A vision for the future of astronomical archives

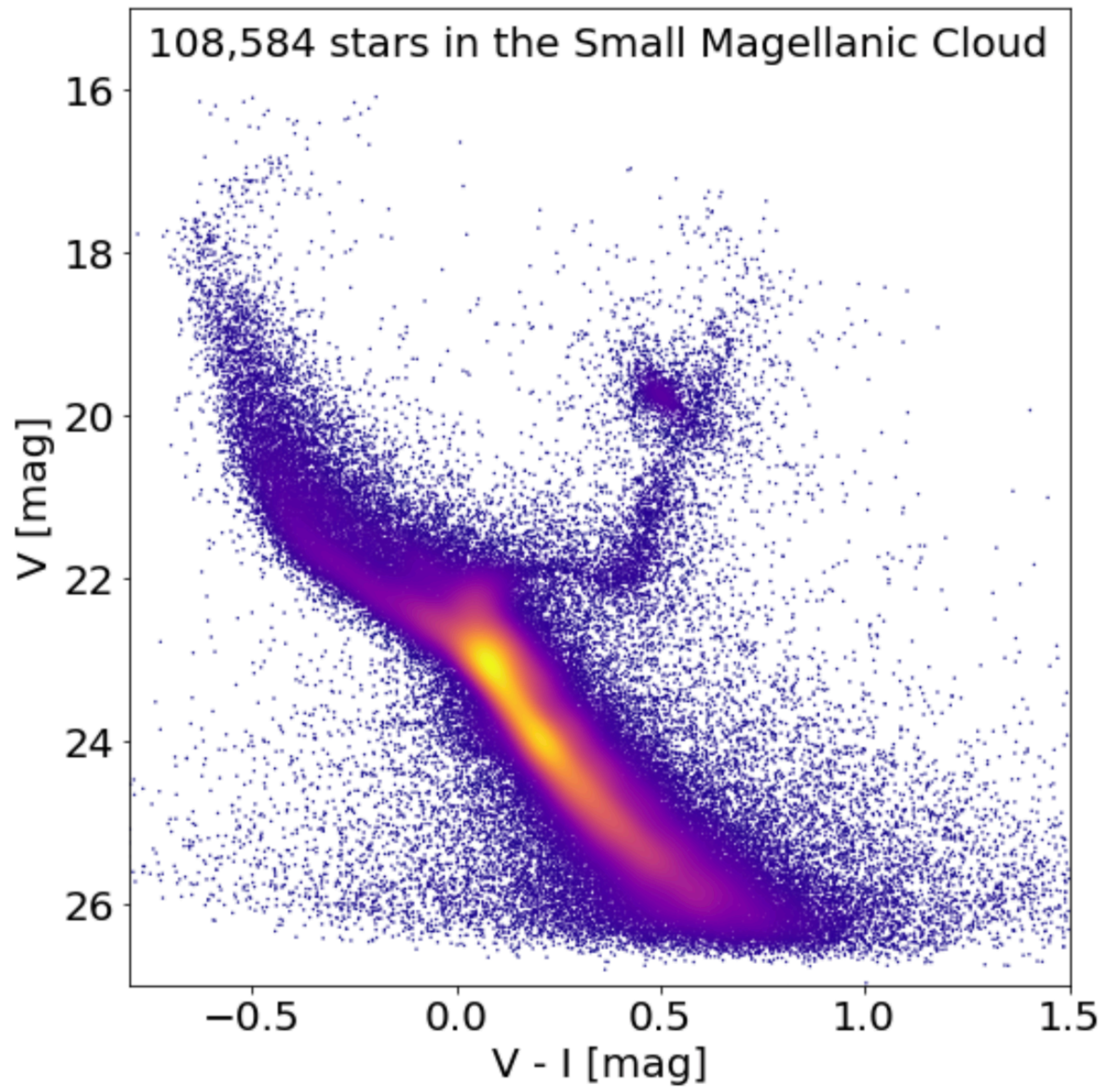
Astro2020: An Enabling Foundation for Research

