L'ús del núvol en l'explotació de dades de la missió Gaia

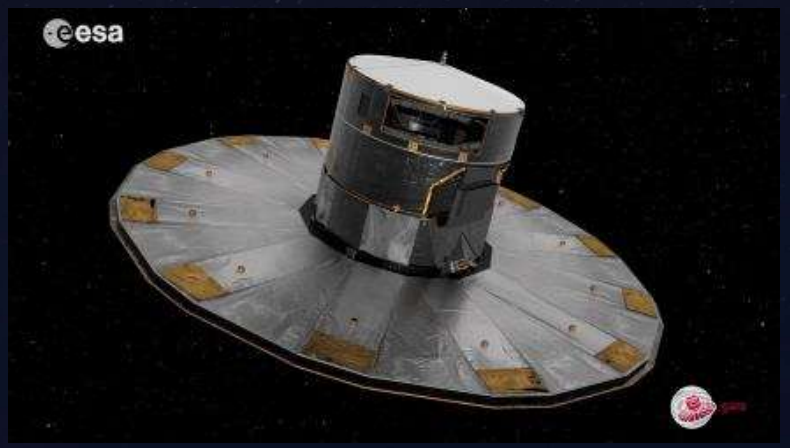
COMPUTACIÓ AL NÚVOL PER A LA RECERCA

Xavier Luri, ICCUB

Context

La missió Gaia

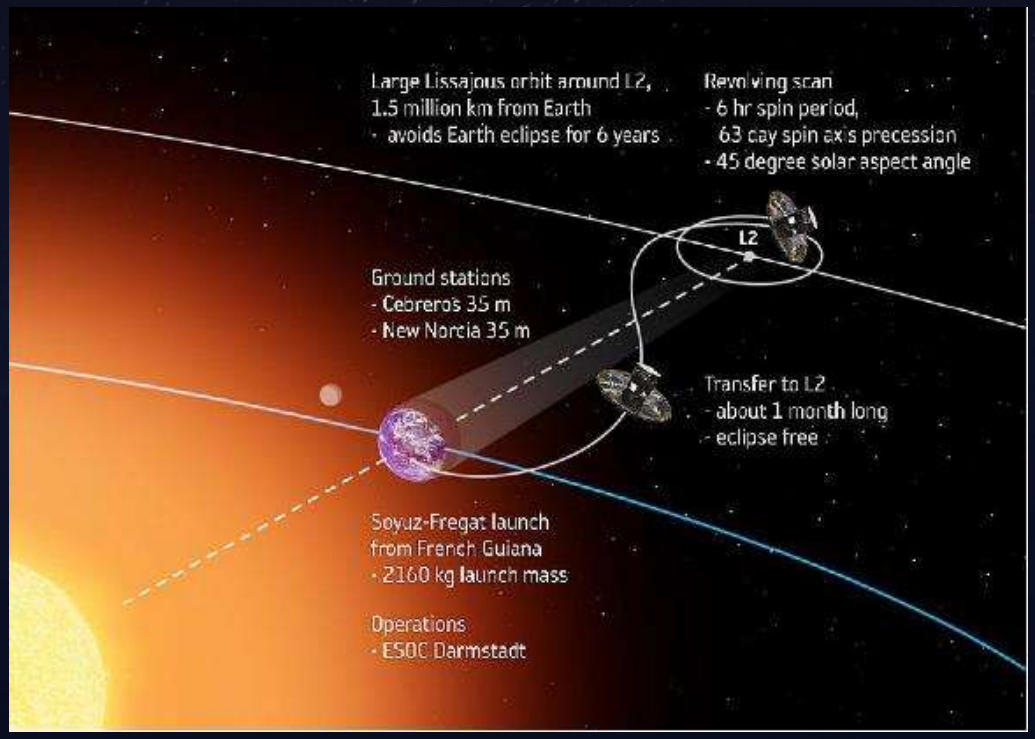
La missió



- Llançament el 19-dec-2013
- Operacions científiques 25-jul-2014
- Finalització prevista: Q1 2025

El satèl·lit

Images courtesy EADS-Astrium



Situat en una òrbita tipus Lissajous al voltant del punt de Lagrange L2 Terra-Sol, a uns 1.5 milions de quilòmetres de la Terra

