The Old Astronomer
Charlie Bowater, 2016

Astronomical data: new challenges and new approaches

Marcella Massardi

(Italian node of the European ALMA Regional Centre Network
Istituto Nazionale d'Astrofisica - Istituto di Radioastronomia - SISSA)

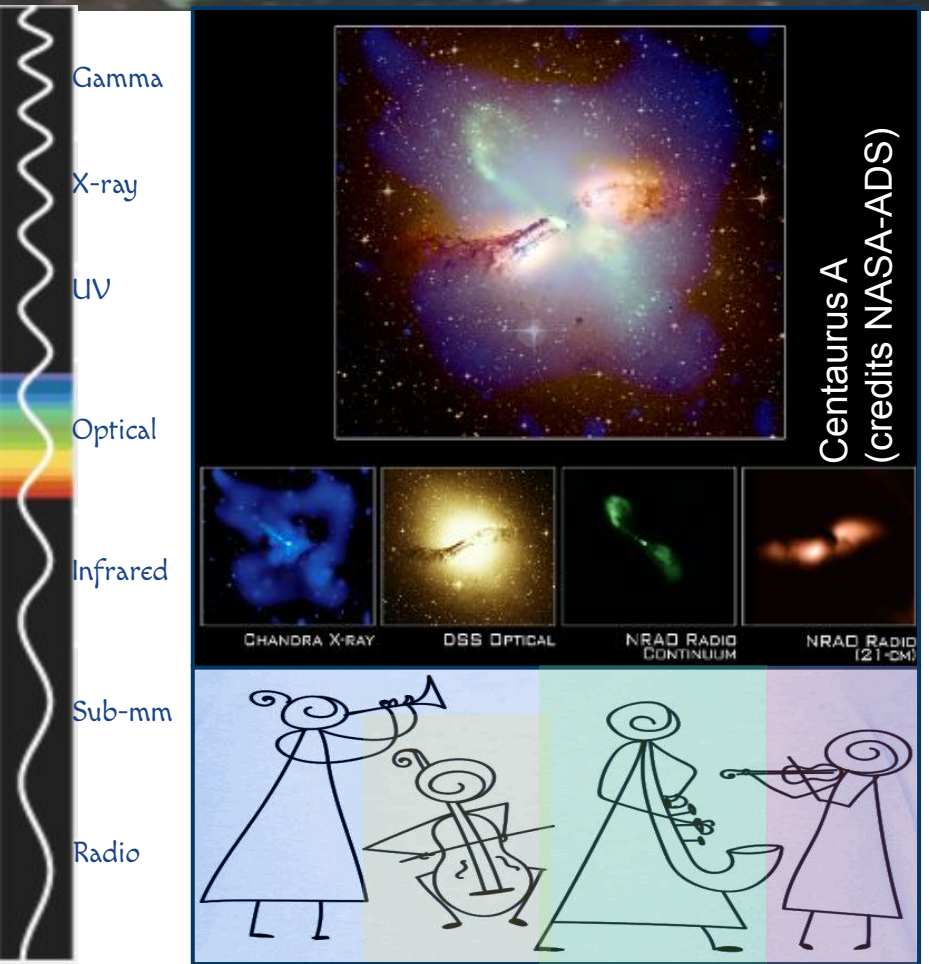
ALMA

EUROPEAN ARC
ALMA Regional Centre || Italian

INAF IRA

SISSA

“Astronomical data”: definition



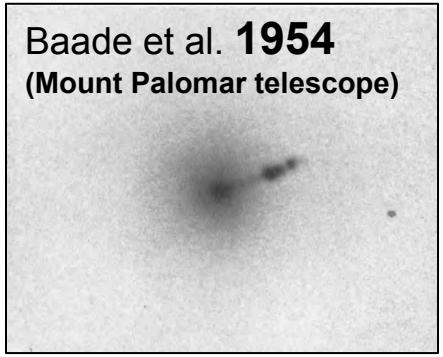
Astronomy combines signals generated by different physical processes, collected with different telescopes in different spectral bands to reconstruct how physics operates in peculiar (not reproducible) environments and across the Universe evolution.

Like an orchestra where you can hear one a few notes of one instrument at the time and you want to reconstruct the whole Symphony.

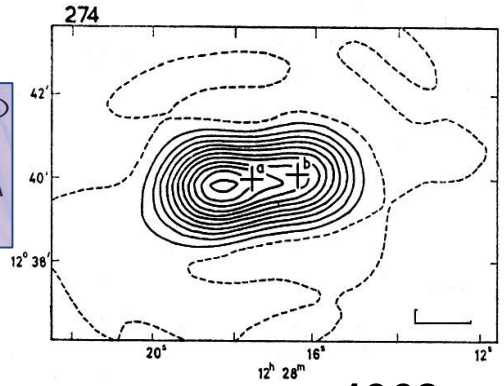
"Astronomical data": evolution (Messier 87)

Optical

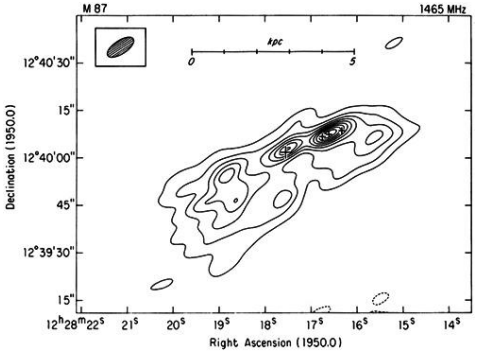
Baade et al. **1954**
(Mount Palomar telescope)



Radio



McDonald et al. **1968**
(One Mile telescope)

