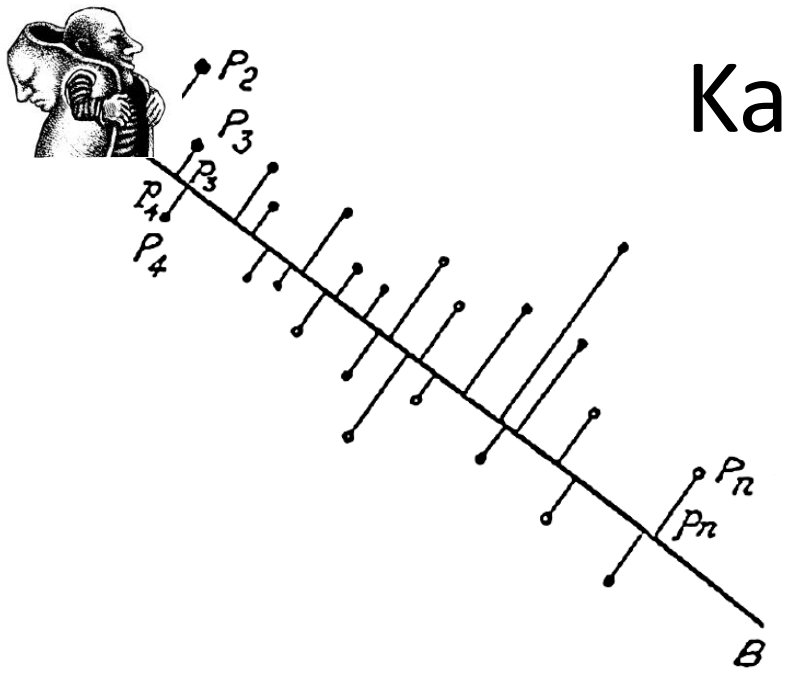
Geometry of Data Sets: Robust Topological Grammars for Unsupervised Learning

Alexander Gorban
University of Leicester, UK
with
Andrei Zinovyev
Institute Curie, Paris, France
and
Evgeny Mirkes
University of Leicester, UK

Plan

- The problem
- Approximation of multidimensional data by low-dimensional objects
- Elastic Manifolds
- Comparison with PCA
- Pluriharmonic graph embedment
- Topological grammars
- Robustness and trimmed springs
- Examples



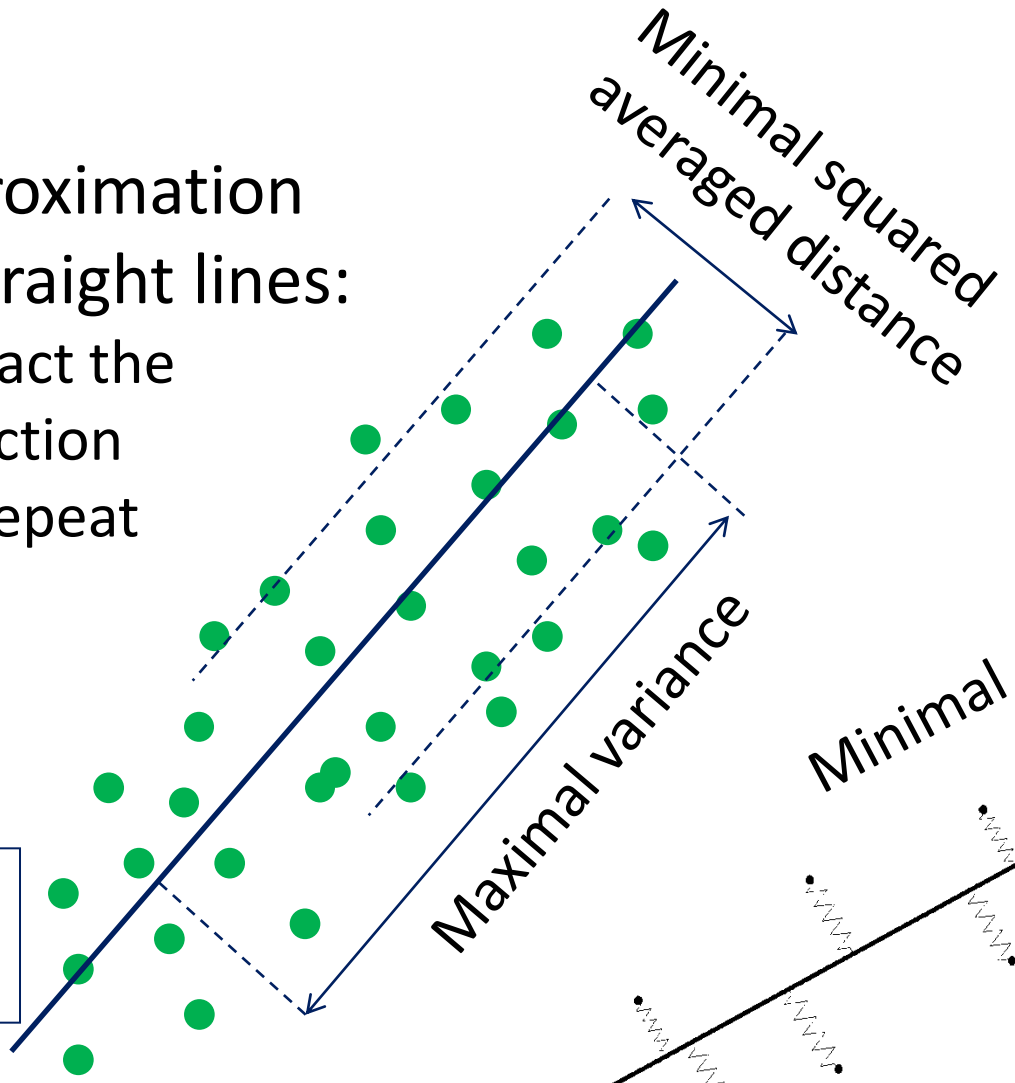
Karl Pearson
1901

LIII. *On Lines and Planes of Closest Fit to Systems of Points in Space.* *By* KARL PEARSON, *F.R.S., University College, London* *.

(1) IN many physical, statistical, and biological investigations it is desirable to represent a system of points in plane, three, or higher dimensioned space by the "best-fitting" straight line or plane. Analytically this

Principal Component Analysis

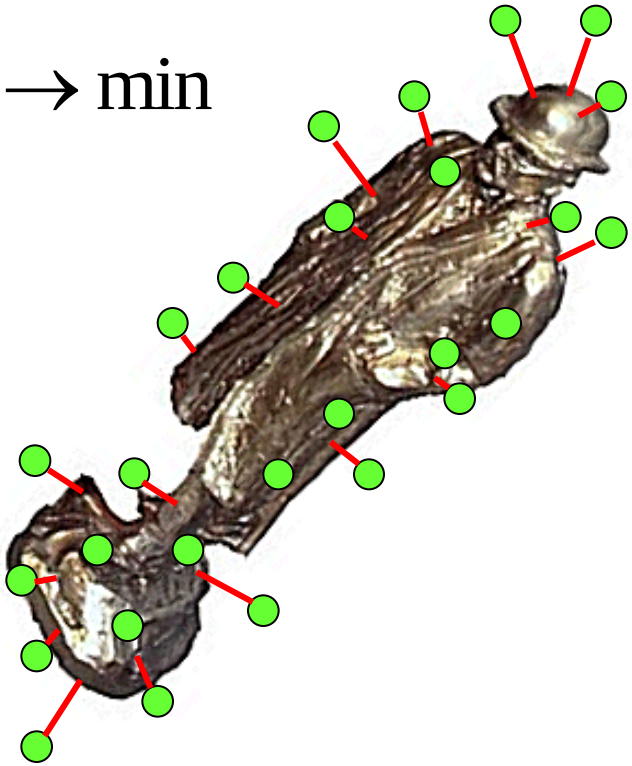
Approximation by straight lines:
Subtract the projection and repeat



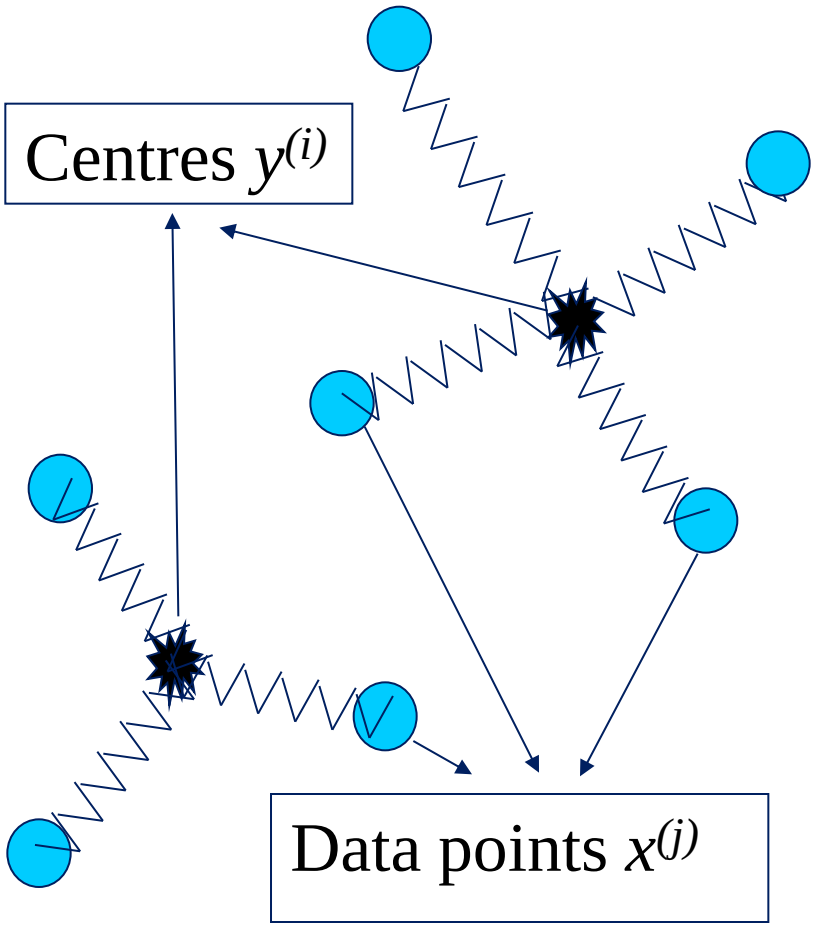
1st Principal axis

Principal “Object”

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



Principal points (K-means)

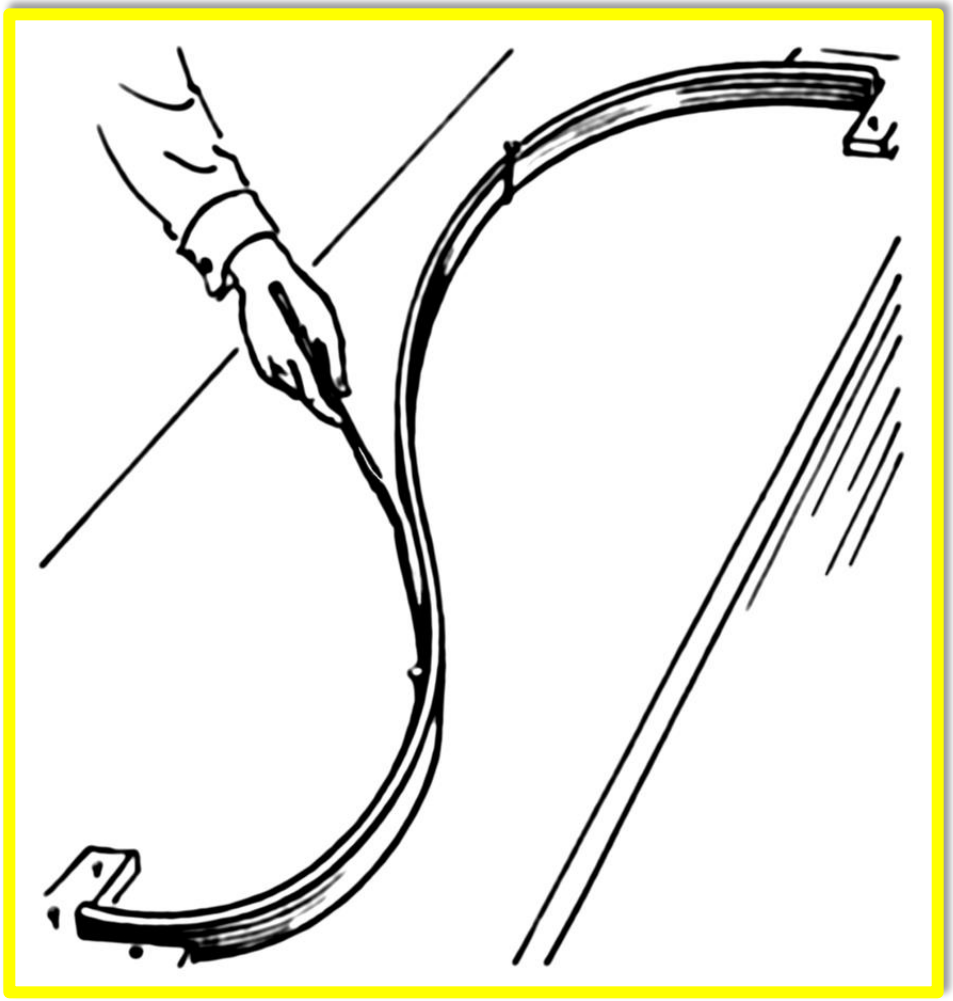


Approximation
by smaller finite sets:

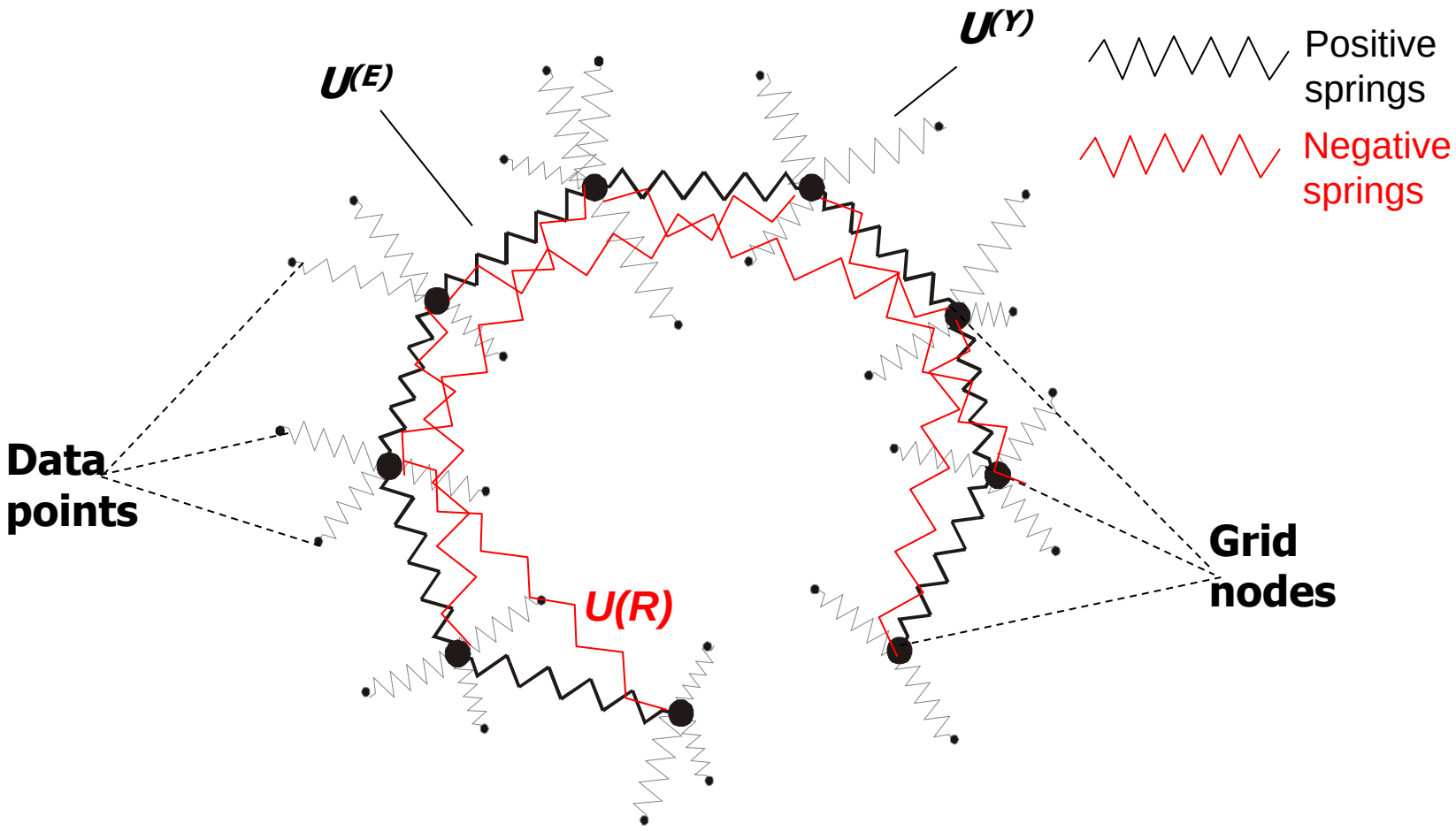
1. Select several centres;
2. Attach datapoints to the closest centres by springs;
3. Minimize energy;
4. Repeat 2&3 until converges.

