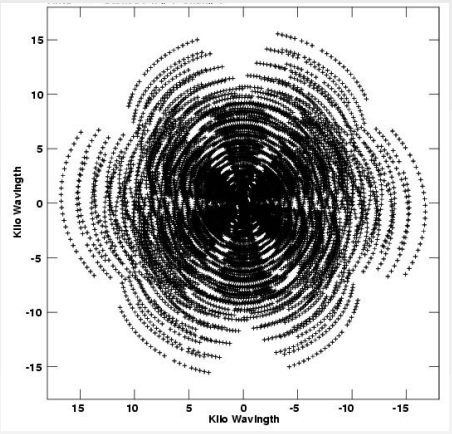
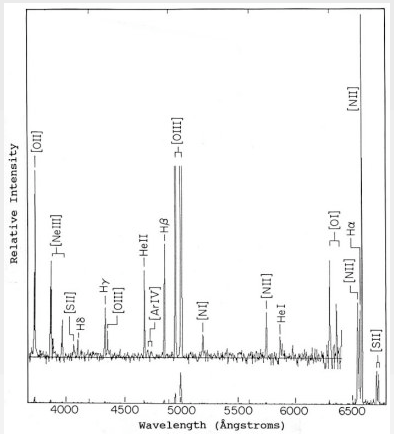
VIRTUAL ASTRONOMICAL OBSERVATORY

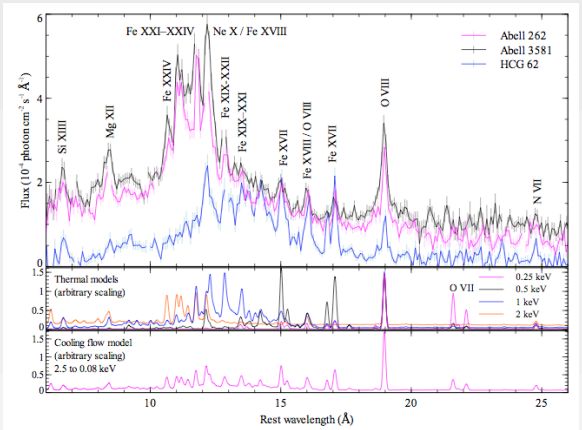
Data Discovery and Access for the Next Decade

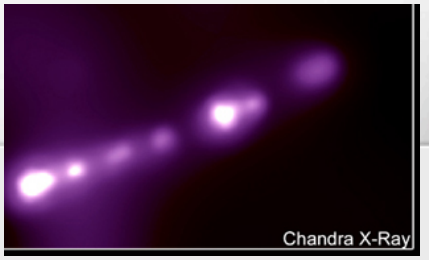
Robert Hanisch
Space Telescope Science Institute
Director, Virtual Astronomical Observatory

