

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

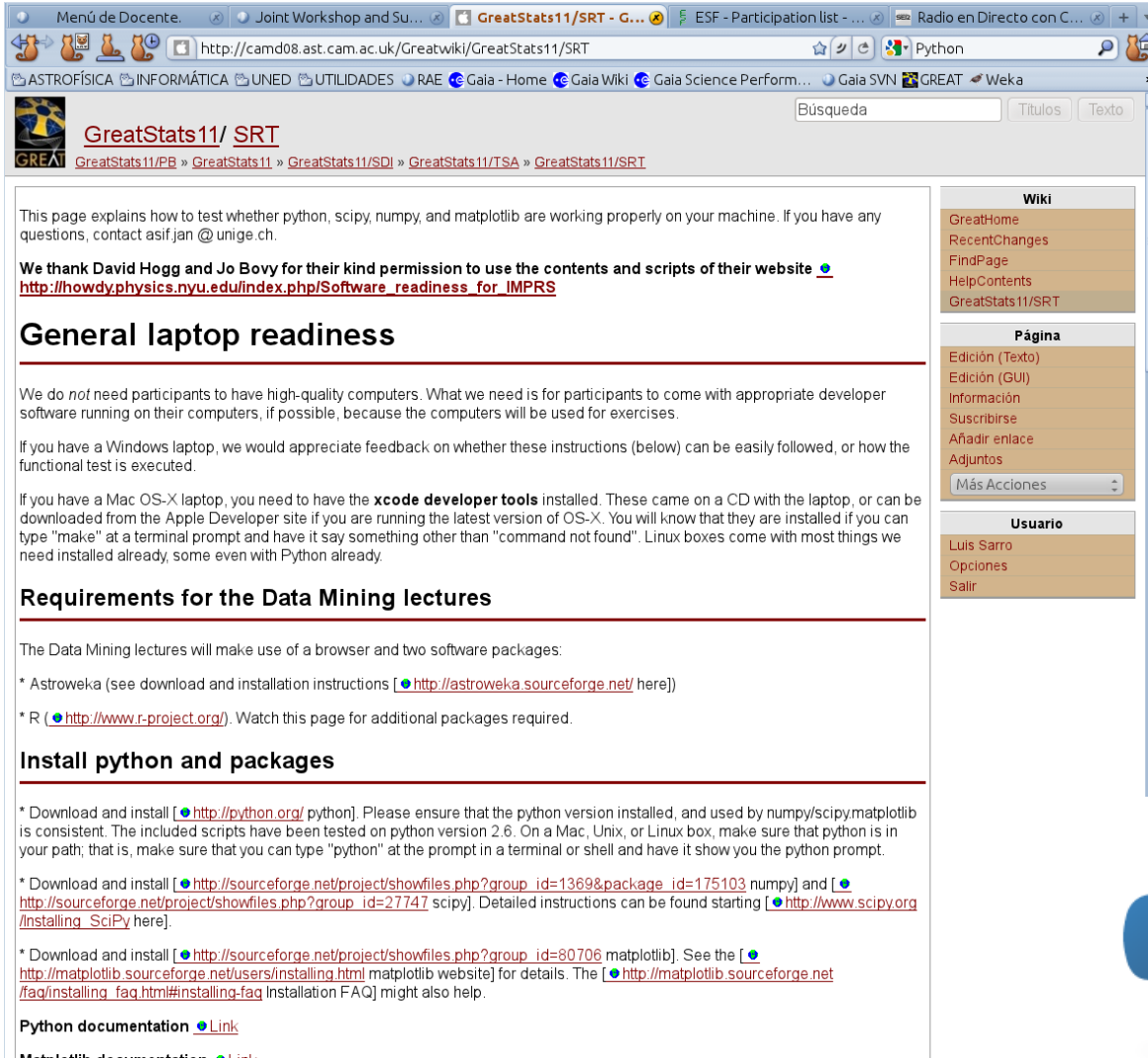
Workshop & School: a novel approach

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

| | | |
|-------------|----------------------|------------------|
| 9:00-11:00 | School Lecture | |
| 11:30-13:30 | School Lecture | |
| 15:00-15:45 | Invited Keynote talk | |
| 15:45-17:00 | Presentations | Coding exercises |
| 17:30-18:15 | Invited Keynote talk | |
| 18:15-19:00 | Presentations | Coding exercises |

SOFTWARE REQUIREMENTS AND READINESS TESTS

DATASETS AND CODE



Menú de Docente. Joint Workshop and Su... GreatStats11/SRT - G... ESF - Participation list - ... Radio en Directo con C...

