Elastic maps with applications in bioinformatics

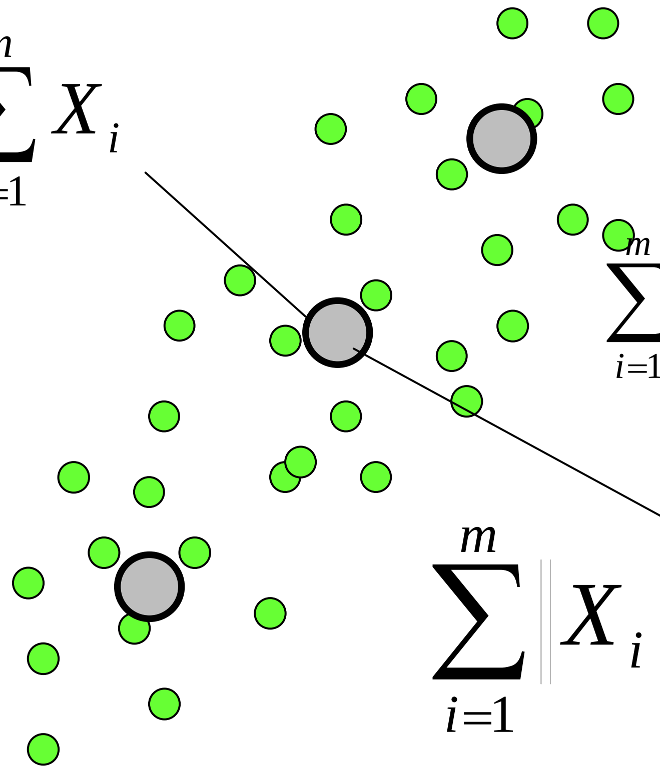
Alexander Gorban
University Of Leicester, UK
Andrei Zinovyev
Institute of Curie, France

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

$$\langle X \rangle = \frac{1}{m}\sum_{i=1}^{m} X_i$$



K-means clustering

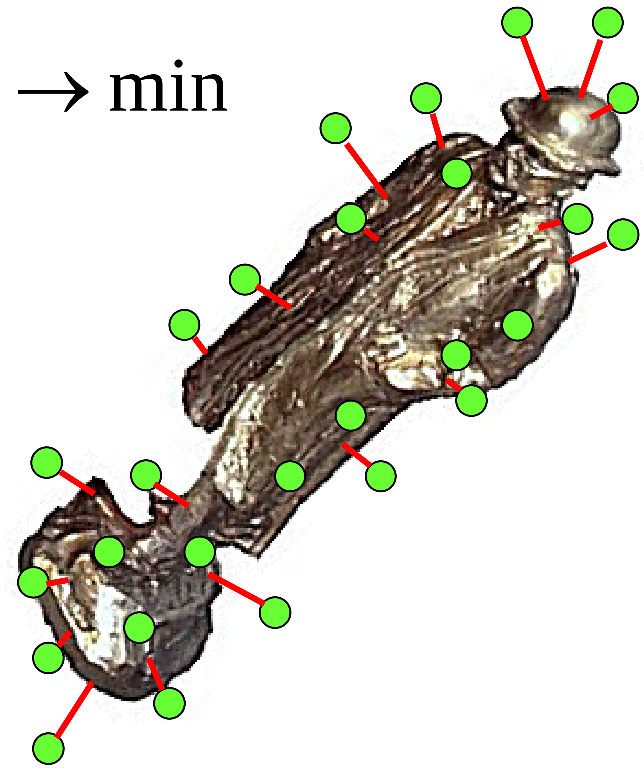
$$\sum_{i=1}^{m} \left\| X_i - closest\ Y \right\|^2 \to \min$$

$$\sum_{i=1}^{m} \left\| X_i - \langle X \rangle \right\|^2 \to \min$$

institutCurie

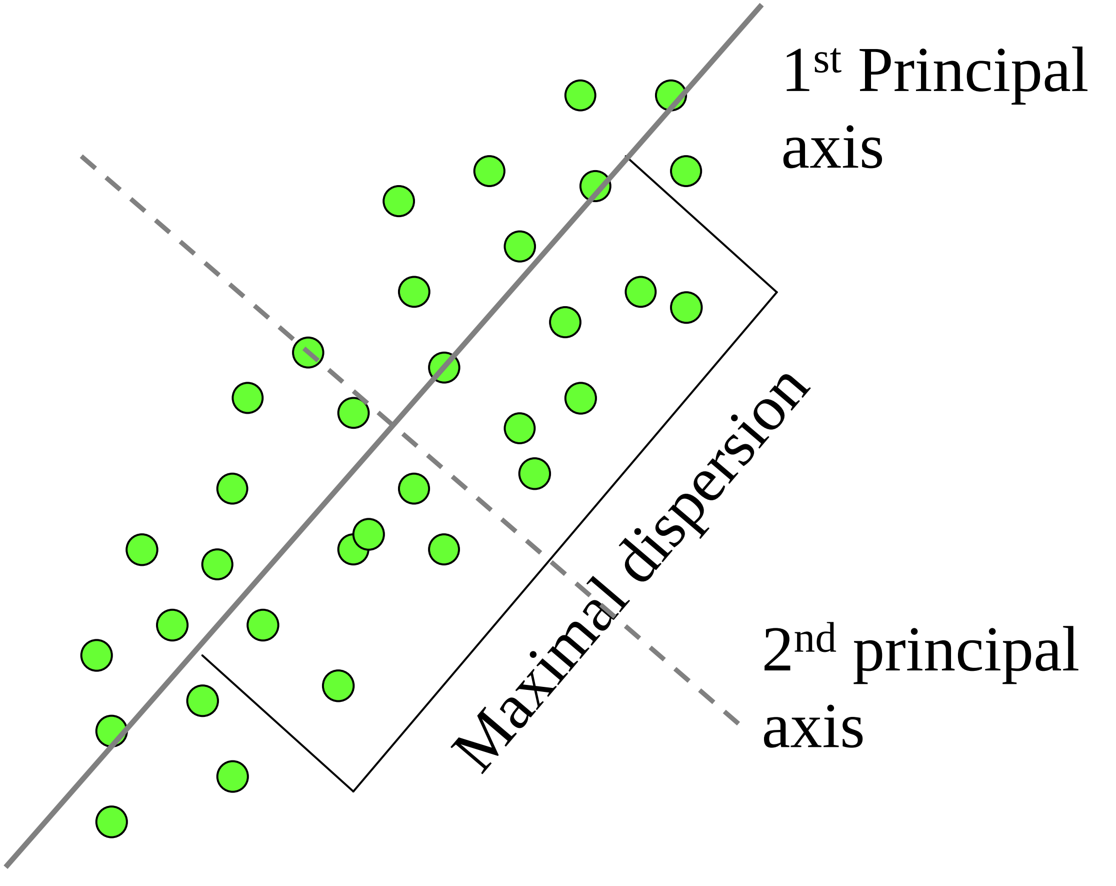
Principal “Object”

$$\sum_{i=1}^{m} \left\| \qquad \right\|^2 \to \min$$

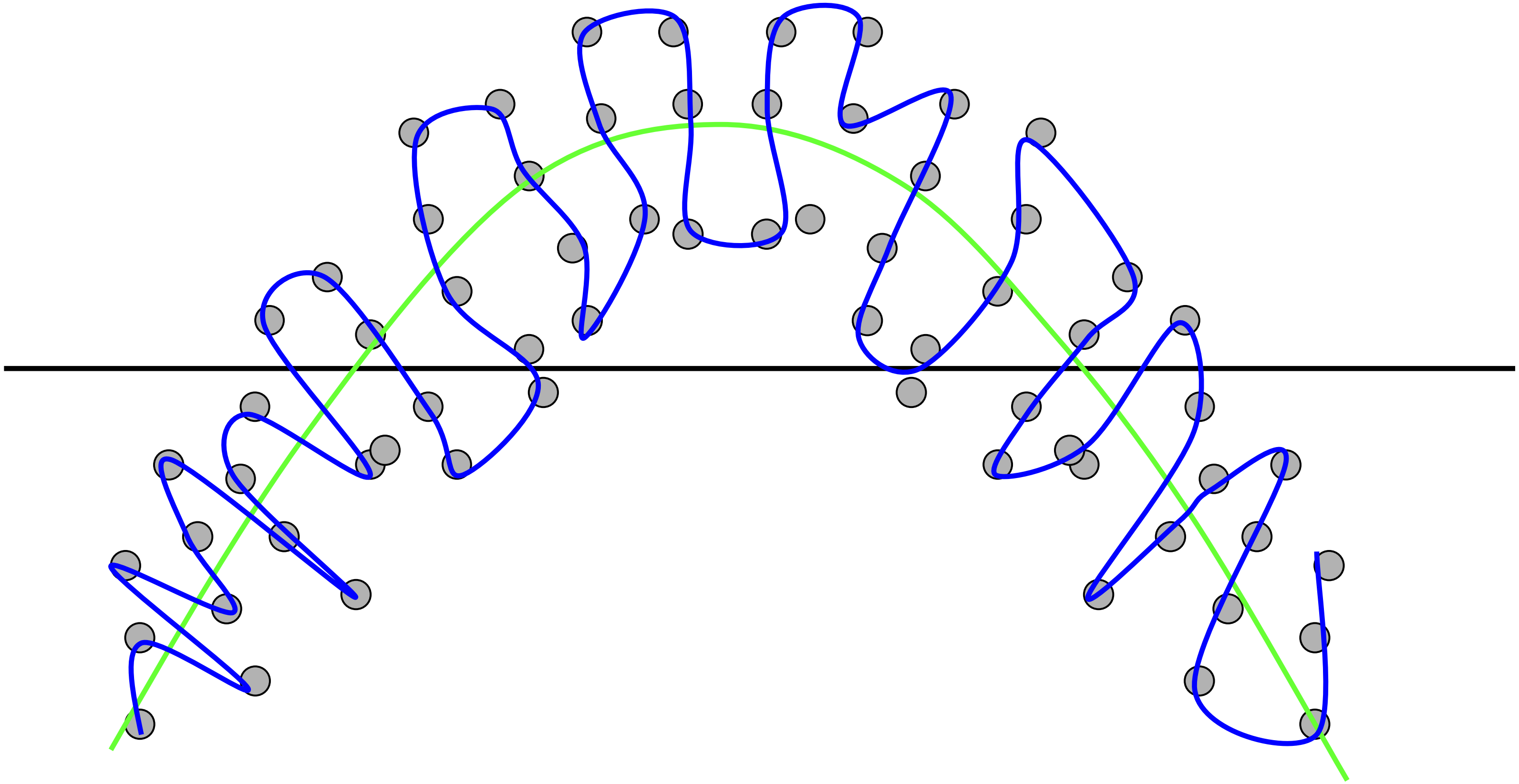


institutCurie

Principal Component Analysis

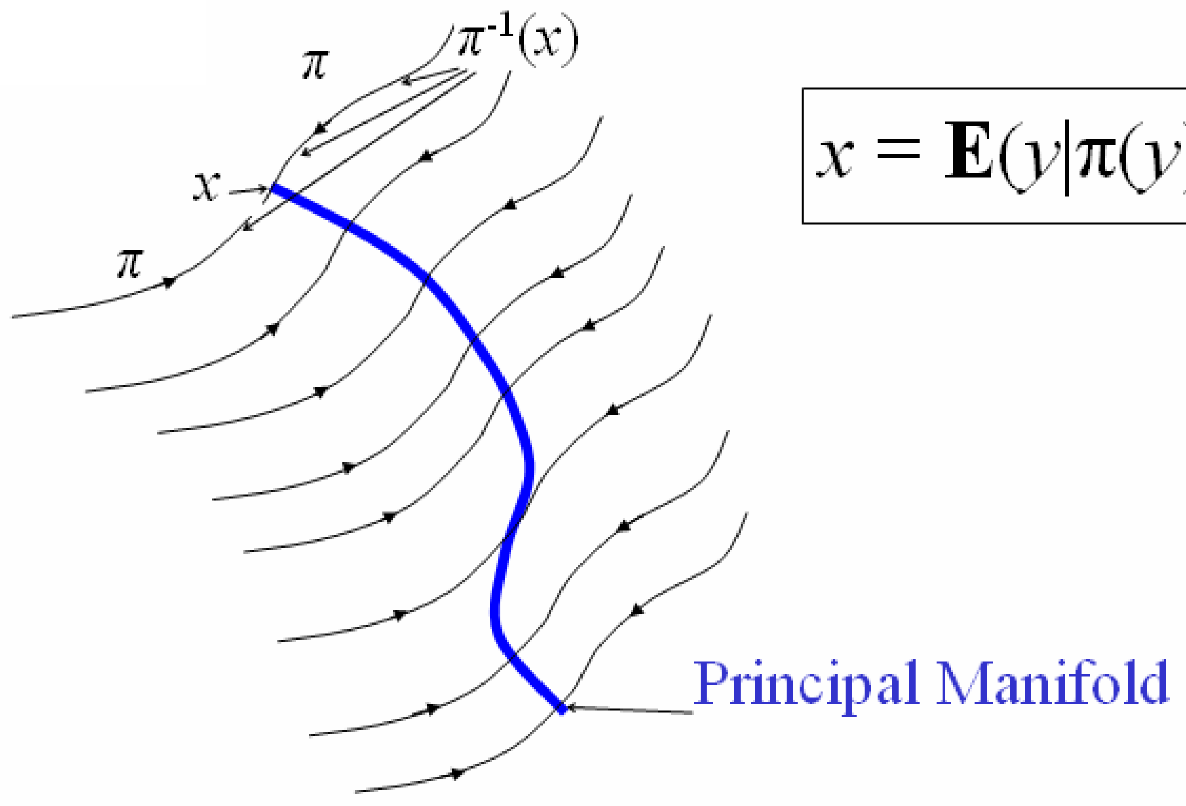


Principal manifold



institutCurie

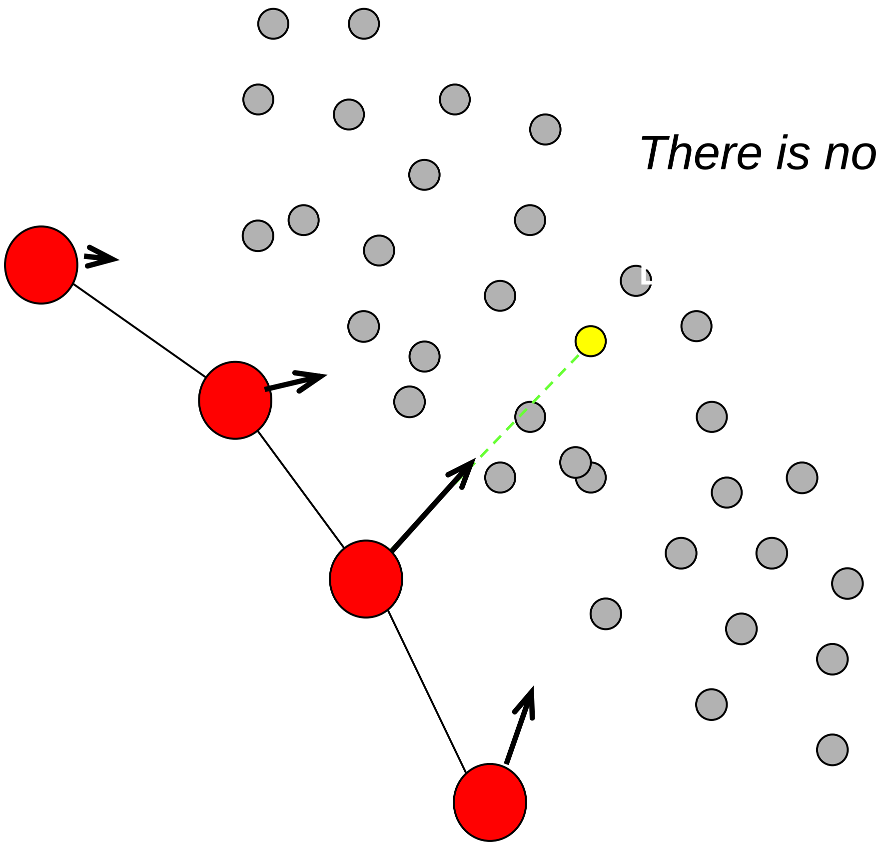
Probability distribution: idea of self-consistency



