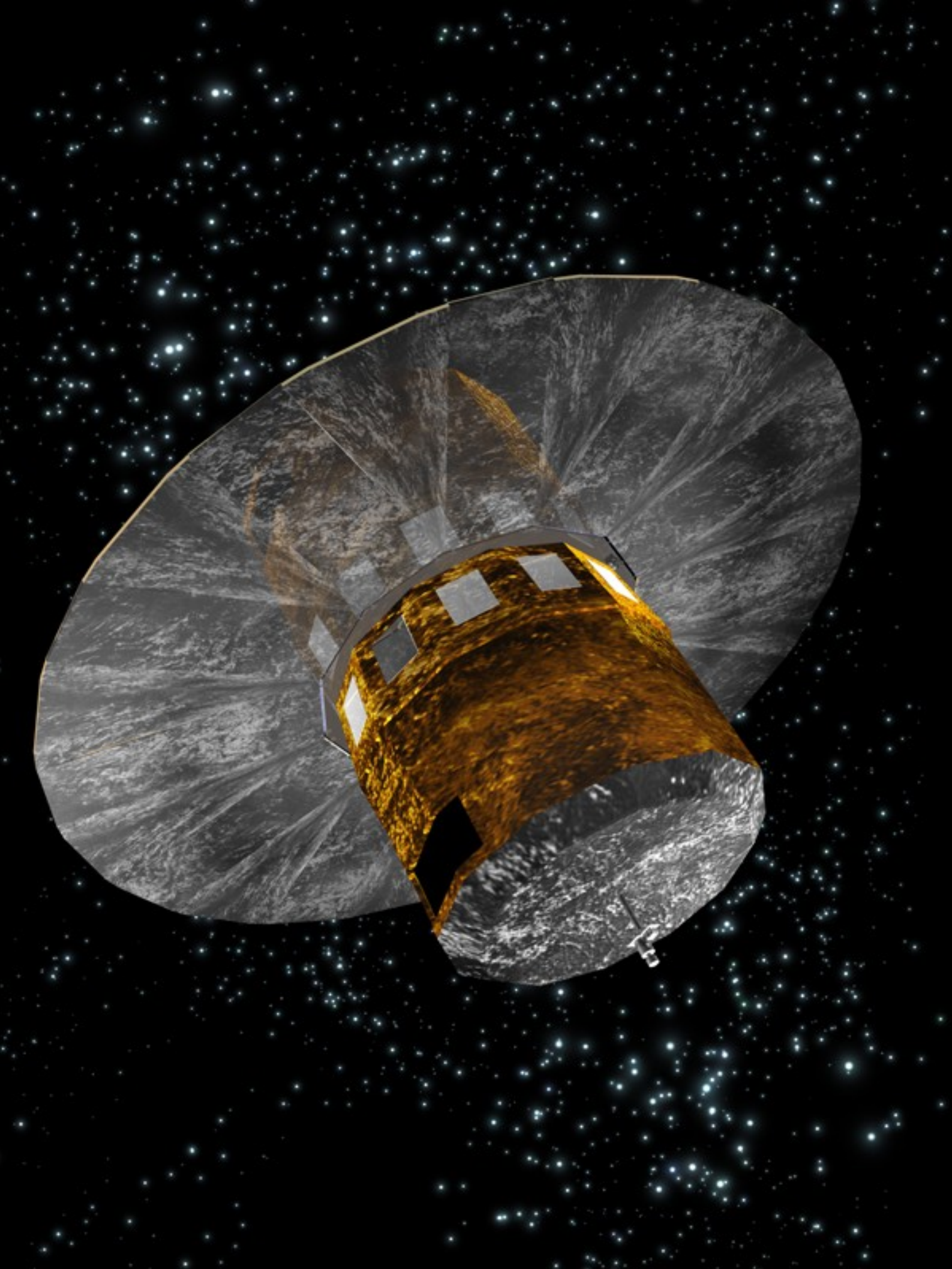
Gaia - Be Stars Classification

Jan Soldan

Astronomical Institute
251 65 Ondrejov
Czech Republic

Gaia satellite

- **Survey satellite** for very precise **astrometry**, **photometry and low / medium resolution spectroscopy**.
- Launch 2011, 5-year lifetime, **Gaia will observe each object (down to 20th mag.) around 50-80 times**.
- Planned observation of **10^9 diverse objects**.

