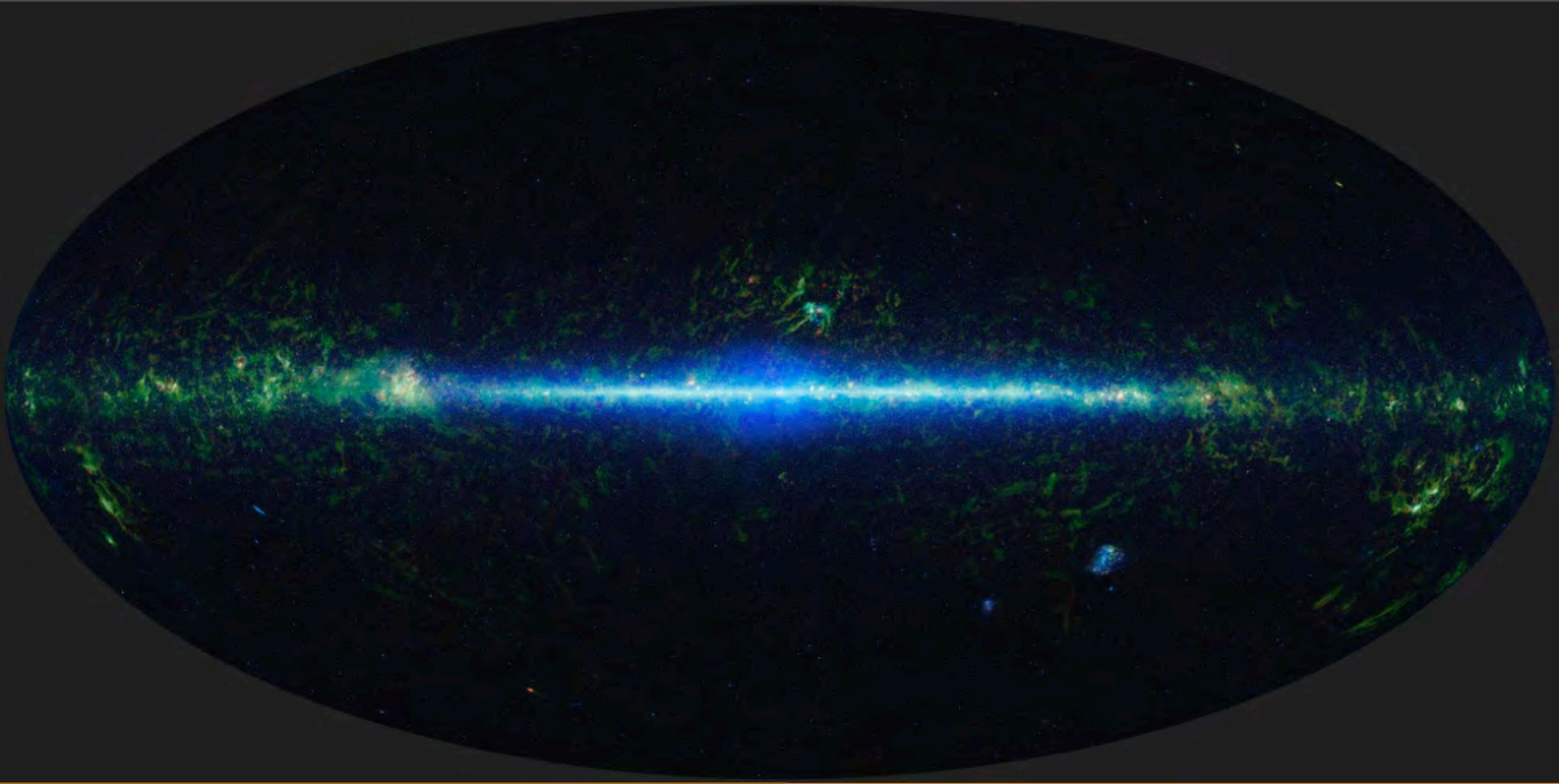


Astronomical Data Archives

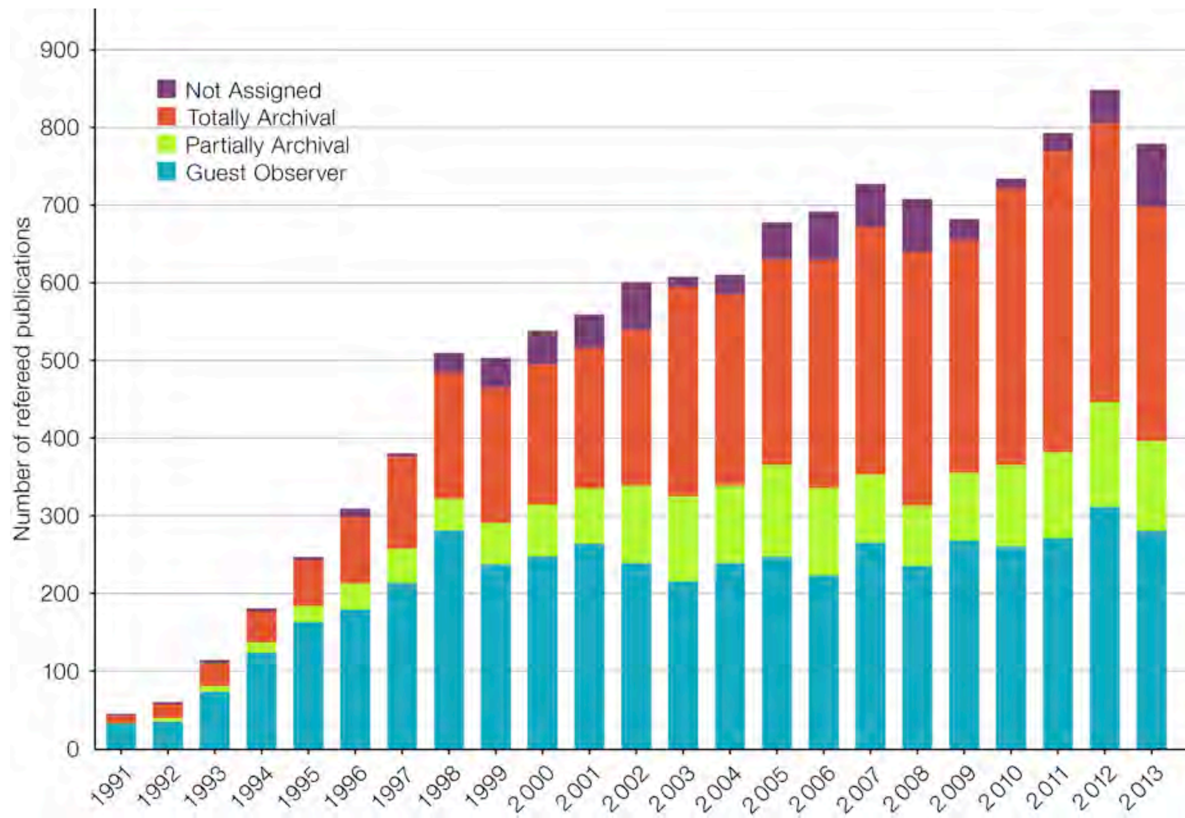
Rachel Akeson
Caltech/IPAC

17 January 2018

1. Archive Goals
2. Current Archives
3. Current and Upcoming Challenges

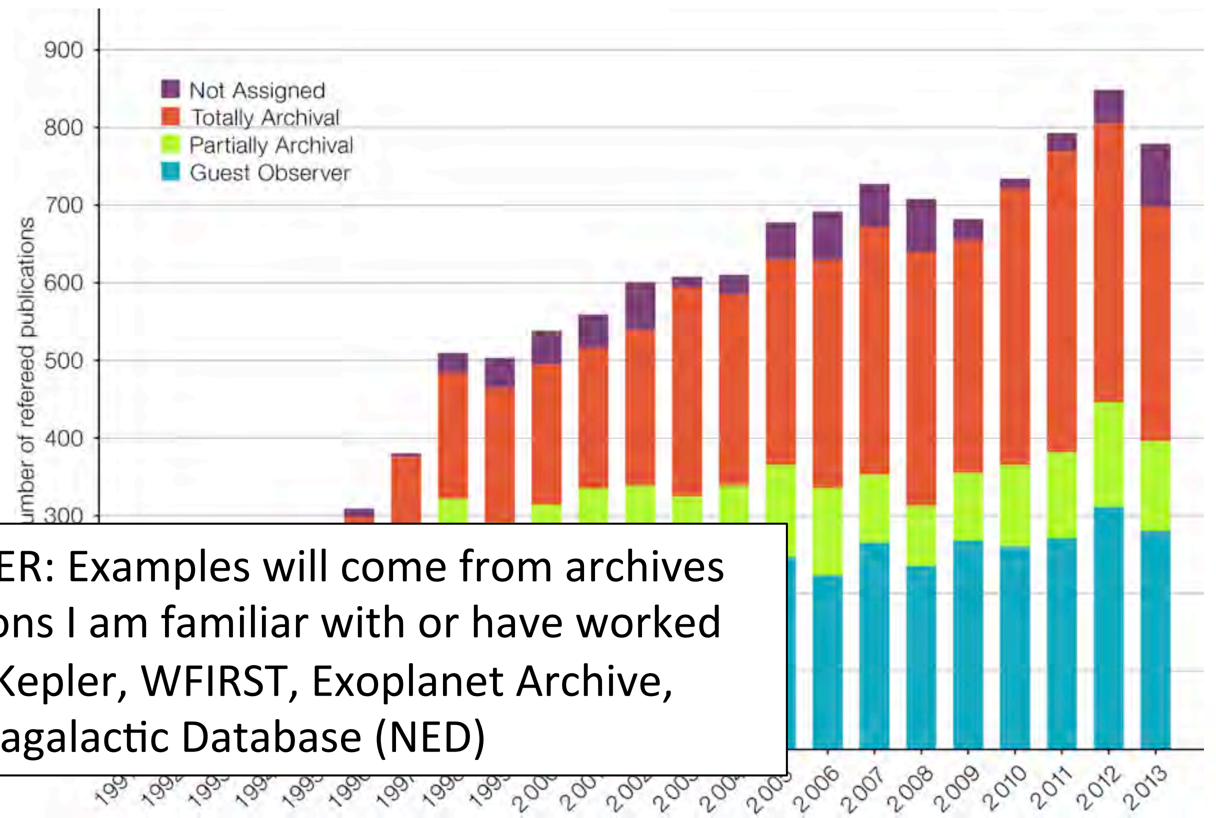
Astronomy Archive Goals

- Data safe-keeping
  - Tremendous financial and community resources are put into gathering data, so some effort in preserving data for the long run is justified
  - Example: Infrared Astronomical Satellite data (IRAS; launched in 1983) was still cited times more than 200 times in 2017
- PI access
- Archival research
  - For Hubble, Archival research (i.e., not from the proposing team, comprises ~half of all publications)



Astronomy Archive Goals

- Data safe-keeping
  - Tremendous financial and community resources are put into gathering data, so some effort in preserving data for the long run is justified
