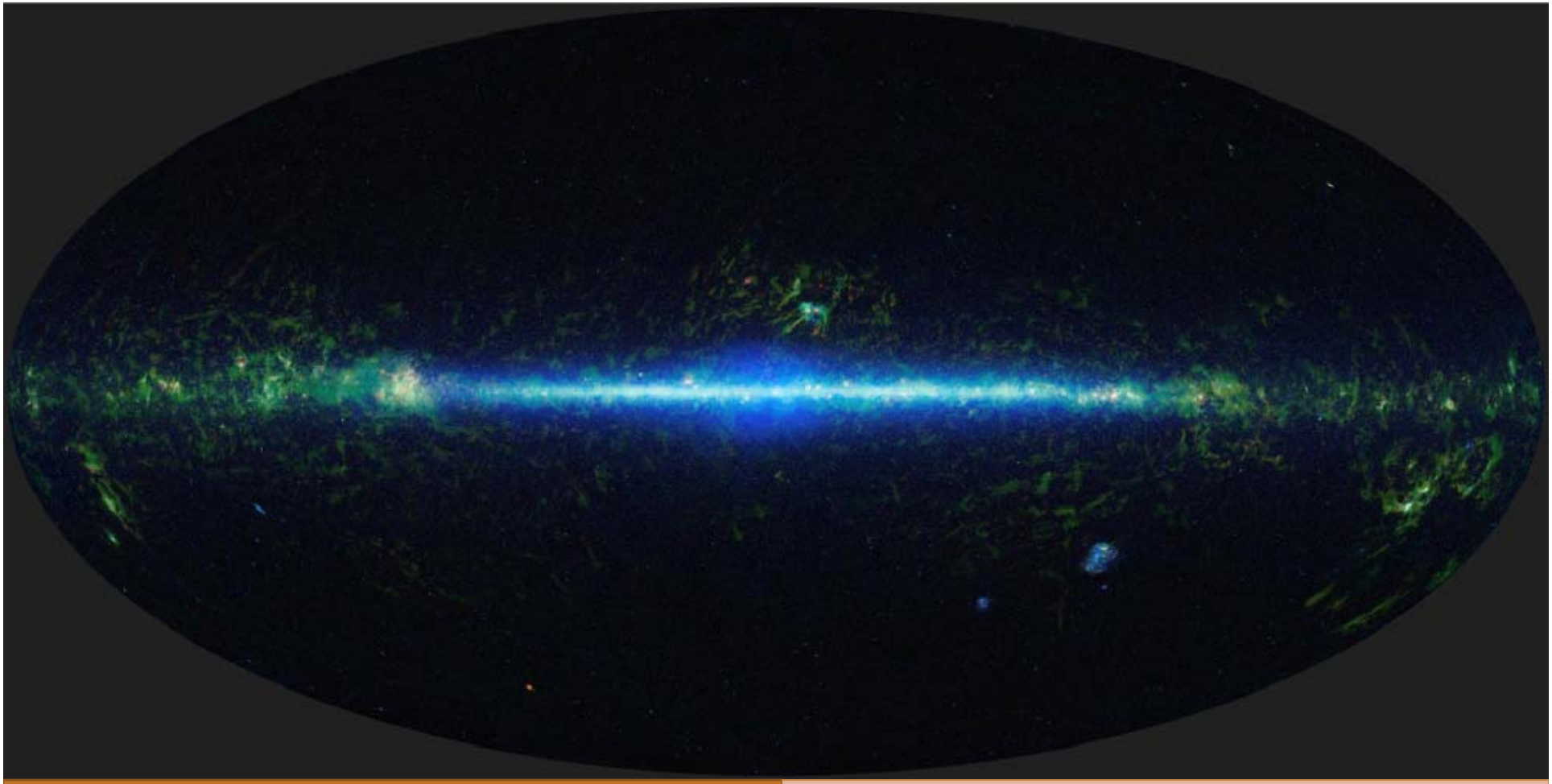


Big Data at IPAC

David A. Imel, Manager
Caltech/IPAC

2 November 2017

1. Introduction to IPAC
2. Big Data in the IPAC Context
3. Lessons Learning at IPAC
4. Opportunities and Challenges

IPAC: Caltech Astrophysics Science Center

- Science Center functions for NASA missions
- Data centers for major projects such as Great Observatories and all-sky surveys
- Supports NASA, NSF and privately funded projects
- Award-winning media, outreach and education support
- Vibrant research environment and staff

Cosmology and galaxy evolution

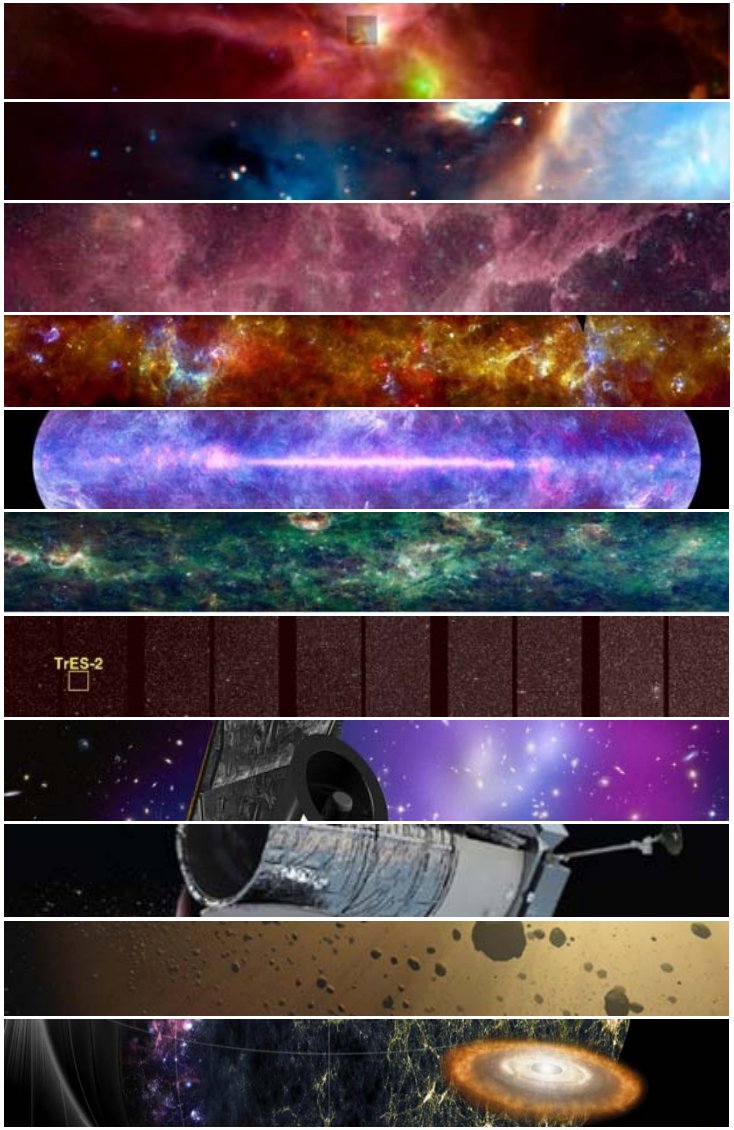
Exoplanets

Asteroids and the solar system

Infrared-submillimeter astrophysics.

32 Years of Science Operations for NASA Missions

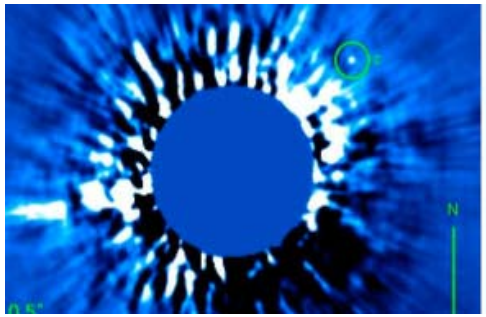
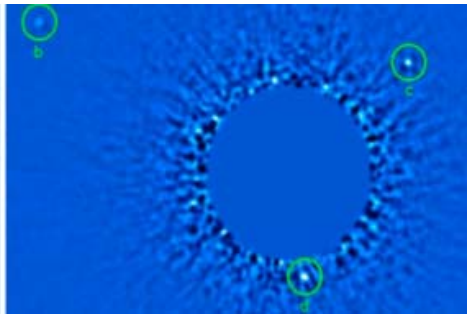
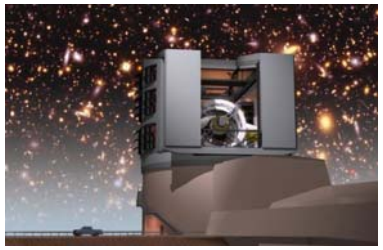
Mission	IPAC Role
IRAS	Science Data Center
ISO	US Data Center
Spitzer	Full Science Operations for a NASA Great Observatory
Herschel	US Science Center
Planck	US Science Data Center
WISE / NEOWISE	Science Data Center
Kepler/K2/TESS	Candidate and Confirmed Planet Archive, Follow-up Observing Program
Euclid	US Science Data Center
WFIRST	Joint Science Operations
NEOCAM	Science Data Center
SPHEREx	Science Data Center



Science Operations for Ground-Based Observatories

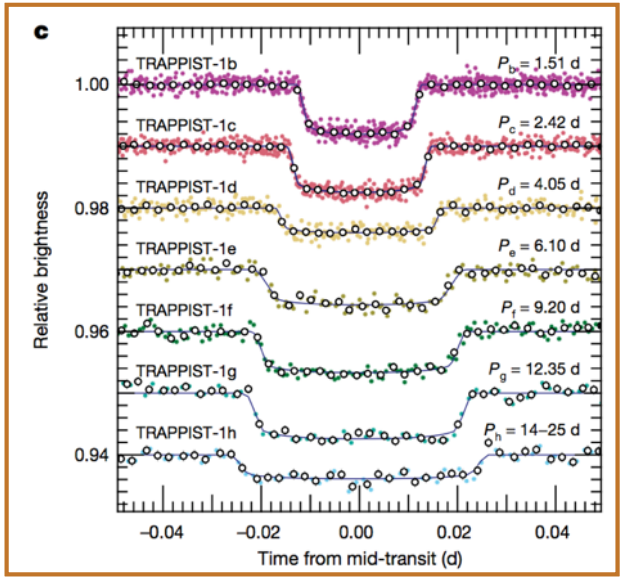
Observatory	IPAC Role
2MASS	Data processing; Archive
Keck Interferometer	Observation planning; Data processing; Archive
Keck Observatory	Browse product pipelines; Archive
Large Binocular Telescope Interferometer	Archive
Palomar Transient Facility	Nightly Ingest; Data processing; Archive
Zwicky Transient Facility	Nightly Ingest; Data Processing; Alerts; Archive
Large Synoptic Survey Telescope	Science Platform & User Interface




Spitzer Finds Seven Earth-Size Planets in One System

- M-Dwarf, 10% R_{sun} 12 pc from here, first observed in 2000
- 2015: ground-based observing yields 3 Earth-sized planets
- System observed by Spitzer in 2016–2017: 7 earth-size planets.
  - Transit observations give planet sizes.
  - With resonances and timing variations, masses can be estimated.
  - Combination gives density: rocky, volatiles, probably water!



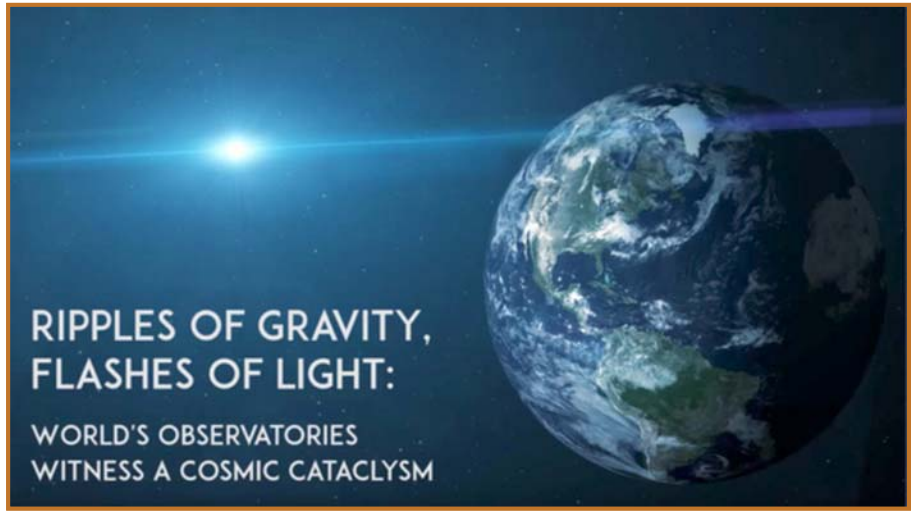
IPAC Astronomy 2017: Debris Disks and Colliding Neutron Stars

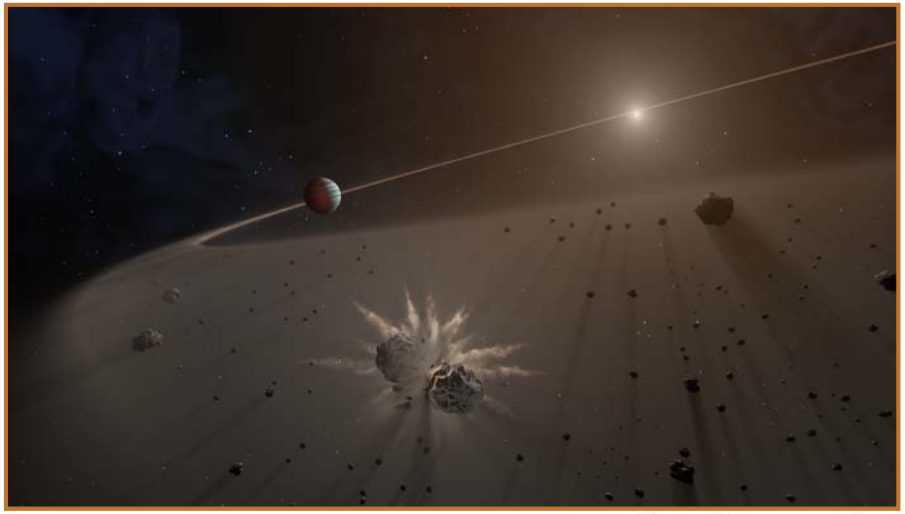
Gravitational Waves!

- IPAC Executive Director Dr. George Helou and IPAC scientist Dr. Lin Yan were part of the team confirming the source of the recent observation of gravity waves from colliding neutron stars.
  - First co-observations of gravity waves with EM-spectrum signature.
  - Data from IPAC's NED cited in discovery papers.
  - ICE team developed key graphics for public communication of results.

Exoplanets in Debris Disks

- IPAC scientist Dr. Tiffany Meshkat finds that giant exoplanets that orbit far from their stars are more likely to be found around young stars that have a disk of dust and debris than those without disks.
  - Spitzer data on debris disk systems vs. non-debris disk systems; scanned for exoplanets
  - Combined with data from Keck and ESO VLT.
- Result helps JWST and other missions plan where to look for exoplanets.

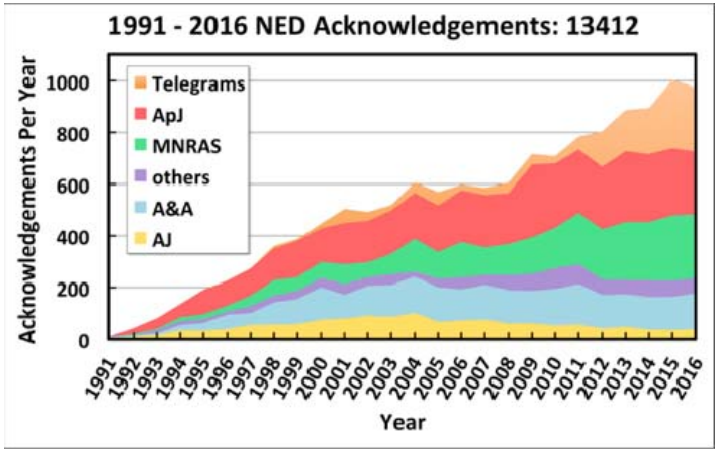




NASA/IPAC Extragalactic Database

NED serves as NASA's "Google for Galaxies"

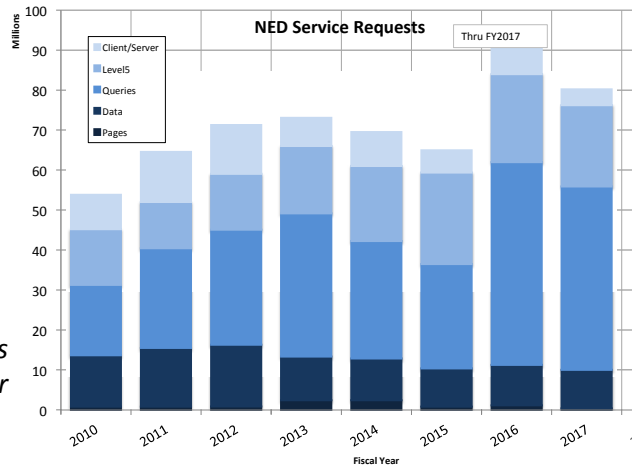
NED citation rate is comparable to Hubble



NED is:

- *Comprehensive*
- *Reliable*
- *Easy-to-use*
- *Synthesis of multi-wavelength data*
- *Content linked to refereed literature*
- *Data augmented with derived physical attributes*

Every day NED serves over 70,000 queries and is acknowledged by 2 peer-reviewed articles.



