



institut**Curie**

# Elastic maps with applications in bioinformatics

**Alexander Gorban**

**University Of Leicester, UK**

**Andrei Zinovyev**

**Institute of Curie, France**

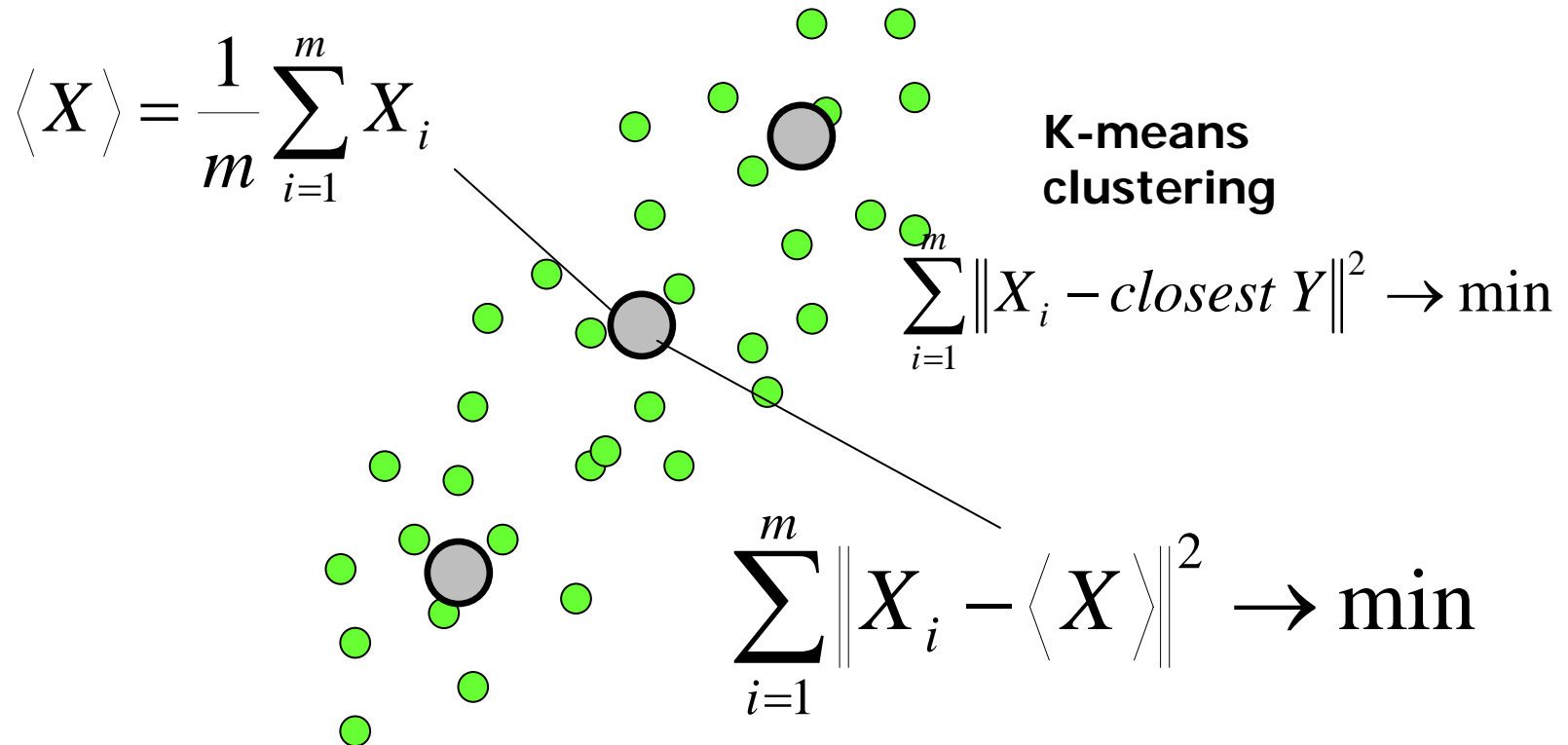
# Outline of the talk

---

- **Principal manifolds as surfaces of minimal elastic energy**
- **Elastic maps: construction and utilization**
- **Examples of use**
- **Microarray datasets**
- **Future development**

