Geometry of data sets

Alexander Gorban
University of Leicester, UK

Plan

- The problem
- Approximation of multidimensional data by low-dimensional objects
- Self-simplification of essentially high-dimensional sets
- Terra Incognita between low-dimensional sets and self-simplified high-dimensional ones.

Change of era

Misha Molibog Graphics

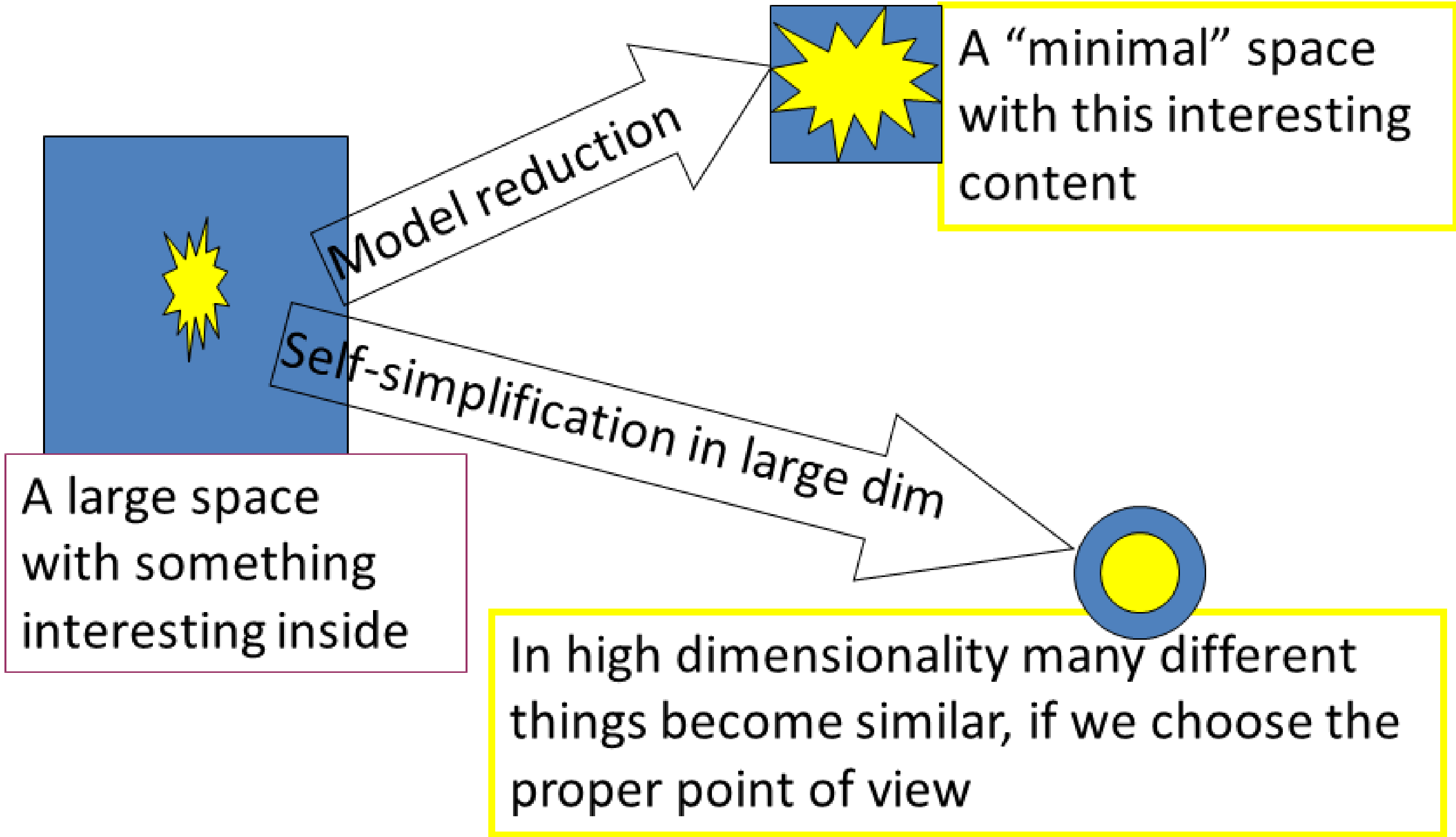
From Einstein's **"flight from miracle."**

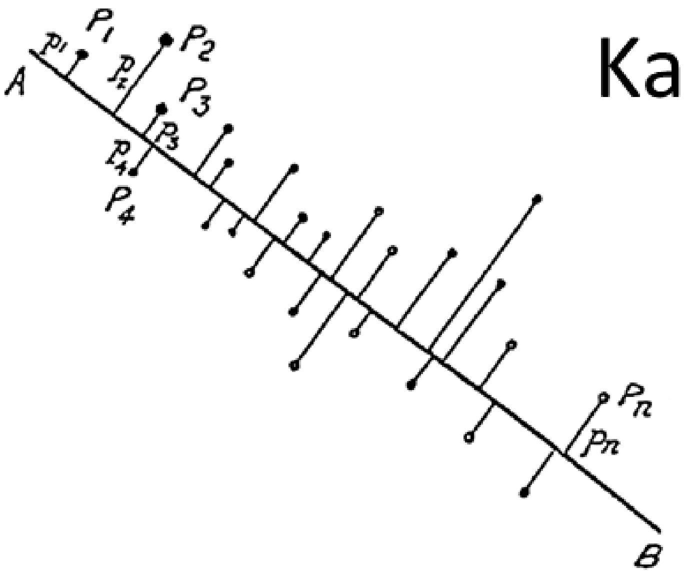
«... The development of this world of thought is in a certain sense **a continuous flight from "miracle"**.»

