Introducción a las bases de datos astronómicas

An introduction to astronomical databases

Sundar Srinivasan

Escuela de Verano, IRyA/UNAM, 2021-07-27

Basic information

Except for the video of this presentation, all files necessary for the workshop are hosted on Github:

https://github.com/sundarjhu/EscueladeVerano_2021

The Moodle platform should mirror the Github repositiory:

https://apps.irya.unam.mx/moodle/course/view.php?id=13

Data is central to astronomy

"Astronomers have invested heavily in knowledge infrastructures – robust networks of people, artifacts, and institutions that generate, share, and maintain specific knowledge about the human and natural worlds."

— Borgman & Wofford, "From Data Processes to Data Products: Knowledge Infrastructures in Astronomy", *Harvard Data Science Review*, July 2021.

Data is central to astronomy

"Astronomers have invested heavily in knowledge infrastructures – robust networks of people, artifacts, and institutions that generate, share, and maintain specific knowledge about the human and natural worlds."

— Borgman & Wofford, "From Data Processes to Data Products: Knowledge Infrastructures in Astronomy", *Harvard Data Science Review*, July 2021.

Three related principles

Data is central to astronomy

"Astronomers have invested heavily in knowledge infrastructures – robust networks of people, artifacts, and institutions that generate, share, and maintain specific knowledge about the human and natural worlds."

— Borgman & Wofford, "From Data Processes to Data Products: Knowledge Infrastructures in Astronomy", *Harvard Data Science Review*, July 2021.

Three related principles

***[Avoiding] Duplication of effort*: Has someone already done this, and is their work readily accessible and reproducible?**

***Automation*: Will I (or someone else) need to use this again?**

***Reproducibility*: Will I (or someone else) be able to repeat my work with the same results?**

How it started

Hipparchos (c.190 BCE—c.120 BCE)
Wikimedia, Public Domain

Catalogue of ~850 star positions

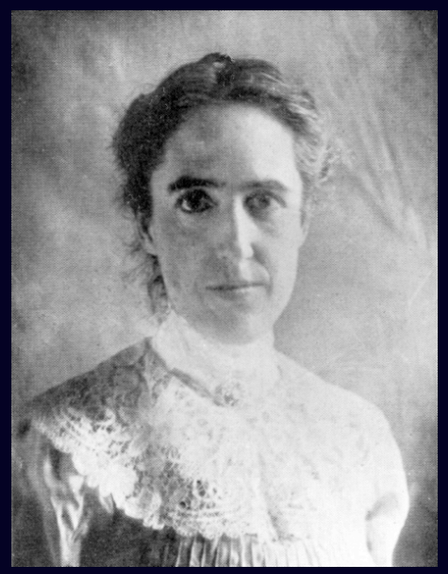
Tycho Brahe (1546-1601)
Edouard Ender/Wikimedia, Public Domain

Positions of naked-eye planets

Charles Messier (1730—1817)
Ansiaux/Stoyan et al. 2008, Public Domain

Catalogue of 110 nebulae/star clusters

Henrietta Swan Leavitt (1868–1921)
Wikimedia, Public Domain



Catalogue of stellar brightnesses

Edwin Hubble (1889-1953)
Mt. Wilson Archive, Carnegie Inst. of Washington

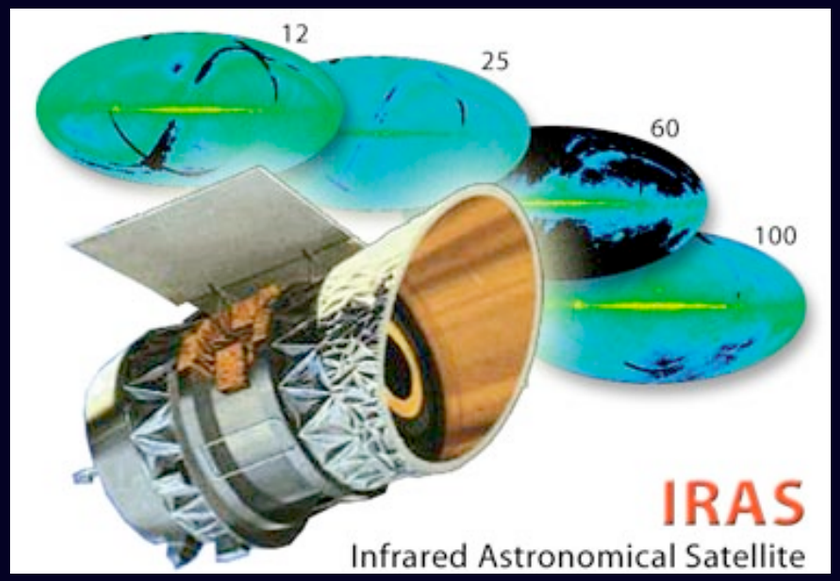
Era of large area/all-sky imaging surveys

Credit: Y. Wadadekar, NCRA/TIFR

How it's going

Infrared Astronomical Satellite (1983-1983)

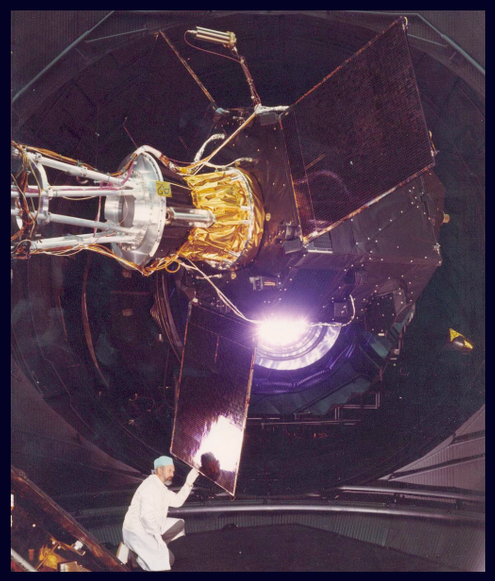
Wikimedia, Public Domain



>250,000 infrared point sources

Hipparcos satellite (1989–1993)

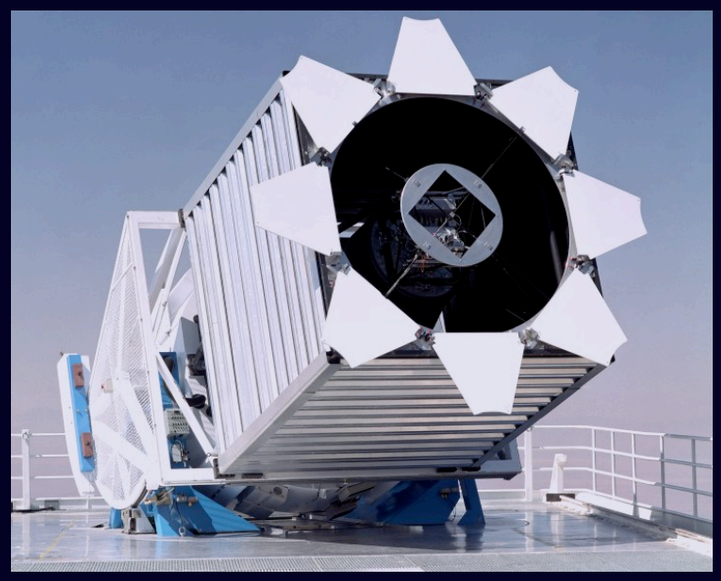
M. Perryman/CC BY-SA 3.0



>10^6 star parallaxes

Sloan Digital Sky Survey (2000—)

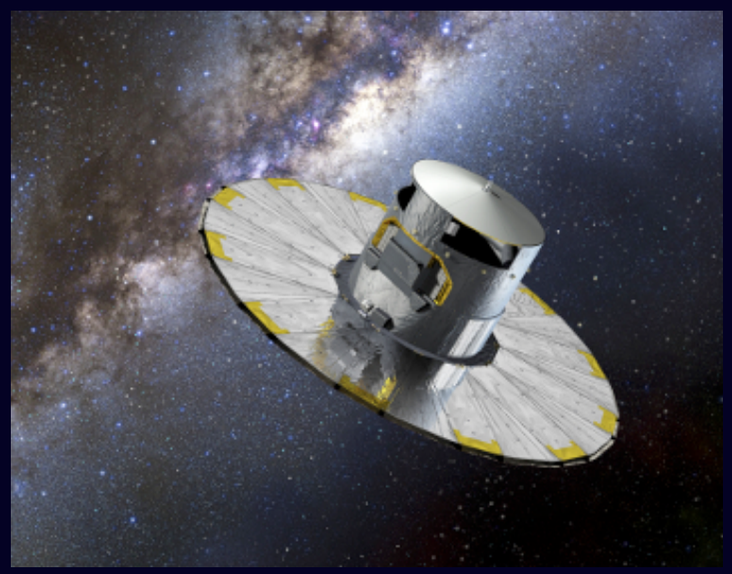
SDSS.org



Photometry/spectra of 4x10^6 objects

Gaia (2013–c. 2022)

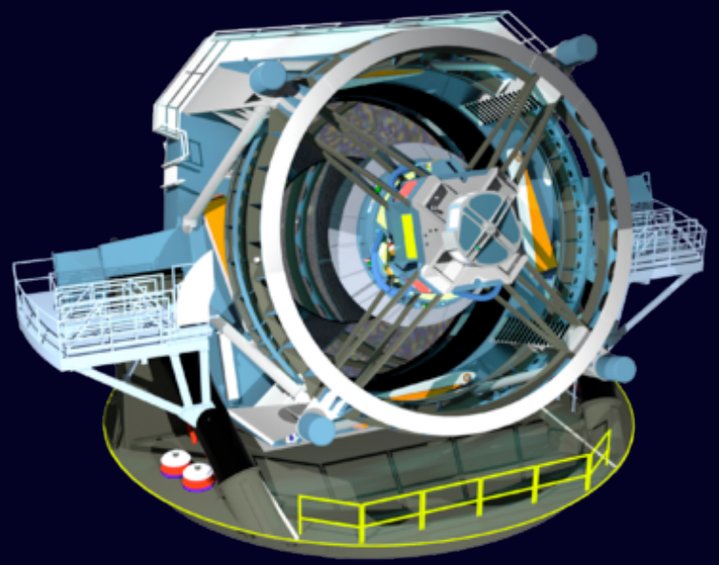
European Space Agency



Astrometry for >10^9 objects, 60 TB @ 1 Mb/s

Vera C. Rubin Observatory (2020s)

