MAST: The Mikulski Archive for Space Telescopes

Richard L. White

Space Telescope Science Institute

2015 April 1, NRC Space Science Week/CBPSS

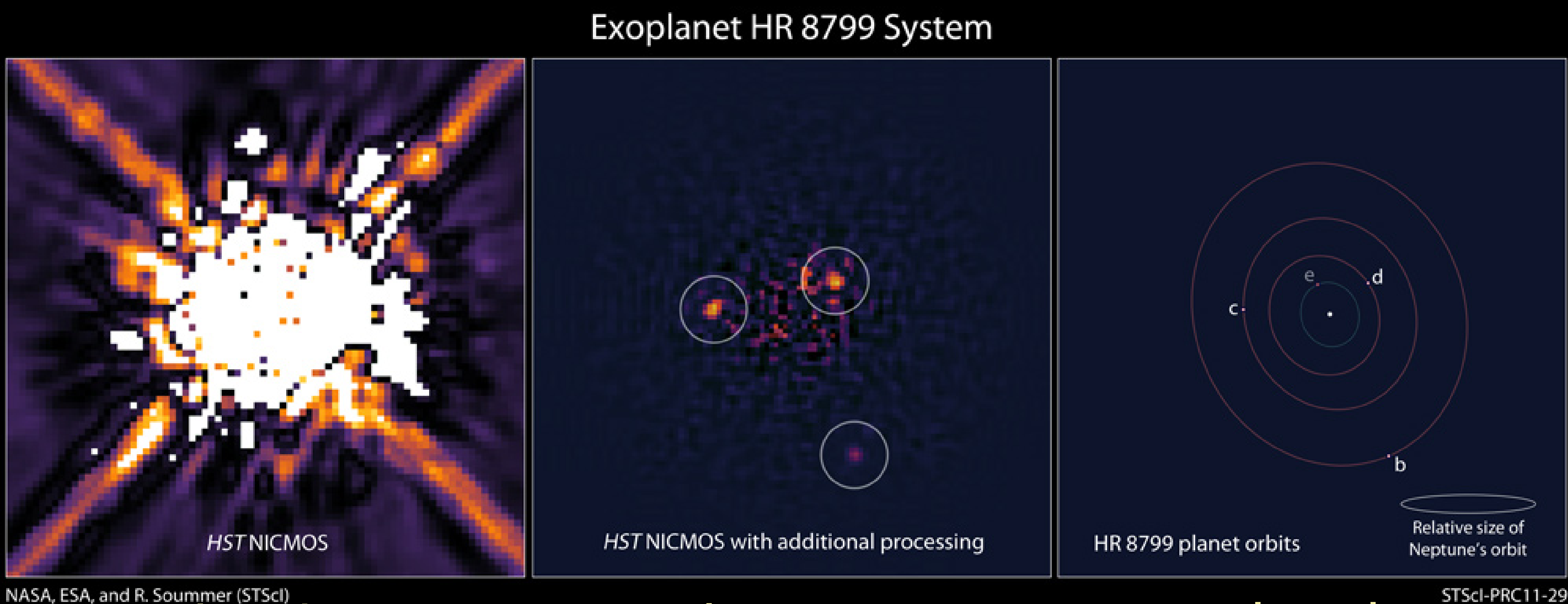
A model for open access

- The NASA astrophysics data archives are a model for open access to data.
  – See white paper by S. Murray & M. Postman on public access to space astronomy data.
- MAST has provided open access to science data for more than 25 years.
  – Hubble changed the paradigm: provide calibrated, science-ready data to the entire community after a proprietary period of 1 year (or less).
- Archival research greatly enhances the science at a small fraction of the total mission cost.

Lessons learned

- Use standard data formats
- Capture complete & accurate metadata
  - Create a data model to enable cross-mission research
- Engage science teams for expertise & support
  - Motivation for the team: increased scientific impact
  - Key for STScI: many scientific staff are advanced users of the data
- Generate high-level, science-ready data products
  - Build expertise into pipelines
  - Community-contributed products are also important
- Provide powerful, web-based tools for data mining

HR8799 b,c,d imaged by HST in 1998



planet b:
83,000x fainter than star
at 1.72 arcsec

planet c:
36,000x fainter than star
at 0.96 arcsec

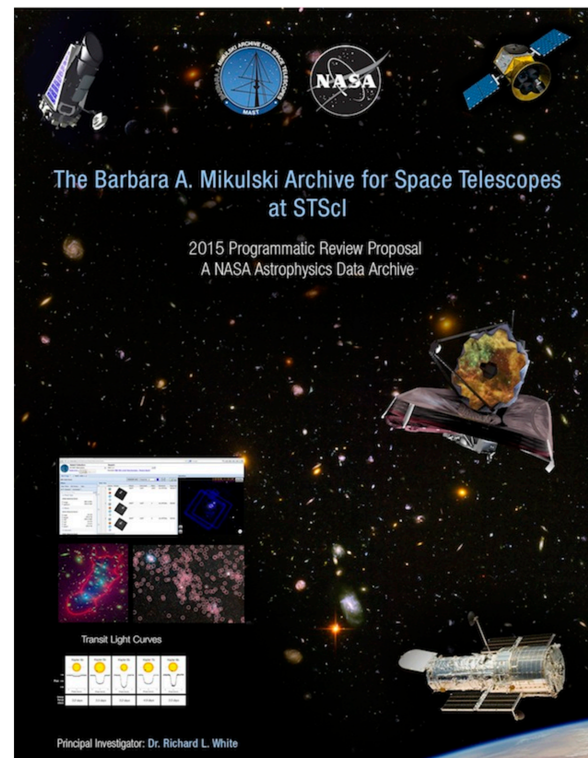
planet d:
33,000x fainter than star
at 0.60 arcsec

These results were made possible by post-processing speckle subtraction and achieve over an order of magnitude contrast improvement over the state of the art when the data was taken in 1998.

