Dynamic Quantum Clustering facilitates visual exploration of Big Data

David Horn

http://horn.tau.ac.il

Halina-Fest Jan. 2014

Big Data: The Challenge

Big Data: Data of all kinds – structured and unstructured - is being collected and warehoused at a tremendous rate.

Problem: Convert all of this *information to understanding.*

Needed: To move beyond our preconceptions of what is in the data and see what is actually there.

Roadmap

- Preprocessing of data: using SVD
- Presenting data points by Gaussians:

  Projecting data space -> Hilbert space
- QC: Potential transform representing data density. Potential minima=cluster centers
- DQC: Embedding into Schrödinger equation and employing quantum gradient descent
- Application to highly complex data in nanochemistry: finding a needle in a haystack
- Application to earthquake data

What Does Data Look Like To SVD ?

Features

Entries

M_{11}	M_{12}	$\cdots$	M_{1n}
M_{21}	$\ddots$		M_{2n}
$\vdots$		$\ddots$	$\vdots$
M_{m1}	M_{m2}	$\cdots$	M_{mn}

$$M = USV^{+}$$

$$M_{ij} = \sum_{\alpha} \lambda_{\alpha} U_{i\alpha} V_{\alpha j}$$

How well does this work ?

Any data matrix is a picture, or at least behaves like a picture!

The potential transform

Represent data points by Gaussians.
Scale-space approach: study the sum of Gaussians

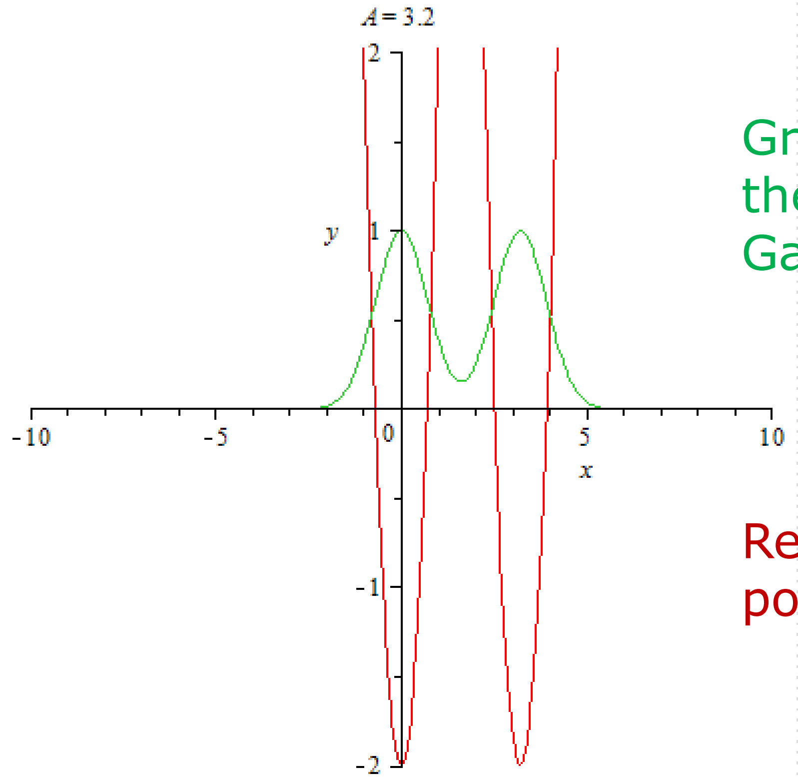
$$\varphi(\vec{x}) = \sum_{i=1}^{n} e^{-\frac{1}{2\sigma^2}(\vec{x}-\vec{x_i})\cdot(\vec{x}-\vec{x_i})}$$

For this probability amplitude we define the potential transform V

$$-\frac{\sigma^2}{2}\nabla^2\varphi + V(\vec{x})\varphi = 0 \qquad V(\vec{x}) = \frac{\sigma^2}{2\varphi}\nabla^2\varphi$$

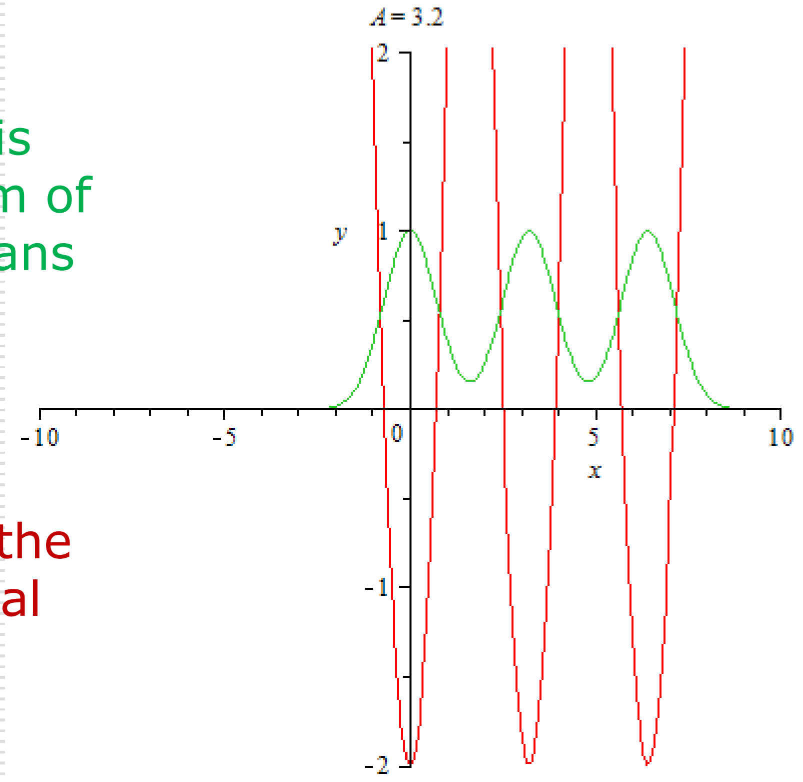
A single Gaussian transforms into a harmonic potential

Comparing sums of Gaussians (centered at 0, A, 2A) and their potentials

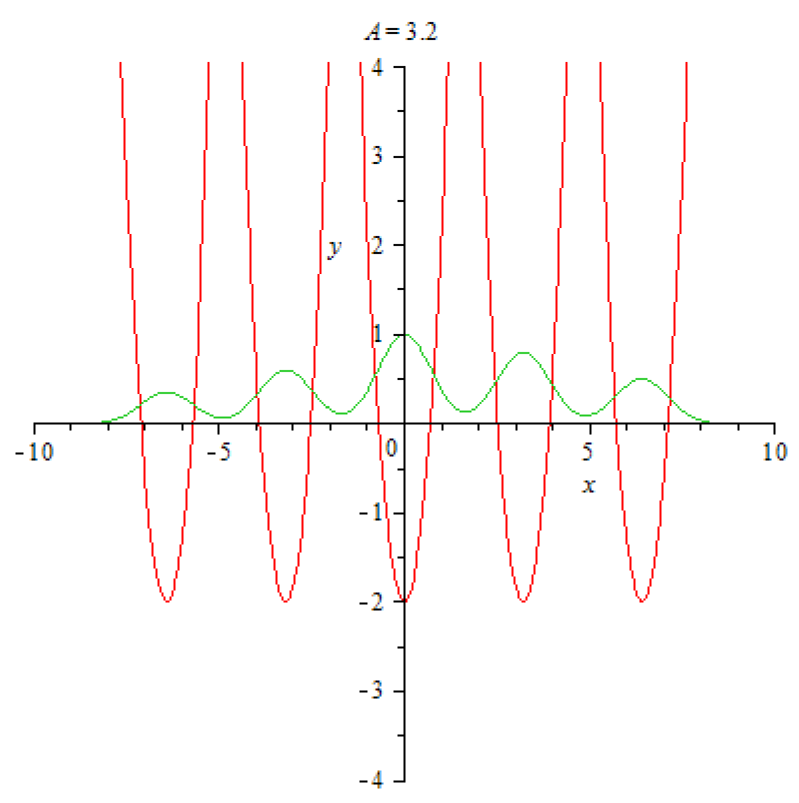


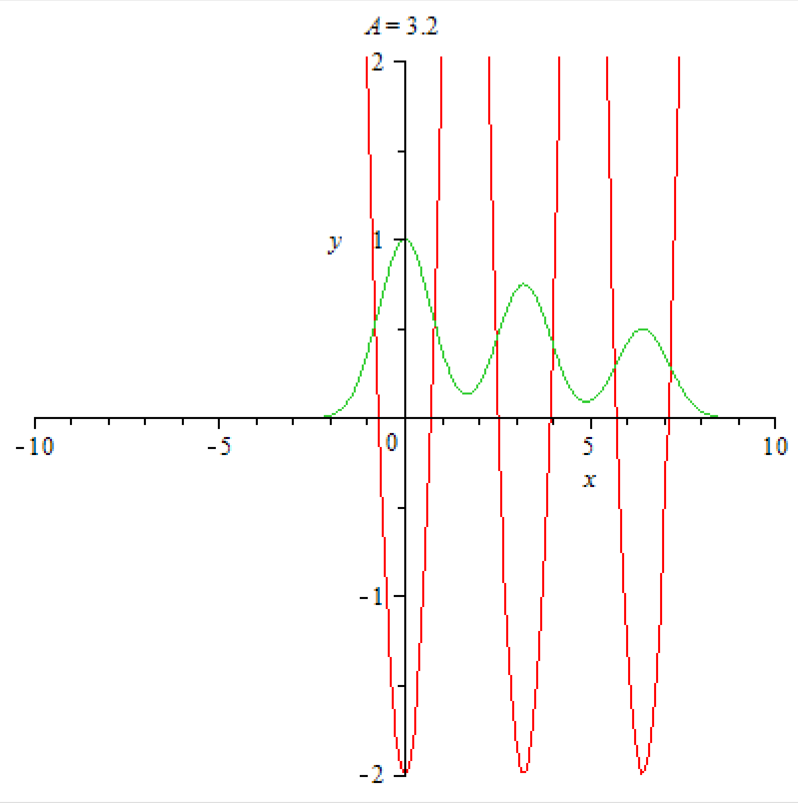
Green is the sum of Gaussians

Red is the potential



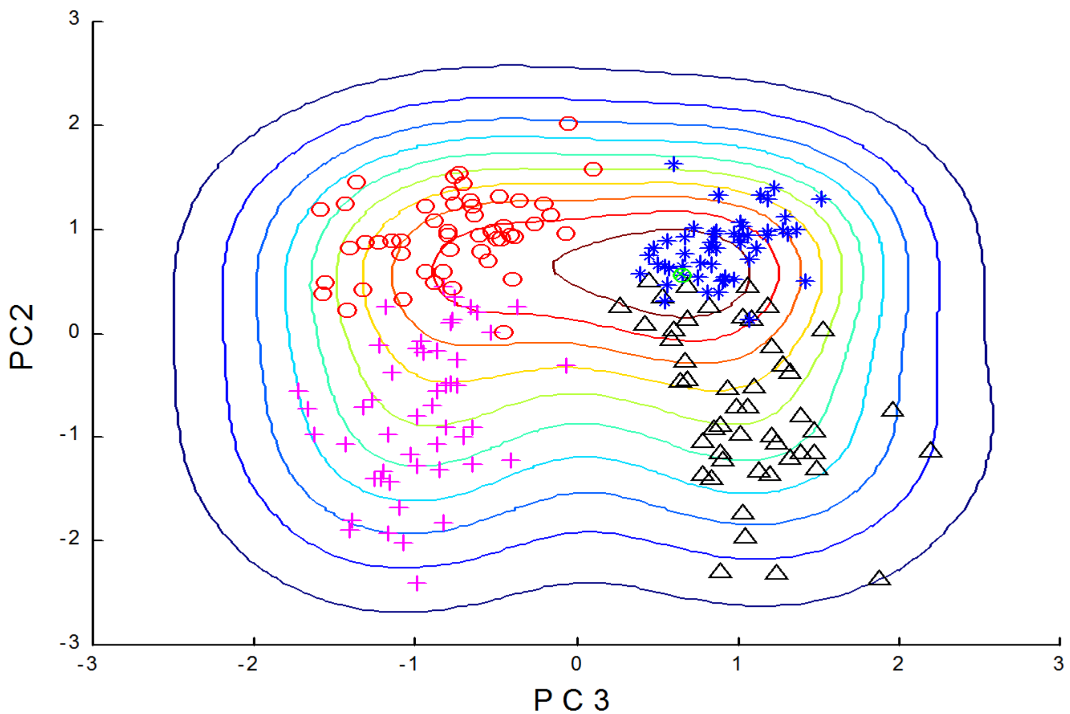
The potential can be thought of as an unbiased way of contrast enhancing the Parzen function to better reveal structure in the data





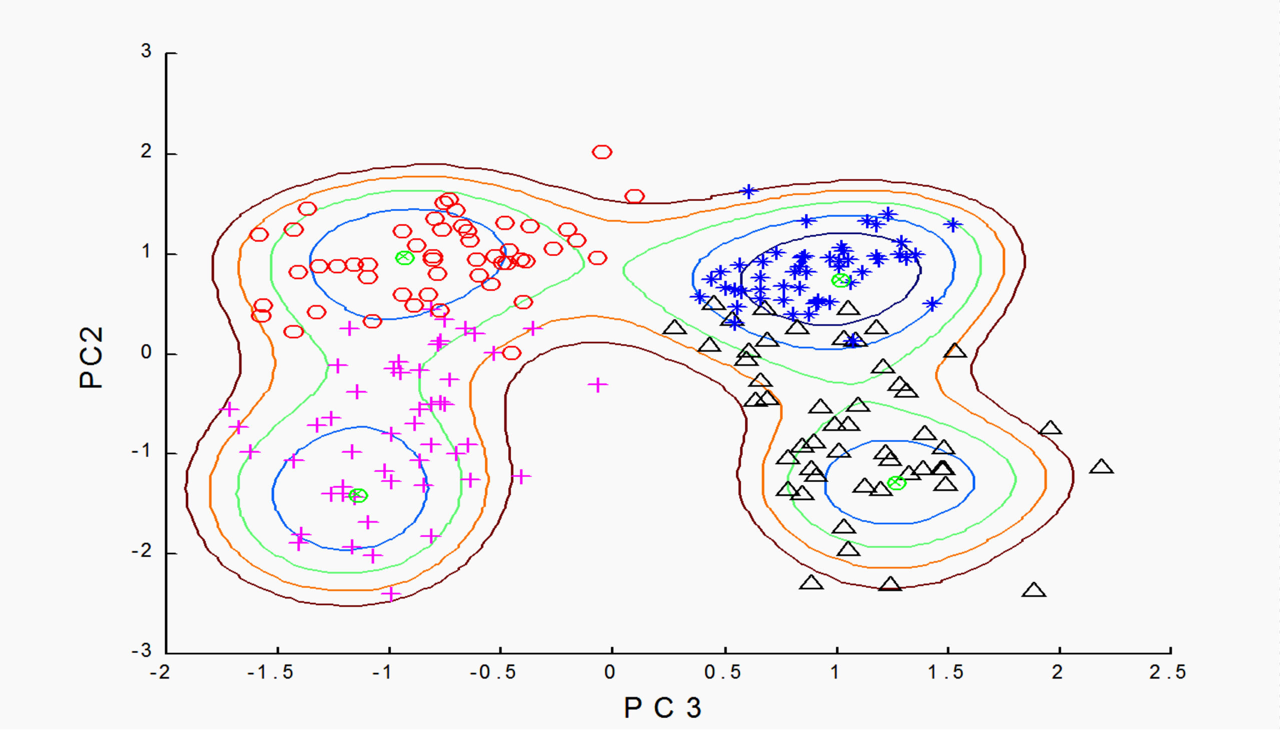
The Crabs Example (from Ripley's textbook)

4 classes, 50 samples each, d=5



A topographic map of the probability distribution for the crab data set with $\sigma=1/\sqrt{2}$. There exists only one maximum although there are 4 classes.

