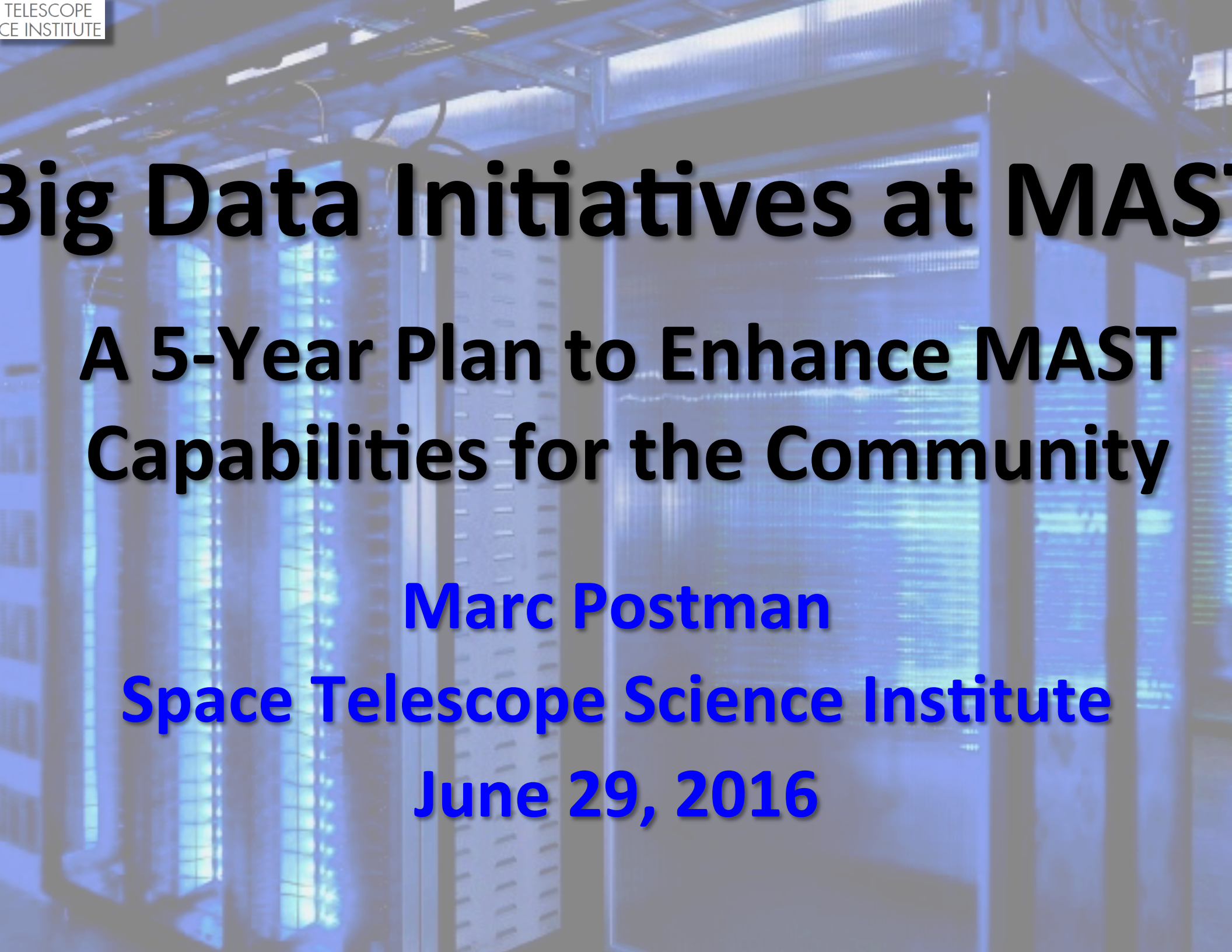


Big Data Initiatives at MAST

A 5-Year Plan to Enhance MAST Capabilities for the Community

Marc Postman

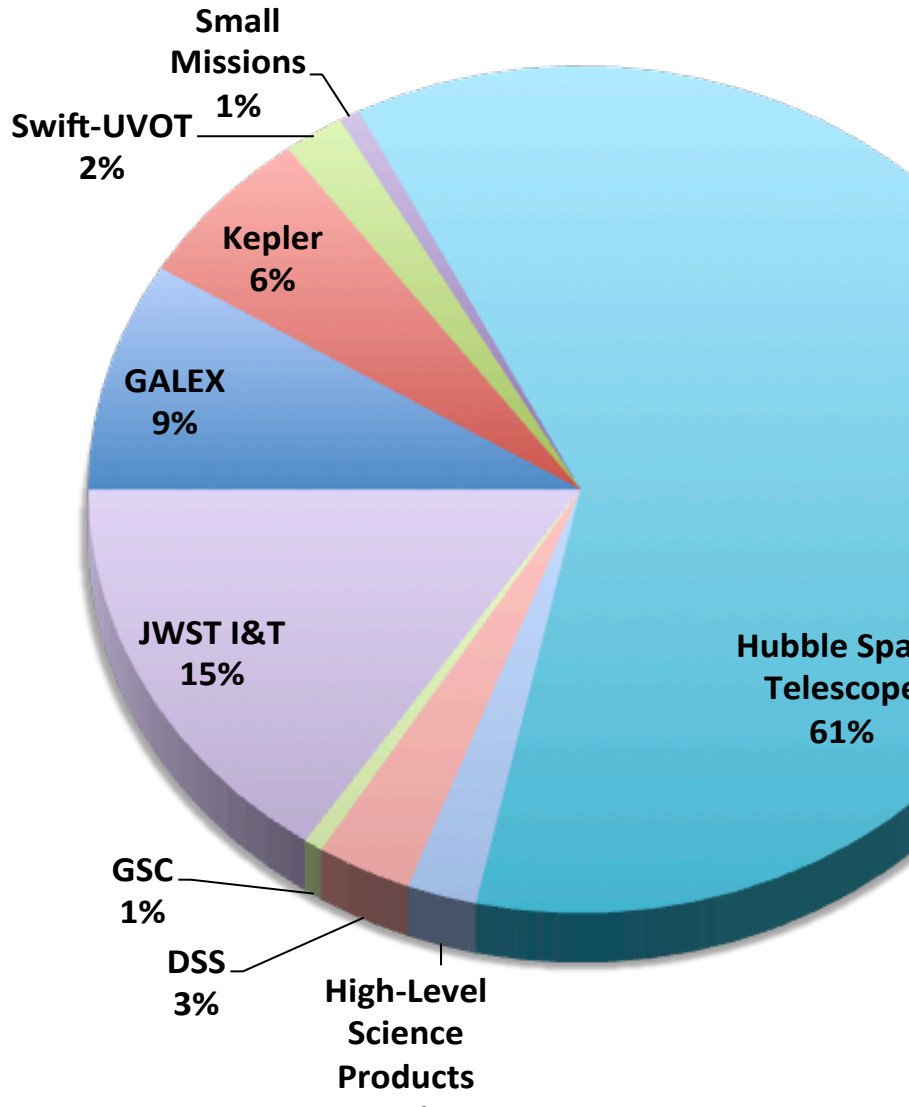
Space Telescope Science Institute

June 29, 2016

Archive for Space Telescopes

AST is a NASA astrophysics data archive cer

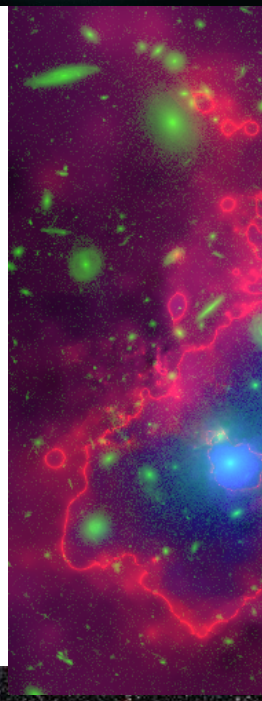
- Archive established with HST launch in 1990
- Multi-mission since addition of IUE in 1998
- 4 active missions including Hubble, Kepler
- Many legacy missions: GALEX, IUE, FUSE, …
- Future: TESS, JWST, WFIRST
- Includes some relevant ground-based archives: GSC, PanSTARRS*



