Gaia Variability Studies: case for Postgres-XC.

Krzysztof.Nienartowicz@unige.ch

Gaia Data Processing Centre in Geneva,

Gaia Coordination Unit 7

PgCon

Ottawa 23rd May 2014

UNIVERSITÉ DE GENÈVE

Gaia DPAC

Includes slides courtesy of Berry Holl and Laurent Eyer

Structure

- My story
- Gaia mission
- Gaia science at CU7, Geneva
- Data model and processing model
- Hardware
- XC role
- Collaboration

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 (5 years); UNIGE/ISDC, Geneva
  - Coordination Unit 7 (CU7) Data Architect
  - Data Processing Centre for Geneva (DPCG) deputy
  - XLDB, timeseries, distributed systems, science, Literature, art, history, politics, basketball, paragliding, architecture, volleyball, sailing, biking,…

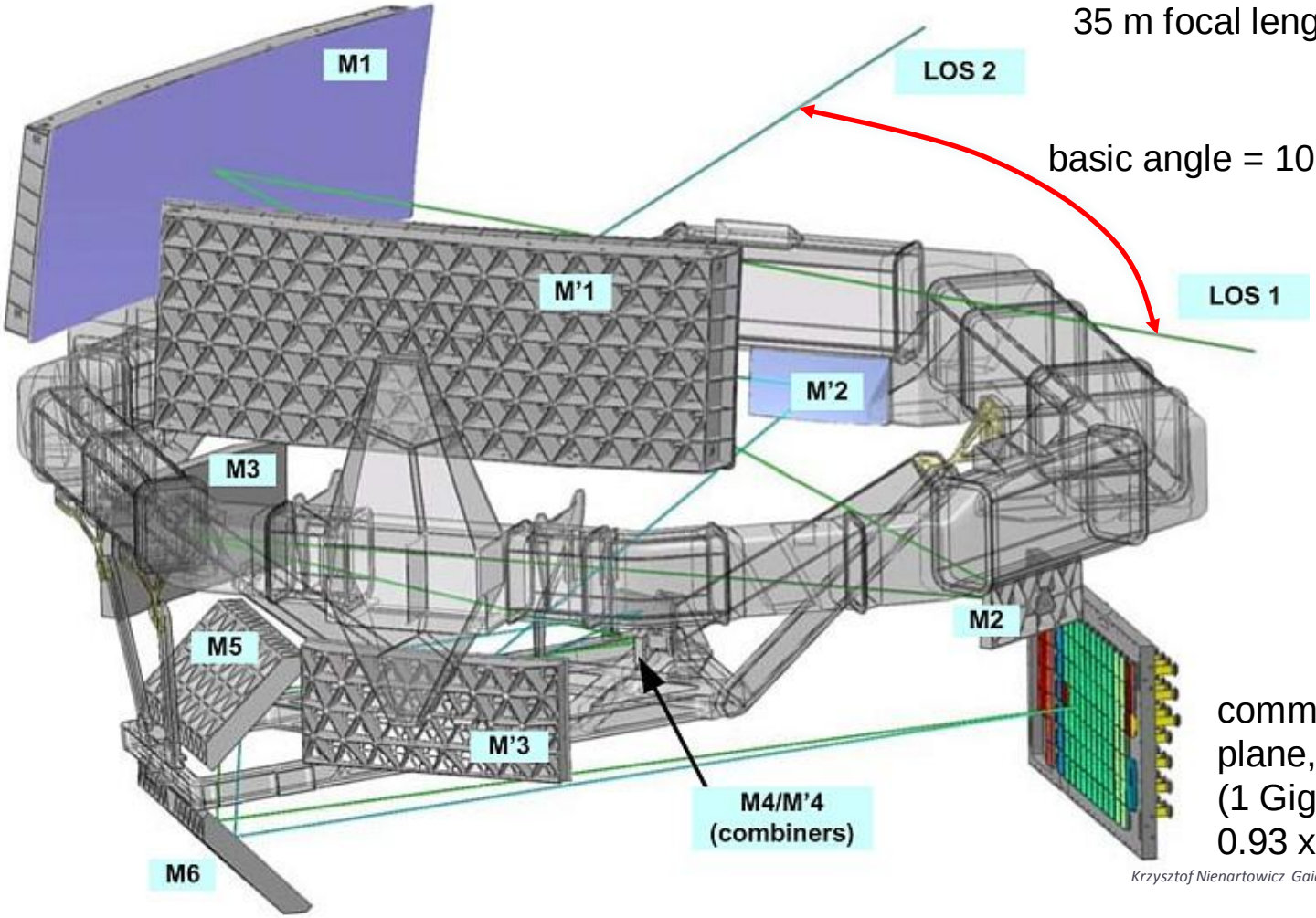
Structure

- My story
- **Gaia mission**
- Gaia science at CU7, Geneva
- Data model and processing model
- Hardware
- XC role
- Collaboration

What is 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 (~1.000.000.000 stars, asteroids, quasars, extragalactic supernovae, etc)**
  - **On average 80 times 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 as many as 10.000 exo-planets
  - **Estimated 10-20% of all population are variables**

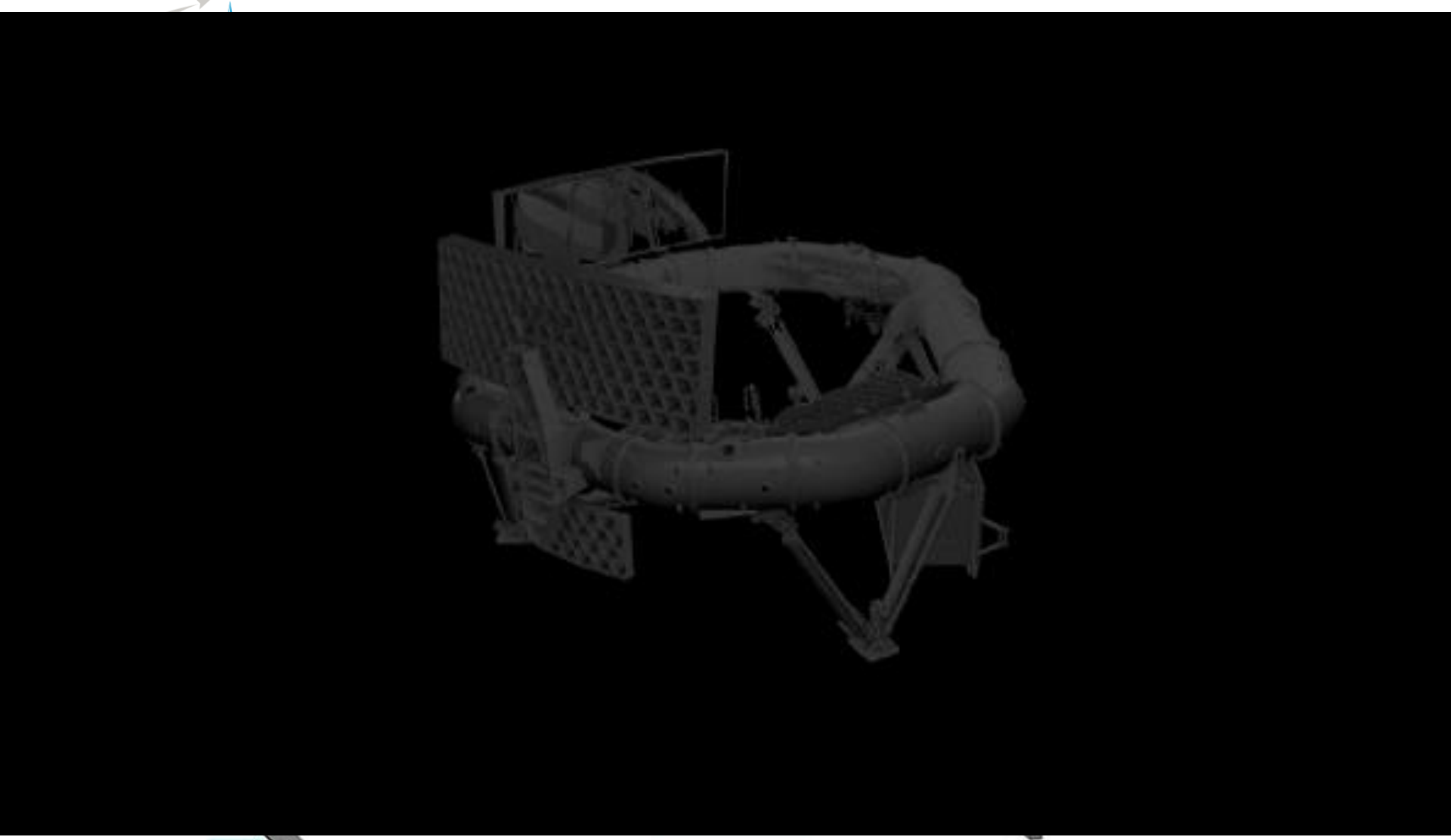
The Gaia instruments



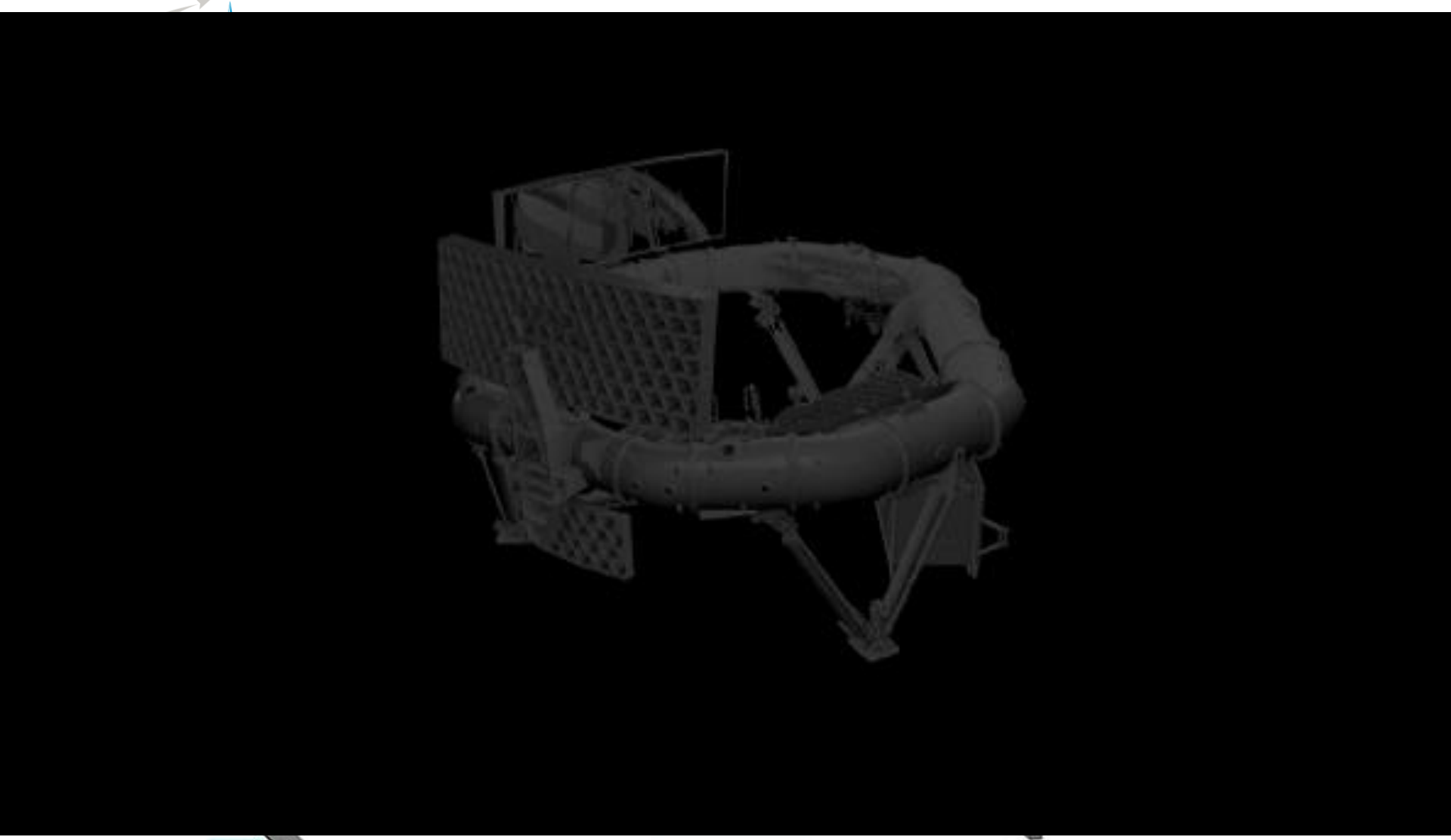
2 off-axis telescopes
1.45 x 0.5 m^2 aperture
35 m focal length

common focal plane, 106 CCDs (1 Gigapixel)
0.93 x 0.42 m^2

The Gaia instruments



The Gaia instruments

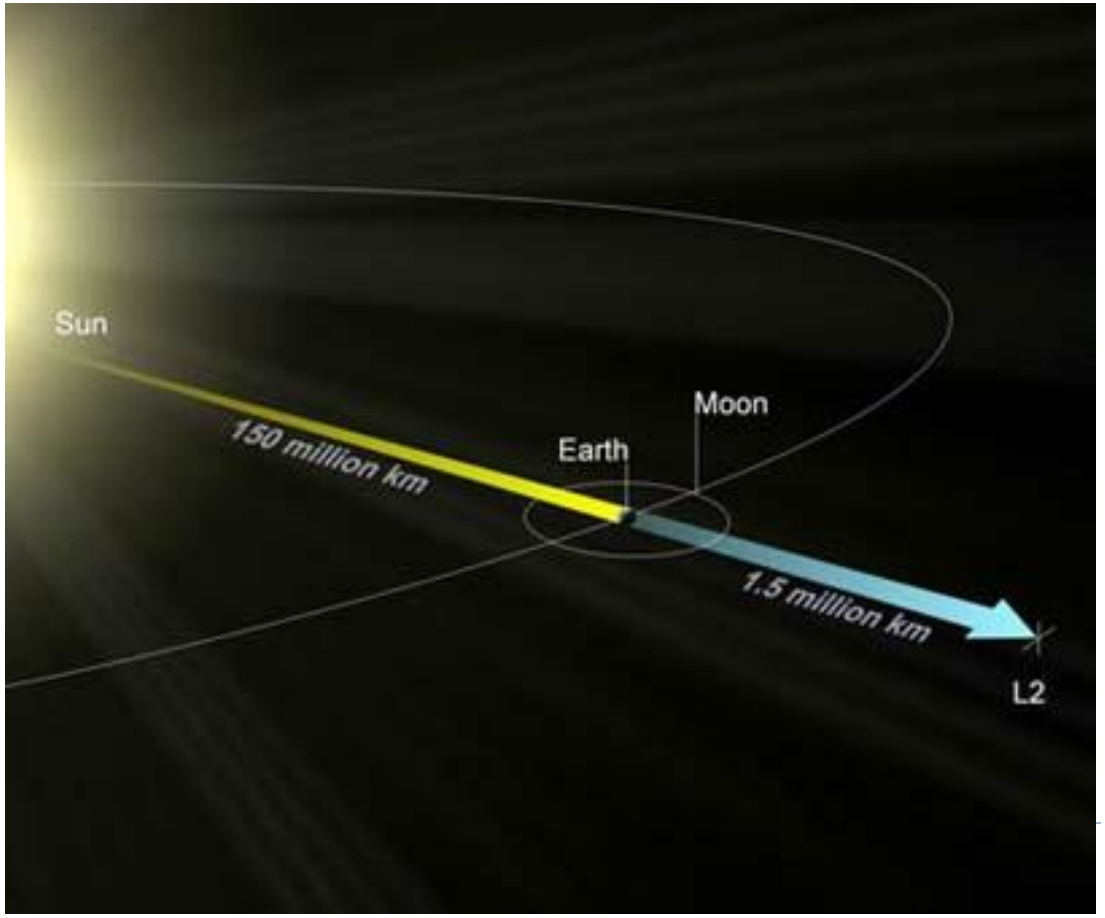


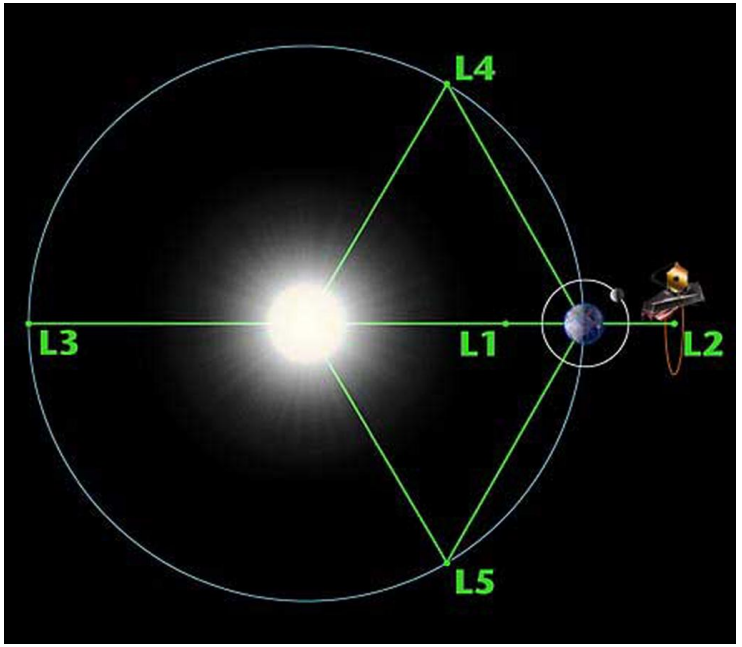
One of the two primary mirrors

esa EADS astrium SAFRAN Sagem

The Gaia satellite

- Location: Lagrange point 2
- Commissioning phase, first calibrated data: Q2/2015





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

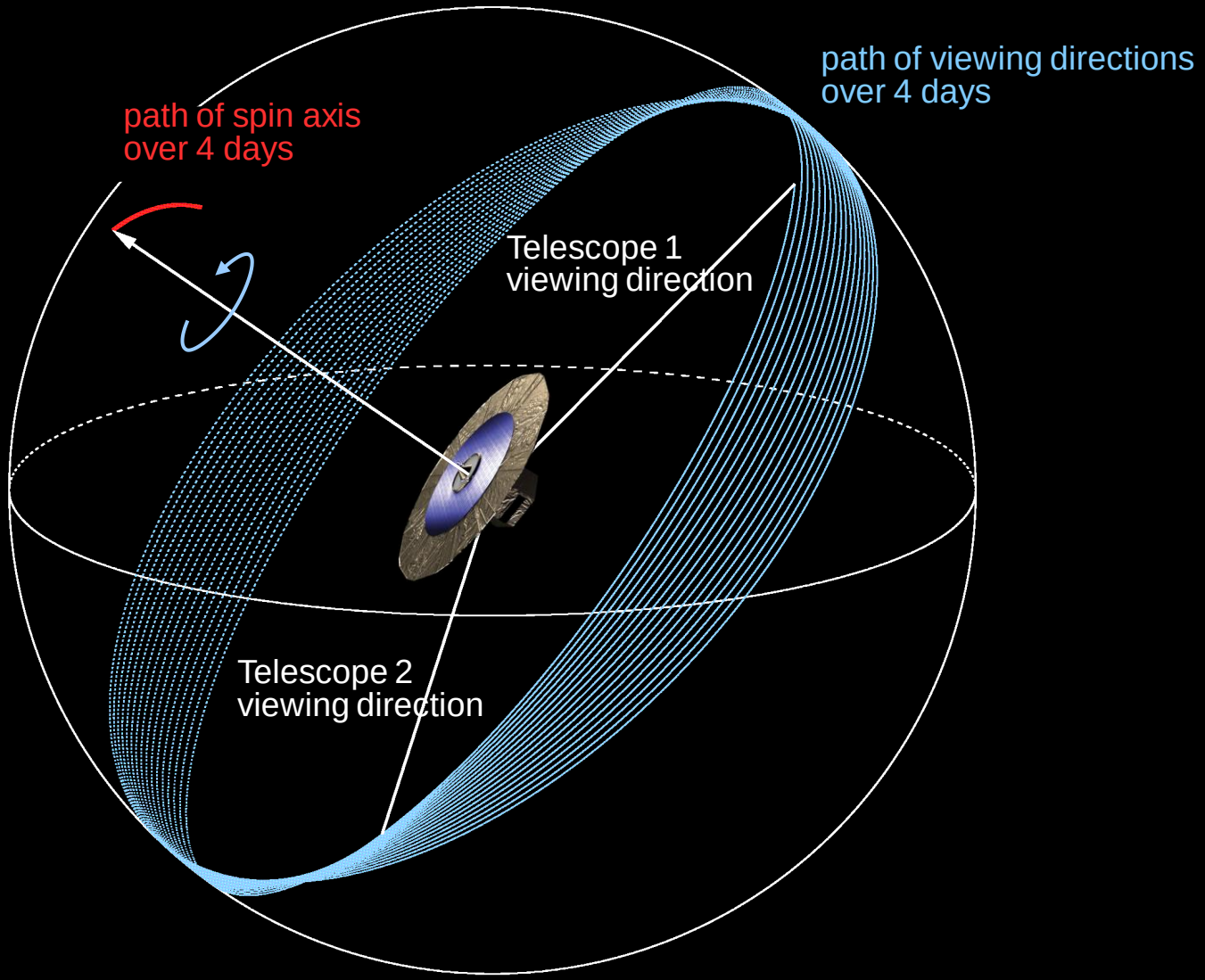


image credit: Lennart Lindegren

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

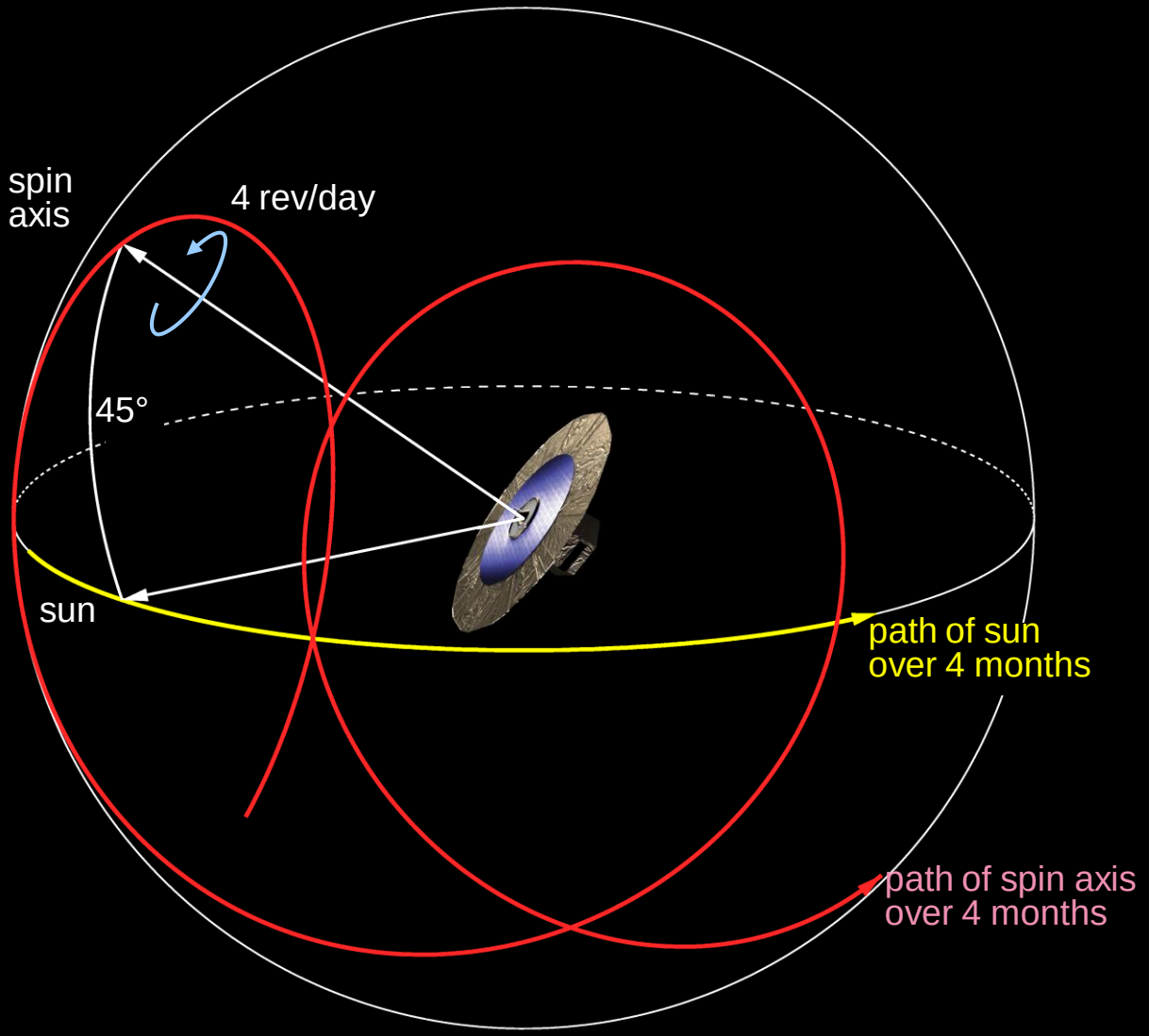
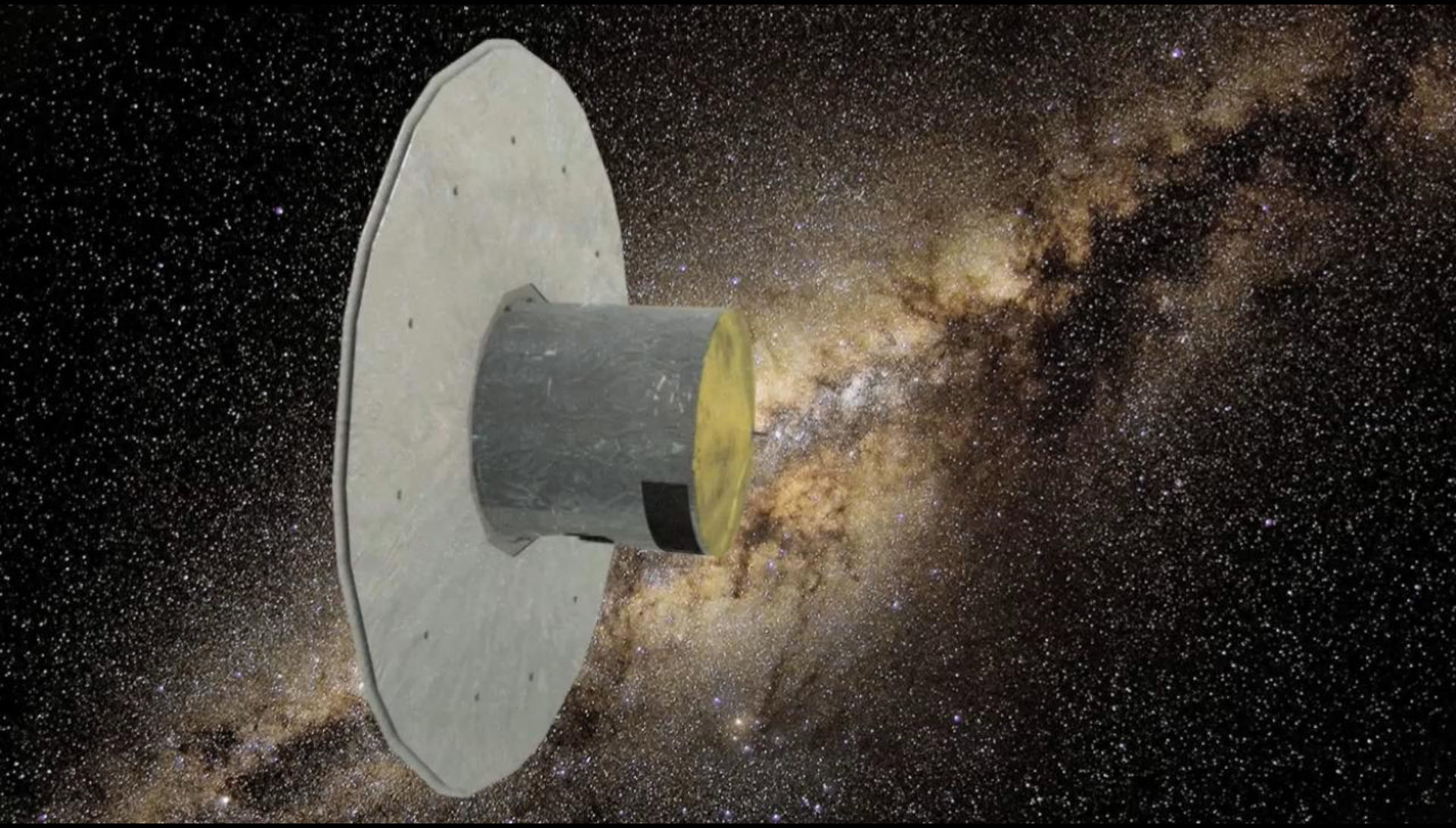


image credit: Lennart Lindegren



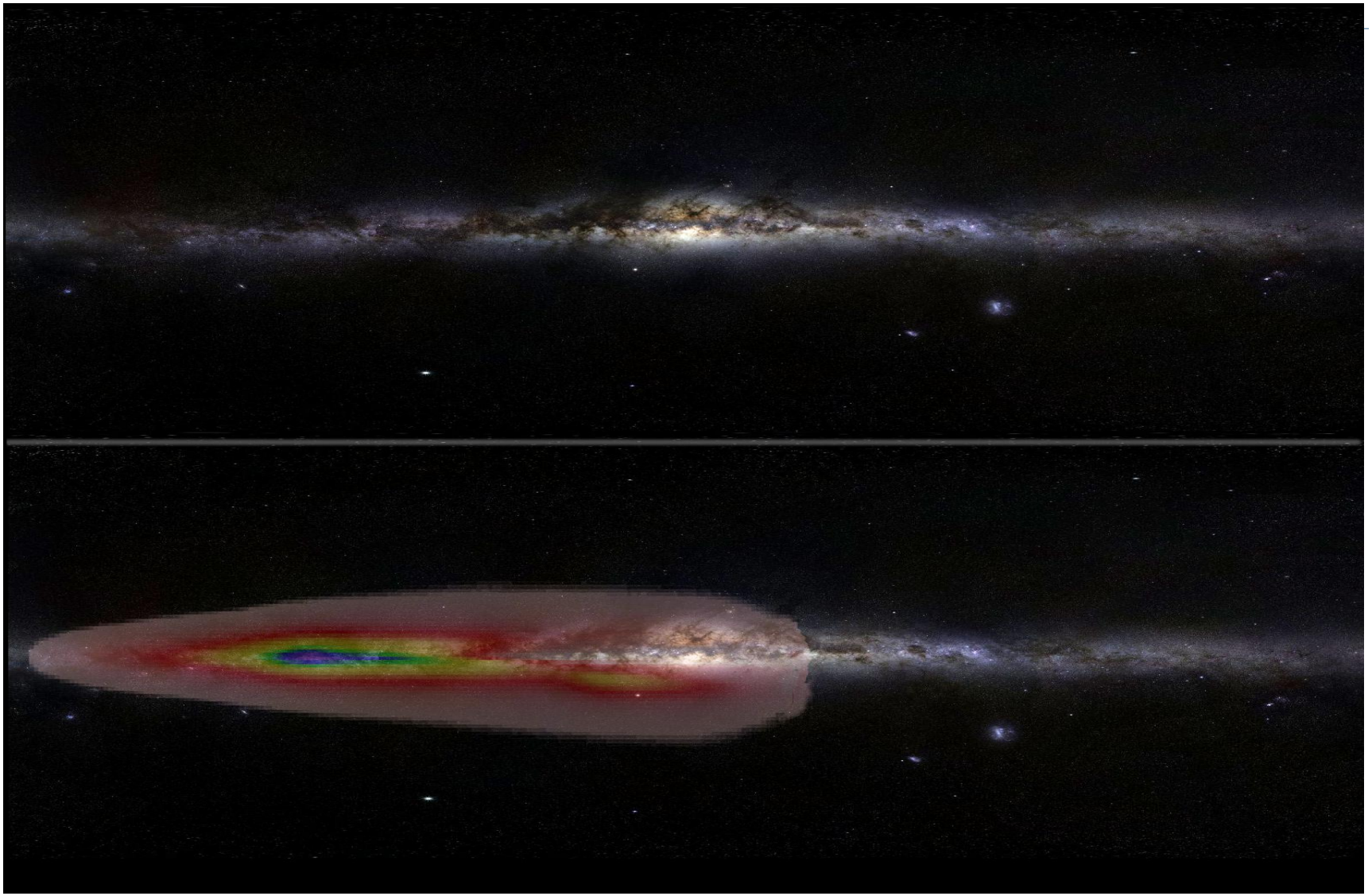
credit: ESA

Structure

- My story
- Gaia mission
- **Gaia science at CU7, Geneva**
- Data model and processing model
- XC role
- Hardware
- Collaboration

