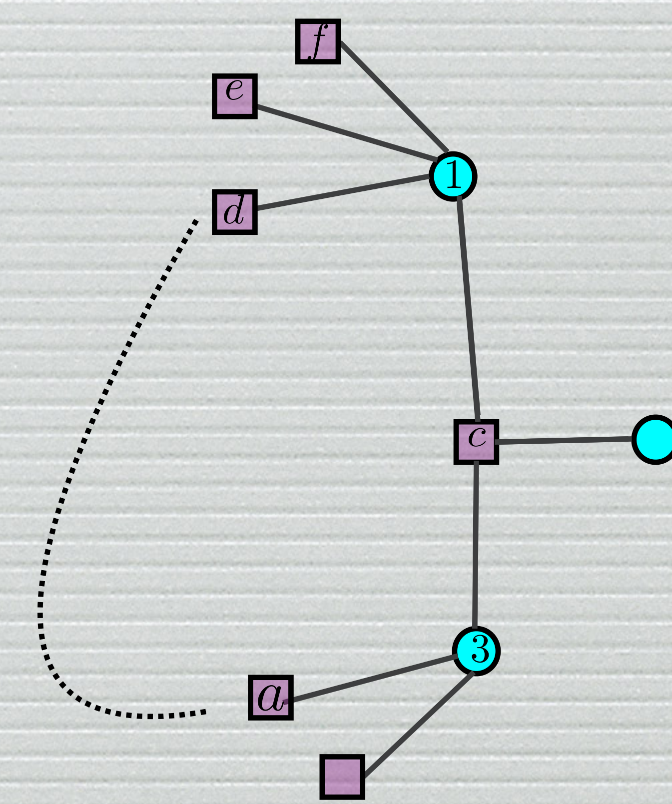


NB: What happens in a glass phase, when there are many pure states, and therefore many solutions ?

BP equations

$$m_{i\to\mu}(x_i) = \prod_{\nu(\neq\mu)} m_{\nu\to i}(x_i)$$

Correct if, in absence of the i-j interaction, the correlations between k and ℓ can be neglected.



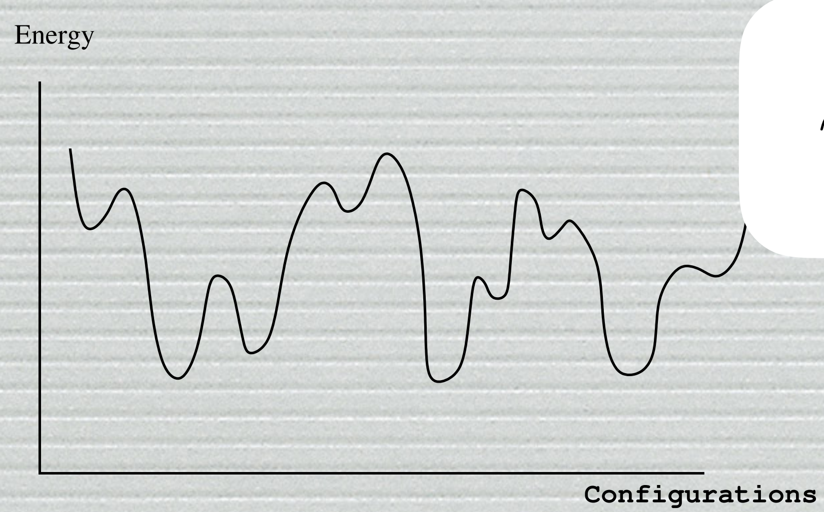
Loop length $O(\log N)$

NB: What happens in a glass phase, when there are many pure states, and therefore many solutions ?

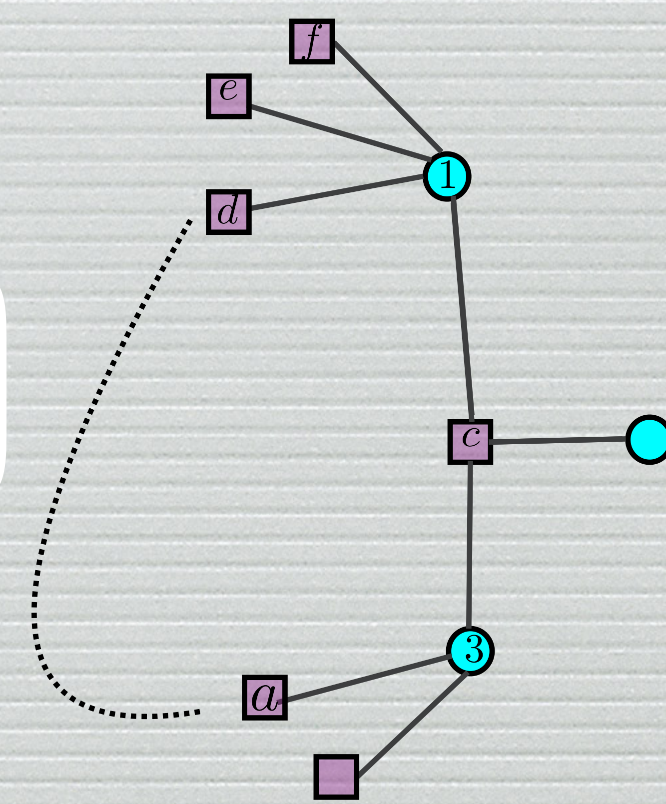
BP equations

$$m_{i\to\mu}(x_i) = \prod_{\nu(\neq\mu)} m_{\nu\to i}(x_i)$$

Correct if, in absence of the i-j interaction, the correlations between k and ℓ can be neglected.



$$m^{\alpha}_{i\to\mu}(x_i) = \prod_{\nu(\neq\mu)} m^{\alpha}_{\nu\to i}(x_i)$$



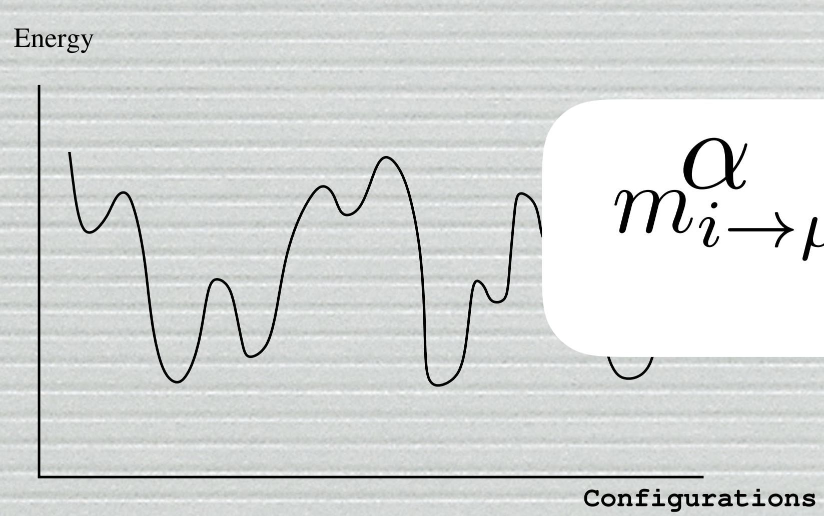
Glassy phase: many states, many solutions of BP

Loop length $O(\log N)$

2) What happens in a glass phase, when there are many pure states, and therefore many solutions ?