MAST data are diverse

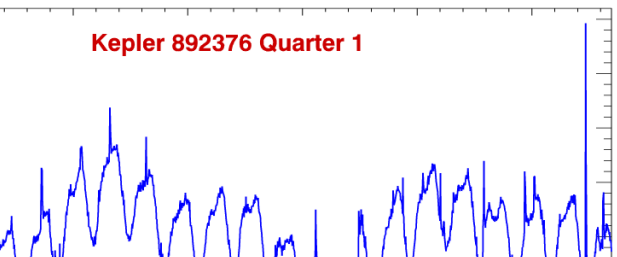
ST data are diverse:

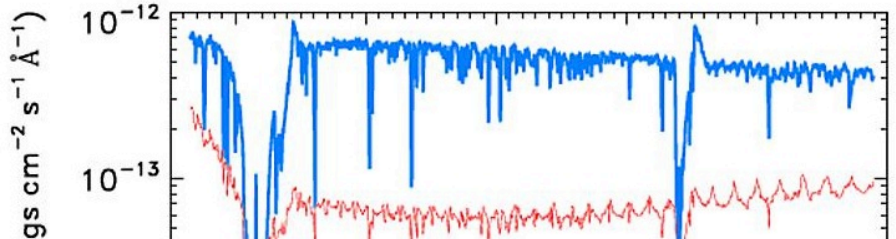
- data types: images, spectra, light curves, catalogs, odels
- scale: from Pan-STARRS (2 PB images + 100 TB atabase) to small shuttle-based missions o community-contributed projects with just a few files
- processing level: from raw data packets delivered rectly from spacecraft to science-ready, high-level cience products

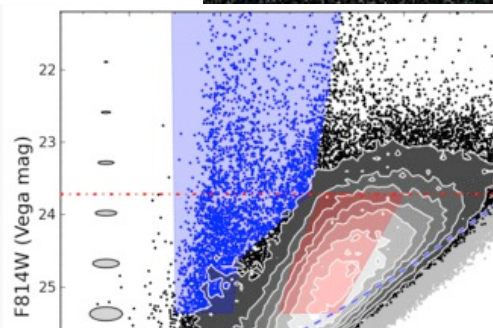


ny different missions and instruments

ubble alone: 12 different instruments, 17 varieties of etector, hundreds of instrument modes/filters/etc.

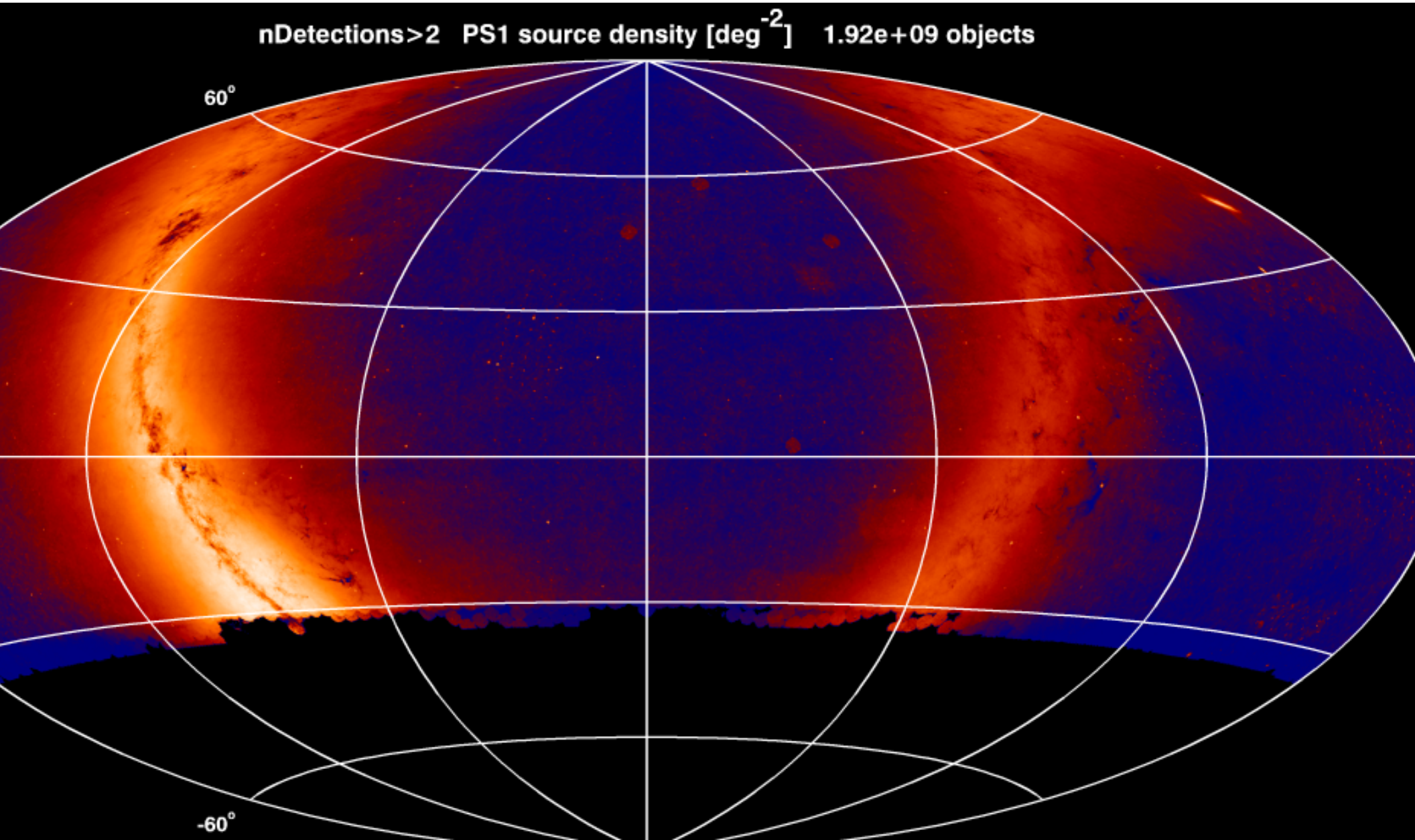






The End of Surveys

hive users are getting more sophisticated every year, and their queries (and analy
easingly crossing over archive and wavelength boundaries. By the time WF
ched this will be even more so, thus **the MAST archive must evolve wi
siderations and capabilities in mind.**



What we mean by "Big Data"

