



## Photometric Redshifts with DAME

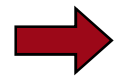
O. Laurino, R. D'Abrusco  
M. Brescia, G. Longo  
&  
DAME Working Group

**VO-Day ... in Tour  
@INAF**



## The general astrophysical problem

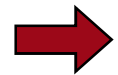
Due to new instruments and new diagnostic tools, the information volume grows exponentially



***Most data will never be seen by humans!***

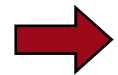
The need for data storage, network, database-related technologies, standards, etc.

Information complexity is also increasing greatly



***Most knowledge hidden behind data complexity is lost***

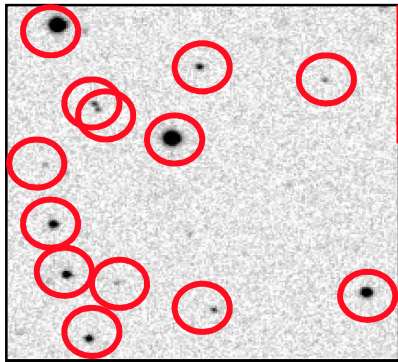
Most (all) empirical relationships known so far depend on 3 parameters ....  
Simple universe or rather human bias?



***Most data (and data constructs) cannot be comprehended by humans directly!***

The need for data mining, KDD, data understanding technologies, hyperdimensional visualization, AI/Machine-assisted discovery





Detect sources and measure their attributes (brightness, position, shapes, etc.)

$p = \{\text{isophotal, petrosian, aperture magnitudes, concentration indexes, shape parameters, etc.}\}$

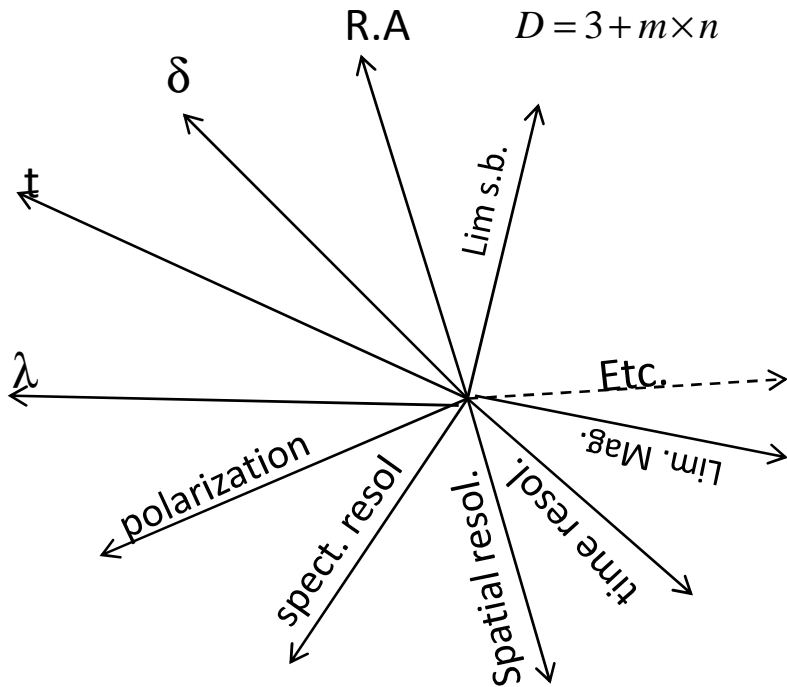
$$p^1 = \{RA^1, \delta^1, t, \{\lambda_1, \Delta\lambda_1, f_1^{1,1}, \Delta f_1^{1,1}, \dots, f_1^{1,m}, \Delta f_1^{1,m}\}, \dots, \{\lambda_n, \Delta\lambda_n, f_n^{1,1}, \Delta f_n^{1,1}, \dots, f_n^{1,m}, \Delta f_n^{1,m}\}\}$$

$$p^2 = \{RA^2, \delta^2, t, \{\lambda_1, \Delta\lambda_1, f_1^{2,1}, \Delta f_1^{2,1}, \dots, f_1^{2,m}, \Delta f_1^{2,m}\}, \dots, \{\lambda_n, \Delta\lambda_n, f_n^{2,1}, \Delta f_n^{2,1}, \dots, f_n^{2,m}, \Delta f_n^{2,m}\}\}$$

.....

$$p^N = \{RA^N, \delta^N, t, \{\lambda_1, \Delta\lambda_1, f_1^{N,1}, \Delta f_1^{N,1}, \dots, f_1^{N,m}, \Delta f_1^{N,m}\}, \dots\}$$

$$D = 3 + m \times n$$



**PARAMETER SPACE**

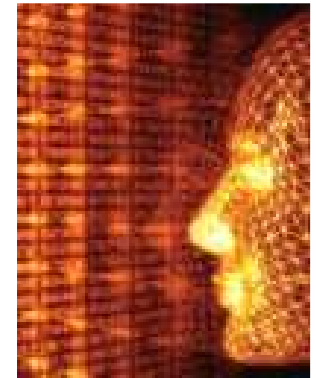
From the Data Mining point of view, any **observed (simulated) datum  $p$**  defines a point (region) in a subset of  $\mathbb{R}^N$ .

$$p \in \mathfrak{R}^N \quad N \gg 100$$



# From data to knowledge: KDD

## *Knowledge Discovery in Databases*



Data Gathering (e.g., from sensor networks, telescopes...)

### Data Farming:

Storage/Archiving  
Indexing, Searchability  
Data Fusion, Interoperability, ontologies, etc.

### Data Mining (or Knowledge Discovery in Databases):

Pattern or correlation search  
Clustering analysis, automated classification  
Outlier / anomaly searches  
Hyperdimensional visualization

### Data understanding

Computer aided understanding  
KDD  
Etc.

## New Knowledge

Database technologies

Key mathematical issues

Ongoing research



## Data Mining in the VO

- A new Interest group on Knowledge Discovery in Massive Data Sets was born inside the IVOA.
- The explosion of Data available (the “Data tsunami”) in the VO can be effectively dealt with using Data Mining, especially when the time variable comes into play.

But...The VO currently lacks general purpose tools for “server side” **massive data sets manipulation.**



### **Dame in the VO Framework**

- To provide the VO with an extensible, integrated environment for **Data Mining and Exploration**;
- Support of the VO standards and formats, especially for application interop (SAMP);
- To abstract the application deployment and execution, so to provide the VO with an “opaque” general purpose computing platform taking advantage of the modern technologies (e.g. Grid, Cloud, etc...).



# What is DAME



DAME is a joint effort between University Federico II, INAF-OACN, and Caltech aimed at implementing (as web application) a scientific gateway for data analysis, exploration, mining and visualization tools, on top of virtualized distributed computing environment.

<http://voneural.na.infn.it/>

**Technical and management info**  
**Documents**  
**Science cases**  
**Newsletter**

The screenshot shows the DAME website home page. The header features the 'DAta Mining & Exploration' logo and the 'DAME' logo. Below the header, there is a navigation menu with links for 'Home', 'Project', 'Documents', 'Download', and 'Contact Us'. The main content area is titled 'Data Mining & Exploration Project' and includes a 'News & Events' section with links to 'New DAME Prototype released', 'DAME Lecture @ IPAC-09', and 'DAME @ ITCAL-09 Conference'. There is also a 'Partners' section listing various institutions and a 'Related links' section. The footer contains copyright information for GMT/UTC: Mer 12:49.

