Neural (and non neural) tools for data mining in massive data sets

Giuseppe Longo & Astroneural

Department of Physical Sciences
University “Federico II” in Napoli

I.N.F.N. – Napoli Unit
I.N.A.F. – Napoli Unit
longo@na.infn.it

January 2005, Edimburgh

The Astroneural Collaboration

Astroneural
DSF - DMI

Giuseppe Longo – DSF
Gennaro Miele – DSF
Roberto Tagliaferri – DMI
Roberto Amato (DSF – student)
Angelo Ciaramella (DMI – post Doc)
Carmine Del Mondo (DSF – fellow)
Lara de Vinco (DMI – fellow)
Ciro Donalek (DSF – Ph.D.)
Omar Laurino (DSF – student)
Gianpiero Mangano (INFN - senior)
Giancarlo Raiconi (DMI – senior)
Antonio Staiano (DMI - post Doc)

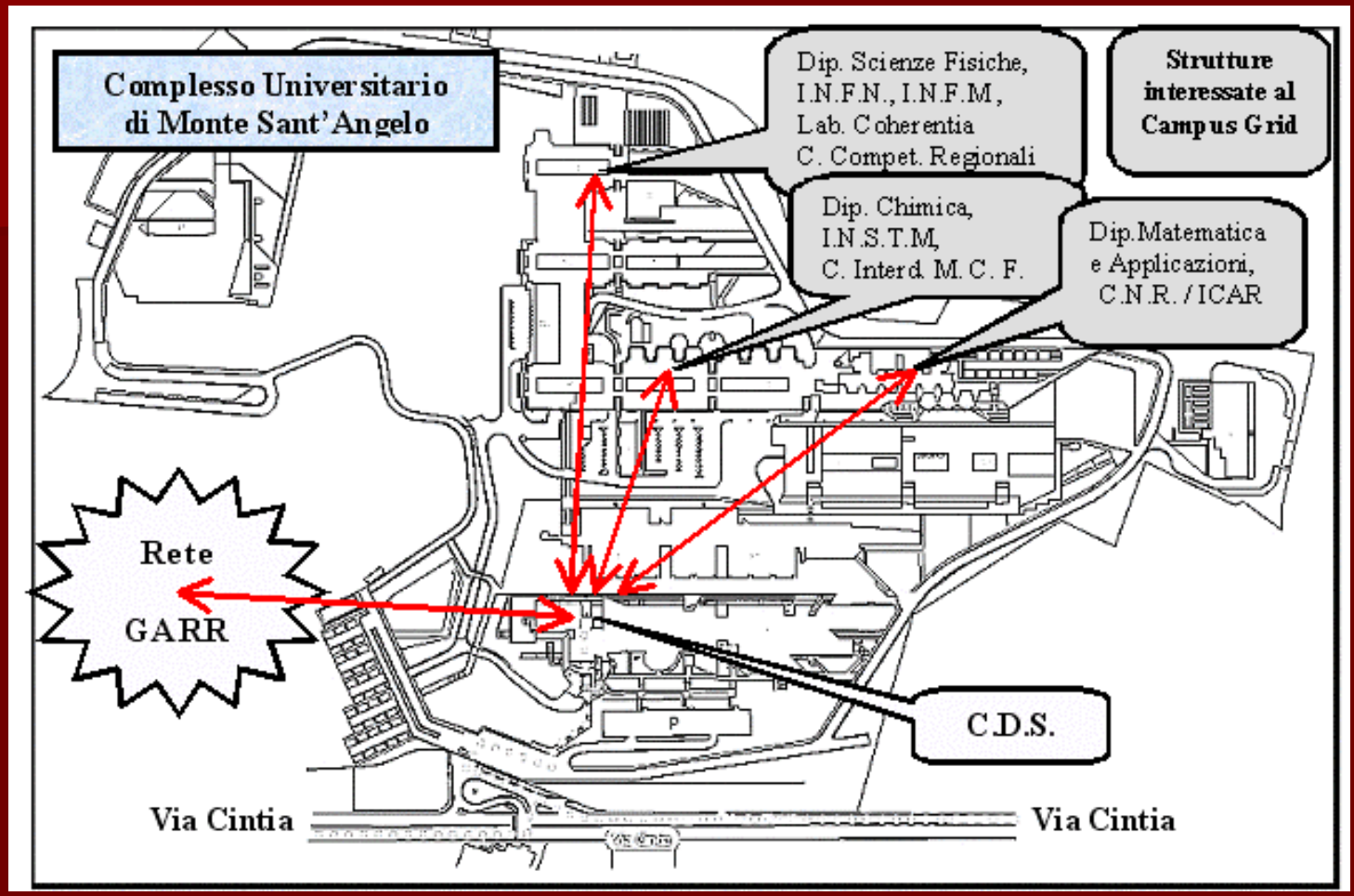
ANFU

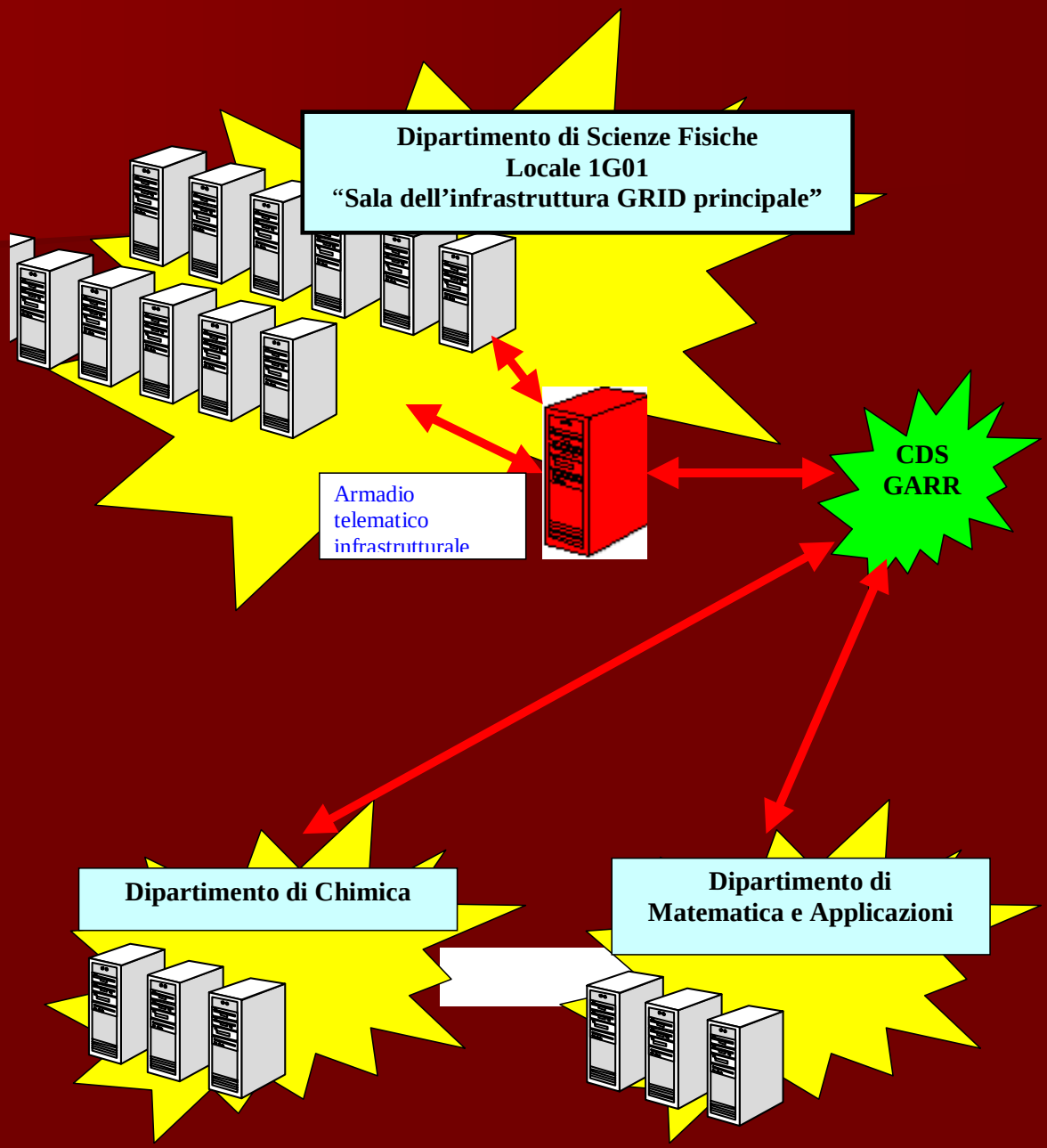
Dipartimento di Scienze Fisiche
Università Federico II - Napoli