BP equations
$$m_{i\to\mu}(x_i) = \prod_{\nu(\neq\mu)} m_{\nu\to i}(x_i)$$

Correct if, in absence of the i-j interaction, the correlations between k and ℓ can be neglected.



$$m^{\alpha}_{i\to\mu}(x_i) = \prod_{\nu(\neq\mu)} m^{\alpha}_{\nu\to i}(x_i)$$

Glassy phase: many states, many solutions of BP

Statistics of $m^{\alpha}_{i\to\mu}(x_i)$ over the many states α

$$P_{i\to\mu}(m)$$

related to

$$P_{\nu\to i}(m)$$

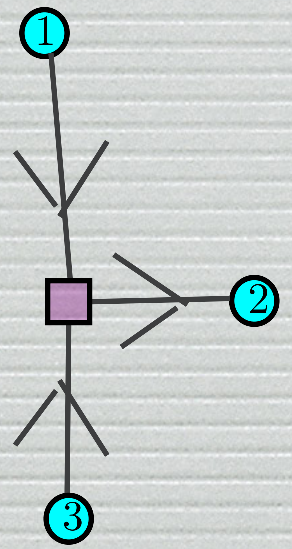
Survey propagation (SP)

MM Parisi Zecchina 2002

Power of message passing algorithms

Approximate solution of very hard, and very large constraint satisfaction problems, ...FAST! (typically linear time)

- BP: Best decoders for LDPC error correcting codes
- SP: Best solver of random satisfiability problems
- BP: Best algorithm for learning patterns in neural networks (e.g. binary perceptron)
- Data clustering, graph coloring, Steiner trees, etc…
- Fully connected networks : TAP (=AMP). Compressed sensing, linear estimation, etc.

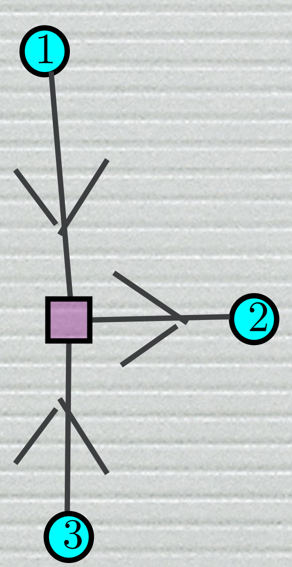


Local, simple update equations:
Each message is updated using information from incoming messages on the same node.
Distributed, solves hard global pb

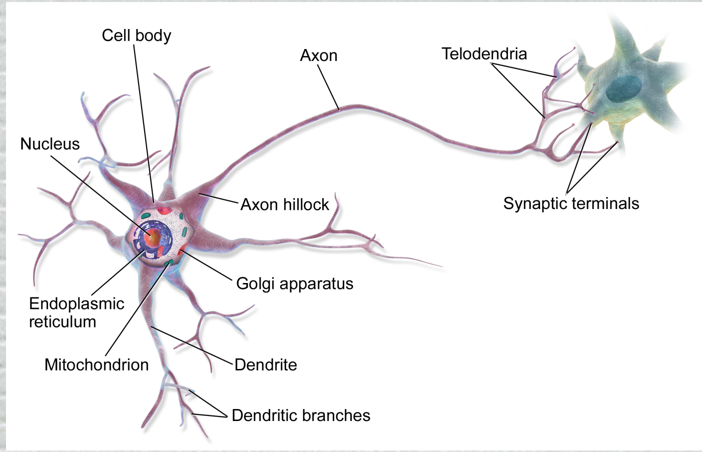
Power of message passing algorithms

Approximate solution of very hard, and very large constraint satisfaction problems, ...FAST! (typically linear time)

- BP: Best decoders for LDPC error correcting codes
- SP: Best solver of random satisfiability problems
- BP: Best algorithm for learning patterns in neural networks (e.g. binary perceptron)
- Data clustering, graph coloring, Steiner trees, etc…
- Fully connected networks : TAP (=AMP). Compressed sensing, linear estimation, etc.



Local, simple update equations:
Each message is updated using information from incoming messages on the same node.
Distributed, solves hard global pb



An example of mean-field based inference: Compressed sensing

Applications:

- Tomography
- MNR
- Single pixel camera
- Satellite images
- ...

Connected to:

- linear regression
- perceptron learning

An example of mean-field based inference: Compressed sensing

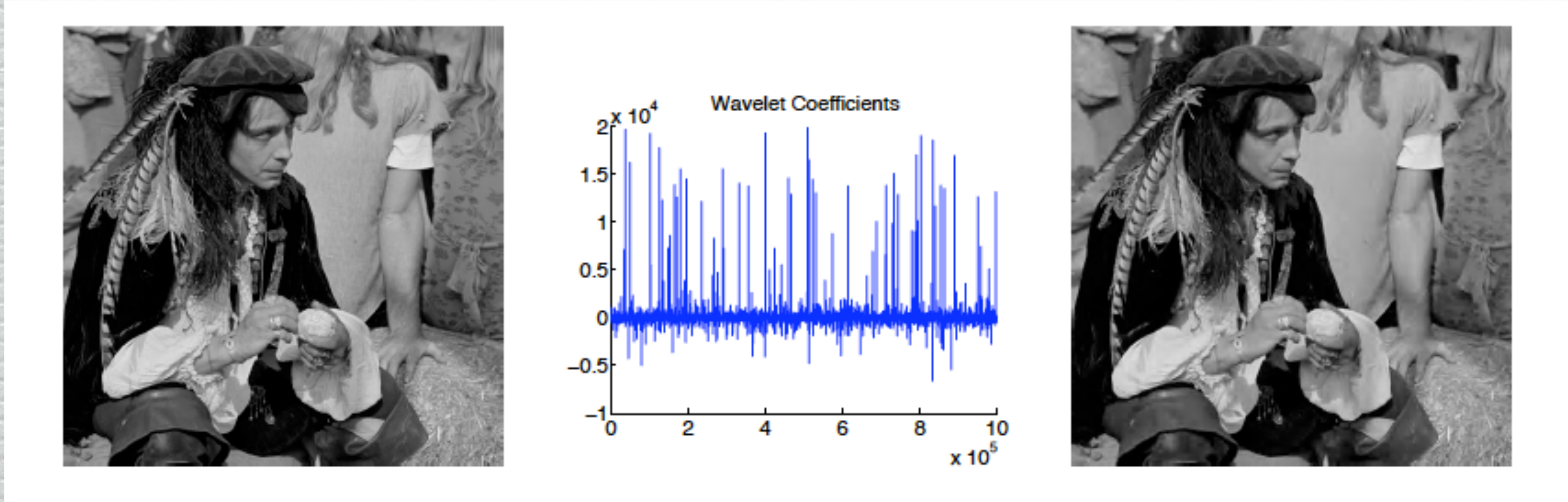
Applications:

- Tomography
- MNR
- Single pixel camera
- Satellite images
- …

Connected to:

- linear regression
- perceptron learning

Sparse data (in appropriate basis)+
linear measurements



Benchmark: noiseless limit of compressed sensing with iid measurements

System of linear measurements

$$y = Fx$$

Measurements

$$y = \begin{pmatrix} y^1 \\ . \\ . \\ y^M \end{pmatrix}$$

Signal $x = \begin{pmatrix} x_1 \\ . \\ . \\ . \\ x_N \end{pmatrix}$

$F = M \times N$ matrix

Random F : «random projections» (incoherent with signal)

Pb: Find x when $M < N$ and x is sparse

Phase diagram

«Thermodynamic limit»

$N \gg 1$	variables
$R = \rho N$	non-zero variables
$M = \alpha N$	equations

- Solvable by enumeration when $\alpha > \rho$ but $O(e^N)$
- ℓ_1 norm approach

  Find a N - component vector x such that the M equations $y = Fx$ are satisfied and $||x||_1$ is minimal
- AMP = Bayesian approach

