# Data mining the many-body problem

Congresso SIF,

16/9/21

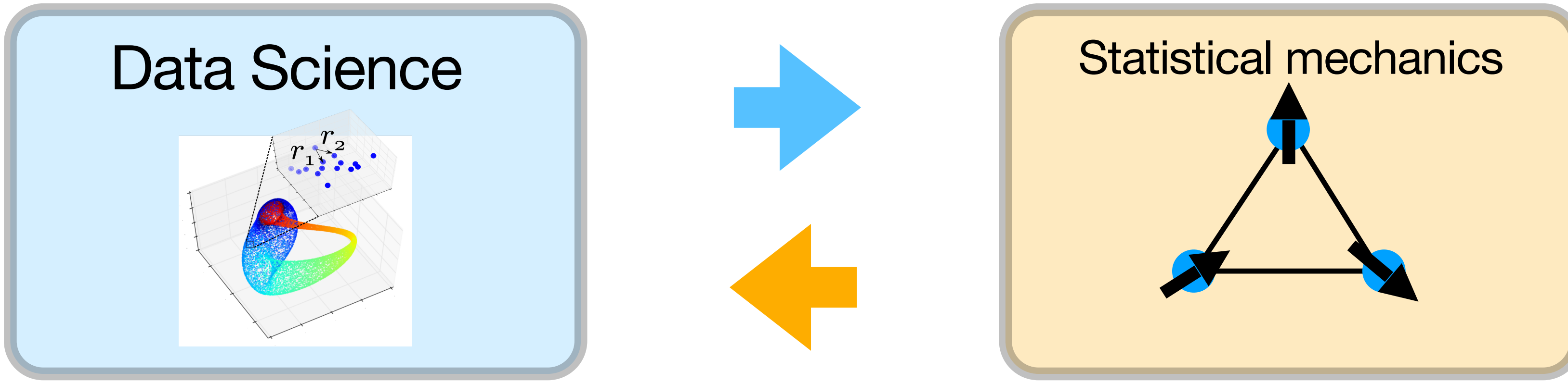
Marcello Dalmonte

ICTP&SISSA, Trieste

Joint work with A. Angelone, R. Fazio, T. Mendes-Santos, A. Rodriguez, X. Turkeshi

Based on: Phys. Rev. X **11**, 011040 and Phys. Rev. X Quantum **2**, 030332 (2021)

# Main idea




<u>Aim of the talk:</u>

- presenting an *informative* data mining viewpoint to fundamental objects of many-body theory
- Explore *universal* behavior of data sets

# Why are we planning to data mine a many body system?

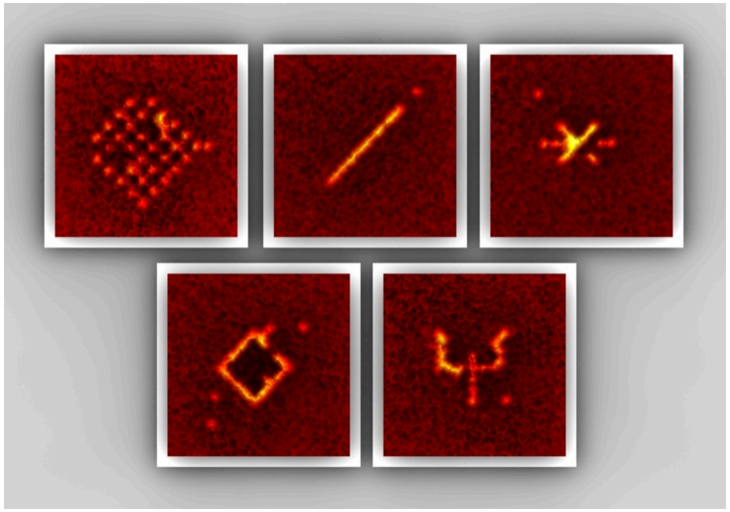
- Motivation 1: large amount of data structures are nowadays ubiquitous in both theory (Monte Carlo, stochastic simulations, etc.) and experiments
- "Motivation" 2: we now have methods to interpret those, that were not there in the past
- Motivation 3: several open questions can be addressed this way, similarly to other fields [theoretical chemistry, data science, spotify]

