Data and Topology (Invitation to "Persistency").

Dan Burghelea

Department of mathematics
Ohio State University, Columbuus, OH

Alba Julia , September 2009

CONTENTS

- **Data analysis**
- **Topology**
- **New Topology- (inspired by Data analysis)**

Data Analysis

1. What is **Data**?
2. How the Data is obtained?
3. Features of the Data
4. What do we want to infer from Data ?
5. A few examples
6. Why topology can help?

1. Data =

POINT CLOUD DATA =

a FINITE (but very large) set of vectors in

an EUCLIDEAN SPACE (most often of high dimension).

hence defines

a metric space (X, d) **of finite cardinality.**

2. Data are obtained :

a. By sampling a geometric objects by points or a probability distribution concentarted near the geometric object (a collection of points (three coordinates) in $\mathbb{R}^3$.)

b. As a collection of two dimensional black-white pictures of a three dimensional object taken by camera ; each 2D picture is regarded as a vector in the pixel space with a gray scale coordinate for each pixel. If camera has 100×100 pixels a collection of vectors in $\mathbb{R}^{10000}$.

c. As a list of measurements of parameters of a collection of oblects/individuals; for example observations on the patients (in a hospital) suspected of diabetes by measuring parameters (in example below 6 parameters) involving insulin response, glucose tolerence, relative weight and others).

3. SPECIAL FEATURES (of data)

The large cardinality (of vectors) and high dimensionality (of the Euclidean space)

a. makes the visualization difficult,

b. makes **noise and incompleteness**
the researcher wants to delete or ignore unavoidable,

c. involves often irrelevant parameters
which the researcher wants to ignore.

4. We want:

- In case of sampled geometrical objects :

  to derive geometric and topological features from the data without reconstructing the object entirely.

- In case of apriory unstructured observations:

  to organize the data as a geometric object or as a collection of geometric objects in order to guess qualitative features. Invariants of such shapes (dimension, connectivity, Betti numbers etc) can hopefully be interpreted as qualitative features.

- to find the number of relevant parameters, detect and eliminate noise, detect incompleteness.

5. Example 1.
Lung cancer imaging.

- 3D radiological images of cancerous lungs shows both tumors and blood vessels as areas of increased density.
- Blood vessels show up as long tunnels in the image
- Tumors show up as balls.
- **Question**: How to distinguish automatically between tumors and blood vessels ?

Example 2. **Diabetes Patients**
(after Miller-Reaven Study) from G Carlsson's paper

- Study carried out in 1976 on 145 patients at Stanford Hospital; Most of patients had symptoms of diabetes although some were normal

- For each patient 6 metabolic variables (involving insulin response glucose tolerence, relative weight .....) were measured and recorded in a 6 dimensional space. Hence a point cloud of 145 points in $\mathbb{R}^6$
- **Questions**: Find the relevant number of metabolic variables needed to detect the diabetes. Find qualitative features (type of diabetes, etc).

6. WHY TOPOLOGY CAN HELP?

TOPOLOGY provides :

1. methods to convert a metric space into "nice shapes"; simplicial complex or family of simplicial complexes.

2. (algorithmically) computable invariants (for simplicial complexes) like homology, Betti numbers, Euler-Poincare characteristic, torsion.

