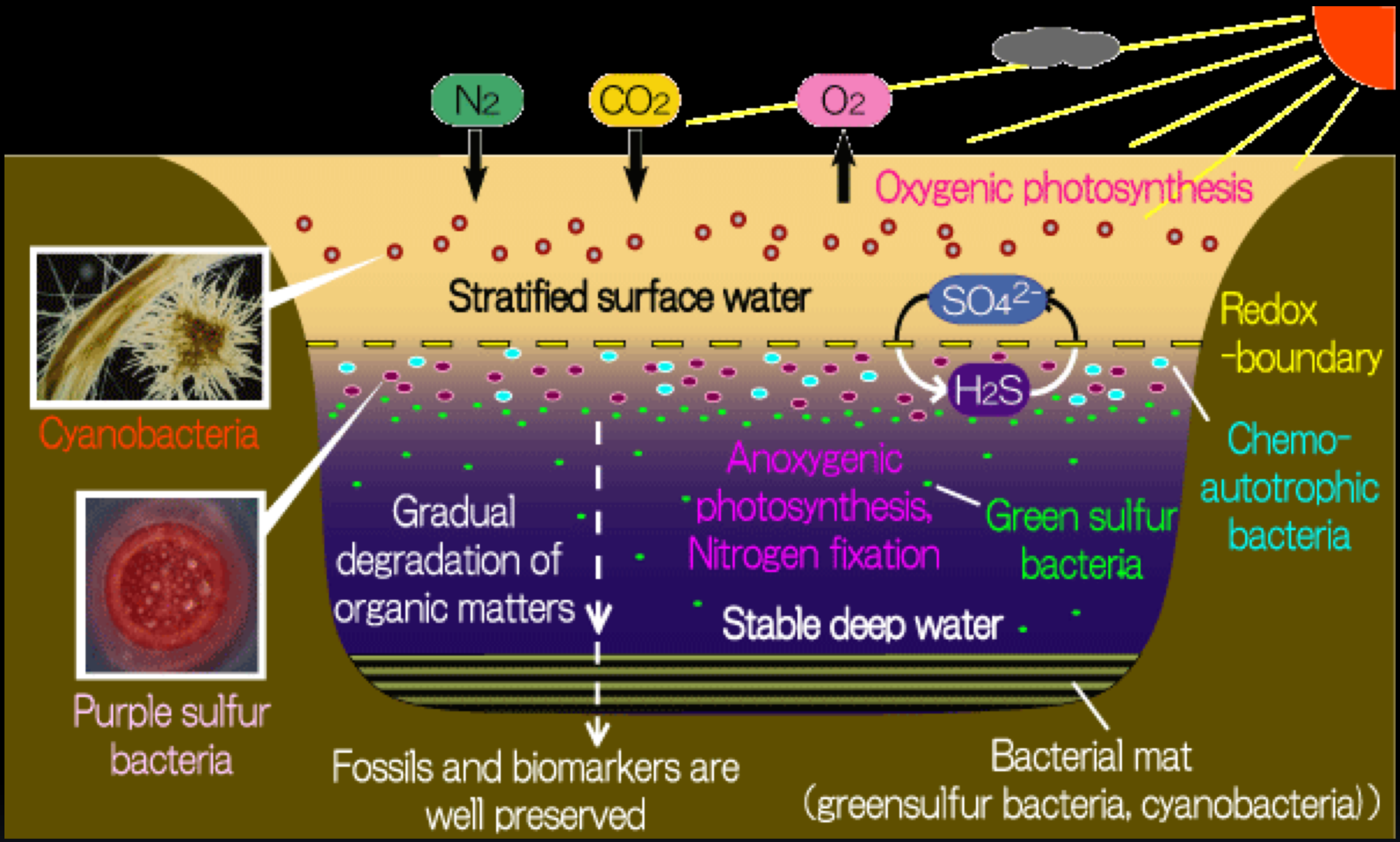
TOPOLOGY AND BIG DATA: A DREAMER’ S OVERVIEW

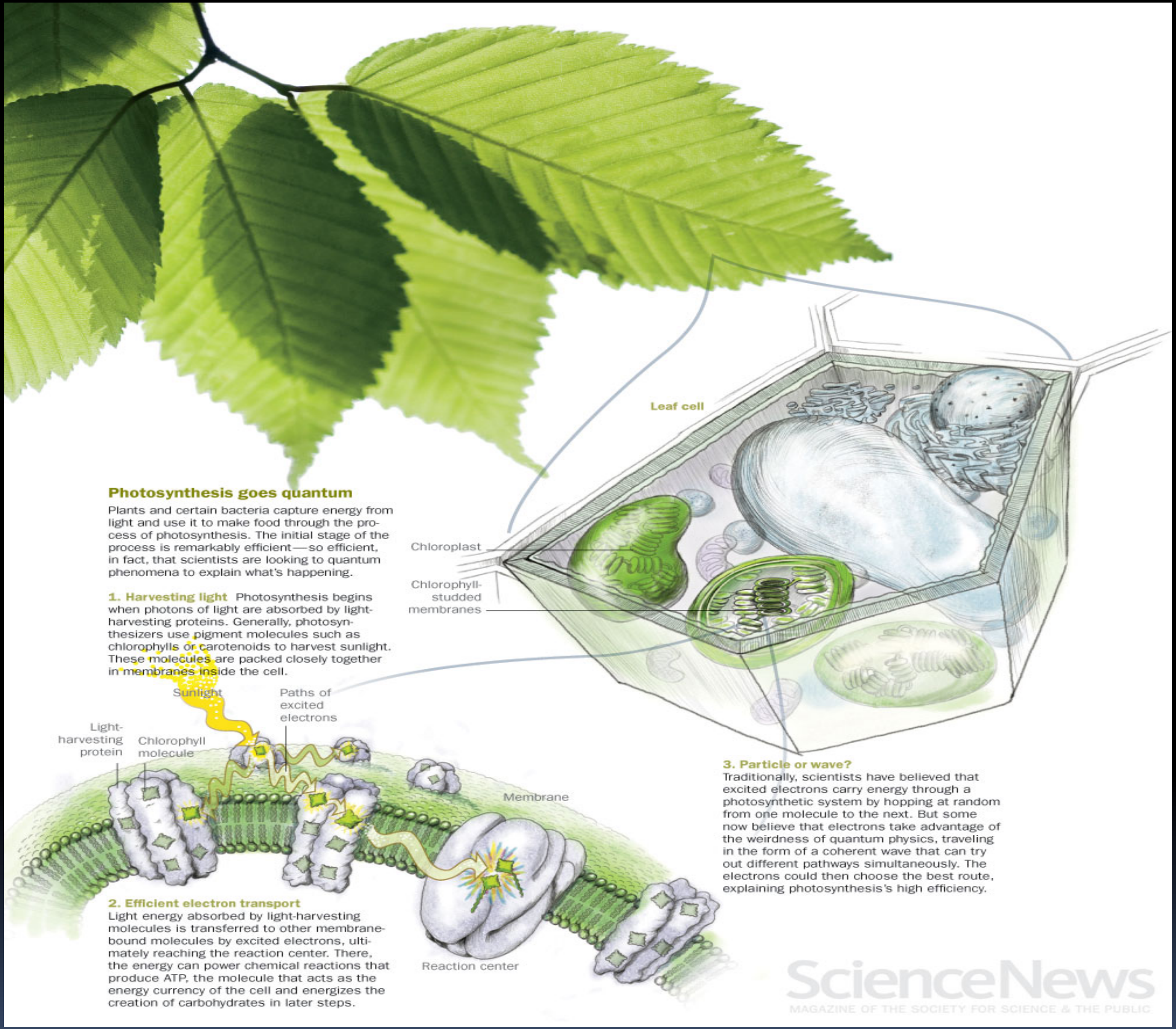
Mario Rasetti

ISI - Torino



Cyano-bacterium sulphureum

Efficient energy transport



Photosynthesis goes quantum

Plants and certain bacteria capture energy from light and use it to make food through the process of photosynthesis. The initial stage of the process is remarkably efficient—so efficient, in fact, that scientists are looking to quantum phenomena to explain what's happening.