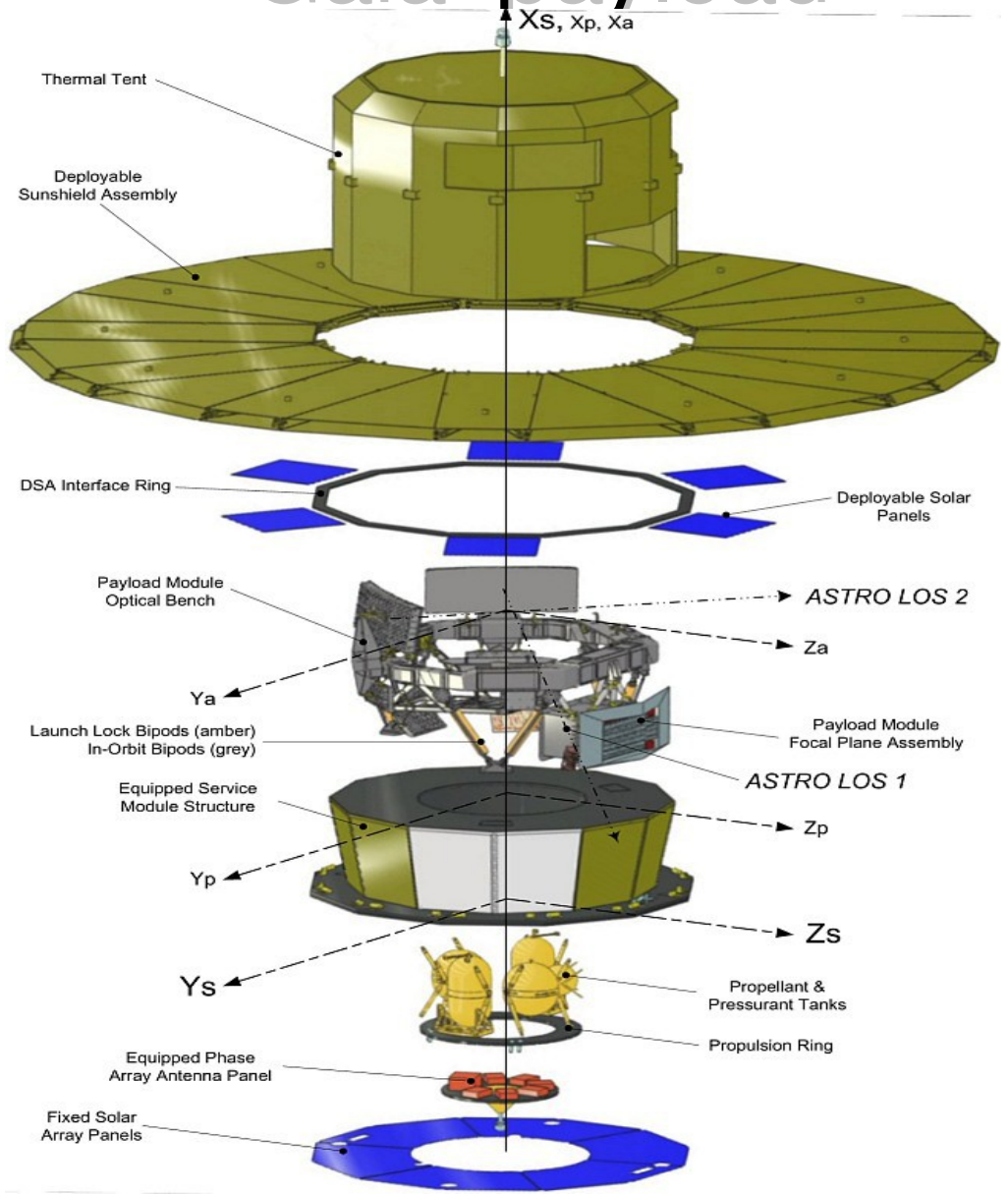


Gaia payload

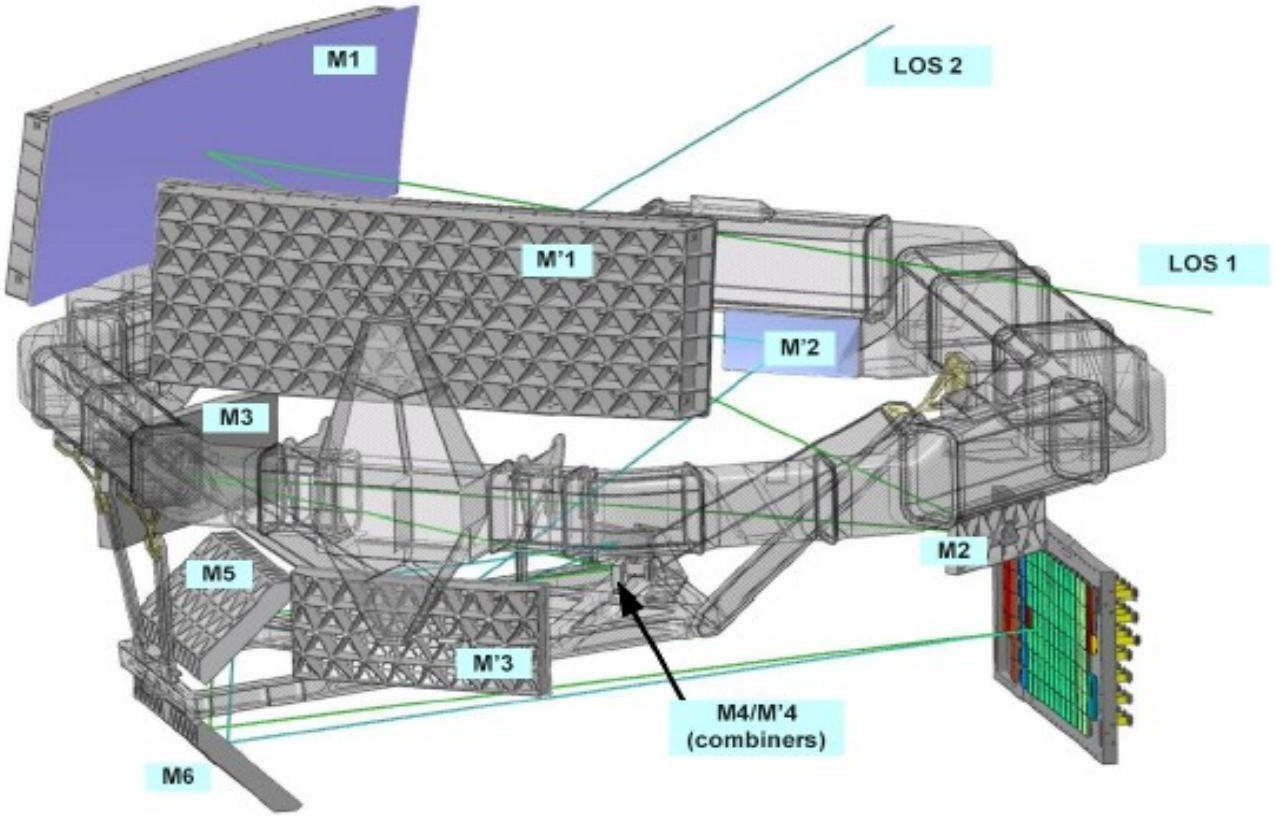
Instr. module

Telescope
Instruments

Service module

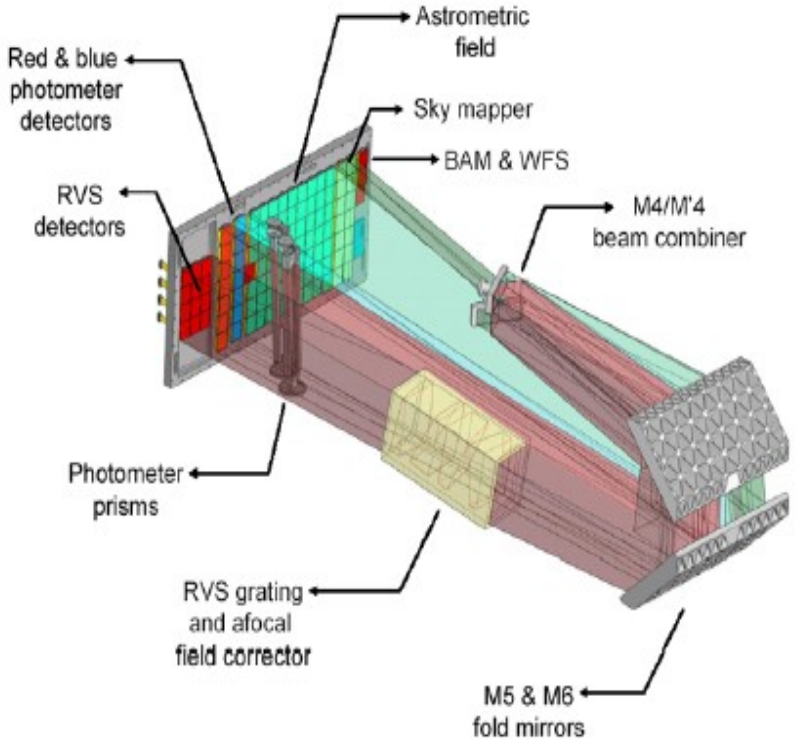


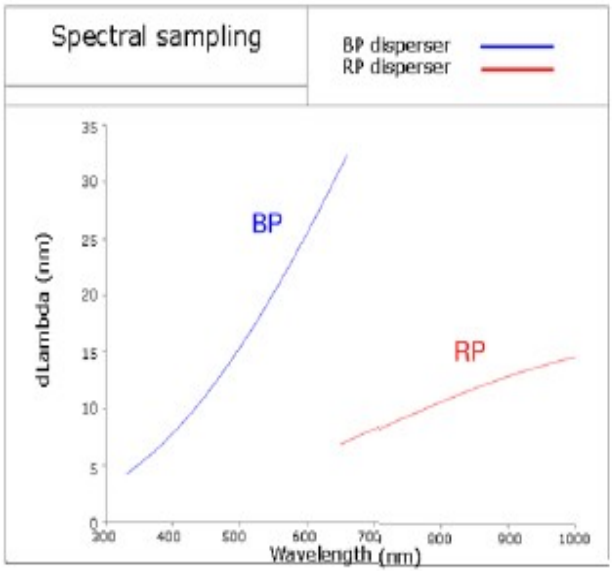
Dual telescope system



Main mirrors: 2 x 1.45 x 0.5 m (ekv. 0.98 m),
f = 35 m, scale 0.170 mm.arcsec^{-1} , FOV 0.7 x 0.7°