80 million queries this year

Published:

- Names
- (α,δ)
- Redshifts
- D_{Mpc}
- Fluxes
- Sizes
- Attributes
- References
- Notes

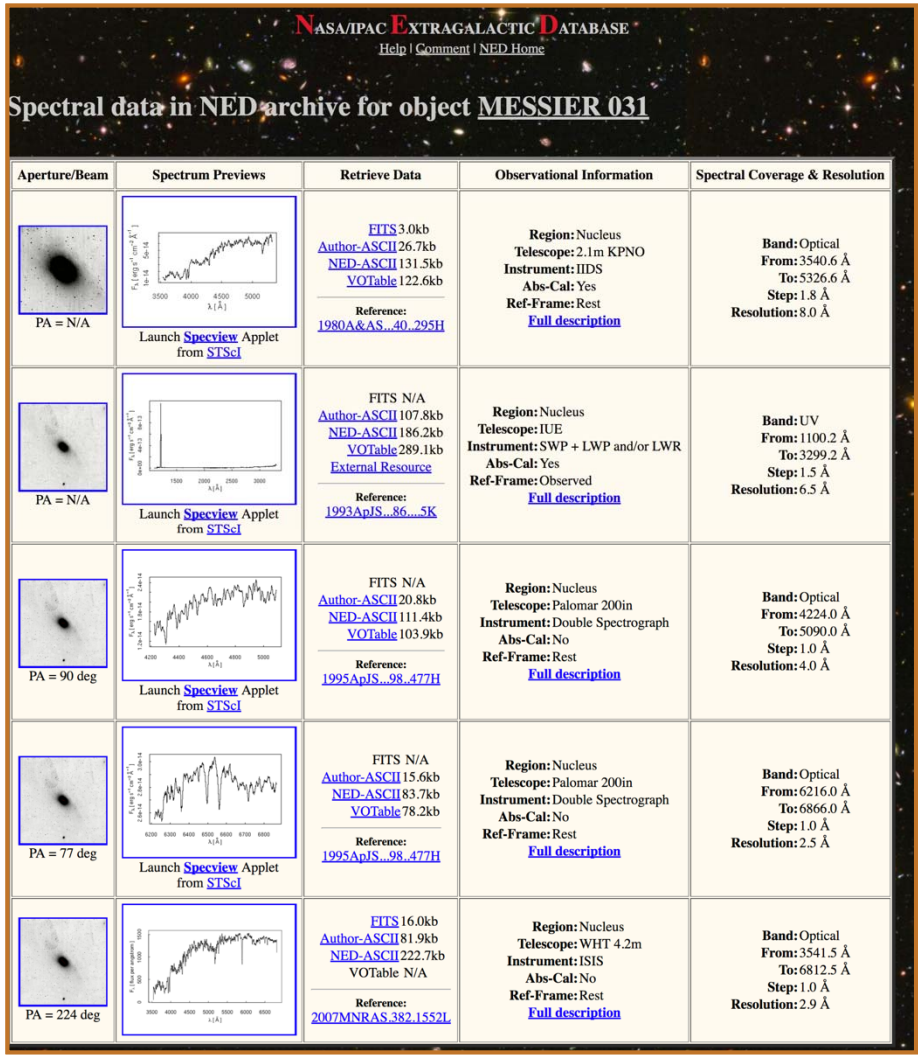
Contributed:

- Images
- Spectra

Derived:

- *Distances*
- *Metric sizes*
- *Luminosities*
- *Velocity corrections*
- *Cosmological corrections*
- *SEDs*
- *A_λ*

NED Is Building the Census of the Universe



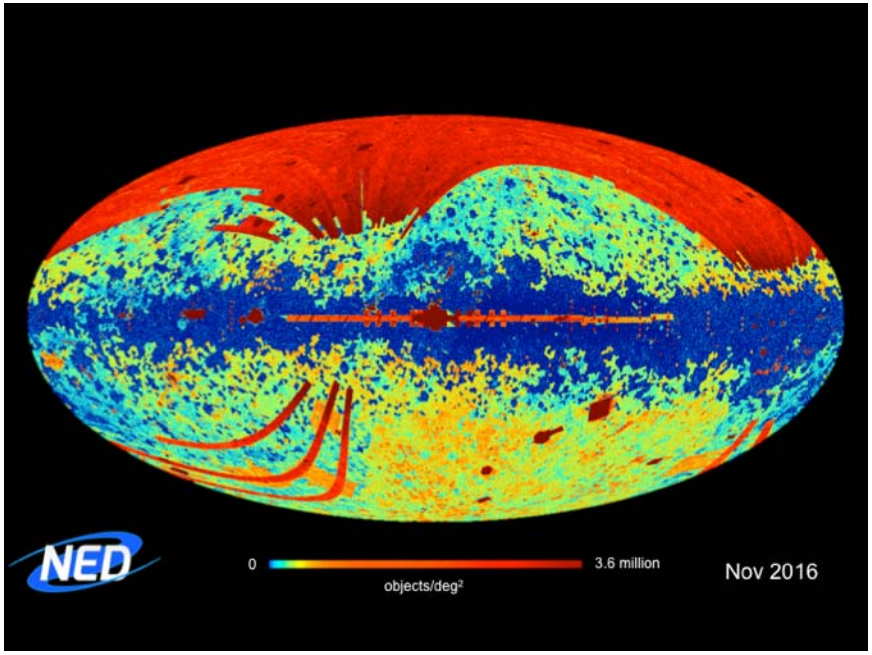
NASA/IPAC EXTRAGALACTIC DATABASE

Help | Comment | NED Home

Spectral data in NED archive for object MESSIER 031

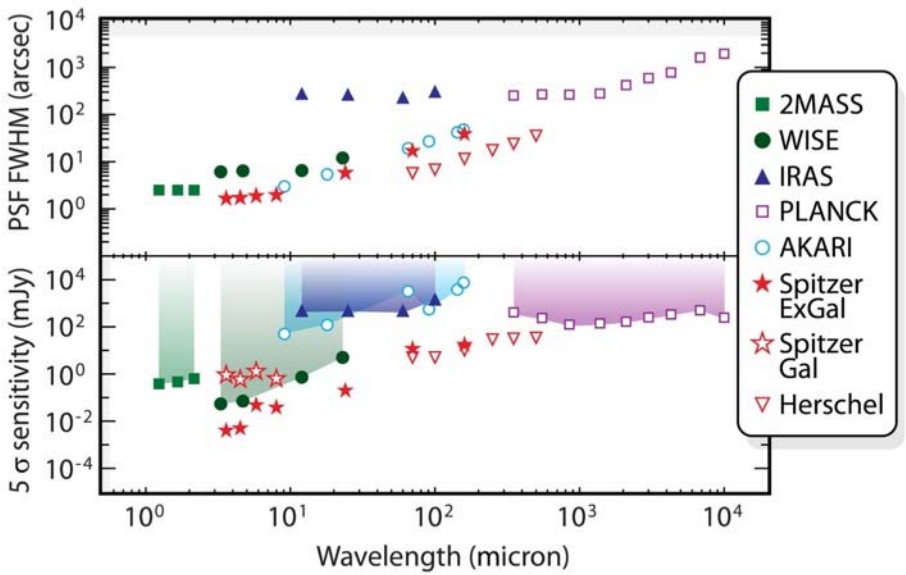
Aperture/Beam	Spectrum Previews	Retrieve Data	Observational Information	Spectral Coverage & Resolution
PA = N/A	Launch Specview Applet from STScI	FITS 3.0kb Author-ASCII 26.7kb NED-ASCII 131.5kb VOTable 122.6kb Reference: 1980A&AS...40..295H	Region: Nucleus Telescope: 2.1m KPNO Instrument: IIDS Abs-Cal: Yes Ref-Frame: Rest Full description	Band: Optical From: 3540.6 Å To: 5326.6 Å Step: 1.8 Å Resolution: 8.0 Å
PA = N/A	Launch Specview Applet from STScI	FITS N/A Author-ASCII 107.8kb NED-ASCII 186.2kb VOTable 289.1kb External Resource Reference: 1993ApJS...86....5K	Region: Nucleus Telescope: IUE Instrument: SWP + LWP and/or LWR Abs-Cal: Yes Ref-Frame: Observed Full description	Band: UV From: 1100.2 Å To: 3299.2 Å Step: 1.5 Å Resolution: 6.5 Å
PA = 90 deg	Launch Specview Applet from STScI	FITS N/A Author-ASCII 20.8kb NED-ASCII 111.4kb VOTable 103.9kb Reference: 1995ApJS...98..477H	Region: Nucleus Telescope: Palomar 200in Instrument: Double Spectrograph Abs-Cal: No Ref-Frame: Rest Full description	Band: Optical From: 4224.0 Å To: 5090.0 Å Step: 1.0 Å Resolution: 4.0 Å
PA = 77 deg	Launch Specview Applet from STScI	FITS N/A Author-ASCII 15.6kb NED-ASCII 83.7kb VOTable 78.2kb Reference: 1995ApJS...98..477H	Region: Nucleus Telescope: Palomar 200in Instrument: Double Spectrograph Abs-Cal: No Ref-Frame: Rest Full description	Band: Optical From: 6216.0 Å To: 6866.0 Å Step: 1.0 Å Resolution: 2.5 Å
PA = 224 deg		FITS 16.0kb Author-ASCII 81.9kb NED-ASCII 222.7kb VOTable N/A Reference: 2007MNRAS.382.1552L	Region: Nucleus Telescope: WHT 4.2m Instrument: ISIS Abs-Cal: No Ref-Frame: Rest Full description	Band: Optical From: 3541.5 Å To: 6812.5 Å Step: 1.0 Å Resolution: 2.9 Å

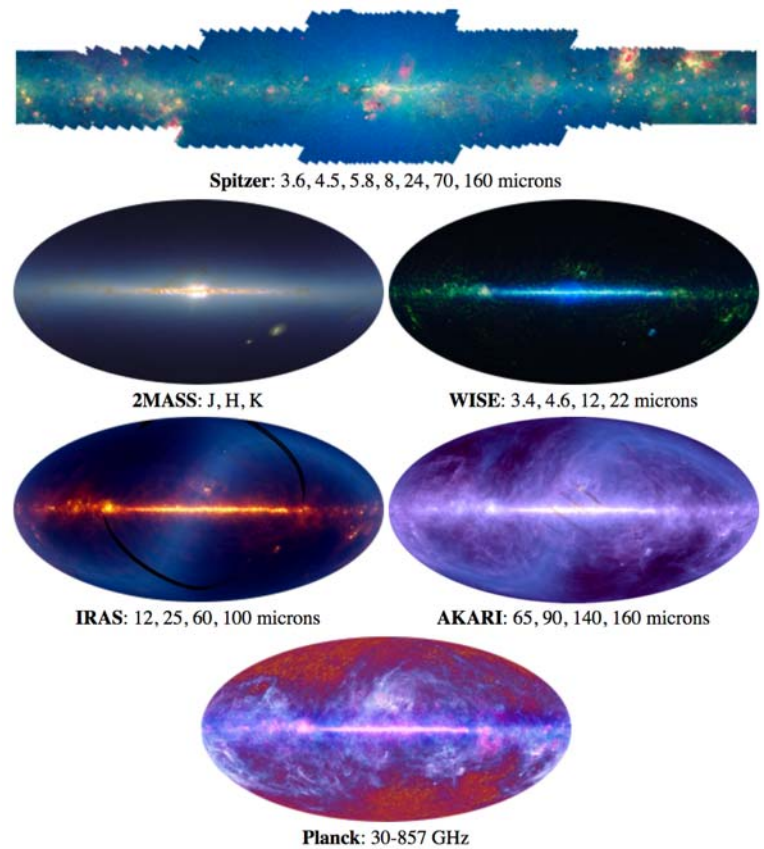
- 0.5 billion catalog sources
- 2.3 billion photometric data points
- 250 million distinct objects
- 40 million links to Journal Articles
- 8 million redshifts
- 2.5 million images



NASA/IPAC Infrared Science Archive

IRSA curates the science products of NASA's infrared and submillimeter missions.

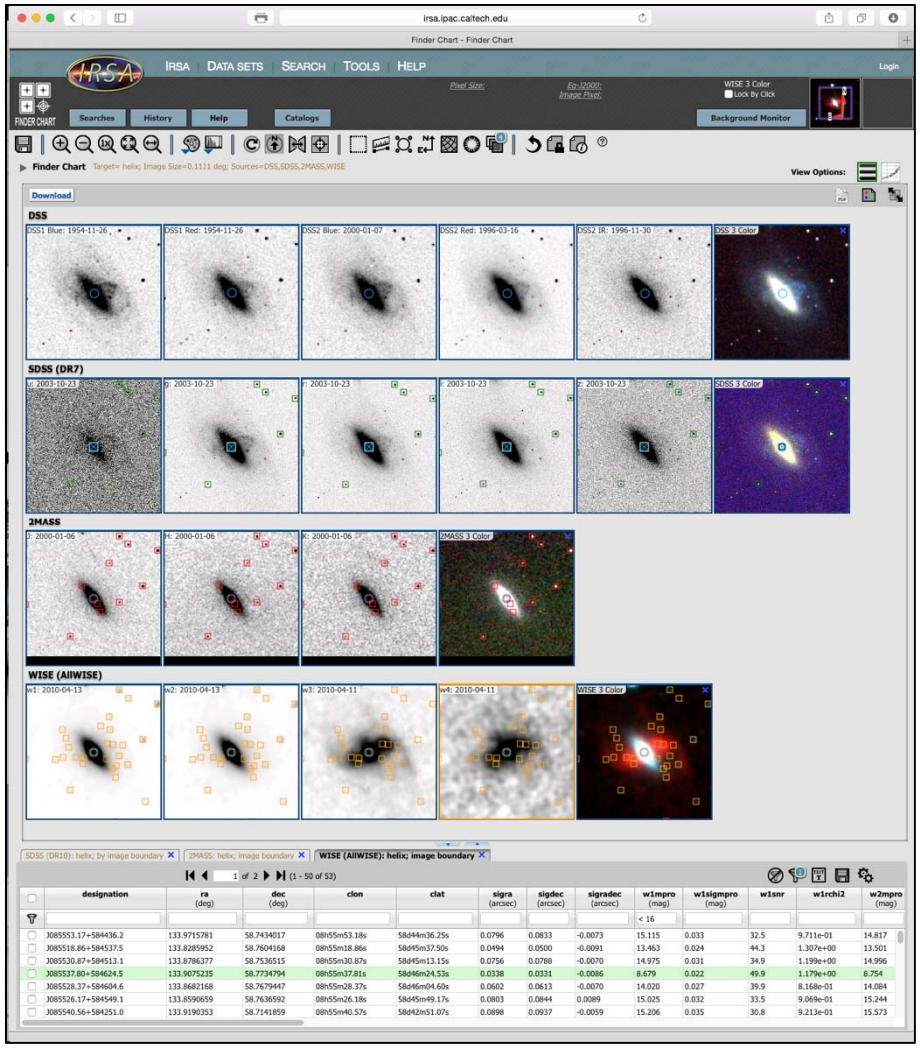




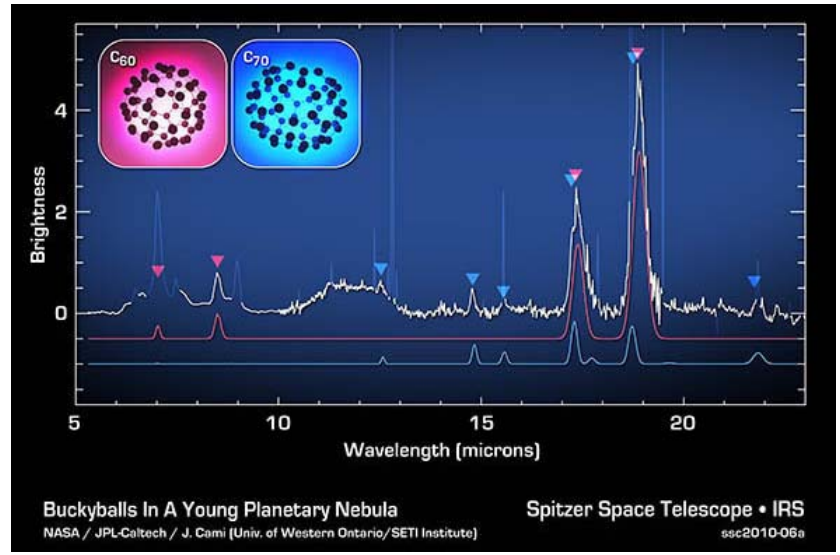
IRSA provides all sky images and catalogs covering 24 wavelength bands.

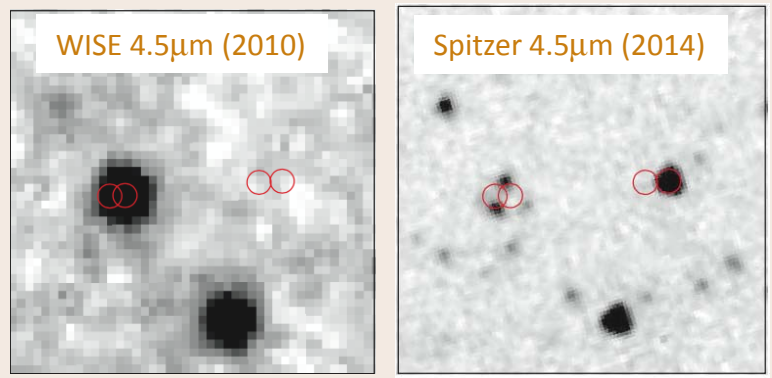
Astrophysics Research with IRSA

IRSA integrates catalogs, images, and spectra from many missions to enable new science discoveries.



Search & display can be tailored to various instrument/science contexts, using reusable visualization components.

