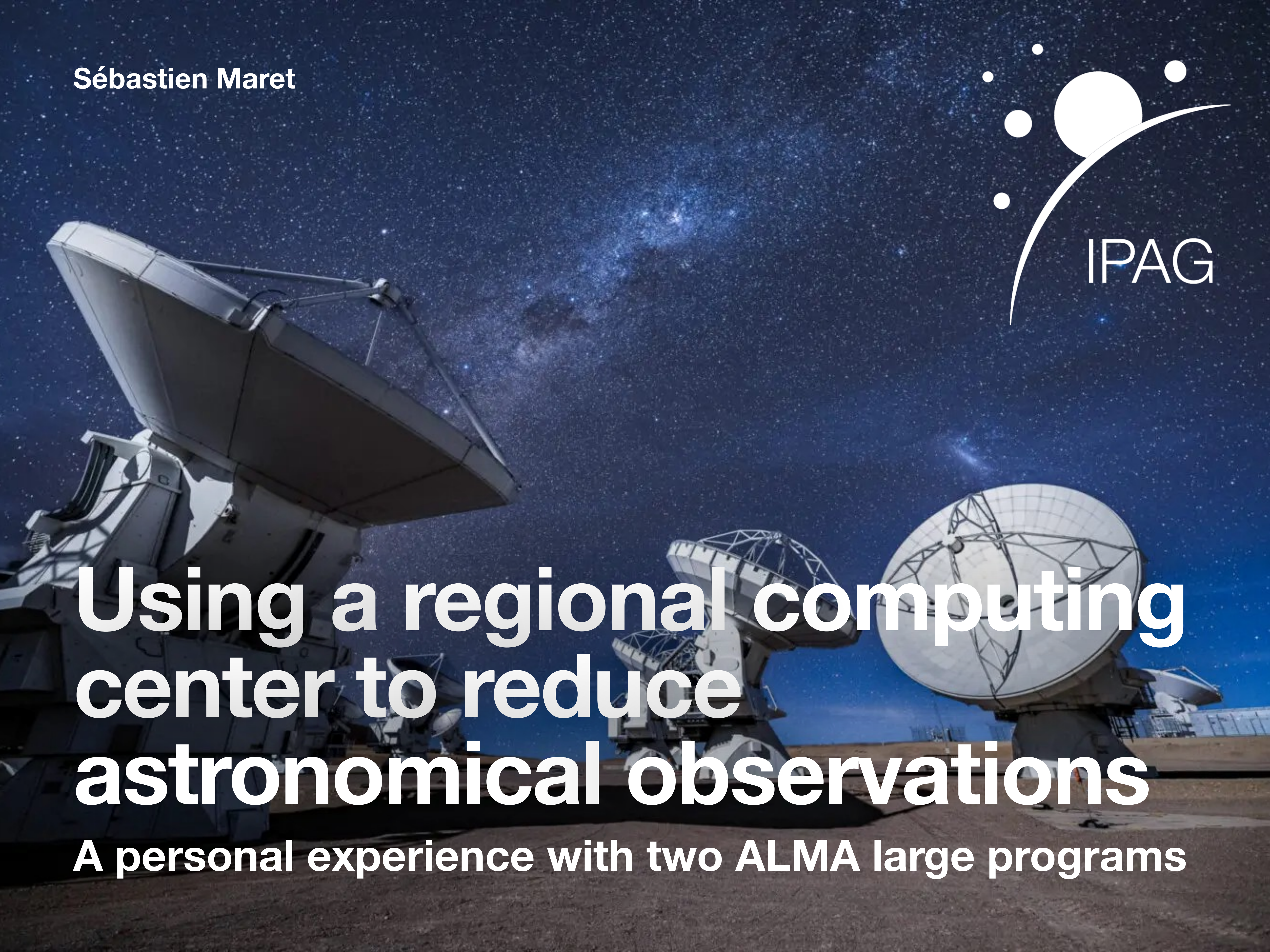


Sébastien Maret

Using a regional computing center to reduce astronomical observations

A personal experience with two ALMA large programs

Content of this presentation

- Reduction of ALMA observations in a nutshell
- Two example of ALMA large programs: COMPASS and Diskstrat
- Implementation and deployment of a data reduction pipeline on the Gricad infrastructure
- Lessons learned from this experiment: advantages and disadvantages of this approach

