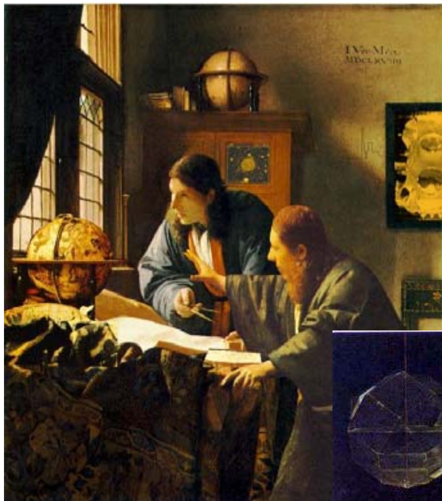
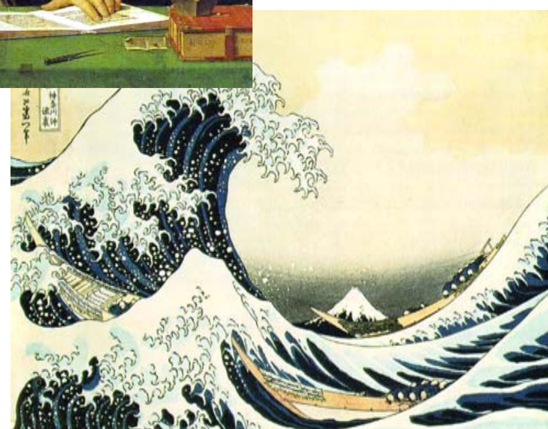
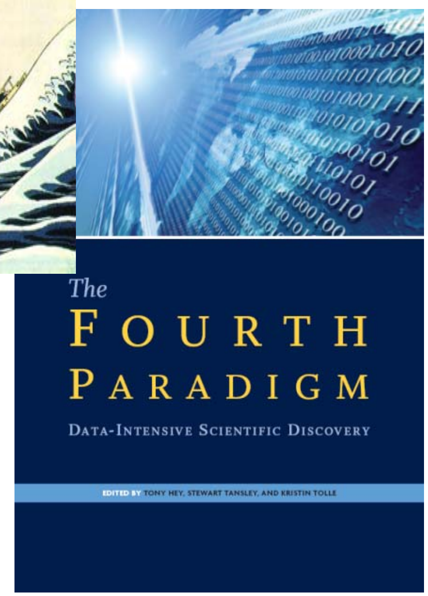


Computationally limited tasks in astronomy ?

We would all testify to the growing gap between the generation of data and our *understanding* of it ...

Ian H. Witten & E. Frank, Data Mining, 2001





Giuseppe Longo

University Federico II in Napoli
& California Institute of Technology

On behalf of the DAME team

The DPOSS/SDSS opened the way to a new methodology and defined what community expects from synoptic surveys

- **SDSS was the right data set at the right moment.**
  - Pioneeristic, yet, manageable with available technology (1 --10 TB of data products)
  - General in purpose, flexible enough to be useful for a large variety of existing problems, yet capable to rise new ones
- **Both data products (e.g. catalogues) and raw data were «immediately» made available to the community**
  - More than 3000 scientific papers came out of the Sloan (most of them from outside the core collaboration…
  - Some of these papers were **from third world countries and/or from small groups** working at small universities
  - **Large number of small technological/methodological** innovations (e.g. citizen science, large reliable KB's, etc.)
  - **Triggered the Interest of KDD community** in playing with a large, publicly available data set complex enough to be interesting from a ML point of view and not protected by any privacy/security issue

LHC like problems…

- **LHC**: among 10^{15} particle events find the only one of interest (Higgs boson)
- **GW**: find optimal algorithm(s) to detect a weak signal in an ocean of noise
- **NEMO**: among a huge number of events find those produced by high energy neutrinos
- **Etc…**

Synoptic sky surveys

In an ocean of complex data find those which are relevant for a huge variety of problems defined by a very large and heterogenous community

We want (*need ?*) to save the SDSS «democratic» approach to the data

BUT

- **Un-movable** data sets
- **Old data centers paradigm cannot be applied and …**
- Need for a large variety of «**user defined**» data products delivered by the data repositories to the final users

With LSST, Kepler, GAIA; Euclid, etc…
we have entered an era where:

- Most data ARE NOT seen by humans!
- Most knowledge hidden behind data complexity is potentially lost
- Most data (and data constructs) cannot be comprehended by humans directly!

