DS6 – Phase 4 Napoli group

Astroneural 1,0 is available and includes tools for supervised and unsupervised data mining:

- Preprocessing & visualization
- Supervised (MLP, RBF)
- Unsupervised (PPS, NEC+dendrogram, SOM)
- PCA and ICA
- Genetic algorithms
- visualzation of results
- http://people.na.infn.it/~longo/

Science is a non negligible part of the work:

- to understand what the r.m.s astronomer needs and to adapt the tool accordingly
- to convince the community that this approach is useful

Hence: Astroneural v 1.0 is being tested on several science cases:

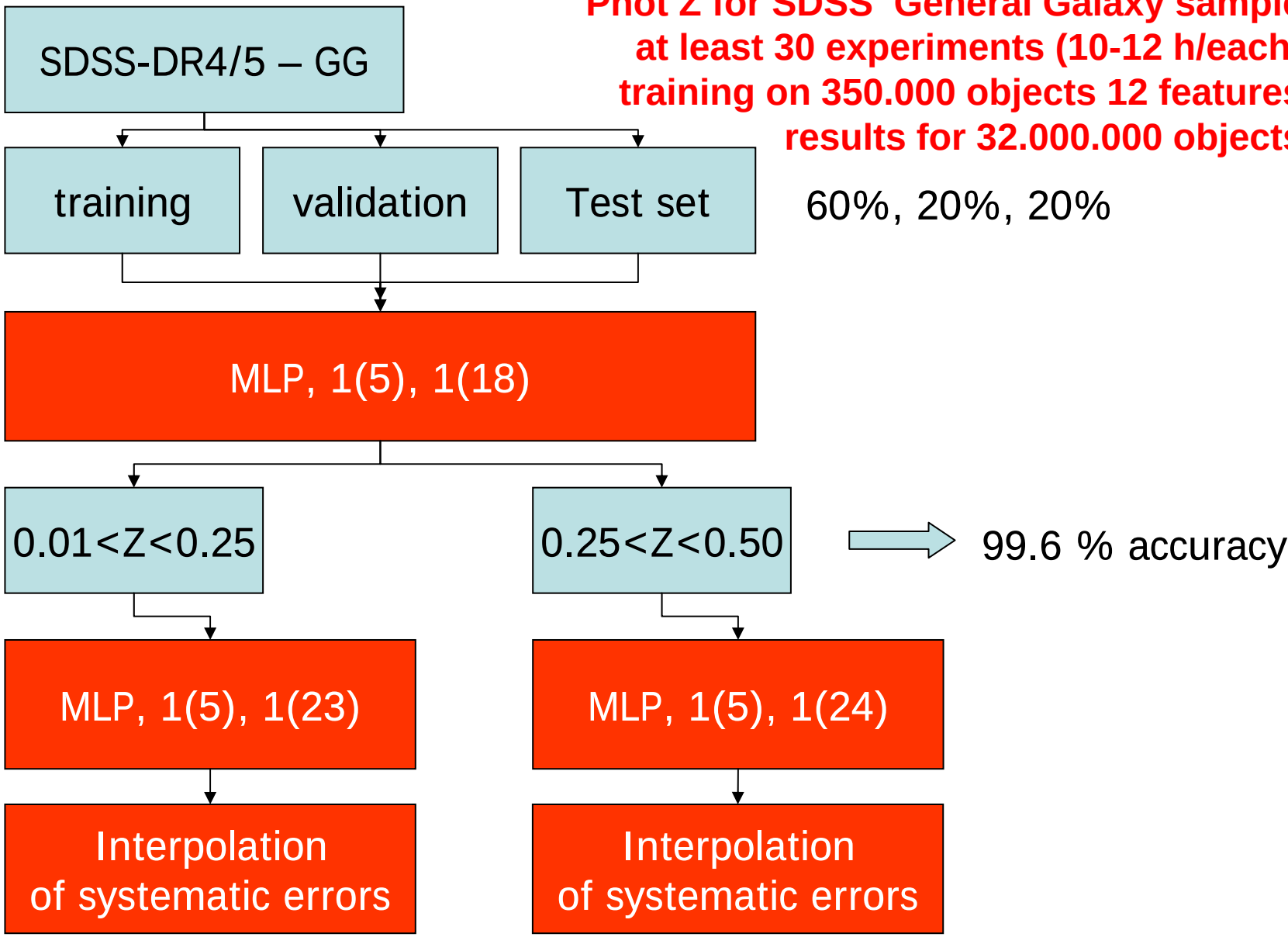
- **photometric redshifts for the SDSS dataset**
- **physical classification of galaxies using photometry (multiband)**
- **Statistical studies of loose groups in 3-D space**
- Star/Galaxy classification on multiepoch survey data (PalomarQuest with Caltech)
- AGN and QSO identification from multiband photometric surveys (in progress)

1. A novel approach to gene expression clustering,, 2005, Bioinformatics, 2006, 22, pp. 589-596
2. Visualization, Clustering and Classification of Multidimensional Astronomical Data,, 2005, CAMP 2005
3. Mining the SDSS data. I. Photometric redshifts for the nearby universe, ApJ. accepted (astro-ph/0703108)
4. Donalek C., 2007, Mining Massive Astronomical Data Sets, Ph.D. Thesis, University Federico II in Napoli & Caltech
5. The use of neural networks to probe the structure of the nearby universe, R. d'Abrusco, G. Longo, M. Paolillo, E. de Filippis, M. Brescia, A. Staiano, R. Tagliaferri, 2007 (astro-ph/0701137)
6. Other papers on Astroneural site ….