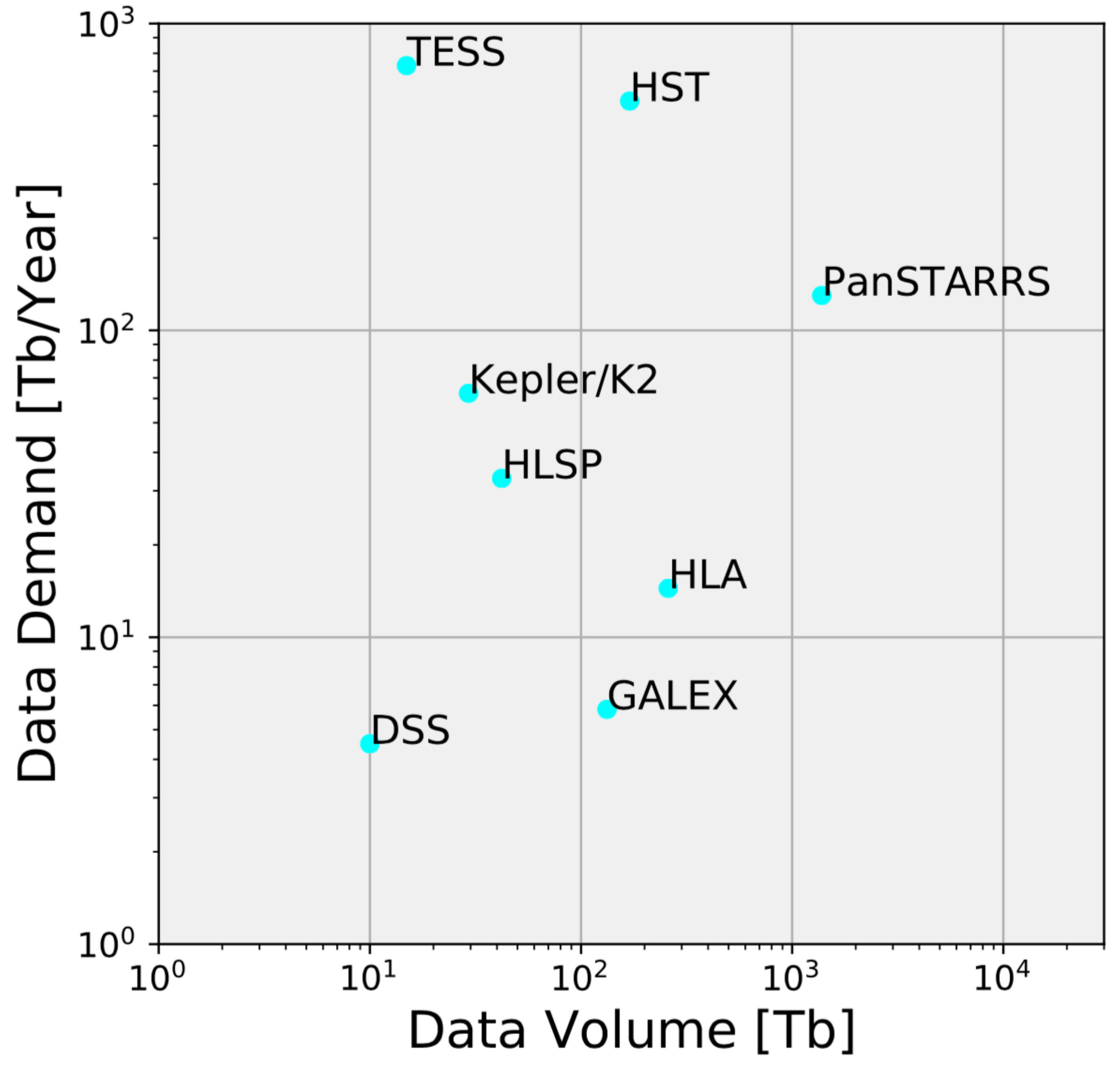
Arfon Smith, Space Telescope Science Institute

What do archives do?

- Store, preserve, and distribute the bits.
- Place core services and interfaces around the data.
- House experts that interface with the community to understand and *anticipate* their needs.
- Build value-added interfaces tailored to specific scientific domains.
- Curate knowledge, including generating value added products.
- Collaborate with other archives.