Steinhaus, 1956;
Lloyd, 1957;
MacQueen, 1967

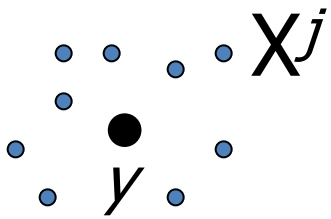
Definition of elastic energy:
we borrow this approach from splines



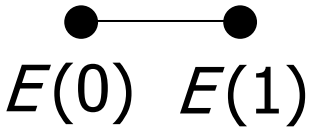
Metaphor of elasticity: elastic net



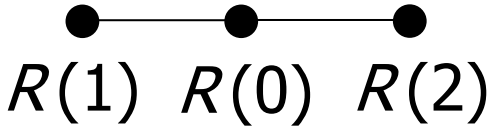
Definition of elastic energy



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



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

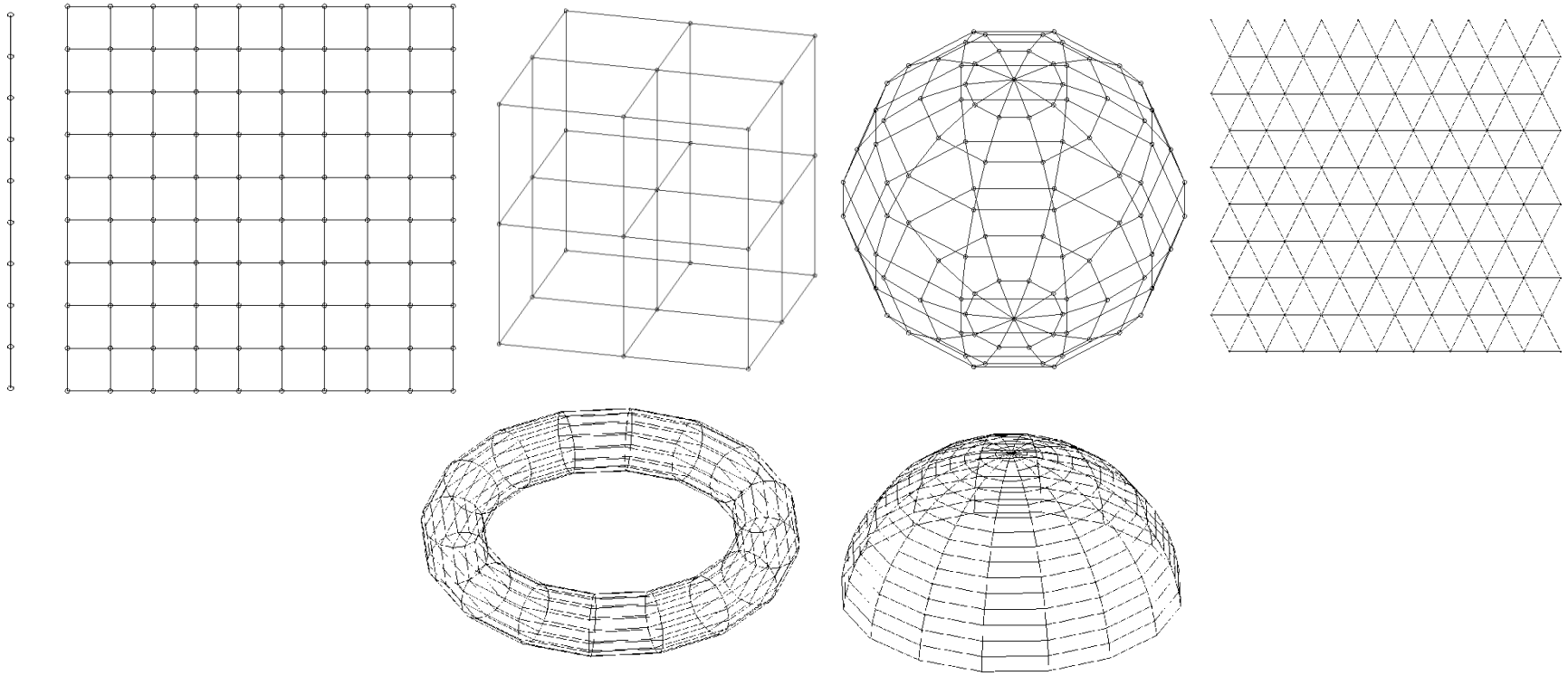


$$U^{(R)} = \sum_{i=1}^{r}\mu_i\left\|R^{(i)}(1) + R^{(i)}(2) - 2R^{(i)}(0)\right\|^2$$

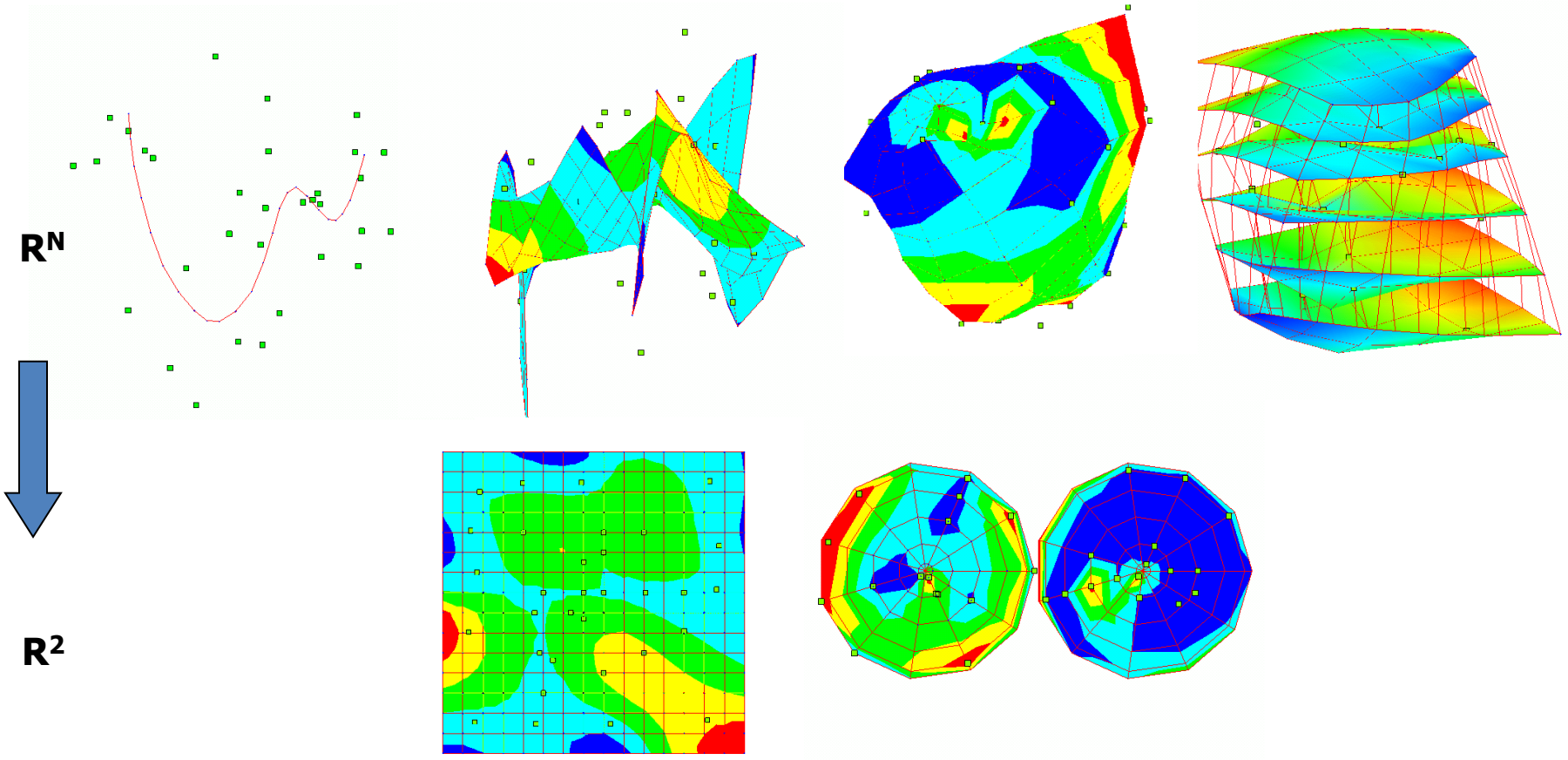
$$\lambda_i = \lambda_0, \quad \mu_i = \mu_0$$

$$U = U^{(Y)} + U^{(E)} + U^{(R)} \to \min$$

Assembling elastic nets



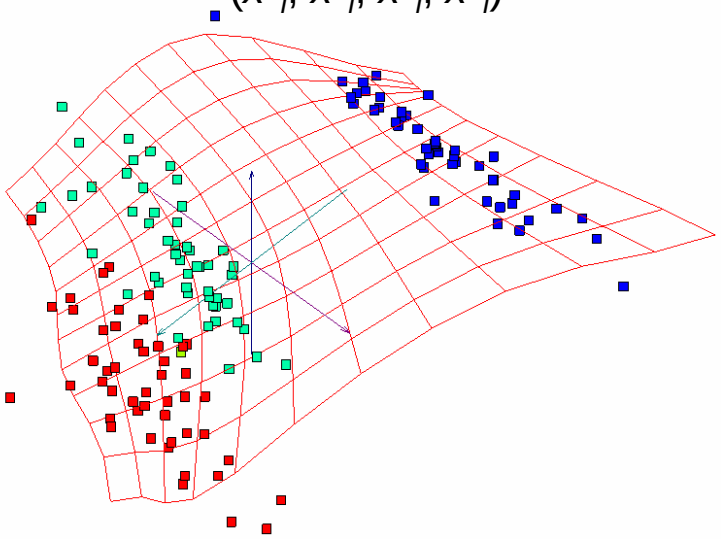
Various manifold topologies



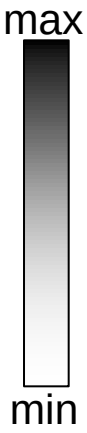
Colorings: visualize any multidimensional function