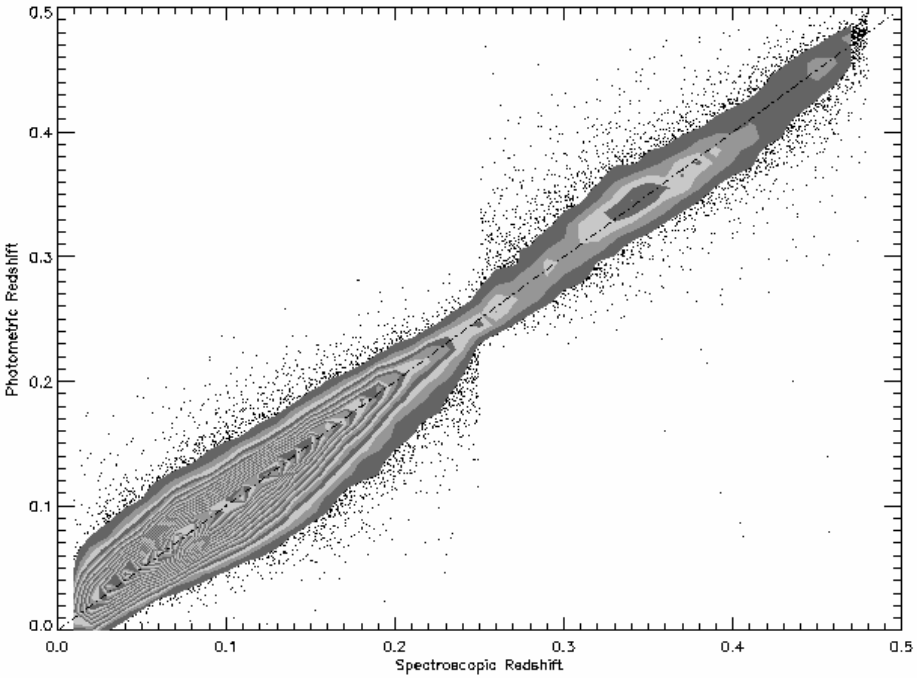
Frequency of Astroneural downloads increased by a factor 10 after publication of papers 1 and 3

VO-Neural

Phot Z for SDSS General Galaxy sample
at least 30 experiments (10-12 h/each)
training on 350.000 objects 12 features
results for 32.000.000 objects

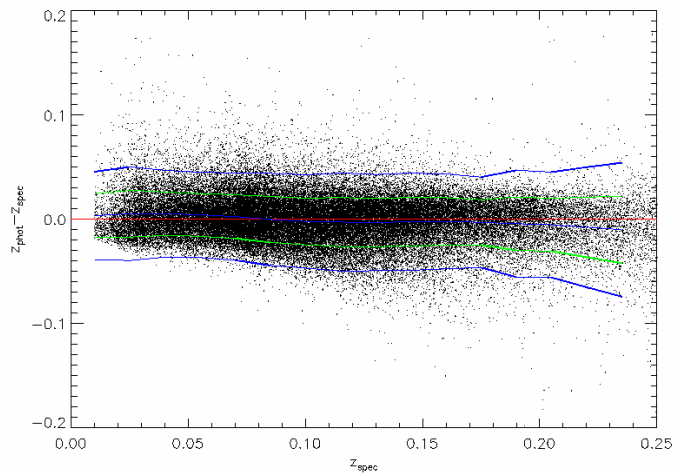


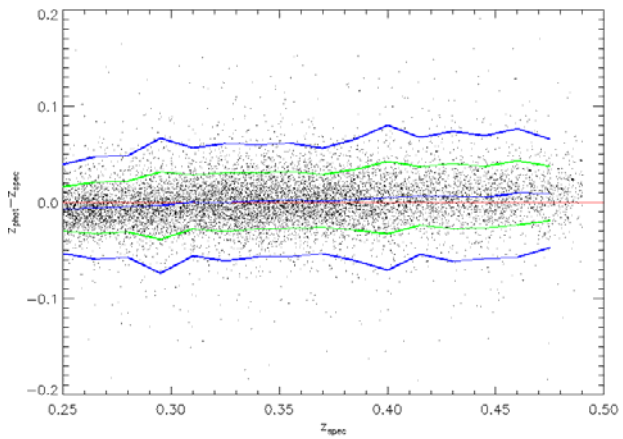
General galaxy sample



$\sigma = 0.0208$

$\Delta z = -0.0029$





VO-Neural

AstroNeural → VO-Neural

Porting of Astroneural as it is, was found to be impossible mainly due to visualization of results and to the fact that software is propietary