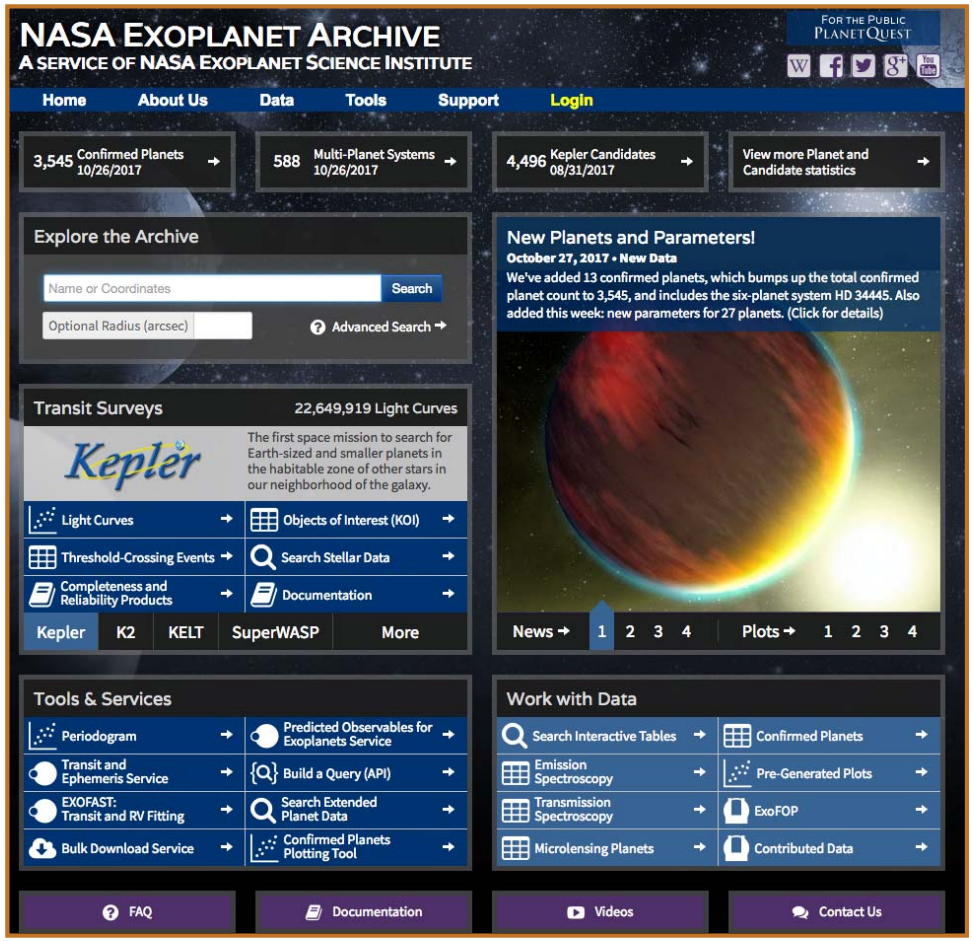




WISE+Spitzer discover the coldest brown dwarf (Luhman 2014, ApJL 786, L18)

NASA Exoplanet Archive

NASA's Official Exoplanet Database



Data Holdings

- **Confirmed exoplanets**
  - 80,000+ planetary and stellar parameters for >3500 exoplanets
  - Weekly updates
- **Kepler**: ~4500 planets and candidates, stellar properties, data validation and occurrence rate products
- Space (CoRoT) and ground-based **transit surveys** (>20 million light curves)
- Community-contributed follow-up observing data (**ExoFOP**)

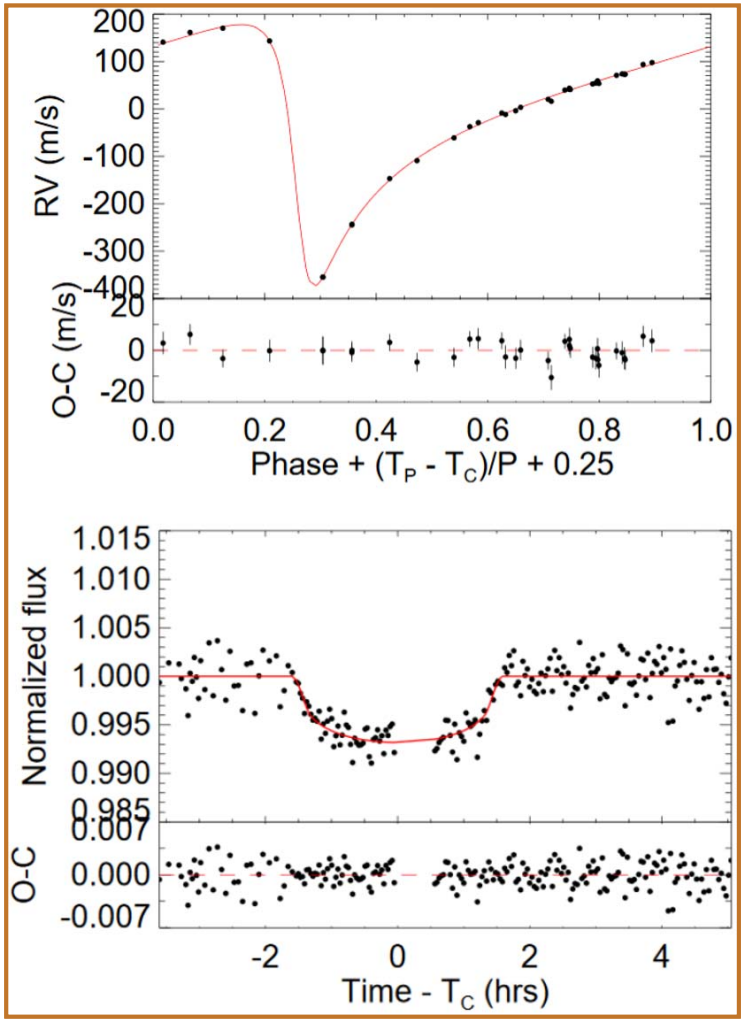
NASA Exoplanet Archive

The Archive supports future Exoplanet missions: TESS, JWST

Capabilities

- Interactive table search; parameter plotting
- Predicted Observables for Exoplanets
- Transit and Ephemeris Service
- Periodograms
- EXOFAST: Transit and Radial-Velocity Fitting
- Auto-generated plots and movies
- API to data

EXOFAST: simultaneous on-demand fitting of radial velocity and transit data for Exoplanet systems. The Archive is in the process of making use of the Commercial cloud for the EXOFAST and Periodogram services.



Caltech

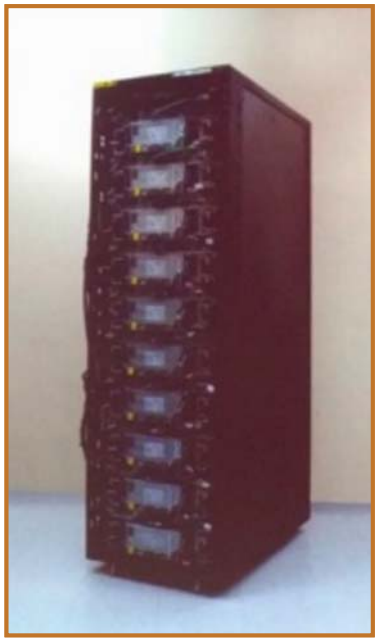
Big Data at IPAC Relative to Elsewhere

Typical Big Data Scales

- Google: >20 Exabytes
- Amazon: >1 million servers
- FINRA: scans PB of financial market data in real time to look for exchange fraud.
- LHC: 600M events/sec (at 1MB/event)
- SKA: data rates of many PB per second

IPAC Datacenter with 12 PB of spinning disk

=



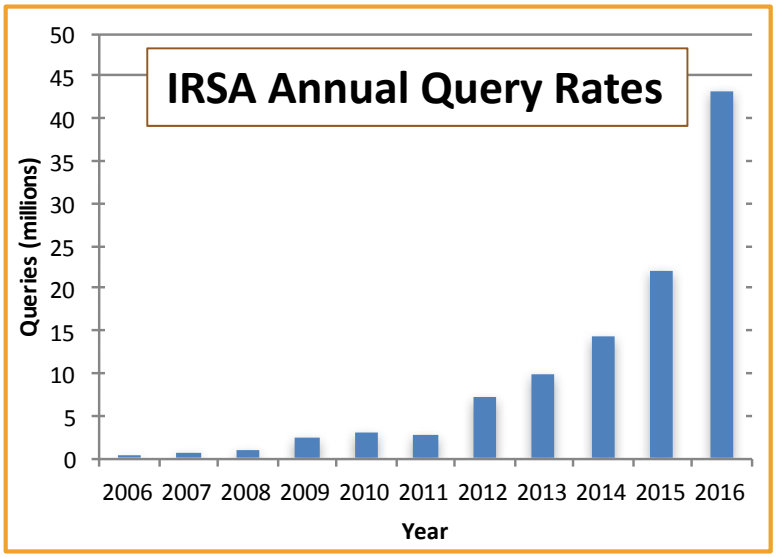
Amazon Rack with 12 PB spinning disk

IPAC Data Center

- 3 rooms, 3500 sq ft
- 76 x 42U racks
- 2500 cores, increasing to 7000 in next few years
- 12 PB disk → 30 PB in next few years
- Robotic tape library with 17 PB capacity
- All on UPS
- Network: 10 Gbps internal; 10 Gbps to commercial internet; 40 Gbps to internet2; planning Core upgrade to 100 Gbps

Scale of Operations at IPAC

Dataset volumes have increased by a factor of 100 in the past decade.



IRSA Data Volume (right) has grown by a factor of ten every few years recently, and is projected to grow even faster.

IRSA queries (left) have exploded with the advent of program interfaces and Virtual Observatory protocols.

Data System Metrics (Ops)	2MASS 1999-2001	WISE 2010-2011	Keck 2004–present	PTF 2009–2017	ZTF 2018–2020+	NEOCAM 5 yr mission	Euclid	WFIRST 5 yr mission
Data Rate	40 GB/night	76 GB/day	~1 GB/night	90 GB/night	1.2 TB/night	154 GB/day	200 GB/day	~25 Tb/day
Data Volume	24.5 TB	32 TB	45 TB	350 TB	3 PB	6 PB	7 PB	3 PB
Complexity	1.5B sources	44B sources	10 instruments, archive only	100B sources, ML*	750B sources; alert system, ML*	0.5 Trillion row DB, ML*	10B sources, 1 of 9 parallel nodes	ML*

**ML = Machine Learning algorithms used in pipeline*

The Context of Operational Systems

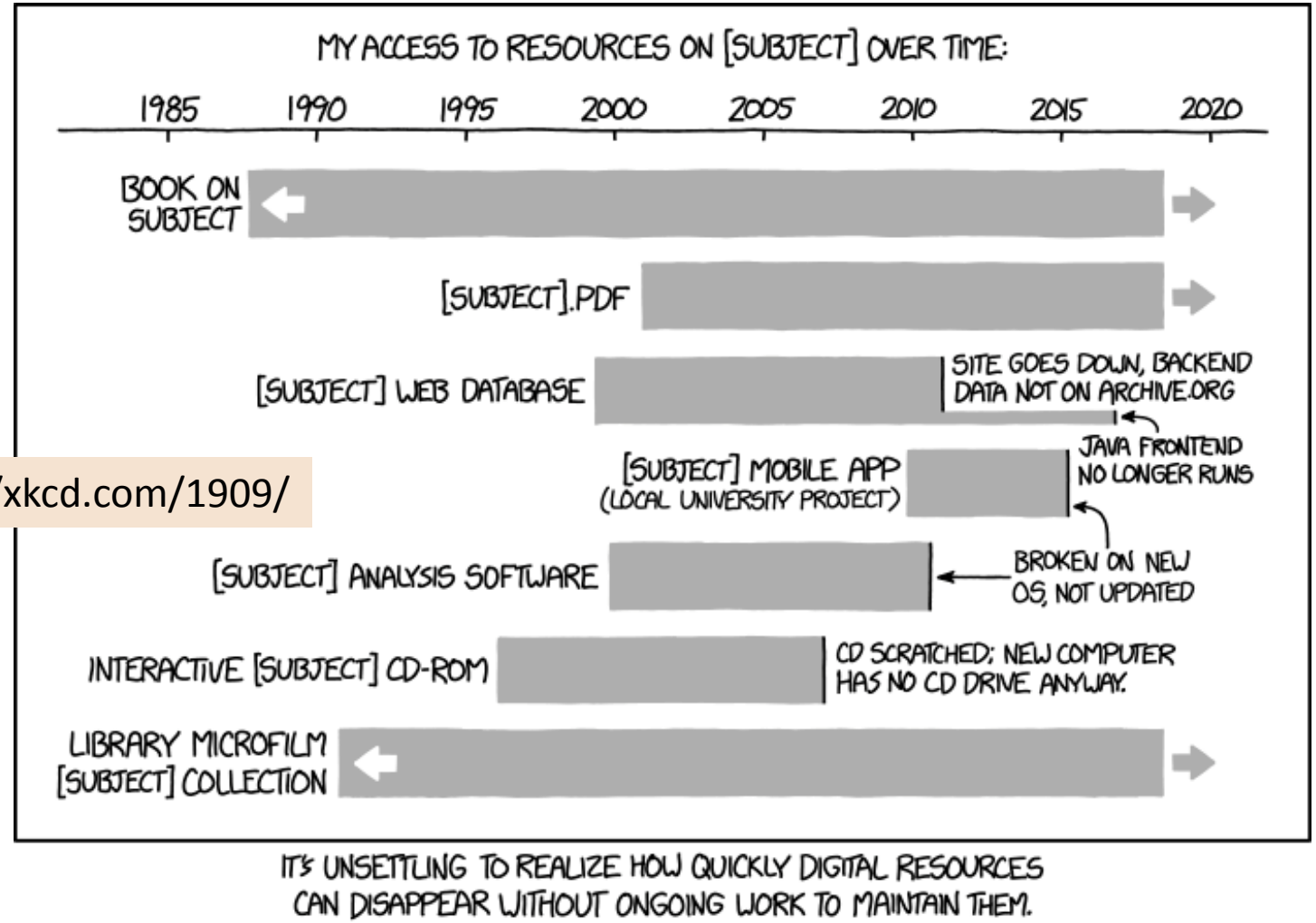
- NASA Archives, like NASA missions, are necessarily conservative, **focusing on reliable operations**:
  - Data integrity, backups, guaranteed uptime, IT security.
  - Support for mission operations; stability of interfaces.
  - Ingestion of datasets is usually a higher priority than new features.
- Data volumes and database table sizes are increasing exponentially, creating challenges for Ingest, query, and download.
  - Successful implementation of program interfaces, e.g. virtual observatory protocols, has led to incorporation of NASA Archives into data processing pipelines, with skyrocketing access rates.

IPAC's highest priority for its resources is supporting NASA missions and the science community. Big Data Technology Innovation must happen in that context.

Summary of FY17 for IRSA and NED:

- IRSA: 30 million queries,
- IRSA: 200 TiB of data downloaded
- IRSA: 20 new datasets
- IRSA: All major data sets available through VO protocols
- IRSA: Time Series Tool
- IRSA: User interface for NASA IRTF
- NED: 80 million database queries
- NED: 83 million new objects and cross-ids
- NED: Redesign of User Interface and supporting infrastructure

https://xkcd.com/1909/



10 Lessons We're Learning at IPAC (with some case studies)

Lesson: Large Holdings

Strategic organization can be more important than new technology.

- Simple solutions with careful planning can get you a long way without much technical magic: ***organization* is the magic!**
- Pay careful attention to data layouts.
- Reduce s/w overheads: small latencies that didn't use to matter now stand out.
- Maximize channel performance I/O, networking.
- Put critical high-demand data on faster storage.
- Increasing storage density worsens bottlenecks.
- 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.

irsabst16-L0/irsa-wise-data00	3.7T	3.4T	228G	94%	/export/irsa-wise-data00
irsabst16-L1/irsa-wise-data01	7.5T	6.9T	573G	93%	/export/irsa-wise-data01
irsabst16-L2/irsa-wise-data02	6.2T	5.9T	352G	95%	/export/irsa-wise-data02
irsabst16-L3/irsa-wise-data03	5.0T	4.8T	264G	95%	/export/irsa-wise-data03
irsabst15-L0/irsa-wise-data04	3.2T	3.0T	201G	94%	/export/irsa-wise-data04
irsabst15-L1/irsa-wise-data05	3.8T	3.6T	241G	94%	/export/irsa-wise-data05
irsabst22-L0/irsa-wise-data06	5.5T	5.2T	286G	95%	/export/irsa-wise-data06
irsabst16-L3/irsa-wise-data07	4.5T	4.2T	235G	95%	/export/irsa-wise-data07
irsabst15-L0/irsa-wise-data08	4.2T	4.0T	257G	95%	/export/irsa-wise-data08
irsabst15-L1/irsa-wise-data09	3.7T	3.5T	205G	95%	/export/irsa-wise-data09
irsabst22-L0/irsa-wise-data10	4.8T	4.6T	294G	95%	/export/irsa-wise-data10
irsabst16-L1/irsa-wise-data11	3.7T	3.5T	212G	95%	/export/irsa-wise-data11
irsabst16-L0/irsa-wise-data12	3.8T	3.5T	232G	94%	/export/irsa-wise-data12
irsabst16-L1/irsa-wise-data13	5.0T	4.6T	400G	93%	/export/irsa-wise-data13
irsabst15-L3/irsa-wise-data14	9.0T	8.5T	444G	96%	/export/irsa-wise-data14
irsabst15-L0/irsa-wise-data15	2.5T	2.4T	144G	95%	/export/irsa-wise-data15
irsabst15-L0/irsa-wise-data16	6.7T	6.3T	345G	95%	/export/irsa-wise-data16
irsabst15-L1/irsa-wise-data17	5.2T	4.9T	285G	95%	/export/irsa-wise-data17
irsabst15-L2/irsa-wise-data18	5.9T	5.6T	337G	95%	/export/irsa-wise-data18
irsabst22-L3/irsa-wise-data19	4.9T	4.7T	218G	96%	/export/irsa-wise-data19
irsabst15-L2/irsa-wise-data20	4.8T	4.5T	259G	95%	/export/irsa-wise-data20
irsabst16-L3/irsa-wise-data22	3.7T	3.4T	211G	95%	/export/irsa-wise-data22
irsabst16-L0/irsa-wise-data23	7.0T	6.6T	428G	95%	/export/irsa-wise-data23
irsabst15-L2/irsa-wise-data24	5.8T	5.5T	339G	95%	/export/irsa-wise-data24
irsabst22-L3/irsa-wise-data25	3.6T	3.4T	218G	95%	/export/irsa-wise-data25
irsabst22-L1/irsa-wise-data26	2.2T	2.1T	150G	94%	/export/irsa-wise-data26
irsabst22-L2/irsa-wise-data28	4.4T	4.2T	254G	95%	/export/irsa-wise-data28
irsabst22-L3/irsa-wise-data29	3.6T	3.4T	213G	95%	/export/irsa-wise-data29
irsabst22-L0/irsa-wise-data30	6.0T	5.7T	345G	95%	/export/irsa-wise-data30
irsabst22-L1/irsa-wise-data31	3.6T	3.4T	192G	95%	/export/irsa-wise-data31
irsabst22-L2/irsa-wise-data32	4.5T	4.2T	254G	95%	/export/irsa-wise-data32
irsabst27-L0/irsa-wise-data33	4.5T	4.3T	235G	95%	/export/irsa-wise-data33
irsabst27-L1/irsa-wise-data34	6.1T	5.8T	322G	95%	/export/irsa-wise-data34
irsabst27-L2/irsa-wise-data35	5.8T	5.4T	312G	95%	/export/irsa-wise-data35
irsabst27-L3/irsa-wise-data36	4.7T	4.4T	286G	95%	/export/irsa-wise-data36
irsabst27-L0/irsa-wise-data37	3.2T	3.0T	213G	94%	/export/irsa-wise-data37
irsabst27-L1/irsa-wise-data38	3.8T	3.2T	588G	85%	/export/irsa-wise-data38
irsabst22-L1/irsa-wise-dbms00	6.7T	6.4T	344G	95%	/export/irsa-wise-dbms00
irsabst16-L3/irsa-wise-dbms01	3.5T	3.1T	400G	89%	/export/irsa-wise-dbms01
irsabst22-L3/irsa-wise-dbms02	4.6T	4.3T	218G	96%	/export/irsa-wise-dbms02
irsabst22-L2/irsa-wise-dbms03	3.6T	3.4T	215G	95%	/export/irsa-wise-dbms03
irsabst15-L1/irsa-wise-dbms04	3.5T	3.3T	212G	95%	/export/irsa-wise-dbms04
irsabst27-L3/irsa-wise-dbms05	3.2T	3.1T	181G	95%	/export/irsa-wise-dbms05
irsabst22-L1/irsa-wise-dbms06	3.8T	3.6T	221G	95%	/export/irsa-wise-dbms06
irsabst15-L3/irsa-wise-dbms07	7.5T	7.1T	443G	95%	/export/irsa-wise-dbms07
irsabst27-L0/irsa-wise-dbms08	8.7T	2.5T	6.2T	30%	/export/irsa-wise-dbms08
irsabst27-L1/irsa-wise-dbms09	3.3T	3.1T	224G	94%	/export/irsa-wise-dbms09
irsabst27-L2/irsa-wise-dbms10	3.7T	3.2T	462G	88%	/export/irsa-wise-dbms10
irsabst1-L0/irsa-wise-dbms11	4.0T	3.7T	324G	93%	/export/irsa-wise-dbms11
irsabst1-L1/irsa-wise-dbms12	4.0T	3.8T	237G	95%	/export/irsa-wise-dbms12
irsabst1-L2/irsa-wise-dbms13	1.9T	1.8T	119G	94%	/export/irsa-wise-dbms13
irsabst1-L3/irsa-wise-dbms14	4.0T	3.9T	136G	97%	/export/irsa-wise-dbms14
irsabst1-L2/irsa-wise-dbms15	4.0T	3.6T	372G	91%	/export/irsa-wise-dbms15
irsabst27-L1/irsa-wise-dbms16	4.0T	3.2T	809G	81%	/export/irsa-wise-dbms16
irsabst1-L1/irsa-wise-dbms17	4.0T	412G	3.6T	11%	/export/irsa-wise-dbms17
irsabst27-L2/irsa-wise-dbms18	16T	3.2T	4.1T	45%	/export/irsa-wise-dbms18
irsabst42-L2/irsa-wise-dbms19	4.0T	3.6T	391G	91%	/export/irsa-wise-dbms19

