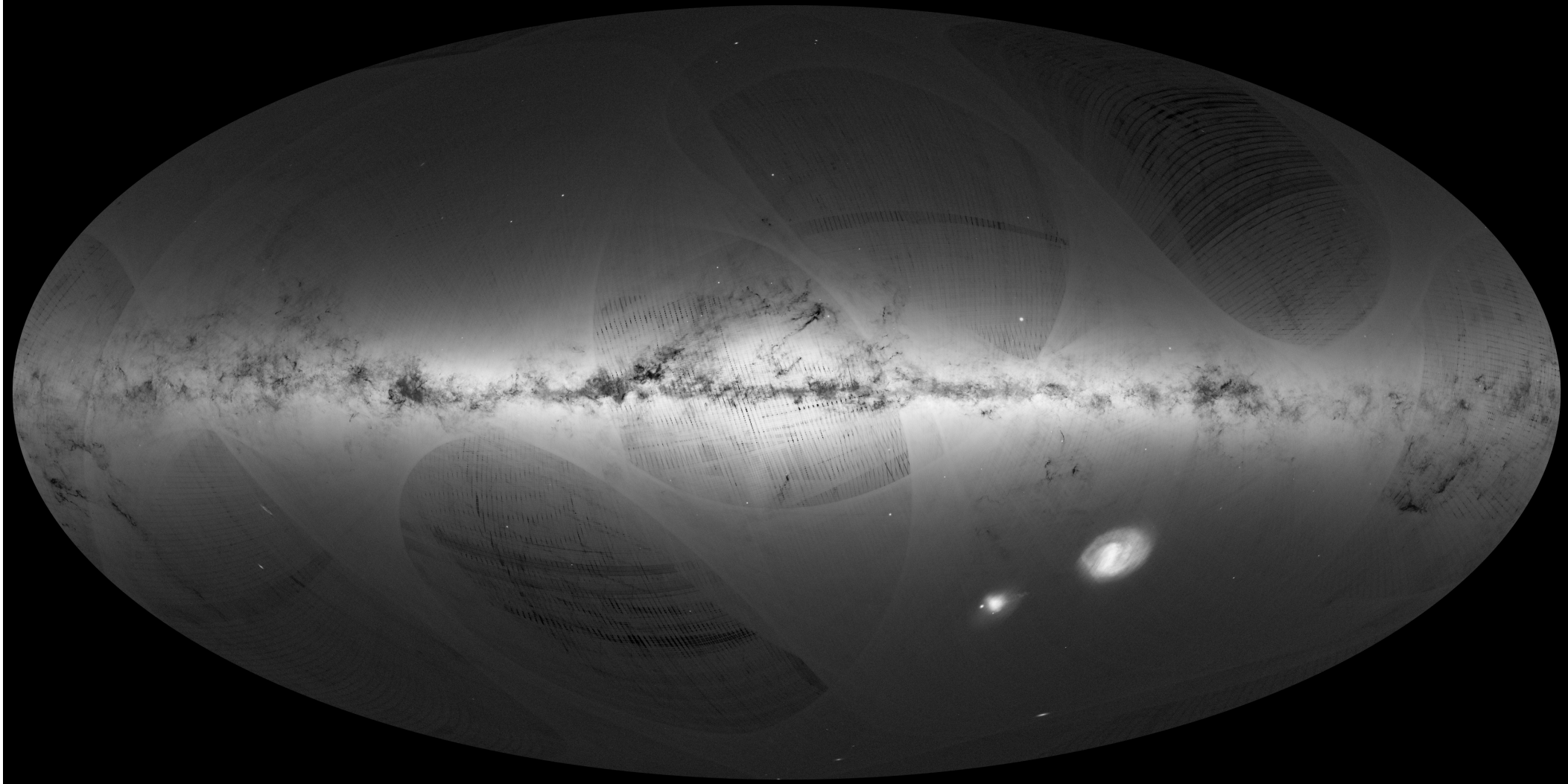
Methods to analyse the one billion time series of Gaia



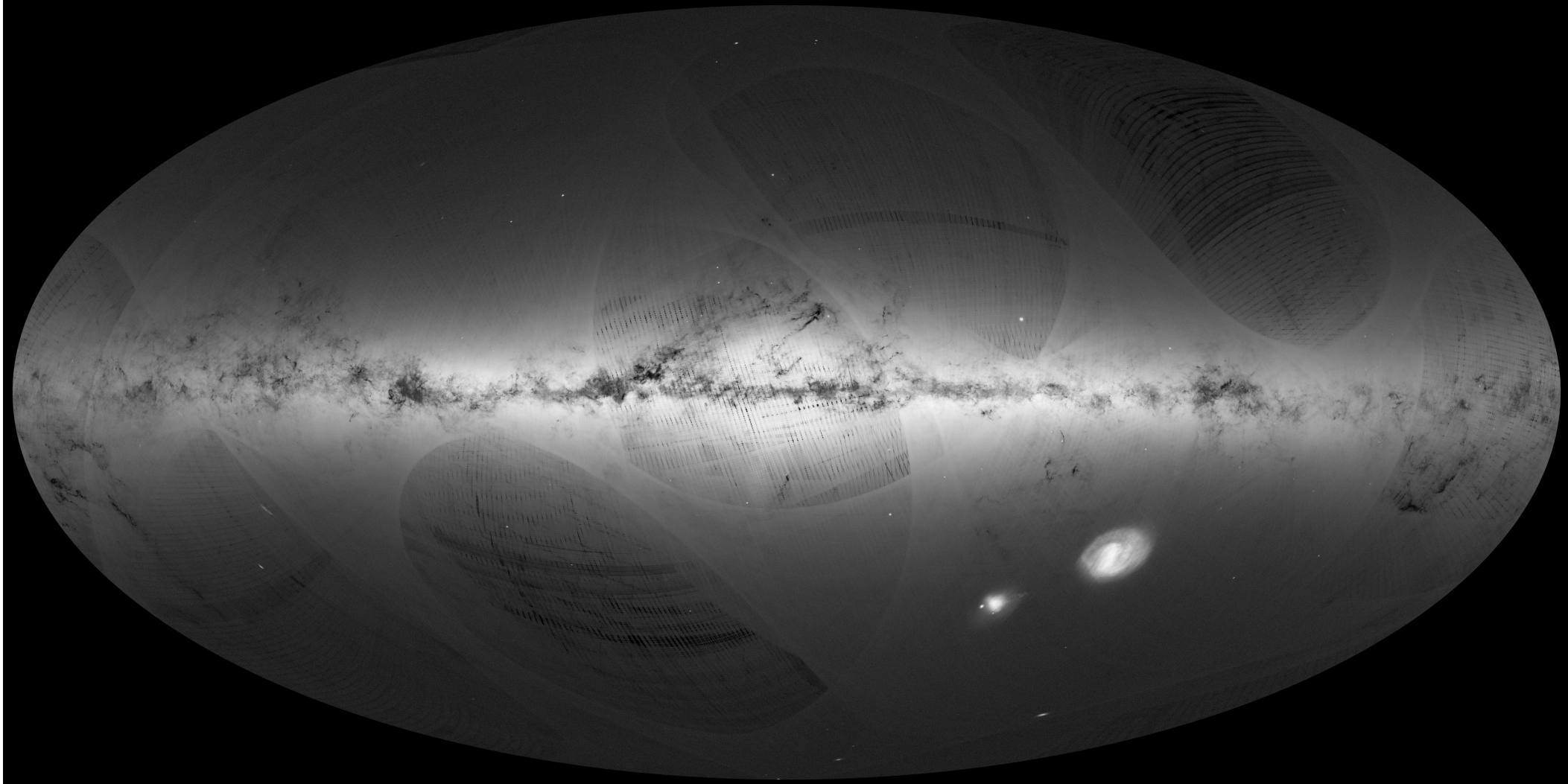
Laurent Eyer, Lorenzo Rimoldini and

Department of Astronomy, University of Geneva, Switzerland

Thursday, June 29, 2017

Prague, Czech Republic

Methods to analyse the one billion time series of Gaia



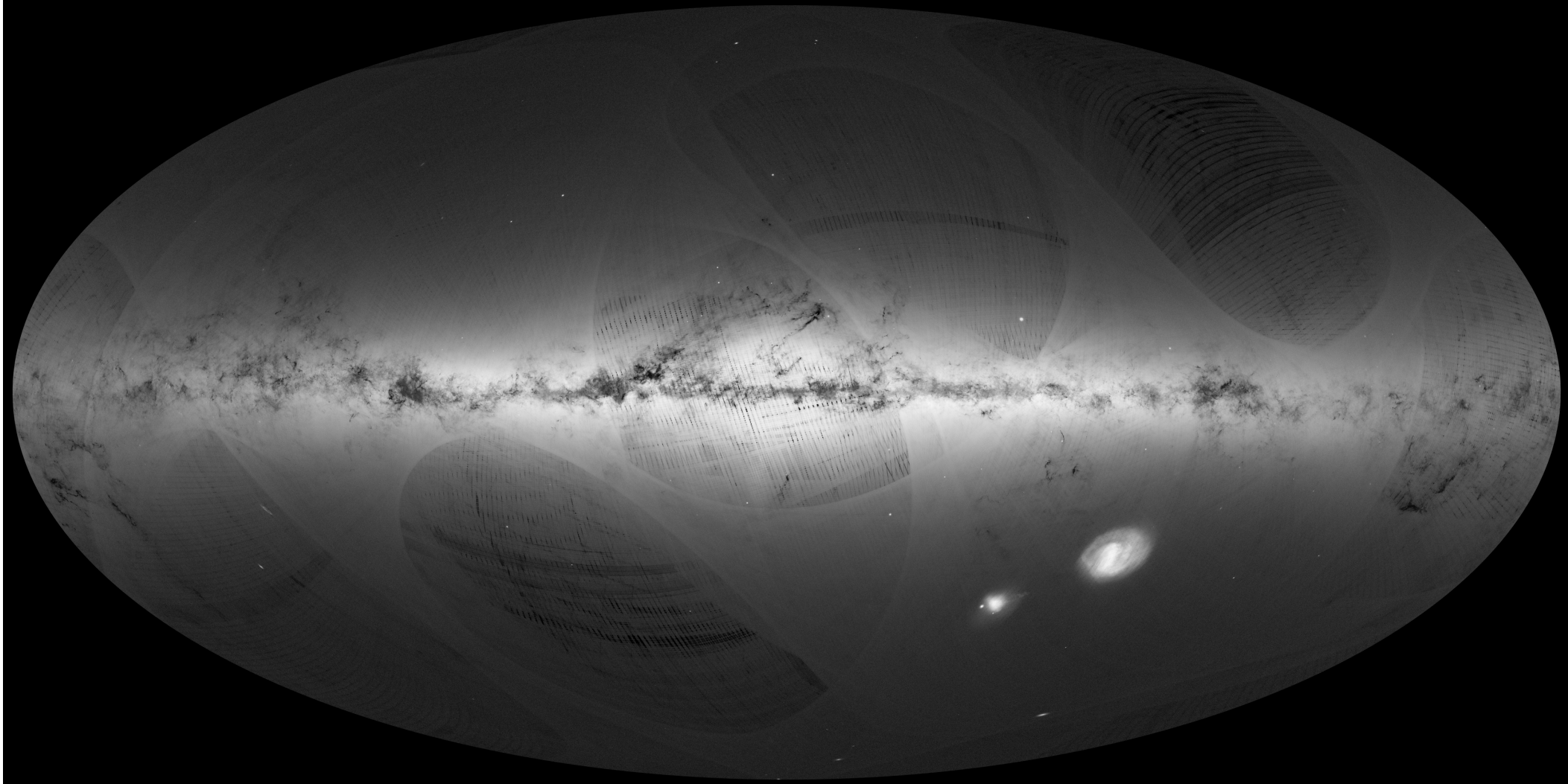
Laurent Eyer, Lorenzo Rimoldini and … CU7/DPCG

Department of Astronomy, University of Geneva, Switzerland

Thursday, June 29, 2017

Prague, Czech Republic

Methods to analyse the one billion time series of Gaia



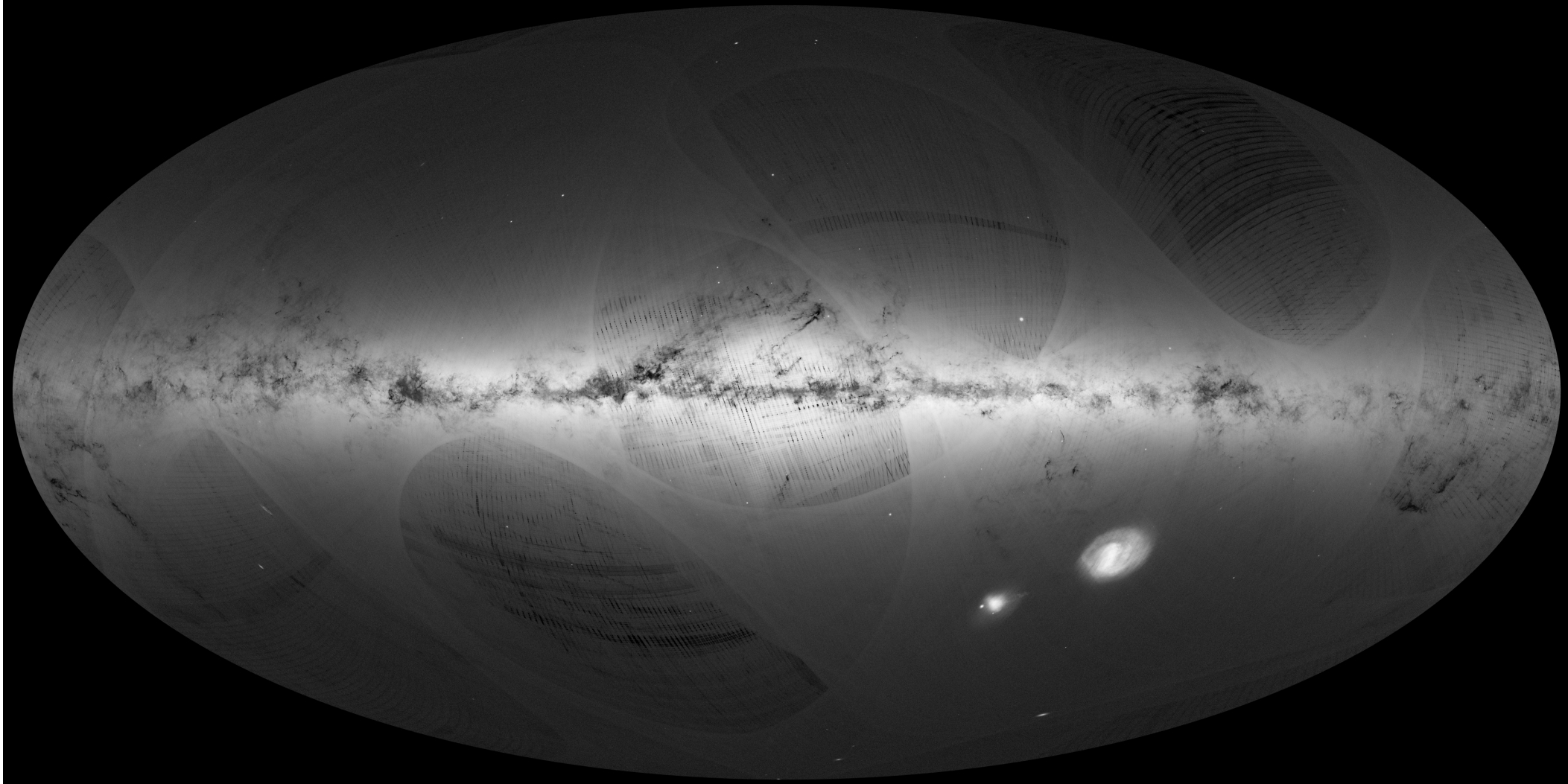
Laurent Eyer, Lorenzo Rimoldini and … CU7/DPCG, CU5/DPCI

Department of Astronomy, University of Geneva, Switzerland

Thursday, June 29, 2017

Prague, Czech Republic

Methods to analyse the one billion time series of Gaia



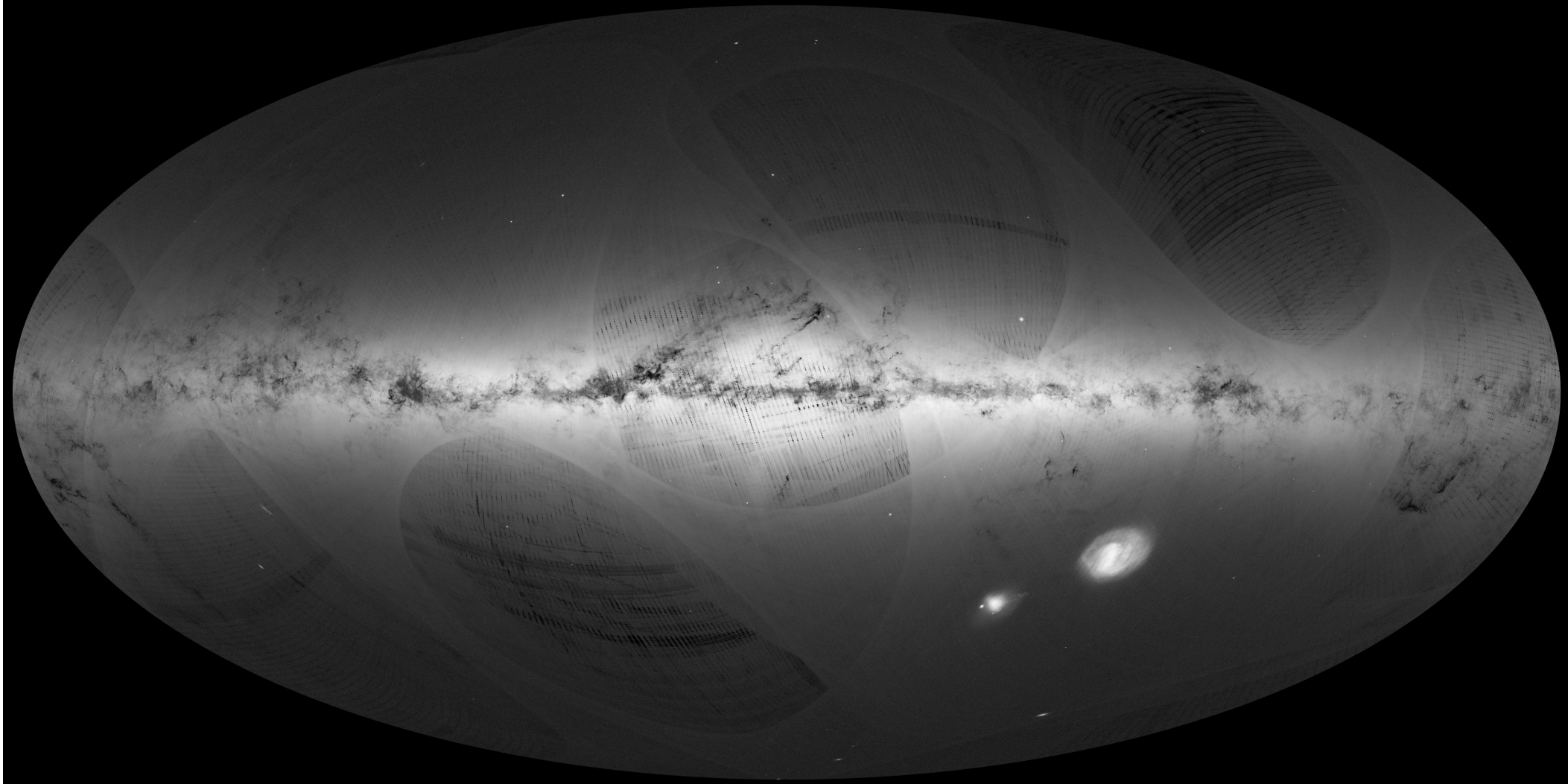
Laurent Eyer, Lorenzo Rimoldini and … CU7/DPCG, CU5/DPCI, CU3/ESAC

Department of Astronomy, University of Geneva, Switzerland

Thursday, June 29, 2017

Prague, Czech Republic

Methods to analyse the one billion time series of Gaia



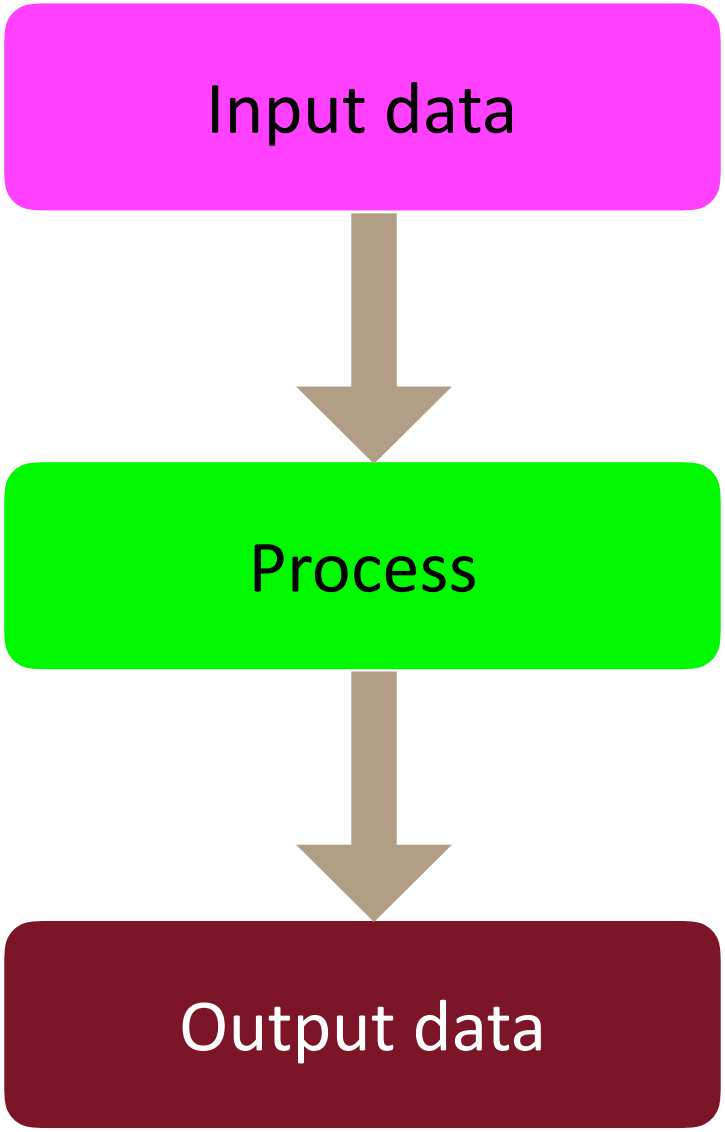
Laurent Eyer, Lorenzo Rimoldini and … CU7/DPCG, CU5/DPCI, CU3/ESAC, DPAC

