iris
Science and Technology Facilities Council
Funded by the European Union
SPACIOUS

Space astronomy science platforms

Nigel Hambly

Institute for Astronomy

THE UNIVERSITY of EDINBURGH

The ‘big data’ challenge

- Support scale and complexity of analysis
  - e.g. Bayesian statistics, Machine Learning, …
- Enable reproducibility of results
  - e.g. large analyses must be independently validated and explored
- Facilitate open science and inclusivity perspectives
  - e.g. ensure widest opportunities and participation
- Ensure sustainability – we should avoid
  - idle hardware
  - accumulation of obsolescent hardware
  - energy waste

Code-to-data (a.k.a. science) platforms

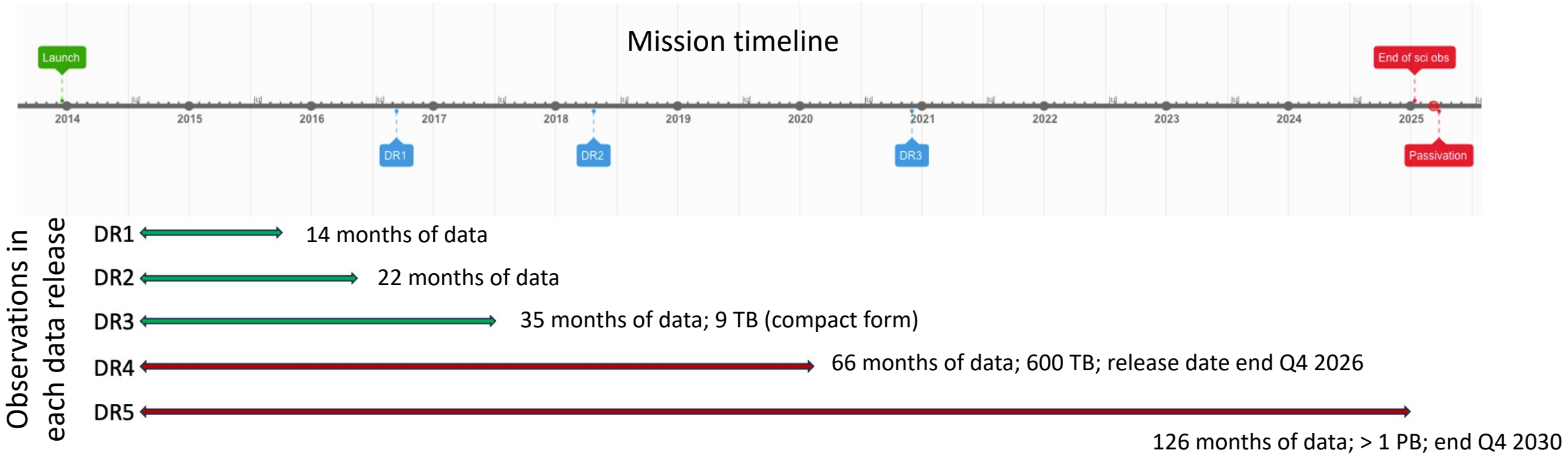
Cloud-based science platforms provide a solution to many of the challenges:

- On-demand
  - and open to all (authorised!)
- Auto-scale in response to `bursty' workloads
  - e.g. unused resources return to a common pool
- Provide a pre-configured environment with all necessary data and SW
  - e.g. enhances reproducibility
- Access to distributed computing
  - e.g. scales to large volumes and heavy analyses

Example astronomy science platform initiatives

- TIKE
  - NASA's TESS and Kepler/K2 mission data
- Rubin Science Platform
  - Process and analyse LSST data
- Astronomy-commons science platform
  - Generic solution illustrated with ZTF
- NADC science platform
  - Generic, Virtual-Observatory initiative
- ESA DataLabs
  - ESA-mission specific services
- SPACIOUS-AstroFlow
  - ESA Gaia/Euclid focussed