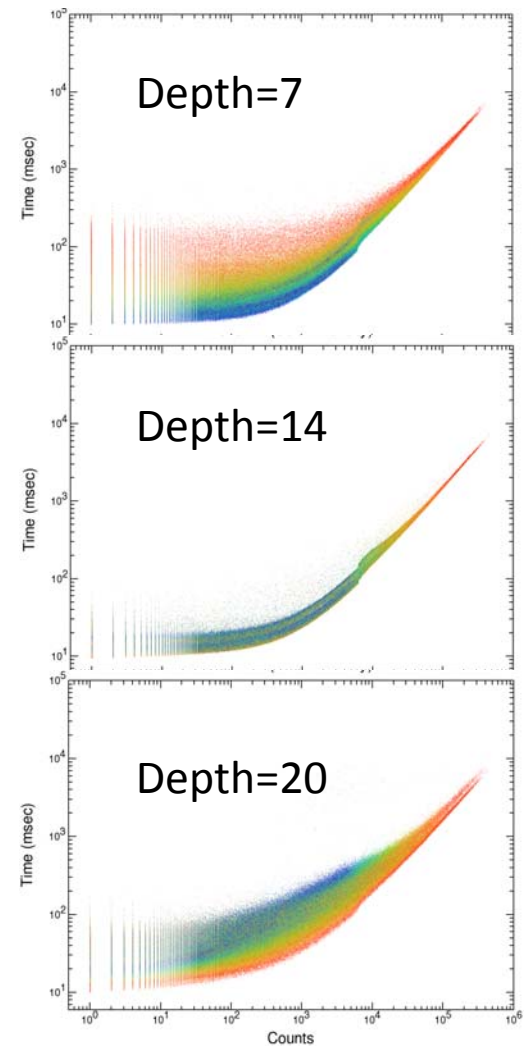
Partial listing of IRSA's WISE data disks, mostly at ~95%.

Lesson: Large Database Tables

Table organization gets even more important as table size grows.

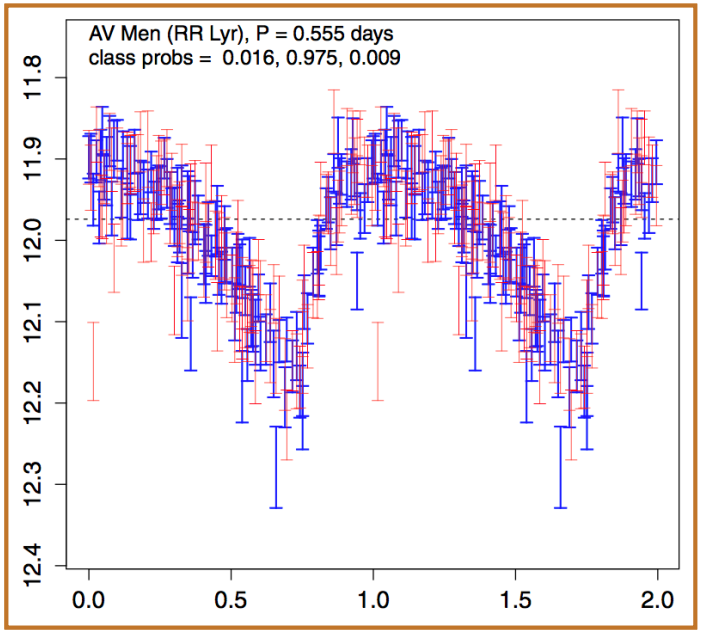
- Optimize data layouts for most common use cases: **But** different use cases require different organizations.
- Increasing interest in summary/statistical queries vs. the typical past use case of simple retrieval:
  - This usage requires more expensive table scan operations.
- May need to consider indexing in space-time, rather than just space, for moving object applications.
- Some tables are outgrowing our ability to handle them with previous techniques: for example, ZTF light curves.

Study of spatial indexing by Good and Berriman shows that which tesselation scheme is used is not as important as optimizing the depth of the scheme. For queries returning <10000 rows, too many cells or two few can increase the query time. Here, color represents the density of objects within a tesselation cell.



Case Study: ZTF Light Curves

- Prior datasets have included a photometry table which includes a record for every measurement of every source.
- ZTF data rate would generate a table with more than a trillion entries.
  - Database servers to handle this scale of data with traditional approach exceeds project budget.
  - Especially difficult with nightly table updates / re-indexing.
- Hybrid approach:
  - Database for objects, photometry in files.
  - Metadata for objects and pre-calculated photometric statistics.
  - Photometry extracted from files on-the-fly in response to queries.



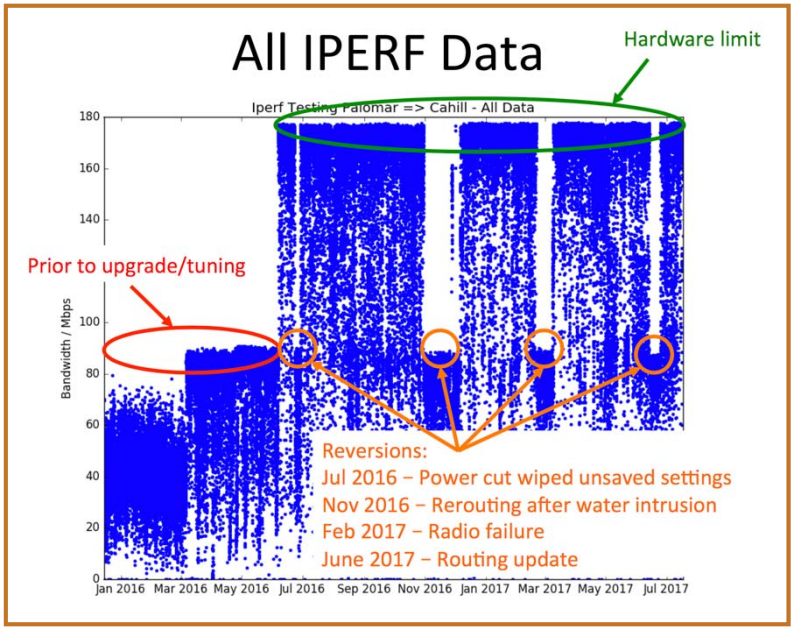
Light curve from WISE photometry database

Cost effective solution, rapid response to queries, but limits queries available.

Lesson: Constant Data Ingest

Invest early in scalable design; follow details end-to-end.

- Reliable and efficient operations requires rigorously verified metadata with tested process for interface changes.
- Tables can grow beyond what is feasible to re-index on a nightly basis: in some cases, we have implemented multi-stage table ingest. But this requires much more operations book-keeping.
- Because an interruption to operations is not acceptable, doing a redesign mid-mission can be very difficult.
  - WISE / NEOWISE / NEOWISE(R) mission has been in operations for 8 years; Spitzer has been in operations for 14 years.

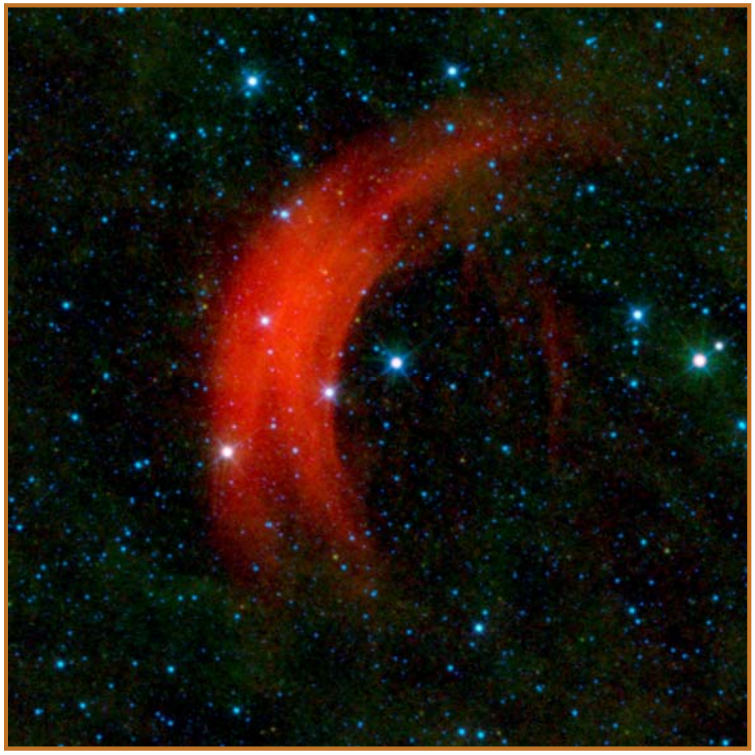


Parameter tuning by application can make a big difference in performance. For ZTF data transfer from Mt. Palomar, optimizing hardware and routing achieves a factor of 2 in transfer rate.

Lesson: Data Complexity and Variability

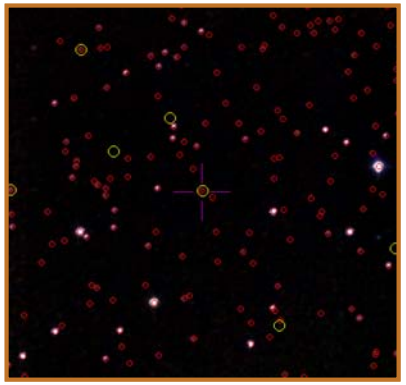
- Metadata:
  - In order to perform searches across an archive, need to have consistent metadata.
  - Need interface and metadata standards to search across archives:
    - We have adopted VO protocols.
    - Working with other archives to adopt the Common Archive Observation Model.
- Objects:
  - Co-registration of objects from one observation to next; moving objects!
  - Cross-identification of objects between datasets: resolution, wavelength.
  - Extended in space.
  - Hierarchical objects (galaxy vs. its components, multiples with planets, planets with moons).
  - Evolving knowledge of the relationships between objects as systems are observed.

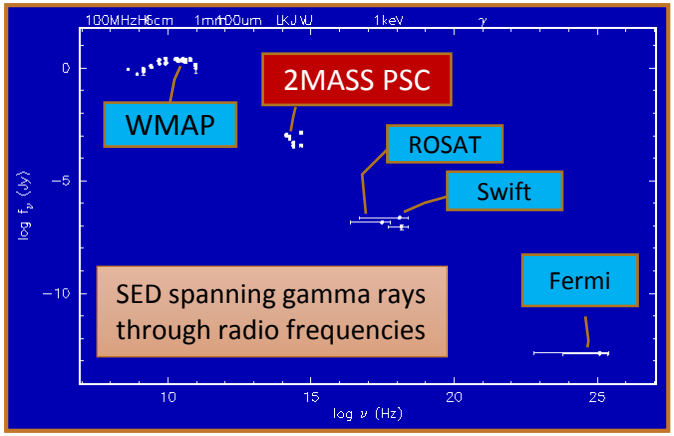
Example of a diffuse, extended structure. This shock wave was observed by WISE, and is caused by the rapid motion of the star in the middle of the image through a nebula.



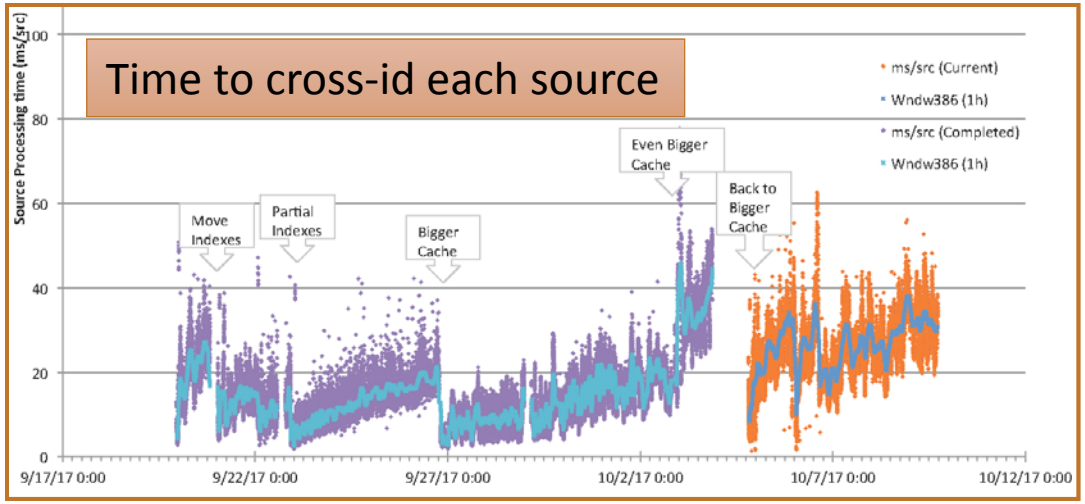
Case Study: Cross-Identification of Large Catalogs

- NED correlate newly ingested catalogs with existing database of ~250 million objects.
- Currently ingesting 2MASS catalog with 470 million sources.
- Next catalog, AllWISE, has 748 million sources.
- Using parallel processing to do cross-ids, but rate slows over time due to database I/O performance.
  - Currently a six-month ingestion process.
  - Database must be parallelized.





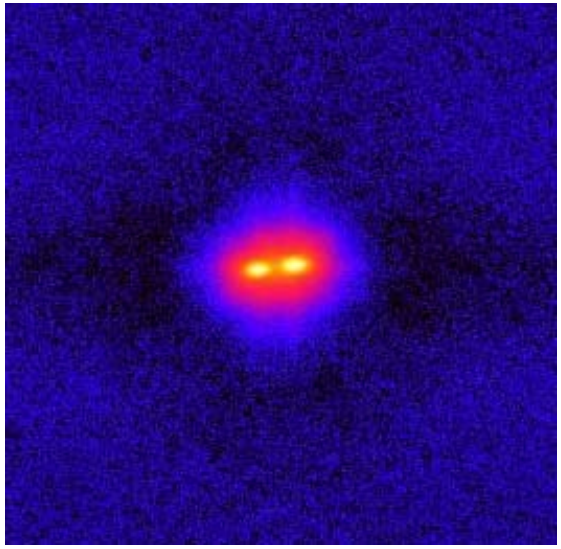
Reliable cross-IDs are required to construct spectral energy distributions (SEDs). *Upper Left*: Cross-matching the 2MASS Point Source Catalog (PSC, red) with prior objects in NED (yellow). Quasar PKS 1057-79 is at center. *Upper right*: Result of fusing 2MASS PSC photometry (JHKs bands) with prior data in NED from Fermi, Swift, ROSAT and WMAP.



Case Study: Complexity and Variability in Exoplanet Archive Objects

Exoplanet systems involve objects with changing M to N mappings.

