Astronomical databases (and how to access them with Python)

Sundar Srinivasan (IRyA/UNAM)

Engineering Week, UACJ, 2021-09-21

Github repository

All files used as part of this presentation are hosted here:

https://github.com/sundarjhu/UACJ_EngineeringWeek2021

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

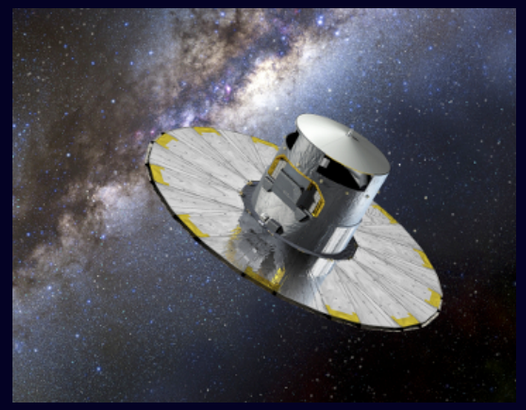
***[Avoiding] Duplication of effort*: Has someone already done this, and is their work readily accessible and reproducible?**

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

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

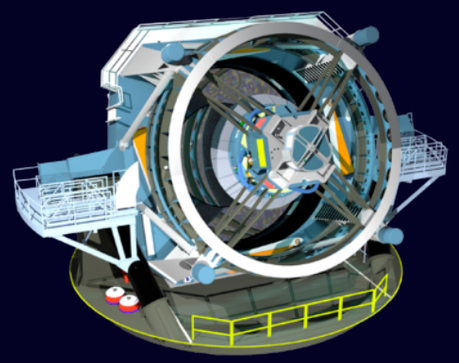
Today's data glut

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



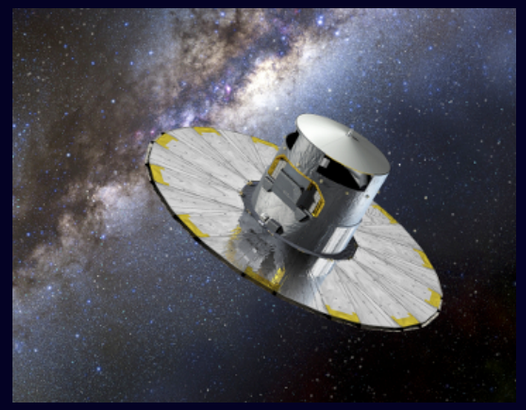
30 TB/night, 100 Gb/s

Square Kilometer Array (2020s)
SKAtelescope.org / CC BY-SA 3.0

~EB of data, ~Pb/s

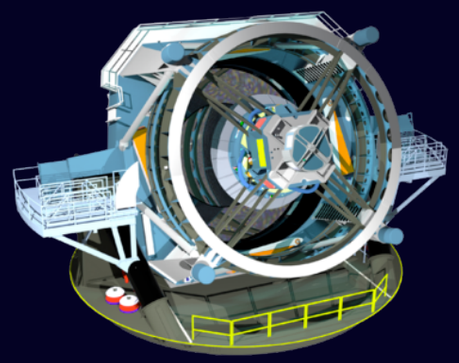
Today's data glut

Gaia (2013–c. 2022)
European Space Agency



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

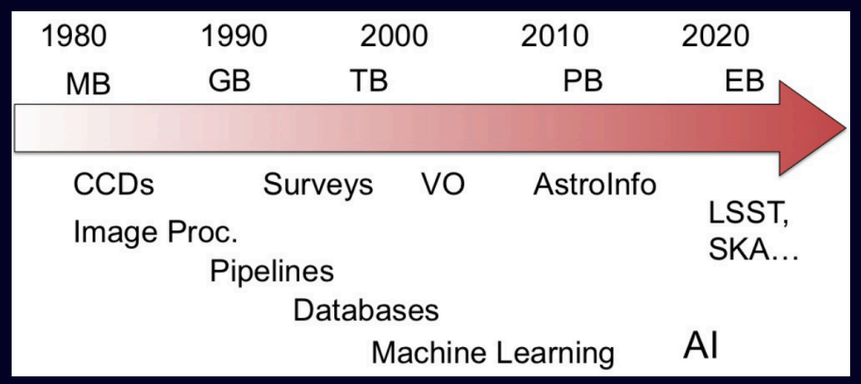
Vera C. Rubin Observatory (2020s)
LSST.org / CC BY-SA 3.0