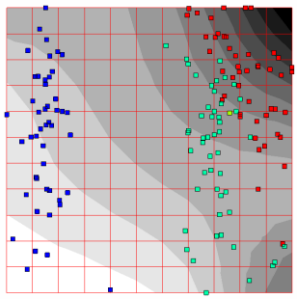
Four dimensional space

$(x^1_i, x^2_i, x^3_i, x^4_i)$

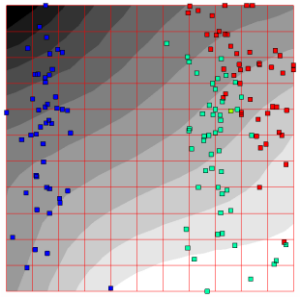


$F = F(x1,x2,x3,x4)$

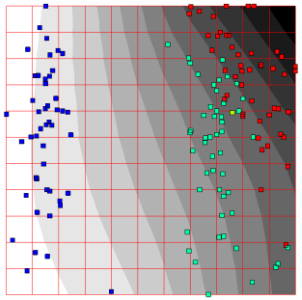




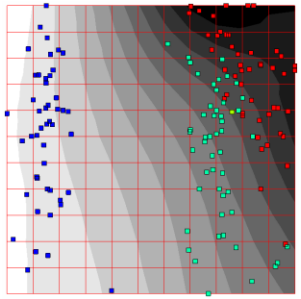
$F(x1,x2,x3,x4) = x1$



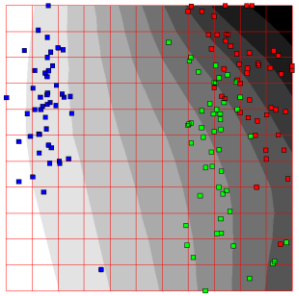
$F(x1,x2,x3,x4) = x2$



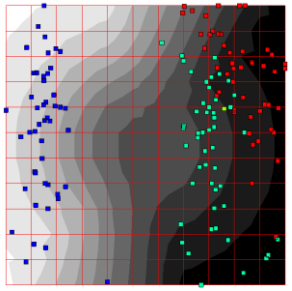
$F(x1,x2,x3,x4) = x3$



$F(x1,x2,x3,x4) = x4$

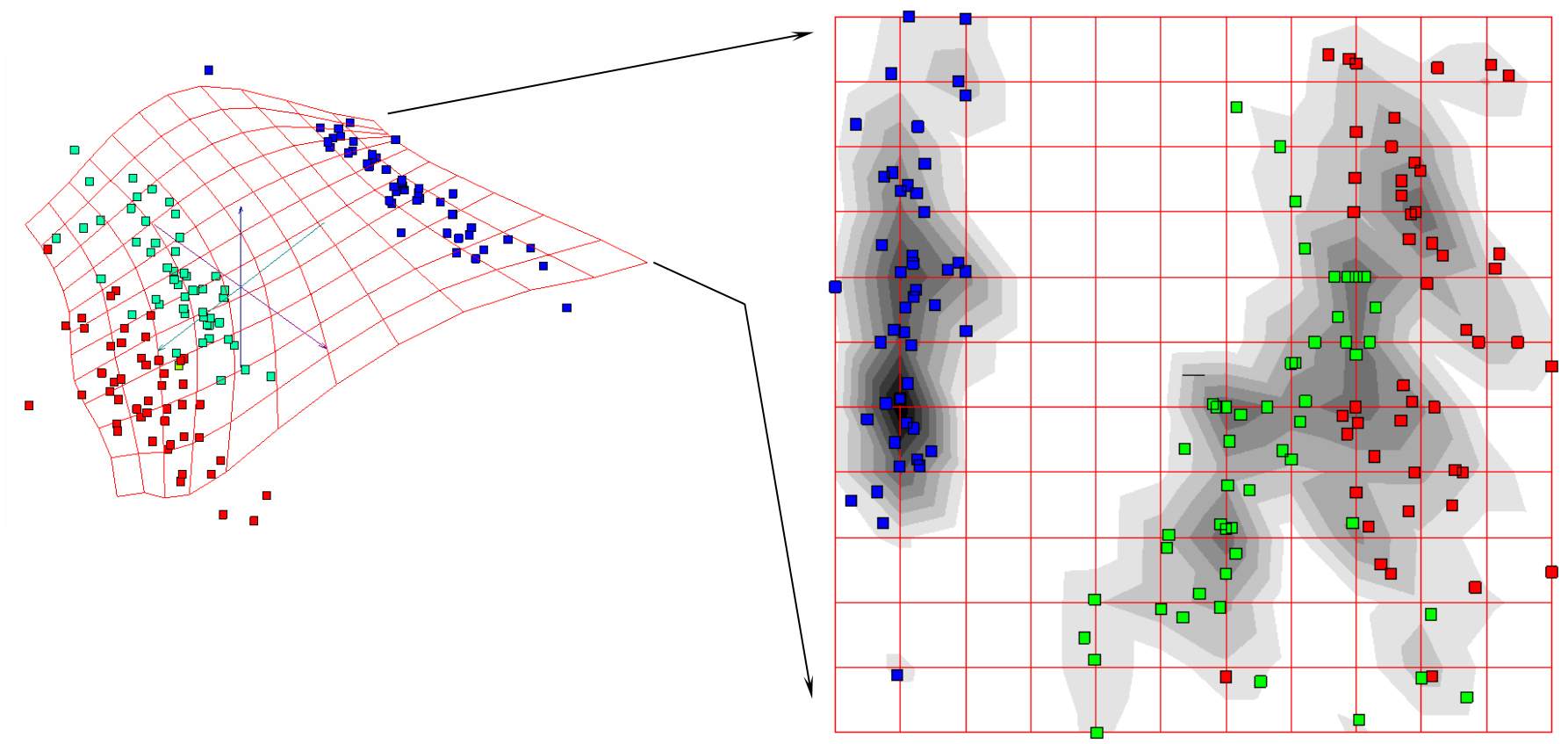


$F= -0.4x1+0.3x2-0.5x3-0.5x4$



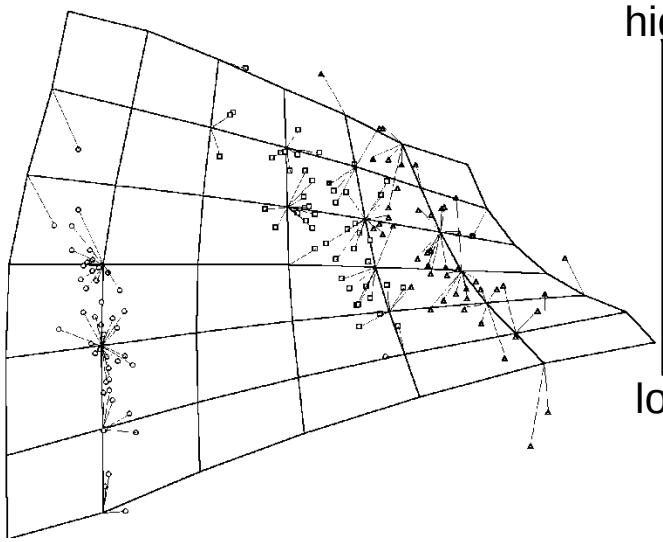
$F= x1^2+x2^2+x3^2+x4^2$

Example of complex function: point density



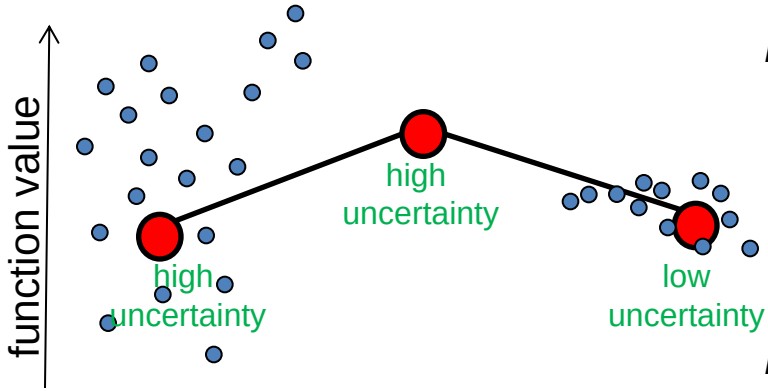
Visualization of *uncertainty*

kNN methodology



F(x1,x2,x3,x4) = x1

F(x1,x2,x3,x4) = x2

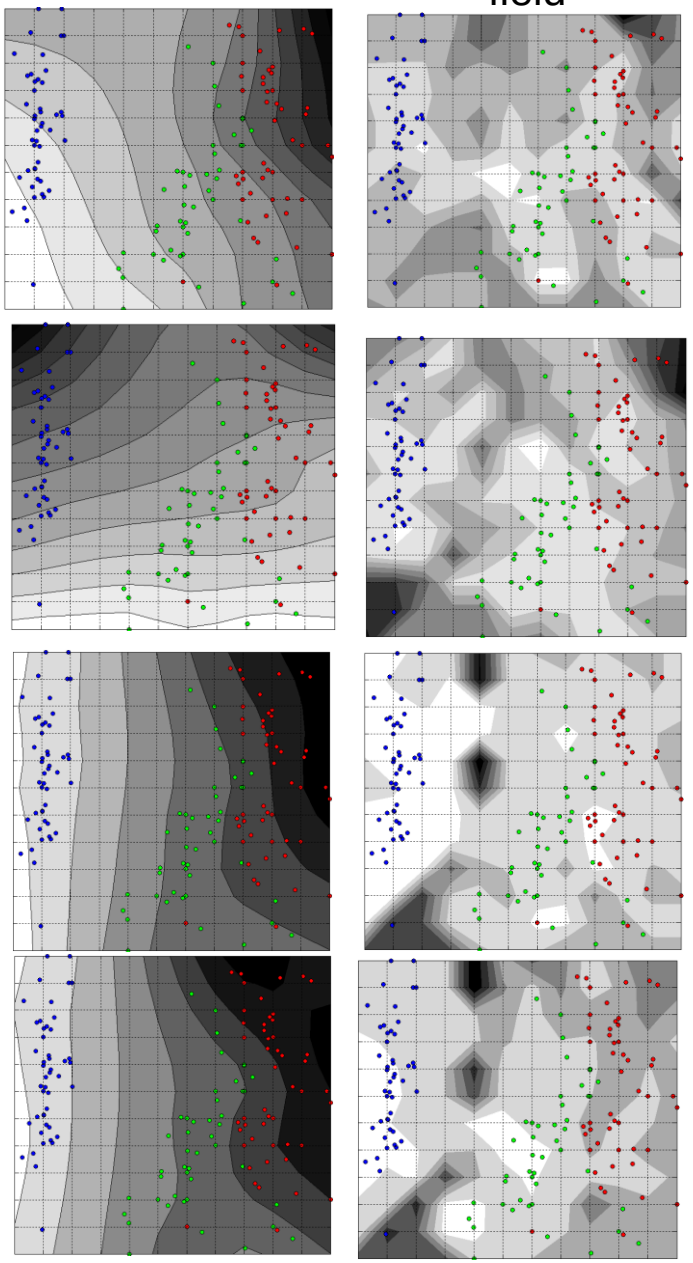


F(x1,x2,x3,x4) = x3

F(x1,x2,x3,x4) = x4

value

uncertainty field



Software

C++ **elmap** package

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

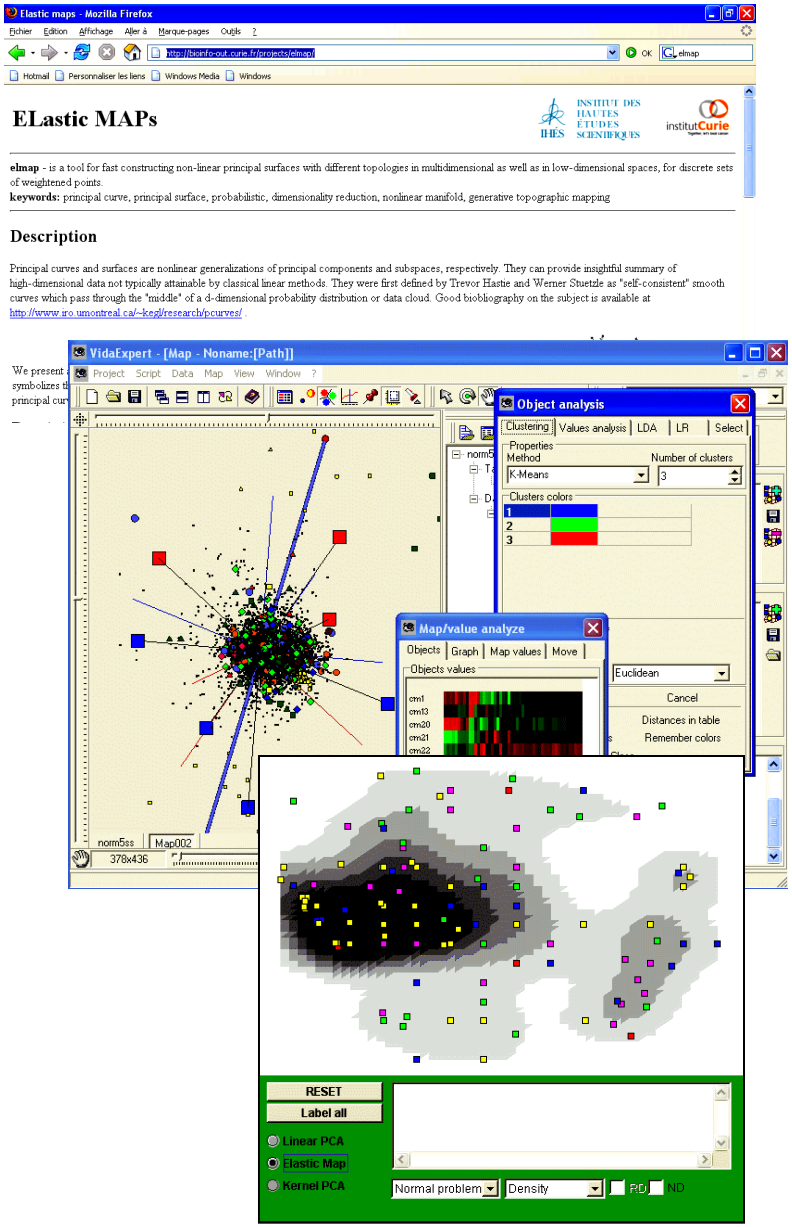
+ Java implementation on demand

VidaExpert end-user data visualization tool

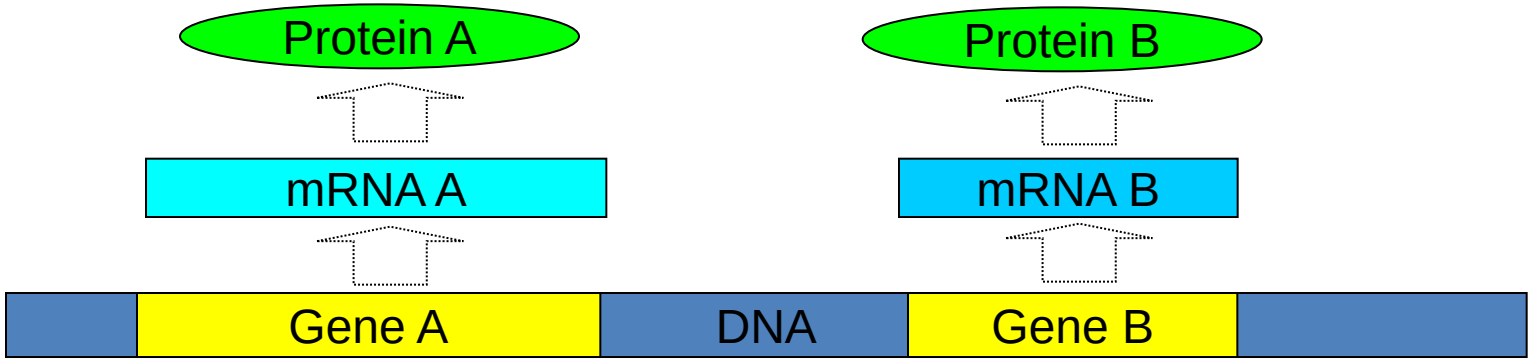
http://bioinfo-out.curie.fr/projects/vidaexpert

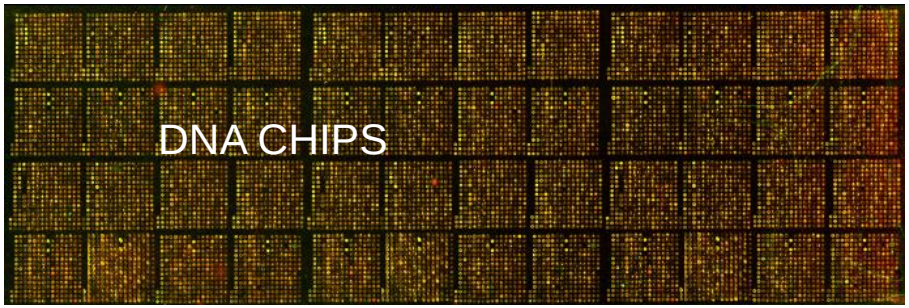
ViMiDa Java-applet

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



Microarray datasets,

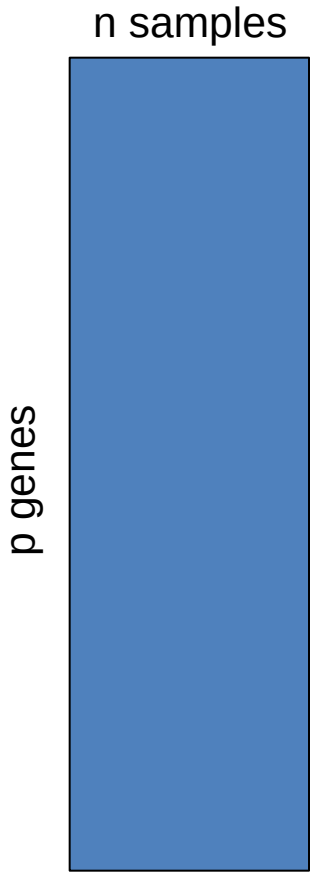




One spot corresponds
to a gene
(mRNA concentration)

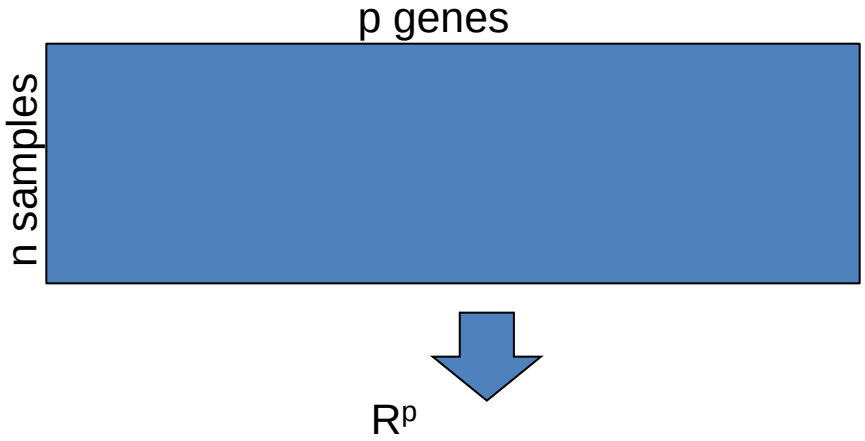
Table of numbers, characteristic size is 10000 genes x100 samples

Large p (~10000), small n (~100)



R^n

Studying relations between genes, their functions, etc.

