TOPOLOGICAL ALGORITHMIC FIELD THEORY OF DATA

A possible new venue for data mining

Mario Rasetti

ISI Foundation - Torino

ISI Global Science Foundation – New York

INSTITUTE
FOR SCIENTIFIC INTERCHANGE
FOUNDATION

The ubiquitous complex systems

Complex, multi-level, multi-scale systems are everywhere:

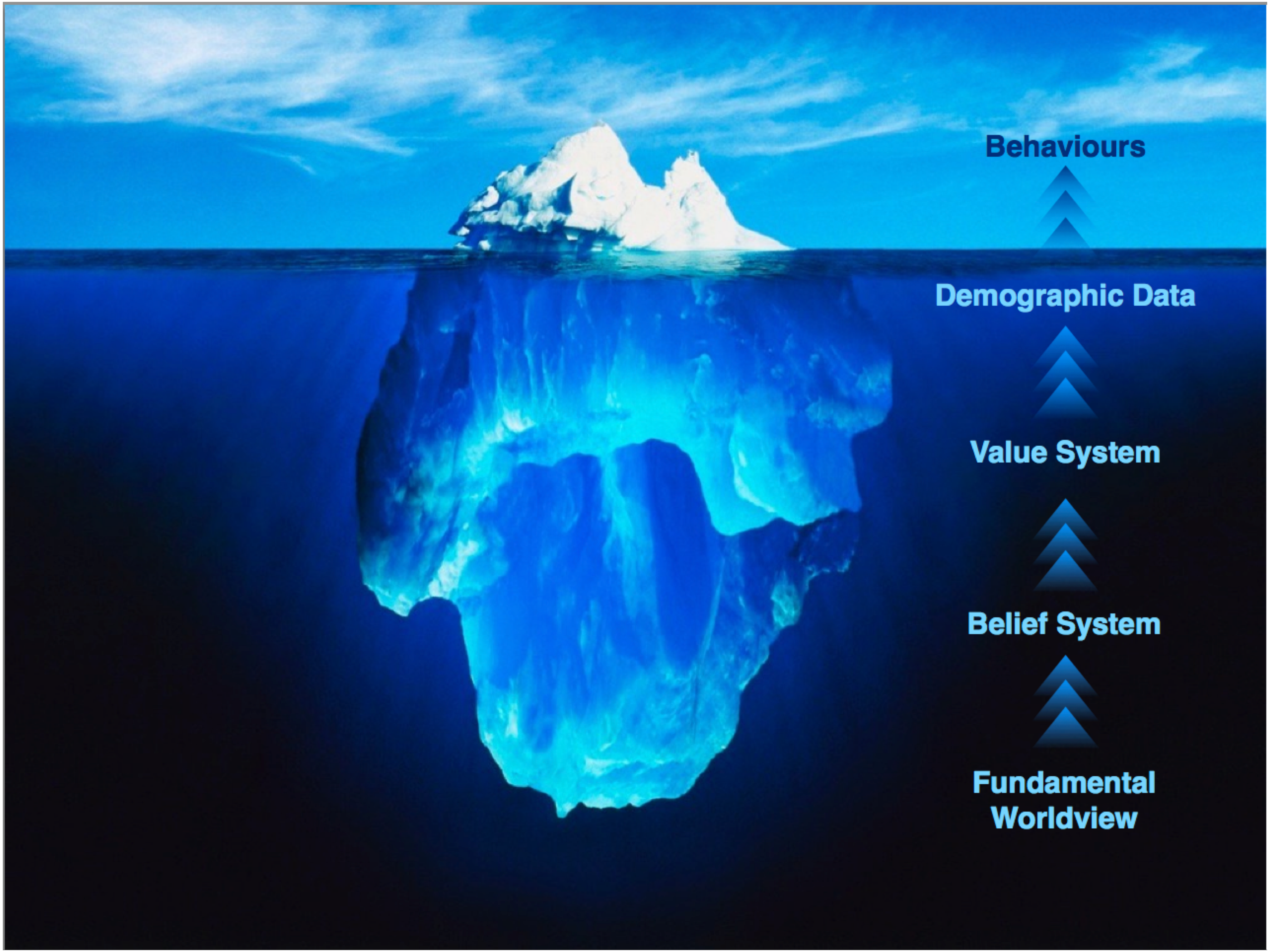
in NATURE ,
but also the Internet, the brain, the climate, the spread of pandemics, economy and finance:
in SOCIETY

The question was asked

can we envisage the construction of a

Complexity Science Theory ?

Does it have sense thinking of a conceptual construct for complex systems playing the same role that statistical mechanics played for thermodynamics ?



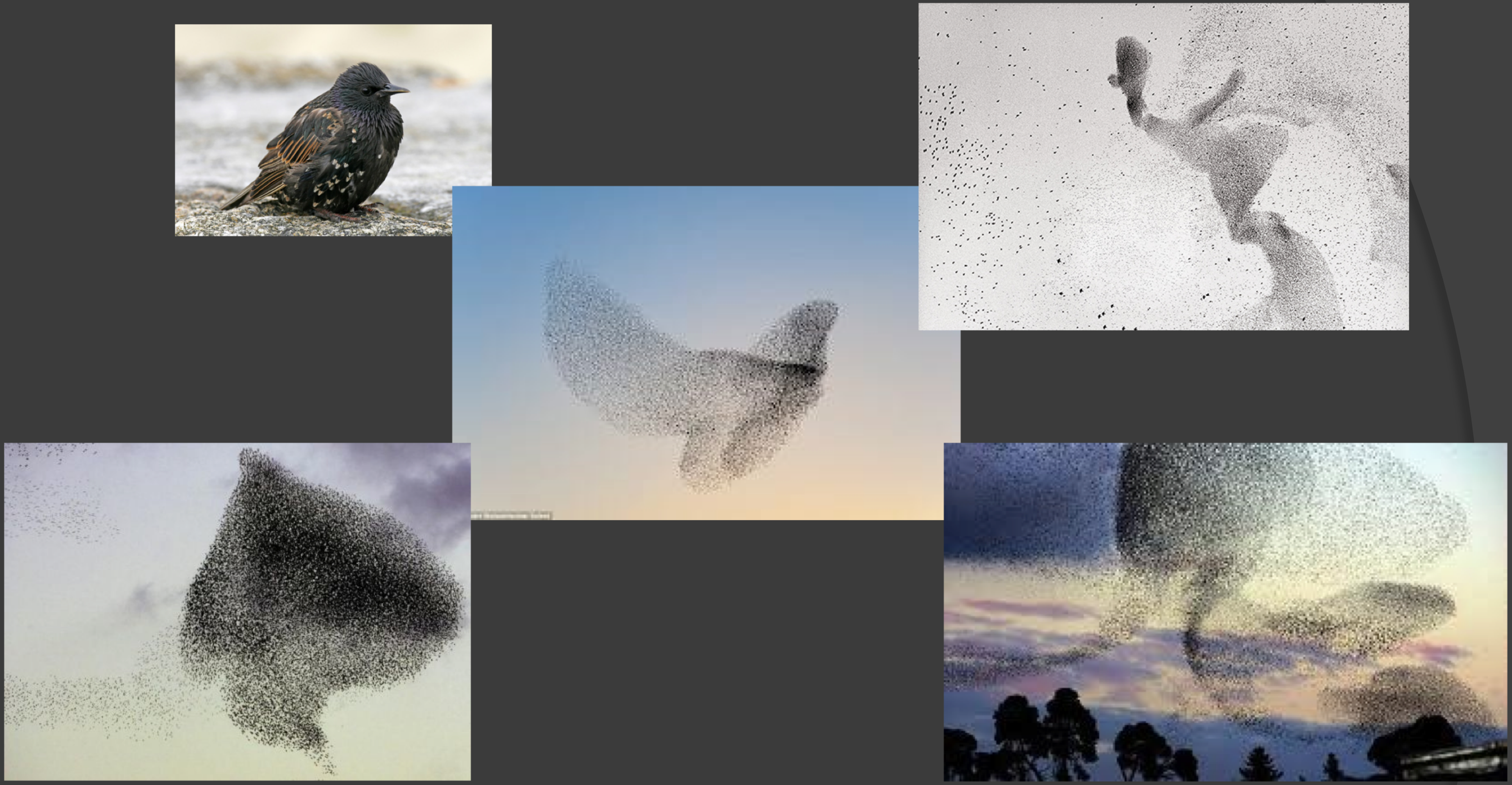
Behaviours

Demographic Data

Value System

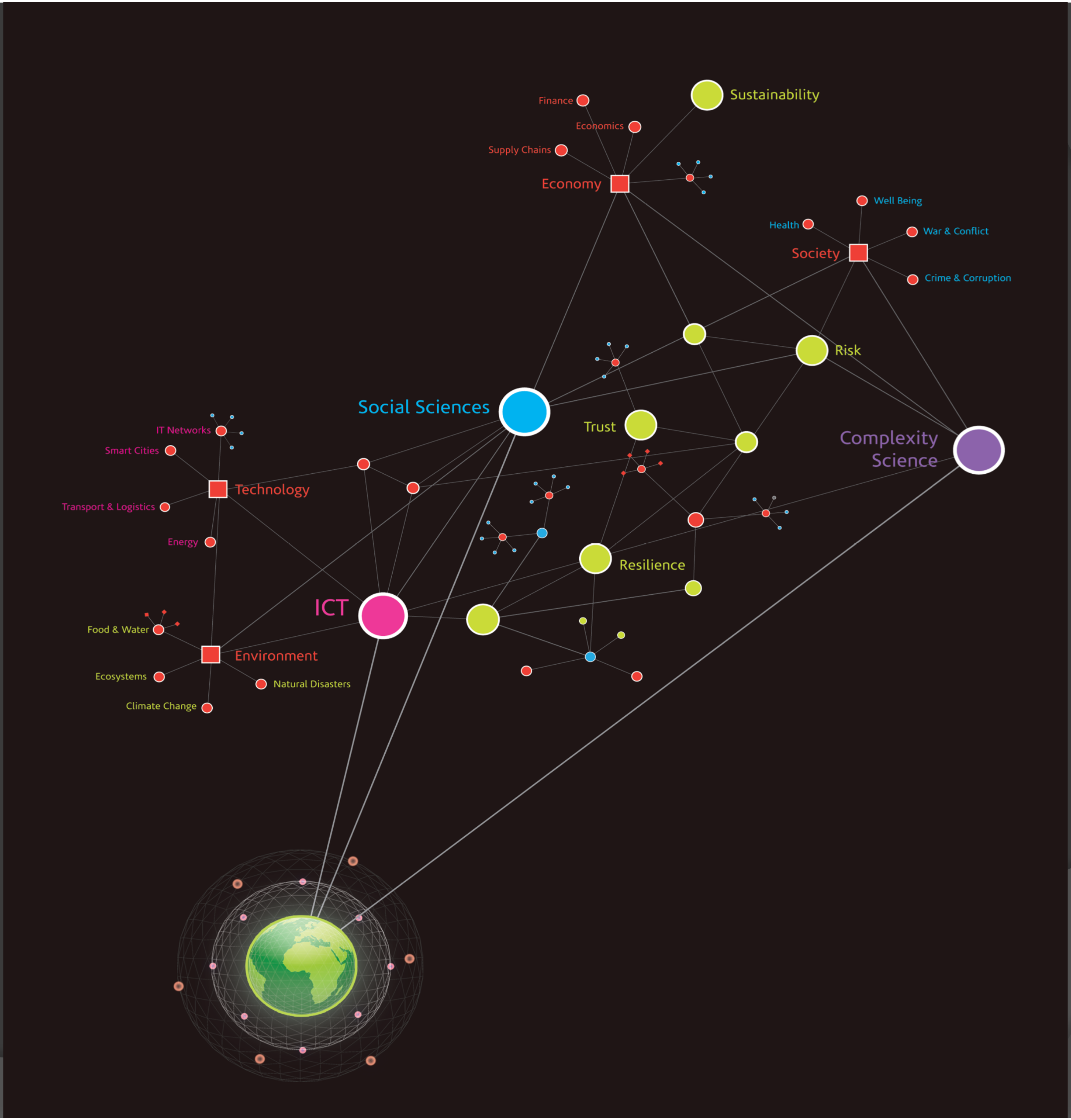
Belief System

Fundamental Worldview



Bevys of Starlings

Multi-agent – Multi-scale – Emergent effects


Finance
Sustainability
Economics
Supply Chains
Economy
Well Being
Health
War & Conflict
Society
Crime & Corruption
Risk
Social Sciences
Trust
Complexity Science
IT Networks
Smart Cities
Technology
Transport & Logistics
Energy
Resilience
ICT
Food & Water
Environment
Ecosystems
Natural Disasters
Climate Change

The challenge is enormous:

In Statistical Mechanics we assume

- Ergodicity (a property shared with number theory: reals *vs.* integers or rationals !)
- Thermodynamic limit (N → ∞)
- Identical / indistinguishable particles
- Repeatable phenomenology;

“experiment – based”

- Analytical structure of underlying dynamics (only phase transitions = singularities)

Only in this way stat-mech is able to provide a reliable representation of several facts of thermodynamics

On the contrary typically complex systems are:

- NOT Ergodic
- Their number of agents N = FINITE (even though it can be quite large)
- Agents are NOT identical (they are distinguishable complex systems themselves, with their strategies)
- They are NEVER representable by analytic (perhaps not even recursive) functions

and, above all,

- They are DATA-based; typically NO repeatable experiment is possible, in reductionist sense

We concentrate on the latter feature

Big Data = Big Challenge

of dealing with the huge amount of information flowing in and around complex systems, endowing ICT with new, more efficient tools to play a role in turning

data into **information**,

information into **knowledge**,

and eventually

knowledge into **wisdom**.

