ESA Gaia Variability studies as Data Science with Postgres-(XL)

Krzysztof Nienartowicz on behalf of
Gaia Data Processing Centre in Geneva,
Gaia Coordination Unit 7
PgConIT 2017

Structure

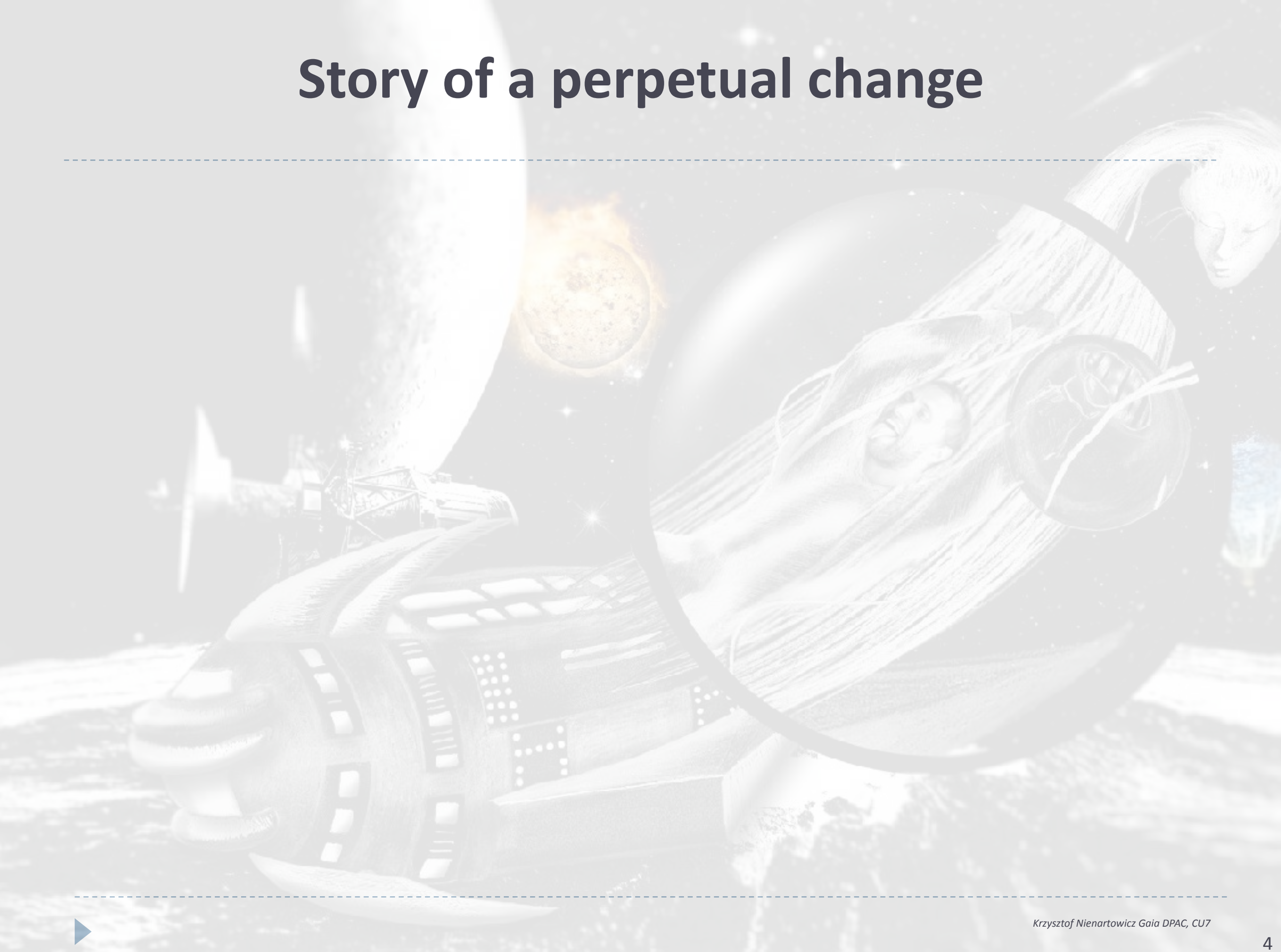
- Story of perpetual change
- Databases in Astronomy
- Gaia mission
- Gaia processing at CU7/DPC Geneva
- Postgres for science
- XL tale
- Collaboration
- Future

Bio

- Corporate software lab; Poland, USA, UK,…
  - Primark Corp-> Thomson Financial -> IHS (4.5 years)
    - The biggest economical timeseries database
    - Global systems' integration
- CERN DB group (6.5 years); Geneva, Switzerland
    - Largest data migration at the time (2002):
      - 400TB moved from Objectivity to hybrid Oracle+in-house platform
      - Largest relational scientific database running at CERN
        - Compass, Harp
    - Biomed secure middleware, Grid
- Gaia Geneva group (8 years); UNIGE/ISDC, Geneva
  - Coordination Unit 7 (CU7) Data Architect
  - Data Processing Centre for Geneva (DPCG) leader/manager

  - *XLDB, timeseries, distributed systems, ML, science, art, literature, paragliding, basketball, volleyball, architecture…*

Story of a perpetual change



Story of a perpetual change

- CERN at the time had the biggest database in the world

Story of a perpetual change

- CERN at the time had the biggest database in the world
  - Objectivity -> ~400TB

Story of a perpetual change

- CERN at the time had the biggest database in the world
  - Objectivity -> ~400TB
  - OODB (can anybody remember ODMG manifesto - 1996?)

Story of a perpetual change

- CERN at the time had the biggest database in the world
  - Objectivity -> ~400TB
  - OODB (can anybody remember ODMG manifesto - 1996?)
  - Page-store, with native C++ binding

Story of a perpetual change

- CERN at the time had the biggest database in the world
  - Objectivity -> ~400TB
  - OODB (can anybody remember ODMG manifesto - 1996?)
  - Page-store, with native C++ binding
- Quite fast

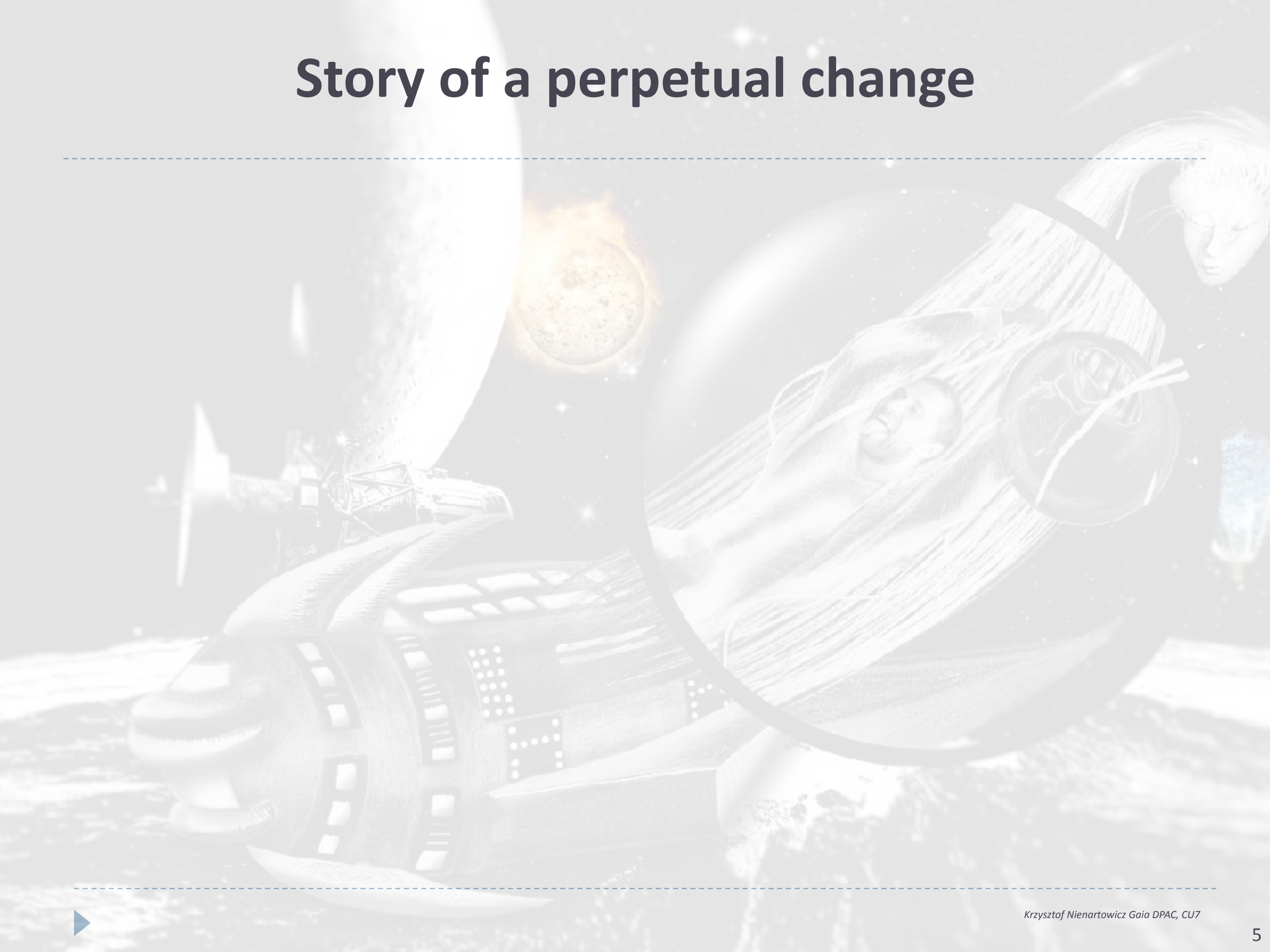
Story of a perpetual change

- CERN at the time had the biggest database in the world
  - Objectivity -> ~400TB
  - OODB (can anybody remember ODMG manifesto - 1996?)
  - Page-store, with native C++ binding
  - Quite fast
  - Expensive

Story of a perpetual change

- CERN at the time had the biggest database in the world
  - Objectivity -> ~400TB
  - OODB (can anybody remember ODMG manifesto - 1996?)
  - Page-store, with native C++ binding
  - Quite fast
  - Expensive
  - Niche..

Story of a perpetual change



Story of a perpetual change

- OODB had been at the time what NoSQL has been recently

Story of a perpetual change

- OODB had been at the time what NoSQL has been recently
  - *Impedance mismatch*

Story of a perpetual change

- OODB had been at the time what NoSQL has been recently
  - *Impedance mismatch*
  - *Graph vs relational*

Story of a perpetual change

- OODB had been at the time what NoSQL has been recently
  - *Impedance mismatch*
  - *Graph vs relational*
    - RAM *pointer-swizzling*

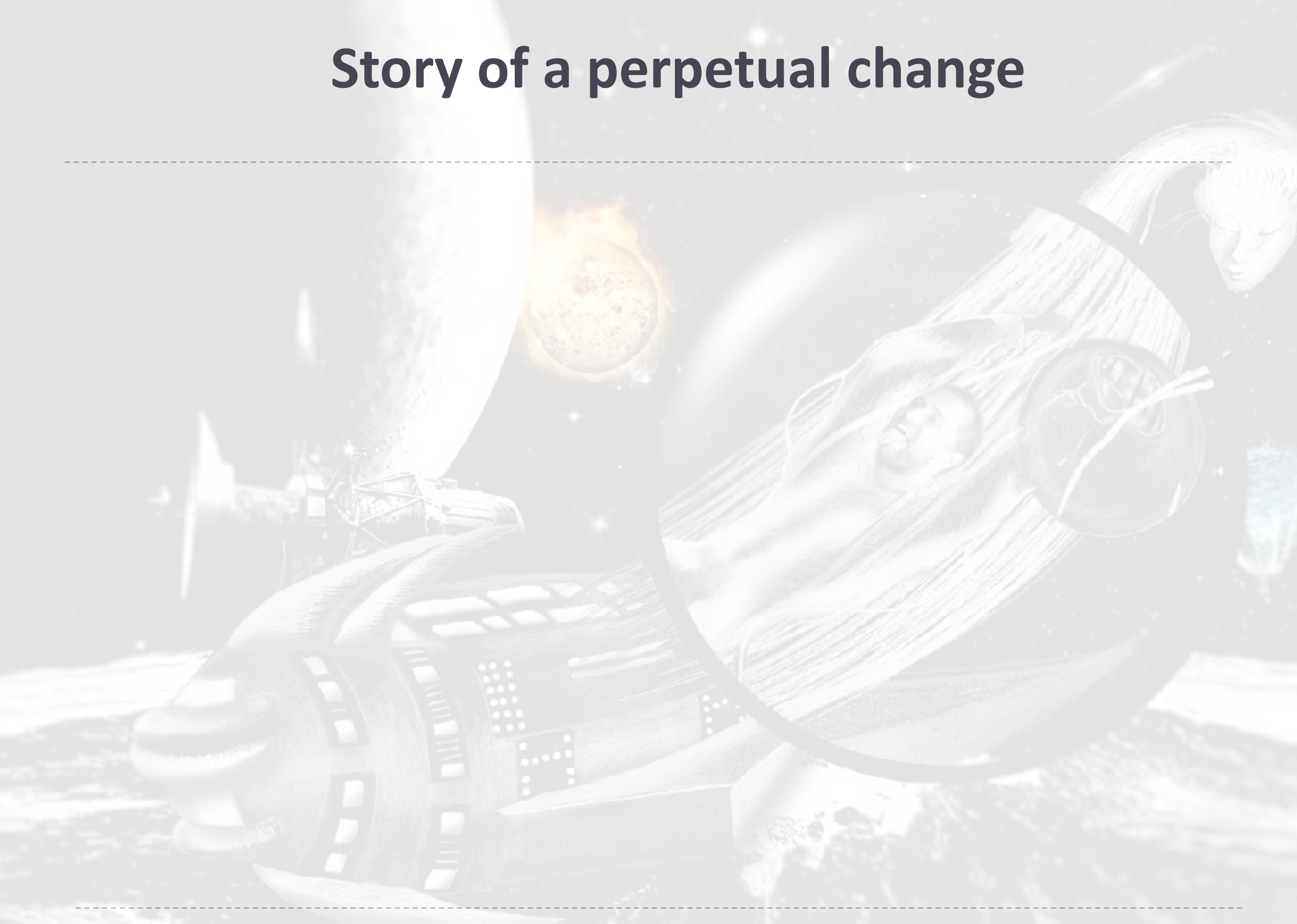
Story of a perpetual change

- OODB had been at the time what NoSQL has been recently
  - *Impedance mismatch*
  - *Graph vs relational*
    - RAM *pointer-swizzling*
- *OODB vs Relational DBMS:*

Story of a perpetual change

- OODB had been at the time what NoSQL has been recently
  - *Impedance mismatch*
  - *Graph vs relational*
    - RAM *pointer-swizzling*
- *OODB vs Relational DBMS:*
  - *Two competing philosophies*

Story of a perpetual change



Story of a perpetual change

- Oracle offered their help

Story of a perpetual change

- Oracle offered their help
  - First RAC installations

Story of a perpetual change

- Oracle offered their help
  - First RAC installations
  - C++ bindings

Story of a perpetual change

- Oracle offered their help
  - First RAC installations
  - C++ bindings
  - Biggest data migration at the time

Story of a perpetual change

- Oracle offered their help
  - First RAC installations
  - C++ bindings
  - Biggest data migration at the time
- SDSS -> Jim Gray -> Sloan Sky Digital Survey (2000-…)

Story of a perpetual change

- Oracle offered their help
  - First RAC installations
  - C++ bindings
  - Biggest data migration at the time
    - SDSS -> Jim Gray -> Sloan Sky Digital Survey (2000-…)
- MapReduce vs RDBMS vs newSQL

Structure

- Story of perpetual change
- Databases in Astronomy
- Gaia mission
- Gaia processing at CU7/DPC Geneva
- Postgres for science
- XL tale
- Collaboration
- Future

Databases in Astronomy

- There have been many
  - File based:
    - FITS: Flexible Image Transport System -> Image container
  - **Sloan Digital Sky Survey** (1998, 40TB raw + 3TB processed)
  - Internal competition between Objectivity (OODB) and MsSql (RDBMS) - Jim Gray
  - *“He was asking questions, then after some time coming back with a SQL solution which was always better than one of Objectivity (to our frustration)…” Peter Z. Kunszt, ~2004*

Databases in Astronomy

Databases in Astronomy

- New wave of Big Data astronomical projects

Databases in Astronomy

- New wave of Big Data astronomical projects
  - Sensor data

Databases in Astronomy

- New wave of Big Data astronomical projects
  - Sensor data
    - ESA **Gaia** 2014-..(1PB - replicated)

Databases in Astronomy

- New wave of Big Data astronomical projects
  - Sensor data
    - ESA **Gaia** 2014-..(1PB - replicated)
    - ESA **Euclid** 2021 - (10PB)

Databases in Astronomy

- New wave of Big Data astronomical projects
  - Sensor data
    - ESA **Gaia** 2014-..(1PB - replicated)
    - ESA **Euclid** 2021 - (10PB)
    - **JWST** 2019-.. (>50TB over 5 years)

Databases in Astronomy

- New wave of Big Data astronomical projects
  - Sensor data
    - ESA **Gaia** 2014-..(1PB - replicated)
    - ESA **Euclid** 2021 - (10PB)
    - **JWST** 2019-.. (>50TB over 5 years)
    - **LSST** *Large Sky Survey Telescope* 2023-…(60PB-> 15PB DB)

Databases in Astronomy

- New wave of Big Data astronomical projects
  - Sensor data
    - ESA **Gaia** 2014-..(1PB - replicated)
    - ESA **Euclid** 2021 - (10PB)
    - **JWST** 2019-.. (>50TB over 5 years)
    - **LSST** *Large Sky Survey Telescope* 2023-…(60PB-> 15PB DB)
    - **SKA** *Square Kilometre Array Telescope, 2022-…*
      (300-500PB per year! - 5 ExaBytes)

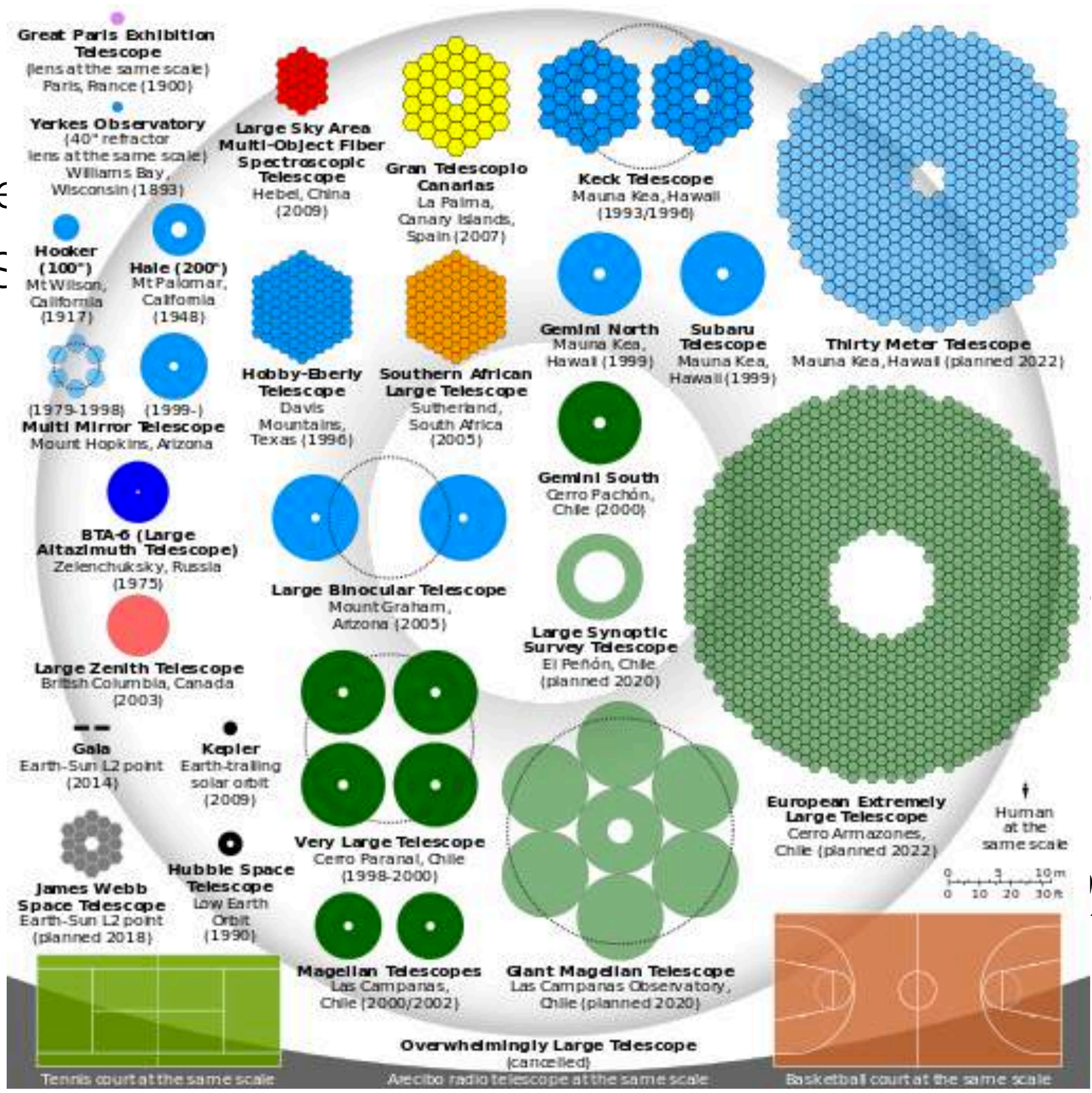
Databases in Astronomy

- New wave of Big Data astronomical projects
  - Sensor data
    - ESA **Gaia** 2014-..(1PB - replicated)
    - ESA **Euclid** 2021 - (10PB)
    - **JWST** 2019-.. (>50TB over 5 years)
    - **LSST** *Large Sky Survey Telescope* 2023-…(60PB-> 15PB DB)
    - **SKA** *Square Kilometre Array Telescope, 2022-… (300-500PB per year! - 5 ExaBytes)*
    - **CTA** *Cherenkov Telescope Array (100PB by 2030)*

Databases in Astronomy

- New wave of Big Data astronomical projects
  - Sensor data
    - ESA **Gaia** 2014-..(1PB - replicated)
    - ESA **Euclid** 2021 - (10PB)
    - **JWST** 2019-.. (>50TB over 5 years)
    - **LSST** *Large Sky Survey Telescope* 2023-…(60PB-> 15PB DB)
    - **SKA** *Square Kilometre Array Telescope, 2022-… (300-500PB per year! - 5 ExaBytes)*
    - **CTA** *Cherenkov Telescope Array (100PB by 2030)*
    - **E-ELT, VLT**… the list goes on..

- Ne
  - S
    -
    -
    -
    -
    -
    -
    -
    -
    -
    -


Great Paris Exhibition Telescope
(lens at the same scale)
Paris, France (1900)
Yerkes Observatory
(40" refractor lens at the same scale)
Williams Bay, Wisconsin (1893)
Hooker (100")
Mt Wilson, California (1917)
Hale (200")
Mt Palomar, California (1948)
(1979-1998)
(1999-)
Multi Mirror Telescope
Mount Hopkins, Arizona
BTA-6 (Large Altazimuth Telescope)
Zelenchuksky, Russia (1975)
Large Zenith Telescope
British Columbia, Canada (2003)
Gaia
Earth-Sun L2 point (2014)
Kepler
Earth-trailing solar orbit (2009)
James Webb Space Telescope
Earth-Sun L2 point (planned 2018)
Hubble Space Telescope
Low Earth Orbit (1990)
Large Sky Area Multi-Object Fiber Spectroscopic Telescope
Hebei, China (2009)
Gran Telescopio Canarias
La Palma, Canary Islands, Spain (2007)
Keck Telescope
Mauna Kea, Hawaii (1993/1996)
Hobby-Eberly Telescope
Davis Mountains, Texas (1996)
Southern African Large Telescope
Sutherland, South Africa (2005)
Large Binocular Telescope
Mount Graham, Arizona (2005)
Very Large Telescope
Cerro Paranal, Chile (1998-2000)
Magellan Telescopes
Las Campanas, Chile (2000/2002)
Gemini North
Mauna Kea, Hawaii (1999)
Subaru Telescope
Mauna Kea, Hawaii (1999)
Gemini South
Cerro Pachón, Chile (2000)
Large Synoptic Survey Telescope
El Peñón, Chile (planned 2020)
Giant Magellan Telescope
Las Campanas Observatory, Chile (planned 2020)
Thirty Meter Telescope
Mauna Kea, Hawaii (planned 2022)
European Extremely Large Telescope
Cerro Armazones, Chile (planned 2022)
Human at the same scale
0 5 10 m
0 10 20 30 ft
Overwhelmingly Large Telescope
(cancelled)
Tennis court at the same scale
Arecibo radio telescope at the same scale
Basketball court at the same scale

Databases in Astronomy

Databases in Astronomy

- Awareness of DB advantages has been growing over recent years

Databases in Astronomy

- Awareness of DB advantages has been growing over recent years
  - Scientist => Data Scientist

Databases in Astronomy

- Awareness of DB advantages has been growing over recent years
  - Scientist => Data Scientist
  - Role of the **newSQL**

Databases in Astronomy

- Awareness of DB advantages has been growing over recent years
  - Scientist => Data Scientist
  - Role of the **newSQL**
- Competition: Document stores, Streaming frameworks, Specialized hardware

Databases in Astronomy

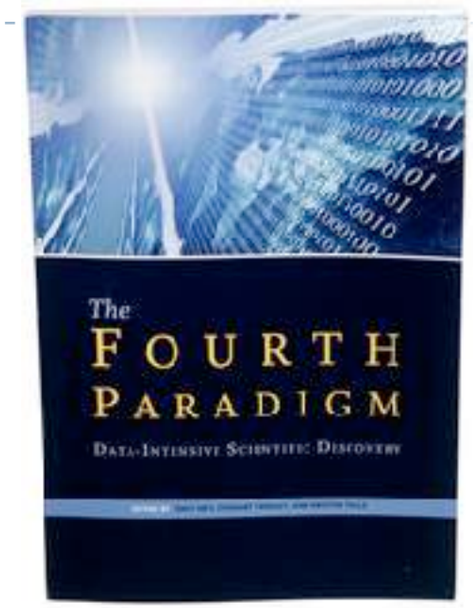
- Awareness of DB advantages has been growing over recent years
  - Scientist => Data Scientist
  - Role of the **newSQL**
    - Competition: Document stores, Streaming frameworks, Specialized hardware
- Future: Polyglot-Hybrids

Databases in Astronomy

- Awareness of DB advantages has been growing over recent years
  - Scientist => Data Scientist
  - Role of the **newSQL**
    - Competition: Document stores, Streaming frameworks, Specialized hardware
- Future: Polyglot-Hybrids
  - Mix of all -> project expertise dependent

Driving force

Information is beautiful…



- [4th Paradigm by Jim Gray]

  … to have a world in which all of the science literature is online, all of the science data is online, and they interoperate with each other...

- ***and methods, including code is online as well***

Gaia premise

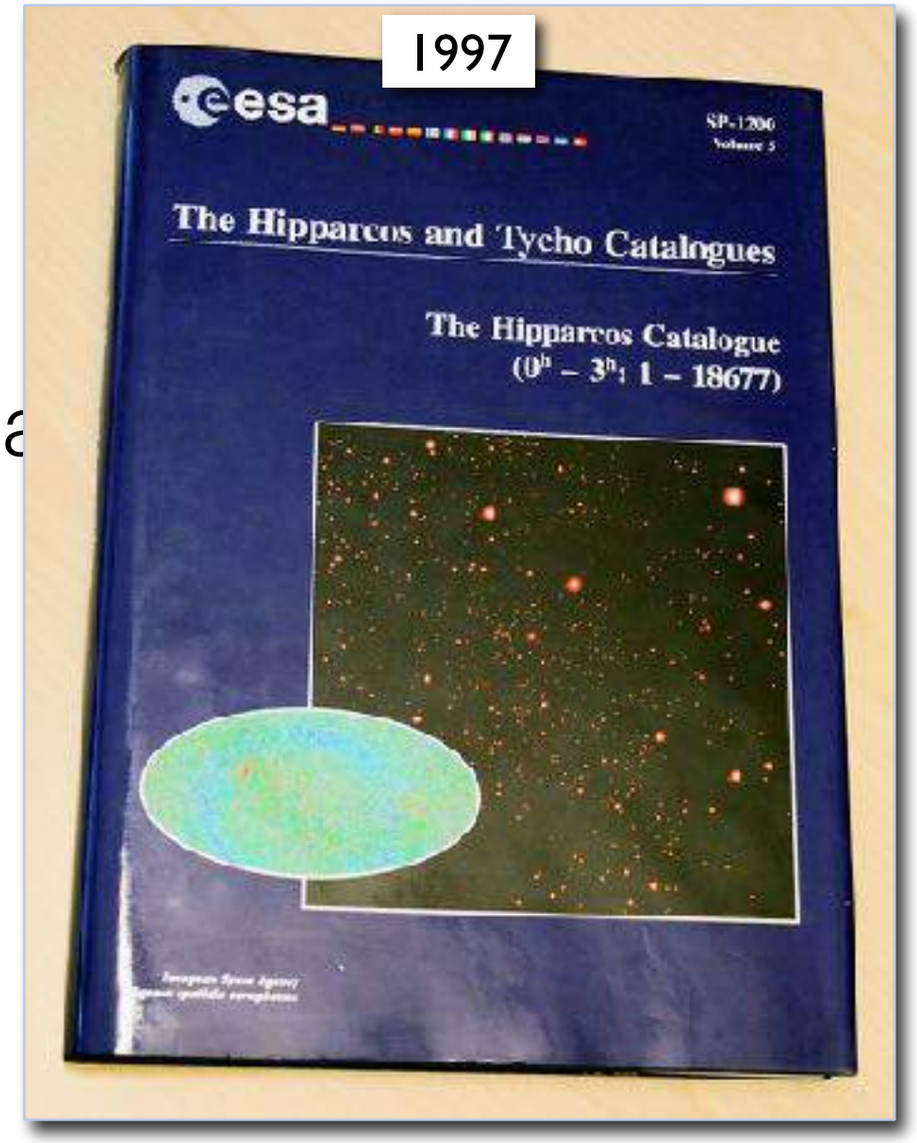
- Publication of all data
- Publication of open Apache-licensed code
  - Publication of methods (i.e. 10+ different period search methods, hierarchical supervised classification, unsupervised classification)
- At the end of the mission
  - Data Release 4, 2022
  - *Data Release 2, April 2018*
- Via Web-archive, service, for offline use, self descriptive..

Gaia premise

- Gaia Catalogue
  - Gaia Archive

Gaia premise

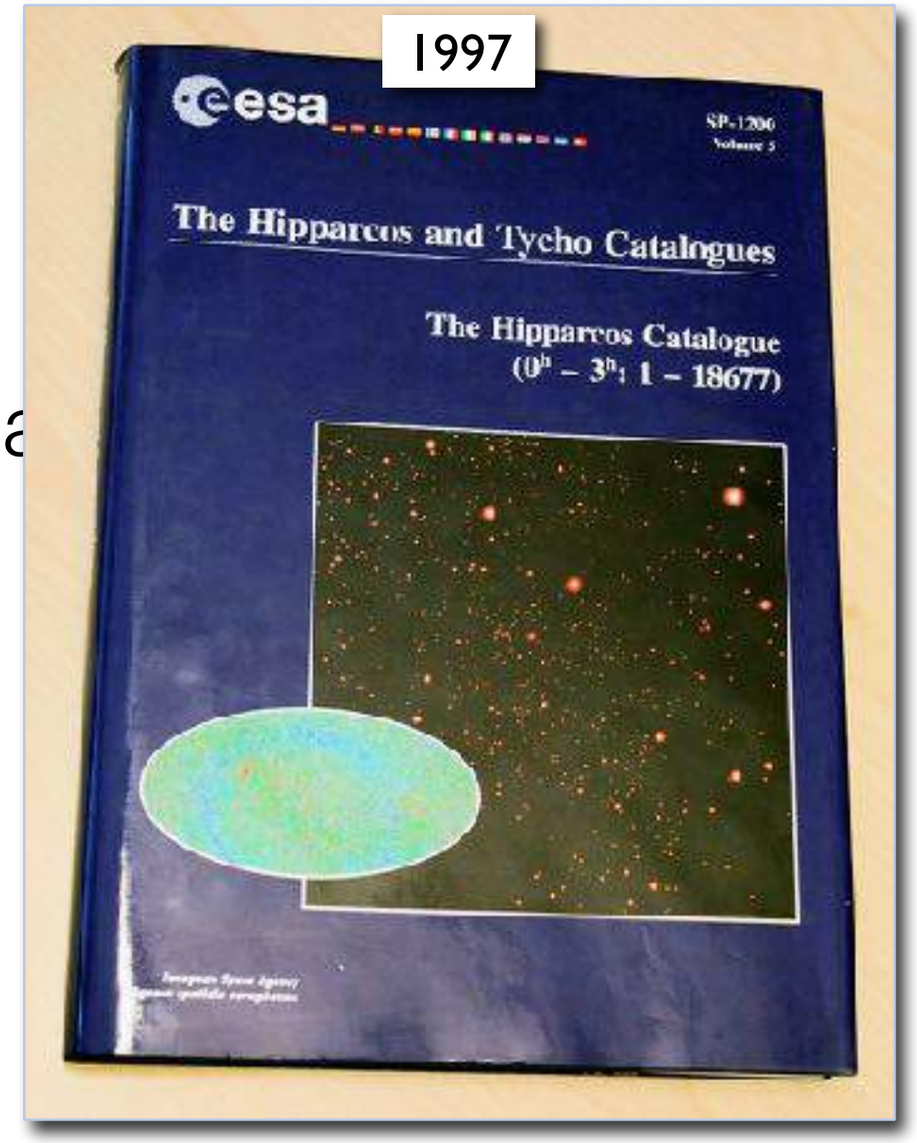
- Gaia
  - Gaia



Gaia premise

Hipparcos vs Gaia catalogue
Ultimate goal

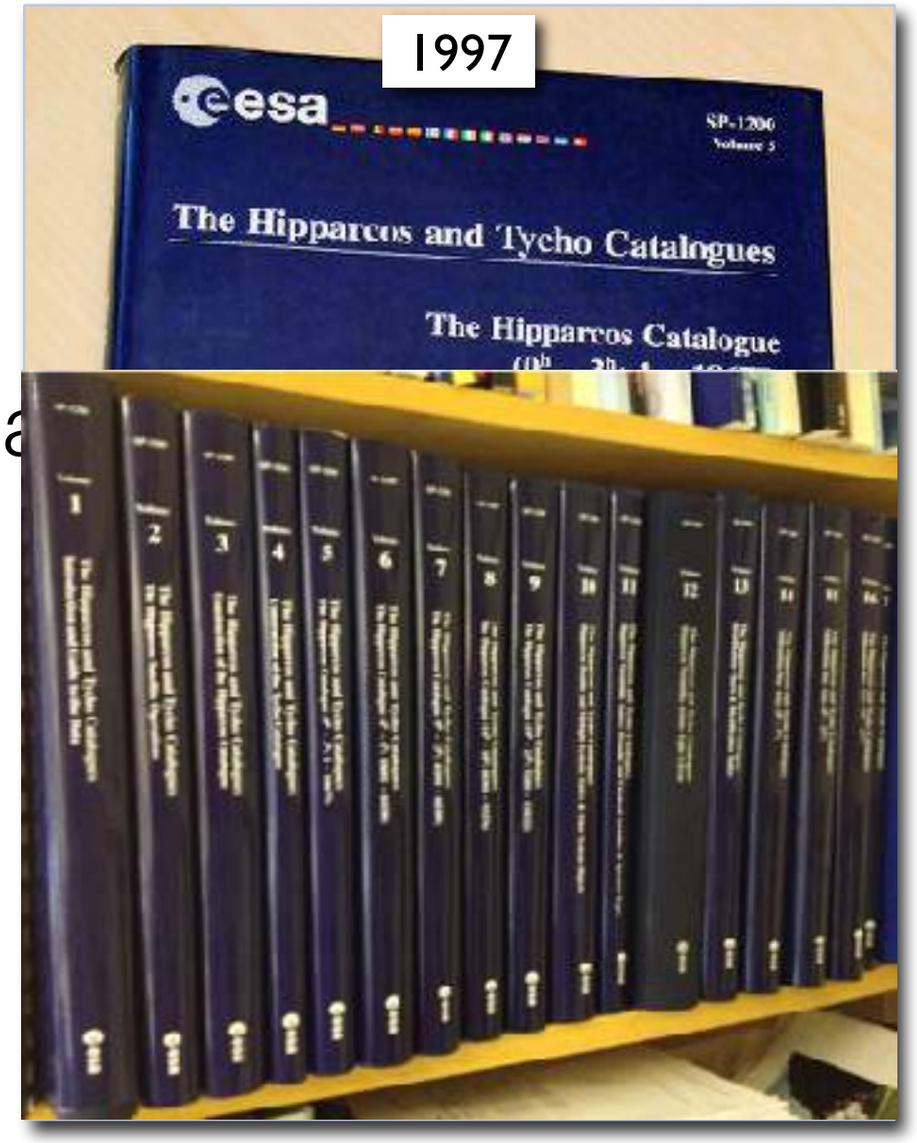
- Gaia
  - Gaia



Gaia premise

Hipparcos vs Gaia catalogue
Ultimate goal

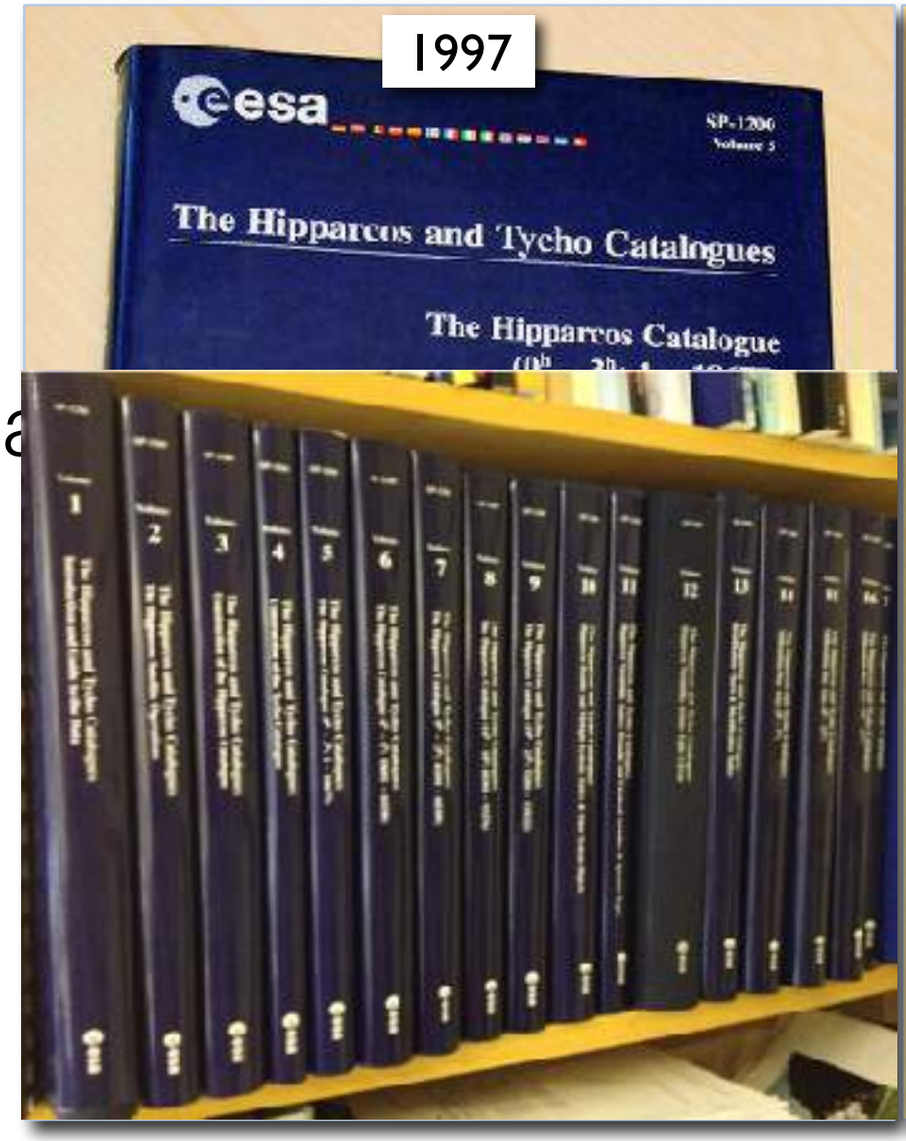
- Gaia
  - Gaia

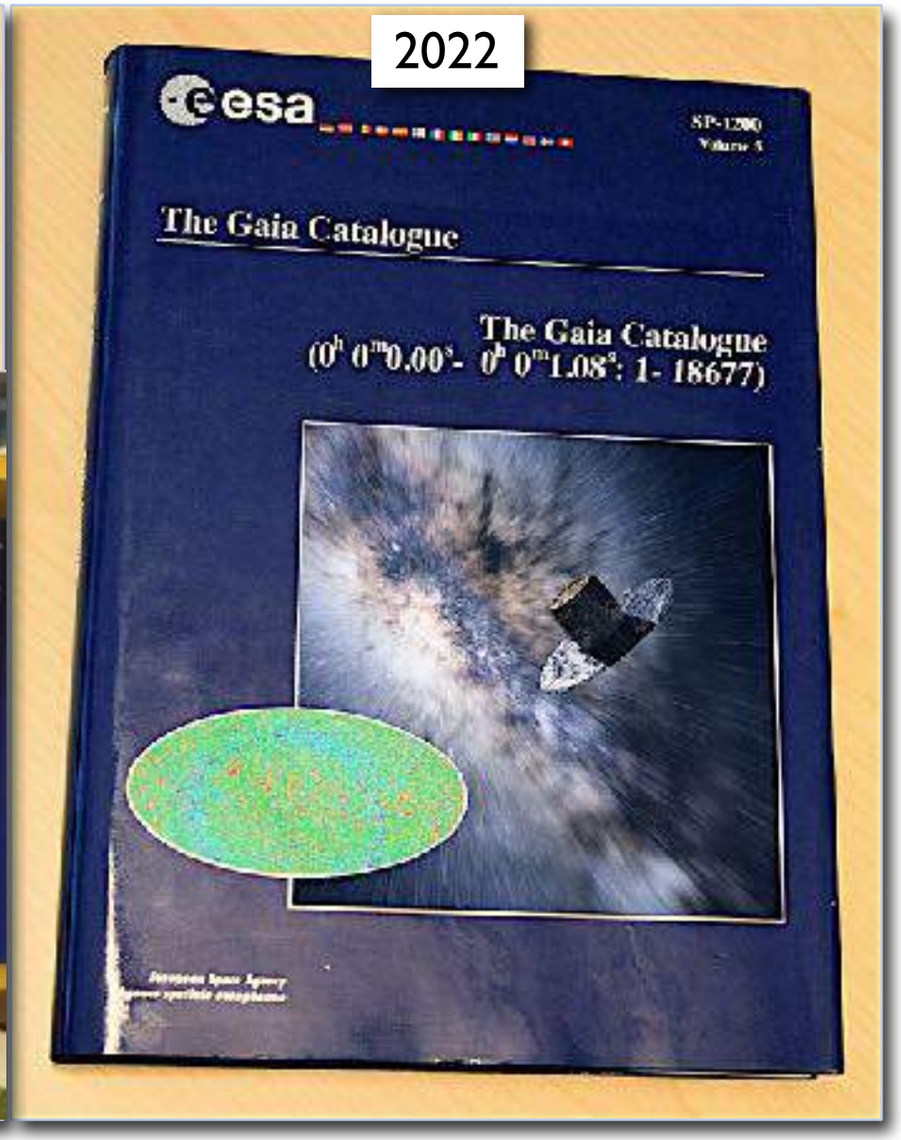


Gaia premise

Hipparcos vs Gaia catalogue
Ultimate goal

- Gaia
  - Gaia

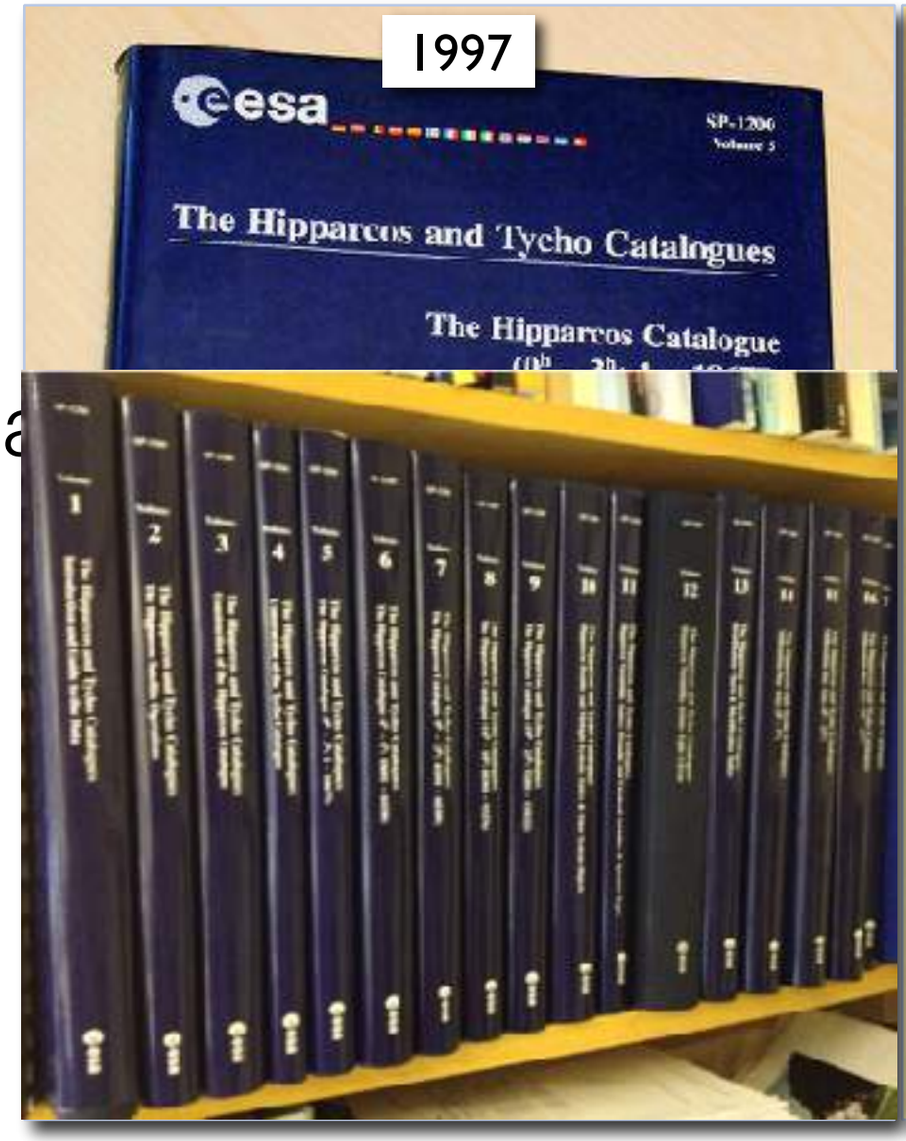


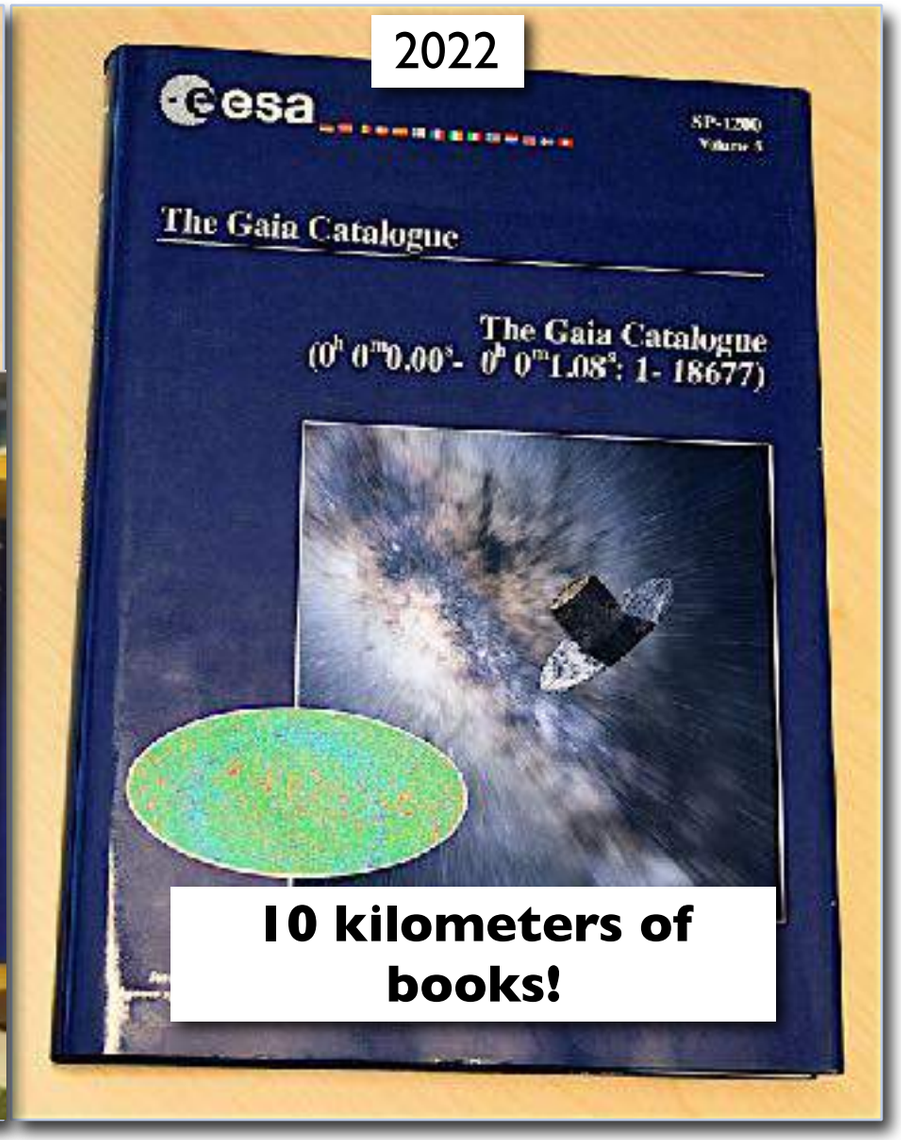


Gaia premise

Hipparcos vs Gaia catalogue
Ultimate goal

- Gaia
  - Gaia





Structure

- Story of perpetual change
- Databases in Astronomy
- Gaia mission
- Gaia processing at CU7/DPC Geneva
- Postgres for science
- Postgres-XL tale
- Collaboration
- Future

What is so special about Gaia?

