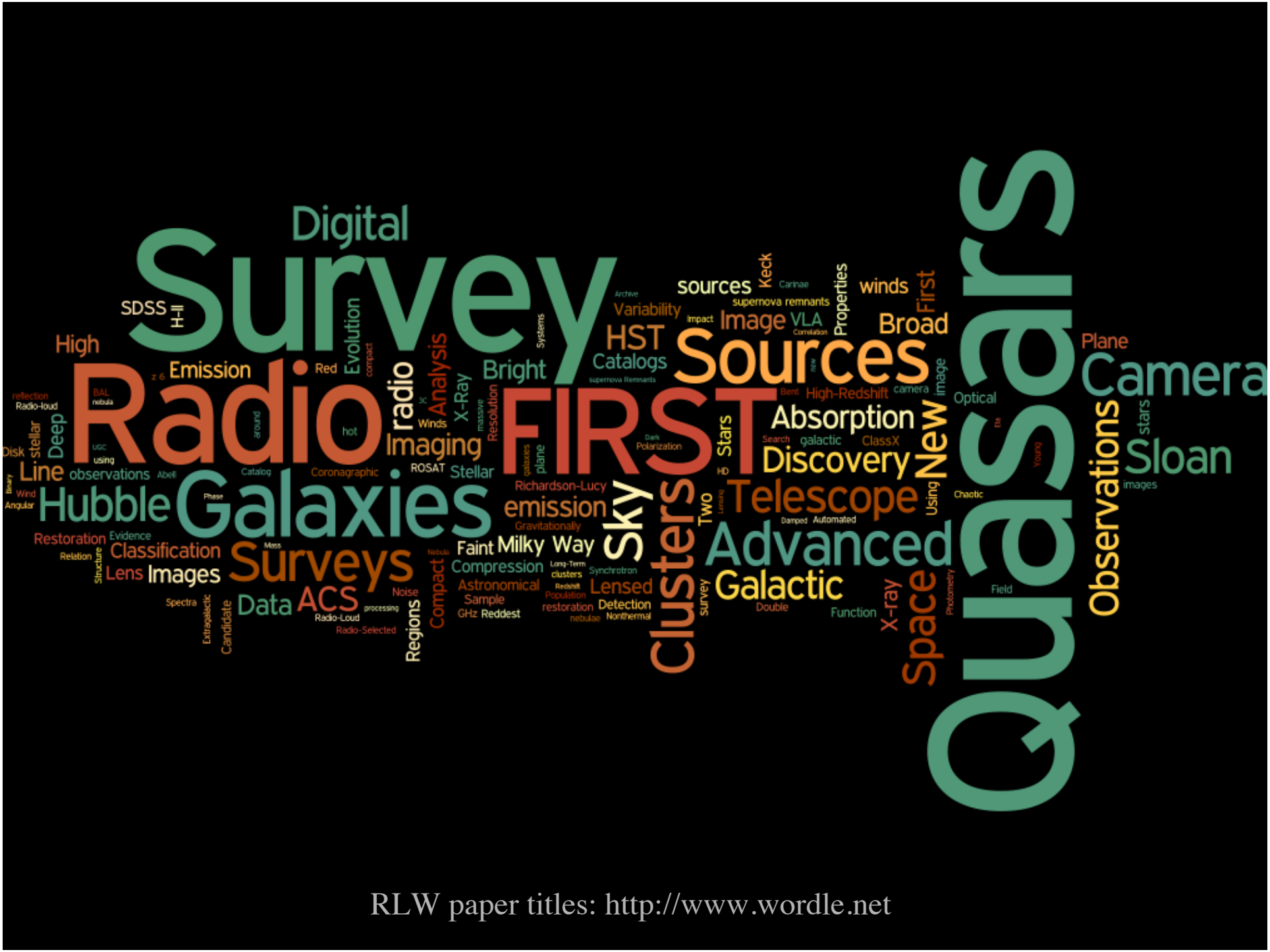


RLW paper titles: http://www.wordle.net

Astronomical Surveys and Data Archives

Richard L. White
Space Telescope Science Institute

HiPACC Summer School, July 2012

Overview

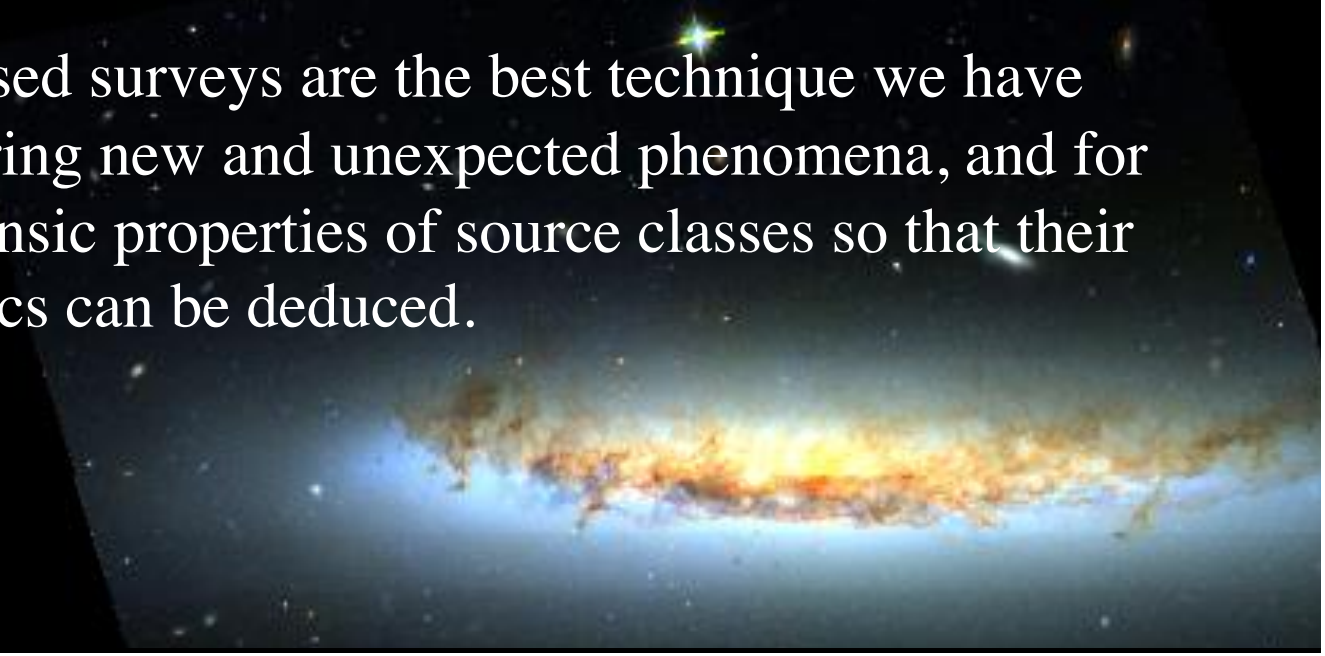
- Surveys & catalogs: Fundamental tools for astronomy

- Mission data archives: Heterogeneous but powerful resources

The Value of Surveys

Astronomy is an observational, rather than an experimental, science. Astronomers carry out the equivalent of experiments by discovering and studying astrophysical systems with a variety of ages and initial conditions. Surveys generate the list of available laboratories for such studies and are central to progress in the discipline.

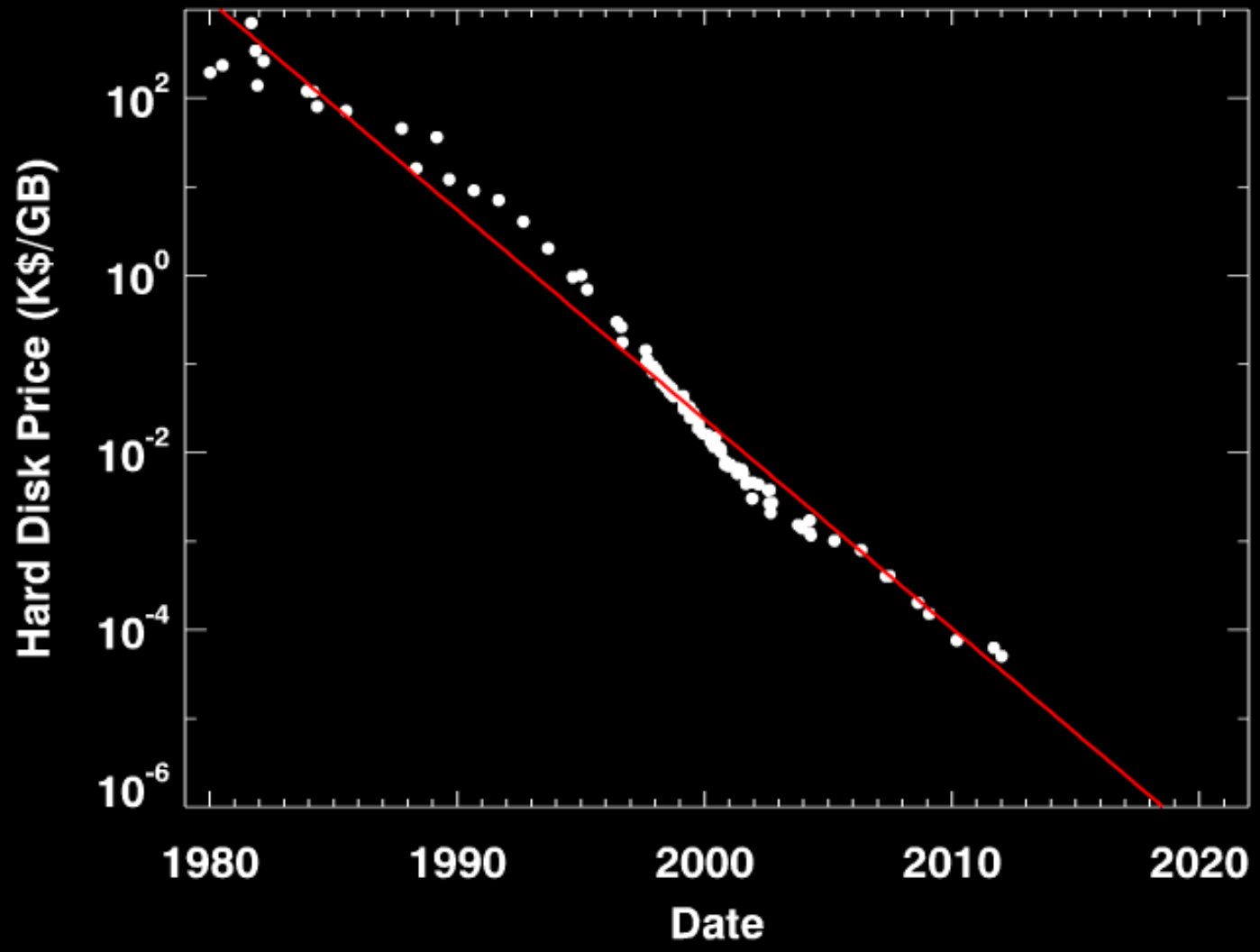
Complete, unbiased surveys are the best technique we have both for discovering new and unexpected phenomena, and for deriving the intrinsic properties of source classes so that their underlying physics can be deduced.