Department of Astronomy, University of Geneva, Switzerland

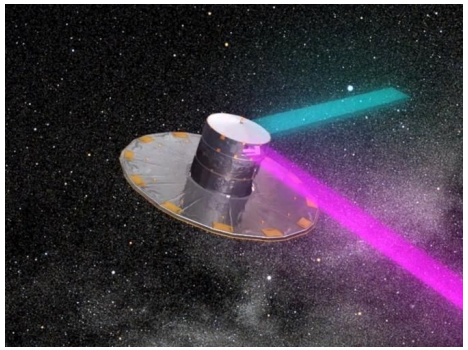
Thursday, June 29, 2017

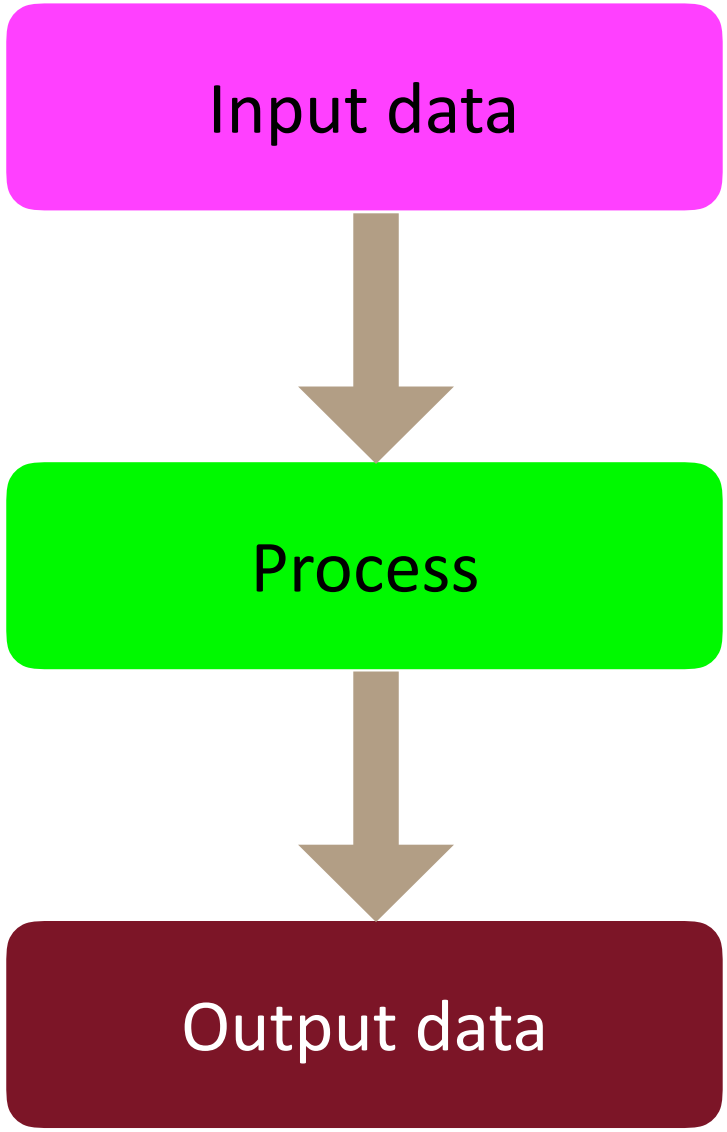
Prague, Czech Republic

Introduction: Problem looks simple

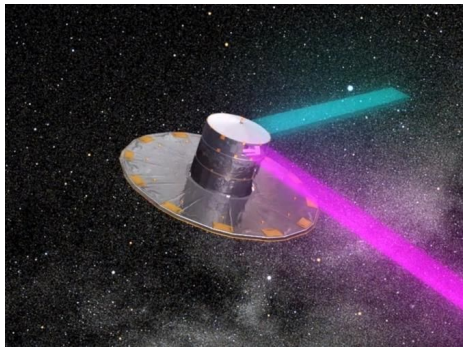


Introduction: Problem looks simple

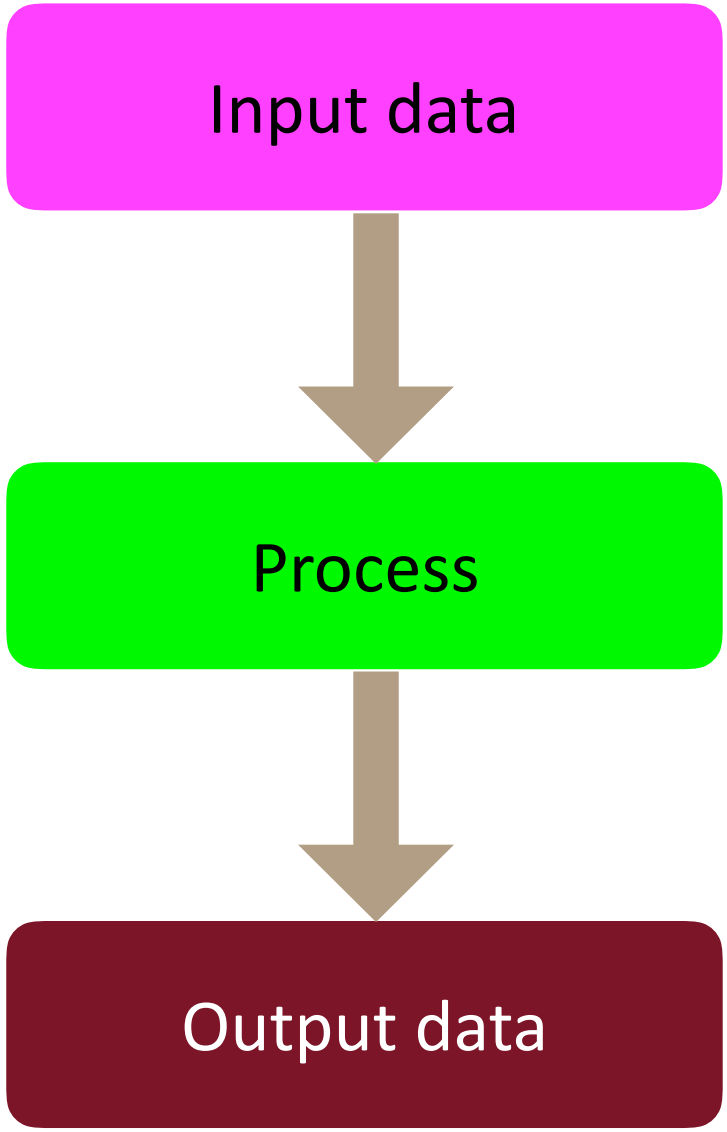




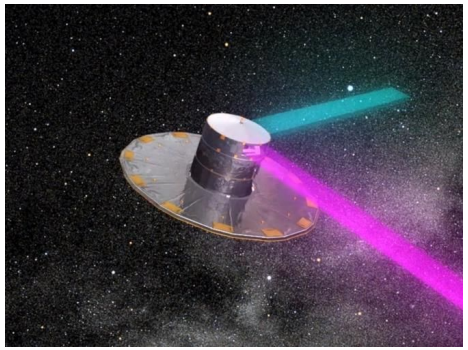
Introduction: Problem looks simple



Methods, algorithms, …

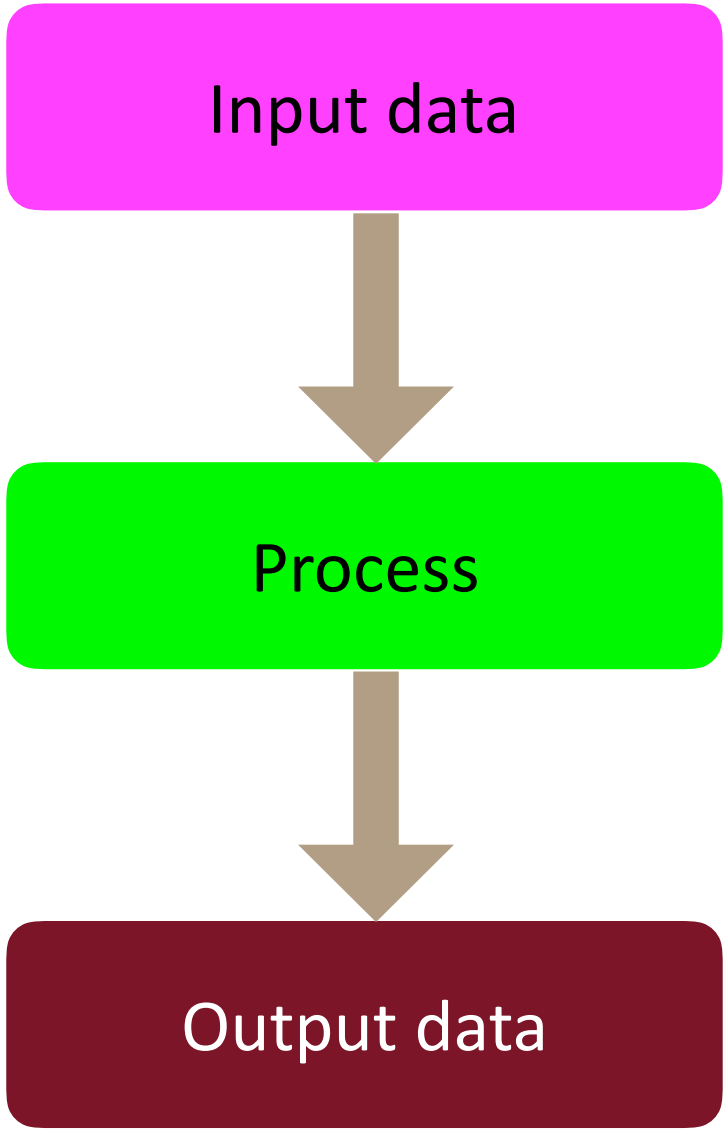


Introduction: Problem looks simple

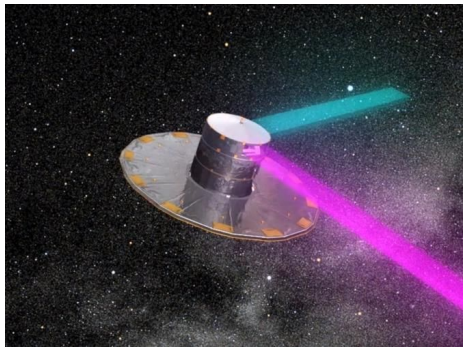


Methods, algorithms, …

$$y = \sum_{n=1}^{N_f} \sum_{k=1}^{N_h(n)} [a_{n,k} \sin(2\pi k f_n t) + b_{n,k} \cos(2\pi k f_n t)] + \sum_{i=0}^{N_p} c_i t^i$$



Introduction: Problem looks simple



Input data

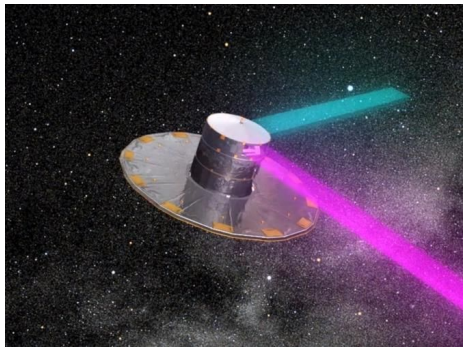
Methods, algorithms, …

$$y = \sum_{n=1}^{N_f} \sum_{k=1}^{N_h(n)} [a_{n,k} \sin(2\pi k f_n t) + b_{n,k} \cos(2\pi k f_n t)] + \sum_{i=0}^{N_p} c_i t^i$$

Process

Output data

Introduction: Problem looks simple

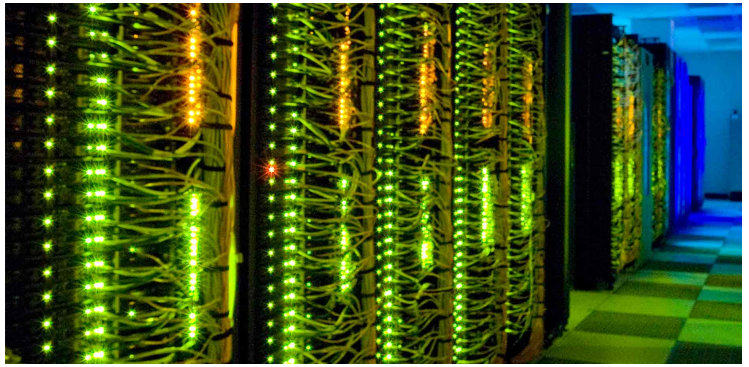


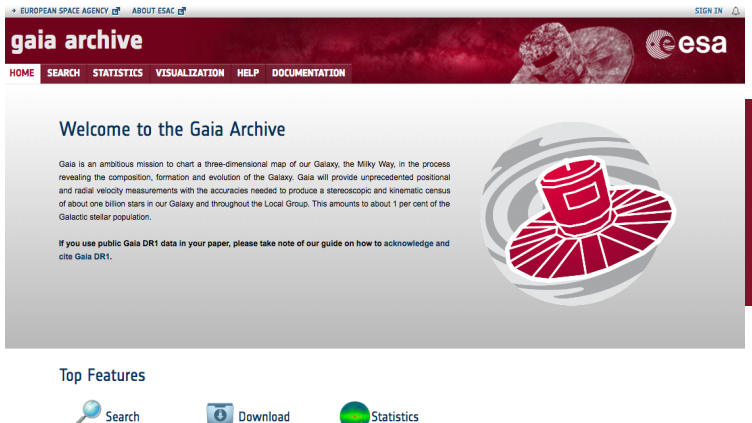
Input data

Methods, algorithms, …

$$y = \sum_{n=1}^{N_f} \sum_{k=1}^{N_h(n)} [a_{n,k} \sin(2\pi k f_n t) + b_{n,k} \cos(2\pi k f_n t)] + \sum_{i=0}^{N_p} c_i t^i$$

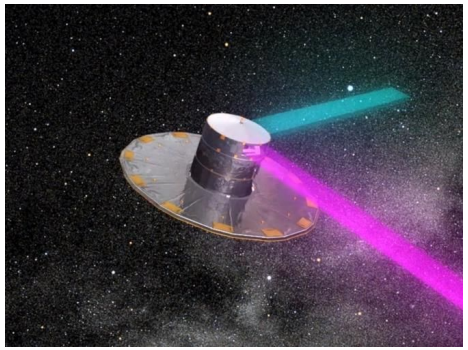
Process





Output data

Introduction: Problem looks simple

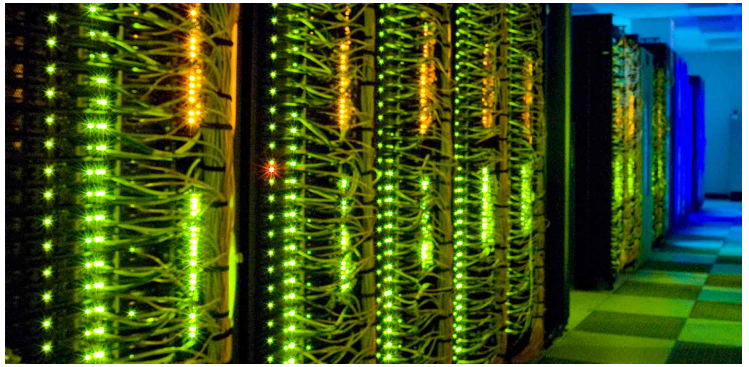


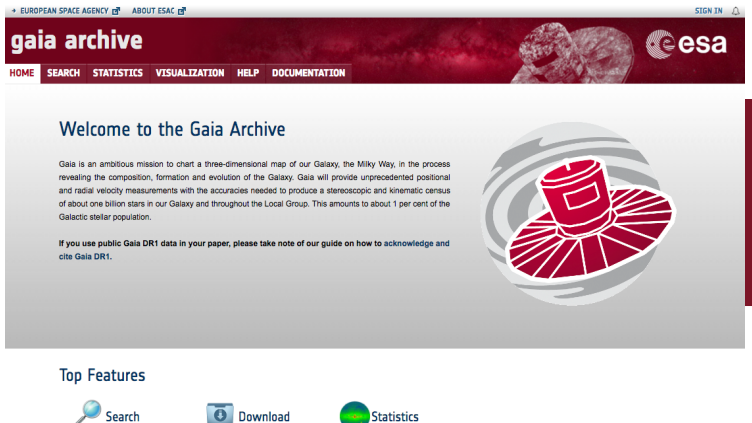
Input data

Methods, algorithms, …

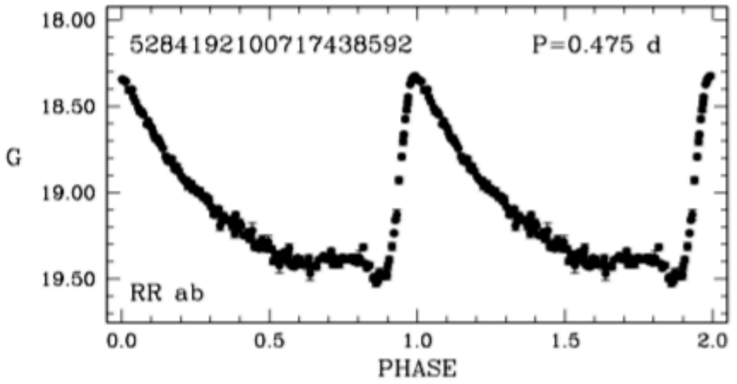
$$y = \sum_{n=1}^{N_f} \sum_{k=1}^{N_h(n)} [a_{n,k} \sin(2\pi k f_n t) + b_{n,k} \cos(2\pi k f_n t)] + \sum_{i=0}^{N_p} c_i t^i$$