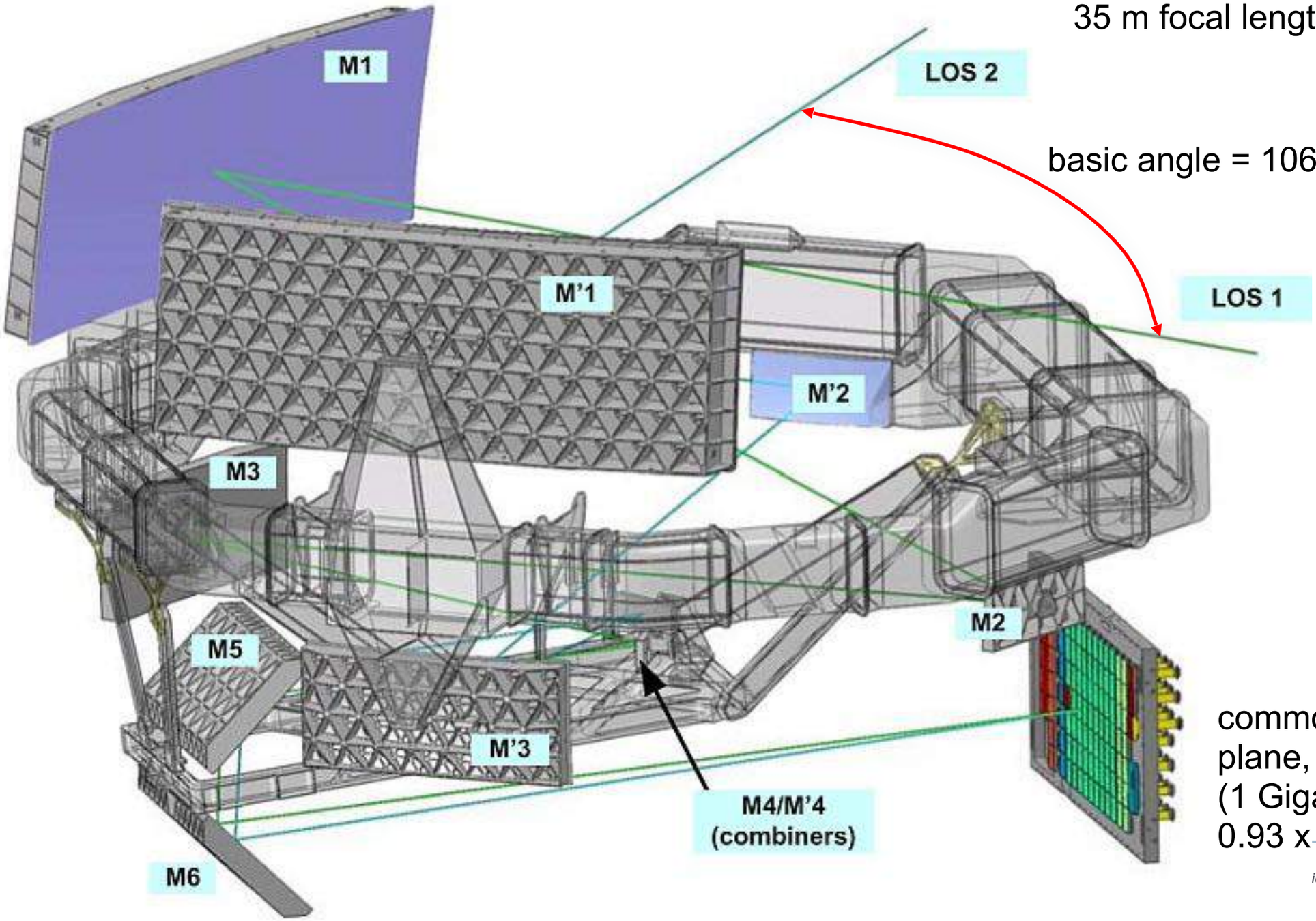
- European Space Agency **cornerstone** mission
  - No equivalent mission for 20-30+ years…
- Census of our Galaxy:
  - **All objects between 6 and 20th magnitude (~1.7B stars, asteroids, quasars, extragalactic supernovae, variables)**
  - **On average 80 measurements during its 5 year mission**
    - positions and parallax with a precision of 20 µasec (at V= 15 mag)
    - Proper motions with a precision of 20 µasec/year (at V= 15 mag)
    - Radial velocities with a precision of 2-10 km/s (for star V<17)
      - **Low resolution spectrum of each star:**
        - allows to determine many stellar properties
          - e.g. temperature, surface gravity, metalicity, age, ...
  - Can potentially discover ~10.000 exo-planets
  - **Estimated 10-20% of all population are variables**

The Gaia instruments

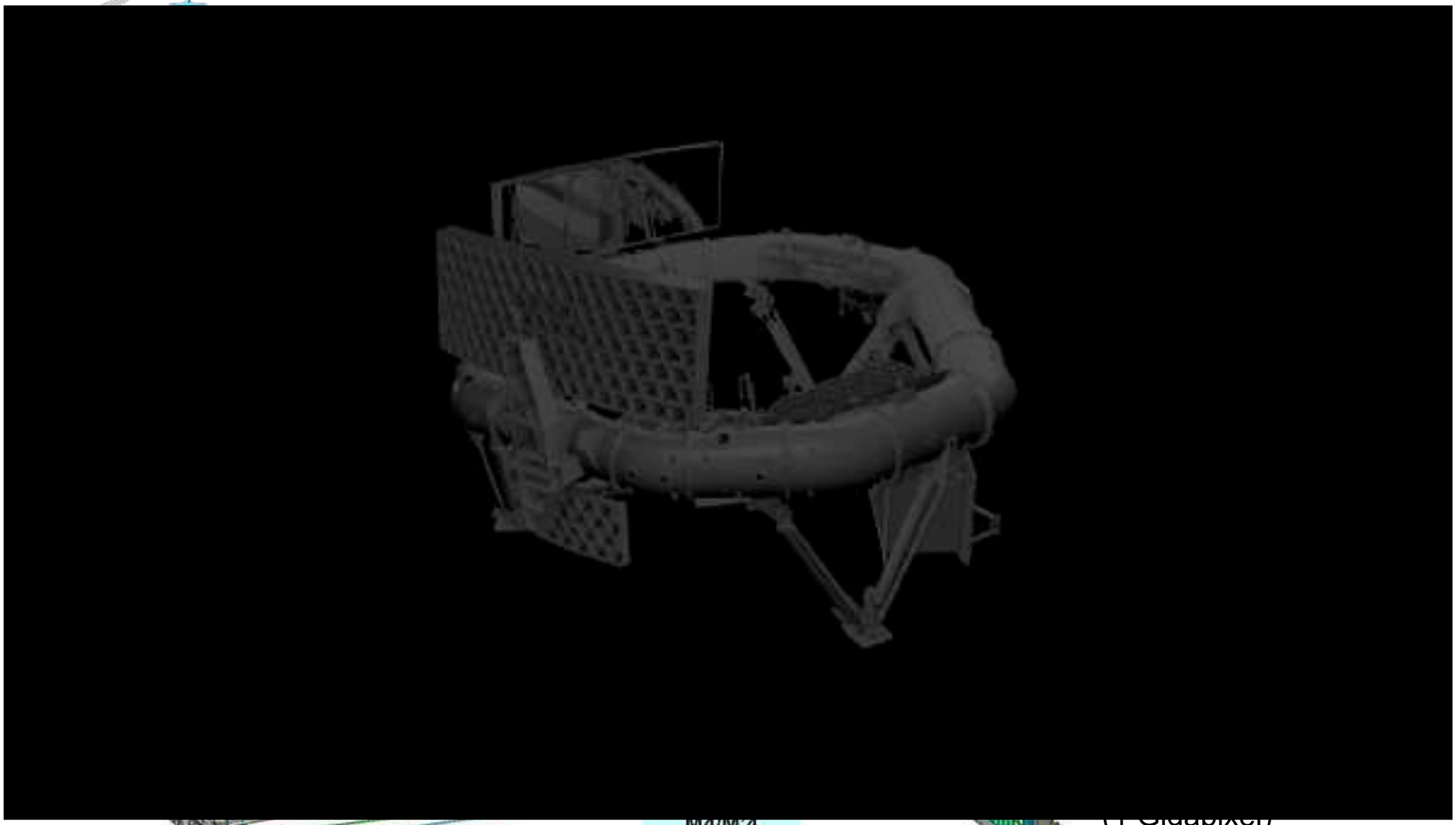
2 off-axis telescopes
1.45 x 0.5 m² aperture
35 m focal length

basic angle = 106.5°

common focal plane, 106 CCDs (1 Gigapixel)
0.93 x 0.42 m²



The Gaia instruments



(1 Gigapixel)
0.93 x 0.42 m²

One of the two primary mirrors

The Gaia satellite

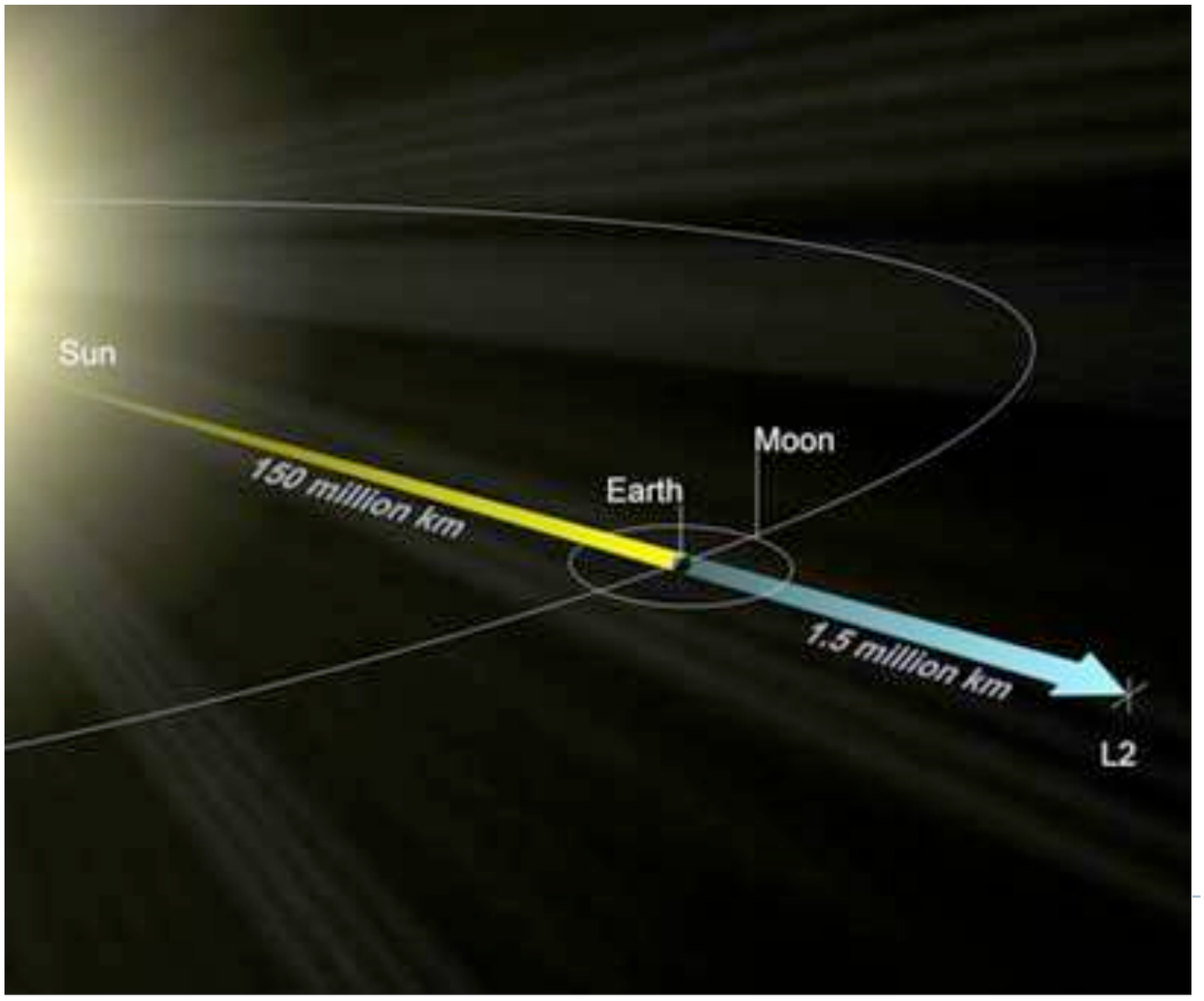
Location: Lagrange point 2

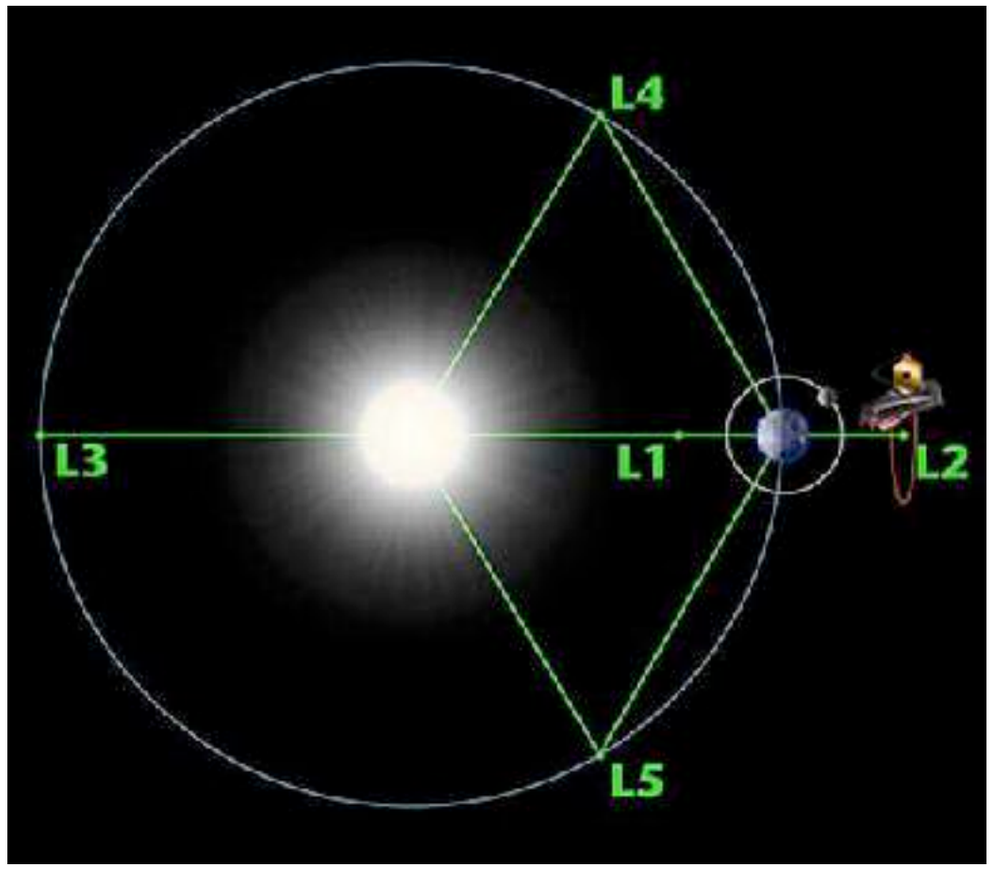
Commissioning phase, first calibrated data: Q2/2015

The Gaia satellite

Location: Lagrange point 2

Commissioning phase, first calibrated data: Q2/2015





Gaia scanning: Motion of viewing directions over **4 days**

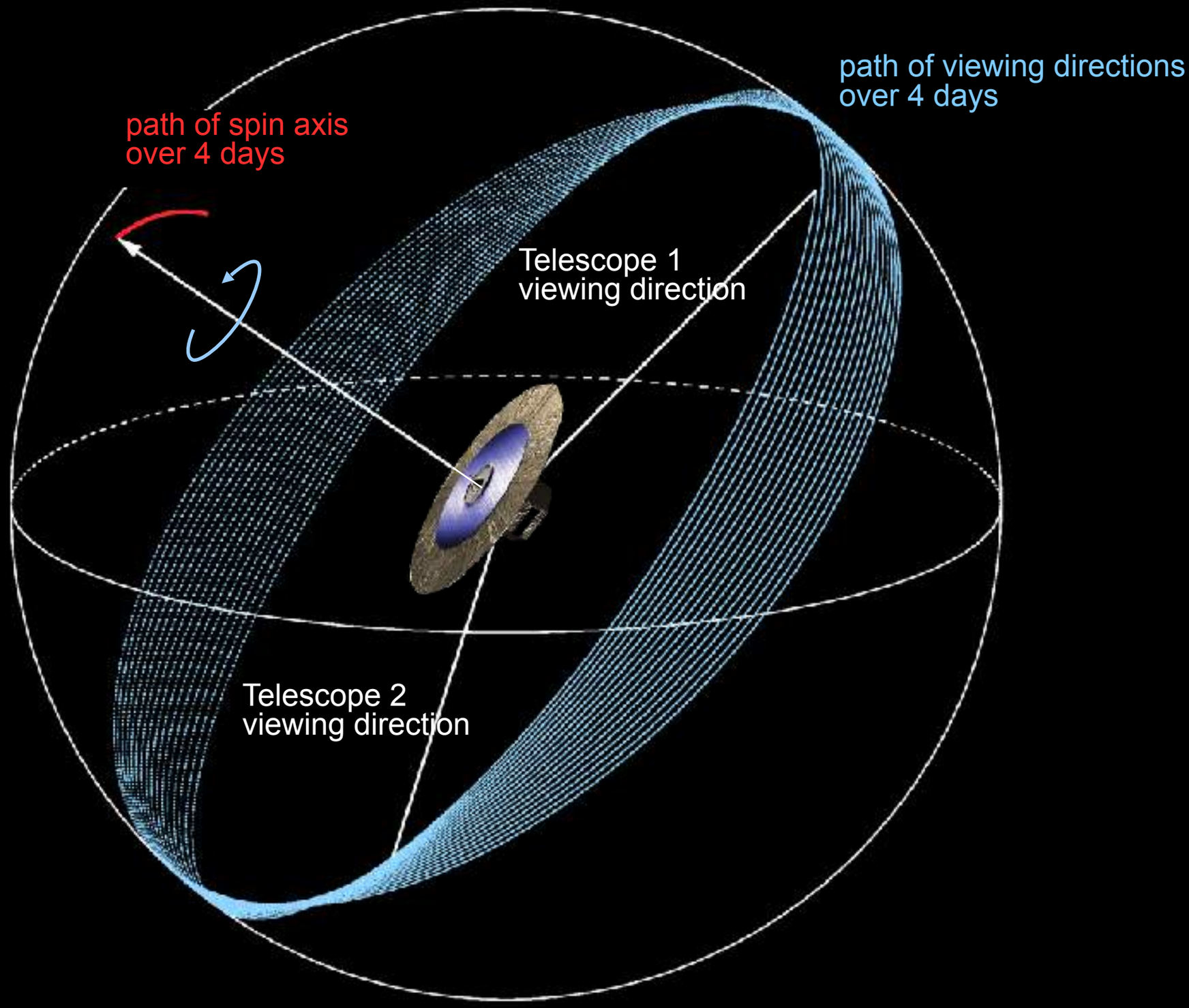


image credit: Lennart Lindegren

Gaia scanning: Motion of the spin axis over **4 months**

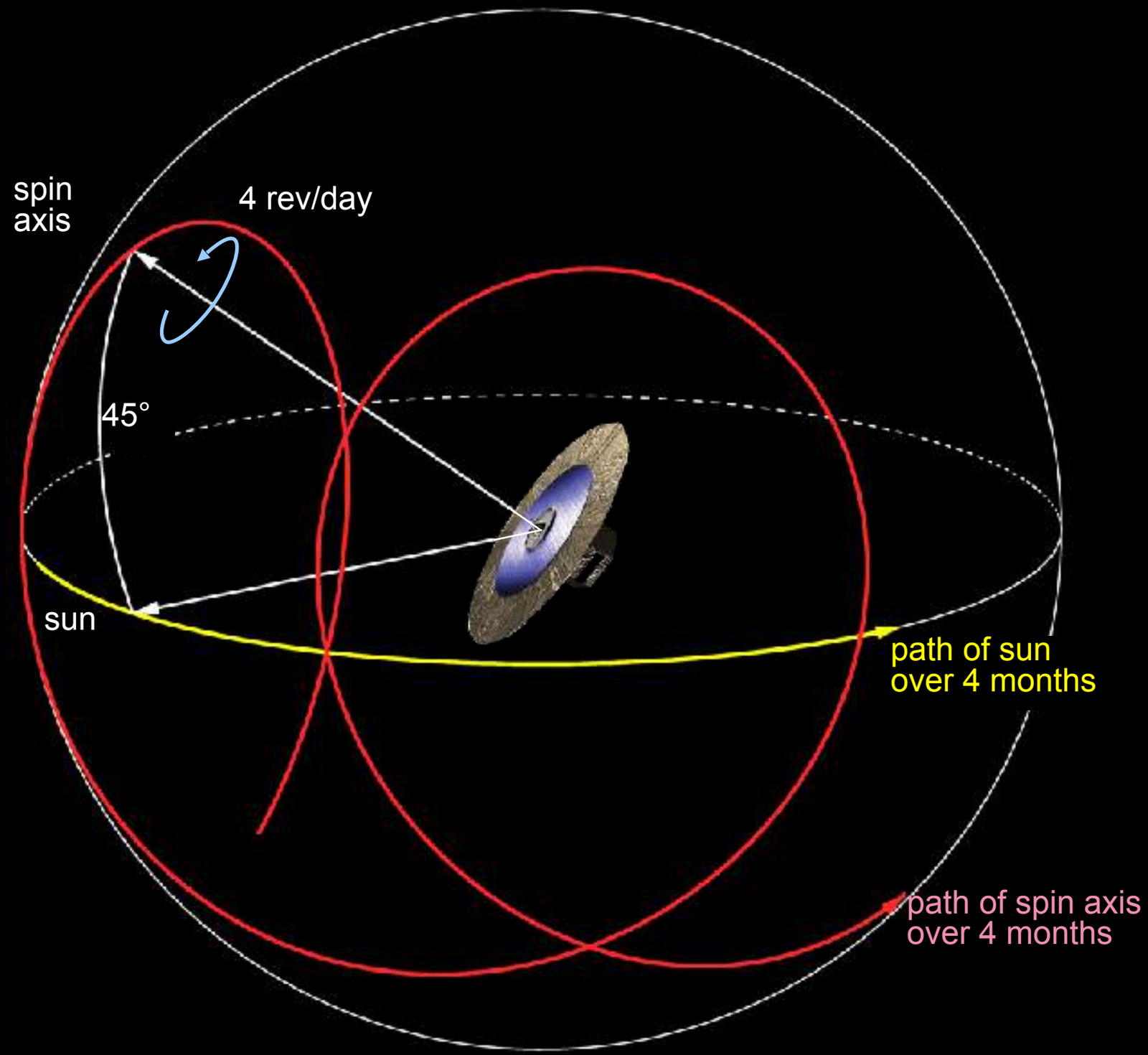
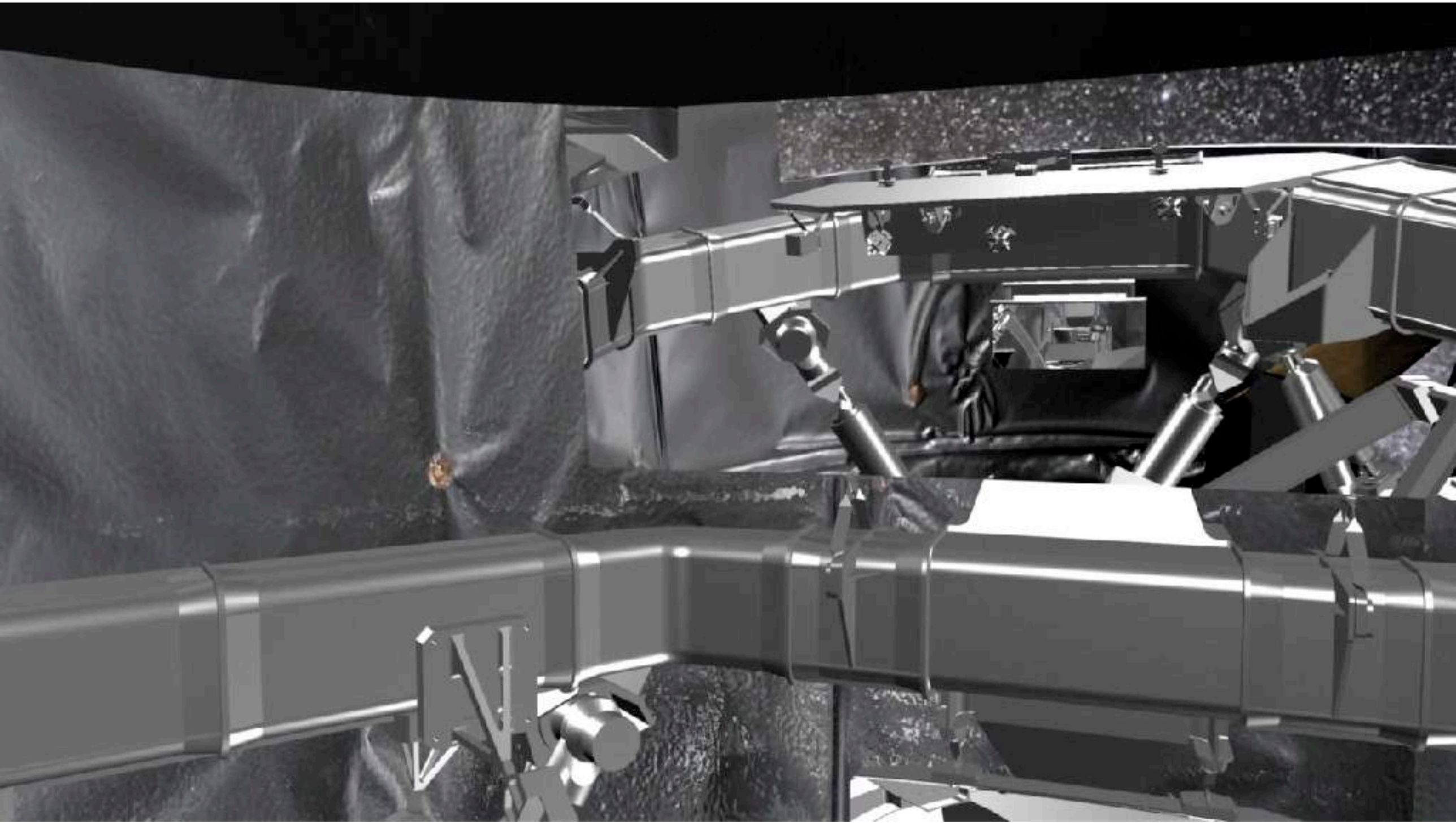
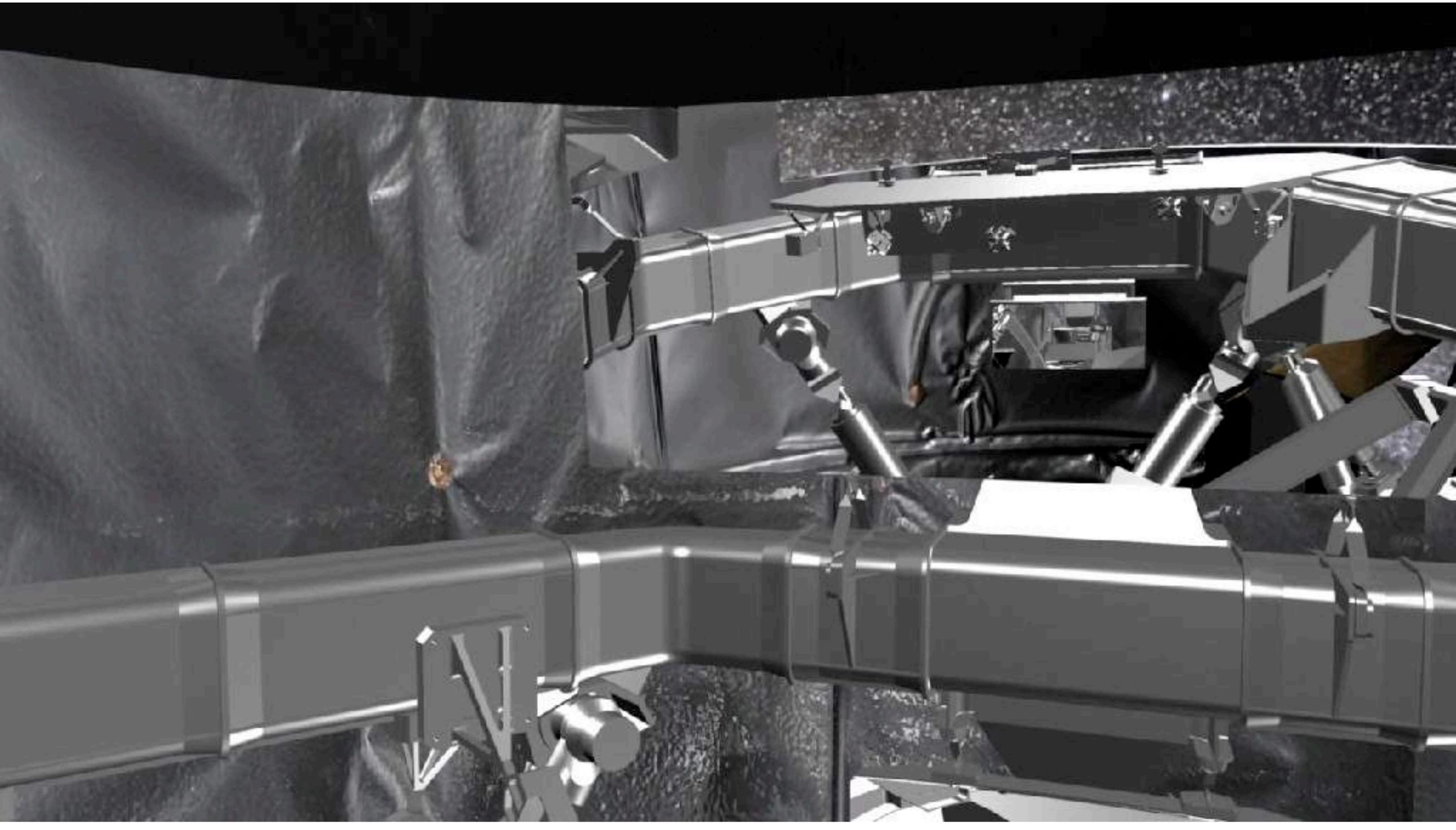
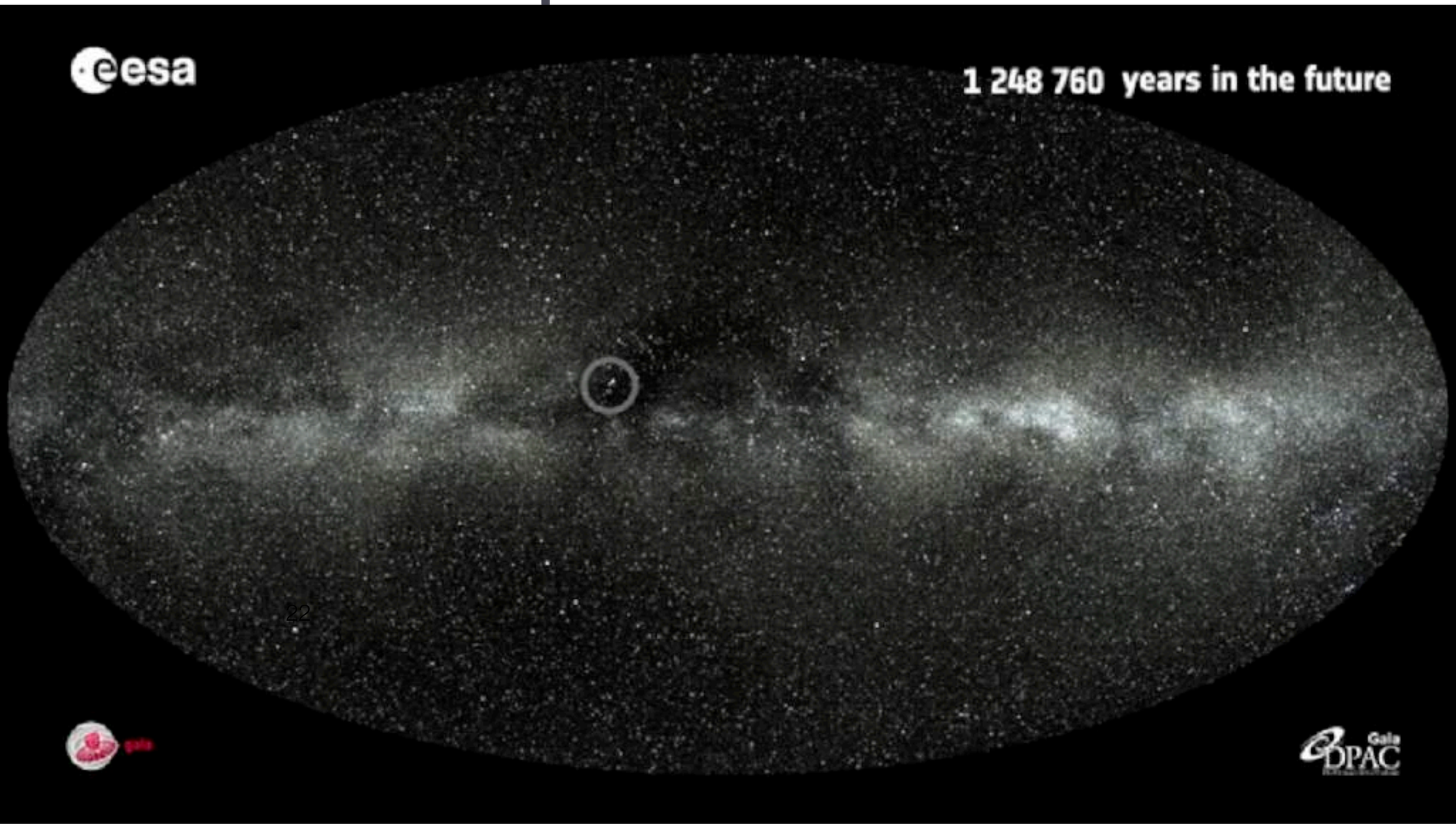


image credit: Lennart Lindegren

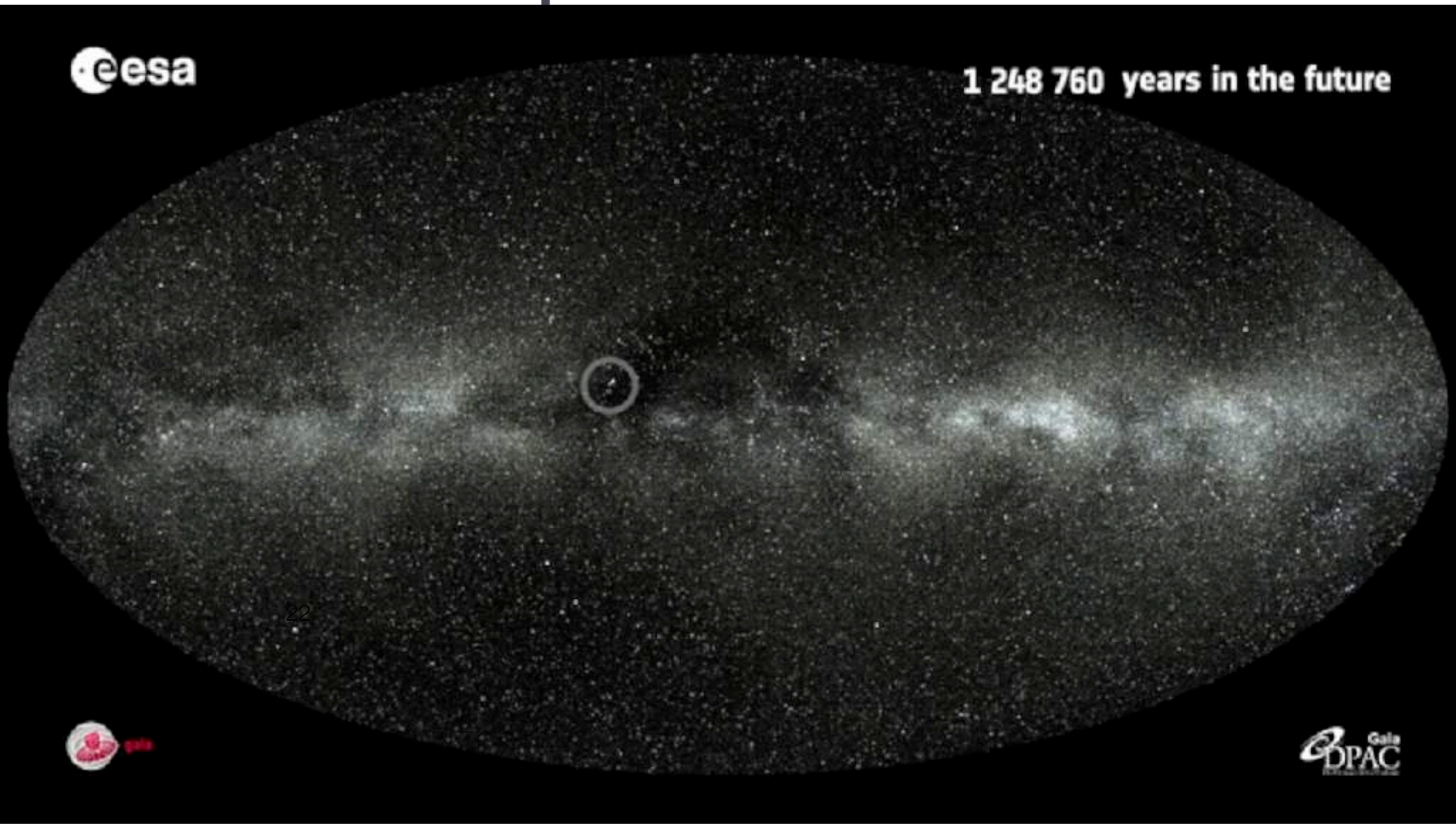




Proper motion in 3D



Proper motion in 3D



Structure

- Story of perpetual change
- Databases in Astronomy
- Gaia mission
- Gaia processing at CU7/DPC Geneva
- Postgres for science
- Postgres-XL tale
- Collaboration
- Future

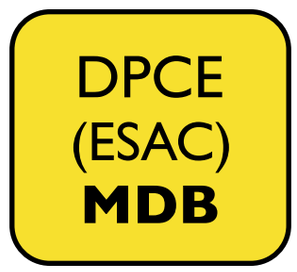
Gaia Consortium - distributed challenge

CU's process data in Data Processing centre(s):

24

Gaia Consortium - distributed challenge

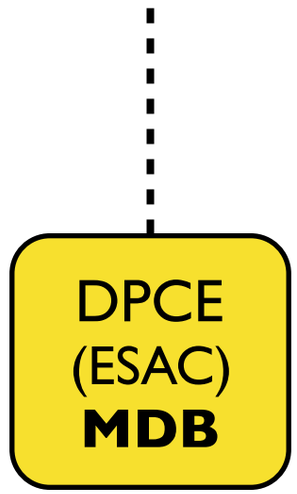
CU's process data in Data Processing centre(s):



24

Gaia Consortium - distributed challenge

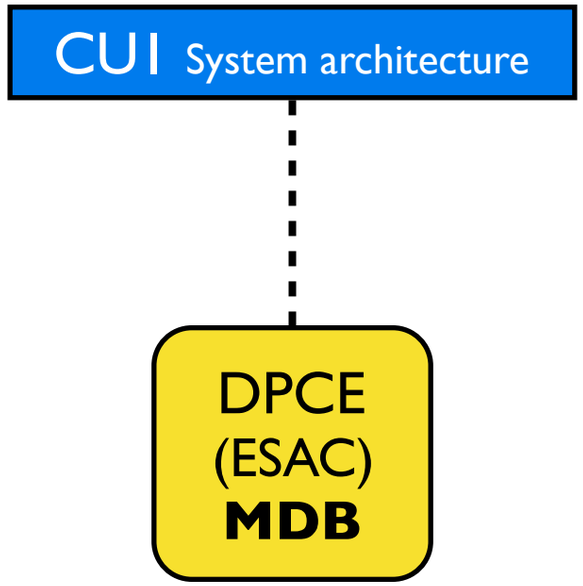
CU's process data in Data Processing centre(s):



24

Gaia Consortium - distributed challenge

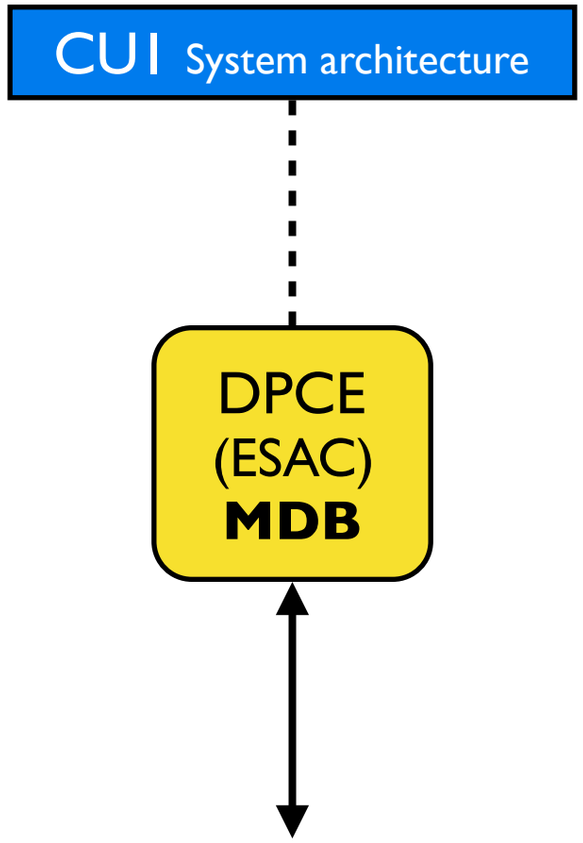
CU's process data in Data Processing centre(s):



24

Gaia Consortium - distributed challenge

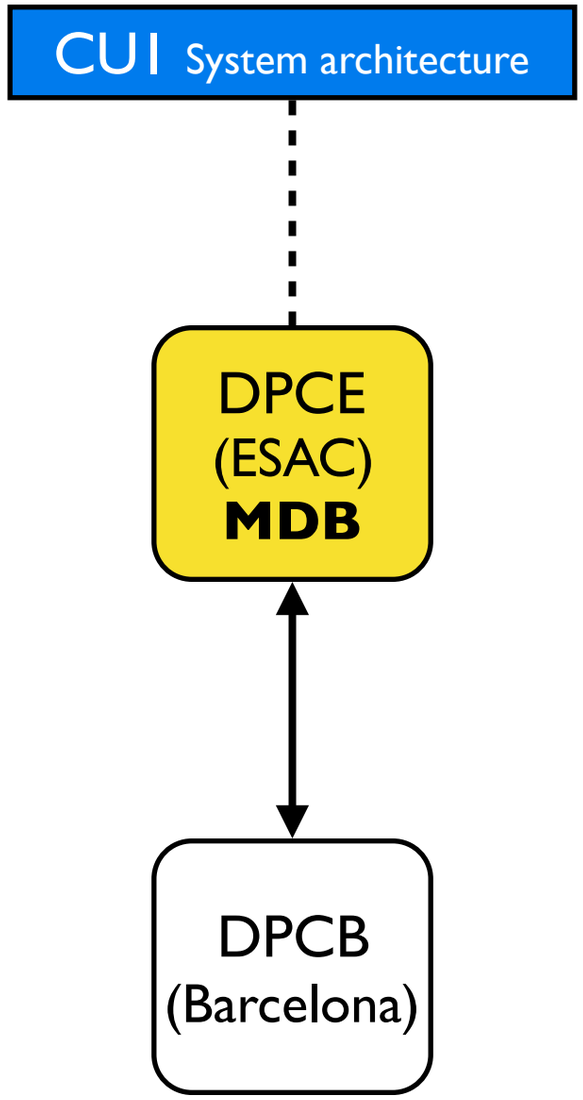
CU's process data in Data Processing centre(s):