LSST.org / CC BY-SA 3.0



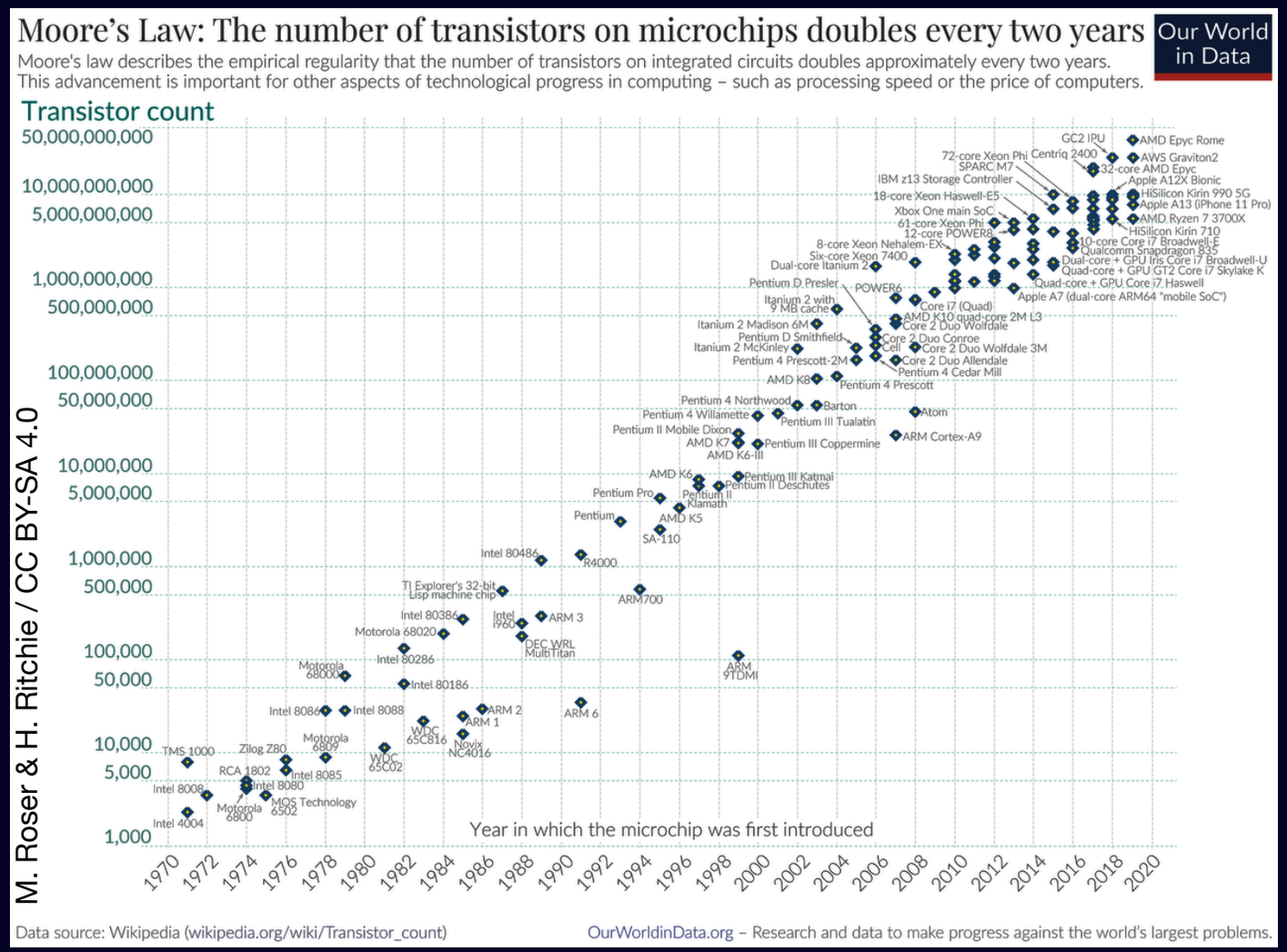
30 TB/night, 100 Gb/s

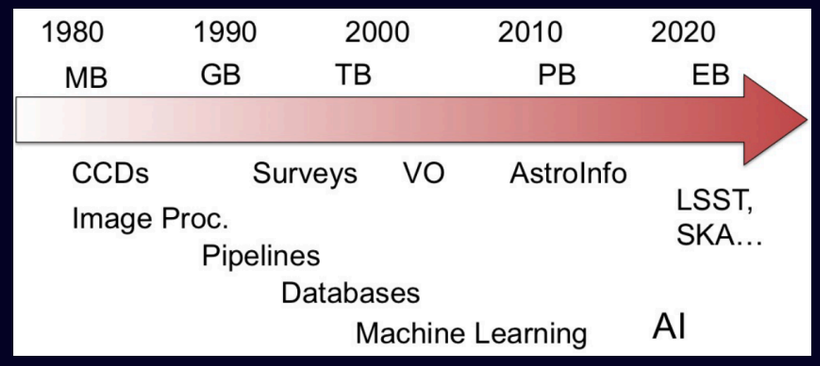
Square Kilometer Array (2020s)

SKAtelescope.org / CC BY-SA 3.0

~EB of data, ~Pb/s

The Big Data Challenge



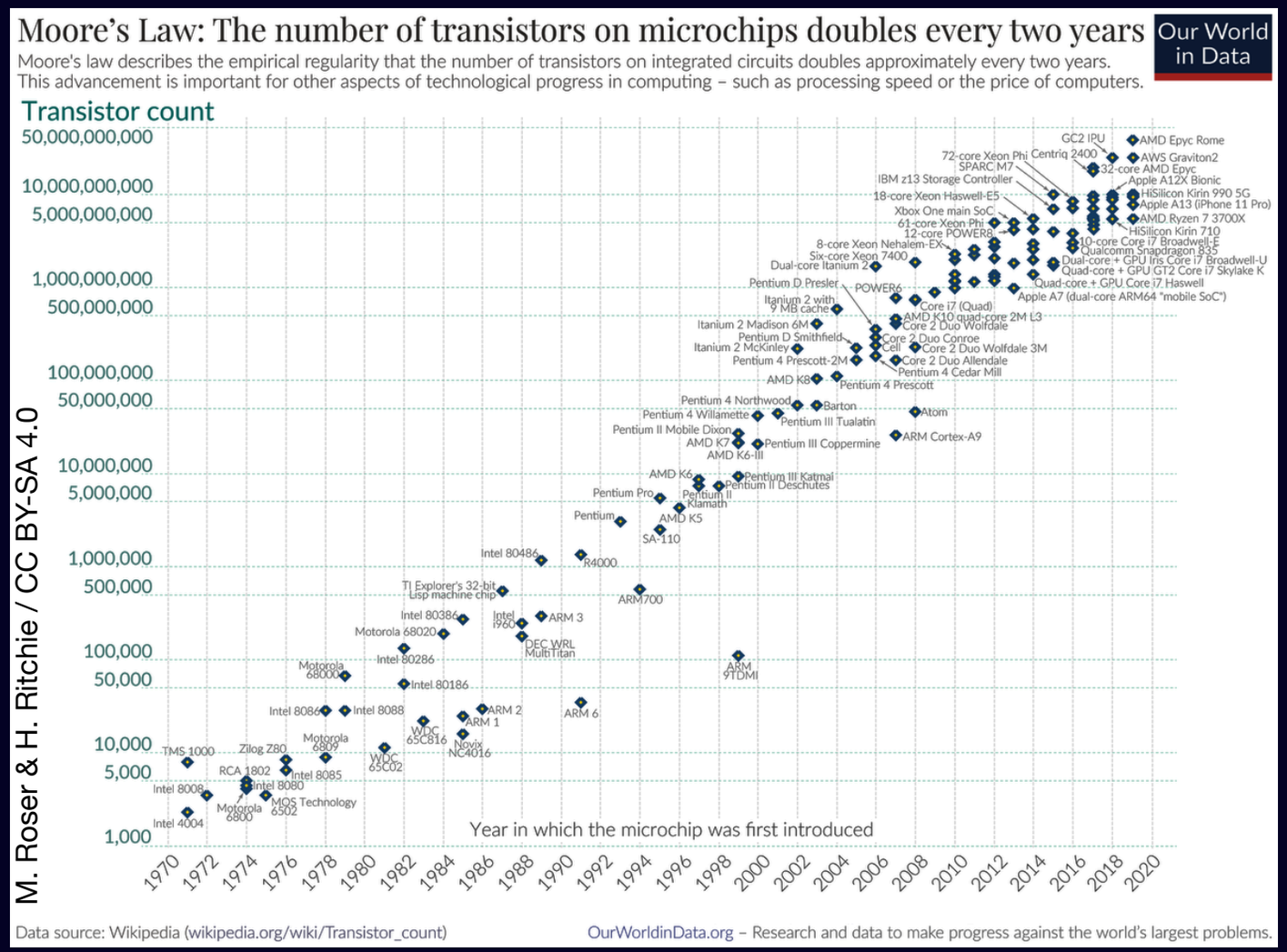


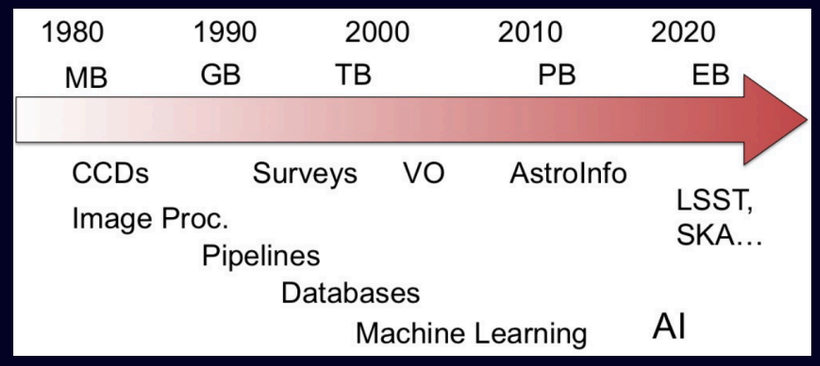
Increasing computing capability, storage, memory, and speed of connectivity has led to increase in quality and quantity of digital archives. Improved internet access and speeds have improved accessibility of these archives.