Process

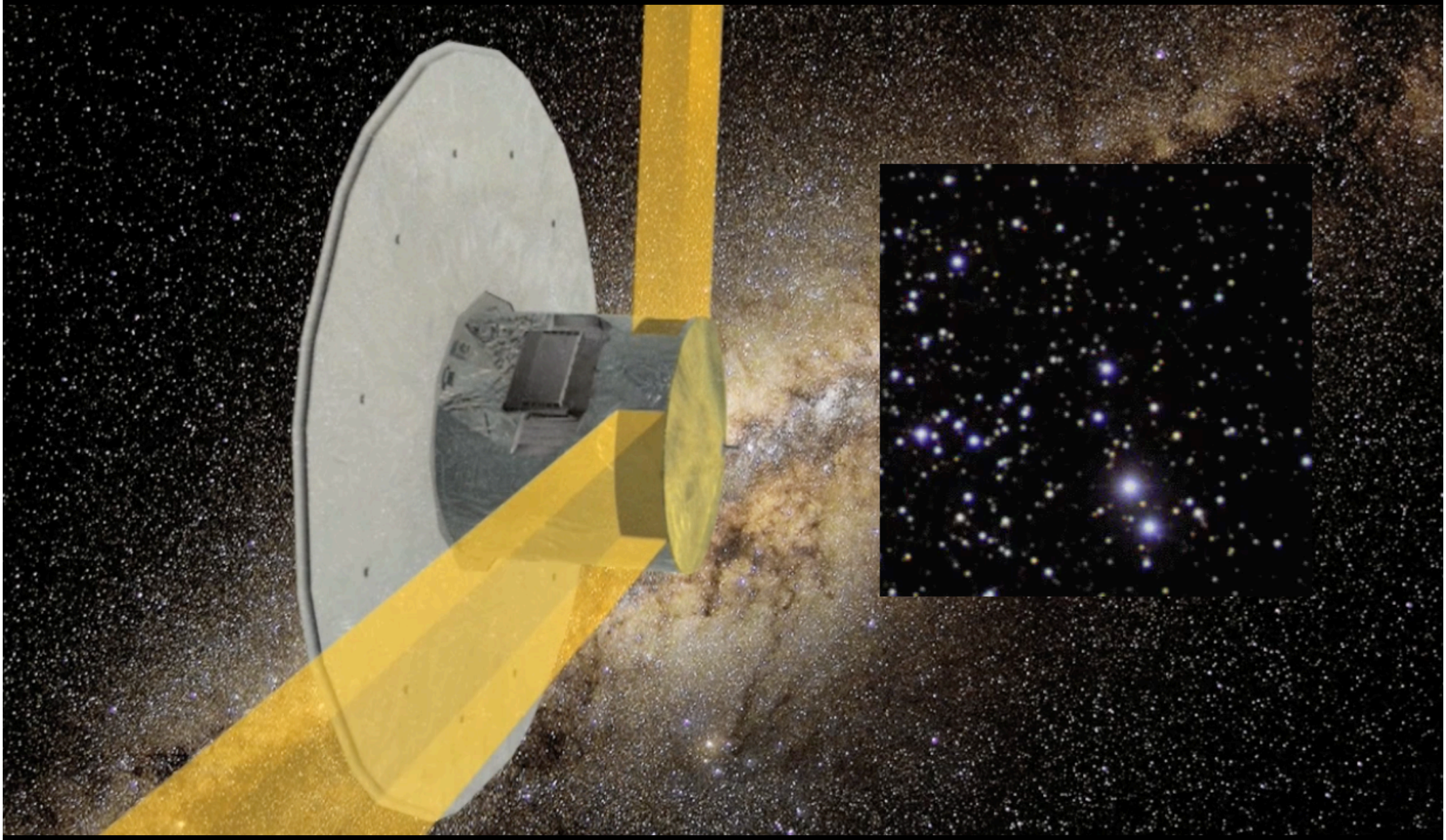




Output data

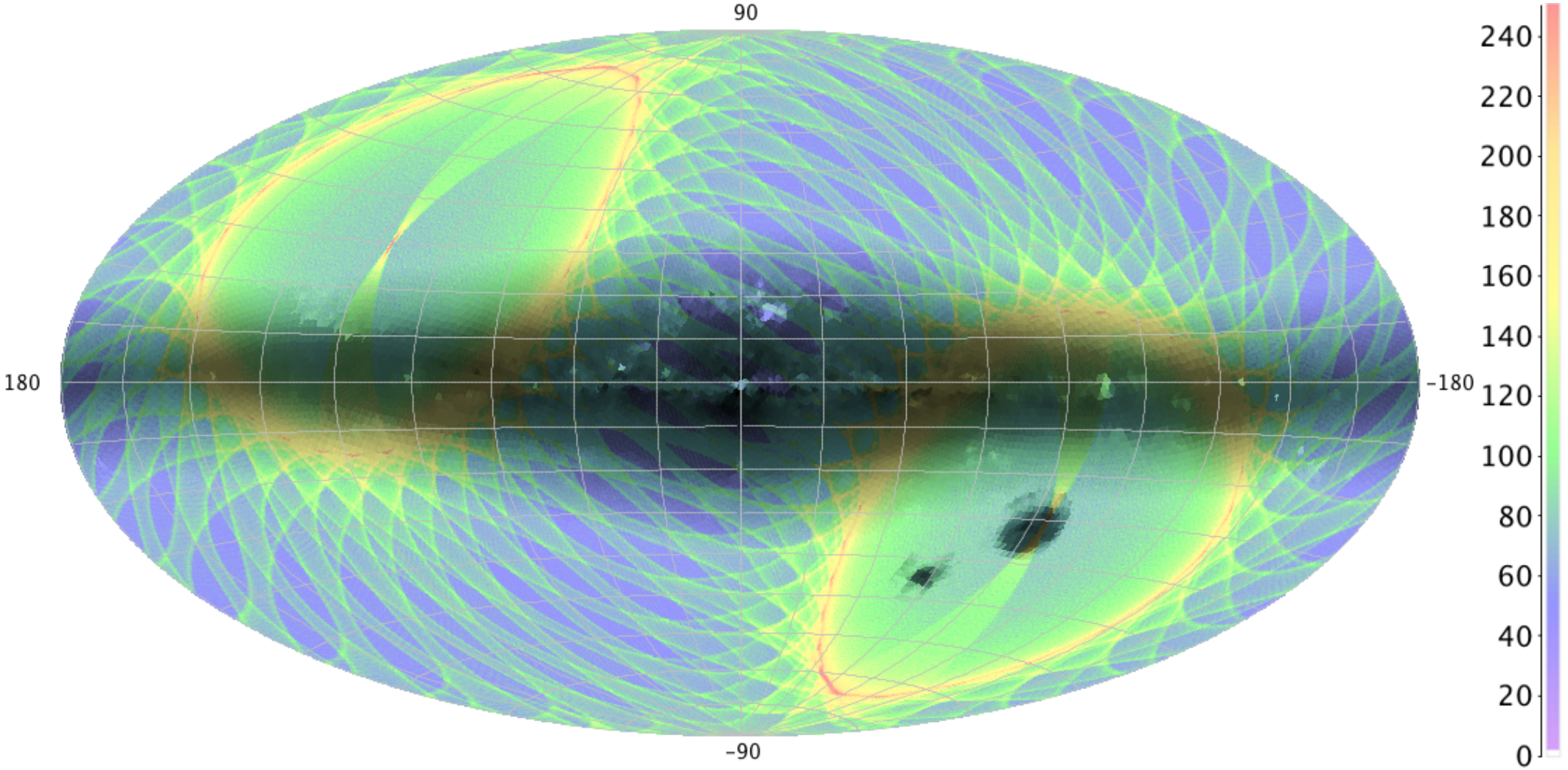


Introduction: Gaia, scanning the whole sky



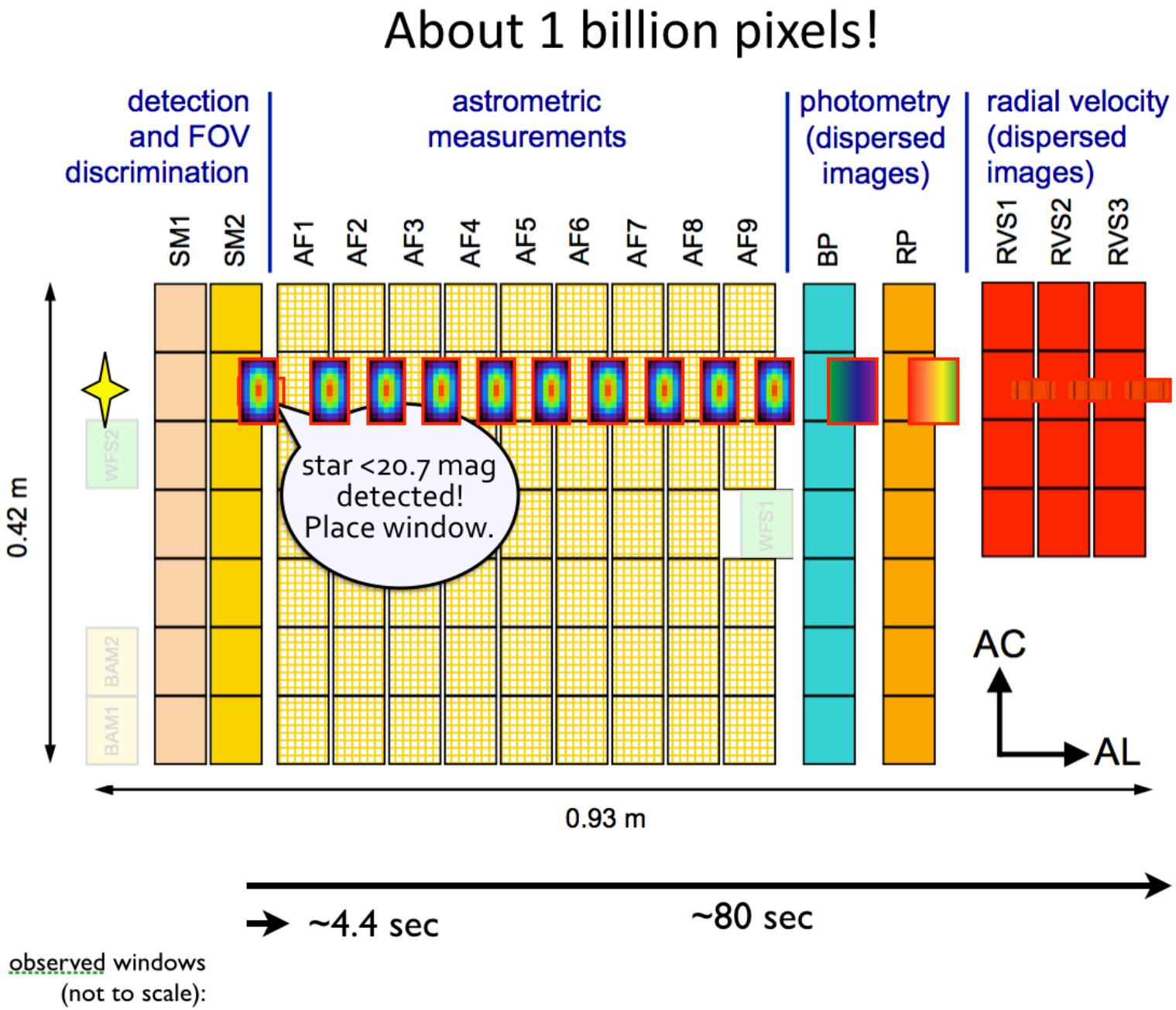
Sampling

Predicted 5 year mission sky coverage (galactic coordinates)



Eyer et al. 2017

The input data: the focal plane



Successive CCDs: 4.85 seconds

Gaia FoV: 0.7 deg x 0.7 deg
pixel: 0.059"(AL) x 0.177"(AC)

Courtesy of Berry Holl

The input data

The input data

Astrometry

Photometry

RVS instrument

The input data

Astrometry

per-CCD positions

Photometry

per-CCD flux

per-CCD spectra in blue and red

RVS instrument

per-CCD spectra in blue and red

The input data

Astrometry

per-CCD positions time series

Photometry

per-CCD flux time series

Field-of-View time series

per-CCD spectra in blue and red time series

BP, RP integrated time series

RVS instrument

per-CCD spectra in blue and red time series

Radial velocities time series

The input data

Astrometry

per-CCD positions — time series → 5 parameter astrometric solution

Photometry

per-CCD flux — time series → Mean G band

Field-of-View time series → Mean G band

per-CCD spectra in blue and red — time series → Mean colour, Mean spectra

BP, RP integrated time series → Mean colour, Mean spectra

RVS instrument

per-CCD spectra in blue and red — time series → Mean Radial Vel.

Radial velocities time series → Mean Radial Vel.

The input data

Astrometry

per-CCD positions time series

parameter

etric solution

Photometry

per-CCD flux time serie

Mean G band

per-CCD spectr eries

egrated time series

Mean colour

Mean spectra

RVS instrum

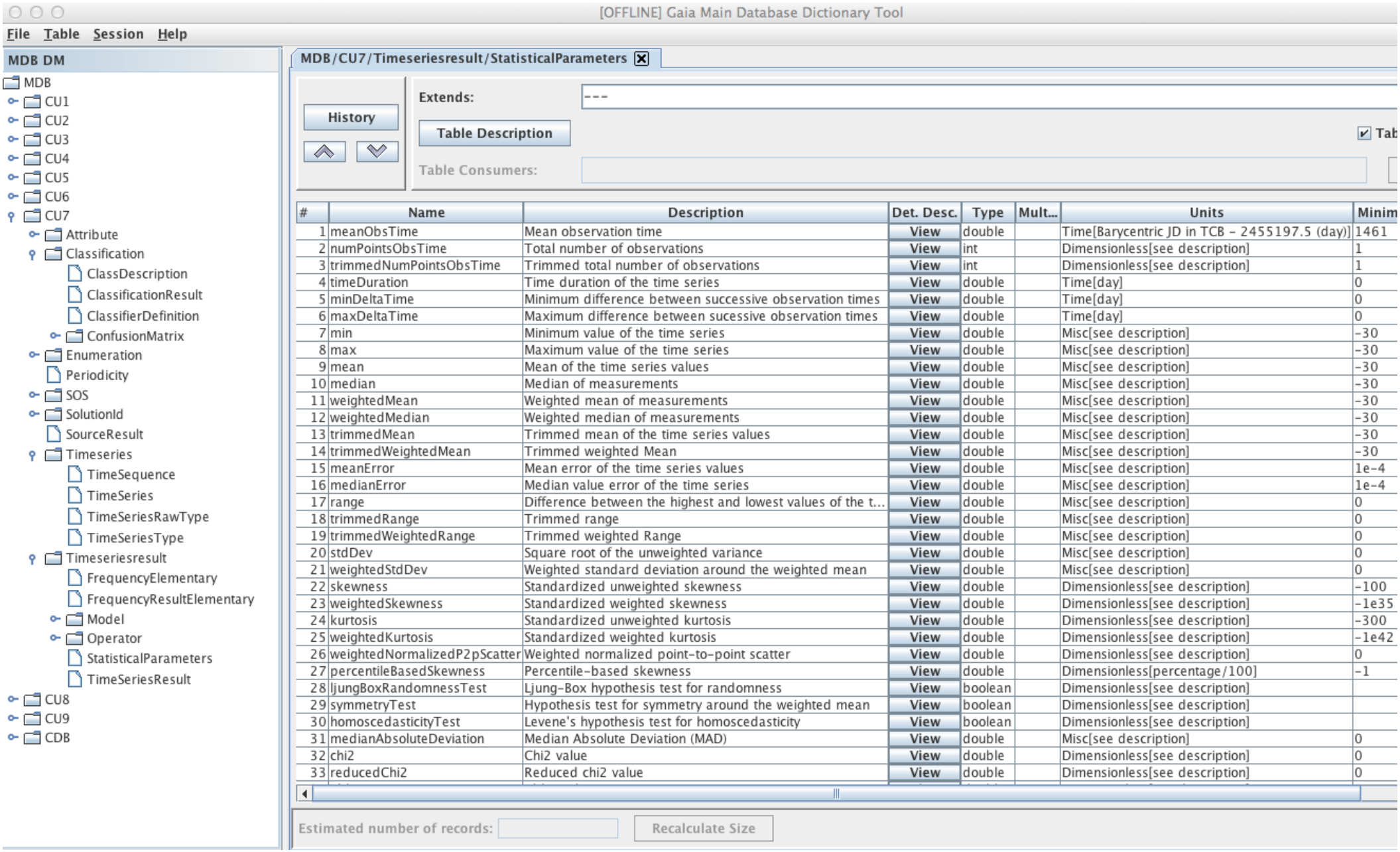
per-CCD spectra in blue and red time series

Mean Radial Vel.

Radial velocities time series

Time Domain Analysis
We (CU7) are interested in variable objects

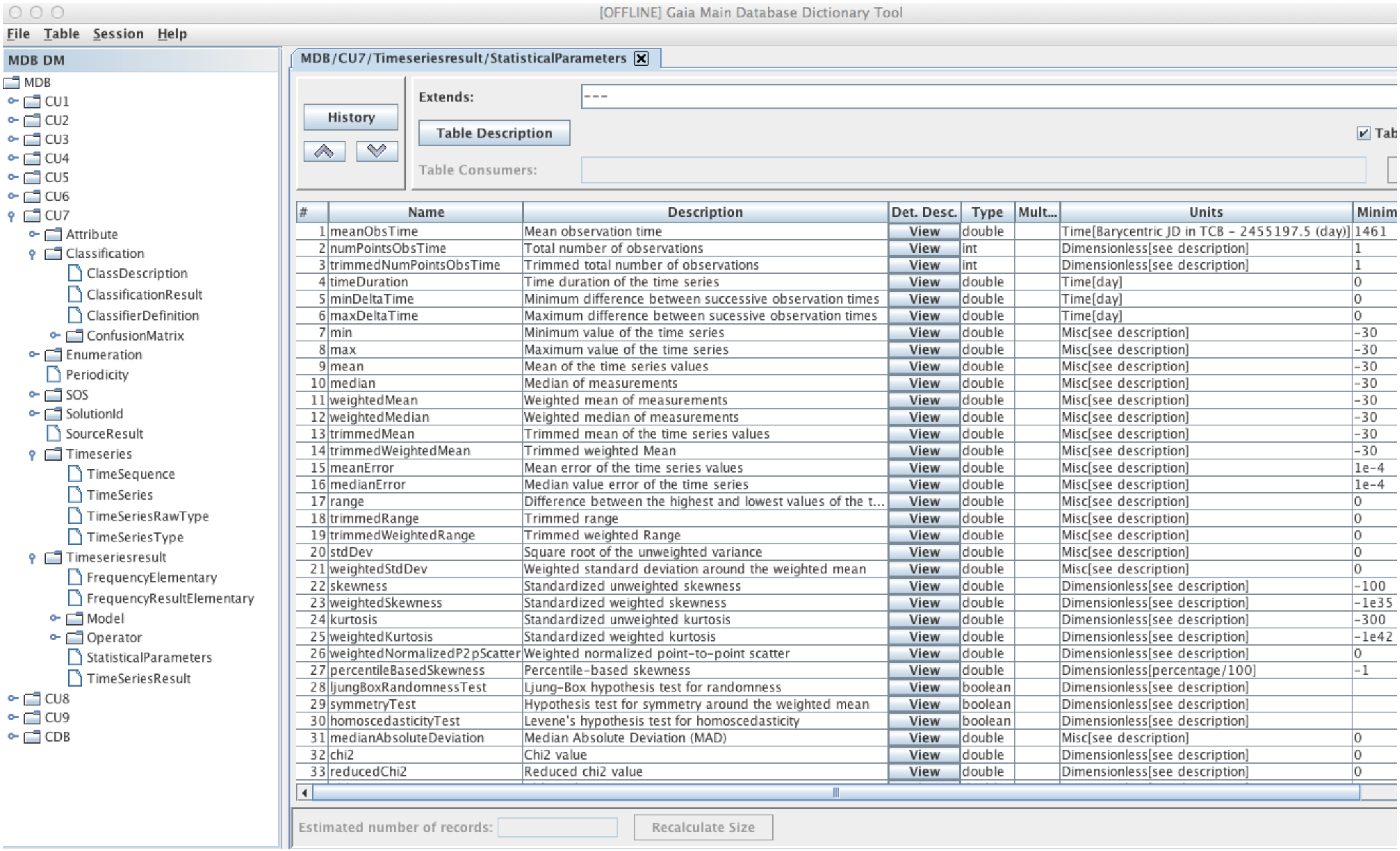
Gaia Main Database Dictionary Tool (in ESAC)



#	Name	Description	Det. Desc.	Type	Mult...	Units	Minim
1	meanObsTime	Mean observation time	View	double		Time[Barycentric JD in TCB - 2455197.5 (day)]	1461
2	numPointsObsTime	Total number of observations	View	int		Dimensionless[see description]	1
3	trimmedNumPointsObsTime	Trimmed total number of observations	View	int		Dimensionless[see description]	1
4	timeDuration	Time duration of the time series	View	double		Time[day]	0
5	minDeltaTime	Minimum difference between successive observation times	View	double		Time[day]	0
6	maxDeltaTime	Maximum difference between sucessive observation times	View	double		Time[day]	0
7	min	Minimum value of the time series	View	double		Misc[see description]	-30
8	max	Maximum value of the time series	View	double		Misc[see description]	-30
9	mean	Mean of the time series values	View	double		Misc[see description]	-30
10	median	Median of measurements	View	double		Misc[see description]	-30
11	weightedMean	Weighted mean of measurements	View	double		Misc[see description]	-30
12	weightedMedian	Weighted median of measurements	View	double		Misc[see description]	-30
13	trimmedMean	Trimmed mean of the time series values	View	double		Misc[see description]	-30
14	trimmedWeightedMean	Trimmed weighted Mean	View	double		Misc[see description]	-30
15	meanError	Mean error of the time series values	View	double		Misc[see description]	1e-4
16	medianError	Median value error of the time series	View	double		Misc[see description]	1e-4
17	range	Difference between the highest and lowest values of the t...	View	double		Misc[see description]	0
18	trimmedRange	Trimmed range	View	double		Misc[see description]	0
19	trimmedWeightedRange	Trimmed weighted Range	View	double		Misc[see description]	0
20	stdDev	Square root of the unweighted variance	View	double		Misc[see description]	0
21	weightedStdDev	Weighted standard deviation around the weighted mean	View	double		Misc[see description]	0
22	skewness	Standardized unweighted skewness	View	double		Dimensionless[see description]	-100
23	weightedSkewness	Standardized weighted skewness	View	double		Dimensionless[see description]	-1e35
24	kurtosis	Standardized unweighted kurtosis	View	double		Dimensionless[see description]	-300
25	weightedKurtosis	Standardized weighted kurtosis	View	double		Dimensionless[see description]	-1e42
26	weightedNormalizedP2pScatter	Weighted normalized point-to-point scatter	View	double		Dimensionless[see description]	0
27	percentileBasedSkewness	Percentile-based skewness	View	double		Dimensionless[percentage/100]	-1
28	ljungBoxRandomnessTest	Ljung-Box hypothesis test for randomness	View	boolean		Dimensionless[see description]	
29	symmetryTest	Hypothesis test for symmetry around the weighted mean	View	boolean		Dimensionless[see description]	
30	homoscedasticityTest	Levene's hypothesis test for homoscedasticity	View	boolean		Dimensionless[see description]	
31	medianAbsoluteDeviation	Median Absolute Deviation (MAD)	View	double		Misc[see description]	0
32	chi2	Chi2 value	View	double		Dimensionless[see description]	0
33	reducedChi2	Reduced chi2 value	View	double		Dimensionless[see description]	0

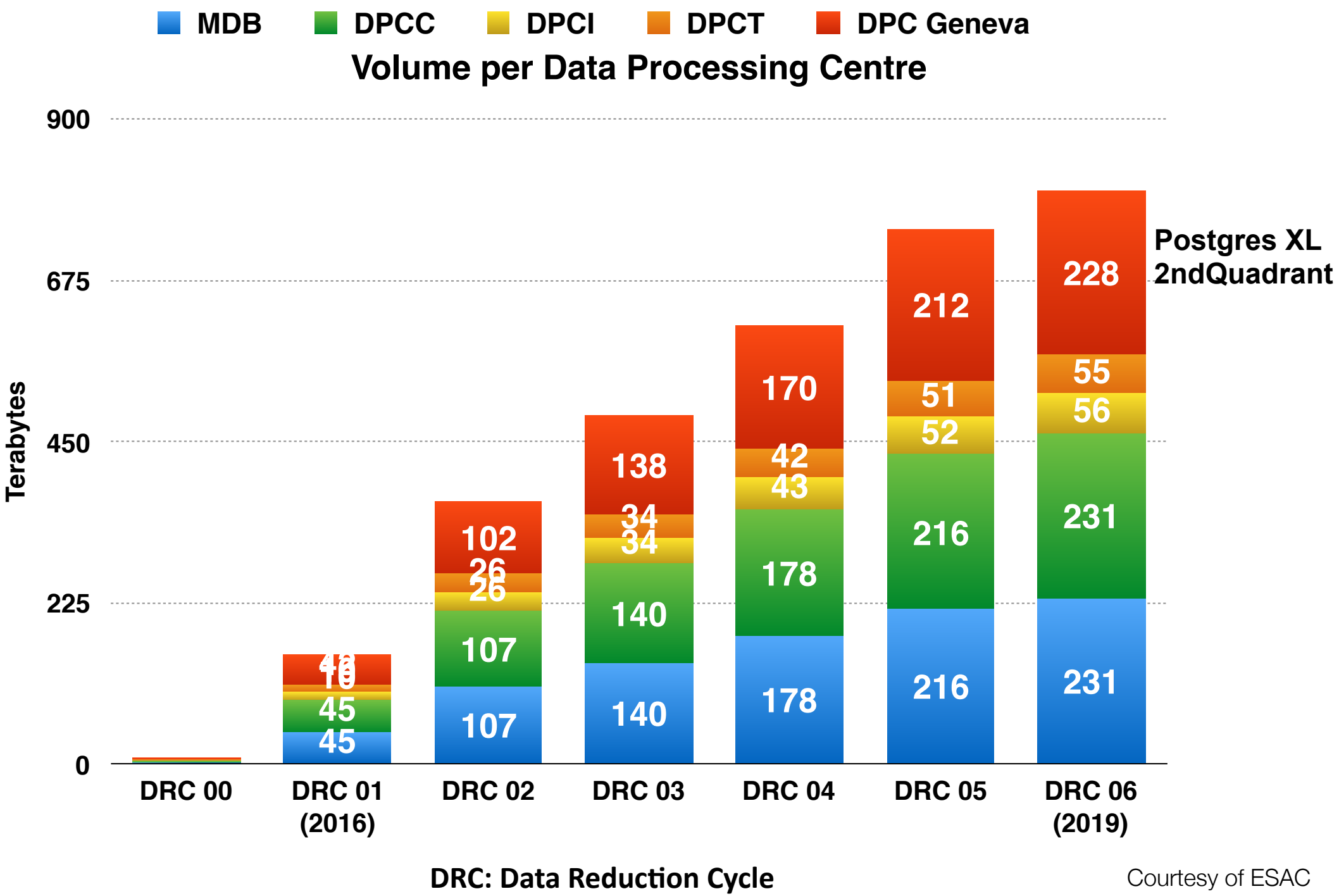
Gaia Main Database Dictionary Tool (in ESAC)

