SPACE TELESCOPE SCIENCE INSTITUTE

BARBARA A. MIKULSKI ARCHIVE FOR SPACE TELESCOPES
MAST

Maximizing the Science in the Era of Data-Driven Astronomy

Alessandra Aloisi
(Program Manager, STScI Data Management)

"Big Questions" in Astronomy and Astrophysics

- USA National Academy of Sciences published in 2010 the decadal survey of Astronomy and Astrophysics
  - the Astro2010 Survey "New Worlds, New Horizons in Astronomy and Astrophysics" identifies the big science questions in astronomy and astrophysics for the decade 2012-2021
  - prioritizes the investments needed
  - recommends a vital and timely scientific program with a balance of small, medium, and large initiatives on the ground and in space
- "Big Questions" to be answered divided into 3 categories:

NEW WORLDS

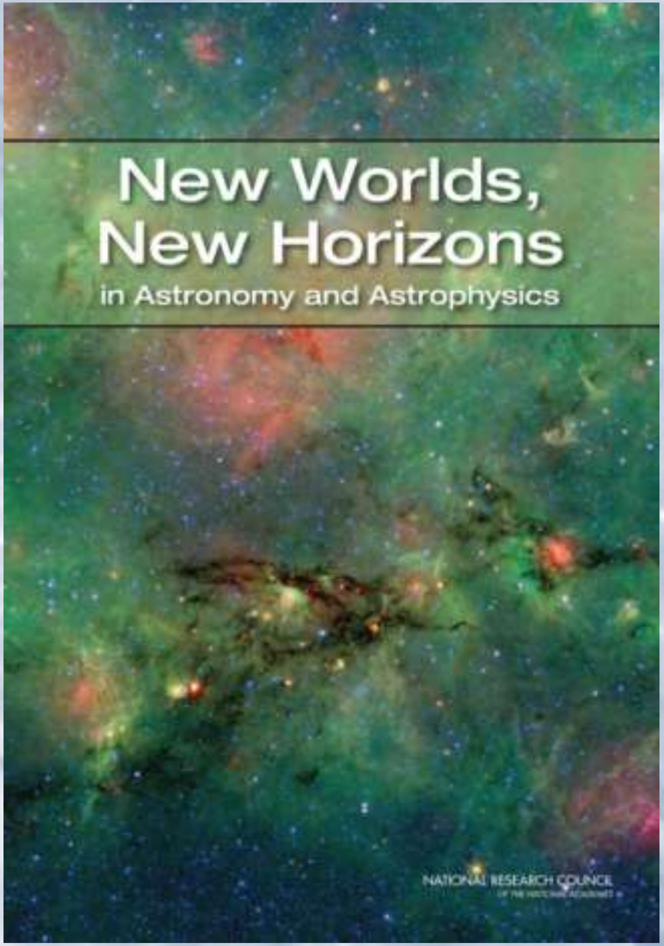
What are planetary systems like?
How do Stars and Planets Form?

FUNDAMENTAL PHYSICS

What happens when stars die?
What are Black Holes?
How Can We Detect Gravitational Waves? What Can They Tell Us?
What are Dark Matter and Dark Energy
What goes on inside Galaxies

