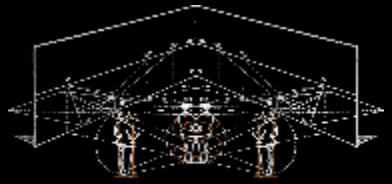


“ Big Data: threath or opportunity? ”

Mario Rasetti

ISI Foundation – Torino
ISI Global Science Foundation – New York

Big data impose a strenuous data manipulation challenge, both in computer science (new computing paradigms; interaction-based computing; beyond Turing machine) and data analytics (new approach to data mining; non-linear causal inference). But the conventional approaches to complex problems in both basic science (mostly the life sciences, climate and earth science) and societal questions as implemented by **data science** (A.I., data mining, machine learning [DL, extreme ML]) and complexity science (network theory), pay perhaps too little attention to basic conceptual and universal features.

In '**topological data analysis**' the foresight is that appropriate mathematical tools, grounded on the profound notion of '***space of data***' – geometrical representation of large data sets – may enable us to incorporate data in a setting (mostly topological) that allows us to control and individuate the hidden information patterns in a more and more effective way.

This means to endow IT with the capacity of more efficiently playing its role in the process:

data → information → knowledge → wisdom.

Topological Data Analysis (TDA) is a theoretical framework allowing for the efficient exploration of large amounts of data. It provides a new data mining method based on a non-linear topological field theory of data space.

The approach is based on the inference of information from **global** rather than **local** data space properties.

It stems out of the integration of deep mathematical aspects of topological analysis of the data space, with formal language theory and theoretical computer science.

TDA goes beyond conventional complex network theory and the constraints of standard methods such as machine learning.

Why is topology the natural tool to handle large, high-dimensional, complex spaces of data?

Because :

Qualitative information is relevant : data users aim to obtain knowledge, i.e., to understand how data is organized on large scale. Global, though qualitative, information is what matters.

Metrics are not theoretically justified : in physical sciences, phenomena support clean theories which tell us exactly what metric to use; in the life or social sciences this is not the case.

Coordinates are not natural : data is typically handled in the form of ‘vectors’ (strings of symbols), yet the ‘components’ or linear combinations or norm of these ‘vectors’ are not natural in any sense: the space of data is not a vector space (Hopcroft): properties of data space depending on the choice of coordinates are not relevant.

Summaries are valuable : the conventional method of handling data is by a graph (**network**) whose vertex set the points of data space; two points being connected by an edge if their ‘proximity measure’ (Grothendieck) is $\leq \varepsilon$, with the optimal choice of ε. This is too local to provide a reliable view. It is much more informative to consider the entire **dendrogram**, getting at once a summary of its relevant features under all possible values of ε by knowing how the global features of data space vary changing ε.

And we simply have to recall that :

Topology is the branch of mathematics that deals with qualitative rather than quantitative 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 (ε)

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

In particular TDA is articulated in the development of:

i) (persistent) Homology methods, to extract, from a **simplicial complex** representation of the data-space manifold, those correlation patterns that encode deep level information;

ii) a Topological Field Theory of Data (TDFT), able to provide all characteristic information about such patterns, non-linearly tangled in the data set, in a way that – due to nonlinearity – also to feedbacks the reorganization of the data set itself;

iii) a Theory of Formal Languages, enabling us to express the semantics of the transformations implied by the field dynamics into automated learning processes.

Its reach is a deep evolution of conventional data mining methods (e.g., ML).

The latter allow for the discovery of unknown features that data conceal resorting to A.I. methods that train the machine to recognize fragments of information. ”**Unsupervised learning**” improves the learner success in the design of algorithms that allow to evolve behaviors based on empirical data.

ML implies some amount of a priori knowledge of what one is looking for. Its typical difficulty is that the set of all possible behaviors given all possible inputs is too large to be covered by the set of known samples (training data).

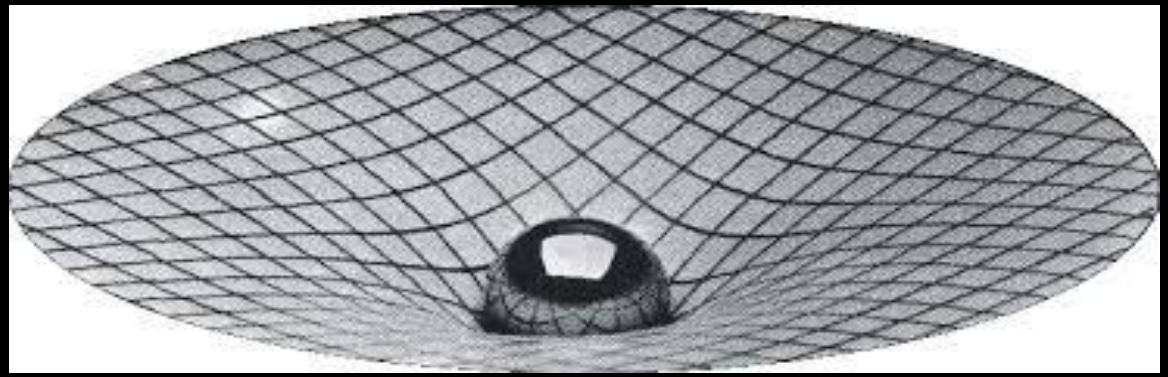
Resorting to topological methods, specifically persistent homology theory – efficient way to distinguish noise from signal – defines a theoretical framework that allows for effective analysis of very large amounts of data, while providing innovative data mining methods that permit mining data spaces with no need of any a priori knowledge of the system.