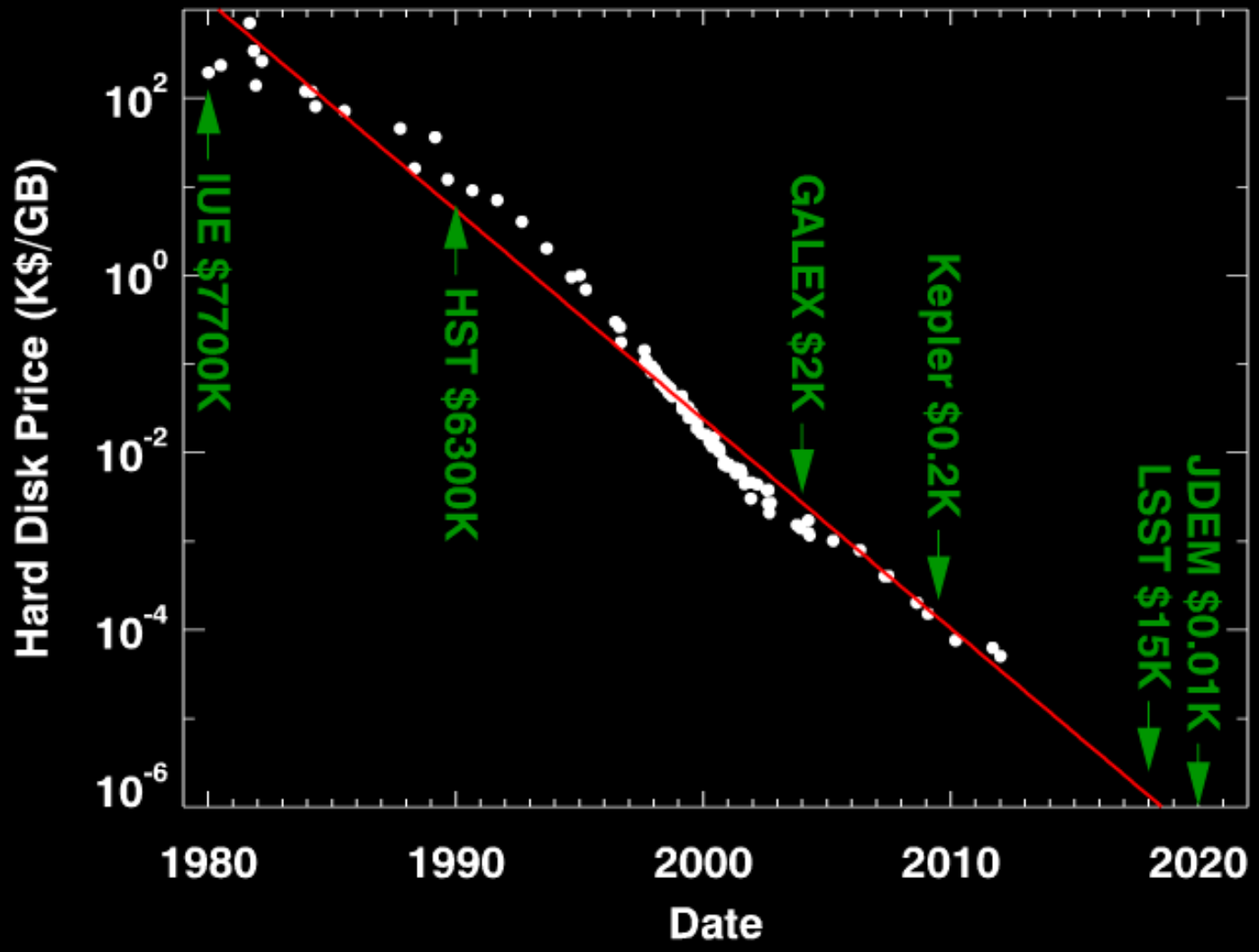
Surveys & Catalogs

- Exponential growth in computing power has enabled vast increases in:
  - Electronic detector sizes
  - CPU computational power
  - Memory for data processing
  - Storage for archives
- Observational astronomy is transitioning to an enterprise dominated by surveys

