Gaia Data Mining Platform

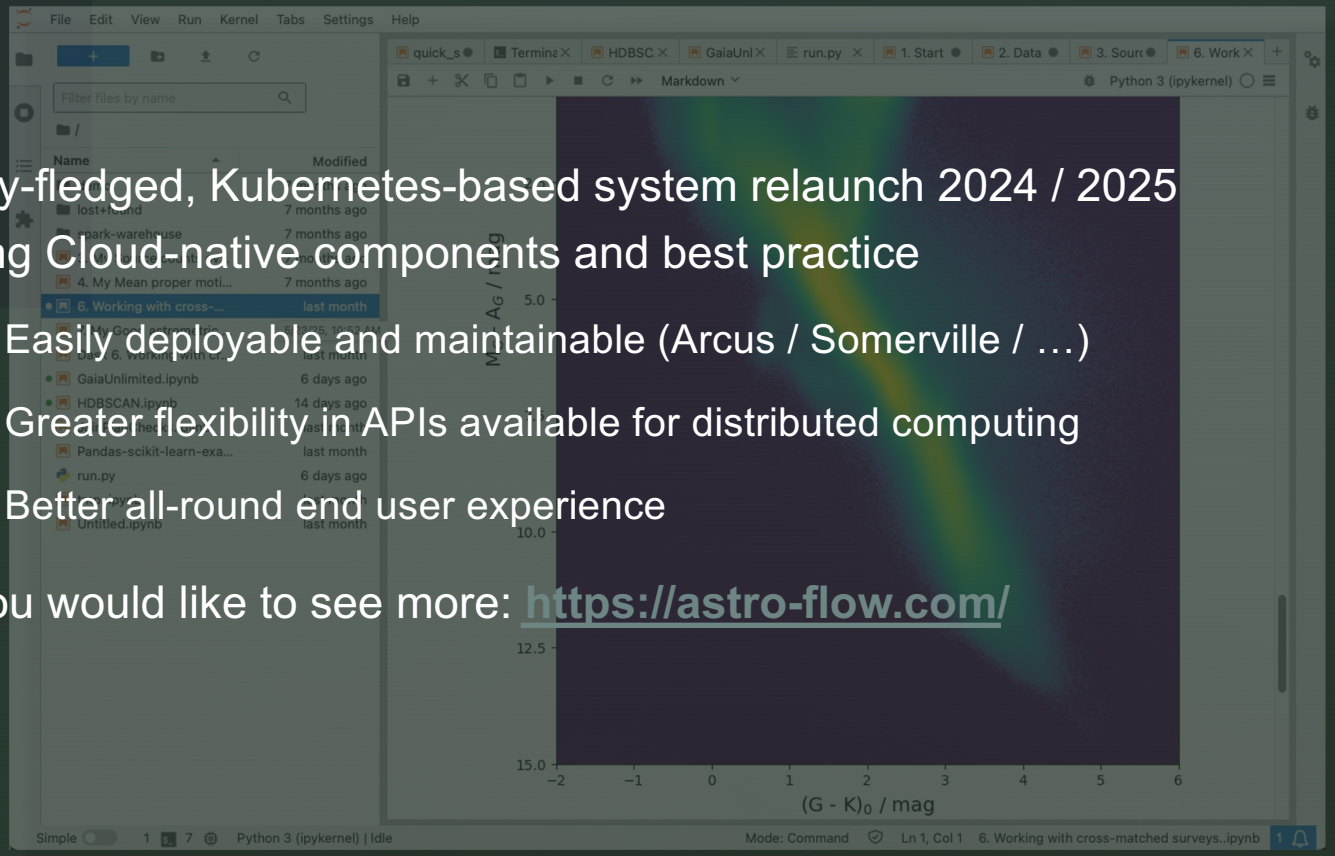
… *an end-user code-to-data science exploitation facility*

- The idea is to set up and maintain a code-to-data platform as a convenient end-user service
  - To facilitate exploitation of richly structured, static release products from ESA's *Gaia* mission
  - To be scalable from < 10 TB to > 1 PB
- Where we were ~ 10 years ago:
  - Bare-metal Apache Spark cluster prototype, 1 TB capability
  - Inflexible and unmaintainable service
- Step in IRIS:
  - First allocations in 2019 / 2020

Gaia DMP 1.0

- Cloud-deployed service (with a lot of learning along the way):
  - OpenStack provisioned clusters
  - A functioning service launched in 2021 (MVP) / 2022 (production)
  - Heavily based on Apache ecosystem (incl. Zeppelin notebooks)
  - Issues with maintainability and flexibility remained …
    - Manual deployment and end-user management
    - Resilience / availability

Gaia DMP 2.0



- Fully-fledged, Kubernetes-based system relaunch 2024 / 2025 using Cloud-native components and best practice
  - Easily deployable and maintainable (Arcus / Somerville / …)
  - Greater flexibility in APIs available for distributed computing
  - Better all-round end user experience
- If you would like to see more: **https://astro-flow.com/**

Thanks to IRIS / STFC …

- Where we are now: we have
  - Easily deployable and maintainable production service in the UK
    - Huge thanks to infrastructure people at Arcus and Somerville
  - ~ 70 registered end-users of the production deployment
    - See these workshop pages for recent illustrative science examples
  - Confidence in scalability from 10 TB (Gaia DR3, 2022) …
    - to Gaia DR4: 0.5 PB (end 2026), and
    - to Gaia DR5: > 1.0 PB (end 2030)