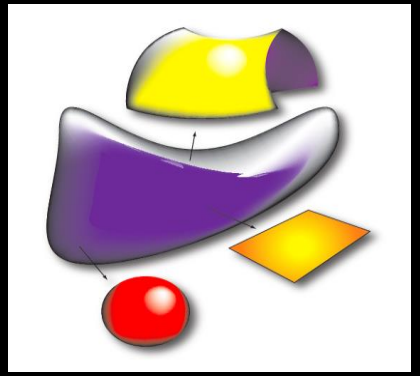
In TDA the natural notion of **Simplicial Complex** is a far-reaching extension of that of complex network in two important respects: 1.) in it relations that are not only 'two-body' but account for the interaction of any number of nodes in the graph; 2.) it involves objects of any number of dimensions, whence the dependence of the correlations from higher order global topological invariants.

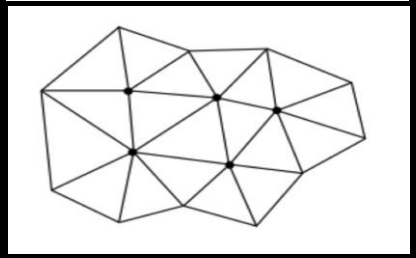
Like in General Relativity:

$$R_{\mu\nu} - \frac{1}{2} g_{\mu\nu} R = \frac{8\pi G}{c^4} T_{\mu\nu}$$

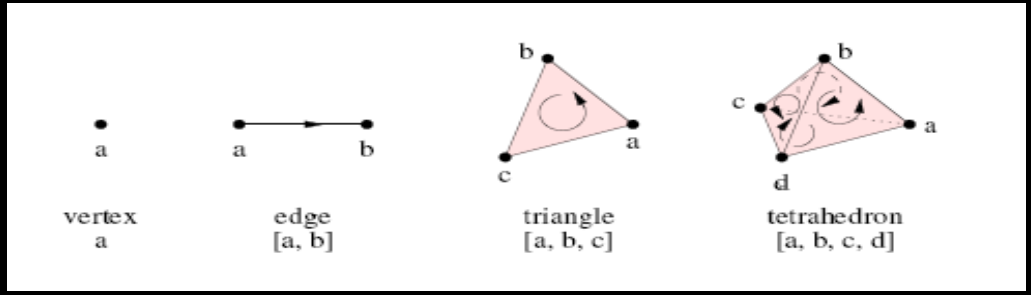
Einstein's equations

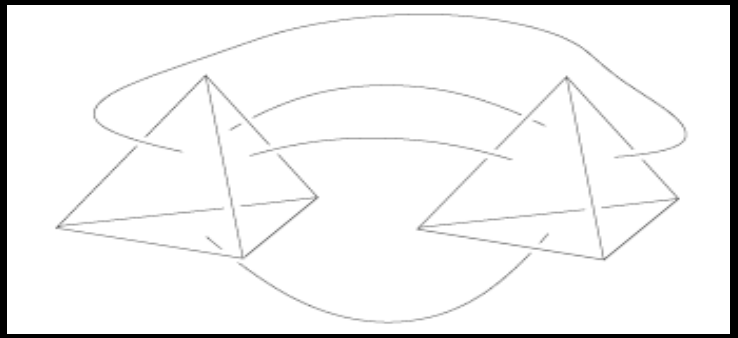


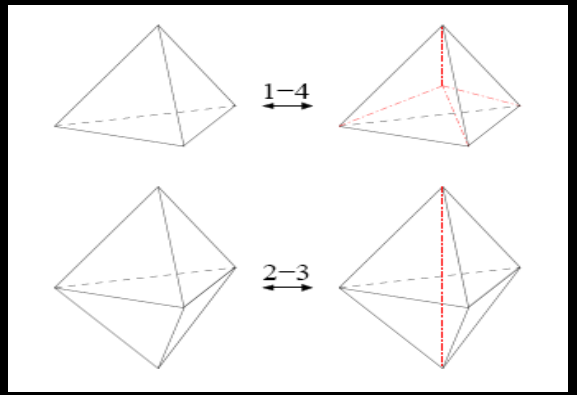


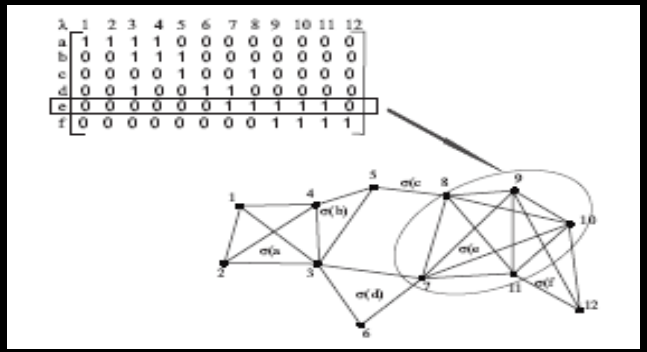


J.A. Wheeler: «Matter tells to space-time how to bend, curved space-time tells matter how to move».





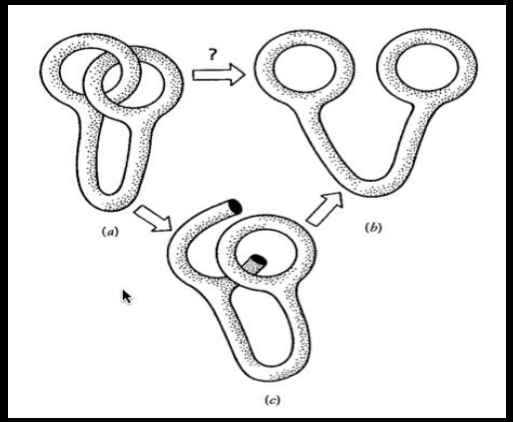


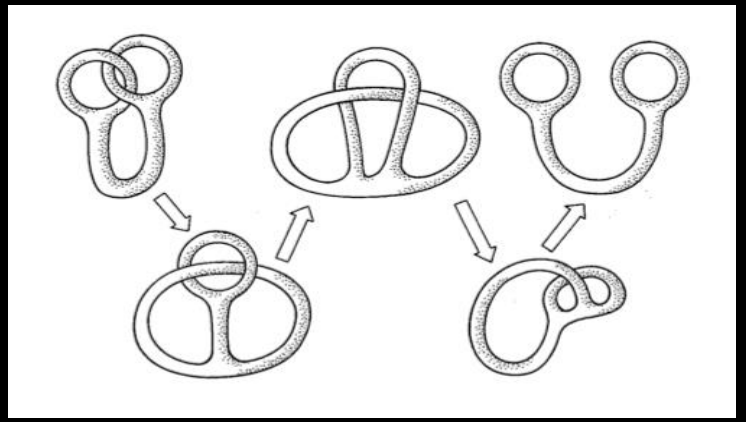


Simplices & Complexes
(in data space are for free!)

