Data panel discussion

NRC OIR Meeting

Irvine, 2014 October 12

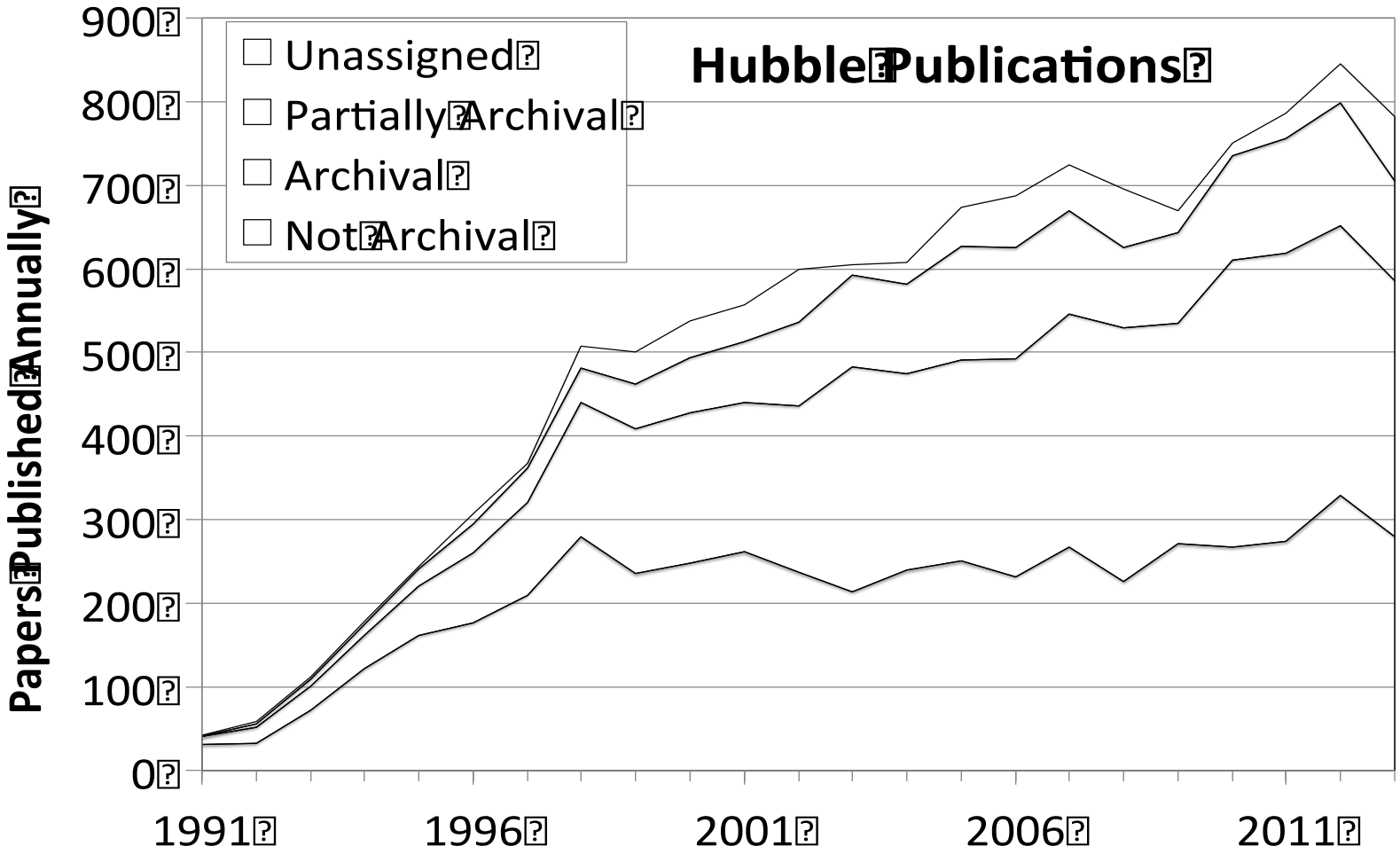
Rick White, STScI

A few comments

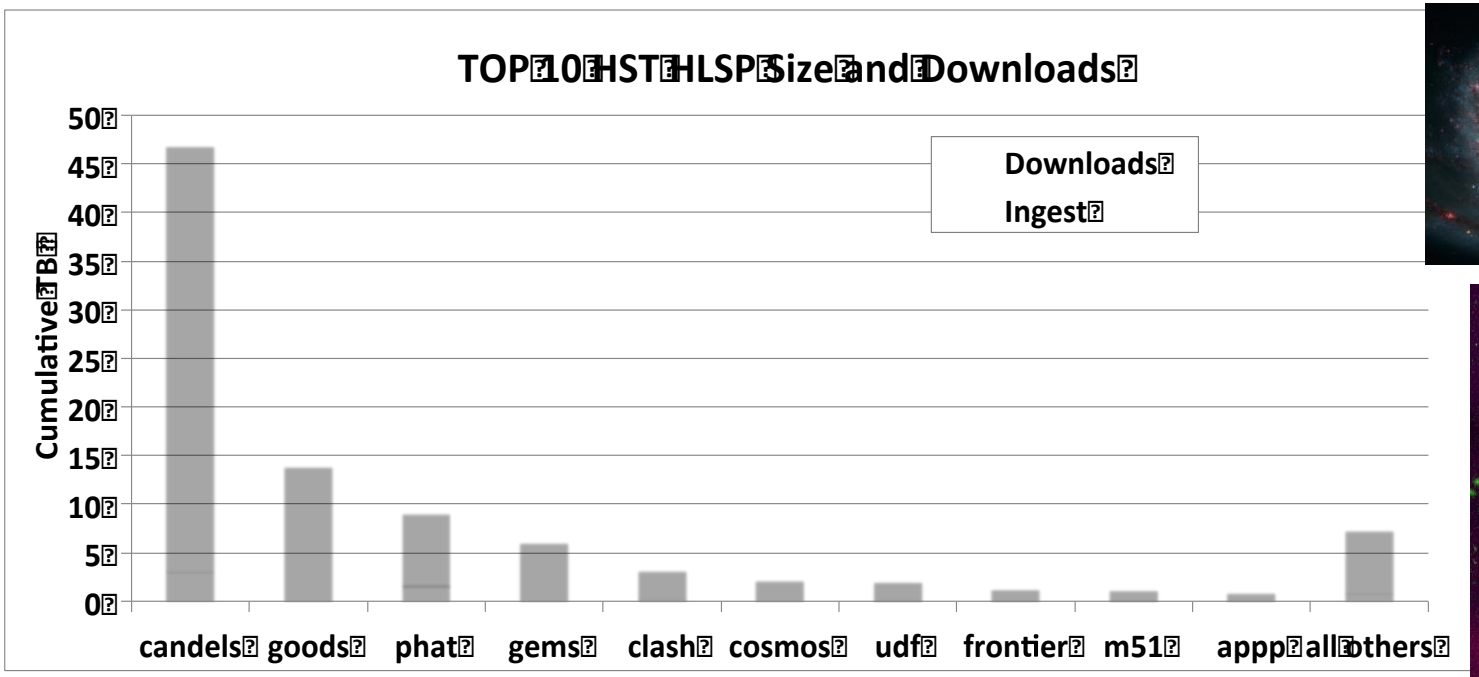
- Archives are an comparatively inexpensive way to increase science output of missions and observatories
  - An archive is not just a collection of data, but is also a collection of services and interfaces to that data
- To prepare for LSST, the ground-based observatories should develop robust archives with science-ready data products & services
  - A unique contribution from existing observatories is past observations of newly discovered variable objects
  - Science-ready: Calibrated images, cutout services, catalogs
- Incorporating community-contributed high-level products is important to maximize archival science
  - High-level science products are much more heavily used than low-level (e.g., raw) data products