The potential transform
exhibits four minima identified with cluster centers



A topographic map of the potential for the crab data set with $\sigma=1/\sqrt{2}$.

Quantum Clustering: Horn and Gottlieb, Phys. Rev. Lett. 88 (2002) 018702

Dynamic Quantum Clustering

Replace the gradient-descent algorithm by a solution of the time-dependent Schrödinger equation, starting with each of the original Gaussians

$$-i\frac{\partial \Psi_i(\vec{x},t)}{\partial t} = \left(-\frac{\nabla^2}{2m} + V(\vec{x})\right)\Psi_i(\vec{x},t)$$

and tracing the convergence of its center-of-mass

$$\langle\, x(\vec{t})\,\rangle = \int d\vec{x}\,\Psi^*(\vec{x},t)\,\vec{x}\,\Psi(\vec{x},t)$$

M. Weinstein and D. Horn, 2009. Dynamic quantum clustering: a method for visual exploration of structures in data. Phys. Rev. E 80, 066117.

Dynamic Quantum Clustering

The differential equation can be solved algebraically by expanding the Hamiltonian within the n Gaussian states defined at the n data-points. Thus, for any dimension, the problem can be reduced to an nXn set of matrix elements.

$$H = \frac{p^2}{2m} + V(x)$$

$$H_{ij} = \langle \psi_i | H | \psi_j \rangle$$

$$N_{ij} = \langle \psi_i | \psi_j \rangle$$

$$\vec{X}_{ij} = \langle \psi_i | \vec{x} | \psi_j \rangle$$

Then exponentiate the finite matrix and compute the time evolution of the expectation values

Application to Sloan Digital Sky Survey of 140K galaxies

Analyzing Big Data with Dynamic Quantum Clustering
M. Weinstein, F. Meirer, A. Hume, Ph. Sciau, G. Shaked, R. Hofstetter, E. Persi, A. Mehta, D. Horn
http://arxiv.org/abs/1310.2700

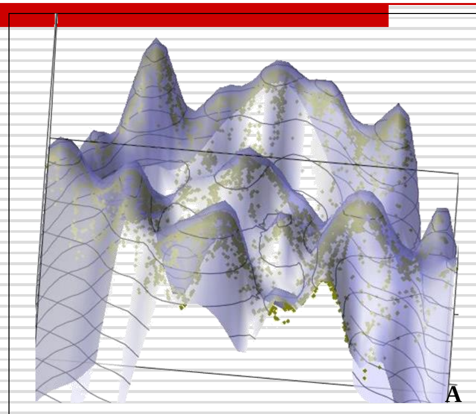


Fig. 2A Comparison of SDSS data points with the derived DQC potential. The potential is plotted upside down, and the yellow data points are slightly shifted in order to increase their visibility.

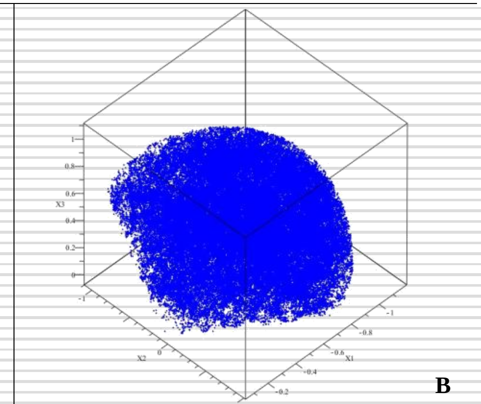


Fig. 2B The distribution of data in a 3D space defined by θ φ and z.

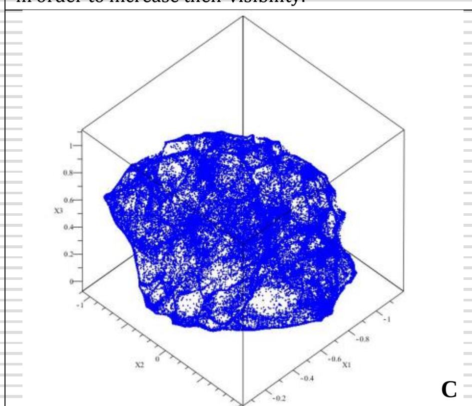


Fig. 2C Early stage of DQC evolution of the data

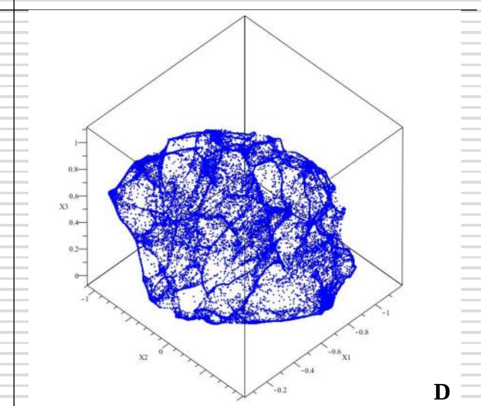


Fig. 2D Further DQC evolution exhibits the clear appearance of string-like structures.

EXAMPLE :NANO-CHEMISTRY

UNBIASED ANALYSIS OF X-RAY ABSORPTION DATA

Data collected at the Stanford Synchrotron Radiation Lightsource (SSRL), using the TXM-XANES microscope, a new device that enables an efficient study of hierarchically complex materials

Marvin Weinstein, Florian Meirer,,
Allison Hume, Phillipe Sciau,
David Horn, Apurva Mehta

Very complex problem: Interface of materials

- Sample data: Roman pottery
  - Red and Black colors are due to different iron oxides
- Similar problems:
  - Lithium-ion batteries
  - Catalyst breakdown

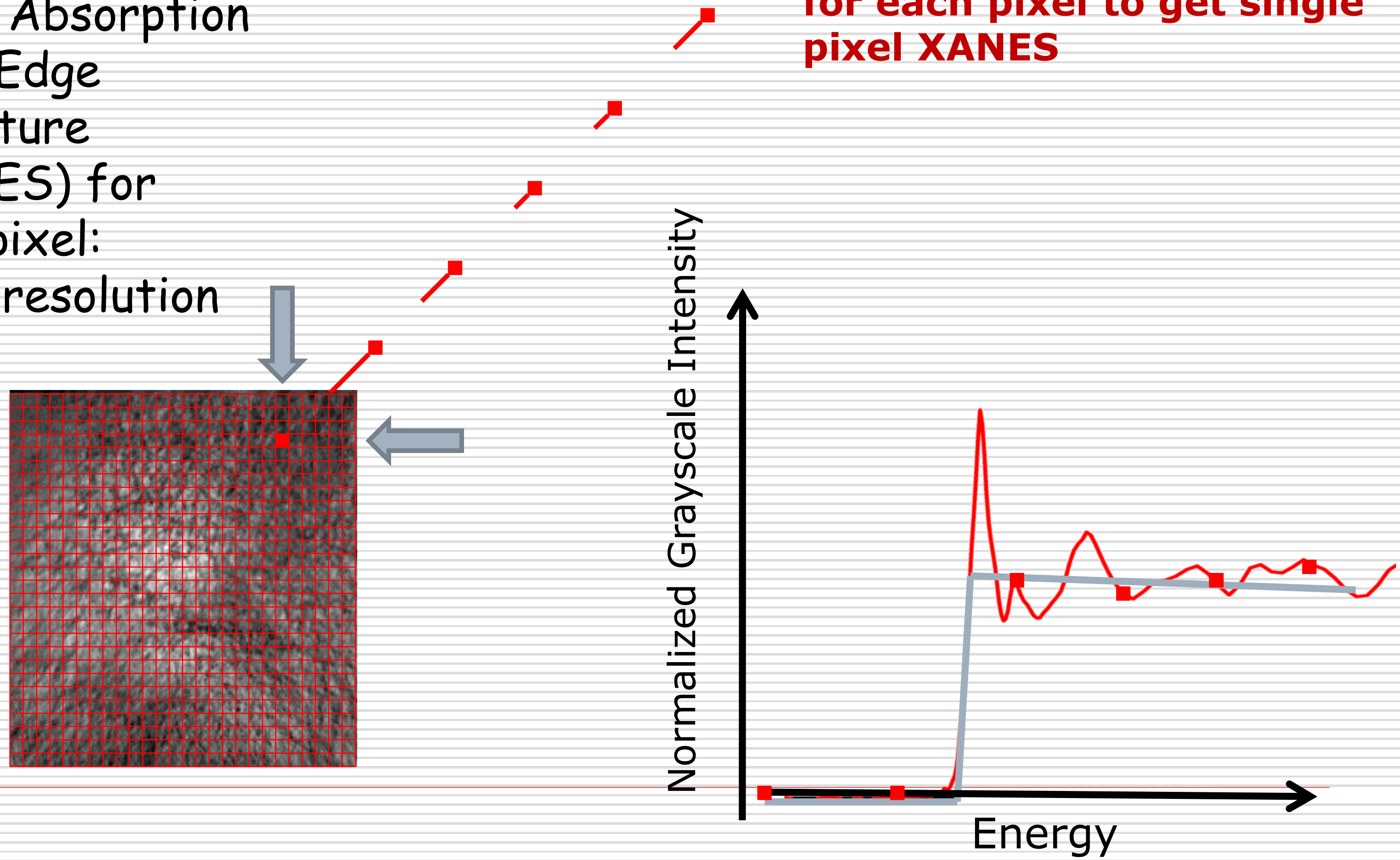
What Will We Learn?

- This is a big, noisy dataset
  - 669,000 x-ray absorption spectra at 148 energies (the energies = features)
  - Full of experimental artifacts
- Goal
  - To group spectra into similar shapes, because the shape correlates with the iron oxide present in the sample
  - There is a needle in this haystack!
- Requirement
  - To do this without assumptions (i.e. in an unsupervised manner)

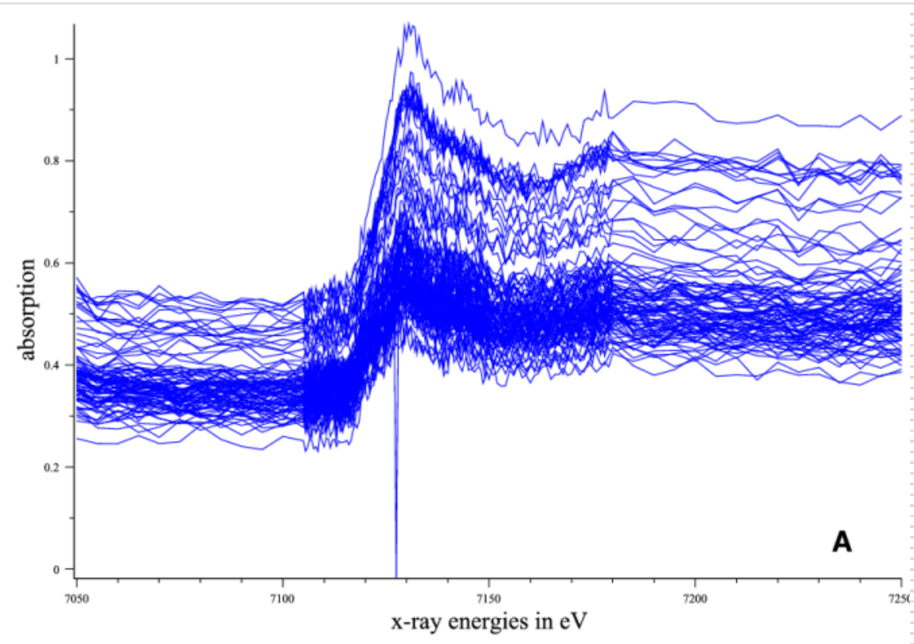
TXM-Xanes

- **Collect one high resolution absorption image at each energy**
- **Trace the absorption value for each pixel to get single pixel XANES**

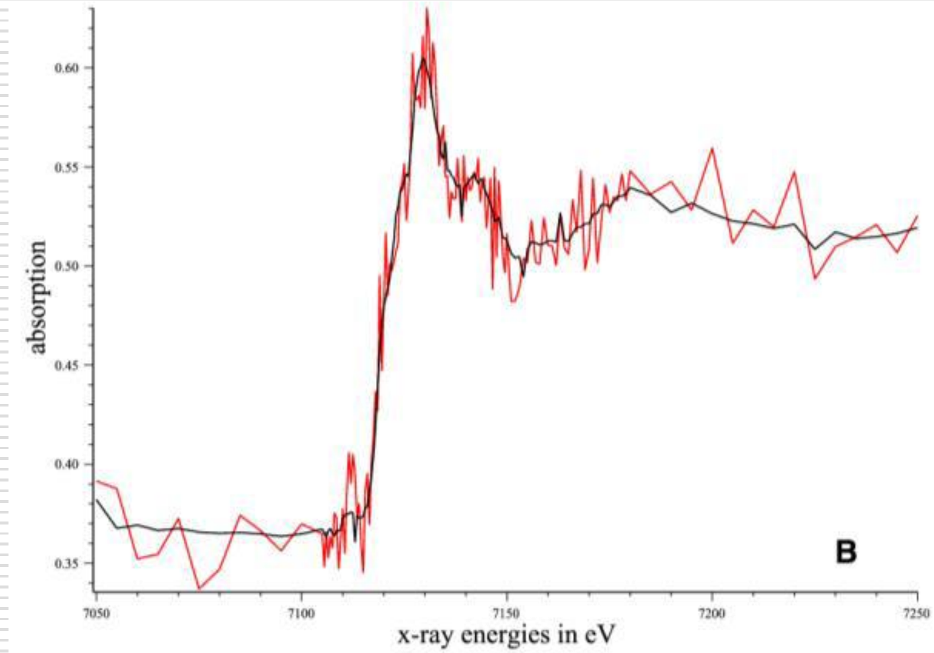
X-ray Absorption Near Edge Structure (XANES) for each pixel: 30nm resolution



Applying SVD reduction from 146 to 5 dimensions reduces much of the noise in the X-ray absorption spectrum