30 TB/night, 100 Gb/s

Square Kilometer Array (2020s)
SKAtelescope.org / CC BY-SA 3.0

~EB of data, ~Pb/s



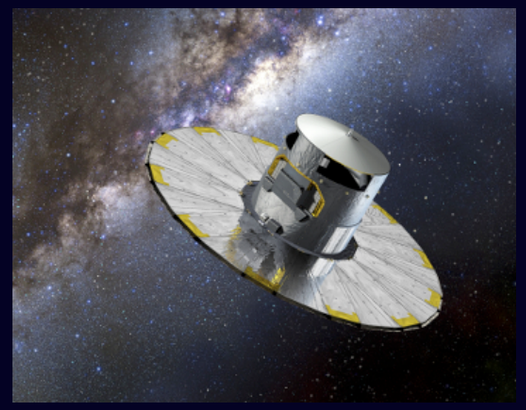
Increasing computing capability, storage, memory, and speed of connectivity has led to increase in quality and quantity of digital archives.

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).

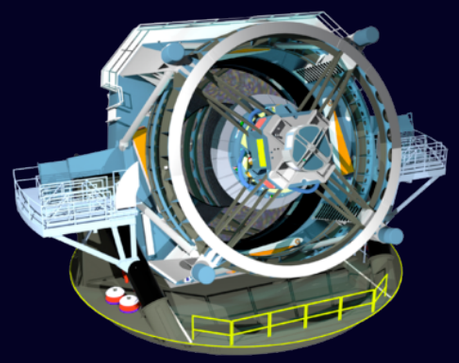
Today's data glut

Gaia (2013–c. 2022)
European Space Agency



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

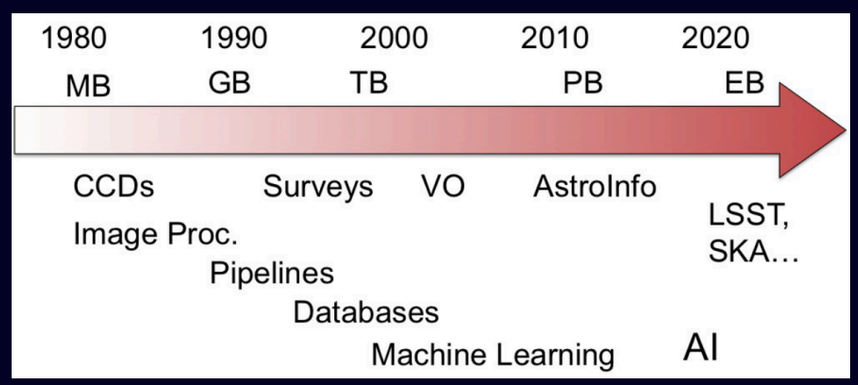
Vera C. Rubin Observatory (2020s)
LSST.org / CC BY-SA 3.0



30 TB/night, 100 Gb/s

Square Kilometer Array (2020s)
SKAtelescope.org / CC BY-SA 3.0

~EB of data, ~Pb/s



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.

