The HCG-16 study

IAA-CSIC

S. Luna, M. Jones, S. Sánchez Expósito, J. Román, J. Garrido, L. Verdes-Montenegro

E-OSSR Onboarding Presentation

19th February 2021

ESCAPE - The European Science Cluster of Astronomy & Particle Physics ESFRI Research Infrastructures has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement n° 824064.

Introduction

- Partner: IAA-CSIC

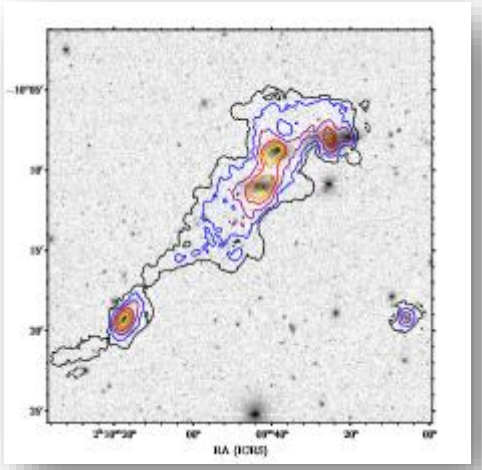
- Science case: *Evolution of compact groups from intermediate to final stages; A case study of the HI content of HCG 16 (DOI)*

- Heads up: The speaker is a sysadmin helping out astronomers to write a reproducible workflow.

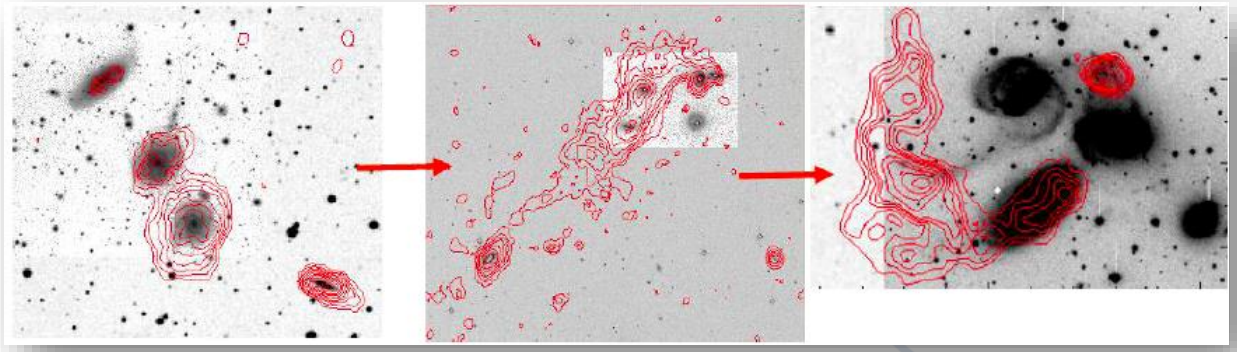
Funded by the European Union's Horizon 2020 - Grant N° 824064

Introduction

- Atomic Gas (HI)
  - HI atoms produce radio emission (1420 MHz)
  - Tracer of galaxy formation and evolution
- Hickson Compact Groups (HCGs):
  - Groups with 4 or more galaxies
  - Nearby galaxies
  - Isolated group
  - There are 100 HCGs, this work focuses on HCG-16



HCG-16
M. Jones et al. 2019



1. HI in galaxies 2. HI in IGM 3. HI out of galaxies

Evolutionary sequence suggested by Verdes-Montenegro et al. 2001

AMIGA

http://amiga.iaa.es

Funded by the European Union's Horizon 2020 - Grant N° 824064

Introduction

- HI analyses to study galaxy formation and evolution is of interest for the SKA community
  - SKA HI Science Working Group

- No SKA data yet… so:
  - VLA: Very Large Array
  - Radio interferometer
  - 27 dishes, 25m diameter