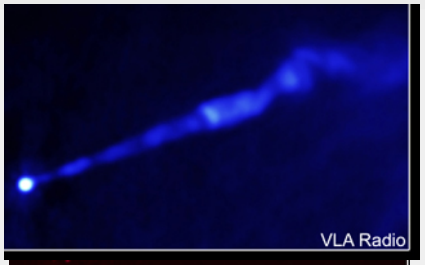
Data in astronomy

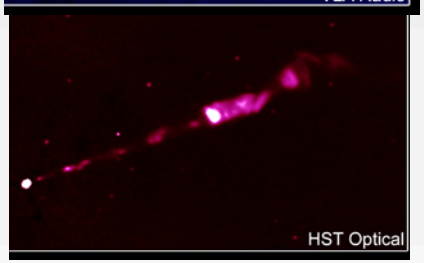


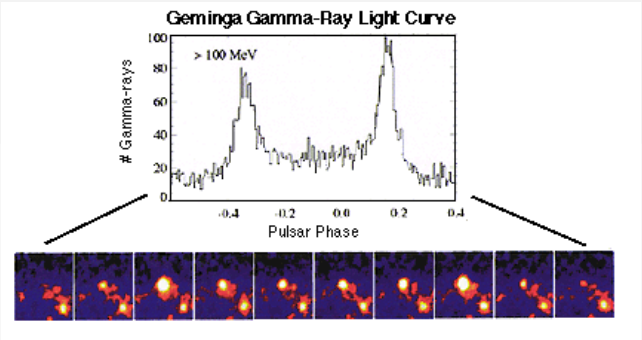


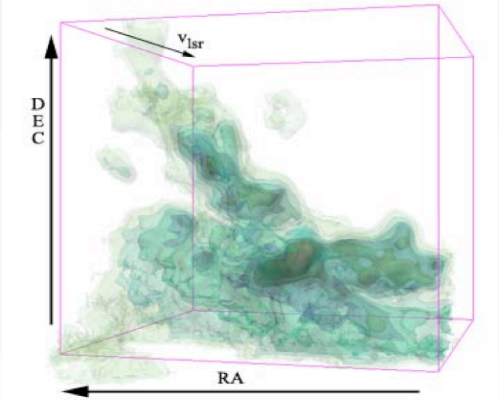












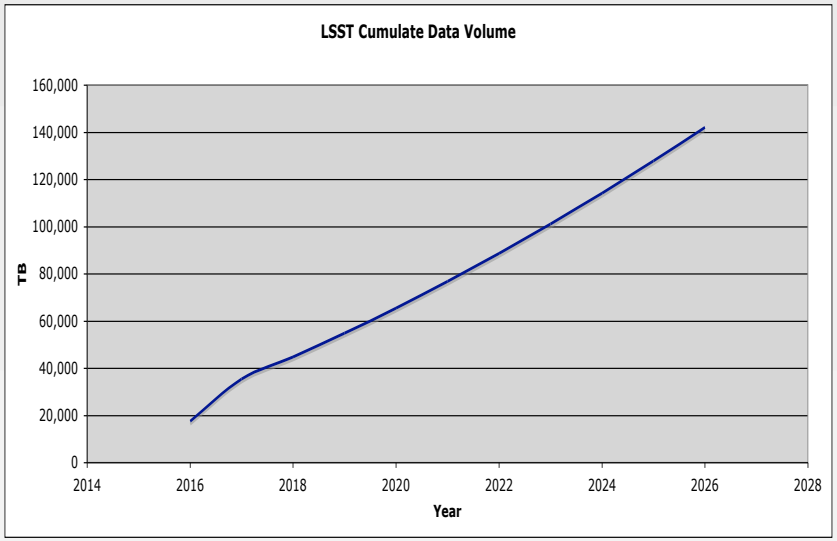
jdate	designation	ra	dec	sup_ra	sup_dec	glon	glat	density	r_k20fe
2451305.6569	12552517+2134339	12 55 25.2	21 34 33.9	193.854874	21.576124	312.417426	84.374259	2.51	5.4
2451700.6751	12554924+2123581	12 55 49.2	21 23 58.2	193.955109	21.399385	313.052422	84.184820	2.66	5.0
2451261.8020	12571719+2120180	12 57 17.2	21 20 18.1	194.321625	21.338383	316.206288	84.058667	2.24	5.9
2451261.8020	12572936+2132520	12 57 29.4	21 32 52.1	194.372269	21.547800	317.124576	84.251945	2.24	5.0
2451261.8020	12572893+2137370	12 57 28.9	21 37 37.1	194.370453	21.627054	317.296215	84.329266	2.35	5.0
2451261.7924	12562741+2131175	12 56 27.4	21 31 17.6	194.114243	21.521484	314.721401	84.277947	2.54	8.1
2451261.8020	12573991+2146420	12 57 39.9	21 46 42.1	194.416367	21.778374	318.097097	84.465693	2.35	5.0
2451261.7972	12564252+2148223	12 56 42.5	21 48 22.4	194.177277	21.806303	315.909776	84.544554	2.51	5.0
null	12564369+2140575	12 56 43.7	21 40 57.6	194.182068	21.682659	315.683000	84.423000	null	214.8
2451261.7924	12561052+2148274	12 56 10.5	21 48 27.5	194.043808	21.807701	314.635368	84.571294	2.78	7.4
2451261.7972	12571196+2146234	12 57 12.0	21 46 23.5	194.299911	21.773294	316.993658	84.486836	2.51	5.5
2451261.8020	12572147+2140450	12 57 21.5	21 40 45.1	194.339417	21.679213	317.136181	84.386833	2.35	5.3
2451700.6751	12554548+2153222	12 55 45.5	21 53 22.2	193.939529	21.889559	313.784434	84.669874	2.66	9.0

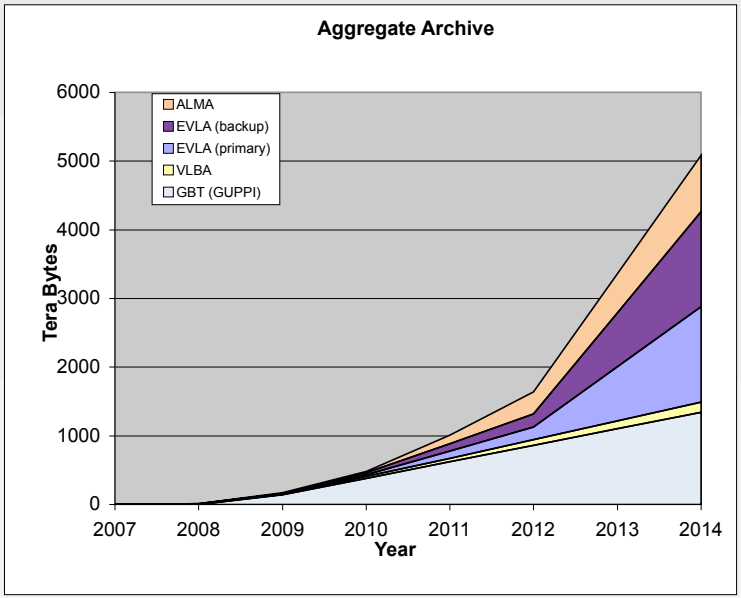
1-d, 2-d, 3-d: intensity/polarization vs. energy, time, position, velocity
tables: catalogs, x-ray event lists, radio visibility measurements

Quantity and distribution

- ~50 major data centers and observatories with substantial on-line data holdings
- ~10,000 data “resources” (catalogs, surveys, archives)
- data centers host from a few to ~100 TB each, currently ~1 PB total
- current growth rate ~0.5 PB/yr, expected to increase soon
- current request rate ~1 PB/yr
- for Hubble Space Telescope, data retrievals are 3X data ingest; papers based on archival data constitute 2/3 of refereed publications