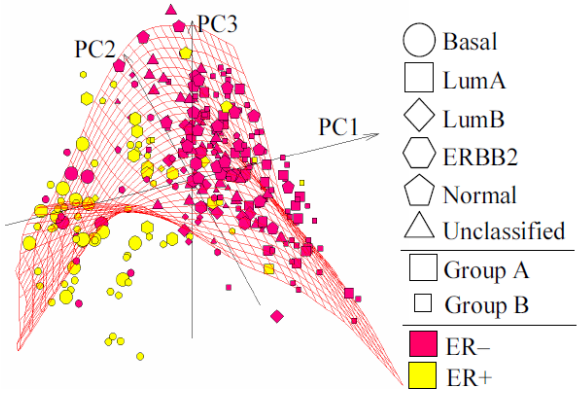


R^p

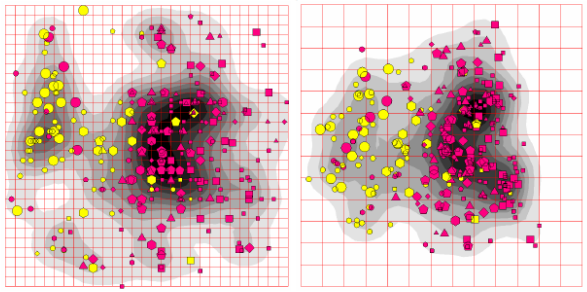
Classification problems, diagnosis, prognosis improvement

Overfitting danger!
Need for regularization!

Are 2D non-linear projections better than 2D linear projections?

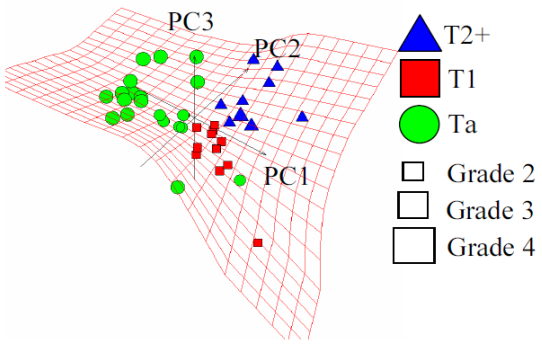


a)

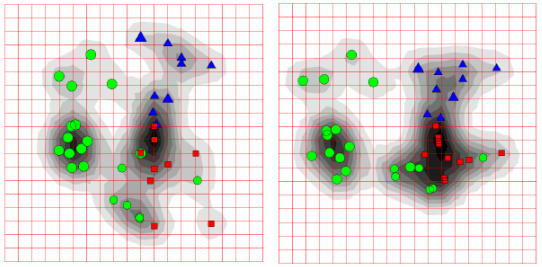


b) ELMAP2D c) PCA2D

Breast cancer
Wang et al., 2005

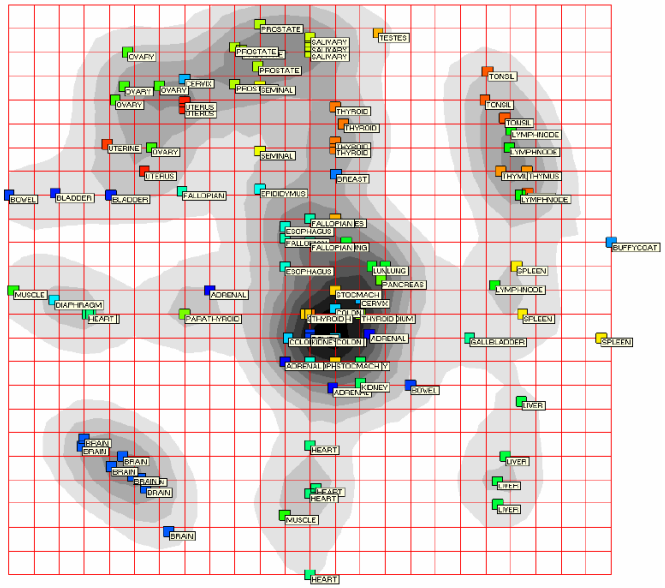


a)

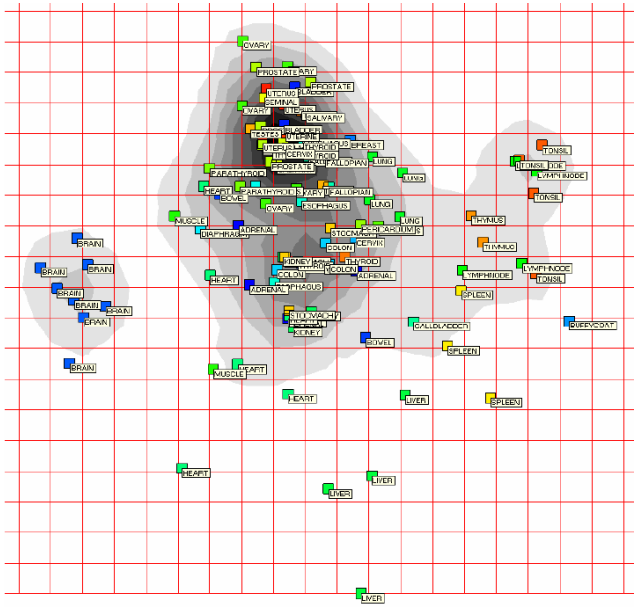


b) ELMAP2D c) PCA2D

Bladder cancer
Dyrskjot et al., 2003



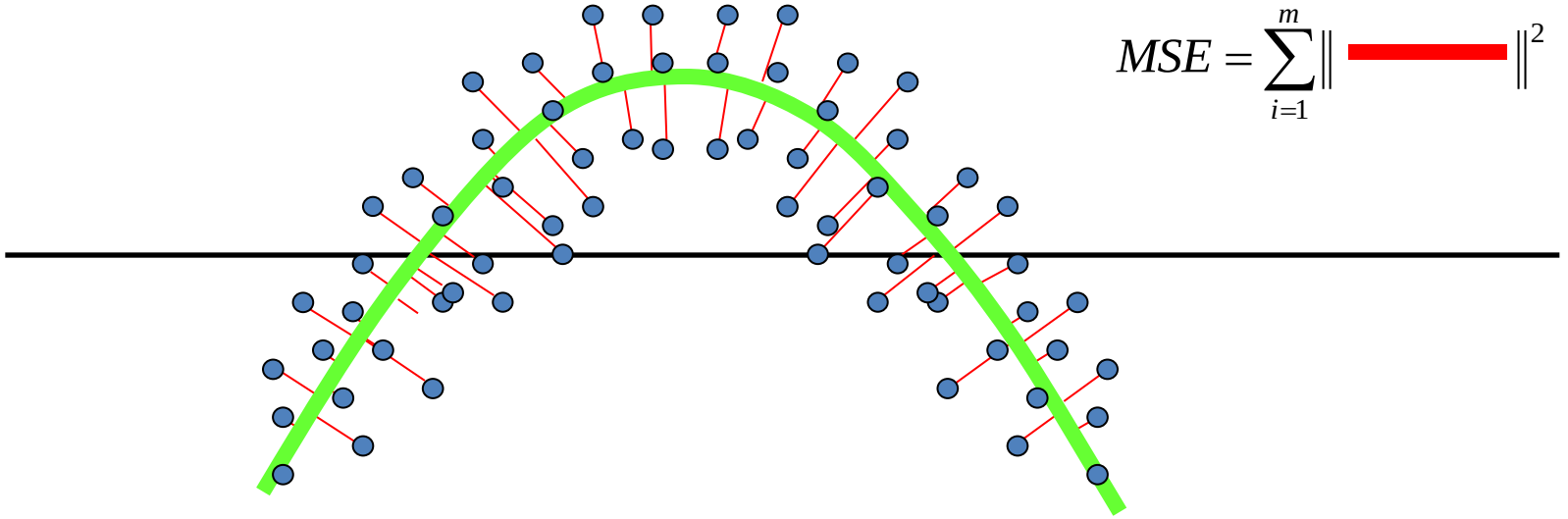
a) ELMAP2D



b) PCA2D

Human tissues, Shyamsundar et al., 2005

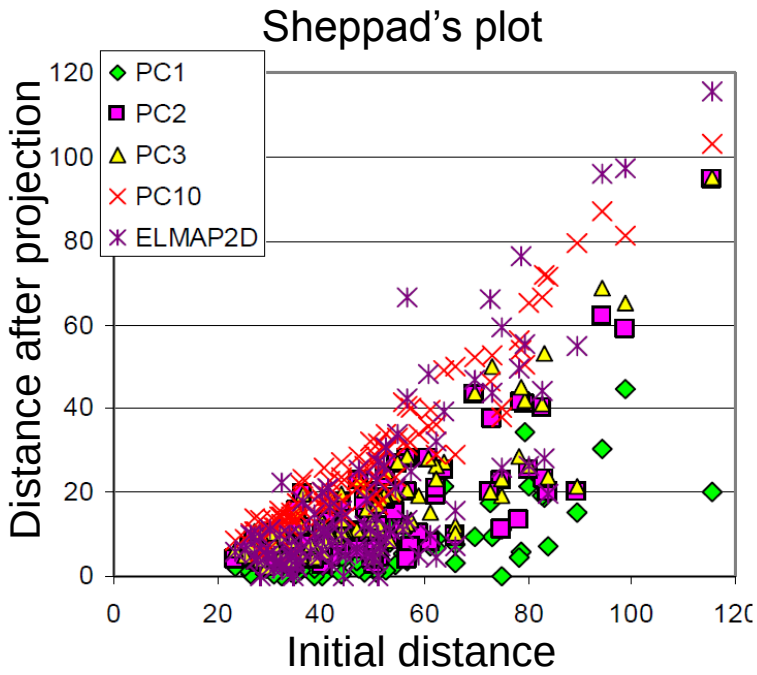
Yes: better approximation, smaller MSE (*as expected*)

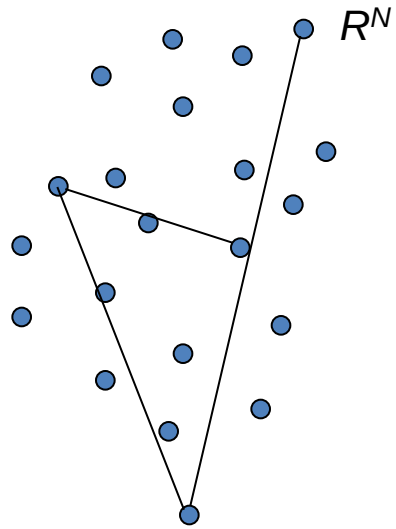


Dataset	ELMAP2D	PC1	PC2	PC3	PC4	PC5	PC10
Breast cancer MSE	48.1	52.6	50.7	49.3	48.4	47.6	45.3
Variation explained	-	7.9%	14.3%	19.0%	22.0%	24.6%	31.5%
Bladder cancer MSE	40.6	48.7	45.4	42.8	40.7	39.2	33.0
Variation explained	-	21.0%	31.4%	38.9%	44.8%	48.8%	63.8%
Normal tissues MSE	36.9	48.8	45.5	42.3	40.1	38.5	32.4
Variation explained	-	10.7%	19.1%	26.0%	30.3%	32.2%	40.7%

7)

Yes: better representation of large distances (*already less trivial*)





K points,
K(K-1)/2 pairwise distances.

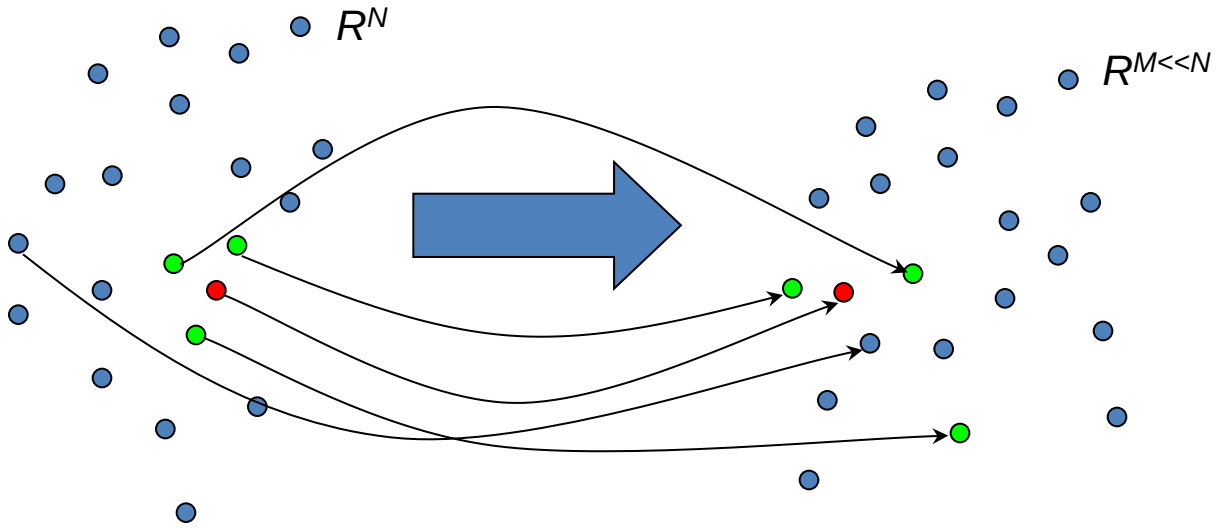
Natural PCA:
Select K most representative pairwise distances

Dataset/method	ELMAP2D	PC1	PC2	PC3	PC4	PC5	PC10
Breast cancer/Pearson	0.60	0.40	0.52	0.61	0.65	0.69	0.75
Breast cancer/Spearman	0.40	0.19	0.32	0.36	0.42	0.49	0.56
Bladder cancer/Pearson	0.87	0.82	0.84	0.88	0.89	0.91	0.96
Bladder cancer/Spearman	0.68	0.57	0.60	0.70	0.70	0.75	0.90
Normal tissues/Pearson	0.80	0.68	0.78	0.82	0.86	0.87	0.95
Normal tissues/Spearman	0.68	0.56	0.69	0.79	0.84	0.86	0.94

Quality of distance mapping (QDM) is a correlation coefficient between the pair-wise distances before and after projection onto the manifold:

(Gorban & Zinovyev, *IJNS*, 2010)

Yes: better point *entourage* preservation (*not necessarily expected*)



Dataset	ELMAP2D	PC1	PC2	PC3	PC4	PC5	PC10	RANDOM
Breast cancer (k=10)	0.26	0.13	0.20	0.28	0.31	0.38	0.47	0.04 ± 0.06
Bladder cancer (k=5)	0.53	0.34	0.53	0.61	0.64	0.70	0.80	0.12 ± 0.14
Normal tissues (k=5)	0.49	0.23	0.33	0.43	0.50	0.54	0.69	0.05 ± 0.09

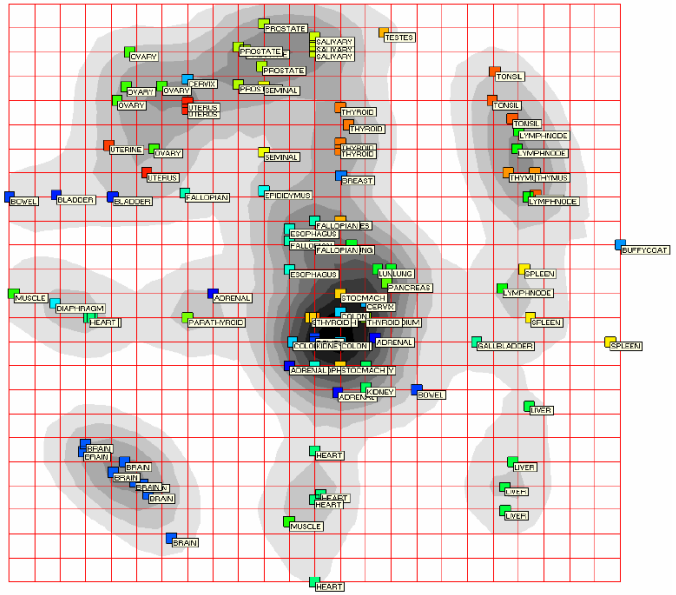
(Gorban & Zinovyev, *IJNS*, 2010)

Quality of point neighborhood preservation(QNP).