1. Harvesting light Photosynthesis begins when photons of light are absorbed by light-harvesting proteins. Generally, photosynthesizers use pigment molecules such as chlorophylls or carotenoids to harvest sunlight. These molecules are packed closely together in membranes inside the cell.

2. Efficient electron transport Light energy absorbed by light-harvesting molecules is transferred to other membrane-bound molecules by excited electrons, ultimately reaching the reaction center. There, the energy can power chemical reactions that produce ATP, the molecule that acts as the energy currency of the cell and energizes the creation of carbohydrates in later steps.

3. Particle or wave? Traditionally, scientists have believed that excited electrons carry energy through a photosynthetic system by hopping at random from one molecule to the next. But some now believe that electrons take advantage of the weirdness of quantum physics, traveling in the form of a coherent wave that can try out different pathways simultaneously. The electrons could then choose the best route, explaining photosynthesis's high efficiency.

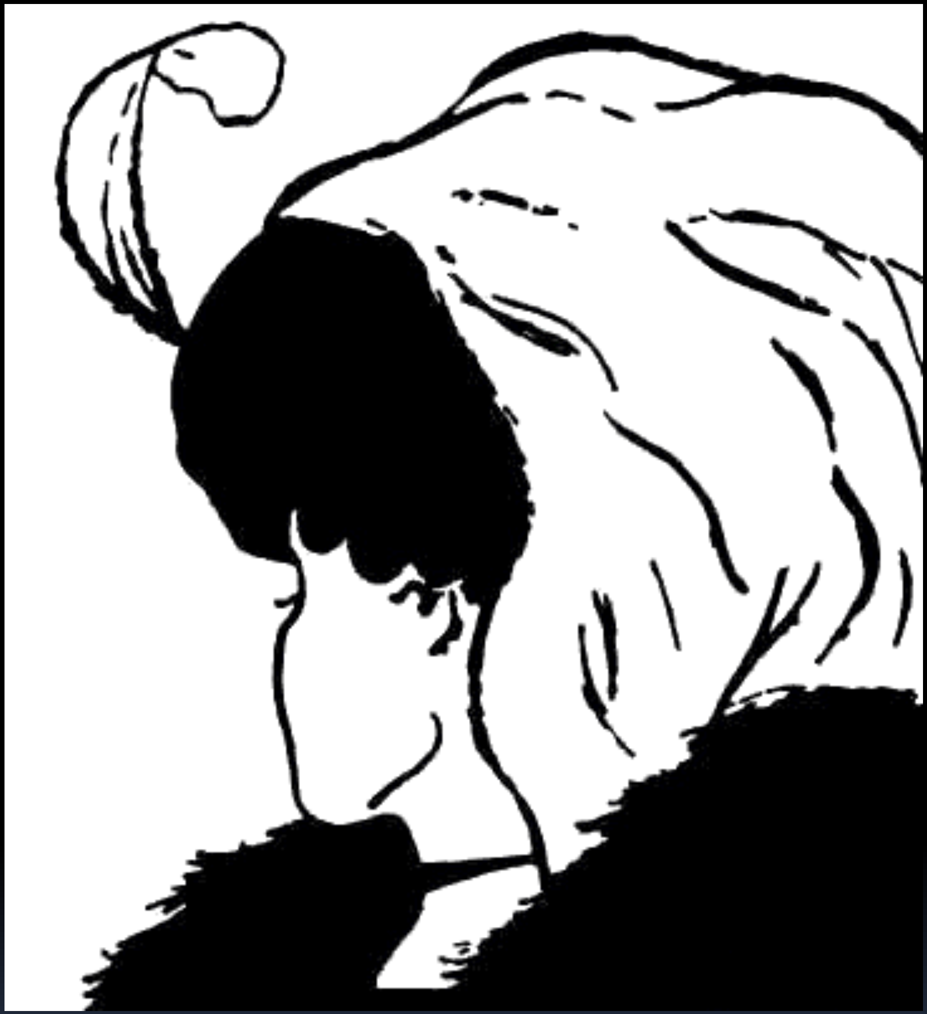
ScienceNews
MAGAZINE OF THE SOCIETY FOR SCIENCE & THE PUBLIC

Photosynthesis

$i\hbar \frac{\partial \Psi}{\partial t} = -\frac{\hbar^2}{2m}\frac{\partial^2 \Psi}{\partial x^2} + V(x)\Psi(x,t)$
ScienceNews
MAGAZINE OF THE SOCIETY FOR SCIENCE & THE PUBLIC

Migrant birds exploit quantum effects in their visual system to 'feel' the magnetic field

Francisco Varela



Ambiguous pictures
Intelligence

Topology and Big Data

- An important feature of modern science (and of society) is that a huge amount of data is produced at unprecedented rates: we recently passed the point where more data is collected than we can physically store.
- Analogously, living matter must have the ability to handle data, in situations where the system is barely able to keep pace with the data produced.

Geometry and topology are very natural tools to handle large, high-dimensional, complex spaces of data:

- *Qualitative information is relevant*: the user aims to obtain *knowledge*, i.e., to understand how data is organized on large scale. Global, even though partly qualitative, information is needed.

- *Metrics are not theoretically justified*: in physics, phenomena support clean theories which tell exactly what metric to use, in biology this is much less clear.
- *Coordinates are not natural*: data is conveyed and received in the form of vectors, whose components are not natural in any sense. One should not consider properties of the data which depend on the choice of coordinates.

- *Summaries are more valuable than parameter choices*: conventional method of handling data is *building* a graph whose vertex set is a set of points (*cloud*) and two points are connected by an edge if their distance is $\leq \varepsilon$, then try to determine the optimal choice of ε.

  It is however much more informative to maintain the entire *dendrogram*, which gives at once a summary of the relevant features of the clustering under all possible values of ε .

We need to develop mechanisms to know how global features vary under changes of parameter.

The idea that emerges is that the methods to adopt should be inspired by *topology*:

- *Topology* is the branch of mathematics which can deal with qualitative geometric information (connectivity, classification of loops and higher dimensional manifolds) in data space.