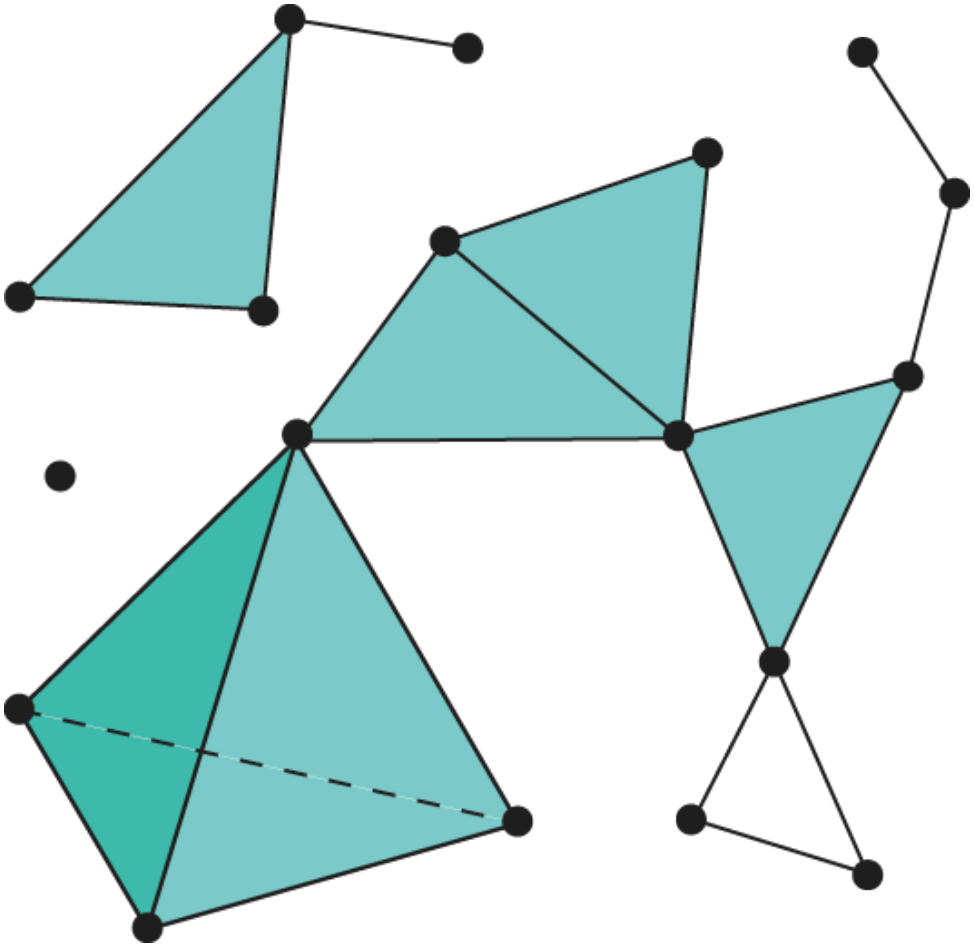
3. new concepts (persitent space) and new invariants (barcodes, persistent diagrams...).

SIMPLICIAL COMPLEXES

Definition

- A **solid** $k-$ **simplex** is the convex hull of $(k+1)$ linearly independent points .
- A **geometric simplicial complex** K is a a "nice subspace of an Euclidean space " precisely a union of **solid simplicies** which intersect each other in faces (subsimplexes) .
- An **abstract simlicial complex** is a pair (V, Σ) with:
  V a finite set,
  Σ a family of nonempty subsets of V,
  so that $\boxed{\sigma \subseteq \tau \in \Sigma \Rightarrow \sigma \in \Sigma}$.

An abstract simplicial complex determines a geometric simplicial complex and vice versa.



- Topology provides methods to assign to a metric space (X, d) and $\epsilon > 0$ simplicial complexes:

- The simplest and most familiar topological invariant is **homology**. The homology can be calculated for simplicial complexes

HOMOLOGY

The homology (with coeficients in $\mathbb{Z}_2$ or $\mathbb{R}$) of a simplicial complex K can be derived from a chain complex whose

- $k-$ component is the vector space $C_k(K)$ generated by $k-$ simplexes (the elements of Σ consisting of exactely $(k+1)$ elements of V)
- and the liner maps $\boxed{\partial_k : C_k(K) \to C_{k-1}(K)}$ defined as a matrix with entries $0, 1, -1$;

The entries indicate IF and HOW a $(k-1)-$ simplex σ is a face of a $k-$simplex τ.

$$\boxed{H_k(K) = \ker(\partial_k)/\mathrm{im}(\partial_{k+1})}$$

Properties of Homology

- Two simplicial complexes (topological spaces) which are homotopy equivalent have isomorphic homology .
- A (simplicial) map $f : K \to K'$ induces a linear map $f_k : H_k(K) \to H_k(K')$ and for f and g homotopic $f_k = g_k$.
- The assignment $K \rightsquigarrow H_k(K)$ is functorial and enjoys a number of properties which make these functors computable.

Simplicial complexes associated with (X, d) and $\epsilon > 0$.

CECH COMPLEX, $\mathcal{C}_\epsilon(X, d) := (\mathcal{X}, \Sigma_\epsilon)$, $X \subset \mathbb{R}^N$.

