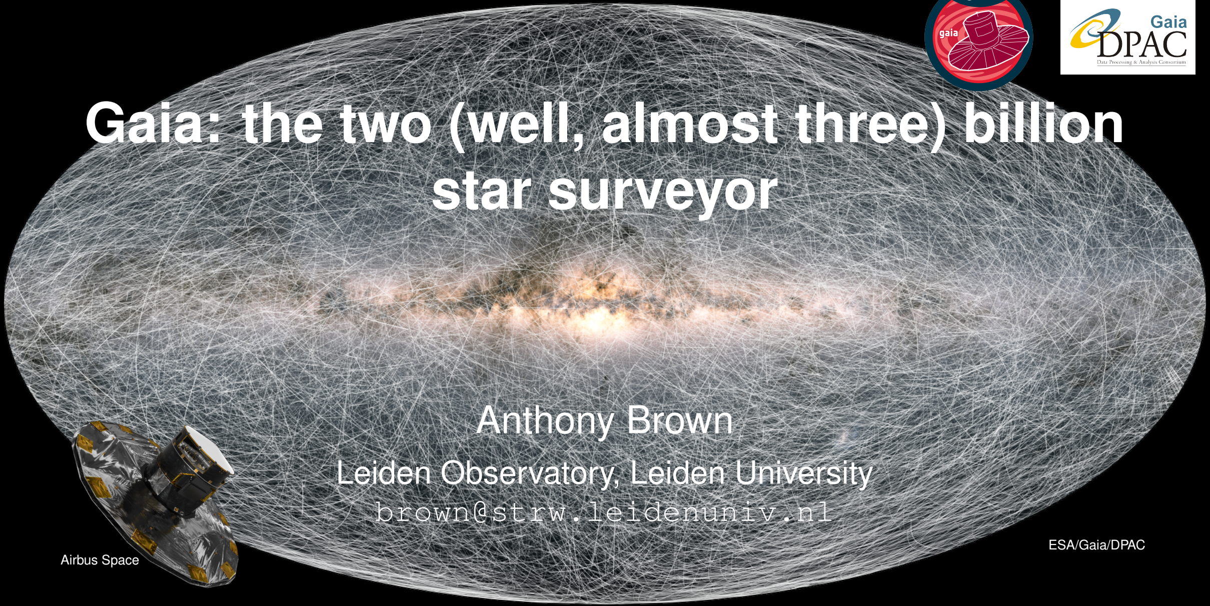


Gaia: the two (well, almost three) billion star surveyor

Anthony Brown

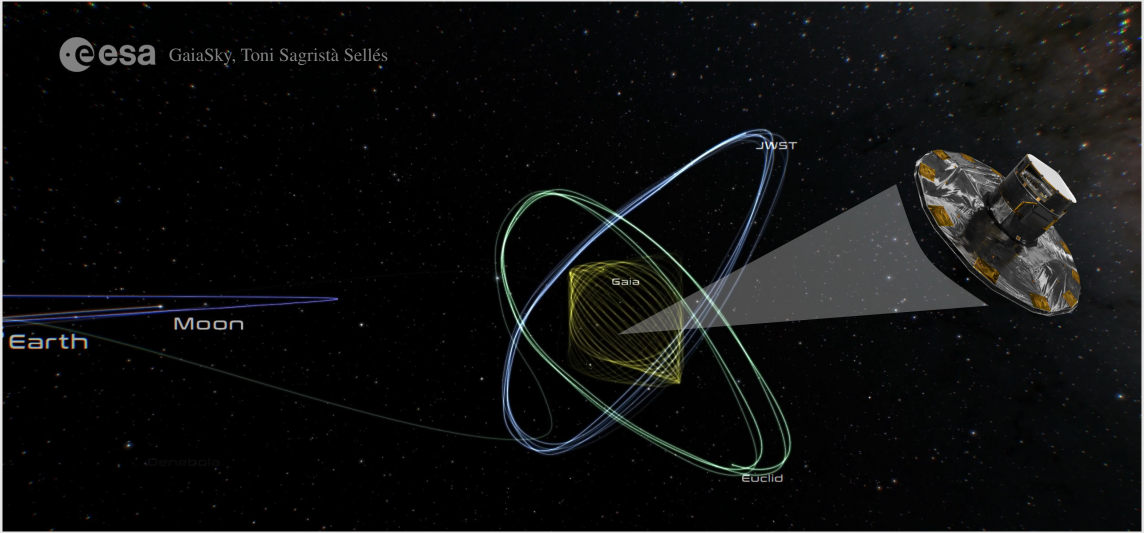
Leiden Observatory, Leiden University

brown@strw.leidenuniv.nl

Airbus Space

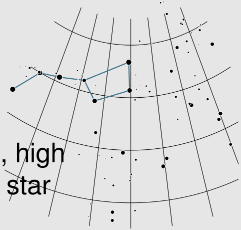
ESA/Gaia/DPAC

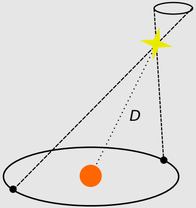
Gaia mission: launched 19.12.2013



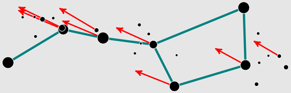
Gaia collects fundamental astronomical data