# Motivation 1: large data structures

**Experiments**: tons of data (amenable to statistical methods)

Example: wave function snapshots

Atom by atom pictures



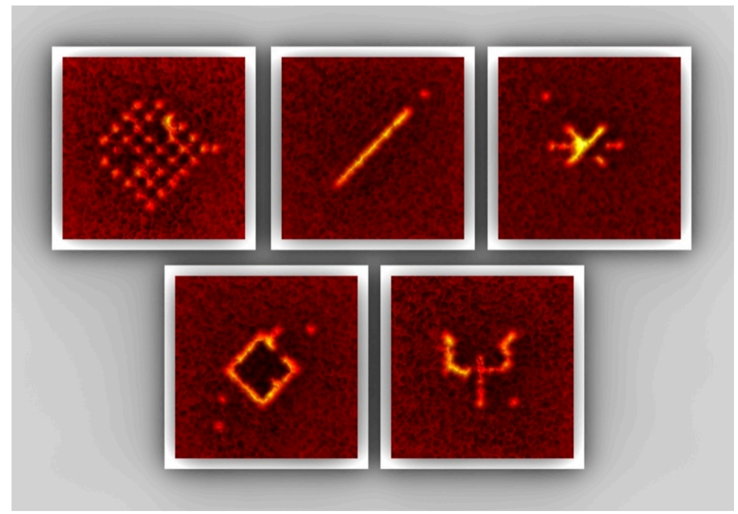
Credit: Bloch's group

# Motivation 1: large data structures

**Experiments**: tons of data (amenable to statistical methods)

Example: wave function snapshots

Atom by atom pictures



Credit: Bloch's group

**Challenge**: exploit naturally available many-body correlations!

# Motivation 1: large data structures

**Theory**: outcome of numerical experiments

$$Z = e^{-\beta H}$$

Example: Monte Carlo simulations (CM, HEP, etc.)

$$X = \{X_1, X_2, ...X_N\}$$

Elements of a Markov chain

# Motivation 1: large data structures

**Theory**: outcome of numerical experiments

$$Z = e^{-\beta H}$$

Example: Monte Carlo simulations (CM, HEP, etc.)

$$X = \{X_1, X_2, \ldots X_N\}$$

Elements of a Markov chain

**Challenge**: agnostic interpretation / e.g., without assumptions

# Outline

A basic tool to data mine datasets: the **intrinsic dimension**

Data mining partition functions:

- Basic idea: a 3-site toy model
- Data mining and classical critical behavior: **universal data sets and emergent simplicity**

Data mining path integrals

Conclusions

# The intrinsic dimension

**The intrinsic dimension**

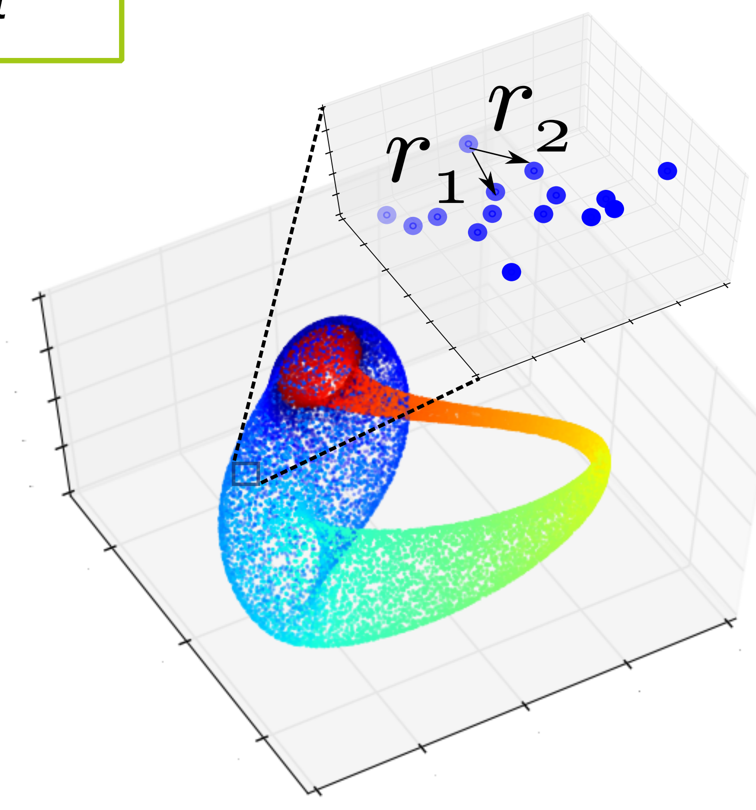
*the minimal number of variables required to describe a dataset*