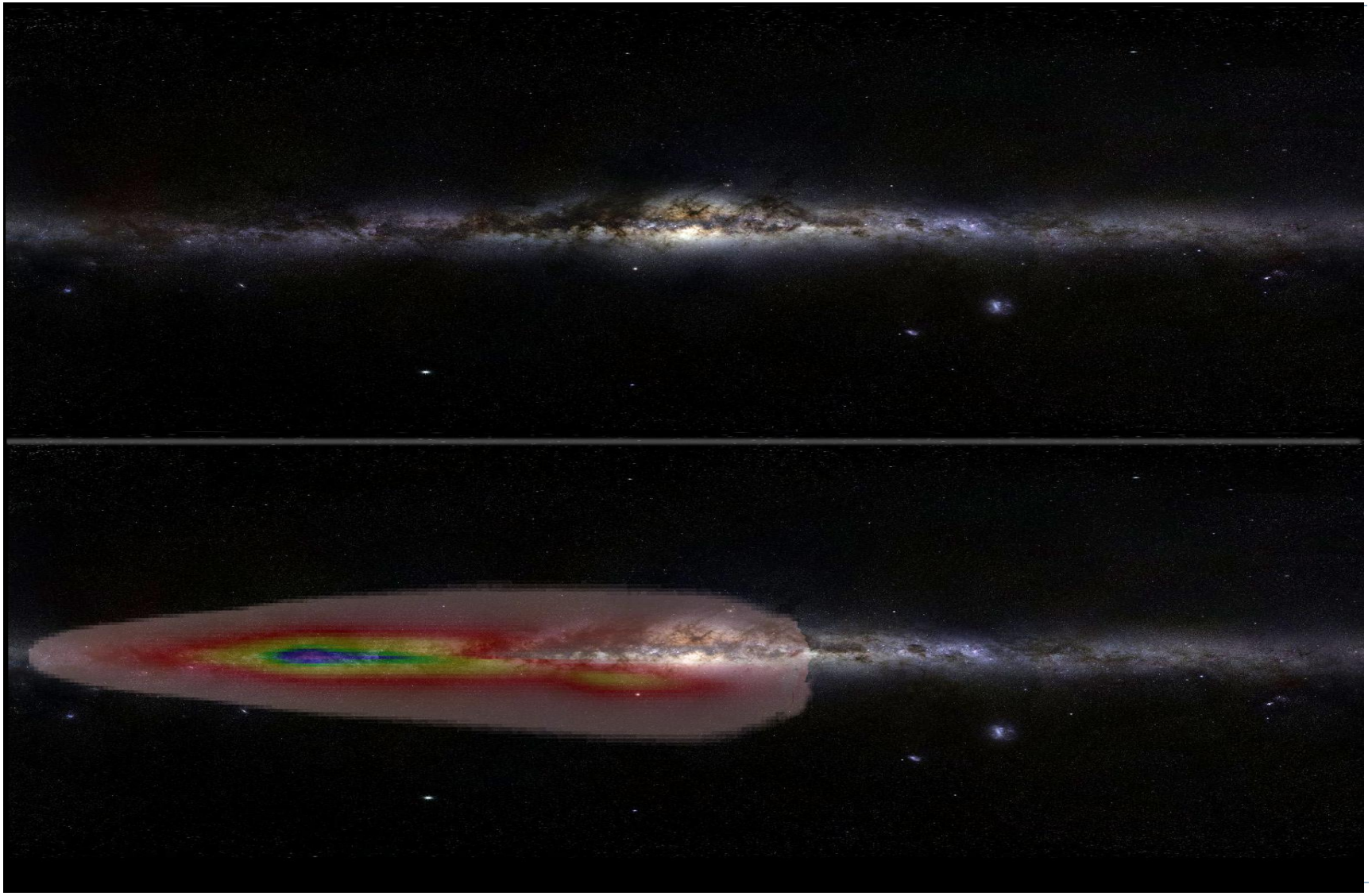
Gaia Science

Simulated coverage



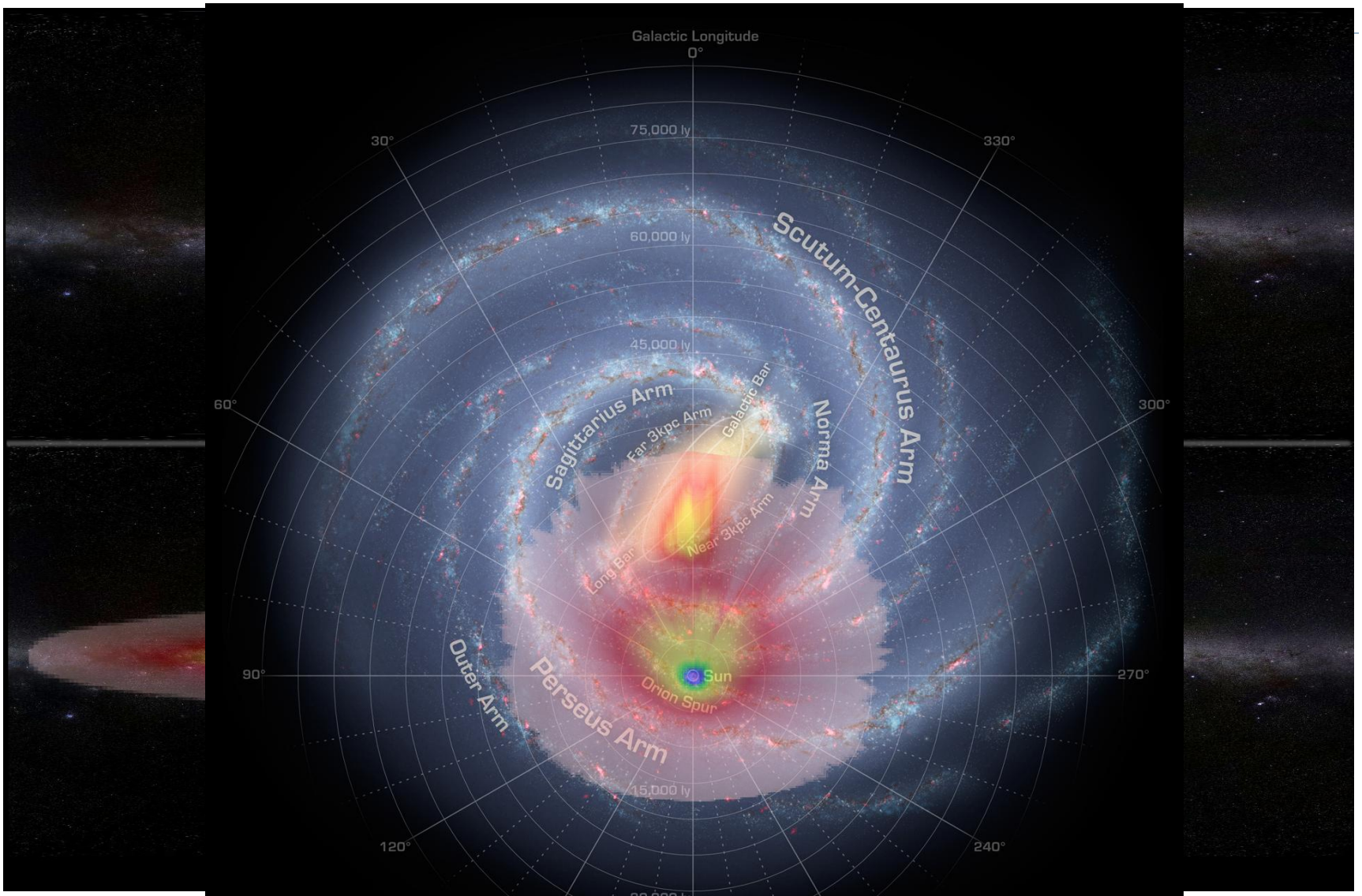
Gaia Science

Simulated coverage



Gaia Science

Simulated coverage



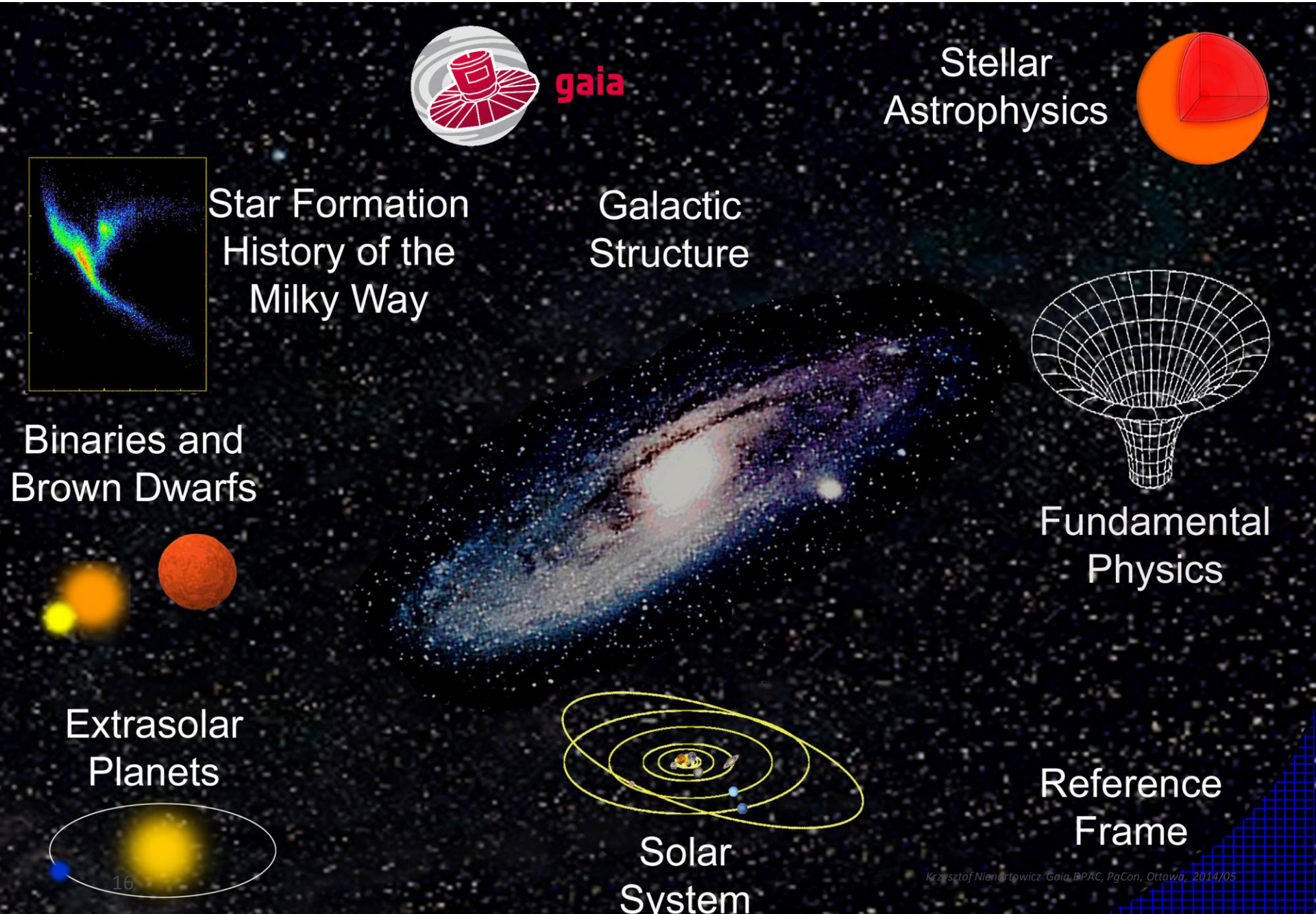
Gaia Science

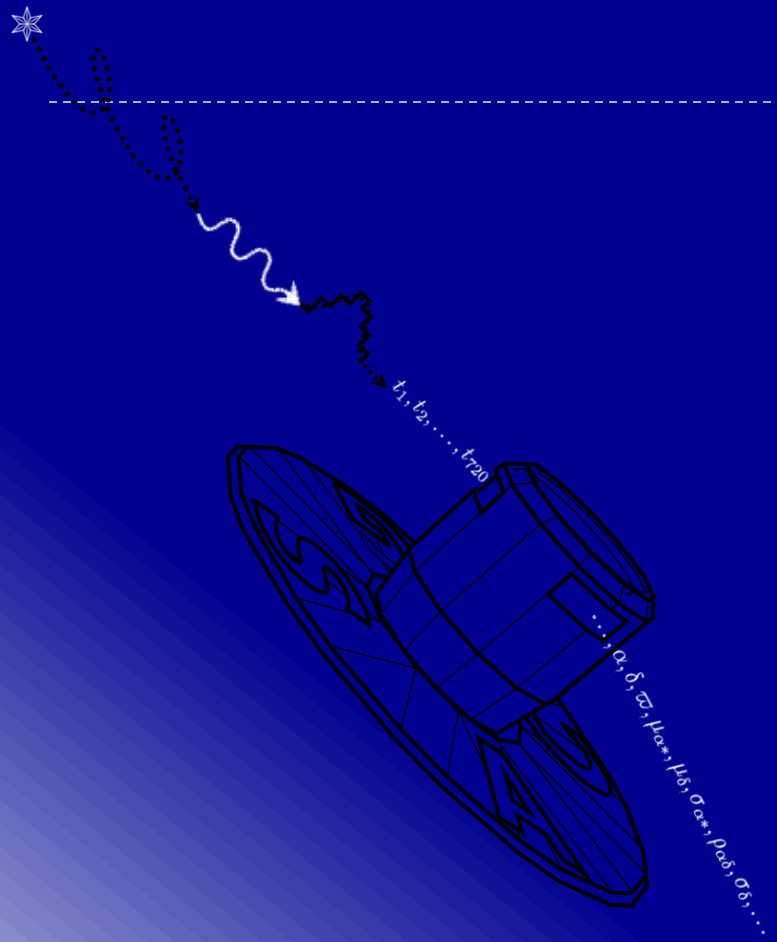
Parallax

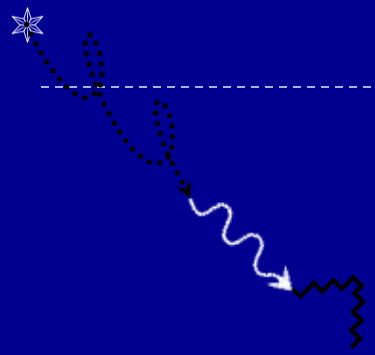
Gaia Science

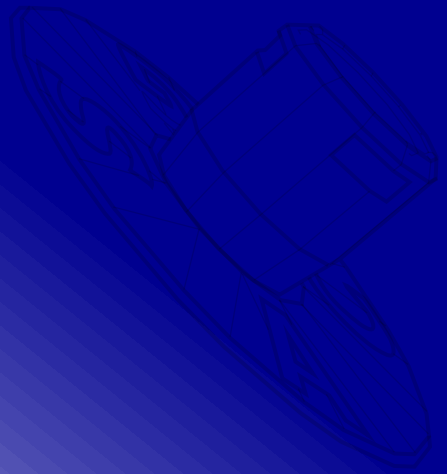
3D Galaxy structure, Galaxy dynamics thanks to parallax, Radial Velocity

Gaia Science




$t_1, t_2, \ldots, t_{720}$
$\ldots, \alpha, \delta, \varpi, \mu_{\alpha*}, \mu_\delta, \sigma_{\alpha*}, \rho_{\alpha\delta}, \sigma_\delta, \ldots$





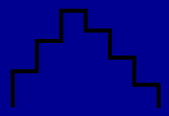
Observing process:

- Apparent source motion on the sky,
- propagation of light to Gaia,
- Astrometric CCDs readout,
typically 6x1 binned counting profile

x 720 x 10^9 = **10^{12}**

Observing process:

- Apparent source motion on the sky,
- propagation of light to Gaia,
- Astrometric CCDs readout,
typically 6x1 binned counting profile



x 720 x 10^9 = $\mathbf{10^{12}}$

Forward modelling:

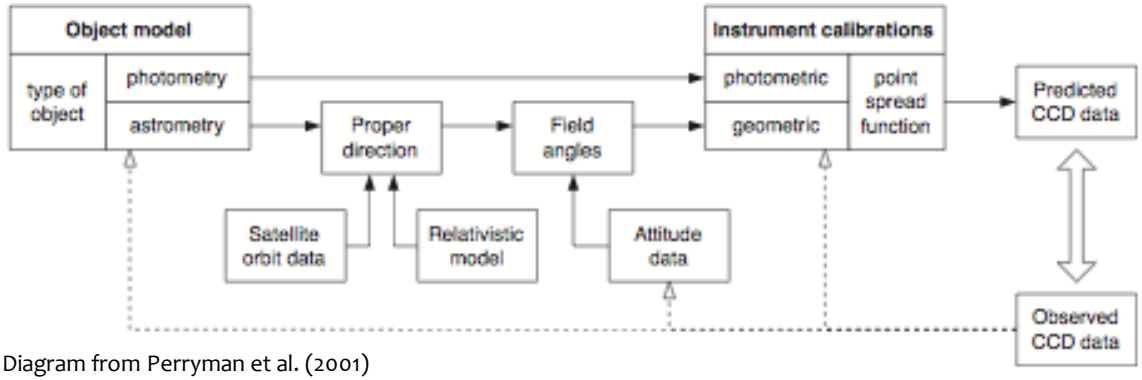
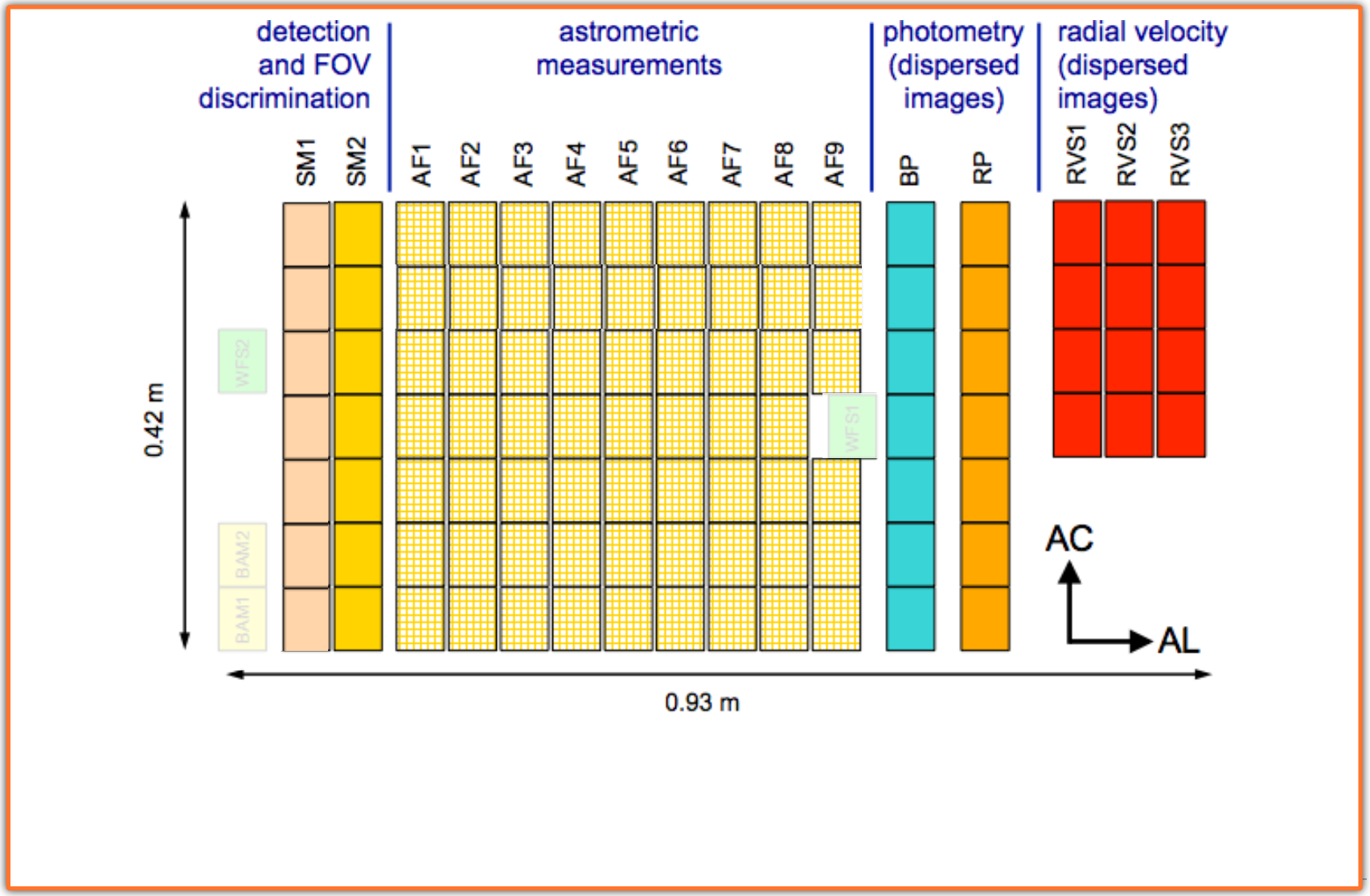


Diagram from Perryman et al. (2001)

Gaia focal plane: CDD, BP RP, RVS observations

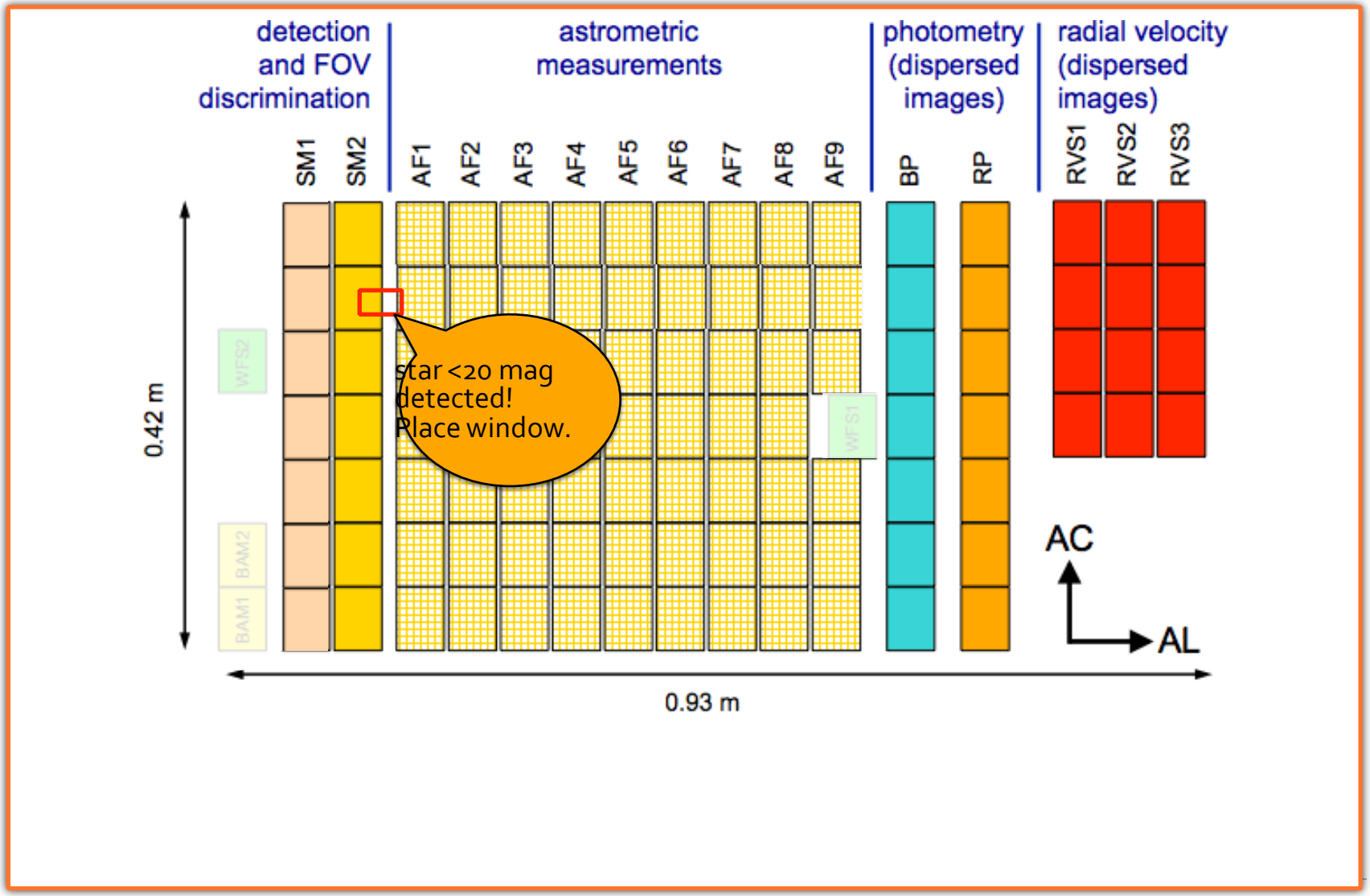
938 million pixel camera



Courtesy Of Berry Holl

Gaia focal plane: CDD, BP RP, RVS observations

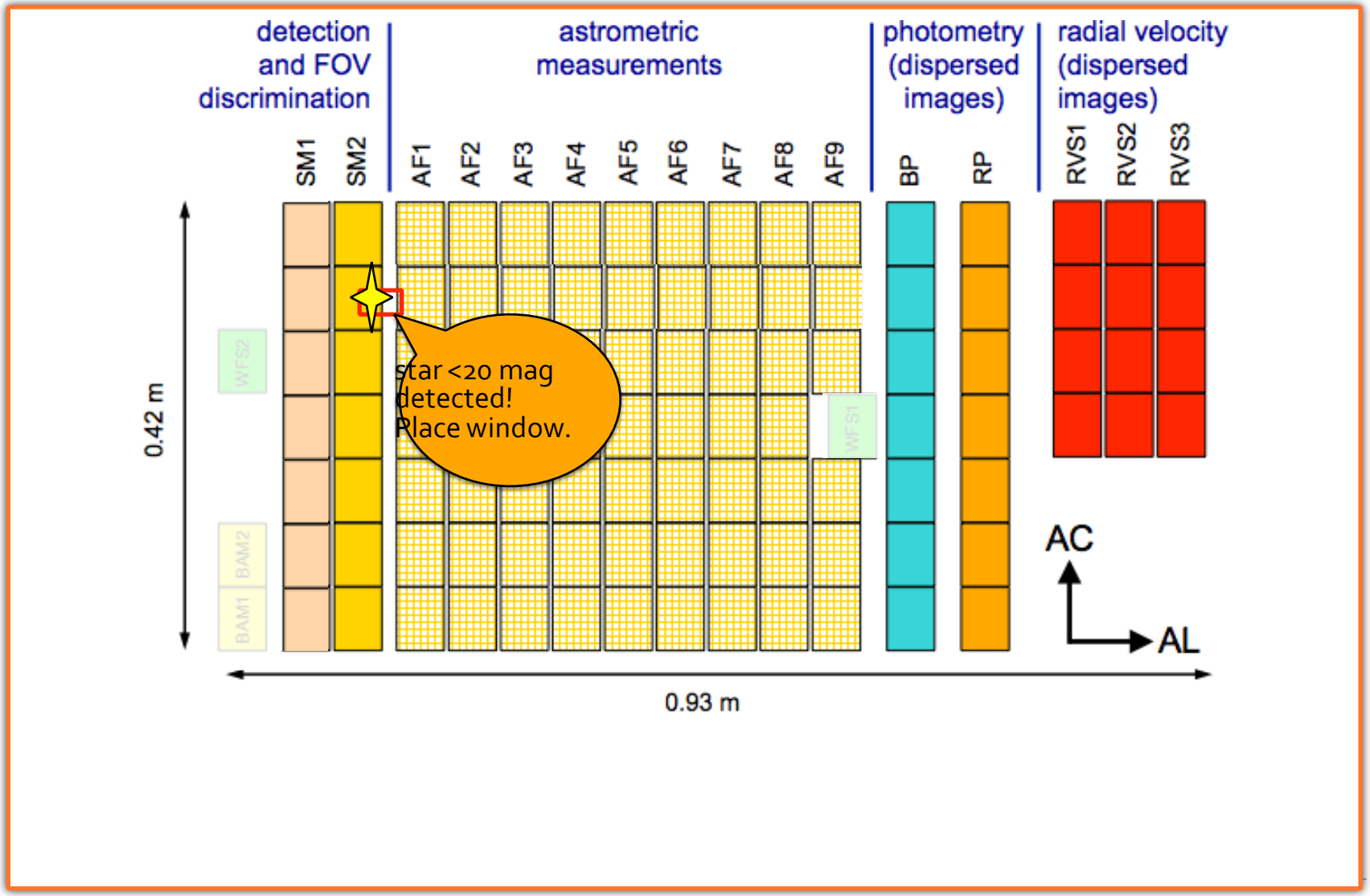
938 million pixel camera



Courtesy Of Berry Holl

Gaia focal plane: CDD, BP RP, RVS observations

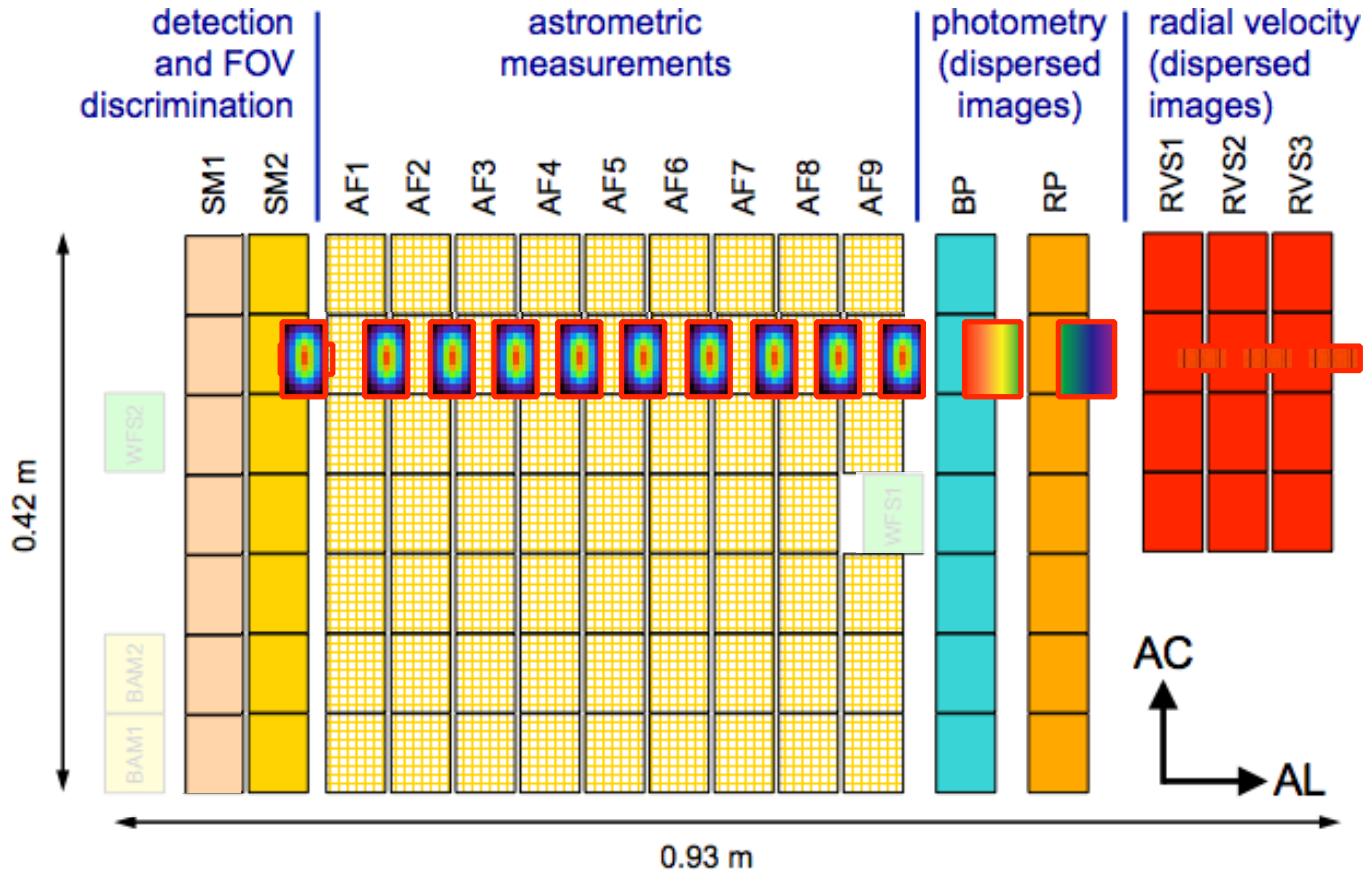
938 million pixel camera



Courtesy Of Berry Holl

Gaia focal plane: CDD, BP RP, RVS observations

938 million pixel camera

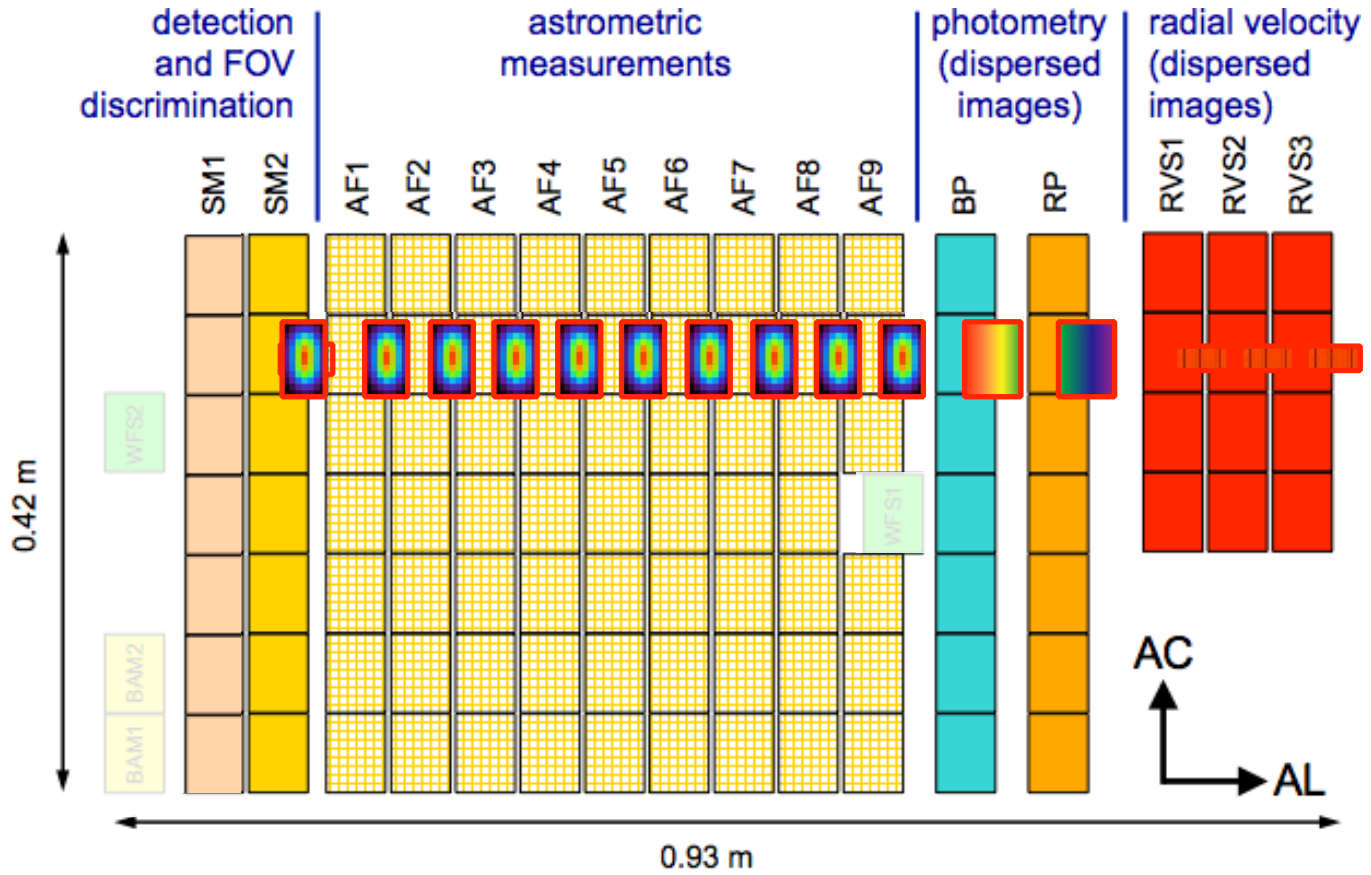


windows observed:

Courtesy Of Berry Holl

Gaia focal plane: CDD, BP RP, RVS observations

938 million pixel camera



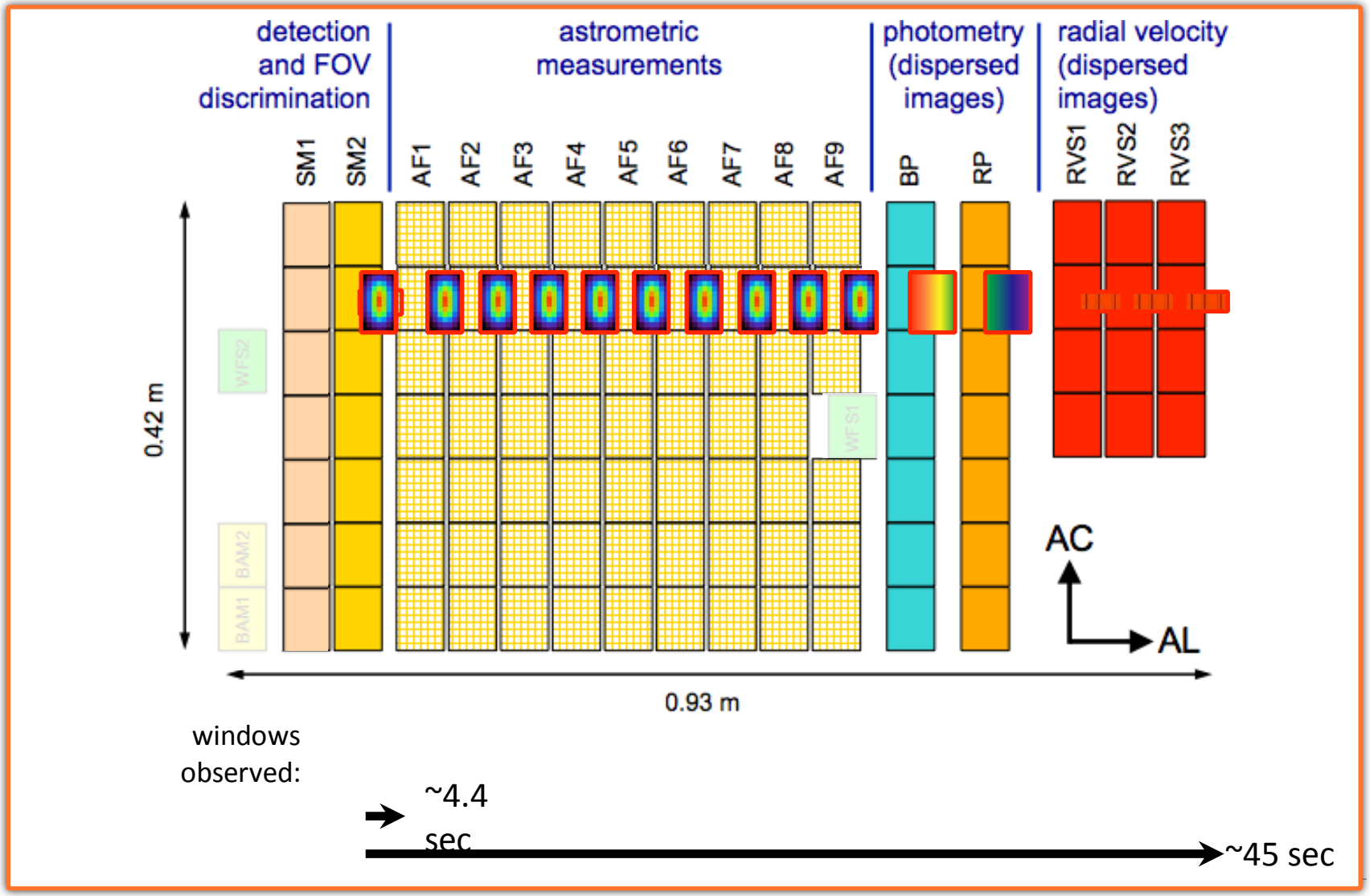
windows observed:

→ ~4.4 sec

Courtesy Of Berry Holl

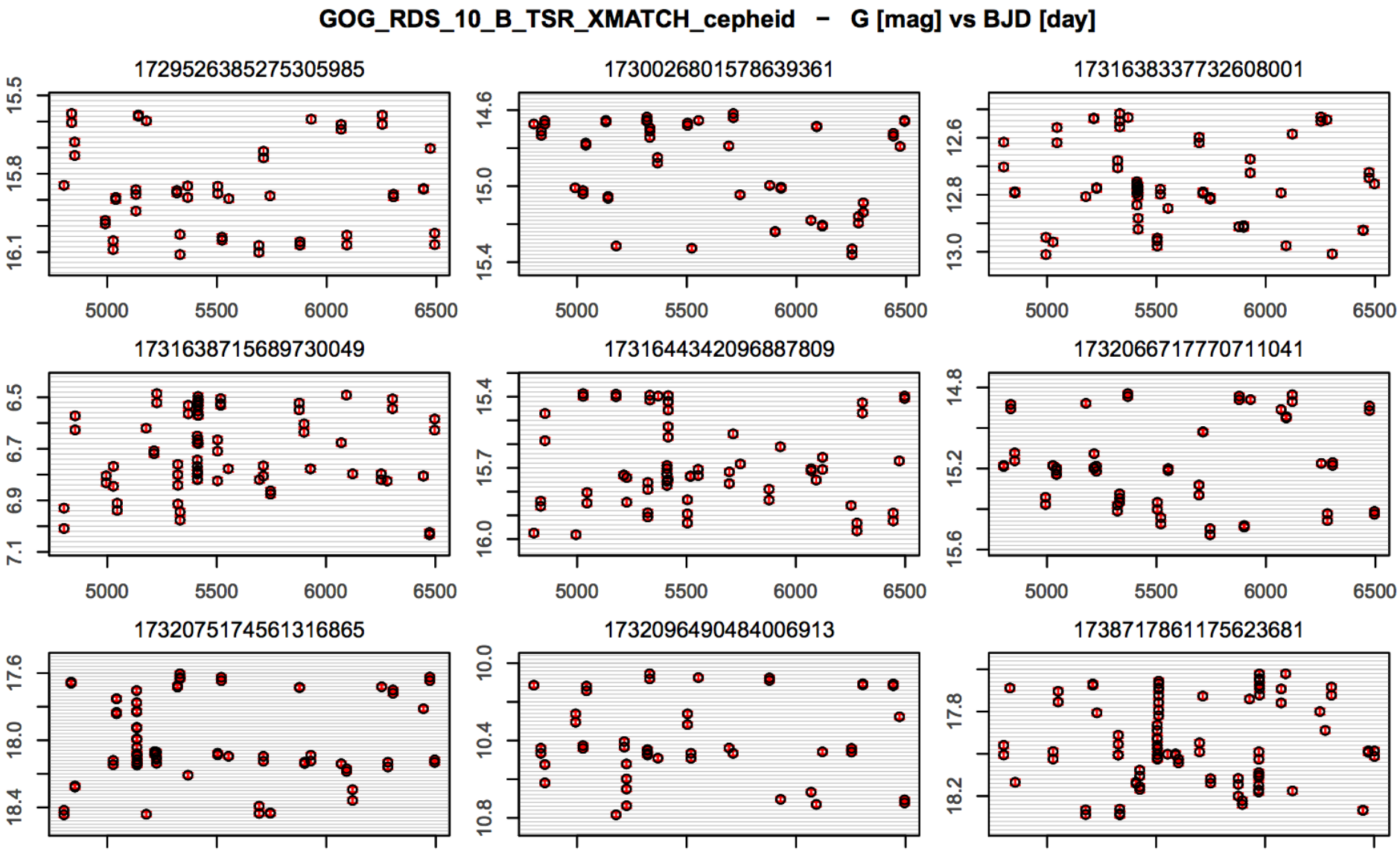
Gaia focal plane: CDD, BP RP, RVS observations

938 million pixel camera

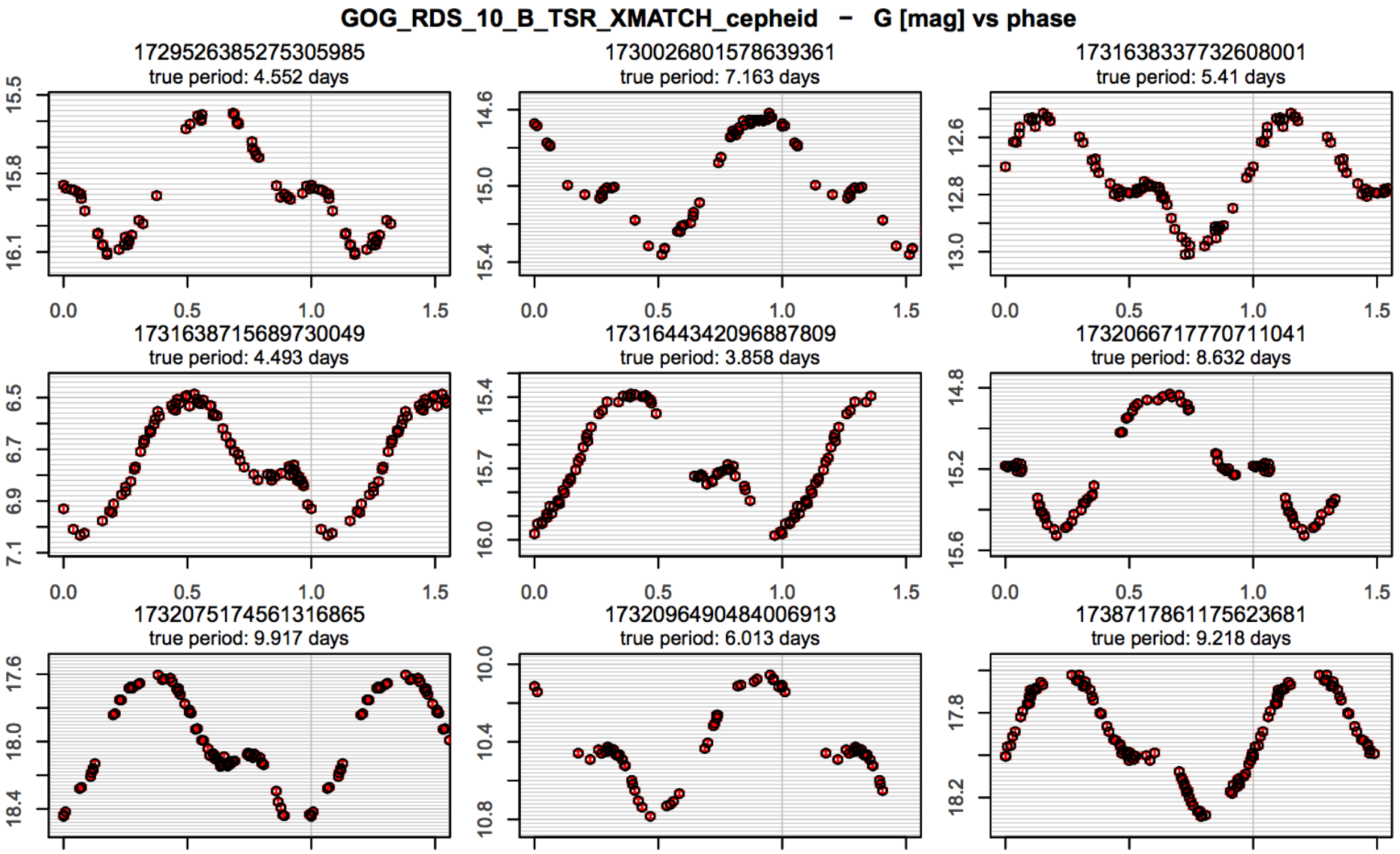


Courtesy Of Berry Holl

Lightcurves == Timeseries

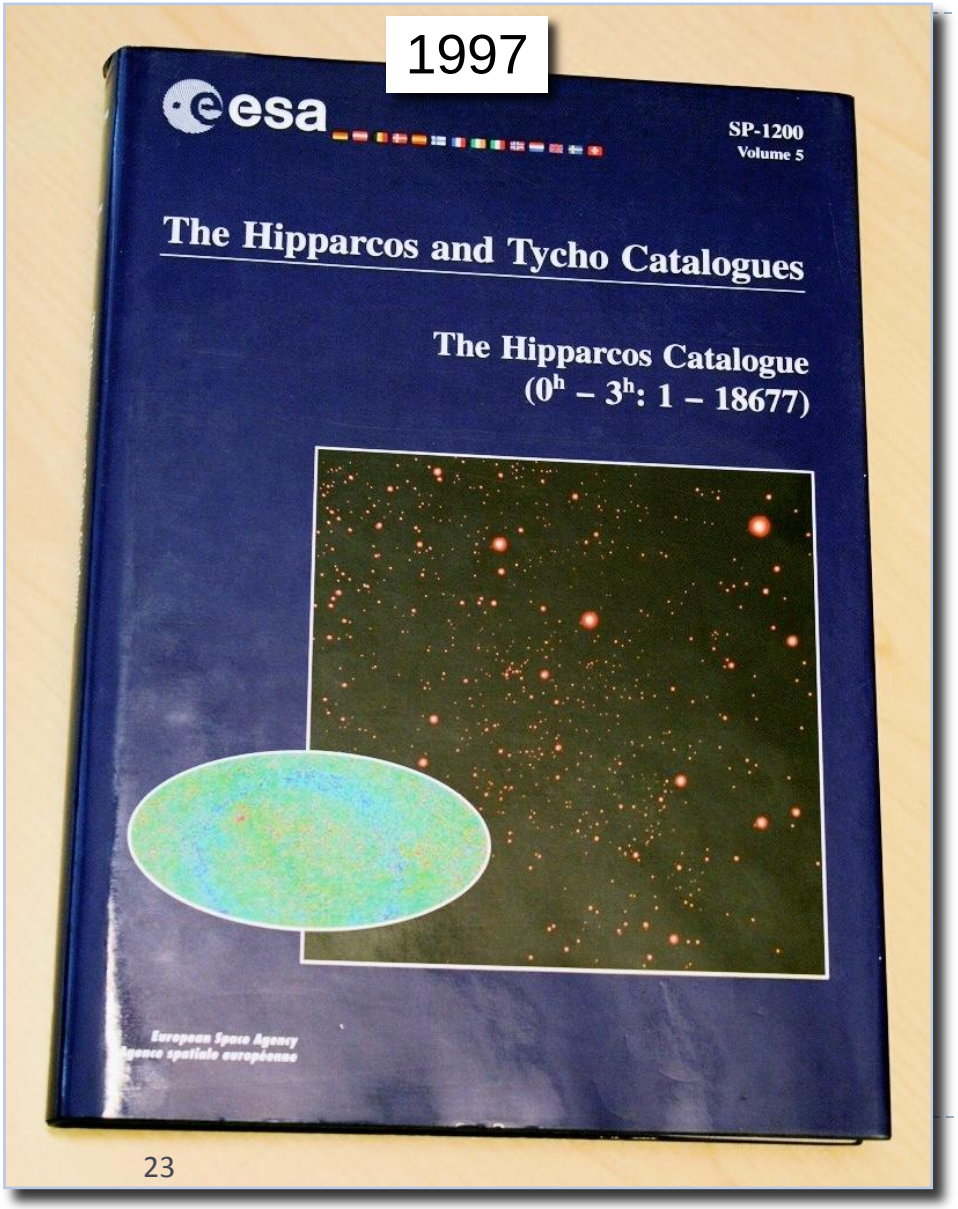


Folded Lightcurves == Timeseries in phase

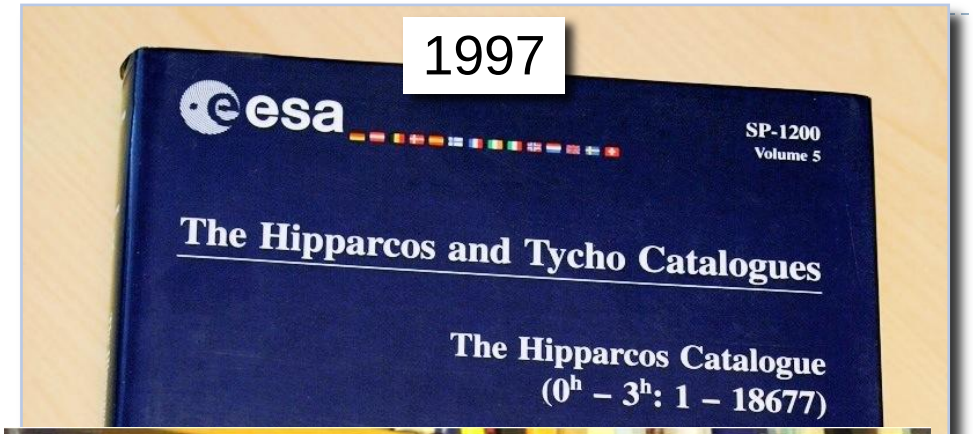


Hipparcos vs Gaia catalogue
Ultimate goal

Hipparcos vs Gaia catalogue
Ultimate goal

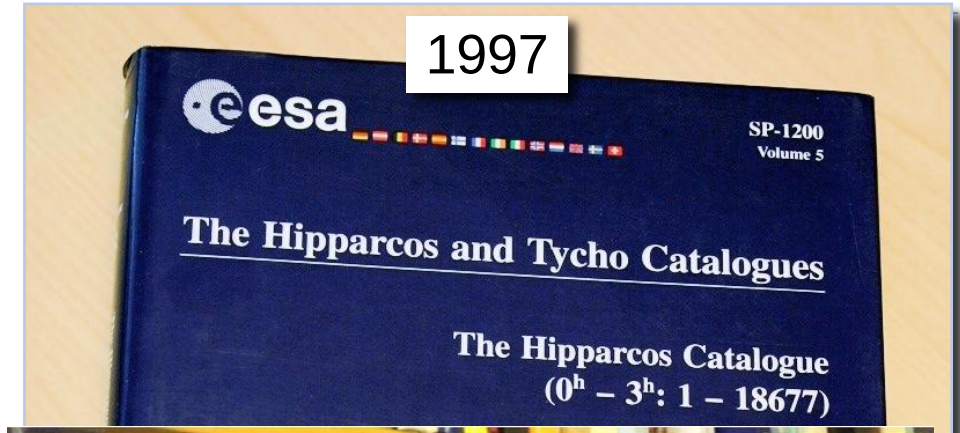


Hipparcos vs Gaia catalogue
Ultimate goal

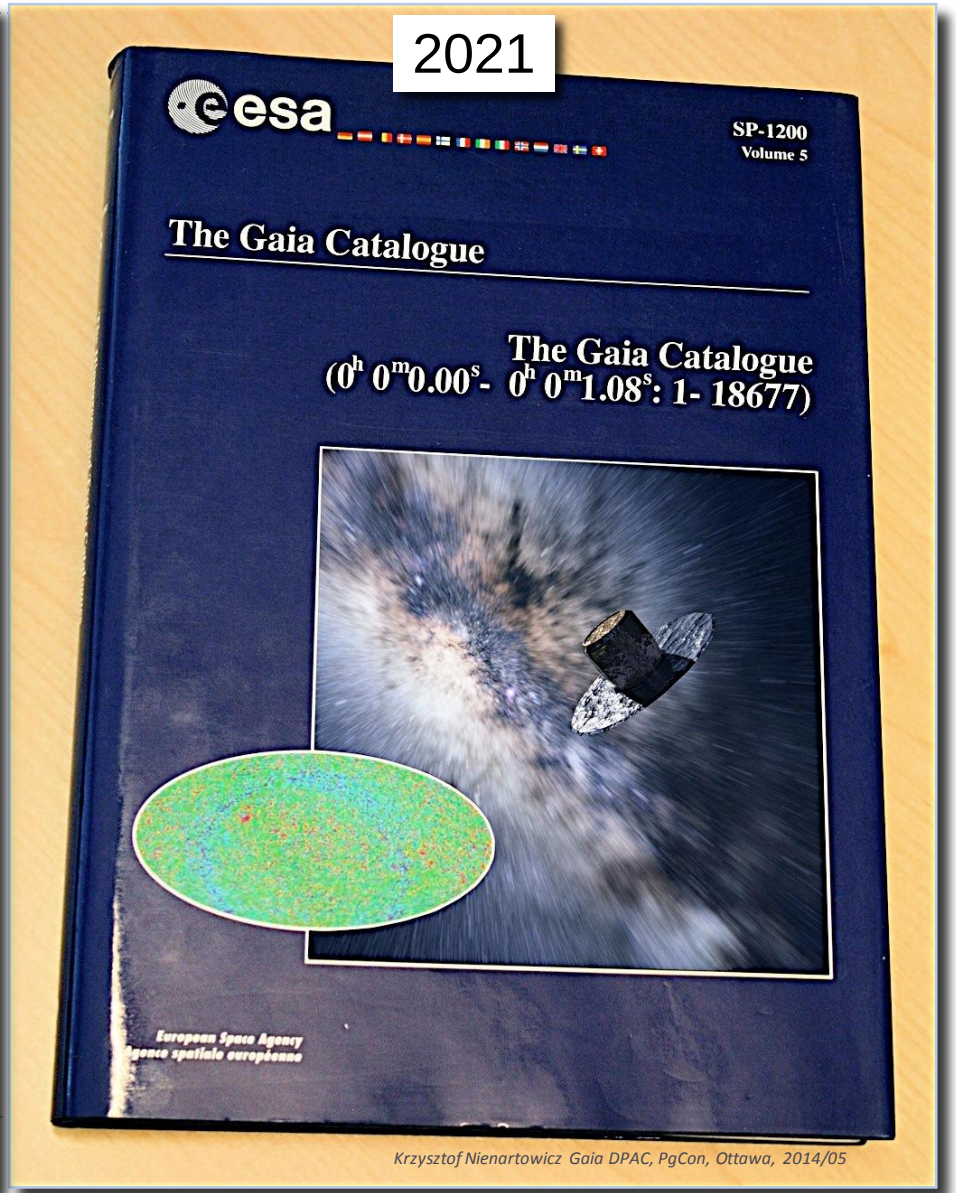


Hipparcos vs Gaia catalogue
Ultimate goal

1997

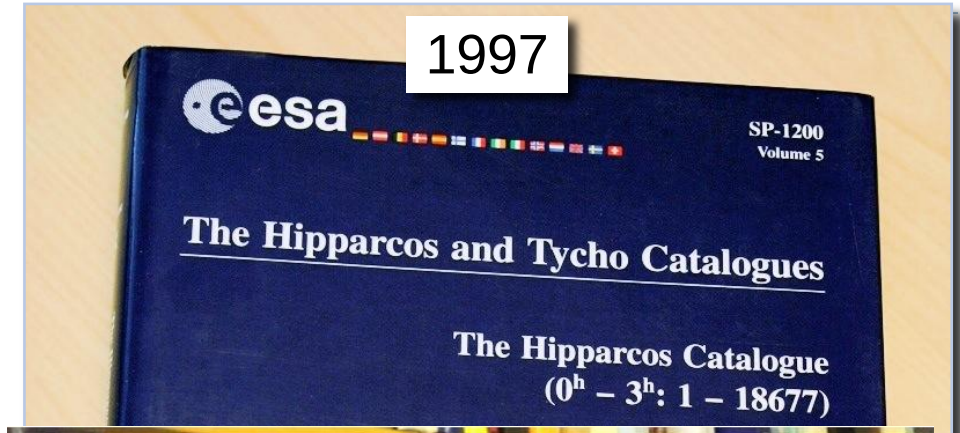


2021

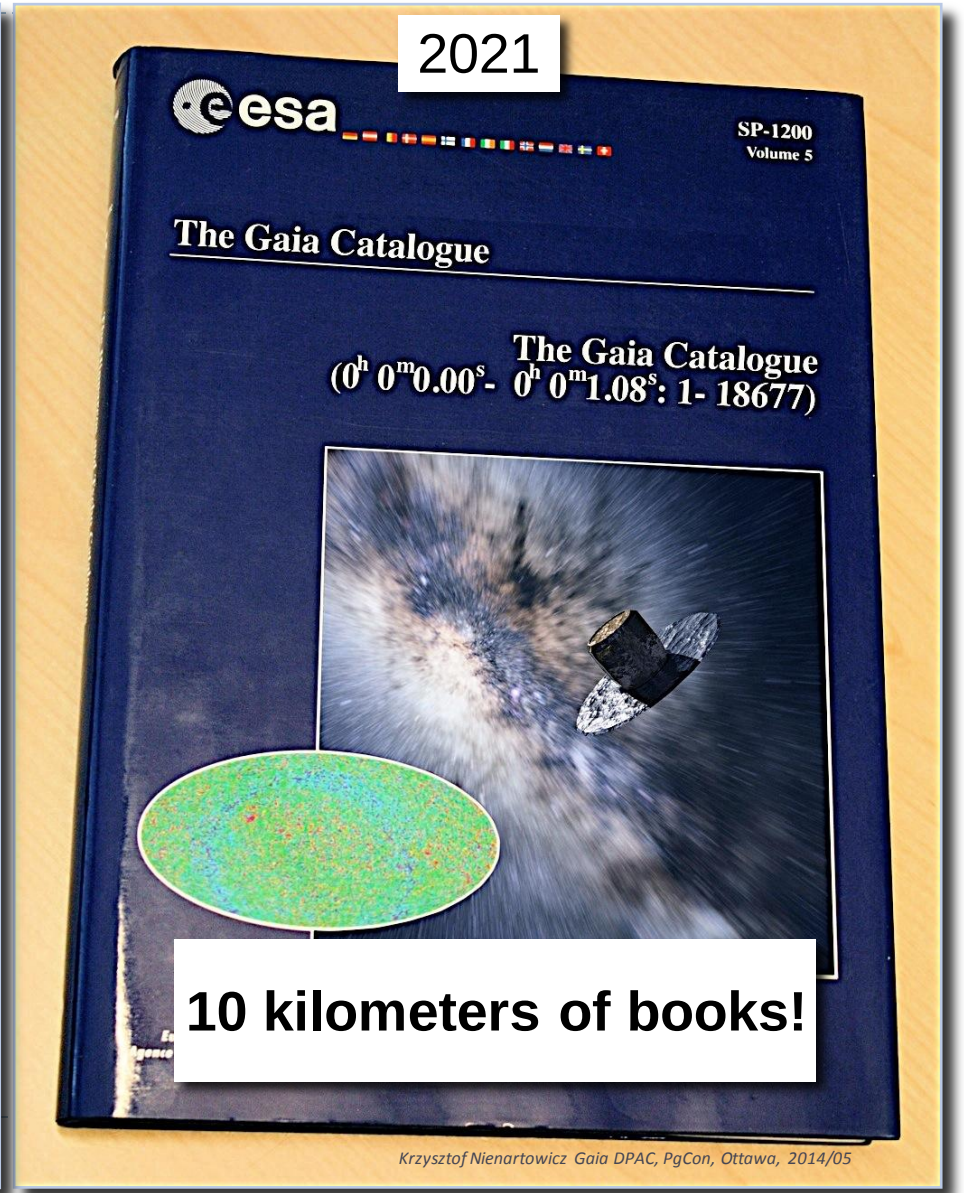


Hipparcos vs Gaia catalogue
Ultimate goal

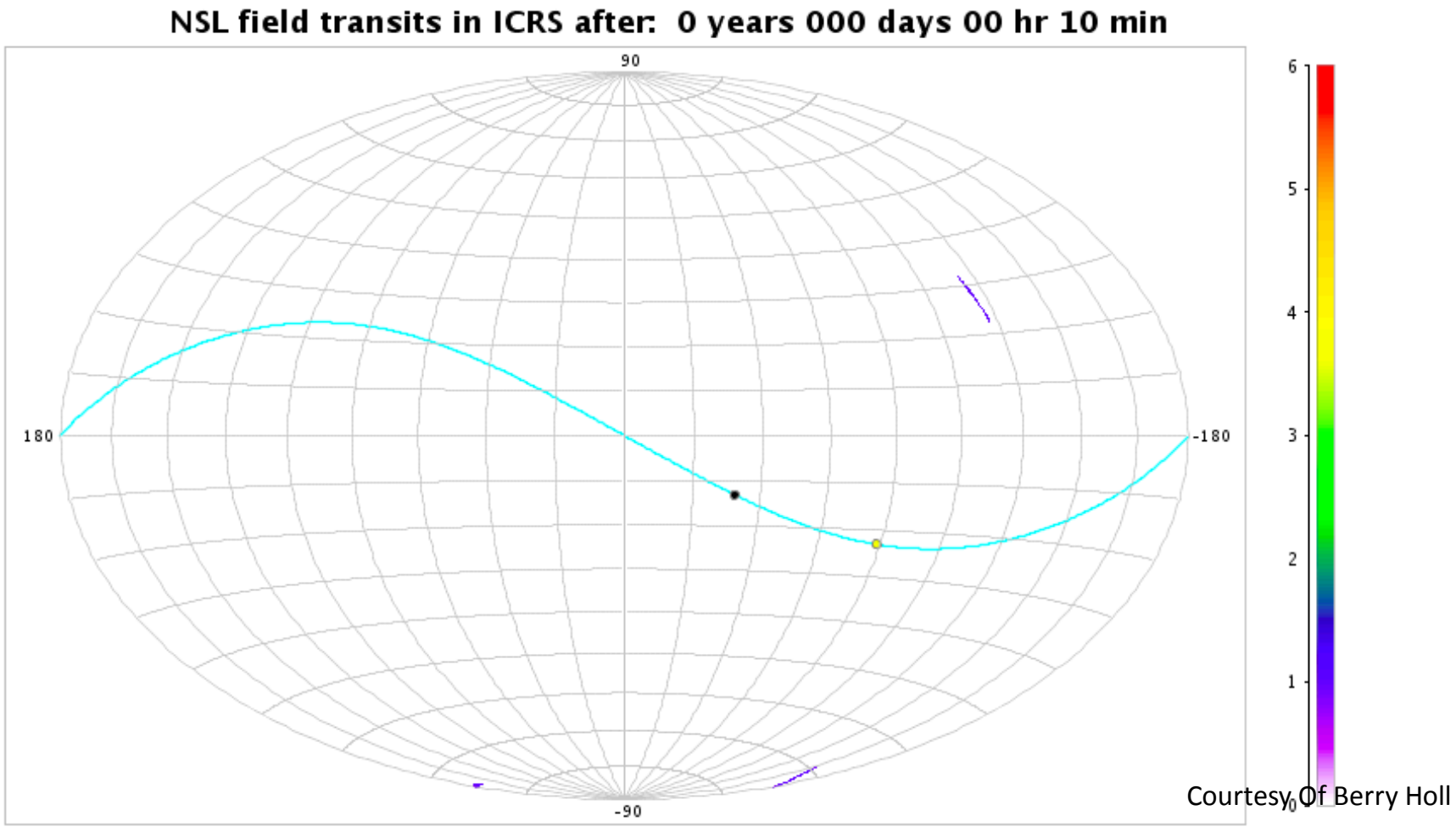
1997



2021



The Gaia scanning law

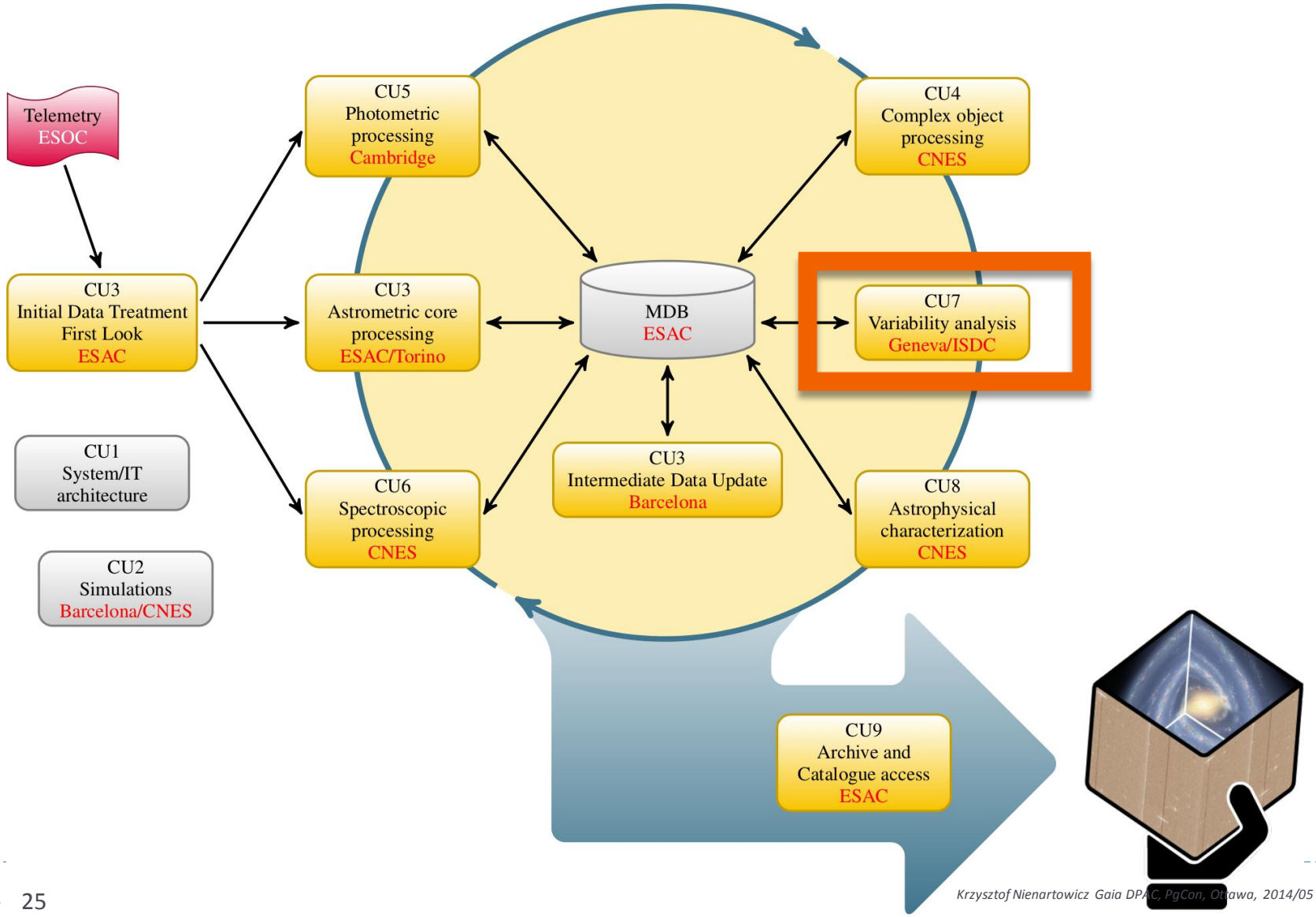


Courtesy Of Berry Holl

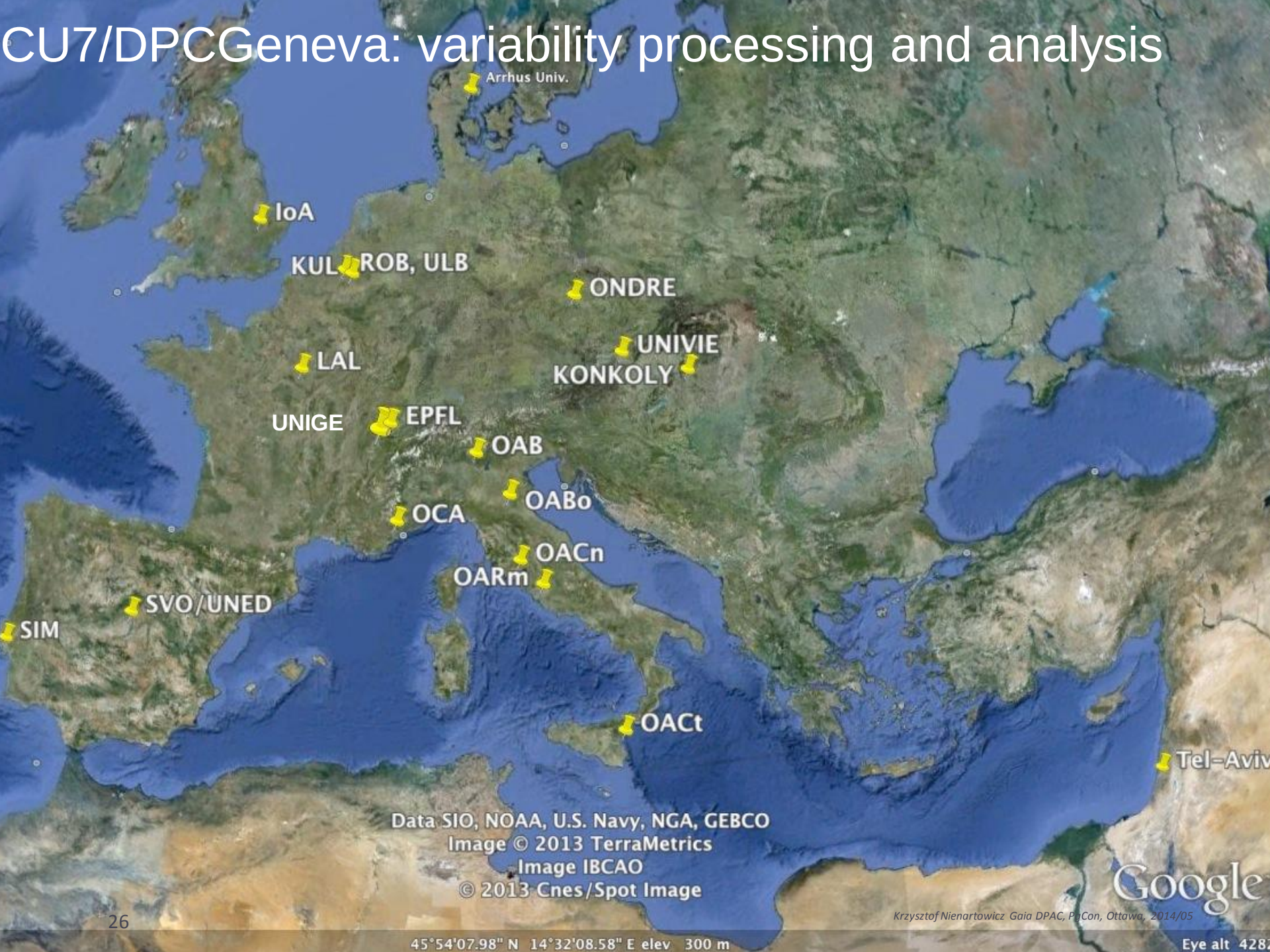
Mean of 70 G-band per-transit observations (40 - 250)