Topology distinguishes among dimensions

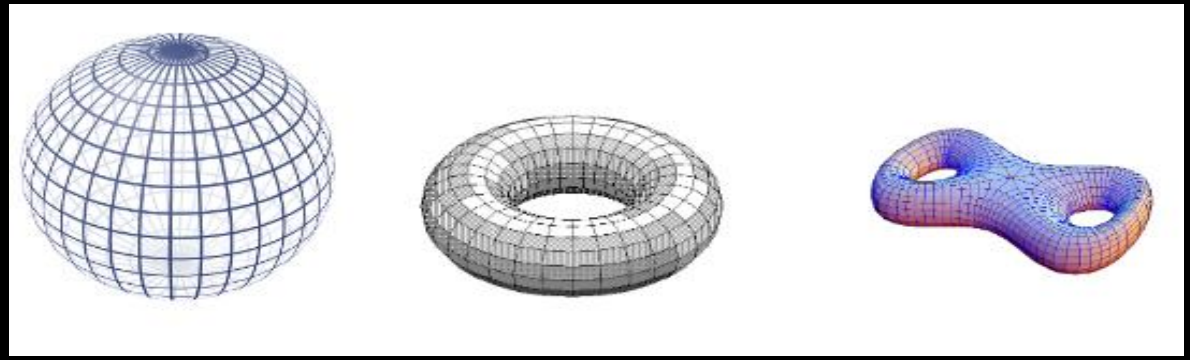
D = 1

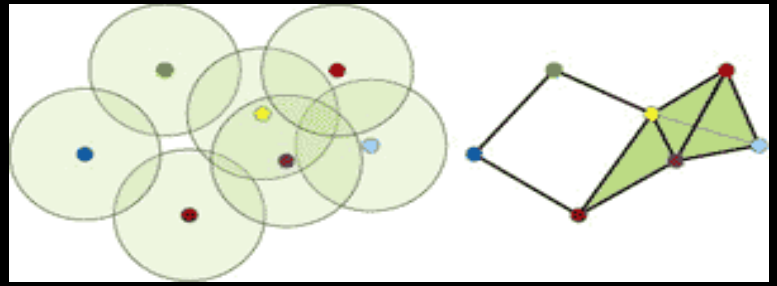




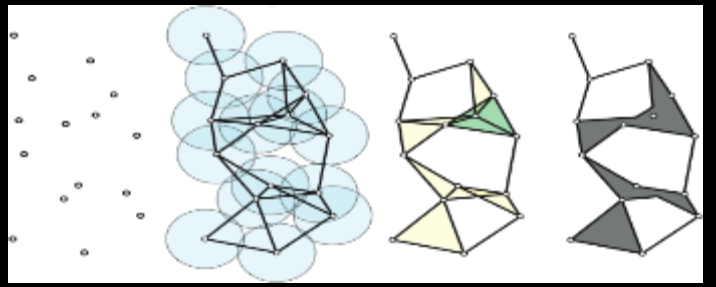
D = 2

Topology distinguishes Shapes

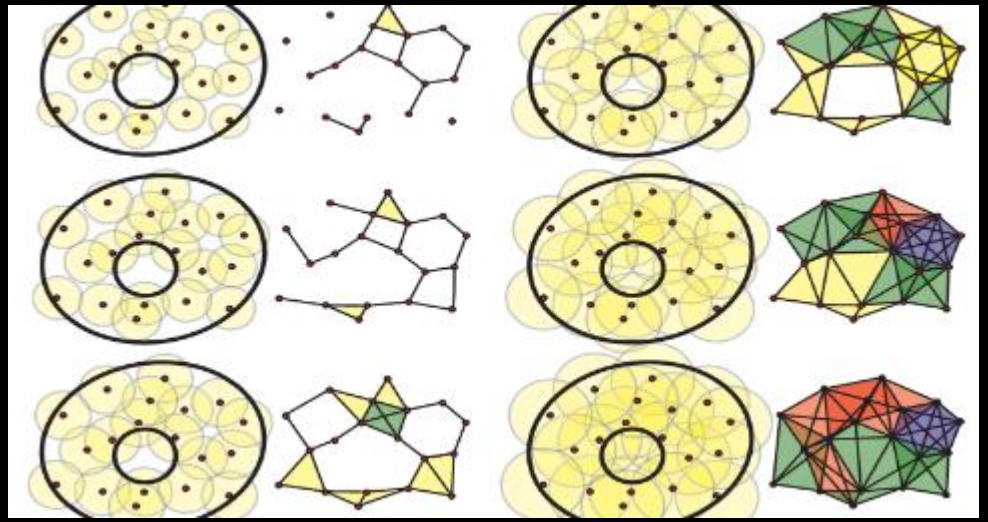




Čech complex

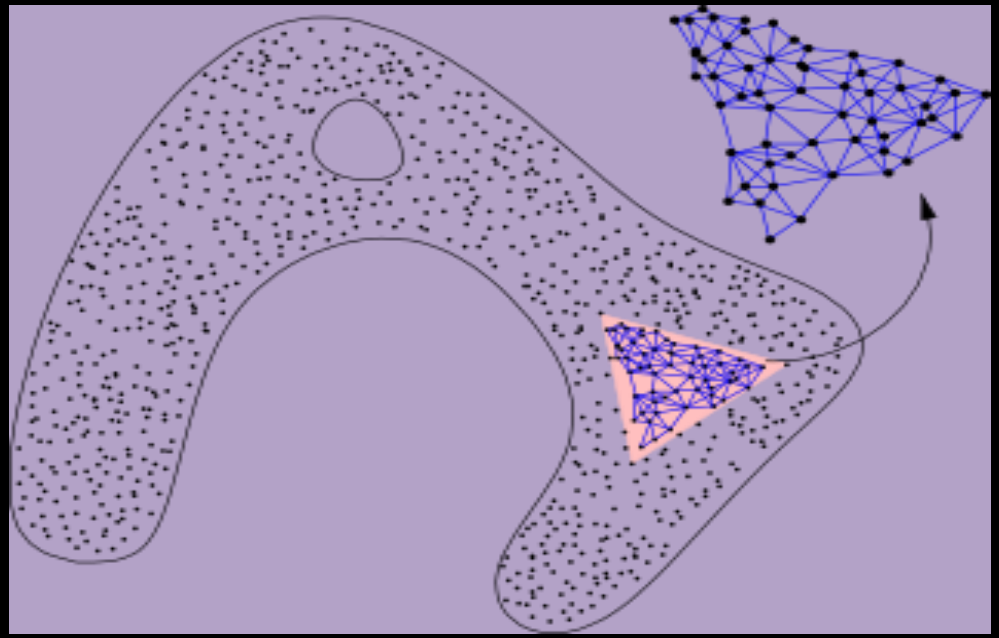


Rips-Vietoris complex

