Advanced Data Products
A perspective

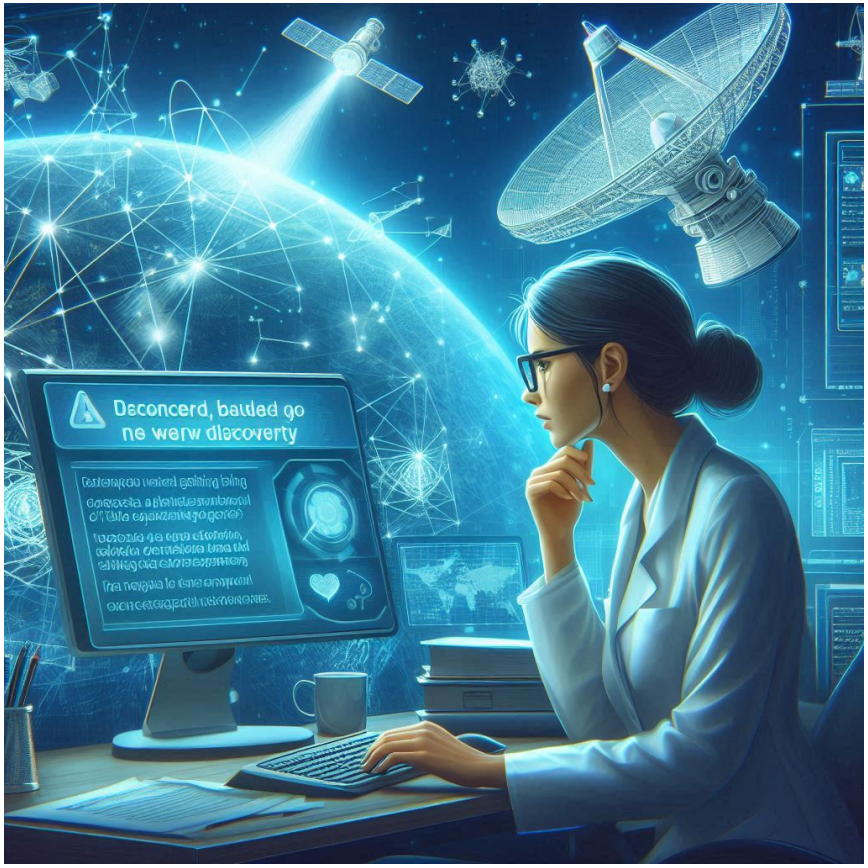
Eric Emsellem

Long term Perspective

Science Discovery Toolkit Search
Real-time Cosmic event detection system
Multi-dimensional astroinformatics layer

Radical transformation
of our understanding of the cosmos

AI-driven astronomical data systems
⇒ real-time and continuous data collecting, processing, interpreting
⇒ discovery alerts, suggests new theories

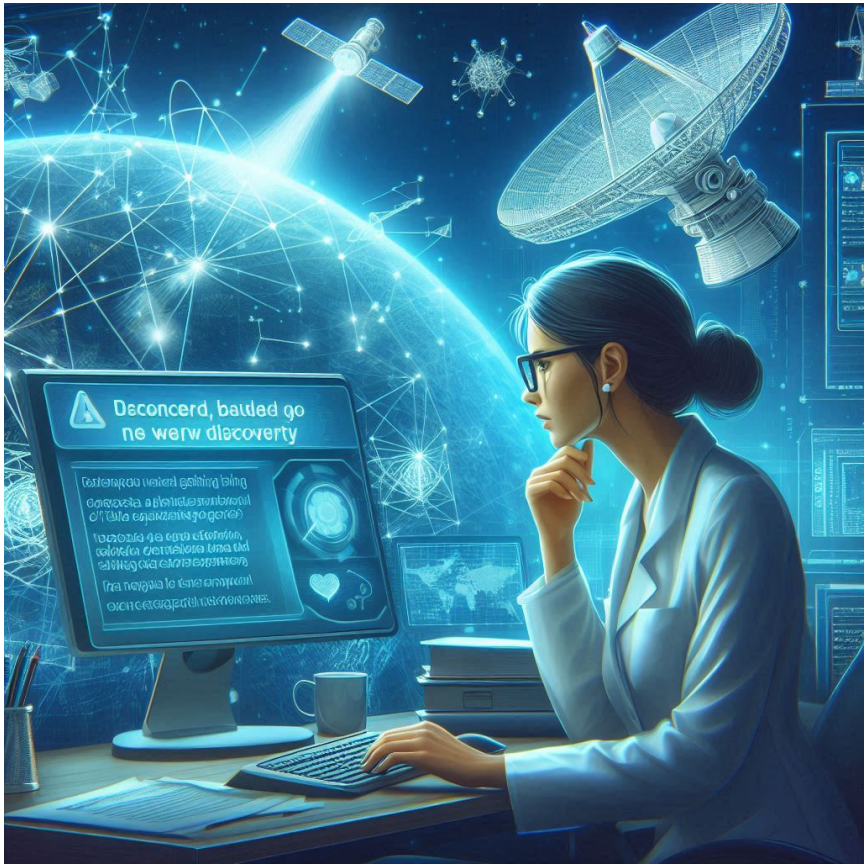


Long term Perspective

Science Discovery Toolkit Search
Real-time Cosmic event detection system
Multi-dimensional astroinformatics layer

Radical transformation
of our understanding of the cosmos

AI-driven astronomical data systems
⇒ real-time and continuous data collecting, processing, interpreting
⇒ discovery alerts, suggests new theories



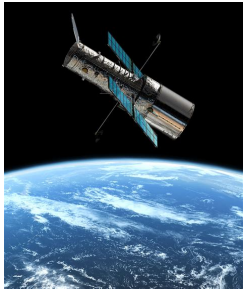
Thanks ChatGPT but ...

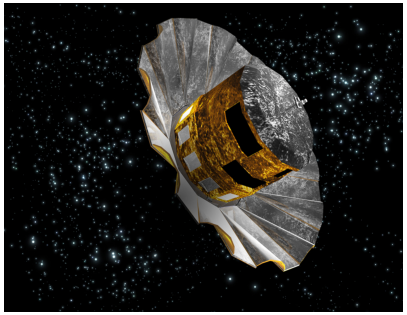
AdP :
what do
you mean ?

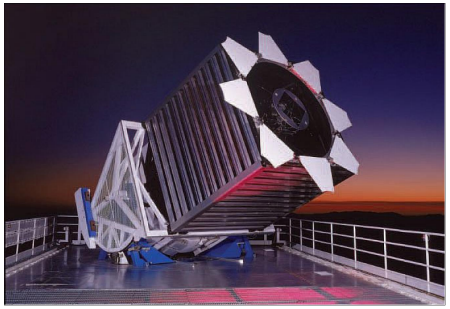
Advanced Data Products – What do you mean ?

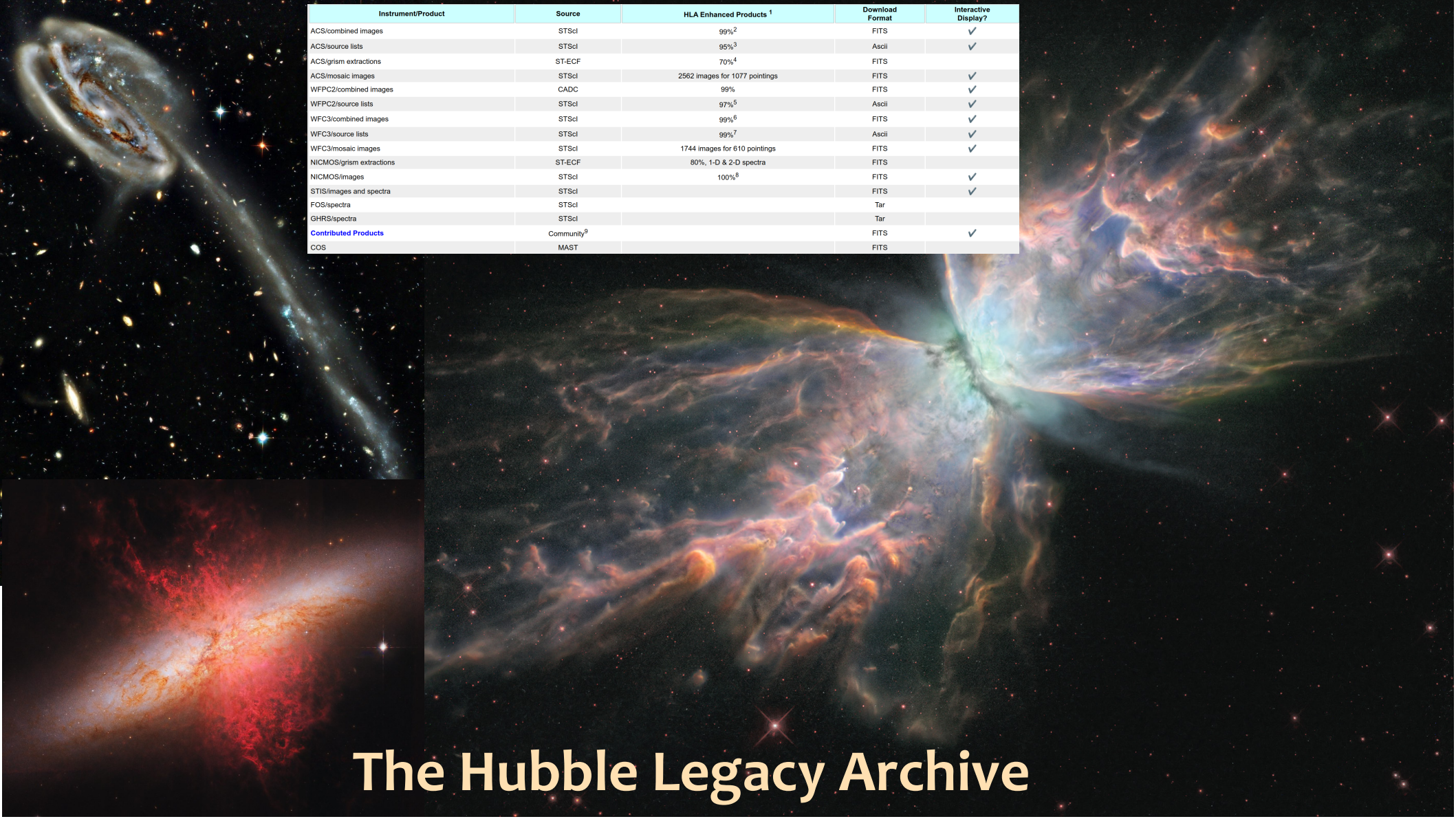
Examples of Data Products

- **The Hubble Legacy Archive** : catalogs and images over 2 decades
- **Gaia Data Releases** : star catalogues, positions, distances, proper motions
- **SDSS Data Releases** : spectra but also redshifts, star formation history
- **CFHT Legacy Survey** : mosaicks of the sky
- **Gemini** : light curve of variable stars