Ex. 1: Klein bottle

Original data in $D = 3$

$$I_d = 2 < 3$$

Operational meaning: one just needs 2 independent variables to properly describe the object of interest



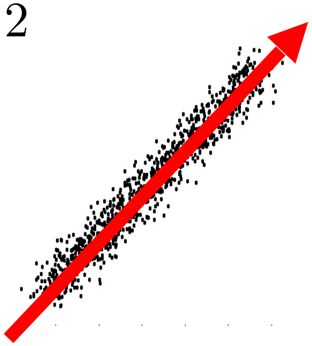
# The intrinsic dimension

**The intrinsic dimension**

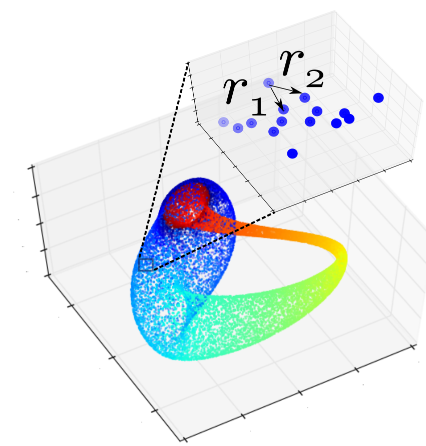
*the minimal number of variables required to describe a dataset*

Widely applied in ab initio molecular dynamics (Laio, Rodriguez, etc.)

$D = 2$



$I_d = 1$




$D = 3$

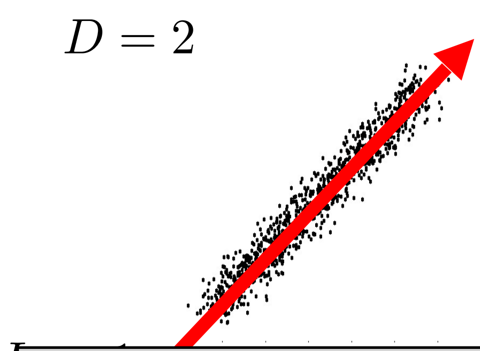
$I_d = 2$

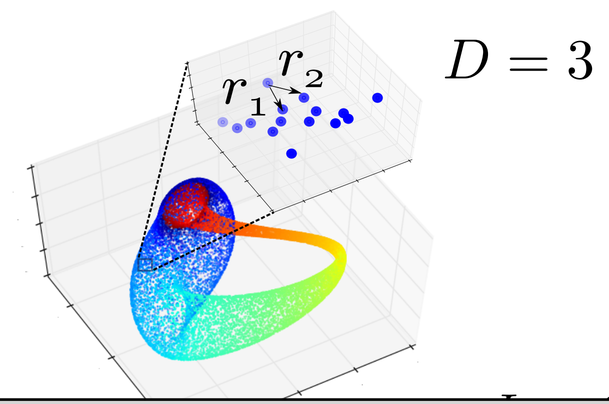
# The intrinsic dimension

**The intrinsic dimension**

*the minimal number of variables required to describe a dataset*

Widely applied in ab initio molecular dynamics (Laio, Rodriguez, etc.)




Intuition: the intrinsic dimension is informative about
(1) number of relevant degrees of freedom, and
(2) **complexity** of a manifold

# A basic example: a 3 site model

Simple example: 3-site Heisenberg model:

$$N_r \simeq 10^4/10^5$$

1) States

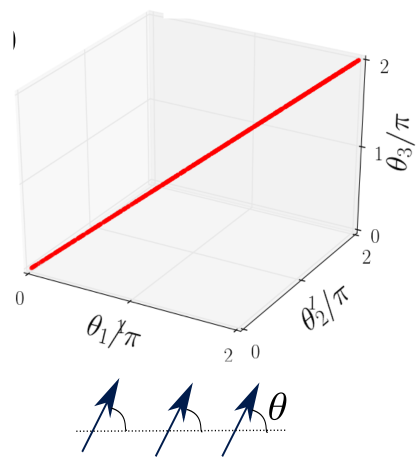
$$\vec{x} = (\vartheta_1, \vartheta_2, \vartheta_3)$$