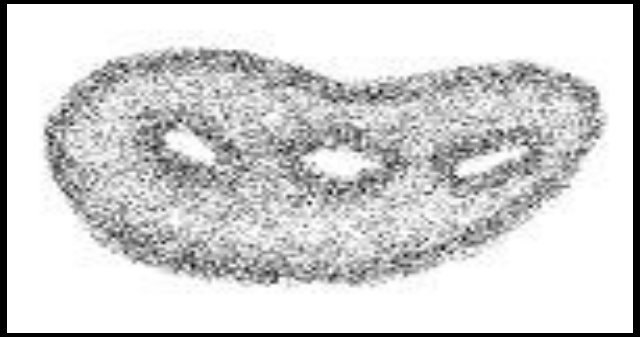


A sequence of Rips complexes for a point set representing an annulus. Increasing ε new holes appear and disappear. Question: which are real, which are noise?

Data as a point cloud



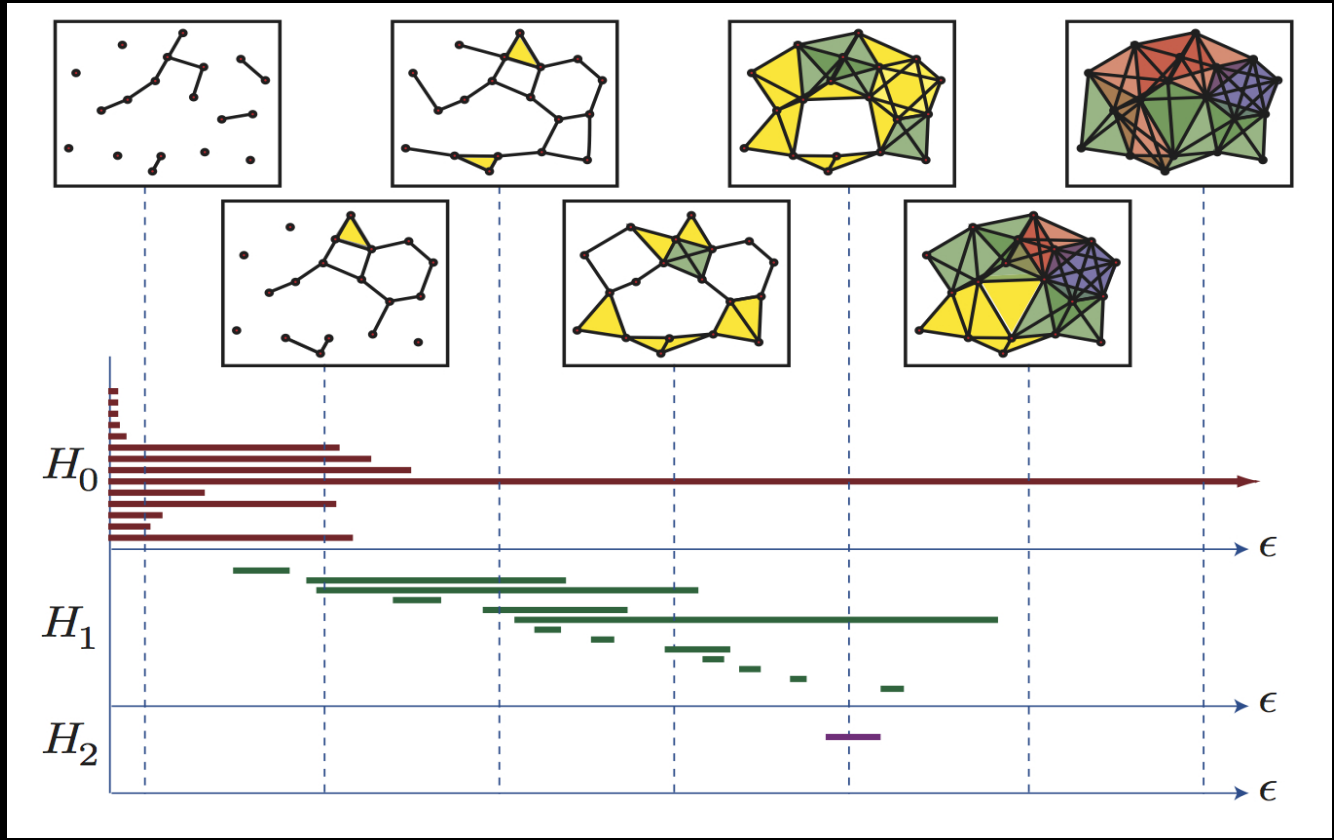
For example, the object below appears to be a genus 3 handlebody:



But what when we zoom in?