Gaia scientific responsibilities

Global dataflow



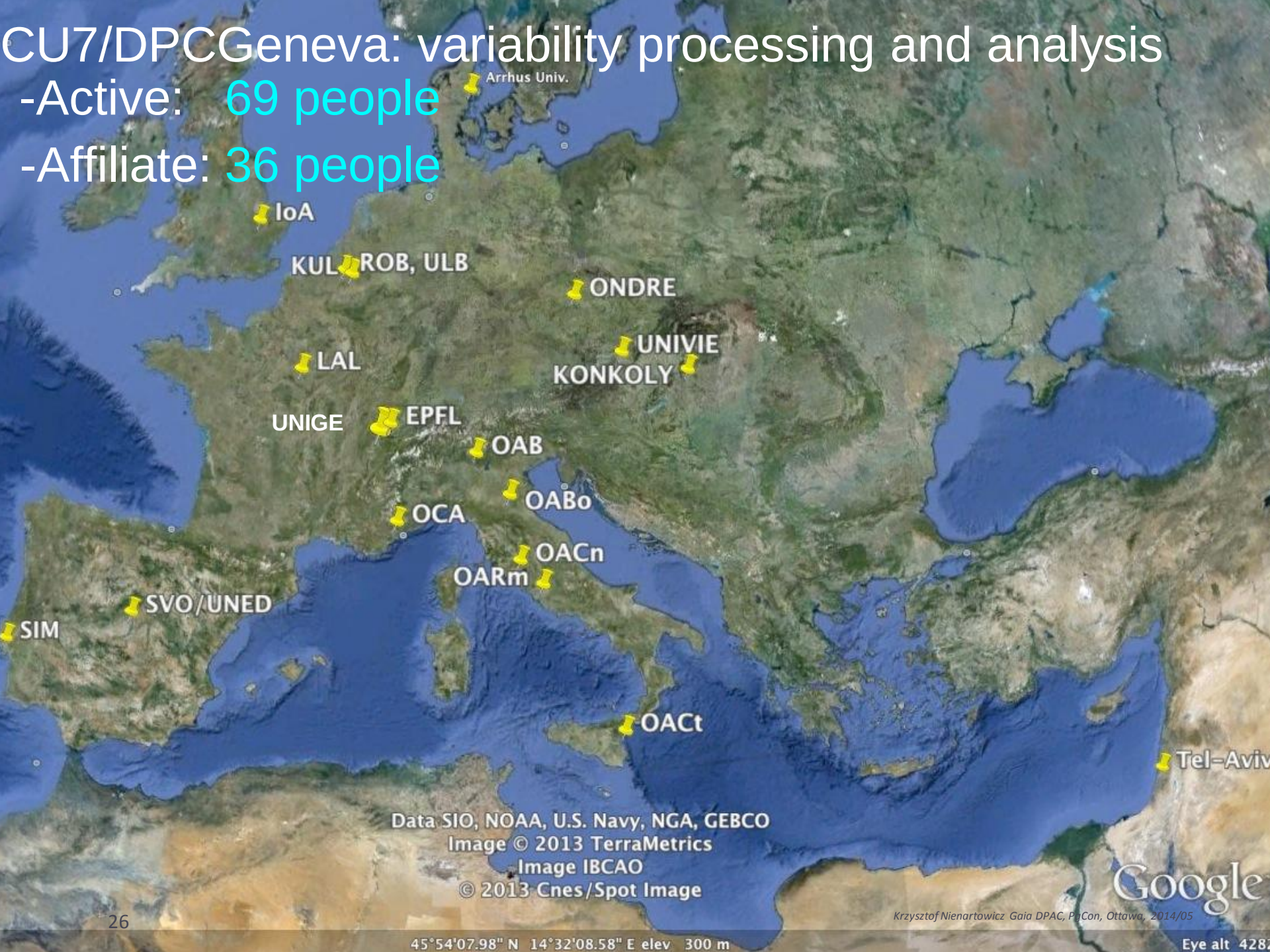
CU7/DPCGeneva: variability processing and analysis



CU7/DPCGeneva: variability processing and analysis

-Active: 69 people

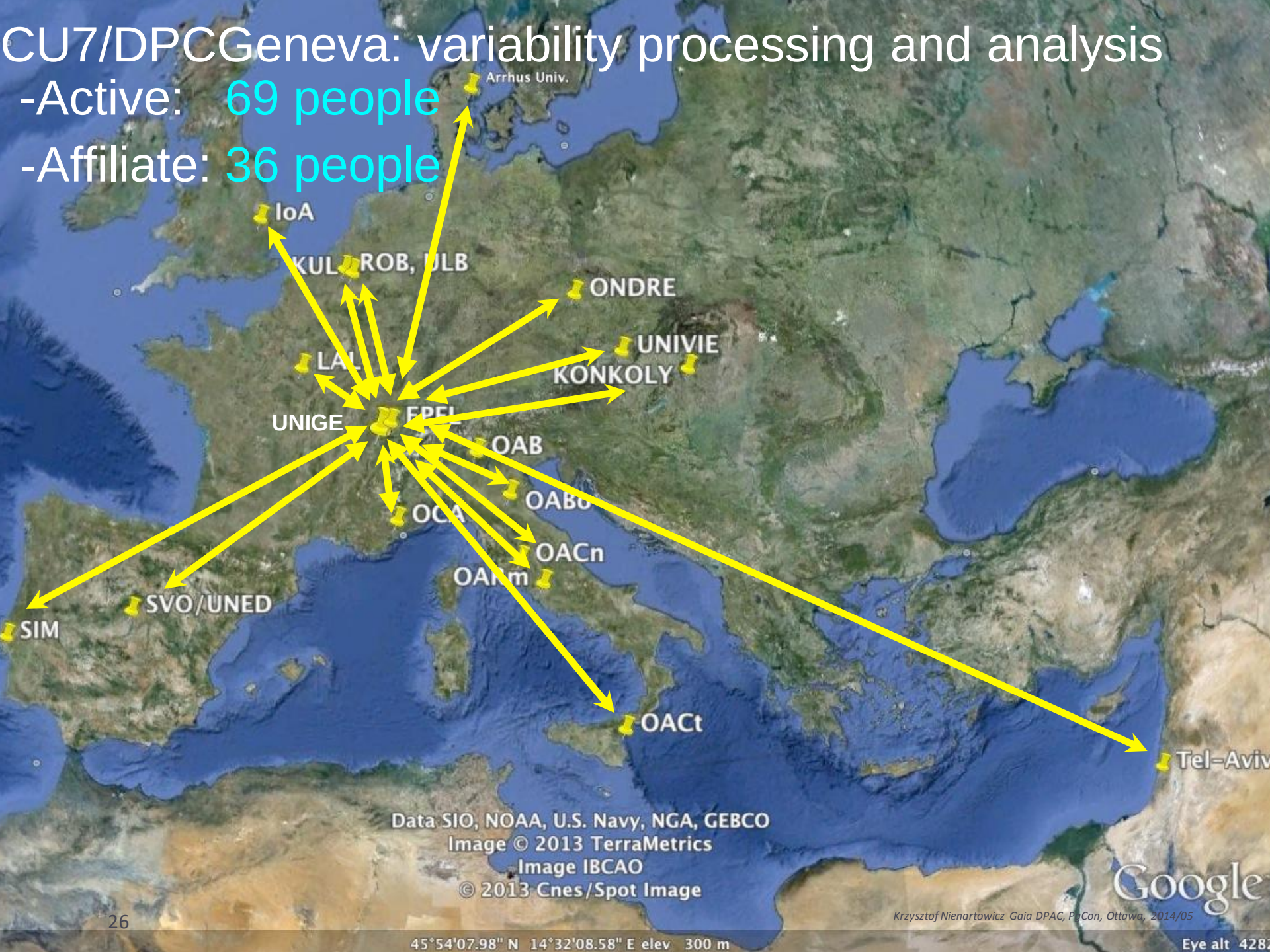
-Affiliate: 36 people



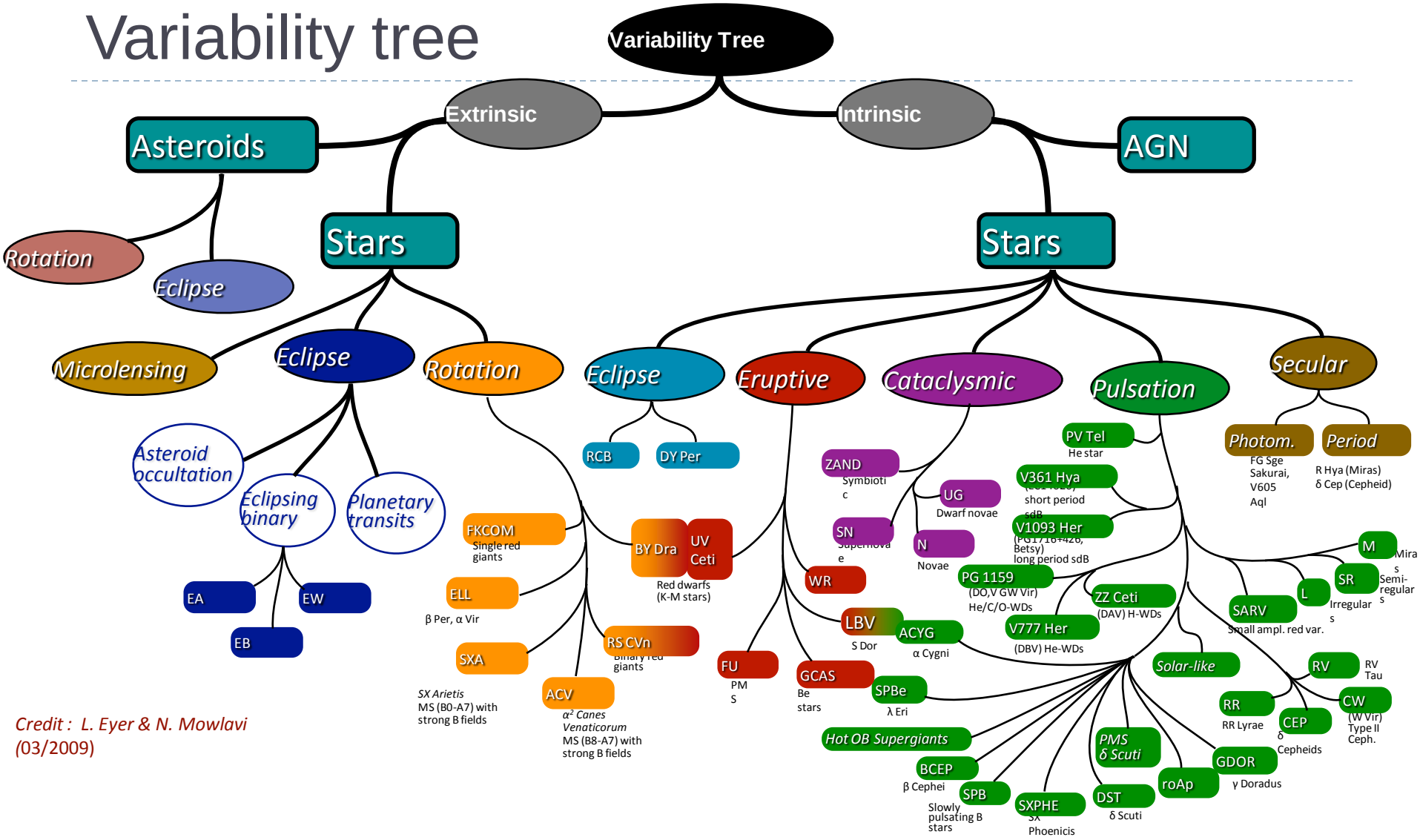
CU7/DPCGeneva: variability processing and analysis

-Active: 69 people

-Affiliate: 36 people

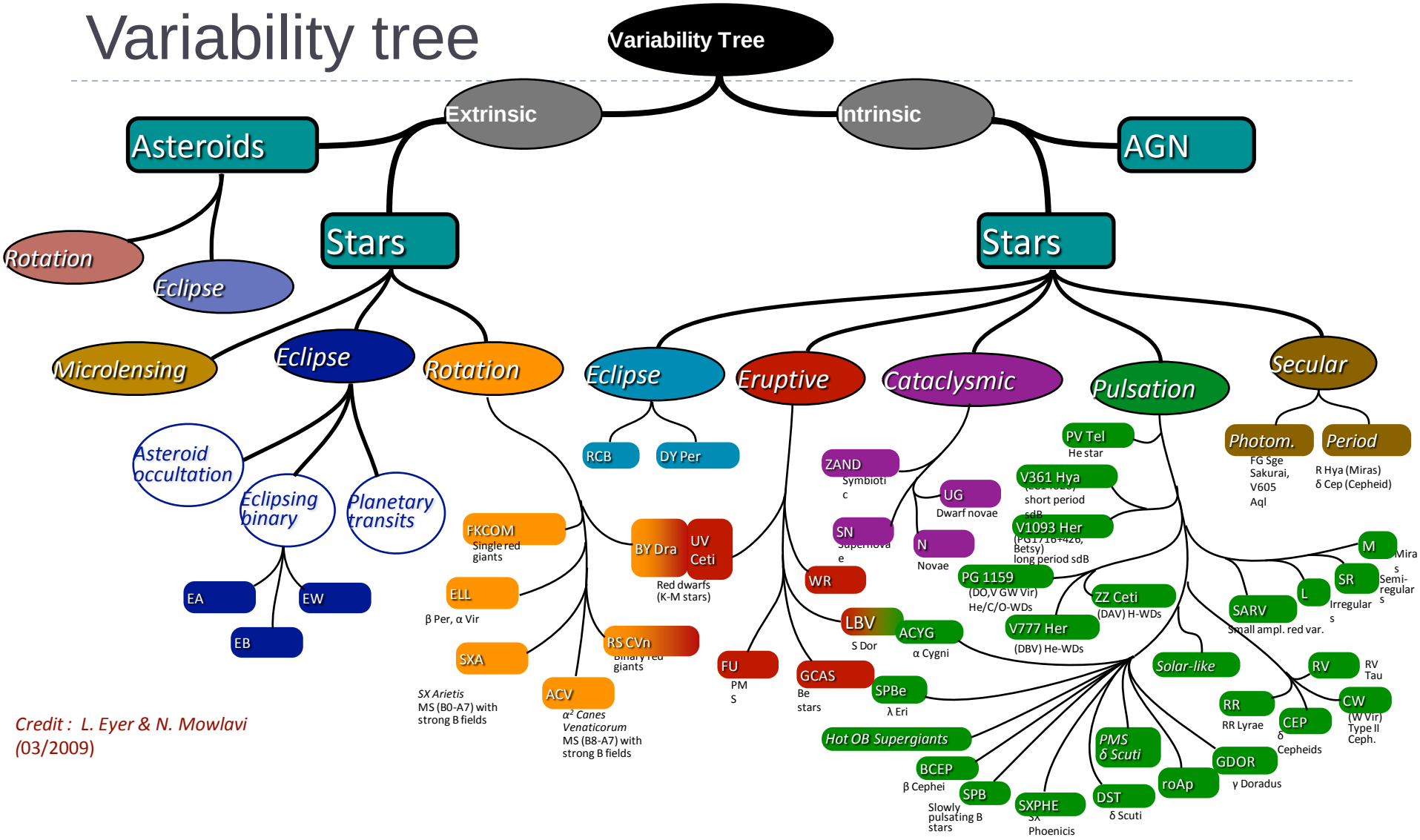


Variability tree



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

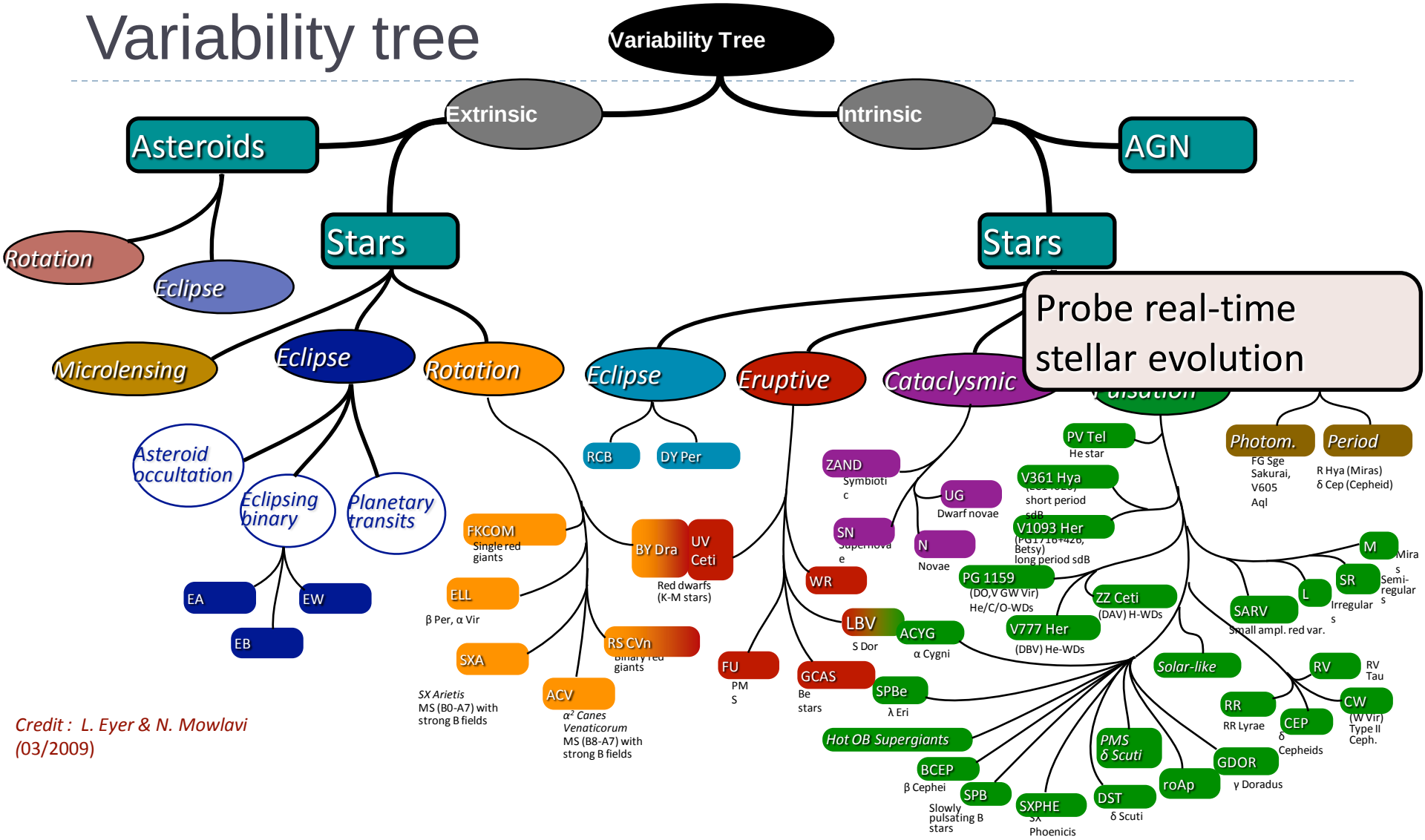
Variability tree



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

Gaia will detect most variable types on this tree

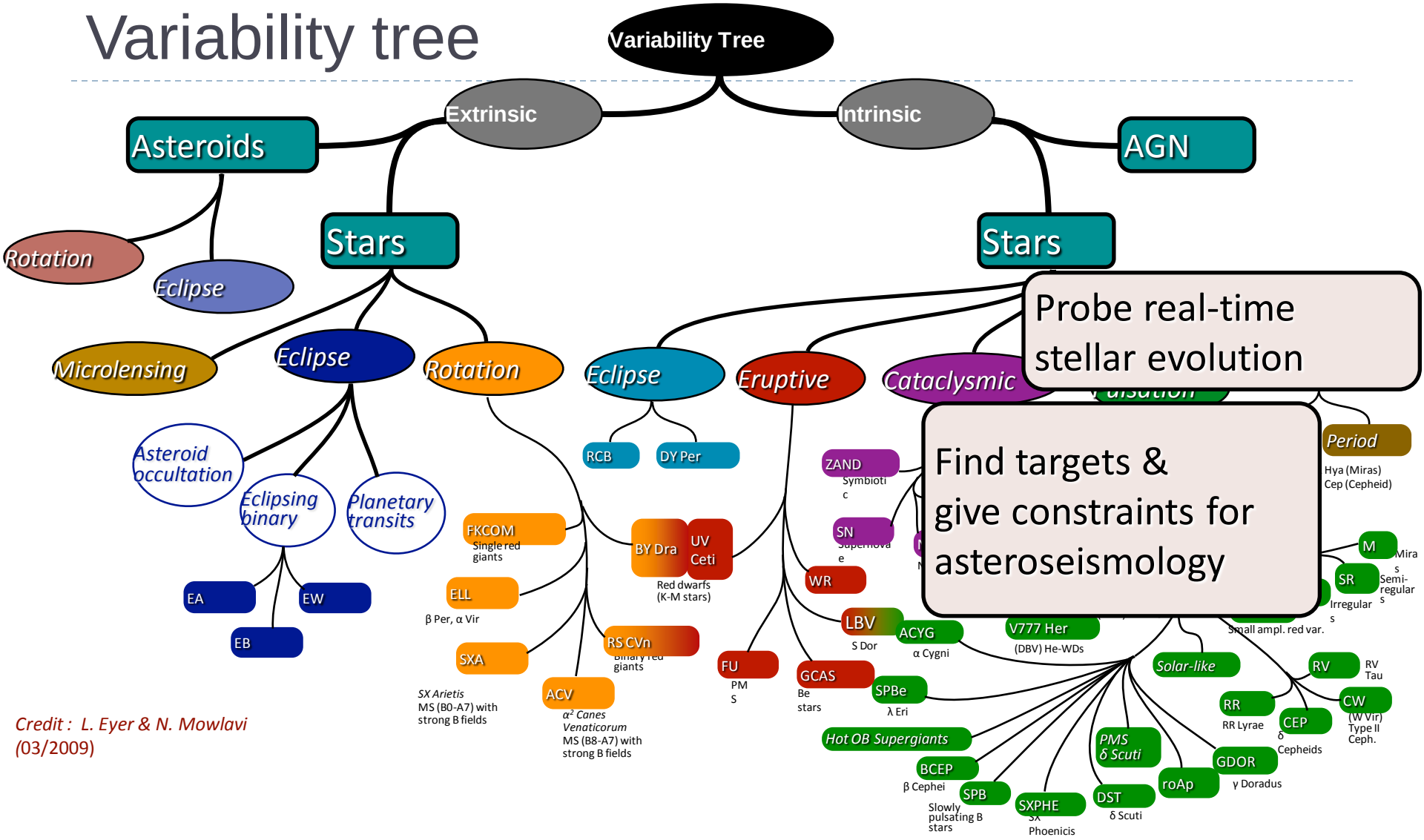
Variability tree



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

Gaia will detect most variable types on this tree

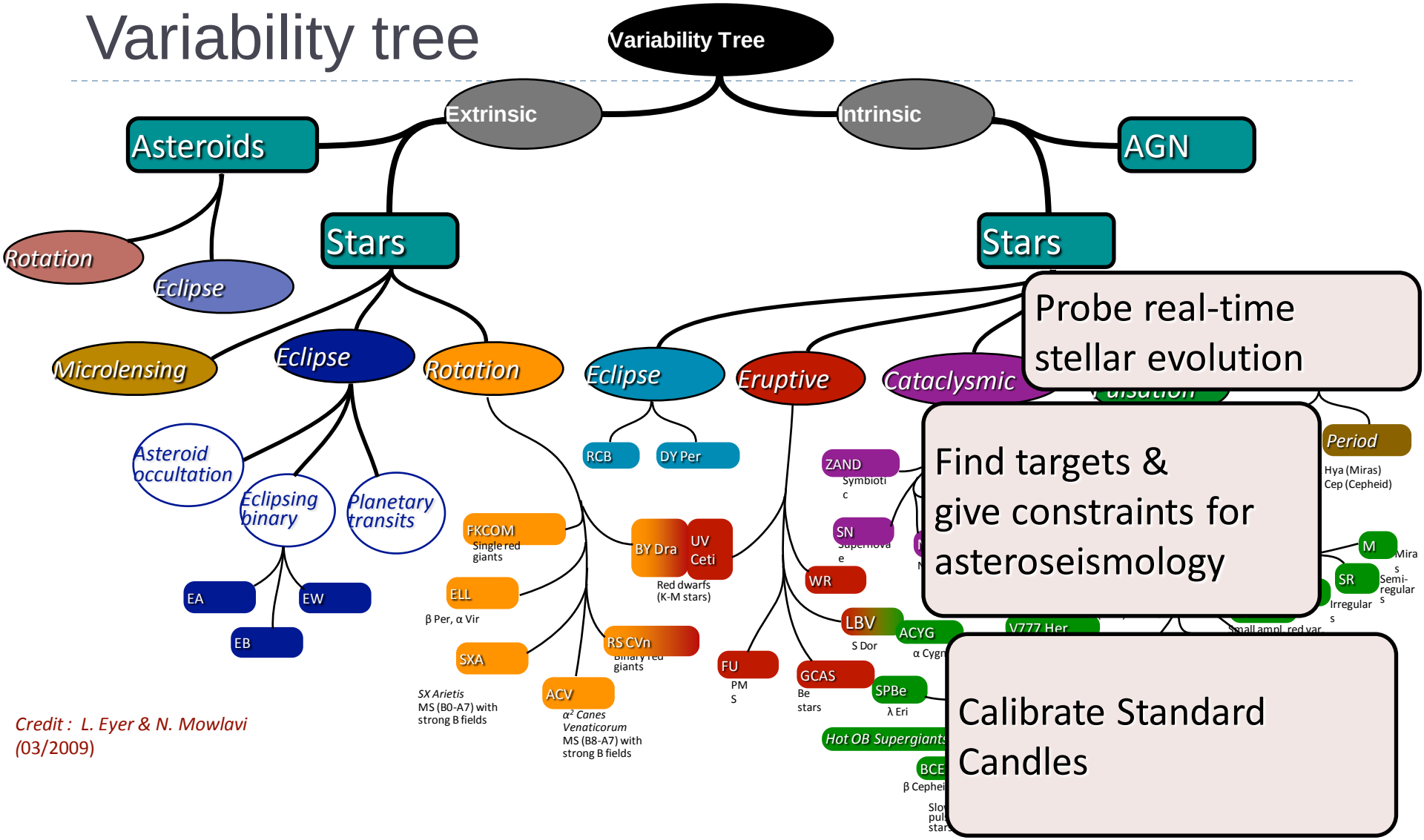
Variability tree



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

Gaia will detect most variable types on this tree

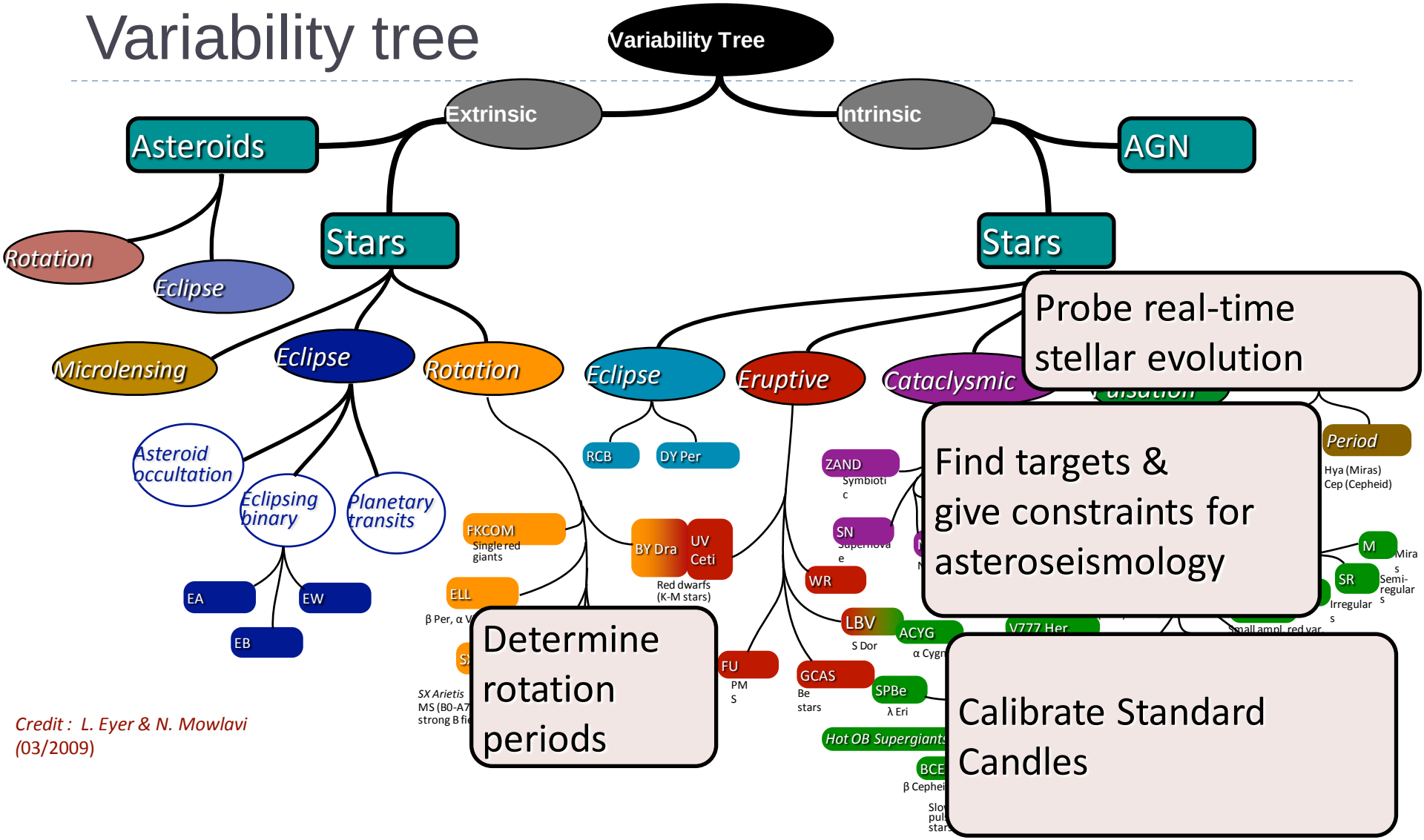
Variability tree



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

Gaia will detect most variable types on this tree

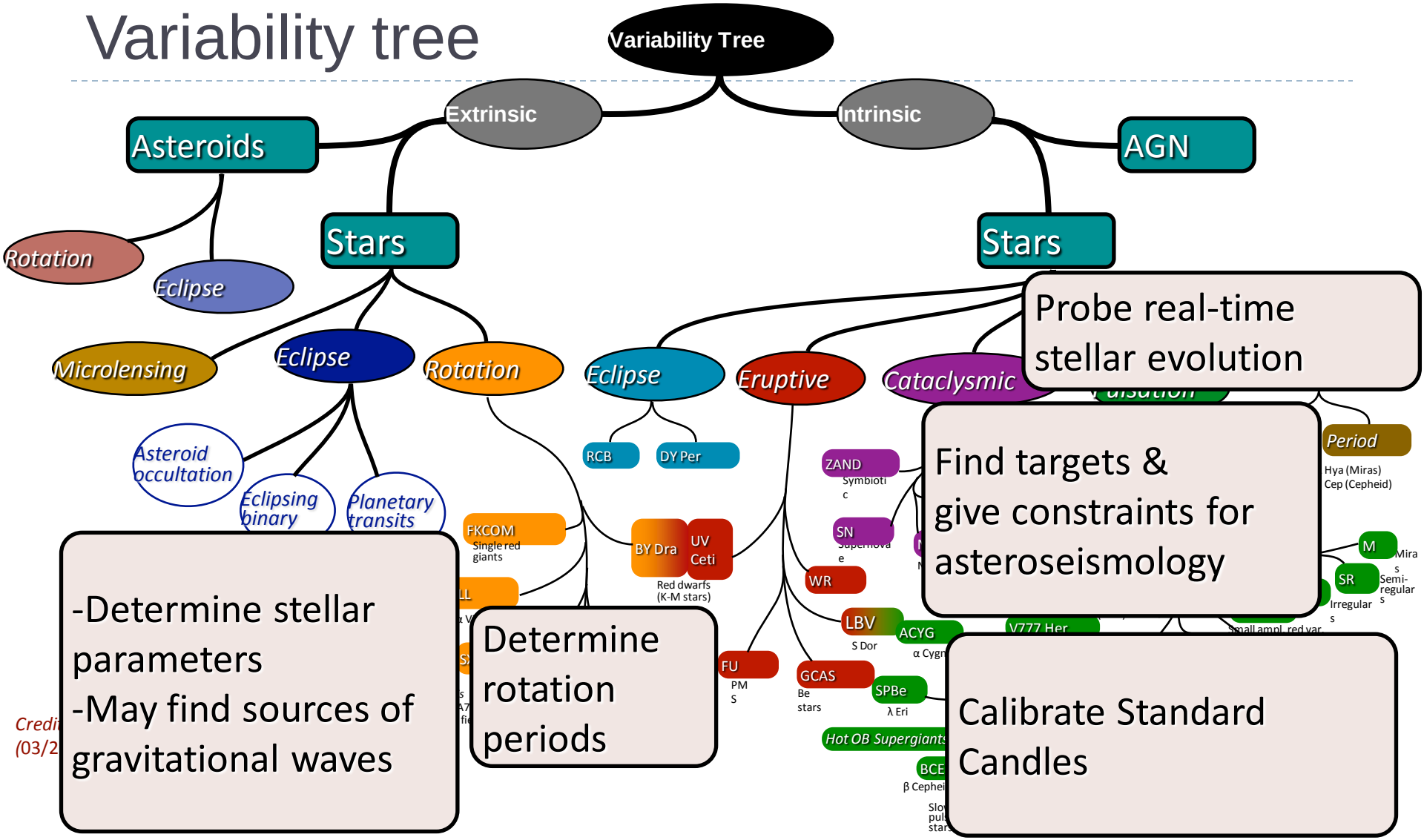
Variability tree



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

Gaia will detect most variable types on this tree

Variability tree



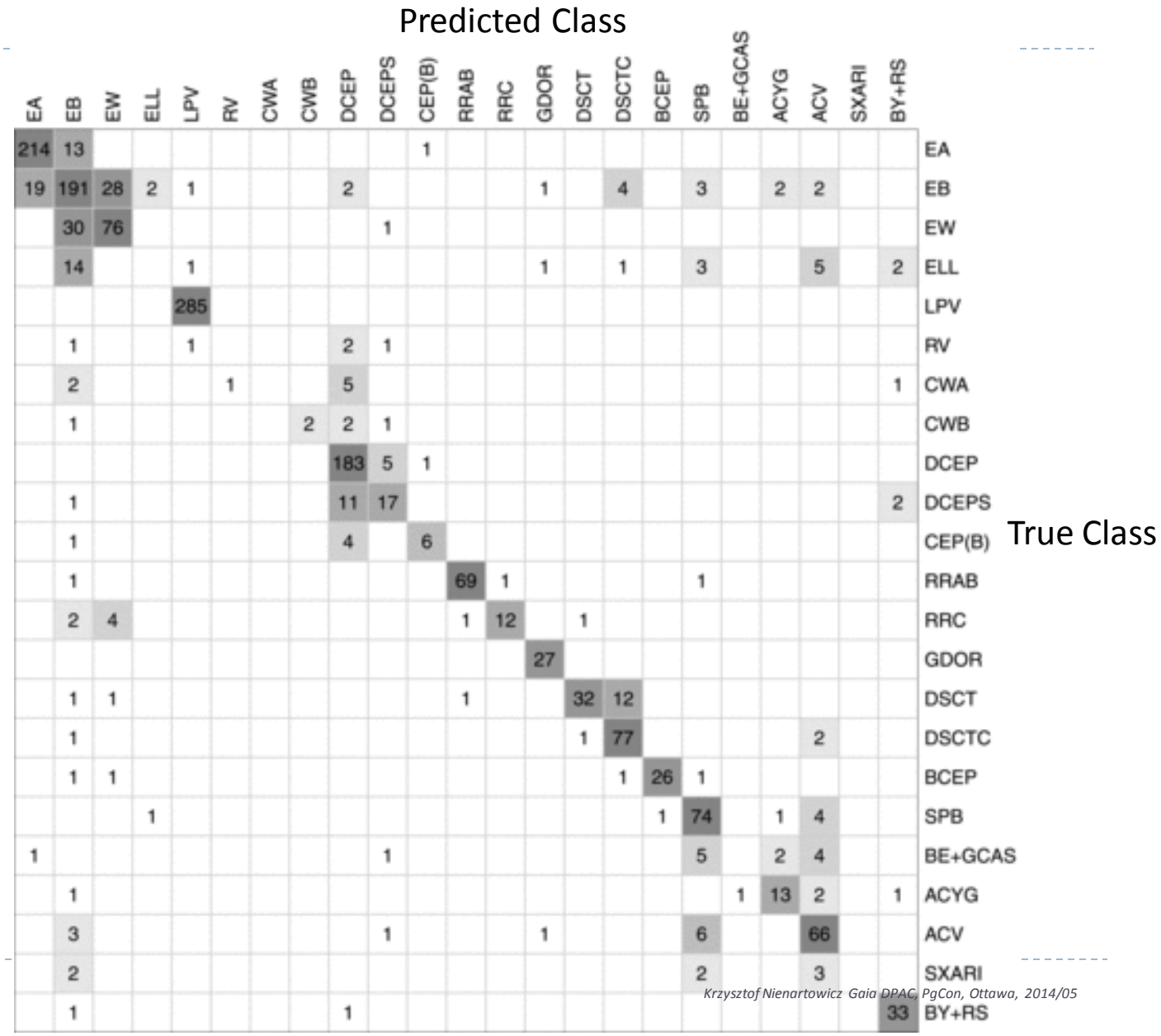
Gaia will detect most variable types on this tree