- *Topology* studies geometric properties in a way which is less sensitive to metrics than geometric methods: it ignores the value of distance functions and replaces it with the notion of connective nearness (proximity).
- *Topology* studies only properties of geometric objects which do not depend on the coordinates, but on intrinsic geometric features. It is coordinate-free.

- Useful relationships in topology naturally involve continuous maps between the objects, hence are a manifestation of *functoriality*. Invariants should be related not just to objects, but also to maps between objects.

  Functoriality is central in algebraic topology because for homological invariants it permits their computation from local information. It reflects a categorical structure.

- We argue that most information about topological spaces can be obtained through *simplicial approximation*: here piecewise linear, discrete realization of functoriality enters into play.

⇒ There ensues the strategic process here proposed :

i) replace the (huge) set of points that constitute the space of data with a family of simplicial complexes, parametrized by a 'proximity parameter ‘, i.e., convert the data set into a global topological object;

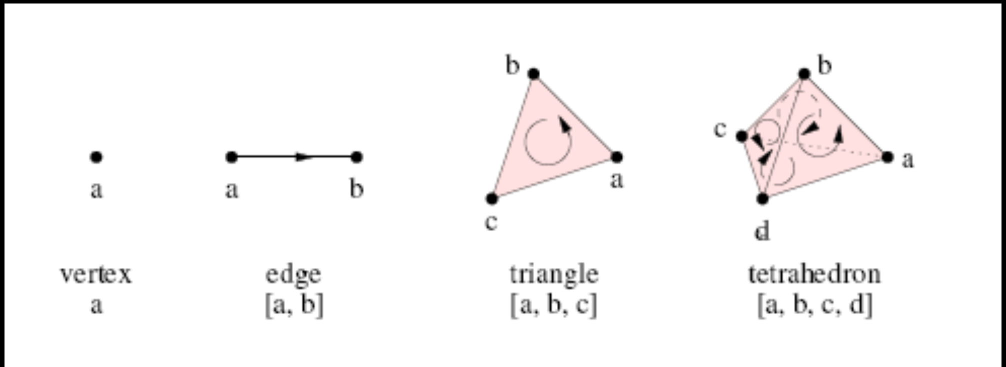
ii) handle topological complexes by the tools of algebraic topology, in particular the theory of *persistent homology*, characterized by a running parameter ε ;

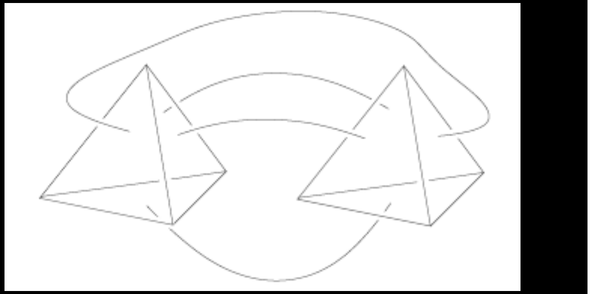
iii) encode the persistent homology of data sets in a parameterized version of Betti numbers.

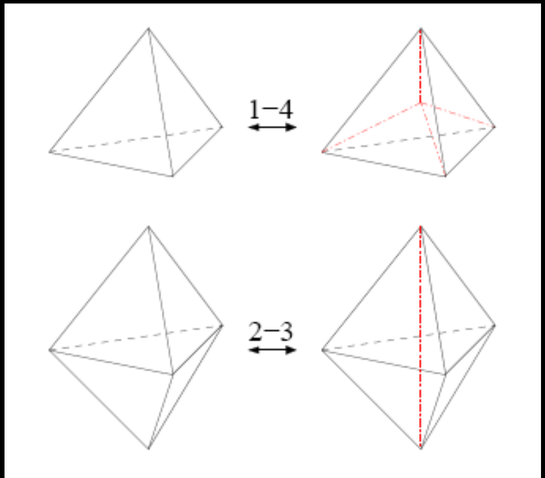
- Data is typically represented by (unordered) sequences of points in some *n*-dimensional 'space of data'. Correlation patterns of data provide the relevant information about the phenomena which data represents.

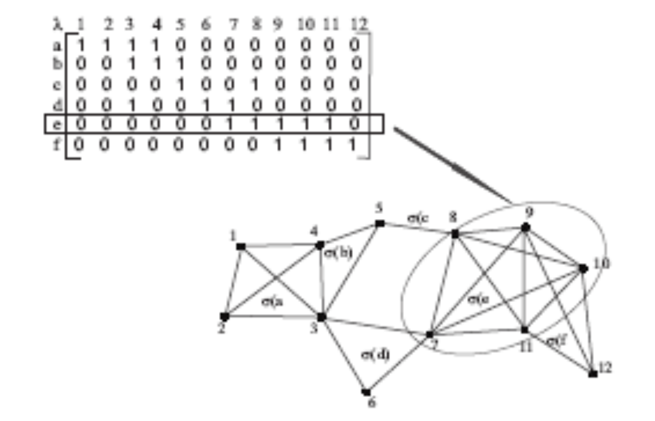
  Point cloud data is a typical instance of data set for which such significant global features are present.

- The conventional way to convert a collection of points in data space into a global object is to use the point cloud as vertex set of a graph Γ, whose edges are determined by proximity.
- Γ captures data connectivity, but ignores a wealth of higher order features, instead well discerned thinking of Γ as the scaffold of a higher-dimensional object: the *simplicial complex* – a PL space built from simple pieces (simplices) identified combinatorially along their faces – obtained by completion of Γ.