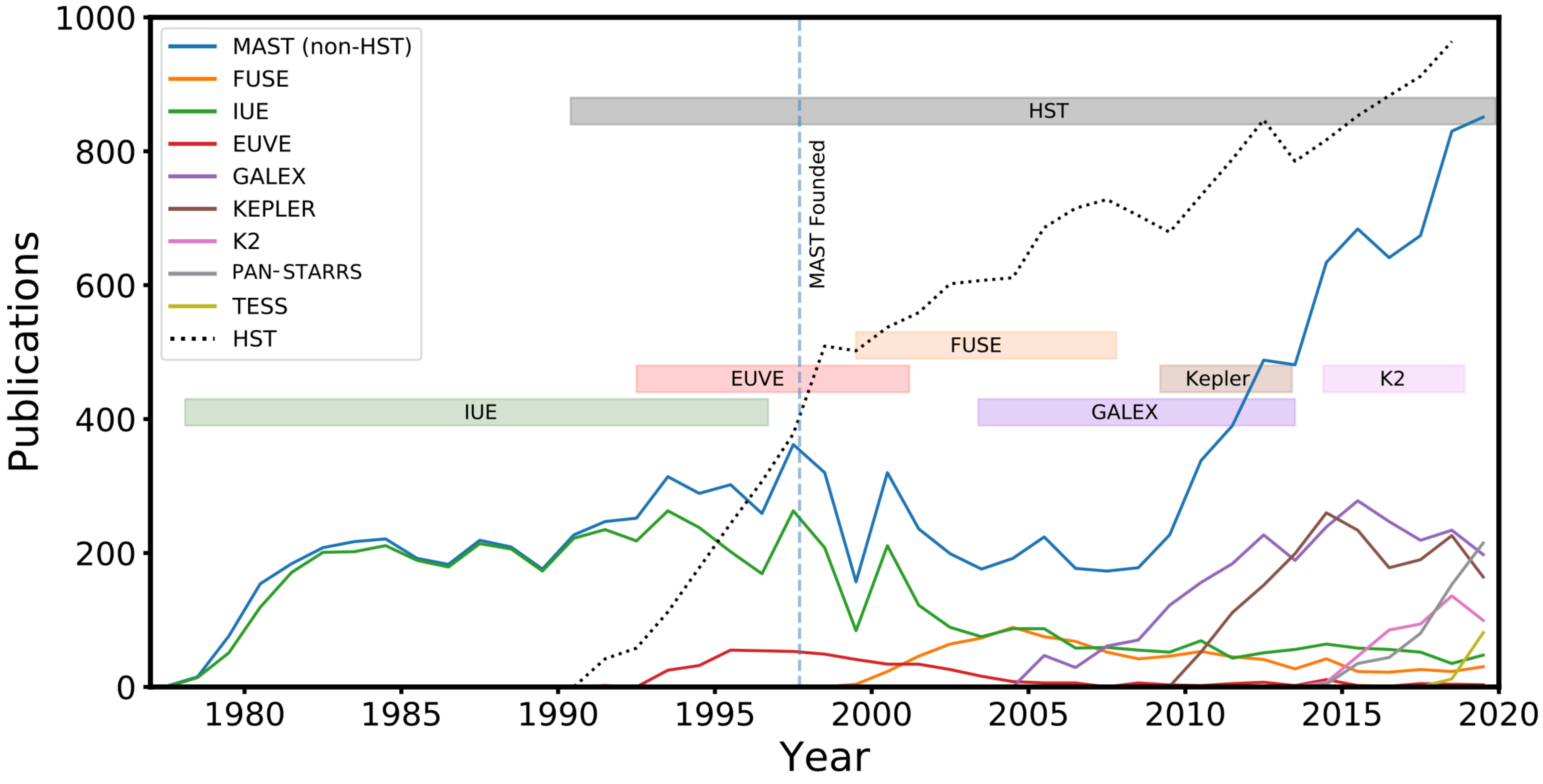
Archives are repositories of expertise and knowledge





