Big Data in astronomy: an overview

SOMACHINE 2020

Federica B. Bianco
University of Delaware
Physics and Astronomy
Biden School of Public Policy and Administration
Data Science Institute
NYU Center for Urban Science and Progress

Rubin Observatory LSST Science Collaborations Coordinator

Rubin LSST Transients and Variable Stars Science Collaborations Chair

this slide deck:

https://slides.com/federicabianco/bdastro

1. historical perspective: BD context
2. BD from astronomical surveys
    - optical
    - radio
    - gravitational waves
3. space-based astronomy BD problem
4. corwdsourcing approach
5. time domain astronomy BD problems
6. platforms
    - computational platforms in astro
    - computational platforms in other fields
    - VO

- shameless plug: the Rubin LSST Science Collaborations

Historical perspective

"Data larger than can be analyzed with typical tool"

"Data that stresses the infrastructure"

"Data that does not fit in memory"

1/6

Historical perspective

Historical perspective

"Data larger than can be analyzed with typical tool"

John R. Mashey Chief Scientist, SGI, mid-1990s

https://static.usenix.org/event/usenix99/invited_talks/mashey.pdf

@fedhere

Historical perspective

"Data larger than can be analyzed with typical tool"

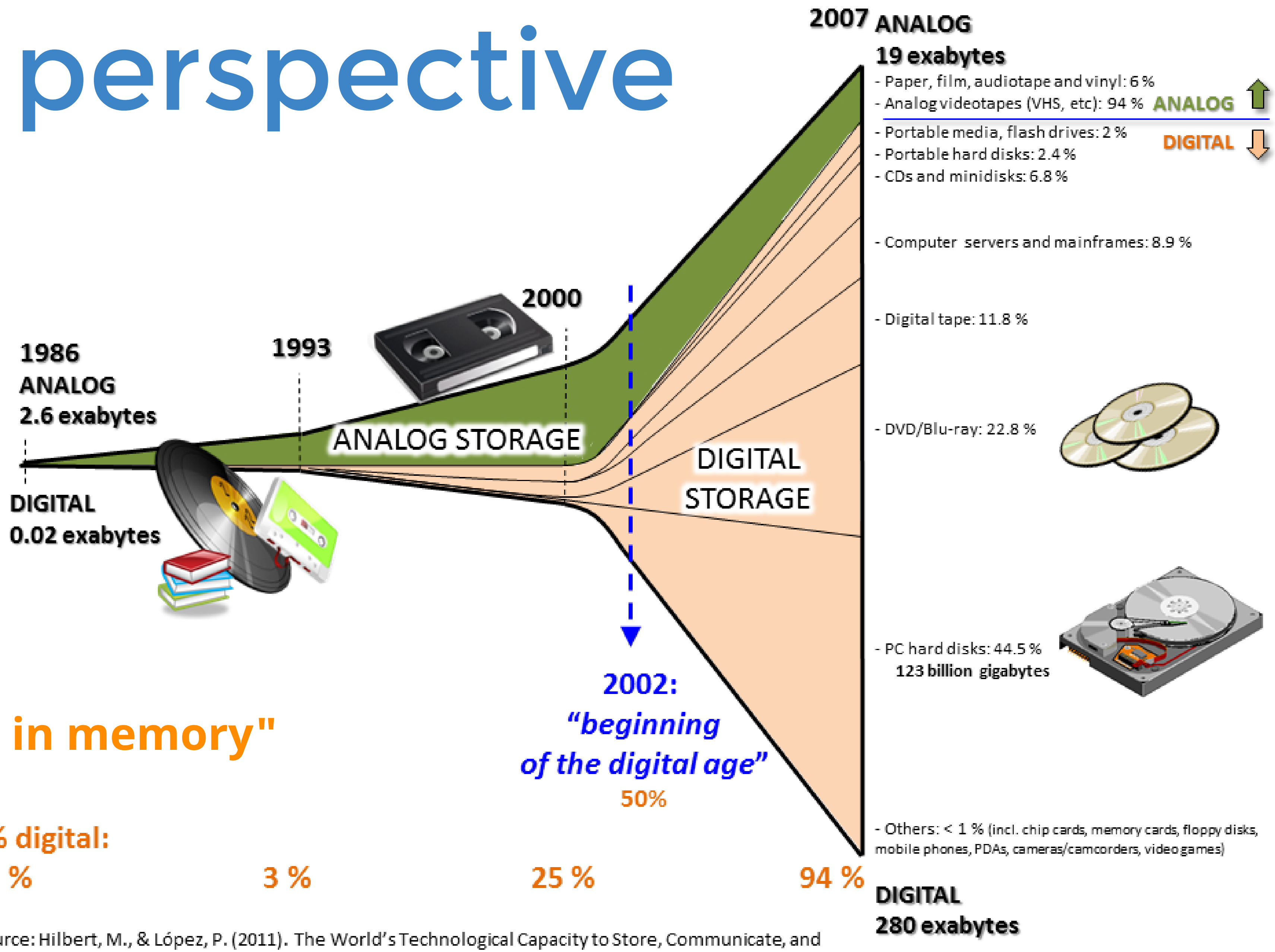
"Data that stresses the infrastructure"

John R. Mashey Chief Scientist, SGI, mid-1990s

https://static.usenix.org/event/usenix99/invited_talks/mashey.pdf

Historical perspective

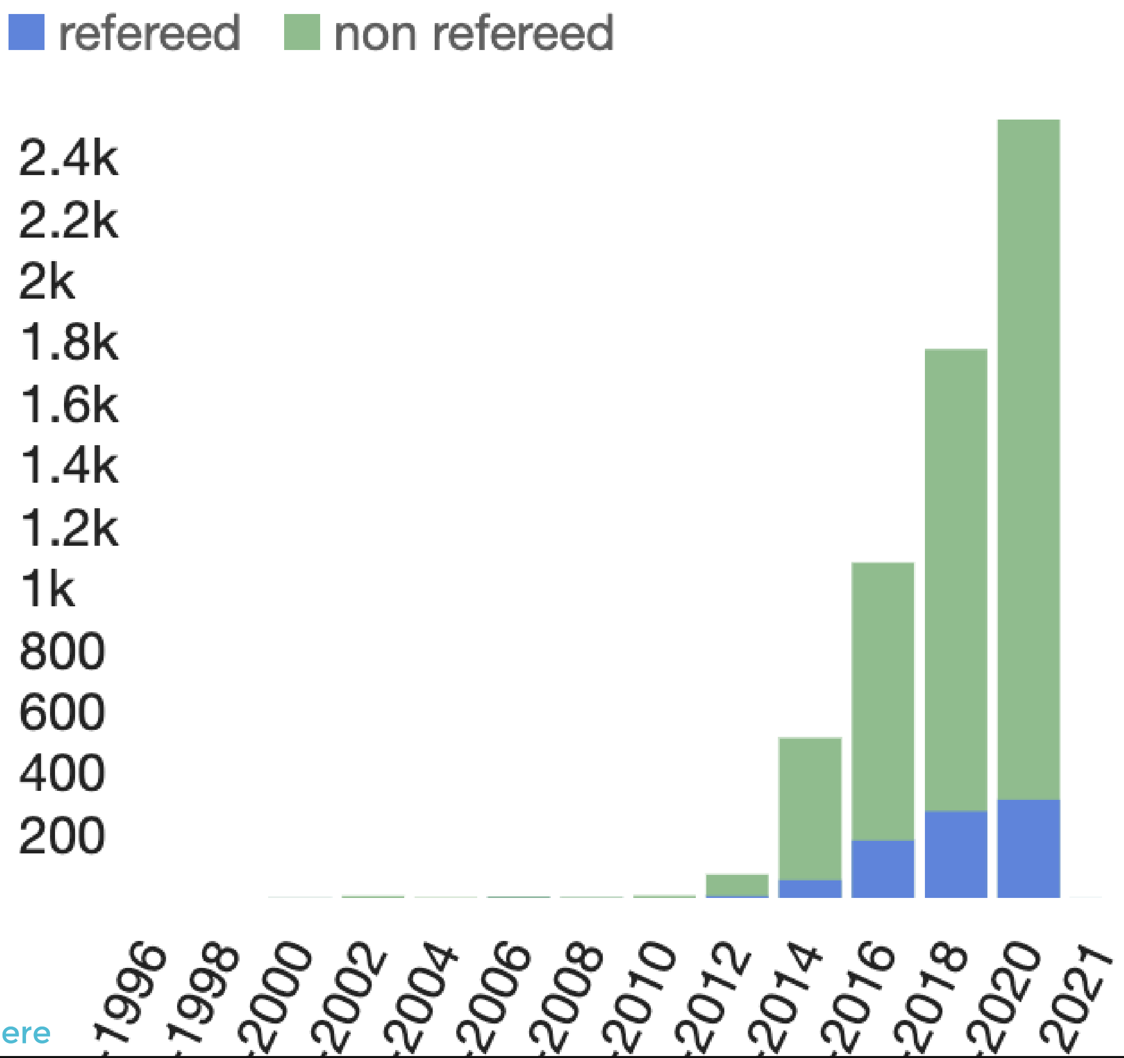
"Data that does not fit in memory"



Source: Hilbert, M., & López, P. (2011). The World's Technological Capacity to Store, Communicate, and Compute Information. *Science*, 332(6025), 60 –65. http://www.martinhilbert.net/WorldInfoCapacity.html

@fedhere

Historical perspective

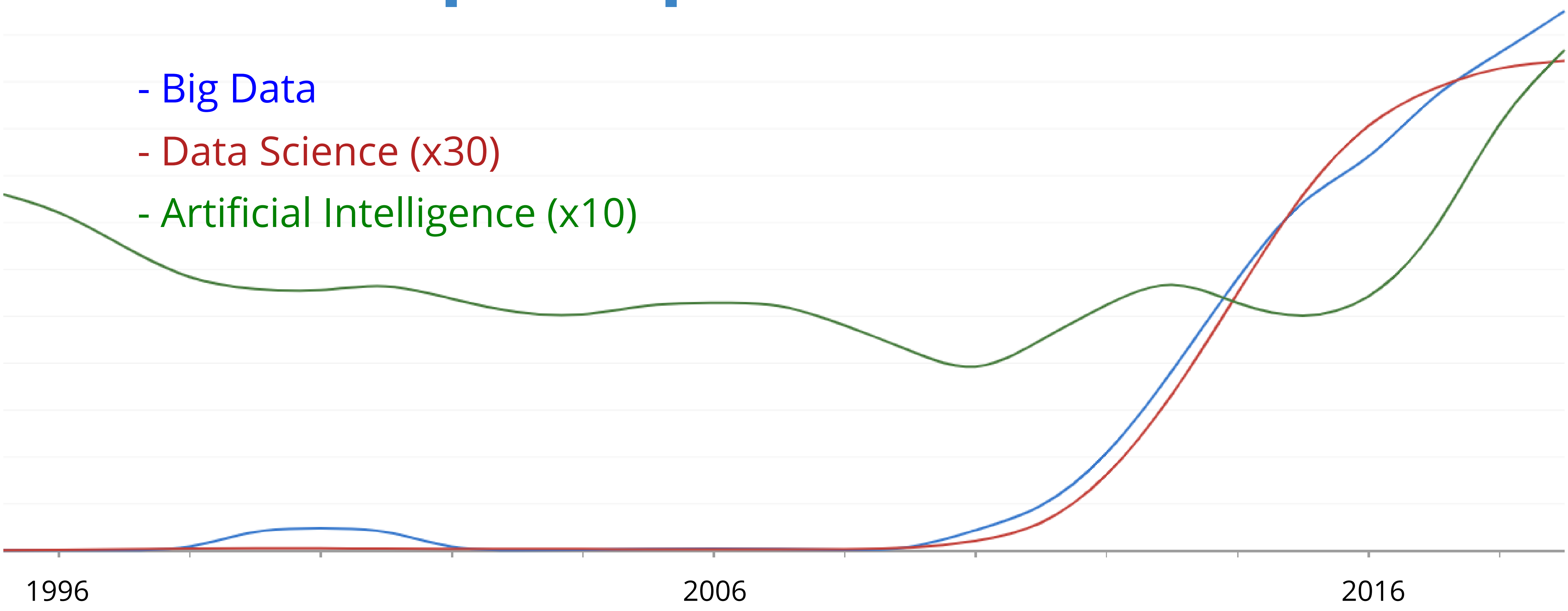


Big Data in astronomy papers

(source: ADS)

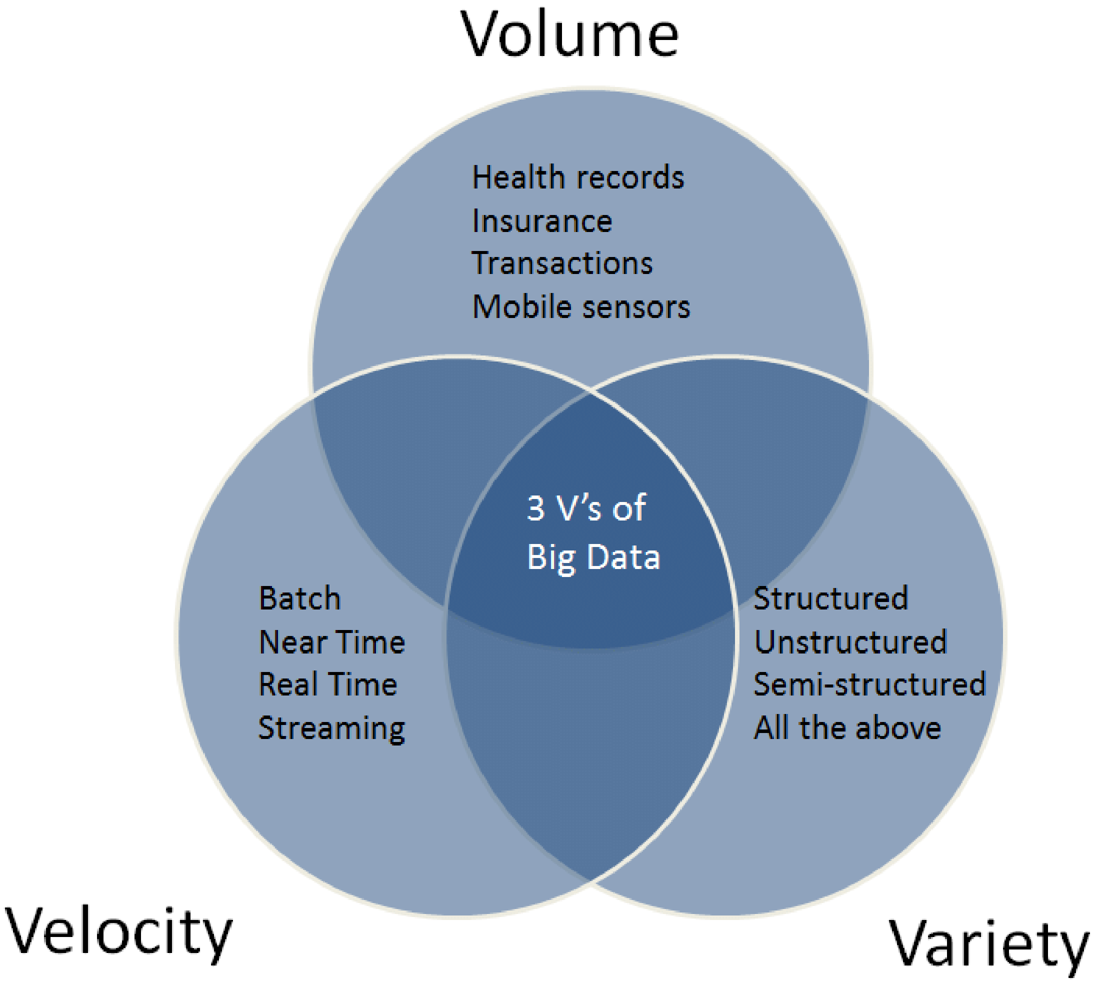
@fedhere

Historical perspective

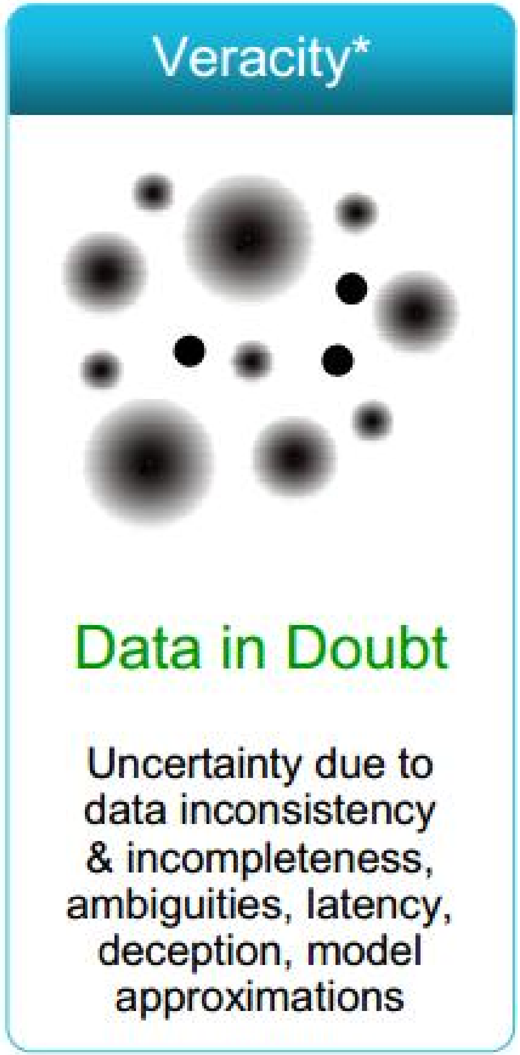


occurrence of term in Google-books corpus https://books.google.com/ngrams

Historical perspective



Gartner report 2001



Michael Walker on 28 November 2012

@fedhere

Big Data: Astronomical or Genomical?

Zachary D. Stephens, Skylar Y. Lee, Faraz Faghri, Roy H. Campbell, Chengxiang Zhai, Miles J. Efron, Ravishankar Iyer,
Michael C. Schatz, Saurabh Sinha, Gene E. Robinson

Published: July 7, 2015 • https://doi.org/10.1371/journal.pbio.1002195

Data Phase	Astronomy	Twitter	YouTube	Genomics
Acquisition	25 zetta-bytes/year	0.5–15 billion tweets/year	500–900 million hours/year	1 zetta-bases/year
Storage	1 EB/year	1–17 PB/year	1–2 EB/year	2–40 EB/year
Analysis	In situ data reduction	Topic and sentiment mining	Limited requirements	Heterogeneous data and analysis
	Real-time processing	Metadata analysis		Variant calling, ~2 trillion central processing unit (CPU) hours
	Massive volumes			All-pairs genome alignments, ~10,000 trillion CPU hours
Distribution	Dedicated lines from antennae to server (600 TB/s)	Small units of distribution	Major component of modern user's bandwidth (10 MB/s)	Many small (10 MB/s) and fewer massive (10 TB/s) data movement

doi:10.1371/journal.pbio.1002195.t001

https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1002195

what drives scientific discovery in astronomy

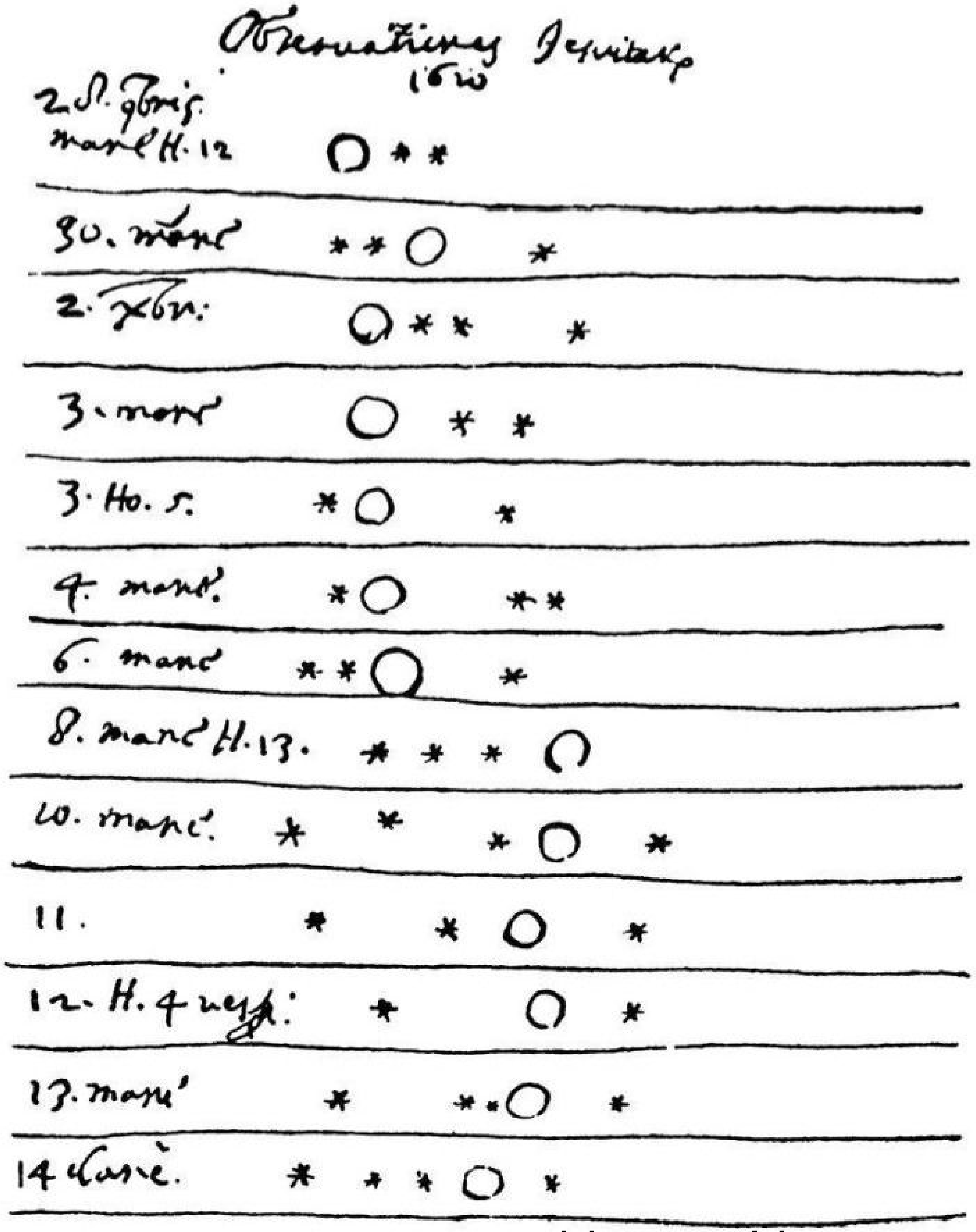
Galileo Galilei 1610

what drives astronomy

Experiment driven

Following: Djorgovski

https://events.asiaa.sinica.edu.tw/school/20170904/talk/djorgovski1.pdf



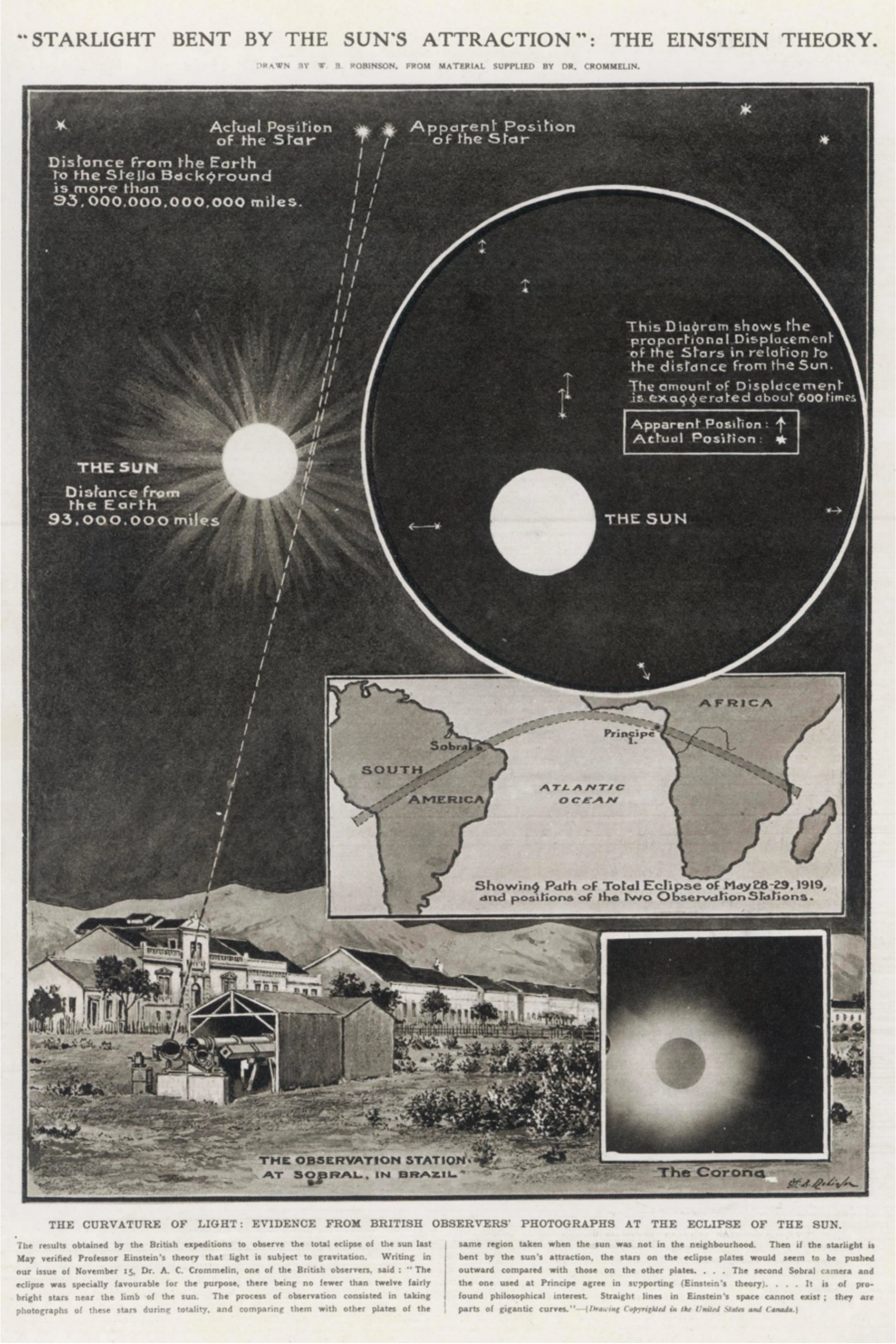
Galileo Galilei 1610

@fedhere

what drives astronomy

Experiment driven

Theory driven | Falsifiability



"STARLIGHT BENT BY THE SUN'S ATTRACTION": THE EINSTEIN THEORY.

DRAWN BY W. B. ROBINSON, FROM MATERIAL SUPPLIED BY DR. CROMMELIN.

