

- State of KDD by Longo
- Inclusion of R routines with examples and datasets in chapter 7 of KDD guide by Mahabal
  - [7: Algorithms and techniques astronomers could benefit from but seldom use](#)
- Bayesian learning with sparse dataset by Philip
- Distributed PCA with distributed tables by Sekhar

# Comments:

- Combining HPC efforts with current KDD efforts will be useful
- We also need to keep an eye on GPUs



- A VAO group led by A. Mahabal and C. Donalek (with contributions from many other people also outside of VAO) has completed an extensive benchmarking of the most commonly used DM packages and WAs.
- ORANGE, Rapid Miner, WEKA (AstroWEka) do not scale to very large data sets, even though some of them are «user friendly». DAME is the only truly scalable package (not so user friendly though)
- In order to deal with MDS, WAs are by far to be preferred with respect to client side solutions.
- There are however some problems which need to be urgently solved (they are more difficult to sampify, problems with visualization, etc.)



- M. Brescia, the VAO group and others, have identified the main requirements which a DM package (or WA) needs to match in order to be useful for the Vobs and have agreed on a series of priorities:
  1. Asynchronous access to data
  2. scriptable (STILTS like) in order to implement and preserve workflows.
  3. Built in pre-processing facilities
  4. Interfaceable with VO-Space
  5. SAMPified (problems in SAMPIfying WA's)



- In a data-centric environment, we need to minimize massive data flows on the network and move applications towards the data centers, especially if they are organized as KDD application warehouses.
- This of course requires a well-defined standardization process, in order to organize applications and SWFs in a fully interoperable way. This step is still “far from the coast” (literal translation from the Italian “*in alto mare* !”)

**The end**