Increase of data-driven science instead of hypothesis-driven science.

Rise of citizen science (e.g., zooniverse.org — classify galaxies on your phone).

The Big Data Challenge





Increasing computing capability, storage, memory, and speed of connectivity has led to increase in quality and quantity of digital archives. Improved internet access and speeds have improved accessibility of these archives.

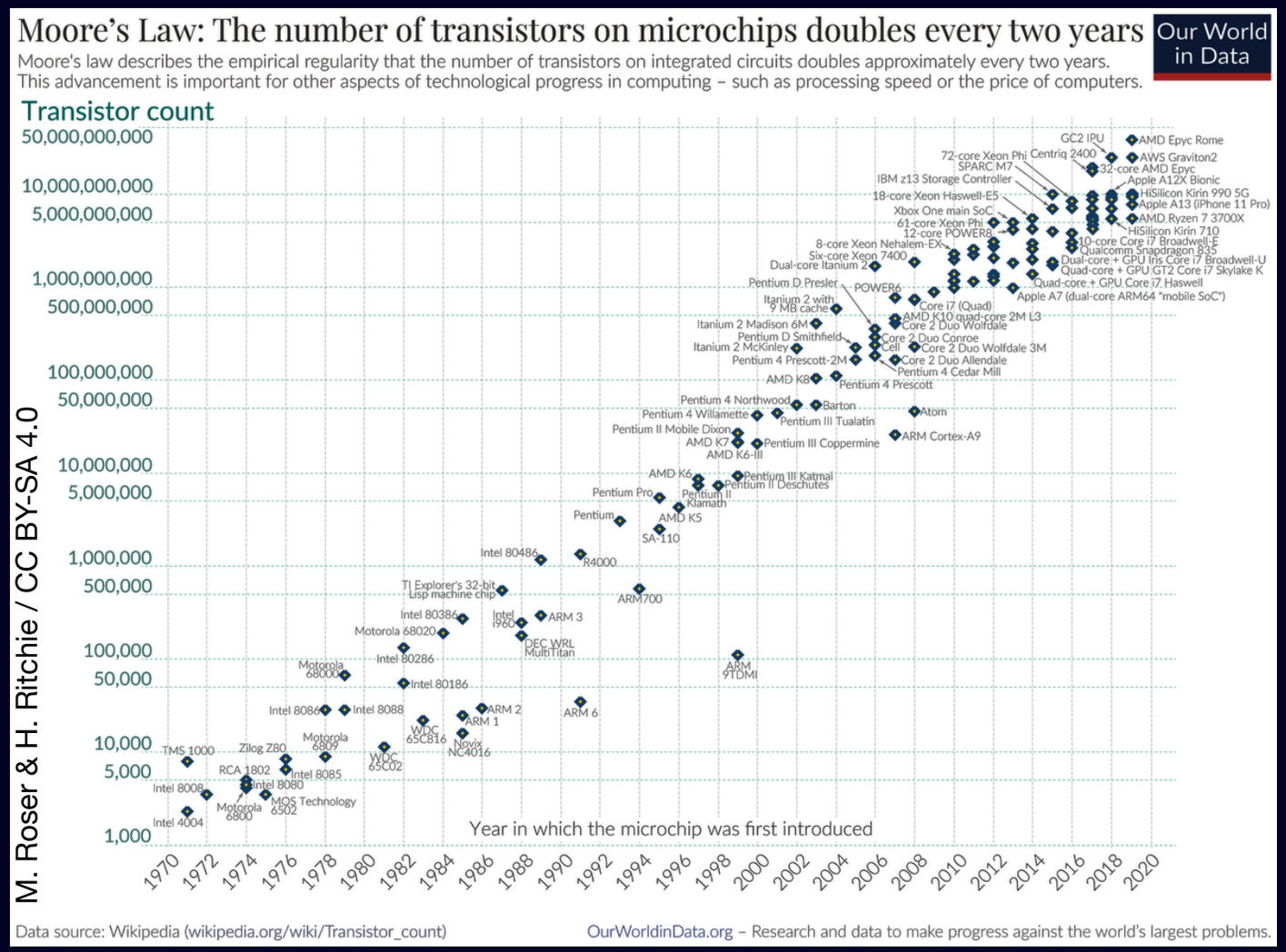
Increase of data-driven science instead of hypothesis-driven science.

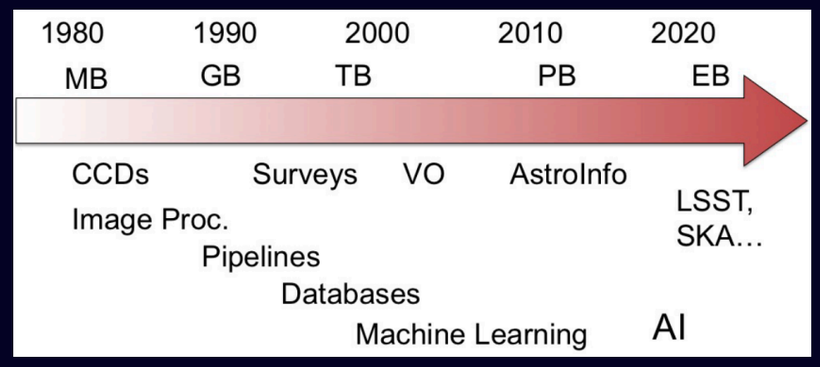
Rise of citizen science (e.g., zooniverse.org — classify galaxies on your phone).

Information content: Most data will never be seen by humans.

Information complexity: Patterns in these data cannot be directly comprehended by humans.

The Big Data Challenge





Increasing computing capability, storage, memory, and speed of connectivity has led to increase in quality and quantity of digital archives. Improved internet access and speeds have improved accessibility of these archives.

Increase of data-driven science instead of hypothesis-driven science.

Rise of citizen science (e.g., zooniverse.org — classify galaxies on your phone).

Information content: Most data will never be seen by humans.

Information complexity: Patterns in these data cannot be directly comprehended by humans.

Need a robust way to store, access, visualise, and analyse such voluminous datasets.

Need to integrate software and web-based tools required to access large archives.

Need to incorporate ways to connect these datasets to relevant publications.

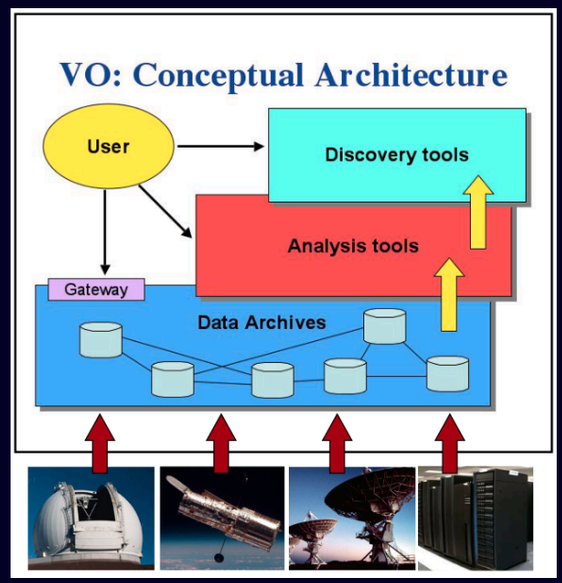
"Interoperability"

Virtual observatories

“A collection of interoperating data archives and software tools which utilise the internet to form a scientific research environment for astronomical research.”
- Wikipedia

Virtual observatories

"A collection of interoperating data archives and software tools which utilise the internet to form a scientific research environment for astronomical research."
- Wikipedia

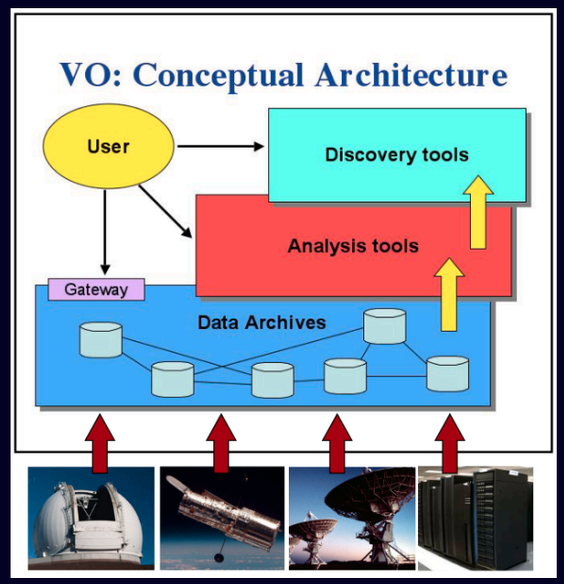


Infrastructure for storage of (and tools for access to) massive/complex datasets. User-friendly interface allowing data discovery, visualisation, and analysis.

~Twenty countries in VO alliance. Education and outreach. Enables science in developing countries through data access.

Virtual observatories

"A collection of interoperating data archives and software tools which utilise the internet to form a scientific research environment for astronomical research."
- Wikipedia



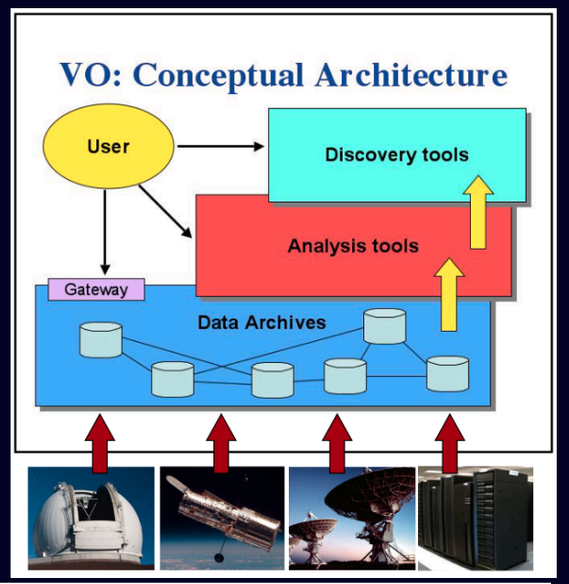
Infrastructure for storage of (and tools for access to) massive/complex datasets. User-friendly interface allowing data discovery, visualisation, and analysis.