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

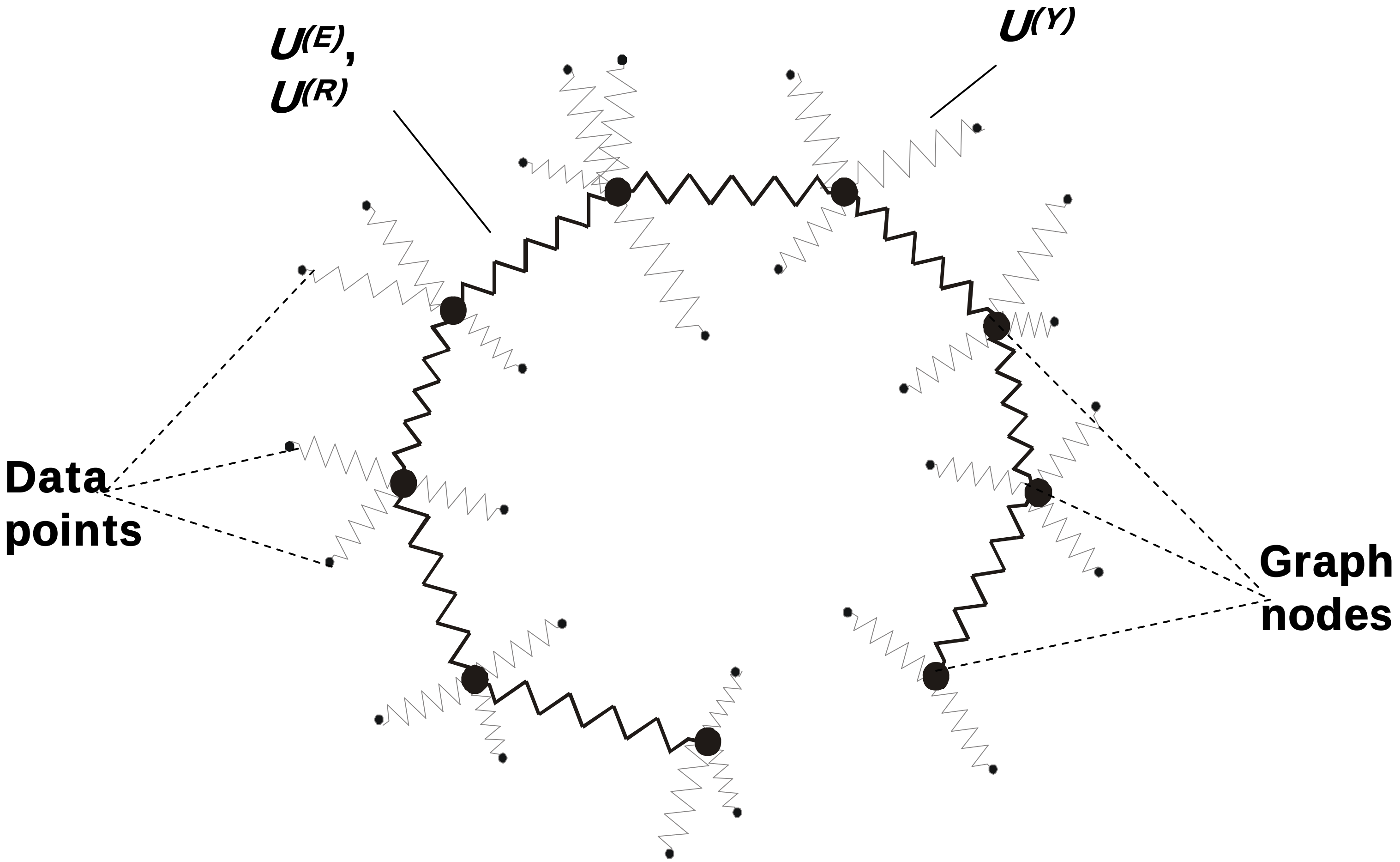
institutCurie

Self-organizing maps

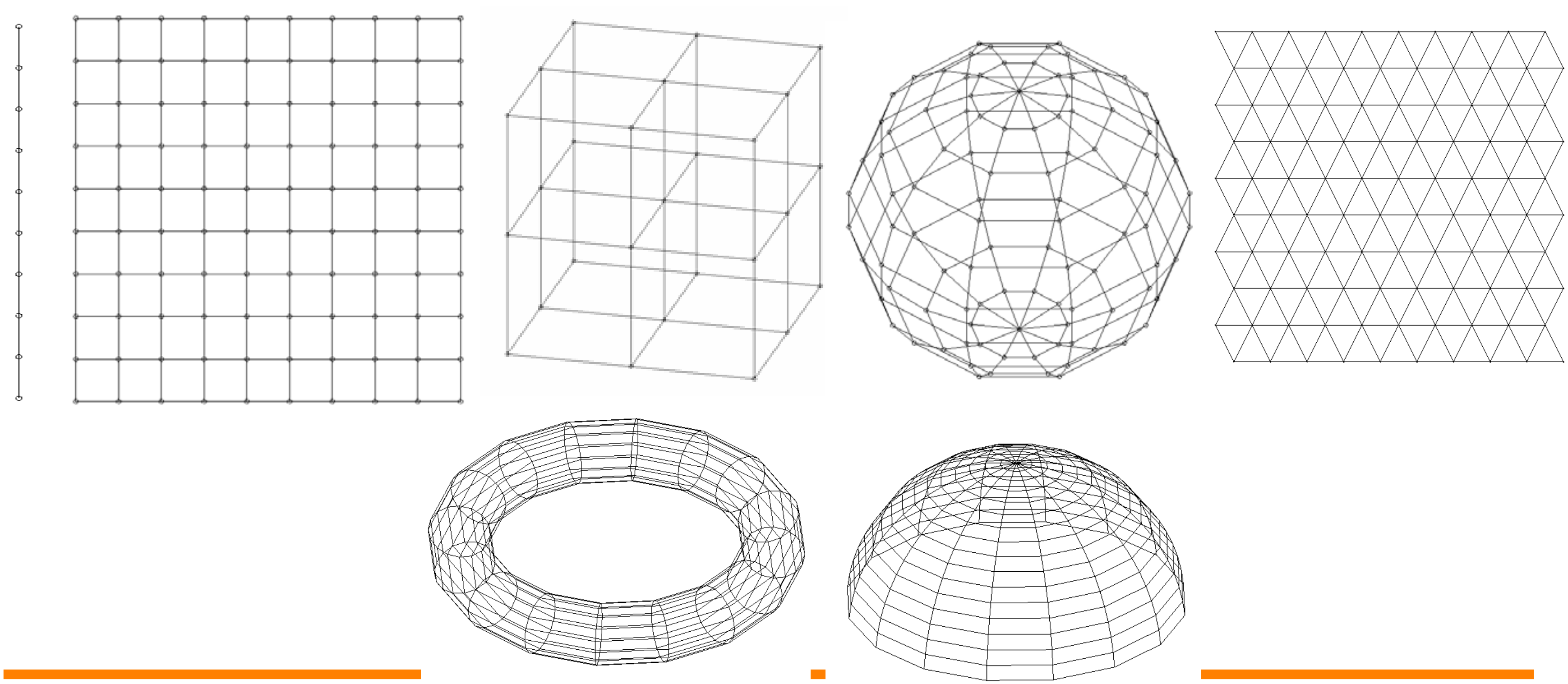


There is no global energy function

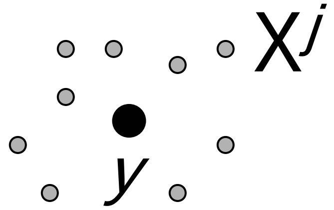
Metaphor of elasticity (energy function proposed by Gorban in 1996 at Russian national neuroinformatics workshop "Neuroinformatics and its applications - 1996")



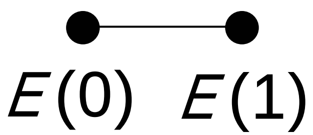
Constructing elastic nets



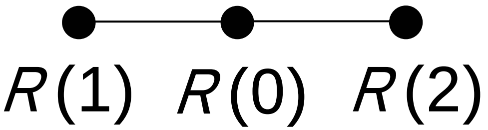
Definition of elastic energy



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



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



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

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

institutCurie

Scaling rules

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

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

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

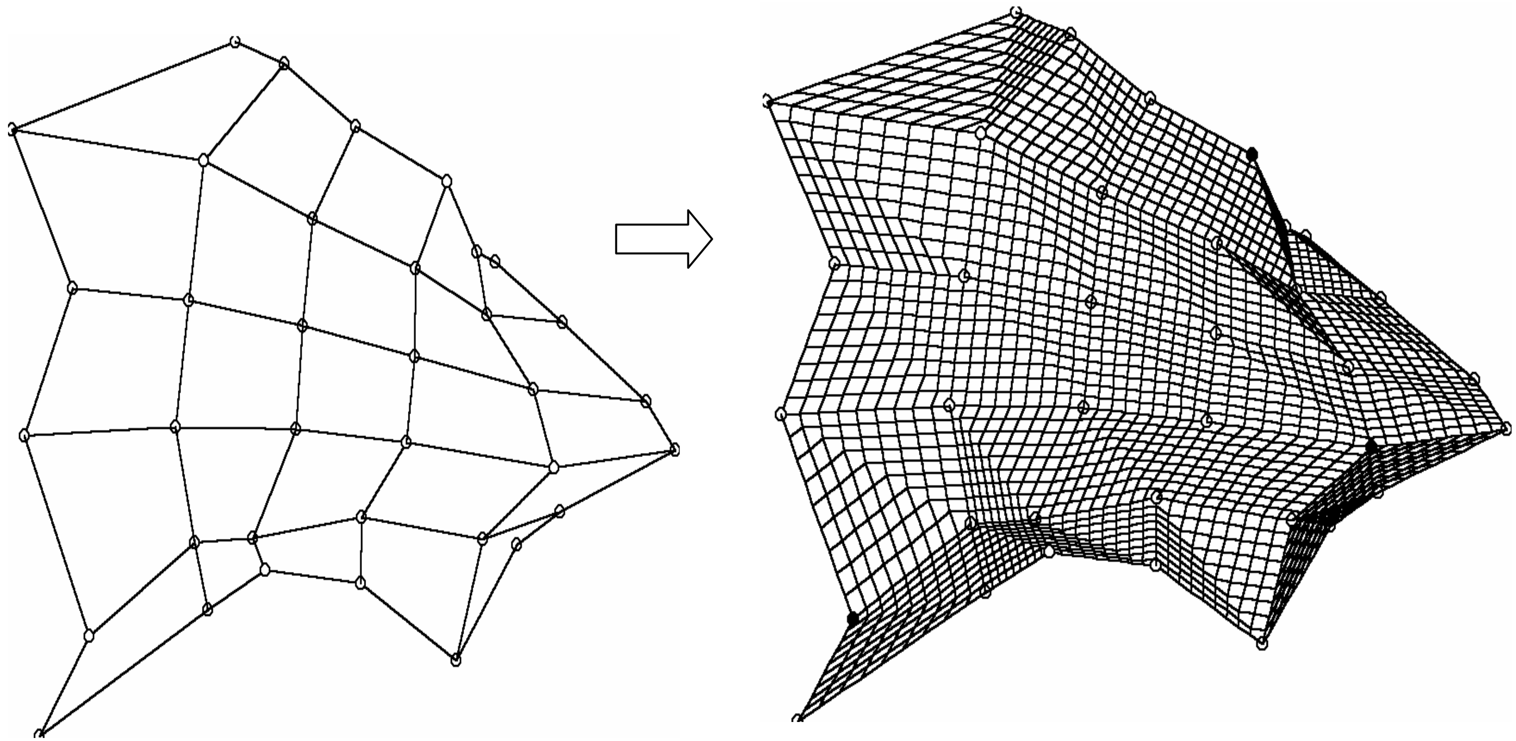
$$\lambda = \lambda_0 s^{\frac{2-d}{d}}$$