The ever-more blurred boundaries between digital and physical worlds will thus progressively fade away, as ICT becomes an integral part of the fabric of nature and society.

Our aim is to explore whether we can tame Big Data with Topology (the geometry of 'shapes')

Fundamental notion, from computer science when dealing with data, is the concept of

'**Space of Data**' :

- the structure in which information is encoded;
- the frame for algorithmic (digital) thinking;
- the lode where to perform Data Mining, i.e., to extract patterns of information

This is the task : find new ways of mining data spaces for information resorting to geometrical (indeed topological, combinatorial) methods.

We claim that topology is the natural tool to handle large, high-dimensional, complex spaces of data

Why? Because :

Qualitative information is relevant : data users aim to obtain *knowledge*: understand how data is organized on large scale. Global, even though partly qualitative, information is what is needed.

Metrics are not theoretically justified : physical phenomena support theories that tell us exactly what metric to use; in the life science or social sciences this is fully uncertain.

Coordinates are not natural : data is typically transmitted in the form of ‘vectors’ (strings of symbols, typically numbers in some field ←→ Gödel numbers), but the ‘components’ or linear combinations of these vectors are not natural in any sense: the space of data is <u>not</u> a vector space

Summaries are most valuable : the conventional method of handling data is *building* a graph (**network**) whose vertex set are data and two points are connected by an edge if their ‘proximity measure’ (in the sense of Grothendieck topology) is, say, $\leq \eta$, and then try and determine the optimal choice of η.

It is however much more informative to consider the entire dendrogram, getting at once a summary of its relevant features under all possible values of η and try and find a way to know how the global features of data space vary changing η.

Data & Topology

If Topology is the natural tool to handle large, high-dimensional, sets of data, how do we deal with it ?

The pillar of computation logic is the Church-Turing Thesis :

> < Any well-defined procedure that can be grasped and performed by the human mind and 'pencil/paper', can be performed on a conventional digital computer with no bound on memory. >

This is NOT a theorem; it is a statement of belief concerning the universe we live in.

There are several intuition-based approaches to Church Turing thesis :

Empirical Intuition

No one has ever given a concrete example of process that humans can compute in a consistent and well defined way, yet cannot be programmed on a computer.
The thesis is true.

Mechanical Intuition

The brain is a machine whose components obey physical laws. As such, in principle, a brain can be simulated on a digital computer, and its functions can be computed by a simulating computer.
The thesis is true.

Quantum Intuition

The brain is a machine, but not a classical one. It has quantum mechanical features, hence there are inherent barriers to its being simulated on a digital computer.
The thesis is false. However, it remains true if we allow for quantum computers.

‘Beyond Turing’ Intuition

The brain is inherently a quantum computing machine but able to compute ‘uncomputable’ functions.
The thesis is false. A new tool is needed: the (quantum) Gandy machine. A Gandy machine is equivalent to a set (finite? countably infinite? uncountable?) of interacting Turing machines: (Topological ≅ non-linear) Quantum Field Theory is necessary.

For all these reasons the methods to adopt should be inspired by *topology*, because :

Topology is the branch of mathematics that deals with qualitative geometric information about a space (connectivity, classification of loops and higher dimensional manifolds, invariants).

Topology – contrary to geometry – studies geometric properties in a way insensitive to metrics : it ignores distance function and replaces it with some measurable notion of 'connective nearness', i.e., *proximity* (η ; think of 'hand-shake')

Topology deals with those properties of geometric objects that do not depend on coordinates but only on intrinsic geometric features. It is *coordinate-free*.

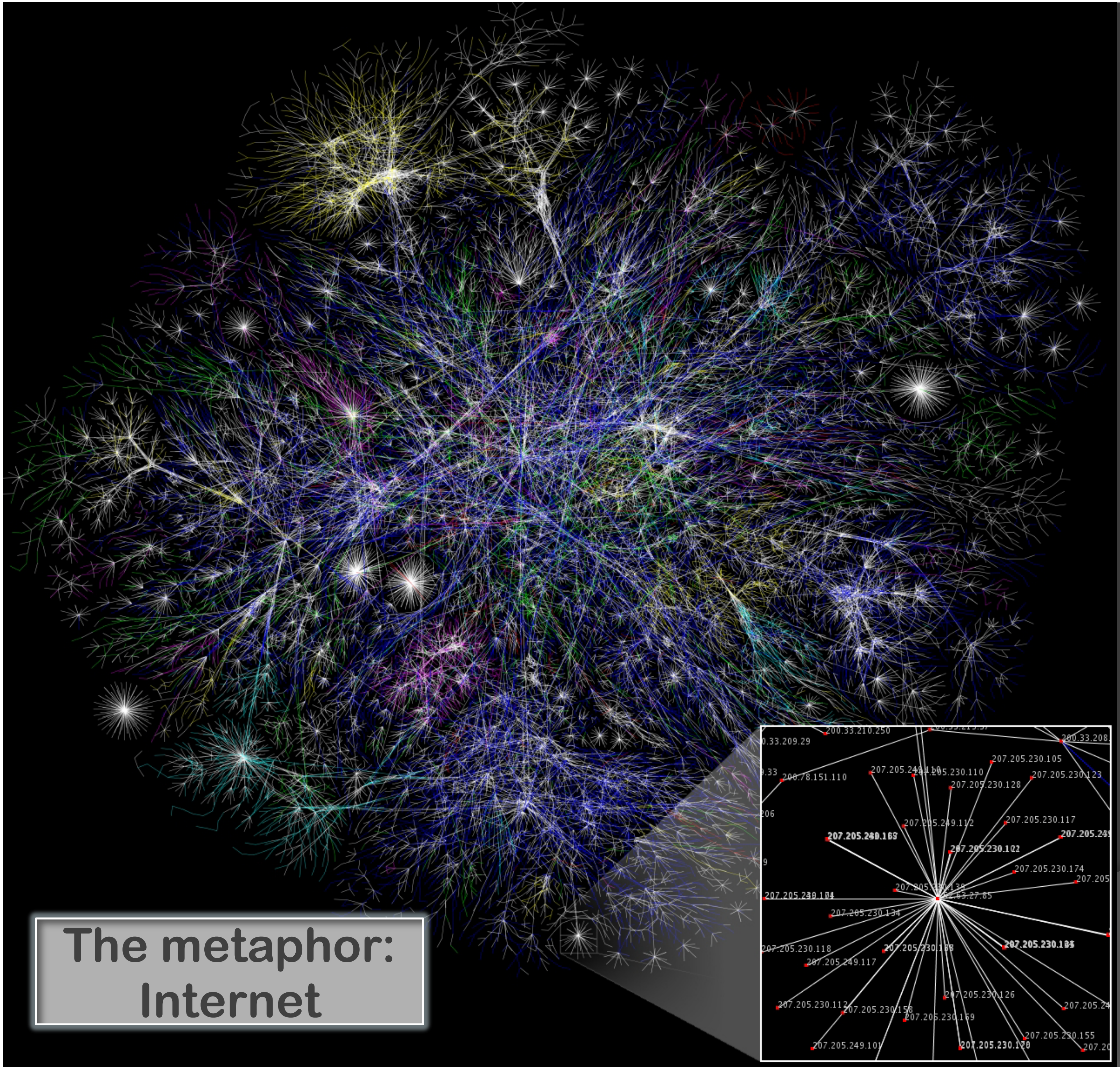
In *Topology* relationships involve maps between objects: they are a manifestation of *functoriality*. Moreover, invariants are related not just to objects but to the maps between them. Functoriality reveals a *categorical* structure enabling the computation of global invariants from local information.

Full information about topological spaces is inherent in their *simplicial representation*, a piece-wise linear, combinatorially complete, discrete realization of functoriality.

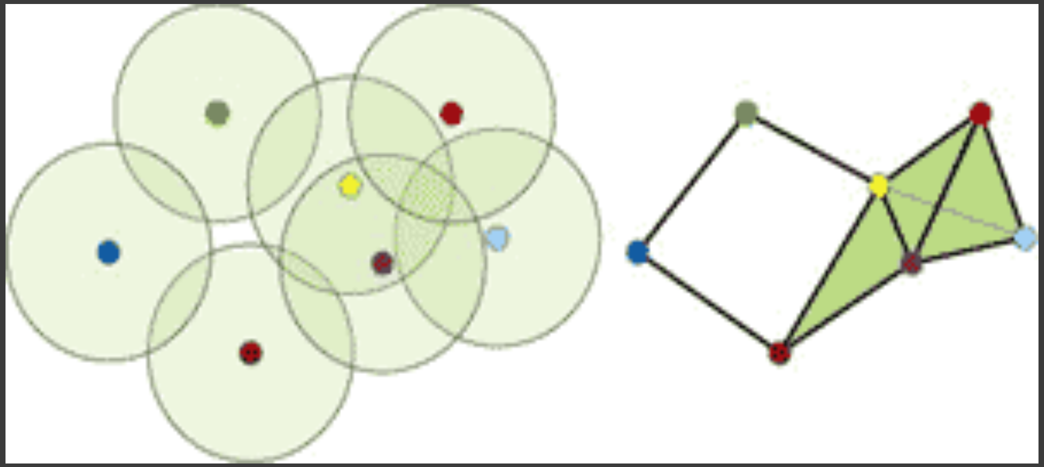
The conventional way to convert the collection of points of data space into a structured object is to use the point cloud X as *vertex set* of a *graph* Θ (a NETWORK), whose edges are determined by proximity.

Θ captures well data connectivity (local), but ignores a wealth of higher order (global) features, well discerned instead considering its completion to a higher-dimensional object, of which Θ is the 1-skeleton: the simplicial complex S≈X, PL space built of simple pieces (simplices) identified combinatorially along their faces.