All-sky, complete, high accuracy star atlas

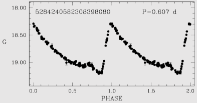


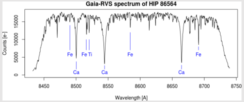


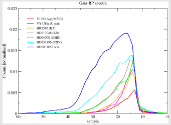
Parallaxes and proper motions

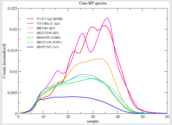


Astrometric, photometric, spectroscopic, radial velocity time series

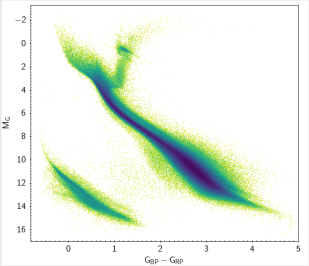






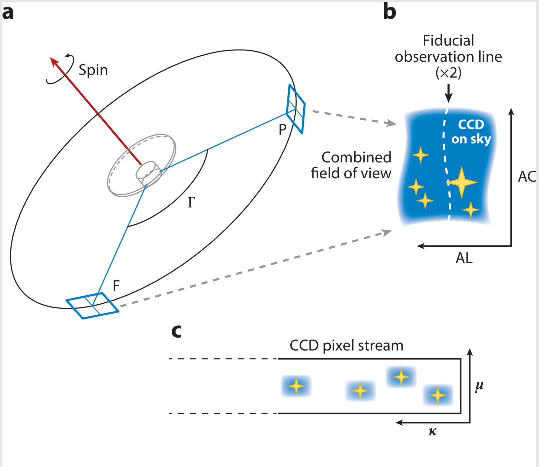


→



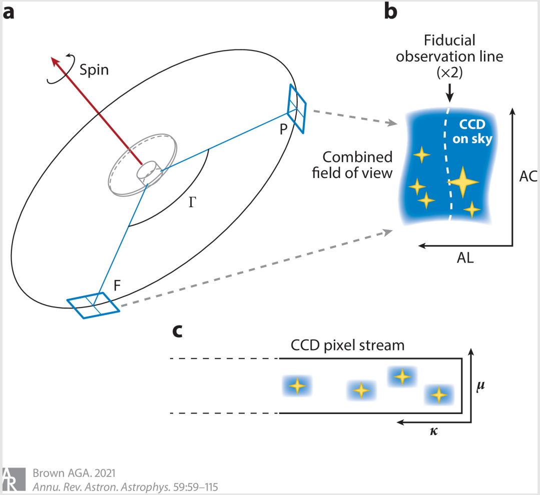
Astrophysical properties

Gaia data collection and processing

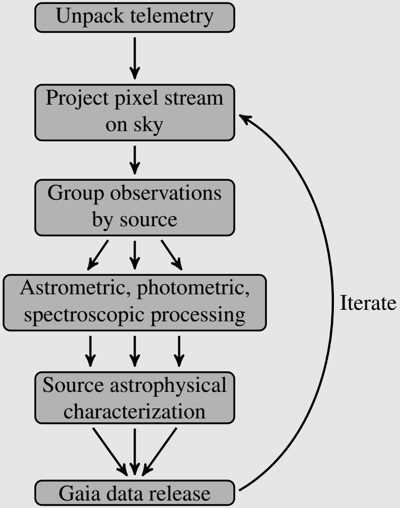


Brown AGA. 2021
Annu. Rev. Astron. Astrophys. 59:59–115

Gaia data collection and processing



Brown AGA. 2021
Annu. Rev. Astron. Astrophys. 59:59–115



Unpack telemetry

Project pixel stream on sky

Group observations by source

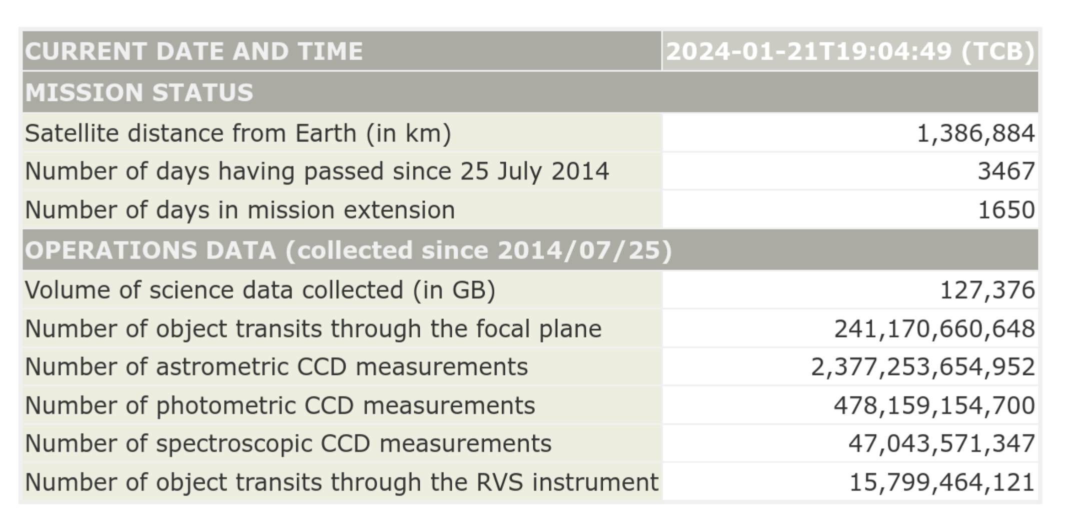