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

ASTROFÍSICA INFORMÁTICA UNED UTILIDADES RAE Gaia - Home Gaia Wiki Gaia Science Perform... Gaia SVN GREAT Weka

Búsqueda Titulos Texto

GreatStats11/ SRT

GreatStats11/PB » GreatStats11 » GreatStats11/SDI » GreatStats11/TSA » GreatStats11/SRT

This page explains how to test whether python, scipy, numpy, and matplotlib are working properly on your machine. If you have any questions, contact asif.jan@unige.ch.

We thank David Hogg and Jo Bovy for their kind permission to use the contents and scripts of their website http://howdy.physics.nyu.edu/index.php/Software_readiness_for_IMPRS

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 came 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 "make" 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 [<http://astroweka.sourceforge.net/> here])
- * R (<http://www.r-project.org/>). Watch this page for additional packages required.

Install python and packages

* Download and install [<http://python.org/> python]. Please ensure that the python version installed, and used by numpy/scipy/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 [http://sourceforge.net/project/showfiles.php?group_id=1369&package_id=175103 numpy] and [http://sourceforge.net/project/showfiles.php?group_id=27747 scipy]. Detailed instructions can be found starting [http://www.scipy.org/installing_SciPy here].

* Download and install [http://sourceforge.net/project/showfiles.php?group_id=80706 matplotlib]. See the [<http://matplotlib.sourceforge.net/users/installing.html> matplotlib website] for details. The [http://matplotlib.sourceforge.net/faq/installing_faq.html#installing-faq Installation FAQ] might also help.

Python documentation [Link](#)

Matplotlib documentation [Link](#)

Wiki

- GreatHome
- RecentChanges
- FindPage
- HelpContents
- GreatStats11/SRT

Página

- Edición (Texto)
- Edición (GUI)
- Información
- Suscribirse
- Añadir enlace
- Adjuntos
- Más Acciones

Usuario

- Luis Sarro
- Opciones
- Salir



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

<http://camd08.ast.cam.ac.uk/Greatwiki/GreatStats11>

Lecturers and invited speakers

David Hogg (New York University)

Models: Specification, complexity and choice

▪ Suzanne Aigrain (Oxford University)

Time series analysis

▪ Giuseppe Longo (Federico II University)

Knowledge Discovery and Data Mining

▪ Matthew Graham (California Institute of Technology)

Technical aspects of the analysis of petabyte-size databases

▪ Robert Lupton (Princeton University)

Statistical Image Analysis

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

Lecturers and invited speakers

- Eilam Gross (Weizmann Institute):

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

- Anthony Brown (Leiden University):