For every data point *I* we calculate the size of the intersection of the set of *k* neighbours calculated for embedding in the multi-dimensional space *S*(*i*; *k*) and in the low-dimensional space *Ś*(*i*; *k*).

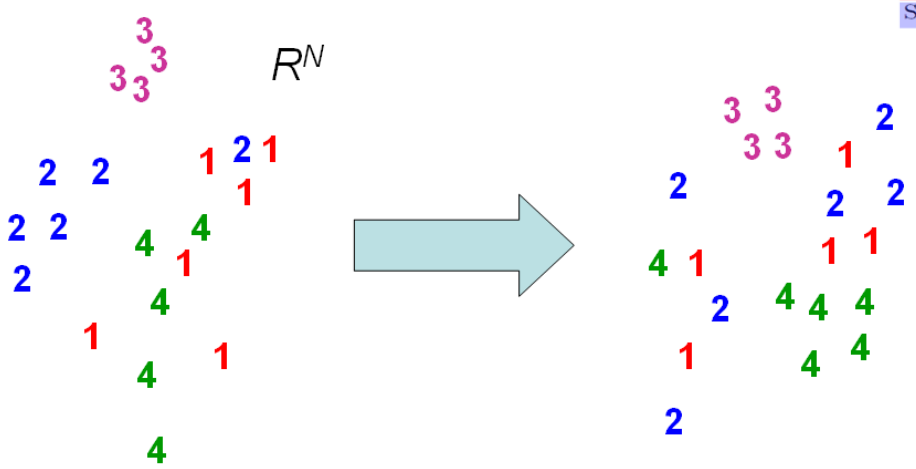
$$\mathrm{QNP}_k = 1/k \sum_{i=1\ldots N} |S(i;k) \cap \hat{S}(i;k)|/N.$$

Yes: better class compactness (*not a trivial property*)



a) ELMAP2D

	Class	ND	ELMAP2D	PC1	PC2	PC3	PC4	PC5	PC10	RANDOM	
	Breast cancer (GROUP), k=5										
	A (193 samples)	0.73	0.67	0.71	0.67	0.68	0.69	0.70	0.70	0.67±0.27	1
	B (93 samples)	0.31	0.30	0.33	0.29	0.35	0.37	0.38	0.31	0.33±0.21	1
	Breast cancer (TYPE), k=3										
-	lumA (95 samples)	0.61	0.60	0.64	0.65	0.67	0.72	0.71	0.74	0.33±0.27	1
++	basal (55 samples)	0.88	0.93	0.83	0.86	0.88	0.90	0.88	0.89	0.19±0.17	3
-	unclas. (42 samples)	0.28	0.20	0.27	0.25	0.27	0.27	0.30	0.29	0.15±0.16	1
+	normal (35 samples)	0.68	0.28	0.14	0.19	0.31	0.41	0.41	0.55	0.12±0.19	2
++	errb2 (34 samples)	0.62	0.36	0.24	0.34	0.32	0.32	0.43	0.59	0.12±0.19	2
-	lumB (25 samples)	0.21	0.08	0.20	0.20	0.21	0.25	0.23	0.36	0.09±0.17	1
	Bladder cancer (TYPE), k=3										
	Ta (20 samples)	0.85	0.83	0.80	0.85	0.85	0.85	0.85	0.85	0.50±0.29	3
	T1 (11 samples)	0.58	0.67	0.45	0.69	0.67	0.70	0.67	0.63	0.27±0.26	4
++	T2+ (9 samples)	0.85	0.89	0.11	0.85	0.81	0.89	0.78	0.81	0.22±0.24	4
	Bladder cancer (GRADE), k=3										
++	Grade 3 (32 samples)	0.94	0.92	0.86	0.83	0.90	0.95	0.94	0.94	0.8±0.23	3
+	Grade 2 (6 samples)	0.5	0.61	0.22	0.39	0.67	0.67	0.67	0.61	0.15±0.23	4
	Normal tissues (TISSUE), k=2										
	BRAIN (8)	1.0	1.0	1.0	0.94	1.0	1.0	1.0	1.0	0.08±0.19	3
+++	HEART (6)	0.92	0.50	0.17	0.25	0.25	0.25	0.33	0.58	0.06±0.16	2
+++	THYROID (6)	0.92	0.67	0.17	0.17	0.25	0.17	0.67	0.75	0.06±0.17	3
+++	PROSTATE (5)	1.0	0.8	0.0	0.0	0.1	0.4	0.5	1.0	0.05±0.16	3
+++	OVARY (5)	0.7	0.6	0.2	0.1	0.3	0.4	0.5	0.7	0.05±0.15	3
-	LUNG (4)	0.88	0.40	0.5	0.63	0.75	0.63	0.63	0.75	0.04±0.13	2
	ADRENAL (4)	0.625	0.0	0.0	0.0	0.0	0.25	0.5	0.625	0.04±0.14	2
++	LIVER (4)	0.88	0.88	0.13	0.75	0.75	0.88	0.88	0.88	0.04±0.14	3
+++	SALIVARY (4)	1.0	1.0	0.25	0.25	0.50	0.88	1.0	1.0	0.04±0.14	3
++	TONSIL (4)	0.88	0.63	0.0	0.13	0.34	0.75	0.75	0.88	0.04±0.14	3
	LYMPHNODE (4)	0.75	0.25	0.25	0.25	0.0	0.25	0.25	0.25	0.04±0.14	2
+++	FALLOPIAN (4)	0.88	0.50	0.0	0.13	0.25	0.25	0.25	0.75	0.04±0.14	3
+	SPLEEN (3)	0.67	0.50	0.50	0.33	0.33	0.89	0.67	0.83	0.03±0.12	3
+	KIDNEY (3)	1.0	0.33	0.0	0.0	0.33	0.33	0.33	0.67	0.03±0.12	2
	UTERUS (3)	0.83	0.33	0.0	0.33	0.33	0.33	0.33	0.83	0.03±0.12	2
	BOWEL (3)	0.17	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.03±0.12	1
	ESOPHAGUS (3)	0.67	0.0	0.0	0.0	0.0	0.0	0.0	0.67	0.03±0.12	2
	COLON (3)	0.33	0.0	0.17	0.0	0.0	0.17	0.17	0.17	0.03±0.12	1
	STOCMACH (3)	0.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.03±0.12	2



"-" – non-linear is worse
"+" – non-linear is better
"++", "+++" – non-linear is MUCH better

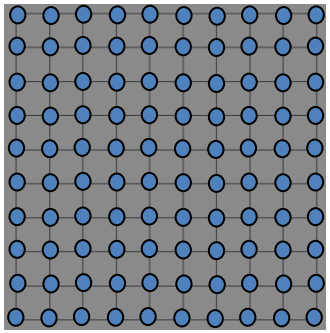
(Gorban & Zinovyev, *IJNS*, 2010)

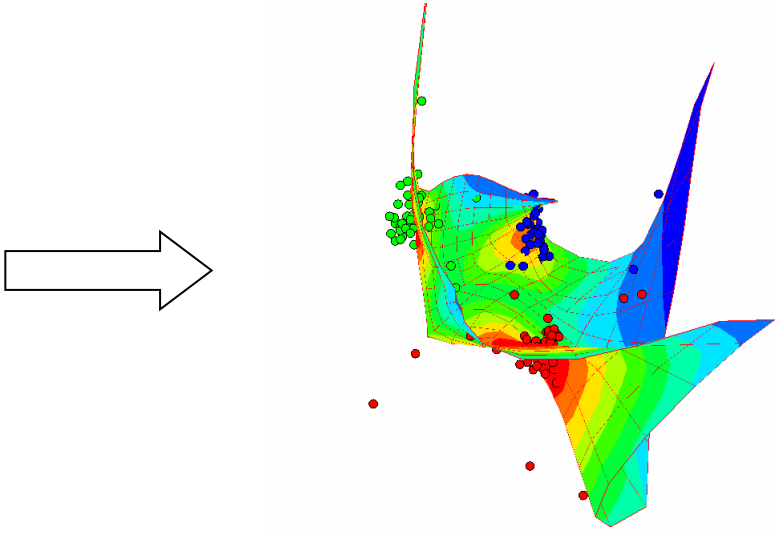
Quality of group compactness (QGC)

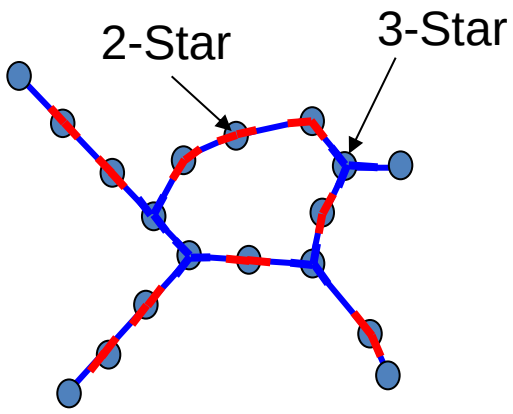
We assume that there is a label $C(i)$ associated with every point i. $N(B)$ is the number of points having the label B. For each label B, we calculate the average number of points with the same label in the k-neighborhood of the points before and after projection. Let us define $c(i; k)$ as the number of points in the k-neighbourhood of the point i having the label $C(i)$. For a label B,

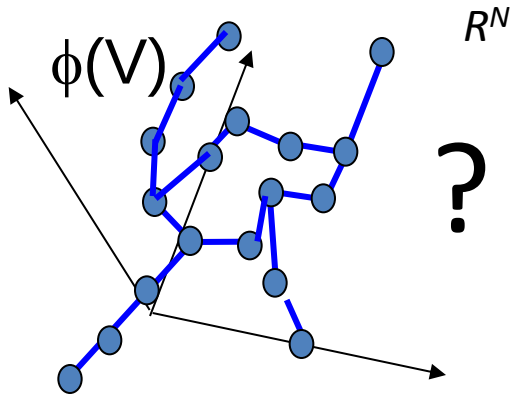
$$\mathrm{QGC}_k(B) = 1/k \sum_{C(i)=B} c(i;k)/N(B)$$

Principal graphs?

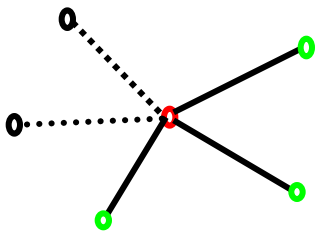


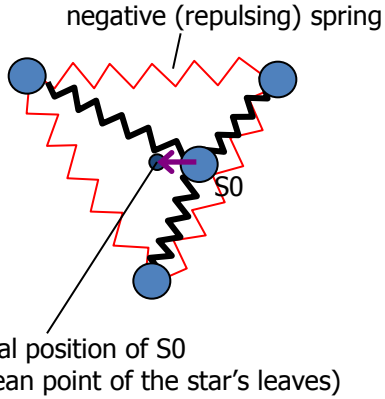






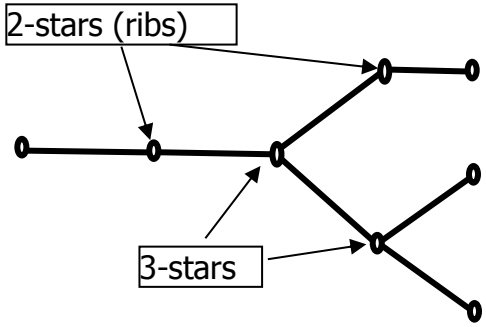
Generalization: what is ***principal graph***?
Ideal object: ***pluriharmonic graph embedment***





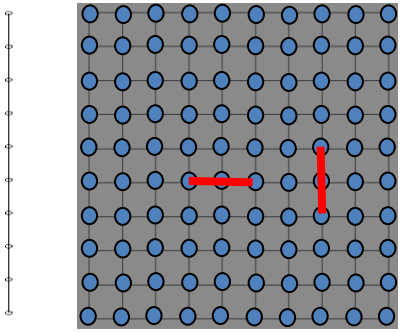
Elastic k-star (k edges, k+1 nodes).
The branching energy is

$$u_{k\text{-star}} = \mu_k \left(y_0 - \frac{1}{k} \sum_{i=1}^{k} y_i \right)^2$$



Primitive elastic graph: all non-terminal nodes with k edges are elastic k-stars.
The graph energy is

$$U_G = \sum_{\text{edges}} u_{\text{edge}} + \sum_{k} \sum_{k-\text{stars}} u_{\text{star}}$$



Pluriharmonic graph embedments generalize
straight line, rectangular grid (with proper choice of k-stars), etc.

Pluriharmonic map

Suppose that for each $k \geq 2$, a family S_k of k-stars in G has been selected. Then we define an ***elastic graph*** as a graph with selected families of k-stars S_k and for which for all $E^{(i)} \in E$ and $S_k^{(j)} \in S_k$ the corresponding elasticity moduli $\lambda_i > 0$ and $\mu_{kj} > 0$ are defined.

