Long term and archivable reproducibility, a summary

Infante-Sainz, R., Akhlaghi, M., Baena-Gallé, R.
Instituto de Astrofísica de Canarias, Universidad de La Laguna

https://maneage.org

Abstract: Scientific data analysis pipelines commonly use high-level technologies that were popular when they were created, only providing an immediate solution which is unlikely to be sustainable or reproducible in the future. We have implemented "Maneage" (***Ma**naging data Lin**eage***), a solution which stores the project in machine-actionable and human-readable plain-text, enabling version-control, cheap archiving, automatic parsing to extract data provenance, and peer-reviewable verification. We show that requiring longevity and reproducibility from scientific data analysis pipeline is realistic, without sacrificing immediate or short-term reproducibility and discuss the benefits of the criteria for scientific progress. For more, see Akhlaghi et al. (arXiv:2006.03018)

RESEARCH DATA ALLIANCE

Fundación BBVA

Universidad de La Laguna

Traces of galaxy formation
stellar populations, dynamics and morphology

Context of the research

How can I trust in the discovery/results announced in a scientific paper?

Would I be able to reproduce my own paper/results published one year ago?

What happen if I change the version of my favourite analysis software?

There is a reproducibility crisis

https://heywhatwhatdidyousay.wordpress.com

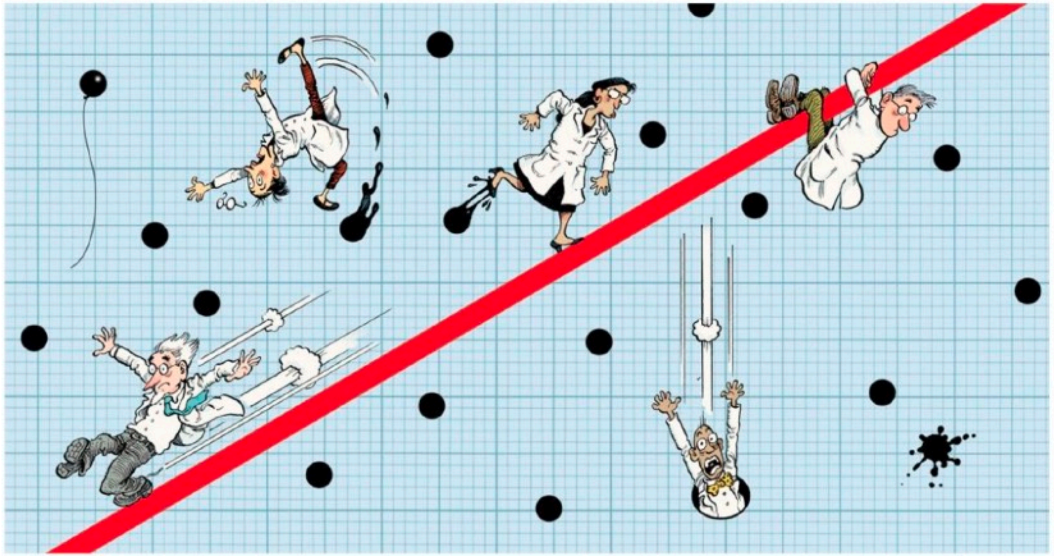


Image from nature.com ("Five ways to fix statistics", Nov 2017)

Science world evolves incredible fast.

Software, methods, tools, etc. get obsolete in a short time.

Being able to reproduce results is in the core of Science.

Science is defined by its METHOD, not its result.

What is the solution?