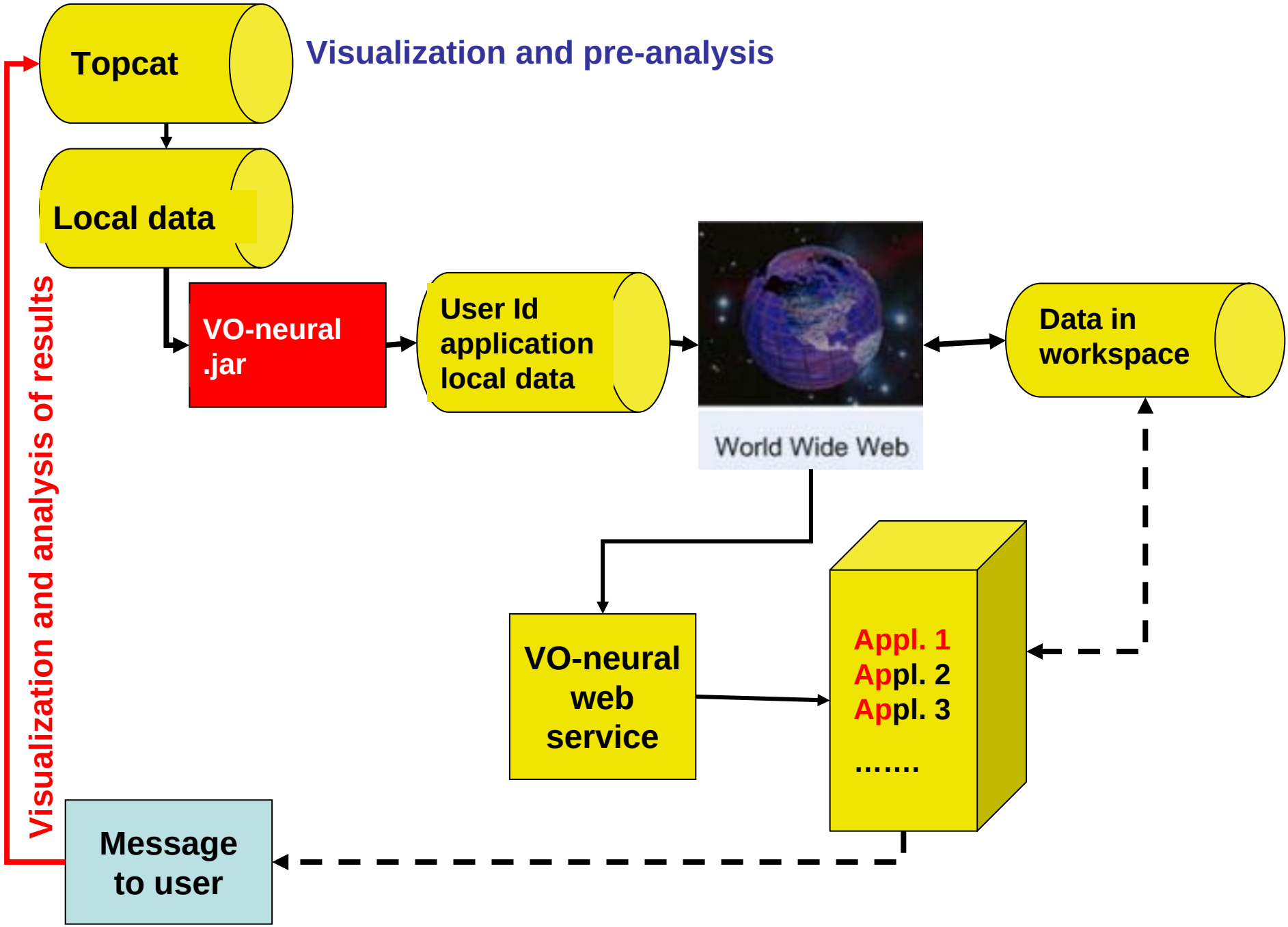
Simple conversion of MatLab code to C is not optimal due to
(i) the need to optimize the code
(ii) the (possibly) long computing time no interactivity
(iii) lack of visualization capabilities

Scalability issue: code needs to be entirely re-written (there is not any C/Java DM library capable to deal with >10^5 records)

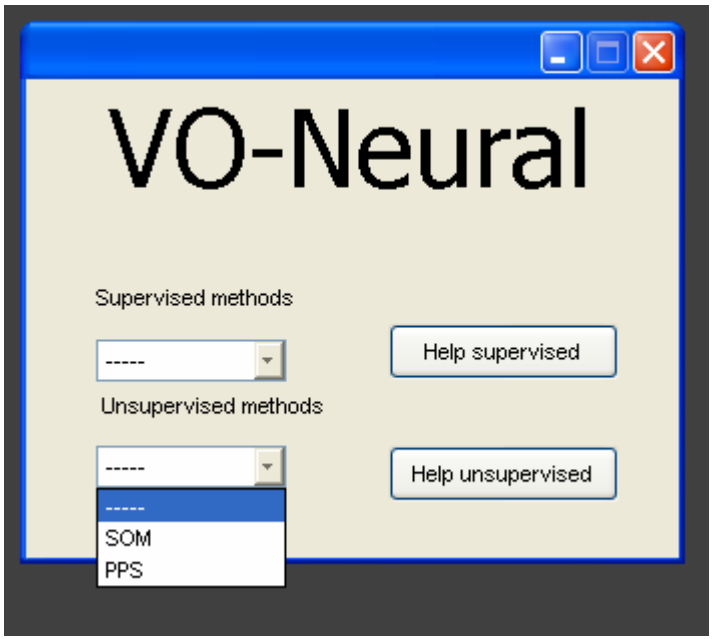
No need to write most of the visualization and pre-analysis tools (TOPCAT does it wonderfully)