Planted: $\phi_T(x)$

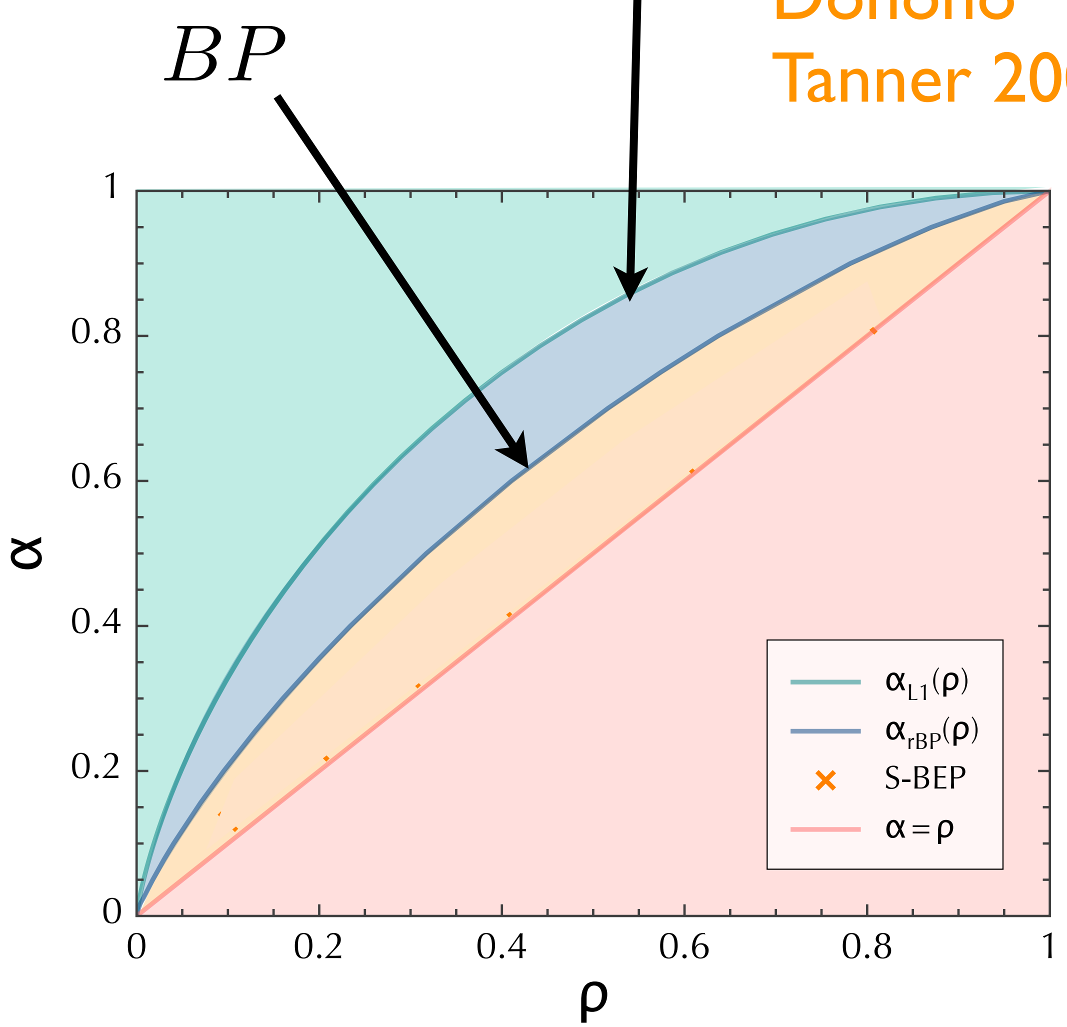
$$P(\mathbf{x}) = \prod_{i=1}^{N} [(1-\rho)\delta(x_i) + \rho\phi(x_i)] \quad \prod_{\mu=1}^{P} \delta\left(y_\mu - \sum_i F_{\mu i} x_i\right)$$