- $\mathcal{X} = X$
- $S_k := \{(x_1, x_2, \cdots x_{k+1})|$ iff $B(x_1; \epsilon) \cap \cdots B(x_{k+1}; \epsilon) \neq \emptyset\}$

If $\epsilon < \epsilon'$ then $\boxed{\mathcal{C}_\epsilon(X, d) \subseteq \mathcal{C}_{\epsilon'}(X, d).}$

If the point cloud data is a sample of a compact manifold embedded in the Euclidean space then:

Theorem

There exists $\alpha > 0$ so that for any $\epsilon-$ dense sample (X, d), $\epsilon < \alpha$, the Cech complex $\mathcal{C}_\epsilon(X, d)$ is homotopy equivalent to the manifold.

VIETORIS- RIPS COMPLEX, $\mathcal{R}_\epsilon(X,d) := (\mathcal{X}, \Sigma_\epsilon)$.

- $\mathcal{X} = X$,
- $S_k := \{(x_1, x_2, \cdots x_{k+1}) |$ iff $d(x_i, x_j) < \epsilon\}$.

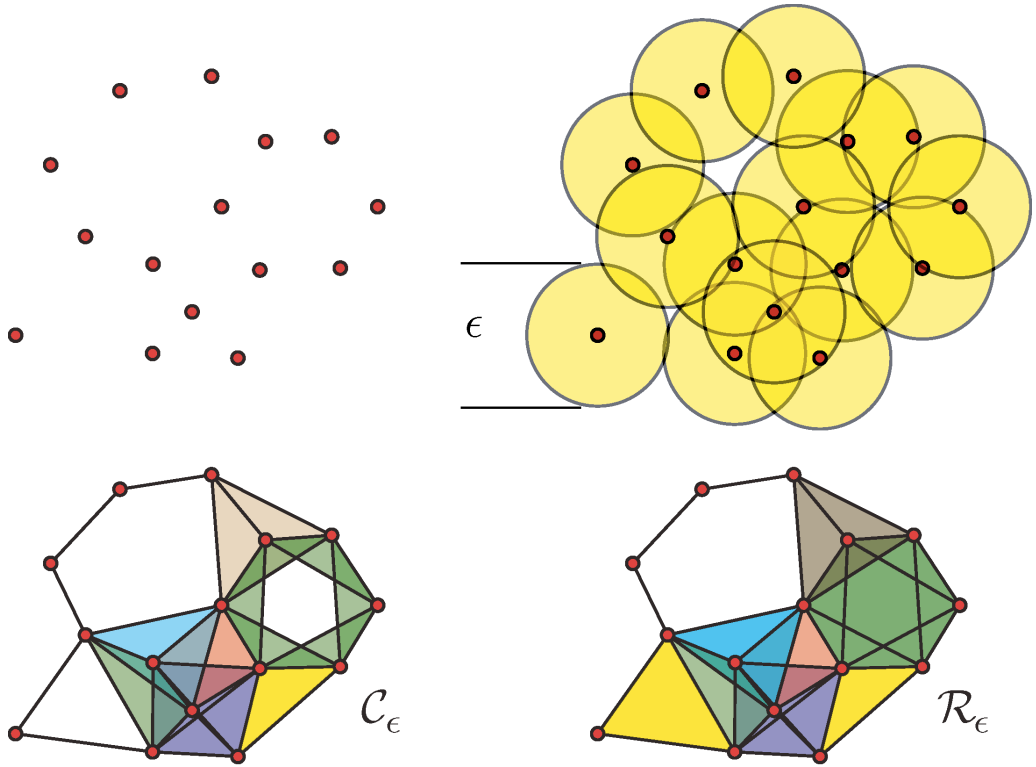
If $\epsilon < \epsilon'$ one has

$$\boxed{\mathcal{R}_\epsilon(X,d) \subseteq \mathcal{R}_{\epsilon'}(X,d).}$$

The topology of $\mathcal{C}_\epsilon(X,d)$ can be very different from $\mathcal{R}_\epsilon(X,d)$

However one has:

$$\boxed{\mathcal{R}_\epsilon(X,d) \subseteq \mathcal{C}_\epsilon(X,d) \subseteq \mathcal{R}_{2\epsilon}(X,d) \subseteq \mathcal{C}_{2\epsilon}(X,d)}$$



A fixed set of points can be completed to a Cech complex C_ε or to a Rips complex R_ε based on a proximity parameter ε. This Cech complex has the homotopy type of the $\varepsilon/2$ cover ($S1 \vee S1 \vee S1$), while the Rips complex has a different homotopy type ($S1 \vee S2$).