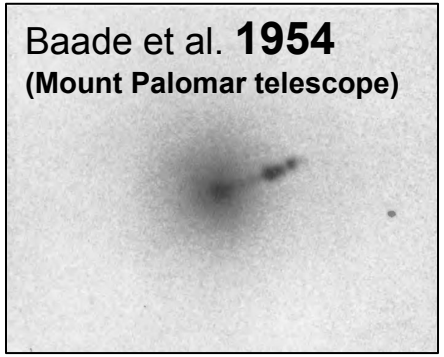


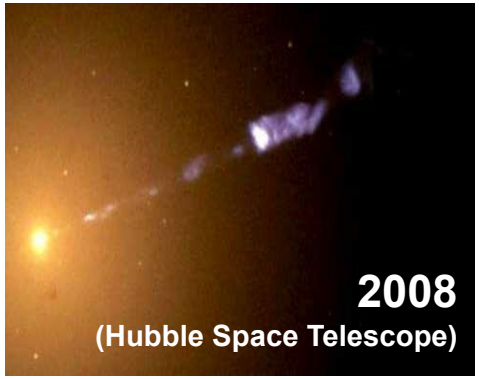
DeYoung et al. **1980**
(Very Large Array)

Existence,
Position,
Structure

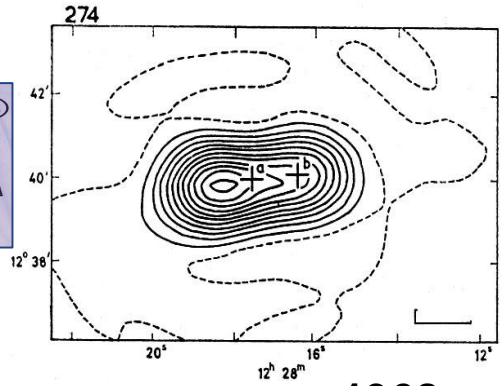
"Astronomical data": evolution (Messier 87)

Optical

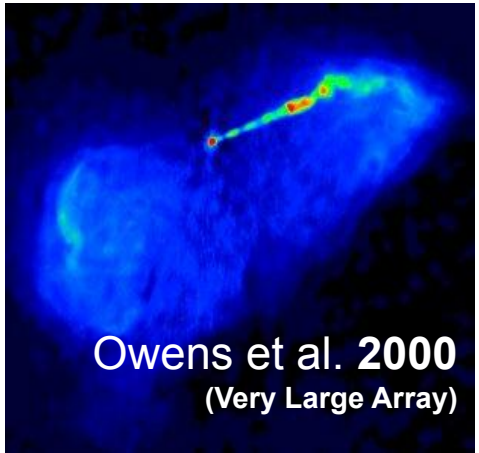




Radio



McDonald et al. **1968**
(One Mile telescope)

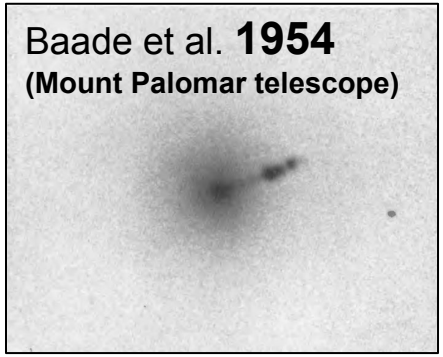


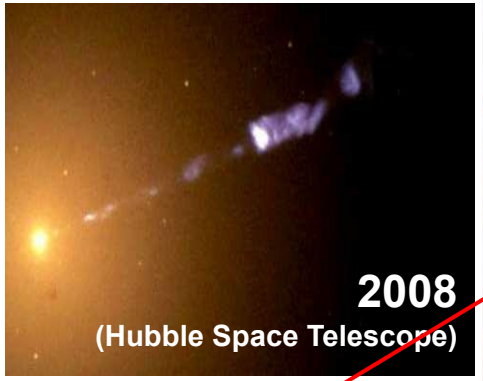
Existence,
Position,
Structure,
Sub-structures,
Chemistry,
Dynamics,

"Astronomical data": evolution (Messier 87)

Optical

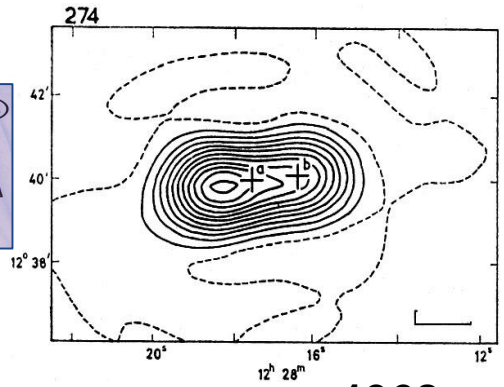
Baade et al. **1954**
(Mount Palomar telescope)



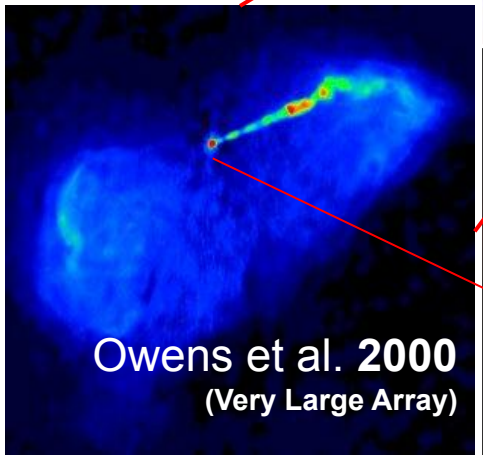


2008
(Hubble Space Telescope)