MAST’s most popular missions

Archives enable science



Archives must evolve to support the science of the 2020s

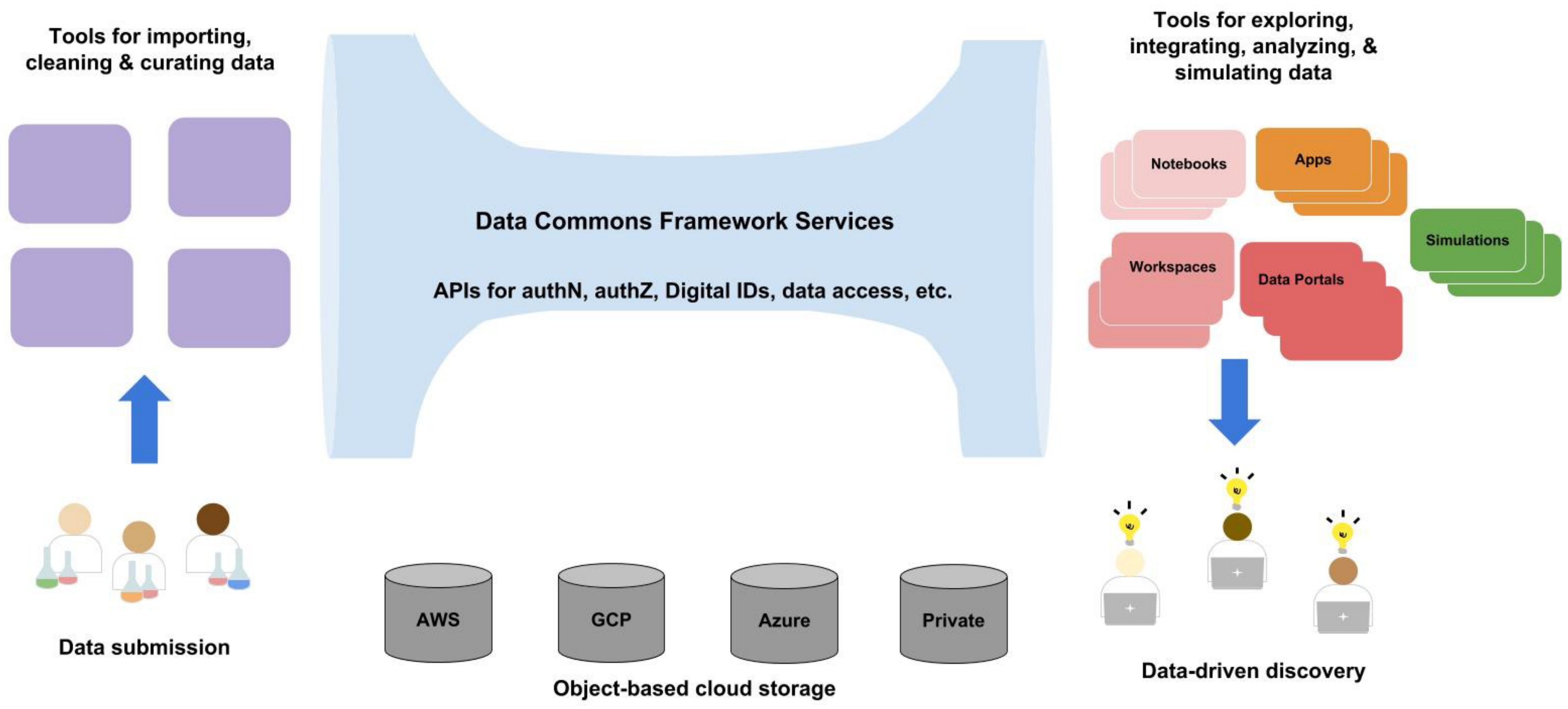
- Current model of search → retrieve → analyze (locally) of data is sub-optimal for peta-scale datasets (and combinations thereof) and sophisticated analyses (e.g. applying machine learning techniques).
- “Color of money” physical siloing of key datasets (e.g. VRO/LSST, WFIRST) likely to be **rate-limiting for science**.
- Standard operating model of requires significant time and expense devoted to important, but non-specialist “data plumbing” problems at centers.
- Most archives and missions lagging significantly behind commercial sector data management (e.g. those using public/commercial cloud computing).
- Very limited adoption of open source and *open development* paradigm for infrastructure resulting in minimal reuse between centers, facilities, and research disciplines.