Soummer et al. 2011, Pueyo et al. 2014

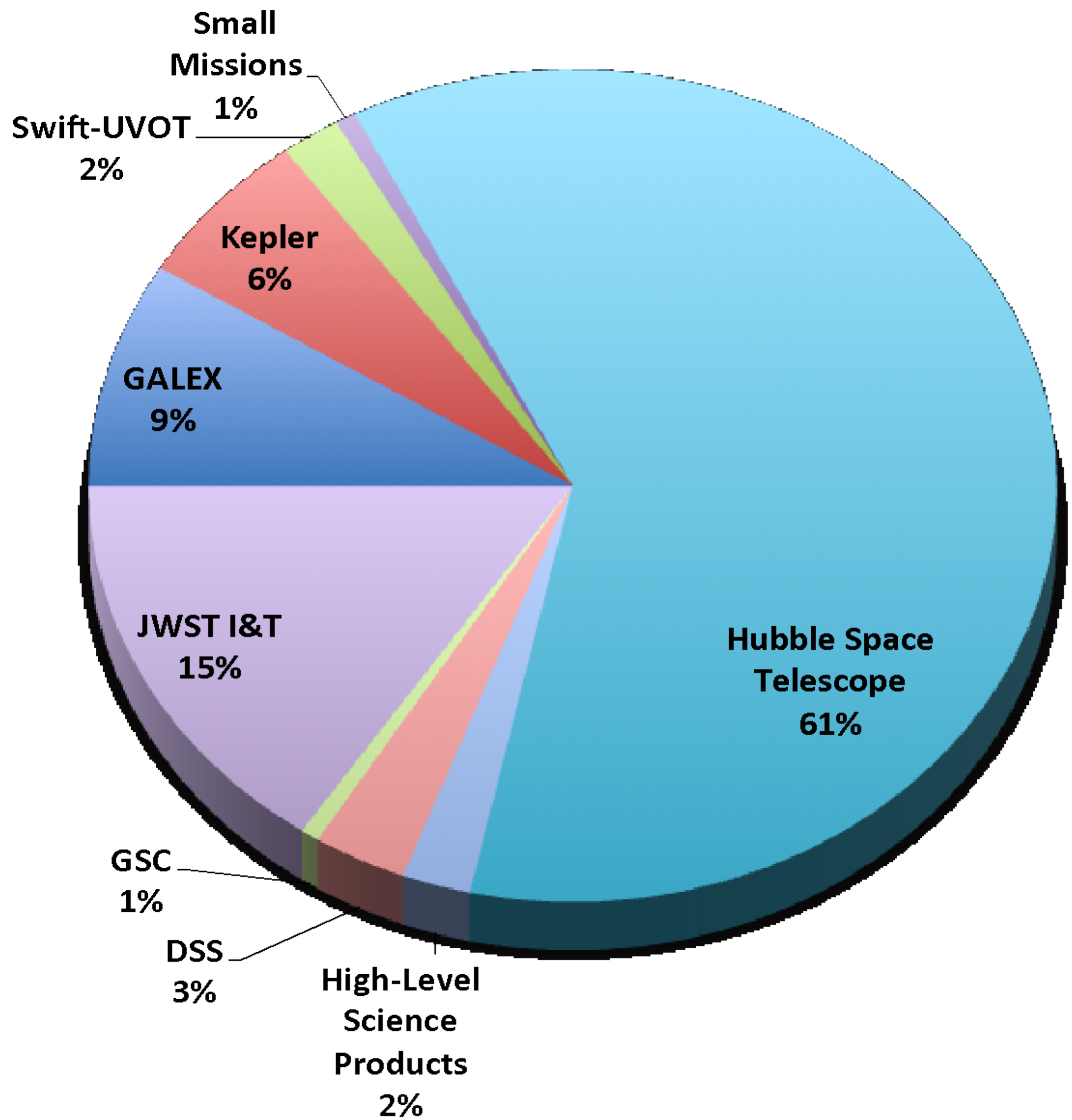
Introduction to MAST

- MAST: a NASA astrophysics data archive center
  - Archive established with HST launch in 1990
  - Multi-mission since addition of IUE in 1998
  - 4 active missions including Hubble, Kepler
  - Many legacy missions: GALEX, IUE, FUSE, …
  - Future: TESS, JWST, WFIRST, …