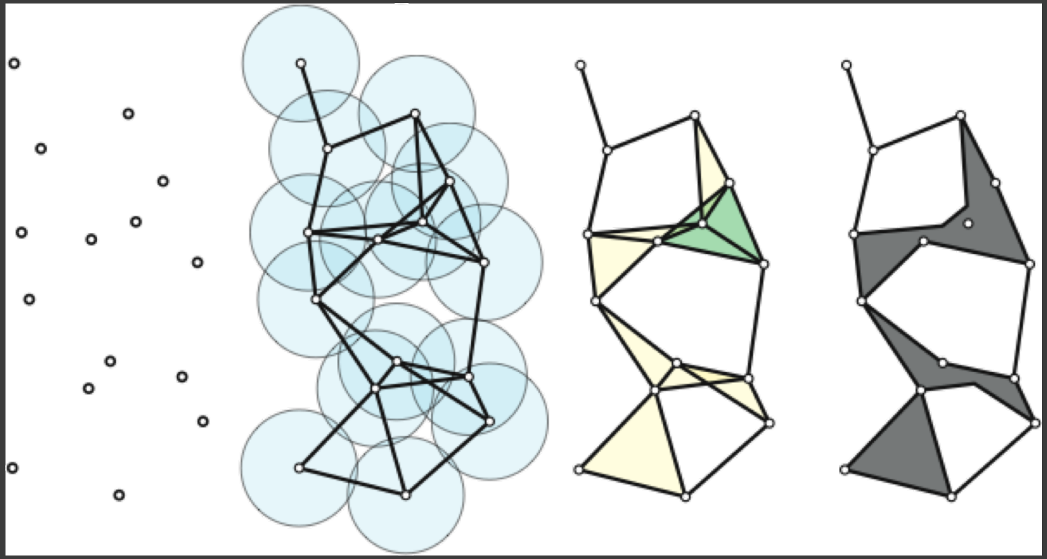
Natural complexes are: i) the Čech complex (k-simplices are ($k+1$)-tuples of points whose $\eta/2$-ball neighborhoods intersect); ii) the Rips-Vietoris complex, whose k-simplices are ($k+1$)-tuples of points pairwise within distance η.



The metaphor: Internet



Čech complex



Rips-Vietoris complex

The necessary tools are :

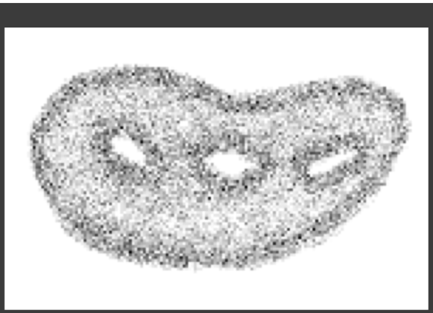
- Persistent homology : to disentangle the global complexity of data sets; (signal *vs*. noise)
- Measure theory / filtration : to perform the process

  graphs (networks) → *simplicial complexes*

  and statistically weight the emerging structure.
- Formal methods : languages – groups – automata; to disentangle the hierarchical structural architecture and lead to the separation of the behavioral level from the structural level (the S[B] system)

Data as a point cloud

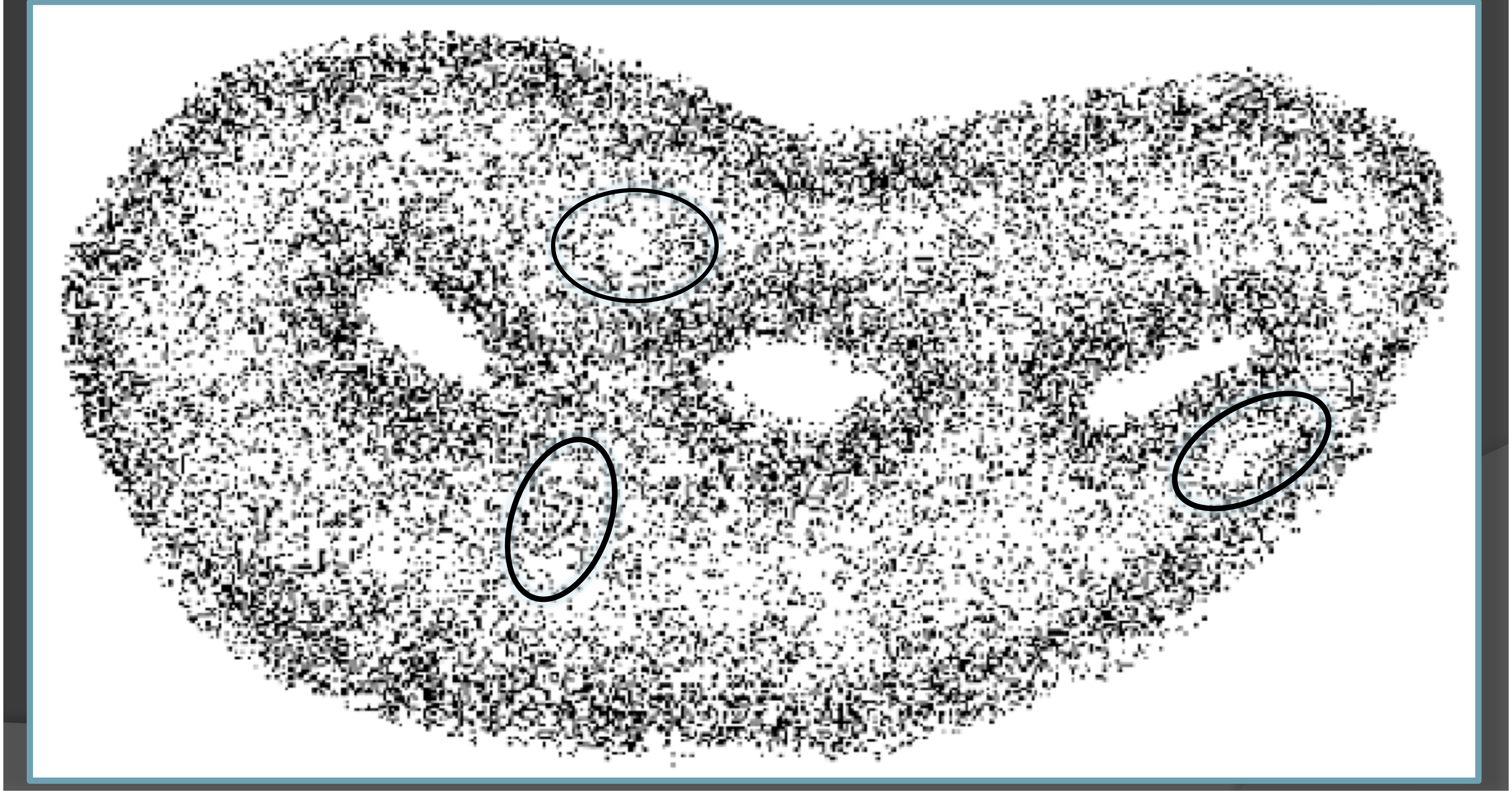
Consider this object:

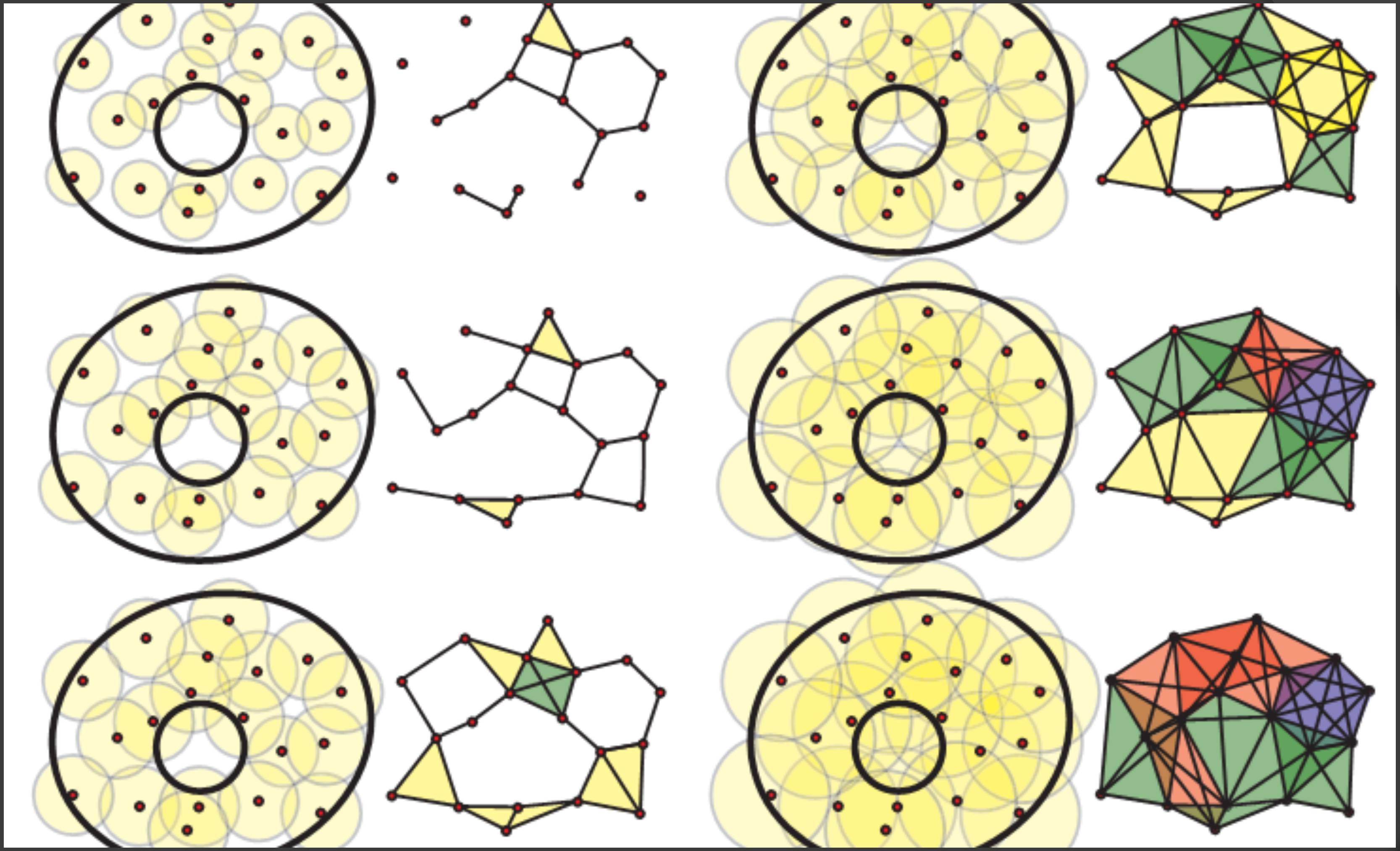


It appears to be a triple torus.

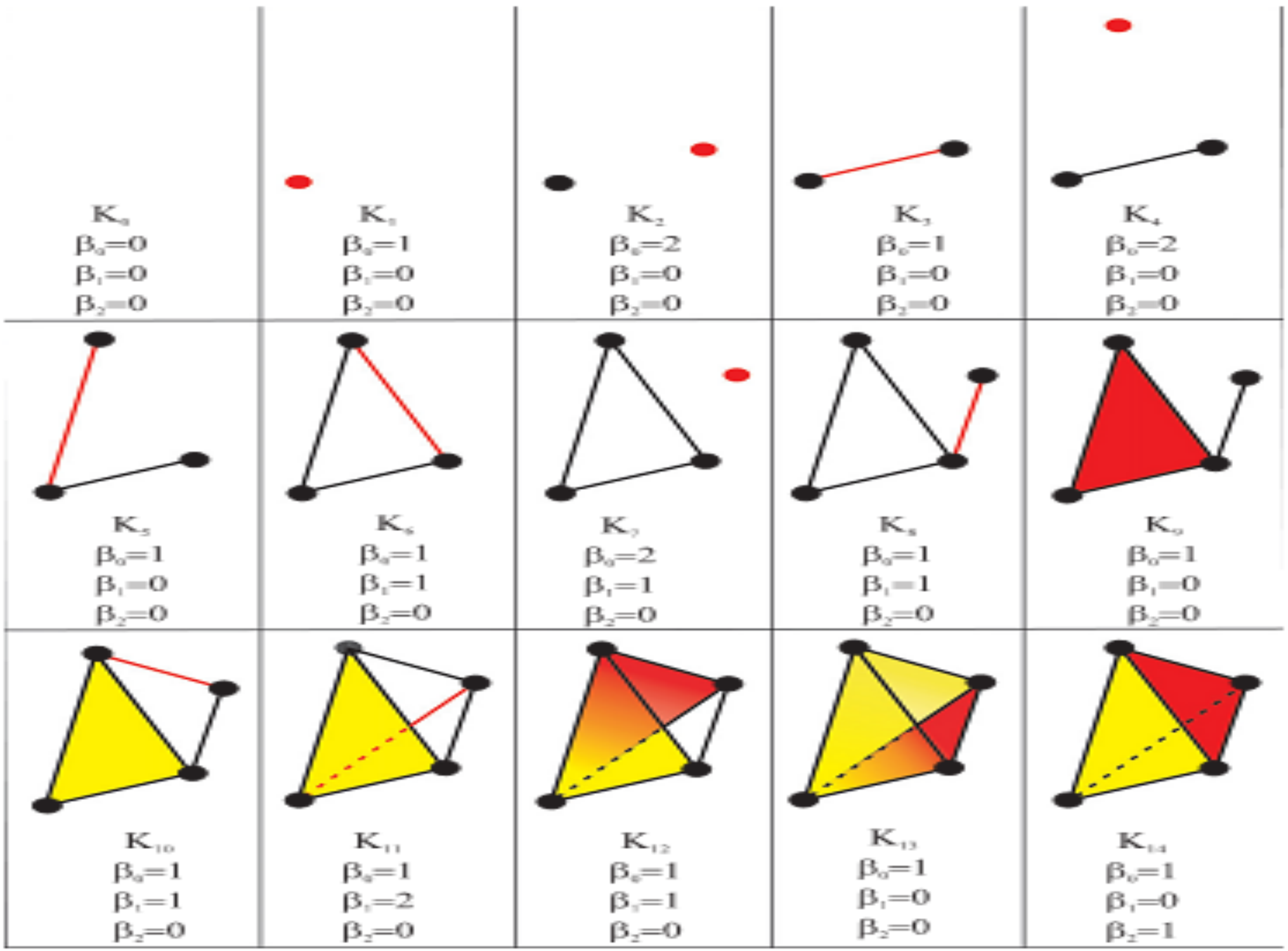
But what when we zoom in?