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

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

institutCurie

Elastic manifolds



institutCurie

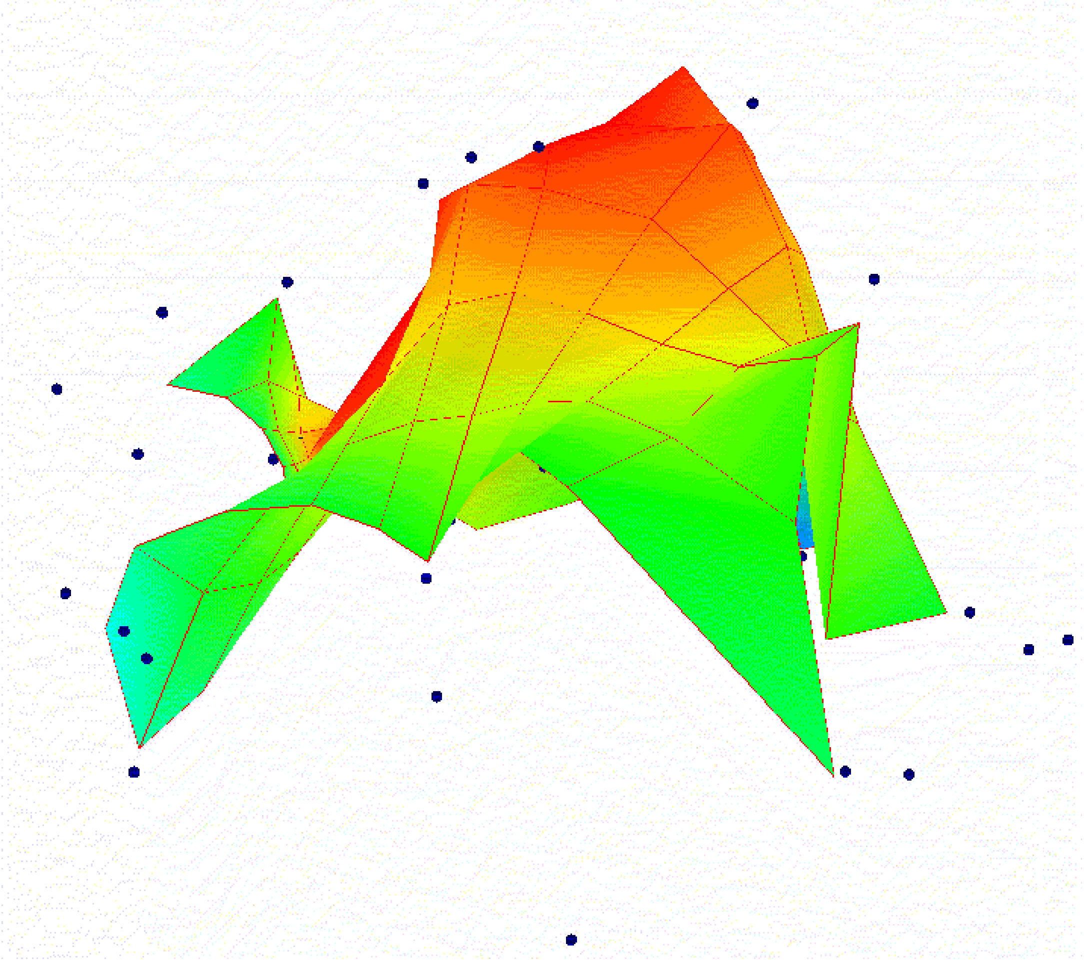
Global minimum and softening

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

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

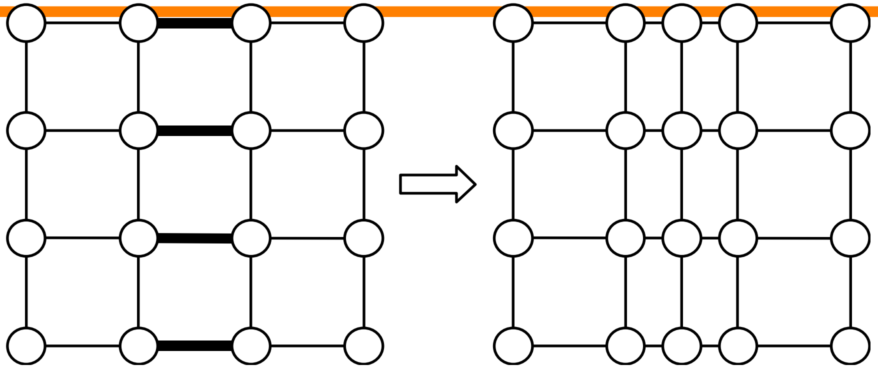
$\lambda_0, \mu_0 \approx 10^1$

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



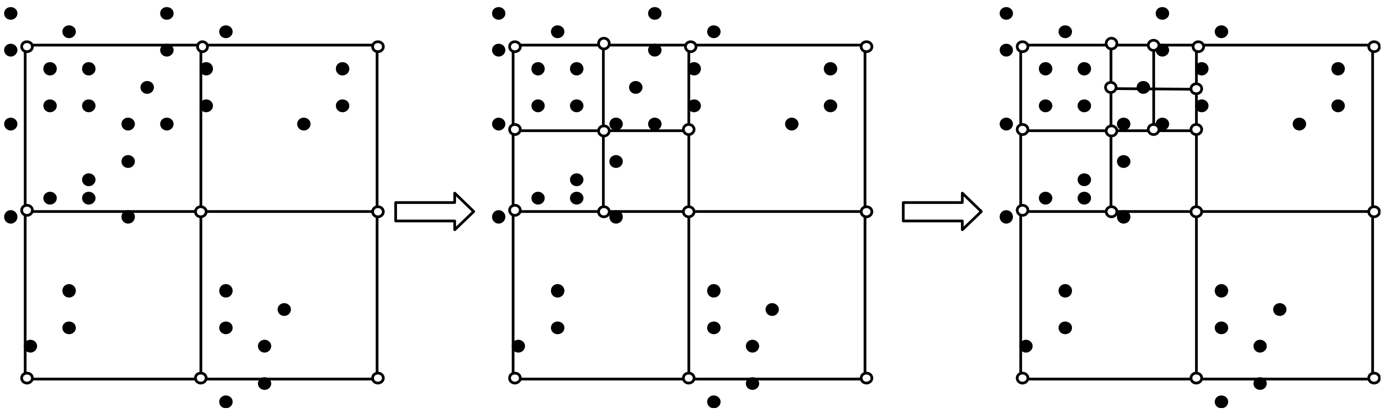
Adaptive algorithms

Refining net:



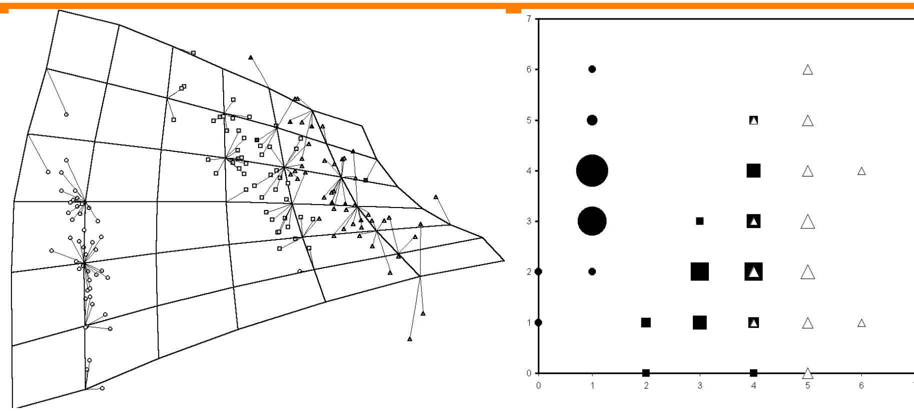
Growing net

Idea of scaling:

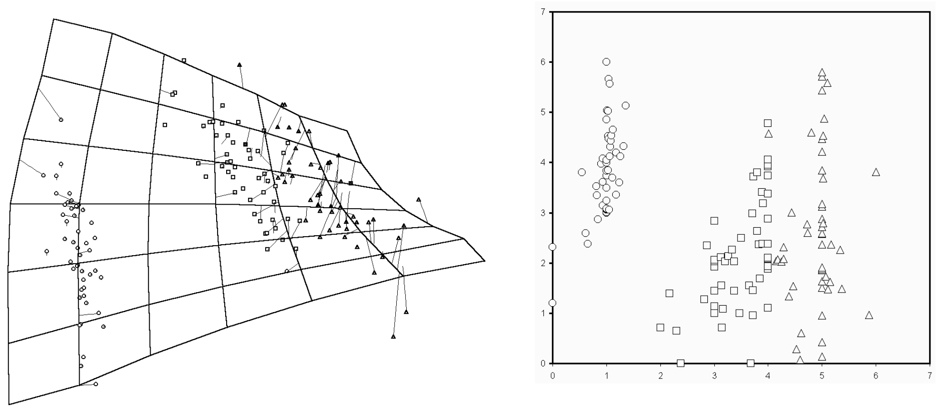


Adaptive net

Projection onto the manifold

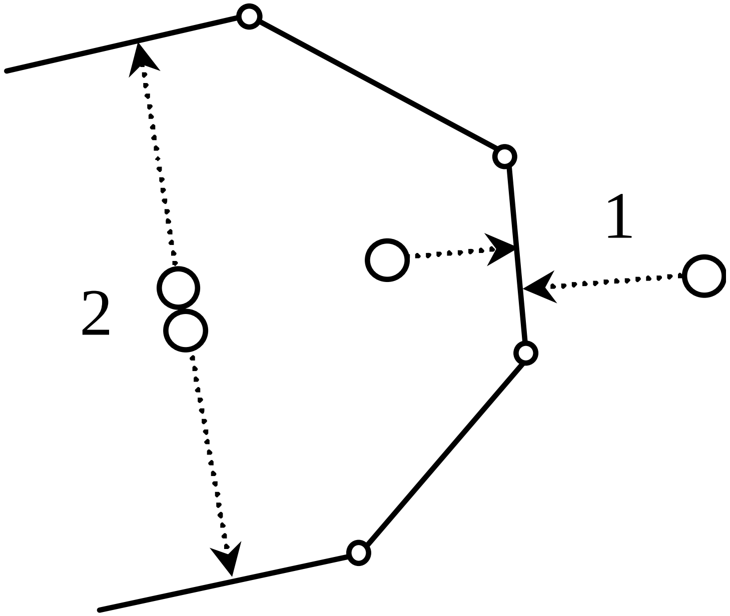


Closest node of the net



Closest point of the manifold

Mapping distortions



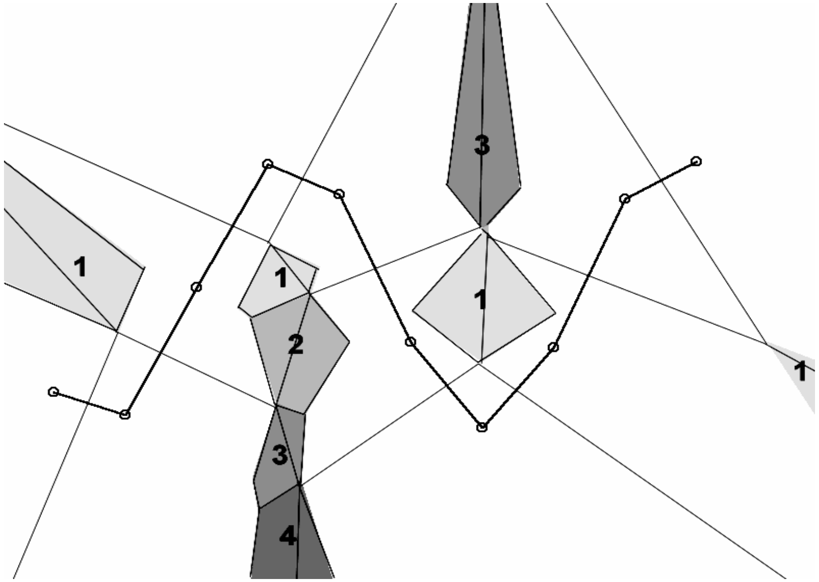
Two basic types of distortion:

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

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

institutCurie

Instability of projection

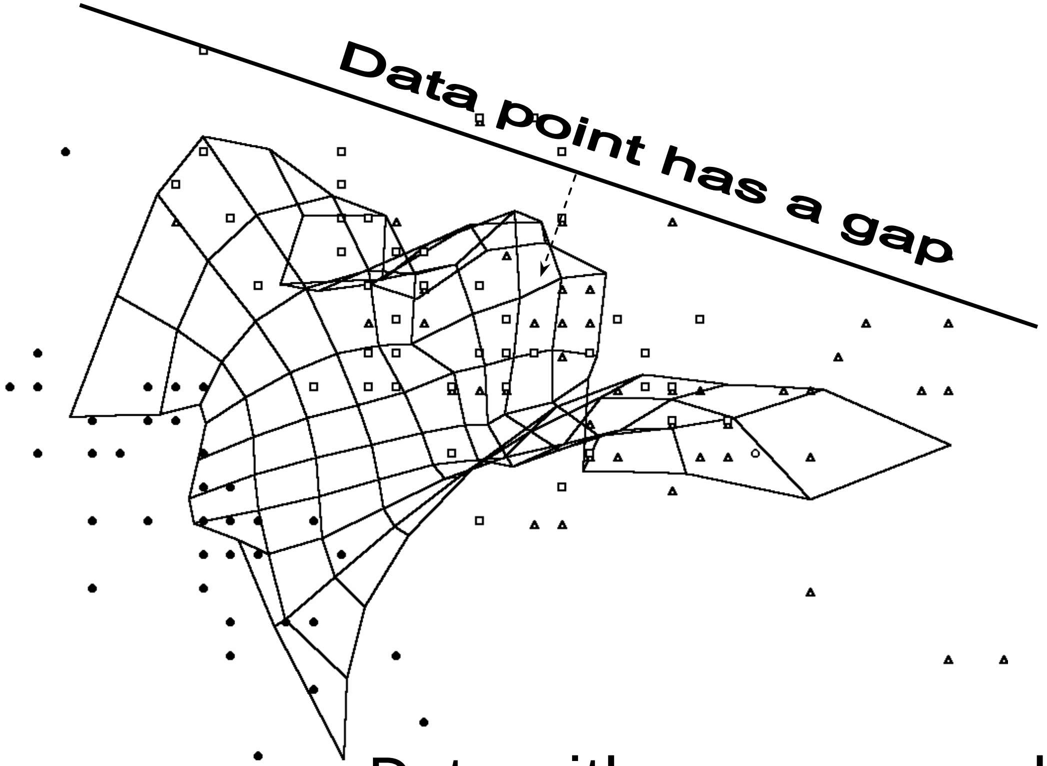


Best Matching Unit (BMU) for a data point is the closest node of the graph, BMU2 is the second-close node. If BMU and BMU2 are not adjacent on the graph, then the data point is *unstable*.