Dip. di Matematica ed Informatica
Università di Salerno

INFN
Istituto Nazionale
di Fisica Nucleare

CAMPUS GRID


Complesso Universitario
di Monte Sant'Angelo
Dip. Scienze Fisiche,
I.N.F.N., I.N.F.M.,
Lab. Coherentia
C. Compet. Regionali
Strutture
interessate al
Campus Grid
Dip. Chimica,
I.N.S.T.M,
C. Interd. M.C.F.
Dip.Matematica
e Applicazioni,
C.N.R./ICAR
Rete
GARR
C.D.S.
Via Cintia
Via Cintia


Dipartimento di Scienze Fisiche
Locale 1G01
“Sala dell’infrastruttura GRID principale”
Armadio
telematico
infrastrutturale
CDS
GARR
Dipartimento di Chimica
Dipartimento di
Matematica e Applicazioni

256 Itanium 2CPU’s 1 GB RAM

16 (2*8) Alpha

512 Pentium IV CPU

128 Opteron

Optical fiber backbone

There are hundreds of different NN's

- MLP (Multi layer perceptron): slow, supervised, non linear
- SOM (self organizing maps) : faster, unsupervised, non linear, great visualization, non physical output
- GTM (generative topographic mapping): slow, unsupervised, great visualization, physical output
- PCA & ICA linear and non linear: terrible visualization, physical output, good performances on uncorrelated data
- Fuzzy C Means: slow on MDSs, effective in "fuzzy problems"
- PPS: great (the best ones for unsupervised clustering, classification and visualization)
- Competitive Evolution on Data (CED): bad visualization, great accuracy as unsupervised clustering tool, ...
- Etc.

Fact:

In VO data sets: $D_D >> 1$, $D_S >> 1$

Advantages:

Data Complexity ⮝ Multidimensionality ⮝ Discoveries

But:

The computational cost of clustering analysis:

K-means: $K \times N \times I \times \mathbf{D}$

Expectation Maximisation: $K \times N \times I \times \mathbf{D^2}$

Monte Carlo Cross-Validation: $M \times K_{max}^2 \times N \times I \times \mathbf{D^2}$

N = no. of data vectors, D = no. of data dimensions

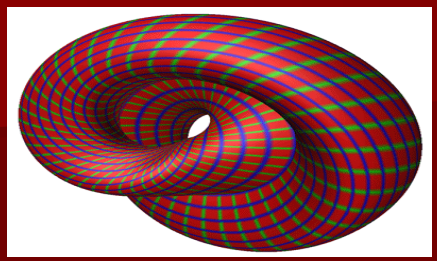
K = no. of clusters chosen, K_{max} = max no. of clusters tried

I = no. of iterations, M = no. of Monte Carlo trials/partitions

Terascale (Petascale?) computing and/or better algorithms

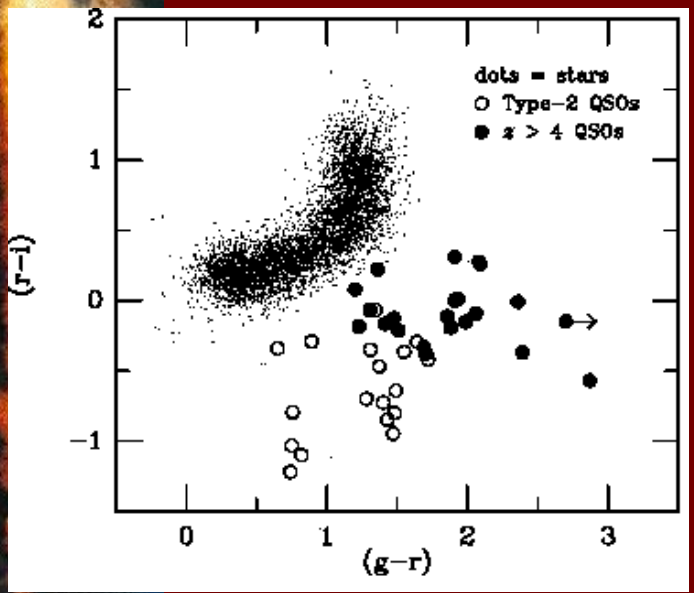
The Curse of Hyperdimensionality

Visualization

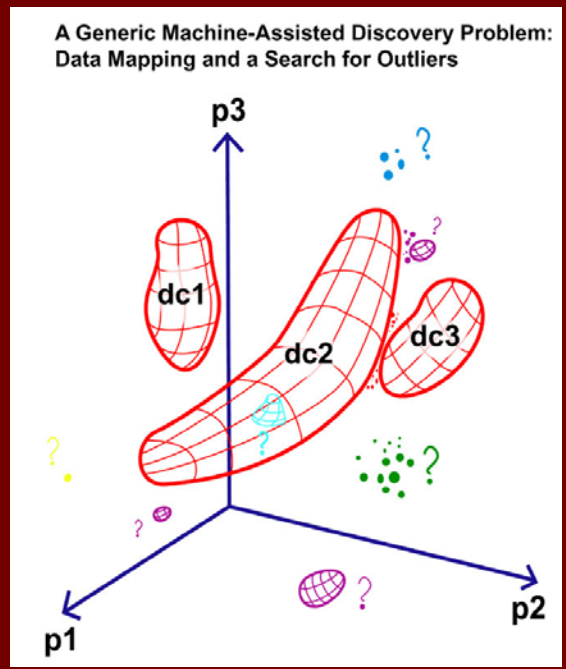


A fundamental limitation of the human perceptionis that we cannot visualize spaces with dimensionality higher than: DMAX = 3? 5?