Programs must be used in iterative way and each code needs its own visualization tool to become user friendly (some problems do exist)

VO-Neural

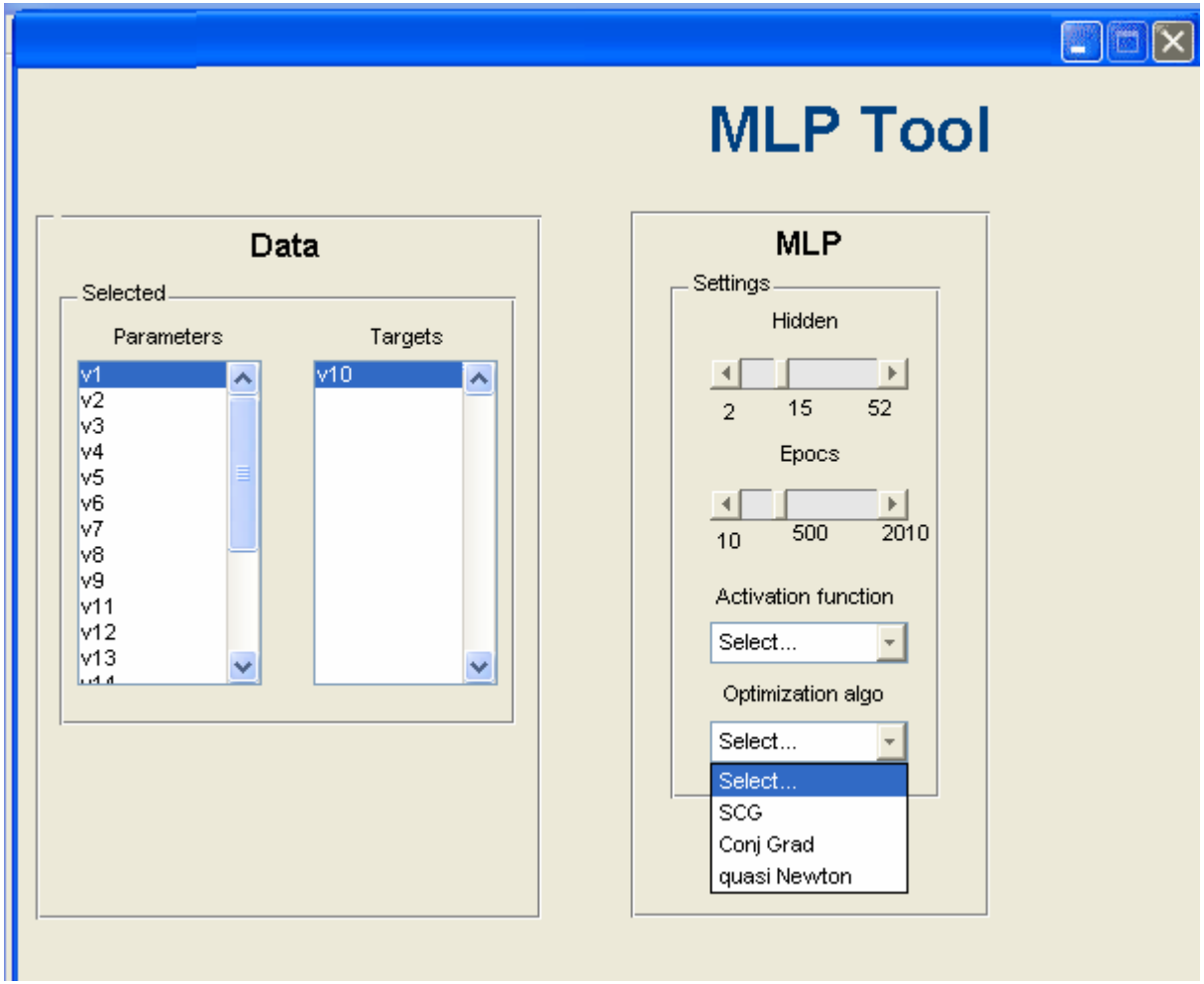

Topcat
Visualization and pre-analysis
Local data
VO-neural
.jar
User Id
application
local data
World Wide Web
Data in
workspace
Visualization and analysis of results
VO-neural
web
service
Appl. 1
Appl. 2
Appl. 3
.....
Message
to user