(Pictures: Peter Saveliev)





A sequence of Rips complexes for a point set representing an annulus. Upon increasing η holes appear and disappear. Which holes are real and which are noise?


K_0 $\beta_0=0$ $\beta_1=0$ $\beta_2=0$
K_1 $\beta_0=1$ $\beta_1=0$ $\beta_2=0$
K_2 $\beta_0=2$ $\beta_1=0$ $\beta_2=0$
K_3 $\beta_0=1$ $\beta_1=0$ $\beta_2=0$
K_4 $\beta_0=2$ $\beta_1=0$ $\beta_2=0$
K_5 $\beta_0=1$ $\beta_1=0$ $\beta_2=0$
K_6 $\beta_0=1$ $\beta_1=1$ $\beta_2=0$
K_7 $\beta_0=2$ $\beta_1=1$ $\beta_2=0$
K_8 $\beta_0=1$ $\beta_1=1$ $\beta_2=0$
K_9 $\beta_0=1$ $\beta_1=0$ $\beta_2=0$
K_{10} $\beta_0=1$ $\beta_1=1$ $\beta_2=0$
K_{11} $\beta_0=1$ $\beta_1=2$ $\beta_2=0$
K_{12} $\beta_0=1$ $\beta_1=1$ $\beta_2=0$
K_{13} $\beta_0=1$ $\beta_1=0$ $\beta_2=0$
K_{14} $\beta_0=1$ $\beta_1=0$ $\beta_2=1$

Filtration of a simplicial complex

Our goal is the construction of a Topological Data Field Theory (TDFT), giving us a new 'camera' for reading complex systems (in nature and society) :

- a camera whose 'photographs' do not consign reality to the past the moment they are shot, but enables us to *predict* a piece of the future;
- a TDFT whose gauge group looks into the transformation properties of the space of data revealing hidden complex patterns, somewhat like the vortices in turbulent water that only Leonardo's eye was ever able to catch.

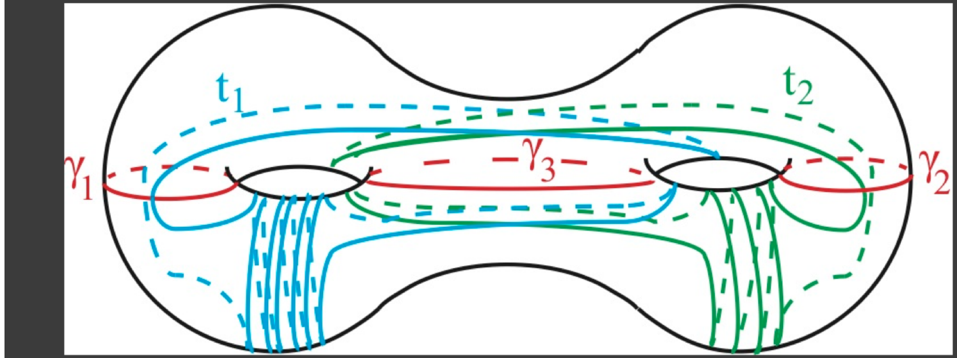
The three pillars the scheme rests on are:

1) Topological Data Analysis (persistent homology driven), based on the global topological (both algebraic and combinatorial) structural features of data space;

1) Statistical/Topological Field Theory of Data Space, as generated by the simplicial structure underlying data space and its self-transformation properties;

1) Semantics, emerging (naturally and autonomously) from the structure of the Formal Language generated by the transformations of data space.

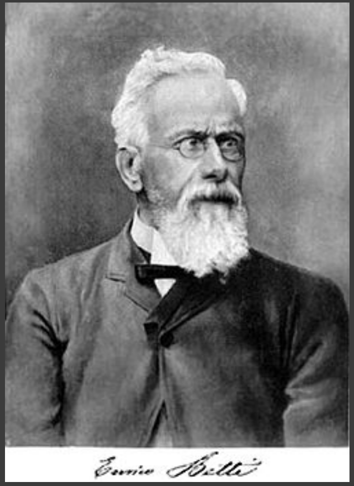
Key ingredients of this construction are the homology groups, H_i (X), i = 0, 1,… , of data space X and the associated Betti numbers, $b_i = b_i(X)$, b_i being the rank (dimension) of $H_i(X)$; basic set of topological invariants of X (when there is no torsion, the b_i's are sufficient).



generators of the homology groups

Intuitively, H_i's are functional algebraic tools to pick up the qualitative features of topological space $S \approx X$ (S = the simplicial complex) connected with the existence in X of holes in various dimensions.

Holes means cycles which don't arise as boundaries of higher-dimensional objects.

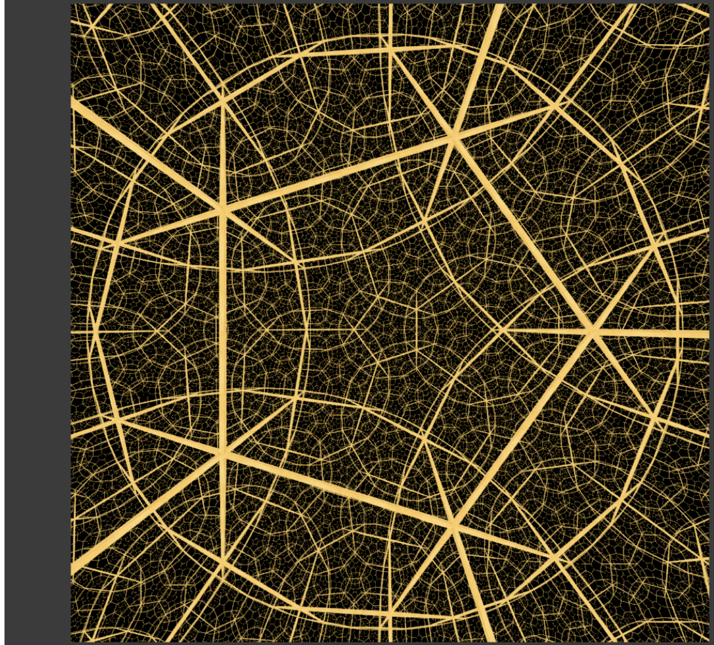


Enrico Betti

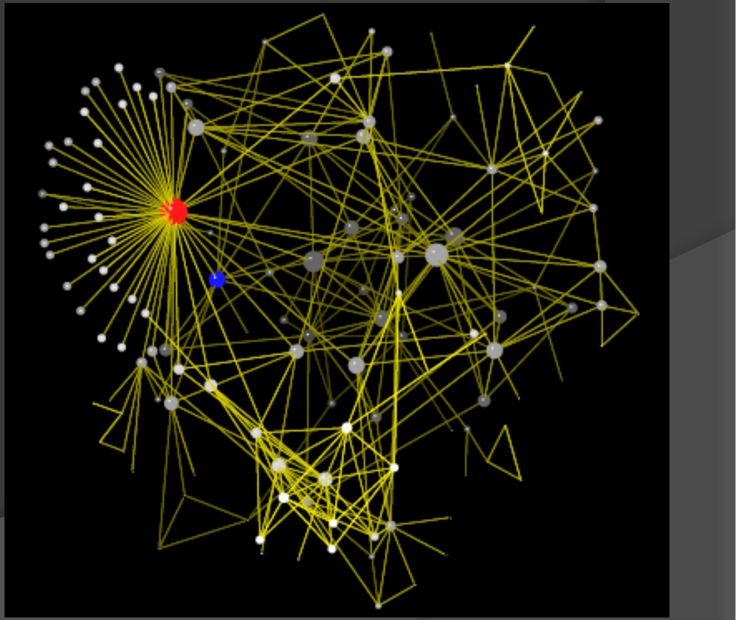
Known, efficient algorithms allow us to compute homology groups.

Most invariants in algebraic topology are difficult to compute efficiently; homology is not, because its invariants arise from quotients of finite-dimensional spaces and some derive from ‘physical’ models.

Traditional topology invariants were constructed out of manifest geometric global properties to distinguish homeomorphically different objects; recently others were discovered based on topological quantum field theory technology. These provide information about topological features one cannot detect, nor hint, based on geometric representation.

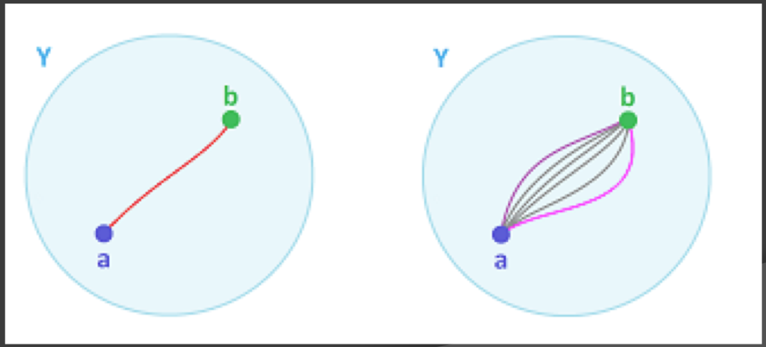


A delicate issue is here that in data space a simplicial complex is not necessarily a manifold. It is only if the links of all k-vertices are simplicial (k+1)-spheres (i.e., are homeomorphic to spheres in $\mathbf{R}^{k+1}$). The difficulty resides in the feature that n-spheres are straightforwardly identifiable only for n = 1, 2. The problem is tractable for n = 3, 4 only in exponential time and it is undecidable for $n \geq 5$.



S. Novikov, however, proved that for $n \geq 5$ the only further obstruction to the simplicial complex being a manifold of given homotopy type is the surgery obstruction : all finite simplicial complexes have the homotopy type of manifolds with boundary.

This leads to expect that the combinatorially different ways of sampling inequivalent structures in the persistence process generate a natural probability measure, consistent with data space invariants and transformation properties.



Finally, besides Vietoris-Rips, Čech, witness or other filtrations, used to implement persistence, another filtration enters into play, Morse filtration.

