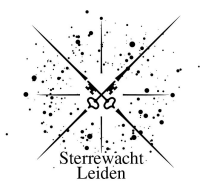
Science from Gaia: how to deal with a complex billion-source catalogue and data archive

Anthony Brown

Sterrewacht Leiden, Leiden University
brown@strw.leidenuniv.nl

- Gaia mission overview
- Complexities of the Gaia catalogue
- Modelling the Milky Way
- Extracting the maximum science from the data archive



- ESA Cornerstone mission within Horizon 2000+ programme
- Create large and highly accurate stereoscopic map of the Galaxy
- Global astrometry concept successfully demonstrated by Hipparcos

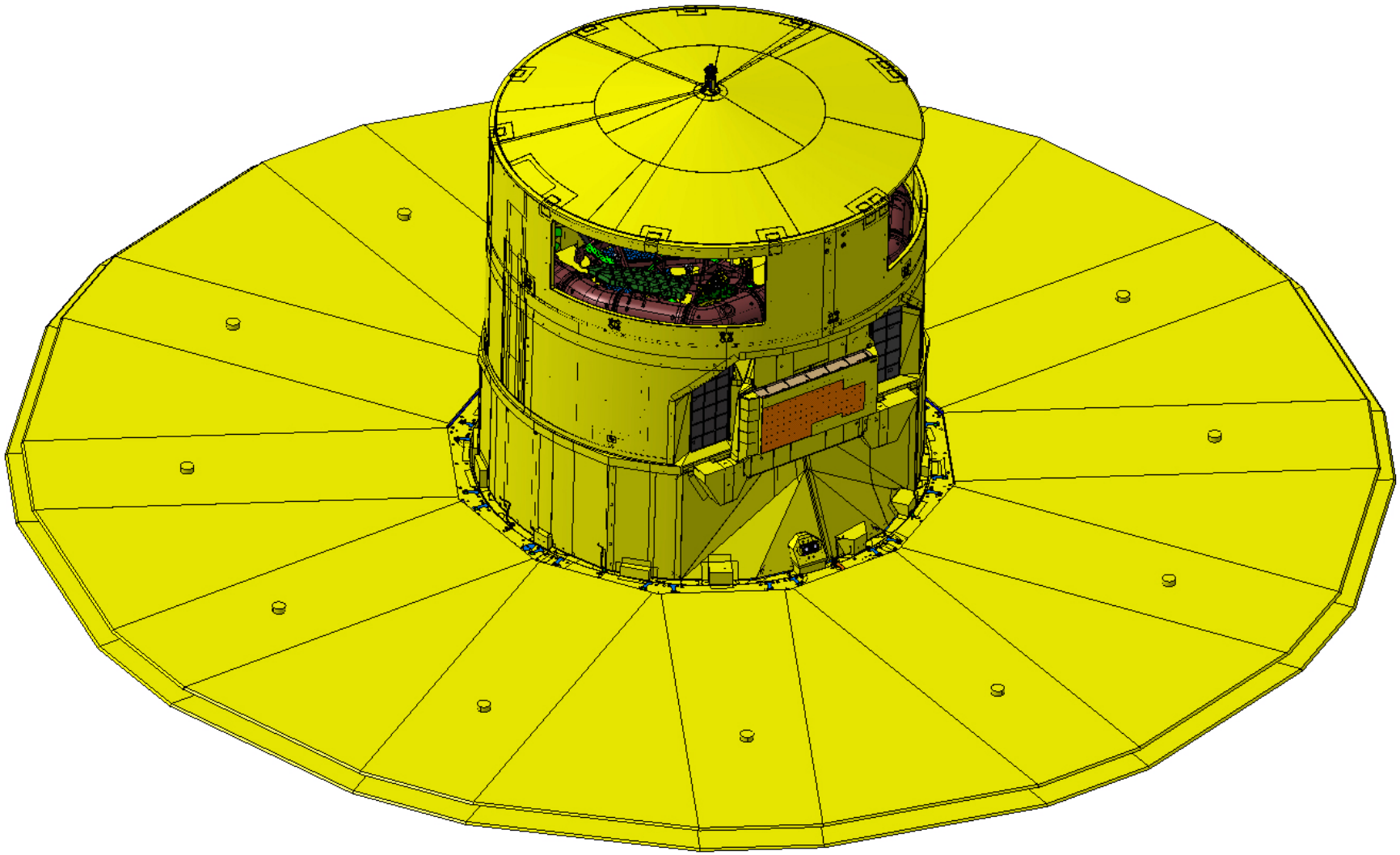
Gaia DPAC

gaia

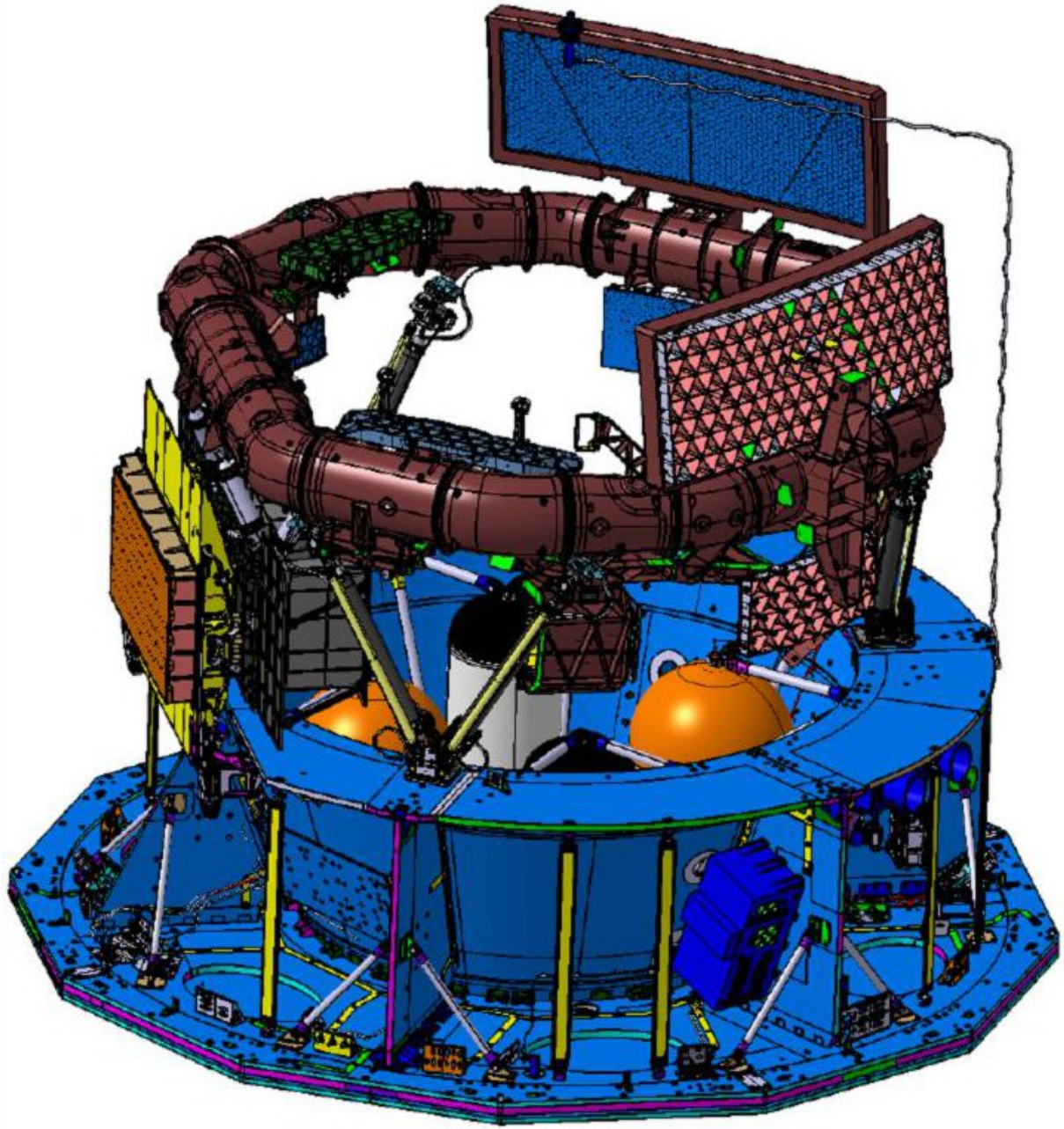
- Launch in 2013 with Soyuz-Fregat from Kourou
- Orbit: vicinity of L2
- Mission duration 5 (+1) years

Image credit: Lund Observatory

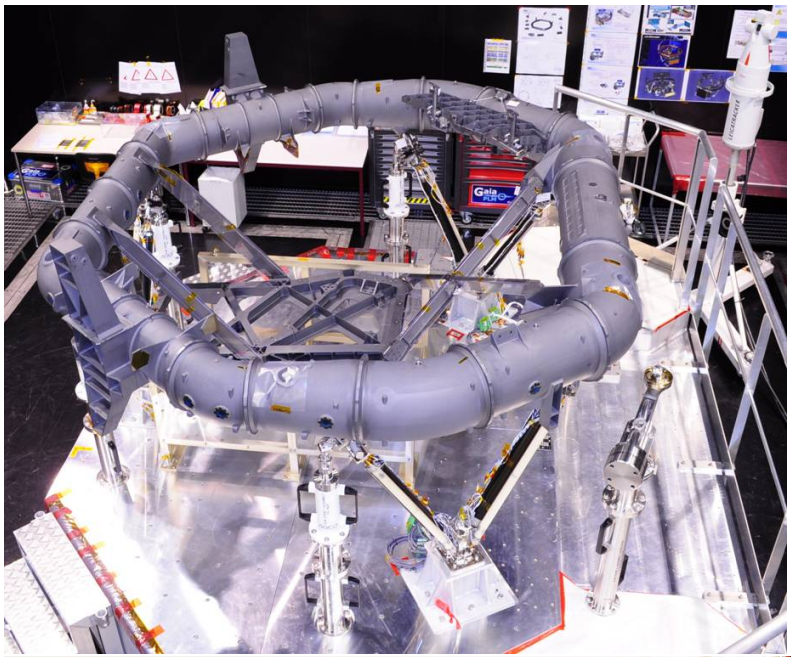
Spacecraft status

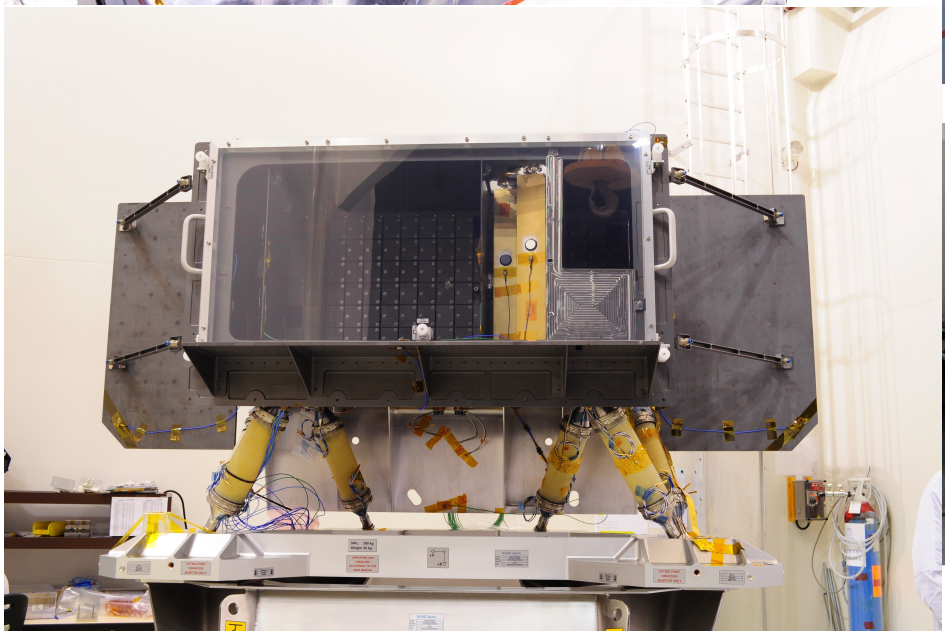


Spacecraft status



Spacecraft status





esa EADS astrium SAFRAN Sagem

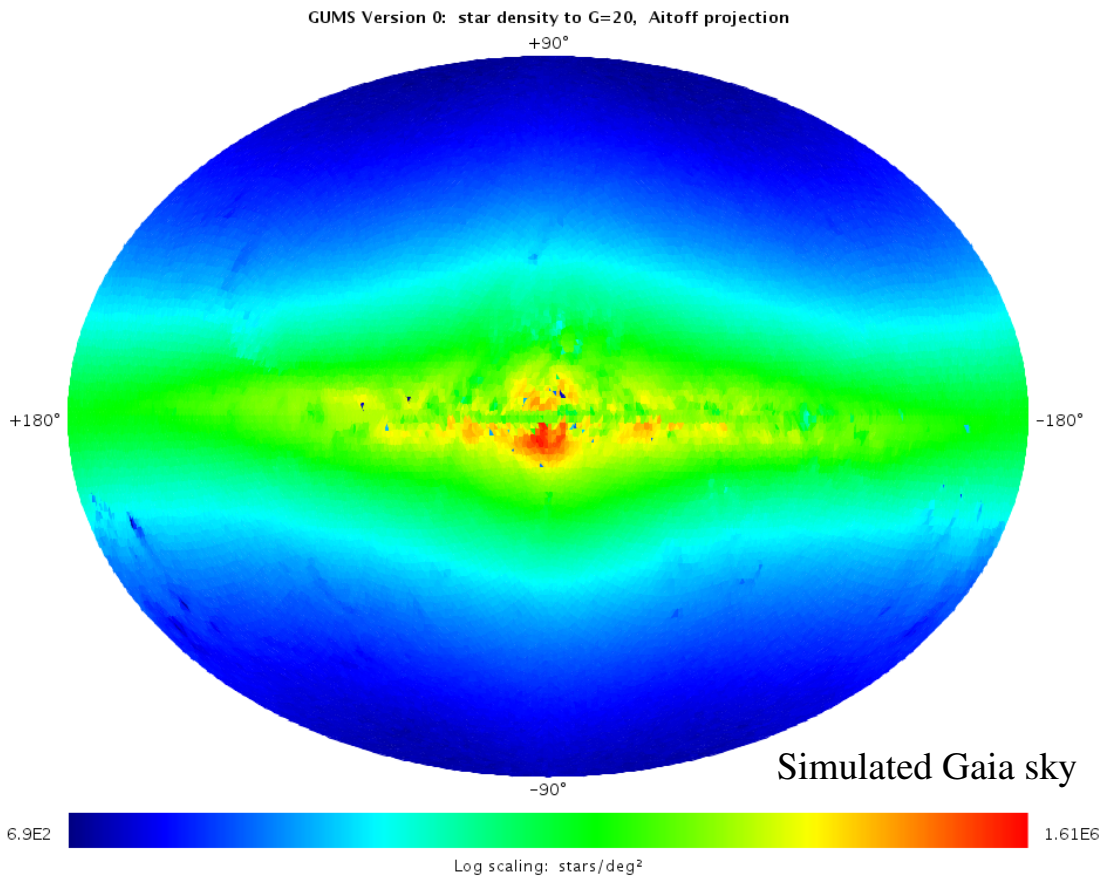
Spacecraft status

Survey capabilities

- Three simultaneous observing modes
- Complete to $G = 20$ ($V = 20$–22)
- Observing programme: autonomous on-board detection and unbiased
- Quasi-regular time-sampling over 5 years ($\sim$ 80 observations)
- Angular resolution comparable to HST

Number of objects

- 1 billion stars to $G = 20$
- 10^6–10^7 galaxies
- 500 000 quasars
- 3×10^5 solar system bodies
- tens of thousands of exoplanets