Maneage

Managing Data Lineage

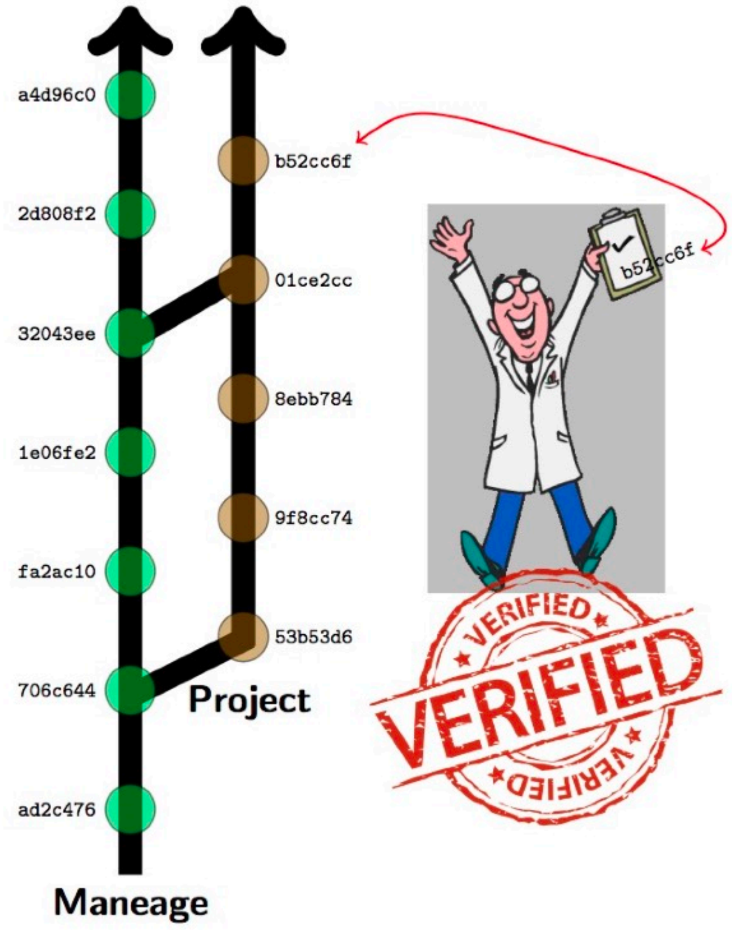
The **solution** should follow the criteria:

- **Complete** / self-contained: No Root, Non-interactive
- **Modularity**: parts should be re-usable
- **Plain text**: version control (Git), archivable
- **Minimal complexity**: never posit pluralities without necessity
- **Verifiable** inputs and outputs: data automatically verified
- **Free** and open source software: non-free software is not configurable

https://natemowry2.wordpress.com/

Maneage is introduced as a customizable template that will:

- Automatically downloads software source and data
- Builds the software in a closed environment
- Run software on data to obtain research results
- Use LaTeX to generate the paper
- The whole project is under version control (Git)




```
$ git clone http://git.maneage.org/project.git
$ ./project configure
$ ./project make
```

HORIZON 2020

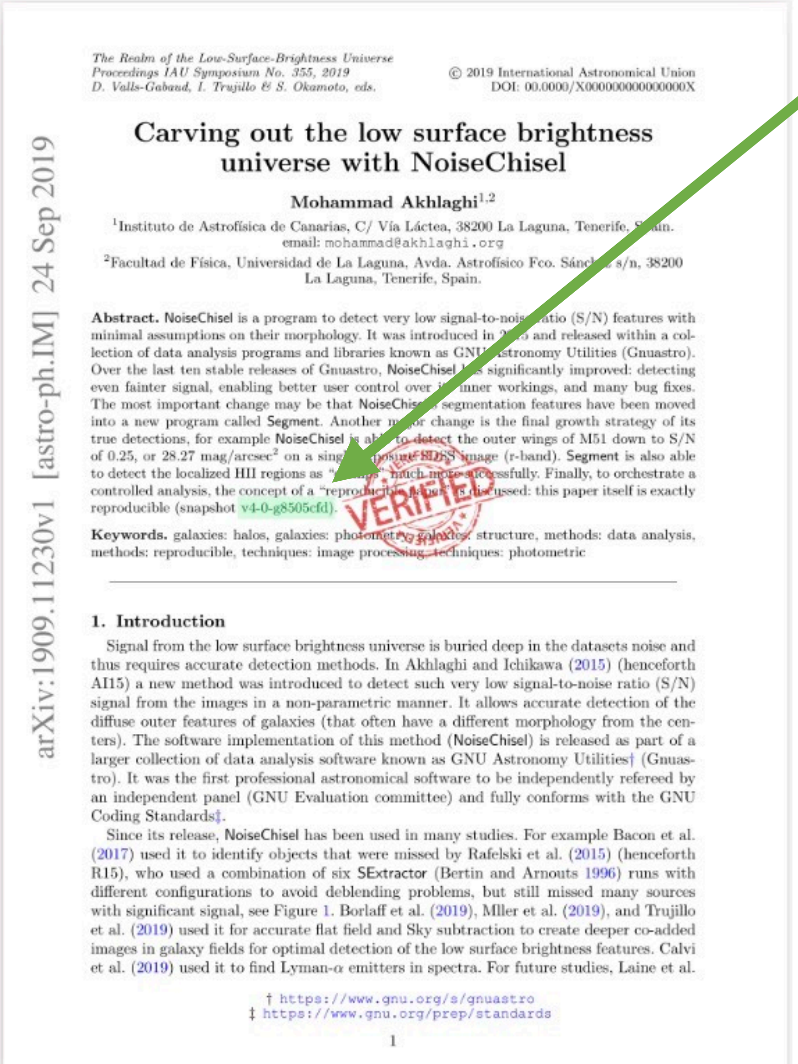
Manage was awarded with a RDA Adoption grant (EU Horizon 2020)

