



The UK Gaia Data Mining Platform

Nigel Hambly (IfA Edinburgh & Gaia DPAC)
National Astronomy Meeting, Hull 2024



THE UNIVERSITY
of EDINBURGH



UNIVERSITY OF
CAMBRIDGE



DR3: “A release of superlatives...”

... yet it represents

- *only* 27% of the likely end-of-mission observation time-line
- ~1% of the likely end-of-mission data release volume

DR1 (Q3 2016; 15 m)



DR2 (Q2 2018; 22 m)



EDR3 (Q4 2020 34 m)



DR3 (Q2 2022; 34 m)



DR4 (not before mid 2026; 66m)



DR5 (not before end 2030 TBC; ≥ 10 yr up to end-of-mission)



Mission timeline

July 2014

May 2017

Jan 2020

July 2024

EOM:
Early 2025

Gaia DR3: a release of superlatives



Beyond the largest and most accurate astrometric and photometric survey to date (Gaia EDR3):

- Largest ever spectrophotometric survey
- Largest ever radial velocity survey
- First space-based all-sky survey of QSO galaxy hosts and of the surface brightness profiles of galaxies in the local universe
- Highest accuracy spectrophotometric-dynamical survey of asteroids
- For many classes of variable stars: largest survey ever
- Largest ever collection of astrophysical data for stars in the Milky Way
- Non-single star survey that surpasses all the work on non-single stars from the past two centuries

GAIA MISSION STATUS

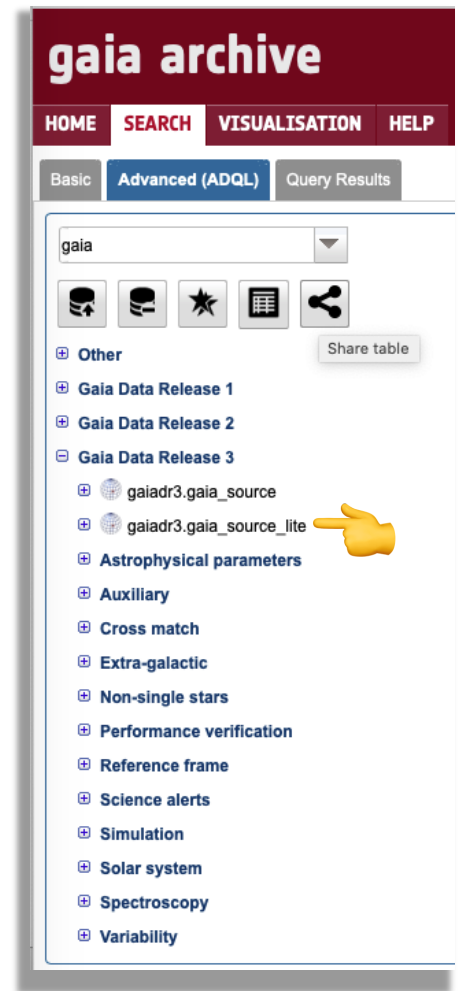
3639 days in science operations
134,606 GB of science data gathered
255,505,012,421 transits observed

Gaia DR3 bulk data volume

- Bulk download of DR3 products is provisioned via a *Content Delivery Network*
 - 8.9 TB of gzipped eCSV (text) files; 25 TB uncompressed
 - Single thread download/decompress the lot in roughly a few days
 - Largest single data sets are (eCSV)
 - XP mean spectra: 8 TB (12% of catalogue sources)
 - MCMC posterior PDF samples output from Apsis General Stellar Parameterization from Photometry: 7 TB (0.1% of catalogued sources)
- DRs 4 (& 5) detailed contents being generated now (DR5 still under discussion) but $\approx 60\times$ (DR4) to **100s** (DR5) times bigger than DR3
 - More spectra, epoch-resolved data, raw and/or intermediate data, ...

Relational systems beginning to groan...