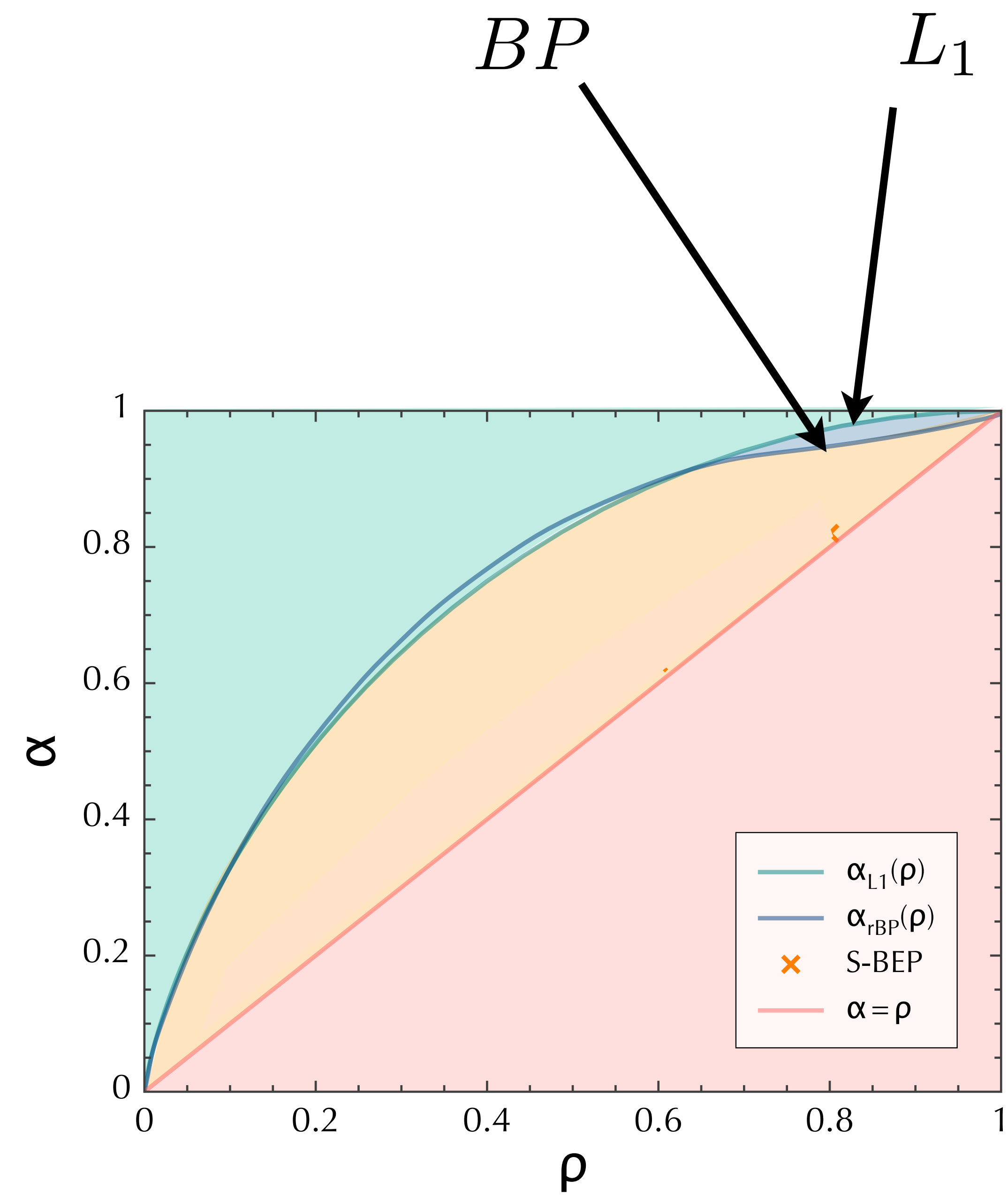
↑

Performance of AMP with Gauss-Bernoulli prior: phase diagram

Krzakala Sausset Mézard Sun Zdeborova 2011

Donoho 2006, Donoho Tanner 2005





Gaussian signal

$$\phi_T(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

Binary signal

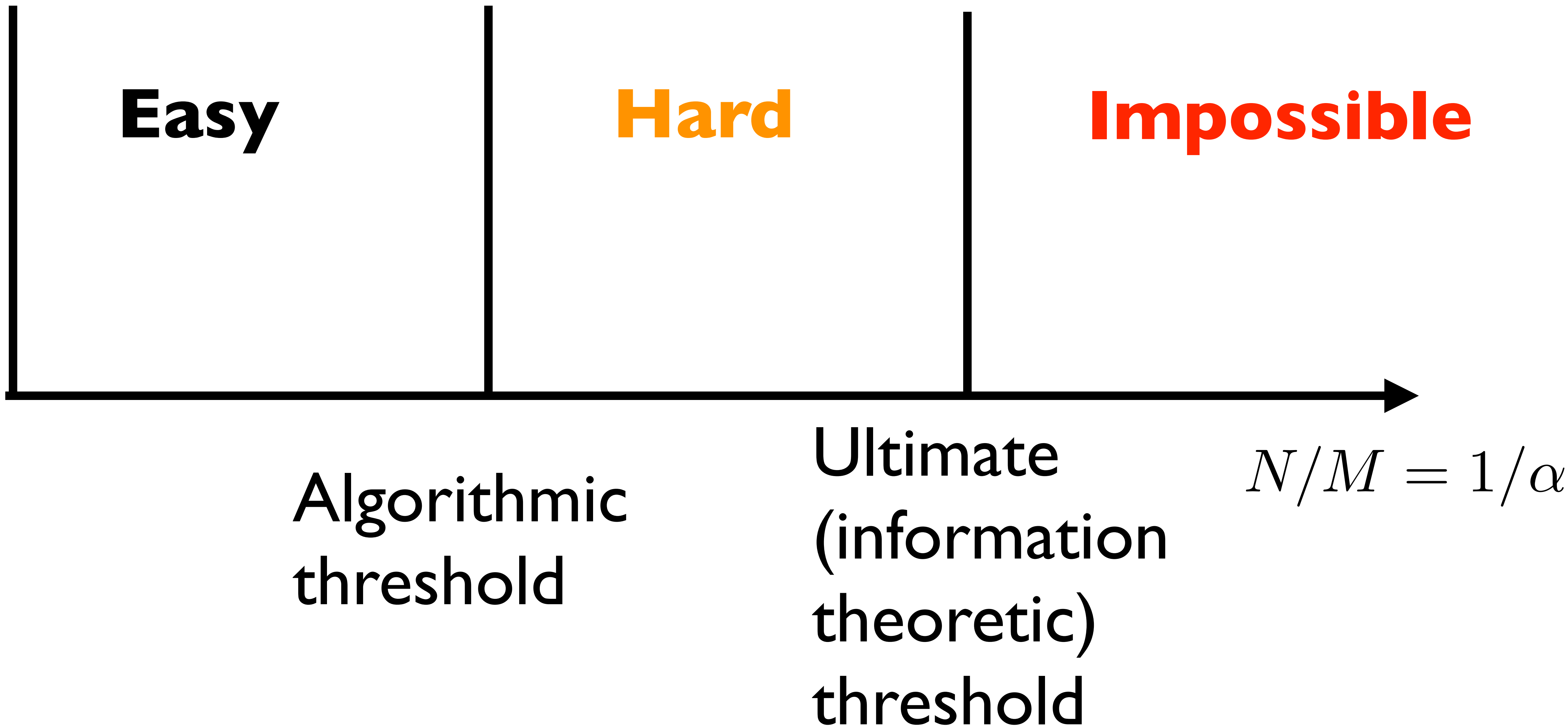
$$\phi_T(x) = \frac{1}{2}\left(\delta_{x,1} + \delta_{x,-1}\right)$$

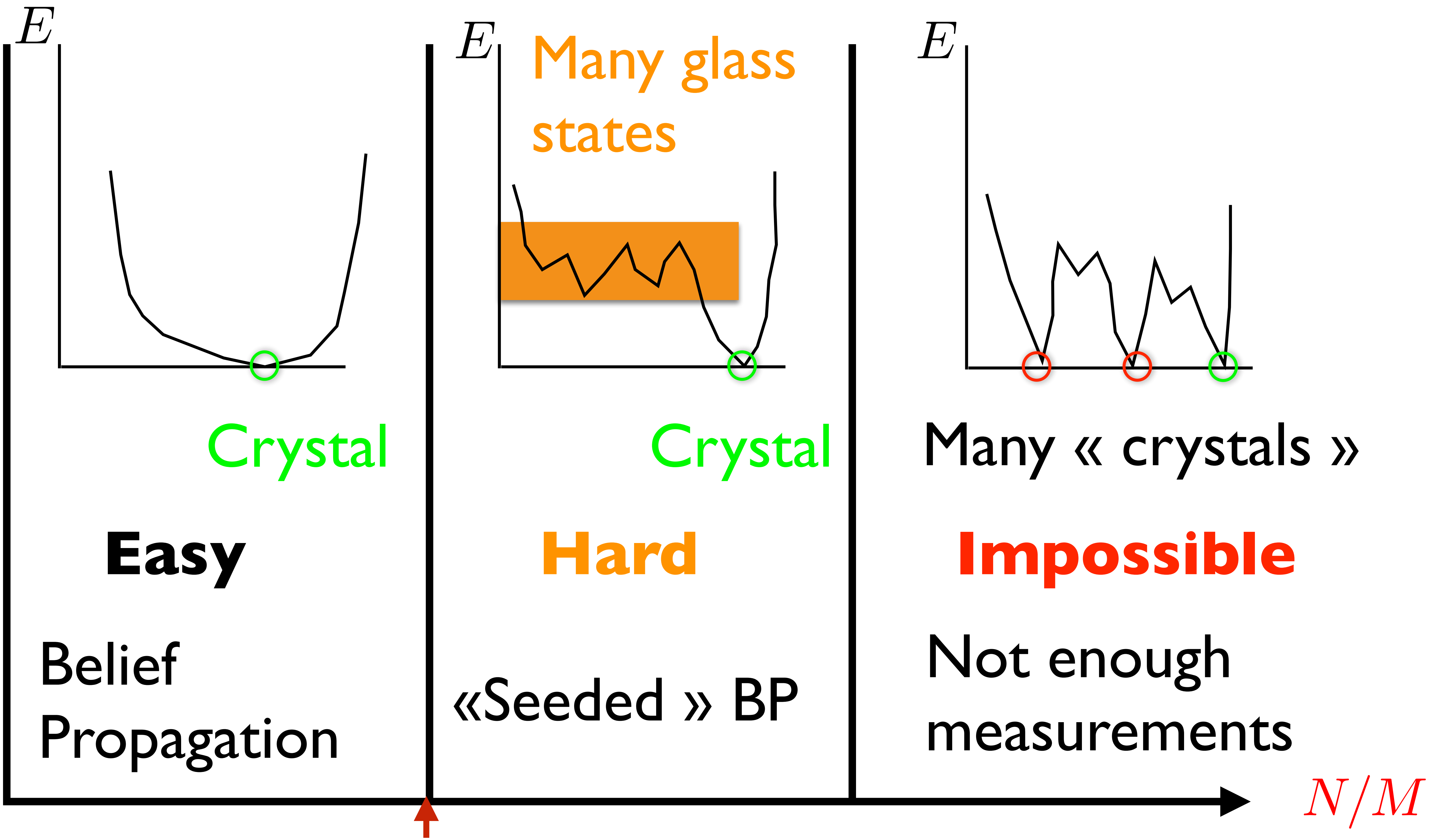
Analysis of random instances : phase transitions

N (real) variables, M measurements (linear functions)

Analysis of random instances : phase transitions

Reconstruction of signal using BP. Fixed ρ , decrease α





Dynamical phase transition. Ubiquitous in statistical inference. Conjecture « All local algorithms freeze »… How universal?