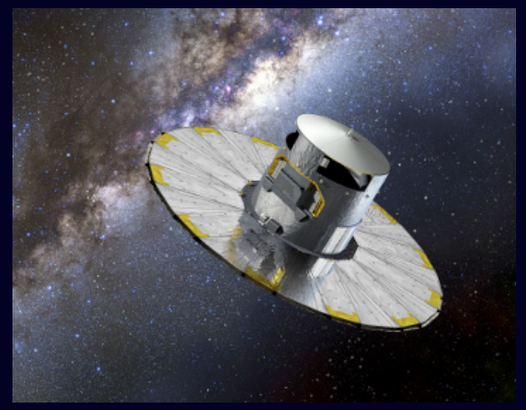
Increase of data-driven science instead of hypothesis-driven science.
Rise of citizen science (e.g., zooniverse.org — classify galaxies on your phone).

Increase in information content: Most data will never be seen by humans.

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

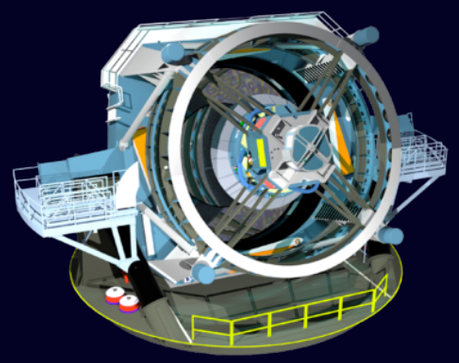
Today's data glut

Gaia (2013–c. 2022)
European Space Agency



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

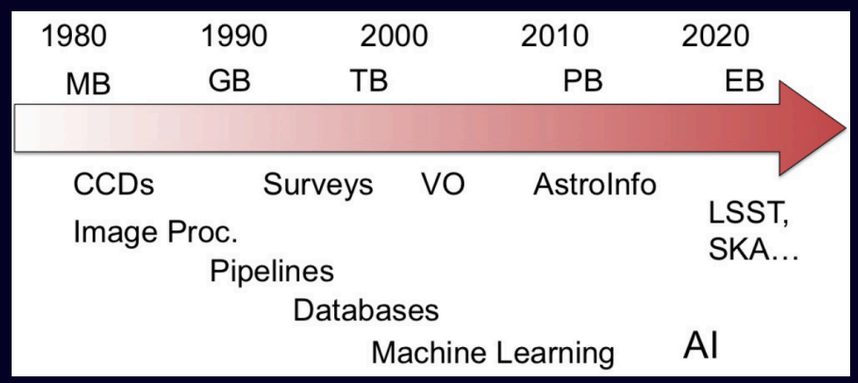
Vera C. Rubin Observatory (2020s)
LSST.org / CC BY-SA 3.0



30 TB/night, 100 Gb/s

Square Kilometer Array (2020s)
SKAtelescope.org / CC BY-SA 3.0

~EB of data, ~Pb/s



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).

Increase in information content: Most data will never be seen by humans.

Increase in 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.

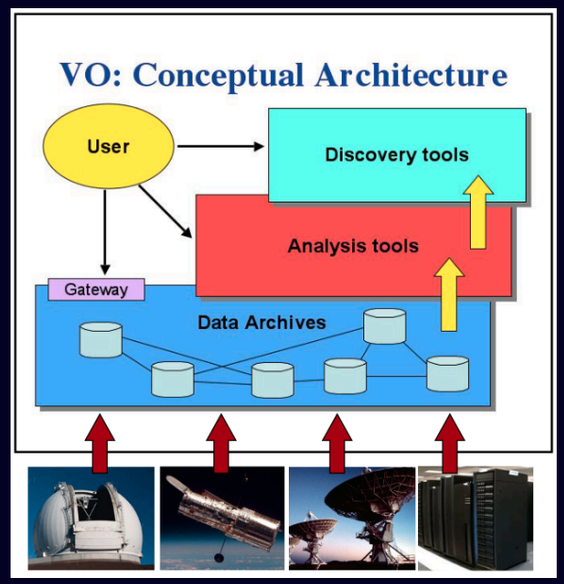
"Interoperability"