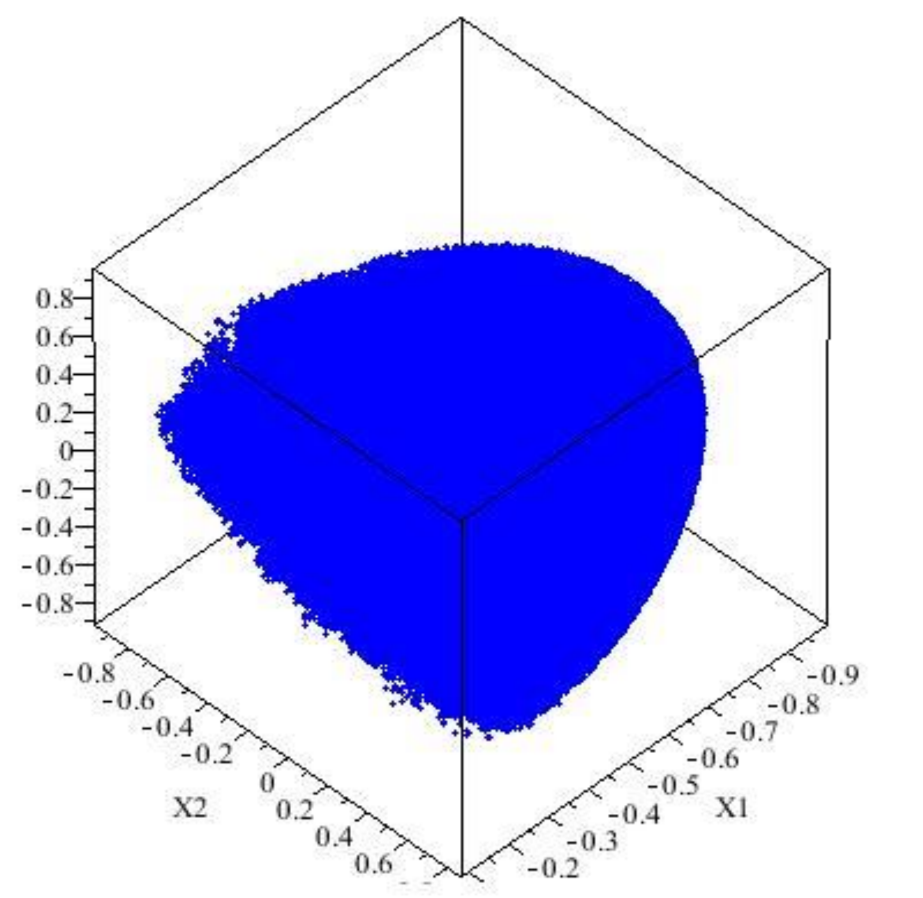
collection of raw data



Red – a typical curve
Black-after noise reduction

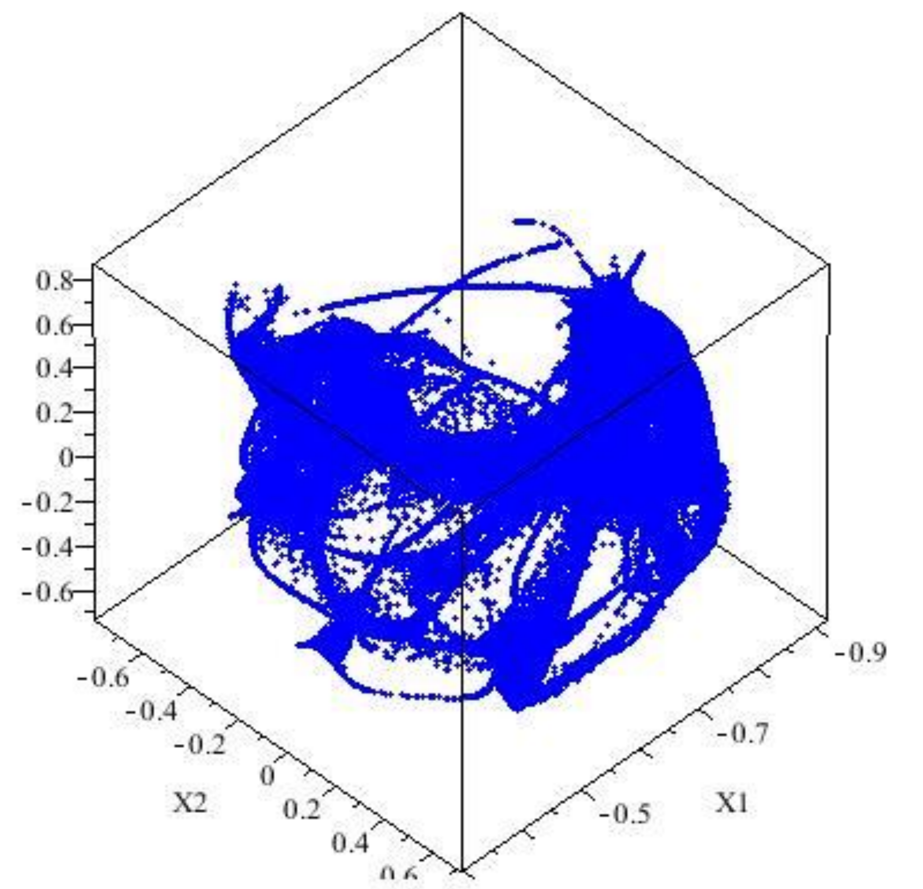
669,000 points in 5 dim feature space, projected onto a unit sphere

Clustering Process:



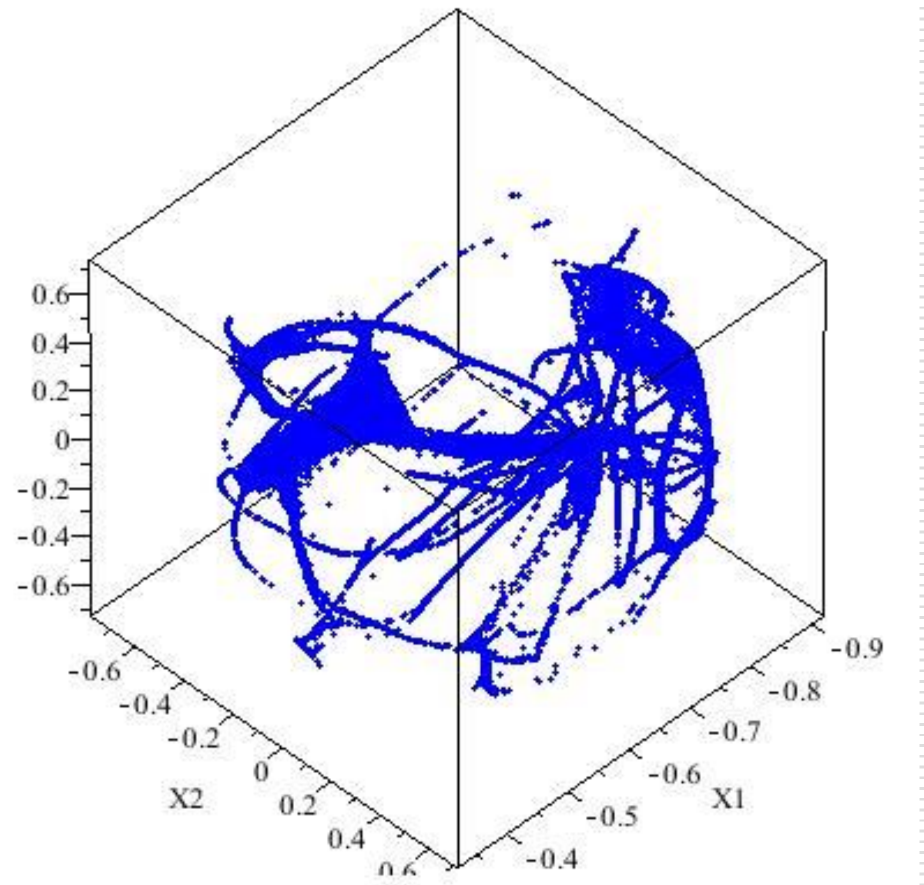
Clustering Process:

Data collapses into clumps and strands



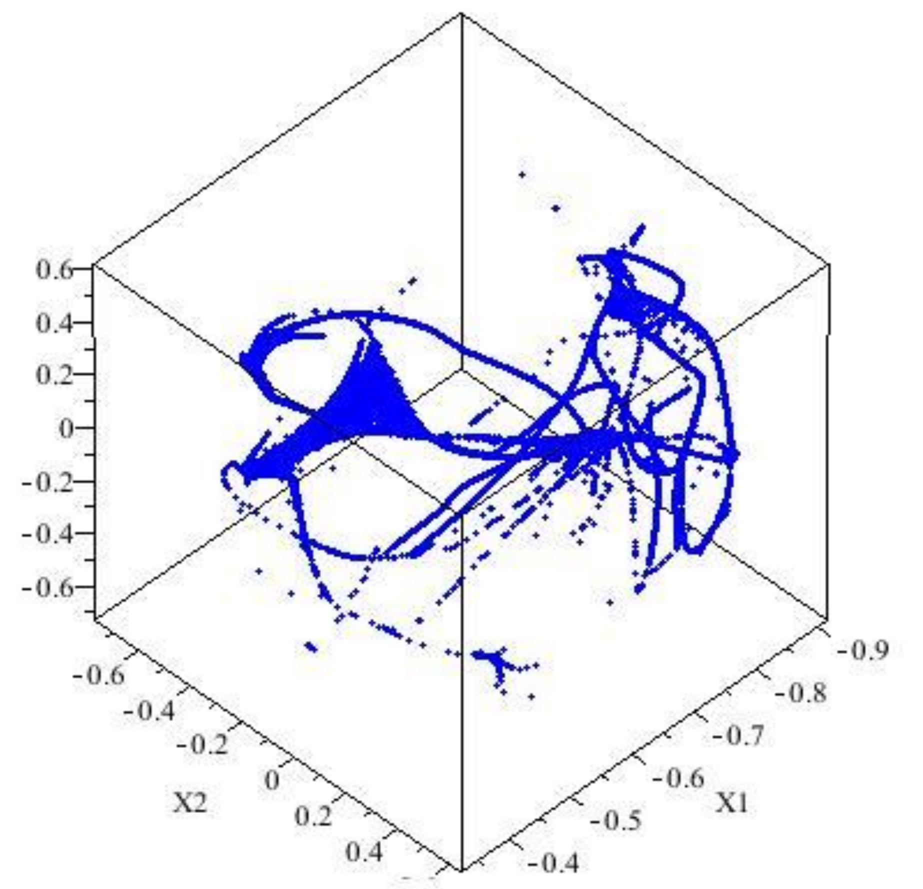
Clustering Process:

Data collapses into clumps and strands



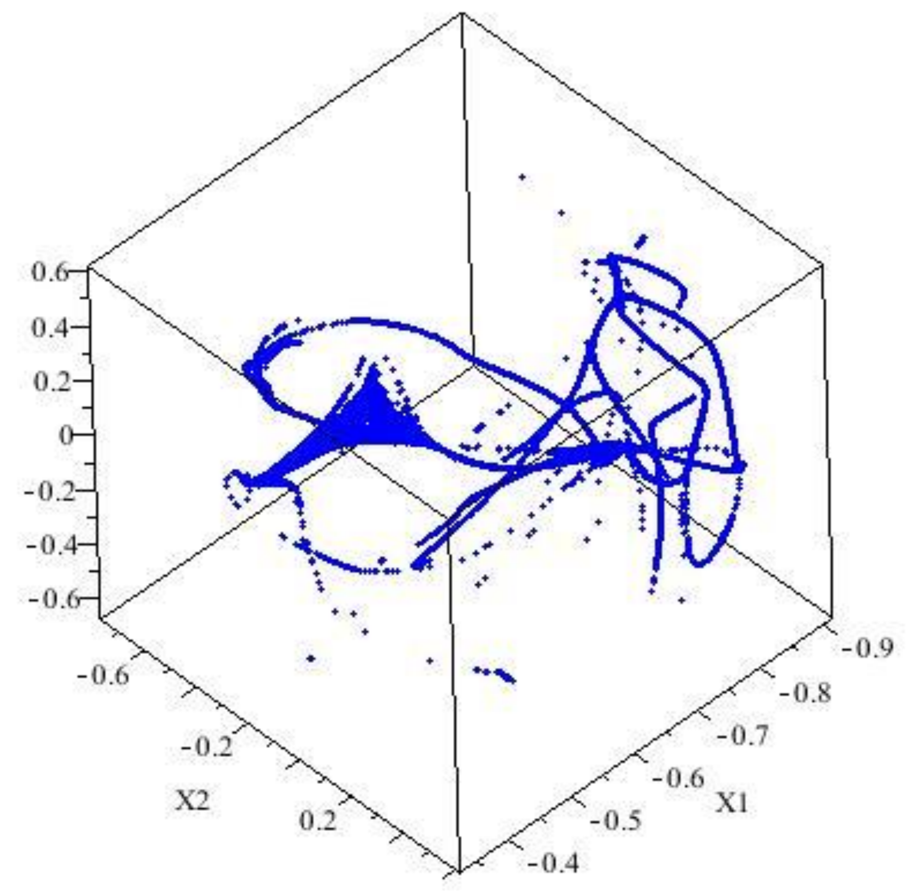
Clustering Process:

Some strands collapse to points, others remain



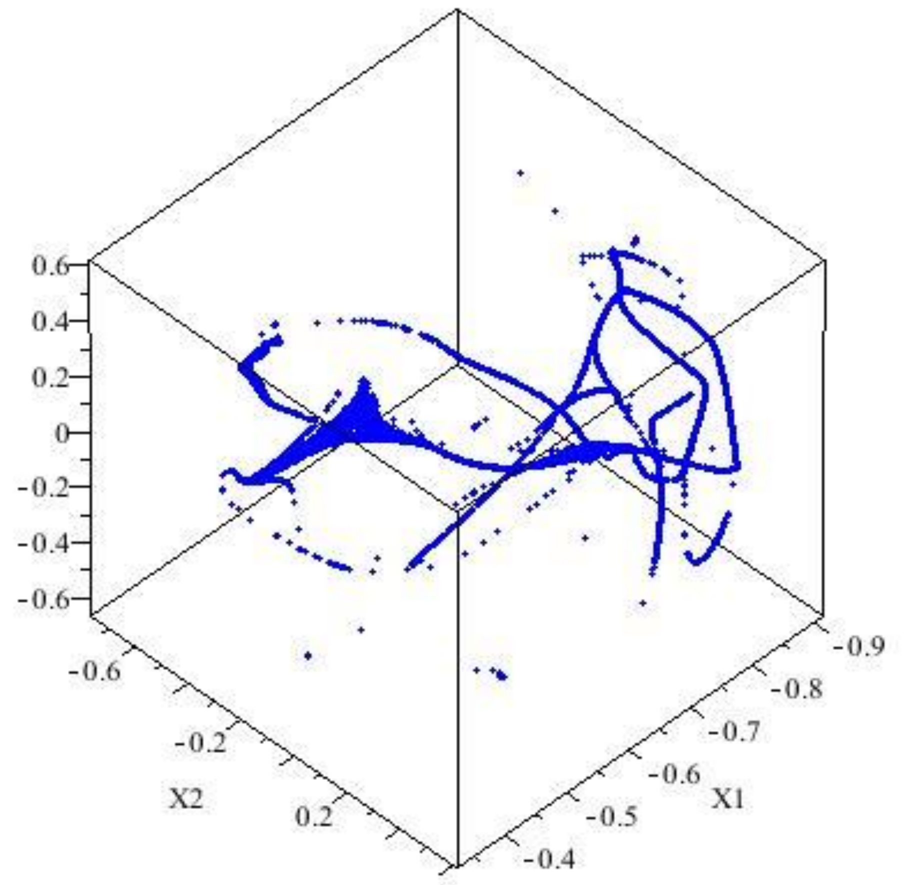
Clustering Process:

Some strands collapse to points, others remain



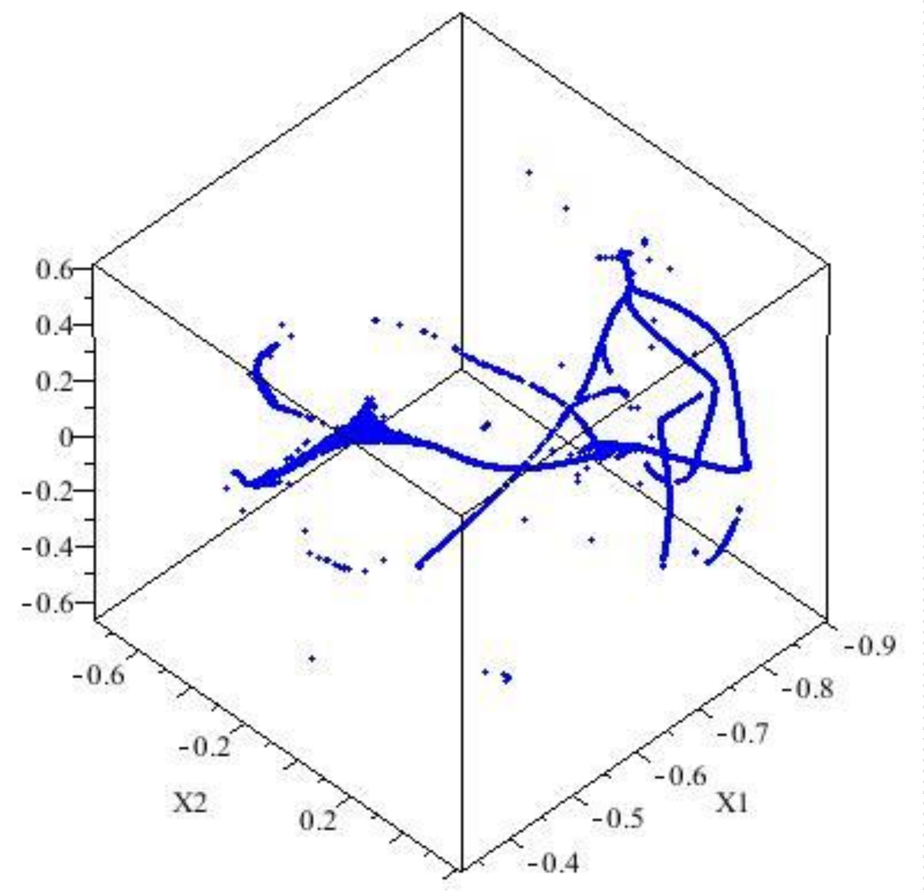
Clustering Process:

Separation continues



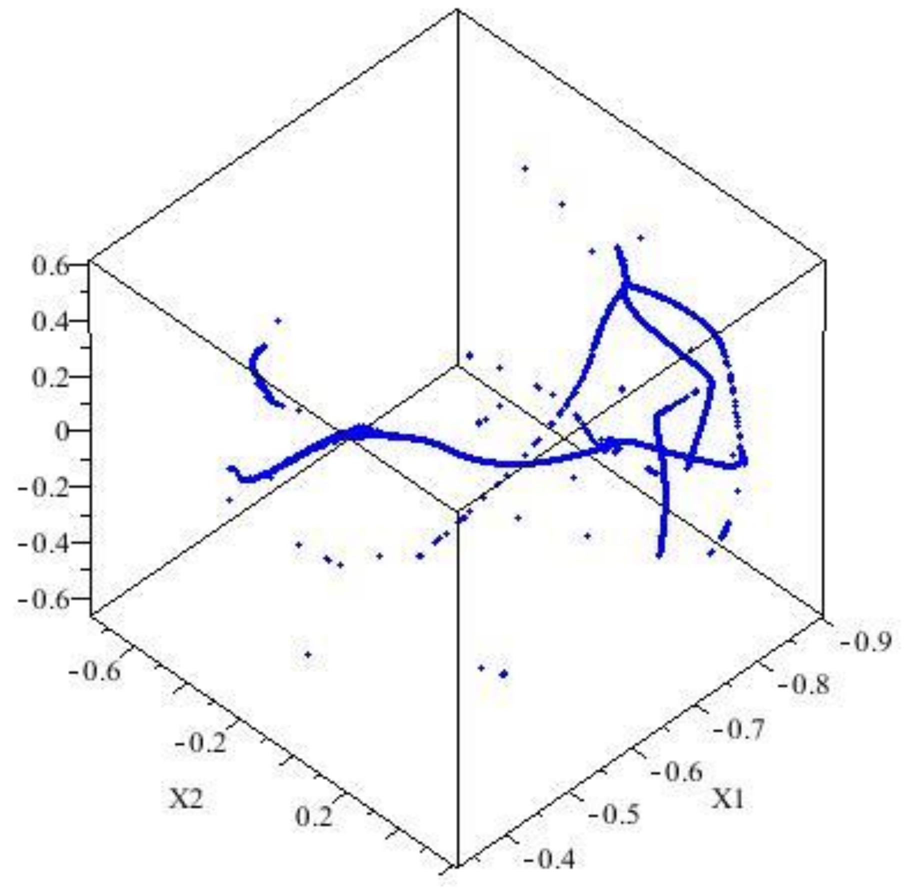
Clustering Process:

Separation continues



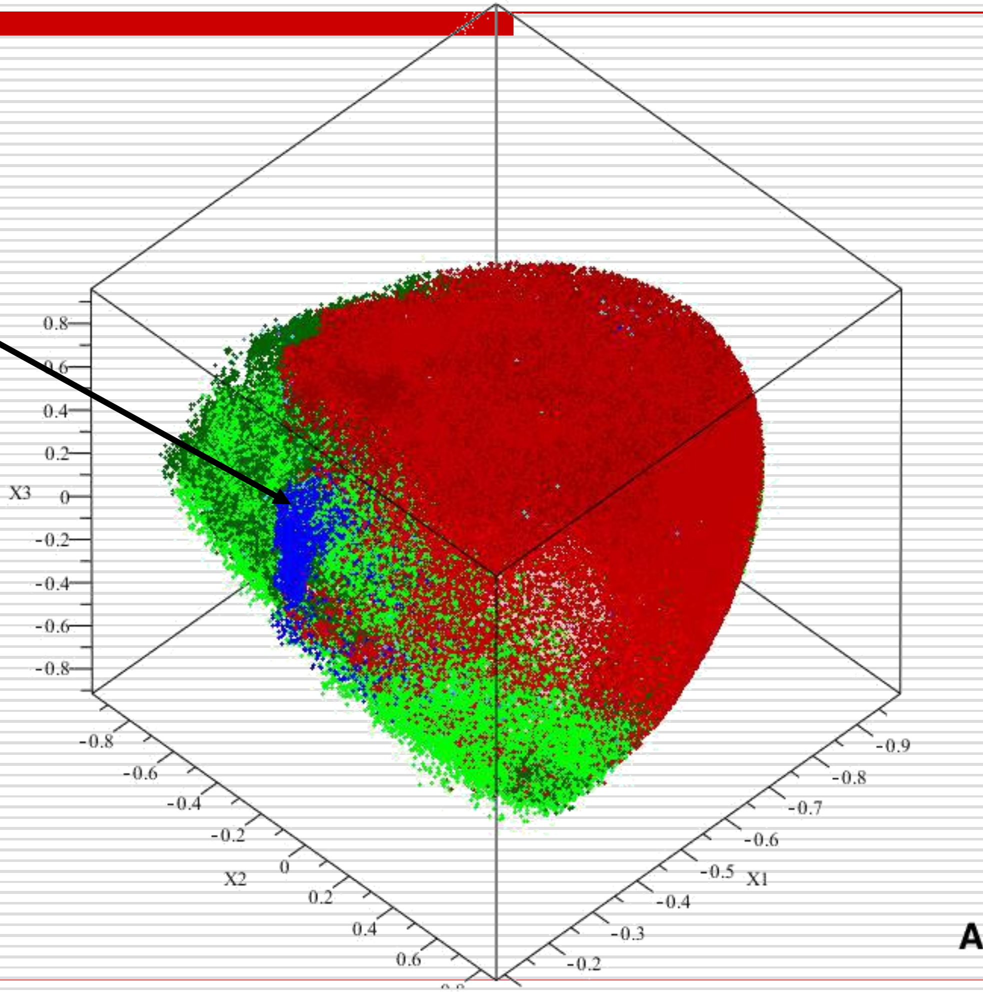
Clustering Process results in Structures and Point Clusters

Identify each connected string as a different structure. Color each structure differently.