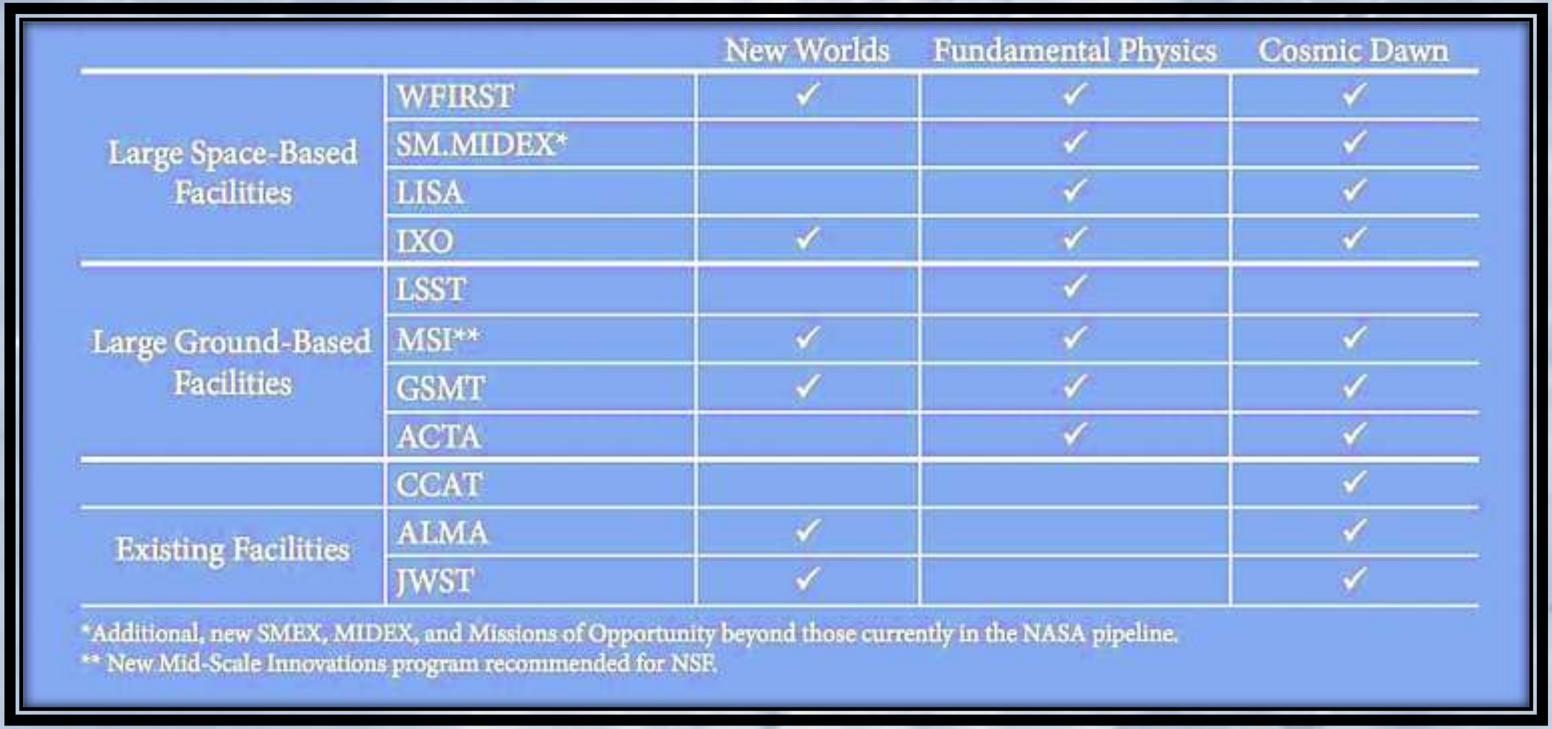
COSMIC DAWN

What Causes Cosmic Inflation?
What Objects First Lit Up The Universe, and When?
How Has The Universe Evolved Over Time?



Telescopes, Instruments, & Programs recommended by Astro2010

In this decade Astronomers are poised to achieve major advances in answering these questions through access to facilities at the forefront of astronomical research:



		New Worlds	Fundamental Physics	Cosmic Dawn
Large Space-Based Facilities	WFIRST	✓	✓	✓
	SM.MIDEX*		✓	✓
	LISA		✓	✓
	IXO	✓	✓	✓
Large Ground-Based Facilities	LSST		✓	
	MSI**	✓	✓	✓
	GSMT	✓	✓	✓
	ACTA		✓	✓
	CCAT			✓
Existing Facilities	ALMA	✓		✓
	JWST	✓		✓

*Additional, new SMEX, MIDEX, and Missions of Opportunity beyond those currently in the NASA pipeline.

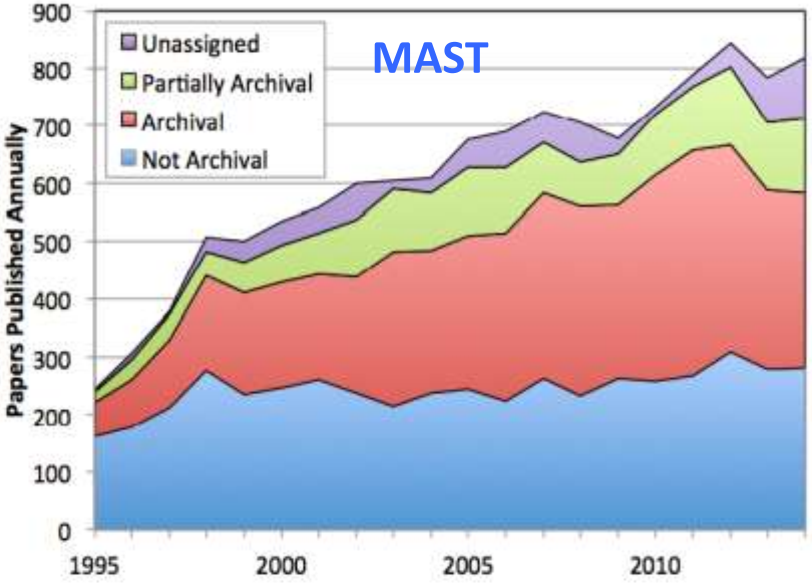
** New Mid-Scale Innovations program recommended for NSF.

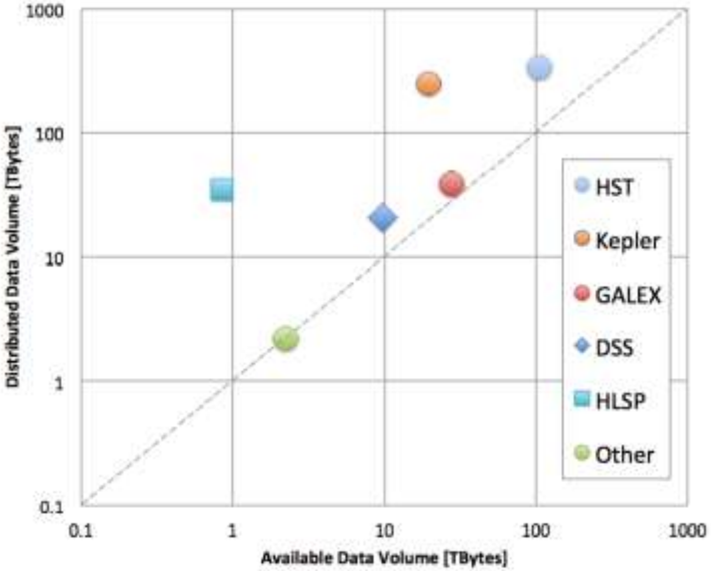
A New Data-Driven Era in Astronomy

Many of the programs recommended will:

- need to hit the ground running because of the limited lifetime of the project (e.g., JWST)
- create massive databases that will be mined for decades (e.g., LSST, WFIRST)
- Produce complex and high-volume databases requiring special tools (e.g. ALMA)

Value of Archival Science & High-Level Science Products (HLSPs)