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

institutCurie

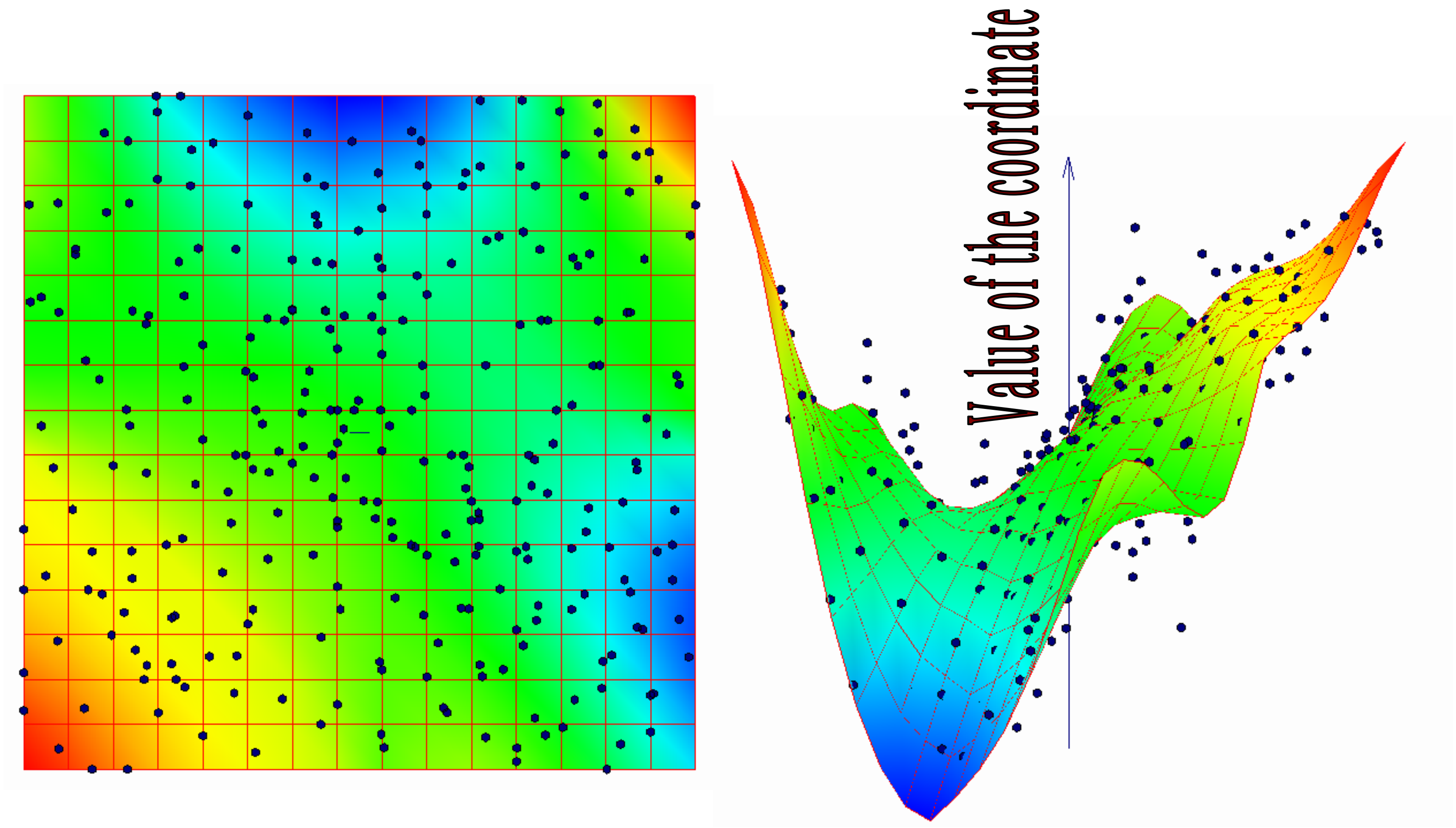
Dealing with missing values in data



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

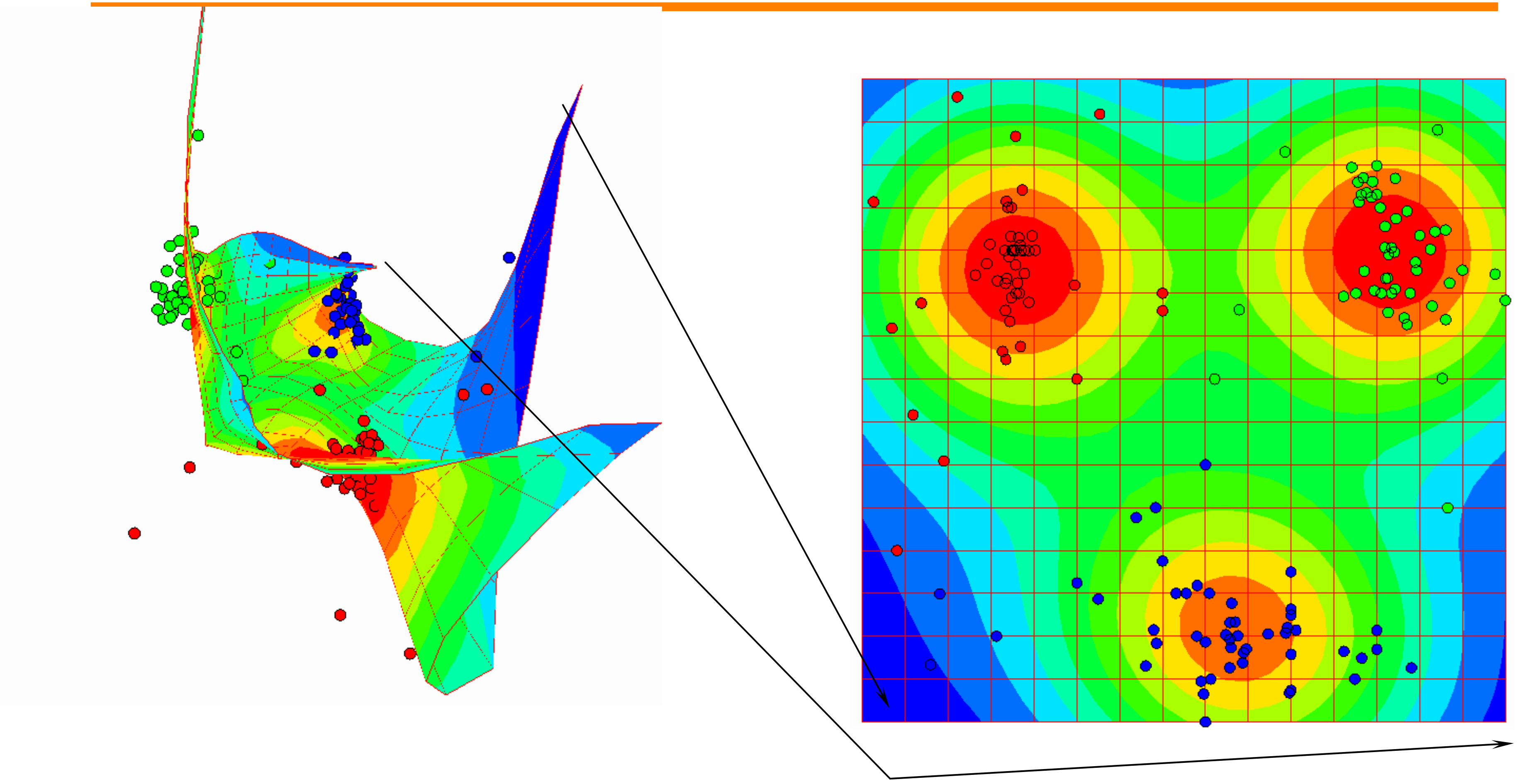
institutCurie

Colorings: visualize any function

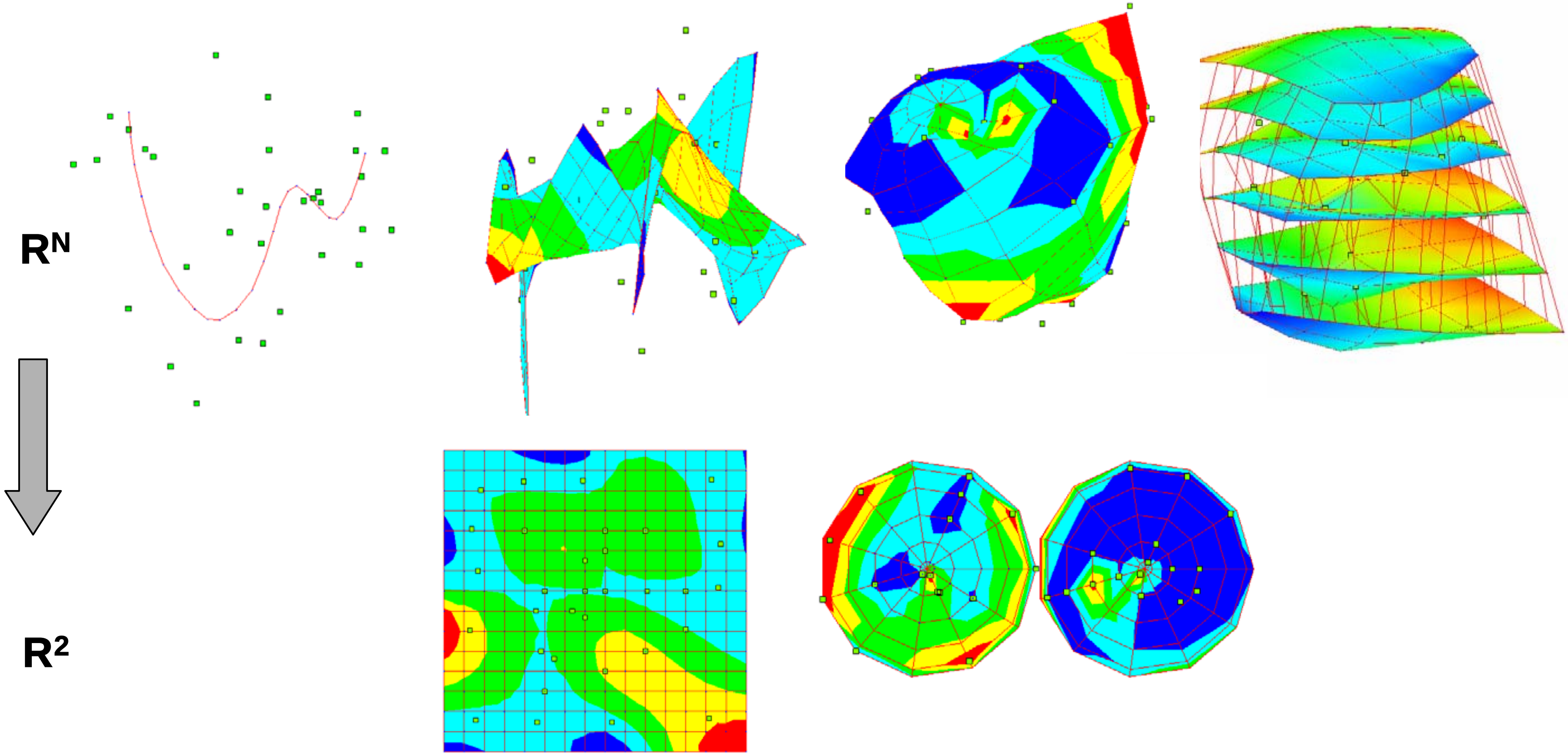


institutCurie

Density visualization

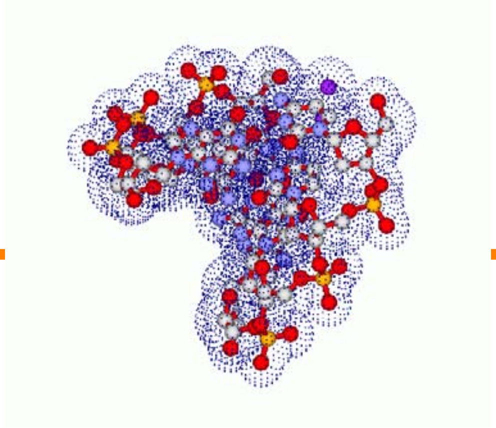


Various manifold topologies



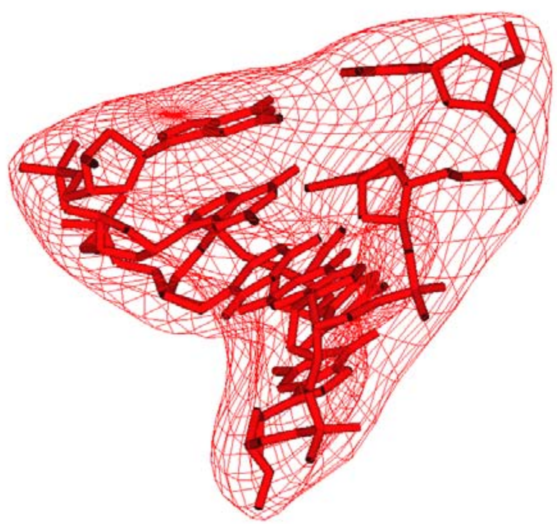
institutCurie

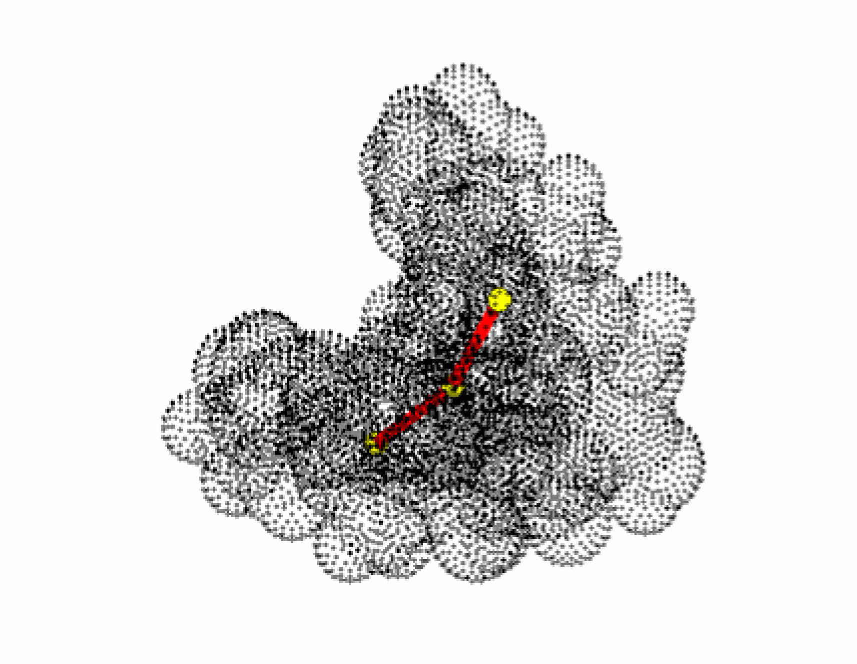
Example 1: Approximating molecular surface



