Astro-NDFI

Astrophysics and Astroparticle Physics
in the National Research Data Infrastructure

Matthias Steinmetz (@GalacticRAVE)

- Big Data and Machine Learning in Astronomy and Astrophysics
- The National Research Data Infrastructure
- Astronomy in the National Research Data Infrastructure

Astronomy from the historical perspective

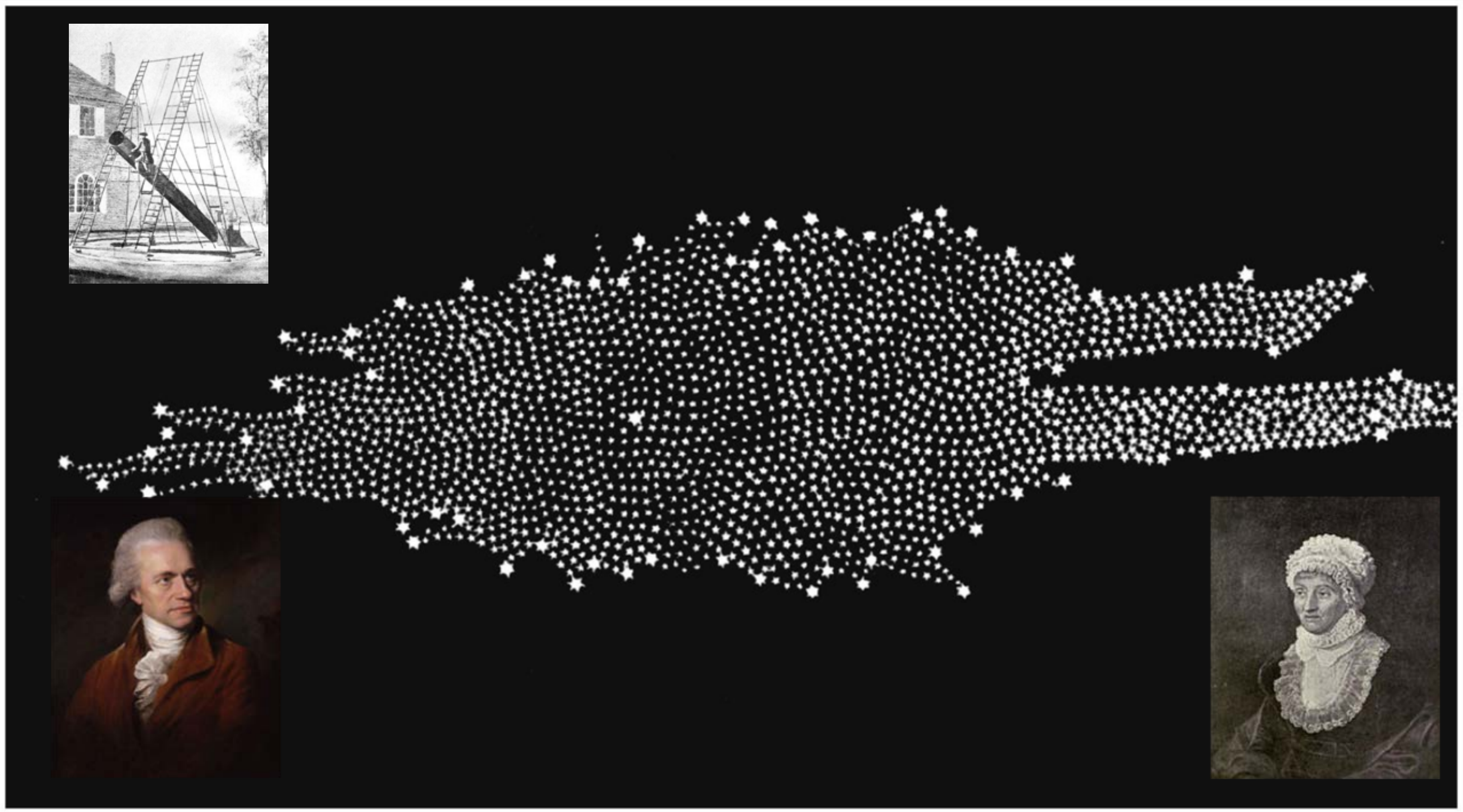
- static sky (fixed stars), about 6000 stars visible with the unaided eye ⇒ navigation

- a few moving objects (Sun, Moon, 5 planets)
⇒ time keeping, calendar

- occasionally unexpected events (comets, novae, supernovae)

Late 18th century: first map of the Galaxy

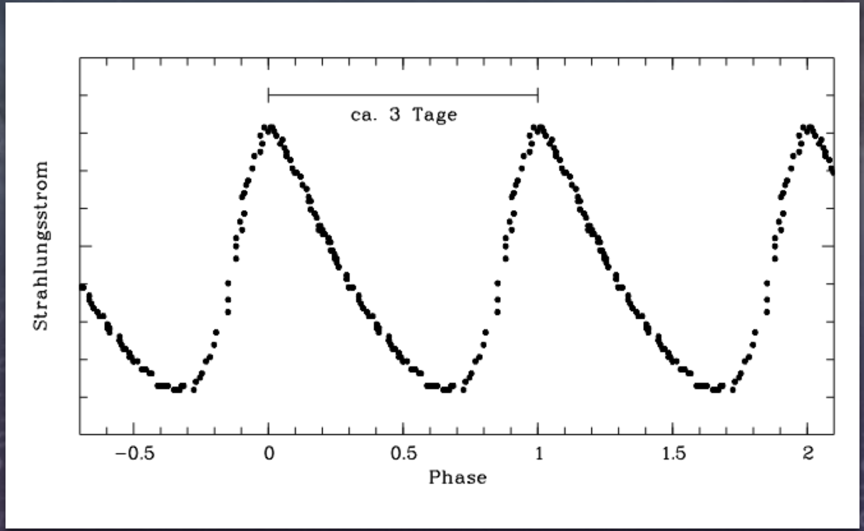
single object ⇒ surveys (Herschel, Bonner Durchmusterung)

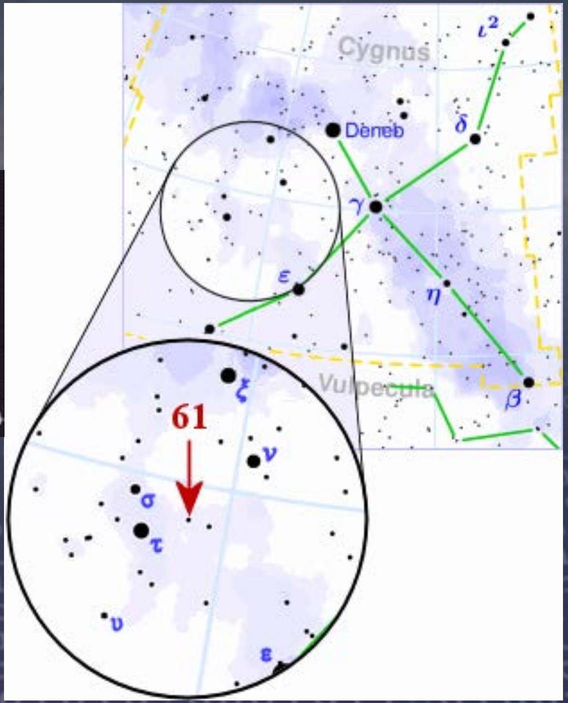


Fixed stars are not fixed!

- proper motions - stars move
- variable stars
- novae (supernovae)
- parallaxes

⇒ Distances, Astronomy becomes 3D





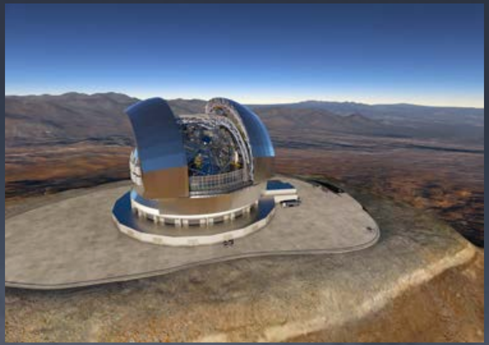
⇒ „time domain“

Major inventions in the 19&20th century

- Spectroscopy (1859) - Astrophysics
  - finding out what objects are made of, line-of-sight motion
  - finding the medium in between stars and galaxies
- Photography (1850s)
  - Recording of observations
  - Analysis after the observation (working with archived data)
  - increasing sensitivity
- CCD (1980s)
  - digital readout
  - huge gain in efficiency
  - linear detectors (working below the sky background)
- New wavelength domains (radio, IR, UV, X-ray, γ-ray)
- New messengers (Cosmic Rays, Neutrinos, Gravitational Waves)
  - Astroparticle Physics