24

Gaia Consortium - distributed challenge

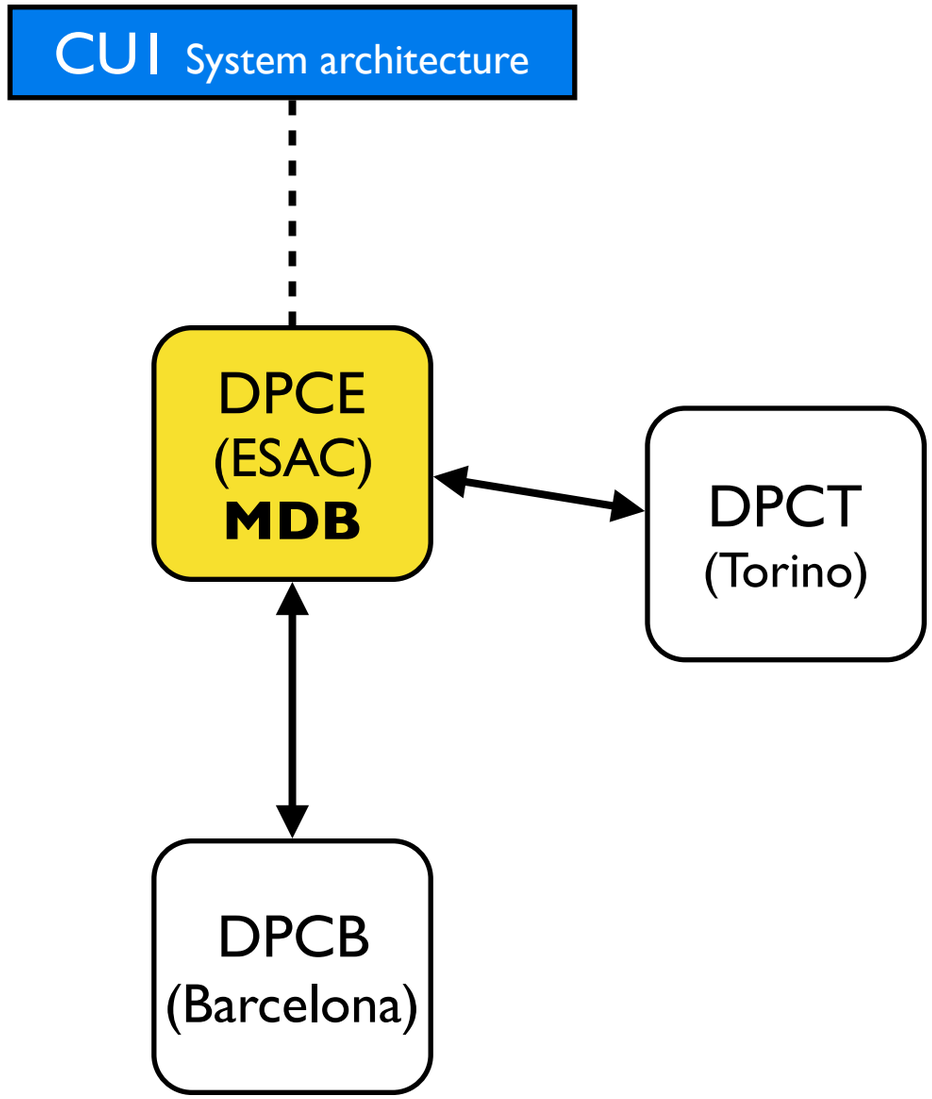
CU's process data in Data Processing centre(s):



24

Gaia Consortium - distributed challenge

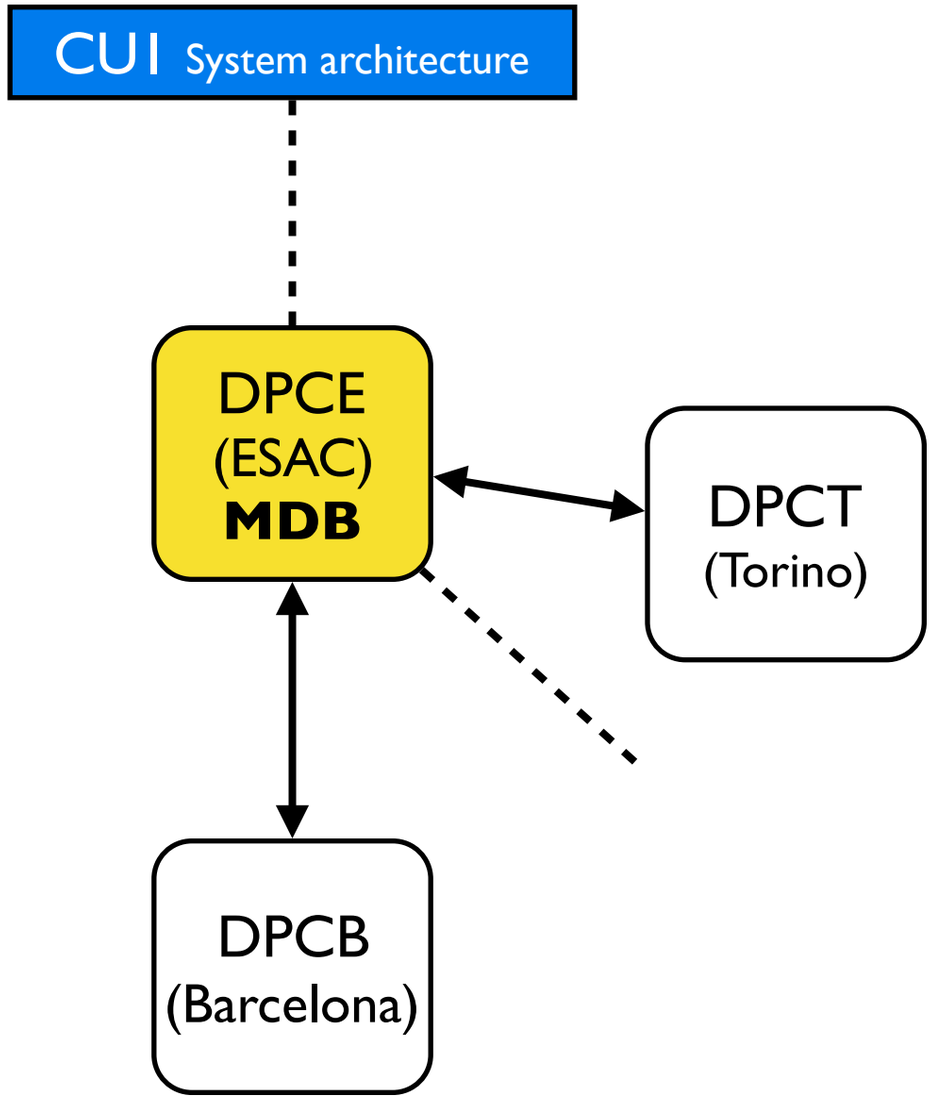
CU's process data in Data Processing centre(s):



24

Gaia Consortium - distributed challenge

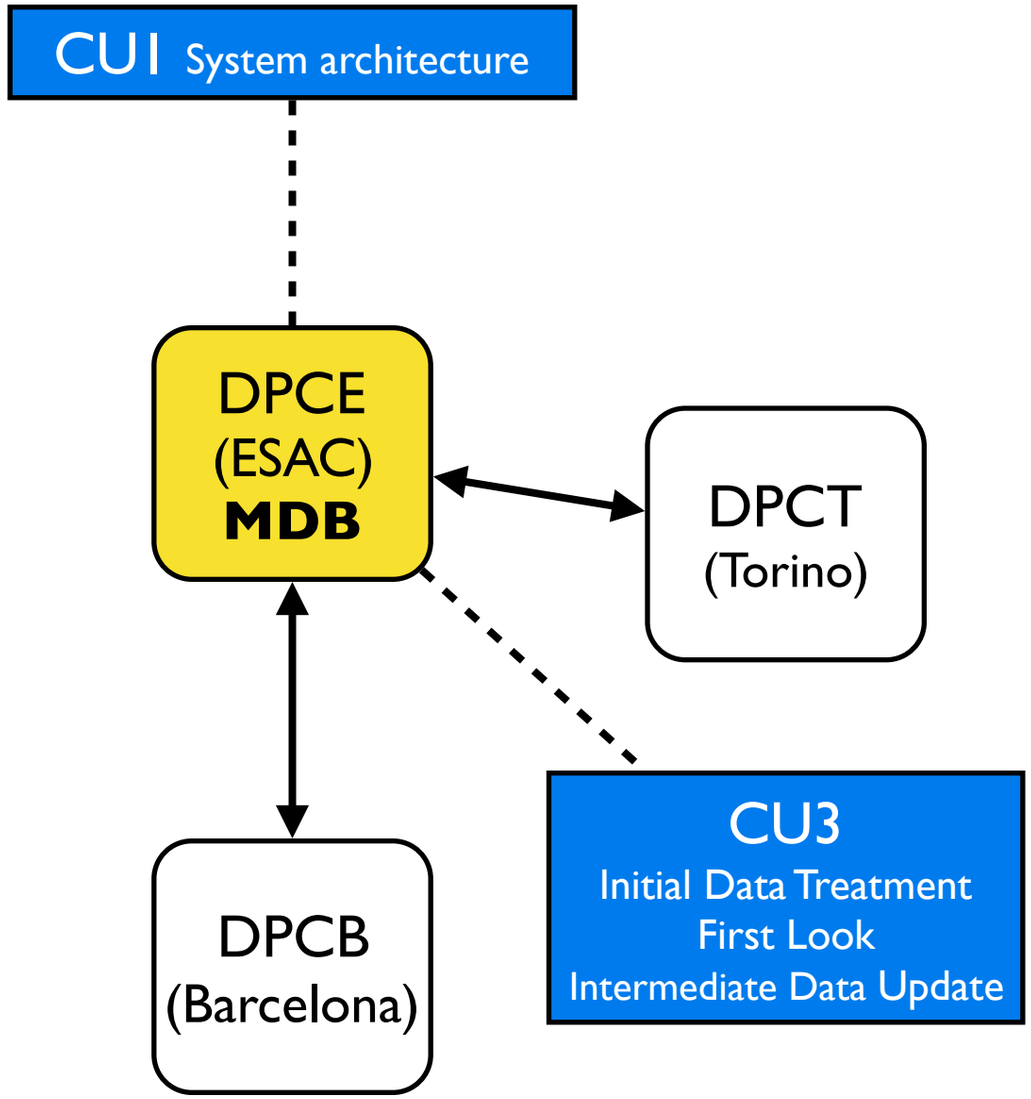
CU's process data in Data Processing centre(s):



24

Gaia Consortium - distributed challenge

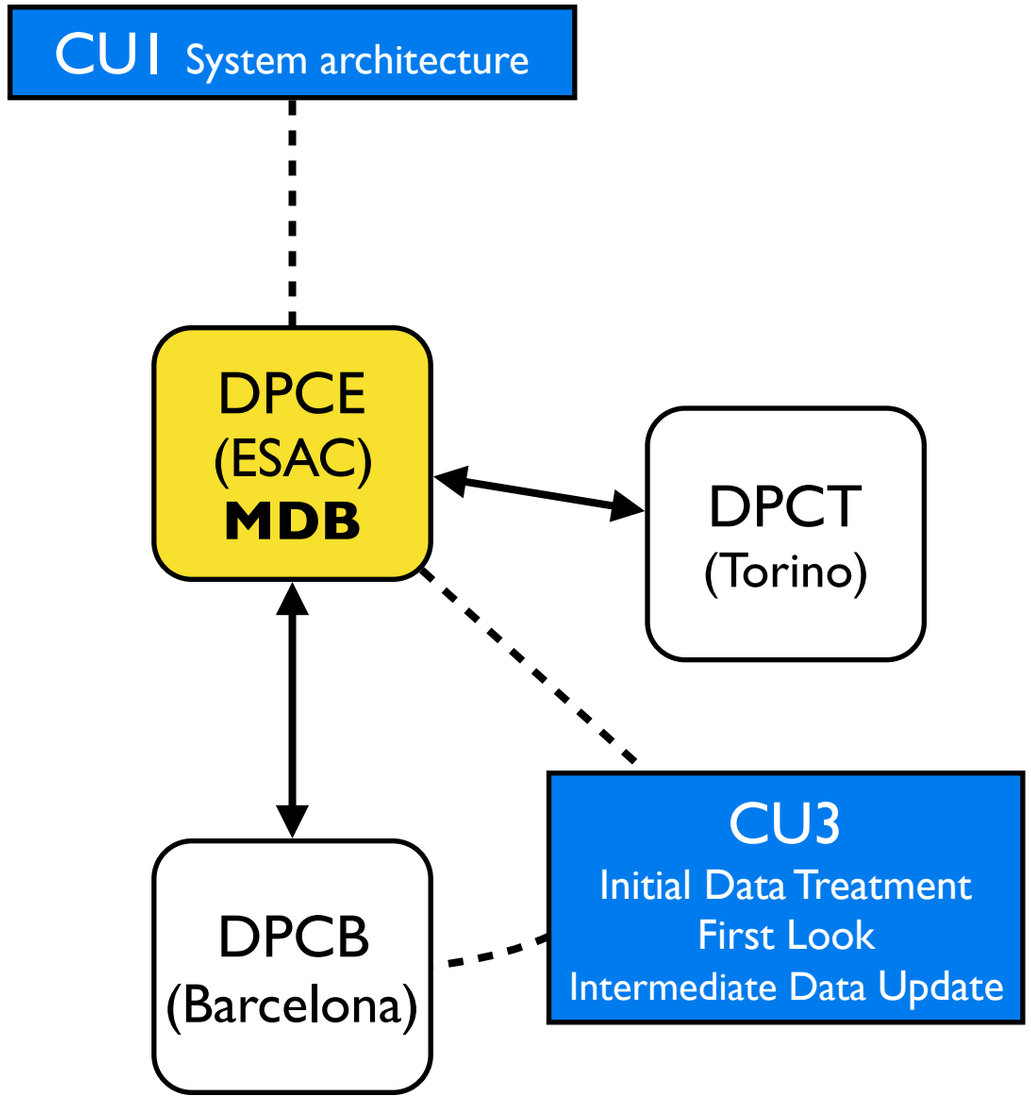
CU's process data in Data Processing centre(s):



24

Gaia Consortium - distributed challenge

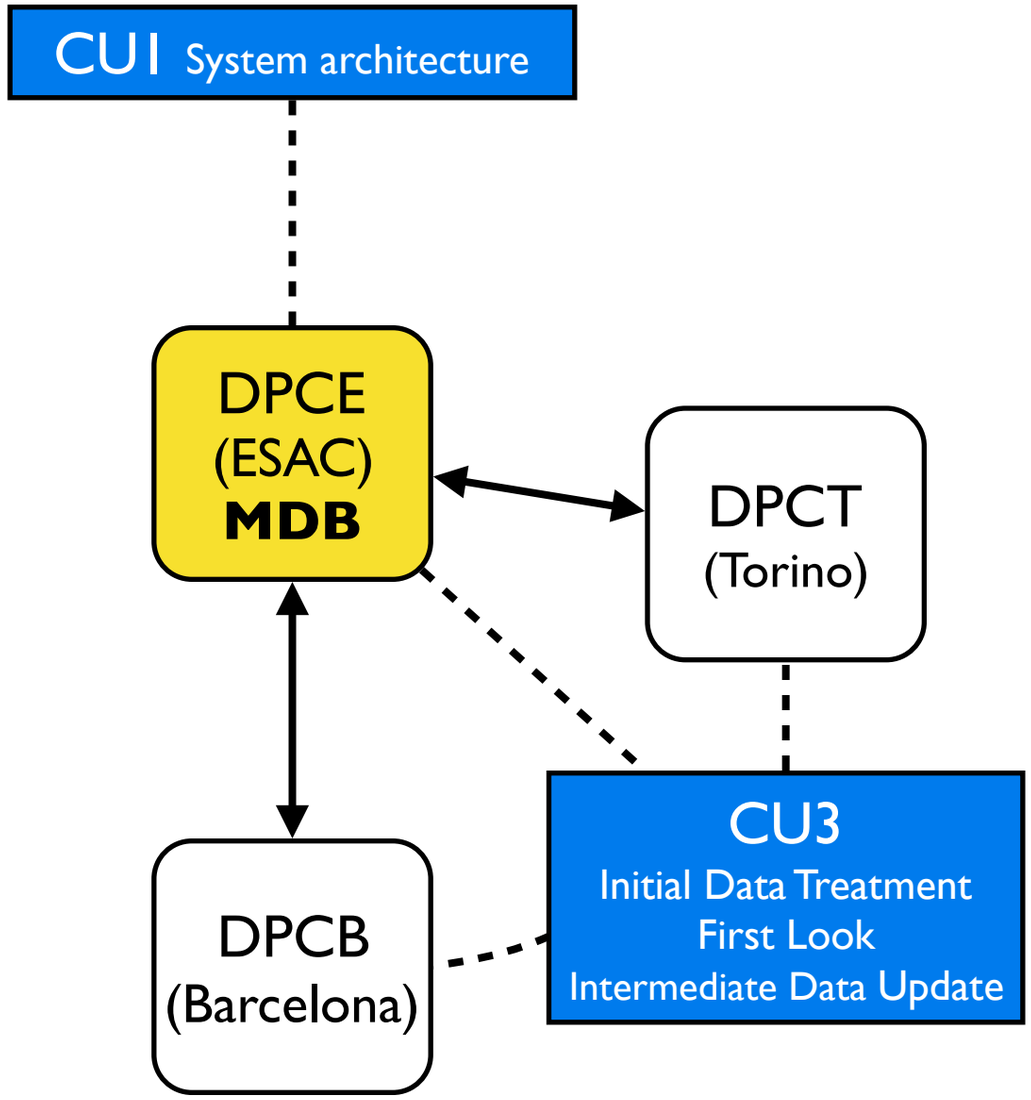
CU's process data in Data Processing centre(s):



24

Gaia Consortium - distributed challenge

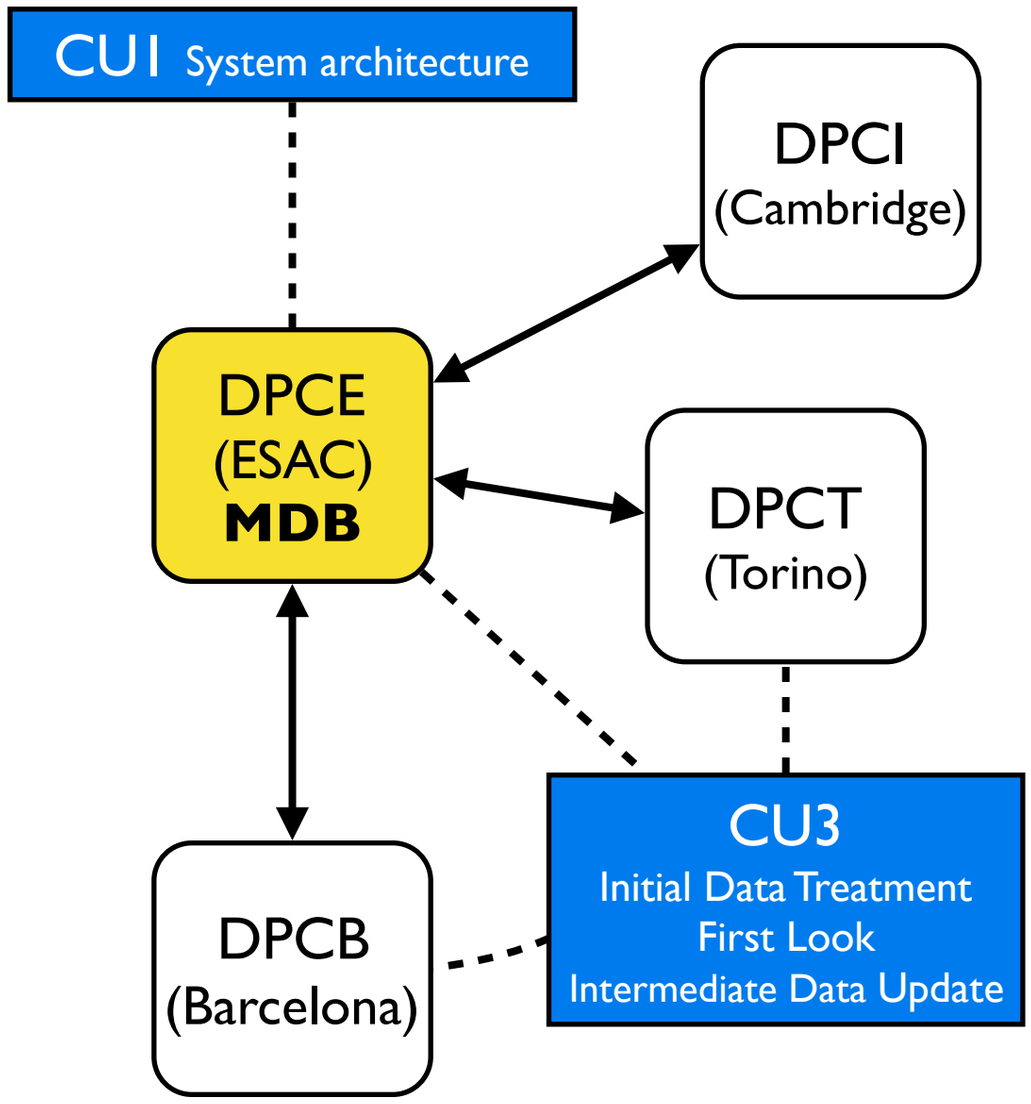
CU's process data in Data Processing centre(s):



24

Gaia Consortium - distributed challenge

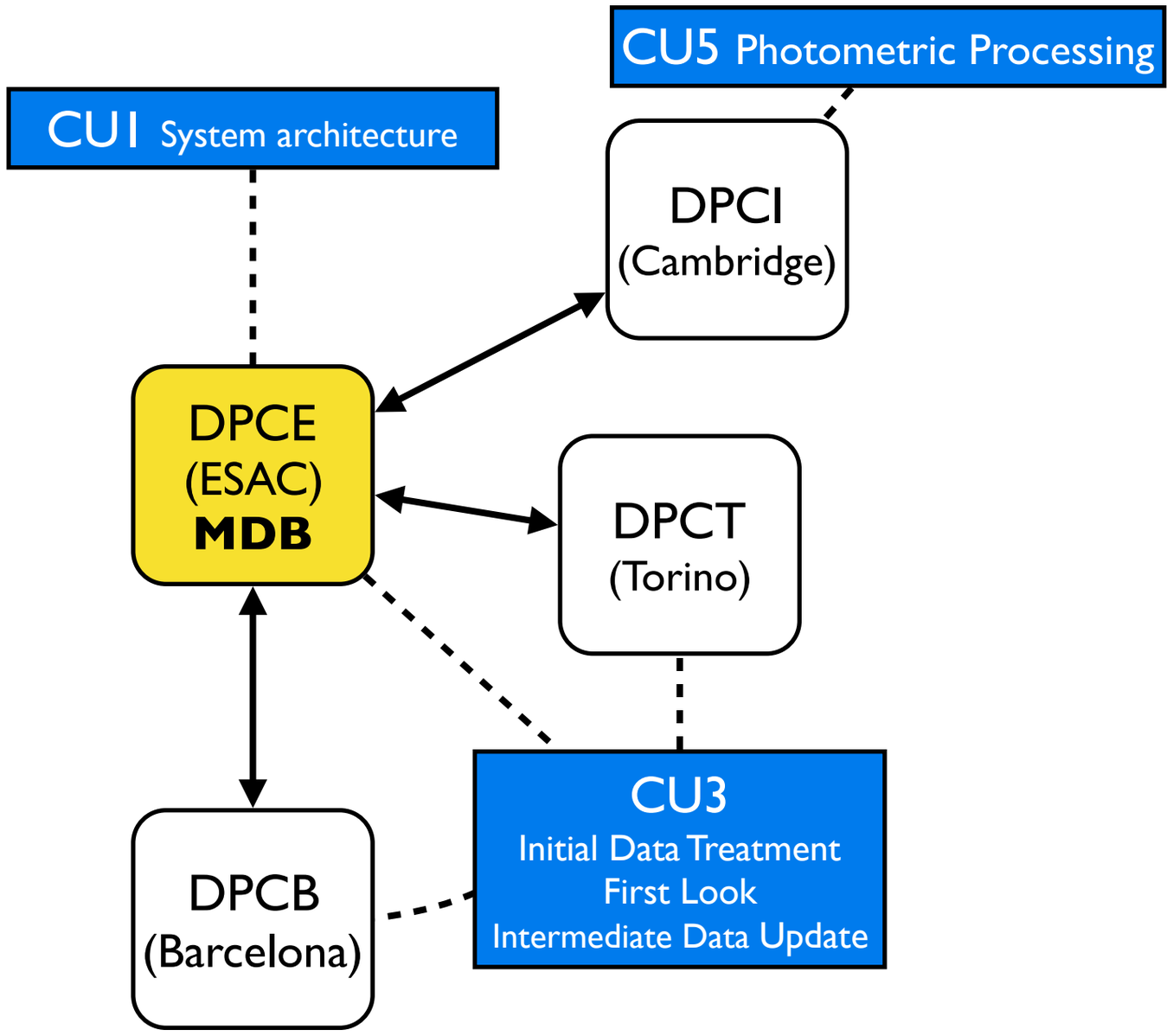
CU's process data in Data Processing centre(s):



24

Gaia Consortium - distributed challenge

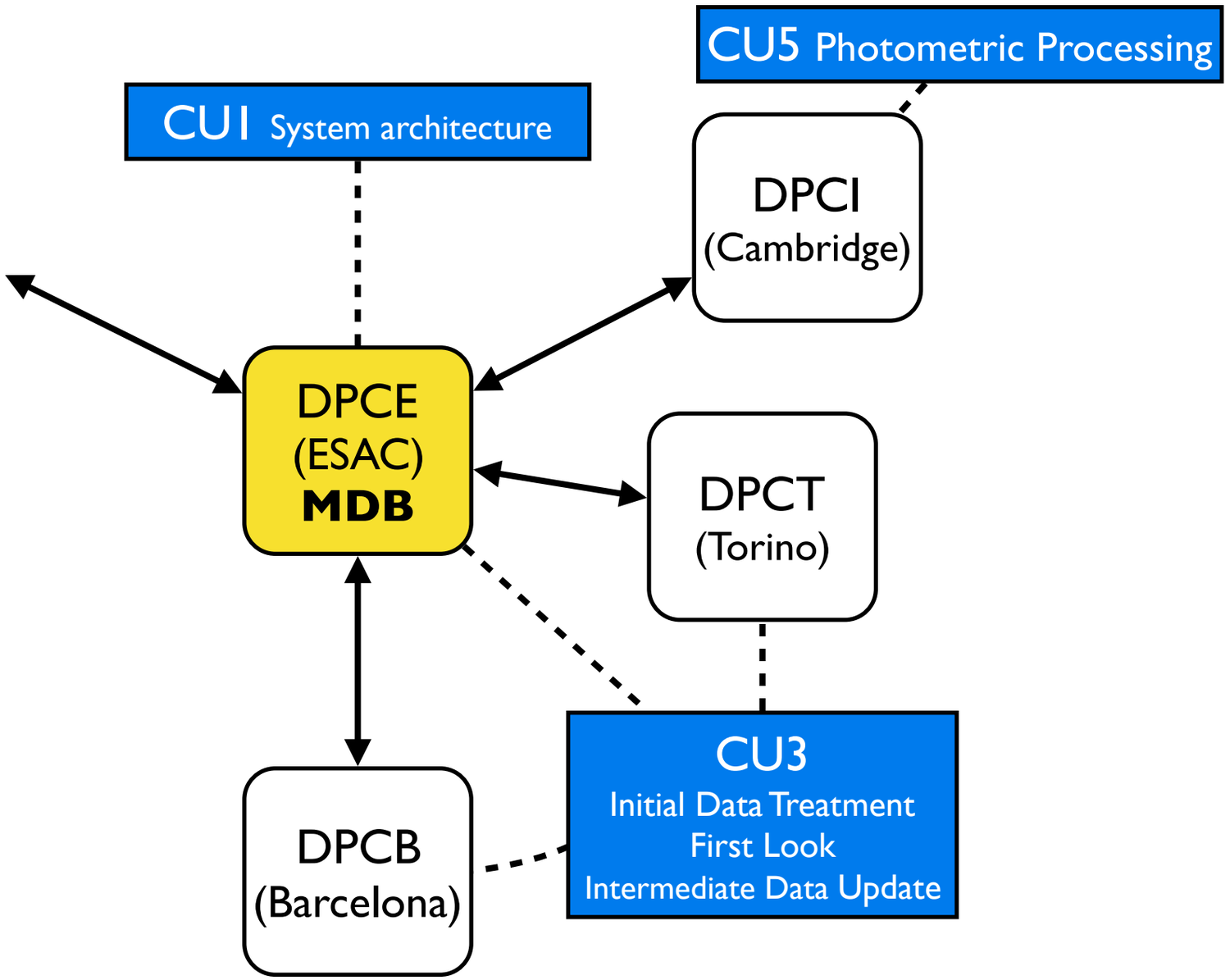
CU's process data in Data Processing centre(s):



24

Gaia Consortium - distributed challenge

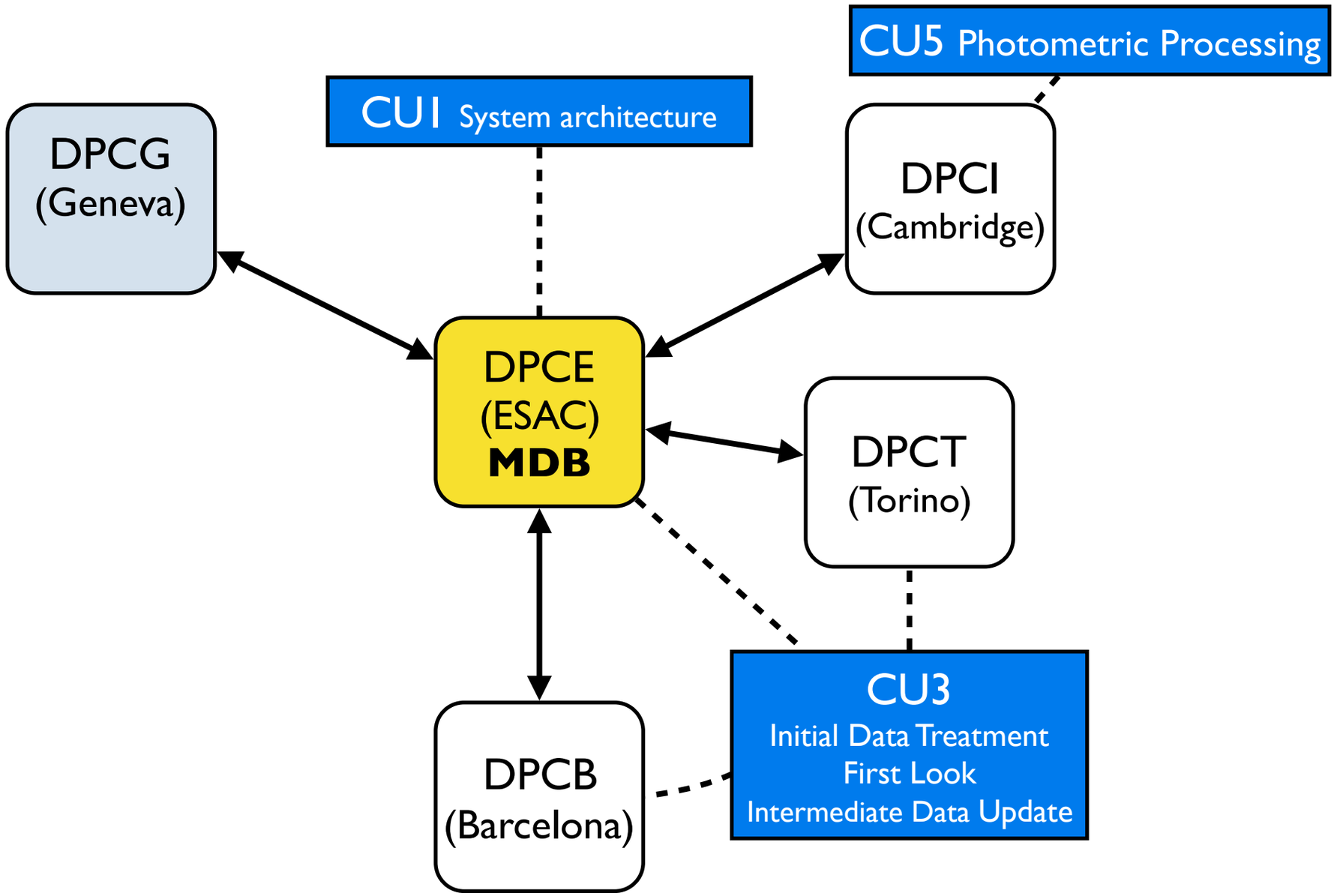
CU's process data in Data Processing centre(s):



24

Gaia Consortium - distributed challenge

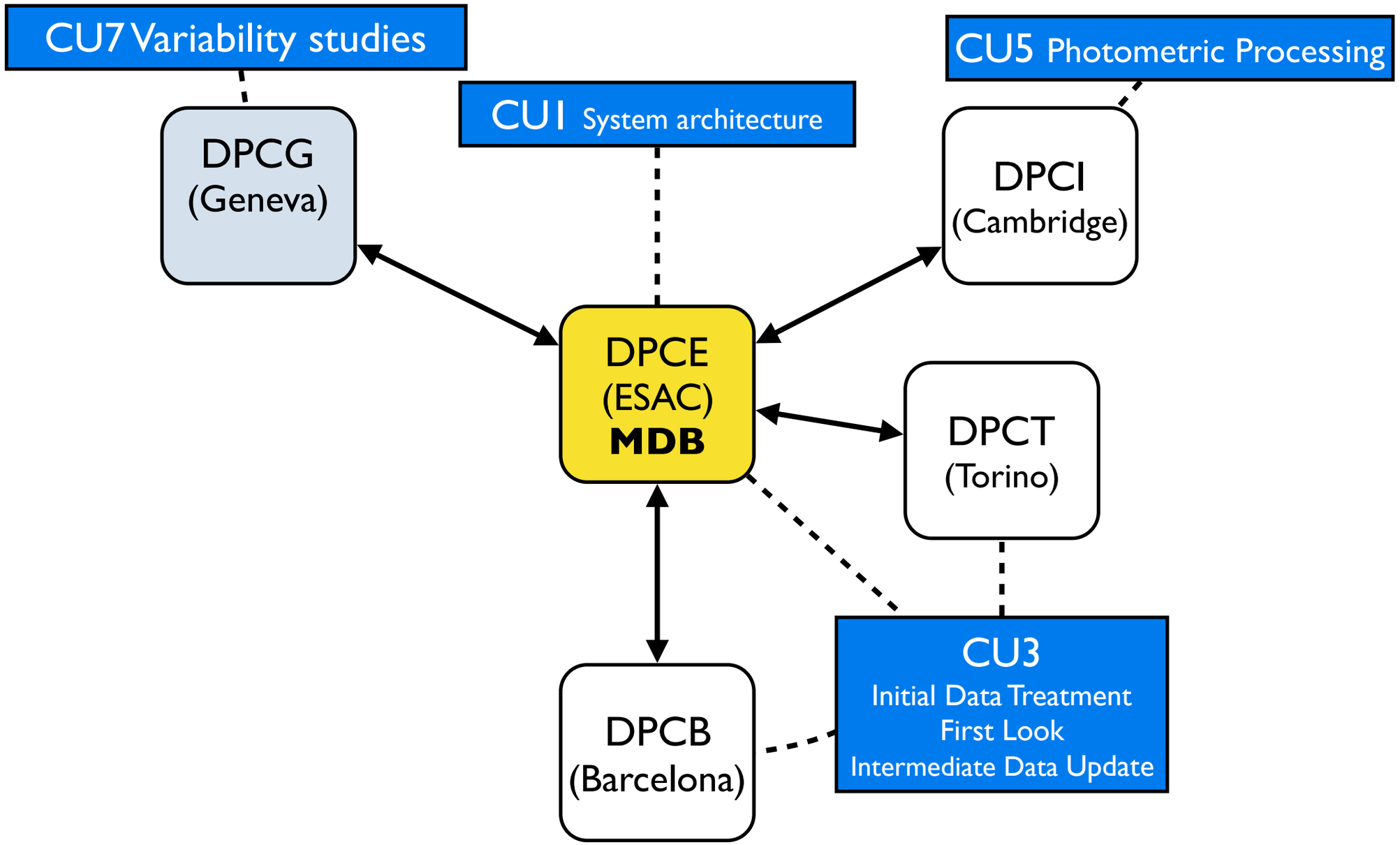
CU's process data in Data Processing centre(s):



24

Gaia Consortium - distributed challenge

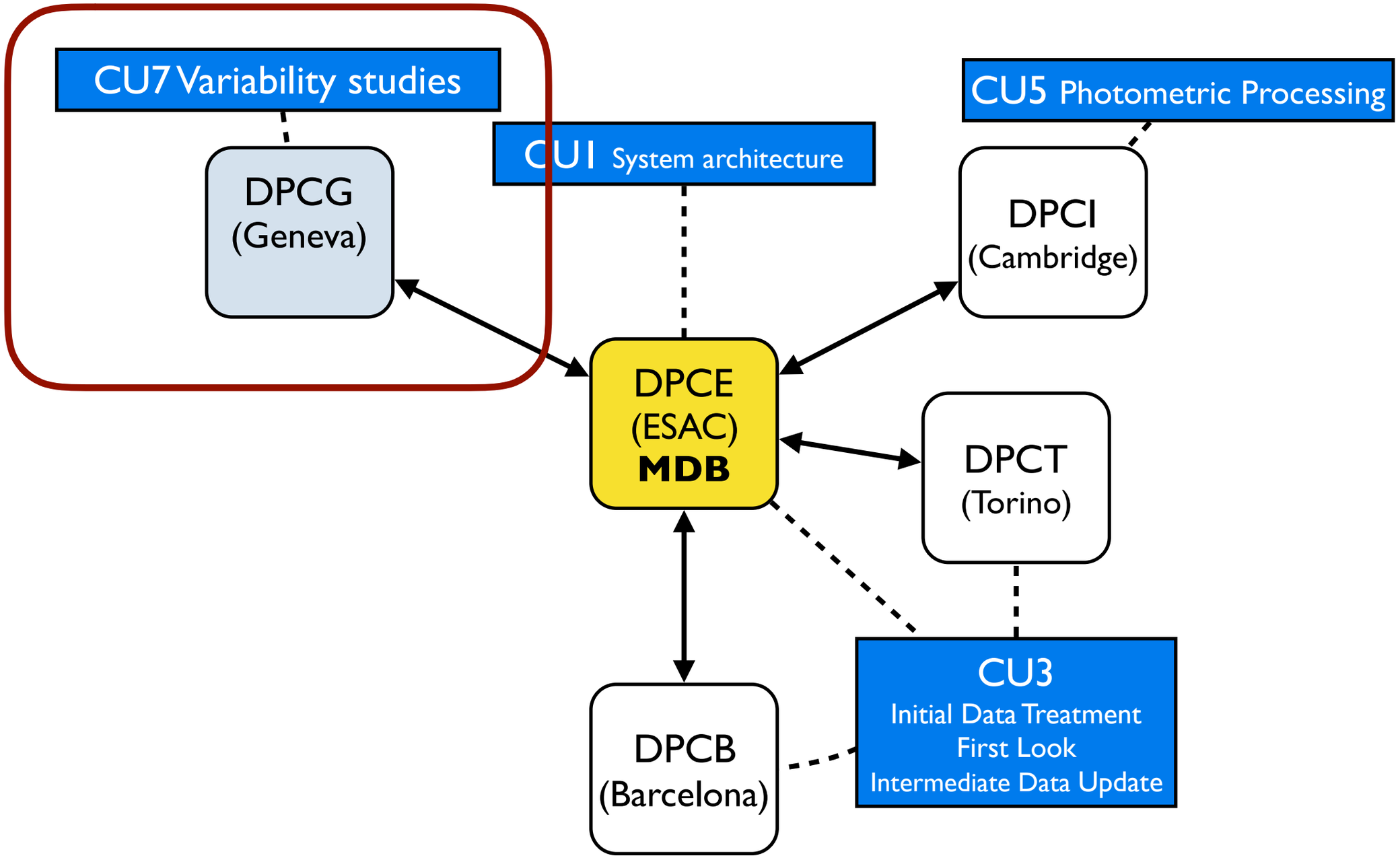
CU's process data in Data Processing centre(s):



24

Gaia Consortium - distributed challenge

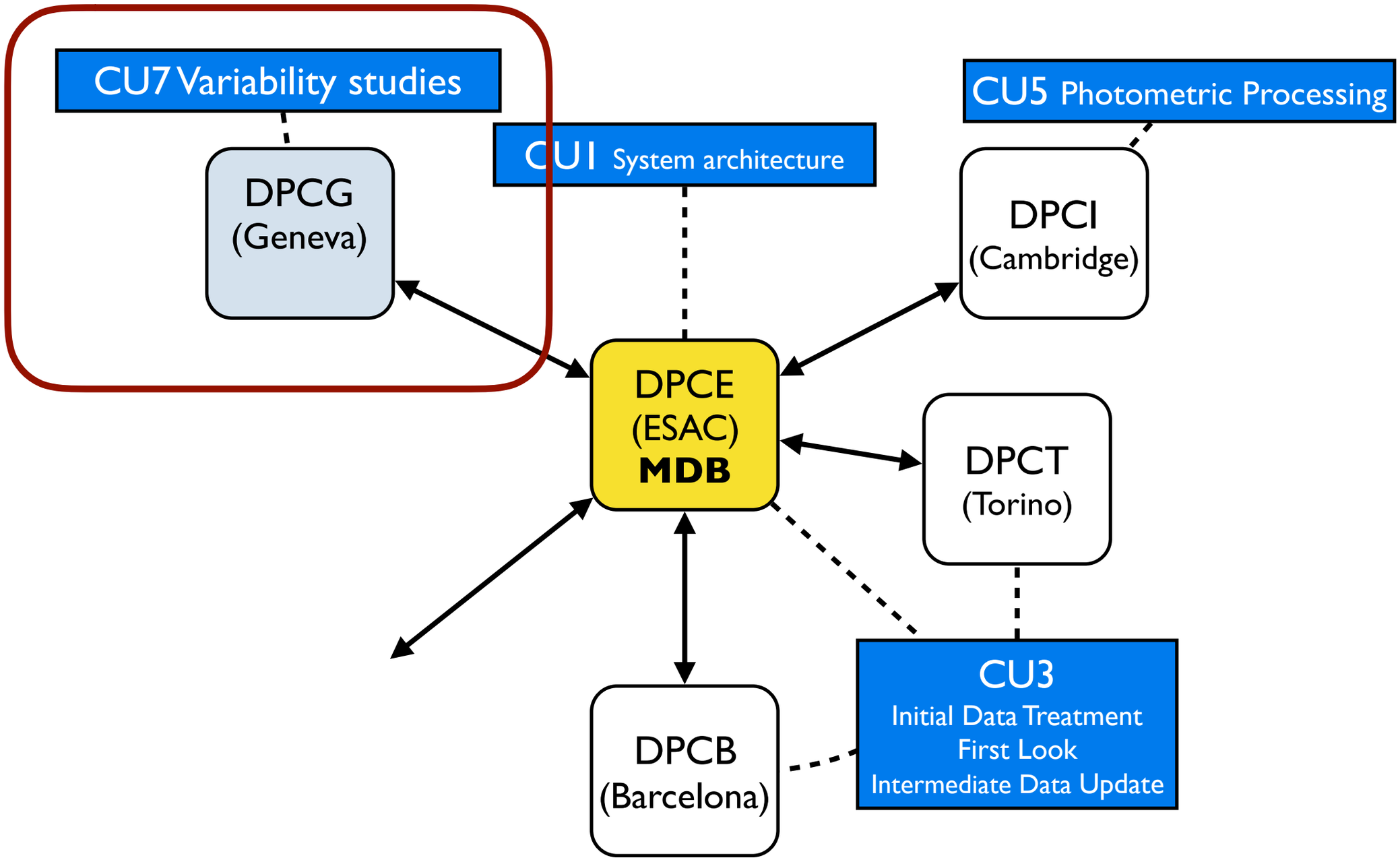
CU's process data in Data Processing centre(s):



24

Gaia Consortium - distributed challenge

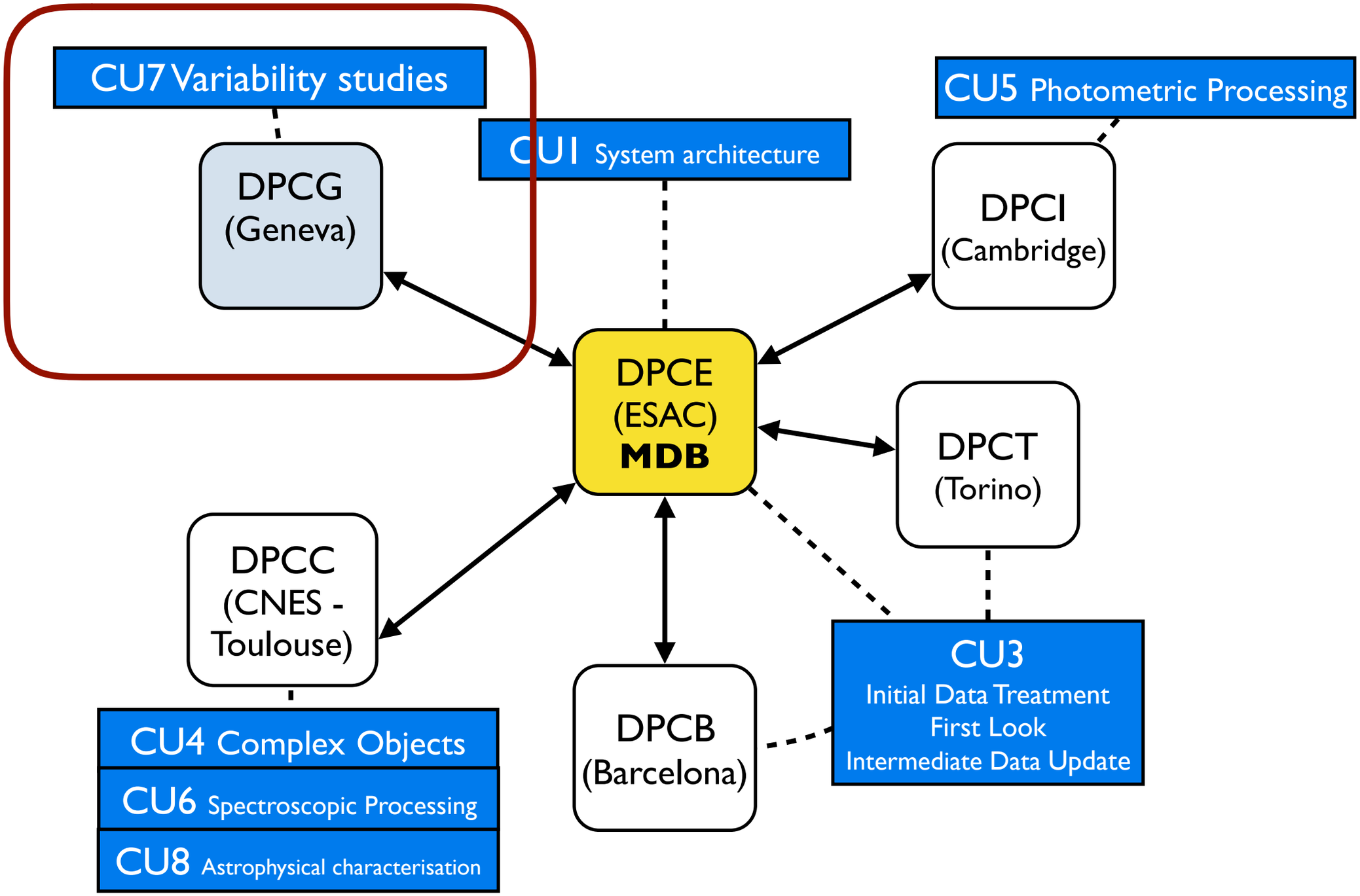
CU's process data in Data Processing centre(s):