Approximating by 2D spherical grid

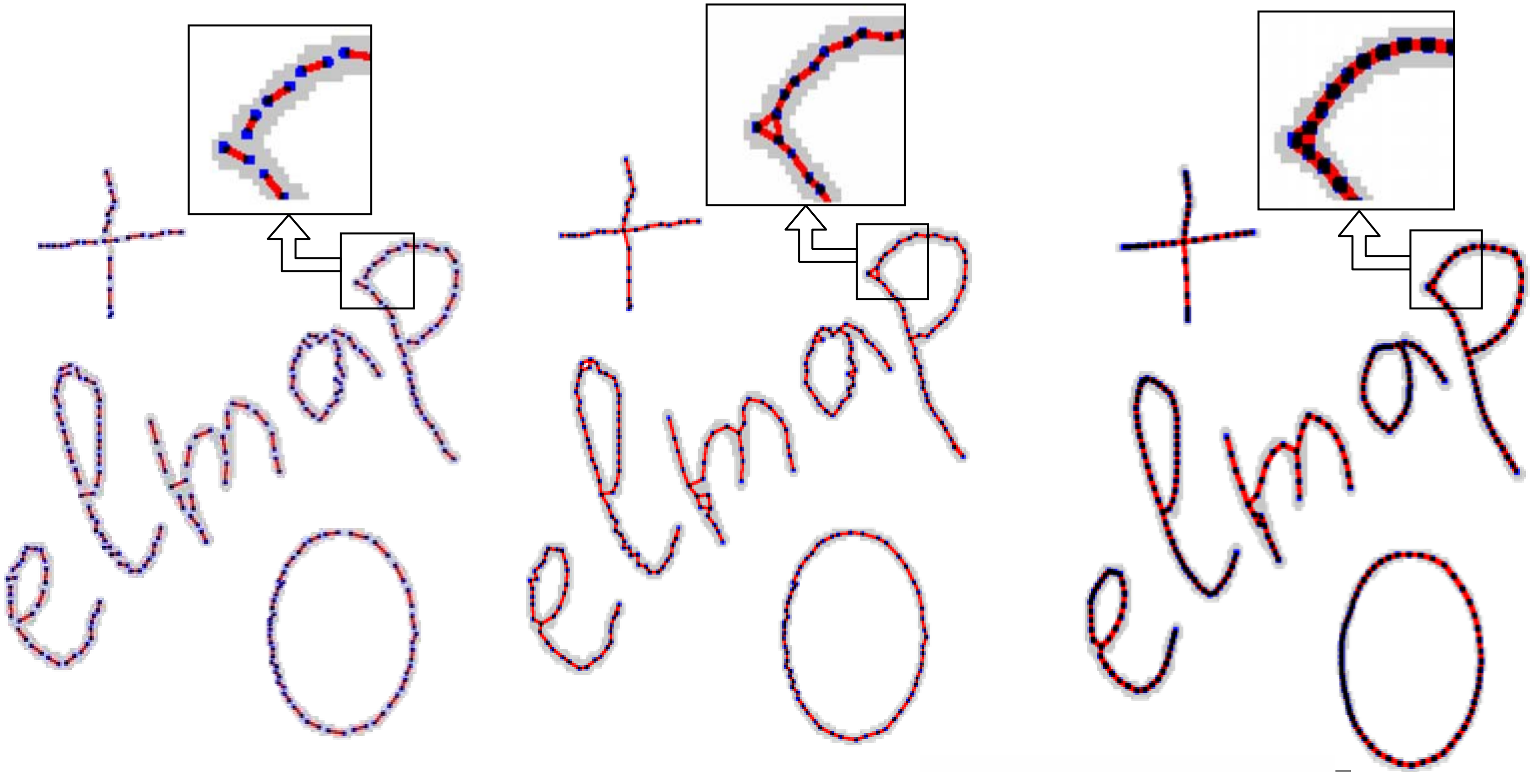
Approximating by 1D curve





institutCurie

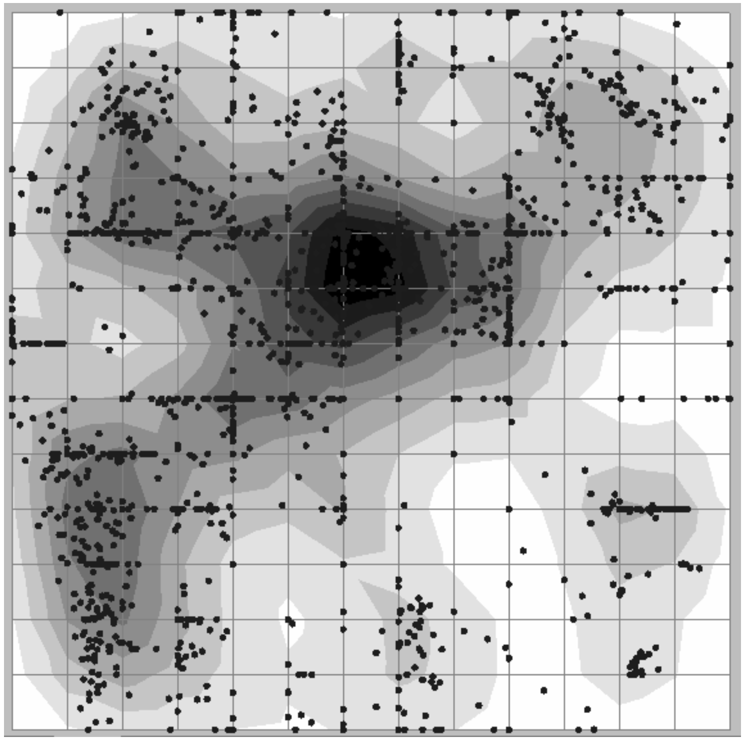
Example 2: Image skeletonization or clustering around curves