Results

Two examples of recent papers using **Maneage**

Last word in the abstract is the **Git hash**

Proceedings of IAU Symposium 355, in press



arXiv:1909.11230v1 [astro-ph.IM] 24 Sep 2019

The Realm of the Low-Surface-Brightness Universe
Proceedings IAU Symposium No. 355, 2019
D. Valls-Gabaud, I. Trujillo & S. Okamoto, eds.

© 2019 International Astronomical Union
DOI: 00.0000/X000000000000000X

Carving out the low surface brightness universe with NoiseChisel

Mohammad Akhlaghi[1,2]

[1]Instituto de Astrofísica de Canarias, C/ Vía Láctea, 38200 La Laguna, Tenerife, Spain.
email: mohammad@akhlaghi.org

[2]Facultad de Física, Universidad de La Laguna, Avda. Astrofísico Fco. Sánchez s/n, 38200 La Laguna, Tenerife, Spain.

Abstract. NoiseChisel is a program to detect very low signal-to-noise ratio (S/N) features with minimal assumptions on their morphology. It was introduced in 2015 and released within a collection of data analysis programs and libraries known as GNU Astronomy Utilities (Gnuastro). Over the last ten stable releases of Gnuastro, NoiseChisel has significantly improved: detecting even fainter signal, enabling better user control over its inner workings, and many bug fixes. The most important change may be that NoiseChisel's segmentation features have been moved into a new program called Segment. Another major change is the final growth strategy of its true detections, for example NoiseChisel is able to detect the outer wings of M51 down to S/N of 0.25, or 28.27 mag/arcsec² on a single-exposure SDSS image (r-band). Segment is also able to detect the localized HII regions as "clumps" much more successfully. Finally, to orchestrate a controlled analysis, the concept of a "reproducible paper" is discussed: this paper itself is exactly reproducible (snapshot v4-0-g8505cfd).

Keywords. galaxies: halos, galaxies: photometry, galaxies: structure, methods: data analysis, methods: reproducible, techniques: image processing, techniques: photometric

1. Introduction

Signal from the low surface brightness universe is buried deep in the datasets noise and thus requires accurate detection methods. In Akhlaghi and Ichikawa (2015) (henceforth AI15) a new method was introduced to detect such very low signal-to-noise ratio (S/N) signal from the images in a non-parametric manner. It allows accurate detection of the diffuse outer features of galaxies (that often have a different morphology from the centers). The software implementation of this method (NoiseChisel) is released as part of a larger collection of data analysis software known as GNU Astronomy Utilities† (Gnuastro). It was the first professional astronomical software to be independently refereed by an independent panel (GNU Evaluation committee) and fully conforms with the GNU Coding Standards‡.

Since its release, NoiseChisel has been used in many studies. For example Bacon et al. (2017) used it to identify objects that were missed by Rafelski et al. (2015) (henceforth R15), who used a combination of six SExtractor (Bertin and Arnouts 1996) runs with different configurations to avoid deblending problems, but still missed many sources with significant signal, see Figure 1. Borlaff et al. (2019), Mller et al. (2019), and Trujillo et al. (2019) used it for accurate flat field and Sky subtraction to create deeper co-added images in galaxy fields for optimal detection of the low surface brightness features. Calvi et al. (2019) used it to find Lyman-α emitters in spectra. For future studies, Laine et al.