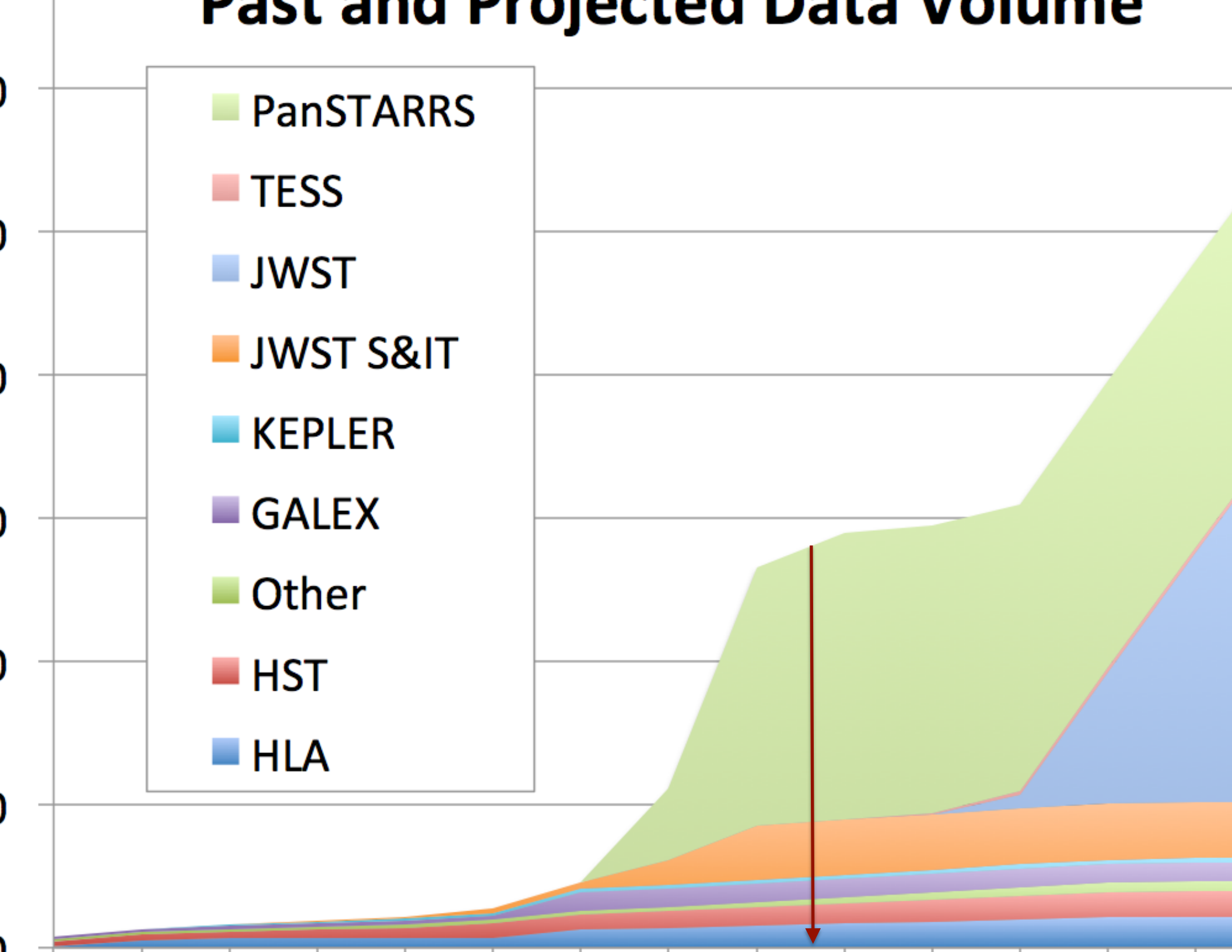
Data" is a term borrowed from industry, which collects data
e scale from users and sensors, and is typically used to serve
ertisements with a higher return on investment than would
erwise be possible.

MAST, we redefine it for astrophysics applications as data that
ets one or more of the following criteria:

Data whose raw form is so large that we must qualitatively cha
he way in which we reduce, store, and access it.

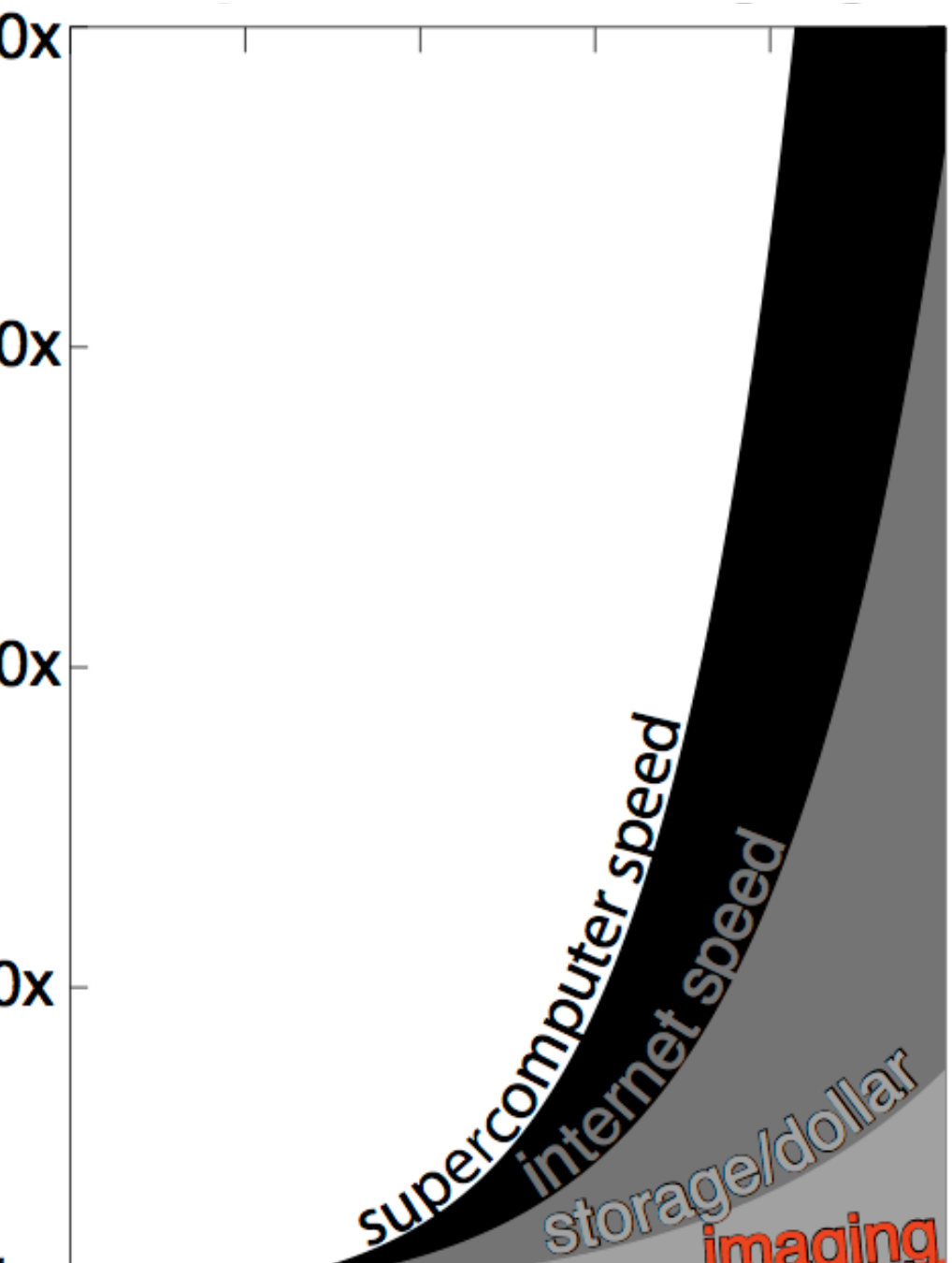
Data whose reduced form is so large that we must qualitativel
hange the way in which we interact with and explore it.

Data whose structure is so complex that our current tools cann
fficiently extract the scientific information we seek.


Past and Projected Data Volume
PanSTARRS
TESS
JWST
JWST S&IT
KEPLER
GALEX
Other
HST
HLA

er 10 years, the rate of UVOIR imaging typically grows by a factor of '

In contrast, the amount of storage one can buy, internet speed, and supercomputer speed grow by factors of dozens to hundreds.