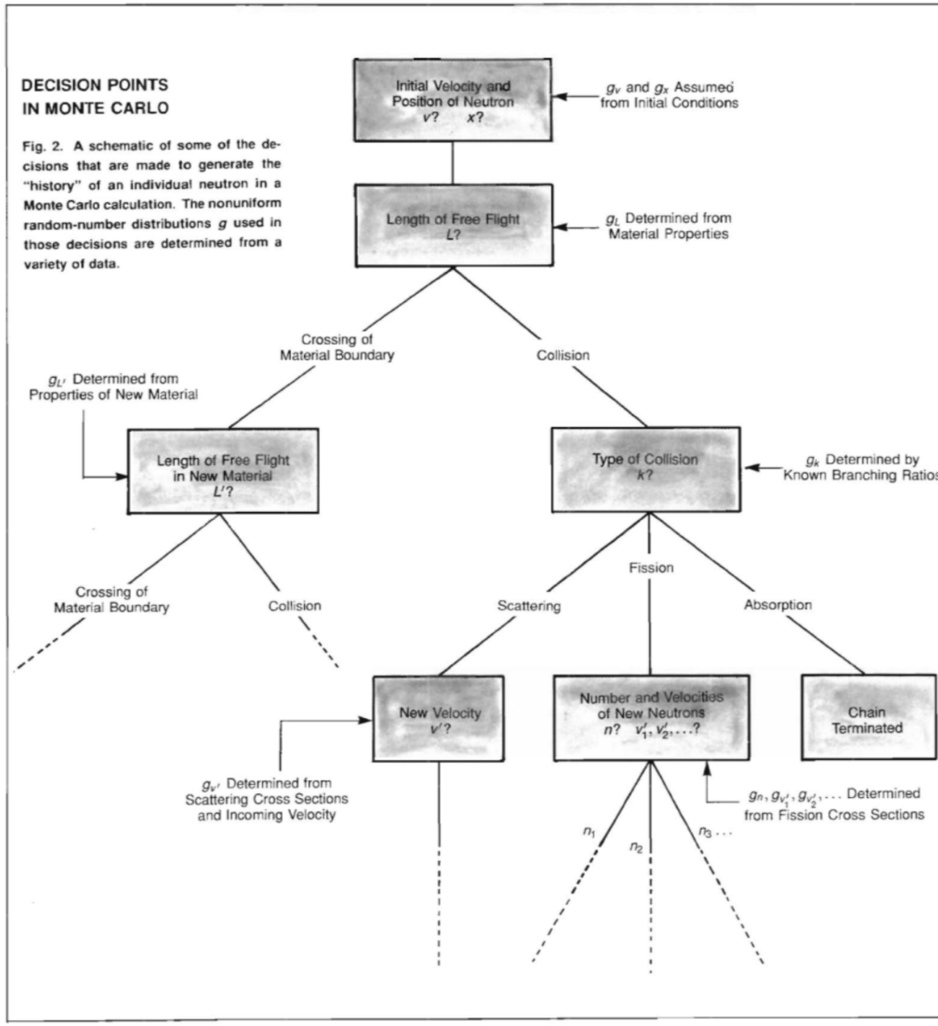
THE CURVATURE OF LIGHT: EVIDENCE FROM BRITISH OBSERVERS' PHOTOGRAPHS AT THE ECLIPSE OF THE SUN.

The results obtained by the British expeditions to observe the total eclipse of the sun last May verified Professor Einstein's theory that light is subject to gravitation. Writing in our issue of November 15, Dr. A. C. Crommelin, one of the British observers, said: "The eclipse was specially favourable for the purpose, there being no fewer than twelve fairly bright stars near the limb of the sun. The process of observation consisted in taking photographs of these stars during totality, and comparing them with other plates of the same region taken when the sun was not in the neighbourhood. Then if the starlight is bent by the sun's attraction, the stars on the eclipse plates would seem to be pushed outward compared with those on the other plates. . . . The second Sobral camera and the one used at Principe agree in supporting (Einstein's theory). . . . It is of profound philosophical interest. Straight lines in Einstein's space cannot exist; they are parts of gigantic curves."—*(Drawing Copyrighted in the United States and Canada.)*

Enistein 1916

what drives astronomy

[Stanislav Ulam](#)



Experiment driven

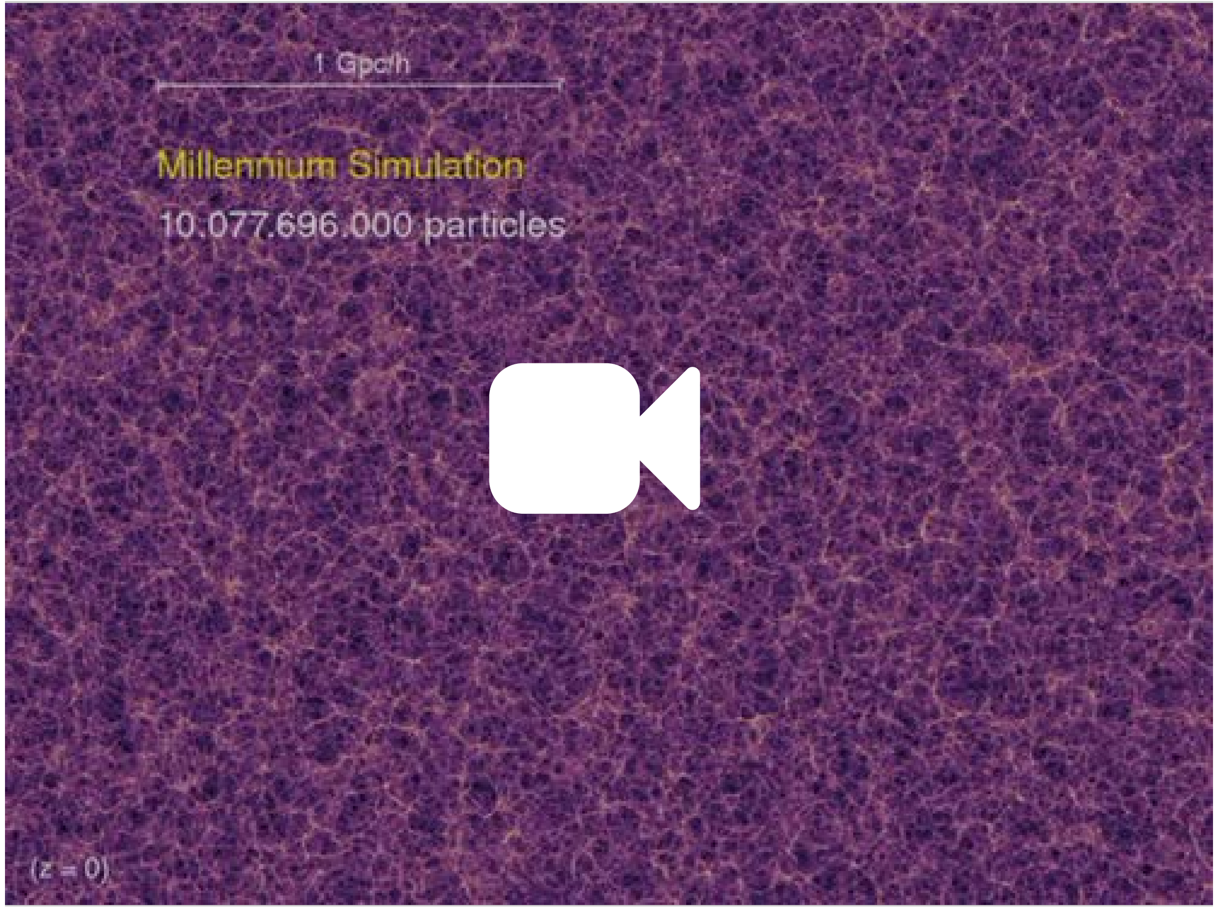
Theory driven | Falsifiability

Simulations | Probabilistic inference | Computation

http://www-star.st-and.ac.uk/~kw25/teaching/mcrt/MC_history_3.pdf

Ulam 1947

what drives astronomy



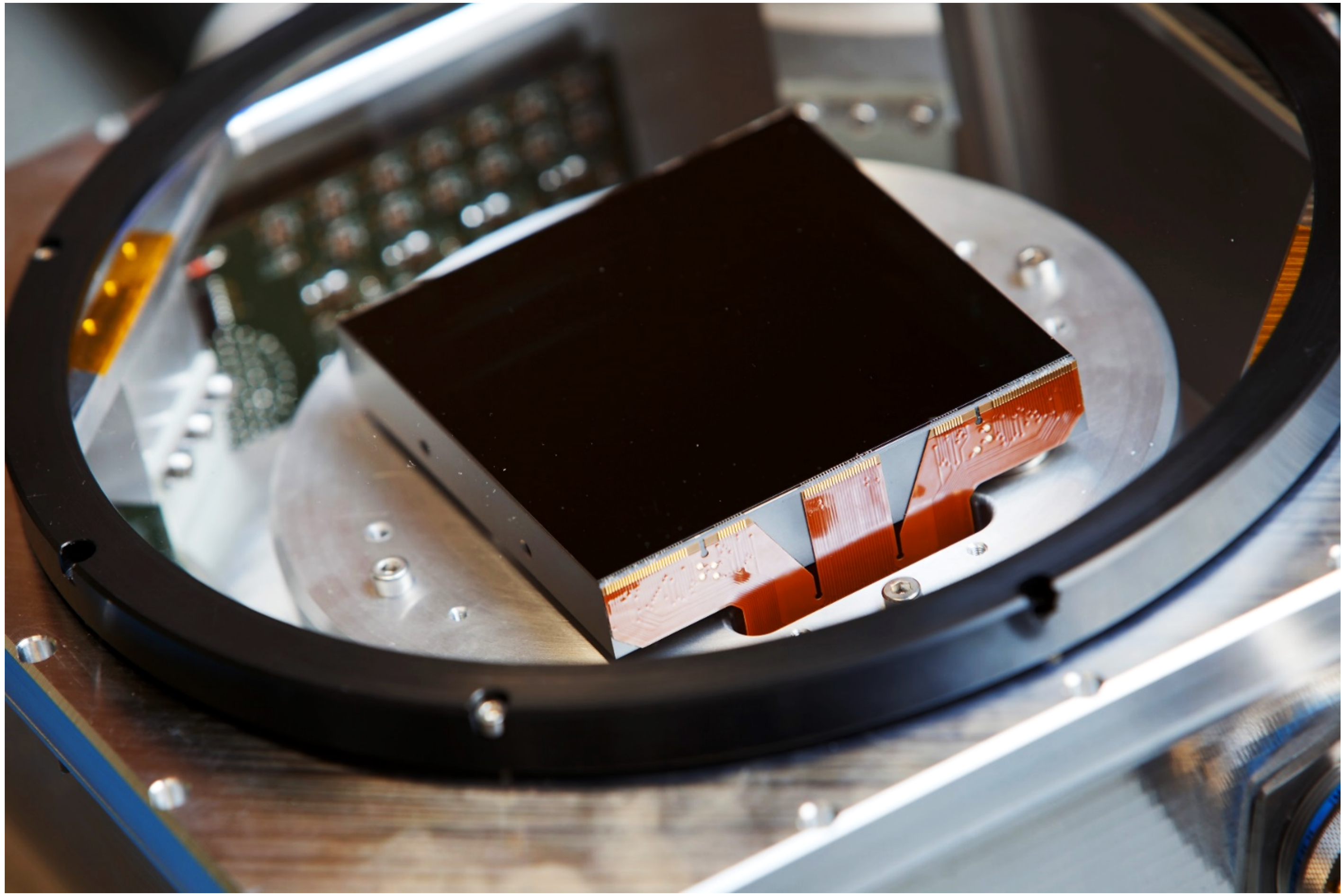
Experiment driven

Theory driven | Falsifiability

Simulations | Probabilistic inference | Computation

Ulam 1947

what drives astronomy



Experiment driven

Theory driven | Falsifiability

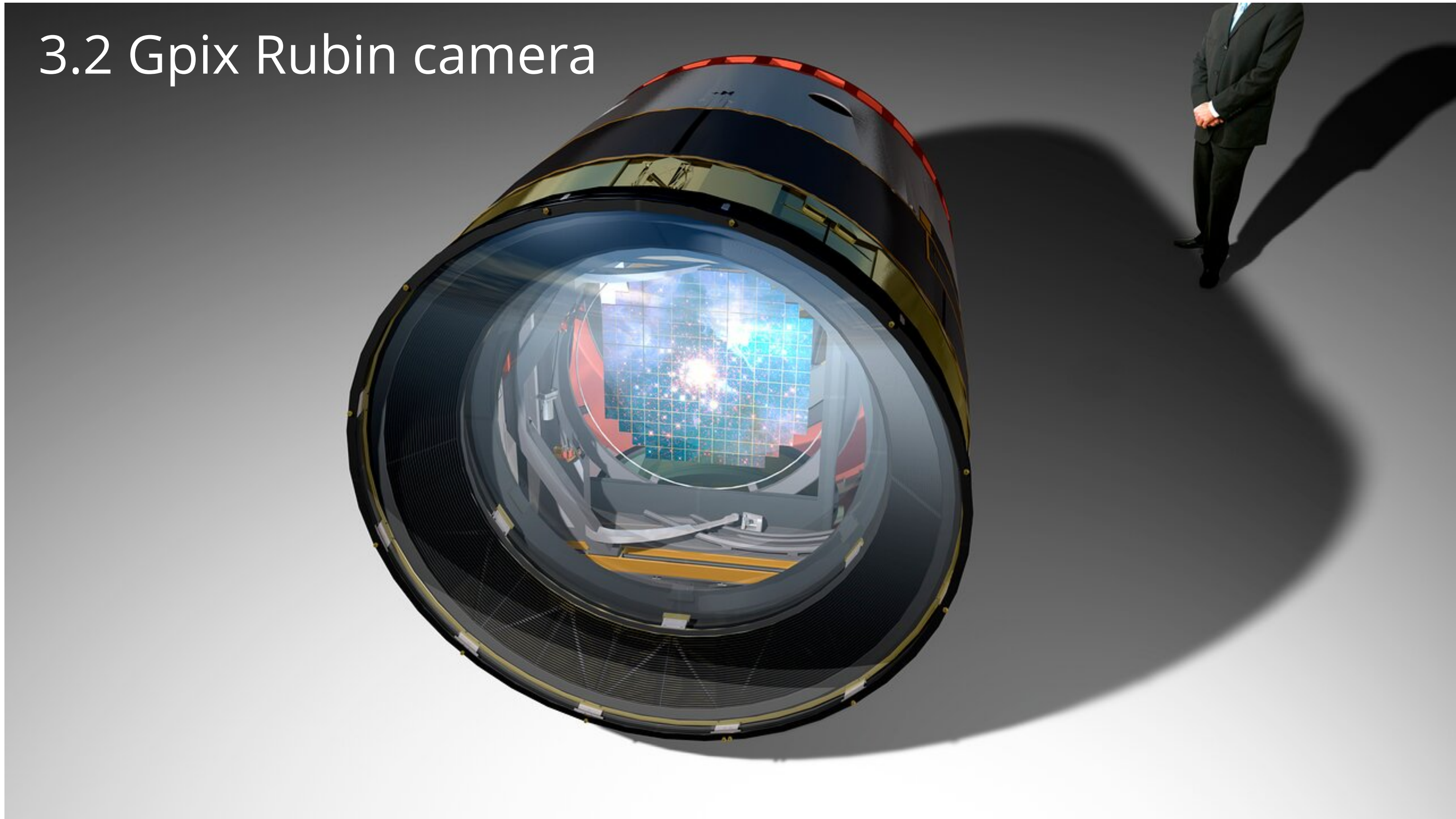
Simulations | Probabilistic inference | Computation

Data | Survey astronomy | Computation | pattern discovery

@fedhere

the 2000s

what drives astronomy



Experiment driven

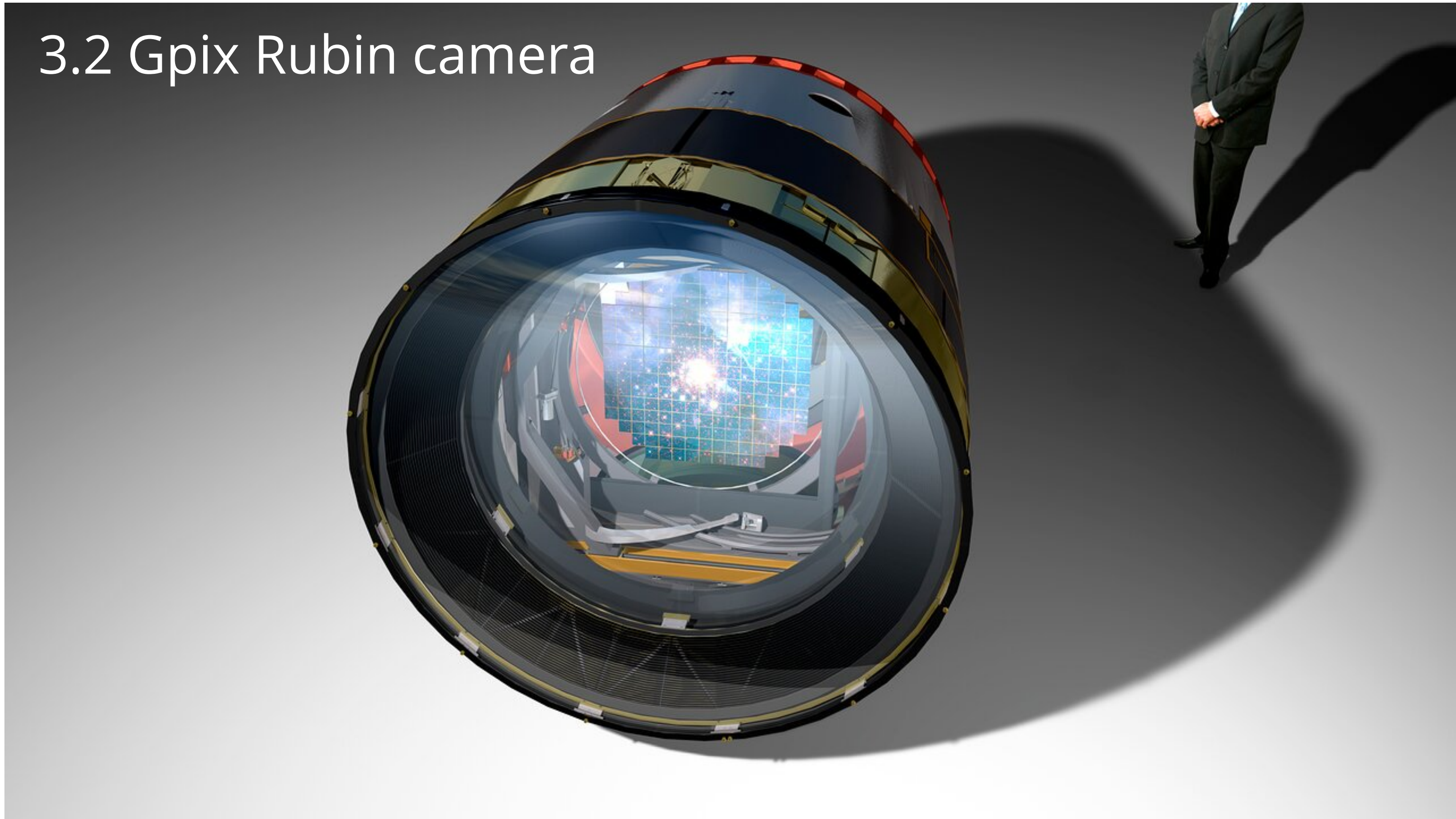
Theory driven | Falsifiability

Simulations | Probabilistic inference | Computation

Data | Survey astronomy | Computation | pattern discovery

the 2000s

what drives astronomy



Experiment driven

Theory driven | Falsifiability

Simulations | Probabilistic inference | Computation

Data | Survey astronomy | Computation | pattern discovery

lazy learning

learning by example (supervised learning)

pattern discovery (unsupervised learning)

4-V of Big Data

V1: Volume

Number of bites

Number of pixels

Number of rows in a data table x number of columns for catalogs

V2: Variety

Diverse science return from the same dataset.

Multiwavelength

Multimessenger

Images and spectra

V3: Velocity

real time analysis, edge computing, data transfer

V4: Veracity

This V will refer to both data quality and availability (added in 2012)

Gartner report 2001

@fedhere

4-V of Big Data in astronomy

V1: Volume

Number of bites

Number of pixels

Number of rows in a data table x number of columns for catalogs

2/6

BD from astronomical surveys

astronomical data production

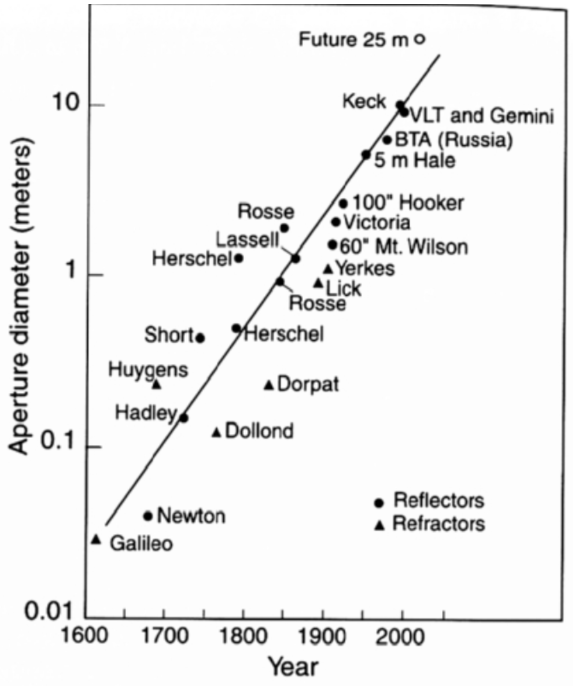


Fig. 1. Evolution of telescope aperture diameter over the last four centuries. According to the trend line shown, the diameter of the largest telescopes doubles about every 40 years. The 20- to 30-meter class telescopes planned for the 2015 time frame display a somewhat faster growth rate than the historical trend.

astronomical data production

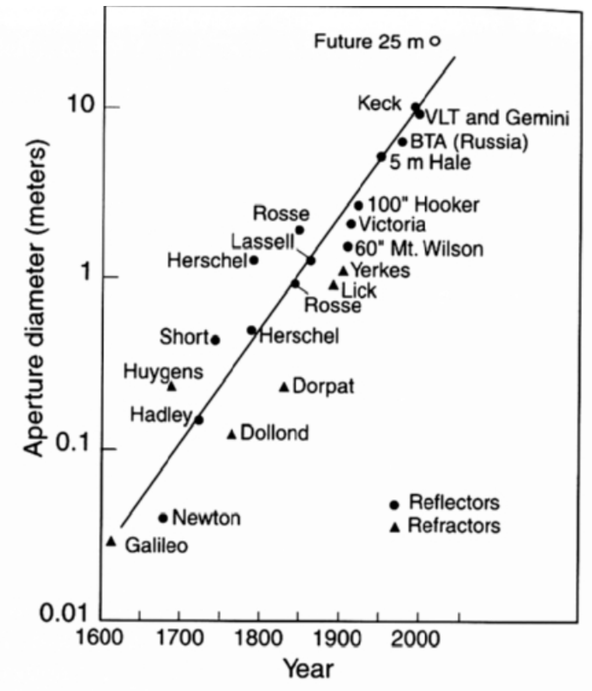
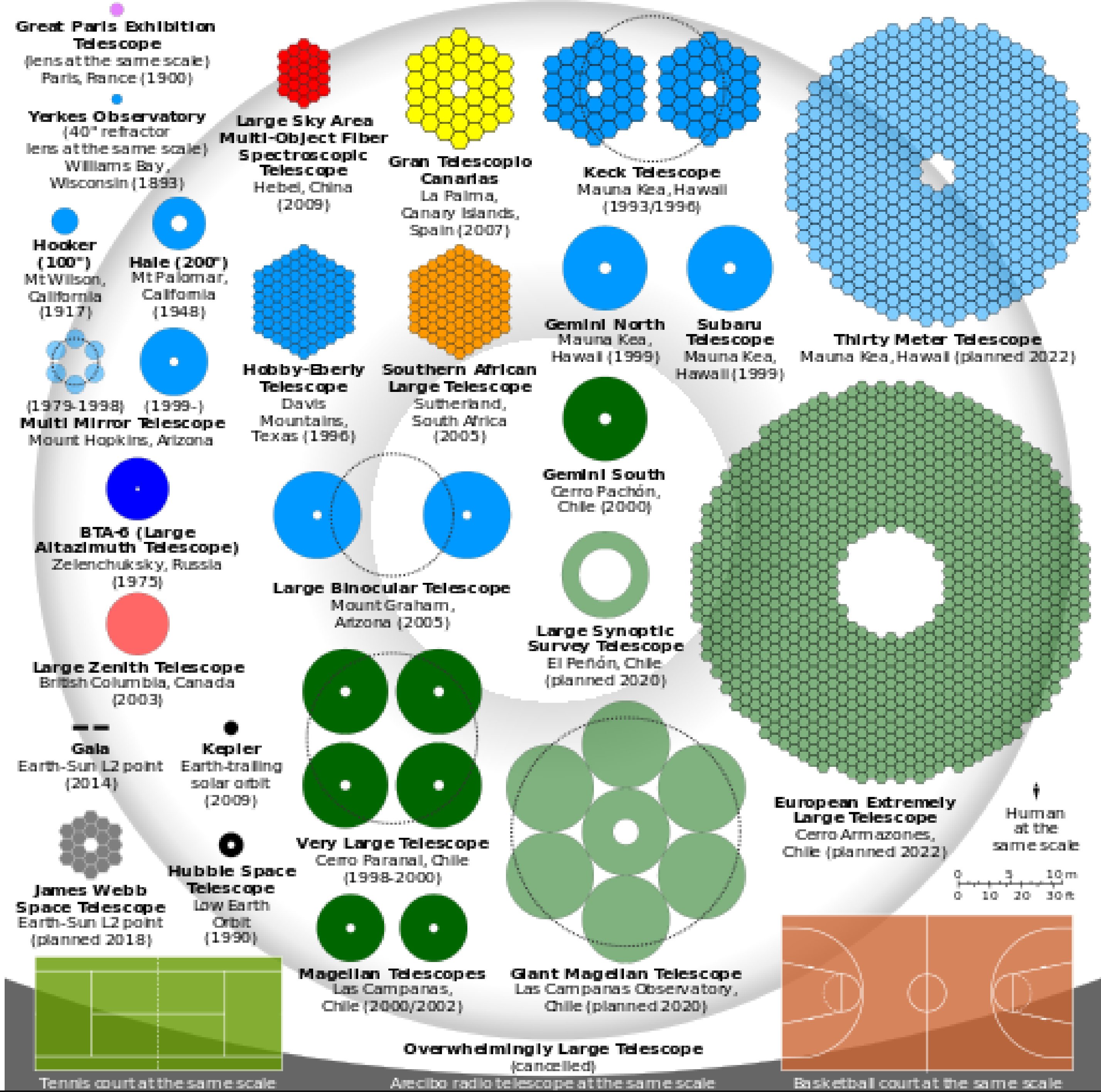


Fig. 1. Evolution of telescope aperture diameter over the last four centuries. According to the trend line shown, the diameter of the largest telescopes doubles about every 40 years. The 20- to 30-meter class telescopes planned for the 2015 time frame display a somewhat faster growth rate than the historical trend.