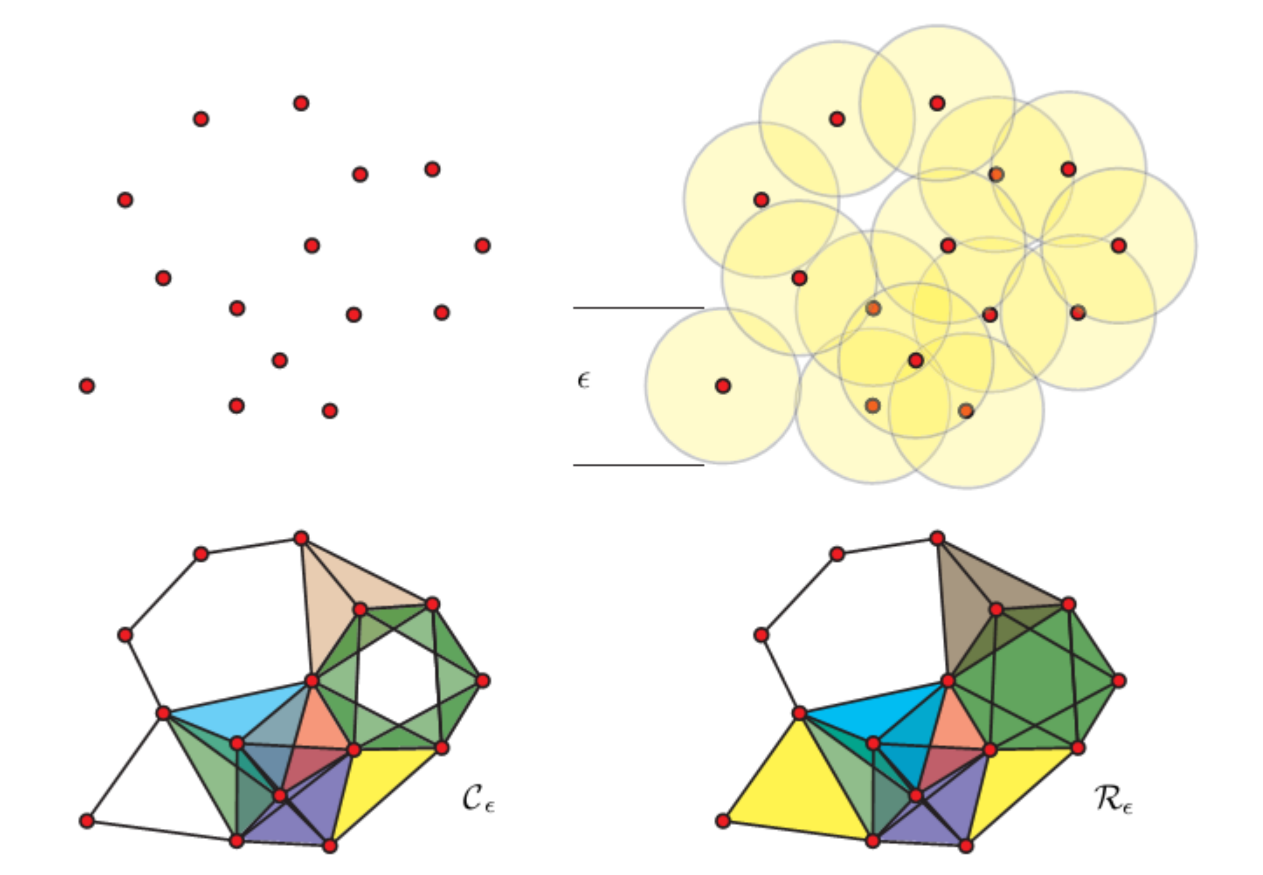


$$\begin{array}{c|cccccccccccc}
\lambda & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 \\
\hline
a & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
b & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
c & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
d & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\
e & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 0 \\
f & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1
\end{array}$$

The two most natural complexes are:

i) the Čech complex [k-simplices are unordered $(k+1)$-tuples of points whose $\varepsilon/2$-ball neighborhoods have a point of intersection]. It has the homotopy type of the union of closed balls of radius $\varepsilon/2$ around the point set.

ii) the Rips complex [k-simplices are unordered $(k+1)$-tuples of points pairwise within distance ε].

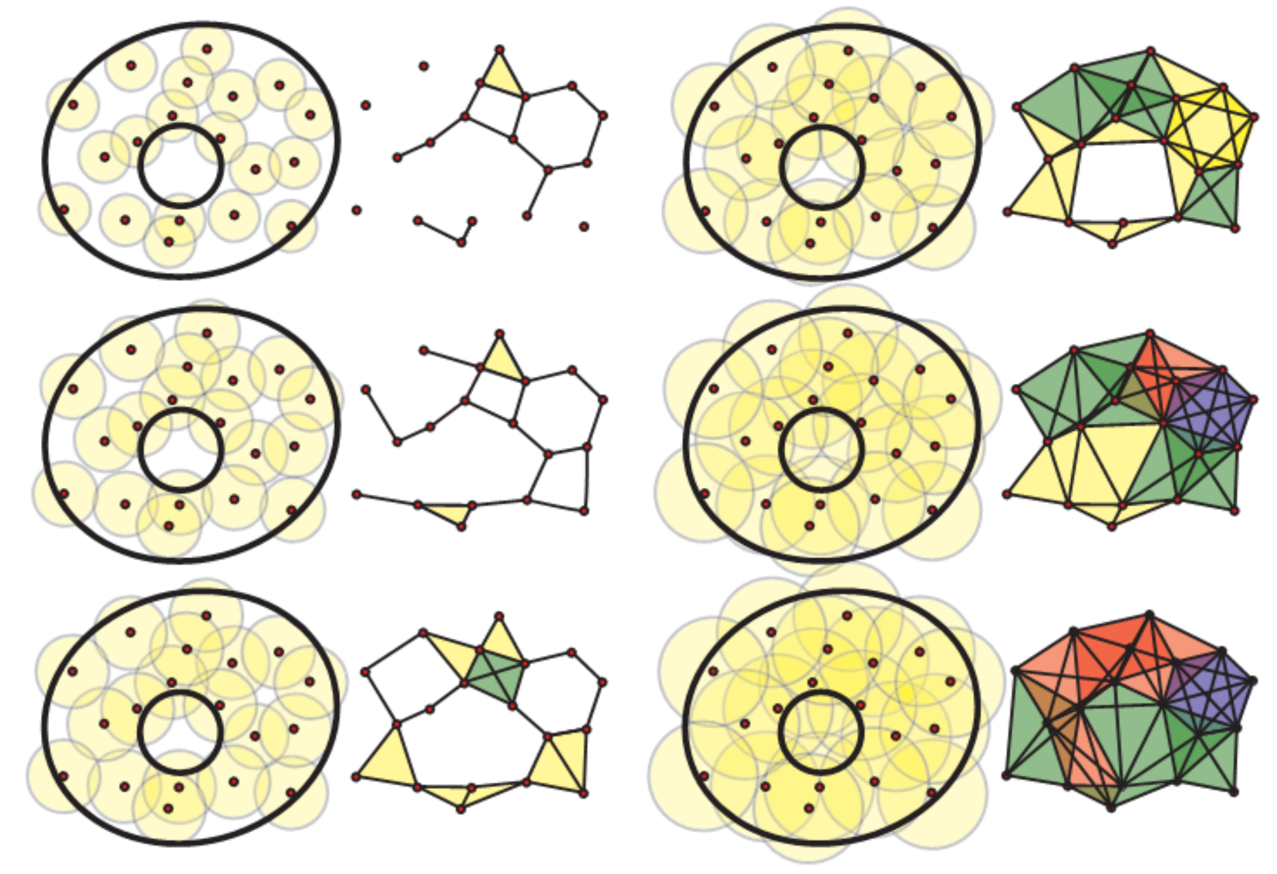


A fixed set of points [upper left] can be completed to Čech complex $\mathcal{C}_\epsilon$ [lower left] or to a Rips complex $\mathcal{R}_\epsilon$ [lower right] based on a proximity parameter ϵ [upper right]. This Čech complex has the homotopy type of the $\epsilon/2$ cover ($S^1 \vee S^1 \vee S^1$), while the Rips complex has a wholly different homotopy type ($S^1 \vee S^2$).

- The most important set of invariants of a topological space X, is its collection of homology groups, $H_i(X)$. Basic ingredients for computing such groups are *Betti numbers*; the i-th Betti number, $b_i = b_i(X)$ being the rank of $H_i(X)$.

- However, Betti numbers *per se* are not enough to decide which invariants are essential and which can be safely ignored.
- An additional notion must be introduced: *homology persistence*.

  Given a parameterized family of spaces, it permits to identify those topological features which persist over a significant parameter range (to be considered as signals with short-lived features as noise).