~Twenty countries in VO alliance. Education and outreach. Enables science in developing countries through data access.

Data in different centres/archives differ in file structure, metadata, and table organisation. Difficult (if not impossible) to access data in a straightforward manner. Even more complicated if we want to combine several data sets.
Needs standardisation!

Virtual observatories

"A collection of interoperating data archives and software tools which utilise the internet to form a scientific research environment for astronomical research."
- Wikipedia

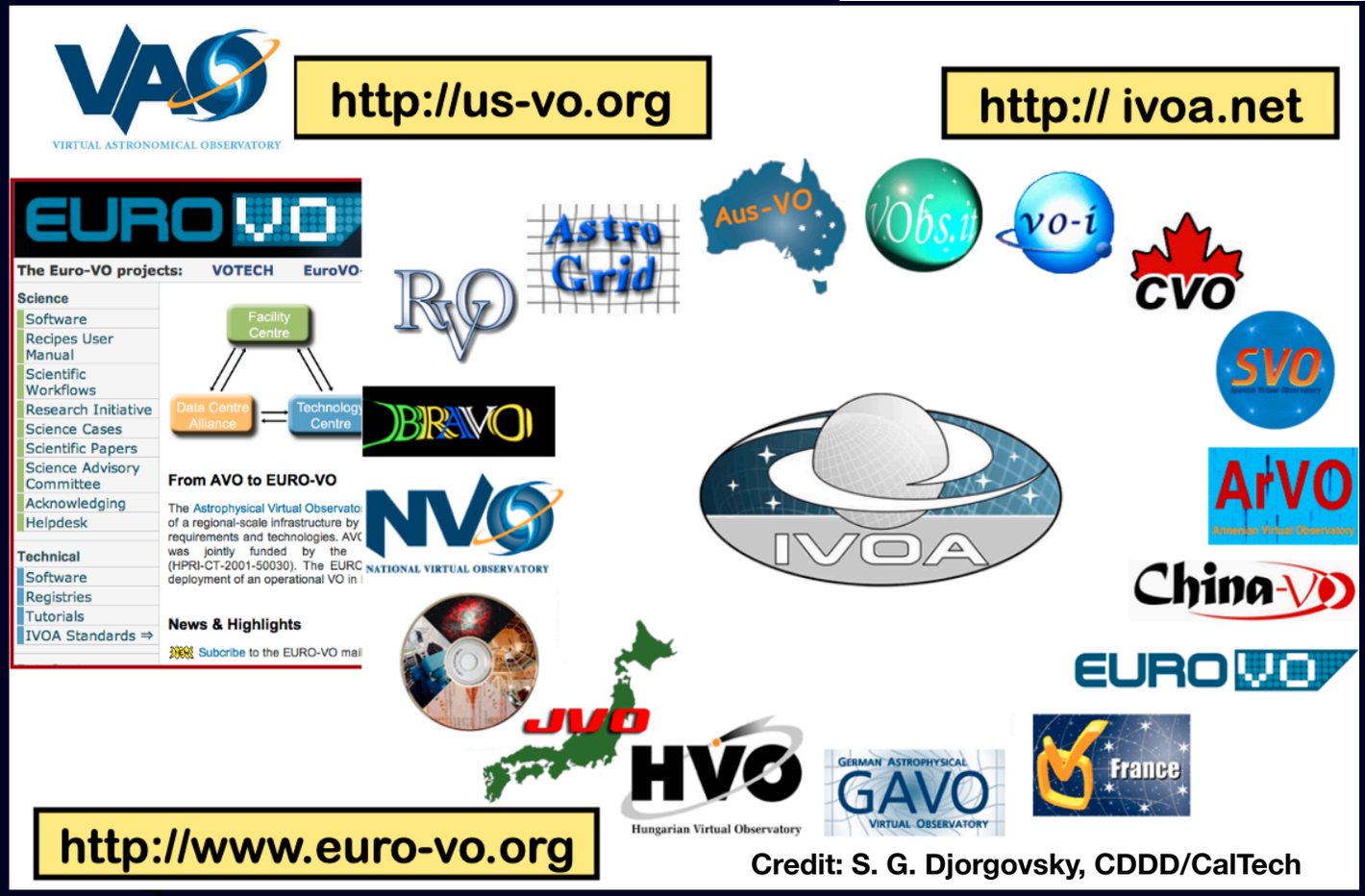


Infrastructure for storage of (and tools for access to) massive/complex datasets. User-friendly interface allowing data discovery, visualisation, and analysis.

~Twenty countries in VO alliance. Education and outreach. Enables science in developing countries through data access.

Data in different centres/archives differ in file structure, metadata, and table organisation. Difficult (if not impossible) to access data in a straightforward manner. Even more complicated if we want to combine several data sets. Needs standardisation!

The IVOA (International Virtual Observatory Alliance):
standards body created by the VO projects to develop the vital interoperability standards upon which the VO implementations are constructed.



Introduction to databases

Databases/Archives — what's there and which of these we'll use

Data — what we will access

Tools — ways to access the data and which of these we'll use

The most popular astronomical databases

The most popular astronomical databases

Astronomical publications: the SAO/NASA Astrophysics Data System (ADS)

"How many papers before 1950 mention white dwarfs?"

The most popular astronomical databases

Astronomical publications: the SAO/NASA Astrophysics Data System (ADS)

"How many papers before 1950 mention white dwarfs?"

Database of objects beyond the Solar System:

The most popular astronomical databases

Astronomical publications: the SAO/NASA Astrophysics Data System (ADS)

"How many papers before 1950 mention white dwarfs?"

Database of objects beyond the Solar System:

SIMBAD (Set of Indentificiations, Measurements, and Bibliography for Astronomical Data)

"What is the radial velocity of α Centauri?"

NED (NASA Extragalactic Database)

"What is the distance to the Triangulum Galaxy M33?"

The most popular astronomical databases

Astronomical publications: the SAO/NASA Astrophysics Data System (ADS)

"How many papers before 1950 mention white dwarfs?"

Database of objects beyond the Solar System:

SIMBAD (Set of Indentificiations, Measurements, and Bibliography for Astronomical Data)

"What is the radial velocity of α Centauri?"

NED (NASA Extragalactic Database)

"What is the distance to the Triangulum Galaxy M33?"

The most popular astronomical databases

Astronomical publications: the SAO/NASA Astrophysics Data System (ADS)

"How many papers before 1950 mention white dwarfs?"

Database of objects beyond the Solar System:

SIMBAD (Set of Indentificiations, Measurements, and Bibliography for Astronomical Data)
"What is the radial velocity of α Centauri?"

NED (NASA Extragalactic Database)
"What is the distance to the Triangulum Galaxy M33?"

Published astronomical catalogues:

Vizier
"What are the 25 μm fluxes of the ten brightest sources in the IRAS point-source catalogue?"

Other popular large databases/archives

Other popular large databases/archives

The InfraRed Space Archive (IRSA), hosted at IPAC (CalTech)

Archive for NASA's infrared and submillimetre projects, including data from the Infrared Astronomical Satellite (IRAS), the Two Micron All-Sky Survey (2MASS), Spitzer Space Telescope (Spitzer), and the Wide-field Infrared Survey Explorer (WISE).

Other popular large databases/archives

The InfraRed Space Archive (IRSA), hosted at IPAC (CalTech)

Archive for NASA's infrared and submillimetre projects, including data from the Infrared Astronomical Satellite (IRAS), the Two Micron All-Sky Survey (2MASS), Spitzer Space Telescope (Spitzer), and the Wide-field Infrared Survey Explorer (WISE).

Cosmos, hosted by the European Space Agency (ESA)

Archive for ESA missions, including the Gaia mission.

Other popular large databases/archives

The InfraRed Space Archive (IRSA), hosted at IPAC (CalTech)

Archive for NASA's infrared and submillimetre projects, including data from the Infrared Astronomical Satellite (IRAS), the Two Micron All-Sky Survey (2MASS), Spitzer Space Telescope (Spitzer), and the Wide-field Infrared Survey Explorer (WISE).

Cosmos, hosted by the European Space Agency (ESA)

Archive for ESA missions, including the Gaia mission.

The Mikulski Archive for Space Telescopes (MAST)

Hosts data from NASA missions such as the Hubble Space Telescope (HST) and the Sloan Digital Sky Survey (SDSS). Will host James Webb Space Telescope (JWST) data soon.

What data will we access during this workshop?

What data will we access during this workshop?

Properties of point sources (stars)

We will focus on the properties of stars. Many of the same methods/tools are also applicable to archives containing extended objects such as galaxies and interstellar clouds.

What data will we access during this workshop?

Properties of point sources (stars)

We will focus on the properties of stars. Many of the same methods/tools are also applicable to archives containing extended objects such as galaxies and interstellar clouds.

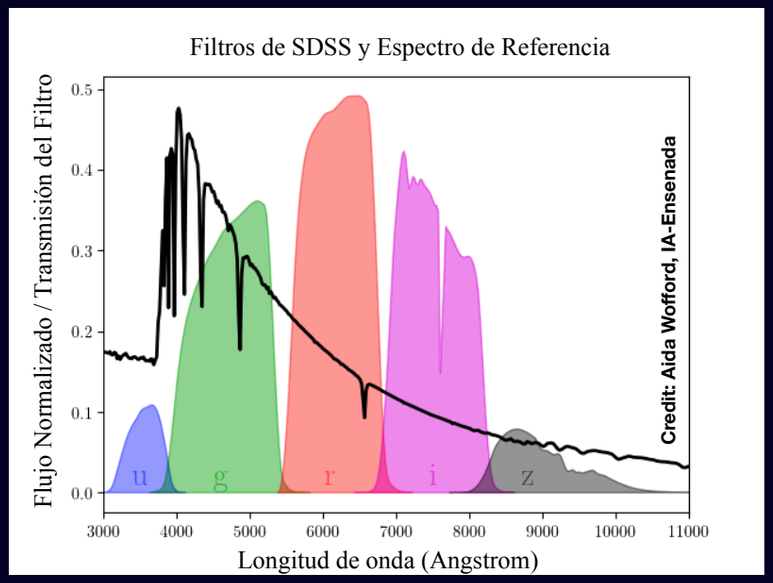
Photometry

We will focus on multi-wavelength photometric data. Similar tools are available for obtaining spectroscopic data.