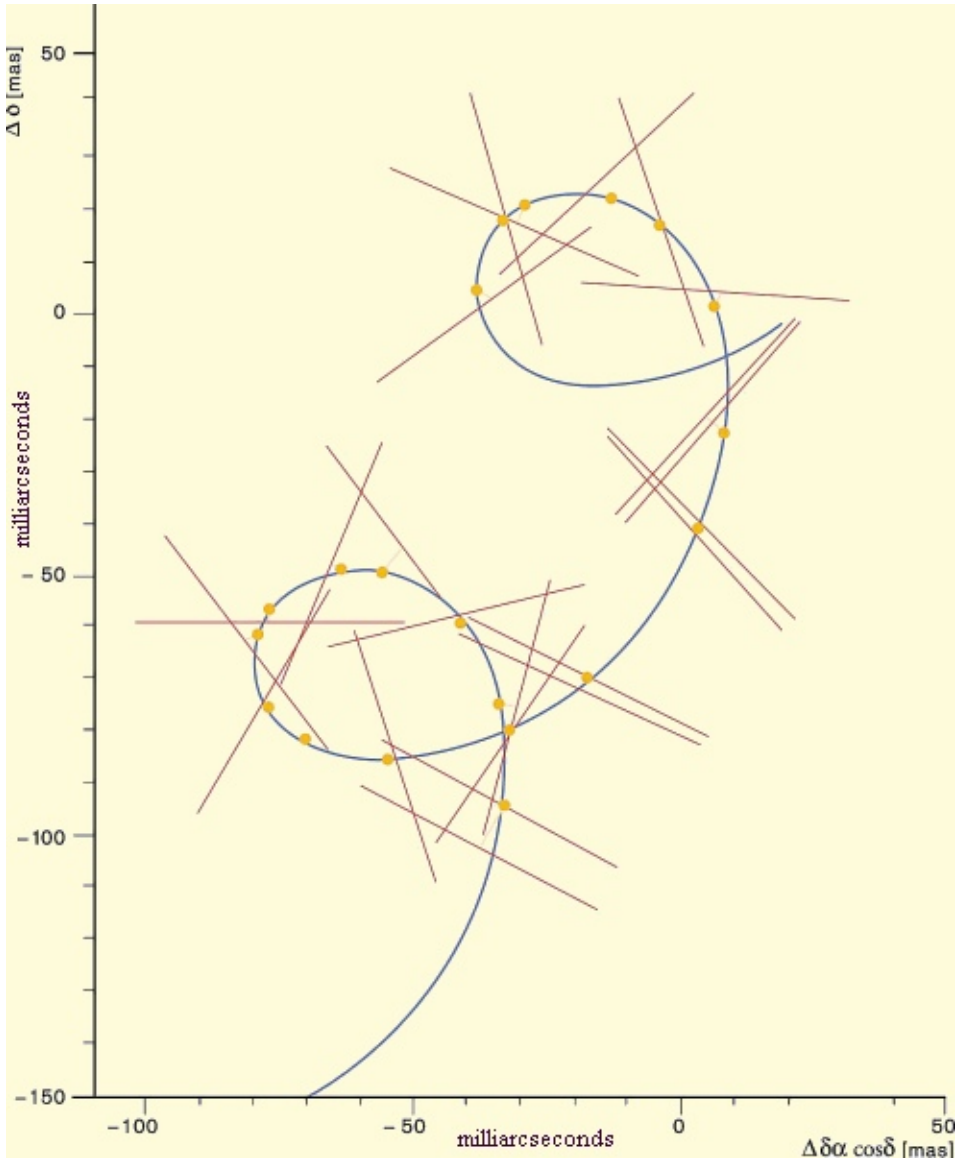


Simulated Gaia sky

Summary of Gaia science harvest

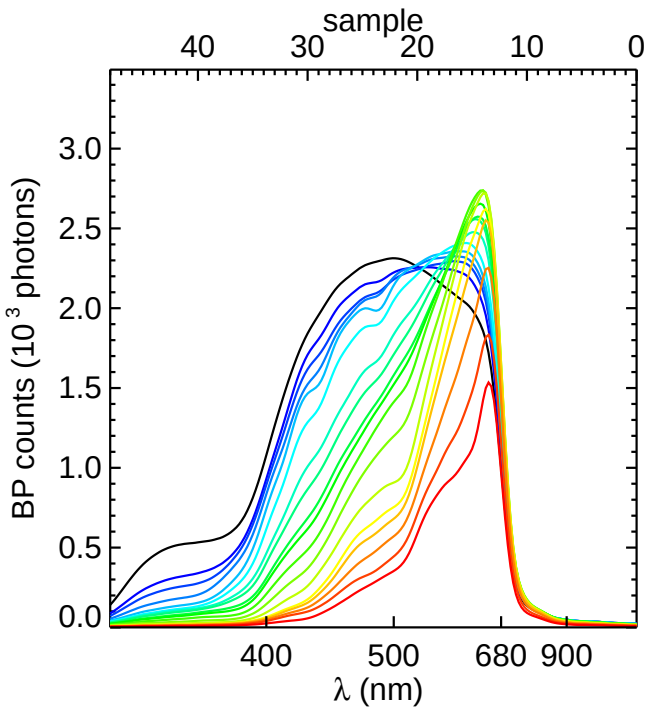
Astrometry

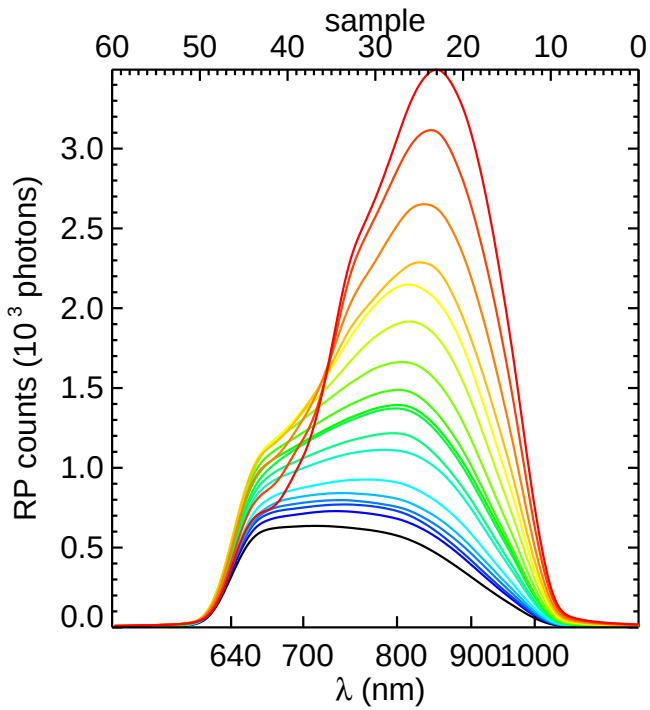
- $\sigma \sim 5$–14 μas $V < 12$, 10–25 μas $V = 15$, 100–300 μas $V = 20$
- $25\,000\ \star/\mathrm{deg}^2$; max $\sim 10^6\ \star/\mathrm{deg}^2$
- Census extra-solar planets to 200 pc
- 3×10^5 minor bodies of the solar system, 100 masses
- PPN γ to $\sim 2 \times 10^{-6}$



Photometry

- Two channels: 330–680 nm (BP), 640–1000 nm (RP)
- Low resolution ($\sim$ 3–30 nm/pixel) prism spectra
- Allows derivation of A_V, $T_{\rm eff}$, $\log g$, [M/H], and [α/H] for brighter stars





Summary of Gaia science harvest

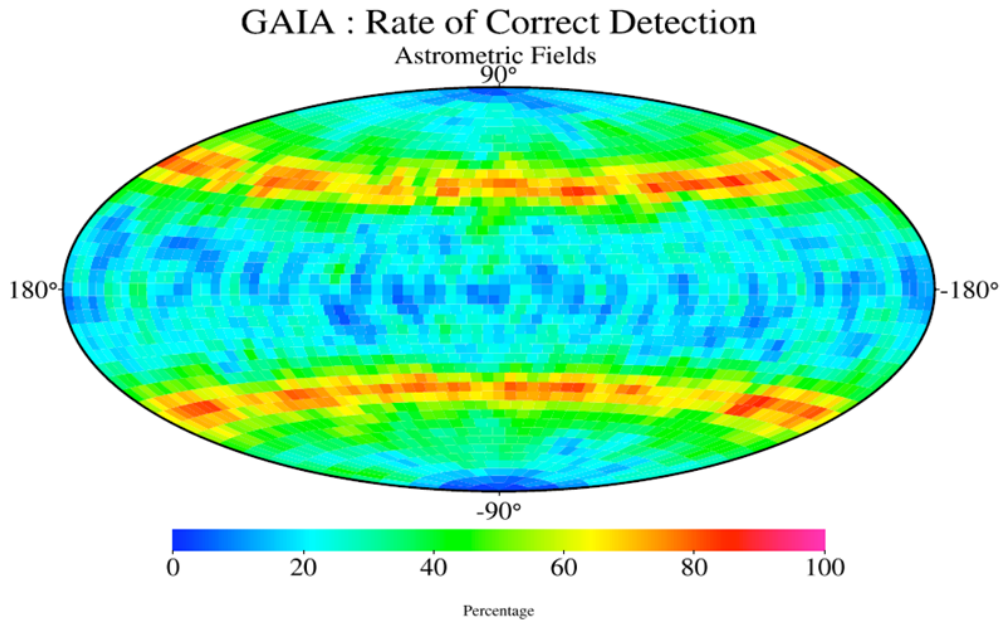
Gaia variable star survey

- $\sim$ 70 epoch survey over 5 years
- mmag accuracy per single observation

Quantitative impact

- 20×10^6 classical variables
- 1–5 million eclipsing binaries
- $\sim$ 5000 Cepheids, 70 000 RR Lyr
  - RR Lyr visible out to $\sim$ 75 kpc

Example: S/N= 0.75, Per=0.2 day

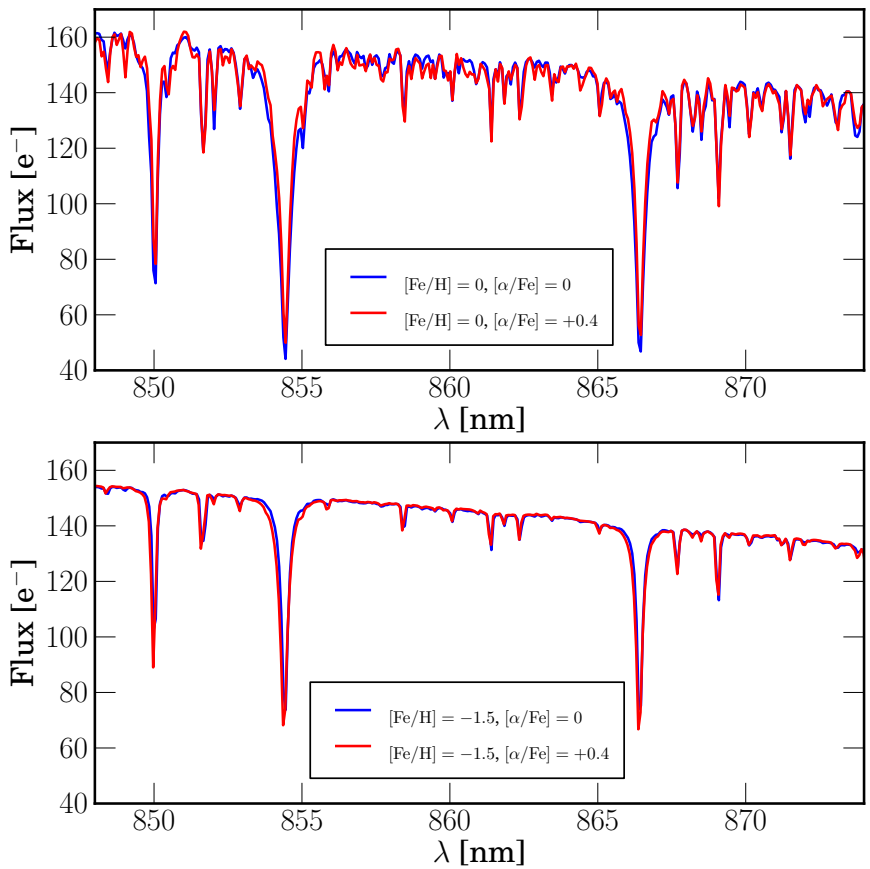


Eyer & Mignard 2005

Summary of Gaia science harvest

Spectroscopy

◆ Radial velocities	$V \leq 17$	$\sim 150 \times 10^6$ stars
◆ Rotational velocities	$V \leq 13$	$\sim 5 \times 10^6$
◆ Atmospheric parameters	$V \leq 13$	$\sim 5 \times 10^6$
◆ Abundances	$V \leq 12$	$\sim 2 \times 10^6$
◆ Interstellar reddening	$V \leq 13$	$\sim 5 \times 10^6$



Diagnostics from spectroscopy

- Binarity/multiplicity, variability
- $\sim 10^6$ spectroscopic binaries
- $\sim 10^5$ eclipsing binaries ($\sim 25\%$ SB2 $\rightarrow$ masses)
- Long period classical Cepheids $\sigma_{v_r} < 7$ km/s $\rightarrow$ 20–30 kpc

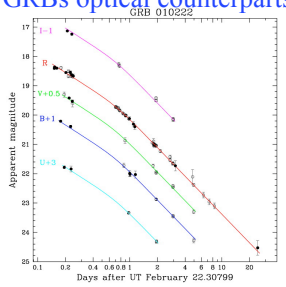
Summary of Gaia science harvest

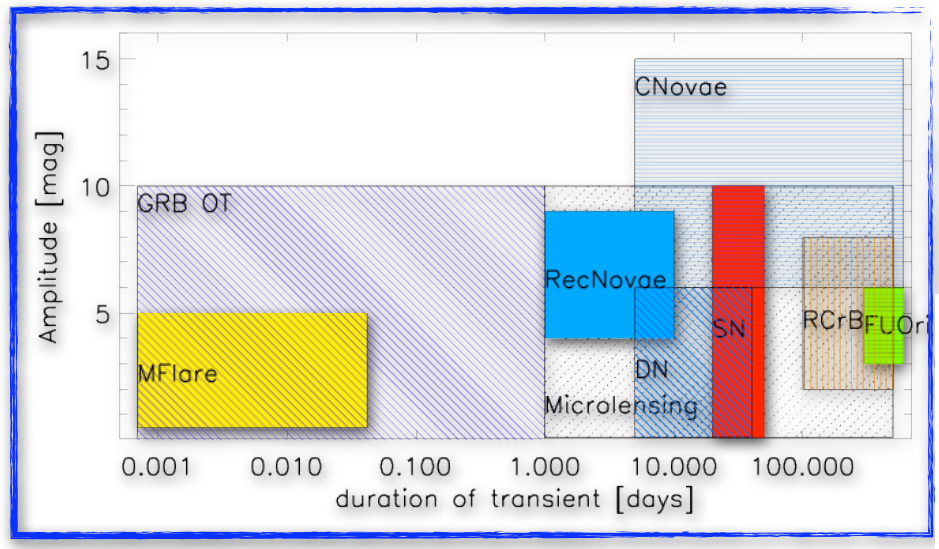
Other

- Accurate stellar classification for all classes and types
- Recalibration of the distance scale
- 10 000 stellar masses $\sigma < 1\%$
- 5×10^5 QSOs + z + photometry, ICRF in the visible