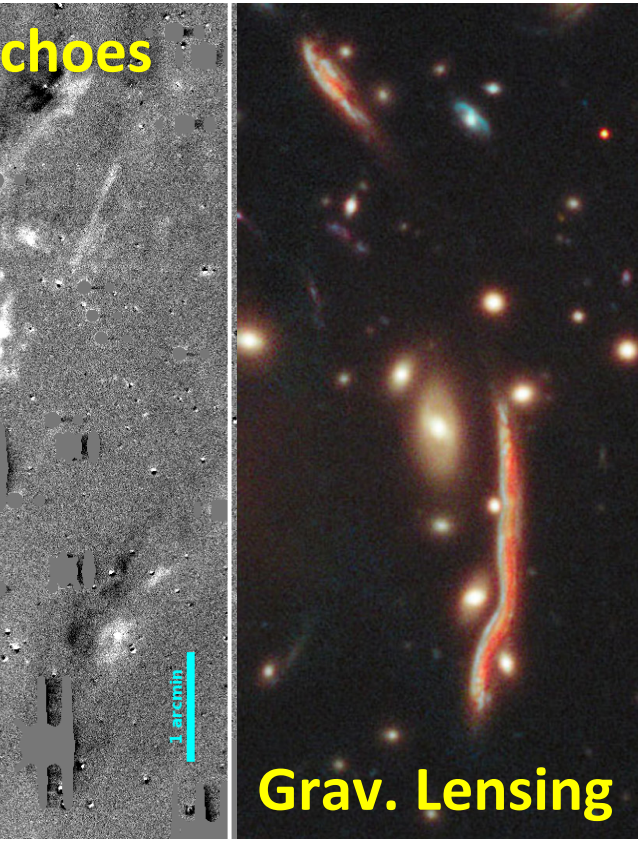


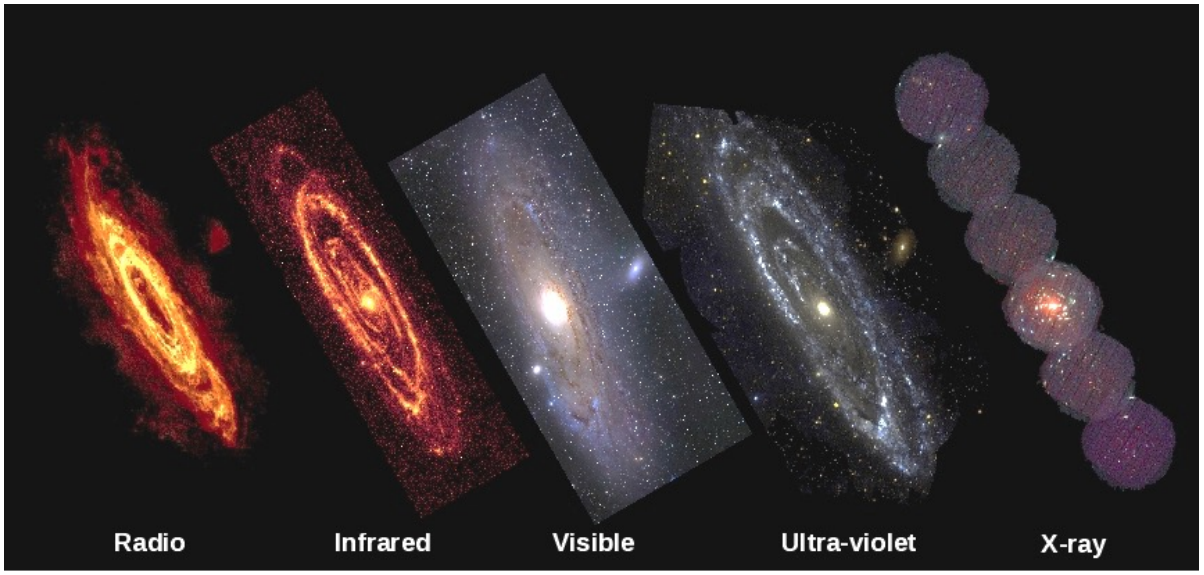
Data source: Barentsen & Peek,
https://github.com/barentsen/tech-progress-data

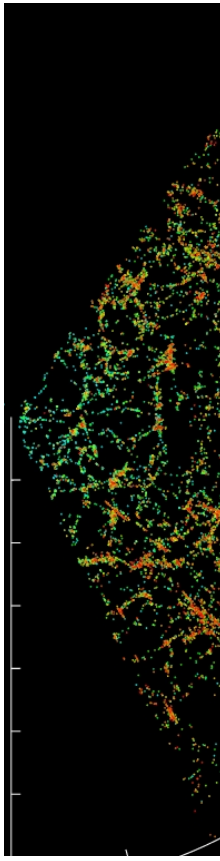
It is not our computers that
keep up, but our brains: our l
ability to explore, compre
assimilate, and infer won't ke
with the flow of data unle
build sophisticated tool
interacting with and commun
the ideas hidden in these eno
datasets.

ated source classification

Multi-wavelength data collection and cross-corre





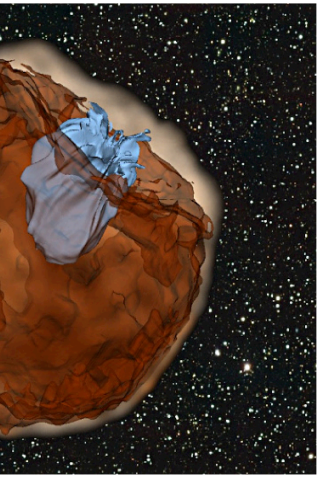


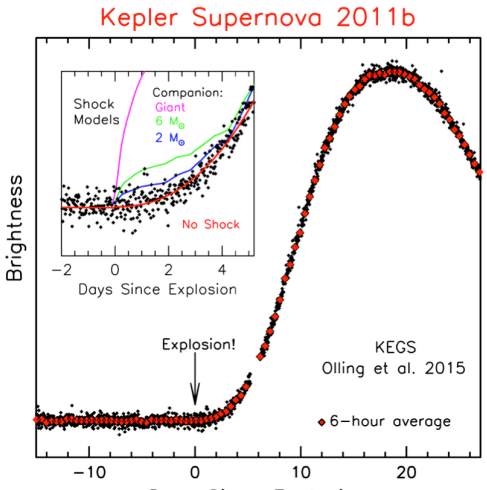
Stellar Populations, Galaxies, Active Galactic Nuclei, Galaxy Photometric Redshifts

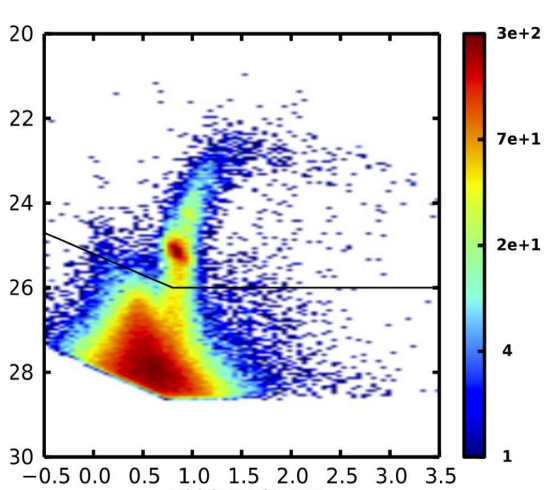
omain analyses

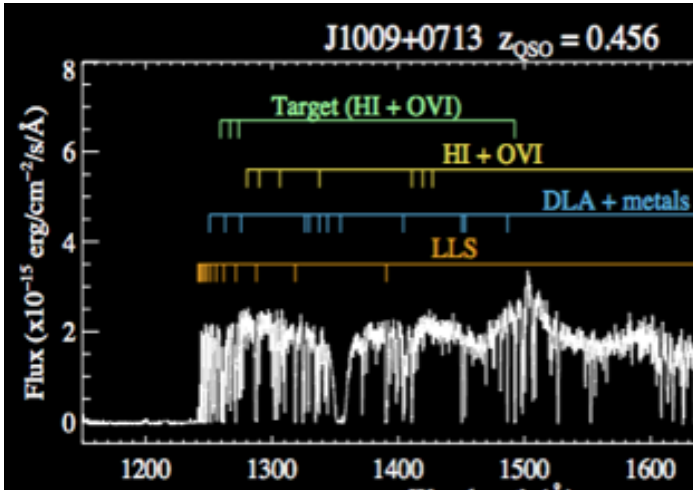
Model fitting to large or complex datasets

Disentangling datasets constituent systems









Advances

Derive Star Formation Histories for 100 Nearest Galaxies (photometr
equired for 10^9 – 10^{10} stars).

Galaxy – Black Hole Co-Evolution Study in Population of 10^7 Galaxies

Automated detection and classification of light echoes (requires time
domain data). Light echo structure is complex and sources are rare.

Automated detection and classification of amorphous astronomical
structures (lensed galaxies, gaseous filaments, etc.). Source structure
complex.