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

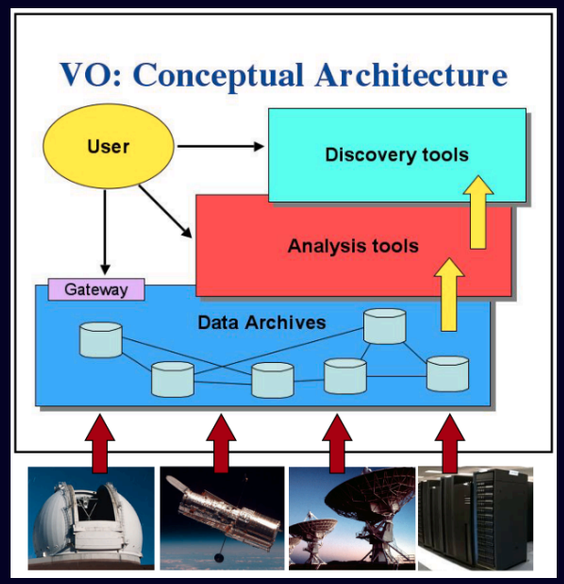


Infrastructure for storage of (and tools for access to) massive/complex datasets.
User-friendly interface allowing data discovery, visualisation, and analysis.

~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



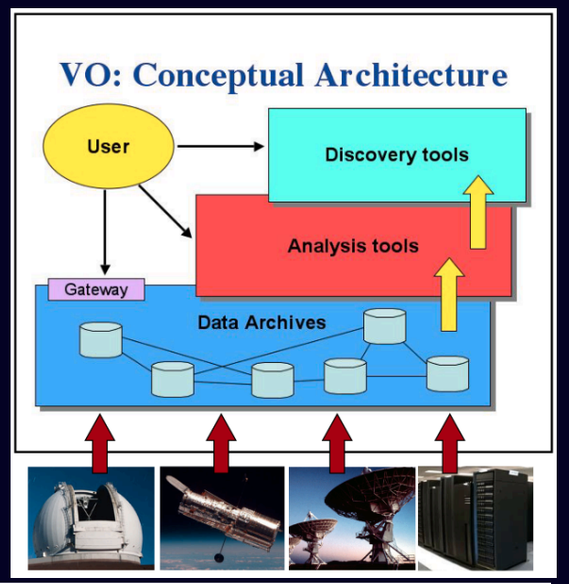
Infrastructure for storage of (and tools for access to) massive/complex datasets.
User-friendly interface allowing data discovery, visualisation, and analysis.