24

Gaia Consortium - distributed challenge

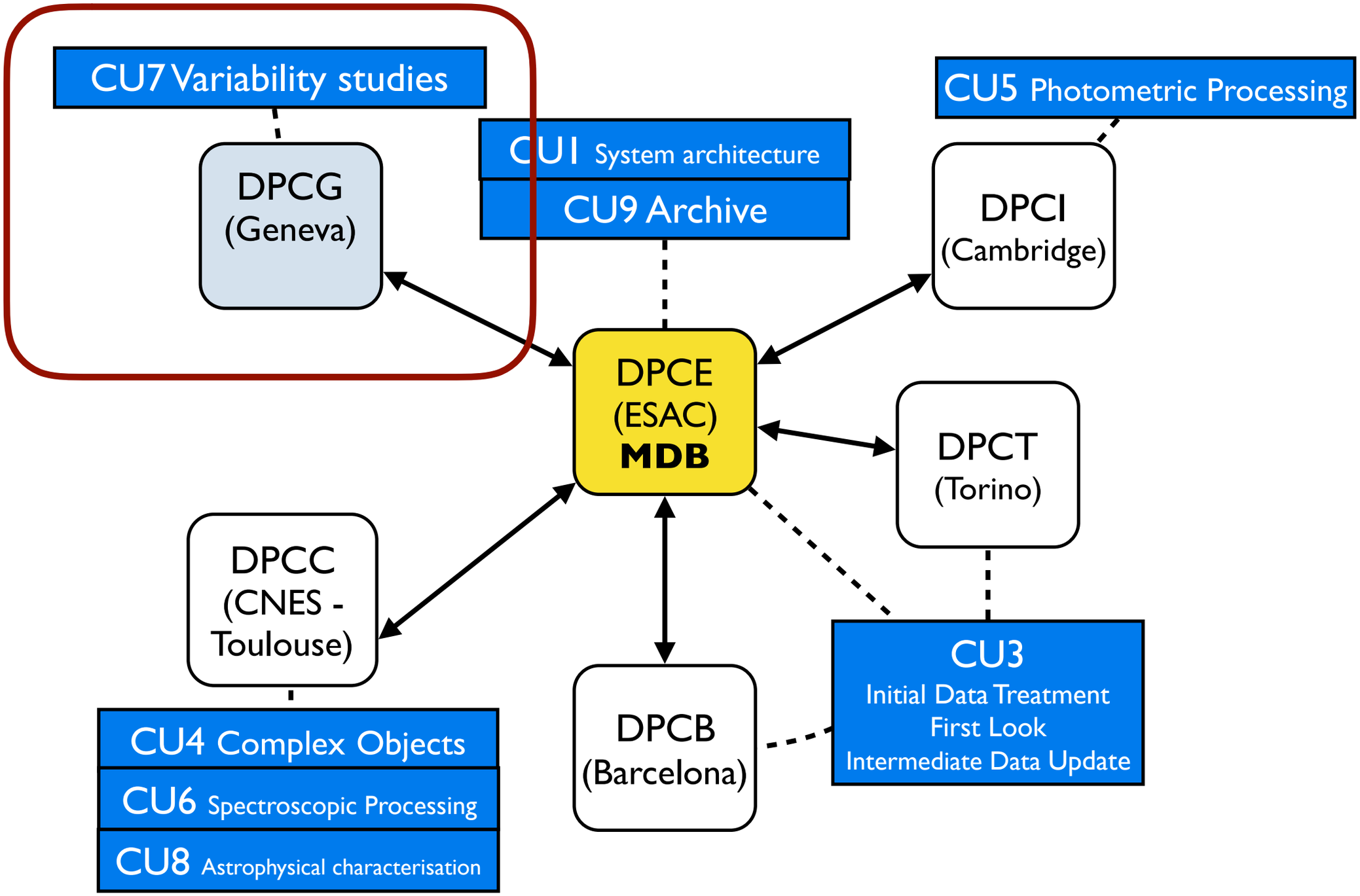
CU's process data in Data Processing centre(s):



24

Gaia Consortium - distributed challenge

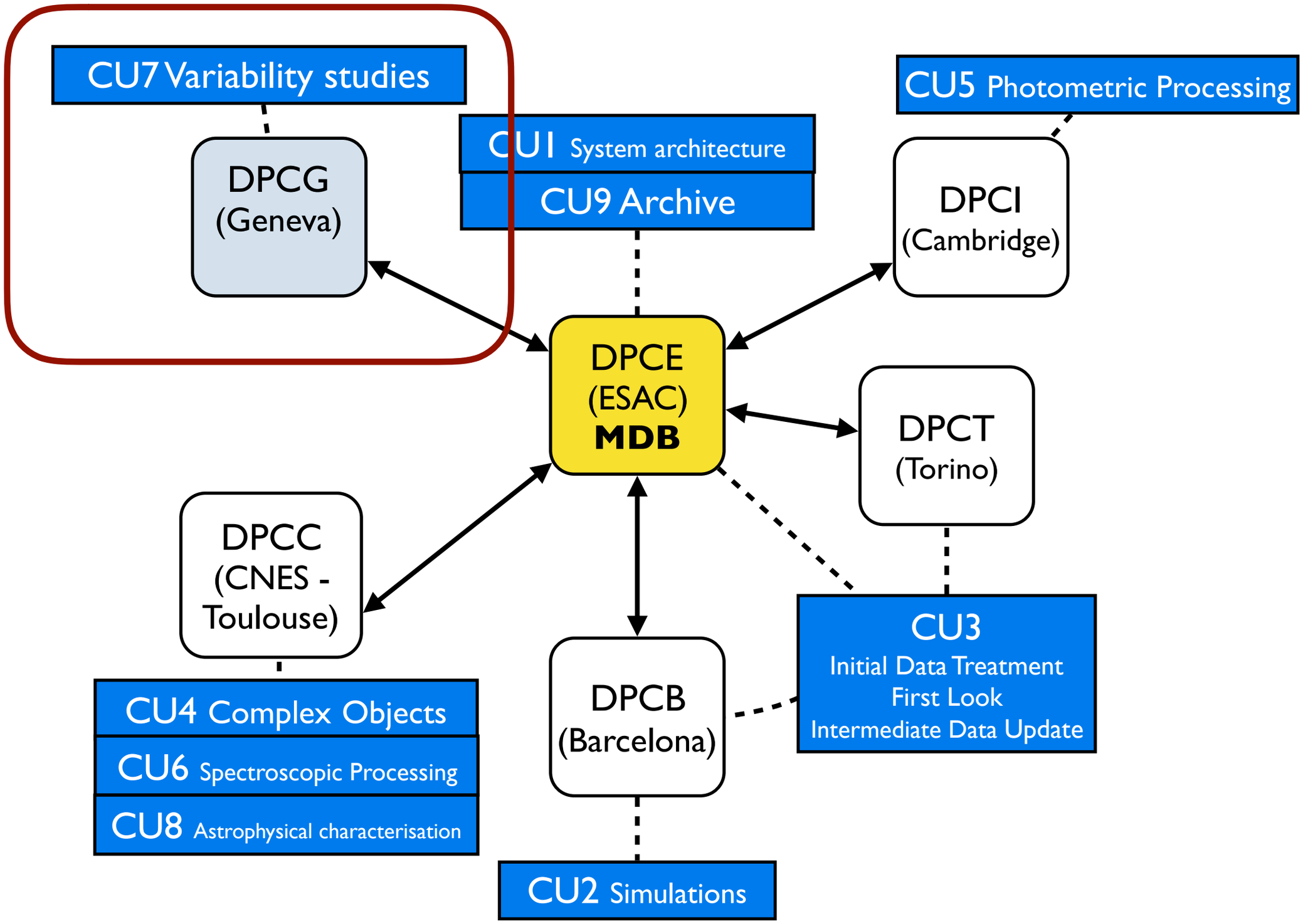
CU's process data in Data Processing centre(s):



24

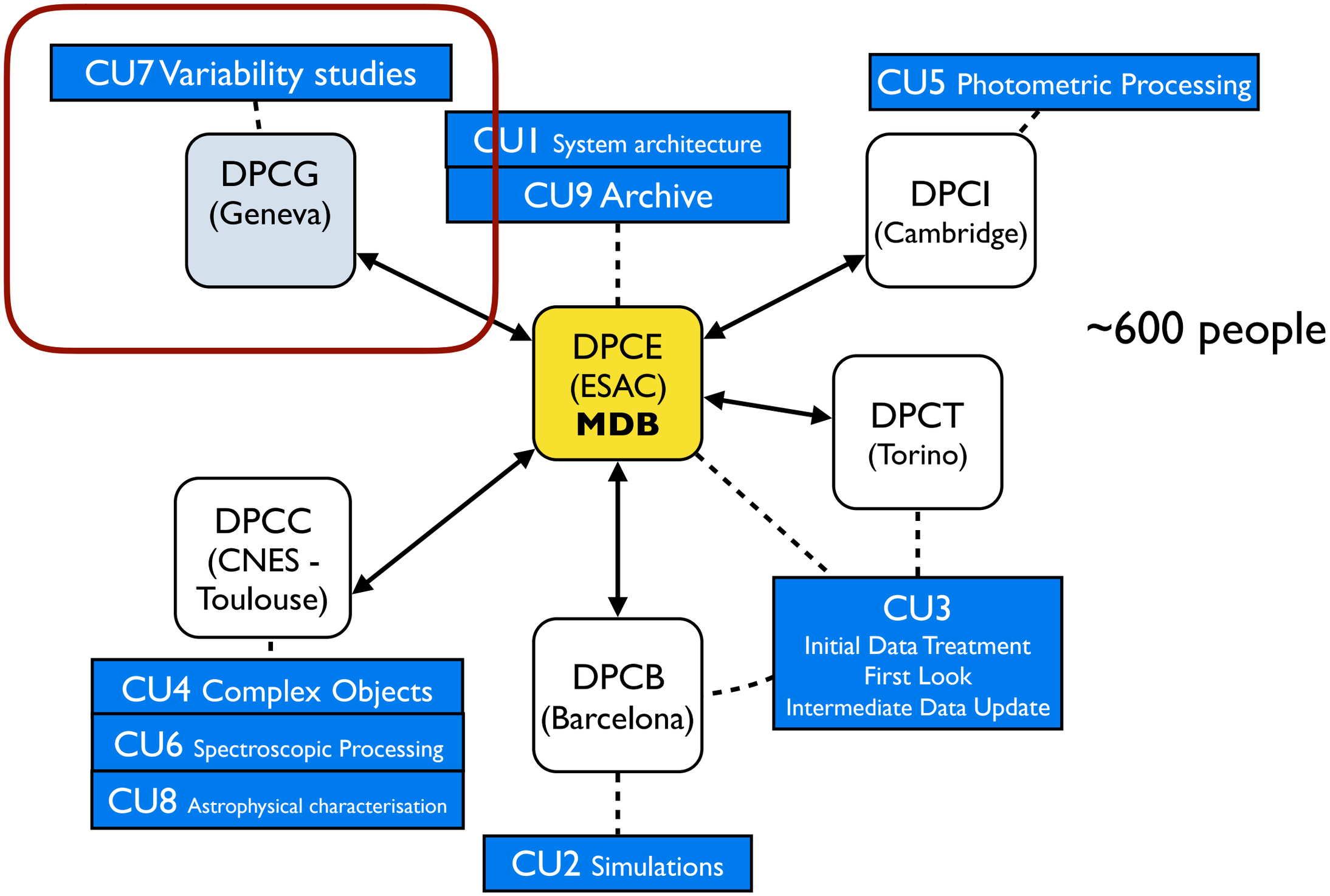
Gaia Consortium - distributed challenge

CU's process data in Data Processing centre(s):



Gaia Consortium - distributed challenge

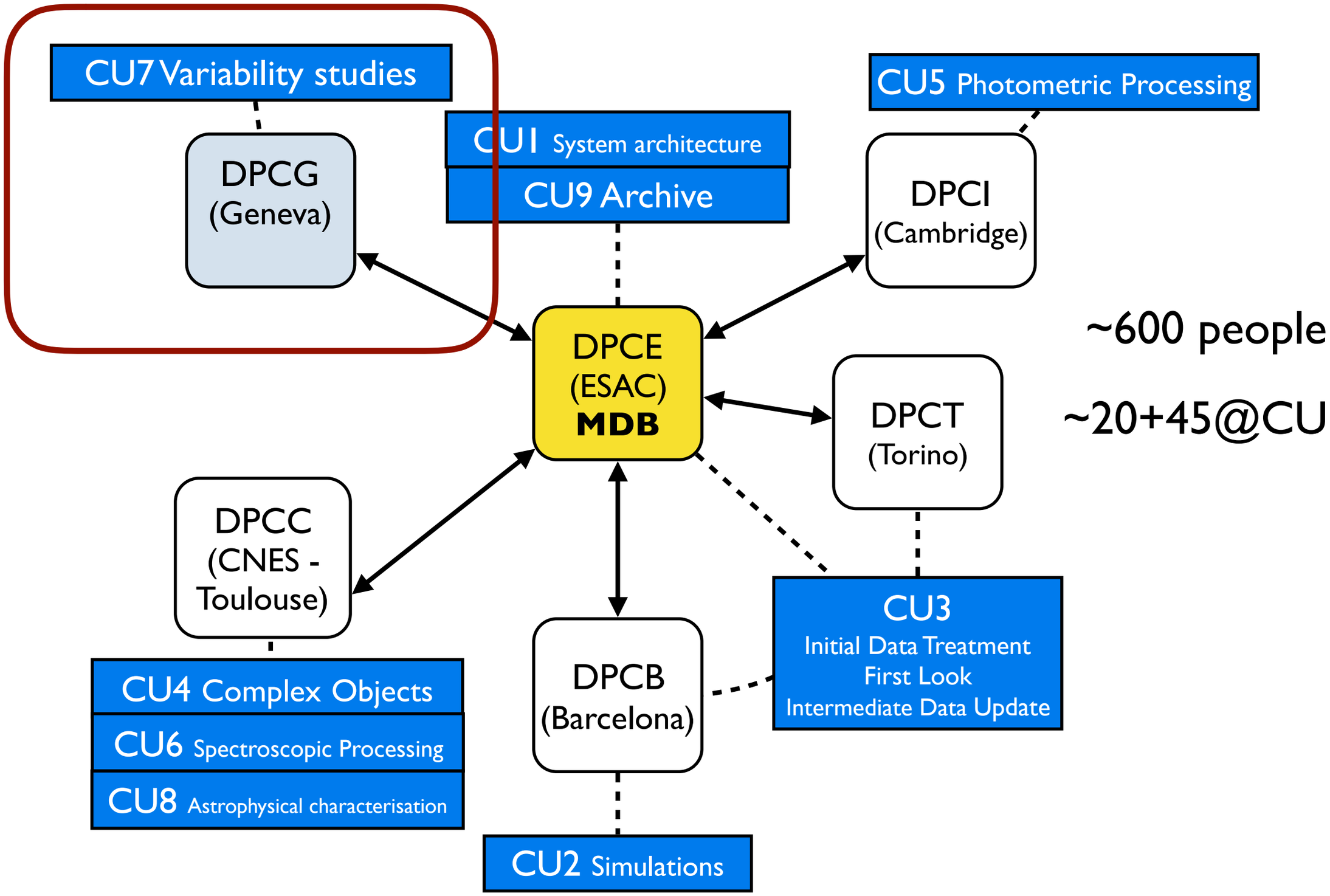
CU's process data in Data Processing centre(s):



~600 people

Gaia Consortium - distributed challenge

CU's process data in Data Processing centre(s):



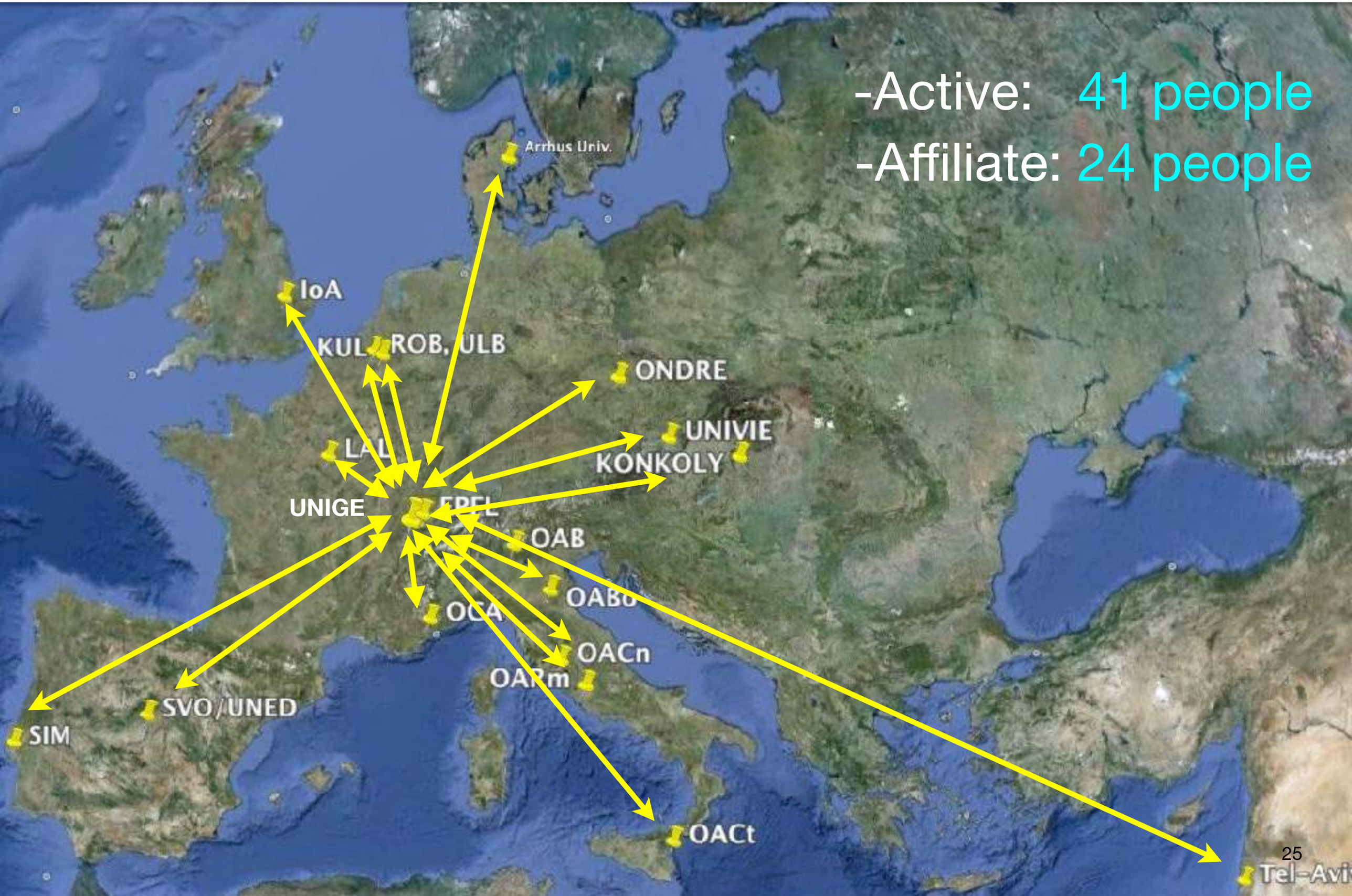
~600 people

~20+45@CU7

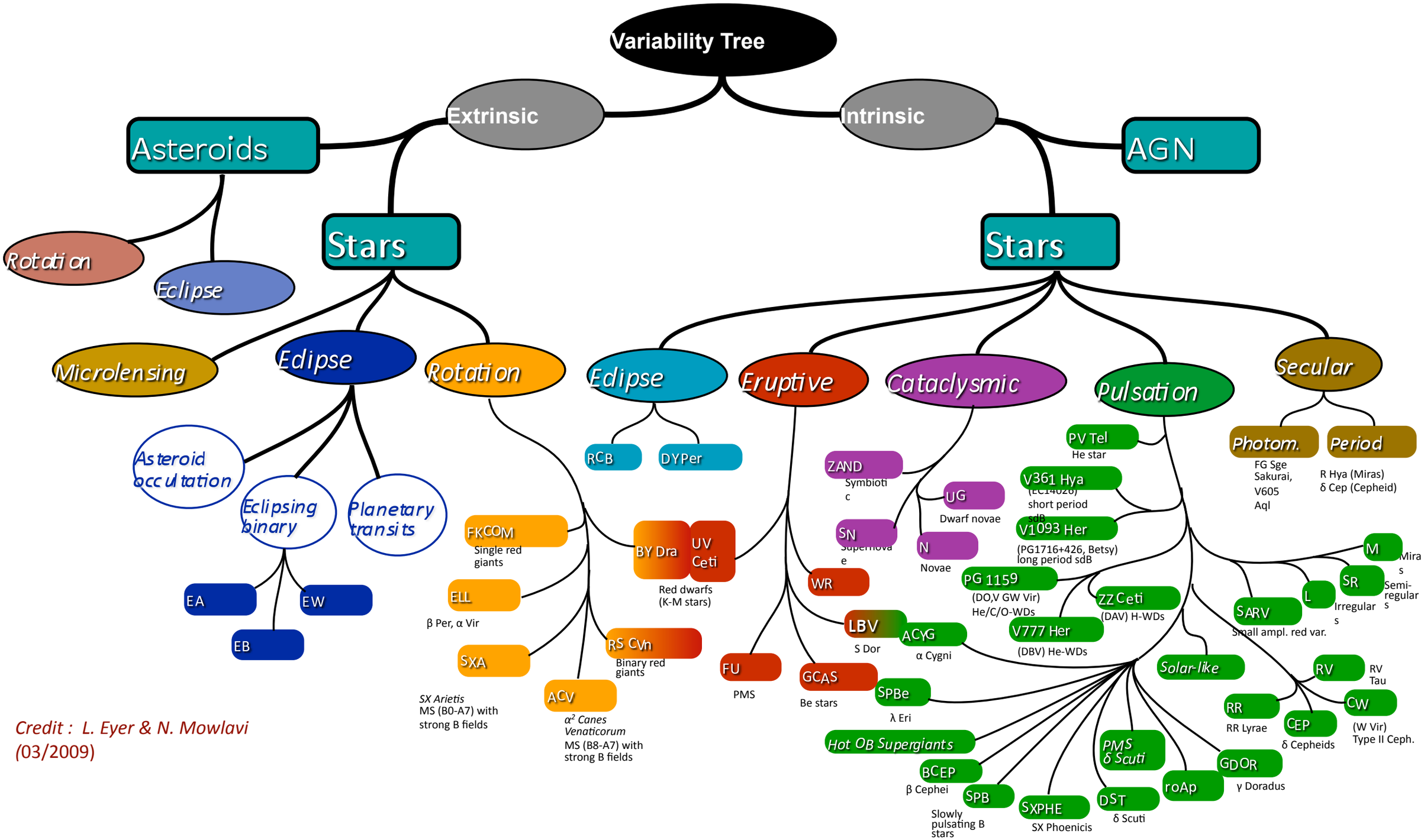
CU7 - distributed challenge

CU7 - distributed challenge

CU7 - distributed challenge

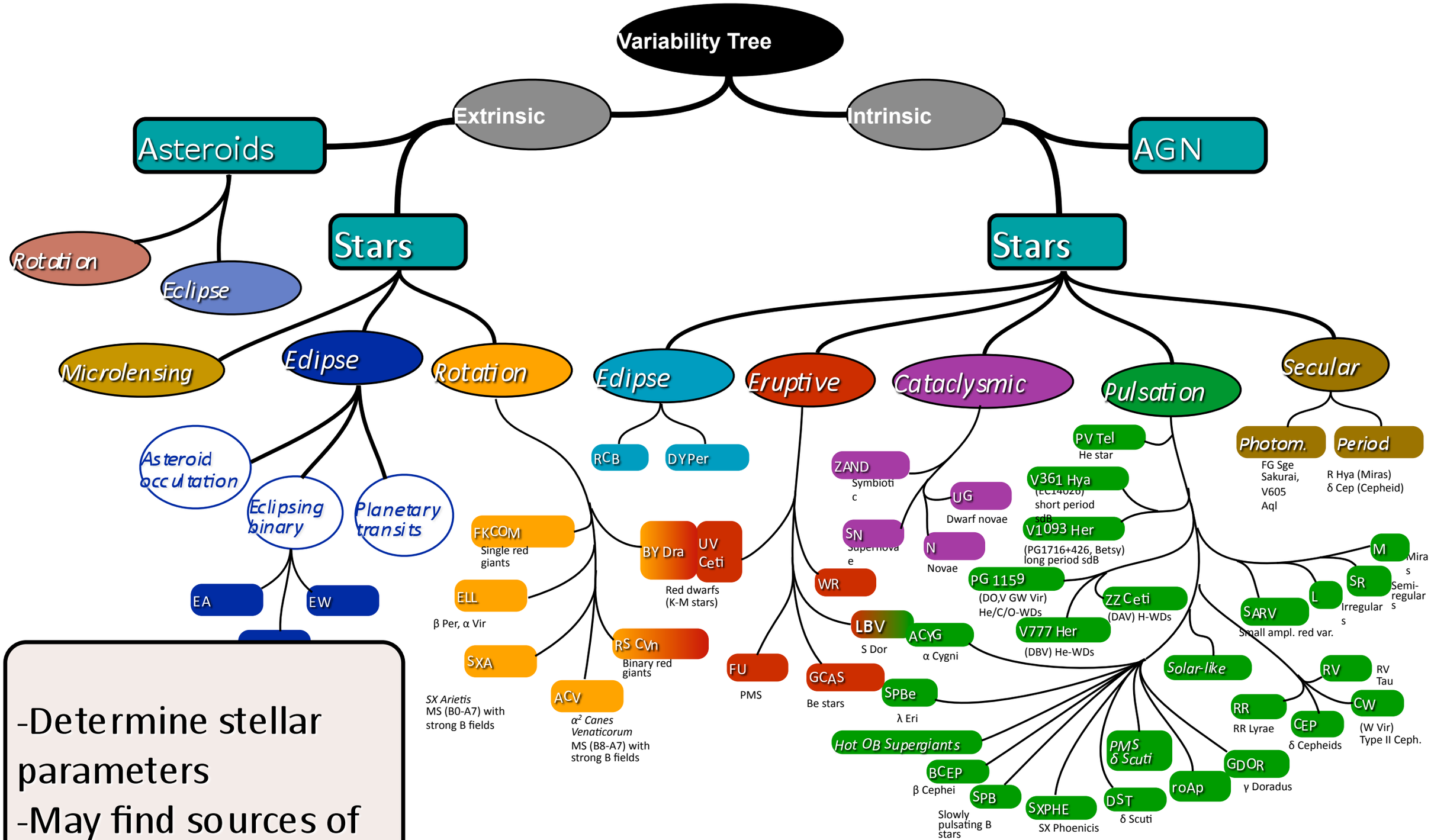


Gaia@Geneva- scientific challenge



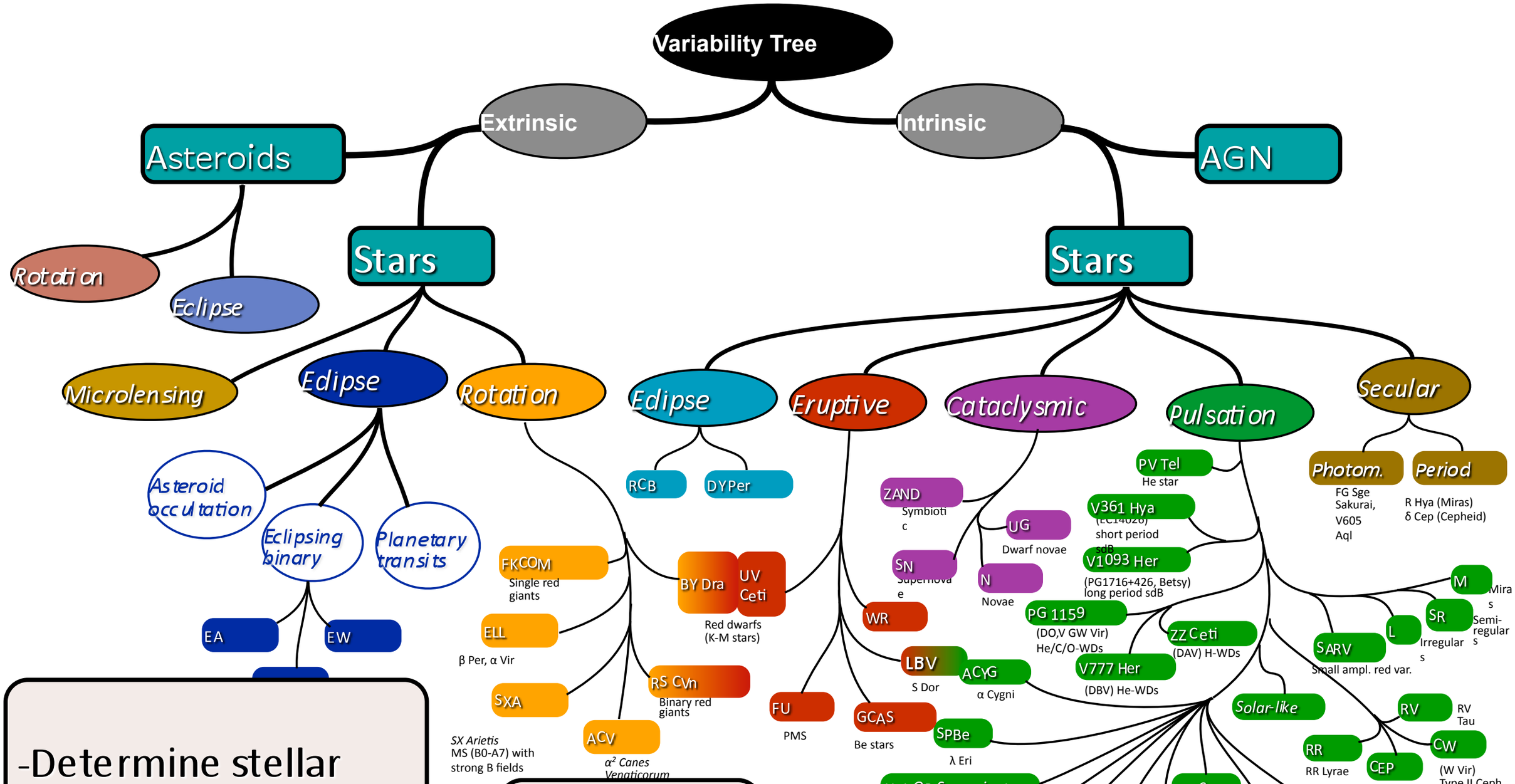
Credit : L. Eyer & N. Mowlavi (03/2009)

Gaia@Geneva- scientific challenge



-Determine stellar parameters
-May find sources of gravitational waves

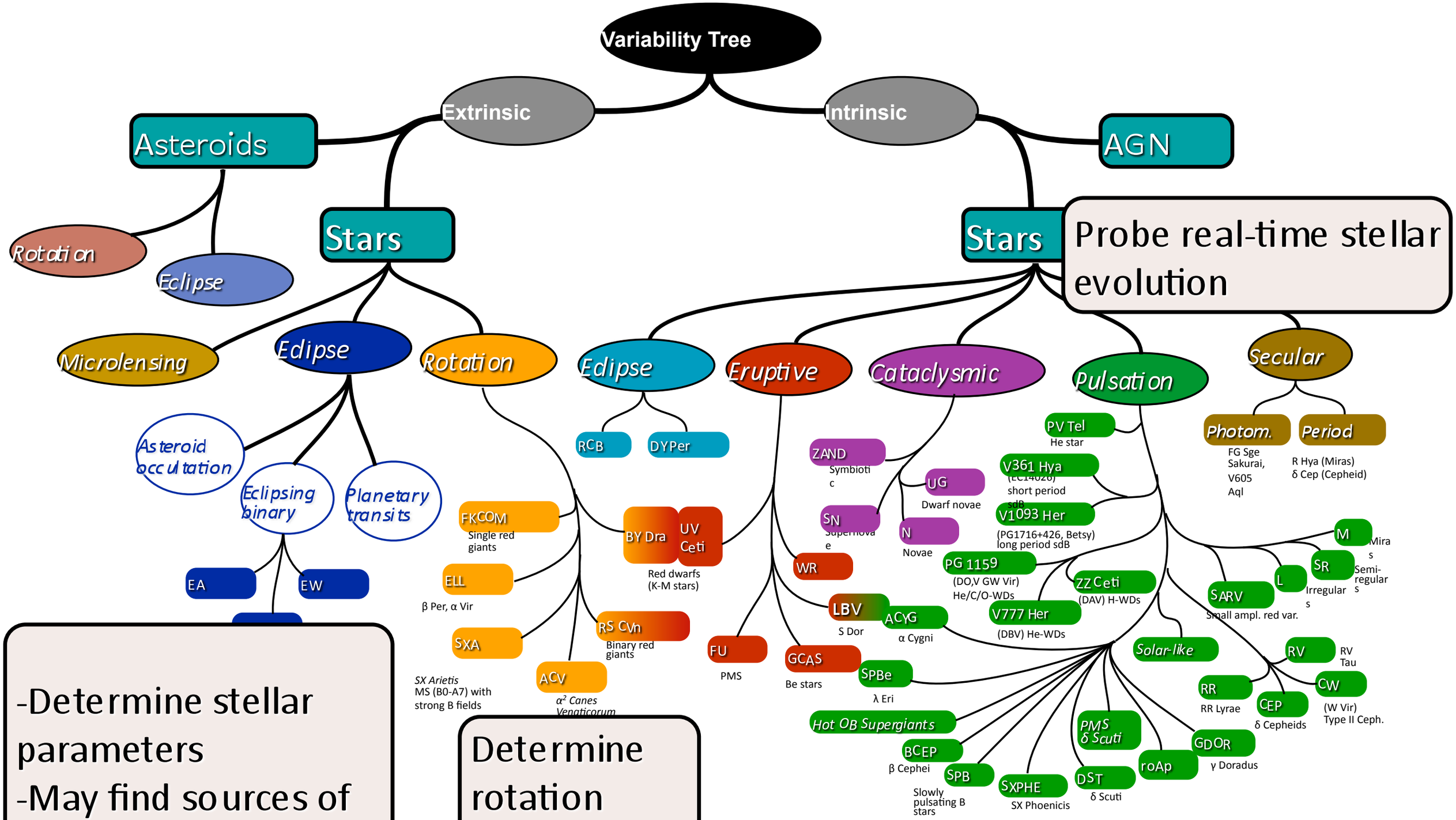
Gaia@Geneva- scientific challenge



-Determine stellar parameters
-May find sources of gravitational waves

Determine rotation periods

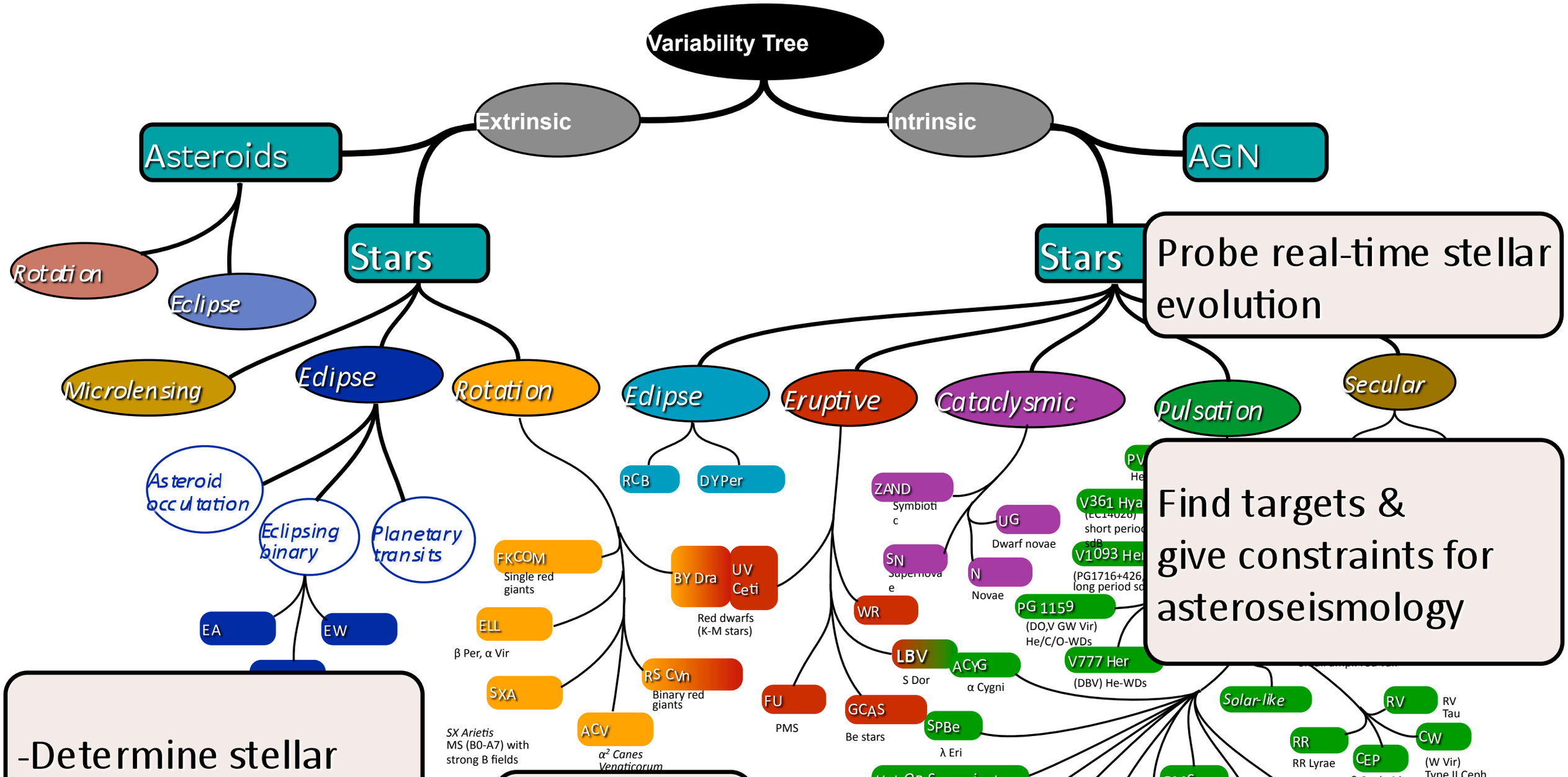
Gaia@Geneva- scientific challenge



-Determine stellar parameters
-May find sources of gravitational waves

Determine rotation periods

Gaia@Geneva- scientific challenge

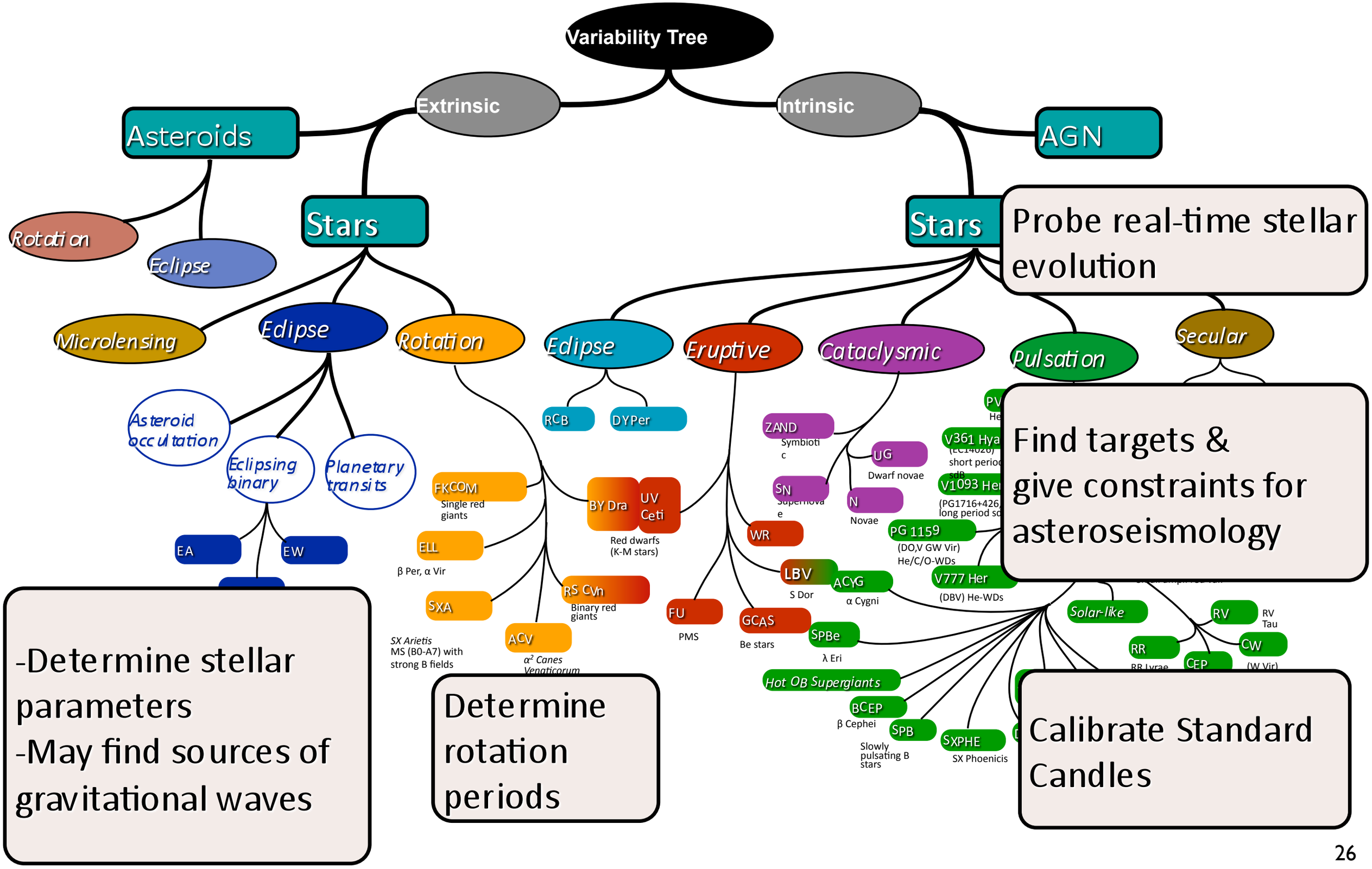


Find targets & give constraints for asteroseismology

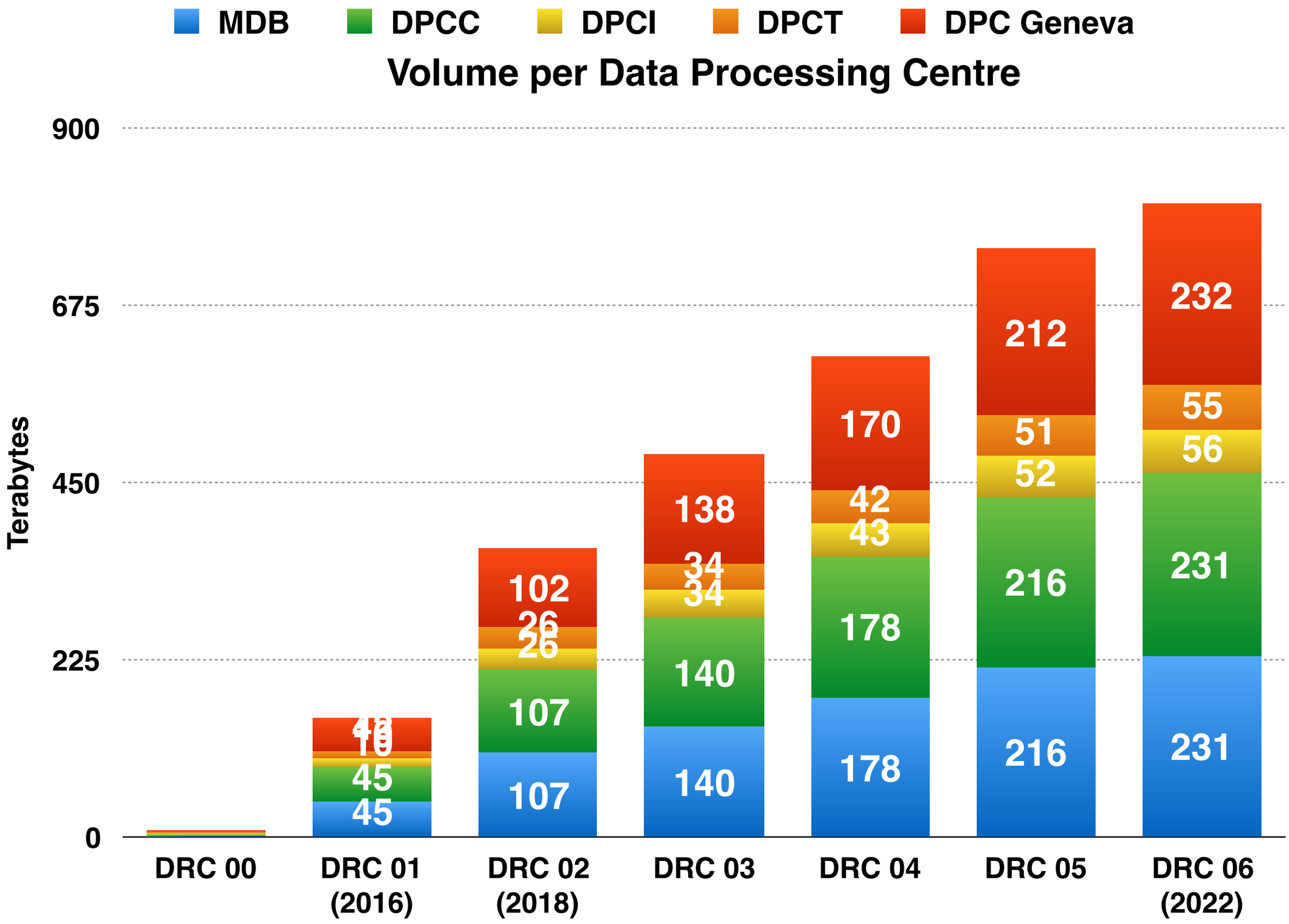
-Determine stellar parameters
-May find sources of gravitational waves