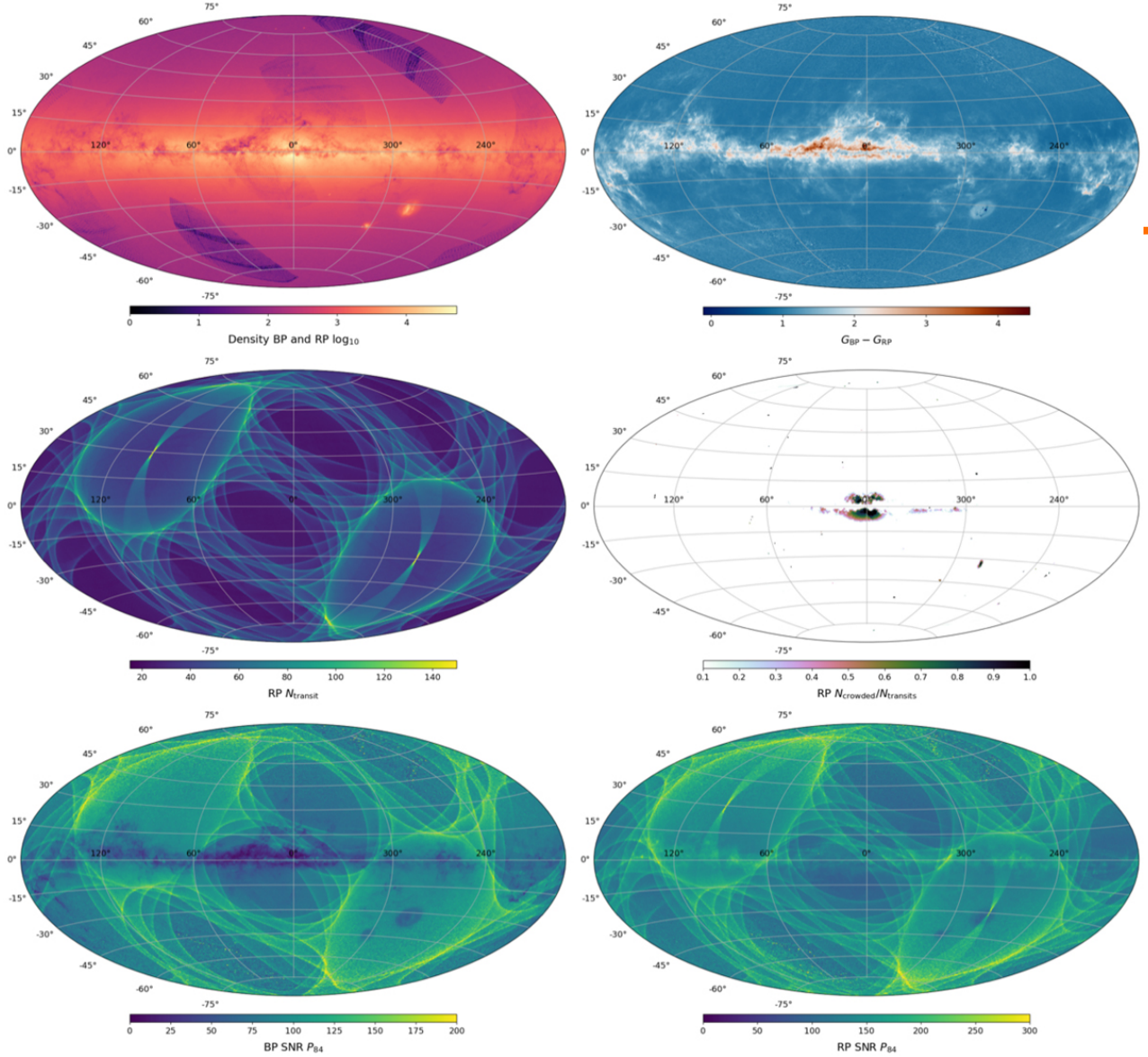




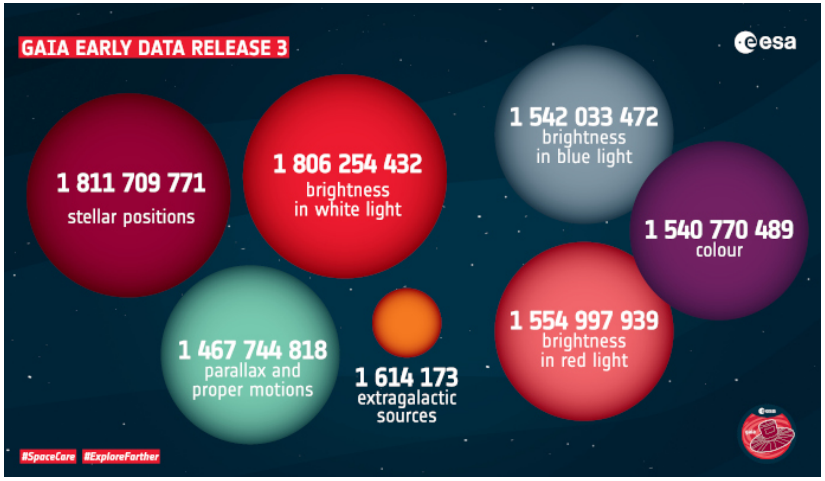


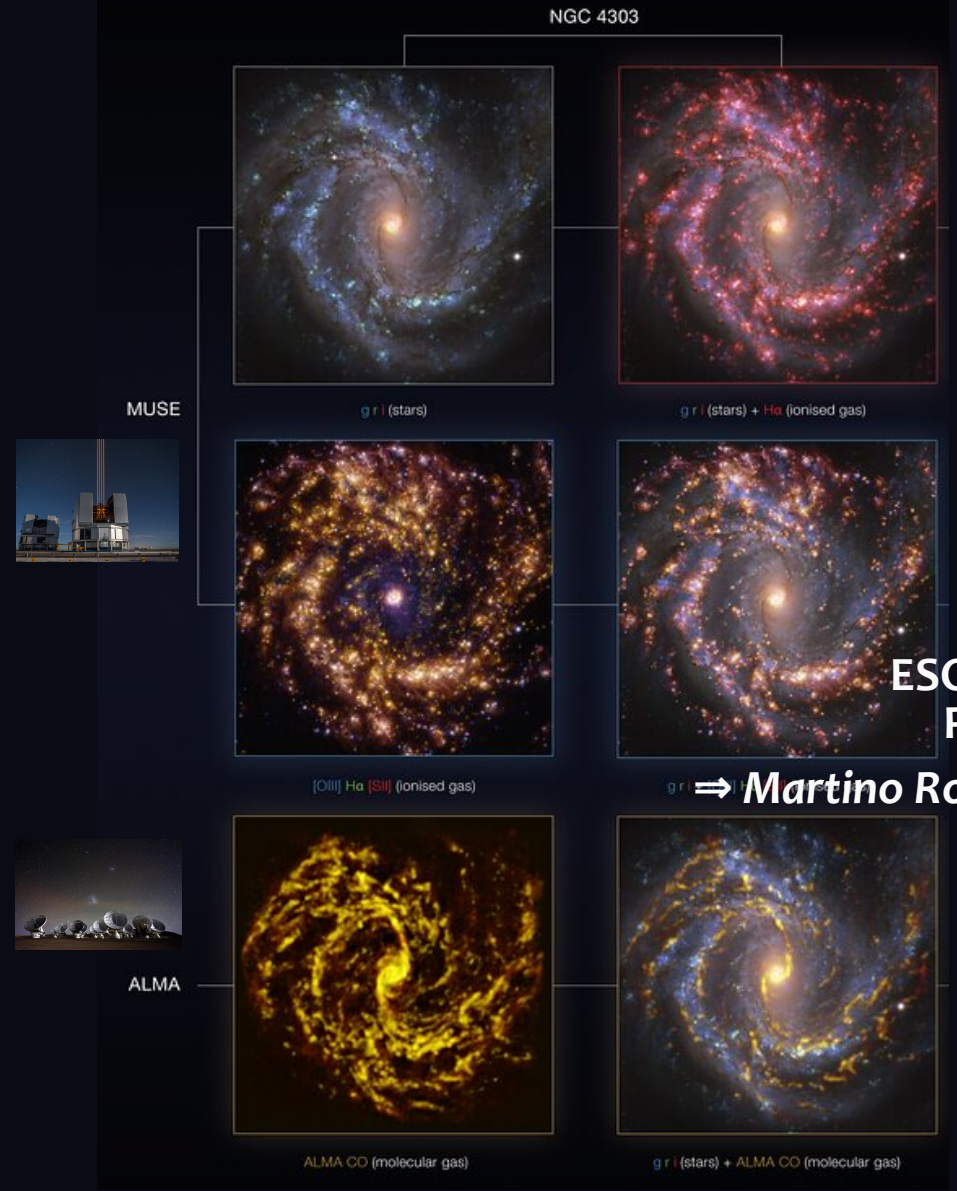
Instrument/Product	Source	HLA Enhanced Products [1]	Download Format	Interactive Display?
ACS/combined images	STScI	99%[2]	FITS	✓
ACS/source lists	STScI	95%[3]	Ascii	✓
ACS/grism extractions	ST-ECF	70%[4]	FITS	
ACS/mosaic images	STScI	2562 images for 1077 pointings	FITS	✓
WFPC2/combined images	CADC	99%	FITS	✓
WFPC2/source lists	STScI	97%[5]	Ascii	✓
WFC3/combined images	STScI	99%[6]	FITS	✓
WFC3/source lists	STScI	99%[7]	Ascii	✓
WFC3/mosaic images	STScI	1744 images for 610 pointings	FITS	✓
NICMOS/grism extractions	ST-ECF	80%, 1-D & 2-D spectra	FITS	
NICMOS/images	STScI	100%[8]	FITS	✓
STIS/images and spectra	STScI		FITS	✓
FOS/spectra	STScI		Tar	
GHRS/spectra	STScI		Tar	
Contributed Products	Community[9]		FITS	✓
COS	MAST		FITS	

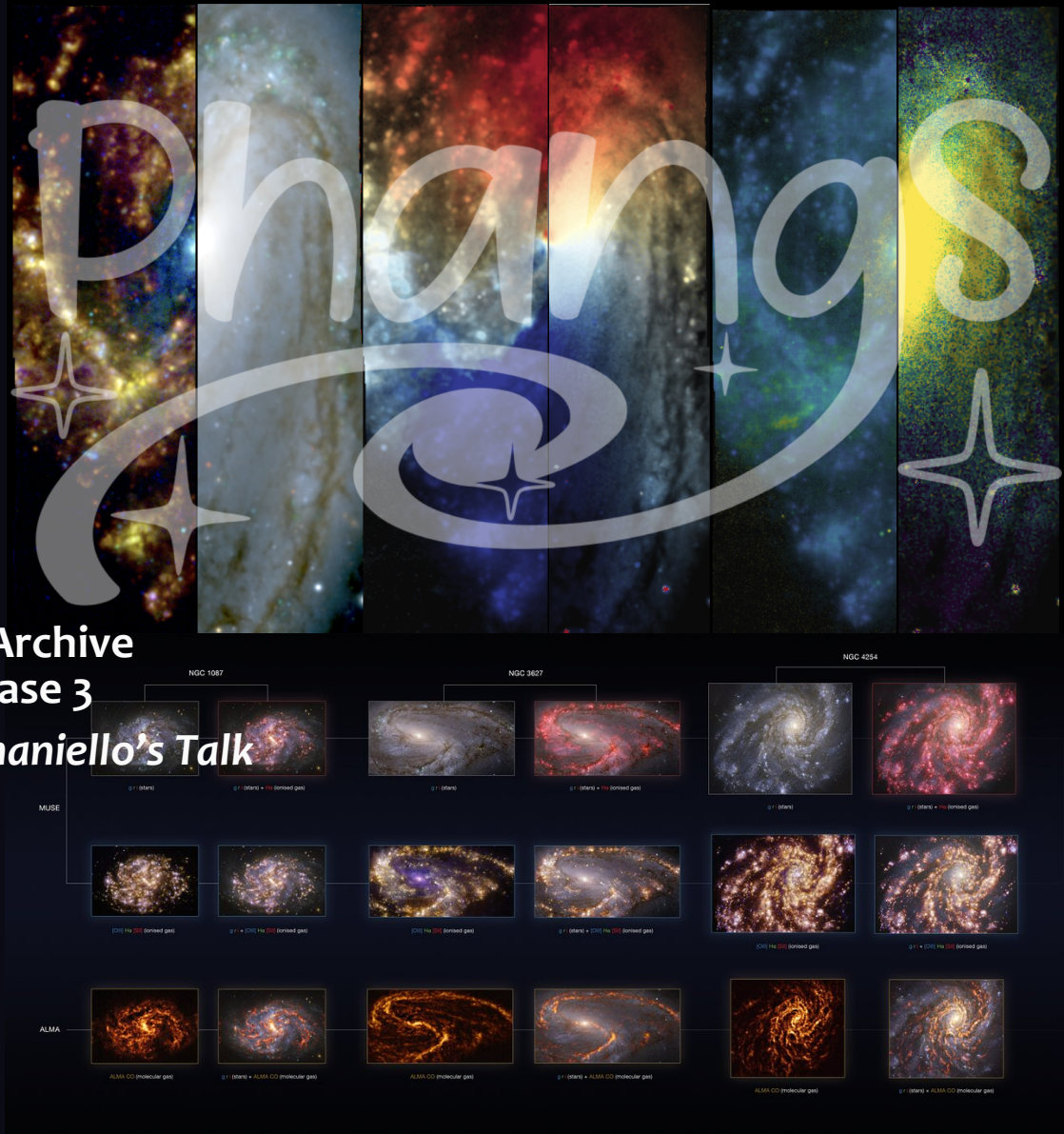
The Hubble Legacy Archive



The Gaia Data Releases







ESO Archive
Phase 3

⇒ ***Martino Romaniello's Talk***

Advanced Data Products – Levels @Geophysics

« Scientific Relevance »

- **Level 0** : Raw data – as measured
- **Level 1** : processed data, in instrument/sensor units, with main instrumental signature removed (sometimes 1A, 1B, 1C…)
- **Level 2** : Calibrated data at the same resolution as Level 1 (no loss)
- **Level 3** : Variables on uniform space/time grids
- **Level 4** : Model output or results of analysis of lower level data, variables derived from multiple measurements