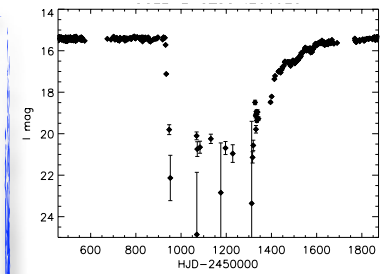
Potential Triggers for Gaia Alerts

GRBs optical counterparts

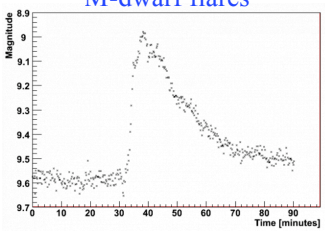




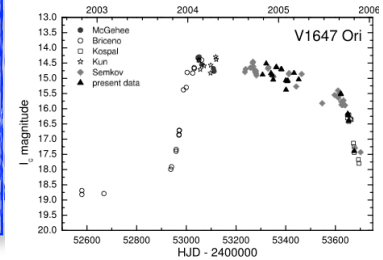
R Coronae Borealis



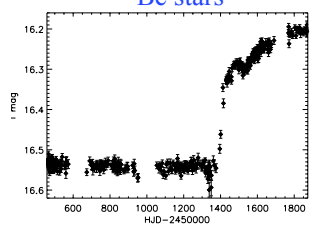
M-dwarf flares



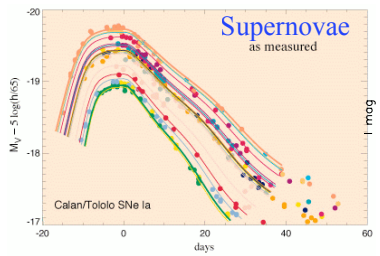
FU Orionis and similar

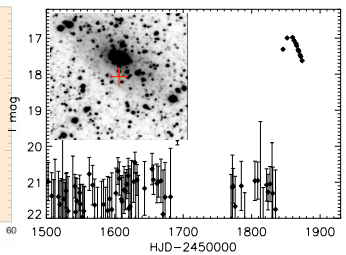


Be stars

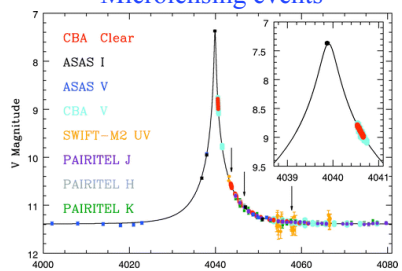


Supernovae
as measured

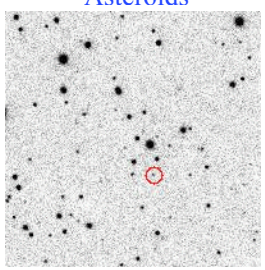




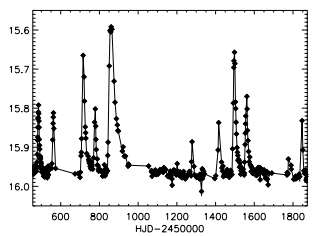
Microlensing events



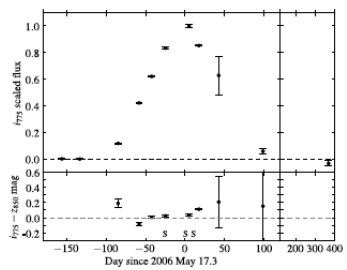
Asteroids



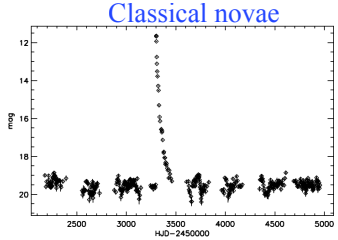
Dwarf novae



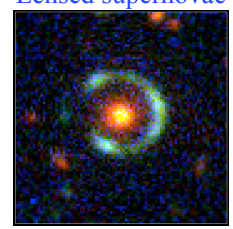
NEW THINGS??



Classical novae

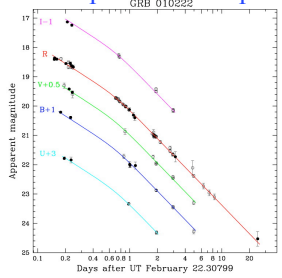


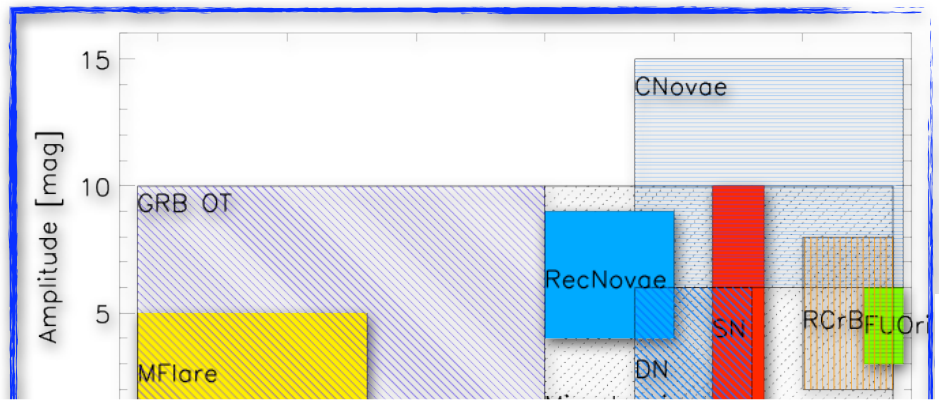
Lensed supernovae



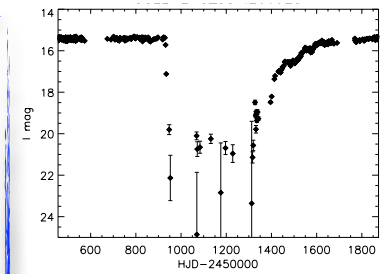
Potential Triggers for Gaia Alerts

GRBs optical counterparts



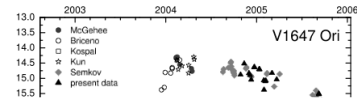


R Coronae Borealis



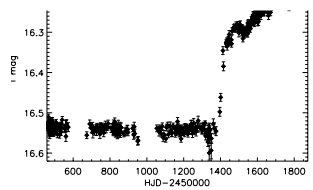
M-dwarf flares

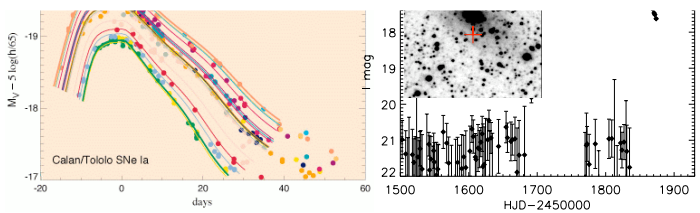
FU Orionis and similar

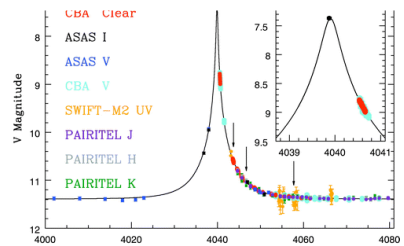


Workshop from June 29th to 1st July 2011

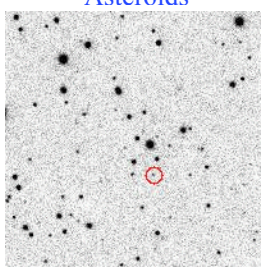
http://www.ast.cam.ac.uk/ioa/wikis/gsawgwiki/index.php/Workshop2011:main



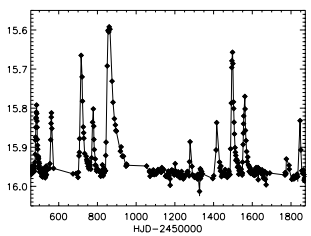




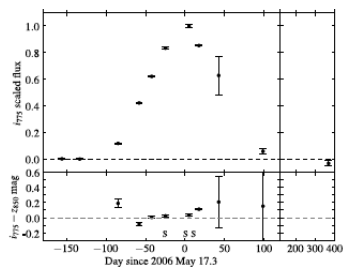
Asteroids



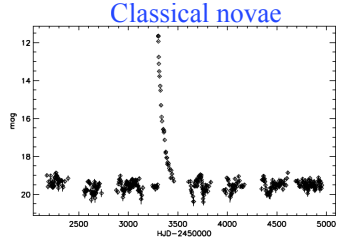
Dwarf novae



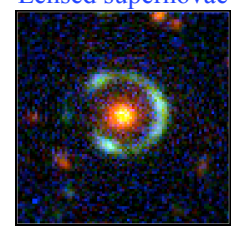
NEW THINGS??