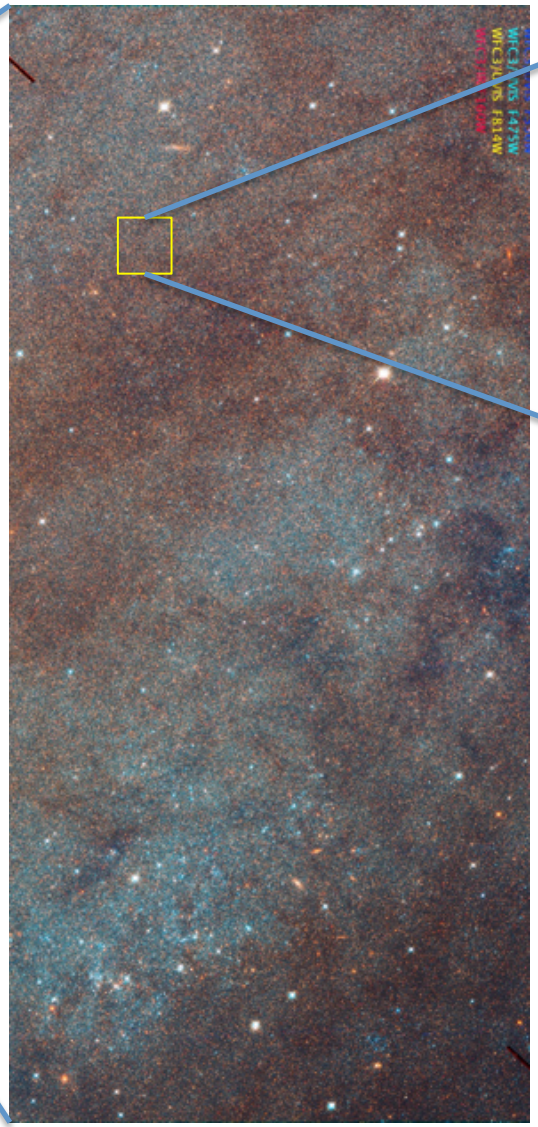
Determine 3D distribution of 10^9 galaxies based on photometric reds
derived from multi-wavelength data.

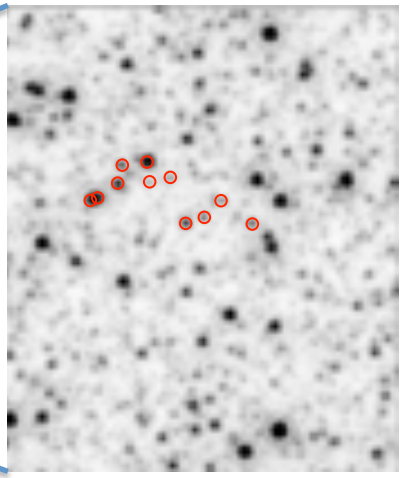
Automated Classification of light-curve data from Kepler/K2/TESS
(hundreds of data points for 10^5 stars).

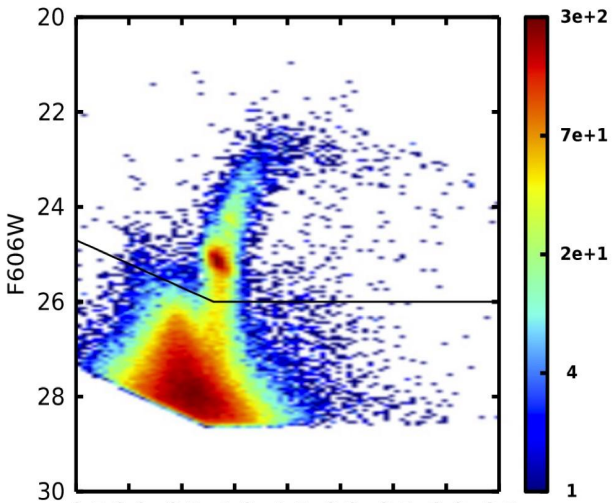
Probabilistic Classification of Transient sources via cross-matching w
arge multi-wavelength catalogs.

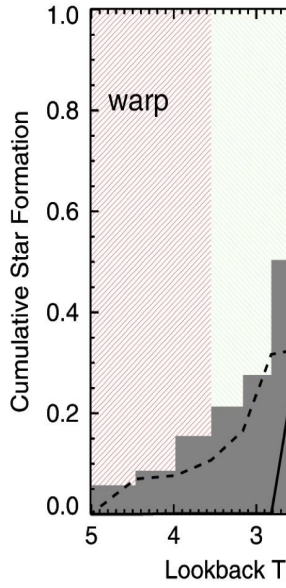
dentification of spectroscopically rare sources via PCA template

racking Star Formation Historie









uality becomes better, larger fraction of users can use automated path for their science.

data

ive – processed

pipeline or HLA)

ive – raw data

3. Do photom (& completeness)

- Use HLA/HSC (DAOphot, SExtractor)
- Use other existing catalog (e.g., HLSP)
- Use SExtractor yourself
- Use DAOphot yourself
- Use other photometry package yourself
- Completeness - not included in HLA/HSC
- Completeness - limited public tools exist

5. SED fitting of single objects

- Limited public tools exist
- TADA
- CHORIZOS (limited support)
- BEAST (not public)
- Use own software

7. Analysis/V

- Use Discovery
  CasJobs - limi
- IDL
- Pyraf (IRAF)
- DS9
- Glue (3D visu
- Use own softv

Typical project timeline

e images

strodrizzle)

rizzle yourself

oftware

4. Crossmatching data (if needed)

- Use HSC
- Use Discovery Portal to do crossmatch
- Use other tools (e.g., IRAF/TMATCH, DAOPHOT/match, …)
- Use own software

6. SFH fitting of populations

- Limited public tools exist
- StarFISH (public – outdated)
- MATCH (not public)
- Use own software

8. Simulati

- Limited pub
- DAOphot/ac
- TFIT - Match
  resolution
- Sunrise (n-b
- Use own soft

STScI's Current Data Science Infrastructure

mpute Servers for Science:

- 6 science servers, ranging from 8 to 32 cores per server. RAM: 32 to 400 GB.
- Utilization of science servers is managed via inform communications. No checking for resource conflic (e.g., running cpu-intensive jobs simultaneously).
- Servers cannot be easily reconfigured to match needs of different kinds of applications.
- STScI has aggregate of ~2,000 cores but they cann be utilized for large-scale computing due to lack o virtualized environment.

rSci's Current Data Science Infrastructu

rrent Large Mission systems:

HST mission systems highly physical (non-virtualiz
nd fixed capacity though use of High Throughput
Computing (HTC) via HTCondor does provide job
cheduling and resource allocation for the Hubble
Data Management System.

WST mission systems are being implemented as
irtualized from ground up. Some in-house
pplications (like DMS) that are primarily CPU-bou
nay stay non-virtualized.

NSCI's Current Data Science Infrastructure

ndwidth:

Mix of 1 Gb/s and 10 Gb/s internal networks.

Upgrades to 10 Gb/s are made during normal refr
ycles or earlier, if needed for specific engineering
equirements (e.g. Connection to storage systems

Connection to internet: 500 Mb/s

Connection to internet2: 1 Gb/s (as of Dec. 2015)

Plans being made to upgrade standard internet
onnection to 1 Gb/s in short term (<12 months?)
nd 10 Gb/s in longer term.

rSci's Current Data Science Infrastructu

rage:

~8 PB of storage using mix of NAS and SAN system

Largest single dedicated system is for PanSTARRS
PB) followed by MAST (1.2 PB).