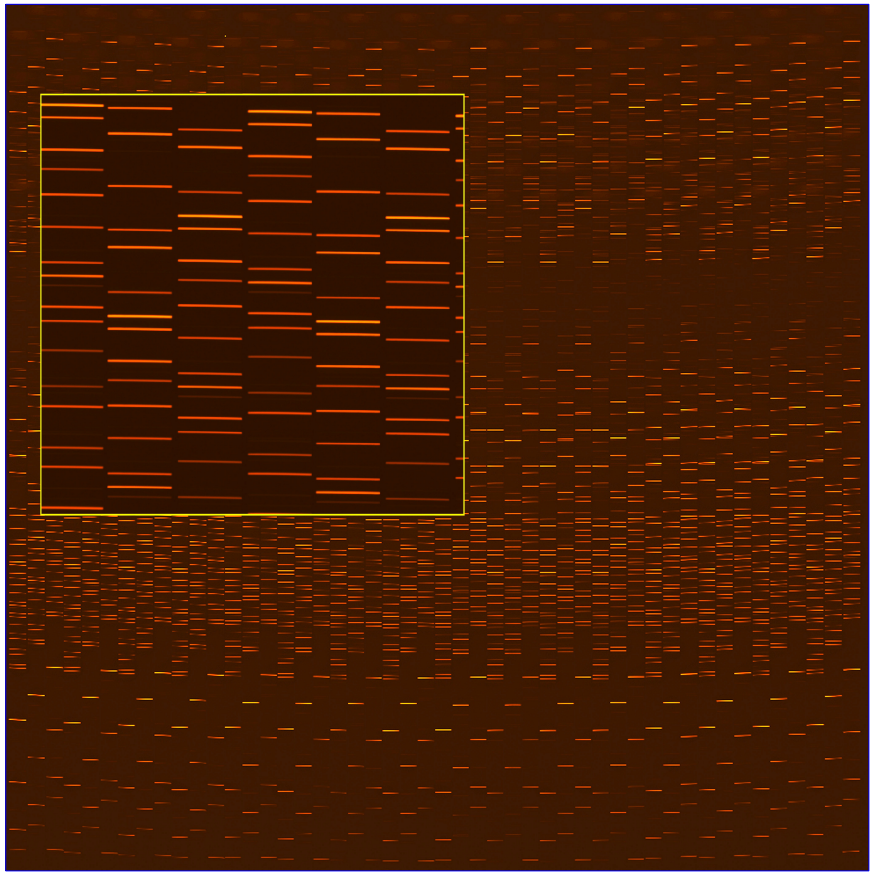
Information loss

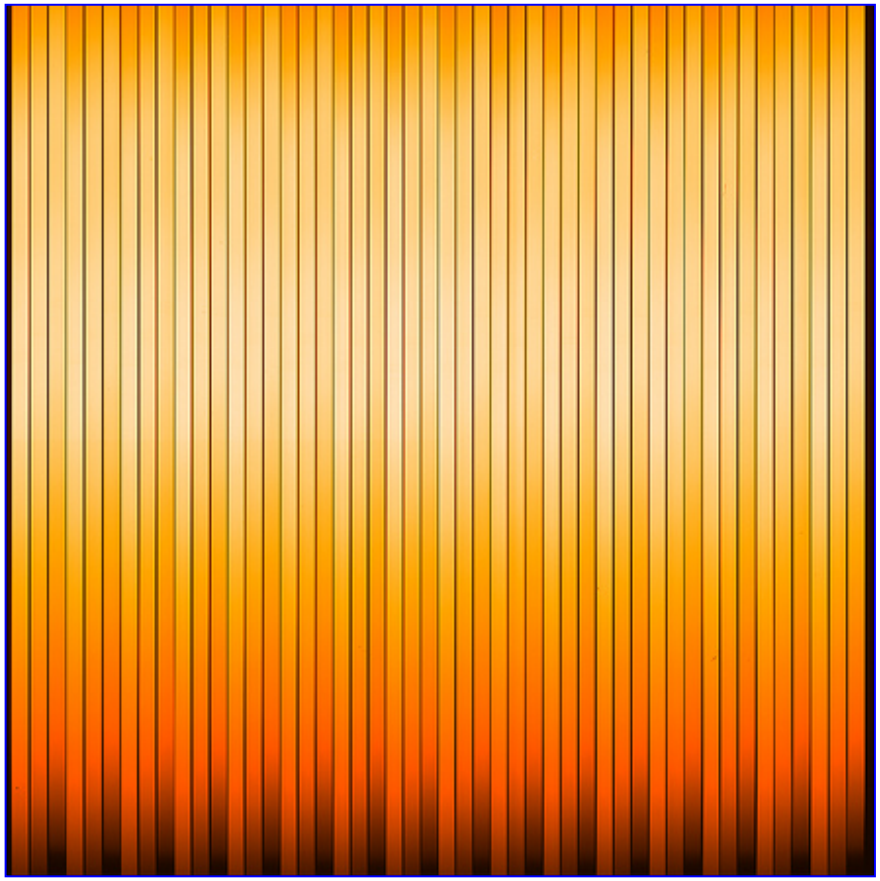
File Size

Level 0

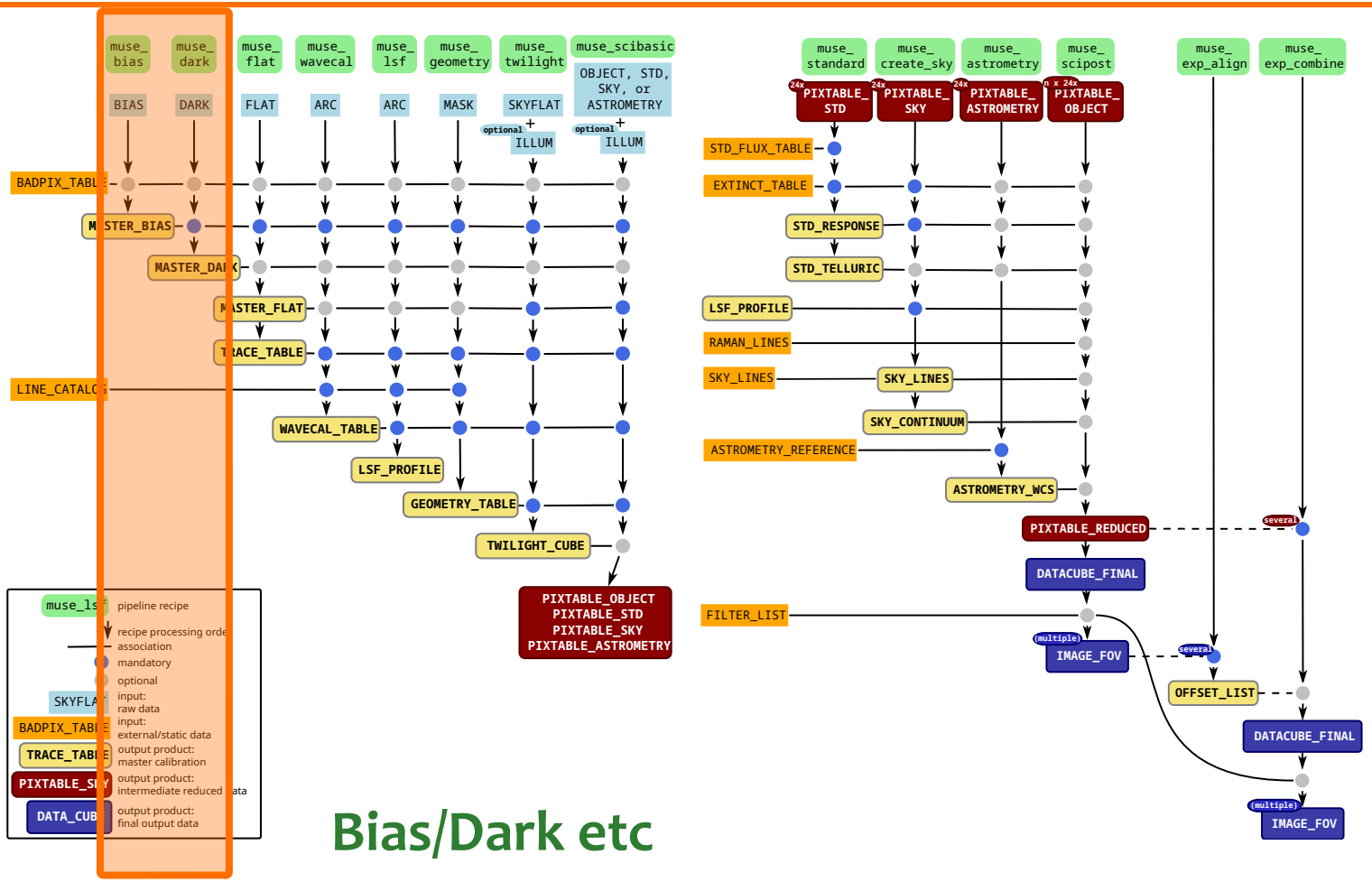
muse
multi unit spectroscopic explorer

24x



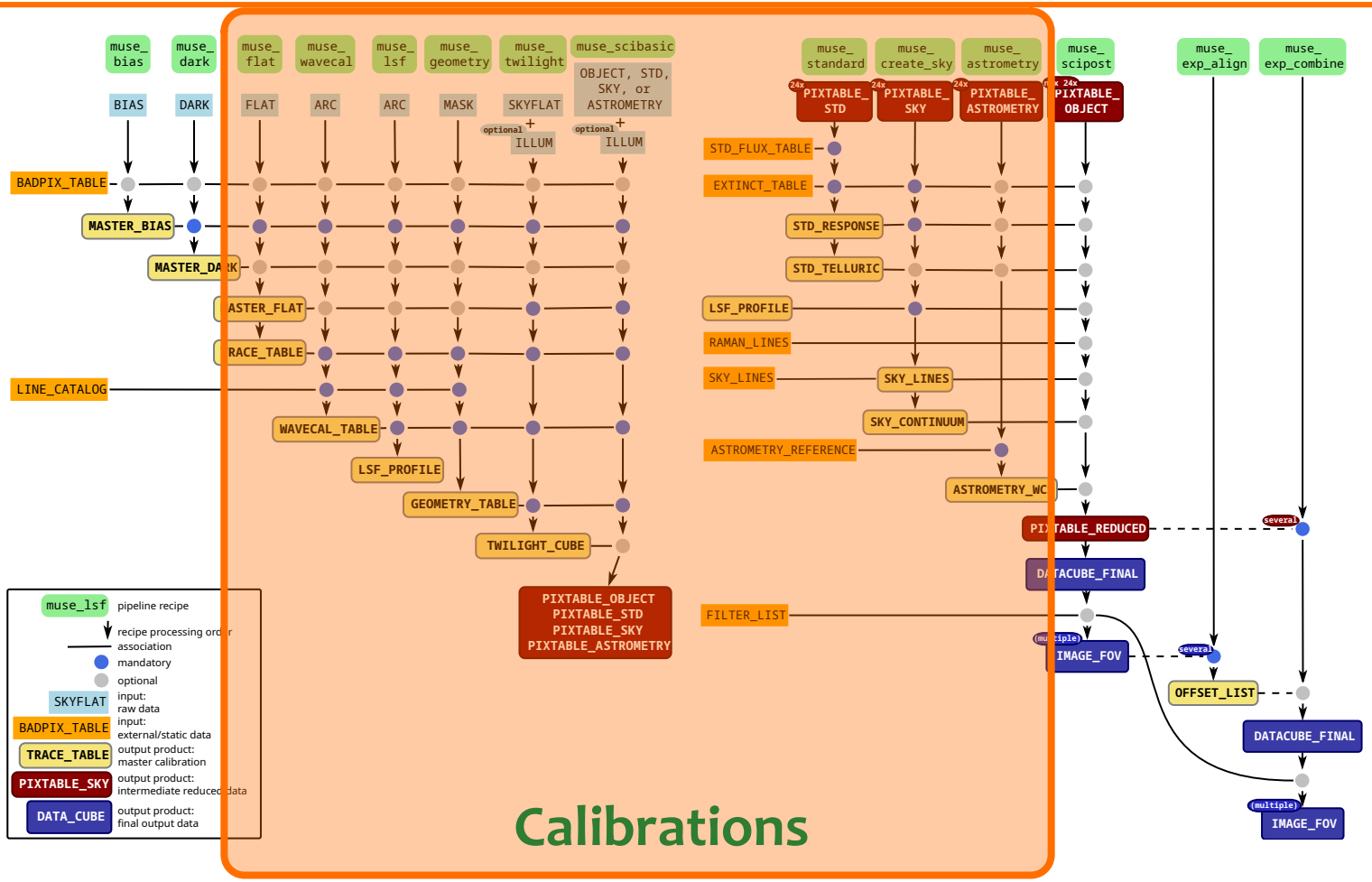


⇒ Level 1

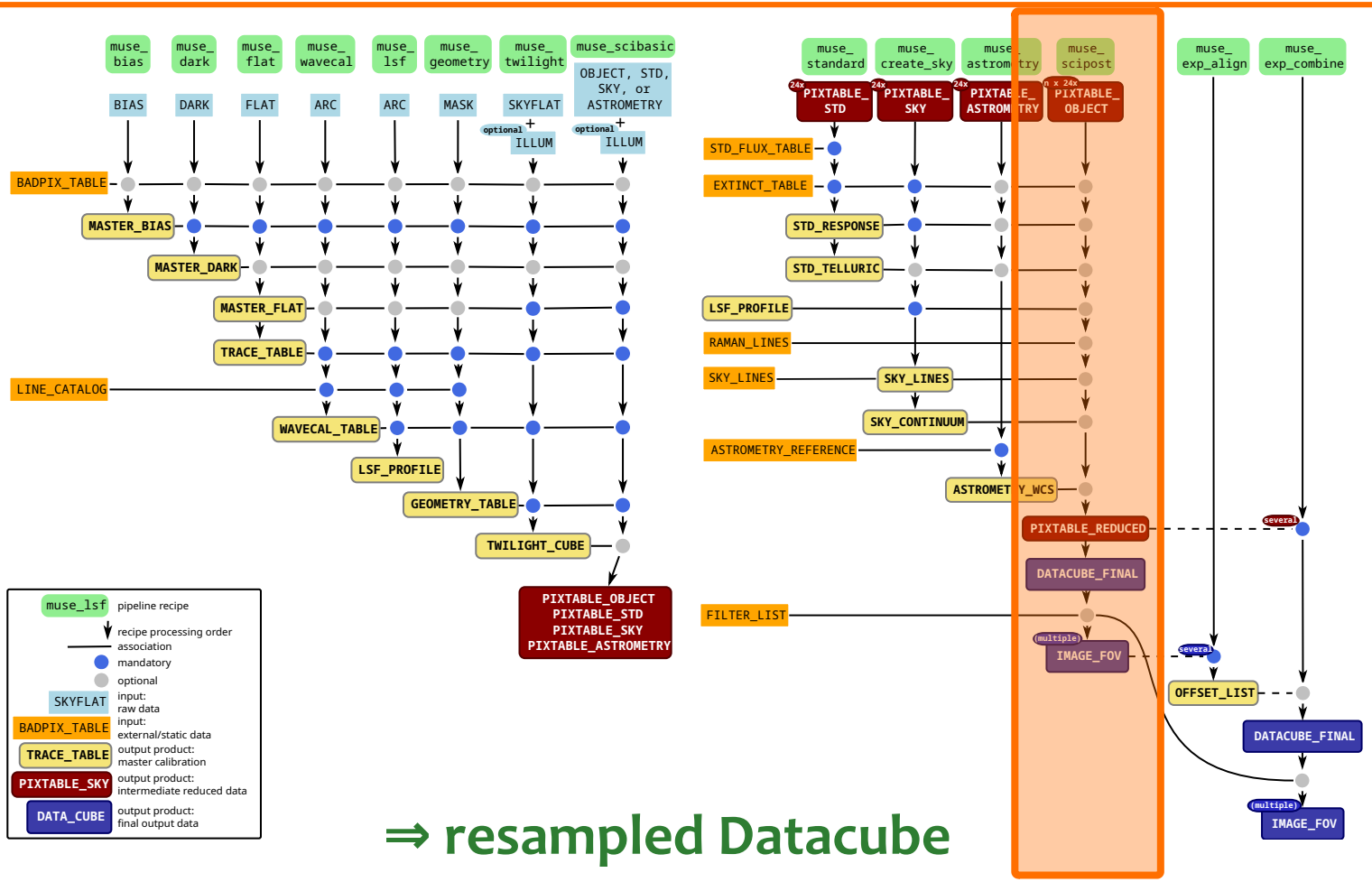


Bias/Dark etc

⇒ Level 2

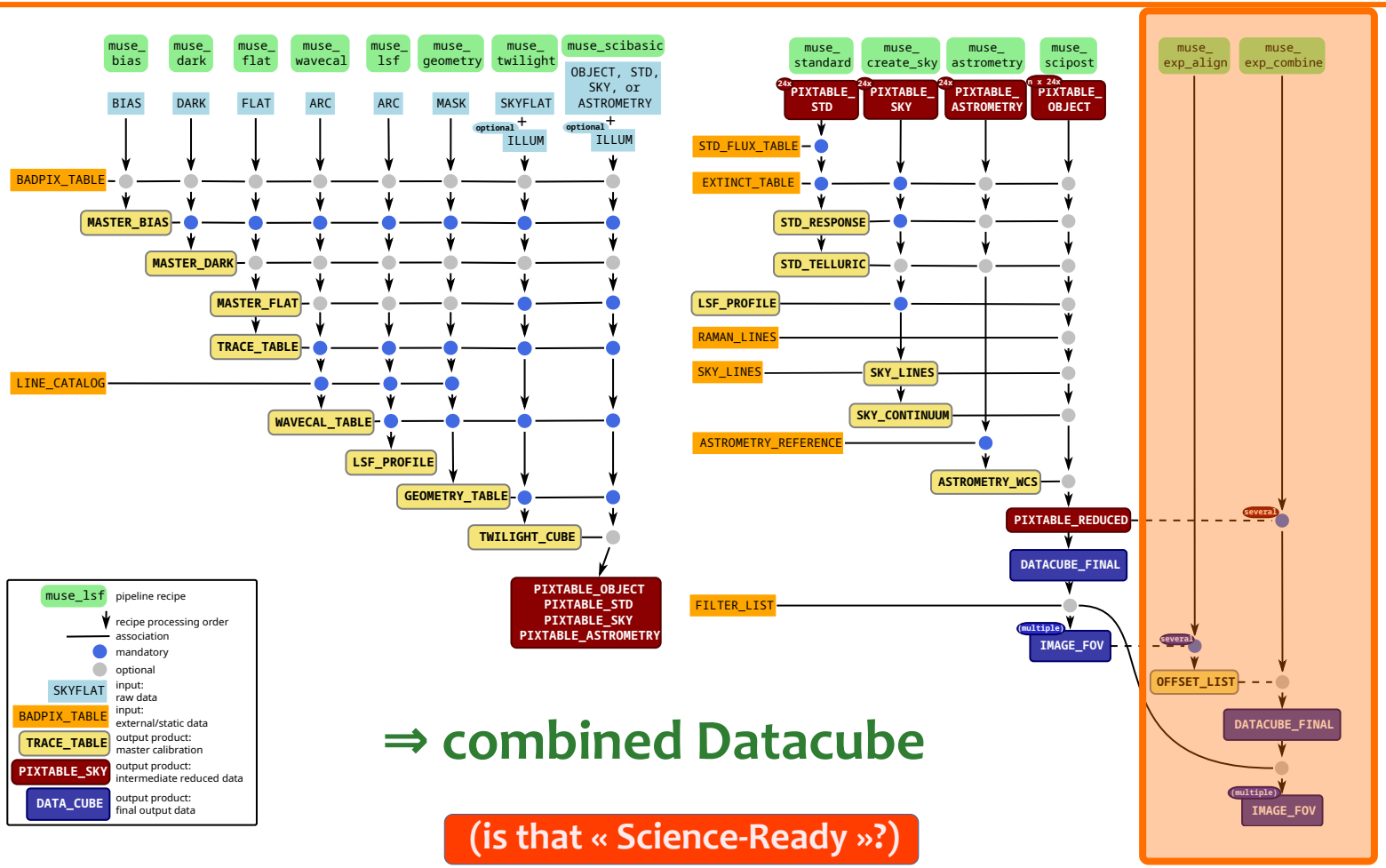

muse_bias
muse_dark
muse_flat
muse_wavecal
muse_lsf
muse_geometry
muse_twilight
muse_scibasic
muse_standard
muse_create_sky
muse_astrometry
muse_scipost
muse_exp_align
muse_exp_combine
BIAS
DARK
FLAT
ARC
ARC
MASK
SKYFLAT
OBJECT, STD, SKY, or ASTROMETRY
optional
ILLUM
optional
ILLUM
24x
PIXTABLE_STD
24x
PIXTABLE_SKY
24x
PIXTABLE_ASTROMETRY
24x
PIXTABLE_OBJECT
BADPIX_TABLE
MASTER_BIAS
MASTER_DARK
MASTER_FLAT
TRACE_TABLE
LINE_CATALOG
WAVECAL_TABLE
LSF_PROFILE
GEOMETRY_TABLE
TWILIGHT_CUBE
PIXTABLE_OBJECT
PIXTABLE_STD
PIXTABLE_SKY
PIXTABLE_ASTROMETRY
STD_FLUX_TABLE
EXTINCT_TABLE
STD_RESPONSE
STD_TELLURIC
LSF_PROFILE
RAMAN_LINES
SKY_LINES
SKY_LINES
SKY_CONTINUUM
ASTROMETRY_REFERENCE
ASTROMETRY_WCS
PIXTABLE_REDUCED
several
DATACUBE_FINAL
FILTER_LIST
(multiple)
IMAGE_FOV
several
OFFSET_LIST
DATACUBE_FINAL
(multiple)
IMAGE_FOV
muse_lsf pipeline recipe
recipe processing order
association
mandatory
optional
SKYFLAT input: raw data
BADPIX_TABLE input: external/static data
TRACE_TABLE output product: master calibration
PIXTABLE_SKY output product: intermediate reduced data
DATA_CUBE output product: final output data
Calibrations

⇒ Level 3



⇒ resampled Datacube

⇒ Level 3 / 4



⇒ combined Datacube

(is that « Science-Ready »?)

Advanced Data Products – Levels (update)

« Scientific Relevance »