Classical novae



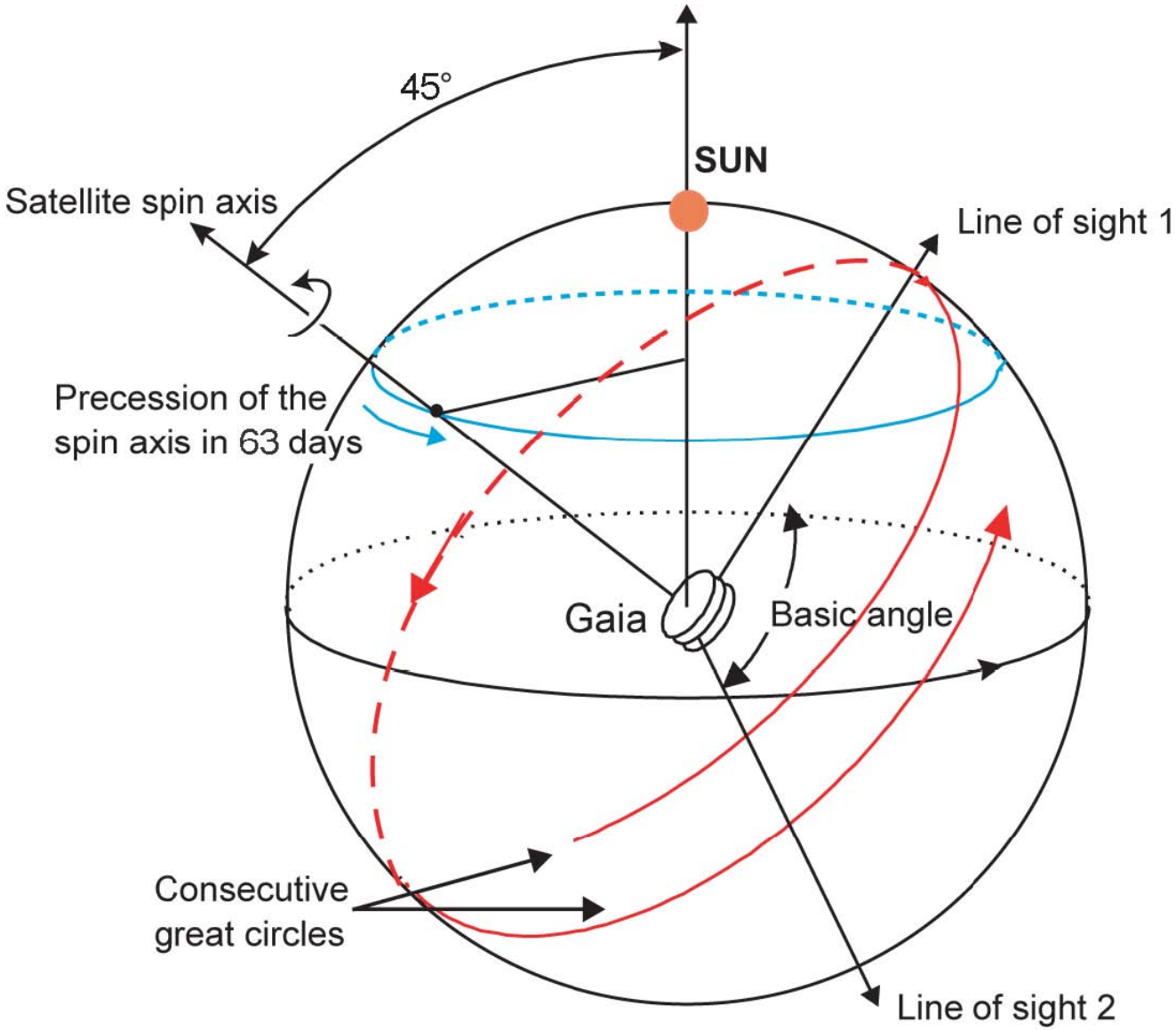
Lensed supernovae



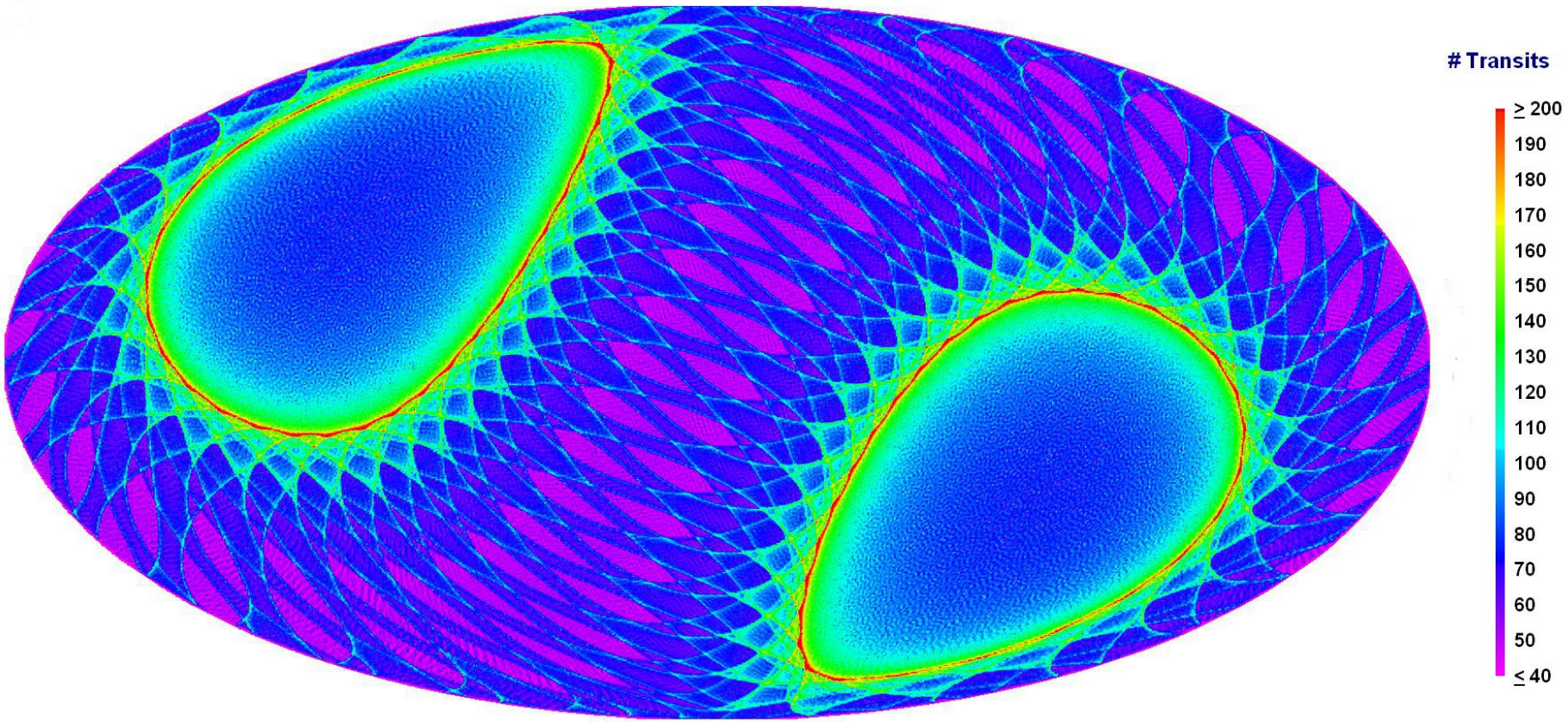
Lukasz Wyrzykowski, IoA Cambridge UK

Transients with the Gaia Mission

Survey strategy

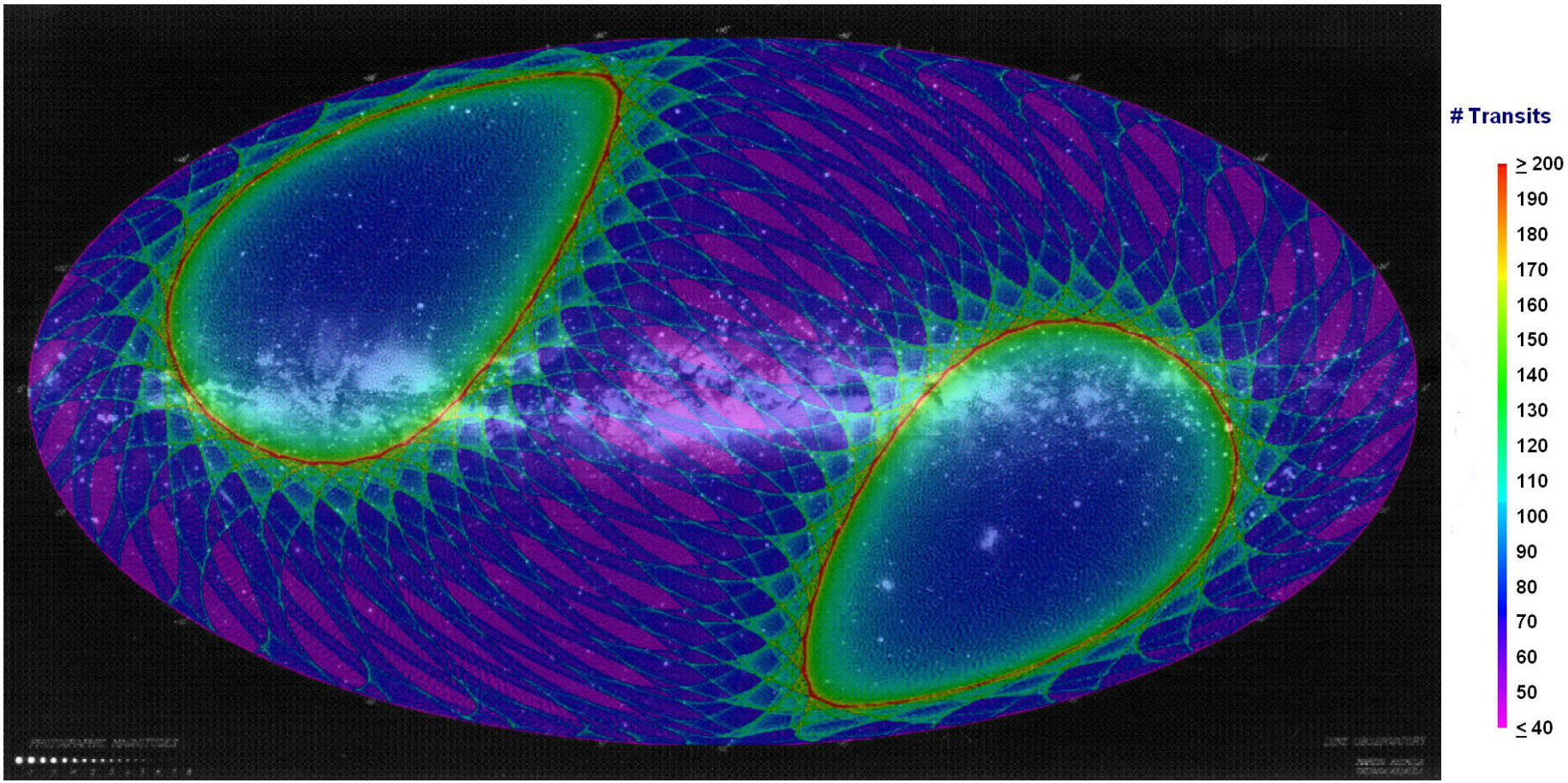


Survey strategy



t_0, $t_0 + 106$ minutes, $t_0 + 6$ hrs, $t_0 + 6$ hrs $+ 106$ minutes, repeated 10–30 days later

Survey strategy



t_0, $t_0 + 106$ minutes, $t_0 + 6$ hrs, $t_0 + 6$ hrs $+ 106$ minutes, repeated 10–30 days later

Hipparcos Catalogue statistics (1)

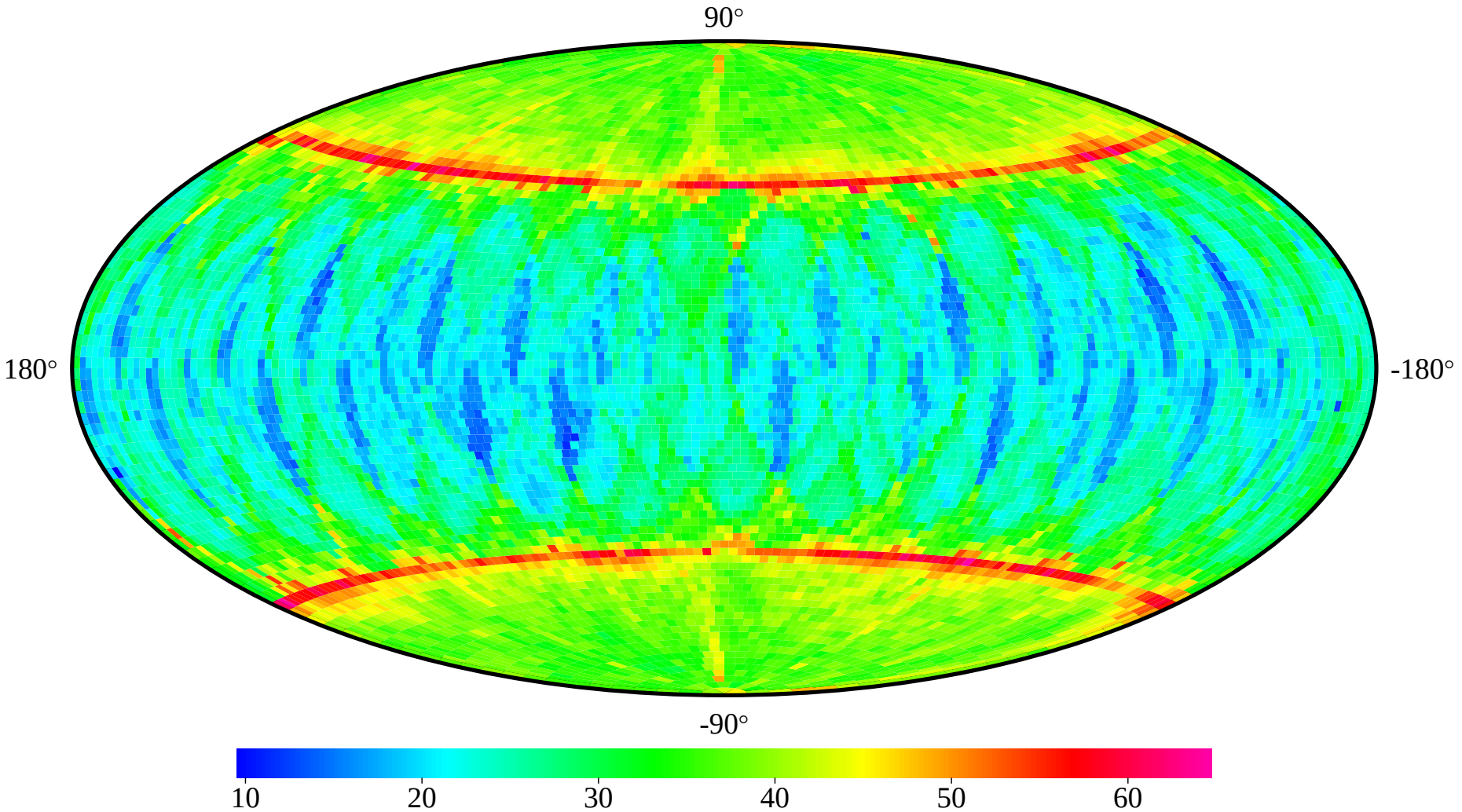


Image credit: ESA

Median number of observations (Ecliptic coordinates)

Hipparcos Catalogue statistics (1)

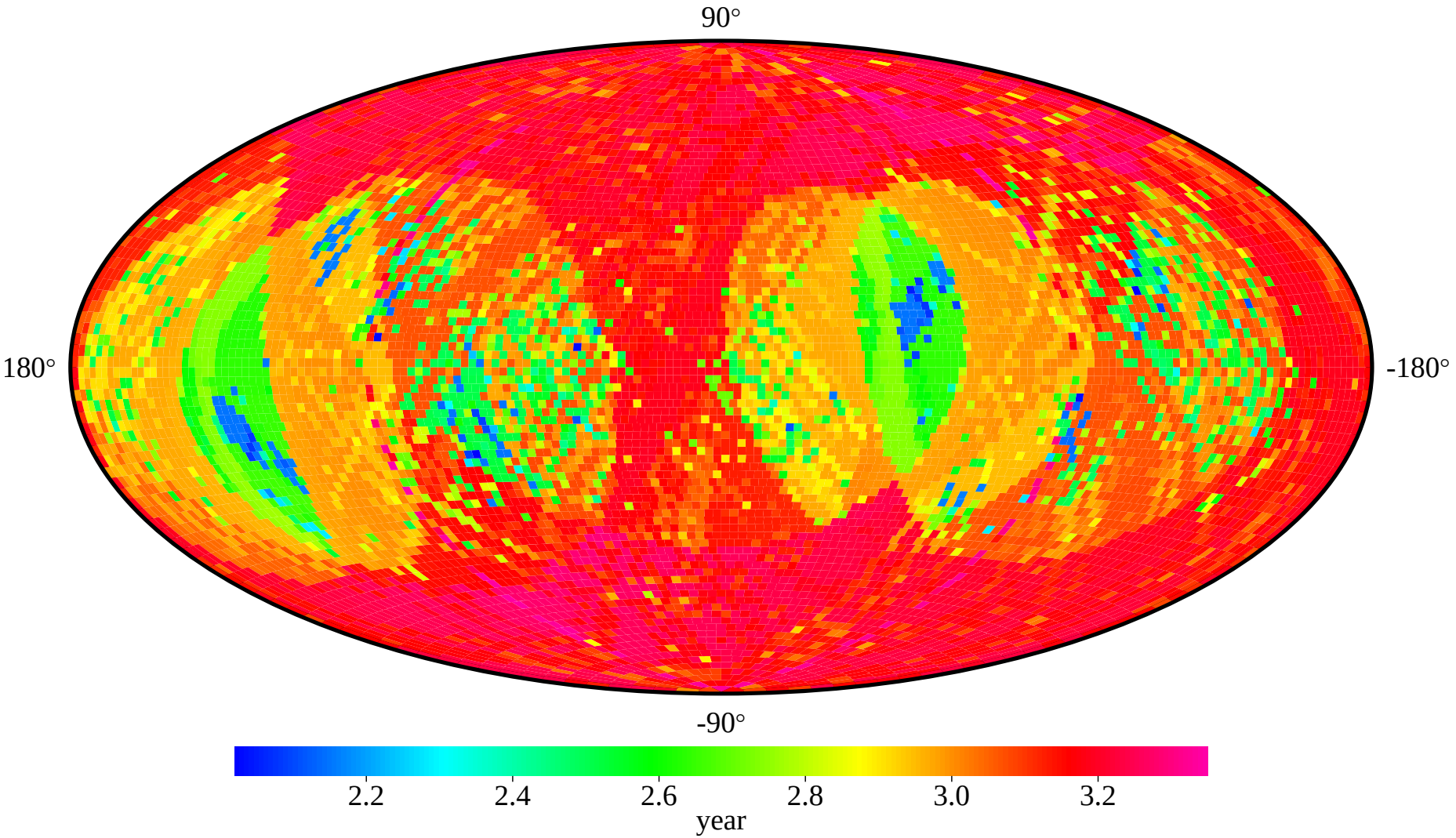


Image credit: ESA

Median duration between first and last observation (Ecliptic coordinates)

- scanning law, mission duration, observation interruptions

Hipparcos Catalogue statistics (1)

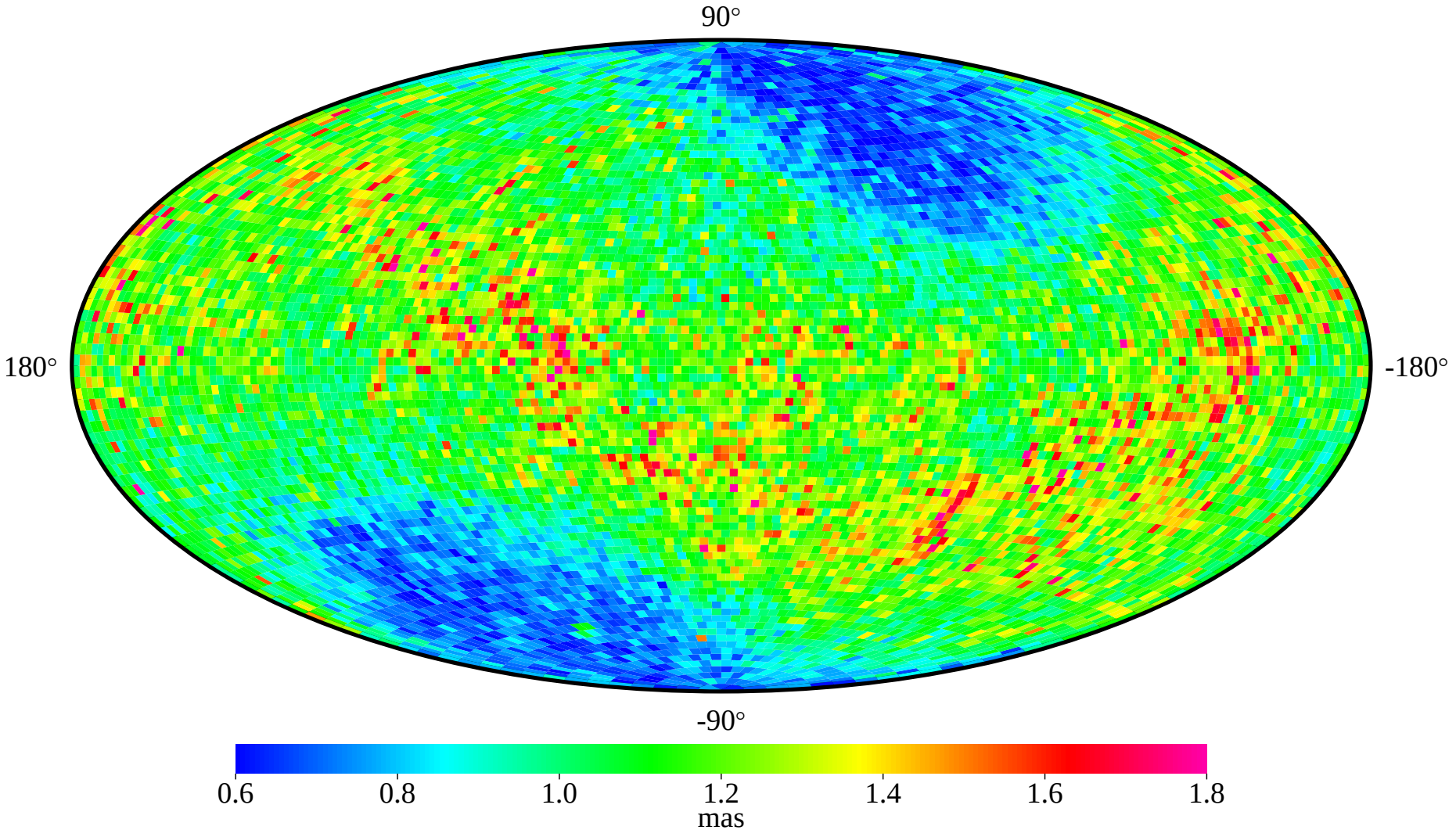


Image credit: ESA

Median standard error on ϖ (Equatorial coordinates)