Reduction of ALMA observations in a nutshell

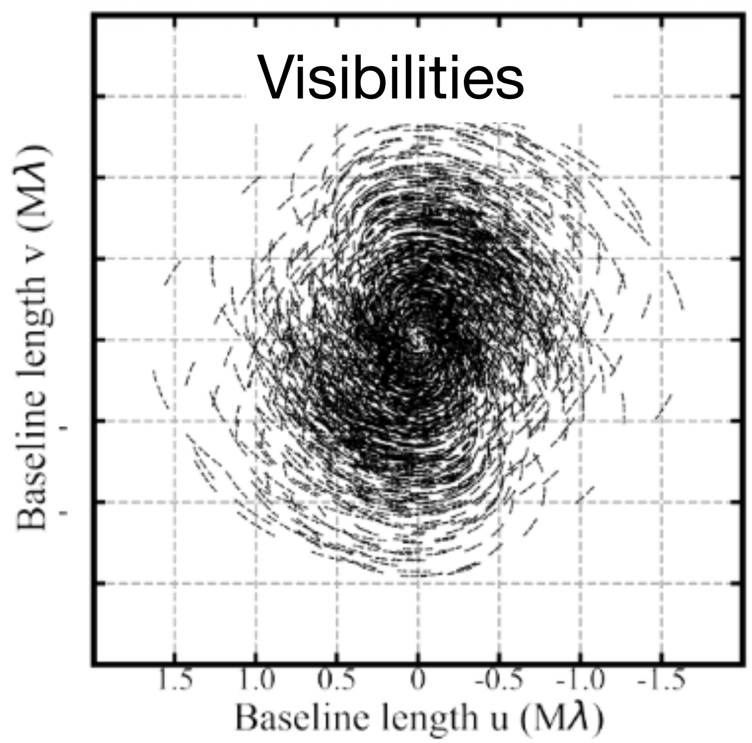
How does a radio-interferometer work?

- A interferometer combines the electric signals from many antennas to produce **complex visibilities**.

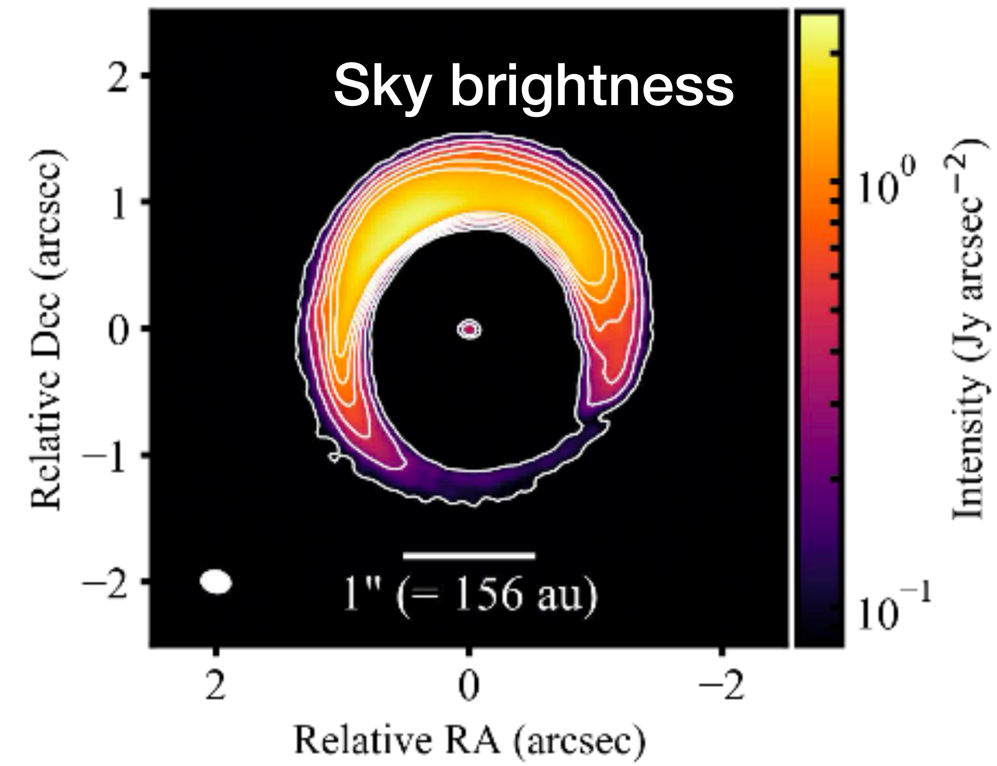
- Each of these visibility correspond to the **Fourier transform of the sky brightness**, in a given position of the uv plane.