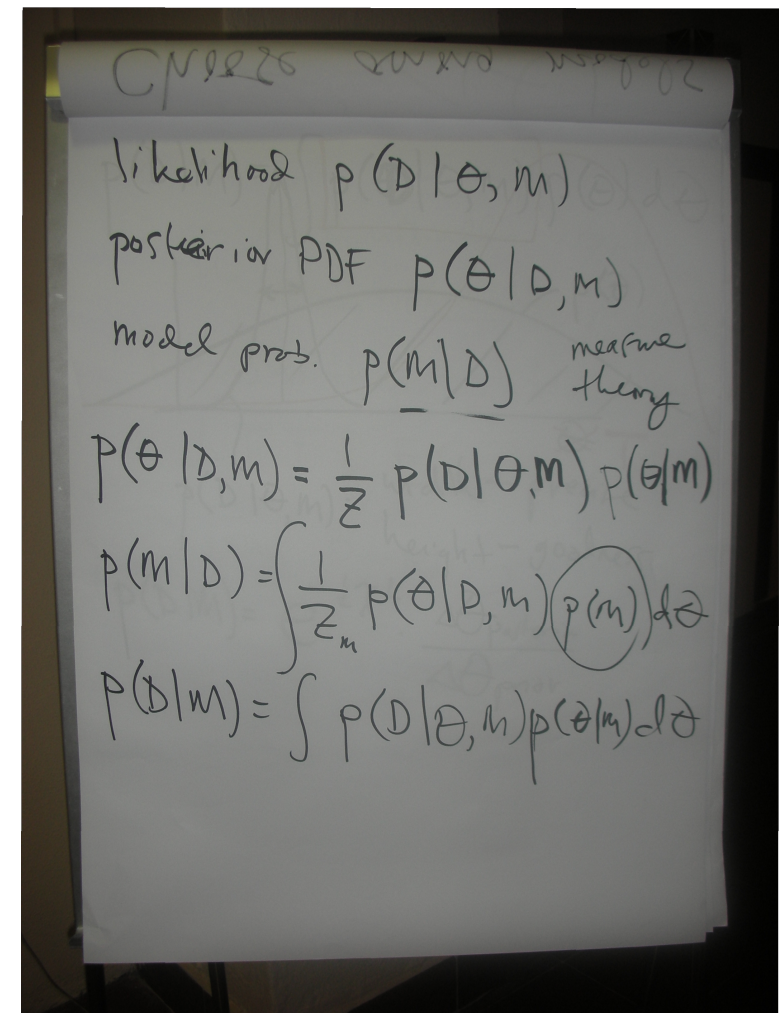
Science with Gaia: how will we deal with a complex billion-source catalogue and data archive?

- Roberto Trotta (Imperial College London):

Recent Advances in cosmological Bayesian model comparison

Models: Specification, complexity, and choice (David Hogg)

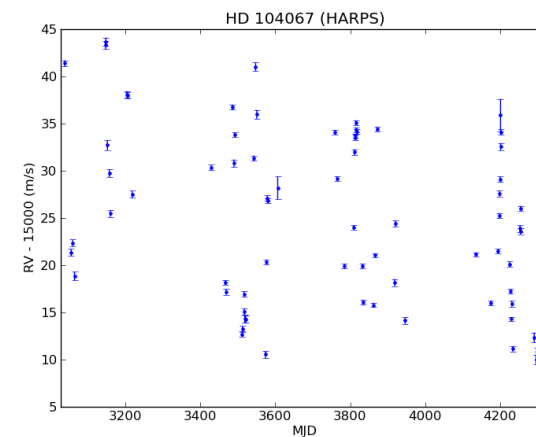
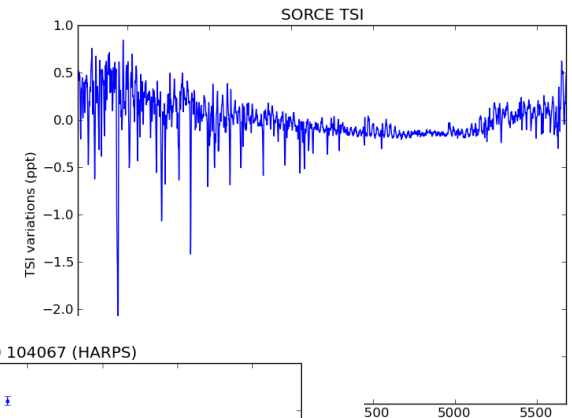
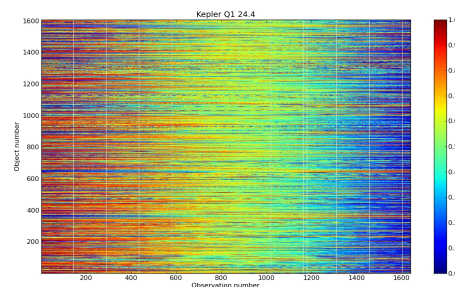
- 1) Model specification and likelihood formulation
- 2) Model complexity and choice
- 3) (pair-coding) Model selection workshop
- 4) (pair-coding) Model selection workshop



Time Series Analysis (Suzanne Aigrain)

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> 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...