Correlated errors

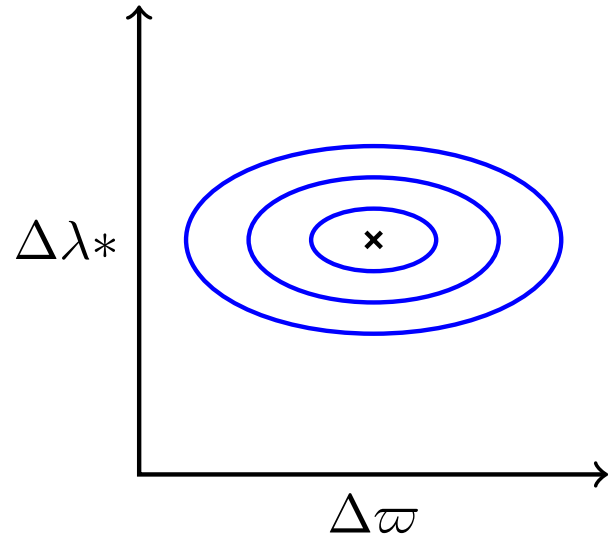
Covariance matrix for measured quantity $\mathbf{v}$:

$$\mathbf{C_v} = \mathrm{E}\left[(\mathbf{v} - \langle\mathbf{v}\rangle)(\mathbf{v} - \langle\mathbf{v}\rangle)^T\right] = \mathrm{E}\left[\Delta\mathbf{v}\Delta\mathbf{v}^T\right]$$

$$c_{ij} = \rho^i_j \sigma_i \sigma_j$$

Confidence region around $\langle\mathbf{v}\rangle$:

$$\Delta\mathbf{v}^T \mathbf{C}_\mathbf{v}^{-1} \Delta\mathbf{v} = z$$

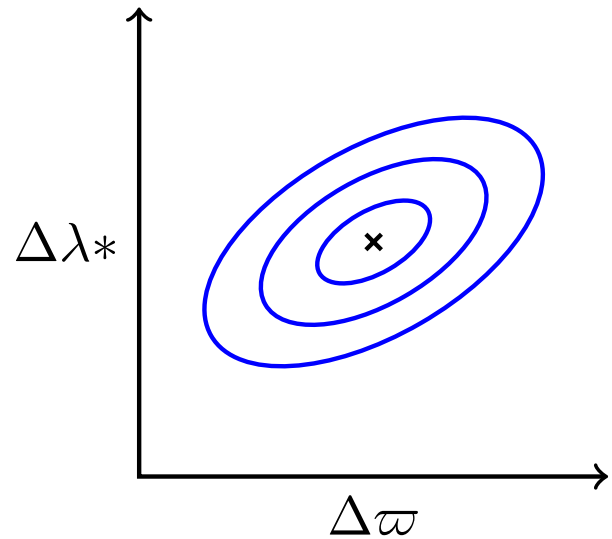


Transformation:

$$\Delta\mathbf{w} = \mathbf{M}\Delta\mathbf{v} \rightarrow \mathbf{C_w} = \mathbf{M}\mathbf{C_v}\mathbf{M}^T$$

Example:

$$\begin{pmatrix} \sigma^2_{\lambda*} & \rho^{\lambda*}_{\varpi}\sigma_{\lambda*}\sigma_{\varpi} \\ \rho^{\lambda*}_{\varpi}\sigma_{\lambda*}\sigma_{\varpi} & \sigma^2_{\varpi} \end{pmatrix}$$



Hipparcos Catalogue statistics (2)

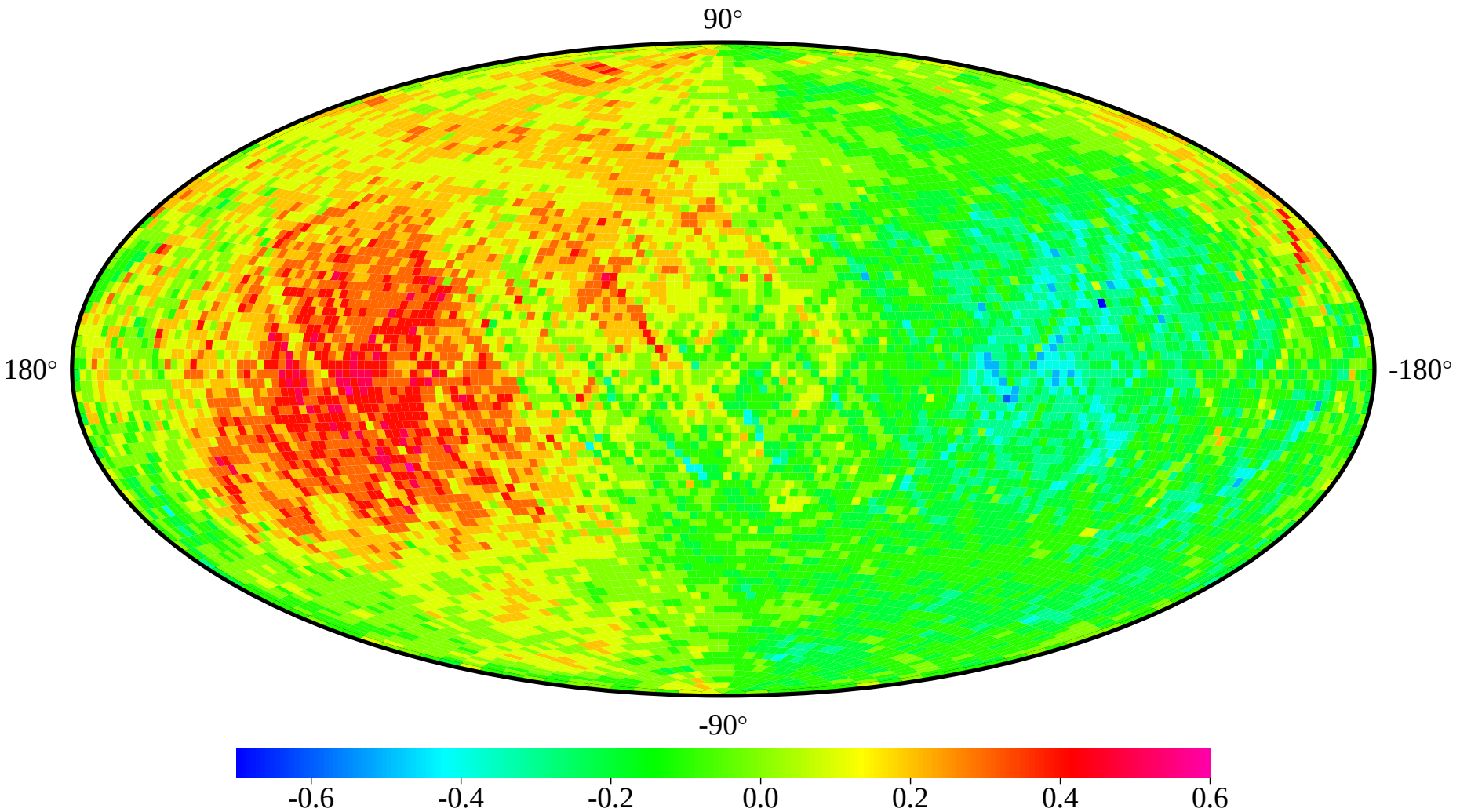


Image credit: ESA

Median correlation between $\lambda*$ and ϖ (Ecliptic coordinates)

- Systematics from asymmetric distribution of observations wrt position of Sun

Hipparcos Catalogue statistics (2)

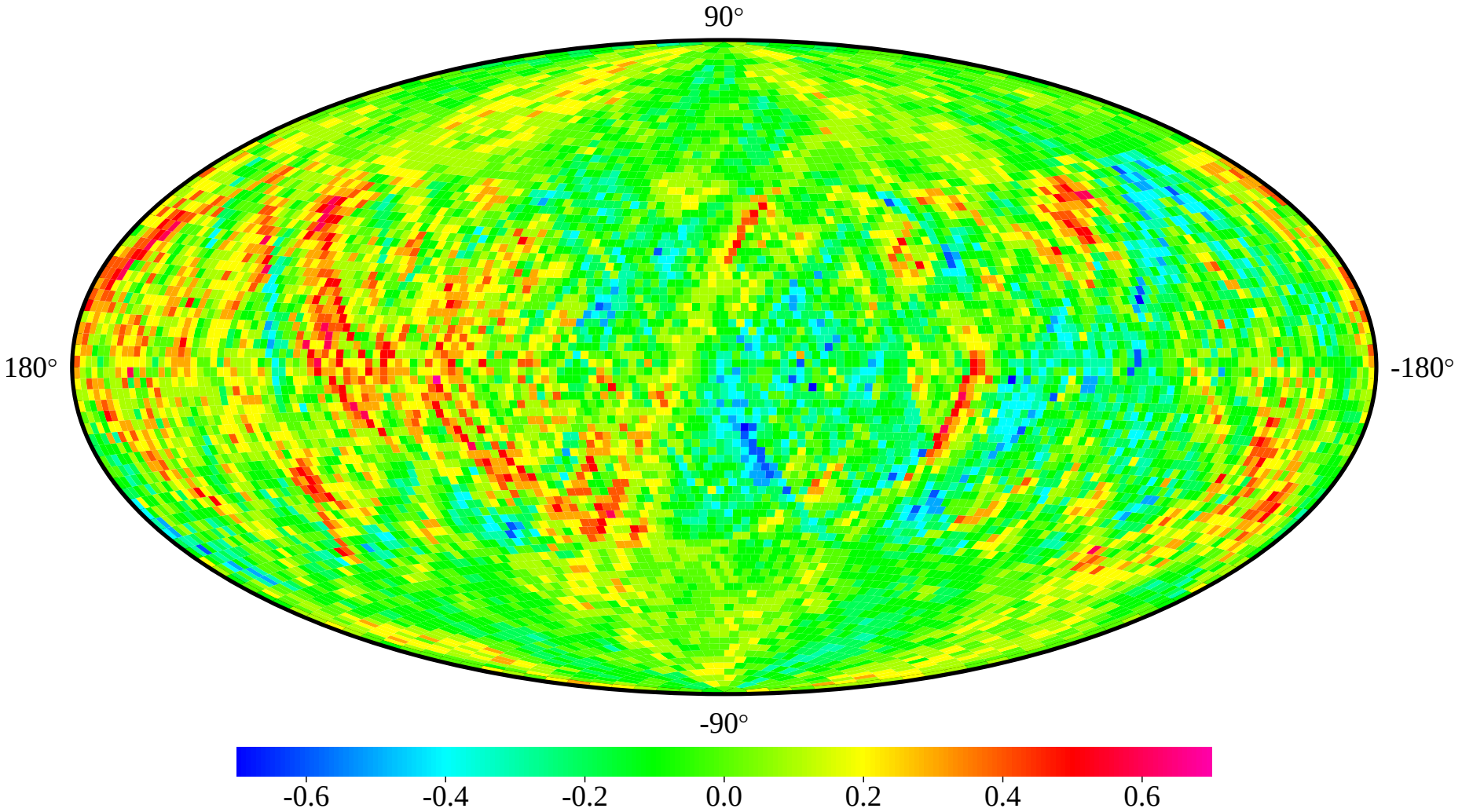


Image credit: ESA

Median correlation between $\mu_{\lambda*}$ and μ_β (Ecliptic coordinates)

Hipparcos Catalogue statistics (2)

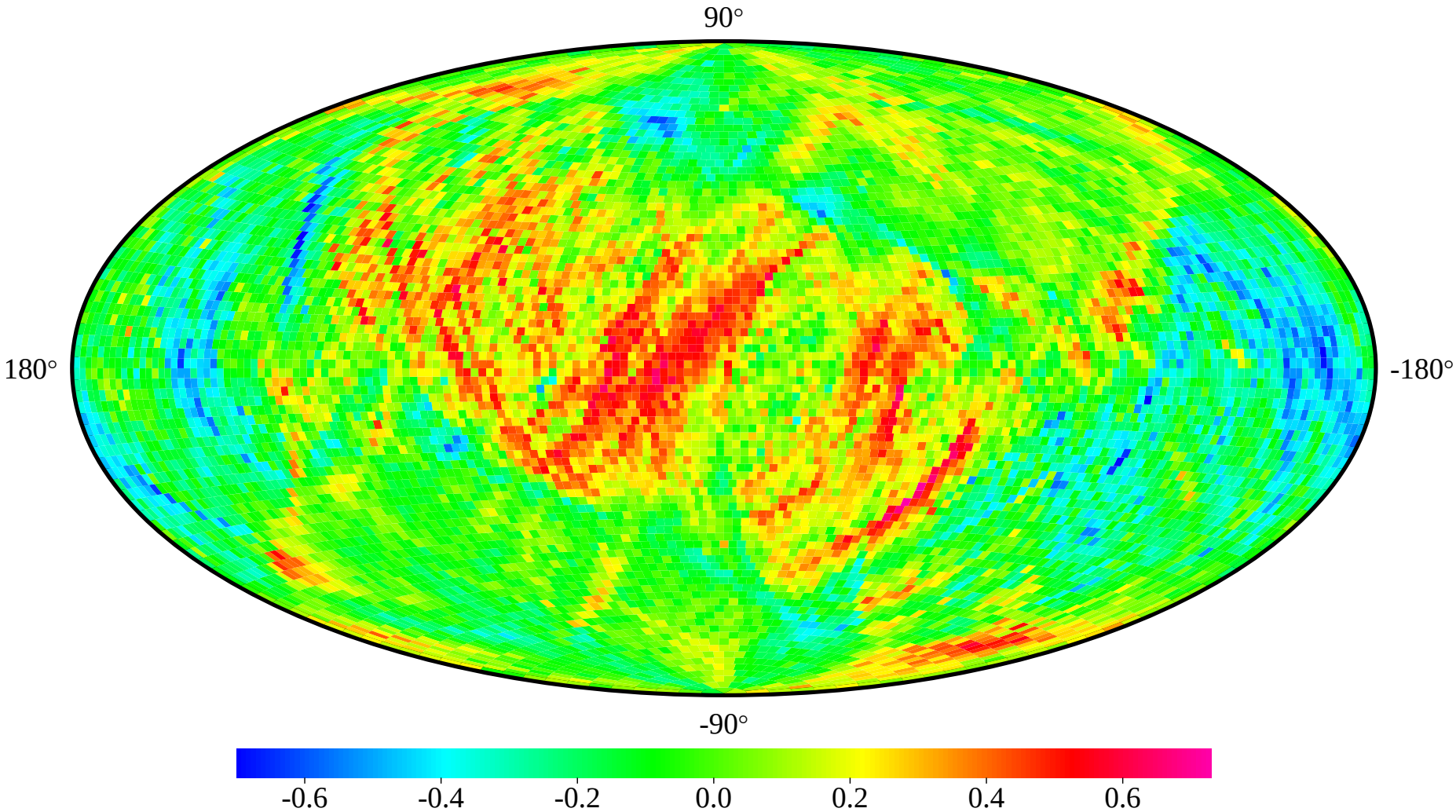
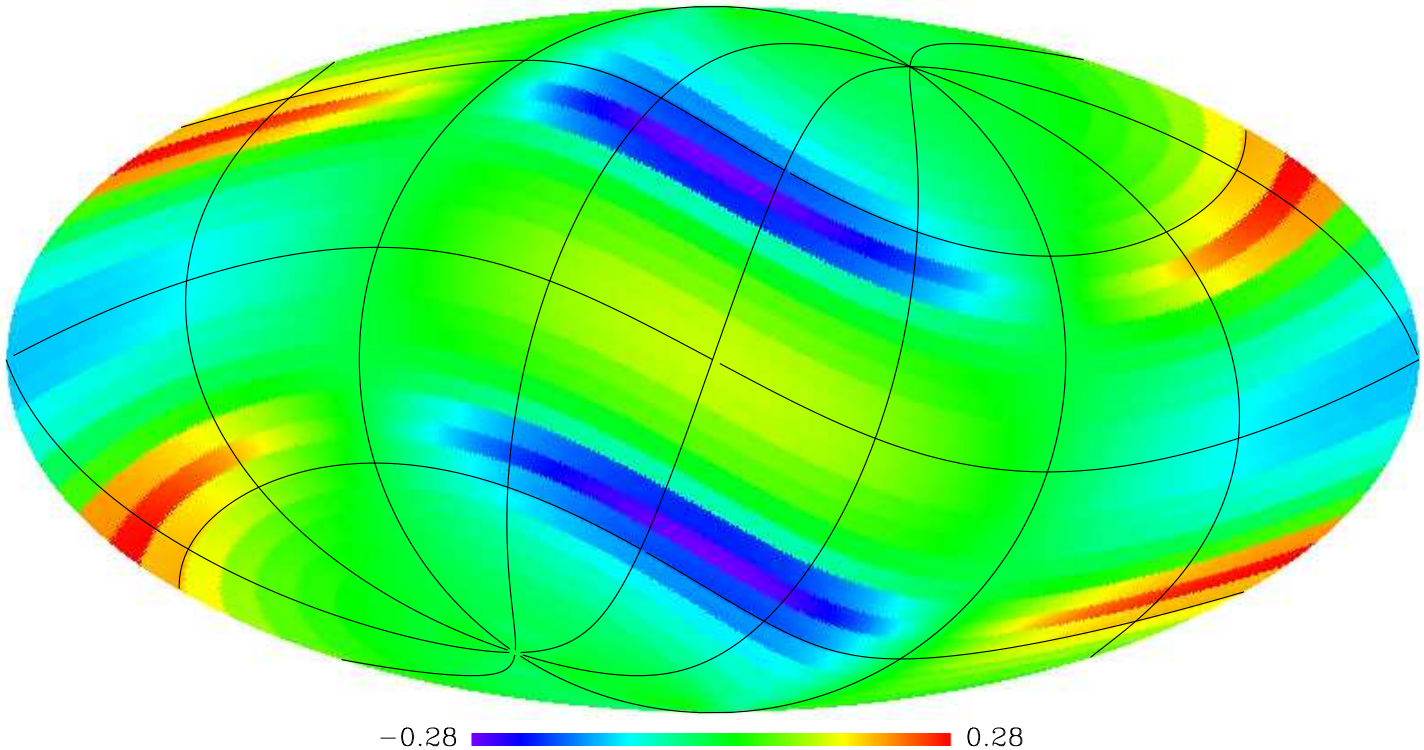