2) Equilibrium weight:

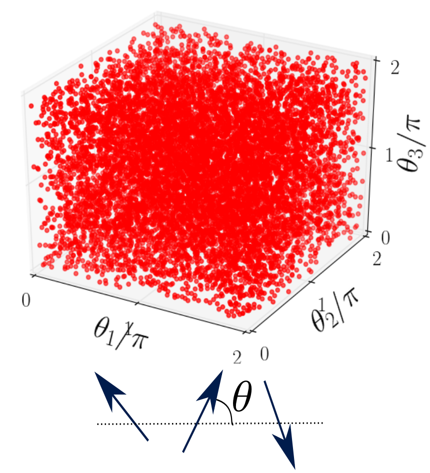
$$\rho(E) \sim e^{-E(\vec{x})/T}$$

3) Hamiltonian:

$$E(\vec{\theta}) = -\sum_{\langle i,j \rangle} \vec{S}_i \cdot \vec{S}_j,$$




$$I_d = 1$$




$$I_d = 3$$

# A basic example: a 3 site model

Simple example: 3-site Heisenberg model:

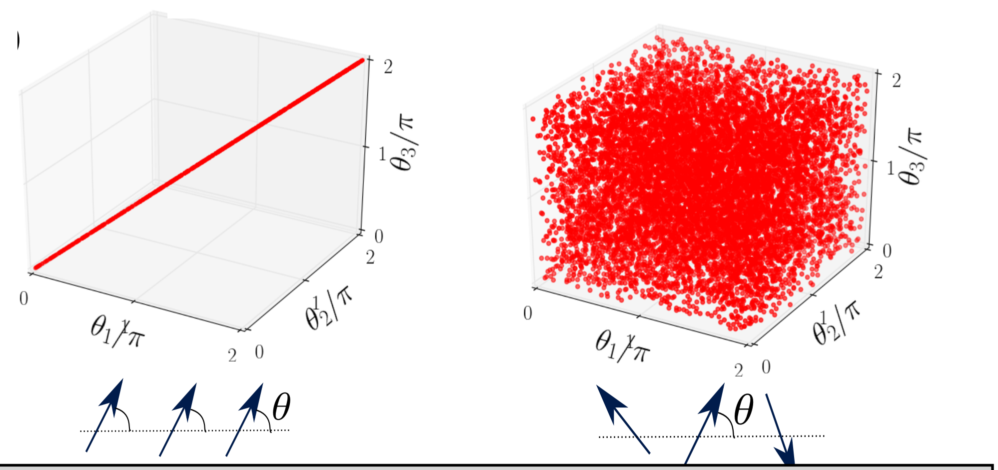
$N_r \simeq 10^4/10^5$

1) States

$$\vec{x} = (\vartheta_1, \vartheta_2, \vartheta_3)$$

2) Equilibrium weight:

$$\rho(E) \sim e^{-E(\vec{x})/T}$$




3) Hamiltonian:

Message 1: Intrinsic dimension can be informative about phases - but what about critical behavior?