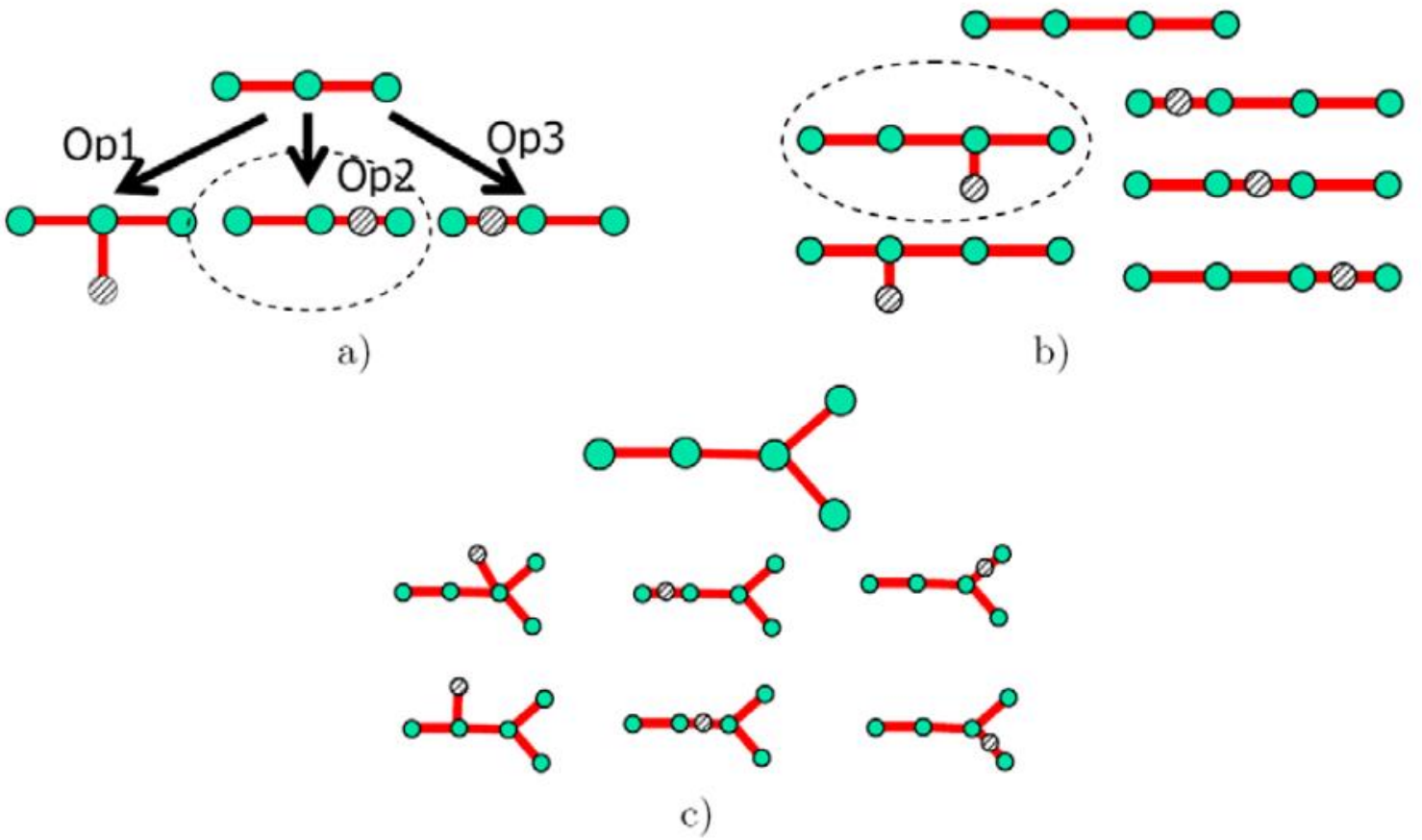
Definition. A map $\phi: V \to \mathbf{R}^m$ defined on vertices of G is ***pluriharmonic*** iff for any k-star $S_k^{(j)} \in S_k$ with the central vertex $S_k^{(j)}(0)$ and the neighbouring vertices, $S_k^{(j)}(i)$ $i = 1...k$, the equality holds:

$$\phi(S_k^{(j)}(0)) = \frac{1}{k}\sum_{i=1}^{k}\phi(S_k^{(j)}(i))$$

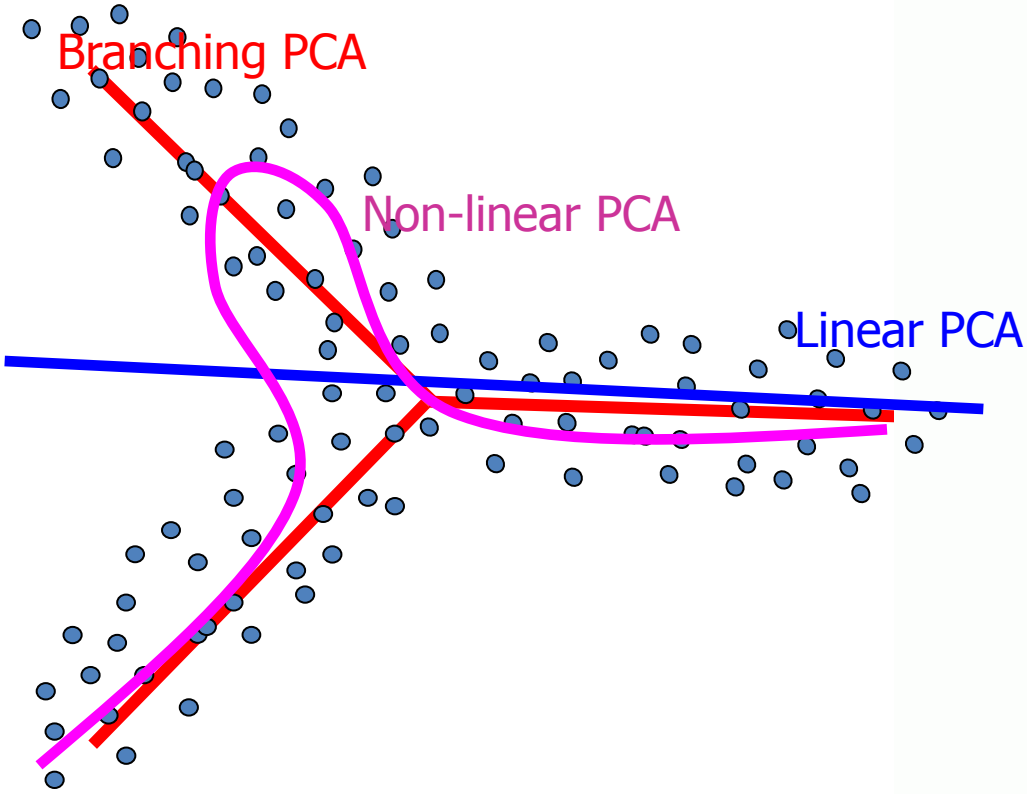
Graph grammars

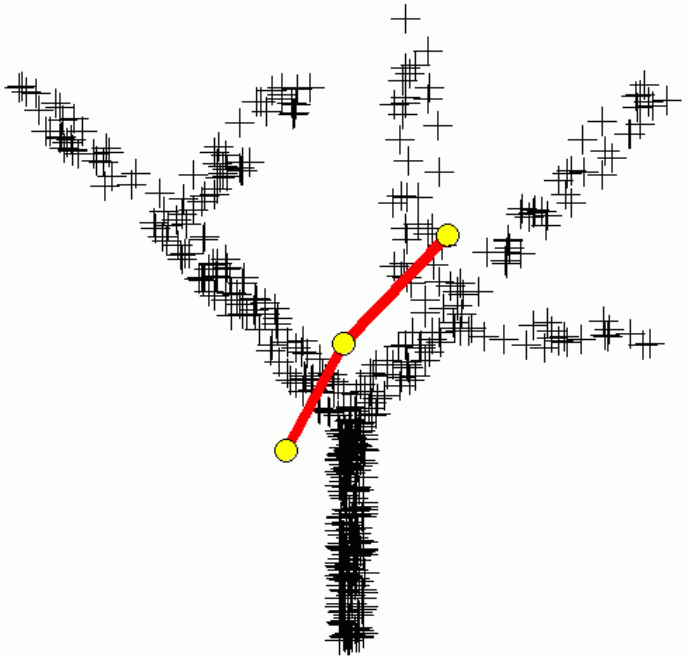
the simplest one: add a node, bisect an edge

Two operations: Operation 1) Add a node to a star Operation 2) Bisect an edge

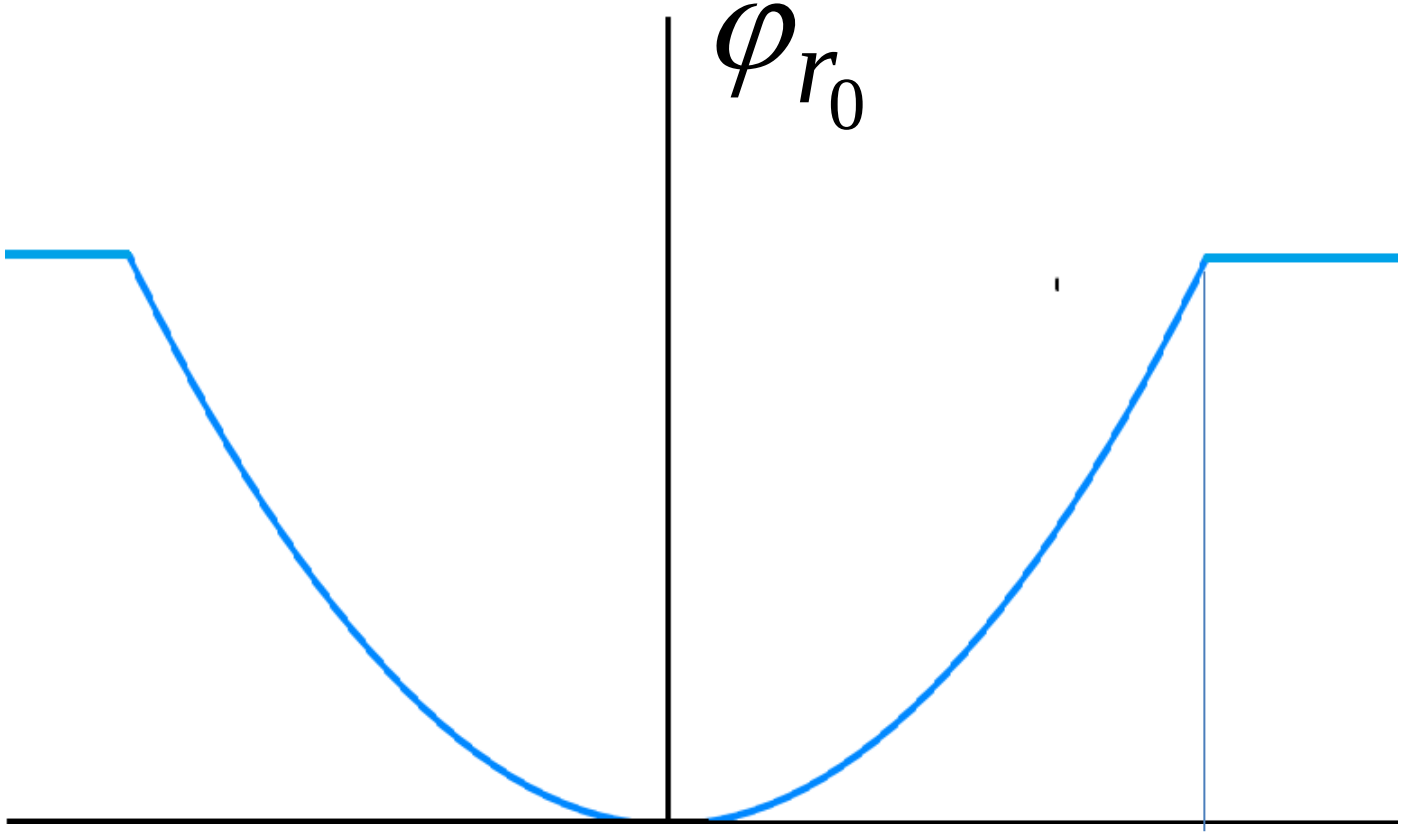


Principal harmonic dendrites (trees) approximating complex data structures



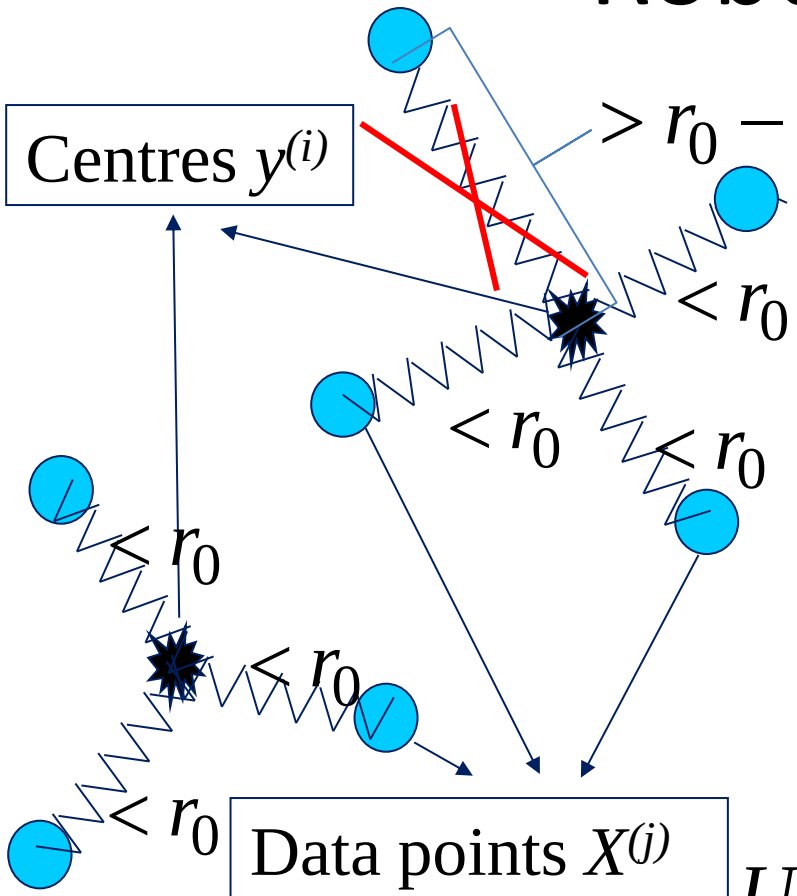


Robustness and trimmed springs



$$U^{(Y)} = \frac{1}{N}\sum_{i=1}^{p}\sum_{X^{(j)}\in K^{(i)}} \varphi_{r_0}\left(\left\| X^{j} - y^{(i)} \right\|\right)$$

Robust owners



Approximation
by smaller finite sets:

1. Select several centres;
2. Attach datapoints to their robust owners by springs;
3. Minimize energy;
4. Repeat 2&3 until converges.

$$U^{(Y)} = \frac{1}{N}\sum_{i=1}^{p}\sum_{X^{(j)}\in K^{(i)}} \varphi_{r_0}\left(\left\|X^{j} - y^{(i)}\right\|\right)$$

$$K^{i} = \left\{X^{j} : \left\|X^{j} - y^{(i)}\right\| < r_0 \;\&\; \forall k \quad \left\|X^{j} - y^{(k)}\right\| \geq \left\|X^{j} - y^{(i)}\right\|\right\}$$

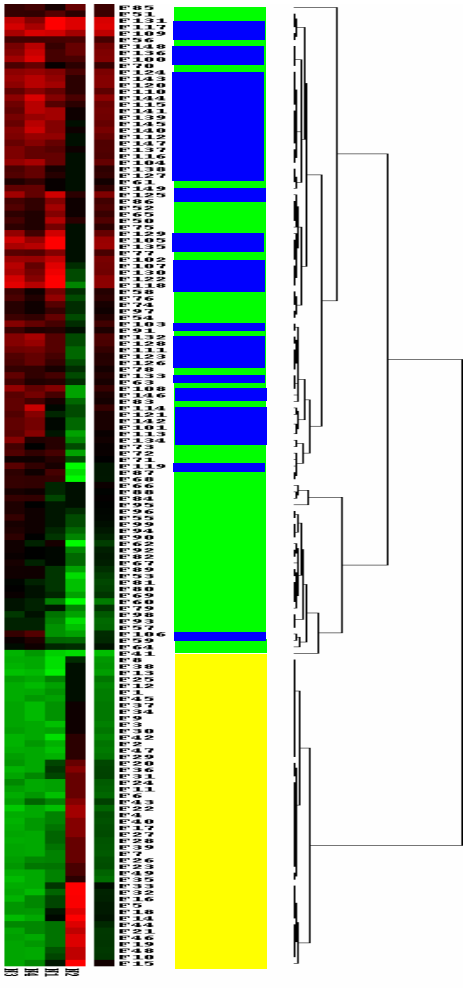
Three types of complexity

The principal graphs can be called *data approximators of controllable complexity*. By complexity of the principal objects we mean the following three notions:

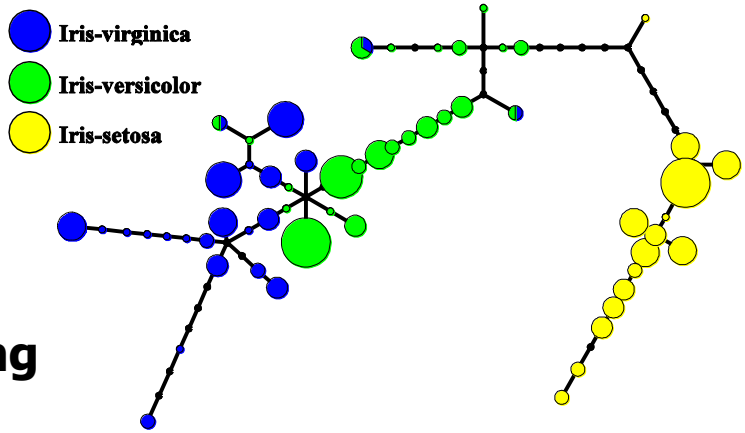
- ***Geometric complexity***: how far a principal object deviates from its ideal configuration; for the elastic principal graphs we explicitly measure deviation from the 'ideal' pluriharmonic graph by the elastic energy $U_\phi(G)$ (3) (this complexity may be considered as a *measure of non-linearity*);
- ***Structural complexity*** : it is some non-decreasing function of the number of vertices, edges and *k*-stars of different orders $SC(G)=SC(|V|,|E|,|S_2|,\ldots,|S_m|)$; this function penalises for number of structural elements;
- ***Construction complexity*** is defined with respect to a graph grammar as a number of applications of elementary transformations necessary to construct given *G* from the simplest graph (one vertex, zero edges).

