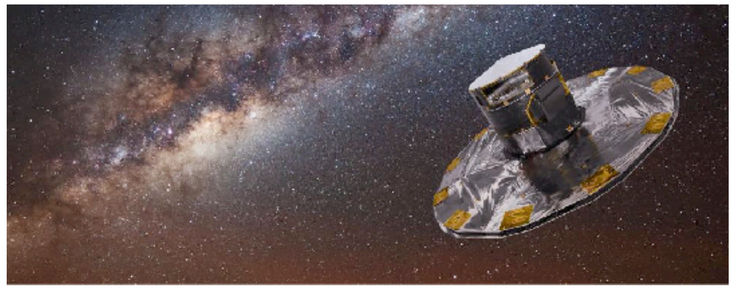
Gaia Core Processing

- Gaia is an ESA space observatory mission to form a precise 3D map of 2 billion stars and other objects within the Milky Way galaxy and beyond. Mapping motion, luminosity, temperature and composition.
- Launched in 2013: 3 trillion observations during 3827 days of science operations between July 2014 and decommissioning in January 2025.
- Gaia Core Processing activity is one of 6 Data Processing Centres around Europe performing on-ground data processing of the satellite data to produce the Gaia Data Releases. Running on dedicated Hadoop cluster from 2012, replaced with a larger version in 2016.
- IRIS resource allocation grew from 600 vCPU, 0.4PB disk in 2021/22 to ~12k vCPU, 4.6PB disk in 25/26.
  - OpenStack resources on Arcus cloud at Cambridge, deployed and configured via Ansible.
  - Running custom data processing software developed within the group built on Apache Spark, Java, Python, Parquet, etc.
- Since May 2024, all data processing at Cambridge has been carried out on IRIS resources.
- Previous dedicated hardware has been decommissioned.



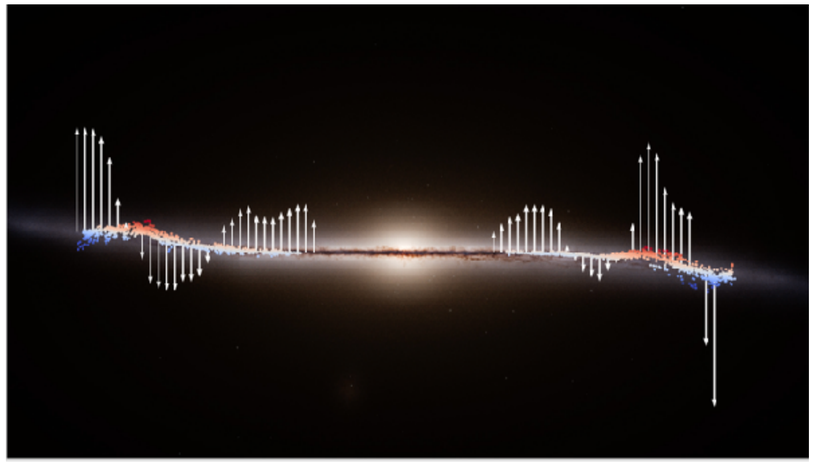
- Third Gaia Data Release DR3 published in June 2022.

- Cambridge data products for DR4 ~245TB delivered by January 2025. Publication of DR4 due in December 2026. Full migration of data processing to IRIS was carried out during the processing for DR4.

- Our processing for the final Gaia data release, DR5, begins in April 2026. Due for publication not before the end of 2030.

- Gaia data referenced in approximately 15k post-launch peer-reviewed publications (as of Dec 2025).

- Enabling research in many areas including solar system objects, stellar evolution, stellar variability, stellar populations, galactic structure and archaeology, galactic black holes, dark matter and more.



The Milky Way's great wave in motion.

ESA/Gaia/DPAC, S. Payne-Wardenaar, E. Poggio et al (2025)