Reduction of ALMA observations in a nutshell

What is data reduction?



HD 142527



Yamaguchi et al. 2020

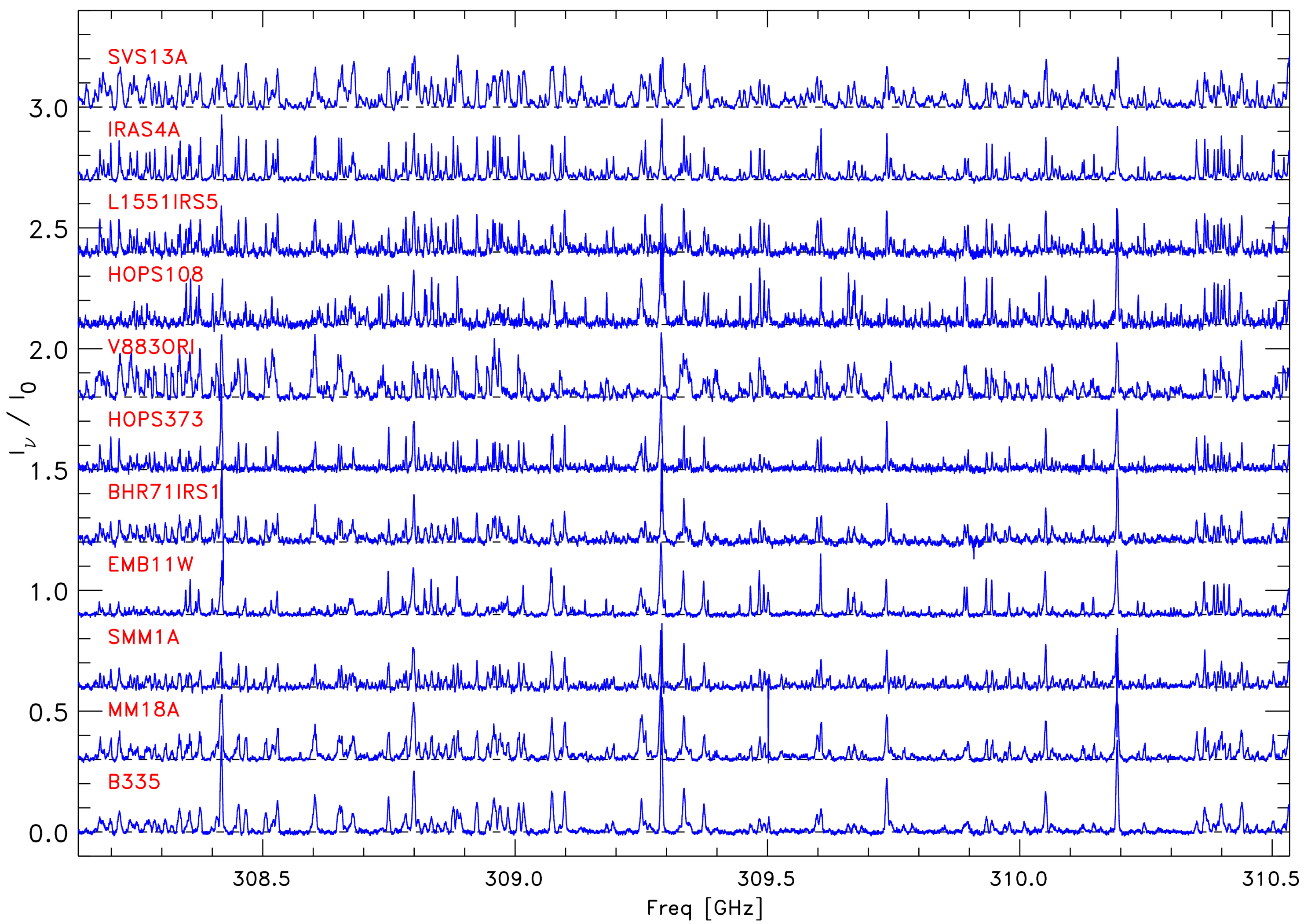
- In order to produce an image of the sky, one first need to **convert the visibilities measured by the instrument into physical units.** This requires to remove the instrumental and atmospheric effets. This step is called **data calibration.**