Statistical Image Analysis (Robert Lupton)

- 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

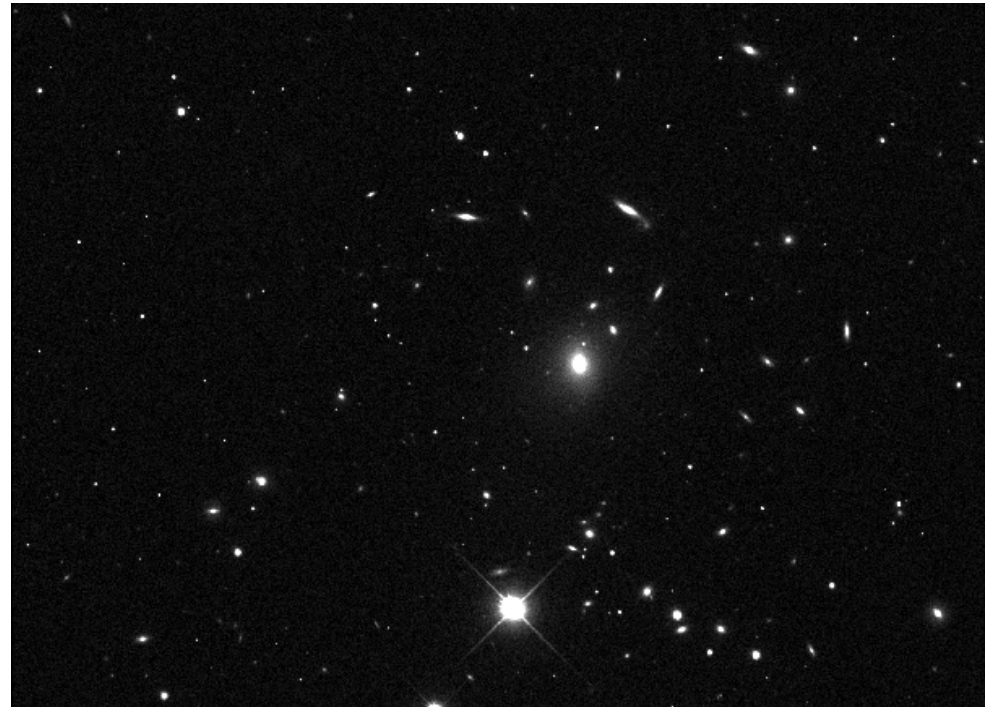
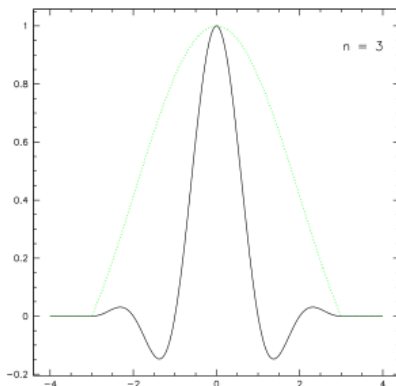
The Sampling Theorem

Sinc Resampling

Lanczos Kernels

A popular modification of the sinc kernel is a *Lanczos*(n) kernel,

$$L_n(x) = \begin{cases} \text{sinc}(x) \times \text{sinc}(x/n) & |x| \leq n \\ 0 & \text{otherwise} \end{cases}$$