Gaia astrometry, photometry and spectroscopy instruments



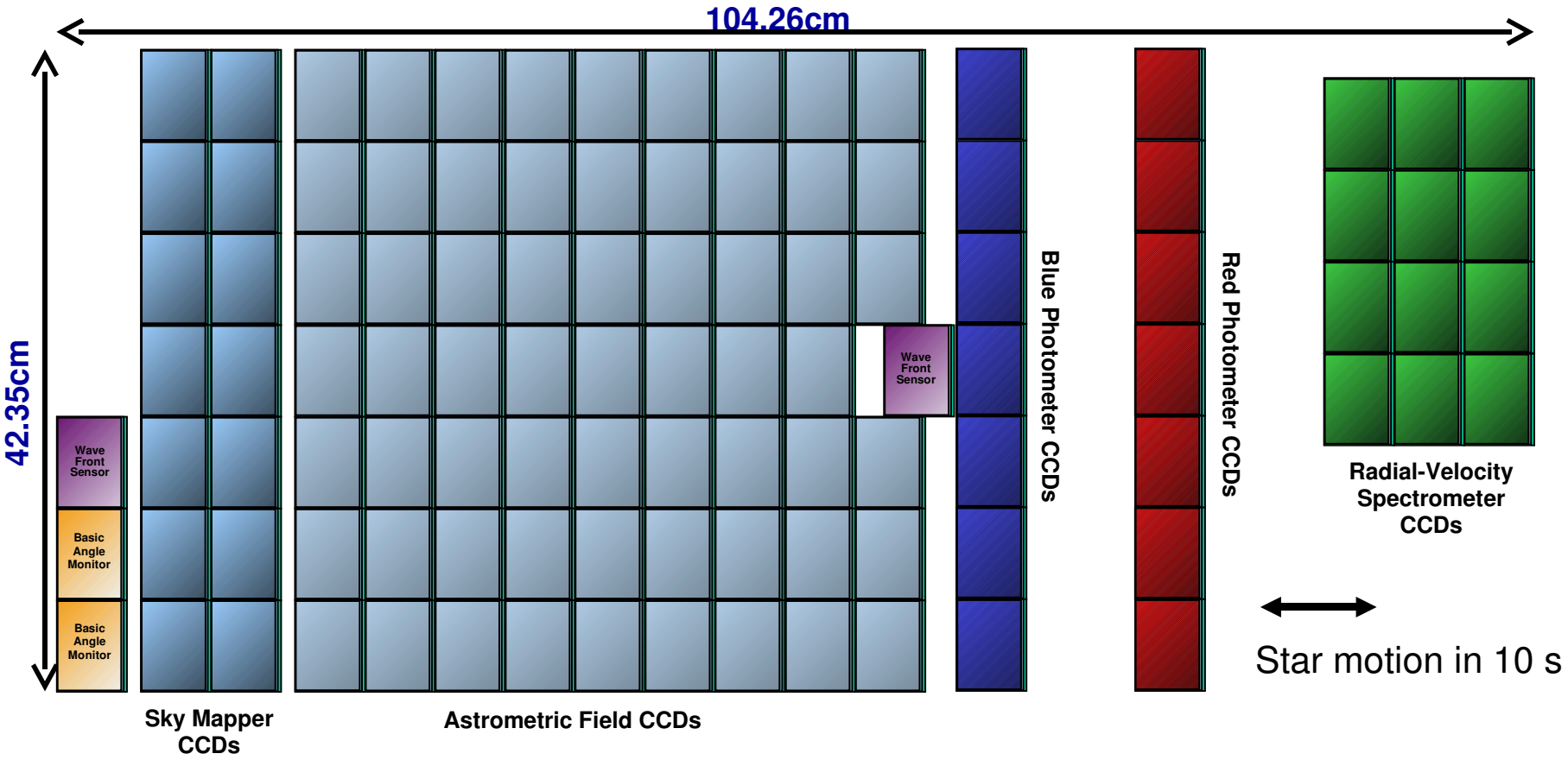


Dispersion in the red channel (650 – 1000 nm):
7 – 15 nm/pixel

Dispersion in the blue chanel (330 – 650 nm):
4 – 32 nm/pixel

Focal plane

Figure courtesy Alex Short



Total field:
- active area: 0.75 deg^2
- CCDs: 14 + 62 + 14 + 12
- 4500 x 1966 pixels (TDI)
- pixel size = 10 µm x 30 µm
  = 59 mas x 177 mas

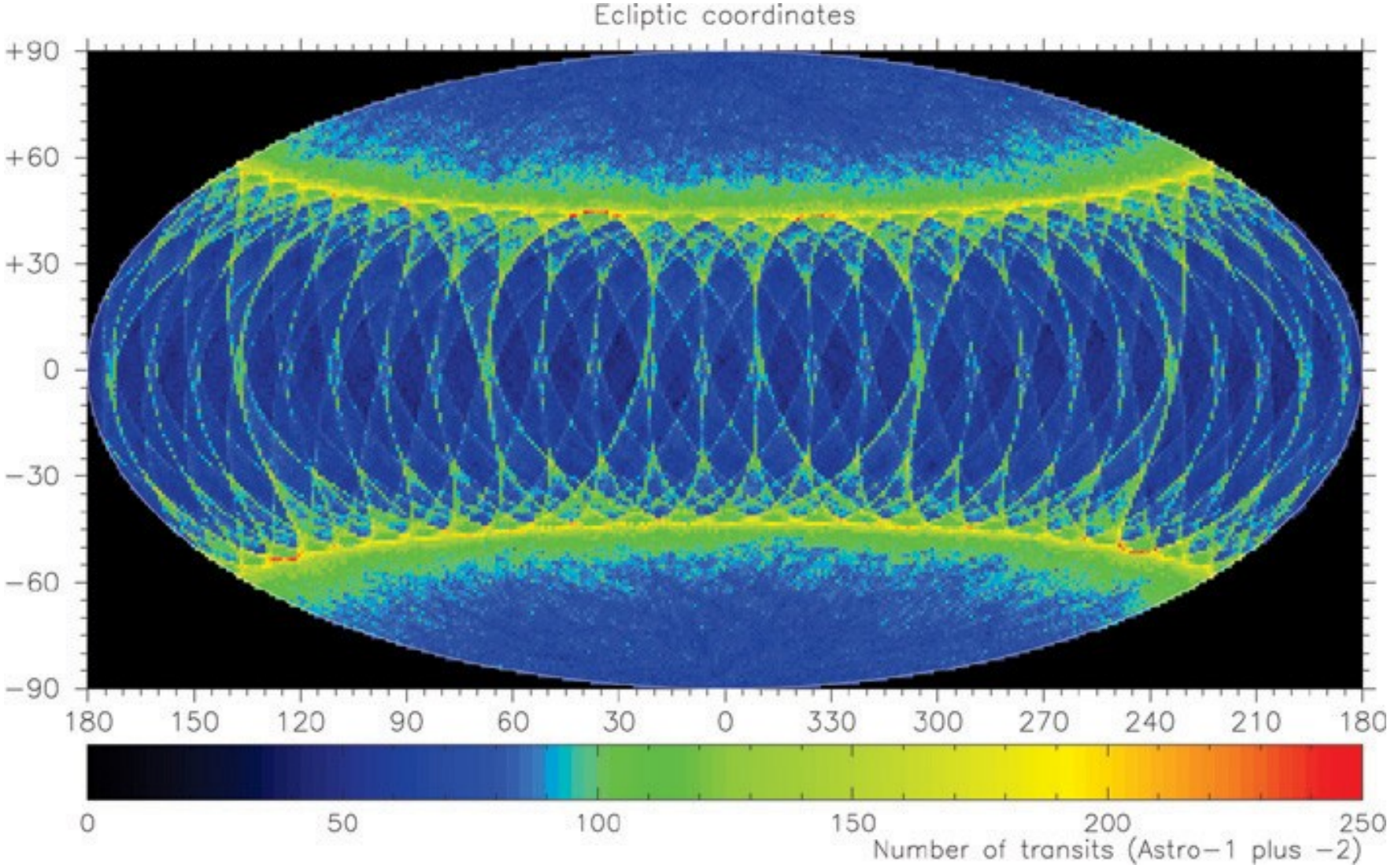
Sky mapper:
- detects all objects to 20 mag
- rejects cosmic-ray events
- FoV discrimination