Funded by the European Union's Horizon 2020 - Grant N° 824064

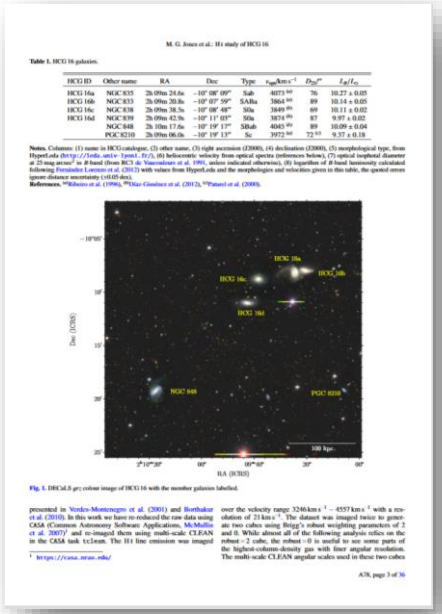
Introduction

- Software: pipeline to reproduce the HCG-16 study
- Purpose: enable end-to-end reproducibility, from initial data to plots in the paper. We open up the whole workflow so researchers should be able to:
  - Understand what was done to the data
  - Verify the analysis
  - Reuse data and code
- Use case: From a researcher's point of view, "can I use the ESCAPE ecosystem to share my analysis with a colleague or a referee?"
- Workflow: https://github.com/AMIGA-IAA/hcg-16

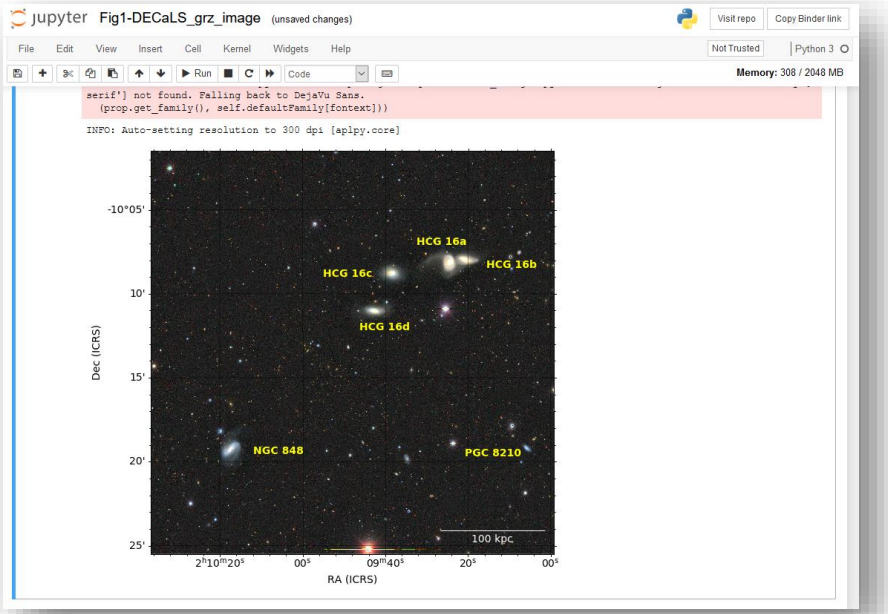
Funded by the European Union's Horizon 2020 - Grant N° 824064

Introduction

- Most figures in the paper can be fully reproduced with a Jupyter notebook [provided in GitHub]



Published PDF



Interactive notebook on [mybinder]

Funded by the European Union's Horizon 2020 - Grant N° 824064

Introduction

- Workflow:
  - First download and install runtime dependencies
    - Note: the exception is docker, which is expected to be already available on the target computer
    - Runtime dependencies are provided in a conda environment
  - Download source code for the analysis and input data.
    - Code from GitHub
    - Data from EUDAT's B2SHARE service
  - Run the pipeline
    1. Flagging and calibration with CASA (docker container)
    2. Imaging with CASA (docker container)
    3. Masking with SoFiA (docker container)
    4. Plotting with Jupyter Notebooks (conda environment)

Funded by the European Union's Horizon 2020 - Grant N° 824064

Software/Service Development

- Software Development Lifecycle Strategies
  - GitHub: git brach/pull request/merge
  - Pin dependencies. Use conda first, docker as last resort.
- Development:
  - Coding style: write understandable code for your future self
  - Versioning: released version 1.0.0 for publication, then point releases for solving issues (e.g. docker's change of ToS)
  - Maintenance: only when necessary to preserve functionality
  - Documentation: comprehensive README file provided
  - Software quality standards: no formal approach
- Testing and efficiency optimization strategies
  - None