- **Level 0** : Raw data – as measured
- **Level 1** : Data with full instrumental signature removed
- **Level 2** : Extracted information, regridding, uniform space/time
- **Level 3** : Model output or results of analysis of lower level data, variables derived from multiple measurements

Information loss

File Size

An enhanced approach : from Level 0 to Level 1… to Level N

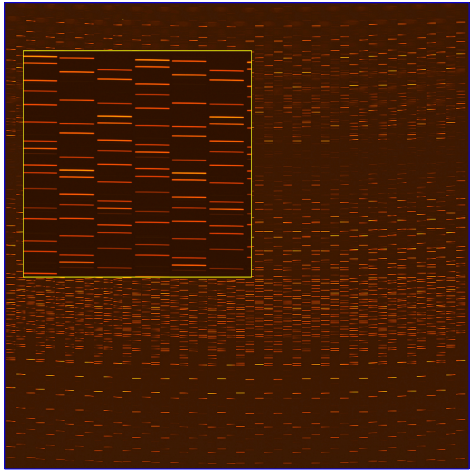


Table creation

x	y	λ	value	flag	variance
-75.004	11.55	8408.7	87.72	0	19785.9
-75.003	11.55	8410.0	10.68	0	19655.3
-75.002	11.55	8411.2	43.55	0	20630.6
-75.000	11.55	8412.5	85.31	5	21929.7
-149.982	7.55	4212.9	1281.37	0	1321482.0
-149.983	7.55	4214.2	482.40	0	166733.0
-149.995	7.55	4215.4	394.27	0	110684.0
-149.997	7.55	4216.8	967.98	0	740456.0

x-y shift/scale

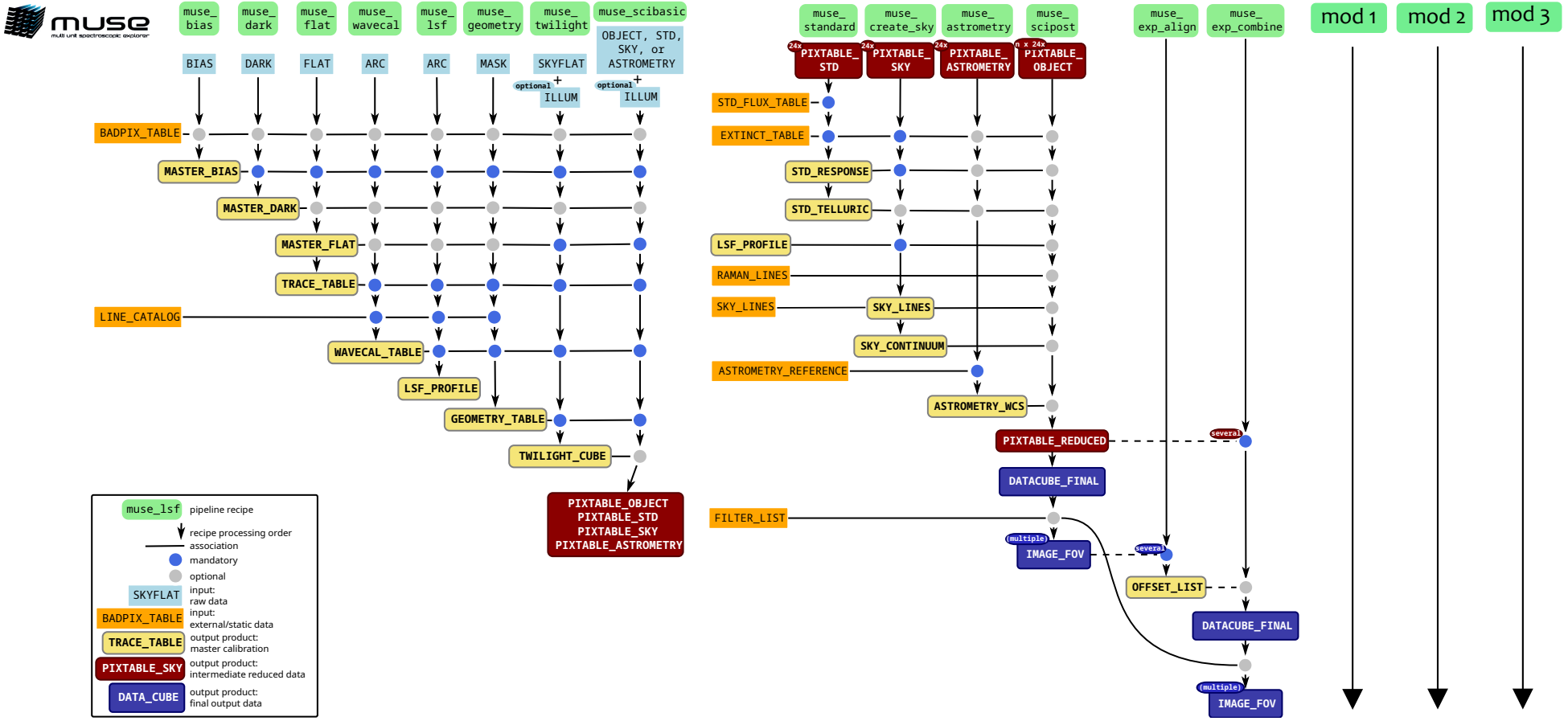
x	y	λ	value	flag	variance
-95.504	17.325	8408.7	87.72	0	19785.9
-95.503	17.325	8410.0	10.68	0	19655.3
-95.502	17.325	8411.2	43.55	a0	20630.6
-95.500	17.325	8412.5	85.31	5	21929.7
-170.482	11.325	4212.9	1281.37	0	1321482.0
-170.483	11.325	4214.2	482.40	0	166733.0
-170.495	11.325	4215.4	394.27	0	110684.0
-170.497	11.325	4216.8	967.98	0	740456.0

astrometric calibration

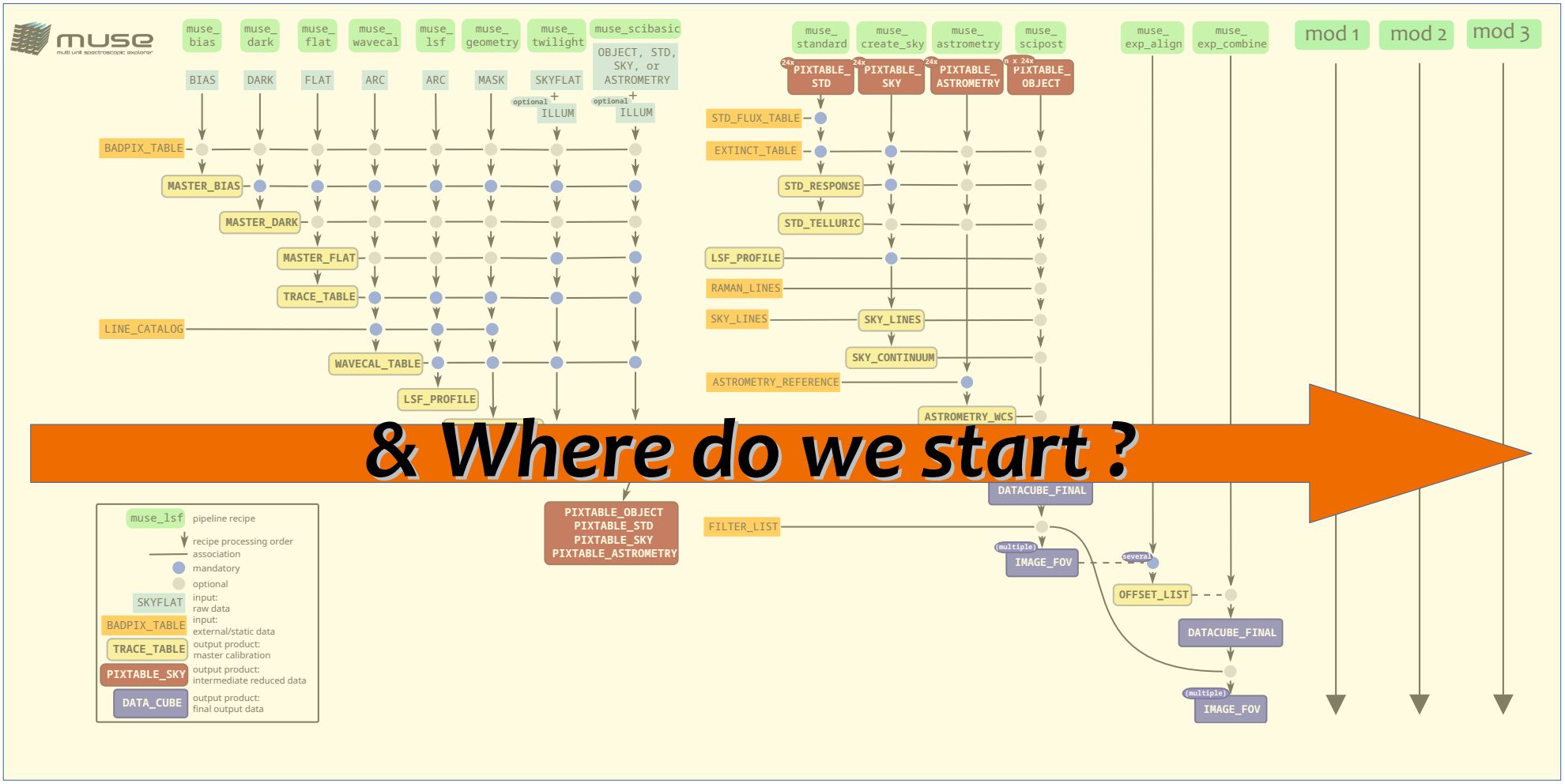
RA	DEC	λ	value	flag	variance
-19.1008	3.465	8408.7	87.72	0	19785.9
-19.1006	3.465	8410.0	10.68	0	19655.3
-19.1004	3.465	8411.2	43.55	0	20630.6
-19.1000	3.465	8412.5	85.31	5	21929.7
-34.0964	2.265	4212.9	1281.37	0	1321482.0
-34.0966	2.265	4214.2	482.40	0	166733.0
-34.0990	2.265	4215.4	394.27	0	110684.0
-34.0994	2.265	4216.8	967.98	0	740456.0

Change of coordinates
No change to the values

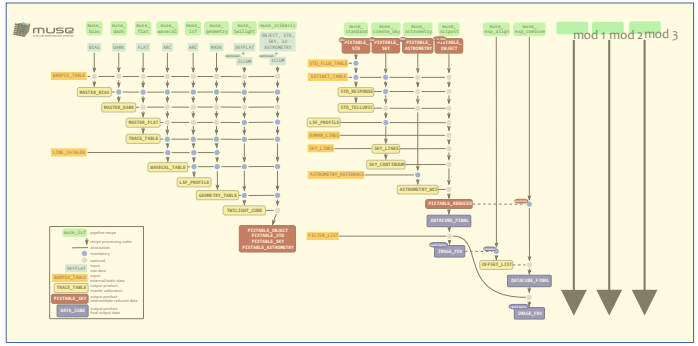
Are those levels relevant for science ?



