Astronomical data Mining:
An application to the photometric redshifts of galaxies and QSOs

Giuseppe Longo
University Federico II in Napoli, longo@na.infn.it

In coll. with
M. Brescia, R. D’Abrusco, O. Laurino & the DAME team
The company which is making the journey...

University Federico II
- Massimo Brescia (project manager)
- Stefano Cavuoti
- Raffaele D’Abru sco
- Giancarlo D’Angelo (GRID)
- Natalia V. Deniskina
- Michelangelo Fiore (student)
- Mauro Garofalo
- Omar Laurino (project engineer)
- Giuseppe Longo (Principal Investigator)
- Francesco Manna (student)
- Alfonso Nocella
- Civita Vellucci (student)

Caltech
- G.S. Djorgovski
- C. Donalek
- A. Mahabal
Summary of the talk

- Data Mining and astronomy
  - Why DAME and what is DAME
    - Photometric redshifts and galaxy phot-z’s in DAME
      - A DM “pipeline” for QSO’s (candidate selection and phot-z’s)
    - Some general considerations on the future
Most of us have done it for their whole life

Compilation of photoelectric multiaperture photometry

Through standard luminosity profile curves to derive “Extrapolation corrections”

.... in order to derive Total Magnitudes of galaxies
Data Mining is not only new astronomy.

In many cases \textit{but NOT ALL} it is just the name we give to rather usual stuff when it needs to be performed fast and on billions of records of COMPLEX data.

Human brain is not sufficient.

Machine learning methods
Data Mining is the activity of extracting **USEFUL** information from **COMPLEX** data using Statistical Pattern Recognition and Machine Learning methods.

### DM Taxonomy

1. To catalogue the known (classification)
2. Characterize the unknown (clustering)
3. Find functional dependencies (regression)
4. Find exceptions (outliers)

### Supervised Methods

Patterns are learnt from extensive set of templates (Base of Knowledge = BoK)

### Unsupervised Methods

Patterns are discovered using the data themselves
The scientific exploitation of a multi band, multiepoch (K epochs) universe implies to search for patterns, trends, etc. among \( N \) points in a \( D \times K \) dimensional parameter space:

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

\[
p^1 = \{ RA^1, \delta^1, t, \{ \lambda_1, \Delta \lambda_1, f_1^{1,1}, \Delta f_1^{1,1}, \ldots, f_1^{1,m}, \Delta f_1^{1,m} \}, \ldots, \{ \lambda_n, \Delta \lambda_n, f_n^{1,1}, \Delta f_n^{1,1}, \ldots, f_n^{1,m}, \Delta f_n^{1,m} \} \}
\]

\[
p^2 = \{ RA^2, \delta^2, t, \{ \lambda_1, \Delta \lambda_1, f_1^{2,1}, \Delta f_1^{2,1}, \ldots, f_1^{2,m}, \Delta f_1^{2,m} \}, \ldots, \{ \lambda_n, \Delta \lambda_n, f_n^{2,1}, \Delta f_n^{2,1}, \ldots, f_n^{2,m}, \Delta f_n^{2,m} \} \}
\]

\[\ldots\]

\[
p^N = \{ RA^N, \delta^N, t, \{ \lambda_1, \Delta \lambda_1, f_1^{N,1}, \Delta f_1^{N,1}, \ldots, f_1^{N,m}, \Delta f_1^{N,m} \}, \ldots \}
\]

\[D = 3 + m \times n\]

\[N > 10^9, D\gg 100, K>10\]
Any observed (simulated) datum \( p \) defines a point (region) in a subset of \( \mathbb{R}^N \). Es:

- RA and dec
- time
- \( \lambda \)
- experimental setup (spatial and spectral resolution, limiting mag, limiting surface brightness, etc.) parameters
- fluxes
- polarization
- Etc.

\[ p \in \mathbb{R}^N \quad N \gg 100 \]

The parameter space concept is crucial to:

1. Guide the quest for new discoveries (observations can be guided to explore poorly known regions), ...
2. Find new physical laws (patterns)
3. Etc,
The computational cost of DM:

N = no. of data vectors, D = no. of data dimensions
K = no. of clusters chosen, $K_{\text{max}}$ = max no. of clusters tried
I = no. of iterations, M = no. of Monte Carlo trials/partitions

K-means: $K \times N \times I \times D$
Expectation Maximisation: $K \times N \times I \times D^2$
Monte Carlo Cross-Validation: $M \times K_{\text{max}}^2 \times N \times I \times D^2$
Correlations $\sim N \log N$ or $N^2$, $\sim D^k$ ($k \geq 1$)
Likelihood, Bayesian $\sim N^m$ ($m \geq 3$), $\sim D^k$ ($k \geq 1$)
SVM $> \sim (NxD)^3$
ASTROINFORMATICS

GRID, CLOUD, TERAGRID, etc.