@fedhere

astronomical data production

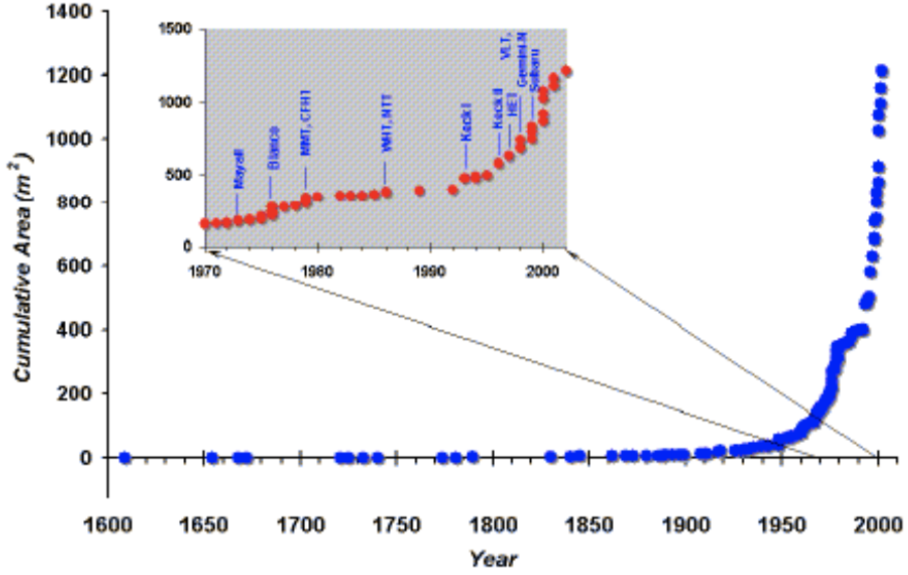
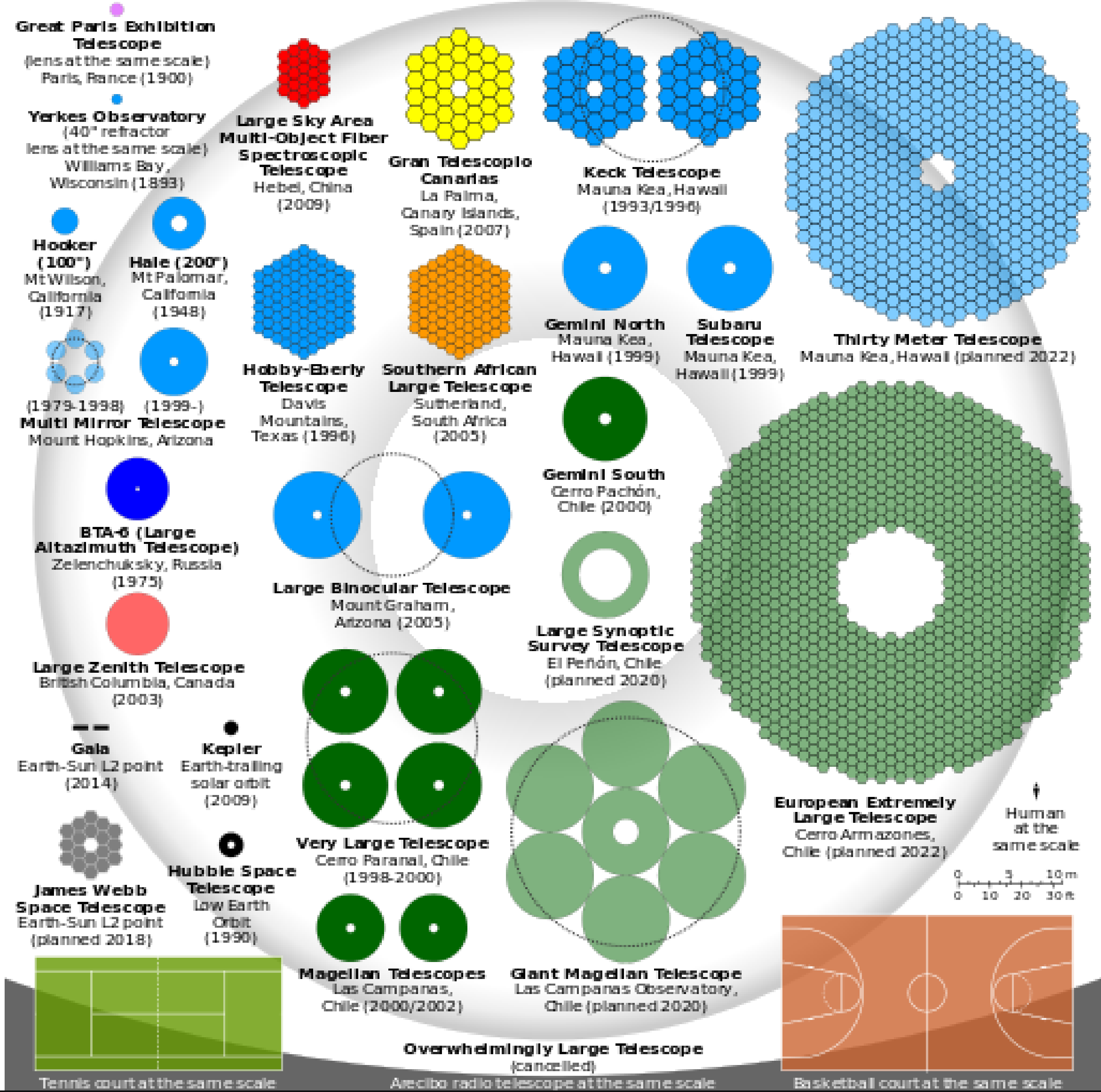


Figure 1 The growth in cumulative telescope collecting area over the past 400 years, with each point representing a completed ground-based telescope. The combination of the ability to manufacture and support large mirrors combined with adaptive optics has given the new generation of large telescopes tremendous scientific gains over the previous 4-m telescopes. For example, an 8-m telescope delivering images of 0.1 arcsec can observe point-like objects at least 20 times fainter than a conventional 4-m telescope delivering 1.0 arcsec images. Will we see such gains in the next generation of telescopes? MMT, Multiple Mirror Telescope; CFHT, Canada-France-Hawaii Telescope; WHT, William Herschel Telescope; NTT, New Technology Telescope; HET, Hobby-Eberly Telescope; VLT, Very Large Telescope; Gem-N, Gemini North Telescope.



@fedhere

astronomical data production

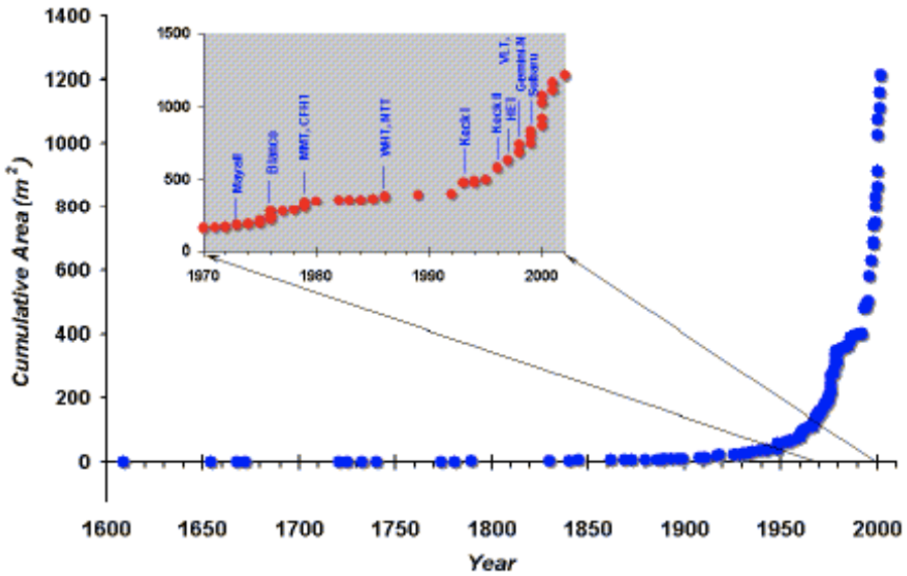
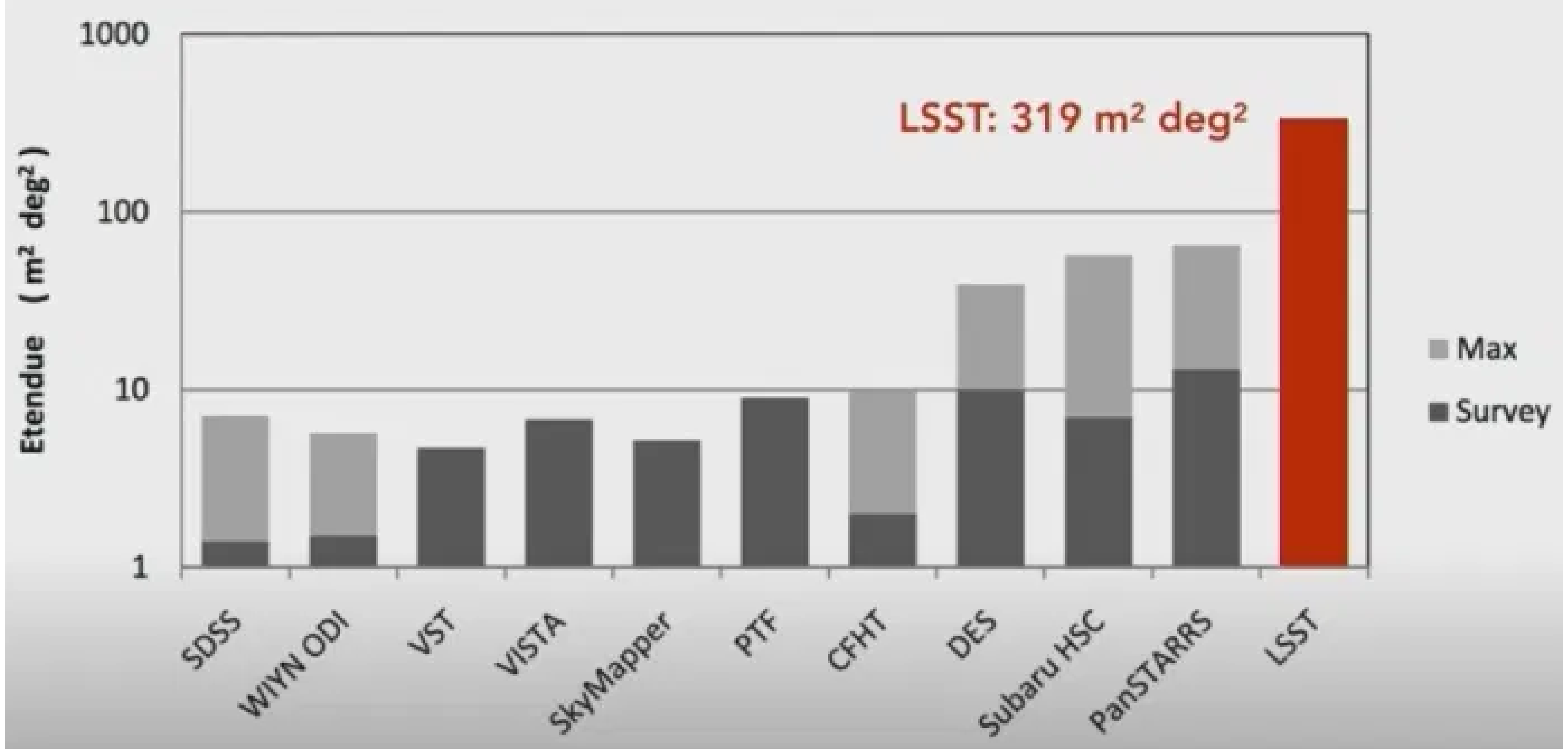


Figure 1 The growth in cumulative telescope collecting area over the past 400 years, with each point representing a completed ground-based telescope. The combination of the ability to manufacture and support large mirrors combined with adaptive optics has given the new generation of large telescopes tremendous scientific gains over the previous 4-m telescopes. For example, an 8-m telescope delivering images of 0.1 arcsec can observe point-like objects at least 20 times fainter than a conventional 4-m telescope delivering 1.0 arcsec images. Will we see such gains in the next generation of telescopes? MMT, Multiple Mirror Telescope; CFHT, Canada-France-Hawaii Telescope; WHT, William Herschel Telescope; NTT, New Technology Telescope; HET, Hobby-Eberly Telescope; VLT, Very Large Telescope; Gem-N, Gemini North Telescope.

Both data volumes and data rates grow exponentially, with a doubling time ~ 1.5 years

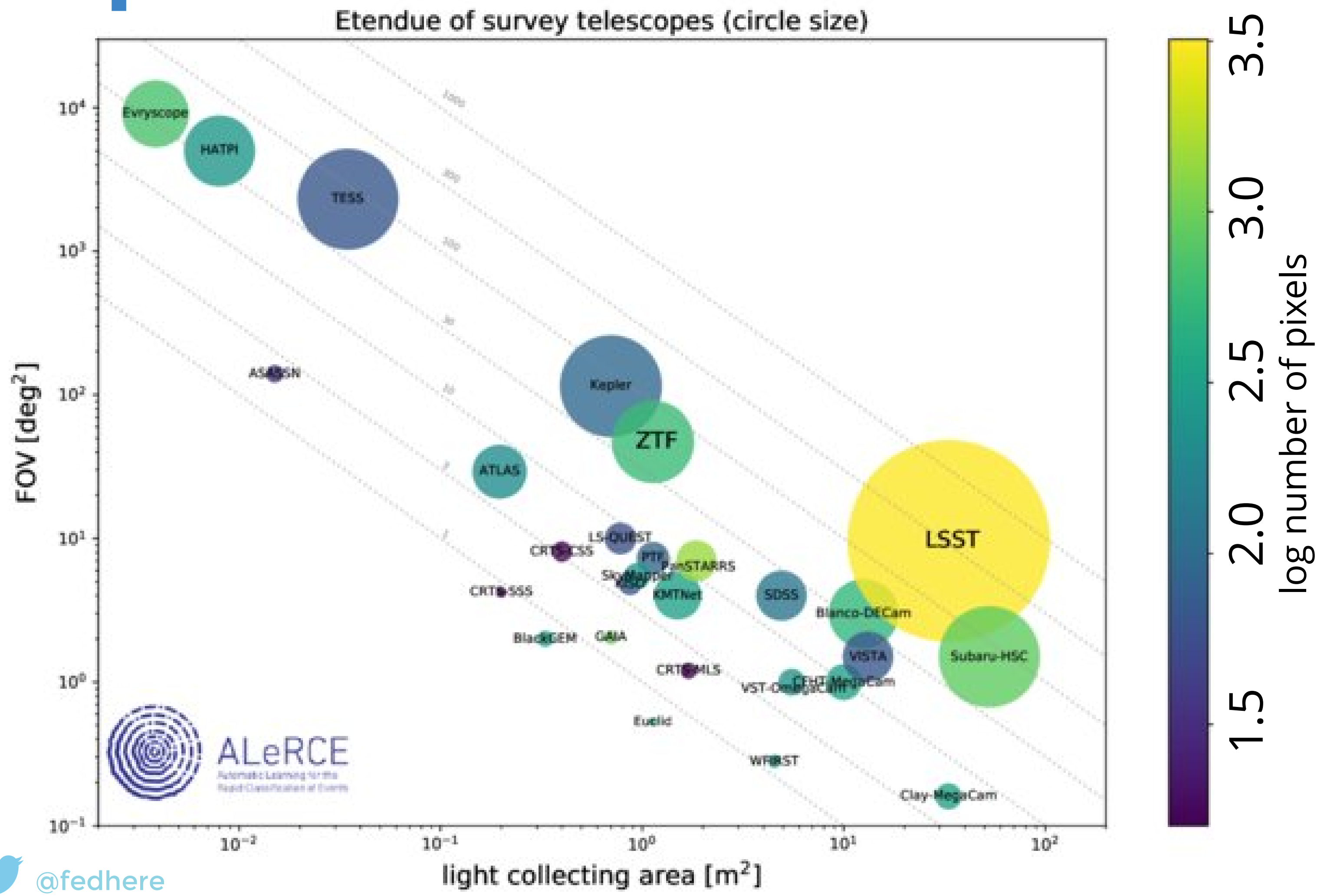
astronomical data production

Etendue: *area x FoV*



@fedhere

astronomical data production



Etendue: *area x FoV*

data volume :
area x FoV
x
resolution
x
sensitivity

@fedhere

astronomical data volume

number of sources

Table 1. Main data for the most important all-sky and large-area astronomical surveys providing multi-wavelength photometric data. Catalogues are given in the order of increasing wavelengths.

Survey, catalogue	Years	Spectral range	Sky area (deg^2)	Sensitivity (mag/mJy)	Number of sources	Density (obj/deg^2)
Fermi-GLAST	2008–2014	10 MeV–100 GeV	All-sky		3033	0.07
CGRO	1991–1999	20 keV–30 GeV	All-sky		1300	0.03
INTEGRAL	2002–2014	15 keV–10 MeV	All-sky		1126	0.03
ROSAT BSC	1990–1999	0.07–2.4 keV	All-sky		18,806	0.46
ROSAT FSC	1990–1999	0.07–2.4 keV	All-sky		105,924	2.57
GALEX AIS	2003–2012	1344–2831Å	21,435	20.8 mag	65,266,291	3044.85
APM	2000	opt b, r	20,964	21.0 mag	166,466,987	7940.61
MAPS	2003	opt O, E	20,964	21.0 mag	89,234,404	4256.55
USNO-A2.0	1998	opt B, R	All-sky	21.0 mag	526,280,881	12,757.40
USNO-B1.0	2003	opt B, R, I	All-sky	22.5 mag	1,045,913,669	25,353.64
GSC 2.3.2	2008	opt j, V, F, N	All-sky	22.5 mag	945,592,683	22,921.79
Tycho-2	1989–1993	opt BT, VT	All-sky	16.3 mag	2,539,913	61.57
SDSS DR12	2000–2014	opt u, g, r, i, z	14,555	22.2 mag	932,891,133	64,094.20
DENIS	1996–2001	0.8–2.4 μm	16,700	18.5 mag	355,220,325	21,270.68
2MASS PSC	1997–2001	1.1–2.4 μm	All-sky	17.1 mag	470,992,970	11,417.46
2MASS ESC	1997–2001	1.1–2.4 μm	All-sky	17.1 mag	1,647,599	39.94
WISE	2009–2013	3–22 μm	All-sky	15.6 mag	563,921,584	13,669.83
AKARI IRC	2006–2008	7–26 μm	38,778	50 mJy	870,973	22.46
IRAS PSC	1983	8-120 μm	39,603	400 mJy	245,889	6.21
IRAS FSC	1983	8–120 μm	34,090	400 mJy	173,044	5.08
IRAS SSSC	1983	8–120 μm	39,603	400 mJy	16,740	0.42
AKARI FIS	2006–2008	50–180 μm	40,428	550 mJy	427,071	10.56
Planck	2009–2011	0.35–10 mm	All-sky	183 mJy	33,566	0.81
WMAP	2001–2011	3–14 mm	All-sky	500 mJy	471	0.01
GB6	1986–1987	6 cm	20,320	18 mJy	75,162	3.70
NVSS	1998	21 cm	33,827	2.5 mJy	1,773,484	52.43
FIRST	1999–2015	21 cm	10,000	1 mJy	946,432	94.64
SUMSS	2003–2012	36 cm	8,000	1 mJy	211,050	26.38
WENSS	1998	49/92 cm	9,950	18 mJy	229,420	23.06
7C	2007	198 cm	2,388	40 mJy	43,683	18.29

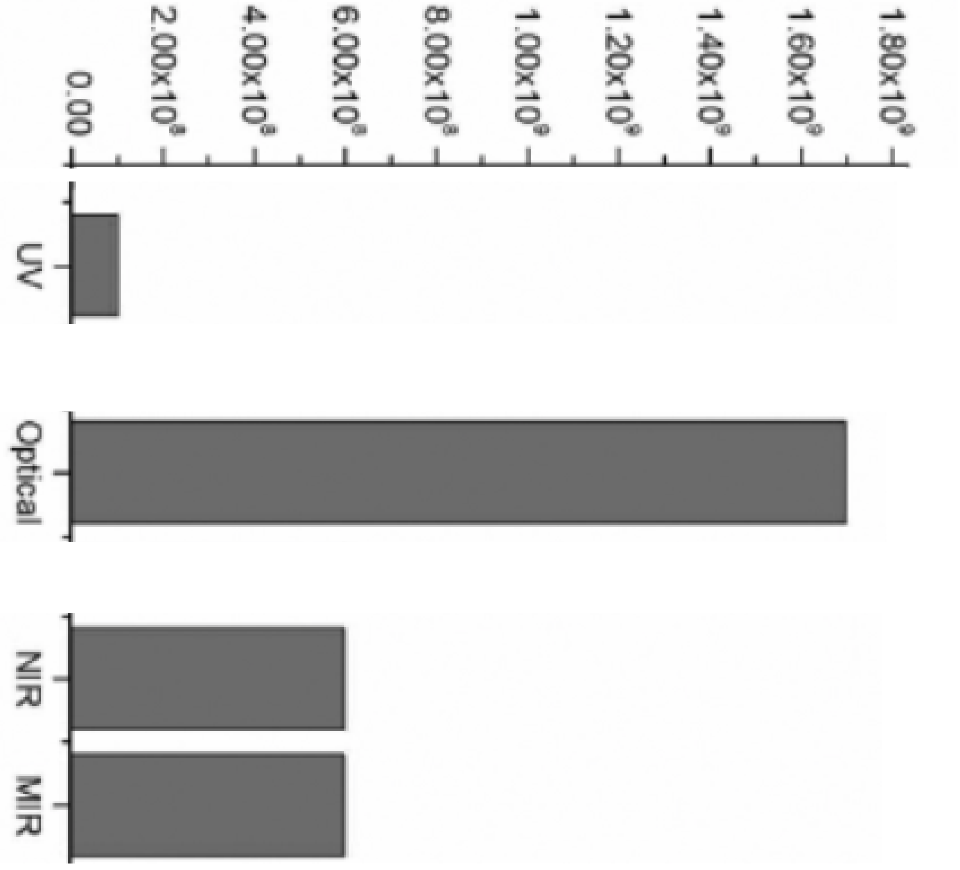
https://www.degruyter.com/view/journals/astro/25/1/article-p75.xml?language=en

astronomical data volume

number of sources

Table 1. Main data for the most important all-sky and large-area astronomical surveys providing multi-wavelength photometric data. Catalogues are given in the order of increasing wavelengths.

Survey, catalogue	Years	Spectral range	Sky area (deg^2)	Sensitivity (mag/mJy)	Number of sources	Density (obj/deg^2)
Fermi-GLAST	2008–2014	10 MeV–100 GeV	All-sky		3033	0.07
CGRO	1991–1999	20 keV–30 GeV	All-sky		1300	0.03
INTEGRAL	2002–2014	15 keV–10 MeV	All-sky		1126	0.03
ROSAT BSC	1990–1999	0.07–2.4 keV	All-sky		18,806	0.46
ROSAT FSC	1990–1999	0.07–2.4 keV	All-sky		105,924	2.57
GALEX AIS	2003–2012	1344–2831Å	21,435	20.8 mag	65,266,291	3044.85
APM	2000	opt b, r	20,964	21.0 mag	166,466,987	7940.61
MAPS	2003	opt O, E	20,964	21.0 mag	89,234,404	4256.55
USNO-A2.0	1998	opt B, R	All-sky	21.0 mag	526,280,881	12,757.40
USNO-B1.0	2003	opt B, R, I	All-sky	22.5 mag	1,045,913,669	25,353.64
GSC 2.3.2	2008	opt j, V, F, N	All-sky	22.5 mag	945,592,683	22,921.79
Tycho-2	1989–1993	opt BT, VT	All-sky	16.3 mag	2,539,913	61.57
SDSS DR12	2000–2014	opt u, g, r, i, z	14,555	22.2 mag	932,891,133	64,094.20
DENIS	1996–2001	0.8–2.4 μm	16,700	18.5 mag	355,220,325	21,270.68
2MASS PSC	1997–2001	1.1–2.4 μm	All-sky	17.1 mag	470,992,970	11,417.46
2MASS ESC	1997–2001	1.1–2.4 μm	All-sky	17.1 mag	1,647,599	39.94
WISE	2009–2013	3–22 μm	All-sky	15.6 mag	563,921,584	13,669.83
AKARI IRC	2006–2008	7–26 μm	38,778	50 mJy	870,973	22.46
IRAS PSC	1983	8-120 μm	39,603	400 mJy	245,889	6.21
IRAS FSC	1983	8–120 μm	34,090	400 mJy	173,044	5.08
IRAS SSSC	1983	8–120 μm	39,603	400 mJy	16,740	0.42
AKARI FIS	2006–2008	50–180 μm	40,428	550 mJy	427,071	10.56
Planck	2009–2011	0.35–10 mm	All-sky	183 mJy	33,566	0.81
WMAP	2001–2011	3–14 mm	All-sky	500 mJy	471	0.01
GB6	1986–1987	6 cm	20,320	18 mJy	75,162	3.70
NVSS	1998	21 cm	33,827	2.5 mJy	1,773,484	52.43
FIRST	1999–2015	21 cm	10,000	1 mJy	946,432	94.64
SUMSS	2003–2012	36 cm	8,000	1 mJy	211,050	26.38
WENSS	1998	49/92 cm	9,950	18 mJy	229,420	23.06
7C	2007	198 cm	2,388	40 mJy	43,683	18.29



https://www.degruyter.com/view/journals/astro/25/1/article-p75.xml?language=en

@fedhere

4-V of Big Data in astronomy

V1: Volume

Number of bites

Number of pixels

Number of rows in a data table x number of columns for catalogs

V2: Variety

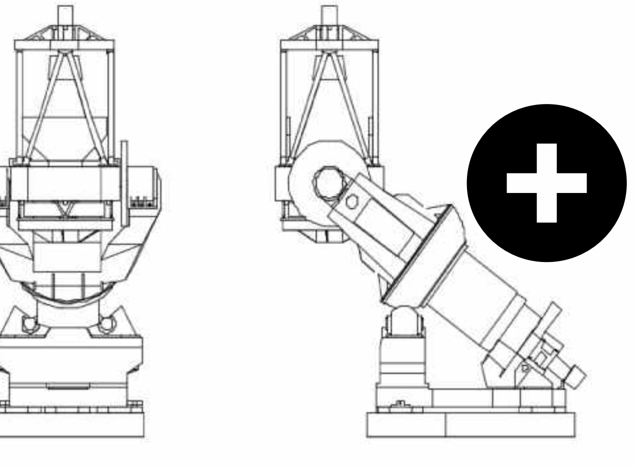
Diverse science return from the same dataset.