- Exoplanet Archive catalogs and curates exoplanets orbiting other stars as published in the literature
- 3 out of every 2 stars are binaries – often unknown at the time of the planetary discovery.
- Often confirmed planets are found to orbit within multi-star systems after the system has been published.
- Archive needs to track
  - System has multiple stars
  - System has planets
  - Which star(s) hosts the planets – if known
  - Changes in derived system parameters

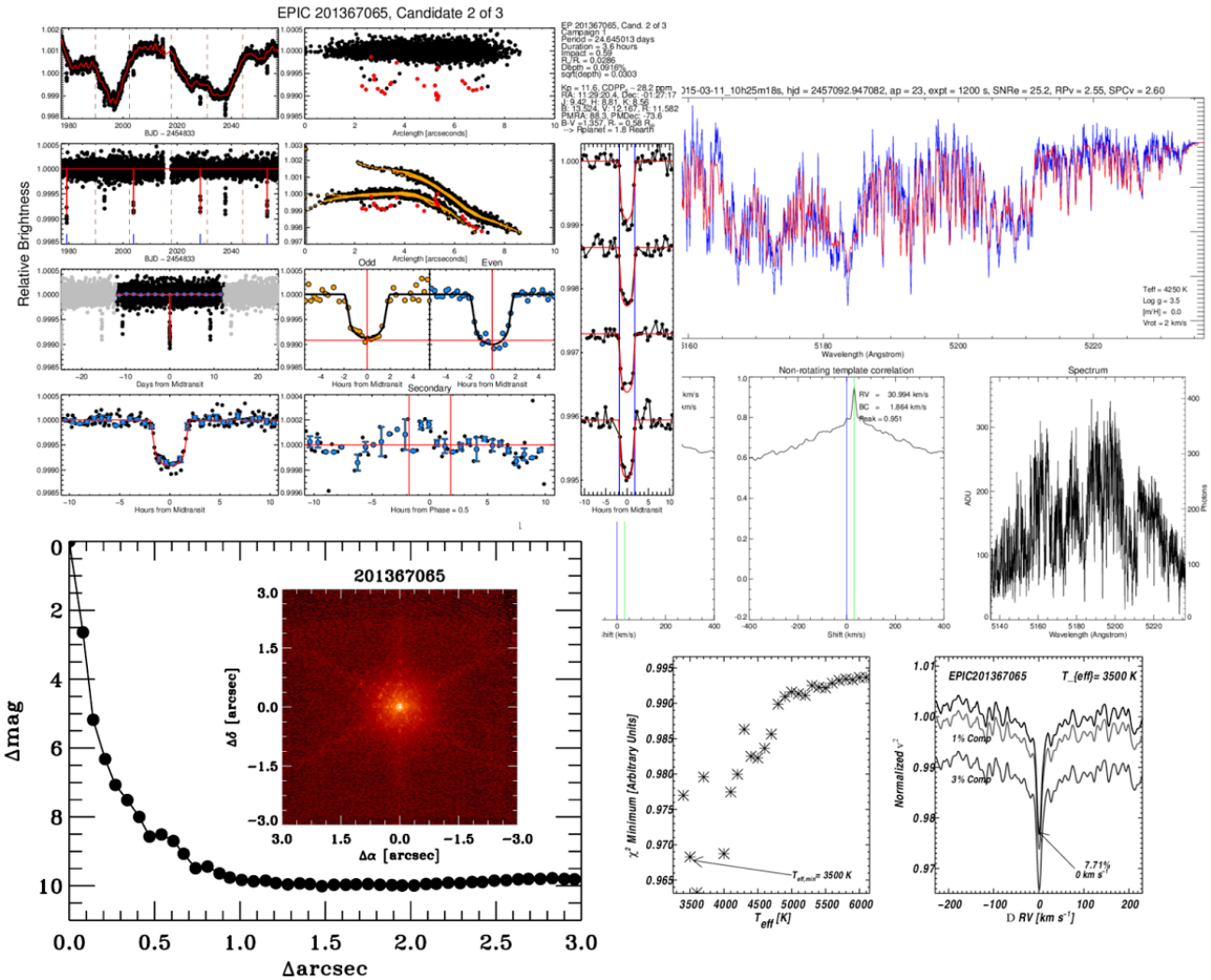


Kepler-132: a confirmed planetary with 4 planets and later found to be a binary star. While the planets are almost certainly real, it is unclear as to which star(s) the planets orbit and what their true planetary radii are.

Case Study: ExoFOP

IPAC Archives must support a large number of data types and formats.

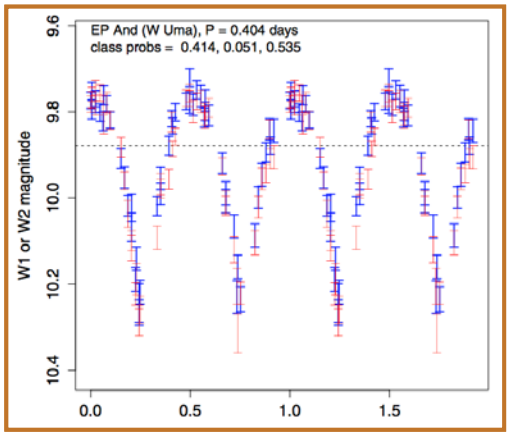
Subset of data files acquired in follow-up observing of a planet candidate system. ExoFOP must ingest and provide both website and API access to a wide variety of data content, including target lists, observing status, observational data, derived parameter data, and notes and comments.

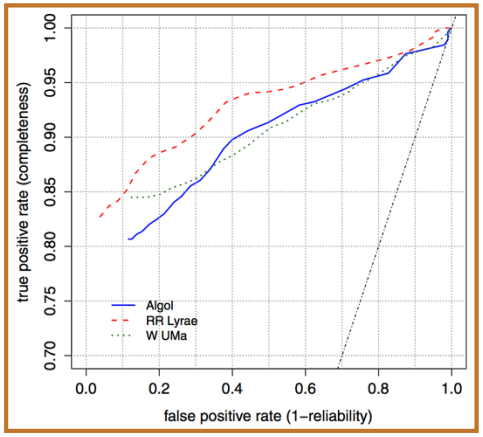


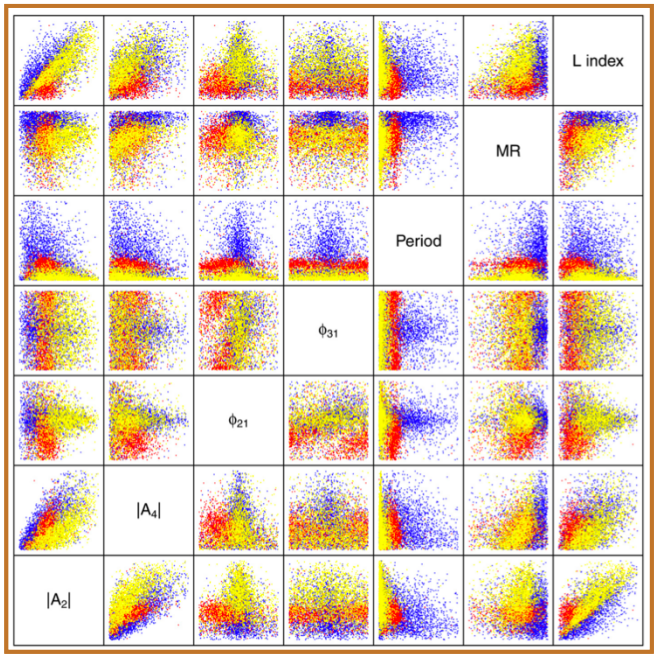
Lesson: Machine Learning

Machine Learning helps IPAC solve Big Data challenges

- **Transient Identification**: used on PTF, and will be used on ZTF, NEOCAM, and WFIRST.
- **Cross-Matching**: used as part of NED ingest of large catalogs. Only way to make associations between catalogs of 100's of millions of objects.
- **Literature Extraction**: NED has evaluated several packages, but with limited success for NED applications to date.
- **Research Applications**:
  - Self-Organizing Maps and t-distributed Stochastic Neighbor Embedding for galaxy colors: ***fast*** estimates of galaxy parameters (redshift, mass, age).
  - Classification of Periodic Variable Stars.



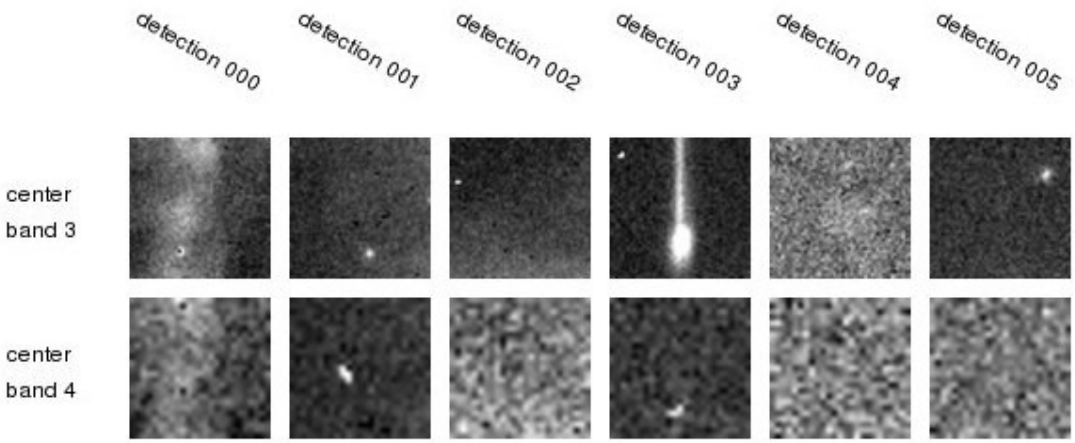




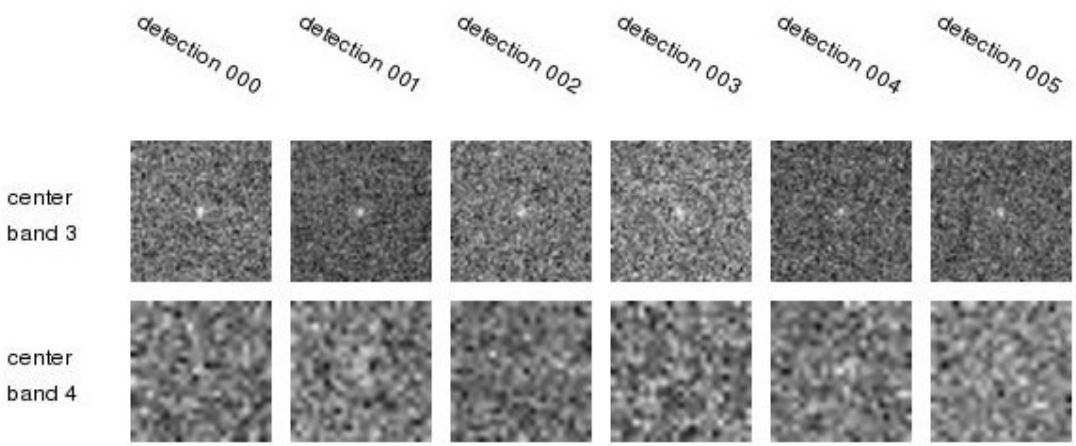
From **Masci et al., ApJ 2014**:
Using Machine Learning to classify periodic variable stars detected by WISE.
Upper Left: example light curves.
Left: Matrix of scatter plots for three variable types for all pairs of metrics.
Upper Right: Accuracy vs. Completeness of classification.

Case Study: ML for NEOCAM Tracklets (1/2)

Training Set Examples



Bad tracklet, comprised of spurious detections

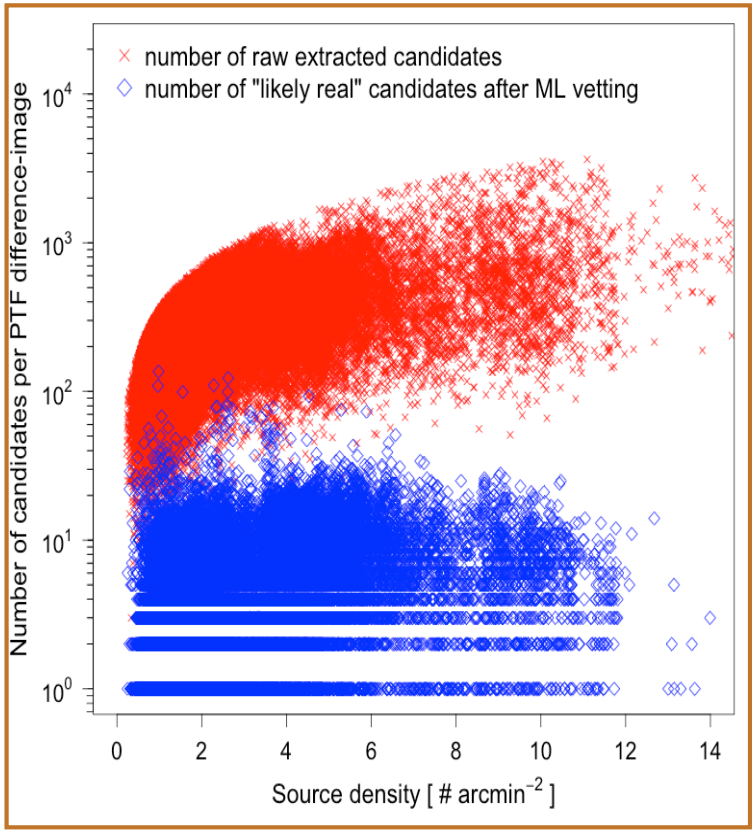


Good tracklet, comprised of reliable detections

Case Study: ML for NEOCAM Tracklets (2/2)

IPAC must use ML to process transient events from its telescopes.

- NEOCam adapts the ML automatic classification algorithms operated in the PTF transient detection pipeline to identify and filter out spurious detections of difference image residuals
  - PTF detected ~1M transient candidates each day
  - The ML classifier filtered out ~94% of these as spurious
  - The remaining candidate detections had a demonstrated reliability of 99% and completeness of ~97%
  - NEOCam will have fewer spurious detections, but 100K real detections daily.

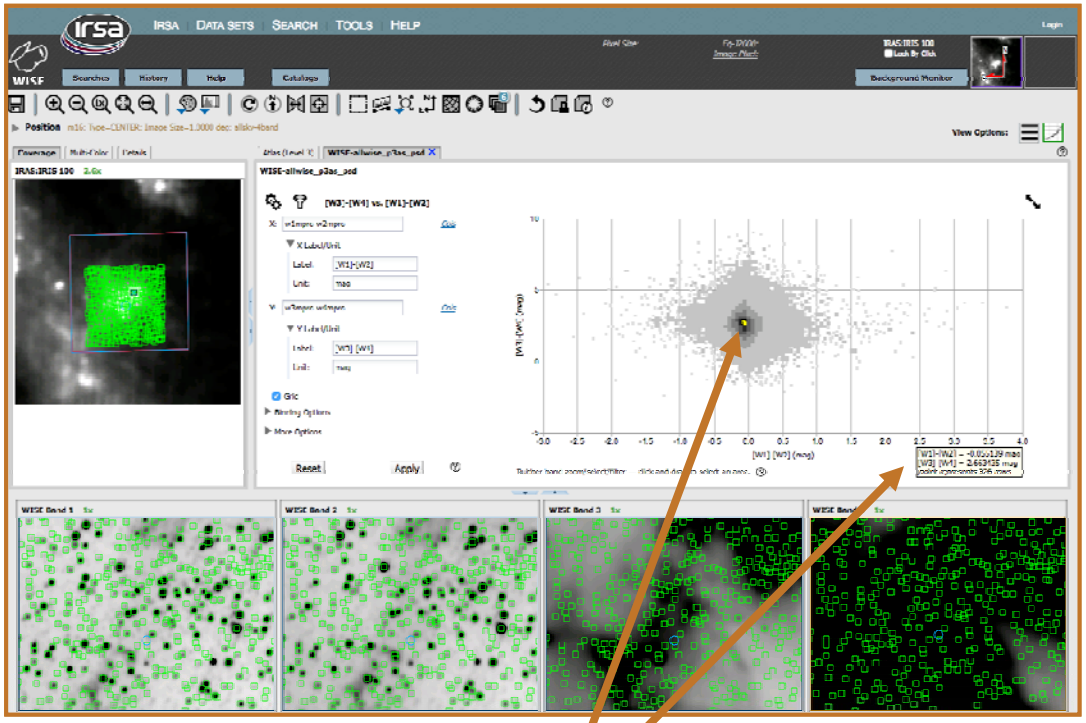


The dominant source of spurious detections for NEOCam is expected to be residuals in the difference images due to small registration errors and PSF mismatches between the visit and static sky images.

Lesson: Data Visualization

Interactive graphics provide intuition about the data.

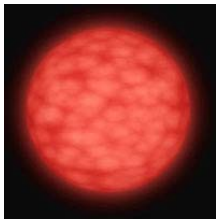
- Co-registration of data sets: IRSA (and soon, NED) allow simultaneous viewing of different data sets.
- Time-domain: light curves, folded-viewing, periodograms, moving objects.
- For massive sets we have to go from symbol representation to continuous quantities: density plots, histograms.
- Data Cubes
- The IPAC Visualization Group (iViz) is exploring data viewing approaches:
  - 3-dim / N-dim representations
  - Will VR be useful?



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.