HC vs Principal Trees

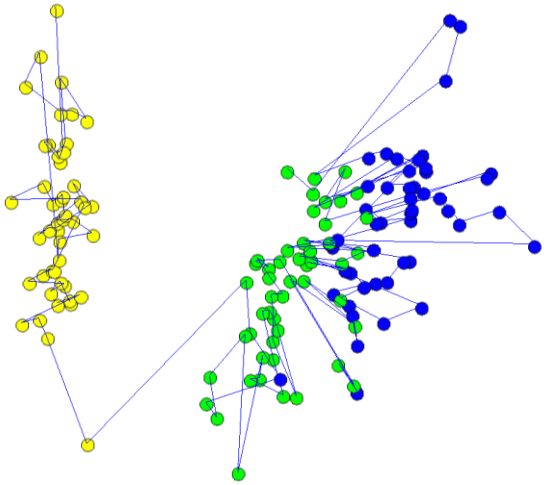
"Genealogy tree"



"Metro map"

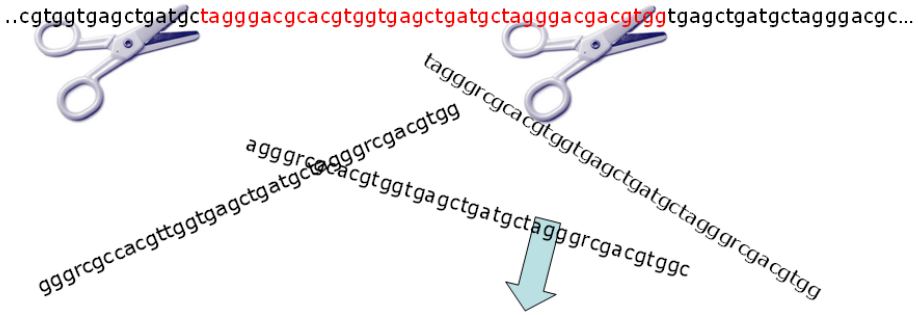


PCA, HC ordering



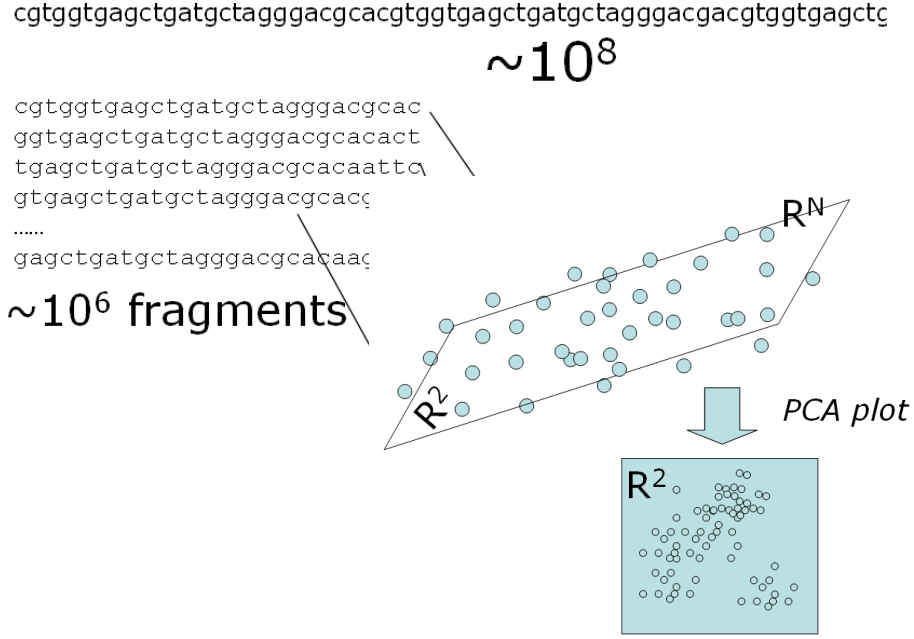
Geometrization of the text: From DNA text to the space of frequency dictionaries

Fragmentation of the DNA text



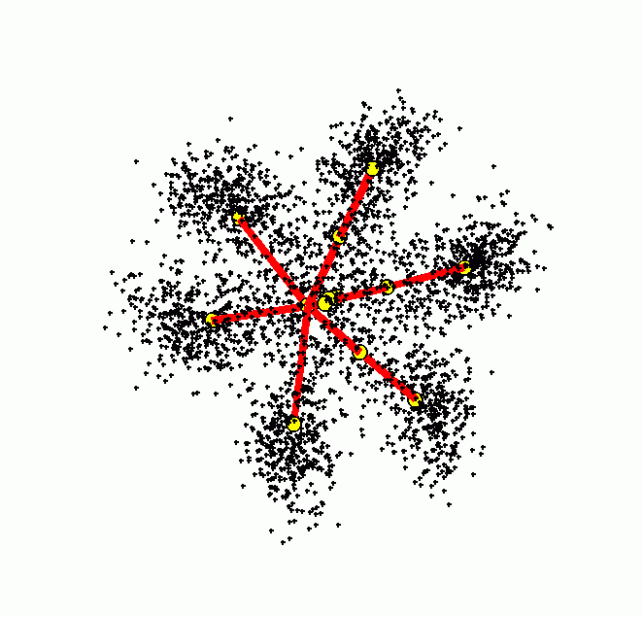
frequency dictionaries:

g g g a c g c a c g t g g t g a g c t g a t g c t a g g g	N = 4=4^1
gg ga cg ca cg tg gt ga gc tg at gc ta gg	N = 16=4^2
gga cgc acg tgg tga gct gat gct agg	N = 64=4^3
g gacg cacg tggt gagc tgat gcta gggr	N=256=4^4

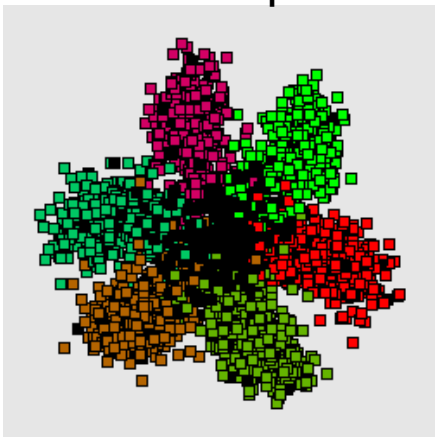


Visualization of 7-cluster genome sequence structure

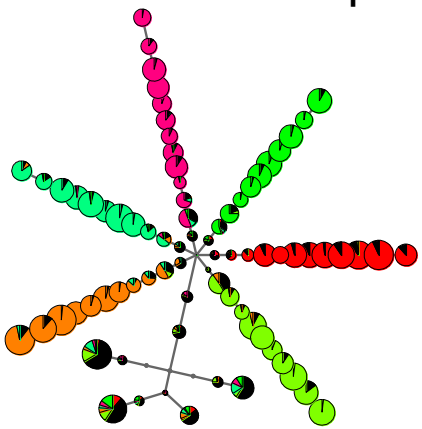
Algorithm iterations

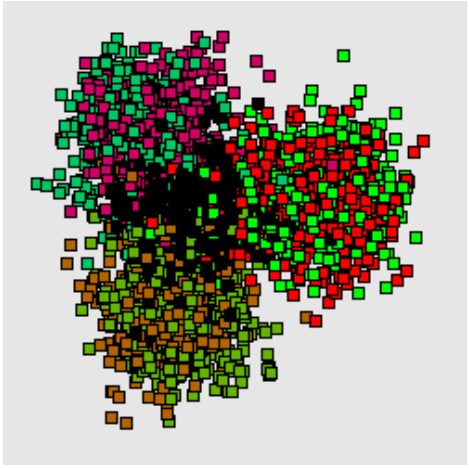


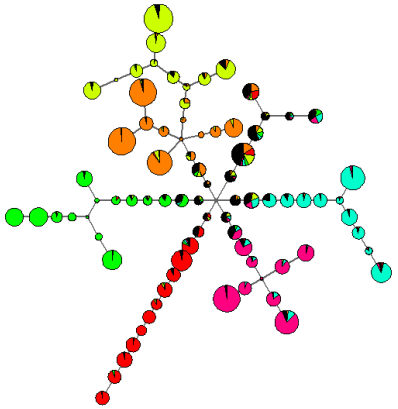
3D PCA plot



Metro map





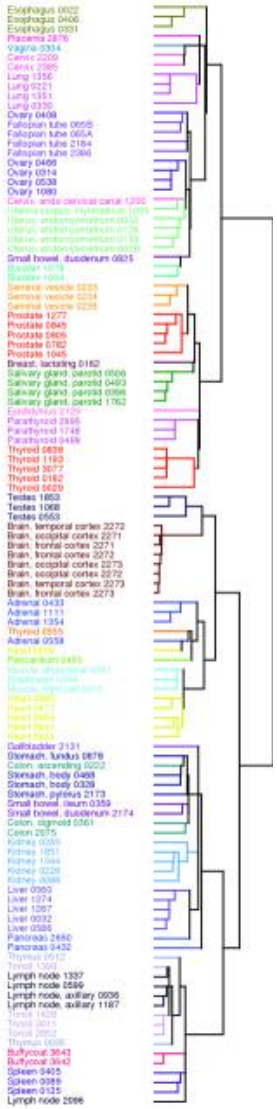


Here clusters overlapping on 3D PCA plot are in fact well-separated and the principal tree reveals this fact

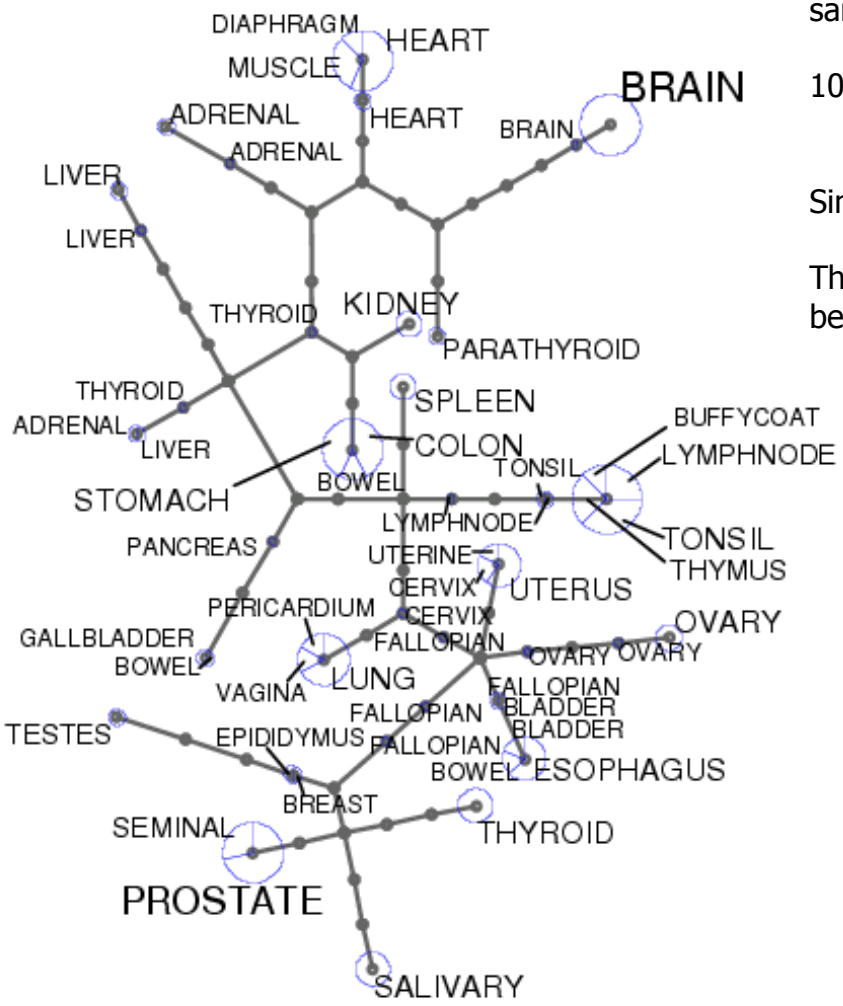
Hierarchical clustering *vs* principal trees, or Genealogy tree *vs* Metro map

(Gorban, Sumner & Zinovyev, *LNCSE*, 2007)

Genealogy tree



Metro map



Microarray data from
Shyamsundar et al. *Genome Biology*, 2005

Gene expression in 103 normal human tissue samples

10383-dimensional space, many missing values

Similar tissues are closely clustered

The tree allows to estimate 'distance' between human tissues

Conclusion

- **Method of elastic maps**: Efficient method and interactive software for constructing low-dimensional non-linear principal manifolds;
- **Principal manifold** as a **screen** for visualizing multidimensional data and functions with their **uncertainties;**
- **Non-linear data visualization** displays are systematically better than linear ones (four quality criteria: *MSE*, *Distance mapping*, *Point entourage*, *Class compactness*);

- **Pluriharmonic graph** (with quadratic energy functionals for deviation from the ideal form) provide us with a rich set of approximants;
- **Topological grammars** and E/M algorithms form an effective technology for datasets approximation;
- **Metro maps** provide us with a nice robust visualisation tool;
- **It works**.