What data will we access during this workshop?

Properties of point sources (stars)

We will focus on the properties of stars. Many of the same methods/tools are also applicable to archives containing **extended objects** such as galaxies and interstellar clouds.



Photometry

We will focus on **multi-wavelength photometric data**. Similar tools are available for obtaining spectroscopic data.

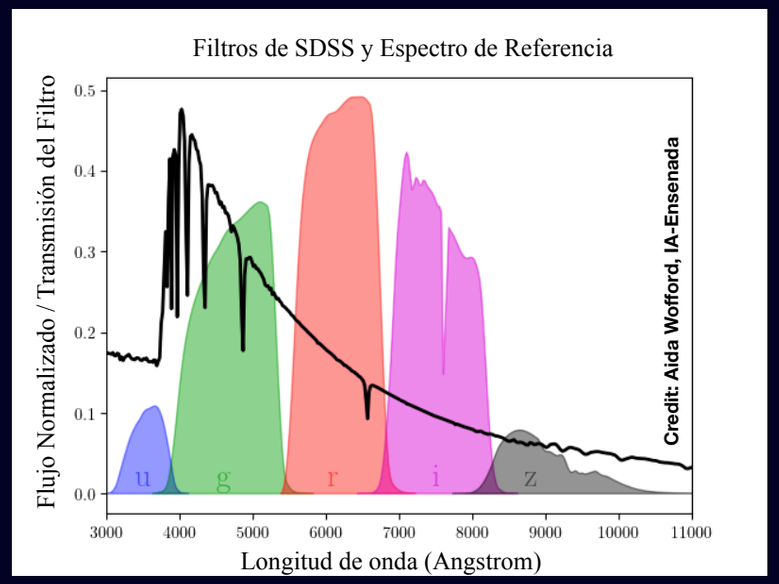
Photometric filters

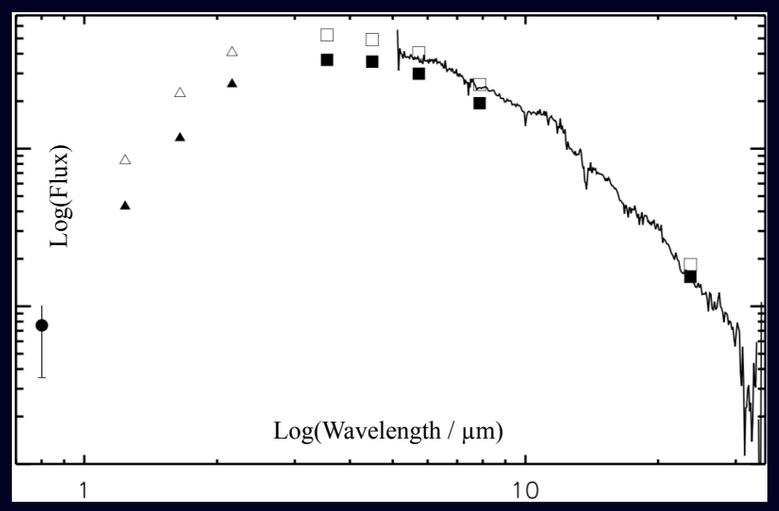
A photometric filter allows photons **within a small range of wavelengths** through to the detector which registers the total flux. This total flux is recorded as the flux in this photometric band.

What data will we access during this workshop?

Properties of point sources (stars)

We will focus on the properties of stars. Many of the same methods/tools are also applicable to archives containing **extended objects** such as galaxies and interstellar clouds.





Photometry

We will focus on **multi-wavelength photometric data**. Similar tools are available for obtaining spectroscopic data.

Photometric filters

A photometric filter allows photons **within a small range of wavelengths** through to the detector which registers the total flux. This total flux is recorded as the flux in this photometric band.

Spectral energy distribution

A collection of photometric fluxes over a **large wavelength range**, it tells us how the energy from the star is distributed as a function of wavelength.

How will we access the data (tools)?

How will we access the data (tools)?

Web interfaces

Allow for simple searches of one source or a small number of objects.

How will we access the data (tools)?

Web interfaces

Allow for simple searches of one source or a small number of objects.

ADQL (Astronomical Data Query Language)

A specialised VO-compliant version of the Structured Query Language (SQL) which allows us to construct “queries” to access specific components of a catalogue. Very flexible, can choose subsets of the data to view/download.

How will we access the data (tools)?

Web interfaces

Allow for simple searches of one source or a small number of objects.

ADQL (Astronomical Data Query Language)

A specialised VO-compliant version of the Structured Query Language (SQL) which allows us to construct “queries” to access specific components of a catalogue. Very flexible, can choose subsets of the data to view/download.

TAP (Table Access Protocol)

A protocol to access tables on the archive independently of the web interface. Communicates with the archive via ADQL queries. Much more freedom than web interfaces.

How will we access the data (tools)?

Web interfaces

Allow for simple searches of one source or a small number of objects.

ADQL (Astronomical Data Query Language)

A specialised VO-compliant version of the Structured Query Language (SQL) which allows us to construct “queries” to access specific components of a catalogue. Very flexible, can choose subsets of the data to view/download.

TAP (Table Access Protocol)

A protocol to access tables on the archive independently of the web interface. Communicates with the archive via ADQL queries. Much more freedom than web interfaces.

Python-based TAP

Python packages like **astroquery** and **PyVO** allow us to access archival data directly from the Python command line. The data are columns viewed/manipulated using **pandas** dataframes or **astropy** tables.

How will we access the data (tools)?

Web interfaces

Allow for simple searches of one source or a small number of objects.

ADQL (Astronomical Data Query Language)

A specialised VO-compliant version of the Structured Query Language (SQL) which allows us to construct “queries” to access specific components of a catalogue. Very flexible, can choose subsets of the data to view/download.

TAP (Table Access Protocol)

A protocol to access tables on the archive independently of the web interface. Communicates with the archive via ADQL queries. Much more freedom than web interfaces.

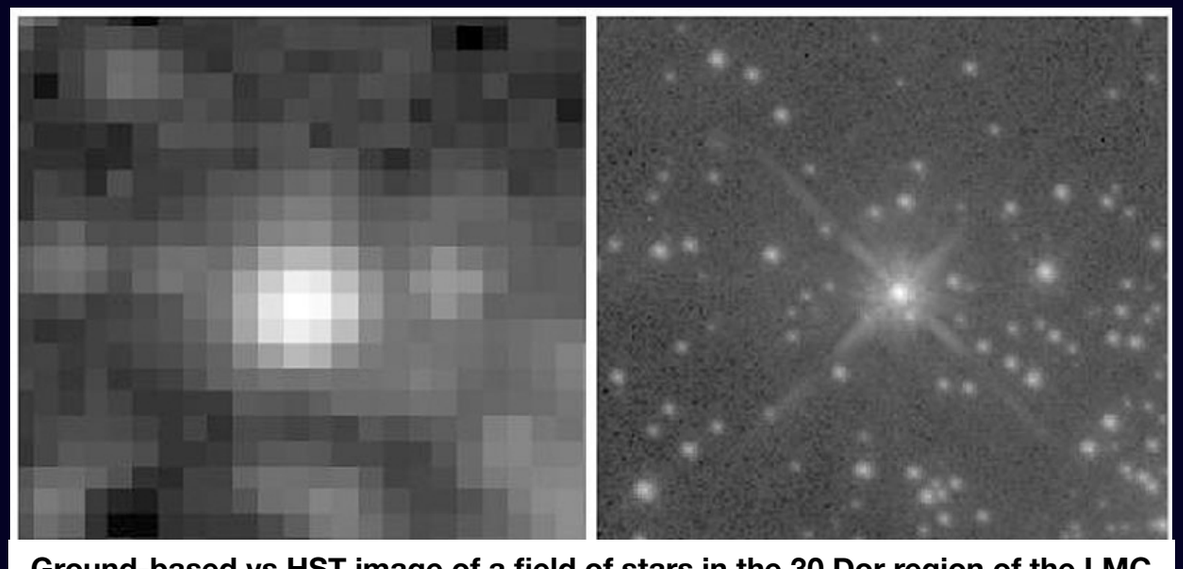
Python-based TAP

Python packages like **astroquery** and **PyVO** allow us to access archival data directly from the Python command line. The data are columns viewed/manipulated using **pandas** dataframes or **astropy** tables.

File format

Communications with the VO are in the form of Virtual Observatory Table (VOTable) files.

Combining data from various surveys and various wavelengths: caveats



Ground-based vs HST image of a field of stars in the 30 Dor region of the LMC

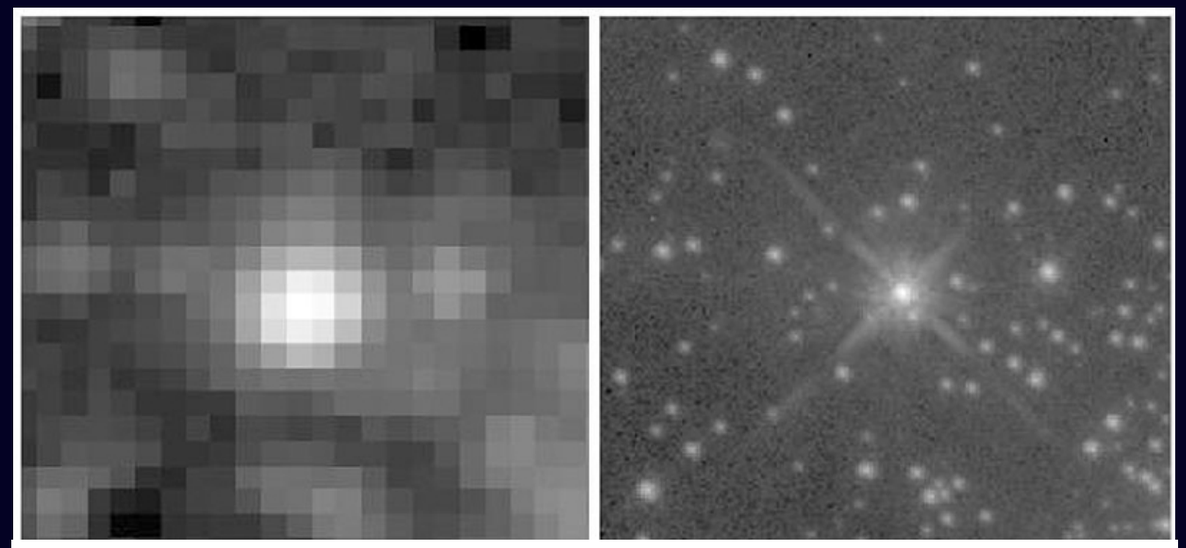
Credit: NASA/ESA, CC BY-SA 4.0

Combining data from various surveys and various wavelengths: caveats

Surveys at the same wavelength but with differing spatial resolution

Spatial resolution improves with technology. A single source may resolve into multiple objects in the higher-resolution image.