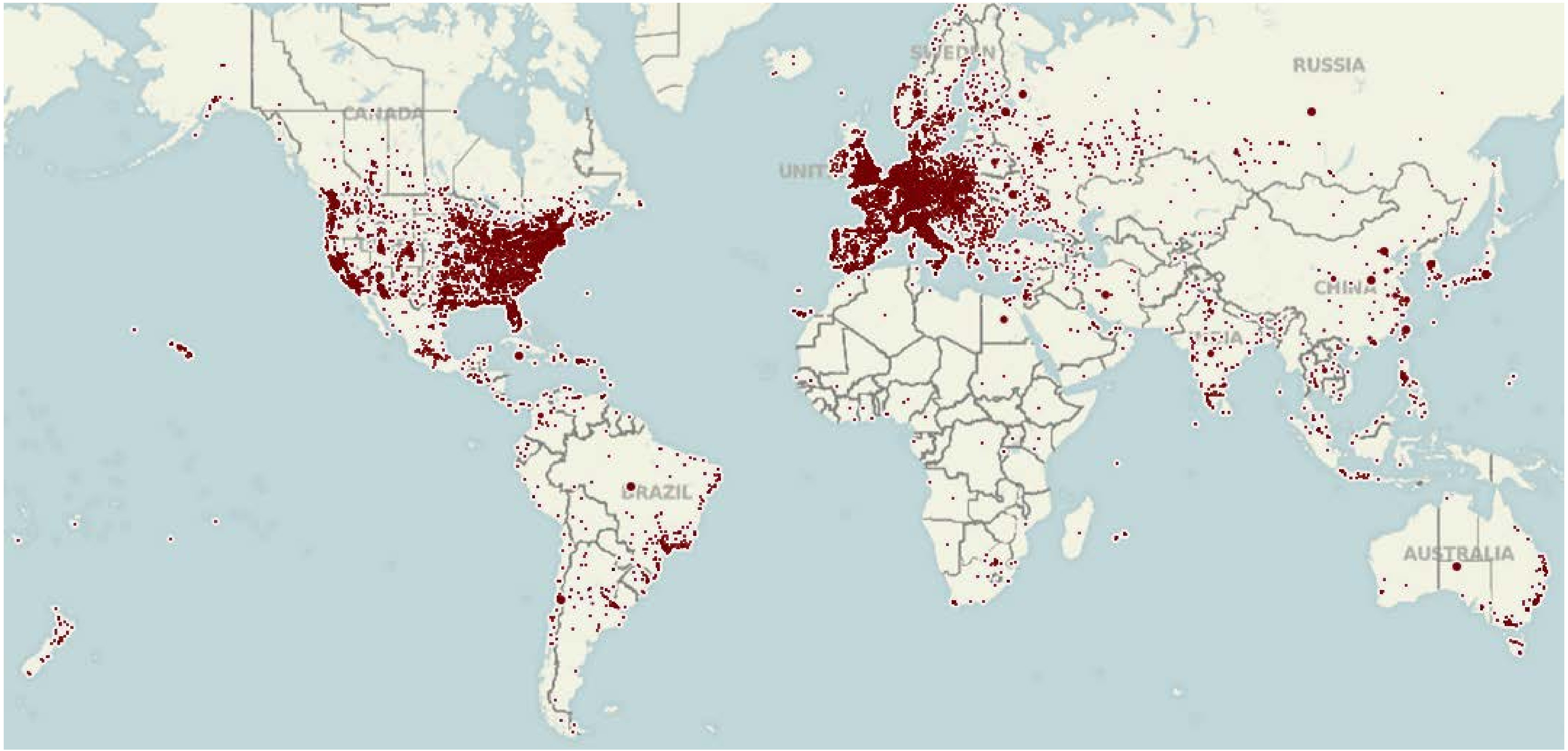


MAST Holdings

- 21 missions/projects (current and planned)
- Multiple wavebands from far ultraviolet to infrared