Astrometric, photometric, spectroscopic processing

Source astrophysical characterization

Gaia data release

Iterate

How big is Gaia data?

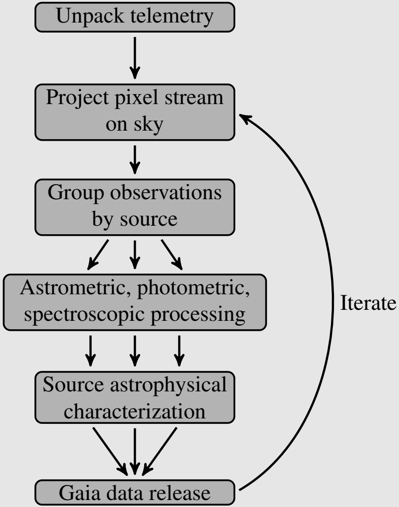


CURRENT DATE AND TIME	2024-01-21T19:04:49 (TCB)
MISSION STATUS	
Satellite distance from Earth (in km)	1,386,884
Number of days having passed since 25 July 2014	3467
Number of days in mission extension	1650
OPERATIONS DATA (collected since 2014/07/25)	
Volume of science data collected (in GB)	127,376
Number of object transits through the focal plane	241,170,660,648
Number of astrometric CCD measurements	2,377,253,654,952
Number of photometric CCD measurements	478,159,154,700
Number of spectroscopic CCD measurements	47,043,571,347
Number of object transits through the RVS instrument	15,799,464,121

HPC, numerical algorithms, machine learning, AI(?)