IRAS observations from the 1970s vs. WISE data from the 2000s.



Ground-based vs HST image of a field of stars in the 30 Dor region of the LMC

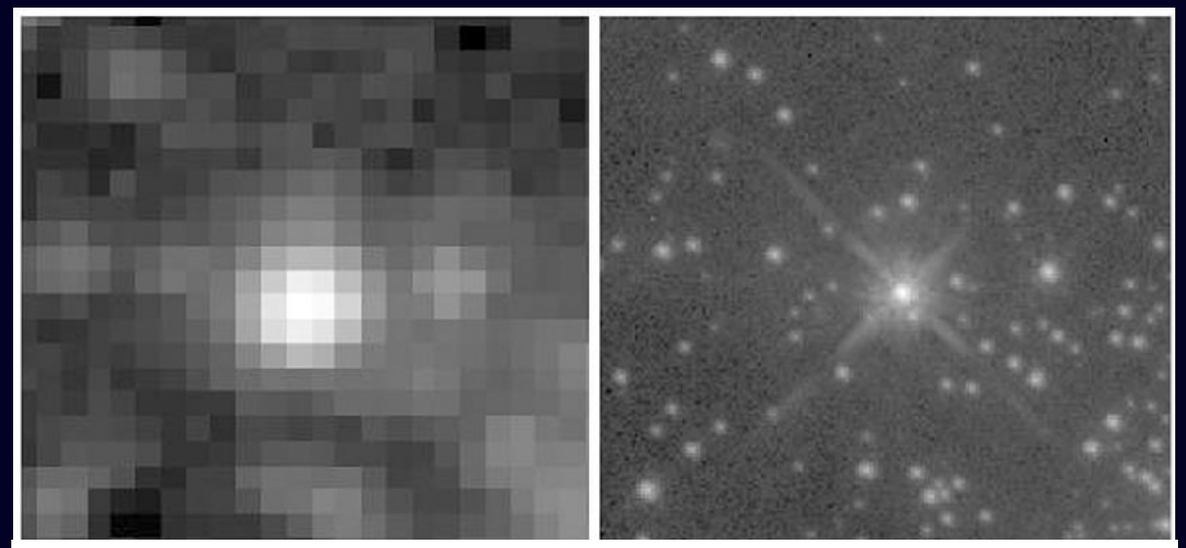
Credit: NASA/ESA, CC BY-SA 4.0

Combining data from various surveys and various wavelengths: caveats

Surveys at the same wavelength but with differing spatial resolution

Spatial resolution improves with technology. A single source may resolve into multiple objects in the higher-resolution image.

IRAS observations from the 1970s vs. WISE data from the 2000s.



Ground-based vs HST image of a field of stars in the 30 Dor region of the LMC

Credit: NASA/ESA, CC BY-SA 4.0

Surveys at different wavelengths

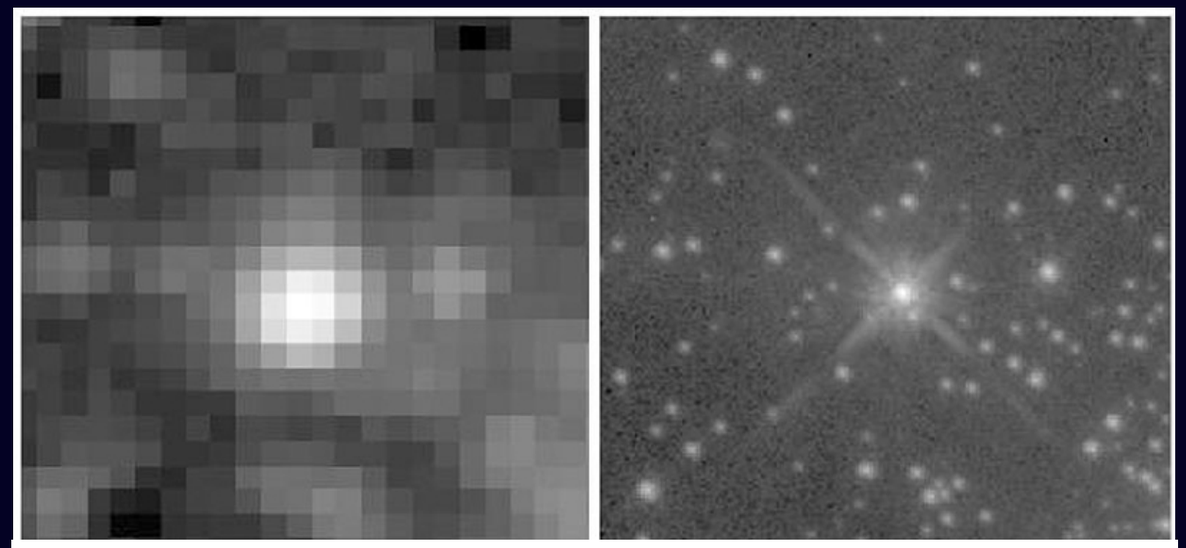
In general, spatial resolution degrades with increasing wavelength. What may look like one object in the mid-IR image may resolve into multiple objects in the optical image!

Combining data from various surveys and various wavelengths: caveats

Surveys at the same wavelength but with differing spatial resolution

Spatial resolution improves with technology. A single source may resolve into multiple objects in the higher-resolution image.

IRAS observations from the 1970s vs. WISE data from the 2000s.



Ground-based vs HST image of a field of stars in the 30 Dor region of the LMC

Credit: NASA/ESA, CC BY-SA 4.0

Surveys at different wavelengths

In general, spatial resolution degrades with increasing wavelength. What may look like one object in the mid-IR image may resolve into multiple objects in the optical image!

Source properties

A very dusty star will appear in the mid-IR image but not in the optical one, and may be misidentified with a nearby blue source in that image.

Cone searches and why they aren’t ideal

Cone searches and why they aren't ideal

Problem: I know the star's position in catalogue #1. I want to find data for the star in catalogue #2 (could be same wavelength or not).

Solution: I search for stars within a circle around that position in catalogue #2. I select the nearest neighbour as the match.

If this match is truly the star from catalogue #1, we call it the *counterpart* in catalogue #2.

Cone searches and why they aren't ideal

Problem: I know the star's position in catalogue #1. I want to find data for the star in catalogue #2 (could be same wavelength or not).

Solution: I search for stars within a circle around that position in catalogue #2. I select the nearest neighbour as the match.

If this match is truly the star from catalogue #1, we call it the *counterpart* in catalogue #2.

Complication: The nearest neighbour is not always the true counterpart due to reasons mentioned on the previous slide.

Cone searches and why they aren't ideal

Problem: I know the star's position in catalogue #1. I want to find data for the star in catalogue #2 (could be same wavelength or not).

Solution: I search for stars within a circle around that position in catalogue #2. I select the nearest neighbour as the match.

If this match is truly the star from catalogue #1, we call it the *counterpart* in catalogue #2.

Complication: The nearest neighbour is not always the true counterpart due to reasons mentioned on the previous slide.

Solution: Use information in addition to just positional matches — e.g., for a red star in the mid-IR, find the reddest match in the optical image.

Cone searches and why they aren't ideal

Problem: I know the star's position in catalogue #1. I want to find data for the star in catalogue #2 (could be same wavelength or not).

Solution: I search for stars within a circle around that position in catalogue #2. I select the nearest neighbour as the match.

If this match is truly the star from catalogue #1, we call it the *counterpart* in catalogue #2.

Complication: The nearest neighbour is not always the true counterpart due to reasons mentioned on the previous slide.

Solution: Use information in addition to just positional matches — e.g., for a red star in the mid-IR, find the reddest match in the optical image.

Complication: Sometimes there can be multiple matches that are equally bright!! Not straightforward to generalise this to large source lists, and not reproducible.

Cone searches and why they aren't ideal

Problem: I know the star's position in catalogue #1. I want to find data for the star in catalogue #2 (could be same wavelength or not).

Solution: I search for stars within a circle around that position in catalogue #2. I select the nearest neighbour as the match.

If this match is truly the star from catalogue #1, we call it the *counterpart* in catalogue #2.

Complication: The nearest neighbour is not always the true counterpart due to reasons mentioned on the previous slide.

Solution: Use information in addition to just positional matches — e.g., for a red star in the mid-IR, find the reddest match in the optical image.

Complication: Sometimes there can be multiple matches that are equally bright!! Not straightforward to generalise this to large source lists, and not reproducible.

Duplication of effort — It is likely that a group of people much smarter than you has identified the true counterpart. Your time is valuable.

Cone searches and why they aren't ideal

Problem: I know the star's position in catalogue #1. I want to find data for the star in catalogue #2 (could be same wavelength or not).

Solution: I search for stars within a circle around that position in catalogue #2. I select the nearest neighbour as the match.

If this match is truly the star from catalogue #1, we call it the *counterpart* in catalogue #2.