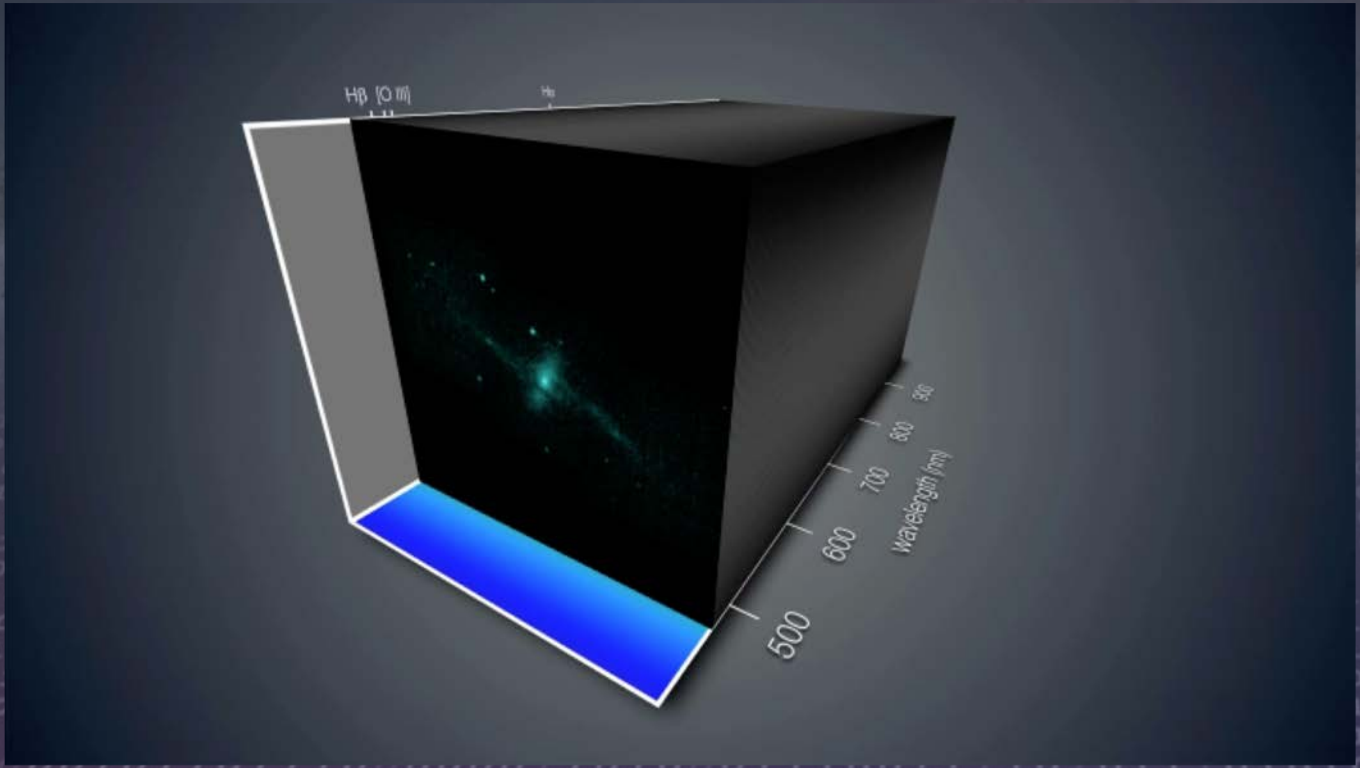
Facilities in the 2010s: 8m telescopes

classical observatories „point and stare“



Facilities in the 2010s: MUSE@VLT

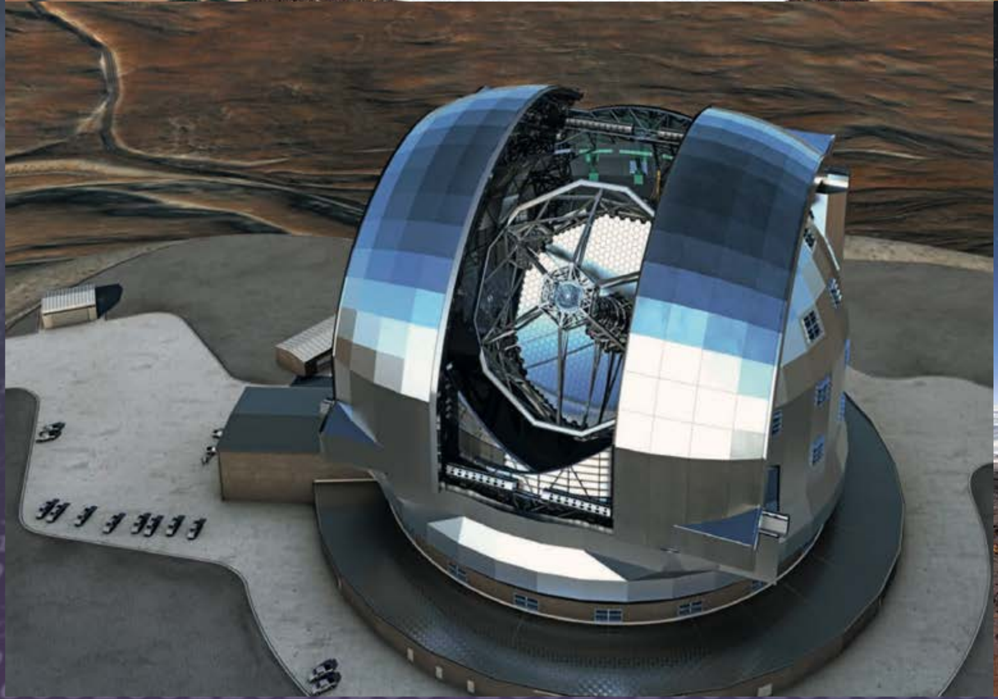
Integral Field spectroscopy:
merging imaging and spectroscopy

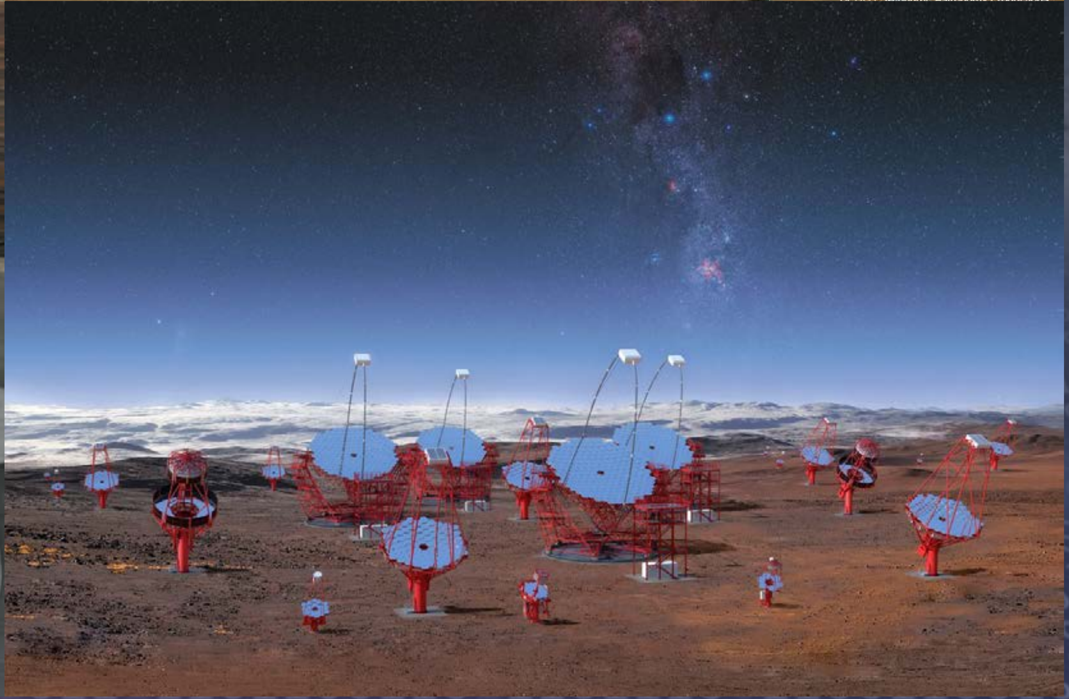


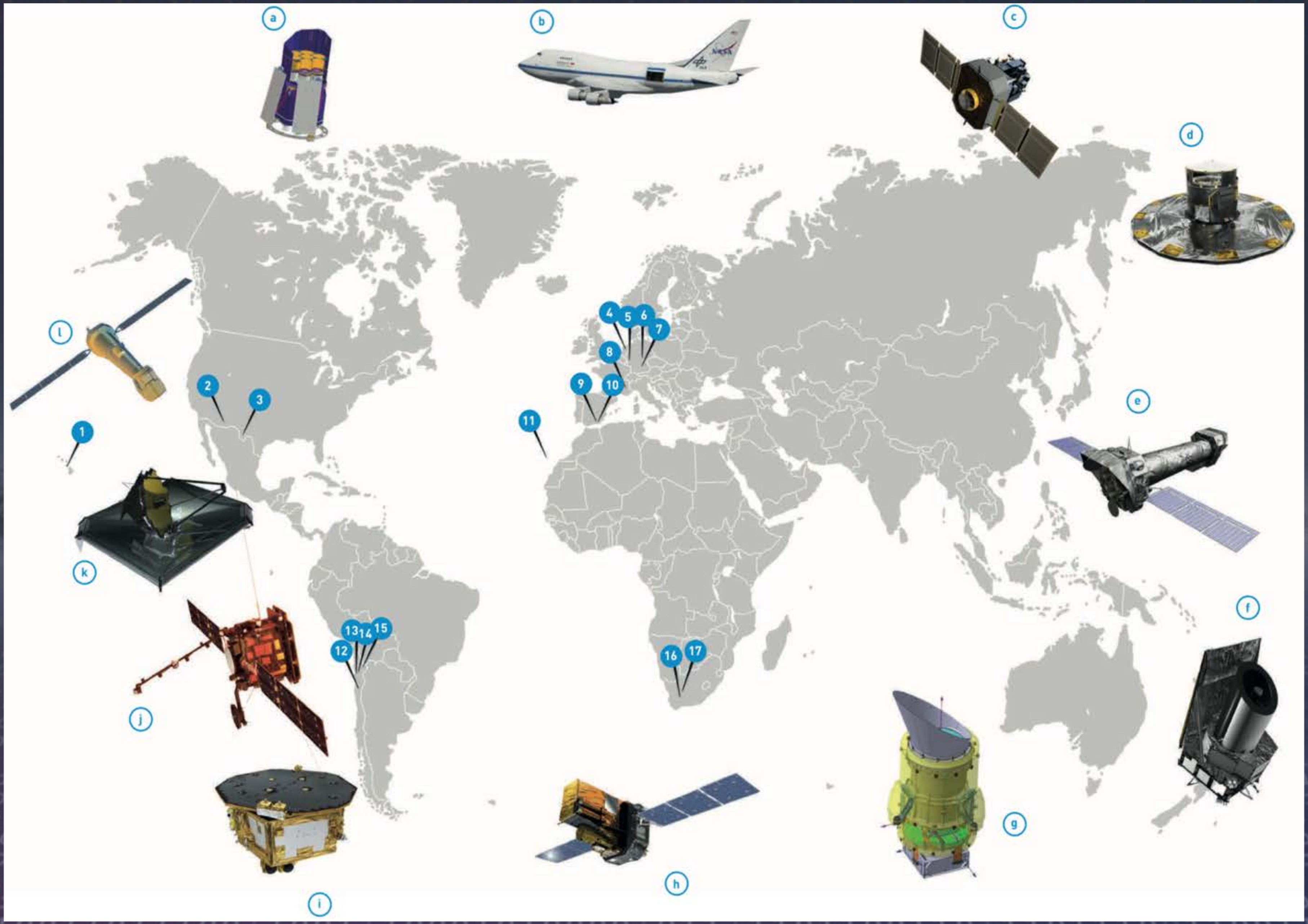
Facilities in the 2010s: Time domain

Monitoring: automated and robotic telescopes

SPDO / Swinburne Astronomy Productions






a
b
c
d
e
f
g
h
i
j
k
l
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17

Example Event Horizon Telescope:

- Rawdata: 20 PB/week
- Actual Image: 1MB

DER SPIEGEL

Nr. 16
13.4.2019

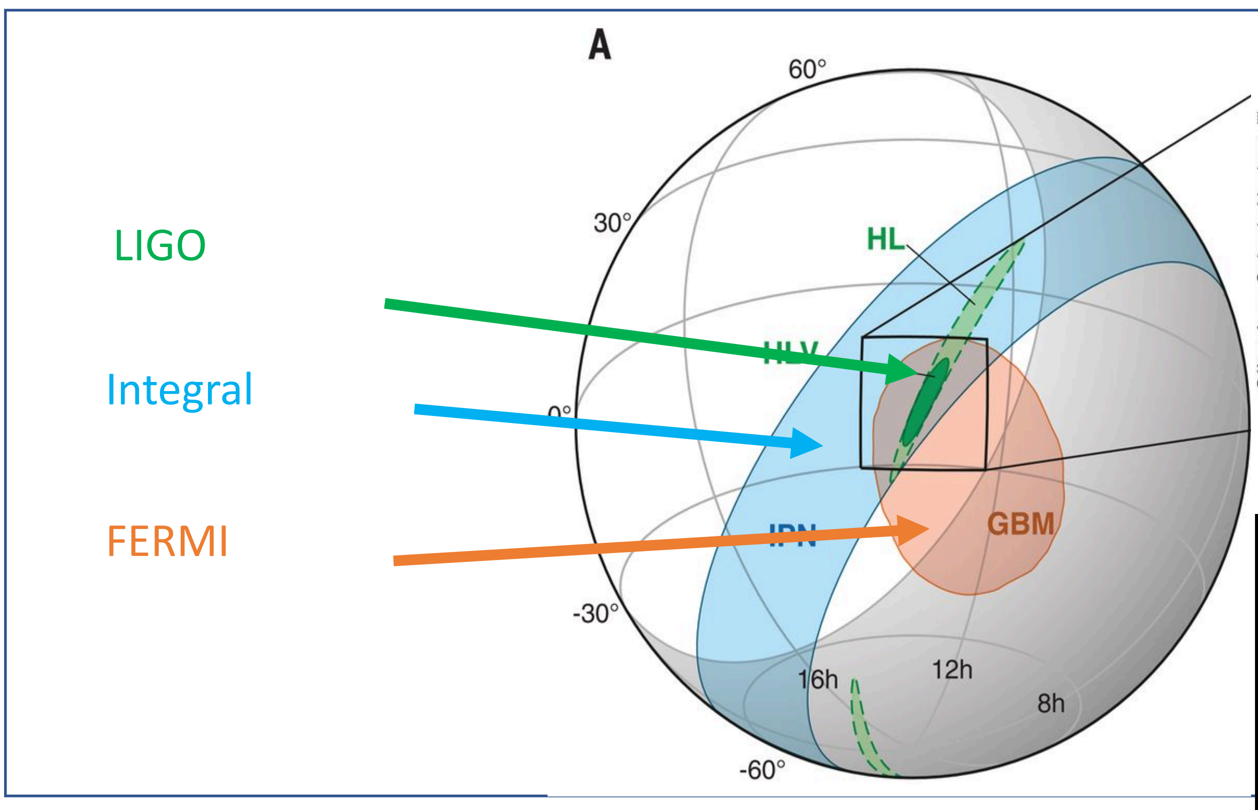
Am Ende von Raum
und Zeit

Was uns schwarze Löcher über
die Geheimnisse des Universums verraten

HÄUSERKAMPF Wie viel Kapitalismus
verträgt der Wohnungsmarkt?

Challenge Square Kilometre Array (SKA)

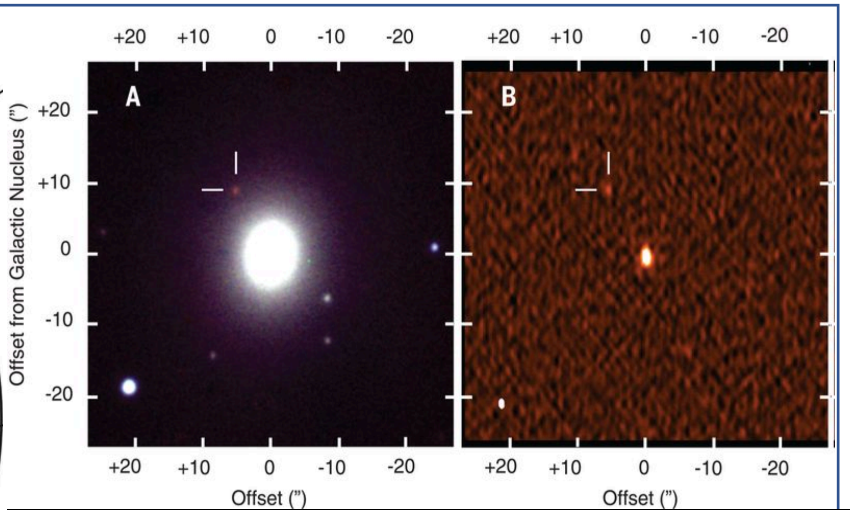
- Rawdata: 5 EB/day
- Archive: 0.3-0.5 EB/yr