- The next step is to **produce astronomical images from these visibilities.** It's independent of the instrument used. This step is called **data reduction,** because the data volume is reduced in the process.

Example: two ALMA large programs

COMPASS: Spectral surveys of young protostars

- **ALMA cycle 9 large program,** P.I. Jes Jørgensen (NL)
- Main goal: **comprehensive inventory of the complex organic molecules (COMs)** composition in a large sample of 11 **young protostars.**
- 33 GHz frequency window in Band 7 at 0.5 km/s spectral resolution, 9 frequency settings.
- Data volume: **35 TB**



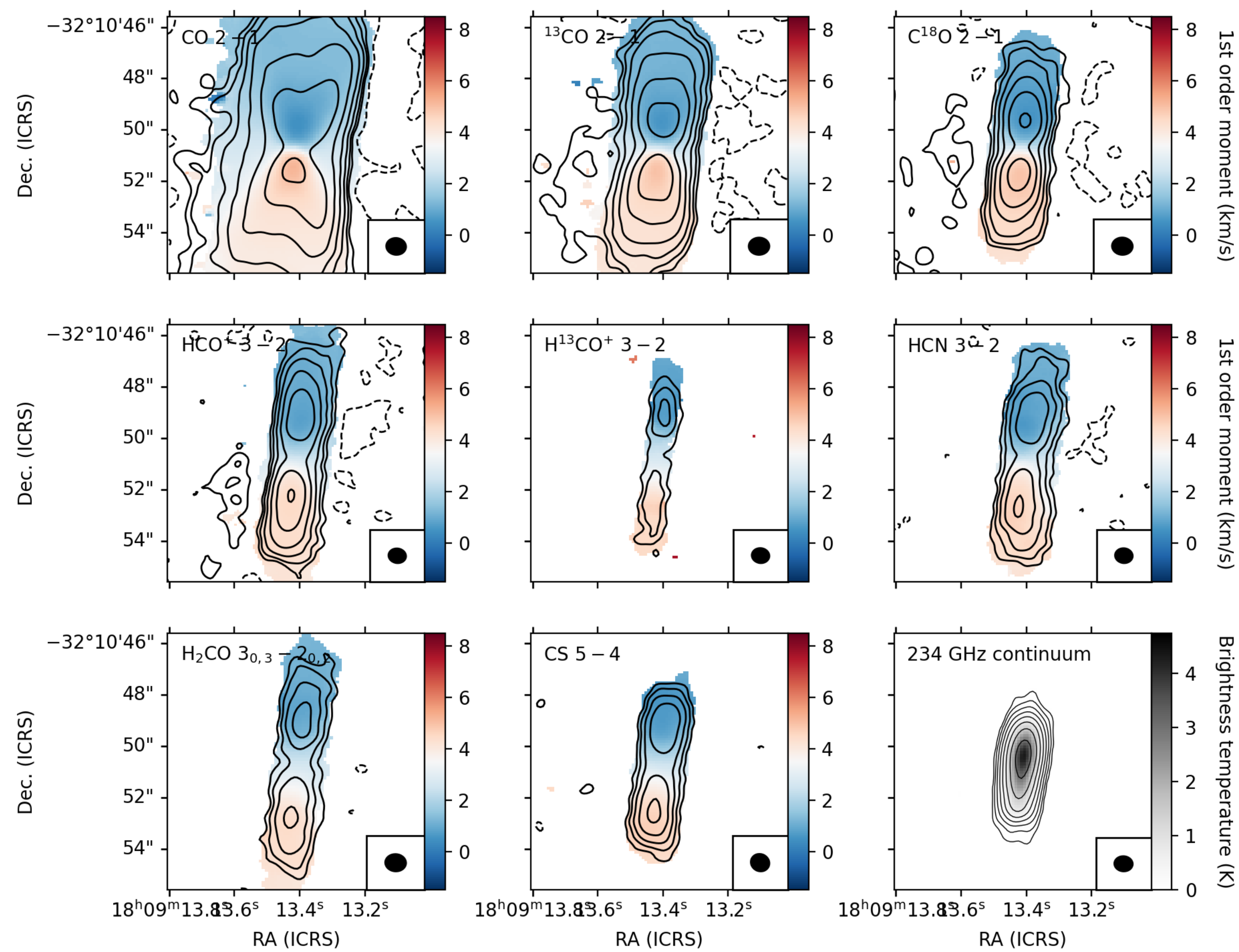
Jørgensen et al. (in prep.)