† https://www.gnu.org/s/gnuastro
‡ https://www.gnu.org/prep/standards

1

Monthly Notices of the Royal Astronomical Society

MNRAS **491**, 5317–5329 (2020) doi:10.1093/mnras/stz3111
Advance Access publication 2019 November 14

The Sloan Digital Sky Survey extended point spread functions

Raúl Infante-Sainz[1,2★], Ignacio Trujillo[1,2] and Javier Román[1,2,3]

[1]Instituto de Astrofísica de Canarias, c/ Vía Láctea s/n, E-38205 La Laguna, Tenerife, Spain
[2]Departamento de Astrofísica, Universidad de La Laguna, E-38205 La Laguna, Tenerife, Spain
[3]Instituto de Astrofísica de Andalucía (CSIC), Glorieta de la Astronomía, E-18008 Granada, Spain

Accepted 2019 October 30. Received 2019 October 29; in original form 2019 September 10

ABSTRACT

A robust and extended characterization of the point spread function (PSF) is crucial to extract the photometric information produced by deep imaging surveys. Here, we present the extended PSFs of the Sloan Digital Sky Survey (SDSS), one of the most productive astronomical surveys of all time. By stacking ~1000 images of individual stars with different brightness, we obtain the bidimensional SDSS PSFs extending over 8 arcmin in radius for all the SDSS filters (u, g, r, i, z). This new characterization of the SDSS PSFs is near a factor of 10 larger in extension than previous PSF characterizations of the same survey. We found asymmetries in the shape of the PSFs caused by the drift scanning observing mode. The flux of the PSFs is larger along the drift scanning direction. Finally, we illustrate with an example how the PSF models can be used to remove the scattered light field produced by the brightest stars in the central region of the Coma cluster field. This particular example shows the huge importance of PSFs in the study of the low-surface brightness Universe, especially with the upcoming of ultradeep surveys, such as the Large Synoptic Survey Telescope (LSST). Following a reproducible science philosophy, we make all the PSF models and the scripts used to do the analysis of this paper publicly available (snapshot v0.4-0-gd966ad0).

Key words: instrumentation: detectors – methods: data analysis – techniques: image processing – techniques: photometric – galaxies: haloes.

1 INTRODUCTION

The point spread function (PSF) describes the response of an imaging system to the light produced by a point source. Real PSFs have complex structures as their shapes depend on the optical path that light takes as it travels through the atmosphere and multiple optical elements, mirrors, lenses, detectors, etc. For the vast majority of astronomical works, only a tiny portion of the PSF (i.e. normally a few inner arcseconds; see e.g. Trujillo et al. 2001a, b) is characterized. In practice, however, the light of both point and extended sources are spread over the entire detector due to the effect of the PSF at large radii. Therefore, it is necessary to have a good understanding of its structure along the entire detector (typically extending over arcminutes or more).

Extended PSFs have become a vital tool to obtain precise photometric information in modern astronomical surveys. For instance, Slater, Harding & Mihos (2009) modelled the extended PSF and the internal reflections produced by the stars of the Burrell Schmidt telescope and showed that virtually all the pixels of the image are dominated by the scattered light by both stars and galaxies at 29.5 mag arcsec^{-2} (V-band). Trujillo & Fliri (2016) also characterized and used the extended PSF of the 10.4 m Gran Telescopio Canarias (GTC) telescope to model and remove the scattered light in ultradeep observations of the UGC 00180 galaxy. Even more troublesome for low-surface brightness studies is the finding (see e.g. Trujillo & Bakos 2013; Sandin 2014, 2015) that the outer regions of astronomical objects are severely affected by their own scattered light produced by the convolution with the PSF. In order to correct this effect, Karabal et al. (2017) generated the PSF and models of the internal reflections from images of the Canada–France–Hawaii Telescope (CFHT) to de-convolve a sample of three galaxies and correct them from instrumental scattered light. More recently, Román, Trujillo & Montes (2019) characterized the PSFs of the Stripe 82 survey and used them to model and correct the scattered light field produced by stars to study the optical properties of the Galactic cirri. All the above works have shown that having an extended PSF is crucial when accurate photometric and structure properties of astronomical objects at low-surface brightness levels are required.

One of the most commonly used surveys for measuring photometric properties of astronomical objects is the Sloan Sky Digital Survey (SDSS; York et al. 2000), covering 14 555 deg^2 on the sky (just over 35 per cent of the full sky) in five photometric bands (u, g, r, i, and z). Although SDSS is a relatively shallow survey compared

★ E-mail: infantesainz@gmail.com

© 2019 The Author(s)
Published by Oxford University Press on behalf of the Royal Astronomical Society

MNRAS 491, 5317