# Many-body: the Ising model in 2D

Hamiltonian

$$E(\vec{s}) = -\sum_{\langle i,j \rangle} s_i s_j ,$$

Data configurations

$$\vec{s} = (s_1, s_2, ..., s_{N_s})$$

Second order (conformal) phase transition at

$$T_c = 2/\ln(1+\sqrt{2}) \simeq 2.26...$$

$$\nu = 1$$

Data structure: $s_1 = (s_{11}, s_{12}, ..) = (0, 1, 0, 1, 1, 1, 0, ...)$

Distance between points:
Hamming distance

$$r(\vec{s^i}, \vec{s^j}) = \sum_p |s_p^i - s_p^j|$$

# Emergence of scale-free network

Hamiltonian

$$E(\vec{s}) = -\sum_{\langle i,j \rangle} s_i s_j \,,$$

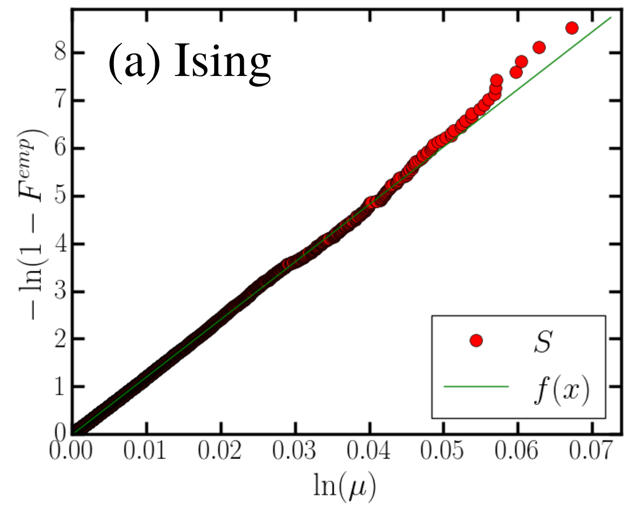
Data configurations

$$\vec{s} = (s_1, s_2, ..., s_{N_s})$$

Second order (conformal) phase transition at

$$T_c = 2/\ln(1+\sqrt{2}) \simeq 2.26...$$

$$\nu = 1$$




$$I_d = -\frac{\ln\left[1 - P(\mu)\right]}{\ln\left(\mu\right)}$$

$$N_r \simeq 10^4/10^5$$

# Emergence of scale-free network

Hamiltonian

$$E(\vec{s}) = -\sum_{\langle i,j \rangle} s_i s_j ,$$

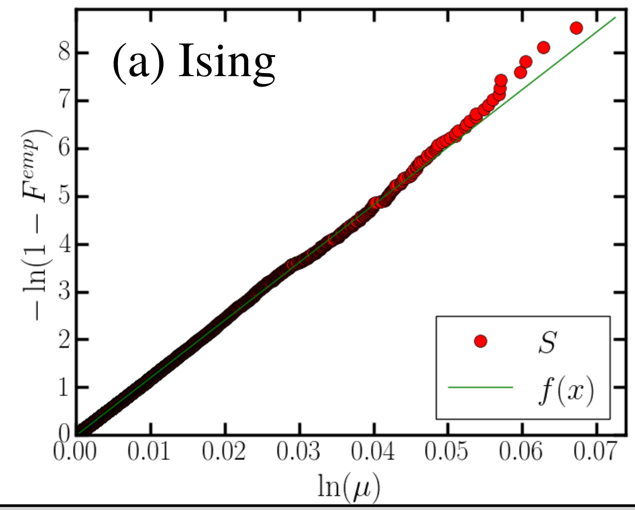
Data configurations

$$\vec{s} = (s_1, s_2, ..., s_{N_s})$$

Second order (conformal) phase transition at

$$T_c = 2/\ln(1+\sqrt{2}) \simeq 2.26...$$

$$\nu = 1$$




Forget about details: if straight line, dataset is parametrized by a Pareto distribution