2-D

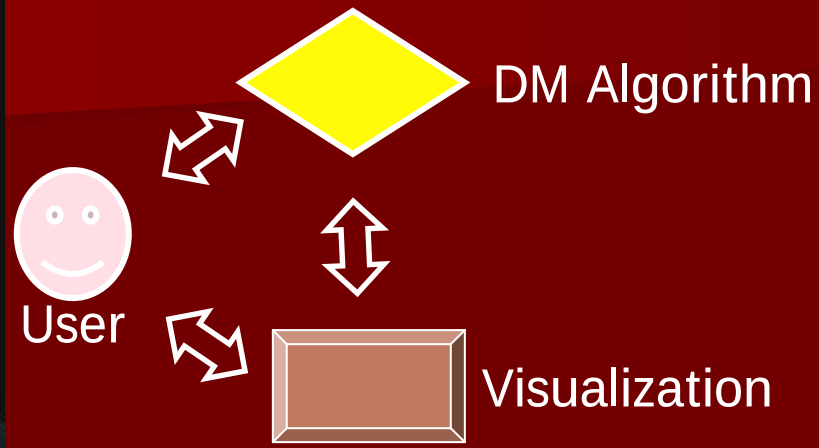


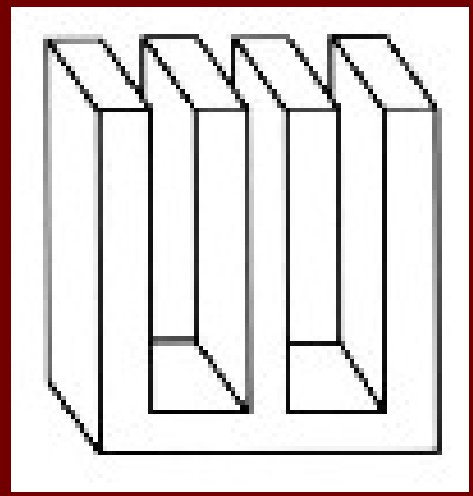
3-D



WHAT DO WE DO WHEN N=50 ?

To visualize and understand we need somehow to compress the relevant information…





- Interactive visualization is a key part of the data mining process:
- Some methodology exists, but **much more is needed**

Aims and applications of AstroNeural

User friendly tool to perform clustering and data mining in high dimensionality spaces

Aims

- Clustering & pattern recognition in high dimensionality spaces
- Visualization
- Classification
- Parametrization of images
- Modeling of massive data sets

Applications

- Astrophysics
- Genetics
- Geophysics
- High energy physics
- Atmospheric physics
- Etc.

Neural Networks are good at:

- performing linear and non linear interpolation
- generalizing
- performing non linear analysis and itentify common trends in data
- for forecasting
- classifying

They learn in two main ways:

- Supervised
- Unsupervised

A priori knowledge needed

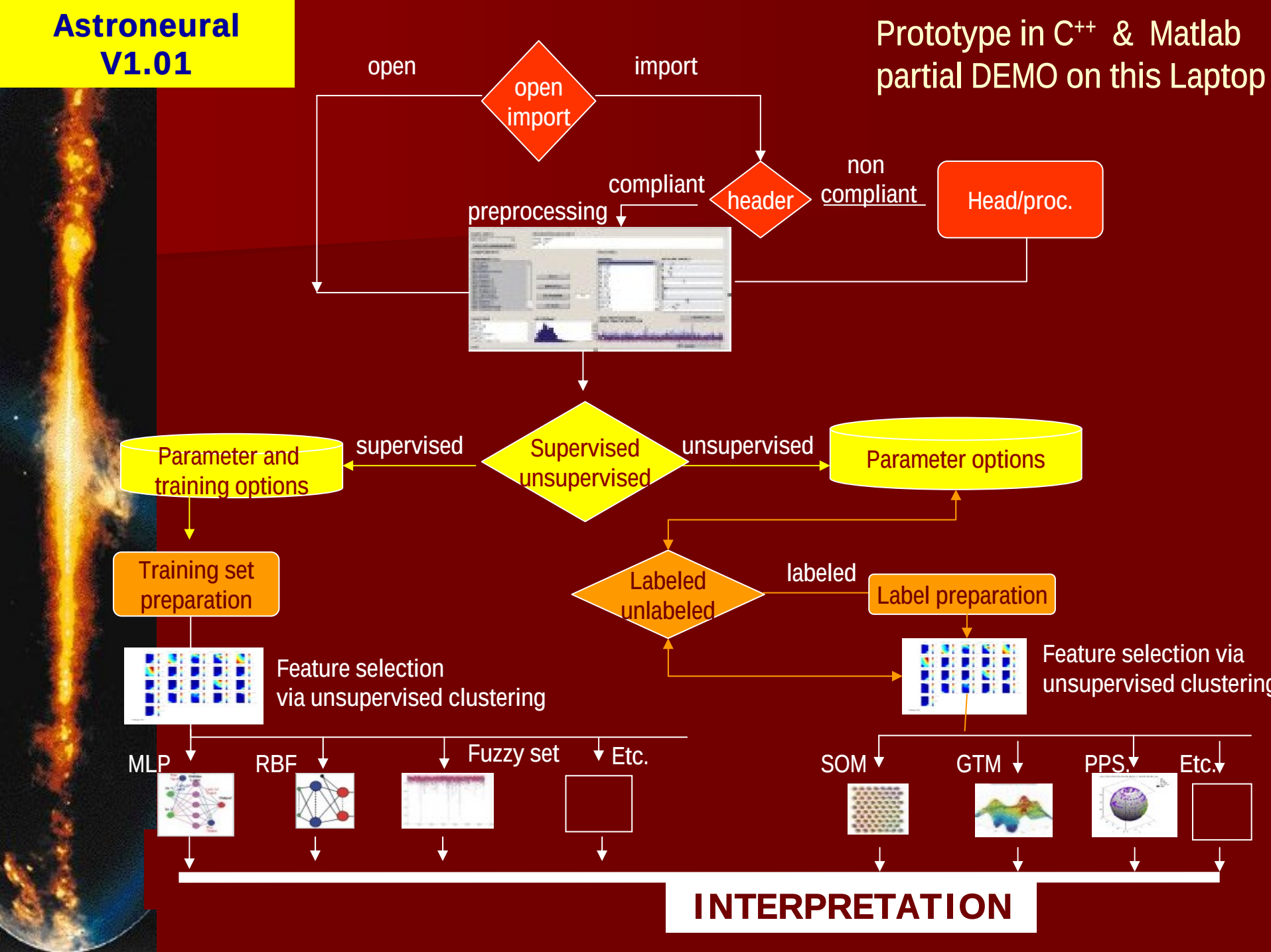
- Training
- Computation of errors

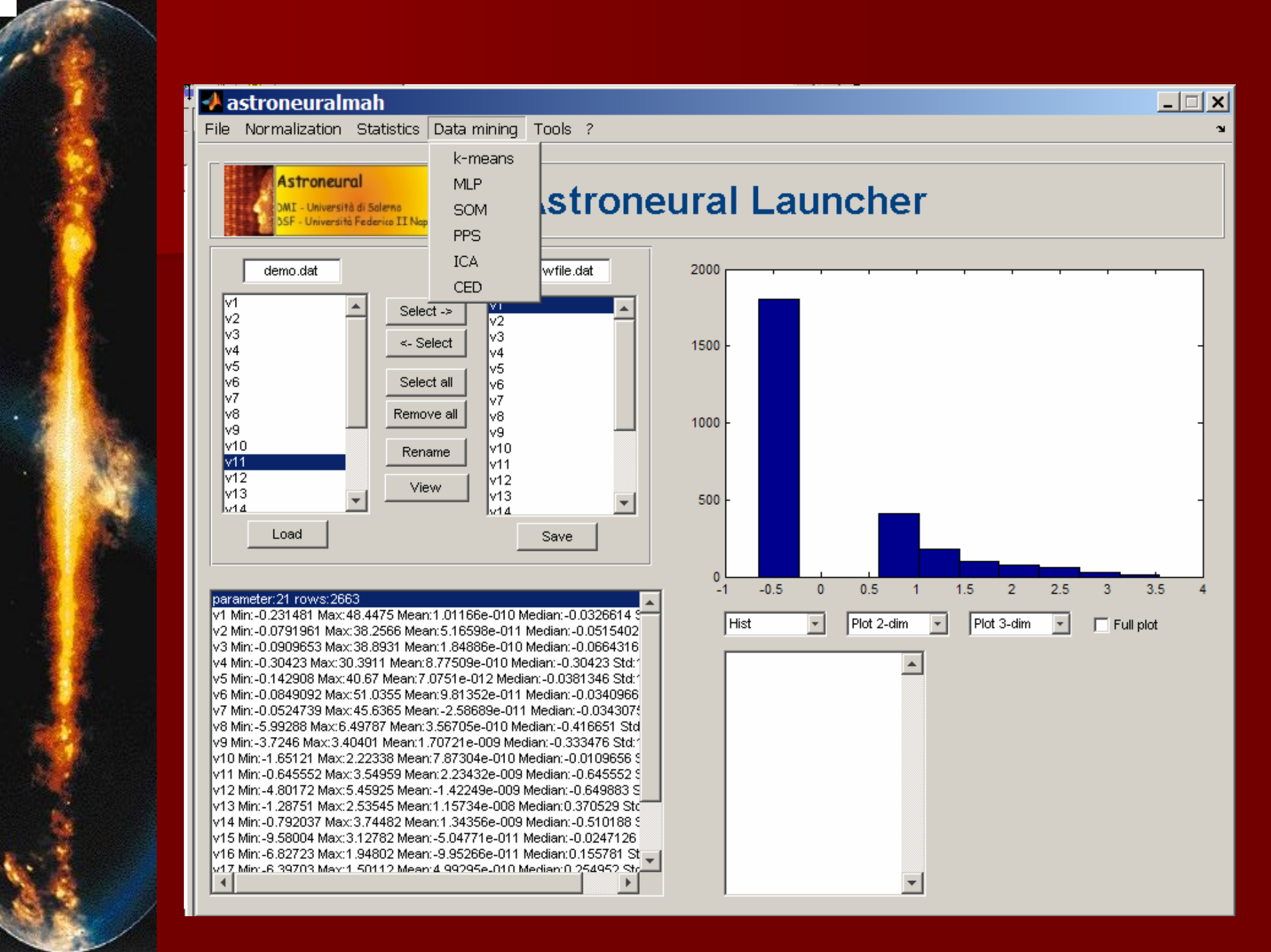
Null/small a priori knowledge

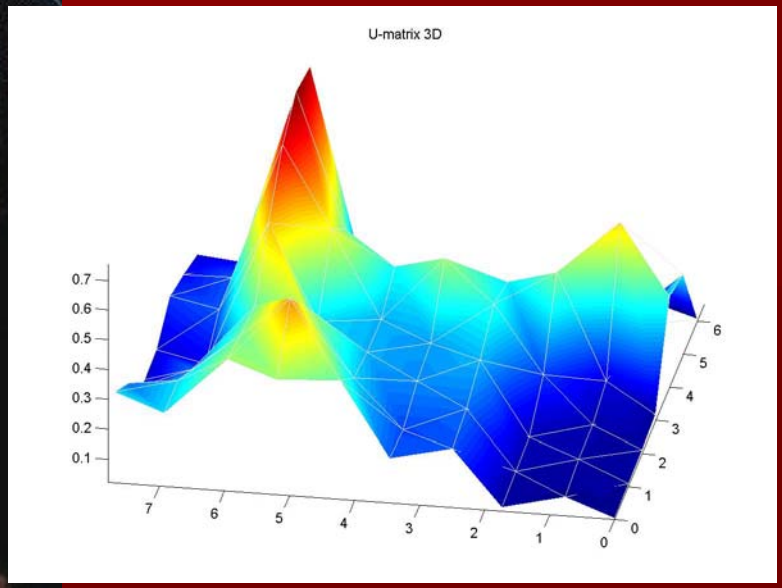
Clustering done on statistical properties of data themeselves

Performances and errors are derived statistically

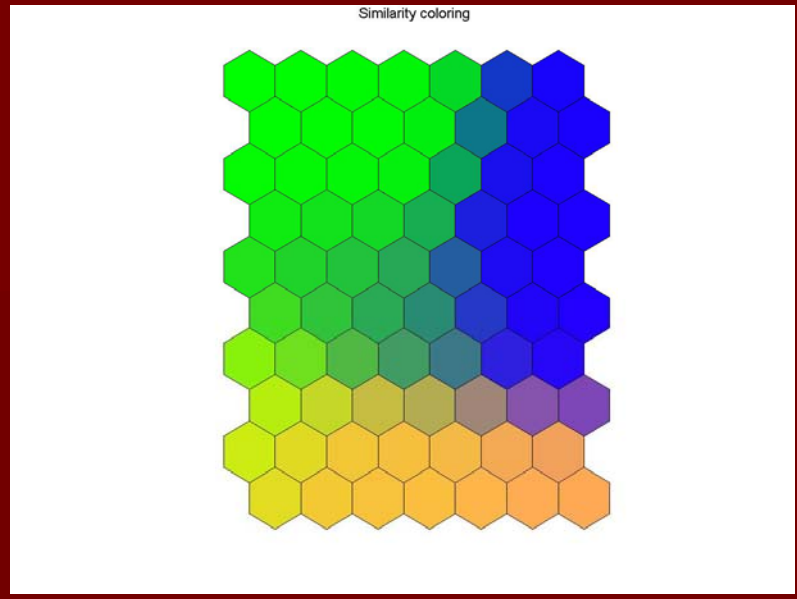
Knowledge comes through labeling


Astroneural
V1.01
Prototype in C++ & Matlab
partial DEMO on this Laptop
open
open
import
import
compliant
header
non
compliant
Head/proc.
preprocessing
Parameter and
training options
supervised
Supervised
unsupervised
unsupervised
Parameter options
Training set
preparation
Labeled
unlabeled
labeled
Label preparation
Feature selection
via unsupervised clustering
Feature selection via
unsupervised clustering
MLP
RBF
Fuzzy set
Etc.
SOM
GTM
PPS.
Etc.
INTERPRETATION