Computing infrastructures

ASTROINFORMATICS (emerging field)

Fast, efficient, innovative algorithms
WEKA, DAME, etc.

Implementation and access to DR

IPAC, CDS, ADSC, etc.
### Variable characteristics and types

<table>
<thead>
<tr>
<th>Variable Type</th>
<th>Characteristics</th>
<th>Type</th>
<th>Operation/example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quantitative</td>
<td>Numerical with ordering relationship and possibility to define a metric</td>
<td>Actual measurement</td>
<td>Regression&lt;br&gt;Photometric redshifts</td>
</tr>
<tr>
<td>Categorical (non-ordered)</td>
<td>Membership into a finite number of classes. No ordering relationship.</td>
<td>Numerical codes (targets) arbitrarily ordered</td>
<td>Classification&lt;br&gt;Search for peculiar objects, QSO’s, Star/galaxy, etc.</td>
</tr>
<tr>
<td>Ordered categorical</td>
<td>Classes ordered by a relationship but there is no metric</td>
<td>Numerical codes non arbitrarily ordered</td>
<td>Classification&lt;br&gt;Morphological and physical classification of galaxies, etc.</td>
</tr>
</tbody>
</table>
DAME is a joint effort between University Federico II, INAF and Caltech aimed at: implementing (as web application) a suite of data exploration, data mining and data visualization tools.

http://dame.na.infn.it/
Web application PROTOTYPE
http://voneural.na.infn.it/
Documents
What is the real DAME

1. The real thing

P.M. Massimo Brescia

1. User friendly
2. Flexible and expandable
3. Running also on HPC or distributed systems

Will substitute the prototype at the end of October 2009
PART II - applications of DAME to observational cosmology
Photometric redshifts of galaxies and QSO’s
Selection of candidate quasars

Laurino et al., 2009, Thesis
1. Photometric redshifts of galaxies

\[ m_U = -2.5 \log_{10} \frac{\int F(\lambda)S_U(\lambda) d\lambda}{\int S_U(\lambda) d\lambda} + c_u \]

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

Etc...

Color indexes

\[ U - B \equiv m_U - m_B \]
\[ B - R \equiv m_B - m_R \]

etc.
Photometric redshifts are always a function approximation hence a DM problem:

$$X \equiv \{x_1, x_2, x_3, \ldots, x_N\} \quad \text{input vectors}$$

$$Y \equiv \{x_1, x_2, x_3, \ldots, x_M\} \quad \text{target vectors} \quad M << N$$

find $$\hat{f}: \hat{Y} = \hat{f}(X)$$ is a good approximation of $$Y$$
Data used in the science cases:

**SDSS:** $10^8$ galaxies in 5 bands;
- BoK: spectroscopic redshifts for $10^6$ galaxies
- BoK: incomplete and **biased**.

**UKIDDS:** overlap with SDSS

**GALEX:** overlap with SDSS
1. Photometric redshifts of galaxies

**SED fitting**
Templates from synthetic colors obtained from theoretical SED’s
Mapping function from simple interpolation

**Interpolative**
Templates from synthetic colors obtained from theoretical SED’s
Mapping function from Bayesian inference

\[ \sigma = 0.051 \]
\[ \Delta z = 0.0144 \]

\[ \sigma = 0.0415 \]
• the color space is partitioned (KD-tree - a binary search tree) into cells containing the same number of objects from the training set
• In each cell a second order polynomial is fit to BoK.

\[ \sigma = 0.023 \]

**Fig. 4.—** On the right we plot a 2 dimensional demonstration of the color space partitioning. In each of these cells we applied the polynomial fitting technique to estimate redshifts. The left figure shows the results.
MLP or Multi Layers Perceptron

- 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.

IPAC-Pasadena, August 5 2009
1. Photometric redshifts of galaxies

SDSS-DR4/5 - SS

Training 60%
Validation 20%
Test set 20%

MLP, 1(5), 1(18)

0.01<Z<0.25
MLP, 1(5), 1(23)

0.25<Z<0.50
MLP, 1(5), 1(24)

99.6% accuracy

IPAC-Pasadena, August 5 2009
Photometric redshifts of galaxies

$\sigma = 0.0183$

SDSS – DR4/5 - LRG

Catalogue can be downloaded from the DAME site.