Astrometry:
- total detection noise: 6 e$^-$

Photometry:
- two-channel photometer
- blue and red CCDs

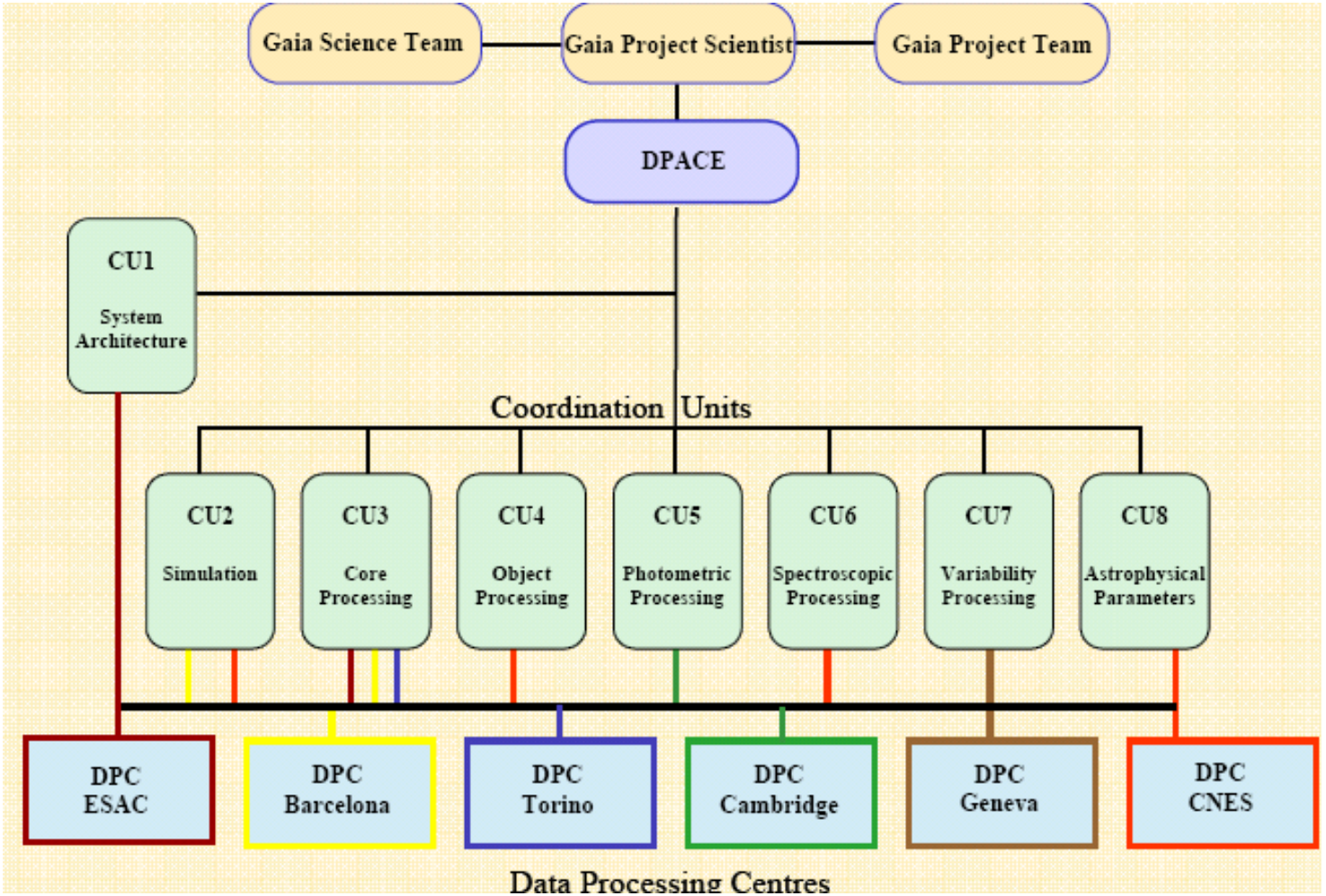
Spectroscopy:
- high-resolution spectra
- red CCDs

Gaia scanning law II



Gaia

DPAC Data Processing and Analysis Consortium

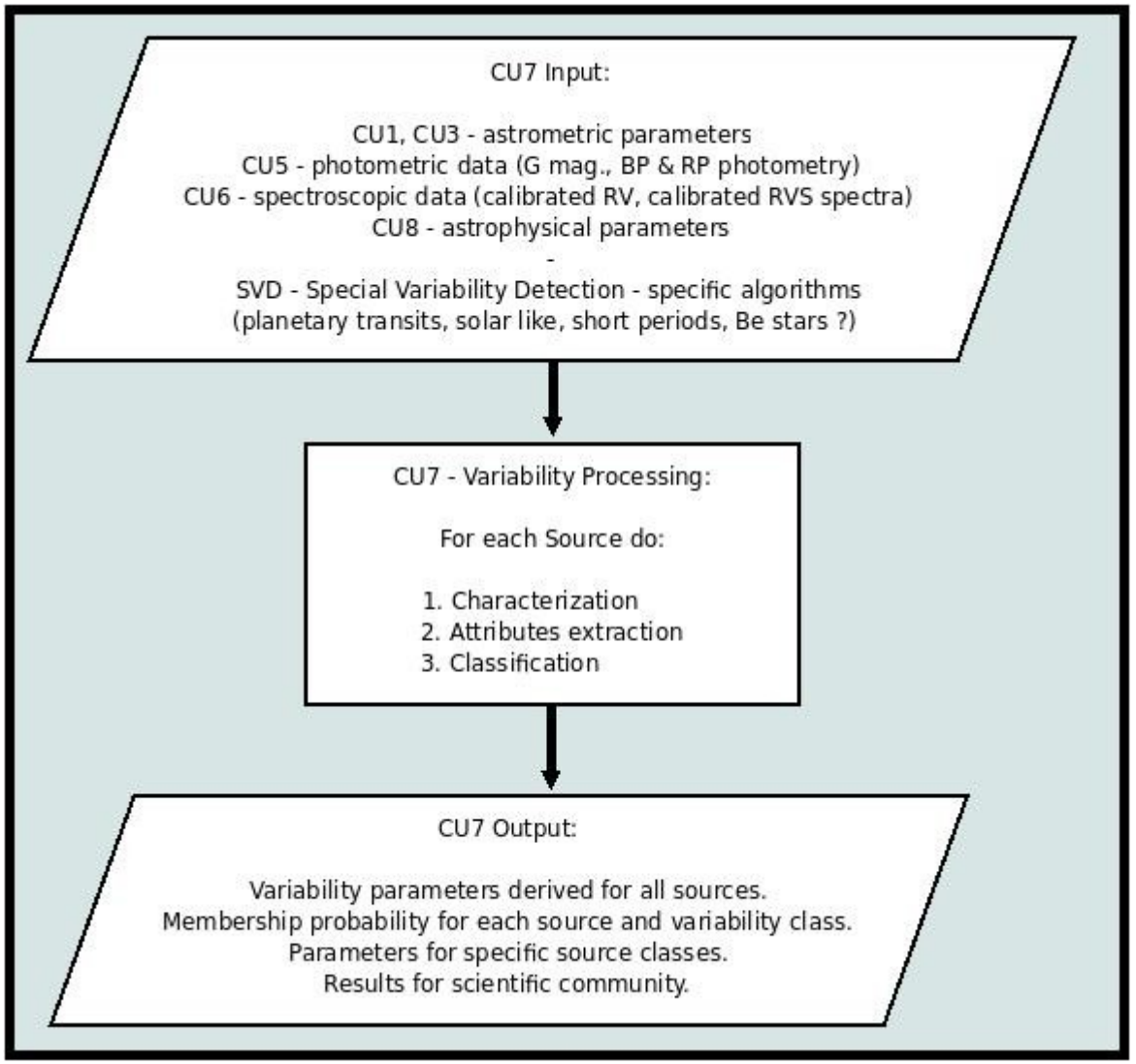


Gaia Be star workpackage

The aims:

- **Implement machine learning classification algorithms** for Be stars light curves.
- **Test classification algorithms** using the OGLE Be stars data.
- **Integrate classification algorithms** inside the CU7 pipeline system.