Total number of parameters (now): 5,540



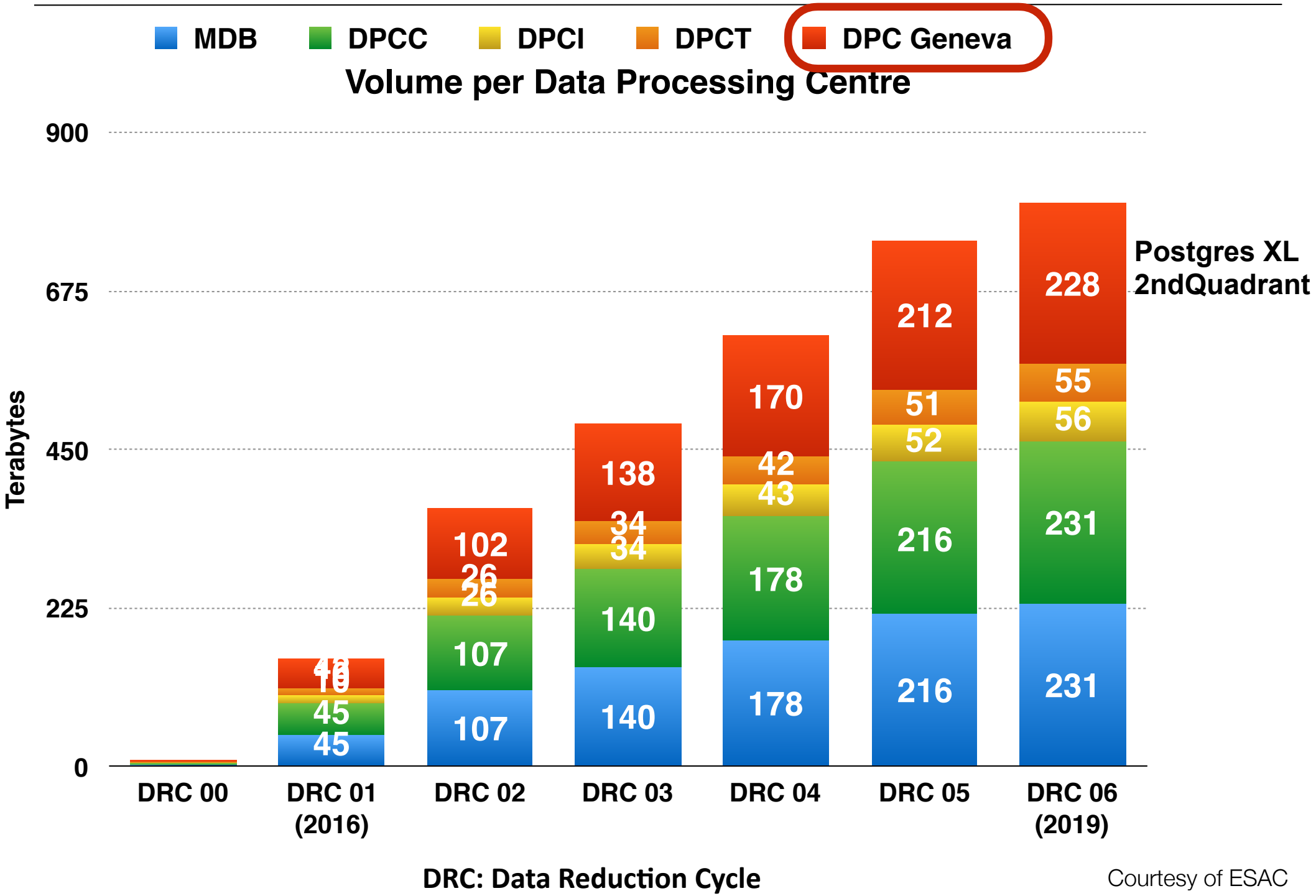
#	Name	Description	Det. Desc.	Type	Mult...	Units	Minim
1	meanObsTime	Mean observation time	View	double		Time[Barycentric JD in TCB - 2455197.5 (day)]	1461
2	numPointsObsTime	Total number of observations	View	int		Dimensionless[see description]	1
3	trimmedNumPointsObsTime	Trimmed total number of observations	View	int		Dimensionless[see description]	1
4	timeDuration	Time duration of the time series	View	double		Time[day]	0
5	minDeltaTime	Minimum difference between successive observation times	View	double		Time[day]	0
6	maxDeltaTime	Maximum difference between sucessive observation times	View	double		Time[day]	0
7	min	Minimum value of the time series	View	double		Misc[see description]	-30
8	max	Maximum value of the time series	View	double		Misc[see description]	-30
9	mean	Mean of the time series values	View	double		Misc[see description]	-30
10	median	Median of measurements	View	double		Misc[see description]	-30
11	weightedMean	Weighted mean of measurements	View	double		Misc[see description]	-30
12	weightedMedian	Weighted median of measurements	View	double		Misc[see description]	-30
13	trimmedMean	Trimmed mean of the time series values	View	double		Misc[see description]	-30
14	trimmedWeightedMean	Trimmed weighted Mean	View	double		Misc[see description]	-30
15	meanError	Mean error of the time series values	View	double		Misc[see description]	1e-4
16	medianError	Median value error of the time series	View	double		Misc[see description]	1e-4
17	range	Difference between the highest and lowest values of the t...	View	double		Misc[see description]	0
18	trimmedRange	Trimmed range	View	double		Misc[see description]	0
19	trimmedWeightedRange	Trimmed weighted Range	View	double		Misc[see description]	0
20	stdDev	Square root of the unweighted variance	View	double		Misc[see description]	0
21	weightedStdDev	Weighted standard deviation around the weighted mean	View	double		Misc[see description]	0
22	skewness	Standardized unweighted skewness	View	double		Dimensionless[see description]	-100
23	weightedSkewness	Standardized weighted skewness	View	double		Dimensionless[see description]	-1e35
24	kurtosis	Standardized unweighted kurtosis	View	double		Dimensionless[see description]	-300
25	weightedKurtosis	Standardized weighted kurtosis	View	double		Dimensionless[see description]	-1e42
26	weightedNormalizedP2pScatter	Weighted normalized point-to-point scatter	View	double		Dimensionless[see description]	0
27	percentileBasedSkewness	Percentile-based skewness	View	double		Dimensionless[percentage/100]	-1
28	ljungBoxRandomnessTest	Ljung-Box hypothesis test for randomness	View	boolean		Dimensionless[see description]	
29	symmetryTest	Hypothesis test for symmetry around the weighted mean	View	boolean		Dimensionless[see description]	
30	homoscedasticityTest	Levene's hypothesis test for homoscedasticity	View	boolean		Dimensionless[see description]	
31	medianAbsoluteDeviation	Median Absolute Deviation (MAD)	View	double		Misc[see description]	0
32	chi2	Chi2 value	View	double		Dimensionless[see description]	0
33	reducedChi2	Reduced chi2 value	View	double		Dimensionless[see description]	0

Cyclic Volume challenge - Petabyte scale



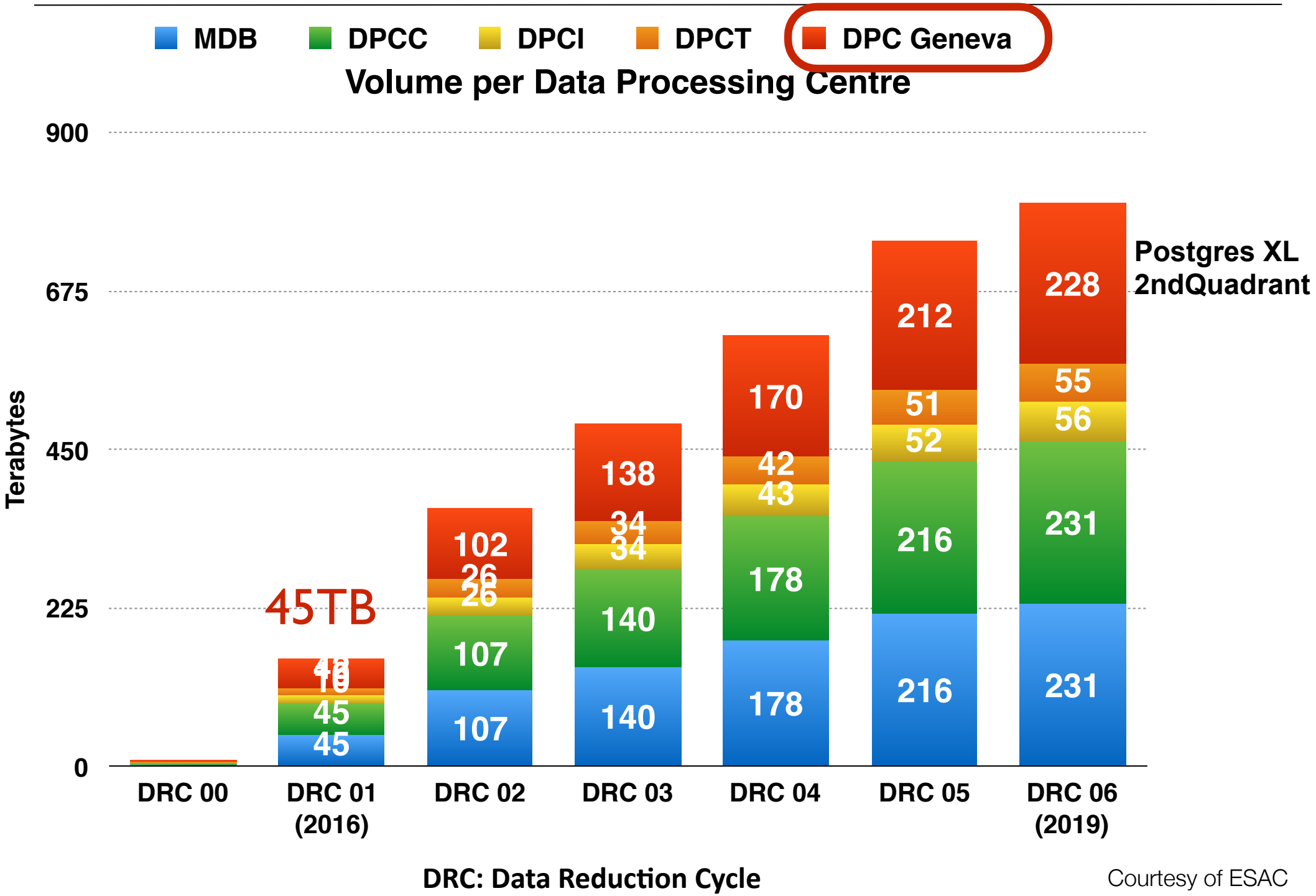
Courtesy of ESAC

Cyclic Volume challenge - Petabyte scale



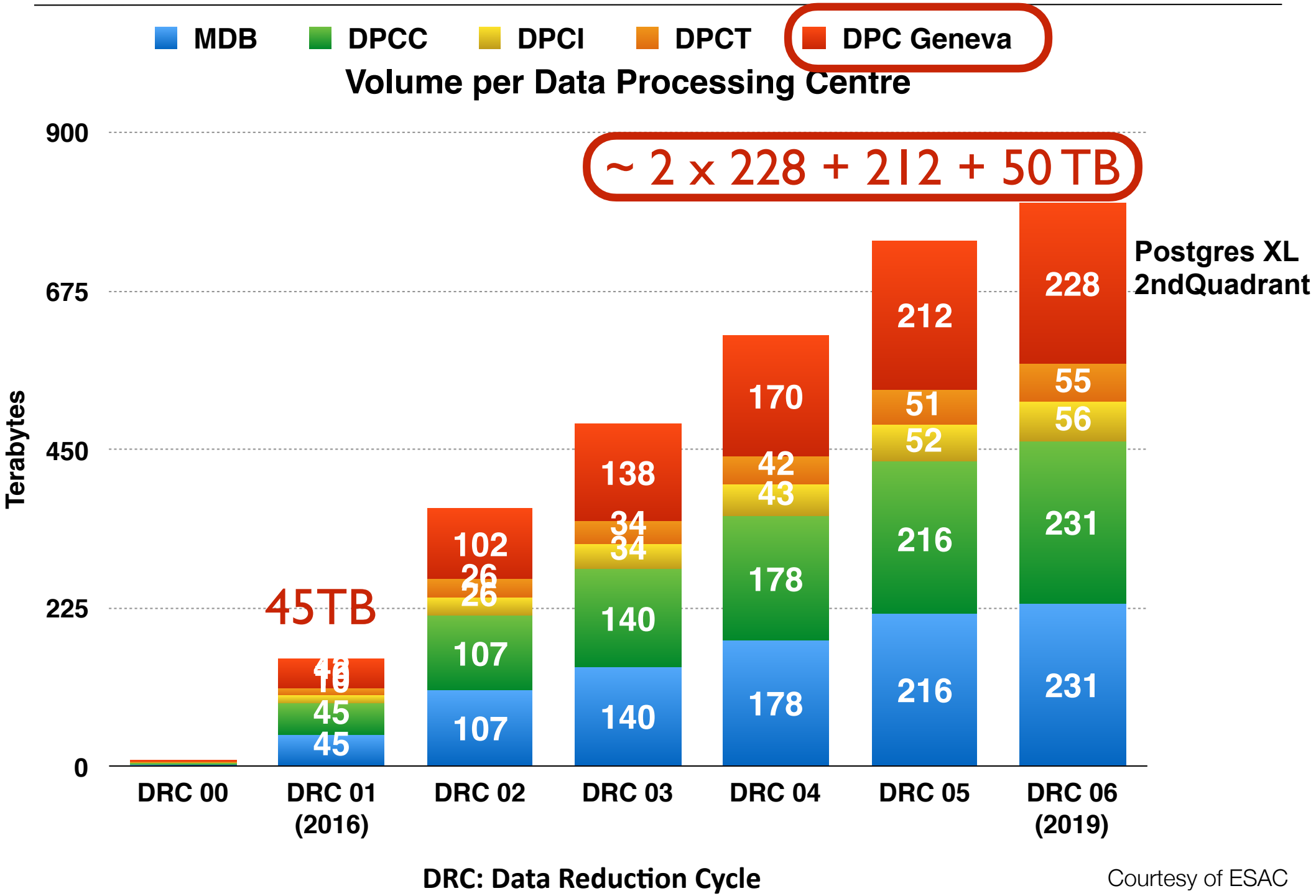
Courtesy of ESAC

Cyclic Volume challenge - Petabyte scale



Courtesy of ESAC

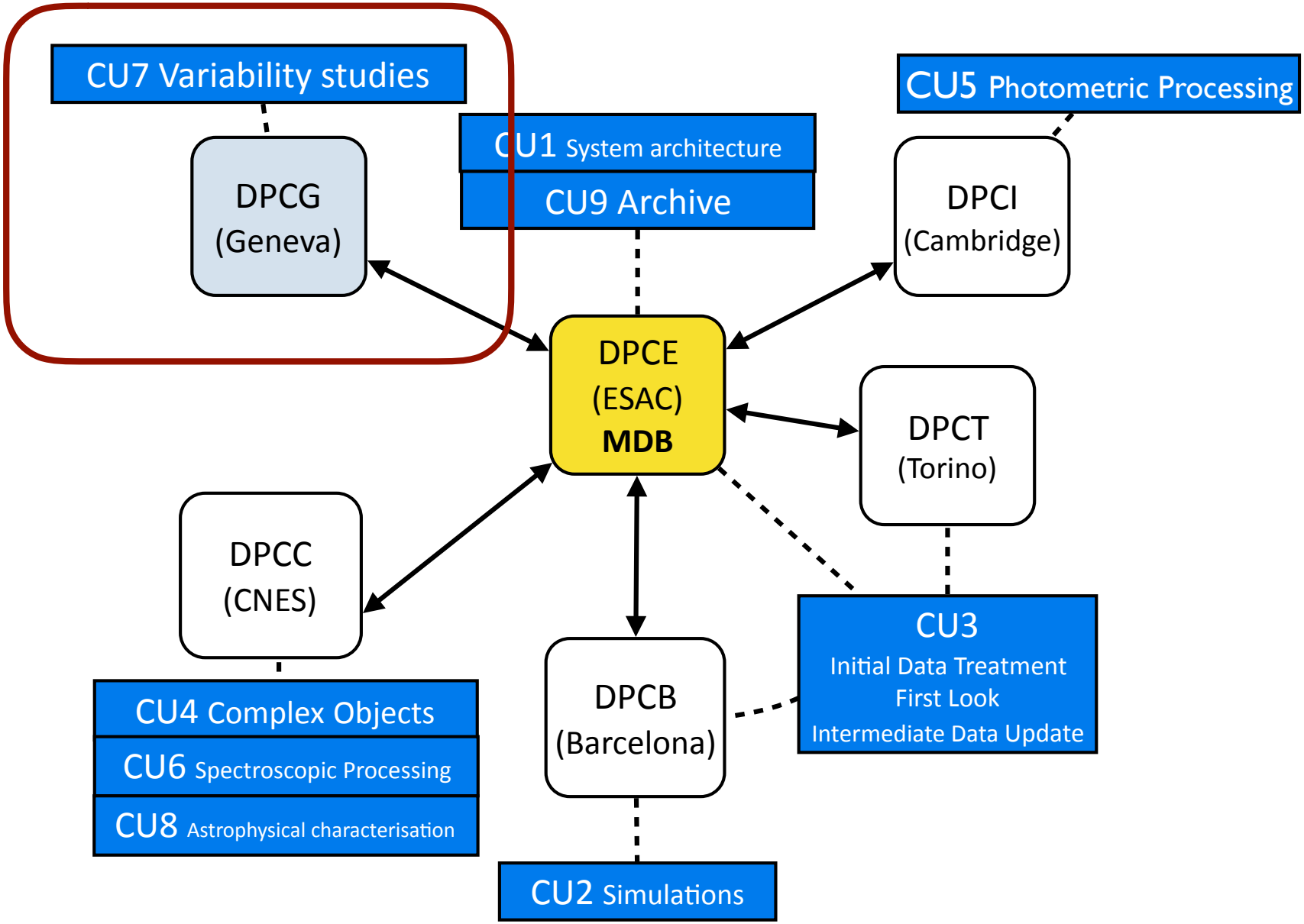
Cyclic Volume challenge - Petabyte scale