  - Example: Infrared Astronomical Satellite data (IRAS; launched in 1983) was still cited times more than 200 times in 2017
- PI access
- Archival research
  - For Hubble, Archival research (i.e., not from the proposing team, comprises ~half of all publications)



DISCLAIMER: Examples will come from archives and missions I am familiar with or have worked on: Keck, Kepler, WFIRST, Exoplanet Archive, NASA Extragalactic Database (NED)

Archival Research Example

- Easily searchable data facilitates innovative uses
  - Example: Ogle et al discover super spiral galaxies using NED (NASA Extragalactic Database)
    - These galaxies have luminosities 10x the Milky Way
  - No new data taken
  - Original goal was to search of properties of 800,000 galaxies ($z < 0.3$) to find luminous elliptical brightest cluster galaxies
  - Database combines information from
    - Galaxy Evolution Explorer (Galex)
    - Sloan Digital Sky Survey
    - Two Micron All-Sky Survey (2MASS)
    - Spitzer
    - Wide-field Infrared Survey Explorer (WISE)

Current Archives

- Archives contain:
  - Raw and processed data from the mission/team
  - High level data products from the community
  - High level products/associations from the archive (e.g. cross identifications)
- Archive levels and support vary greatly, from non-searchable raw data to processed data with visualization and tools
- Congressional mandate on federally-funded data has accelerated the trend of PI-produced data being made available to the community
- Data release models
  - Continuous: most GO driven telescopes/missions, with or without proprietary period
  - Episodic: Generally when processing is needed on sets of data (spatial or temporal)

For recent updates and developments: Astronomical Data Analysis Software & Systems conference series