# Mean point

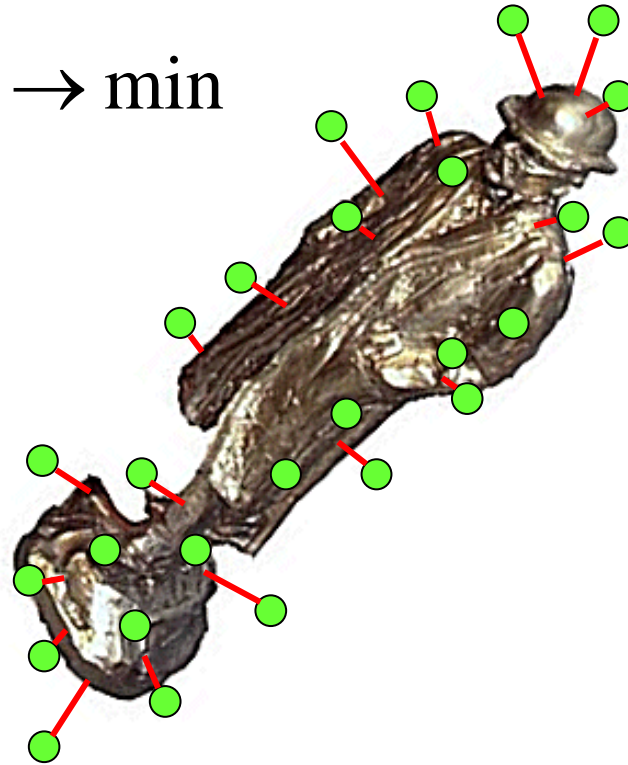
---



# Principal "Object"

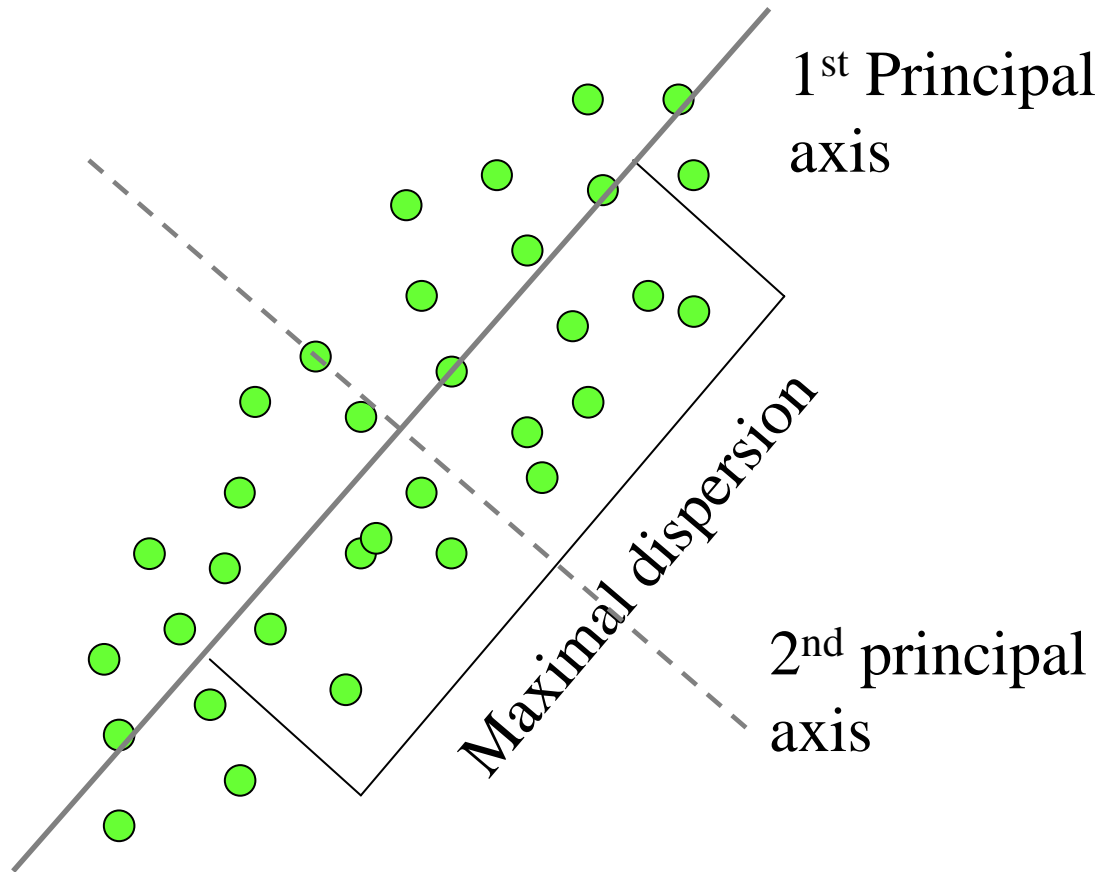
---

$$\sum_{i=1}^m \left\| \text{---} \right\|^2 \rightarrow \min$$



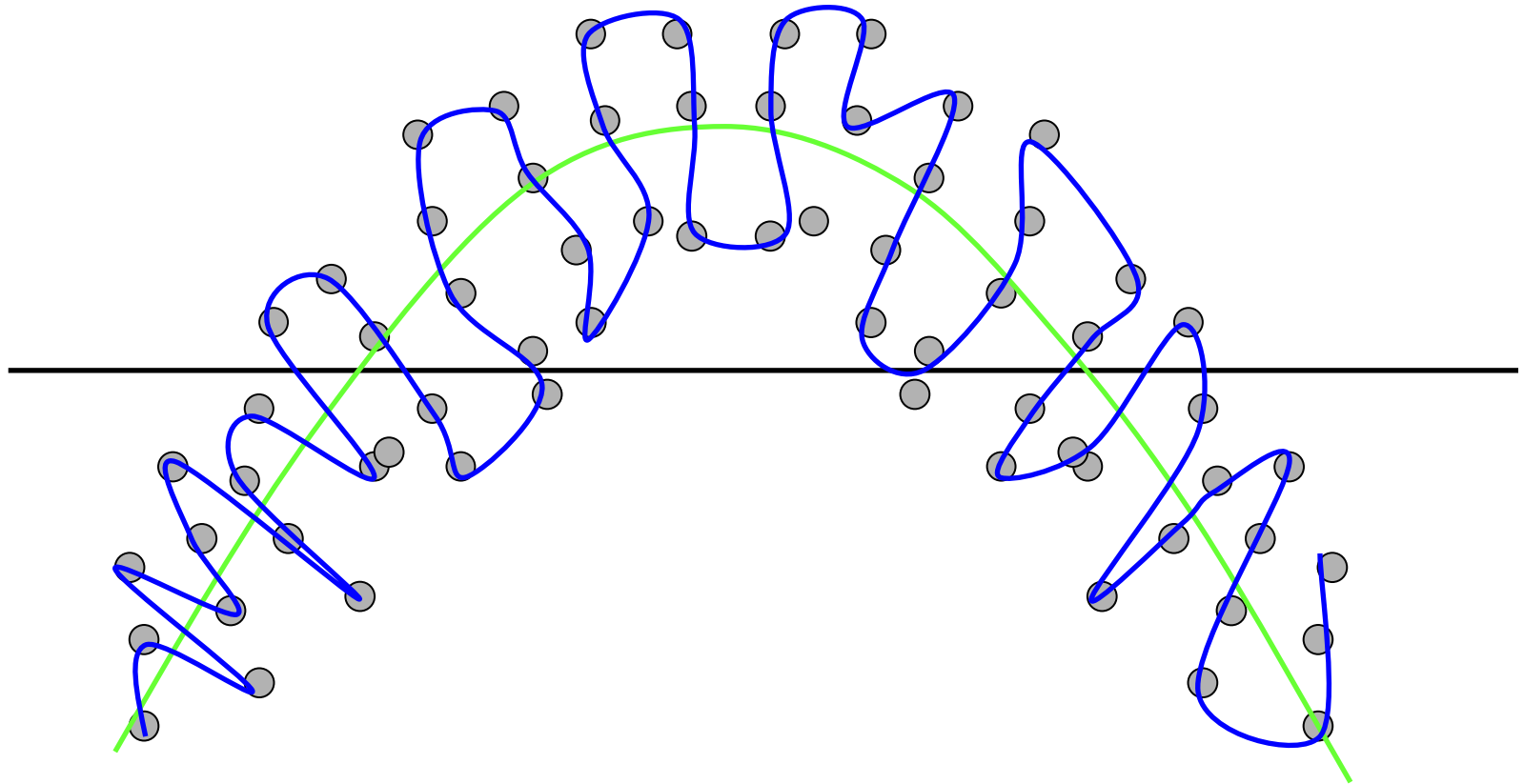
# Principal Component Analysis

---



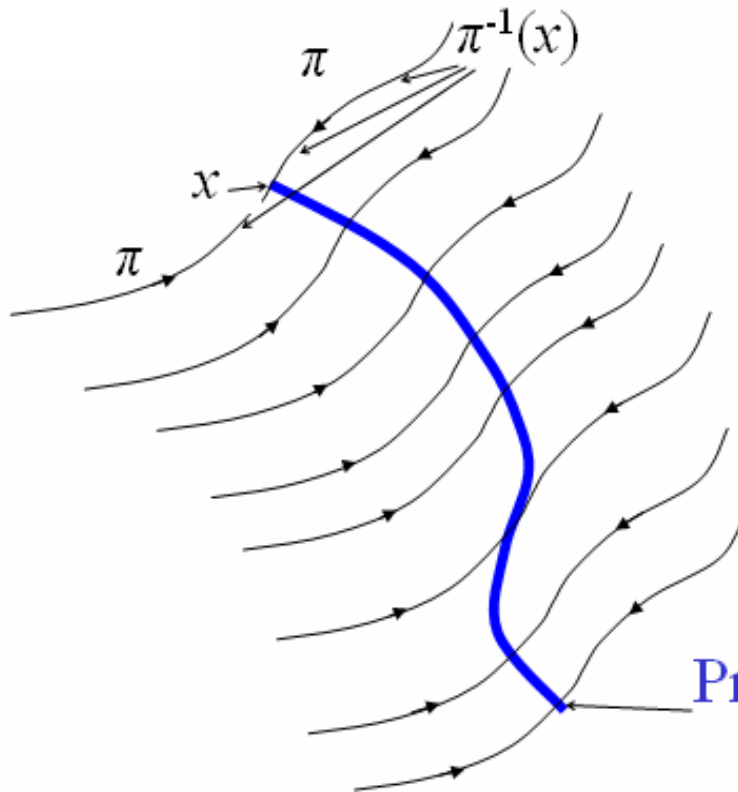
# Principal manifold

---



# Probability distribution: idea of self-consistency

---

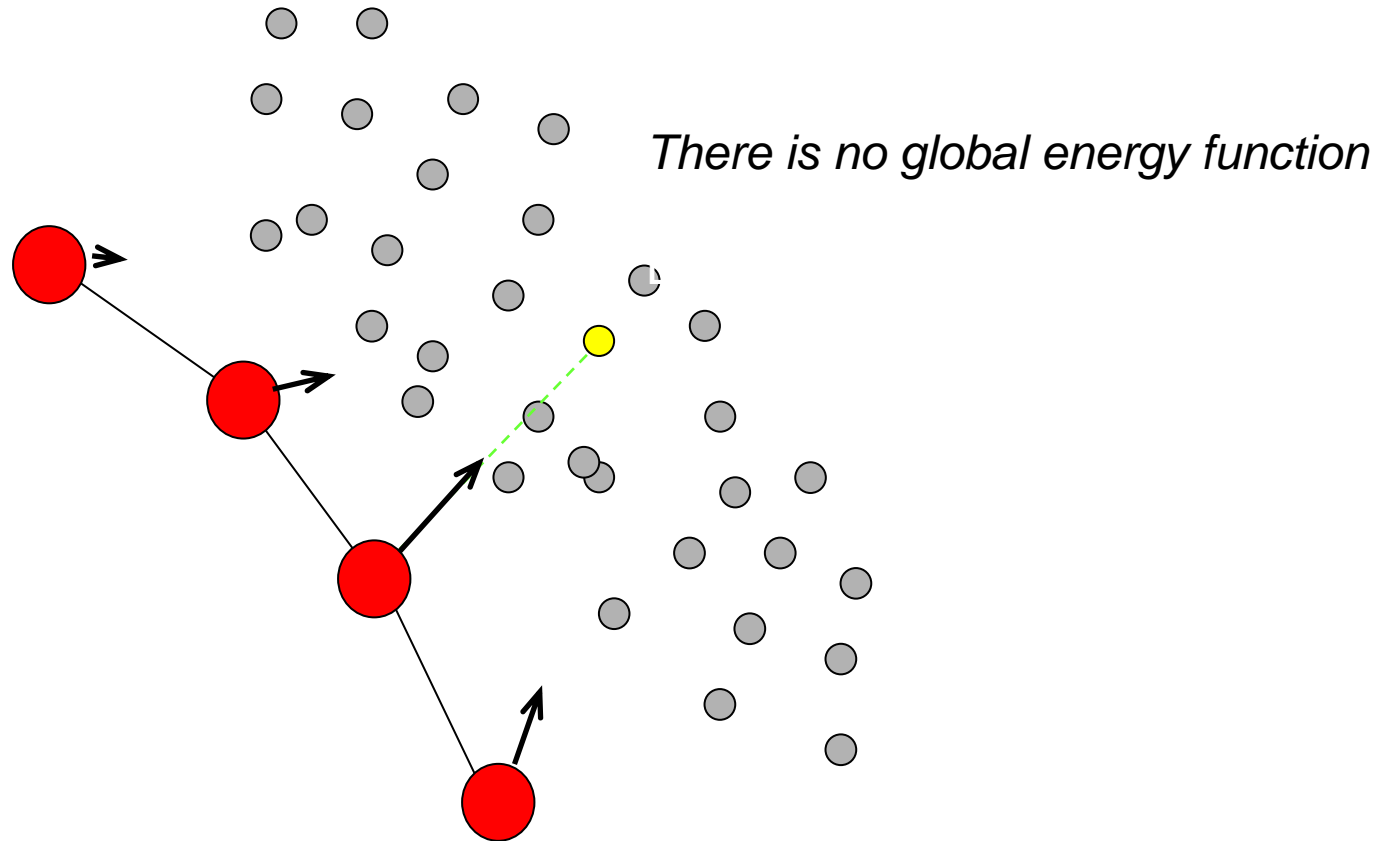


$$x = \mathbf{E}(y | \pi(y) = x)$$

Principal Manifold

# Self-organizing maps

---

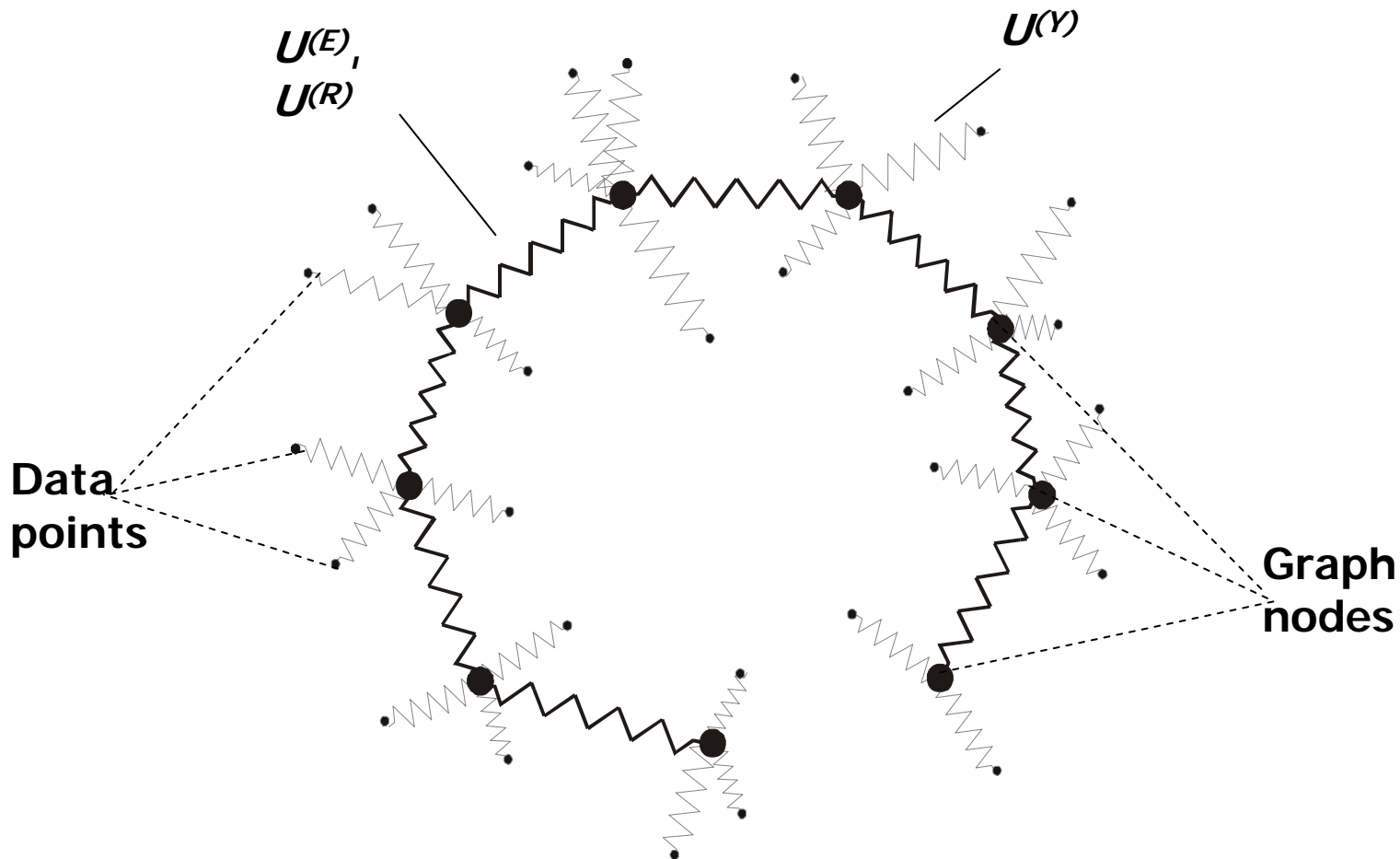




# Metaphor of elasticity

(energy function proposed by Gorban in 1996  
at Russian national neuroinformatics workshop  
“Neuroinformatics and its applications - 1996”)

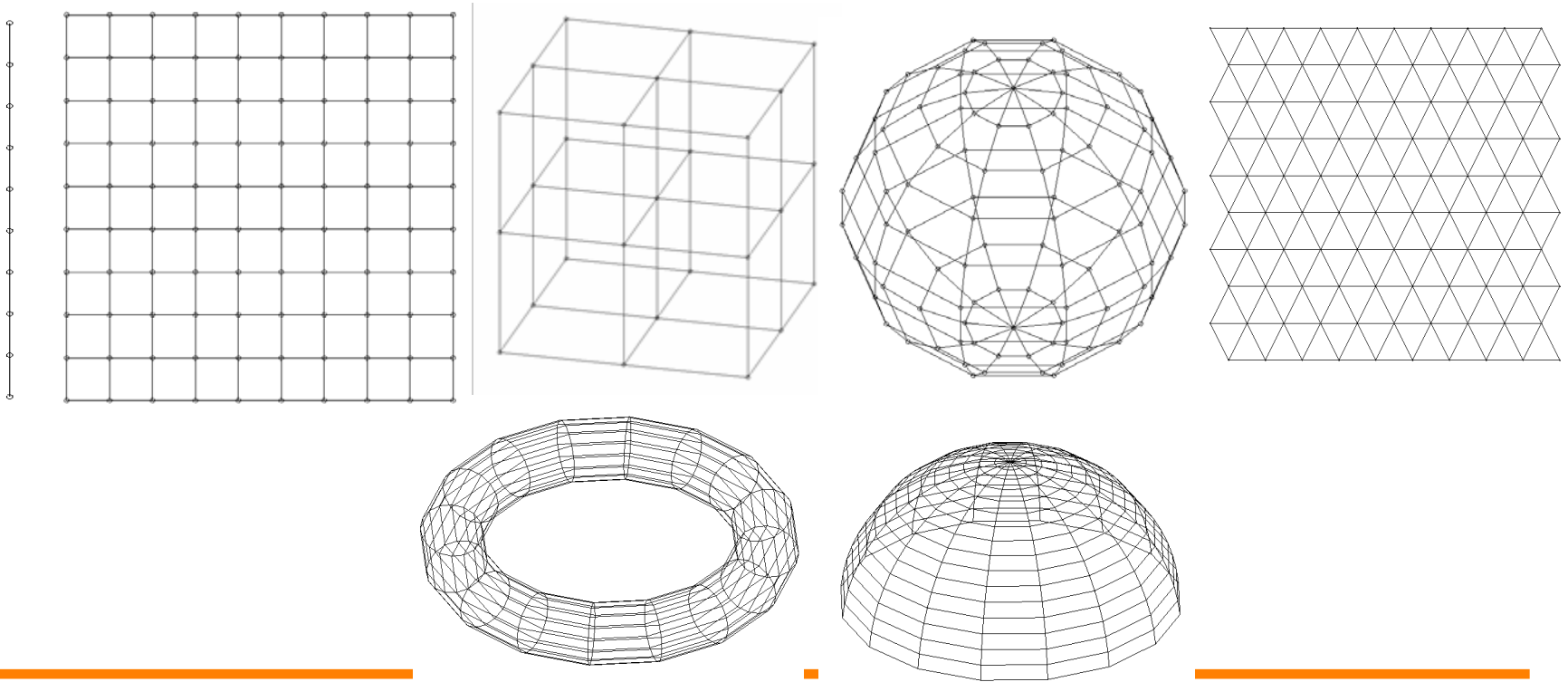
---



# Constructing elastic nets

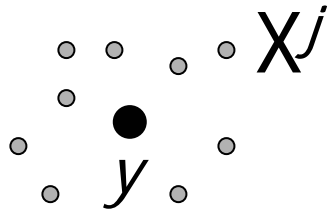
---

●  $\mathcal{Y}$     ●—●  $E(0)$   $E(1)$     ●—●—●  $R(1)$   $R(0)$   $R(2)$

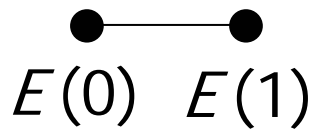


# Definition of elastic energy

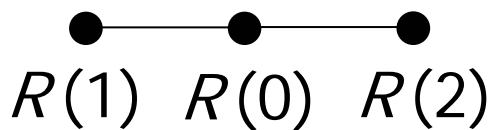
---



$$U^{(Y)} = \frac{1}{N} \sum_{i=1}^p \sum_{x^{(j)} \in K^{(i)}} \|X^j - y^{(i)}\|^2$$



$$U^{(E)} = \sum_{i=1}^s \lambda_i \|E^{(i)}(1) - E^{(i)}(0)\|^2$$



$$U^{(R)} = \sum_{i=1}^r \mu_i \|R^{(i)}(1) + R^{(i)}(2) - 2R^{(i)}(0)\|^2$$

$$U = U^{(Y)} + U^{(E)} + U^{(R)} \quad \lambda_i = \lambda_0, \quad \mu_i = \mu_0$$