Step 3: design the measurement matrix in order to get around the glass transition

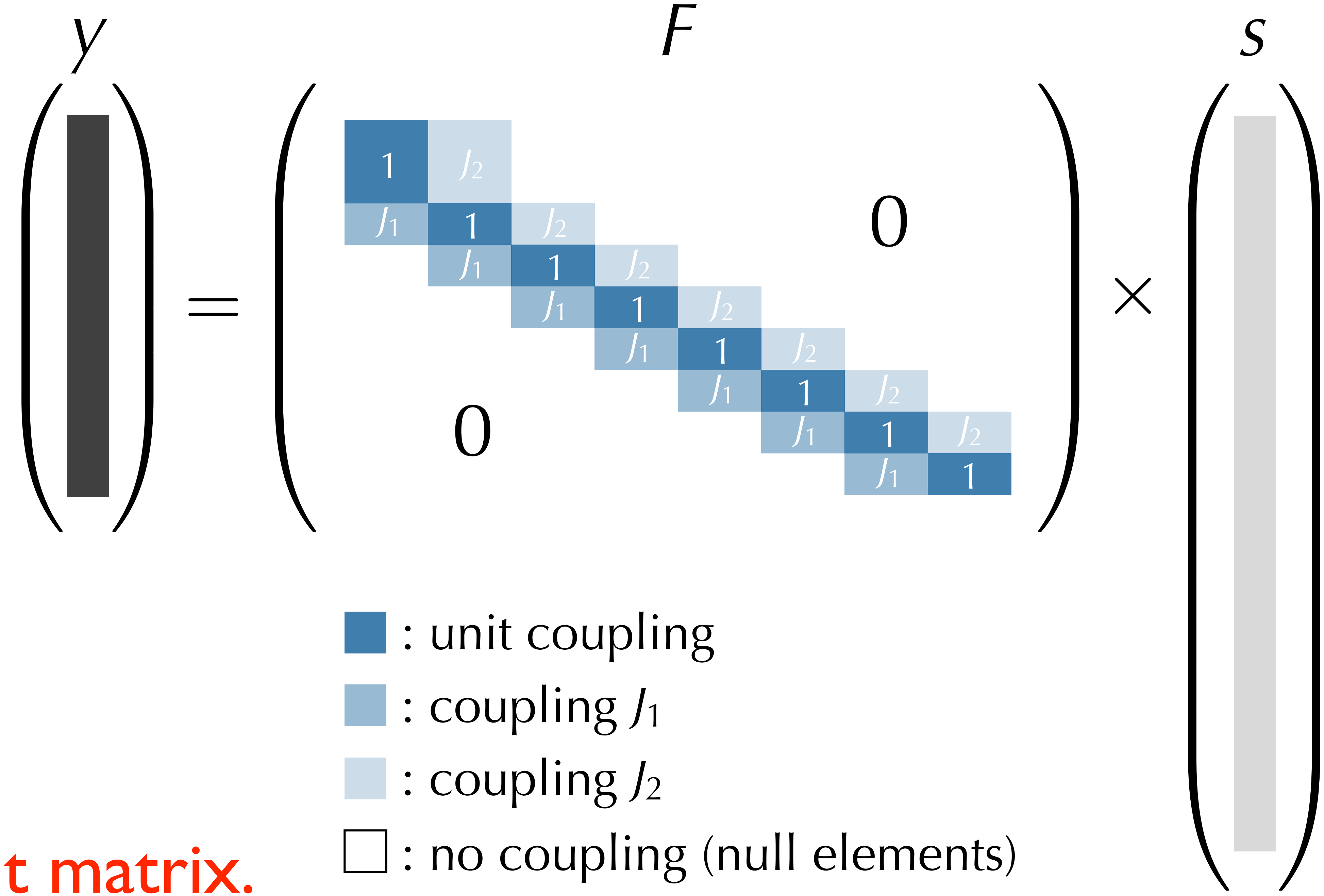
Getting around the glass trap: design the matrix F so that one nucleates the naive state (crystal nucleation idea,
...borrowed from error correcting codes : « spatial coupling »)

Felström-Zigangirov,
Kudekar Richardson Urbanke,
Hassani Macris Urbanke,
...

«Seeded BP»

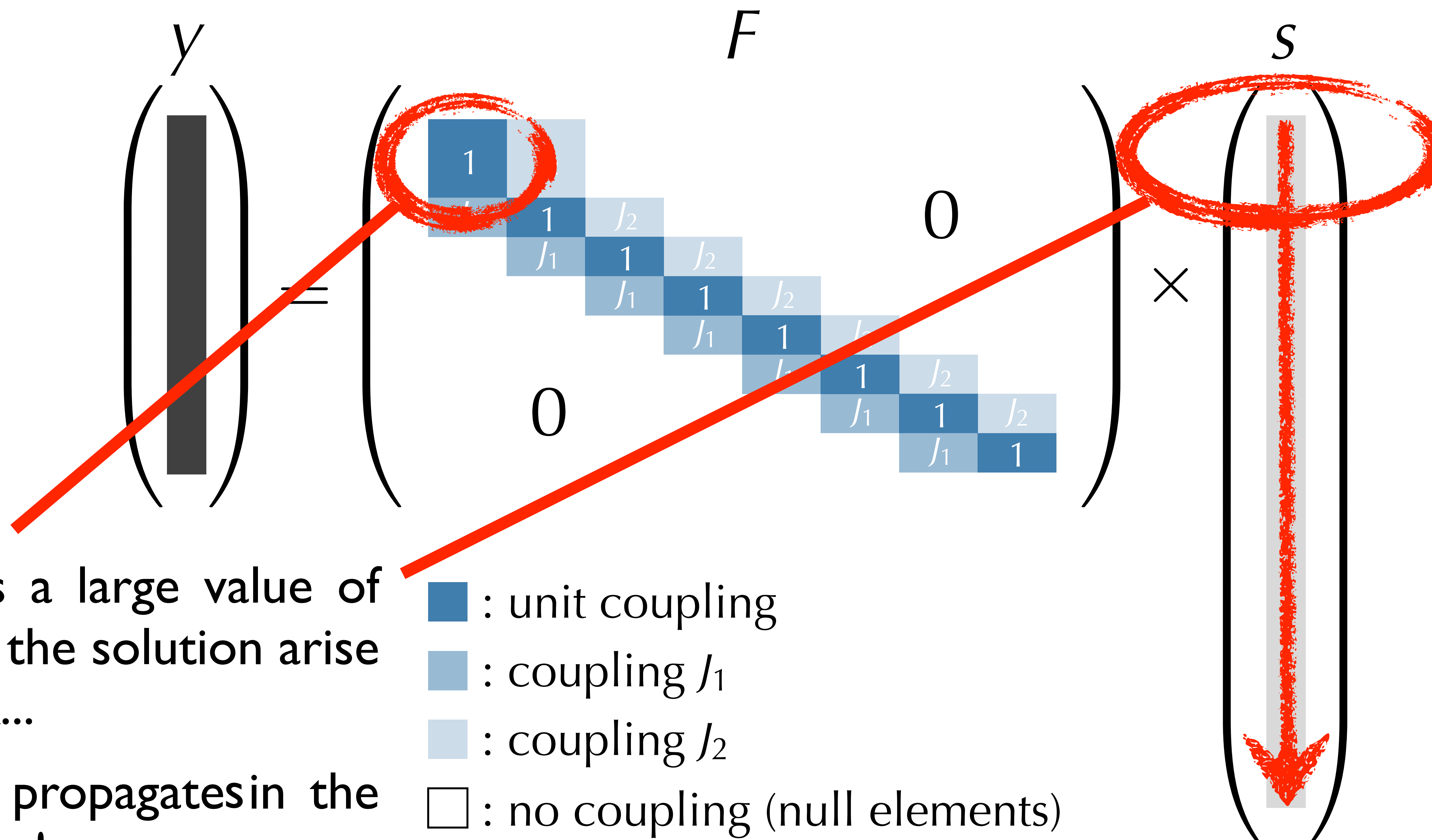
Nucleation and seeding

Nucleation and seeding



Structured measurement matrix. Variances of the matrix elements

$F_{\mu i} =$ independent random Gaussian variables, zero mean and variance $J_{b(\mu)b(i)}/N$



Block 1 has a large value of M such that the solution arise in this block...

