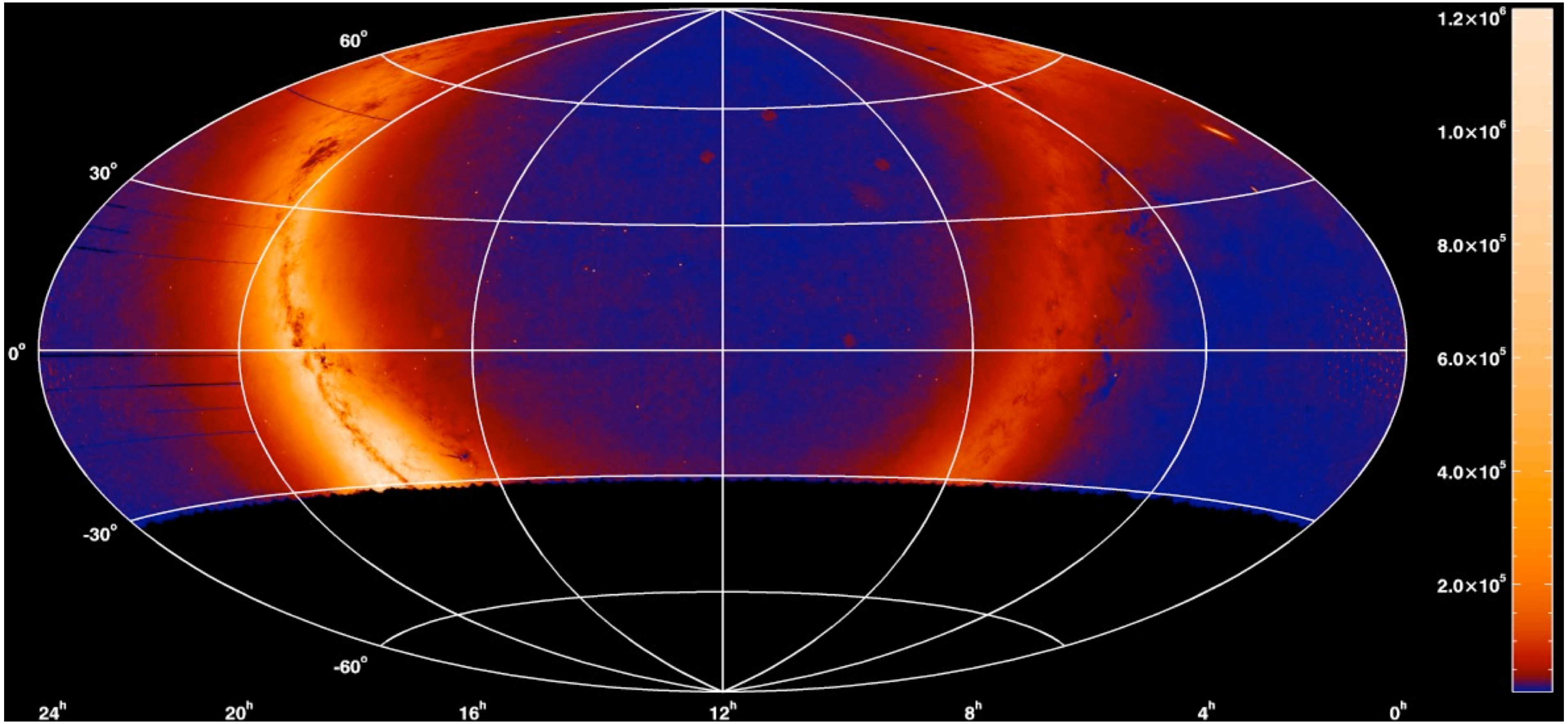
The WFIRST Science Archive: An Astrophysics Discovery Machine

Marc Postman (STScI), Alex Szalay (JHU)
and the SIT-F Team

Wide-Field Infrared Survey Telescope
WFIRST
Astrophysics • Dark Energy • Exoplanets