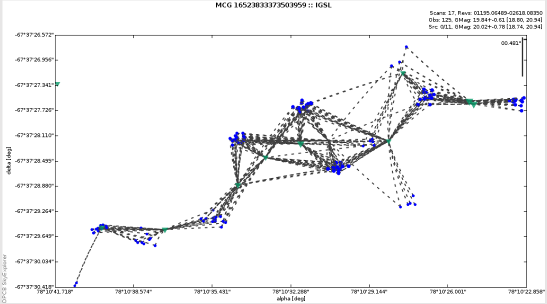
- Computing is done on large clusters in six data processing centres
- Data centres communicate via central hub: ‘Main data base’
- Data releases are extracted from MDB contents

MareNostrum 4 @ Barcelona Supercomputing Center

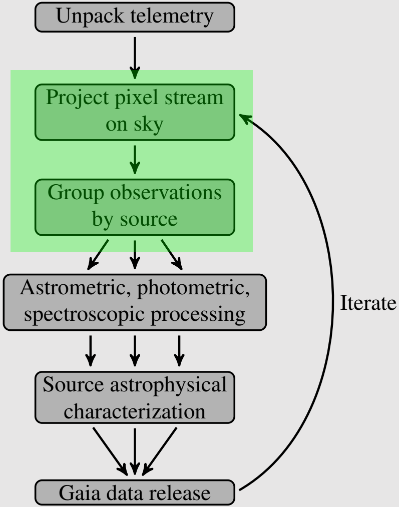


Unpack telemetry
Project pixel stream on sky
Group observations by source
Astrometric, photometric, spectroscopic processing
Source astrophysical characterization
Gaia data release
Iterate

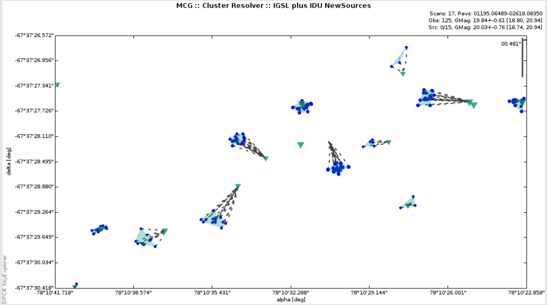
HPC, numerical algorithms, machine learning, AI(?)



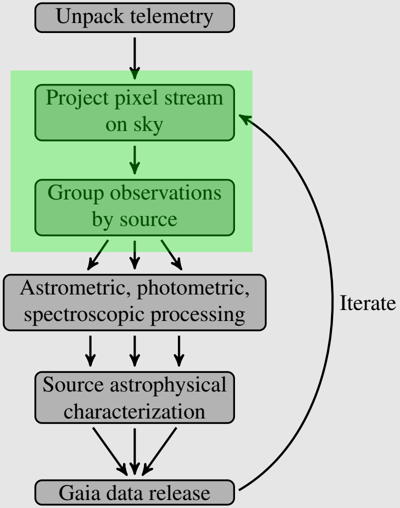
- Sophisticated clustering algorithm to associate observations with sources
- Needs to account for conflicting matches, fast moving sources, variable sources, spurious sources


Unpack telemetry
Project pixel stream on sky
Group observations by source
Astrometric, photometric, spectroscopic processing
Source astrophysical characterization
Gaia data release
Iterate

HPC, numerical algorithms, machine learning, AI(?)



- Sophisticated clustering algorithm to associate observations with sources
- Needs to account for conflicting matches, fast moving sources, variable sources, spurious sources


Unpack telemetry

Project pixel stream on sky

Group observations by source

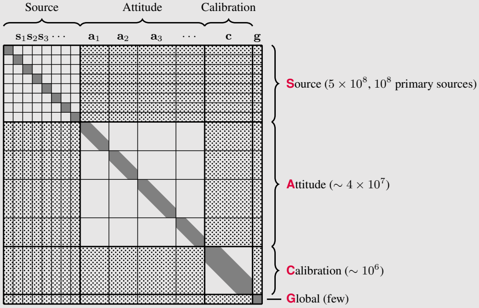
Astrometric, photometric, spectroscopic processing

Source astrophysical characterization

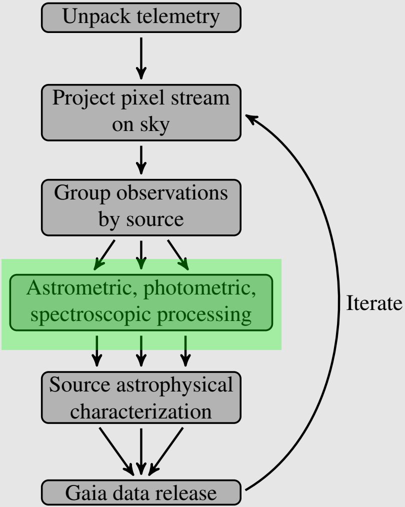
Gaia data release

Iterate

HPC, numerical algorithms, machine learning, AI(?)