Determine rotation periods

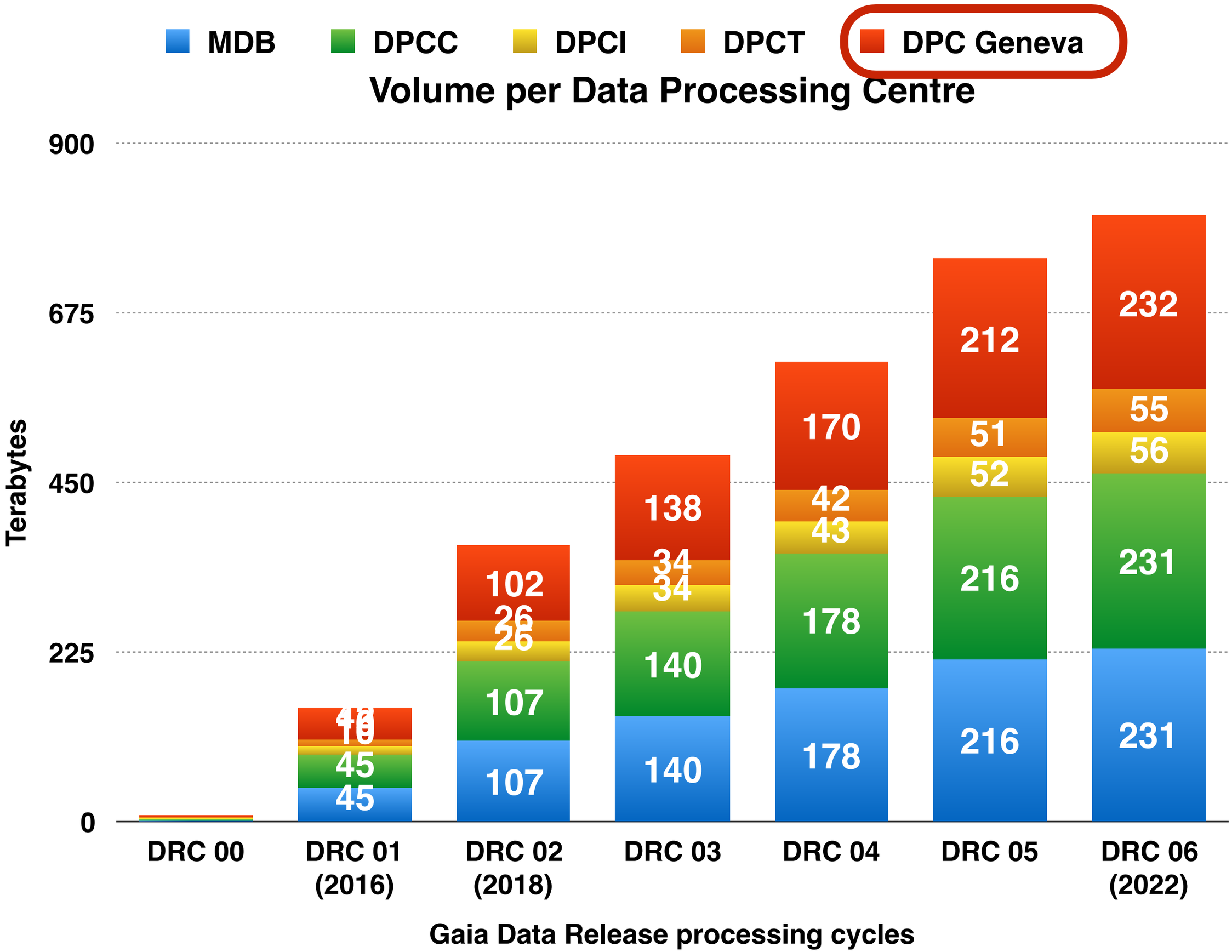
Gaia@Geneva- scientific challenge



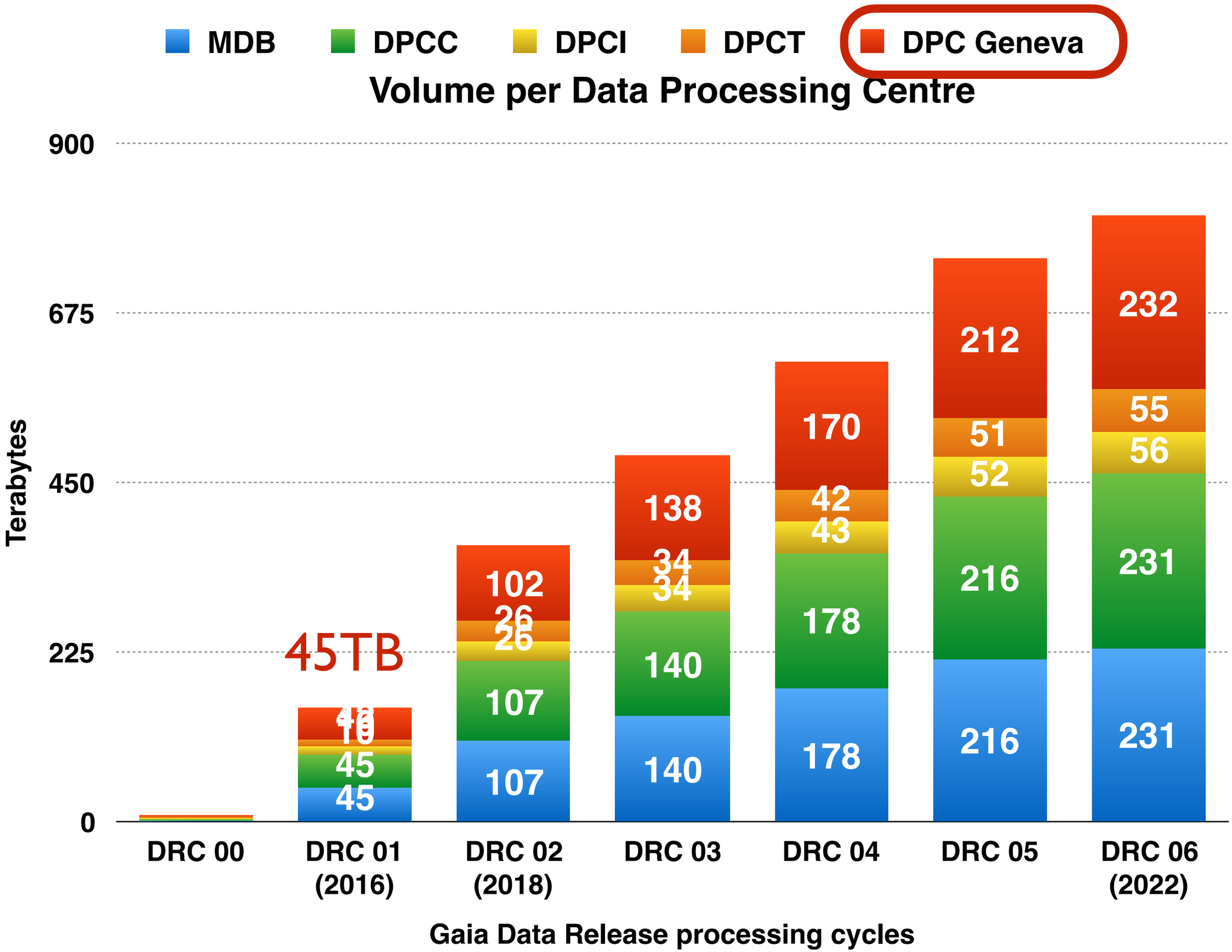
Cyclic Volume challenge - Petabyte scale



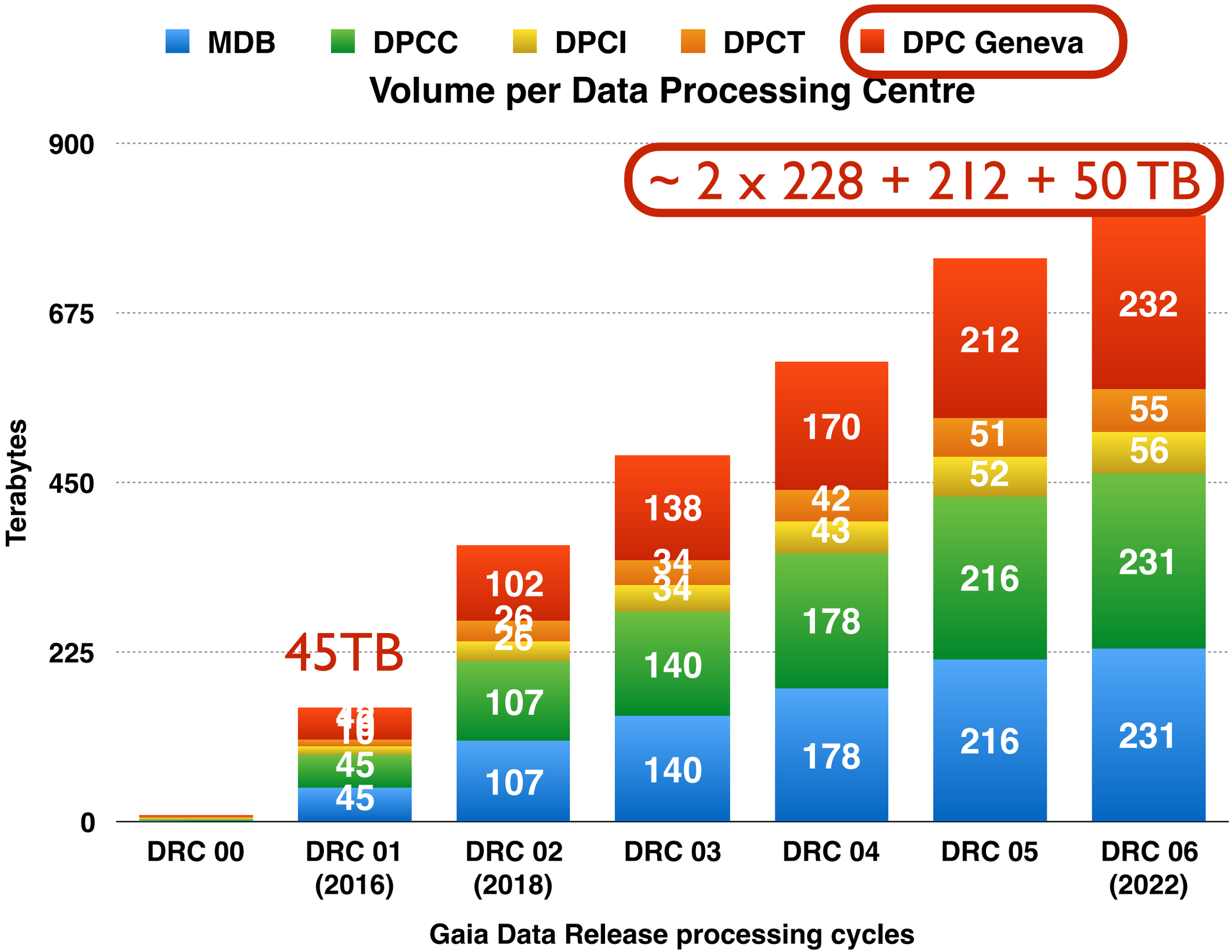
Cyclic Volume challenge - Petabyte scale



Cyclic Volume challenge - Petabyte scale



Cyclic Volume challenge - Petabyte scale



Processing challenge

Processing challenge

Data Import and transformation of photometry

Processing challenge

Data Import and transformation of photometry

Data curation

Processing challenge

Data Import and transformation of photometry

Data curation

Machine Learning

Processing challenge

Data Import and transformation of photometry

Data curation

Machine Learning

Sampling

Processing challenge

Data Import and transformation of photometry

Data curation

Machine Learning

Sampling

Batch

Processing challenge

Data Import and transformation of photometry

Data curation

Machine Learning

Sampling

Batch

Streaming

Processing challenge

Data Import and transformation of photometry

Data curation

Machine Learning

Sampling

Batch

Streaming

Ad-Hoc analysis:

Processing challenge

Data Import and transformation of photometry

Data curation

Machine Learning

Sampling

Batch

Streaming

Ad-Hoc analysis:
NewSQL, R, Java, Groovy,

Processing challenge

Data Import and transformation of photometry

Data curation

Machine Learning

Sampling

Batch

Streaming

Ad-Hoc analysis: NewSQL, R, Java, Groovy, Python

Processing challenge

Data Import and transformation of photometry

Data curation

Machine Learning

Sampling

Batch

Streaming

Ad-Hoc analysis: NewSQL, R, Java, Groovy, Python

Results curation, Data Export

Processing challenge

Data Import and transformation of photometry

Data curation

Machine Learning

Sampling

Batch

Streaming

Ad-Hoc analysis: NewSQL, R, Java, Groovy, Python

Results curation, Data Export

Software stack:

Processing challenge

Data Import and transformation of photometry

Data curation

Machine Learning

Sampling

Batch

Streaming

Ad-Hoc analysis: NewSQL, R, Java, Groovy, Python

Results curation, Data Export

Software stack:

Hundreds of Java classes

newSQL

Processing challenge

Data Import and transformation of photometry

Data curation

Machine Learning

Sampling

Batch

Streaming

Ad-Hoc analysis: NewSQL, R, Java, Groovy, Python

Results curation, Data Export

Software stack:

Hundreds of Java classes

newSQL

R - modules

Processing challenge

Data Import and transformation of photometry

Data curation

Machine Learning

Sampling

Batch

Streaming

Ad-Hoc analysis: NewSQL, R, Java, Groovy, Python

Results curation, Data Export

Software stack:

Hundreds of Java classes

newSQL

R - modules

Data-grid:

Processing challenge

Data Import and transformation of photometry

Data curation

Machine Learning

Sampling

Batch

Streaming

Ad-Hoc analysis: NewSQL, R, Java, Groovy, Python

Results curation, Data Export

Software stack:

Hundreds of Java classes

newSQL

R - modules

Data-grid: distributed RAM

Processing challenge

Data Import and transformation of photometry

Data curation

Machine Learning

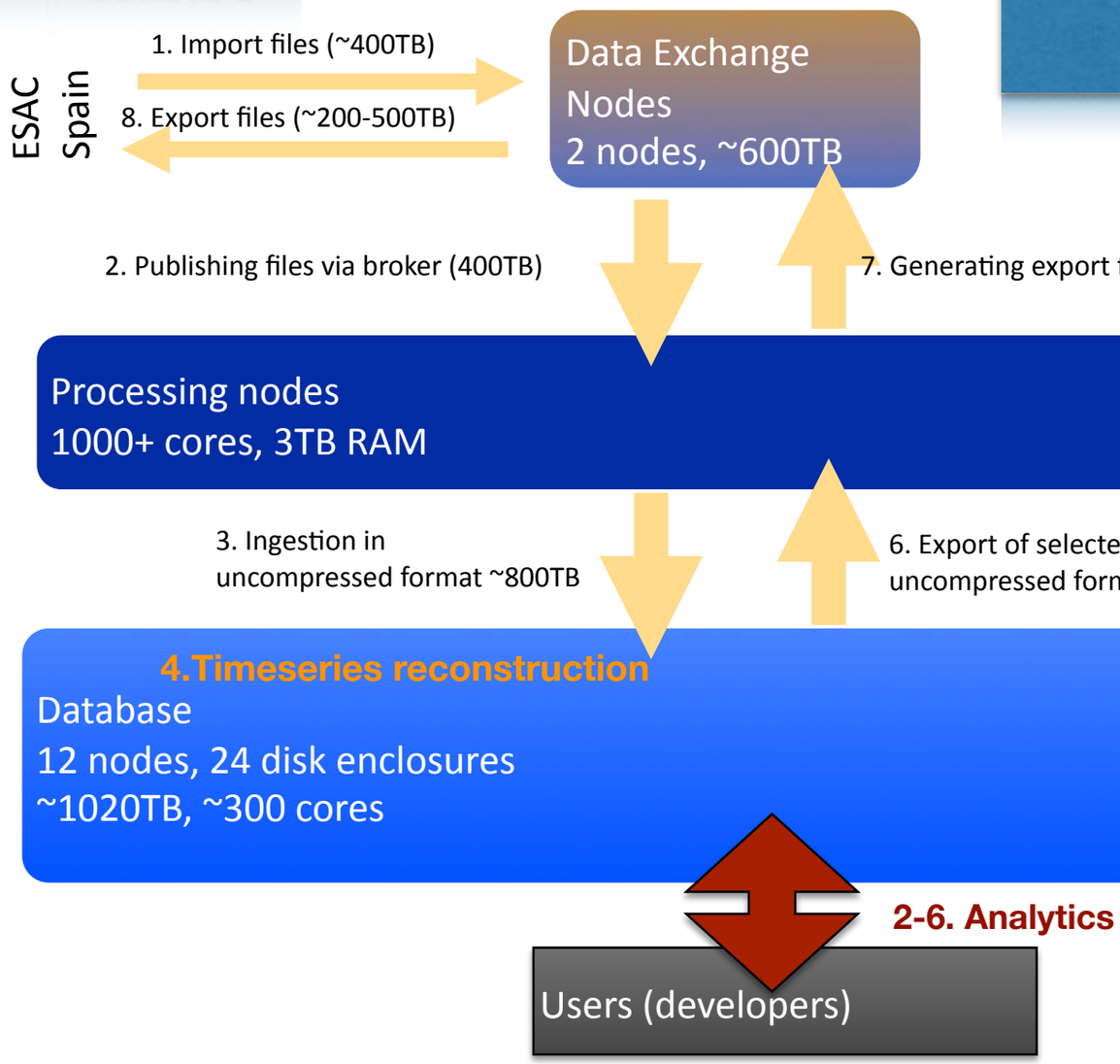
Sampling

Batch

Streaming

Ad-Hoc analysis: NewSQL, R, Java, Groovy, Python

Results curation, Data Export



Software stack:

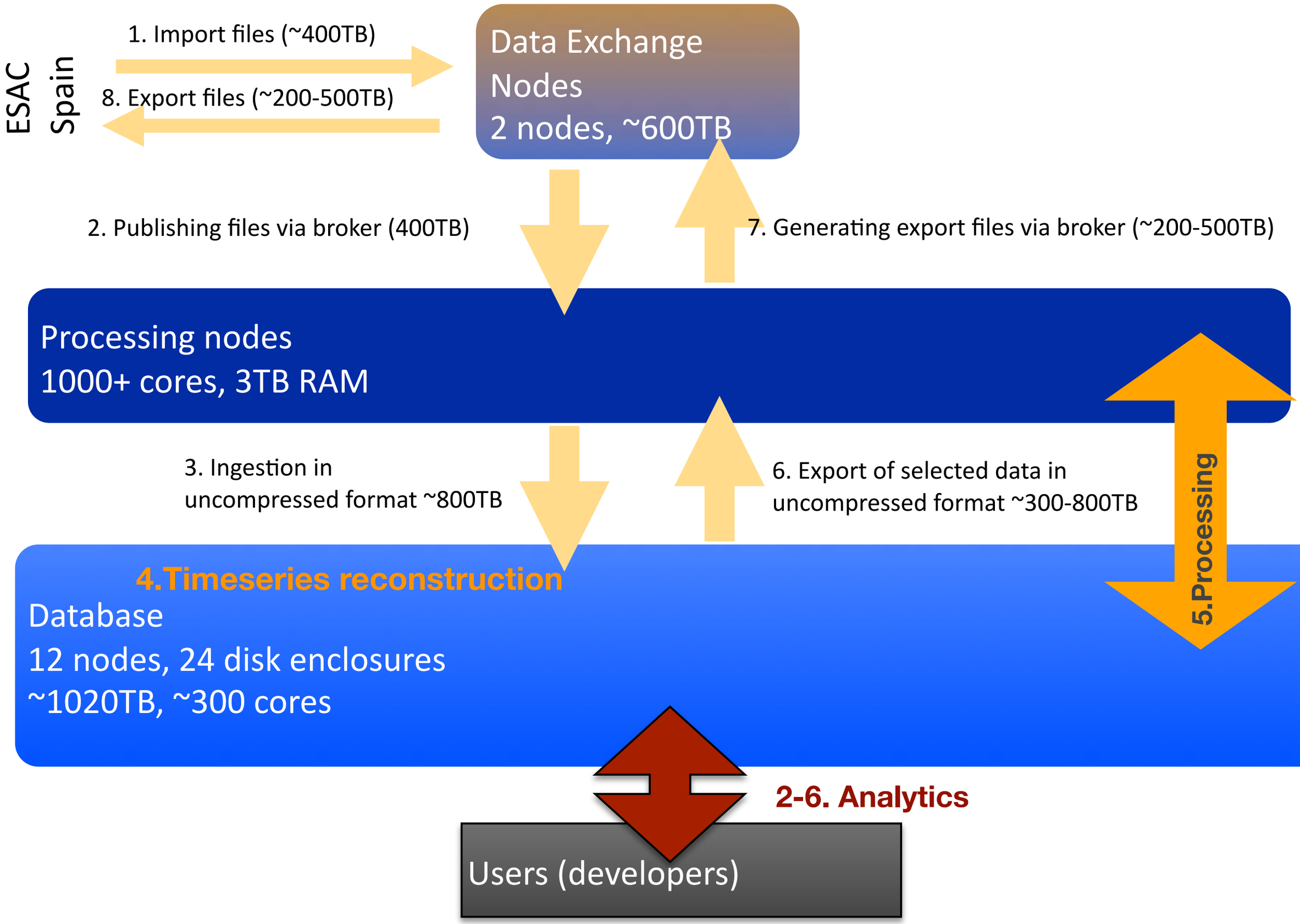
Hundreds of Java classes

newSQL

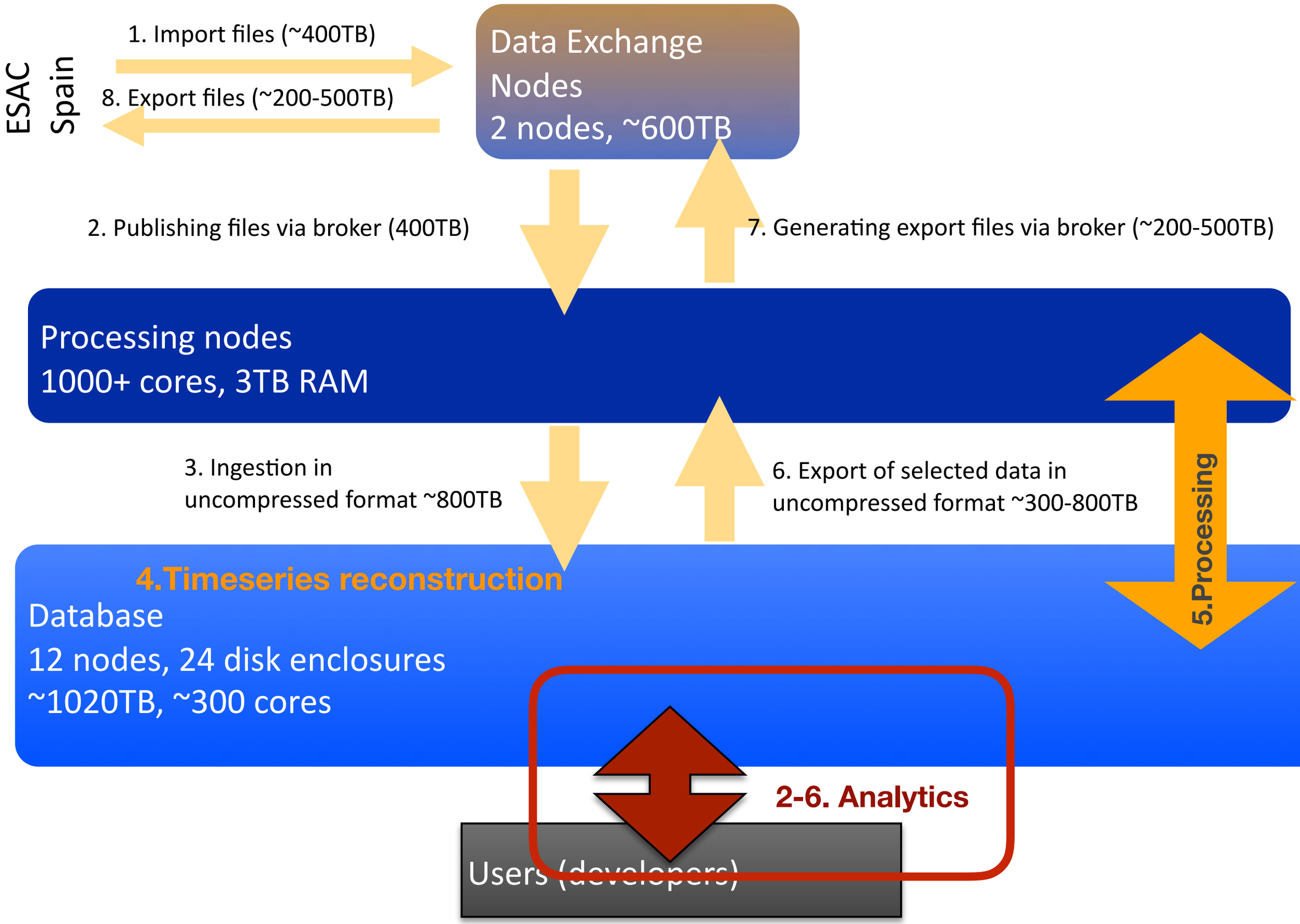
R - modules

Data-grid: distributed RAM

Processing challenge



Processing challenge



Data access philosophy: Needs

Data access philosophy: Needs

Free, Direct access to any tuple from 2 (4) x 10^{10}

Data access philosophy: Needs

Free, Direct access to any tuple from 2 (4) x 10^{10}

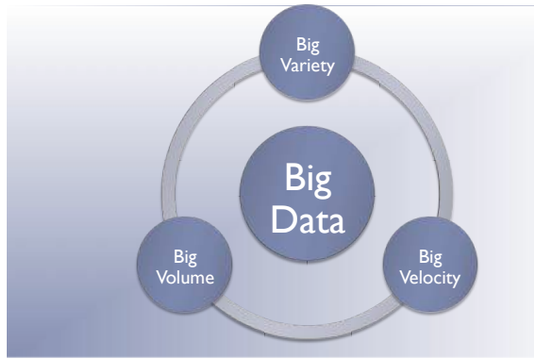
Quick, iterative environment

Data access philosophy: Needs

Free, Direct access to any tuple from 2 (4) x 10^{10}

Quick, iterative environment

3 x V vs Big Volume problems

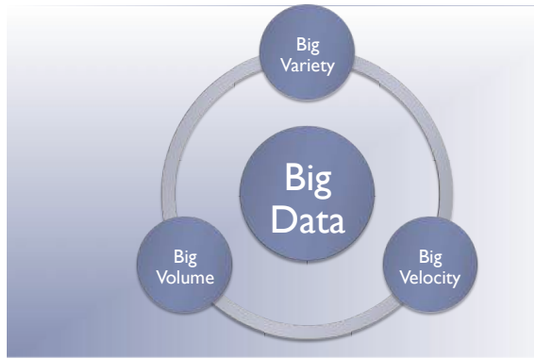


Data access philosophy: Needs

Free, Direct access to any tuple from 2 (4) x 10^{10}

Quick, iterative environment

3 x V vs Big Volume problems



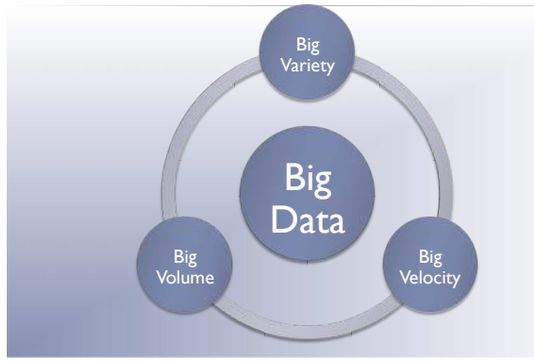
Hybrid solutions:
Spark-like streaming + Data grids +

Data access philosophy: Needs

Free, Direct access to any tuple from 2 (4) x 10^{10}

Quick, iterative environment

3 x V vs Big Volume problems



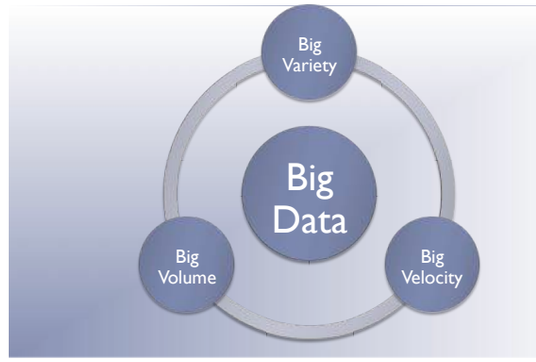
Hybrid solutions:
Spark-like streaming + Data grids +
[Columnar store and/or Distributed relational

Data access philosophy: Needs

Free, Direct access to any tuple from 2 (4) x 10^{10}

Quick, iterative environment

3 x V vs Big Volume problems



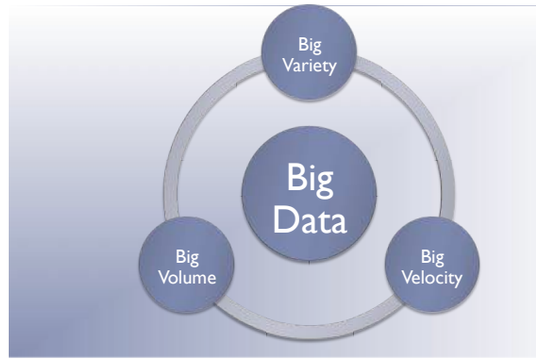
Hybrid solutions:
Spark-like streaming + Data grids +
[Columnar store and/or Distributed relational and/or document DBs]
CPU + data affinity for *part* of operations

Data access philosophy: Needs

Free, Direct access to any tuple from 2 (4) x 10^{10}

Quick, iterative environment

3 x V vs Big Volume problems



Hybrid solutions:
Spark-like streaming + Data grids +
[Columnar store and/or Distributed relational
and/or document DBs]
CPU + data affinity for *part* of operations

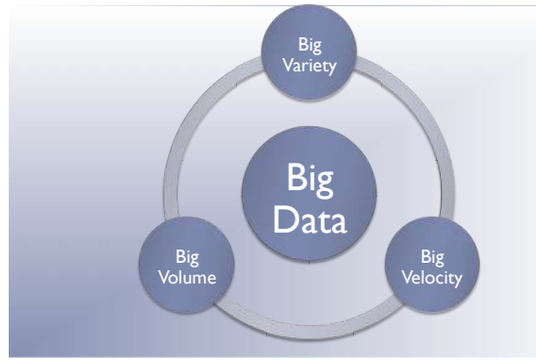
Domain Specific Language: extended SQL, Groovy, MVEL, R

Data access philosophy: Needs

Free, Direct access to any tuple from 2 (4) x 10^{10}

Quick, iterative environment

3 x V vs Big Volume problems



Hybrid solutions:
Spark-like streaming + Data grids +
[Columnar store and/or Distributed relational
and/or document DBs]
CPU + data affinity for *part* of operations

Domain Specific Language: extended SQL, Groovy, MVEL, R

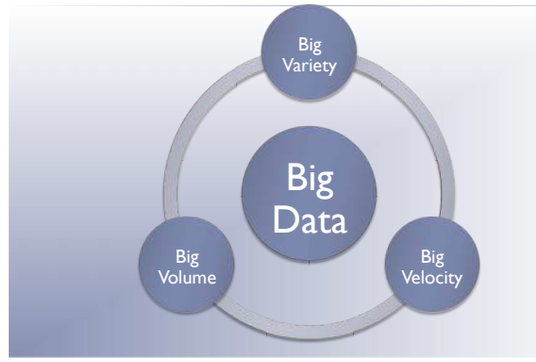
Role of NewSQL - concise data access code

Data access philosophy: Needs

Free, Direct access to any tuple from 2 (4) x 10^{10}

Quick, iterative environment

3 x V vs Big Volume problems



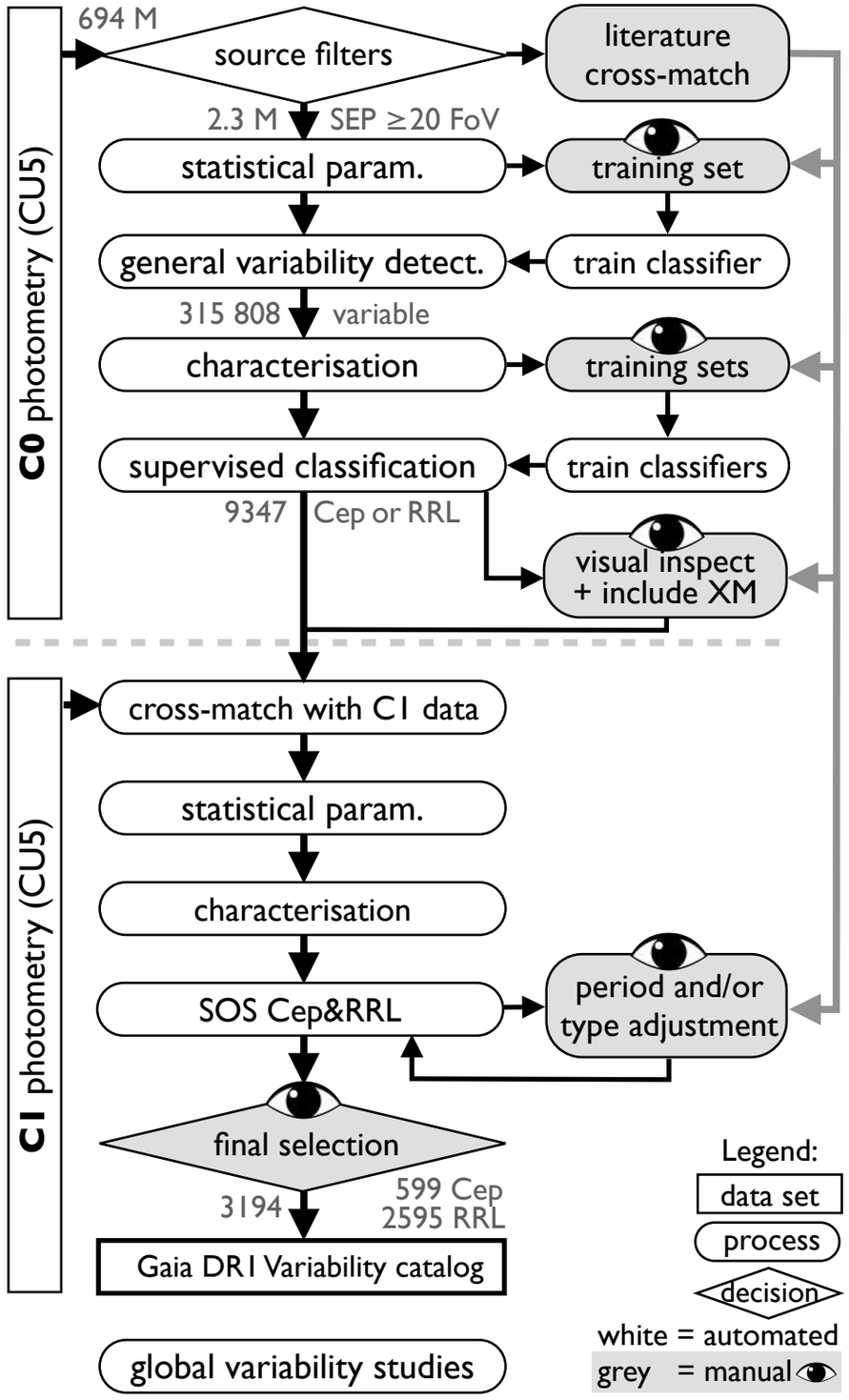
Hybrid solutions:
Spark-like streaming + Data grids +
[Columnar store and/or Distributed relational
and/or document DBs]
CPU + data affinity for *part* of operations

Domain Specific Language: extended SQL, Groovy, MVEL, R

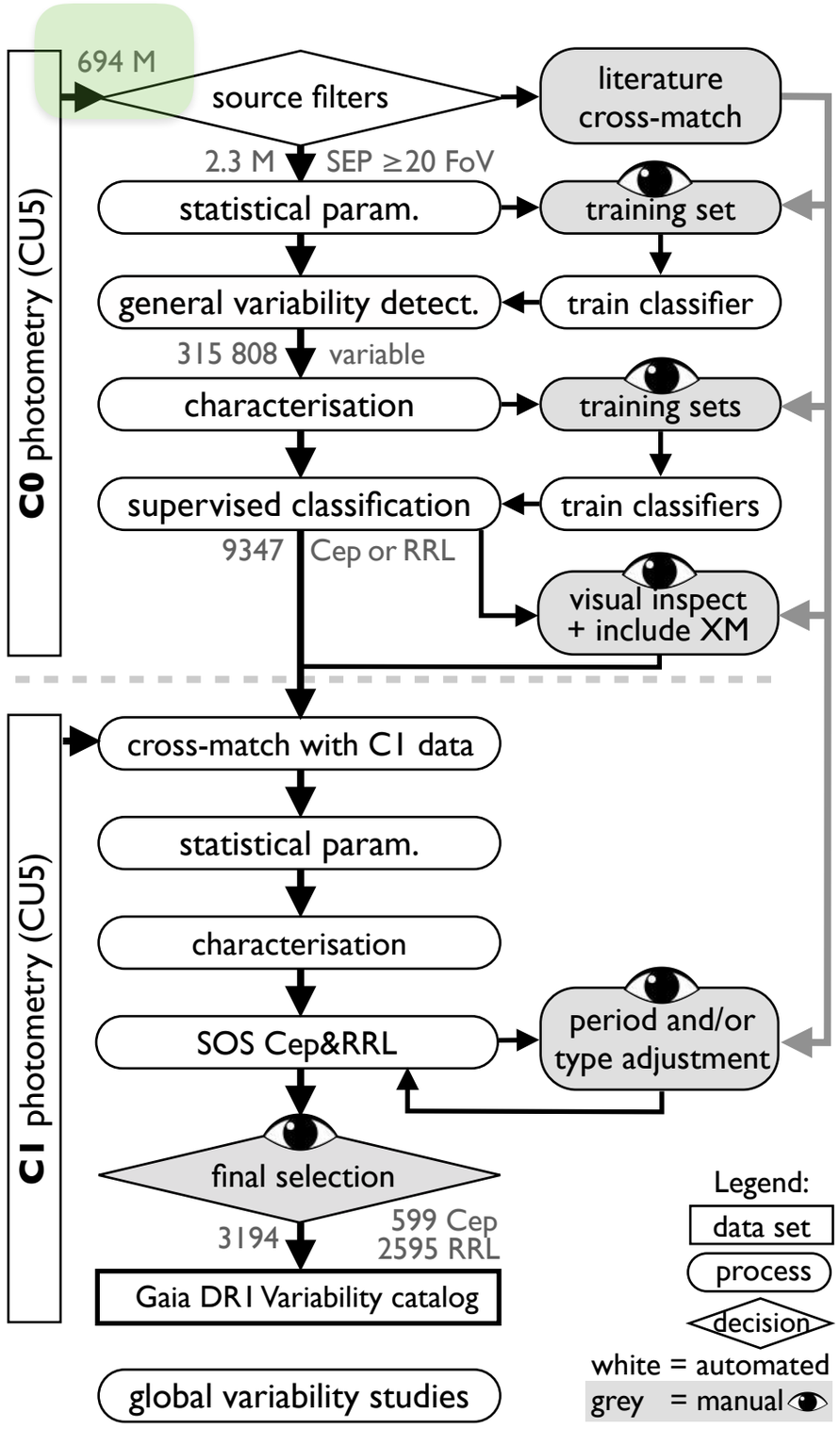
Role of NewSQL - concise data access code