dvanced Software and Analysis Too

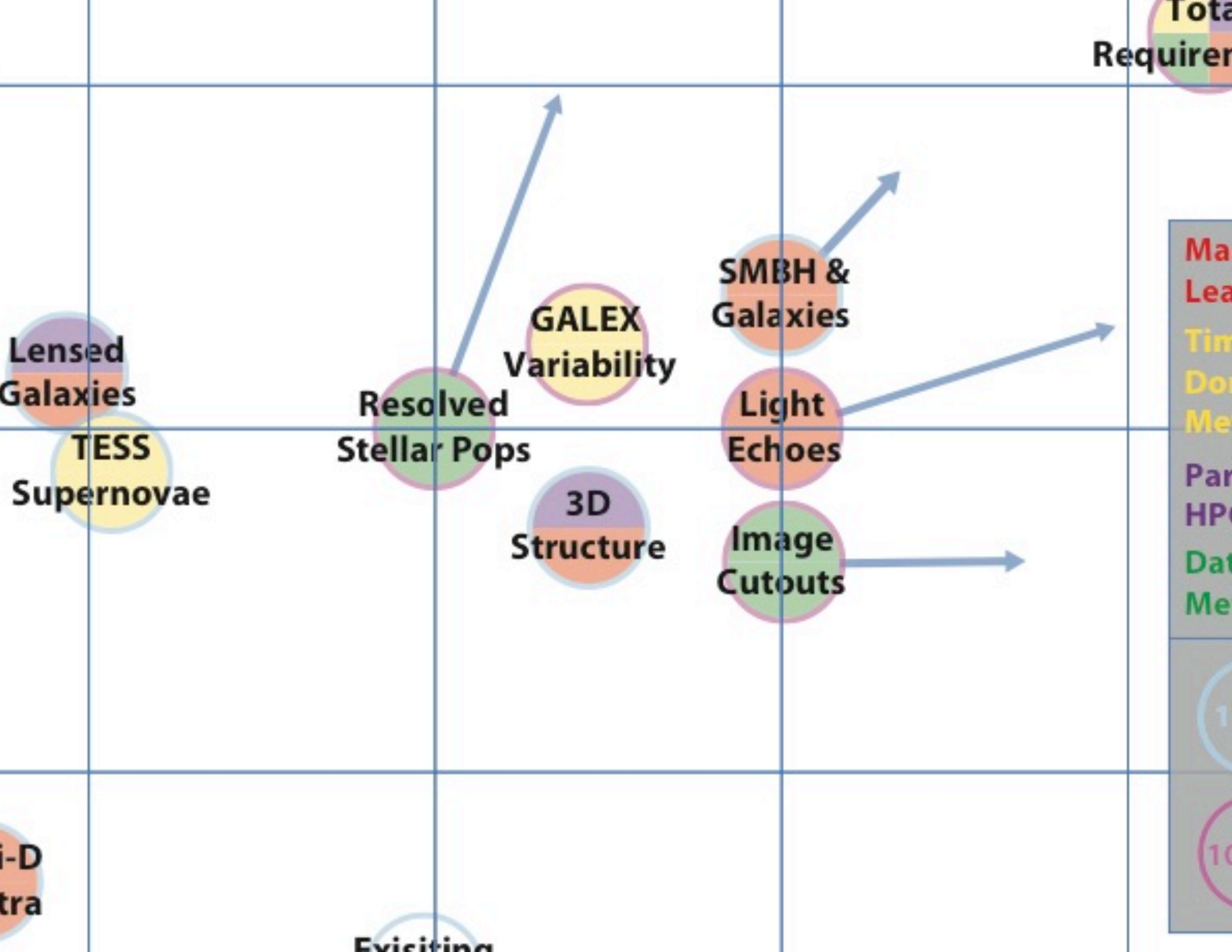
Archive users are getting more sophisticated every ear, and their queries (and analyses) are increasing rossing over archive and wavelength boundaries.

Archives of the near future will be more than a sim passive file or database servers. Users will rather nteract with them algorithmically though well- lefined and well-designed Applications Programmi nterfaces (APIs).

Cloud computing (as well as dedicated local clouds) hould be ubiquitous in the next 5 to 10 years.

Given the amount of astronomical data in MAST in ear future, we expect that server-side analyses wi e commonplace for the users, thus an advanced cripting capability must be supported.

ence Case	Computing	Storage	Bandwidth	Software
ho detection ssification (PS1,)	>1000 cpu core	~1 to 10 PB	~10 Gb/s needed for citizen science/ data transfer.	Machine Learning (ML) classification, feature v (FV) mapping.
galaxy on (HST, JWST,)	~1000 cpu core	~10 TB (if image cutout service avail.)	~1 Gb/s	Parallelizable code management, ML code
d stellar tion studies VST, WFIRST)	>1000 cpu core (10K core better)	Few hundred TB	~10 Gb/s (uses non-local data)	Automated pipeline, ef database query tools
AGN / Galaxy ution studies	>2000 cpu core	>1 PB	>1 Gb/s	ML classification / FV mapping / efficient x-co
cture in	~500 cpu core	>200 TB	>1 Gb/s	Parallelizable code; hig connectivity to other ar
pernovae	~1000 cpu core	~10 TB for raw data	~1 Gb/s (if data local)	Transient detection cod highly undersampled da
nter-visit ity study	Several hundred cpu core	Few hundred TB	High bw to local db (>10 Gb/s)	Transient detection cod capable of running on 1 curves.
imensional scopic datasets	Not a driver	~100 GB	Current OK	FV mapping, ML codes, Efficient x-corr, PCA ana


Lensed Galaxies
TESS Supernovae
Resolved Stellar Pops
GALEX Variability
3D Structure
SMBH & Galaxies
Light Echoes
Image Cutouts

ecommendations: Infrastructur

Significant improvements in internal and exter
network bandwidths are needed. External bw:
Gb/s asap; 10 Gb/s by 2018-2019. Internal bw:
10 Gb/s by 2018, 100 Gbps by 2021.

Multiple virtualized, dynamically configurable
nodes with upwards of 1000 cores each are
needed in next 2 years. Multiple 10K core
systems will likely be needed in 5 years.

Storage: Increase storage capacity by 6 PB by
end of 2018; increase to 30 – 50 PB by 2021.

ecommendations: System Desig

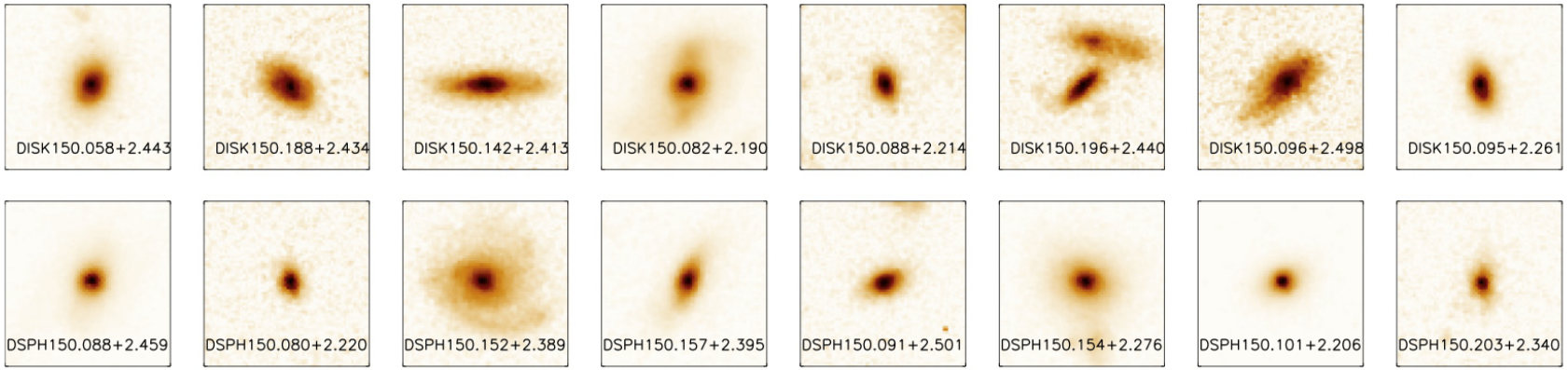
- Data workspaces needed to bring "scientists to the data."
- Workspace must support a variety of tools, scriptin languages and software.
- Must provide adequate computing resources (cpu, memory, storage)
- Must have good internal bandwidth between workspace, data sources, and cpu cores.
- n 5 years timeframe, STScI will need to invest in ybrid cloud (Flexible datacenter/private + public loud/partner institution [e.g., JHU/MARCC = Mid-Atlantic Regional Computing Center])

ecommendations: Science Softwar

Develop integrated data visualization tools tha
it server-side to enable researchers to efficien
explore all our data holdings (e.g., *GLUE*).