Source
Attitude
Calibration
$s_1 s_2 s_3 \cdots$
a_1
a_2
a_3
$\cdots$
c
g
Source (5×10^8, 10^8 primary sources)
Attitude ($\sim 4 \times 10^7$)
Calibration ($\sim 10^6$)
Global (few)

- Large systems of linear equations; iterative least squares solution
  - calibration terms lead to non-sparse matrices
- Also classical astronomical data processing methods (e.g., cross correlation of spectra with templates to derive radial velocity)



Unpack telemetry

Project pixel stream on sky

Group observations by source

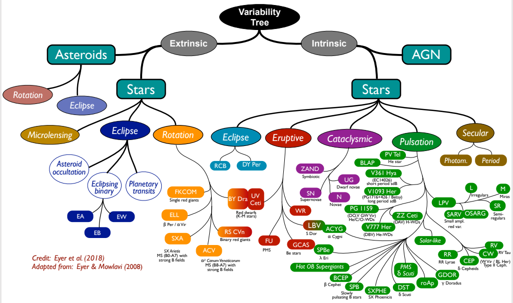
Astrometric, photometric, spectroscopic processing

Source astrophysical characterization

Gaia data release

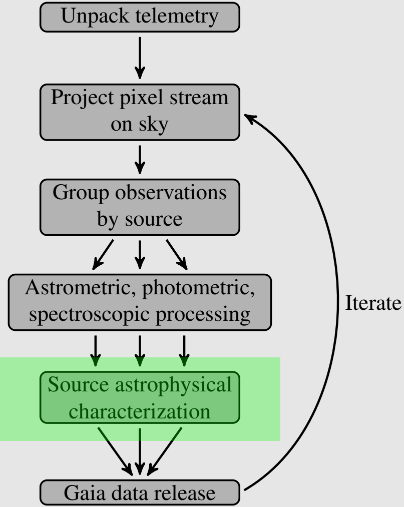
Iterate

HPC, numerical algorithms, machine learning, AI(?)



Credit: Eyer et al. (2018)
Adapted from: Eyer & Mowlavi (2008)

- Machine learning algorithms: classification, clustering, outlier detection, period searching in time series, …
- Inference algorithms: MCMC, likelihood optimization, …



Unpack telemetry

Project pixel stream on sky

Group observations by source

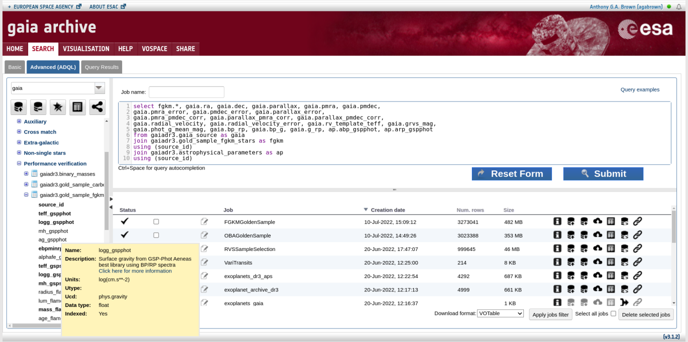
Astrometric, photometric, spectroscopic processing

Source astrophysical characterization

Gaia data release

Iterate

HPC, numerical algorithms, machine learning, AI(?)