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

- In the evolution of the complex as it is constructed starting from the empty set and adding simplices, the progressive sequence of sub-complexes is referred to as *filtration*.
- Aim of persistence homology is to measure the 'lifetime' of the topological properties of a simplicial complex under filtration.
- Finally, when the (asymptotic) stable complex is identified, patterns in data space are derived obtaining Morse numbers (in discrete context) from Betti numbers.

- M compact n-dimensional smooth manifold;
- $f : M \rightarrow \mathrm{R}$ smooth function over manifold M;
- $p \in M$ *critical point* of f : in local coordinates around p

$$\partial f/\partial x_1 = \cdot \cdot \cdot = \partial f/\partial x_n = 0 \ ;$$

- p *non degenerate* : *Hessian matrix* **H** [elements $H_{ij} = (\partial^2 f/\partial x_i\, \partial x_j)$] non-singular;
- for p non degenerate, *Morse index* of p = number of negative eigenvalues of **H** at p.
- *Morse function* a smooth function f with only non-degenerate critical points

- **Morse Lemma** If $p \in M$ is a non-degenerate critical point of f then $\exists$ a neigbourhood U, $p \in U$, and local coordinates $y_1, \ldots, y_n$ such that $y_i(p) = 0$ and

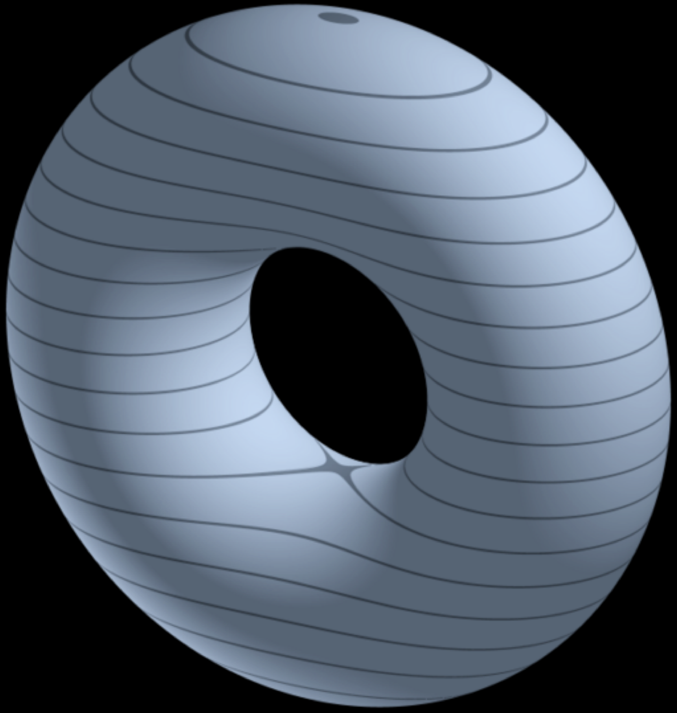
$$f(q) = f(p) - y_1(q)^2 - \cdots - y_\lambda(q)^2 + y_{\lambda+1}(q)^2 + \cdots$$
$$\cdots + y_n(q)^2, \quad \forall\, q \in U,$$

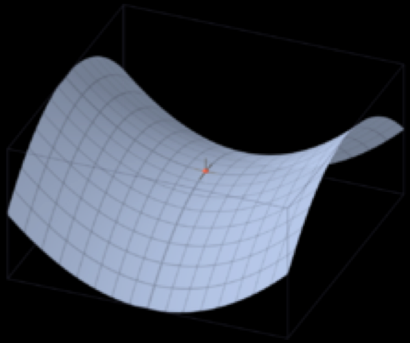
where λ is the Morse index of f at p.

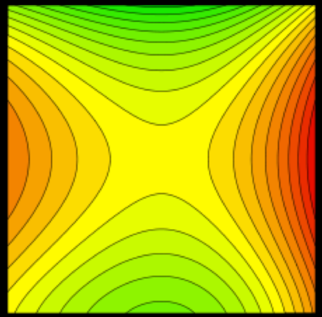
- Basic idea of Morse theory is that the homotopy type of submanifold $M^a = \{p \in M \mid f(p) \leq a\}$ changes only at critical points of f. If there is no critical value in $[a, b]$, then gradient flow of f provides a diffeomorphism between M^a and M^b.
- At critical value a one can suppose, under small perturbation of f, that there is only one critical point p, $f(p) = a$. The result means that one can get $M^{a+\varepsilon}$ from $M^{a-\varepsilon}$ by attaching a handle B_λ, a cell of dimension of the index λ of f at the critical point.

- This handle B_λ can be constructed via the description given by Morse Lemma and the attaching map is between ∂B_λ and $M^{a\text{-}\varepsilon}$.
- Using results of *Whitehead* in homotopy theory, the following theorem is proved:

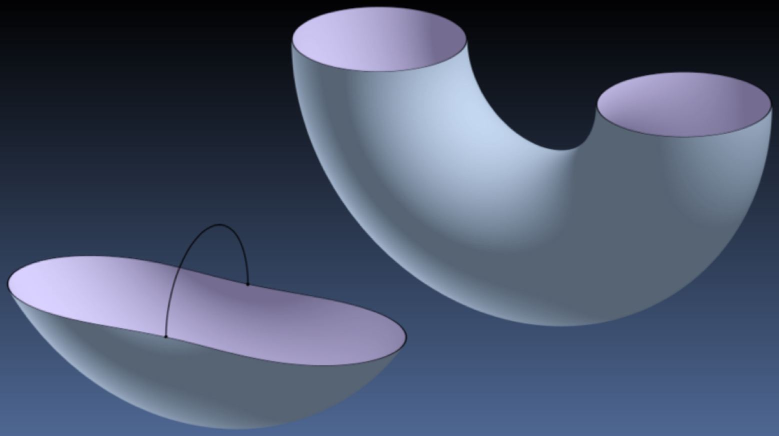
- **Theorem 1.** If f is a Morse function on M such that M^a is compact for each $a \in$ R then M has the homotopy type of a cell complex with one λ-dimensional cell for each critical point of index λ.

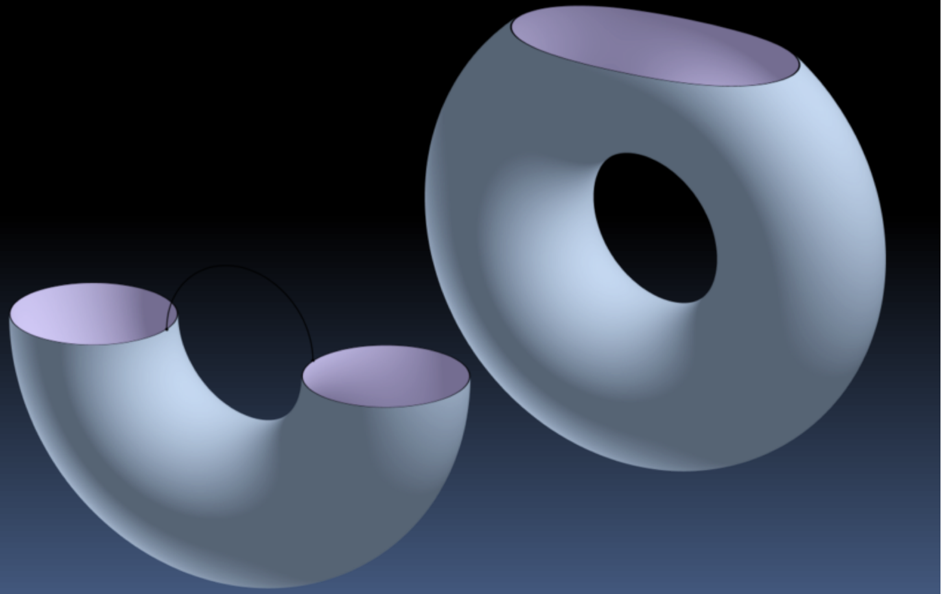






Morse





- If b_k denotes the *k*-th Betti number of *M*, dimension of the (co)homology group $H^k(M;\mathrm{R})$, and m_k is the number of critical points of Morse function *f* with index *k*, then
- **Theorem 2. (Morse inequalities)** $b_k \leq m_k$. Moreover,