VOneural.jar



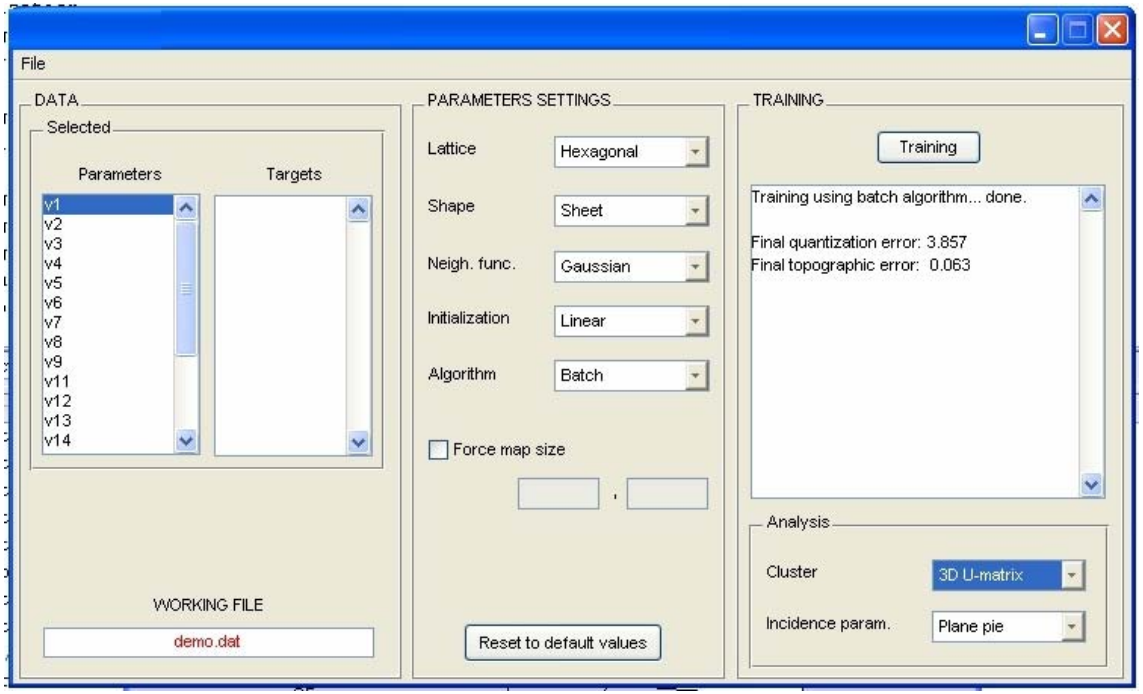
- preprocessing, visualization and data selection using TOPCAT
- Supervised (MLP, RBF)
- Unsupervised (PPS, NEC + dendrogram, SOM)
- Genetic algorithms?
- Clustering via K-means
- PCA & ICA

VOneural.MLP



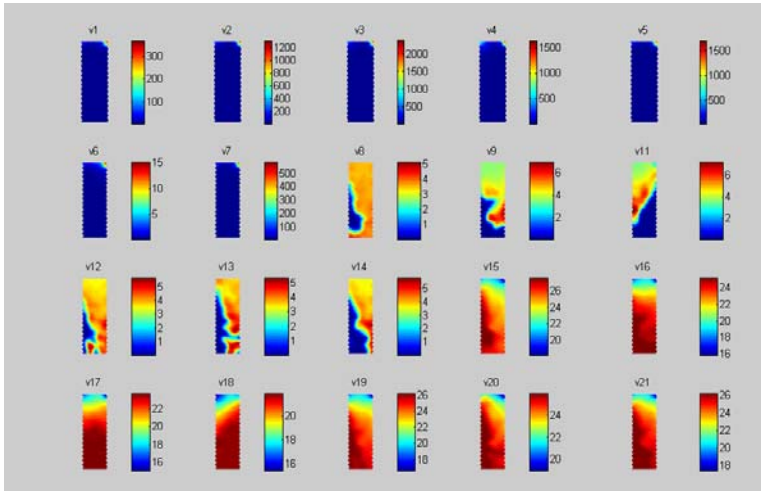
Done: under test as local implementation