Astronomy should take a more unified approach to data management operations (including archives).

1.

Adopt conventions from other disciplines (e.g. biomedical)

1. **Modular:** composed of functional components with well-specified interfaces.
2. **Community-driven:** created by many groups to foster a diversity of ideas.
3. **Open:** developed under open-source licenses that enable extensibility and reuse, with users able to add custom, proprietary modules as needed.
4. **Standards-based:** consistent with standards developed by coalitions such as the ~~Global Alliance for Genomics and Health~~ IVOA.



End-to-end design principle

GitHub

specutils
An astropy package for spectroscopy

astropy
A Community Python Library for Astronomy

https://medium.com/@rgrossman1/a-proposed-end-to-end-principle-for-data-commons-5872f2fa8a47 • https://doi.org/10.1016/j.tig.2018.12.006

Astronomy should adopt a cloud-hosted data commons model for storing data.

2.

*Data commons co-locate **data** with **cloud computing** infrastructure and commonly used **software services, tools & apps** for managing, analyzing and sharing data to create an **interoperable resource** for the research community.*

Robert L. Grossman, Allison Heath, Mark Murphy, Maria Patterson and Walt Wells, A Case for Data Commons Towards Data Science as a Service, IEEE Computing in Science and Engineer, 2016. Source of image: The CDIS, GDC, & OCC data commons infrastructure at a University of Chicago data center.

Cloud + open infrastructure as a science accelerator

- Major projects/facilities should agree to co-locate data (e.g., pixels & catalogs) in the commercial/public cloud.
- Development and sharing of machine-readable infrastructure templates would enable significantly easier **infrastructure reuse between projects** (including older facilities) and between research disciplines (e.g. astronomy and earth science).
- Commoditization of low level infrastructure would allow individual projects to focus on the unique aspects of their mission (e.g. instrument calibration, custom interfaces, user support).
- Adoption of public cloud would make academia and industry more aligned, thereby enabling more routine technology transfer (from industry) and easier hiring.
- Open source infrastructure raises the profile and prestige of mission-critical work.

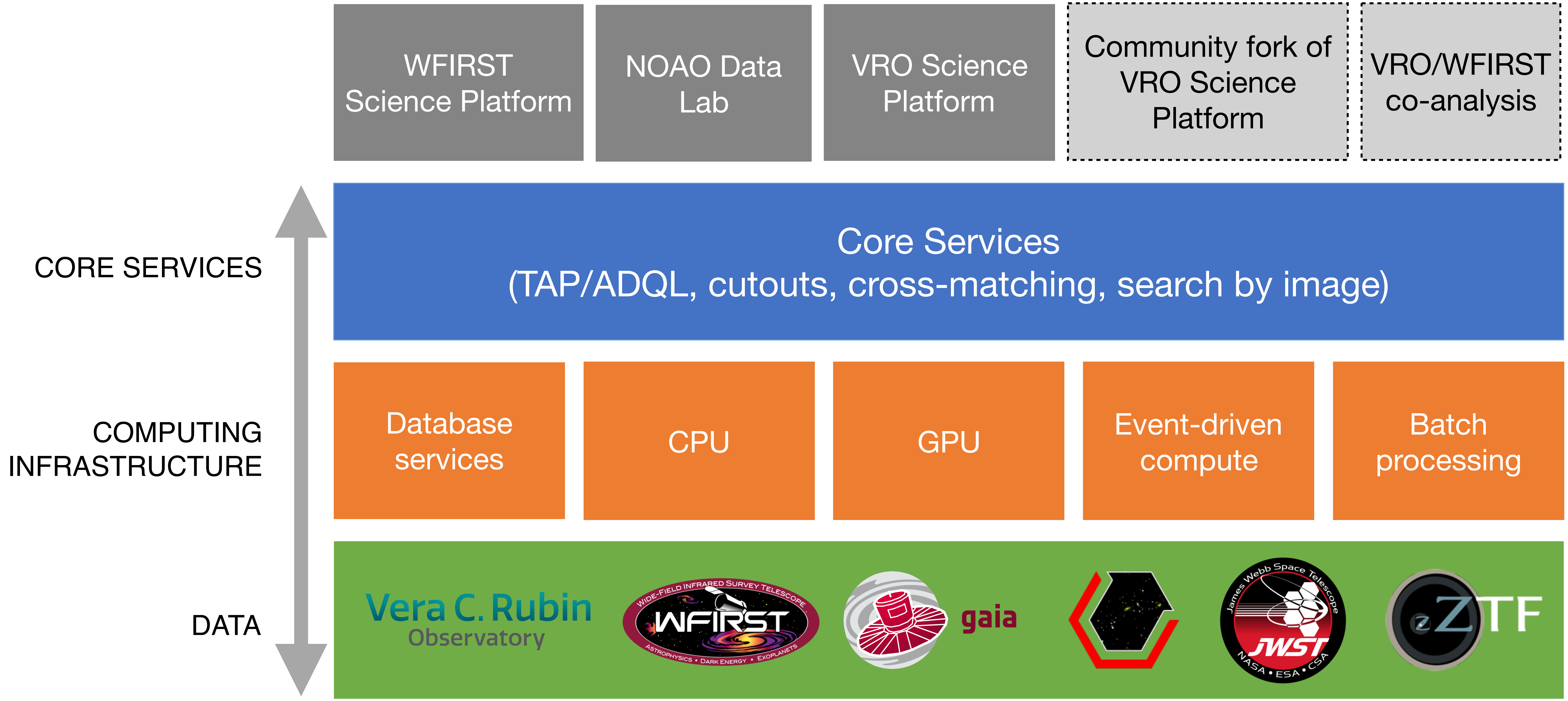
* https://medium.com/descarteslabs-team/thunder-from-the-cloud-40-000-cores-running-in-concert-on-aws-bf1610679978