<http://dame.na.infn.it/>

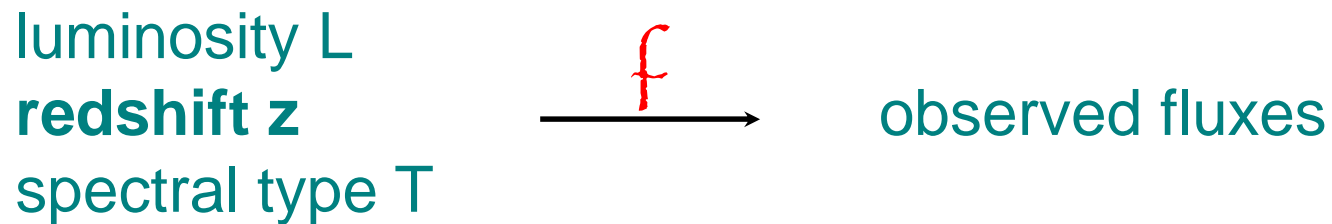
**Web application PROTOTYPE**

The screenshot shows the DAME web application interface. The user is logged in as Giuseppe Iengo. The interface includes a navigation menu, a 'My Experiments' section with a table of experiments, and a 'My Filestore' section. The 'My Experiments' table has columns for Name, Science case, and Mode. The 'My Filestore' section shows a directory listing for /Iengo/MAE with files like MAE.log, catiousie\_photons.cvx, and robust\_siams.txt.

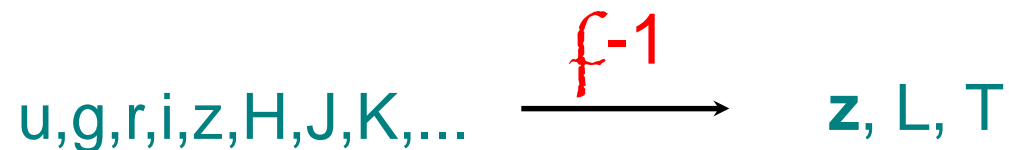
Name	Science case	Mode
gippo	psqso	psqso
giovavietto	psqso	psqso
giovavietto	psqso	psqso
giovavietto	psqso	psqso
MAE	psqso	psqso
psqso	mlpregression	mlpregression
psqso	mlclassification	mlclassification
psqso	psqso	psqso

# Photometric redshifts?

Multicolour photometry maps physical parameters:



If the relation can be inverted then:



The function can be approximated by regression in the photometric space. The accuracy of the photometric redshifts depend on two aspects:

**How the photometric filters cover the SED of the sources...**

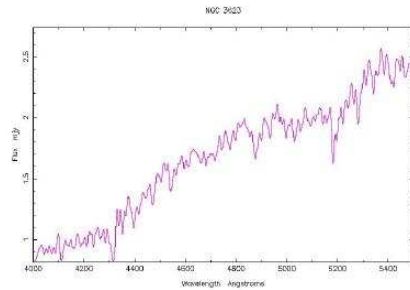
**How absorption and emission features relate to the photometric filters...**



# A Science case: photometric redshifts (sample from SDSS galaxies)

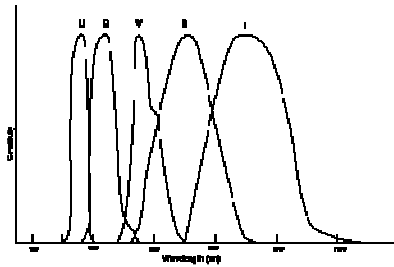
one of the main tools to investigate the spatial distribution of galaxies, i.e. to reconstruct the 3-D position of very large number of sources using only their photometric properties.

Spectral Energy Distribution convolved with band filters



Galaxy spectrum -  $F(\lambda)$

X



Photometric system -  $S_i(\lambda)$

=

$$m_U = -2.5 \log_{10} \frac{\int F(\lambda) S_U(\lambda) d\lambda}{\int S_U(\lambda) d\lambda} + c_U$$

$$m_B = -2.5 \log_{10} \frac{\int F(\lambda) S_B(\lambda) d\lambda}{\int S_B(\lambda) d\lambda} + c_B$$

Etc...

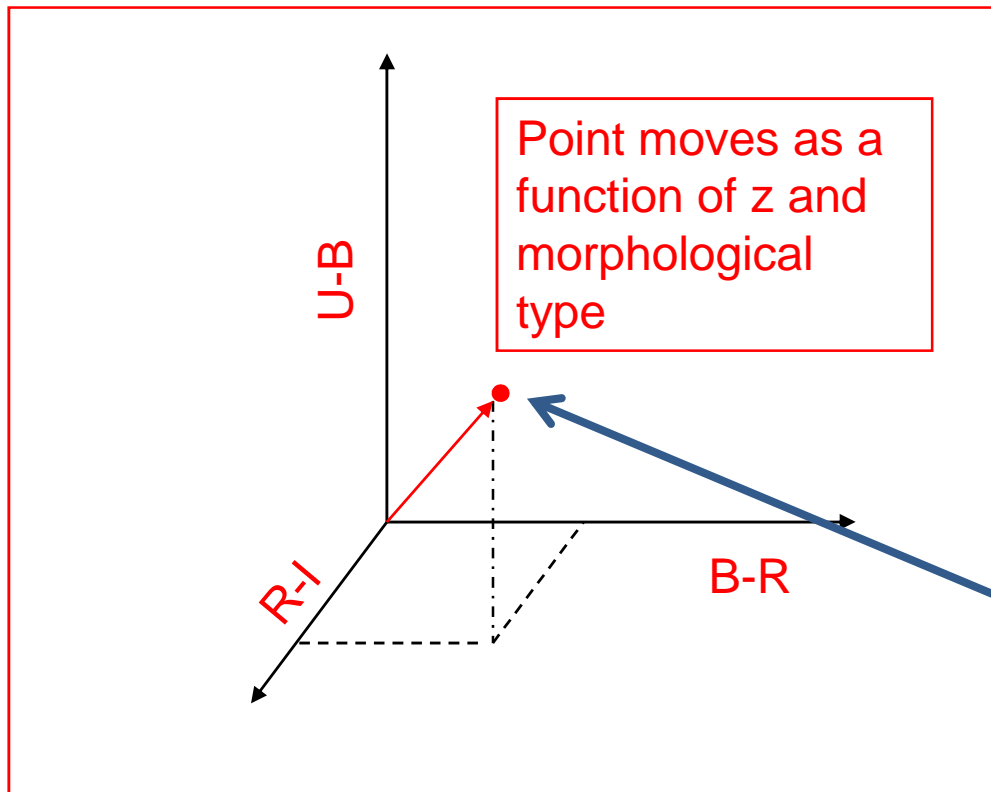


Color indexes

$$U - B \equiv m_U - m_B$$

$$B - R \equiv m_B - m_R$$

etc.