VOneural.SOM tool

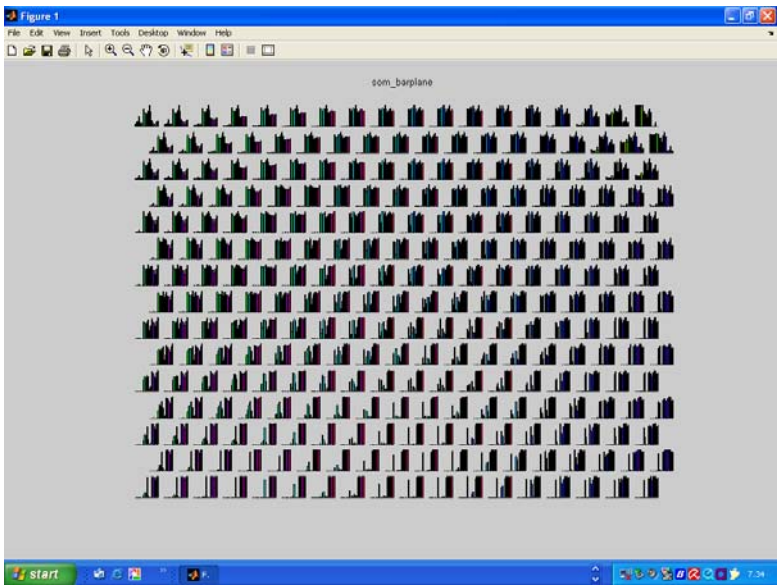


Partially done: Problems with visualization of some features

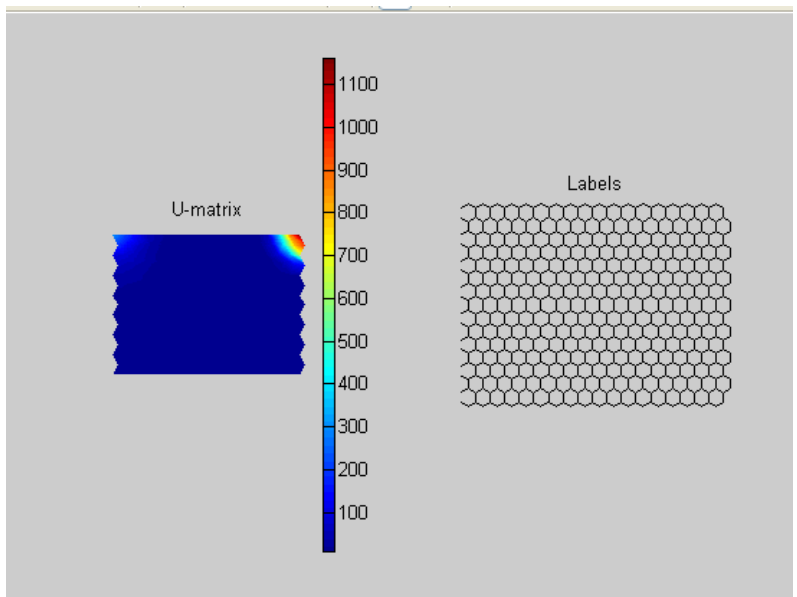
Examples of needed visualizations



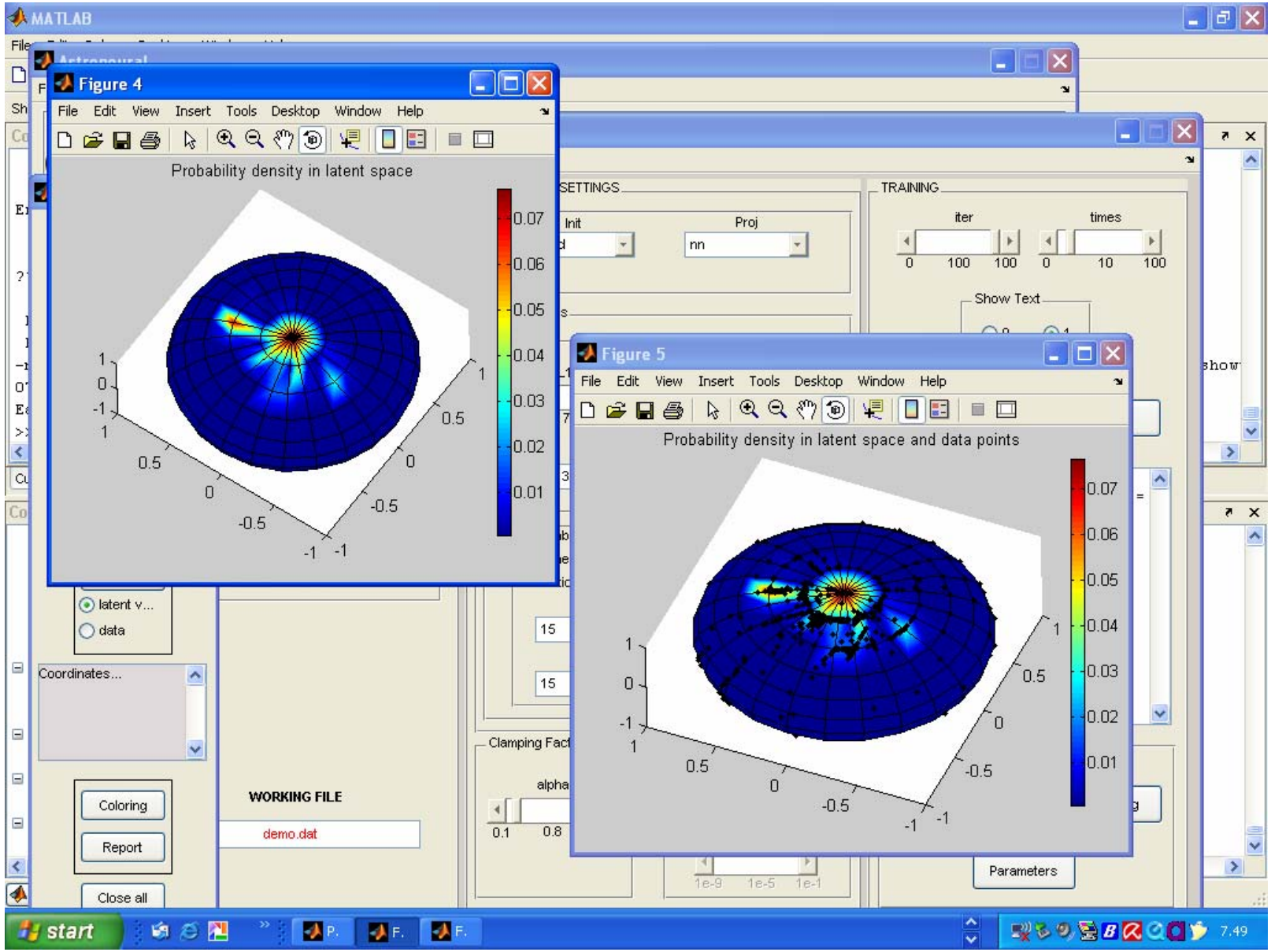