---

## Scaling rules

---

For uniform d-dimensional net from the condition of constant energy density we obtain:

$$\lambda_1 = \lambda_2 = \dots = \lambda_s = \lambda(s);$$

$$\mu_1 = \mu_2 = \dots = \mu_r = \mu(r)$$

$$\lambda = \lambda_0 s^{\frac{2-d}{d}}$$

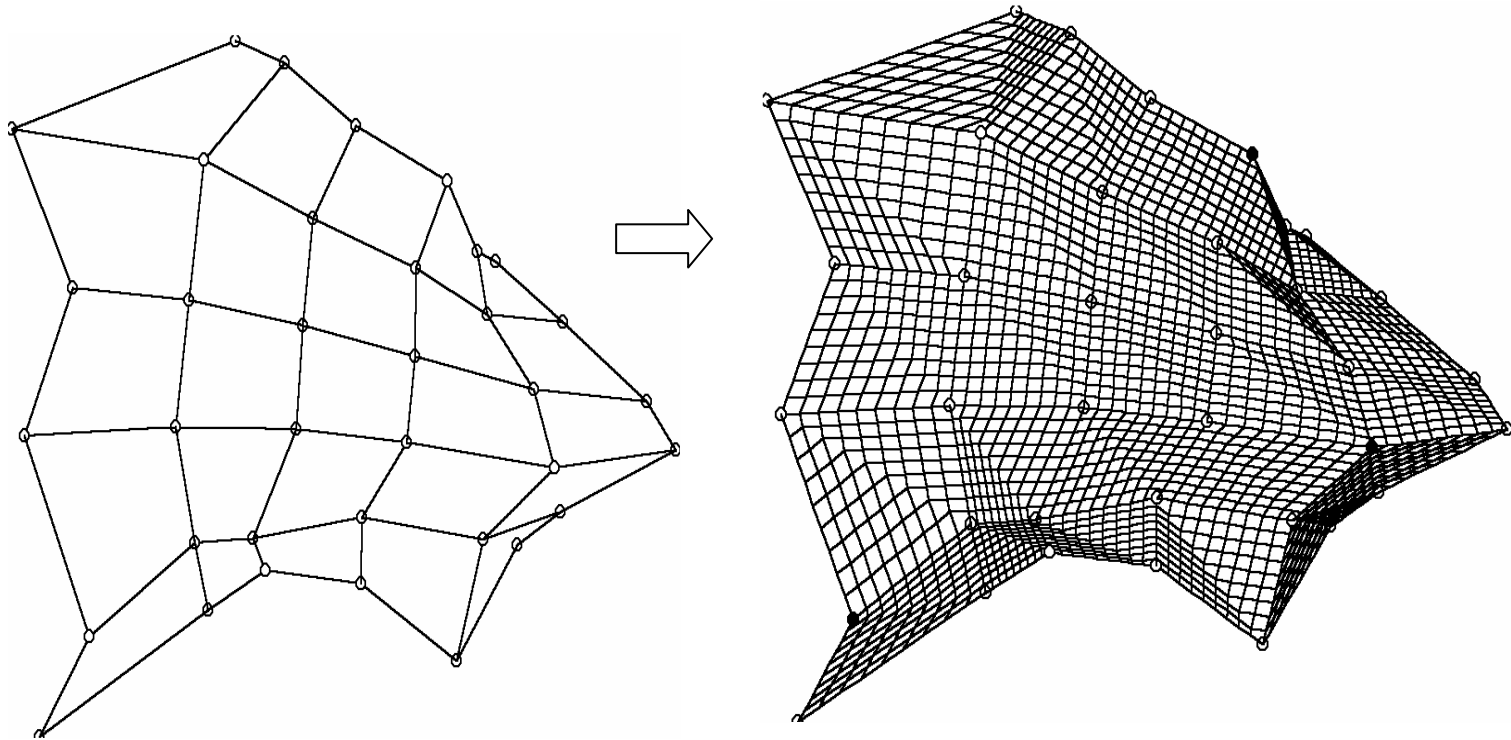
$s$  is number of edges,  
 $r$  is number of ribs  
in a given volume

$$\mu = \mu_0 r^{\frac{4-d}{d}}$$

---

# Elastic manifolds

---



# Global minimum and softening

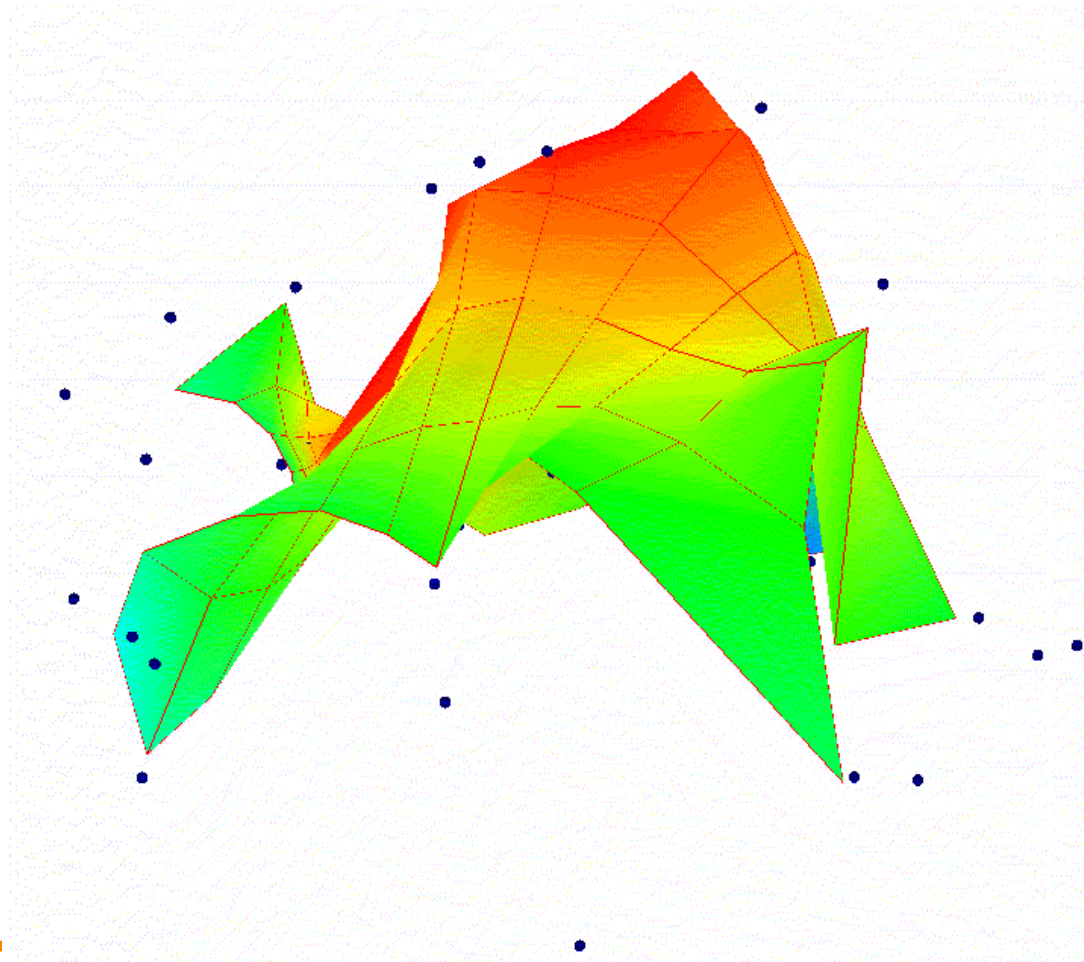
---

$$\lambda_0, \mu_0 \approx 10^3$$

$$\lambda_0, \mu_0 \approx 10^2$$

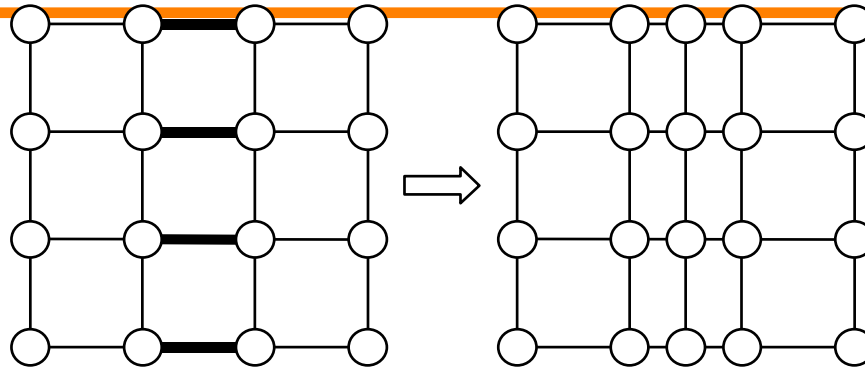
$$\lambda_0, \mu_0 \approx 10^1$$

$$\lambda_0, \mu_0 \approx 10^{-1}$$

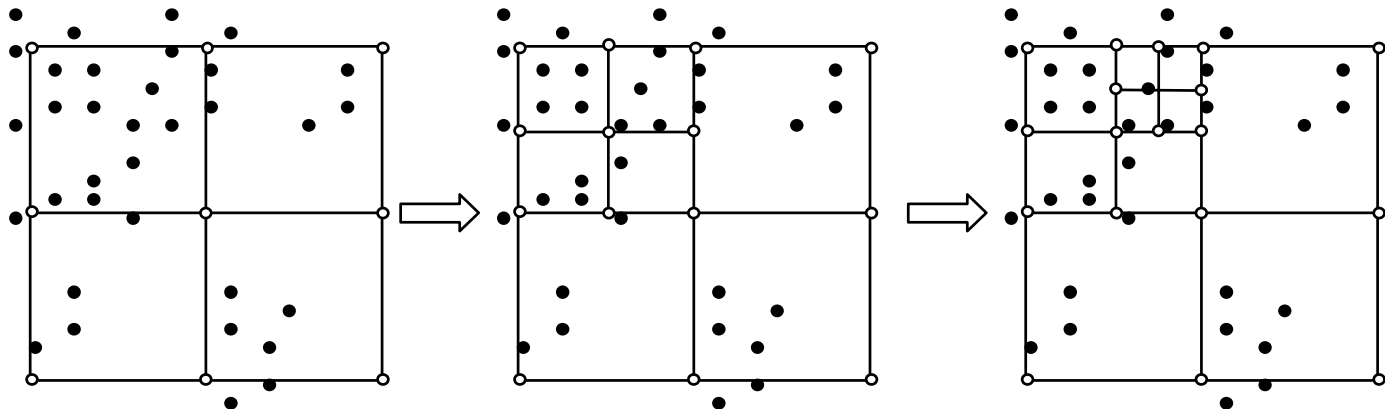


# Adaptive algorithms

Refining net:

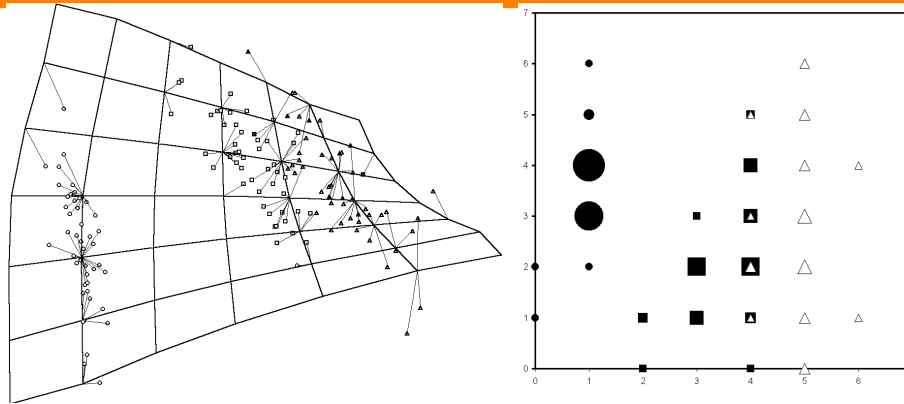


Growing net

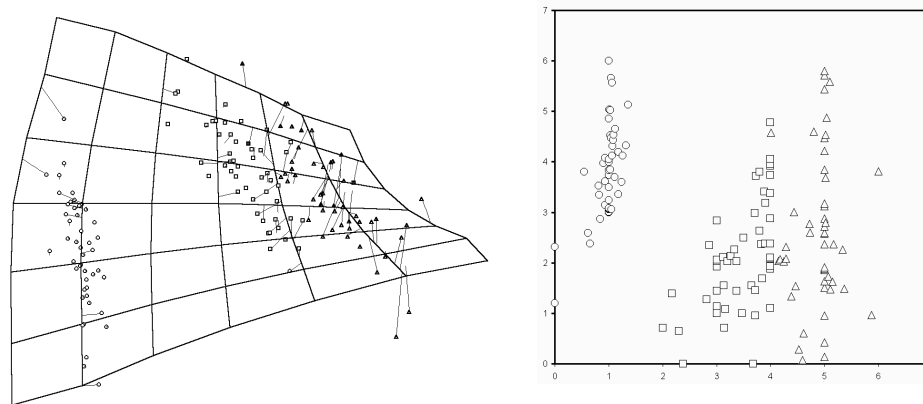


Adaptive net

# Projection onto the manifold



**Closest node of the net**

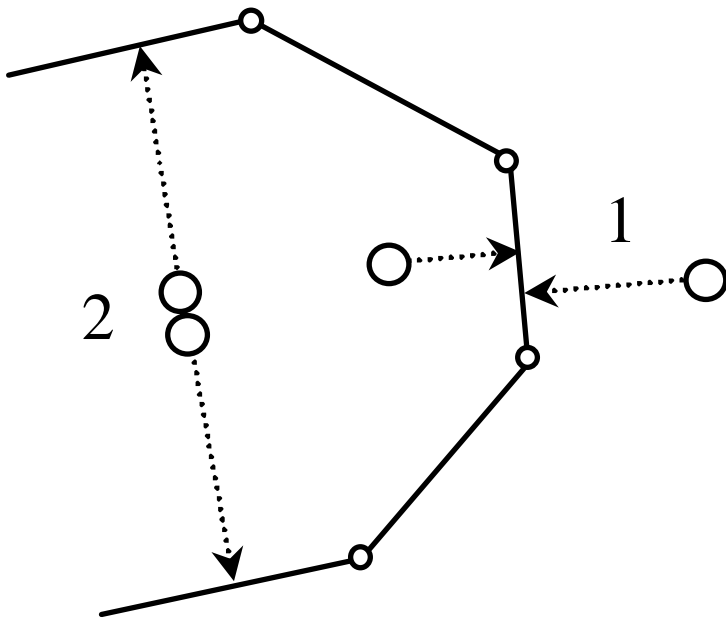


**Closest point of the manifold**



# Mapping distortions

---



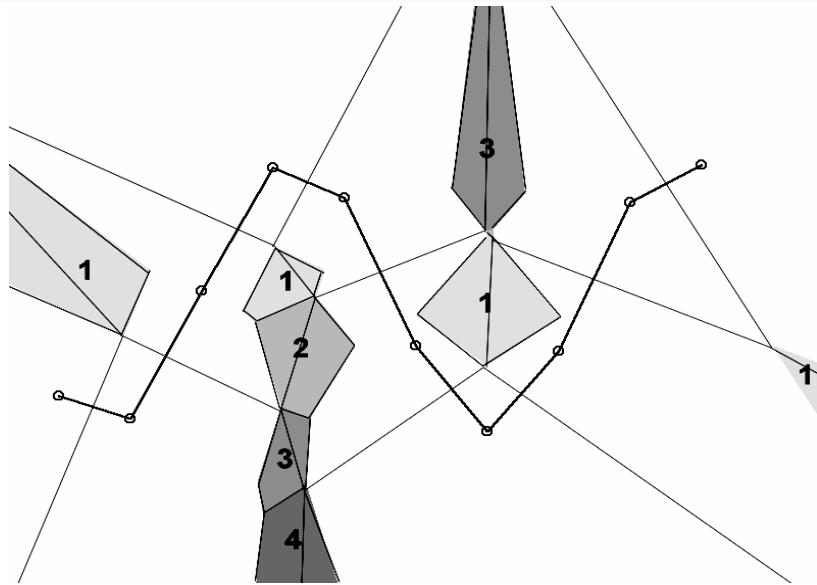
Two basic types of distortion:

1) Projecting distant points in the close ones (**bad resolution**)

2) Projecting close points in the distant ones (**bad topology compliance**)

# Instability of projection

---



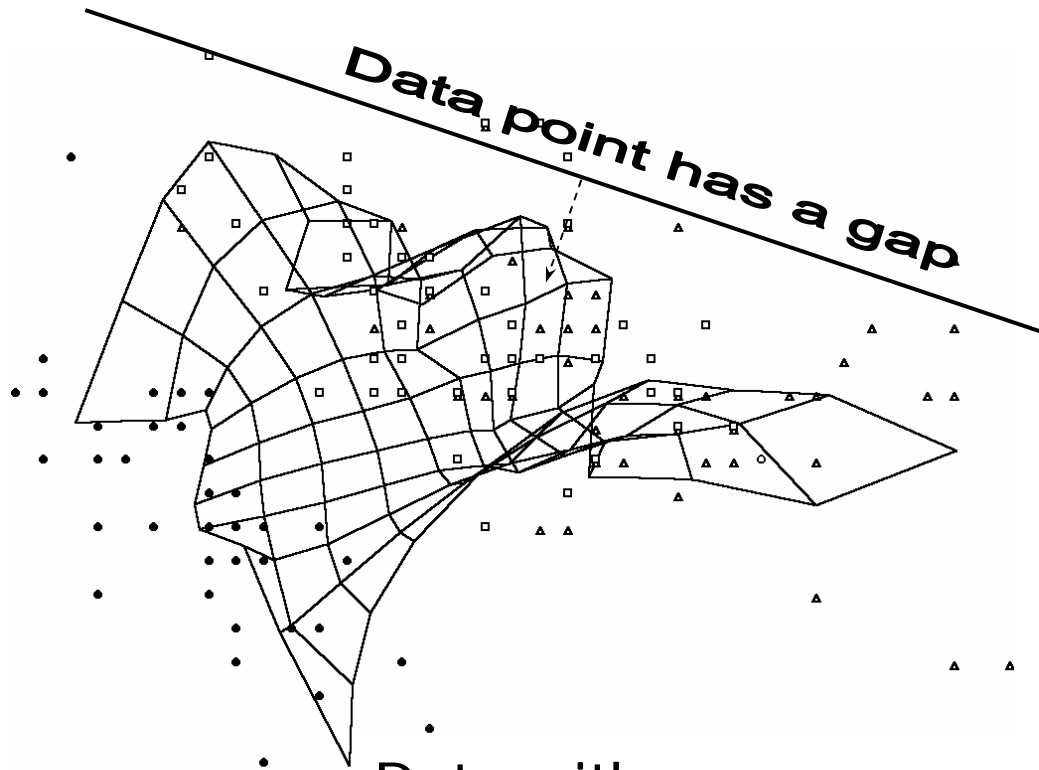
Best Matching Unit (BMU) for a data point is the closest node of the graph, BMU2 is the second-close node. If BMU and BMU2 are not adjacent on the graph, then the data point is *unstable*.

Gray polygons are the areas of instability. Numbers denote the degree of instability, how many nodes separate BMU from BMU2.

---

# Dealing with missing values in data

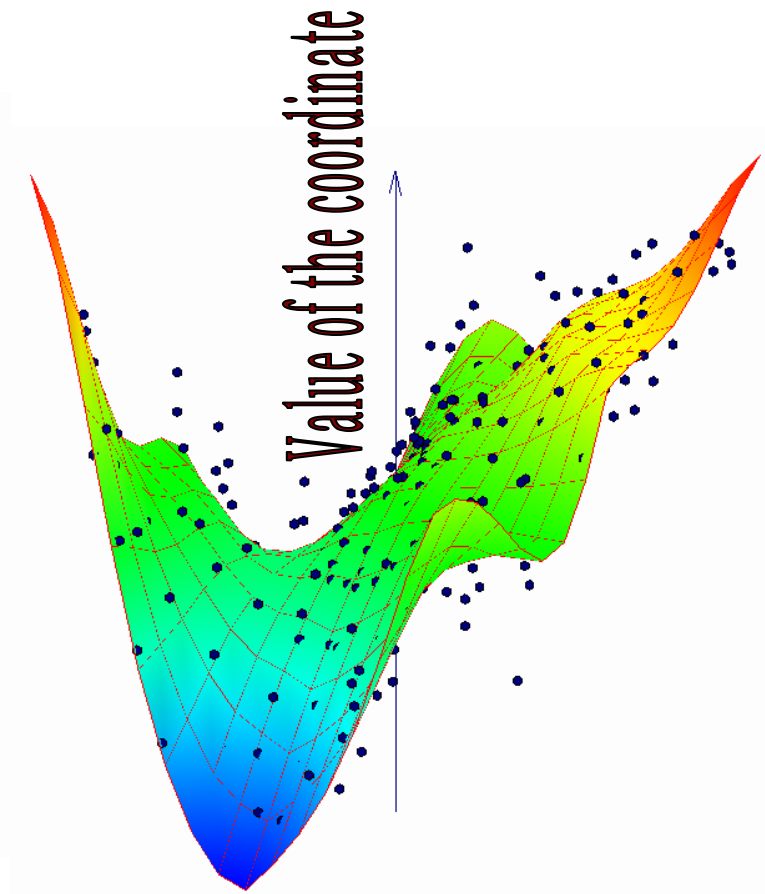
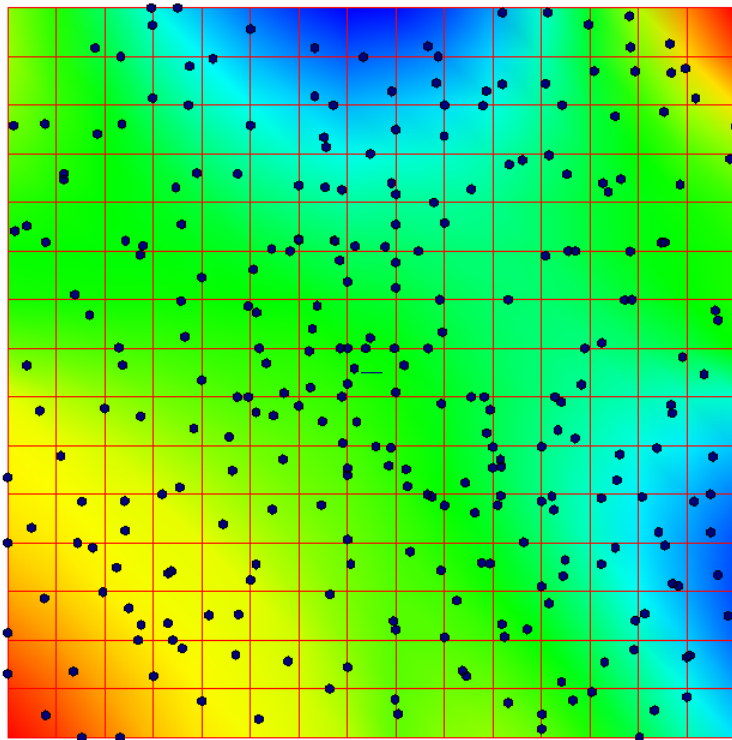
---



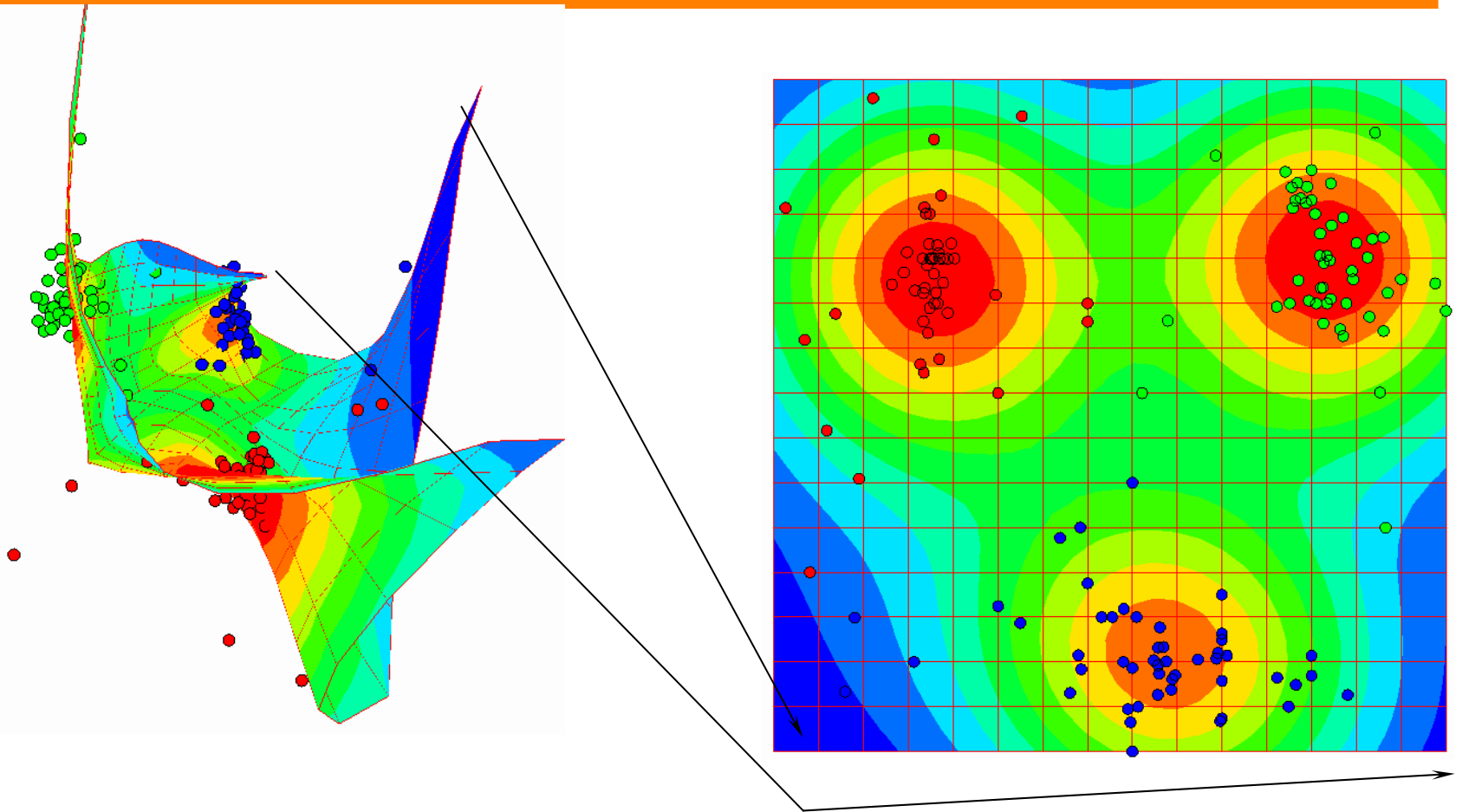
- Data with gaps are modelled as affine manifolds, the nearest point on the manifold provides the optimal filling of gaps.
-