https://erda.ku.dk/vgrid/COMPASS/

Example: two ALMA large programs

Diskstrat: Line maps of edge-on protoplanetary disks

- **ALMA cycle 11 large program**, PI Romane Le Gal.
- Main goal: map the vertical chemical structure of 9 edge-on **protoplanetary disks.**
- **3 frequency settings in band 6 at 0.04 km/s resolution**, covering 27 lines, and 3 continuum bands.
- Data volume: **40 TB**



Diskstrat collab. (in prep)

Data reduction of COMPASS and Diskstrat

Implementation (1/2)

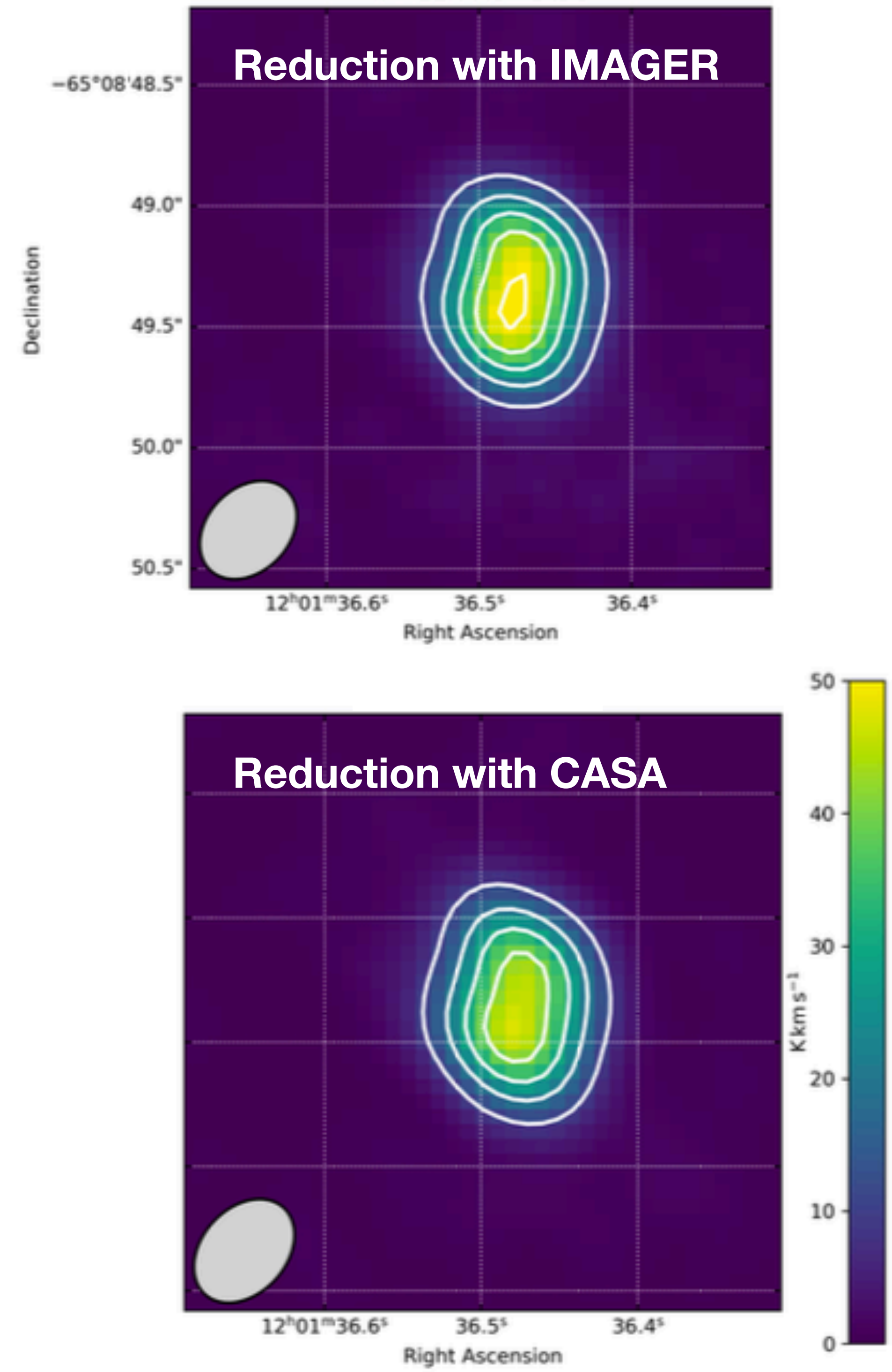
- The first COMPASS observations were reduced independently with <u>CASA</u> (developed by the NRAO) and <u>IMAGER</u> (developed in Bordeaux by Stéphane Guilloteau, and based on Gildas, developed by IRAM)