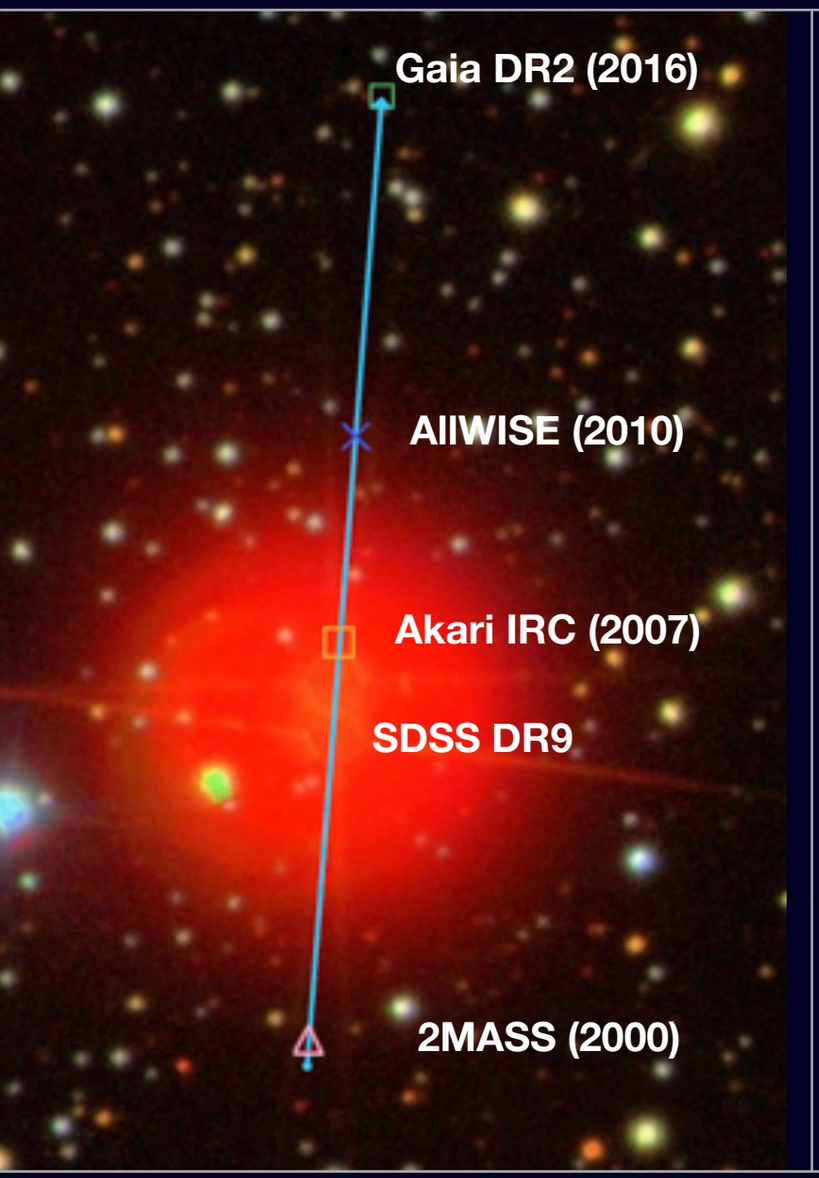
Complication: The nearest neighbour is not always the true counterpart due to reasons mentioned on the previous slide.

Solution: Use information in addition to just positional matches — e.g., for a red star in the mid-IR, find the reddest match in the optical image.

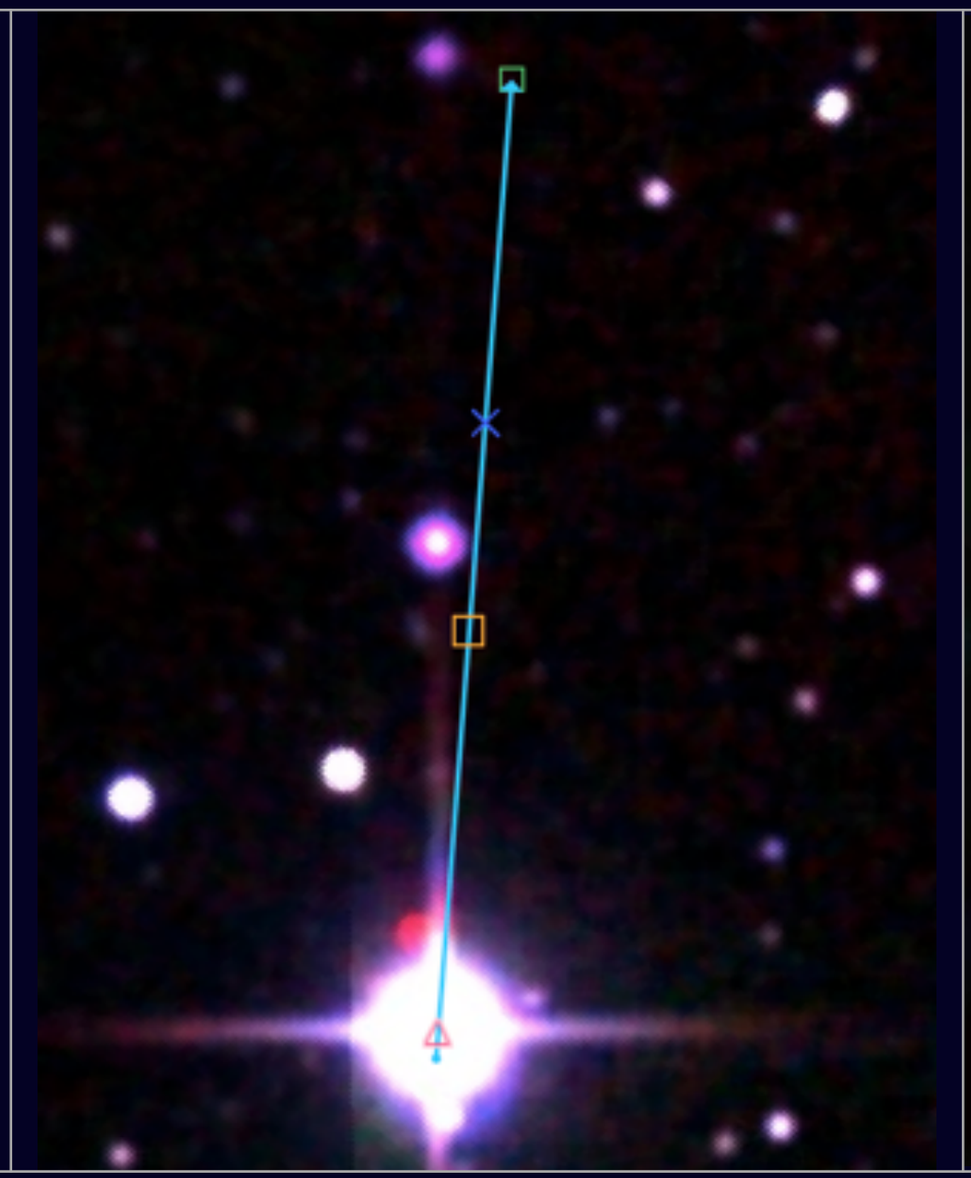
Complication: Sometimes there can be multiple matches that are equally bright!! Not straightforward to generalise this to large source lists, and not reproducible.

Duplication of effort — It is likely that a group of people much smarter than you has identified the true counterpart. Your time is valuable.

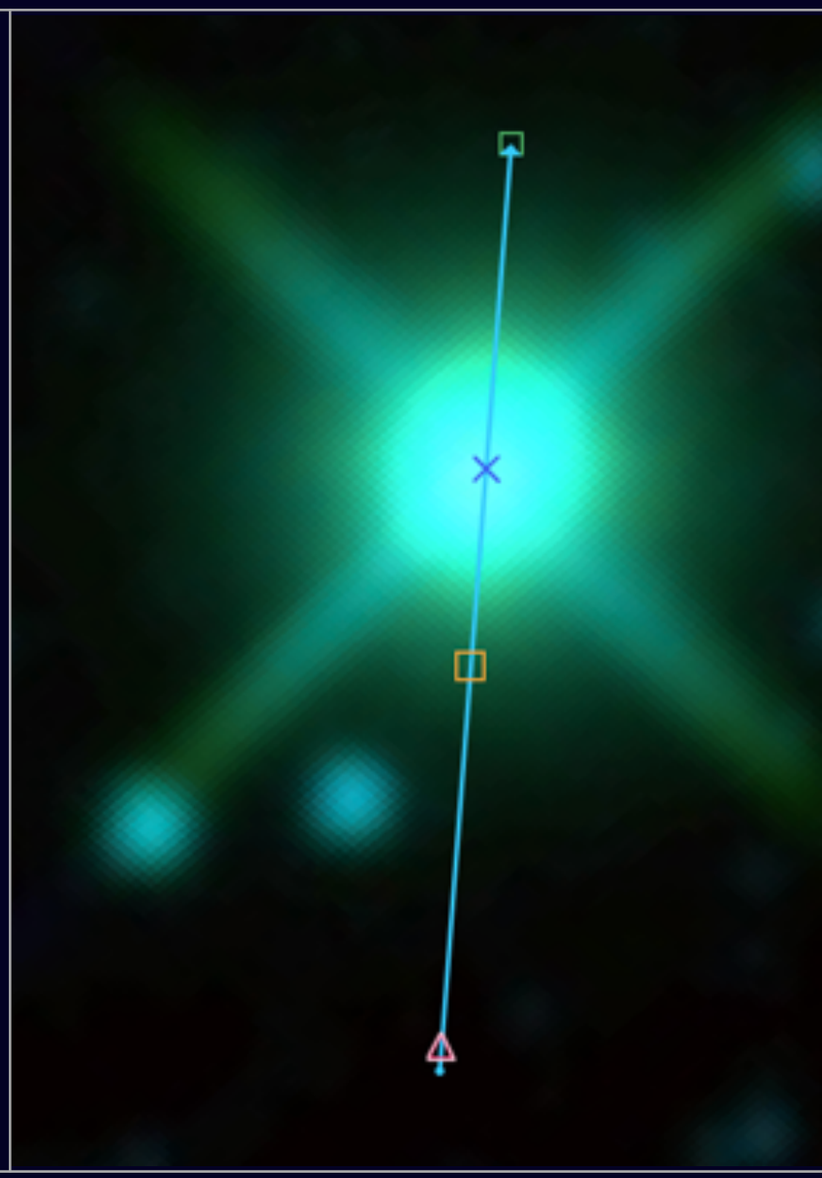
Extreme example: Barnard's Star (proper motion 10″ yr^{-1})



Optical image (SDSS DR9)



Near-IR image (2MASS)



Mid-IR image (AllWISE)

AllWISE beam size: 6″.
Cone search around AllWISE location with 6″ radius will find many matches in both 2MASS and SDSS DR9, but won't find true counterpart in either catalogue!

