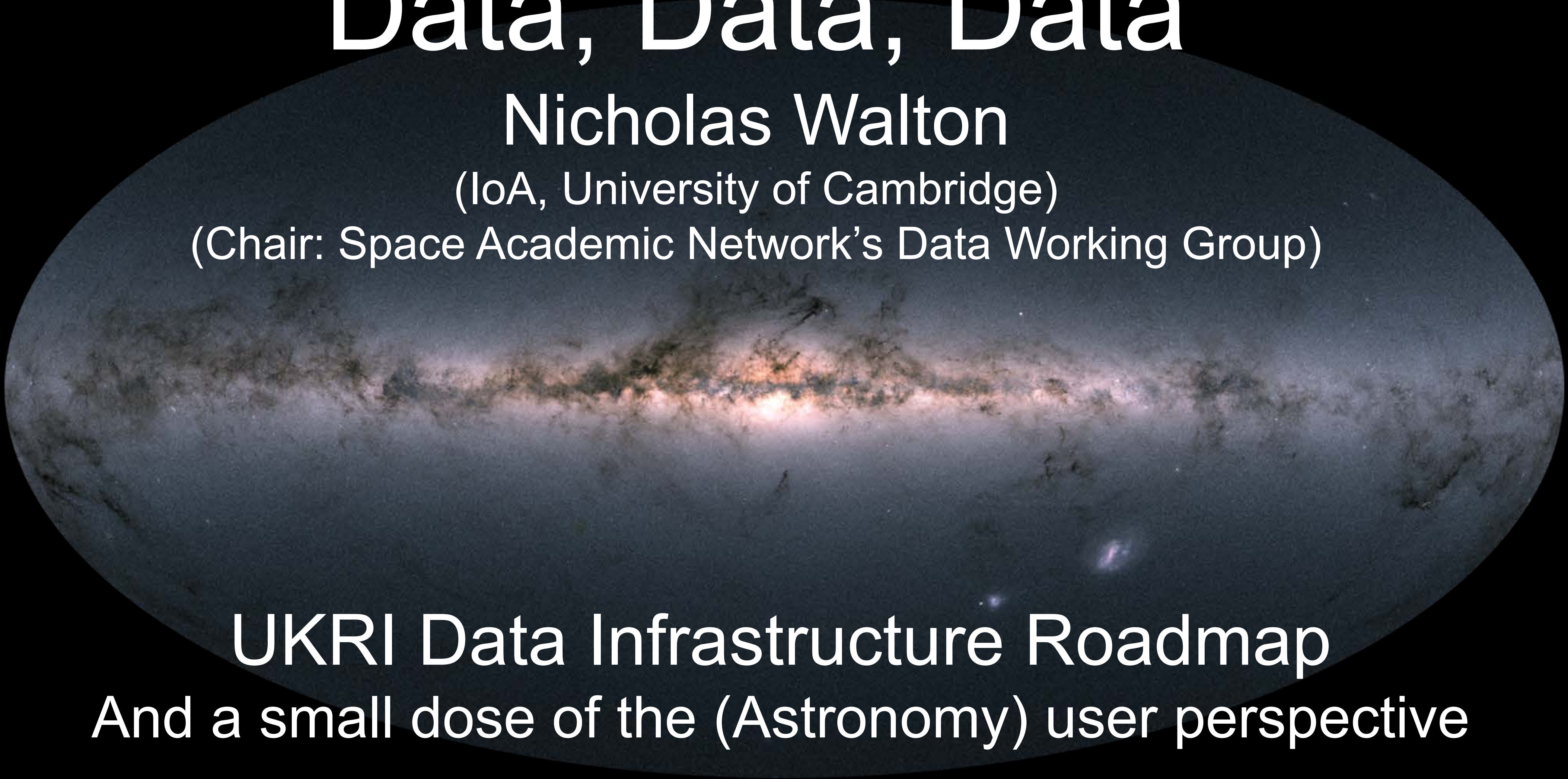
Data, Data, Data

Nicholas Walton
(IoA, University of Cambridge)
(Chair: Space Academic Network’s Data Working Group)



UKRI Data Infrastructure Roadmap

And a small dose of the (Astronomy) user perspective

UKRI Data Infrastructure Roadmap

White Paper led by Jeremy Yates + RC experts

The UKRI Data Infrastructure Roadmap White Paper is currently in an (advanced) draft format: some aspects may still be revised

> UKRI urgently needs to restore the foundations upon which such exploitation of data can happen. [..] put in place the physical compute and storage capacity needed to host and exploit the data across UKRI. Without this all other discussions are moot.

19 Key Recommendations covering: Research Data Infrastructure: Research Data Exploitation and sharing; International Collaboration and Leadership; People and skills

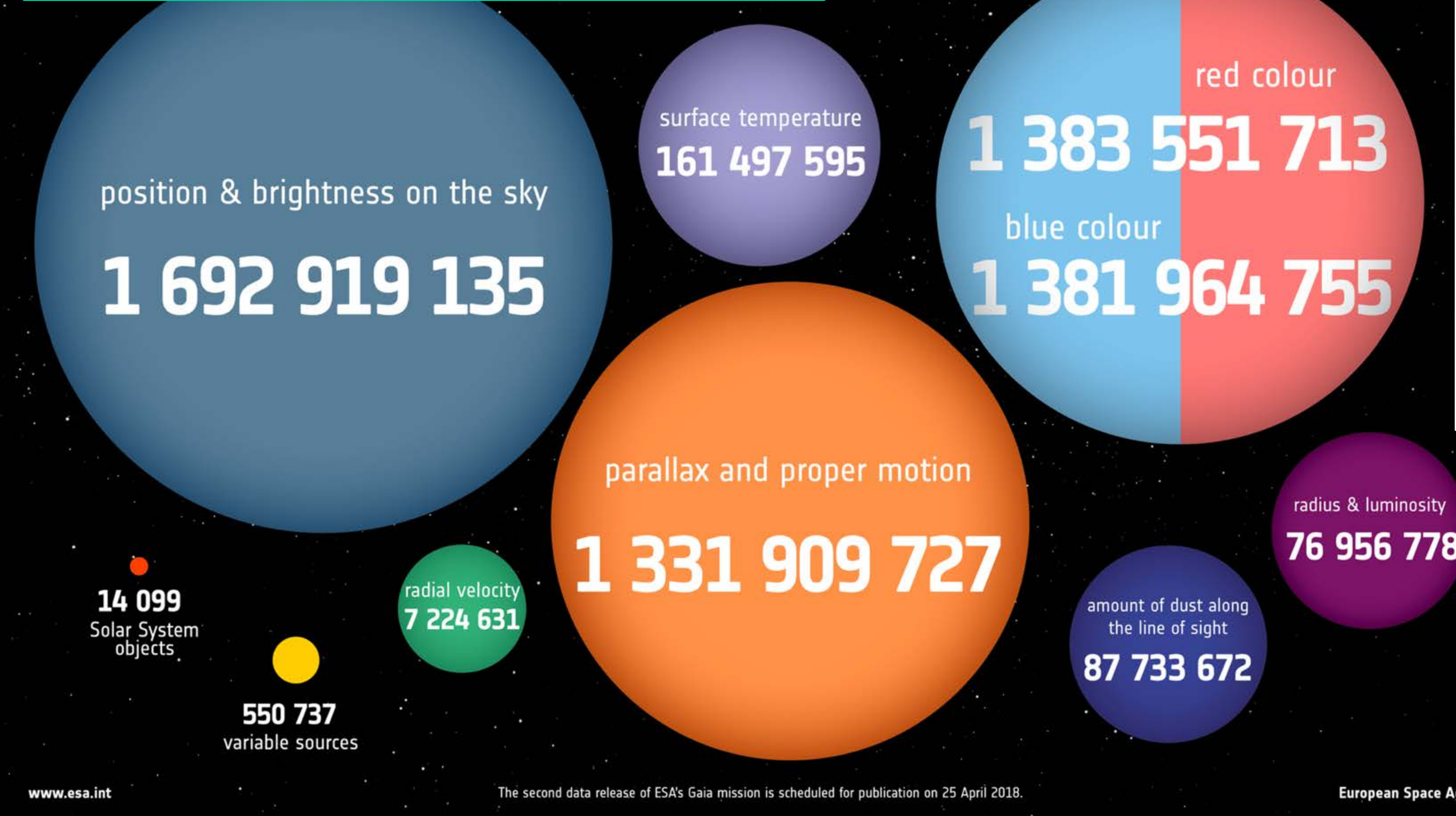
Near term priorities for 2020-2022:

Maintaining and operating both existing and imminently deployed infrastructures; Transformation to new capabilities, Establishing coordination activities at the UKRI, RDI and International level

UKRI investment of £200M-£300M per annum is necessary to deliver this transformation and the required level of infrastructure and services.

Gaia: an example Big Data Challenge

Gaia DR2: with larger to come



The second data release of ESA's Gaia mission is scheduled for publication on 25 April 2018.

Gaia
1 Trillion observations reached on 14 April 2018 … and counting …

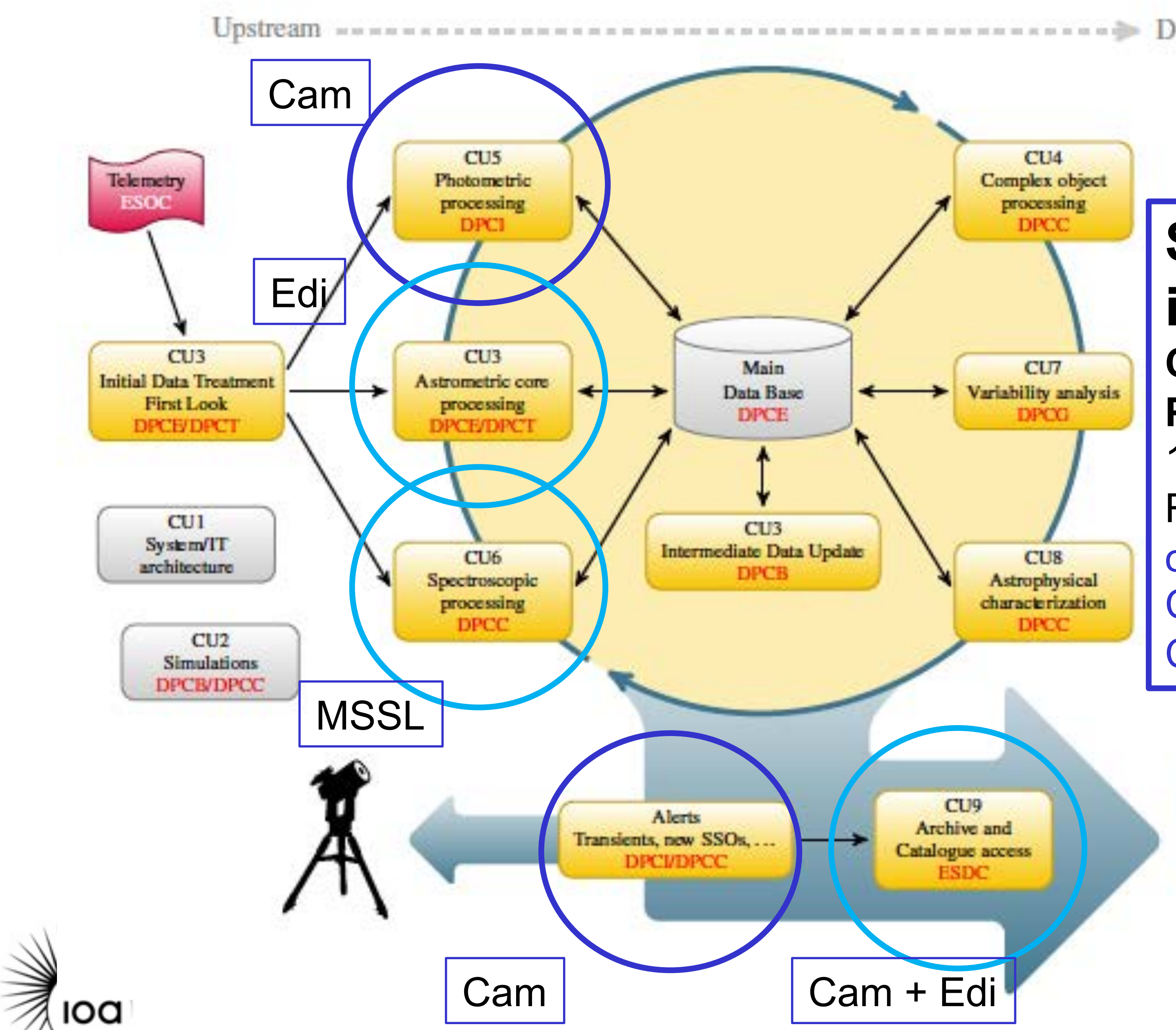
2 Billion sources / 1 Billion images/day/ 5 million spectra/day / **main database 1PB**