- Comparison between the two: same results, but **IMAGER is x20 faster** (hours vs days for one frequency setting per source)

    - **I/O are a bottleneck**; IMAGER does most of the treatment in the RAM.

    - IMAGER is parallelized (at the node level) with OpenMP.

BHR 71
Methyl formate (287.8 GHz)

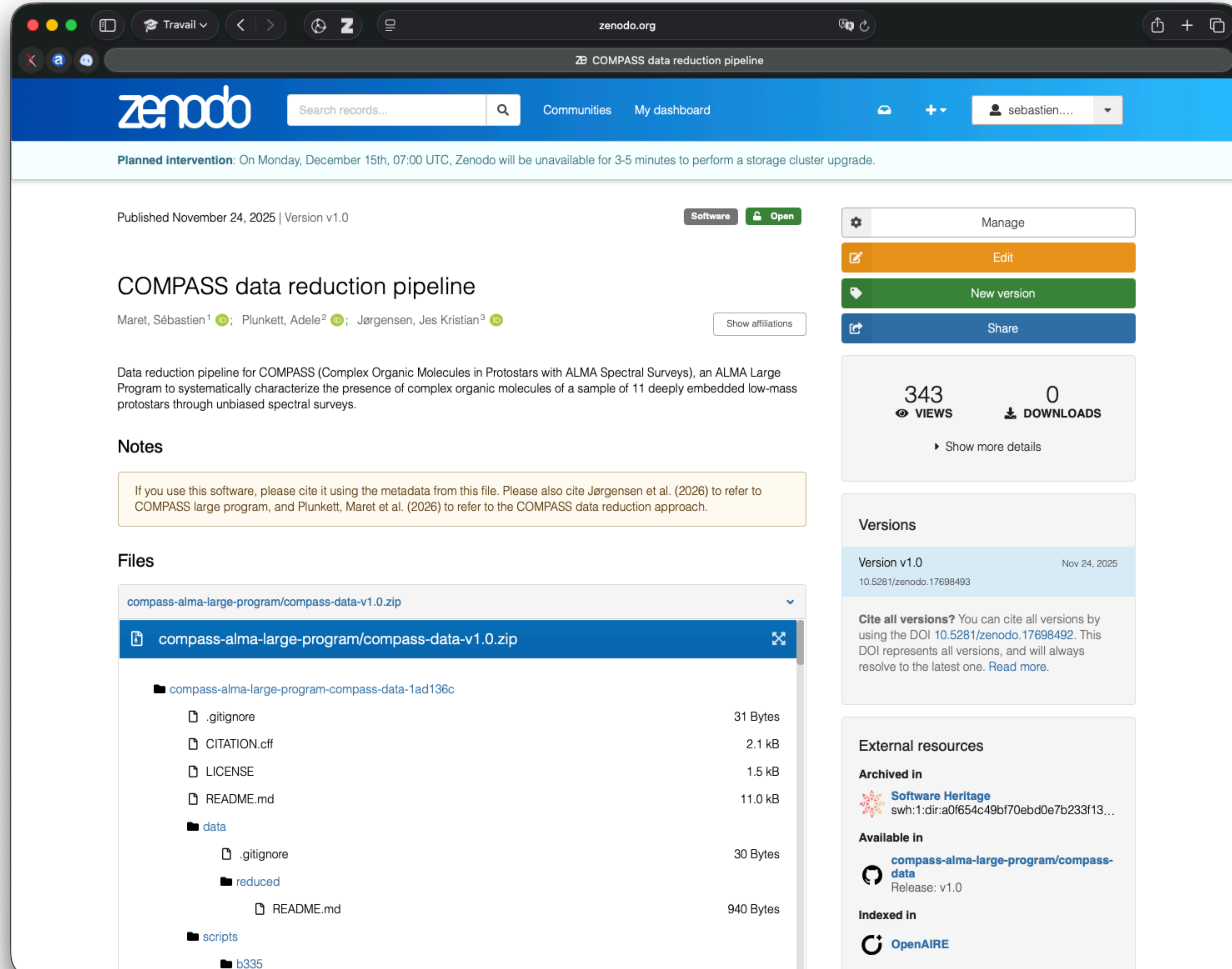


Plunkett, Maret et al. (in prep)

Data reduction of COMPASS and Diskstrat

Implementation (2/2)

- We developed a r**eduction pipeline for COMPASS based on IMAGER**, with additional scripts in Python (for post-processing and visualization of the results).
- **Parallelization** is trivial because the reduction of each source and frequency setting is independent of the others.
- The data reduction can span over months, so it's important to have a controlled software environnement. We use Nix to ensure **reproducibility** (we wrote a package for IMAGER)
- The same pipeline was later adapted for the **Diskstrat** large program.



https://doi.org/10.5281/zenodo.17698493

Data reduction of COMPASS and Diskstrat

Deployment on the Gricad infrastructure

- We deployed the COMPASS pipeline on the Dahu cluster.
  - The reduction for each frequency setting and source is **run on parallel on a several nodes** (typically 36 cores, 192 GB RAM)
  - Some of the observations required to use « fat » nodes (16 cores, 1.5 TB RAM)
  - Total computing time: ~ **10 000 CPU hours**
- **Storage:** Bettik distributed high performance scratch (based on BeeGFS)
- Long term **archival and distribution** of the reduced data: MANTIS (based on iRods)

GRICAD

Grenoble Alpes Recherche | Infrastructure de calcul intensif et de données

Conclusions

Lessons learned (1/2)

- Using a regional computing center for reducing data has one major **advantage**:
  - **Clusters with ample ressources (computing but most importantly storage) are readily available**; the cost of such ressources would be prohibitive for most institutes.
- It also have **disadvantages**:
  - Pipelines needs to be designed to be run in **batch mode.**
  - Requires to learn to use a **job scheduler** (OAR, SLURM…).
  - Some tasks requires user’s interaction (e.g. **visualization**).
  - **Deployment** of the reduction software can be an issue.

Conclusions

Lessons learned (2/2)

- Having **efficient data reduction softwares** is key !
  - Related question: how to support the development of these?
  - IMAGER is developed as part of the **SNO Radioastronomie millimétrique.**
- With the new planned facilities (e.g. ALMA WSU, ELT, SKA…) **the data volume and the computational needed for the data reduction will increase dramatically**. It will become inevitable to use computing center for this.
- The **community is not ready for this shift**; the ASUM could perhaps help (training ?)