CU7 Data Flow



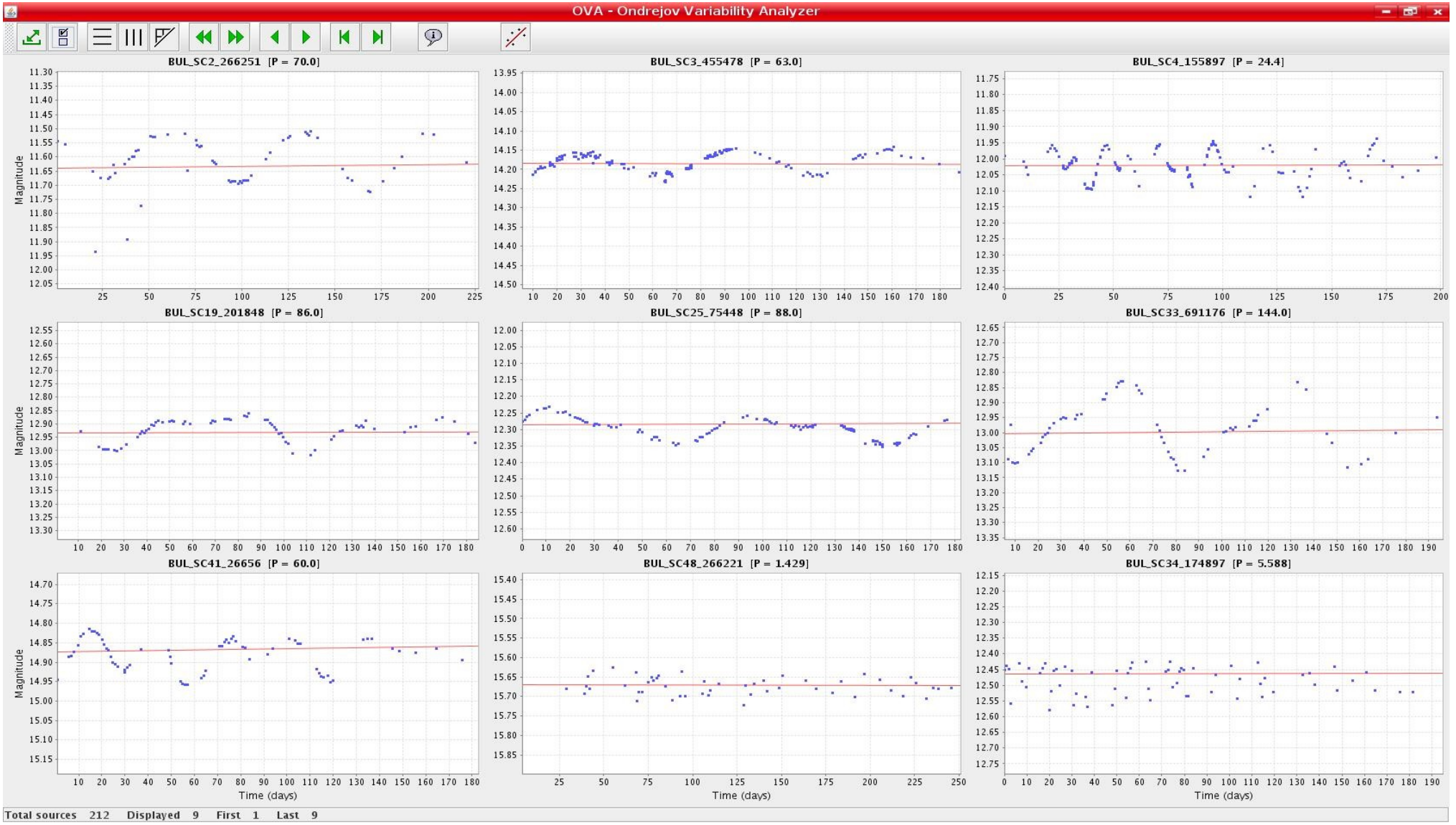
OGLE Be stars

- **Paper: A catalogue of Be stars in the direction of the Galactic Bulge**, **A&A 478, 659-665 (2008)**. This paper describes the first systematic search for Be stars candidates in the direction of the Galactic Bulge, based on specific criteria of magnitude, colour and variability in the B & I bands on 48 OGLE II GB fields (The **O**ptical **G**ravitational **L**ensing **E**xperiment).
- **There are 29053 possible Be stars candidates.**
- **There are 1488 almost certainly Be stars.**
- **There were found 196 periodic Be stars**.

