Report: GREAT workshop & Summer School on Astrostatistics and Data Mining in large astronomical databases

Luis Sarro (SVO-UNED)
on behalf of the SOC:
Coryn Bailer-Jones
Laurent Eyer
William O'Mullane
Joris De Ridder
Workshop & School: a novel approach

- Rationale: we wanted to combine classical training with examples of current research in the field.

<table>
<thead>
<tr>
<th>Time</th>
<th>Activity</th>
</tr>
</thead>
<tbody>
<tr>
<td>9:00-11:00</td>
<td>School Lecture</td>
</tr>
<tr>
<td>11:30-13:30</td>
<td>School Lecture</td>
</tr>
<tr>
<td>15:00-15:45</td>
<td>Invited Keynote talk</td>
</tr>
<tr>
<td>15:45-17:00</td>
<td>Presentations</td>
</tr>
<tr>
<td>17:30-18:15</td>
<td>Invited Keynote talk</td>
</tr>
<tr>
<td>18:15-19:00</td>
<td>Presentations</td>
</tr>
</tbody>
</table>

Coding exercises
SUMMARY

- 34 school students from CZ, LT, SP(9), IT(5), VE(2), DE(4), UK(2), SL, ZA, PL, BE(2), AU, CH(2), NL, RU

- 56 workshop participants

- We offered 20 additional seats in the school for workshop participants, under restricted conditions (we had to find seats for virtually all workshop participants).
SOFTWARE REQUIREMENTS AND READINESS TESTS
DATASETS AND CODE

General laptop readiness

We do not need participants to have high-quality computers. What we need is for participants to come with appropriate developer software running on their computers, if possible, because the computers will be used for exercises.

If you have a Windows laptop, we would appreciate feedback on whether these instructions (below) can be easily followed, or how the functional test is executed.

If you have a Mac OS-X laptop, you need to have the `xcode developer tools` installed. These come on a CD with the laptop, or can be downloaded from the Apple Developer site if you are running the latest version of OS-X. You will know that they are installed if you can type `makai` at a terminal prompt and have it say something other than `command not found`. Linux boxes come with most things we need installed already, some even with Python already.

Requirements for the Data Mining lectures

The Data Mining lectures will make use of a browser and two software packages:

* Astroweka (see download and installation instructions [here](http://astroweka.sourceforge.net/)).
* R ([http://www.r-project.org/](http://www.r-project.org/)). Watch this page for additional packages required.

Install python and packages

* Download and install [Python](http://python.org). Please ensure that the python version installed, and used by `numpy/copy`/`matplotlib` is consistent. The included scripts have been tested on python version 2.6. On a Mac, Unix, or Linux box, make sure that python is in your path; that is, make sure that you can type "python" at the prompt in a terminal or shell and have it show you the python prompt.

* Download and install [NumPy](http://sourceforge.net/projects/numpy/files?group_id=1369&package_id=175193) and [Matplotlib](http://sourceforge.net/projects/matplotlib/files?group_id=121721). Detailed instructions can be found starting [here](http://www.scipy.org/Installing_ScPy/here).

* Download and install [Makai](http://sourceforge.net/projects/makai/files?group_id=380706) `matplotlib` See the [Matplotlib documentation](http://matplotlib.sourceforge.net/) for details. The [IDLE documentation](http://idler.readthedocs.org/) might also help.

Python documentation [here](http://python.org).
http://www.iwinac.uned.es/Astrostatistics/w/program.html

http://camd08.ast.cam.ac.uk/Greatwiki/GreatStats11
Lecturers and invited speakers

<table>
<thead>
<tr>
<th>Lecturer and Affiliation</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>David Hogg (New York University)</td>
<td><strong>Models: Specification, complexity and choice</strong></td>
</tr>
<tr>
<td>Suzanne Aigrain (Oxford University)</td>
<td><strong>Time series analysis</strong></td>
</tr>
<tr>
<td>Giuseppe Longo (Federico II University)</td>
<td><strong>Knowledge Discovery and Data Mining</strong></td>
</tr>
<tr>
<td>Matthew Graham (California Institute of Technology)</td>
<td><strong>Technical aspects of the analysis of petabyte-size databases</strong></td>
</tr>
<tr>
<td>Robert Lupton (Princeton University)</td>
<td><strong>Statistical Image Analysis</strong></td>
</tr>
</tbody>
</table>
Lecturers and invited speakers

- David Hogg:
  Exoplanet demography, quasar target selection, and probabilistic redshift estimation: Hierarchical models for density estimation, classification, and regression.