- TAP/ADQL provides **limited** access to bulk data via *VO Datalink*
 - Relational DBs don't handle array types easily (i.e. time series, spectra, sampled PDFs, ...)
- Even straight tabular data is becoming challenging
 - Mitigate in the short term via e.g. column subset “gaiadr3.gaia_source_lite” in Gaia archive
- No facility for end-user programmability
 - e.g. ADQL has only very basic statistical aggregates
 - ADQL is not a programming language!



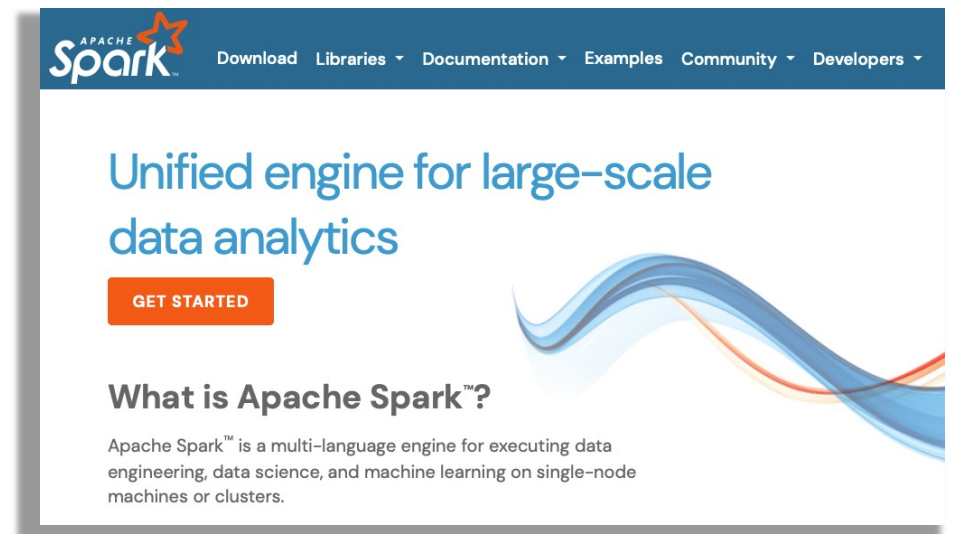
Advanced, scale-out usage scenarios

Community requirements gathered by DPAC via GREAT network and documented in [Brown+ \(2012\)](#)


- Higher order, robust statistical aggregates (e.g. GDAS-OA-03)
- Analysis of per-CCD photometry for short timescale variability (e.g. GDAS-ST-19)
- Searches in Fourier-analysed time domain data (e.g. GDAS-ST-12)
- Wholesale dataset trawls (e.g. GDAS-ST-11)
 - e.g. Spectral twins
- Pattern queries (e.g. GDAS-ST-08)
 - increasingly requiring Machine Learning techniques
- General CPU-intensive analysis (e.g. GDAS-OA-01)
- Efficient searching for pairs (or higher multiples) of associated objects, e.g.
 - Lensed QSOs
 - Wide binaries
- Searches in time-resolved astrometric data, e.g. detect plane gravitational wave(s) or primordial stochastic GW background
 - Requires local plane coordinate residuals from epoch astrometry

Introducing the *UK Gaia Data Mining Platform*

- The (obvious) solution: code-to-data platforms
 - Bring end-user code to lots of CPU co-located with the data
 - Employ distributed computing to mitigate increases in data volume and scale of processing
 - cf. *Rubin Science Platform* for LSST, STScI/MAST *TIKE*, ESA *Datalabs*, ...
- The UK Gaia DMP
 - Deployed on the STFC IRIS *Cloud*
 - Flexibility and scalability
 - Employs Apache Spark ecosystem
 - Python notebook interface
 - Friendly APIs to access distributed processing
 - Familiar libraries for vectorized operations
 - Machine Learning and many other libraries



<https://dmp.gaia.ac.uk/>

**Gaia DMP**

Notebook ▾Job

NHambly ▾

Welcome to the Gaia Data Mining Platform!