Zspec





# Photometric redshifts: the Data Mining approach

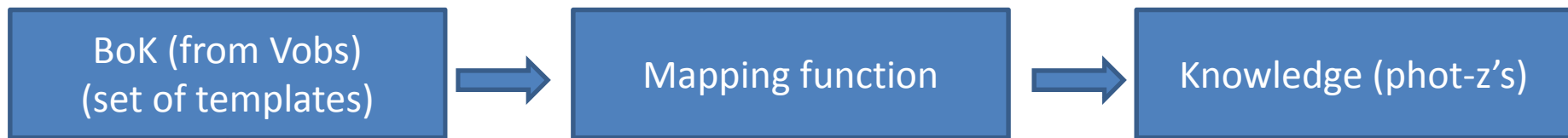
Photometric redshifts are treated as a regression problem (i.e. function approximation) in a multidimensional parameter space, hence a DM problem:

$\mathbf{X} \equiv \{x_1, x_2, x_3, \dots, x_N\}$  **input vectors**

$\mathbf{Y} \equiv \{x_1, x_2, x_3, \dots, x_M\}$  **target vectors**  $M \ll N$

**find**  $\hat{f}$ :  $\hat{\mathbf{Y}} = \hat{f}(\mathbf{X})$  **is a good approximation of Y**

**BoK = Base of Knowledge**



Observed Spectroscopic Redshifts  
Synthetic colors from theoretical SEDs  
Synthetic colors from observed SED's  
.....

Knowledge always reflects the biases in the BoK.

## **Interpolative**

Uneven coverage of parameter space

## **SED fitting**

Unknown or oversimplified physics?

Unjustified assumptions!

Computational intensive!

Possibly accurate, but is it worth?

.....



# Photometric Redshifts with Neural Networks

- Spectroscopic observations are the most accurate method to determine redshifts, but time consuming;
- Photometric sources often outnumber Spectroscopic ones up to 3 orders of magnitude (it may depend on the BoK);
- If we build a reliable BoK with spectroscopic data we can reproduce the functional mapping between photometric parameters and redshift;
- Zphot accuracy is adequate for several astronomical applications;

## Neural Networks advantages:

Fast;

Scale much better than any other method;

Learn by examples (BoK, in this case SDSS) and adapts easily to new data;

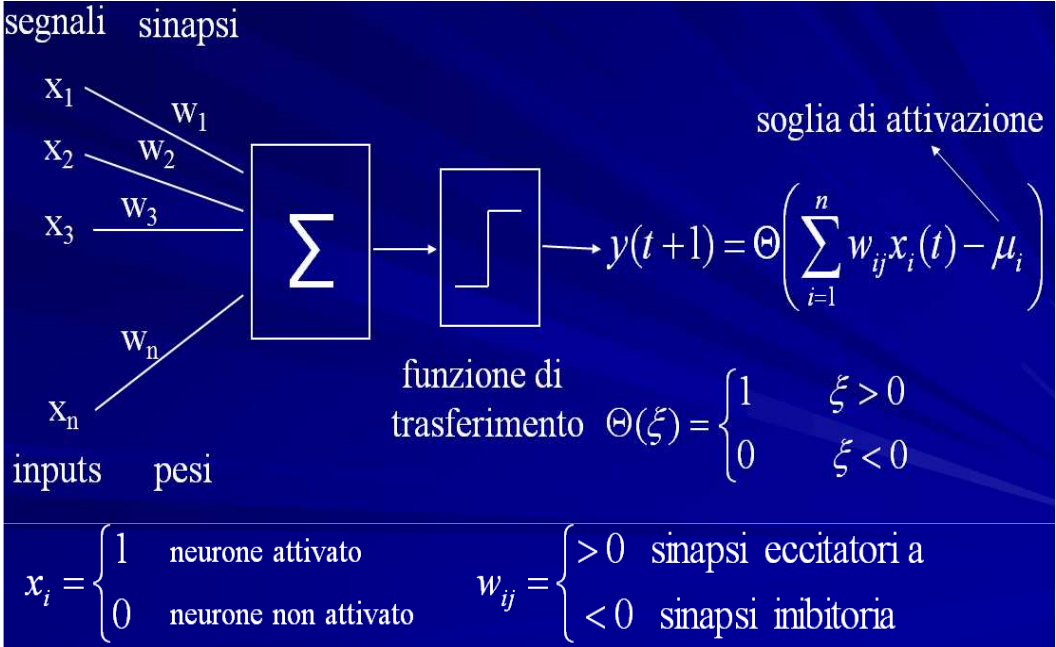
Do not require *a priori* assumption on the Spectral Energy Distributions of sources;

May be applied to all classes of extragalactic sources;



# MLP – The Architecture

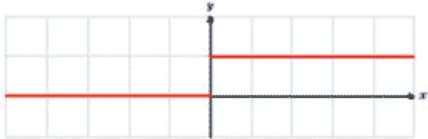
## McCulloch & Pitts artificial Neuron model



## Neuron activation function

Gradino

$$\Phi(A) = \begin{cases} 1 & A > \theta \\ 0 & \text{altrimenti} \end{cases}$$

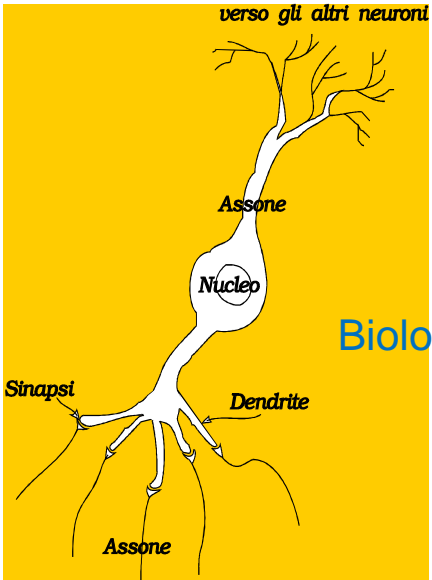
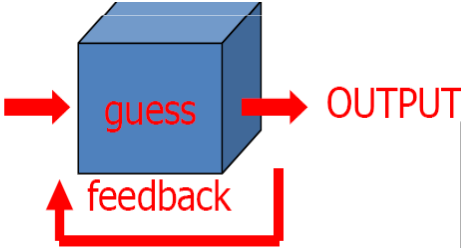
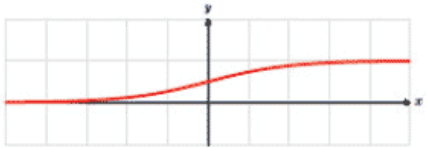