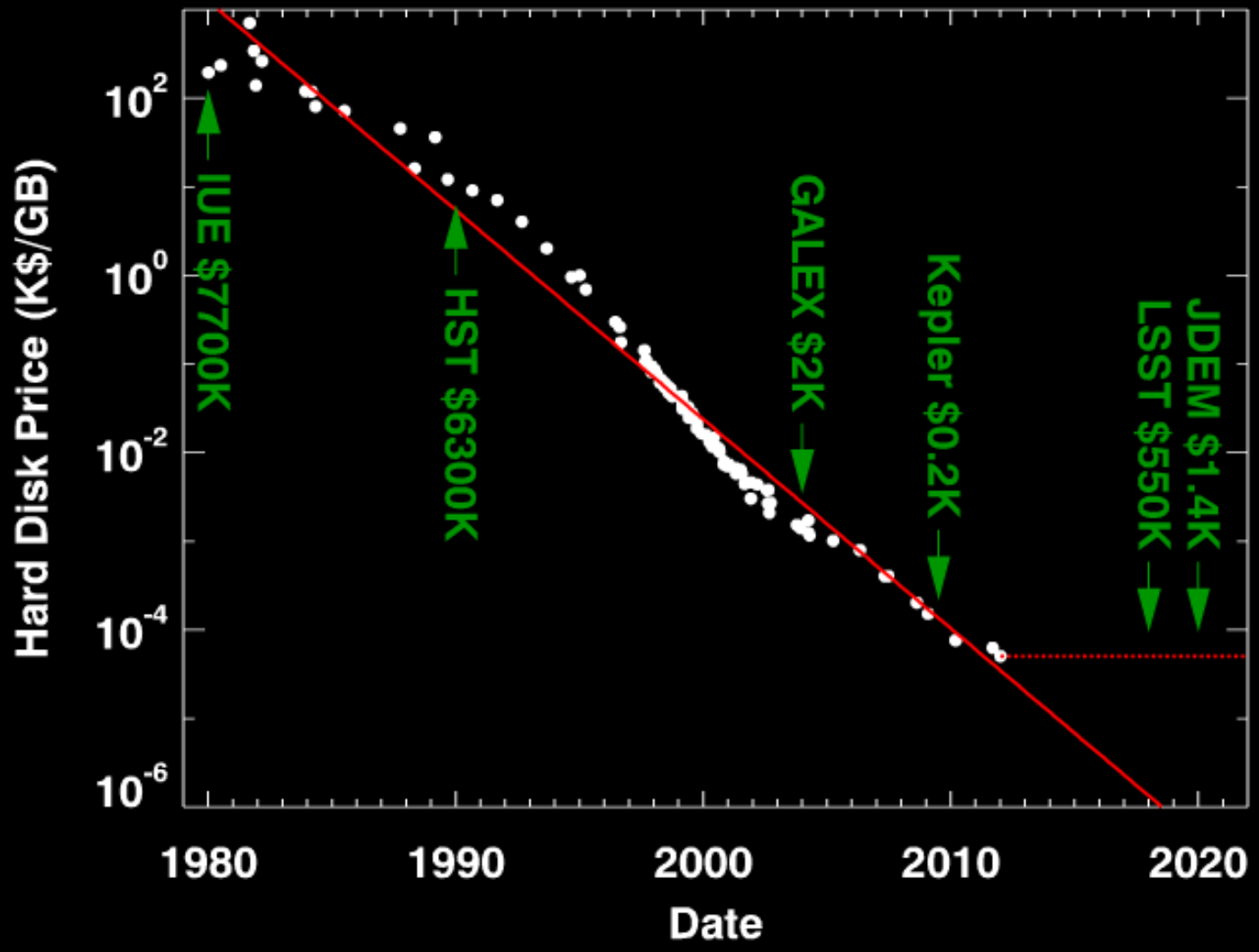
Disk Cost per Gigabyte



Disk Cost per Gigabyte



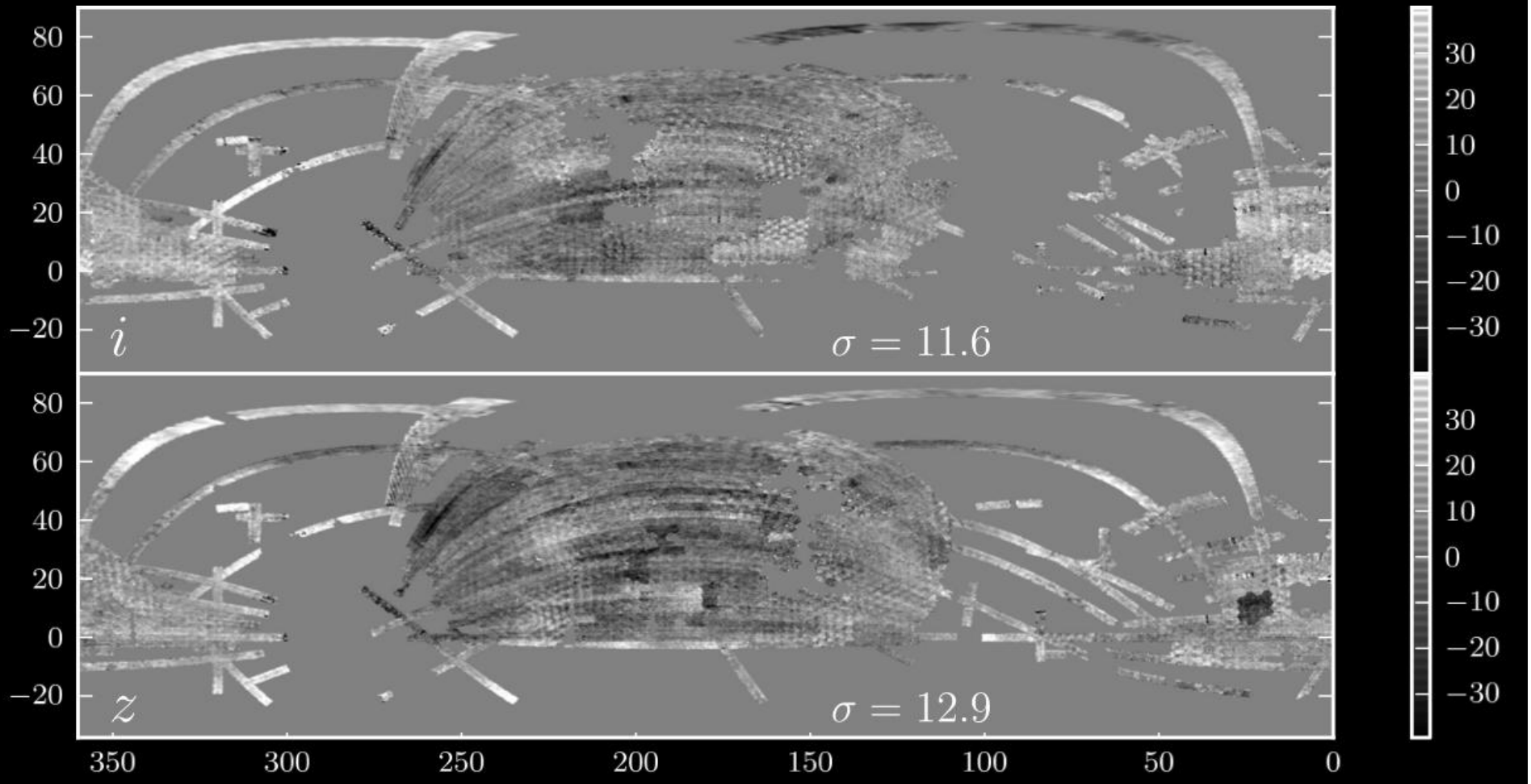
Disk Cost per Gigabyte



Catalogs

- Creating a catalog is an integral part of the survey
  - Survey reduction is iterative: info from initial catalog feeds back into improved calibration
    - E.g., Pan-STARRS ubercal
  - Early science with catalog is key to improving survey quality
    - Searching for SDSS objects without 2MASS counterparts is a great way to find flaws in both catalogs

Pan-STARRS Ubercal



Comparison of Pan-STARRS photometric calibration to SDSS
Schlafly, Finkbeiner, et al.

Tools for Catalog Access

- Essential catalog functions:
  - Search on position or other parameters
  - Cross-match with other catalogs
- Primary access tools:
  - Virtual Observatory (VO) services
    - Cone search (position), ObsTAP (more flexible, but not widely supported)
  - Direct database access: CasJobs (SQL)
    - Far more powerful, allows large queries & results
  - Custom tools & services
  - Just download the catalog (not practical for largest)

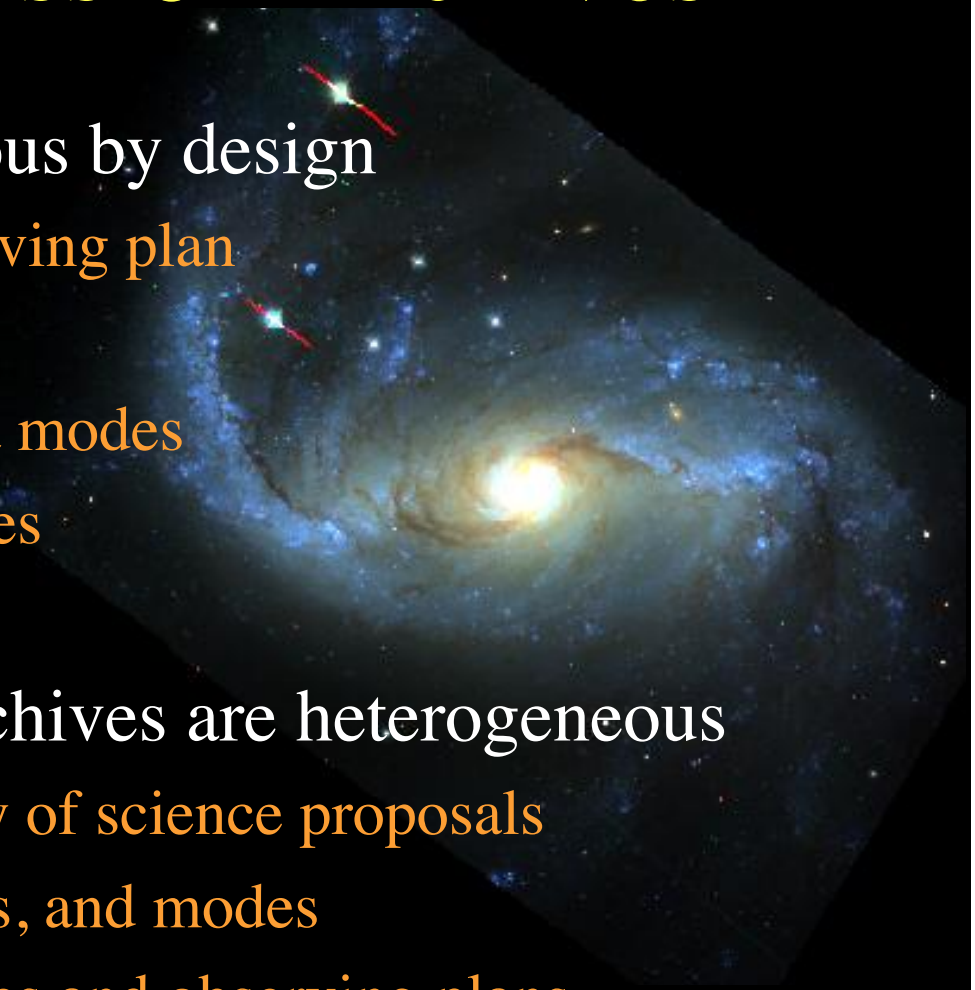
Selected Catalogs

Catalog	Wavelength	Sky Area	Comments
2MASS	Near-IR 1.1–2.3 μm	All sky	Bright, mainly stars
GSC2/USNO-B	Optical	All sky	From photographic plates!
WISE	Mid-IR 3–22 μm	All sky	
ROSAT	X-ray	All sky	
Hipparcos/Tycho	Optical	All sky	Astrometric
GALEX	Ultraviolet	30,000 deg^2	
NVSS	Radio 20 cm	30,000 deg^2	Very low resolution 45”
Pan-STARRS	Optical 0.4–1.1 μm	30,000 deg^2	Incomplete, not public yet
SDSS	Optical 0.35–1.1 μm	10,000 deg^2	Images & spectra
FIRST	Radio 20 cm	10,000 deg^2	Matches SDSS area
Medium deep	Optical, IR	~ 10^3 deg^2	UKIDSS, CFHTLS, Kepler, …
Small very deep	Many, X-ray to radio	< 2 deg^2	Cosmos, UDF, GOODS, …

The Sloan Digital Sky Survey

- SDSS set the standard that all current & future surveys should aspire to
  - High quality, highly uniform data products
  - Well documented, early, and frequent public data releases
  - Powerful tools for public data access (including the creation of CasJobs)
  - Enormous science impact
- SDSS is the model for future projects such as LSST

Surveys vs. Mission Archives



- Surveys are homogeneous by design
  - Driven by a single observing plan
  - One instrument
  - Few filters or instrument modes
  - Consistent exposure times
  - Uniform sky coverage
- Mission/observatory archives are heterogeneous
  - Driven by a great variety of science proposals
  - Many instruments, filters, and modes
  - Highly variable exposures and observing plans
  - Very uneven sky coverage