Powered by Apache Spark & Zeppelin

The UK Gaia Data Mining Platform is a science platform for large-scale exploitation of the publicly released Gaia data products. It provides a notebook-based environment and distributed compute facilities via Apache Zeppelin & Spark. Particularly suited to scale-out workflows requiring some level of high performance and/or high throughput.

Notebook ↗

- [Import note](#)
- [Create new note](#)

- [Users](#)
- [Trash](#)

Help


Get started with [Gaia DMP documentation](#)


Community


Please feel free to contact us with any questions or to report issues
Any contribution are welcome!

galadmp-support@roe.ac.uk


[Github](#)





**Gaia in the UK**
Taking the Galactic Census


**UKRI**

Science and
Technology
Facilities Council

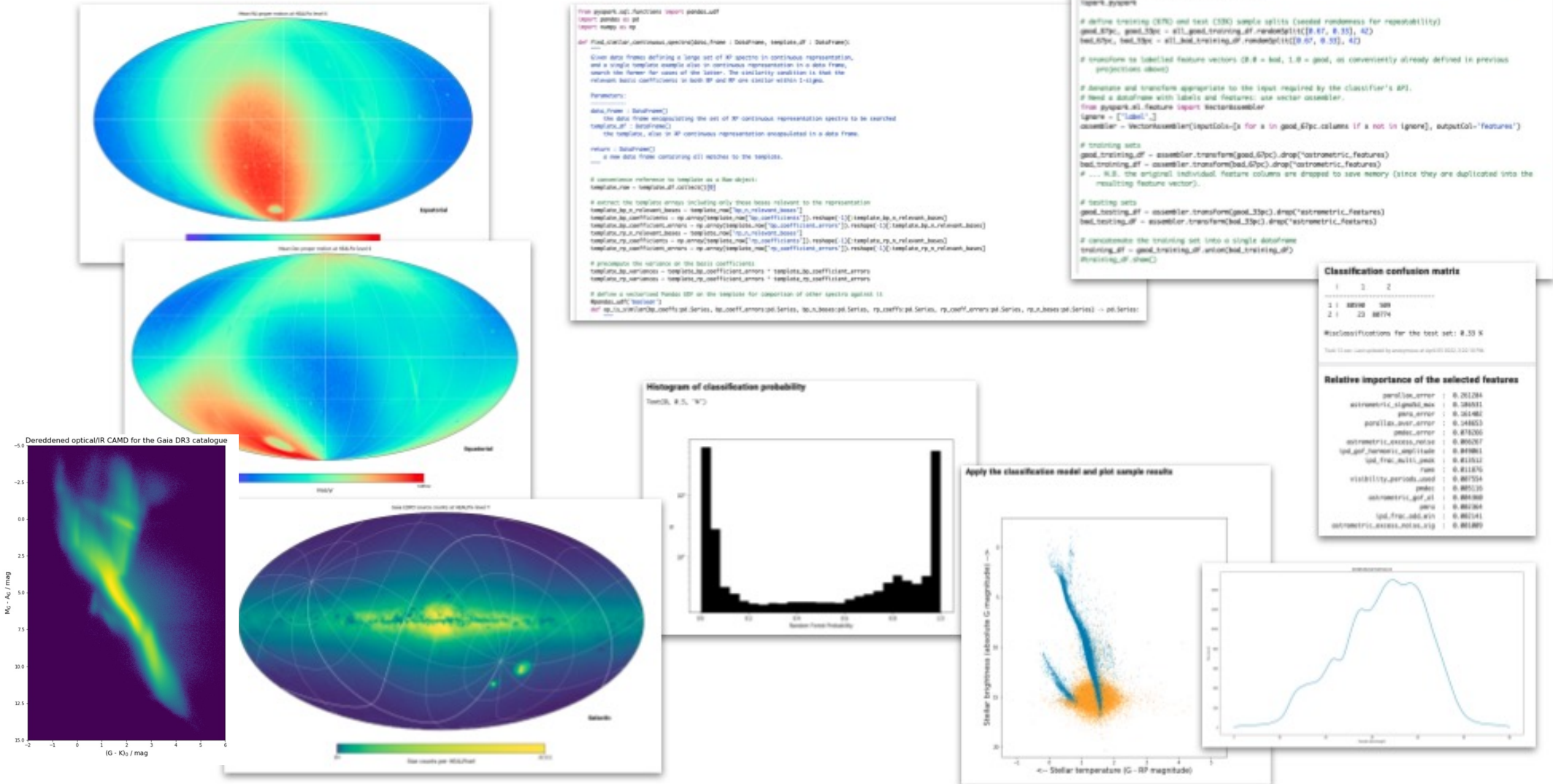
**iris**

**Gaia DPAC**
Data Processing & Analysis Consortium

**APACHE Spark™**

**Apache Zeppelin**

<https://dmp.gaia.ac.uk/>



A detailed example: searching 2×10^8 spectra

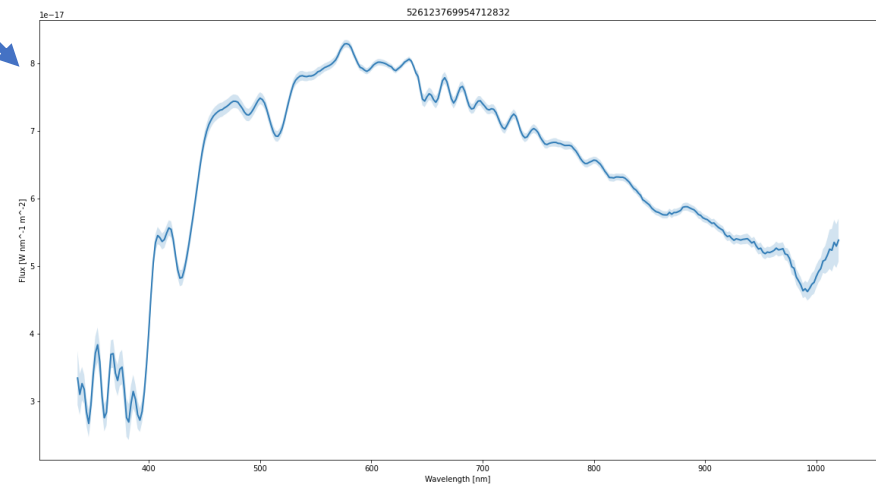
