The Big Bang of astronomical data. How to use Python to survive the data flood.

Jose Sabater Montes
Institute for Astronomy, University of Edinburgh

P. Best, W. Williams, R. van Weeren, S. Sanchez, J. Garrido, J. E. Ruiz, L. Verdes-Montenegro and the LOFAR surveys team
The new astronomy
The new astronomy

ALMA correlator
Astronomy and Python

- Python is currently the main language used for astronomy
- General Python computing libraries: numpy, scipy, matplotlib, pandas, emcee...
- Specific astronomical libraries (see http://www.astropython.org/packages/)
  - astroML: machine learning and data mining
  - astropy: main general library for astronomy
  - etc.
The future of astronomy

- New state of the art astronomical infrastructures that produce an overwhelming amount of data

- Examples:
  - ESA Gaia
  - Large Synoptic Survey Telescope
  - ESA Euclid
  - The Square Kilometre Array and its pathfinders (LOFAR, ASKAP, Meerkat...)
  - Etc.
Large Synoptic Survey Telescope

- 8.4 m mirror
- Covers the full visible sky every two nights
- Under construction - operational in 2022
Large Synoptic Survey Telescope

- Camera 189x16 Mpix
- Pipeline preprocessing: 3GB/s
- 30 TB per night during 10 years
- 2 M events triggered per night
Square Kilometre Array (SKA)

- Radio telescope with 1 km² of collecting area
- Phase 1 - 2020

2016-10-09  PyConES 2016 Almería
SKA data

- Phase 1:
  - 10 TB/s from the antennas to the correlator
  - 40 GB/s of data → 70 PB per year
  - 1 MW infrastructure and 10 MW processing
SKA data

- Phase 2:
  - 160 TB/s from the antennas to the correlator
  - > 100 GB/s of data → 4.6 EB per year
  - 200 to 2000 dishes
  - 130K to 1M antennas
LOFAR

- Low Frequency Array
- Software defined radio-interferometer working at low frequencies (30 to 240 MHz)
- One of the Square Kilometre Array pathfinders
LOFAR Stations
LOFAR frequencies

- LBA 30-80 MHz
- HBA 120-240 MHz
LOFAR science

- Origin and evolution of galaxies and supermassive black holes
- Epoch of reionization
- Solar science and space weather
- Transients
- Map the galaxy using pulsars
- Exoplanets, SETI
Radio galaxies

Hercules A. Credits: NASA and the NRAO
LOFAR aperture synthesis

- field of view diameter of \(~5\) deg at 150 MHz
- resolution \(< 5\) arcsec (up to 0.1 arcsec)
LOFAR imaging

In 8 hours
~40 sq. deg.
5000 sources

Calibration on IAA (Granada) cluster
LOFAR imaging

In 8 hours
~40 sq. deg.
5000 sources

Calibration on
IAA (Granada) cluster
Extended sources
Ionosphere

- Effect depends on frequency, length of the baselines and f.o.v.
- LOFAR, worst case:
  - Wide field of view
  - Long distance baselines
  - Low frequency

H. Intema
Ionosphere

- Effect depends on frequency, length of the baselines and f.o.v.
- LOFAR, worst case:
  - Wide field of view
  - Long distance baselines
  - Low frequency

H. Intema
Ionosphere
Challenges for the astronomer

- User data calibration (remove the effect of the ionosphere and the RFI)
  - 8 hours full resolution $\rightarrow$ ~20 TB
  - Minimum of 2 CPU years to run the calibration
  - Experimental pipeline

- LOFAR calibration software
  - Difficult to install
  - Continuous development
Computational solution needed

• Parallelizable:
  – Deal with a large amount of data in a reasonable time.

• Flexible:
  – Adapt the infrastructure ("hardware") to different calibration strategies
  – Deal with quickly changing temperamental software
  – On-demand (optional but very useful)
HPC, HTC and cloud computing

- Tests in different infrastructures: clusters, GRID, cloud, etcetera.
- SKA-AWS astrocompute proposal
  - Preparation of the base infrastructure (virtual machine images, check provisioning of spot instances, etc.)
  - Data transfer: 50 TB
  - Adapt calibration pipeline and run http://www.lofarcloud.uk
Experimental calibration pipeline

Calibrator data
360 chunks (1 sb)

Main target data
360 chunks (1 sb)

Preprocessed target data
36 chunks (10 sb)

Calibration solutions

Combined data: 9 chunks (40 sb)

Facet calibration

Self-cal and subtraction

~30 iterations

Data split:
- field
- observation
- frequency

Pre-processing

Processing

Final image
The role of Python

LOFAR software

- libraries
- prog. 1 ...
- prog. n
- script 1 ...
- script n
The role of Python

LOFAR software

- libraries
- prog. 1 ...
- script 1 ...
- prog. n
- script n

Experimental pipelines

Pipeline 1

- step 1 → step 2
- step 3 → step 4
- step 5 → step 6

Pipeline n ...

2016-10-09 PyConES 2016 Almería
The role of Python

LOFAR software

- libraries
- prog. 1
- ...
- prog. n
- script 1
- ...
- script n

Infrastructure

- pipeline chunk 1
- pipeline chunk 2
- ... 
- pipeline chunk 3
- pipeline chunk 4
- pipeline chunk 5
- ... 
- pipeline chunk n

Experimental pipelines

- Pipeline 1
  - step 1
  - step 2
  - step 3
  - step 4
  - step 5
  - step 6

- Pipeline n
  - ...

control
The role of Python

LOFAR software

- libraries
- prog. 1
- script 1
- ... (repeated for n)

Experimental pipelines

- Pipeline 1
  - step 1 -> step 2
  - step 3 -> step 4
  - step 5 -> step 6

- Pipeline n
  - ... (repeated for n)

Infrastructure

- pipeline chunk 1
- pipeline chunk 2
- ... (repeated for n)

Ansible

- control

- Python
- Python wrapper
- Python mixed
- Other

Cython

2016-10-09 PyConES 2016 Almería
Summary

- Big software and data managing challenges associated to new astronomical infrastructures, even for final users.

- The role of Python:
  - Quick prototyping - fundamental for experimental pipelines and testing.
  - Multi-domain - Can be used for a wide range of problems.
  - Robust - Enough to write “real” efficient software.
  - Unifying tool - that holds all together.