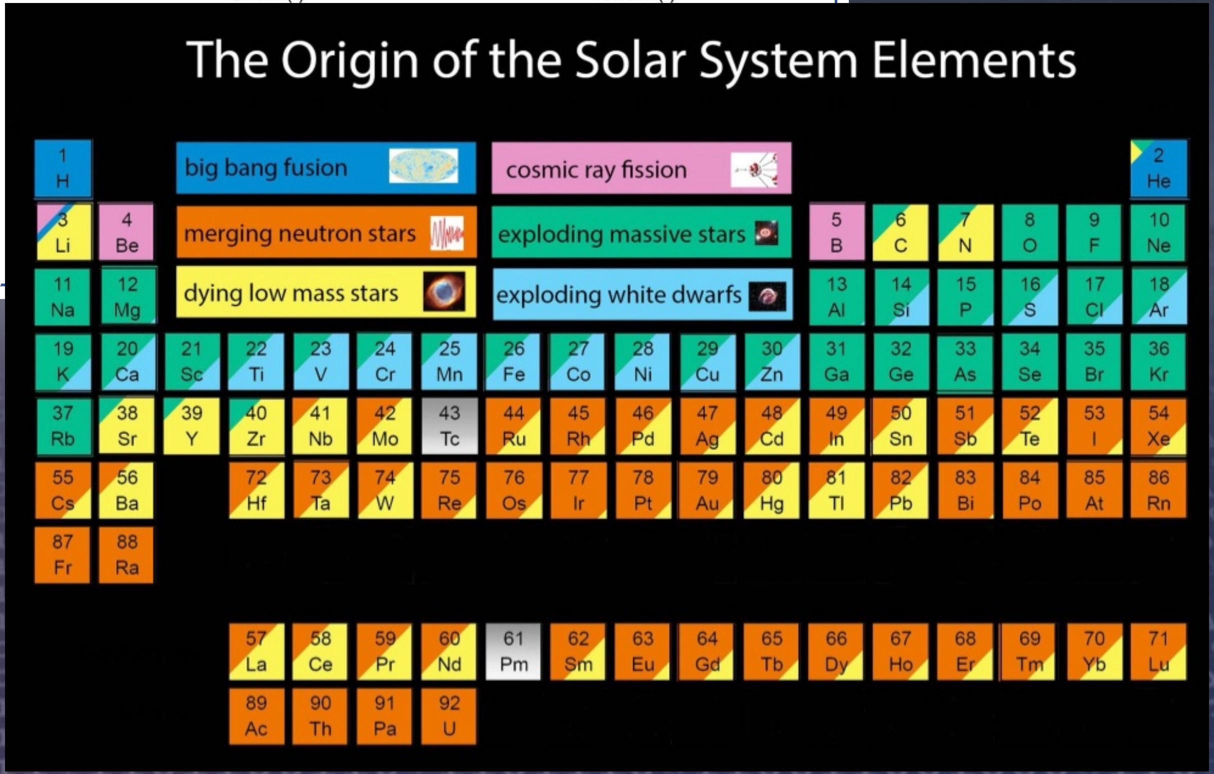


LIGO

Integral

FERMI





The Origin of the Solar System Elements

big bang fusion

cosmic ray fission

merging neutron stars

exploding massive stars

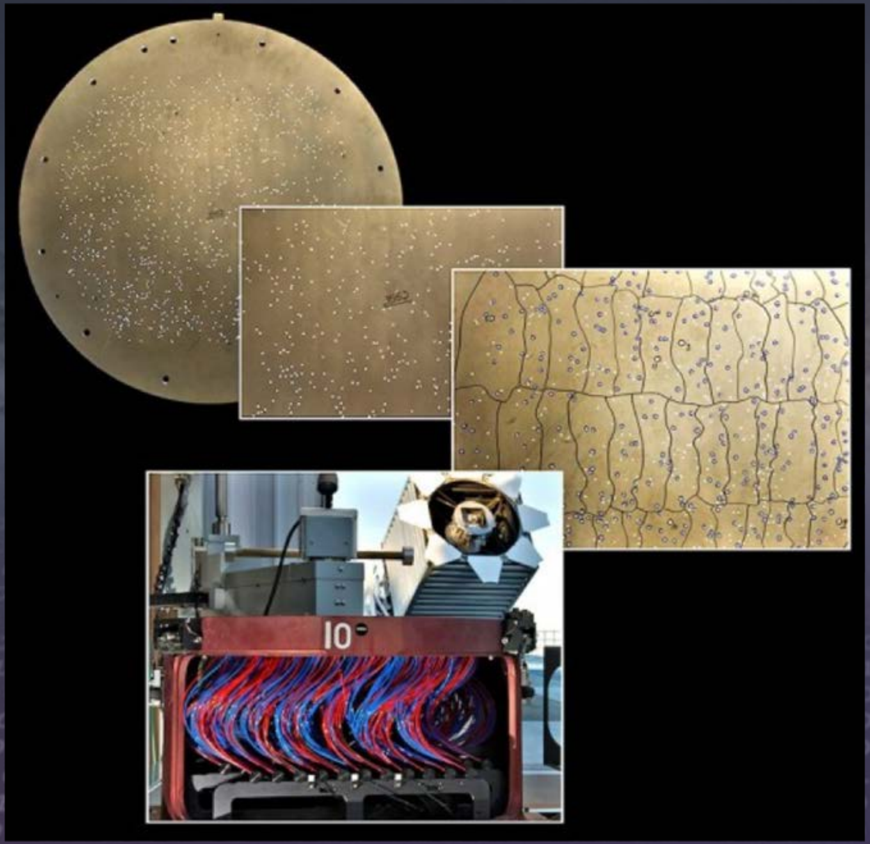
dying low mass stars

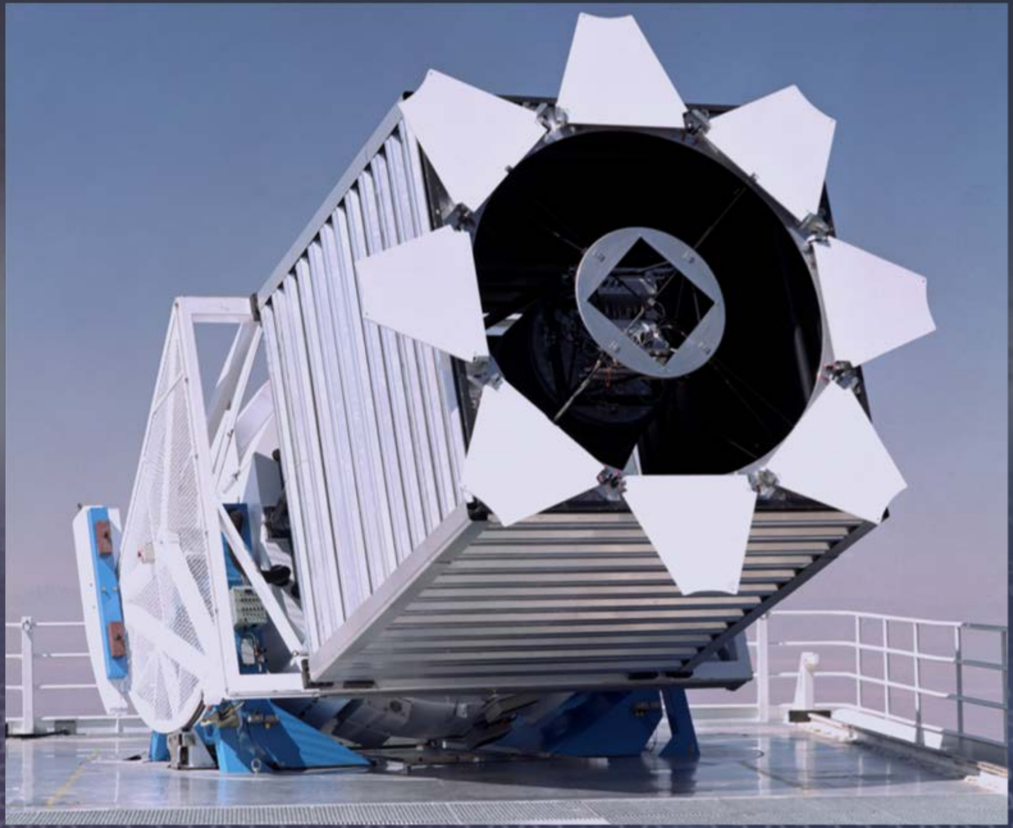
exploding white dwarfs

1 H																	2 He
3 Li	4 Be											5 B	6 C	7 N	8 O	9 F	10 Ne
11 Na	12 Mg											13 Al	14 Si	15 P	16 S	17 Cl	18 Ar
19 K	20 Ca	21 Sc	22 Ti	23 V	24 Cr	25 Mn	26 Fe	27 Co	28 Ni	29 Cu	30 Zn	31 Ga	32 Ge	33 As	34 Se	35 Br	36 Kr
37 Rb	38 Sr	39 Y	40 Zr	41 Nb	42 Mo	43 Tc	44 Ru	45 Rh	46 Pd	47 Ag	48 Cd	49 In	50 Sn	51 Sb	52 Te	53 I	54 Xe
55 Cs	56 Ba		72 Hf	73 Ta	74 W	75 Re	76 Os	77 Ir	78 Pt	79 Au	80 Hg	81 Tl	82 Pb	83 Bi	84 Po	85 At	86 Rn
87 Fr	88 Ra																
			57 La	58 Ce	59 Pr	60 Nd	61 Pm	62 Sm	63 Eu	64 Gd	65 Tb	66 Dy	67 Ho	68 Er	69 Tm	70 Yb	71 Lu
			89 Ac	90 Th	91 Pa	92 U											

Facilities in the 2010s: Surveys

Sloan Digital Sky Survey: Digital Revolution in Astronomy





18 years of Sloan Digital Sky Survey

- 2000-2005 SDSS-I
- 2005-2008 SDSS-II
- 2008-2014 SDSS-III
- 2014-2020 SDSS-IV (incl. southern hemisphere
- 2020-2025 SDSS-V (tbc)

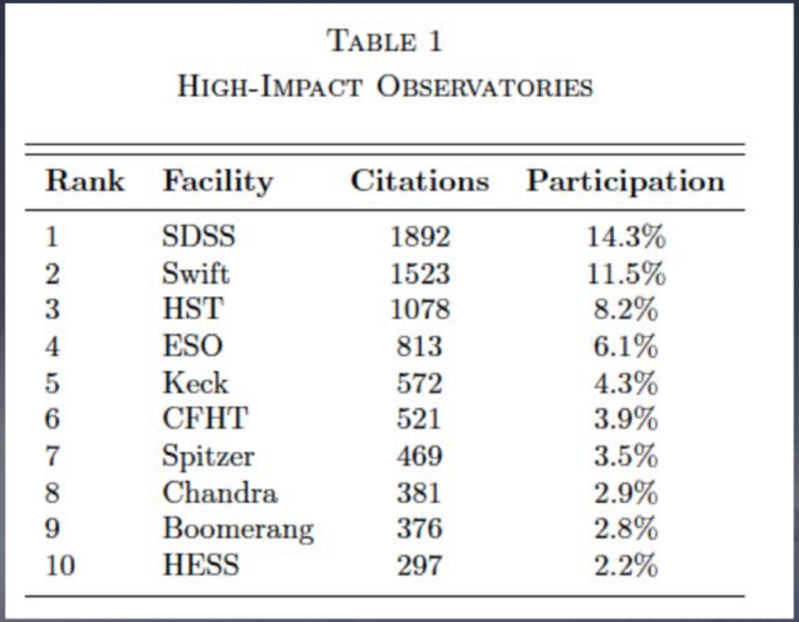
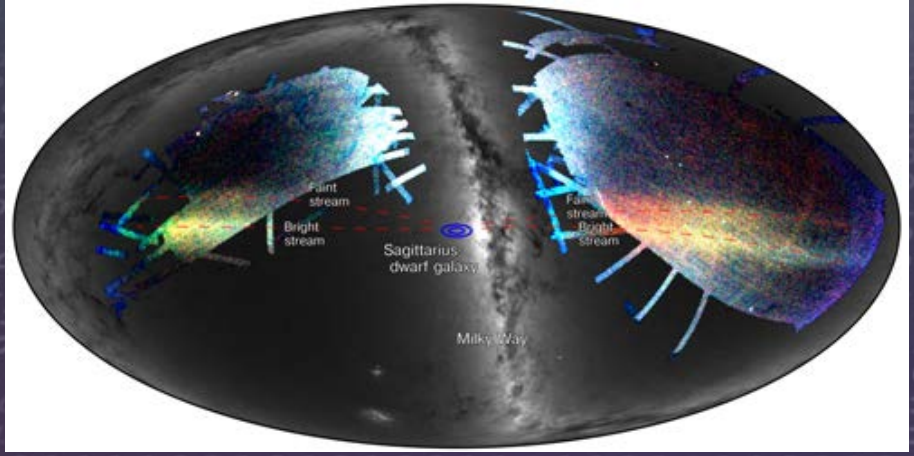


TABLE 1
HIGH-IMPACT OBSERVATORIES

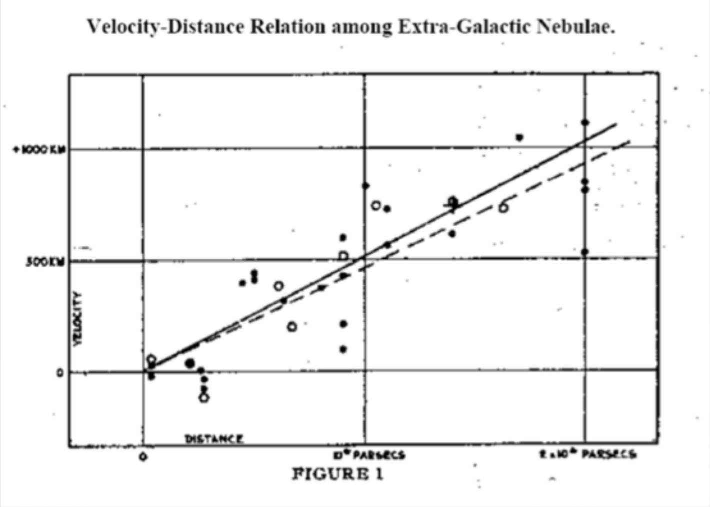
Rank	Facility	Citations	Participation
1	SDSS	1892	14.3%
2	Swift	1523	11.5%
3	HST	1078	8.2%
4	ESO	813	6.1%
5	Keck	572	4.3%
6	CFHT	521	3.9%
7	Spitzer	469	3.5%
8	Chandra	381	2.9%
9	Boomerang	376	2.8%
10	HESS	297	2.2%



among the highest impact astronomical observatories:

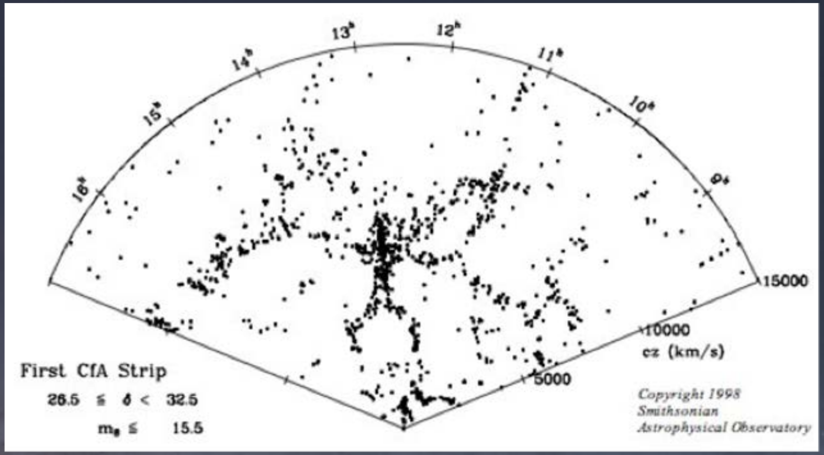
- 13 data releases
- 5000+ scientific papers
- cited collectively more than 200000 times
- h-index: 180

# of galaxies for which we have redshifts



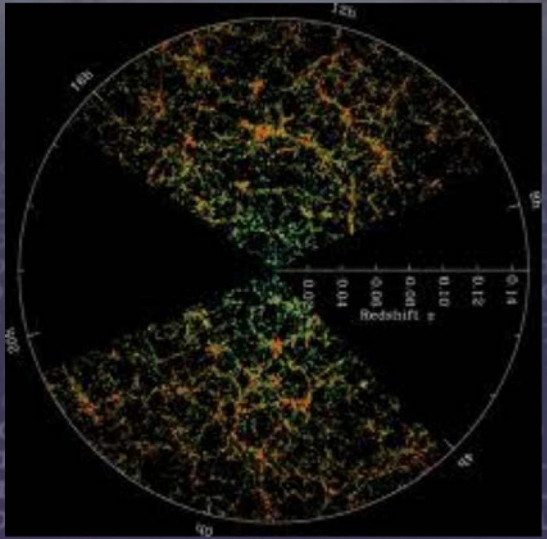
1929: 24 galaxies

x50 in 60 years



1986: 1100 galaxies

x 1000
in 25 years



2008: 1 million galaxies

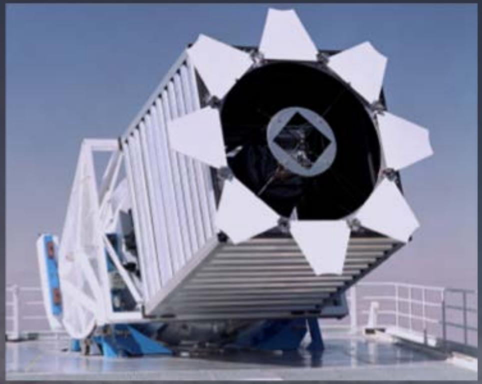
SDSS: high level data products (catalogues)

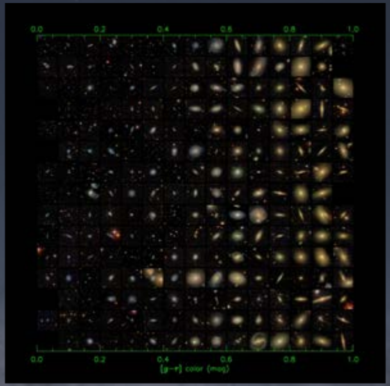
- Catalogue:
  - a list of all detected objects (stars, galaxies)
  - measured parameters (size, color, brightness)
- The utility of sky maps:
  - Discovery: Is this a new asteroid or is it already catalogued?
  - Classification: What type of galaxies exist?
  - Populations: Do quasars change their properties with time?
  - Rare Objects: Is this a very weird object?
  - Cosmology: How fast does the universe expand?
- Recipe for success:
  Science ready database - measurements can be analyzed w/o need for complex image processing