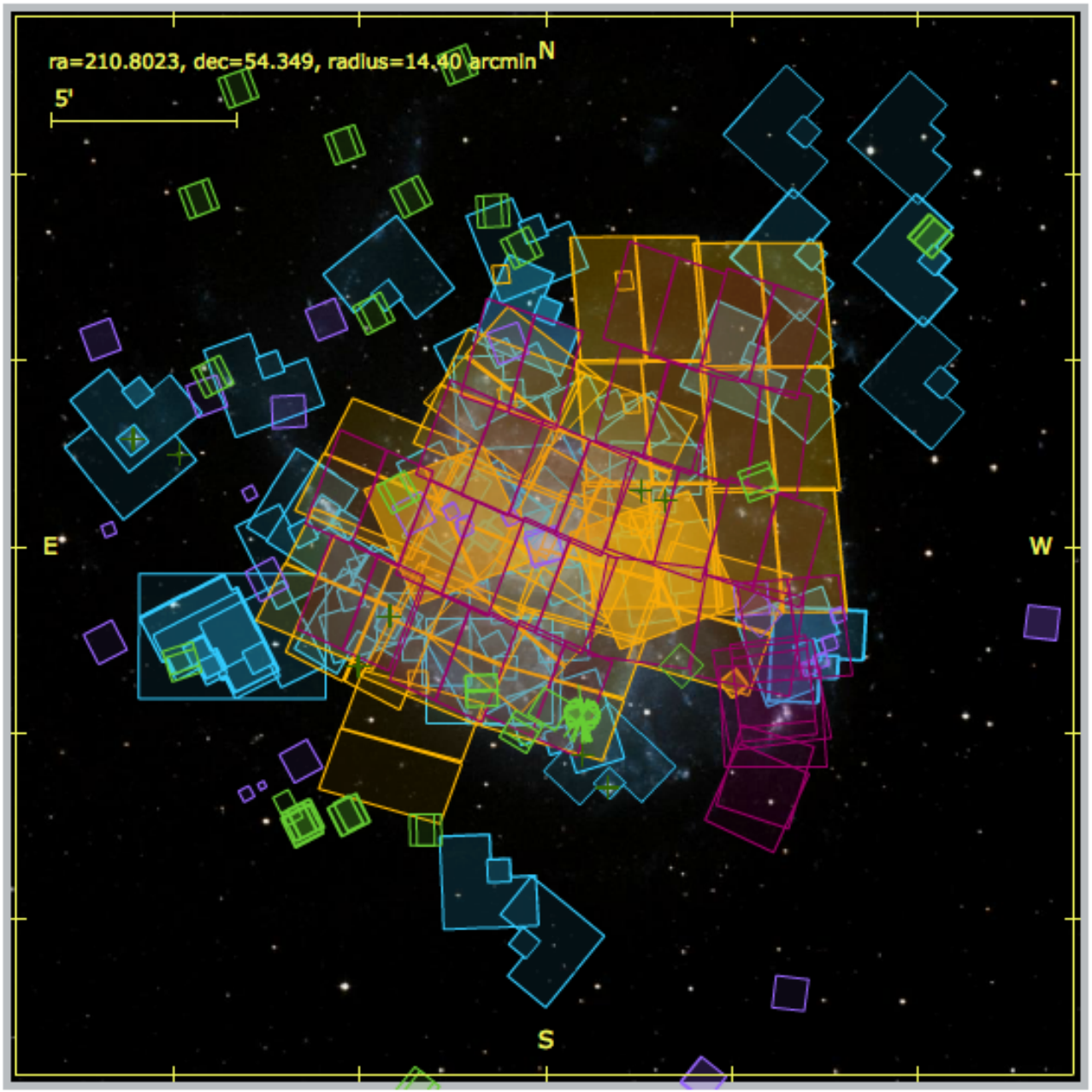
HST Observations of M101

Instruments	#Footprints
ALL	1391
ACS	205
ACSGrism	0
WFPC2	357
WFPC2-PC	349
NICMOS	122
NICGrism	0
WFC3	49
COS	0
STIS	249
FOS	60
GHRS	0

From Hubble Legacy Archive
http://hla.stsci.edu

HST Observations of M101

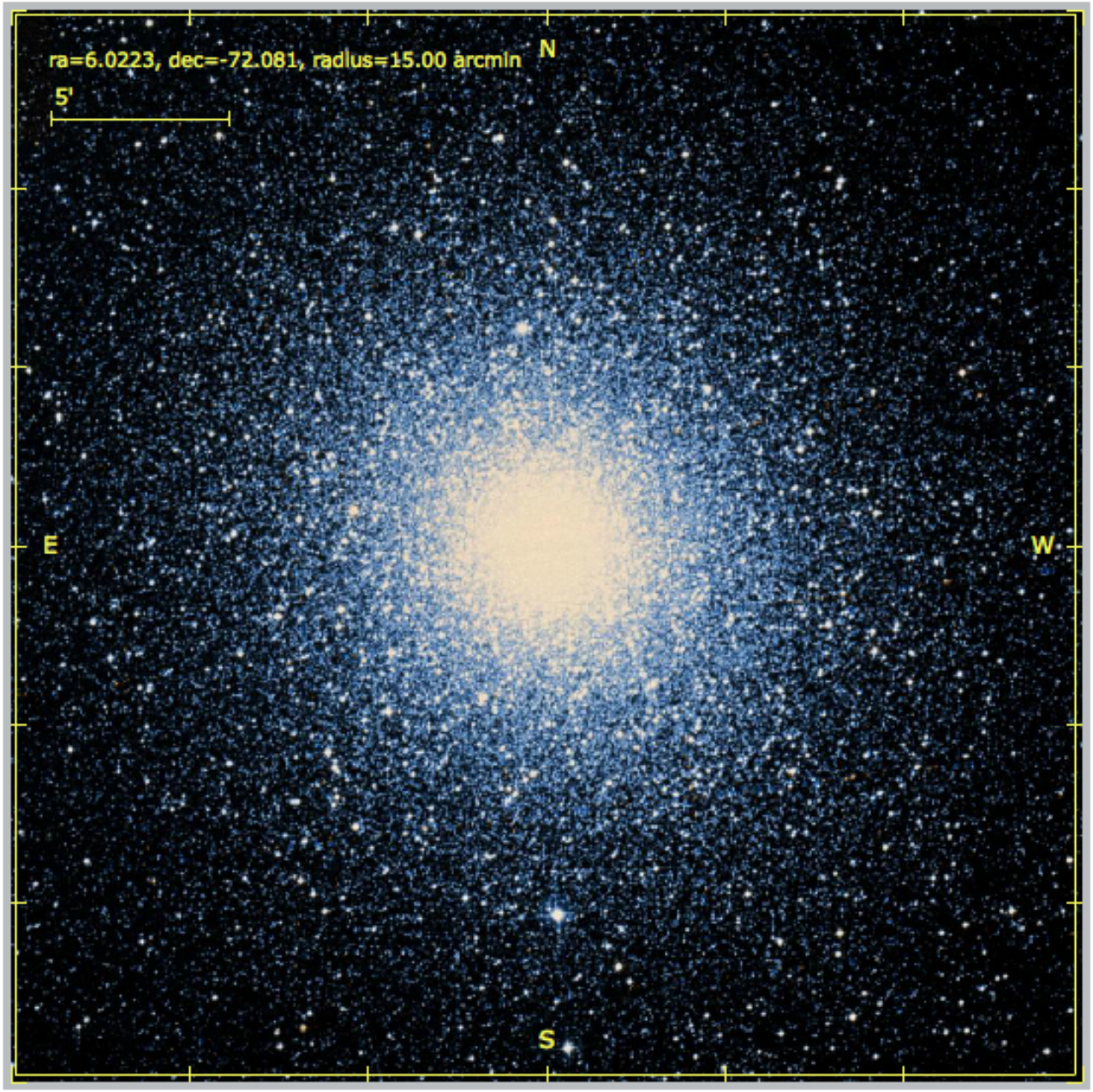
Instruments	#Footprints
☑ ALL	1391
☑ ACS	205
☑ ACSGrism	0
☑ WFPC2	357
☑ WFPC2-PC	349
☑ NICMOS	122
☑ NICGrism	0
☑ WFC3	49
☑ COS	0
☑ STIS	249
☑ FOS	60
☑ GHRS	0