Classification of variable objects

Random Forest

Dubath et al. 2011

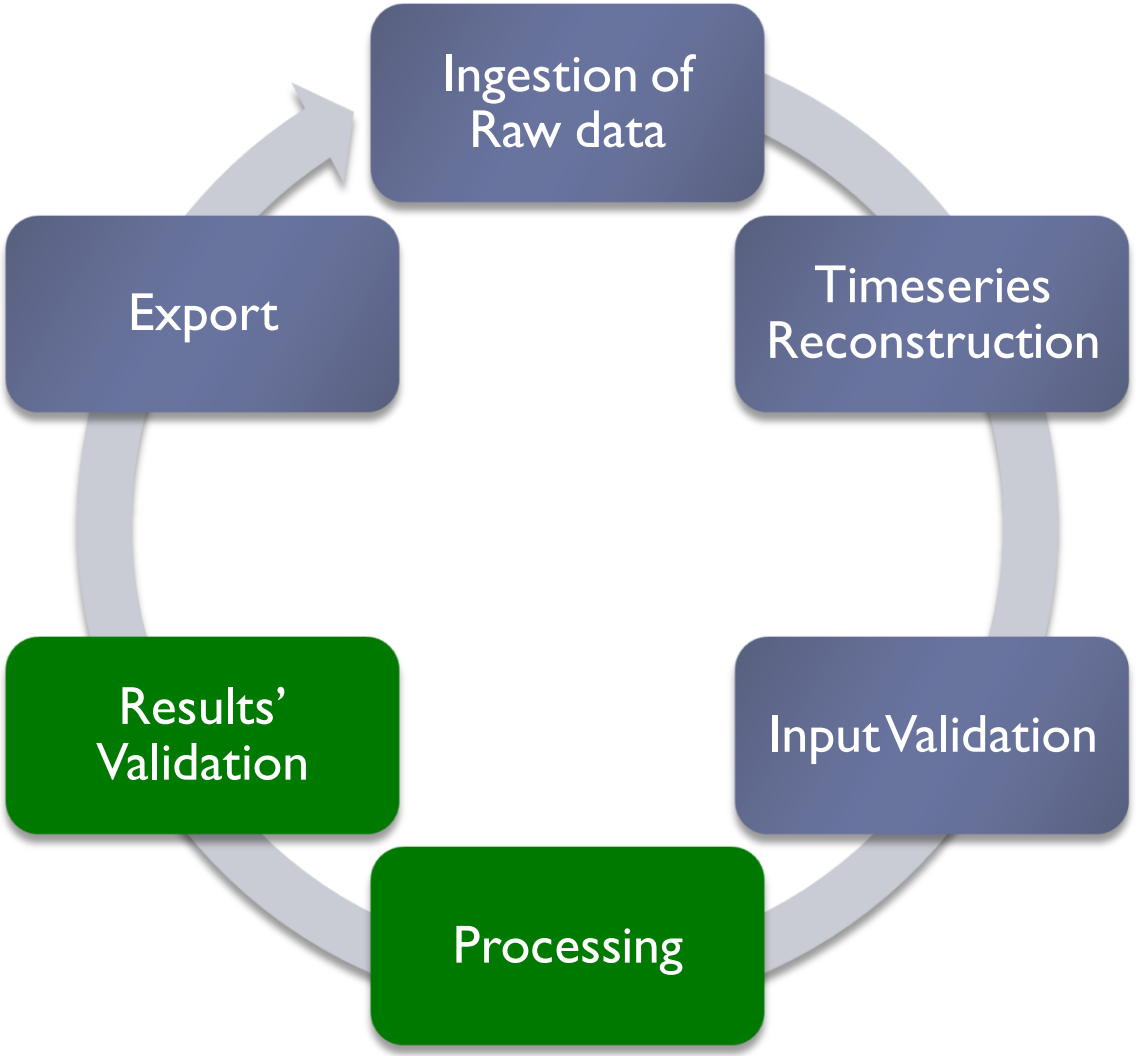
Similar study by Richards et al. 2011



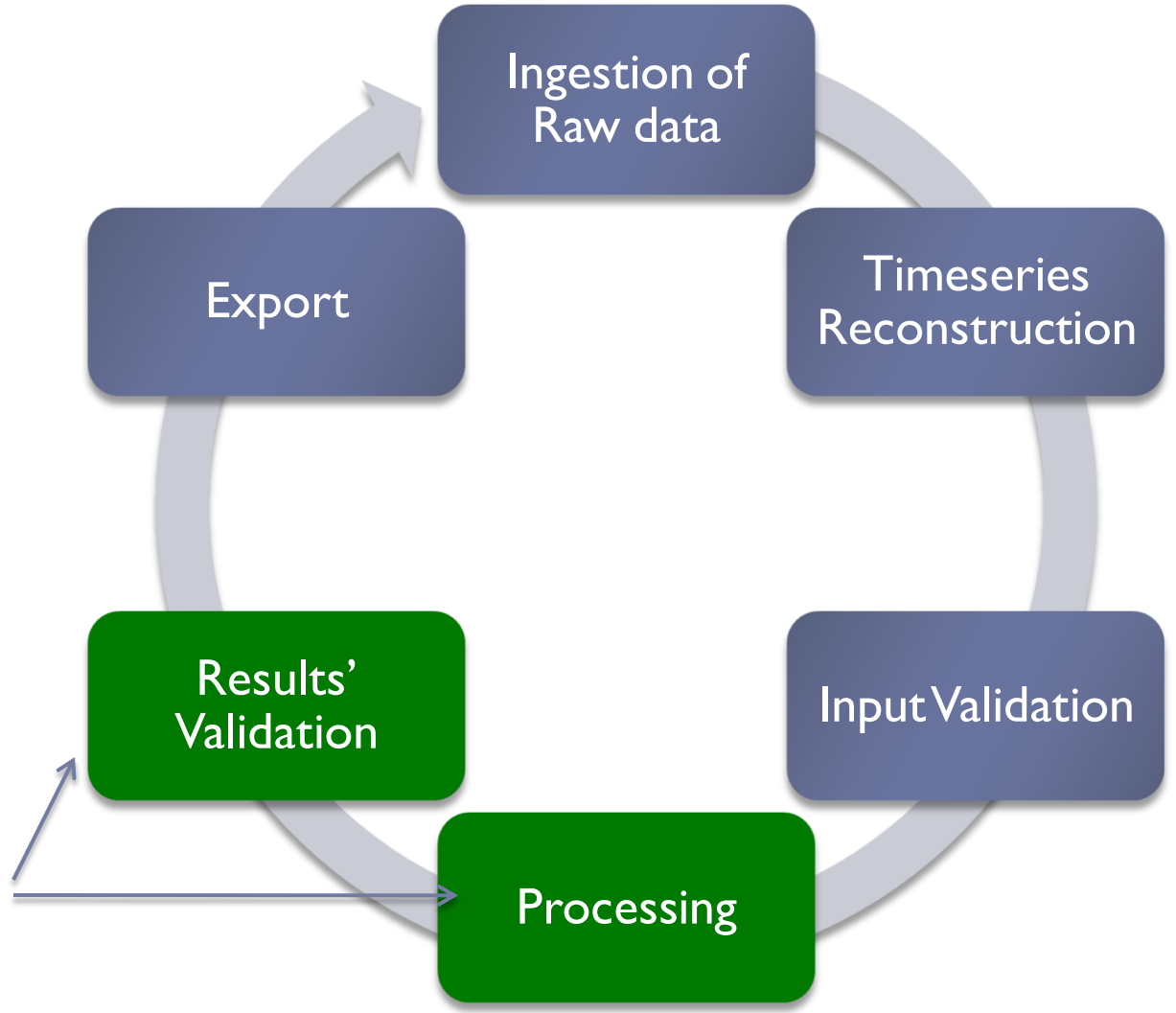
Structure

- My story
- Gaia mission
- Gaia science at CU7, Geneva
- **Processing model and Data model**
- XC role
- Hardware
- Collaboration

Cyclic Processing model
six months windows



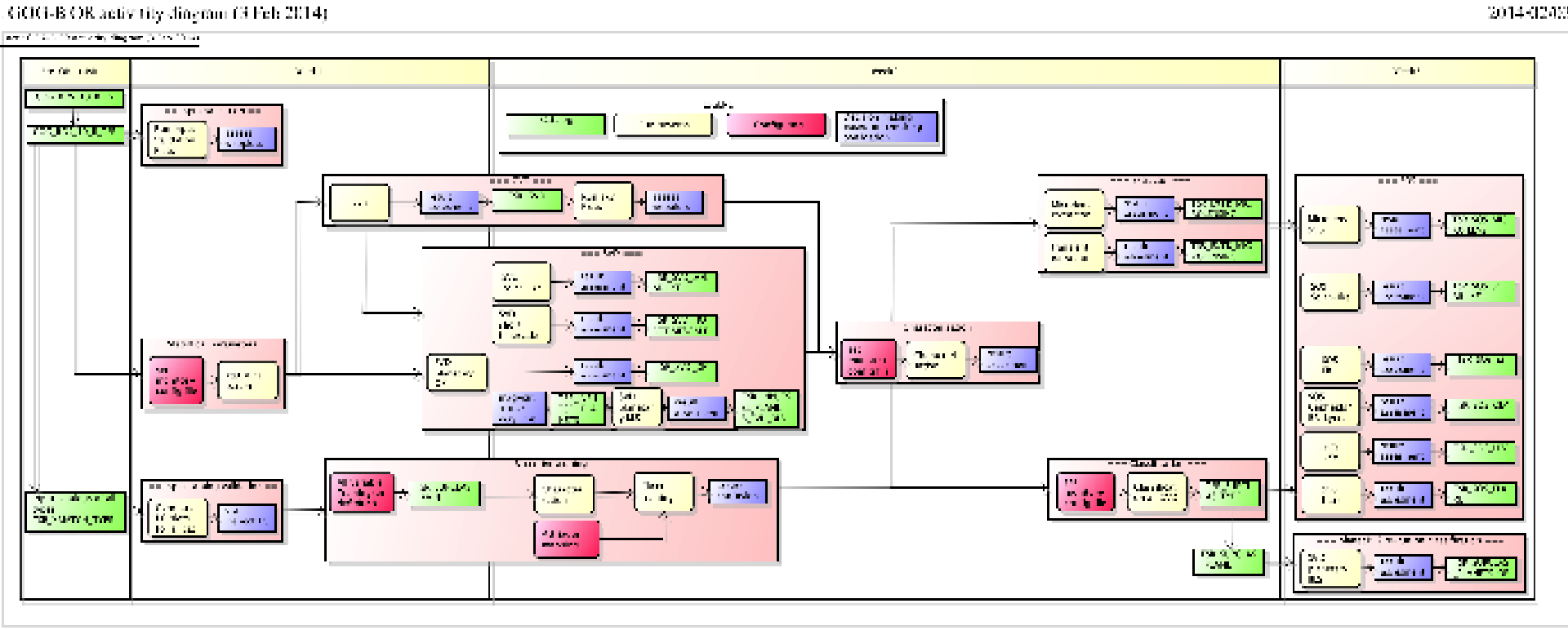
Cyclic Processing model
six months windows



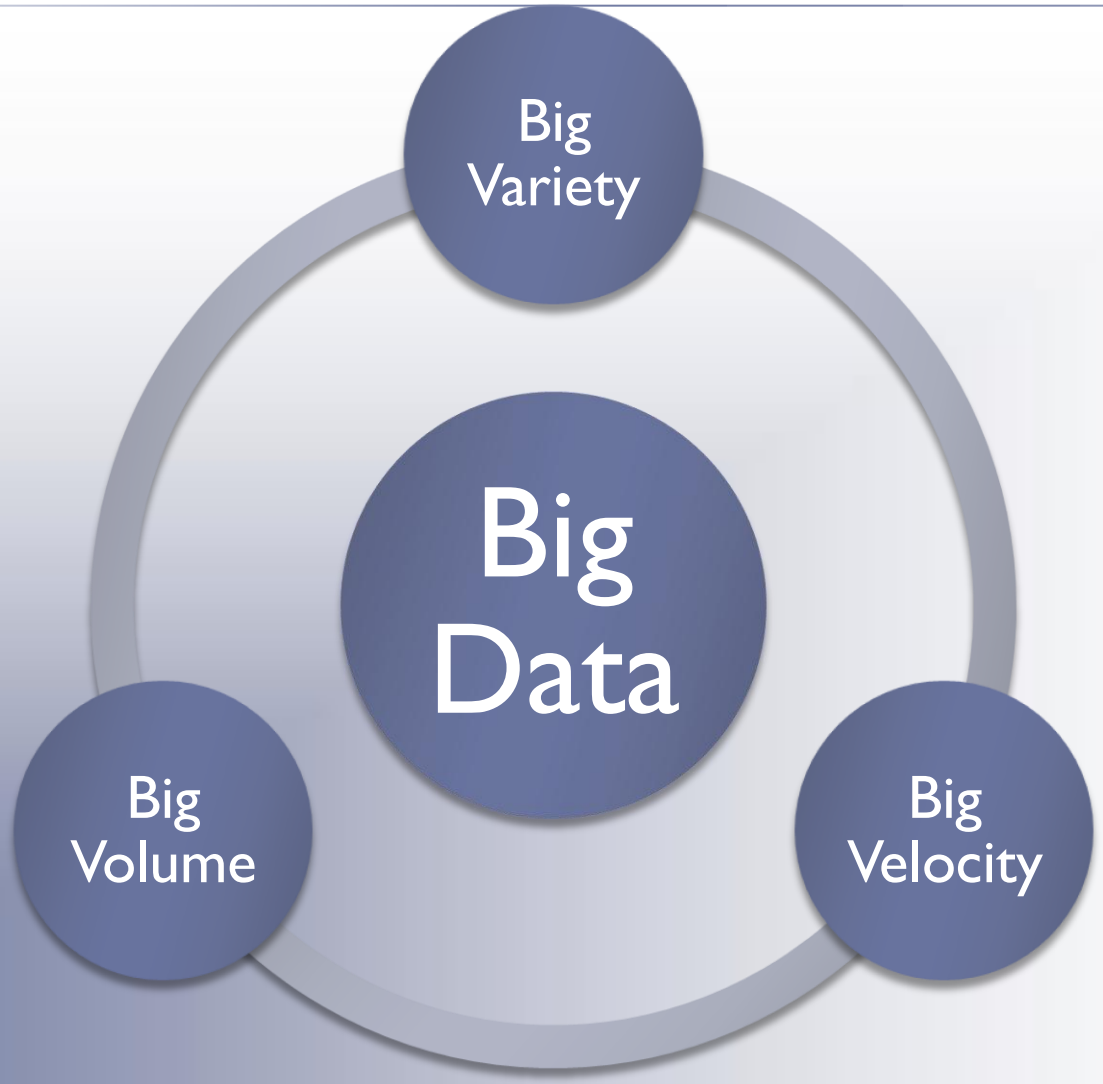
k-times during cycle on full dataset and samples < 30 days

Cyclic Processing model

one month processing sub-cycle

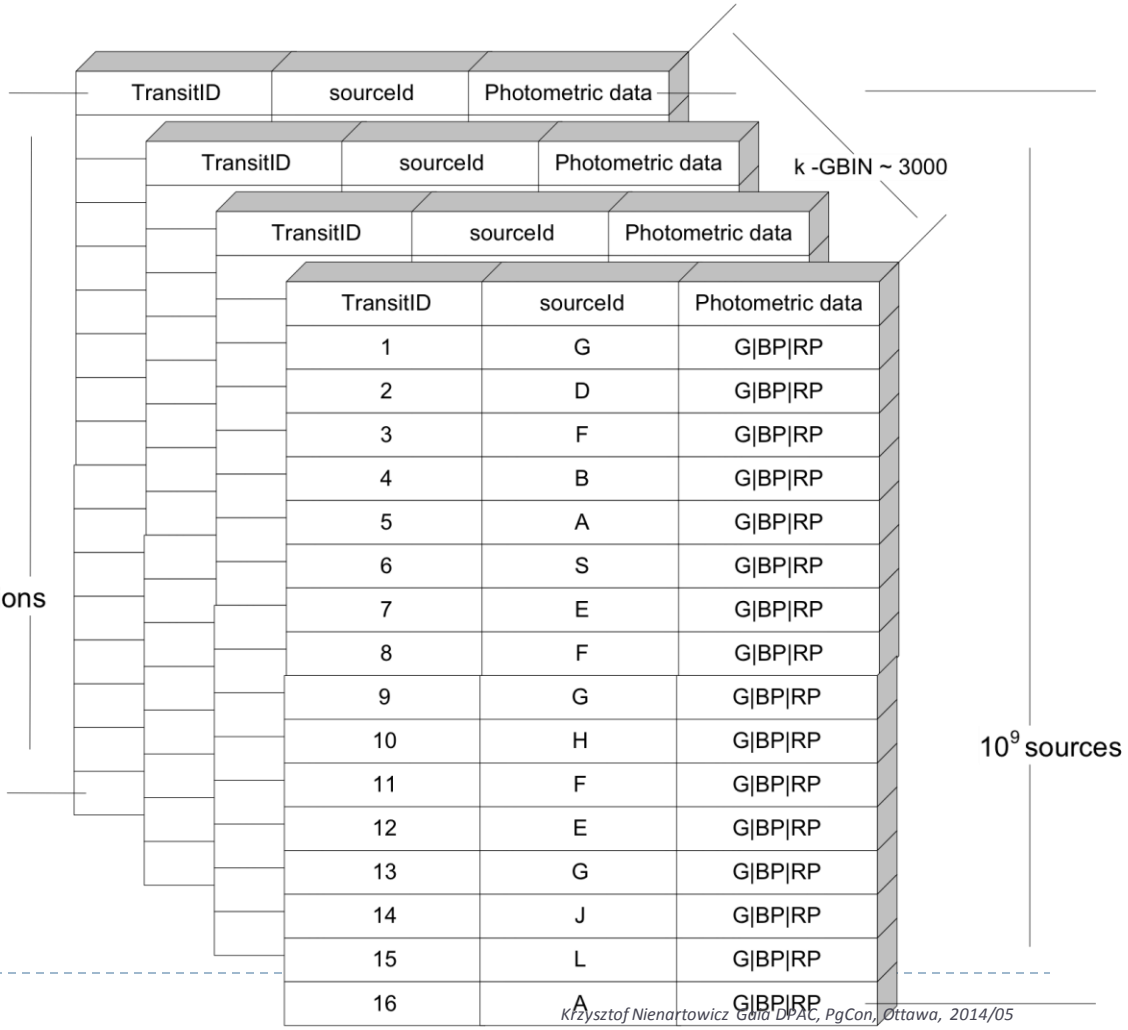


Big Data triangle
by M. Stonebreaker



Big Velocity
i.e. Ingestion

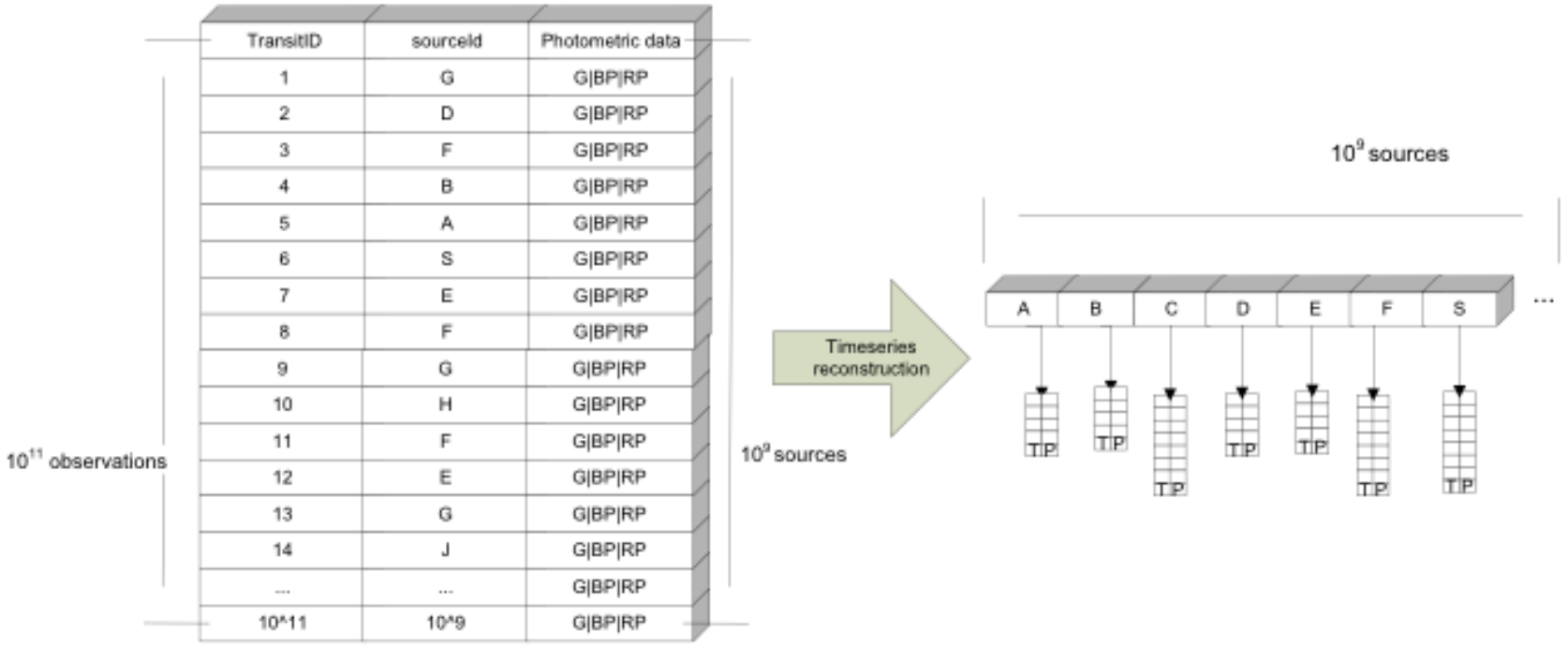
- All AstroElementaries distributed over multiple files.
  - We do not know how exactly
- GBINs are slow. CPU bound, not IO bound. <1MB/s
- 900 sec/1GB
- ~ 500h just to read 10^9*80 entries (single threaded)
- ~ 3-7TB of FovTransits (depends on the compression, we have 1/3 of fields now => G band only => 2.4TB of AE)
- ~3000 – 1GB compressed files to process for 10^9*80 tuples
- **Solution: Load into intermediate partitioned (distributed) table and run aggregating SQL reconstructing the timeseries**



Timeseries reconstruction

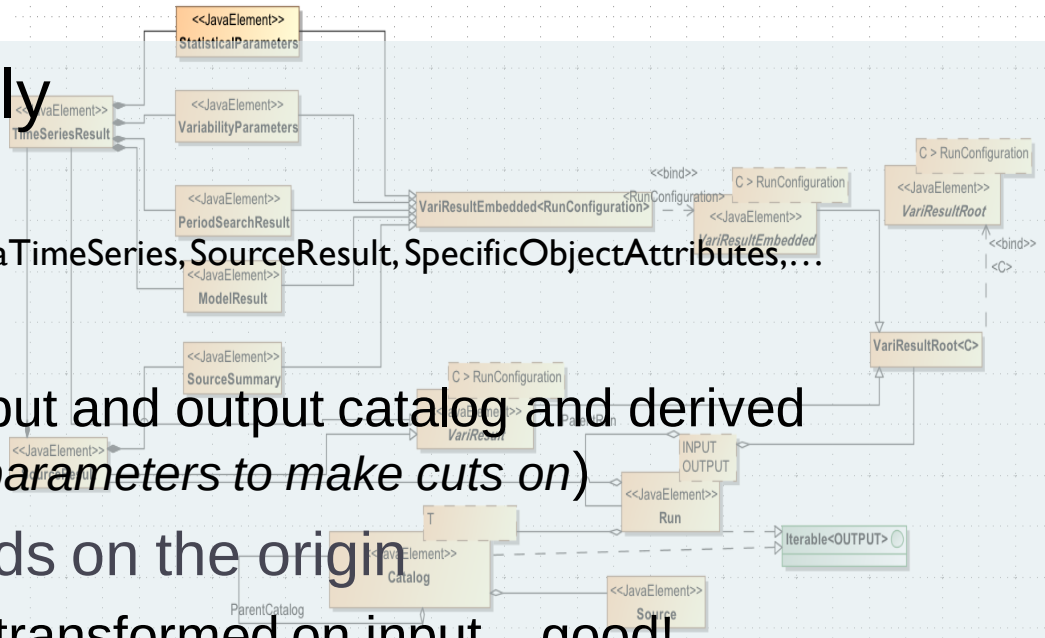
```
insert into gog_rds_10_c.ts(catalogid,sourceid,fvaluetype,obstimes,ftimeseriestype,vals,valserr,flags)
        -- convert and pivot data
        select
    gog_rds_10_c.getcatalogId('GOG_RDS_10_C') catalogid,
    sourceid,
    0 fvaluetype, -- flux type
    obs,
    (unnest(v)).*,
    StatusFlag from (
         select
         sourceid, converttobytea(array_agg(convertedTransitid)) obs,
         array[
         (getTsTypeId('GOG_RDS_10_C','GAIA_PHOT_G'),  converttobytea(array_agg(gflux)), converttobytea(array_agg(GFLuxError)) )::ts_buckets_bin
         (getTsTypeId('GAIA','BP'), converttobytea(array_agg(BPFlux)),converttobytea(array_agg(BPFluxError)) )::ts_buckets_bin ,
         (getTsTypeId('GAIA','RP'), converttobytea(array_agg(RpFlux)),converttobytea(array_agg(RpFluxError)) )::ts_buckets_bin
        ] v
         ,convertinttobytea(array_agg(StatusFlag)) StatusFlag
        from (select ( ((transitid >> 5) & CAST(x'003FFFFFFFFF000' AS bigint))::bigint * 50 / 10^9 / 60 / 60/ 24) convertedTransitid, calphot.*
            from gog_rds_10_C.calphotfovtransitimpl calphot join source s on (calphot.sourceid = s.sourceid) order by calphot.sourceid,1 ) a
        group by sourceid
    ) pivoted
    ;
```

Timeseries reconstruction



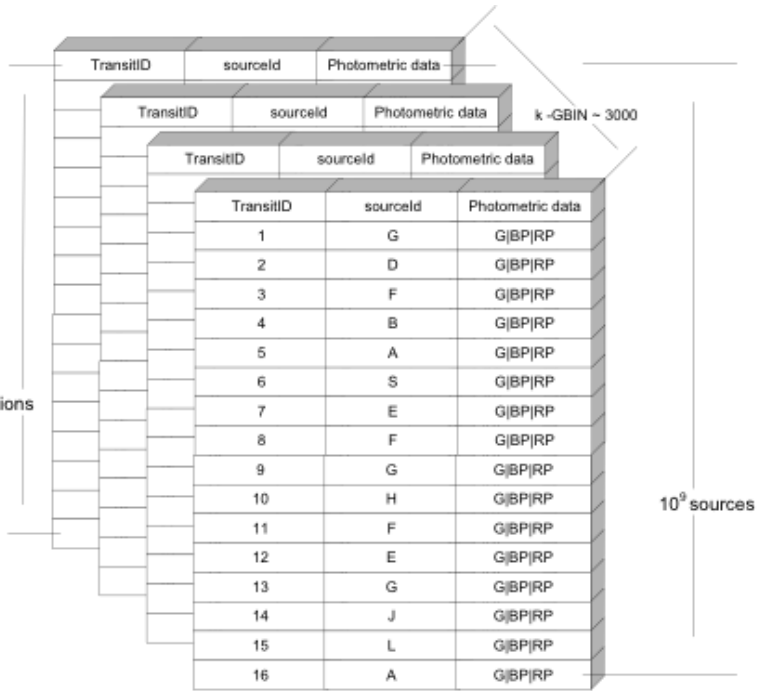
Big Variety

- CU7 OM has approximately
  - ~ 70 domain classes
    - Source, **Timeseries**, SpectraTimeSeries, SourceResult, SpecificObjectAttributes,…
  - Attributes
    - Possibly tens to hundreds input and output catalog and derived attributes used in analytics (*parameters to make cuts on*)
  - Input transformation depends on the origin
    - Gaia ~ 5-10 MDB tables are transformed on input – good!
    - Other surveys
      - From semi/quasi structured to structured formats -EAV
  - Production of the output is rather simple
    - **Analyzing it: Not!**

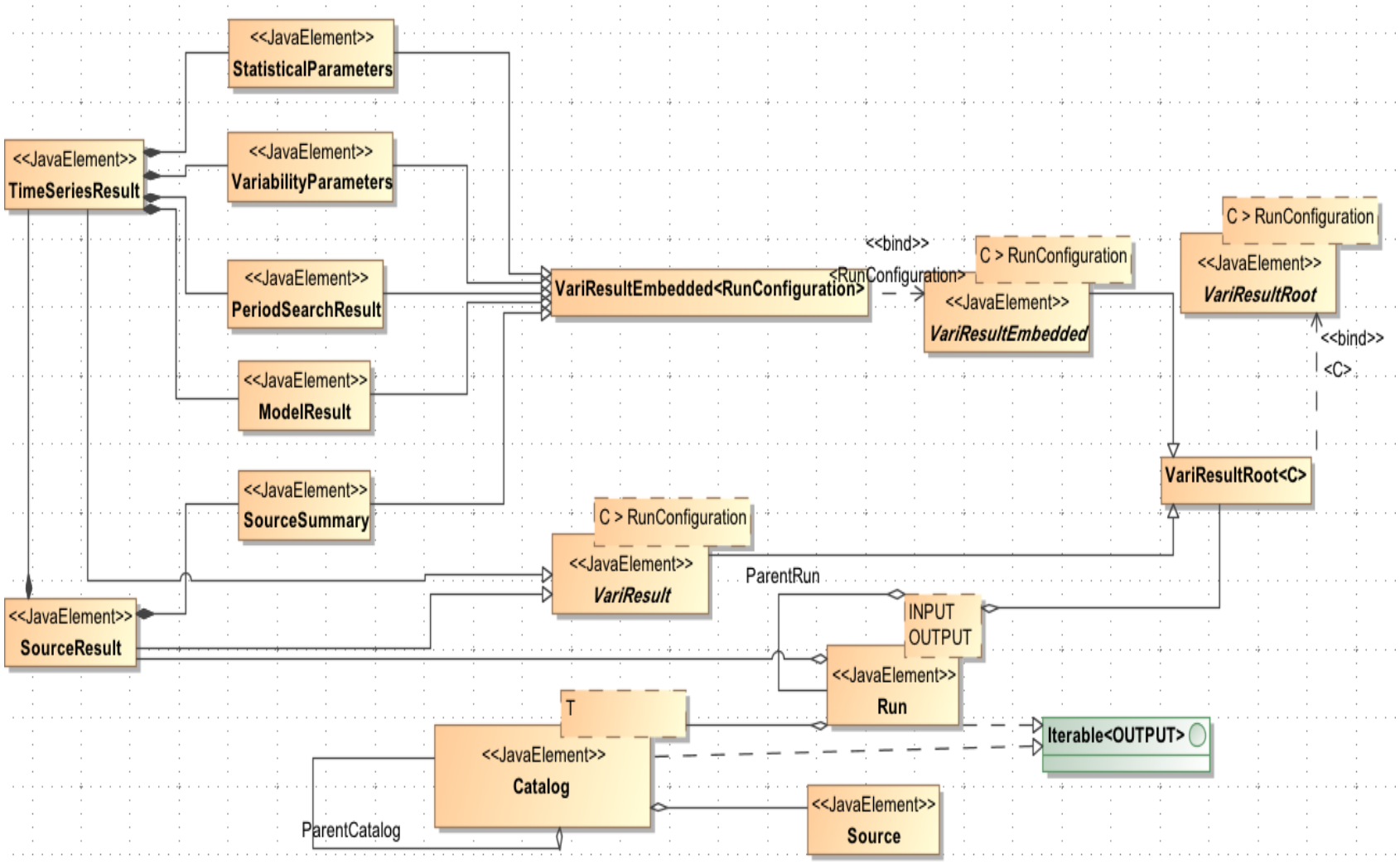


Big Volume

- ~50TB … 800TB
- Input
  - Hundreds catalogs,10^9 Sources, 5 x 10^9 Timeseries
  - Existing relatively large surveys
    - EROS, SuperWASP
  - Gaia input via GBINs
    - Deserialization
    - Reconstruction
  - Analysis of the input to find outliers,
    - Finding bugs in data in minutes – possible scanning 10^{11} entities in 10's of seconds
- Output
  - k-iterations on input: 10^9 source results, 5 x 10^9 TimeseriesResults
    - Results on all processed Sources/Timeseries
      - Re-run multiple times
    - Intermediate analytical results
    - Ad-hoc analytical results
  - Export of data into compressed GBINs



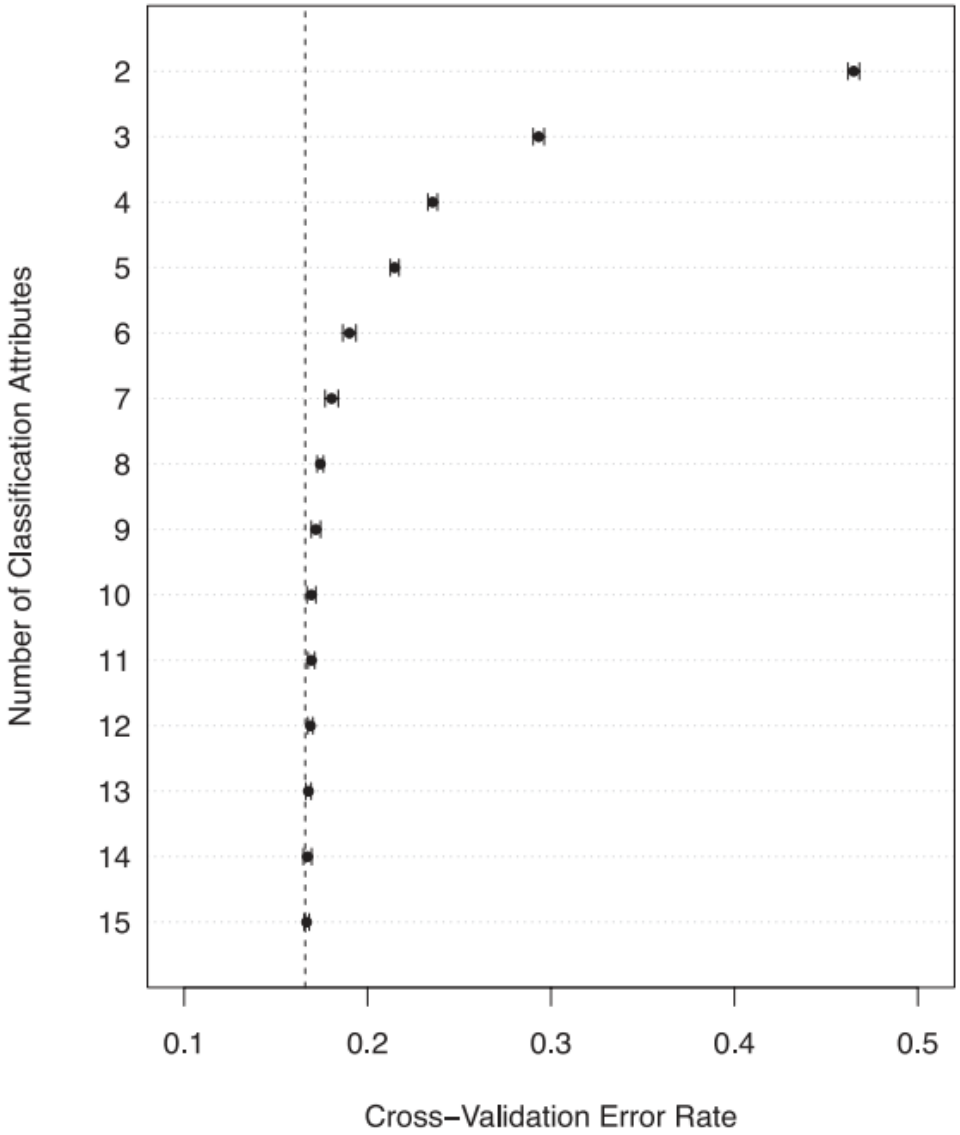
Data model
Java Persistence API, custom PG mappings