Lineare

$$\Phi(A) = kA$$

Sigmoide

$$\Phi(A) = \frac{1}{1 + e^{-kA}}$$

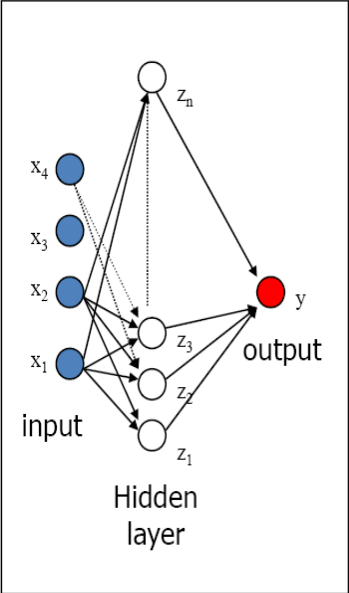


Biological Neuron

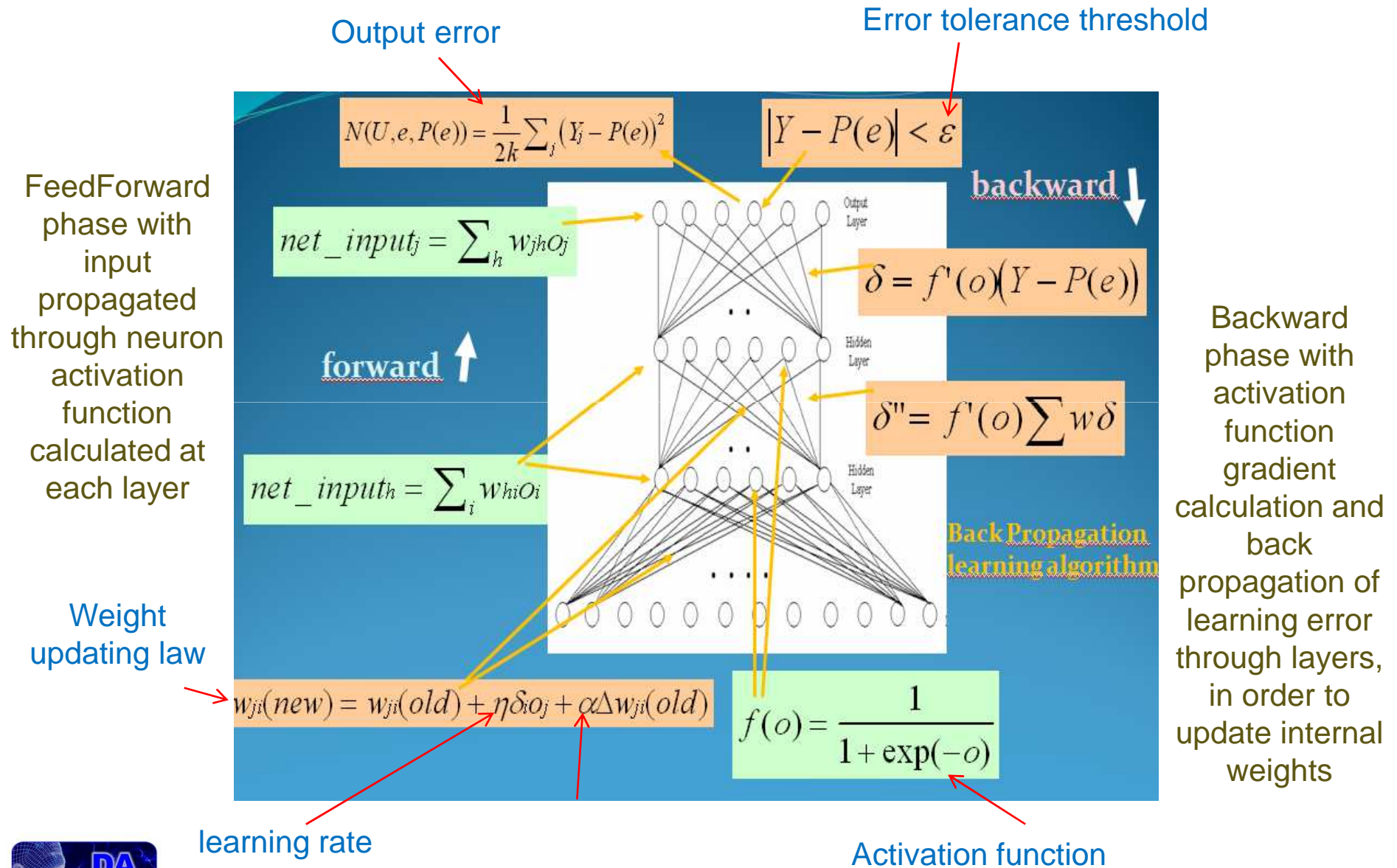


- input layer (n neurons)
- M hidden layer (1 or 2)
- Output layer (n' < n neurons)

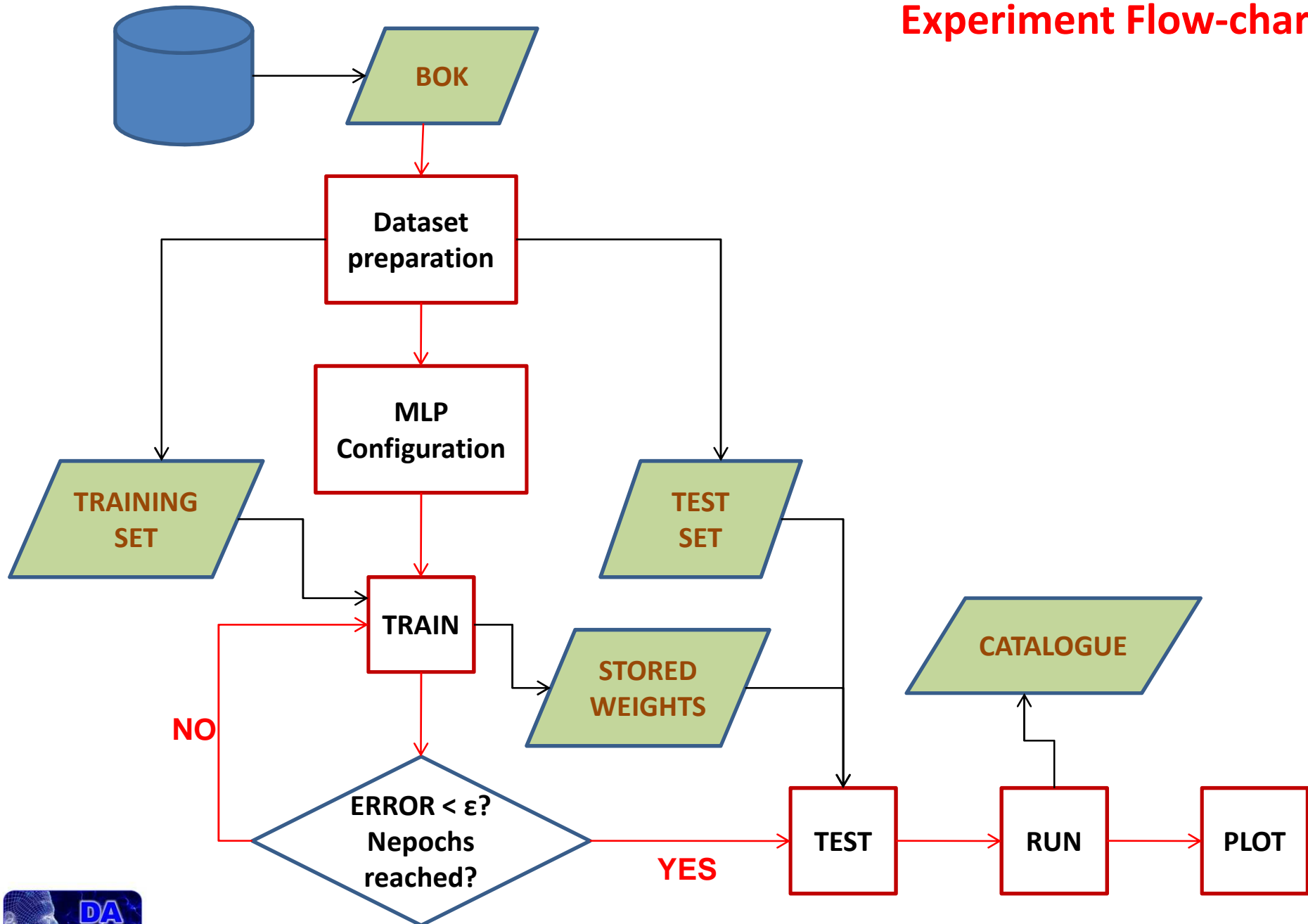
Neurons are connected via activation functions  
 Different NN's given by different topologies,  
 different activation functions, etc.