Image credit: ESA

Median correlation between $\mu_{\alpha*}$ and μ_δ (Equatorial coordinates)

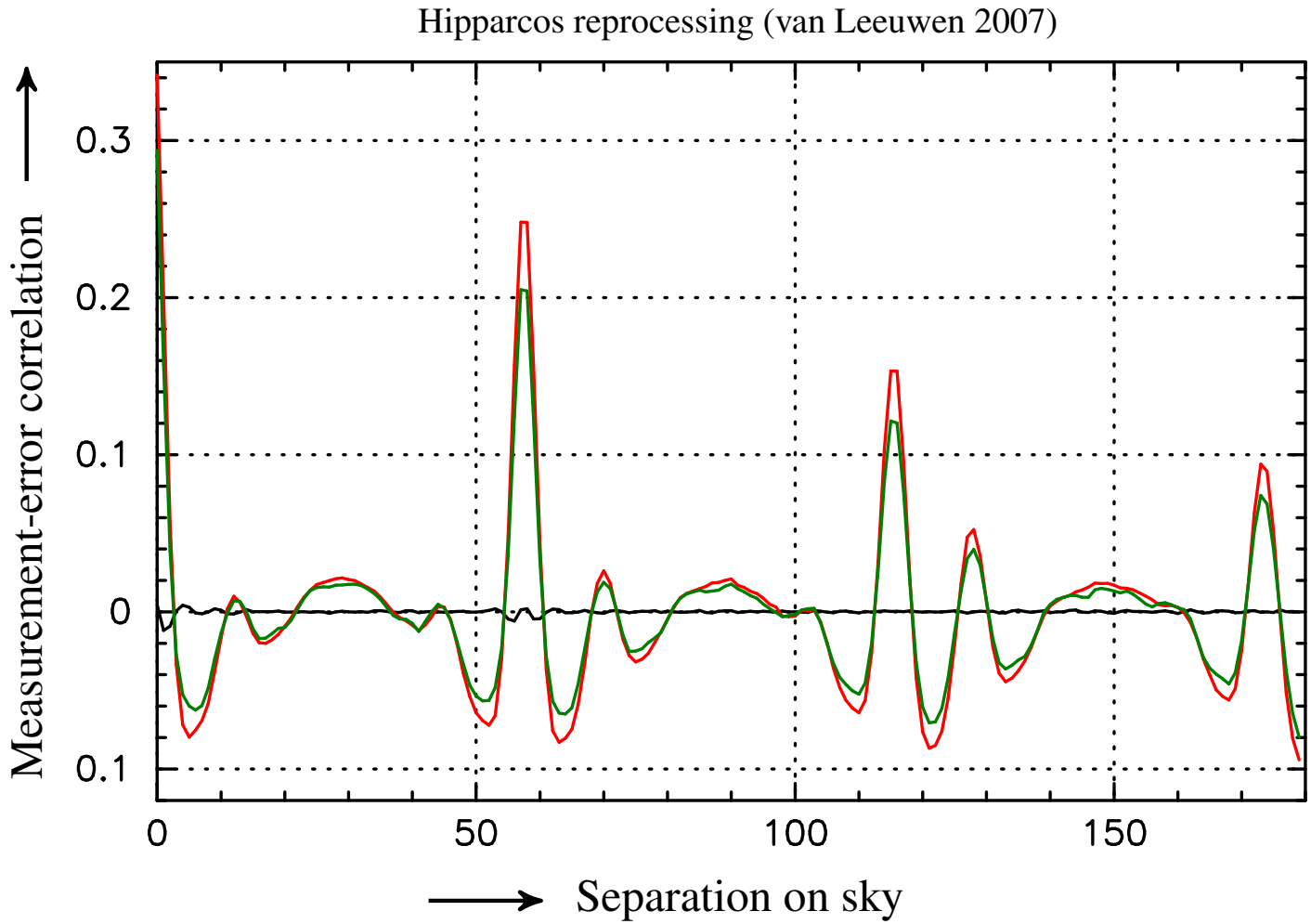
Effect of coordinate transformation



Correlation between $\alpha*$ and δ due to coordinate transformation (Ecliptic to Equatorial)

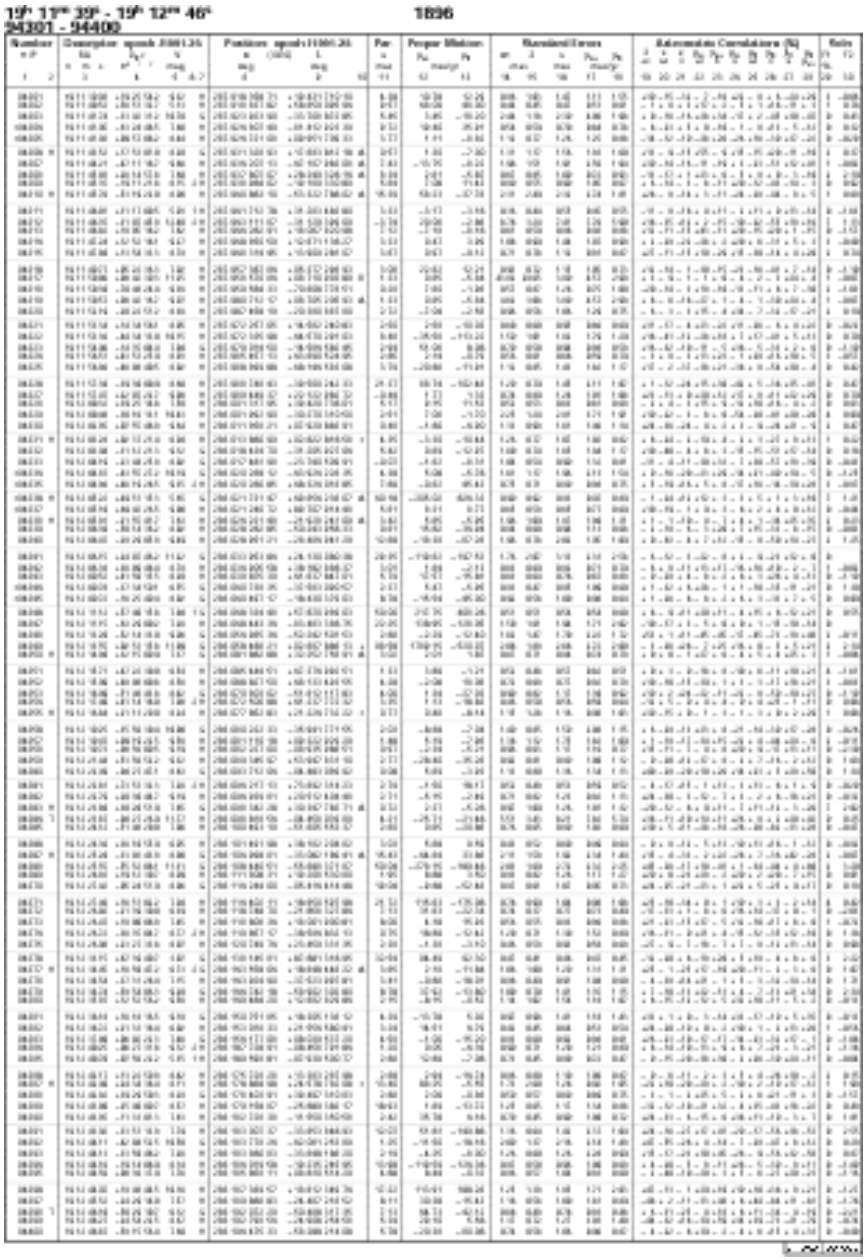
- Assume $\lambda*$ and β uncorrelated
- $\sigma_{\lambda*}/\sigma_{\beta}$ varies with β

Star to star correlations



How to deal with these in Gaia case: presentation by Berry Holl

Turn this



into. . .

The Milky Way from Gaia

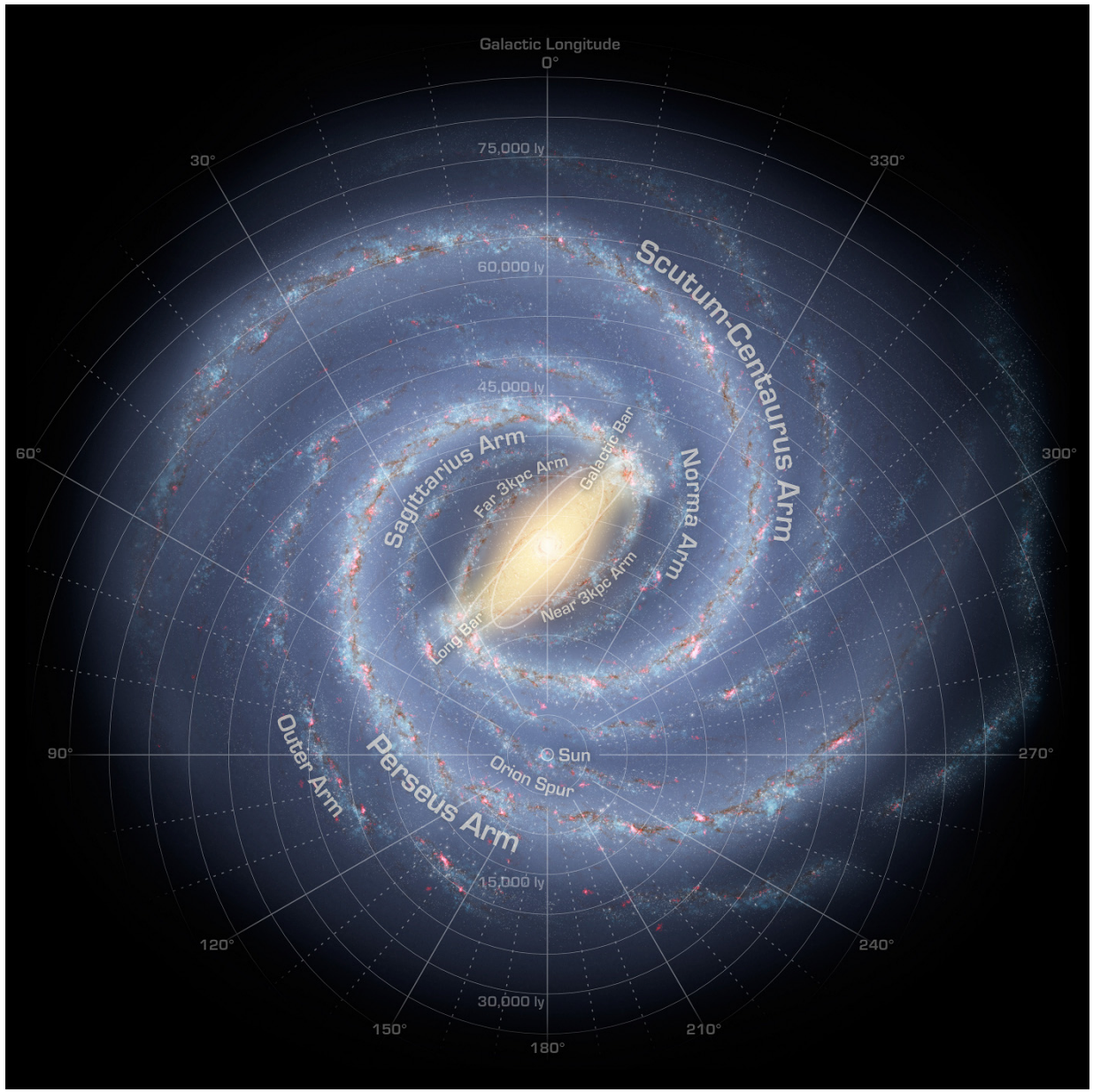


Image credit: NASA/JPL-Caltech/R. Hurt (SSC)

Determining the best Galaxy model

Goal

- Dynamical model of the Galaxy capable of 'explaining' the entire Gaia catalogue

Dynamical model

- gravitational potential
- distribution functions for each stellar population
  - probability distributions in mass, abundances, ages
- large number of parameters

Basic predictions:

$$f_\Theta(\mathbf{r}, \mathbf{v}, t|\text{stellar population})$$

1. *In what space* do we work to determine the parameter vector Θ?
2. *How* do we determine the parameter vector Θ?