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





Karl Pearson
1901

[559]

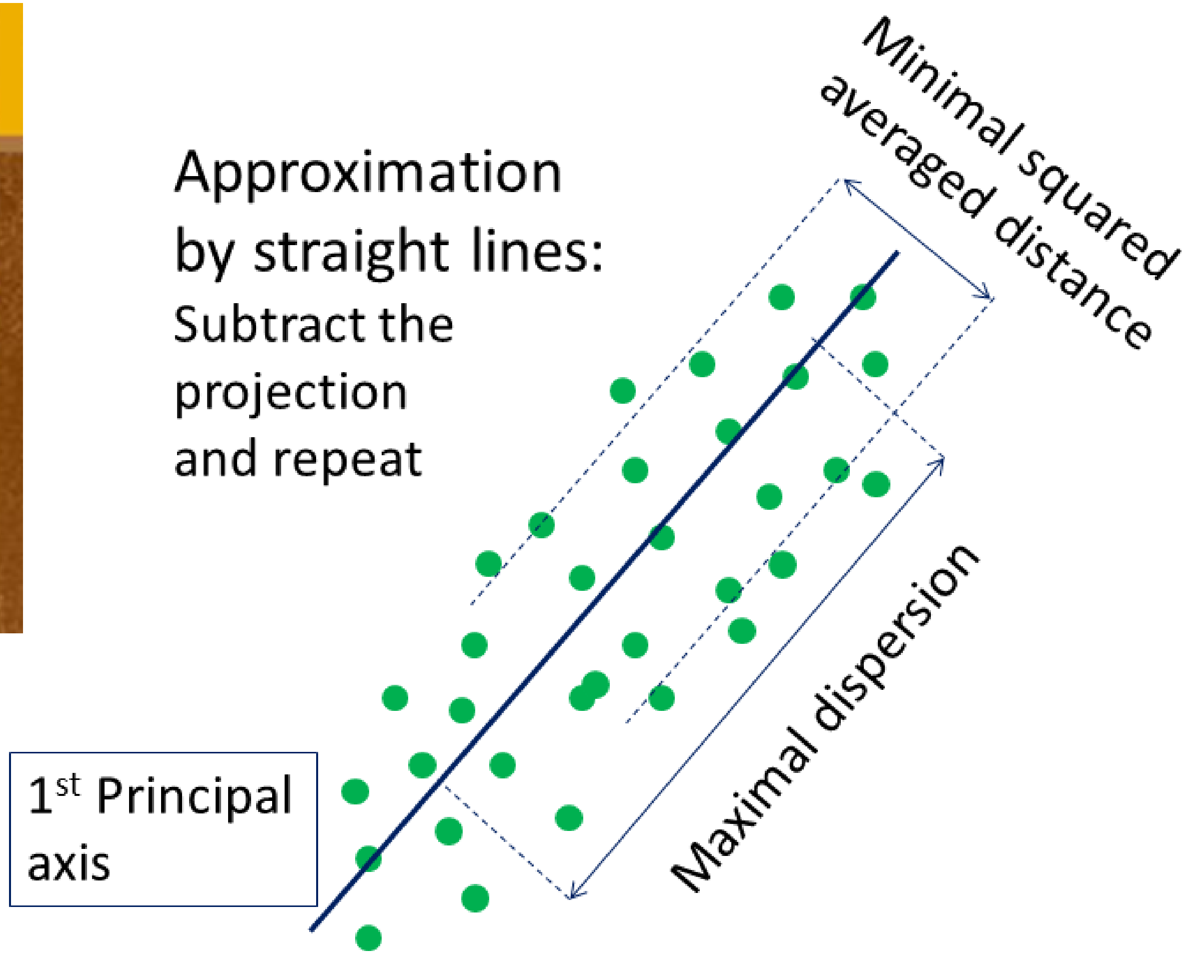
LIII. *On Lines and Planes of Closest Fit to Systems of Points in Space.* By KARL PEARSON, *F.R.S., University College, London* *.

(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

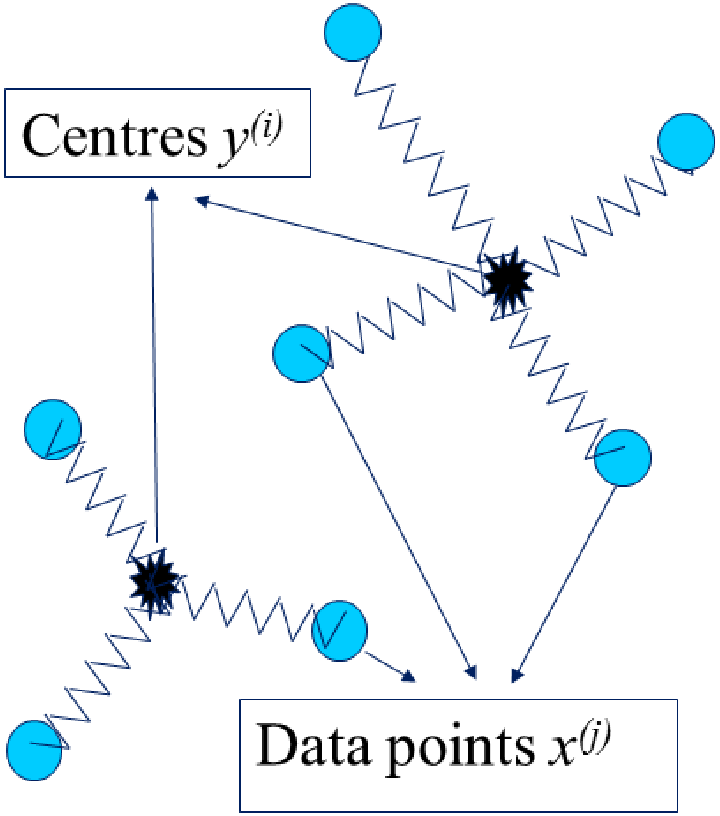
Principal Component Analysis

Approximation by straight lines:

Subtract the projection and repeat



Principal points (K-means)

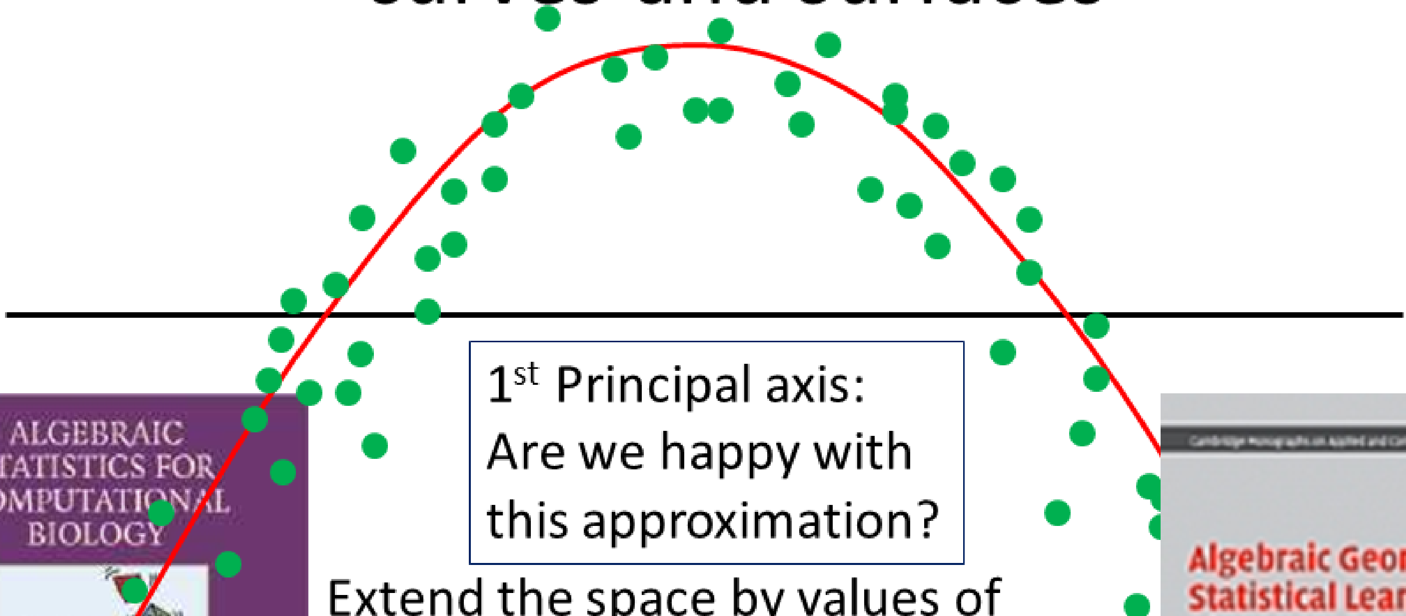


Approximation
by smaller finite sets:

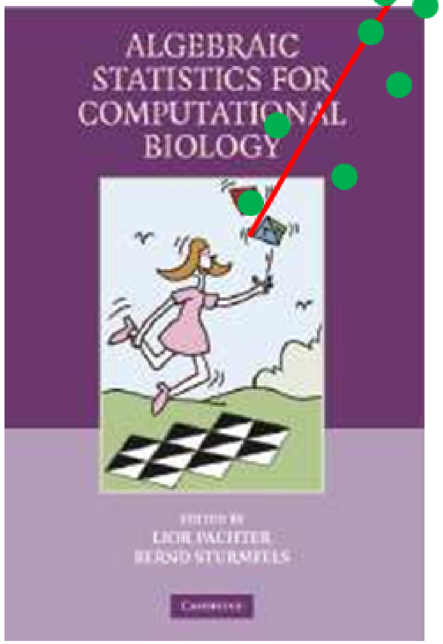
1. Select several centres;
2. Attach datapoints to the closest centres by springs;
3. Minimize energy;
4. Repeat 2&3 until converges.

Steinhaus, 1956;
Lloyd, 1957;
MacQueen, 1967

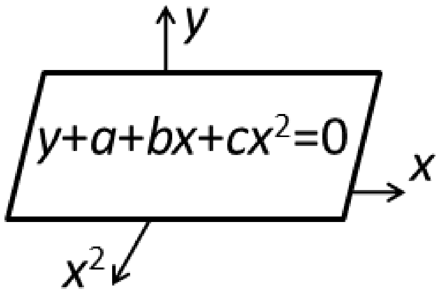
Approximation by algebraic curves and surfaces



1st Principal axis:
Are we happy with
this approximation?