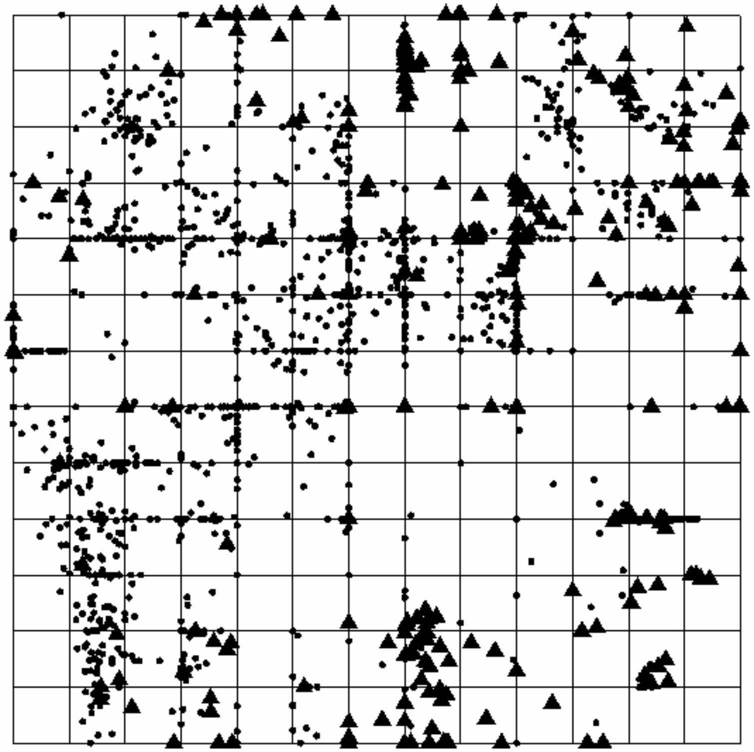
institutCurie

Example 3: Medical table

1700 patients with infarctus myocarde



Patients map, density

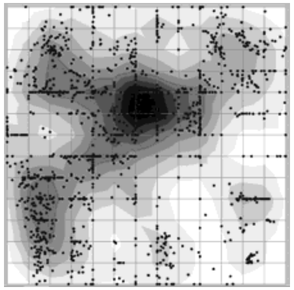


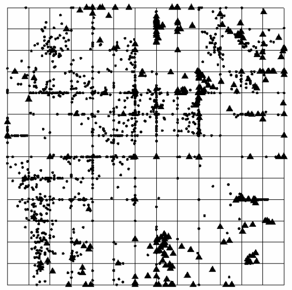
Lethal cases

institutCurie

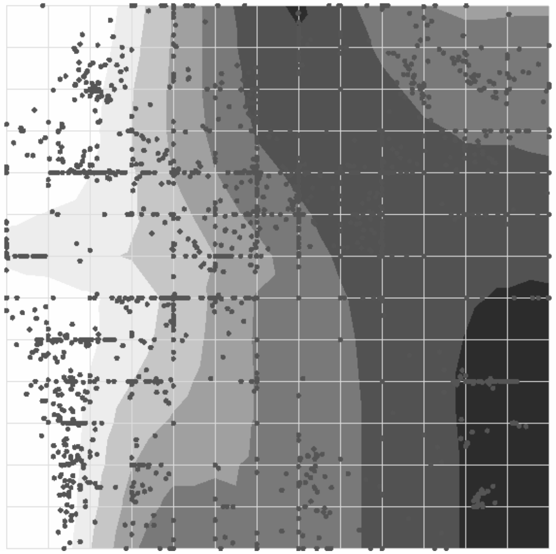
Example 3: Medical table

1700 patients with infarctus myocarde

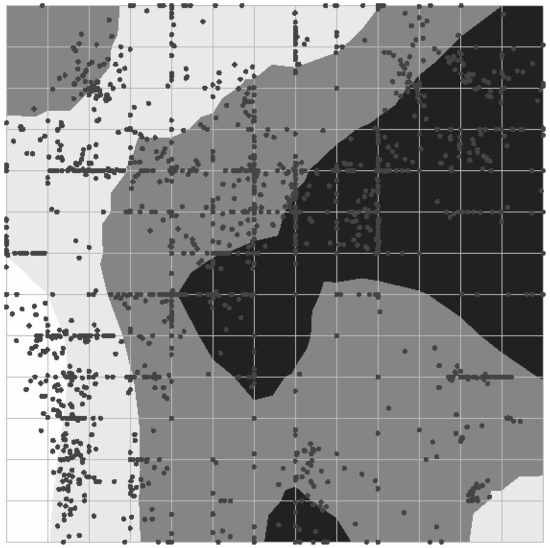




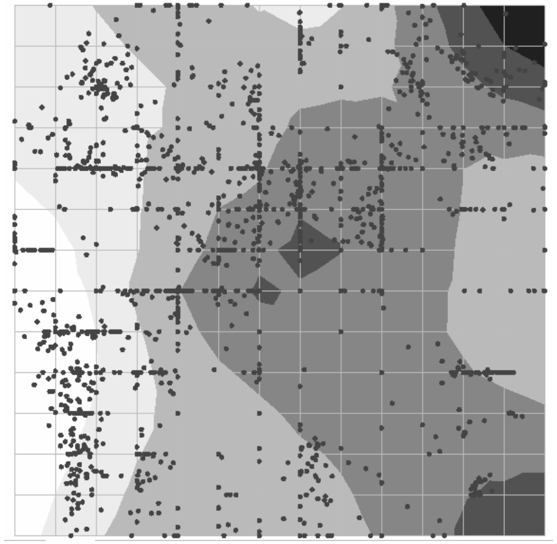
128 clinical variables



Age

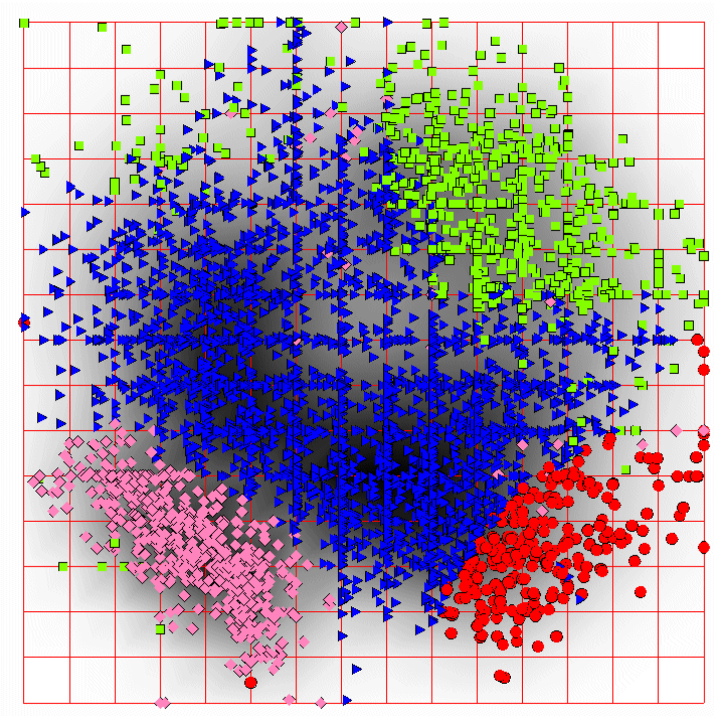


Numberof infarctus in anamnesis

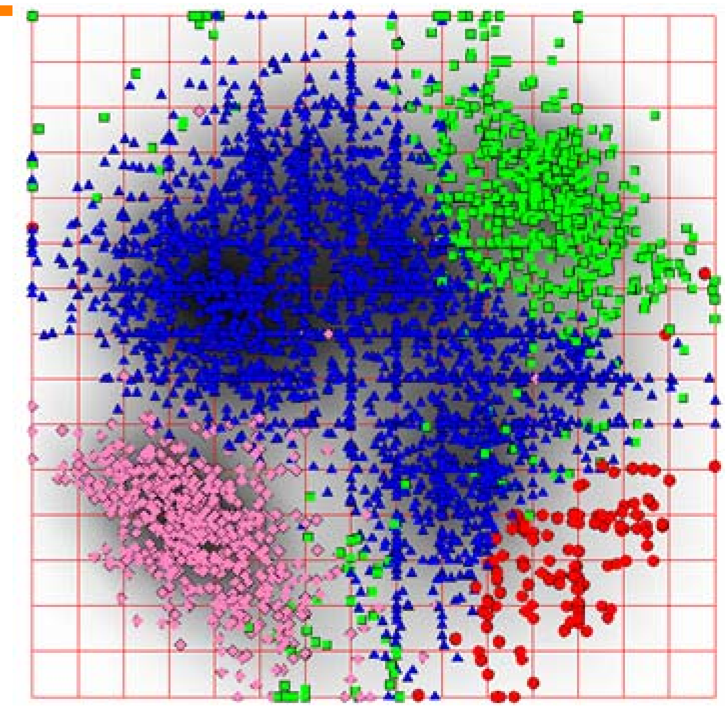


Stenocardia functional class

Example 4: Codon usage in all genes of one genome



Escherichia coli



Bacillus subtilis

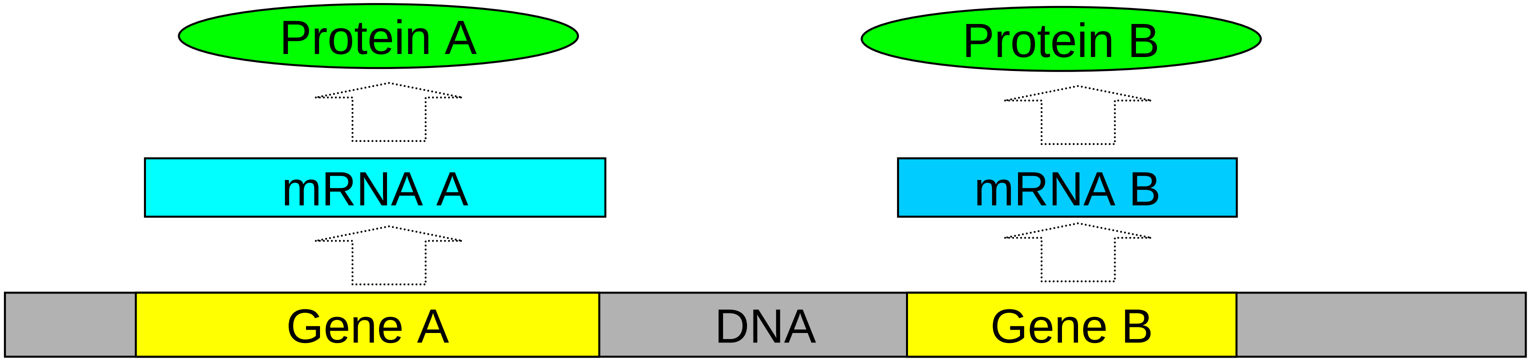
Majority of genes

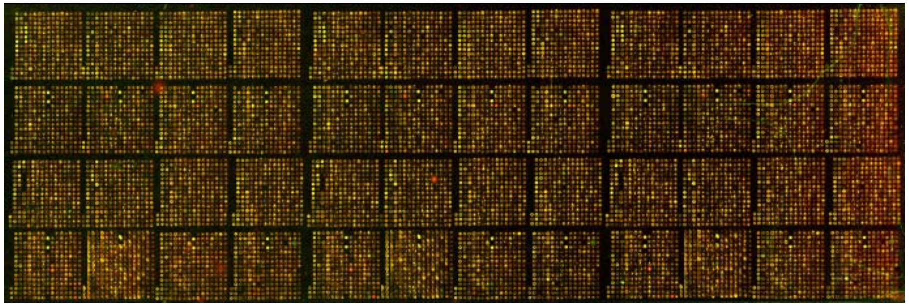
"Foreign" genes

Highly expressed genes

"Hydrophobic" genes

Microarray technology and microarray datasets



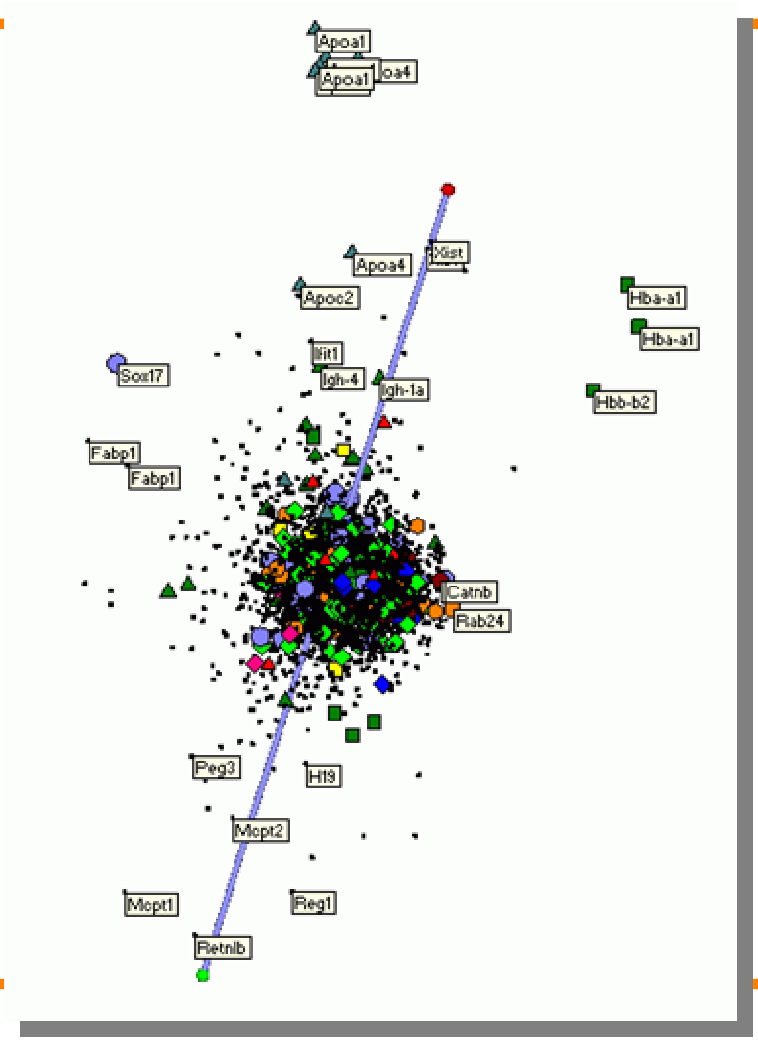


One spot corresponds
to a gene
(mRNA concentration)

Table of numbers, characteristic size is 10000 genes x100 samples

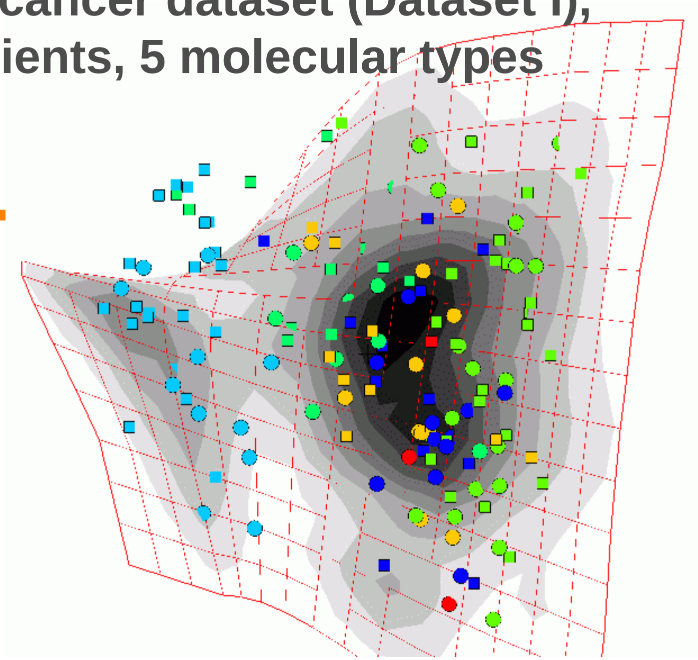
institutCurie

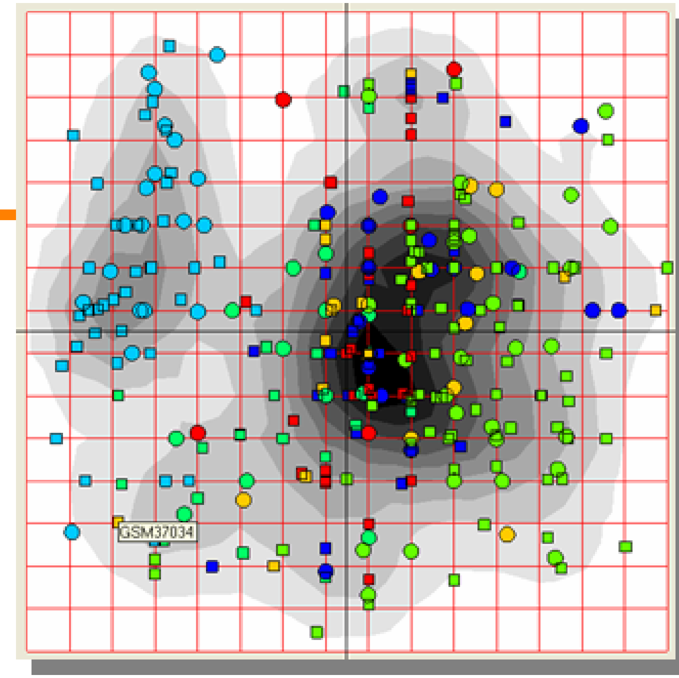
Gene space:
every point correspond to a gene characterized by its expression in m samples



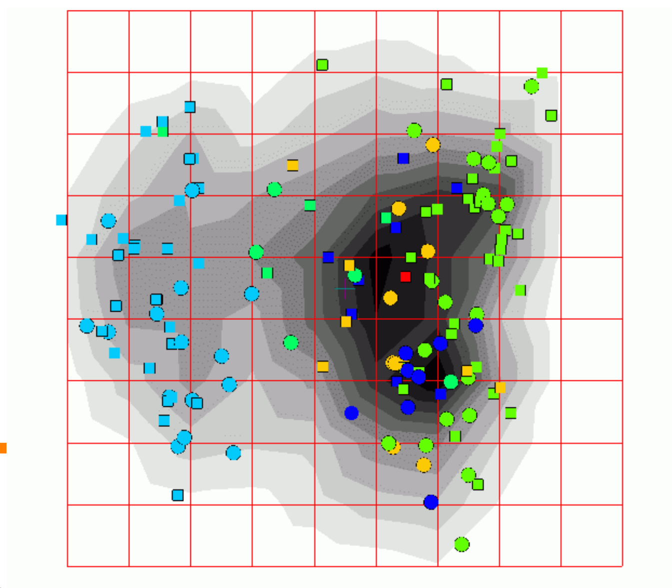
institutCurie

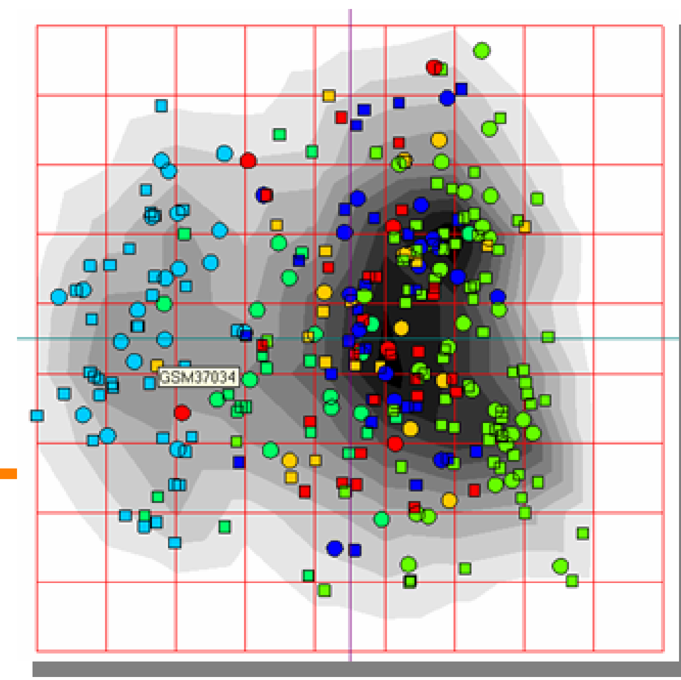
Breast cancer dataset (Dataset I), 286 patients, 5 molecular types



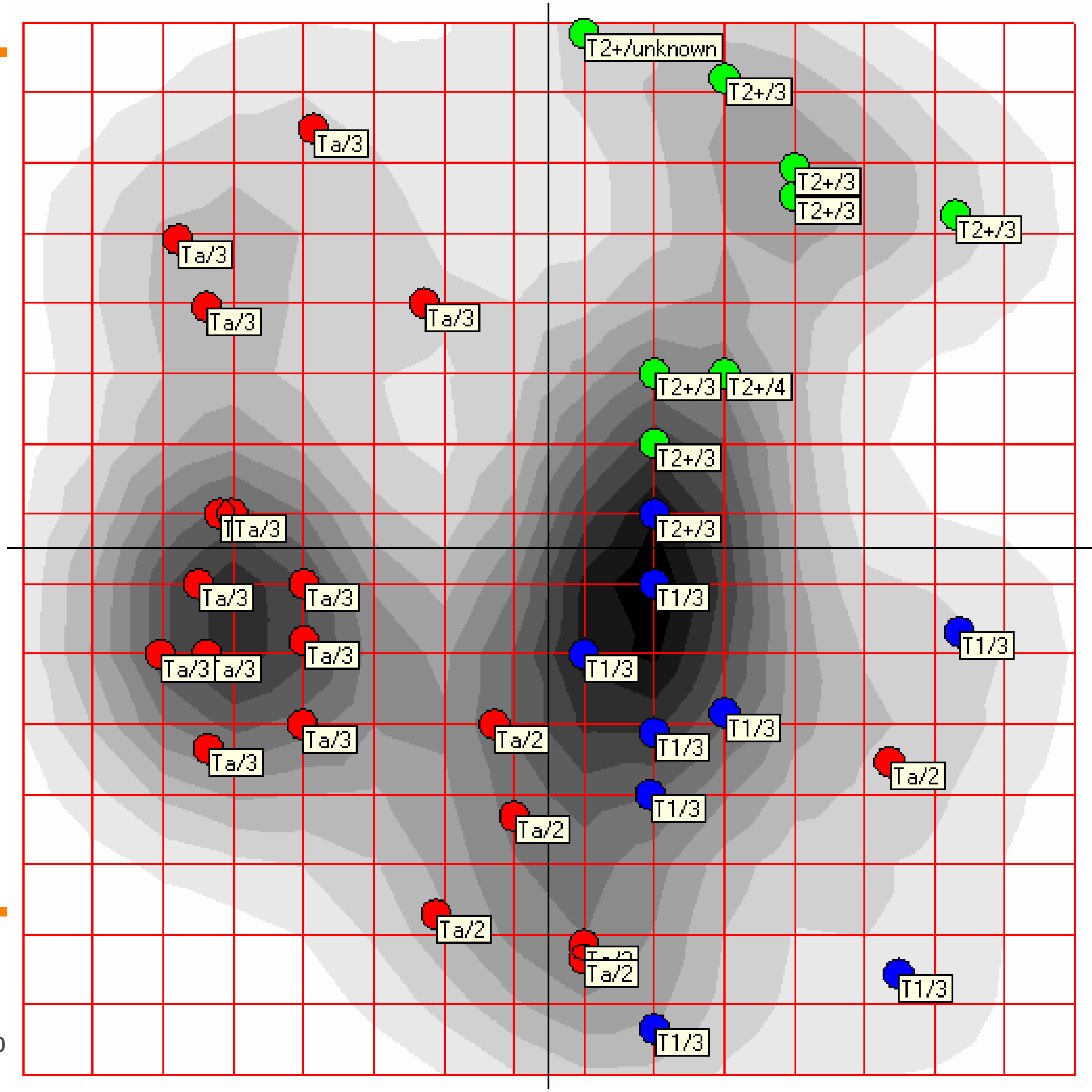


Gain in Mean Square Error ~ 30%



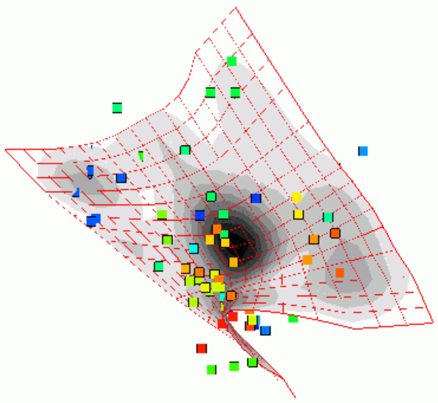


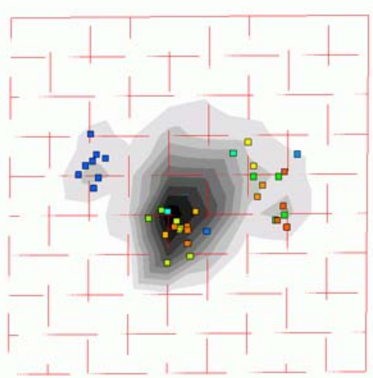
Bladder cancer dataset (Dataset II), 40 patients