Courtesy of ESAC

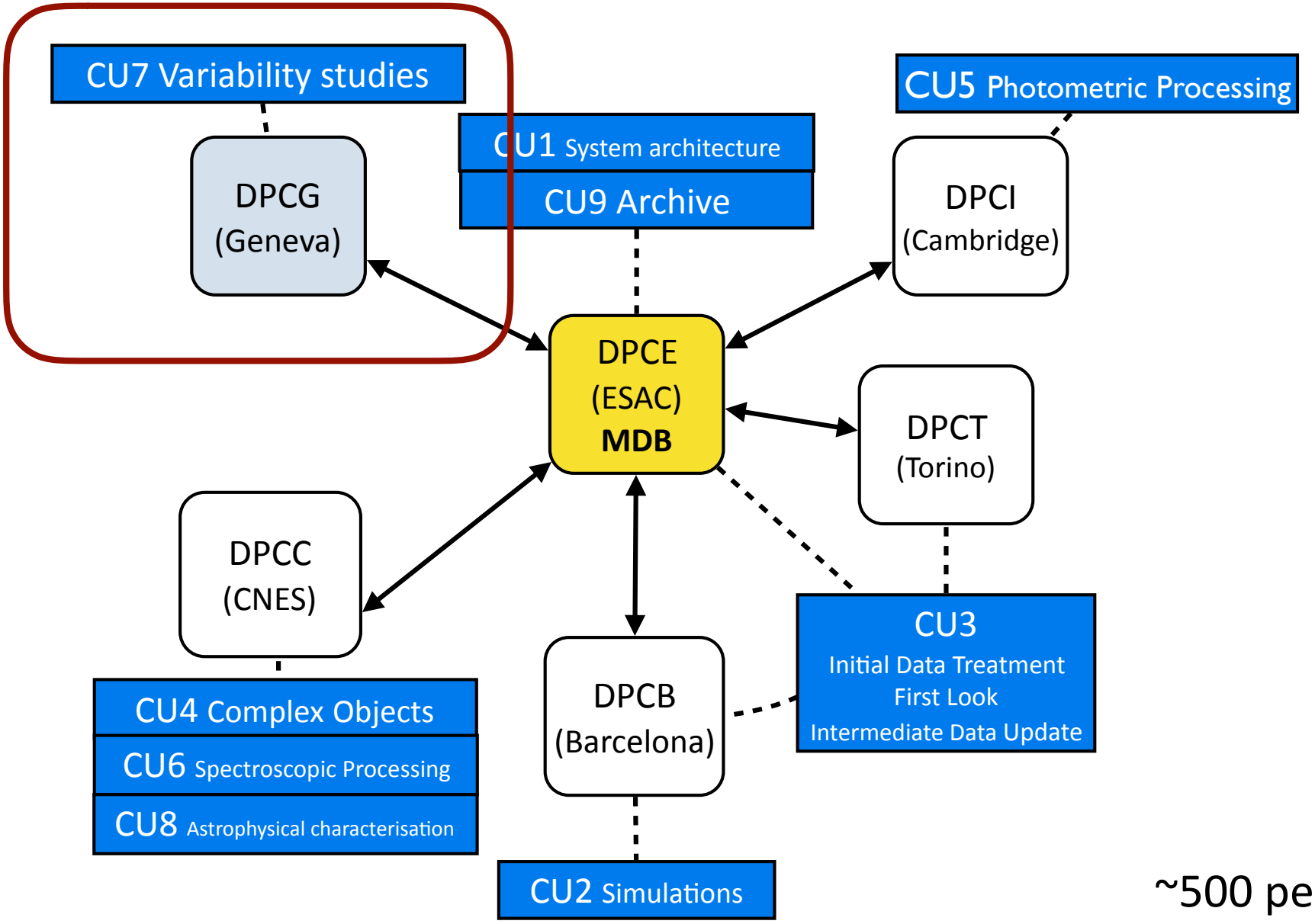
Gaia Consortium - distributed challenge

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



Gaia Consortium - distributed challenge

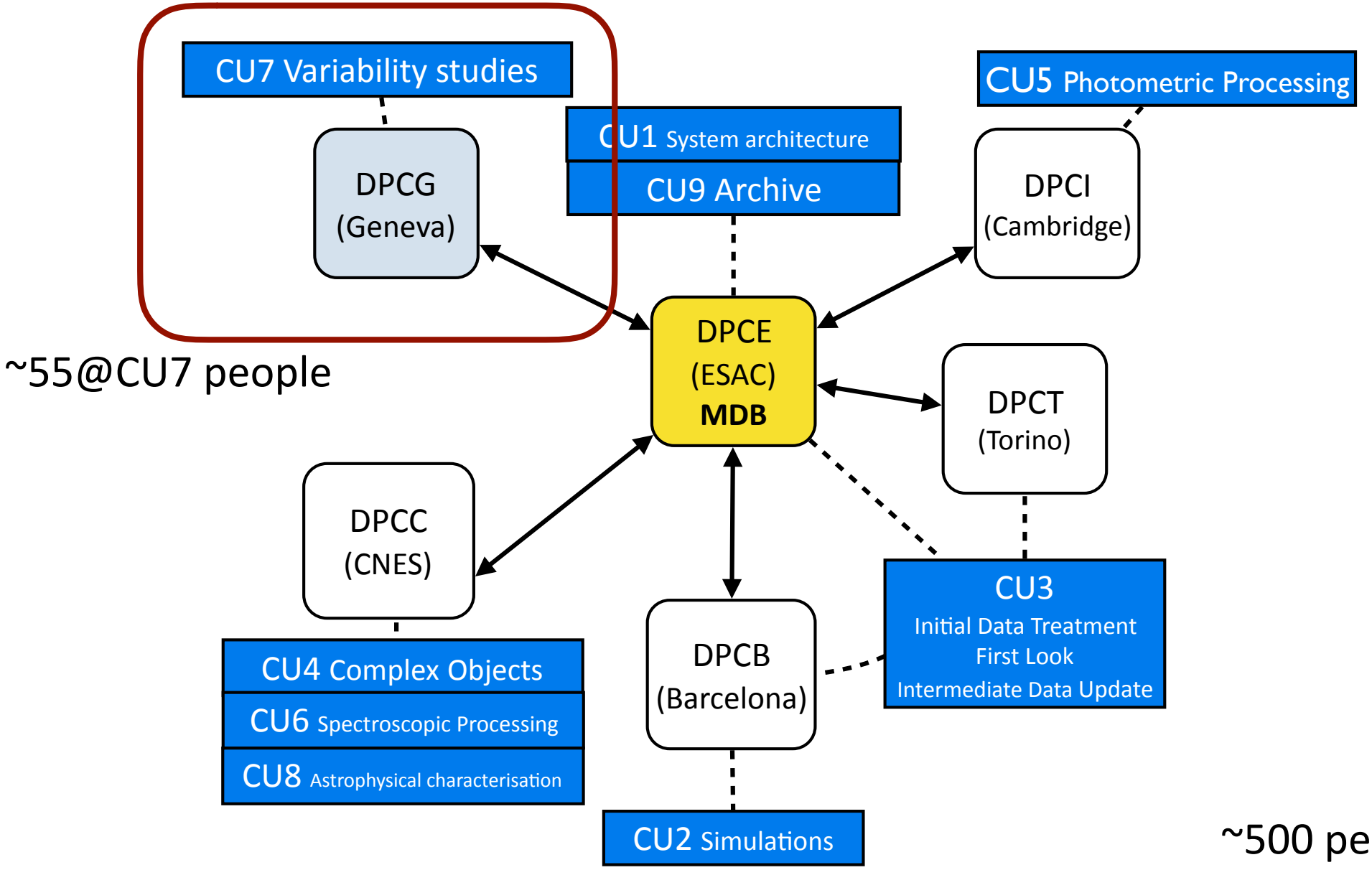
CU's process data in Data Processing centre(s):



~500 people

Gaia Consortium - distributed challenge

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



Approach for the variability analysis

Approach for the variability analysis: Iterative and pragmatic

Approach for the variability analysis: Iterative and pragmatic

High amplitude
Regular objects
Smaller list of Attributes

Approach for the variability analysis: Iterative and pragmatic

High amplitude
Regular objects
Smaller list of Attributes

Cepheids
RR Lyrae stars
+…

Approach for the variability analysis: Iterative and pragmatic

High amplitude
Regular objects
Smaller list of Attributes

Cepheids
RR Lyrae stars
+…

G FoV photometry

Approach for the variability analysis: Iterative and pragmatic

High amplitude
Regular objects
Smaller list of Attributes

Cepheids
RR Lyrae stars
+…

G FoV photometry

Approach for the variability analysis: Iterative and pragmatic

Small amplitudes
Complicated signals

High amplitude
Regular objects
Smaller list of Attributes

Cepheids
RR Lyrae stars
+...

G FoV photometry

Approach for the variability analysis: Iterative and pragmatic

High amplitude
Regular objects
Smaller list of Attributes

Small amplitudes
Complicated signals

Exo-planetary transits
Short time scale
+ …

Cepheids
RR Lyrae stars
+…

G FoV photometry

Approach for the variability analysis: Iterative and pragmatic

High amplitude
Regular objects
Smaller list of Attributes

Small amplitudes
Complicated signals

Exo-planetary transits
Short time scale
+ …

Cepheids
RR Lyrae stars
+…

Per-CCD data

G FoV photometry

At the Gaia precision not all stars are variable

In the Galaxy, variable objects represent about few percent to 15% of the population

At the Gaia precision not all stars are variable

In the Galaxy, variable objects represent about few percent to 15% of the population

10s to 150 million sources to analyse

At the Gaia precision not all stars are variable

In the Galaxy, variable objects represent about few percent to 15% of the population

10s to 150 million sources to analyse

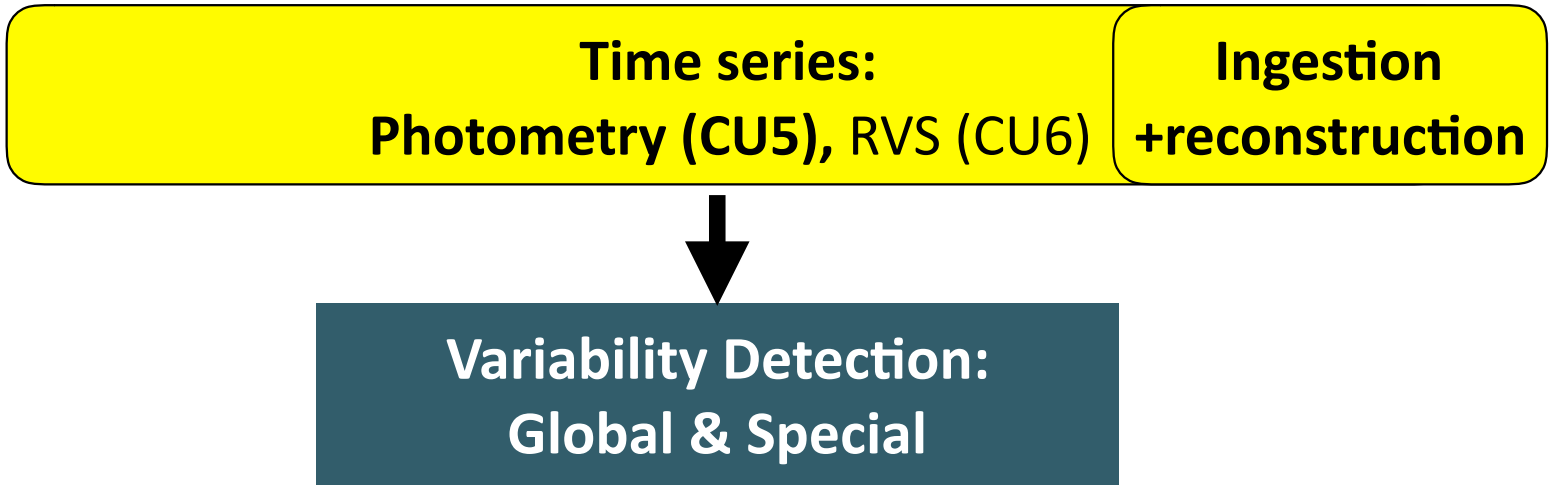
Reduce your problem by one order of magnitude!

Variability Processing and Analysis: A global, comprehensive approach

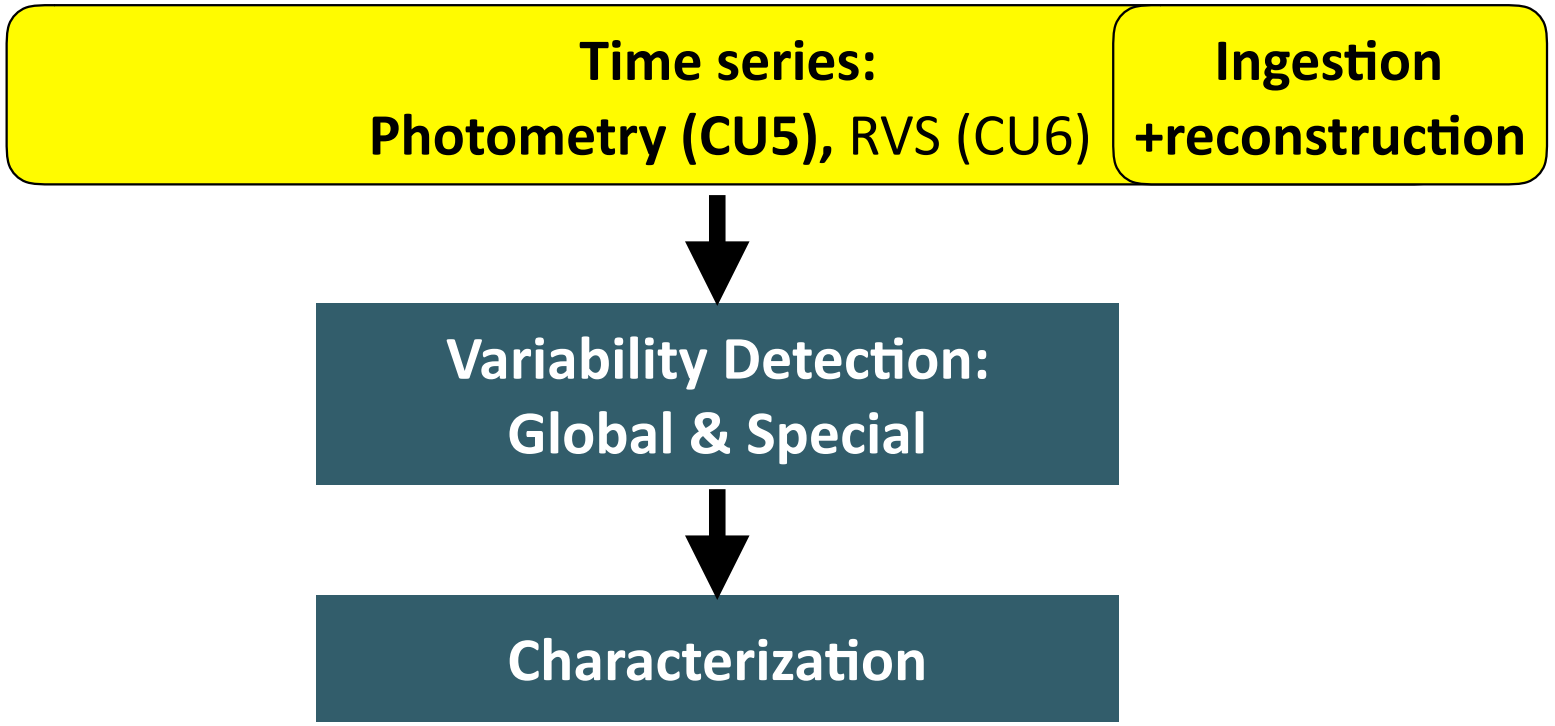
Time series:
Photometry (CU5), RVS (CU6)

Ingestion +reconstruction

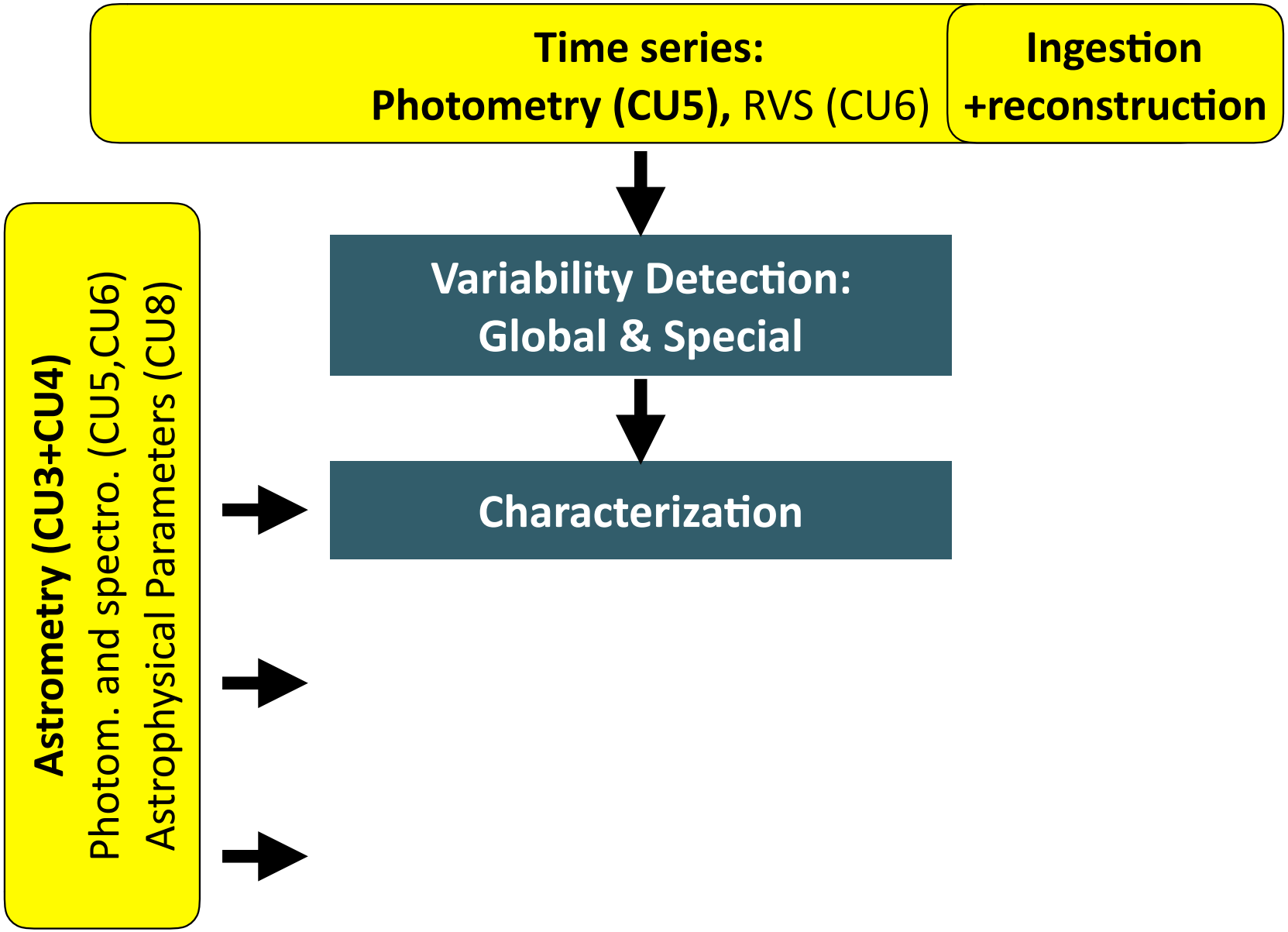
Variability Processing and Analysis: A global, comprehensive approach



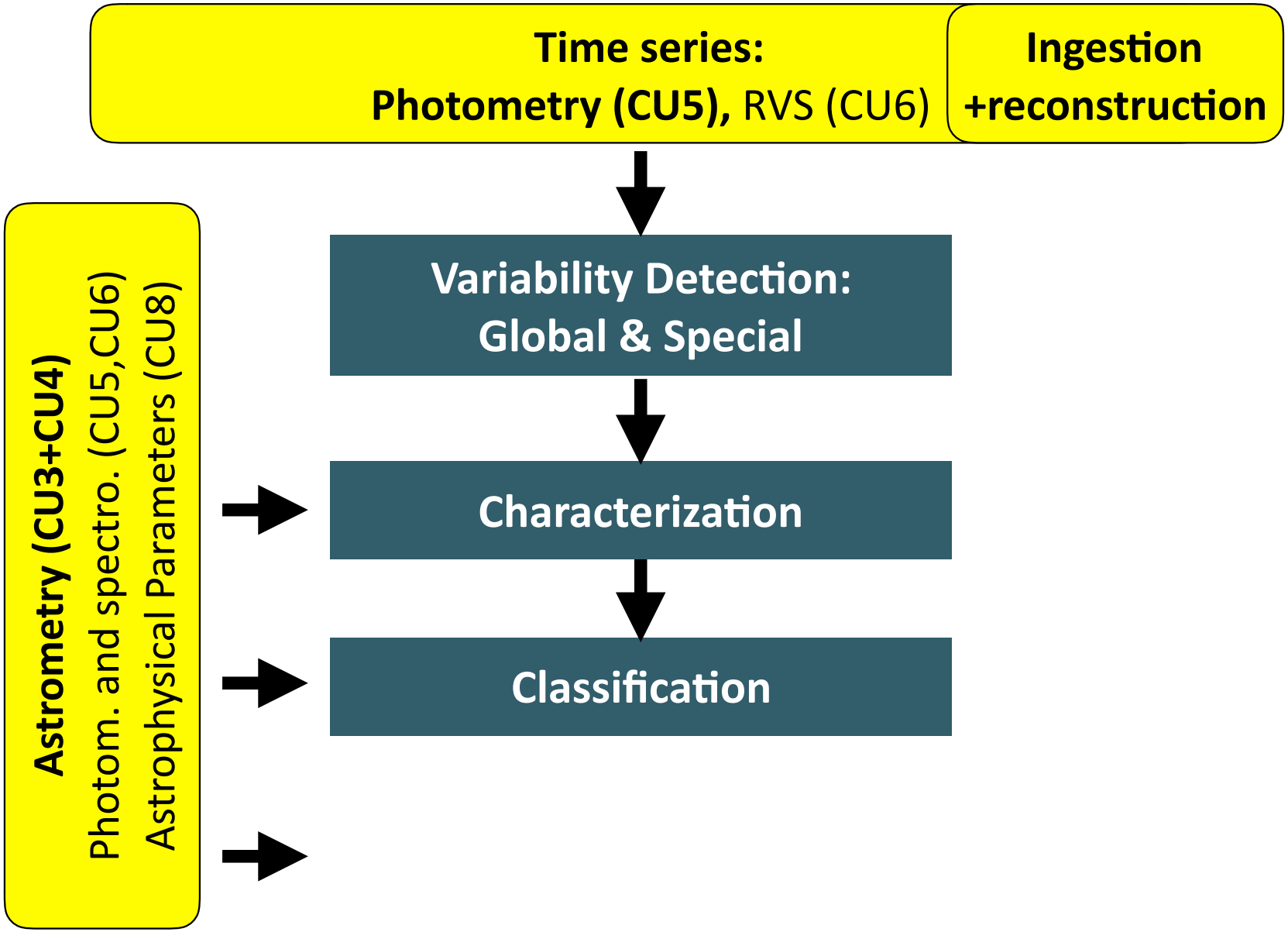
Variability Processing and Analysis: A global, comprehensive approach