El consorci DPAC

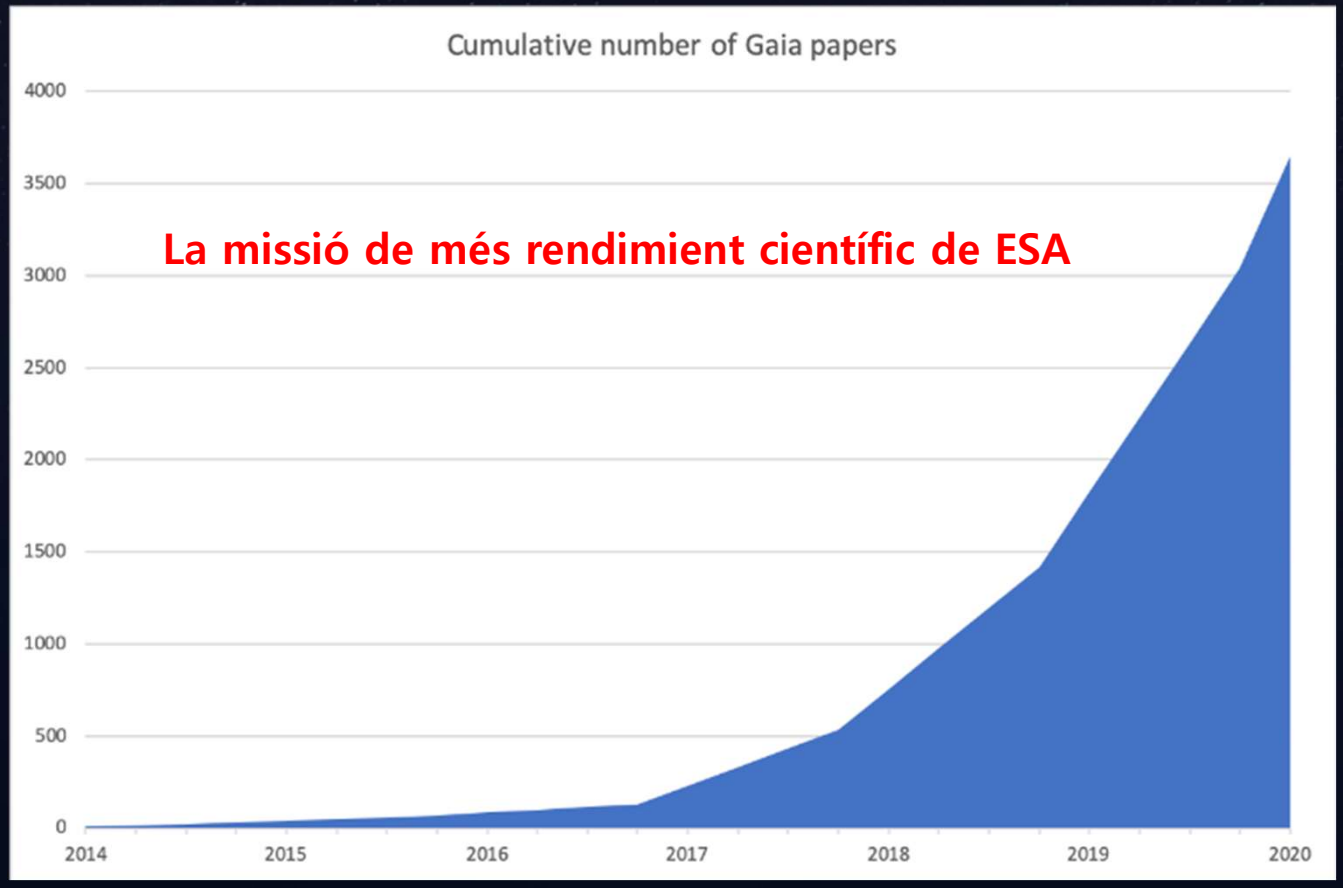
Gaia Data Processing and Analysis Consortium