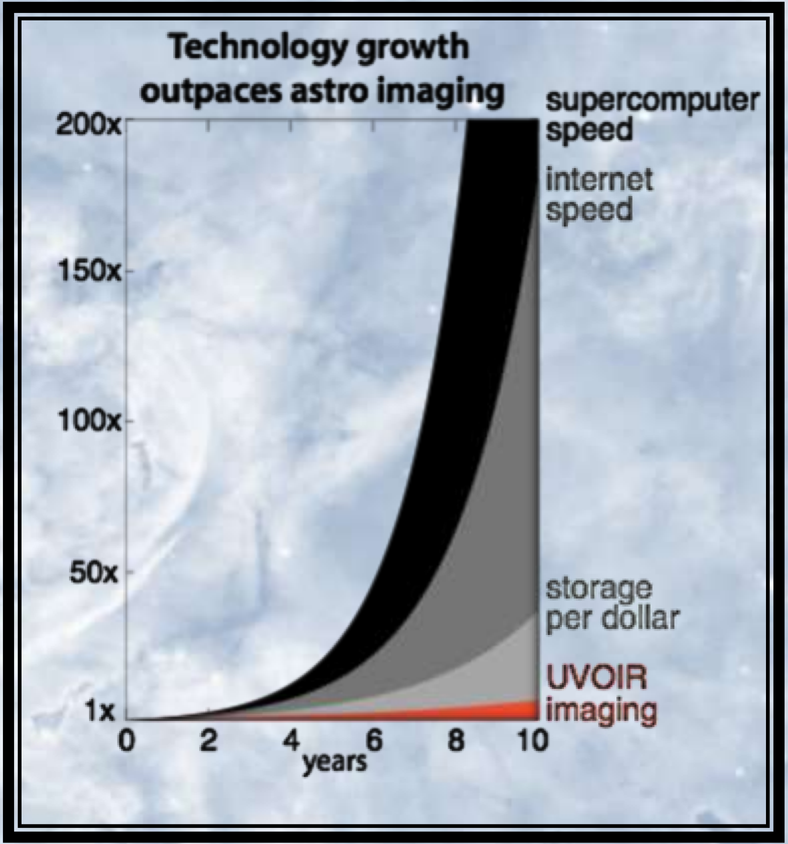




- 60% of MAST papers based, in whole or in part, on archival (HST) data!
- Missions with HLSP in MAST (e.g., HST and Kepler) show a distribution in volume ~ 2 mag higher than available data volume.
- Kepler HLSPs are mostly light curves.
- HST HLSPs are mostly products produced by the community (e.g., multi-cycle treasury programs) or HLA (imaging) products.

Archives as new Scientific Opportunity

- Very large and/or highly complex datasets will come on line over the next several years
- Archives will offer an untapped scientific opportunity, particularly when different data collections will be simultaneously available
- Move research from a small-size sample to a full population



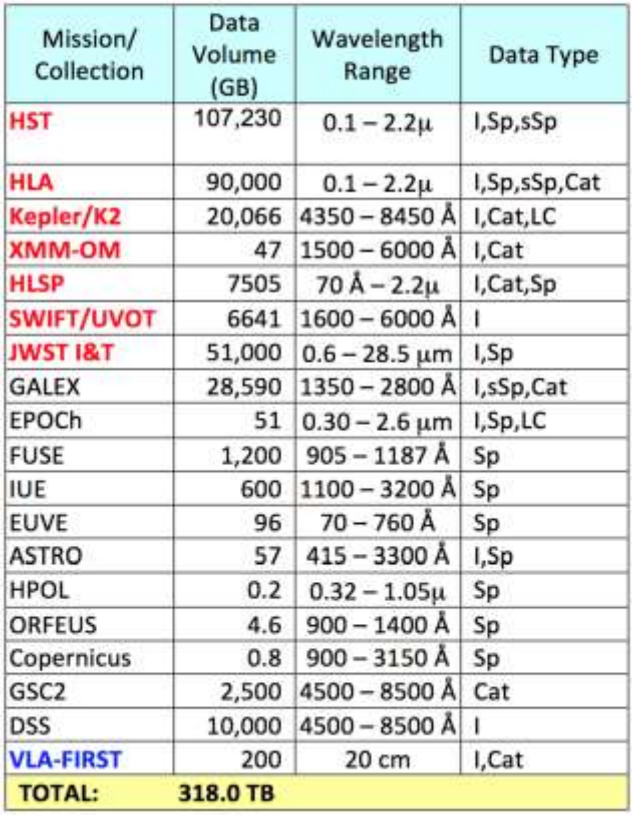
Courtesy J.Peek

Challenge: How do we take full advantage of this new Astronomical revolution ?

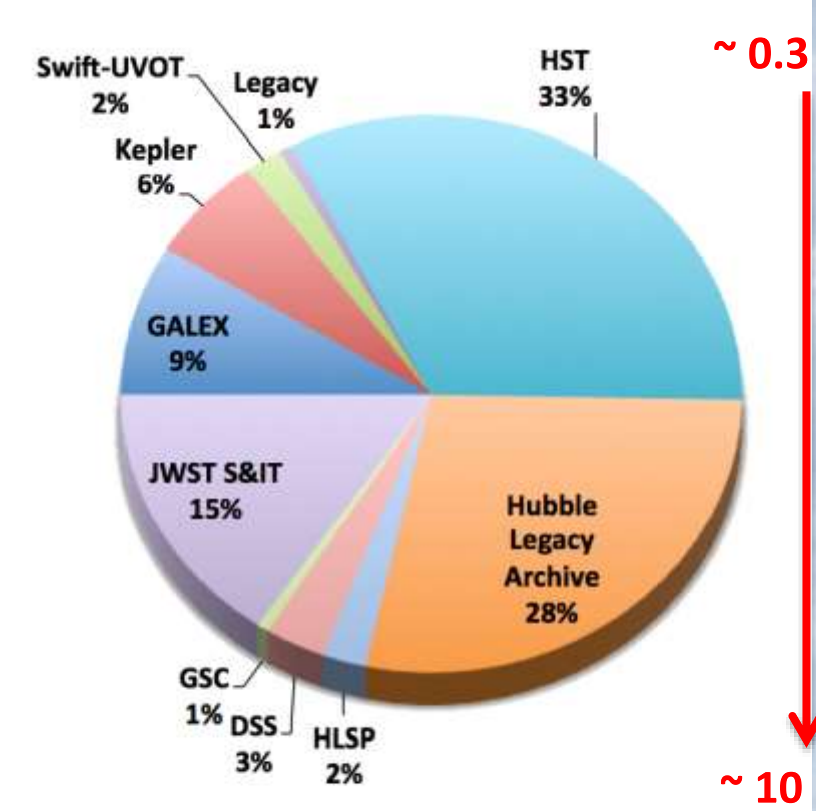
MAST Current and Future Data Collections