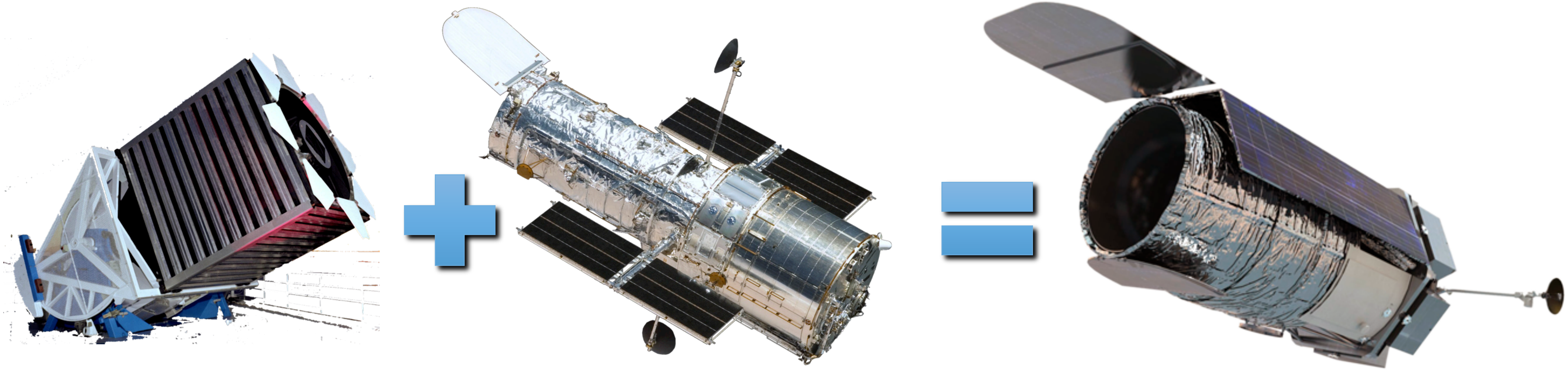
WFIRST and The Era of Surveys

Archive users are getting more sophisticated every year, and their queries (and analyses) are increasingly crossing over archive and wavelength boundaries. By the time WFIRST is launched this will be even more so, thus **the WFIRST archive must be built with such considerations and capabilities in mind.**