OBSERVATIONS:

- An element in $H_k(\mathcal{C}_\epsilon(X,d))$ which survives in $H_k(\mathcal{C}_{2\epsilon}(X,d))$ provides a nonzero element in $H_k(\mathcal{R}_{2\epsilon}(X,d))$

  An element in $H_k(\mathcal{R}_\epsilon(X,d))$ which survives in $H_k(\mathcal{R}_{2\epsilon}(X,d))$ provides a nonzero element in $H_k(\mathcal{C}_\epsilon(X,d))$

- The algorithm which inputs the point cloud data and outputs the abstract complex $\mathcal{C}_\epsilon$ is considerably more expensive than the algorithm which outputs the abstract complex $\mathcal{R}_\epsilon$.

- For ϵ very small both the Cech and Rips complexes are both disjoint unions of $\sharp(X)$ points and for ϵ large enough are $(\sharp(X)-1)-$solid simplicies.

- The topology of the ϵ complexes differ, for different ϵ's .

  It is therefore desirable to consider all these complexes together.

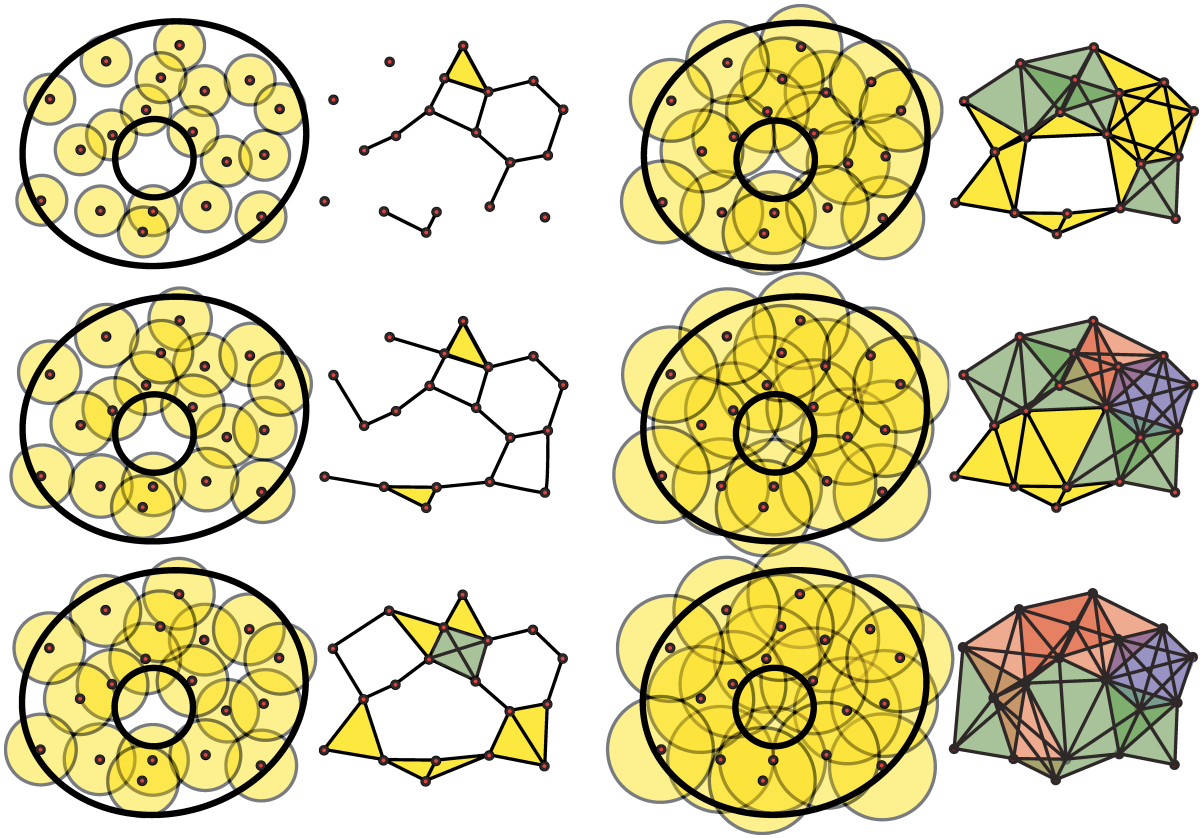
- The homology of all complexes (Chech or Rips for any ϵ)

  can be efficiently collected into

  PERSISTENT HOMOLOGY

  introduced by **Edelsbrunner, Letcher, Zomorodian** and

  algebraized by **Carlsson and Zomoradian**.