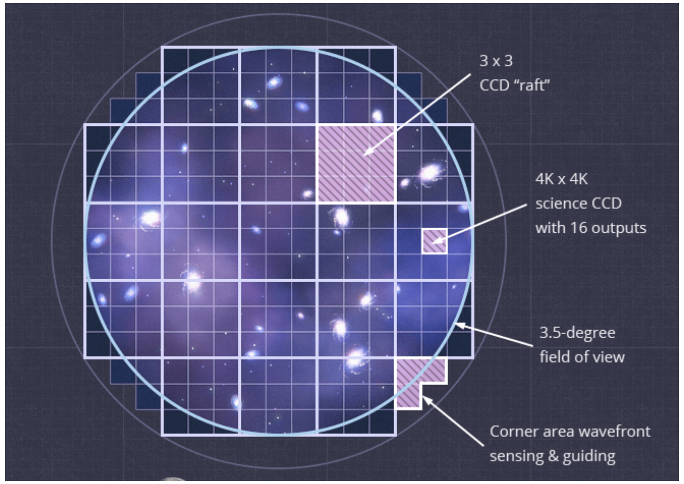
Multiwavelength

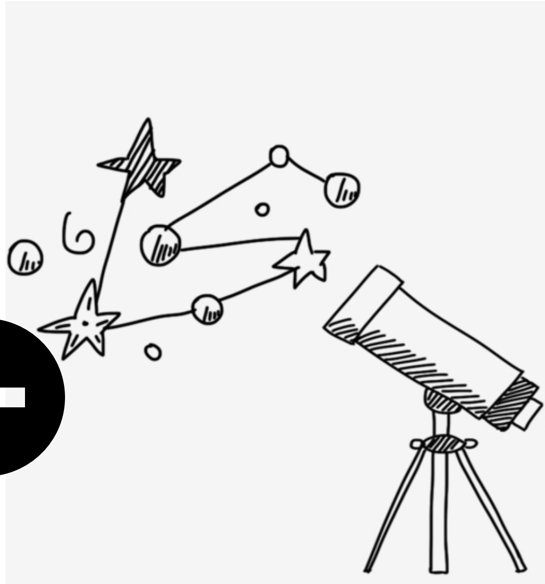
Multimessenger

Images and spectra

ground based
how do the data get big?







filters → variety (complexity)

telescope size → fainter, more distant

FoV → more sky area at once

camera size → more data units

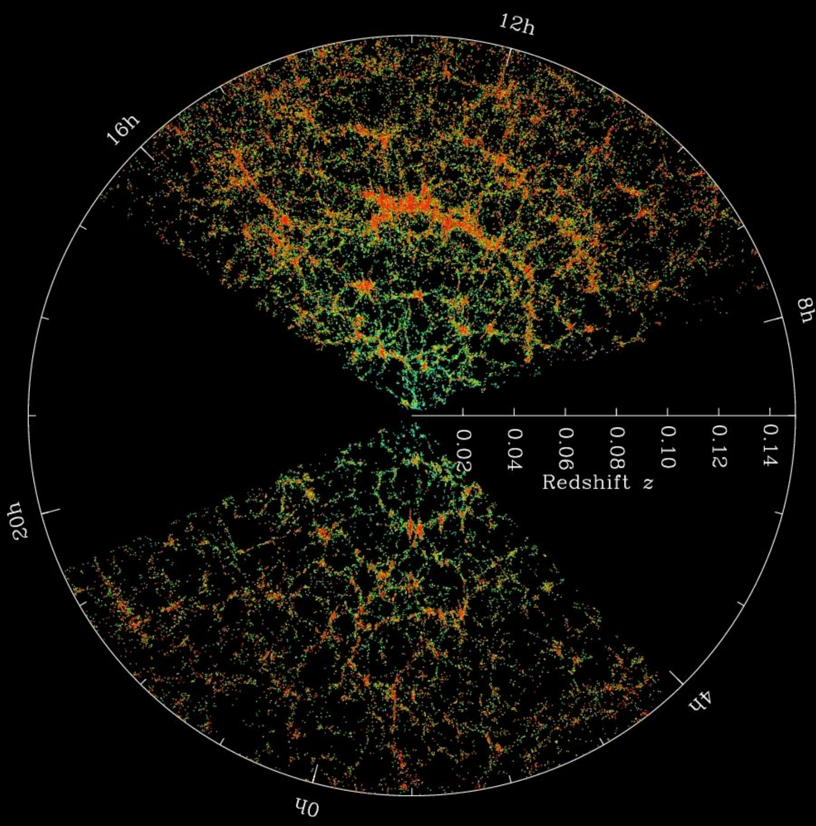
resolution → more objects/details

optical

SDSS

"The Sloan Digital Sky Survey has created the most detailed three-dimensional maps of the Universe ever made, with deep multi-color images of one third of the sky, and spectra for more than three million astronomical objects. Learn and explore all phases and surveys—past, present, and future—of the SDSS."

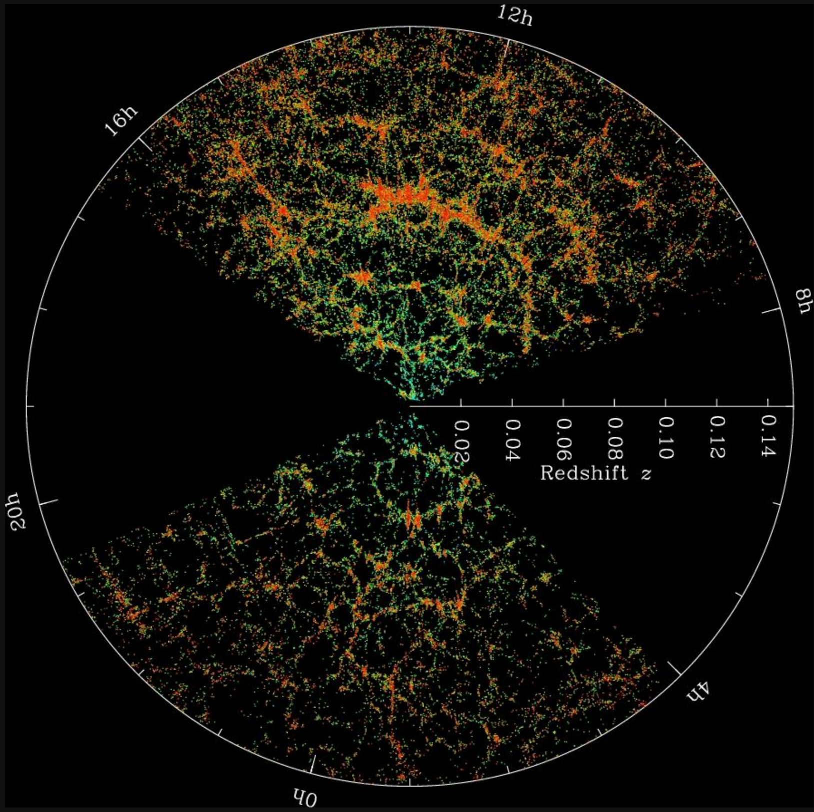
- 5 bands
- 2.5m
- 6 sq degree
- 4Mpix
- 1"/pix



optical: data releases

photometric parameters for 53 million unique objects.

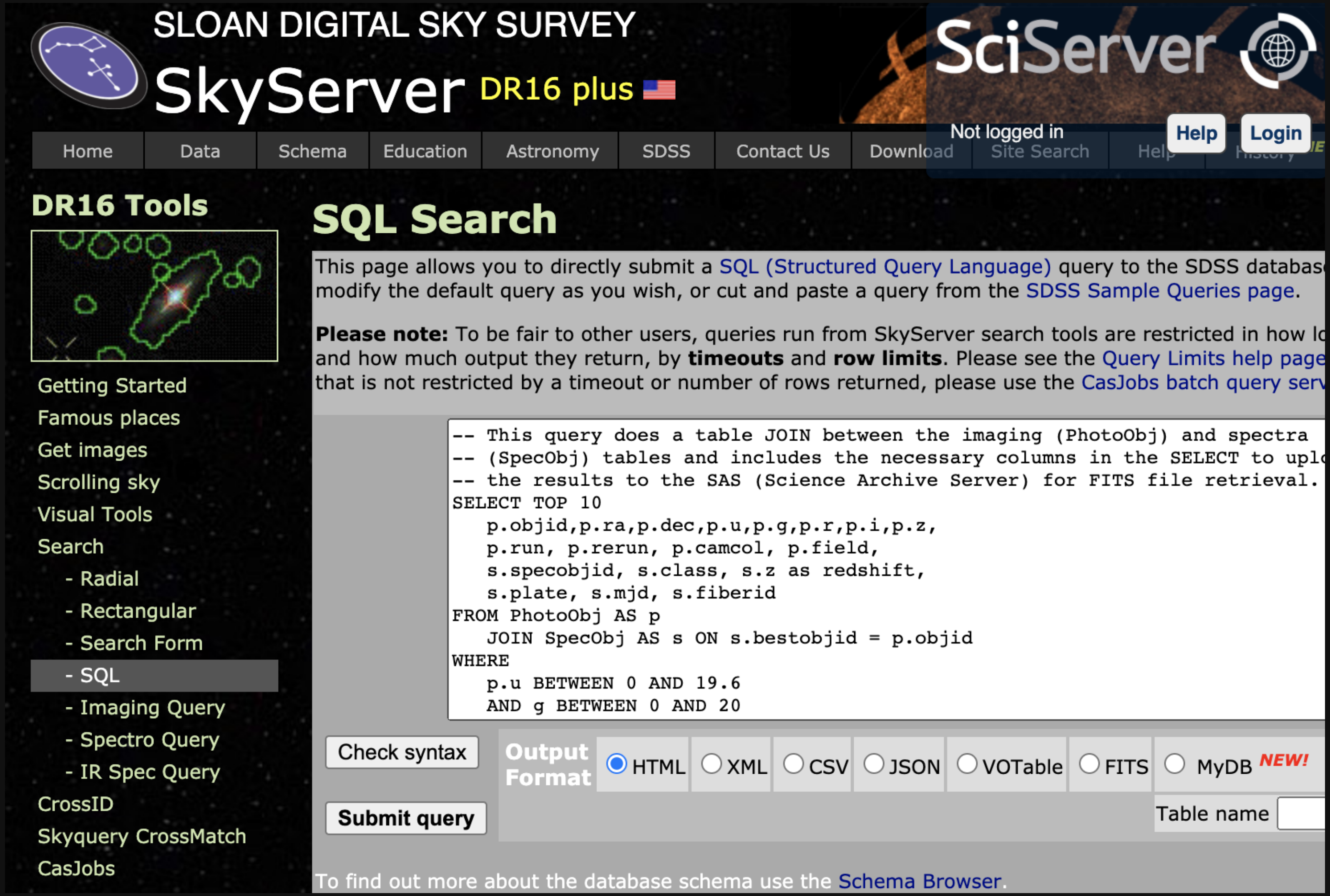
SDSS DR	images	catalog	1D+2D spectra
2003 DR1	2.3Tb	0.5Tb	
2003 DR2	5Tb	0.7Tb	
2004 DR3	6Tb	1.2Tb	
…			
2009 DR7	15.7Tb	18Tb	3.5 TB
…			
2019 DR16			273 TB



The SDSS map of the Universe. Each dot is a galaxy; the color bar shows the local density.

@fedhere

optical: data releases



@fedhere

optical PS1

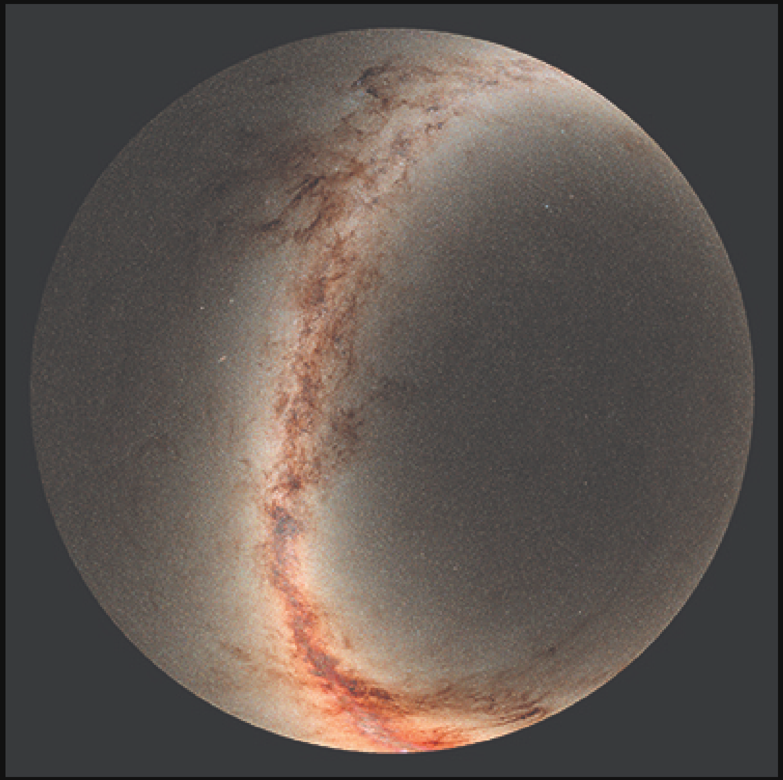
Pan-STARRS
PS1 Science Consortium

2019 DR2 PanSTARRS

1.6 Pb

The amount of imaging data is equivalent to two billion selfies, or 30,000 times the total text content of Wikipedia. The catalog data is 15 times the volume of the Library of Congress.

- 6 bands
- 1.8m
- 7 sq degree
- 1.4Gpix
- 0.26''/pix



@fedhere

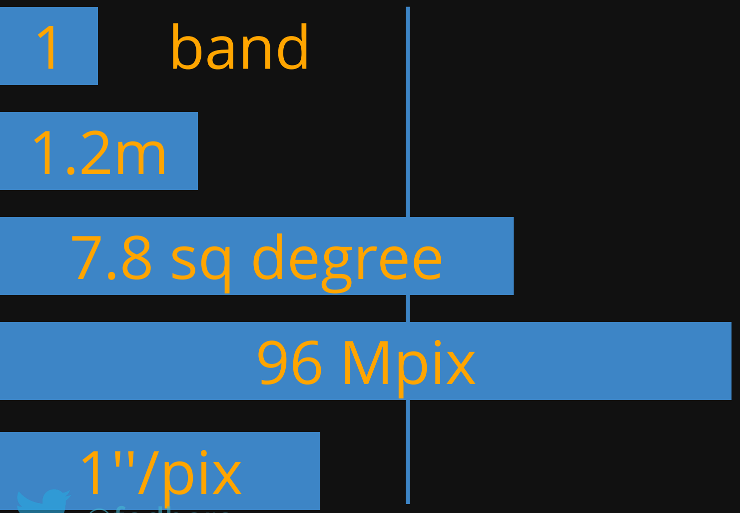
optical: data releases

Number of raw on-sky camera exposures ingested:	**40,000**	**23 TB**
Volume (with ancillary files):		**195 TB**
Number of lightcurves	319	**110 GB**

Pan-STARRS
PS1 Science Consortium

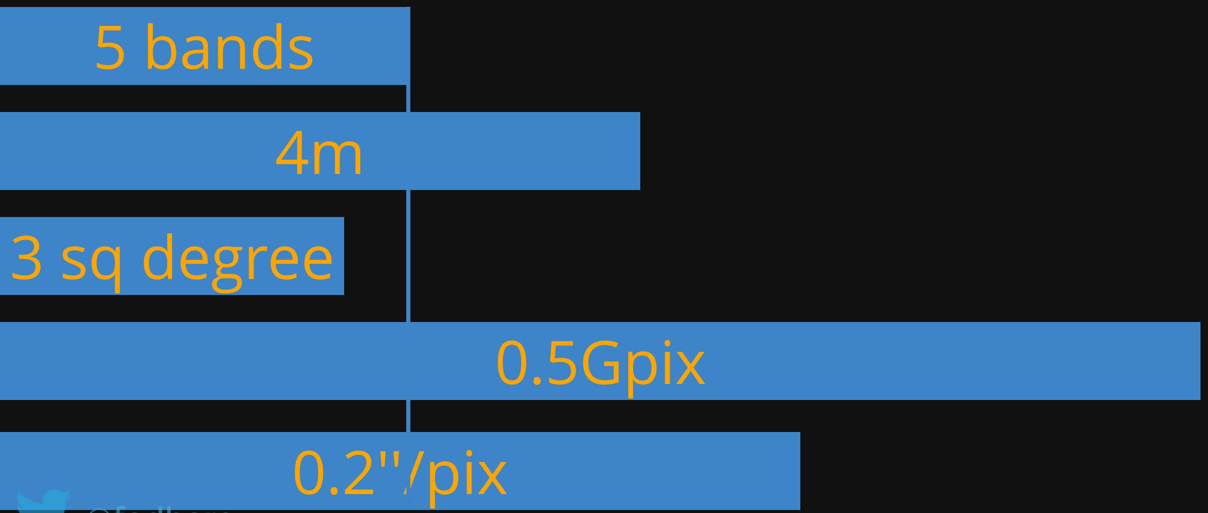
optical: PTF

- Catalina Sky Survey
- Digitized Sky Survey
- VIMOS-VLT Deep Survey
- Palomar Distant Solar System Survey
- Dark Energy Survey
- Pan-STARRS
- DESI Legacy Imaging Surveys

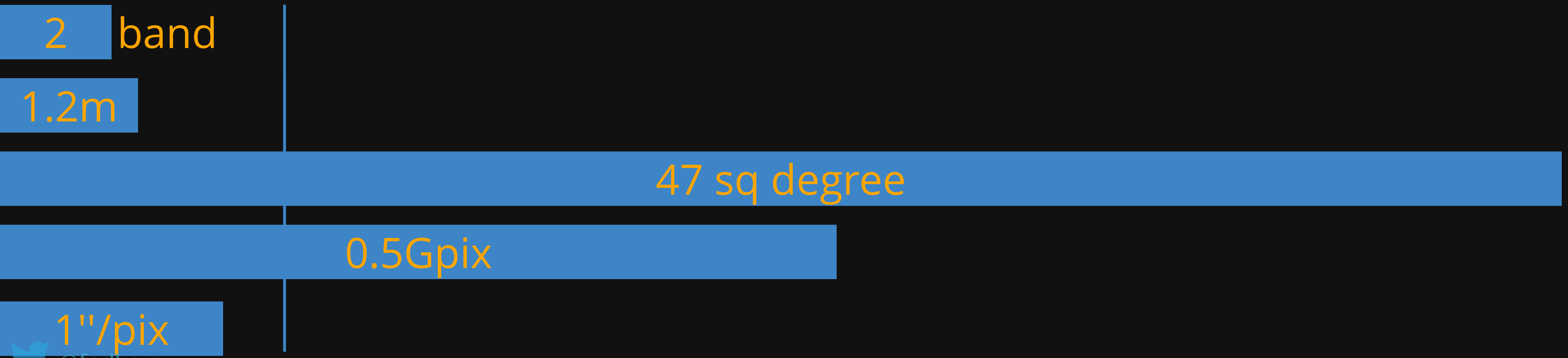


optical: DES

570 MP camera with a 3 deg2 field of view installed at the prime focus of the Blanco 4 m



optical: ZTF



optical: Rubin LSST

6 bands

8m (6.5 effective)

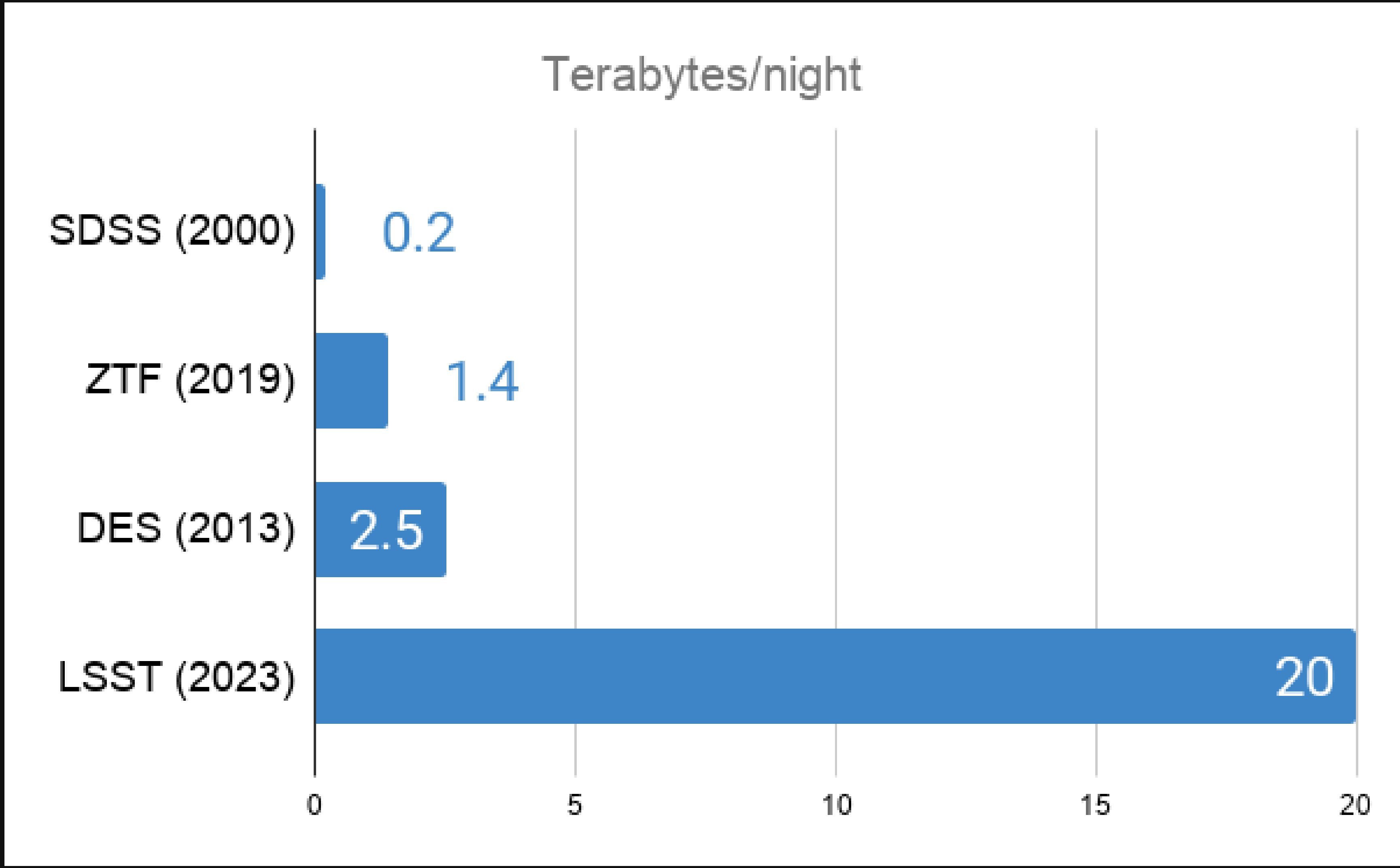
9 sq degree

3.2Gpix

0.2''/pix

@fedhere

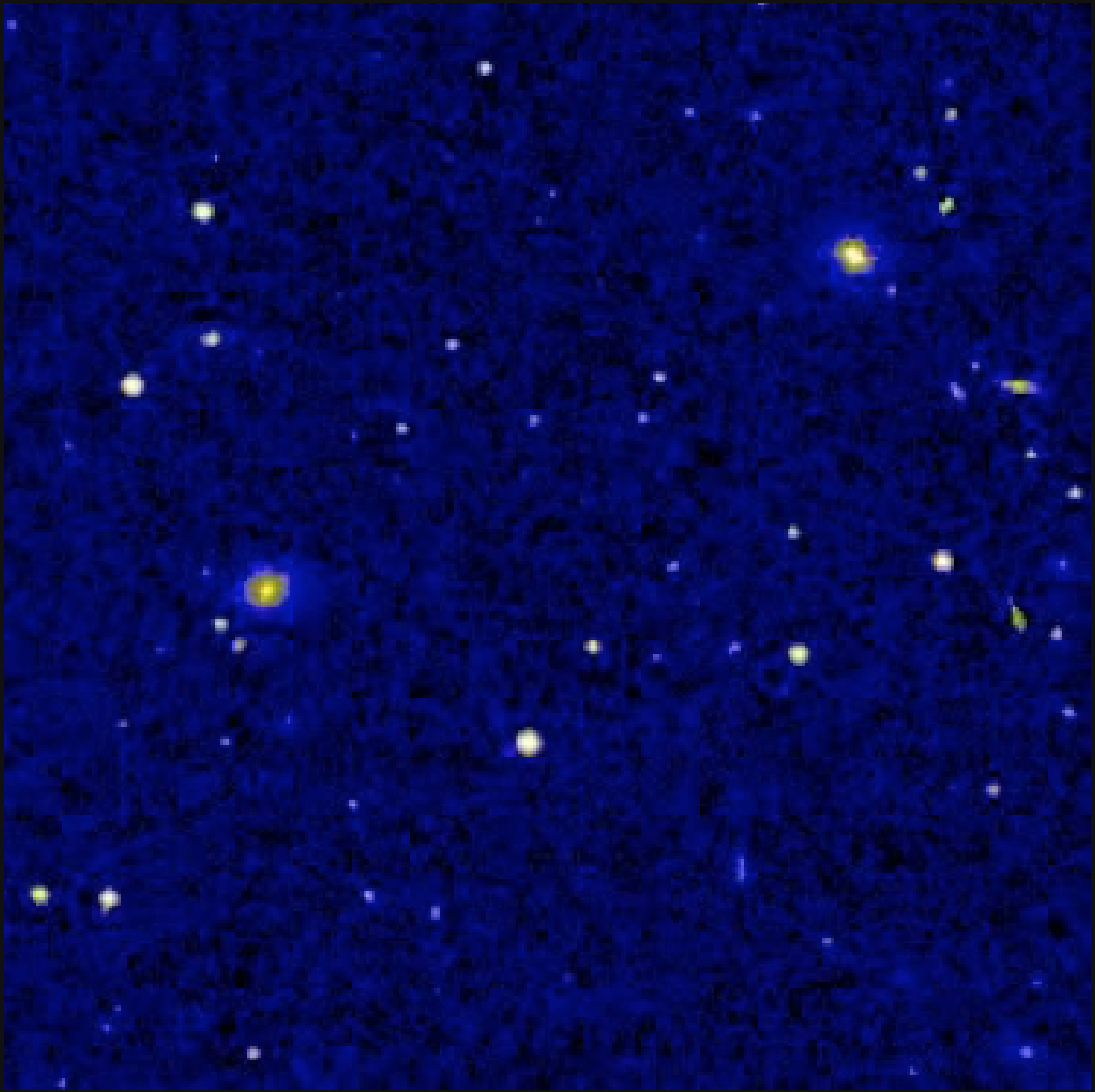
optical: Rubin LSST



@fedhere

DSS:

digitized photographic plates



One quarter the diameter of the moon

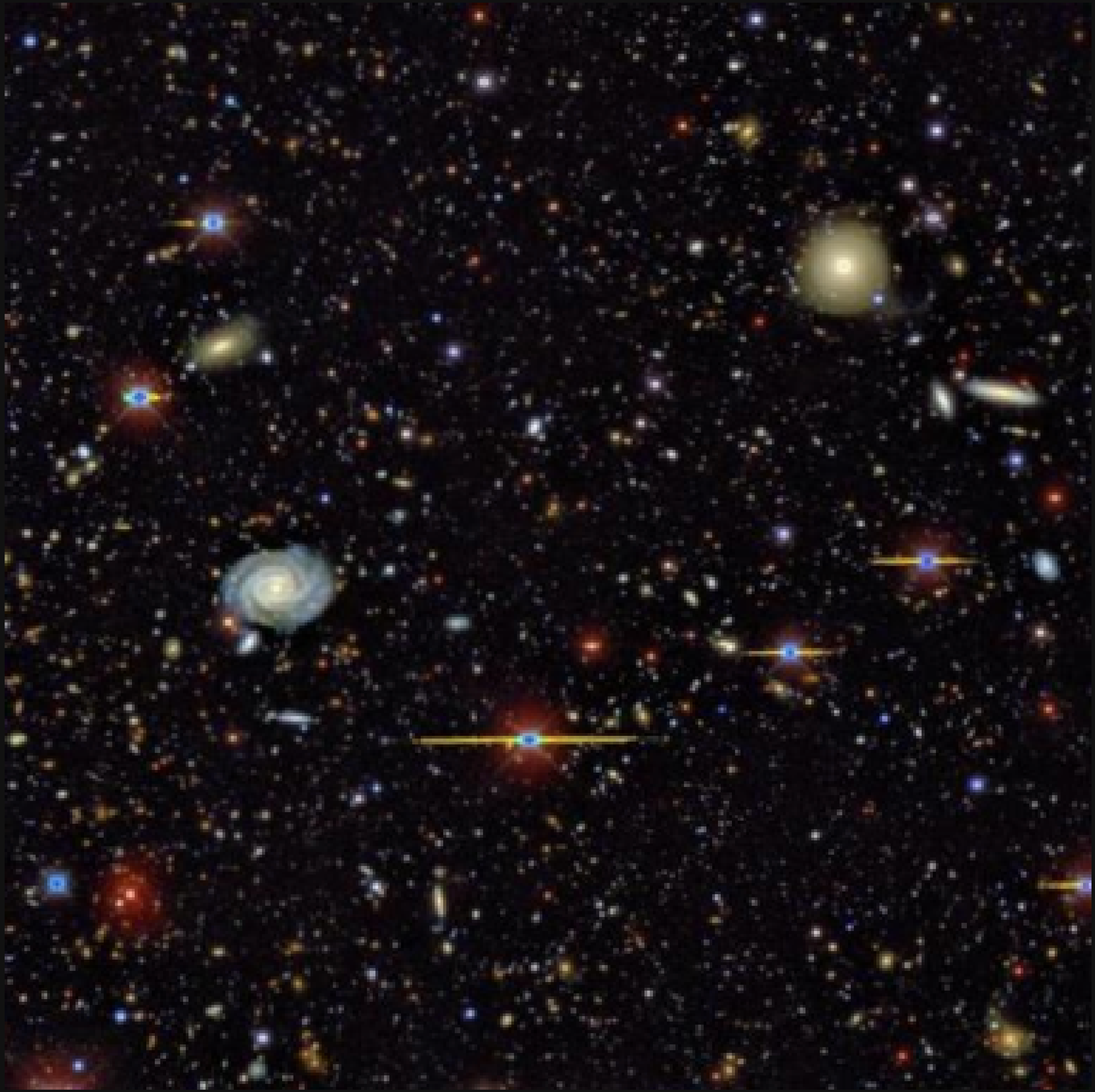
@fedhere

SDSS

One quarter the diameter of the moon

DLS

20 sq deg ultra-deep multi-band sky survey.