Data model
Java Persistence API, custom PG mappings

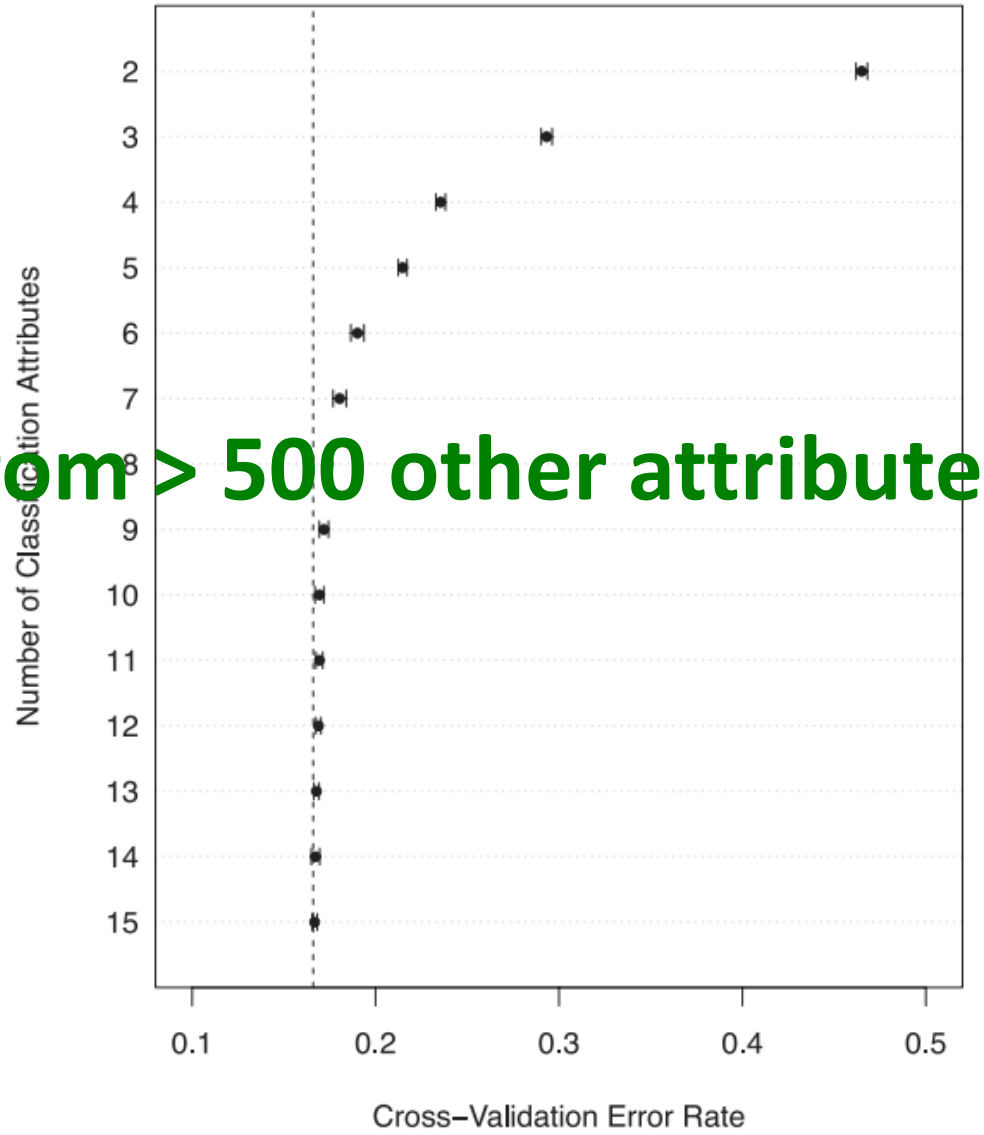
- Native ORM mappings:
  - 1-> *
  - 1-1:
    - Embedded
- Custom OO mappings:
  - hstore + metadata: EAV
  - PG objects <compositeType>
  - Arrays of PgObjects
  - Arrays of Arrays of Arrays of … PG Objects

Attributes



L. Rimoldini et al 2012

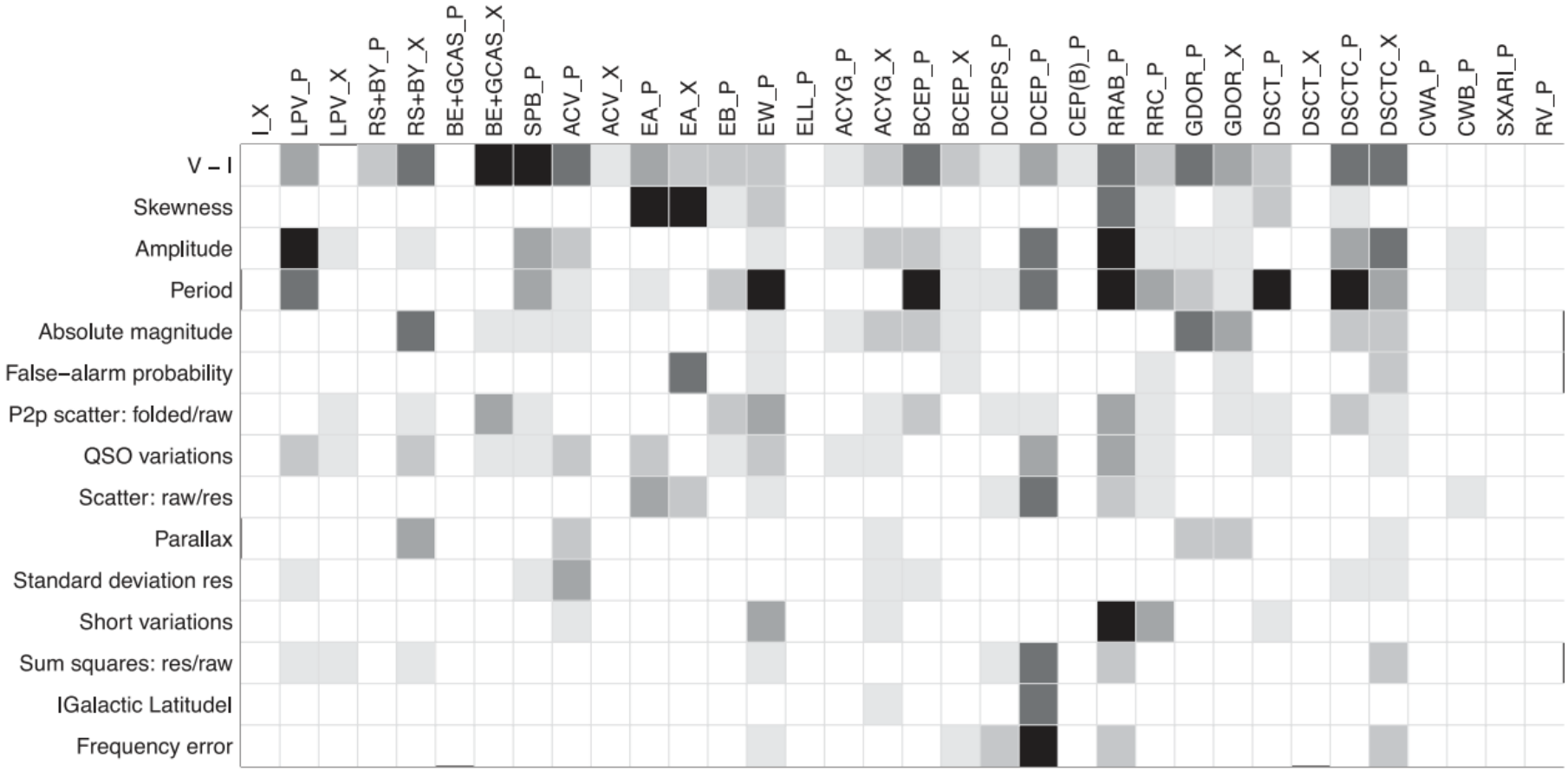
Attributes



From > 500 other attributes….

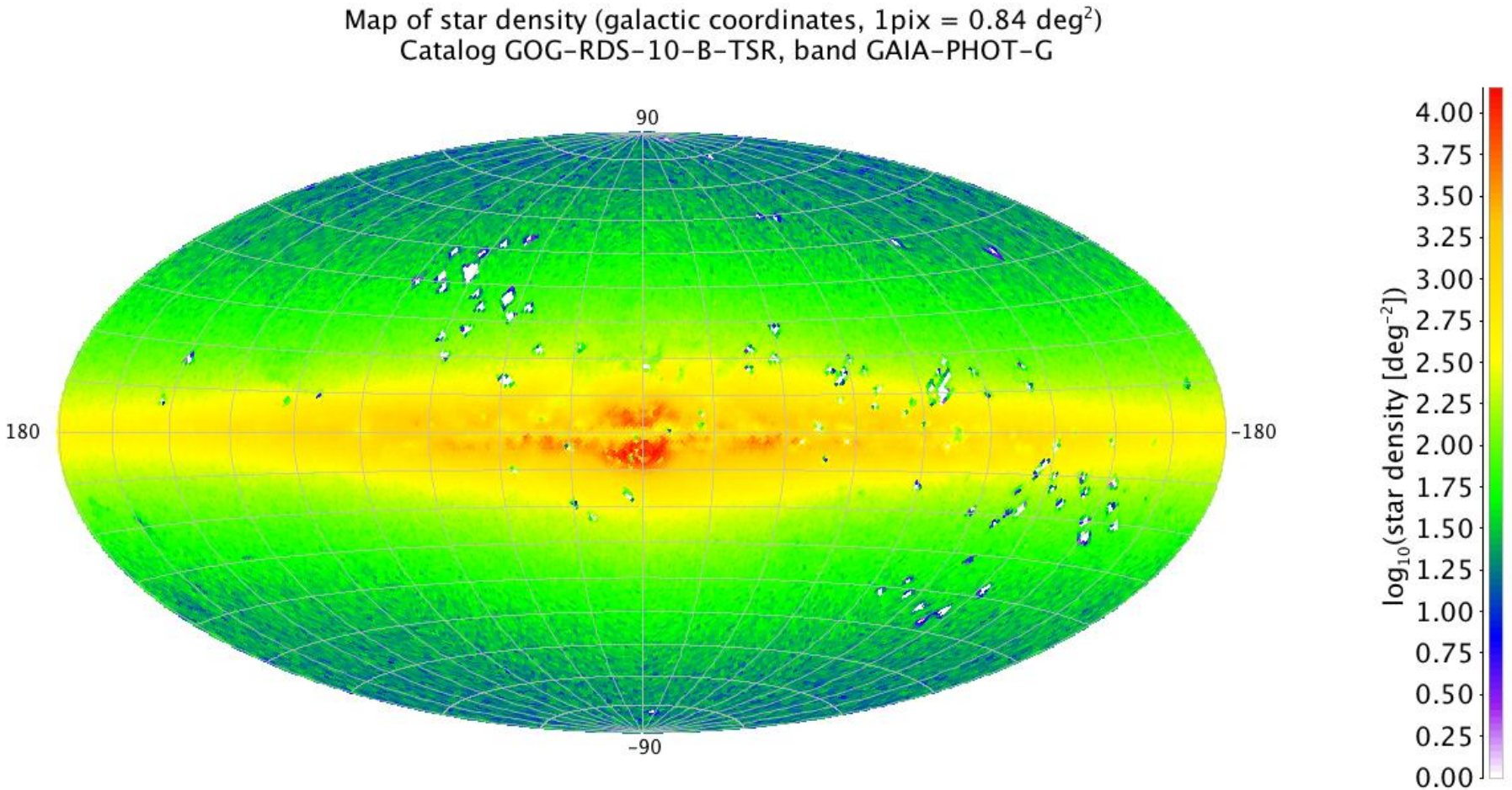
L. Rimoldini et al 2012

Attributes
(columnX, array[i], (pgType[j]).type2[i]…)

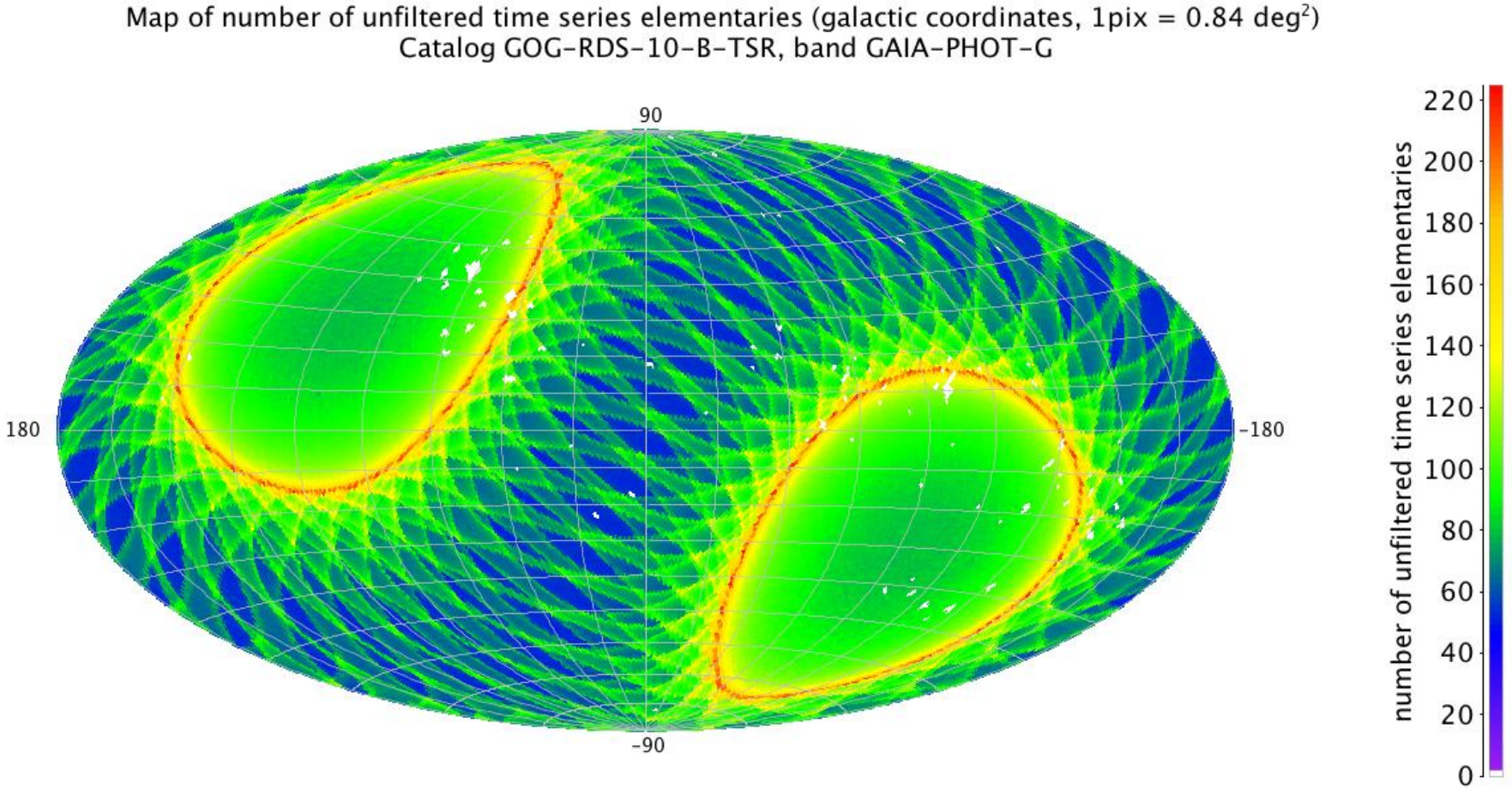


L. Rimoldini et al 2012

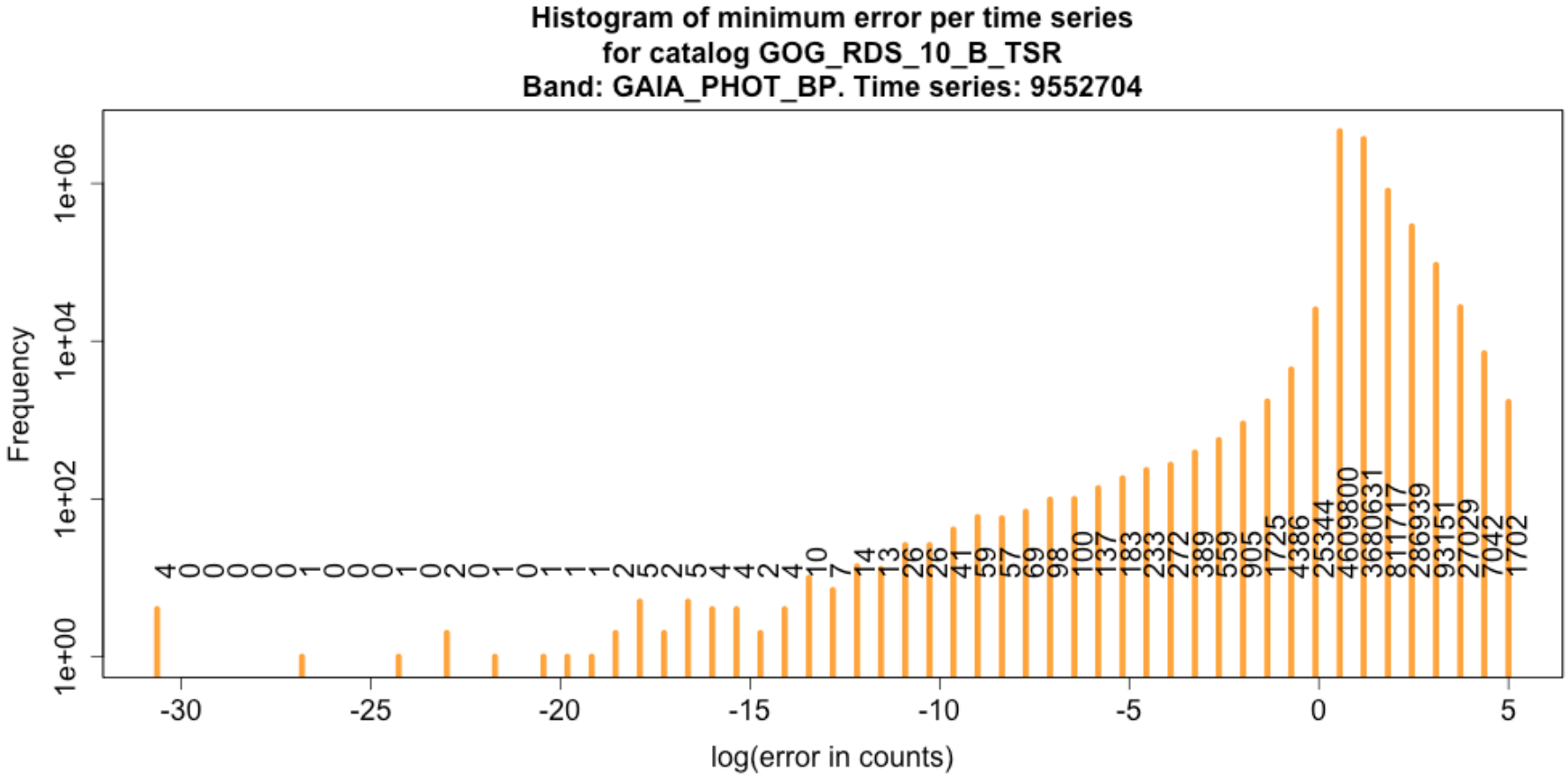
Analytics via SQL



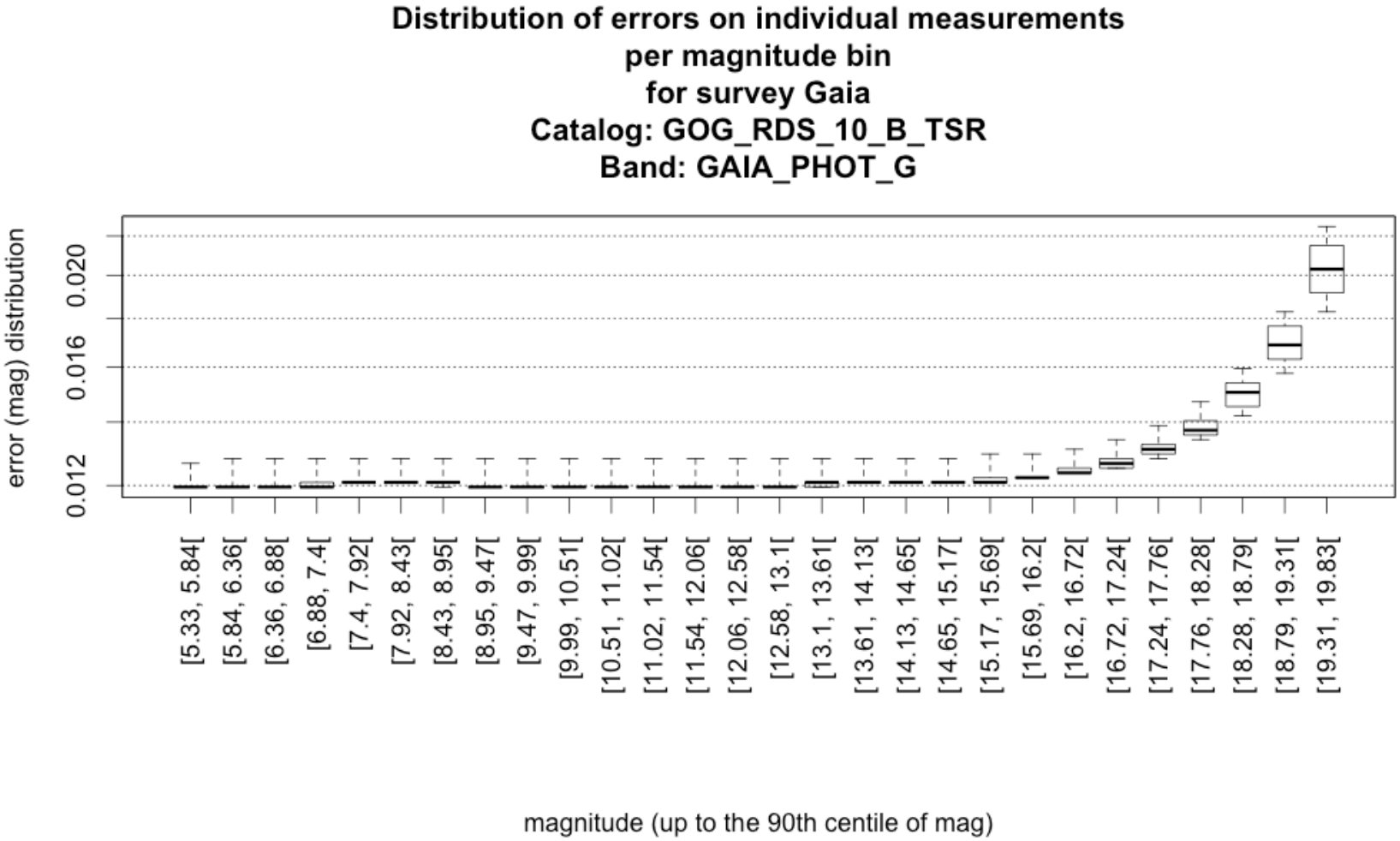
Analytics via SQL



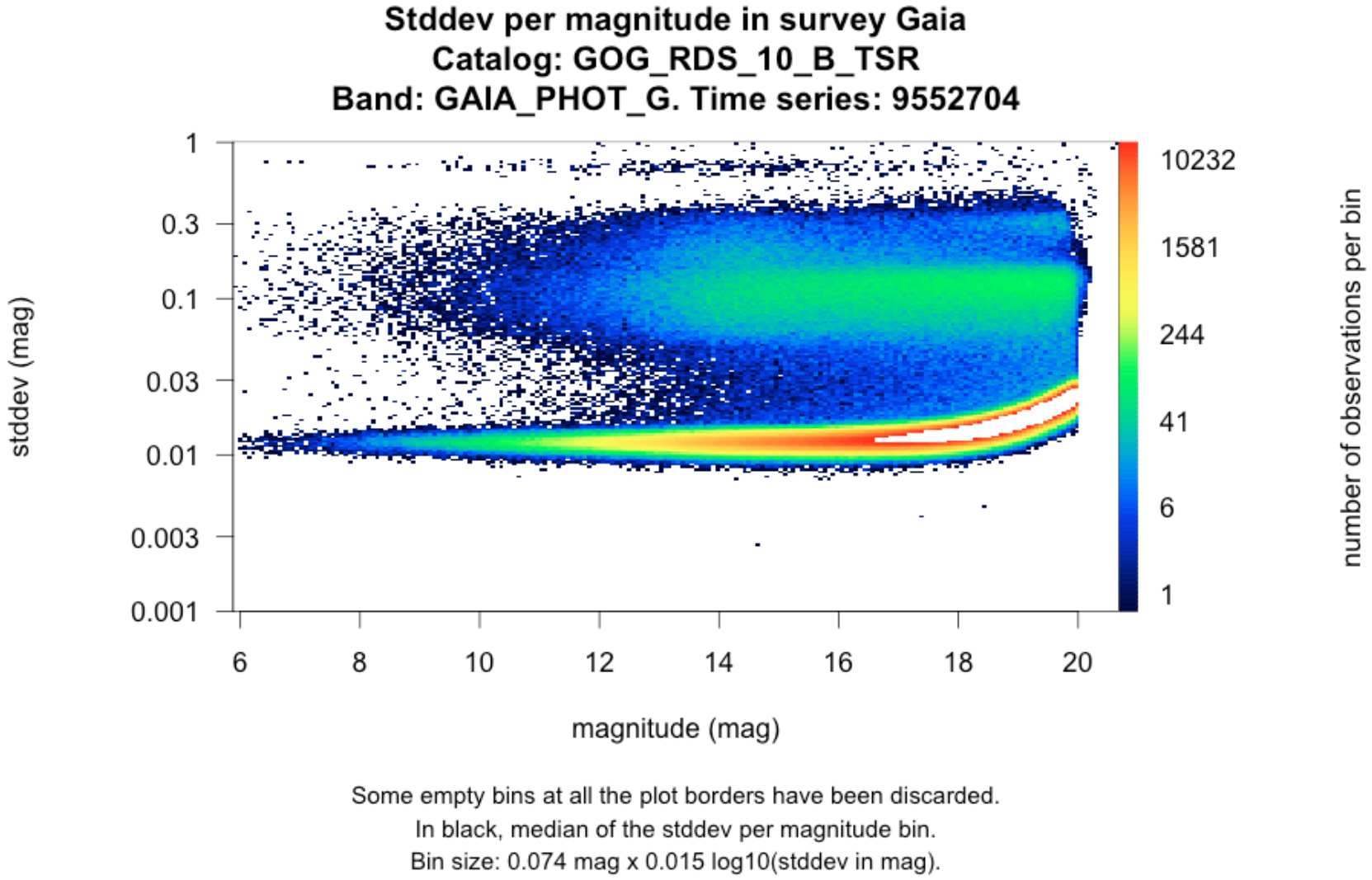
Analytics via SQL



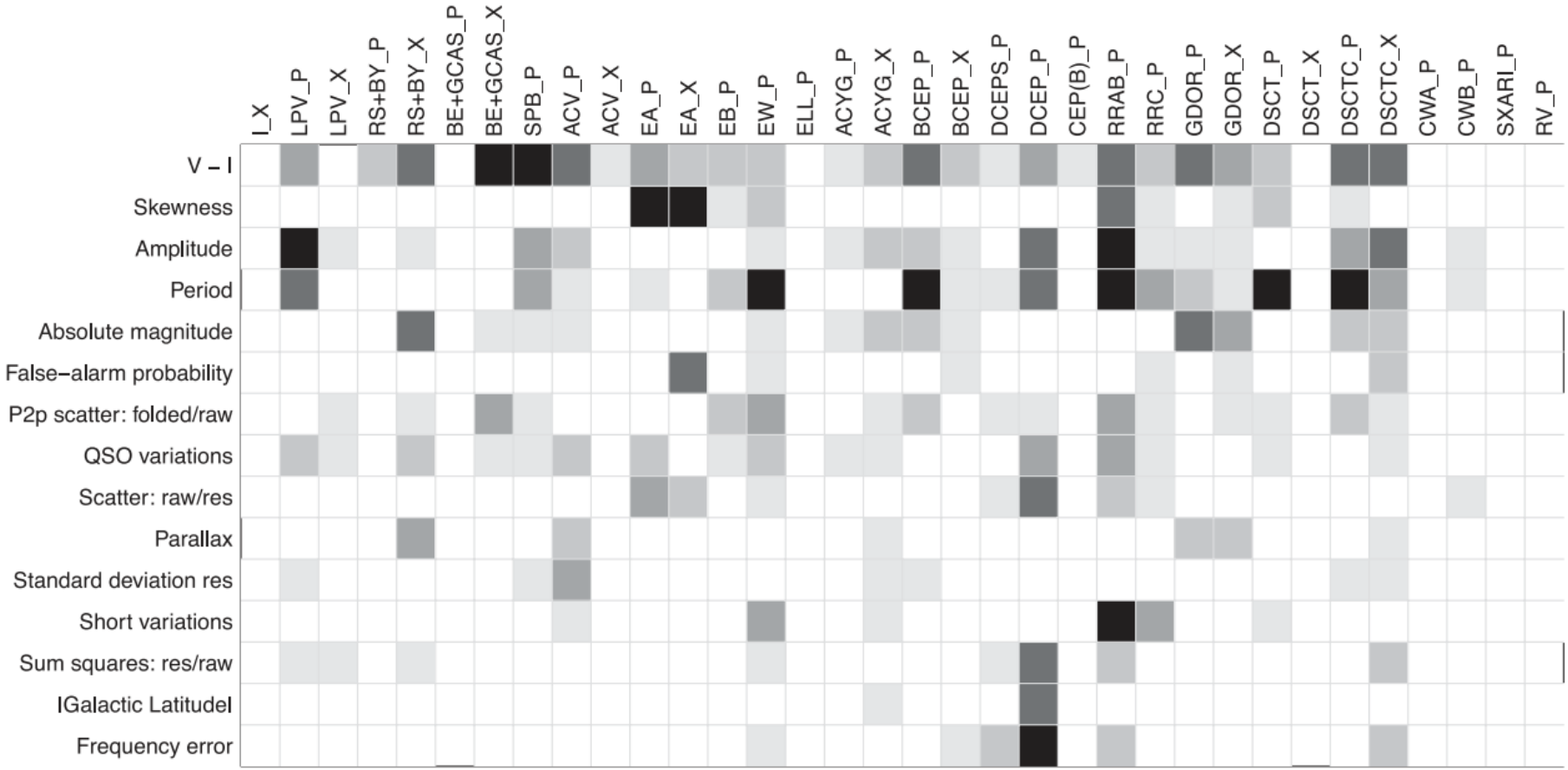
Analytics via SQL



Analytics via SQL

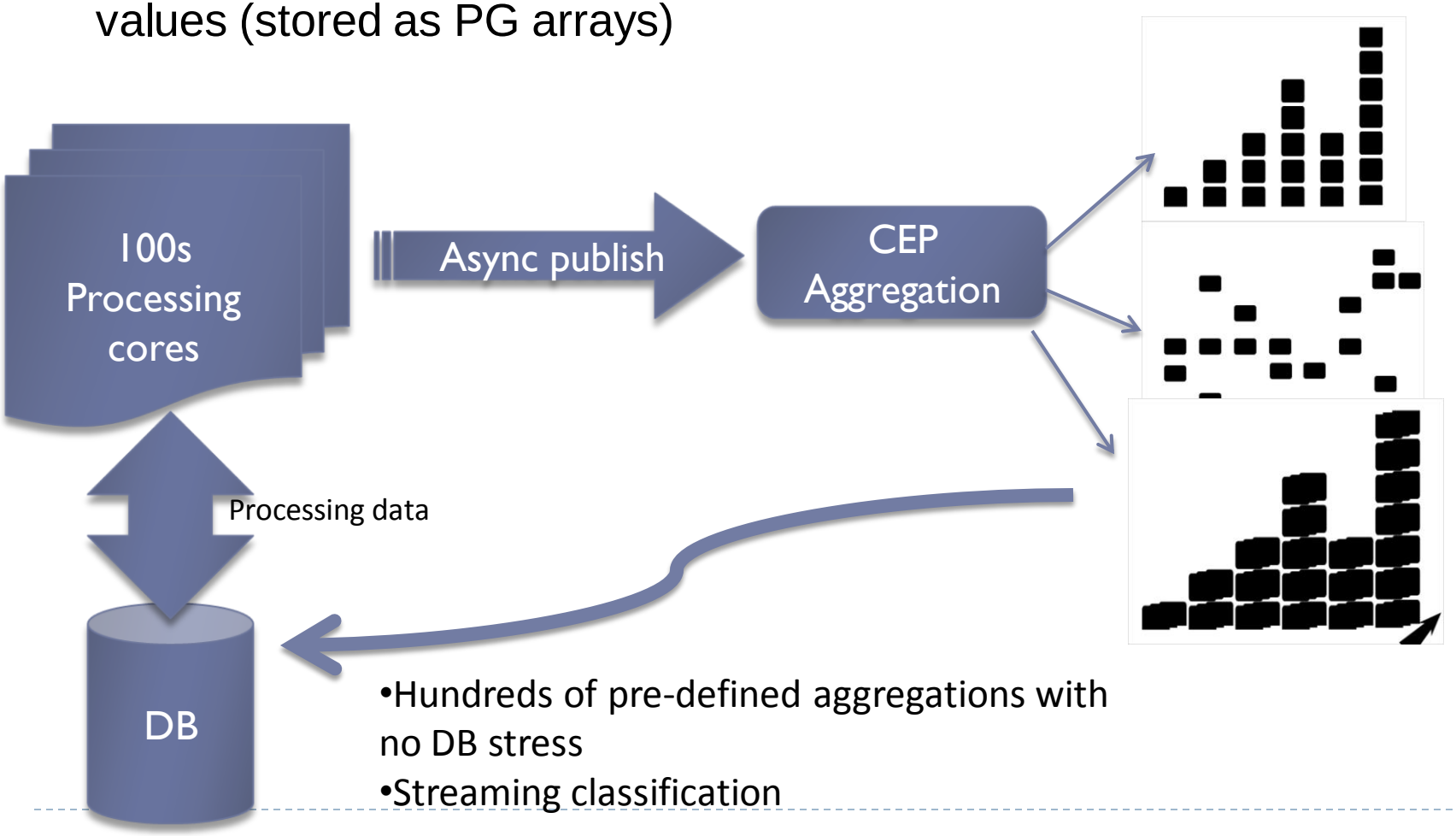


Attributes
(columnX, array[i], (pgType[j]).type2[i]…)



Complex Event Processing

- Stream of results is continuously aggregated externally via subpackage based on Camel/Esper and stored as scalar, 1D, 2D values (stored as PG arrays)



- Hundreds of pre-defined aggregations with no DB stress
- Streaming classification

Complex Event Processing
EQL with Apache Camel

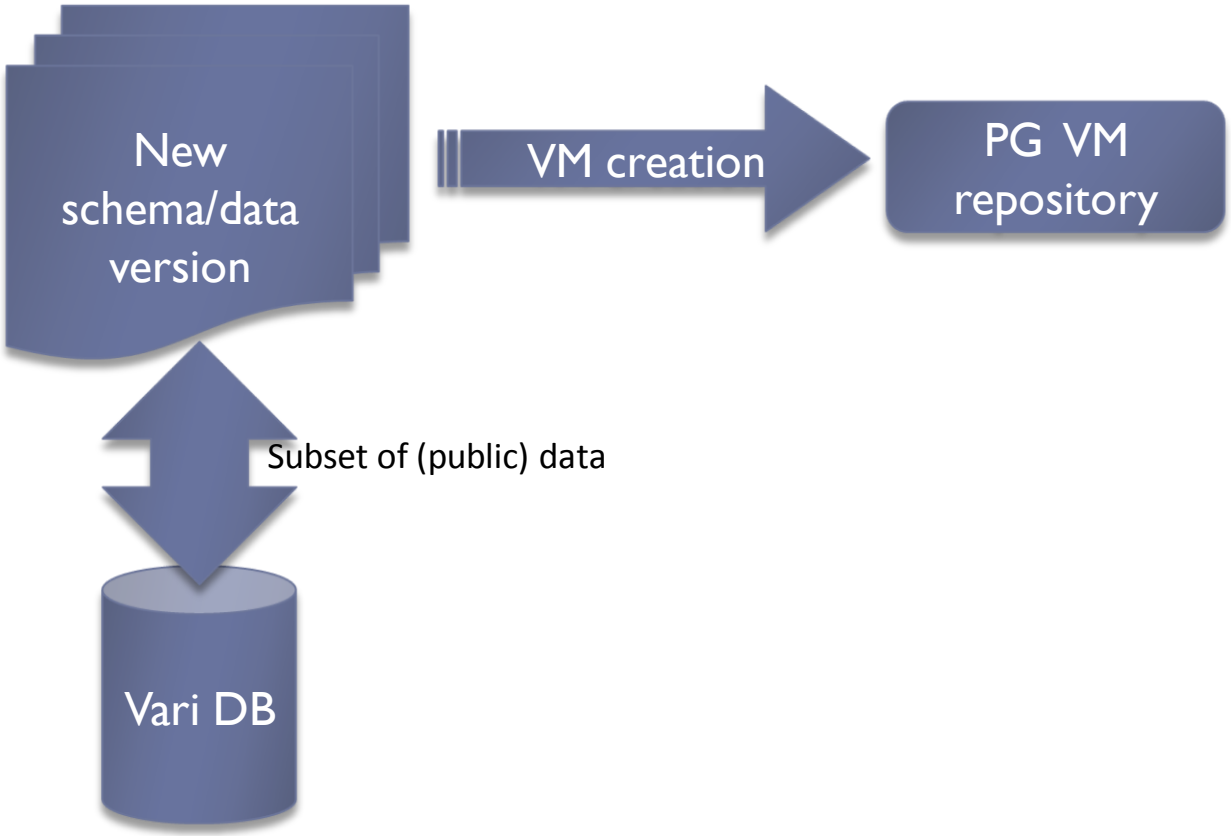
```
from("esper://sourceresult?eql=insert into PeriodSearchResultList select run.runId as runid,  " +
        "hist('skewness', -5.0, 5.0, 0.5, "
        + "((timeSeriesResult('HIPPARCOS')).getStatisticalParameters()).getSkewness()) as histogram " +
        "from SourceResult.win:time_batch(10 sec) " +
        "where ((timeSeriesResult('HIPPARCOS')).getStatisticalParameters()).getSkewness() > 0 "+
        "group by run.runId")
        .process(new Processor() {
            @SuppressWarnings("unchecked")
            public void process(Exchange exchange) throws Exception {
                com.espertech.esper.event.map.MapEventBean ev = (com.espertech.esper.event.map.MapEventBean) exchange
                        .getIn().getBody();
                Map map = (Map) ev.getUnderlying();
                exchange.getOut().setHeaders(exchange.getIn().getHeaders());
                exchange.getOut().setHeader("runid", map.get("runid"));
                exchange.getOut().setBody(map.get("histogram"));
            }
        }).log("${body}").to("activemq:varisystem.monitoring.histogram");
```