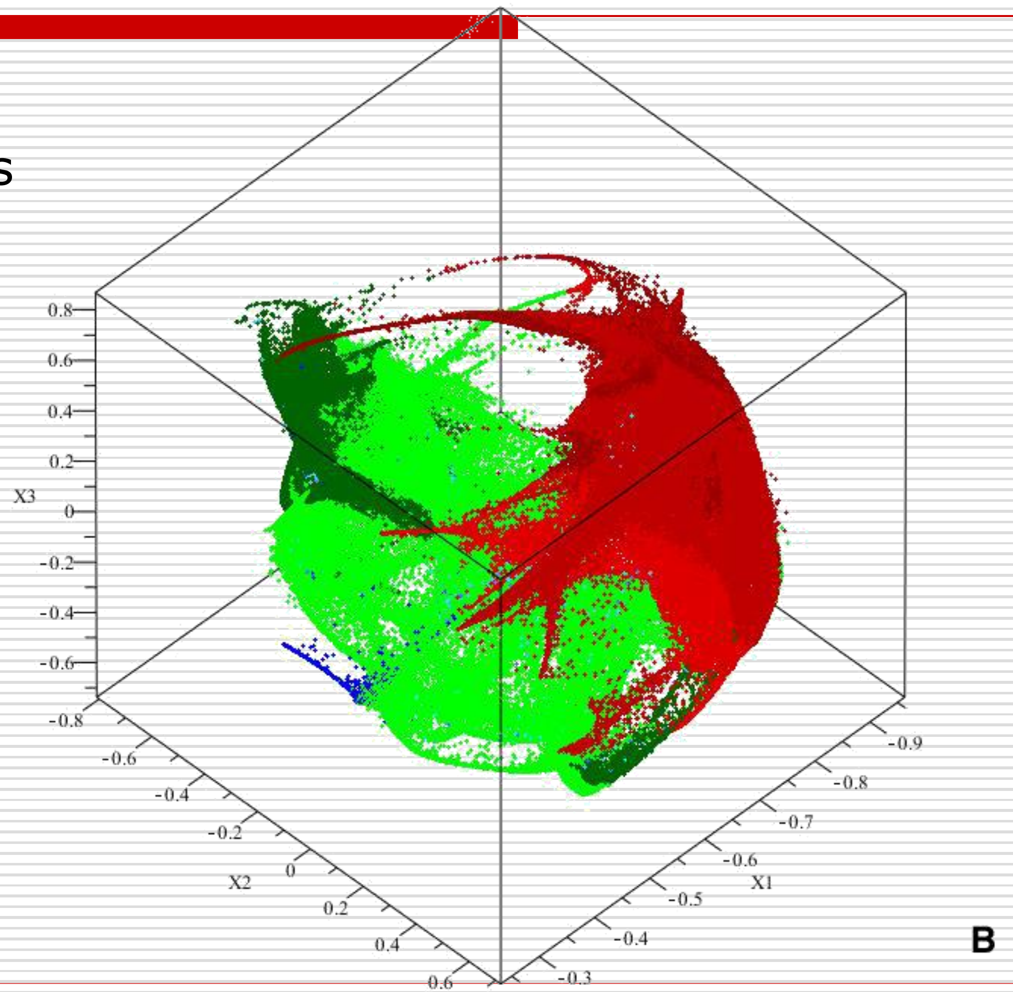
Same after coloring

The needle in the haystack



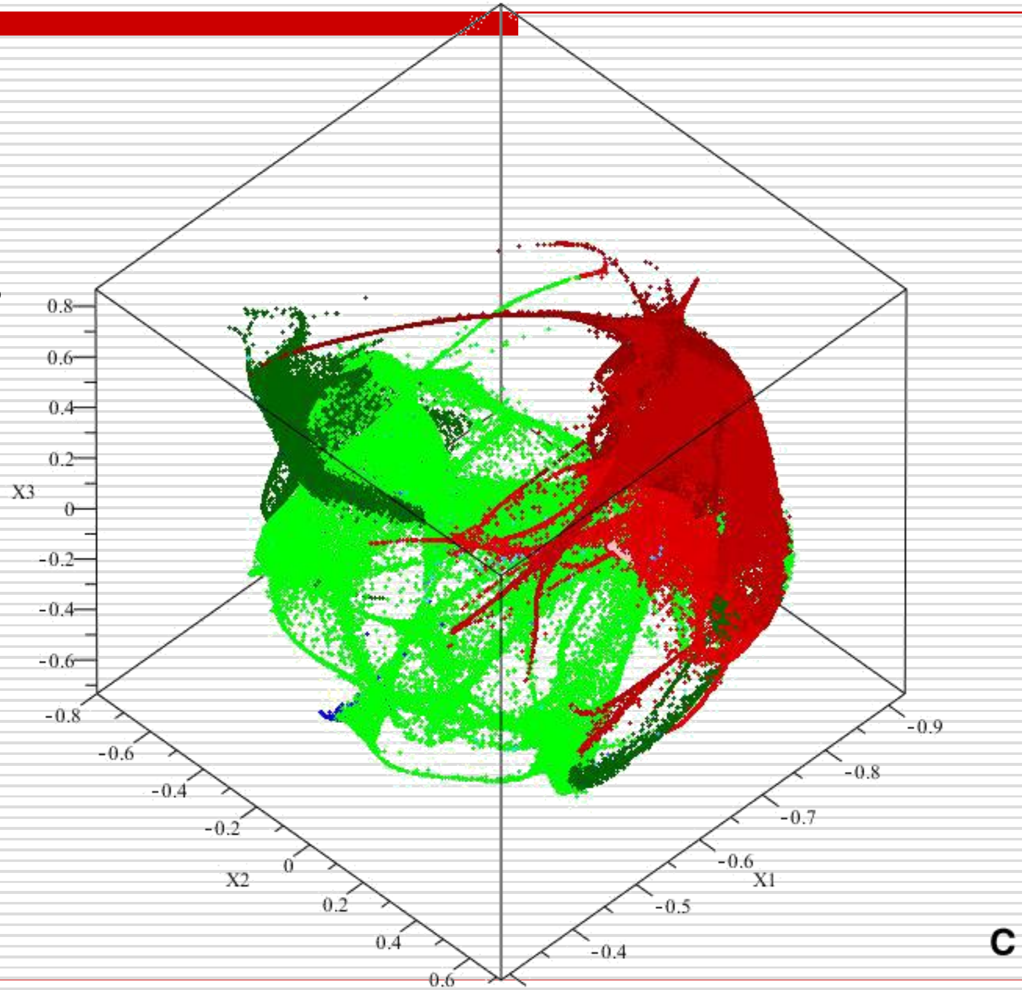
Clustering Process

Data collapses into clumps and strands. The blue data swiftly separates



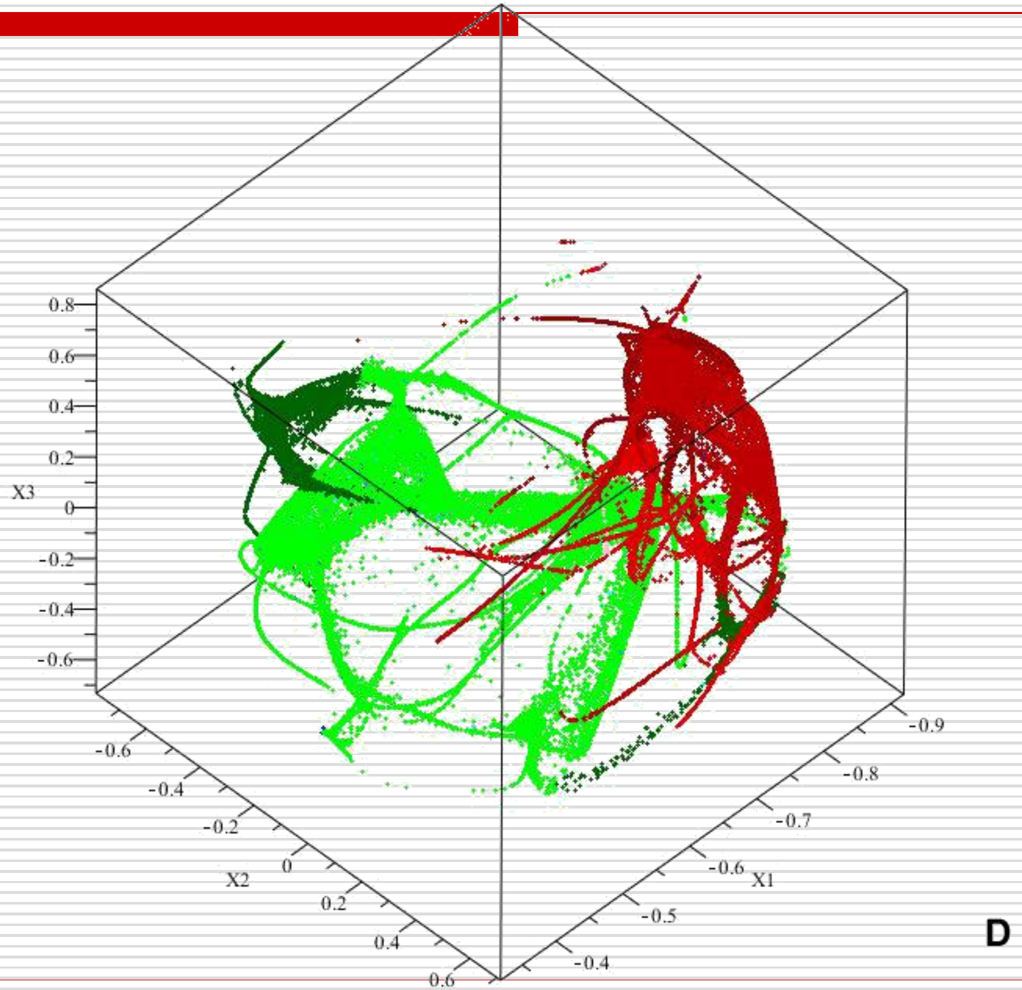
Clustering Process

Data collapses into clumps and strands



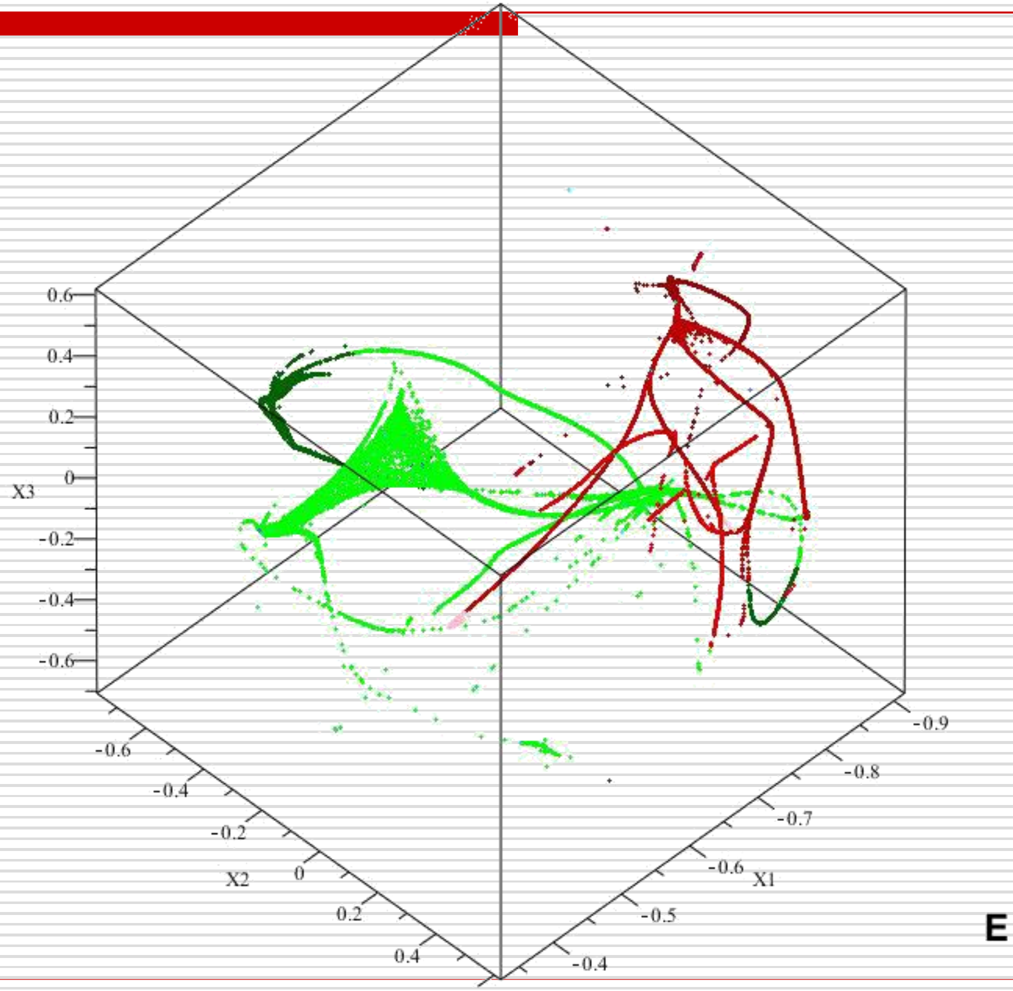
Clustering Process

Some strands collapse to points, others remain



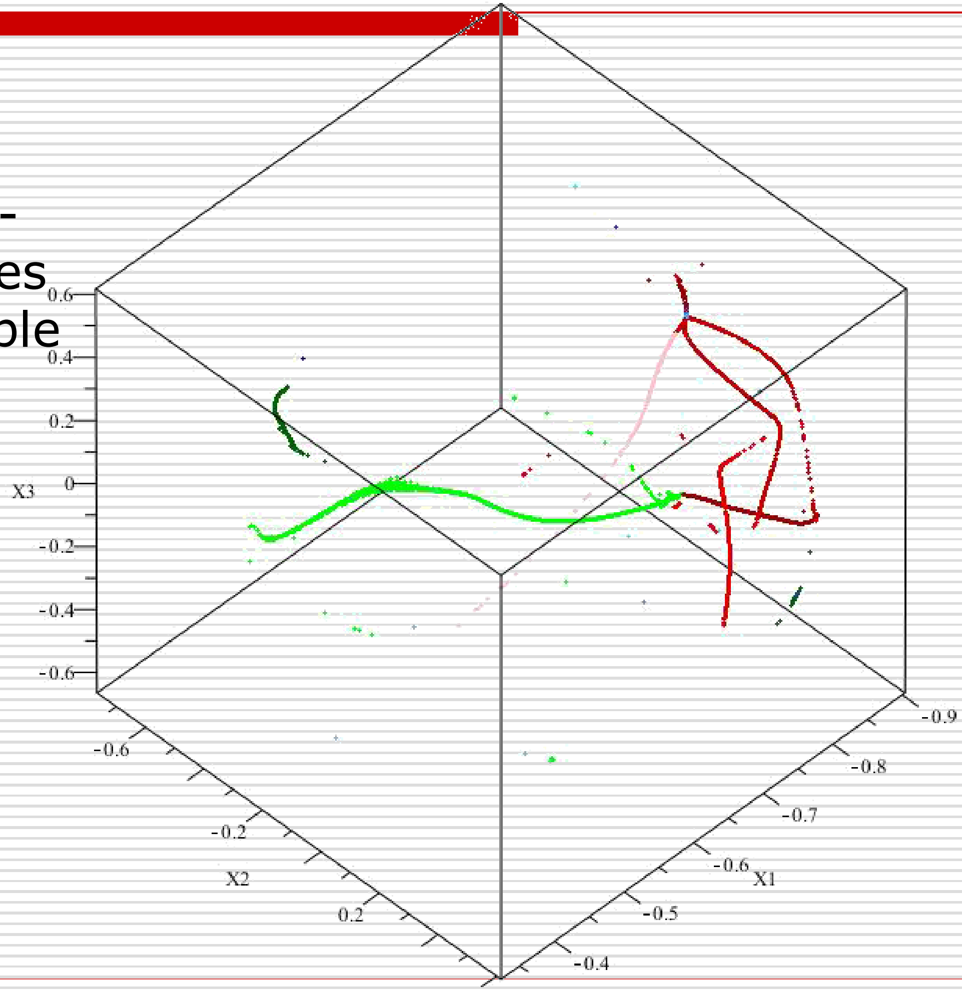
Clustering Process

Some strands collapse to points, others remain



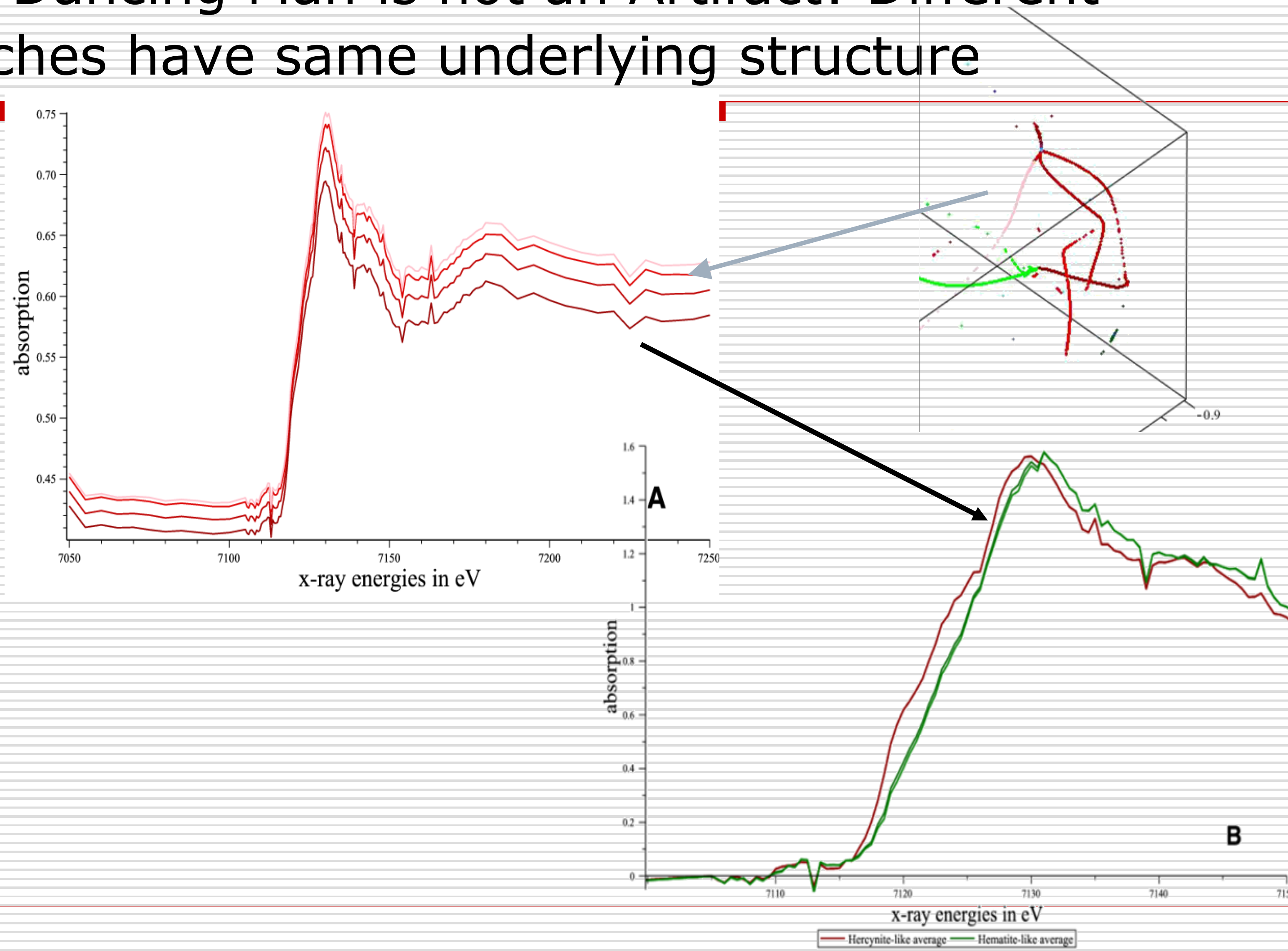
Clustering Process

Separation continues leading to non-trivial structures as well as simple clusters

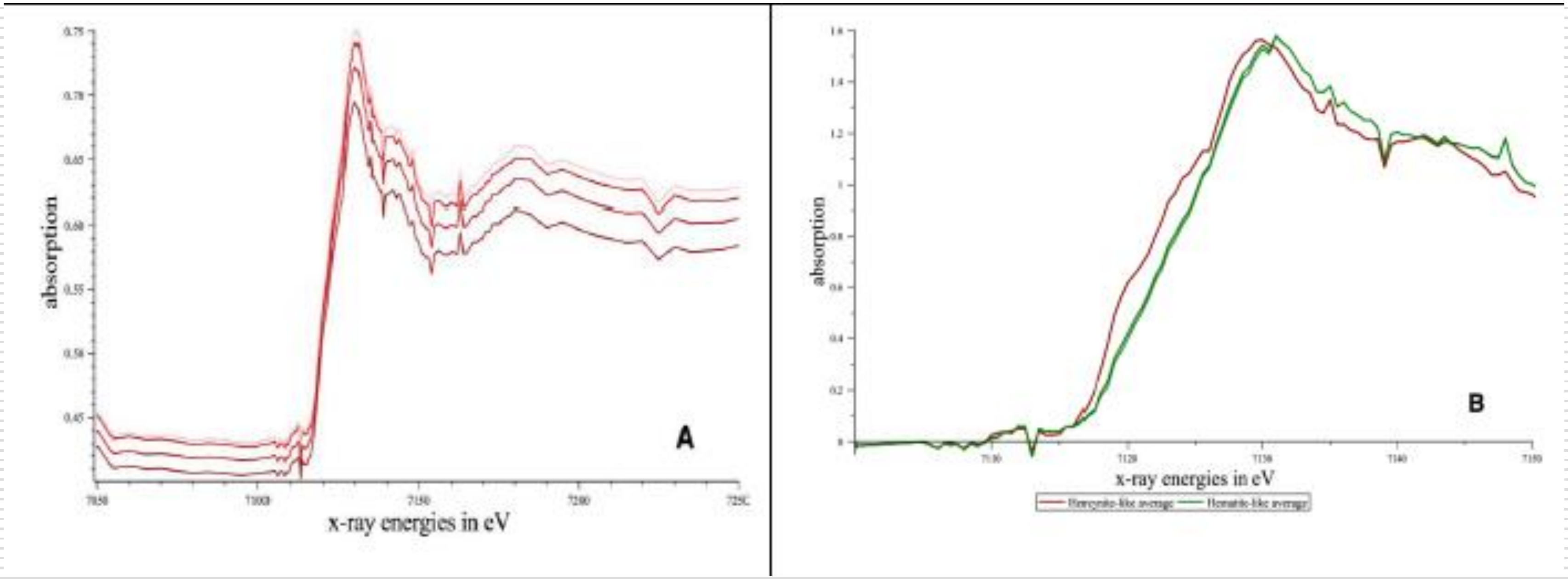


F

The Dancing Man is not an Artifact: Different branches have same underlying structure



