Gaia Data Processing: the challenges

F. Mignard

Observatory of the Côte d'Azur, Nice.

Outline

- The Gaia data
- Overview of the processing
- Dependencies & Complexity: two illustrations

The Gaia Data

Volume and Time sampling

Sources of data

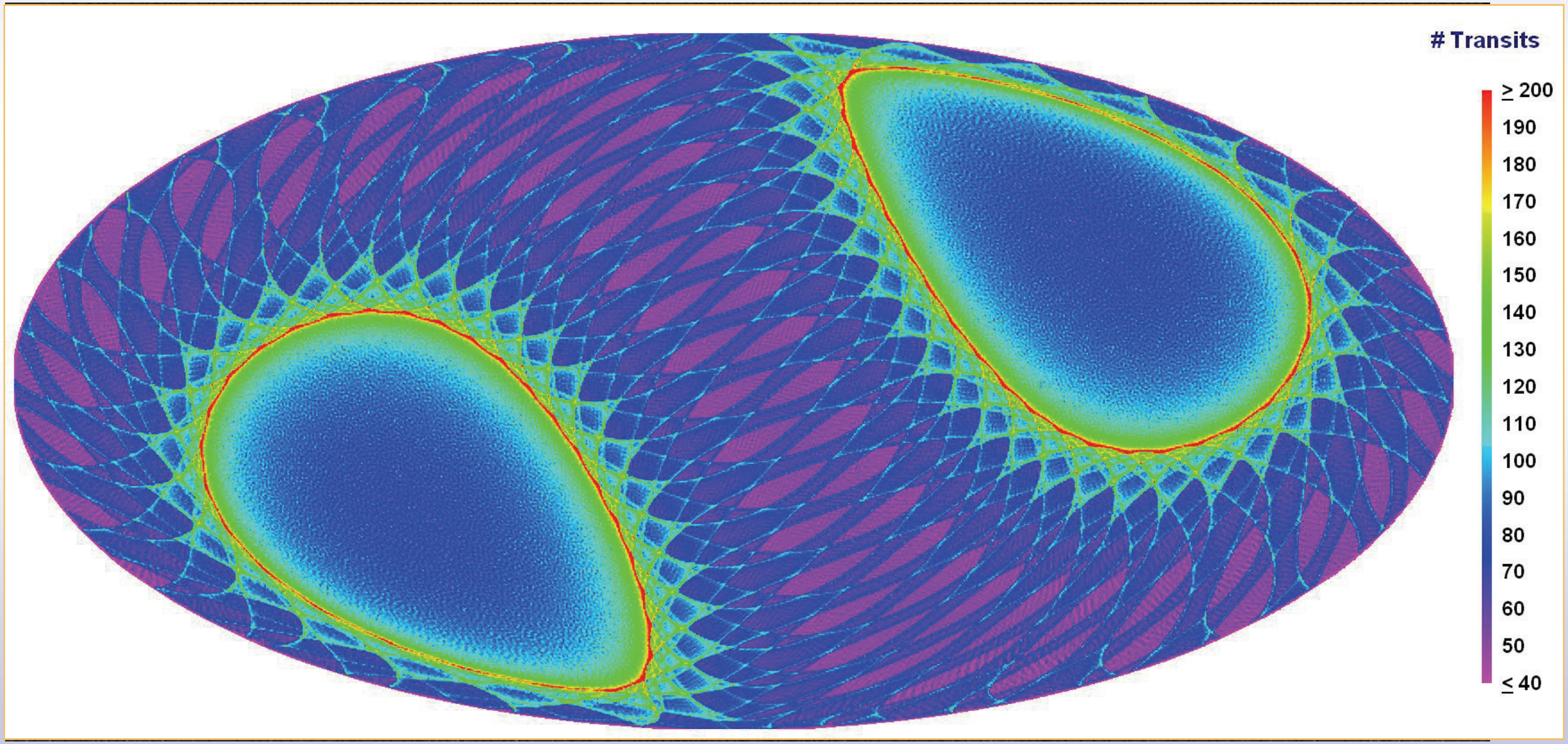
- Gaia has three instruments with three data flow
  - Astrometric CCDs
  - Photometric CCDs in the BP/RP bands
  - Spectroscopic data
- Data is organised in form of telemetry packets
  - astrometry & photometry
  - spectroscopy
  - one must also add house-keeping data and orbit data
- DPAC provides also auxiliary data from the ground

Raw data volume

- Main field
  - Astrometric data
    - Sky-mappers :: 14 CCds fully read
    - Astrometric CCDs :: 62 CCDs read with windows
  - Photometric data in BP/RP
    - photon counts of dispersed images :: 14 CCDs read with wide windows
- Spectroscopic field
  - RVS spectra :: 12 CCDs read with ultra wide windows
- Additional data
  - on-board metrology (time, WFS, BAM)
  - on-board attitude and detection data
- 1 billion source ~ 25,000 /deg^2
  - average makes sense for DPAC, not for the on-board S/W
  - time average more important than space average