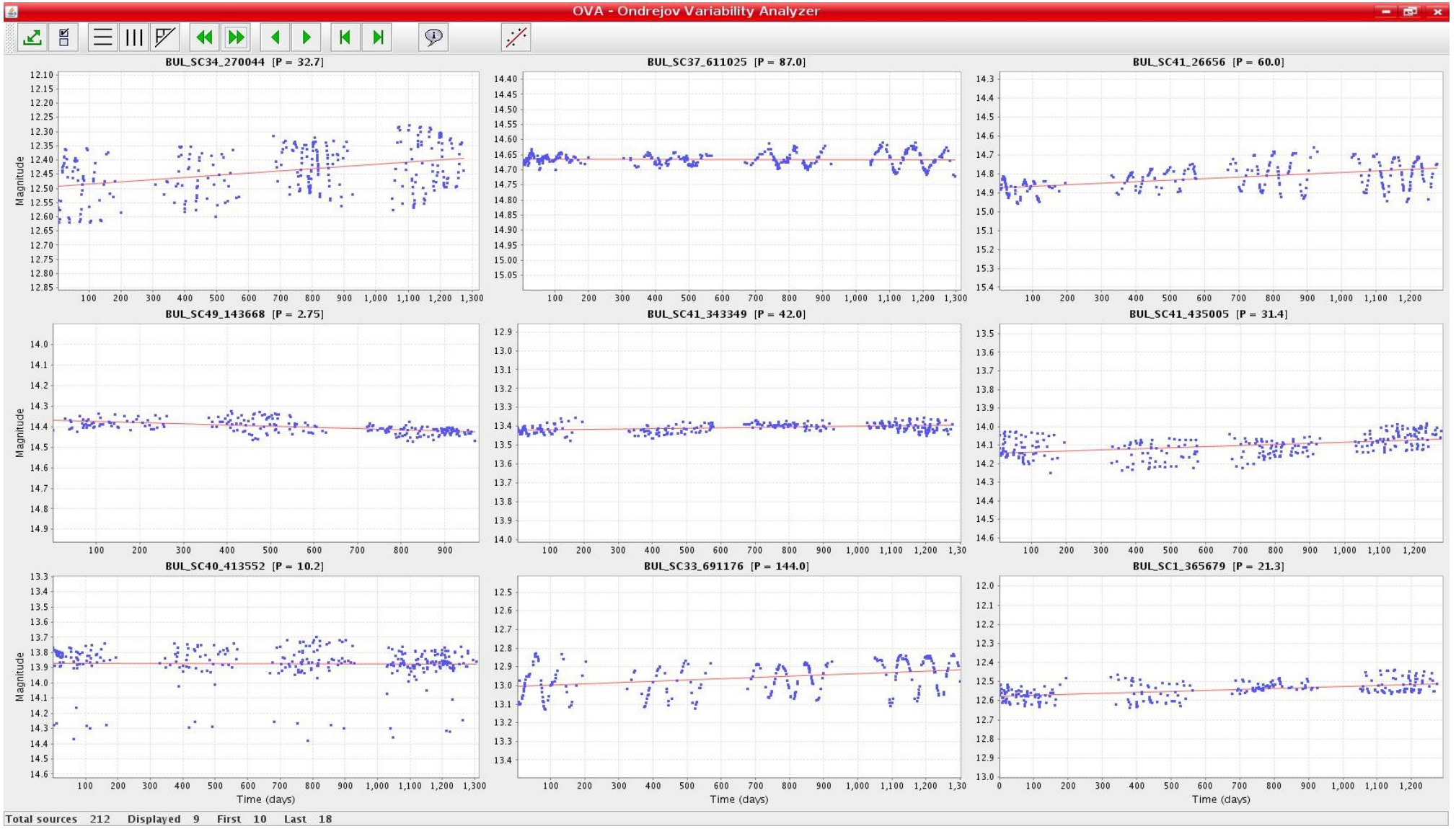
OGLE Be Stars

- **The subset of 1488 OGLE Be stars light curves represents an ideal and large enough sample of data** used for our classification.
- **OGLE data covers ~4 years of observation**, similar to Gaia mission.
- **OGLE data have a good sampling**, each light curve contains 250-350 data points compared to 50-80 measurements from Gaia mission.