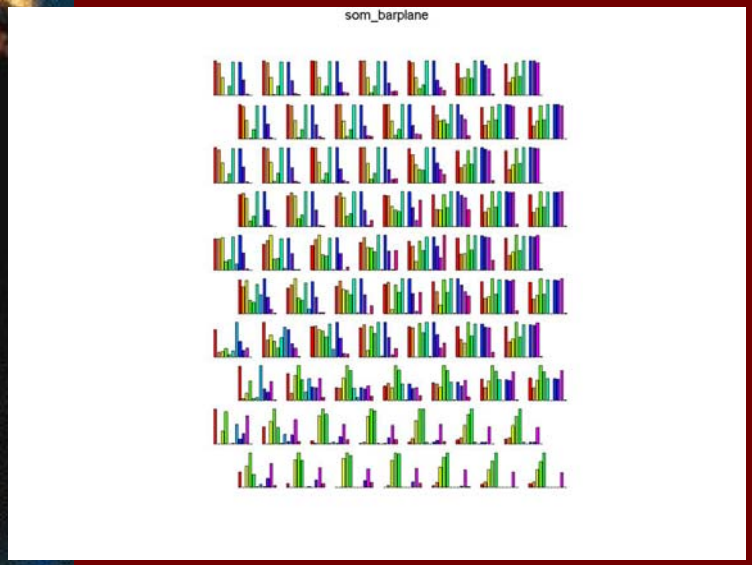

astroneuralmah
File
Normalization
Statistics
Data mining
Tools
?
k-means
MLP
SOM
PPS
ICA
CED
Astroneural
DMI - Università di Salerno
DSF - Università Federico II Nap
Astroneural Launcher
demo.dat
wfile.dat
v1
v2
v3
v4
v5
v6
v7
v8
v9
v10
v11
v12
v13
Select ->
<- Select
Select all
Remove all
Rename
View
Load
Save
2000
1500
1000
500
0
-1
-0.5
0
0.5
1
1.5
2
2.5
3
3.5
4
Hist
Plot 2-dim
Plot 3-dim
Full plot
parameter:21 rows:2663
v1 Min:-0.231481 Max:48.4475 Mean:1.01166e-010 Median:-0.0326614
v2 Min:-0.0791961 Max:38.2566 Mean:5.16598e-011 Median:-0.0515402
v3 Min:-0.0909653 Max:38.8931 Mean:1.84886e-010 Median:-0.0664316
v4 Min:-0.30423 Max:30.3911 Mean:8.77509e-010 Median:-0.30423
v5 Min:-0.142908 Max:40.67 Mean:7.0751e-012 Median:-0.0381346
v6 Min:-0.0849092 Max:51.0355 Mean:9.81352e-011 Median:-0.0340966
v7 Min:-0.0524739 Max:45.6365 Mean:-2.58689e-011 Median:-0.034307
v8 Min:-5.99288 Max:6.49787 Mean:3.56705e-010 Median:-0.416651
v9 Min:-3.7246 Max:3.40401 Mean:1.70721e-009 Median:-0.333476
v10 Min:-1.65121 Max:2.22338 Mean:7.87304e-010 Median:-0.0109656
v11 Min:-0.645552 Max:3.54959 Mean:2.23432e-009 Median:-0.645552
v12 Min:-4.80172 Max:5.45925 Mean:-1.42249e-009 Median:-0.649883
v13 Min:-1.28751 Max:2.53545 Mean:1.15734e-008 Median:0.370529
v14 Min:-0.792037 Max:3.74482 Mean:1.34356e-009 Median:-0.510188
v15 Min:-9.58004 Max:3.12782 Mean:-5.04771e-011 Median:-0.0247126
v16 Min:-6.82723 Max:1.94802 Mean:-9.95266e-011 Median:0.155781



3-D U Matrix



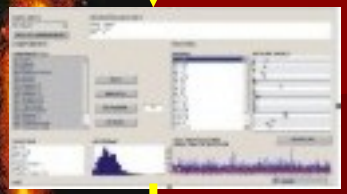
Similarity coloring



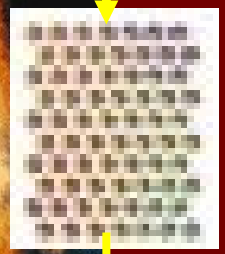
Feature significance maps

Example n. I: photometric redshifts for SDSS

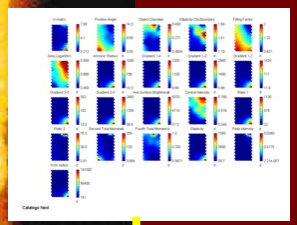
SDSS-EDR DB



Unsupervised SOMS's + supervised MLP's



SOM unsup.
Set construction

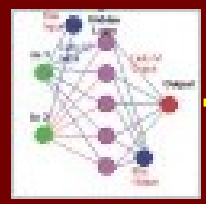


SOM supervised
Feature selection

SOM unsup.
completeness

Reliability
Map

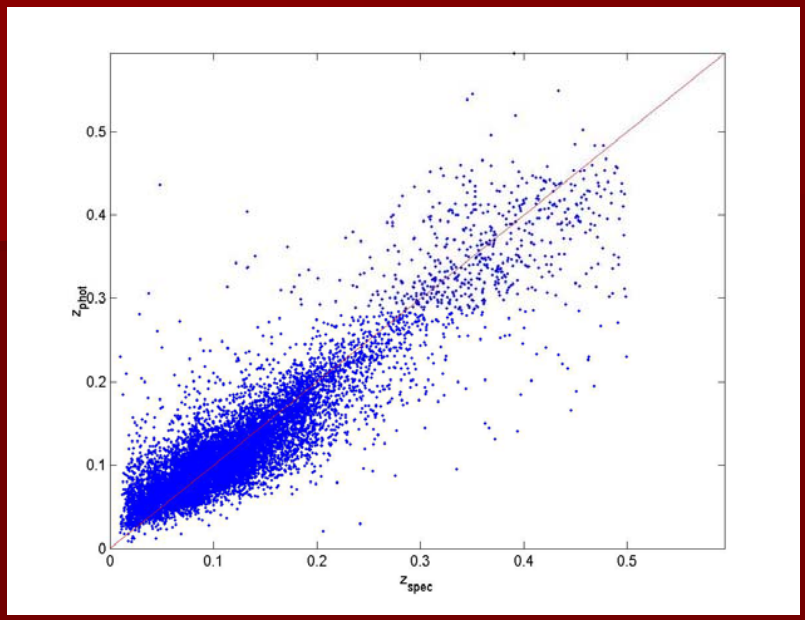
MLP supervised
experiments



Best MLP
model

- Input data set:
  SDSS – EDR photometric data (galaxies)
- Training / validation / test set:
  SDSS-EDR spectroscopic subsample

Unsupervised SOMS's + supervised MLP's



Robust error: 0.02176
16 millions P.R.