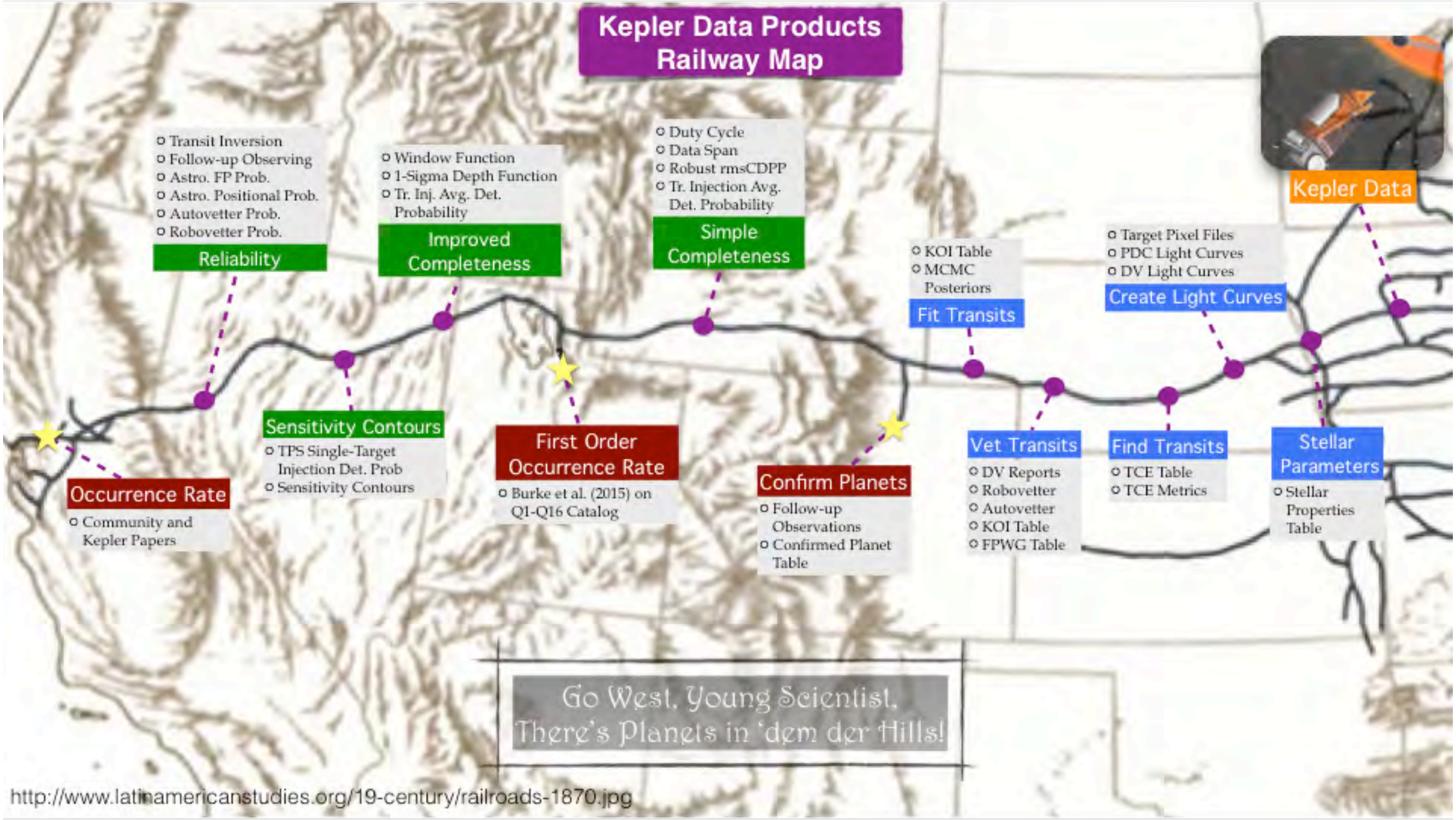
Current Archives: General Categories

- Space missions
  - NASA, ESA, JAXA etc. generally provide high level data archives including processed data and often analysis tools
  - Funding to develop and operate the archive is generally allocated as part of the mission planning process
- Ground-based telescopes
  - Optical
    - Most large telescopes have archives (Keck, VLT/ESO, Gemini etc)
    - Many small and medium sized telescopes do not have open archives
  - Radio/millimeter
    - NRAO, ALMA