Machine learning is no longer a viable option, it is a must…

- Data quality assessment
- ML aided data understanding
- Feature selection
- Data compression (delivery of specific products to the community and groups)
- Etc.

But ML is neither a simple nor an user friendly task

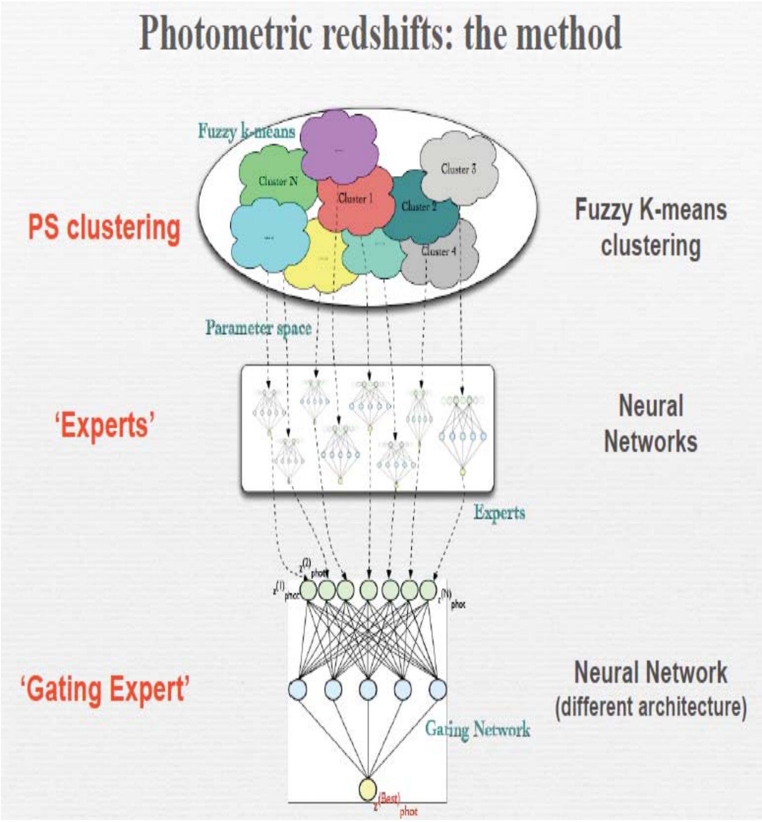
ML and KDD algorithms do not scale well with N and D

- **Querying:** spherical range-search $O(N)$, orthogonal range-search $O(N)$, spatial join $O(N^2)$, nearest-neighbor $O(N)$, all-nearest-neighbors $O(N^2)$
- **Density estimation:** mixture of Gaussians, kernel density estimation $O(N^2)$, kernel conditional density estimation $O(N^3)$
- **Regression:** linear regression, kernel regression $O(N^2)$, Gaussian process regression $O(N^3)$
- **Classification:** decision tree, nearest-neighbor classifier $O(N^2)$, nonparametric Bayes classifier $O(N^2)$, support vector machine $O(N^3)$
- **Dimension reduction:** principal component analysis, non-negative matrix factorization, kernel PCA $O(N^3)$, maximum variance unfolding $O(N^3)$
- **Outlier detection:** by density estimation or dimension reduction $O(N^3)$
- **Clustering**: by density estimation or dimension reduction, k-means, meanshift segmentation $O(N^2)$, hierarchical (FoF) clustering $O(N^3)$
- **Time series analysis:** Kalman filter, hidden Markov model, trajectory tracking $O(N^n)$
- **Feature selection and causality:** LASSO, L1 SVM, Gaussian graphical models, discrete graphical models
- **2-sample testing and testing and matching:** bipartite matching $O(N^3)$, n-point correlation $O(N^n)$….

Things are even worse if D is taken into account

Courtesy of A. Gray

Machine learning methods, in order to be effective need to be complex enough to capture the hidden knowledge



- Not methods, but workflows combining many methods
- Lenghty fine tuning is required
- Complex evaluation of results, with complex visualization issues, etc..

Laurino, D'Abrusco et al. 2011

Computing intensive tasks in astronomy?

…. For a Data Miner it is a piece of cake….