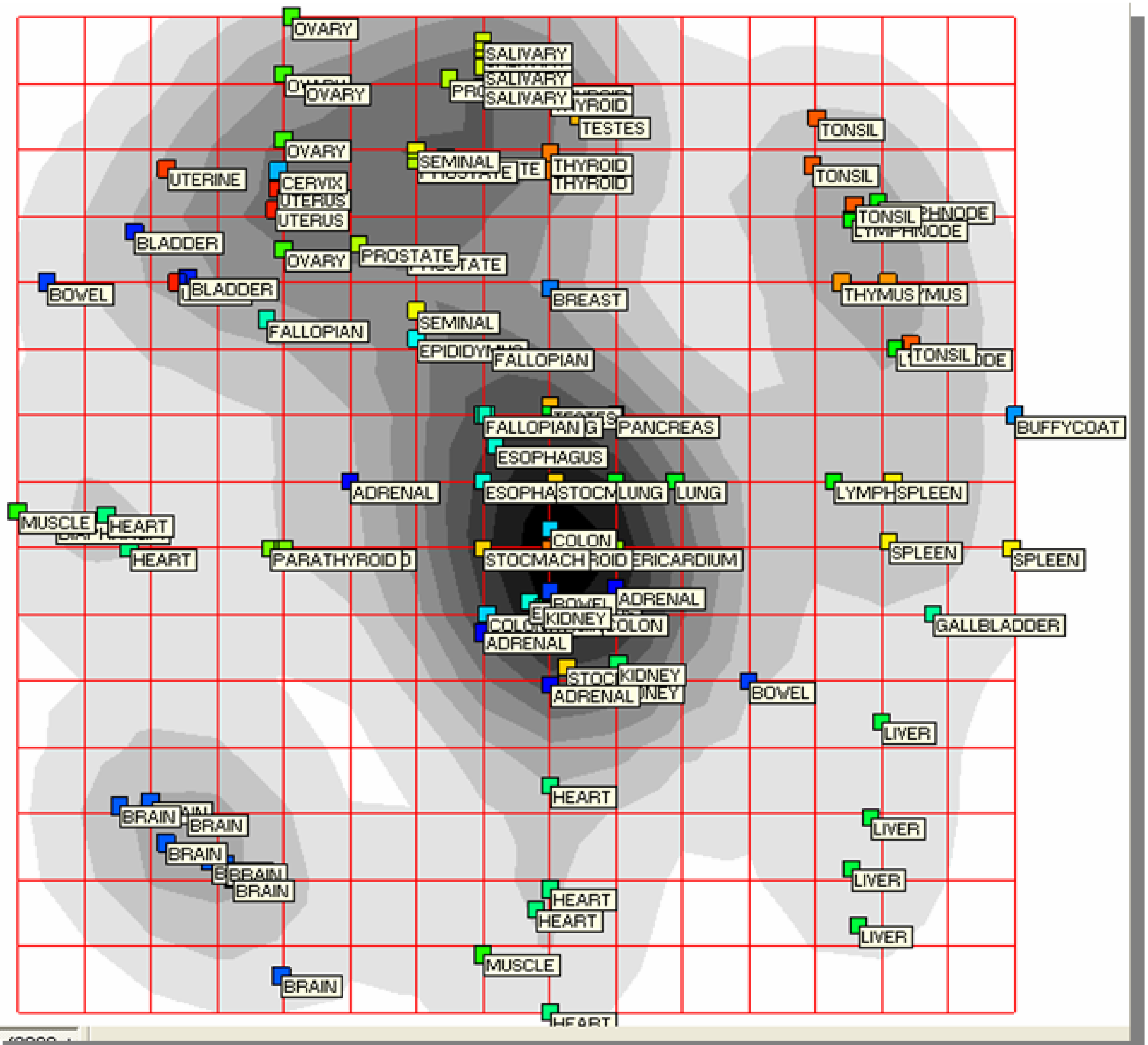


institutCurie

102 healthy tissues (Dataset III)

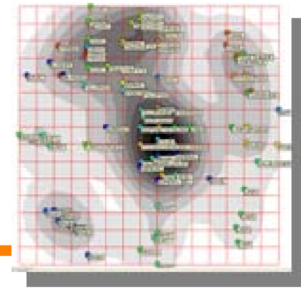


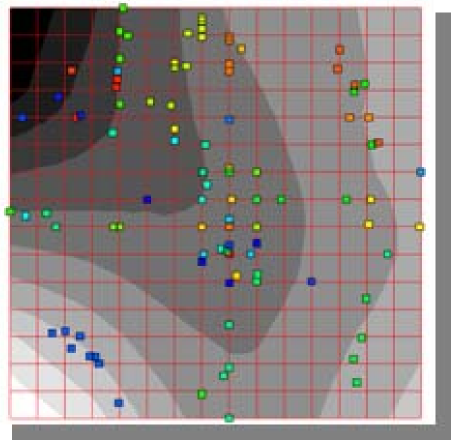




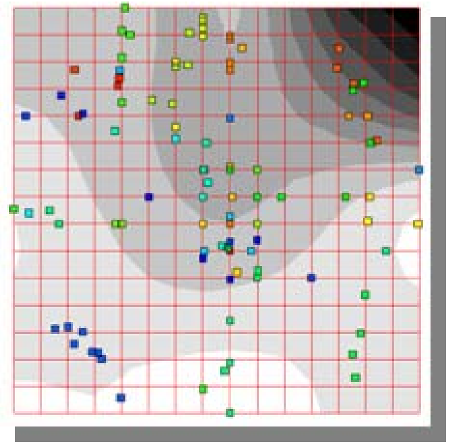
institutCurie

102 healthy tissues (Dataset III)

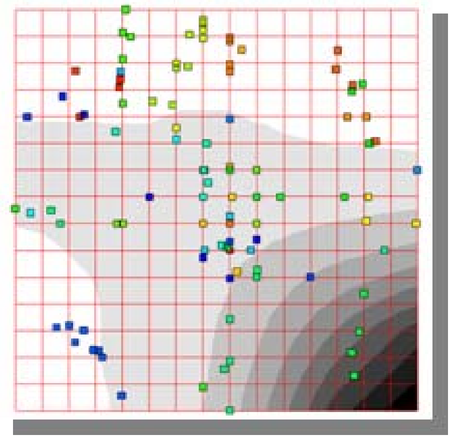




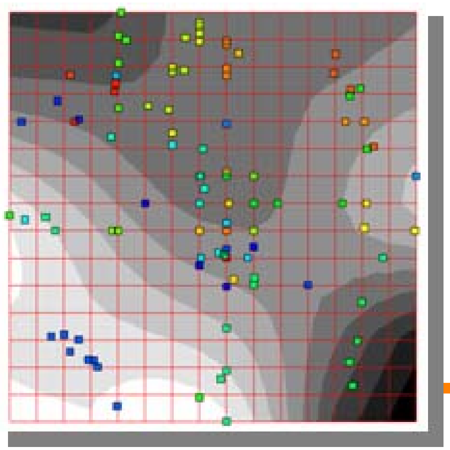
actin gamma 2



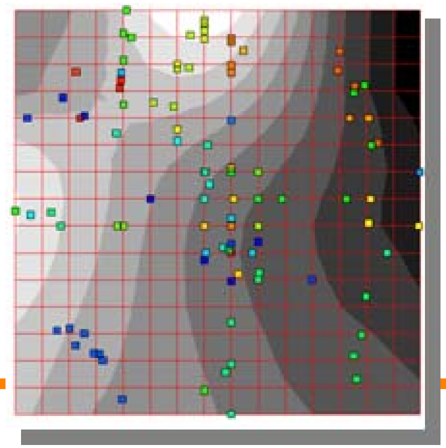
keratin 5



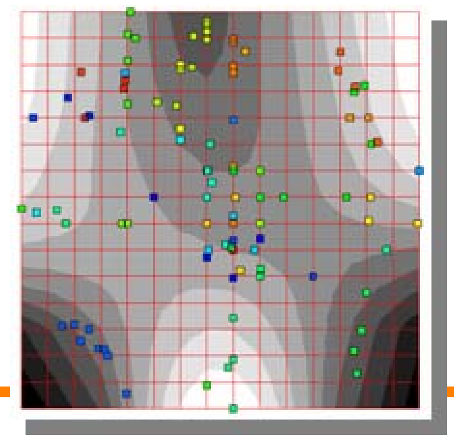
aldolase B



claudin 1



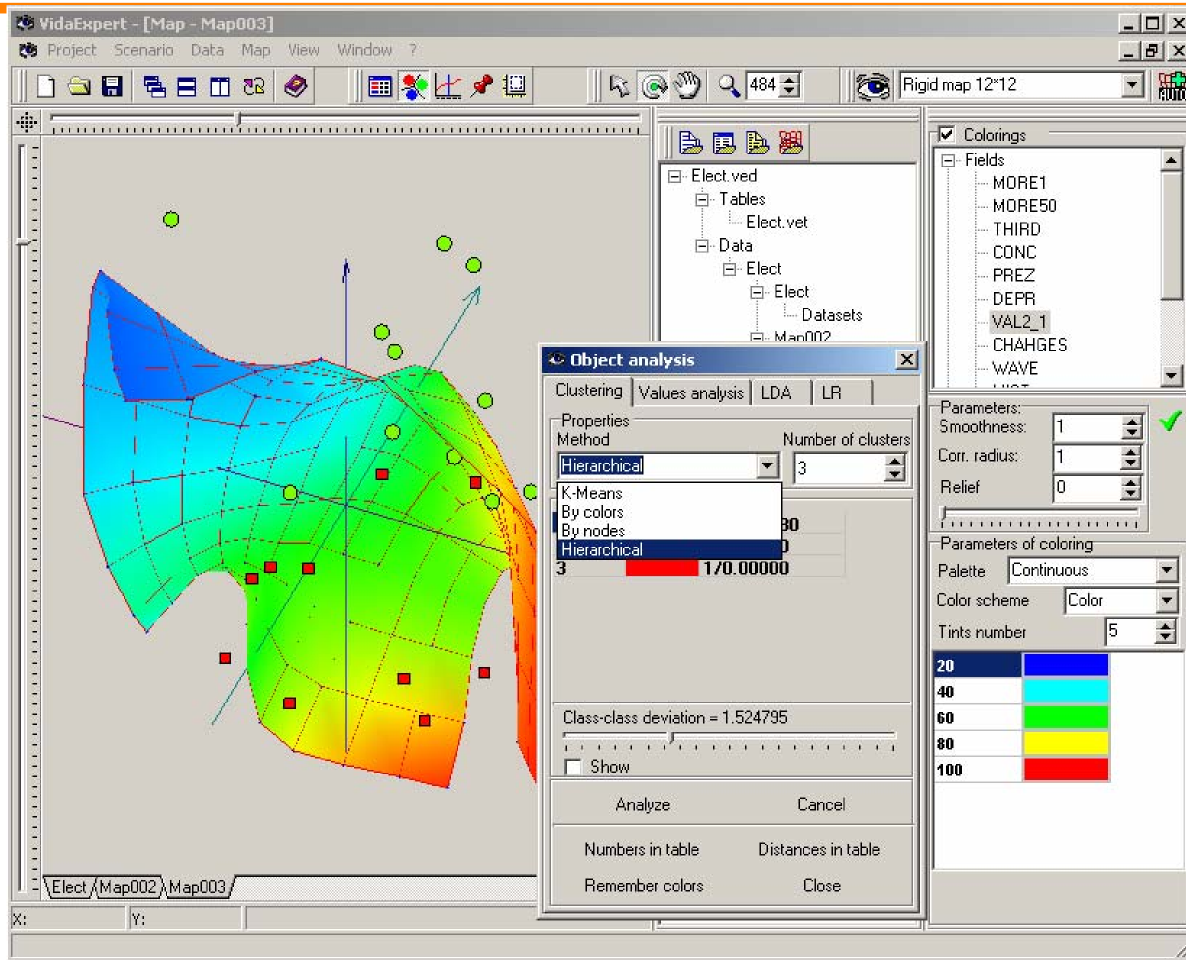
calgranulin B



ERBB3

institutCurie

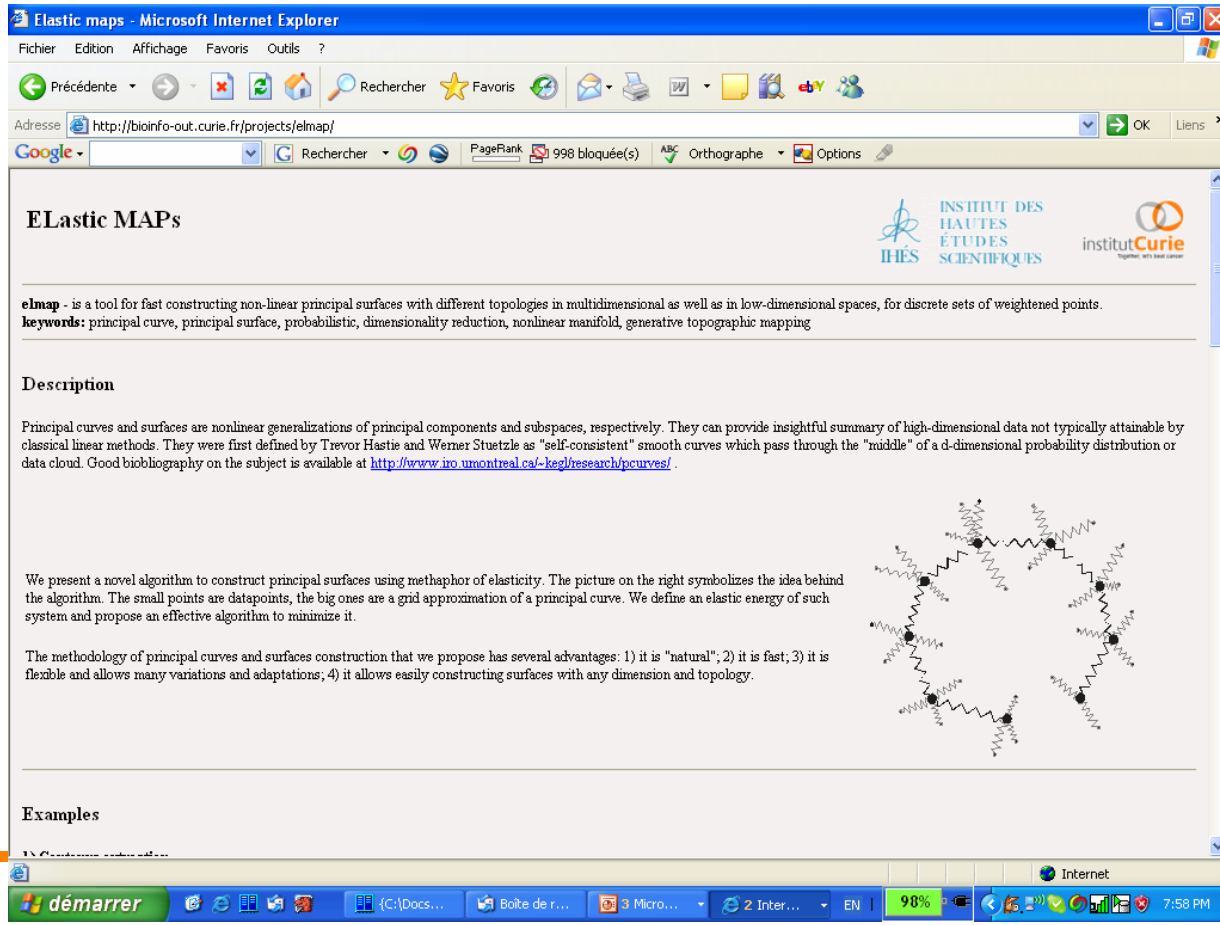
Implementation of the idea: VidaExpert tool



