Summary

The School on Astrostatistics and Data Mining (hereafter SADM) took place from May 29th to June 3rd in the Hotel Taburiente in the island of La Palma (Spain). Initially, a registration fee of 150 Euros was requested, which was reimbursed in the form of travel grants to all those students that requested financial support in their application forms. The registration fee was set to ensure that all accepted applicants actually attended the school and that there would be no vacant seats.

The school comprised 5 main blocks, each with a total of 4 lectures of one hour each. The areas and lecturers are listed below:

1. David Hogg (New York University, USA) *Models: Specification, complexity and choice*
2. Suzanne Aigrain (Oxford University, UK) *Time series analysis*
3. Giuseppe Longo (Federico II University, Italy) *Knowledge Discovery and Data Mining*
4. Matthew Graham (California Institute of Technology, USA) *Technical aspects of the analysis of petabyte-size databases*
5. Robert Lupton (Princeton University, USA) *Statistical Image Analysis*

The program was such that school lectures took place in the morning (from 9:00 to 11:00 and from 11:30 to 13:30) while the keynote talks took place in the afternoon in the context of the workshop. There were 8 keynote talks (two each afternoon from Monday to Thursday):

1. David Hogg: Exoplanet demography, quasar target selection, and probabilistic redshift estimation: Hierarchical models for density estimation, classification, and regression.
2. Suzanne Aigrain: Learning to disentangle Exoplanet signals from correlated noise
3. Giuseppe Longo: Astroinformatics and data mining: how to cope with the data tsunami
4. Matthew Graham: The Art of Data Science
5. Robert Lupton: Astronomical Surveys: from SDSS to LSST
6. Eilam Gross (Weizmann Institute, Israel): Statistical methods in High Energy Physics and their implementation for Higgs Search and Dark Matter Search
7. Anthony Brown (Leiden University, Netherland): Science with Gaia: how will we deal with a complex billion-source catalogue and data archive?
8. Roberto Trotta (Imperial College London, UK): Recent Advances in cosmological Bayesian model comparison

Students were adviced to occupy the rest of the afternoon working on the exercises provided by the lecturers (see below) and selecting the talks that might be of direct interest for ther thesis projects.

Further details on the school program and materials can be found in the next sections of the report.

Event description

On November 17th 2010 the first announcement of the SADM was released. The SOC received 50 applications before the deadline set on March 1st. On March 11th 2011 the acceptance letters were sent to 38 applicants. 34 of them finally registered for the school. The selection of the applicants was carried out based on the relevance of the school topics for the applicants research projects.

The application form can be downloaded from

http://www.iwinac.uned.es/Astrostatistics/ss/reg-form.pdf

An registration fee of 150 Euros was set in order to ensure the attendance of accepted/registered applicants. This fee was reimbursed in the form of a travel grant to all those students that requested financial support. In addition to this, extended travel grants were awarded to 5 applicants (Iván Cabrera, Michelle Knights, Cecilia Mateu, Dmitry Svinkin and Jaroslav Vazny). The final list of travel grants is enclosed below (in Euros).

Simona Bekeraite	150
Nadejda Blagorodnova	150
Donata Bonino	150
Iván Cabrera	150+400
Elena Carolo	150
Hywel John Farnhill	150
Ginevra Favole	150
Apoorva Jayaraman	150
Michelle Knights	150+250
Cecilia Mateu	150+400
Pieter Neyskens	150
David salvetti	150
Gaetano Scandariato	150
Dmitry Svinkin	150 + 200
Pau Vallés	150
Jaroslav Vazny	150 + 250
Benedetta Vulcani	150
Zenghua Zhang	150

Furthermore, the School organisation funded the meals of all participants in the hotel where the school and workshop took place.

School material

In the following, we will often refer to school materials made available to the students through the School web pages
(http://www.iwinac.uned.es/Astrostatistics/ss/school.html)
and through the GREAT network wiki
(http://camd08.ast.cam.ac.uk/Greatwiki/GreatStats11)

Software & readiness tests.

In the application form, the applicants committed to having a laptop ready for coding with the software recommended by the school organisers. The SOC together with the lecturers decided to use python as the programming language of the school. This was announced to the students on March 15th thus leaving 2.5 months for getting aquainted with the python language.

Furthermore, a software readiness test was set up in the School wiki pages hosted in the GREAT network wiki with python programs kindly provided by one of the lecturers (D. Hogg) at

http://camd08.ast.cam.ac.uk/Greatwiki/GreatStats11/SRT

The software readiness test web also included a section with Frequently Asked Questions and installation instructions.

The Data Mining lecturer (G. Longo) recommended the use of R and astroweka.

Materials and wiki

All the information relative to the SADM was provided through the school web pages, including links to the GREAT network wiki pages with specific information on individual lectures.

This includes information on

i) the application and registration processes, deadlines, forms
ii) logistics (Venue, hotels, booking, getting there...)
iii) scope and committes
iv) lecturers and the program
v) software requirements
vi) contacting the SOC/LOC

The GREAT wiki pages (linked from the SADM web pages) were used to provide lecturers with an interactive space that could be taylored to their respective areas. The goal was to have a detailed program with links to relevant bibliography, python code and datasets. The profiting of this tool was inhomogeneous across areas. We recommend a visit of the wiki pages of the Time Series Analysis Lectures

http://camd08.ast.cam.ac.uk/Greatwiki/GreatStats11/TSA

for an example of a good use of the tool. Unfortunately, not all of the lecturers made an

intensive use of the wiki, and some delivered the materials short before the school. Further below in this report we suggest ways to avoid these problems.