From Hubble Legacy Archive
http://hla.stsci.edu

HST Observations of 47 Tuc

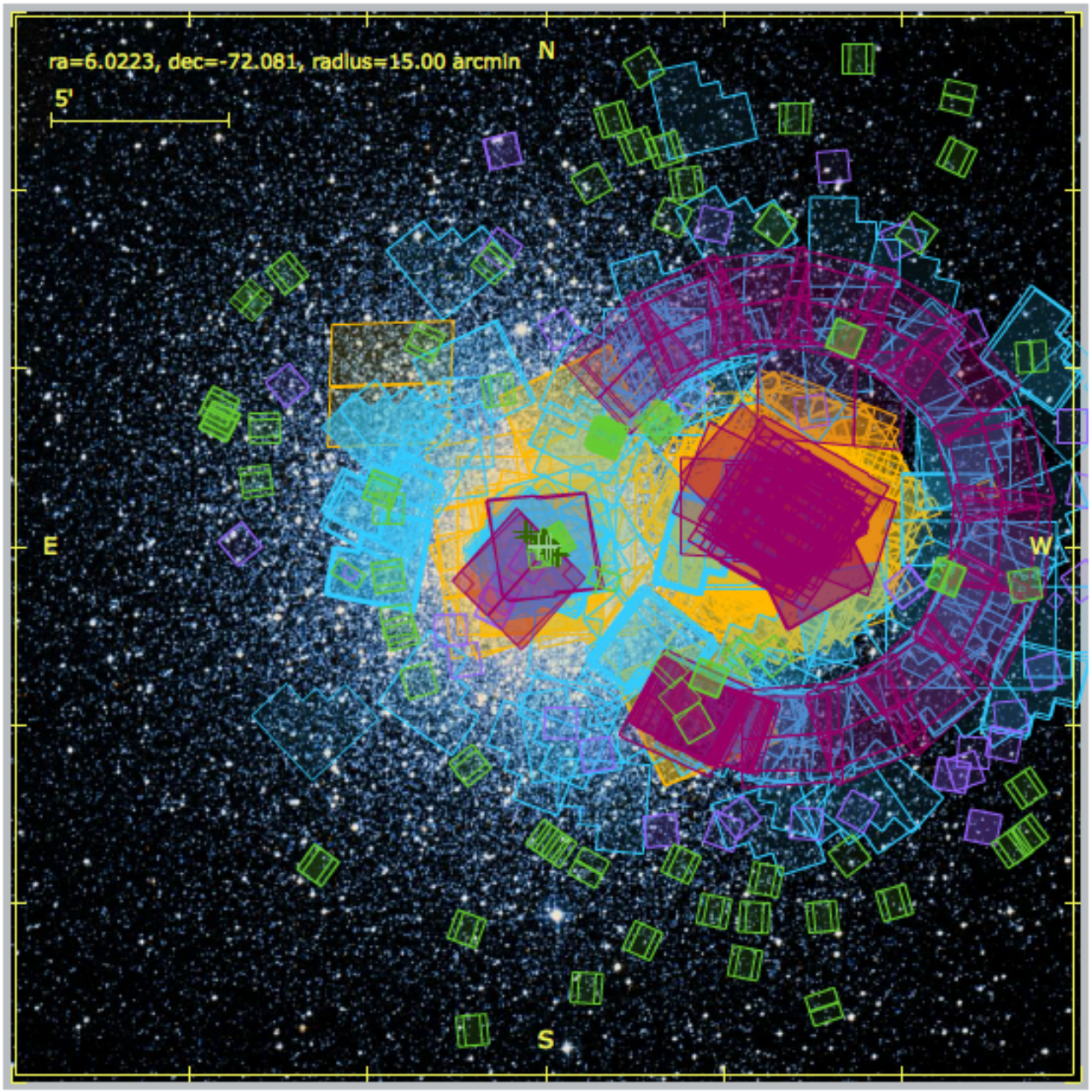
Instruments	#Footprints
ALL	3965
ACS	1768
ACSGrism	0
WFPC2	754
WFPC2-PC	0
NICMOS	139
NICGrism	0
WFC3	361
COS	5
STIS	882
FOS	49
GHRS	7



From Hubble Legacy Archive
http://hla.stsci.edu

HST Observations of 47 Tuc

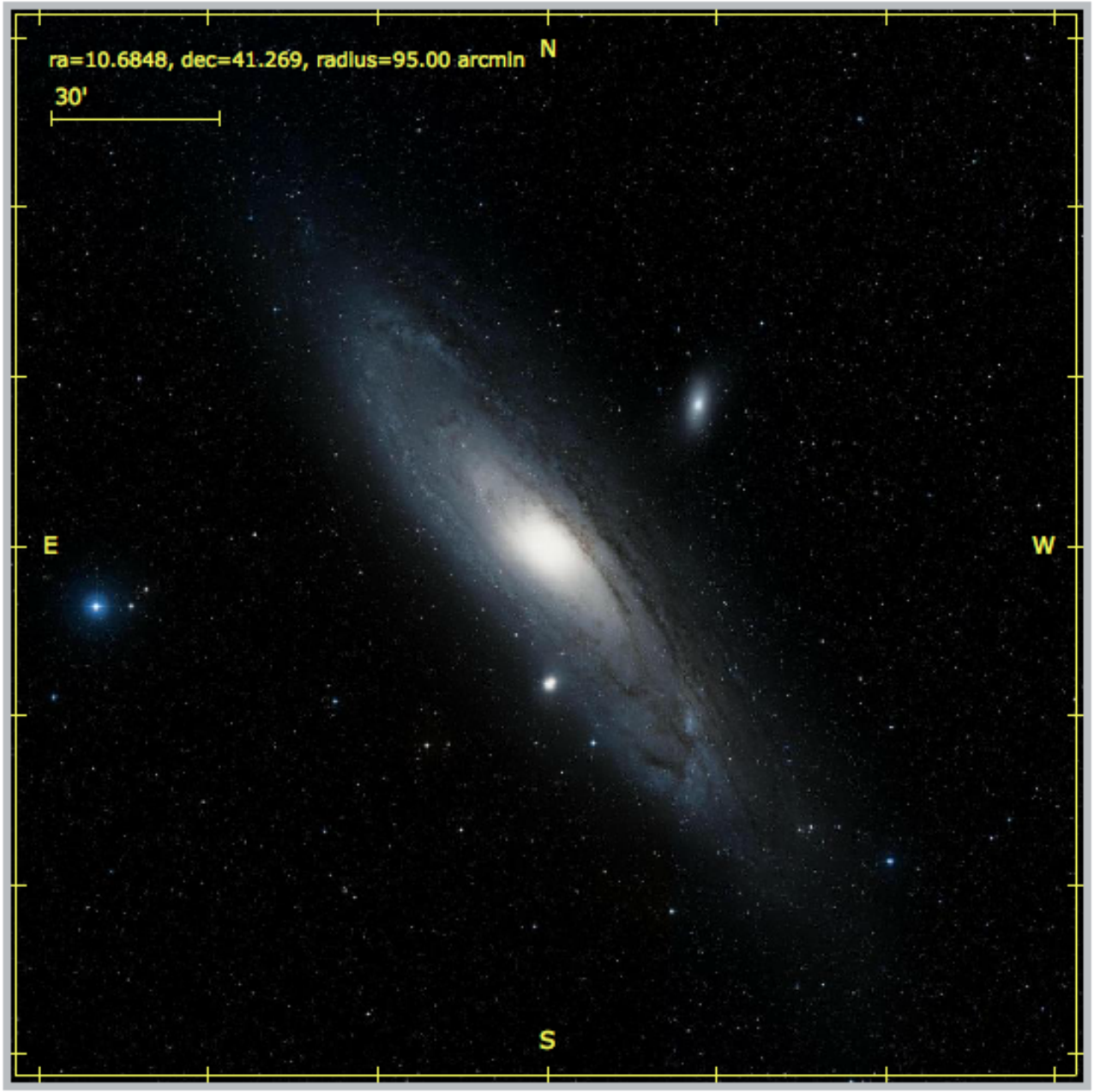
Instruments	#Footprints
☑ ALL	3965
☑ ACS	1768
☑ ACSGrism	0
☑ WFPC2	754
☐ WFPC2-PC	0
☑ NICMOS	139
☑ NICGrism	0
☑ WFC3	361
☑ COS	5
☑ STIS	882
☑ FOS	49
☑ GHRS	7



From Hubble Legacy Archive
http://hla.stsci.edu

HST Observations of M31 (Andromeda)

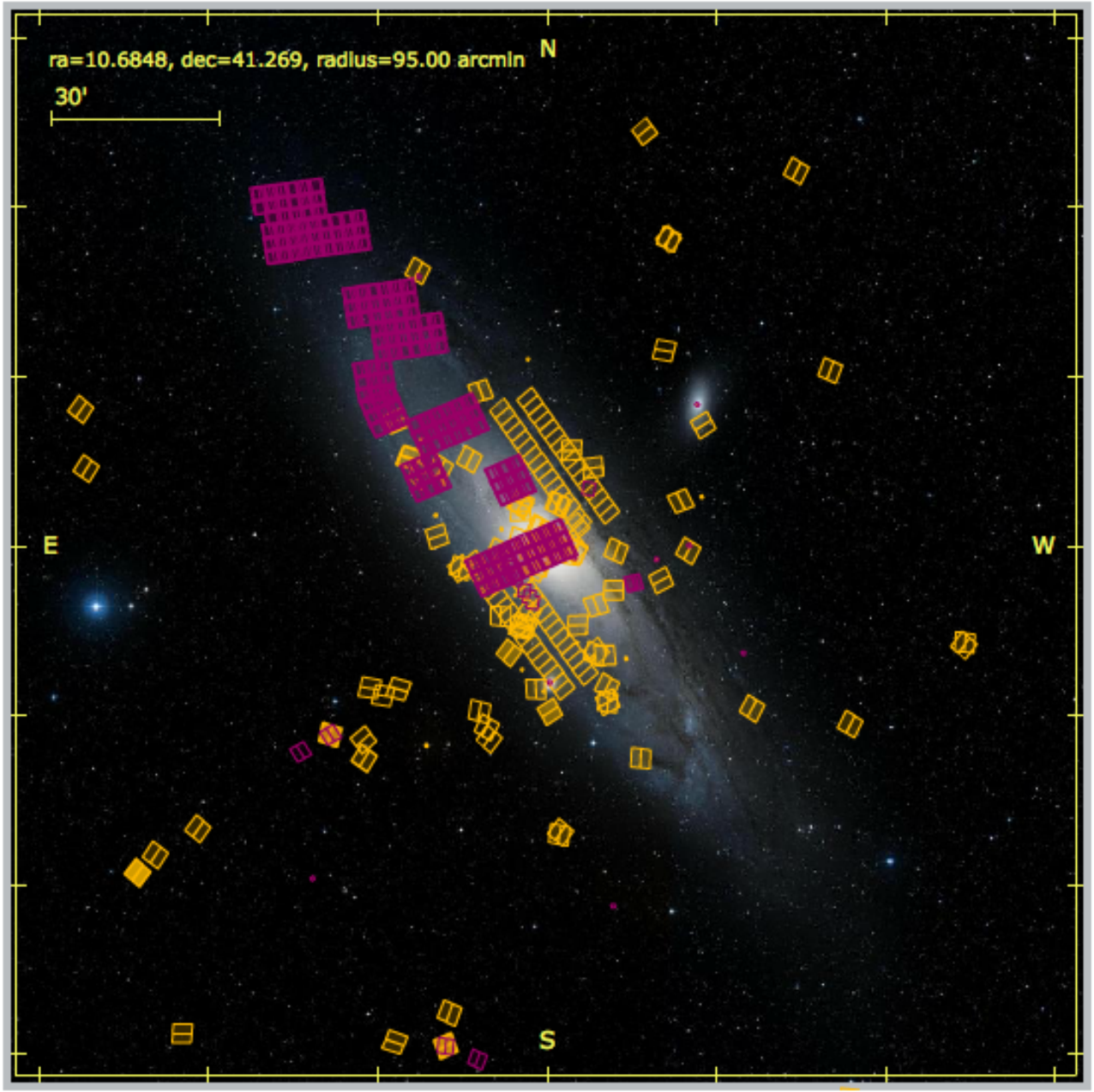
Instruments	#Footprints
ALL	2135
ACS	749
ACSGrism	0
WFPC2	0
WFPC2-PC	0
NICMOS	0
NICGrism	0
WFC3	1386
COS	0
STIS	0
FOS	0
GHRS	0