Signal

Noise



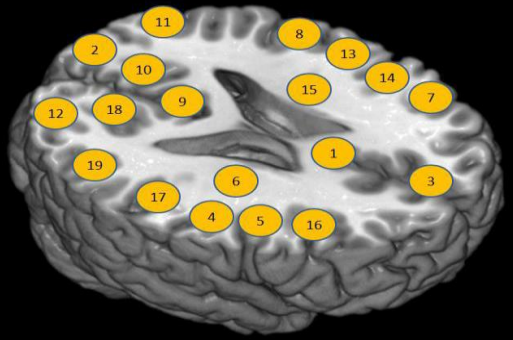
Barcodes (persistence of Betti numbers)

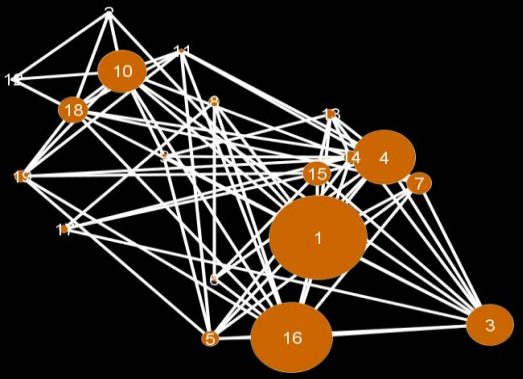
What was done

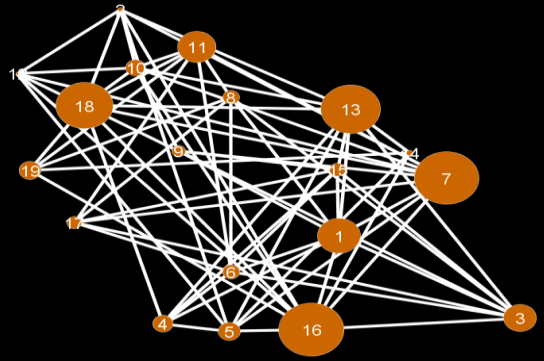
G. Petri, P. Expert, F. Turkheimer, R. Carhart-Harris, D. Nutt, P. J. Hellyer, F. Vaccarino,
Homological scaffolds of brain functional networks
Journal Royal Society Interface **11**: 20140873

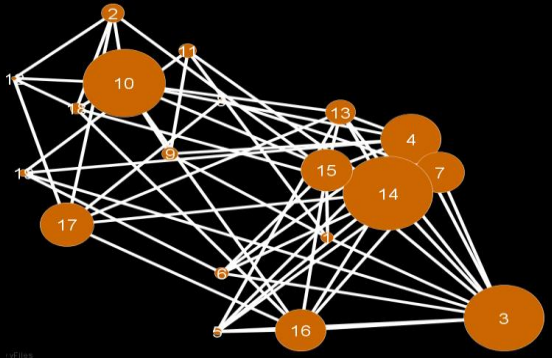
Networks have long permeated many areas of science, including neuroimaging. Traditionally, the structure of complex networks has been studied through their statistical properties and the metrics of the node and link system: degree-distribution, node centrality, modularity.

Here the characteristics of functional brain networks (**SC**) at the mesoscopic level was studied instead from the TDA perspective, which highlights the role of inhomogeneities in the fabric of functional connections.


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


10
18
4
15
7
1
16
3
5


11
18
13
7
1
16
3
4
5

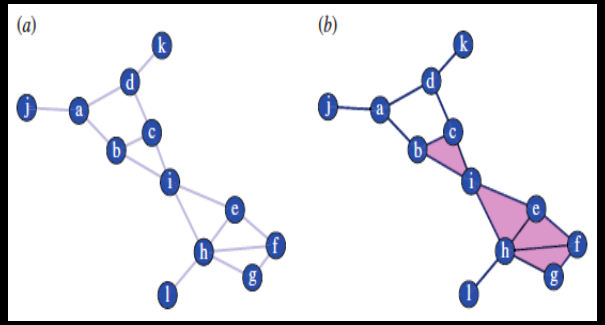

10
13
4
15
7
14
17
16
3
5

f-MRI Data

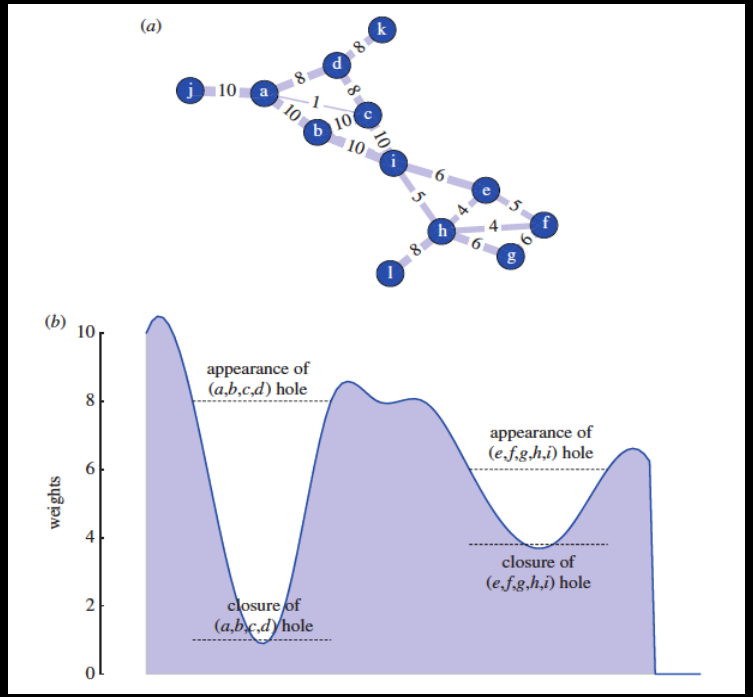
This was done by focusing on the features of a set of topological objects – **homological cycles** – associated with the weighted functional network. Topological information defines the homological **scaffolds**, a set of objects designed to represent the features of the correlation network and make homological properties amenable to graph statistical methods. These tools were applied here to compare resting-state functional brain activity in a sample of healthy volunteers after intravenous infusion of **placebo** or of **psilocybin**.