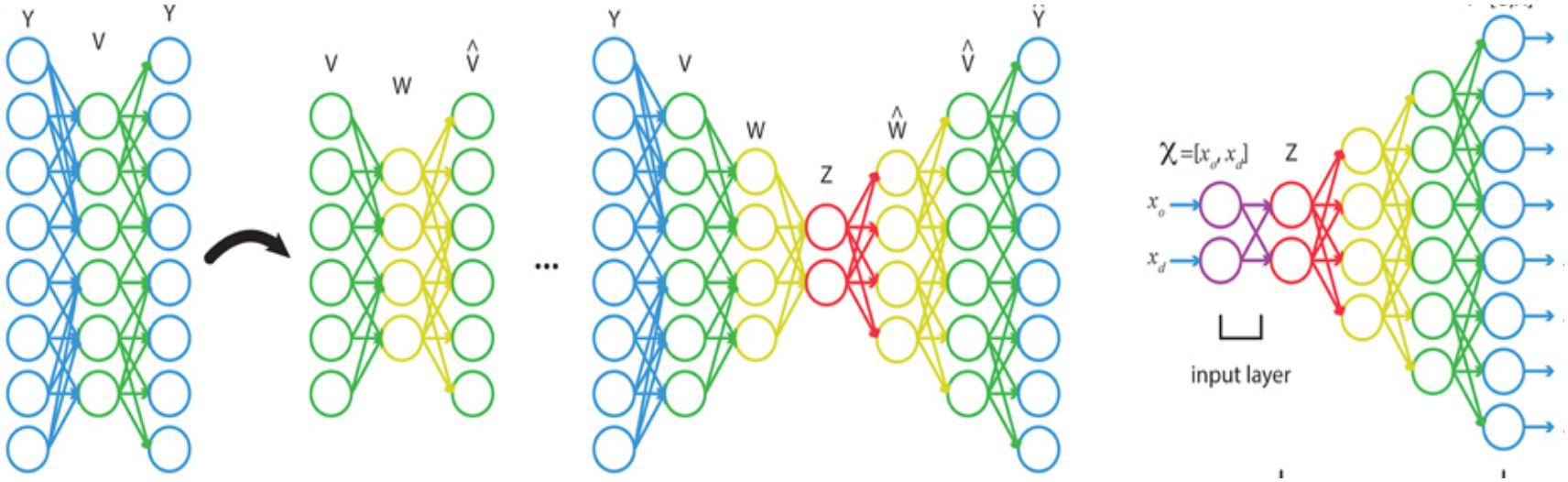
Explore, select and deploy machine-learning
architectures to support archive researchers w
will, more frequently, require advanced
classification and regression analyses.

nitiate development of automated spectral
eature classification tools. The software shoul
be adaptable to data from all current and futu
pectroscopic data in MAST.



Server–Side Scripting Machine Learning (ML) & visualization

- Allows users to refine (in real-time) their queries and generate new hypotheses without having to download, query, investigate, and repeat.
- On-the-fly ML clustering and dimensionality reduction of search results ca generate easy to investigate data visualizations

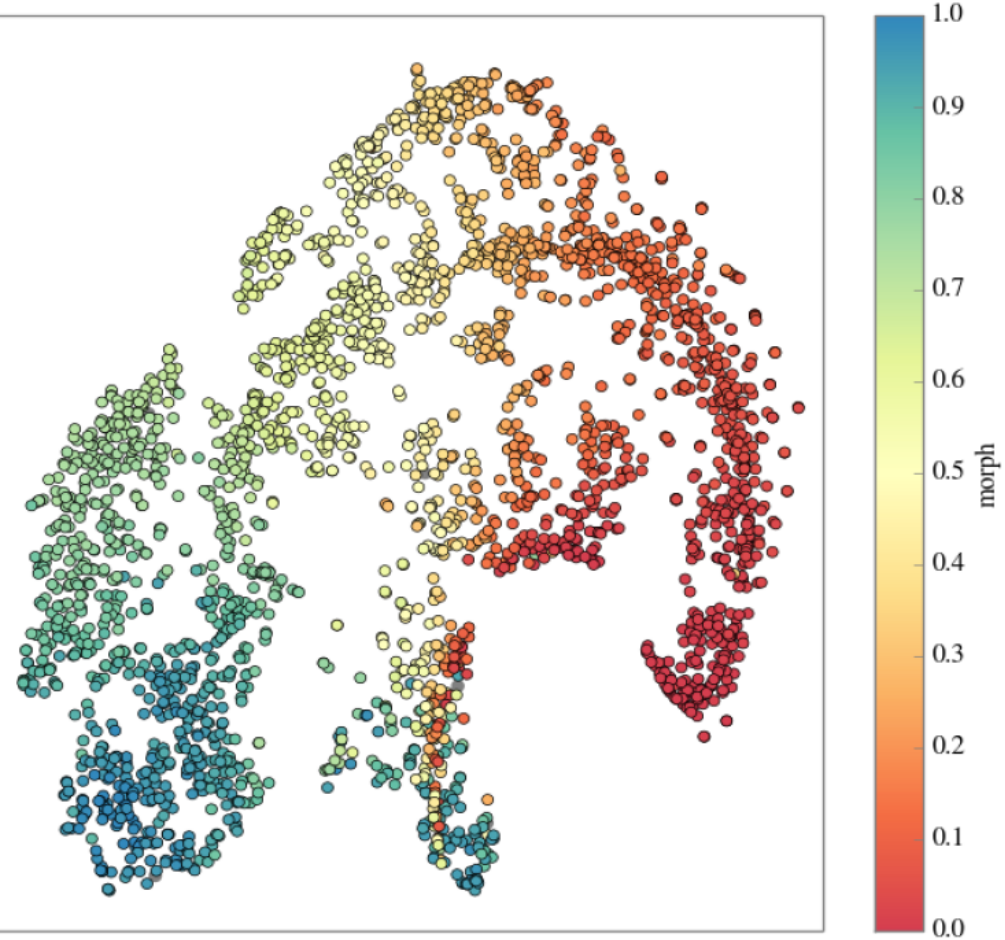


Deep auto-encoder dimensionality reduction: *unsupervised learning methods take very high dimensional data and reduce them to 2 or 3 key dimensions h*

erver–Side Scripting Machine Learning and visualization

- t-SNE attempts to cluster objects into like groups. If we can enable users to perform this clustering and dimensionality reduction "on the fly" on search results, it will allow users to further refine their searches and do more advanced hypothesis generation before downloading any large data sets.