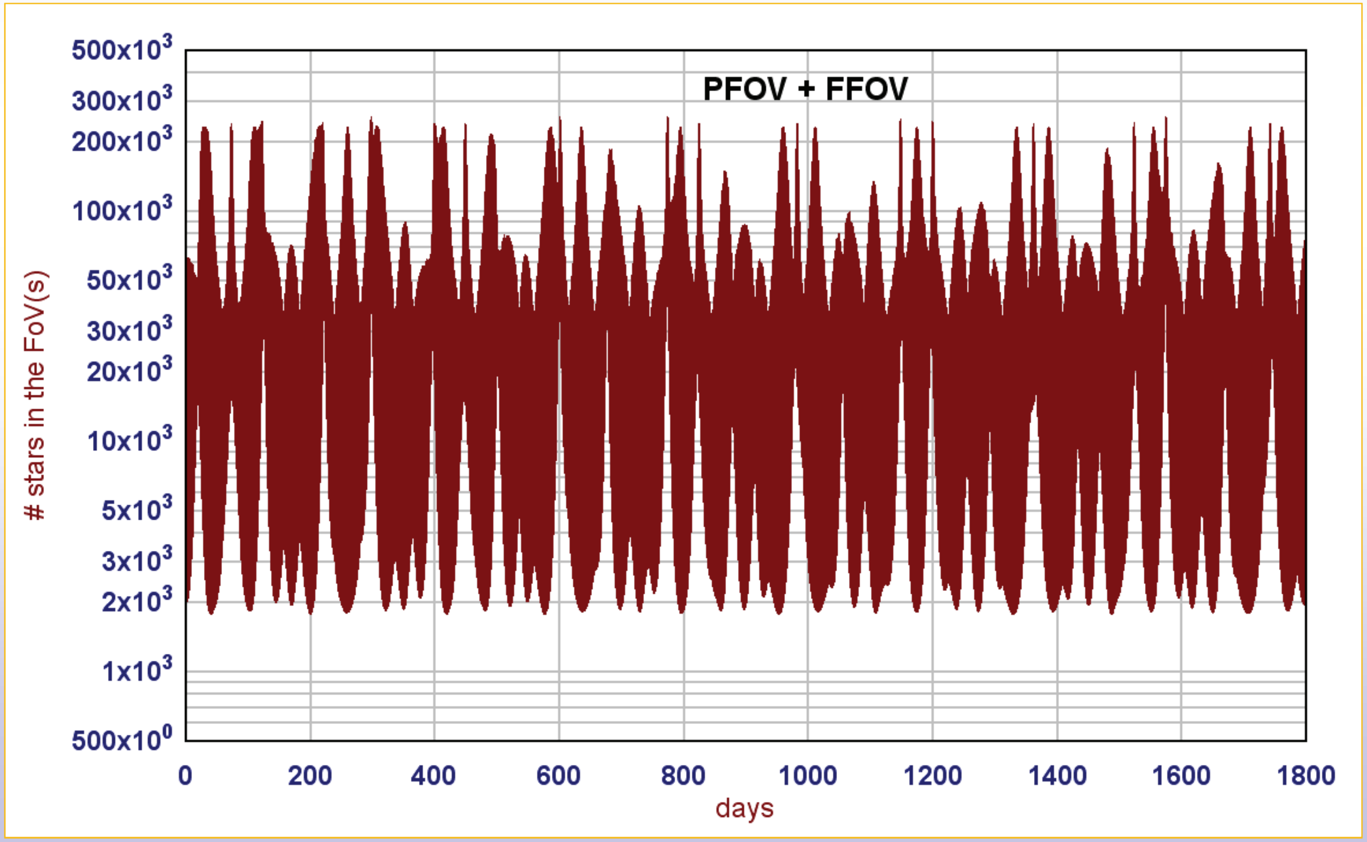
Sky average vs. Time average

- Time average is a combination of the sky distribution and the scanning law
  - two different symmetries: galactic plane and eclitpic plane



How many stars in the FoVs ?