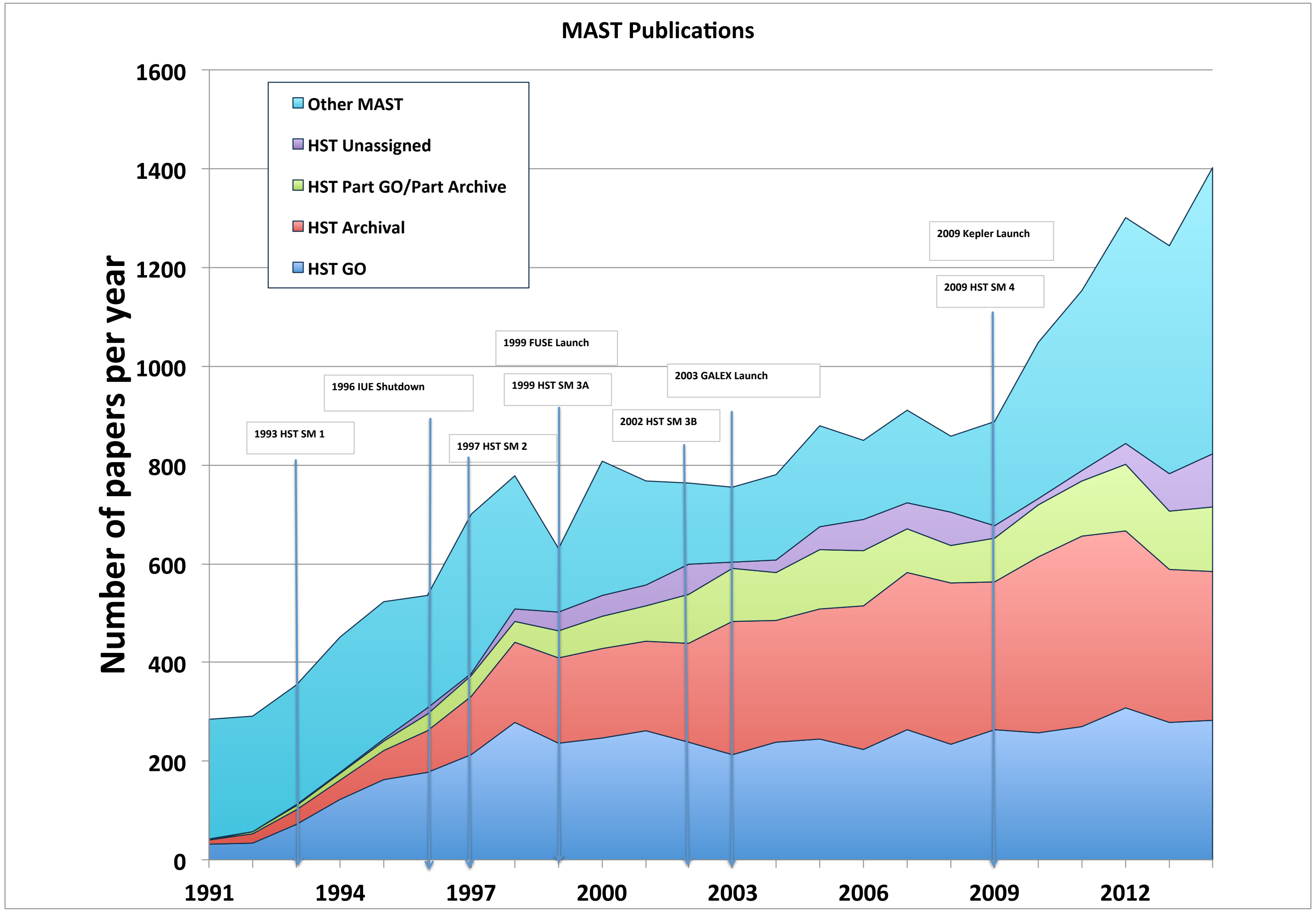


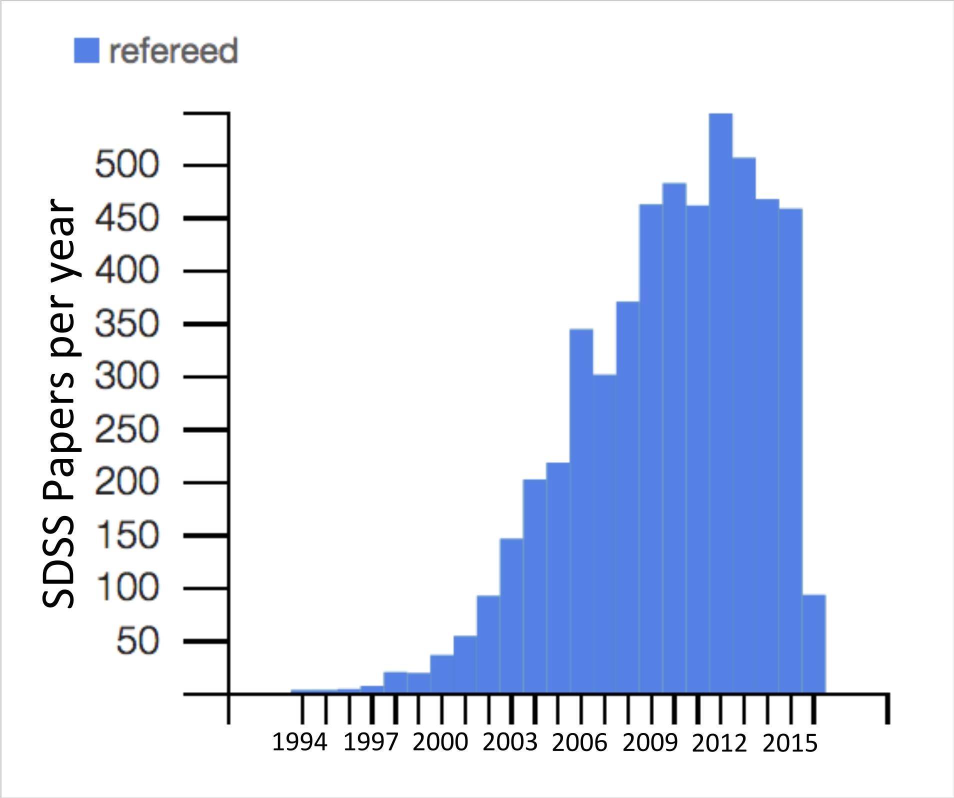
Map of 1.8 billion objects in PanSTARRS 3PI survey (R. White, 2016)

Our view of WFIRST:



- A combination of a large survey telescope with a Great Observatory.
- Key Challenge for our SIT: how do you design the WFIRST database, advanced machine learning algorithms, and the science archive interfaces to enable a broad community to extract a wide array of science from the WFIRST datasets?

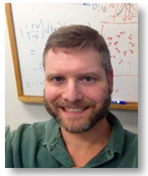




The WFIRST Science Archive will play a key role in maximizing scientific discovery in the 2020s.

The WSA will serve a broad community: Survey Science Teams, Guest Observers, and Guest Investigators (*a.k.a.,* Archival Researchers)

Our Team



- Jay Anderson (STScI)

- Tamas Budavari (JHU)

- Andrew Connolly (UW)

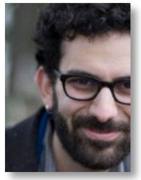
- Mike Fall (STScI)

- Sarah Heap (GSFC)

- Gerard Lemson (JHU)

- Tom McGlynn (GSFC)

- Brice Menard (JHU)



- Joshua Peek (STScI)

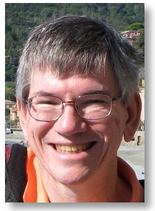
- Marc Postman (STScI)

- Swara Ravindranath (STScI)

- Greg Snyder (STScI)

- Alexander Szalay (JHU), P.I.