Advanced indexing techniques, Sketches - data digests

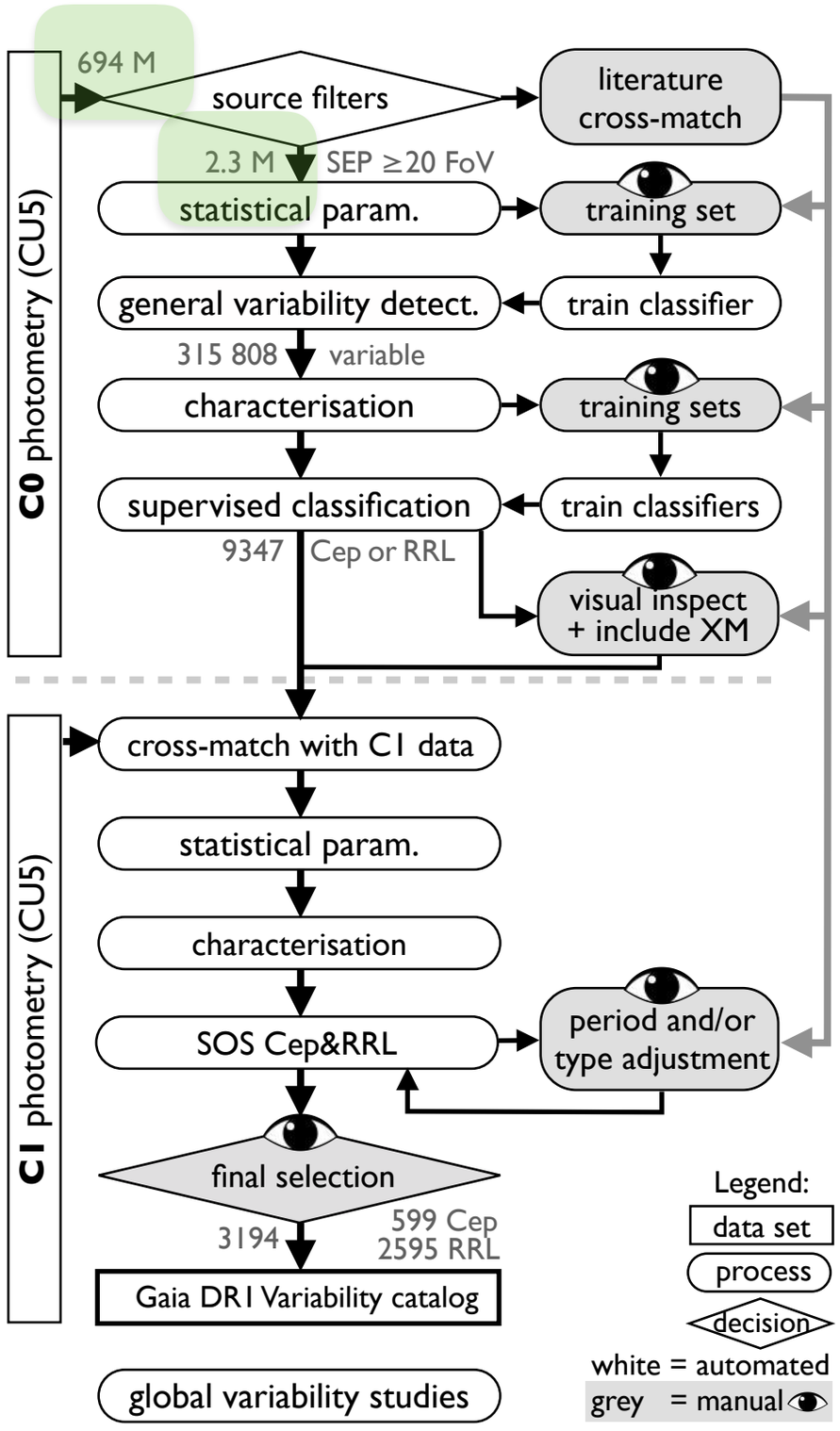
Cyclic Processing: Dynamic workflows



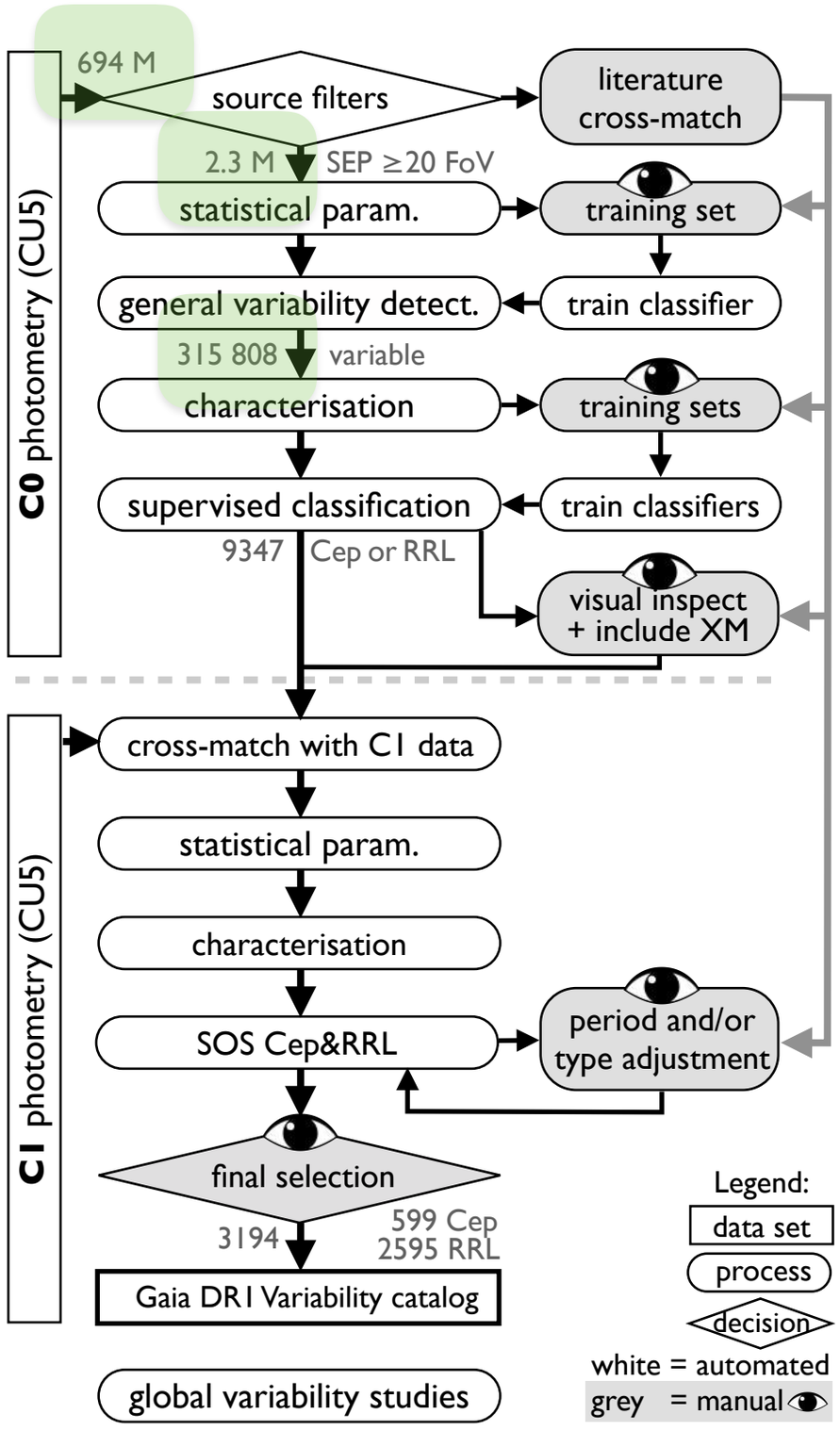
Cyclic Processing: Dynamic workflows



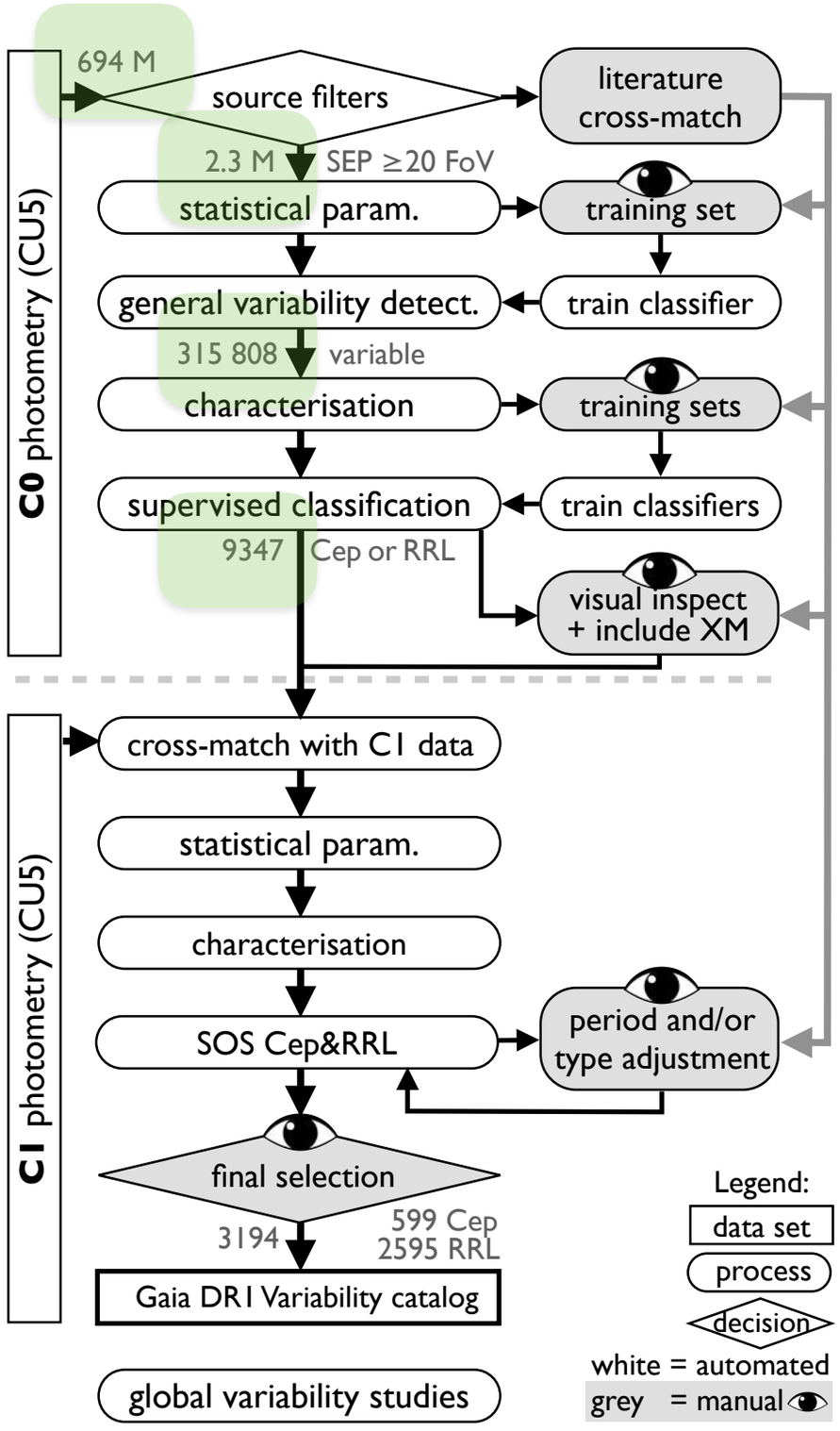
Cyclic Processing: Dynamic workflows



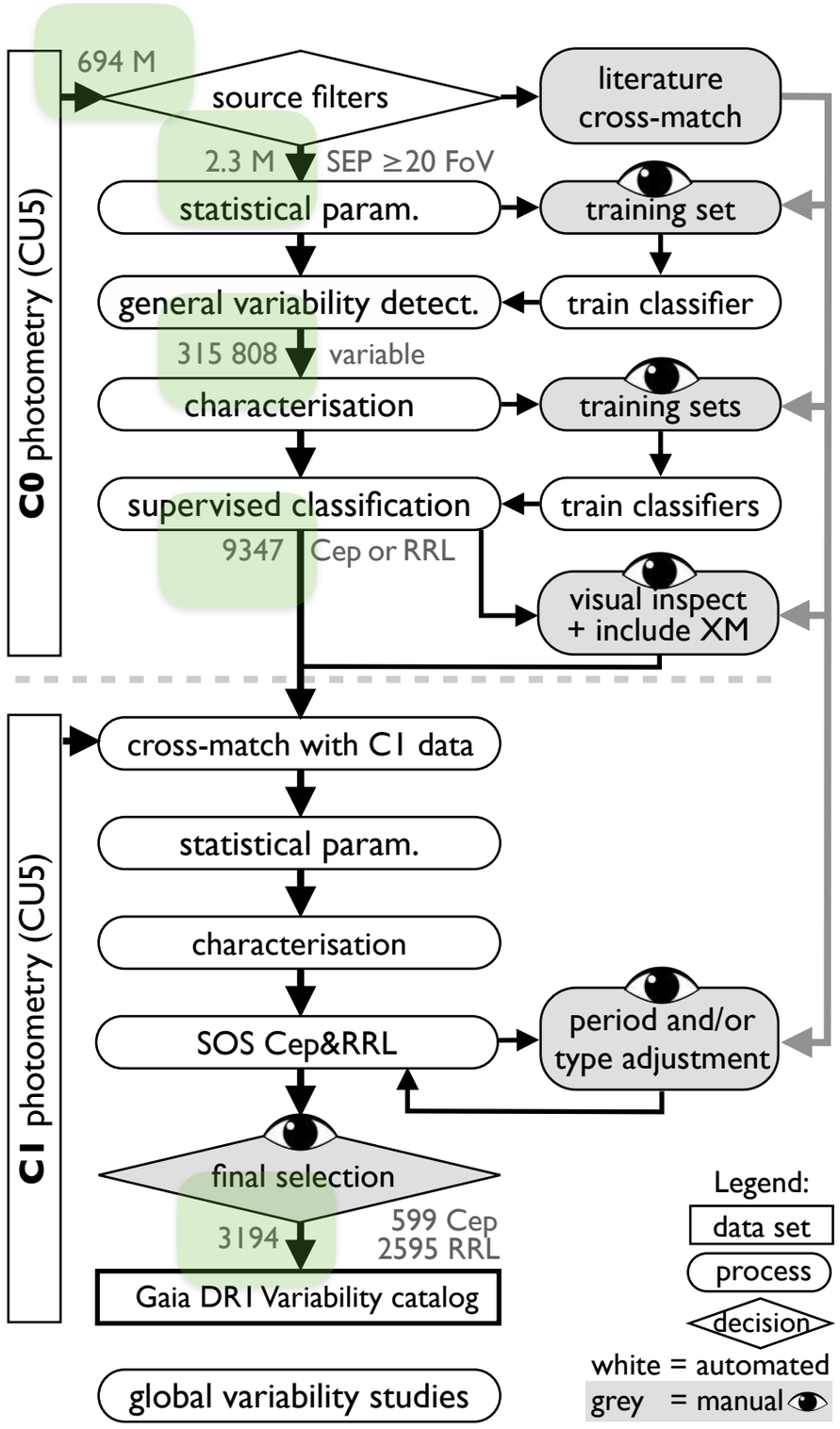
Cyclic Processing: Dynamic workflows



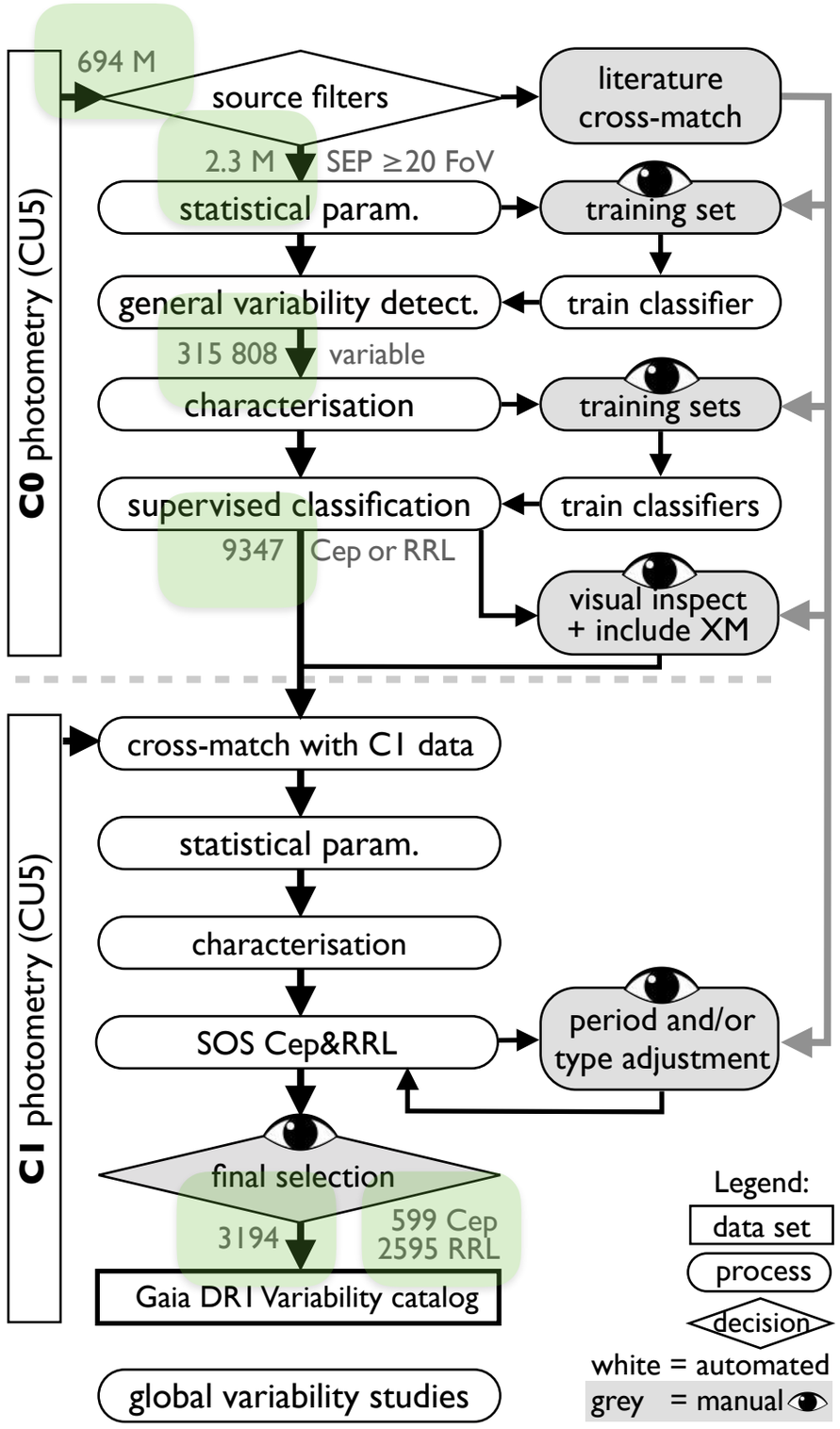
Cyclic Processing: Dynamic workflows



Cyclic Processing: Dynamic workflows



Cyclic Processing: Dynamic workflows



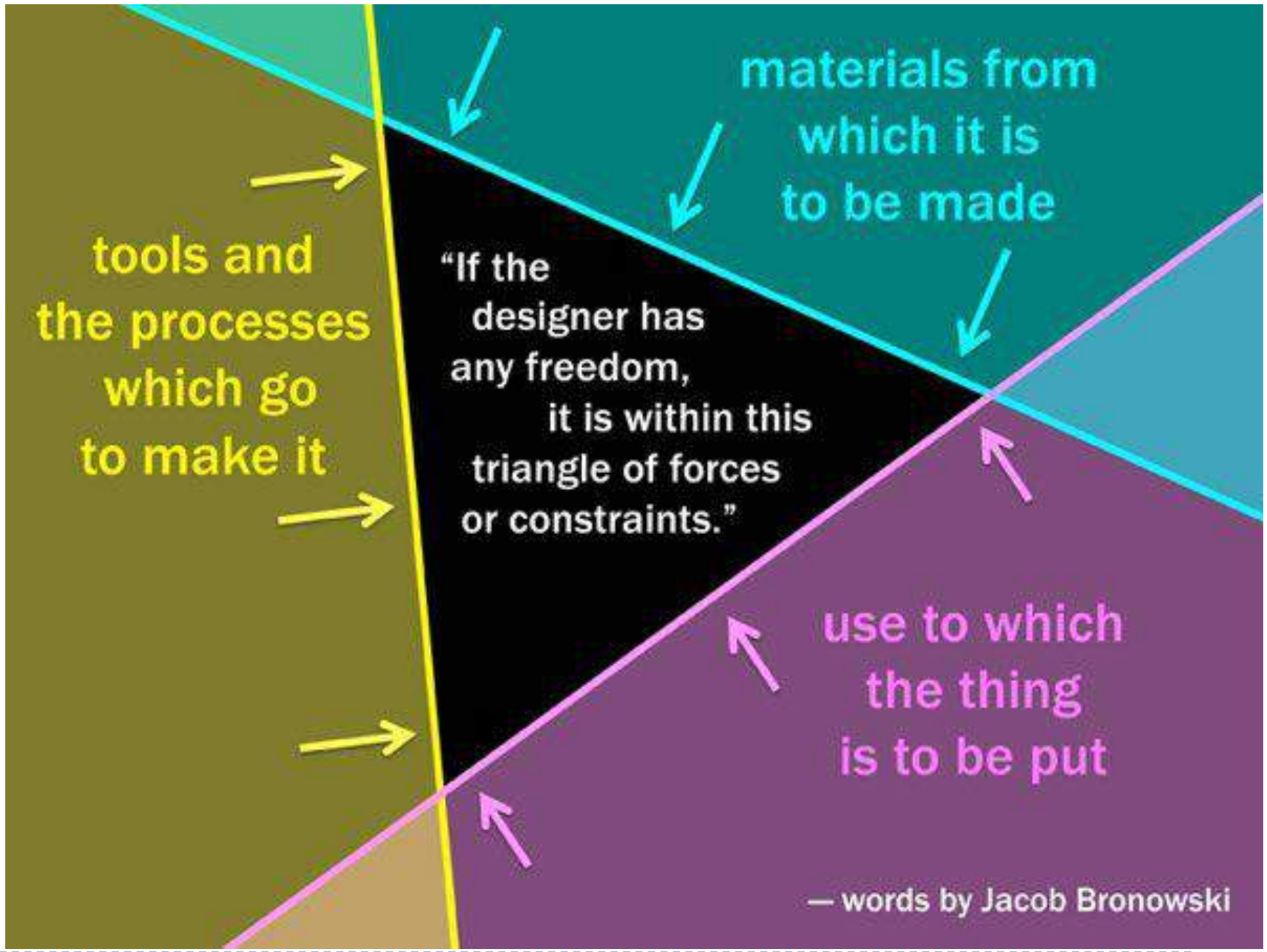
Structure

- Story of perpetual change
- Databases in Astronomy
- Gaia mission
- Gaia processing at CU7/DPC Geneva
- P[rolost]gres for science
- Postgres-XL tale
- Collaboration
- Future

Science as a design exercise..

SciAm 2015/2/10 - Visualization constraints

by Jacob Bronowski



Context as a shaping force

If the designer has any freedom, it is within this triangle of forces or constraints.

[*The Shape of Things, The Observer,1952*]

Postgres for science

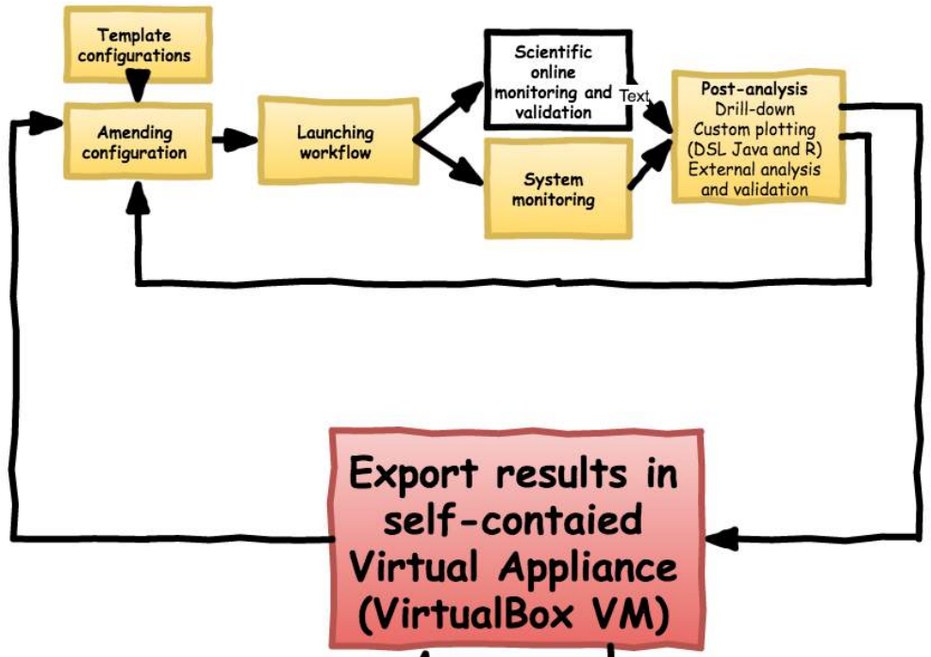
- Need for [software] tools
  - Open source
  - Stable *enough*
  - Scaling well *enough*
    - *Vertical vs **horizontal***
      - *Columnar & Parallel*

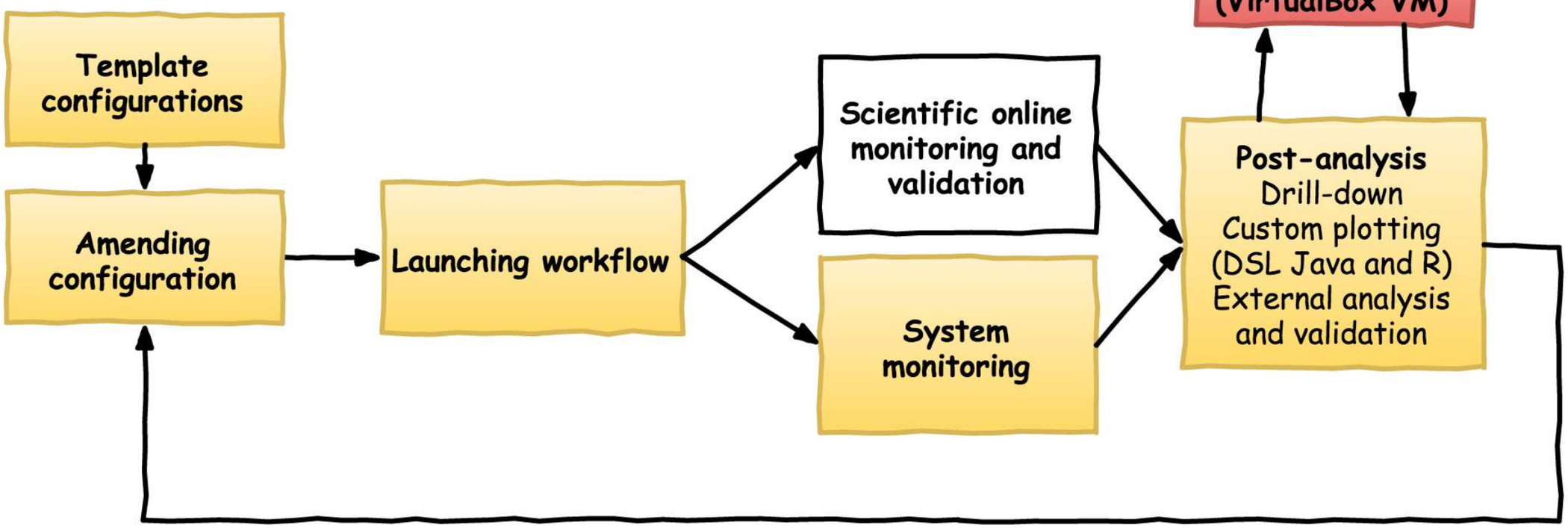
Postgres for science

- Needs
  - Extensive *enough*
    - Spatial index (q3c, pgSphere)
    - Bloom, Brin, R-Tree, GIN, GIST.., KNN-..
    - plJava, plR, …
  - Extendible
    - PG Extensions

Workflow Scaling - PG as an appliance

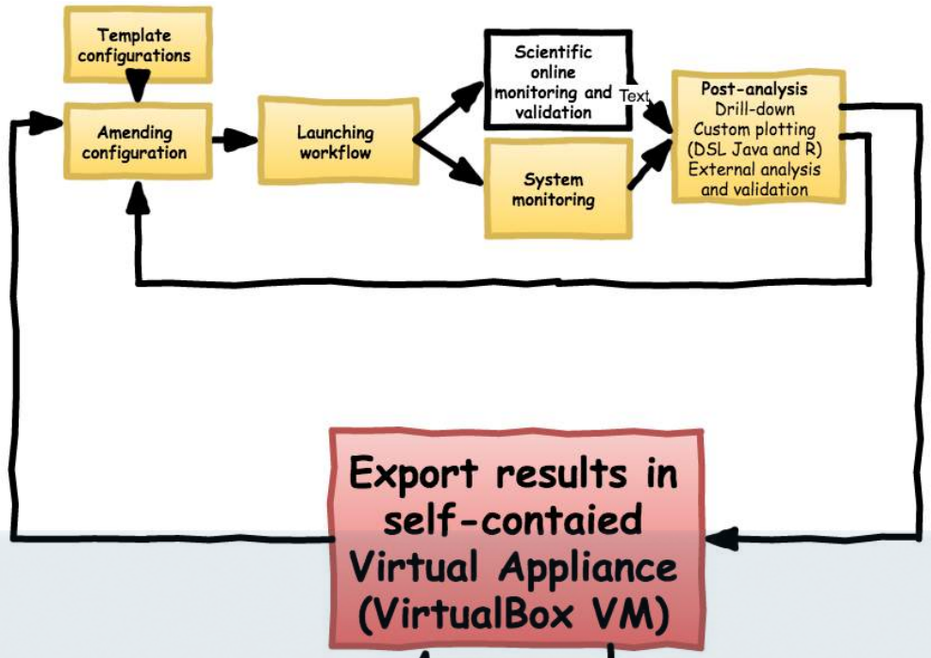
- Distributed iterative process, repeated on small selections or samples of sources outside Geneva.
  - Too few resources to re-run, used for results analysis and tagging via Visualisation tool mostly.
  - Virtual Appliances issued daily at some point.
  - 5GB of plots generated, 100s of them with 1000s of details
  - Very intense communication



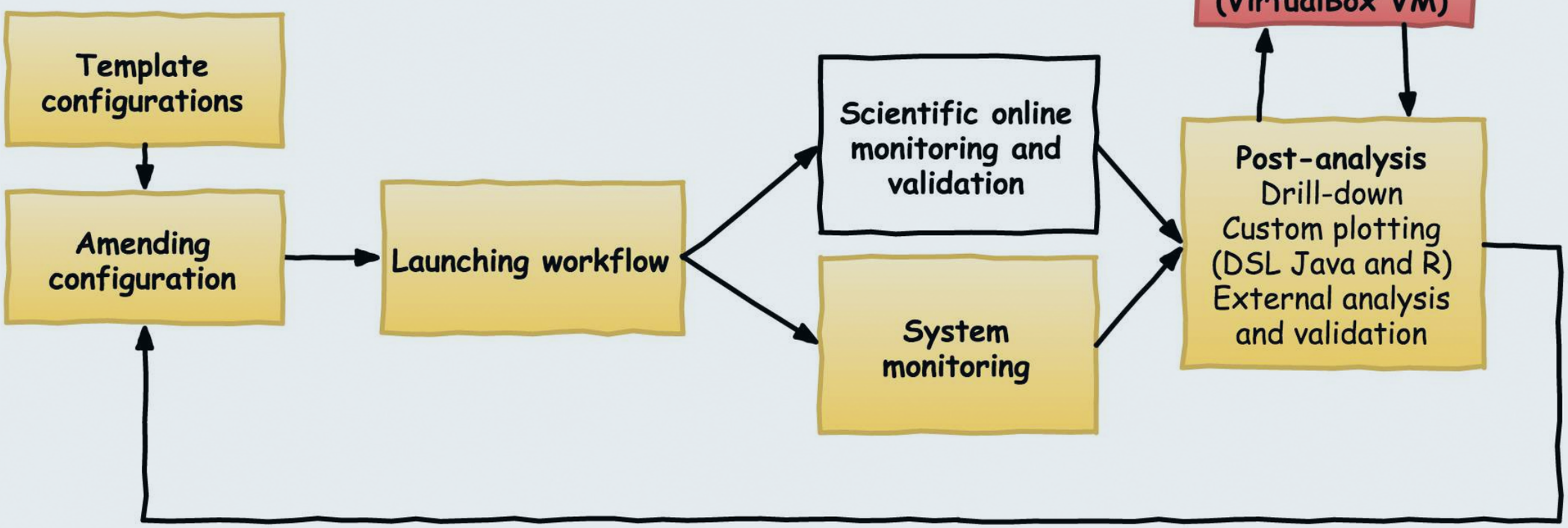


Workflow Scaling - PG as an appliance

- Distributed iterative process, repeated on small selections or samples of sources outside Geneva.
  - Too few resources to re-run, used for results analysis and tagging via Visualisation tool mostly.
  - Virtual Appliances issued daily at some point.
  - 5GB of plots generated, 100s of them with 1000s of details
  - Very intense communication

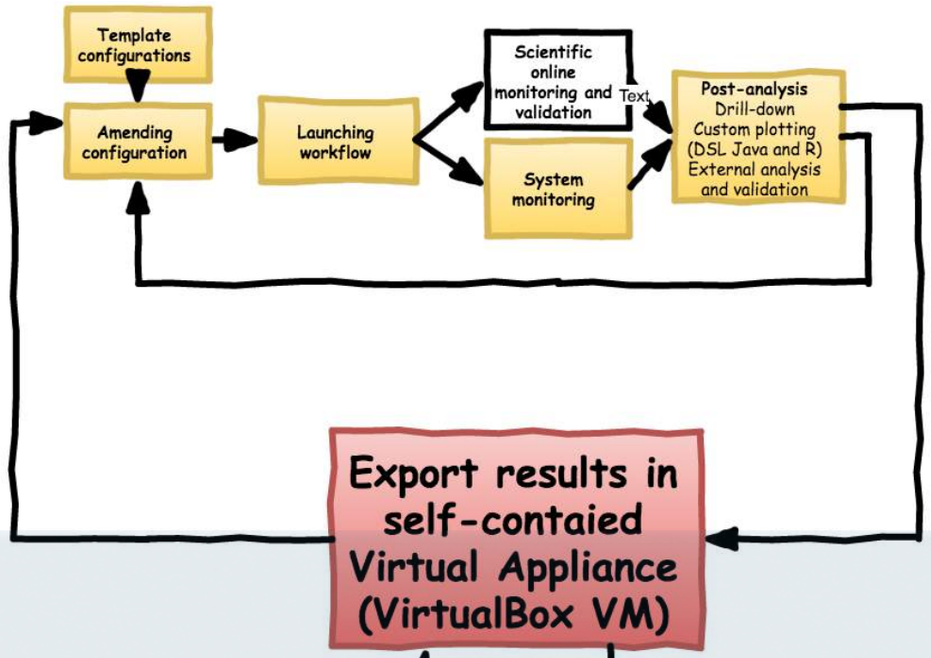


DPCG

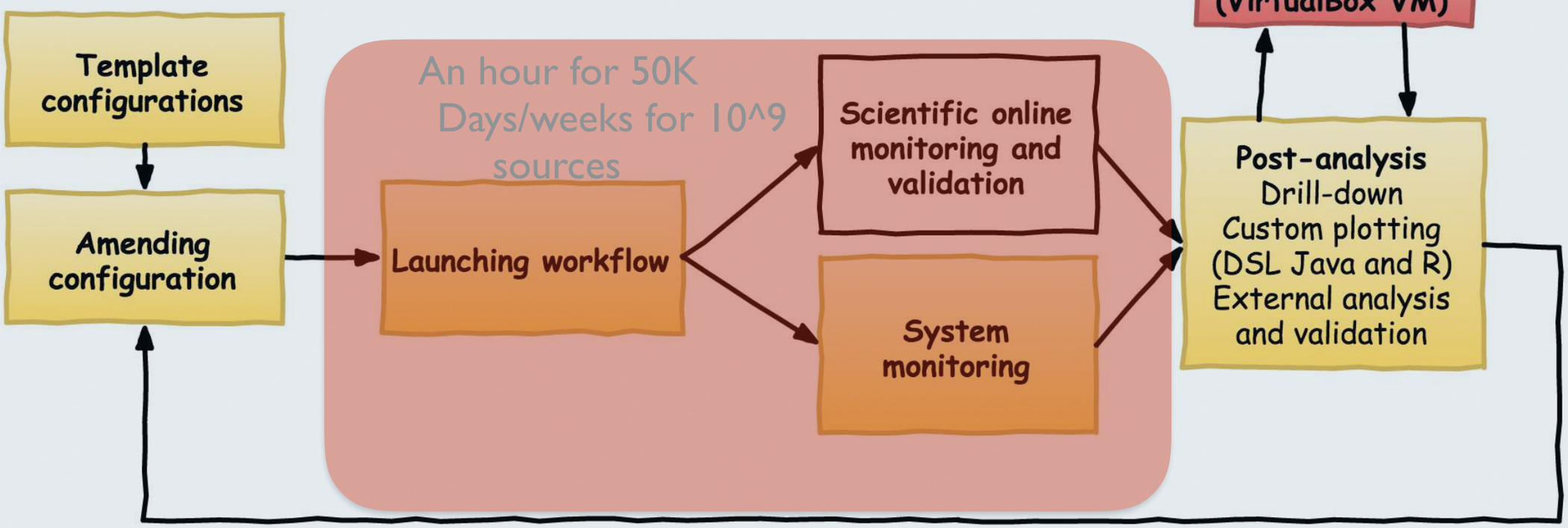


Workflow Scaling - PG as an appliance

- Distributed iterative process, repeated on small selections or samples of sources outside Geneva.
  - Too few resources to re-run, used for results analysis and tagging via Visualisation tool mostly.
  - Virtual Appliances issued daily at some point.
  - 5GB of plots generated, 100s of them with 1000s of details
  - Very intense communication

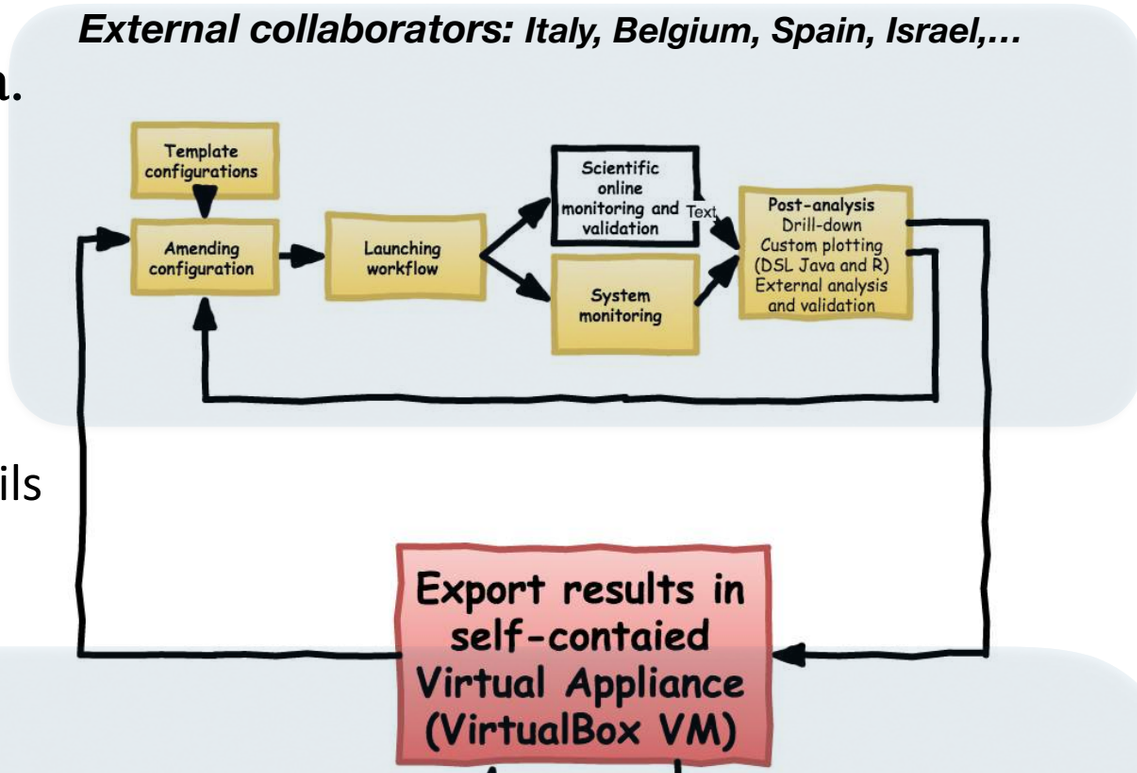


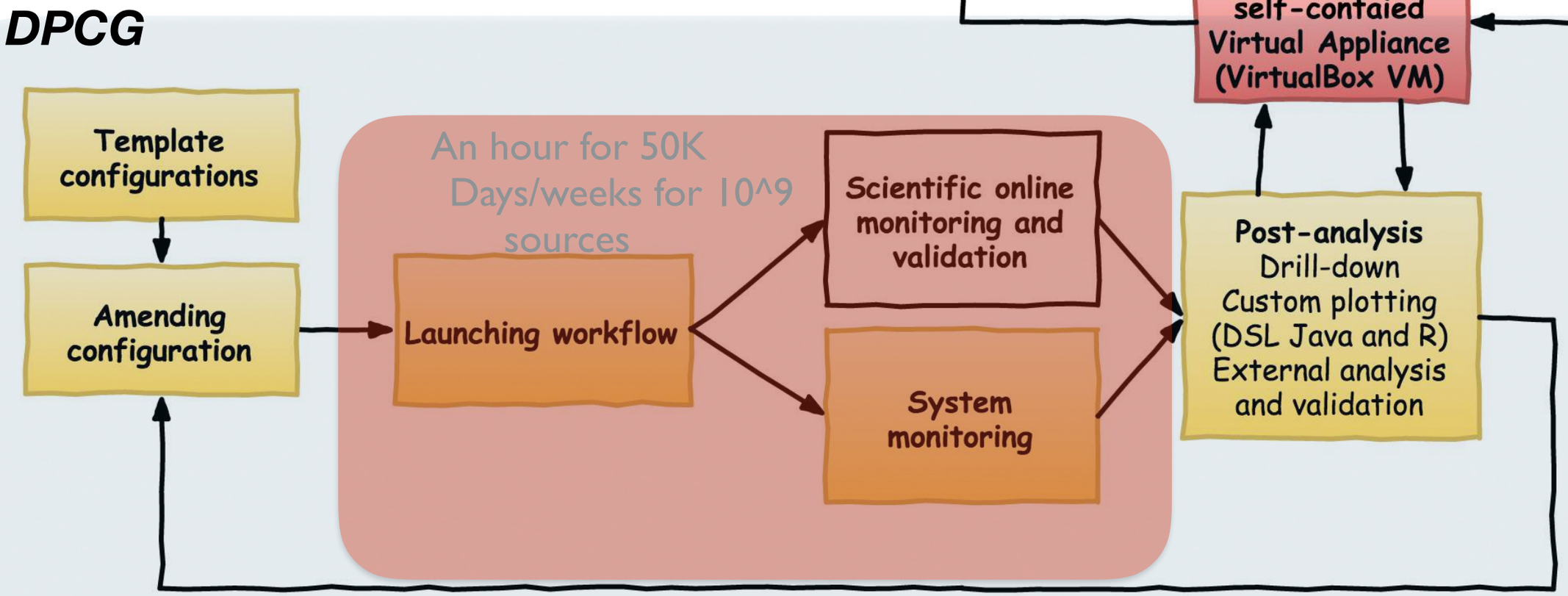
DPCG



Workflow Scaling - PG as an appliance

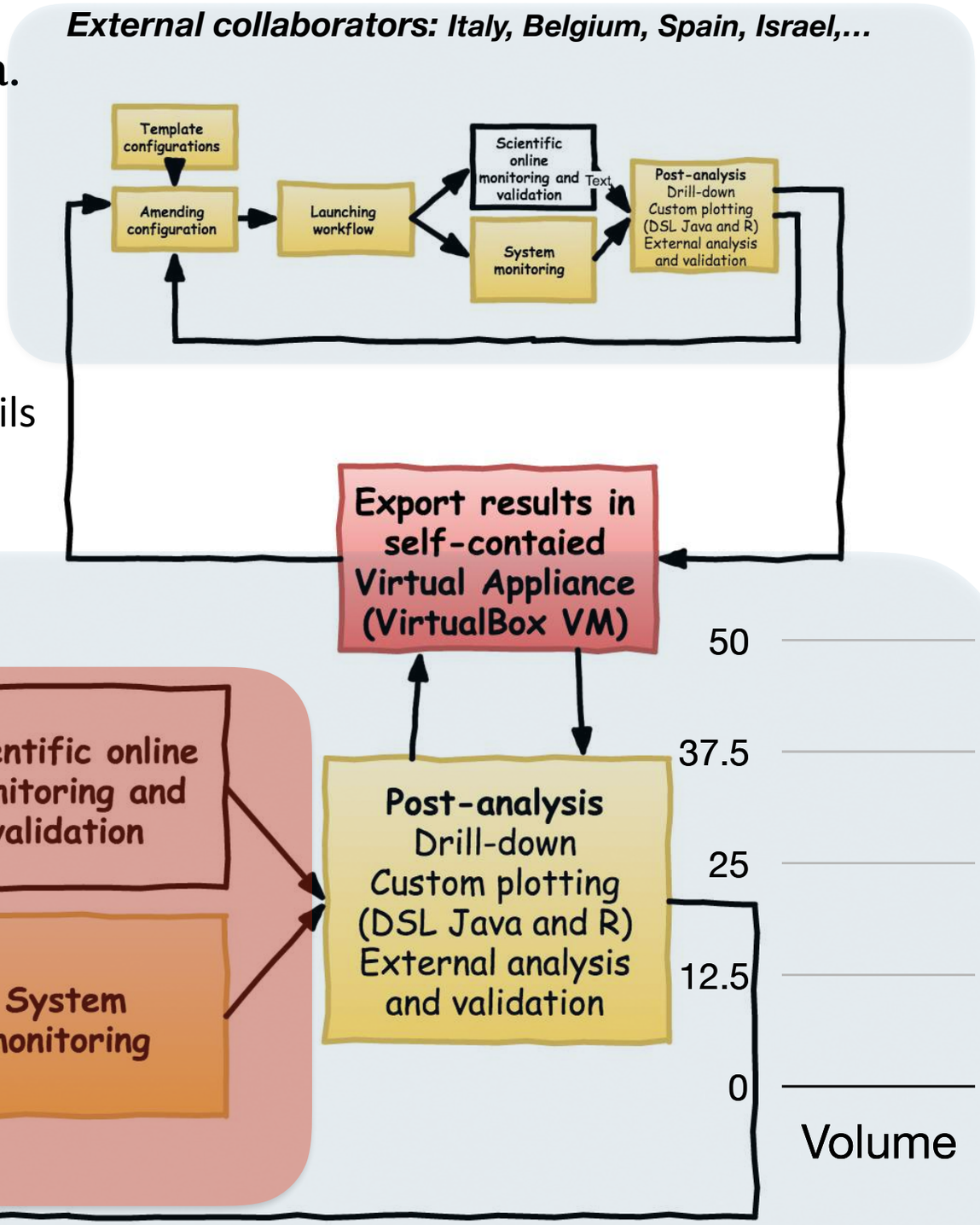
- Distributed iterative process, repeated on small selections or samples of sources outside Geneva.
  - Too few resources to re-run, used for results analysis and tagging via Visualisation tool mostly.
  - Virtual Appliances issued daily at some point.
  - 5GB of plots generated, 100s of them with 1000s of details
  - Very intense communication





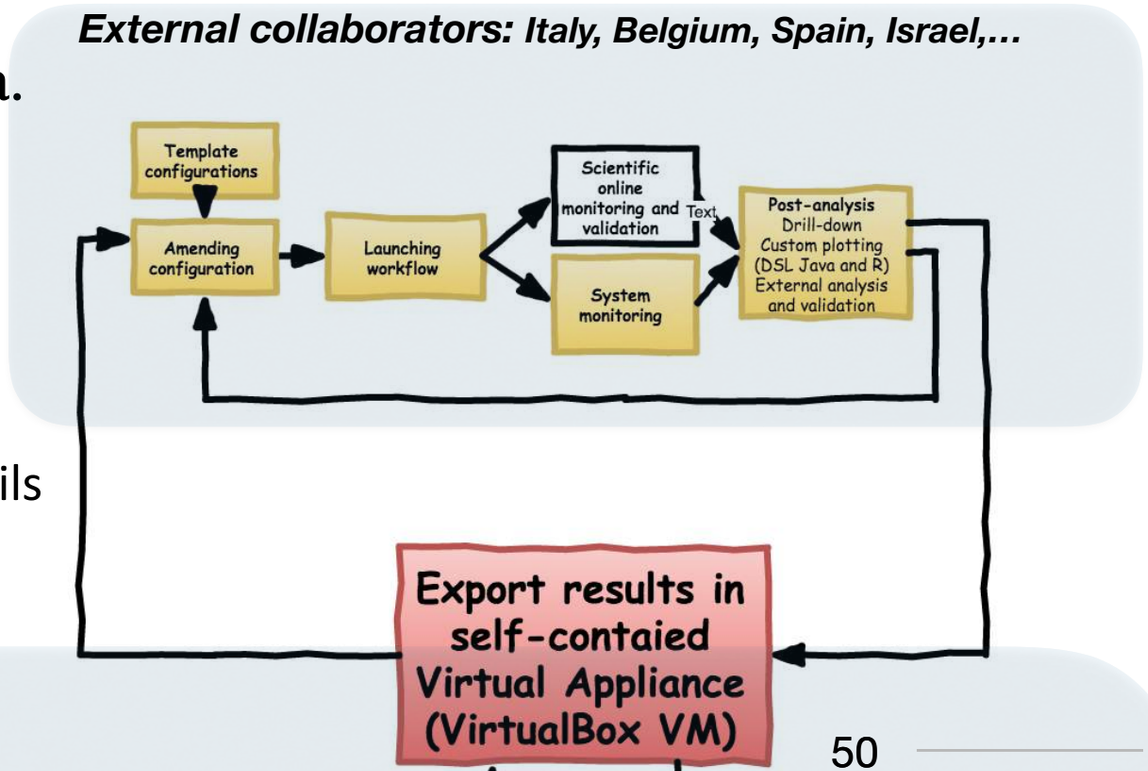
Workflow Scaling - PG as an appliance

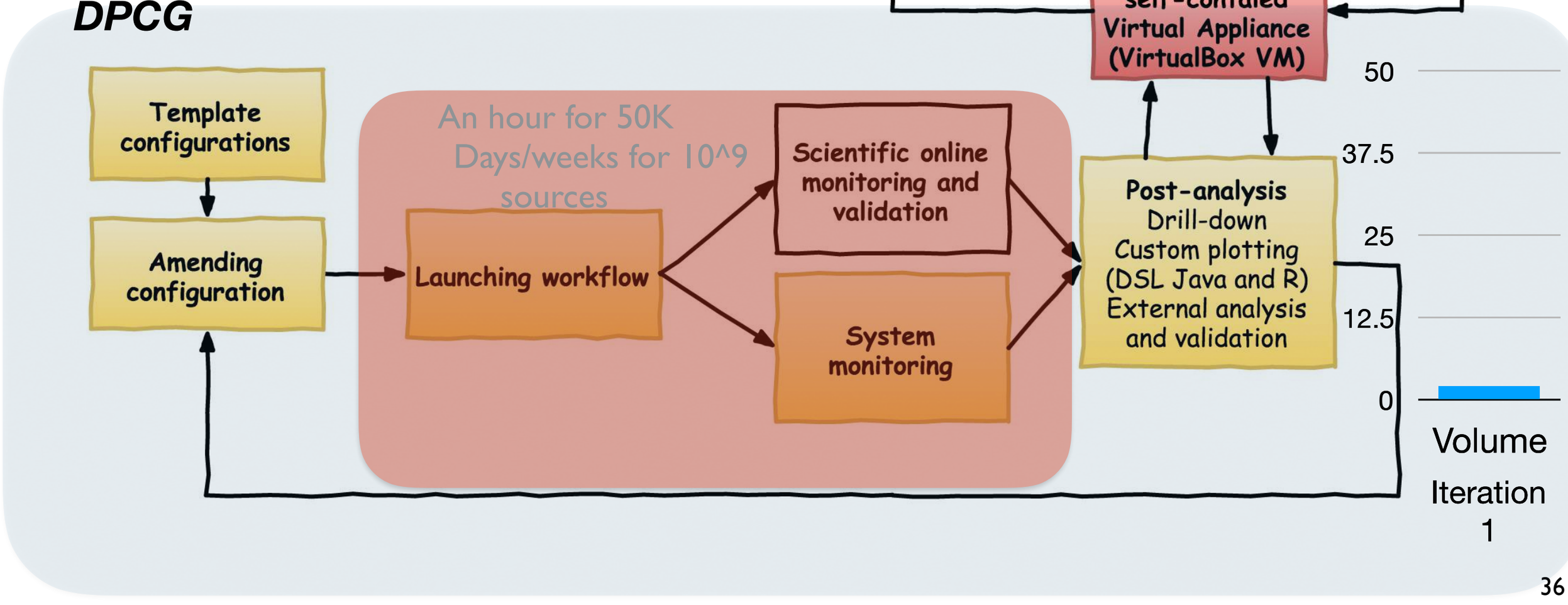
- Distributed iterative process, repeated on small selections or samples of sources outside Geneva.
  - Too few resources to re-run, used for results analysis and tagging via Visualisation tool mostly.
  - Virtual Appliances issued daily at some point.
  - 5GB of plots generated, 100s of them with 1000s of details
  - Very intense communication



Workflow Scaling - PG as an appliance

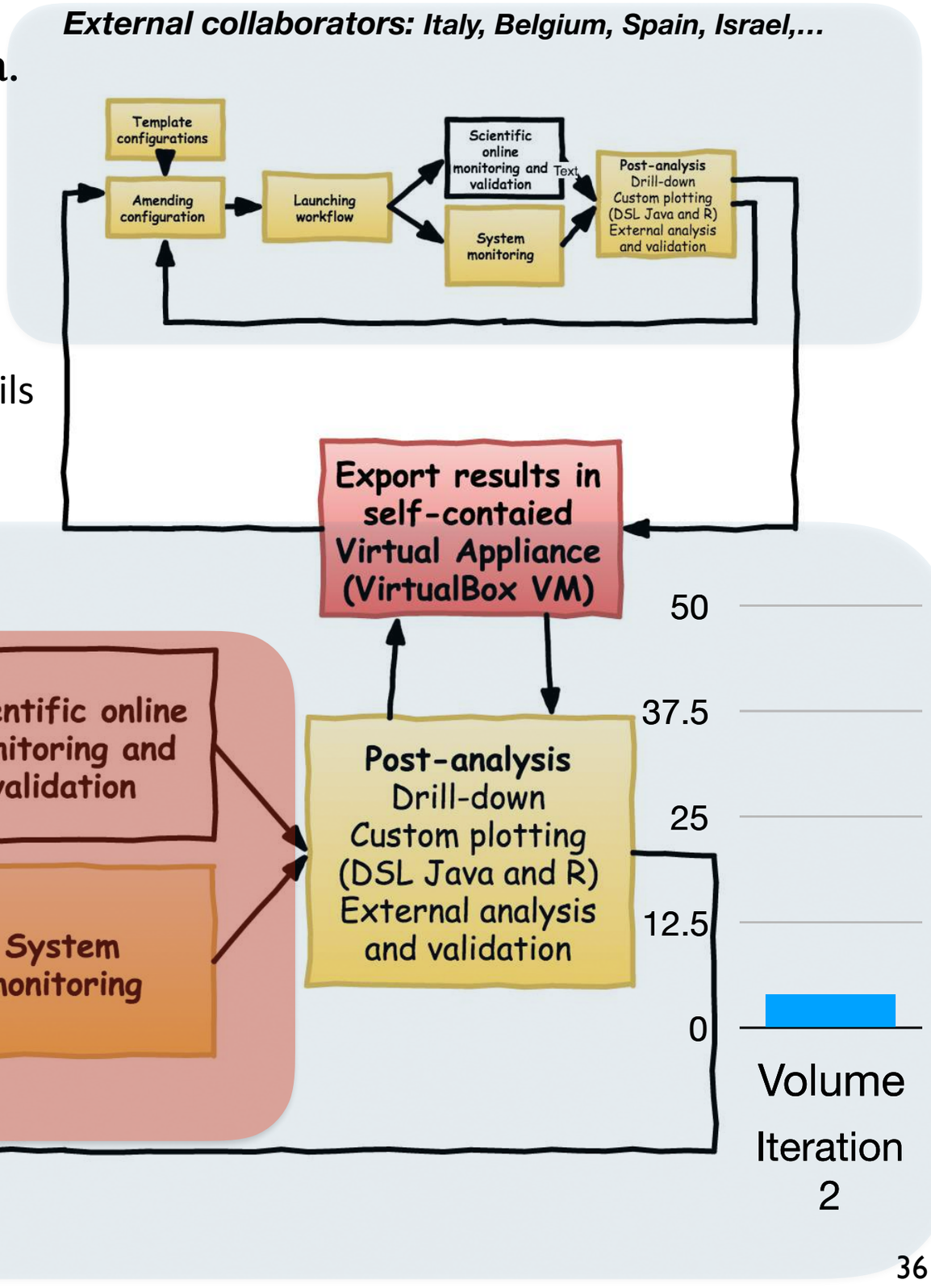
- Distributed iterative process, repeated on small selections or samples of sources outside Geneva.
  - Too few resources to re-run, used for results analysis and tagging via Visualisation tool mostly.
  - Virtual Appliances issued daily at some point.
  - 5GB of plots generated, 100s of them with 1000s of details
  - Very intense communication





