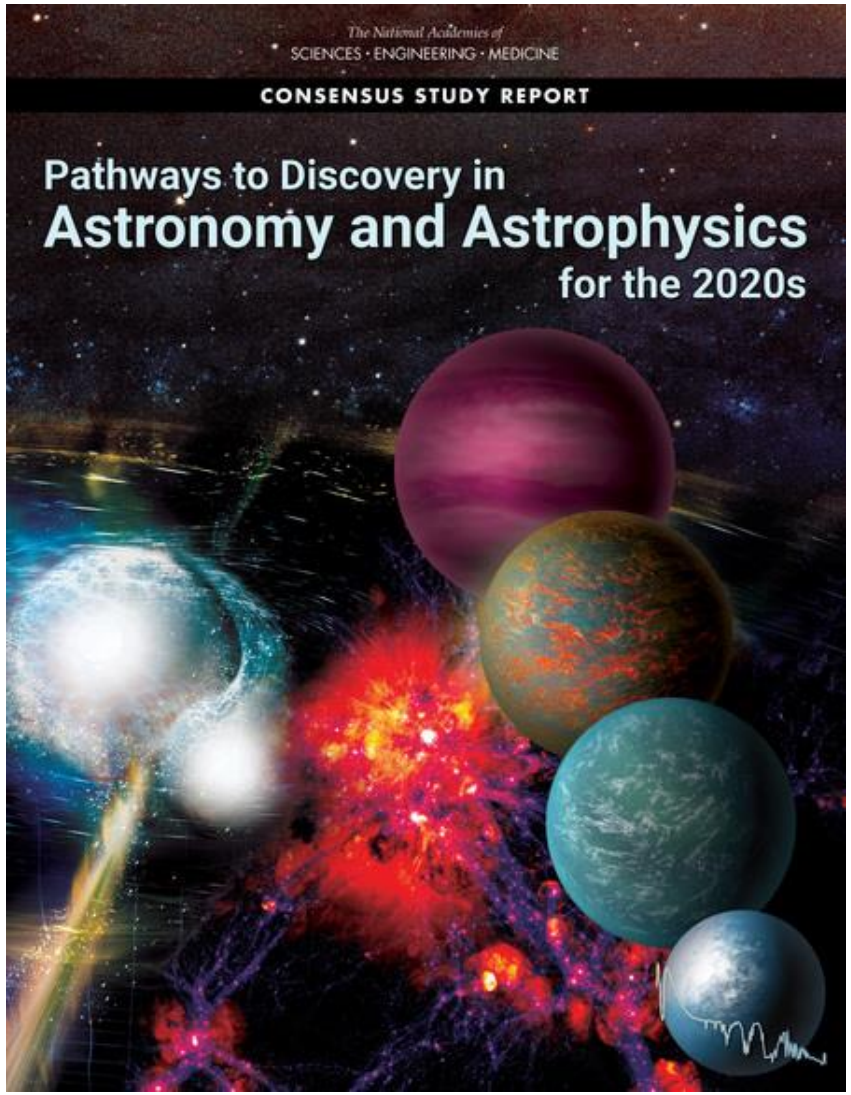
Report on the workshop: "The Future of Astronomical Data Infrastructure" and future plans.

Luca Rizzi, Program Director, MPS/AST



Astro2020: Findings

- Software development has become an essential part of every sub-field of astronomy. However, software developers and large software development efforts are not adequately funded or supported by existing structures
- Progress will come from an **end-to-end approach** that considers the entire flow of data from the instrument, to the archive, to analysis and publication. Increasing the prevalence of both science-ready data products and effective archives is best achieved if done hand-in- hand with each other. Making codes publicly available will help to **minimize redundancy, encourage the adoption of common standards, and promote applications using multiple data sets.**

Astro2020: Recommendation

- The National Science Foundation and stakeholders should develop a plan to address how to **design, build, deploy, and sustain pipelines** for producing science-ready data across all general-purpose ground-based observatories (both federally and privately funded), providing funding in exchange for ensuring that all pipelined observations are archived in a standard format for eventual public use.