- Ani Thakar (JHU)



- Rick White (STScI)

Goals for Designing WFIRST Archive Science Tools

- Study and evaluate the best archival practices from across all of astronomy.
- Design and build simple end-to-end simulations. Incorporate these simulations into a scalable, queryable database.
- Establish a common platform for data query interfaces and data exploration.
- Design and prototype a unified object catalog using a simulated catalog, with cross-matches to external surveys.
- Investigate and develop capabilities needed for a GO+GI hybrid archive.
- Identify forward-looking technologies not in production today anywhere (e.g., scripting, fast parallel analysis tools).
- Develop and build prototypes of highly scalable parallel tools (e.g. cross-correlations inside the archive database).
- Implement novel object classification codes based on machine learning techniques.

Scientific Requirements

Derived from specific science use cases:

- Identify scaling relations between galaxy properties detectable when cross-correlating information in multiple databases of millions of galaxies with multiband photometry and grism spectroscopy.
- Establish constraints on galaxy evolution over the range 0.5 < z < 7 using a combination of simulated WFIRST morphological, photometric and spectroscopic data with a novel approach to modeling stellar assembly in galaxies using realistic imaging and grism simulations.
- Determine the early (z > 1) evolution of galaxy morphology using existing cosmologically simulated galaxy images.
- Determine how well WFIRST data constrains the accretion histories of SMBHs and their influence on galaxy growth using population models and mock grism data to test various scenarios.
- Identify significant stellar structures in the Galactic halo via simultaneous archive queries of all WFIRST datasets (HLS + GO data).