Cloud + open infrastructure as a science accelerator (cont.)

- Public data staged in the cloud combined with open source infrastructure enables **permissionless innovation** by anyone in the global astronomical community including:
  - Ability to ‘rent’ vast computational resources by the the hour, minute, or second*.
  - Gain access to the latest hardware (e.g. GPUs for deep learning).
  - Easily leverage new computational paradigms in astronomy (e.g. serverless).
  - Vastly reduces barriers (financial & logistical) for those wishing to carry out joint analyses of key datasets (e.g. VRO/LSST + WFIRST).
  - Enable the community to build custom science interfaces.

* https://medium.com/descarteslabs-team/thunder-from-the-cloud-40-000-cores-running-in-concert-on-aws-bf1610679978

Astronomy data commons



*…co-locate **data** with **cloud computing** infrastructure and commonly used **software services, tools & apps** for managing, analyzing and sharing data to create an **interoperable resource**…*

Unification ≠ Centralization

3.

True centralization is a risk to science

- Some consolidation especially in a public cloud-hosted ***data commons***, for low-level data management functions is a good idea.
- **Archives preserve expertise**: true centralization of archives puts science at major risk through loss of scientific expertise and domain knowledge.
- Experts housed at archives act as "product managers” interacting with users, and designing custom interfaces to enable their science (e.g. ExoMAST).
- *Some* duplication of effort is good. Existing archives in friendly competition, leading to increased innovation.
- Data calibration pipelines (and resulting products) are **major intellectual contributions** in their own right and this expertise often resides with a small number of individuals close to the mission/instrument team.

Conclusions & potential actions

Conclusions & potential actions

- Fund open infrastructure and their collaborations (e.g. Jupyter, Pangeo, Astropy).
- Fund cross-collaboration between existing centers (e.g. NASA ADCAR overguides).
- Enact existing federal mandates for open source and open data.
- Adopt an implementer-led process for developing open infrastructure.
- Engage with commercial cloud vendors to secure best possible pricing for cloud usage and egress waivers.