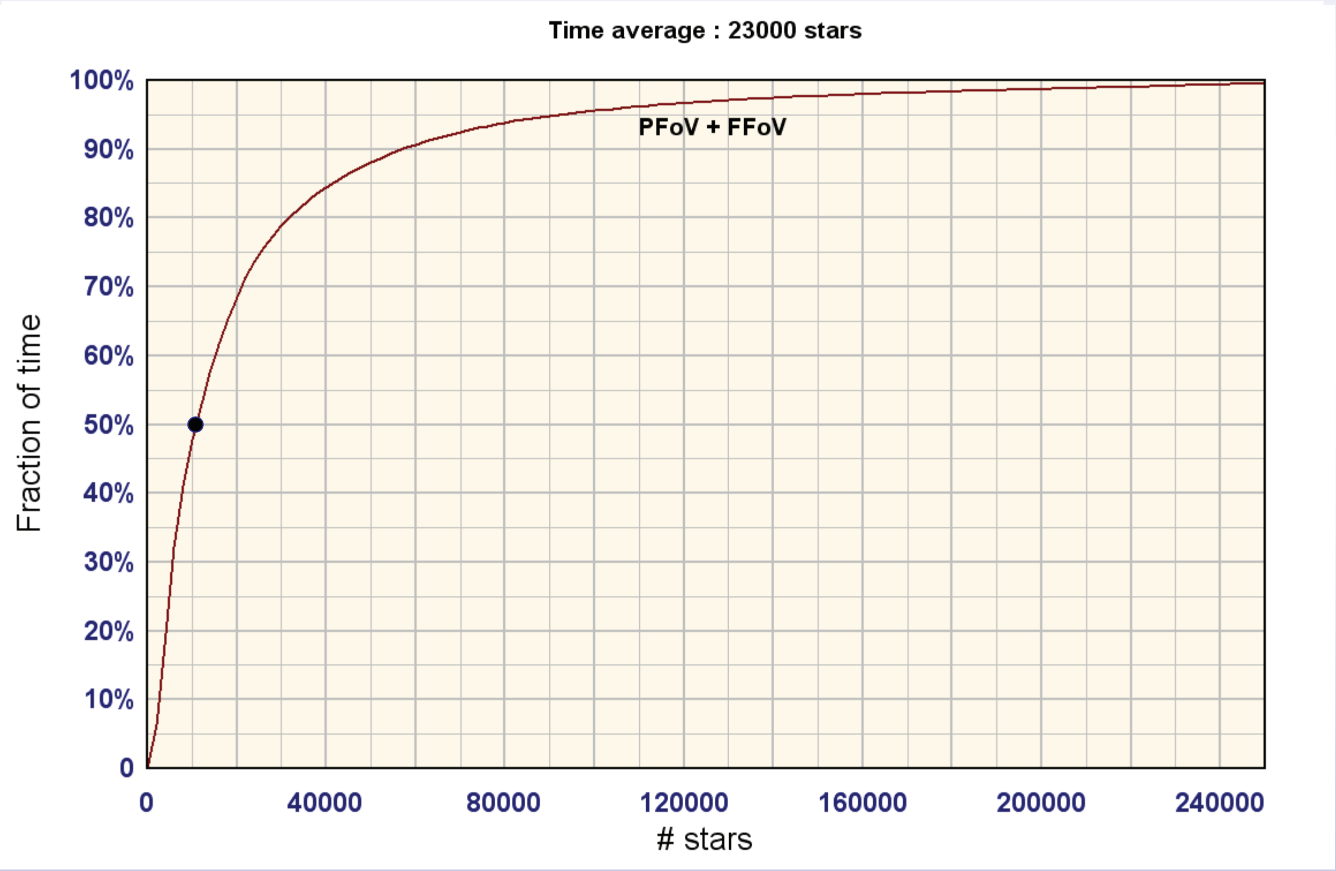
- Computation with the Nominal scanning law
  - time sampling = 7.5 mn over 5 years
- Pointing directions of each FoV in galactic coordinates
- Galaxy model for the stellar density
- The two FoVs are mapped on the same detector and densities added



Time distribution

- Time distribution of the stellar density on the focal plane
  - > 50% of time with combined area < 12000 stars in the Astro CCDs

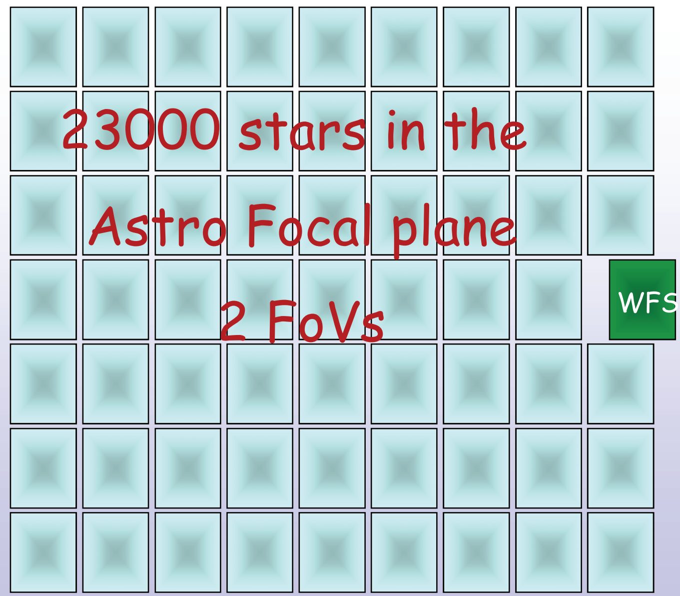
median



The sky mapped onto the focal plane

- Sky average density to V = 20 : 25000 */deg^2
  - But with large concentration near the galactic plane
- However Gaia spends more time in low-density areas
  - Time average is smaller ➔ sky is "empty" outside the galactic plane
- But the two FoVs are not superimposed as independent samplings

On the average on the sky one has:



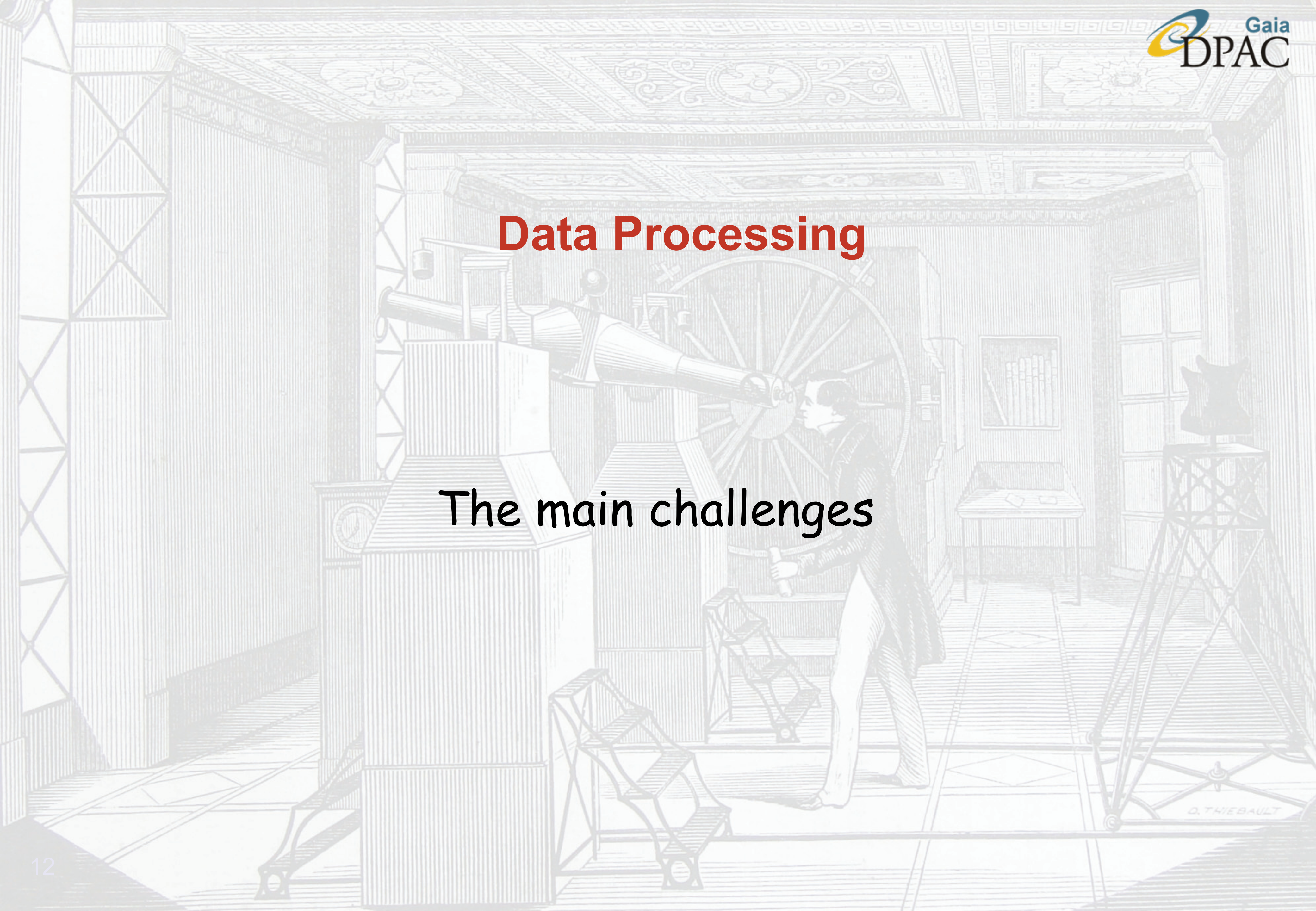
Astro FP ~ 1 deg^2

Elementary data on ground

- One billion sources to be observed
  - Stars, galaxies, QSOs, SSOs
- Average of 80 transits per sources during the mission
  - min ~ 50   max ~ 200
- Every source transits over 9 AF CCDs
  - 700 individual measurements per source
- 0.7 x 10^{12} CCD images produced by the mission
  - ~ one trillion elementary data
- Most images are 1D with have 6 samples
  - But some have 12 or 18
  - Bright sources are 2D
  - BP/RP have window 60-pixel wide

Summary numbers

- Data volume
  - compressed telemetry 250 Tb
  - raw data 100 TB
  - processed data and archives ~ 0.5 to 1 PB
- Computational size
  - 1.5×10^{21} FLOPs ➔ crude estimate
- Computational power expected in the DPAC in ~ 2012
  - > 10 TFLOP/s ➔ 2 yr CPU for 10^{21} FLOPs
- Data transfer
  - Downlink 50 GB/day
  - Data exchange between ESAC and DPCs :
    - challenging but workable solution being tested (W. O'Mullane presentation)
    - could ultimately rely to physical shipment !