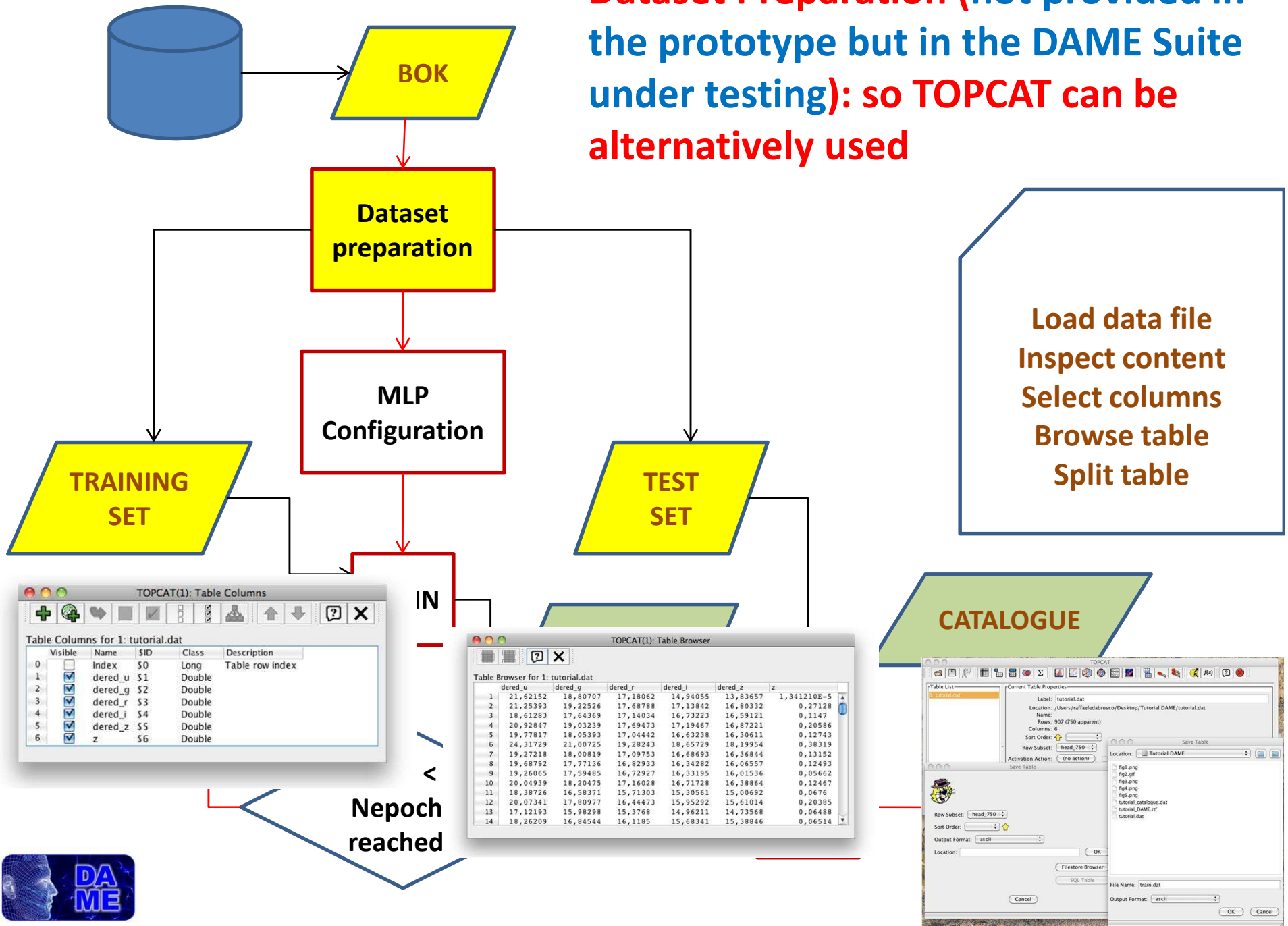
# MLP – Back Propagation learning algorithm



# Experiment Flow-chart



**Dataset Preparation (not provided in the prototype but in the DAME Suite under testing): so TOPCAT can be alternatively used**



TOPCAT(1): Table Columns

Visible	Name	\$ID	Class	Description
<input type="checkbox"/>	Index	\$0	Long	Table row index
<input checked="" type="checkbox"/>	dere_d_u	\$1	Double	
<input checked="" type="checkbox"/>	dere_d_g	\$2	Double	
<input checked="" type="checkbox"/>	dere_d_r	\$3	Double	
<input checked="" type="checkbox"/>	dere_d_i	\$4	Double	
<input checked="" type="checkbox"/>	dere_d_z	\$5	Double	
<input checked="" type="checkbox"/>	z	\$6	Double	

TOPCAT(1): Table Browser

	dere_d_u	dere_d_g	dere_d_r	dere_d_i	dere_d_z	z
1	21,62152	18,80707	17,18062	14,94055	13,83657	1,341210E-5
2	21,25393	19,22526	17,68788	17,13842	16,80332	0,27128
3	18,61283	17,64369	17,14034	16,73223	16,59121	0,1147
4	20,92847	19,03239	17,69473	17,19467	16,87221	0,20586
5	19,77817	18,05393	17,04442	16,63238	16,30611	0,12743
6	24,31729	21,00725	19,28243	18,65729	18,19954	0,38319
7	19,27218	18,00819	17,09753	16,68693	16,36844	0,13152
8	19,68792	17,77136	16,82933	16,34282	16,06557	0,12493
9	19,26065	17,59485	16,72927	16,33195	16,01536	0,05662
10	20,04939	18,20475	17,16028	16,71728	16,38864	0,12467
11	18,38726	16,58371	15,71303	15,30561	15,00692	0,0676
12	20,07341	17,80977	16,44473	15,95292	15,61014	0,20385
13	17,12193	15,98298	15,3768	14,96211	14,73568	0,06488
14	18,26209	16,84544	16,1185	15,68341	15,38846	0,06514

CATALOGUE

TOPCAT

Table List: tutorial.dat

Current Table Properties:

Label: tutorial.dat  
Location: Users/raffaellaabrusco/Desktop/Tutorial DAME/tutorial.dat  
Name: tutorial.dat  
Rows: 907 (750 apparent)  
Columns: 6  
Sort Order: head\_750  
Row Subset: head\_750  
Activation Action: (no action)