... and then propagates in the whole system!

$$L = 8$$

$$N_i = N/L$$

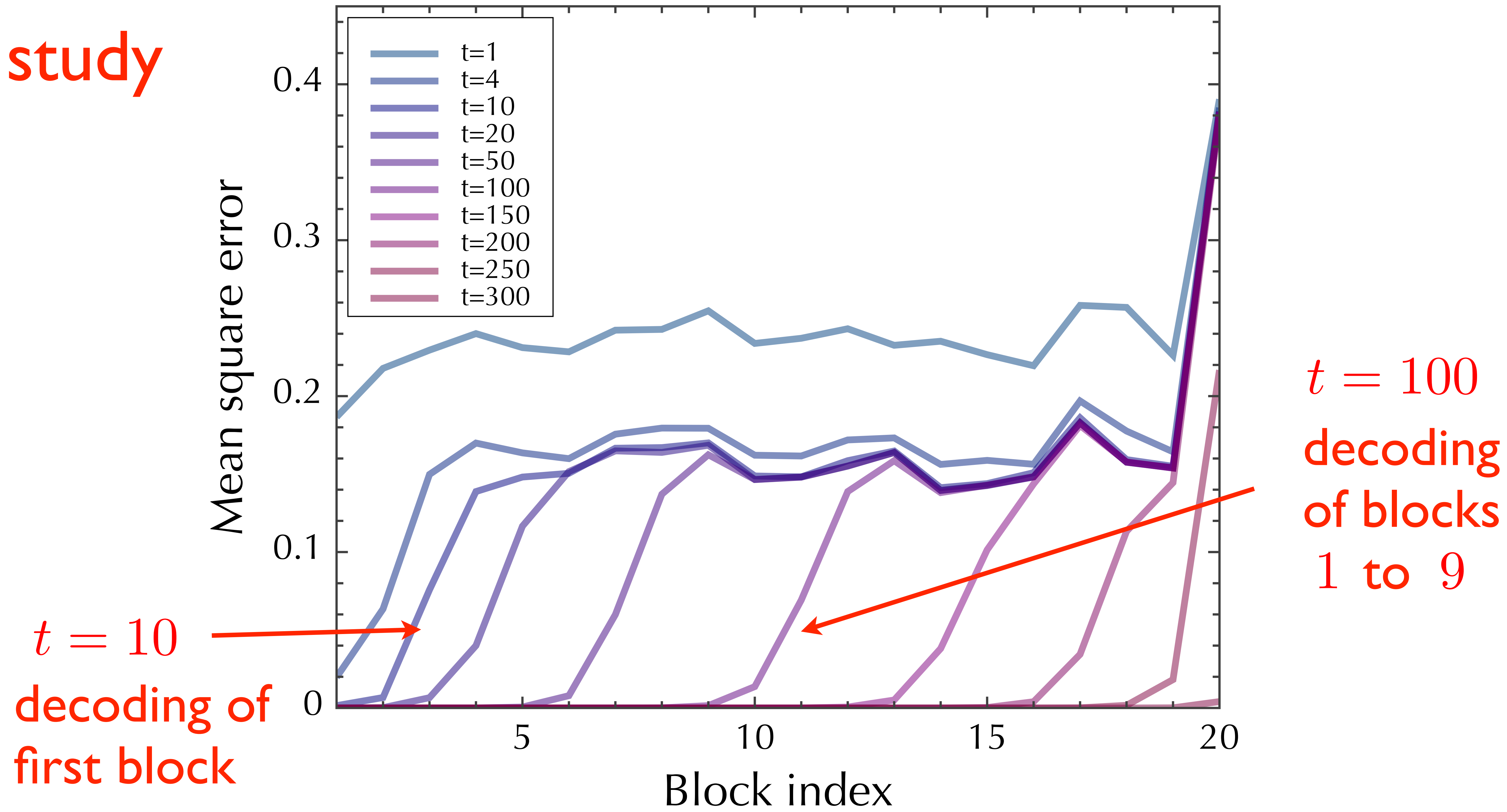
$$M_i = \alpha_i N/L$$

$$\alpha_1 > \alpha_{BP}$$

$$\alpha_j = \alpha' < \alpha_{BP} \qquad j \geq 2$$

$$\alpha = \frac{1}{L}\left(\alpha_1 + (L-1)\alpha'\right)$$

Numerical study

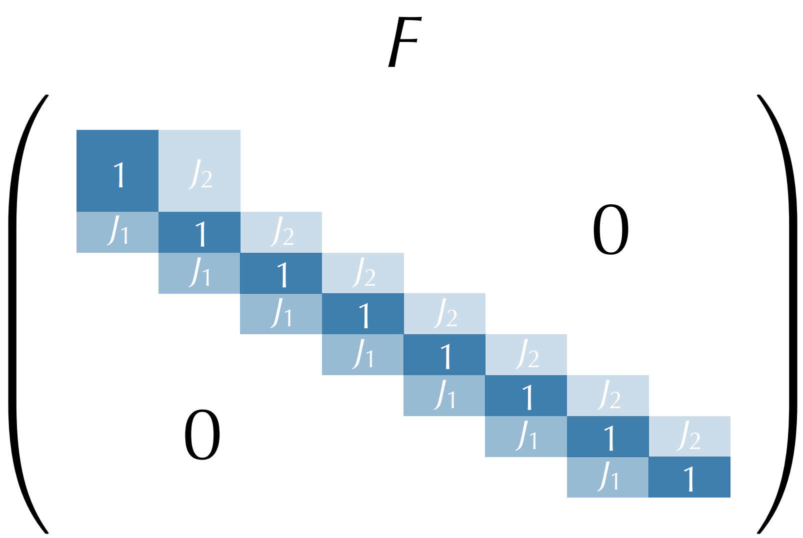


$L = 20$ $N = 50000$ $\rho = .4$ $J_1 = 20$ $\alpha_1 = 1$

$J_2 = .2$ $\alpha = .5$

Performance of the probabilistic approach + message passing + parameter learning+ seeding matrix

$$Z = \int \prod_{j=1}^{N} \mathrm{d}x_j \prod_{i=1}^{N} [(1-\rho)\delta(x_i) + \rho\phi(x_i)] \prod_{\mu=1}^{M} \delta\left(y_\mu - \sum_{i=1}^{N} F_{\mu i} x_i\right)$$