~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

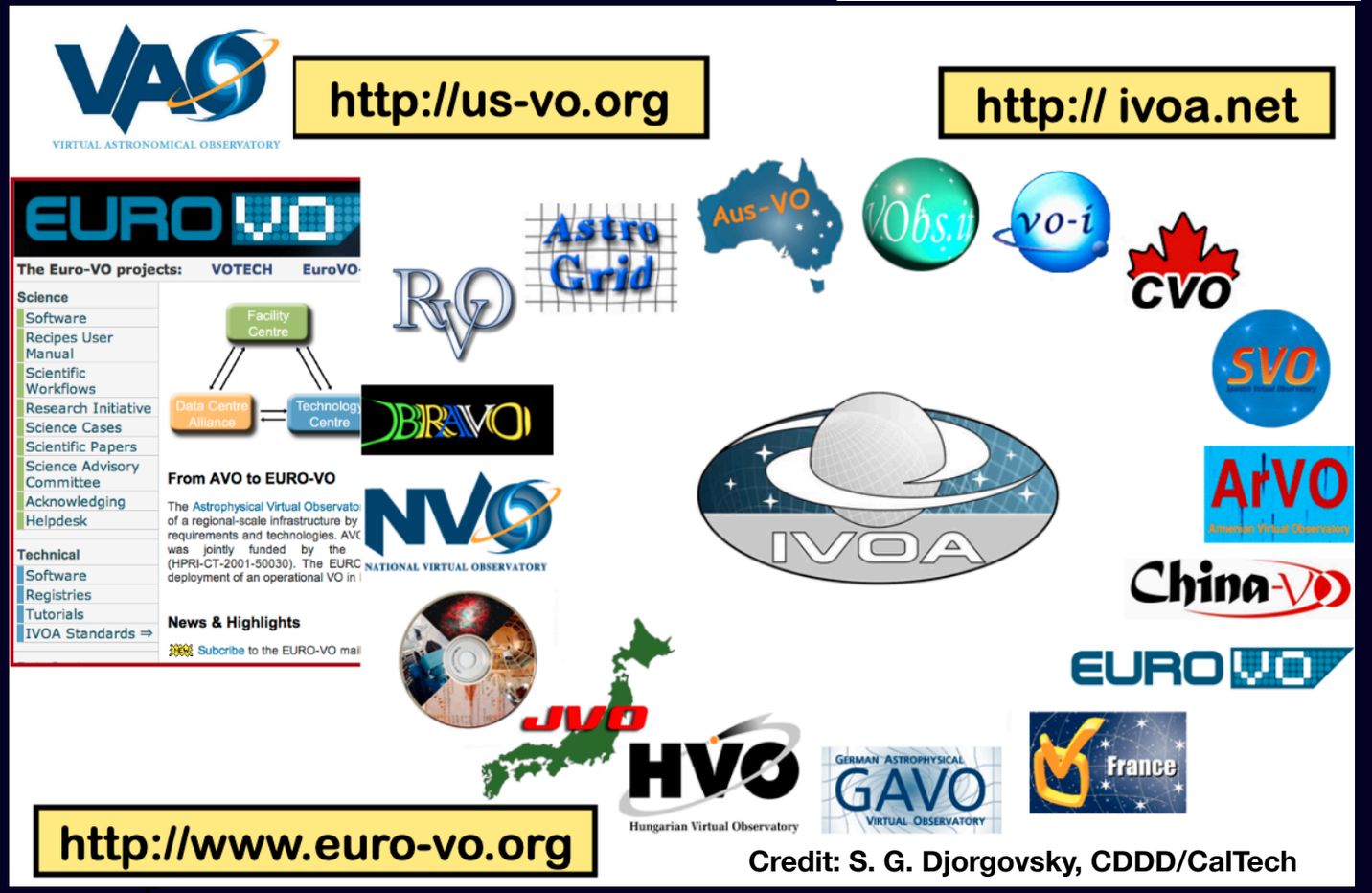


Infrastructure for storage of (and tools for access to) massive/complex datasets.
User-friendly interface allowing data discovery, visualisation, and analysis.

~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/visualise the data

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?"

Published astronomical catalogues:

VizieR

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

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?"

Some other databases:

The InfraRed Space Archive (IRSA; CalTech)

Cosmos (ESA)

The Mikulski Archive for Space Telescopes (MAST; STScI)

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.