FUV

MIR



Mission/ Collection	Data Volume (GB)	Wavelength Range	Data Type
HST	107,230	0.1 – 2.2μ	I,Sp,sSp
HLA	90,000	0.1 – 2.2μ	I,Sp,sSp,Cat
Kepler/K2	20,066	4350 – 8450 Å	I,Cat,LC
XMM-OM	47	1500 – 6000 Å	I,Cat
HLSP	7505	70 Å – 2.2μ	I,Cat,Sp
SWIFT/UVOT	6641	1600 – 6000 Å	I
JWST I&T	51,000	0.6 – 28.5 μm	I,Sp
GALEX	28,590	1350 – 2800 Å	I,sSp,Cat
EPOCh	51	0.30 – 2.6 μm	I,Sp,LC
FUSE	1,200	905 – 1187 Å	Sp
IUE	600	1100 – 3200 Å	Sp
EUVE	96	70 – 760 Å	Sp
ASTRO	57	415 – 3300 Å	I,Sp
HPOL	0.2	0.32 – 1.05μ	Sp
ORFEUS	4.6	900 – 1400 Å	Sp
Copernicus	0.8	900 – 3150 Å	Sp
GSC2	2,500	4500 – 8500 Å	Cat
DSS	10,000	4500 – 8500 Å	I
VLA-FIRST	200	20 cm	I,Cat
TOTAL:	318.0 TB		



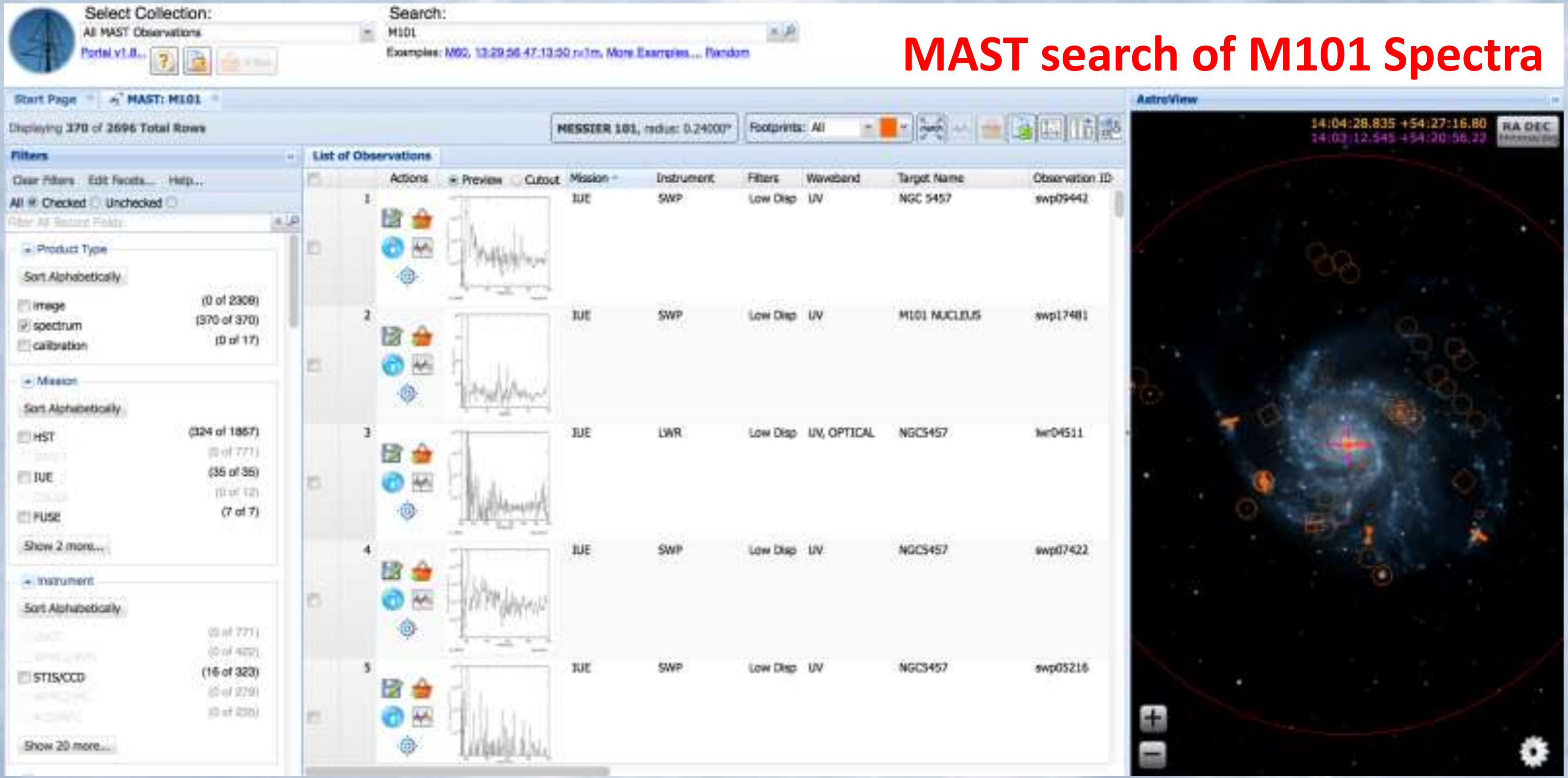
~ 0.3 PB

~ 10 PB

- MAST includes data from active (**RED**) and legacy (**BLACK)** missions covering the whole spectral range from FUV to MIR. Non-NASA funded projects are also included (**BLUE)**.
- Data volume of current MAST holdings is ~ 300 TB. This will significantly grow into the PB scale over the next several years due to upcoming new missions that MAST will support.
- JWST data product mission 10-year baseline is ~ 1PB (not including working datasets or HLSP).
- MAST is playing or will be playing a significant role in the following ongoing/future missions: **PanSTARRs** (~ 2 PB), **TESS** (~ 20 TB), **WFIRST/AFTA** (3-9 PB), and **GAIA** (US-Mirror site; total of 300 TB by 2022).

MAST Discovery Portal

1. Unify MAST missions with a common discovery interface.
2. Provide instant access to Virtual Observatory collections.
3. Build a framework for astronomy data interchange.