- Simulations
- Analytic approaches (replicas and cavity)

$$\rightarrow \alpha_c = \rho_0$$

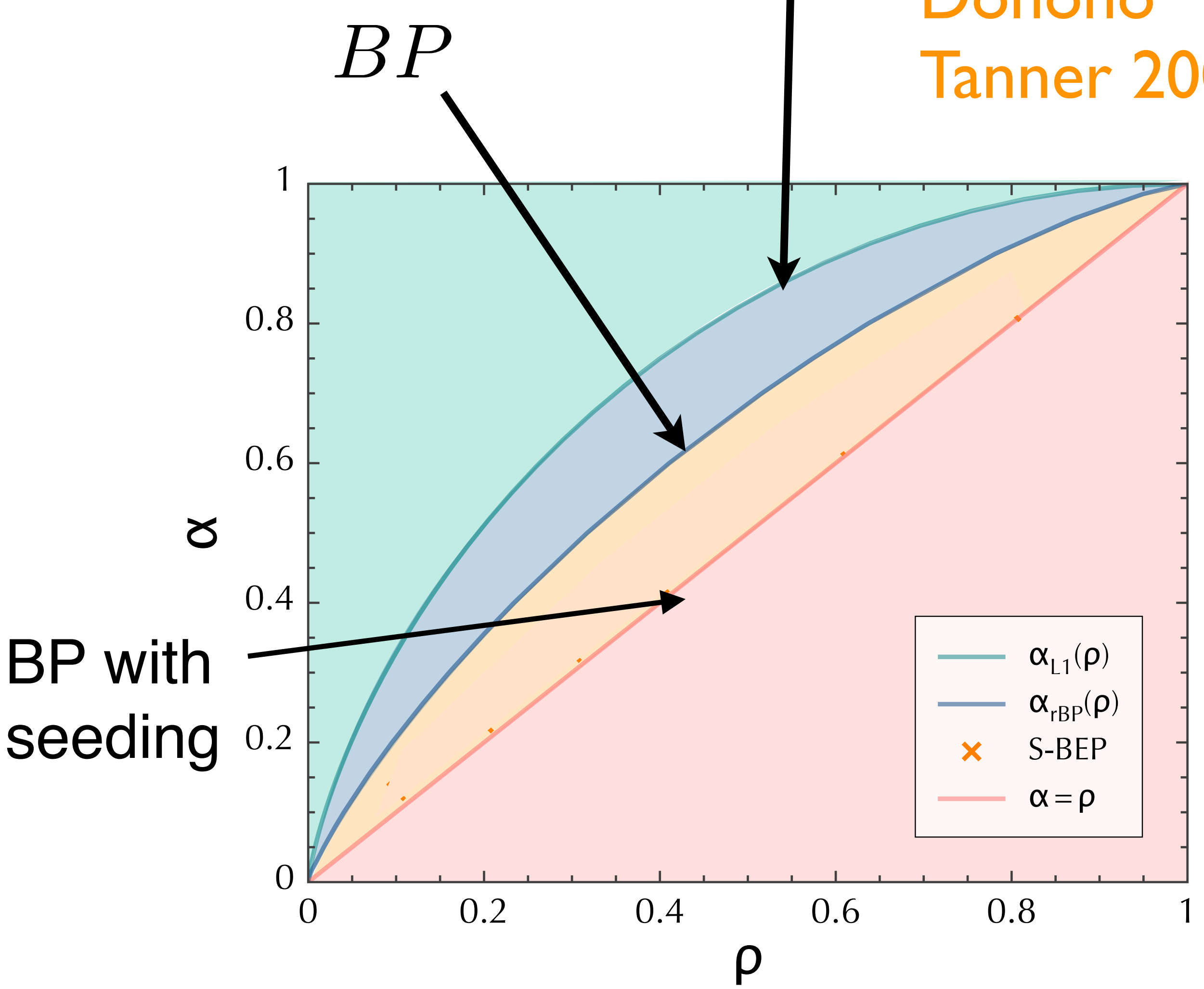
Reaches the ultimate information-theoretic threshold