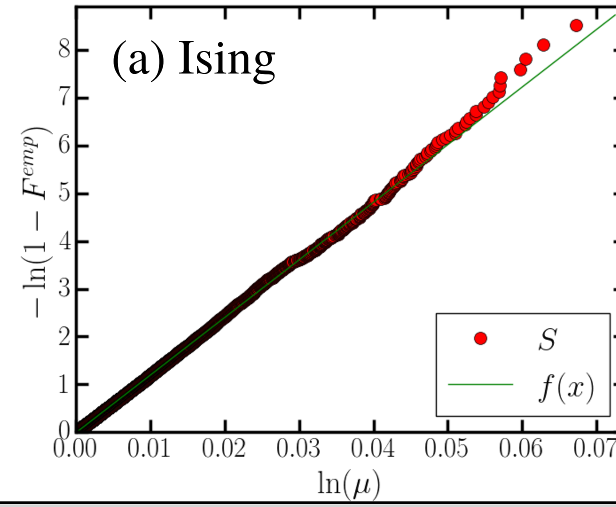
# Emergence of scale-free network

Hamiltonian

$$E(\vec{s}) = -\sum_{\langle i,j \rangle} s_i s_j \,,$$

Data configurations

$$\vec{s} = (s_1, s_2, ..., s_{N_s})$$




Second order (conformal) phase transition at

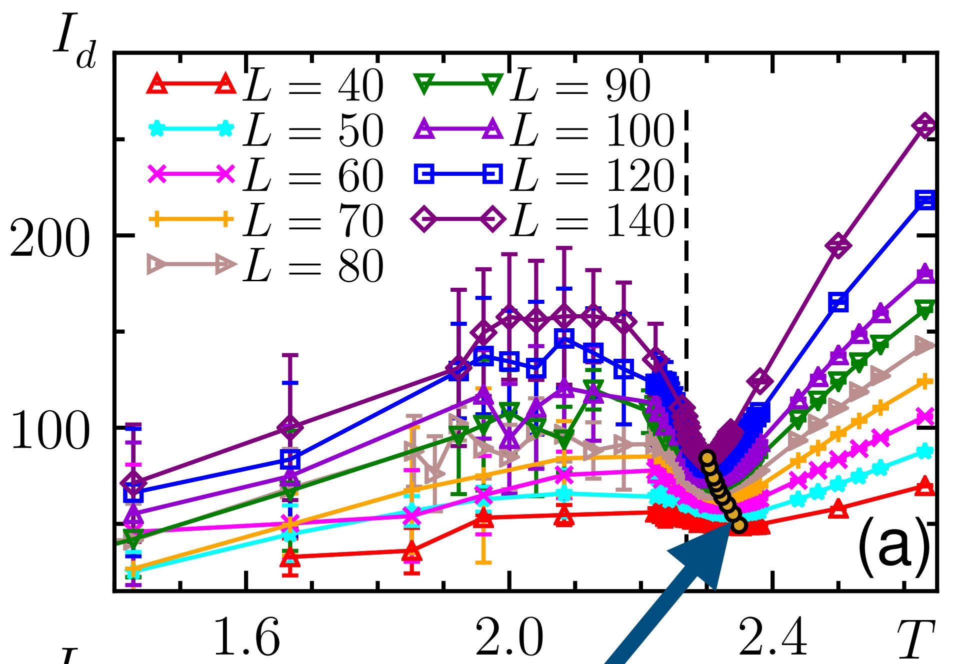
$T_c$

Forget about details: if straight line, dataset is

to

First observation of scale free network in the present context - those have widespread applications in many fields (neural networks, internet, etc.)

# Intrinsic dimension: emergent simplicity




**emergent simplicity**:
manifold simplifies at transition points!

# Intrinsic dimension: universal behavior

Universal collapse scaling from renormalization group

Scaling function

$$I_d = L^{\zeta} \boxed{f(\xi/L)}$$

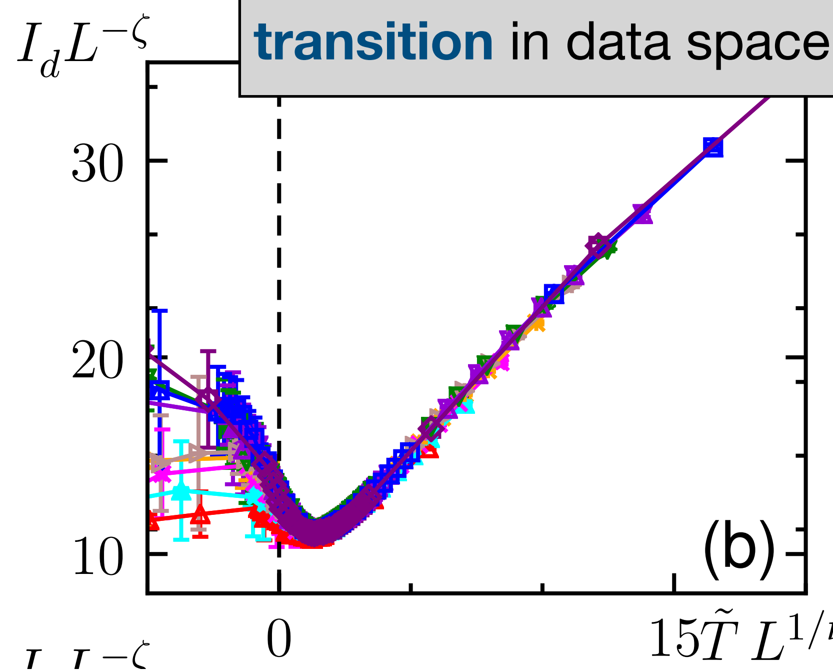
$$\xi \sim (T - T_c)^{-\nu}$$

Second order (conformal) phase transition at

$$T_c = 2/\ln(1+\sqrt{2}) \simeq 2.26...$$

$$\nu = 1$$

**Universality** of data structure: **structural transition** in data space




Free parameters: $T_c, \nu, \zeta$

$$T_c = 2.283(2), \nu = 1.02(2), \zeta = 0.410(5)$$

Review: A Pelissetto & E Vicari, Physics Reports (2002)