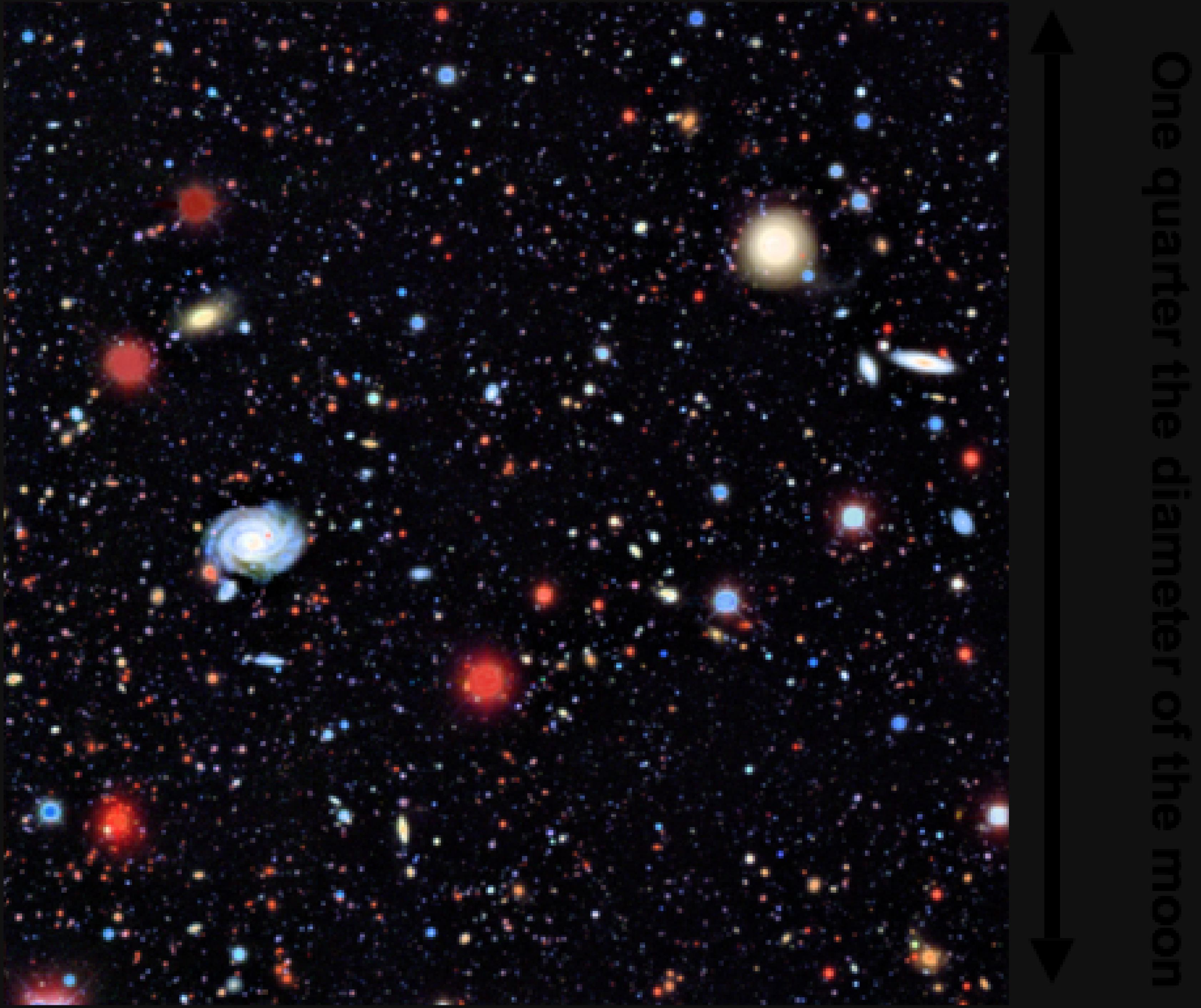
One quarter the diameter of the moon

@fedhere

Rubin

LSST

(simulated)



@fedhere

optical: other surveys

MACHO (Microlensing)

Catalina Sky Survey

Digitized Sky Survey

SNLS (Supernovae)

OGLE (Microlensing)

DLA (weak lensing)

VIMOS-VLT Deep Survey

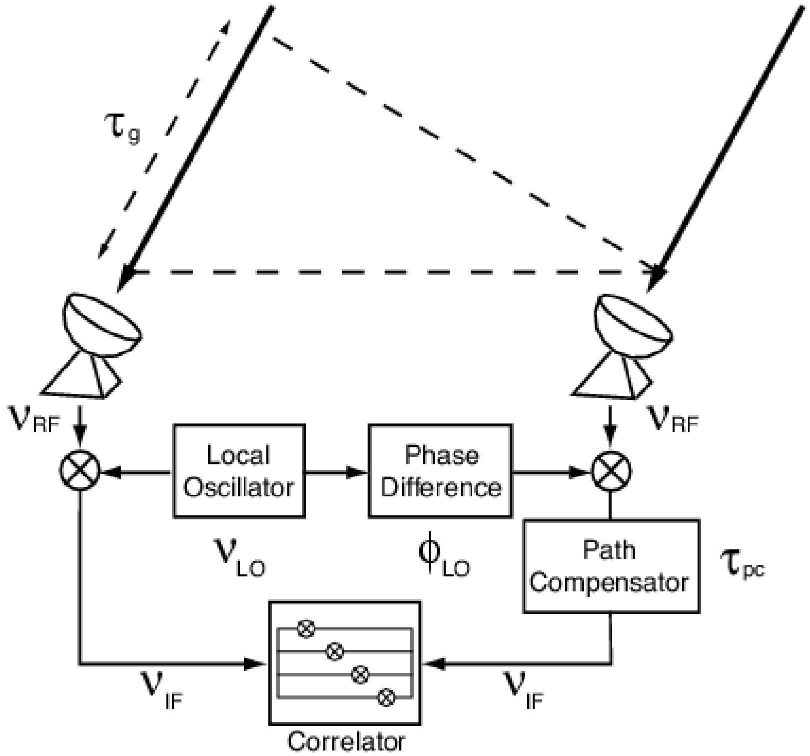
Palomar Distant Solar System Survey

DESI Legacy Imaging Surveys

...

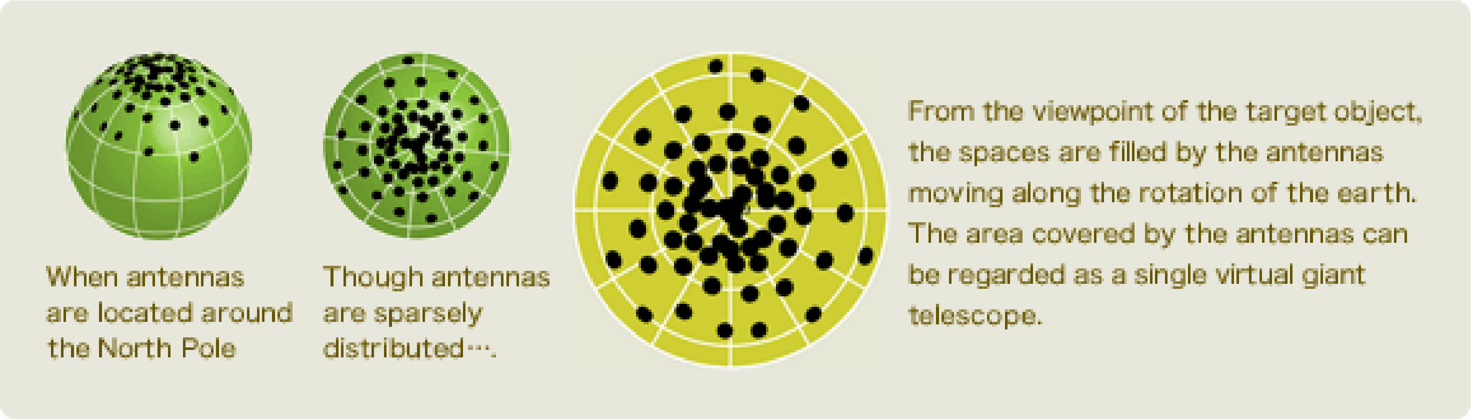
radio

Radio interferometers do not image the sky directly. Instead they measure the amount of power on different angular scales,



interferometry:

Create a virtual telescope by combining multiple antennae.



* The actual ALMA antenna location differs from the figure above. The figure is a conceptual illustration to explain the principle of the "aperture synthesis" technique (interferometric imaging method) in a very simple way.

Cross correlating the signal from each antenna produces coherent interpretable radio images of the sky

The directed use of the Earth's rotation for filling the Fourier plane is known as *Earth rotation aperture synthesis* and was the subject of the 1974 Nobel Prize in Physics

https://archive.briankoberlein.com/2015/10/14/how-does-interferometry-work/

@fedhere

5Pb data

Event Horizon Telescope, 2019
https://iopscience.iop.org/journal/2041-8205/page/Focus_on_EHT

Extreme imaging via physical model inversion: seeing around corners and imaging black holes,

Dr. Katie Bouman

5Pb data

Event Horizon Telescope, 2019
https://iopscience.iop.org/journal/2041-8205/page/Focus_on_EHT

Extreme imaging via physical model inversion: seeing around corners and imaging black holes,

Dr. Katie Bouman

reconstruct images and video from a sparse telescope array distributed around the globe. Additionally, it presents a number of evaluation techniques developed to rigorously evaluate imaging methods in order to establish confidence in reconstructions done with real scientific data.

5Pb data

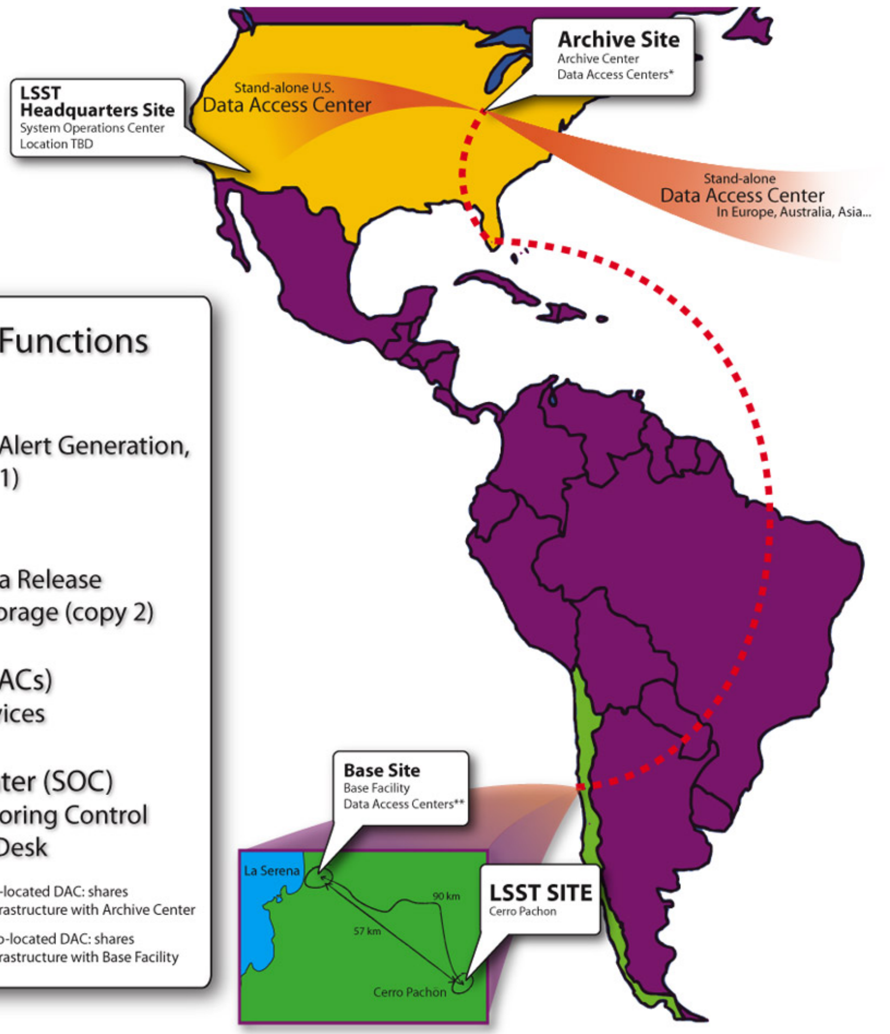
Event Horizon Telescope, 2019

https://iopscience.iop.org/journal/2041-8205/page/Focus_on_EHT

Extreme imaging via physical model inversion: seeing around corners and imaging black holes,

Dr. Katie Bouman

@fedhere



Site Roles and their Functions

- Base Facility
  Real-time Processing and Alert Generation,
  Long-term storage (copy 1)

- Archive Center
  Nightly Reprocessing, Data Release
  Processing, Long-term Storage (copy 2)

- Data Access Centers (DACs)
  Data Access and User Services

- System Operations Center (SOC)
  System Supervisory Monitoring Control
  & End User Support/Help Desk

* Co-located DAC: shares infrastructure with Archive Center

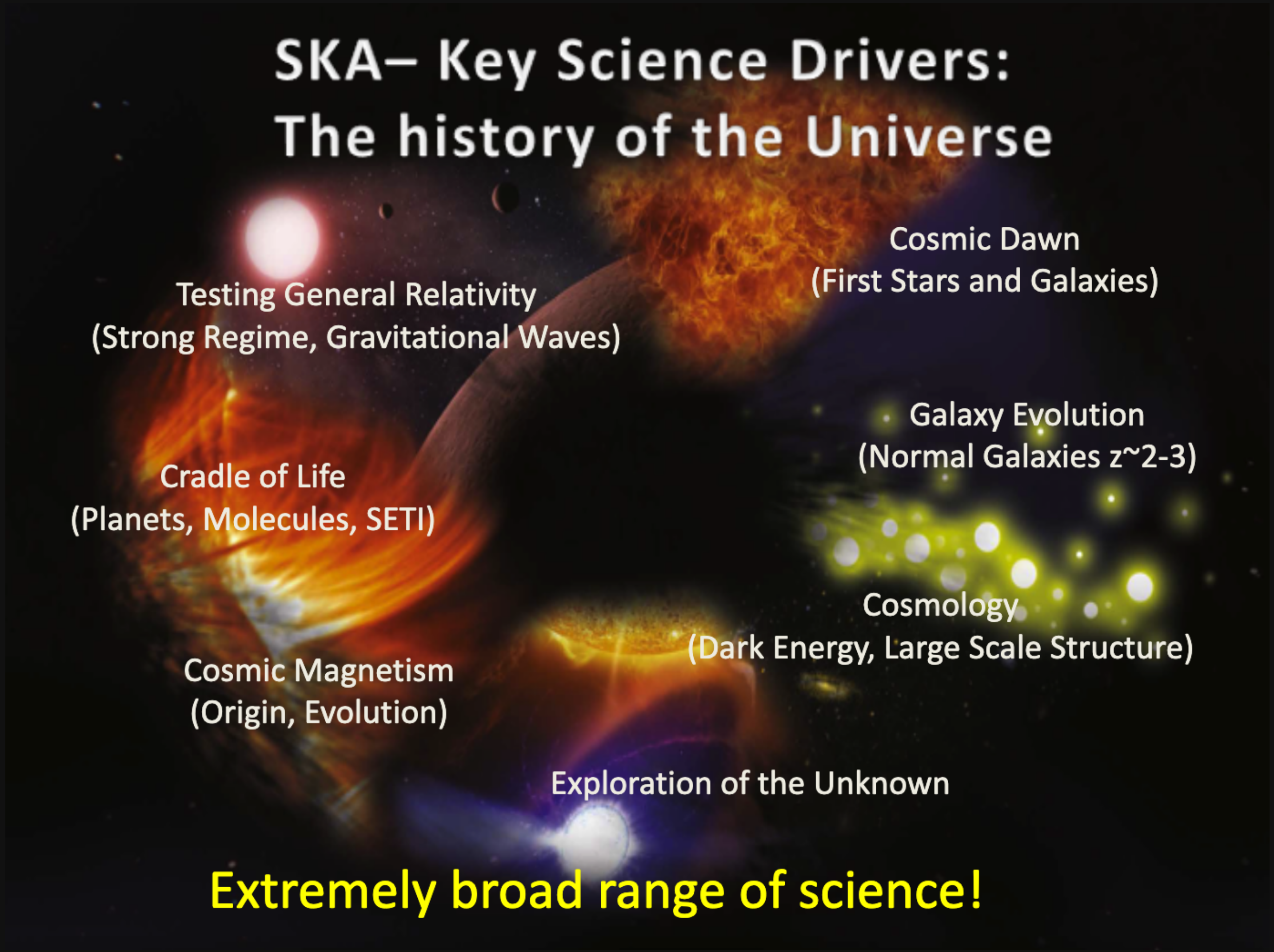
** Co-located DAC: shares infrastructure with Base Facility

... if you thought LSST was Big Data...SKA

@fedhere

Square Kilometer Array

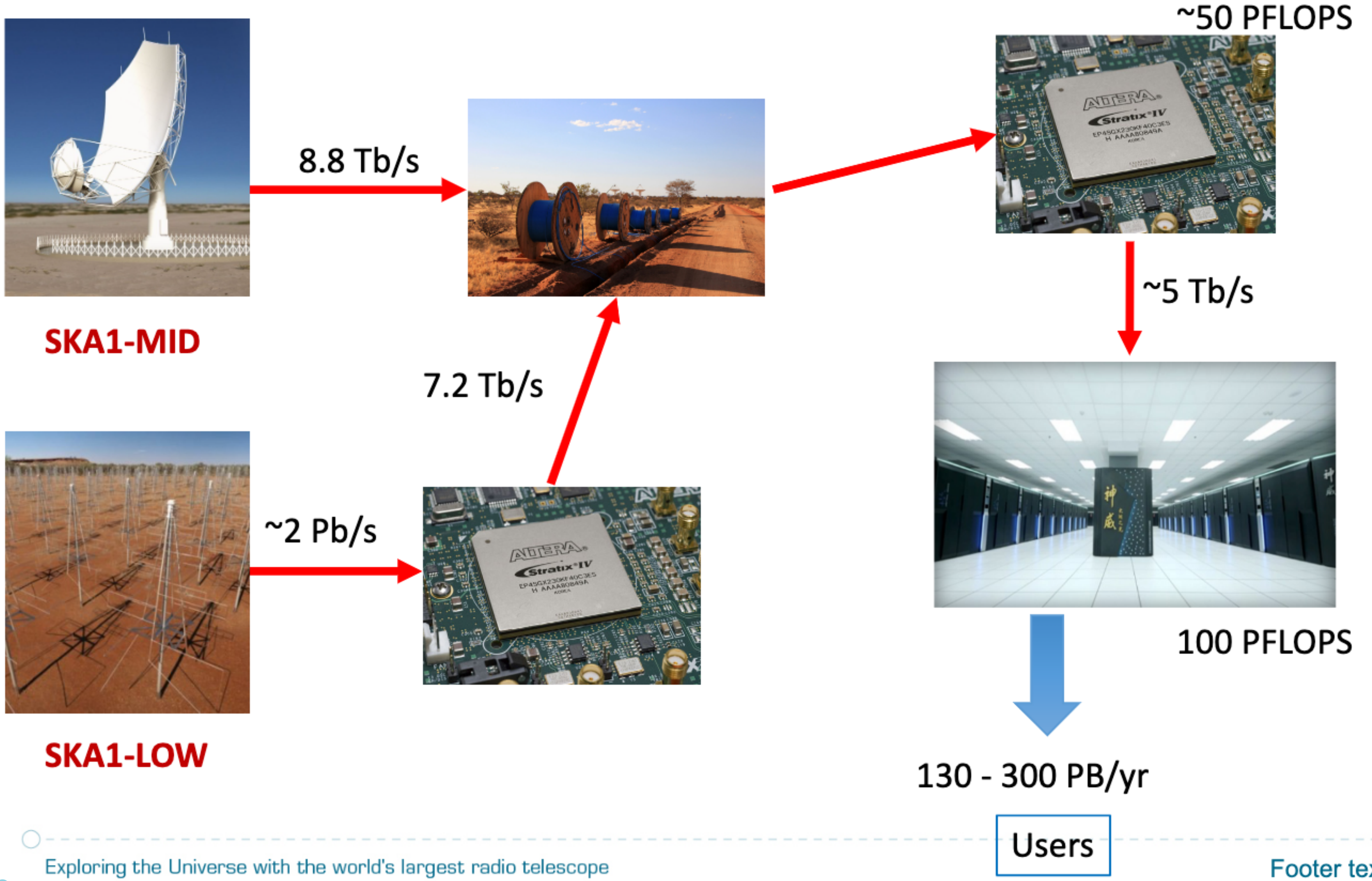
https://www.skatelescope.org/ska-community-briefing-18jan2017/



@fedhere

Square Kilometer Array

Data Flow through the SKA



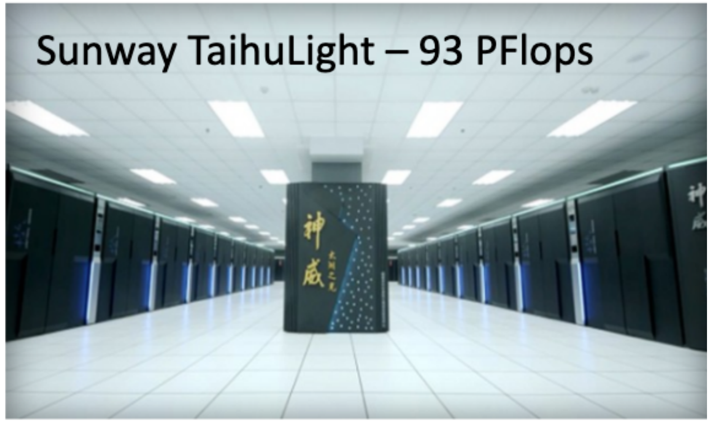
https://www.skatelescope.org/ska-community-briefing-18jan2017/

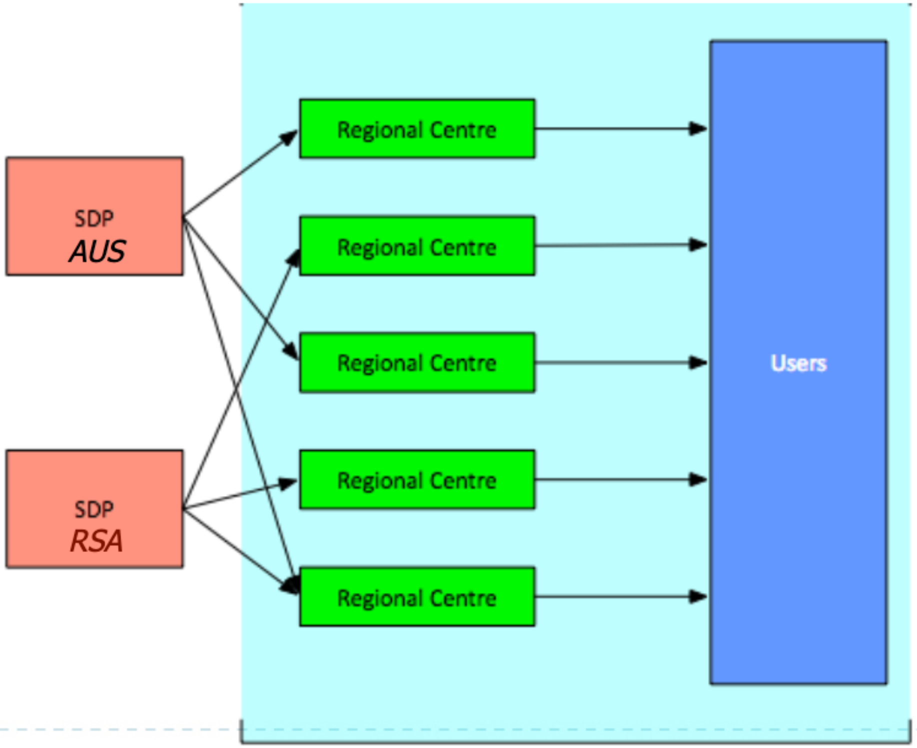
Square Kilometer Array

SKA Regional Centres – outside SKAO scope

- Required
  - capacity for reprocessing data and their analysis
  - storage for a long-term archive
  - local user support

- Intent
  - SKA partner countries planning SKA regional Centres
  - National super-computing centres
  - Provide local support to scientists
  - Development of new techniques, new algorithms
  - Deliver SKA science





https://www.skatelescope.org/ska-community-briefing-18jan2017/

Data rate: 0.5–1 TB s−1.