The results show that the homological structure of the brain's functional patterns undergoes a dramatic change post-psilocybin, characterized by the appearance of many transient structures of low stability and of a small number of persistent ones not observed in the case of placebo.

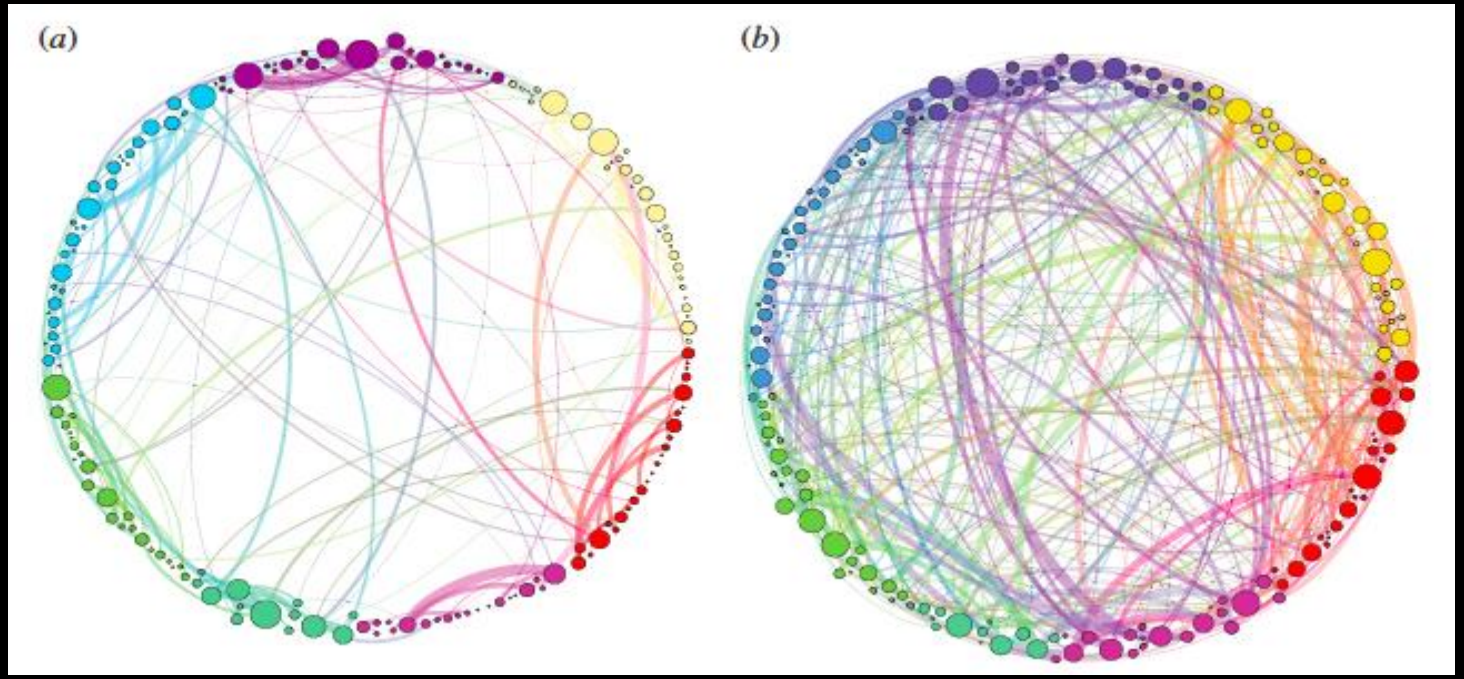
In this way one do away with (arbitrary) averaging of neuroimaging data in space/time or across participants (e.g., patients vs. control).



An unweighted network (a), and its clique complex, (b), obtained by promoting cliques to simplices.



(a) A weighted network;
(b) Its intuitive representation in terms of stratigraphy in the weight structure as for to weight filtration.