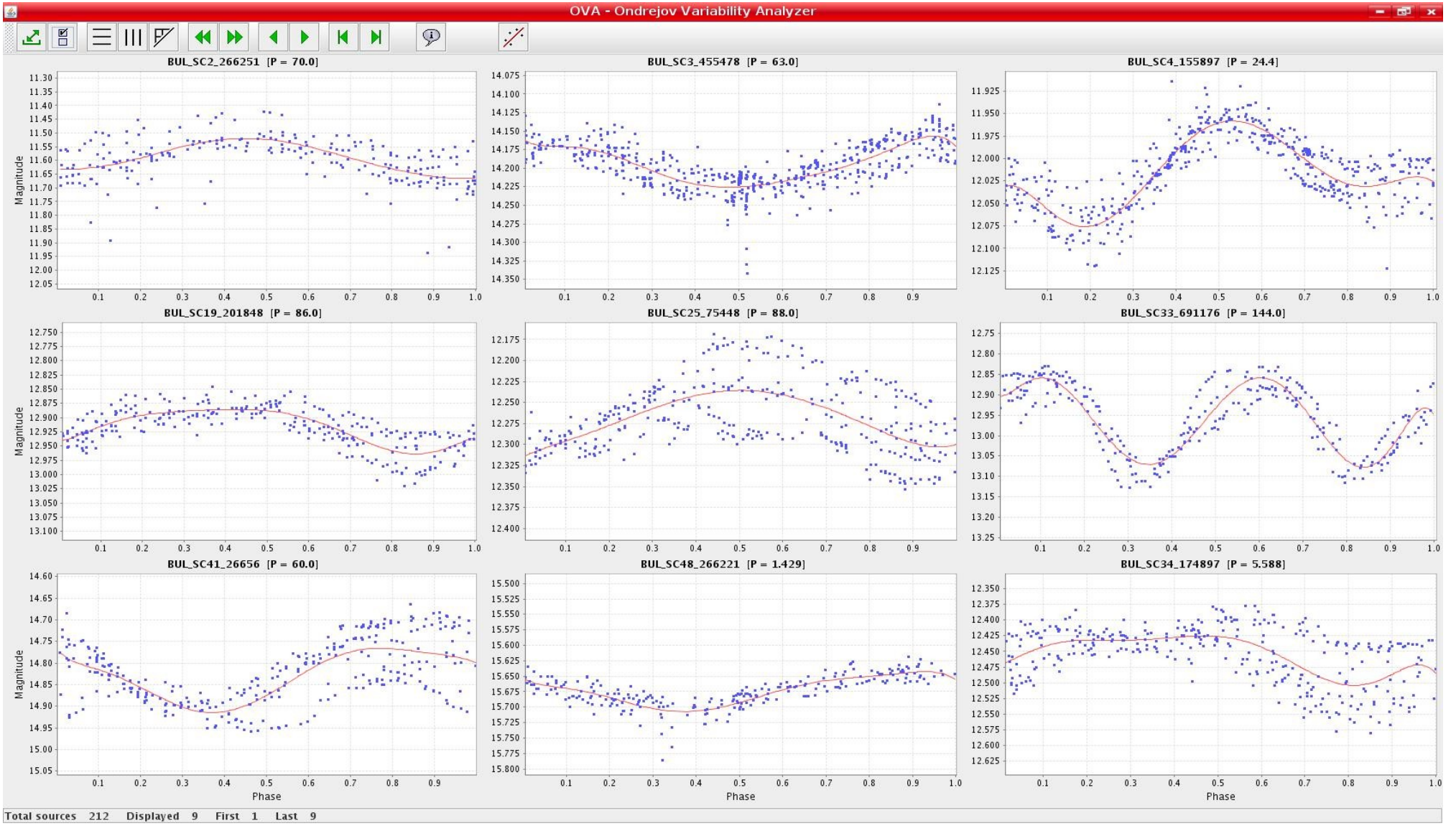
OGLE Be Stars



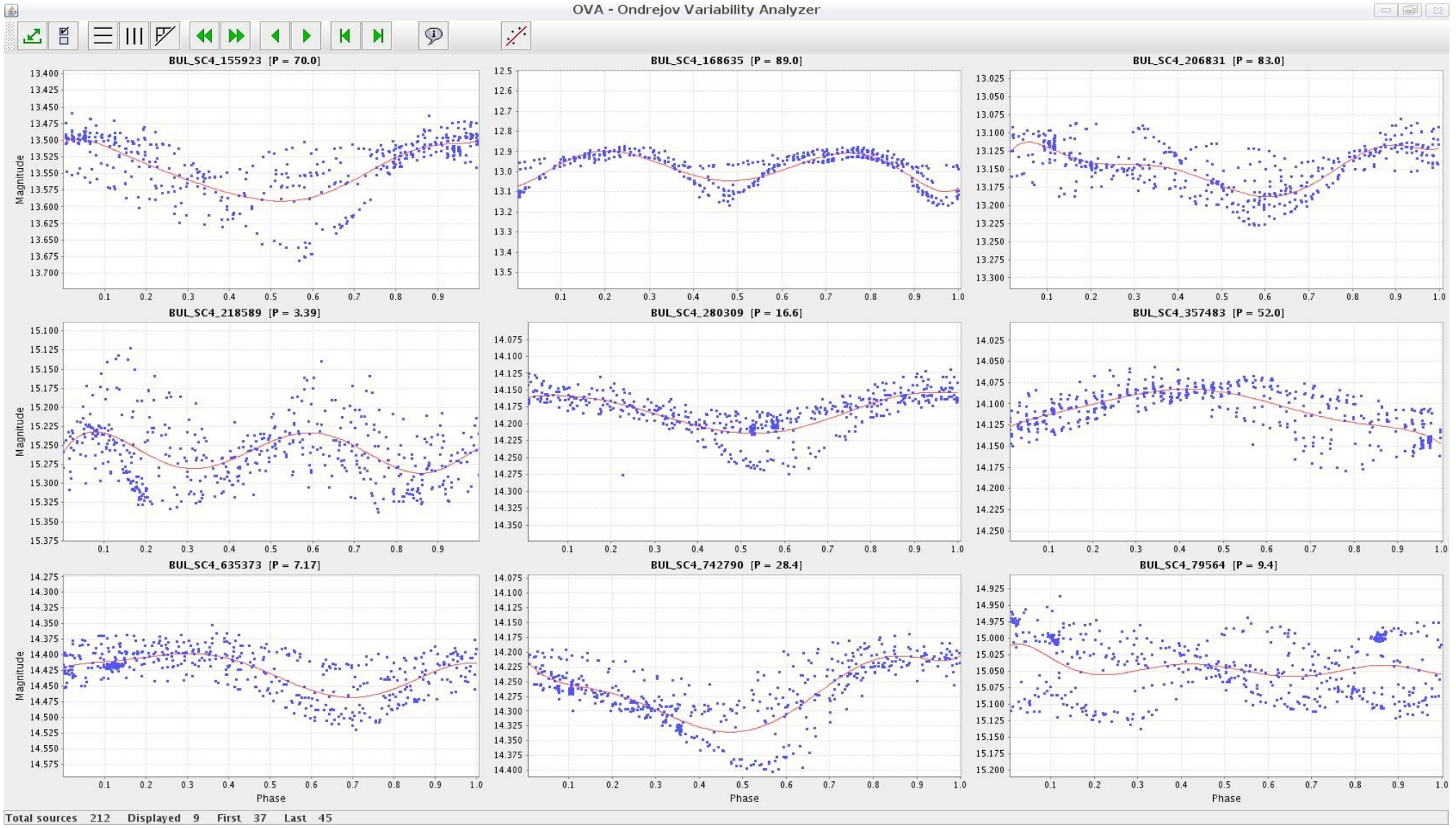
OGLE Be Stars



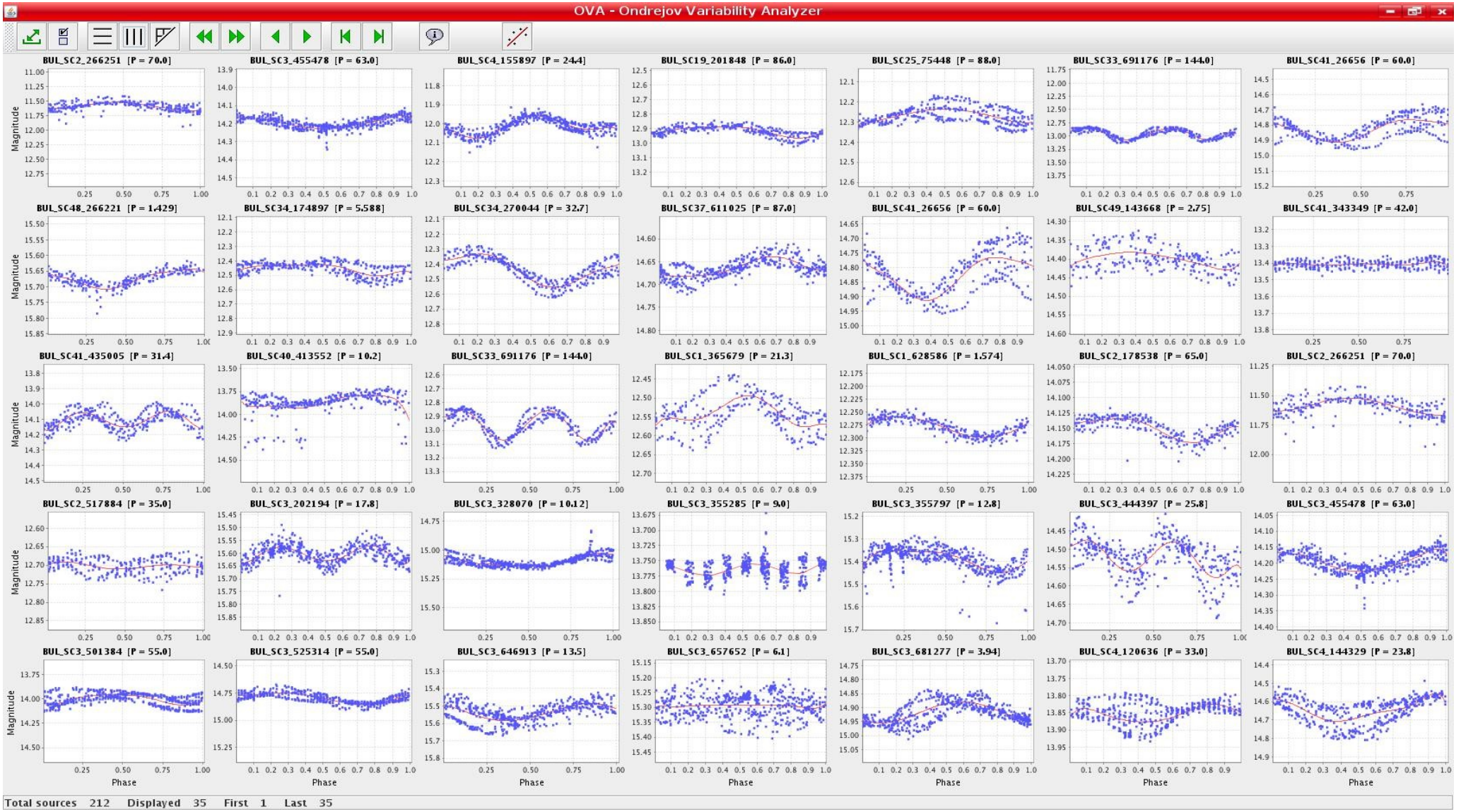
OGLE Be Stars



OGLE Be Stars



OGLE Be Stars



Basic Be star properties

- **Extremely luminous and blue stars**.
- **Have a prominent emission lines of hydrogen in its spectrum.**
- **Rapidly rotated.**
- **Be stars may vary in brightness on 3 time scales:**

  1st class - short var. 0,2–2 days with 0.1 mag,

  2nd class - med. var. weeks-months, up to 0.2 mag,

  3rd class - long var. years-decades up to 0.8 mag.

Be stars classification strategy

- **1st - classify the first 1488 Be stars** (training set) based on their similar properties using supervised and/ or unsupervised analysis.
- **2nd - classify the rest of possible Be stars** (~27500 stars) to such groups.
- **3rd- resample the OGLE Be stars** similarly to the Gaia light curves and try to classify them again.
- **4th - implement Be stars classification** algorithms in CU7 framework.

CU7 - Characterization

- **Derives the basic statistical parameters** for each Source (light curve) on its input (weighted mean, min, max, median, stdev, skewness, kurtosis, polynomial models, variability flag, etc.).
- **Derives the period search results** (number of periods, period values and their associated probabilities, etc.).

CU7 - Attribute Extraction

- **This module extracts attributed** from the source object which are requested by a machine learning classificator.
- **These attributes are also used for training** of machine learning classification methods to build the class models.

CU7 - Classification

- **Compute a set of membership probabilities**, i.e. the probability that a source belong to a given variability class.
- **Results are inside the 3 dim. array**: Map<ClassificationMethod, Map<SourceClass, ClassMembership>>
- **All sources of a given type should be sent** to some specific processing (SOS).

CU7- **S**pecific **O**bject **S**tudies

- **There will be a SOS package** for each detected source type on its input (AGN, Be stars, Cataclysmic variables, etc.).
- **The SOS should confirm** the classified source type.
- **The SOS should derive specific parameters** for this particular source type.