Case: Parallax for Ultra-Cool Dwarf Stars

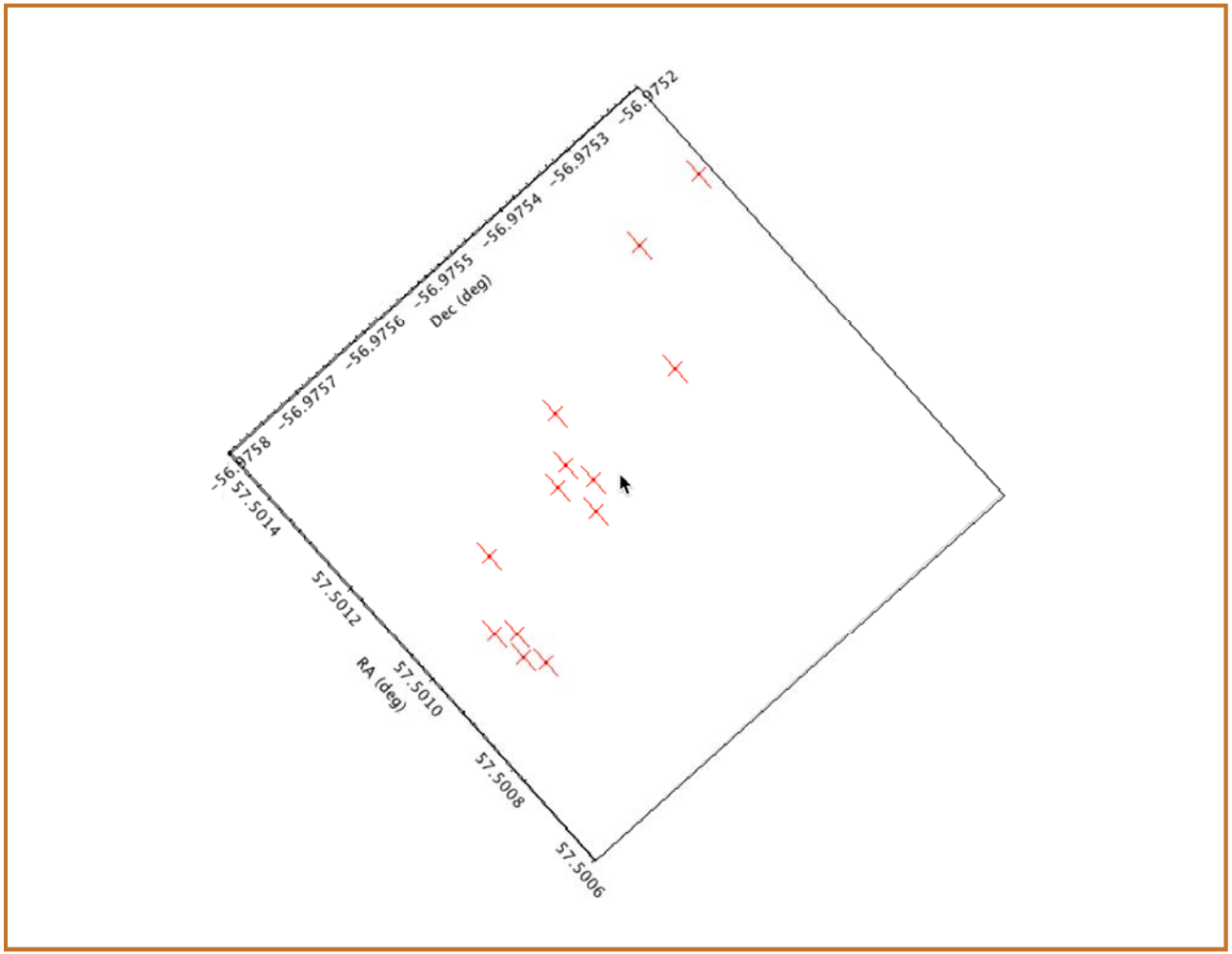
M



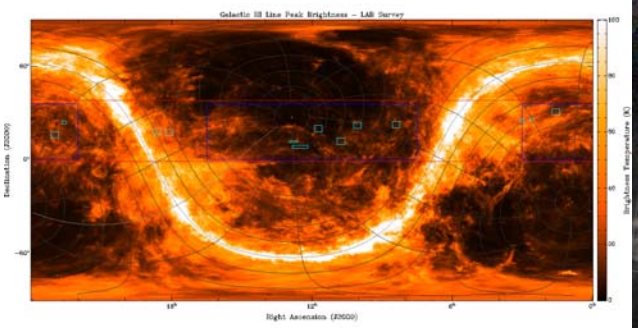
L

T

Y



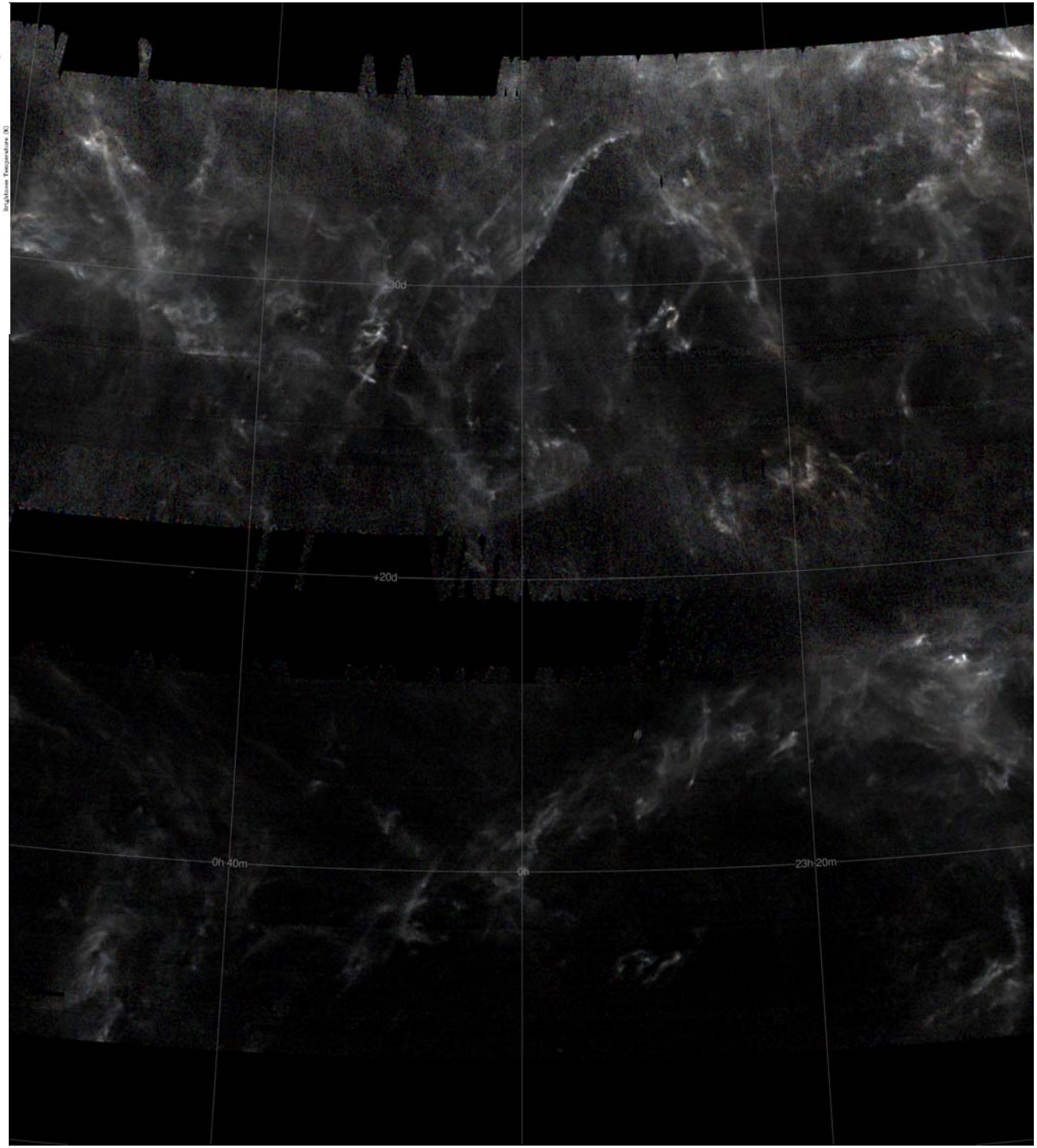
Davy Kirkpatrick, 2016, fitting parallax and proper motion using Spitzer on WISE Y0 dwarf: 6 pc



Data Cubes

Generated by IPAC's Montage Open-Source Toolkit running on AWS.

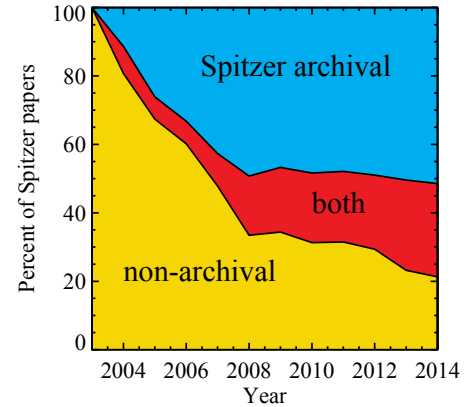
Full-resolution mosaic of the central 256 frequency planes of 30 GALFA-HI images, centered on 0h Right Ascension. The RGB color is derived by combining 3 adjacent frequency planes. All gaps, such as that around 20 degrees declination, are due to incomplete coverage in the input images.



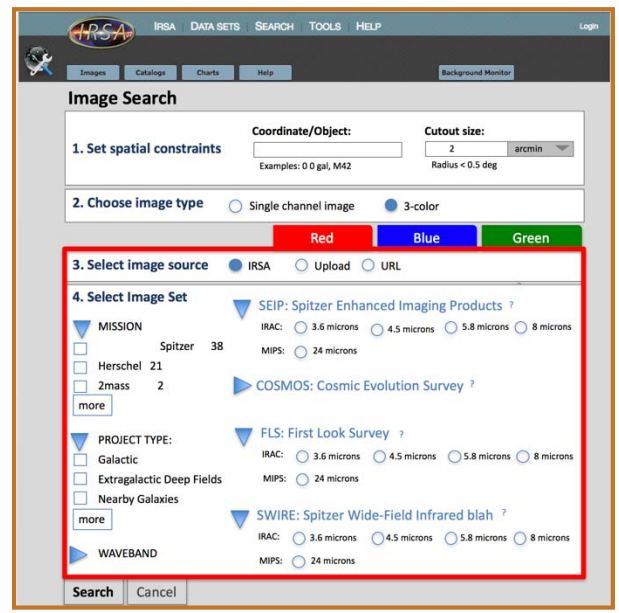
Lesson: Data Discovery

Ease of access is more important than a single point of access.

- The use of data in NASA archives can **double the science** of the original mission.
- NAVO is a collaboration of NASA archives to provide uniform access to data via VO protocols.
  - NAVO has evaluated having a single portal for access to all NASA astrophysics data.
  - A single portal is less effective and more expensive than archives dedicated to supporting a specific community with tools, formats, services, and expertise.
  - NAVO provides the machinery to do data discovery via its Registry.
- NED and Exoplanet archive are data discovery engines:
  - Attempt to provide all known information about a particular object and/or region in the sky.
  - All listings have links back to published research and to primary observational data (e.g. other NASA archives).
- IRSA provides comprehensive data discovery for its holdings, and plans to increase integration with VO discovery.

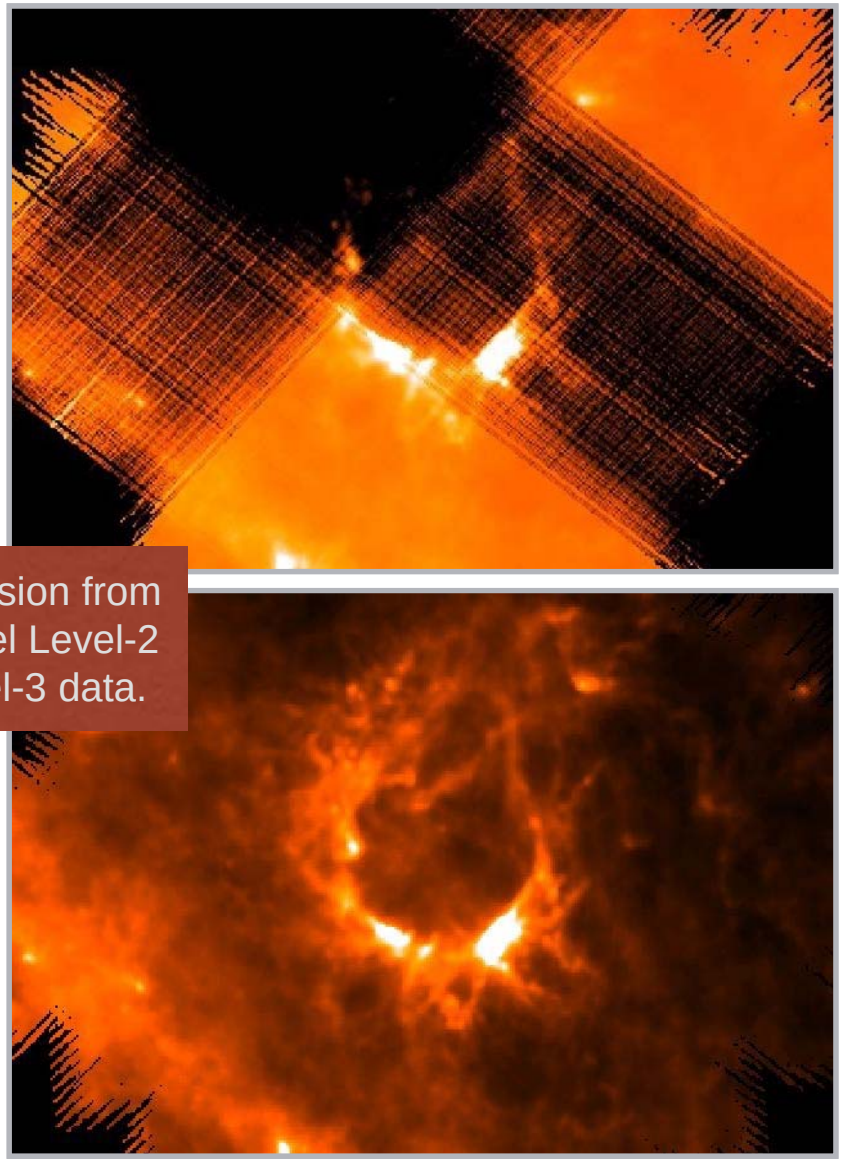


New IRSA data discovery service will provide an "Amazon-like" functionality to selecting and displaying images and catalogs from IRSA holdings as well as other NASA Archives.



Lesson: Virtualization

- Important step in making analysis available near the data: safe environment.
  - IPAC Herschel provided virtual machines for US scientists to run the very memory-intensive and complicated HIPE analysis software.
- Improves reliability for pipeline processing, system maintenance in context of 24/7 operations.
- Helpful for moving to cloud, or cloud-hybrid datacenter.
- Euclid is virtualizing the entire processing system: identical science data centers at 9 locations around the world, including IPAC.
- May use as data-delivery mechanism for Joint Data Processing activity.



Progression from Herschel Level-2 to Level-3 data.

Lesson: The Commercial Cloud

The Cloud does not necessarily reduce system administration costs.

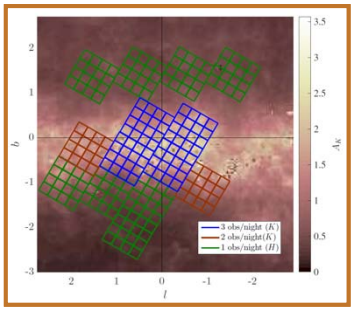
- The IPAC Datacenter is cost-effective for systems in long-term and mostly full-time use.
  - Backups to AWS Glacier would currently cost ~$2M/year, and even more for data on S3.
  - Processing pipelines are long-term 24/7 operations.
- Cloud-computing is cost-effective for ephemeral processing:
  - Preliminary development
  - Sandboxes & experiments
  - Debugging and automated integration testing
  - Surge computing needs: reprocessing, urgent and parallelizable tasks, one-off simulations, Sagan Workshop

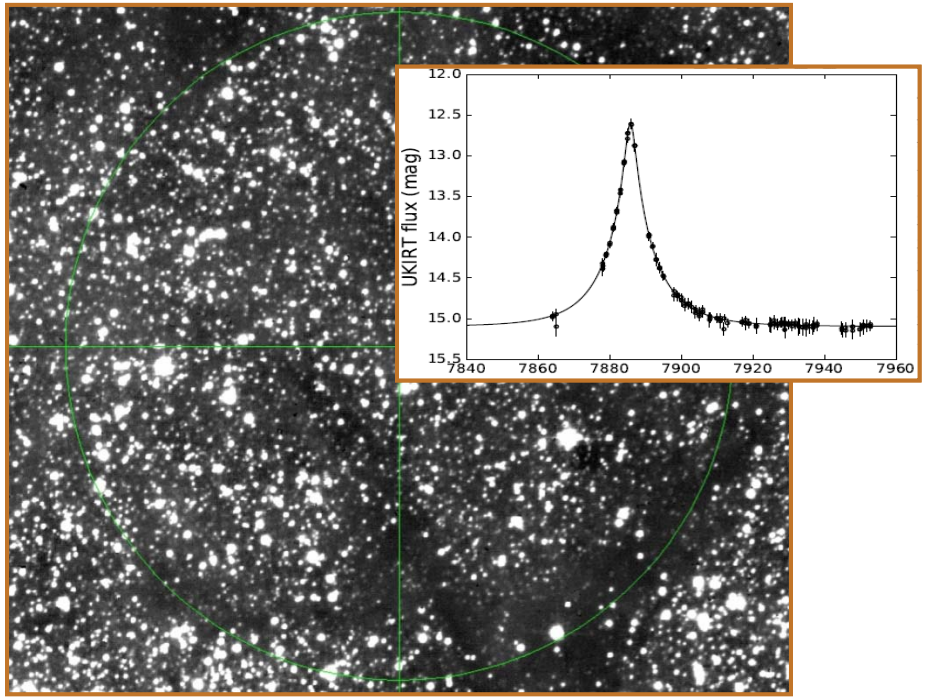
5°×5° 18k x 18k pixel section (1.2, 3.4 and 8.8 μm) of 16-wavelength Infrared Atlas of Galactic Plane using Montage on the Amazon Cloud.

Case Study: Surge Computing

The Cloud can provide an agile approach to compute-intensive tasks.

- Generate ~20M light curves from UKIRT survey imagery to estimate microlensing event rate for WFIRST.
- Research project with a deadline: would have taken weeks to run on original 4-core system. Easily parallelizable task: separable analysis of multiple datasets.
- Rather than order expensive hardware, configured AWS AMI on compute-optimized system to execute analysis.
- After verifying function on a single VM, duplicated twelve times: analysis ran over 1–2 days.
- 1 of the 12 crashed: apparently the allocated memory was insufficient. Took 5 min to create VM with double the memory, which succeeded. Very agile approach!
- Total bill was ~$300; 2/3 was for TB data transfers, 1/3 for CPU.
  - Longer term data storage would have been $50/month (about 2 TB).
  - Subsequent work done using JPL Supercomputing cluster—better price for keeping data longer, though less optimized compute systems.