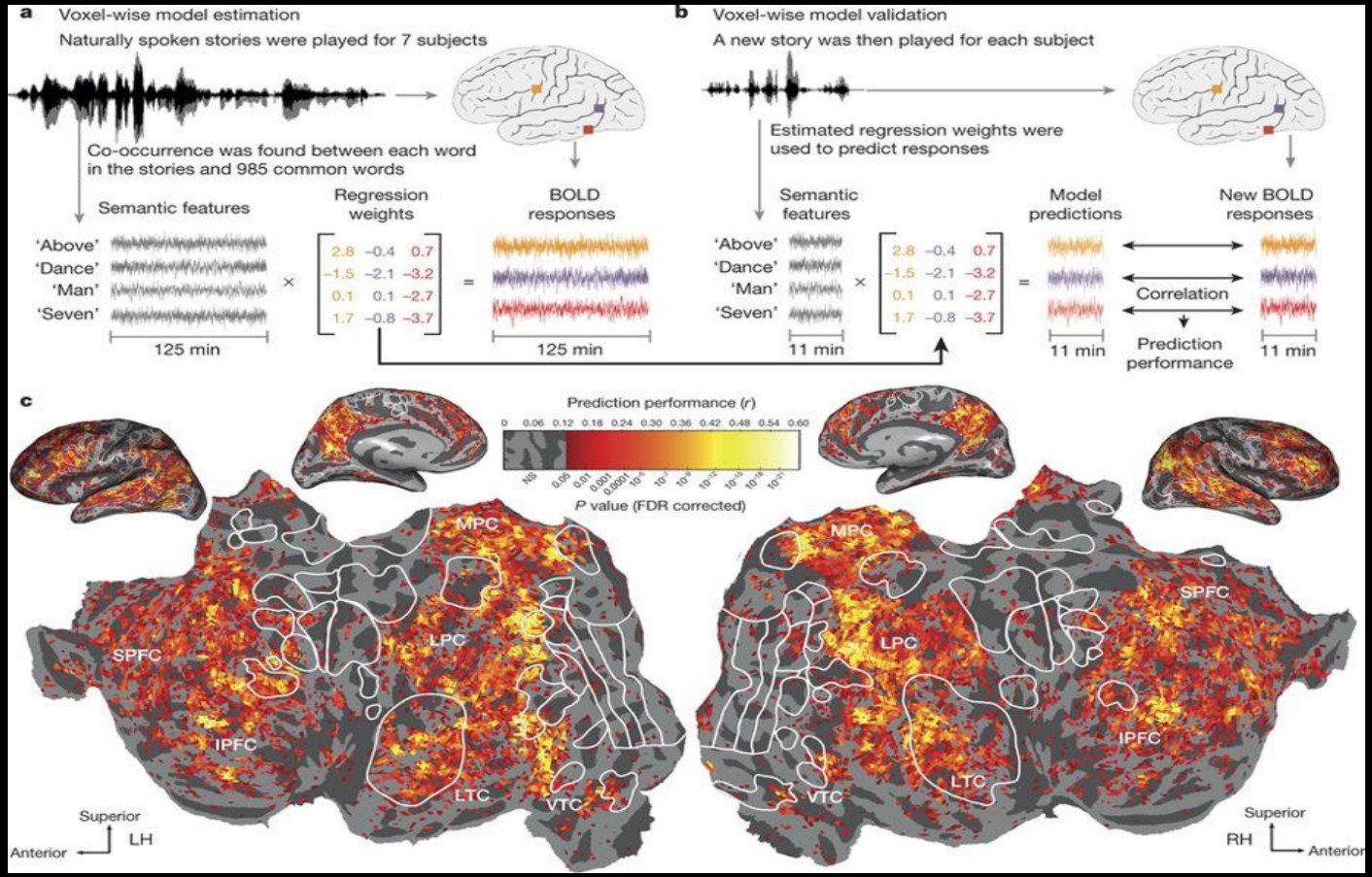


Simplified visualization of the persistence homological scaffolds: (a) placebo, (b) psilocybin

Alexander G. Huth, Wendy A. de Heer, Thomas L. Griffiths, Frédéric E. Theunissen, Jack L. Gallant, **Natural speech reveals the semantic maps that tile human cerebral cortex**, *Nature* **vol. 532**, pages 453–458 (28 April 2016)

The content of language is represented in regions of the cerebral cortex collectively known as the '**semantic system**' (**SSy**).

Semantic selectivity across the cortex is systematically mapped using a voxel-wise model of fMRI data, collected while subjects listened to hours of words. The **SSy** appears organized into intricate patterns that seem consistent across individuals and even languages. A generative model based on TDA was used to create the model resulting in the semantic atlas. Results suggest that most areas within **SSy** are consistent with specific semantic domains or groups of related concepts. Data-driven methods – common neuroanatomy and functional connectivity – provide an efficient way for mapping functional representations in the brain.


a
Voxel-wise model estimation
Naturally spoken stories were played for 7 subjects
Co-occurrence was found between each word in the stories and 985 common words
Semantic features
Regression weights
BOLD responses
'Above'
'Dance'
'Man'
'Seven'
125 min
125 min
b
Voxel-wise model validation
A new story was then played for each subject
Estimated regression weights were used to predict responses
Semantic features
Model predictions
New BOLD responses
Correlation
Prediction performance
11 min
11 min
11 min
c
Prediction performance (r)
P value (FDR corrected)
MPC
LPC
SPFC
IPFC
LTC
VTC
Superior
Anterior
LH
RH

The emerging picture

In the neurosciences TDA allows us to create a unifying language to understand brain function as it emerges from a multi-scale topological machine model.

One may use its language to model the overall brain's dynamical organization using the modeled topological dynamics as a 'lens' to improve diagnostic and predictive nosology, e.g., in healthy ageing and neurodegenerative diseases, providing a mechanistic insight on how coding at neuronal population level reverberates through scales up to brain regional level.

Brain is a complex organization of several units and each of its states can be defined as a particular configuration of individual brain units.

The space of brain states, '**Activity Landscape**' (AL), can be defined considering all possible configurations of all different brain units; each point of it corresponds also to a specific instant in time.