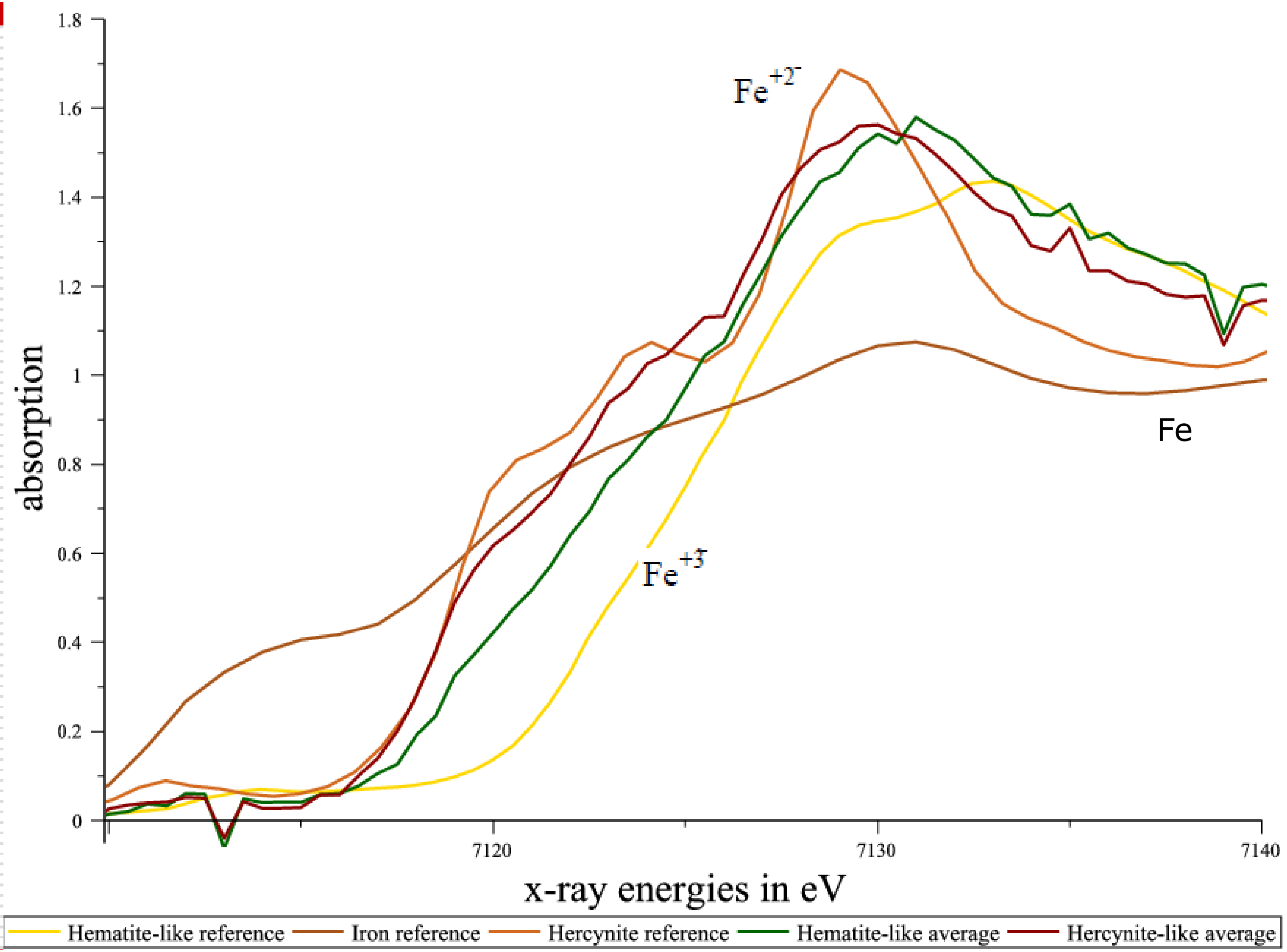
Averaging and normalizing original spectra



Left: **averages** of absorption spectra of the **raw data** for the different arms of the “dancing man” structure. The averages remove most of the noise. The resulting spectra differ by the value of the pre-edge and the edge-jump.
Right: The overlay of all **normalized** averages of red structures
(arms of the “dancing man”) lie to the left of the averages constructed from the two different green structures.

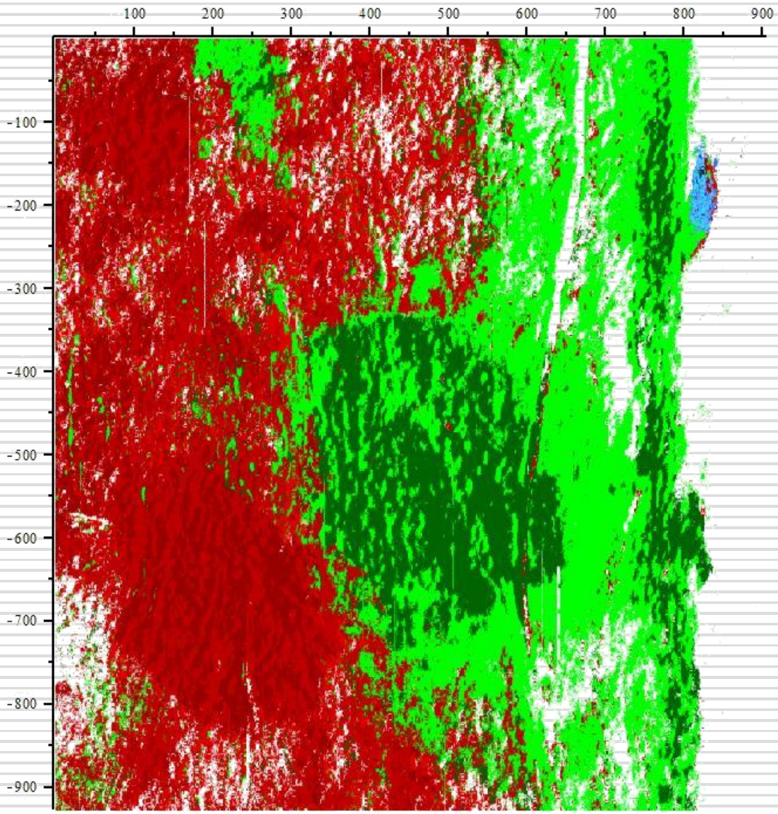
Normalization procedure: subtract average of 20 lowest energy points, rescale so that average of 20 highest energy points is 1.

A comparison of average spectra for the hematite-like (green) and hercynite-like (red) clusters to reference spectra for hematite (Fe^{+3}) hercynite (Fe^{+2}), and Fe. All x-ray energies are in eV.

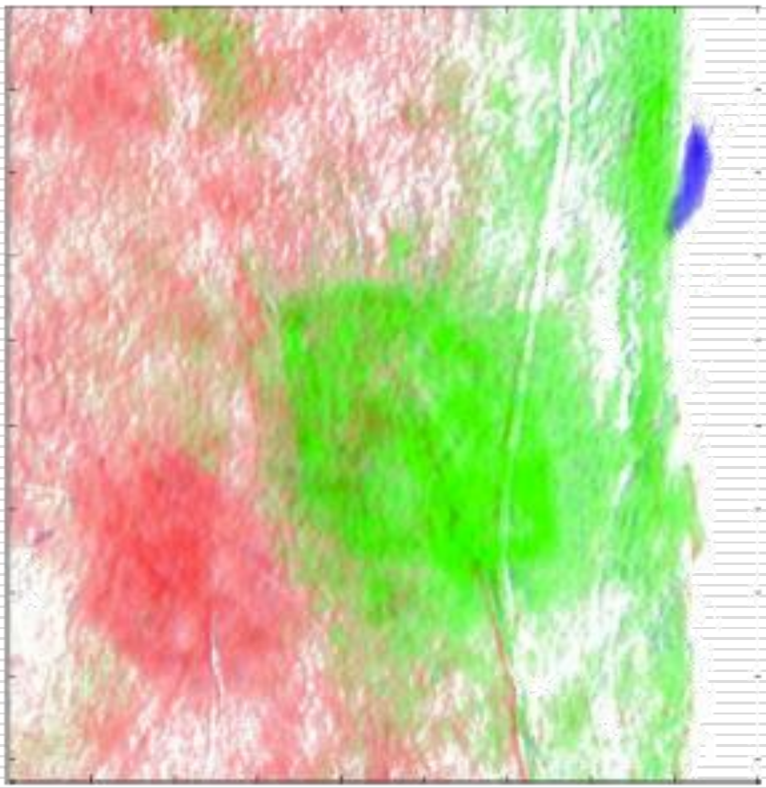


Hematite=Fe_2O_3 Hercynite=$FeAl_2O_4$

Comparing to Supervised Fit

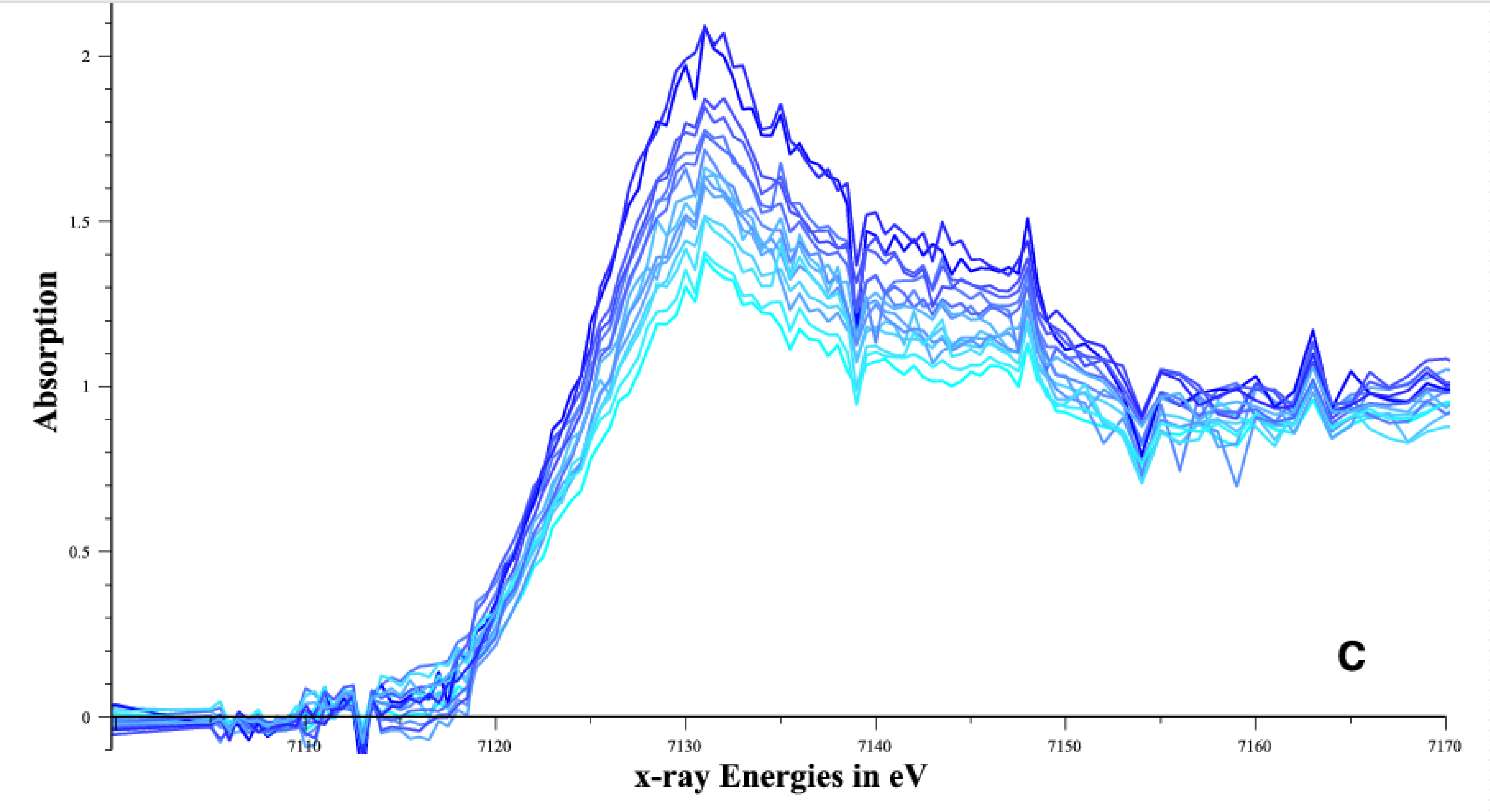


DQC results



supervised best fits to sum of 3 spectra

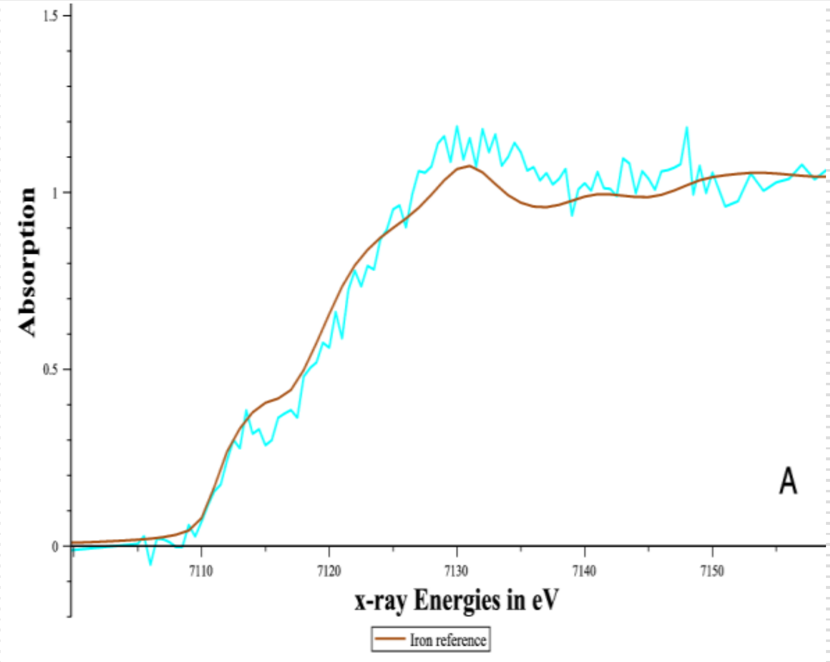
Other sub-clusters found in the blue cluster, showing the range of variation of the averages.