gaia

Gaia Data Processing

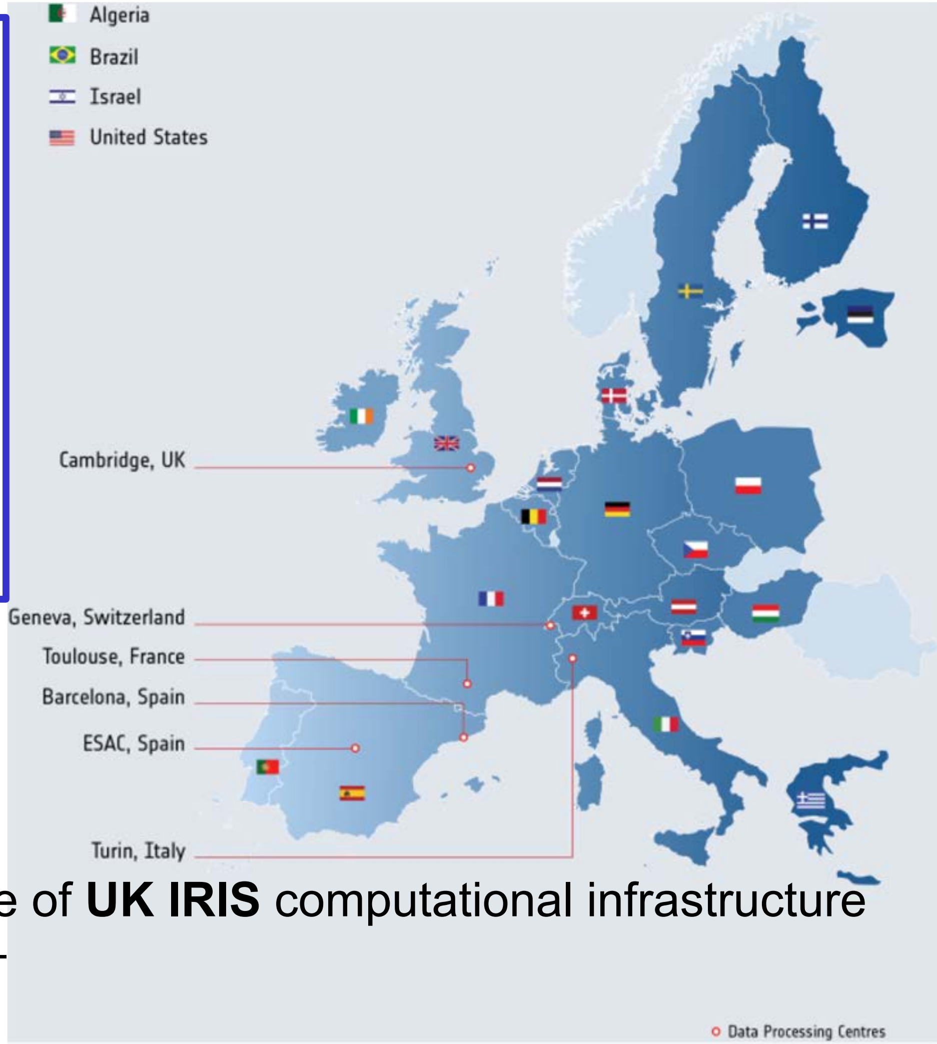
Typical modern big data science

a pan European effort: ~450 specialists from 24 countries



Significant UK involvement:
Lead Photometric processing and Flux alerts
Contribute to pre-processing, RVS, archive, beta testing

Significant UK involvement:
Cambridge Data Processing Centre
10500 CPUs, 55TB RAM, 2.3PB disk
operated at West Cambridge Data Centre



use of **UK IRIS** computational infrastructure

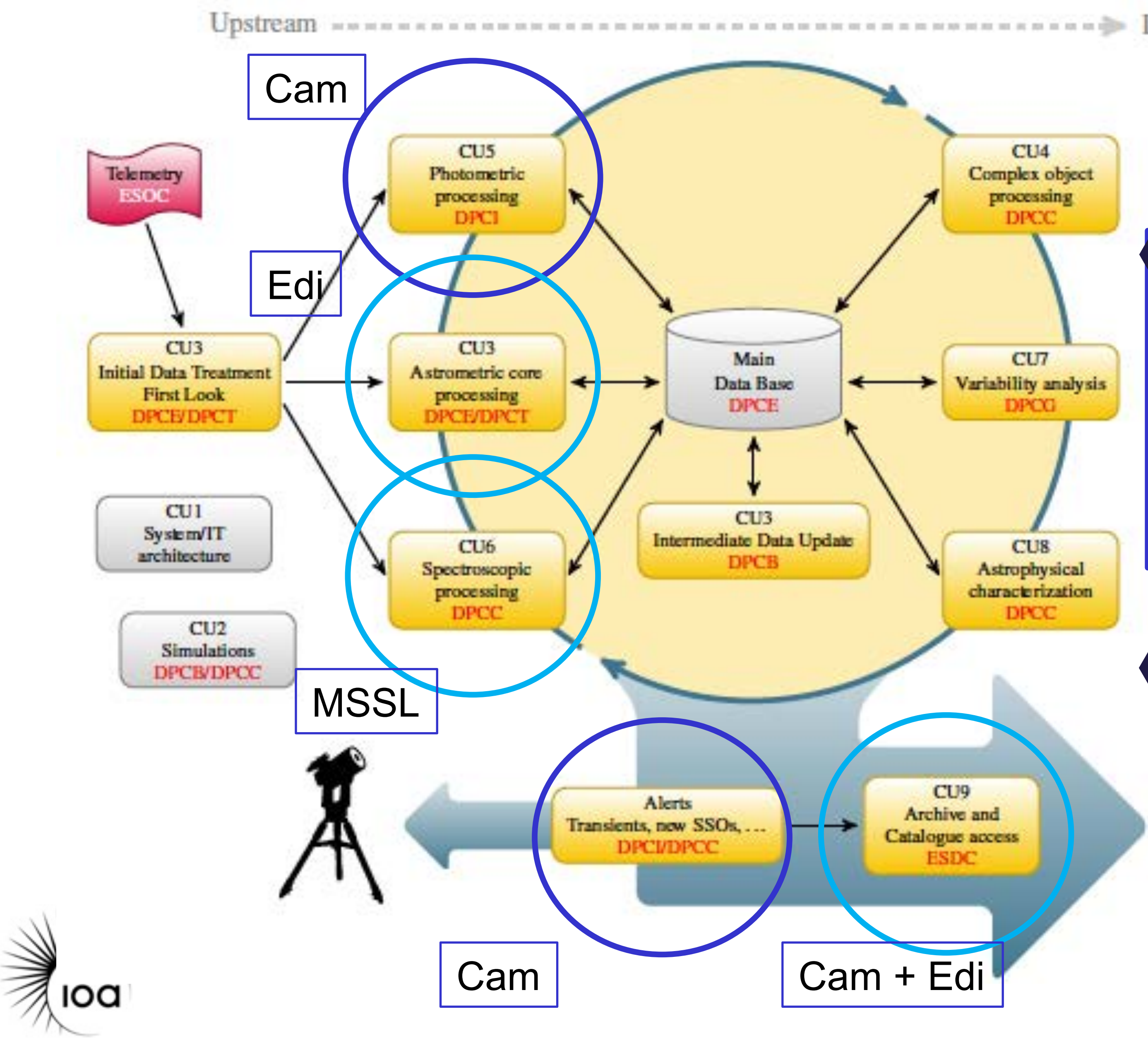
mputing -

ioa

Gaia Data Processing

A good example of a research and innovation infrastructure

a pan European effort: ~450 specialists fr… …ountries



Significa… …t:

…ton… …Fl…

…to p… …S, … …ting

i…

Ca…

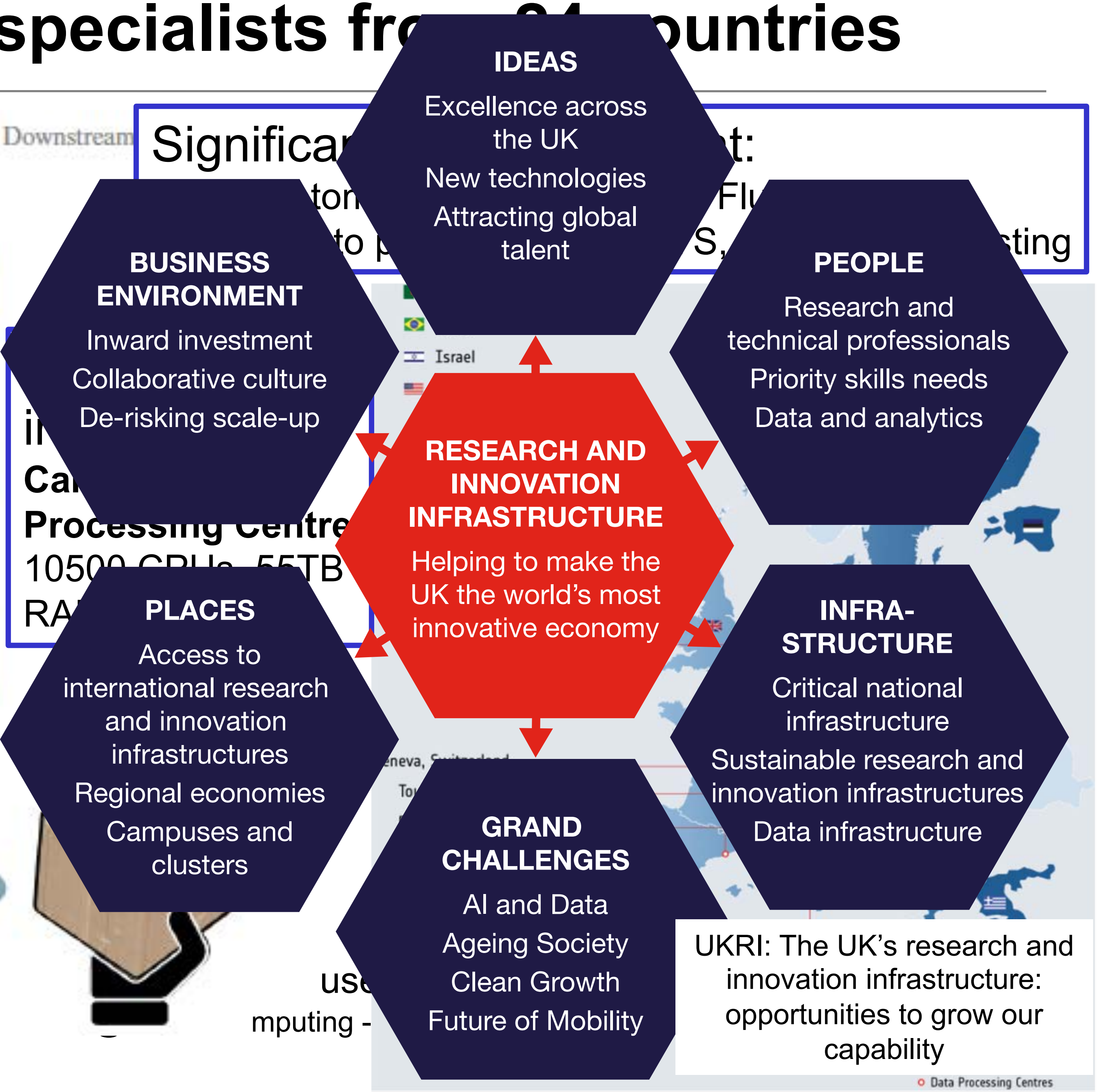
Processing Centre

10500 CPUs, 55TB

RA…

us…

…mputing -…



IDEAS

Excellence across the UK

New technologies

Attracting global talent

BUSINESS ENVIRONMENT

Inward investment

Collaborative culture

De-risking scale-up

PEOPLE

Research and technical professionals

Priority skills needs

Data and analytics

RESEARCH AND INNOVATION INFRASTRUCTURE

Helping to make the UK the world's most innovative economy

PLACES

Access to international research and innovation infrastructures

Regional economies

Campuses and clusters

INFRA-STRUCTURE

Critical national infrastructure

Sustainable research and innovation infrastructures

Data infrastructure

GRAND CHALLENGES

AI and Data

Ageing Society

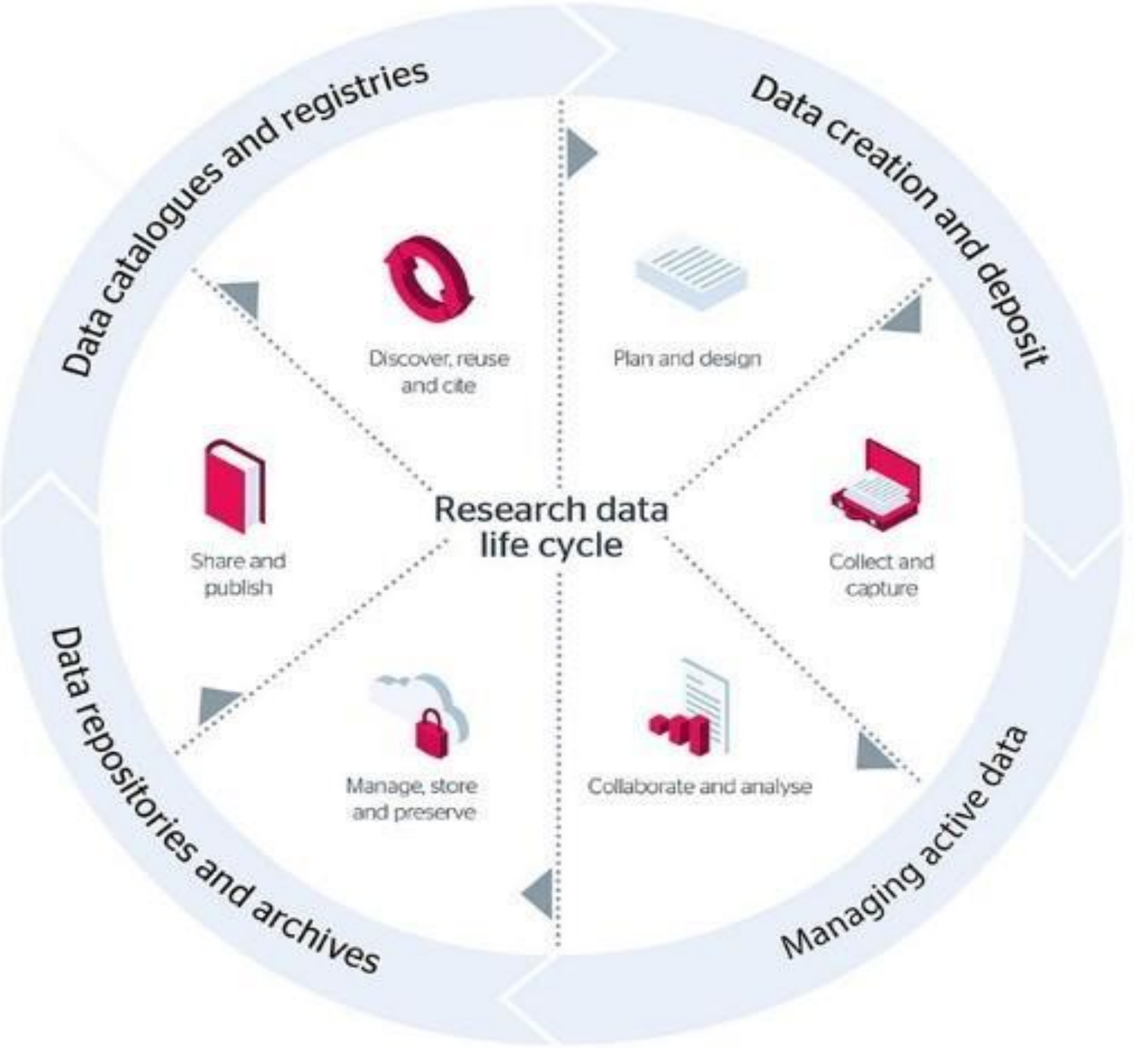
Clean Growth

Future of Mobility

UKRI: The UK's research and innovation infrastructure: opportunities to grow our capability

The Vision for a UKRI Research Data Infrastructure and Services ecosystem

Create a thriving, strategically coordinated, and federated UKRI Research Data Infrastructure (RDI) ecosystem, which will be an essential cross-cutting theme of the UKRI Research Infrastructure Roadmap. Components of the RDI ecosystem from particular research disciplines will be interoperable. Only in this way will the benefits and impacts of UK's rich data resources will be maximised, and the effectiveness of funding will be assured through appropriate coordination, consolidation and co-location.

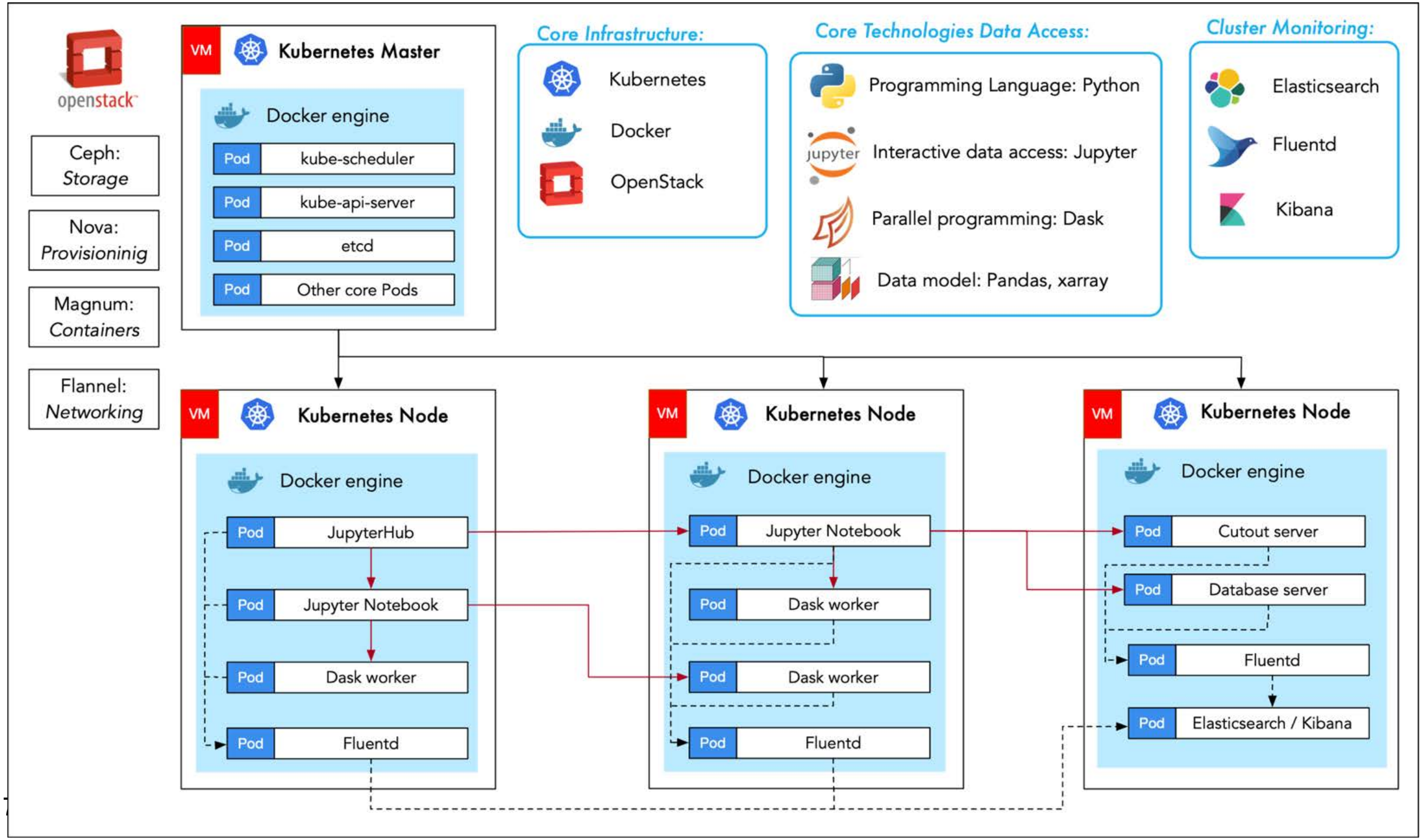


The Research Data Life Cycle in terms of Policy Requirements and Outcomes.