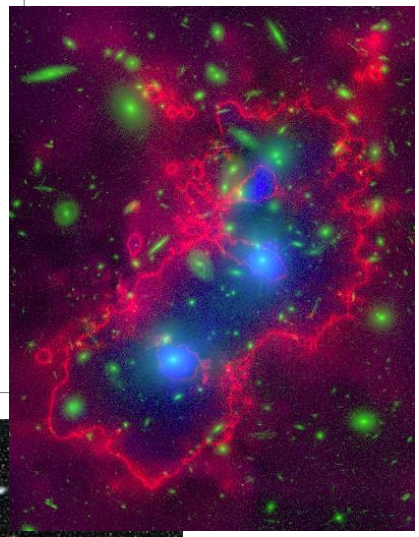
Archives increase science

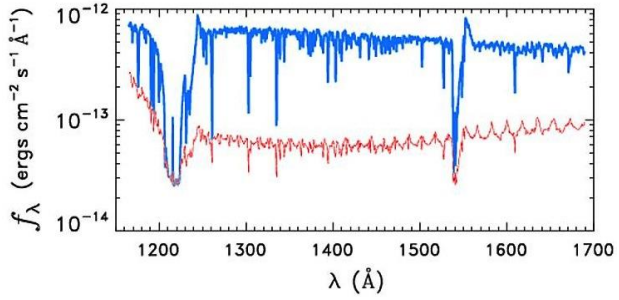
> 1100 papers published in 2013 using data from MAST