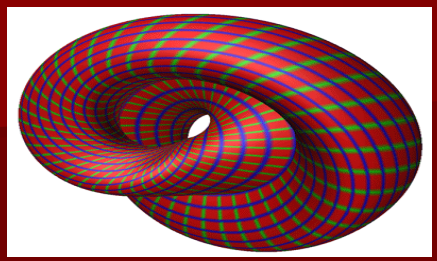
SOM output (each exhagon is a neuron)

Numbers above frame: redshift range

Numbers in the cells:
number of input data activating that neuron

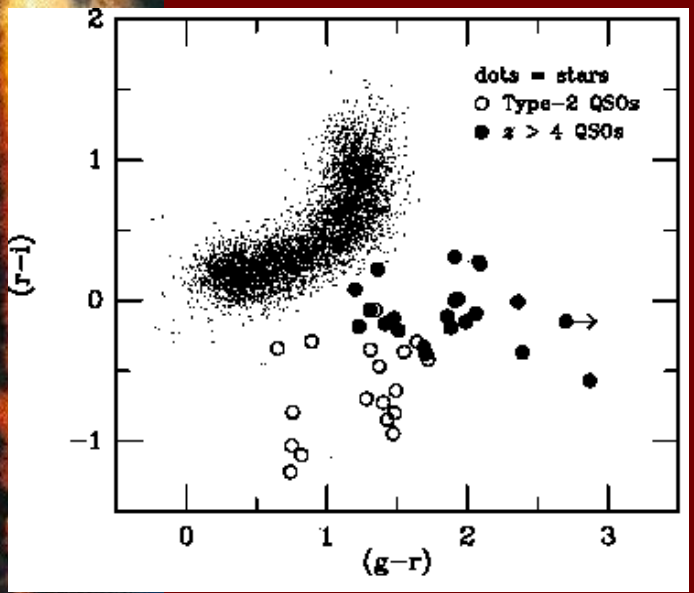
The Curse of Hyperdimensionality

Visualization

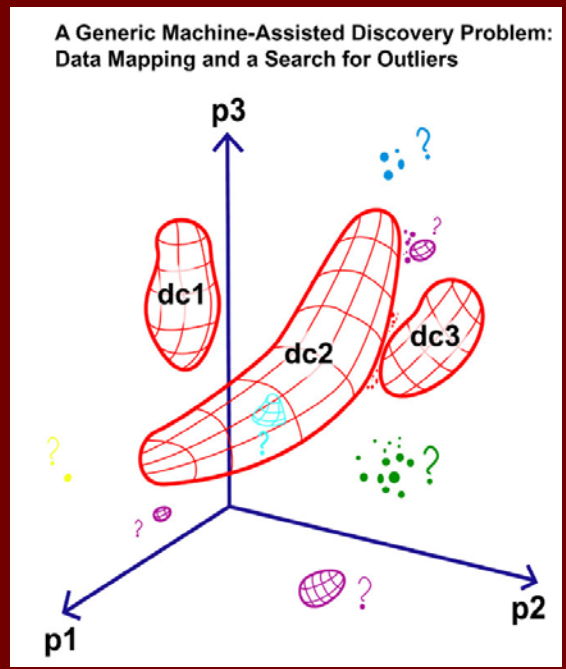


A fundamental limitation of the human perceptionis that we cannot visualize spaces with dimensionality higher than: DMAX = 3? 5?

2-D



3-D

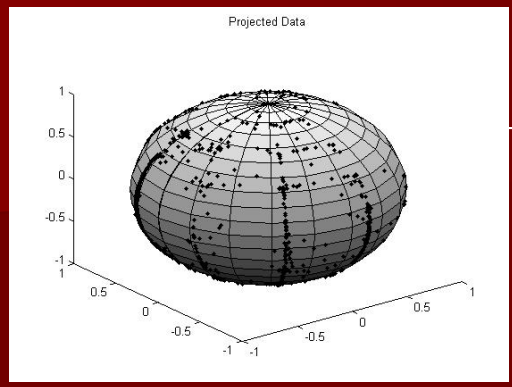


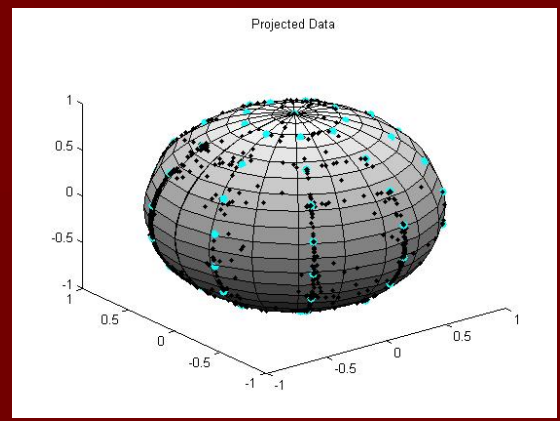
WHAT DO WE DO WHEN N=50 ?

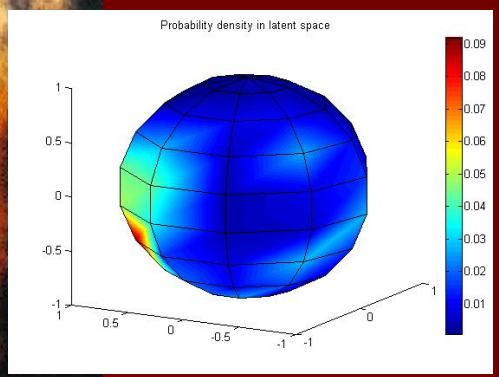
Probabilistic Principal Surfaces (PPS)

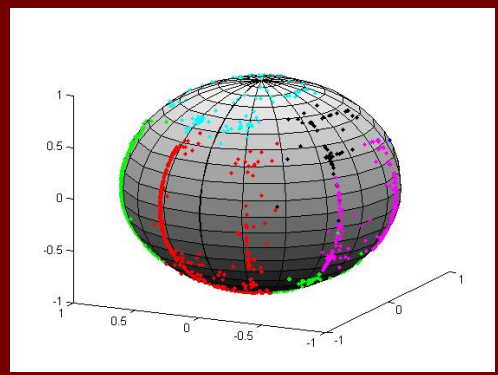
GOODS

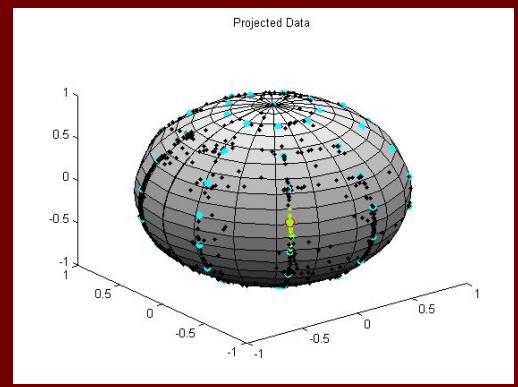
Data set











report - WordPad

File Edit View Insert Format Help

```
-------------------------------------------------
16-Nov-2004 17:34:41

Latent Variable n.  65
Number of points 15
Point index          Confidence
--------------------------------
    33             0.999977650         1
    113             0.900503866
    159             0.999999997
    172             0.920108462
    185             0.999996577
    197             0.999999167         1
    208             0.999999687
    233             0.998609813
    250             1.000000000
    284             0.999999999         1
    324             0.999999009
    1502             0.999836633
    1546             0.964281584
    1615             0.999999839         1
    1686             0.999948042
-------------------------------------------------
16-Nov-2004 20:51:16

Latent Variable n.  56
Number of points 27
Point index          Confidence
--------------------------------
    1759             0.999285812         2
    1799             0.980303983
    1811             0.996048119
    1828             0.997007775         1
    1869             0.971020113
    1935             0.998875283         2
    1999             0.995662703
    2045             0.982838862         2
    2065             0.996552019
```