Data Processing

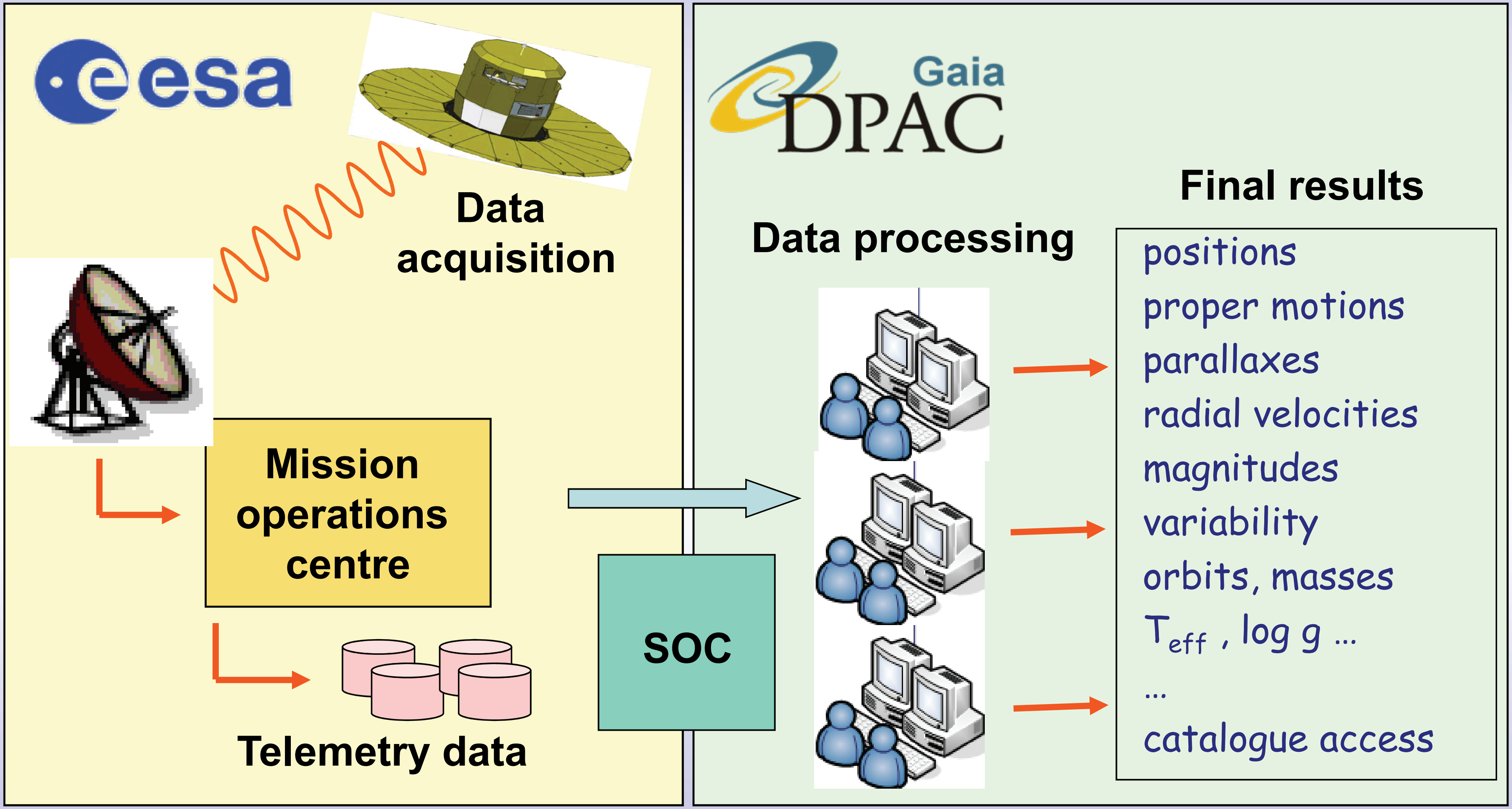
The main challenges

Different complexities in the DP

- Data volume
- Computational volume
- Data entanglement
- Institutional constraints

The sheer complexity of Gaia DP results from the combination of these four elements