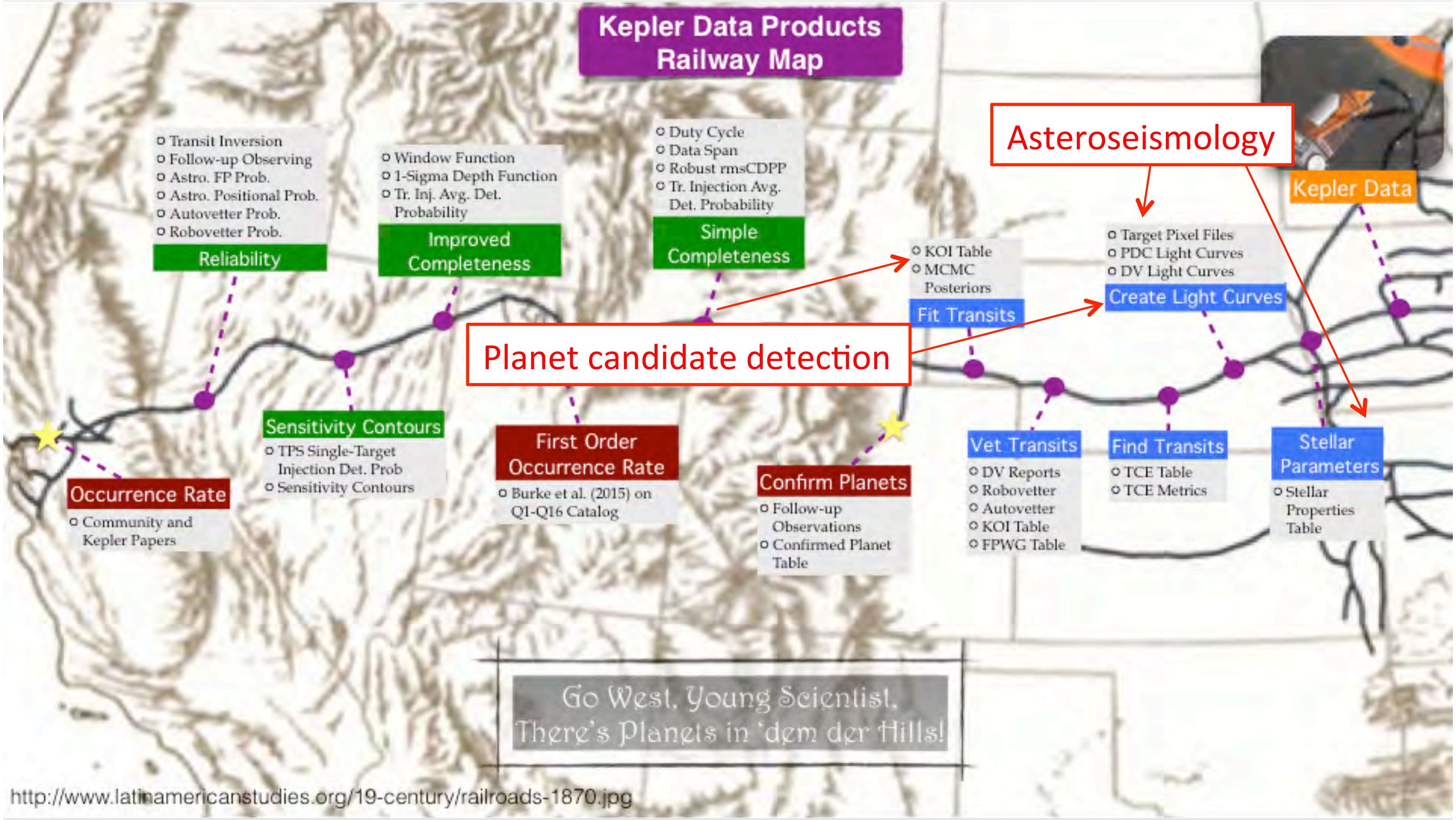
Ancillary Data and Additional Products

- Need to understand what additional data is crucial for data processing or interpretation
  - For processing
    - Spacecraft data (even if it doesn't seem immediately relevant)
      - Example: Spitzer developed an exoplanet transit mode ~8 years after launch using pointing data not originally used in processing pipeline
  - For interpretations/astrophysics
    - Example: Source categorizations
    - More likely to come from community, issues with standardization and reliability
    - LISA example: Input waveforms
- Other kinds of data products
  - Completeness and reliability or pipeline characterization data
  - Simulated data and inputs
  - Probabilities and posteriors

Example: Kepler



Example: Kepler



Current and Upcoming Challenges: Data Volume (1)

- By Earth Sciences or Silicon Valley standards, Astronomy is not "Big Data"
- BUT, data volume is still an issue
  - Current examples
    - WISE all sky survey
      - Source catalog ~ 800 million sources
      - Photometry measurements ~ 42 billion
    - IPAC total holdings = 12 PB
  - Upcoming mission examples
    - LSST max data rate = 20 TB/night (2020)
    - WFIRST max data rate = 1.5 TB/day (2026)
    - SKA (Square Kilometer Array) = 20 TB/day processed data (2024, partial)
- Issues with volume:
  - Retrieval
  - Organization/Searching
  - Processing