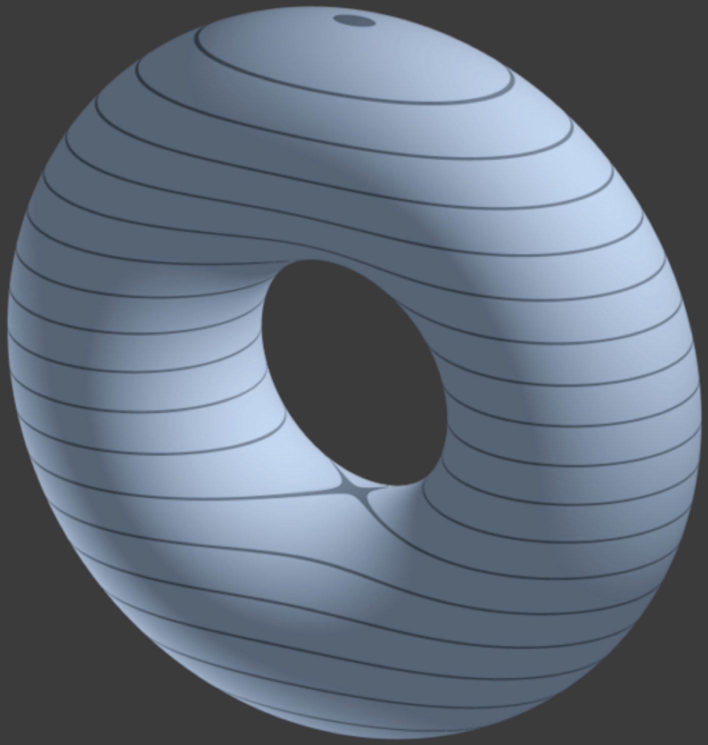
For simplicial complexes that are manifolds, this is a filtration by excursion sets; for the non-smooth, discrete, intrinsic, metric-free version thereof, proper to the wild simplicial complex that is data space, one needs to recall that Morse theory generates inequalities between (alternating sums of) Betti numbers and the numbers of critical points of the Morse function for each index (# of *Hessian* negative eigenvalues):

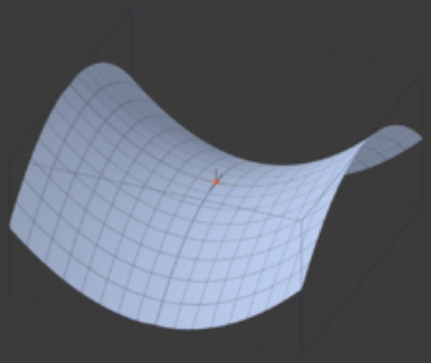
$$b_k \le m_k$$
$$b_1 - b_0 \le m_1 - m_0$$
$$b_2 - b_1 + b_0 \le m_2 - m_1 + m_0 \quad \dots$$

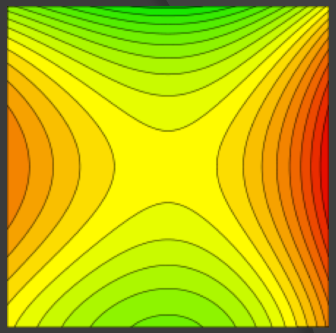
The Morse complex, built out of the critical points of a Morse function with support on the vertices of the simplicial complex, has the same homotopy as this underlying structure.

Morse stratification, exactly like the Harder-Narasimhan stratification, provides around the Morse strata a negative normal bundle to the critical sets, built, e.g., by the simplicial / combinatorial analog of Hodge's construction.

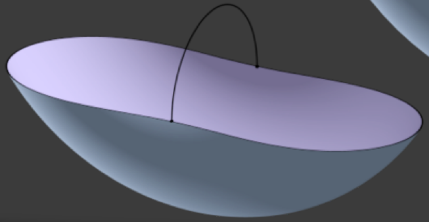
What relates Morse with homology is the property that the number of critical points of index k of a function f_M is equal to the number of k cells in the simplicial complex obtained climbing f_M, that bears on b_k.

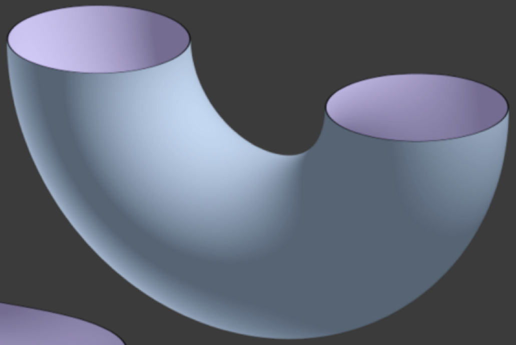


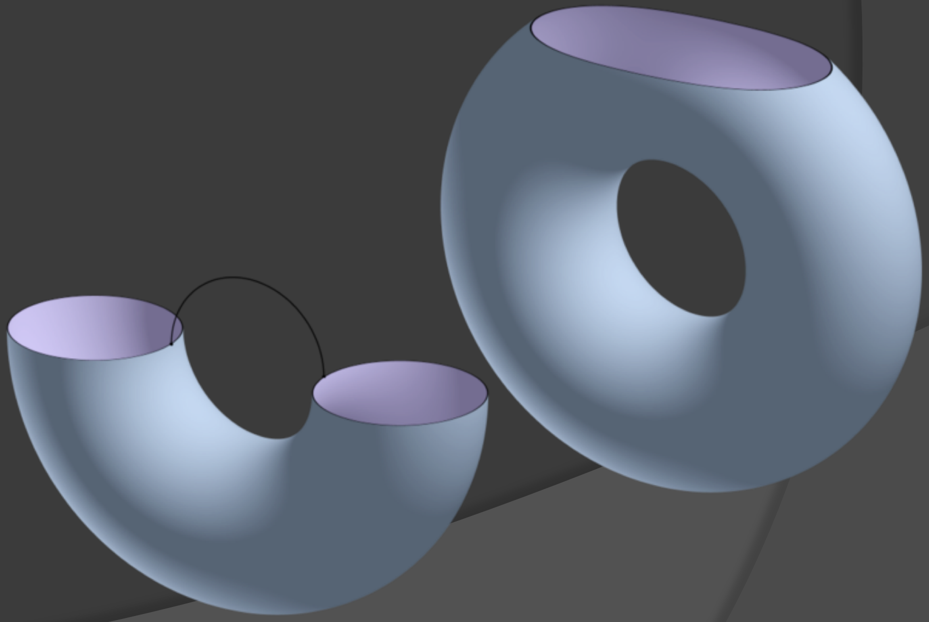




Morse







Morse homology, defined using a generic f_M and the local induced metric is a true topological invariant (i.e., independent of the function and the metric), and is isomorphic to singular homology : Morse and Betti numbers encode the same information, yet Morse numbers allow us to think of an underlying ‘manifold’.

Gromov’s spaces of bounded geometries provide a natural framework for addressing the measure questions posed by high-dimensional simplicial geometry and establishing entropy estimates to characterize the distribution of inequivalent simplicial configurations.

This leads to the construction of a statistical field theory of data, as the statistical features of GH topology are fully determined by the homotopy types of data space. Complexity and randomness of the emerging structure here can be large, as the number of coverings of a simplicial complex of bounded geometry grows exponentially with the volume (≈ thermodynamic limit – swiss cheese ↔ foam).

Yet, as growing filtrations of simplicial complexes are more and more random, it is possible to extend the notion of Gibbs field to the case where the substrate is not a graph but a simplicial complex, leading to a well defined statistical field theory.

Emerging scenario : the deep connection between the simplicial complex structure of data space and the information it encodedes resides in the property that data can be partitioned in a variety of equivalence classes, classified by their homotopy type, all elements of each of which can be assumed to encode ‘similar’ information. Thus in X information behaves as a sort of ‘order parameter’.

A single object encompasses most of the information about the global topological structure of X : P(z), the Hilbert-Poincaré series, generating function of the Betti numbers of S.

P(z) can be constructed through a field theory, as one of the functors of the theory for an appropriate choice of the field action.

A ‘Topological Field Theory’ of data space can be constructed mimicking conventional TFT, though with deep structural differences : discrete vs. continuous, wild vs. tame, infinite vs. finite gauge group.

The ingredients for a TFT are:

i) a base space, M. The structure of M allows us to do calculus with the appropriate type of field by the vector bundle obtained attaching to each point of M a fiber F;

ii) an action acting consistently over F;

iii) a ‘gauge’ group G.

Field equations are a ‘variational machine’ that takes as input the symmetry constraint imposed by G-invariance, and generates as output a field satisfying that constraint. The field at a point of M is an element of the fiber F that is now a G-bundle.

In view of the data space structure the construction requires specific tools:

i. vector bundles. They are proper to the differential category but have a PL category analogue, block bundles, that allow us to reduce geometric and transformation problems over simplicial complexes to homotopy theory for the groups and complexes involved. They provide a natural tool to construct the moduli space of G-bundles in a discretized setting.

Since the homotopy class of a map fully determines its homology class, the simplicial block-bundle construction furnishes all necessary tools to compute, e.g., the Poincaré series.

ii. The (exponentiated) action. Needed to construct the field theory propagators. A natural candidate is the Heat Kernel K, but constructed with the intrinsic combinatorial Laplacian over the simplicial complex. The Heat Kernel's trace gives the Poincaré series.

ii. gauge group G. Data space X is fully characterized by its topological properties, then there is one symmetry it has to satisfy: invariance under all homeomorphisms of X that don't change its topology and are consistent with the constraints. G must then be the semidirect product $P_{\wedge}G_{MC}$ of the group P associated with the characteristic process algebra of the data set and G_{MC} , the simplicial analog of the mapping class group for X ($\mathrm{Diff} / \mathrm{Diff}_0$).

Using the block bundle approach for X and given G, all topological invariants can be computed in the context of the TDFT through the subsets of symmetries of G :

i) The cosets of G order data in equivalence classes with respect to isotopy and canonical equivalence under the process algebra P

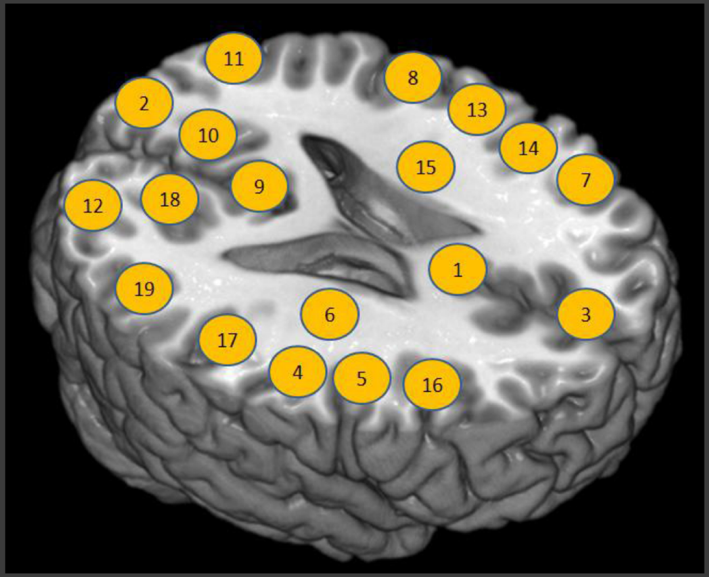
i) The choice among the several possible theories (actions) can be made unique by self-consistency, comparing the coefficients of P(z) with the Betti numbers outcome of the 'phenomenological' persistent homology analysis of data

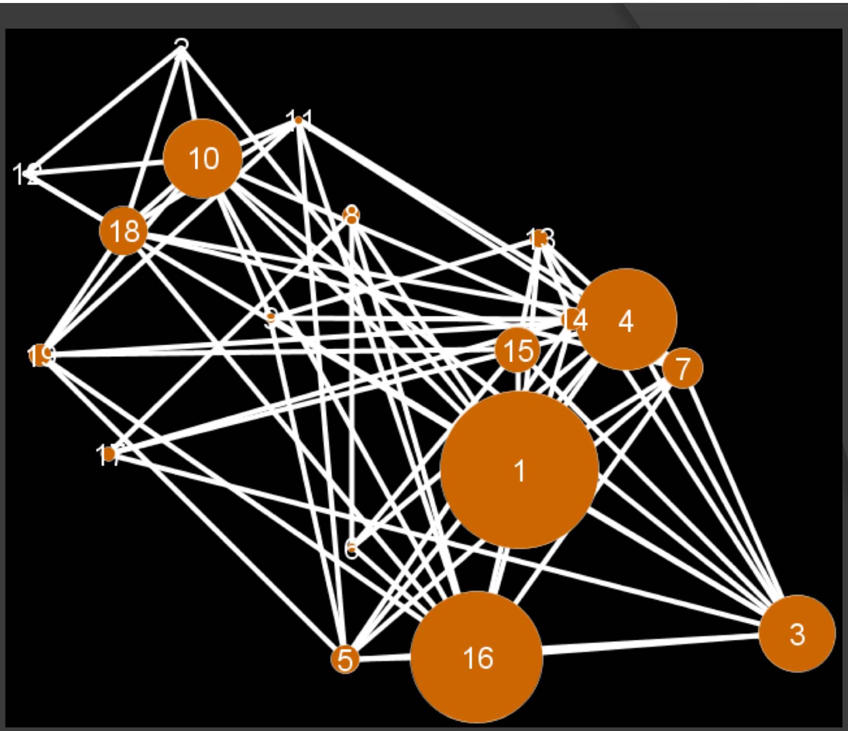
i) Correlation functions of the resulting field theory fully describe the pattern system in data space