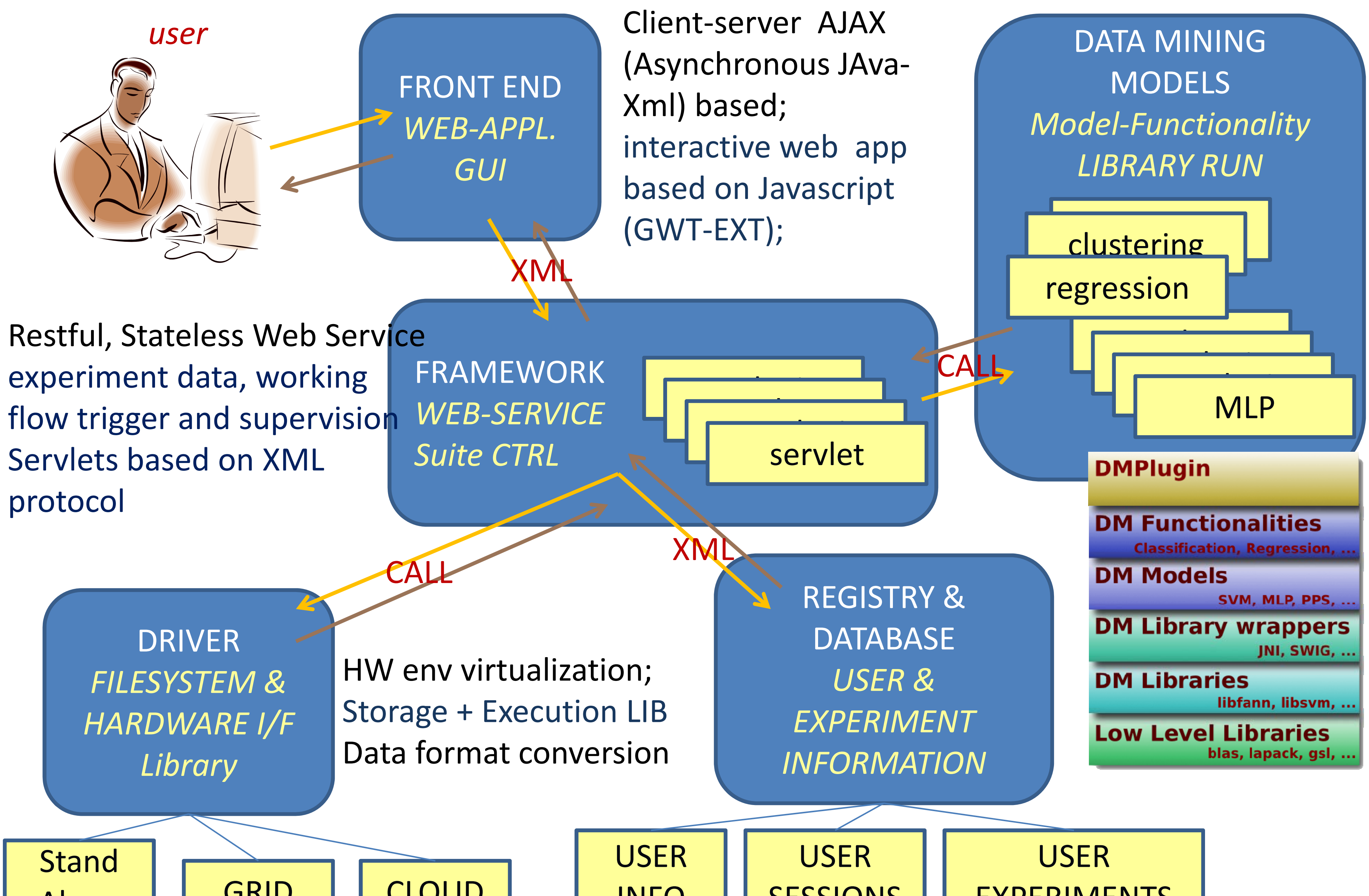
- **Every ML problem is potentially a data intensive one and can push to the limits any available HW and SW…**

- **We cannot move the data to the final users, but we need to move «user defined apps» where the data are (still a largely unexplored field in astronomy)**

- **Final users need to have «transparent» access to large computing facilities (better horses than chickens…)**

- **To implement effective ML methods we need to address a wide selection of «collateral problems» in parallelization of existing codes, visualization, benchmarking of algoriths, etc…**

The DAME architecture



brescia@na.astro.it

Topics I think should be addressed during the discussion (s):

- Standards for implementing «user defined» ML applications at the data repositories
- Visualization of complex data sets: what is available and what needs to be done.
- Template data sets for bench-marking of ML algorithms
- Identification of one or more «killer-like» problem (time domain) where to test the whole machinery