Current and Upcoming Challenges: Data Volume (1)

- By Earth Sciences or Silicon Valley standards, Astronomy is not "Big Data"
- BUT, data volume is still an issue
  - Current examples
    - WISE all sky survey
      - Source catalog ~ 800 million sources
      - Photometry measurements ~ 42 billion
    - IPAC total holdings = 12 PB
  - Upcoming mission examples
    - LSST max data rate = 20 TB/night (2020)
    - WFIRST max data rate = 1.5 TB/day (2026)
    - SKA (Square Kilometer Array) = 20 TB/day processed data (2024??)
- Issues with volume:
  - Retrieval → Some users want (or think they want) the entire survey. Adds significant logistics and data management work.
  - Organization/Search
  - Processing

Data Volume: Organization/Searching

- Lessons learned
  - Simple solutions with careful planning can get you a long way without much technical magic: ***organization*** **is the magic!**
  - Reduce s/w overheads: small latencies that didn't use to matter now stand out.
  - Beware complexity - things that are complicated when they are small explode on you when they grow big.
  - Large datasets are hard to move: try to get it right the first time, and consider moves carefully.
  - Large databases are hard to change or update, so plan the content carefully before loading.
  - Optimize data layouts for most common use cases: **But** different use cases require different organizations.
  - May need to consider indexing in space-time, rather than just space, for moving object applications.

Data Volume: Organization/Searching

- Lessons learned
  - Simple solutions with careful planning can get you a long way without much technical magic: ***organization* is the magic!**
  - Reduce s/w overheads: small latencies that didn't use to matter now stand out.
  - Beware complexity - things that are complicated when they are small explode on you when they grow big.
  - Large datasets are hard to move: try to get it right the first time, and consider moves carefully.
  - Large databases are hard to change or update, so plan the content carefully before loading.
  - Optimize data layouts for most common use cases: **But** different use cases require different organizations.
  - May need to consider indexing in space-time, rather than just space, for moving object applications.

Data Volume: Processing

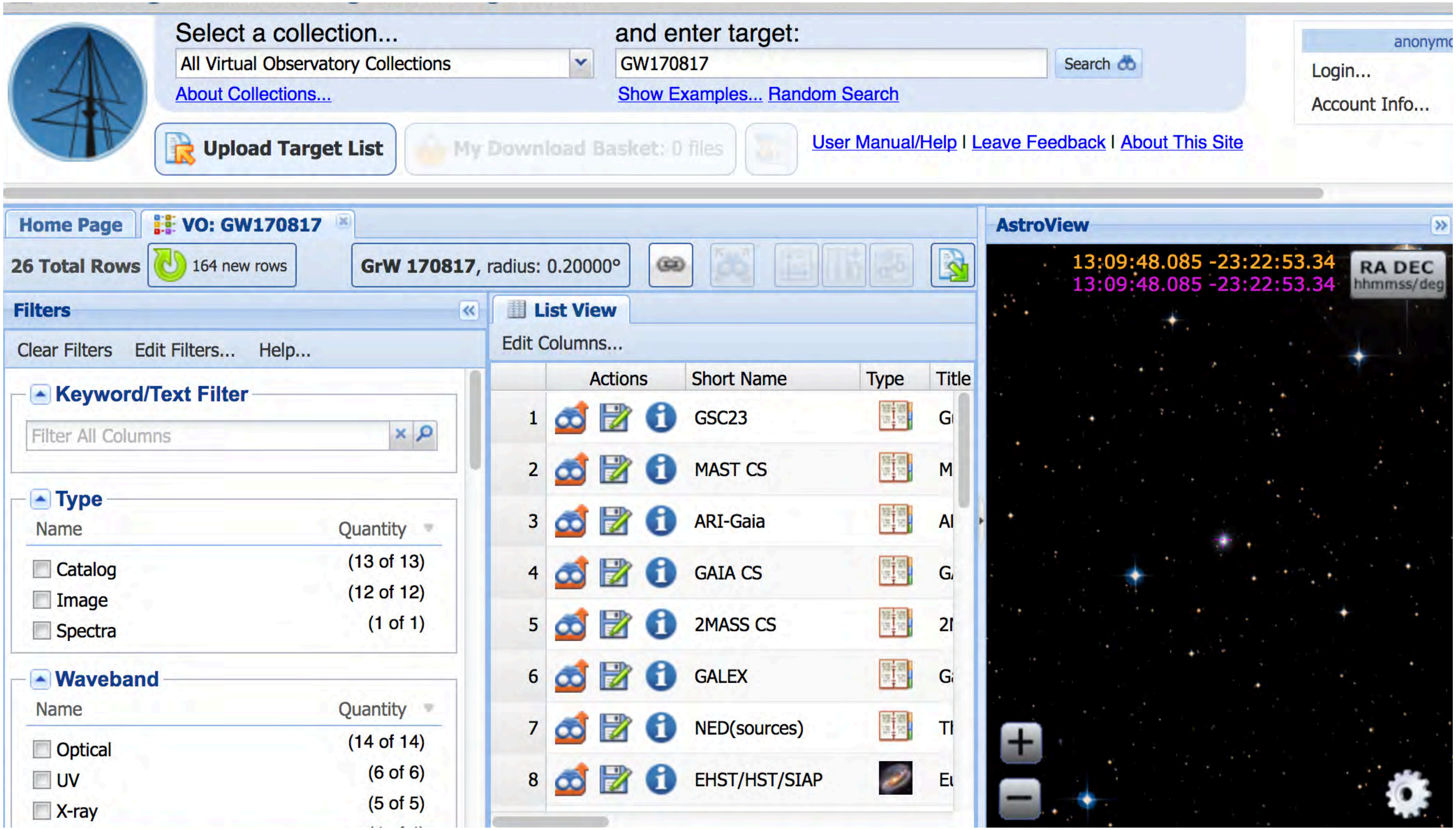
- Large data volumes drive the need to process the data where they are stored
  - Not all users will have the resources to store and/or process large volumes
  - Also an issue for heavy processing on mid-sized data
- Different models
  - Give users CPU resources on same system as archive, generally comes with a version of the pipeline which may may editable by user
  - Store copy of the data somewhere with large processing capabilities
    - Cloud
    - Super computer center
  - Users often want to start at intermediate processing level