Funded by the European Union's Horizon 2020 - Grant N° 824064

Software/Service Development

- Information on how to run the software
  - Step-by-step guide on the repository's README file
- software licenses
  - MIT license
- General guidelines that are followed
  - Document as much as possible to help others reuse it
  - Automate every possible step
  - Streamline deployment to enhance portability between systems

Funded by the European Union's Horizon 2020 - Grant N° 824064

Software/Service Requirements

- Operating System, compilation environment
  - It has been tested on Ubuntu 18.04
  - In theory you just need bash and docker pre-installed
  - Everything else installs automatically with conda
- Hardware requirements
  - Minimum of 10GB disk space on the working directory
  - Recommended minimum of 2 CPU cores, 2GB RAM
- Containerisation and portability requirements
  - Password-less “sudo docker” is a pre-requisite
- Workflow / interface requirements to other software/services
  - Internet access: docker hub, conda packages, input data

Funded by the European Union’s
Horizon 2020 - Grant N° 824064

OSSR Integration

- What is available?
  - Code, data, scientific publication
  - https://github.com/AMIGA-IAA/hcg-16
- What will be onboarded (source code, container, test workflow incl. data)?
  - Code, data
- Are there open points and requirements?
  - From the OSSR integration point of view, we don't know

Funded by the European Union's Horizon 2020 - Grant N° 824064

OSSR Integration

- What is the ”user story” of a EOSC user taking on the software/service?
  - From the data side (what data can be analysed and how)
    - All the data and the code are publicly available, so as long as the installation works, the analysis workflow should be fully reproducible.
  - From the OSSR side (how to find data and easy use demos, tutorials, documentation, …)
    - Our aim is to provide a self-explanatory README, with steps to reproduce the workflow, and links to everything: code, data and scientific publication.

Funded by the European Union’s Horizon 2020 - Grant N° 824064

Time for a short demo (~10 min)

- Show how the software is used and what is the outcome
  - Let's go to the README (showcase mybinder link)
- What should and can a EOSC user do with the software?
  - Understand what was done to the data
  - Verify the analysis
  - Reuse data and code

Funded by the European Union's Horizon 2020 - Grant N° 824064

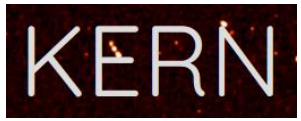
Open Points and Discussion Time

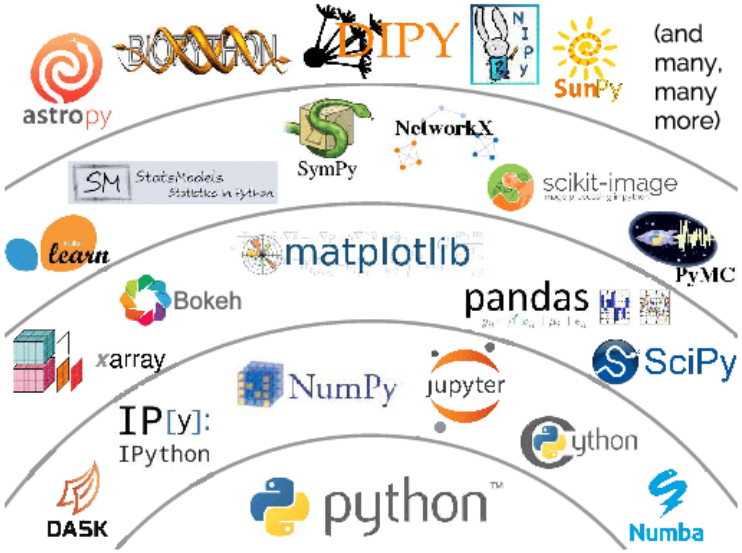
- Which of your questions have not been covered so far?
- What do you want to discuss?
  - Our main goal here is to check whether end-to-end reproducibility is feasible, and how much is the effort
    - What are the blockers?
      - Overhead associated when trying to achieve reproducibility
      - The original data could not be downloaded automatically so we had to re-host it on the EUDAT's B2SHARE service
      - SlicerAstro: interactive tool whose commands could not be scripted in the pipeline
      - IDL: proprietary software hinders Open Science. Hopefully for this specific case we are looking into the alternative, open source GDL

Funded by the European Union's Horizon 2020 - Grant N° 824064

Open Points and Discussion Time

- This work would have not been possible without:





Funded by the European Union's Horizon 2020 - Grant N° 824064