Boundary conditions

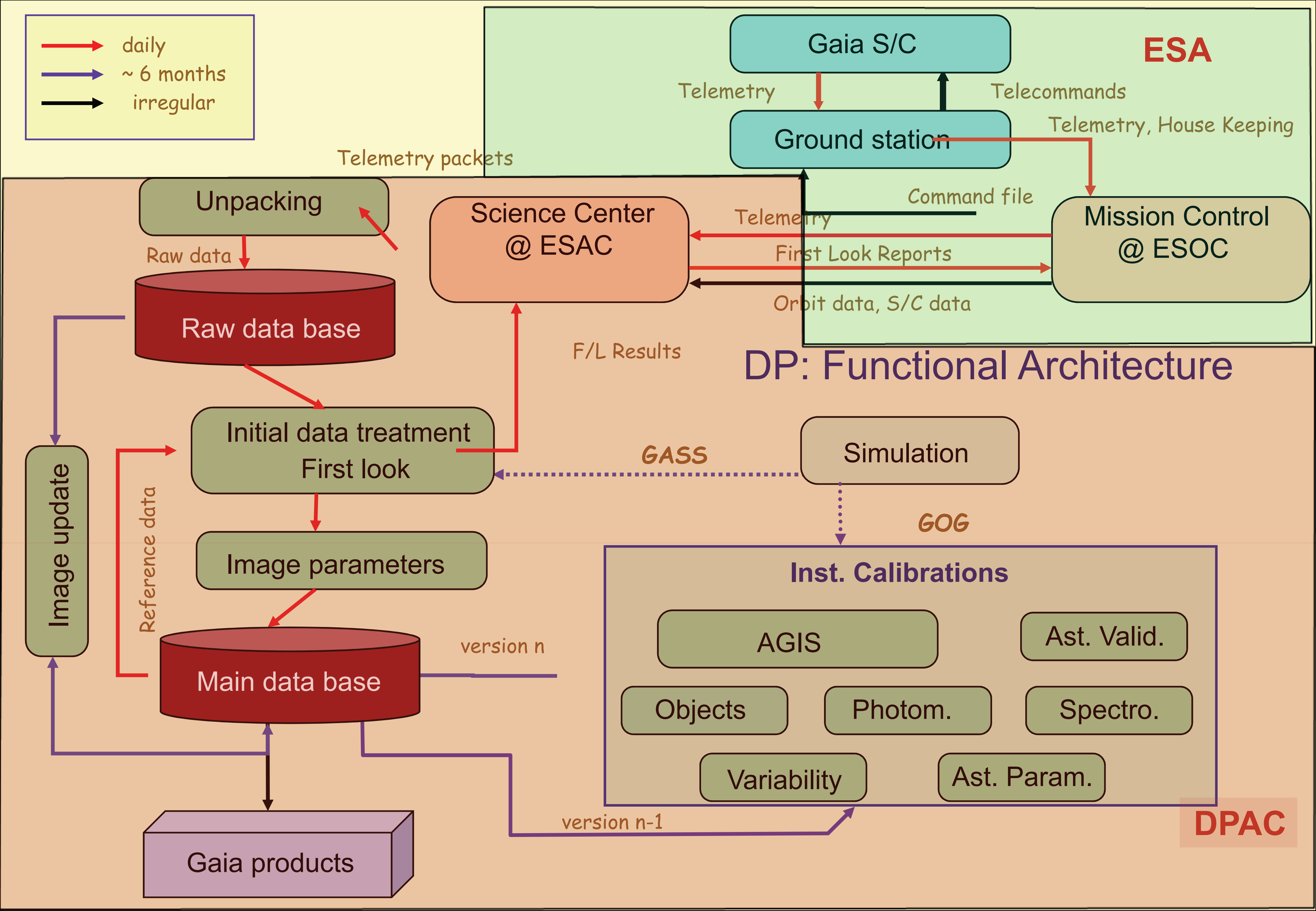


Major steps in the data processing 1/2

- Initial and global processing :
  - Data reception, preliminary attitude, source identification
  - Calibration, attitude, reference system
    - a global iterative processing is performed in this step
    - solution over ~ 10^8 primary stars
  - Update of the Main Database
    - astrometric solution on *x* months
    - satellite attitude
    - instrument parameters like CCD scales, Basic Angle

Major steps of the data processing 2/2

- Object based processing :
  - Processing for well-behaved sources
    - astrometric solution for secondary stars
    - photometry, variability detection and analysis
    - analysis of spectroscopic data
      - partly iterative for the wavelength calibration
  - Special sources
    - double and multiple stars (> 10^8 sources)
    - unresolved galaxies
    - quasars (~ 5×10^5)
    - solar system objects (~ 3×10^5)
  - Astrophysical parameters extraction



daily
~ 6 months
irregular
Gaia S/C
ESA
Telemetry
Telecommands
Ground station
Telemetry, House Keeping
Telemetry packets
Command file
Unpacking
Science Center @ ESAC
Telemetry
Mission Control @ ESOC
First Look Reports
Raw data
Raw data base
Orbit data, S/C data
F/L Results
DP: Functional Architecture
Initial data treatment First look
GASS
Simulation
Image update
Reference data
GOG
Image parameters
Inst. Calibrations
AGIS
Ast. Valid.
version n
Main data base
Objects
Photom.
Spectro.
Variability
Ast. Param.
version n-1
DPAC
Gaia products

Computational Size

	F = FLOPs	Total FLOPs
Initial treatment	2×10^{15} F/days	4×10^{18}
Iterative astrometry	2×10^{19} F/cycle	2×10^{20}
Image update	*1×10^{20} F/cycle*	1×10^{21}
Spectroscopy	*4×10^{17} F/cycle*	4×10^{18}
Photometry	*5×10^{18} F/cycle*	5×10^{19}
Non single sources	*~ 10^{16} F/cycle*	1×10^{17}

Total for a 5-year data processing:

1.5×10^{21} FLOPs

How big are 10^{21} FLOPs ?

10^{21} FLOPs is big, but achievable with a good organisation, but …

Comparison to Gaia

- Big difference with today large computations ~ 10^{20} – 10^{21} FLOPs
  - all in distributed computing over thousands units
  - all with virtually no data handling or big storage needs
  - all can be broken down into small independent pieces

Gaia

- big computation for today standards
- cannot be fully setup into many parallel computations
- involves a large data handling
- data must be accessed chronologically or per source
- computing power must be available in few centres

Dependency and Complexity

Two examples

Nature of the algorithms

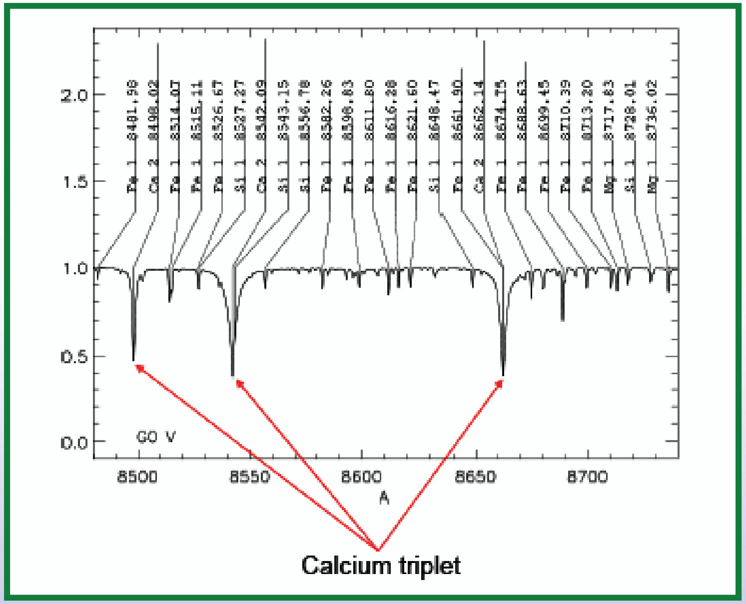
- The processing is comprised of hundreds of algorithms
- Some are genuine numerical procedure
  - Inverse computations: typically model fitting to observed data
    - Astrometric solution, orbit determination, calibration, attitude determination
  - Direction computation
    - Prediction of an observation, ephemeris computation, astrometric model
    - Statistical validation, plotting
    - Synthetic spectral libraries
- Many are closer to data handling with more combinatoric than numeric
  - Data compression, automated classification,
  - Object identification
    - Match an observation to a catalogue star
    - Detect any previously observed solar system object