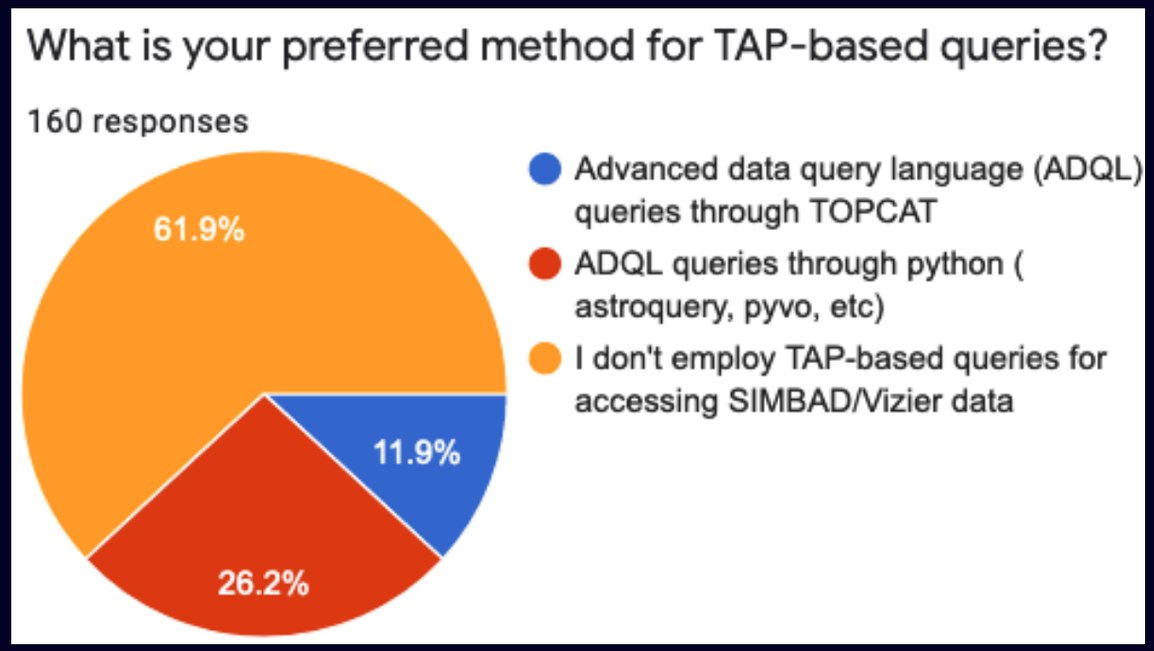
File format

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

There’s a gap in the market that needs filling

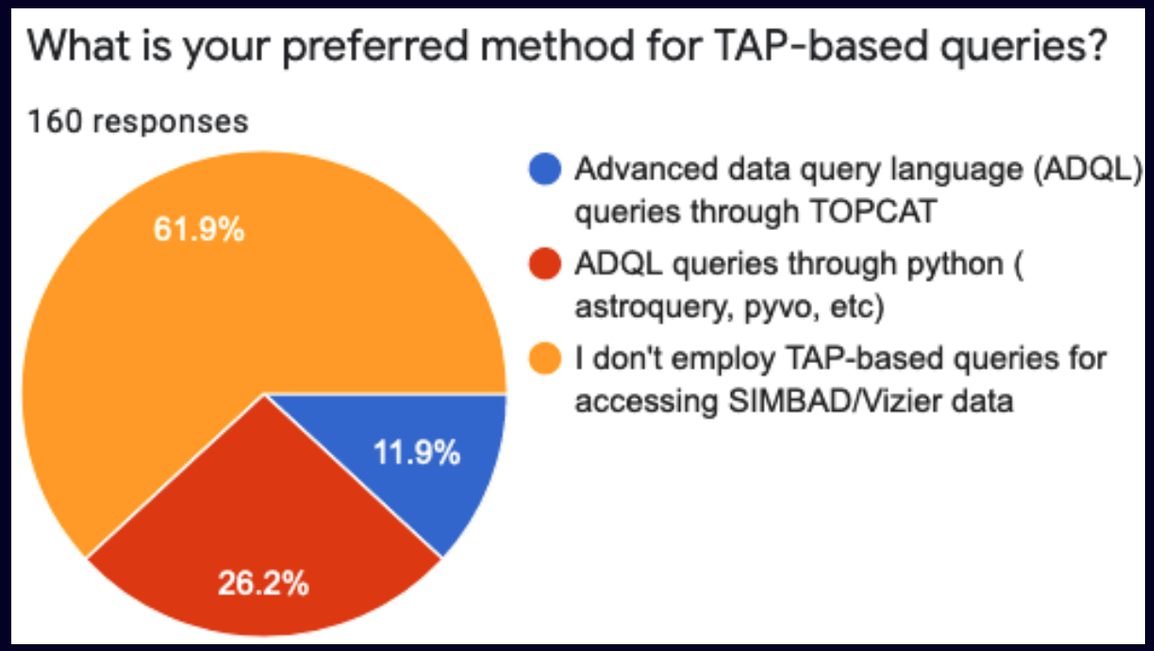
There’s a gap in the market that needs filling