Progenitors of the Milky Way galaxy from z=3 to z=0.5

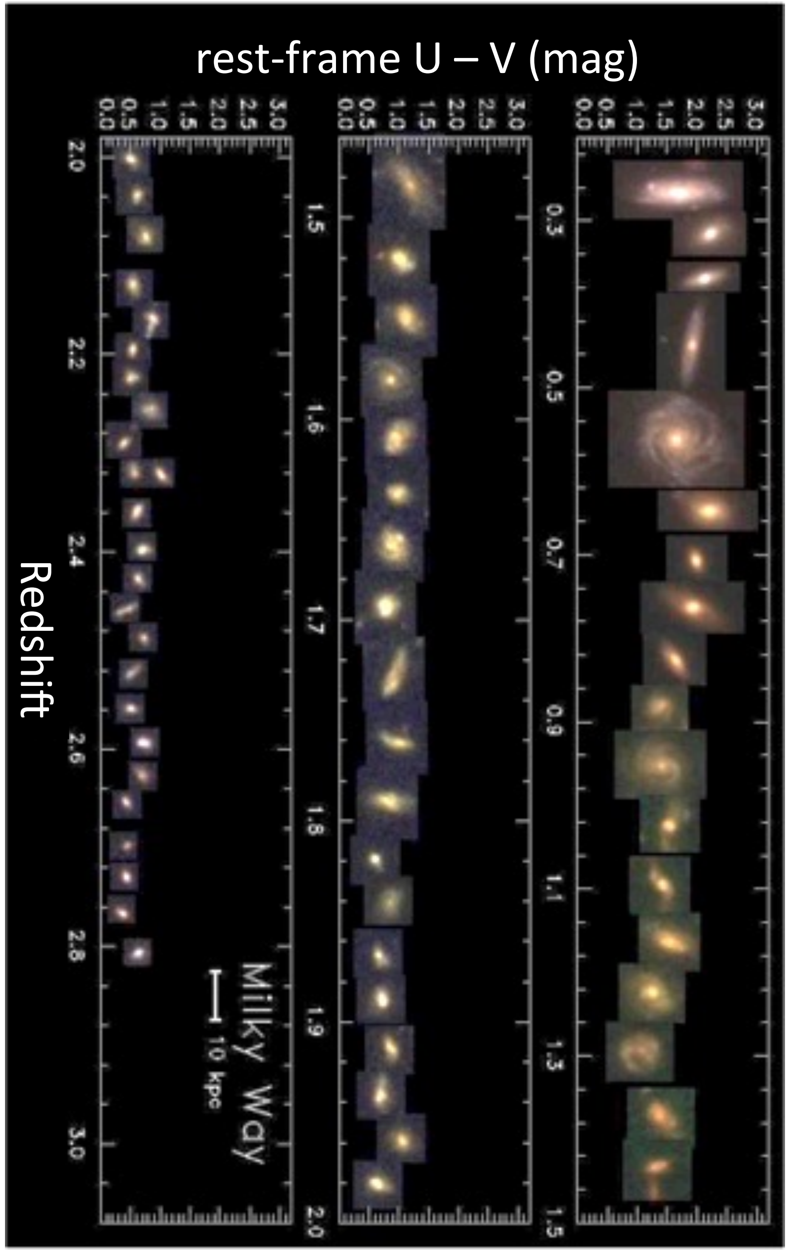


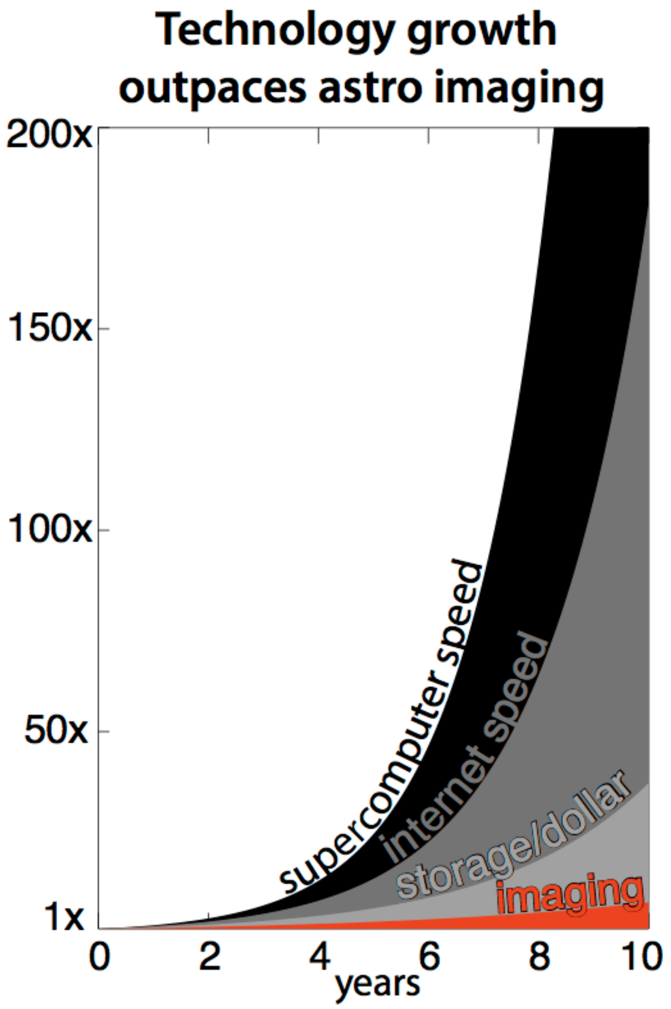
Image credit: Papovich et al. (2015)

Design Philosophy

- WSA must Integrate data from a wide variety of sources, as it is likely that user queries will span not only the WFIRST database but other databases, particularly LSST (but also PanSTARRS, Gaia, WISE, VISTA, DES, HSC, Euclid, eRosita, etc.).
- Apply the lessons learned from archives across astronomy, from GO-based (like MAST) to GI-based (like SDSS), to find best practices and identify potential solutions to the particular issues relevant to WFIRST.
- Make the archive tools and data easily accessible to non-specialist astronomers.

Our Approach

- How to best move the analysis to the data?
- Use “20 queries” methodology (Gray et al 2002).
- Identify and adapt advanced technologies for use with the WFIRST science archive.
- Supporting a heterogeneous user base:
    - Experts vs. General users
    - GI’s vs. GO’s
    - NIR vs. other wavelengths
    - Power users vs. quick queries
- Work with the community and the WFIRST SWG to develop archive tools to aid GO+GI investigations.



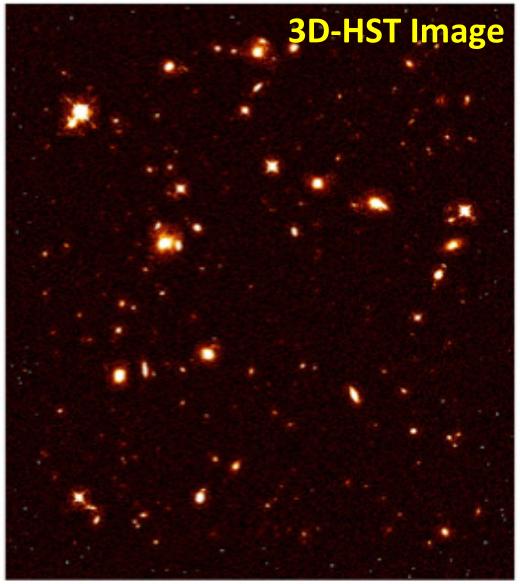
Data source: Barentsen & Peek, https://github.com/barentsen/tech-progress-data