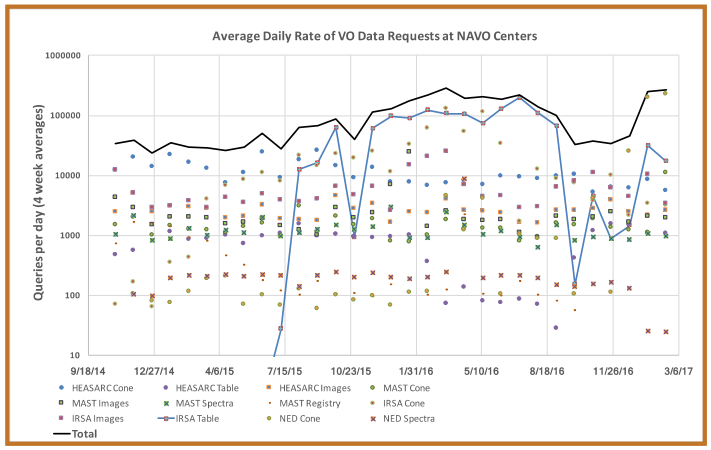




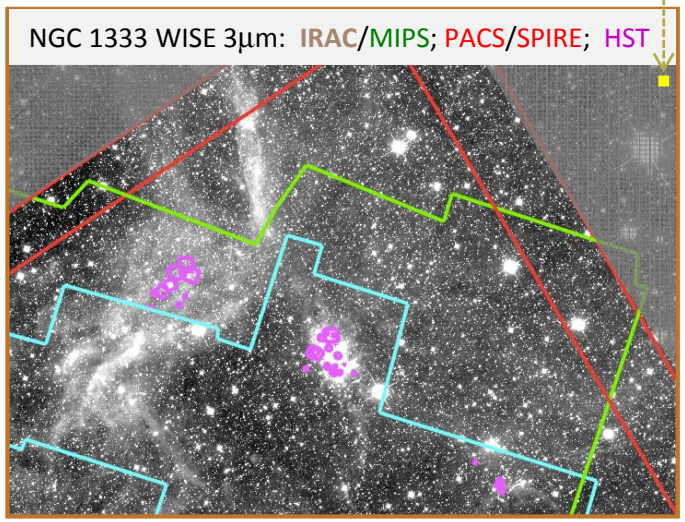
Lesson: Interoperability

Increasing interoperability can allow a divide-and-conquer approach to Big Data.

- Take advantage of work that has already been done:
  - NED now using ADS for bibliographic info
  - NED using Image Viewer from IRSA / LSST (Firefly): comes not only with image manipulation and coordinate info, but also source table & plotting
  - IRSA using NED data tables (enabled by VO protocols).
- Publishing archives to VO registry and providing protocols greatly expanded access through VO discovery tools.
- Collaboration between IRSA and MAST: overlay of Spitzer/Hubble/JWST footprints on data for observation planning.



JWST/NIRCAM 1 module

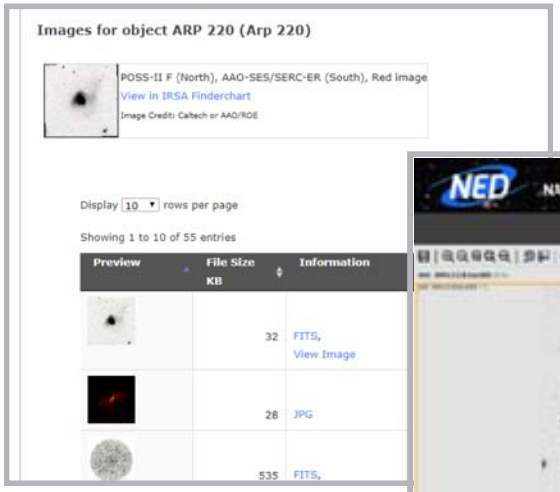


This figure shows overlays of areas surveyed by Spitzer (blue and green), Herschel (red),and HST (purple) in the region of NGC 133. The background image is the WISE 3 micron image of the region.

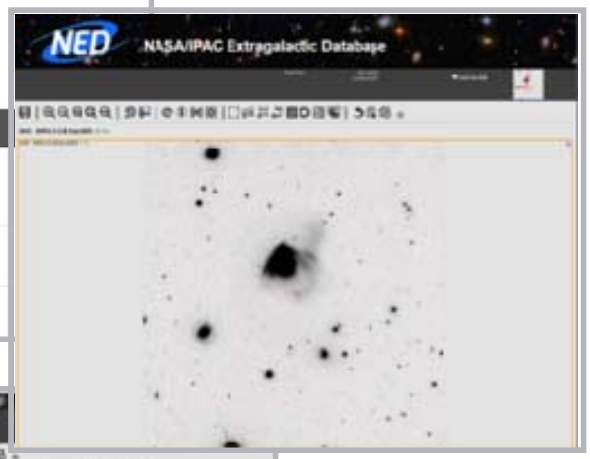
Case Study: NED Adopts IRSA Firefly Visualization

A broken service leads to better functionality and common look-and-feel across IPAC archives.

- NED used Aladin Java applet for FITS image viewing; not supported by modern browsers.
- For release next month: NED adopted open-source Firefly image viewer service, developed for IRSA, LSST, WISE, PTF, ZTF and others.
- Firefly capability brings not only image viewing, but will add catalog overlays, selection, and plotting.



Images available for this NED object



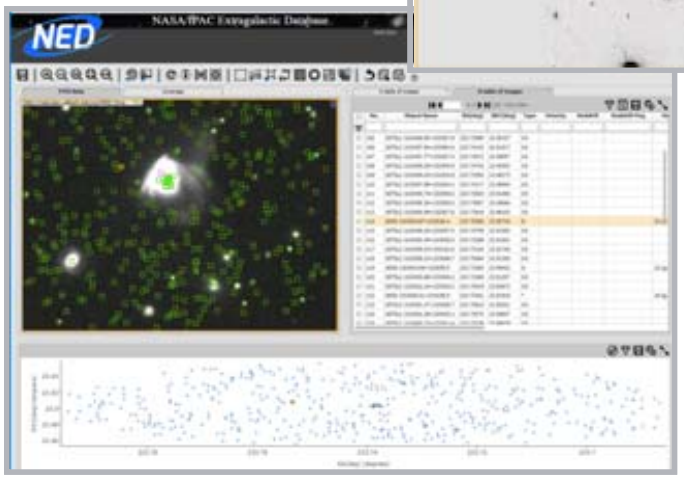


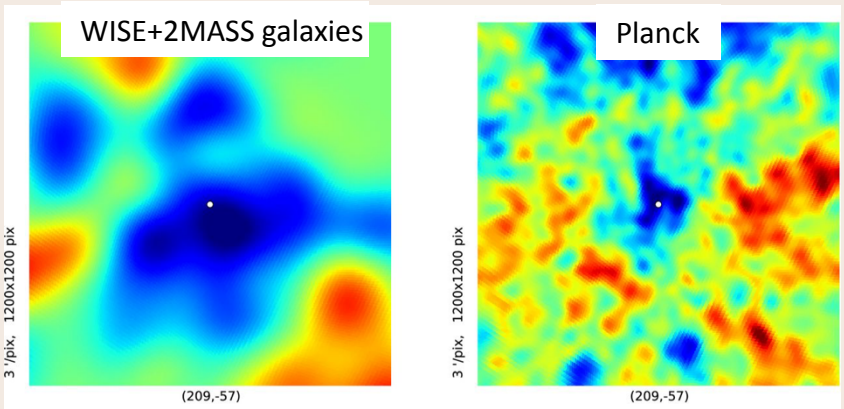
Image Viewer

Tri-View with all NED sources selected

Big Data at IPAC: Opportunities and Challenges

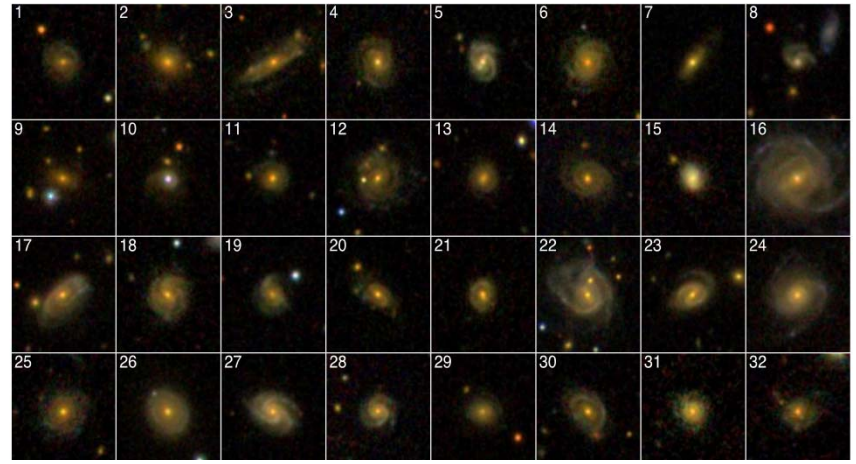
Opportunity: Archive as Observatory

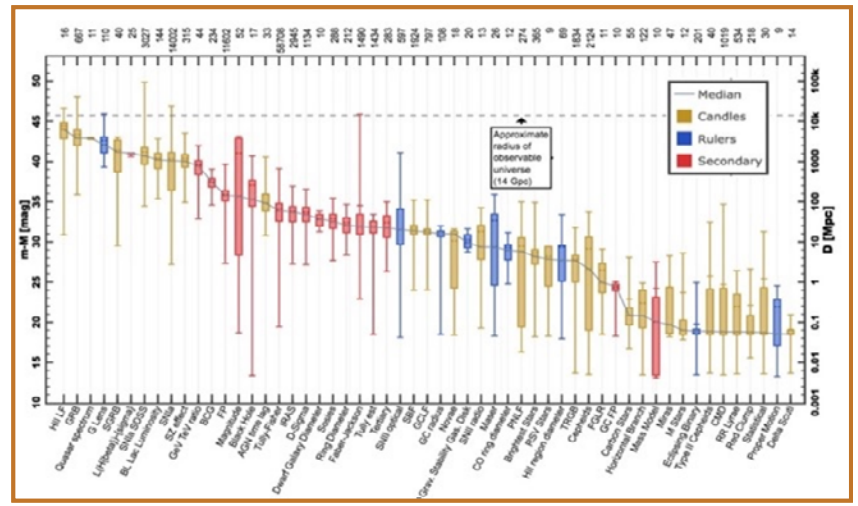
- The quantity and complexity of archival data now available has led to the use of Astronomy Archives as Virtual Observatories.
- NASA Archives are cooperating to create a synergistic "Virtual Observatory" across individual archives using VO protocols.



Combining data from multiple missions (WISE+2MASS+PanSTARRS) data may reveal super-void in CMB cold spot seen by Planck; (Szapudi et al. 2015, MNRAS, 450, 288)

Discovery of a new class of super-luminous spiral galaxies (Ogle et al. 2016) based entirely on data synthesized within NED demonstrates its power as a discovery engine.

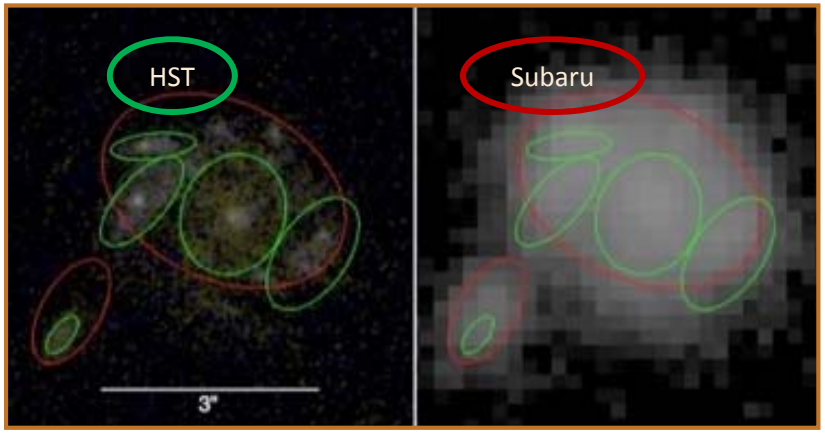




A visualization of the redshift-independent distance measurement techniques used in extragalactic research, as available in NED, from Steer et al. (2017, AJ, 153, 37). The data are sorted by median distance, showing the 25th and 75th percentiles (boxes) and full range of each distribution.

Opportunity: Joint Processing for Large Surveys

- The science opportunities from joint analysis of data from Euclid, LSST, and WFIRST go well beyond the science enabled by each survey alone.
  - Many of the goals of a joint analysis require pixel-level co-processing to address the complexity and subtlety of systematics confusion, and astrophysics.
  - The resources for joint analysis are beyond NSF/DOE budget for LSST and NASA budget for Euclid & WFIRST.
- IPAC is leading to scope an approach for this kind of processing:
  - Target specific science goals to scope requirements: tools and architecture will pay off well beyond those goals.
  - Products of joint processing may be "objects" rather than just data sets: data embedded in environments (VMs or Docker containers) with methods to access and analyze them—compatible with cloud-based distribution for on-demand scalability.



Left: HST view of a galaxy cluster. Right: Subaru Suprime Cam view of the same cluster. Green ellipses are HST extracted sources, red are the Subaru extracted sources. Combining data sets of different resolutions and wavelengths can lead to science results not available from analysis of the individual sets alone.

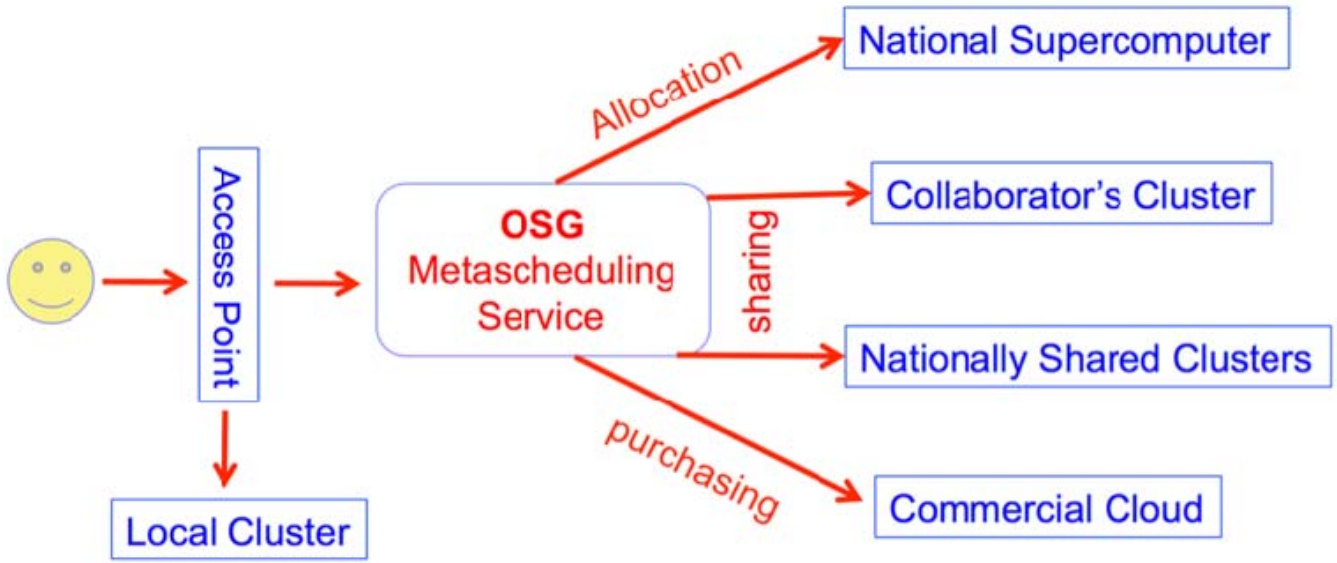
Joint Processing Enables:

Improved Photo-Z estimates
Improved weak-lensing shear field estimates
Better galaxy cluster mass estimates
Star-formation history for millions of galaxies
Cross-mission systematics checks
Suppress spurious objects
Separating blended sources

Opportunity: Pacific Research Platform

- IPAC is implementing high-performance connectivity (10-100 Gbps) via the NSF-funded Pacific Research Platform (PRP) node at Caltech
  - Deployed "perfsonar" performance testing endpoints on IPAC's PRP and regular networks
  - Deploying two data transfer nodes on PRP for prototyping and project experimentation
  - Developing a plan for properly integrated access to PRP from IPAC's core networks
- This will be important to fulfill IPAC's role as the Euclid U.S. Science Data Center in a widely distributed system
  - **Update:** Demonstrated the use of our PRP node to transfer 60 TB of reprocessed Planck data from NERSC. First attempt yielded factor of 3 speedup.
  - Current limitation in rate is now from disk I/O rather than network bandwidth.
- Evaluating the option of implementing a permanent Globus endpoint on our PRP node.