Alternative to cone searches: unique identifiers

1. Assign a unique identifier to each target/star/observation in a catalogue/ dataset.

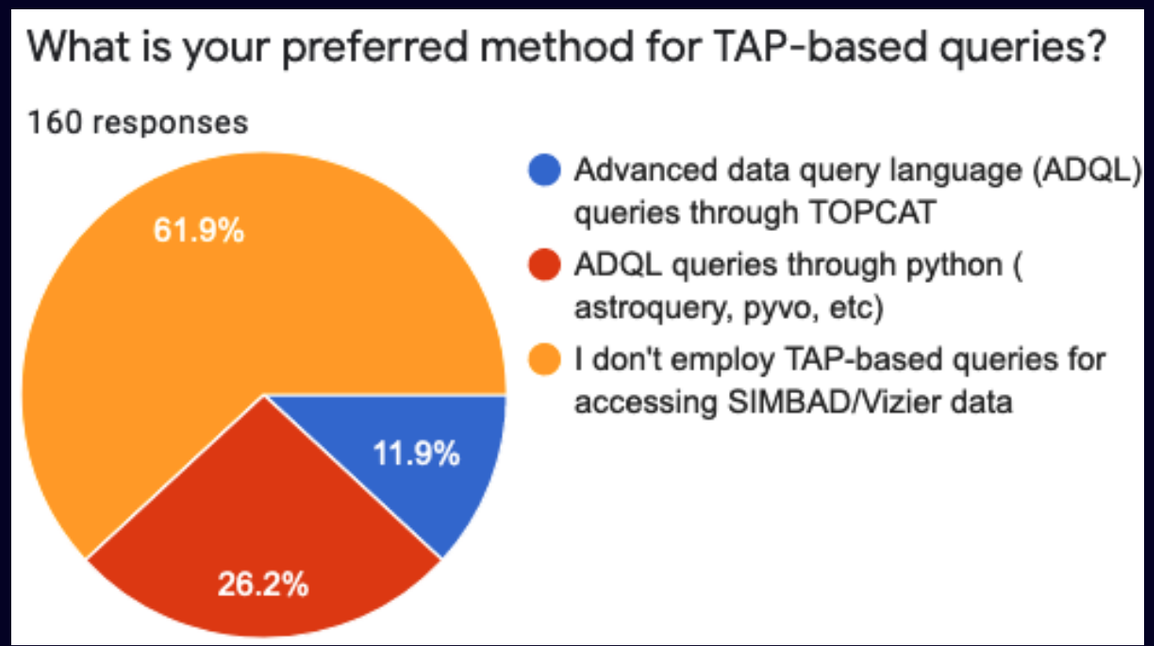
2. When matching two catalogues, produce a table of unique identifiers from catalogue #1 and the matching unique identifier from catalogue #2

3. If you've done the matching correctly (taking into account the problems discussed on previous slides), this work should be made available to the scientific community.

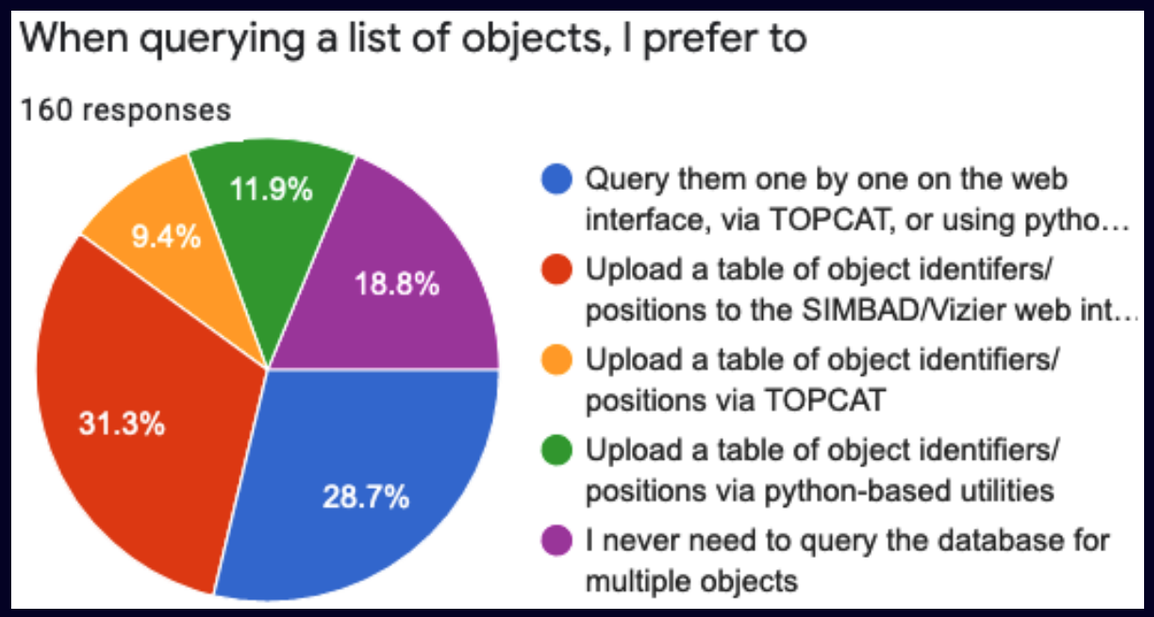
4. Once hosted on an archive, other users can avoid cone searches and simply search for true counterparts in catalogue #2 as long as they use the same unique identifiers as in catalogue #1.

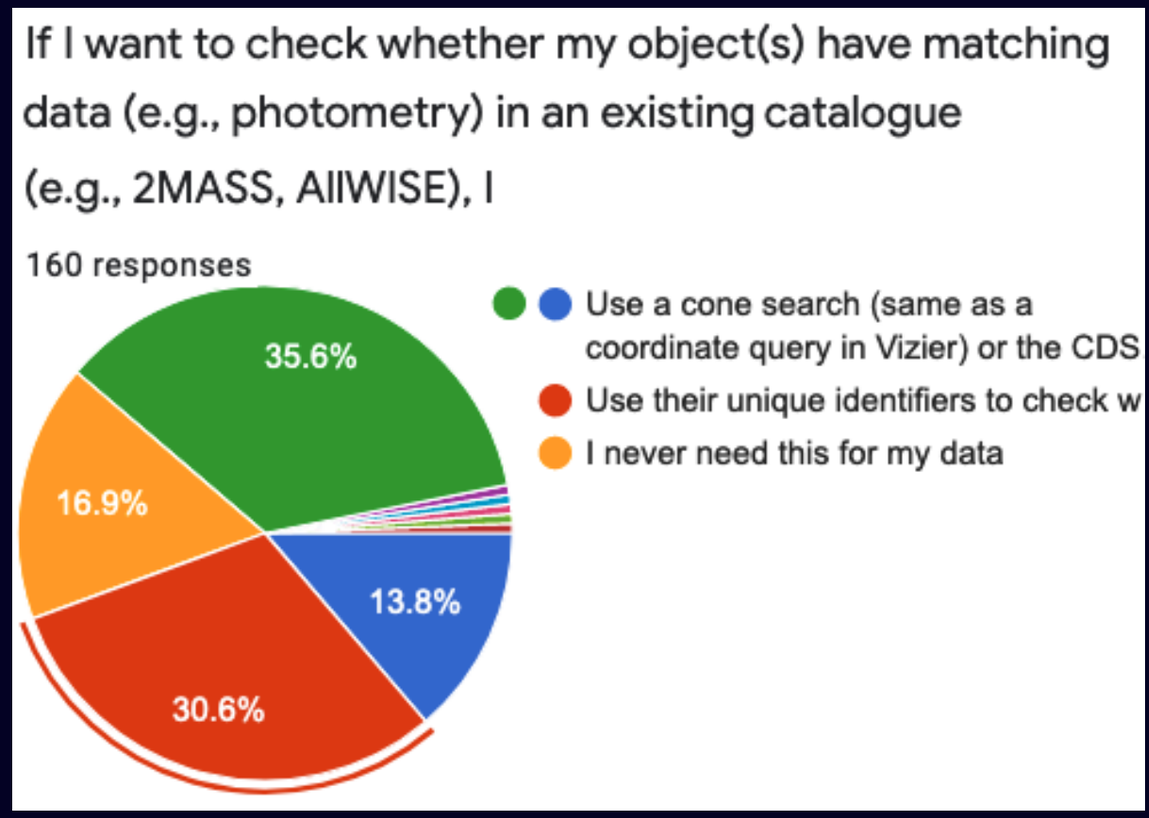
There's a gap in the market that needs filling

Most astronomers don't use TAP queries



Most astronomers would rather perform repetitive tasks than take some time to automate them.





Most astronomers use cone searches (duplication of effort at the very least).

We need you!

Contribute to the science — whether as a student or as a citizen. A degree in astronomy concentrating on results from large datasets will prepare you for careers in science as well as technology! Astrostatistics, big data.

Contribute to the software development — improve your Python skills and prepare yourself for careers in support science and technology! Astroinformatics.

Opportunities (Let's Make Lots of Money)

s.srinivasan@irya.unam.mx, https://bit.ly/32DhjOgOg

The Nearby Evolved Stars Survey (NESS; https://evolvedstars.space)

An inventory of the nearest ~850 evolved stars to study the properties of their circumstellar dust and gas.

How you can help:

- Reduce and analyse sub-millimetre/radio data of nearby (<2 kpc) dusty evolved stars.
- Assist with automated classification of evolved stars from their photometry/spectra.
- Investigate properties of the dust around evolved stars.
- Query large sets of existing data.

Skills you will learn:

- Software (Python code) development.
- Analysis of large datasets of photometry and spectra.
- Querying large astronomical archival databases.
- Astrostatistics, machine learning, data science, reproducible research practices.

These skills will help you with future careers within astronomy as well as in data science!

Some resources

Astronomy using archival data (Y. Wadadekar, Radio Astronomy Winter School 2020, IUCAA)

Astronomy in the Era of Big Data (S. G. Djorgovski, TIARA Summer School 2017)

SIMBAD, Vizier, and Alladin: the CDS astronomical tool suite (Pierre Fernique, New Year Lectures from Astronomical Software Masters)

Virtual Observatory Tools for Astronomers (Justyn Campbell-White, University of Kent)

NASA Virtual Observatory (NAVO) workshop

Citizen science with Zooniverse