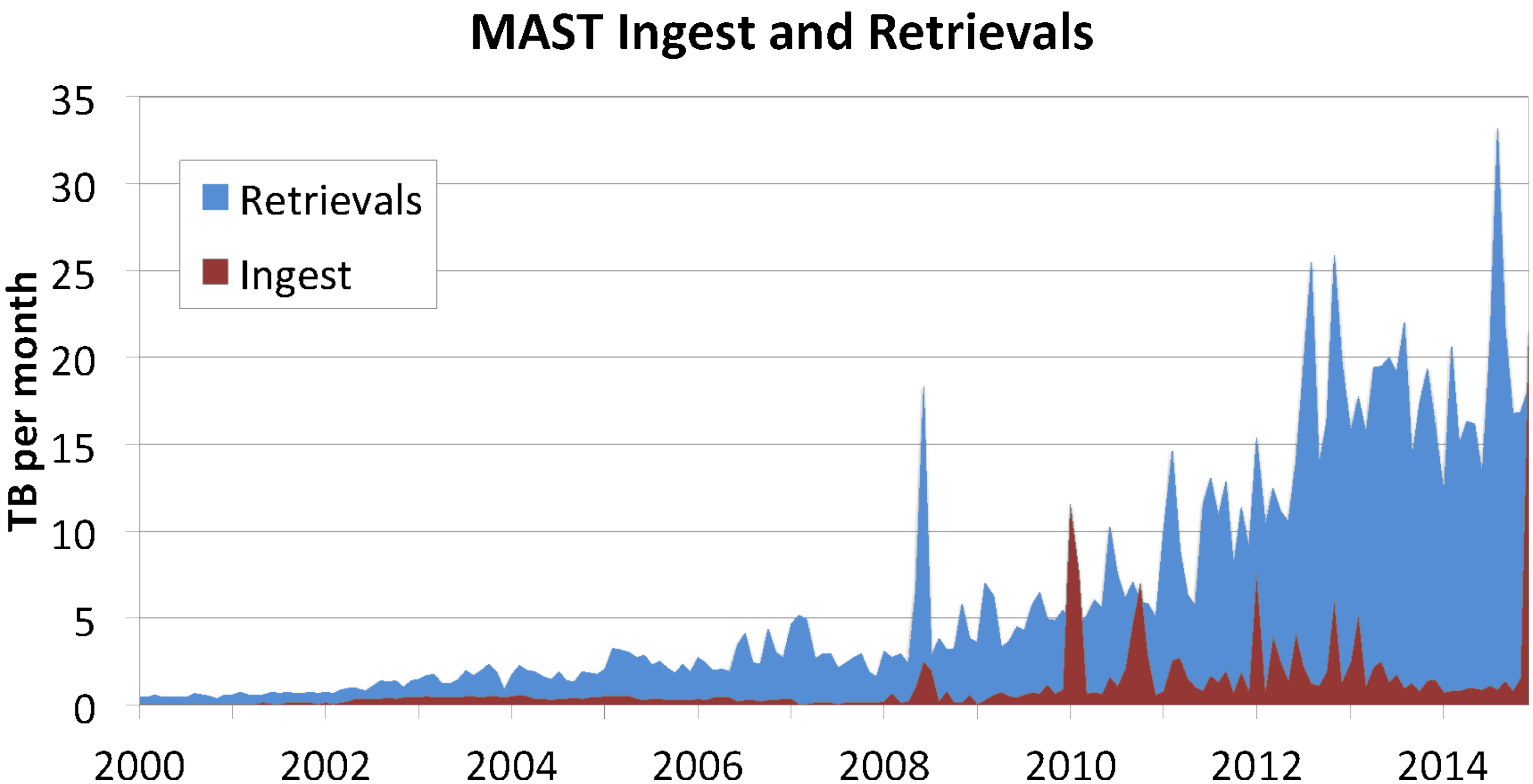
MAST Searches



2.2 million searches per month in 2014

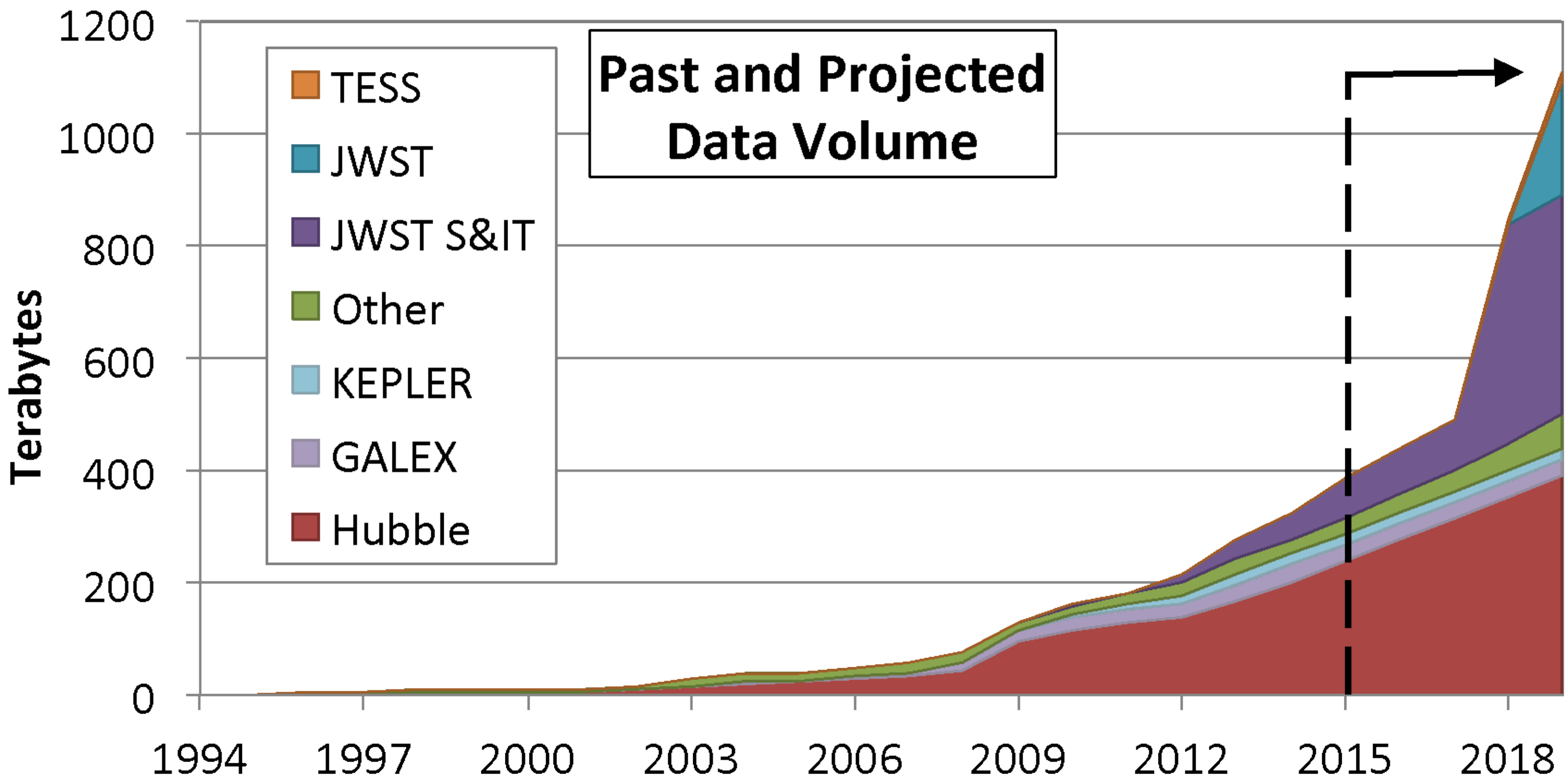
> 12,000 registered archive users

MAST Data Distribution



The archive distributes 7–10x more data than is archived each month.
The average for 2014 was 18.5 TB/month.