From Hubble Legacy Archive
http://hla.stsci.edu

HST Observations of M31 (Andromeda)

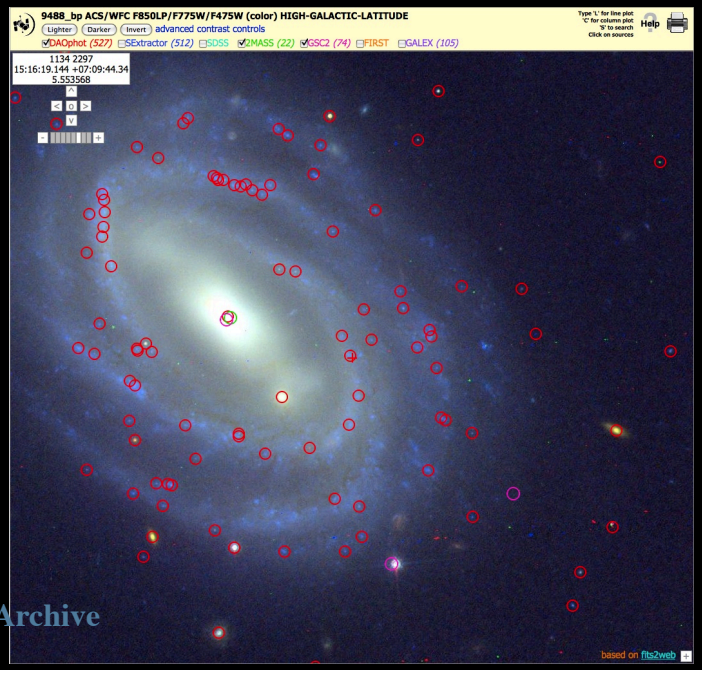
Instruments	#Footprints
☑ ALL	2135
☑ ACS	749
☐ ACSGrism	0
☐ WFPC2	0
☐ WFPC2-PC	0
☐ NICMOS	0
☐ NICGrism	0
☑ WFC3	1386
☐ COS	0
☐ STIS	0
☐ FOS	0
☐ GHRS	0



From Hubble Legacy Archive
http://hla.stsci.edu

STScI Archive & Data Center

- STScI hosts archives and data processing for multiple missions
  - The big active missions: HST, Kepler, GALEX
- HST archive has been in operation since Hubble launch in 1990
  - All HST data are retrieved through the archive
  - Hubble Legacy Archive with enhanced data products open since 2008

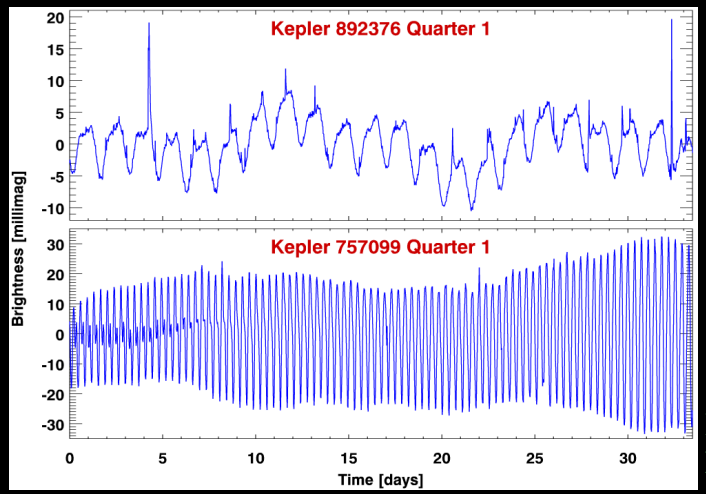


ACS image
& catalogs
Hubble Legacy Archive

M51 ACS

MAST: Mikulski Archive for Space Telescopes

- Formerly the Multi-mission Archive at Space Telescope
  - Established in 1997 as NASA's Optical/UV archive
  - Supports both active (HST/HLA, GALEX, Kepler, XMM-OM) and legacy missions (IUE, FUSE, EUVE, …)
  - The other NASA archive centers:
    - HEASARC (GSFC): X-ray, gamma ray
    - IRSA/IPAC/NED/… (Caltech/JPL): Infrared
    - ADS (CfA): Astronomical literature



Kepler light curves

What is the STScI archive?

- Data

  ~185 TB of images, spectra, catalogs, time series

- Metadata

  ~10^6 HST observations (plus other missions)

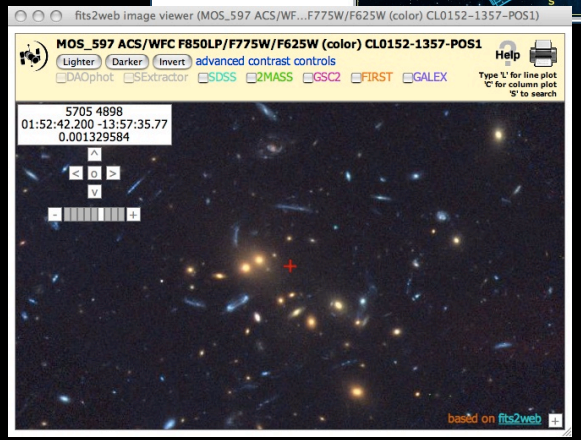
  Documentation, publication links, …

- User interfaces

  Search, browse, plot, explore

  Browser-based interfaces

  Help desk/User support




```
<VOTABLE>
<DESCRIPTION>STScI Hubble Legacy Archive SIAP</DESCRIPTION>
  <INFO name="QUERY_STATUS" value="OK"></INFO>
  <RESOURCE type="results">
    <PARAM datatype="char" name="INPUT:POS" value="210.802458,54
    <PARAM datatype="double" name="INPUT:SIZE" value="0.240000"
    <PARAM datatype="char" name="INPUT:FORMAT" value="FITS" arra
    <PARAM datatype="char" name="INPUT:imagetype" value="best" a
    <PARAM datatype="char" name="INPUT:inst" value="acs,wfpc2,ni
    <PARAM datatype="int" name="INPUT:hrcmatch" value="0"></PARA
    <PARAM datatype="double" name="INPUT:zoom" value="1.000000">
    <PARAM datatype="double" name="INPUT:autoscale" value="99.50
    <PARAM datatype="int" name="INPUT:asinh" value="1"></PARAM>
    <PARAM datatype="char" arraysize="*" name="refframe" ucd="VO
    <PARAM datatype="char" arraysize="*" name="projection" ucd="
    <TABLE>
```

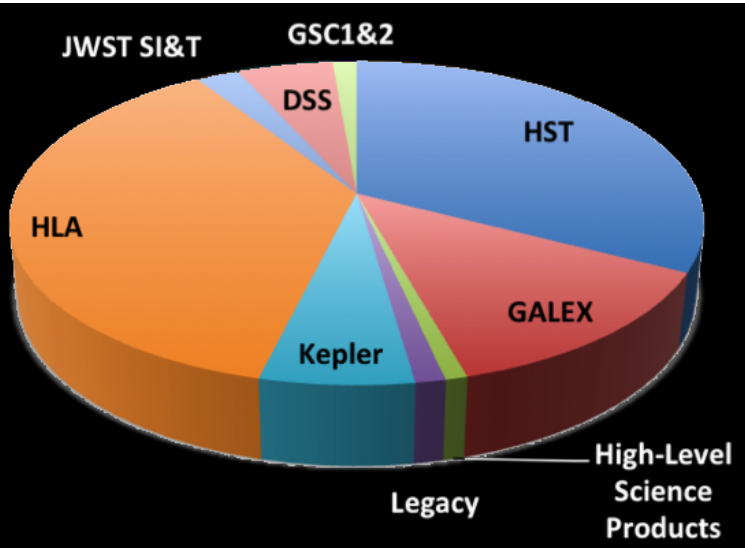

- Services

  VO services, data retrieval, image cutouts, …

  (UIs are built around VO services)