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

elmap C++ package

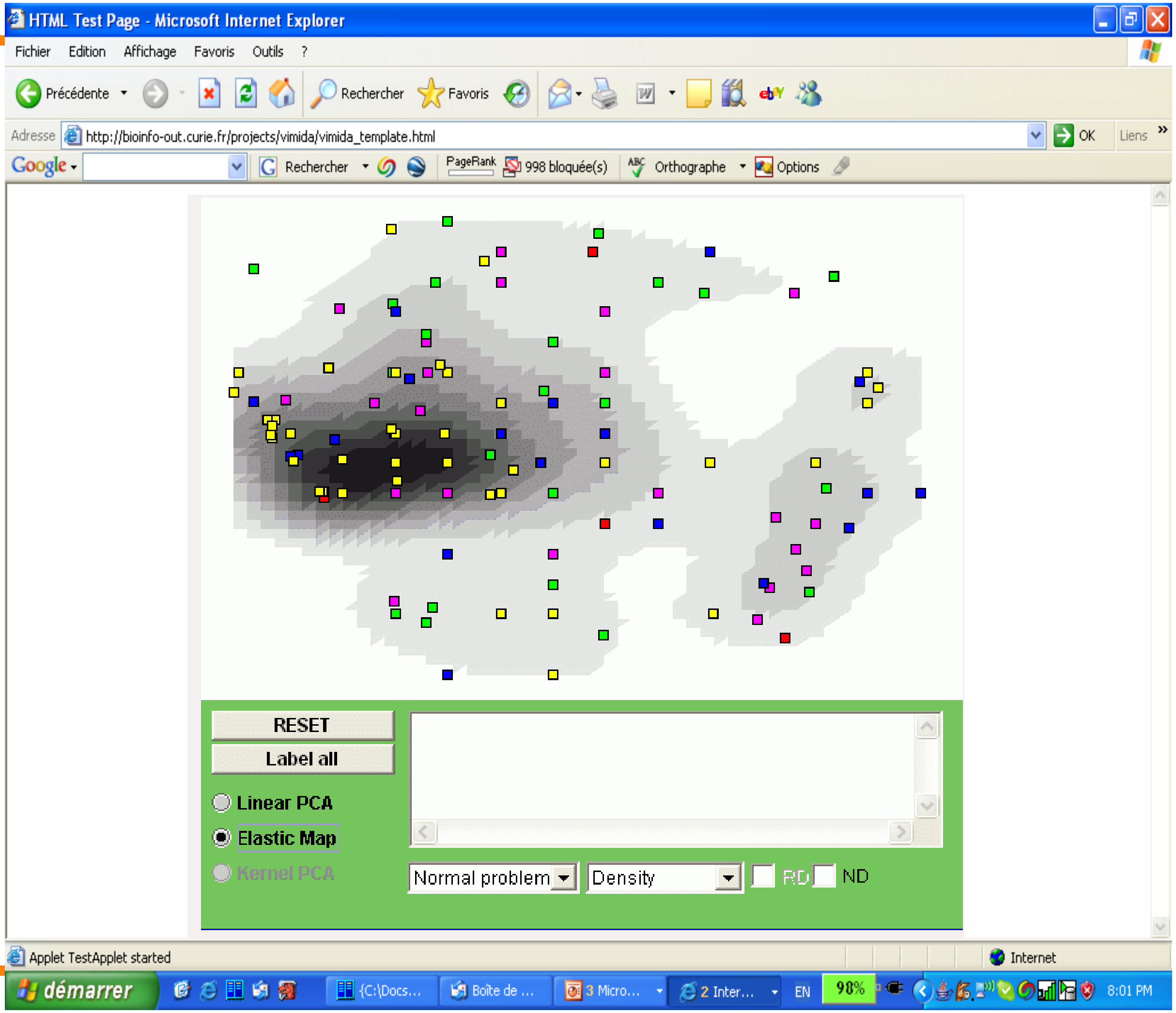
http://bioinfo.curie.fr/projects/elmap/



institutCurie

VIMIDA: Java-applet for multidimensional data visualization

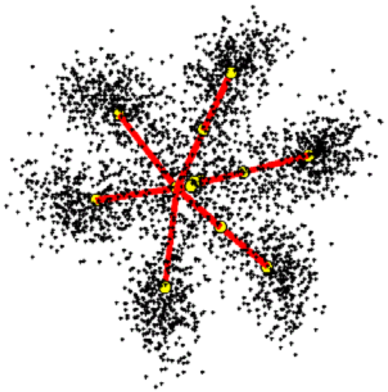
Part of PLATAN :
Microarray data analysis pipeline developed
In Institut Curie

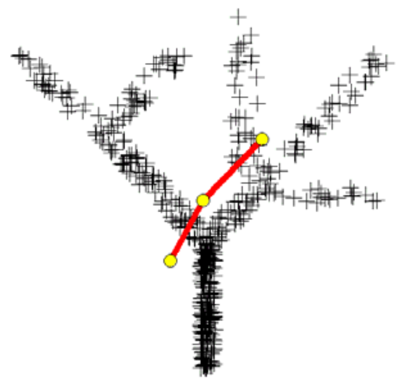


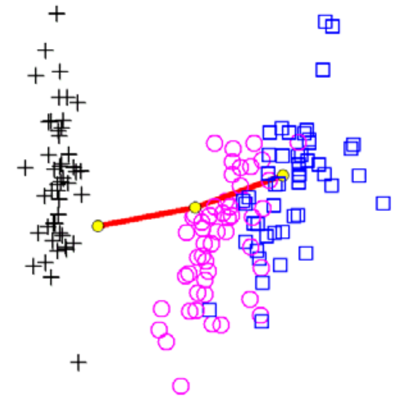
http://bioinfo.curie.fr/projects/vimida/

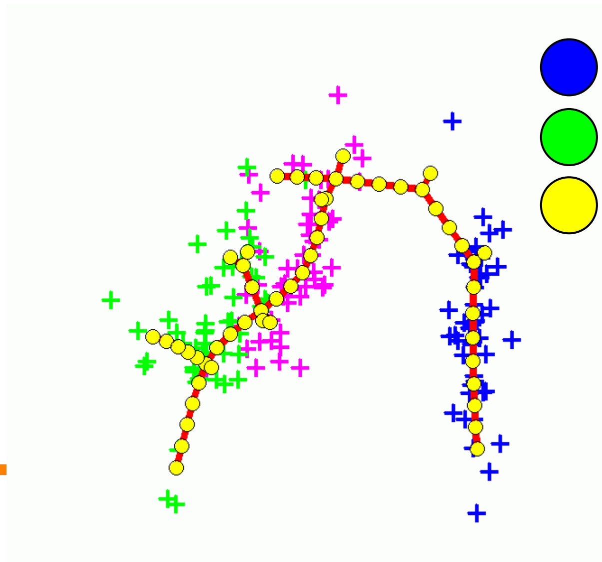
institutCurie

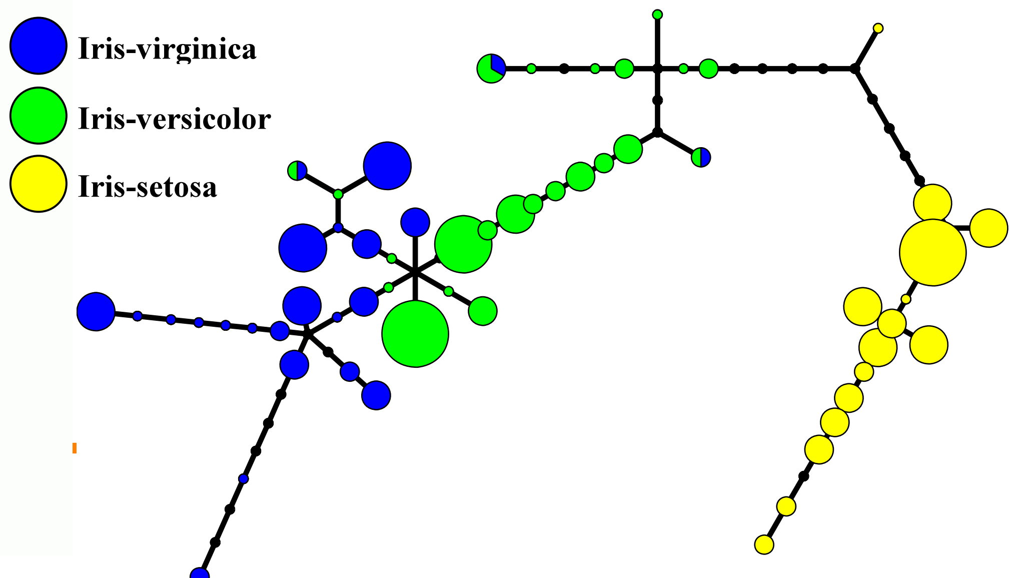
Anonce
Topological grammars: principal trees, cubic complexes, etc. in the talk of Professor Gorban (26 August)



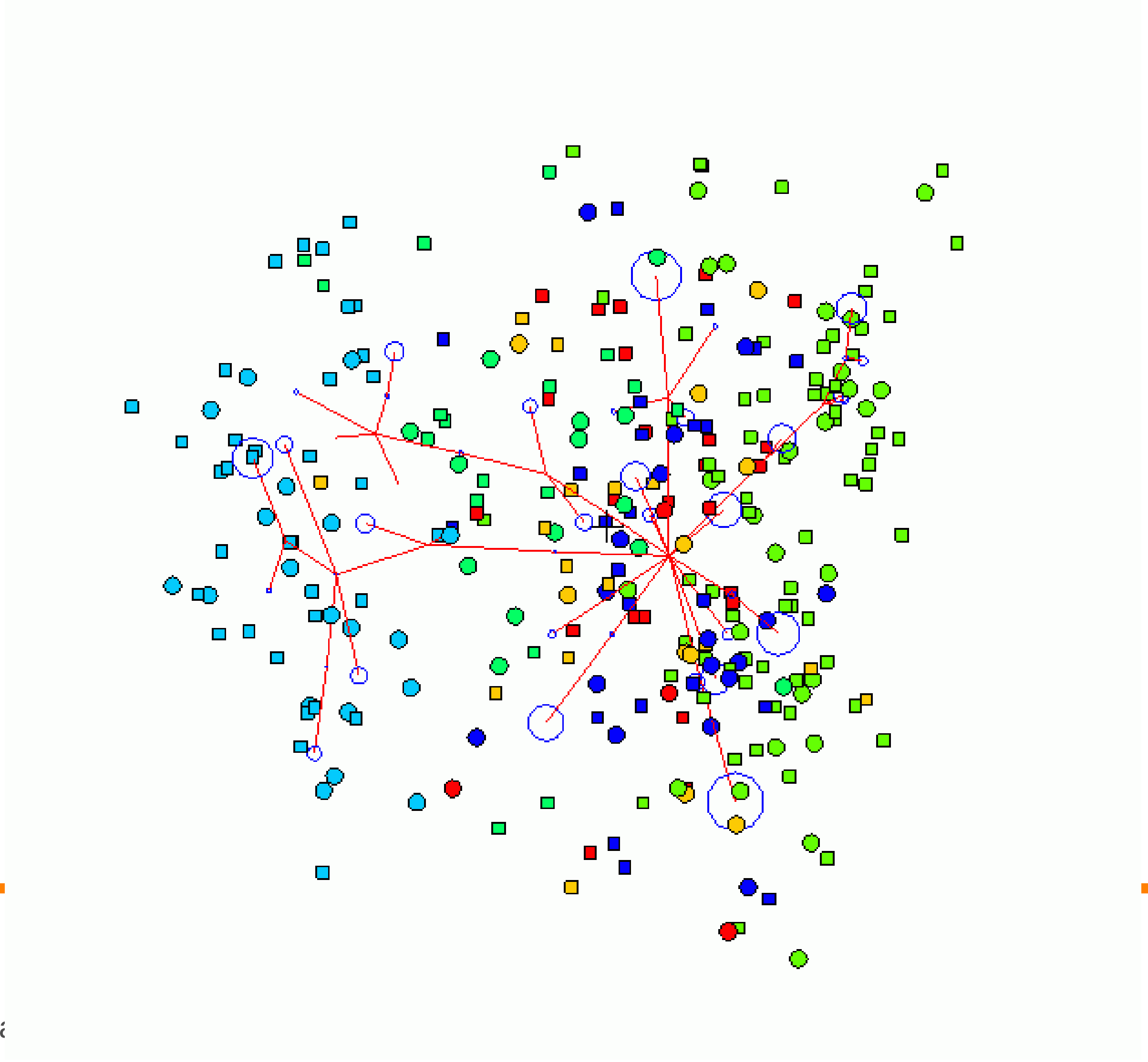








Branching principal components for bioinformatics data: alternative for hierarchical clustering?



institutCurie

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

institutCurie