Obvious dependencies

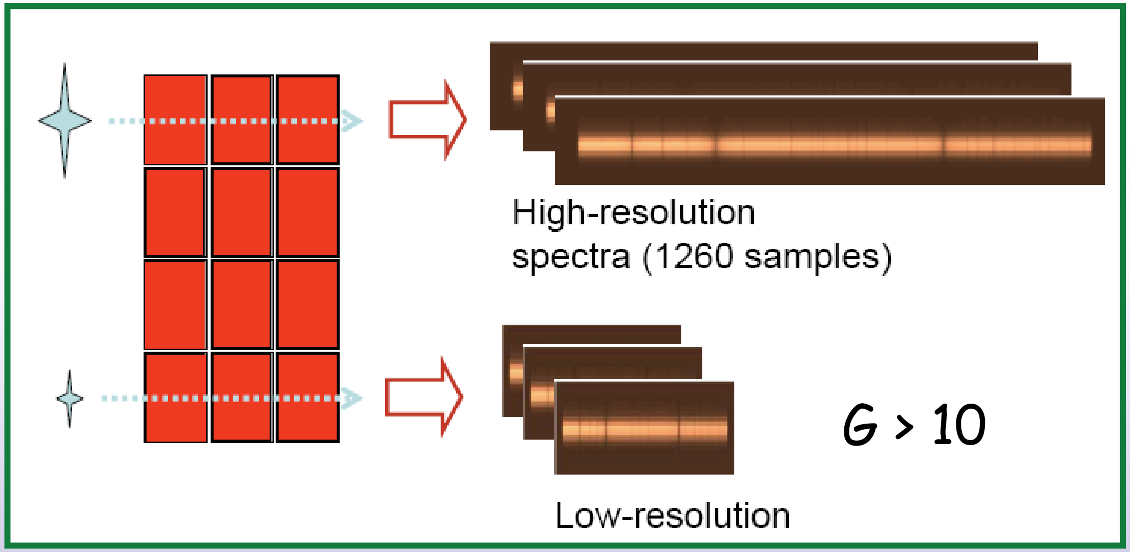
- Satellite attitude is needed almost everywhere
- Instrument calibration parameters depend on the source
  - primarily on its spectral type
  - colours are provided by the photometric solution
- Astrometry needs some knowledge of radial velocity
  - Generally not known before Gaia
  - Even for Hipparcos stars , only 50% have a known Vr
  - Most will be derived from the spectroscopic data (up to G ~ 16)
- Photometric and spectroscopic wavelength calibration need astrometric data
- Calibration results from AGIS used by every other chain

A more detailed example: spectroscopy

- Works in TDI mode
- The spectrum is not an instantaneous view
  - integrated over 4 s
- The star have crossed a full CCD chip

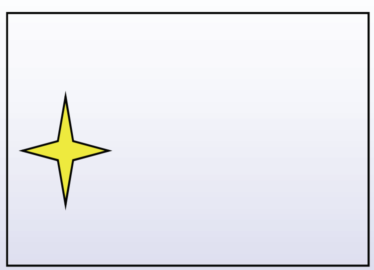


$G < 10$

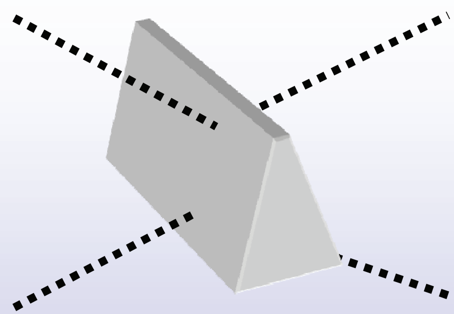


Relevant figures for the RVS

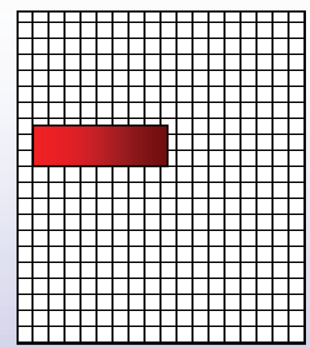
- Spectrum covers 27 nm
- It is spread over ~ 1100 pixels (HR)
  - 1 px ~ 0.025 nm
- $\sigma(v_r) = 1$ km/s ➔ $\delta\lambda \sim 0.003$ nm = 0.12 pixel
- Therefore wavelength calibration is a big issue
  - radiation damage bias too !