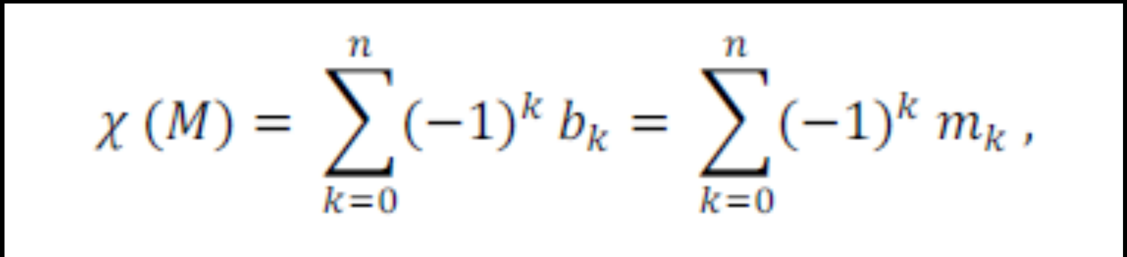
$$b_0 \leq m_0$$

$$b_1 - b_0 \leq m_1 - m_0$$

$$b_2 - b_1 + b_0 \leq m_2 - m_1 + m_0$$

$$\cdots$$

and

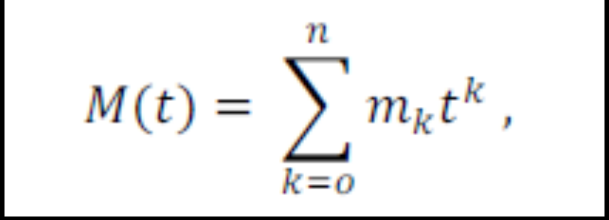


$$\chi\,(M) = \sum_{k=0}^{n} (-1)^k\, b_k = \sum_{k=0}^{n} (-1)^k\, m_k\,,$$

where χ is Euler characteristic of *M*.

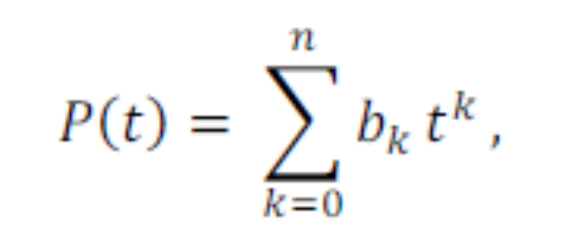
This is a special case of a stronger form:

in terms of *Morse polynomial*:



$$M(t) = \sum_{k=o}^{n} m_k t^k ,$$

and *Poincaré polynomial*:



$$P(t) = \sum_{k=0}^{n} b_k t^k ,$$

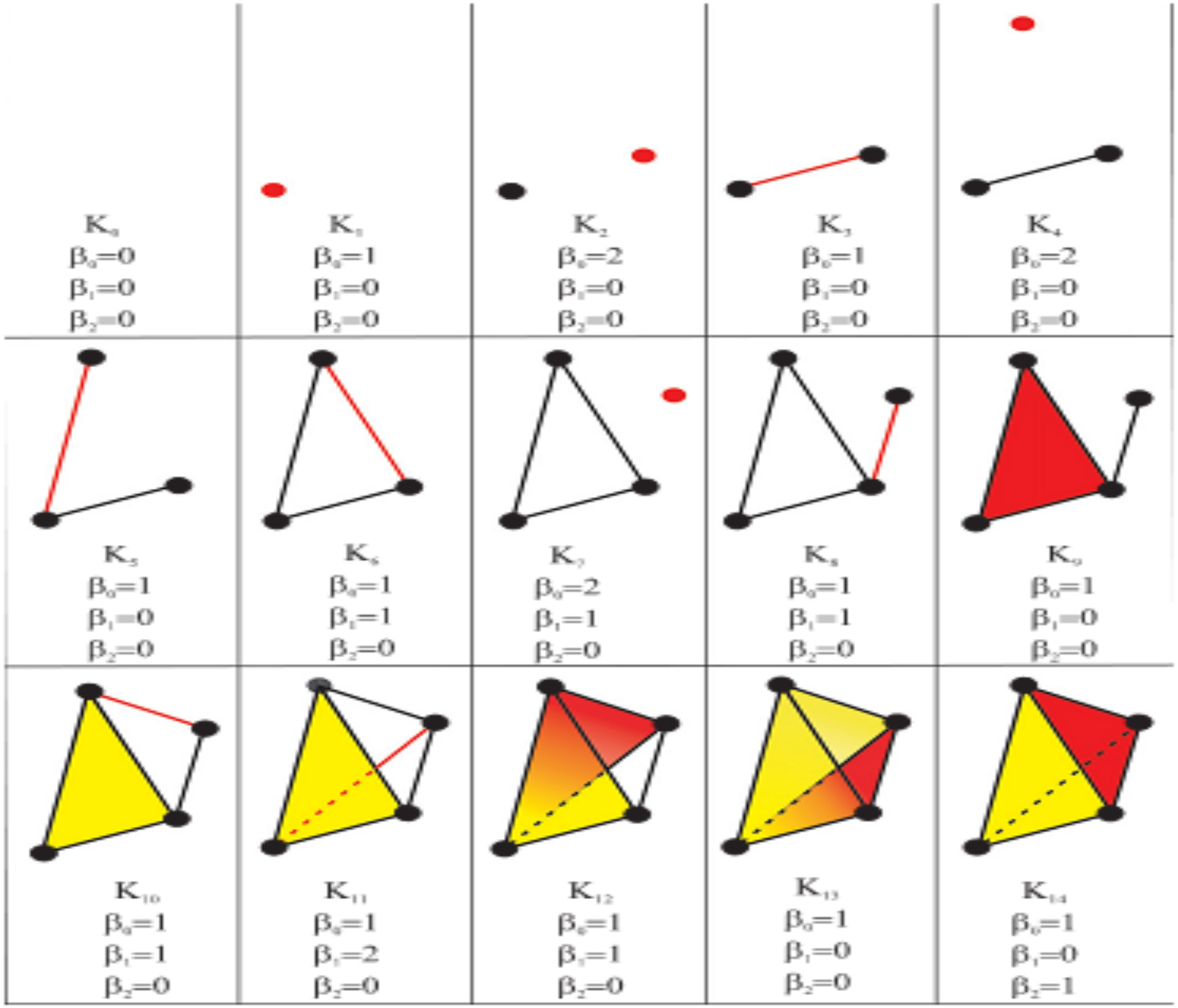
one may express Morse inequalities symbolically as $M(t) \geq P(t)$.