Data growth



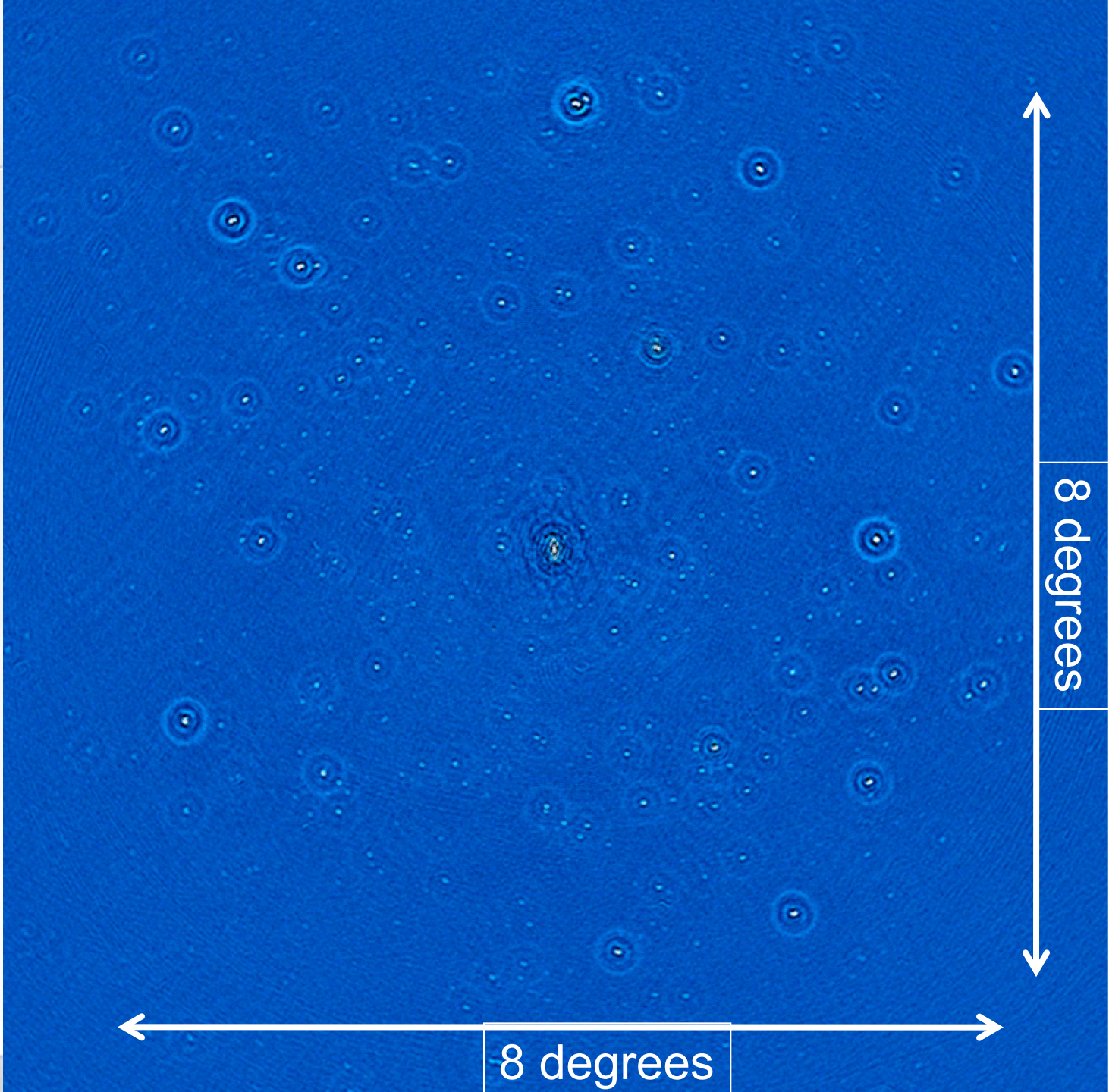


	Expected number of galaxies/stars per band (end of LSST Survey)				# epochs
	Universal Sky		Galactic plane		
	Galaxies	Stars	Galaxies	Stars	
U	2.50E+09	1.50E+09	5.00E+08	5.10E+09	70
G	4.00E+09	3.00E+09	2.00E+09	3.10E+09	100
R	4.00E+09	3.00E+09	2.00E+09	3.10E+09	230
I	5.50E+09	4.50E+09	3.50E+09	4.60E+09	230
Z	1.00E+10	9.00E+09	8.00E+09	9.10E+09	200
Y	2.50E+09	1.50E+10	5.00E+08	1.60E+10	200
max	**1.00E+10**	**1.50E+10**	**8.00E+09**	**1.60E+10**	
avg					**171.67**
sum	**2.85E+10**	**3.60E+10**	**1.65E+10**	**4.10E+10**	**1,030**