Data from the individual antennas of the SKA1-MID, or the individual stations of the SKA1-LOW, are transported to the central signal processing facility, where the data from each pair of antennas/stations are correlated to produce the visibility data

300PB/telescope/year = 8.5EB

Typical images will have spatial axes with 2e15 x 2e15 pixels up to 2e16 frequency channels. This dimensionality results in petabyte scale volumes for individual data products,

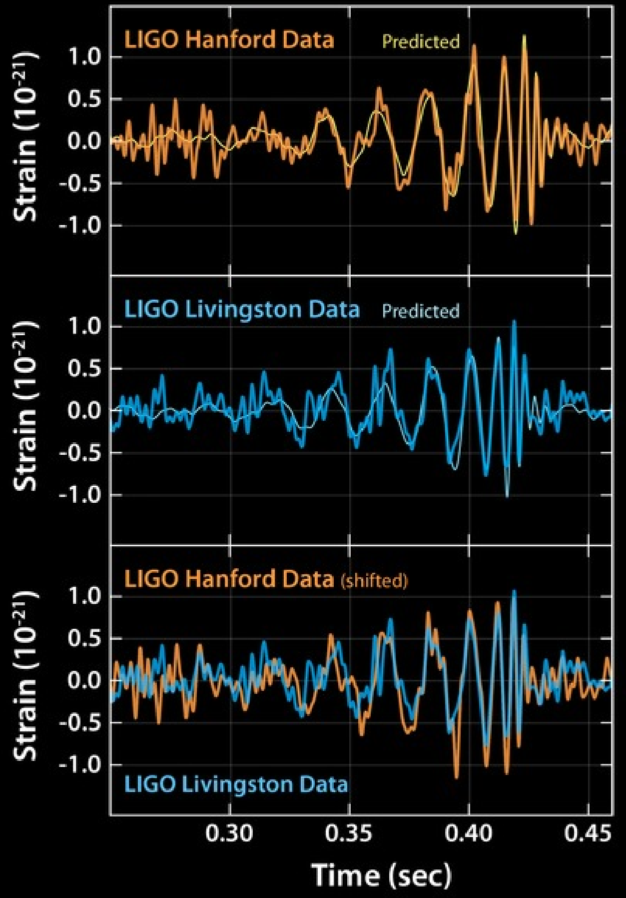
Performing multiple Fourier transforms on data with image sizes as large as 2e15 or even 2e16 pixels on a side is computationally prohibitive.

[...] In spite of these challenges, the computing for the SKA over the coming decade is certainly achievable, and it is likely that its implementation will be instrumental in driving the next generation of global e-infrastructure.

The computers must be able to make decisions on objects of interest, and remove data which is of no scientific benefit, such as radio interference from things like mobile phones or similar devices, even with the remote locations which will host the SKA.

https://royalsocietypublishing.org/doi/10.1098/rsta.2019.0060

Gravitational Wave and MMA



Computation and Data Collection

Computers are required both to run the LIGO instruments and to process the data that it collects.

When it is in 'observing' mode, LIGO generates terabytes (*1000's of gigabytes*) of data *every day*. All of this information must be transferred to a network of supercomputers for storage and archiving. Such supercomputers are located at each of the observatories, at Caltech, at MIT, and at various other institutions. Once the data is secured, scientists can use customized computer programs to scour the data for gravitational waves.

The amount of data LIGO collects is as incomprehensively large as gravitational wave signals are small. LIGO's archive already holds the equivalent over 1-million DVDs of data and will add the equivalent of about 178-thousand DVDs each year to its archive. In actual numerical terms, the data archive at Caltech holds over 4.5 Petabytes (Pb) of data, and will grow at a rate of about 0.8 Pb (800 terabytes) per year. What's a petabyte? If you wanted to count up to a petabyte by counting one byte per second, it would take you 35.7 million ***years*** to reach one petabyte!

Storing information is one thing; processing it is another. Processing and analyzing all of LIGO's data requires a vast computing infrastructure. For LIGO's first observing run in 2015, the LIGO Lab will provide 35 MSU (**million service units**) worth of computing cycles/time. This is equivalent to running a modern 4-core laptop computer for 1,000 years! The amount of computing time is expected to grow by a factor of 10 to around 400 MSU by the time LIGO has completed its third observing run.

If you'd like to learn even more about all of LIGO's remarkable technology and engineering, visit **Look Deeper**.

@fedhere

Gravitational Wave and MMA

To search for binaries with components more massive than m_{min}=0.2$M_{\odot}$ while losing no more than 10% of events the initial LIGO interferometers will require 1x10e11flops for data analysis to keep up with data acquisition.

Advanced LIGO will require 7.8e11 flops, and VIRGO 4.8e12 flops.

If the templates are stored rather than generated as needed, storage requirements range

1.5e11 (TAMA) - 6.2e14 (VIRGO)

https://journals.aps.org/prd/abstract/10.1103/PhysRevD.60.022002

Owen+1999

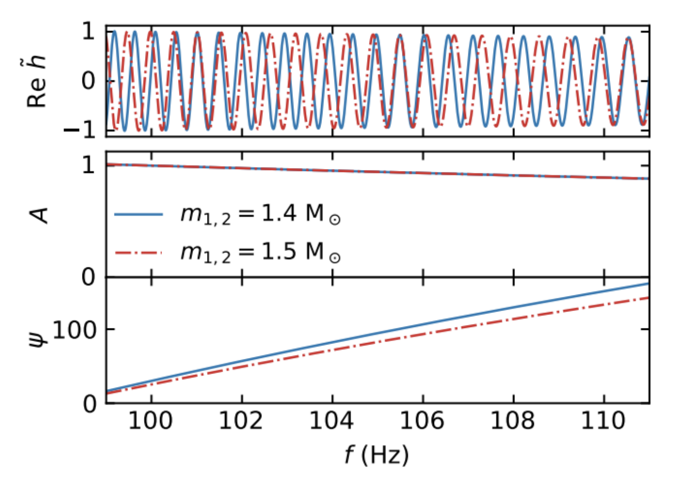


FIG. 1: An example of two waveforms that look very different to each other in the frequency domain (top panel) but have very similar amplitude and phase profiles (middle and bottom panels). The amplitude and phase profiles can be well captured by a low-dimensional linear space spanned by a few basis functions. Waveform amplitudes are shown in arbitrary units.

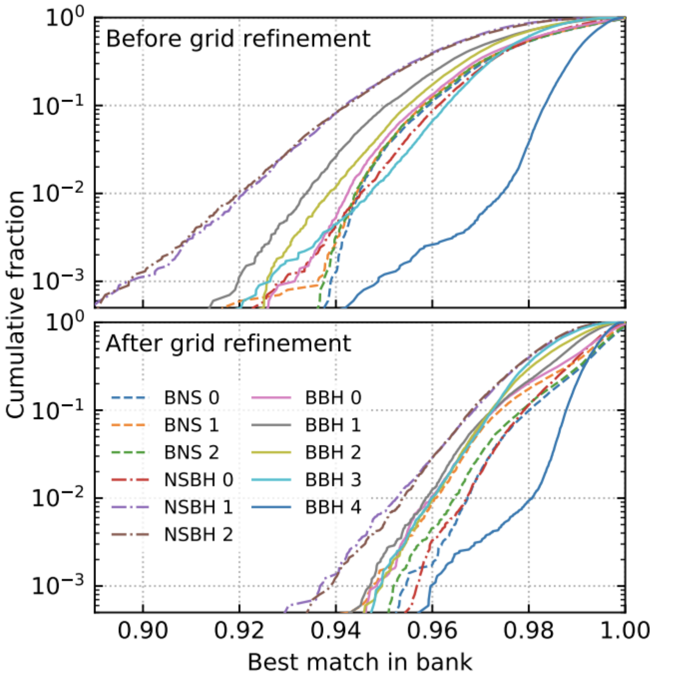


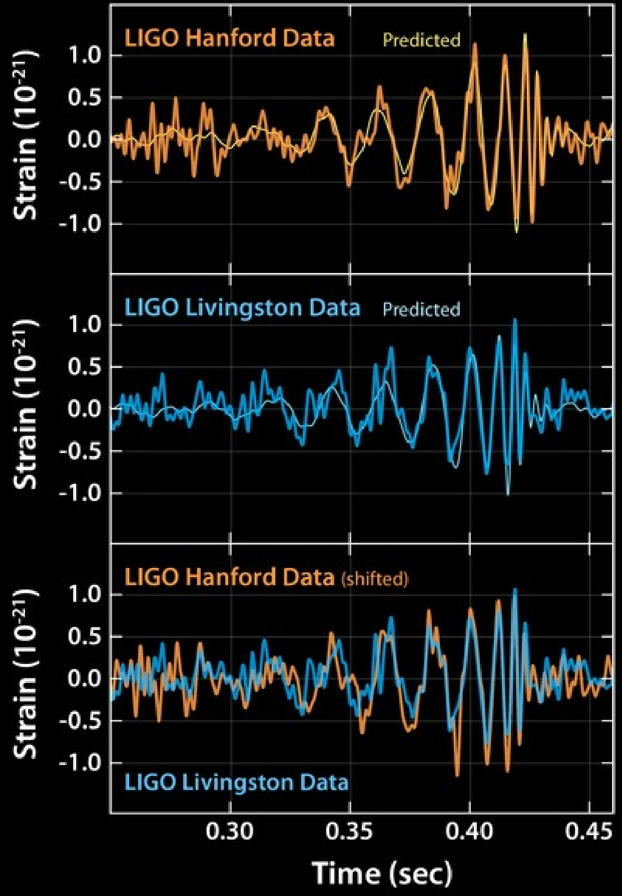
FIG. 5: Effectualness of our template banks, tested on random waveforms drawn from a distribution uniform in individual masses and aligned spins. The vertical axis shows the fraction of the random trials that do not achieve a given match in the bank.

https://arxiv.org/pdf/1904.01683.pdf

Roulet+2019

@fedhere

Gravitational Wave and MMA

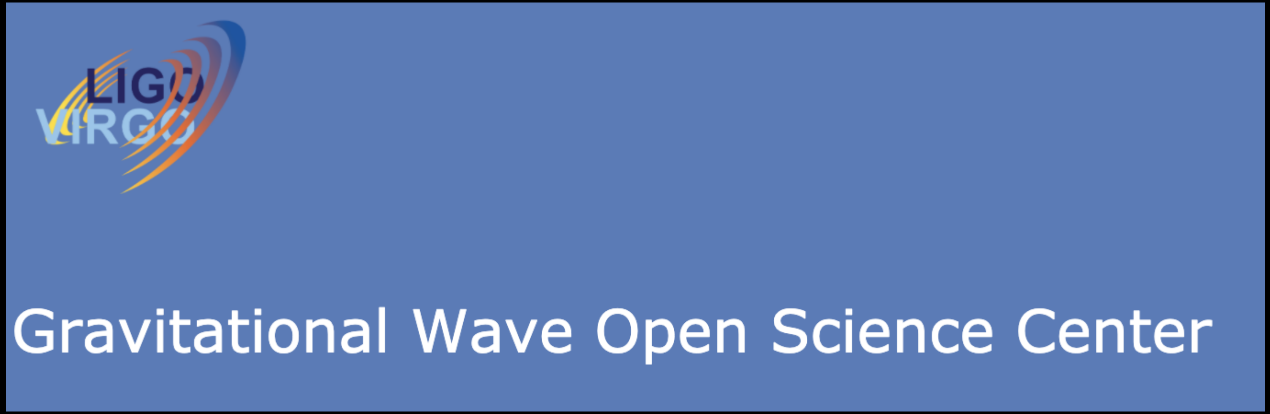


Historical data releases

All data from Advanced LIGO can be found at GWOSC. GWOSC also hosts data releases from initial LIGO (S5 and S6 observing runs; constraints on GRB051103, and data from a blind injection study).

Some additional historical data (which are not formally LSC data products) are listed below:

Nov 18, 2015	Localization of Short Duration Gravitational-wave Transients with the Early Advanced LIGO and Virgo Detectors
Apr 23, 2014	The First Two Years of Electromagnetic Follow-Up with Advanced LIGO and Virgo



https://www.gw-openscience.org/eventapi/html/O3_Discovery_Papers/GW190521/v2/

@fedhere

3/6

space-based astronomy BD problems

4-V of Big Data in astronomy

V1: Volume

Number of bites

Number of pixels

Number of rows in a data table x number of columns for catalogs

V2: Variety

Diverse science return from the same dataset.

Multiwavelength

Multimessenger

Images and spectra

V3: Velocity

real time analysis, edge computing, data transfer

space based

Hubble deep field:

When in 1993 Dr. Robert "Bob" Williams became the new director of the Space Telescope Science Institute

The new director realized that a facility so powerful could maximize its scientific return only if its unique data were made available to the whole world community immediately after acquisition, a dramatic paradigm shift in the way astronomical observatories had been operated until then. In a bold move, in December 1995 Dr. Williams used his "Director Discretionary Time", the fraction of a telescope's observing time that the director can use for special projects, for an unprecedented experiment: stare on the same spot on the sky for ten consecutive days to image the faintest and most distant astronomical sources ever unveiled by humans: the Hubble Deep Field was born. He made the images available to anyone who had the curiosity to look and study them and the findings turned out to be nothing short of transformative.

Bob decided that this data was so overwhelmingly powerful, in terms of what it was telling us about the universe, that it was worth it for the community to be able to get their hands on the data immediately. And so the original deep field team processed the data, found the objects in it, and then catalogued each of them, so that every object in the deep field had a description in terms of size, distance, color, brightness and so forth. And that catalogue was available to researchers from the very start---it started a whole new model, where the archive does all the work.

https://www.umass.edu/events/talk-robert-williams-probing

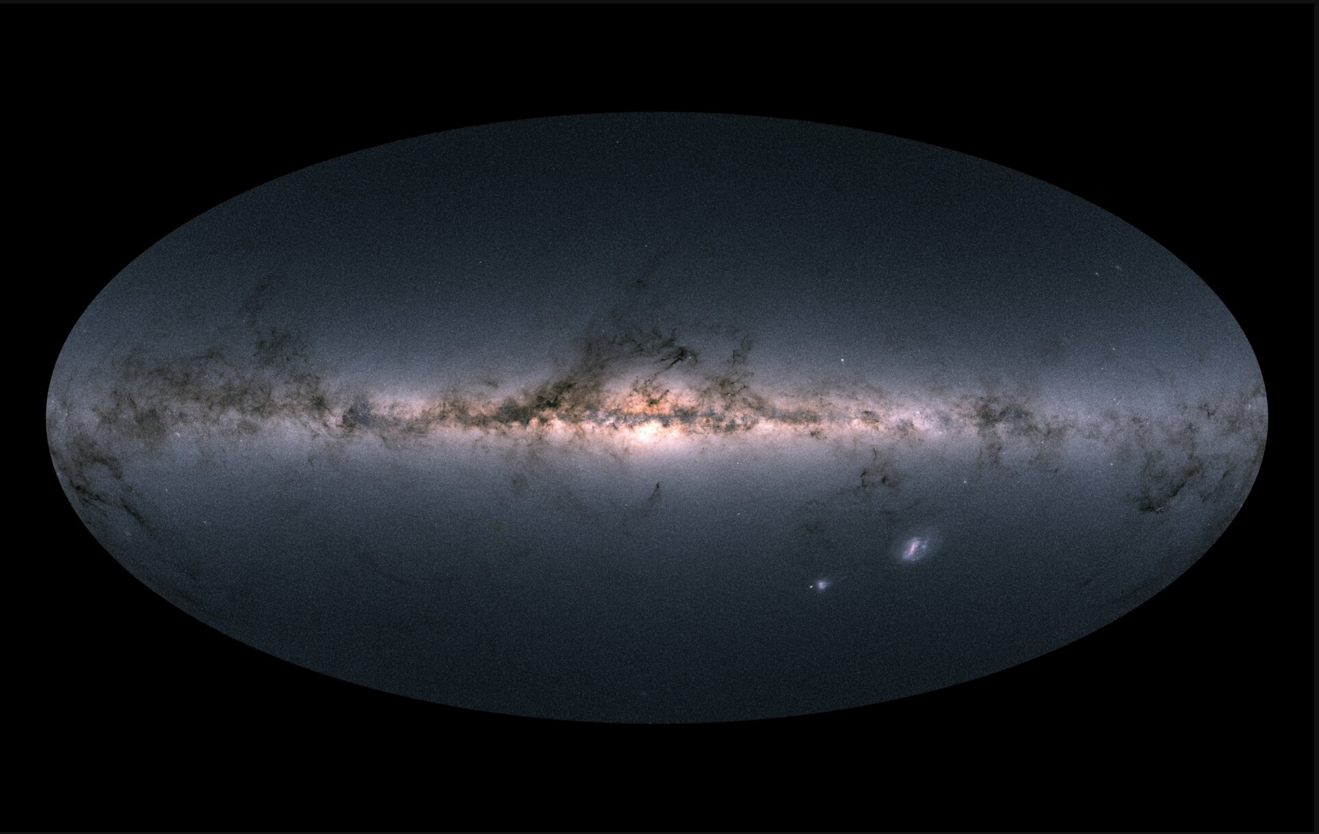
@fedhere

γ- and X-ray

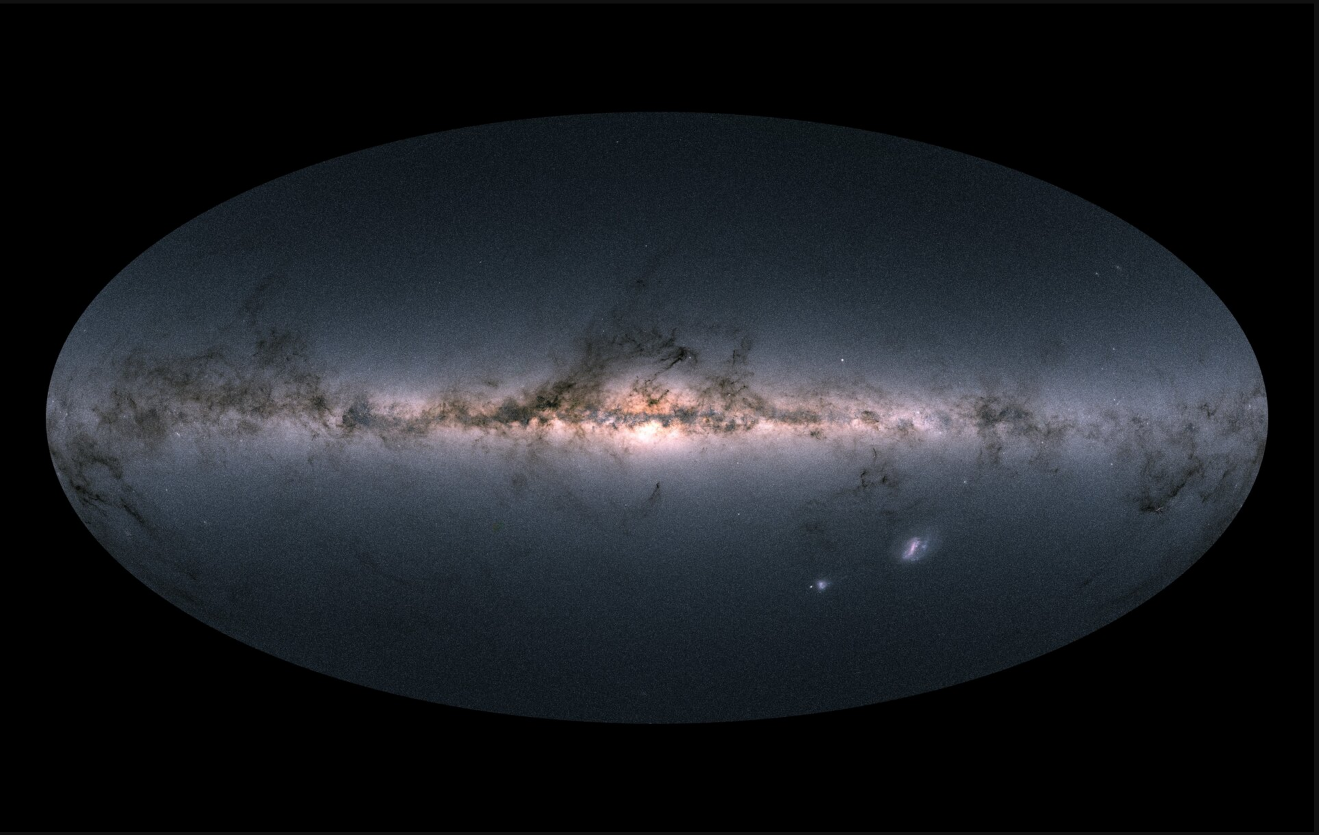
Table 1. Main data for the most important all-sky and large-area astronomical surveys providing multi-wavelength photometric data. Catalogues are given in the order of increasing wavelengths.

Survey, catalogue	Years	Spectral range	Sky area (deg^2)	Sensitivity (mag/mJy)	Number of sources	Density (obj/deg^2)
Fermi-GLAST	2008–2014	10 MeV–100 GeV	All-sky		3033	0.07
CGRO	1991–1999	20 keV–30 GeV	All-sky		1300	0.03
INTEGRAL	2002–2014	15 keV–10 MeV	All-sky		1126	0.03
ROSAT BSC	1990–1999	0.07–2.4 keV	All-sky		18,806	0.46
ROSAT FSC	1990–1999	0.07–2.4 keV	All-sky		105,924	2.57
GALEX AIS	2003–2012	1344–2831Å	21,435	20.8 mag	65,266,291	3044.85

Gaia: big data and telemetry do not get along...



Gaia: big data and telemetry do not get along...



Besides, the whole content of all CCDs cannot be downloaded because of the telemetry bottleneck: the content of the Astrometric focal plane CCDs only would already amount to about 6000 Mbps. while the actual bandwidth reaches a few Mbps only.

Let us assume that about 55 stars/s will be observed on the average in each of the CCDs. *Detecting* then *windowing* each star with say 6×12 pixels of 16 bits each, imply that 10 Mbps would be needed, already gaining a factor 600. Then, because one-dimensional measurements allow for achieving the astrometric performances, a factor 12 can be gained by a 1×12 binning[1]. Besides, this *sampling* greatly improves the signal to noise ratio. Finally, a factor > 2 can perhaps be gained with a lossless compression scheme Portell et al. (2005).

http://articles.adsabs.harvard.edu/pdf/2005ESASP.576..335A

4/6

crowd-sourcing

4-V of Big Data in astronomy

V1: Volume

Number of bites

Number of pixels

Number of rows in a data table x number of columns for catalogs

V2: Variety

Diverse science return from the same dataset.

Multiwavelength

Multimessenger

Images and spectra

V3: Velocity

real time analysis, edge computing, data transfer

V4: Veracity

This V will refer to both data quality and availability (added in 2012)

crowd sourcing

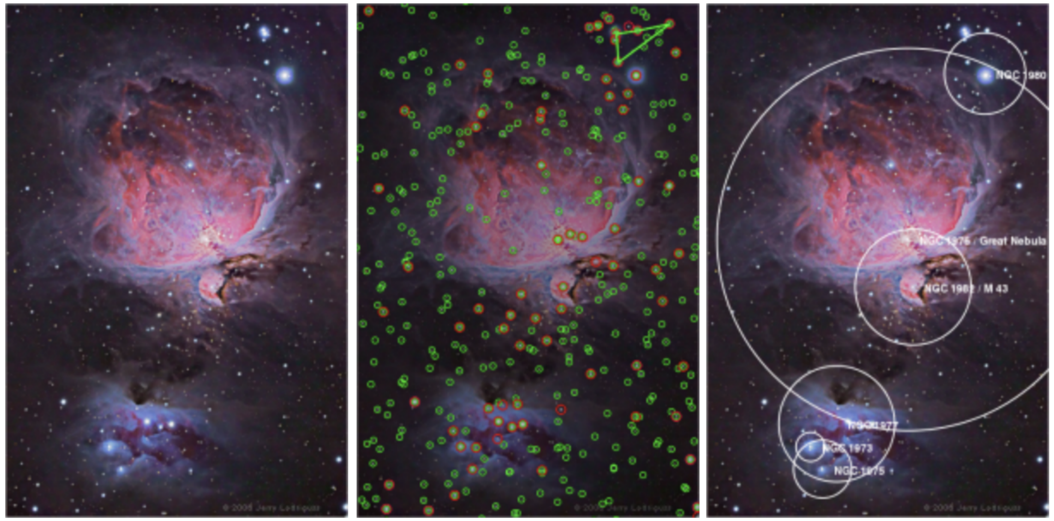
Astrometry.net

home | project summary | people | gallery | news | related links | bibliography | data | use | download | forum

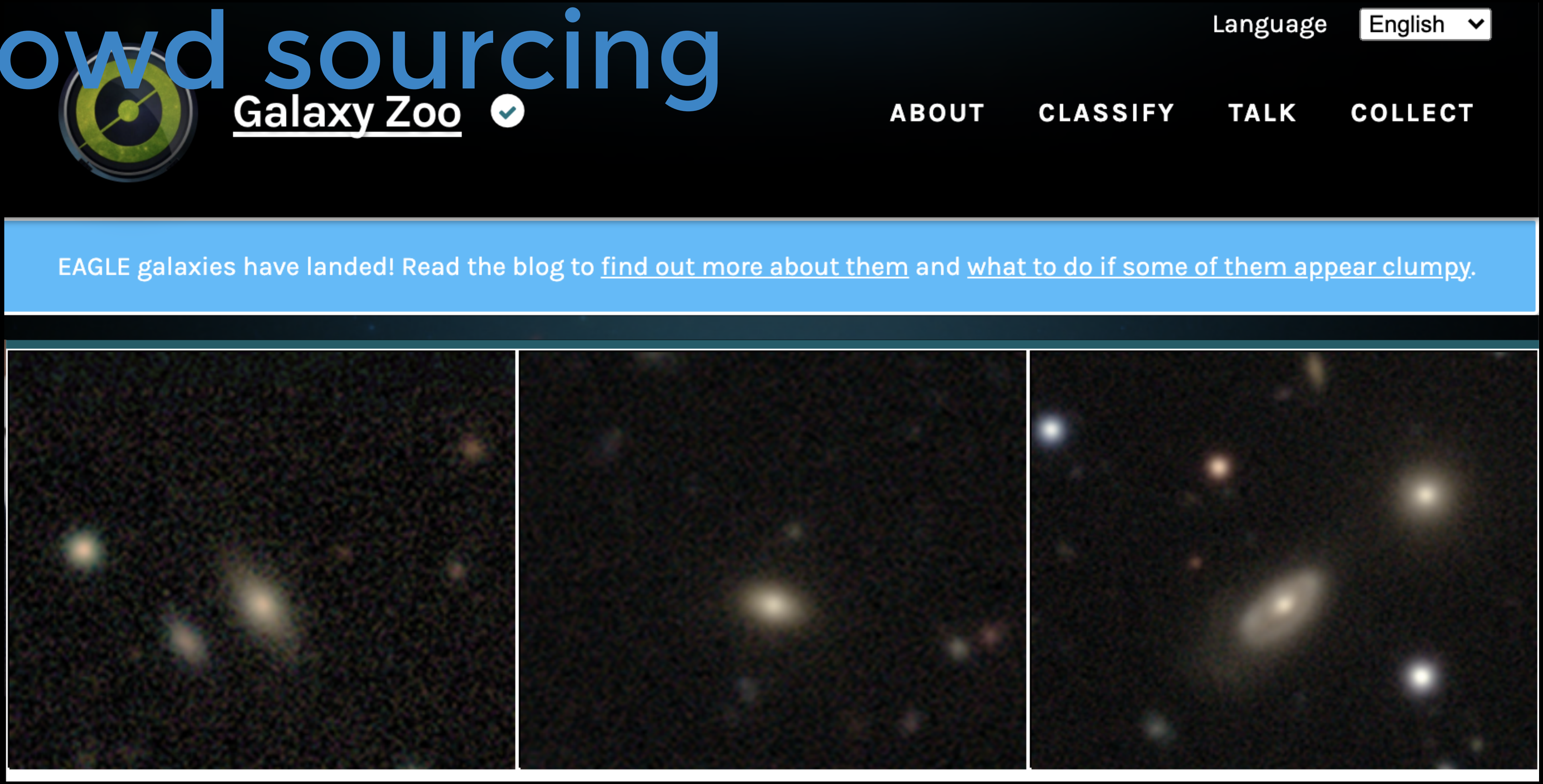
Gallery of Solved Images

In the images below, the red circles are stars our algorithm automatically detects in the image, and the green circles are stars from our master index which appear in the query image. Nebulae, constellations and other objects can be automatically overlayed on the image after it has been solved.