Most astronomers don’t use TAP queries

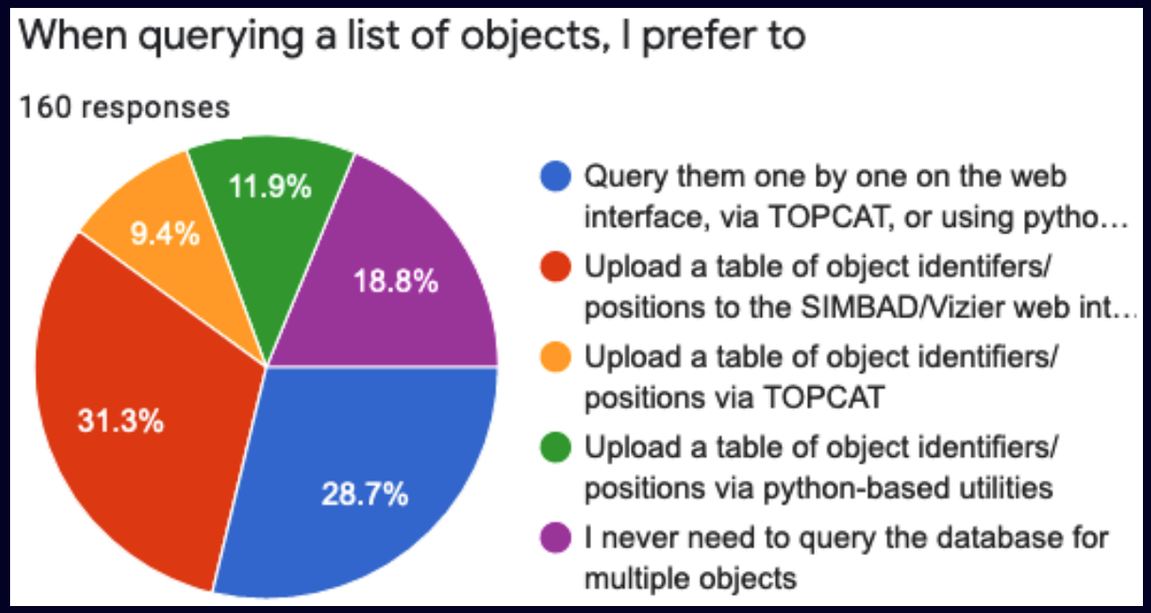


There’s a gap in the market that needs filling

Most astronomers don’t use TAP queries

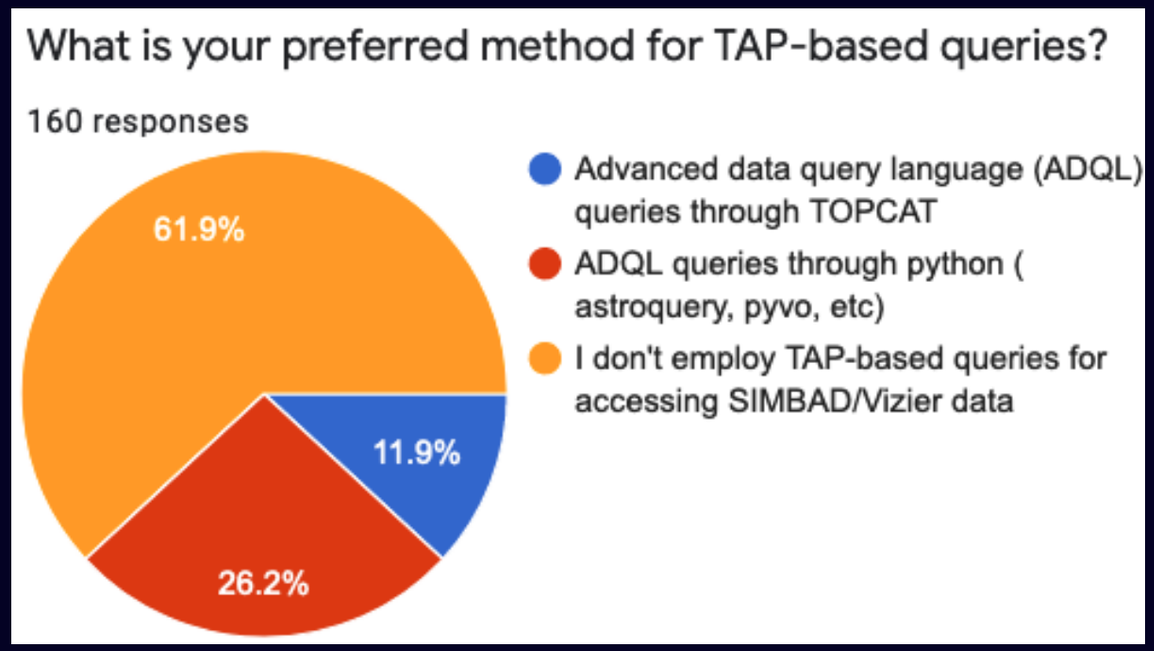


Most astronomers would rather perform repetitive tasks than take some time to automate them.

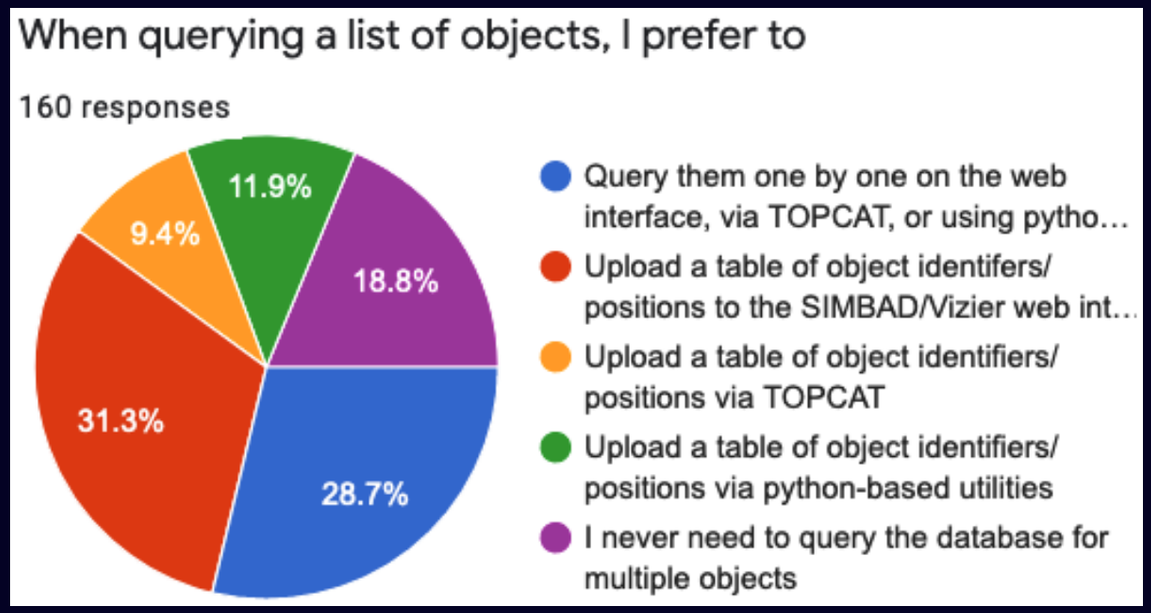


There’s a gap in the market that needs filling

Most astronomers don’t use TAP queries



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



We need you!

Contribute to the science

An astronomy degree concentrating on results from large datasets will prepare you for careers in science as well as technology!
Astrostatistics, big data.

Contribute to 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/32DhjOg

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 (< 3 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

Demo: how to query SIMBAD tables

We will now access Jupyter notebooks from the Github repository.

A similar Jupyter notebook on the Github repository also shows examples of queries to VizieR.