ASTRONOMERS AT WORK
Let's dispel some common beliefs.
I NEVER USE A TELESCOPE
I NEVER GO NEAR A TELESCOPE
I NEVER EVEN LOOK UP.
S.Harris

The next decade: LSST

- (Large Synoptic Survey Telescope): combine survey and time domain
- SDSS did a color picture of the sky





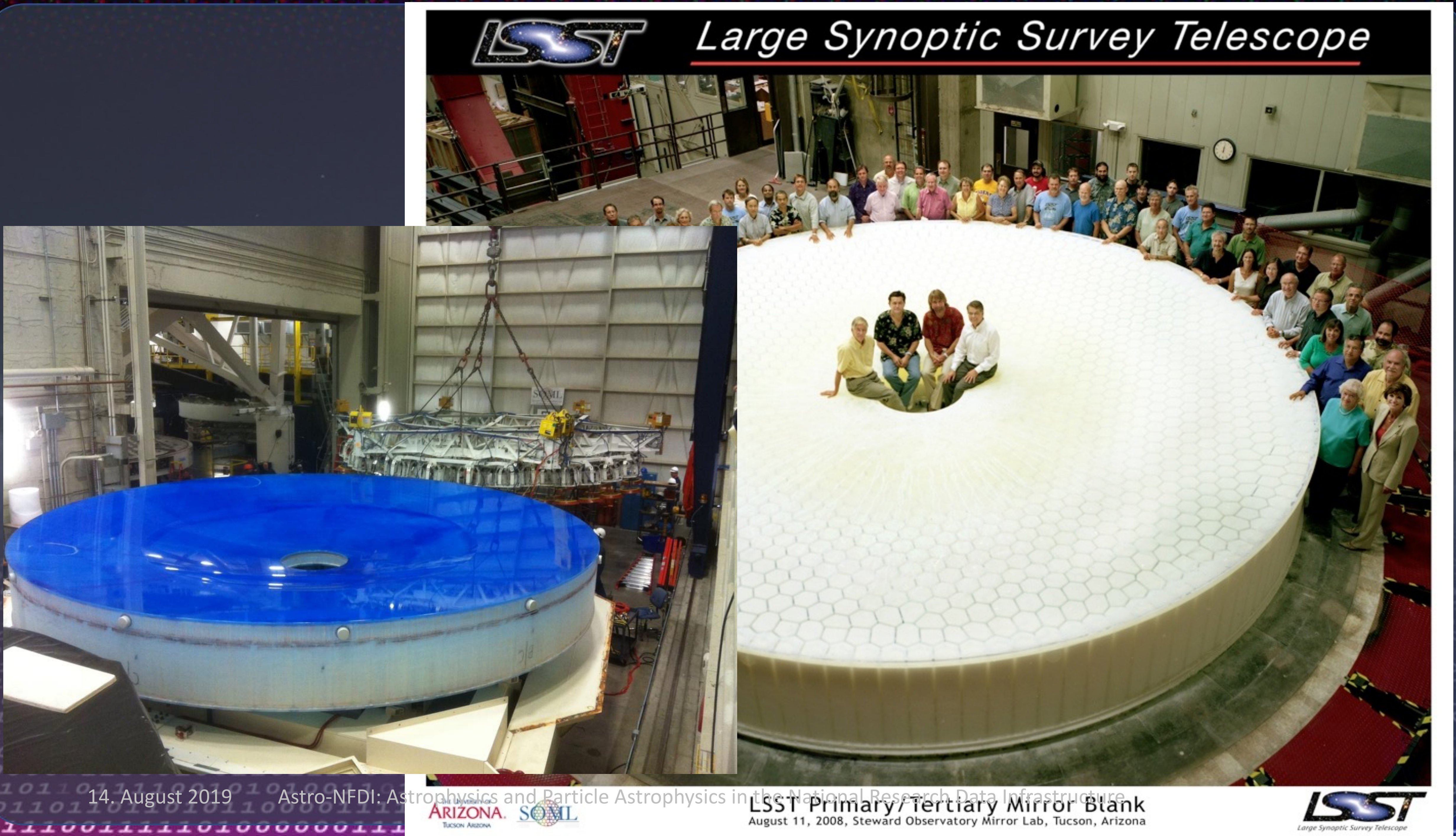
- LSST will do a color movie of the sky

An optical/near-IR survey of half the sky in ugrizy bands to r~27.5 based on ~1000 visits over a 10-year period

A catalog of 20 billion stars and 20 billion galaxies with exquisite photometry, astrometry and image quality!

First light: 2020

Mar 10, 2019

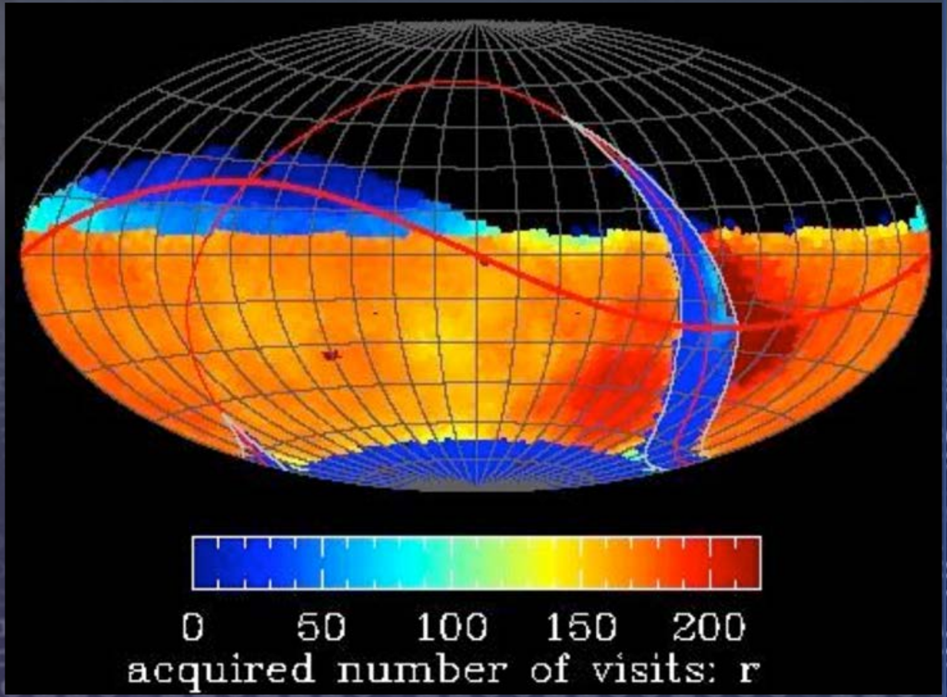

LSST
Large Synoptic Survey Telescope
SOML
The University of Arizona
Tucson Arizona
Large Synoptic Survey Telescope

LSST Primary/Tertiary Mirror Blank
August 11, 2008, Steward Observatory Mirror Lab, Tucson, Arizona

Basic idea behind LSST: a uniform sky survey

- 90% of time will be spent on a uniform survey: every 3-4 nights, the whole observable sky will be scanned twice per night
- after 10 years, half of the sky will be imaged about 1000 times (in 6 bandpasses, ugrizy): a digital color movie of the sky
- ~100 PB of data: about a billion 16 Mpix images, enabling measurements for *40 billion objects!*

Right: a 10-year simulation of LSST survey: the number of visits in the r band (Aitoff projection of eq. coordinates)



LSST From the User's Perspective: A Data Stream, a Database, and a (small) Cloud

Prompt Data Products

- A stream of ~10 million time-domain events per night, detected and transmitted to event distribution networks within 60 seconds of observation.
- A catalog of orbits for ~6 million bodies in the Solar System.

Level 1

Yearly Data Releases

- A catalog of ~37 billion objects (20B galaxies, 17B stars), ~7 trillion single-epoch detections ("sources"), and ~30 trillion forced sources, produced annually, accessible through online databases.
- Deep co-added images.

Level 2

User Contributed Data Products

- Services and computing resources at the Data Access Centers to enable user-specified custom processing and analysis.
- Software and APIs enabling development of analysis codes.

Level 3

Big data is open data!

© Zeljko Ivezic

1. Early data releases greatly improve the final product
2. Early data releases enable coeval science
3. More science is extracted from the same dataset
4. Enables reproducibility of science results
5. Synergy between different datasets
6. Cross-disciplinary science
7. More citations and prestige to the team
8. Education and public outreach

The National Research Data Infrastructure

Some Background on NFDI

- So-called „Digital Agenda“ is a priority for the past and current federal government
- Joint Research Conference (GWK = federal and state research ministries) earmarked 90M€ p.a. for the next 10 years to establish a national research data infrastructure (NFDI)
- Funds will be dispersed in 3 rounds, AO for the first one issued in June 2019
- Infrastructure aspect: networking and quality assurance rather than competition:

  „Im Auswahlverfahren für die antragstellenden Konsortien wird mit Blick auf eine interagierende, nutzerorientierte Informationsinfrastruktur ein Schwerpunkt auf die Vernetzung und die Abstimmung der Konsortien untereinander und auf eine Weiterentwicklung und Verbesserung der Anträge im Laufe des Auswahlprozesses gelegt“
- Consortium not well defined (b/c heterogeneity of the research landscape)
- Applicants: University and publicly funded research institutions
  - Funds go to the coordinating institution and will be dispersed from there
- Expectation: ~30 consortia
- GWK (joint federal + state science conference) involvement

  ⟹ may lead to institutional funding (?)