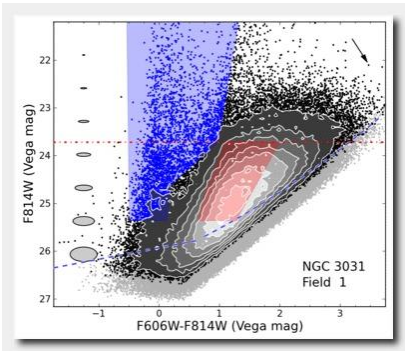


Community Contributed High Level Science Products



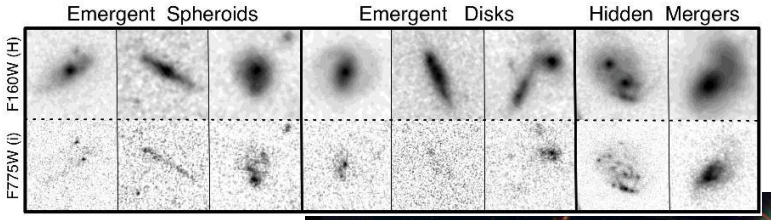


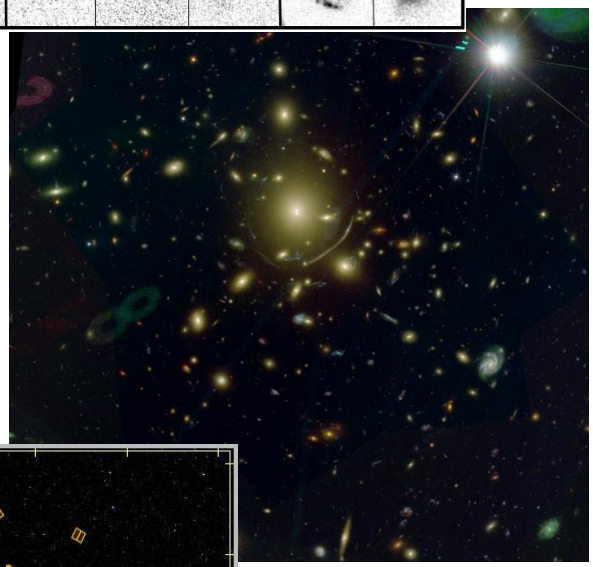


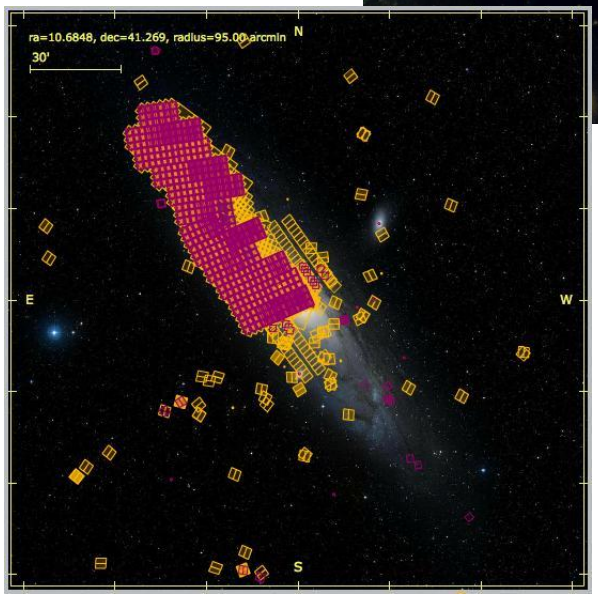


Sample High-Level Science Products: HST Multi-Cycle Treasury Programs

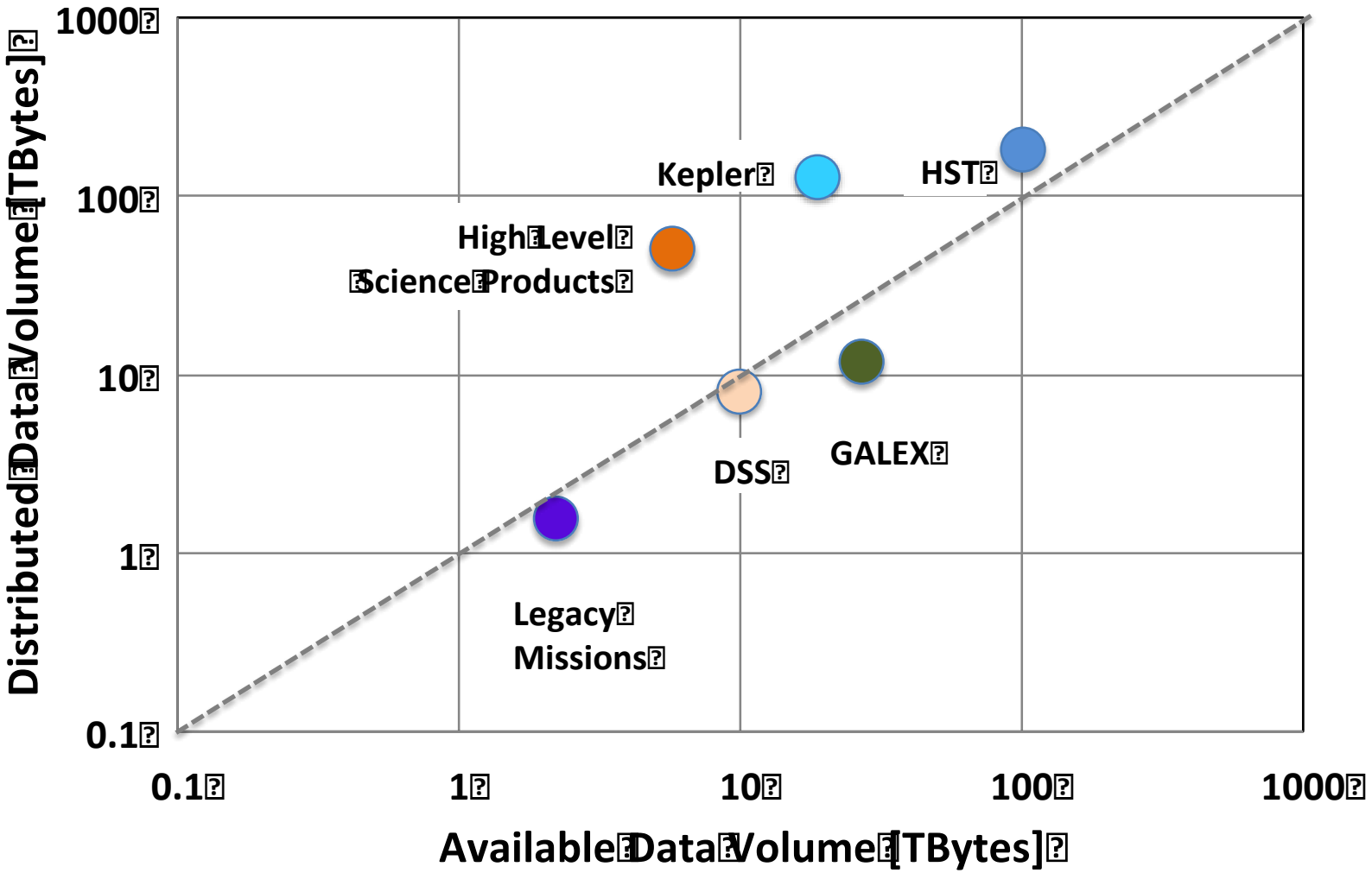
- CANDELS (Faber/Ferguson)
  - 3.1 TB of data
  - ~ 48 TB distributed to 1680 IP addresses
- CLASH (Postman)
  - 516 GB of data
  - ~ 3.9 TB distributed to 1489 IP addresses
- PHAT (Dalcanton)
  - 1.7 TB of data
  - ~ 7 TB distributed to 618 IP addresses
- PHAT, CLASH catalogs are in databases and are being integrated for user access