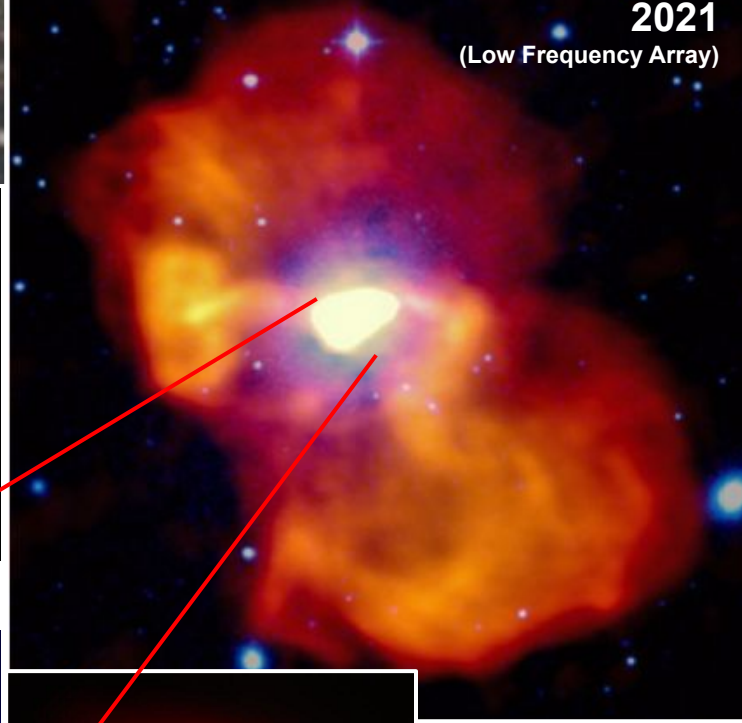
Radio



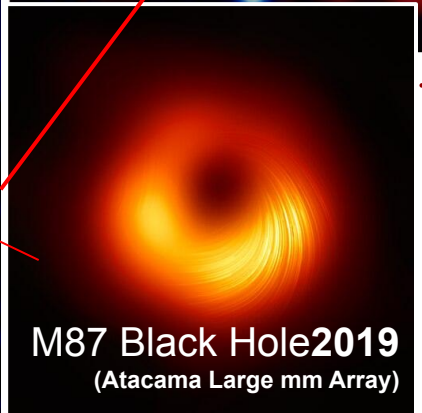
McDonald et al. **1968**
(One Mile telescope)



Owens et al. **2000**
(Very Large Array)



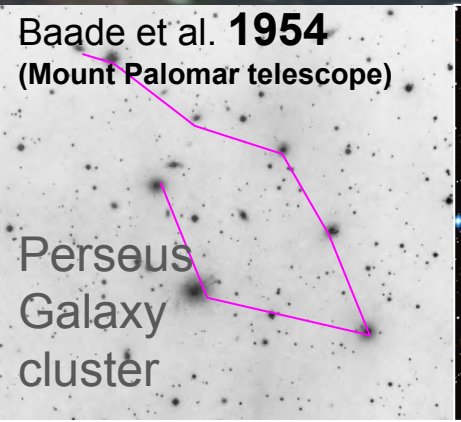
2021
(Low Frequency Array)

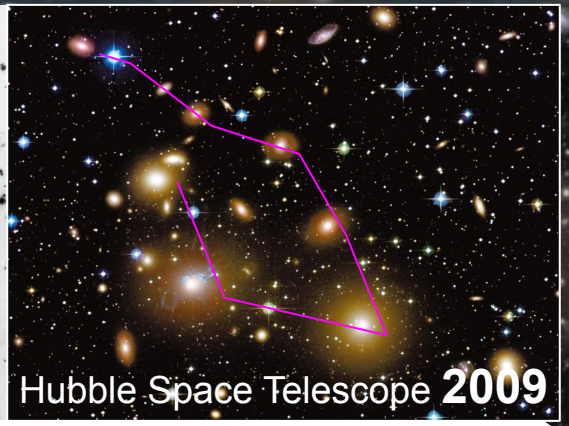


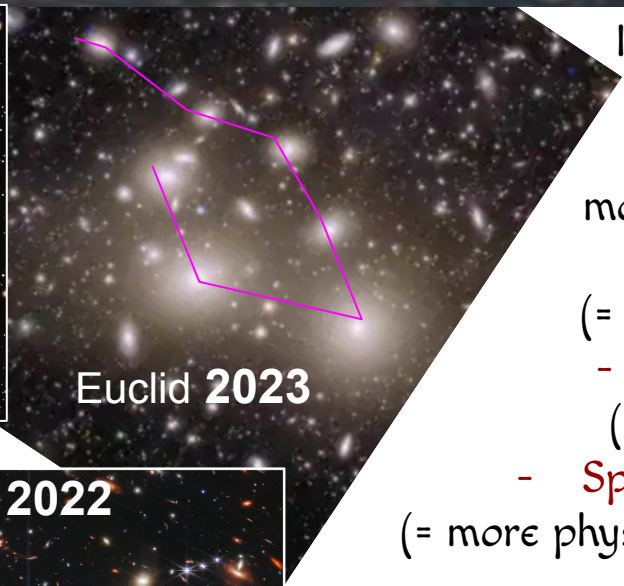
M87 Black Hole**2019**
(Atacama Large mm Array)