Example: Gaia data release volumes



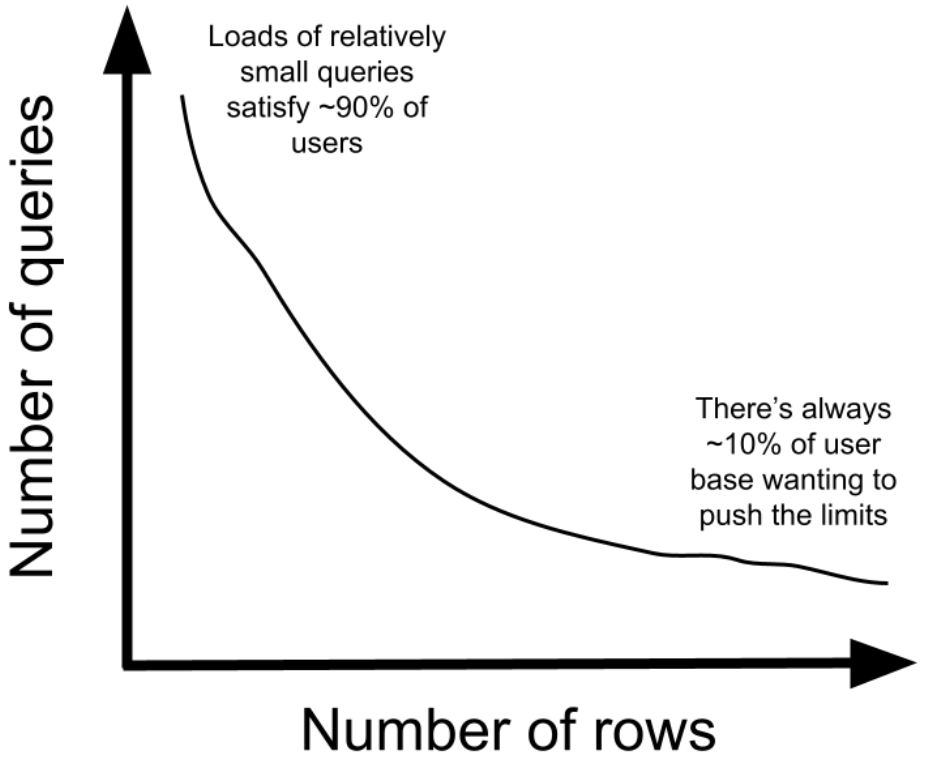
Notes:

- Time lag between end of observations and DR: *all data reprocessed from scratch with updated calibrations*
- Exponential growth in DR size: new high-level products*, and more time-resolved data (as well as mean products)*
- DRs 1, 2 & 3 bulk distributed in CSV.gz*: DRs 4 & 5 to be bulk distributed in Parquet format*

The long tail of scale-out usage scenarios

e.g. `Gaia data access scenarios summary'

- 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)
  - some 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



Example: *Gaia* epoch photometry

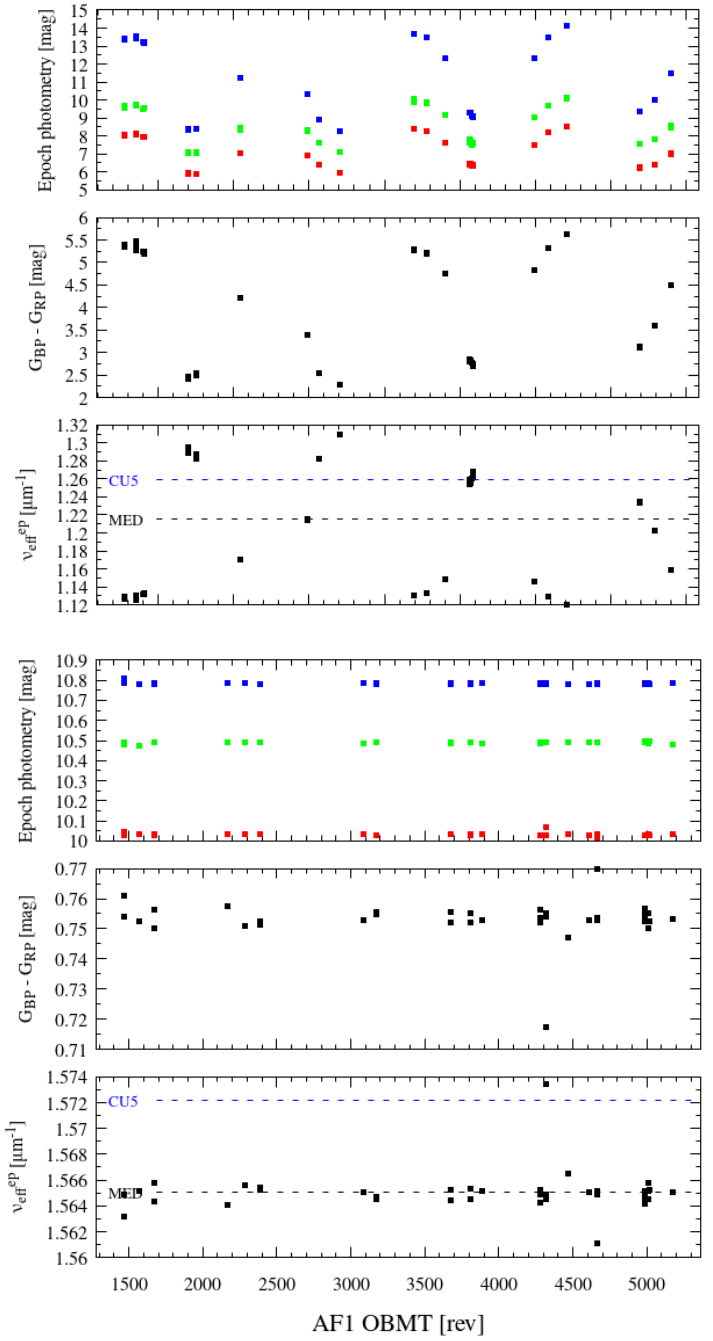
Time-resolved transit level G, BP and RP photometric analysis:

- DR3: 11.8 million records, total data volume 61 GB
- (DR4: ≈ 2 billion records, total data volume ≈ 10 TB)

Using the SPACIOUS-AstroFlow science platform

- Modest level of parallelism and memory (50 cores, 1.4 GB memory per core by default)
- DR3 analysis in around 15 seconds
  - Pre-filtered to 3 million variable candidates
  - Robust statistics: median colour and scatter

N.B. DR4 CCD-level (G only) photometry:
122 billion records, 21 TB !



Example: *Gaia* XP mean 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: 220 million records, total data volume 2.7 TB (8 TB in uncompressed eCSV…)
- (DR4: ≈ 2 billion records, total data volume 25 TB)
- basis-set ‘continuous’ representation
  - N basis coefficients
  - N coefficient uncertainties
  - N(N-1)/2 correlation coefficients
  - N = 55

Gaia XP mean spectra (cont.)

Large-scale use case: given an example template, find similar spectra

- Statistical rigour: compute the *Mahalanobis distance* (e.g. De Angeli et al. 2022) between the template and all others

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

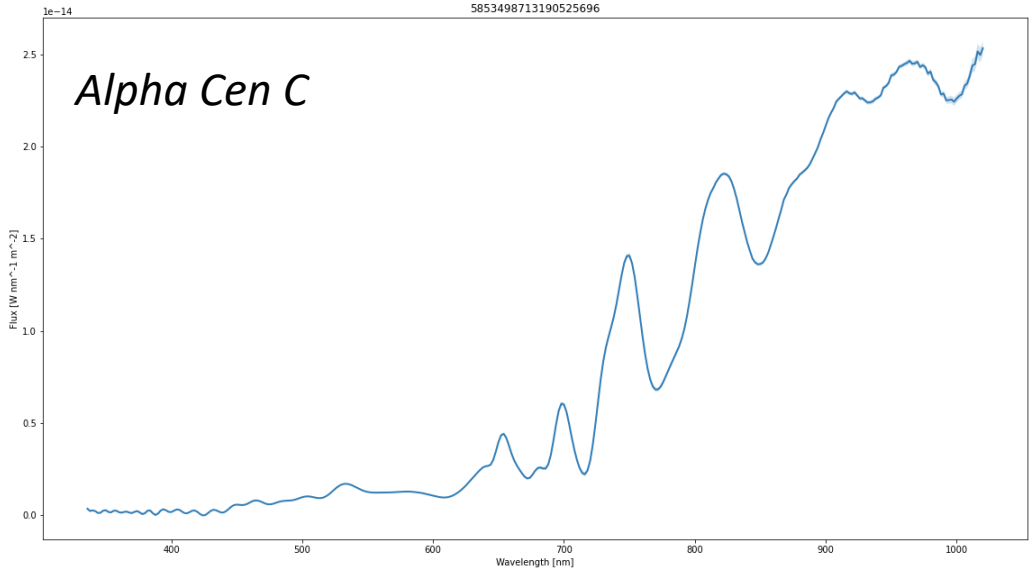
linear array follows a column-major storage scheme, i.e.:

$$\mathbf{M} = \begin{bmatrix} 1 & C[0] & C[1] & C[3] & C[6] & \cdots & C[S-(n-1)] \\ & 1 & C[2] & C[4] & C[7] & \cdots & C[S-(n-2)] \\ & & 1 & C[5] & C[8] & \cdots & C[S-(n-3)] \\ & & & 1 & C[9] & \cdots & C[S-(n-4)] \\ & & & & \ddots & \ddots & \vdots \\ & & & & & 1 & C[S-1] \\ & & & & & & 1 \end{bmatrix}$$