For Help, press F1

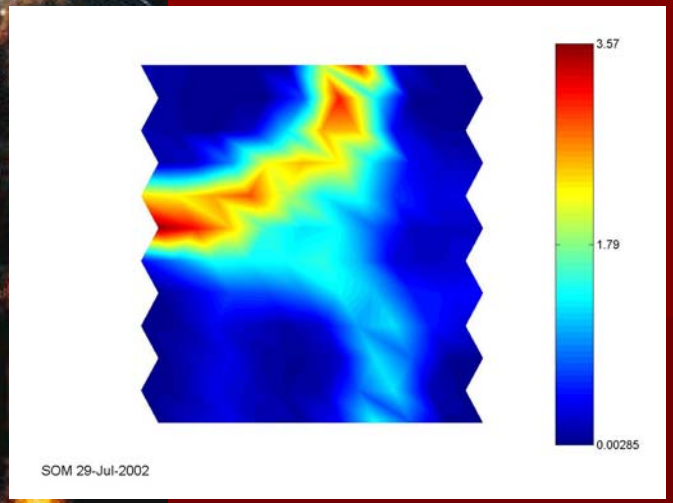
Astroneural Version 1.01 ready

Portation on freeware software in progress

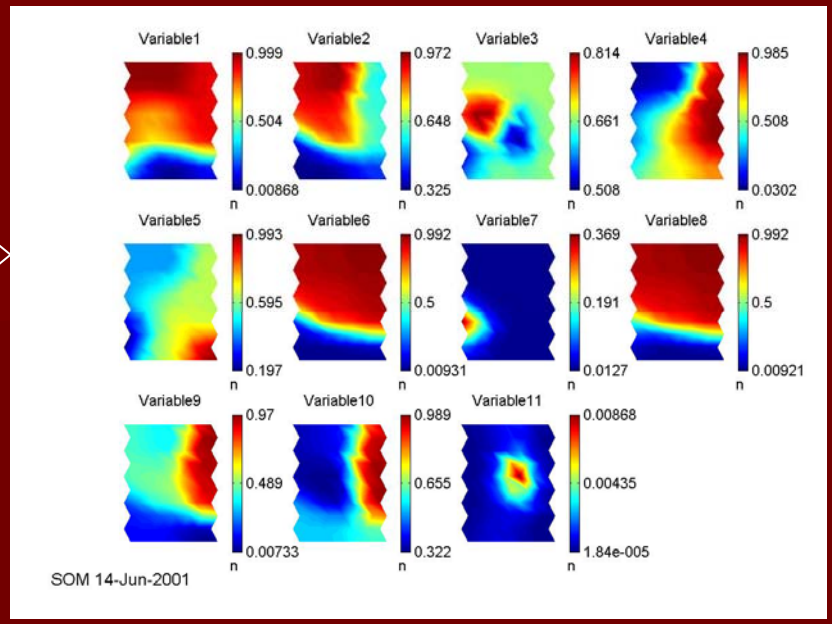
Web interface ready

Implementation of specific tasks for GRID use is in progress (NA, TS, CT, etc.)

Implementation of backwards connection to pixels is in progress

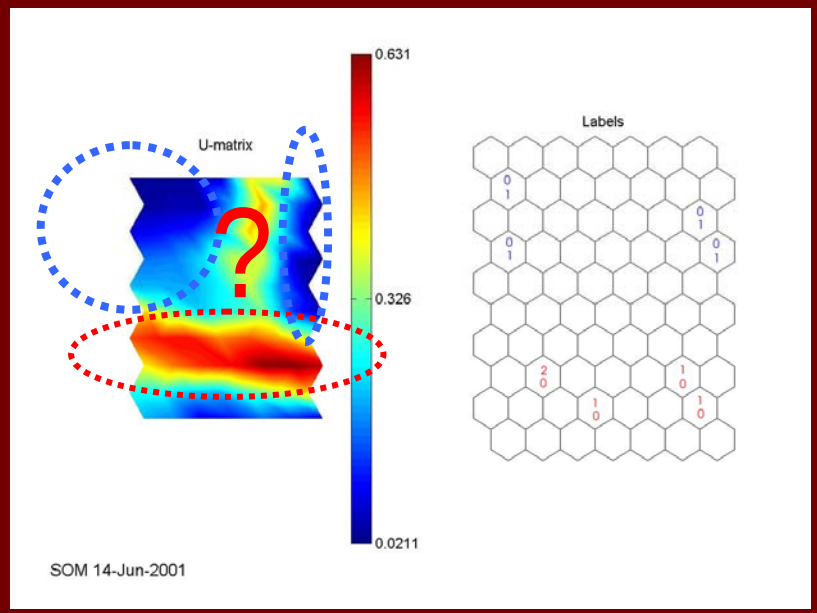


B.E.S.



U-Matrix
278 parameters

Compressed feature space
BMU matrices

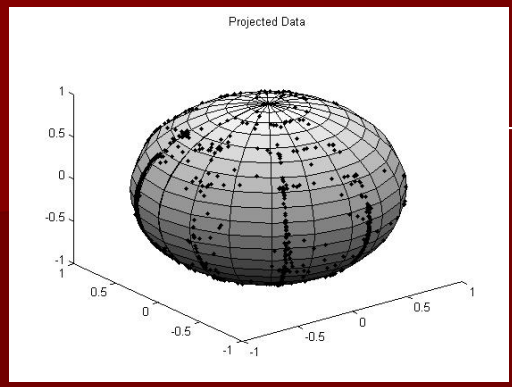


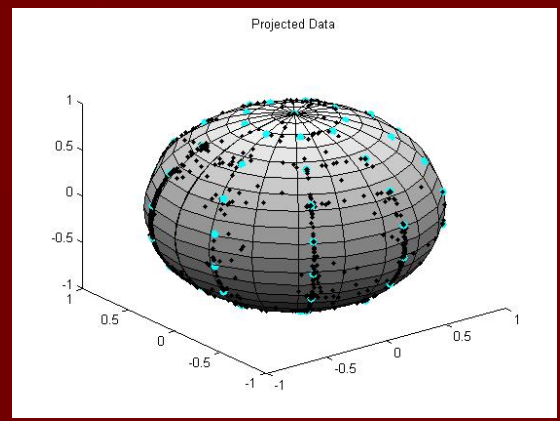
UP: good tracking
Below: bad tracking

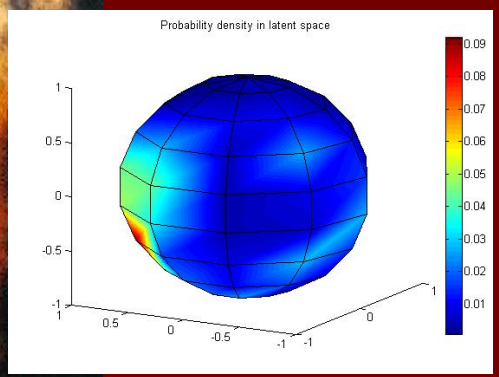
Probabilistic Principal Surfaces (PPS)