D’Abrusco et al. 2007
<table>
<thead>
<tr>
<th>type</th>
<th>method</th>
<th>data</th>
<th>$\Delta z_{rms}$</th>
<th>Notes</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CWW</td>
<td>EDR</td>
<td>0.0666</td>
<td></td>
<td>(Csabai et al. 2003)</td>
</tr>
<tr>
<td>SEDF</td>
<td>Bruzual-CHarlot</td>
<td>EDR</td>
<td>0.0552</td>
<td></td>
<td>(Csabai et al. 2003)</td>
</tr>
<tr>
<td></td>
<td>Interpolated</td>
<td>EDR</td>
<td>0.0451</td>
<td></td>
<td>(Csabai et al. 2003)</td>
</tr>
<tr>
<td></td>
<td>Polynomial</td>
<td>EDR</td>
<td>0.0318</td>
<td></td>
<td>(Csabai et al. 2003)</td>
</tr>
<tr>
<td></td>
<td>KD-tree</td>
<td>EDR</td>
<td>0.0254</td>
<td></td>
<td>(Csabai et al. 2003)</td>
</tr>
<tr>
<td></td>
<td>ANNz</td>
<td>EDR</td>
<td>0.0229</td>
<td></td>
<td>(Collister &amp; Lahav 2004)</td>
</tr>
<tr>
<td>ML</td>
<td>SVM</td>
<td>EDR</td>
<td>0.027</td>
<td></td>
<td>(Wadadekar 2004)</td>
</tr>
</tbody>
</table>
General galaxy sample

\( \sigma = 0.0208 \)
\( \Delta z = -0.0029 \)

LRG sample

\( \sigma = 0.0178 \)
\( \Delta z = -0.0011 \)

Non LRG only

\( \sigma = 0.0363 \)
\( \Delta z = -0.0030 \)

D'Abrusco et al. 2007

1. Photometric redshifts of galaxies
1. Photometric redshifts of galaxies

[Graphs showing distribution of photometric redshifts for different redshift ranges for both general galaxy sample and LRG sample.]
What do we learn if the BoK is biased:

- At high z LRG dominate and interpolative methods are not capable to “generalize” rules
- An unique method optimizes its performances on the parts of the parameter space which are best covered in the BoK

Step 1: unsupervised clustering in parameter space

Step 2: supervised training of different NN for each cluster

Step 3: output of all NN go to WGE which learns the correct answer

Laurino et al. 2009a,2009b
**Laurino et al. 2009a, 2009b**

$$\sigma = 0.0172$$

No systematic trends
PART II - applications to observational cosmology
Photometric selection of candidate QSO’s
(as a clustering problem)

Traditional way to look for candidate QSO in 3 band survey

- Cutoff line
- Candidate QSOs for spectroscopic follow-up’s
- Ambiguity zone

Adding one feature improves separation…

PPS projection of a 21-D parameter space showing as blue dots the candidate quasars. Notice better disentanglement
SDSS QSO candidate selection algorithm (Richards et al, 2002) targets star-like objects as QSO candidate according to their position in the SDSS colours space (u-g,g-r,r-i,i-z), if one of these requirements is satisfied:

- QSOs are supposed to be placed >4σ far from a cylindrical region containing the “stellar locus” (S.L.), where σ depends on photometric errors.

OR

- QSOs are supposed to be placed inside the inclusion regions, even if not meeting the previous requirement.

\[ c = 95\%, \ e = 65\% \] locally less
1. **inclusion regions** are regions where S.L. meets QSO’s area (due to absorption from Lyα forest entering the SDSS filters, which changes continuum power spectrum power law spectral index). All objects in these areas are selected so to sample the [2.2, 3.0] redshift range (where QSO density is also declining), but at the cost of a worse efficiency (Richards et al, 2001).

2. **exclusion regions** are those regions outside the main “stellar locus” clearly populated by stars only (usually WDs). All objects in these regions are discarded.

**Overall performance of the algorithm:** completeness $c = 95\%$, efficiency $e = 65\%$, but locally (in colours and redshift) much less.
Step 1: Unsupervised clustering

**PPS** determines a large number of distinct groups of objects: nearby clusters in the colours space are mapped onto the surface of a sphere.

NegE=750  
NegE=4

Step 2: Cluster agglomeration

**NEC** aggregates clusters from PPS to a (a-priori unknown) number of final clusters.

1. **Plateau analysis**: final number of clusters N(D) is calculated over a large interval of D, and critical value(s) D\( \text{th} \) are those for which a plateau is visible.