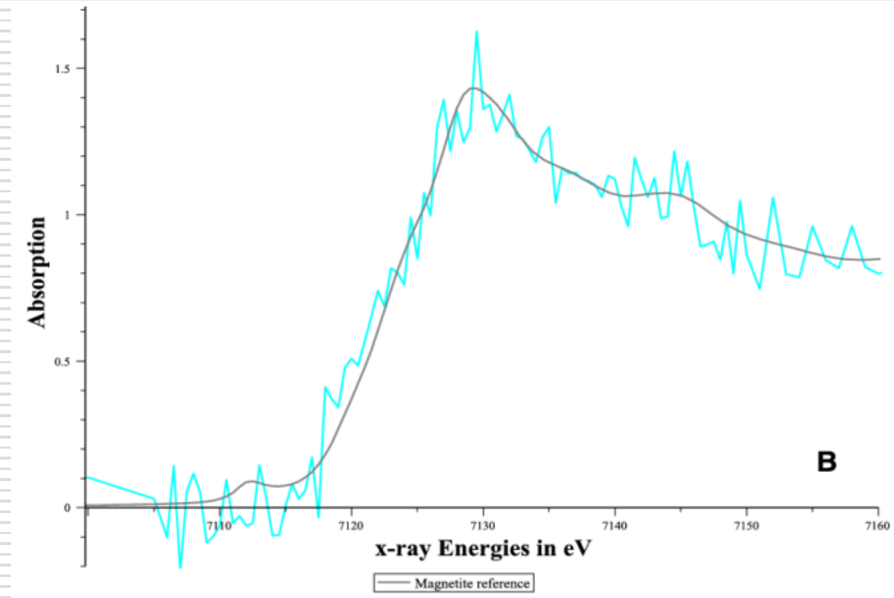


Zooming into the blue cluster:

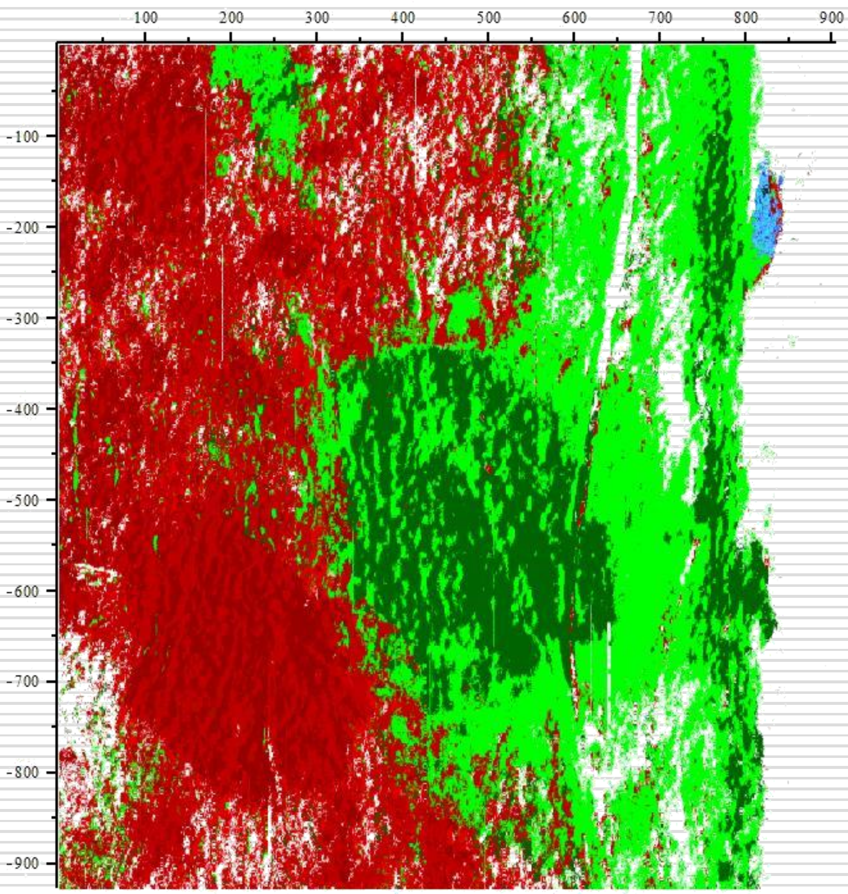
A) A reference iron spectrum compared to the spectrum of one of 60 points contained in an *iron*-like sub-cluster;

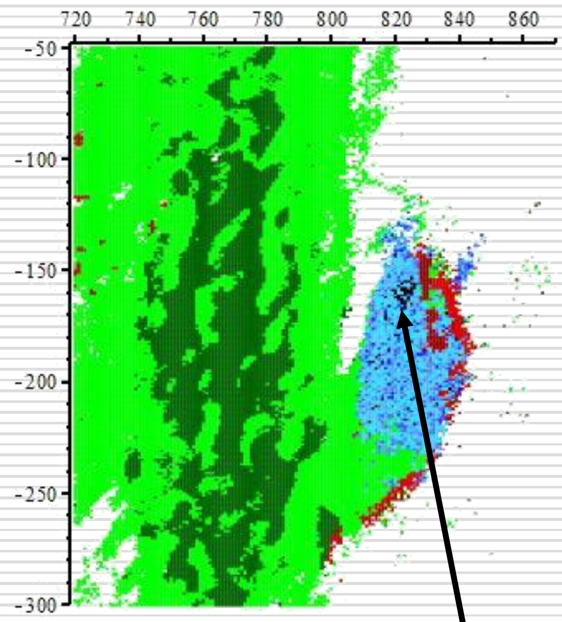
B) A reference magnetite spectrum compared to the one point of a small *magnetite*-like (Fe_3O_4)sub-cluster;





Back to the Sample





The real needle in the haystack: 69 points out of 669,000

Summing It Up

- DQC is a powerful, visual paradigm for exploring and analyzing big, complex datasets for *hidden structures/information*
  - Structures, as defined by DQC, are much more complex and information rich objects than simple clusters
- DQC is data agnostic, unbiased, doesn’t find structure in random data and works on data that resists analysis by simple clustering algorithms that preclude elongated structures

Another example: Analysis of earthquakes in the Middle East

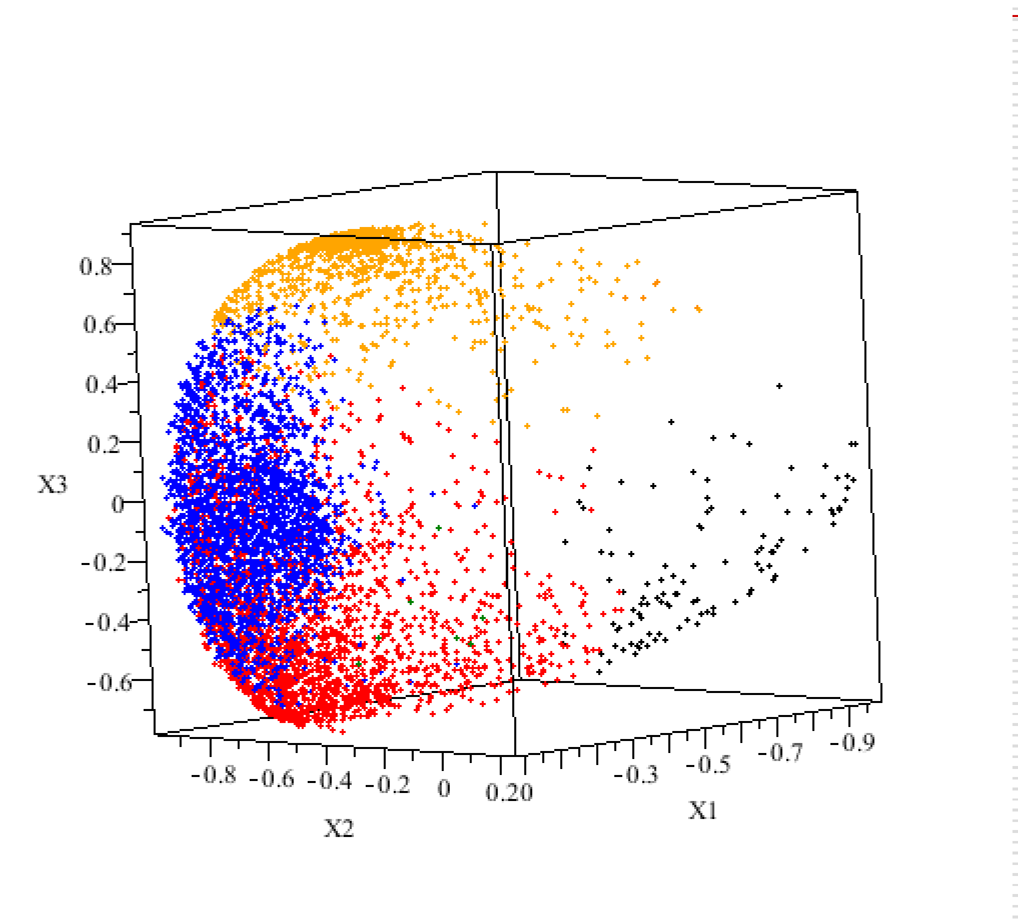
Data presented as list with five features:

- M_d – the (coda duration) magnitude of the earthquake
- M_0 – seismic moment of the earthquake
- Stress drop (difference in stress before and after)
- Radius of the fault broken by the earthquake
- f_0 - corner frequency beyond which the spectrum decreases like white noise

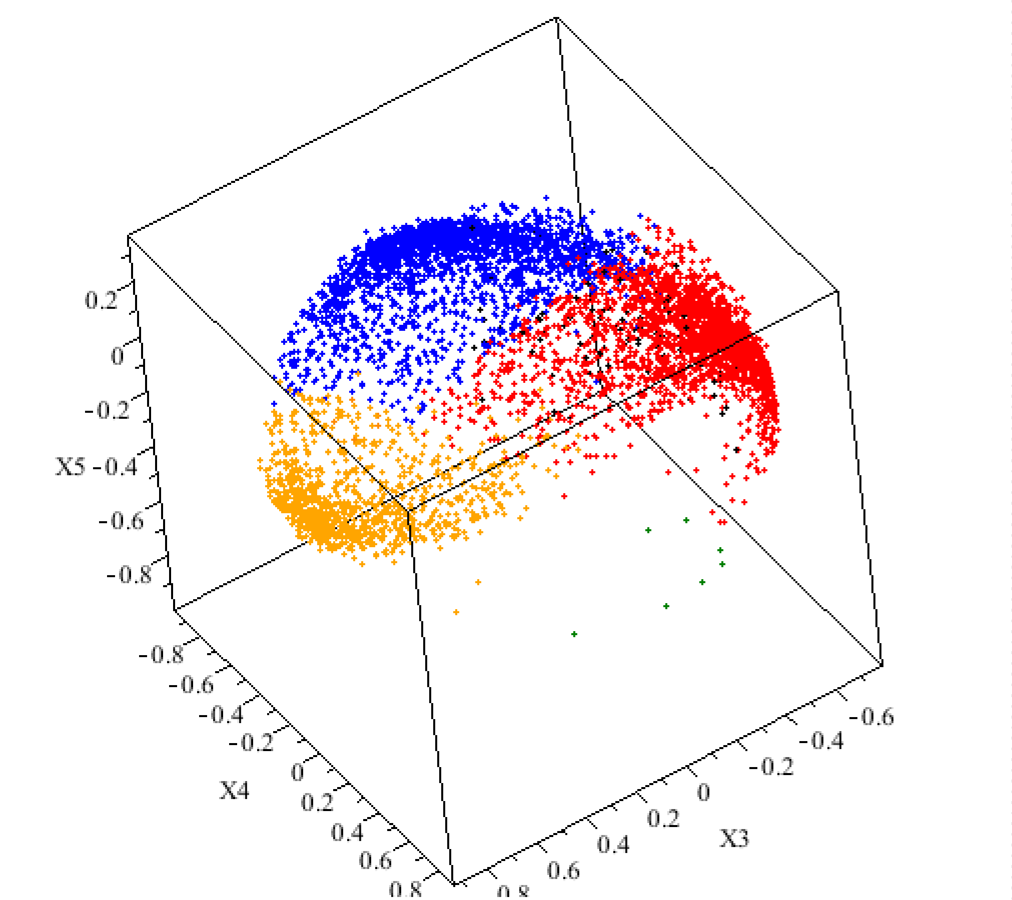
There exist 5700 events having all parameters.

Additional information: Events are located geographically in dimensions XY and are assigned a depth value Z. The recording time of each event T is specified.

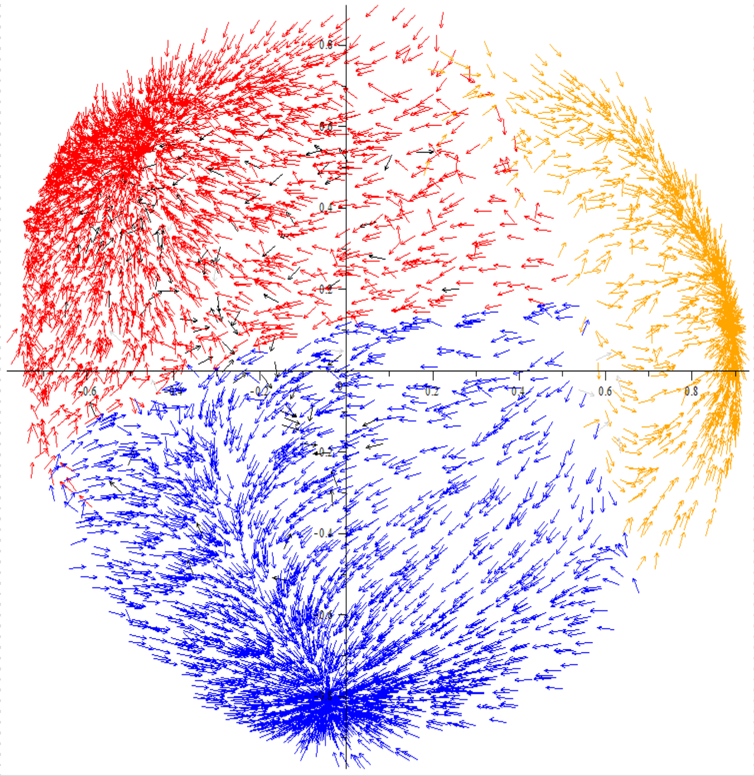
Convergence in SVD space (colors defined a-posteriori)



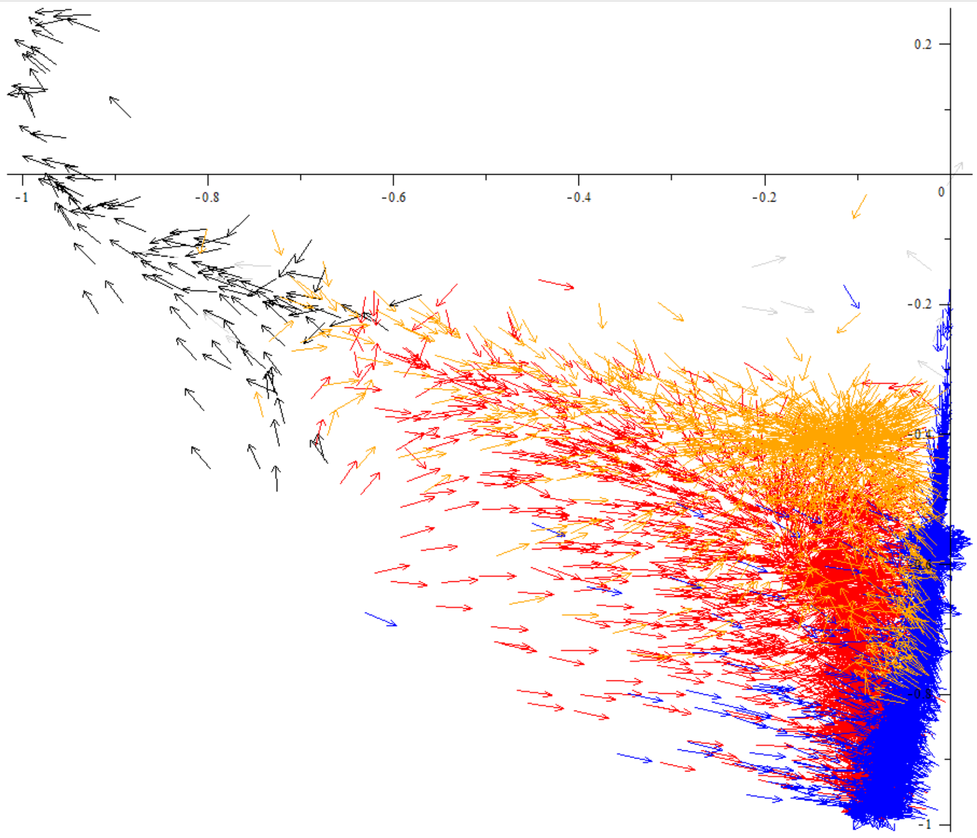
Convergence in SVD space



Unit vectors of –gradV colored according to final cluster classifications. Shown are PC=3 and 4