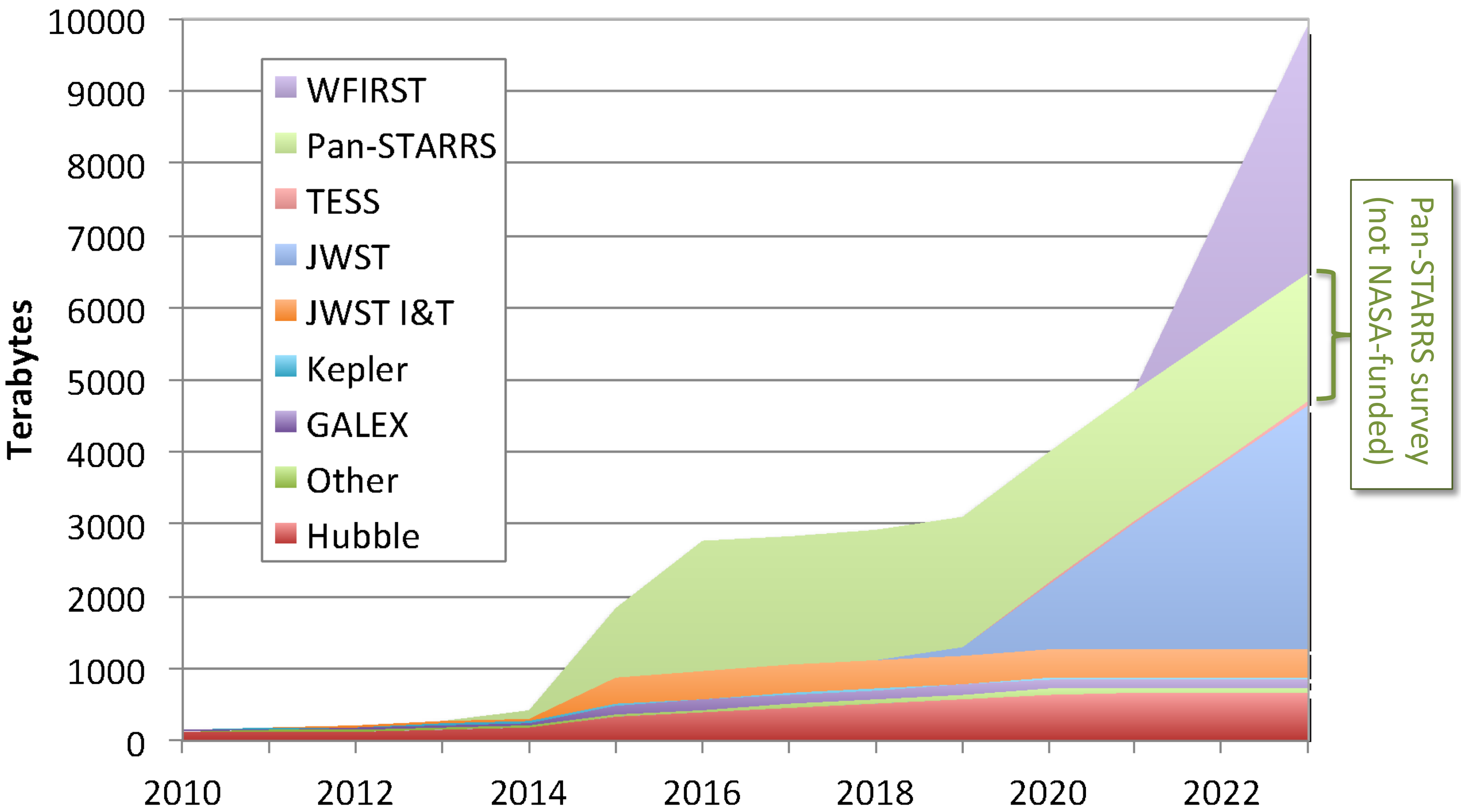
MAST Data Growth



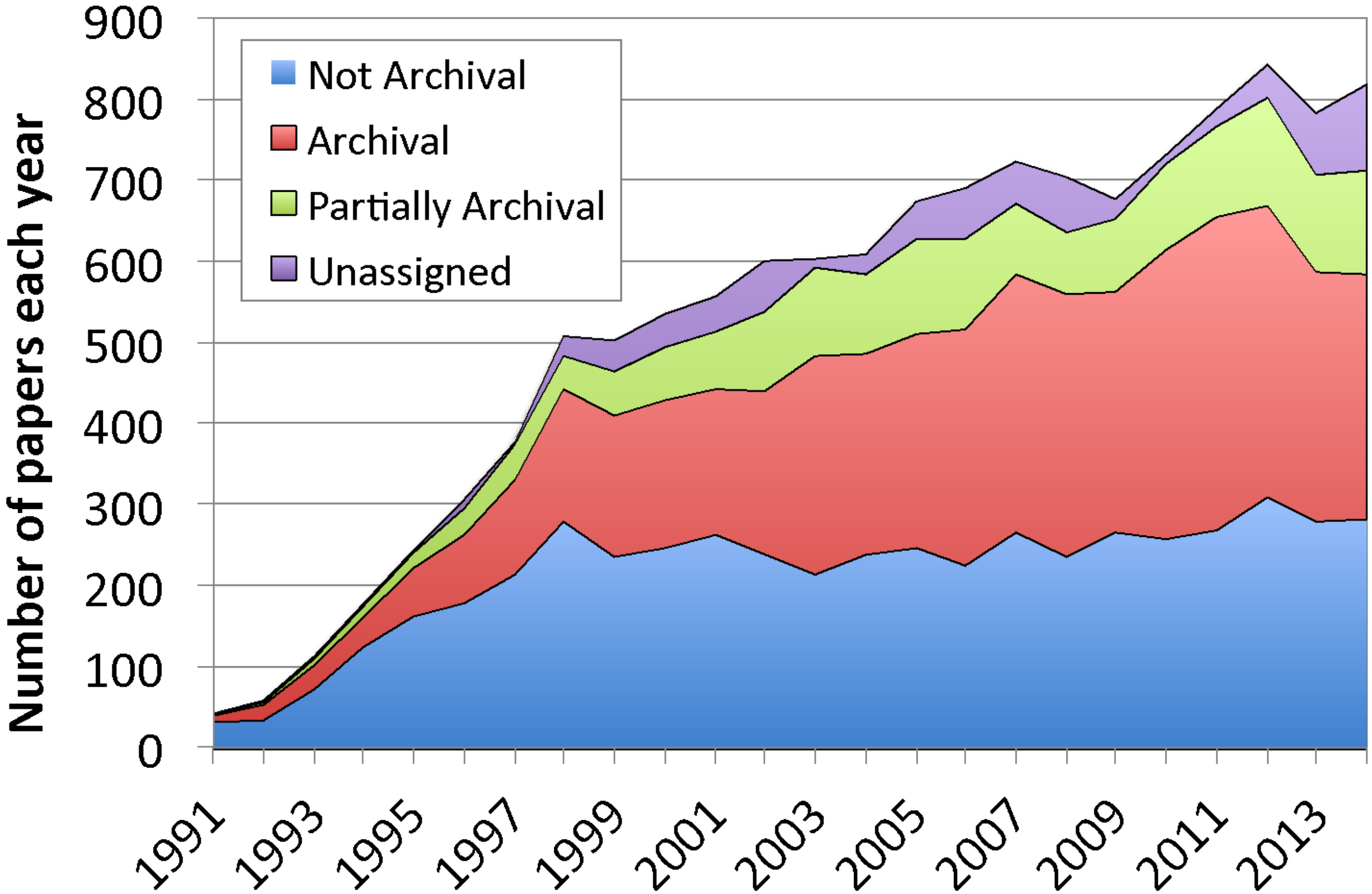
320 TB of data products

MAST ingest rate 2.7 TB / month (2014)

Long-term MAST Data Growth



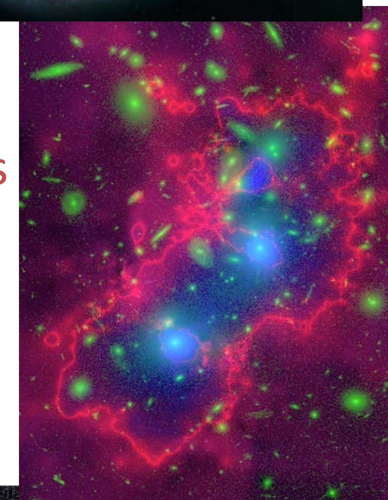
HST Publication Rate



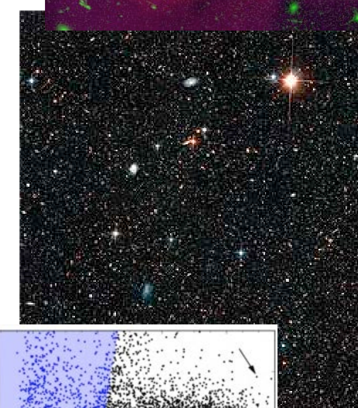
Over the last 4 years, 1200 papers were published each year using MAST data.

The publication rate for totally archival Hubble papers has exceeded the non-archival (GO/PI) publication rate every year since 2003.

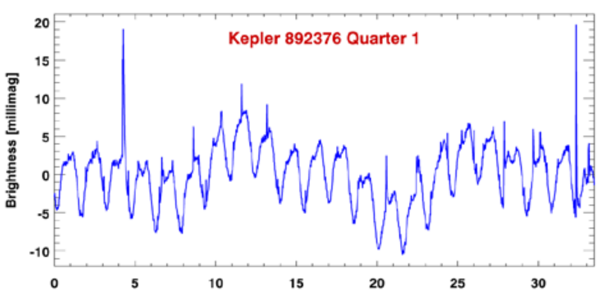
MAST data are diverse

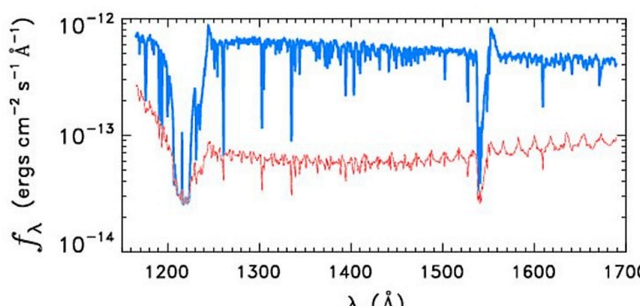


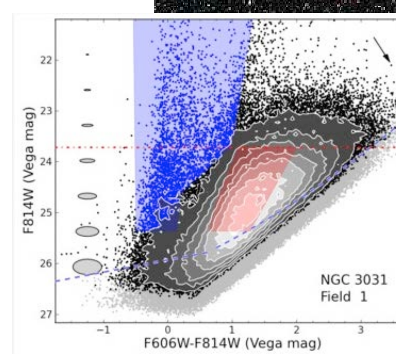
- MAST data are diverse:
  - in date of observations: from 1972 to now
  - in data types: images, spectra, light curves, catalogs, models
  - in scale: from Pan-STARRS (2 PB images + 100 TB database) to small shuttle-based missions to community-contributed projects with just a few files
  - in processing level: from raw data packets delivered directly from spacecraft to science-ready, high-level science products