MAST search of M101 Spectra

Select Collection: All MAST Observations

Search: M101

Examples: M60, 13:29:56 47:13:50 r=1m, More Examples... Random

Portal v1.8...

Start Page | MAST: M101

Displaying 370 of 2696 Total Rows

MESSIER 101, radius: 0.24000°

Footprints: All

Filters

Clear Filters Edit Facets... Help...

All ○ Checked ○ Unchecked ○

Filter All Record Fields

Product Type

Sort Alphabetically

- □ image (0 of 2308)
- ☑ spectrum (370 of 370)
- □ calibration (0 of 17)

Mission

Sort Alphabetically

- □ HST (324 of 1867)
- [illegible] (0 of 771)
- □ IUE (35 of 35)
- [illegible] (0 of 12)
- □ FUSE (7 of 7)

Show 2 more...

Instrument

Sort Alphabetically

- [illegible] (0 of 771)
- [illegible] (0 of 422)
- □ STIS/CCD (16 of 323)
- [illegible] (0 of 279)
- [illegible] (0 of 235)

Show 20 more...

List of Observations

	Actions	● Preview ○ Cutout	Mission	Instrument	Filters	Waveband	Target Name	Observation ID
1			IUE	SWP	Low Disp	UV	NGC 5457	swp09442
2			IUE	SWP	Low Disp	UV	M101 NUCLEUS	swp17481
3			IUE	LWR	Low Disp	UV, OPTICAL	NGC5457	lwr04511
4			IUE	SWP	Low Disp	UV	NGC5457	swp07422
5			IUE	SWP	Low Disp	UV	NGC5457	swp05216

AstroView

14:04:28.835 +54:27:16.80

14:03:12.545 +54:20:56.22

Do not miss T. Donaldson Demo 5 !

MAST and the Partner Archives

- Partnerships with other Archives is key:
  - It establishes common data interchange models:
    - CAOM (Common Archive Observation Model)
      - originally created by CADC
      - fully adopted for ingestion of all STScI supported missions into MAST
      - under consideration/implementation at ESAC & IPAC
    - VO standards via the NASA-VO, USA-VO, and IVOA collaborations

      McGlynn poster P071

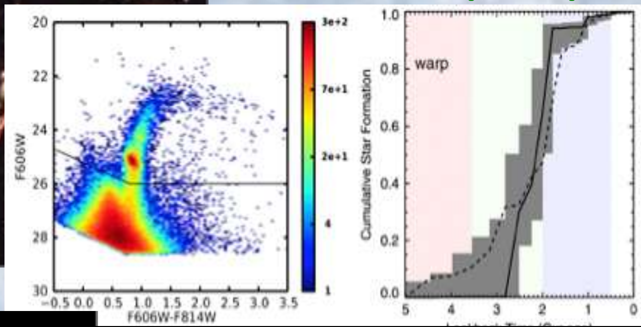
      Arviset talk O11.3
  - It allows for technology exchanges that facilitate data discovery and mining:
    - complex non-positional searches
    - VO registry
    - TAP services
    - Indexing
    - Astrotag project
    - MARC/DOI initiative

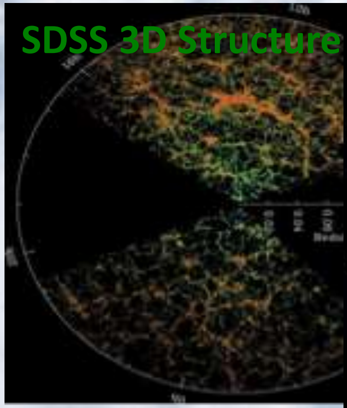
New Science Drivers @ STScI

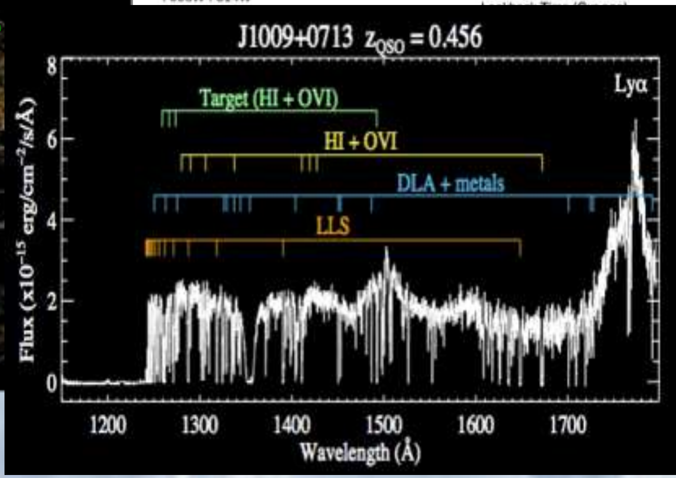
- STScI is going through process to identify new science opportunities with present and future MAST data holdings
  Postman et al. 2015, White Paper
- Several science cases identified that tap into STScI scientific areas of expertize:
  1. Automated Identification of Gravitationally Lensed Galaxies
  2. Classification of Amorphous Sources (e.g., star clusters, galaxies, etc.)
  3. Resolved Stellar Populations and Star-Formation Histories
  4. Mapping the Cosmos in 3D with Multi-wavelength Data
  5. Black Hole and Host Galaxy Co-Evolution
  6. Time-Domain Astronomy
  7. Multi-dimensional Exploration of Spectroscopic Datasets

Xu et al. 2015

Williams et al. 2015 (M31)