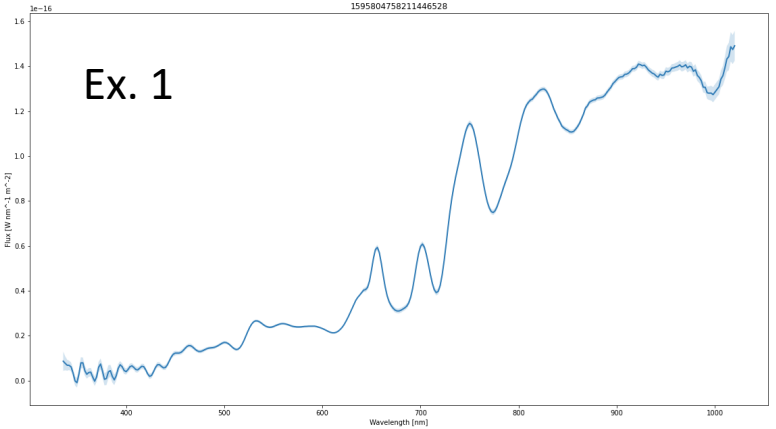
- In each case reconstruct the full 2d covariance matrix from the (flattened, 1d) correlation matrix and uncertainties vector
- matrix & vector multiplications implemented as a Pandas (vectorized) User Defined Function for execution on Spark cluster worker nodes

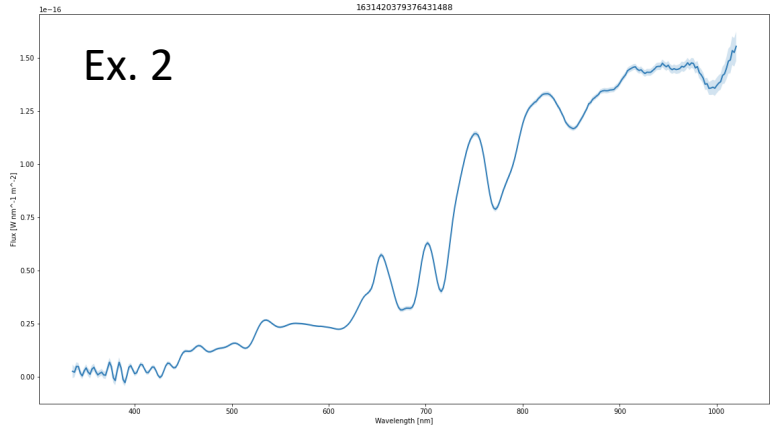
Gaia XP mean spectra (cont.)

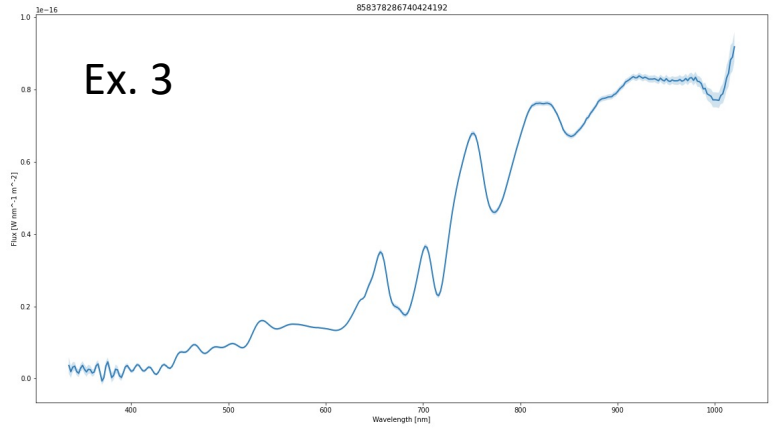
Top 3 Proxima Cen lookalikes in Gaia DR3 taking full account of variance / covariance in calibrated mean spectra:




```
# Request more memory than default import to handle large XP spectra data
spark_setup = SparkSetup(pods=10, cpu=5, mem=70, data_setup=SetupDataGaiaDR3)
spark = spark_setup.get_spark_session()
```







See demo Notebook 5 on the AstroFlow platform

- Configurable memory feature
- Takes 15 to 20 minutes

Programme summary

The workshop aim is for you to invest a little time to try out a code-to-data platform; we have spaced out the sessions:

- Facilities
  - The focus is on SPACIOUS-AstroFlow and ESA DataLabs
- Example usage scenarios
  - Focus on Gaia workflows
- Surgery sessions
  - **Sign up** to discuss your astrophysics and astronomy interests, usage scenario(s) and / or have a personal walk-through of existing usage examples at your own pace
  - If no slots left then email me for ad-hoc bookings following the workshop
- Wrap-up: **single show-and-tell slide please** from all participants, Friday morning
  - A little bit about you and your research
  - What you've learnt that is useful
  - What you're planning to do next

Meeting etiquette

Please

- Video and mic off unless you are speaking
- Raise hand if you would like to ask a question at the end of each presentation
- Upload slides (and/or any other material you'd like to share) to Indico
  - You will need to create an account to do this

Feel free to use the chat facility to communicate during the meeting

- Zoom meeting will remain open during the week