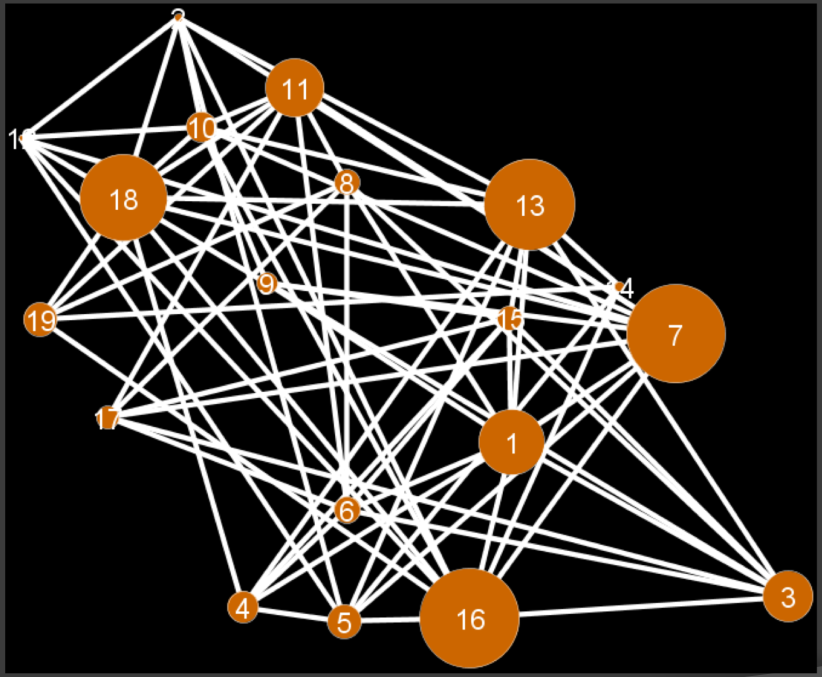
An example : brain functions from MRI data

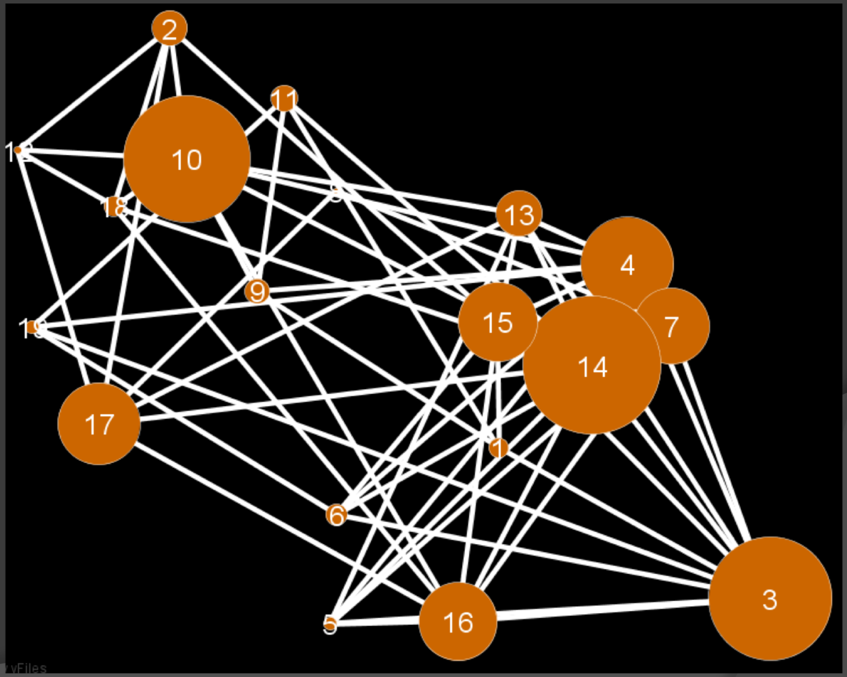
Some of the tools of TDFT were applied to compare resting-state fMRI data of brain activity in a sample of healthy volunteers half of whom infused a placebo, the other half a psychoactive drug (psilocybin).

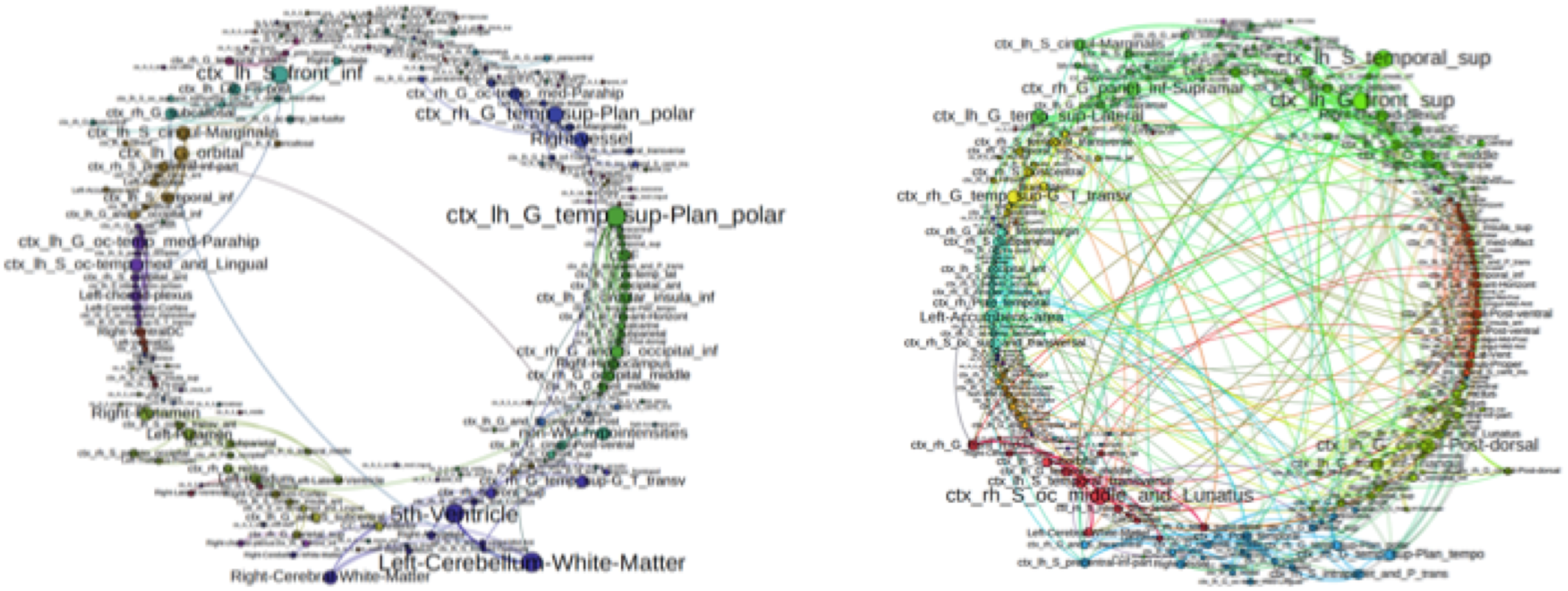
The homological structure of the brain's functional patterns undergoes a dramatic change due to psilocybin, with a large number of transient structures of low stability and of a small number of highly stable ones, not observed in the placebo case. The corresponding cycles (H_1) indicate enhanced brain functional integration under psilocybin as compared to the normal resting-state.


11
8
2
13
10
14
15
7
9
18
12
1
19
3
6
17
4
5
16


2
11
12
10
18
8
13
9
14
4
15
7
19
1
17
6
5
16
3


2
11
12
10
18
8
13
9
14
19
15
7
17
1
6
4
5
16
3


2
11
12
10
18
8
13
4
9
15
7
19
14
17
1
6
5
16
3
yFiles

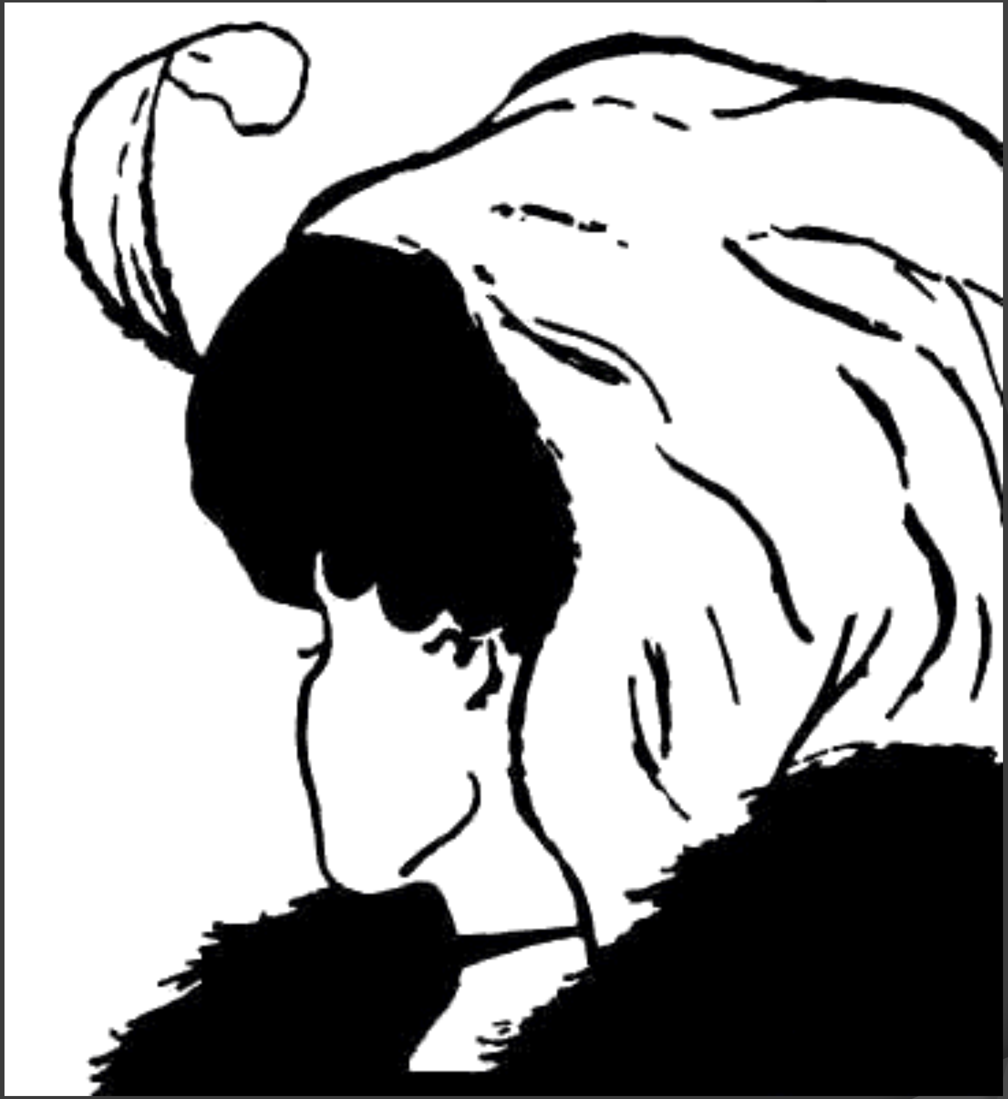

ctx_lh_S_front_inf
ctx_rh_G_oc-temp_med-Parahip
ctx_rh_G_temp_sup-Plan_polar
Right-vessel
ctx_lh_S_cingul-Marginalis
ctx_lh_G_orbital
ctx_lh_S_temporal_inf
ctx_lh_G_temp_sup-Plan_polar
ctx_lh_G_oc-temp_med-Parahip
ctx_lh_S_oc-temp_med_and_Lingual
Left-choroid-plexus
ctx_lh_S_circular_insula_inf
ctx_rh_G_and_S_occipital_inf
ctx_rh_G_occipital_middle
Right-Putamen
non-WM-hypointensities
ctx_rh_G_temp_sup-G_T_transv
5th-Ventricle
Left-Cerebellum-White-Matter
Right-Cerebral-White-Matter
ctx_lh_S_cingul-Marginalis
ctx_lh_S_temporal_sup
ctx_rh_G_pariet_inf-Supramar
ctx_lh_G_front_sup
ctx_lh_G_temp_sup-Lateral
ctx_rh_G_temp_sup-G_T_transv
Left-Accumbens-area
ctx_lh_G_cingul-Post-dorsal
ctx_rh_S_oc_middle_and_Lunatus