SDSS 3D Structure



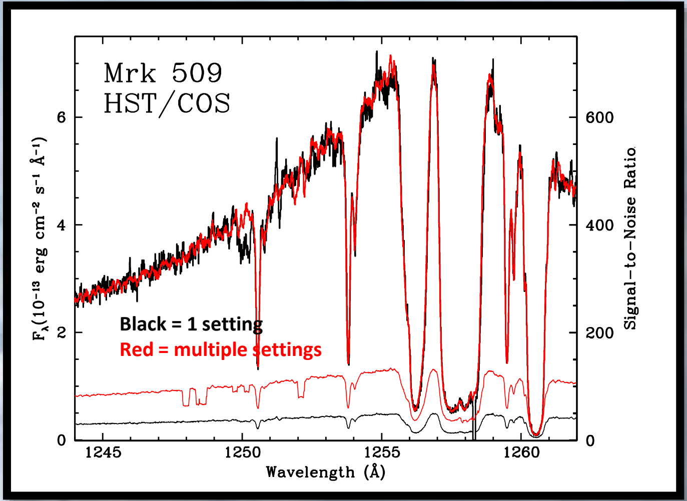
Tumlinson et al. 2013

Common Threads Identified

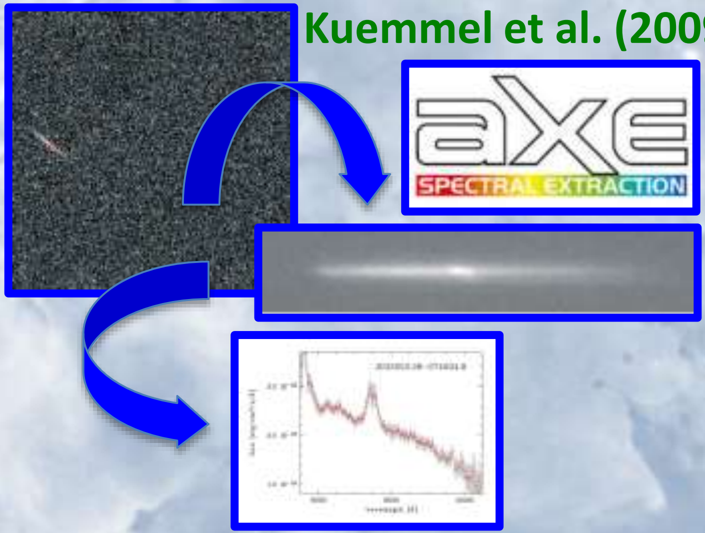
- *High-level science products*
- *Capability to create and execute data analysis processes workflows*
- *High-performance computing*
- *Multi-dimensional data visualization tools for data discovery*
- *Publicly available and/or open source software tools for data reduction and data analysis*
- Automated detection/classification/recognition algorithms
  Liang talk O2.5
  Hampton talk O8.3
- Machine learning tools
  Durrent-Whyte talk O1.1

High-Level Science Products @ STScI (1)

- Several ongoing HST projects at STScI to create new HLSPs:
  - HLSPs already exist for HST imaging through HLA
  - Hubble Source Catalog (HSC) v.1 released in Winter 2015 **Whitmore talk O13.1**
  - HLSPs for HST spectroscopy currently under development and implementation
- HST+JWST+WFIRST working Group at STScI currently investigating new algorithms to perform optimal extraction of grisms (MOS) spectra



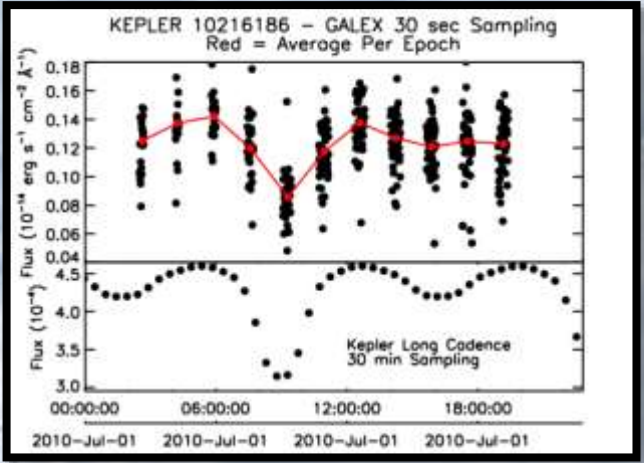
Kriss et al. (2011)



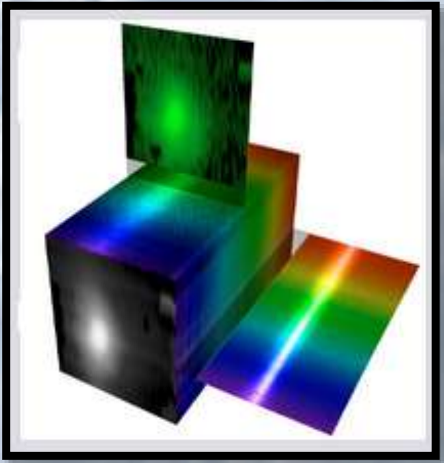
Kuemmel et al. (2009)

High-Level Science Products @ STScI (2)

- Kepler data and HLSPs (light curves) distributed by MAST
- New gPhoton database of GALEX time-tagged photon events released to the public in Summer 2015 with software to create light curves, data cubes and images
- JWST will produce HLSPs as part of regular science operations. These include, but are not limited to:
  - Mosaics/dithers
  - Time series/light curves
  - photometric, astrometric, and morphological source catalogs
  - single and multichannel IFU 3D data cubes
  - 2D maps derived from IFU data cubes (e.g., dynamical moment maps of intensity, relative velocity, and line widths)
  - 1D spectral extractions of MSA/grisms and IFU observations and combination of spectra taken with different gratings/channels