- Lesson learned: Best if planned and developed with rest of telescope
  - Even if not funded immediately, having a design minimizes the chance that later capabilities will be precluded or cost much more to develop

Current and Upcoming Challenges: Multi-mission Research and Interoperability

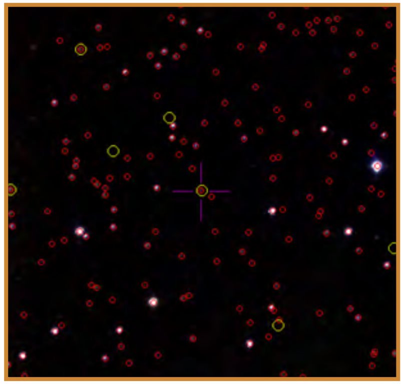
- Many areas of research require coming several wavelengths/mission/ measurement types to make progress on the astrophysics
- Given the many different sources of archive funding (and different mandates/ requirements) it is unlikely that any group would be funded to store ALL astronomy data
- Virtual Observatory efforts
  - IVOA and national VO organizations have worked to set standards and provide tools and services to allow users to access data from multiple host archives
  - Defining data standards gets progressively harder as data types become more complicated, e.g. data tables vs spectral cubes from an integral field spectrograph
  - Implementation within archive done with resources from individual archives
- Lesson learned: Need to design good interfaces for users to retrieve data with complex constraints as science cases will always be incomplete

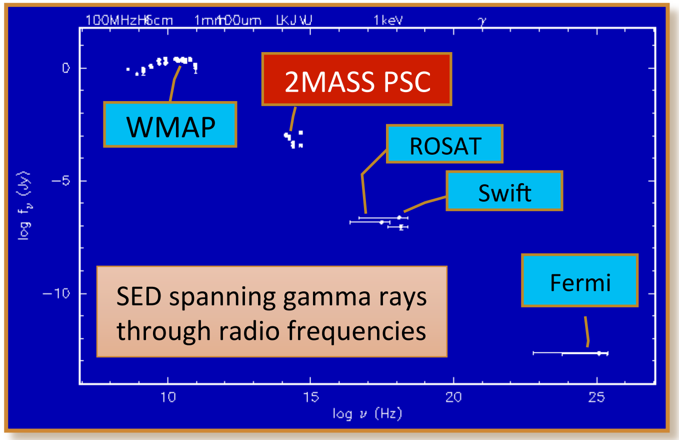
VO Example: GW170817



Current and Upcoming Challenges: Cross Matching/Identification

- One of the most important functions of an archive is matching sources between catalogs
  - Many different levels
    - Coordinate only
    - Coordinate + astrophysics
    - Measure of reliability