- Many different missions and instruments
  - Hubble alone: 12 different instruments, 17 varieties of detector, hundreds of instrument modes/filters/etc.







Dealing with data diversity – 1

- Accept diversity as a fact
  - Metadata, telemetry, calibration data, etc., are always going to be specific to the mission/instrument
- Implement diverse storage & access mechanisms
  - Create separate mission databases
  - Supply custom interfaces for advanced mission searches
    - Store mission-specific info in standard configuration files
    - Build interfaces automatically from config data
  - Allow direct database (DB) access by users (CasJobs)
    - Supports large queries using all DB parameters
    - User work areas for uploading and creating user DBs

Pros:

✓ Preserves all details of mission-specific data

Cons:

- Significant manual effort required for new missions & projects
- Mission-specific user expertise required to use advanced features

Dealing with data diversity – 2

- Create homogeneous views to integrate diverse data
  - Common Archive Observation Model (CAOM)
    - Data model defining common subset of metadata for all missions
  - Virtual Observatory (VO) protocols
    - Common scriptable interfaces for data access
    - Used by users, between archives, and inside the archive
  - MAST Discovery Portal interface
    - Single interface with access to all MAST data
    - Tools for previewing, selecting, analyzing, & downloading data
    - Can also access other archives through VO protocols
- Most current MAST work focuses on this unifying approach

Pros:

✓ New features are usable across missions

✓ Easier for users to learn

Cons:

\- Some mission-specific info not accessible (but still exists)

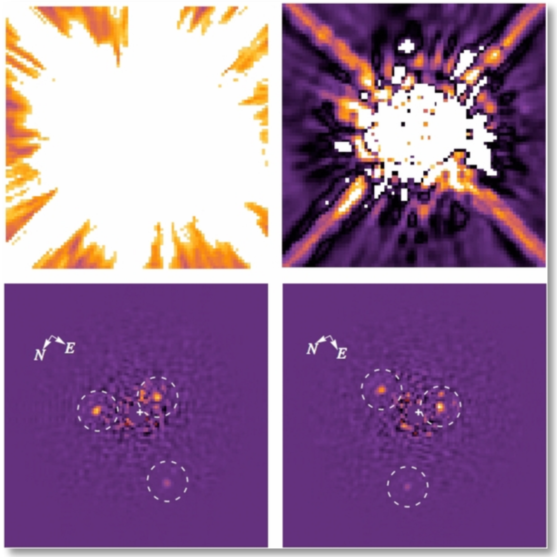
MAST lessons learned: data

- Use standard data formats
  - Astronomy has used FITS for decades
    - Self-describing, open data format for images, tables, etc. with embedded simple metadata description
    - FITS is showing its age (not good for hierarchical data)
  - Expect formats and conventions to evolve
    - HST used 4 different FITS data formats for 4 spectrographs!
- Capture complete & accurate metadata
    - But metadata will evolve too for most missions

MAST lessons learned: data products – 1

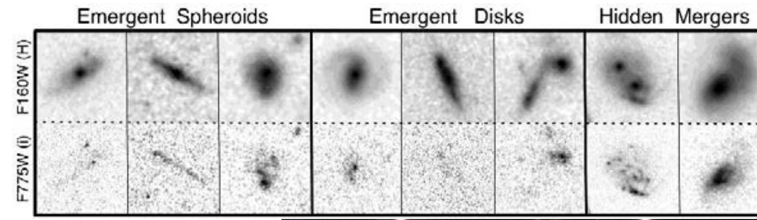
- Generating science-ready, high-level science products (HLSPs) is key to enabling archival science.
  - There is a false savings in delivering only raw data products: the costs of processing data then are incurred many times over by different users in different locations.
  - Archives that deliver only raw data are much less useful – and are much less used – than archives that deliver science-ready products.
- HST data processing pipelines generate calibrated datasets that are science-ready.
  - On-the-fly reprocessing from raw data uses current instrument calibrations
  - Astronomical data is difficult: single-photon counting with extreme demands on calibration & sensitivity
    - If we can do it, you can generate science-ready data too!

MAST lessons learned: data products – 2

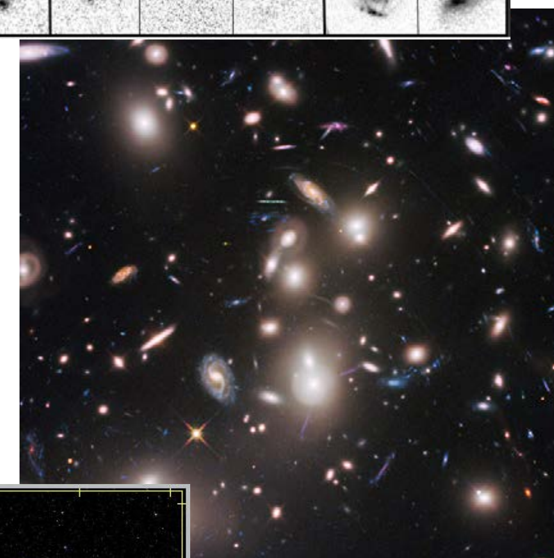


- Community-contributed HLSP are even better
  - Used 10x as much as typical pipeline products
  - Rely on scientific refereeing process to ensure data quality
  - Many advanced in data processing algorithms originate in these projects
    - Improvements are used in advanced MAST pipelines (e.g., Hubble Legacy Archive) to generate higher level products modeled on community projects
- Build as much expertise as possible into sophisticated pipelines & well-documented data products
  - These may be the only products that remain easily usable when community expertise fades after the mission ends