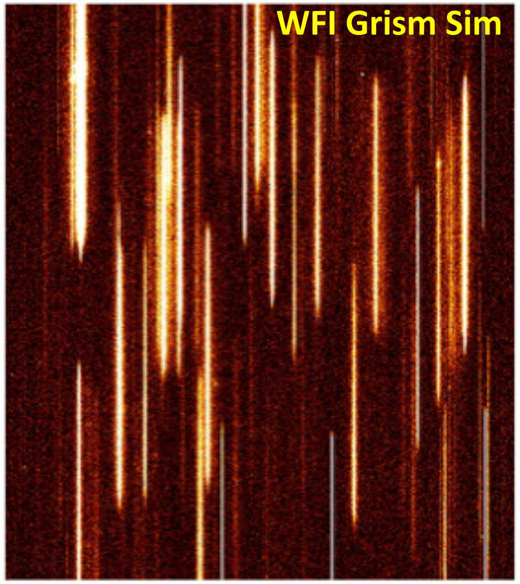
Initial SIT Activities

- Build a standardized and homogeneous (simulated) WFIRST object catalog
  - Stage 1 creates a Level 1 “raw” catalog, and Level 2 for major object types
  - Stage 2 combines these , correlates them with external sources and runs classification tools, clarifies additional requirements on Stage 1 data
- Key outcomes:
  - Quantify how the quality of the object catalog impacts key science measurements;
  - Maximize overlap with other catalogs, characterize how their inclusion improves the science;
  - Learn how to capture and track the proper window functions and the exposure-time maps;
  - Assess the impact of including external multi-wavelength data using simulated and real data.

Development of the Data Analysis Techniques

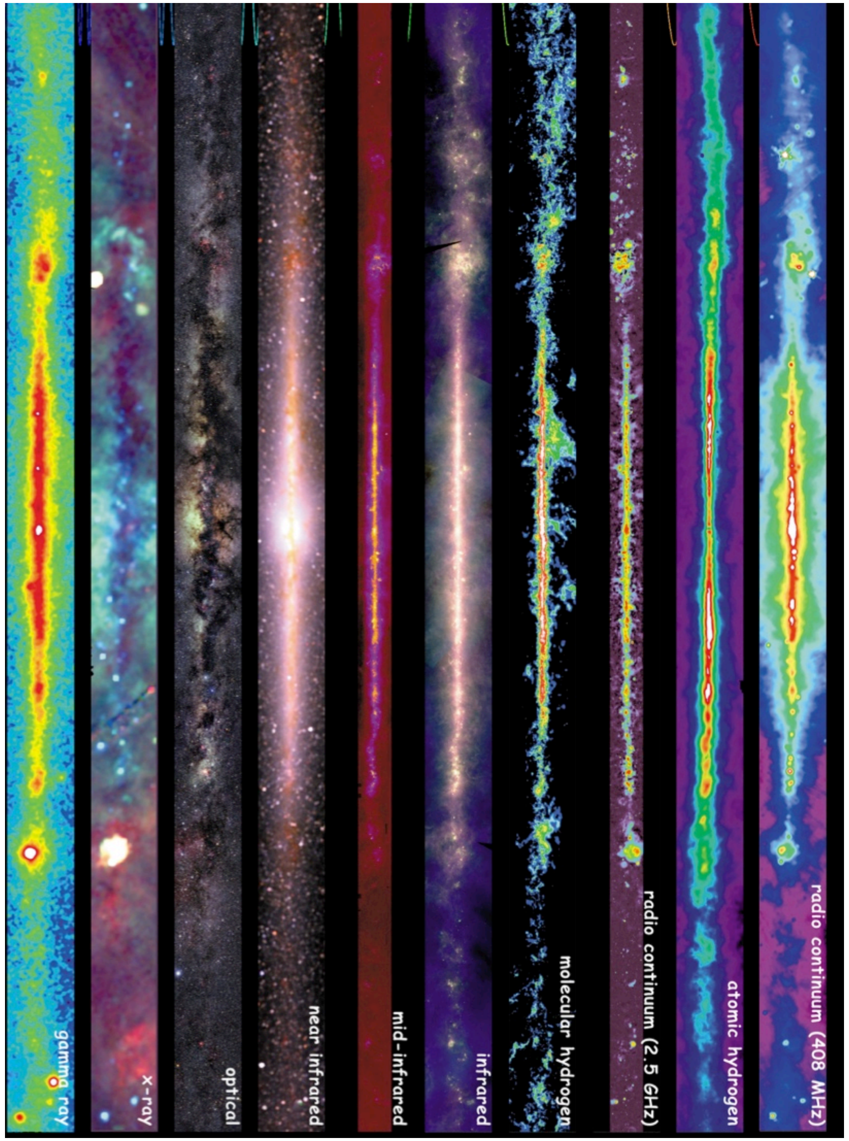
- Advanced Database Framework: ultimately need to go beyond what the typical astronomical data systems offer but start with what we know…
  - Adapt SDSS archive framework to WFIRST parameters to assess challenges.
  - Re-use CasJobs/SciServer system, add advanced scripting to upload analysis.
- Large Simulation Databases
  - Millennium Simulation Database and Millennium Run Observatory – already have access to these.
  - Combine with smaller simulations to mock galaxy morphologies and SEDs
  - Collaborate with other SITs!
- Advanced Object Classification and new "feature vectors"
  - Optimize for kpc-scale galaxy morphology and SED classification (grism emission line) using WFIRST mock data
  - Build and share two feature vector libraries for the above investigations.