Important Points for a Consortium Proposal*

- Consortia and NFDI as a whole:
  - Comprehensive research data management
  - increase efficiency throughout the scientific system
- Consortia in NFDI:
  - What is the added value we bring to the overall structure?
  - Mandate properly – the essential needs of a community should not be distributed over several consortia
- Consortia provide a set of services for their community/domain:
  - selected, maintained and operated in joint responsibility by the consortium partners
  - services that – demonstrably! – are solutions for specific methodological requirements
  - generic services with added value to the NFDI or
  - own services and tools, or integration of services operated elsewhere

*after P. Gehring (RfII): Putting the NFDI into Practice

Important Points for a Consortium Proposal*

- Define what “science-driven” means in and for the community
- Identify a range of essential services
- Integrate all relevant players/research “nodes” in the domain at an early stage
- Ensure the coordinated advancement of the selected (and future) services
- Establish procedures to prioritize future services and developments
- It’s a structure, not a funding program - joint and potentially difficult responsibilities must be taken over the long term

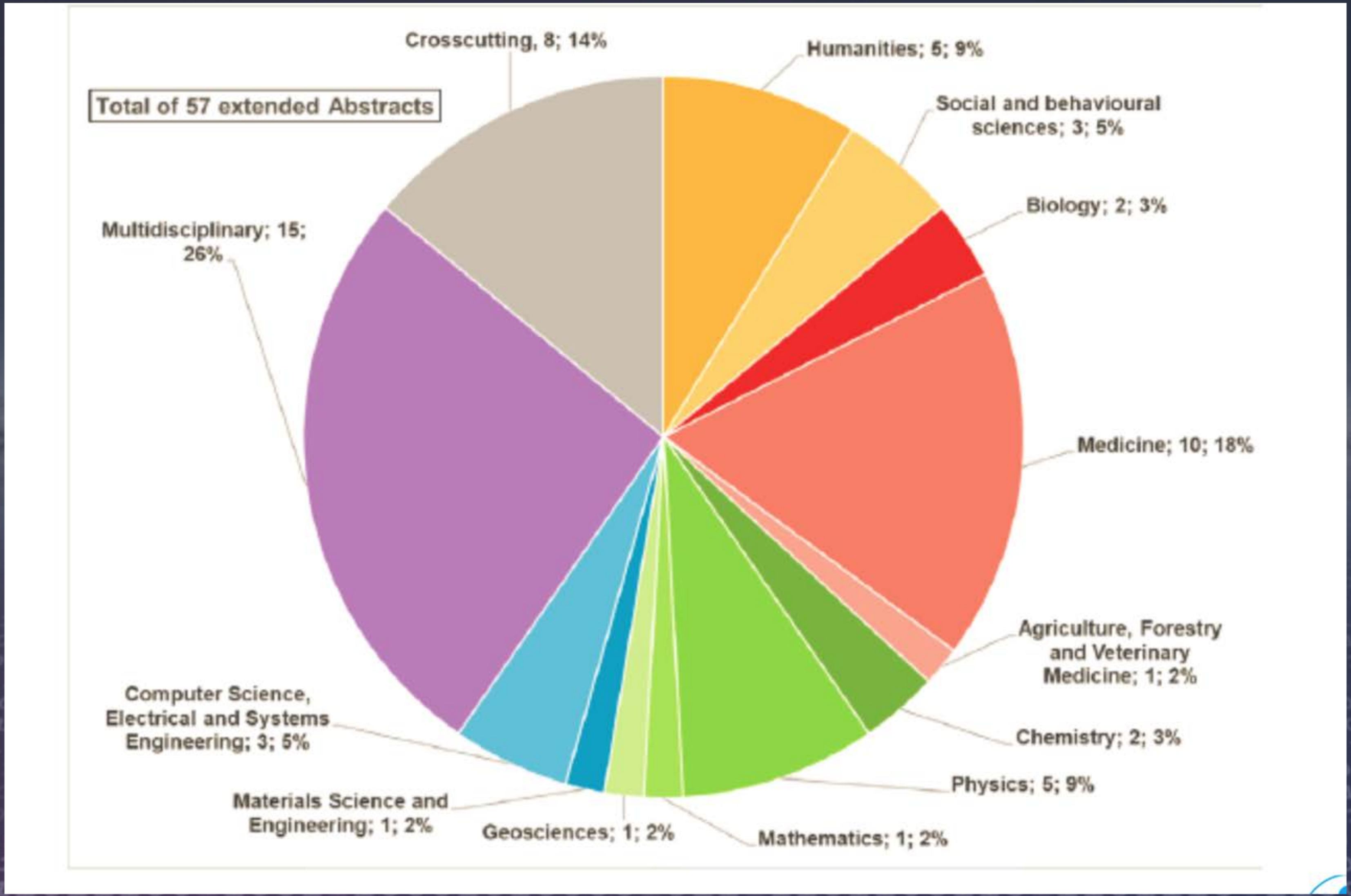
*after P. Gehring (RfII): Putting the NFDI into Practice

Important points for a Consortium Proposal*

- Participation appears sufficiently important and rewarding from a researcher’s perspective
- Efficient participation structure for the researchers who use the services
- Divergent requirements of data users and data producers are managed
- Different groups within the user community have a balanced voice

*after P. Gehring (RfII): Putting the NFDI into Practice

First round of LoIs (DFG, Eickhoff)



What NFDI is not:

- a centralized archive a la CDS
- a Verbundforschung-like project of a sub-community with particular interests

  „eine interagierende, nutzerorientierte Informationsinfrastruktur … Schwerpunkt auf die Vernetzung und die Abstimmung der Konsortien untereinander“
- the funding for your favourite facility or research project
- everything related to concrete
- clusters and supercomputers

Roadmap towards NFDI

- 29.3.2019 Letters of interest due for consortia
- 13./14.5.2019 NFDI Conference at DFG in Bonn
  - Expert committee: 20 members
- June 2019: Announcement of Opportunity
  - Applicant, co-Applicants, Participants
  - Eligible costs include cost for personnel, operations, travel, project-related external contracts. Contributions by the applicants and participants in terms of research data management are expected. Hardware etc. can only be financed in exceptional cases. ⟹ people, not silicon or concrete
  - Funding is for 5 years
- June/July: binding letters of intent
- October: applications due
- June 2020: GWK decides on first grants
- July 2020, 2021 conference for round 2 and 3

Astro-NDFI

Astrophysics and Astro-Particle Physics in the National Research Data Infrastructure

Current situation

- Two major initiatives in the area of particle physics, nuclear physics, astro-particle physics and astrophysics, supported by their respective committees: KET, KHuK, KAT and RDS
    - KET, KHuK and KAT
    - NFDI-Astro RDS and KAT
- each representing ~2000-3000 PhD scientists
- these two consortia are in regular contact
- further interactions, e.g., geosciences (planetology)

Who we are:

- About 3000 researchers in the area of astronomy, astrophysics and particle astrophysics ("astrophysics") in Germany
- About 50 university and non-university institutions (Helmholtz, Leibniz, Max-Planck) + HPC Centers
- Tightly linked to international and intergovernmental organization and facilities (e.g. ESO, SKA, CTA, EST)
- Organized in the Council of German Observatories (RDS) and the Committee for Astro-Particle Physics (KAT)

RDS

KAT. Komitee für Astro. Teilchen. Physik

Denkschrift 2017

Perspektiven der Astrophysik in Deutschland 2017-2030

Von den Anfängen des Kosmos bis zu Lebensspuren
auf extrasolaren Planeten

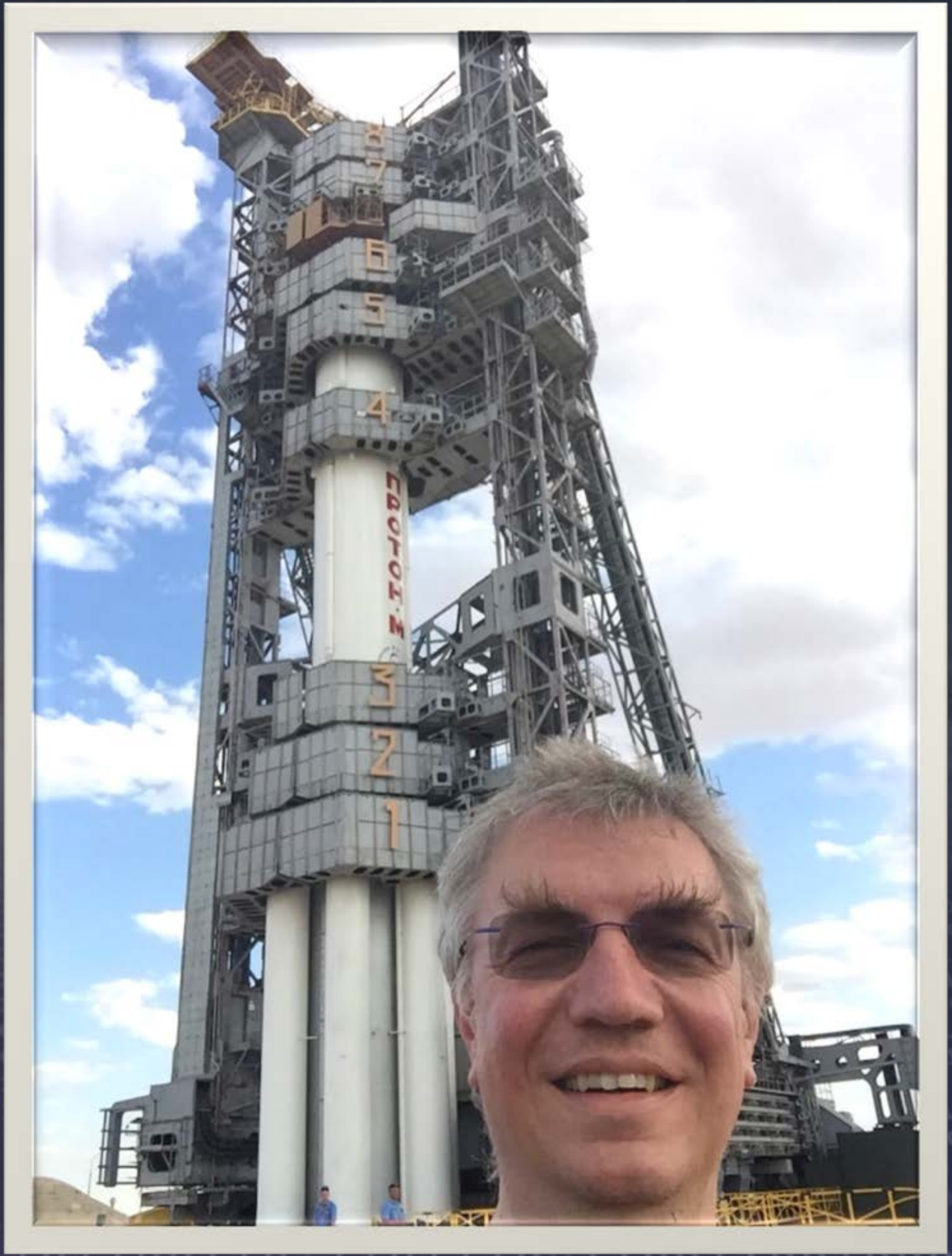
Matthias Steinmetz, Marcus Brüggen, Andreas Burkert, Eva Schinnerer, Jürgen Stutzki,
Linda Tacconi, Joachim Wambsganß, Jörn Wilms (Redaktionskomitee des Rats deutscher Sternwarten)

Astrophysics is a Data and Discovery Science

- FAIR: Findable, Accessible, Interoperable and Reusable is key!
- Historically:
  - Copernican revolution (long term observations)
  - Halley’s comet (inference from different records)
  - Neptune’s discovery (combining theoretical predictions & excellent maps)
- Recent examples
  - Hubble Deep Fields (from exoplanets to cosmology)
  - 50% of publications based on Hubble Telescope data based on archival work
  - Gravitational Waves by merger of two neutron stars (multi-messenger approach)
- Future:
  - “Data Avalanche” (Exabyte/yr) by large Surveys and next Generation Telescopes & next generation of simulations. This has already started! (e.g. LOFAR, EHT)
  - Data mining and deep learning

Our Data …

- Diverse in nature and multi-use
  - Electromagnetic (radio, submm, Optical/IR, UV, X-ray, Gamma-ray)
  - High energy particles (protons, hadrons, electrons, neutrinos)
  - Gravitational Waves
  - Time Domain Data
  - Computer Simulations
  - Laboratory Astrophysics (Astrochemistry, Astrobiology)
- Large data volumes (hitting technological boundaries)
  - Distributed & dynamic, non-conservative compression
  - Deep learning methods for archive coherence and fast & flexible interoperation
- What makes this interesting for others:
  - Mostly open data (except proprietary periods of ~1 year)
  - Lots of it (currently some 100 PBytes), in particular imaging data