Science User Data Access and Analysis

Astronomy Deployment Example with IRIS@Cambridge

Cambridge Service for Data-Driven Discovery (CSD3)



openstack

Ceph: *Storage*

Nova: *Provisioninig*

Magnum: *Containers*

Flannel: *Networking*

VM Kubernetes Master
- Docker engine
- Pod kube-scheduler
- Pod kube-api-server
- Pod etcd
- Pod Other core Pods

Core Infrastructure:
- Kubernetes
- Docker
- OpenStack

Core Technologies Data Access:
- Programming Language: Python
- Interactive data access: Jupyter
- Parallel programming: Dask
- Data model: Pandas, xarray

Cluster Monitoring:
- Elasticsearch
- Fluentd
- Kibana

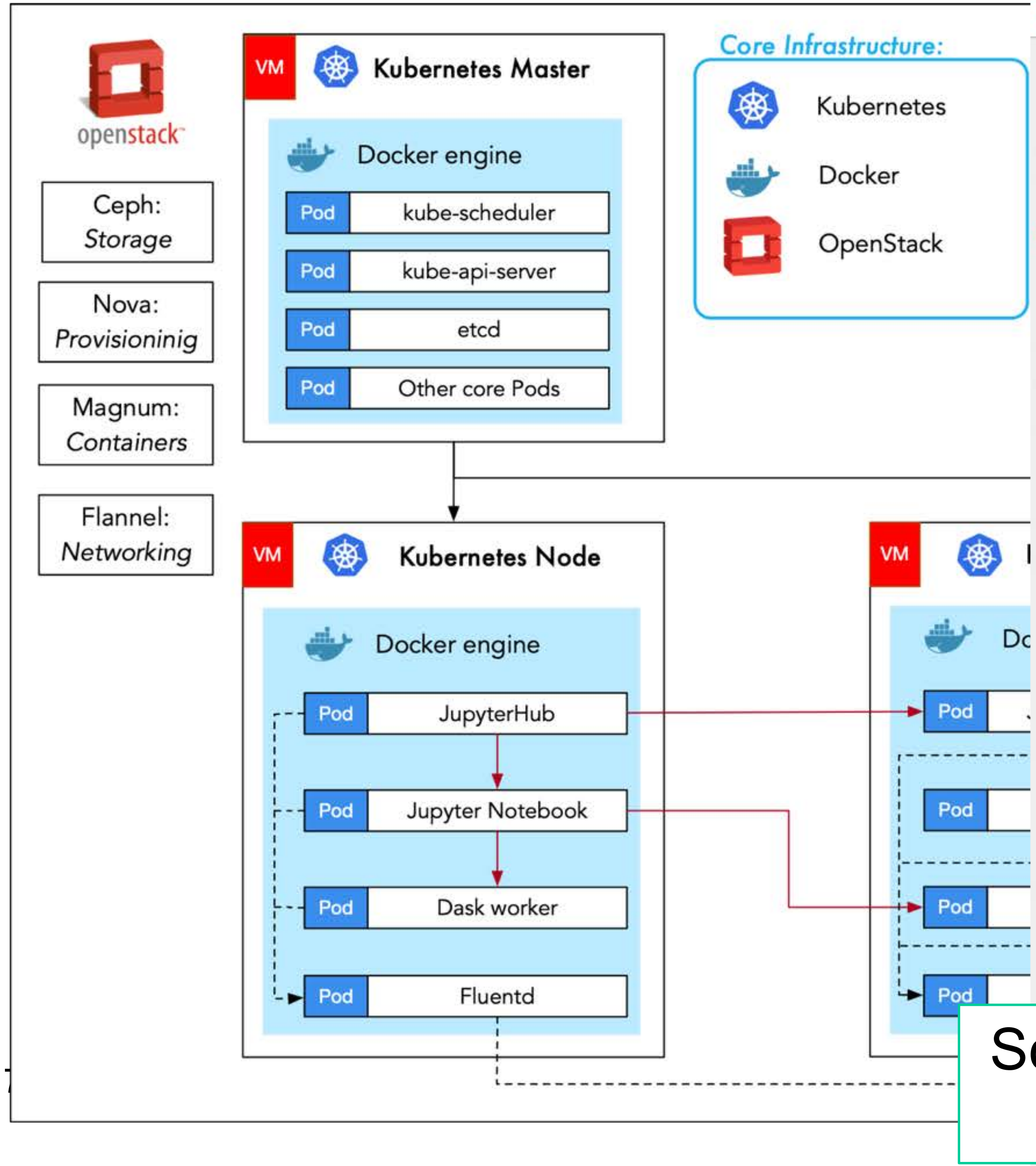
VM Kubernetes Node
- Docker engine
- Pod JupyterHub
- Pod Jupyter Notebook
- Pod Dask worker
- Pod Fluentd

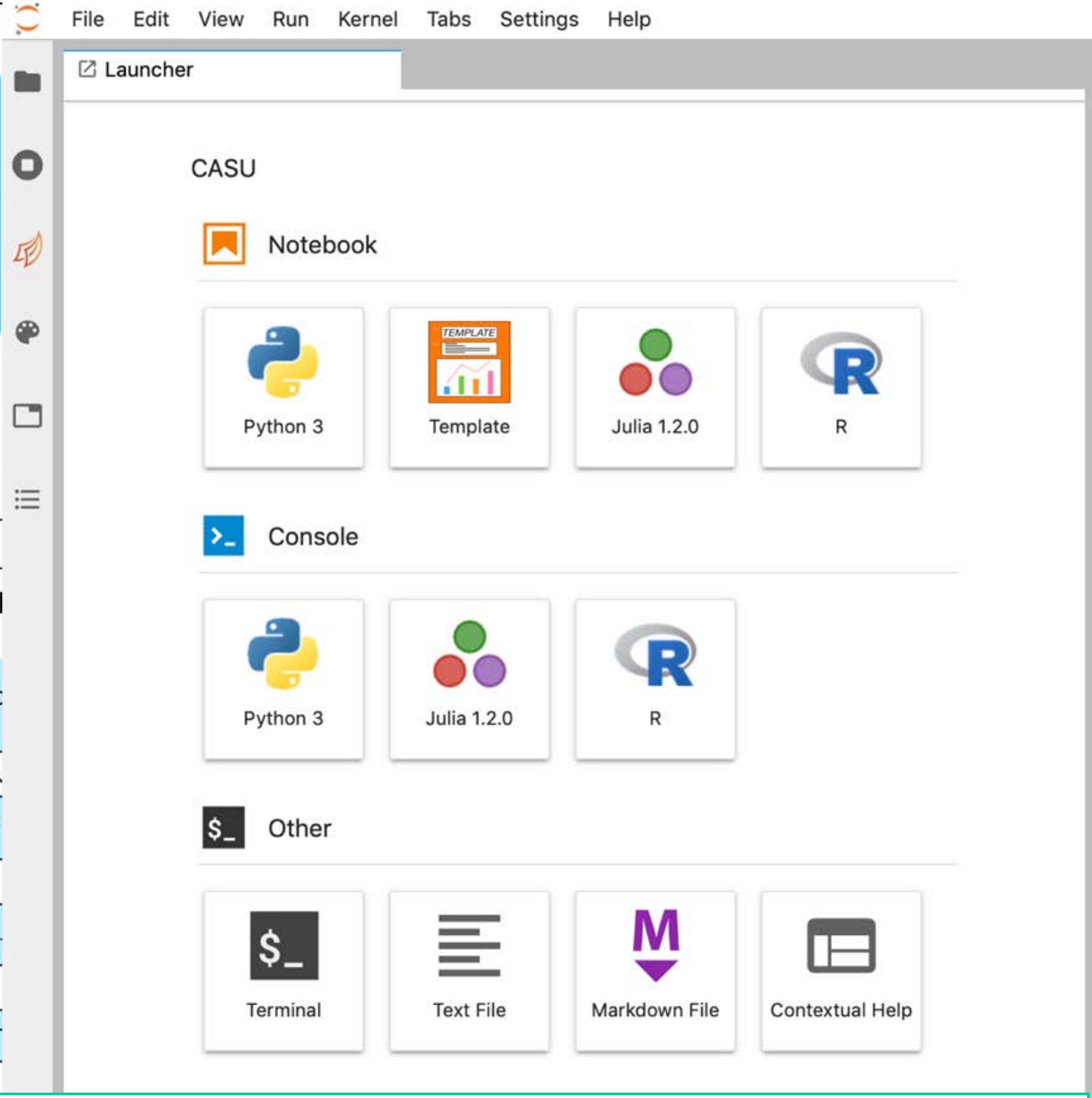
VM Kubernetes Node
- Docker engine
- Pod Jupyter Notebook
- Pod Dask worker
- Pod Dask worker
- Pod Fluentd

VM Kubernetes Node
- Docker engine
- Pod Cutout server
- Pod Database server
- Pod Fluentd
- Pod Elasticsearch / Kibana