# Intrinsic dimension: quantitative predictions

Universal collapse scaling from renormalization group

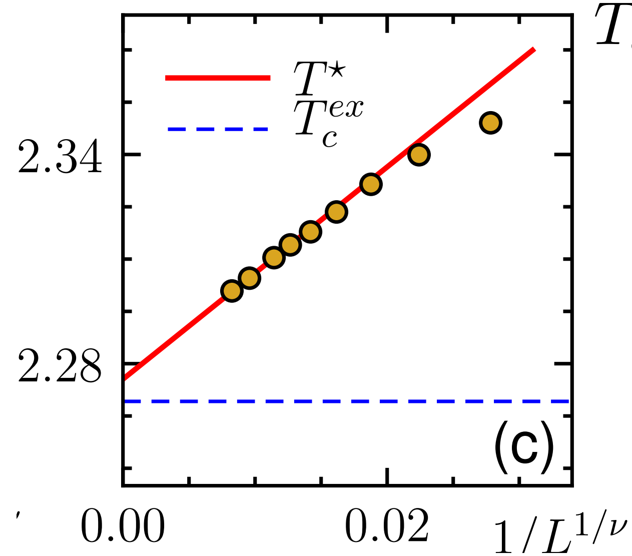
$$T^*(L) - T_c \sim \frac{1}{L^{1/\nu}}$$

Second order (conformal) phase transition at

$$T_c = 2/\ln(1+\sqrt{2}) \simeq 2.26...$$

$$\nu = 1$$

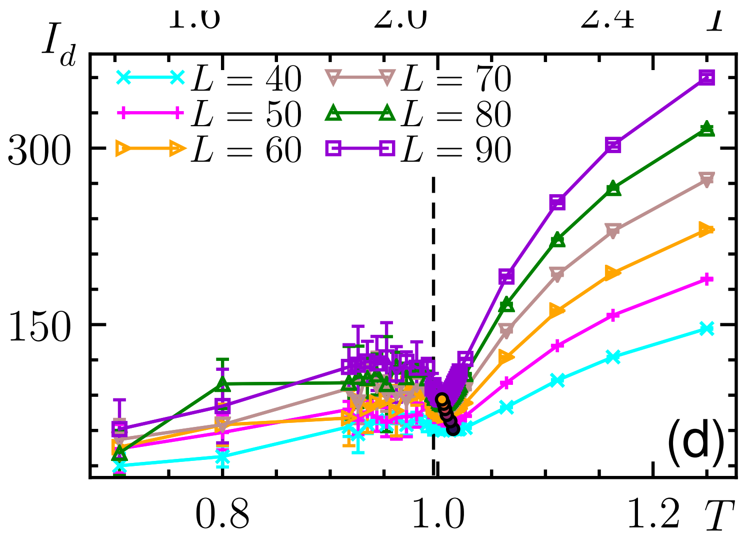
Useful to find **critical temperature and exponents** *without* any assumption on order parameters!