Direct comparison in phase-space

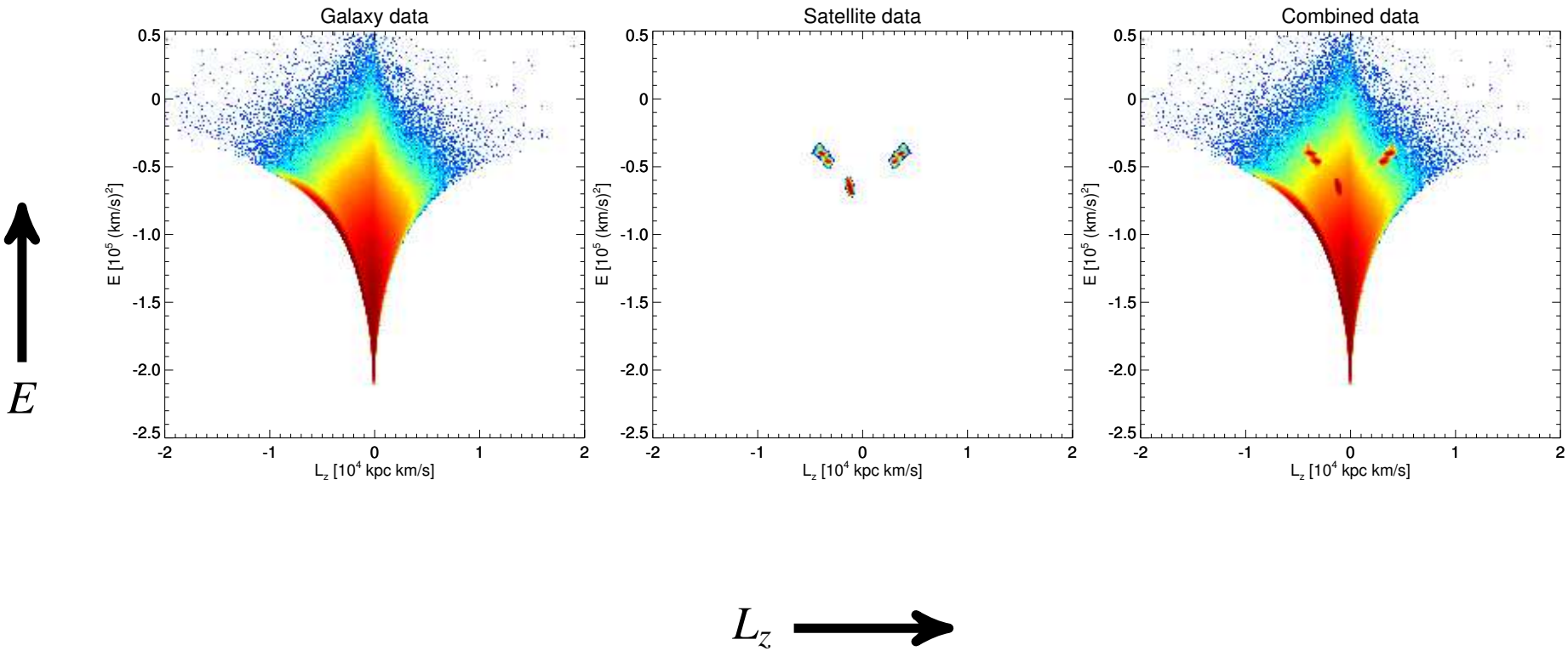
Directly comparing predicted $(\mathbf{r}, \mathbf{v})$ to observed values is most 'natural' approach. However:

- Effects of dust to be corrected
- Incomplete phase space data (lack of $v_{\rm rad}$)
- Parallax to distance conversion is non-linear ($r = 1/\varpi$)
  - For true and observed parallax ϖ_0 and ϖ:

$$E[\varpi] = \varpi_0 \quad \text{but} \quad E[\frac{1}{\varpi}] \neq \frac{1}{\varpi_0} \; !$$

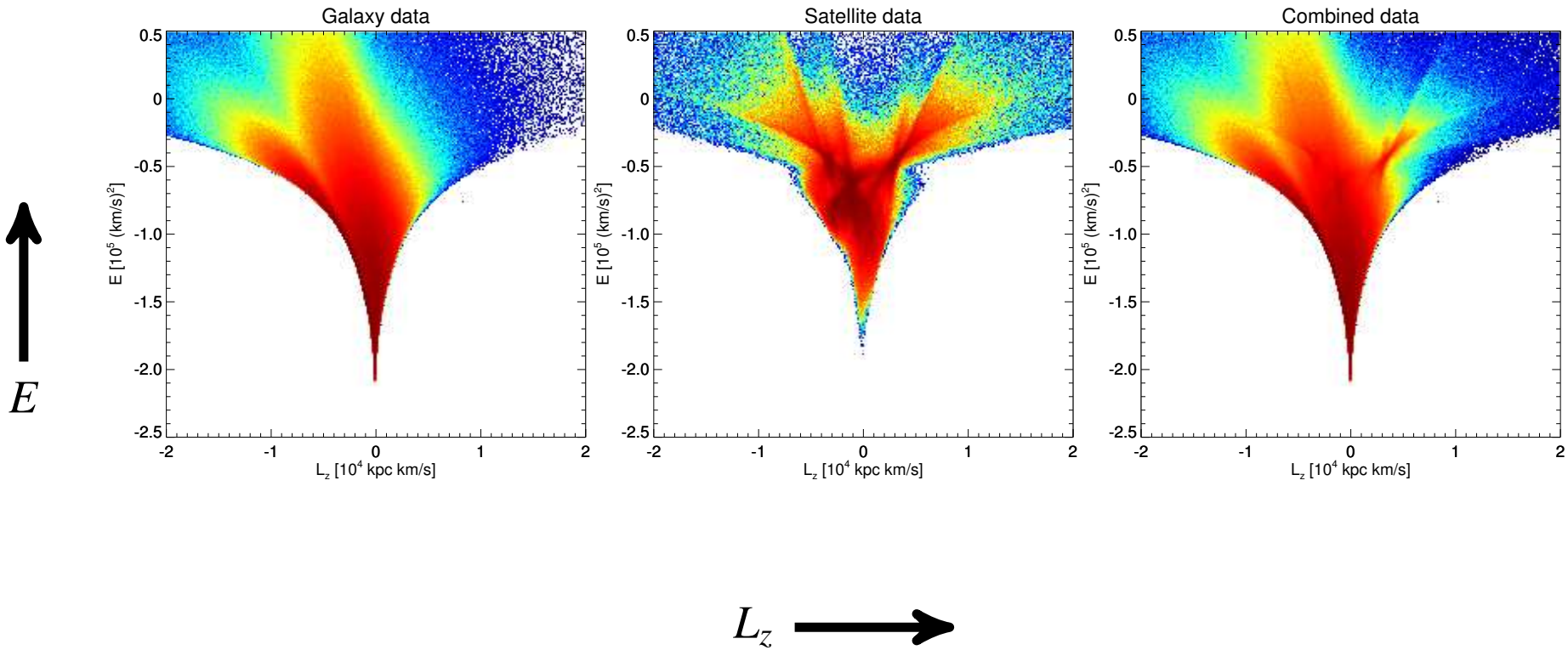
  - For large σ_ϖ/ϖ expect large biases and spurious features in $\mathbf{r}$, $\mathbf{v}$ E, L, etc
    - E and L both depend on $1/\varpi^2$
  - Selection on σ_ϖ/ϖ can introduce severe truncation biases
- Correlated observables and non-linear transformation of parallax will produce strongly non-Gaussian errors with complicated correlations

Direct comparison in phase-space



See: Brown, Velázquez & Aguilar, 2005, arXiv:astro-ph/0504243

Direct comparison in phase-space



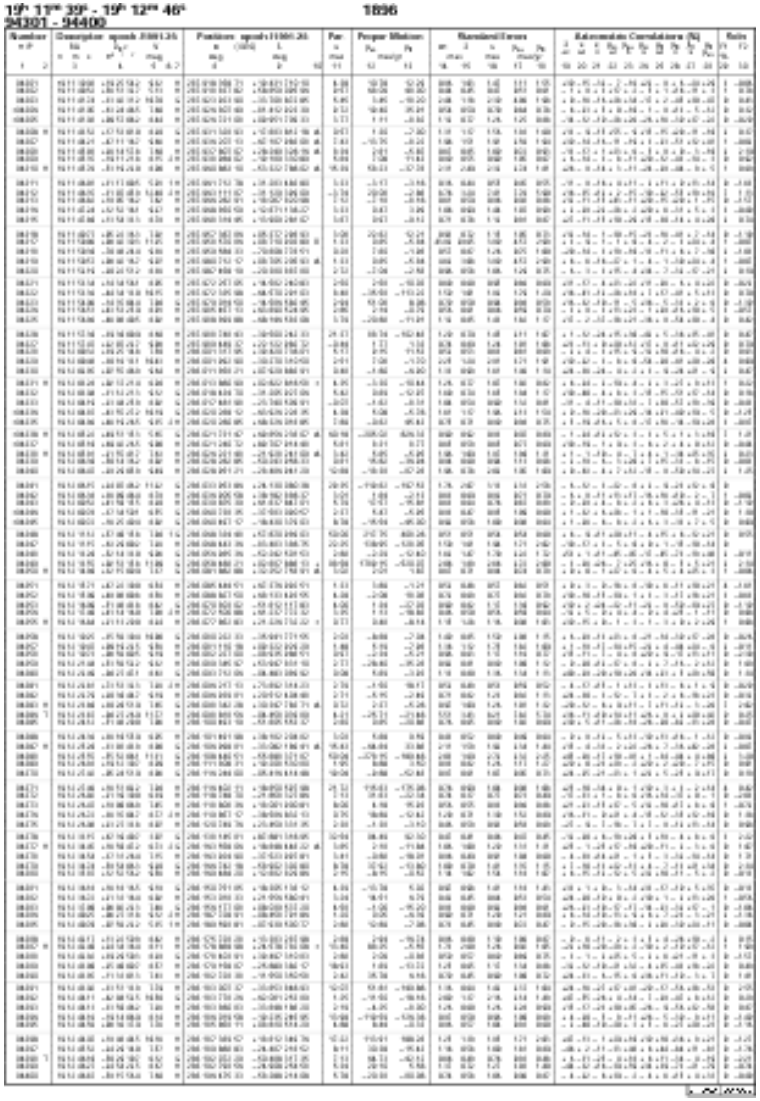
See: Brown, Velázquez & Aguilar, 2005, arXiv:astro-ph/0504243

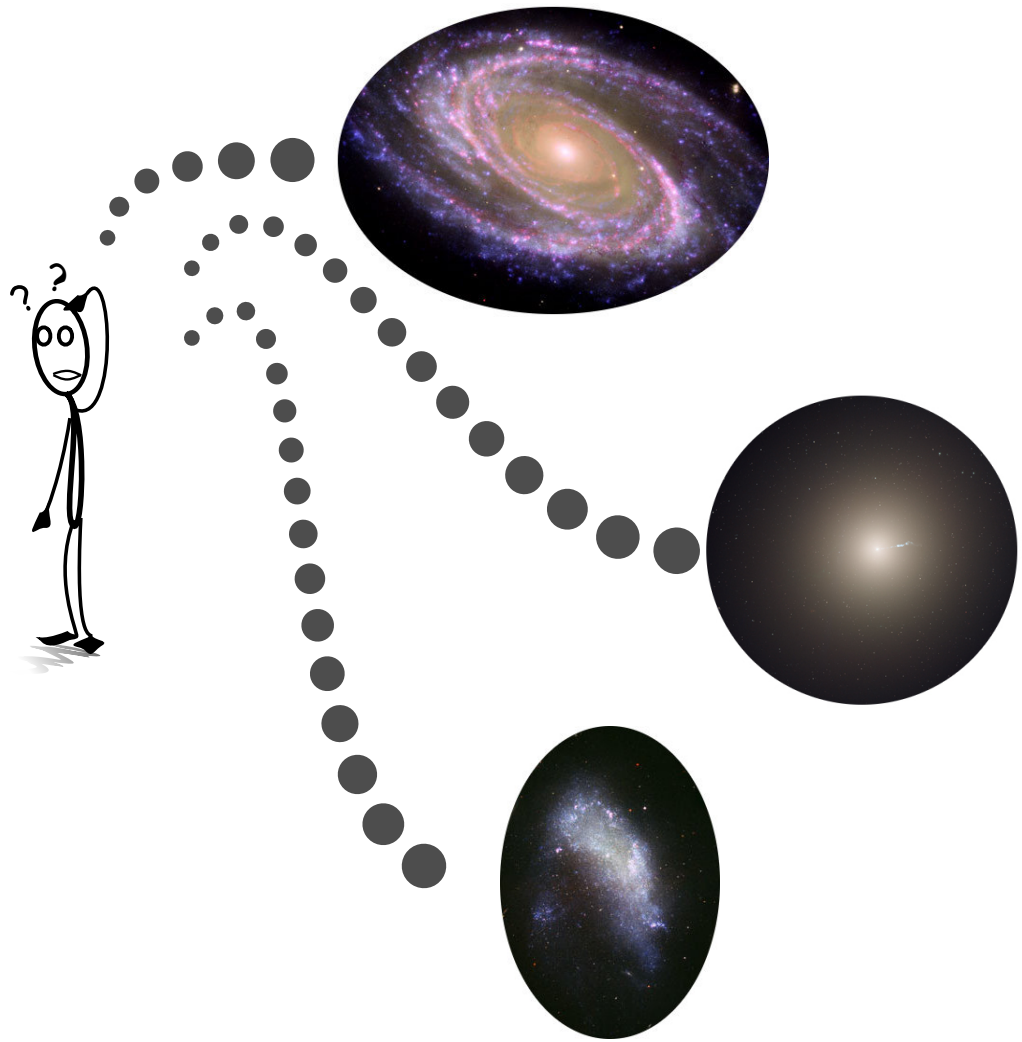
Project Galaxy model into the data space:

$$f_\Theta(\mathbf{r}, \mathbf{v}, t|\text{stellar population}) \rightarrow$$
$$g_\Theta(\alpha, \delta, \varpi, \mu_{\alpha*}, \mu_\delta, v_{\rm rad}, G, \text{colour}, A_V, [\text{M/H}], \ldots |\text{stellar population})$$

- No non-linear parallax transformation
- All parallax data can be used (including negative parallaxes)
- Easy accounting for incomplete phase space data (i.e., no $v_{\rm rad}$)
- Selection effects and incompleteness can be modelled
- Extinction can be modelled
- Correlations in the measurement errors are more easily accounted for in the data space

‘Deciding’ on the best parameter values





‘Deciding’ on the best parameter values

Bayesian inference through maximization of likelihood:

$$\phi = \sum_{\text{all stars}} \ln p(\alpha, \delta, \varpi, \mu_{\alpha*}, \mu_\delta, v_{\text{rad}}, G, \text{colour}, A_V, [\text{M/H}], \ldots | \text{stellar population}, \Theta)$$

- Allows precise hypothesis testing
- Result will be a probability distribution over the parameter space
  - not all aspects of Galaxy model will be uniquely determined

Challenges:

- Large amount of data and large number of model parameters
- Full catalogue comparison needed
  - constraints from all sky directions and full Gaia volume
  - Star-to-star correlations to be accounted for

Facilitating precise hypothesis testing

Three proposals in Hogg & Lang (2011, arXiv:1008.0738)

- Careful definition of the uncertainty of each catalogue entry
- Provide a sampling of possible Gaia catalogues
  - approximate the full covariance matrix of the catalogue
- Expose the likelihood function
  - provide access to the full covariance matrix

Definition of catalogue entry uncertainty

- Catalogue entry $\mathbf{v}^T = (\alpha, \delta, \varpi, \mu_{\alpha*}, \mu_\delta, v_{\rm rad}), \mathbf{C}^{-1}$
- Two hypotheses (alternatives) for $\mathbf{v}$; $\mathbf{V}_1$ and $\mathbf{V}_2$
- define

$$\Delta\chi^2 \equiv (\mathbf{V}_2 - \mathbf{v})^T \mathbf{C}^{-1} (\mathbf{V}_2 - \mathbf{v})^T - (\mathbf{V}_1 - \mathbf{v})^T \mathbf{C}^{-1} (\mathbf{V}_1 - \mathbf{v})^T$$

- *Define* $\mathbf{v}$, $\mathbf{C}^{-1}$ such that the result is as close as possible to what you would have computed for $\Delta\chi^2$ in the *raw image pixels*, marginalizing over all nuisance (calibration) parameters.

- For practical example see 'Spectroperfectionism': Bolton & Schlegel, arXiv:0911.2689
- Not far from existing plans for publication of Gaia data (see also Hipparcos Catalogue)
- How to do this transparently for the great variety of catalogue entries?
  - differences in 'distance to the raw data'

Make $K + 1$ Gaia catalogues

- Release one 'primary' catalogue
- Other K catalogues sample posterior distribution of possible catalogues
  - astrometric, radial velocity, and astrophysical parameter variations
  - calibration parameter variations
  - allow catalogues of different complexity (eg., binary vs. single star)
- Average of a quantity over the K samples should approximate marginalization over all probabilistic quantities

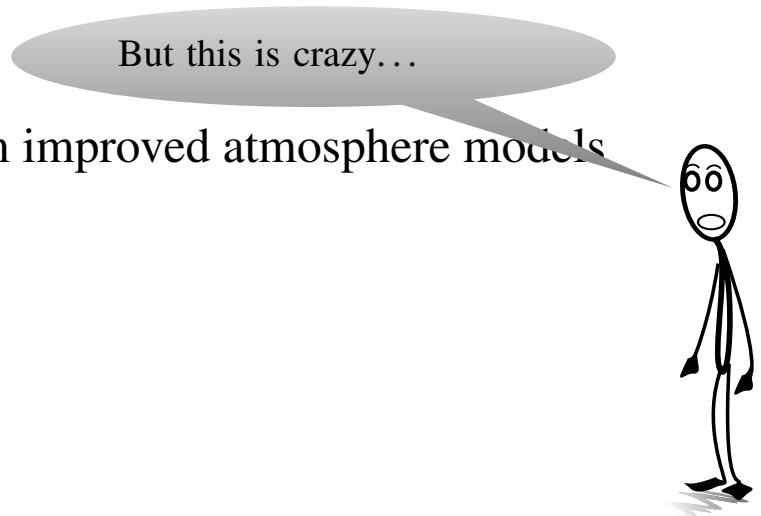
- Approximates the full covariance matrix of the catalogue
  - star-to-star correlations can be calculated
- For a practical example see presentation by Berry Holl
- How large should K be?
- How to sample 'catalogue space' efficiently such that the result is representative of the posterior distribution?
  - allowing for complexity variations is challenging

Exposing the likelihood function

- Allow users to submit a Δcatalogue
  - difference between primary Gaia catalogue and an alternative
  - Δ can be for science data and/or calibration parameters
- Return $\Delta \log \mathcal{L}$
  - likelihood re-evaluated against the raw pixel data
  - provides access to any covariance matrix element
  - computational burden on the user (performs his/her own uncertainty analysis)
- Allows handling different complexity choices
  - alternative deblending in crowded regions
  - single vs. multiple star solutions
  - incorporation different priors
  - different astrophysical parameters from improved atmosphere models
  - ...