Science User Data Access and Analysis

Astronomy Deployment Example with IRIS@Cambridge

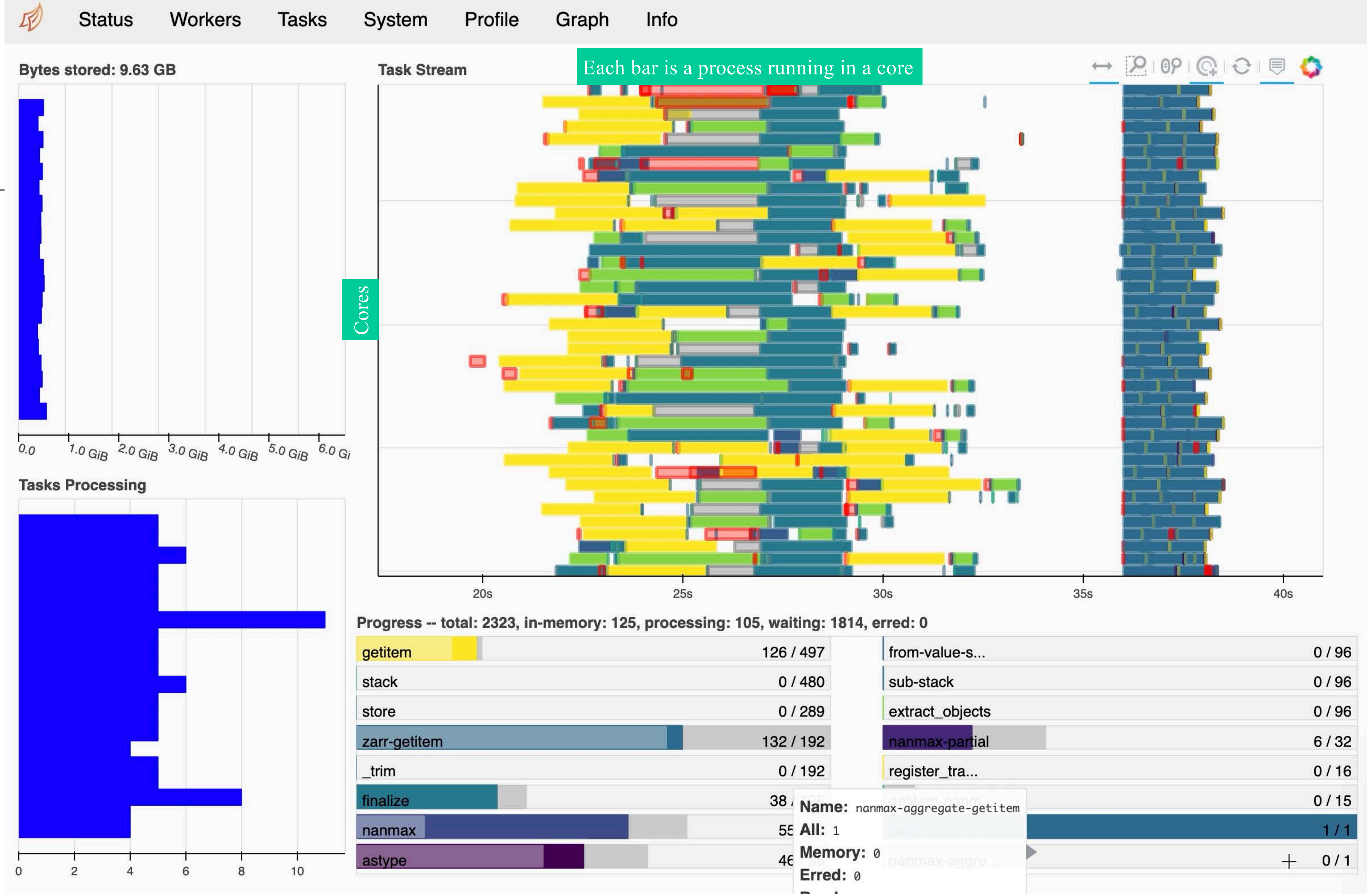




Science user interface provides access to code, data, visualisation, sharing

Easy user access to scalable underlying resources.

Algorithms at the data at the servers



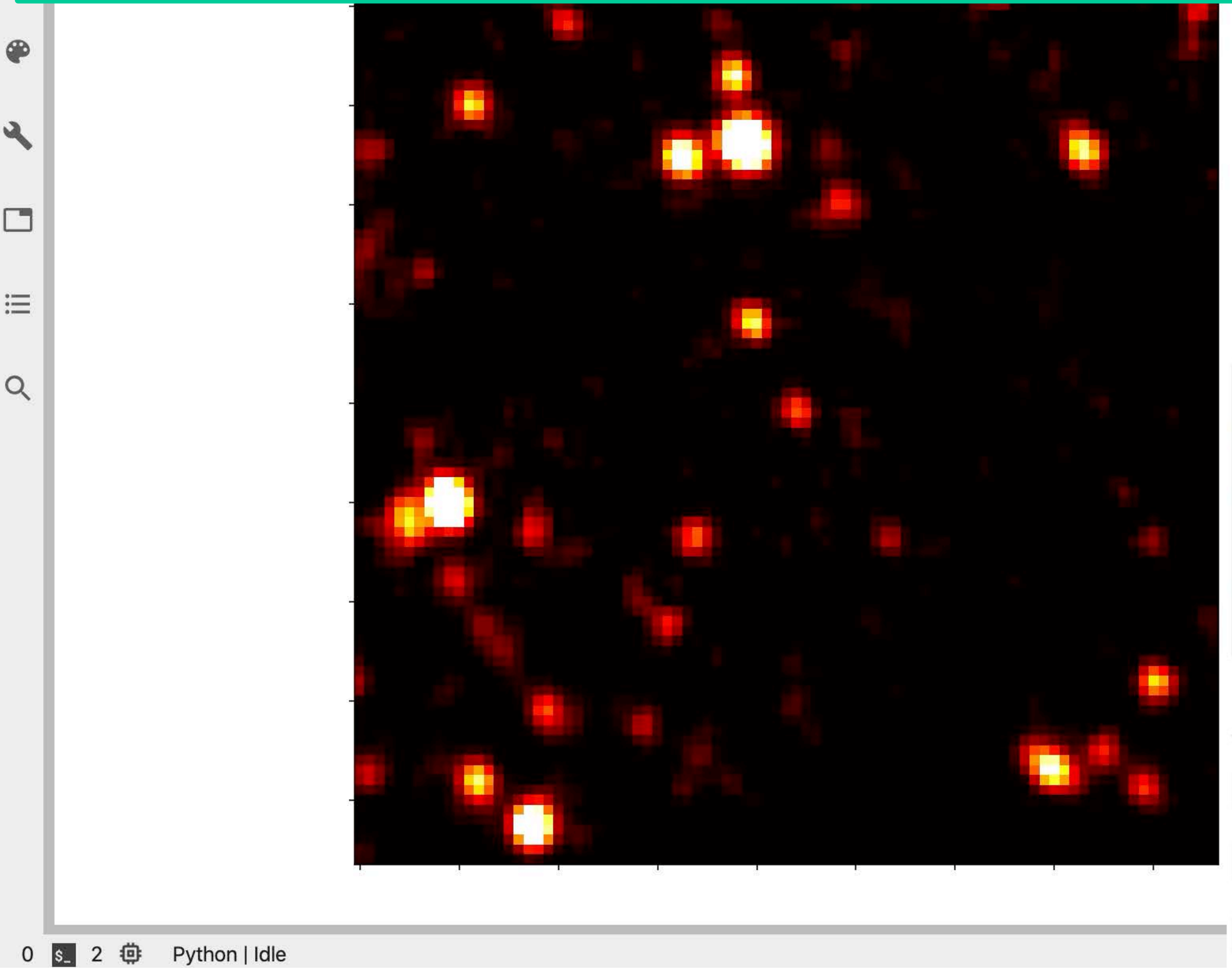
Each bar is a process running in a core

Cores

```
Progress -- total: 2323, in-memory: 125, processing: 105, waiting: 1814, erred: 0
getitem          126 / 497        from-value-s...     0 / 96
stack              0 / 480        sub-stack           0 / 96
store              0 / 289        extract_objects     0 / 96
zarr-getitem     132 / 192        nanmax-partial      6 / 32
_trim              0 / 192        register_tra...     0 / 16
finalize          38 /                                0 / 15
nanmax            55                                  1 / 1
astype            46                                  0 / 1
```

```
Name: nanmax-aggregate-getitem
All: 1
Memory: 0
Erred: 0
```

Dynamic user access to entire VISTA pixel & catalogue data set / direct user access to the processing pipeline outputs



No: **1**

Filename v20100228_00495_st_tl.fit
Filter Ks
Survey VVV
Exp Time 8.0

Cambridge Astronomy Survey Unit

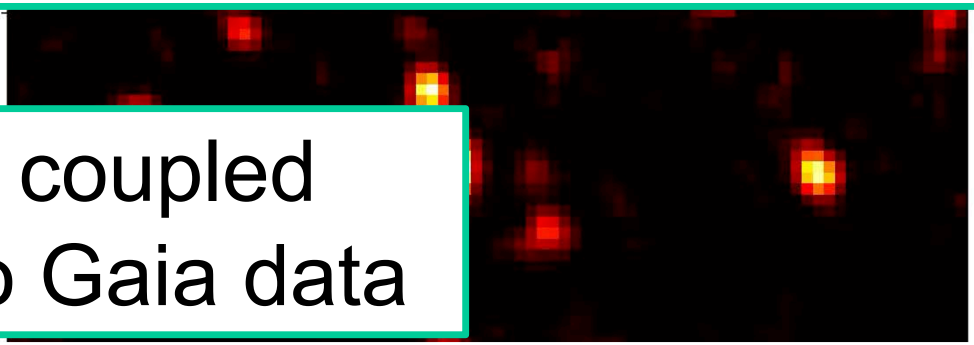
Query VISTA database around position

```
vistadb = VISTADB()

# Coordinates to search for
ra, dec = 194.30, -64.75

# Columns to print
columns = ['filename','coords','filtername', 'surveyname', 'nigh
          'totexptime', 'obsfwhm', 'obstatus', 'qcstatus']

# Execute query and dissplay first 10 results
res = vistadb.query_radec(ra, dec)
res[columns].head(n=10)
```

	filename	coords	filtername	surveyna					
id									
62124	v20100218_00330_st_tl.fit	13:02:18.02 -64:35:28.7	Ks	VVV	20100218	8.0	0.8	Completed	A
68702	v20100228_00495_st_tl.fit	13:02:18.02 -64:35:28.7	Ks	VVV	20100228	8.0	0.8	Completed	A
70725	v20100304_00444_st_tl.fit	13:02:18.02 -64:35:28.7	Ks	VVV	20100304	8.0	0.8	Completed	A
81752	v20100315_00327_st_tl.fit	13:02:18.02 -64:35:28.7	Ks	VVV	20100315	8.0	0.8	Completed	A
83062	v20100316_00361_st_tl.fit	13:02:18.02 -64:35:28.7	Ks	VVV	20100316	8.0	0.8	Completed	A
83235	v20100316_00499_st_tl.fit	13:02:18.02 -64:35:28.7	H	VVV	20100316	40.0	0.8	Completed	A
83254	v20100316_00511_st_tl.fit	13:02:18.22 -64:35:29.9	Ks	VVV	20100316	40.0	0.8	Completed	A
83273	v20100316_00523_st_tl.fit	13:02:18.22 -64:35:29.9	J	VVV	20100316	40.0	0.8	Completed	A
120735	v20100422_00425_st_tl.fit	13:02:18.02 -64:35:28.7	Y	VVV	20100422	40.0	1.0	Completed	A
120754	v20100422_00437_st_tl.fit	13:02:18.22 -64:35:29.9	Z	VVV	20100422	40.0	1.0	Completed	A