ADPs ?


muse
multi unit spectroscopic explorer
muse_bias
muse_dark
muse_flat
muse_wavecal
muse_lsf
muse_geometry
muse_twilight
muse_scibasic
muse_standard
muse_create_sky
muse_astrometry
muse_scipost
muse_exp_align
muse_exp_combine
mod 1
mod 2
mod 3
BIAS
DARK
FLAT
ARC
ARC
MASK
SKYFLAT
OBJECT, STD, SKY, or ASTROMETRY
optional
ILLUM
BADPIX_TABLE
MASTER_BIAS
MASTER_DARK
MASTER_FLAT
TRACE_TABLE
LINE_CATALOG
WAVECAL_TABLE
LSF_PROFILE
PIXTABLE_STD
PIXTABLE_SKY
PIXTABLE_ASTROMETRY
PIXTABLE_OBJECT
STD_FLUX_TABLE
EXTINCT_TABLE
STD_RESPONSE
STD_TELLURIC
LSF_PROFILE
RAMAN_LINES
SKY_LINES
SKY_CONTINUUM
ASTROMETRY_REFERENCE
ASTROMETRY_WCS
& Where do we start ?
DATACUBE_FINAL
FILTER_LIST
IMAGE_FOV
OFFSET_LIST
DATACUBE_FINAL
IMAGE_FOV
PIXTABLE_OBJECT
PIXTABLE_STD
PIXTABLE_SKY
PIXTABLE_ASTROMETRY
muse_lsf pipeline recipe
recipe processing order
association
mandatory
optional
SKYFLAT input: raw data
BADPIX_TABLE input: external/static data
TRACE_TABLE output product: master calibration
PIXTABLE_SKY output product: intermediate reduced data
DATA_CUBE output product: final output data

ADPs : where do we stop ?



« Observations become scientific when used as evidence of phenomena »

Borgman 2015

What are we trying to solve ?

Observational science :
Transform a set of observations into a **coherent picture**

Objectives and Gains

⇒ **Optimise** and expand exploitation of facilities / data

⇒ **Lower** barriers

⇒ **Respond** to potential changes in our ecosystem (agile)

⇒ Use it as a **leverage for progress**

*[but probably **not** : to shorten the way to papers]*

An evolving landscape

Era of « Big » & « multi-messenger » Data

- **LSST / SKA** : petabytes or even exabytes during its mission
  (LSST = **20 Tb / night**)

- **Analysis speed** : e.g., LIGO that requires processing **within minutes**

- **Richness** : multi-**lambda** / **instruments** / **formats** / **dimensions**

- **Complexity** : data sources, data formatting, nature of the information, need for forward modelling (and instrument simulators)

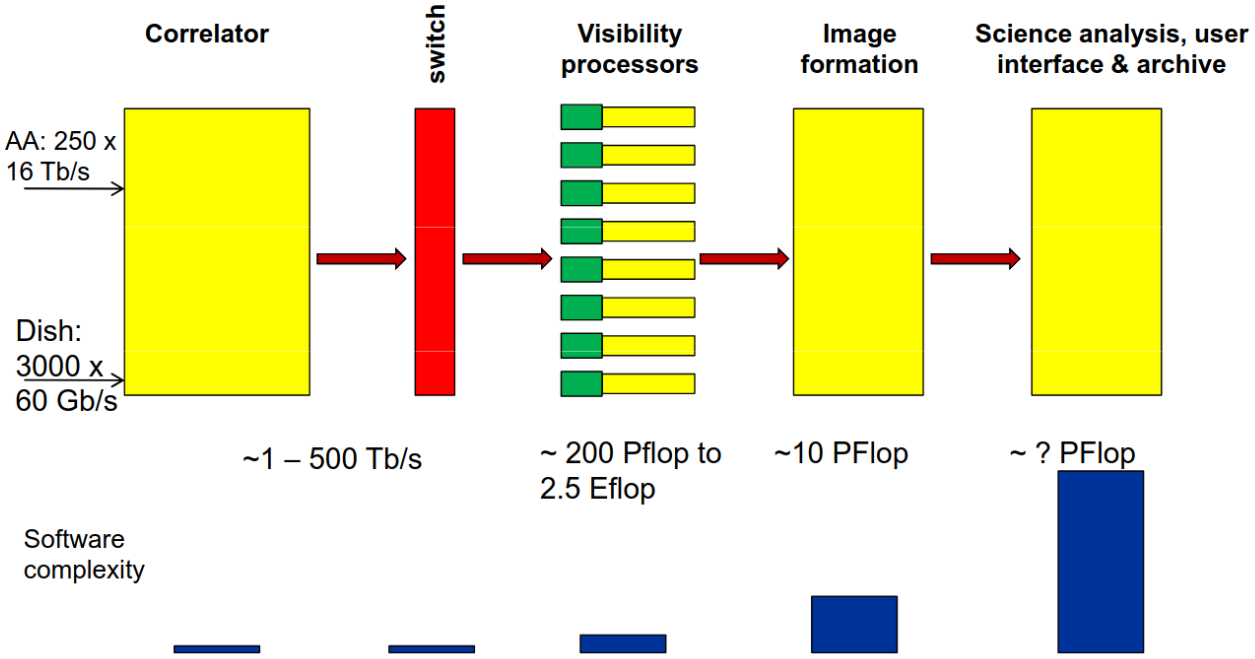
- **Accuracy** : LIGO, or ANDES/ELT, errors (e.g., covariance)

- **Cost & Efficiency of Processing** : SKA !

The Scientific perspective

- Technical / Computing assessment

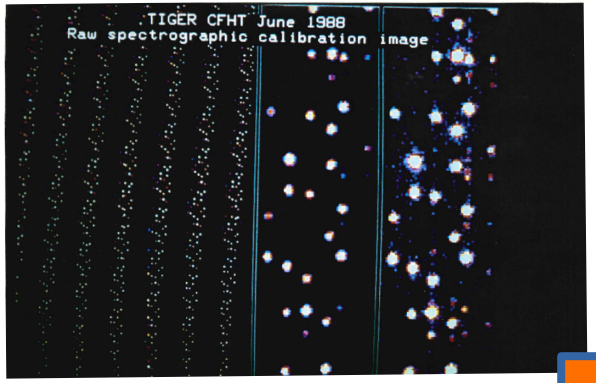
  ⇒ ***what about the end product = scientific discovery ?***

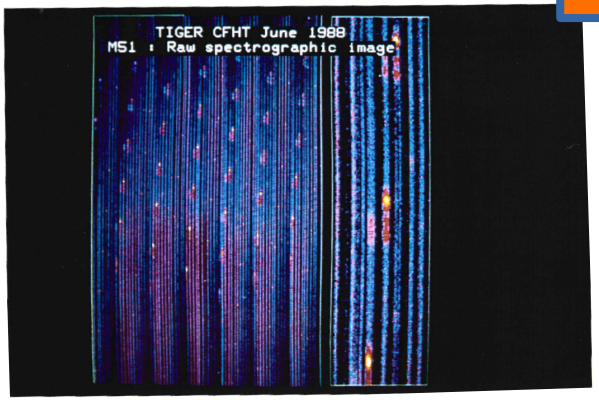


©Paul Anderson @ SKA

Emerging from the Dark Ages

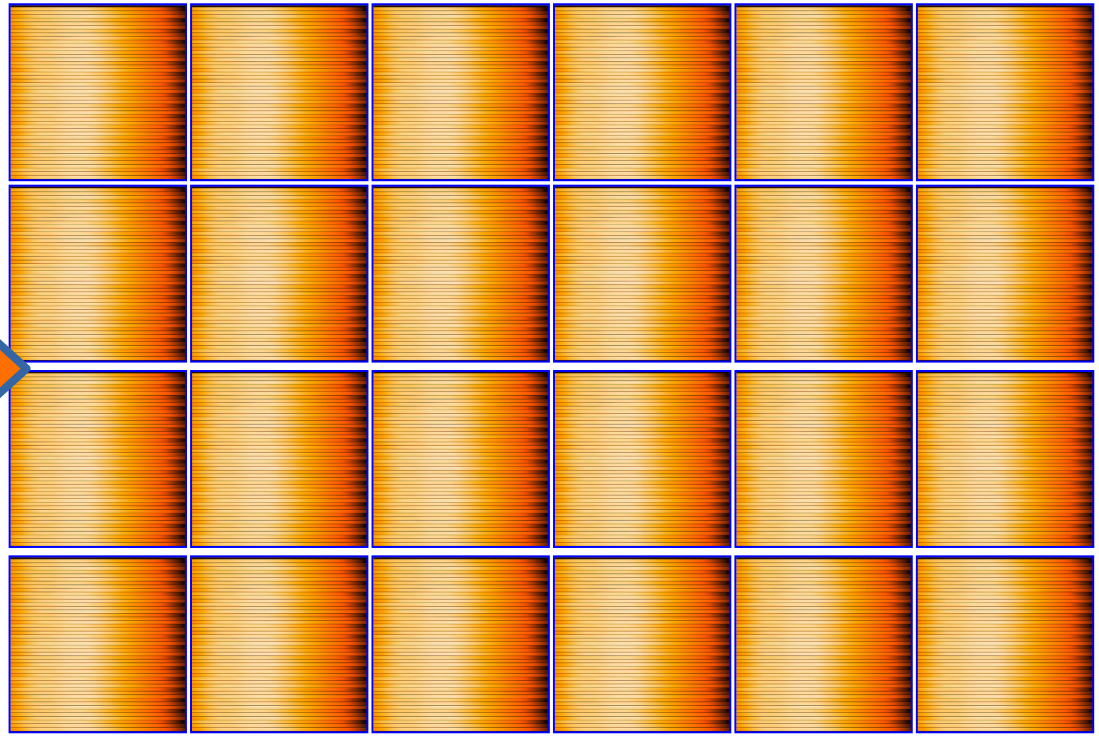
50 x 400





1988

90,000 x 4000

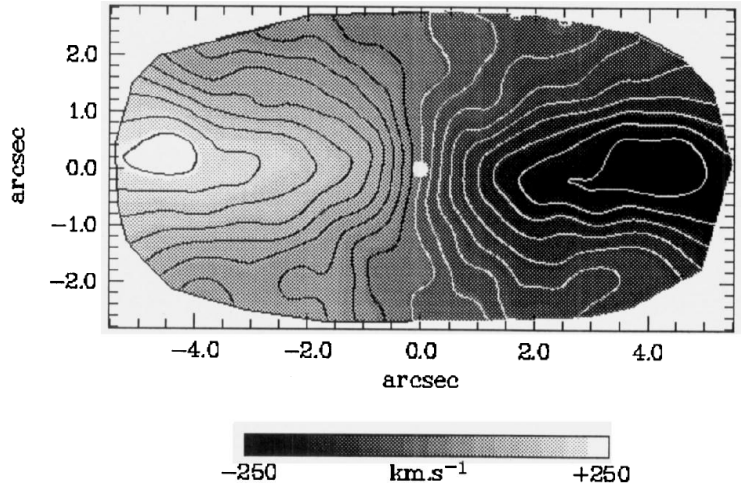


2014

Emsellem, Bacon, Monnet, Poulain 1996

TIGER

V_*

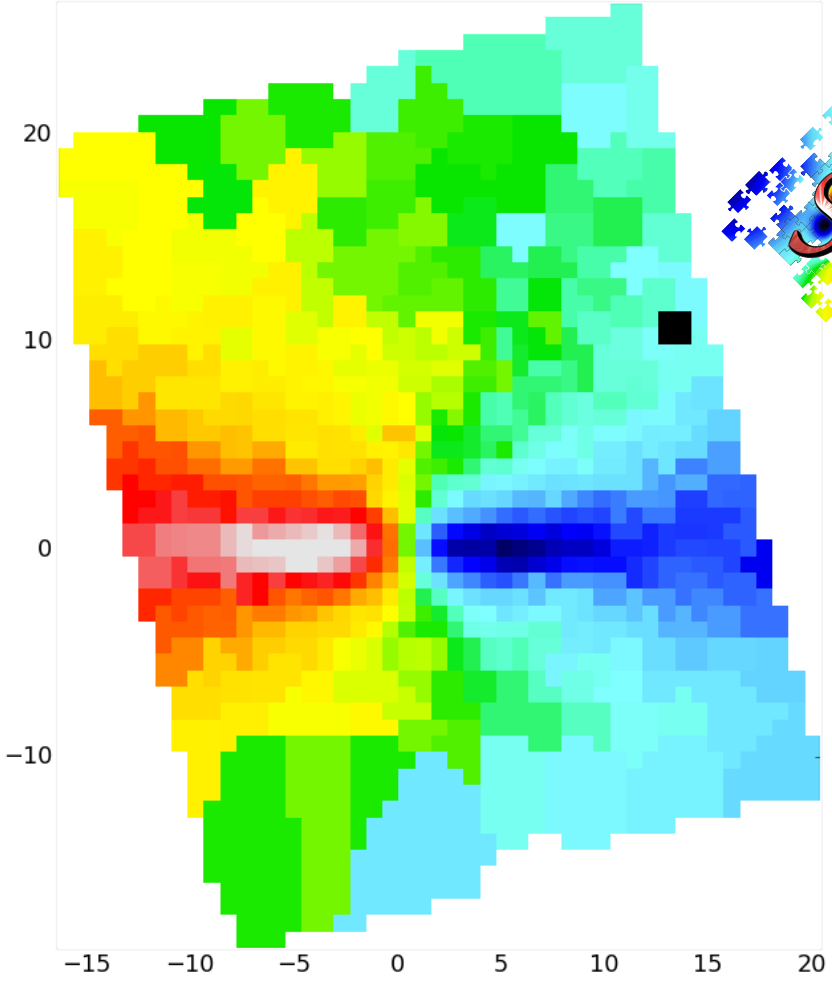


1996

400 spectra / 2 hours

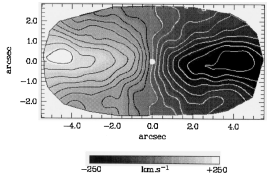
Emsellem et al., unpublished

V*



1500 spectra / 30’