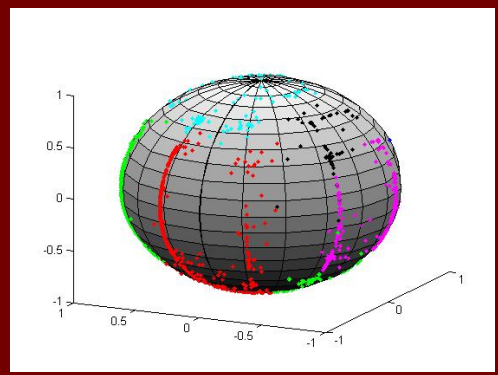
GOODS

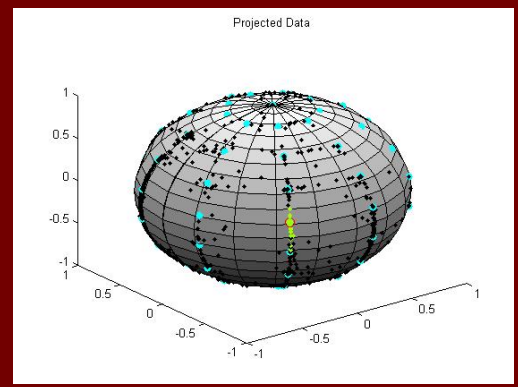
Data set











report - WordPad

File Edit View Insert Format Help

```
-------------------------------------------------
16-Nov-2004 17:34:41

Latent Variable n.  65
Number of points 15
Point index          Confidence
--------------------------------
    33             0.999977650         1
    113             0.900503866
    159             0.999999997
    172             0.920108462
    185             0.999996577
    197             0.999999167         1
    208             0.999999687
    233             0.998609813
    250             1.000000000
    284             0.999999999         1
    324             0.999999009
    1502             0.999836633
    1546             0.964281584
    1615             0.999999839         1
    1686             0.999948042
-------------------------------------------------
16-Nov-2004 20:51:16

Latent Variable n.  56
Number of points 27
Point index          Confidence
--------------------------------
    1759             0.999285812         2
    1799             0.980303983
    1811             0.996048119
    1828             0.997007775         1
    1869             0.971020113
    1935             0.998875283         2
    1999             0.995662703
    2045             0.982838862         2
    2065             0.996552019
```

For Help, press F1

Astroneural Version 1.01 ready

Portation on freeware software in progress

Web interface ready

Implementation of specific tasks for GRID use is in progress (NA, TS, CT, etc.)

Implementation of backwards connection to pixels is in progress

A multifrequency analysis of radio variability of blazars

A. Ciaramella[1,7], C. Bongardo[2], H. D. Aller[3], M. F. Aller[3], G. De Zotti[2], A. Lähteenmaki[4], G. Longo[5,6], L. Milano[5,6], R. Tagliaferri[1,7], H. Teräsranta[4], M. Tornikoski[4], and S. Urpo[4]

A&A 419, 485–500 (2004)
DOI: 10.1051/0004-6361:20035771

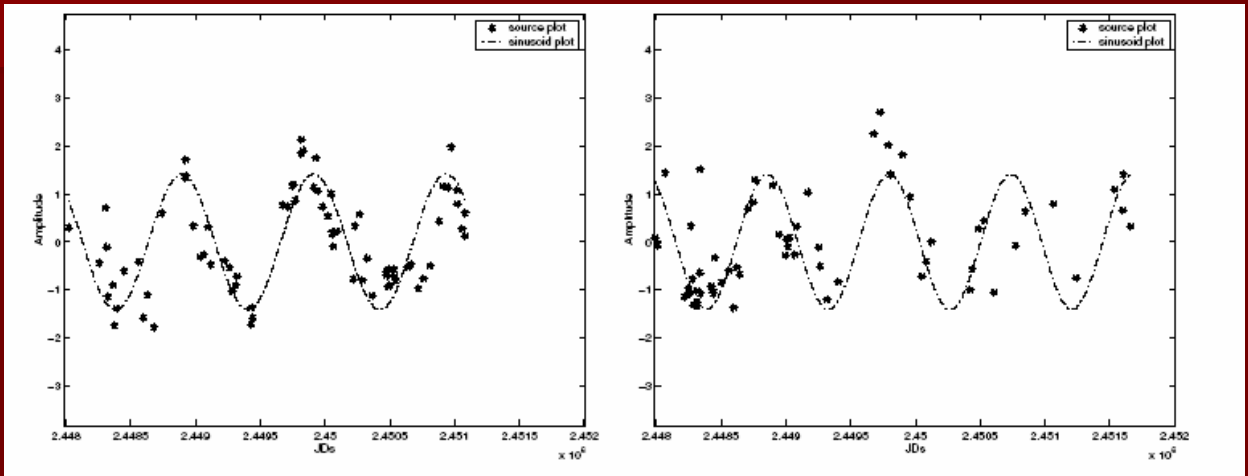


Table 1. Objects for which we have positive detection of periodicity in Metsähovi daily averaged data. Column 1: object identification. Columns 2–5: 22 GHz data; Cols. 6–9: 37 GHz data. Columns 2 and 6: number of data points; Cols. 3 and 7: maximum admissible period to avoid aliasing; Cols. 4 and 8: period (in units of 10^3 days) obtained by STIMA; Cols. 5 and 9: period obtained from the Lomb's Periodogram.

Name	N	Mx. P. ($\times 10^3$)	STIMA ($\times 10^3$)	Lomb ($\times 10^3$)	N	Mx. P. ($\times 10^3$)	STIMA ($\times 10^3$)	Lomb ($\times 10^3$)	Notes
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
0224 + 671	76	3.062	1.021	1.021	63	3.668	0.950	0.970	
0945 + 408	47	3.742	1.386	1.336	26	4.467	1.313	1.240	few points
1226 + 023	694	6.665	3.029	2.777	716	7.822	3.260	1.261	
2200 + 420	644	6.676	3.034	3.034	715	7.873	2.811	3.028	
2251 + 158	571	6.676	2.384	2.384	549	7.538	2.217	2.512	

Status

- **Matlab Version 1.0 ready and available**

  (no documentation)
- **Porting on C++** in progress
- **Web access** (upload data and download results) done (to be redone)

- **VO conversion** (to be done)
- **Integration with high performance visualization tools (VISIVO)** beginning with CT/BO

VO started in 2000

VO Tech is a new project ..... And we shall be judged for it

Where are we

Where do we want to go

Why

?

Census of existing

Sharing know how

Choice of test applications

Who is going to say whether they are correct and useful?

Templates

Simulations vs real data?