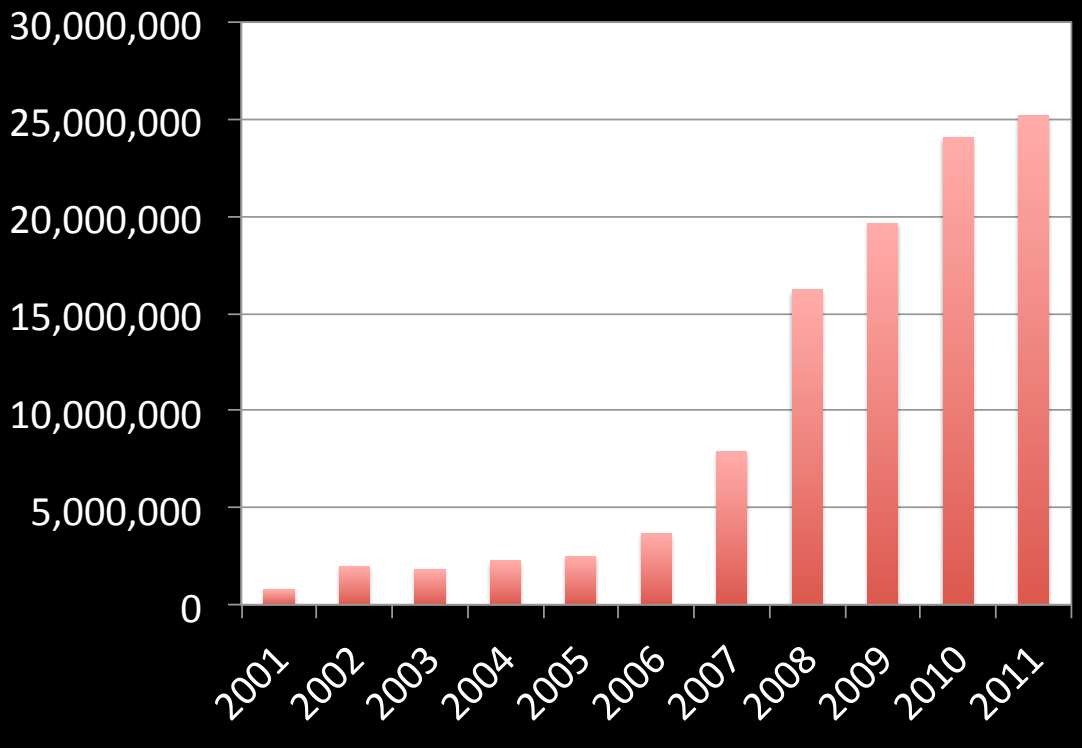
The MAST Archive: 2 minute summary

- ~ 185 TBytes (62 TB HST, 79 TB HLA)
- Ingest rate: > 25 TB/yr
- Retrievals: > 100 TB/yr
  - Distributed volume ~4x ingest