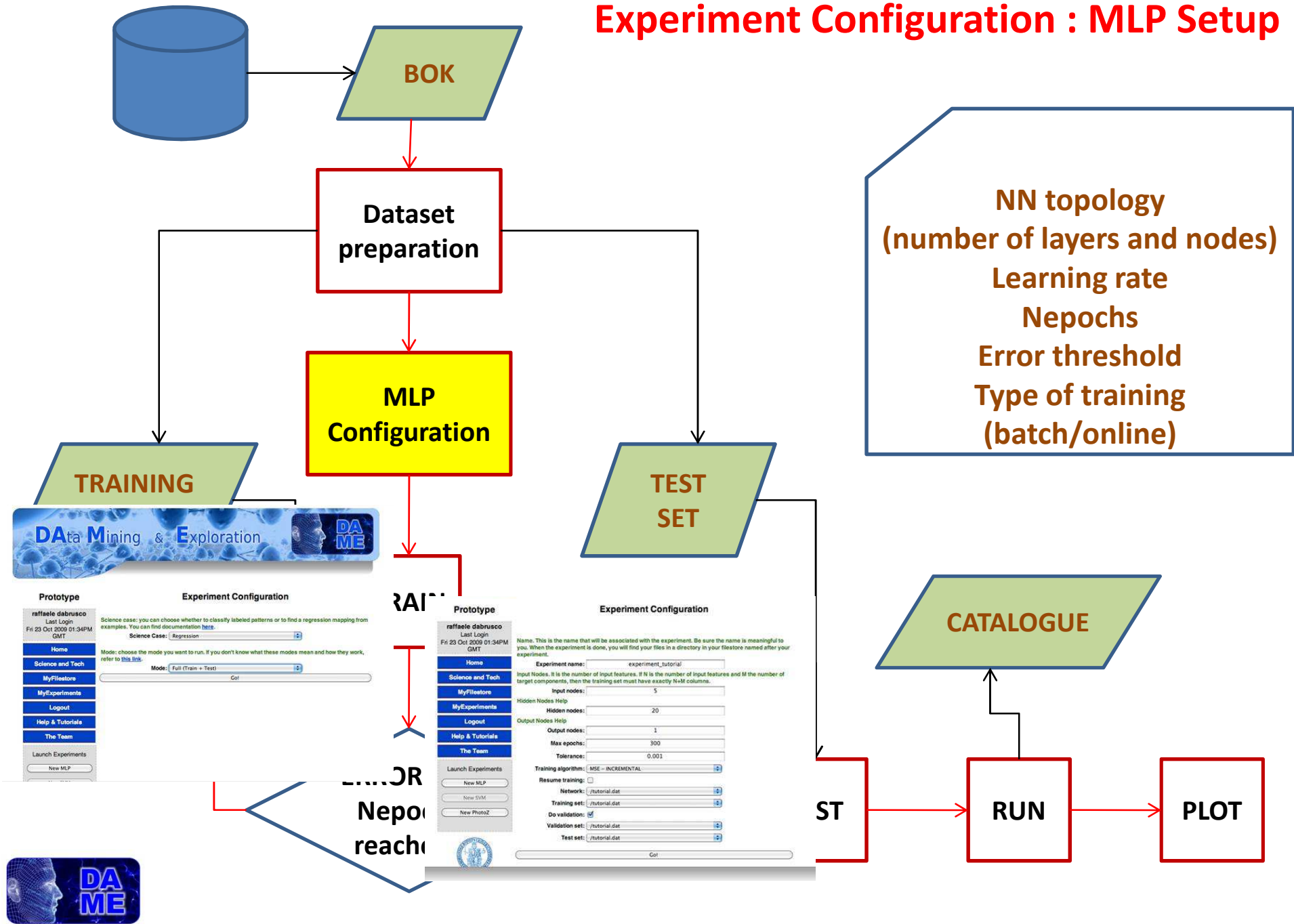
Save Table

Location: Tutorial DAME

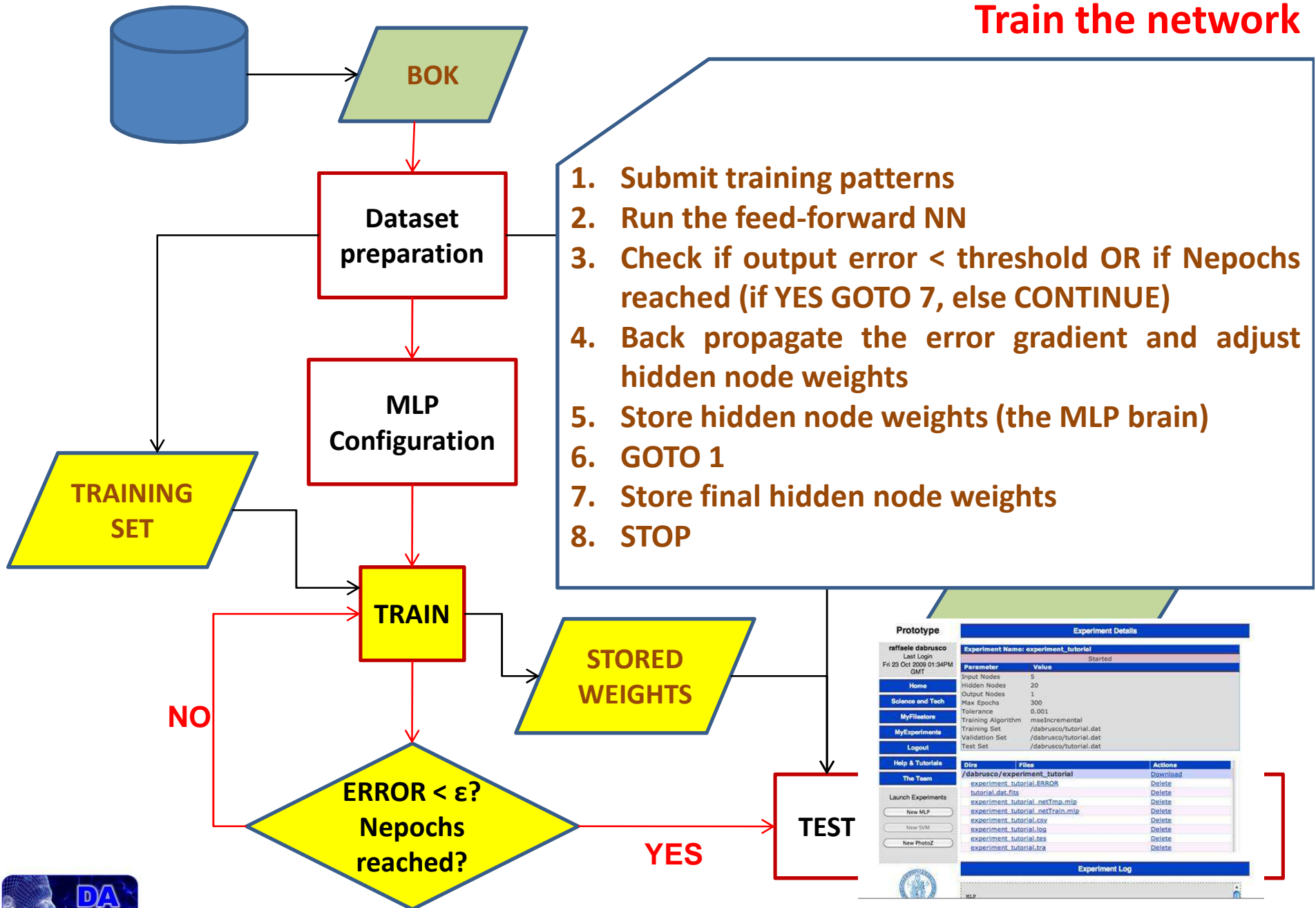
File Name: train.dat  
Output Format: ascii



# Experiment Configuration : MLP Setup



# Train the network



Experiment Details	
Experiment Name:	experiment_tutorial
Parameter	Value
Input Nodes	5
Hidden Nodes	20
Output Nodes	1
Max Epochs	300
Tolerance	0.001
Training Algorithm	mseIncremental
Training Set	/dabrusco/tutorial.dat
Validation Set	/dabrusco/tutorial.dat
Test Set	/dabrusco/tutorial.dat