ALMA: 5 PB by 2014
LSST: 60 PB by 2020,
~40B row databases

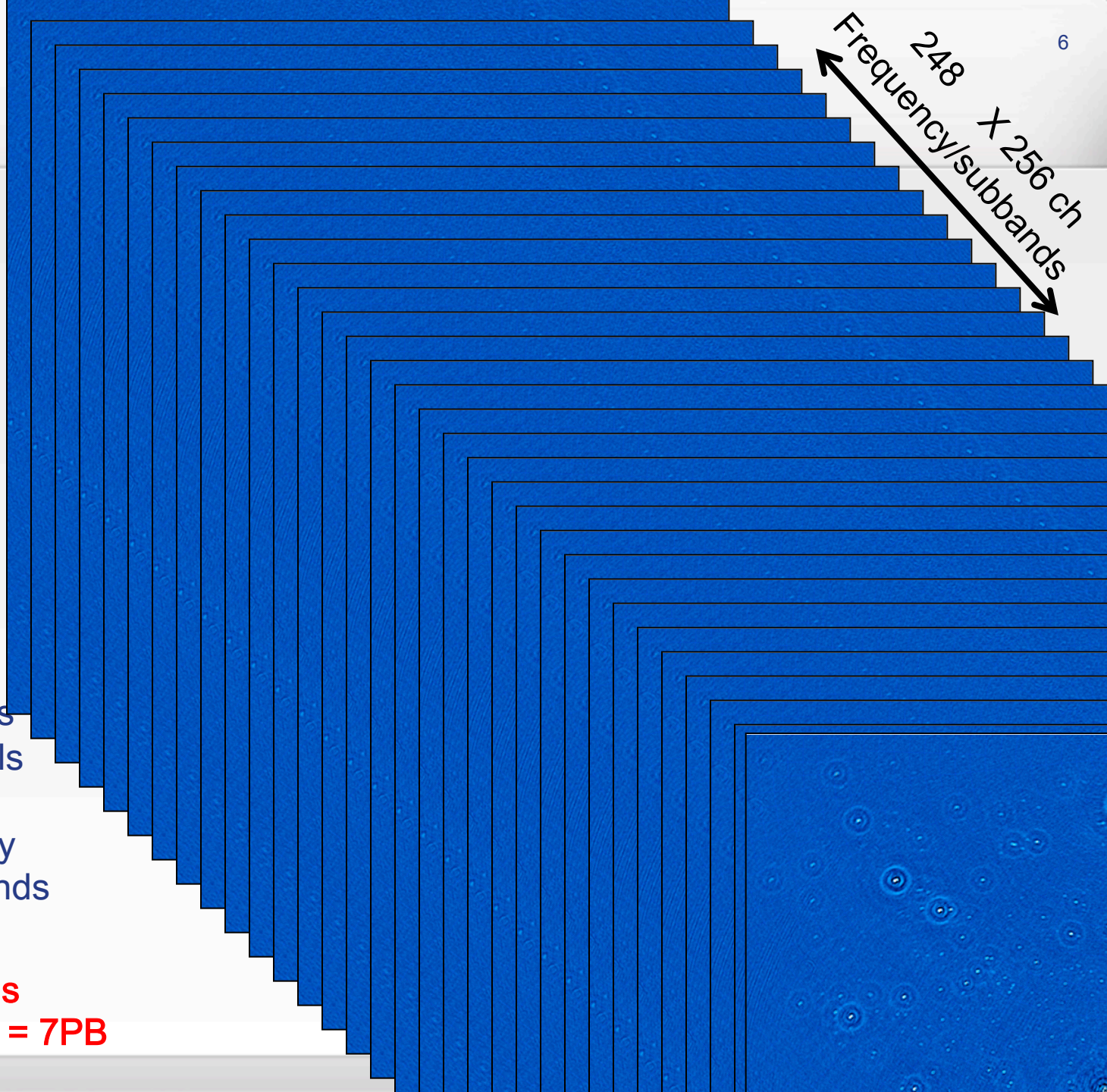
LOFAR Data Examples

Source:
3C61.1
Frequency (HBA):
115-185 MHz
Image size:
8x8 degrees
Resolution:
10 arcsec
Pixel size:
4 arcsec pixels
Total # of pixels:
5.184e+07 pixels
51,840,000 pixels
Total data size:
200GB/frequency
(c/o A. Alexov, M. Wise)



Source:
3C61.1
Frequency (HBA):
115-185 MHz
Image size:
8x8 degrees
Resolution:
10 arcsec
Pixel size:
4 arcsec pixels
Total # of pixels:
5.184e+07 pixels
51,840,000 pixels
Total data size:
200MB/frequency
50GB all subbands

14.3TB all channels
x 500 HBA pointings
of the Northern sky = 7PB



Expectations (c/o A. Szalay)

- Scientific data is doubling every year, reaching PBs
- Data is everywhere, never will be at a single location
- Need randomized, incremental algorithms
  - Best result in 1 min, 1 hour, 1 day, 1 week
- Architectures increasingly CPU-heavy, IO-poor
- Data-intensive scalable architectures needed
- Most scientific data analysis done on small to midsize BeoWulf clusters, from faculty startup
- Universities hitting the “power wall”
- Soon we cannot even store the incoming data stream
- Not scalable, not maintainable…

Astro2010

- Notes dependence of astrophysics research on high performance computing
  - Simulations
  - Data processing
  - Data storage and access
- Move toward many-core architectures
  - How to make effective use?
  - Legacy software
  - Compiler technology
- Can expect a massive re-tooling of the computational infrastructure in the coming decade

Astro2010

- Data archives
  - "Central to astronomy today"
  - HST, 2MASS, and SDSS archival research is major contributor to scientific productivity

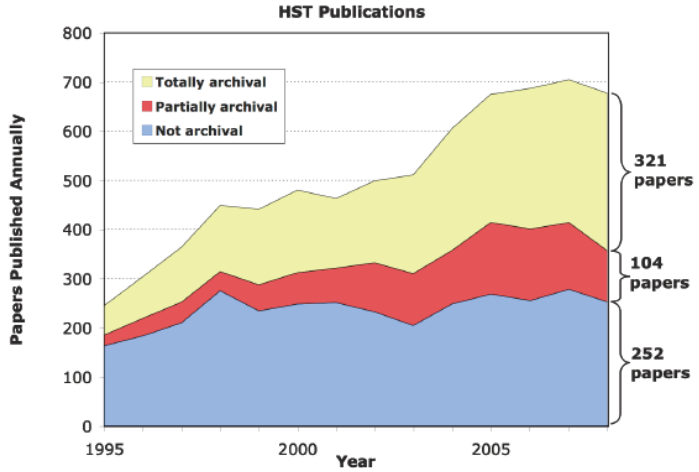


Figure 1: Number of annual publications using Hubble Telescope data. The publications have been divided into non-archival papers written by the original investigators (blue), totally archival publications not involving none of the original proposers (yellow), and papers that include data from multiple proposals with some being archival and some not (red). The number of archival papers has exceeded the number of PI-led papers since 2006.