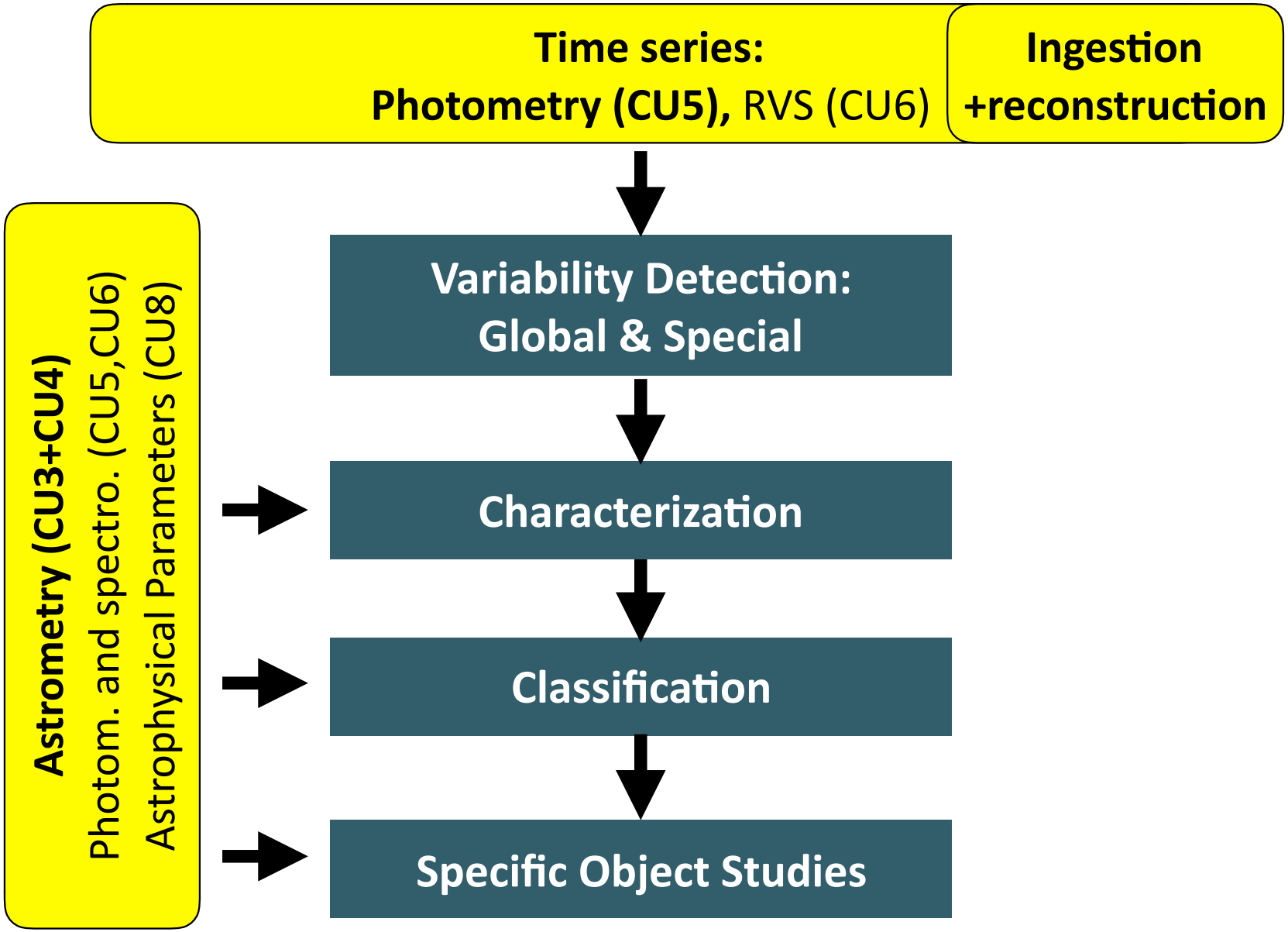
Variability Processing and Analysis: A global, comprehensive approach



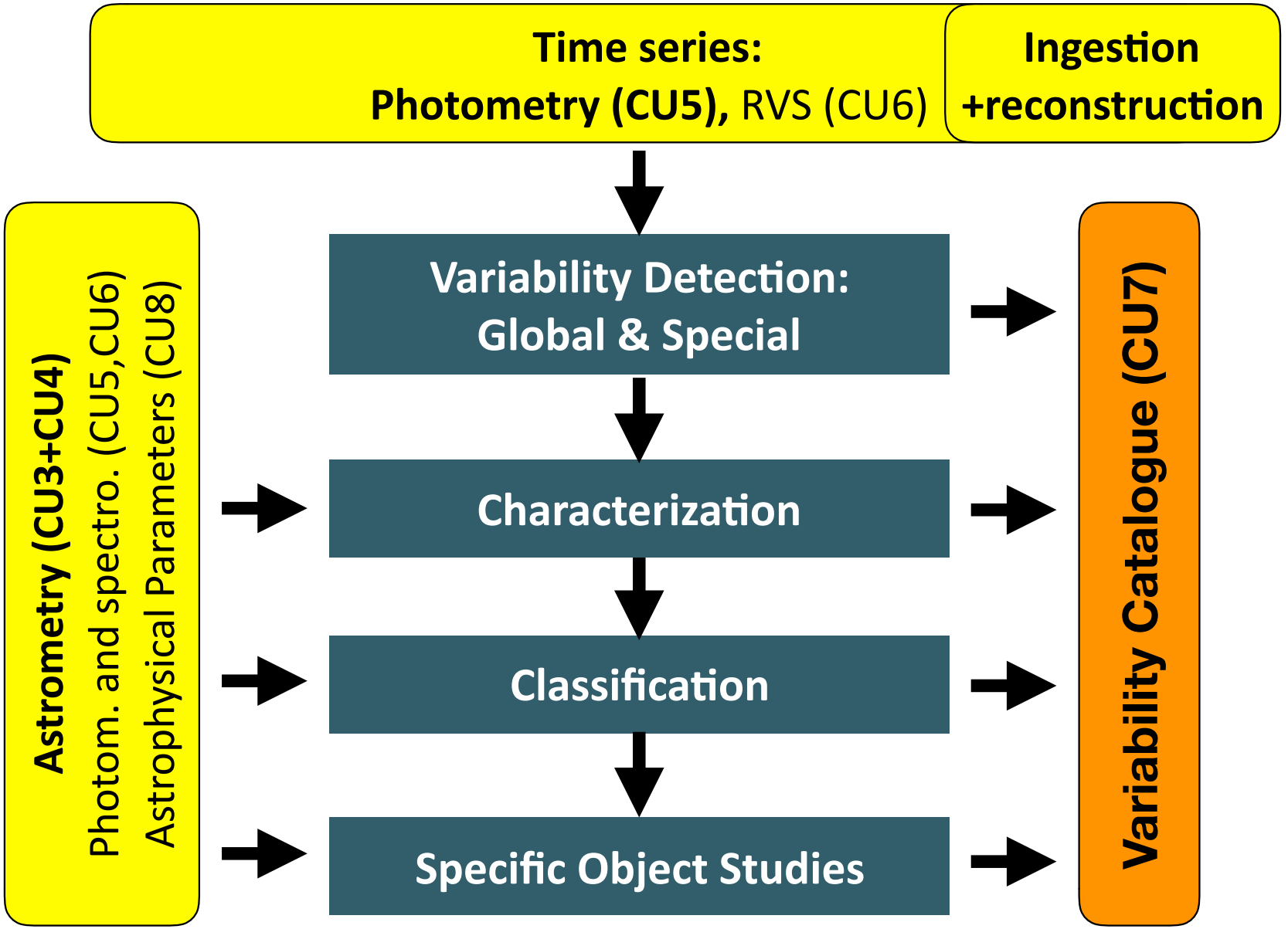
Variability Processing and Analysis: A global, comprehensive approach



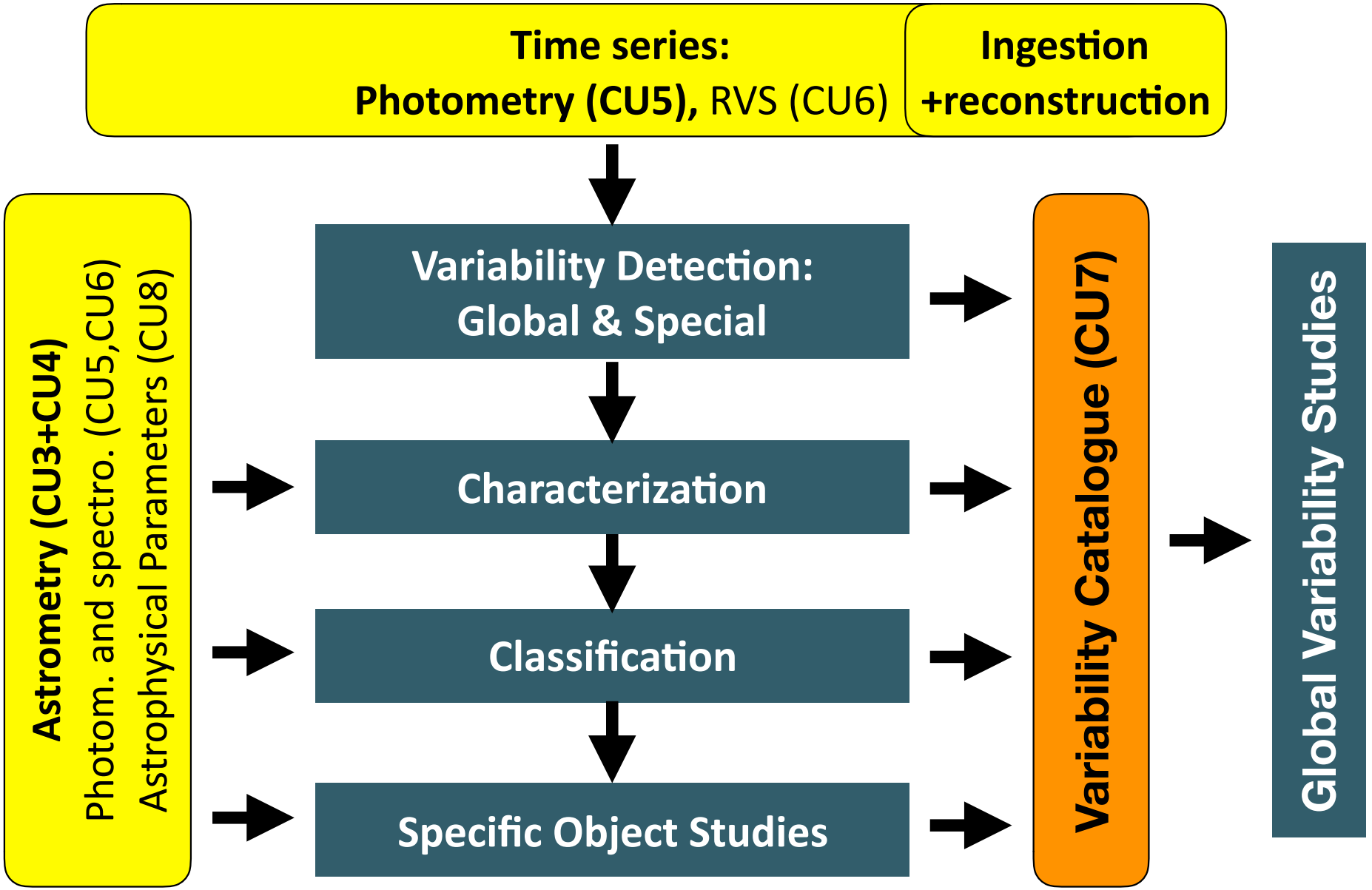
Variability Processing and Analysis: A global, comprehensive approach



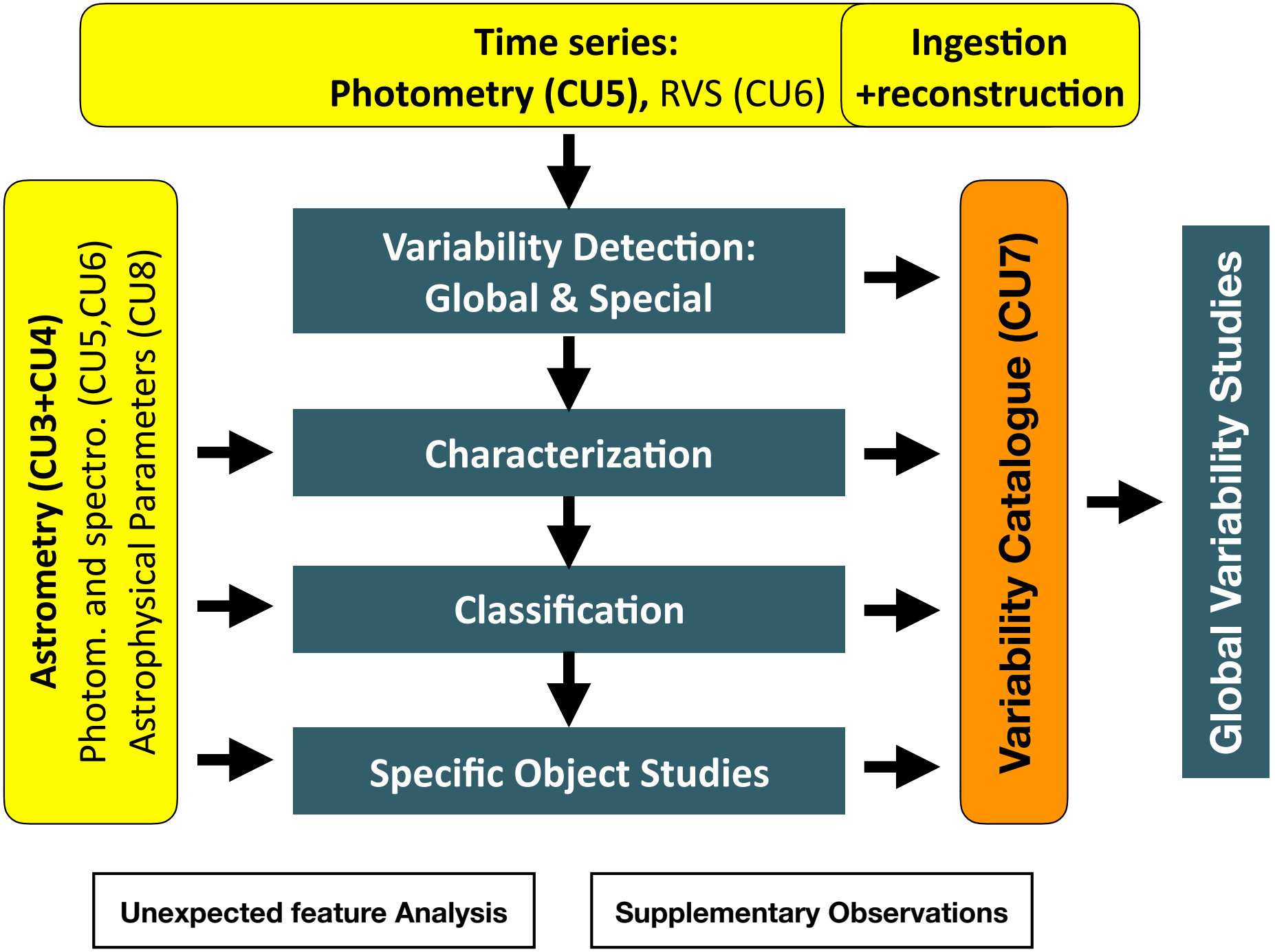
Variability Processing and Analysis: A global, comprehensive approach



Variability Processing and Analysis: A global, comprehensive approach



Variability Processing and Analysis: A global, comprehensive approach



Methods: General Variability Detection

Classical method

Hypothesis testing:

Null hypothesis

Alternative hypothesis

Use some statistics e.g. Chi2

Methods: General Variability Detection

Classical method

Hypothesis testing:

Null hypothesis

Alternative hypothesis

Use some statistics e.g. Chi2

P-value

Methods: General Variability Detection

Classical method

Hypothesis testing:

Null hypothesis

Alternative hypothesis

Use some statistics e.g. Chi2

P-value

— allows us to compare stars with different precision, different number of measurements

Methods: General Variability Detection

Classical method

Hypothesis testing:

Null hypothesis

Alternative hypothesis

Use some statistics e.g. Chi2

P-value

— allows us to compare stars with different precision, different number of measurements

— allows us also to rescale empirically the uncertainties

Methods: General Variability Detection

Classical method

Hypothesis testing:

Null hypothesis

Alternative hypothesis

Use some statistics e.g. Chi2

HOWEVER

P-value

— allows us to compare stars with different precision, different number of measurements

— allows us also to rescale empirically the uncertainties

Methods: General Variability Detection

Classical method

Hypothesis testing:

Null hypothesis

Alternative hypothesis

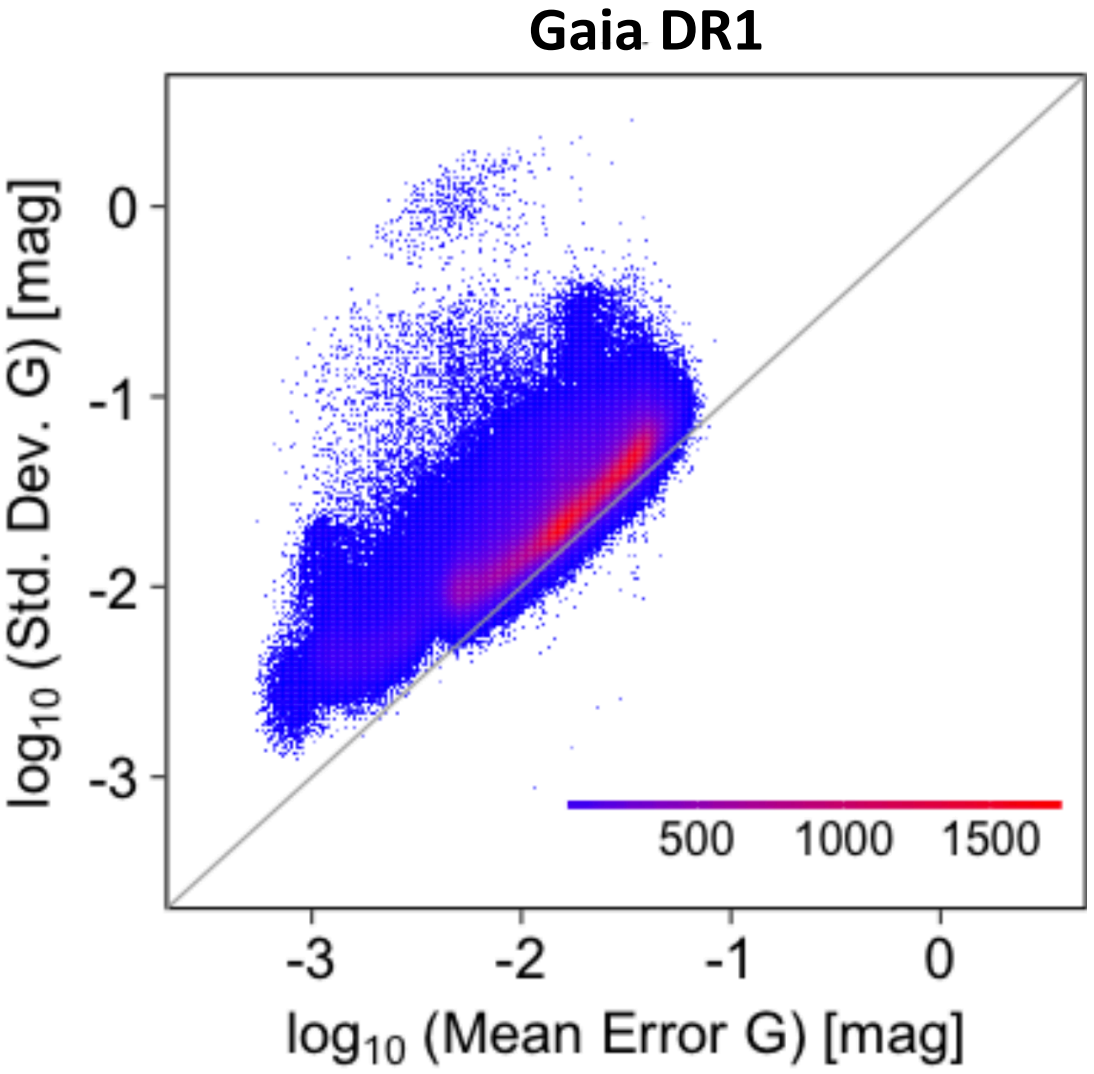
Use some statistics e.g. Chi2

P-value

— allows us to compare stars with different precision, different number of measurements

— allows us also to rescale empirically the uncertainties

HOWEVER



see Eyer et al. 2017

General Variability Detection

Alternative: use a supervised classification scheme (e.g. Gaia Data Release 1)

Attributes: General + those derived from time series

Algorithm: Random Forest (Breiman 2001)