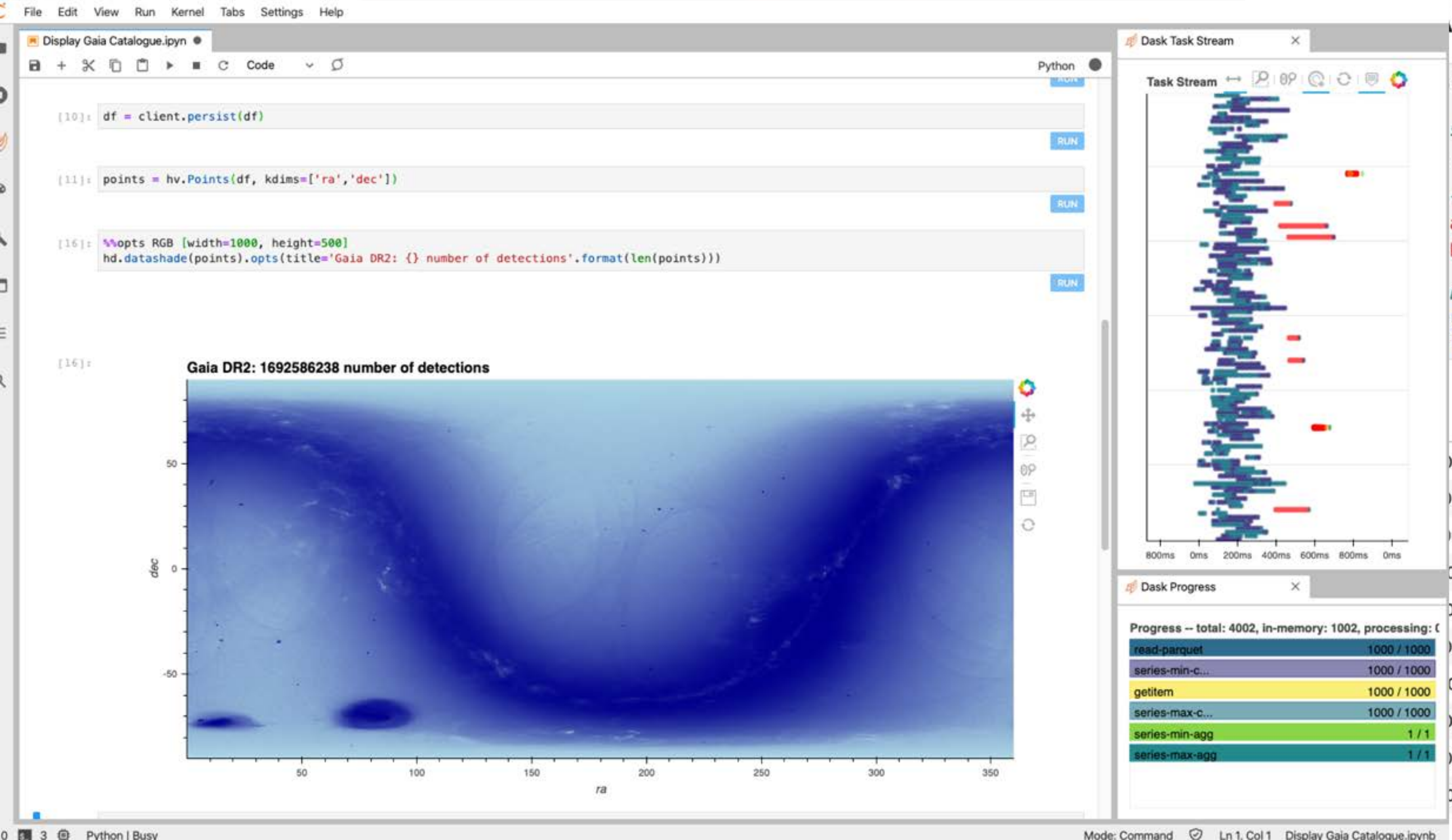
Dynamic user access to entire VISTA pixel & catalogue data set / direct user access to the processing pipeline outputs

Tightly coupled access to Gaia data



No: 1

Filename v20100228_00495_st_tl.fit
Filter Ks
Survey VVV
Exp Time 8.0



```
[10]: df = client.persist(df)

[11]: points = hv.Points(df, kdims=['ra','dec'])

[16]: %%opts RGB [width=1000, height=500]
      hd.datashade(points).opts(title='Gaia DR2: {} number of detections'.format(len(points)))

[16]:
```

Gaia DR2: 1692586238 number of detections

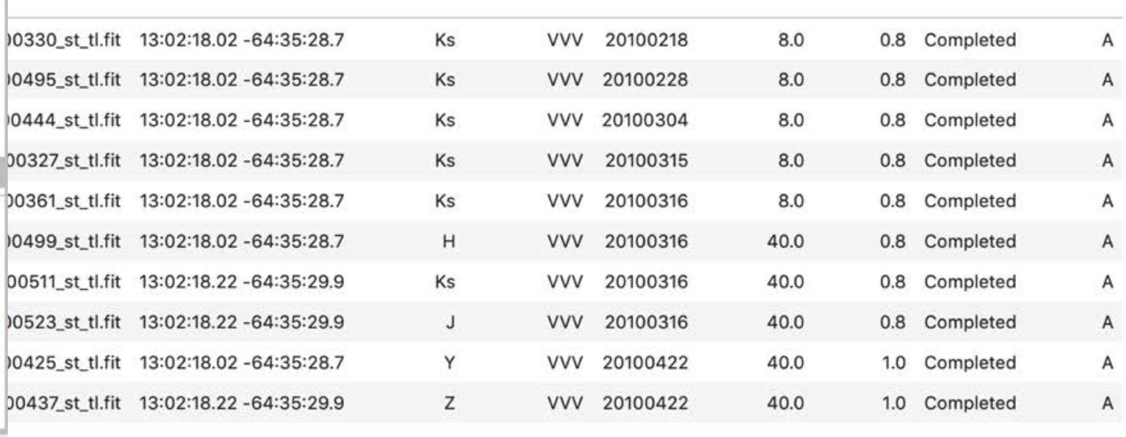
Progress -- total: 4002, in-memory: 1002, processing: (

```
read-parquet      1000 / 1000
series-min-c...   1000 / 1000
getitem           1000 / 1000
series-max-c...   1000 / 1000
series-min-agg    1 / 1
series-max-agg    1 / 1
```

Mode: Command   Ln 1, Col 1   Display Gaia Catalogue.ipynb

database around position

```
()

search for
 -64.75

t
ame','coords','filtername', 'surveyname', 'nigh
ptime', 'obsfwhm', 'obstatus', 'qcstatus']

nd dissplay first 10 results
ry_radec(ra, dec)
(n=10)
```



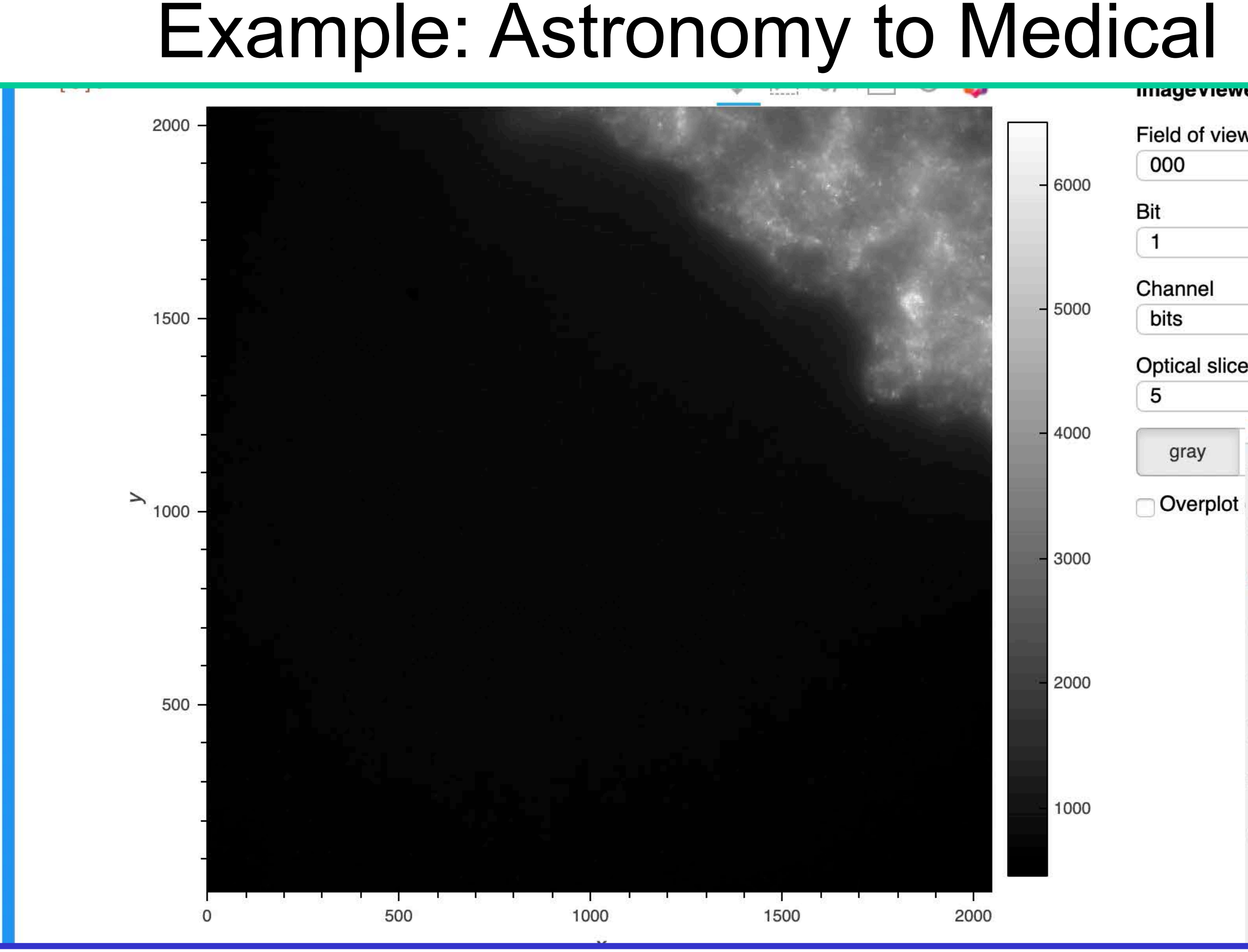
filename	coords	filtername	surveyna					
)0330_st_tl.fit	13:02:18.02 -64:35:28.7	Ks	VVV	20100218	8.0	0.8	Completed	A
)0495_st_tl.fit	13:02:18.02 -64:35:28.7	Ks	VVV	20100228	8.0	0.8	Completed	A
)0444_st_tl.fit	13:02:18.02 -64:35:28.7	Ks	VVV	20100304	8.0	0.8	Completed	A
)0327_st_tl.fit	13:02:18.02 -64:35:28.7	Ks	VVV	20100315	8.0	0.8	Completed	A
)0361_st_tl.fit	13:02:18.02 -64:35:28.7	Ks	VVV	20100316	8.0	0.8	Completed	A
)0499_st_tl.fit	13:02:18.02 -64:35:28.7	H	VVV	20100316	40.0	0.8	Completed	A
)0511_st_tl.fit	13:02:18.22 -64:35:29.9	Ks	VVV	20100316	40.0	0.8	Completed	A
)0523_st_tl.fit	13:02:18.22 -64:35:29.9	J	VVV	20100316	40.0	0.8	Completed	A
)0425_st_tl.fit	13:02:18.02 -64:35:28.7	Y	VVV	20100422	40.0	1.0	Completed	A
)0437_st_tl.fit	13:02:18.22 -64:35:29.9	Z	VVV	20100422	40.0	1.0	Completed	A

Cambridge Astronomy Survey Unit

Effectively exploiting common standards, common infrastructure, common analysis: Example: Astronomy to Medical