1996

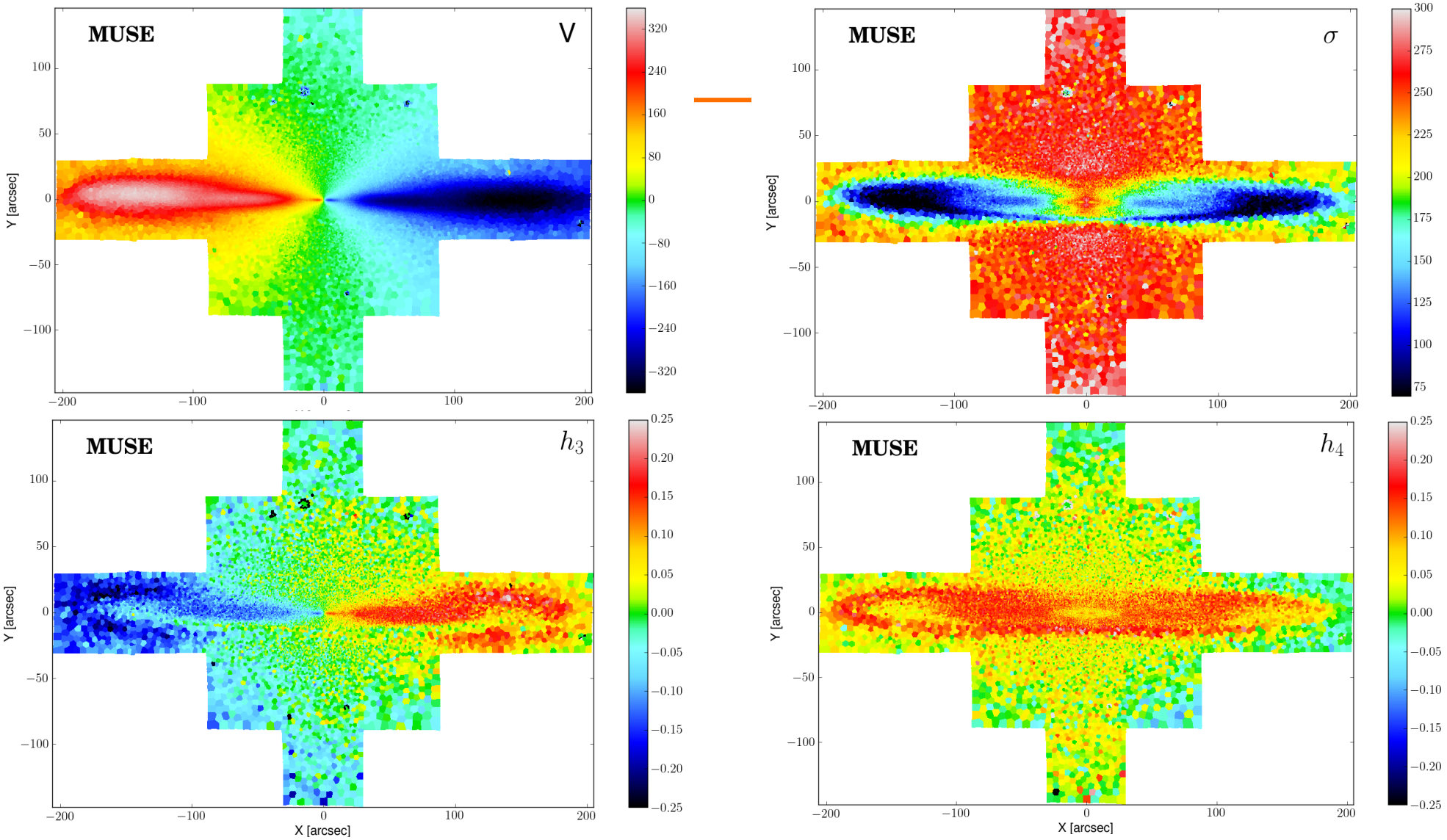


2003

Emsellem et al., SV data

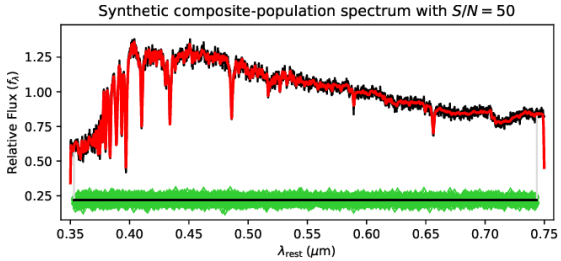
1.4 10^6 spectra, < 4h

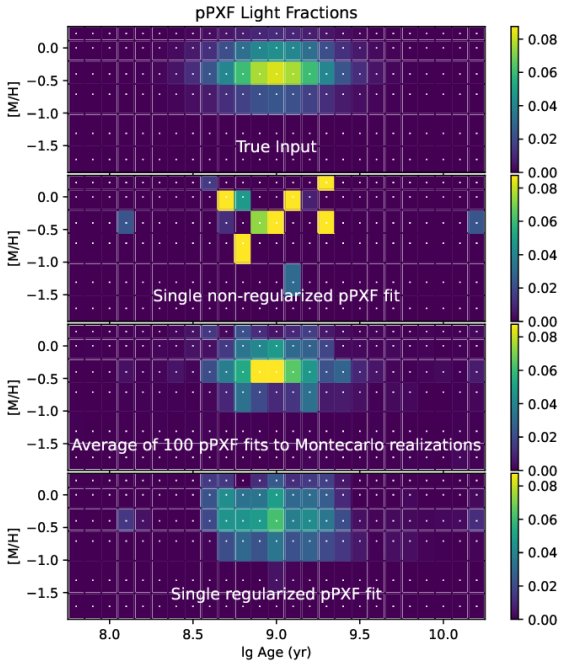
2014


MUSE
V
MUSE
σ
MUSE
h3
MUSE
h4
Y [arcsec]
X [arcsec]

What it entails

Star formation history (Cappellari, 2022)



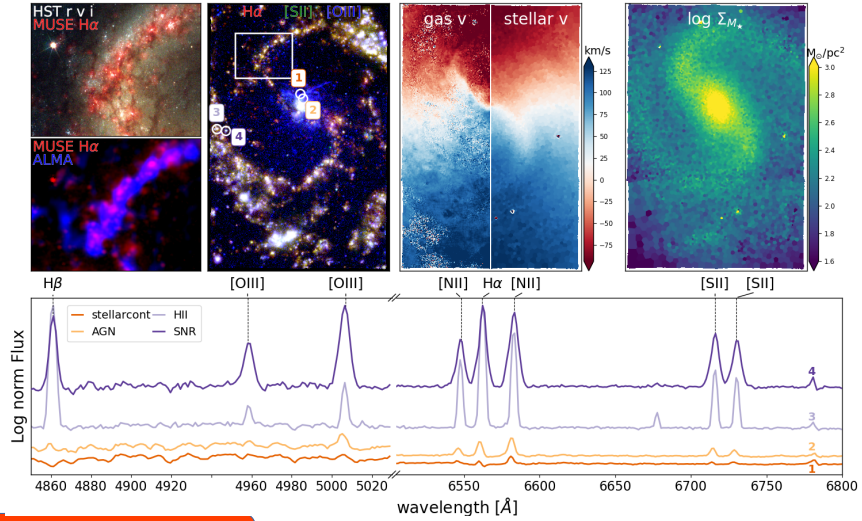


- Input spectrum and variance (& covariance)
- A complex code with algorithmic branches
- Many Input parameters :
  - Polynomial degrees, windows, potential emission lines, regularisation matrix
- Extension law
- Full stellar library
  - More spectra
  - More assumptions (Simple Stellar Population)

Data do not work in isolation

- **Data: catalogs, maps, key astrophysical parameters**

ESO Archive
Phase 3 data products



Emsellem et al. 2022

- **Simulated Universe and predictive models**
  - Simulations & models go hand in hand with data

    e.g., applied libraries, theoretical assumptions, numerical models and simulations

    (e.g., MAST High-Level Science Products, HLSPs ⇒ see Scott Fleming's talk)

A technical paradigm change

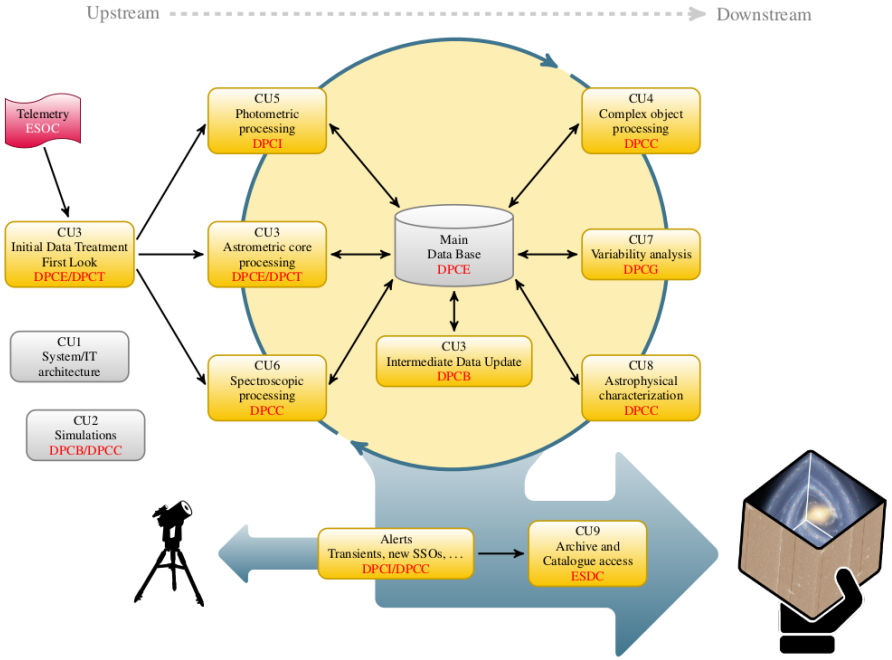
- Technological developments
  - hardware and associated frameworks & infrastructures
  - Often software + hardware hand in hand (Supercomputers, GPUs)
- **Algorithmic** improvements : machine learning/AI, optimisation algorithms
- Automated Discovery (some of the fastest algorithmic improvements)
  - Automated classification, anomaly detection
  - Predictive models
  - Optimisation of telescope time

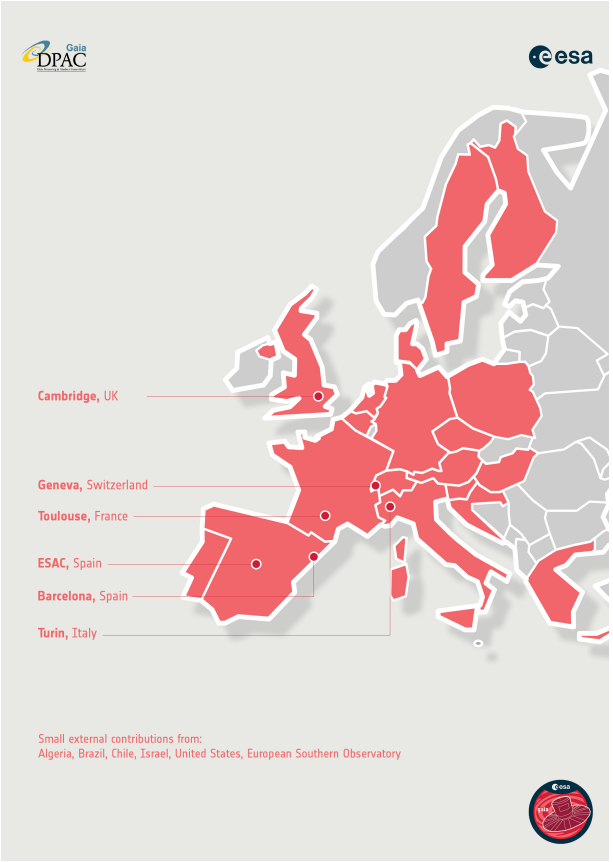
⇒ New partnerships with AI scientists
⇒ Importance of scrutinising its ethical impact

- New development/**communication** platforms and **networks**

An Era of collaborations and Interoperability

- Networks of observatories, institutes, organisations
  ⇒ Requires Standards
  - Data formats
  - Collaborative tools and protocols
  - Portals, Platforms, Clouds





A societal paradigm change

- **The next generation (?)**
  - Those who will use SKA, the ELTs

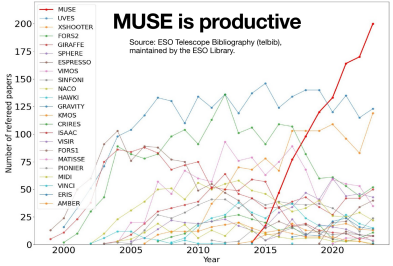
    ⇒ are barely coming out of high school



- **Public engagement**
  - Citizen science
  - What about citizen developments ?

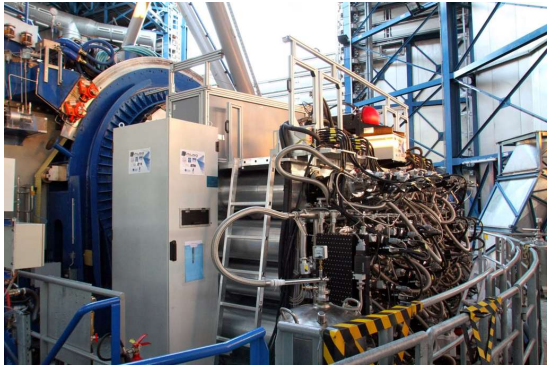
- **Public opinion**
  - Everyone has an opinion, and they are visible, impactfull
  - Harder to discriminate levels of relevance

Why is MUSE at the VLT so “successful” ?



BTW : *what does this mean ?*
Which metric should we use ?
[popularity, production, impact ?]

- Design and development approach
  - Broad community service
  - But restricted set of modes
  - Robust pipeline : based on simulations, collaborative updates
  - **3-ways conversation : ESO/Consortium/Community**
  - **Early testing phase of information extraction, simulations**
  - **Parallel development of ADP modules**



⇒ Very complex instrument, but « easy » to use
⇒ Relatively easy **to transform raw data into a coherent scientific picture**

Risks and Threats

The boundary limit

- Limitations come **from the community** (size, expertise, coordination)
  - *not so much from the data* **[we are data rich]**
- If data cannot be used easily **(high cost! Or low efficiency)**

  ⇒ the community will switch to other datasets
- The origin of the data does not matter, e.g.,

  ⇒ ESO could turn into the **best data provider** and ***have low impact***

Risk assessment

Big Data / Multi-source

- Being late : data is there, tools are not, no one is ready
- Accessibility becomes a myth
  - « who » can address data ?
  - « who » can use the tools ?