Courtesy S. Fleming



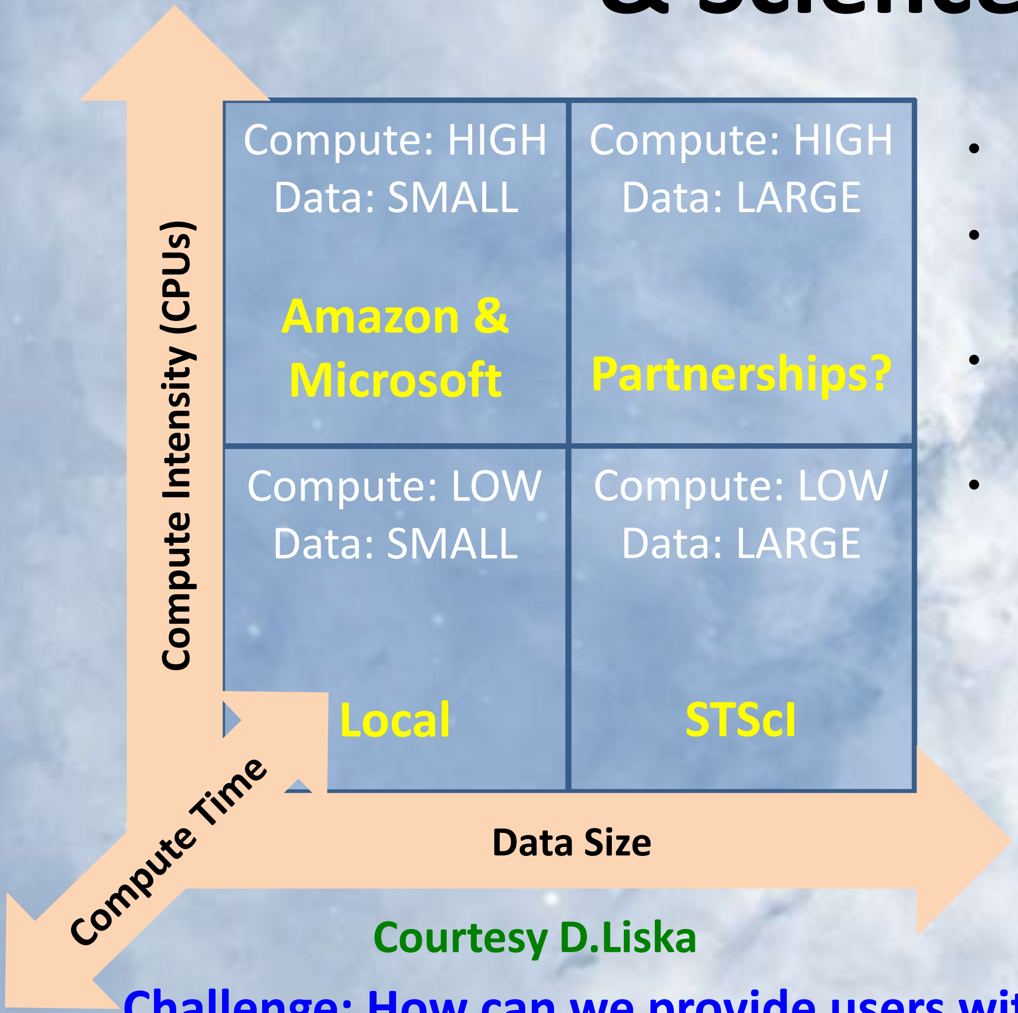
IFU Data Cubes

New Scalable Architecture for Multi-Missions Operations

- STScI recently undergone a couple of major upgrades to DMS infrastructure for modernization and added flexibility
  - Single-Sign On (SSO) upgrade
    - It allows users to use STScI SSO Portal credentials to log on to many services throughout the Institute. DMS area was the first one to move to this new STScI service.
    - Once fully implemented, SSO will allow users to only need one user name and password to access all STScI services. **Alexov poster P0002**
  - Upgrade of data processing and distribution:
    - CDBS reference file system replaced with new CRDS
    - Old OPUS pipeline infrastructure replaced with new Condor/OWL distributed workflow processing and networked storage solutions
      - New workflow allows for easily manageable HST and support/ancillary pipelines
      - Scalable architecture allows for affordability of new missions
    - OTFR replaced with online cache that is updated as needed, e.g., based on availability of new reference files and/or software
      - Data accessible through URL
    - Operational workflow now automatically includes CAOM population and preview creation for view in the MAST portal
  - New data workflow manager will be used for JWST data processing

STScI will adopt new workflow manager for creation of HLSPs !

Computing Resources & Science Cloud



Compute: HIGH Data: SMALL **Amazon & Microsoft**	Compute: HIGH Data: LARGE **Partnerships?**
Compute: LOW Data: SMALL **Local**	Compute: LOW Data: LARGE **STScI**

Courtesy D.Liska

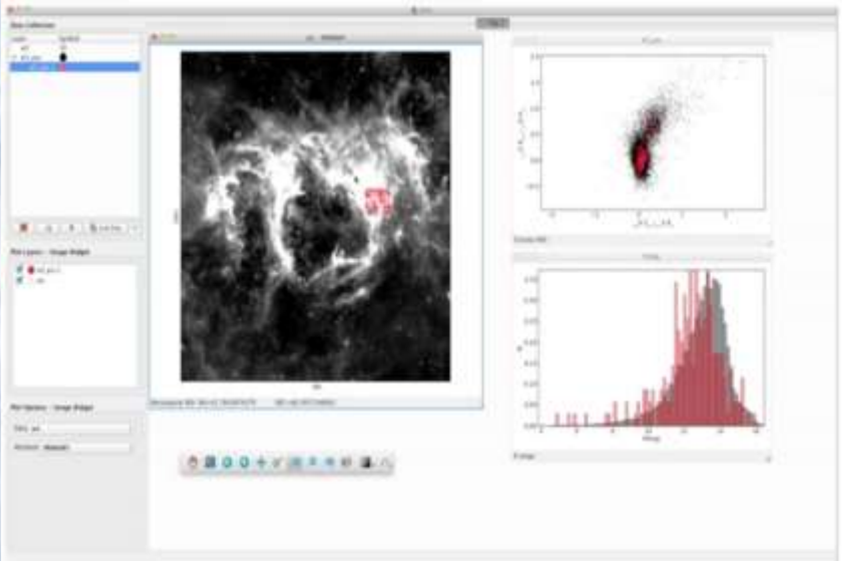
- Solution multi-faceted based on amount of data & compute needs
- Fully-virtualized compute environment called Flexible Data Center (FDC) currently under development at STScI
- FDC will need to accommodate each of the three dimensions in an integrated fashion both for internal and external (MAST) users
- Even after STScI upgrade of both Internet and Internet 2 to 10Gb over the next year or so, bottleneck will still be user bandwidth
  - Implementation of new architectures that enable Astronomer's compute environment to be "close" to the data
  - Examples of science clouds under implementation include CADC and ESAC

    O'Mullane talk O1.4
    Kinney talk O12.1
    Durand talk O12.2
    Vinsen talk O12.6

Challenge: How can we provide users with access to Science Cloud Services, including Amazon/Microsoft and Partner Institution Supercomputing Centers ?