2. **Dendrogram analysis**: the stability threshold(s) D\( \text{th} \) can be determined observing the number of branches at different levels of the graph.
To determine the critical dissimilarity $D_{th}$ threshold we rely not only on a stability requirement.

A cluster is successful if fraction of confirmed QSO is higher than assumed fractionary value ($Th$)

$D_{th}$ is required to maximize $NSR$

$$NSR = \frac{\text{Number of successful clusters}}{\text{Number of total clusters}}$$

The process is recursive: feeding merged unsuccessful clusters in the clustering pipeline until no other successful clusters are found.

The overall efficiency of the process $e_{tot}$ is the sum of weighed efficiencies $e_i$ for each generation:

$$e_{tot} = \sum_{i=1}^{n} e_i$$
To assess the reliability of the algorithm, the same objects used for the “training” phase have been re-processed using photometric informations only. Results have been compared to the BoK.

\[
\begin{array}{|c|c|}
\hline
\text{algorithm} & \text{QSOs} & \text{not QSOs} \\
\hline
\text{QSOs} & 759 & 72 \\
\text{not QSOs} & 83 & 1327 \\
\hline
\end{array}
\]

\[e = 83.4 \% \quad c = 89.6 \%\]
Only a fraction (43%) of these objects have been selected as candidate QSO’s by SDSS targeting algorithm in first instance: the remaining sources have been included in the spectroscopic program because they have been selected in other spectroscopic programmes (mainly stars).
In this experiment the clustering has been performed on the same sample of the previous experiment, using only optical colours.
Experiment 2: local values of $\varepsilon$
Experiment 2: local values of $c$
### Table: Sample Parameters

<table>
<thead>
<tr>
<th>Sample</th>
<th>Parameters</th>
<th>Labels</th>
<th>$\epsilon_{\text{tot}}$</th>
<th>$\epsilon_{\text{tot}}$</th>
<th>$n_{\text{gen}}$</th>
<th>$n_{\text{succ,clus}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Optical QSO candidates (1)</strong></td>
<td>SDSS colours</td>
<td>‘specClass’</td>
<td>83.4 % ($\pm 0.3 %$)</td>
<td>89.6 % ($\pm 0.6 %$)</td>
<td>2</td>
<td>(3,0)</td>
</tr>
<tr>
<td><strong>Optical + NIR star-like objects (2)</strong></td>
<td>SDSS colours + UKIDSS colours</td>
<td>‘specClass’</td>
<td>91.3 % ($\pm 0.5 %$)</td>
<td>90.8 % ($\pm 0.5 %$)</td>
<td>3</td>
<td>(3,1,0)</td>
</tr>
<tr>
<td><strong>Optical + NIR star-like objects (3)</strong></td>
<td>SDSS colours</td>
<td>‘specClass’</td>
<td>92.6 % ($\pm 0.4 %$)</td>
<td>91.4 % ($\pm 0.6 %$)</td>
<td>3</td>
<td>(3,0,1)</td>
</tr>
</tbody>
</table>

The catalogue of candidate quasars is publicly available at the URL:

http://voneural.na.infn.it/catalogues_qsos.html

**BUT … LET’S GO BACK TO PHOT-Z**
1. Photometric redshifts of QSOs

No need for fine tuning !!!
Only New BoK !!!

\[ \sigma = 0.154 \]

\[ \sigma = 0.104 \]

\[ \sigma = 0.089 \]
Degeneracy induced by lines exiting photometric bands

Distribution of $Z_{\text{spec}}$ (solid) and $Z_{\text{phot}}$ (dashed) for test set !!!!
1. Photometric redshifts of QSOs

Laurino et al. 2009a, 2009b
Errors:

- **Input noise**: error propagation on the input parameter (Ball et al. 2008)
- **Model variance**: different models make differing predictions (Collister & Lahav 2004)
- **Model bias**: different models may be affected by different biases.
- **Target noise**: in some regions of the parameter space, data may represent poorly the relation between featured and targets (**Laurino 2009**).
1. Number of technical/algorithmic papers increases with new funding opportunities. Number of refereed papers remains constant.

2. Most of the work, so far, remains at the implementation stage (computer Science and algorithm development) and does not enter the “science production” stage...

3. Out of one thousand papers checked (galaxies, observational cosmology, survey) over the last two years: DM could be applied or involved in at least 30% of them leading to better results
Machine Learning based Data Mining is unavoidable when working on huge data sets.

The extraction of BoK’s offers challenges to good data repositories and data archives.

Accuracy of results depends on accuracy of BoK !!!!

Reliability and completeness of information
(no data is better than bad data)
Compliance with ontologies
Advanced queries in natural language