Extend the space by values of additional functions and apply PCA



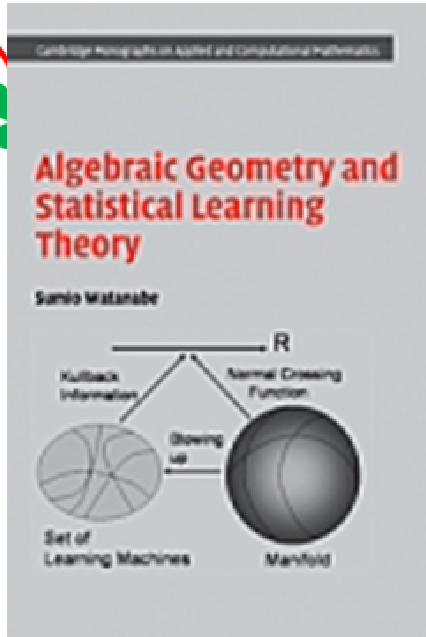
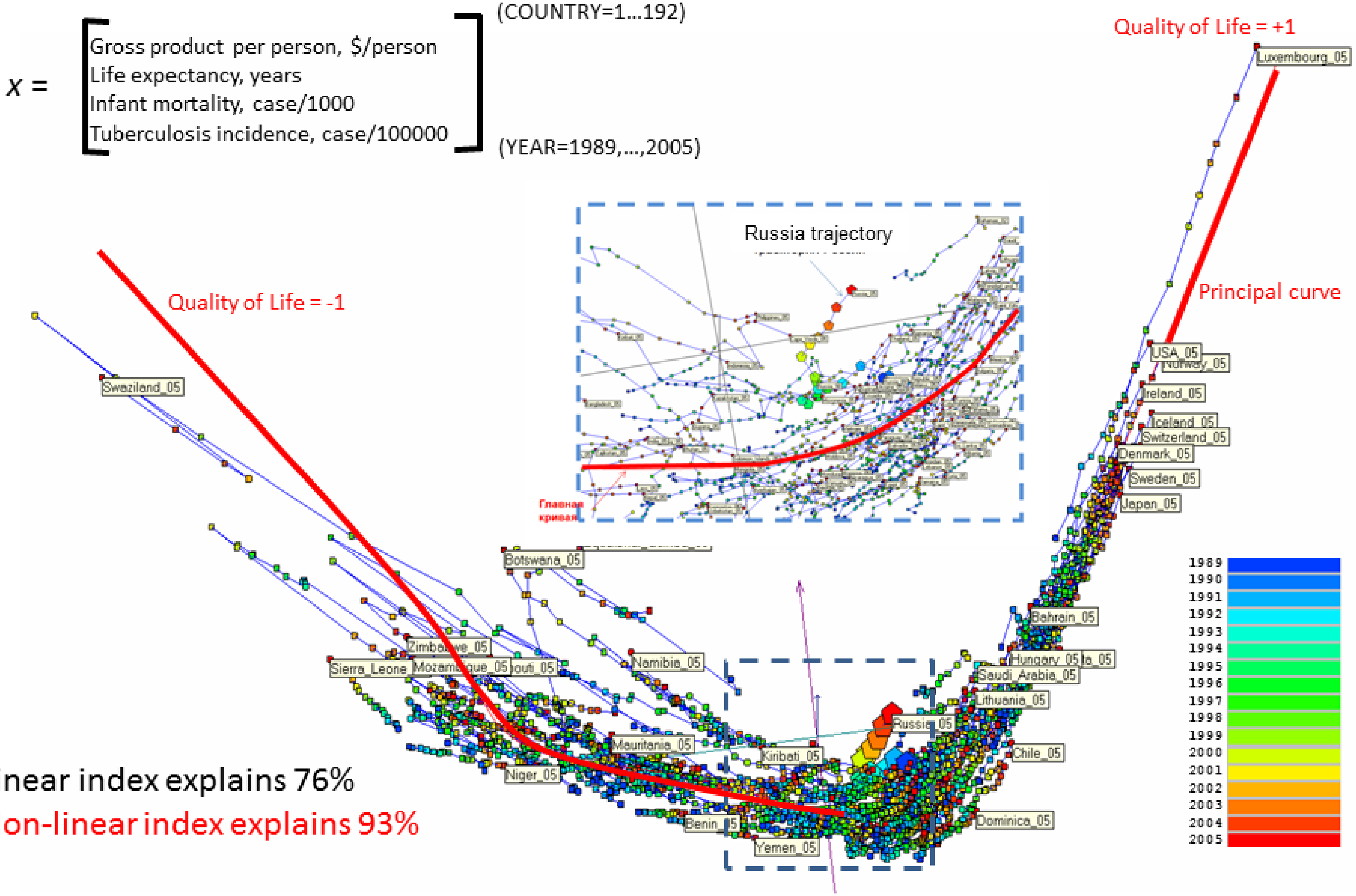