High-Level Science Products are heavily used



Q1: Data management costs

- **Besides how to manage data in an integrated way moving forward, what is the appropriate ratio of software/hardware cost? What's the key expense?**
- Overall software dominates the costs of data management but hardware becomes more important as the data volume ramps up.
  - Software costs scale with number of different data products (e.g., number of instruments, operating modes, etc.)
  - Hardware costs scale with the data volume.
- Hardware cost factors:
  - Storage remains a significant cost. We spend less on CPU than on storage.
  - Database machines are important and cost more per unit storage.
  - Network costs are important.
- Software costs can often be shared across missions.
  - E.g., the MAST portal project is developing interfaces that are shared by many different missions.
  - MAST databases rely on commercial DB systems (MS SQL Server) but costs of licenses and support are shared across missions.

Q2: Raw versus pipeline data

- **Discuss raw data vs. pipeline data, and the need for pipelines to be done by experts on the particular instruments.**
- Generating science-ready, high-level science products is key to enabling science.
  - There is a false savings in delivering only raw data products: the cost 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.
- On the other hand, it is important to enable contributions from the community to improved algorithms for data processing.
  - Advances in data processing originate from scientists trying to improve the quality of the data for their own science.
  - A healthy system will fund a robust pipeline for the mission/project while also supporting advanced research on improved data processing in the science community.

Q3: How & where to reprocess

- **There's a lack of standardization for ground-based data because of the need to reprocess, specific to different instruments, different conditions, different science goals. How will this intermediate-level processing be possible? Where will it occur? (that is, data will be too large to download on individual computers for reprocessing). Is a coordinated effort needed – data centers, etc.?**
- This is true for HST data too: the science program uses different instruments driven by different science goals, etc.
- The difference from ground-based data is not that the data are taken more uniformly, but that:
  - The calibration is maintained by the observatory
  - Extensive metadata describing the intent of observations is fully described in the observation database
- Going forward, better capture of this metadata is key to enabling pipeline processing at the observatories.

Q4: Where to archive?

- **How/where will archiving take place?**
- The answer depends on the project scale:
  - For small observatories should contribute data to an existing archive.
  - Large observatories/projects should keep the archive close to the location of the experts in the instrumentation and data.
- Ground-based observatories have rarely created excellent archives.
  - I attribute that to inadequate funding (and a tendency to steal funding from archives when it is needed for instrument development and operations.)
  - Ideally: Fund the observatories to develop strong archive centers going forward.
    - They can learn from the NASA archive centers, which have been generally well funded and have shown that a strong archive is essential to enable the best science.
- Finite-lifetime missions should partner with long-term archive centers to have a transition plan from active missions to legacy data archives.
  - NASA does a good job of this: a few domain-specific archive centers work with the active missions and then adopt the data products when the missions end.
  - The archive centers have expertise in data that should be captured as part of the close-out process.