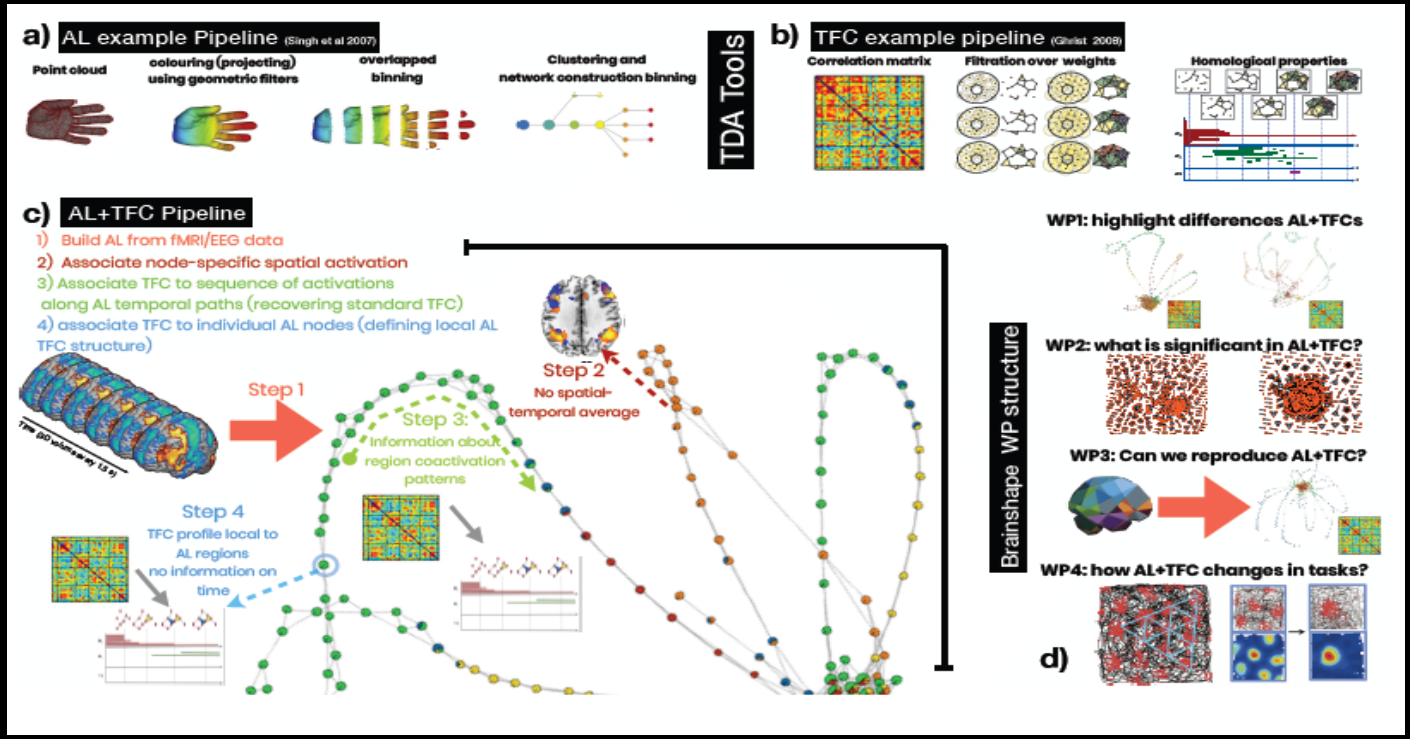
Then one can aim to model the entire **AL** of brain's trajectory of states in a given population, correlated with internal dynamics, '**Functional Connectivity**' (FC), of brain regions (or units, depending on scale).

The **AL** alone contains little information on how different units or regions exchange informations or coactivate: such patterns are instead well described in terms of complex networks (→ simplicial complexes) whose nodes and weighted edges (→ simplices of various dimensions) represent brain regions and the strength of interaction of pairs (→ multiplets) of regions.

The adoption of these discrete, combinatorial topological entities as descriptive language for **FC** is natural, as they are easily interpretable and also amenable to powerful analytical and statistical techniques.

Functional networks (**SC**s) require, to be dealt with, more sophisticated analysis (N.B. while **CN**s only encode pairwise interactions, **SC**s encode modulation of brain connectivity due to multiple regions interactions; of which there is mounting evidence in the case of multitasking actions).

The availability of a new description framework able to join temporal resolution and spatial (voxel-level) description of the **AL** with the biologically relevant information provided by **FC** is crucial to move beyond purely statistical characterization, towards mechanistic explanation of neuroimaging data, and provide biologically grounded nosology and biomarkers.



a) landscape, b) connectivity, c) combine landscape and connectivity; d) 1. highlight differences, 2. find significance, 3. Model, 4. How tasks change the combination.

Perspectives

Promoting the description of **FC** to '**topological**', i.e., the building blocks from networks to simplicial complexes, permits to describe and to quantitavely 'compare' the shape of spaces, grasp the structure of complex high-dimensional spaces in which neuronal/brain activity lives, and encode explicitly interactions of more than two elements. As the presence of topological effects is a signature for resting state **FC**, altered states, mental dysfunctions – invisible in a **CN** description also at small scales, we foster topological language at all scales.

Open questions

How does **AL**+**topoFC** change our picture of ageing and neuro-degenerative diseases [Parkinson, Alzheimer, Fronto-temporal Dementia, Amyotropic Lateral Sclerosis, Asperger]? **which** tools assess the significance of the observed topological alterations? **why** does a particular topology emerges? **what** is the link with topological encoding during tasks of increasing complexity?

BIOMEDIA
La condivisione del sapere

Per quanto concerne i moderatori, relatori, formatori, tutor, docenti è richiesta dall'Accordo Stato-Regioni vigente apposita dichiarazione esplicita dell'interessato, di trasparenza delle fonti di finanziamento e dei rapporti con soggetti portatori di interessi commerciali relativi agli ultimi due anni dalla data dell'evento. La documentazione deve essere disponibile presso il Provider e conservata per almeno 5 anni.

Dichiarazione sul Conflitto di Interessi

Il sottoscritto_____Mario RASETTI_____ in qualità di:

□ responsabile scientifico □ moderatore □ docente x relatore □ tutor

dell'evento "X CONGRESSO AIRMM - RISONANZA MAGNETICA IN MEDICINA 2019: DALLA RICERCA TECNOLOGICA AVANZATA ALLA PRATICA CLINICA"
Milano, 28-29 marzo 2019

da tenersi per conto di **Biomedia srl Provider n. 148,**

ai sensi dell'Accordo Stato-Regione in materia di formazione continua nel settore "Salute" (Formazione ECM) vigente,

Dichiara

χ che negli ultimi due anni NON ha avuto rapporti anche di finanziamento con soggetti portatori di interessi commerciali in campo sanitario

χ che negli ultimi due anni ha avuto rapporti anche di finanziamento con soggetti portatori di interessi commerciali in campo sanitario (indicare quali):

_____ _____
_____ _____
_____ _____
_____ _____