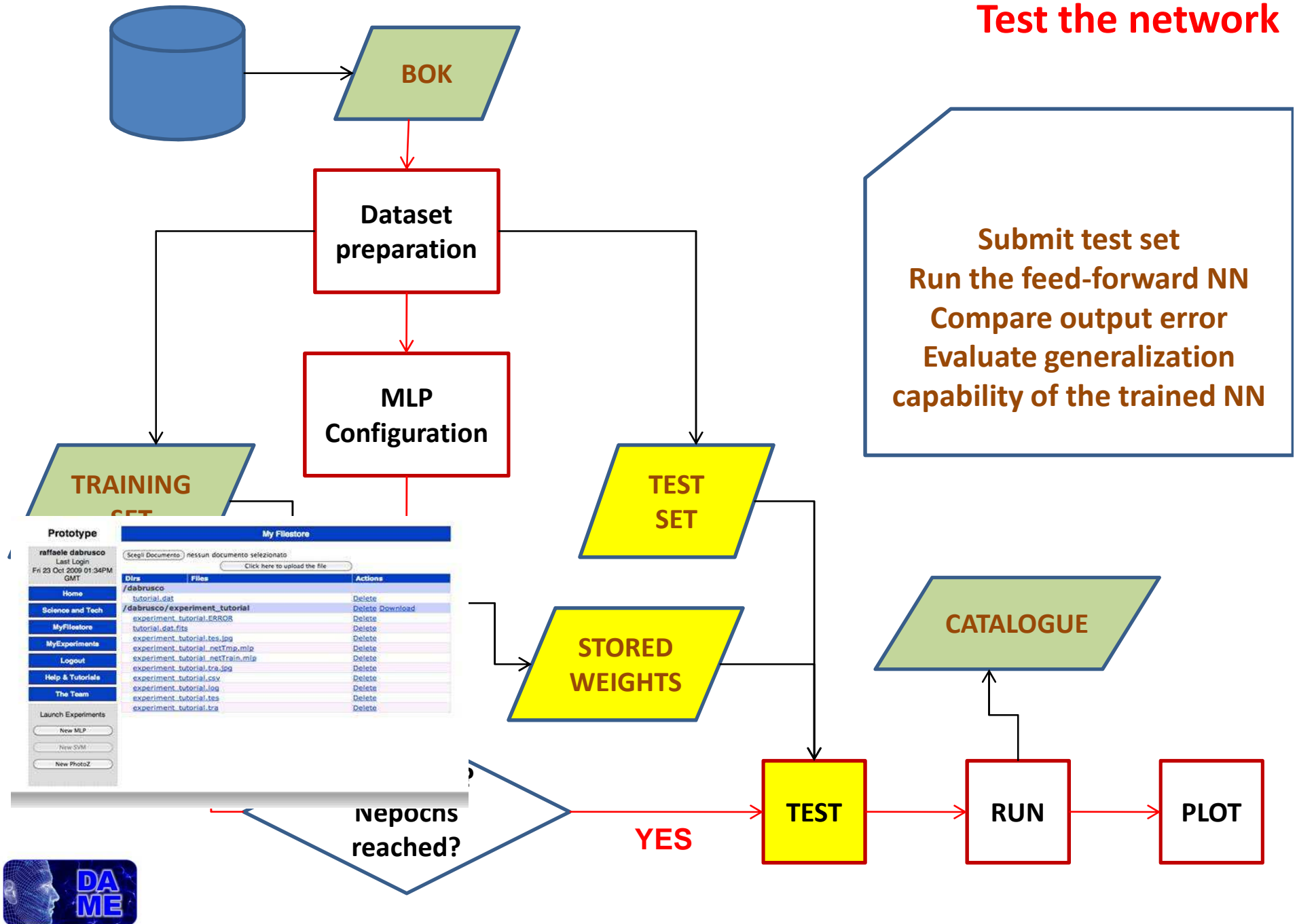
  

Dir	Files	Actions
/dabrusco/experiment_tutorial	experiment_tutorial.ERRORS	Download
	tutorial.dat.files	Delete
	experiment_tutorial_netTmp.mlp	Delete
	experiment_tutorial_netTrain.mlp	Delete
	experiment_tutorial.csv	Delete
	experiment_tutorial.log	Delete
	experiment_tutorial.tes	Delete
	experiment_tutorial.tsa	Delete





# Test the network



# Run the network

Run NN like a generic deterministic one-shot algorithm  
Generate output Catalogue  
Plot and analyze output

Prototype  
raffaële dabrusco  
Last Login: Fri 23 Oct 2009 01:34PM GMT  
Home  
Science and Tech  
MyFilestore  
MyExperiments  
Logout  
Help & Tutorials  
The Team  
Launch Experiments  
New MLP

Experiment Configuration  
Science case: you can choose whether to classify labeled patterns or to find a regression mapping from examples. You can find documentation [here](#).  
Science Case: Regression  
Mode: Run  
Go!

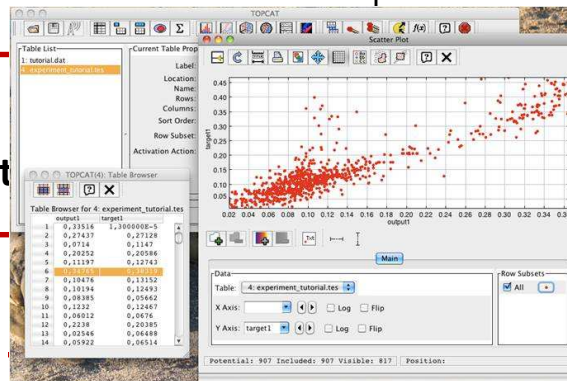


Dataset preparation

Prototype  
raffaële dabrusco  
Last Login: Fri 23 Oct 2009 01:34PM GMT  
Home  
Science and Tech  
MyFilestore  
MyExperiments  
Logout  
Help & Tutorials  
The Team  
Launch Experiments  
New MLP

Experiment Configuration  
Name: This is the name that will be associated with the experiment. Be sure the name is meaningful to you. When the experiment is done, you will find your files in a directory in your filestore named after your experiment.  
Experiment name: catalogue\_tutorial  
Network: /experiment\_tutorial/experiment\_tutorial\_netTra  
Data set: /tutorial\_catalogue.dat  
Go!

Preparation



## TRAINING

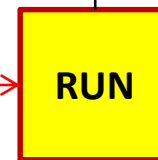
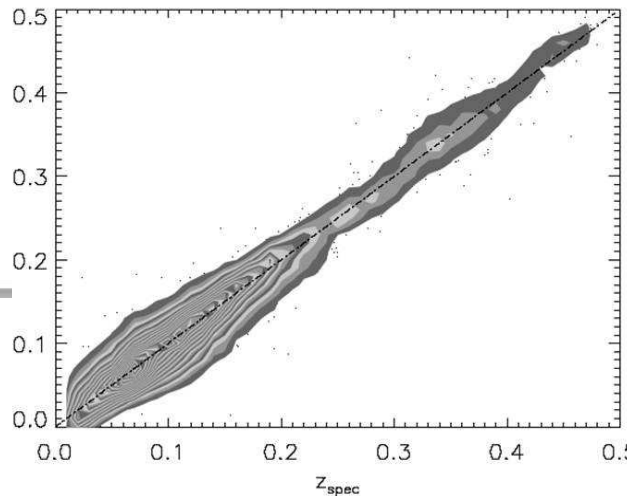
Prototype  
raffaële dabrusco  
Last Login: Fri 23 Oct 2009 02:21PM GMT  
Home  
Science and Tech  
MyFilestore  
MyExperiments  
Logout  
Help & Tutorials  
The Team  
Launch Experiments  
New MLP  
New SVM  
New PhotoZ

Experiment Details  
Experiment Names: tutorial\_catalogue  
Parameter: Finished  
Data Set: /dabrusco/tutorial\_catalogue.dat