Multi-Dimensional Data Visualization Tools

- Astronomers need to be able to more easily explore large-volume data
- Can be accomplished by building powerful, flexible, and integrated data visualization tools
- One working example is GLUE:
  - Python framework to link visualizations of multiple related datasets for easy exploration and simultaneous manipulation of data across several files.
  - Developed at Harvard through NASA funding in conjunction with the JWST project at STScI
  - Designed modularly allows astronomers to add their own custom importing, viewing, and manipulation tools
  - At present developed for JWST IFU data to only work on desktop clients and with relatively small data sets
  - Could be adapted to a web interface for large datasets

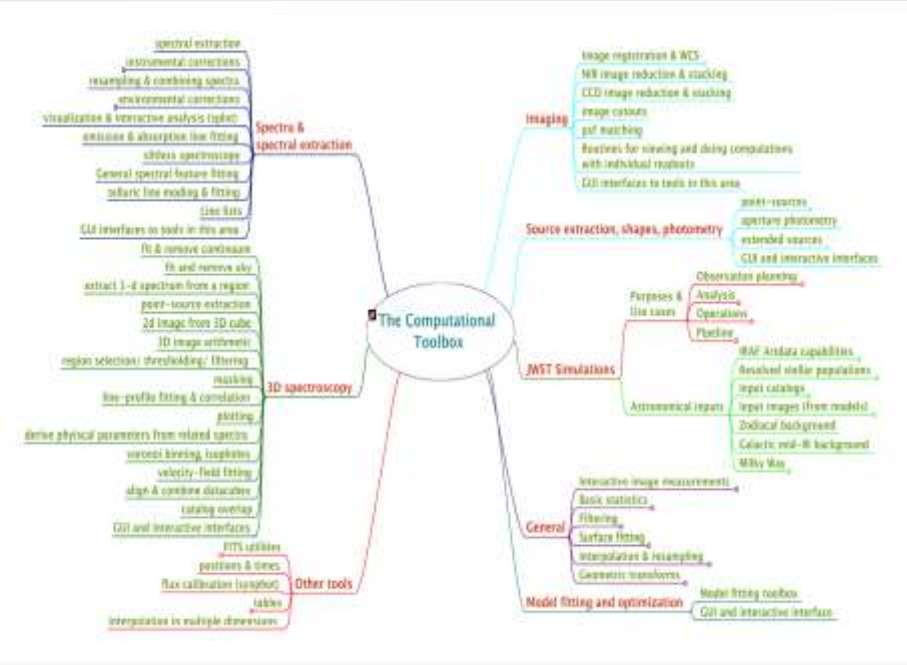


Beaumont et al. 2013

Challenge: How can we adapt a Data Visualization Tool like GLUE to work on the Web and have scalability ?

Open-Source Tools for Data Analysis

- STScI is investing in development of Data Analysis tools for JWST:
  - Open source software (Astropy)
  - Easy to install
  - Well documented
  - Easy to extend
  - Multiple interfaces (GUI, command line, & scripting)
  - Built on stable, widely adopted languages (Python and C for speed)
  - Built on stable, widely adopted code libraries
  - Leverage existing codes and algorithms.
- Concept should be extended to include repository of in-house developed and contributed on-line software accessible to archival users through science cloud


The Computational Toolbox
Spectra & spectral extraction
spectral extraction
instrumental corrections
resampling & combining spectra
environmental corrections
visualization & interactive analysis (splat)
emission & absorption line fitting
slitless spectroscopy
General spectral feature fitting
stellar line moding & fitting
Line lists
GUI interfaces to tools in this area
3D spectroscopy
fit & remove continuum
fit and remove sky
extract 1-d spectrum from a region
point-source extraction
2d image from 3D cube
3D image arithmetic
region selection/ thresholding/ filtering
masking
line-profile fitting & correlation
plotting
derive physical parameters from related spectra
voronoi binning, isophotes
velocity-field fitting
align & combine datacubes
catalog overlay
GUI and interactive interfaces
Other tools
FITS utilities
positions & times
flux calibration (synphot)
tables
interpolation in multiple dimensions
Imaging
Image registration & WCS
NIR image reduction & stacking
CCD image reduction & stacking
image cutouts
psf matching
Routines for viewing and doing computations with individual readouts
GUI interfaces to tools in this area
Source extraction, shapes, photometry
point-sources
aperture photometry
extended sources
GUI and interactive interfaces
JWST Simulations
Purposes & Use cases
Observation planning
Analysis
Operations
Pipeline
Astronomical inputs
IRAF Artdata capabilities
Resolved stellar populations
Input catalogs
Input images (from models)
Zodiacal background
Galactic mid-IR background
Milky Way
General
Interactive image measurements
Basic statistics
Filtering
Surface fitting
Interpolation & resampling
Geometric transforms
Model fitting and optimization
Model fitting toolbox
GUI and interactive interface

Ferguson et al. (2014)

Robitaille talk O8.1 on Astropy

Data Management Contacts @ STScI

Alessandra Aloisi (DMS Program Manager)
Anastasia Alexov (JWST DMS & Archive SSO)
Howard Bushouse (JWST Calibration Pipelines)
Tom Donaldson (MAST Portal & VO)
Perry Greenfield (JWST Data Analysis Tools)
Mark Kyprianou (JWST DMS)
Karen Levay (Archive Sciences Branch)
John MacKenty (Grisms WG)
Joshua Peek (MARC/DOI)
Marc Postman (Community Mission Office)
Jason Tumlinson (HST Spectroscopic Data Products WG)
Sarah Weissman (AstroTag)
Rick White (MAST)
Brad Whitmore (HSC)