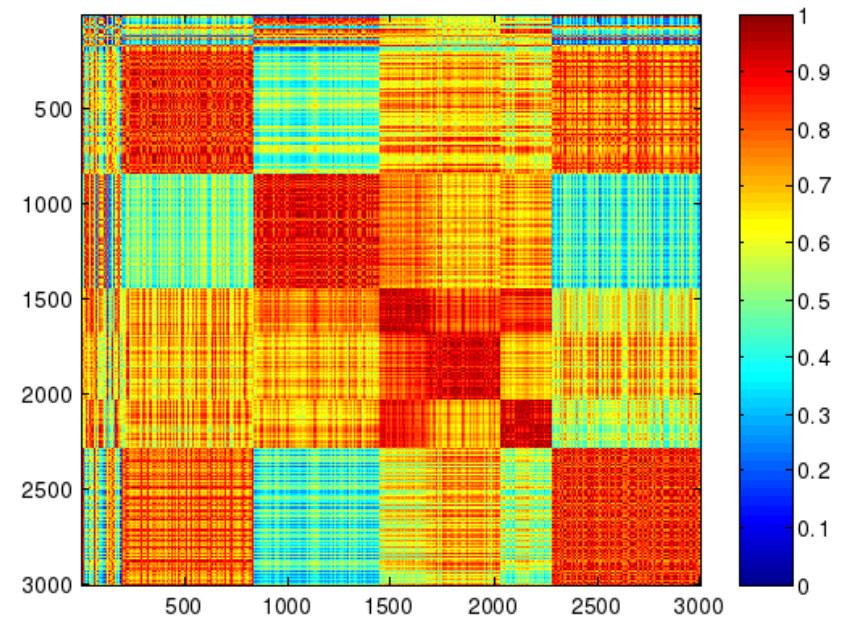
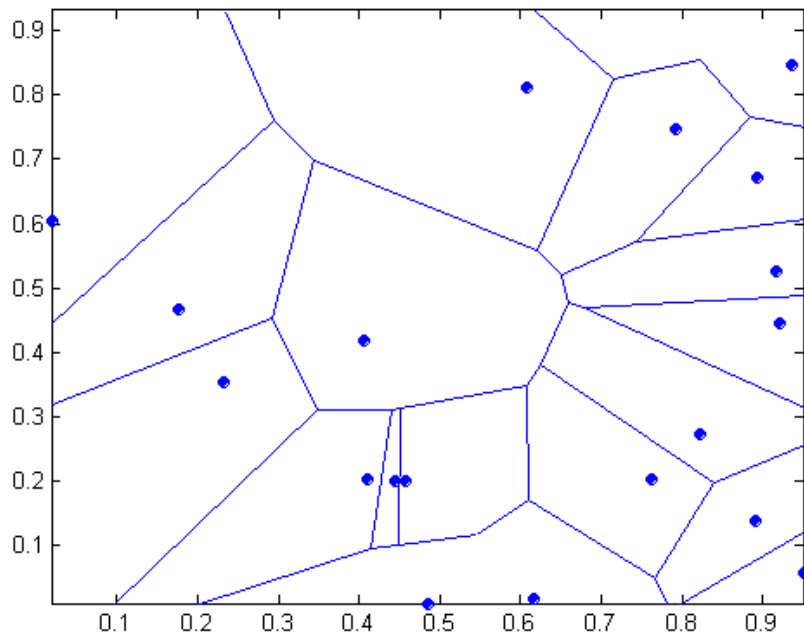


Lecture 1: what is data mining

Lecture 2: feature selection and dimensionality reduction

Lecture 3: classification tasks and supervised methods

Lecture 4: clustering methods



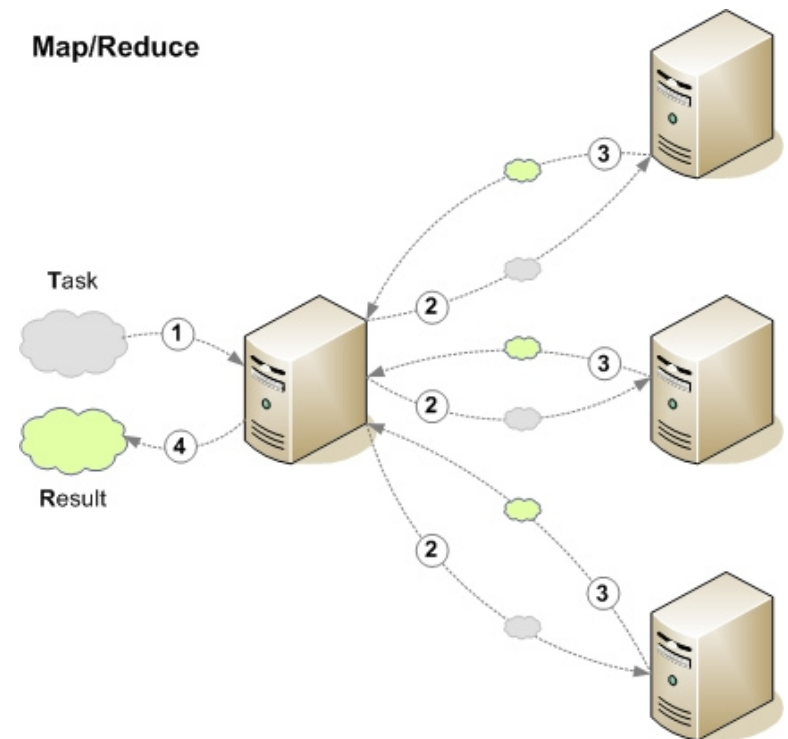
Technical Aspects of the analysis of petabyte size databases (M. Graham)

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.