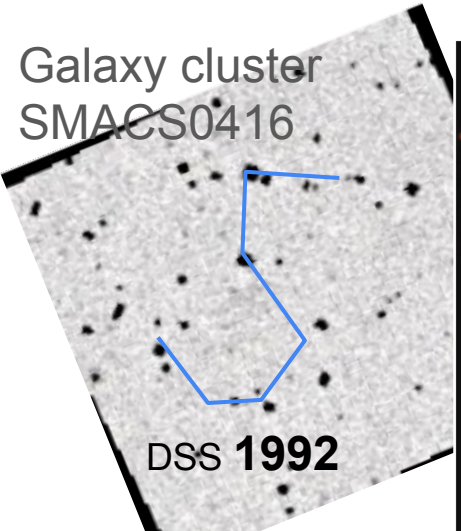
Environment, Details

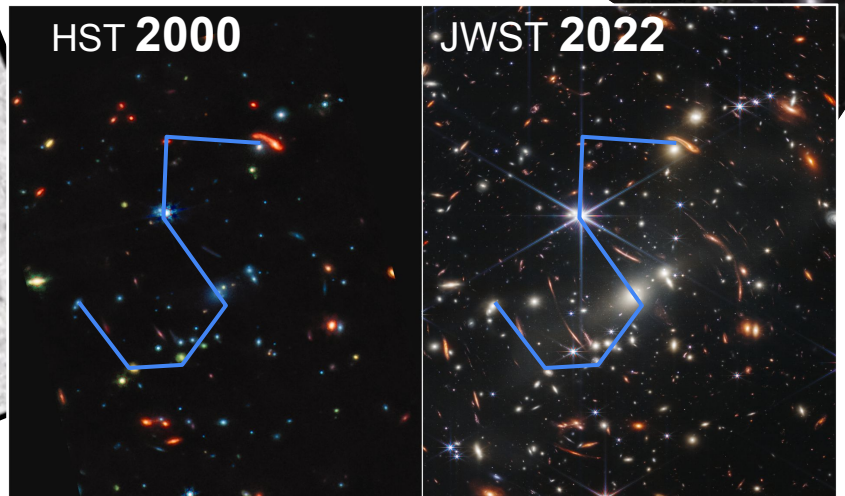
“Astronomical data”: definition





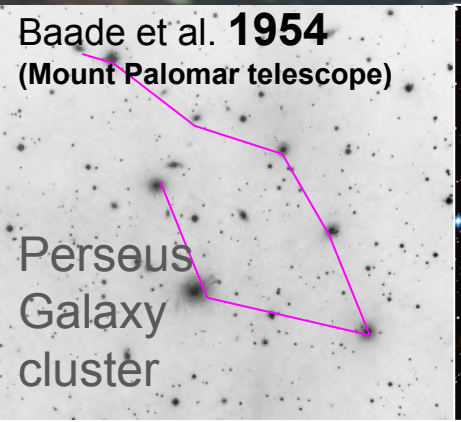


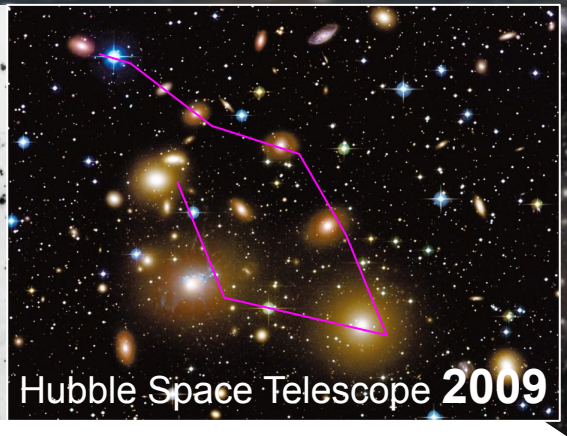


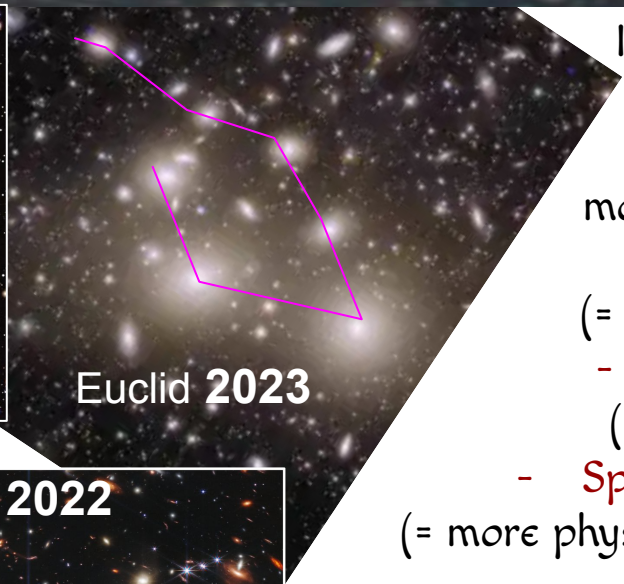


Improvement in
- Sensitivity
(= fainter or more distant obj)
- Resolution
(= smaller details)
- Sky coverage
(= more objects)
- Spectral coverage
(= more physical processes)

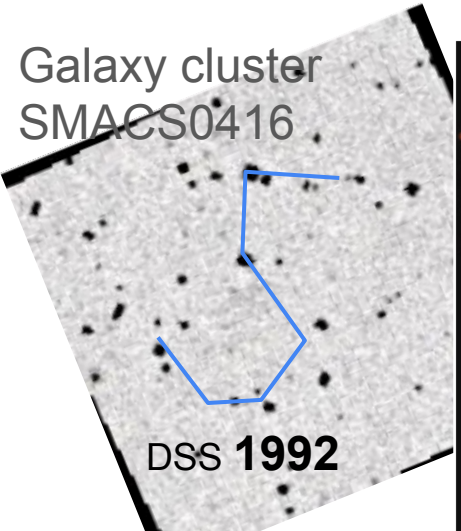
"Astronomical data": definition

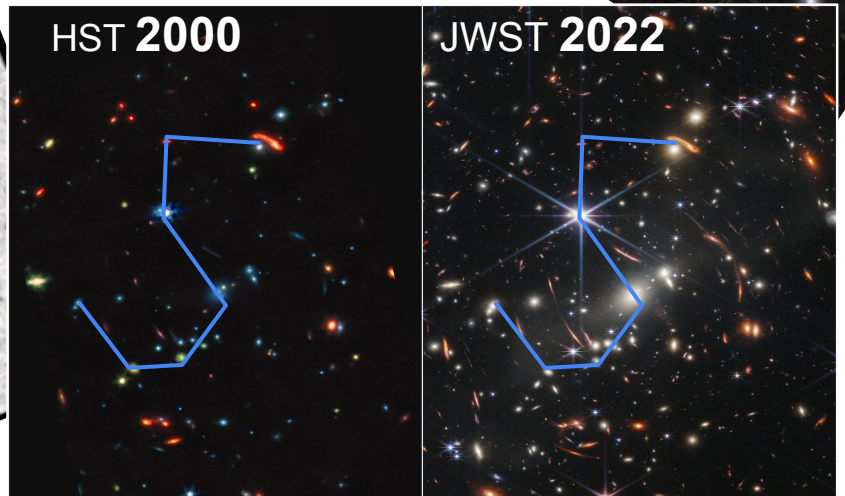






Improvement in
- Sensitivity
(= fainter or more distant obj)
- Resolution
(= smaller details)
- Sky coverage
(= more objects)
- Spectral coverage
(= more physical processes)





Growing data volume!!