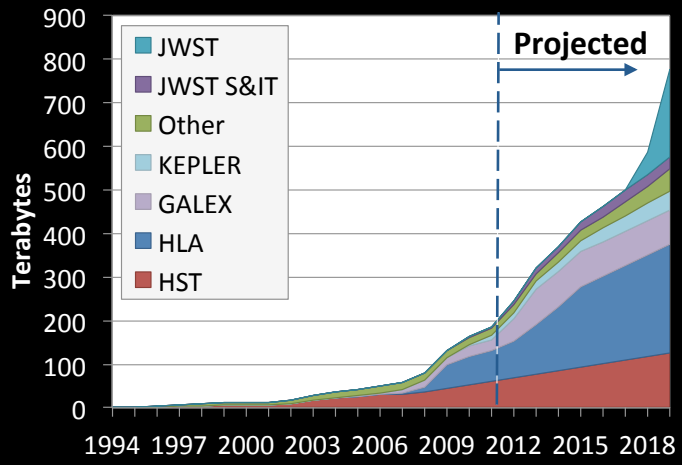


Data holdings

Number of searches per year

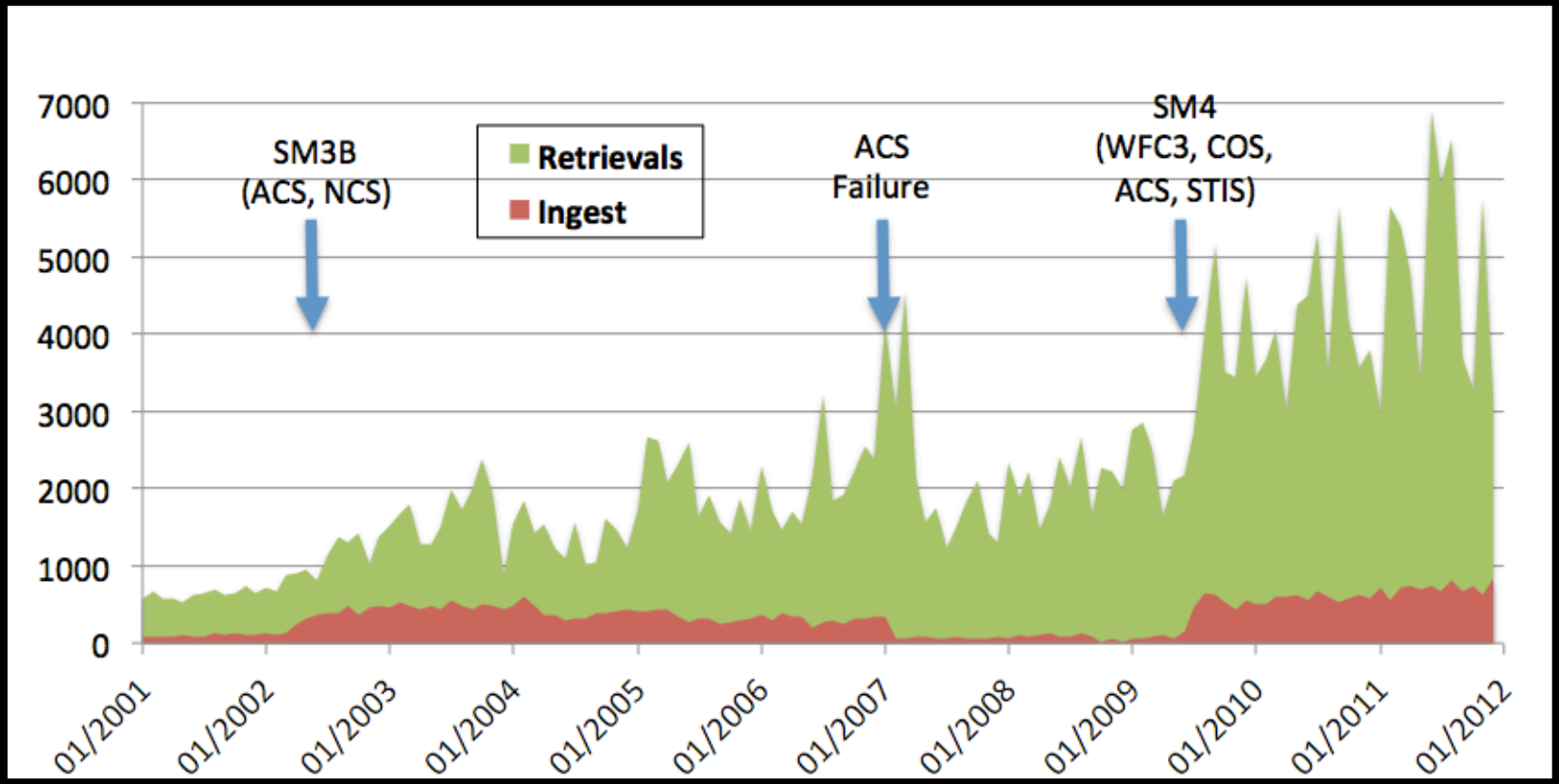


Past & projected volume

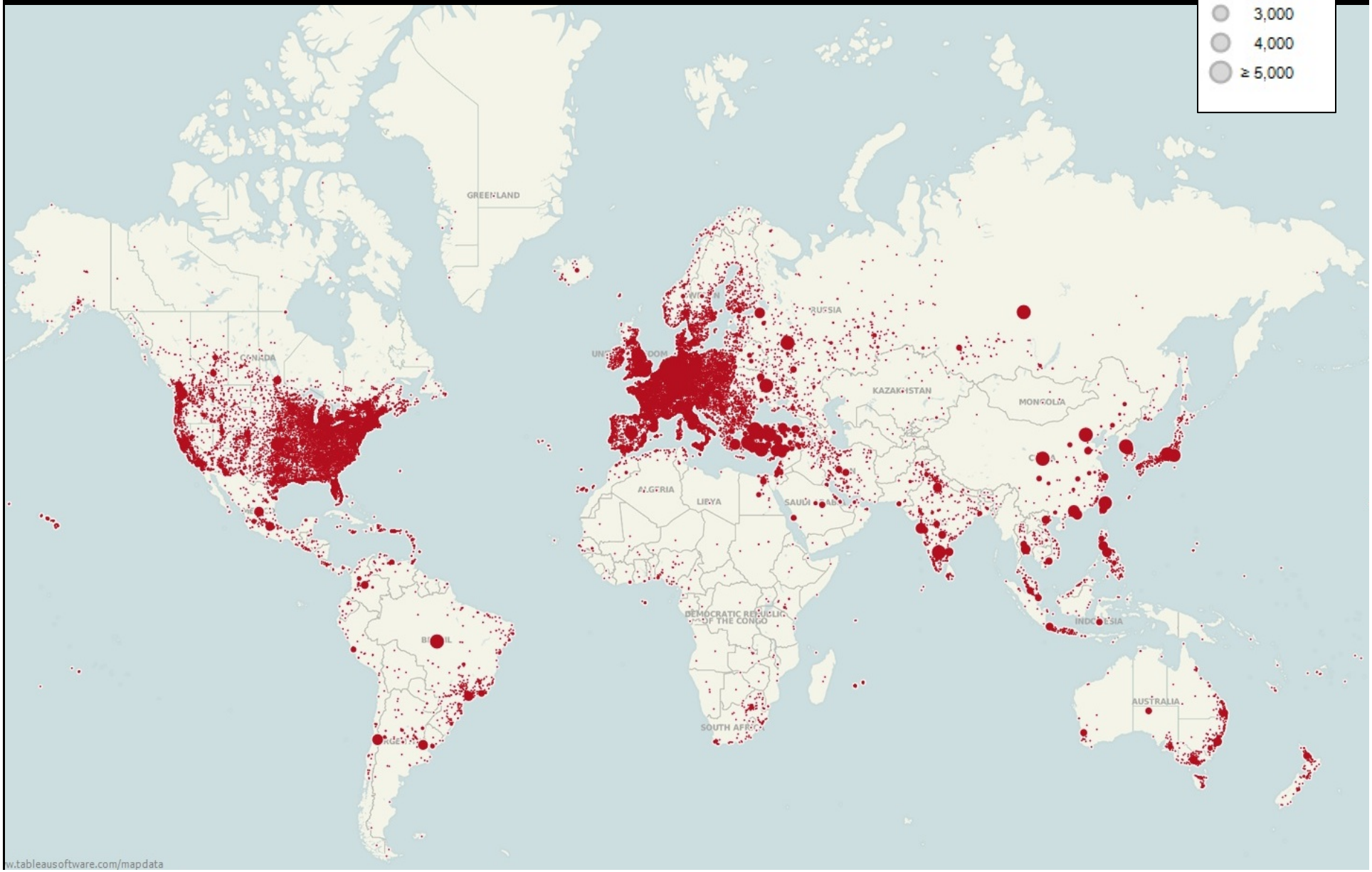


HST Archival Data Demand is Increasing

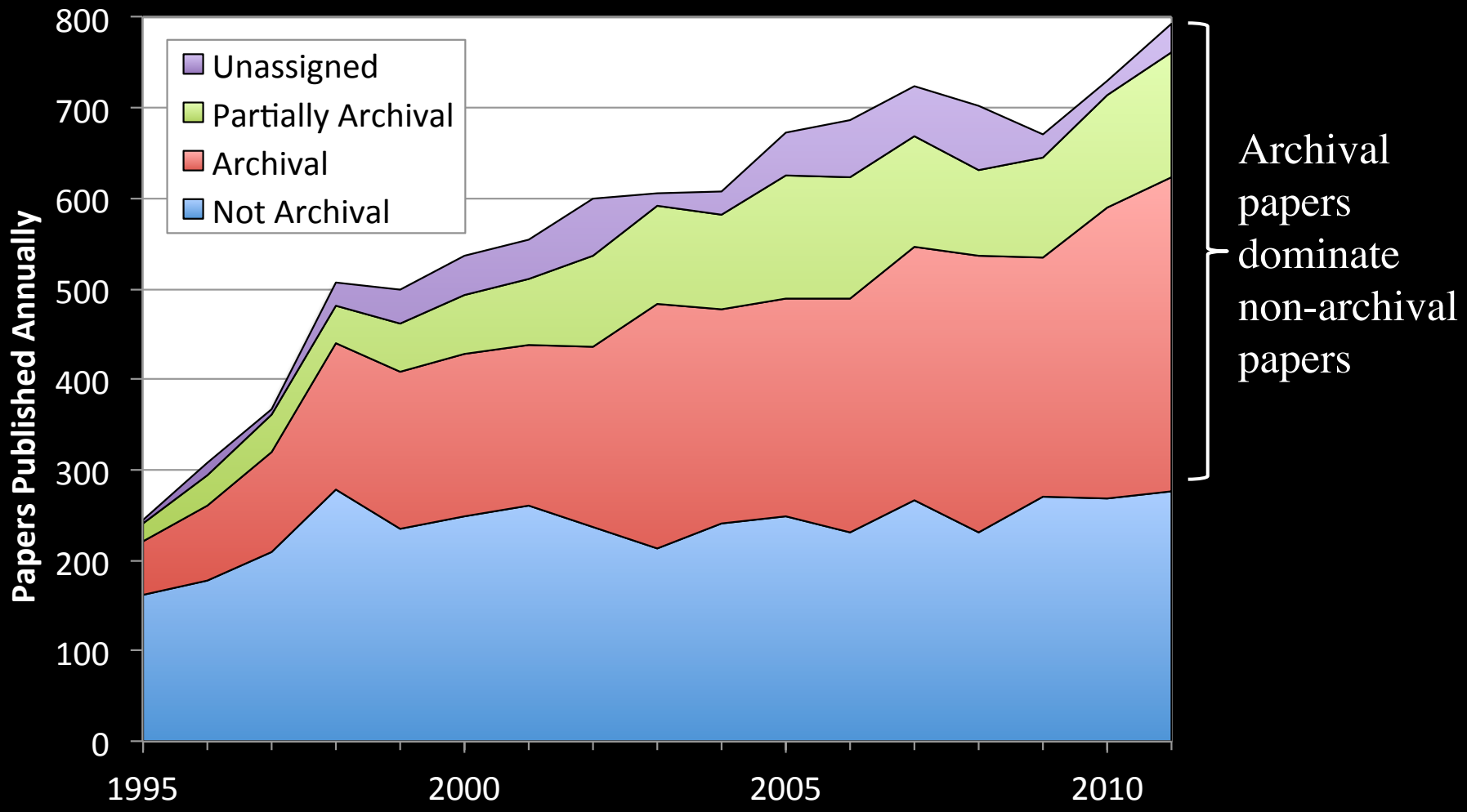
- HST archive retrievals doubled after Servicing Mission 4
- >10,000 registered archive users (85 countries)



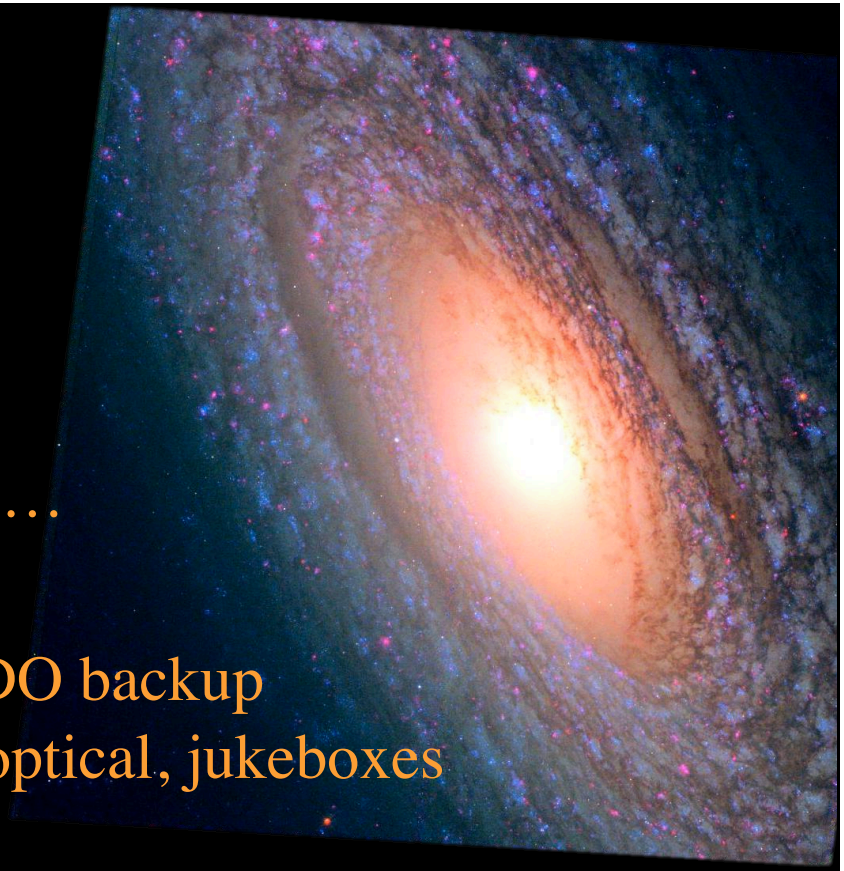
Location of Identifiable IP Addresses



HST Publication Statistics



Key Archive Technologies



NGC 2841
WFC3

- Databases
  - Current: MS SQL Server
  - Past: Sybase, Objectivity, …
- Data Storage
  - Current: Magnetic disks + UDO backup
  - Past: Optical & magneto-optical, jukeboxes
- Data delivery
  - Current: Almost 100% Internet
  - Past: 9-track tape, Exabyte tape, CD-ROM, DVD
- User interfaces & services
  - Current: Browser/HTTP-based: Javascript, PHP, Flash, …
  - Past: Custom C/C++/Java applications, paper forms!

1990

- There is no Internet – data on tapes in the mail
- Disks cost $10 million per terabyte
- Archive computers are DEC VAX/VMS

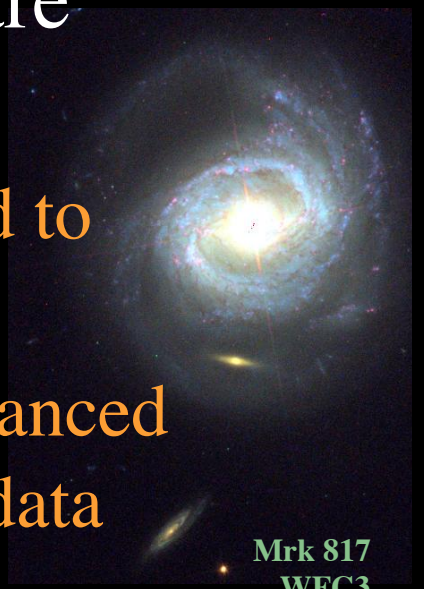
2012

- Nearly all data are delivered via Internet
- Disks cost $100 per terabyte
- Archive computers are many-core Linux systems

Data Product Generation

The archive is not a collection of static data.

- Data processing is continuously improving
- New "science-ready" data products are being created:
  - GALEX raw data are being reprocessed to generate 50 TB photon database
  - Hubble Legacy Archive is creating enhanced images and source catalogs from HST data
    - http://hla.stsci.edu



Summary

- Surveys are the equivalent of experiments for the observational science of astronomy
  - Catalogs are essential both as data products and for (self-) calibration
  - SDSS is the model for future surveys: rapid public access to high quality products with great tools
- Observatory archives are highly heterogeneous compared with surveys
  - Very heavily used & highly productive
    - Archival science will dominate ultimate Hubble science
  - Technologies: databases, storage, interfaces