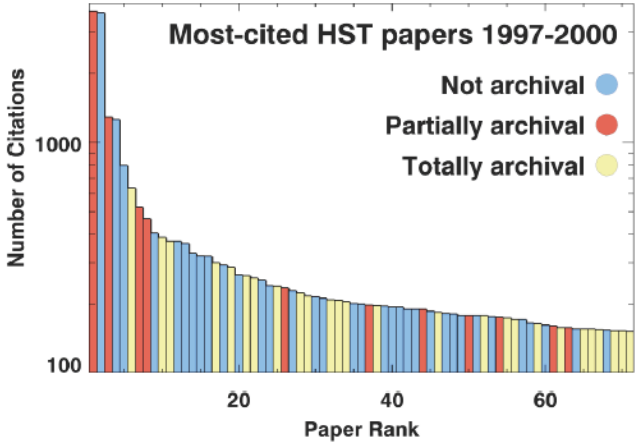


Figure 2: Highly cited HST publications between 1997 and 2000. All 71 papers with more than 150 citations (as of March 2009) are included in the sample. Note the *y*-axis is logarithmic. As in Figure 1, the publications have been divided into non-archival (blue), totally archival (yellow), and partially archival (red) depending on whether the original proposers were authors on the paper. Totally archival papers make up 37% of the highly cited sample, which is slightly above the rate expected based on their frequency of publication during this period.

c/o R. White (STScI) and pp. 5-11, 5-12 of NWNHAA

- Radio facilities can broaden impact by providing VO-compliant archives

Astro2010

- Virtual Observatory
  - *"The National Virtual Observatory [with international VO collaboration] …has produced widely accepted standards for data formatting, curation, and the infrastructure of a common user interface."*

  [Note: VO not explicitly reviewed in Astro 2010, as it was an approved program in the 2000 Decadal Survey and already being implemented as Astro 2010 was in progress.]
- Data preservation and curation
  - *"It is…necessary for NSF to adopt NASA's model of long-lived data archive centers…for long-term curation of data."*
- Software
  - *"New packages capable of handling large datasets are urgently needed. These are likely to be created and employed within a common-use environment."*

Astro2010

- Facility planning and data management
  - *"Recommendation: Proposals for new major ground-based facilities and instruments with significant federal funding should be required as a matter of agency policy to include a plan and if necessary a budget for ensuring appropriate data acquisition, processing, archiving, and public access after a suitable proprietary period."*
- But note CODMAC (1982, NAS) report:
  - *"Generally, data-system and data-analysis activities are not adequately funded. Underfunding results from at least three related causes: when there is insufficient planning in the early mission phases, the required funding will often be underestimated; overruns that occur during mission system development may absorb the funds allocated for data handling and analysis; and because of imperfections in the flight and ground hardware and software, the data processing may be more extensive than originally estimated."*

Data reduction & analysis software

- Packages in widest use today date from 1980s
- Institutional support has waned
- Planning for next generation systems is weak
- CASA is notable exception
- New approaches may come from outside the astronomy community; "disruptive" technologies (e.g., *R*)
- Plan/budget for DR&A software: integral component of facilities
- Next-gen systems should be
  - Component-based
  - Language-flexible
  - Open
  - Conducive to astronomer-developers and code-sharing
  - Able to take advantage of multi-core processors, GPUs

Data management

- Well-characterized archival data enormously valuable, both from dedicated surveys and heterogeneous collections
- Data discovery/federation enabled by the Virtual Observatory; challenges remain
  - Need database technology capable of managing 10^9 – 10^{12} rows; potentially disruptive technology change
  - Need increases in network bandwidth, ability to move algorithm to data
  - Metadata management critical
  - Support for long-term access to survey data, other heritage data products, unclear
- Plan/budget for comprehensive archiving, long-term curation, VO-compatible access

Observation and simulation

- Unprecedented opportunity for bringing together simulation and data faces us now
- Interoperability fostered by VO protocols/standards
- Need to improve access, transparency, reproducibility, return on investment, efficiency, and infrastructure
- Visualization tools essential for understanding simulations, large datasets, and relationships
- Simulations and observations must be made interoperable, facilitated by VO protocols and standards

Professional development & education

- Data-literacy increasingly important skill, but students have little interaction with real data
- Software development in astronomy under-appreciated
- Software development skills very important for astronomy graduate students
- Deployment of data and data analysis tools to university classroom difficult
- Software developers do not fare well in traditional career paths (usual citation metrics are inadequate measure of merit)
- Improve software skills for astronomy grad students (Comp. Sci. training, summer schools)
- Recognize software professionals as integral to astronomy enterprise; provide appropriate merit-based career paths
- Bring data and appropriate data analysis tools into undergraduate education