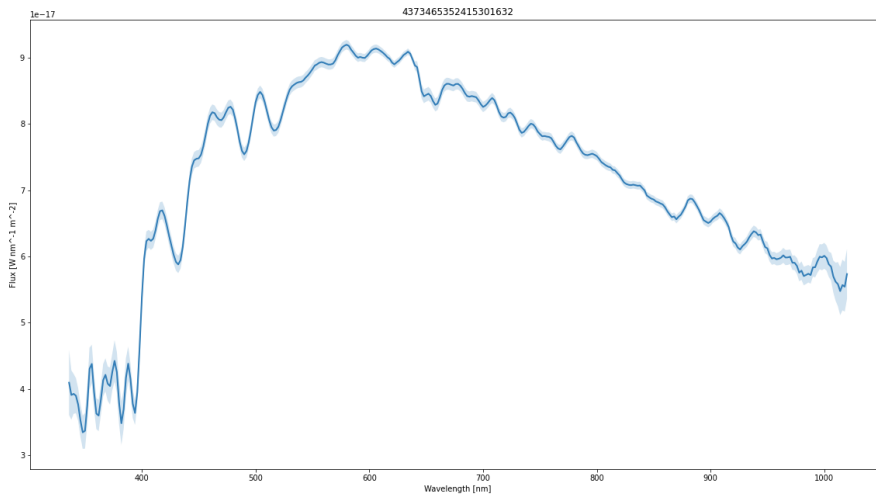
```
xp_continuous_mean_spectrum_schema = StructType([
    StructField('source_id', LongType(), False), # Unique source identifier (unique within a particular Data Release)
    .
    .
    .
    StructField('bp_coefficients', ArrayType(DoubleType()), True), # Basis function coefficients for the BP spectrum representation
    StructField('bp_coefficient_errors', ArrayType(FloatType()), True), # Basis function coefficient errors for the BP spectrum representation
    StructField('bp_coefficient_correlations', ArrayType(FloatType()), True), # Correlation matrix for BP coefficients
    .
    .
    .
    StructField('rp_coefficients', ArrayType(DoubleType()), True), # Basis function coefficients for the RP spectrum representation
    StructField('rp_coefficient_errors', ArrayType(FloatType()), True), # Basis function coefficient errors for the RP spectrum representation
    StructField('rp_coefficient_correlations', ArrayType(FloatType()), True), # Correlation matrix for RP coefficients
    .
    .
    .
])
```