Job name:

```
select fgkm.*, gaia.ra, gaia.dec, gaia.parallax, gaia.pmra, gaia.pmdec,
gaia.pmra_error, gaia.pmdec_error, gaia.parallax_error,
gaia.pmra_pmdec_corr, gaia.parallax_pmra_corr, gaia.parallax_pmdec_corr,
gaia.radial_velocity, gaia.radial_velocity_error, gaia.rv_template_teff, gaia.grvs_mag,
gaia.phot_g_mean_mag, gaia.bp_rp, gaia.bp_g, gaia.g_rp, ap.abp_gspphot, ap.arp_gspphot
from gaiadr3.gaia_source as gaia
join gaiadr3.gold_sample_fgkm_stars as fgkm
using (source_id)
join gaiadr3.astrophysical_parameters as ap
using (source_id)
```

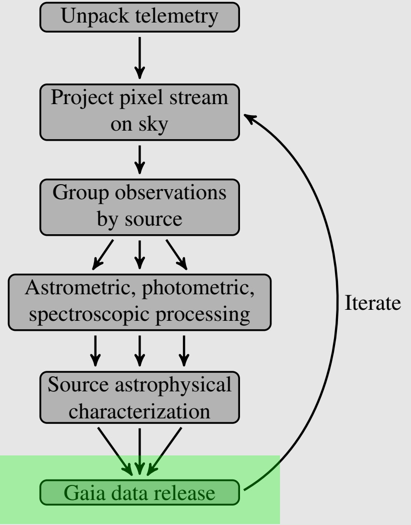
Ctrl+Space for query autocompletion

Status			Job	Creation date	Num. rows	Size
✓	□		FGKMGoldenSample	10-Jul-2022, 15:09:12	3273041	482 MB
✓	□		OBAGoldenSample	10-Jul-2022, 14:49:26	3023388	353 MB
			RVSSampleSelection	20-Jun-2022, 17:47:07	999645	46 MB
			VariTransits	20-Jun-2022, 12:25:00	214	8 KB
			exoplanets_dr3_aps	20-Jun-2022, 12:22:54	4292	687 KB
			exoplanet_archive_dr3	20-Jun-2022, 12:17:13	4999	661 KB
			exoplanets_gaia	20-Jun-2022, 12:16:37		1 KB

Download format: VOTable

Name: logg_gspphot
Description: Surface gravity from GSP-Phot Aeneas best library using BP/RP spectra
Click here for more information
Units: log(cm.s**-2)
Utype:
Ucd: phys.gravity
Data type: float
Indexed: Yes

- Data access through ADQL, TAP, Python modules, bulk download
- Extensive documentation, tutorials, code examples
- ESA + Several partner data centres



Unpack telemetry

Project pixel stream on sky

Group observations by source

Astrometric, photometric, spectroscopic processing

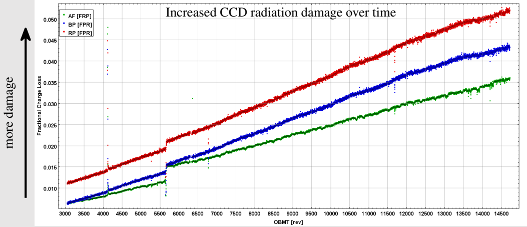
Source astrophysical characterization

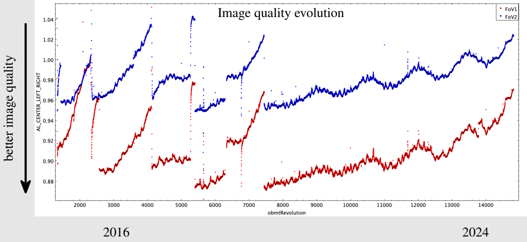
Gaia data release

Iterate

Specifics of the Gaia data processing

- Of order 3 × 10^{12} individual observations
  - ▶ they all count!
- Observations are connected over large angles on the sky
- Data access: spatially or temporally grouped, time series per source
- All instruments are self-calibrated
  - ▶ Complex calibration models with millions of parameters
- Instrument characteristics evolve over time
- Division of data processing tasks and physical location was partly political
- Most of this is not unique to Gaia…





Gaia and Big Data: the practical challenges

- Develop and maintain a well-documented data model
- Coordination across DPAC units (JIRA, e-mail, many telecons…)
- Organizing data ahead of a processing run
- Data transfer between processing centres
- Develop/ test/ validate new or improved pipeline code
- Validation of the data products
  - are they scientifically correct?
  - no missing or out of bounds values?
  - what to do if we find a problem: fix, or document?
  - …
- Transfer of data from the main data base to the public archive
  - make sure data mapping is done carefully and without mistakes
- Documentation for a data release