Academic/industry collaboration

- Largest-scale computation and data storage facilities now in corporate hands (Google, Amazon, Microsoft, …)
- Corporate entities interested in astronomy and astronomy data (GoogleSky, WorldWide Telescope)
- Other research communities making extensive use of web-based collaboratories, workflow systems, component sharing
- Astronomy community uncomfortable with reliance on commercial providers of storage and processing; concerns about control and commitment
- Impedance mismatch between astronomy and industry: data as data, data as profit center (Cloud-based storage currently not affordable for astro research community)
- Enormous potential for astronomy as customers; benefits of collaboration are less certain
- Engage in collaborations with industry partners and test storage and computational facilities

NSF MPSE interests

- "Data-Enabled Science in the Mathematical and Physical Sciences"
  - Commissioned by Ed Seidel
  - Chaired by James Berger (UNC)
  - 29-30 March 2010
  - Two representatives from each of astronomy, chemistry, materials research, mathematics, physics
- Two major challenges identified
  - Data management itself
  - Scientific inference from massive data
- Recommendations re/ astronomy
  - Support comprehensive data management for all major facilities and large data producing projects.
  - Close the gaps in astronomy data archiving.
  - Invest in data mining, analysis, and visualization algorithm development and corresponding scientific applications that are tailored to research with large astronomical data sets and that foster the emerging field of astroinformatics.
  - Develop programs that capitalize on archival research, possibly in collaboration with NASA.
  - Support professional communication through workshops and conferences focused on data-intensive astronomical research.
  - Develop stronger programs for education and outreach, in collaboration with EHR, highlighting data-enabled science and citizen science.

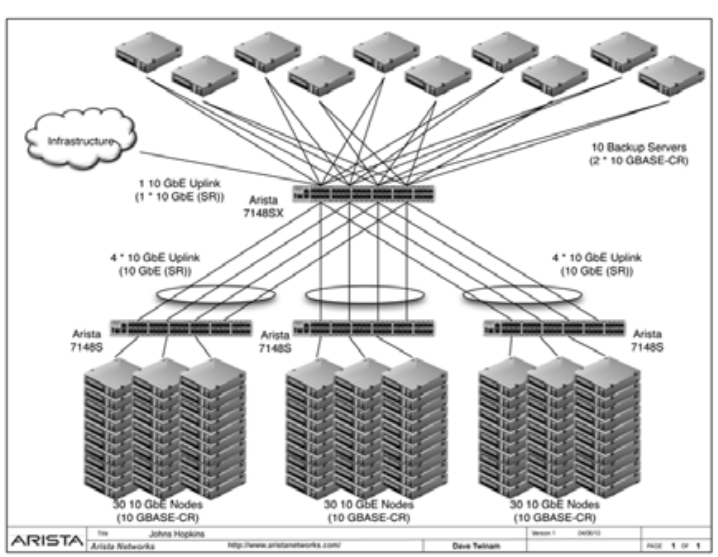
- NRAO Green Bank
- May 3: Data understanding
  - J. Lazio, A. Connolly, D. Pugmire
- May 4: Data processing
  - A. Szalay, J. Tarter, B. Berriman
- May 5: Data management
  - R. Hanisch, M. Wise
- Industry leaders invited
- *Workshop format*, evening topical discussions

Other meetings like this:

- "The Growing Demands on Connectivity and Information Processing in Radio Astronomy from VLBI to the SKA", Aveiro, Portugal, May 24-25, sponsored by RadioNet et al.
- "Big Data", Calgary, May 31, sponsored by ISIS

Data-optimized computing at JHU

- A. Szalay leading construction of “Data-Scope”
- $2M from NSF + $1M from Johns Hopkins
- Features
  - 5 PB storage
  - 90 computational servers
  - 12 storage servers
  - 540 Tflop aggregate
  - 460 GB/s I/O
  - 116 kW power consumption
- Will support targeted problems in astrophysics, genomics, turbulence, environmental studies, et al.



The research record and data

- Journals and preprints in astronomy are themselves data
- Data underlying the images and graphics published in journals not systematically preserved
- Without full stewardship of the research record, key elements of scientific process missing: reproducibility, integrity
- Develop data-friendly publication policies and long-term data stewardship solutions
- Monitor intellectual property, copyright, and open access policies and re-examine publishing business model
- VAO collaborating with NSF CISE-funded project, the Data Conservancy (DataNet program)
- Note: NSF policy now requires data management plans with all proposals

The Virtual Observatory

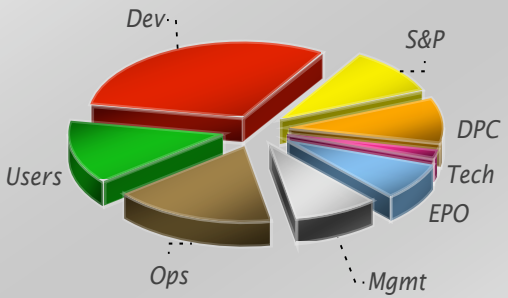
- The VO is foremost a data discovery, access, and integration facility
- International collaboration on metadata standards, data models, and protocols
  - Image, spectrum, time series data
  - Catalogs, databases
  - Transient event notices
  - Software and services
  - Distributed computing (authentication, authorization, process management)
  - Application inter-communication
- International Virtual Observatory Alliance established in 2001, patterned on WorldWideWeb Consortium (W3C)

US VO efforts

- National Virtual Observatory (NVO) development effort, 2001-08
  - $14M, 17 organizations
  - NSF Information Technology Research program
- Virtual Astronomical Observatory (VAO) operational facility, 2010-2015
  - Funding is $5.5M/year for five years, subject to annual performance review, 9 organizations
  - $4M/year from NSF/AST
  - $1.5M/year from NASA
  - Covers ~27 FTE over the ten organizations
- VAO is managed by the VAO,LLC (limited liability company) co-owned by AUI (operates NRAO and ALMA) and AURA (operates NOAO and STScI)
  - VAO has its own Board of Directors (J. Gallagher, chair)
  - R. Hanisch, director; B. Berriman, program manager, D. De Young, project scientist, A. Szalay, technology advisor
  - G. Fabbiano, chair of Science Council