feature significance maps



U-matrix with labels

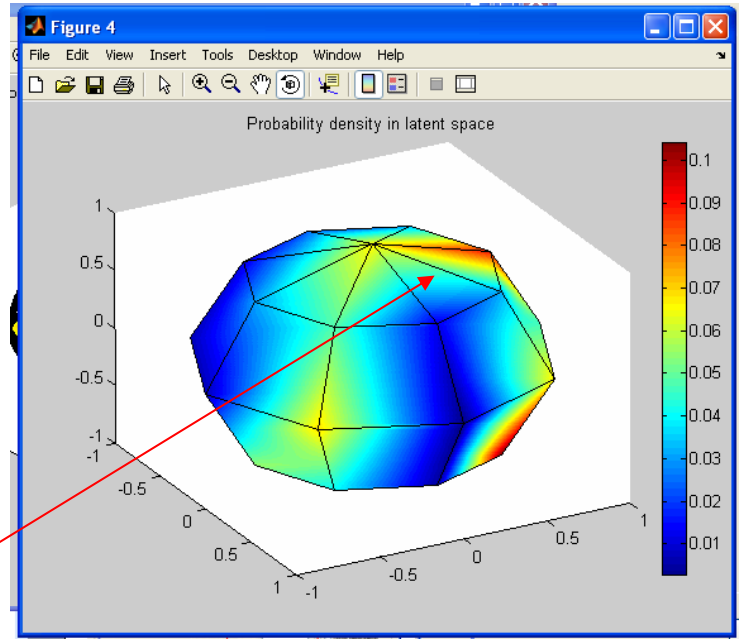


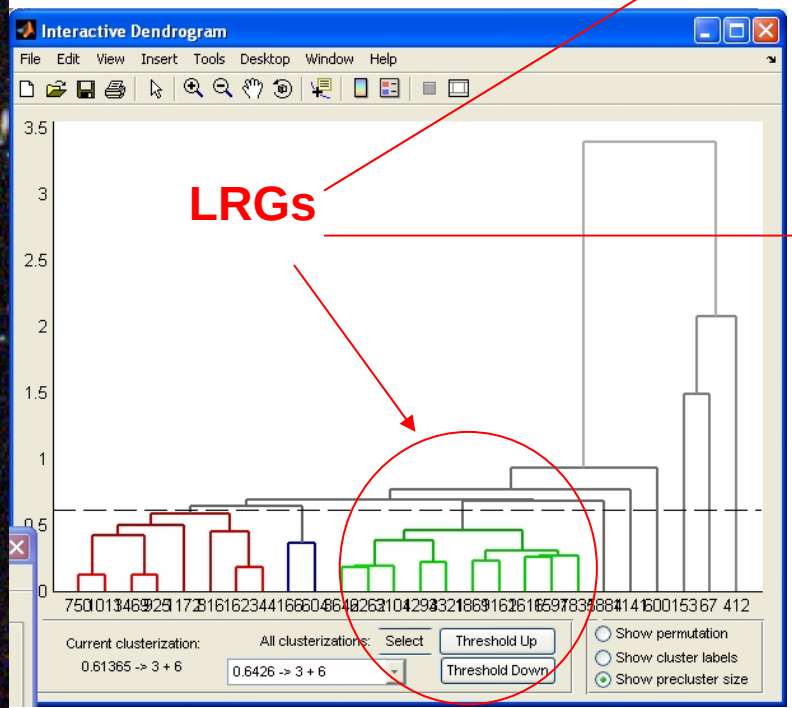
Probabilistic Principal Surfaces: probably solved (in part) within TOPCAT