- Suzanne Aigrain:
  Learning to disentangle Exoplanet signals from correlated noise

- Giuseppe Longo:
  Astroinformatics and data mining: how to cope with the data tsunami

- Matthew Graham:
  The Art of Data Science

- Robert Lupton:
  Astronomical Surveys: from SDSS to LSST
Eilam Gross (Weizmann Institute):

Statistical methods in High Energy Physics and their implementation for Higgs Search and Dark Matter Search

Anthony Brown (Leiden University):

Science with Gaia: how will we deal with a complex billion-source catalogue and data archive?

Roberto Trotta (Imperial College London):

Recent Advances in cosmological Bayesian model comparison
Models: Specification, complexity, and choice (David Hogg)

1) Model specification and likelihood formulation
2) Model complexity and choice
3) (pair-coding) Model selection workshop
4) (pair-coding) Model selection workshop
Table of contents

- Lecture 0 (to be provided in advance as links or bibliography if needed)
  - To first order, time-series are like any other 1-D dataset, and the same principles apply when trying to model them. Therefore, reading the document provided by D. Hogg [http://arxiv.org/abs/1008.4686](http://arxiv.org/abs/1008.4686) will also be helpful in preparing for these lectures.
  - Three example datasets have been provided (see below), which the students are encouraged to download and try to read in and plot in advance of the school. These datasets will be used as part of hands-on workshop sessions by several of the lecturers.

- Topics to be covered in the lectures:
  - What’s special about time-series?
  - Bayesian spectral analysis
  - Some notes about using the discrete Fourier transform
  - Autocorrelation functions
  - ARMA models
  - Gaussian processes
  - Systematics in ensembles of time-series

- Topics to be covered if time permits
  - State-space models and quasi-periodic systems
  - Empirical mode decomposition and the Hilbert-Huang transform
  - Change-point detection

- Exercises
  - See the lectures...
1. The Sampling Theorem and Image Resampling
2. Object Detection and Measurement as Statistical Estimation
3. Hands-on session: object detection and measurement
4. Hands-on session continued: object detection and measurement

Lanczos Kernels

A popular modification of the sinc kernel is a \textit{Lanczos}(n) kernel,

\[
L_n(x) = \begin{cases} 
\text{sinc}(x) \times \text{sinc}(x/n) & |x| \leq n \\
0 & \text{otherwise}
\end{cases}
\]
Lecture 1: what is data mining
Lecture 2: feature selection and dimensionality reduction
Lecture 3: classification tasks and supervised methods
Lecture 4: clustering methods
Technical Aspects of the analysis of petabyte size databases (M. Graham)

Technical aspects of the analysis of petabyte-size databases

It would take over 33 years to watch a 1 PB MP3 movie yet, within the decade, data sets of this size will be as everyday a feature of astronomical life as astro-ph or APOD. This section will cover the practical aspects of handling petascale (and larger) data sets and streams including new computational approaches needed to work with them from an astronomer’s perspective.

Table of contents

- Lecture 0 (to be provided in advance as links or bibliography if needed)
  - How big is a petabyte?
  - Big data sets en route: astronomy, other sciences
- Lecture 1: How to store a petabyte
  - What do you store?
  - Cost and performance of storage
  - Databases: relational vs non-relational, indexing
- Lecture 2: How to work with a petabyte
  - Distribution
  - Divide and conquer: MapReduce, Hadoop (how to sort 1 PB)
  - Putting things together: FIG
- Lecture 3: How to analyze a petabyte
  - Random access
  - Characterizing data
  - Streaming statistics
- Ideas for pair-coding examples (to be discussed with SOC / other lecturers)
  - Coding up a simple analysis routine using Hadoop
The (anonymous) questionnaire

Which format do you prefer

- At least one - 47.4%
- At least one + Whisperer - 32.9%

How was the pace of the event (-1= too slow, 0= just right, 1= too fast)?

- 0 - 31.4%
- -1 - 22.9%
- 1 - 18.0%

How was the length of the event (-1= too long, 0= just right, 1= too short)?

- 1 - 16.7%
- 0 - 15.9%
- -1 - 13.0%

Overall Presenters

- 3 - 8.9%
- 2 - 5.2%
- 1 - 4.8%
- 0 - 21.7%
- -1 - 4.1%
- -2 - 17.4%
- -3 - 17.4%

Overall Event

- 3 - 4.9%
- 2 - 1.0%
- 1 - 4.1%
- 0 - 12.4%
- -1 - 39.7%
- -2 - 56.8%
Wrong

- More/longer hands-on sessions
- Allow for choice of programming language/problems with python
- Hands-on sessions more guided.

Right

- Video recording of lectures and keynote talks
- Presentations, manuscripts, python code, datasets available online
- Hands-on sessions
- Thought provoking keynote talks
Funding and grants

- Fee reimbursed to all school participants that requested financial support
- Lunch at the venue for all participants
- Travel grants (1500 Euros)