# Colorings: visualize any function

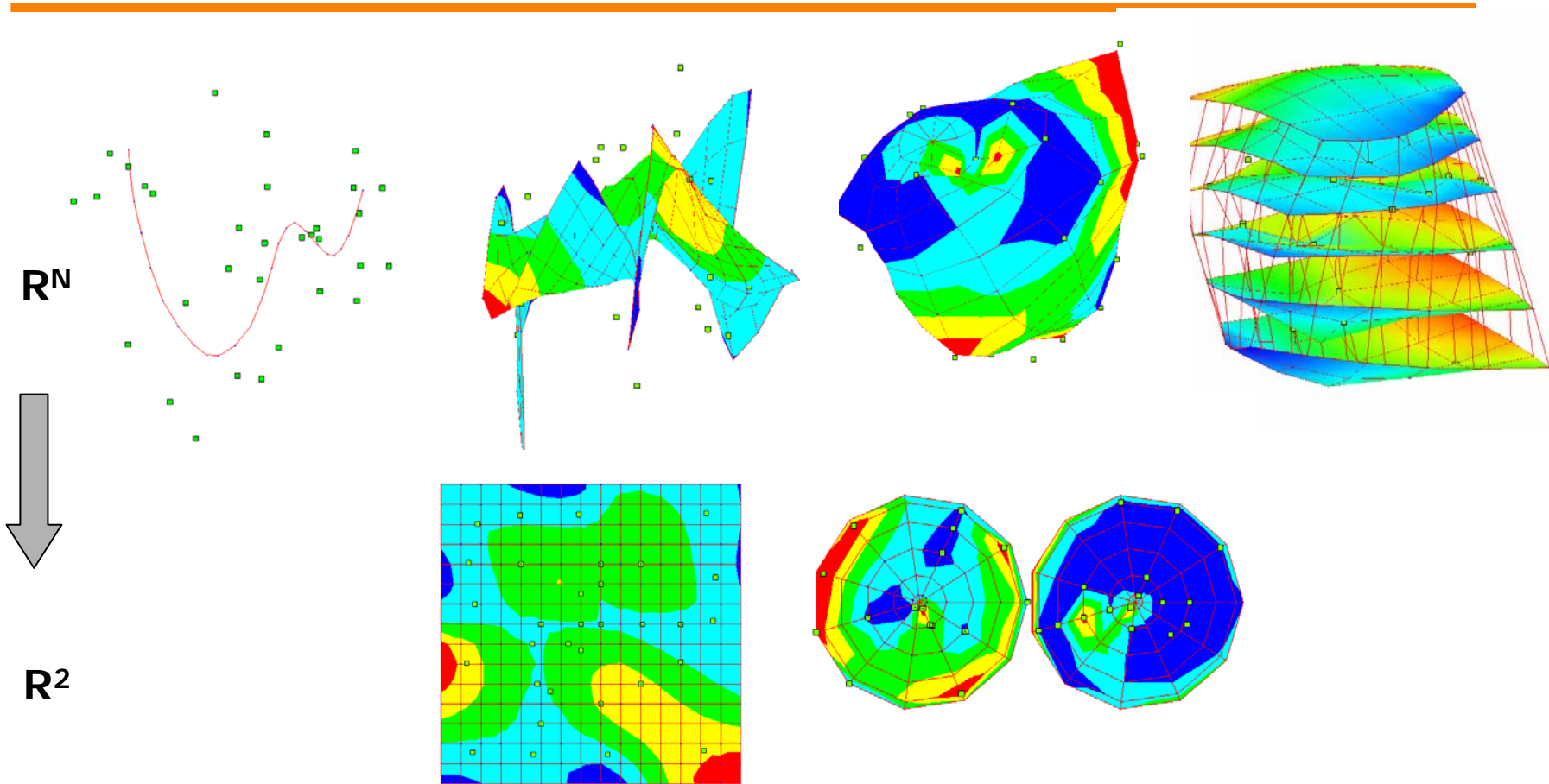
---



# Density visualization

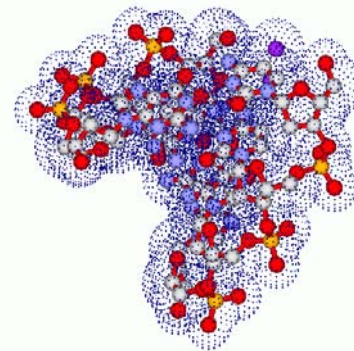


# Various manifold topologies

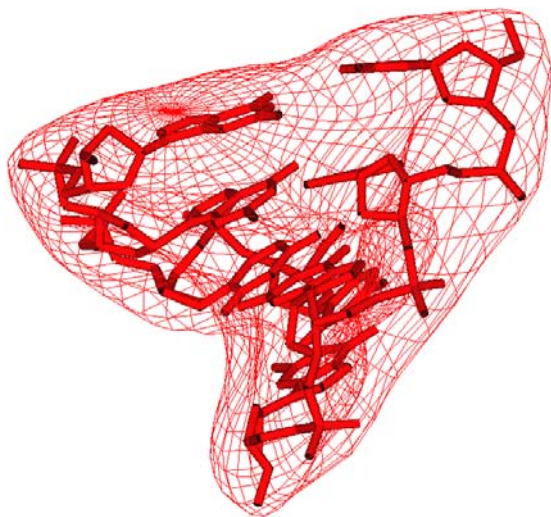


# Example 1: Approximating molecular surface

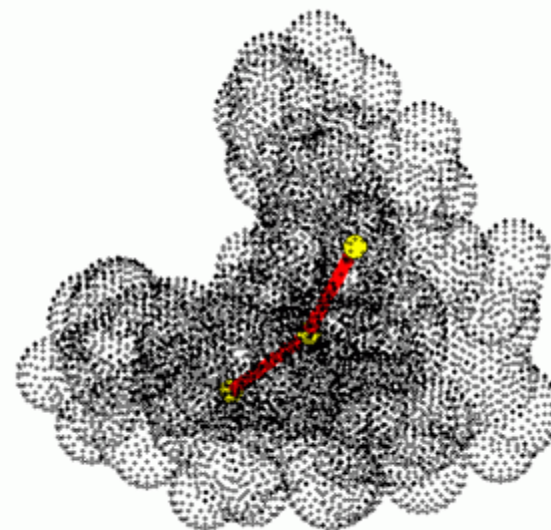
---



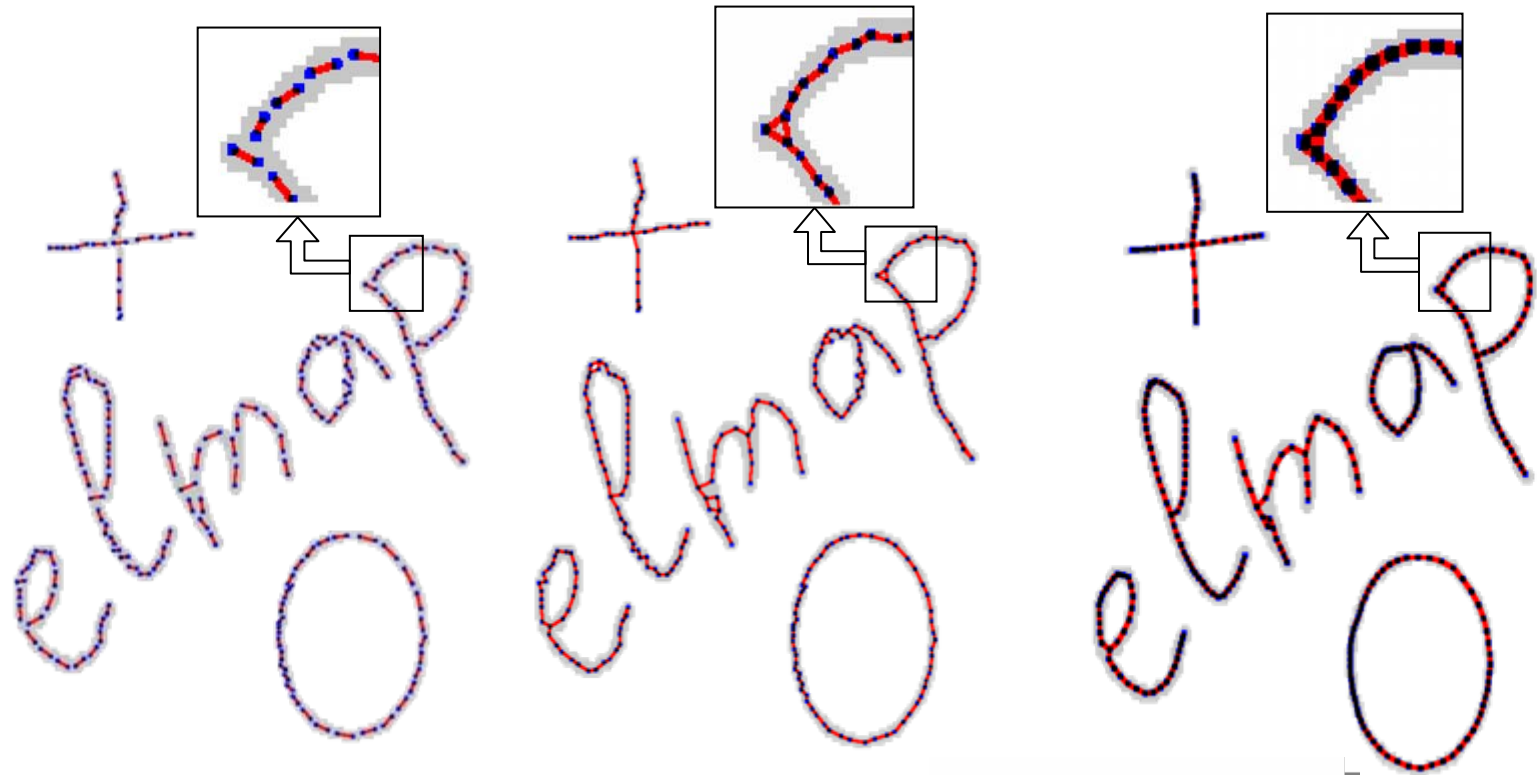
Approximating by 2D spherical grid



Approximating by 1D curve



## Example 2: Image skeletonization or clustering around curves



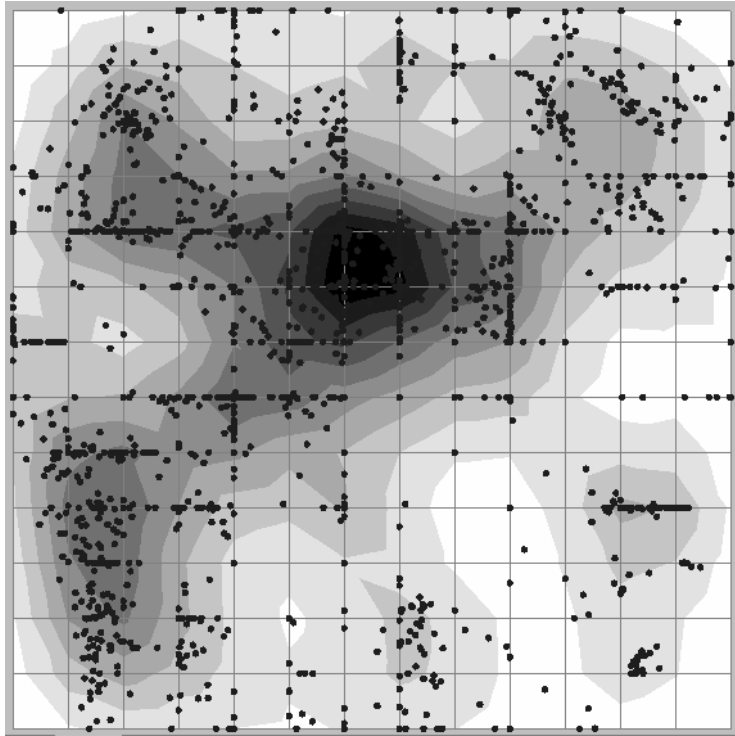
+  
elmap  
O



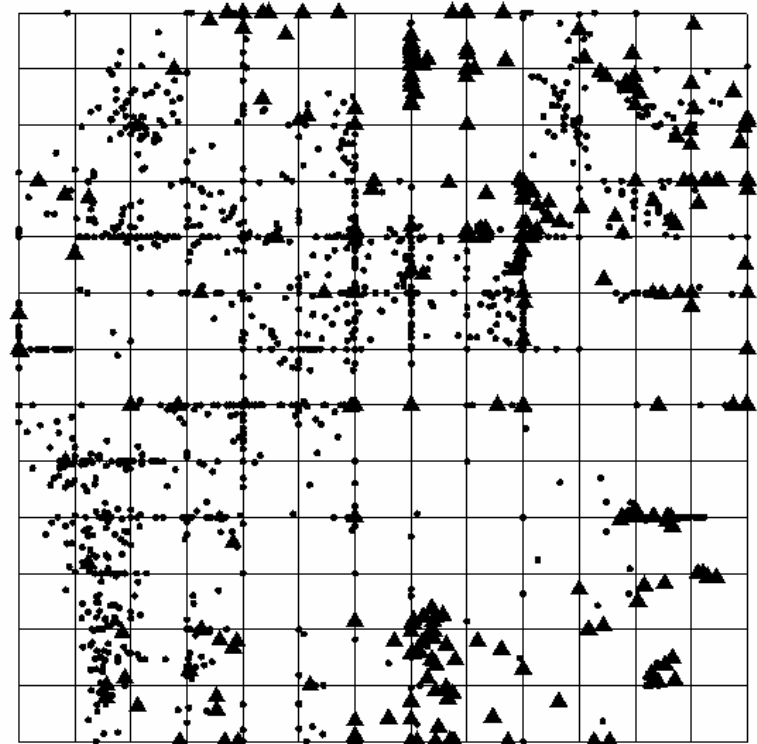
# Example 3: Medical table

1700 patients with infarctus myocarde

---



Patients map, density

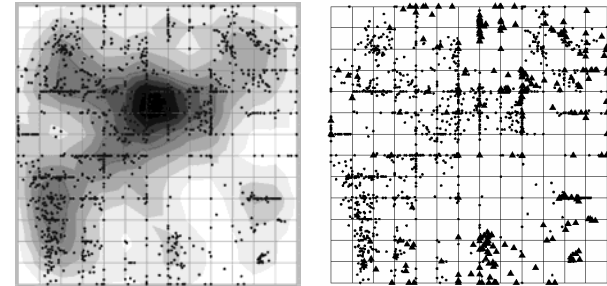


Lethal cases

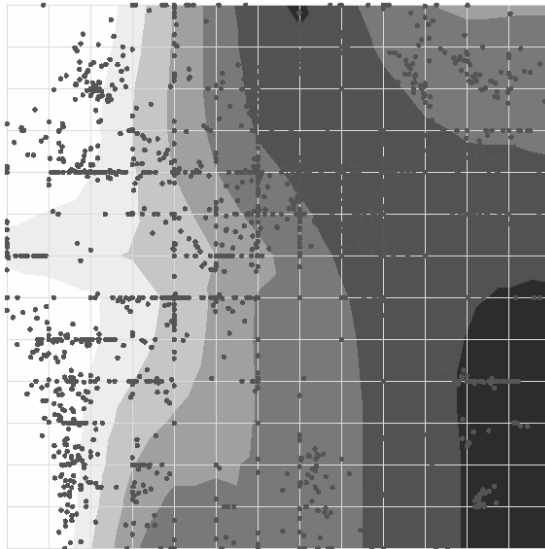
---

# Example 3: Medical table

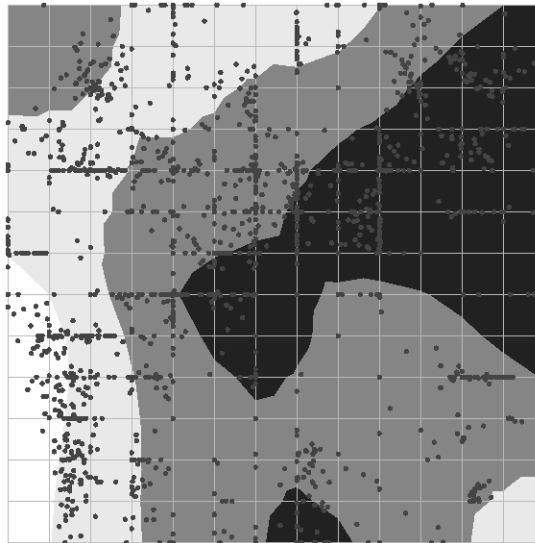
1700 patients with infarctus myocarde



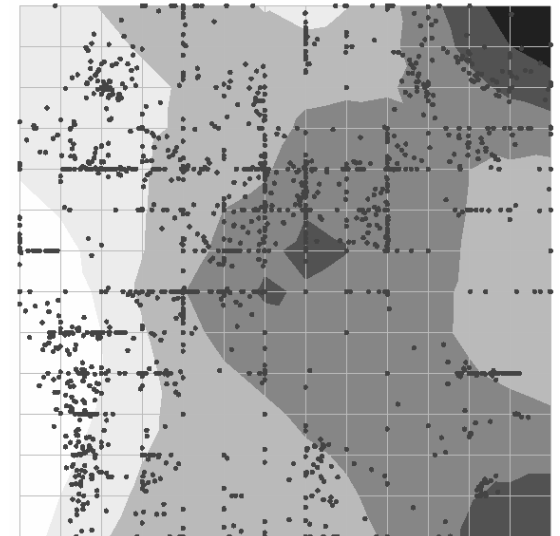