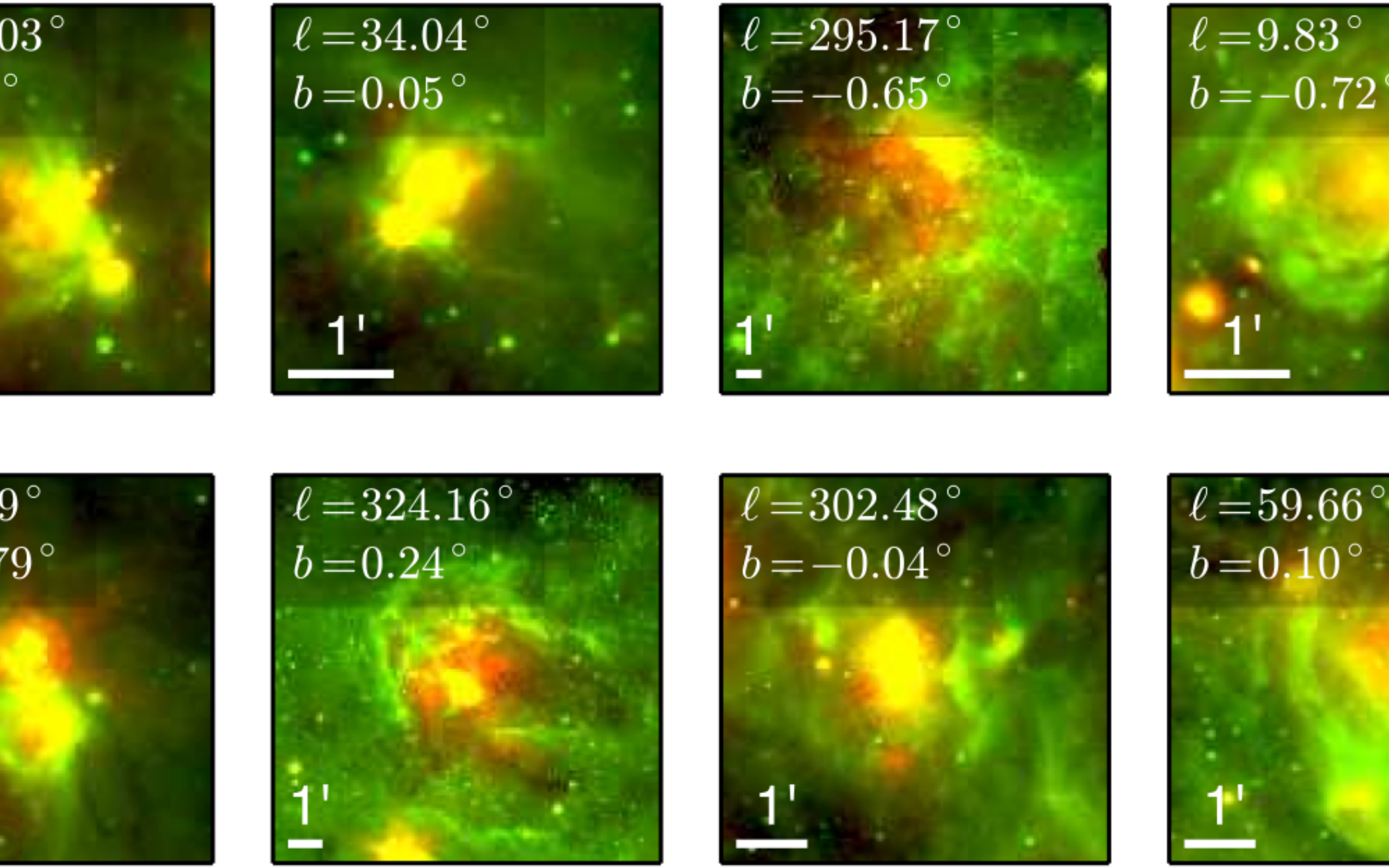
http://keplerebs.villanova.edu/tsne



t-SNE* clustering *applied to Kepler eclipsing binary star data.*
(Matijevic et al. 2012)

**t-Distributed Stochastic Neighbor Embedding*

fying bubbles in interstellar gas (due to star forming re



The ML algorithm "Brut" was trained on citizen science classifications

Recommendations: Skill sets

nitiate the ***Barry M. Lasker Data Science Postdoctoral Fellowship***: focused specifical on early-career researcher whose interests nd expertise are centered on big data cience.

– Program has been initiated. First fellow selecte (41 applied).

Recommendations: Skill sets

Expand staff expertise with server-side API' and scripting environments.

Expand staff expertise in the areas of machine-learning and automated classification techniques.

Expand staff expertise in the development and application of code to reduce the dimensionality of highly complex datasets.

ecommendations: Organizational Structu

onsolidate leadership and oversight of our current and upcoming scie
rchive initiatives (HST, Kepler, TESS, JWST, MAST, PanSTARRS, WFIRST
ith goal of maximizing science potential, supported with state-of-the-
ols and interfaces.

stablish a Data Science Mission Office (DSMO) at STScI.

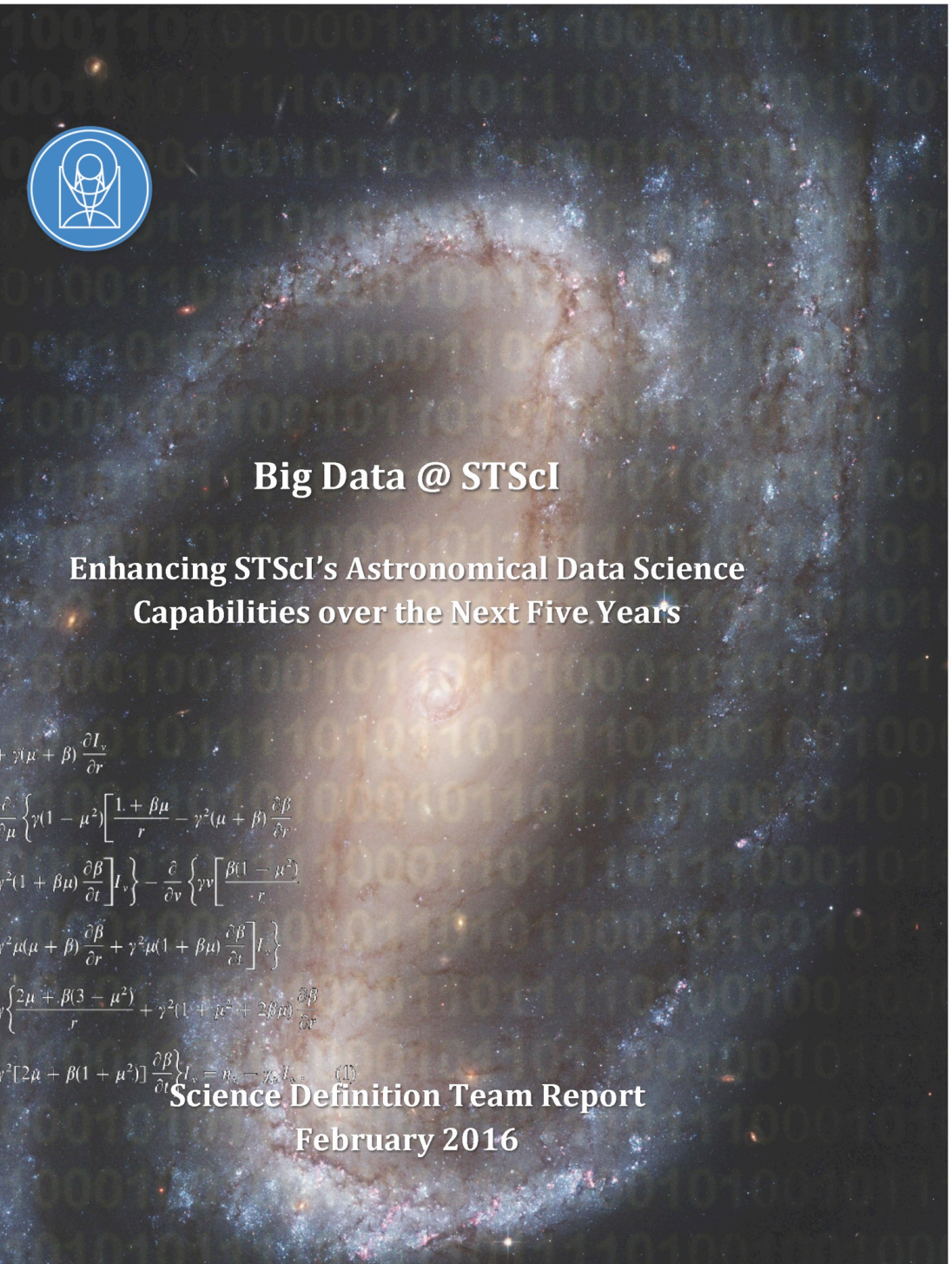
rchive project scientist and Archive project technologist will be reside
SMO.

- Project scientist responsible for developing and maximizing scientific impact of STScI archives.
- Project technologist responsible for developing and prioritizing utilization of appropriate technologies to support archival research and big data science.

taff members from other missions and divisions will be matrixed with
SMO to support its activities.

Big Data Report

http://archive.stsci.edu/reports/BigDataReport_Final.pdf



Discovery in Astronomically Big Data

An AURA (STScI – NOAO – LSST) workshop and hands-on explorati

February 27 - March 2, 2017 at STScI

Astro Domains		**Data Methods**	
xy Beyond 3D	Big Spectroscopy	Citizen Science	Data In
ts & the Time Domain	Morphology	Machine Learning	Visu

Goals

n about new scientific discoveries in each astronomy domain

ange skills for discovery methods in current datasets

ss and reduce cultural barriers to work in the discovery space

elop partnerships for common visualization and exploration to