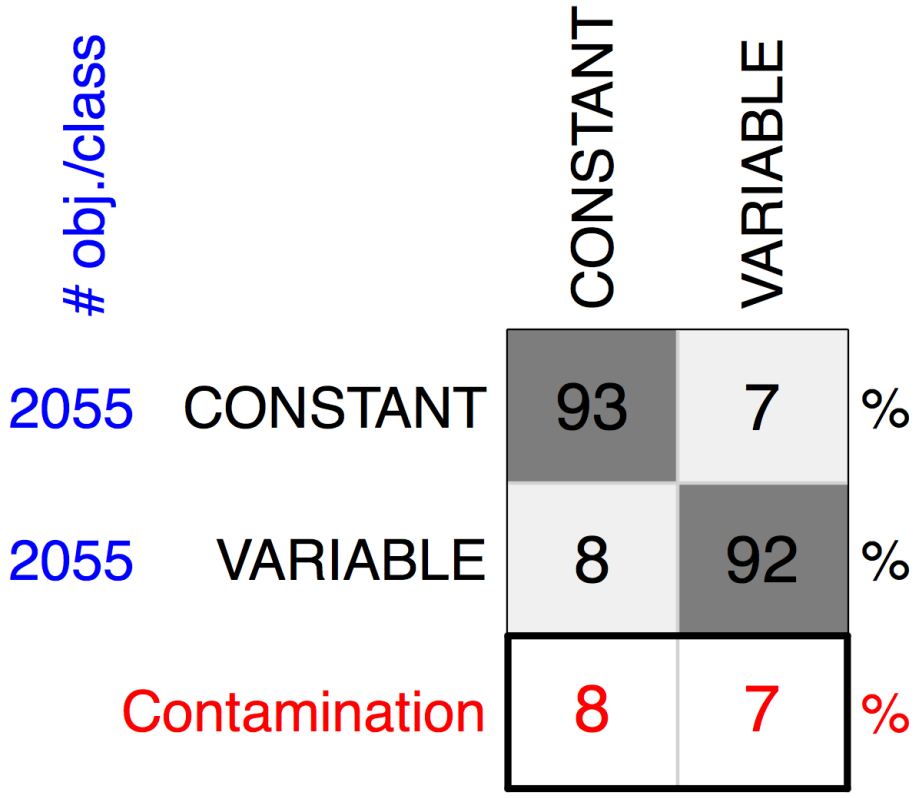
Training-set: based on OGLE

Result of the classification

Completeness

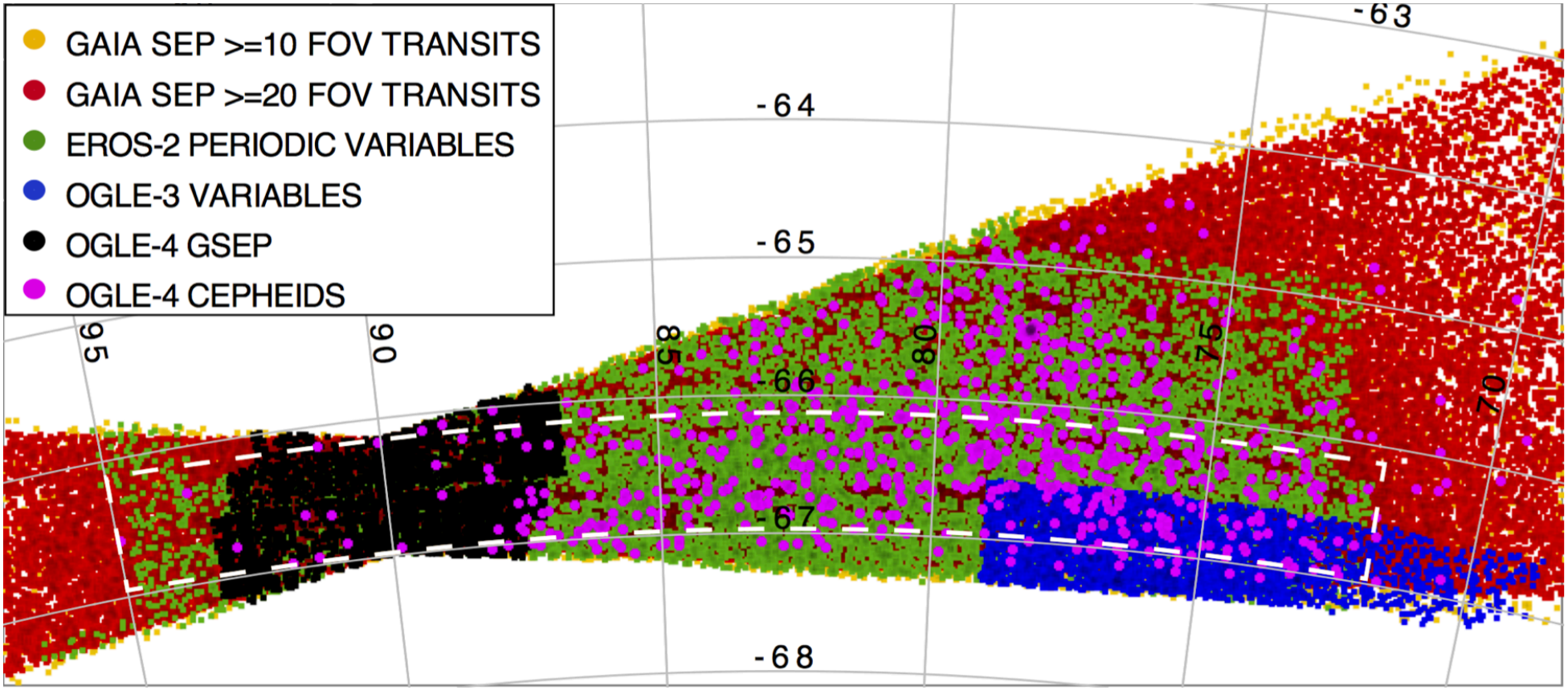
Contamination

Confusion Matrix:



Eyer et al. 2017

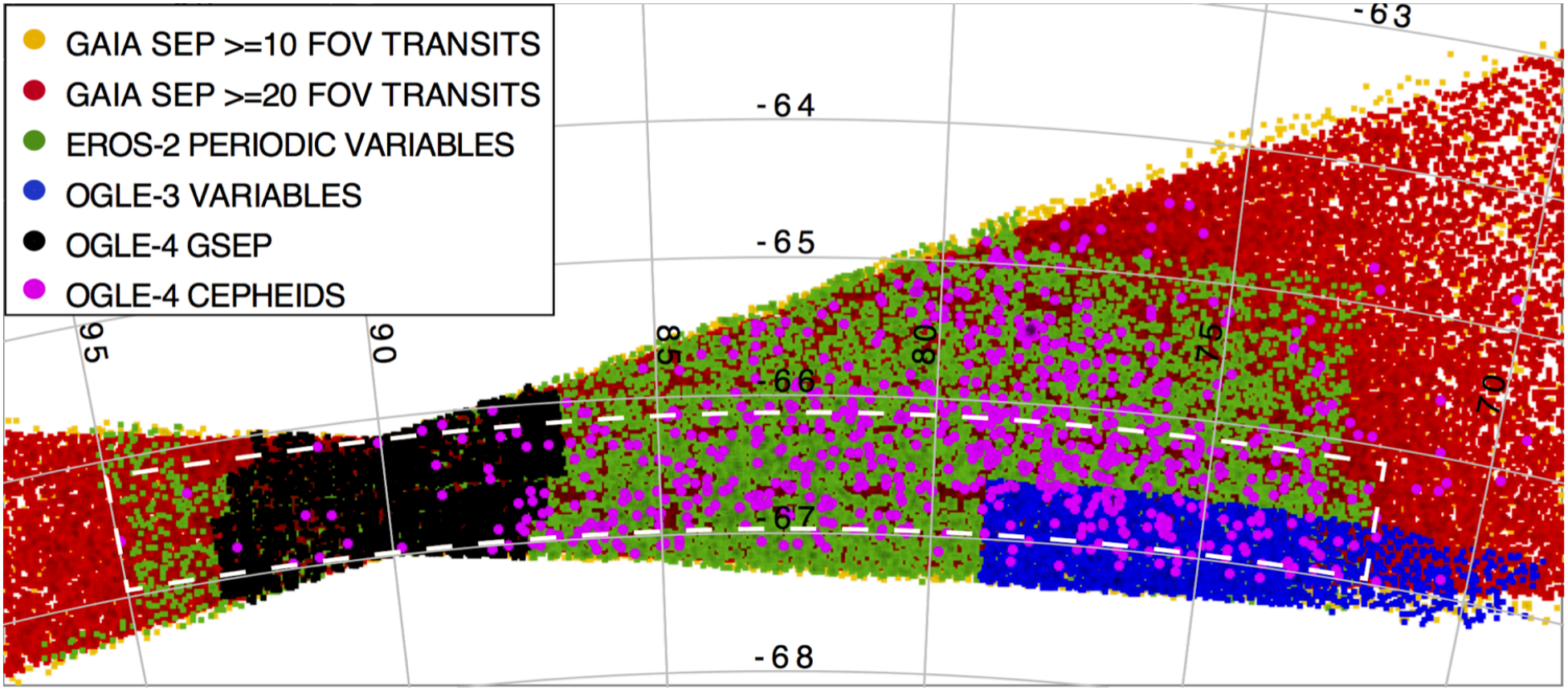
Region for the Gaia Data Release 1: well covered by OGLE



Gaia DR1 footprint in Equatorial Coordinates

Eyer et al. 2017

Region for the Gaia Data Release 1: well covered by OGLE

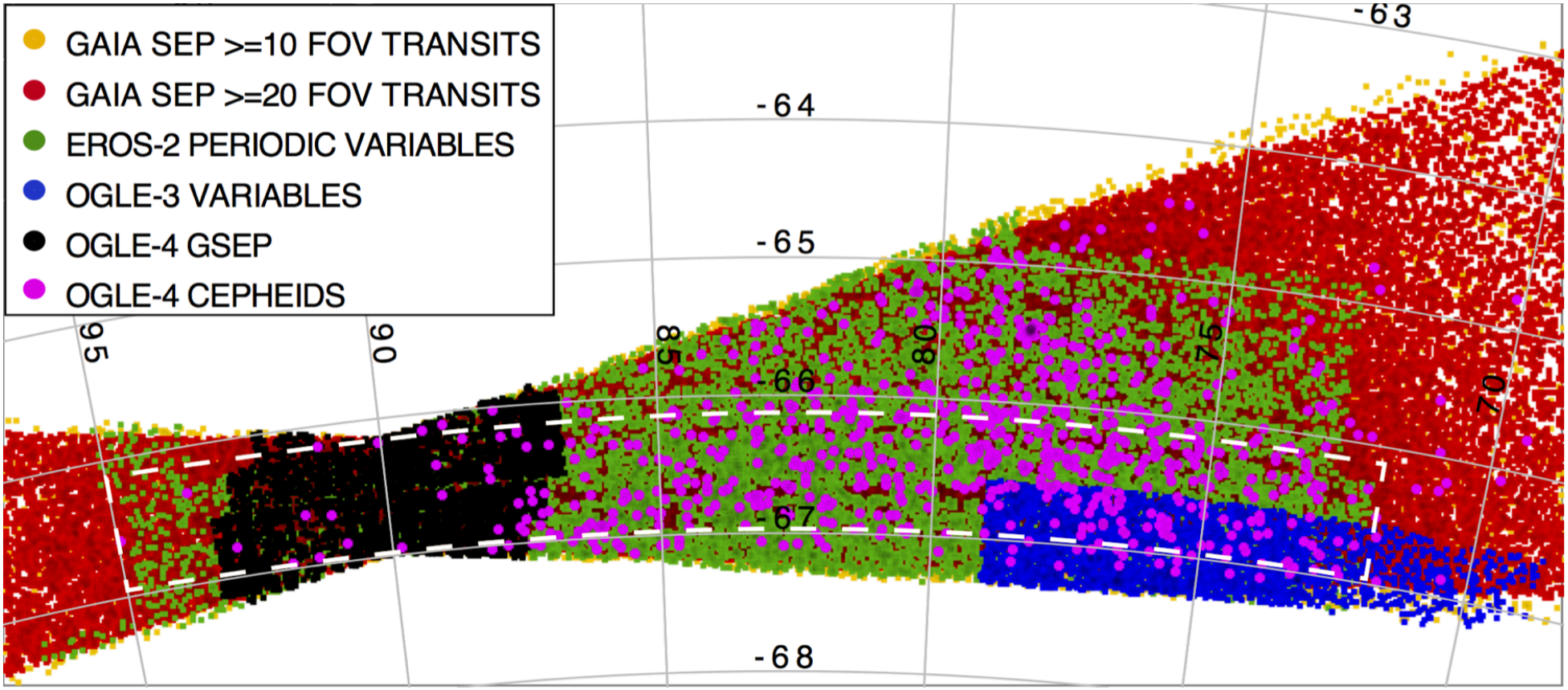


Gaia DR1 footprint in Equatorial Coordinates

Eyer et al. 2017

HOWEVER

Region for the Gaia Data Release 1: well covered by OGLE



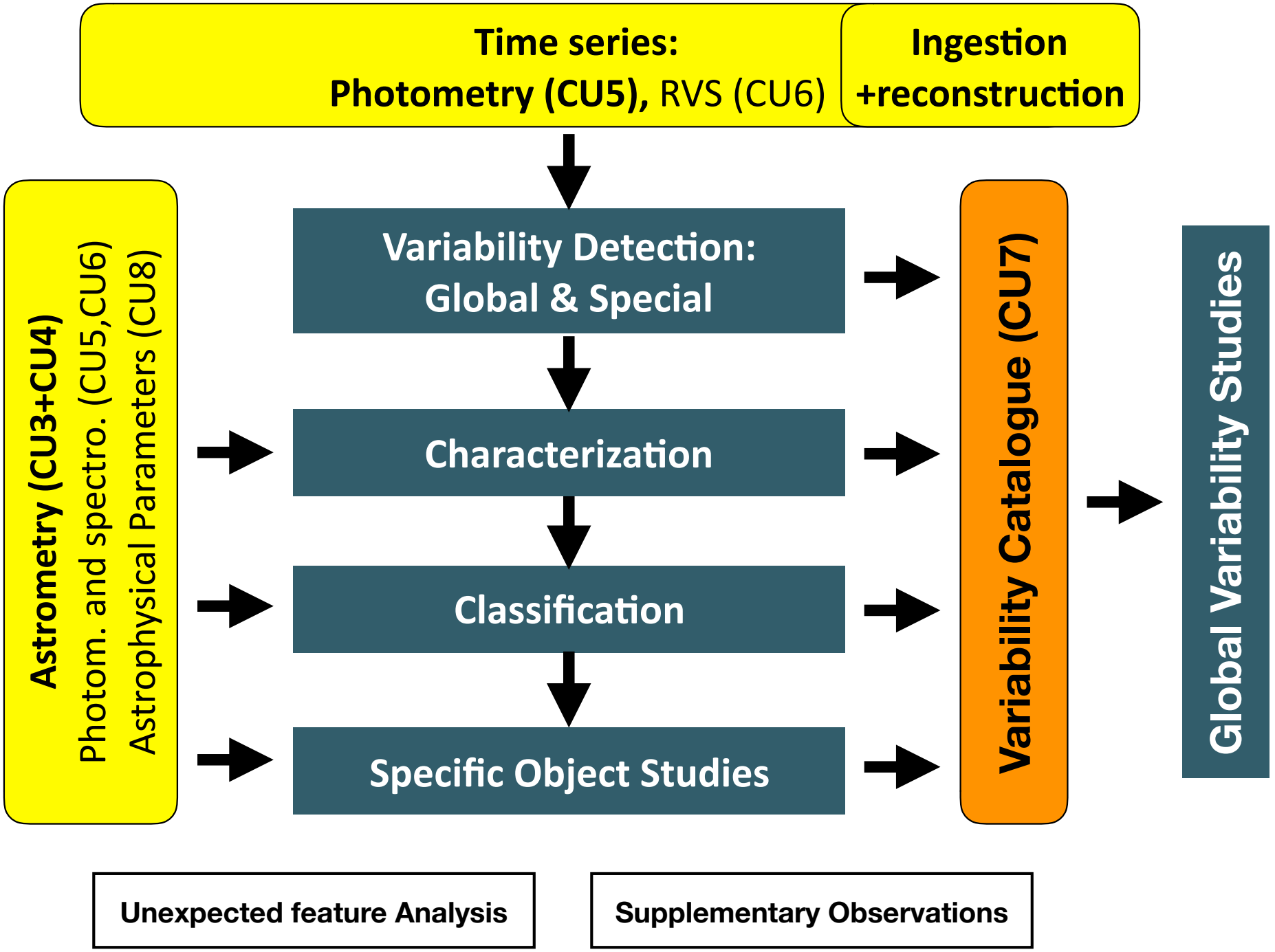
Gaia DR1 footprint in Equatorial Coordinates

Eyer et al. 2017

HOWEVER

Gaia Data Release 2 is for the whole sky, no training-set is really representative

Variability Processing and Analysis: A global, comprehensive approach



Classification: Iterative approach (semi-supervised) for Gaia Data Release 2

An obvious fact: the classification depends critically on the training-set

Iterative work to improve (generalise) the training-set

a. Coverage and distribution of training-set similar to real data

— Sky position, number of measurements, magnitude (signal/noise)

b. Relative fraction of variables in the training-set representative of the true one (e.g. this is true for Random Forest)

Classification: Several classifiers

Supervised classification (several methods):

Classification: Several classifiers

Supervised classification (several methods):

Multistage tree:
Bayesian networks

Multistage tree:
Gaussian mixture

Random Forest

Classification: Several classifiers

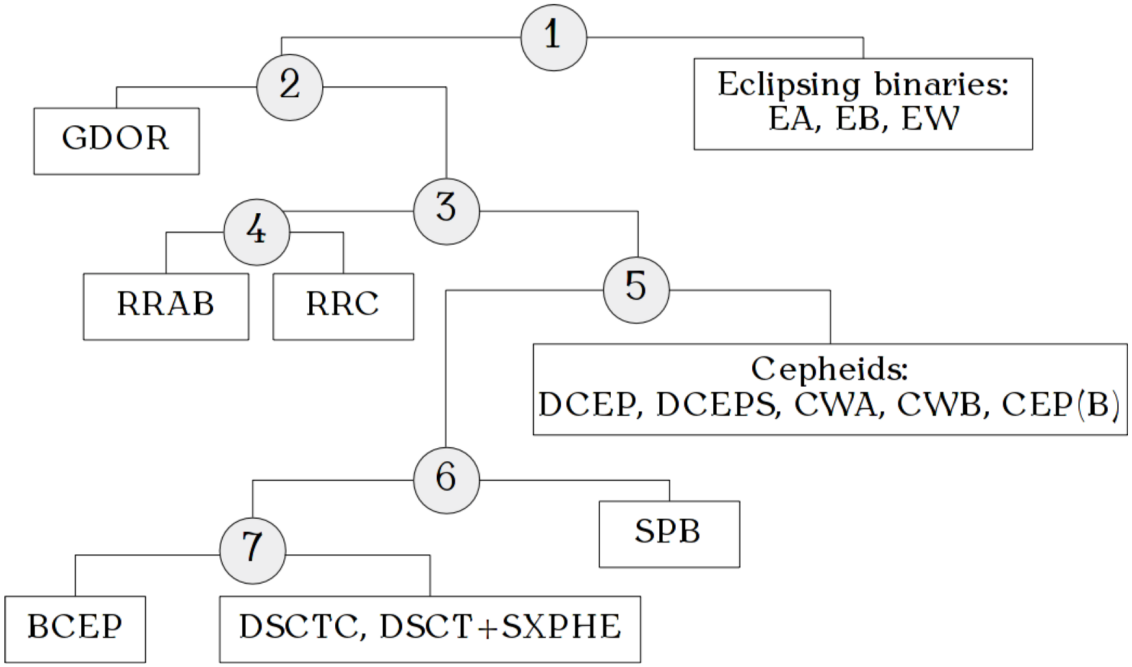
Supervised classification (several methods):

Multistage tree: Bayesian networks

Multistage tree: Gaussian mixture

Random Forest

Multi stage tree example



Classification: Several classifiers

Supervised classification (several methods):

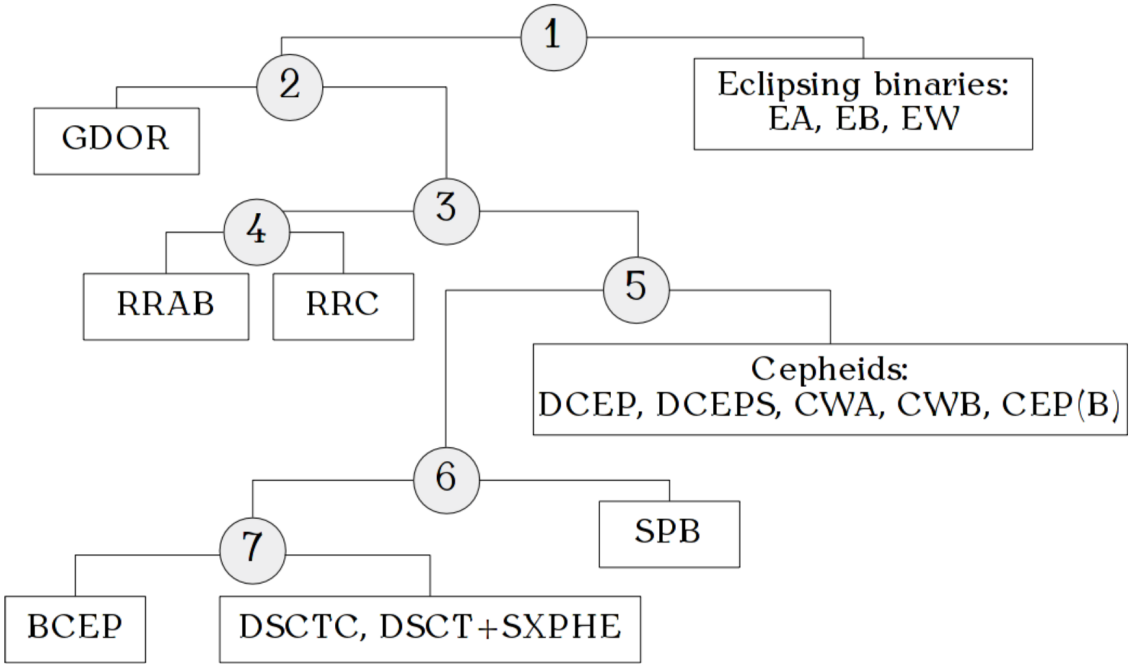
Multistage tree:
Bayesian networks

Multistage tree:
Gaussian mixture

Random Forest

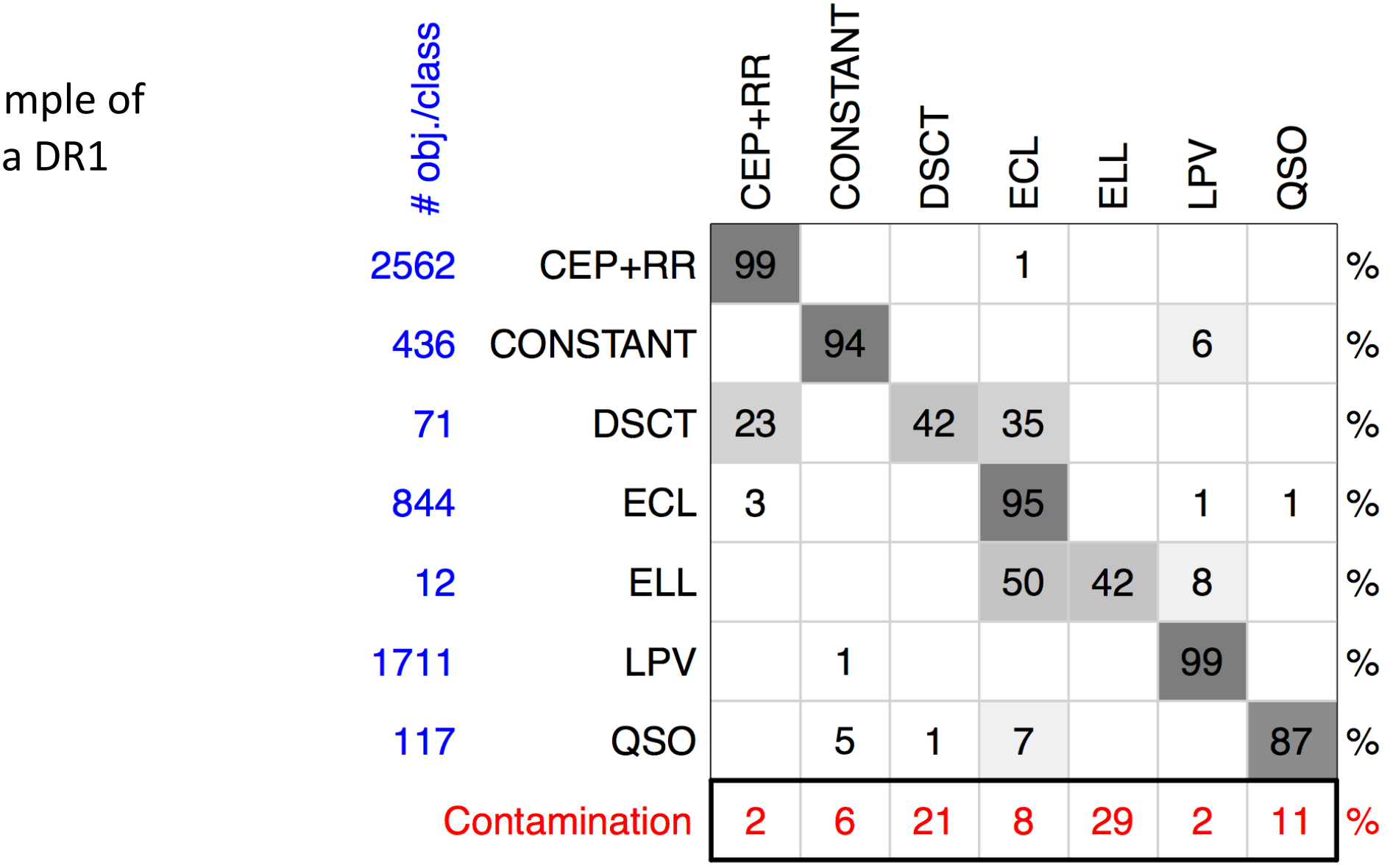
Furnish training-set
built from Crossmatched data