- [1] A. N. Gorban, A. Y. Zinovyev, Principal Graphs and Manifolds. In: Handbook of Research on Machine Learning, IGI Global, 2009. 28-59.
  E-print: http://arxiv.org/abs/0809.0490
- [2] A. N. Gorban, A. Zinovyev, Principal manifolds and graphs in practice: from molecular biology to dynamical systems, International Journal of Neural Systems, Vol. 20, No. 3 (2010) 219-232, E-print: http://arxiv.org/abs/1001.1122

- More: just go to arXiv and look for gorban

Change of era

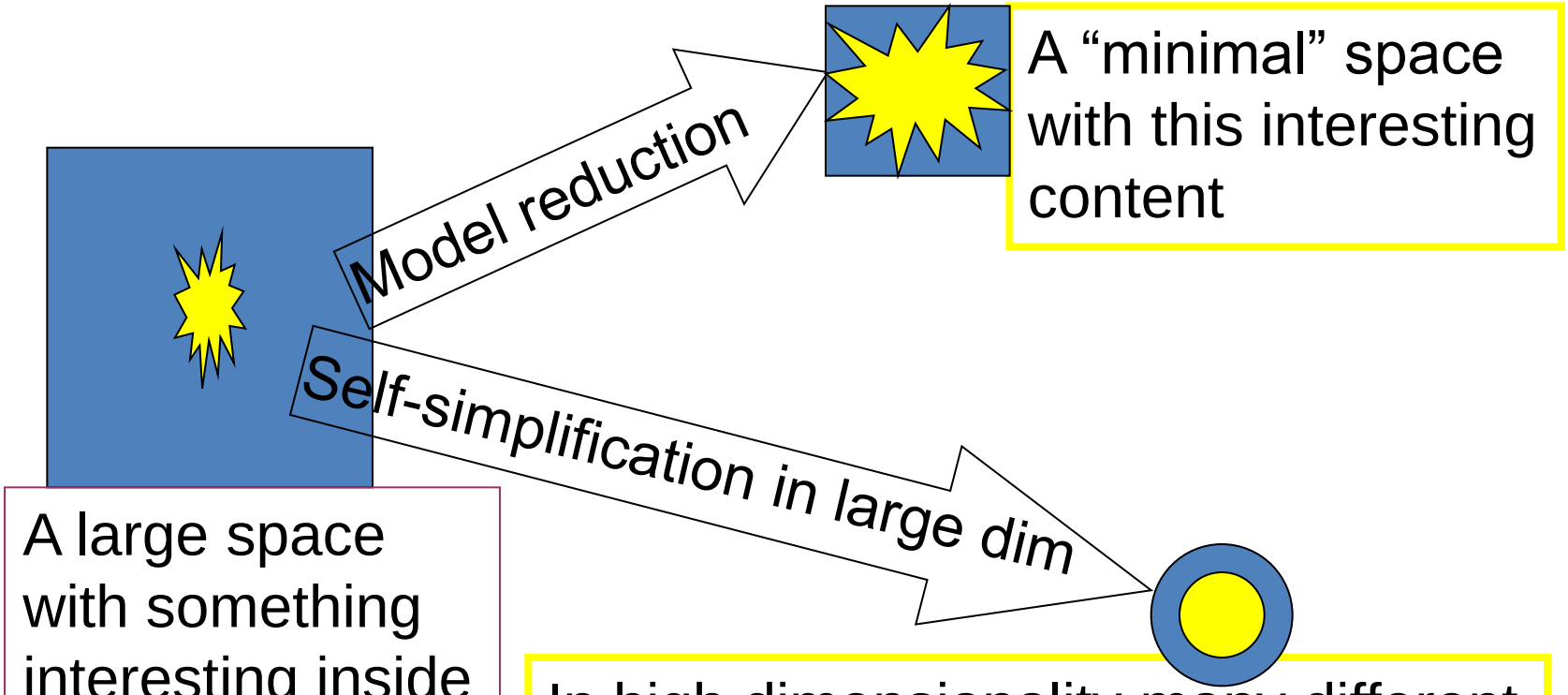
From Einstein's **"flight from miracle."**

«… The development of this world of thought is in a certain sense **a continuous flight from "miracle**".»

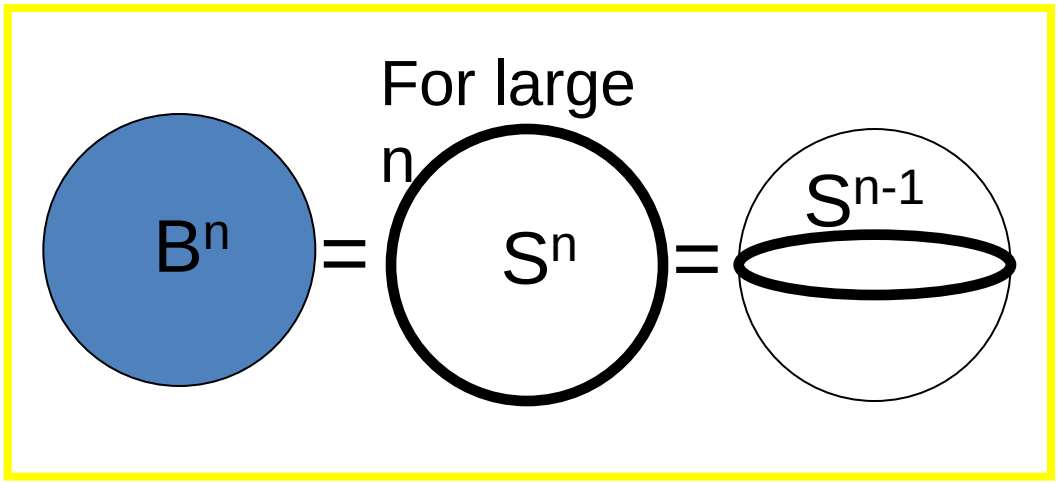
To **struggle with complexity**

"I think the next century will be the century of complexity."
Stephen Hawking

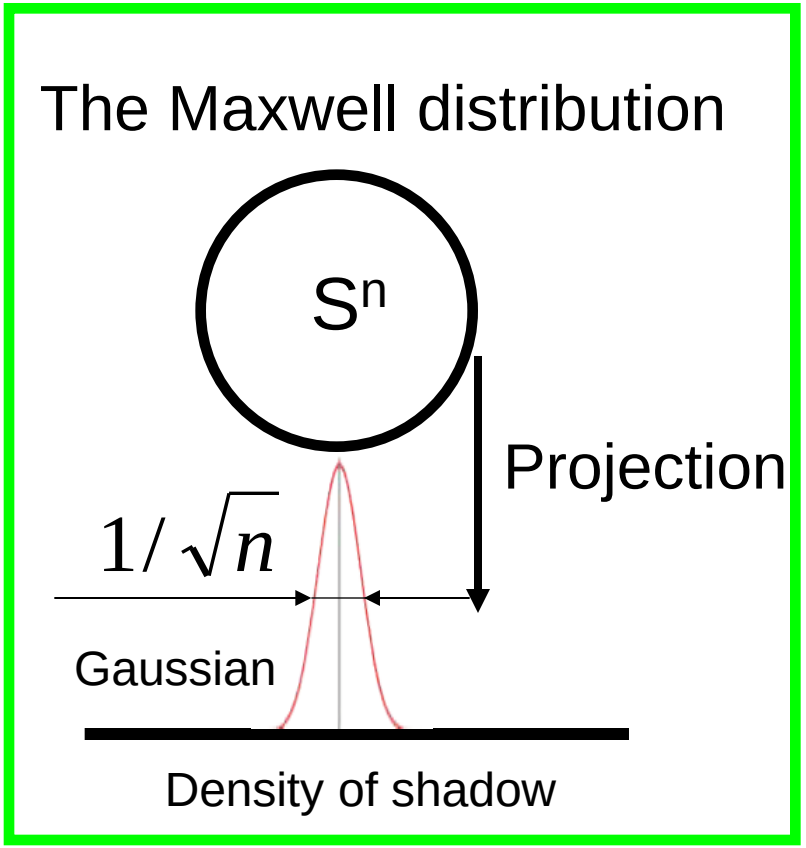
Two main approaches in our struggle with complexity



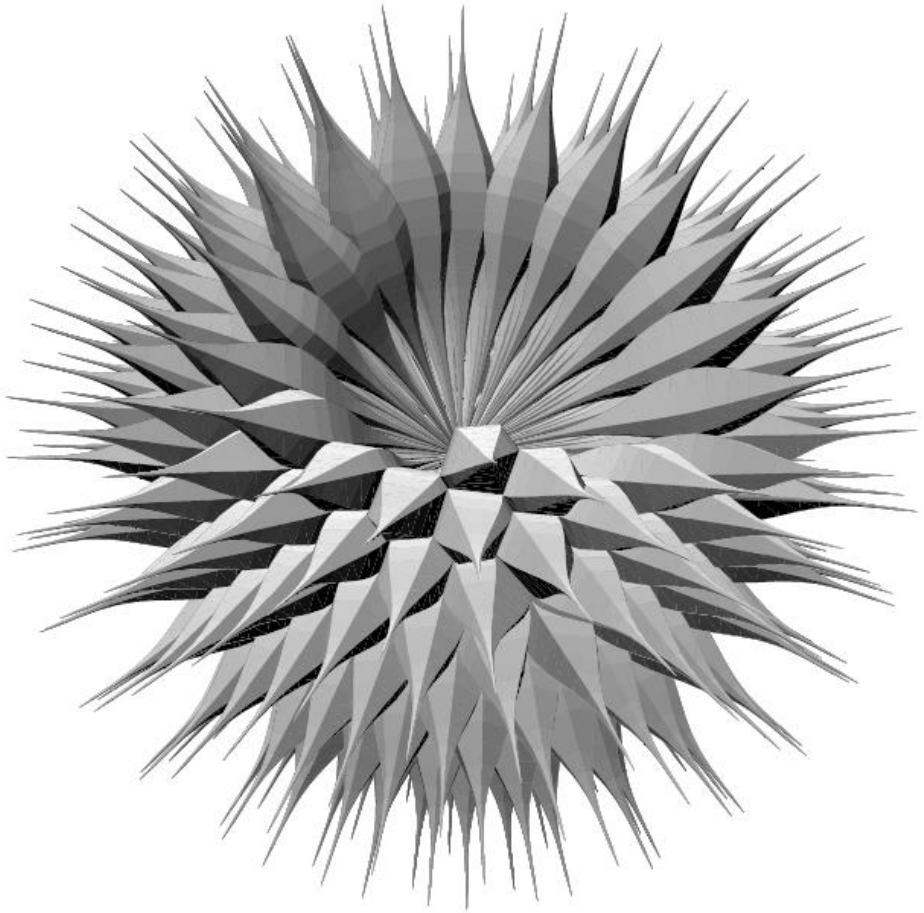
Measure concentration effects



Maxwell
Gibbs
Milman
Talgrand
Gromov
……



A 3D representation of an 8D hypercube



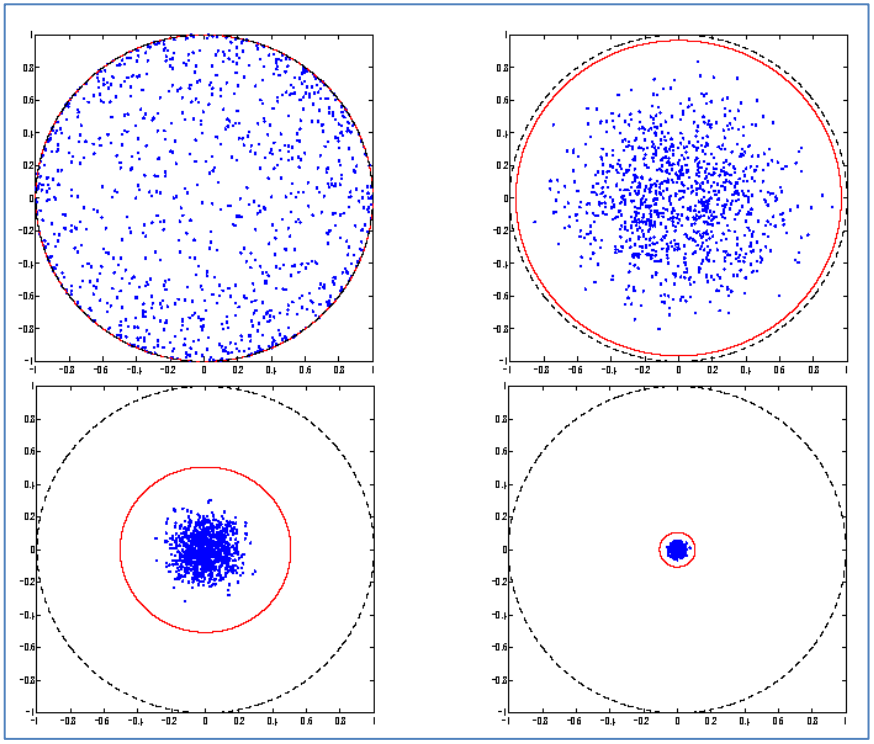
The body has the same radial distribution and the same number of vertices as the hypercube.

A very small fraction of the mass lies near a vertex.

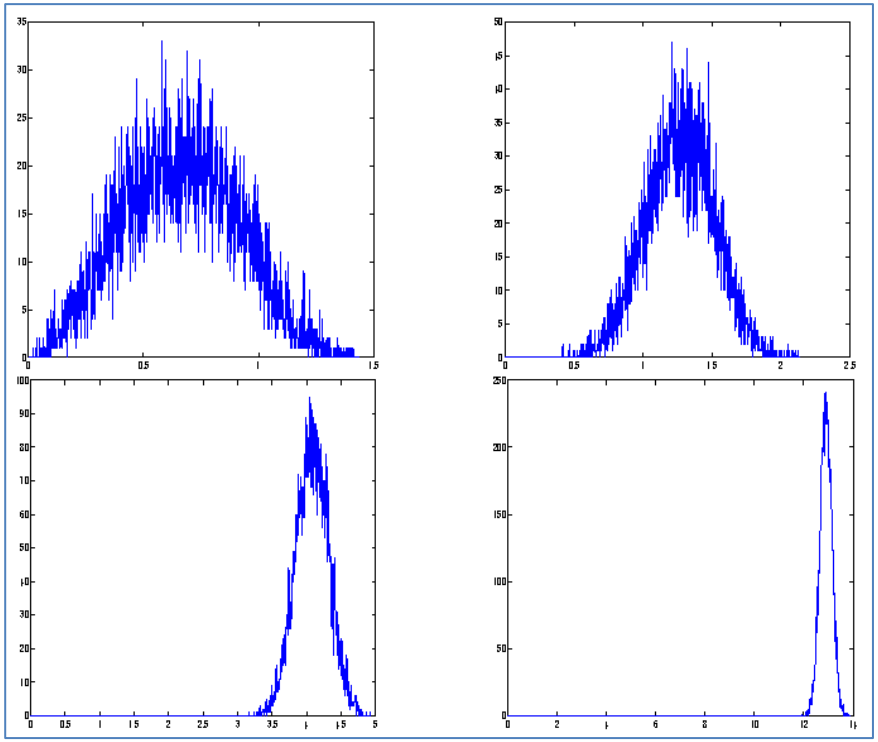
Also, most of the interior is void.

(Illustration by Hamprecht & Agrell, 2002)

Strange properties of high dimensional sets



Observable diameter of the sphere $\boldsymbol{S}^n$, n = 3, 10, 100, 2500.



Distribution of distances for pairs of points in the unit hypercube $\boldsymbol{I}^n$, n = 3, 10, 100, 1000. (For random samples of

Three provinces of the Complexity Land