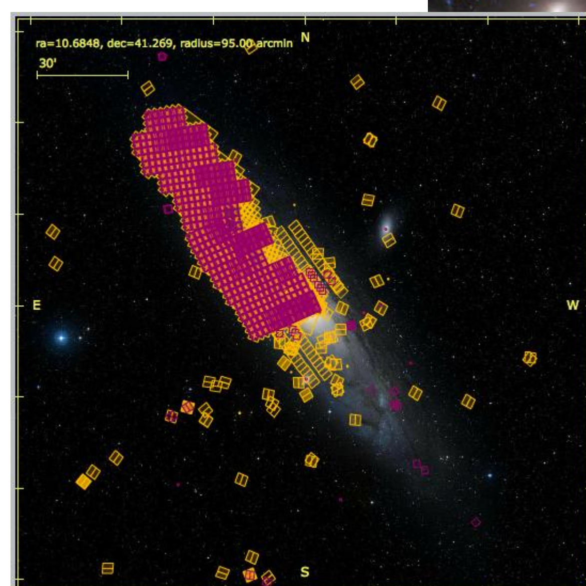
Sample HLSPs: HST Multi-Cycle Treasury Programs



- CANDELS (Faber/Ferguson)
  - 3.1 TB of data
  - ~ 53 TB distributed to 2178 IP addresses



- CLASH (Postman)
  - 0.6 TB of data
  - ~ 5.7 TB distributed to 1637 IP addresses

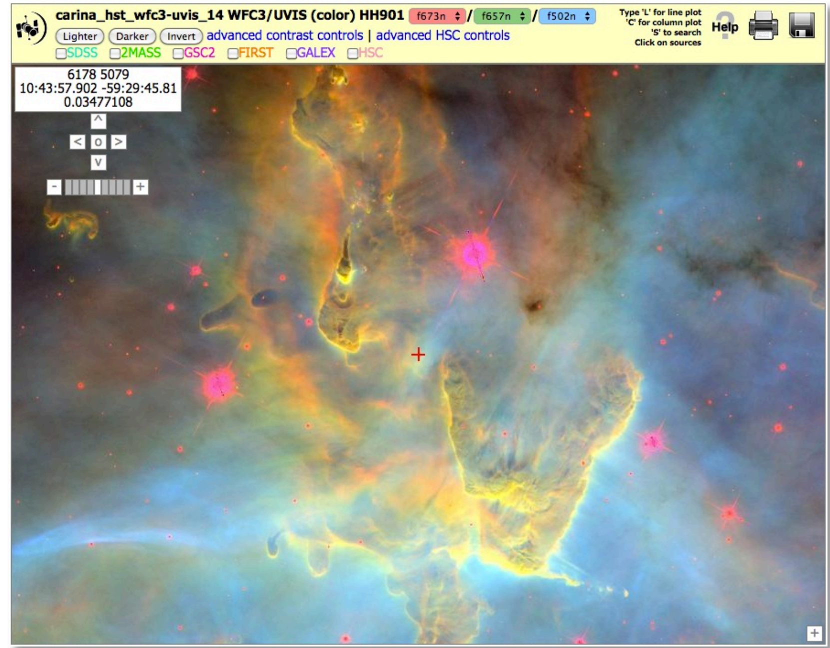


- PHAT (Dalcanton)
  - 2.0 TB of data
  - ~ 8.9 TB distributed to 2485 IP addresses

MAST lessons learned: capturing expertise

- Capture the knowledge of experts on data and instruments
  - Keep data close to science team while mission is active
    - MAST provides support ranging from archive/database specialists to software developers to instrument scientists
    - A key for STScI is that many of our scientific staff are both calibration experts and advanced users of our data who push the archive capabilities and data products
  - Plan for closeout at end of mission

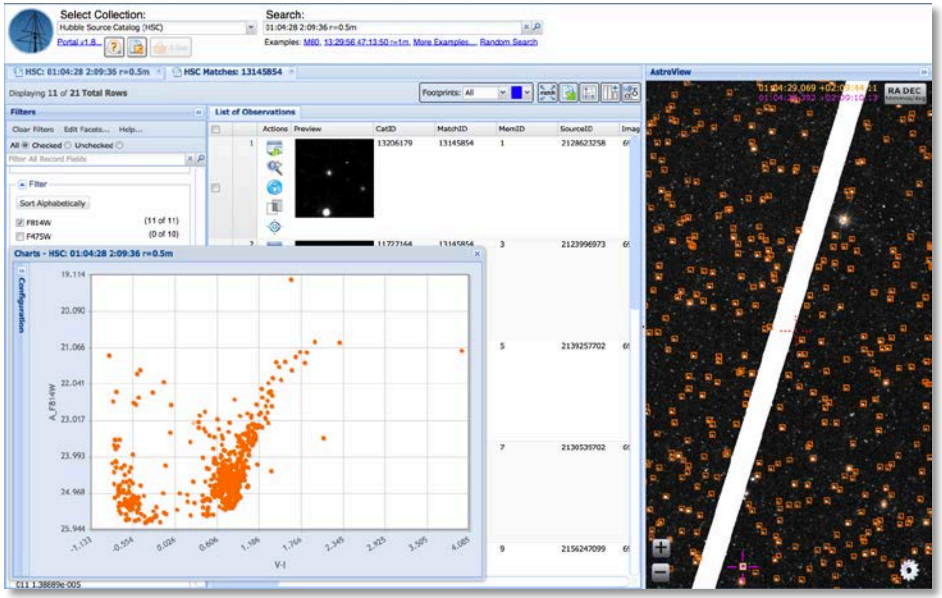
MAST lessons learned: engage the teams



- The best advocates for any data set are the team that designed & built the experiment and collected the data.
  - Teams working on archival projects are also a valuable resource.
  - If you can engage the team, they will produce the higher-level products that are necessary to enhance the archival value of the data.
- Why would the science teams want to expend that effort?
  - We show them that they will increase the scientific impact of their work (and increase the citations for their papers!) by contributing high-level science products. It is good for them and good for science.