Illustration: Nonlinear happiness

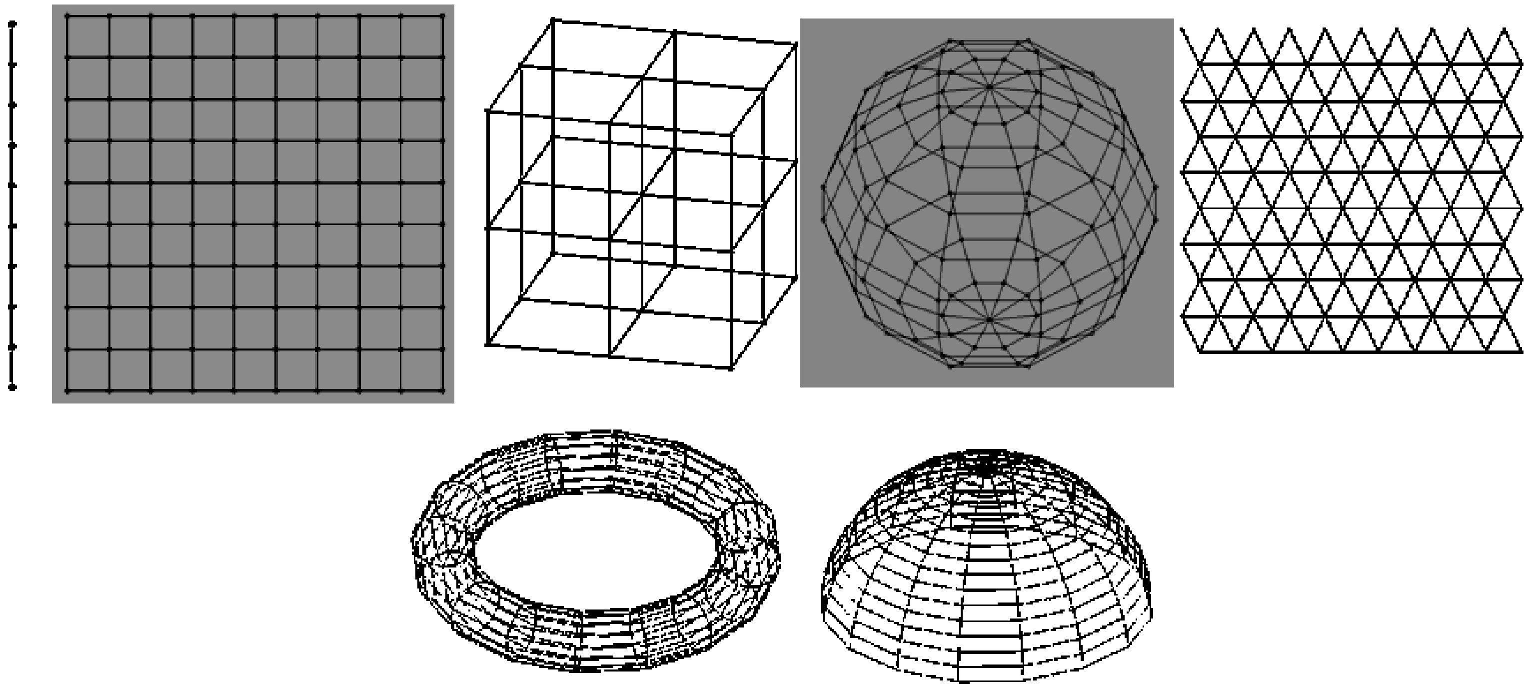
$$x = \begin{bmatrix} \text{Gross product per person, \$/person} \\ \text{Life expectancy, years} \\ \text{Infant mortality, case/1000} \\ \text{Tuberculosis incidence, case/100000} \end{bmatrix}^{(\text{COUNTRY}=1\ldots192)}_{(\text{YEAR}=1989,\ldots,2005)}$$