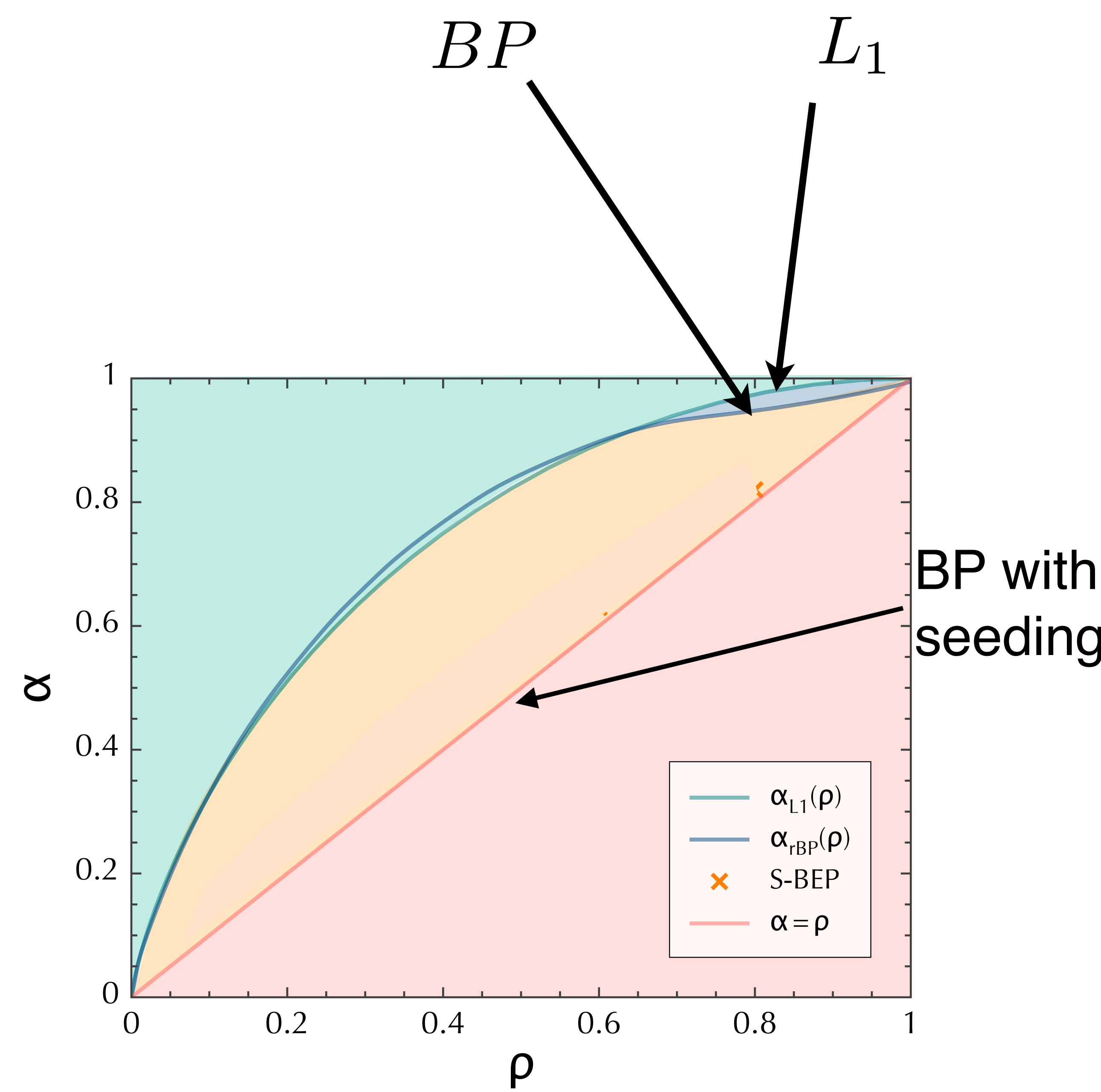
Proof: Donoho Javanmard Montanari

Performance of AMP with Gauss-Bernoulli prior: phase diagram

Krzakala Sausset Mézard Sun Zdeborova 2011

Donoho 2006, Donoho Tanner 2005



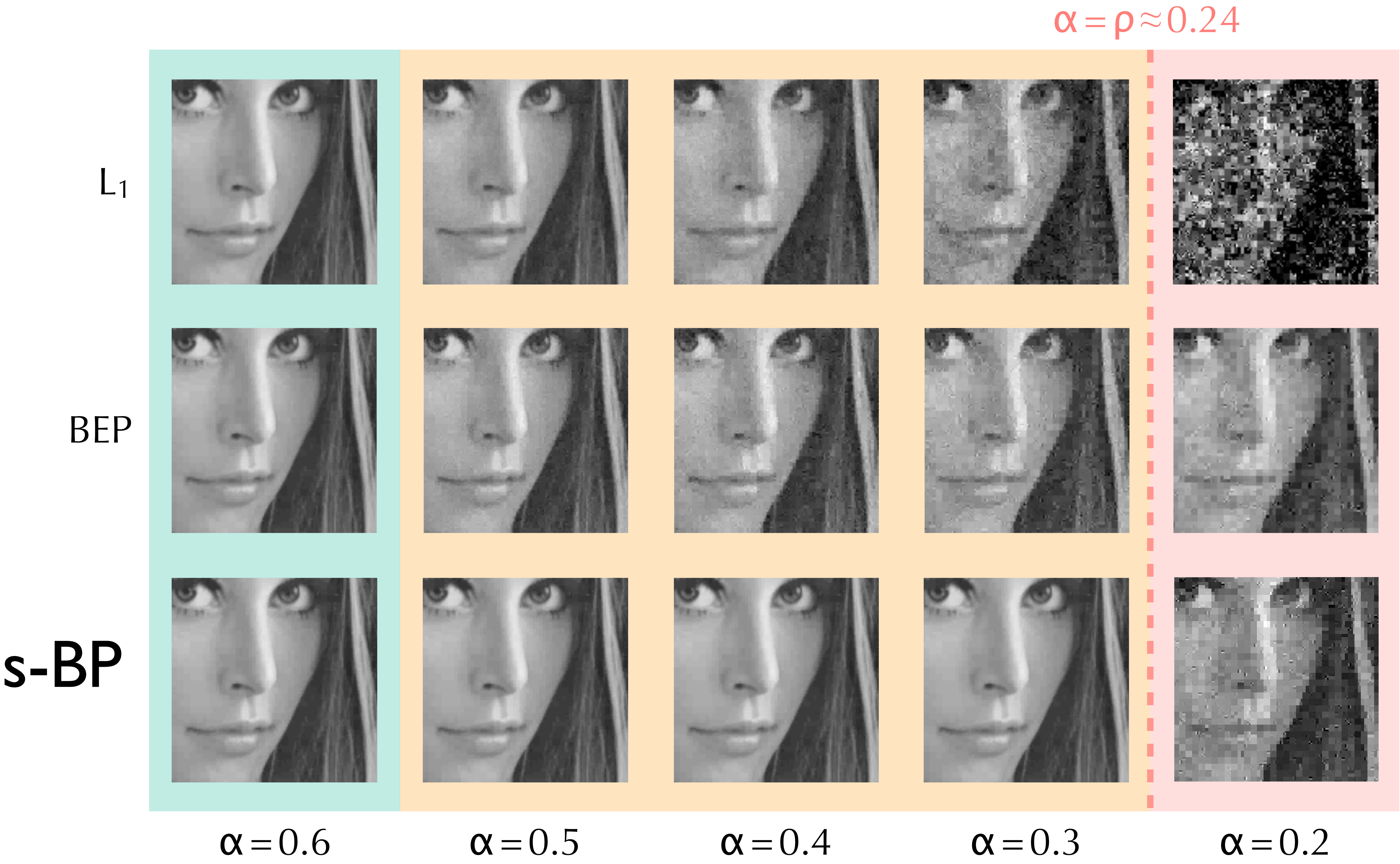


Gaussian signal

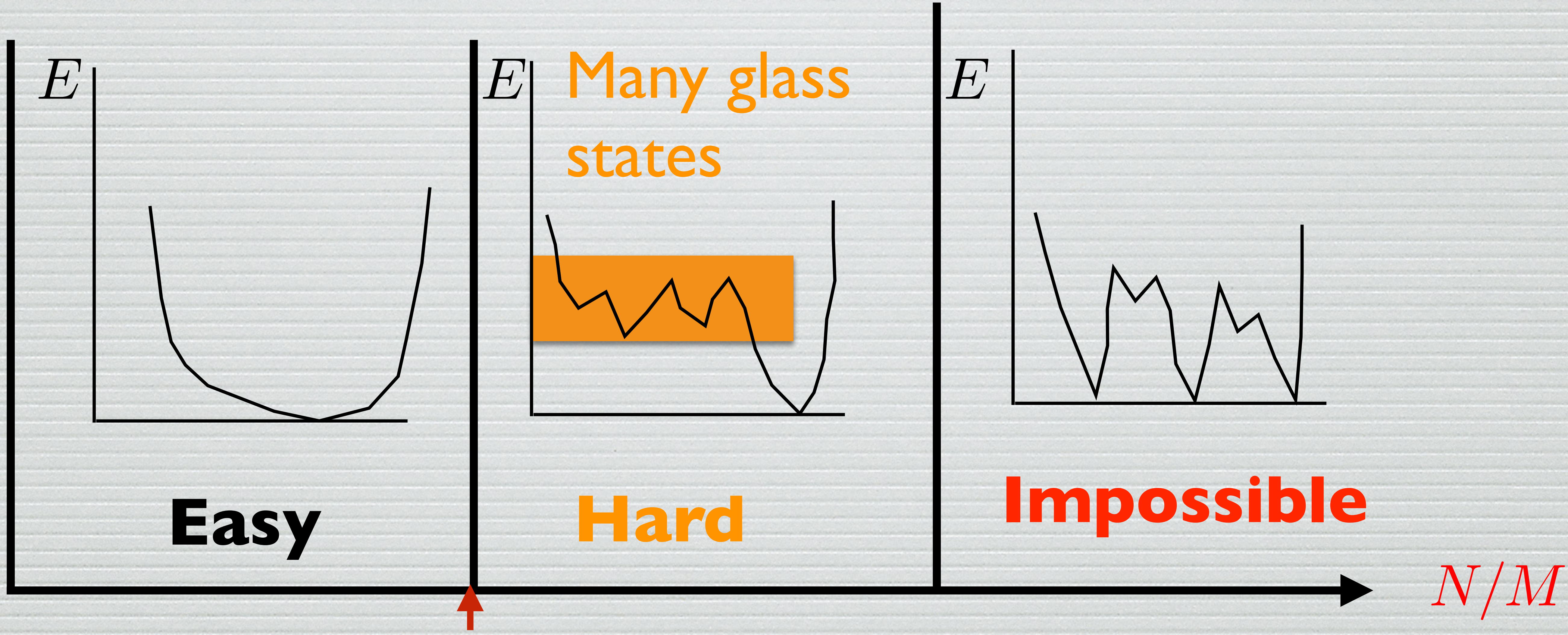
$$\phi_T(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

Binary signal

$$\phi_T(x) = \frac{1}{2}\left(\delta_{x,1} + \delta_{x,-1}\right)$$



α = ρ ≈ 0.24
L1
BEP
s-BP
α = 0.6
α = 0.5
α = 0.4
α = 0.3
α = 0.2



Phase transitions are crucial in large inference problems

Hard-Impossible = absolute limit (Shannon-like)

Easy- Hard = limit for large class of algorithms (local)

The spin glass cornucopia

A very sophisticated and powerful corpus of conceptual and methodological approaches has been developed (replicas, cavity, TAP,….) mostly in the years 1975-2000, and has found applications in many different fields of information theory and computer science

Portrait of Ottavio Strada,

Tintoretto, Venice 1567

Rijk's Museum Amsterdam

Thanks

Jean Barbier, Emmanuelle Gouillart, Yoshiyuki Kabashima, **Florent Krzakala**, Ayaka Sakata, François Sausset, Yifan Sun, **Lenka Zdeborova,** Pan Zhang,…

The spin glass cornucopia

A very sophisticated and powerful corpus of conceptual and methodological approaches has been developed (replicas, cavity, TAP,….) mostly in the years 1975-2000, and has found applications in many different fields of information theory and computer science

Portrait of Ottavio Strada,

Tintoretto, Venice 1567

Rijk's Museum Amsterdam