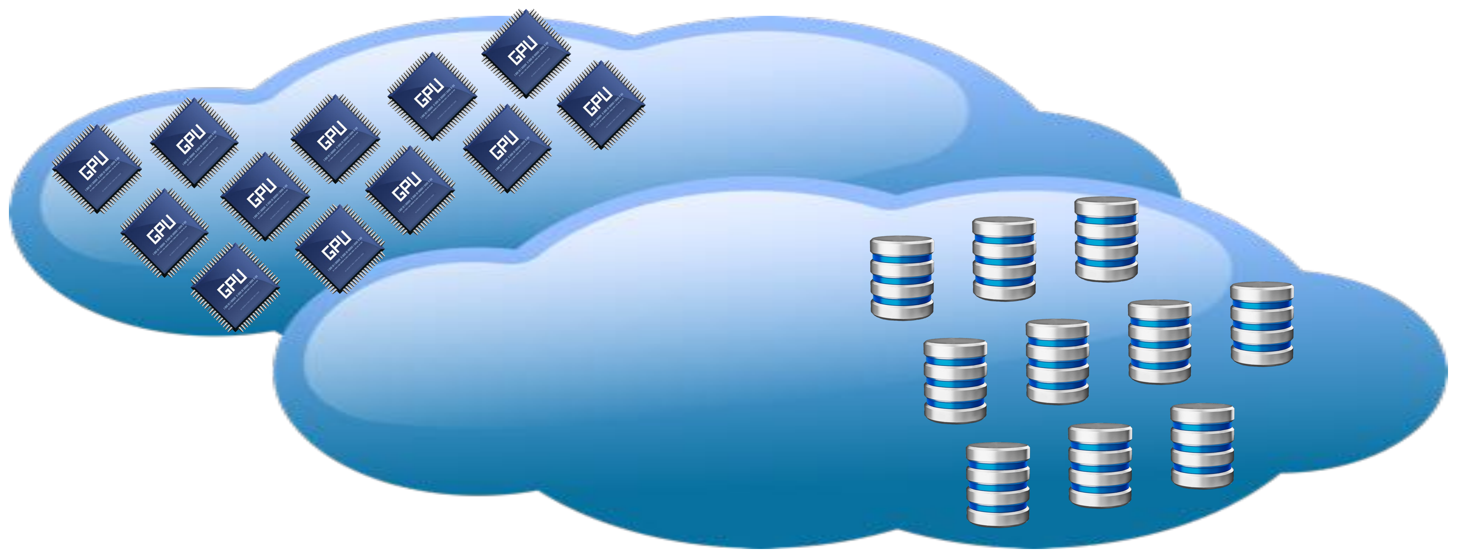
1979

2017

256 Gigabytes

≤ 2050



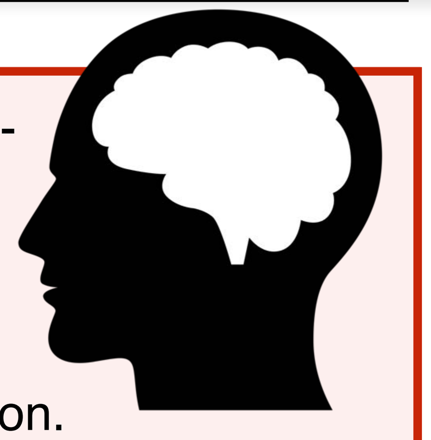
Archival Data and Computational Power in Planetary Astronomy: Lessons Learned 1979–2016 and a Vision for 2020–2050

Mark R. Showalter, Matthew S. Tiscareno, and Robert S. French, SETI Institute, Mountain View, CA

A Future Data Analysis Scenario

1. **A scientist has a hypothesis and wishes to test it.**
2. **The relevant data sets are already on the ground, having been obtained by NASA spacecraft and Earth-based facilities over a time span of many decades.**
3. **The scientist quickly codes up a data processing algorithm, building upon a large library of tested, documented, open-source building blocks contributed by her fellow scientists.**
4. **She prepares a virtual machine (VM) that contains the program and any relevant support files, and submits the VM to a cloud computing service.**
5. **An arbitrary number of instances of the VM execute in the cloud, assimilate the results, and deliver them back to her quickly.**
6. **Using visualization tools based on other open-source algorithms, she explores the results and reaches her conclusions.**
7. **She publishes her results as a journal article, plus supplemental data and VMs. Results are immediately available to the rest of the community.**

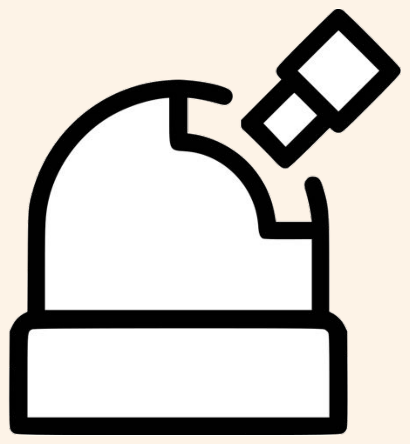
1.