Linear index explains 76%

Non-linear index explains 93%

Constructing elastic nets



Definition of elastic energy: we borrow this approach from splines

X^j

y

$\left\| X^j - y^{(i)} \right\|^2$

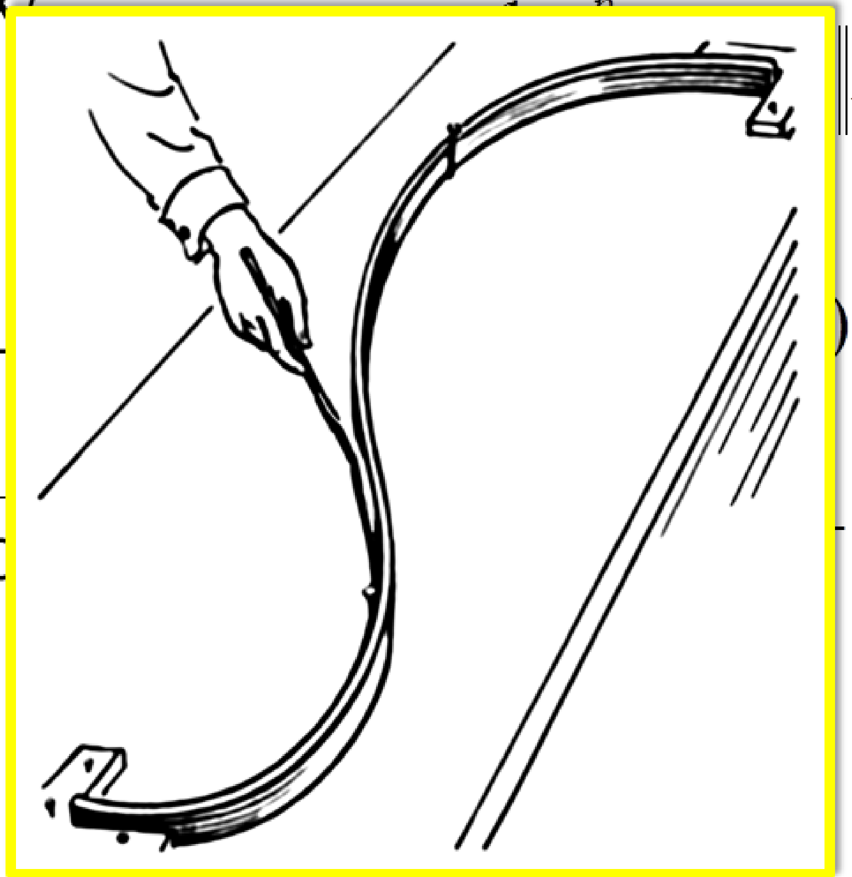
$E(0)$ $E(1$

$) - E^{(i)}(0) \right\|^2$

$R(1)$ $R(0$

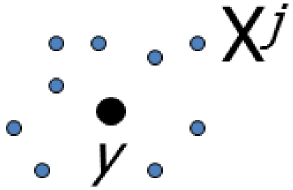
$R^{(i)}(2) - 2R^{(i)}(0) \right\|^2$

$\mu_i = \mu_0$

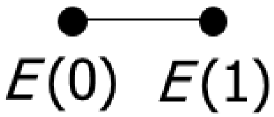


$\to \min$

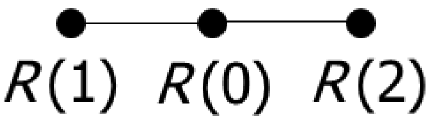
Definition of elastic energy



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



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

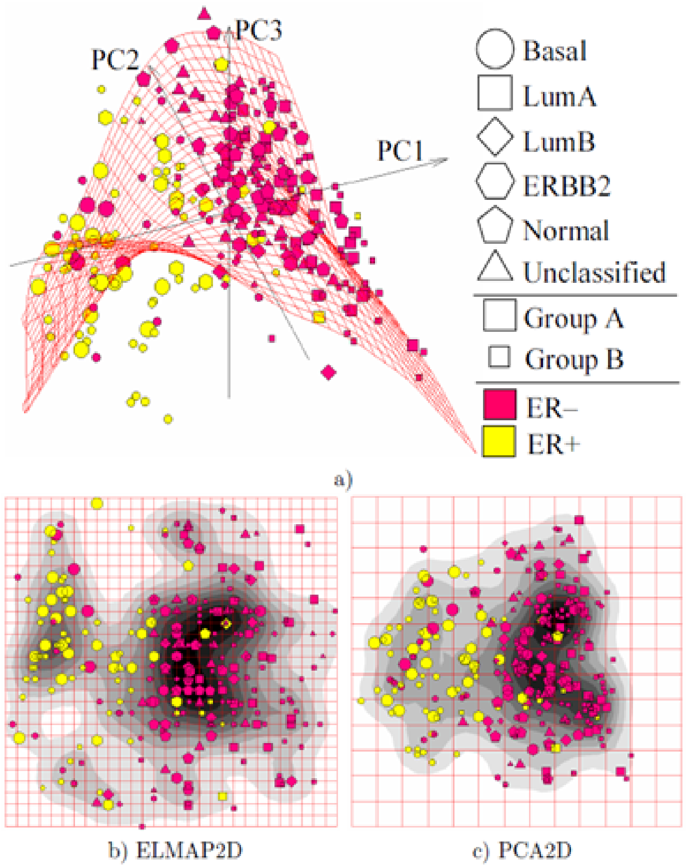


$$U^{(R)} = \sum_{i=1}^{r} \mu_i \left\| R^{(i)}(1) + R^{(i)}(2) - 2R^{(i)}(0) \right\|^2$$

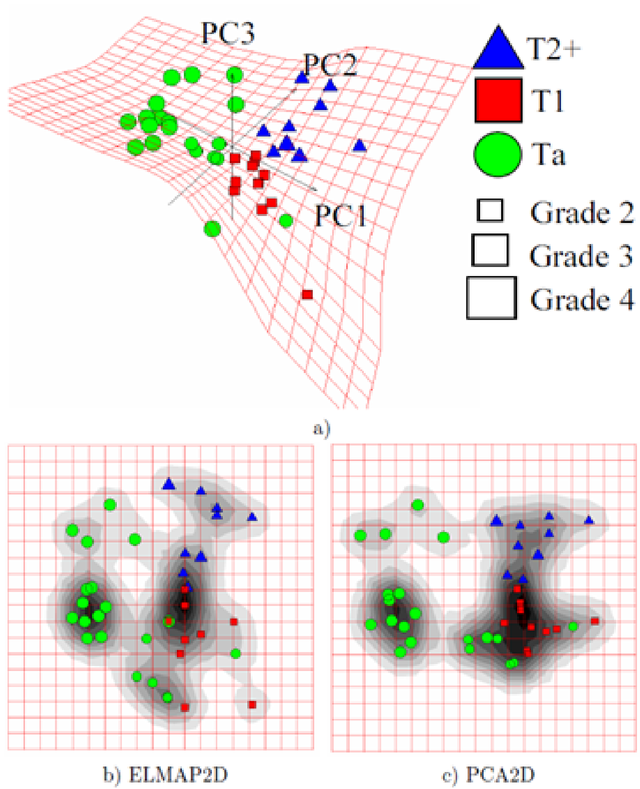
$$\lambda_i = \lambda_0, \quad \mu_i = \mu_0$$

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

Are non-linear projections better than linear projections?