Resources

- Astronomy should be in the clouds: https://arxiv.org/abs/1907.06320
- Elevating the Role of Software as a Product of the Research Enterprise: https://arxiv.org/abs/1907.06981
- Public dataset programs: Amazon Web Services • Google Cloud • Microsoft Azure
- Data Lakes, Clouds, and Commons: A Review of Platforms for Analyzing and Sharing Genomic Data https://doi.org/10.1016/j.tig.2018.12.006
- Sustaining Community-Driven Software for Astronomy in the 2020s: https://ui.adsabs.harvard.edu/abs/2019BAAS...51g.180T/abstract

Appendix

What do you see as the future of archives?

- Reliable, trusted sources of data and expertise.
- Facilitators of server-side analysis and critical infrastructure for joint analysis of peta-scale datasets and enabling any astronomer to work with them.
- Places to collaborate with peers.
- Heavily invested in open source, reusable infrastructure components.
- Taking a more unified, cross-mission approach to data management such as sharing Infrastructure as Code components to facilitate individual applications like Science Platforms.
- Support for long term storage of mission data.
- Making heavy use of the public/commercial cloud.

What are the challenges to enabling interoperability between archives?

- General siloing challenges (color of money). No funded mandate to collaborate.
- Infrastructure lifecycle/refresh cycle of missions (hard to get money to update old systems).
- Limited adoption of open source and open development ‘mentality’ of sharing technologies between centers.
- Use of physical (rather than virtual/cloud-based) data centers.

How can we facilitate interoperability between NASA and NSF-funded archives.

- Fund explicit collaborations between them (e.g. shared infrastructure projects, NASA overguides).
- Fully implement existing federal mandates for open source software.
- Require existing archives to development and *share* machine-readable infrastructure templates for their missions.

What are the advantages/disadvantages of a central archive?

Advantages

- Potential economies of scale through consolidation.
- Co-location of data permits more performant (and cheaper) cross-mission analyses.
- Potential increased technology synergies (if done right) between missions.
- More routine sharing of expertise between missions.
- Democratization of technology/collaborative model with community if done right (Infrastructure as Code technologies + public cloud).
- More seamless collaboration with industry, if implemented using industry-standard technologies such as the public cloud.

What are the advantages/disadvantages of a central archive?

Disadvantages

- Loss of expertise with missions and therefore science opportunities.
- Forcing heterogeneous datasets into common data models potentially risks losing nuance of individual mission data.
- Loss of innovation through ‘competition' between centers.
- At extreme, if centralization included co-locating all staff then loss of key critical staff.
- Introducing a single point of failure into critical infrastructure.
- Fundamentally puts scientific productivity of archives at major risk.

How should we assure long-term access to data?

- Create organizations with permanent archiving as part of their mission.
- Keep some money aside (~5%) from construction/operations for long-term archiving (see Szalay & Barish).
- Identify organizations that can help astronomy achieve this mission (e.g. university libraries, cloud vendors).
- Ensure that negotiated deals with public cloud vendors include provisions for long term access *and* an exist clause should they stop participating in public cloud activities.

What information do we need to maintain from projects?

- People:
  - Context.
  - Nuance.
  - Expertise with data.
- Raw bits, catalogs, software, documentation.

Should we archive simulations?

- Should we store and serve simulations? YES
- Should we archive simulations in perpetuity? Probably not.

How do we enable astronomers to process some/all of the relevant data on local machines?

Depends what we mean by ‘local’… Processing all VRO/LSST pixels on my laptop seems unrealistic. However, assuming we mean ‘being able to use the mission tools locally then:

- Fund high-quality research software. Insist that it’s open source.
- Adopt a tiered approach for infrastructure i.e. allow astronomers to use similar/same tools that are used at scale, just locally.
- Data management/data processing is pleasingly parallel, that is, a local machine can be one processing ‘node’ out of many. Frameworks such as Dask potentially allow for a hybrid data processing where command/control is on local machine but large clusters are being accessed for bulk of computing.

What should the relationship be between astronomical archiving and other national-scale archiving challenges? e.g. in geophysics, biology, physics, etc.

- Developing common business models, especially negotiated rates with e.g. commercial cloud vendors for storage and egress (distribution).
- Develop common ‘people operation’ models, i.e. what is the right balance between a core workforce component and key staff employed at distributed centers.
- Develop and fund common community models (e.g. see Pangeo right now in Earth science).