- DR3 has 220 million blue + red spectra in basis-set representation
 - N basis coefficients
 - N coefficient uncertainties
 - $N(N-1)/2$ correlation coefficients
 - where $N = 55$ in each passband
- 2.7TB in compact (Parquet) binary format
- Simple use case: given one example template, find similar spectra ...

... for example, a Solar (G-) type star

$$D_M = \sqrt{(c_1 - c_2)^T (\Sigma_1 + \Sigma_2)^{-1} (c_1 - c_2)}$$

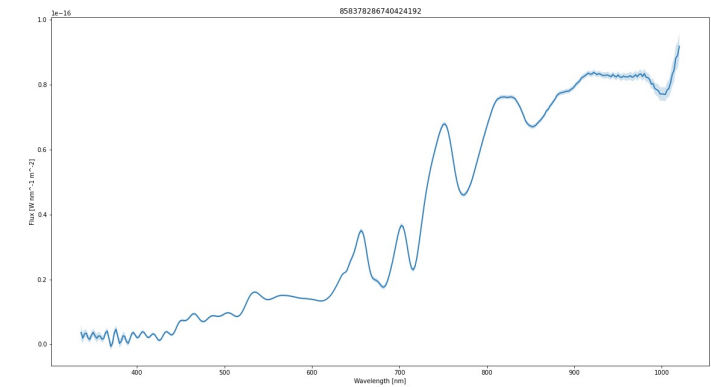
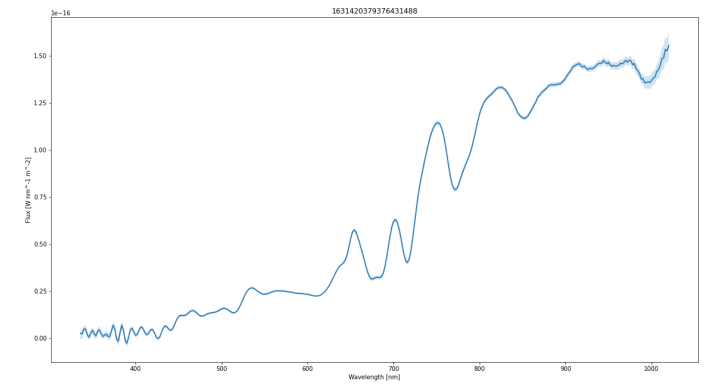
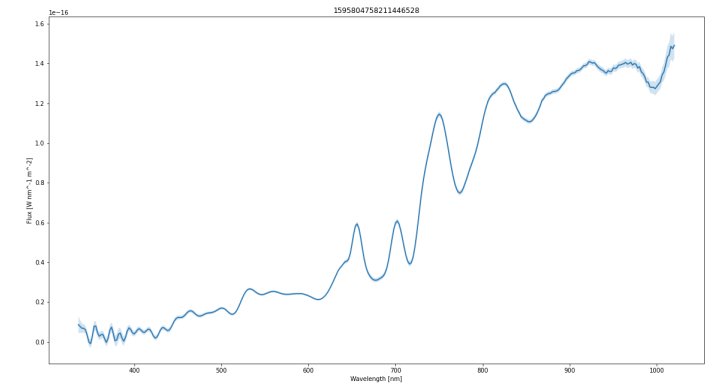
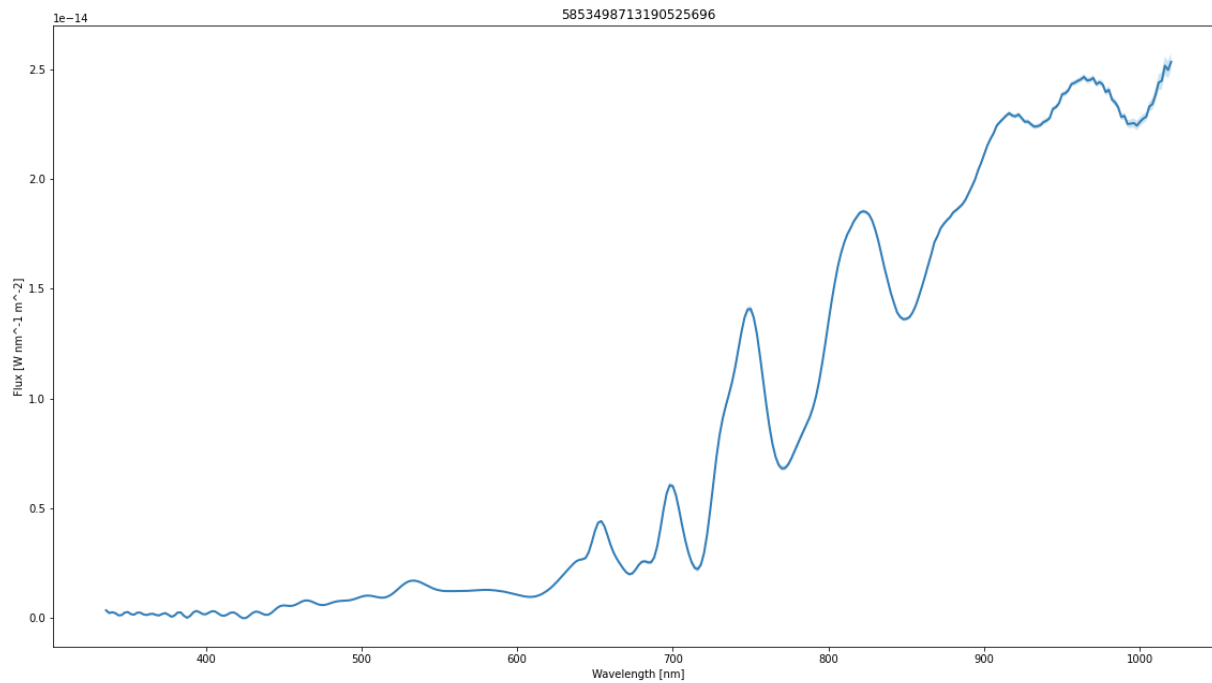
- Statistical rigour: compute the *Mahalanobis distance* (e.g. De Angeli et al. 2022) between the template and all others
 - In each case reconstruct the full 2d covariance matrix from the (flattened, 1d) correlation matrix and uncertainties vector
 - matrix & vector multiplications implemented as a “Pandas” (vectorized) User Defined Function for execution on Spark cluster worker nodes



Example matches at similar S/N in a couple of hours

- Modest level of parallelism in (virtual) Spark cluster
- I/O bound (CPU wait time typically 50%)

Another example: Proxima Cen look-alikes



Further information

- If you ...
 - ... find your Gaia science hobbled by the existing archive access, and / or
 - ... are interested in Data Mining / Machine Learning in Gaia astronomy, and / or
 - ... wish to learn more about industry-standard “big data” technology

then please get in touch (speak to me today at NAM 2024 and see live demos!)

Email: gaiadmp-support@roe.ac.uk

for an account and further information.