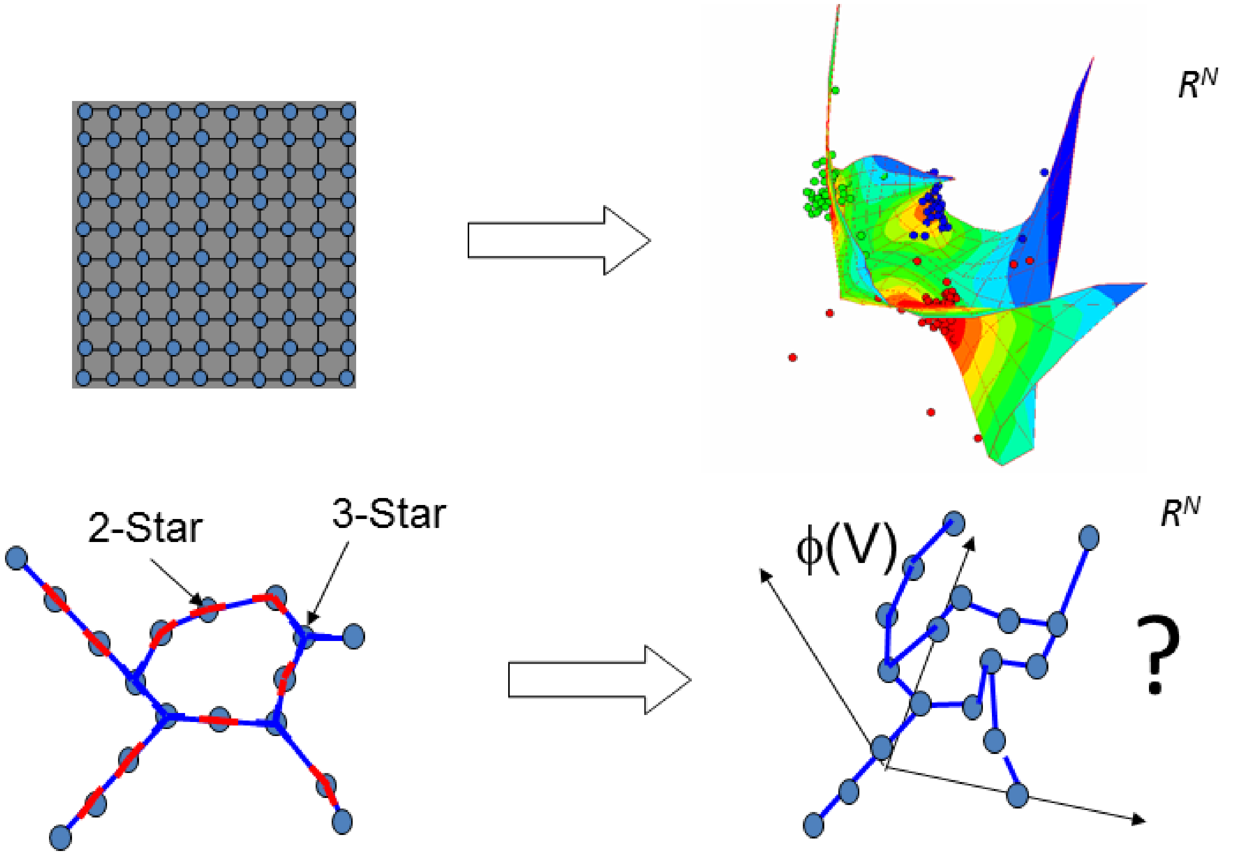


Breast cancer
Wang et al., 2005

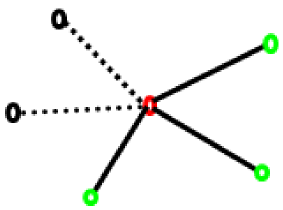


Bladder cancer
Dyrskjot et al., 2003

Principal graphs?

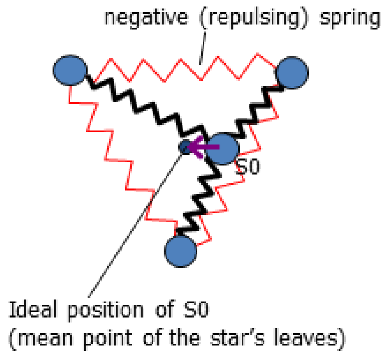


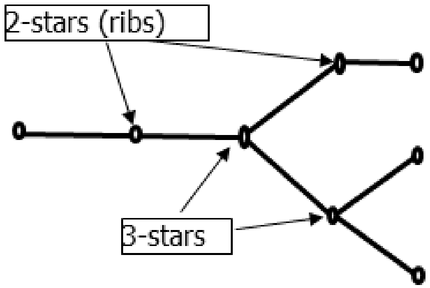
Generalization: what is ***principal graph***?
Ideal object: ***pluriharmonic graph embedment***



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

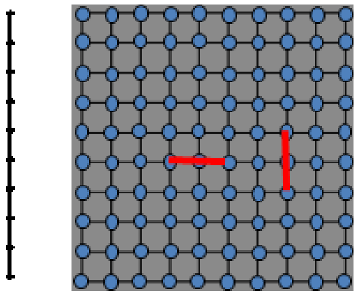
$$u_{k\text{-star}} = \mu_k \left(y_0 - \frac{1}{k} \sum_{i=1}^{k} y_i \right)^2$$