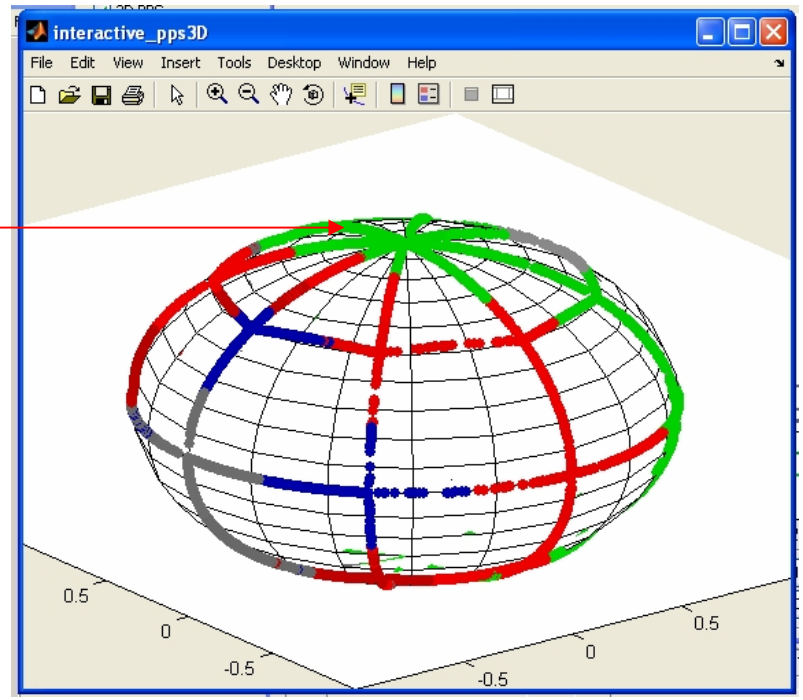
Tricky interactive visualization

To be solved





LRGs



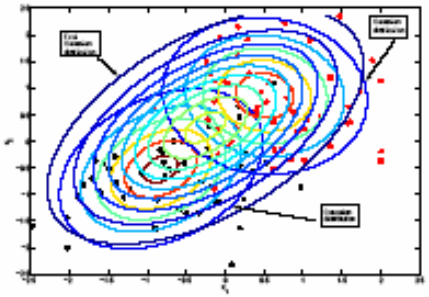
Mon. Not. R. Astron. Soc. **000**, 1–?? (2006) Printed 9 March 2007 (MN LATEX style file v2.2)

An universal classifier for the virtual observatory. I. The methods.

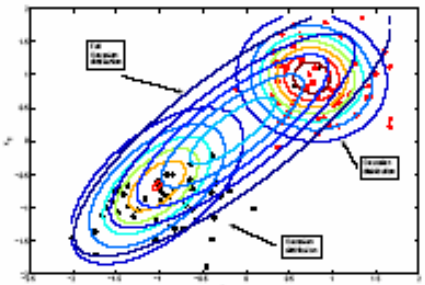
Longo Giuseppe[1,3,4], Brescia Massimo[3,4], D'Abrusco Raffaele[1,2],

De Filippis Elisabetta[1,3],Paolillo Maurizio[1,3,4], Staiano Antonino[5],

Tagliaferri Roberto[6]

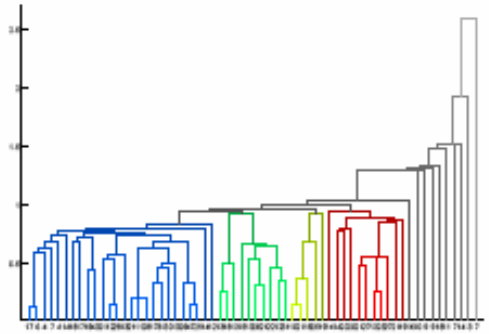


a)

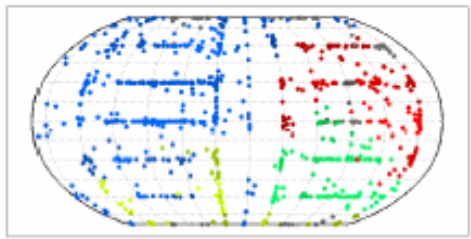


b)

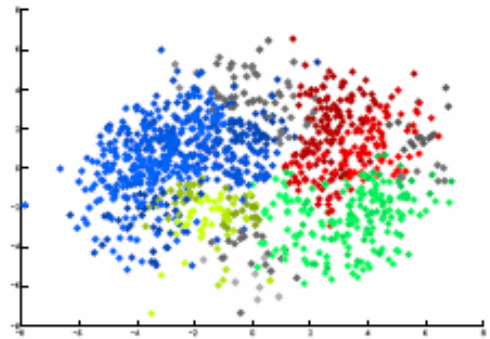
Figure 7. Gaussian distribution examples. Upper panel: Negentropy = 2.6261 using the G_1 function with $a_1 = 0.1$; Lower panel: Negentropy = 0.005 using the G_1 function with $a_1 = 0.1$



a)



b)



c)

Figure 9. NEC colored dendrogram.PPS 2-dimensional mapping and labeling.MDS 2-dimensional projection and labeling.

No DEMO:

if you wish you can play with Astroneural (tbx) on this computer

The remaining year will be entirely devoted to the following tasks:
Completion of the release 1.0 of the VO-Neural package (plasticized, etc)

ALL data mining tools are of limited use unless the MISSING DATA problem is solved

This will include: MLP, PPS, SOM and NEC modules together with all the needed visualization tools. We also plan to include immediately after also the RBF module.

MLP	**March 2007**
SOM	**June 2007**
PPS	**September 2007**
NEC + dendrogram	**September 2007**
NexT	**December 2007**

Several Other Science cases will be completed (all done within the VO):

- **3-D structure of nearby (z<0.25) universe using SDSS data and SOM clustering techniques**
- **AGN/QSO identification from SDSS and UKIDS surveys (both unsupervised and supervised MLP)**
- **Star/Galaxy separation with a priori information**