Control group

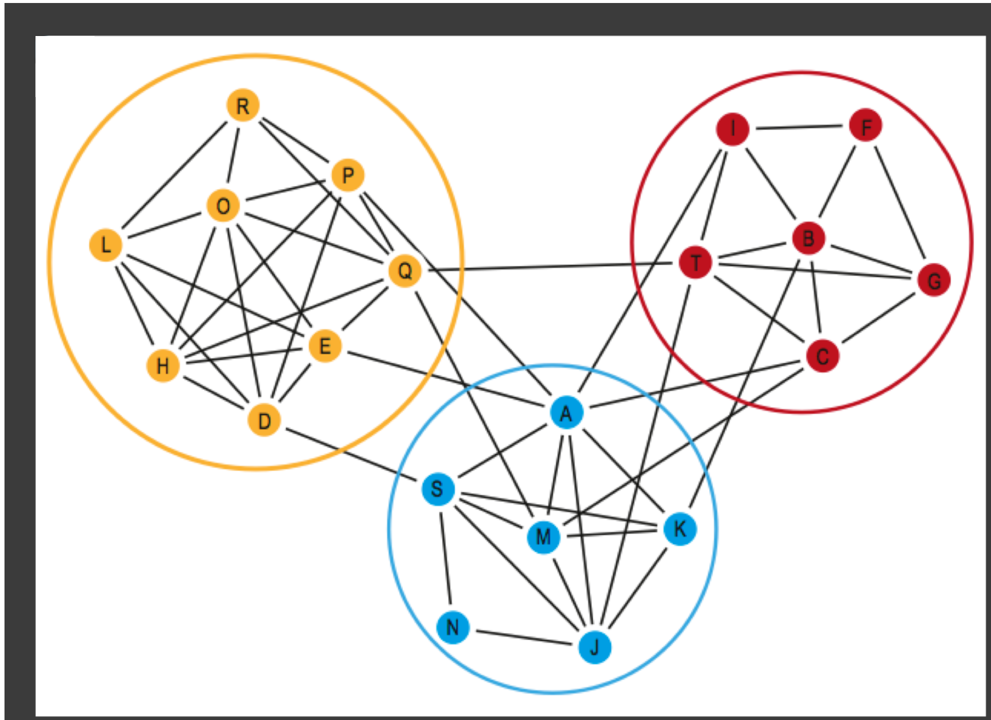
'Psilocybin' group

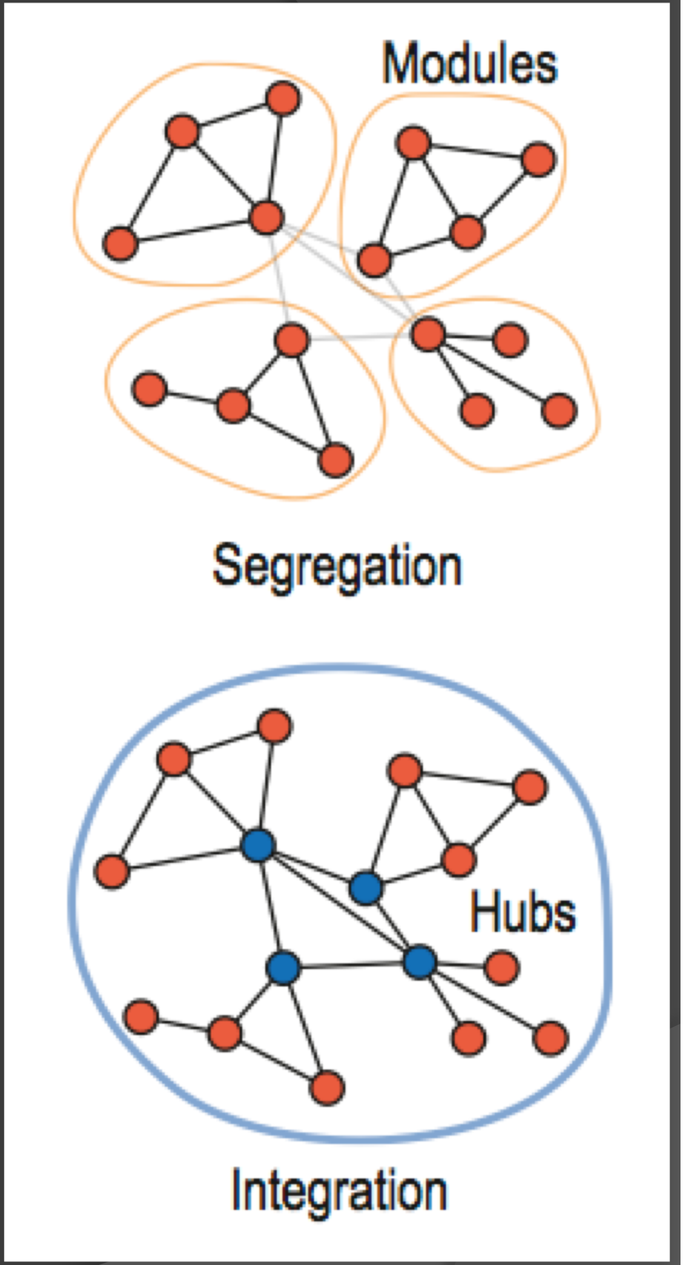
The figures evidence the method's capacity to describe coexisting mesoscopic patterns at various intensity scales, complementing the information relative to the cluster structure of the brain's functional circuits.

Francisco Varela



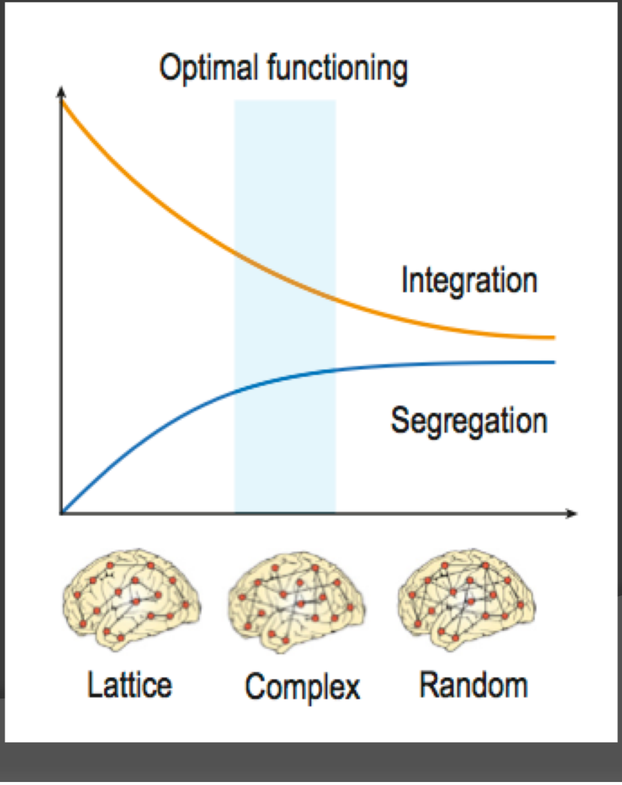
Ambiguous pictures



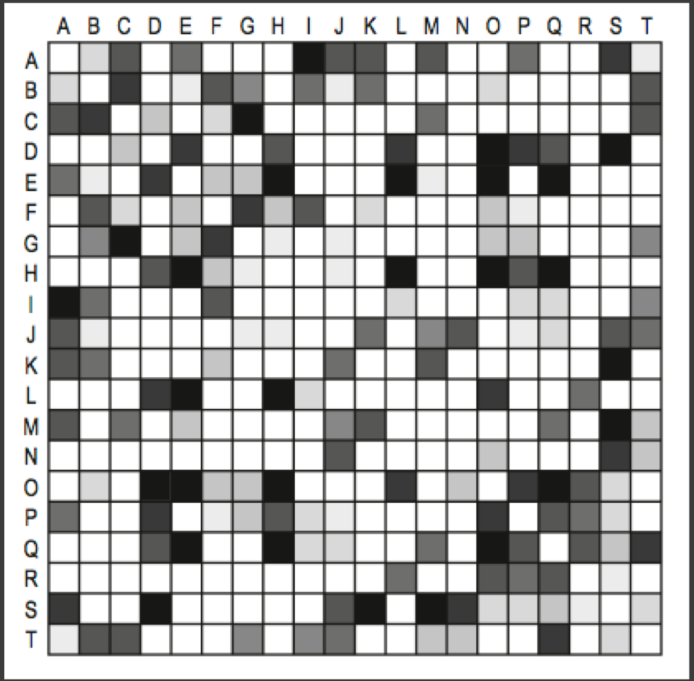


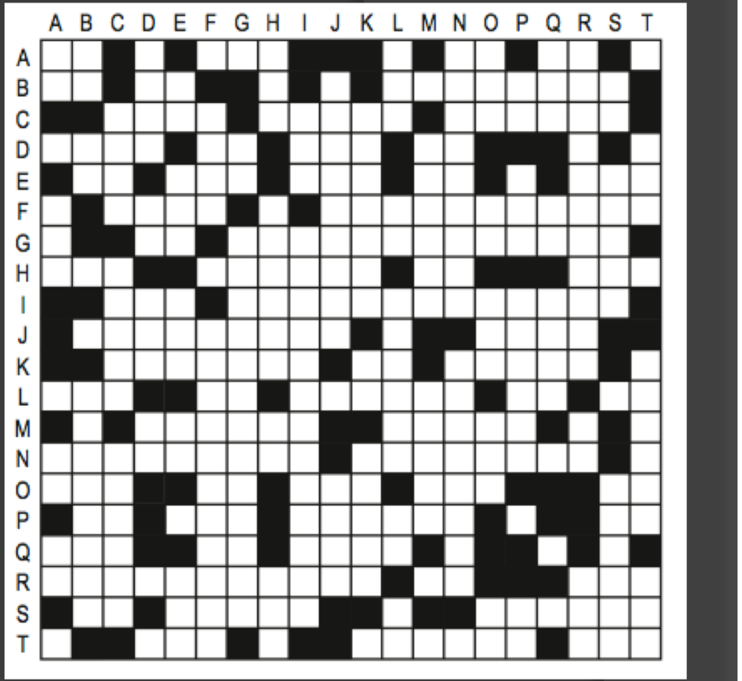
Connectomics

optimal functioning of any brain is to balance spatial integration and segregation



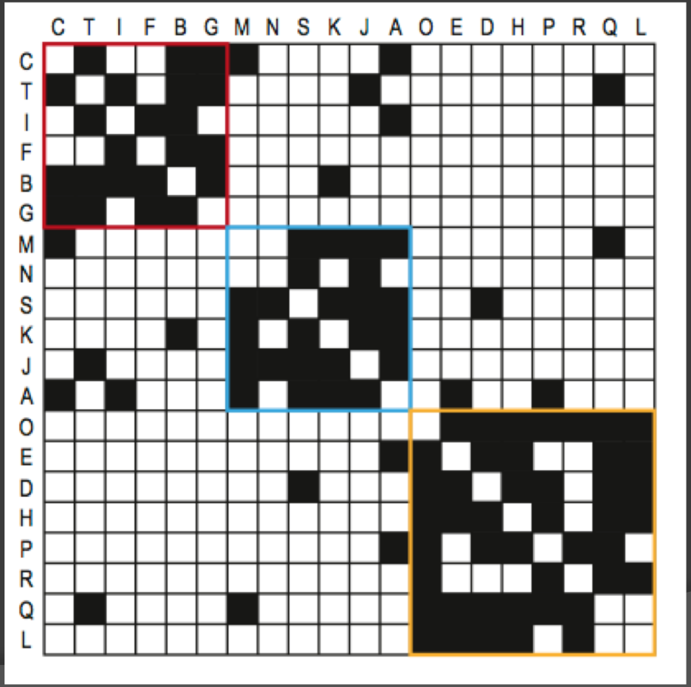
Connection strengths (shades of gray; 20 brain regions)