A shot of the Great Nebula, by Jerry Lodriguss (c.2006), from astropix.com



crowd sourcing



LSST: 2x3.2 GPix images/minute for 10 years:

scaling the Galaxy Zoo the entire population of the Earth would be insufficient to study the full dataset.

crowd sourcing

An all-sky search for continuous wave signals in the frequency range 50-1190 Hz + with frequency derivative range from

-20e-10 Hz/s to 1.1e-10 Hz/s collected in 2005-2007 during the fifth LIGO science run.

Hundreds of thousands of host machines, that contributed a total of approximately 25000 CPU

What is Einstein@Home?

Einstein@Home uses your computer's idle time to search for weak astrophysical signals from spinning neutron stars (often called pulsars) using data from the LIGO gravitational-wave detectors, the Arecibo radio telescope, and the Fermi gamma-ray satellite.

Learn more

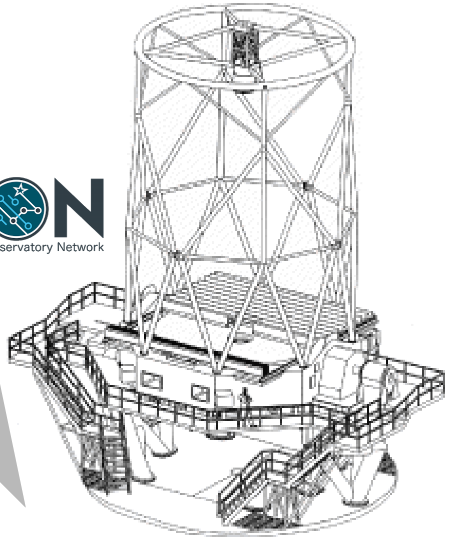
JOIN NOW

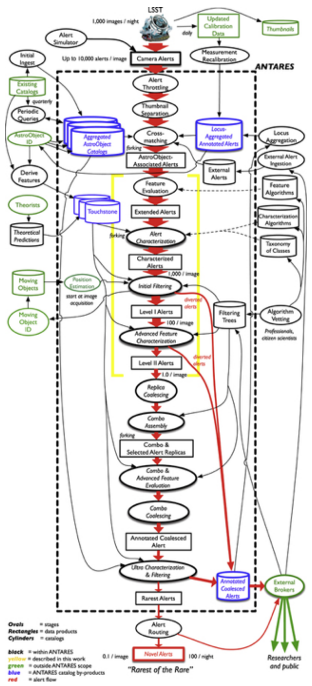
5/6

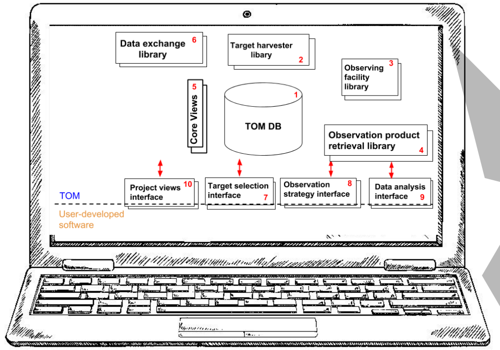
time-domain astronomy BD problems

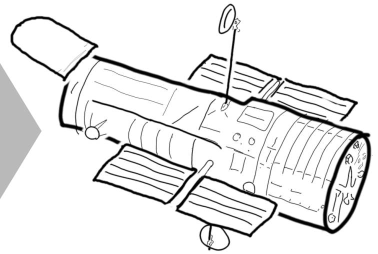
Big data and time domain astrophysics

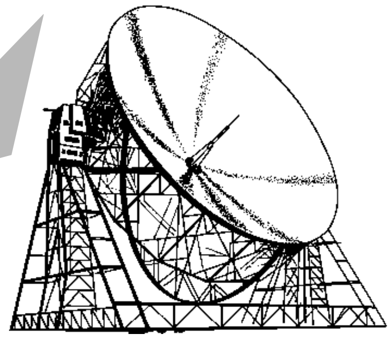
AEON
Astronomical Event Observatory Network











Rubin
Observatory

LSST
Legacy Survey of Space and Time

ALeRCE
Automatic Learning for the Rapid Classification of Events

ANTARES
beta

Lasair

TOM
TOM TOOLKIT
Las Cumbres Observatory LCO

the astronomy discovery chain

@fedhere

Big data and time domain astrophysics
Discovery

ZTF

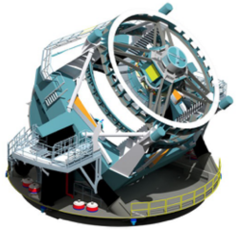
Rubin Observatory

LSST
Legacy Survey of Space and Time

Data Management System Overview

Raw Data: 20TB/night

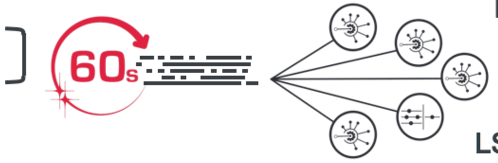
Sequential 30s images that cover the entire visible sky every few days.



Prompt Data Products

Alerts: up to 10M/night

via nightly alert streams



Community Brokers

LSST Alert Filtering Service

@fedhere

https://www.youtube.com/embed/ZdvEGPt4s0Y?enablejsapi=1

Big data and time domain astrophysics

Discovery

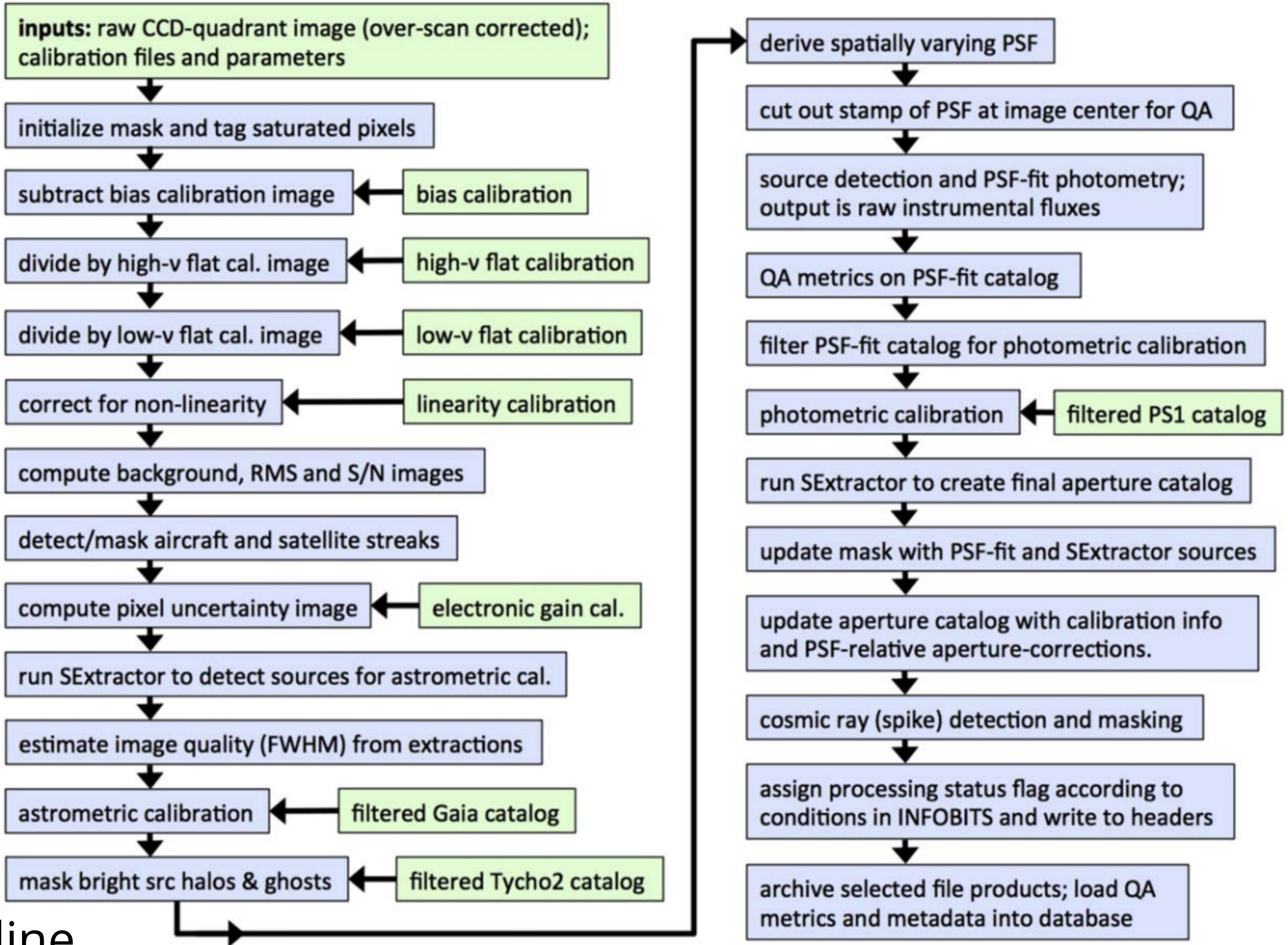


Figure 3. Processing flow in the instrumental calibration pipeline. This represents the first phase of the real-time pipeline.

ZTF realtime pipeline

Rubin Observatory

Legacy Survey of Space and Time

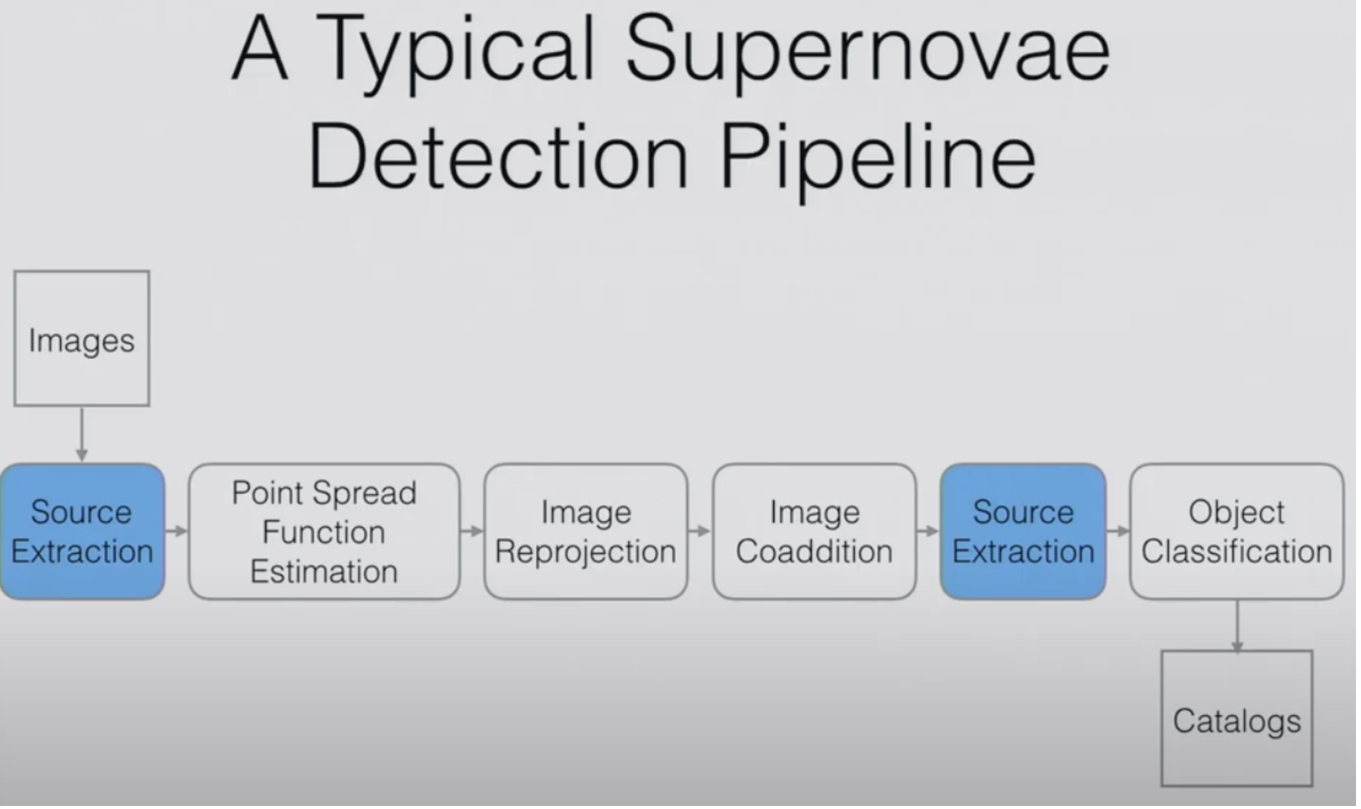
@fedhere

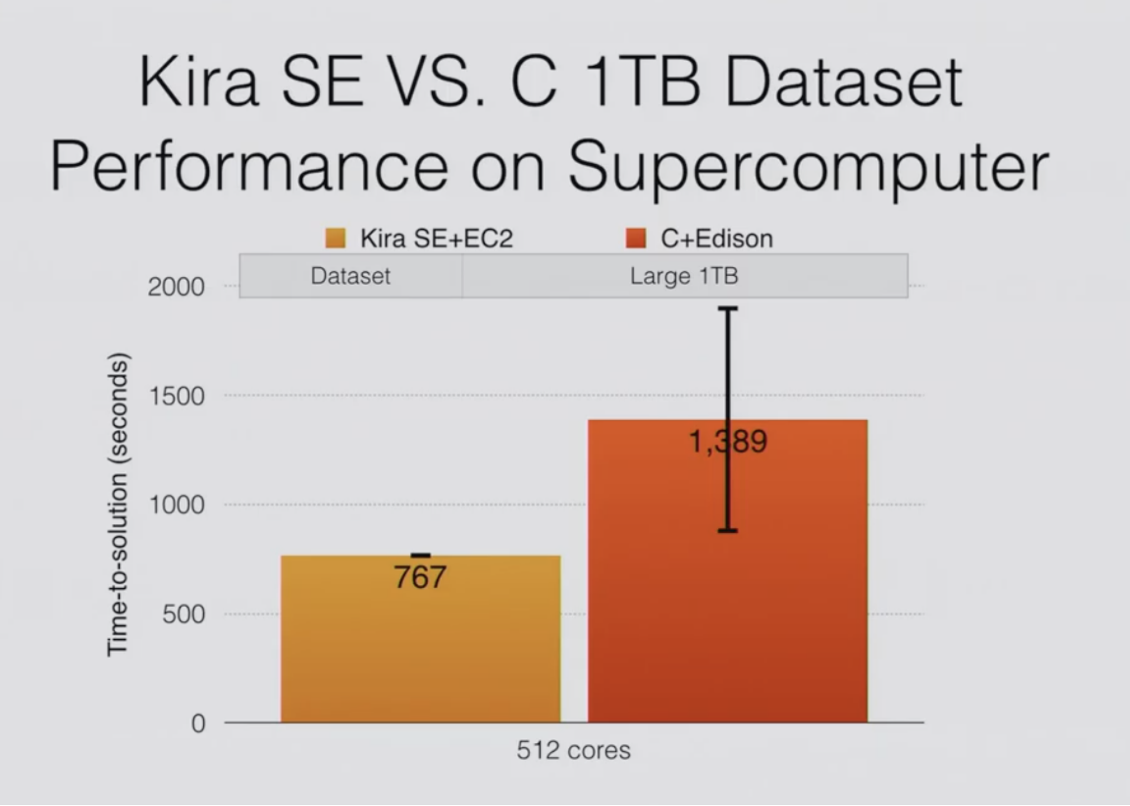
Demonstrated computational gain by using Big Data platforms

Kira: Processing Astronomy Imagery Using Big Data Technology

BIDS Spring 2016 Data Science Faire | May 3, 2016 UC Berkeley Zhao Zhang

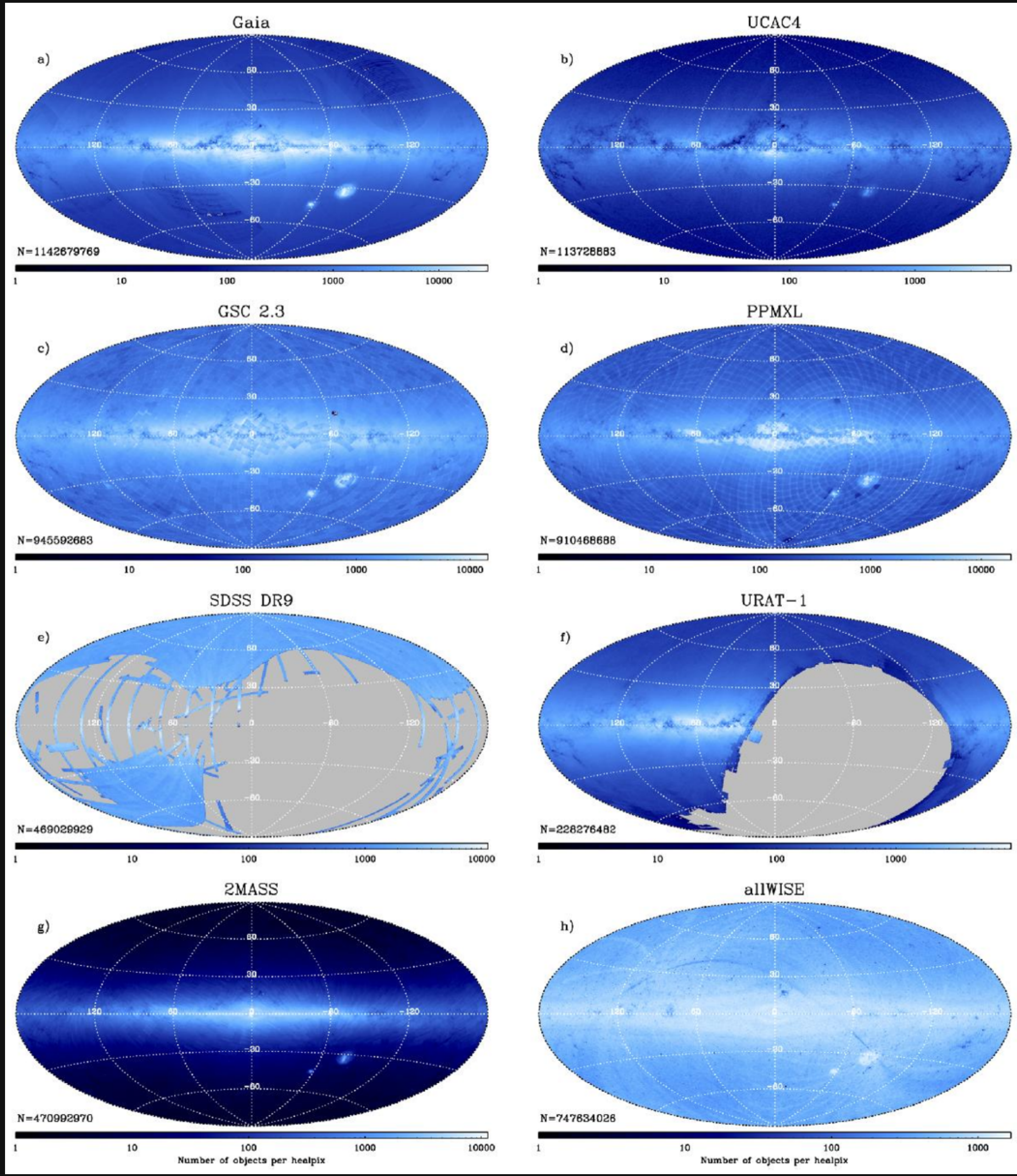
Particularly because the data (IO) intensive applications are





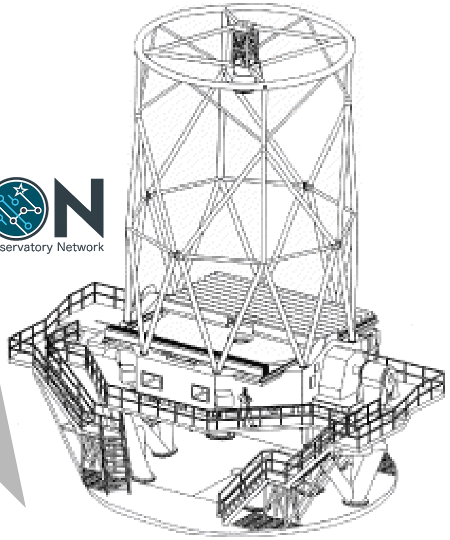
Cross-matching

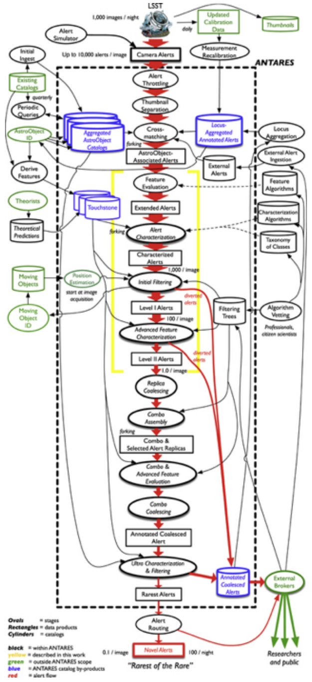
Cross matching catalogs is vital for physical inference. The complexity of the data scales ~with the square of the data sources

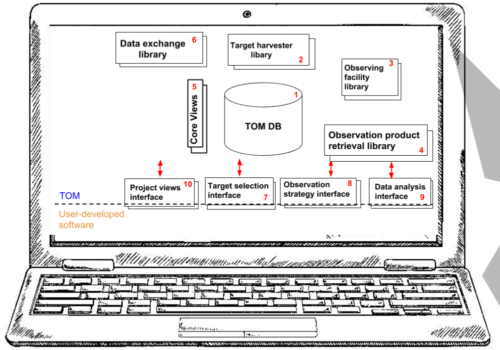


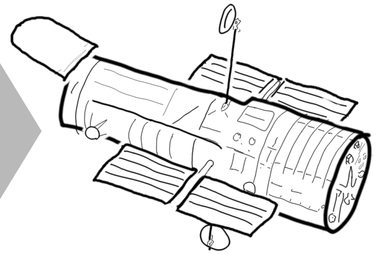
Big data and time domain astrophysics

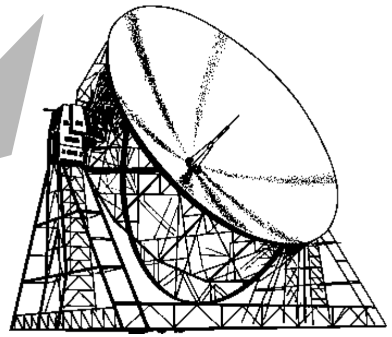
AEON
Astronomical Event Observatory Network











Rubin
Observatory

LSST
Legacy Survey of Space and Time

ALeRCE
Automatic Learning for the Rapid Classification of Events

ANTARES beta

Lasair

TOM
TOM TOOLKIT Las Cumbres Observatory LCO

the astronomy discovery chain

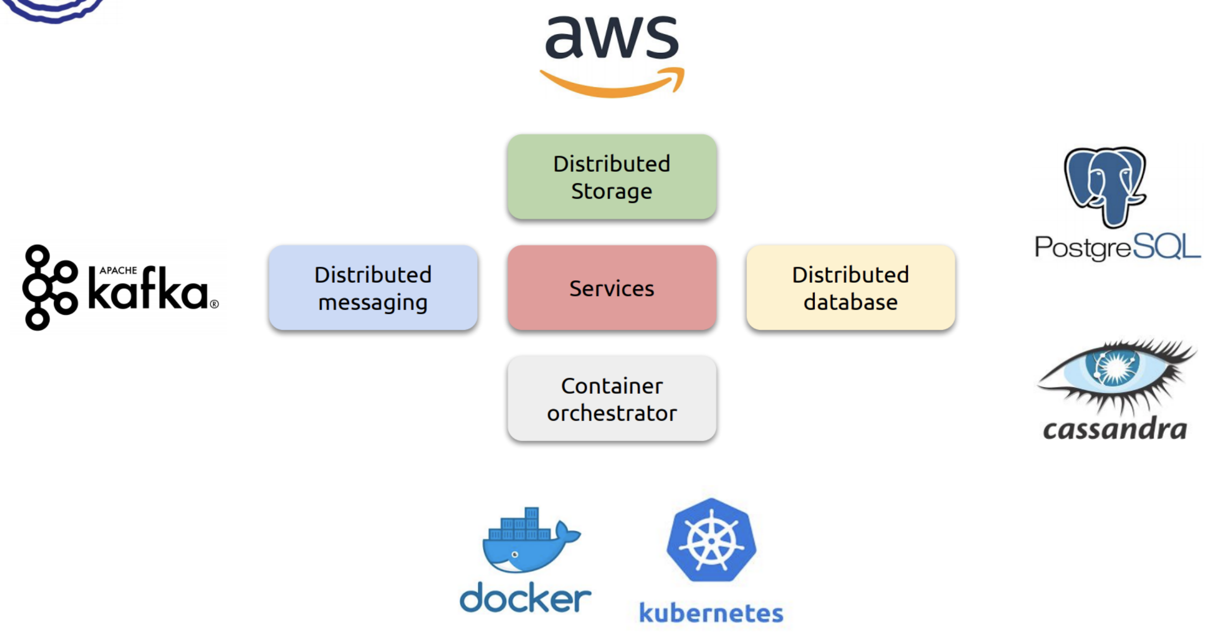
@fedhere

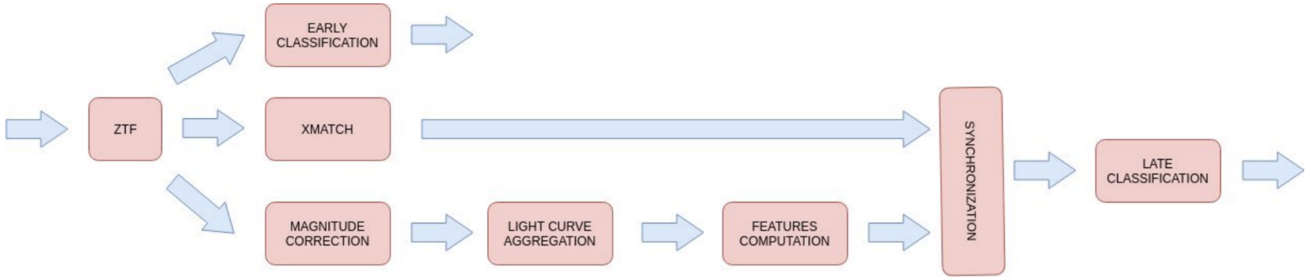
Big data and time domain astrophysics

Alert brokers augmentation and classification

https://cpb-us-e1.wpmucdn.com/sites.northwestern.edu/dist/a/2770/files/2019/08/CASTILLO.pdf

ALeRCE infrastructure





ANTARES

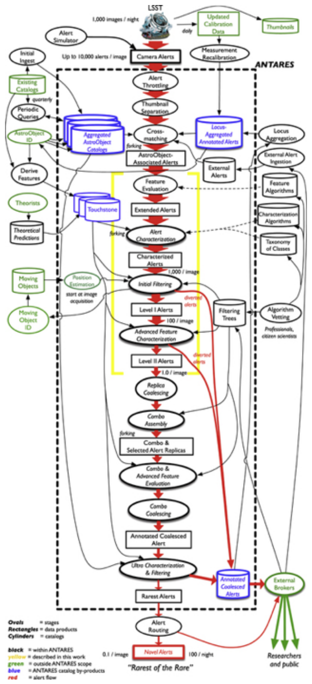
ALeRCE
Automatic Learning for the Rapid Classification of Events