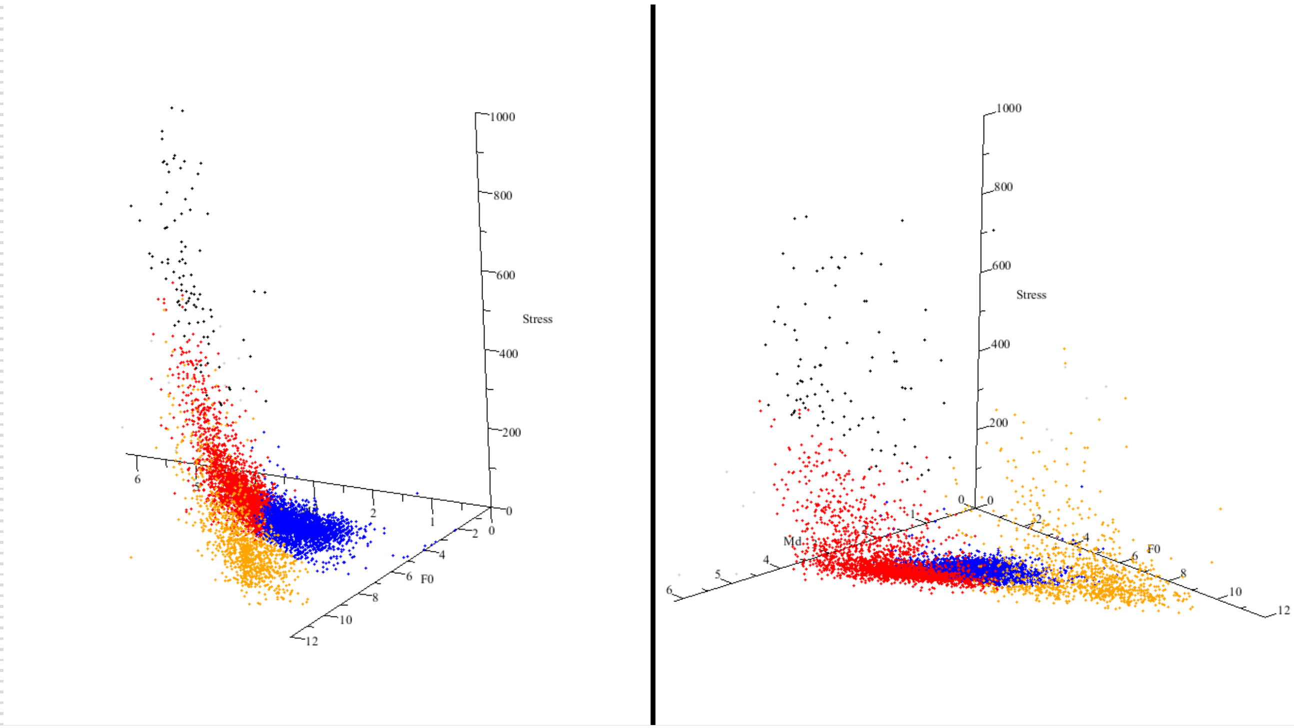
Unit vectors of –gradV colored according to final cluster classifications. Shown are PC=1 and 2



Different spaces:

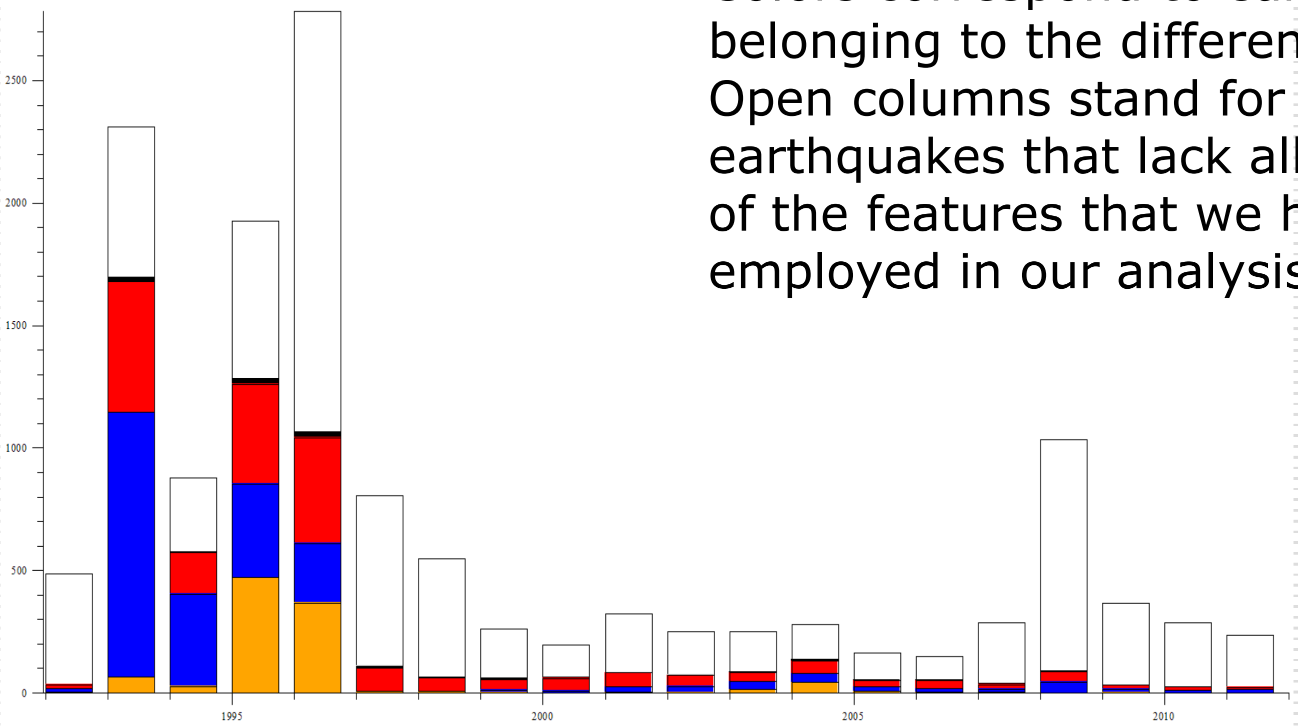
- 5-parameter SVD sphere, used for QC analysis.
- original 5-dimensional parameter space, used to identify meanings of the different clusters.
- XYZ indicates geometrical positions of events, classified into the different clusters as indicated by colors.
- the temporal dimension.

Distribution of earthquakes within the original feature space of M_d, f_0 and stress, shows that clusters touch each other.



Red: large magnitude. Blue: medium to low magnitude, low stress. Orange: medium to low magnitude, high stress.

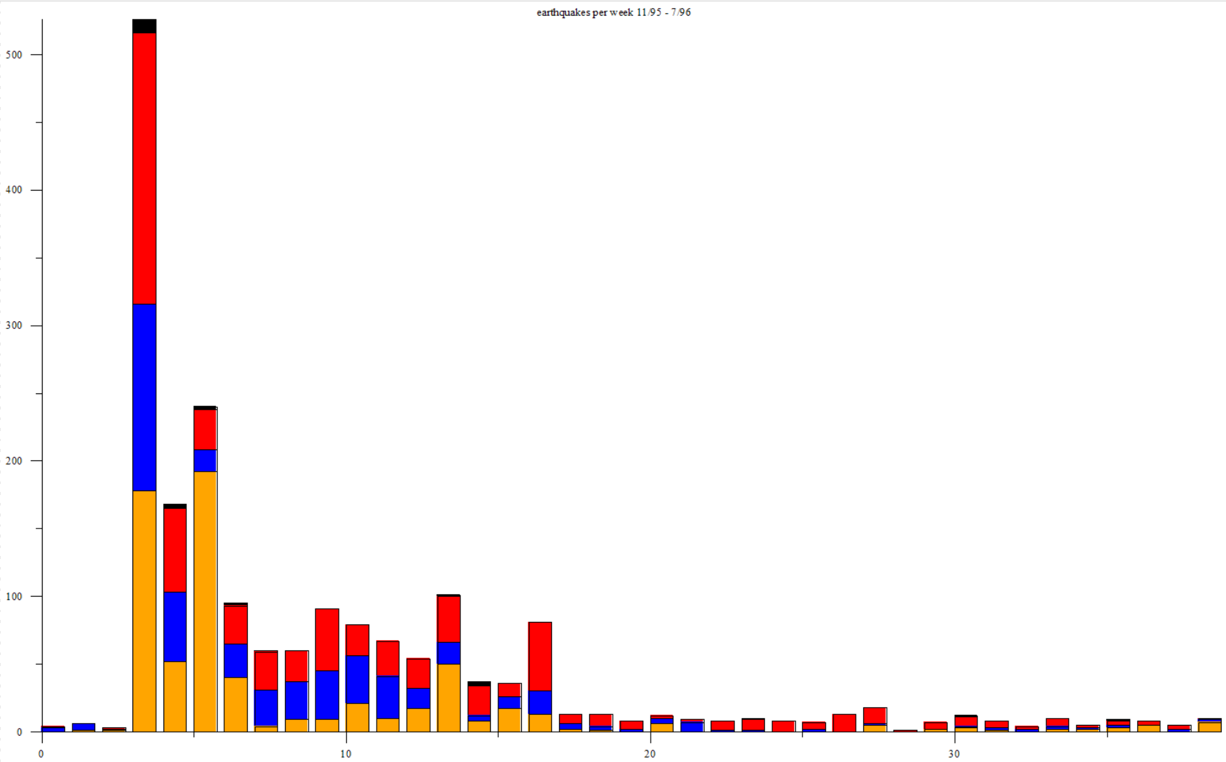
Temporal distribution: note orange concentration after the Nov 1995 earthquake



Colors correspond to earthquakes belonging to the different clusters. Open columns stand for registered earthquakes that lack all or some of the features that we have employed in our analysis.

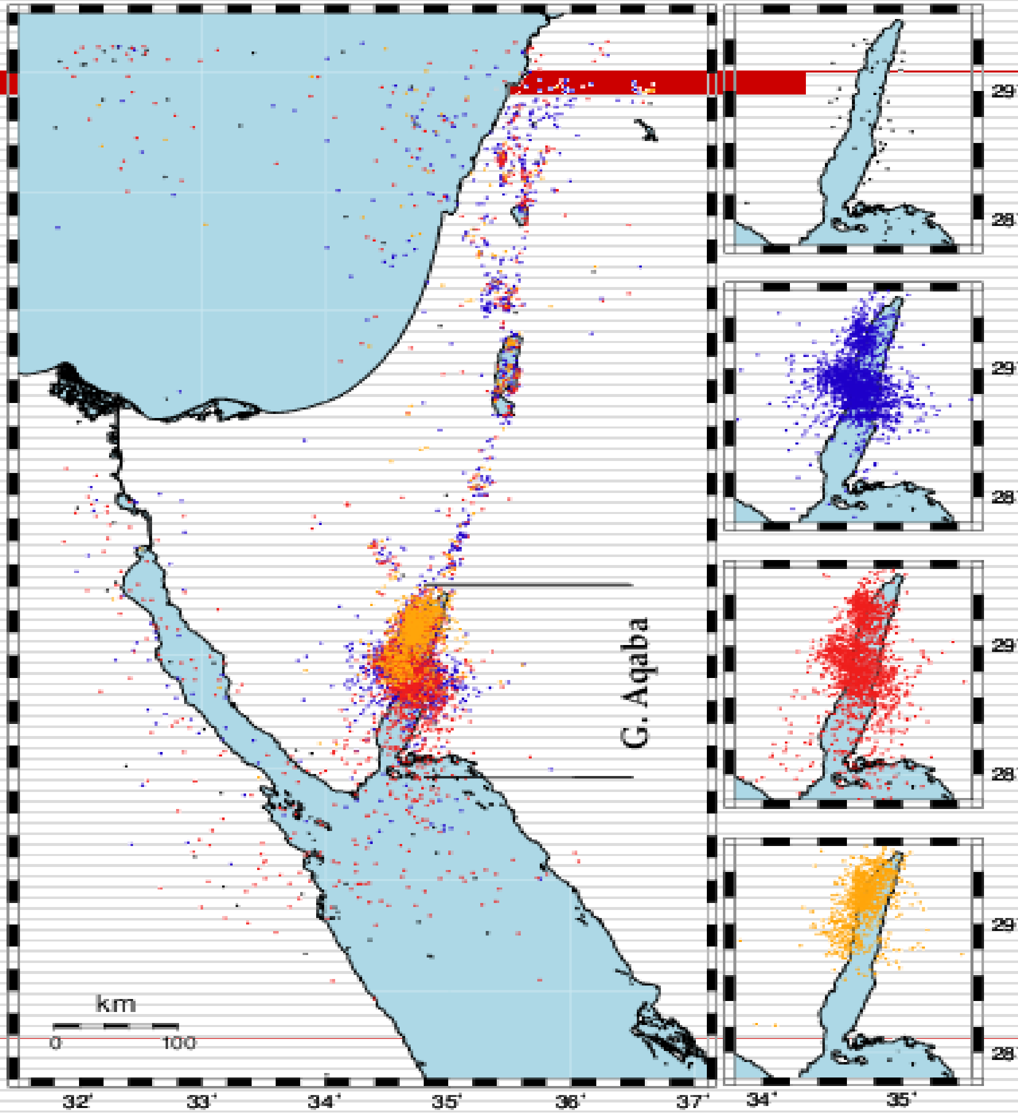
Yearly histogram of earthquakes throughout 1992-2011.

Weekly scale of the same distribution

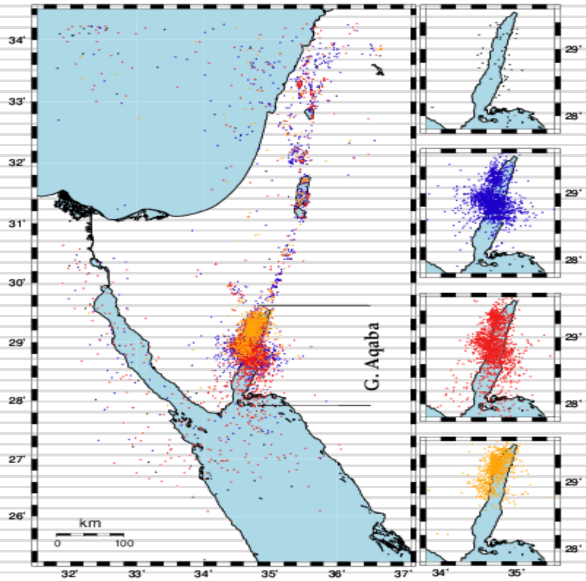


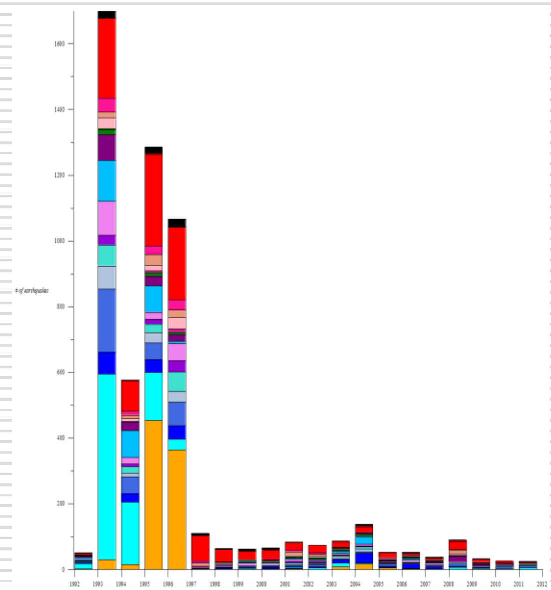
Nov 1995 – June 1996

Putting it on a geographic map



Conclusion of this analysis





Conclusion: orange events represent ruptures that have occurred following the major earthquake of November 1995. They have not been observed in such quantities in other geological activities in this region in many decades.

Thank you

Congratulations to Halina