In this notation:

- **Theorem 3.** Let $f : M \to \mathrm{R}$ be a Morse function, on a compact manifold *M*. Then

$$M(t) - P(t) = (1 + t)Q(t),$$

for some polynomial $Q(t)$ such that $Q(t) \geq 0$.


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 (tetrahedron) and its topological characterization. At each stage a vertex or a face is added. represented in red.

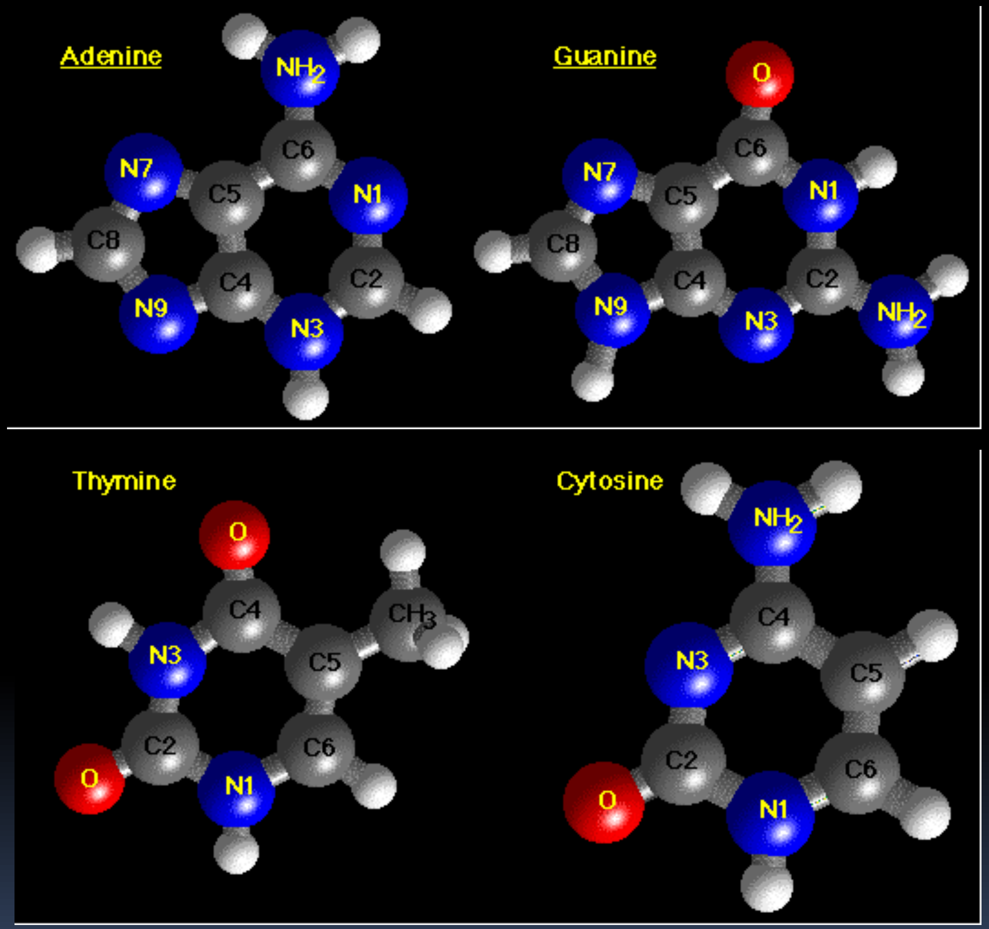
- Standard invariants were *created* to distinguish between things: it is their intrinsic definition that makes clear what kind of (manifest) properties they reflect.
- But other invariants were instead *discovered*, whose construction, based on TQFT, provides information about purely topological properties we were unable to detect, nor even to hint.

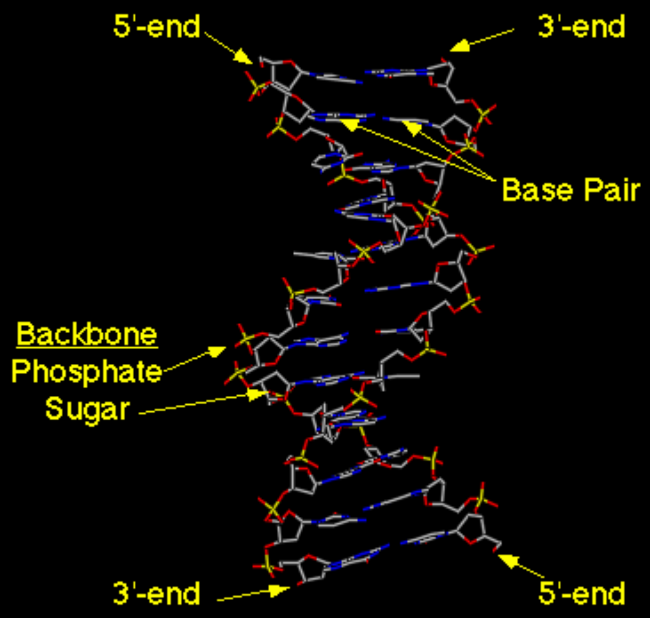
- For example, the topological type of a smooth, closed, oriented surface *S* is fully determined by its genus *g* $(\equiv b_1)$, which is but the number of handles of *S*. *g* is obtained from Euler number, the topological invariant $\chi = 2 - 2g$, which can be easily evaluated upon tessellation of S by Euler's formula $\chi = V + F - E$; with V= # Vertices; F = # Faces; E = # Edges.

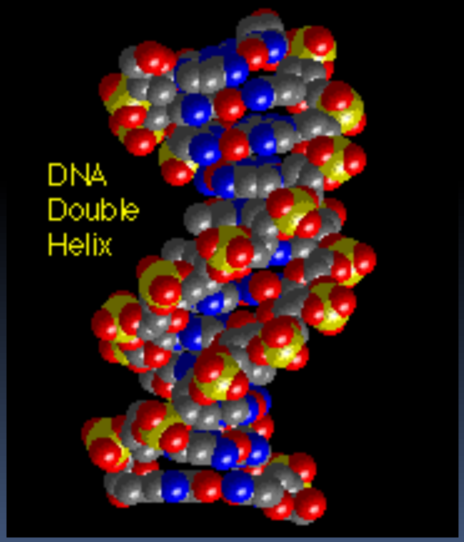
The dreamer’ s view

- A symmetry over the space of quantum deforms of a finite-dimensional configuration manifold *M*, in the form of a quotient action of the 'motivic' Galois group has a non-trivial realization – linked with the action of $\mathbf{G}_{\text{G-T}}$ over the extended moduli space of some generalized quantum field theories. These can be observed only when *q* is a root of unity, because then the center of $U_q(sl_2)$ is much more extended and is strongly non-trivial with respect to the case of generic *q* (and $q=1$).