Opportunity: Open Science Grid



Open Science Grid: "distributed High-Throughput Computing"

- Sort of a "SETI-at-Home" for data centers: keeps CPUs busy.
- Advantage over commercial cloud: data transfer is "free".
- Access to computing related to computing resources provided

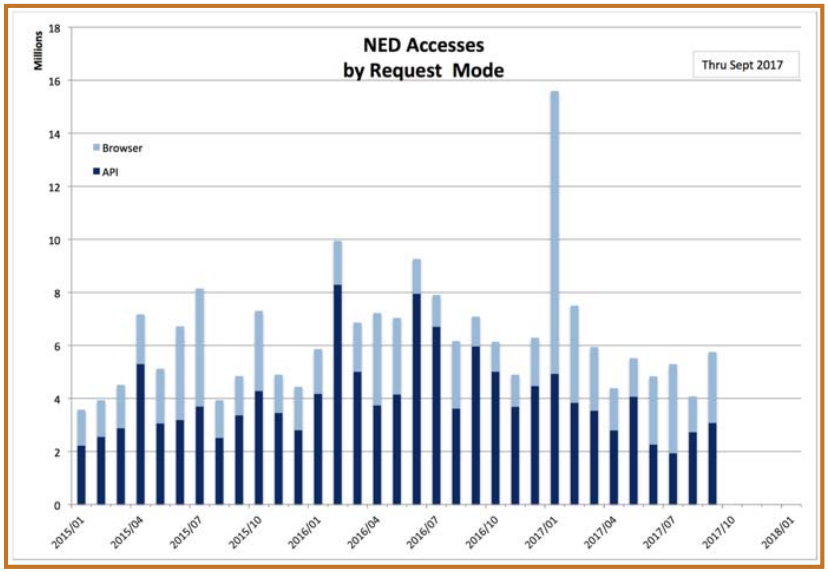
IPAC Evaluating Participation

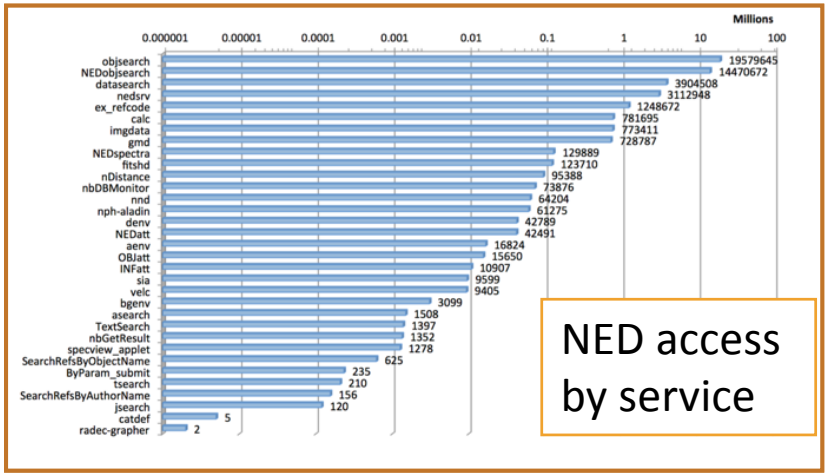
- Meetings with OSG participants at IPAC and at OSG Conference; exchange of presentations.
- Implementing general purpose shared VM compute cluster
- Need to plan security policies (i.e., firewall rules) to allow outside users to run tasks on internal IPAC compute cluster.

Challenge: Analytics

Data analytics can inform strategic investment.

- IPAC uses both google/apache analytics as well as query logging:
  - Metrics reported to funding agencies: number of hits, unique IPs.
  - Identification of "most popular archive queries", bad-actor IP addresses, popularity by archive API "service", geographic location of hits (easily spoofed, however), most popular IPAC website pages.
- We have yet to fully use data analytics related to IPAC data services:
  - Not yet a priority from funding agency and science community
  - Once those use cases are identified, services may need to be modified to record useful analytics data.

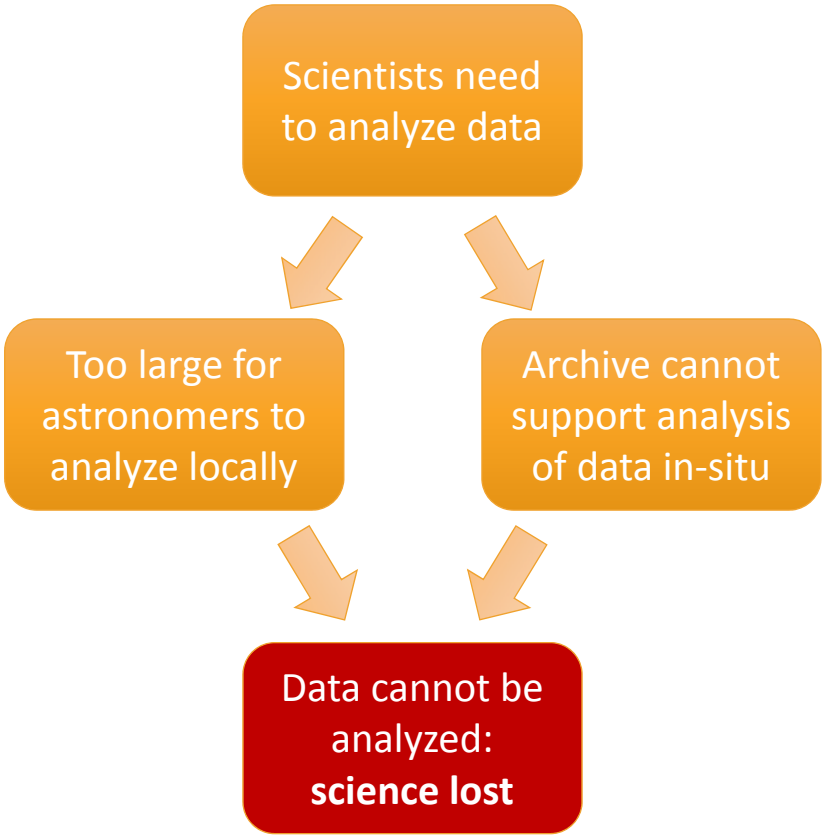




Reminder: The Archive Dilemma

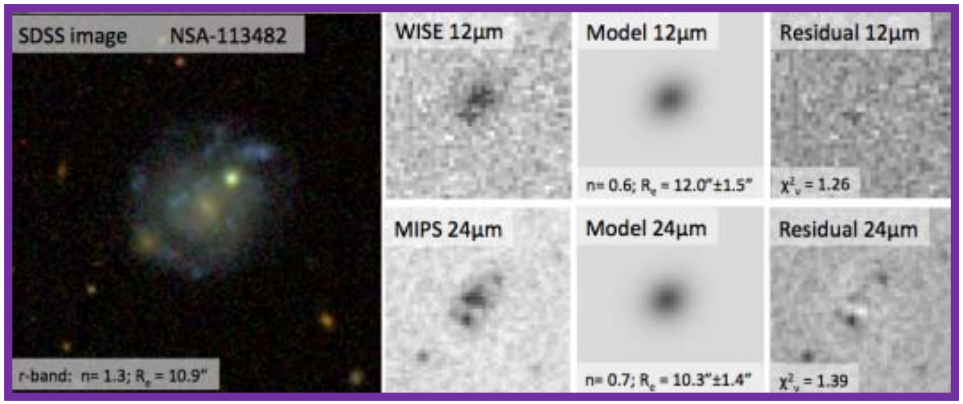
As part of IPAC's briefing to the Big Data Task Force Subcommittee of the NASA Advisory Council Science Committee in 2016, we highlighted a paradox confronting NASA Archives:

- NASA's Astrophysics Archives have focused on curation and online query/access services for science datasets. Resources available for custom processing and analysis are limited.
- Operating Missions are primarily focused on generation and distribution of standard products for many applications.
- Many users have indicated interest in deeper analysis and mining of those products.
- Most users do not have resources to download entire data sets for special-purpose analysis.

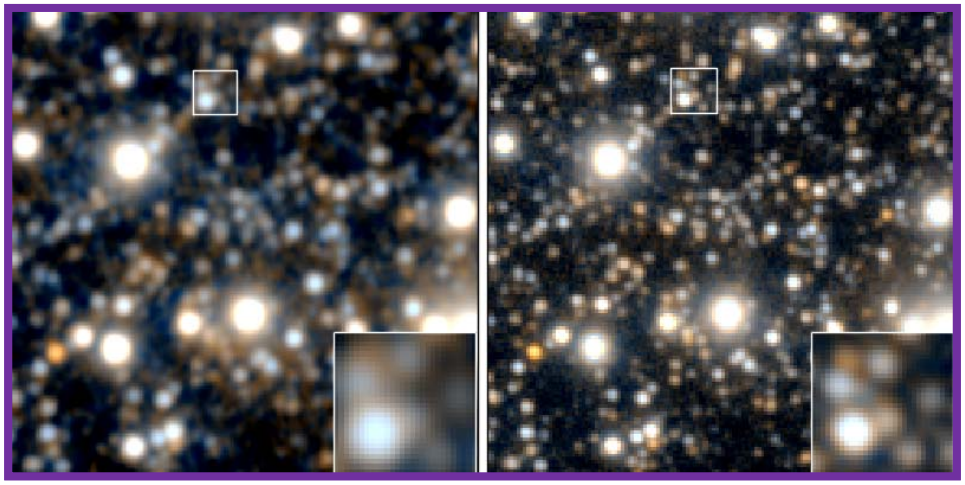


Challenge and Opportunity: Analysis Near the Data

- Complex and high-impact queries
  - Efficient billion-row multi-table queries, with VO protocols and optimized local performance
  - Enhanced statistical views of query results; leverage ongoing visualization work
  - Sustainable implementation (queuing, asynchronous TAP)
  - Alternate DBMS options, e.g. LSST's Qserv, multiple instances of DB
- Test integrating standard pixel analysis packages (source extraction, PSF-fitting, quantitative morphology)
- User-optimization of algorithms: Python notebooks, e.g. IPython, Jupyter, for collaboration, record keeping, and publishing to the cloud
- Coordination of user access to intensive computation on Archive hardware (VMs, Docker)



Finn et al. measure cluster galaxy gas disks using Spitzer and WISE, and stellar disks in the optical. Expanding to the entire WISE all-sky data set, this technique will inform models of star formation truncation as a function of environmental density.

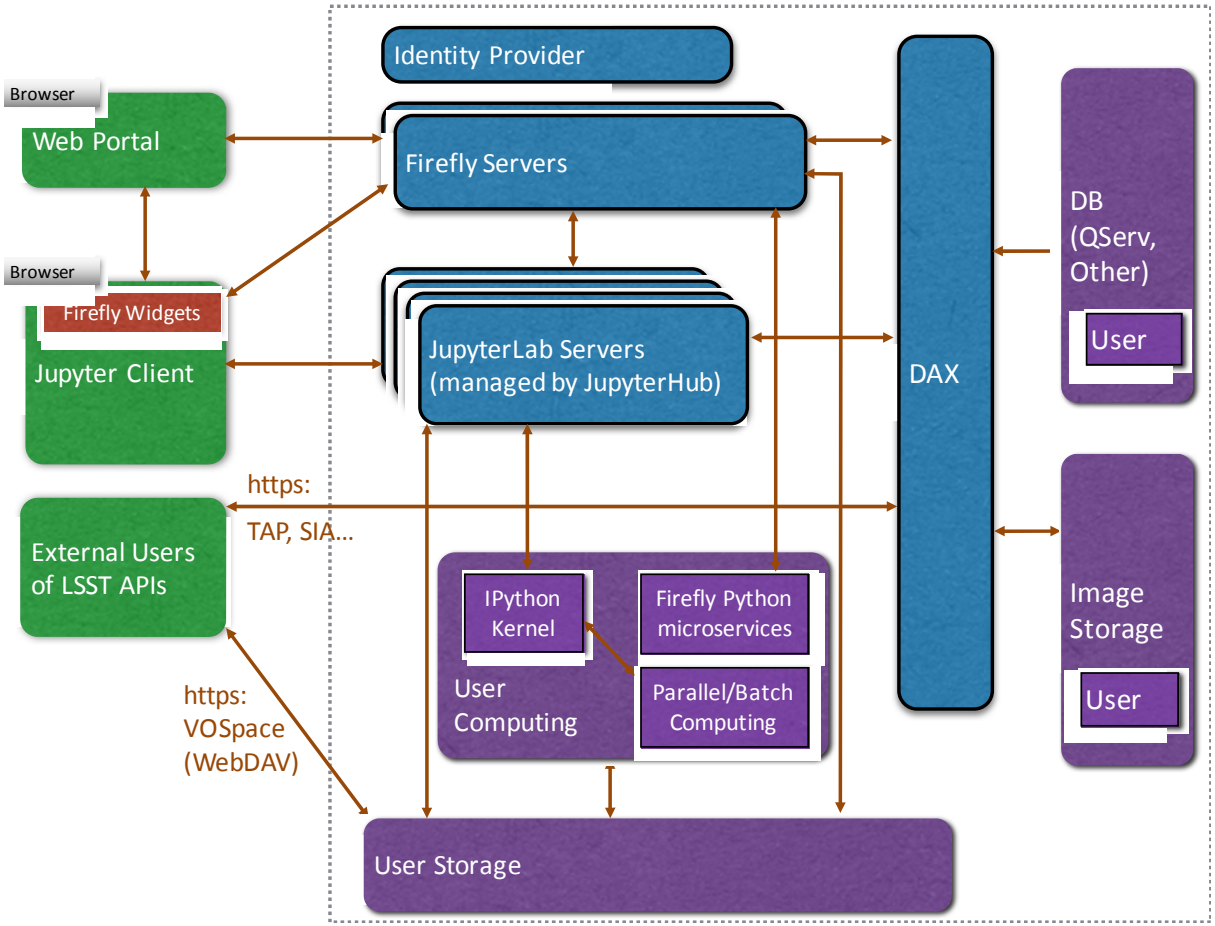


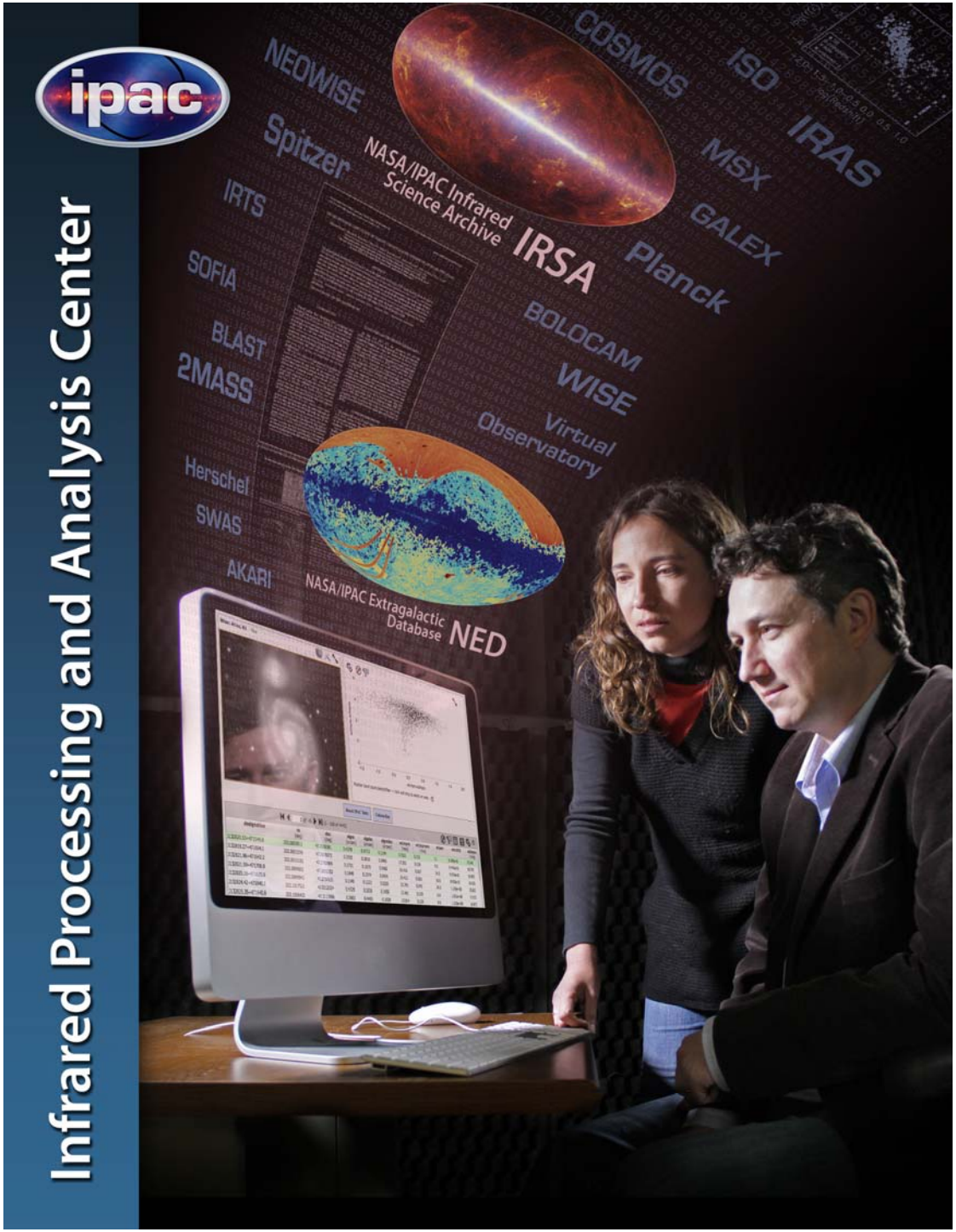
Lang (2014) reprocessed WISE single-exposure images to optimize measurements of extended sources. Other science objectives will require alternate processing.

The LSST Science Platform

IPAC Team is integrating the three LSST Science Platform Aspects.

- Provides access to LSST data via three "Aspects":
  - API: IVOA-standard access to catalogs and images; support for user data
  - Web Portal: Structured access to all data with viz and discovery tools
  - Notebooks: Interactive Python environment based on JupyterLab
- Aspects are integrated:
  - Near-data user computing and storage
  - Workflows can cross aspects





Summary

- Archives double or more the science return from a mission: data discovery and archive interoperability enhance this.
- IPAC is learning to use Big Data techniques to serve modern astronomical data sets and support their exploitation.
- Organization is the key to managing Big Data: must account for how data are transferred, queried, accessed, and analyzed.
- Processing at the Archive is already under development. The next challenge will be processing at multiple archives.

Acknowledgements

This update describes the work of scores of scientists, developers, and engineers.

The following* directly contributed to this update:

- George Helou, IPAC Executive Director
- IRSA: Steve Groom, Harry Teplitz, Vandana Desai, Justin Howell, Luisa Rebull, and the IRSA team.
- NED: Joe Mazzarella, Rick Ebert, Jeff Jacobson, and the NED team.
- Exoplanet Archive: Rachel Akeson, David Ciardi, and the Exoplanet Archive team
- ZTF: Frank Masci, Ben Rusholme, and the ZTF team.
- Montage: Bruce Berriman and John Good
- Firefly: Trey Roby, Loi Ly
- LSST: Xiuqin Wu, Gregory Dubois-Felsmann
- Science Research: Yossi Shvarzvald, Davy Kirkpatrick, Peter Capak, Dan Masters
- Spitzer: Sean Carey, Jim Ingalls, and the Spitzer Science Center Team
- NEOCAM: Roc Cutri, Carrie Nugent
- ICE: Gordon Squires, Janice Lee, Robert Hurt, Tim Pyle, Jake Llamas, and the ICE Team
- IPAC Systems Engineering: Dave Flynn and the ISG SysEng Team

**Many IPAC staff work on multiple activities. All are named only once here.*