Dirs	Files	Actions
/dabrusco/tutorial_catalogue		Download
tutorial_catalogue_log		Delete
tutorial_catalogue_run		Delete
tutorial_catalogue_dat.fits		Delete
tutorial_catalogue_ERROR		Delete

Experiment Log  
MLP  
Executing option: RUN  
Input nodes: 0  
Output nodes: 0  
Nodes in hidden layer: 0  
Maximum epochs: 0  
Problem case: Regression  
Error tolerance: 0  
Input network name: /var/www/dameMLP/data/users/dabrusco/omar\_train/omar\_train\_netTrain.mlp  
Training dataset: empty  
Validation dataset: empty  
Testing dataset: tutorial\_catalogue/tutorial\_catalogue.dat.fits  
Output profile filename: tutorial\_catalogue

Neuro reach



# Follow on-line tutorial

**DAME help & tutorials - HowTo - Mozilla Firefox**

File Modifica Visualizza Cronologia Segnalibri Strumenti Aiuto

http://pcdevauc.na.infn.it:9000/photoztut/

Più visitati Come iniziare Ultime notizie HotMail gratuita Personalizza collegam... Personalizzazione coll... Windows WindowsMedia

EuroVO-AIDA - VO-Day ... in Tour Webmail DAME help & tutorials - HowTo DAME help & tutorials - HowTo

## DAta Mining & ExploratiON

### Prototype

Photometric redshifts with DaME

#### 1. The Scientific Problem

Photometric redshifts have become one of the main tools to investigate the spatial distribution of galaxies, since they are necessary to reconstruct the 3-dimensional position of very large number of sources using only their photometric properties. One application of zphot are amazing maps of the Universe like this:

Home

Science and Tech

GMT/UTC: Lun 09:25 Italia: Lun 10:25 Los Angeles: Lun 01:25 Cile: Lun 06:25 Completato

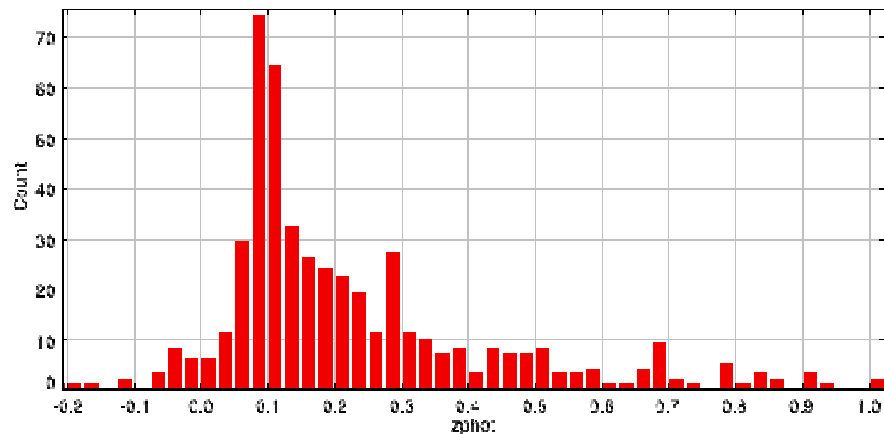
start DAME help & tutorials... ICFDT\_novembre2009 Microsoft PowerPoint ... IT 10:25

## Use case II

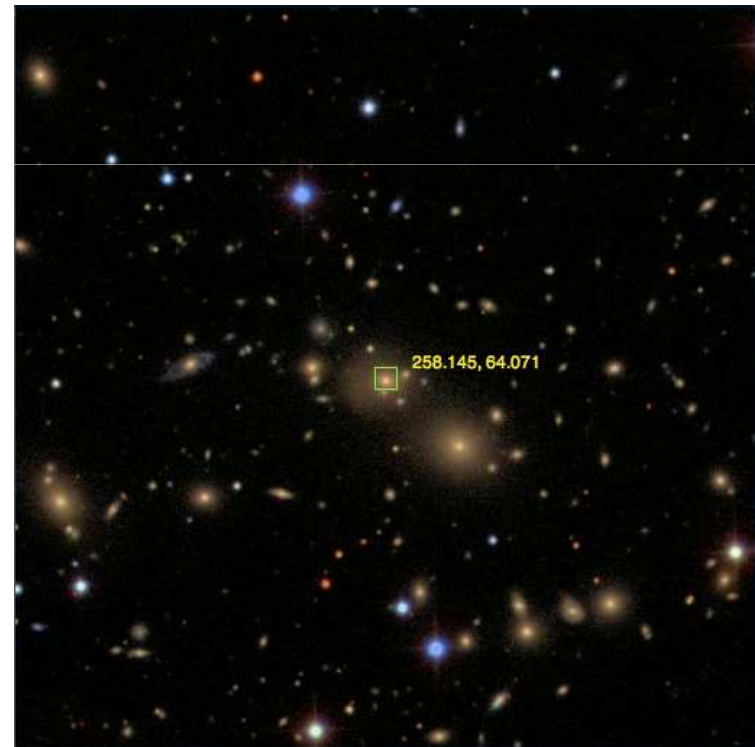


A possible application of photometric redshifts is the selection of galaxies belonging to bound structures like clusters or groups (see, for example, [Capozzi et al. 2009]). Let's see how we can find out if a given observed "overdensity" of galaxies in a field is likely associated to a real structure.

**Photometric redshifts can be used as a valid alternative technique or side-by-side to other methods based** on the photometric properties of galaxies (for example, the "red-sequence" method [Gladders & Yee 2000]).



We will consider a **well known rich cluster of galaxies (Abell 2255)** and **explore a field containing this cluster using SDSS photometry and our own photometry redshifts**



You will use **DAME, Topcat and Aladin**, with which you should already be **comfortable!**

# Tutorial – Part II



Follow the step-by-step instructions on the DAME prototype web page...



## Prototype

Username

Password

Home

Science and Tech

Sign Up!

Help & Tutorials

The Team



## Using photometric redshifts to spot galaxy clusters


A possible application of photometric redshifts relates to the selection of sources belonging to bound structures like clusters or groups of galaxies.


A close projected group of galaxies on the sky can be either produced by sources which are gravitationally connected to each other (i.e. physically close in both projected and redshift spaces), or can be the result of a chance superposition of physically unrelated galaxies (sources which are close in the sky projection but have different redshifts).

In order to check this out, **photometric redshifts can be used as a valid alternative technique or side-by-side to other classical methods** based on the photometric properties of galaxies (for example, the "red-sequence" method [Gladders & Yee 2000]).

Let's see how we can find out if a given "overdensity" of galaxies in a field is likely associated to a real structure. We will consider a **well known rich cluster of galaxies called Abell 2255** (after the name of the astronomer who compiled a large catalogue of clusters of galaxies) and **explore a field containing this cluster using photometry redshifts**.

We will use **DAME, Topcat and Aladin**, but only the main steps of the process will be described, so you'll have to work the details out by yourself!

1. First of all, **you need a catalogue of photometric sources** in the area of the sky where the cluster of galaxy is more likely to be found. We will use the querying service provided by Topcat to look, on VizieR, for sources in the Abell 2255 field. Click on the "Open new table" icon 

in the main bar of Topcat, then click on the VizieR service icon ; now you can perform a cone search around the Abell 2255 cluster position in the sky by filling the fields in the VizieR window with the following values: RA 258.145 (deg), DEC 64.071 (deg), Radius: 200" (change