Complex Event Processing

real time monitoring

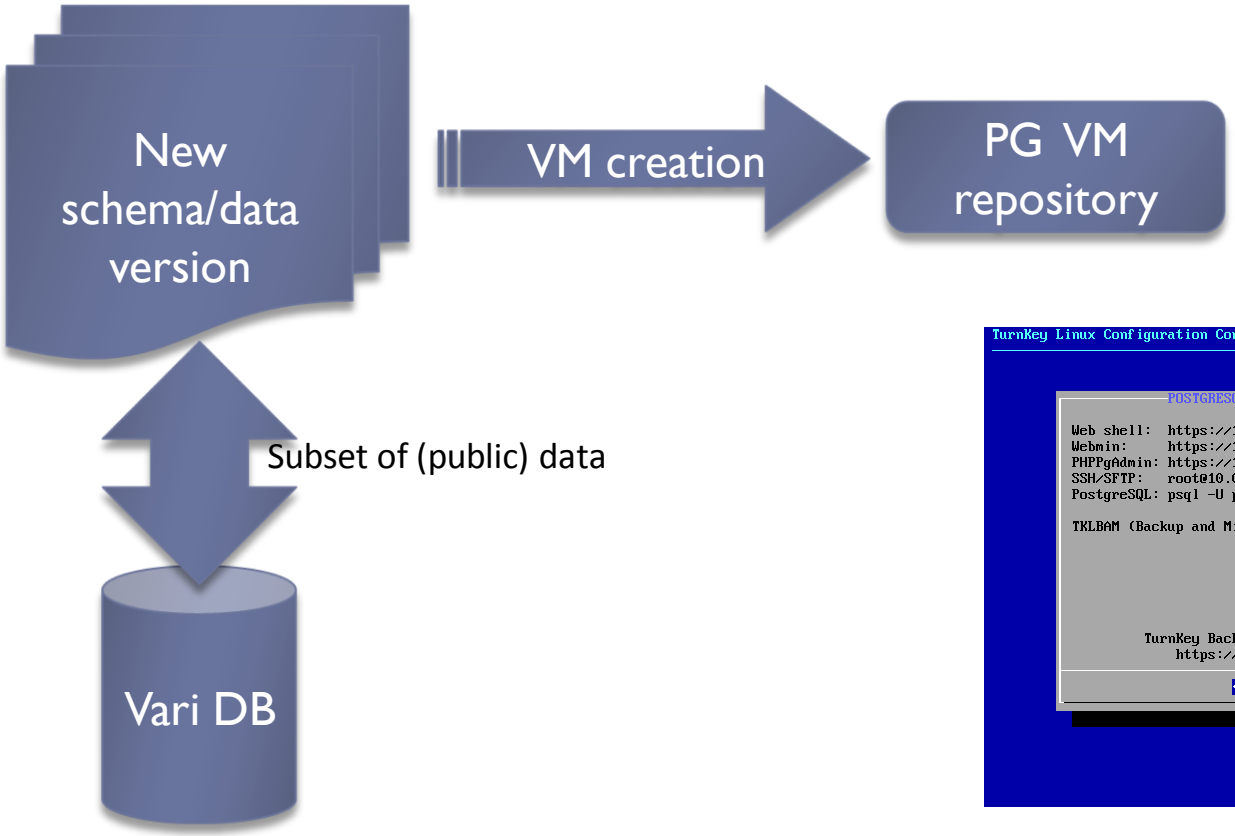
Data distribution for development

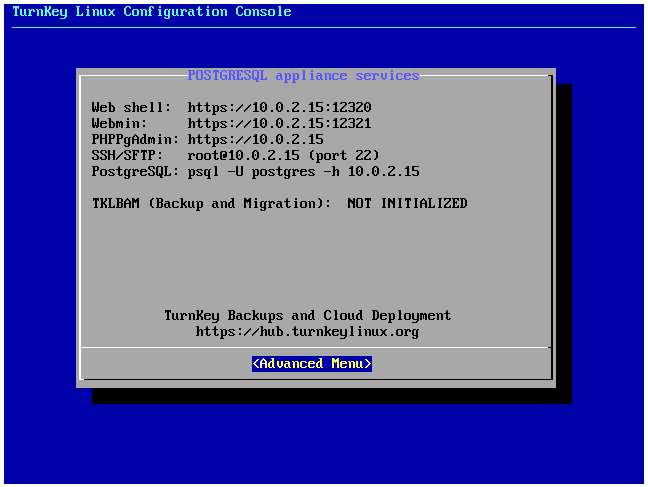
- Development, public data is distributed via PG Appliances
  - 5-15GB
  - ~0.5-5M sources



Data distribution for development

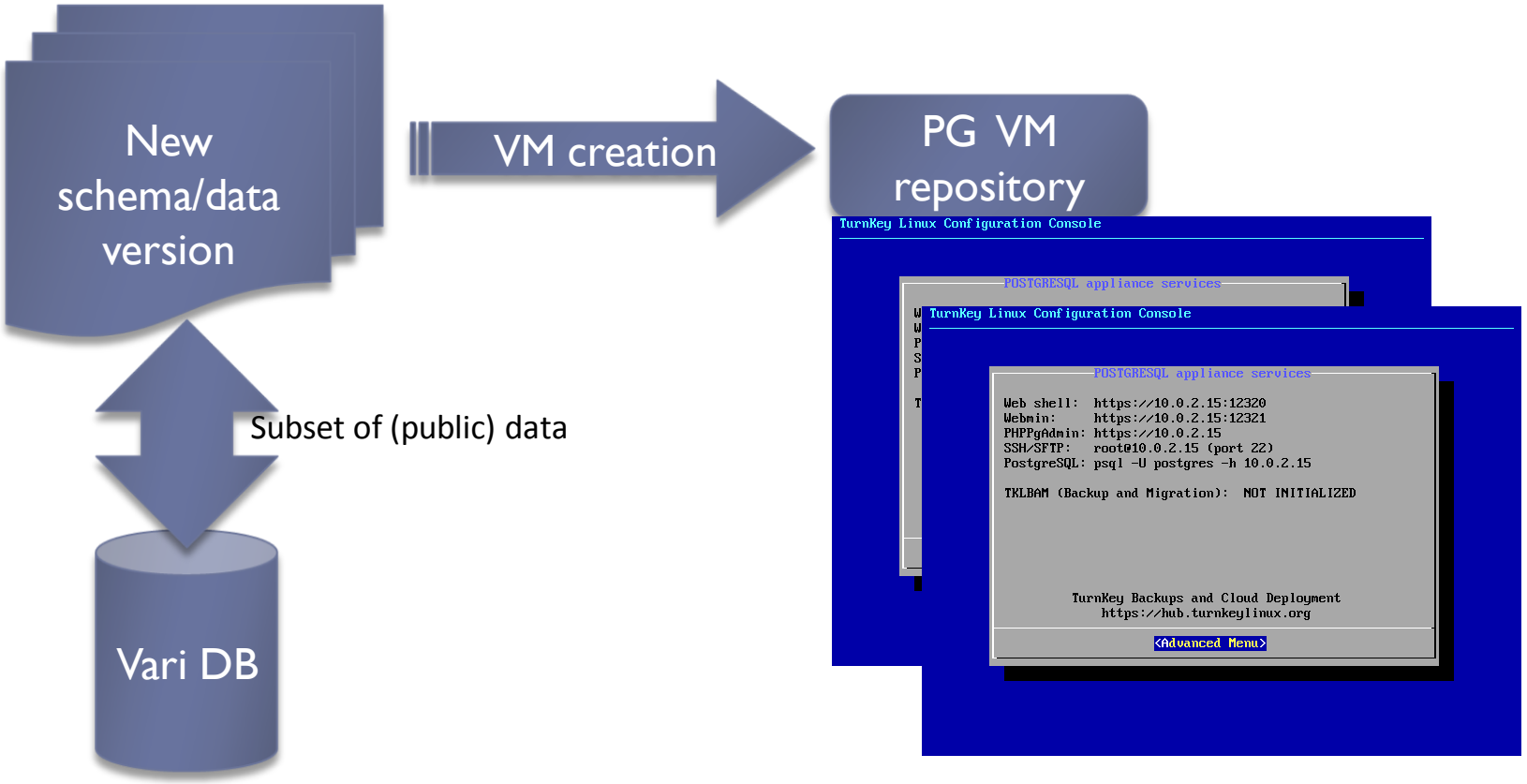
- Development, public data is distributed via PG Appliances
  - 5-15GB
  - ~0.5-5M sources





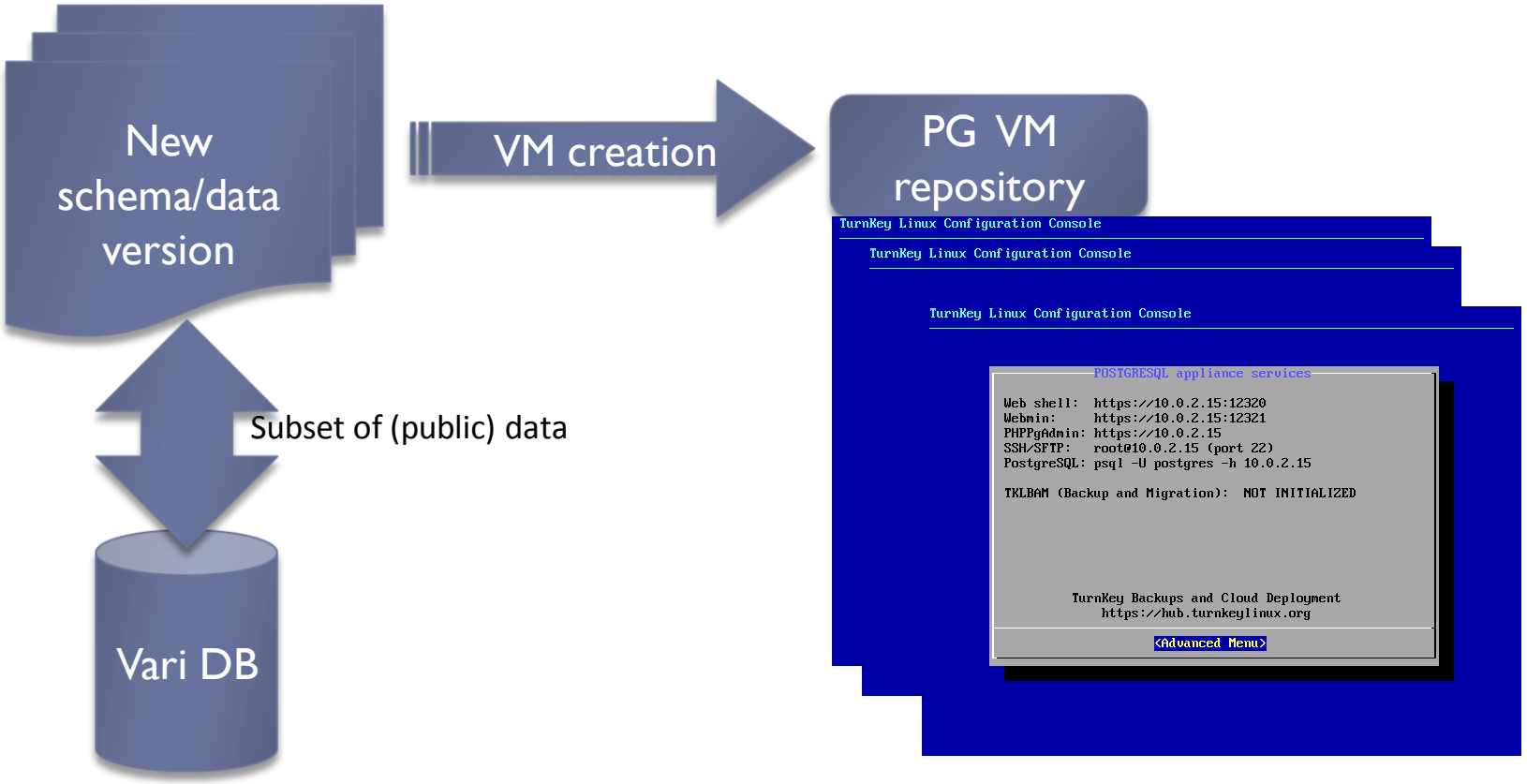
Data distribution for development

- Development, public data is distributed via PG Appliances
  - 5-15GB
  - ~0.5-5M sources



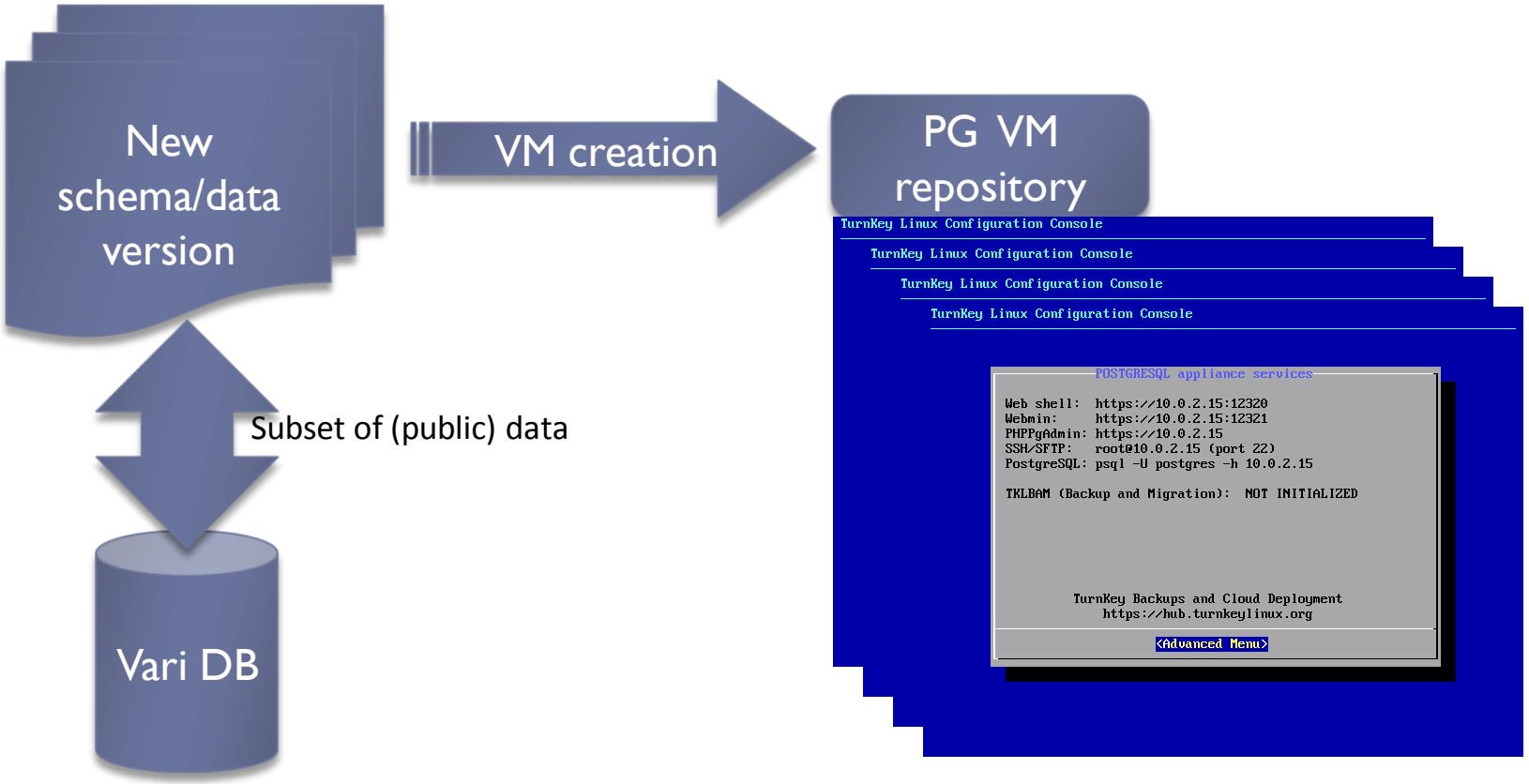
Data distribution for development

- Development, public data is distributed via PG Appliances
  - 5-15GB
  - ~0.5-5M sources



Data distribution for development

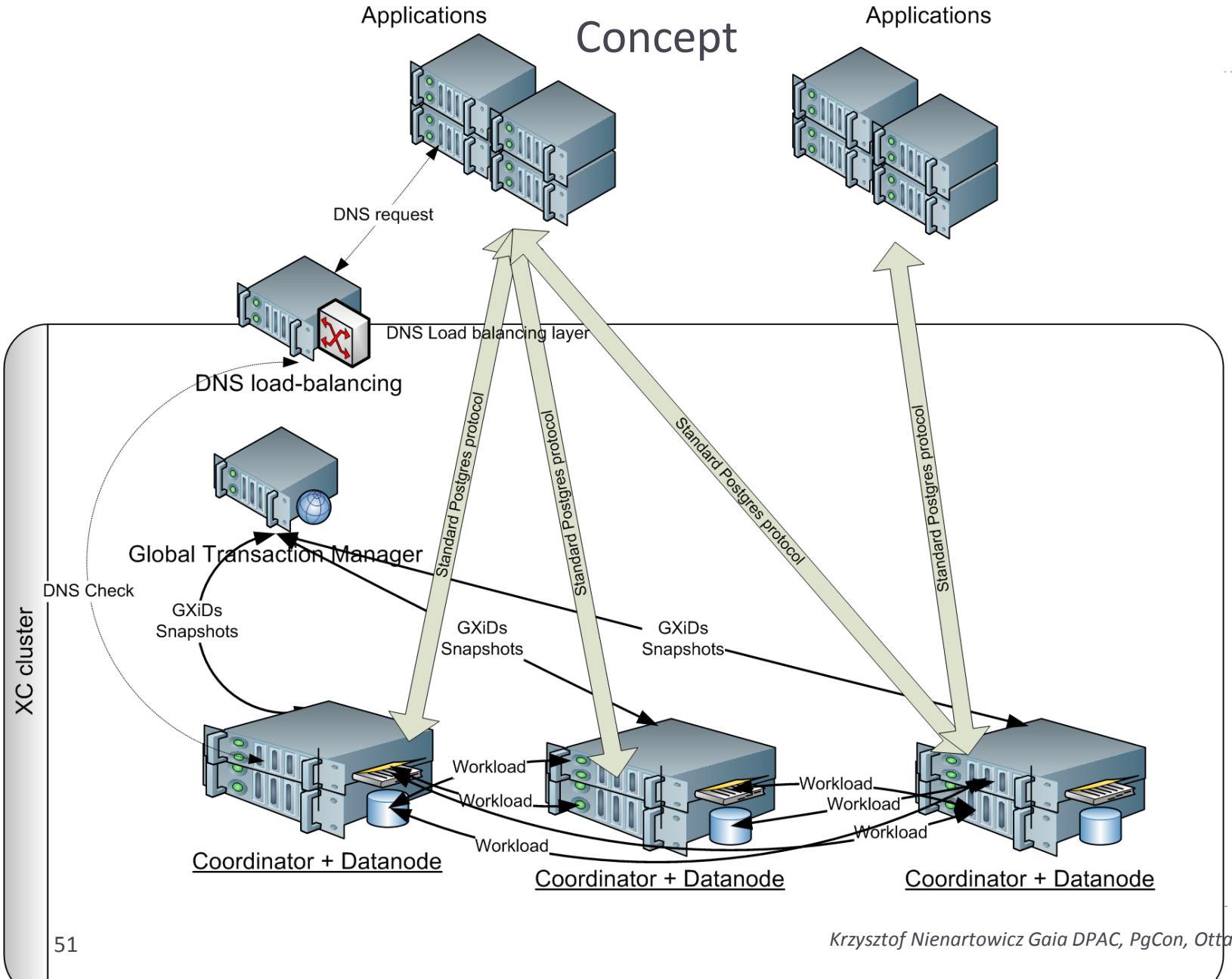
- Development, public data is distributed via PG Appliances
  - 5-15GB
  - ~0.5-5M sources



Structure

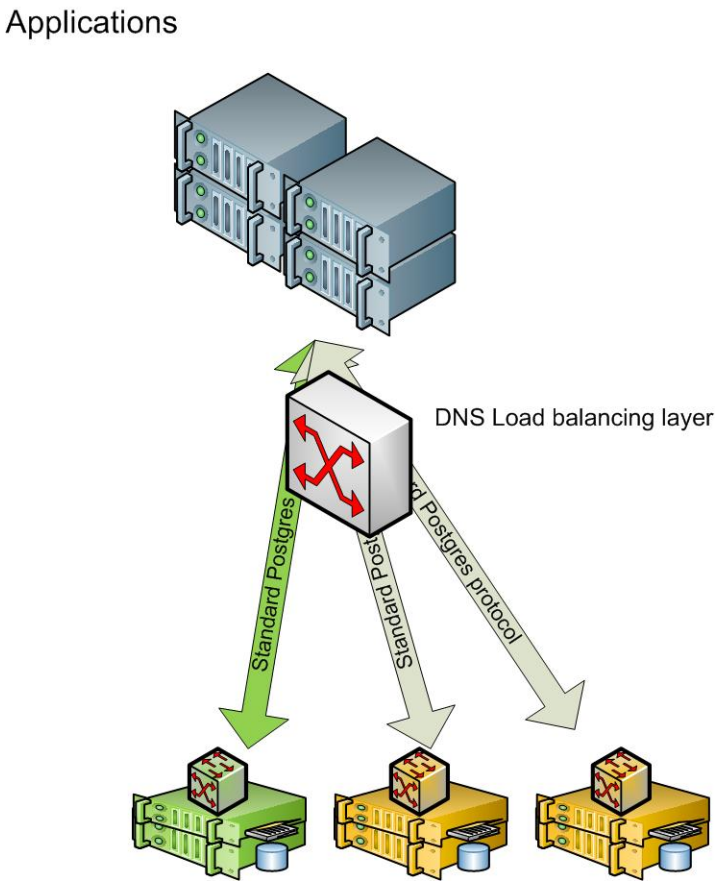
- My story
- Gaia mission
- Gaia science at CU7, Geneva
- Processing model
- Data model
- **XC role**
- Hardware
- Collaboration

Data platform: Postgres – XC/XL



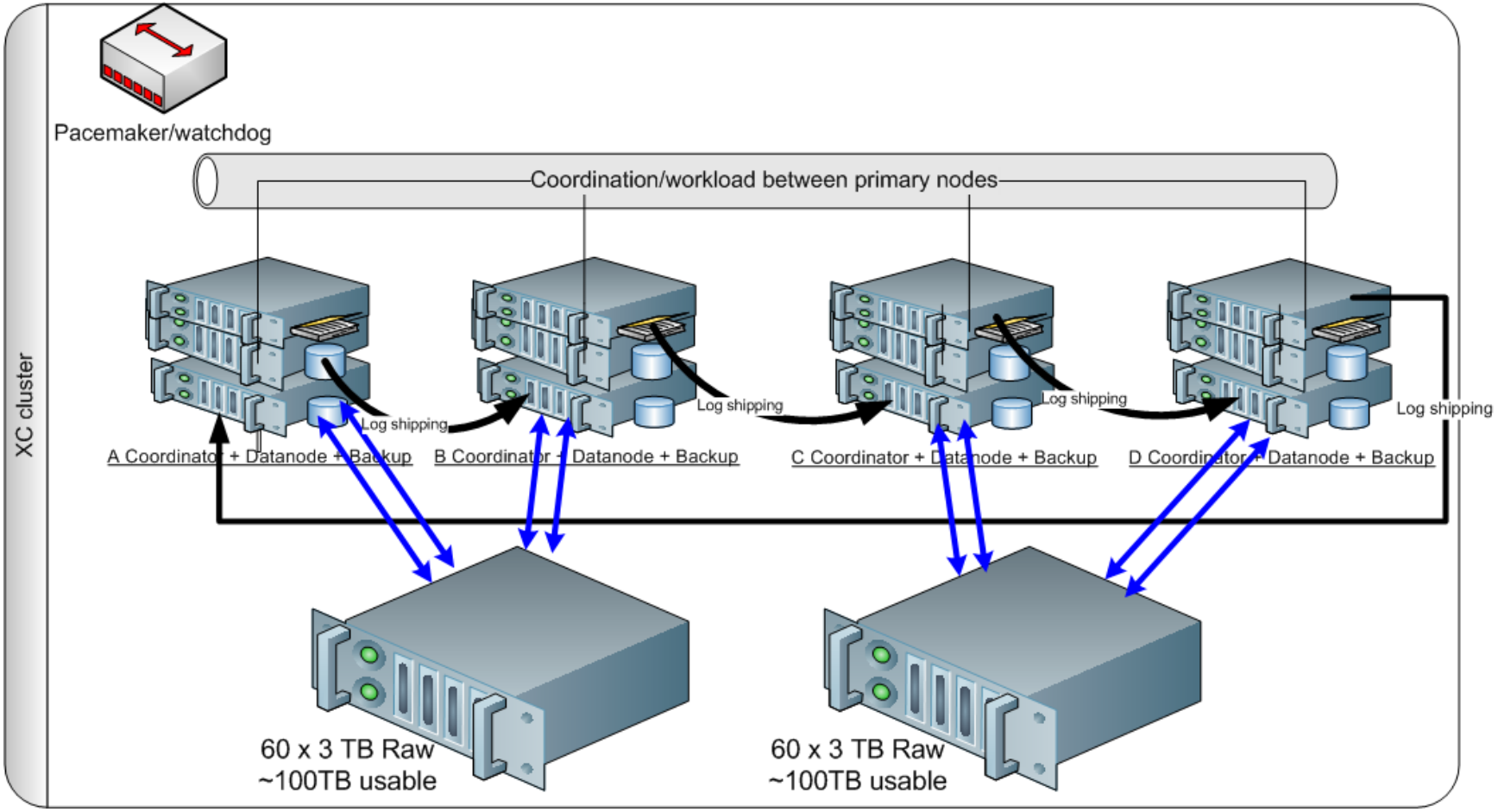
Postgres XC/XL

Load balancing – Jim Mlogdenski



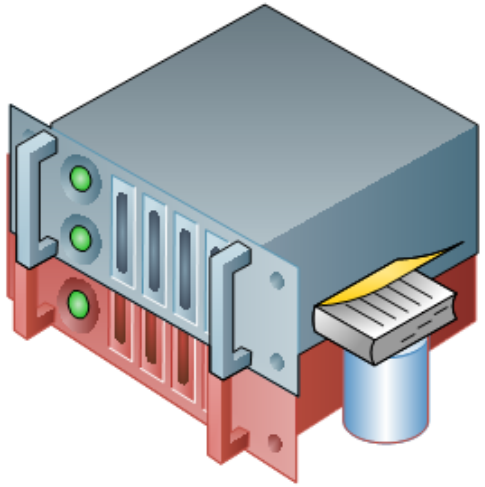
Postgres XC/XL (Koichi)

RAID1E



Postgres XC/XL

DB node



Coordinator + Datanode+Backup combined

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

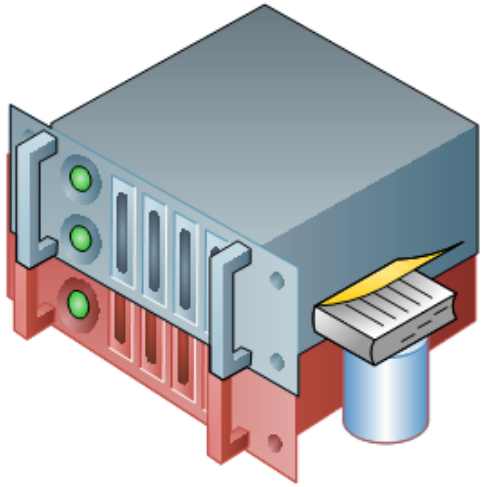
Storage for backup operations
WAL – Write Ahead Log: min 2x15K 300GB, RAID0(1 optional), quasi-sequential
Main Data Partition: 15-20TB usable per box, RAID6, quasi sequential + sequential
Data+seldomly used index
~1GB/s sequential
> 500 real IOPS
Separate IO channel for backup and primary.
Needs Min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

DB node



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

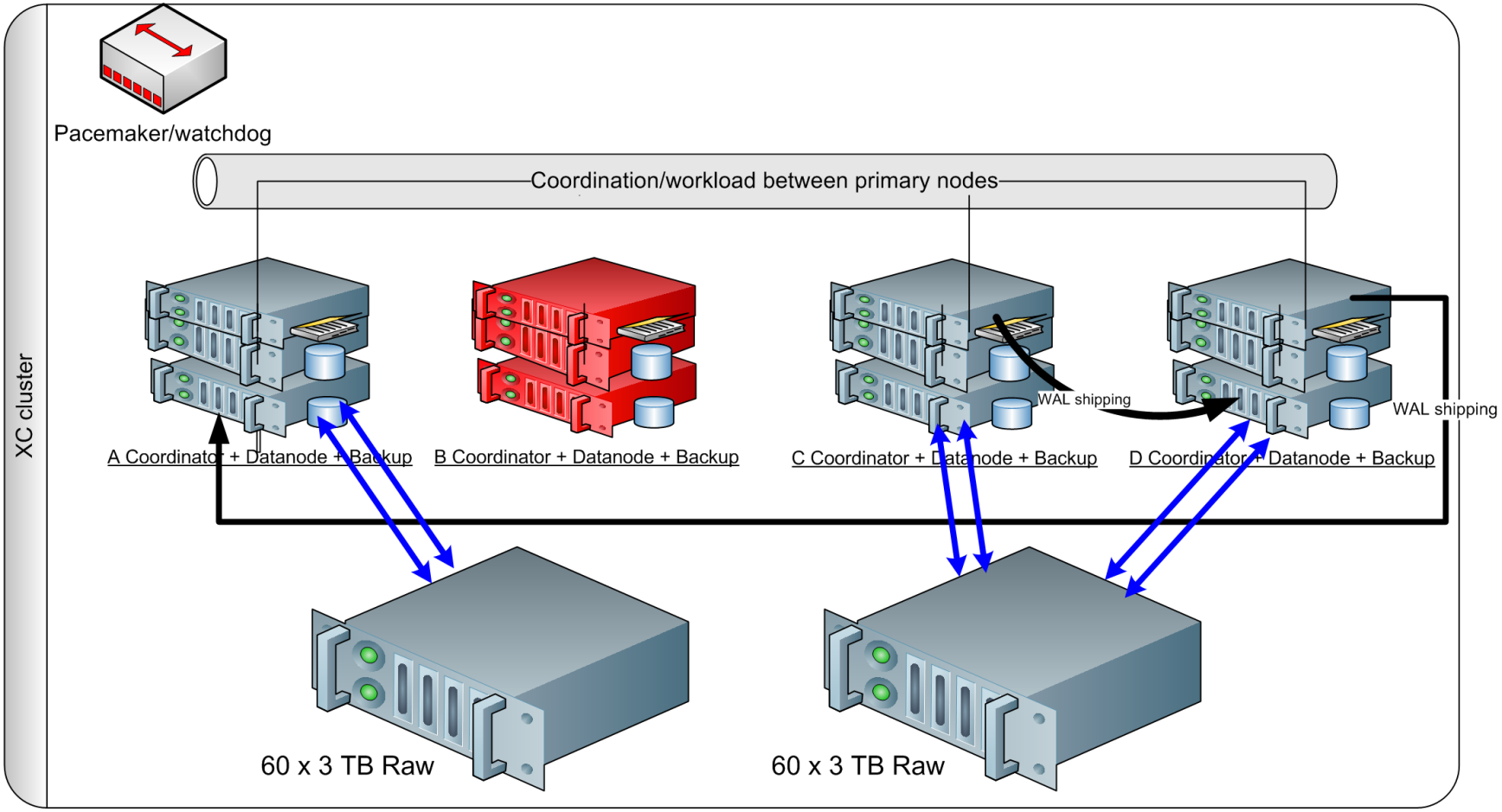
Storage for backup operations
WAL – Write Ahead Log: min 2x15K 300GB, RAID0(1 optional), quasi-sequential
Main Data Partition: 15-20TB usable per box, RAID6, quasi sequential + sequential
Data+seldomly used index
~1GB/s sequential
> 500 real IOPS
Separate IO channel for backup and primary.
Needs Min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