Future of Astronomical Data Infrastructure

NSF, NASA, and Center for Computational Astrophysics of the Flatiron Institute

4 days workshop, in person and remote, February 13-16 2023

Goal: identify technical solutions, implementation options, a long-term vision, and a tentative roadmap to address the recommendations of Astro2020. Participants were encouraged to think big and be ambitious.

110 applicants; 40 attended in person, 26 remotely

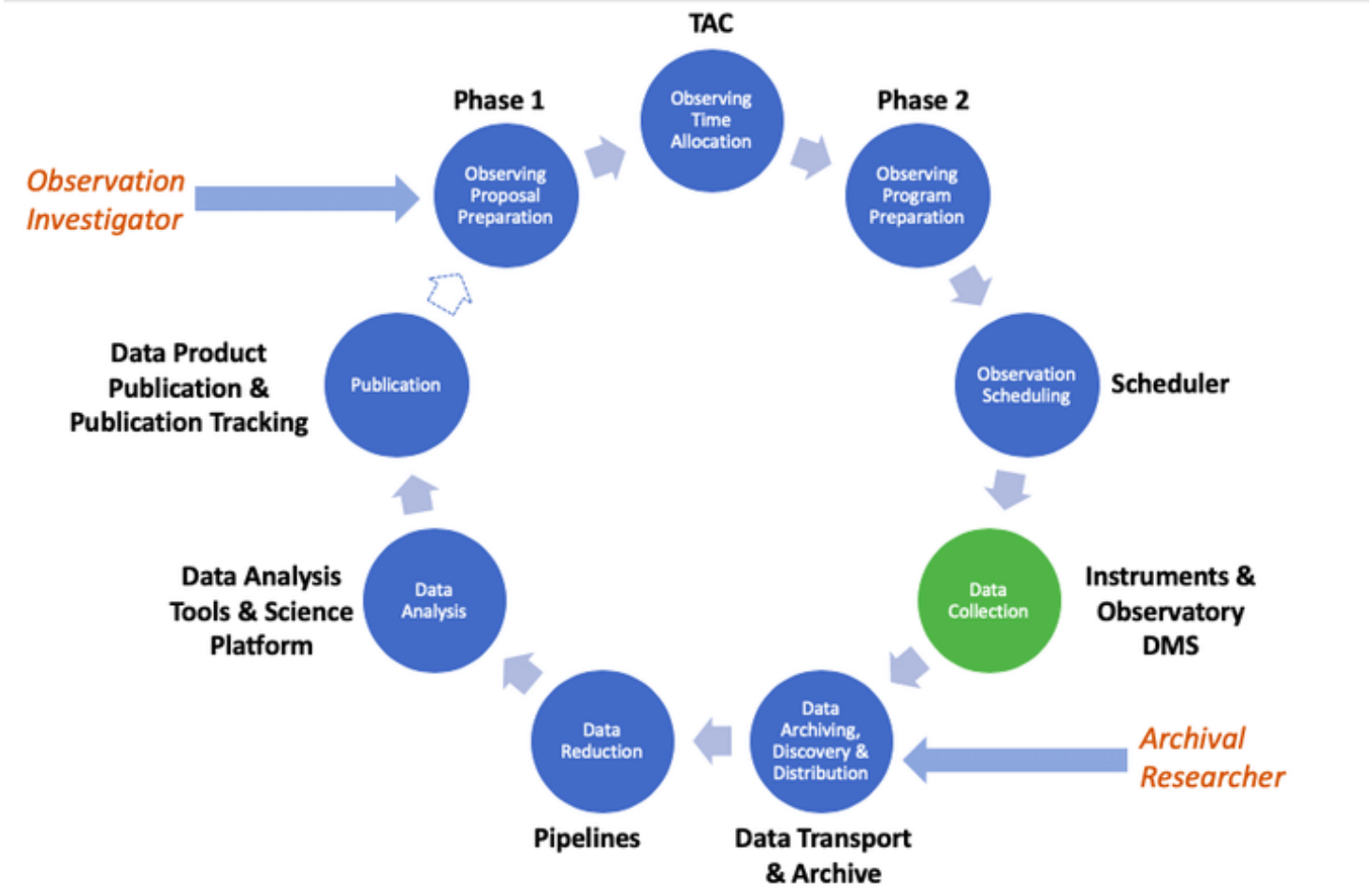
Broad representation from all data centers (US and abroad), surveys, ESO, universities, individual developers, NASA, ADS, observing facilities, community-led projects, high performance computing