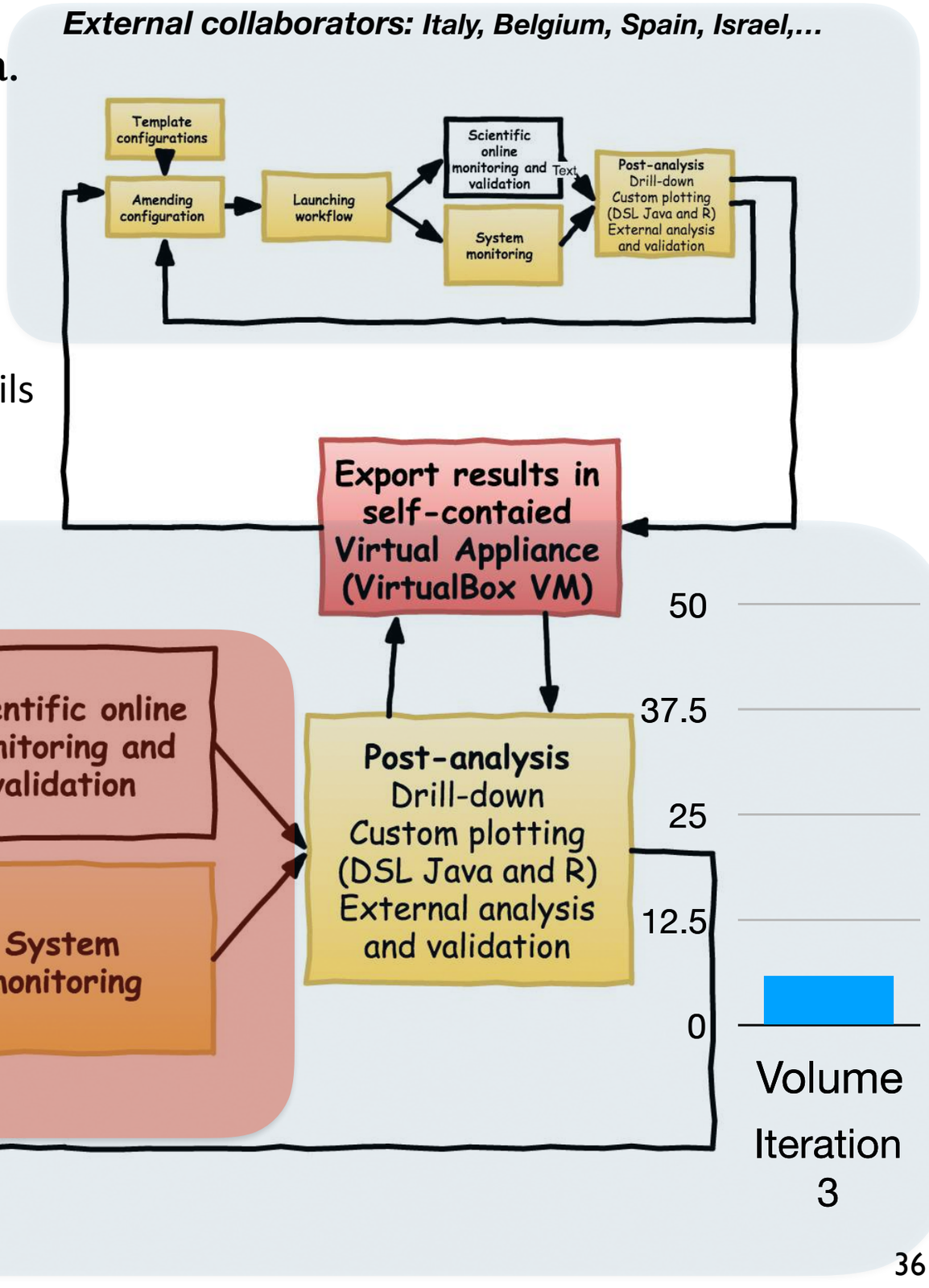
Workflow Scaling - PG as an appliance

- Distributed iterative process, repeated on small selections or samples of sources outside Geneva.
  - Too few resources to re-run, used for results analysis and tagging via Visualisation tool mostly.
  - Virtual Appliances issued daily at some point.
  - 5GB of plots generated, 100s of them with 1000s of details
  - Very intense communication



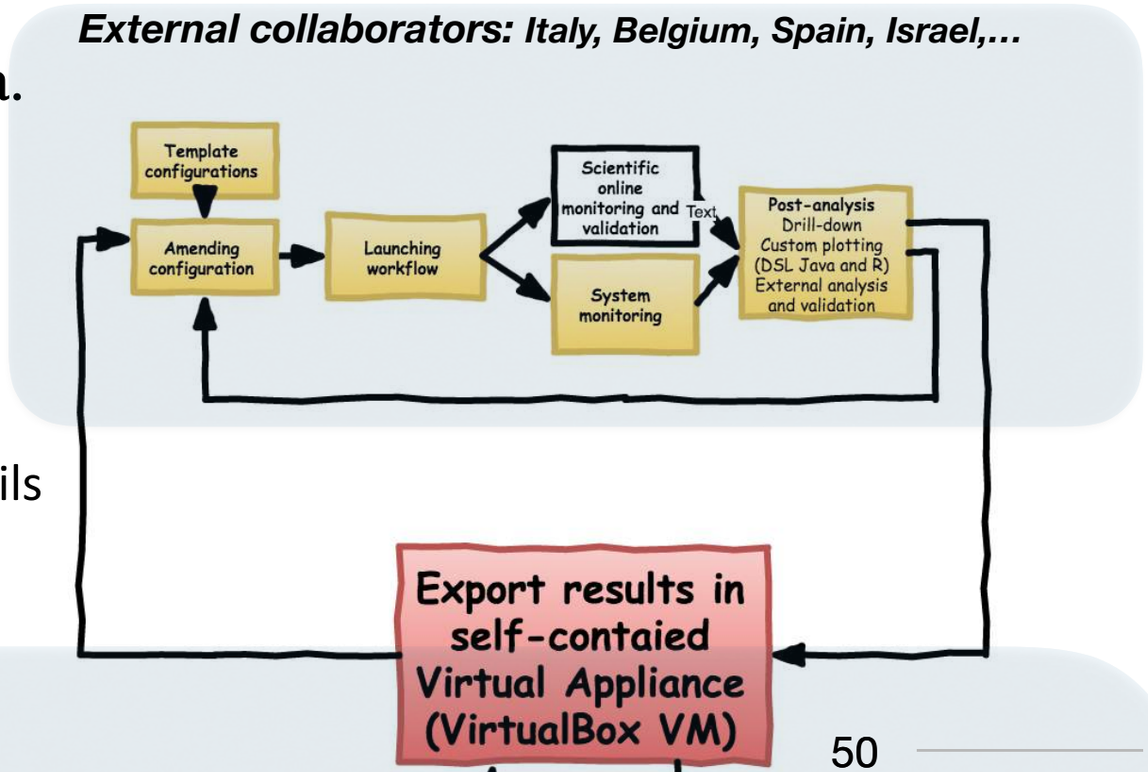
Workflow Scaling - PG as an appliance

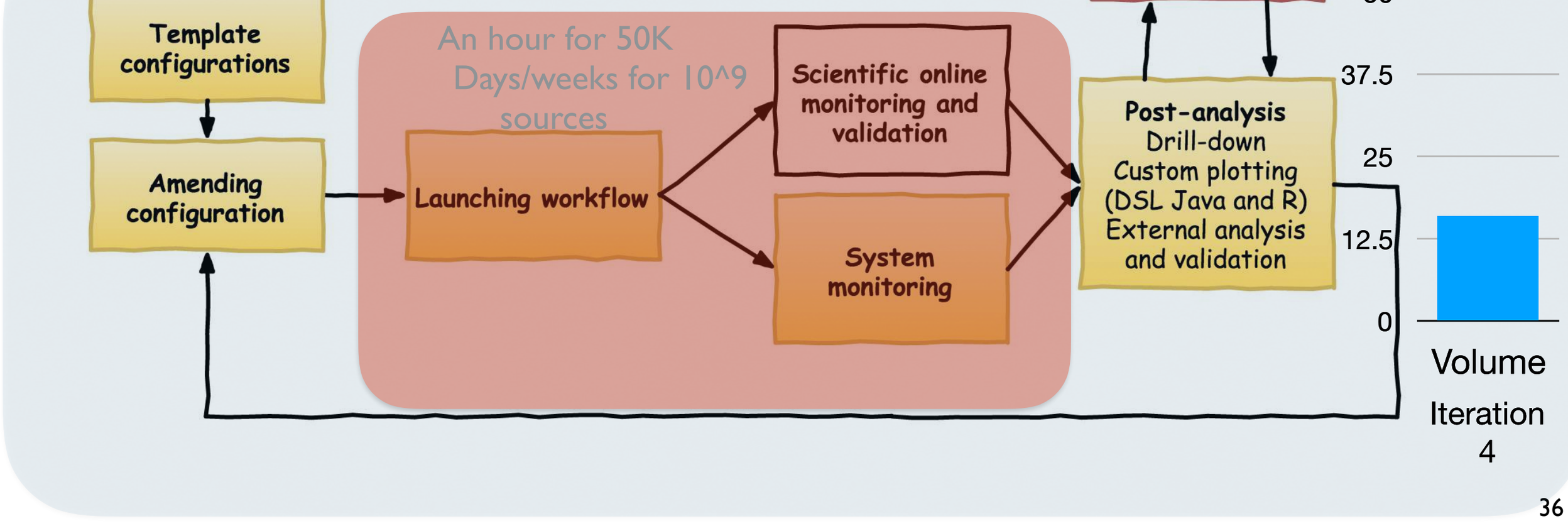
- Distributed iterative process, repeated on small selections or samples of sources outside Geneva.
  - Too few resources to re-run, used for results analysis and tagging via Visualisation tool mostly.
  - Virtual Appliances issued daily at some point.
  - 5GB of plots generated, 100s of them with 1000s of details
  - Very intense communication



Workflow Scaling - PG as an appliance

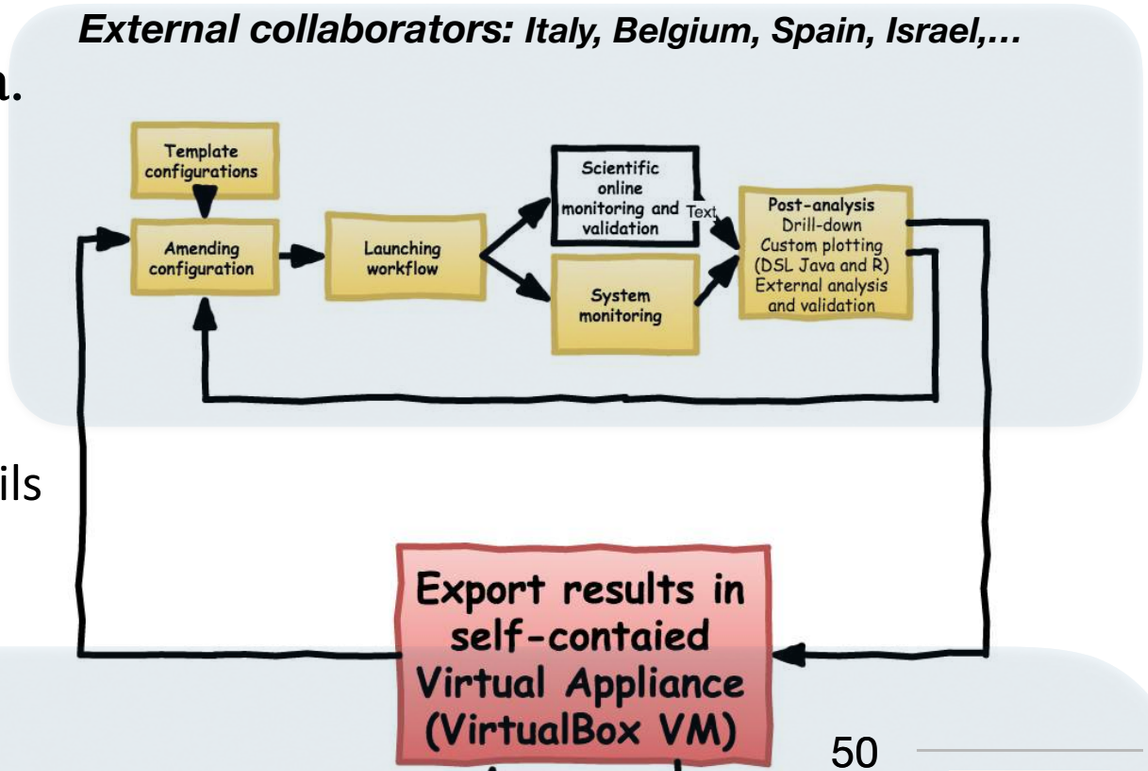
- Distributed iterative process, repeated on small selections or samples of sources outside Geneva.
  - Too few resources to re-run, used for results analysis and tagging via Visualisation tool mostly.
  - Virtual Appliances issued daily at some point.
  - 5GB of plots generated, 100s of them with 1000s of details
  - Very intense communication

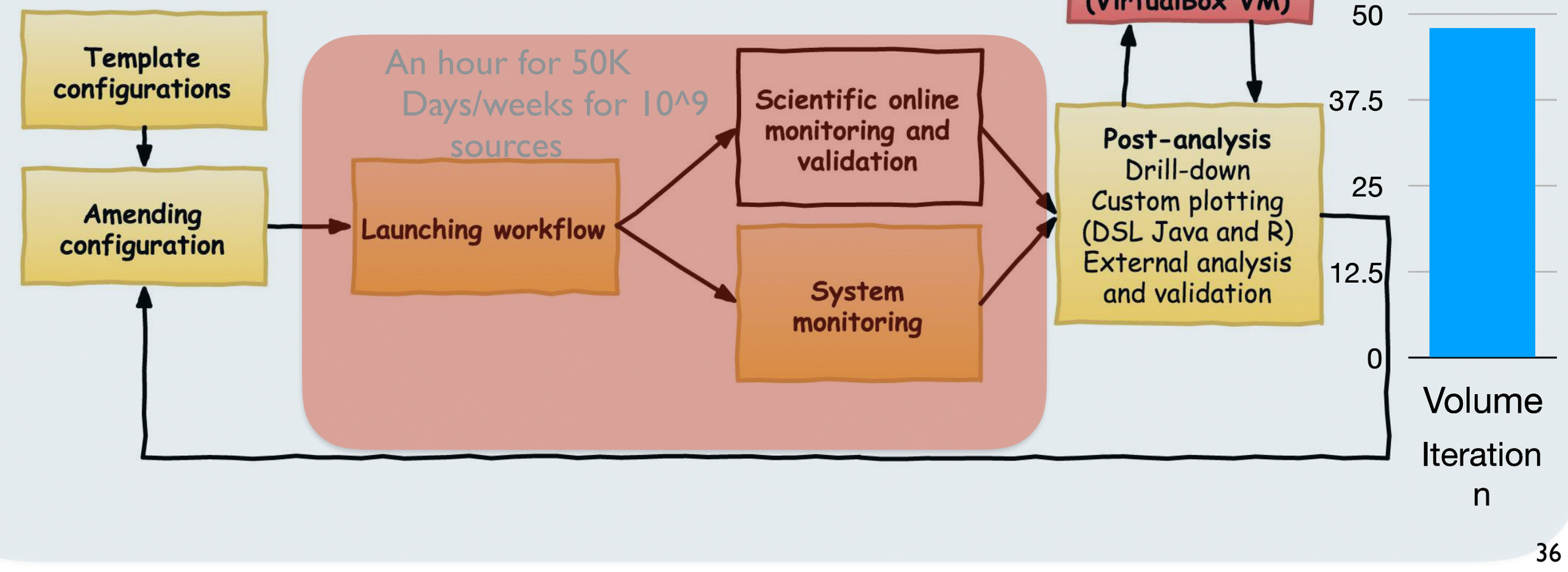




Workflow Scaling - PG as an appliance

- Distributed iterative process, repeated on small selections or samples of sources outside Geneva.
  - Too few resources to re-run, used for results analysis and tagging via Visualisation tool mostly.
  - Virtual Appliances issued daily at some point.
  - 5GB of plots generated, 100s of them with 1000s of details
  - Very intense communication





Postgres for science infrastructure

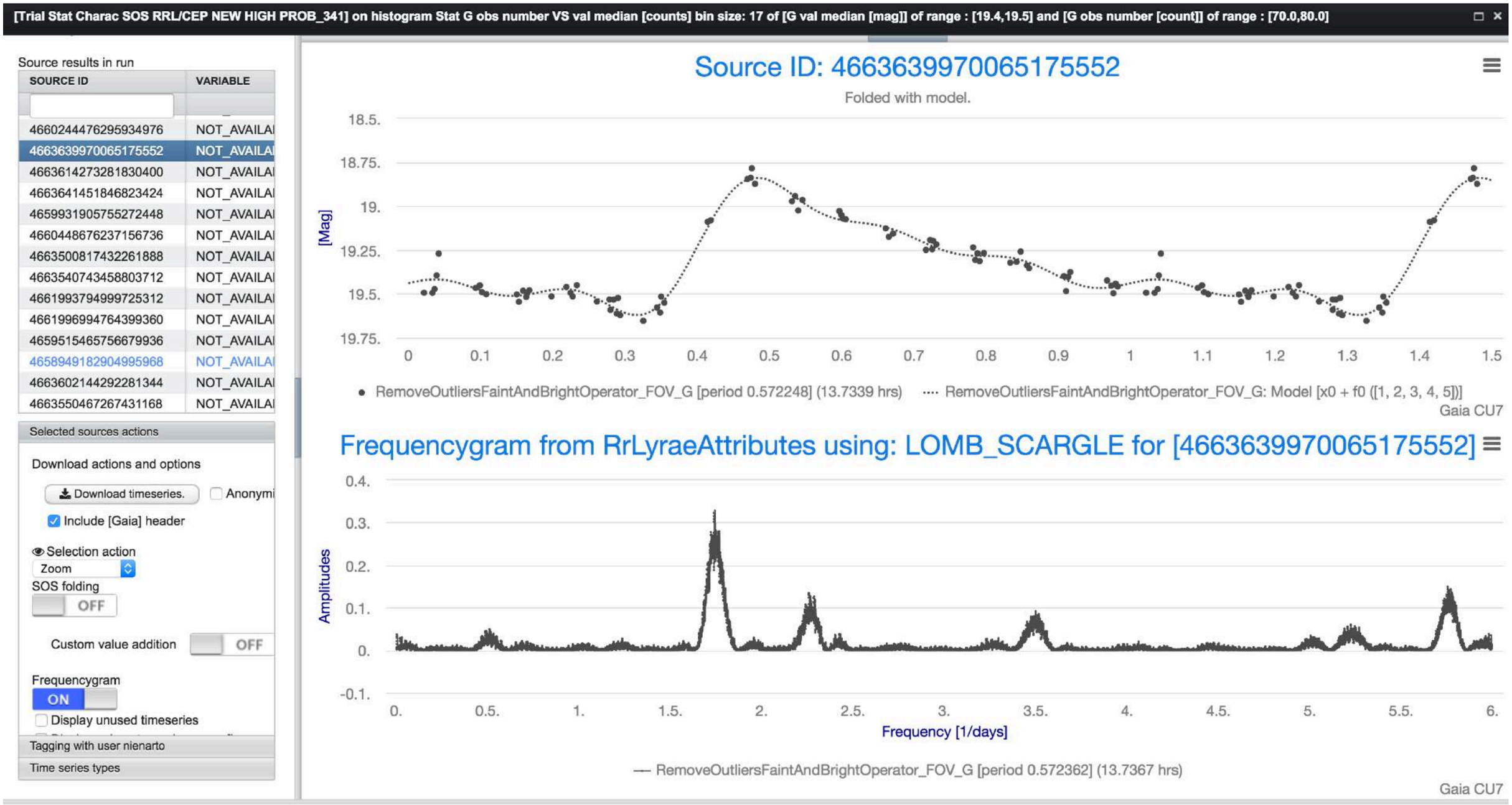
- All our Collaborative tools run on Postgres(-XL)
  - Mattermost (Free Slack alternative) (PG)
  - Owncloud (Free Dropbox alternative) (PG)
  - VariDashboard DPCG (being integrated with both above) (PG-XL)

Postgres for science infrastructure

- All our Collaborative tools run on Postgres(-XL)
  - Mattermost (Free Slack alternative) (PG)
  - Owncloud (Free Dropbox alternative) (PG)
  - VariDashboard DPCG (being integrated with both above) (PG-XL)
- OK, we have ELK deployed as well…

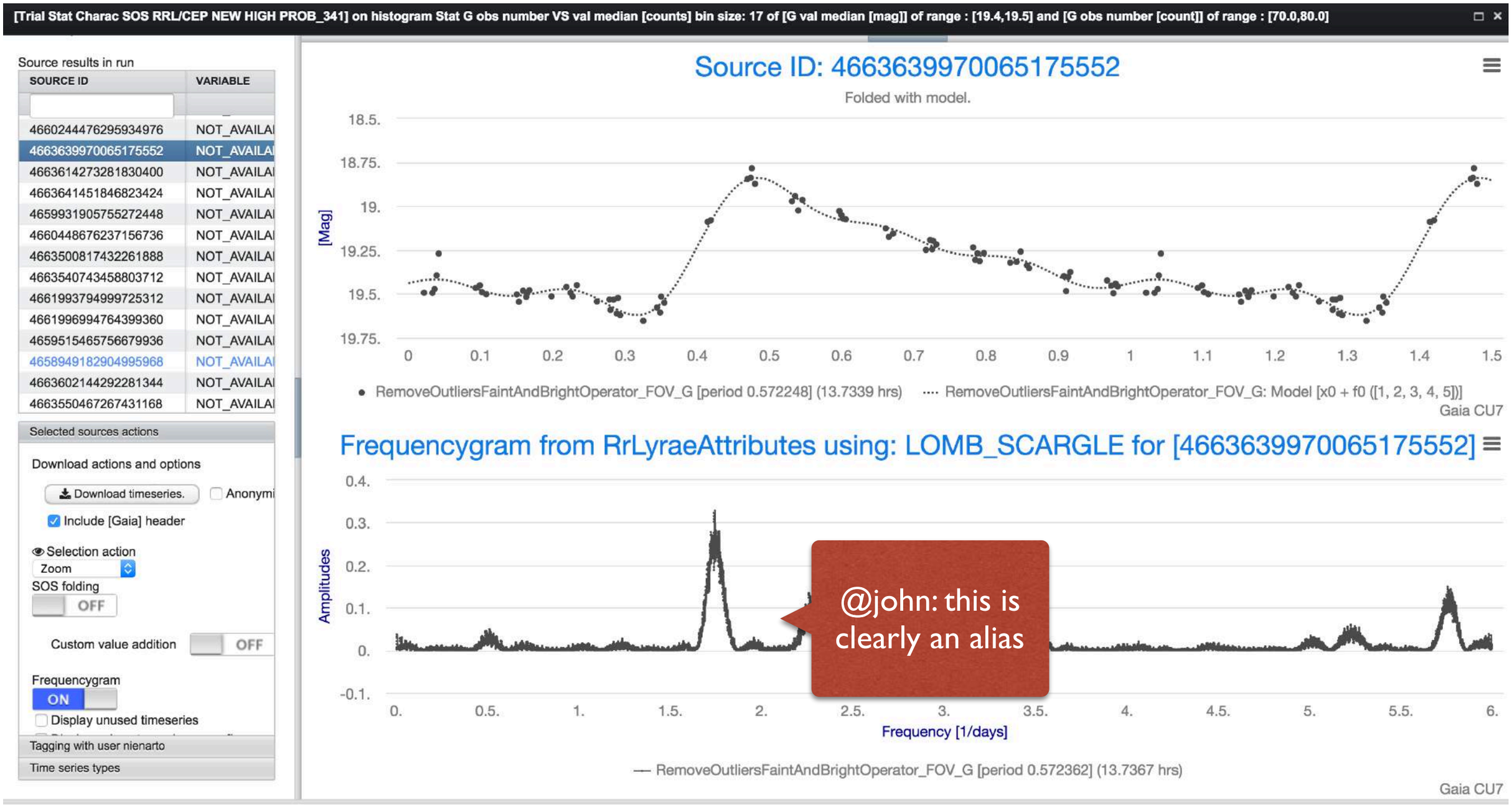
Capturing soft knowledge..

- Tags as organic part of a data model
  - The more of the soft-knowledge
    - tags, discussions, annotations, both verbal and visual…



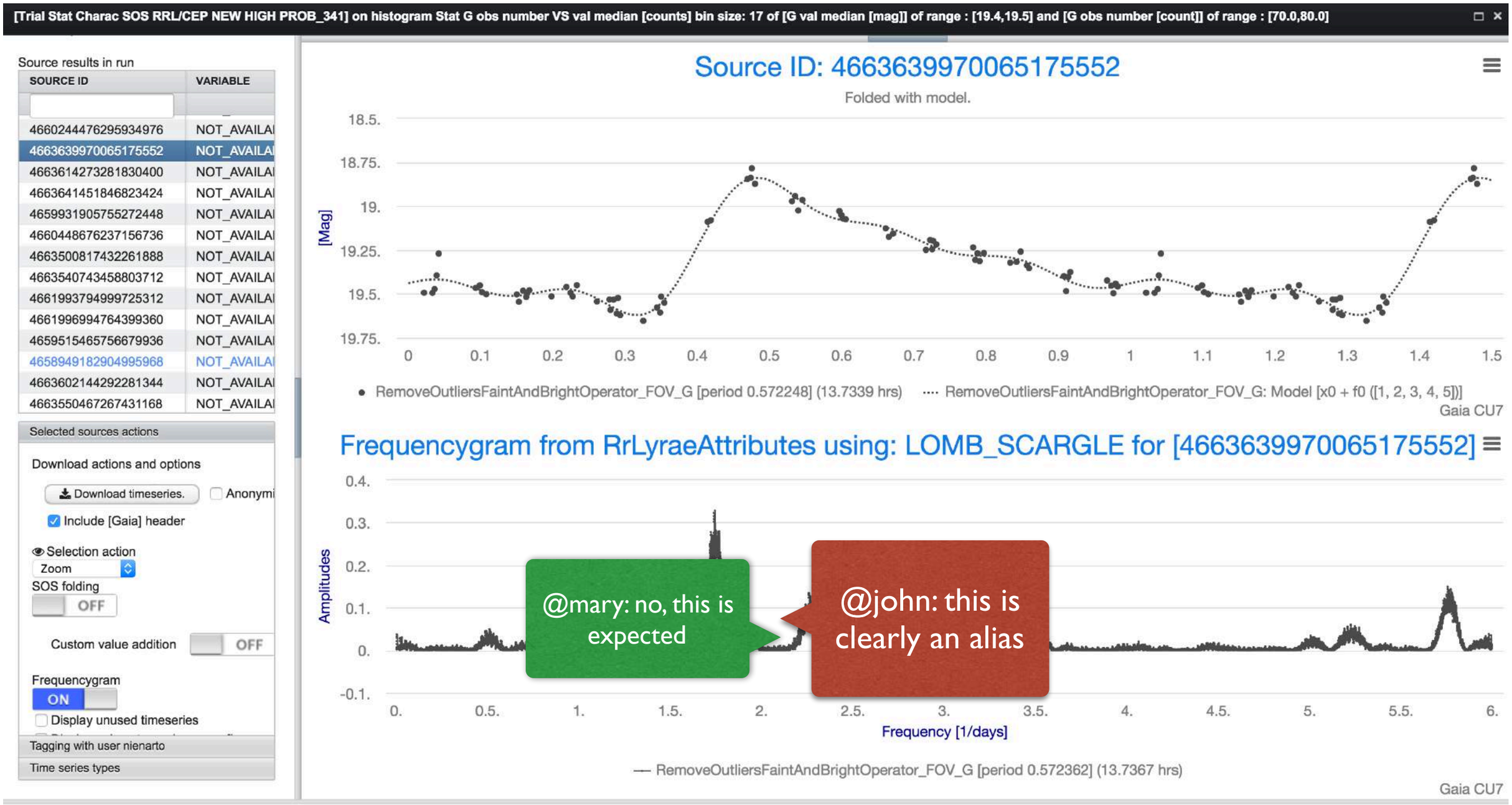
Capturing soft knowledge..

- Tags as organic part of a data model
  - The more of the soft-knowledge
    - tags, discussions, annotations, both verbal and visual…



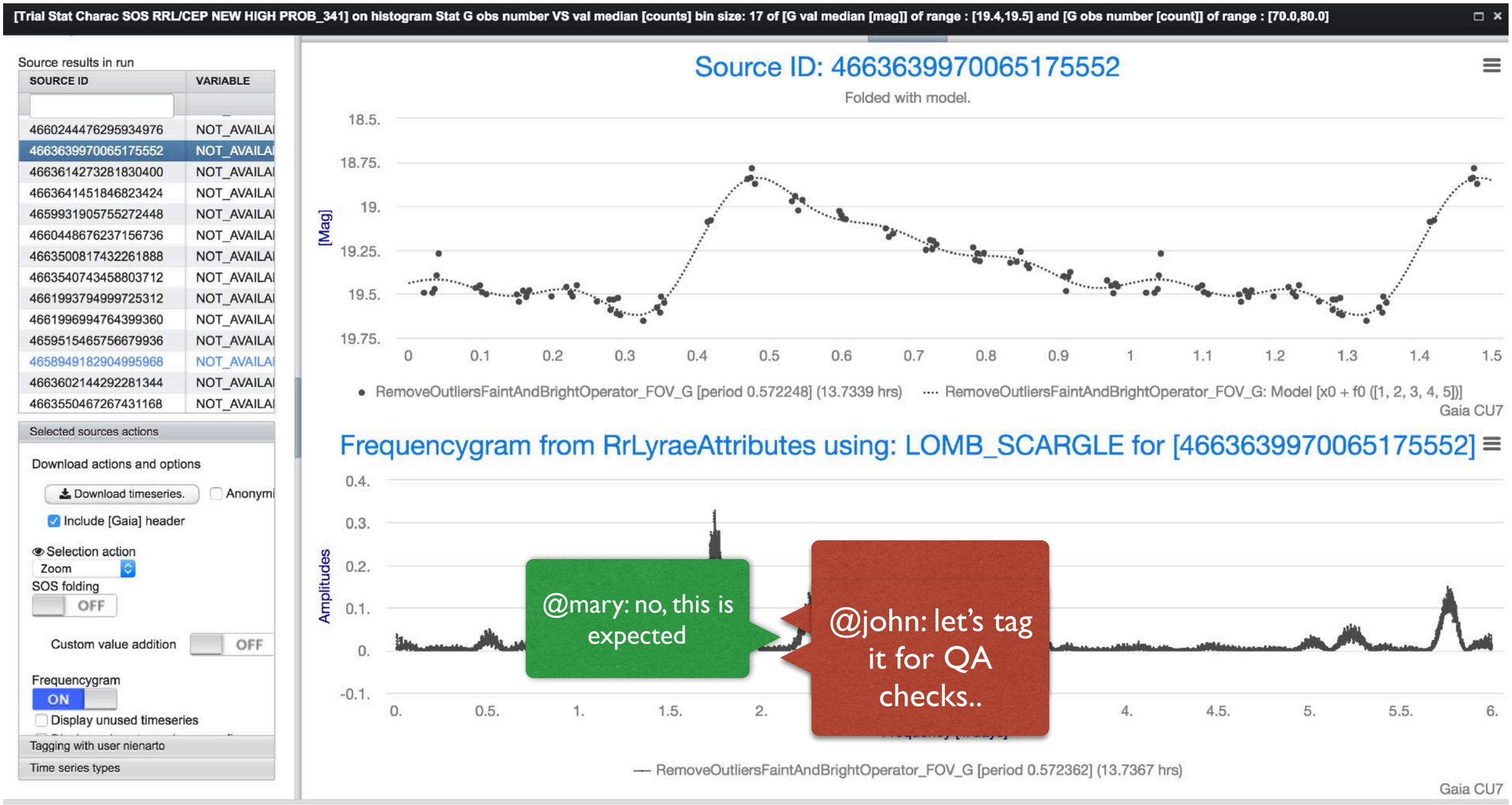
Capturing soft knowledge..

- Tags as organic part of a data model
  - The more of the soft-knowledge
    - tags, discussions, annotations, both verbal and visual…



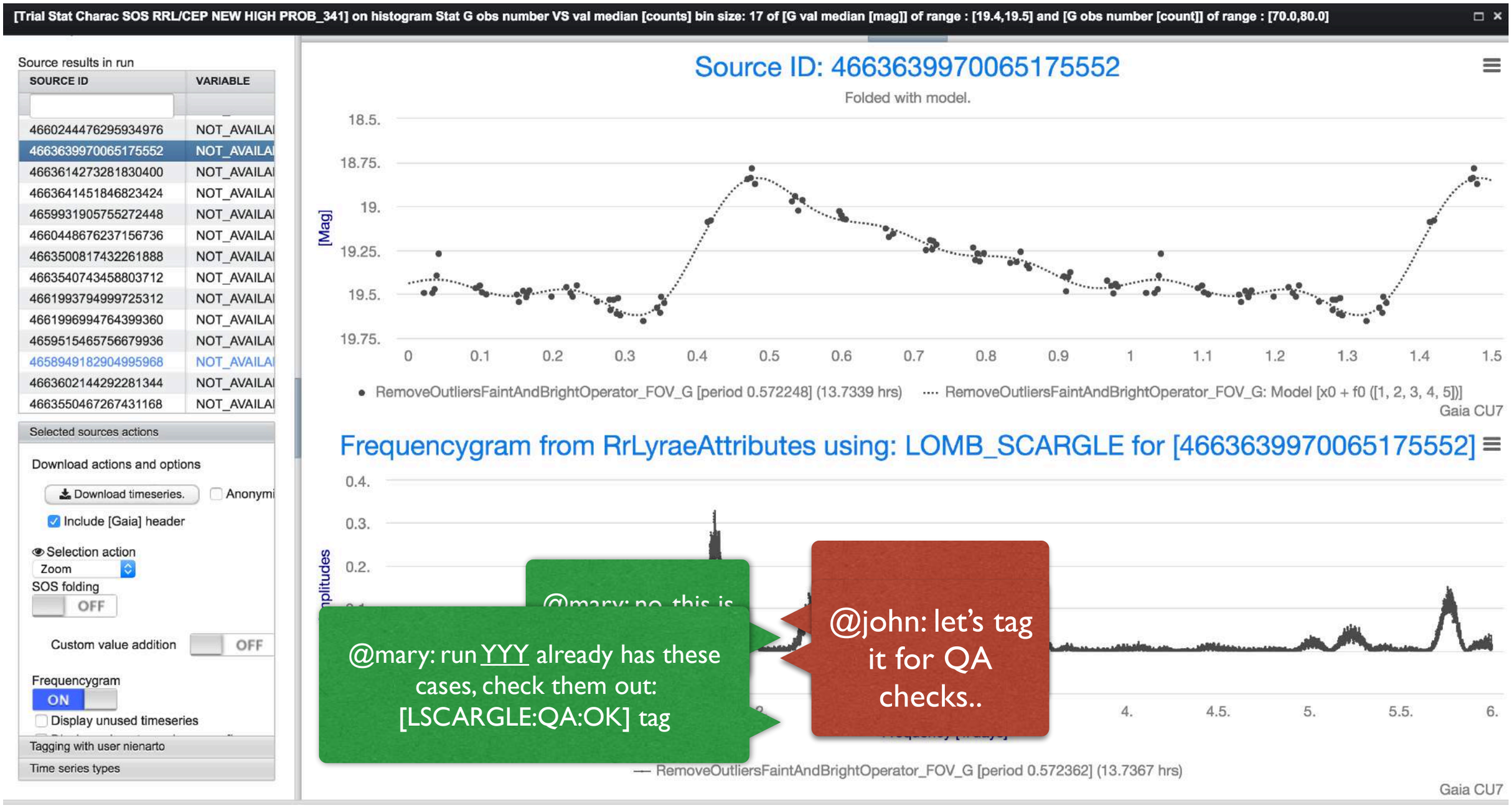
Capturing soft knowledge..

- Tags as organic part of a data model
  - The more of the soft-knowledge
    - tags, discussions, annotations, both verbal and visual…



Capturing soft knowledge..

- Tags as organic part of a data model
  - The more of the soft-knowledge
    - tags, discussions, annotations, both verbal and visual…



Validation via Analytics, Visualisation

Validation via Analytics, Visualisation

How to reduce 80 x 4 billion records to a single screen?

Validation via Analytics, Visualisation

How to reduce 80 x 4 billion records to a single screen?

How to show 500+ derived values from each of 1.5 billion sources, 6 billion timeseries from 100Ks of CPU hours?

Validation via Analytics, Visualisation

How to reduce 80 x 4 billion records to a single screen?

How to show 500+ derived values from each of 1.5 billion sources, 6 billion timeseries from 100Ks of CPU hours?

How to correlate results with calibrated data from Spacecraft?

Validation via Analytics, Visualisation

How to reduce 80 x 4 billion records to a single screen?

How to show 500+ derived values from each of 1.5 billion sources, 6 billion timeseries from 100Ks of CPU hours?

How to correlate results with calibrated data from Spacecraft?

How to not get lost in all generations of data?

Validation via Analytics, Visualisation

How to reduce 80 x 4 billion records to a single screen?

How to show 500+ derived values from each of 1.5 billion sources, 6 billion timeseries from 100Ks of CPU hours?

How to correlate results with calibrated data from Spacecraft?

How to not get lost in all generations of data?

And plentitude of software versions?

Validation via Analytics, Visualisation

How to reduce 80 x 4 billion records to a single screen?

How to show 500+ derived values from each of 1.5 billion sources, 6 billion timeseries from 100Ks of CPU hours?

How to correlate results with calibrated data from Spacecraft?

How to not get lost in all generations of data?

And plentitude of software versions?

How to:

Validation via Analytics, Visualisation

How to reduce 80 x 4 billion records to a single screen?

How to show 500+ derived values from each of 1.5 billion sources, 6 billion timeseries from 100Ks of CPU hours?

How to correlate results with calibrated data from Spacecraft?

How to not get lost in all generations of data?

And plentitude of software versions?

How to: Inspect

Validation via Analytics, Visualisation

How to reduce 80 x 4 billion records to a single screen?

How to show 500+ derived values from each of 1.5 billion sources, 6 billion timeseries from 100Ks of CPU hours?

How to correlate results with calibrated data from Spacecraft?

How to not get lost in all generations of data?

And plentitude of software versions?

How to: Inspect

Validate

Validation via Analytics, Visualisation

How to reduce 80 x 4 billion records to a single screen?

How to show 500+ derived values from each of 1.5 billion sources, 6 billion timeseries from 100Ks of CPU hours?

How to correlate results with calibrated data from Spacecraft?

How to not get lost in all generations of data?

And plentitude of software versions?

How to: Inspect

Validate

Act on unexpected

Structure

- Story of perpetual change
- Databases in Astronomy
- Gaia mission
- Gaia processing at CU7/DPC Geneva
- Postgres for science
- **Postgres-XL tale**
- Collaboration
- Future

Postgres-XL - Scalability

- Started by Koichi Suzuki at NTT@Japan as Postgres-XC
  - Coordinators and datanodes.
  - Evolved into Postgres-X2
- Postgres-XL: fork of XC with stress on robustness
  - Some changes in the architecture, introduction of shared queues for execution of queries in **scatter-gather** pattern
- Most of the recent work done by Pavan Deolasee and Tomas Vondra of 2ndQuadrant

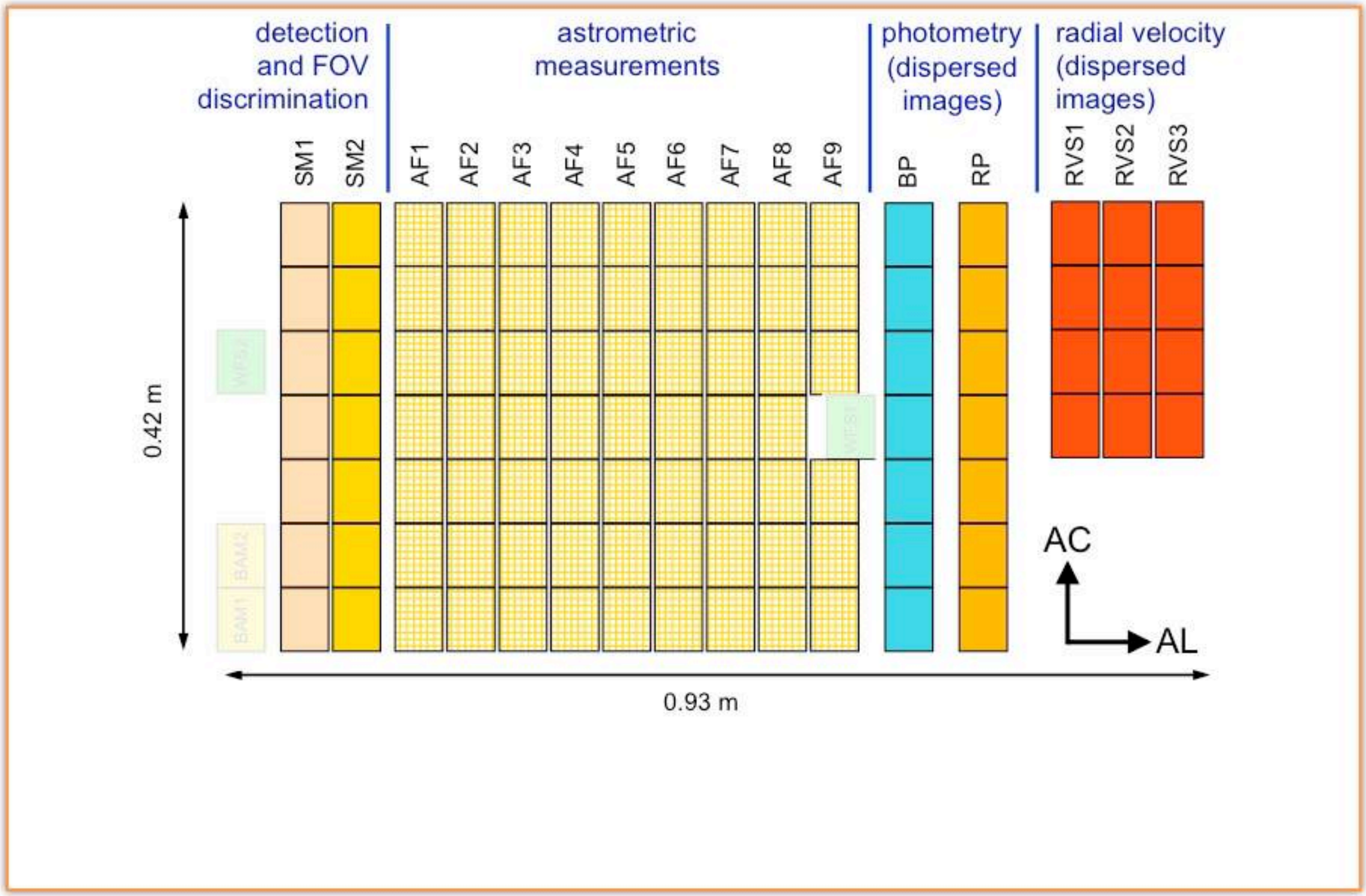
Postgres-XL - our philosophy

Postgres-XL - our philosophy

- Gaia perspective on Open source community with esp. 2ndQuadrant support
  - ***Do ut des***
    - *I give that you may give…*
      - *Belief that byproduct of the public science should be generic Open Source with mutual gain.*
      - *A win-win, but at a cost, as:*

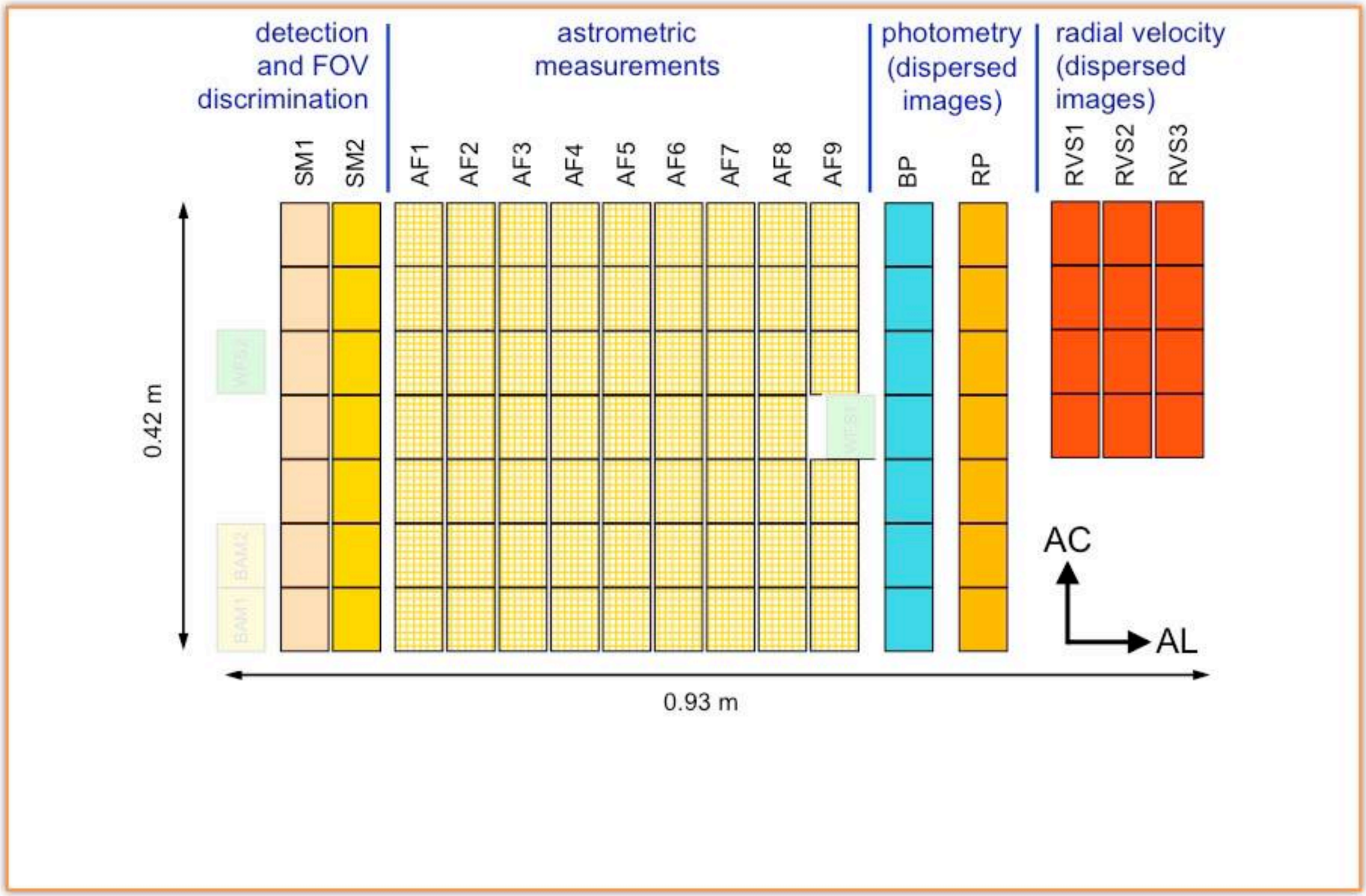
- Gaia is the **extreme** use-case for any data management platform in the academic frame and would need special care whatever the platform used..