- Scientific research will always be a human-centered enterprise.
  - HAL will not help us.
- Our goal should be to minimize obstacles along the path from hypothesis to conclusion.

2.

- NASA archiving must be comprehensive, including data from Earth-based telescopes, lab experiments and published theoretical models.
- All data sets must include raw and calibrated data, plus a rich and complete set of metadata.
- **Approaches:**
  - Standardize planetary metadata (“SPICE 2050”?).
  - Provide pointing geometry at pixel-scale precision or finer.
    - Hardware solution (best): Develop better spacecraft tracking and telemetry.
    - Software solution (if needed): Apply geometric reconstruction to images via pattern recognition.
  - Find a way to preserve mission software so it will continue to run forever.
    - Currently, the lifetime of any program is ≲ 10 years.
  - Archive easy-to-use derived data products (maps, etc.), when backed up by scientific publications.
  - Employ “scraping” of the literature, press release images and captions, and supplementary data from journals, to keep the archive complete.



3.

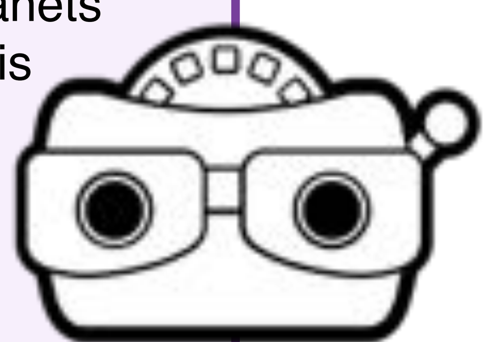
- Widespread availability of open-source software will prevent NASA-funded scientists from constantly needing to “re-invent the wheel”.
  - Eventually, the accumulated tools will allow most scientists to focus on research, not programming.
- **Approaches:**
  - NASA should continue to encourage the sharing of open source software by its scientists.
  - NASA should reward (and fund) scientists for well-designed, well-documented libraries.
    - Scientists have no incentive to share their tools if publications are the only metric of success.
  - We should consider identifying a limited set of “recommended” programming languages.
  - Discourage or perhaps even forbid the use of proprietary languages (e.g., IDL and Matlab).
    - Long-term costs to NASA are very high.
    - Proprietary languages cannot be used in VMs.
  - Find ways to harness the latest developments in computing, e.g., neural nets, GPU programming, and whatever else may be coming in the decades ahead.

7.

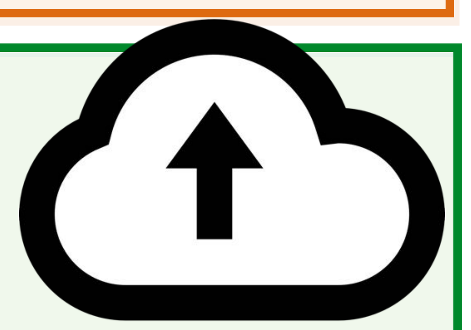
- Immediate, automatic archiving can speed up the pace of NASA discoveries, because each new result will be instantly available to the entire community.
- **Approach**:
  - Develop automated tools and templates that ensure every data set is accompanied by suitable documentation and metadata.
  - Employ intelligent “scraping” of new publications.

6.

- Data visualization is one area where planetary scientists can benefit from development going on at for-profit companies.
  - Example: “Google Earth” has now been used for “Google Mars”, “Google Moon” and “Google Pluto”.
- Many companies are currently investing in Virtual Reality (VR).
- **Approaches:**
  - We should encourage for-profit companies to support planets within their VR tools, noting the educational value that this support will provide to the broader public.
  - NASA should encourage and support development of open-source VR tools for the exploration and visualization of planetary data.

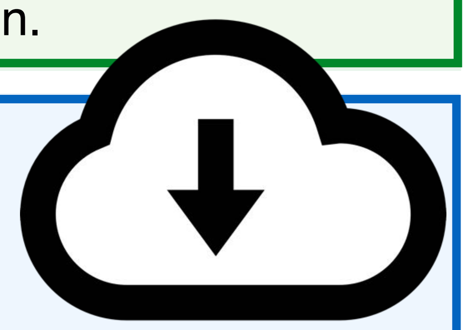


4.



- VMs are self-contained entities holding the OS, software and support files.
  - They can be shared, executed in the cloud (in arbitrary numbers), and also used locally.
- VMs are only useful if for processing can be performed without human intervention.
- **Approaches**:
  - Ensure that all metadata is accurate enough for “blind” data processing.
  - Encourage the use of VMs as a mechanism to distribute and document analysis procedures and models.
  - Spearhead efforts to ensure that VMs can be archived and reused decades after their creation.

5.



- Cloud computing has the potential to revolutionize planetary data analysis and modeling.
  - Highly parallel problems (e.g., process every image that…) can execute very quickly.
- **Approaches**:
  - Permanently store all data and metadata within cloud services where VMs can have direct access.
  - Encourage and fund the use of cloud computing; spearhead efforts to reduce its cost.