ПРОТОН-М

Why is Astronomy „special“ for Big Data

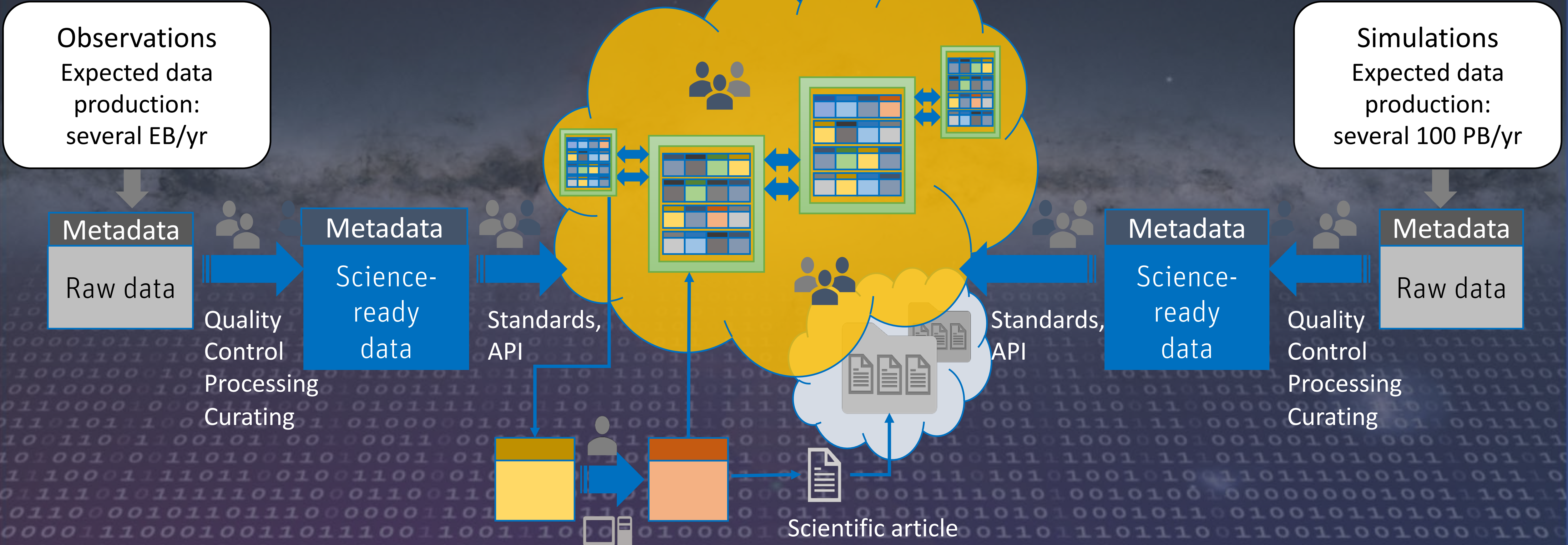
© Jim Gray/Alex Szalay

- high appeal for the public
- no commercial value
  - no major data rights or privacy issues (but security: satellite positions and orbits)
  - ideal to experiment with algorithms
- actual data with all its problems and issues
  - multi-dimensional
  - distributed in space and time
- diverse and distributed
  - many observatories
  - at different sites
  - at different times
- There are lots of it (soon 100s of petabytes)

Our experience …

- Long tradition in data archives and data mining
  - Now: deep learning & ever increasing data sets & data rates
- FITS: evolving quasi-standard for data and meta data for 40 years
  - TIFF image format is a well known offspring
  - Interesting synergies (e.g. FITS is used for digitization of the Vatican library)
  - Non-trivial converter in other communities (e.g. particle physics data)
- Virtual Observatory: ongoing development since 20 years
  - Standardized access, interoperability (partially FAIR)
- involved in major international data initiatives,
  - EOSC and ESCAPE research data management for ESFRI projects
- Cooperation with commercial sector (Microsoft, Google, amazon, SAP …)

Eco-system of scientific data



Observations
Expected data production:
several EB/yr

Metadata

Raw data

Quality Control
Processing
Curating

Metadata

Science-ready data

Standards, API

Scientific article

Standards, API

Metadata

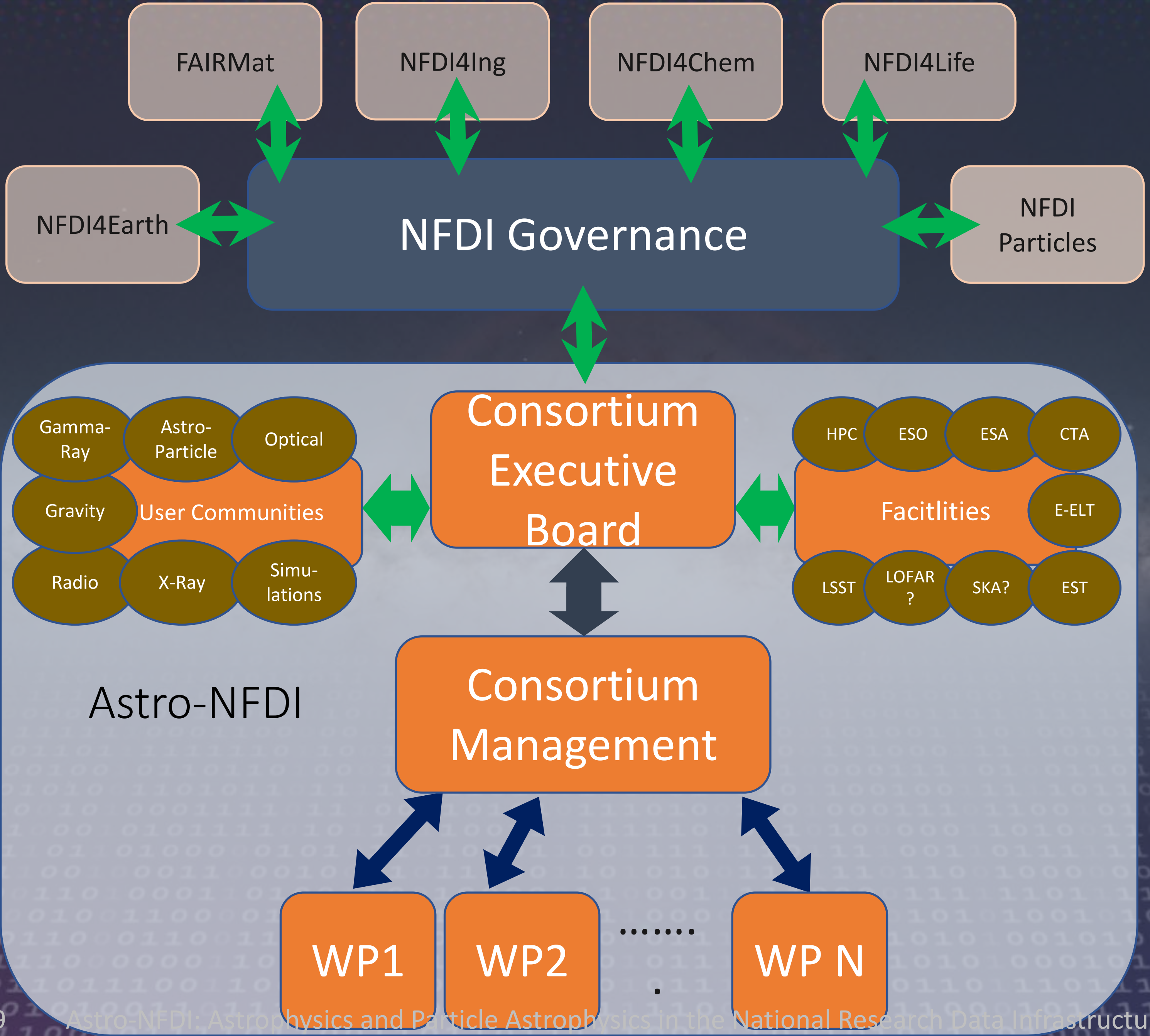
Science-ready data

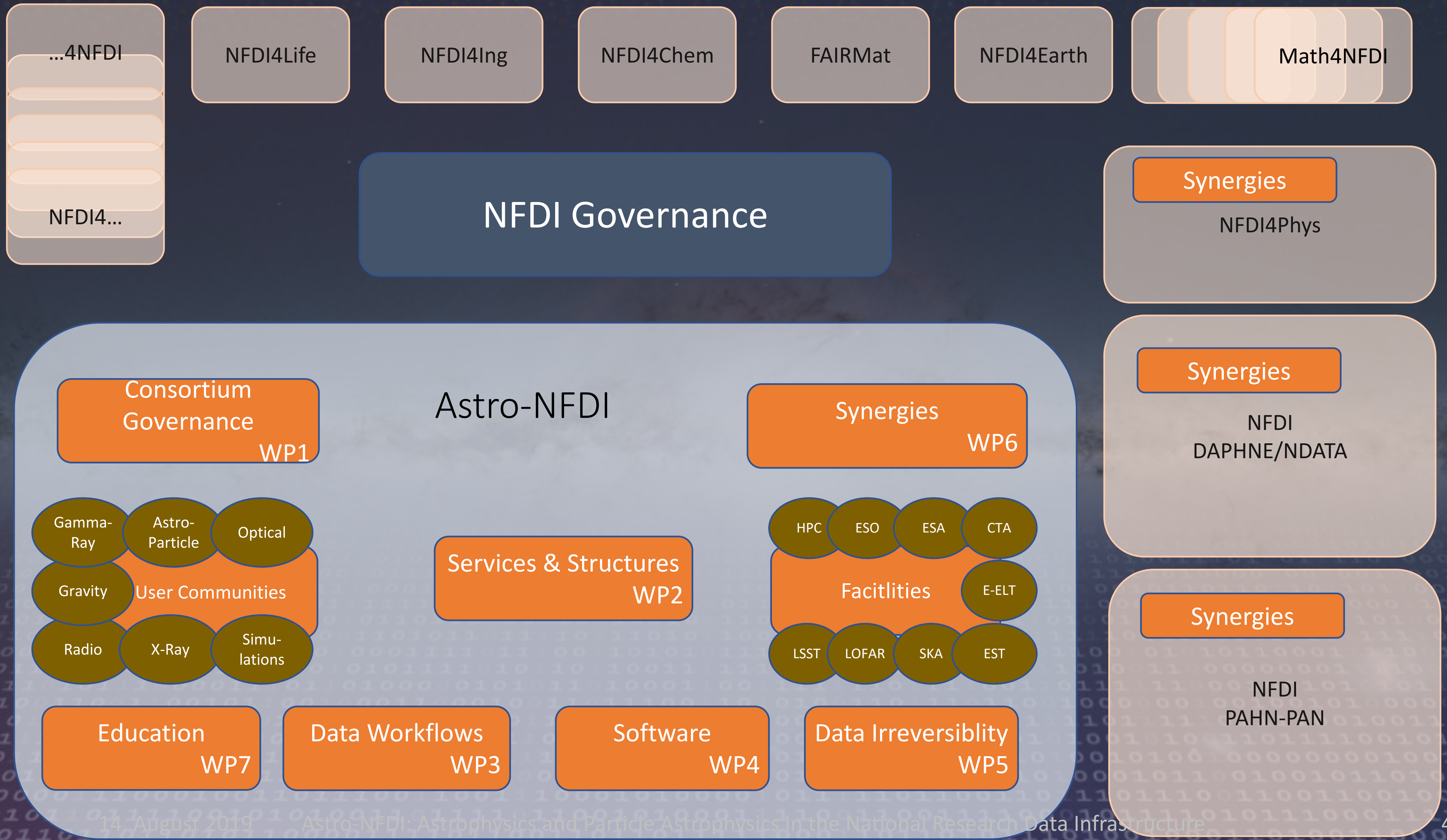
Quality Control
Processing
Curating

Metadata

Raw data

Simulations
Expected data production:
several 100 PB/yr


FAIRMat
NFDI4Ing
NFDI4Chem
NFDI4Life
NFDI4Earth
NFDI Governance
NFDI Particles
Gamma-Ray
Astro-Particle
Optical
Gravity
User Communities
Radio
X-Ray
Simu-lations
Consortium Executive Board
HPC
ESO
ESA
CTA
Facitlities
E-ELT
LSST
LOFAR ?
SKA?
EST
Astro-NFDI
Consortium Management
WP1
WP2
…….
.
WP N



...4NFDI

NFDI4...