Growing data volumes

Past

Typical file ~ 0.3 GB/hr
Stored ~ 30 TB (~40yr)
Growth rate ~ 9 GB/day

50 million AU dollars (only construction)
6 million AU dollars per year

Present

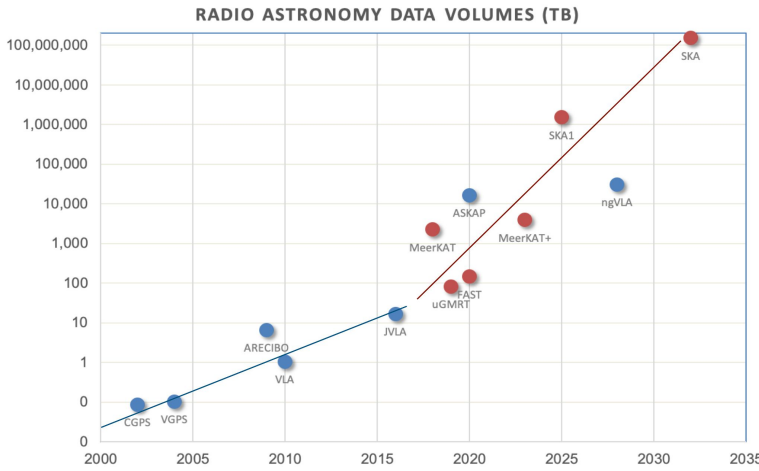
Typical file ~ 0.1 TB/hr
Stored ~ 2 PB (~10yr)
x50-100 by 2030 upgrade!

1.4 billion dollars (only construction)
100 million dollars for global operation per yr

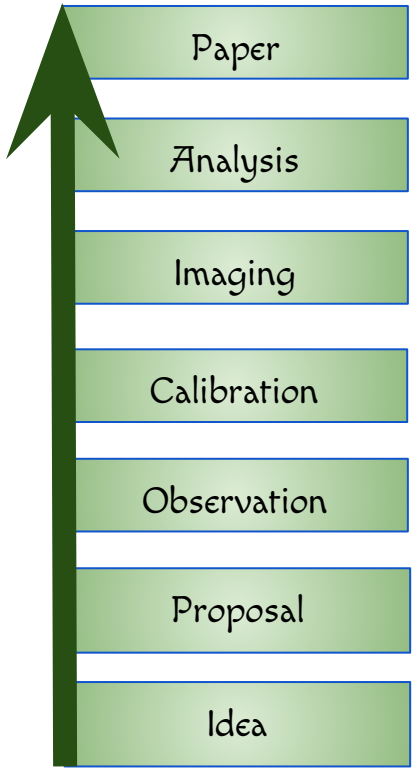
Future

Growth rate ~ 2PB/day
pushed at 100 Gigabit/s

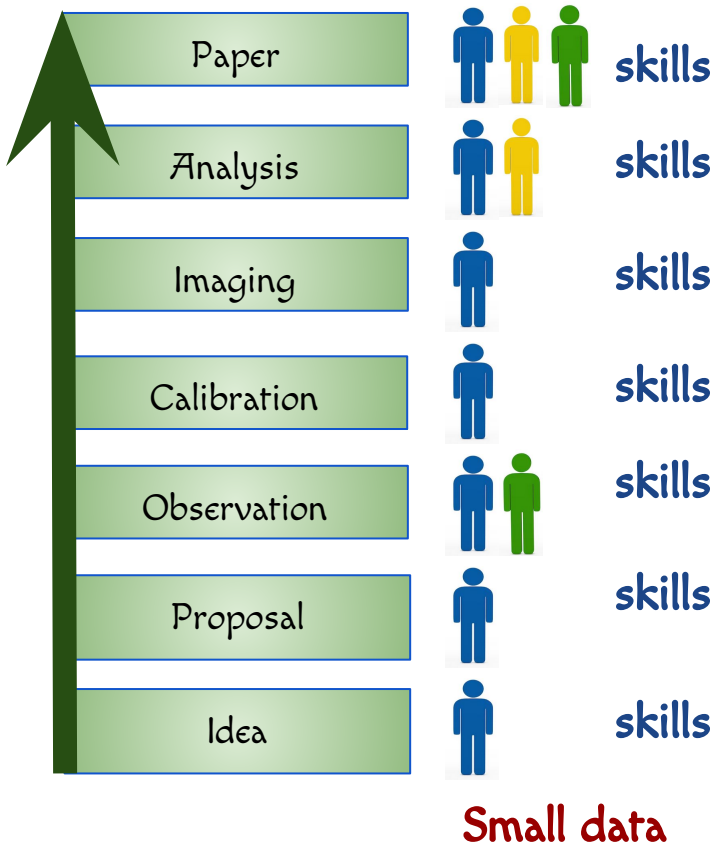
1.9 billion Euros (only construction)
data transfer 2MEuro/yr per 100 Gbit/s