MAST lessons learned: user interfaces

- Web interfaces are much preferred to downloaded software
  - Web software is always up-to-date
  - Everybody has a browser
    - Example: old Starview MAST interface (Java application) was used by only 1% of users despite additional functionality
- Challenges:
  - Evolving web technology
  - But HTML5 + Javascript + server-side software is now standard and widely supported
    - See MAST portal as example http://mast.stsci.edu

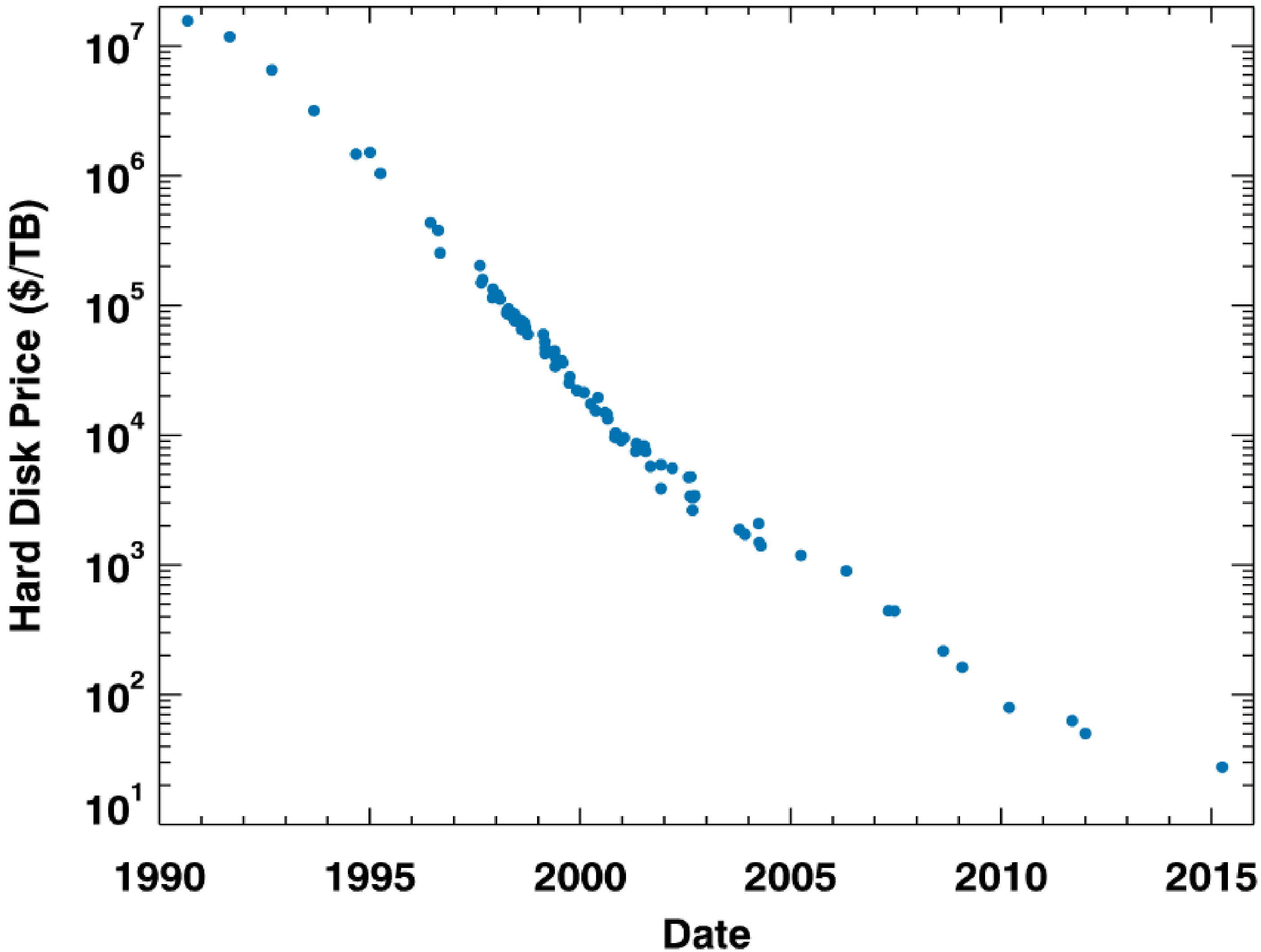


MAST lessons learned: plan for change

- An archive is an evolving collection of data products, interfaces, and access services.
  - It is not simply a fixed collection of data!

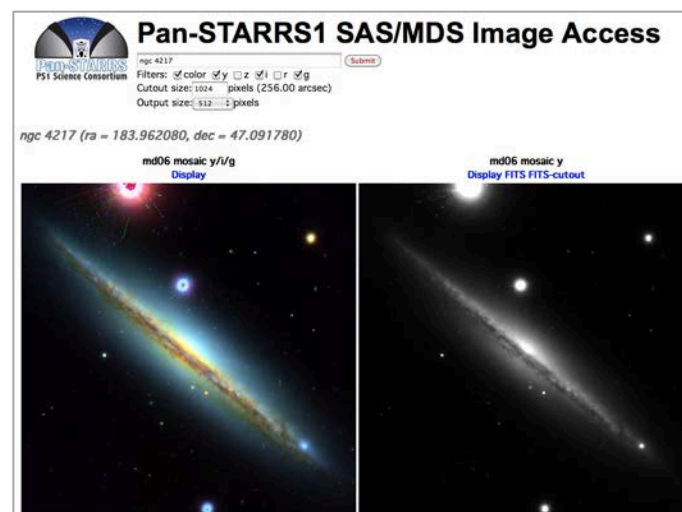
- A long-lived data archive must evolve along with the world and the user community.
  - At Hubble launch in 1990:
    - There was no internet
    - Disk storage cost $20 million per terabyte
    - Operating an up-to-date archive in the midst of rapid change is hard!

Disk storage prices since Hubble launch



Large data products challenge the open access model

- Downloading some datasets is simply not practical
  - Pan-STARRS has 2 petabytes of images along with a 100 terabyte catalog database
  - The GALEX photon database is a 150 TB table with all 1 trillion photons collected by the GALEX ultraviolet sky survey mission
- Essential to provide tools that allow users to query, browse, and mine big data without downloading it
- Allow downloads of selected data and dynamically generated products (image cutouts, catalog subsets, light curves, movies)
  - … but does that satisfy all requirements for open data access?



Summary

- An archive is not a bit bucket – it is a living, evolving science machine.
- Archives can greatly enhance science at a relatively small cost. Once the infrastructure is in place, the incremental cost for new missions is modest.
- Keep the archive close to the scientists, who are needed to support and enhance the data & tools and who push the data to the limit.
- NASA's long-term funding for a network of archive data centers that support astrophysics data from gamma-rays to the microwave background has played a key enabling role.

> Successful research using archival data sets is dependent on the resident expertise and corporate memory that reside at the science centers.
>
> – *Portals to the Universe: The NASA Astronomy Science Centers*, National Academies Press, 2007