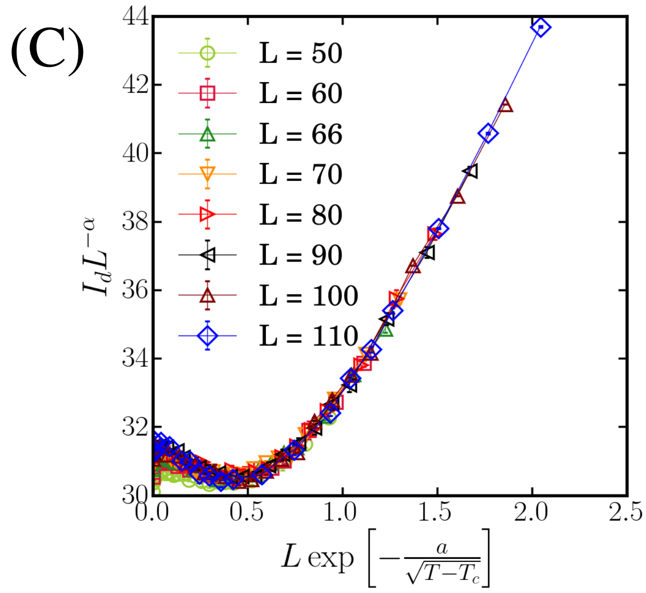

$$T_c = 2.2784(2)$$

# Kaleidoscope of applications

Same behavior for other second order and topological transitions




3-state Potts model




XY model

# Main messages

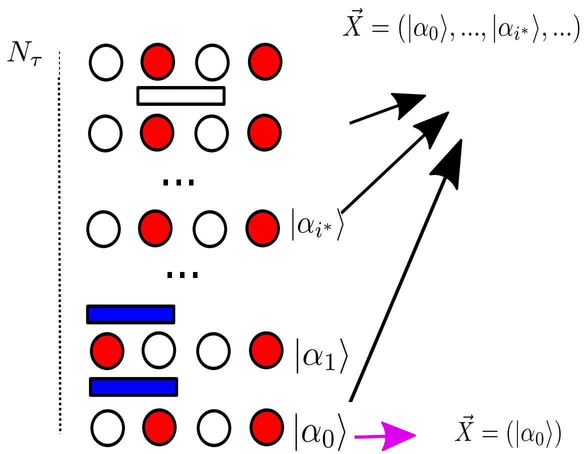
Partition functions of classical many-body models remarkably describe scale free-networks

Such scale free-networks exhibit emergent simplicity at transition points

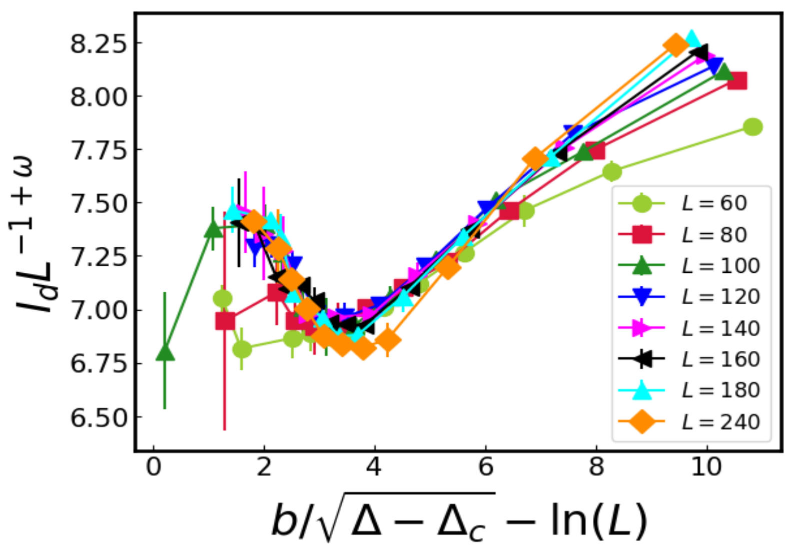
The data structures undergo a transition as long as the model does, with the same critical exponents!

# What's next?

Application to quantum mechanics: Monte Carlo methods again




1D Heisenberg model




Also quantum structures seem to exhibit the same universal behavior!

# What's next?

Other perspectives:

- experiments on quantum matter: working of a quantum computer and simulator
- variational Monte Carlo methods (first results with Jastrow unclear)
- reverse engineer our finding to discover universal data structures in, e.g., neural networks

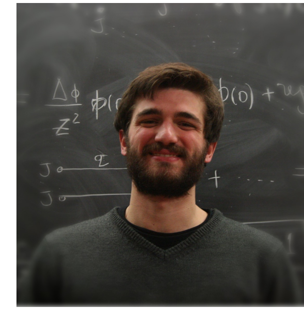
ICTP and SISSA

Adriano Angelone (-> Paris)

Saro Fazio

Tiago Mendes-Santos (-> MPIPKS Dresden)

Alex Rodriguez



Xhek Turkeshi

# Thank you!

Mendes-Santos, Turkeshi, MDM, Rodriguez, Phys. Rev. X 11, 011040 (2021)
Mendes-Santos, Angelone, Rodriguez, Fazio, MDM, Phys. Rev. X Quantum 2, 030332 (2021)