CANCER RESEARCH UK GRAND CHALLENGE

cruk.org/grandchallenge



ImageViewer

Field of view: 000

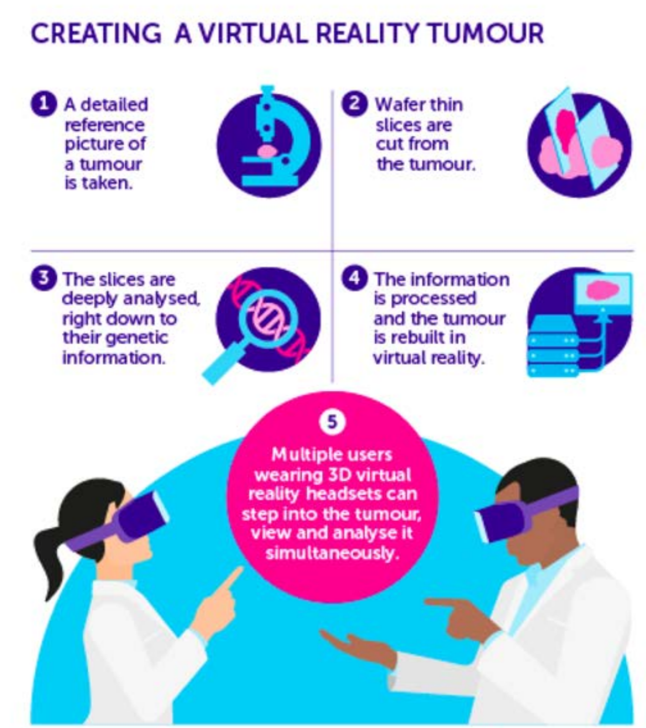
Bit: 1

Channel: bits

Optical slice: 5

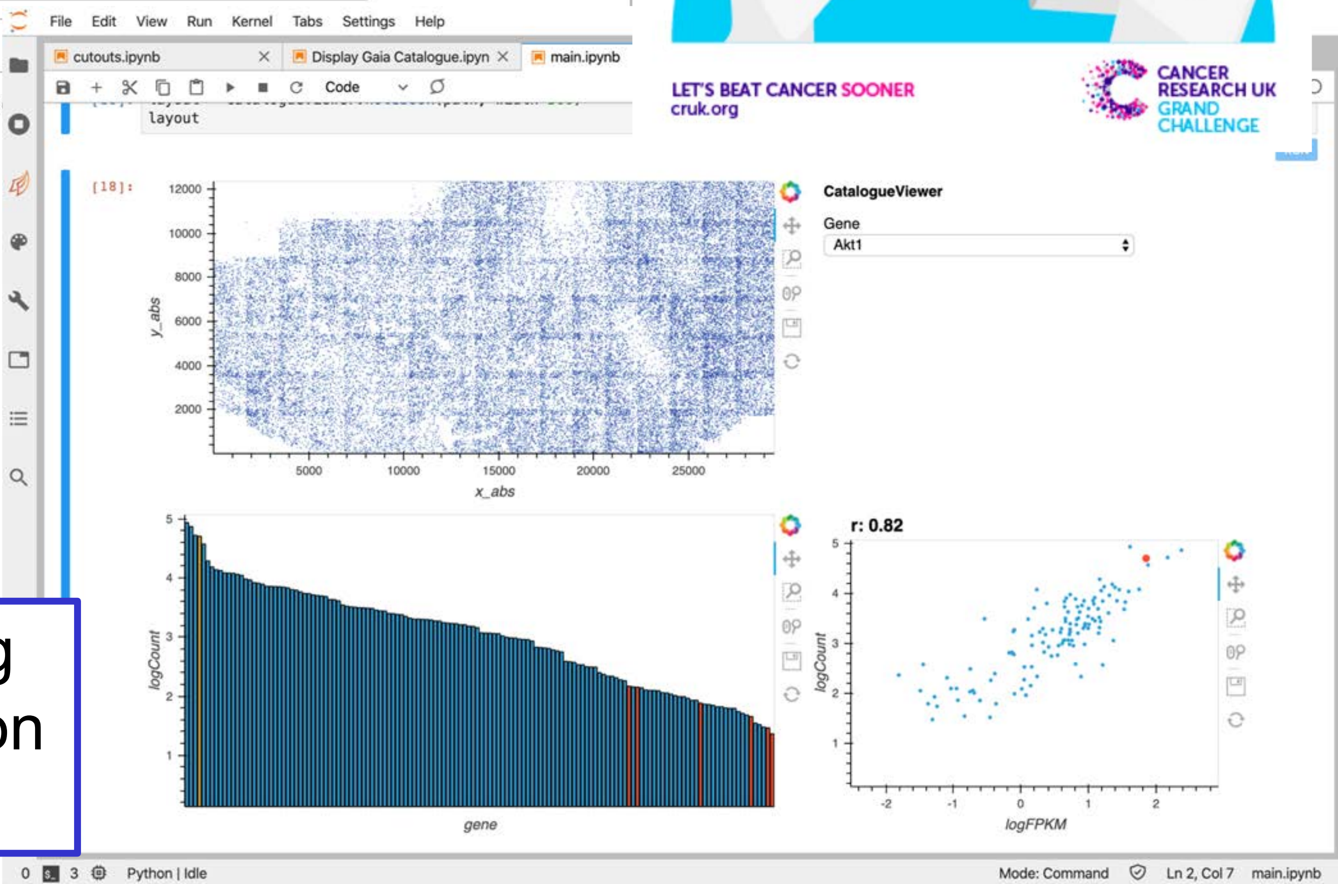
gray

□ Overplot



CREATING A VIRTUAL REALITY TUMOUR

1. A detailed reference picture of a tumour is taken.
2. Wafer thin slices are cut from the tumour.
3. The slices are deeply analysed, right down to their genetic information.
4. The information is processed and the tumour is rebuilt in virtual reality.
5. Multiple users wearing 3D virtual reality headsets can step into the tumour, view and analyse it simultaneously.



LET'S BEAT CANCER SOONER
cruk.org

CANCER RESEARCH UK GRAND CHALLENGE

cutouts.ipynb | Display Gaia Catalogue.ipyn | main.ipynb

```
layout
```

[18]:

CatalogueViewer

Gene: Akt1

r: 0.82

0 3 Python | Idle   Mode: Command   Ln 2, Col 7   main.ipynb

Commonality of approach. Large medical imaging data analysis pipelines deployed at CASU based on astronomy (VISTA/ Gaia) system

Research Data Infrastructure (RDI)

Investment is urgently needed now, in the period 2020-22, to put the UK on a world class footing in respect of physical infrastructure and software infrastructure, reversing the significant gap that has arisen over the last few years.

Research Data Exploitation and sharing

Each Sector should refine and update its RDI requirements, in terms of its own Research Data Life Cycle, such that data are supported at each stage in the life cycle and can be readily analysed, discovered, combined, reused and repurposed.

International Collaboration and Leadership

Coordination structures are needed commensurate with the fact that the creation and use of Resources for research data are increasingly an international activity, with major subject-specific repositories having a global reach.

People and Skills

Investment is needed in people needed to create, engineer and apply the advanced computing techniques to the data to extract knowledge and innovate.

People: jobs & career path

Research Use Case: Square Kilometer Array
white paper informed by examples e.g. from STFC/UKRI domain

The SKA project is an international effort to build the world's largest radio telescope in order to image huge areas of the sky on a scale and with a level of sensitivity no survey telescope has ever achieved before. To enable the science fully, there are major data and data infrastructure issues to be addressed:

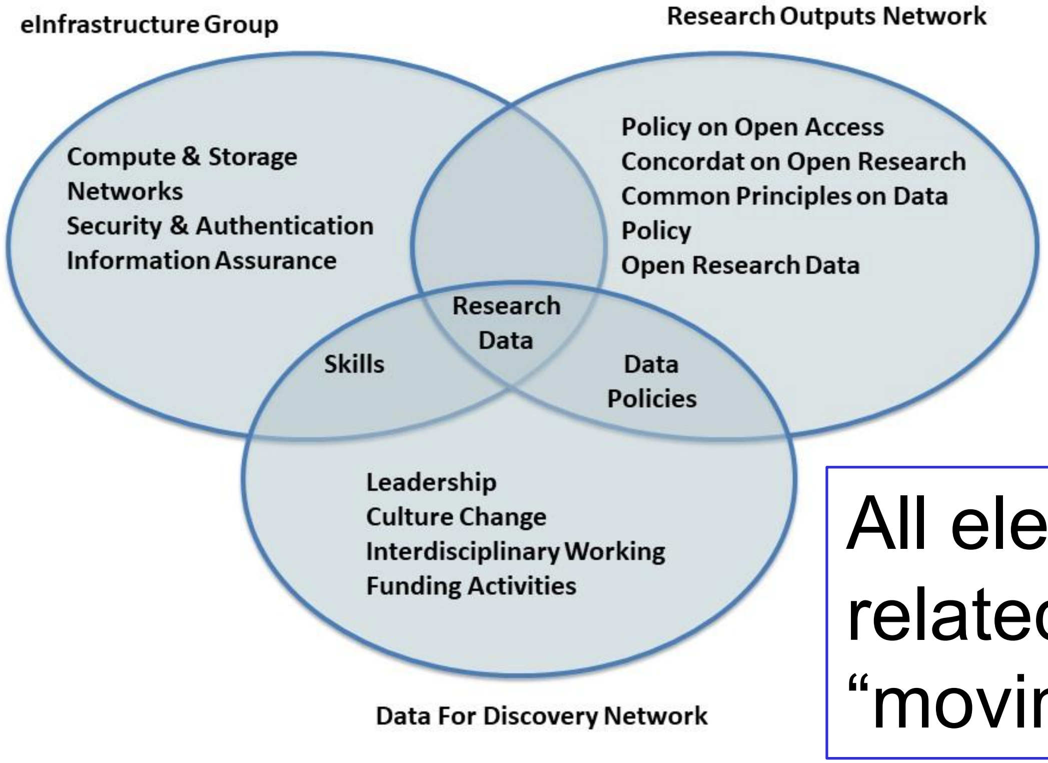
- ~1 PB/day into the science archive → significant data volumes
- Archive: search ability on the individual data products/ meta-data + curation
- User authentication & authorisation must be enforced → data rights
- Multiple secondary data products derived from the primary data → storage implications
- Analysis of data products →large number of astrophysical sources
- Range of analysis algorithms run on the data → compute needs
- Individual SKA image data products are so large (250TB on average) → move the algorithms to the data
- Interoperability of SKA data with other astronomy data