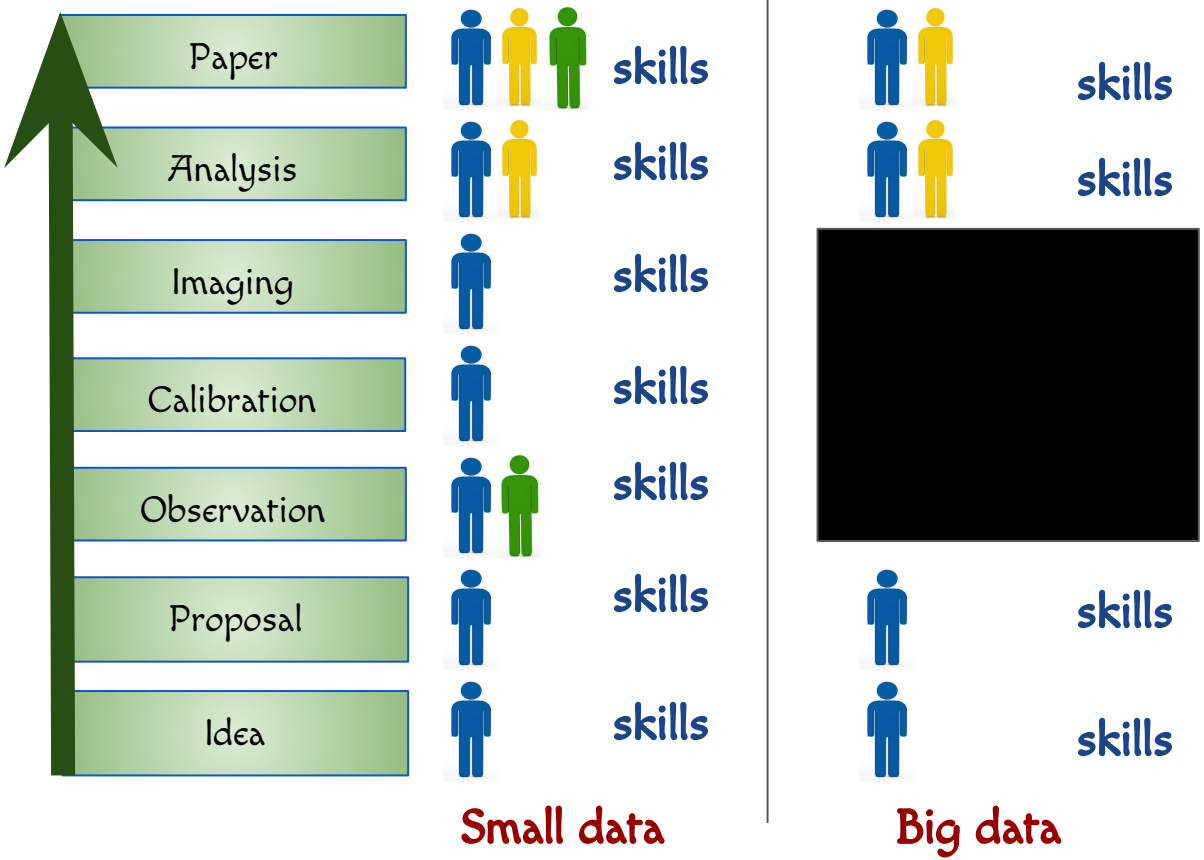
Astronomical Project lifetime



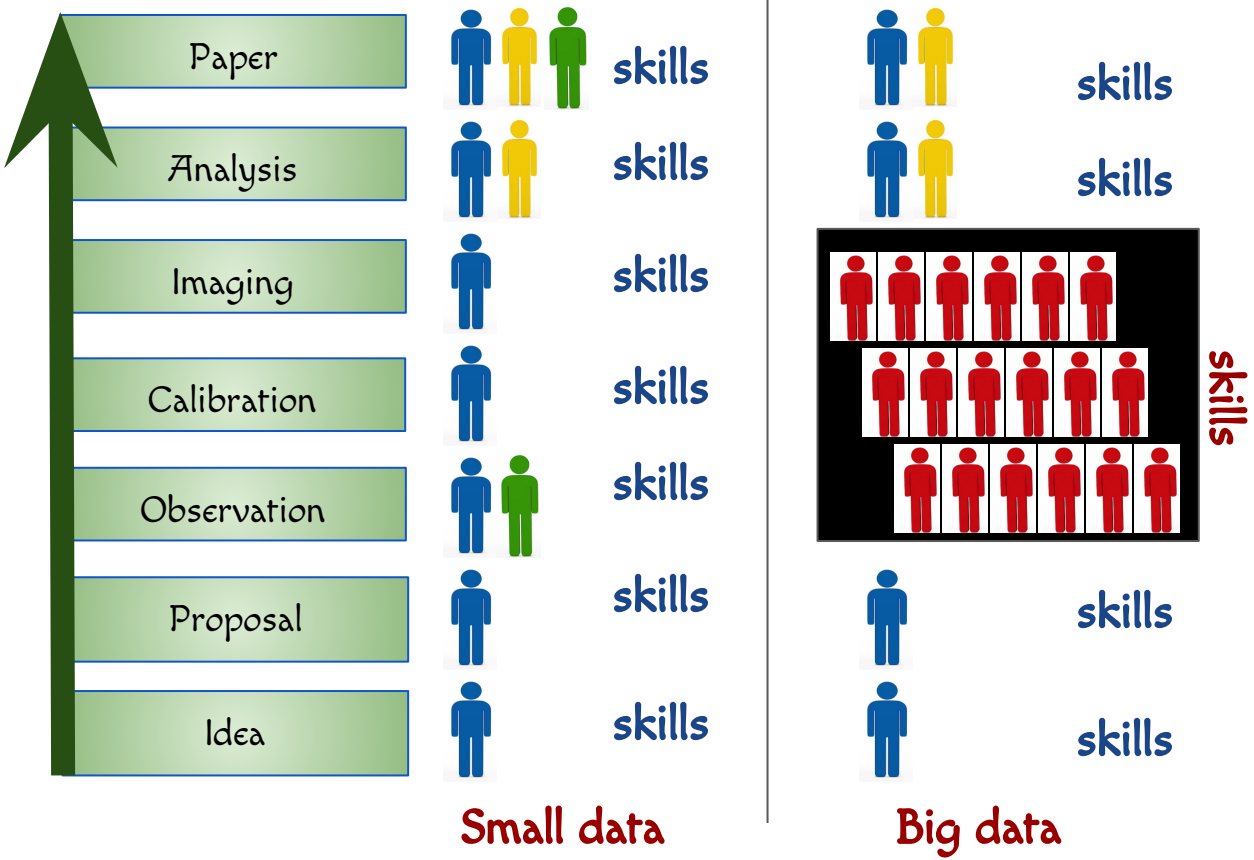
Implications of big data



Implications of big data



Implications of big data

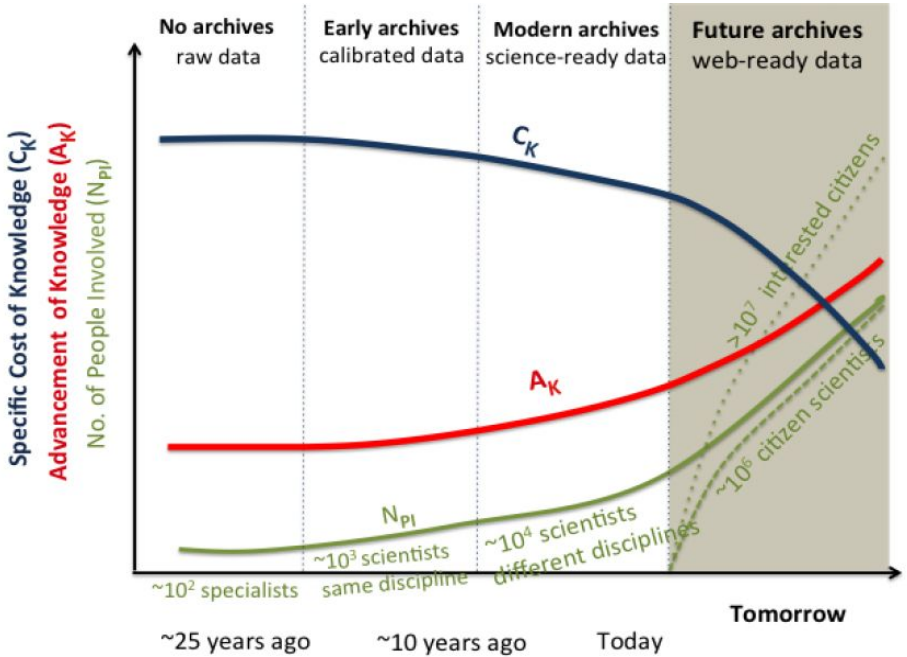


Issues:

- **loss/change of skills**
- **ownership** of data
- maximize data exploitation through **archives** (preservation and integrity)

Trust in the facility

Astronomical Archives

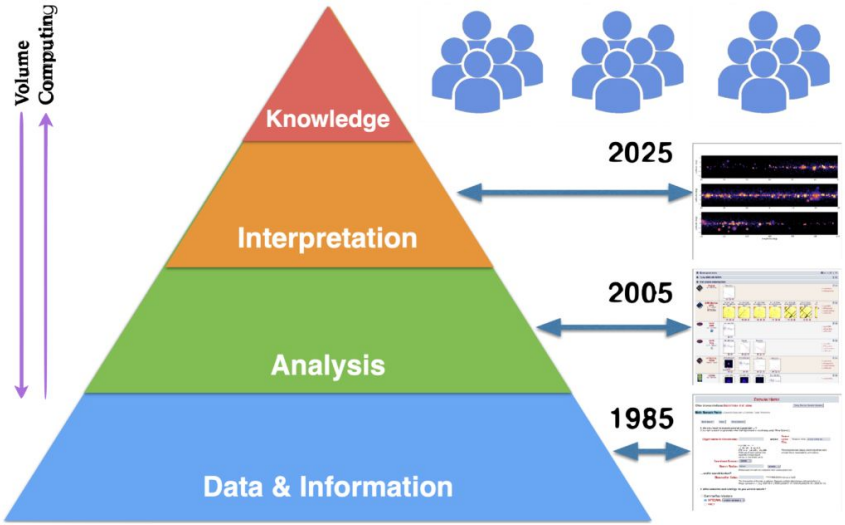


"Open Universe" is an initiative proposed to the United Nations Committee to stimulate a dramatic increase in the availability and usability of space science data, extending the potential of scientific discovery to new participants in all parts of the world.
(Barres de Almeida et al. 2021)

- **Data belongs to the telescopes**
- Investigators have a **proprietary period** for privileged access
- Afterwards data are **public through Archives** for preservation and distribution

- **Each facility manages its archive,** typically with different policies, products and principles
- More recent archives are more than storage units, allowing for interaction and analysis, moving towards **scientific platforms.**

Astronomical Archives



- **The proposal defines the data integrity** limits (i.e. the facility could decide to preserve only portions that are scientifically significant, or maximize the usage preserving the full data)