## 128 clinical variables



Age

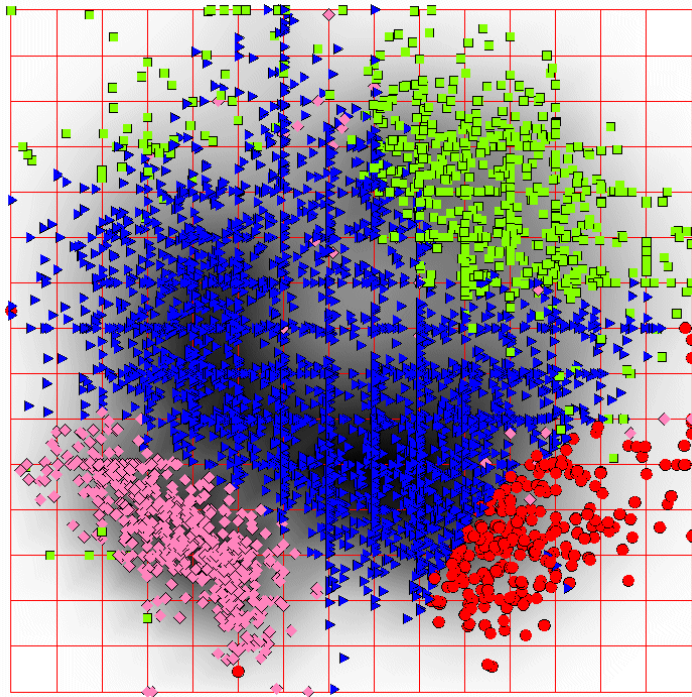


Number of infarctus  
in anamnesis

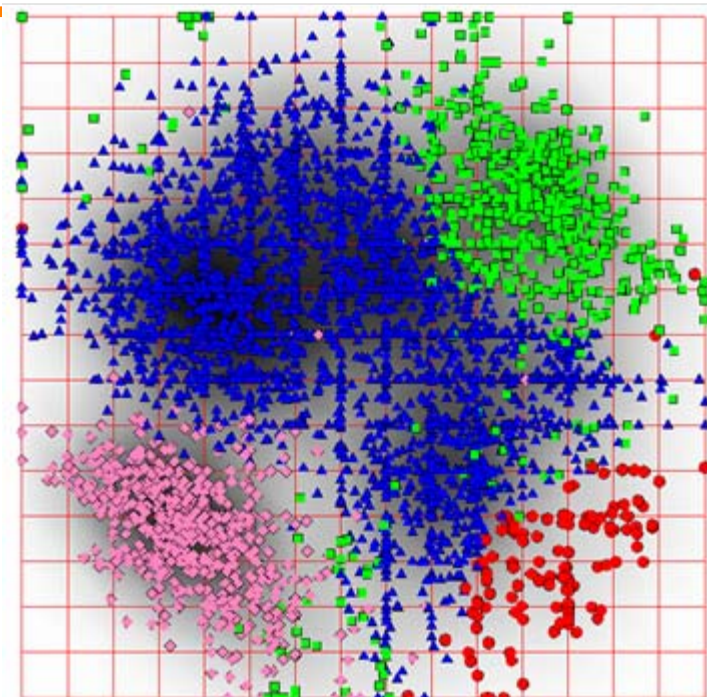


Stenocardia functional  
class

## Example 4: Codon usage in all genes of one genome



*Escherichia coli*



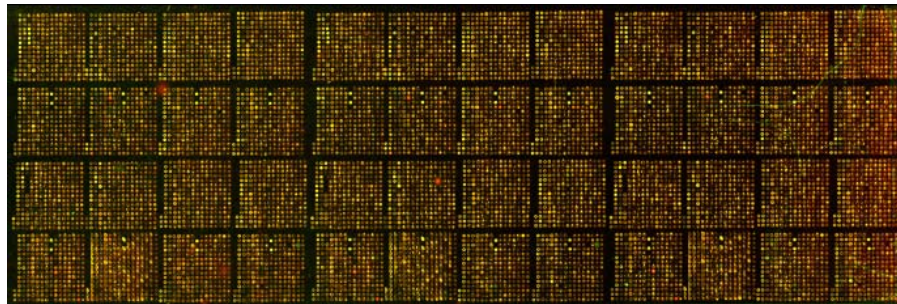
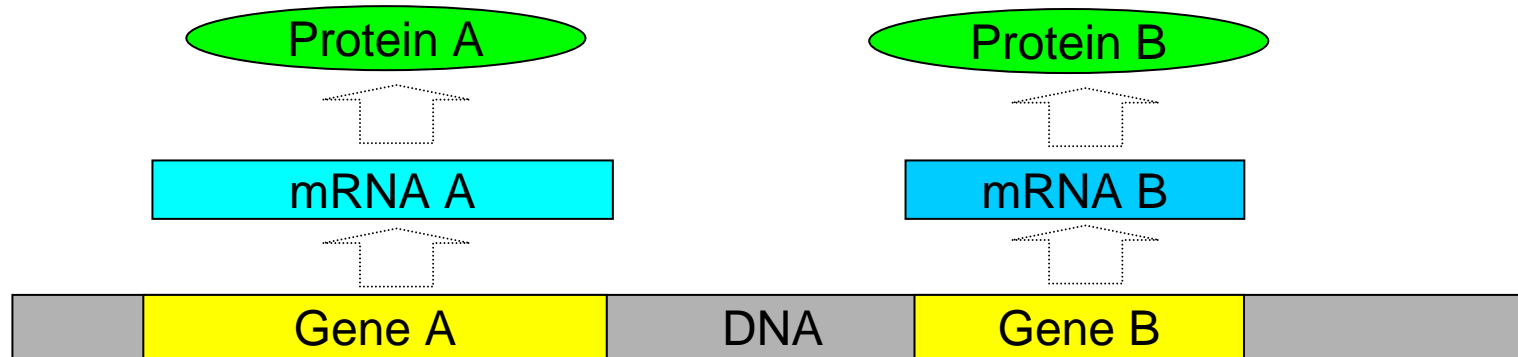
*Bacillus subtilis*

- Majority of genes
- Highly expressed genes

- "Foreign" genes
- "Hydrophobic" genes

# Microarray technology and microarray datasets

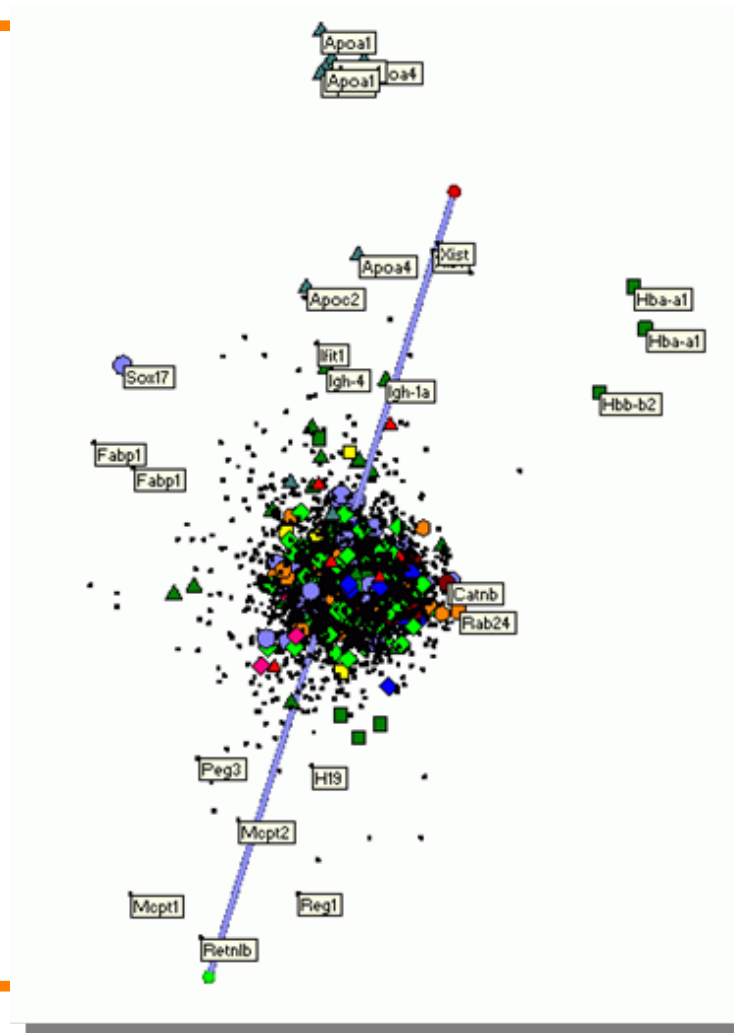
---



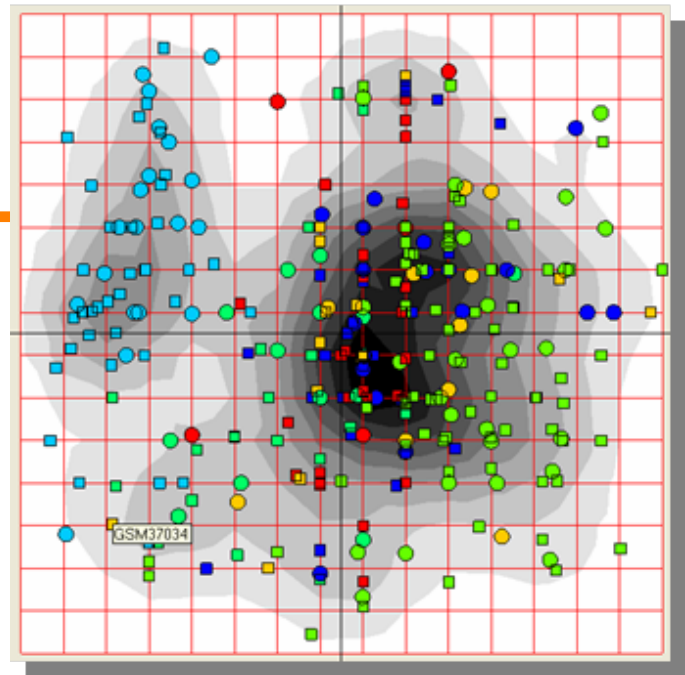
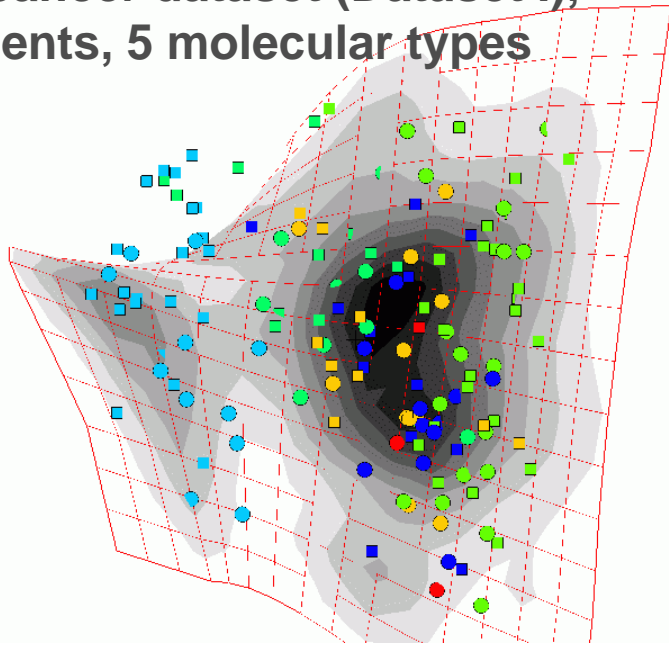
One spot corresponds to a gene (mRNA concentration)

Table of numbers, characteristic size is 10000 genes x100 samples

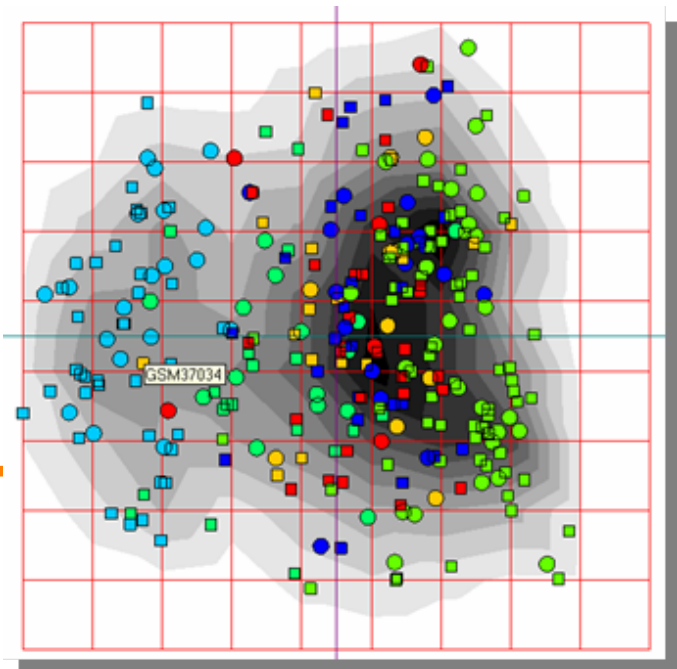
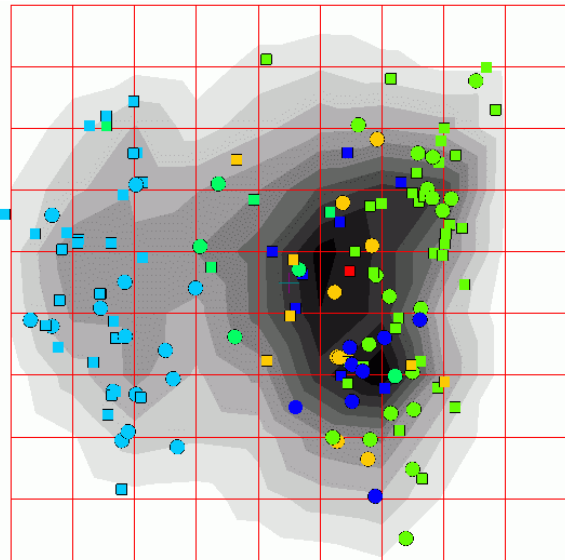
Gene space:  
every point correspond to a gene characterized by its expression  
in m samples