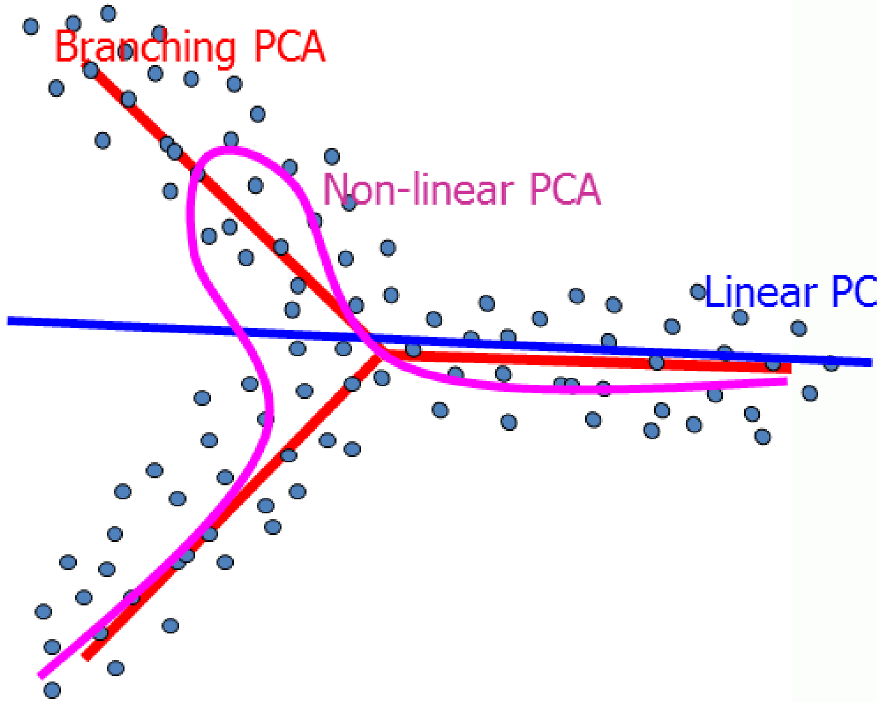
Primitive elastic graph: all non-terminal nodes with k edges are elastic k-stars.
The graph energy is

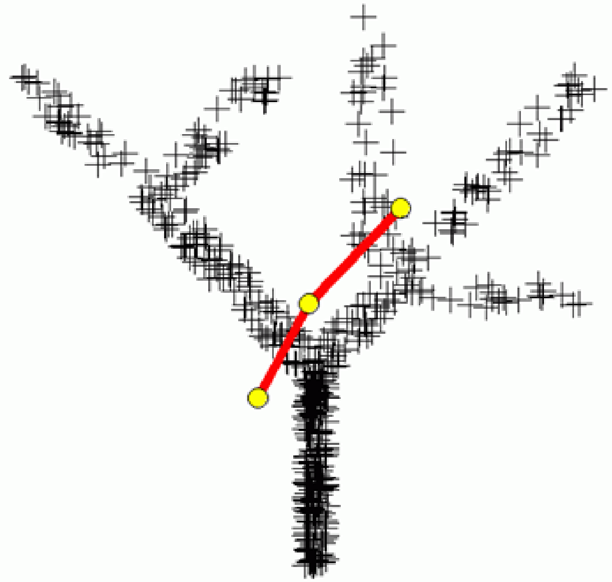
$$U_G = \sum_{\text{edges}} u_{\text{edge}} + \sum_{k} \sum_{k-\text{stars}} u_{\text{star}}$$



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

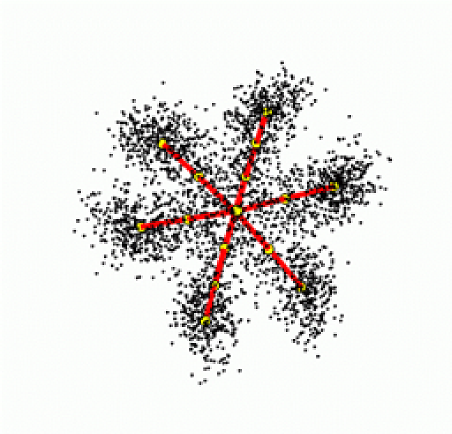
Principal harmonic dendrites (trees) approximating complex data structures



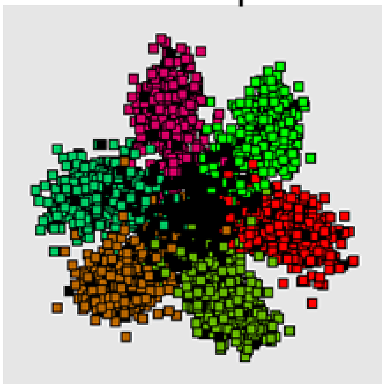


Visualization of 7-cluster genome sequence structure

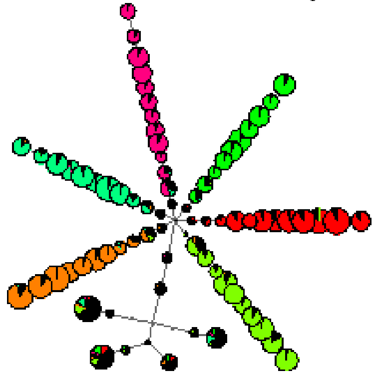
Algorithm iterations



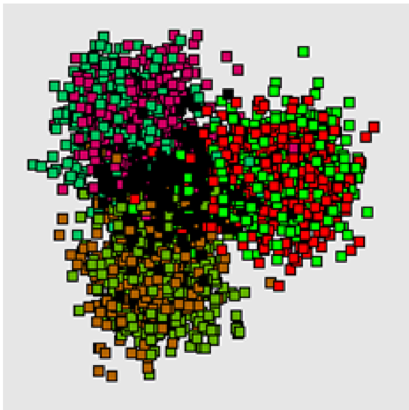
3D PCA plot

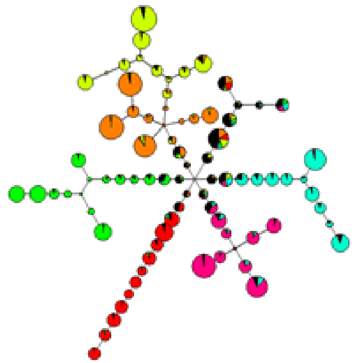


Metro map



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





And much more for low-dimensional subsets:

- Local Linear Embedding
- Isomap
- Laplace Eigenmaps
- Nonlinear Multidimensional Scaling
- Independent Component Analysis
- Persistent cohomology
- ....

Three provinces of the Complexity Land