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

ALGEBRA

- A **persistent vector space** $\mathcal{V}$, is a sequence

  $\{V_n, \varphi_{n,n+1} : V_n \to V_{n+1} | n \in \mathbb{Z}_{\geq 0}\}$,with V_n vector space

  over a field κ and $\varphi_{n,n+1}$ linear maps .

  Put $\varphi_{k,k+p} := \varphi_{k+p-1,k+p} \circ \cdots \varphi_{k+1,k+2} \circ \varphi_{k,k+1}$

- The vector $x \in V_k$ is *born* in $V_p, p \leq k$ if

  in the image of $\varphi_{p,k}$ and not in the image of $\varphi_{p-1,k}$,

  and *dies* in V_r, $r \geq k$, if $\varphi_{k,r}(x) = 0$ but $\varphi_{k,r-1}(x) \neq 0$.

- Morphism and isomorphisms are obvious

A Theorem

Definition

1. A persistent vector space is tame iff each V_n has finite dimension and $\varphi_{n,n+1}$ is an isomorphisms for n large enough.

2. **Bar code** is a finite collection of intervals $\mathcal{B}(\mathcal{V}) = \{[i, \alpha],\ i \leq \alpha\}$ with $i \in \mathbb{Z}_{\geq 0},\ \ \alpha \in \mathbb{Z}_{\geq 0} \cup \infty,$

Theorem

a. A tame persistent vector space has a barcode such that :

$\dim(im(\varphi_{k,k+r}) =$ *number of intervals which contain* $[k, k+r]$.

b. Two tame persistent vector spaces are isomorphic iff the bar codes are the same.

Proof (Carlson- Zomorodian)

Based on the structure of f.g. (graded) modules over a (graded) PID (graded) ring.

κ a field , $\kappa[t]$ the polynomial ring, viewed as a graded ring with t^n of degree n.

For $\mathbb{M} = \bigoplus_{i \geq 0} M_i$ denote by $\Sigma^p \mathbb{M}$ the graded vector space with $i-$ component $= 0$ if $i \leq p-1 = M_{i-p}$ if $i \geq p$.

Regard $\mathbb{V} := \bigoplus_{k \geq 0} V_k$ as graded vector space. $\mathbb{V}$ is a f.g. $\kappa[t]-$ graded module with $tx = \varphi_{k,k+1}(x)$ if $x \in V_k$.

$$\mathbb{V} = \bigoplus_{1 \leq i \leq s} \Sigma^{r_i} \kappa[t] \oplus \bigoplus_{1 \leq j \leq t} \Sigma^{m_j} (\kappa[t]/t^{n_j})$$

The intervals $[r_i, \infty), 1 \leq i \leq s$ and $[m_j, m_j + n_j]$ define the barcode $\mathcal{B}(\mathcal{V})$ of $\mathcal{V}$.

TOPOLOGY

Definition

1. A tame persistent space (complex) is a filtered space (complex) $\mathcal{K} := \{K_0 \subset K_1 \subset \cdots K_i \subset K_{i+1} \subset \cdots\}$ s.t.: $K_i \subset K_{i+1}$ homotopy equivalence for i large enough and $H_k(K_i)$ a finite dimensional vector space for any i, k.
2. For any k one associates the persistent vector space $\mathcal{H}_k(\mathcal{K})$ whose components are $H_k(K_i)$ and liner maps induced by inclusions. The barcodes $\mathcal{B}(\mathcal{H}_k(\mathcal{K}))$ are called the barcode of $\mathcal{K}$ and denorted by $\mathcal{B}(\mathcal{K})$
3. The persistent homology of the persistent space $\mathcal{K}$ is the collection of vector spaces

$$H_k^{i,j} := \mathrm{im}(H_k(K_i) \to H_k(K_j)),\ i \leq j.$$

Source 1.

A function $f : M \to \mathbb{R}$, M a compact ANR (cell complex)

is called tame if the homology of the excursion

sets $f^{-1}((\infty, c])$ changes only for a finite number of values

$c_0 < c_1 < c_2 \cdots c_N$.

Then

$$\mathcal{K} = \{K_0 \subset K_1 \subset \cdots \subseteq K_r \subseteq\},$$

$K_i = f^{-1}((-\infty, c_i])$, is a tame persistent space.

NOTE: A Morse function on a closed manifold is a tame function.

Source 2.