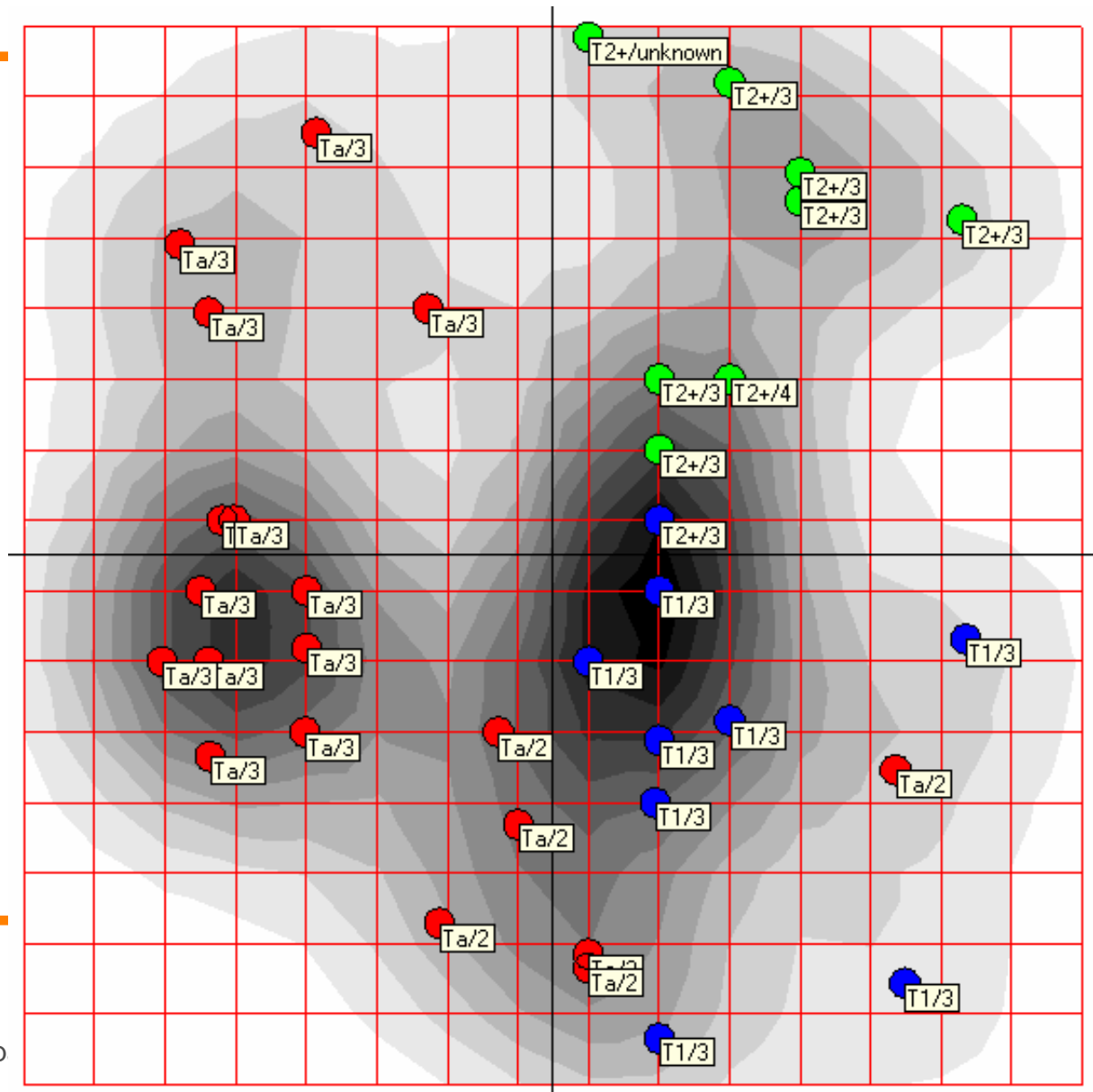
# Breast cancer dataset (Dataset I), 286 patients, 5 molecular types



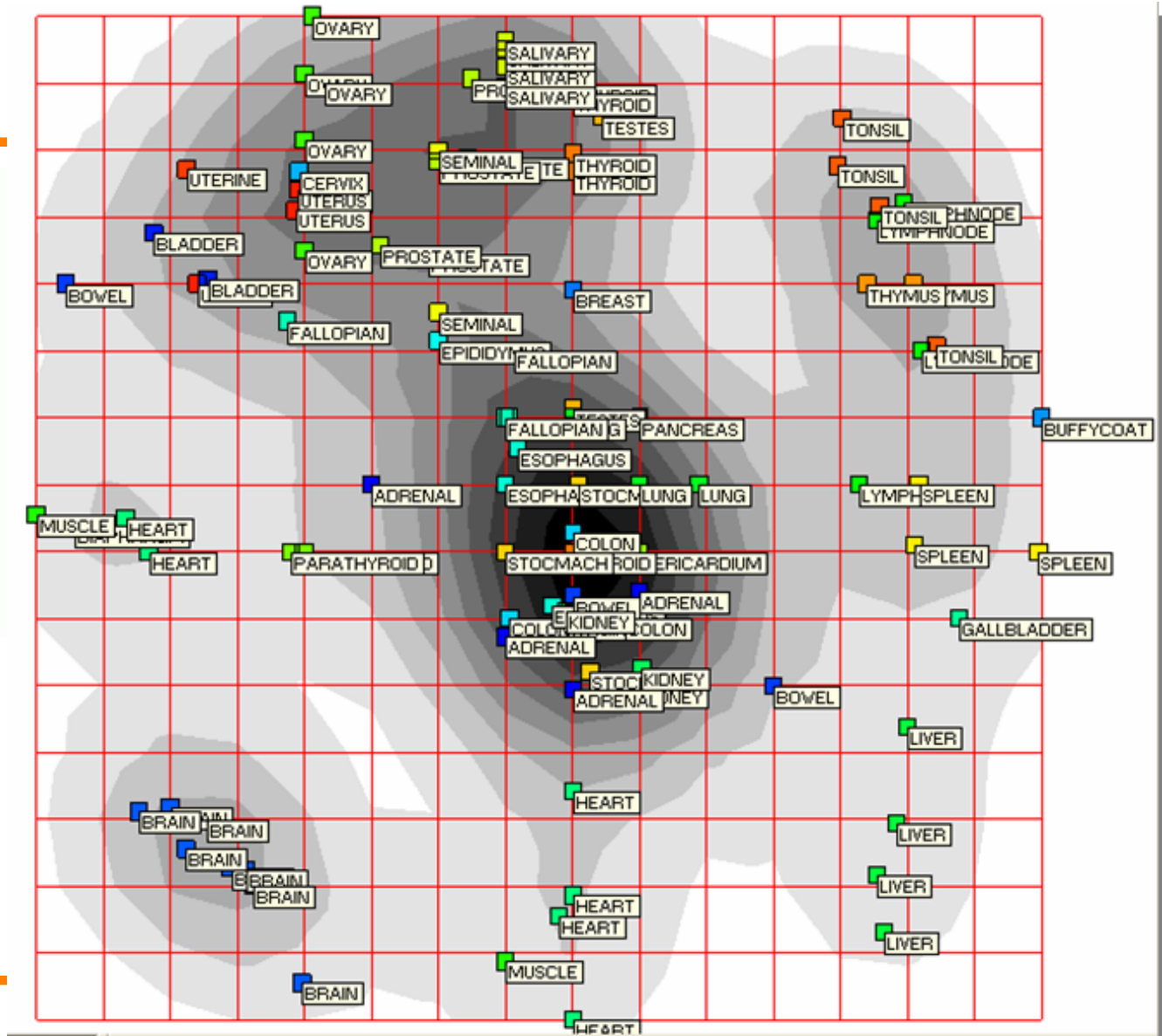
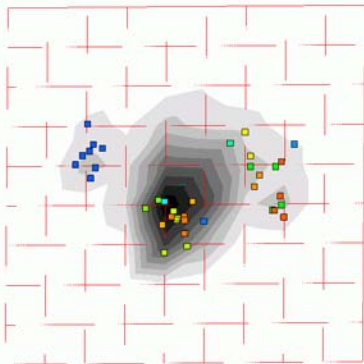
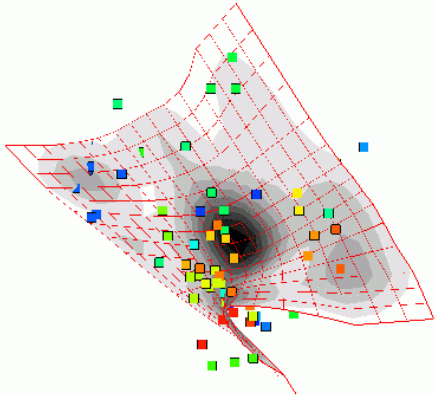
Gain in Mean Square Error ~ 30%



# Bladder cancer dataset (Dataset II), 40 patients

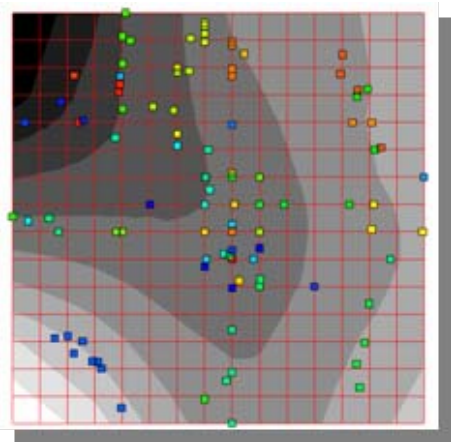
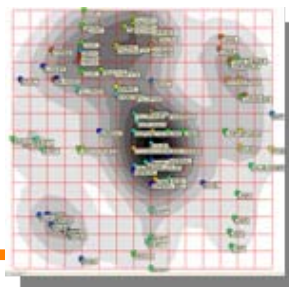


# 102 healthy tissues (Dataset III)

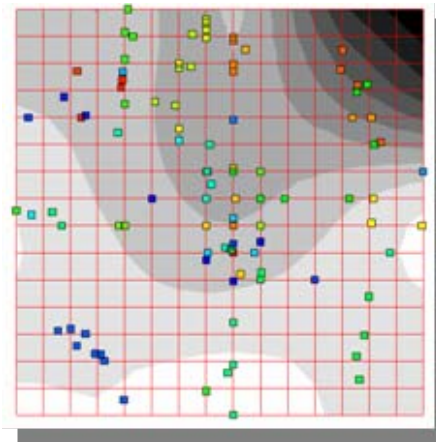




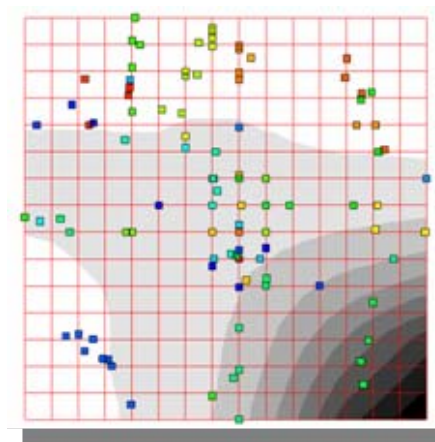
# 102 healthy tissues (Dataset III)