Lasair

Big data and time domain astrophysics

Rubin
Observatory

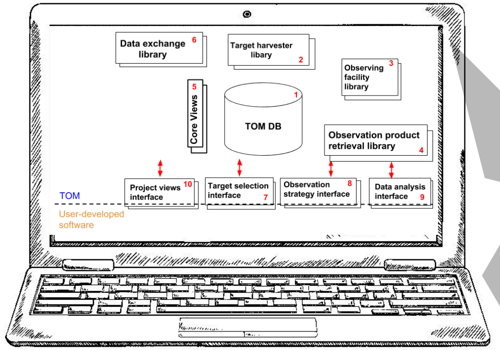
LSST
Legacy Survey of Space and Time



ALeRCE
Automatic Learning for the Rapid Classification of Events

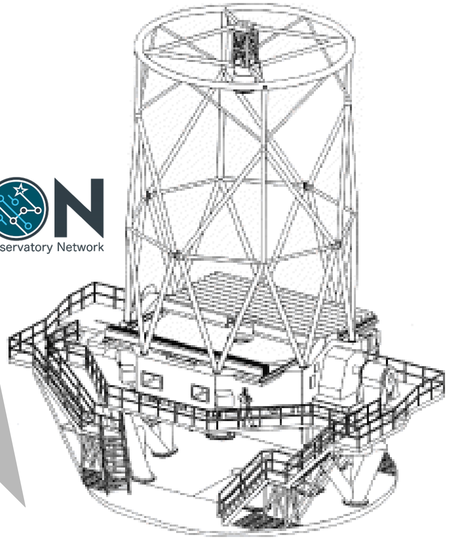
ANTARES beta

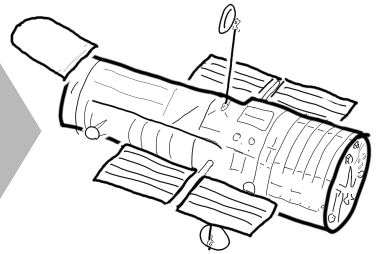
Lasair

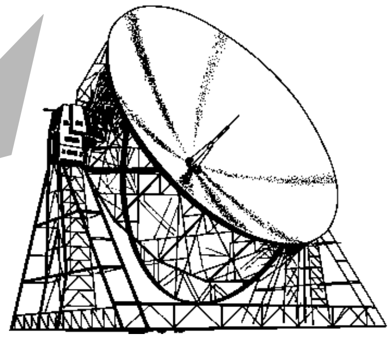


TOM
TOM TOOLKIT
Las Cumbres Observatory LCO

AEON
Astronomical Event Observatory Network







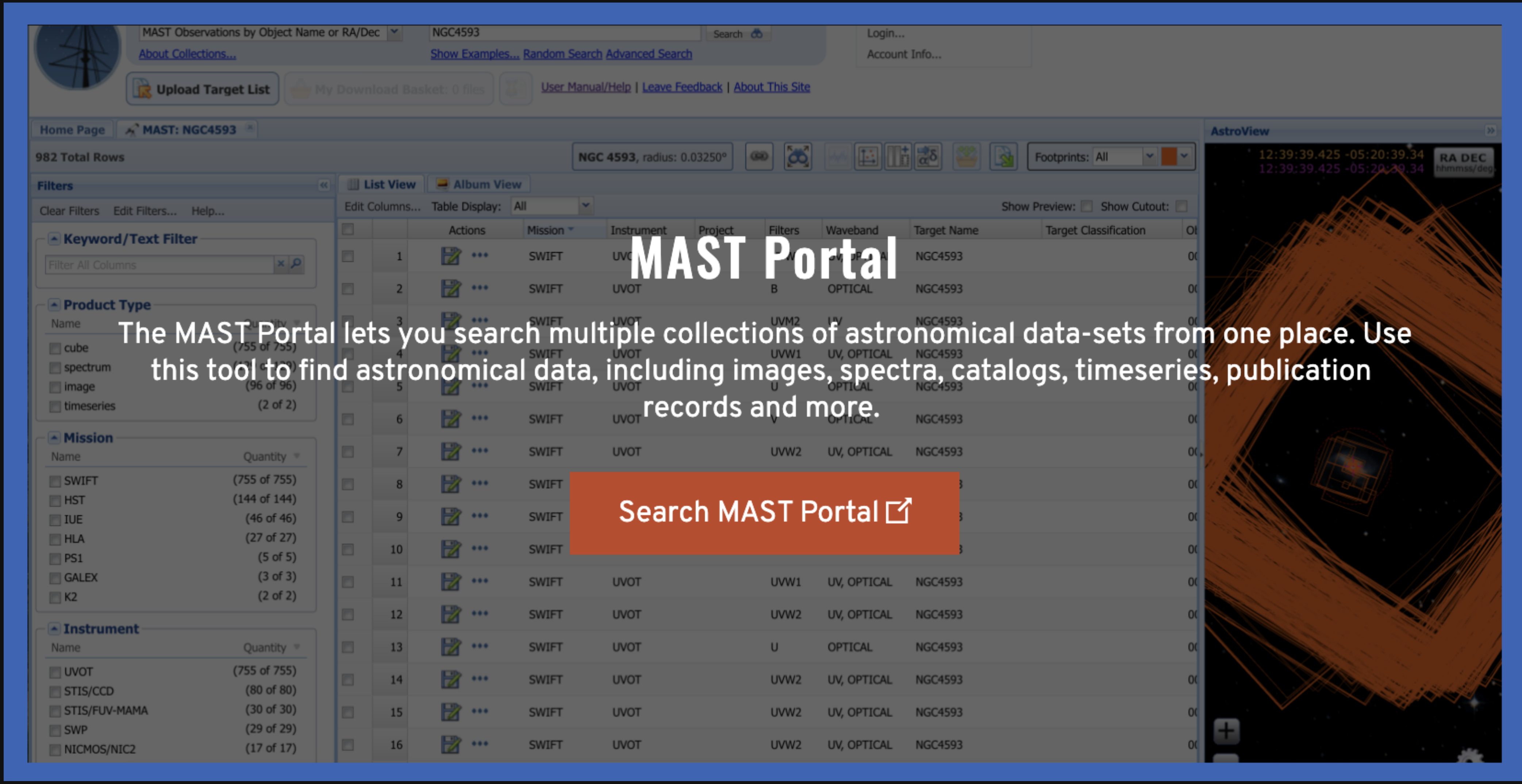
the astronomy discovery chain

@fedhere

6/6

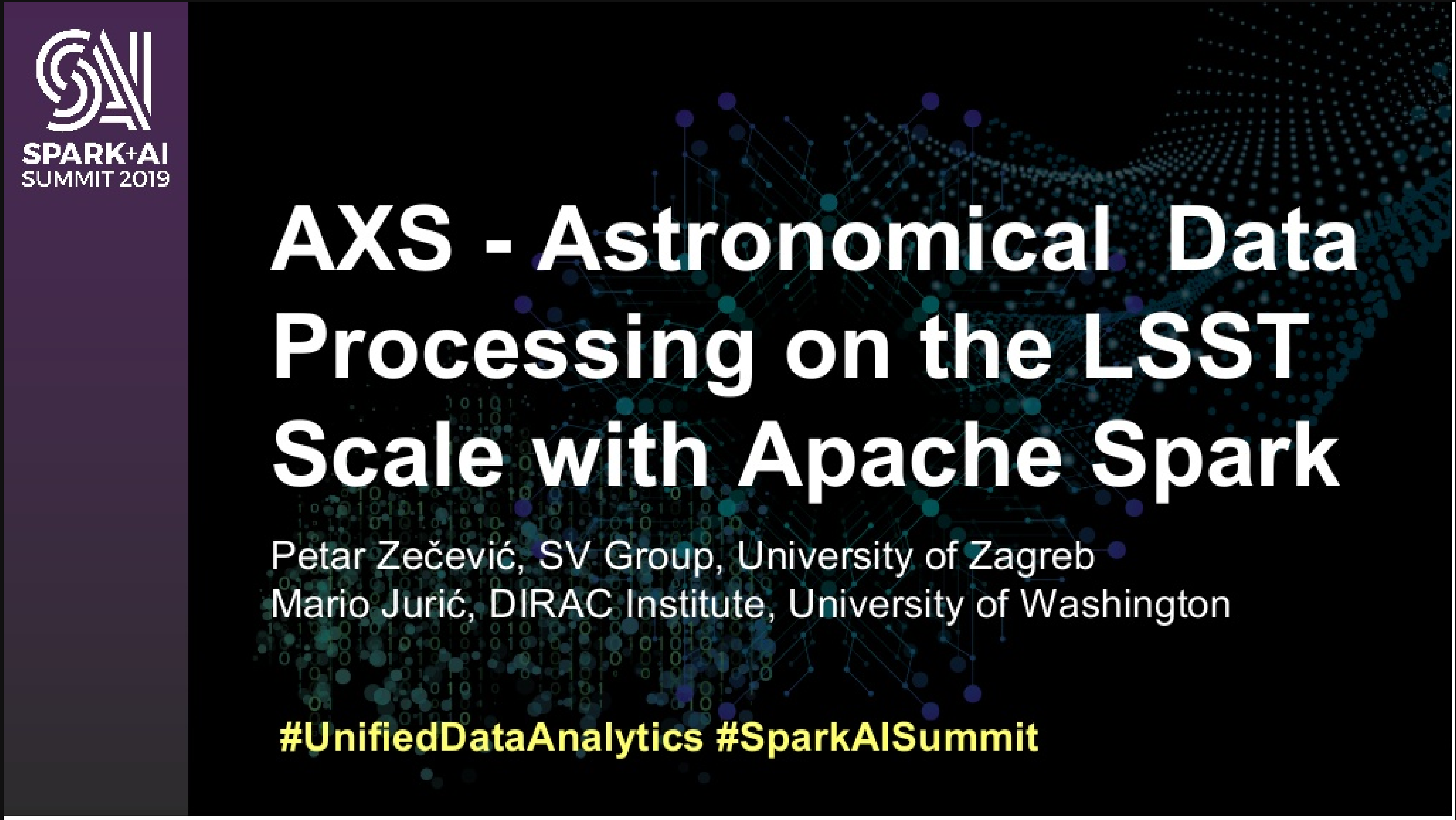
platforms

data platforms



data platforms

https://www.slideshare.net/databricks/astronomical-data-processing-on-the-lsst-scale-with-apache-spark



data platforms

https://www.slideshare.net/databricks/astronomical-data-processing-on-the-lsst-scale-with-apache-spark

SPARK+AI SUMMIT 2019

Enter Spark, AXS

- AXS: Astronomy eXtensions for Spark
- The main idea:
  - Spark is a proven, scalable, cloud-ready and widely-supported analytics framework with full SQL support (legacy support).
  - Extend it to exploratory data analysis.
  - Add a scalable positional cross-match operator
  - Add a domain-specific Python API layer to PySpark
  - Couple to S3 API for storage, Kubernetes for orchestration…
- … A scalable platform supporting an arbitrarily sized dataset and a large number of users, deployable on either public or private cloud.

data platforms

https://www.sciserver.org/

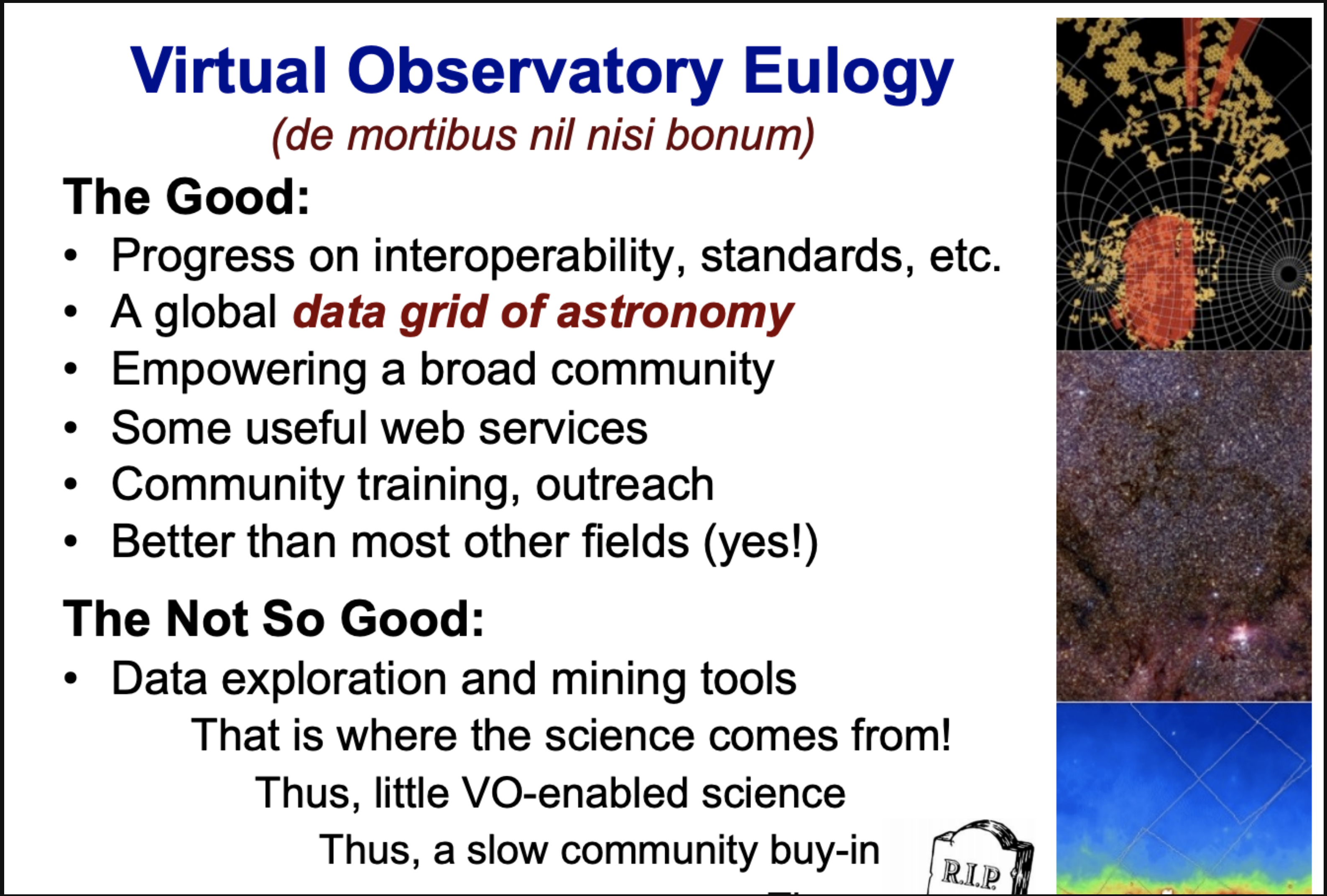
VO

Provide and federate content (data, metadata) services, standards, and analysis/compute services – Develop and provide data exploration and discovery tools

https://ivoa.net/astronomers/getting_started.html

VO

Provide and federate content (data, metadata) services, standards, and analysis/compute services – Develop and provide data exploration and discovery tools



https://events.asiaa.sinica.edu.tw/school/20170904/talk/djorgovski1.pdf

Djorovski (Caltech) 2017

Are VO's successful in other fields...

Google Earth Engine

https://www.youtube.com/embed/MnCf9Gjz720?enablejsapi=1

https://events.asiaa.sinica.edu.tw/school/2
0170904/talk/djorgovski1.pdf

OPINION:
what astro did right about BD

FITS files: universal data storage

Strong pressure on making data public

Strong tradition of collaboration

OPINION:
what astro did right about BD

Still lack of trust in cloud services

sparse collaboration between institutes generating solutions, a ton of platforms that work differently

slow integration of methods

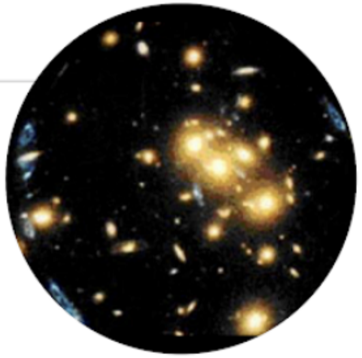
Shameless plug: Rubin LSST Science Collaborations

Four Science Goals

Rubin
Observatory

Dark Matter, Dark Energy

- Weak Lensing
- Baryon acoustic oscillations
- Supernovae, Quasars

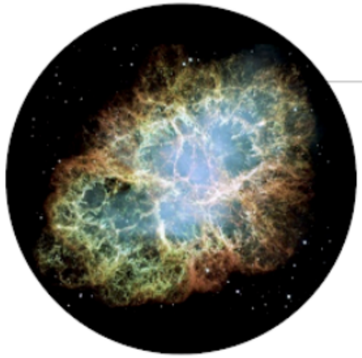


Cataloging the Solar System

- Potentially Hazardous Asteroids
- Near Earth Objects
- Object inventory of the Solar System

Milky Way Structure & Formation

- Structure and evolutionary history
- Spatial maps of stellar characteristics
- Reach well into the halo



Exploring the Transient sky

- Variable stars, Supernovae
- Fill in the variability phase-space
- Discovery of new classes of transients

"From Science Drivers to Reference Design", Ivezić et al. (2008), arXiv:0805.2366

Rubin LSST Science Collaborations

TVS

DESC

GalaxiesSC

SLSC

SMWLV

SSSC

AGN

ISSC

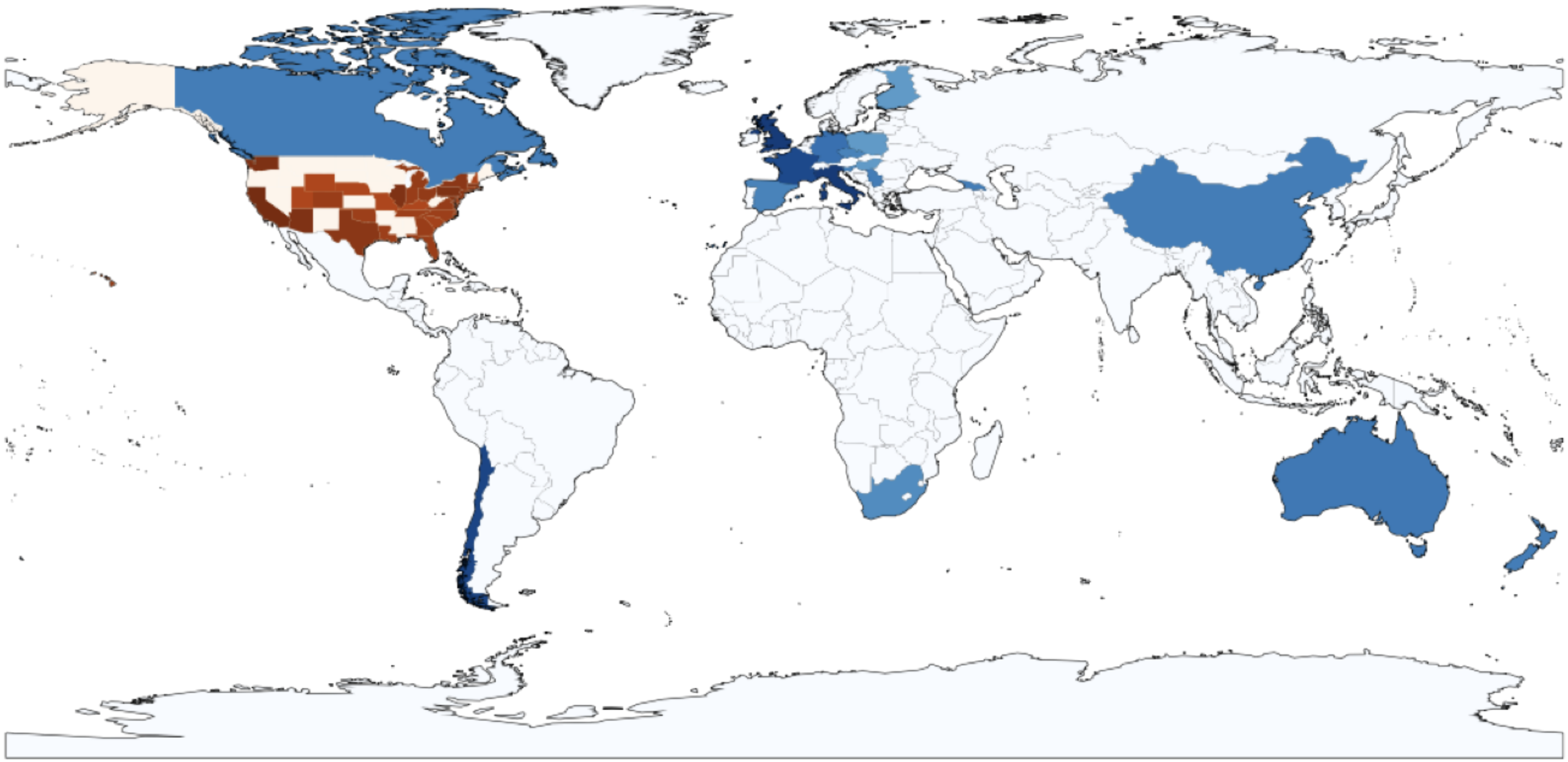
No federally funded LSST science

No science is reserved for any one group

Since there is no science team, science preparation is done by the Science Collaborations... for free!

1500+ members

Rubin Observatory LSST SCs

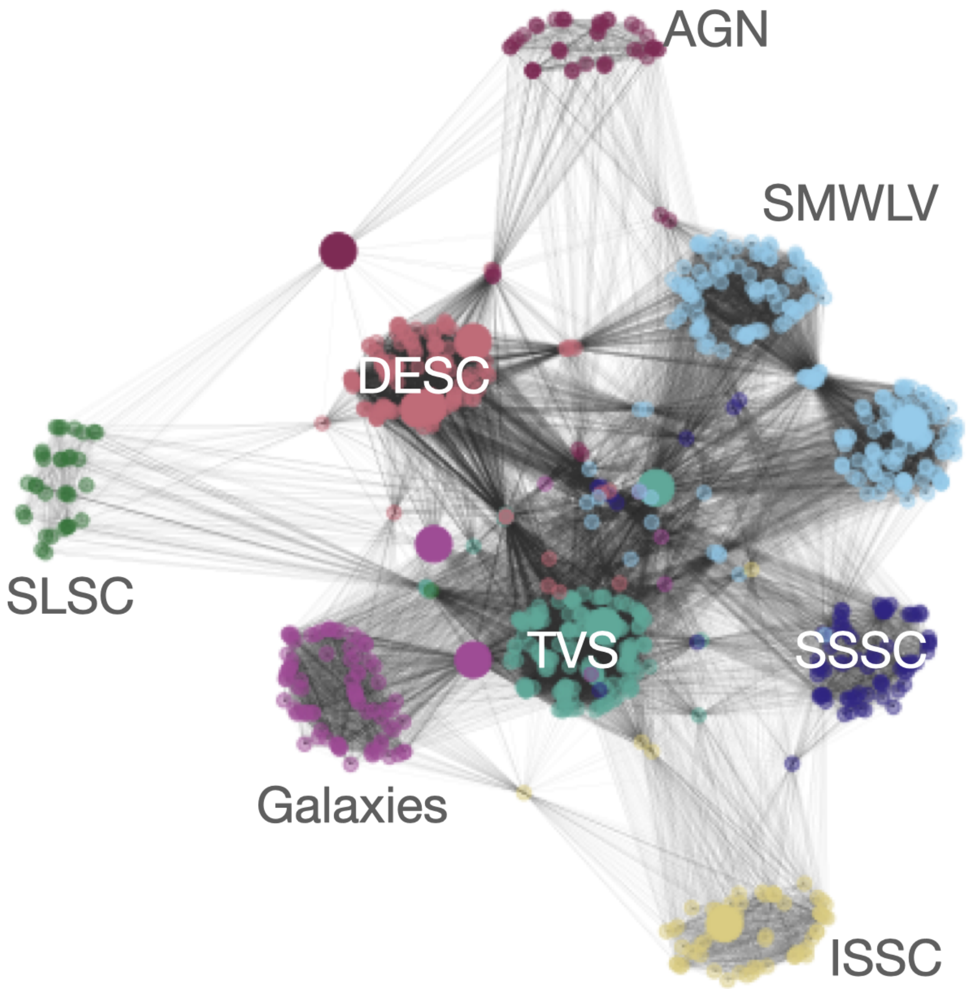


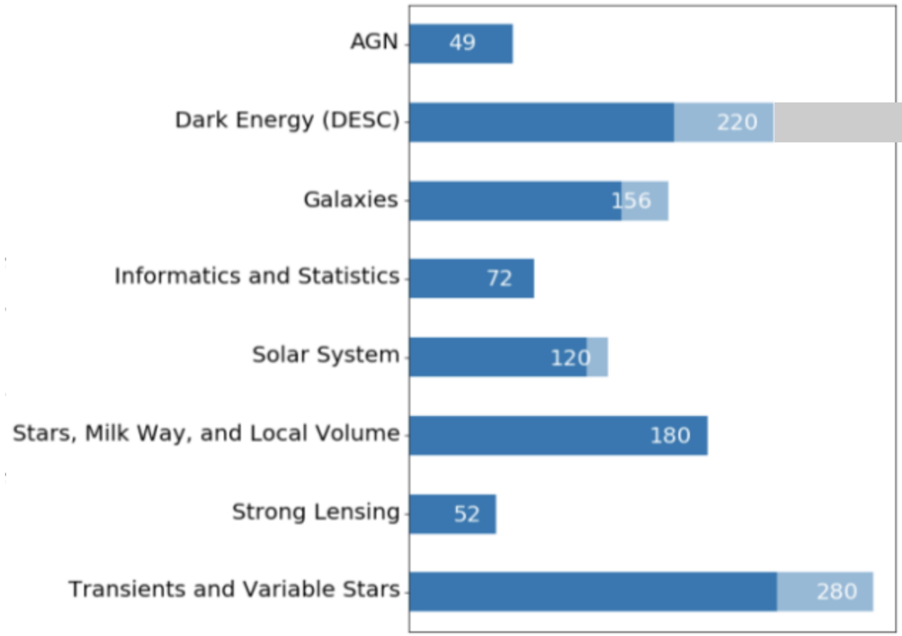
Rubin Observatory

Science Collaborations

LSST Legacy Survey of Space and Time

Rubin Observatory LSST SCs





Rubin Observatory

Legacy Survey of Space and Time

Rubin LSST Science Collaborations

TVS
DESC
GalaxiesSC
SLSC

We aspire to be an inclusive, equitable, and ultimately just group and we are working with renewed vigor in the wake of the recent event that exposed inequity and racism in our society to turning this aspiration into action.

https://lsst-tvssc.github.io/calltoaction.html

SMWLV
SSSC
AGN
ISSC

#desc-for-black-lives

@heather999 created this channel on June 9th. This is the very beginning of the **#desc-for-black-lives** channel. Description: Dialogues about how each of us as individual DESC members and our collaboration as a whole can help eradicate anti-Black racism. (edit)

Diversity Equity and Inclusion council of the SCs

Thank you!

Federica B. Bianco
University of Delaware
Physics and Astronomy
Biden School of Public Policy and Administration
Data Science Institute
NYU Center for Urban Science and Progress

Rubin Observatory LSST Science Collaborations Coordinator

Rubin LSST Transients and Variable Stars Science Collaborations Chair

please email me if you have questions!
fbianco@udel.edu