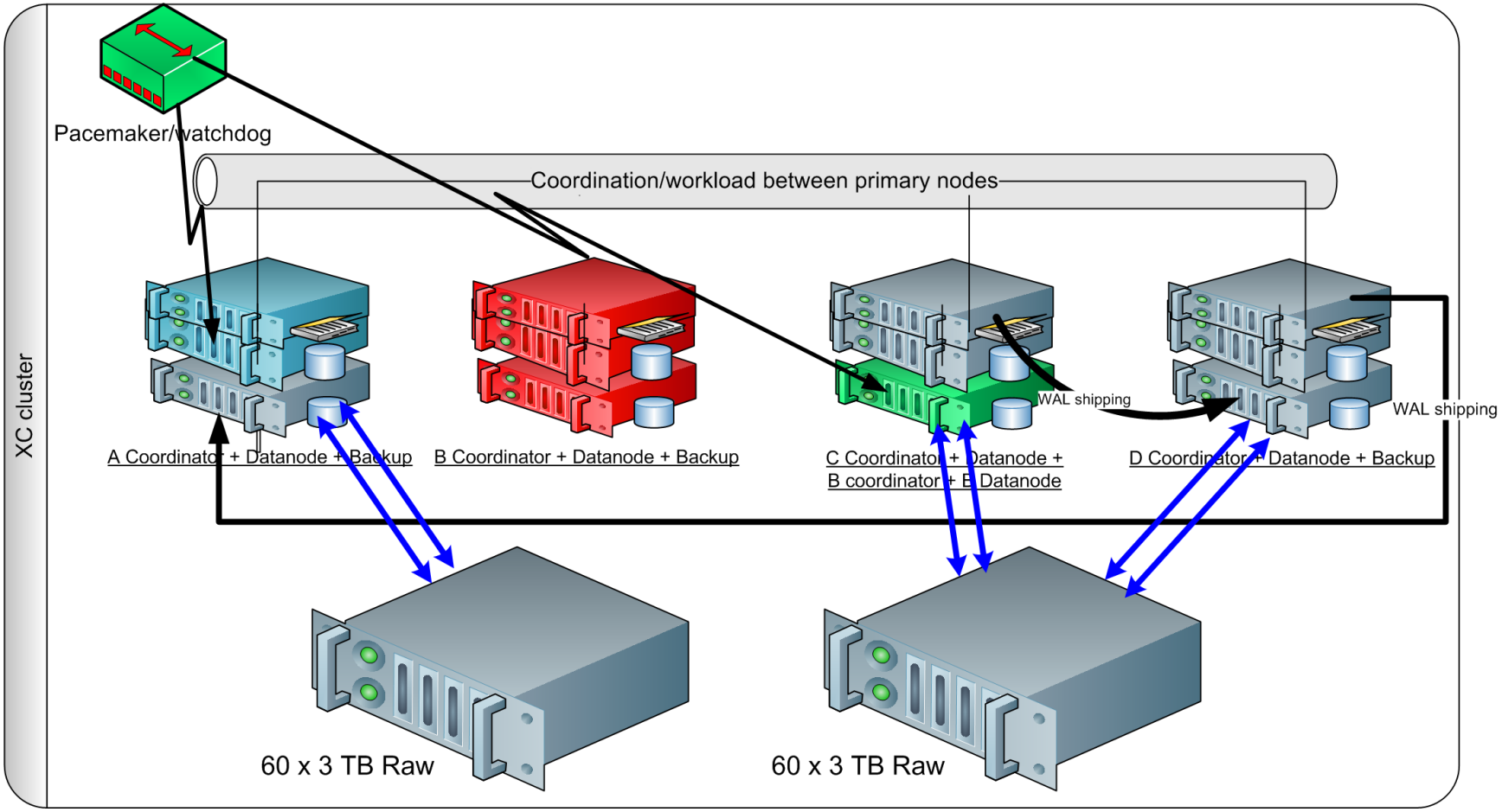
Postgres XC/XL

High availability - Failure



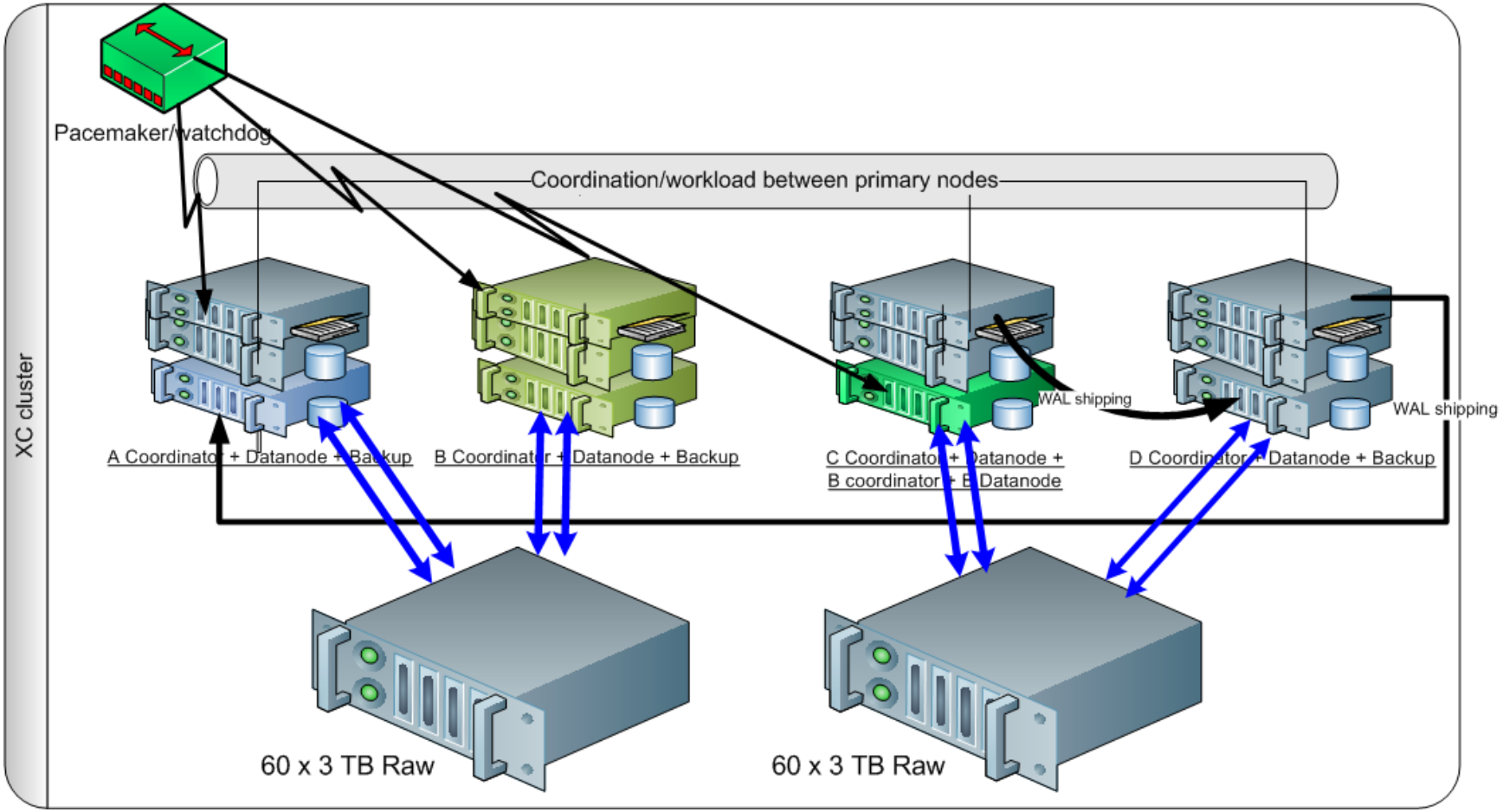
Postgres XC/XL

HA – Backup promotion



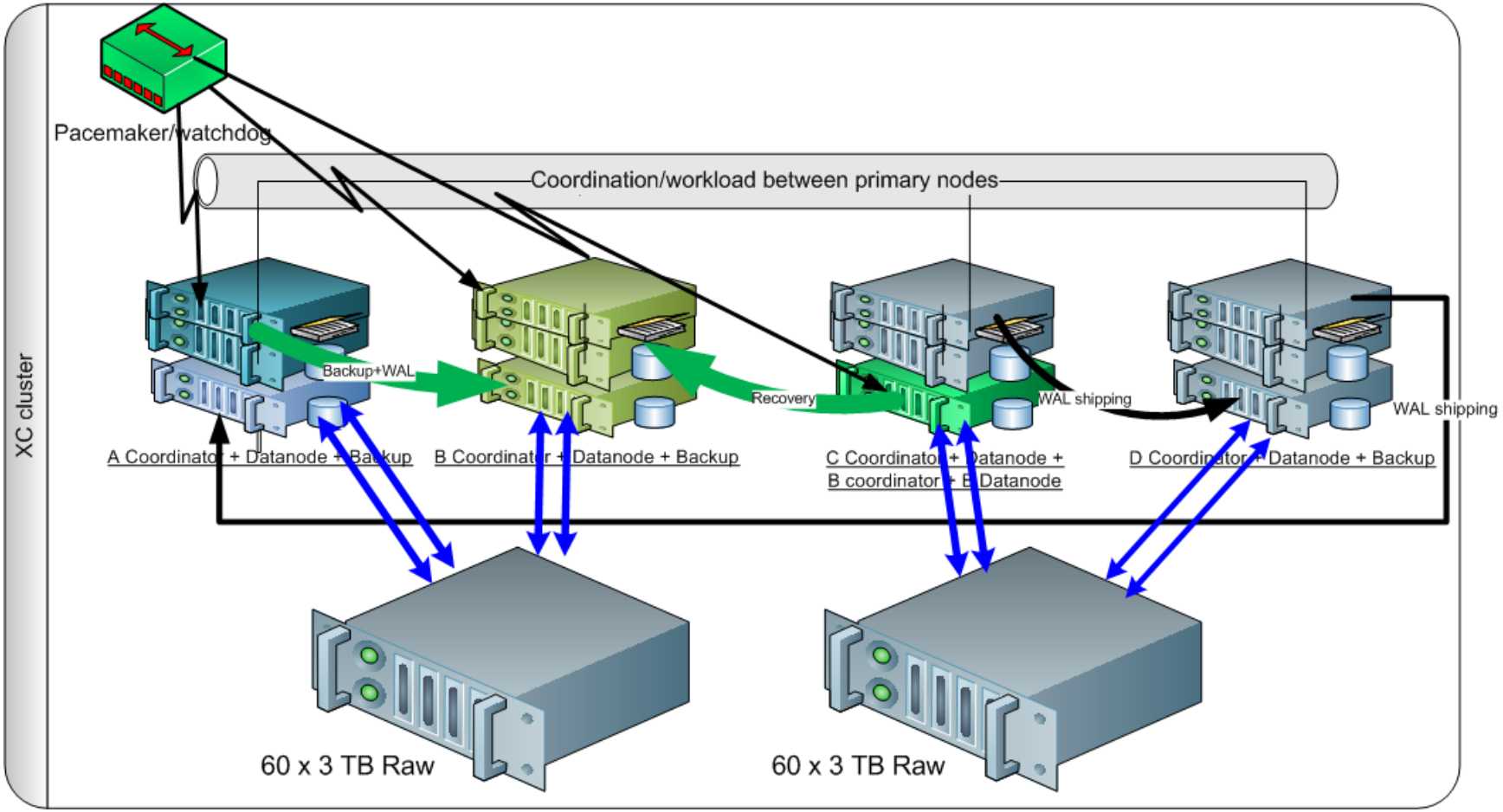
Postgres XC/XL

HA: recovery and backup demotion



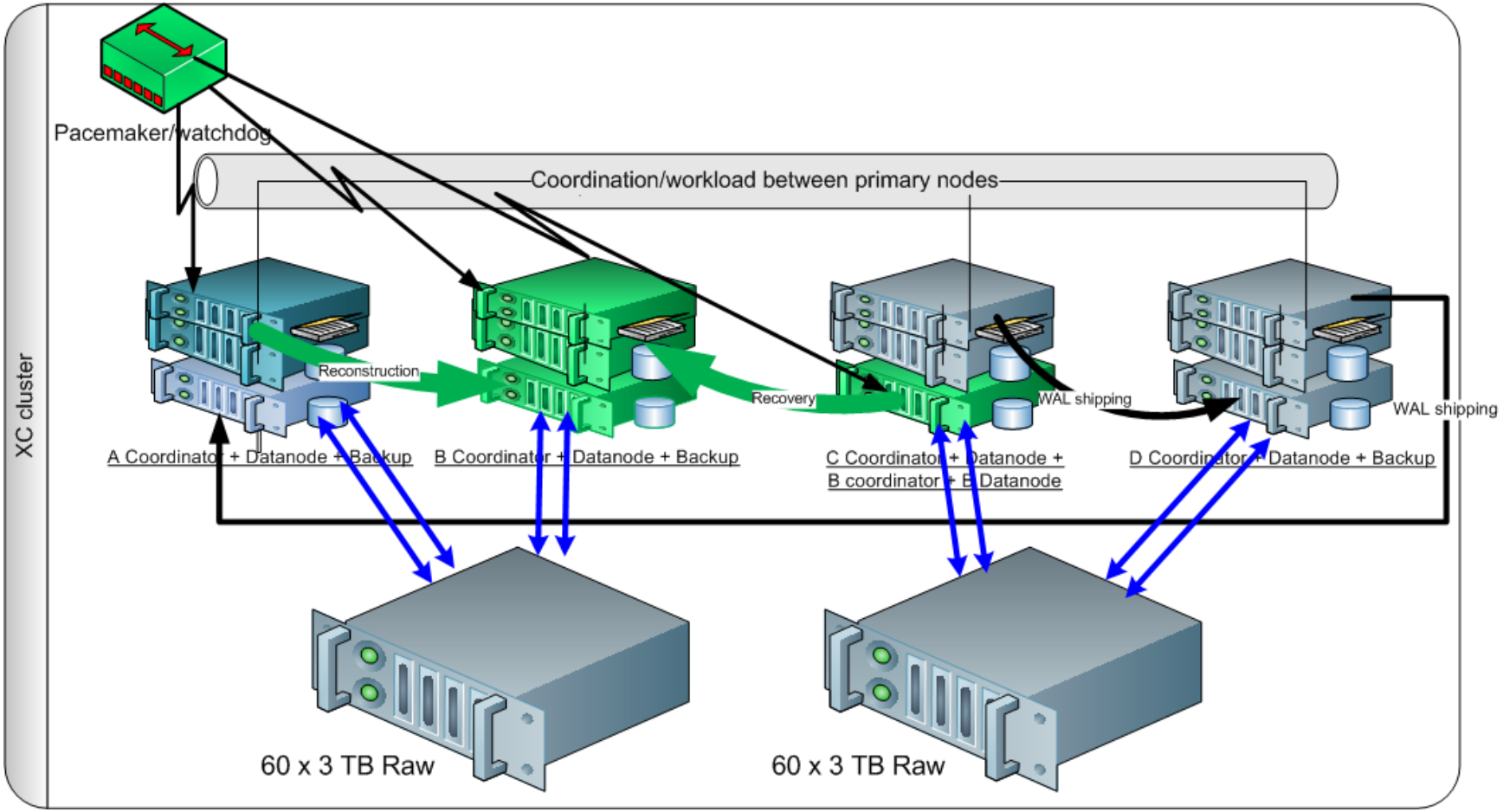
Postgres XC/XL

HA: recovery and backup demotion



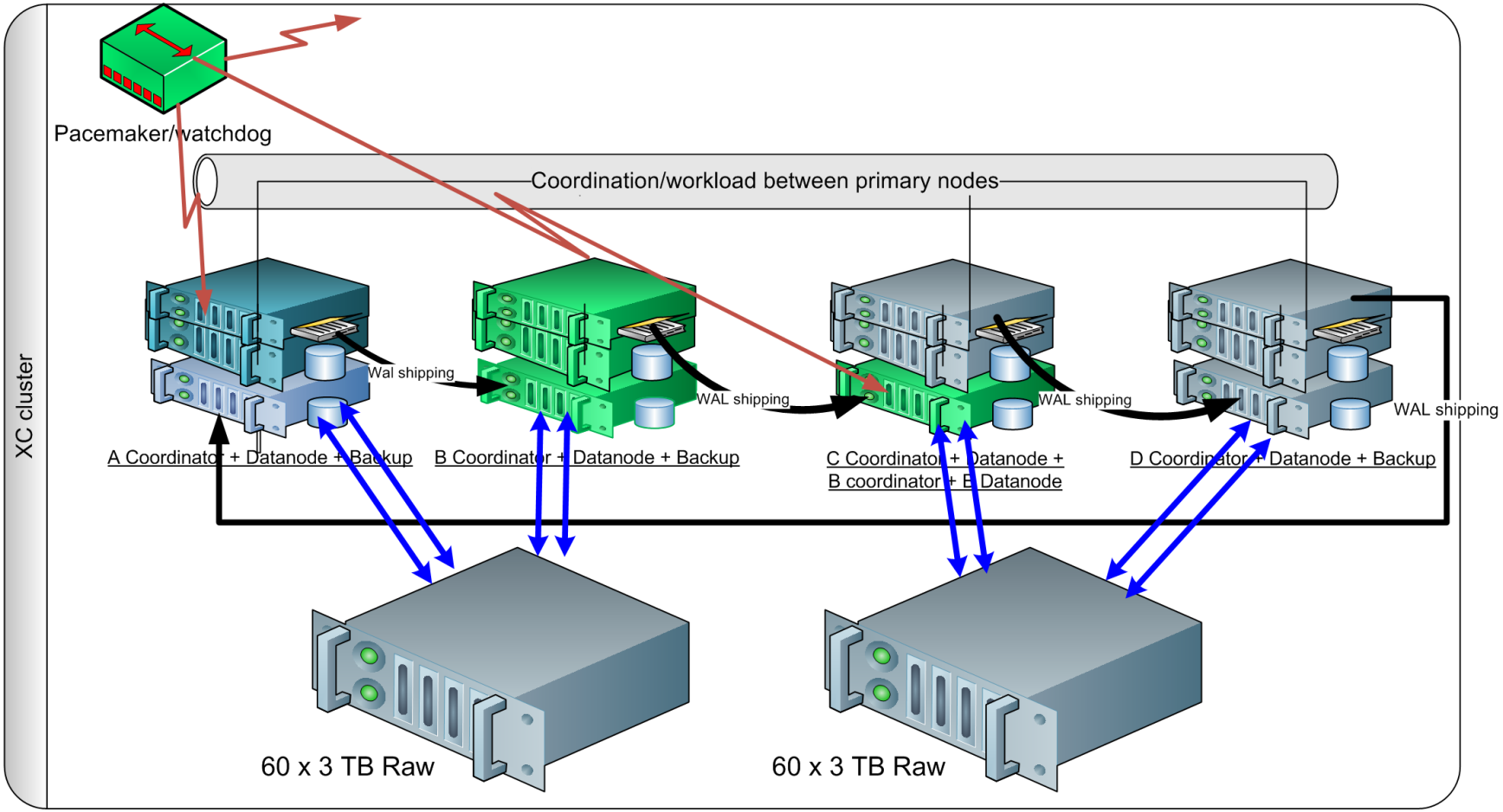
Postgres XC/XL

HA: recovery and backup demotion



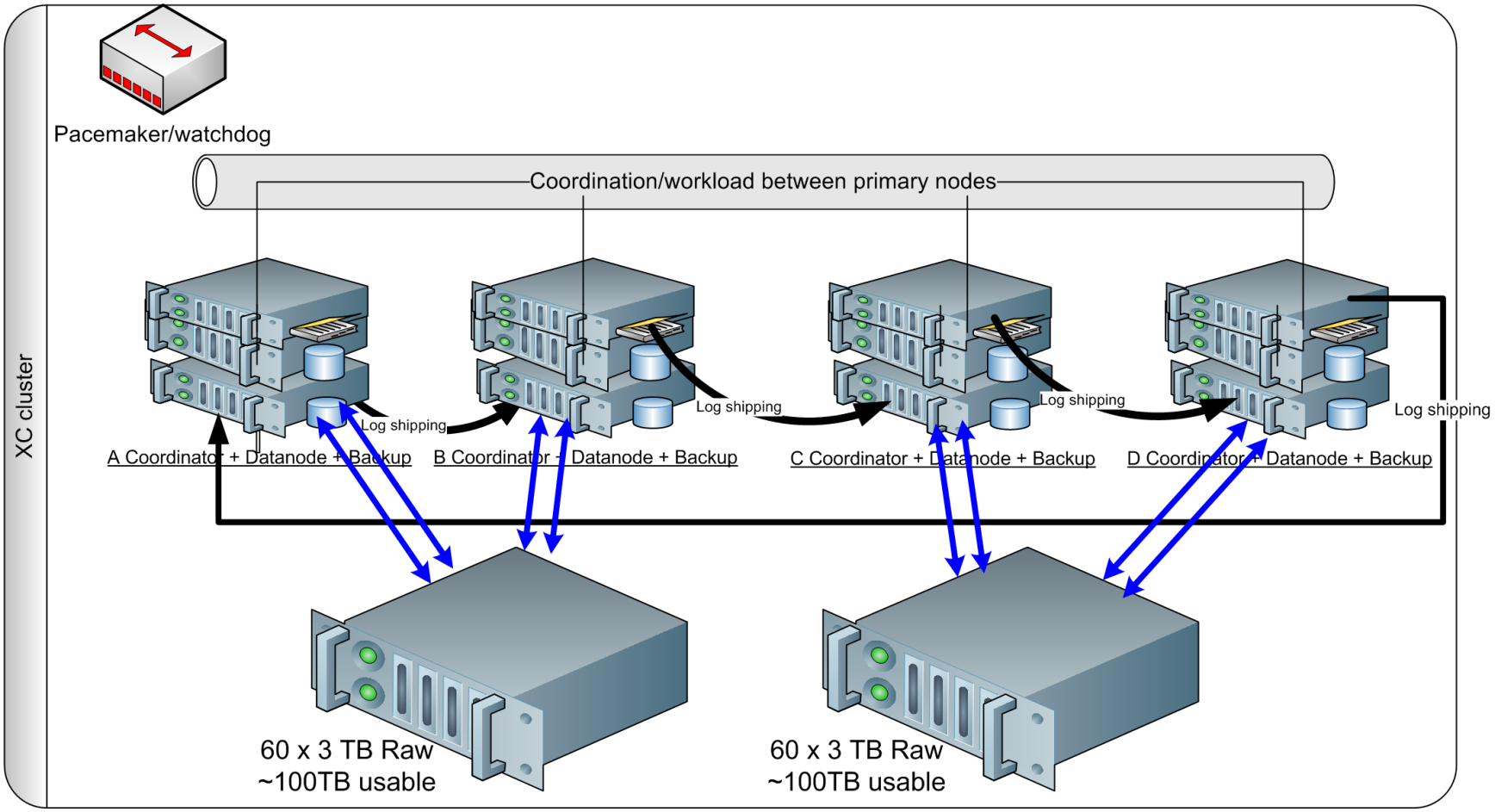
Postgres XC/XL

HA: recovery and backup demotion



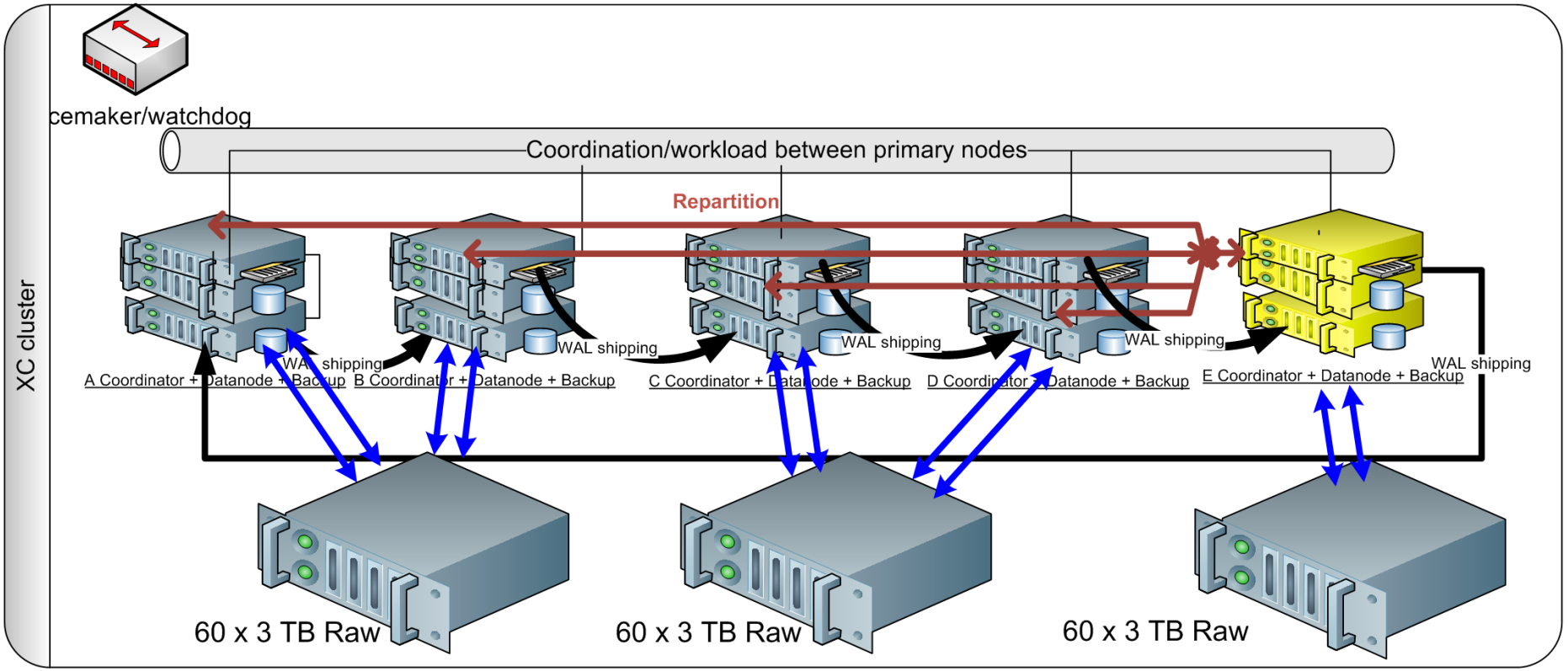
Postgres XC/XL

HA: recovery and backup demotion



Postgres XC/XL

Scalability: expansion, repartitioning



Postgres XC/XL role

- Postgres-XC integration seamlessness
  - Postgres compatible
    - PG SQL (psql, pgAdmin, JPA, hstore, object types, spatial indicies, n-dimensional knn-searches)
    - R: RPostgreSQL
  - CU7 Framework compatible
    - OpenJPA
    - Java (JDBC, PG copy protocol)
  - Scalability
    - Horizontal + veritcal - 50TB …. ~ 1PB, 10^{12} entities
    - Timeseries reconstruction
    - Batch processing, source oriented
    - Analytical processing, ad hoc
  - High Availability

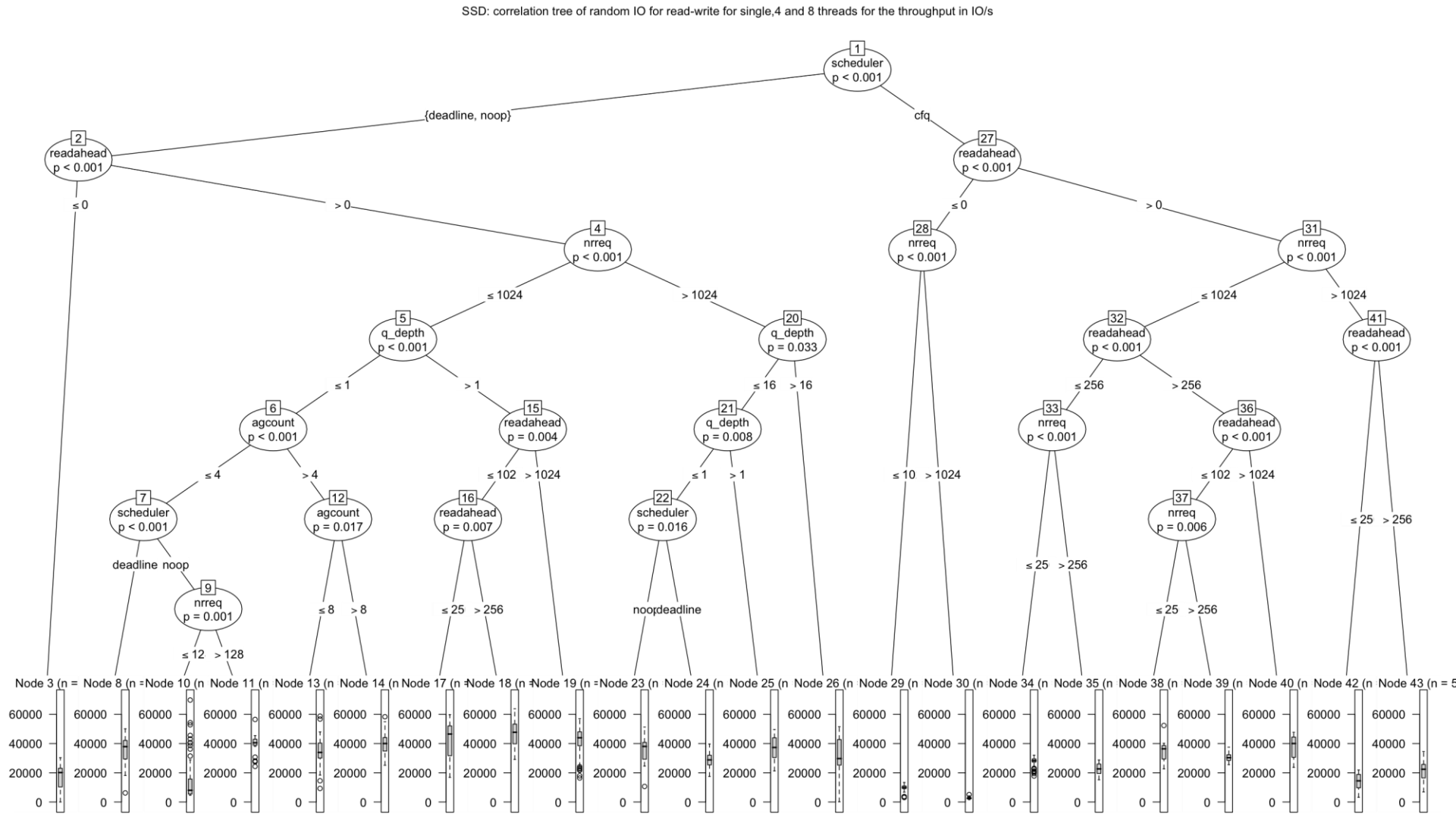
Postgres XC/XL role – what do we need

- Very High priority
  - Stability
  - Performance
  - Partitioning
  - PG type model (objects, arrays of objects)
  - JDBC working (pooling)
  - R working
  - Parallel loading into **partitions** of XC distributed tables
    - Support for COPY (via JDBC or other Java driver)
    - UDF for custom distribution logic (or as external API)
  - Custom aggregates
  - Load balancing
  - Support for existing index extensions
  - PLJava, plR working
- High priority
  - Query estimates available on the coordinator
  - OLAP features:
    - In-node query parallelization
    - Index organized tables
    - Working Bitmap, Bloom indicies
    - Column store
    - Estimators
    - Effective knn-searches
    - Compression
    - Read-only tables(paces)

Structure

- My story
- Gaia mission
- Gaia science at CU7, Geneva
- Processing model
- Data model
- XC role
- **Hardware**
- Collaboration

OS IO tuning


SSD: correlation tree of random IO for read-write for single,4 and 8 threads for the throughput in IO/s
1 scheduler p < 0.001
{deadline, noop}
cfq
2 readahead p < 0.001
27 readahead p < 0.001
≤ 0
> 0
4 nrreq p < 0.001
28 nrreq p < 0.001
31 nrreq p < 0.001
≤ 1024
> 1024
5 q_depth p < 0.001
20 q_depth p = 0.033
32 readahead p < 0.001
41 readahead p < 0.001
≤ 1
> 1
≤ 16
> 16
6 agcount p < 0.001
15 readahead p = 0.004
21 q_depth p = 0.008
33 nrreq p < 0.001
36 readahead p < 0.001
≤ 4
> 4
≤ 102
> 1024
≤ 1
> 1
≤ 10
> 1024
≤ 256
> 256
7 scheduler p < 0.001
12 agcount p = 0.017
16 readahead p = 0.007
22 scheduler p = 0.016
37 nrreq p = 0.006
deadline
noop
9 nrreq p = 0.001
≤ 8
> 8
≤ 25
> 256
noop
deadline
≤ 12
> 128
Node 3 (n = 60000 40000 20000 0
Node 8 (n = 60000 40000 20000 0
Node 10 (n = 60000 40000 20000 0
Node 11 (n = 60000 40000 20000 0
Node 13 (n = 60000 40000 20000 0
Node 14 (n = 60000 40000 20000 0
Node 17 (n = 60000 40000 20000 0
Node 18 (n = 60000 40000 20000 0
Node 19 (n = 60000 40000 20000 0
Node 23 (n = 60000 40000 20000 0
Node 24 (n = 60000 40000 20000 0
Node 25 (n = 60000 40000 20000 0
Node 26 (n = 60000 40000 20000 0
Node 29 (n = 60000 40000 20000 0
Node 30 (n = 60000 40000 20000 0
Node 34 (n = 60000 40000 20000 0
Node 35 (n = 60000 40000 20000 0
Node 38 (n = 60000 40000 20000 0
Node 39 (n = 60000 40000 20000 0
Node 40 (n = 60000 40000 20000 0
Node 42 (n = 60000 40000 20000 0
Node 43 (n = 5 60000 40000 20000 0

Big Volume vs Constraints

Big Volume vs Constraints

- Growing number of cores

Big Volume vs Constraints

- Growing number of cores
- Growing number of users

Big Volume vs Constraints

- Growing number of cores

Procurement Plan

	Today	2013	2014	2018
Cores	O(100)	O(100)	O(300)	O(1000+)
DB Nodes	1	3+2	4 +1	8+2
Storage	24TB	50TB	75-315 TB	200-950 TB
		Launch	IDR(*1)	FDR(*2)

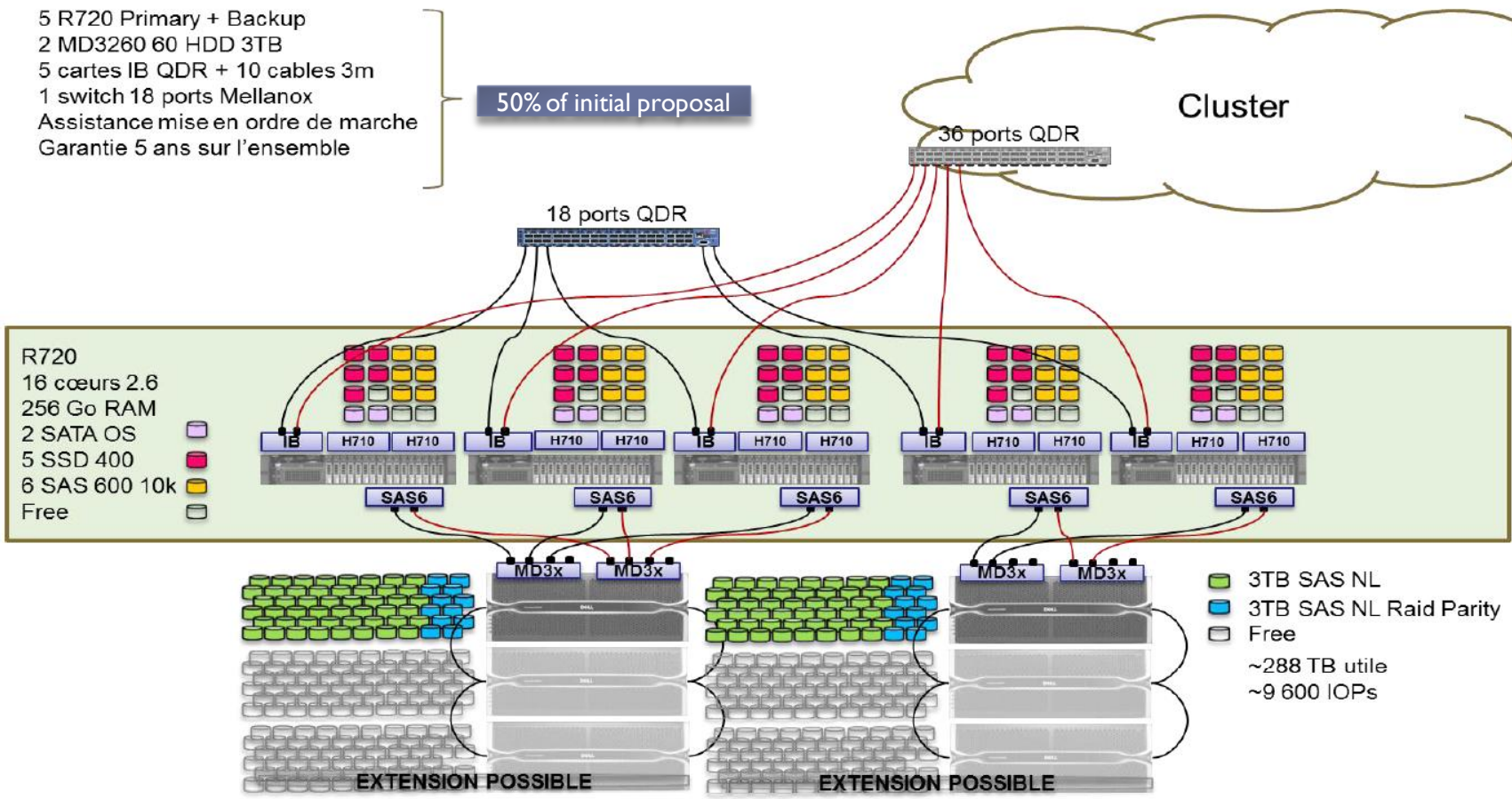
(*1) IDR: Intermediate **D**ata **R**eduction

(*2) FDR: Final **D**ata **R**eduction

- Growing number of users

Hardware
Dell, 180 disks, 540TB raw
WAL space, SSDs: 2TB per node…

Ref.	Description	Qté.	PU € HT	Disc	PU remisé € HT	Prix total € HT



Collaboration - Conclusion

- We are few of milestone astronomical experiments
  - Relying on Postgres already
    - Made few strategic choices because of XC existence
  - Willing to invest:
    - time in Postgres XC/XL
  - **Hoping XC/XL will merge their efforts into a holistic solution!!!**
  - Encouraging Tokyo meeting in 2014/03

  - Invested into:
    - testing and
    - into relevant hardware already
    - ported some extensions, i.e. plJava, tested many more..

  - Visible in the Gaia consortium
    - Linked to the final Gaia catalogue, possibly influencing their choice
  - Veeeery long term user
    - **production starts in mid 2015! – we will be deploying it next weeks!**

QA…