Functional Requirements for Federated RDIs

Physical & Stewardship Infrastructures

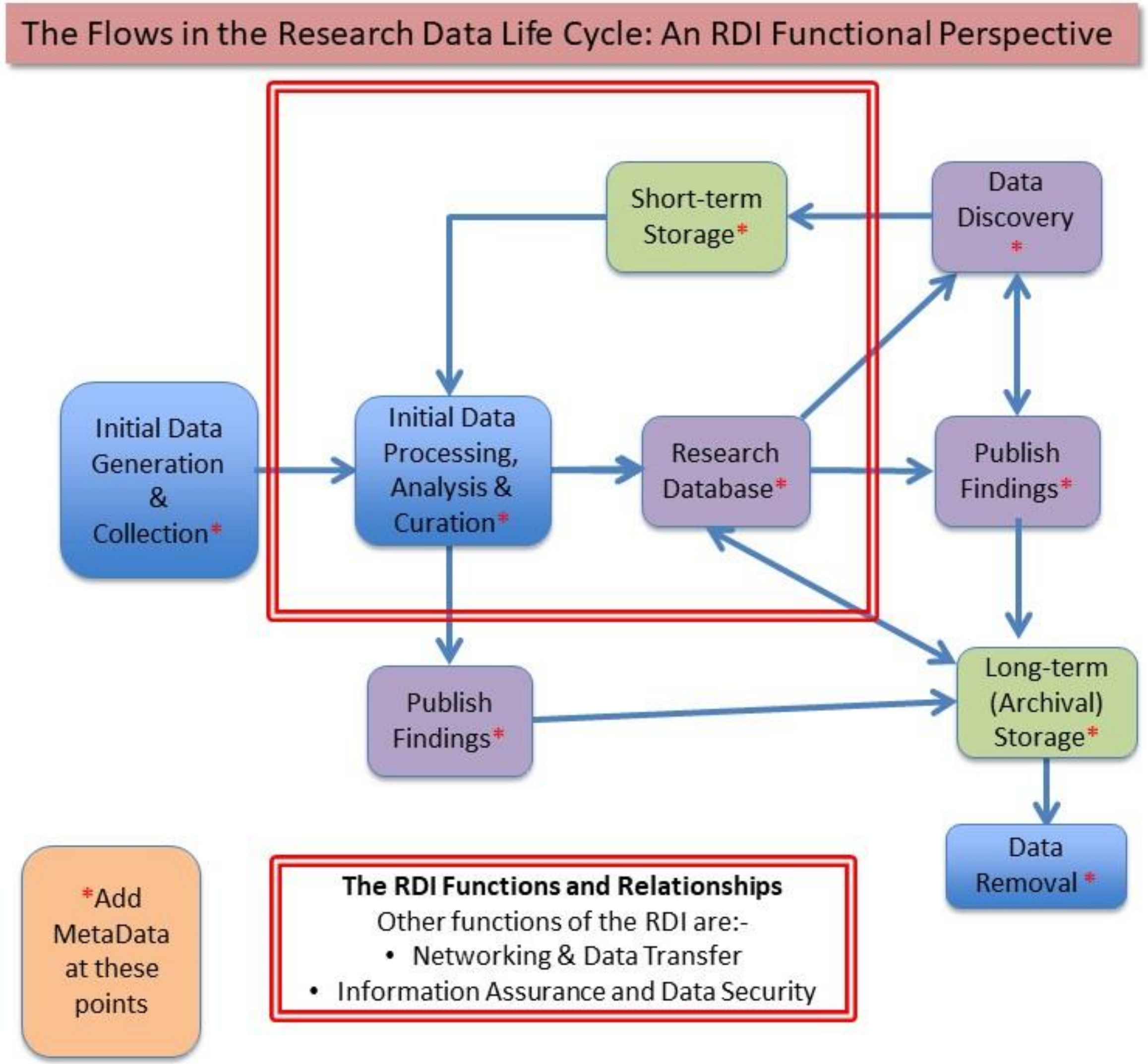
Physical Infrastructure: Storage/ Compute/ Networks/ Software
Stewardship Infrastructure: People and Skills/ Metadata, Data Curation, and Data Integration

Research Data Infrastructure



All elements related, lots of "moving parts"

Focus of the RDI White Paper


The Flows in the Research Data Life Cycle: An RDI Functional Perspective
Short-term Storage*
Data Discovery *
Initial Data Generation & Collection*
Initial Data Processing, Analysis & Curation*
Research Database*
Publish Findings*
Publish Findings*
Long-term (Archival) Storage*
Data Removal *
*Add MetaData at these points
The RDI Functions and Relationships
Other functions of the RDI are:-
• Networking & Data Transfer
• Information Assurance and Data Security

The Data Infrastructure Roadmap

timely investment and action needed now

2019: Establishing the UKRI RDI - Governance, Co-Ordination and Review

2020-2022: Maintaining the Competitiveness of the UKRI RDI

2020-2022: Transforming the UKRI RDI

2022-27: Maintaining Competitiveness and adding new Capability to the UKRI RDI

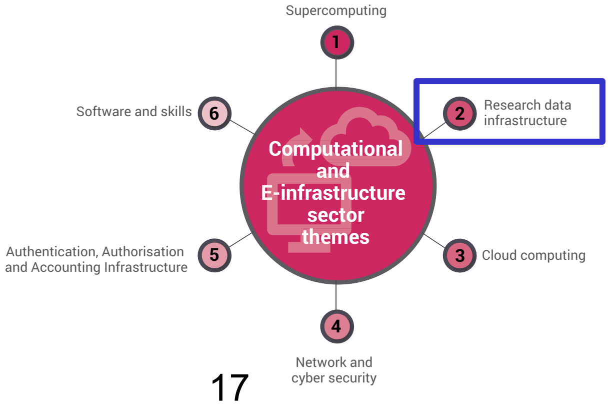
RDI Function	Roadmap Activity								
Physical Infrastructure: Hardware	Review {1}	Transformation {2}		Continued incorporation of new capabilities {3}					
	Emergency Investments to maintain competitiveness {4}			Investments to maintain competitiveness (5}					
Physical Infrastructure: Software	Review {6}	Transformation {7}		Continued incorporation of new capabilities {8}					
	Emergency Investments to maintain competitiveness {9}			Investments to maintain competitiveness {10}					
Stewardship Infrastructure: Data Curation and Data Management	Review {11}	Transformation {12}		Continued incorporation of new capabilities {13}					
	Emergency Investments to maintain competitiveness {14}			Investments to maintain competitiveness {15}					
Co-ordination of the UKRI data infrastructure	UKRI eInfrastructure governance {16}								
	Co-ordination of the Data Management and Curation activities {17}								
	Co-ordination of International Activities {18}								
Start of:	2019	2020	2021	2022	2023	2024	2025	2026	2027

The Data Infrastructure Roadmap

Recommendations for Action: pre-requisites

Physical and Stewardship Infrastructure Dependencies

- common approach to Authentication, Authorisation and resource Accounting Infrastructure (AAAI) → AAAI White Paper
- common policy framework supporting the federation of services and resources;
- end-to-end networking capability →Networking White Paper
- Collaboration tools enabling delegated management of user communities (e.g. VREs)
- Integrated approach to data anytime/anywhere
- Use of clouds / commercial or non-commercial


Supercomputing
1
Research data infrastructure
2
Cloud computing
3
Network and cyber security
4
Authentication, Authorisation and Accounting Infrastructure
5
Software and skills
6
Computational and E-infrastructure sector themes

Actions for establishing, transforming and sustaining the UKRI RDI Federation

Initial Actions

A1. Deploy new Compute and Storage capacity in annual cycles

A2. Set up the Coordination Structures for UKRI e-Infrastructure

A3. Review of current capabilities and requirements

Recommendations: The actions for maintaining and transforming the UKRI be executed on the suggested timescales with investments of **£200-300M p.a.*** beginning in 2020.

Transformation Activities

A4. Facilities/Large Projects data stewardship & science tools development/ maintenance

A5. Data integration and metadata tool development / A6. API and standards

A7. Investigate use of Commercial Cloud / A8. Ensure training activities drive **FAIR** take-up

A9. JISC capability for data research storage and re-use for data based in HEIs.

A10. Fellowships in data science

A11. AAAI, Networking and Security development

Sustained Investment in Hardware, Software & **People**

* UKRI Annual budget 18/19 ~£7.5B → £250M ~3%

FAIR: Findable, Accessible, Interoperable and Re-usable

Data, Data, Data

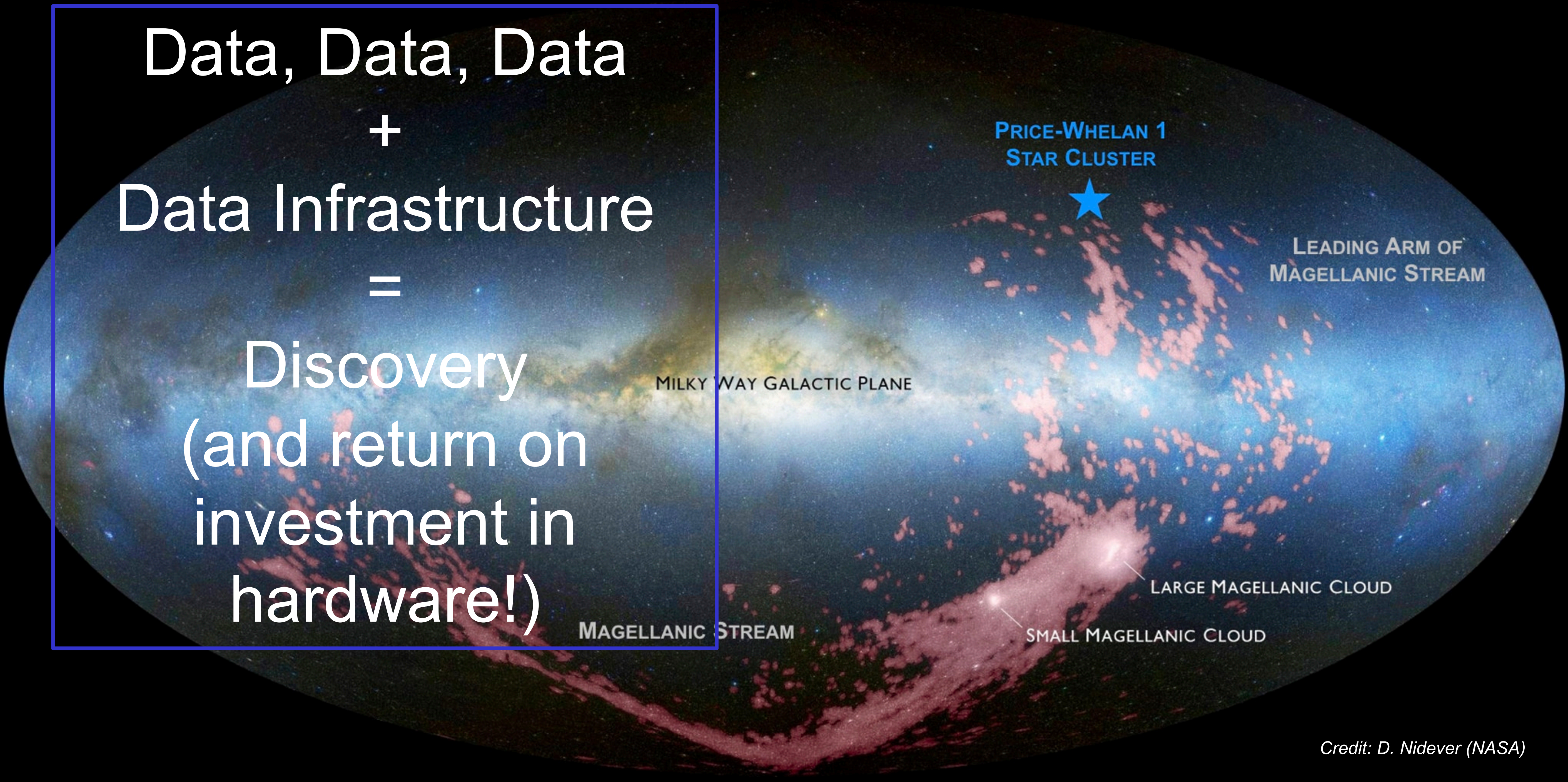
+

Data Infrastructure

=

Discovery

(and return on investment in hardware!)



Credit: D. Nidever (NASA)