NFDI4Life

NFDI4Ing

NFDI4Chem

FAIRMat

NFDI4Earth

Math4NFDI

NFDI Governance

Synergies

NFDI4Phys

Synergies

NFDI
DAPHNE/NDATA

Synergies

NFDI
PAHN-PAN

Astro-NFDI

Consortium Governance
WP1

Synergies
WP6

Gamma-Ray

Astro-Particle

Optical

Gravity

User Communities

Radio

X-Ray

Simu-lations

Services & Structures
WP2

HPC

ESO

ESA

CTA

Facitlities

E-ELT

LSST

LOFAR

SKA

EST

Education
WP7

Data Workflows
WP3

Software
WP4

Data Irreversiblity
WP5

Our NFDI needs

- Large internationalization (like many other communities)
  - NFDI must not be a national island solution
- Building sustainable competence centers for (astronomical) data
  - Data management (curation, provenance, publication) & Data publication software
  - Solutions for ‚last dirty mile‘ (small data collections)
  - Well defined, internationally compatible interfaces to big projects
- Extension of FAIR data policies in our discipline
  - Interoperable, interdisciplinary standards und metadata, DOI
- Code to the Data:
  - Deep learning
  - Distributed data processing and Data Mining
  - Dynamical Archives: Resolving lossy data problems and managing ‚live data‘
- Scientific Software Curation
  - supporting generic Open Source Software (e.g. astropy, gammapy)
  - managing data and connected software as units

What we can bring to the NFDI

- Experience with across the discipline data and meta-data formats
- Scientific & methodical proximity:
  - particle and nuclear physics (PAHN-PaN)
  - Photon Science (DAPHNE)
  - Physics: NFDI4Phys
- Imaging data: many possible synergies with medical imaging, geosciences (also libraries?)
- Memory Driven computing: Human Genome Archive (GHGA)

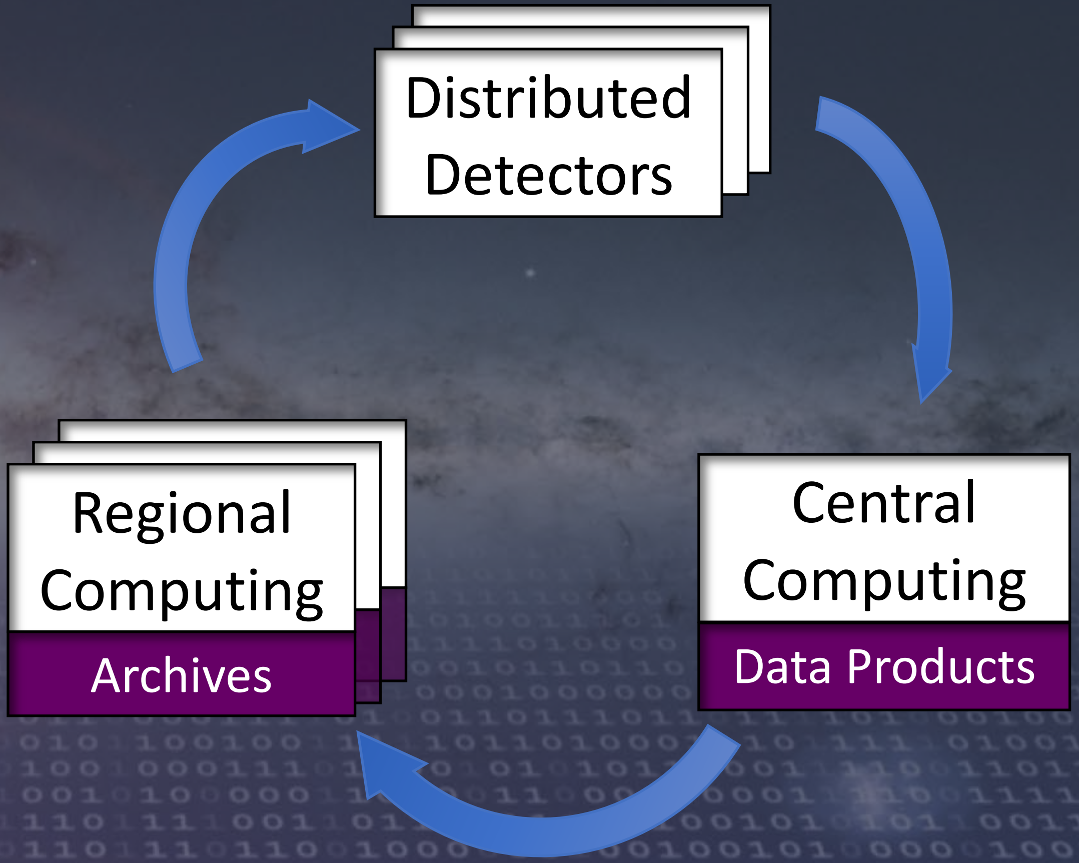
Challenge: Data Irreversibility

- Resolution (time/frequency/spatial) of detectors increases steadily
  - ⟹ Exponential grow of data volumes
  - ⟹ Only small extracts of raw data can be archived
  - ⟹ Data reduction by orders of magnitude

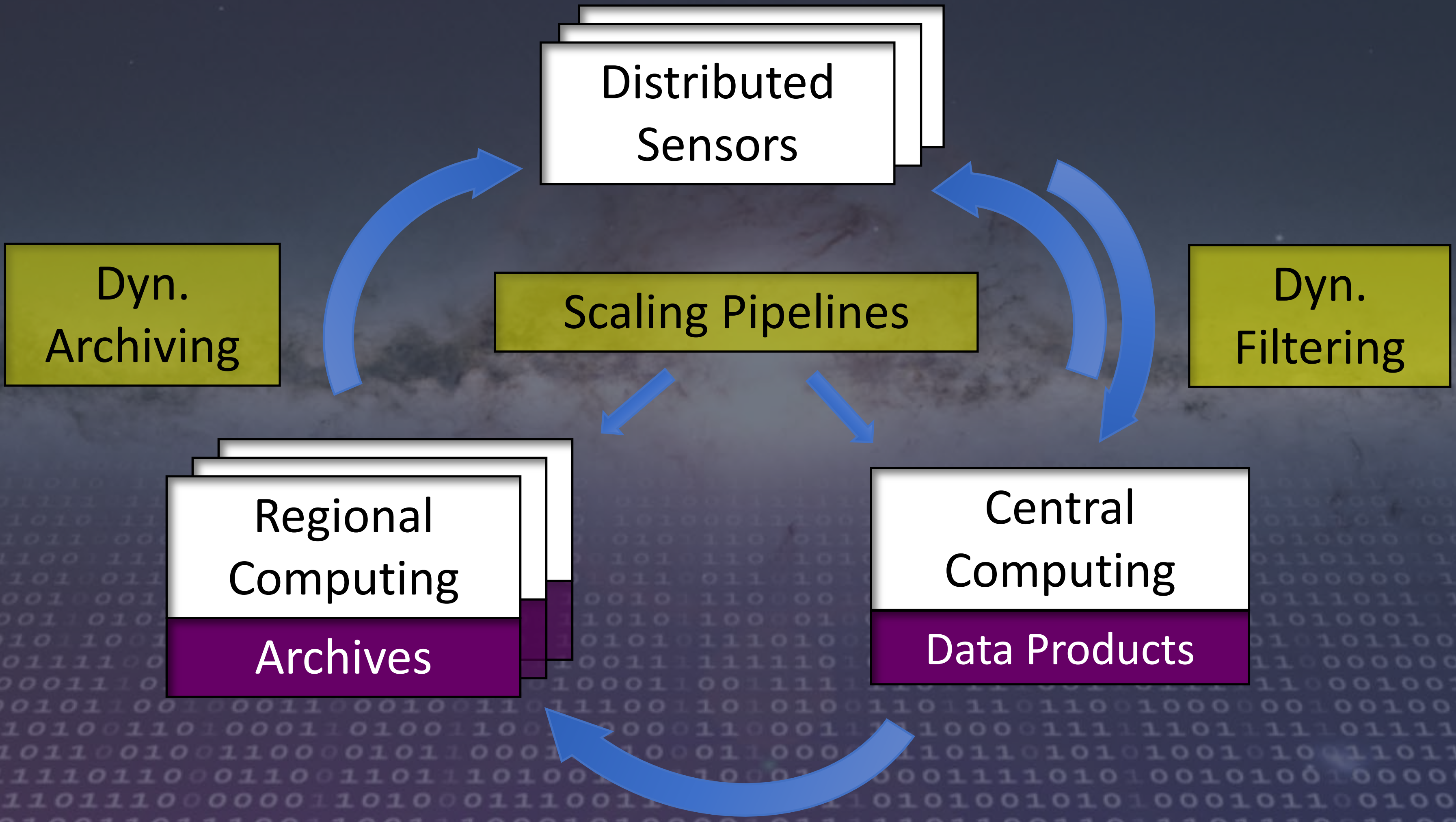
- Data irreversibility
  - Only “derived data products” are stored in archives
  - Data reduction ⟹ irreversible loss of information
  - Fundamental problem: reproducibility of results!

Challenge: Data Irreversibility

- Reduction of raw data streams by orders of magnitude
    - Cherry picking in real-time
    - Irreversible loss of information
- Data products transported to archives (= a few ”**data lakes**”)



Challenge: Data Irreversibility



Astro-NFDI: Consortium Workpackages

- Governance, Consortium Management
- Distributed Services and Structures
- Data Workflows
- Software for Data
- Data Irreversibility Challenges
- Synergies / Interaction with other consortia
- Training, Summerschools etc., (Education)

“The future is already here. It’s just not very evenly distributed”

William Gibson