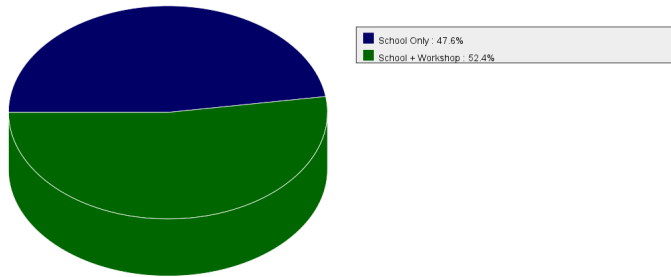
Table of contents

- 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: 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

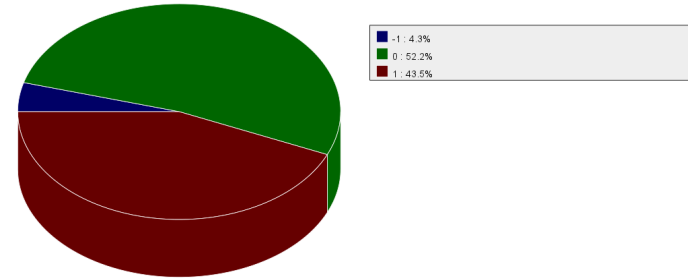


The (anonymous) questionnaire

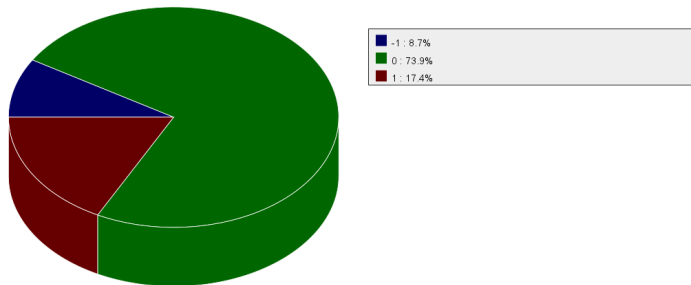
Which format do you prefer



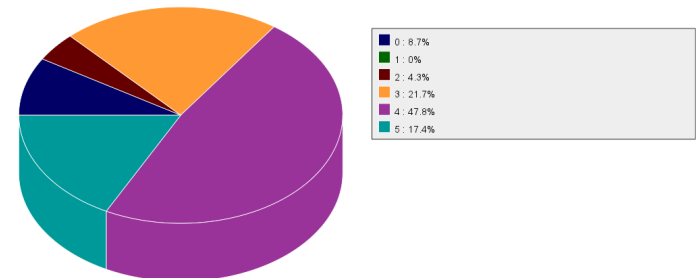
How was the pace of the event (-1= too slow, 0= just right, 1= too fast)?



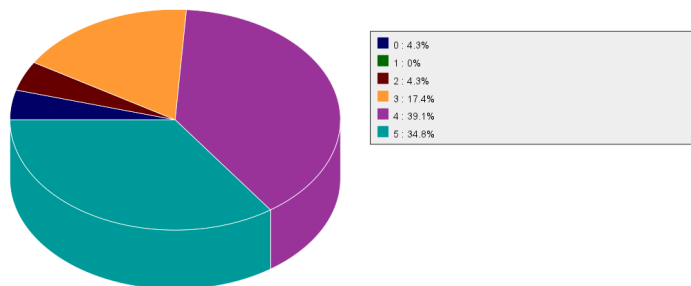
How was the length of the event (-1= too long, 0= just right, 1= too short)?



Overall Presenters



Overall Event



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)