CDD, BP RP, RVS observations



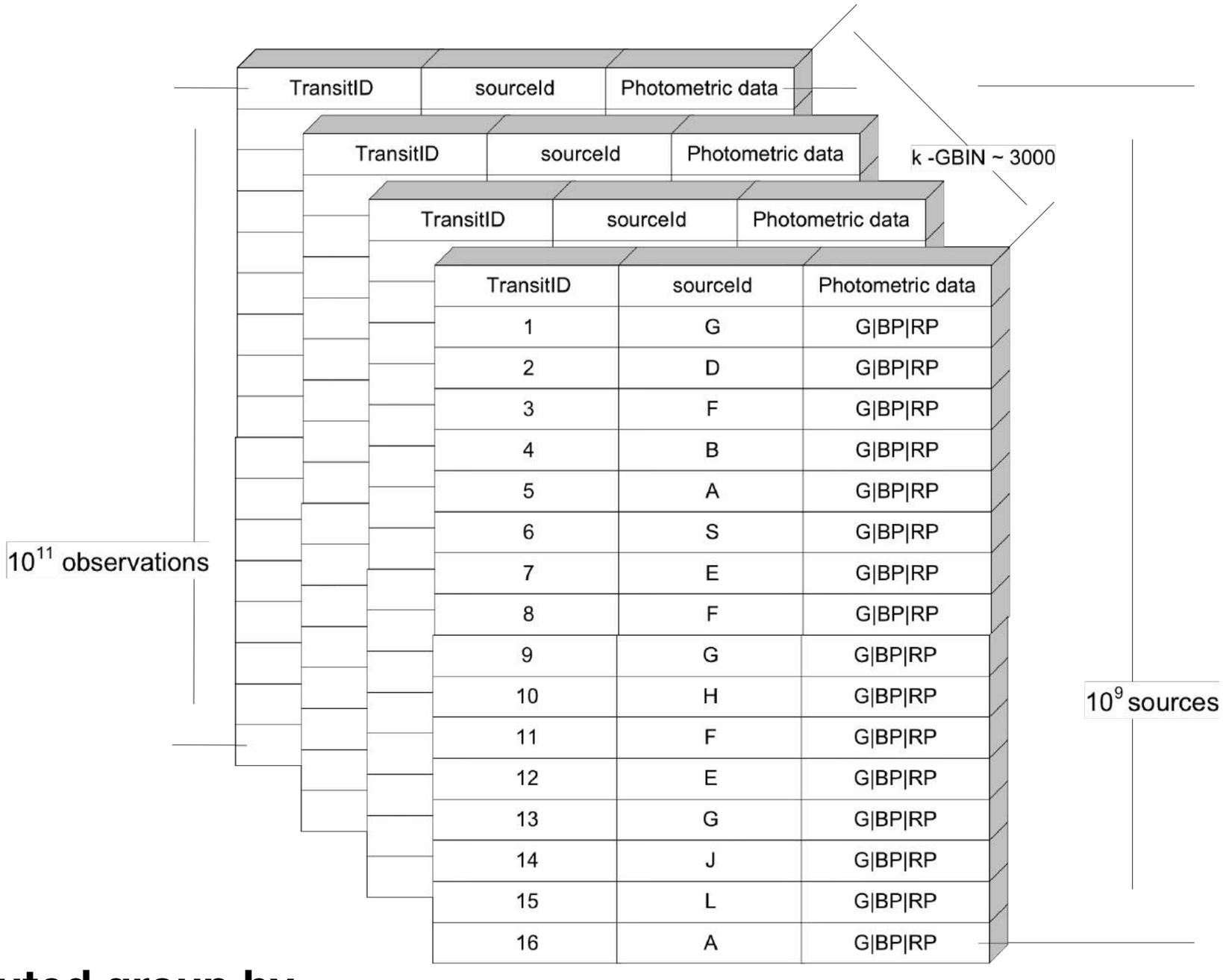
Courtesy: Berry Holl

CDD, BP RP, RVS observations



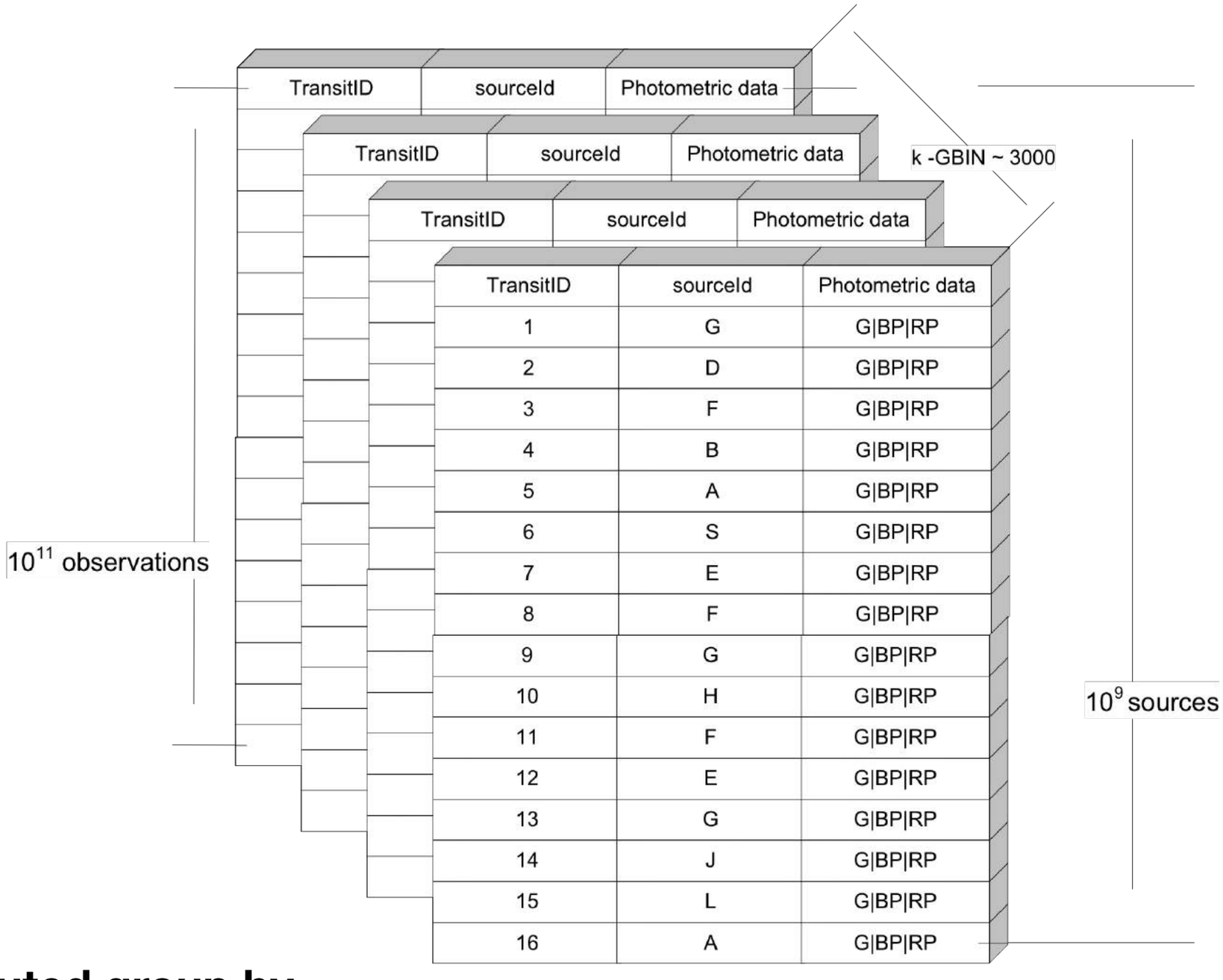
Courtesy: Berry Holl

Data mapping, photometry reconstruction



Distributed group by

Data mapping, photometry reconstruction



Distributed group by

Data mapping, photometry reconstruction

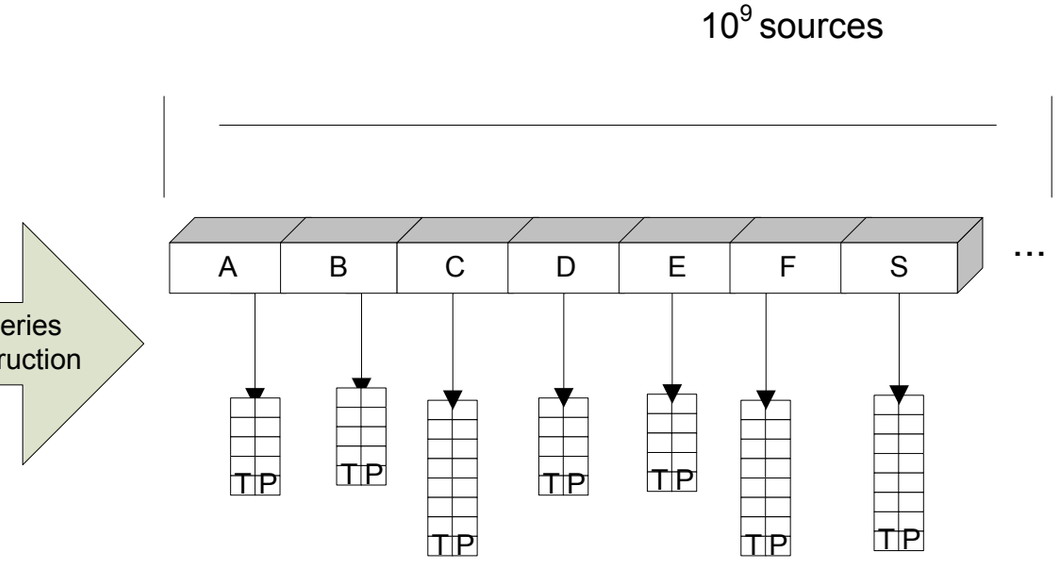
TransitID	sourceId	Photometric data
1	G	G|BP|RP
2	D	G|BP|RP
3	F	G|BP|RP
4	B	G|BP|RP
5	A	G|BP|RP
6	S	G|BP|RP
7	E	G|BP|RP
8	F	G|BP|RP
9	G	G|BP|RP
10	H	G|BP|RP
11	F	G|BP|RP
12	E	G|BP|RP
13	G	G|BP|RP
14	J	G|BP|RP
...	...	G|BP|RP
10^11	10^9	G|BP|RP

10^{11} observations

10^9 sources

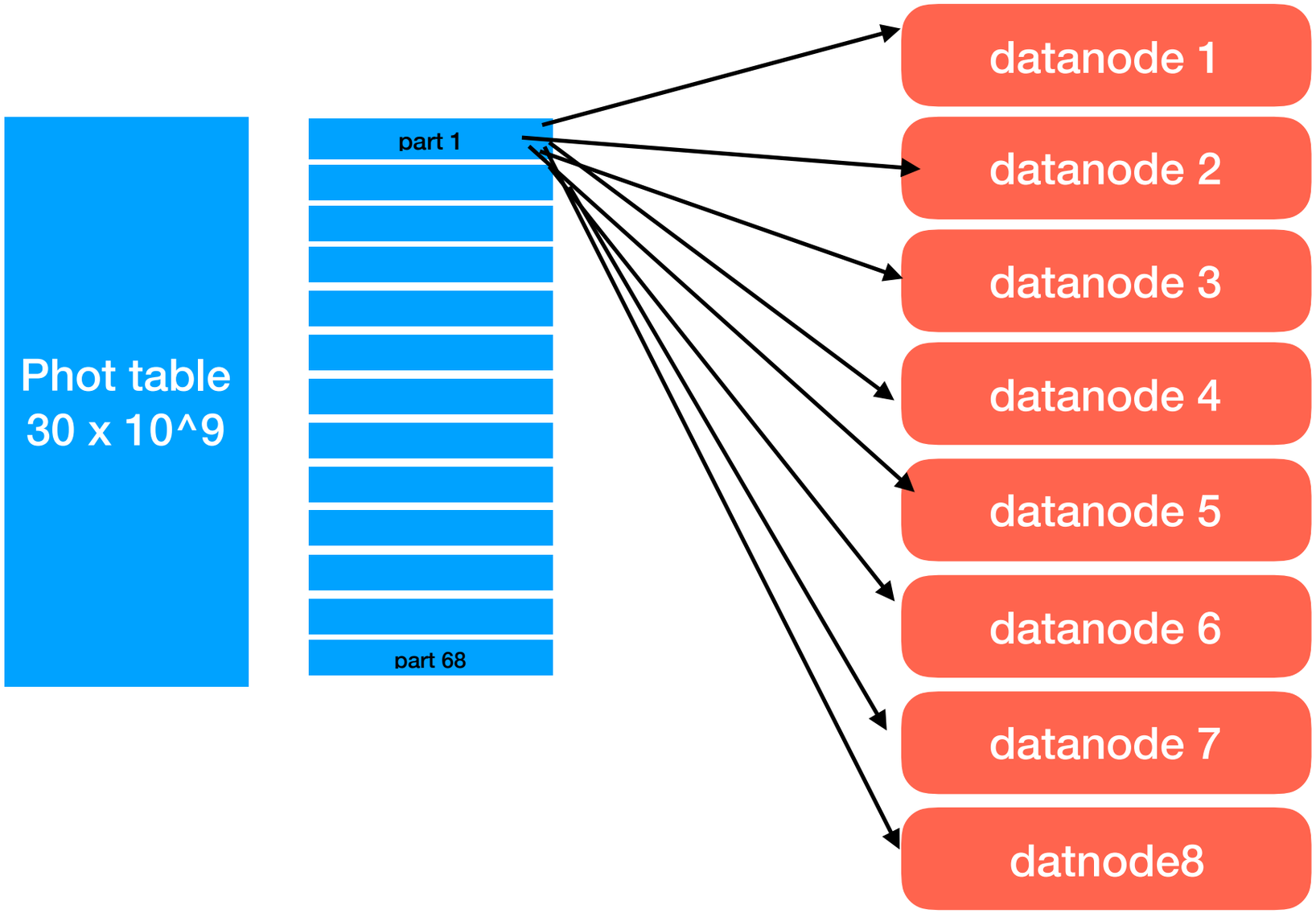
Timeseries reconstruction

10^9 sources



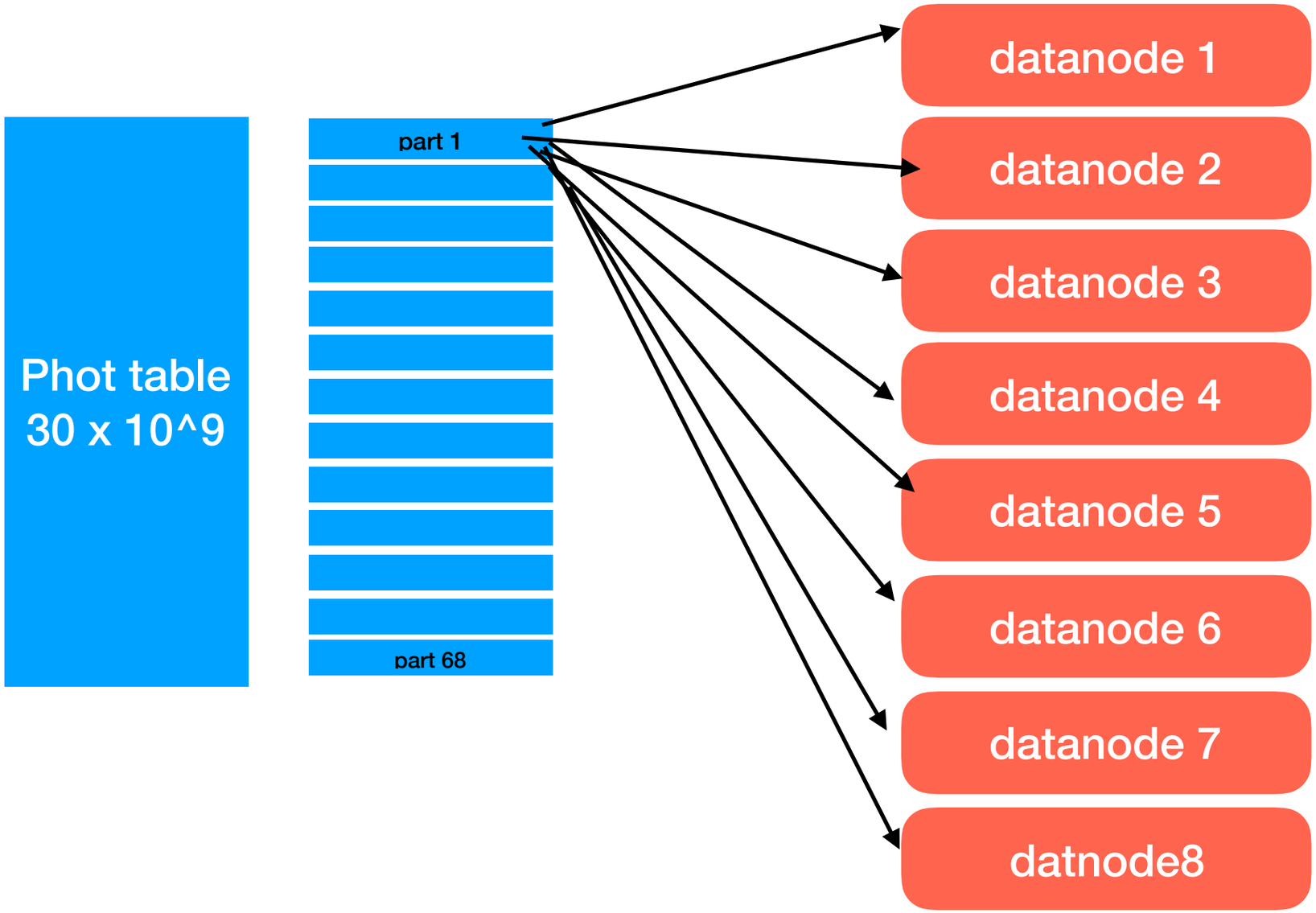
Distributed group by

Linear scalability



Distributed **group by** at each partition.
Arbitrary chosen 68 partitions by sourceid.
By scattering load on all the cluster
we can get linear scalability, **100x** faster than with a naive approach.
Takes 12 hours for 15TB of DB volume generated

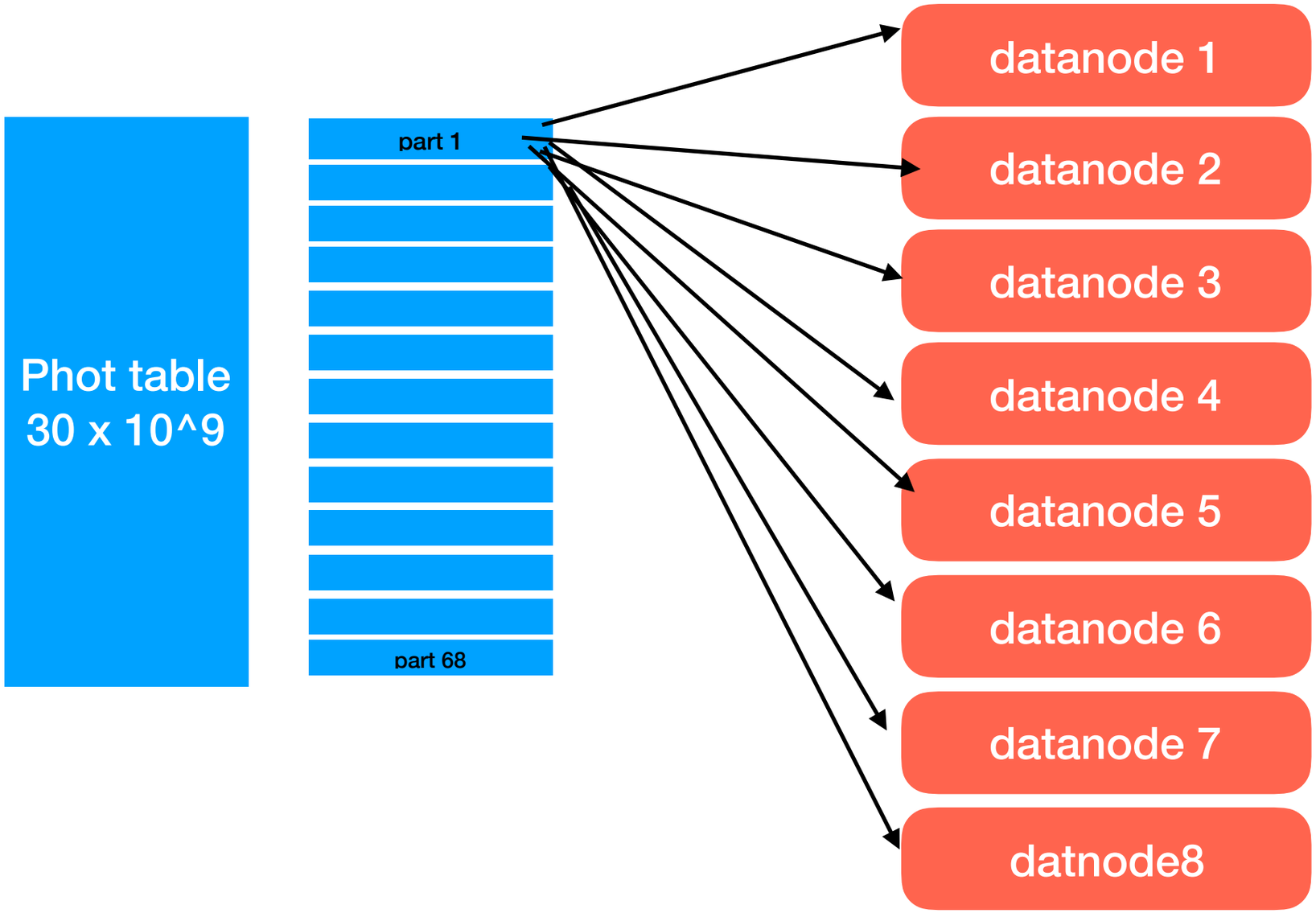
Linear scalability



Total 130TB now

Distributed **group by** at each partition.
Arbitrary chosen 68 partitions by sourceid.
By scattering load on all the cluster
we can get linear scalability, **100x** faster than with a naive approach.
Takes 12 hours for 15TB of DB volume generated

Linear scalability

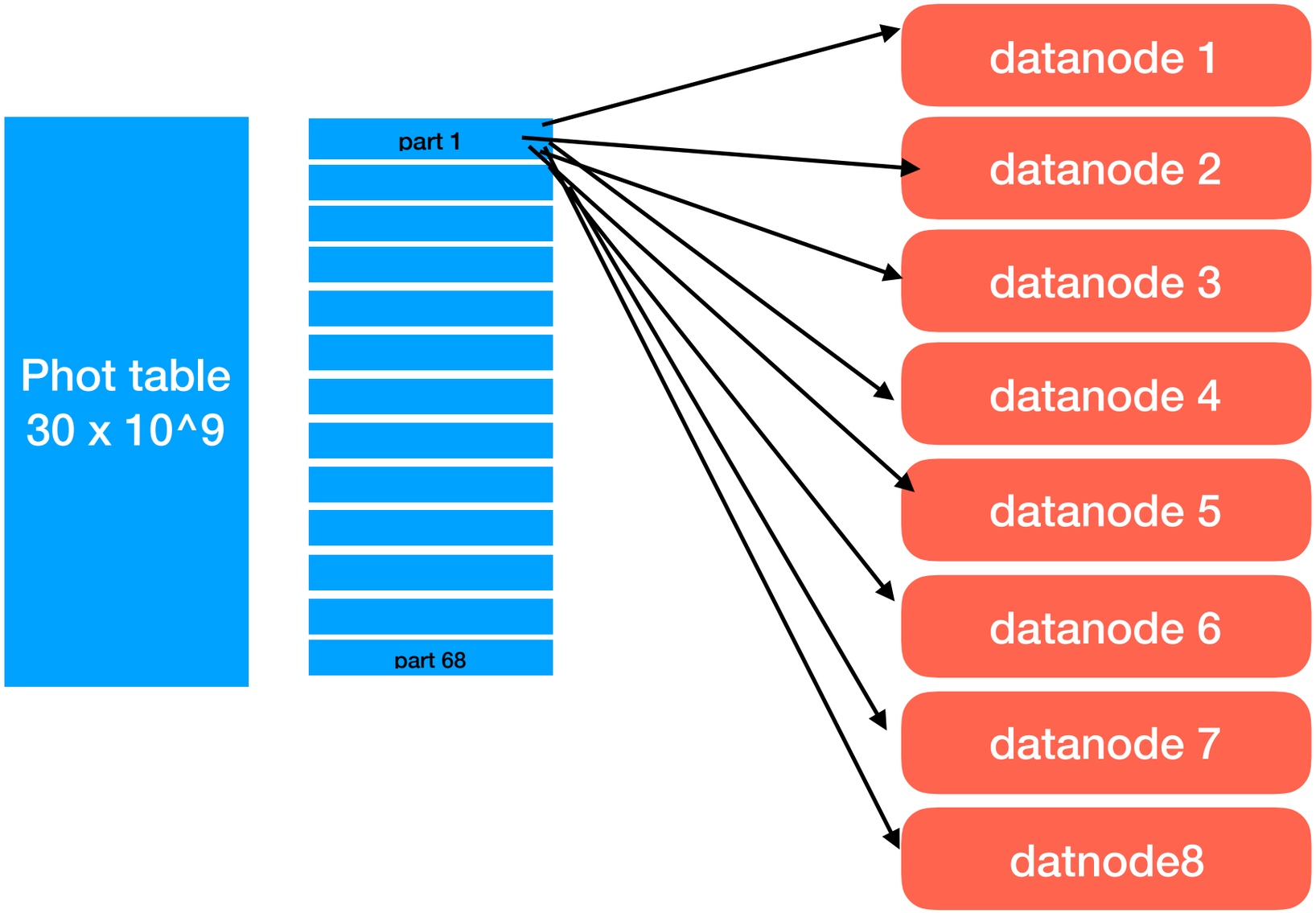


Total 130TB now

60TB in live tables

Distributed **group by** at each partition.
Arbitrary chosen 68 partitions by sourceid.
By scattering load on all the cluster
we can get linear scalability, **100x** faster than with a naive approach.
Takes 12 hours for 15TB of DB volume generated

Linear scalability



Total 130TB now

60TB in live tables

Will double every year..

Distributed **group by** at each partition.
Arbitrary chosen 68 partitions by sourceid.
By scattering load on all the cluster
we can get linear scalability, **100x** faster than with a naive approach.
Takes 12 hours for 15TB of DB volume generated

Linear scalability

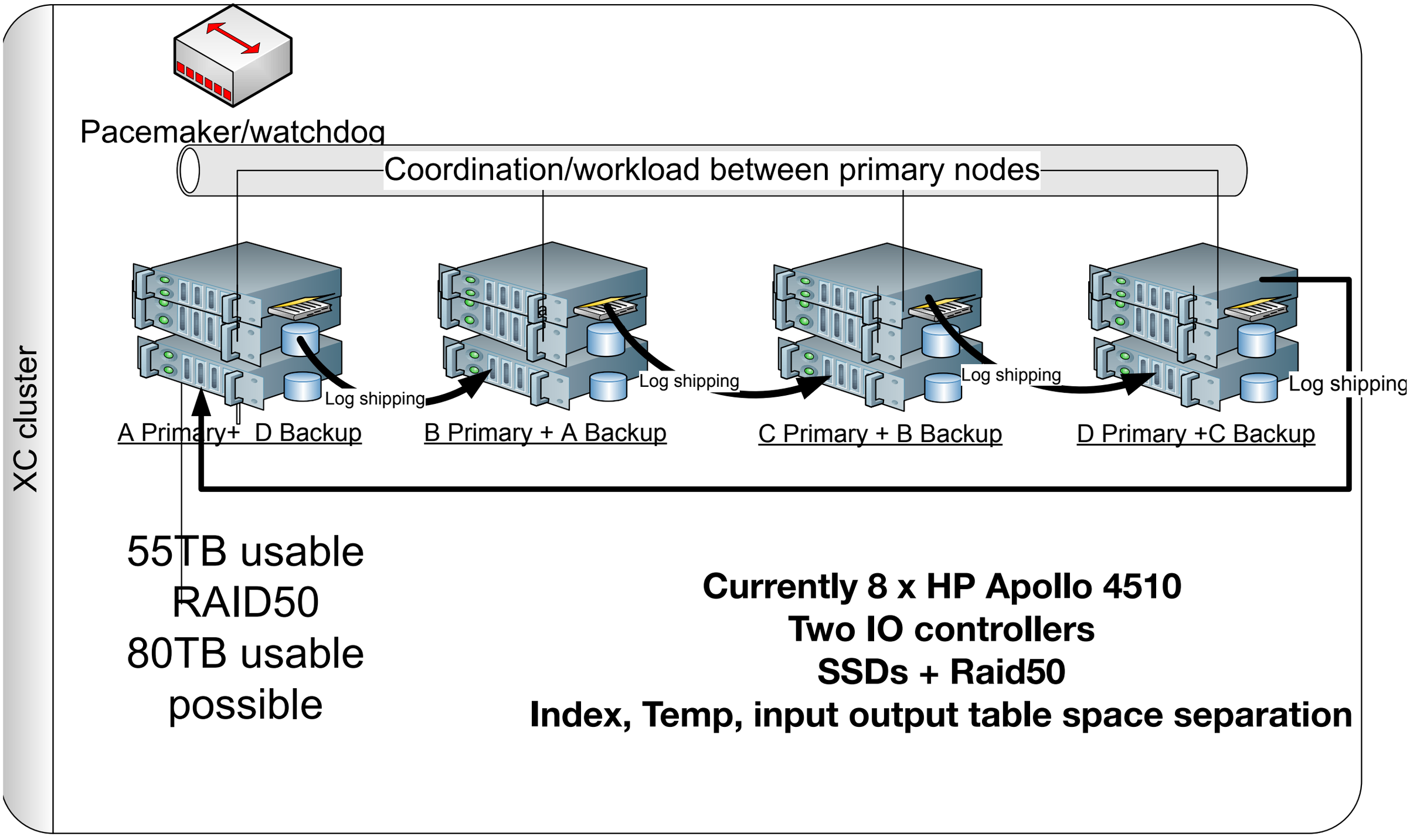
```
psql -AtXq  -U xl_dpac_c2  -d surveys -c "
select E'set work_mem=''4GB''';
set search_path to xl_dpac_c2,xl_dpac_c2_part,xl_dpac_c2_mdb, public;
SET DEFAULT_TRANSACTION_ISOLATION TO ''REPEATABLE READ'';
explain insert into xl_dpac_c2.ts (catalogid, sourceid, fvaluetype, omtype, obstimes, ftimeseriestype, vals, valserr, flags, transitids, length)
select xl_dpac_c2.getcatalogId(''GAIA_C2_ALL'') catalogid,
(a).sourceid sourceid,
0 as fvaluetype,
  ''P'' as omtype,
    convertFromObmtToTcbTime( obstimes, (a).sourceid, ''TCB''::text, 1577880000000000000 ,
        (a).alpha, (a).alphastarerror, (a).delta, (a).deltaerror, (a).varpi, (a).varpierror, (a).mualphastar, (a).mualphastarerror, (a).mudelta, (a).mudeltaerror, (a).r
    tstype,
val::bytea,
valerr::bytea,
flag::bytea,
transitid::bytea,
cardinality(transitid)
from (
select  (cu5.sourceid, alpha, alphastarerror, delta, deltaerror, varpi, varpierror, mualphastar, mualphastarerror, mudelta, mudeltaerror, radialvelocity, radialvelocity
(unnest( ARRAY[
                (xl_dpac_c2.getTsTypeId(''Gaia'',''GAIA_PHOT_G''),    array_agg(transitid order by gobstime) filter (where gobstime>0),  array_agg(gobstime order b
                (xl_dpac_c2.getTsTypeId(''Gaia'',''GAIA_PHOT_BP''),   array_agg(transitid order by bpobstime) filter (where bpobstime>0), array_agg(bpobstime order
                (xl_dpac_c2.getTsTypeId(''Gaia'',''GAIA_PHOT_RP''),   array_agg(transitid order by rpobstime) filter (where rpobstime>0), array_agg(rpobstime order
                (xl_dpac_c2.getTsTypeId(''Gaia'',''GAIA_PHOT_G_CCD''),array_agg(transitid order by smobstime) ,xl_dpac_c2.array_accum_cat(astrotimes_flatter(transi
                      xl_dpac_c2.array_accum_cat(smflux ||  afflux order by smobstime ) , xl_dpac_c2.array_accum_cat(smfluxerror || affluxerror order by smobstime)
                  ])).*
from
xl_dpac_c2_c1_photometry.mdb_cu3_source_019_' || num ||' s join xl_dpac_c2_c1_photometry.mdb_cu5_finalcalphotfovtransit_011_' || num || ' cu5 using(sourceid)
group by 1
order by 1
) as rowls;
#'
    from (select lpad(i::text, 3, '0') num, datnum from generate_series(1,68) i order by i) part " |  xargs --delimiter=# -I {} -n 1 -P 64 psql -1 -d surveys -c "{}"
```

Linear scalability

```
psql -AtXq  -U xl_dpac_c2  -d surveys -c "
select E'set work_mem=''4GB''';
set search_path to xl_dpac_c2,xl_dpac_c2_part,xl_dpac_c2_mdb, public;
SET DEFAULT_TRANSACTION_ISOLATION TO ''REPEATABLE READ'';
explain insert into xl_dpac_c2.ts (catalogid, sourceid, fvaluetype, omtype, obstimes, ftimeseriestype, vals, valserr, flags, transitids, length)
select xl_dpac_c2.getcatalogId(''GAIA_C2_ALL'') catalogid,
(a).sourceid sourceid,
0 as fvaluetype,
  ''P'' as omtype,
    convertFromObmtToTcbTime( obstimes, (a).sourceid, ''TCB''::text, 1577880000000000000 ,
        (a).alpha, (a).alphastarerror, (a).delta, (a).deltaerror, (a).varpi, (a).varpierror, (a).mualphastar, (a).mualphastarerror, (a).mudelta, (a).mudeltaerror, (a).r
    tstype,
val::bytea,
valerr::bytea,
flag::bytea,
transitid::bytea,
cardinality(transitid)
from (
select  (cu5.sourceid, alpha, alphastarerror, delta, deltaerror, varpi, varpierror, mualphastar, mualphastarerror, mudelta, mudeltaerror, radialvelocity, radialvelocity
(unnest( ARRAY[
                (xl_dpac_c2.getTsTypeId(''Gaia'',''GAIA_PHOT_G''),    array_agg(transitid order by gobstime) filter (where gobstime>0),  array_agg(gobstime order b
                (xl_dpac_c2.getTsTypeId(''Gaia'',''GAIA_PHOT_BP''),   array_agg(transitid order by bpobstime) filter (where bpobstime>0), array_agg(bpobstime order
                (xl_dpac_c2.getTsTypeId(''Gaia'',''GAIA_PHOT_RP''),   array_agg(transitid order by rpobstime) filter (where rpobstime>0), array_agg(rpobstime order
                (xl_dpac_c2.getTsTypeId(''Gaia'',''GAIA_PHOT_G_CCD''),array_agg(transitid order by smobstime) ,xl_dpac_c2.array_accum_cat(astrotimes_flatter(transi
                      xl_dpac_c2.array_accum_cat(smflux ||  afflux order by smobstime ) , xl_dpac_c2.array_accum_cat(smfluxerror || affluxerror order by smobstime)
                  ])).*
from
xl_dpac_c2_c1_photometry.mdb_cu3_source_019_' || num ||' s join xl_dpac_c2_c1_photometry.mdb_cu5_finalcalphotfovtransit_011_' || num || ' cu5 using(sourceid)
group by 1
order by 1
) as rowls;
#'
    from (select lpad(i::text, 3, '0') num, datnum from generate_series(1,68) i order by i) part " |  xargs --delimiter=# -I {} -n 1 -P 64 psql -1 -d surveys -c "{}"
```

- Meta-queries generating queries per partition queries executed in parallel
  - Poor-man-parallelism

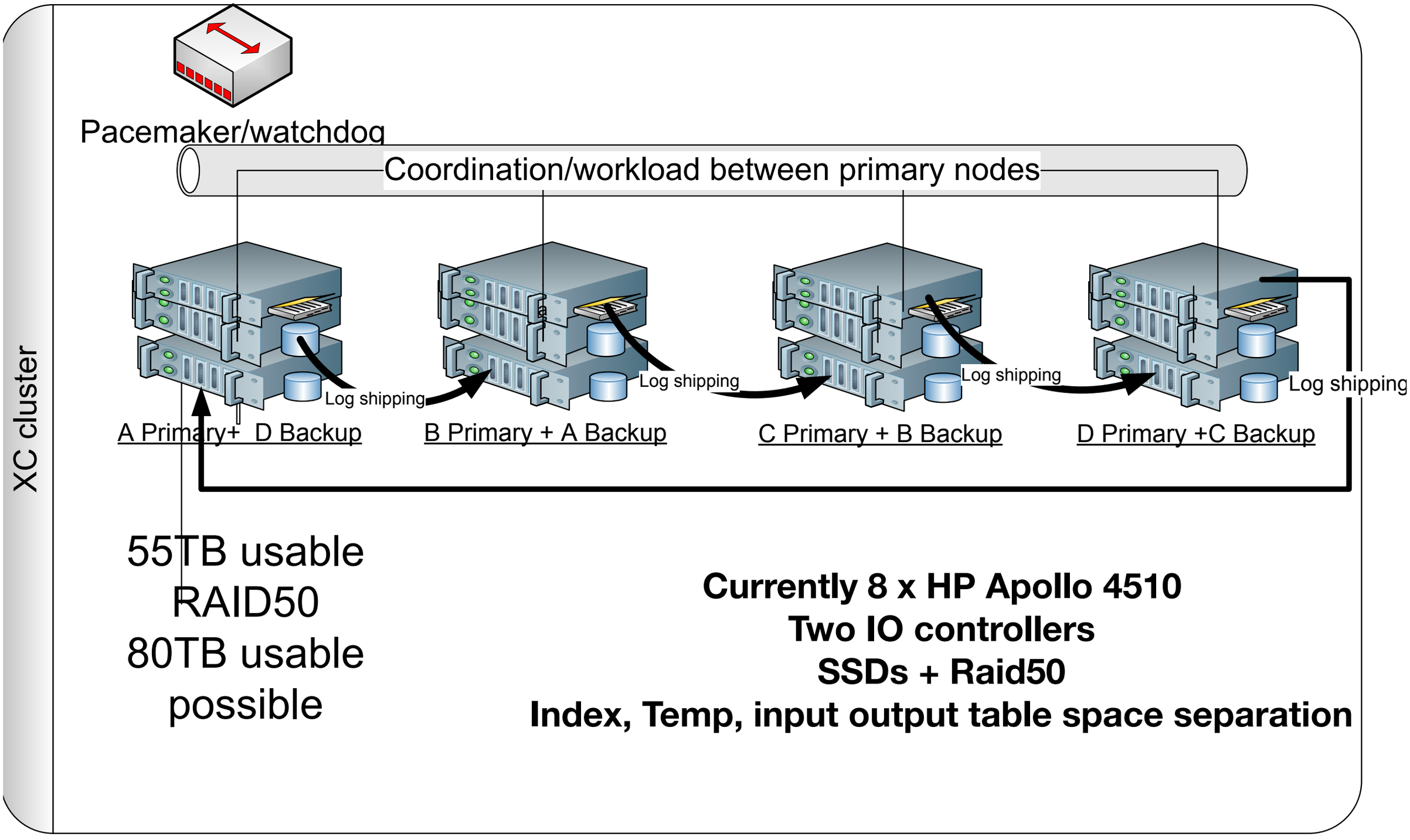
Postgres-XL Journey



55TB usable
RAID50
80TB usable
possible

Currently 8 x HP Apollo 4510
Two IO controllers
SSDs + Raid50
Index, Temp, input output table space separation

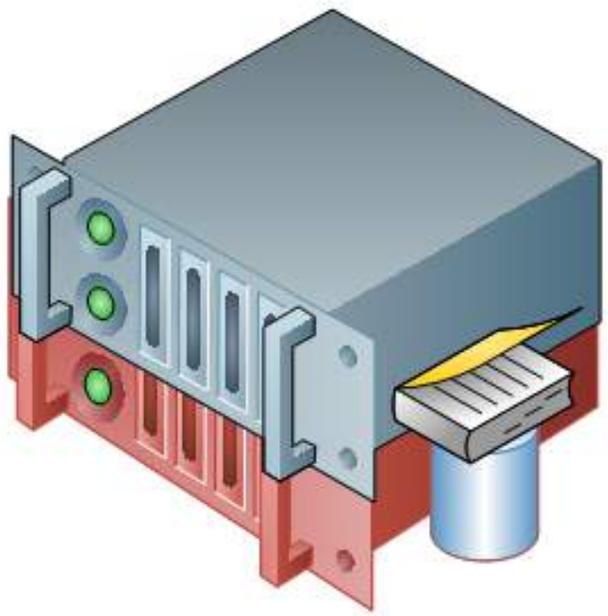
Postgres-XL Journey



55TB usable
RAID50
80TB usable
possible

Currently 8 x HP Apollo 4510
Two IO controllers
SSDs + Raid50
Index, Temp, input output table space separation

Postgres-XL collocated coord + datanode



Coordinator + Datanode+Backup combined

Primary storage
Primary coordinator
Backup storage
Backup coordinator

DataNode storage

Storage for primary operations
Separate OS disk
WAL – Write Ahead Log: min 2x15K 300GB, RAID0(1 optional), quasi-sequential
Main Data Partition: 15-20TB usable per box, RAID10, mixed random + sequential
Data+seldomly used index
~1GB/s sequential
> 2000 real IOPS
Index and reconstruction space: SSD ~2TB, RAID10
>1GB/s sequential
>40000 real IOPS
Separate IO channel for Main data partition and index partition.
Needs Min 128GB RAM, lots of cores 32+
2x infiniband

Coordinator storage

Coordination:
In principle can sit on the same partitions as DataNode
or one dedicated RAID10 with relatively fast IOPS,
<500GB of storage.
Needs min ~16GB of RAM,
symmetric number of cores (shared with datanode),
mostly idling if done properly

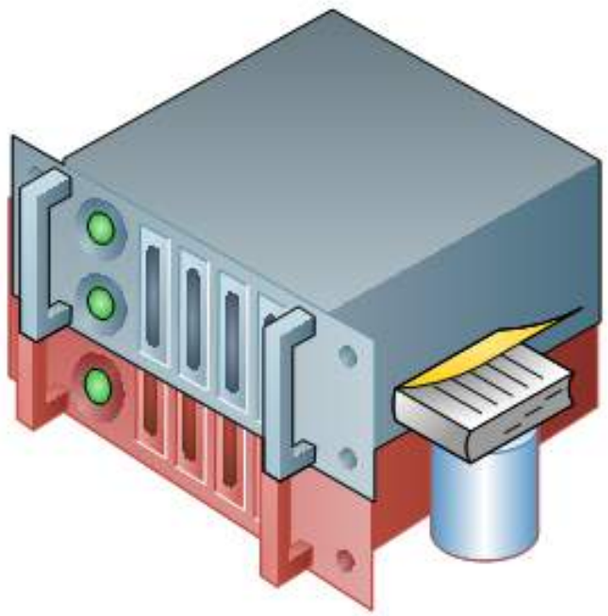
Backup DataNode storage

Storage for backup operations
WAL – Write Ahead Log: min 2x15K 300GB, RAID0(1 optional), quasi-sequential
Main Data Partition: 15-20TB usable per box, RAID6, quasi sequential + sequential
Data+seldomly used index
~1GB/s sequential
> 500 real IOPS
Separate IO channel for backup and primary.
Needs Min 16GB RAM,few cores

Backup Coordinator storage

Coordination:
In principle can sit on the same partitions as Backup DataNode
Needs min ~8GB of RAM,
symmetric number of cores (shared with backup datanode),
mostly idling if done properly

Postgres-XL collocated coord + datanode



Coordinator + Datanode+Backup combined

Primary storage
Primary coordinator
Backup storage
Backup coordinator

DataNode storage

Storage for primary operations
Separate OS disk
WAL – Write Ahead Log: min 2x15K 300GB, RAID0(1 optional), quasi-sequential
Main Data Partition: 15-20TB usable per box, RAID10, mixed random + sequential
Data+seldomly used index
~1GB/s sequential
> 2000 real IOPS
Index and reconstruction space: SSD ~2TB, RAID10
>1GB/s sequential
>40000 real IOPS
Separate IO channel for Main data partition and index partition.
Needs Min 128GB RAM, lots of cores 32+
2x infiniband

Coordinator storage

Coordination:
In principle can sit on the same partitions as DataNode
or one dedicated RAID10 with relatively fast IOPS,
<500GB of storage.
Needs min ~16GB of RAM,
symmetric number of cores (shared with datanode),
mostly idling if done properly

Backup DataNode storage

Storage for backup operations
WAL – Write Ahead Log: min 2x15K 300GB, RAID0(1 optional), quasi-sequential
Main Data Partition: 15-20TB usable per box, RAID6, quasi sequential + sequential
Data+seldomly used index
~1GB/s sequential
> 500 real IOPS
Separate IO channel for backup and primary.
Needs Min 16GB RAM,few cores

Backup Coordinator storage

Coordination:
In principle can sit on the same partitions as Backup DataNode
Needs min ~8GB of RAM,
symmetric number of cores (shared with backup datanode),
mostly idling if done properly

Postgres-XL Journey

- We pushed few patches improving scalability
  - ~900 active computing cores, no pooling would work well
- Reported ~40 issues, from corruption to minor annoyances
- Created a global system views extension stub
  - https://github.com/yazun/xl_global_views

```
select * from pgxl_stat_activity where  state = 'active' ;
 node_name | node_type | datid | datname |  pid  | usesysid  |    usename     | application_name |
-----------+-----------+-------+---------+-------+-----------+----------------+------------------+---
 coord1    | C         | 16395 | surveys |  7628 |        10 | postgresxl     |                  | [r
 coord1    | C         | 16395 | surveys | 34736 |        10 | postgresxl     | psql             | [r
 coord2    | C         | 16395 | surveys | 39046 |        10 | postgresxl     | pgxc             | 19
 coord3    | C         | 16395 | surveys | 18188 |        10 | postgresxl     |                  | [r
 coord3    | C         | 16395 | surveys | 23589 |        10 | postgresxl     | pgxc             | 19
 coord4    | C         | 16395 | surveys |  8713 |        10 | postgresxl     | pgxc             | 19
 coord4    | C         | 16395 | surveys | 37982 |   2154309 | rimoldin_local |                  | 16
 coord5    | C         | 16395 | surveys | 14039 |        10 | postgresxl     | pgxc             | 19
```

Future

Future

- We are finishing Data Release Cycle 2 now
- Postgres XL 10
  - Helping with hardware synthetic testing on our hw
  - Deployment and testing on v. 10. November-…
    - Moving to native v10 partitioning
    - Continuing being part of the Postgres-XL effort
- Some 10+ issues should be fixed.
- Expansion of the cluster.

Thank you, Q & A

- Special thanks for Pavan Deolasee and Tomas Vondra for their dedication and 2ndQuadrant (Simon Rigs) for recognition of the synergy.