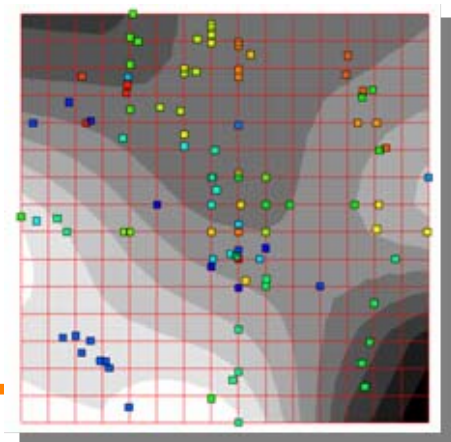
**actin gamma 2**



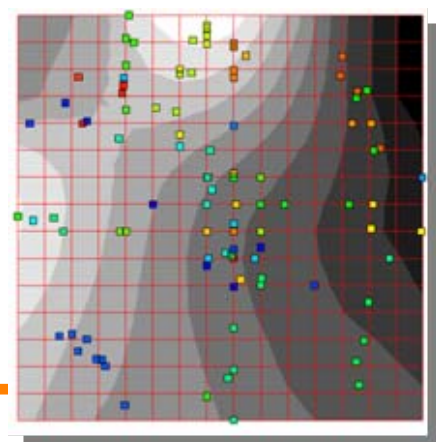
**keratin 5**



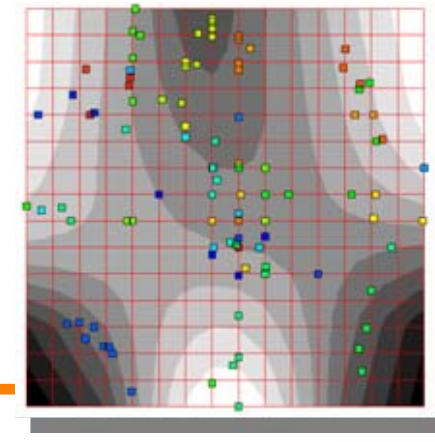
**aldolase B**



**claudin 1**

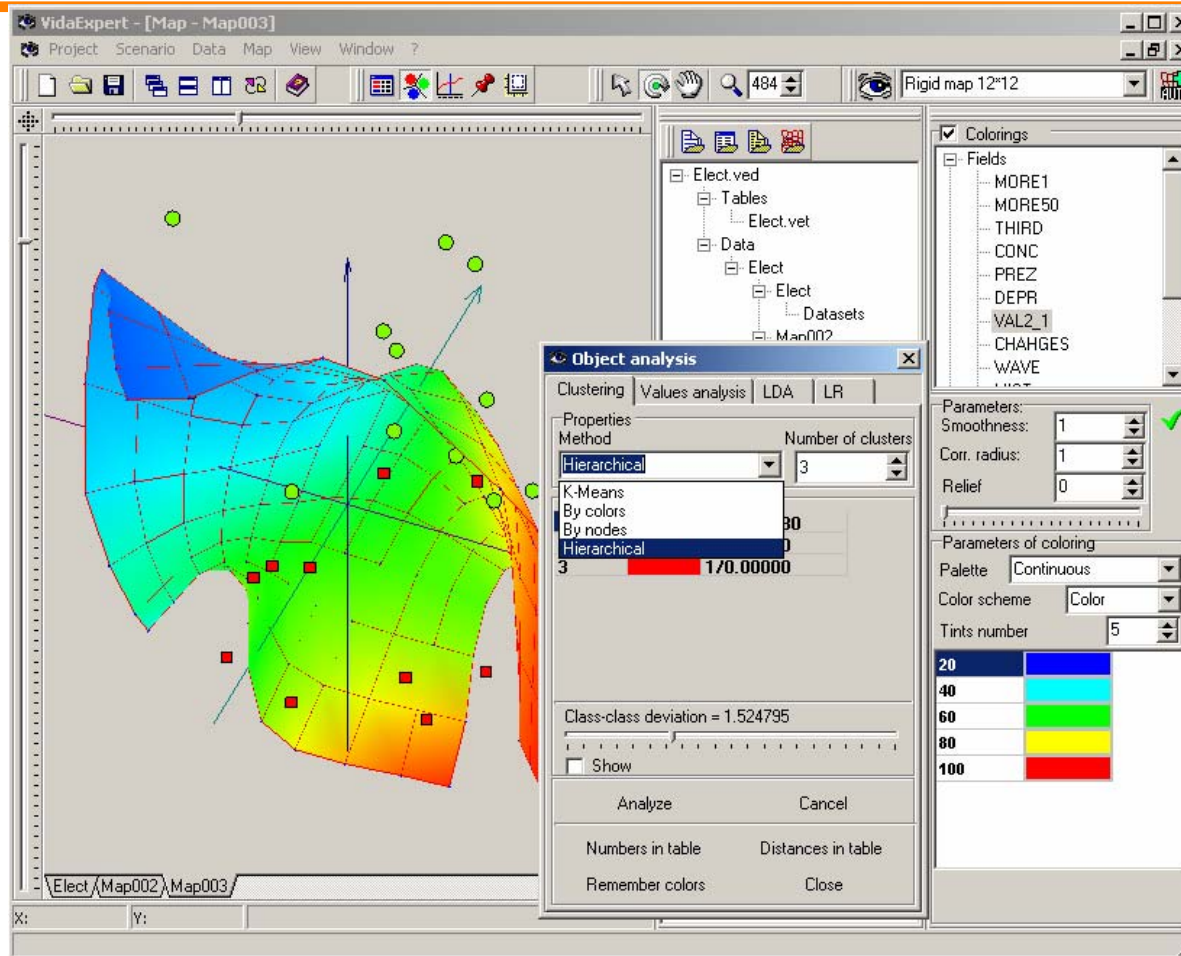


**calgranulin B**



**ERBB3**

# Implementation of the idea: VidaExpert tool



<http://bioinfo.curie.fr/projects/vidaexpert/>

# elmap C++ package

<http://bioinfo.curie.fr/projects/elmap/>

The screenshot shows a Microsoft Internet Explorer browser window with the address bar set to <http://bioinfo-out.curie.fr/projects/elmap/>. The page title is "Elastic maps". The header includes logos for IHÉS, INSTITUT DES HAUTES ÉTUDES SCIENTIFIQUES, and institutCurie. The main content area is titled "ELastic MAPs" and contains the following text:

**elmap** - is a tool for fast constructing non-linear principal surfaces with different topologies in multidimensional as well as in low-dimensional spaces, for discrete sets of weighted points.  
**keywords:** principal curve, principal surface, probabilistic, dimensionality reduction, nonlinear manifold, generative topographic mapping

**Description**

Principal curves and surfaces are nonlinear generalizations of principal components and subspaces, respectively. They can provide insightful summary of high-dimensional data not typically attainable by classical linear methods. They were first defined by Trevor Hastie and Werner Stuetzle as "self-consistent" smooth curves which pass through the "middle" of a d-dimensional probability distribution or data cloud. Good bibliography on the subject is available at <http://www.iro.umontreal.ca/~kegl/research/pcurves/>.

We present a novel algorithm to construct principal surfaces using metaphor of elasticity. The picture on the right symbolizes the idea behind the algorithm. The small points are datapoints, the big ones are a grid approximation of a principal curve. We define an elastic energy of such system and propose an effective algorithm to minimize it.

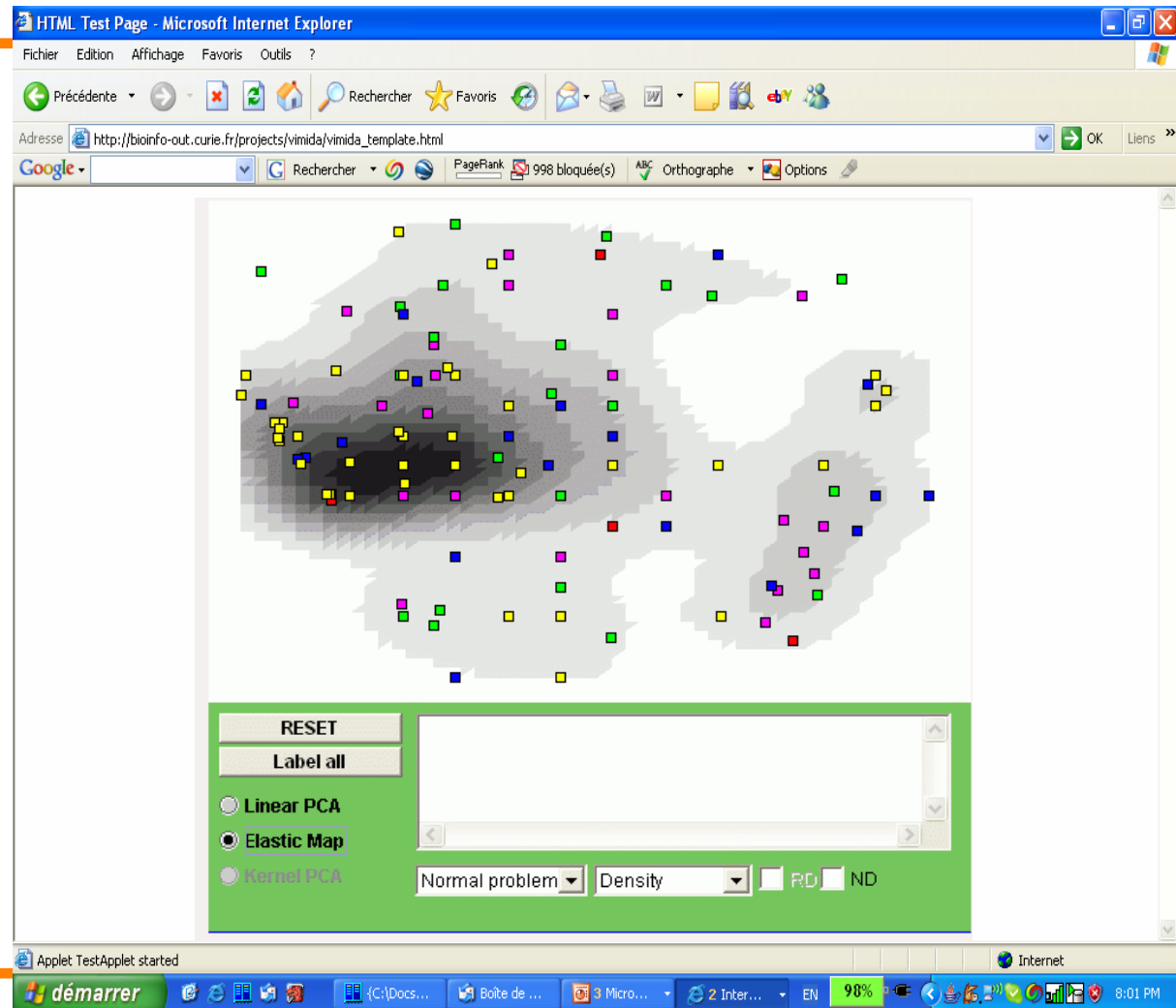
The methodology of principal curves and surfaces construction that we propose has several advantages: 1) it is "natural"; 2) it is fast; 3) it is flexible and allows many variations and adaptations; 4) it allows easily constructing surfaces with any dimension and topology.

**Examples**

The screenshot also shows a diagram on the right side of the page, which is a grid of points connected by wavy lines, representing the concept of an elastic energy system used in the algorithm.

# VIMIDA: Java-applet for multidimensional data visualization

Part of PLATAN :  
Microarray data analysis  
pipeline developed  
In Institut Curie

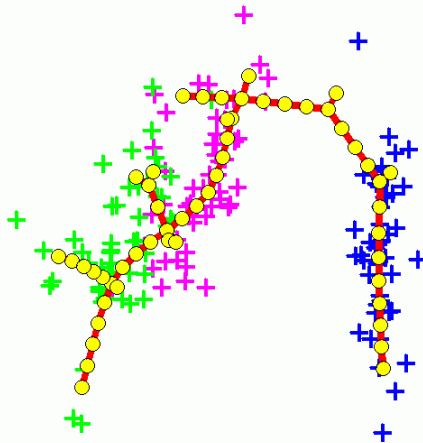
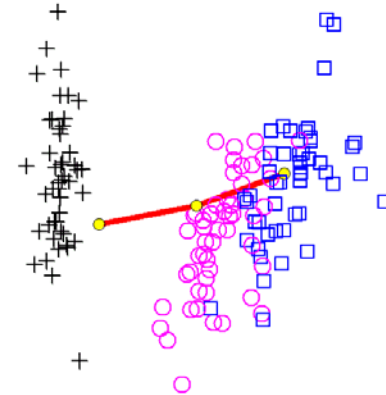
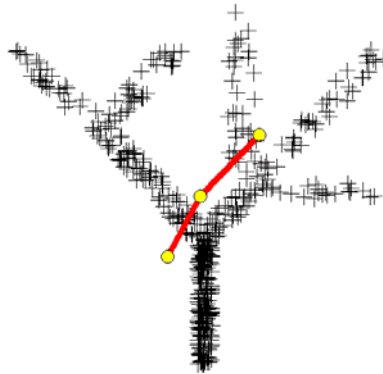
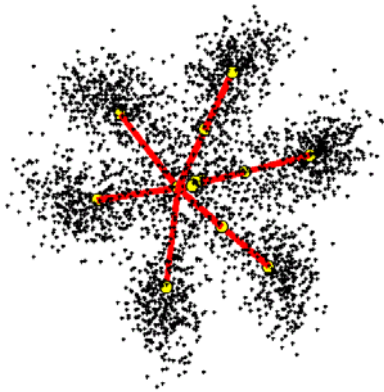





<http://bioinfo.curie.fr/projects/vimida/>

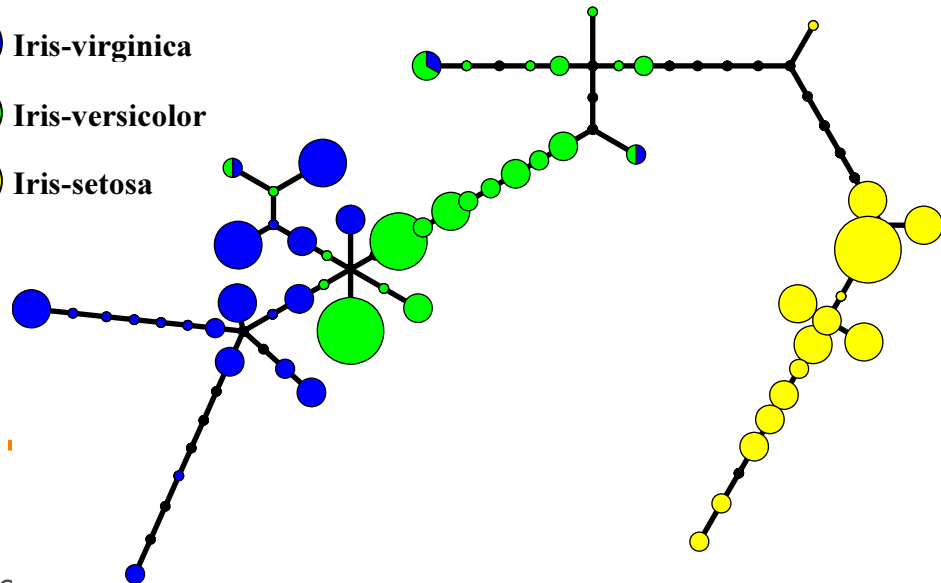
# Anonce

## Topological grammars: principal trees, cubic complexes, etc. in the talk of Professor Gorban (26 August)

---

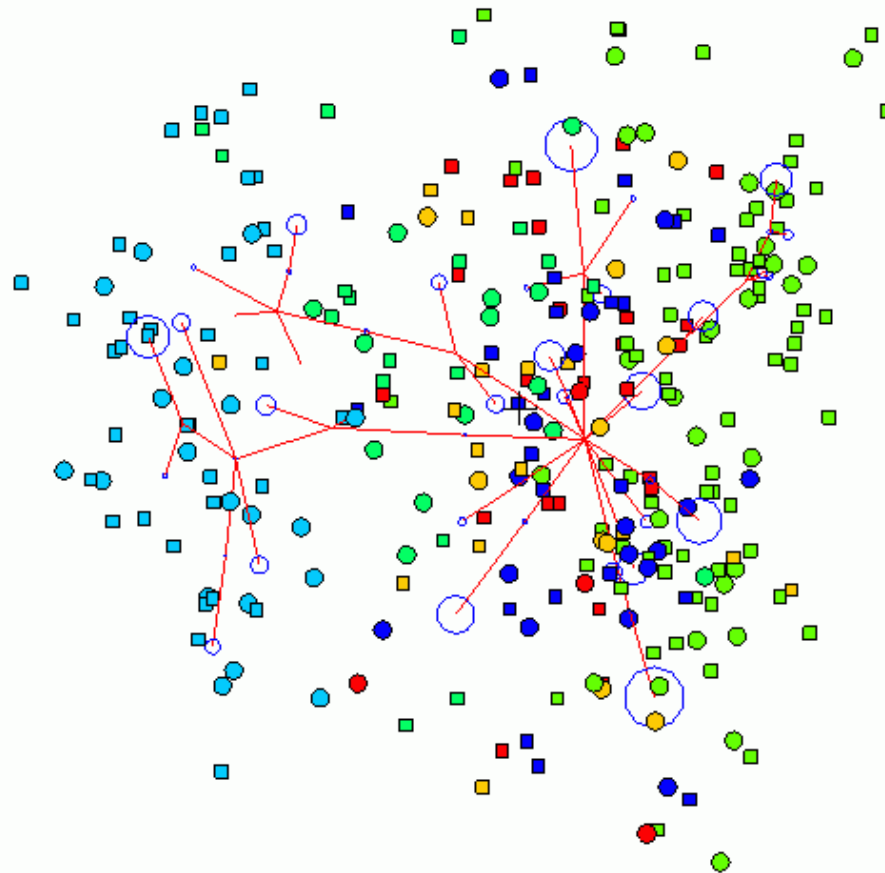


-  **Iris-virginica**
-  **Iris-versicolor**
-  **Iris-setosa**



# Branching principal components for bioinformatics data: alternative for hierarchical clustering?

---



# Papers

---

# Acknowledgements

---

**Prof. Misha Gromov (France)**

**Dr. Alexei Rossiev (Moscow)**

**Dr. Alexander Pitenko (Krasnoyarsk, Russia)**

**Neil Sumner (Leicester, UK)**

**Laboratory of neuroinformatics of  
Institute of computational Modeling,  
Russian Academy of Science**