Results

- It’s not about pipelines and archives, it’s about building an end-to-end ecosystem (the circle of discovery)
- Individual stakeholders are building individual pieces, often with a lot of duplication; nobody is tasked with building shared, end-to-end solutions
- There is a strong desire and a sense of urgency in the community: they want to work better and together
- Structural issues of the profession and in the funding models prevent effective collaborations
- The community supports the creation of a coordination structure, but the actual implementation needs more work

It’s an end-to-end ecosystem

- Software in all parts of this cycle (tools) and in the arrows between parts (standards).
- Development of the system should be inspired by broadening participation



Structural issues, inadequate funding models, nobody is tasked with building shared solution, duplication of effort

- We have been largely failing to converge behind sustainable collaborative software development for decades
- Not a problem with individuals (we all mean well), funding (it's not the total that is inadequate), or NSF/NASA (other STEM fields have the same problems)
- Both single-mission funding and joint infrastructure funding are inadequate models. No real incentive to re-use existing solutions or to build for the community.
- How do we get to a place where doing the right thing is easier and better rewarded than doing the wrong thing? How do we define "good results" (re-use as a measure of success)

The community supports the creation of a coordinating structure – no consensus on the preferred solution

POSSIBLE MODELS

- A **coordinating committee** to identify and prioritize critical software infrastructure development for the benefit of the broad astronomical community. This would involve brokering, facilitating, and empowering collaborations across teams who may be funded by independent agencies or projects.
- An **inter-agency program office**, tasked to oversee policies and efforts that lead to a flourishing astrophysics data and software ecosystem. The program office would fund and oversee activities around developing policies, structures, and incentive systems that support getting the most science done by the broadest number of people, within the context of the data and software ecosystem.
- A **new organization** to support the development of core software tools, staffed with engineers, scientists, project managers and support staff. Could be made of a core engineering pool plus a network of engineers embedded in individual projects.
- A **multi-site, federated organization** that serves as a layer on top of the existing facilities. Each site would be responsible for one (or more) work-packages that address a component of the end-to-end system. All software created in the work packages would be deployed identically at all sites, and each work package provider must consider all requirements and needs of all sites to be fully service-oriented.

Next steps

A report

Draft already circulated and accepted – Release in the next few weeks

Phase 0

Engagement, discussion, presentations to achieve buy-in from funding agencies.

Phase 1

The NSF, NASA and other stakeholders empower and fund a small steering committee

Level 1 requirements; Governance model; Develop use cases with the community to drive the processes.

Phase 2

Agencies prepare for the implementation plan delivered in Phase 1.

Summary

- The community embraces the recommendations of Astro2020
- We cannot build pipelines and produce science-ready data in isolation: it's a complex ecosystem (cycle of discovery)
- Efforts to build the ecosystem have failed due to structural issues
  - Funding models
  - Incentive models
- The community supports the creation of a "coordinating structure"

A world where the policies, culture, and incentive systems exist that lead to a flourishing astrophysics data and software ecosystem that supports getting the most science done by the broadest number of people

Thank you!