Field of view



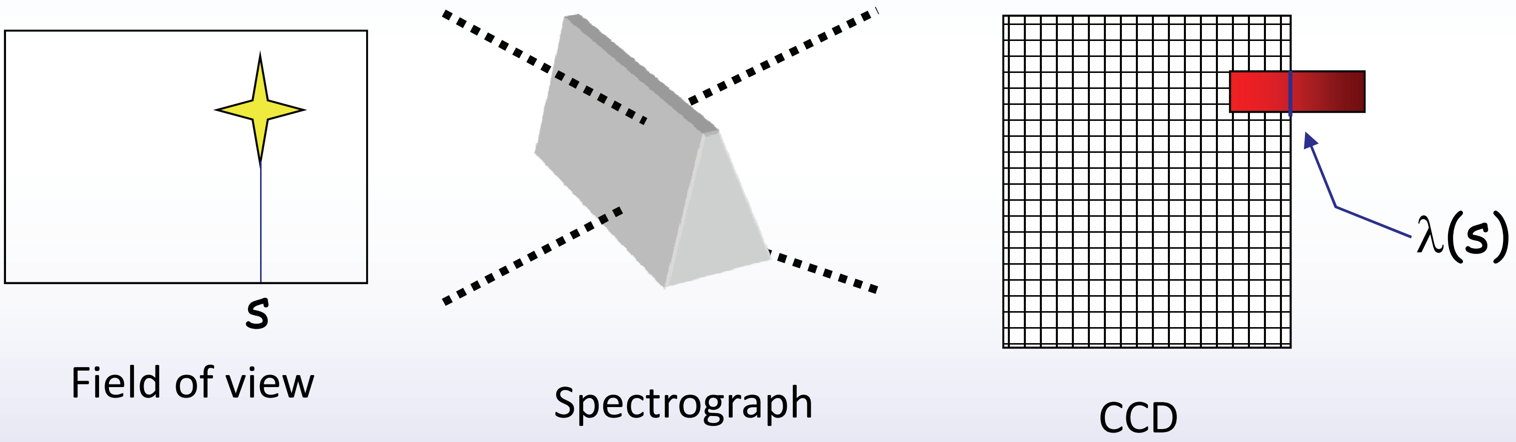
Spectrograph



RVS CCD

Wavelength calibration

- The Gaia spectro is a slitless instrument
- No internal reference for λ
- Good connection:
  - position of the image in the FoV and λ in the readout register



- Therefore: the position must be known or computed
  - remember: the RVS field if much outside the Astro field

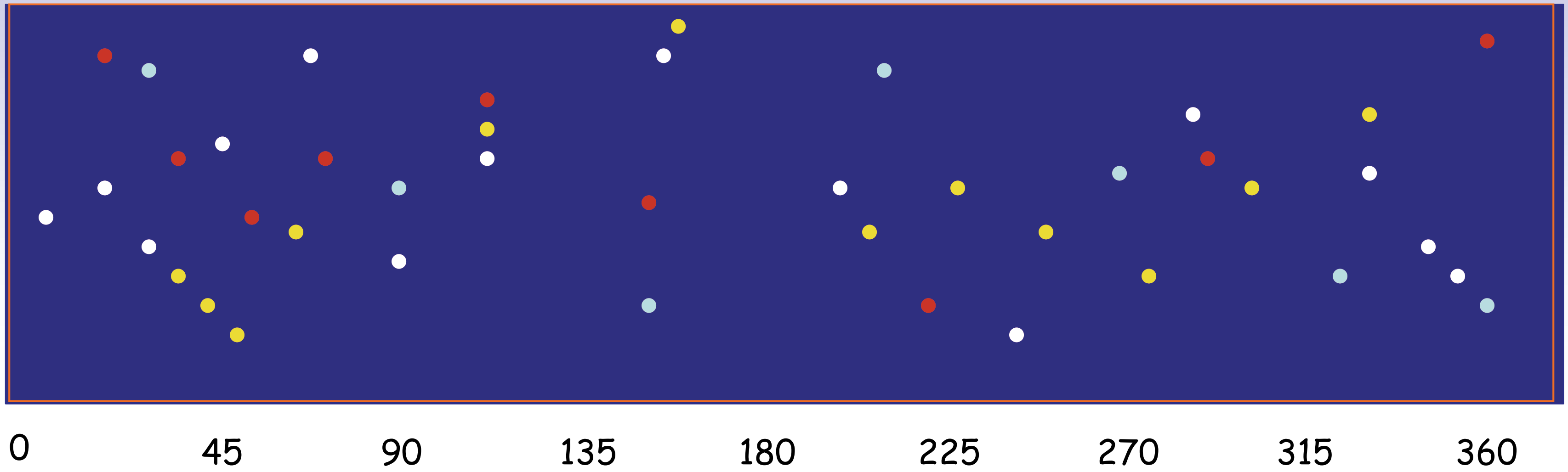
More dependencies for RVS

- Object identification, observation time
- Geometric calibration of the Focal Plane
- Magnitude in the RVS band is needed to correct for contamination
  - it is determined during the general photometric processing
- Characteristic of stars (spectral type) is needed to select masks
  - this is determined in the Astrophysical parameters processing
    - which needs also the spectra to complete the task

Identification of Moving objects

- Observations of stars are matched to a source
  - at each transit one ID is created and then associated to a source
  - this task is done in the Initial Data Treatment
- Then, in a well organised DB, it is easy to collect together, the ~ 80 observations of a particular object
- Solar system objects are observed as regularly as stars
- They have a motion relative to stars and cannot be matched easily to a source

Identification of Moving objects



- one must look first at all the known solar system objects and try to match a source position to Gaia observation
  - looks simple, but there are about 500,000 possible sources
- if this fails the object is probably new
  - then it becomes very hard to match its ~ 70 observations to a single source
  - here we have a real complex problem

From Transits to Sources

Transits

5 Sources