Scope and functions

- Seven major areas of activity
  - Operations: T. McGlynn, HEASARC, A. Thakar, JHU
  - User Support: E. Stobie, NOAO, M. Nieto-Santisteban, JHU
  - Product Development: R. Plante, NCSA, G. Greene, STScI
  - Standards and Protocols: M. Graham, Caltech, D. Tody, NRAO
  - Data Preservation and Curation: A. Rots, SAO, J. Mazzarella, NED (A. Accomazzi, SAO/ADS)
  - Technology Evaluation: A. Mahabal, Caltech
  - Education and Public Outreach: B. Lawton, STScI



Science initiatives

- The VAO has selected seven science initiatives that were endorsed by the Science Council as providing maximal scientific impact in the astronomy community:
  1. Development of a dedicated VAO Portal
  2. Scalable cross-matching between catalogs of sources
  3. Building and Analyzing Spectral Energy Distributions
  4. Time Domain Astronomy: (a) Periodograms and light curve analyses; (b) Transient event services
  5. Data Linking and Semantic Astronomy
  6. Desktop Tool Integration
  7. Data Mining and Statistical Analysis

Science deliverables

- Four areas selected for science deliverables in Year 1 (assume start date = Oct 1 2010).

Science Deliverable	Delivery Date	Lead
Portal that supports search, visualization, filtering and data access across all data sets accessible to the VAO	Jun 30, 2011	Tom Donaldson, STScI
SED service that collects and plots multi-wavelength data and supports interactive visualization attributes of data	July 30, 2011	Janet Evans, SAO
Deliver **cross-matching engine** that supports cross-matches across at least two large catalogs	August 30, 2011	John Good, IPAC & Tamas Budavari, JHU
Time Series Astronomy: Deliver periodogram service and light curve classification service for data sets at NStED, TSC (Harvard)	September 30, 2011	John Good, IPAC