If (X, d) is a point cloud data there exists only finitely many $(0 = \epsilon_0 < \epsilon_1 < \cdots \epsilon_N)$ so that the Cech or Rips complexes change the homotopy type. By taking the ϵ_i-Cech or Rips complex as K_i, one obtains a tame persistent complex . We write $\mathcal{B}(X, d) := \mathcal{B}(\mathcal{K}(\mathcal{R}(X, d)))$.

NOTE: Source 2. can be viewed as a particular case of Source 1, where the space $X = K_0 \times [0, 1] \cup K_1 \times [1, 2] \cup \cdots$ and the function is the projection on the second component.

OBSERVATIONS

- For a POINT CLOUD data (X, d) there are reasonably effective algorithms to calculate the Rips complex $\mathcal{R}_\epsilon(X, d)$ and effective algorithms to directly calculate the bar code of $\mathcal{B}(X, d)$.
- In a bar code the long intervals are significant qualitative features observable at various level of resolution, while the small intervals indicate "Noise".
- The barcode based on Rips complex and Cech complex are almost the same.

GENERAL QUESTIONS

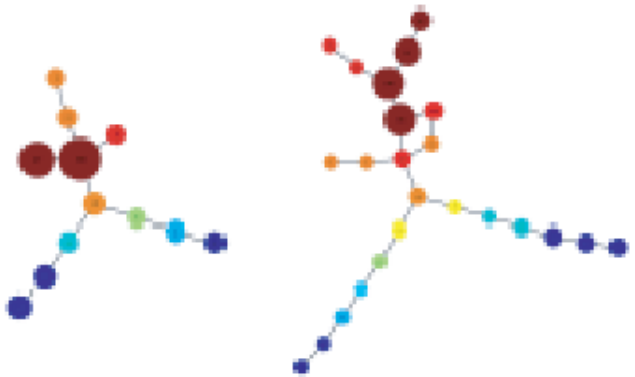
- What does it mean for two point cloud data to have the same, almost the same, close enough bar codes?
- What features (of a point cloud data) can be recovered from bar codes?

Some answers:

- (Example 1.) Presence of long intervals in the bar code corresponding to $k = 2$ sugests presence of tumors. Lack of barcodes corresponding to $k = 2$ rules out tumors and presence of long intervals in the bar code corresponding to $k = 1$ indicates different abnormalities (inflated blood vessels) .
- (Example 2) There were essential three relevant parameters (so the point clout data can be recognized as a potential three dimensional space and there were two type of diabetes (Type 1 and type 2) as long as the sick patients were concerned.

The recent paper of G.Carlsson (Topology and data, BAMS 2009) provides additional examples in various other fields.

Here is a pictorial representation of the diabetes study .



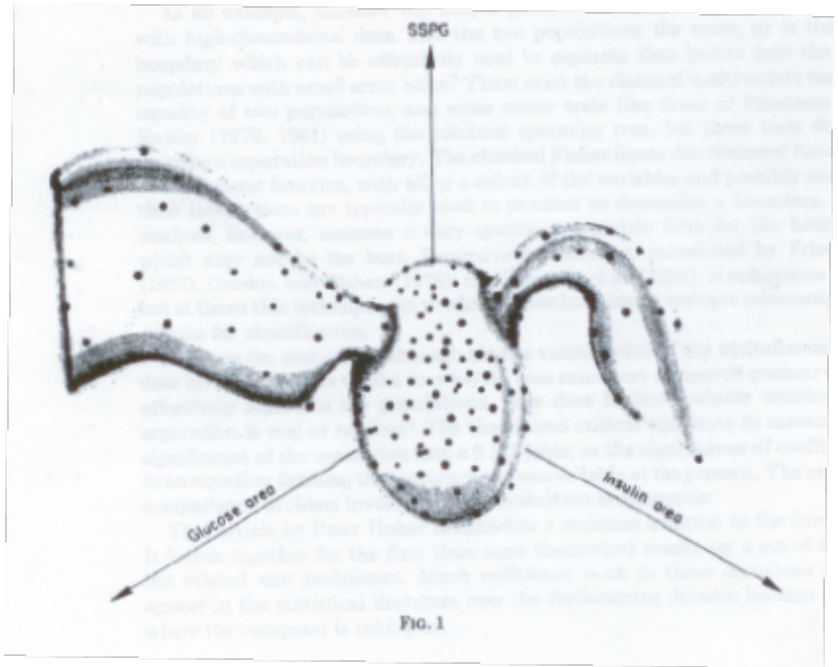


Fig. 1