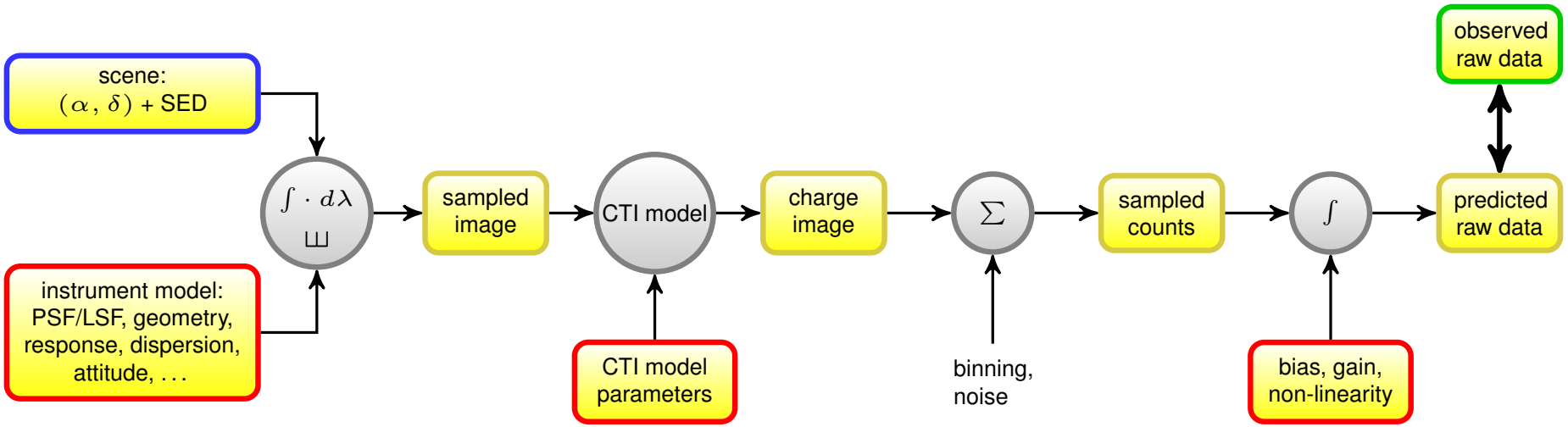
Exposing the likelihood function

- Allow users to submit a Δcatalogue
  - difference between primary Gaia catalogue and an alternative
  - Δ can be for science data and/or calibration parameters
- Return $\Delta \log \mathcal{L}$
  - likelihood re-evaluated against the raw pixel data
  - provides access to any covariance matrix element
  - computational burden on the user (performs his/her own uncertainty analysis)
- Allows handling different complexity choices
  - alternative deblending in crowded regions
  - single vs. multiple star solutions
  - incorporation different priors
  - different astrophysical parameters from improved atmosphere models
  - ...



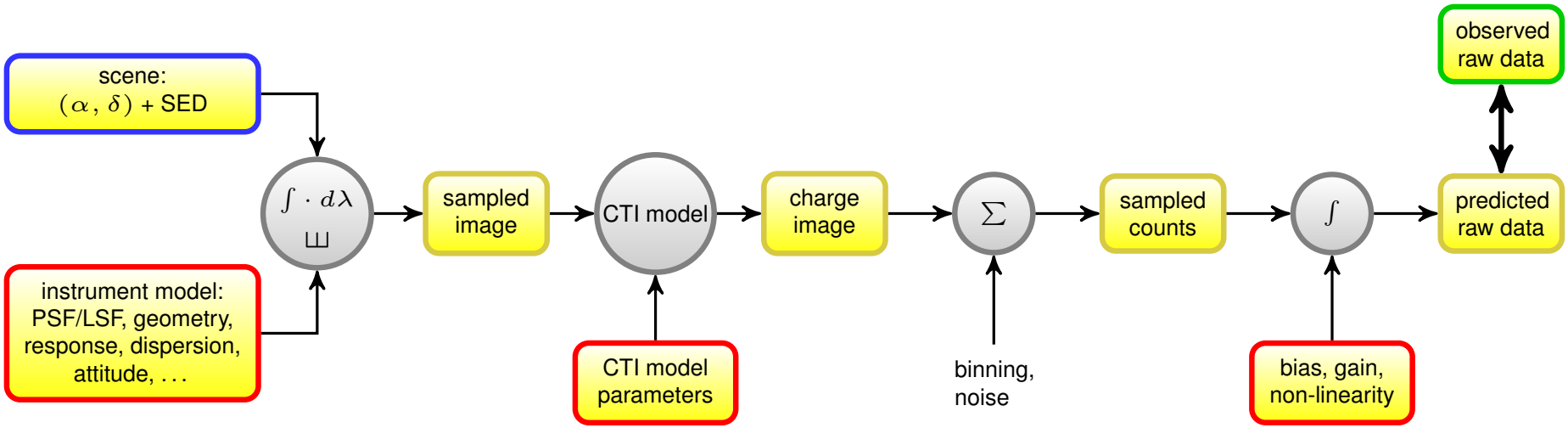
Exposing the likelihood function

(Parts of) Gaia data processing chain already use forward modelling:



Exposing the likelihood function

(Parts of) Gaia data processing chain already use forward modelling:

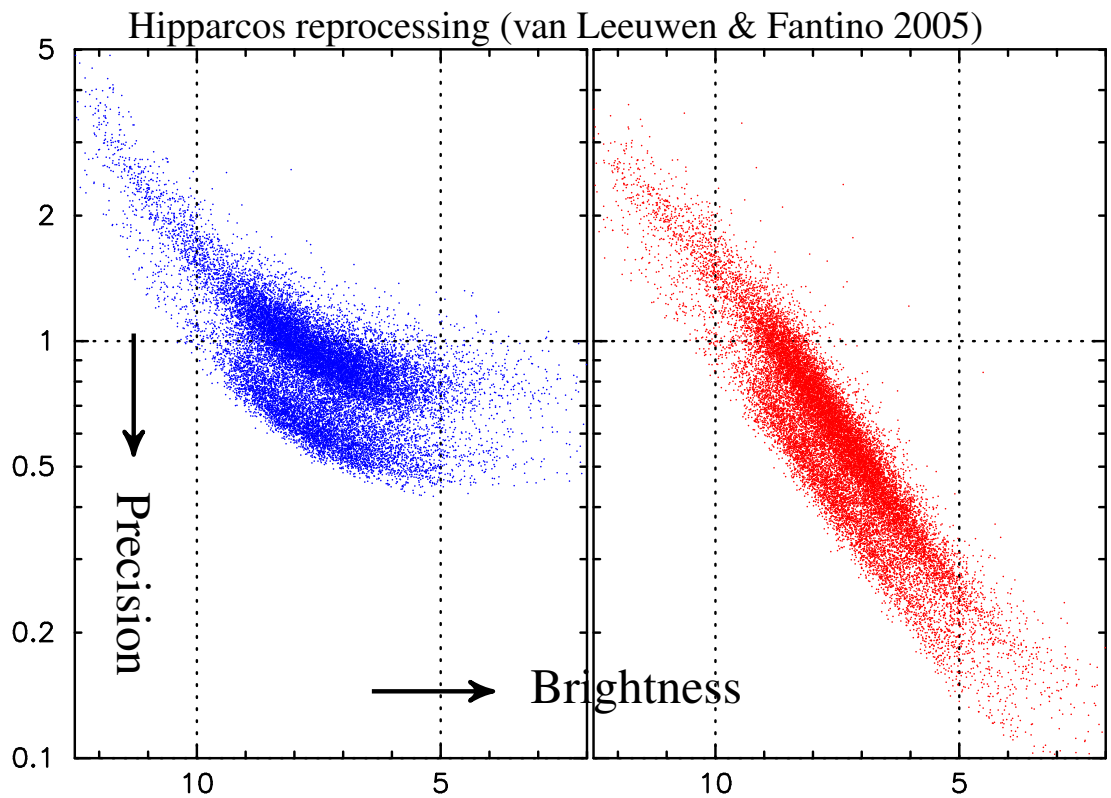


Why not publish this machinery? Think 2020 and beyond!

Preserve the raw data and processing software

Data curation

- All raw data
  - ∼ 60 TB uncompressed
- Calibration data and models
- Intermediate data products
- All processing software
- Implement data lineage concept

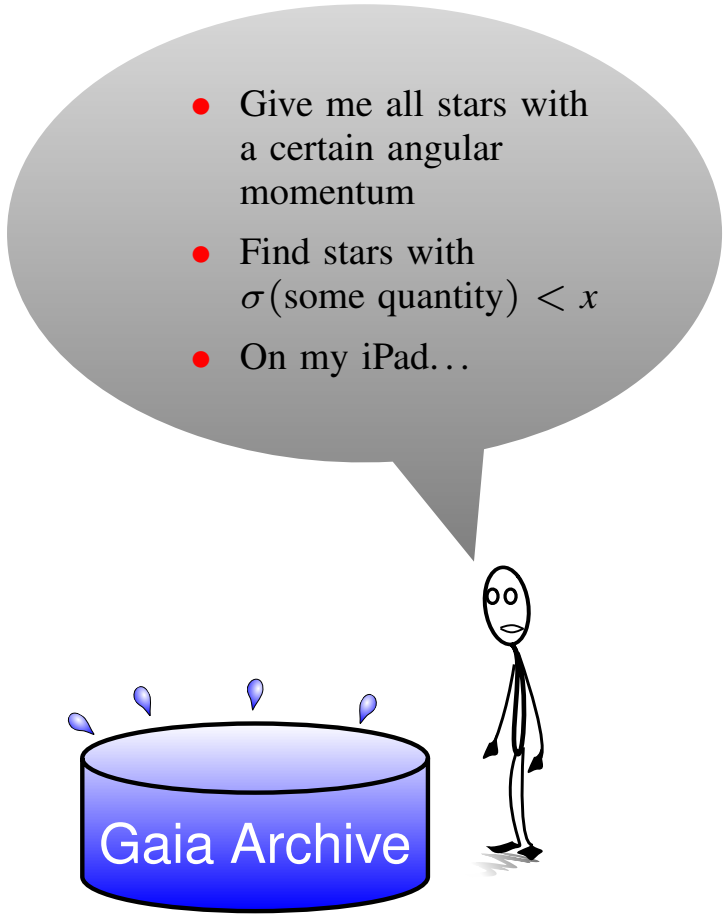


Science goals

- Raw data reprocessing based on better algorithms, better calibration models etc
- Alternative processing of specific stars, groups of stars, or even entire catalogue
- Reprocessing data based on new and independent information

Database and data archive

- General public oriented with advanced features for professionals
- Allow arbitrary queries
  - with option to account for star-to-star covariances
- Data mining
- *Visualisation* (data space is $>$ 10-dimensional!)
- Transparent integration with other catalogues or sky surveys
- Access with variety of ‘devices’
- All of this has to be fast of course . . .



Living archive concept

Gaia catalogue and archive released in $\sim$ 2021 should not be 'final'

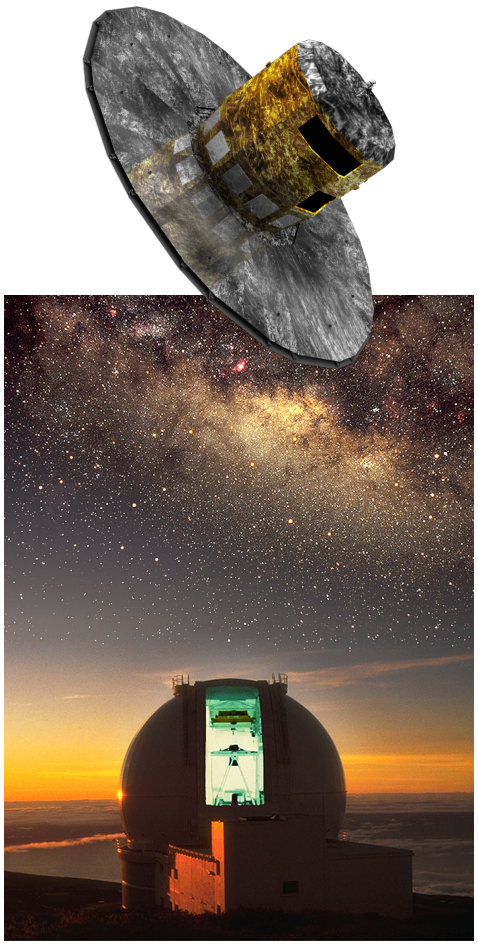
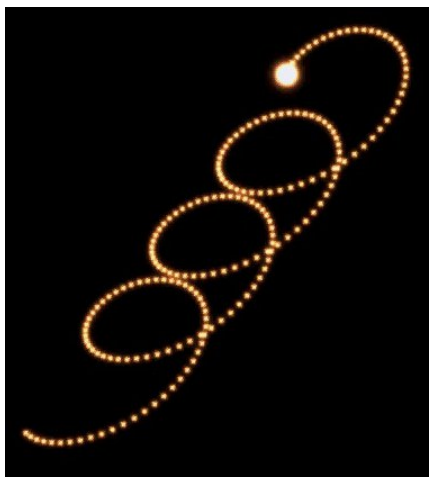
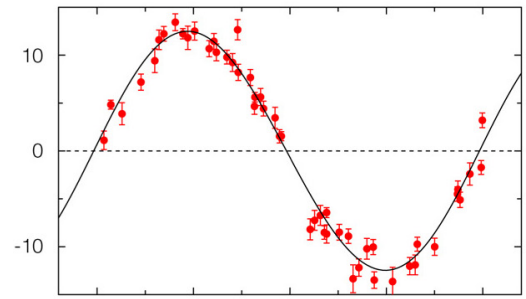


Photo Courtesy of the Isaac Newton Group of Telescopes, La Palma

- Updates should be allowed so as to incorporate:
  - updated classification or parametrization of stars
  - better distance estimates for faint stars
  - ground-based follow-up observations
  - independent information on, e.g., double stars
- Implications for maintenance, quality, security, keeping mirrors in sync



+



Bring the processing to the data

- Szalay (Sloan Digital Sky Survey) has advocated this for large archives
  - allow arbitrarily complex processing of archive data
  - example: dynamical model of the Milky Way that best explains the catalogue

- Virtualisation (O'Mullane, ESAC) could allow users a virtual machine *in the Data Centre with the Archive*
  - code what you want and specify how you want run it

Let's work to make the Gaia archive future proof!

- Gaia will provide an unprecedented stereoscopic map of our Milky Way and the nearby universe
  - 1 billion stars, 300 000 solar system objects, millions of galaxies, 500 000 quasars, 10 000 exo-planets, ...
  - catalogue 'finished' in 2021
- It will be *the* astronomical data archive for decades to come
  - tremendous discovery potential when combined with other archives
- Research or invest effort in the following:
  - keep raw data, calibration data, and processing software available
  - facilitate reprocessing
  - make the archive 'live'
  - bring the processing to the data
  - precise hypothesis testing agains the raw image pixels