⇒ We lose part of our community

- **& Low efficiency** : less than optimal data exploitation

  & Bad use of money and efforts (low optimisation)

  ⇒ Be Open : Engage from the start of the design phase

  ⇒ On site and distributed services (nodes, platforms)

Risk assessment

Tech paradigm change

- Running after trains (Python! Platforms ?)
- **Loss of expertise, attractiveness and visibility**
- **Low efficiency : less than optimal data exploitation**

  & Bad use of money and efforts (low optimisation)

⇒ Keep an engaged perspective = active participation

⇒ Keep an attractive ecosystem (careers, projects)

⇒ Be open

Risk assessment

Societal paradigm change

- Research approach **not in sync** with new generation of researchers
- Lack of a community identity
- **Losing « costumers »** (both scientists **&** the public)
  - Not fully serving our community
  - Far from optimal Investment
  - Drift and disconnect

⇒ Monitor changes, embed yourself, adopt modern standards

⇒ Promote career paths, hire junior staff

Risk assessment

Collaborations / Interoperability

- Wrong standards, or … no Standards
- **Culture gap**, conversation between stakeholders too costy
- Disconnect ⇒ not lowering the barriers, just **offsetting** them
- Loss of visibility, credibility
- **Ivory tower effect** (we are the best)

⇒ **engage, expose yourself, embed yourself**

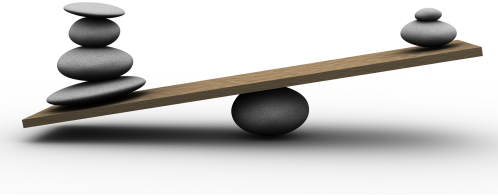
⇒ **Be Open** [see the coming HOWTO slide]

Black box effect

- **Telescope → remote observing**
  - Losing expertise, experience, understanding
- **Tools and AdPs**
  - Losing expertise, experience, understanding

  ⇒ ***Should we care ?***

Lost value **VS** **Added value**



⇒ Importance of being aware, doing it consciously

⇒ Embed methods and configurations in metadata ⇒ reproducibility

Challenges

Challenges

- Mid/Long-term maintenance
- Mid/Long-term preservation : of data & software + configurations
- Mid/Long-term sustainability : accessible, usable
  - New models may be needed
- Relevance of service/products w.r.t. science, and our communities
- Funding and resources (in difficult times)

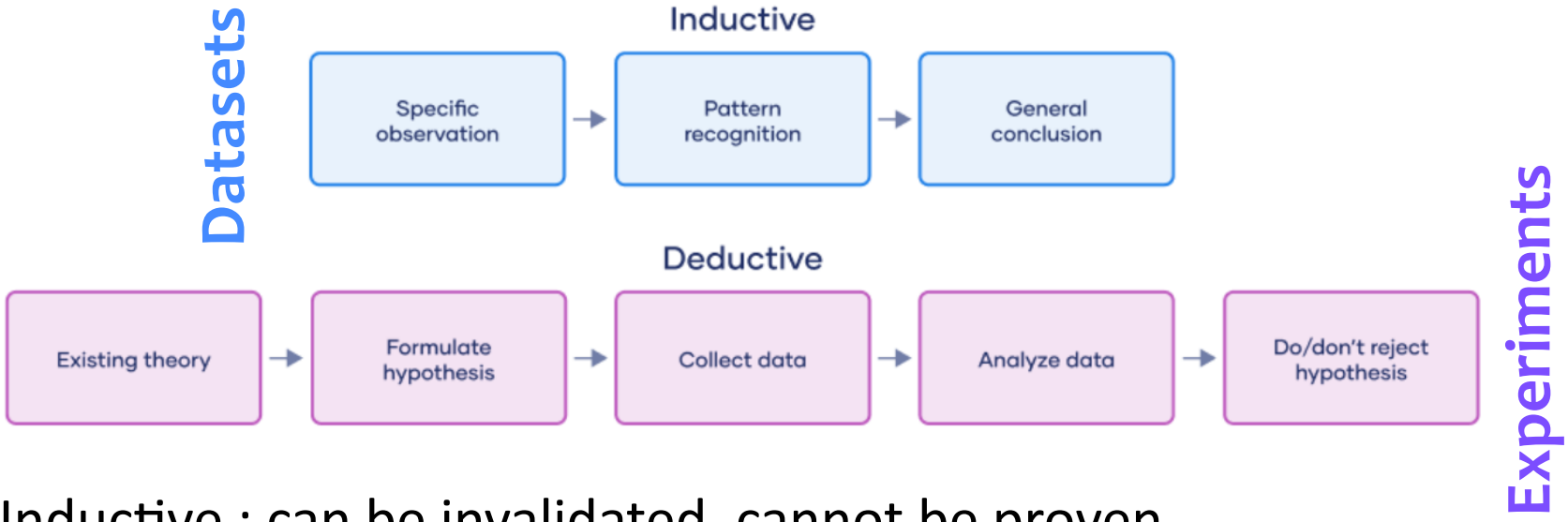
⇒ Tendency to be satisfied with the current « working » model

⇒ Tendency to focus development within on one silo (my own)

[while inter-connected multi-source data is the new paradigm]

Which Advanced Data Products ?

- Inductive vs Deductive Research



  - Inductive : can be invalidated, cannot be proven

    ⇒ **Risk is** to develop useless ADPs

  - Deductive : needs to be based on validated input

    ⇒ **Risk is** to develop too specific/constrained ADPs

Opportunities

Which Advanced Data Products ? No rigid answer

- **Guided developments**
  - Sense of purpose : science-driven, problem solving
  - Injecting knowledge : e.g., instrument, methods, science overview

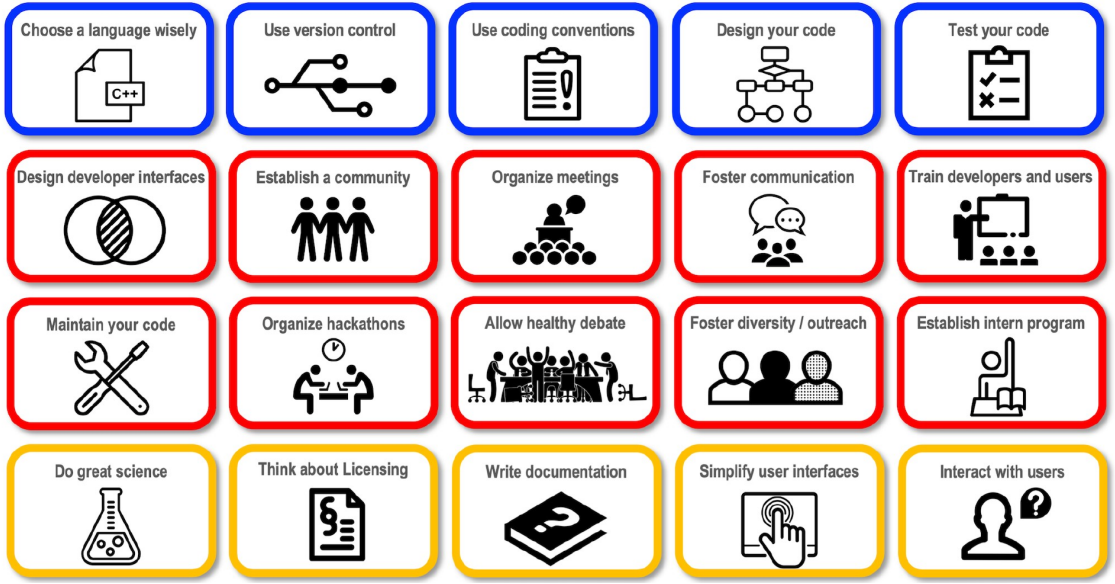
⇒ The observatories/consortia have unique expertise

- **Collaborative and open development**
  - Open framework : allowing entries from the start
  - Flexible framework : lowering the barriers for contributions
  - Expert reviews : can be distributed
  - Observatories can be gate keepers and guides

⇒ **Let them be « The light houses »**

HOWTO = collaborative developments

- **Standards on all sides**
  - Synergies from facilities
    - VO is great ! But not enough.
  - Standards for the development
  - Modularity
  - Tests : version, unit, data
  - Code Review requests
- **Community interactions**
  - Development Meetings, knowledge spread
  - Hackathons, Data Challenges
  - Training



Koehler Leman et al. 2020 ; Rosetta Commons

Skills, resources, people

- Not reinventing the wheel
- **Embed observatory staff in relevant forums**
  - Make them visible actors
  - Let them contribute, guide, expose
    - **And credit them for that**
- Creating bridges between experts
  - Need for embedded professionals
    - The model of collaborative University departments does not work well

« Software/tools cannot be sustained without a community around it »

Koehler Leman et al. 2020

Crediting system

- **Recognition and reward system needs a revolution**
  - Promote existing talents, pro-actively sponsor career paths
- **Incentivise**
  - Creative thinking, diversity of contributions
  - By making such contributions part of our astronomy port-folio
- Some practical things we can do ***(Christopher Steven Marcum)***
  - « Jettison journal titles in CVs and letters of references »
  - « Substitute impact factors with narrative statement »
    - E.g. : free narrative CV templates
  - « Provide parity to non-publication output for values-based assessments »
    - Diversity of contributions

Preservation

- Of Data = not anymore a bit threat (mid-term)

  ⇒ but still an unknown (long-term)

- What about Software ?

Black Box Effect

- ***Describe the methods, algorithms, configuration files***
- Metadata associated with data

  ⇒ Reproducibility (different than reusability)

> ⇒ *The journey is more important than the destination*
> ⇒ **The method is more important than the product**

[But there is no universal answer that applies to all
Sometimes = **better to reobserve, recompute, store ?**]

From hardware, to software, to AdPs

1980's Deliver an instrument → **working hardware**

1990's Space-born facilities = software is key
Ground-based facilities = « after the fact »

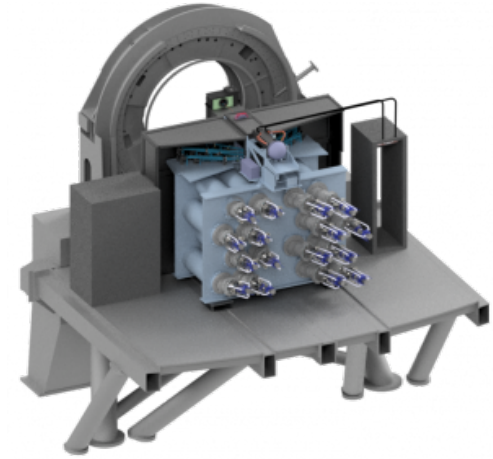
2010's ***New paradigm***
Instrument ≡ hardware + reduction software + operational scheme

Time for

2020's ***An Updated paradigm***
Instrument ≡ hardware + software + operational scheme
≡ ecosystem fostering science results

Opportunities

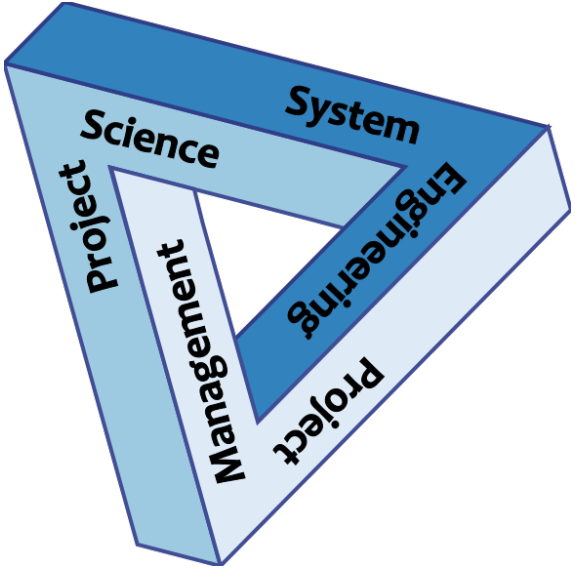
- **MUSE lessons learned**
- E.g., ⇒ **BlueMUSE @ VLT** : PhaseA (**not approved**)
  - How do we incentivise ?
  - How do we connect ?
  - How do we sustain ?



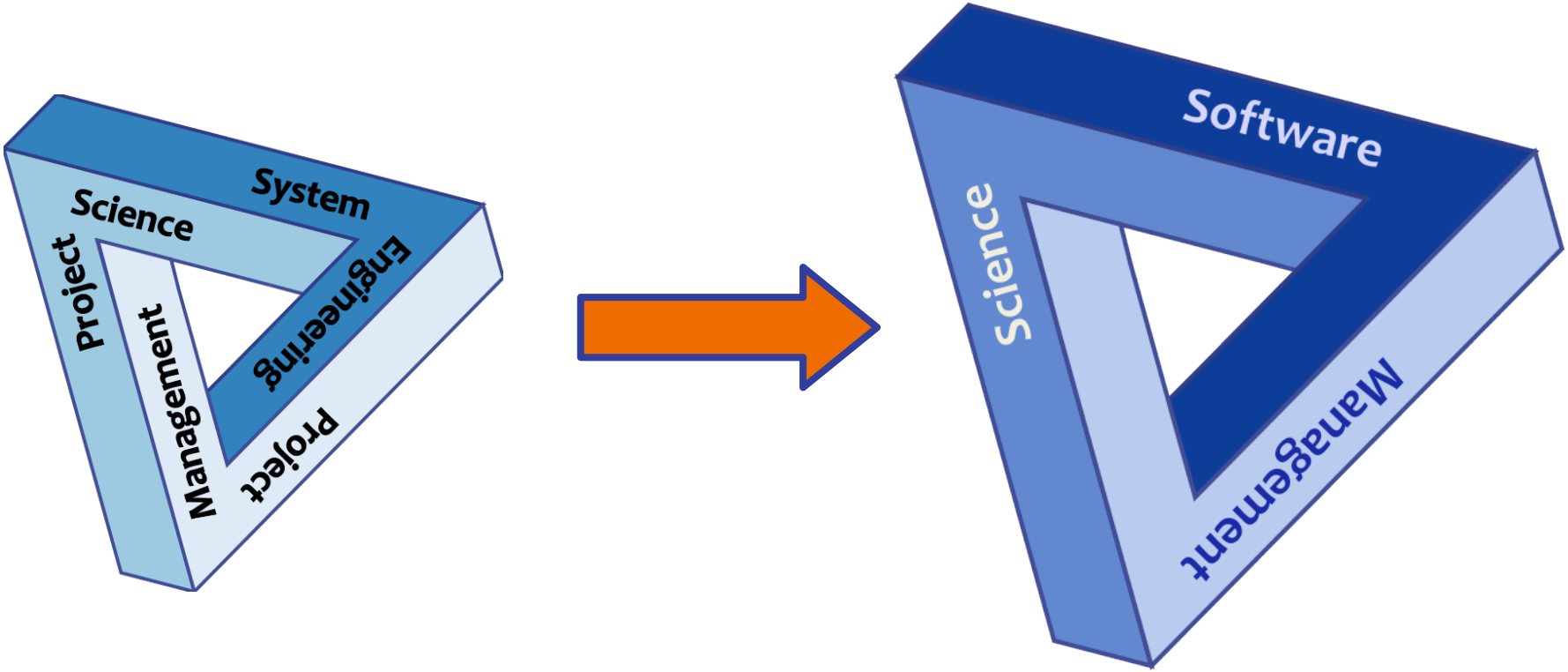
- How do we shape a Data Analysis System for a given facility ?
  - Risk to freese a product, make it irrelevant
  - Risk not to deliver