A B C D E F G H I J K L M N O P Q R S T
A B C D E F G H I J K L M N O P Q R S T


A B C D E F G H I J K L M N O P Q R S T
A B C D E F G H I J K L M N O P Q R S T

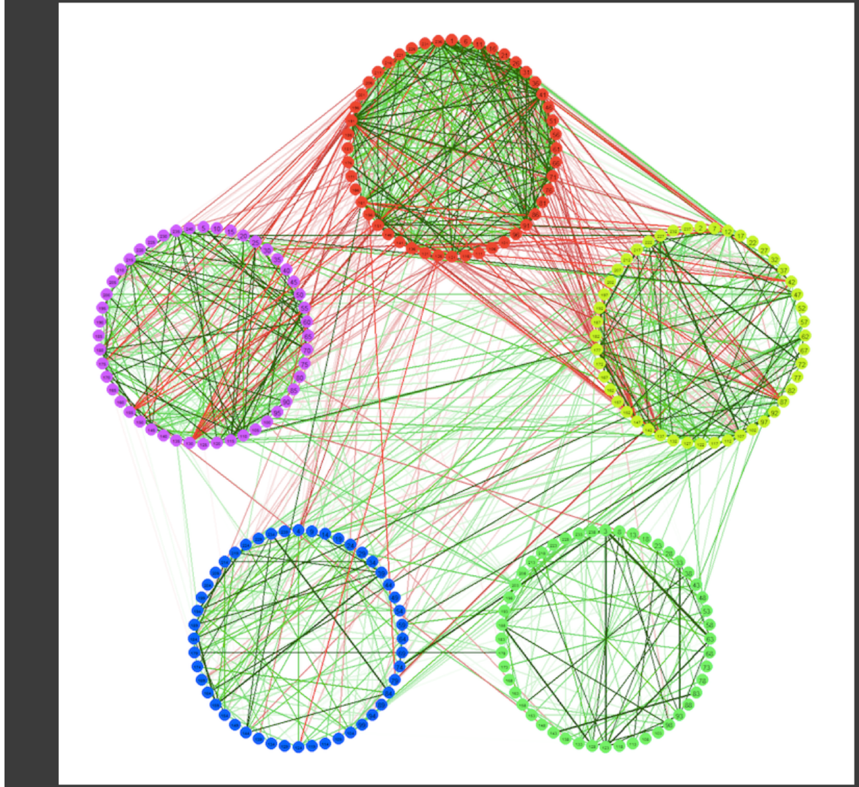
Binarization (threshold)

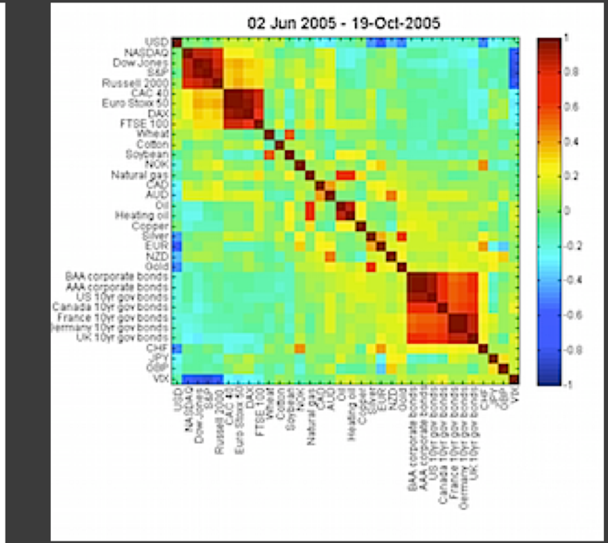
Reordering and modularization


C T I F B G M N S K J A O E D H P R Q L
C T I F B G M N S K J A O E D H P R Q L

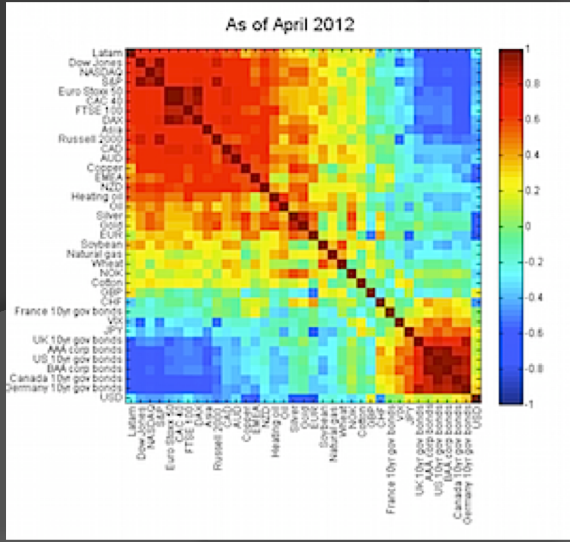
Brain Network

The data space splits into the direct sum of irreps of G. The general ‘covariance matrix’ of a generic machine learning algorithm becomes ‘block-diagonal’: all zeroes are pushed to the upper-right / lower-left corners.



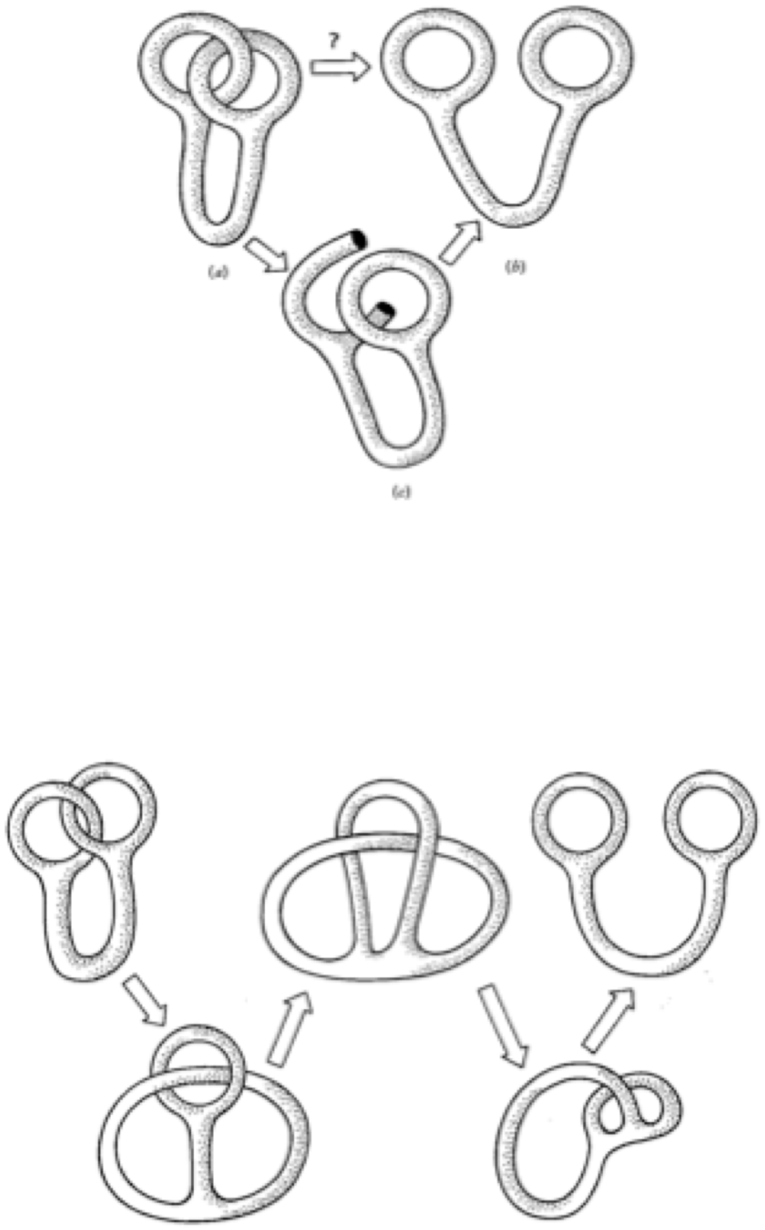


02 Jun 2005 - 19-Oct-2005

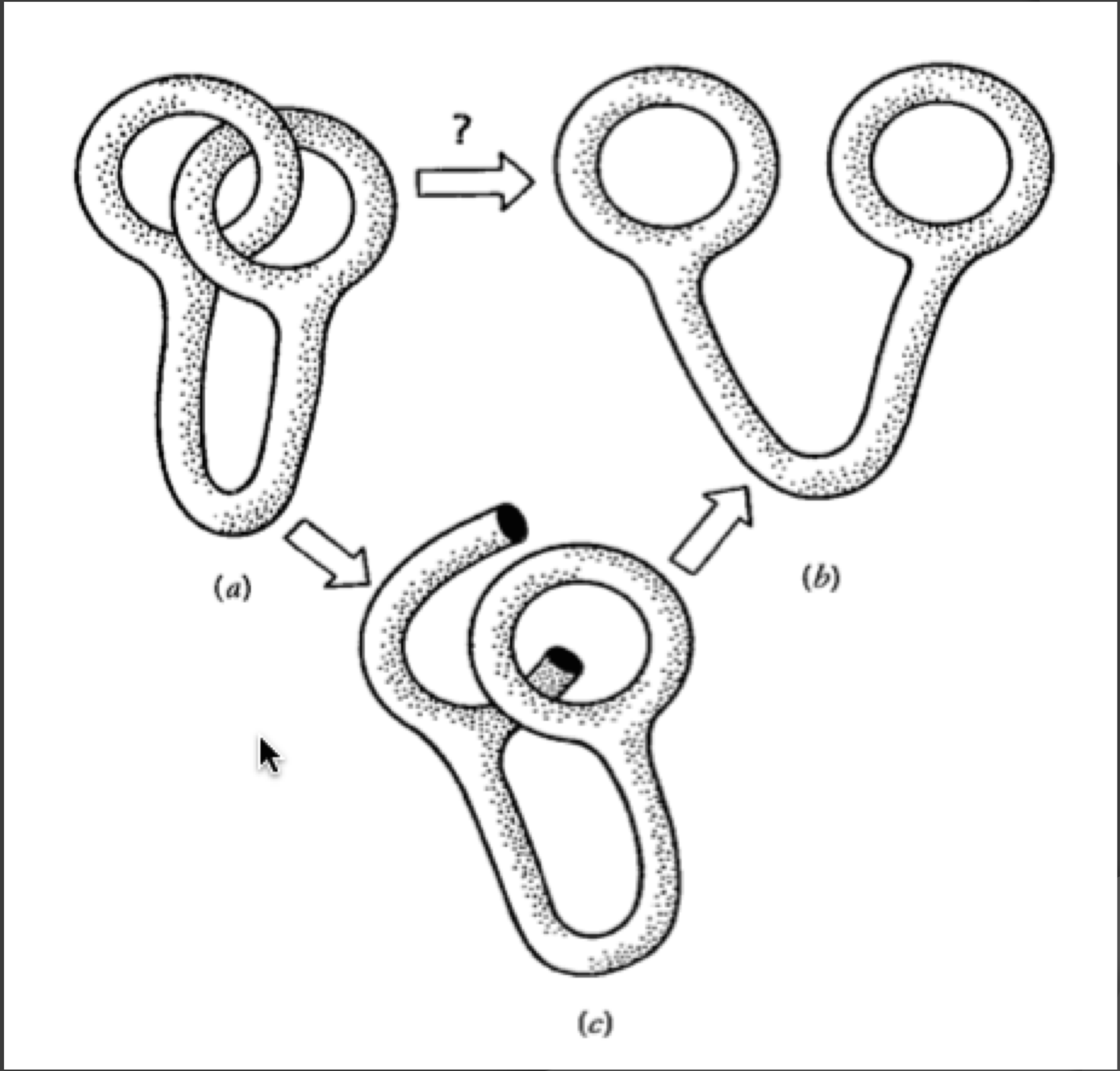


As of April 2012

Topology and dimensions



D = 1



D = 2