- **The metadata identifies the proposal idea ownership and data goals** and all the descriptions of observations and of processes applied to the data.

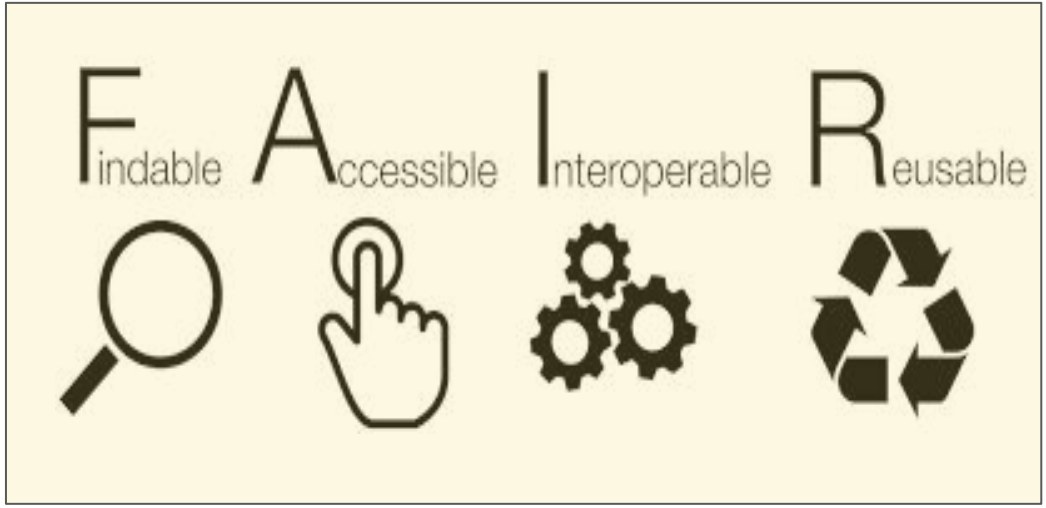
- **Bigger and bigger data require data centers** for processing and new professional figures to guarantee trustability all through the process

Open Science and IVOA principles

Open science *is defined as an inclusive construct … aiming to make multilingual scientific knowledge openly available, accessible and reusable for everyone, to increase scientific collaborations and sharing of information for the benefits of science and society …* ***and it builds on the following key pillars: open scientific knowledge, open science infrastructures, science communication, open engagement of societal actors and open dialogue with other knowledge systems.***
(UNESCO Recommendation on Open Science https://www.unesco.org/en/open-science/about)

The Virtual Observatory (VO) is the vision that astronomical datasets and other resources should work as a seamless whole. Many projects and data centres worldwide are working towards this goal. The International Virtual Observatory Alliance (IVOA) is an organisation that debates and agrees the technical standards that are needed to make the VO possible.