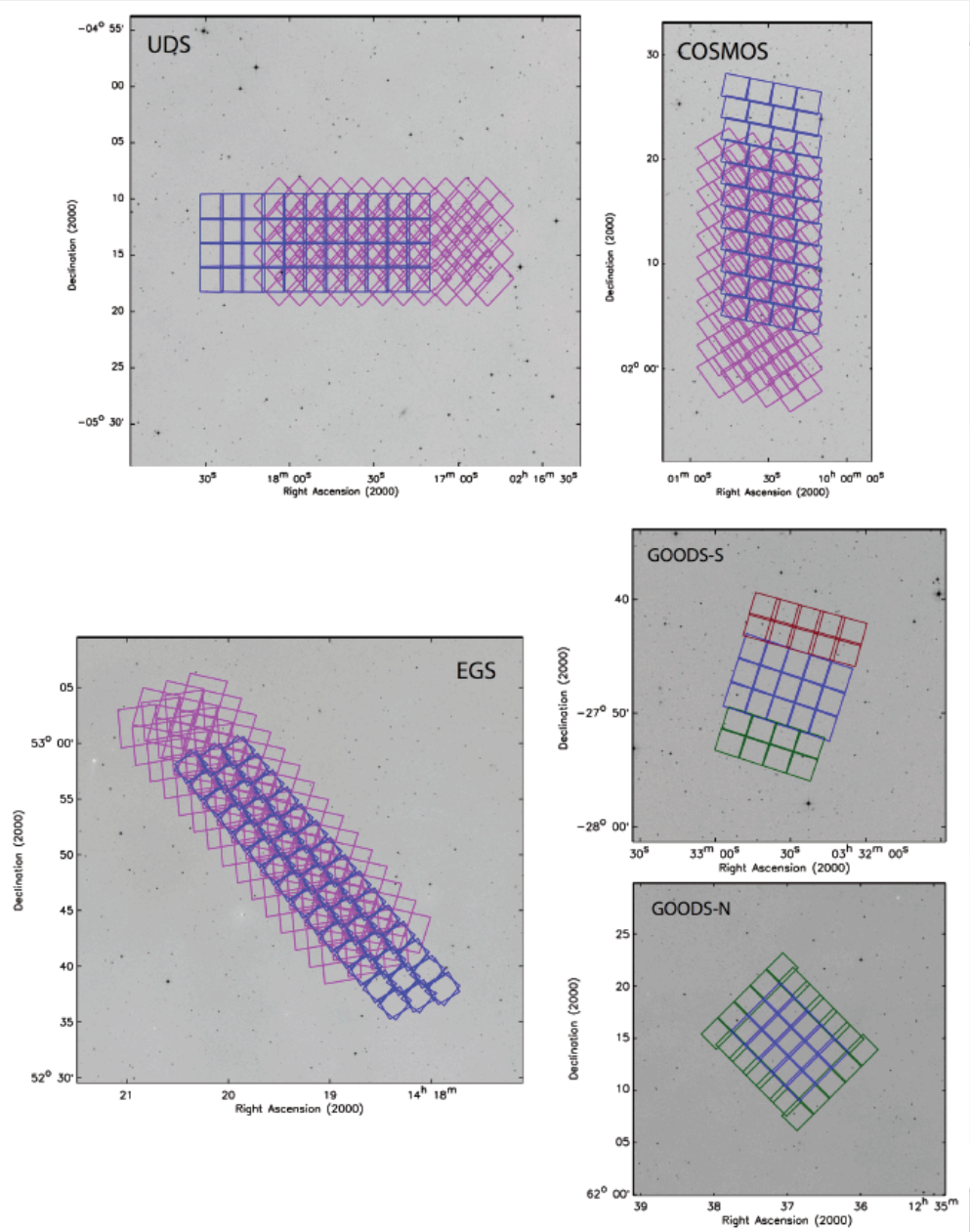
Science deliverables

- Four science initiatives will undergo a study period during Year 1:
  - Time Domain Astronomy (Transients)
  - Data Linking and Semantic Astronomy
  - Desktop Tool Integration
  - Data Mining and Statistical Analysis
- The goals of these studies are to make recommendations on science deliverables for Year 2+ that will be evaluated by the Science Council.

Science collaborations

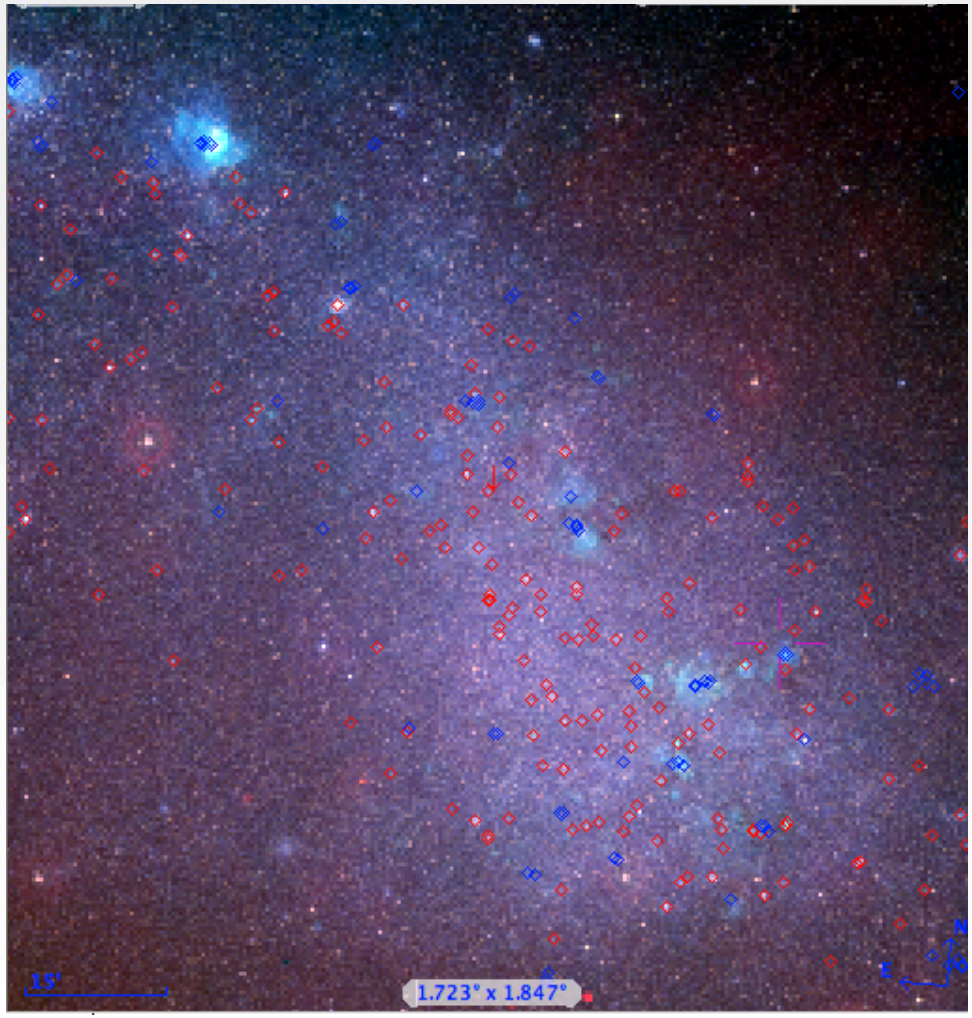
- CANDELS: Cosmic Assembly Near-infrared Deep Extragalactic Legacy Survey
  - HST multi-cycle (3-year) treasury program, S. Faber and H. Ferguson, CoPIs, >100 members of science team
  - Multi-wavelength (radio to x-ray) study of >250k galaxies with $1.5 < z < 8$
  - Understand initial epoch of star formation, disk formation, first generation of interactions and mergers, role of AGN formation in galaxy evolution
- SED-informed cross-matching
- VOEvent notices (supernovae)
- Image cut-out services

CANDELS fields



Small Magellanic Cloud

- Construct 3-dim model of SMC based on period-luminosity data on 3,000+ Cepheid variables
  - Construct SEDs for ~100M objects in 10x10 deg FOV
  - Stellar population study of a dwarf galaxy
  - Effects of galaxy interactions in dwarf systems
  - B. Madore (Carnegie) PI
- Test of scalable cross-matching and large-scale SED construction



Summary

- Advanced facilities of the coming decade will produce unprecedented volumes of data, complex data
- Sound data management practices must be integrated into facility / instrumentation design and implementation
- We will live in a world of distributed data, distributed services
- Data discovery, access, re-use, and comparison, is enabled by adherence to VO standards and protocols
- New and/or potentially disruptive technologies will be needed to manage and understand massive data sets