  ⇒ Give a sense of direction, keeping relevance, adaptability

Management / System Engineering / Science



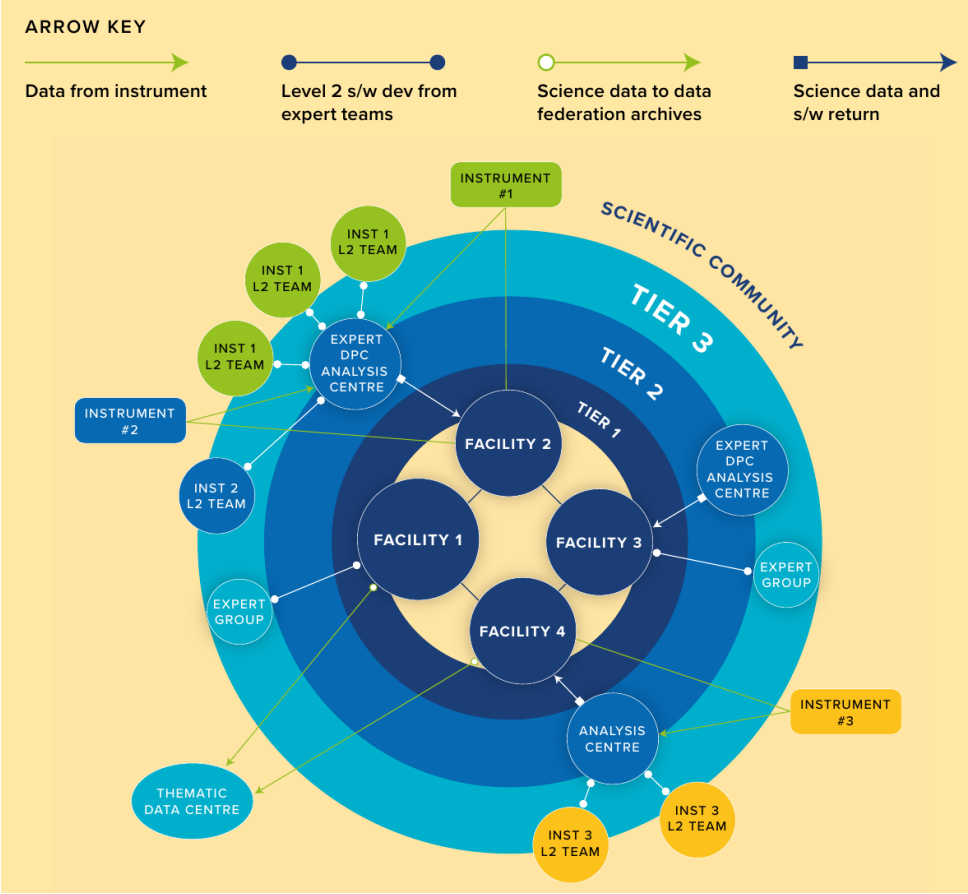
Management / Software Engineering / Science



A Tiered approach

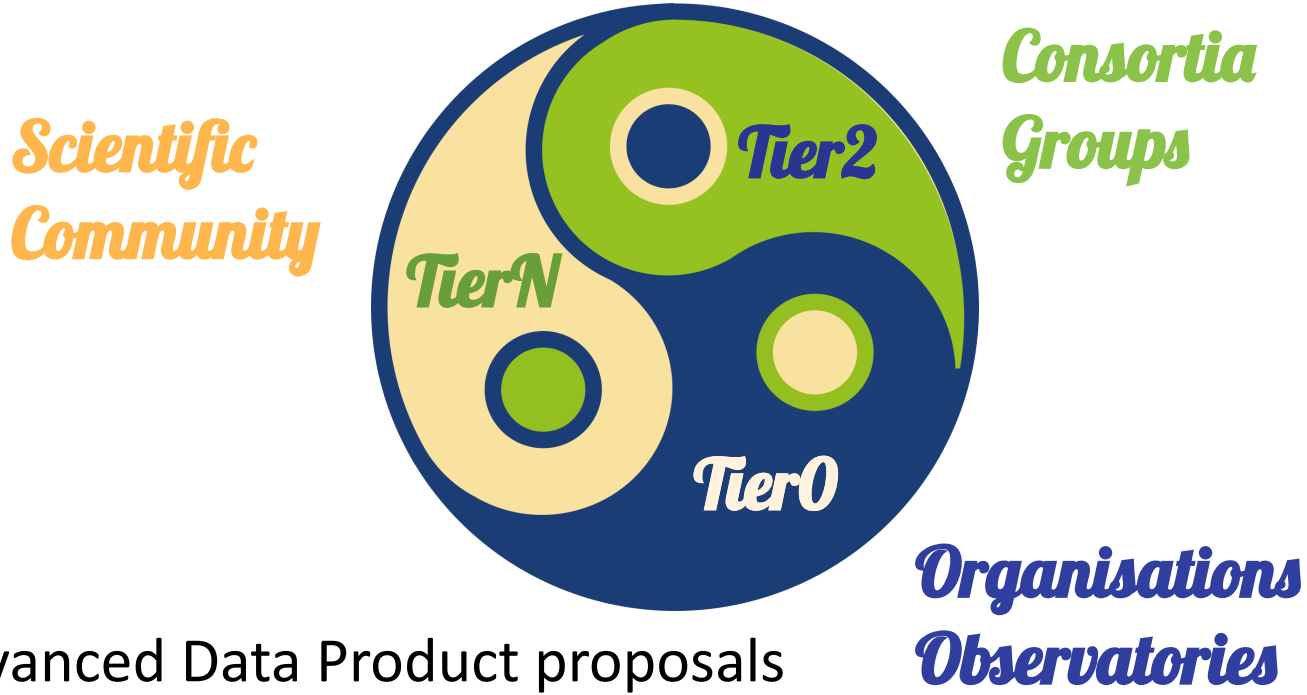
- **Lessons learned**
  - Gaia ⇒ DPAC
  - ALMA ⇒ ARCs
  - … SKA ⇒ SRCs
- **Identified centers of expertise**
  - Mission specific tools
  - More agile development
- **Coordination of a guided effort**
  - Integrated communication between the builders, the coordinators & the scientific community

    ⇒ improved common standard



Emsellem et al. (Fig. 1, Panel Report A)

Engage with the community



- Advanced Data Product proposals
  - Direct connections

Curation ⇒ **[Code and Science Case] Review requests**

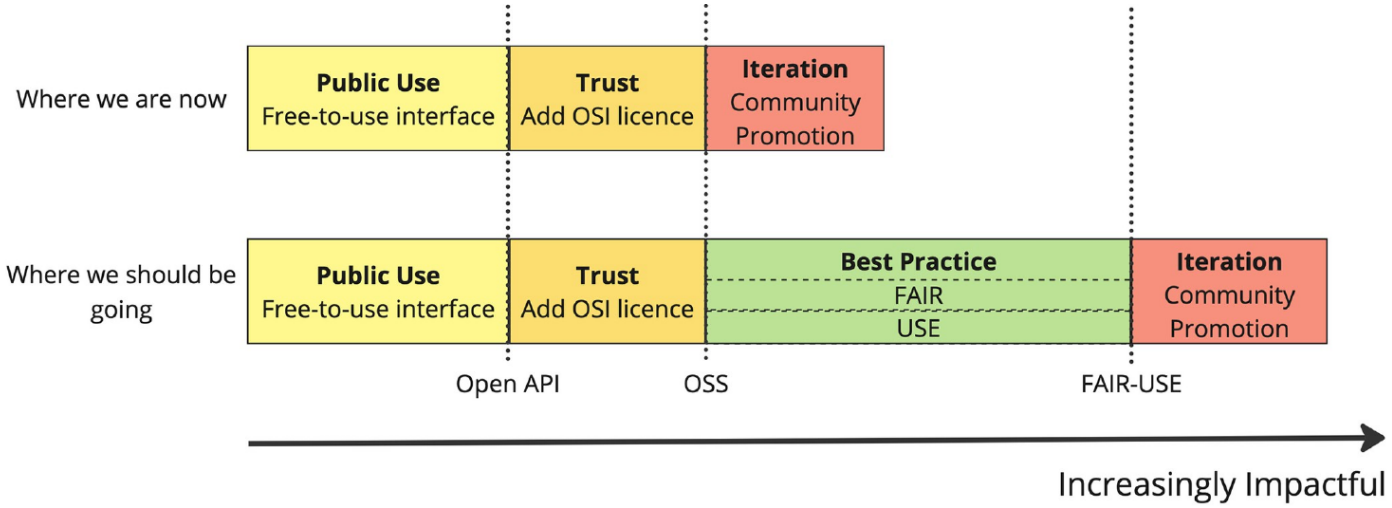
  - Distributed / shared peer review

Lowering the barriers

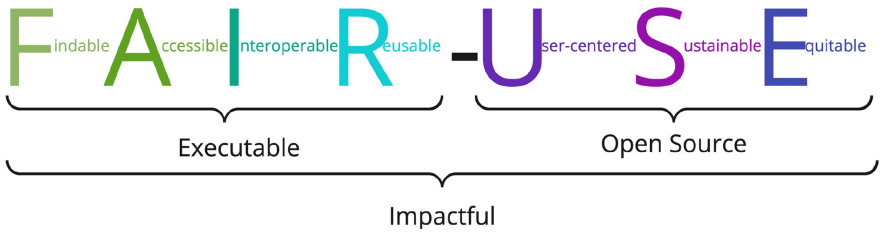
- Challenge of ***inclusiveness & equitable access***
  - The **Open Science** umbrella
- Provide **Data Products & the Tools** to use them
  - Break the « in-house » view of tool development
  - Break the « we cannot do it, let the community do it » view
  - Let's provide a framework for development and guide it
    - Meaning : make sure it exists and is relevant
- **Provide the infrastructure**
  - Integrate datasets into international **data-sharing platforms**
  - Advertise it and provide training, tutorials

    ⇒ will the community adopt it ?

Best practices and more



Sonabend et al. 2024

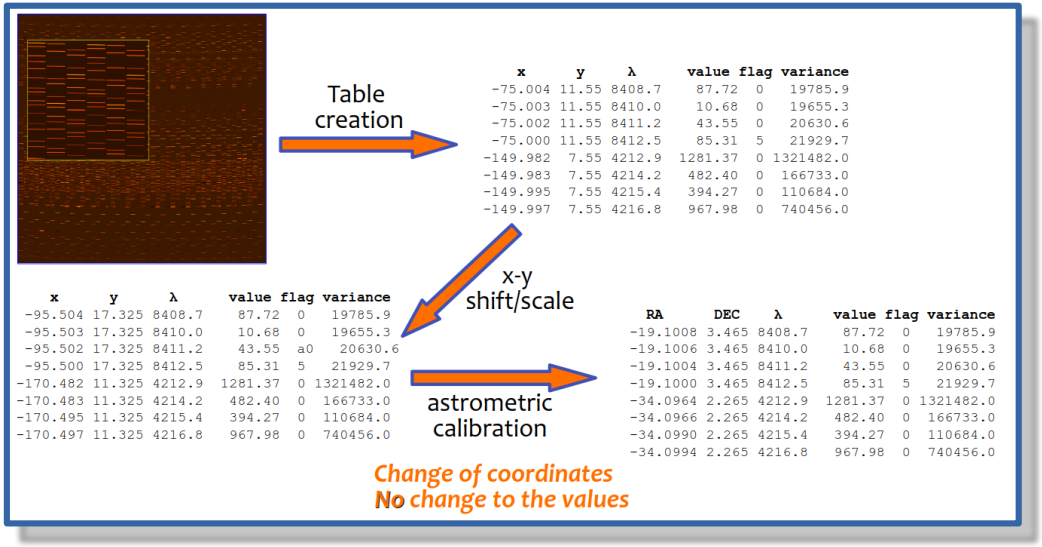


All stakeholders = involved from the start

⇒ to ensure the product genuinely meets the key user needs

One more = the enhanced (spectro)perfectionism

⇒ See the ADPs as an integral part of the data journey



⇒ ADPs : may be a misleading wording

[advanced w.r.t. what?]

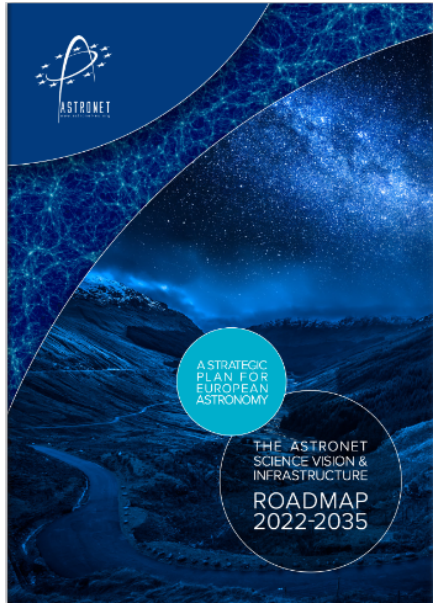
Wrapping up

- ADPs as a way to
  - **Optimise the scientific outcome** of our facilities
  - Add value (legacy), **foster collaborations**, **create an identity**
- A conversation, lowering barriers
  - **Embed observatory**, organisation staff in the community
  - Consider a **Tiered** approach as a supporting, collaborative framework
  - Expose and promote tools
  - **Develop** ADPs and their methods **with** the community **from the start**
- **Upgrade/Change our crediting system**
  - Acknowledge the value of ADP development
  - Create concrete and rewarding career paths
  - Create long-term funding opportunities

Wrapping up

- We need standards, we need testing frameworks
- We need expertise (resources) : Obs. as unique knowledge base
- ADPs (as pipelines) should not be developed in isolation

  (organisation may not wish to turn into data or service providers)



Take-home messages

ADPs acknowledged as

an **integral part of**

a mission design and development

Make

the development and preservation of ADPs

a part of our **astronomical port-folio**