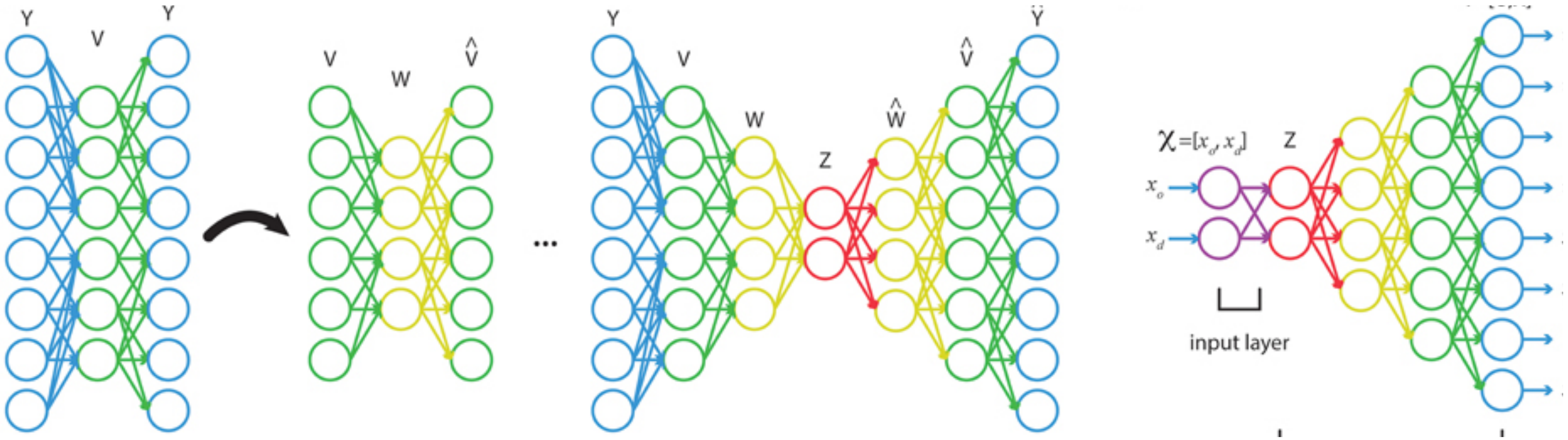
Development of the Data Analysis Techniques

- Large Scale Cross-Correlations
  - Many investigations require auto/cross-correlations
  - Database should be able to perform queries involving 2-point statistics
- Parallel Bayesian Cross-Matching
  - Multi-wavelength properties essential to modern astrophysical analyses: provides the SED
  - Already built parallel Bayesian on-the-fly cross matching tool for SDSS, GALEX, 2MASS, WISE, HSC
  - Runs in minutes on 10^8 object databases
- Signal Injection Architecture
  - Allows users to add specified signals (point source with this brightness and color at this position) to the pipeline and recover results from the Archive for sensitivity and completeness tests.



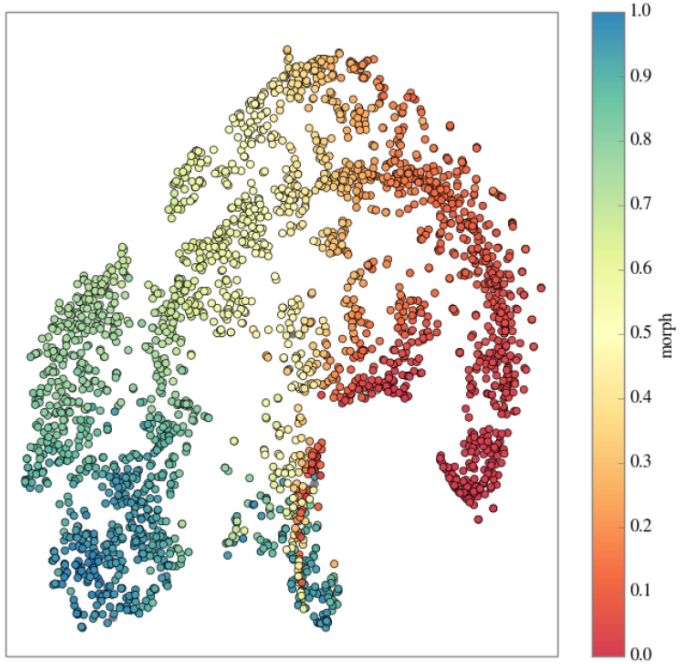
Development of the Data Analysis Techniques