Multi stage tree example



Classification

Example of
Gaia DR1

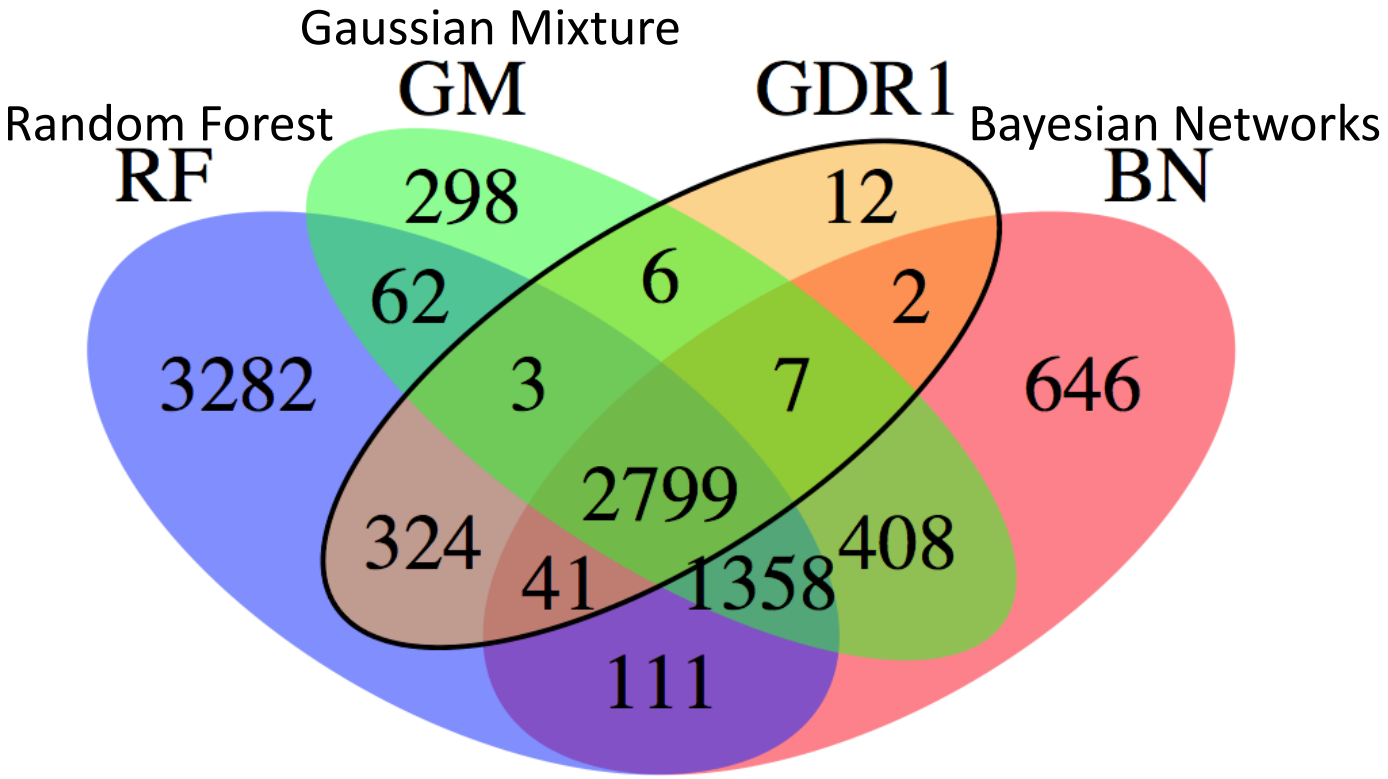


Random Forest

Eyer et al. 2017

Classification: Several classifiers

Venn Diagram of the different classifiers
of Cepheids and RR Lyrae stars for the Gaia DR1



Eyer et al. 2017

Improve the classification: Meta-classifier

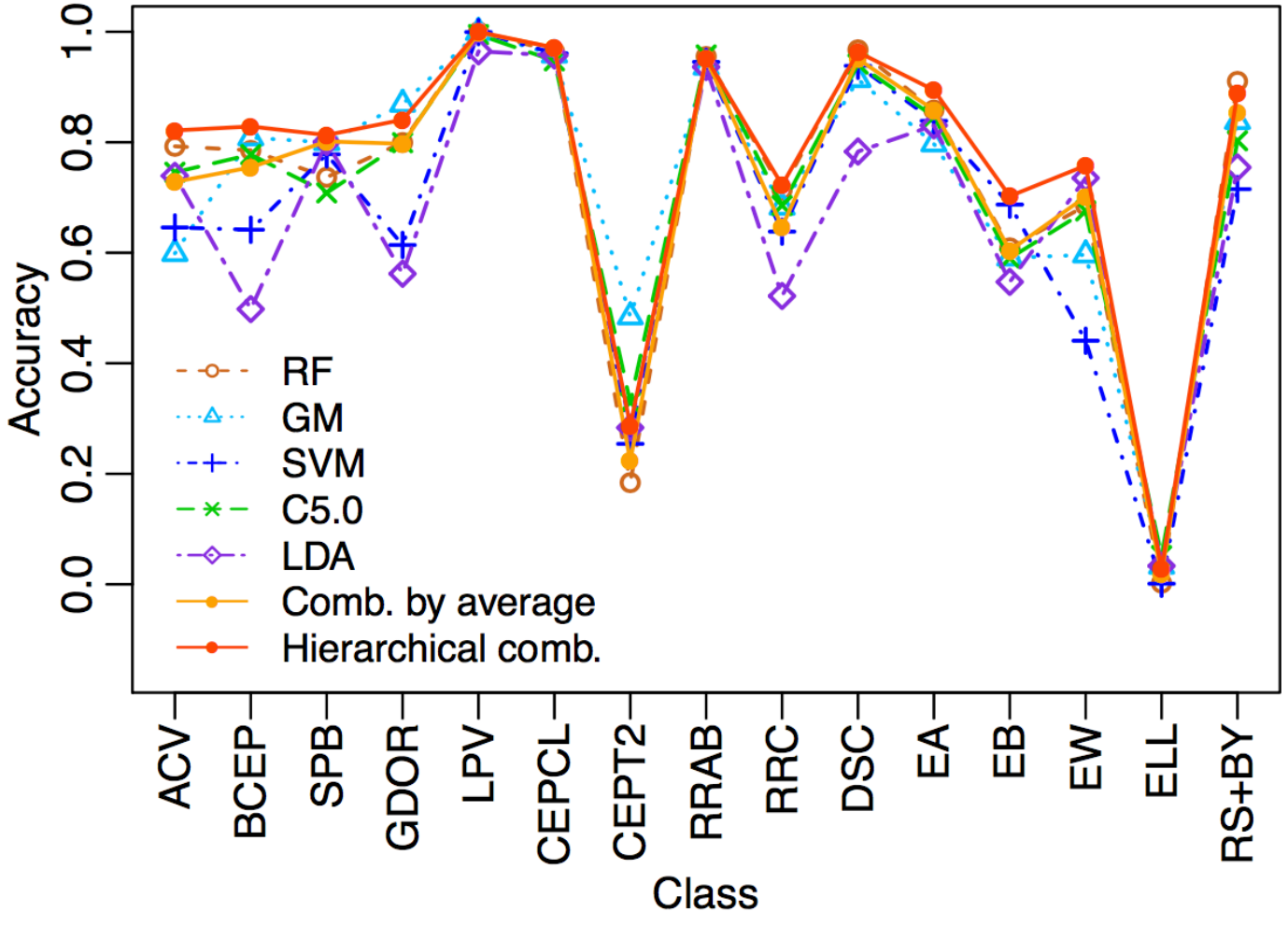
Combination of several classifiers: a two level hierarchy learners

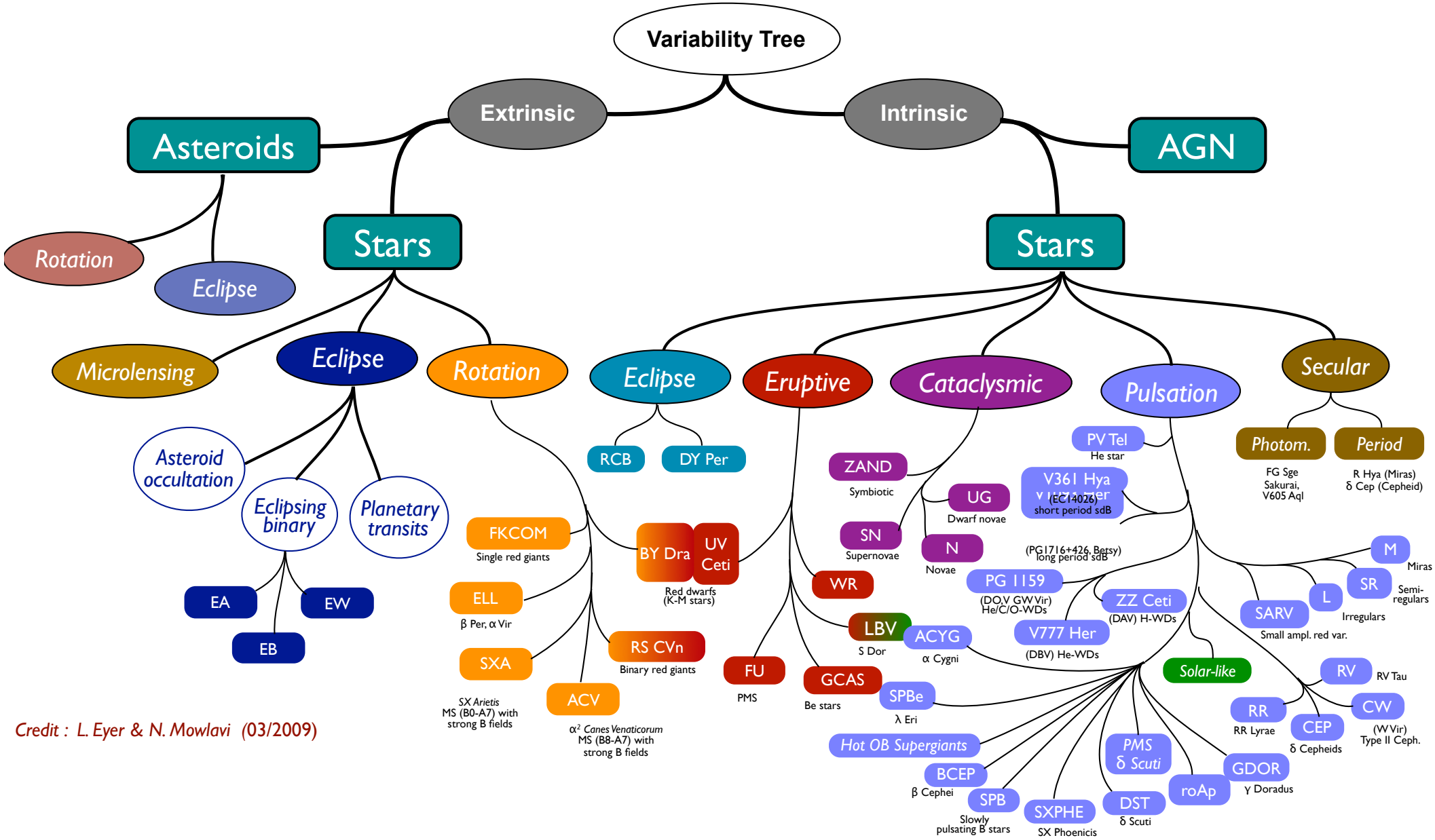
Tests were done on Hipparcos by Süveges et al. 2017

Improve the classification: Meta-classifier

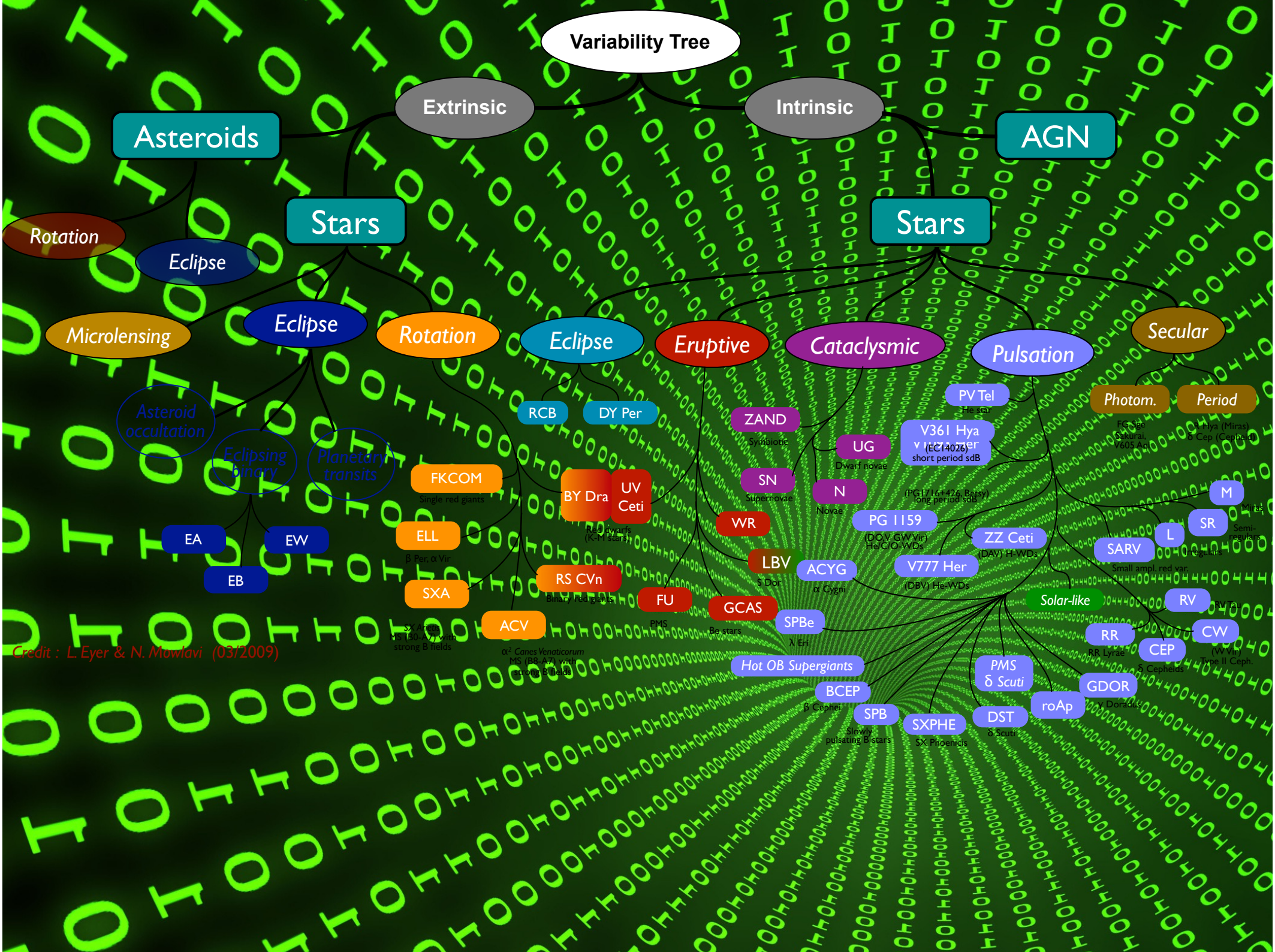
Combination of several classifiers: a two level hierarchy learners

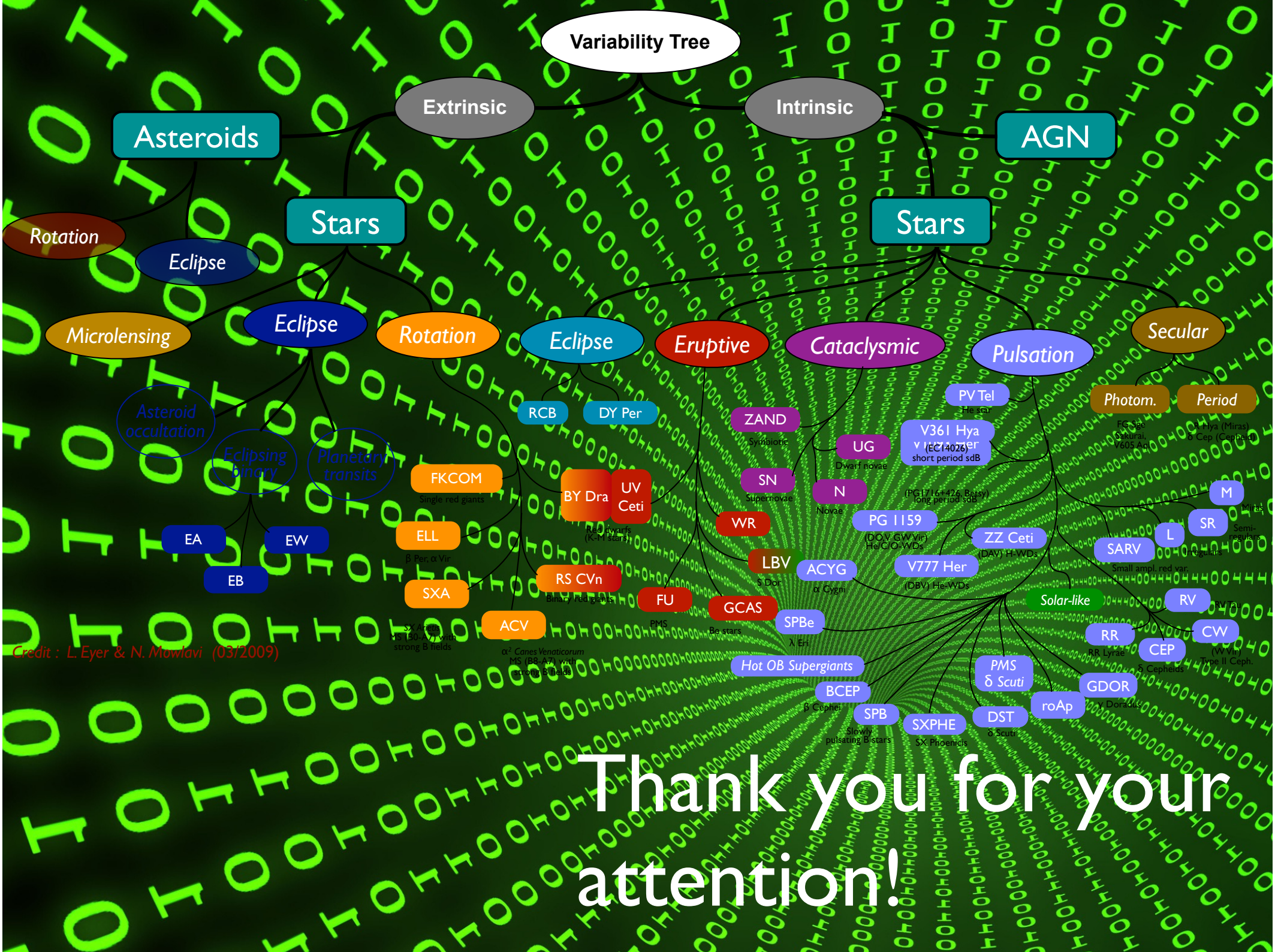
Tests were done on Hipparcos by Süveges et al. 2017





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


Variability Tree
Extrinsic
Intrinsic
Asteroids
AGN
Stars
Stars
Rotation
Eclipse
Microlensing
Eclipse
Rotation
Eclipse
Eruptive
Cataclysmic
Pulsation
Secular
Asteroid occultation
Eclipsing binary
Planetary transits
EA
EW
EB
FKCOM
Single red giants
ELL
β Per, α Vir
SXA
ACV
RCB
DY Per
BY Dra
UV Ceti
RS CVn
Binary Red giants
FU
PMS
GCAS
Be stars
SPBe
λ Eri
ZAND
Symbiotic
SN
Supernovae
WR
LBV
S Dor
ACYG
α Cygni
UG
Dwarf novae
N
Novae
PG 1159
(DO,V,GW Vir)
He/C/O-WDs
PV Tel
He star
V361 Hya
(EC14026)
short period sdB
(PG1716+426 Betsy)
long period sdB
ZZ Ceti
(DAV) H-WDs
V777 Her
(DBV) He-WDs
Photom.
Period
M
Mira
SR
Semi-regulars
L
SARV
Small ampl. red var.
RV
RV Tau
CW
(W Vir)
Type II Ceph.
RR
RR Lyrae
CEP
δ Cepheids
GDOR
γ Doradus
roAp
DST
δ Scuti
SXPHE
SX Phoenicis
PMS
δ Scuti
Solar-like
Hot OB Supergiants
BCEP
β Cephei
SPB
Slowly pulsating B-stars
Credit : L. Eyer & N. Mowlavi (03/2009)



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

Thank you for your attention!