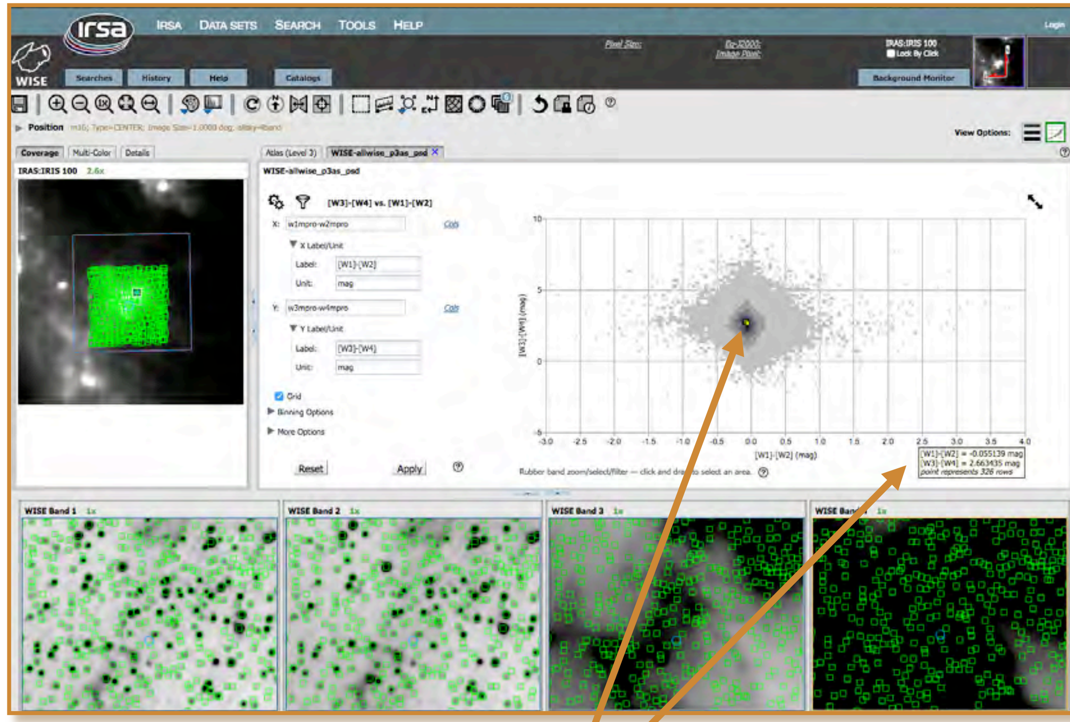




- Example:
  - NED: (NASA Extragalactic Database)
- Must correlate newly ingested catalogs with existing database of ~250 million objects.
- Currently ingesting 2MASS catalog with 470 million sources.
- Machine learning used for assigning statistical probabilities to matches

Current and Upcoming Challenges: Data Visualization

- Interactive graphics provide intuition about the data.
- Co-registration of data sets: Example: IRSA allows simultaneous viewing of different data sets.
- Time-domain: light curves, folded-viewing, periodograms, moving objects.
- For massive sets we have to go from symbol representation to continuous quantities: density plots, histograms.
- Data Cubes



IRSA Viewer uses a density plot when the number of points becomes too great to show individually. The number of points in each bin in the plot is provided on hover.

What does LISA need?

- Is the LISA catalog a table with the same parameters for all sources/detections?
- Consider hierarchical or on-demand products
  - Fermi
    - User driven generation of products based on current data
    - See info from Anne on Google drive
  - Kepler
    - Different products for different use cases
    - Additional/ancillary data varies depending on object
- LISA can use a combination of previous and current formats and structures and custom ones
- Beware of overly complex products
  - i.e. if a table is mostly empty, it may be more useful as multiple products
- High level archive design considerations
  - Collection of products at each processing level vs single series of products
  - What is generated by project/science team and what is contributed by community

Takeaway Points

- A well designed archive enhances what the community can do
  - Design depends on both on raw and processed data types, but also the most common use cases
- There are many different archive types and organizational models to choose from
- Be optimistic in archive plan but be prepared for reality
  - Include many use cases, but prioritize
  - Plan for future expansion
  - Even if processing development has to be delayed, put effort into design and use cases
- Use whatever you can from the community
  - It's nice to think about doing "everything right" but resources are always finite