CU7 - classification tool

- **WEKA** – The **W**aikato **E**nwironment for **K**nowledge **A**nalysis tool from Waikato University New Zealand.
- **Contains supervised, unsupervised classification** algorithms, neural nets, visualization, etc..

What is statistical classification ?

- **Classification is the task of learning a target function *f* that maps each attribute set X to one of the predefined class labels Y**.
- Training data: {($\mathbf{X_1}$,Y_1),...($\mathbf{X_n}$,Y_n)} produce a classifier *f*: **X**->Y which maps an object **X** to its classification class Y.
- We create classification models from an input data set.
- Examples includes: linear & quadratic classifiers, decision tree, rule-based, neural networks and Bayes classifiers.
- Each technique employs a **learning algorithm** to identify a model that best fits the relationship between the attribute set and and class label of the input data.

Concept of probability

- The probability of a sample point (outcome) is the proportion of occurrences of the sample point in a long series of experiments.
- **1. Mutually exclusive events** E1, E2 – no common sample points. $P(E1+E2) = P(E1) + P(E2)$
- **2. Not mutually exclusive events** E1, E2 – contains one or more common sample points. $P(E1+E2) = P(E1) + P(E2) – P(E1,E2)$
- **2a. Independent events:** $P(E1,E2) = P(E1) * P(E2)$
- **2b. Dependent events:** $P(E1,E2) = P(E1) * P(E2 \mid E1)$
- Dependence between events is treated by the notion of conditional probability.

$$P(E1|E2) = \frac{P(E1, E2)}{P(E2)}$$

Bayesian classifiers

- In many applications the relationship between the attribute set and the class variable is **non-deterministic**, *i.e. the class record cannot be predicted with certainty – for example* ***noisy data***.
- **Bayesian classifiers allow modeling of probabilistic relationships between the attribute set and the class variable**.
- Based on Bayes Theorem (conditional probability):

$$P(Y|X) = \frac{P(X|Y) * P(Y)}{P(X)}$$

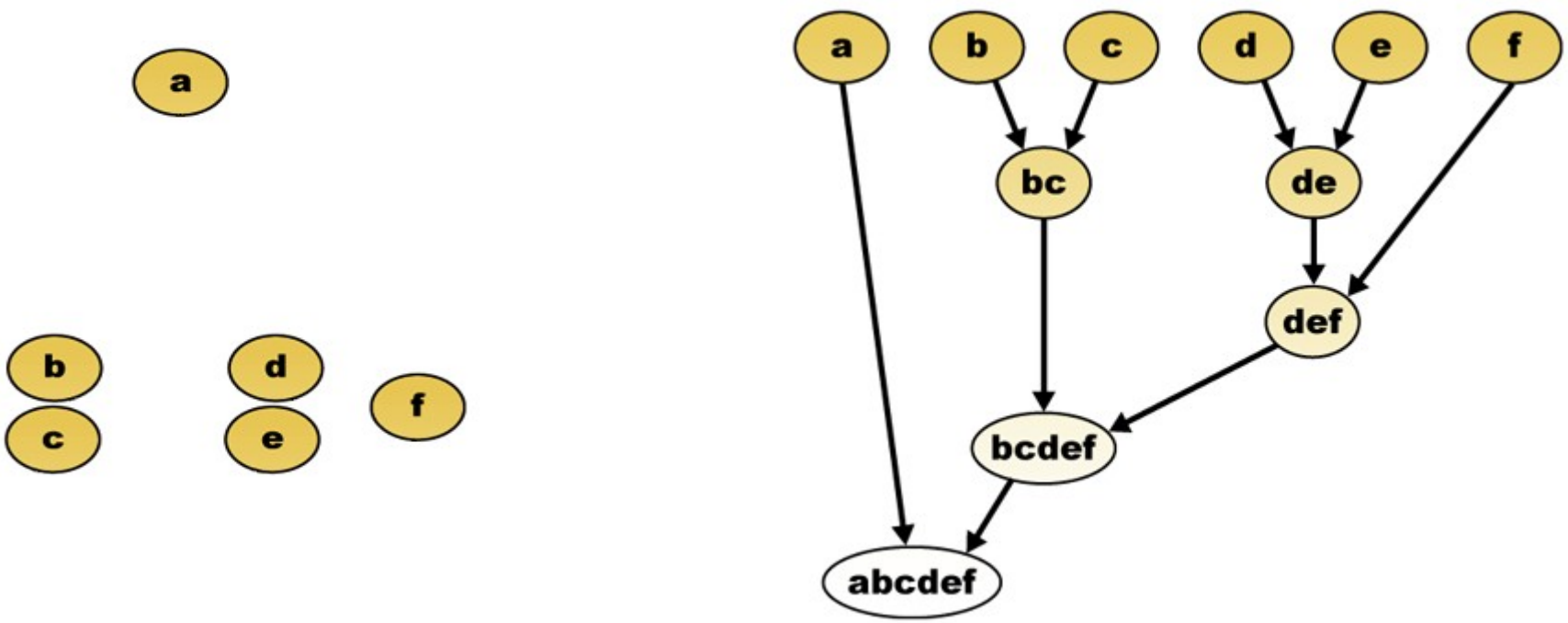
Bayesian classifiers

- During the training phase we need to learn these P(Y|**X**) conditional (posterior) probabilities for each combination of **X** and Y based on information from the training data.
- By knowing these probabilities, a new record **X'** can be classified by finding the class Y' that maximizes the posterior probability P(Y'|**X'**).
- BT allows us to express the posterior probability in terms of prior probability P(Y), the **class-conditional probability** P(**X**|Y), and the evidence, P(**X**).

Clustering

- **Unsupervised classification – no classification rules are known**.
- Instances are divided into natural groups.
- Groups may be mutually exclusive or overlapping.
- Groups may be probabilistic – instance belongs to each group with a certain probability.
- Groups may be hierarchical.

Hierarchical clustering



We try to minimize total intra-cluster variance, or, the squared error function

$$V = \sum_{i=1}^{k} \sum_{x_j \in S_i} (x_j - \mu_i)^2$$

where there are ***k*** clusters Si, i = 1, 2, ..., k, and μi is the centroid or mean point of all the points xj ∈ Si.

Clustering - ***k-means***

- Simple & effective clustering technique.
- Specify ***k*** – how many clusters will be sought.
- Then ***k*** points are randomly selected as cluster centers.
- All instances are assigned to their closest cluster center - the ordinary Euclidean distance: $r(\mathbf{x}i,\mathbf{x}j) = \mathrm{sqrt}(\mathrm{sum}(x_{ik} - x_{jk})^2)$

  $k=1..N$
- Then new cluster centers are calculated (centroid, mean).
- New iteration is done using these new cluster centers.
- Cluster center errors are calculated.
- Iteration is stopped when the computed error is less or equal that requested one.

Clustering - ***k-means***

- Centroid minimizes the total squared distance from each of the cluster's points to its center.
- But the **minimum is local one**.
- The final clusters are quite sensitive to the initial cluster centers.
- Difficult to find globally optimal clusters.
- **Recommendation**: run the algorithms several times with different initial choices and the select the one with the smallest total squared distance.

Current status

- **The OGLE Be stars dataset delivered** to CU7 group in Geneva for its integration into their framework.
- **The first version of BeStarsAttrExtractor delivered** to ESA CU7 SVN system in Feb. 2009. This application generates the WEKA's ARFF file.
- **Work on Be stars WEKA model in progress.**

Conclusion

- Software development for Be stars classification in progress in close cooperation with CU7 group in Geneva Observatory.