All presentations used in the lectures (and those corresponding to the workshop keynote and contributed talks) were made available immediately as links in the school program web pages. Also, the manuscripts submitted for the workshop proceedings were made available in the same way. Furthermore, all of the lectures and keynote talks were recorded in video and these are now being edited for web publication.

The SOC made a strong emphasis in the reuse of datasets and cases across lectures, when and if possible. As a result, some of the lecturers used the same datasets for exemplifying various techniques (e.g. the HARPS radial velocity data of HD104067 used by D. Hogg and S. Aigrain or the k-means algorithm by M. Graham and G- Longo).

Proceedings.

There is an agreement with the Springer Publishing Company (New York) for the edition of the workshop proceedings in the newly created series on Astrostatistics. The deadline for manuscript submission was set to May 22nd (before the school/workshop) and was subsequently extended until June 15th. The proceedings will be edited as an e-book free of charge for school and workshop participants.

Questionnaire

One week after the SADM closure, the school organisers set up a web-interface to an anonymous questionnaire aimed at evaluating several aspects of the entire event. The link to the questionnaire

http://gaia.esac.esa.int/qi/questionaire.gen_form?p_workshop=222

was made public to the school students on June 9th, and the answers were compiled in an online report on June 19th. The full report (including the comments) is available at

http://gaia.esac.esa.int/apex/f?p=101:1:6546257185672768:::::

It is impossible to summarise the entire report here, but we would like to remark several points. We will concentrate on the ways to improve future schools, but we would like to remark that the overall impression is that the school caused an excellent impression in the students.

First of all, the school students are divided with respect to the evaluation of the novelty introduced here with the school+workshop format. Half of the student prefer this scheme while the other half prefers the classical school only design. Second, the average evaluation of the lecturers is very variable, and we even find very different evaluations for a given lecturer between the school lectures and the keynote talk.

In general, the students express a very positive attitude towards the practical component of the lectures (hands-on sessions, pair-coding, practical exercises) but some of them also express difficulties with the programming language (python). It is our impression, that even though the announcement was made 2.5 months in advance of the school, few of the students had got acquainted with the basics of the language by the end of May. It is a lesson to be learnt from this school, that more emphasis/time/care

has to be placed in the preparation of the practical exercises both on the lecturer and on the students sides. One recurrent topic in the comments was the necessity for *longer and more guided hands-on sessions*.

The pace of the event was considered by 45% of the students too fast, while 50% of them considered it to be just right.

Programe of the event

The final program of the event can be consulted at

http://www.iwinac.uned.es/Astrostatistics/w/program.html

which includes also the workshop keynote and contributed talks. Here we include only the school lectures:

- **Models: Specification, complexity and choice (David Hogg)**
  What is a model? What freedoms does a model have and how can we capture that? Are qualitatively different models comparable? What is the difference between a likelihood and a probability for a model or for model parameters? How do we decide among models that are qualitatively similar but quantitatively different? How do we decide among models that are qualitatively different? The most important content will be conveyed through a lab session in which participants pair-code solutions to some model selection problems.
  Table of contents
    - Lecture 0: (to be provided in advance as links or bibliography if needed)
    - Lecture 1: Model specification and likelihood formulation
    - Lecture 2: Model complexity and choice
    - Lecture 3: (pair-coding) Model selection workshop
    - Lecture 4: (pair-coding) workshop continued
- **Knowledge Discovery and Data Mining (Giuseppe Longo)**
  Feature selection: filter approach, wrapper approach, PCA, Diffusion Maps. Supervised classification: the curse of dimensionality, bias-variance trade-off, the kernel trick, support vector machines, cross-validation, evaluation of classifiers. Unsupervised classification taxonomy, evaluation measures.
  Table of contents:
    - Lecture 0: (to be provided in advance as links or bibliography if needed)
    - Lecture 1: what is data mining
    - Lecture 2: feature selection and dimensionality reduction
    - Lecture 3: classification tasks and supervised methods
    - Lecture 4: clustering methods
- **Statistical Image Analysis (Robert Lupton)**
  The source detection problem, source modelling, catalogue cross correlations, combination of images...
  Table of contents
    - Lecture 0 (to be provided in advance as links or bibliography if needed)
    - Lecture 1 The Sampling Theorem and Image Resampling
    - Lecture 2 Object Detection and Measurement as Statistical Estimation
    - Lecture 3 Workshop: object detection and measurement
    - Lecture 4 (workshop continued, if needed)

- **Technical aspects of the analysis of petabyte-size databases (Matthew Graham)**
  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: PIG
    - 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
- T**ime series analysis (Suzanne Aigrain)**
  This section will cover common tool for exploring and characterising time-series and ensembles thereof. The first two lectures are devoted to time- and frequency domain techniques respectively, and cover some frequently used exploratory . Particular attention will be devoted to the treatment of stochastic processes and mixtures of stochastic and periodic processes.
  Table of contents
    - Lecture 0 (to be provided in advance as links or bibliography if needed)
      - stationarity, autocorrelation function, (discrete) Fourier transform, window function
      - properties of the Gaussian distribution
    - Lecture 1: Time-domain analysis
      - autocorrelation techniques
      - common time-domain filters
      - stochastic processes: ARIMA models, Gaussian processes
    - Lecture 2: Frequency analysis
      - noise properties in the frequency domain
      - periodic signal detection
      - time-frequency analysis, wavelet transforms
    - Lecture 3: Ensembles of time series
      - principal component analysis in the time and frequency domains
      - classification and clustering