(https://ivoa.net/)



Carbon footprint (ALMA example)

Storage of 1TB of data -> 2000kg/yr of CO2
Transfer of 1 GB of data -> 3kg

A median dataset in ALMA archive has 100 GB size

-> 1yr storage generates 200kg of CO2 per copy (3Archive +1PI)
-> at least 300kg per each data transfer (at least 6 times)

A 100 GB dataset in ALMA Archive in 5 yr generates 13000kg of CO2
corresponding to CO2 generated by 3 cars driven continuously per 1yr
CO2 absorbed by 215 trees in 10 yr

In the ALMA archive we have more than 60000 datasets (10GB-1TB size)
= CO2 absorbed by trees covering the whole area of Florence for 10yr

Only less than 20% of the archive has an associated publication!!!
By 2030 data size is expected to grow by a factor x50–100

RESEARCH DATA USAGE MUST BE MAXIMIZED

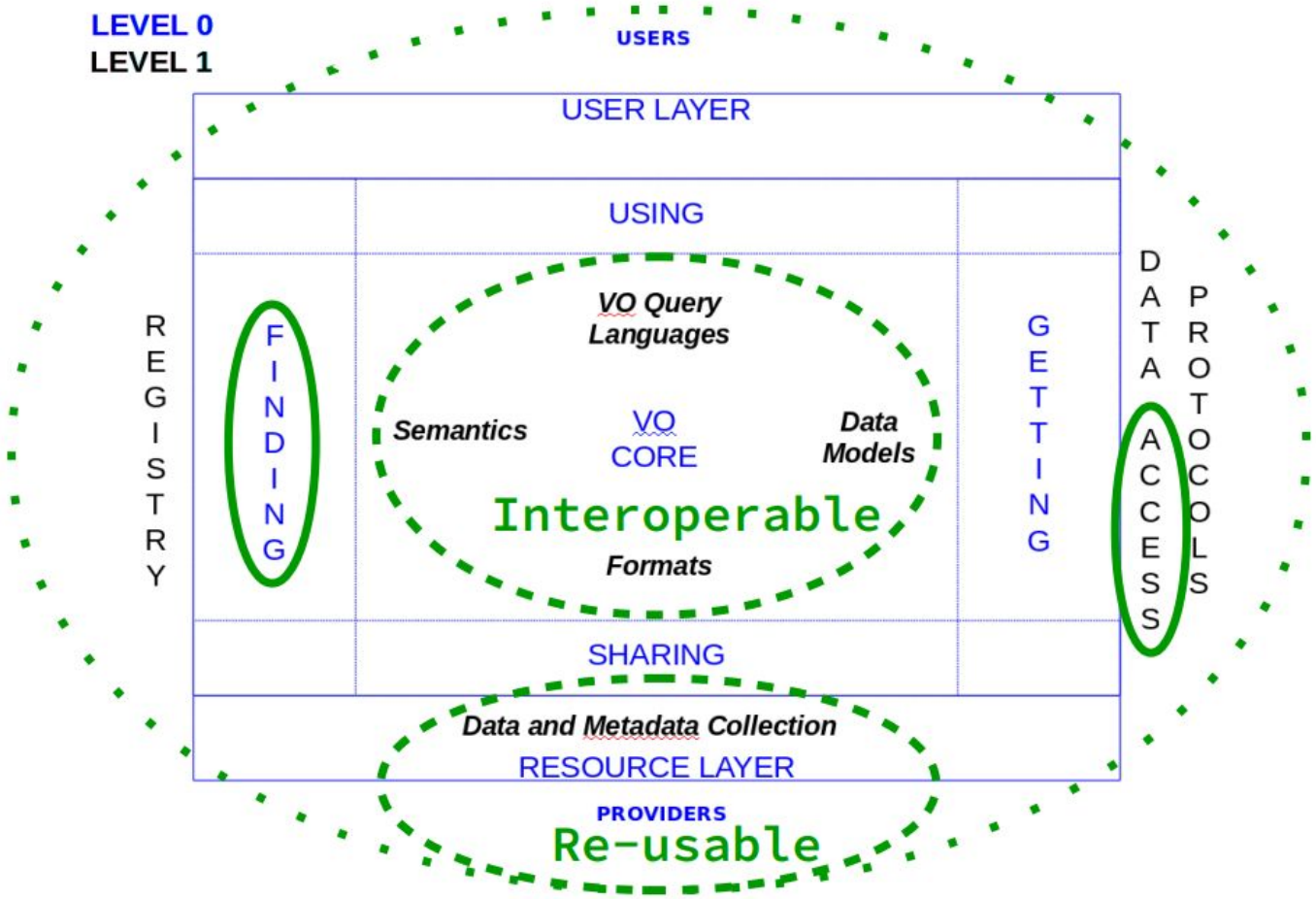
Summary

The evolution in data size implies/requires

- improvements in our research results
- change of framework
- change of mentality in the research community
- opening to collaborative approach/commensality

- more responsibility in building the environment
    - user friendly infrastructures
    - FAIR principles
    - sustainability in the process
    - new professional figures
        (with proper career paths)

IVOA


LEVEL 0
LEVEL 1
USERS
USER LAYER
USING
REGISTRY
FINDING
VO Query Languages
Semantics
VO CORE
Data Models
Interoperable
Formats
GETTING
DATA ACCESS
PROTOCOLS
SHARING
Data and Metadata Collection
RESOURCE LAYER
PROVIDERS
Re-usable