- Server–Side Scripting Machine Learning and 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 can generate easy to investigate data visualizations



***Deep auto-encoder dimensionality reduction**: unsupervised learning methods to take very high dimensional data and reduce them to 2 or 3 key dimensions but without the need to assume linearity.*

http://keplerebs.villanova.edu/tsne



***t-SNE** clustering** applied to Kepler eclipsing binary star data.*

(Matijevic et al. 2012)

**t-Distributed Stochastic Neighbor Embedding*

Our Planned Deliverables over next 5 years:

- A high level unified object catalog across multiple datasets.
- End-to-end simulations from cosmology to grism spectra integrated with a WFIRST archive database (done in collaboration with other SITs).
- Parallel cross matching and cross correlation tools scaling to billions of objects.
- Data visualization tools utilizing ML and dimensionality reduction techniques.
- Tools for precise tracking of footprint overlaps and tracking edge effects.
- Development of complex feature vectors over the unified catalog.
- Development of an advanced classification scheme on top of the unified catalog.
- Signal injection utility for WFIRST galaxy completeness studies. Should be adaptable to other applications (e.g., SNe, stellar populations, etc.)
- Schema requirements, impact of emerging technologies, lessons learned.

Summary

- WFIRST science output will be heavily derived from users interacting with a Petabyte-scale data archive. New regime for a NASA mission.
- We will use a handful of GO/GI WFIRST science cases to identify and develop the best tools and practices needed to extract the science.
- We will use our own simulations as well as those provided by SITs to perform our study. ***Collaborations welcome!***
- Technical focus will be on testing advanced machine-learning classification and scalable, multi-wavelength cross-matching and cross-correlation utilities coupled to data-integrated visualization software.
- Produce and share a prototype, unified WFIRST object catalog suitable for assessing GO+GI archival science use cases with community.

20 Queries Methodology (Gray et al. 2002)

1. Find all galaxies without saturated pixels within 1’ of a given point.
2. Find all galaxies with blue surface brightness between and 23 and 25 magnitude per square arcseconds, and super galactic latitude between (-10, 10), and declination less than zero.
3. Find all galaxies brighter than magnitude 22, where the local extinction is >0.175.
4. Find galaxies with an isophotal surface brightness (SB) larger than 24 in the red band, with an ellipticity>0.5, and with the major axis of the ellipse between 30” and 60”arc seconds (a large galaxy).
5. Find all galaxies with a deVaucouleours profile and the photometric colors consistent with an elliptical galaxy.
6. Find galaxies that are blended with a star and output the deblended galaxy magnitudes.
7. Provide a list of star-like objects that are 1% rare.
8. Find all objects with unclassified spectra.
9. Find quasars with a line width >2000 km/s and 2.5 < redshift < 2.7.
10. Find galaxies with spectra that have an equivalent width in H-alpha > 40Å.
11. Find all elliptical galaxies with spectra that have an anomalous emission line.
12. Create a grided count of galaxies with u-g>1 and r<21.5 over -5<Dec.<5, and 175 < R.A. < 185, on a grid of 2’ arc minutes. Create a map of masks over the same grid.
13. Create a count of galaxies for each of the HTM triangles which satisfy a certain color cut, like 0.7u-0.5g-0.2i<1.25 and r<21.75, output it in a form adequate for visualization
14. Find stars with multiple measurements that have magnitude variations >0.1.
15. Provide a list of moving objects consistent with an asteroid.
16. Find all objects similar to the colors of a quasar at 5.5<redshift<6.5.
17. Find binary stars where at least one of them has the colors of a white dwarf.
18. Find all objects within 30 arcseconds of one another that have very similar colors: that is where the color ratios u-g, g-r, r-i are less than 0.05m.
19. Find quasars with a broad absorption line in their spectra and at least one galaxy within 10 arcseconds. Return both the quasars and the galaxies.
20. For each galaxy in the LRG data set (Luminous Red Galaxy), in 160<right ascension<170, count of galaxies within 30"of it that have a photoZ within 0.05 of that galaxy.