- The depth of these rules has unimaginable reach. First, it implies that actually the whole supporting structure not only of living matter, but of nature itself as a whole, is affected by quantization by deformation. Even space and time themselves can no longer be thought of as an immutable stage over which events take place, because in turn they are directly involved in the process of quantization, which 'deforms' the structure that every classical observer intuitively attributes to them.


Adenine
NH2
C6
N7
C5
N1
C8
C4
C2
N9
N3
Guanine
O
C6
N7
C5
N1
C8
C4
C2
N9
N3
NH2
Thymine
O
C4
CH3
N3
C5
C2
C6
O
N1
Cytosine
NH2
C4
N3
C5
C2
C6
O
N1


5'-end
3'-end
Base Pair
Backbone
Phosphate
Sugar
3'-end
5'-end


DNA
Double
Helix

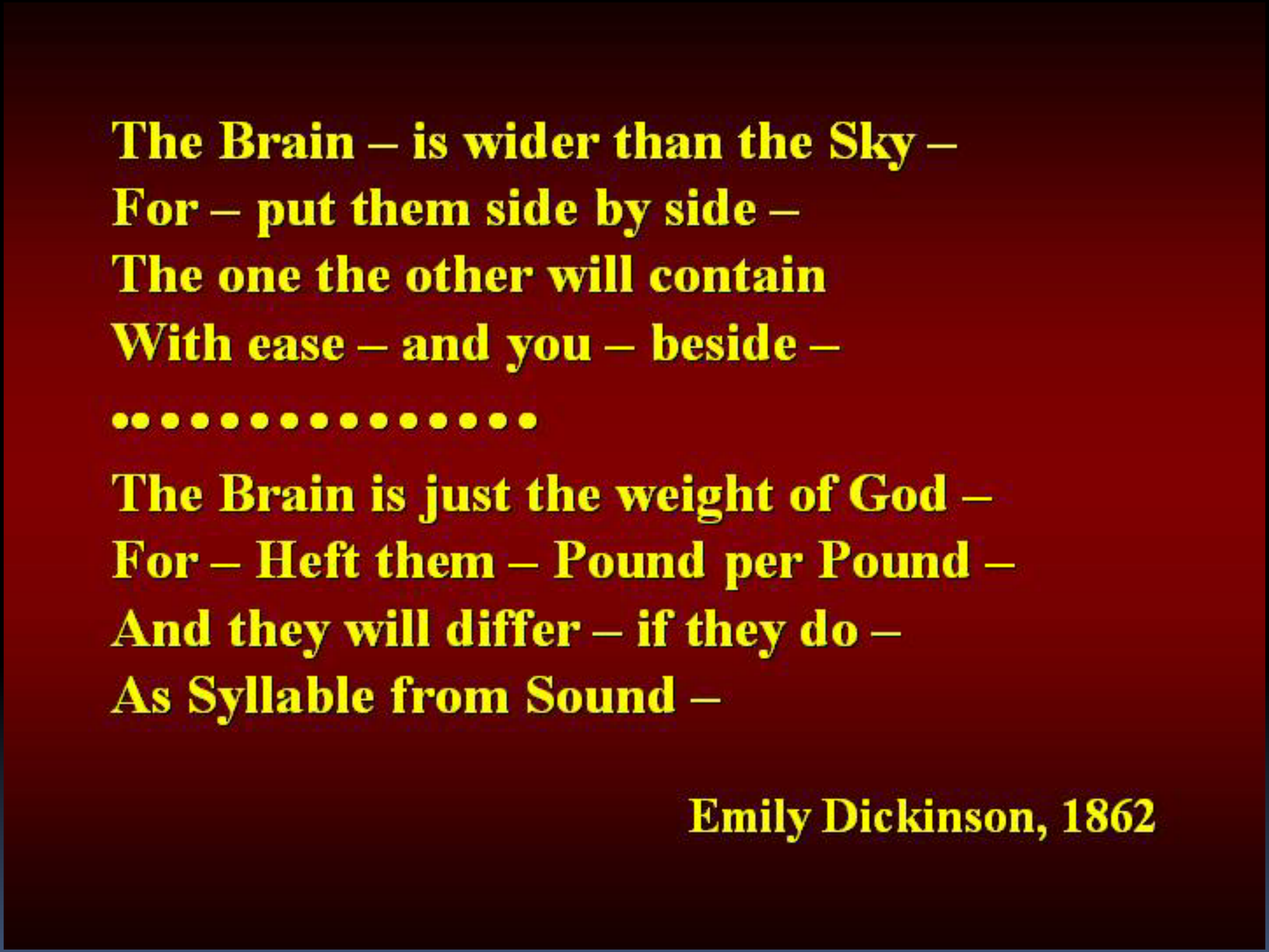
- We claim that it is deformation by quantization of the finite-dimensional manifold "ambient" space of living matter which gives rise to the infinite-dimensional space necessary to realize the action of $\mathbf{G}_{G\text{-}T}$. Also the algebra of periods of Tate is generated by this same process of quantization by deformation. Thus, even if a quantum computer (the automaton) is thought of as derived by a procedure of quantization by deformation of a classica model, the algebra of numbers physically 'natural' for it is not ordinary arithmetic, but a wider structure, indeed just $\Pi_{\mathbf{z}}$.

In a conventional quantum computer only rationals can be realized in a natural way, but the periods of Π_z emerge when an observer wants to probe non-local sub-systems, namely verify the long-time behavior of the machine for all its possible types of input data, exploring its global topology. I.e., the 'numbers' of Π_z, which do not have any classical analogue, emerge from an operation similar to what computer scientists refer to as "testing of the (quantum) software". In the language of formal logics they belong to a meta-level, as they refer to tests of the computer performed from outside, not to the system itself.

- In a conceptual scheme of this sort not only there is no contradiction with the perspective of quantum information, but the latter is encompassed in a much wider and more general structure: because also "testing" a device is, under any point of view, a possible, permitted and necessary physical procedure.
- One can finally thus concisely synthesize the perspectives derived from the framework discussed:

- i) all physical systems, including living ones, must be representable "in real time" by a quantum computer (a quantum automaton) and, vice-versa, quantum computers should be describable as physical systems which realize the quantization by deformation of a classical machine; ii) observables of the (quantum) living physical world are those and only those which can be determined by the observation of quantum computers by quantum computers.

- It is this set of dynamical properties, together with the structural ones characterizing the automaton, that arguably provide the complete conceptual reference frame to answer - giving a solution to the problem of a description consistent with quantum physics and complexity science of living matter, its properties and functions - the question posed at the beginning:



The Brain – is wider than the Sky –
For – put them side by side –
The one the other will contain
With ease – and you – beside –

••••••••••••••••

The Brain is just the weight of God –
For – Heft them – Pound per Pound –
And they will differ – if they do –
As Syllable from Sound –

Emily Dickinson, 1862

There is no Frigate like a Book
To take us Lands away
Not any Coursers like a Page
Of prancing Poetry –
This Travel may the poorest take
Without oppress of Toll –
How frugal is the Chariot
that bears the Human soul.

Perception of an Object costs
Precise the Object’s loss –
Perception in itself a Gain
Replying to it’s Price –
The Object Absolute – is nought –
Perception sets it fair
And then upbraids a Perfectness
That situates so far –