- ~500 membres
- 24 agències de finançament

Duració MLA: 2007-2022+

Cost mitjà anual: ~30M€
Cost global: ~500M€

L'èxit de Gaia

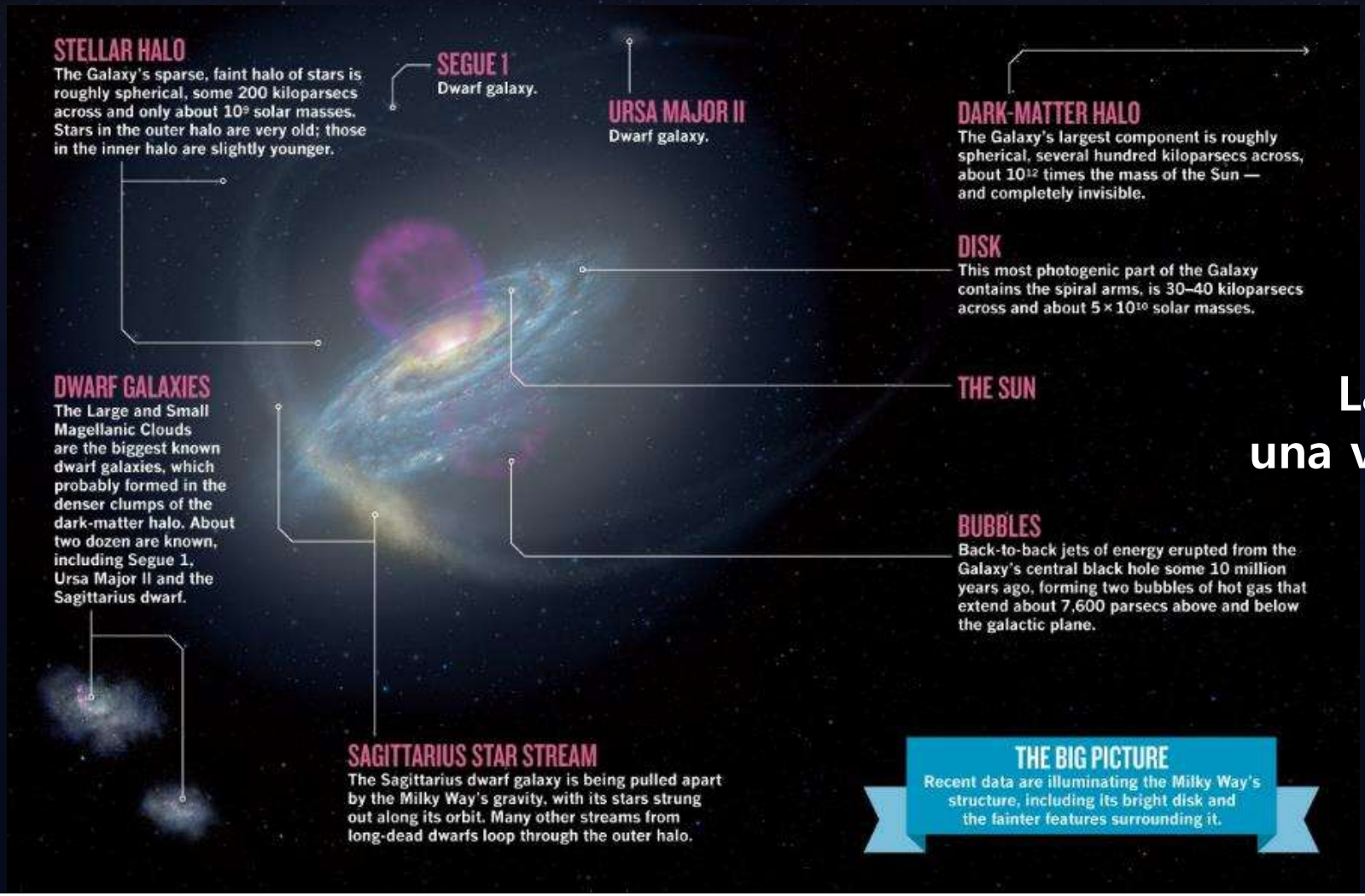


La missió de més rendimient científic de ESA

(NASA-ADS: 5549 articles a 3/10/2021)

Objetiu científic: la nostra Galaxia

Composició, evolució, dinàmica i estructura de la Vía Láctea



La Vía Làctia: una visió actualizada

Context

El grup Gaia a la UB

Gaia@UB

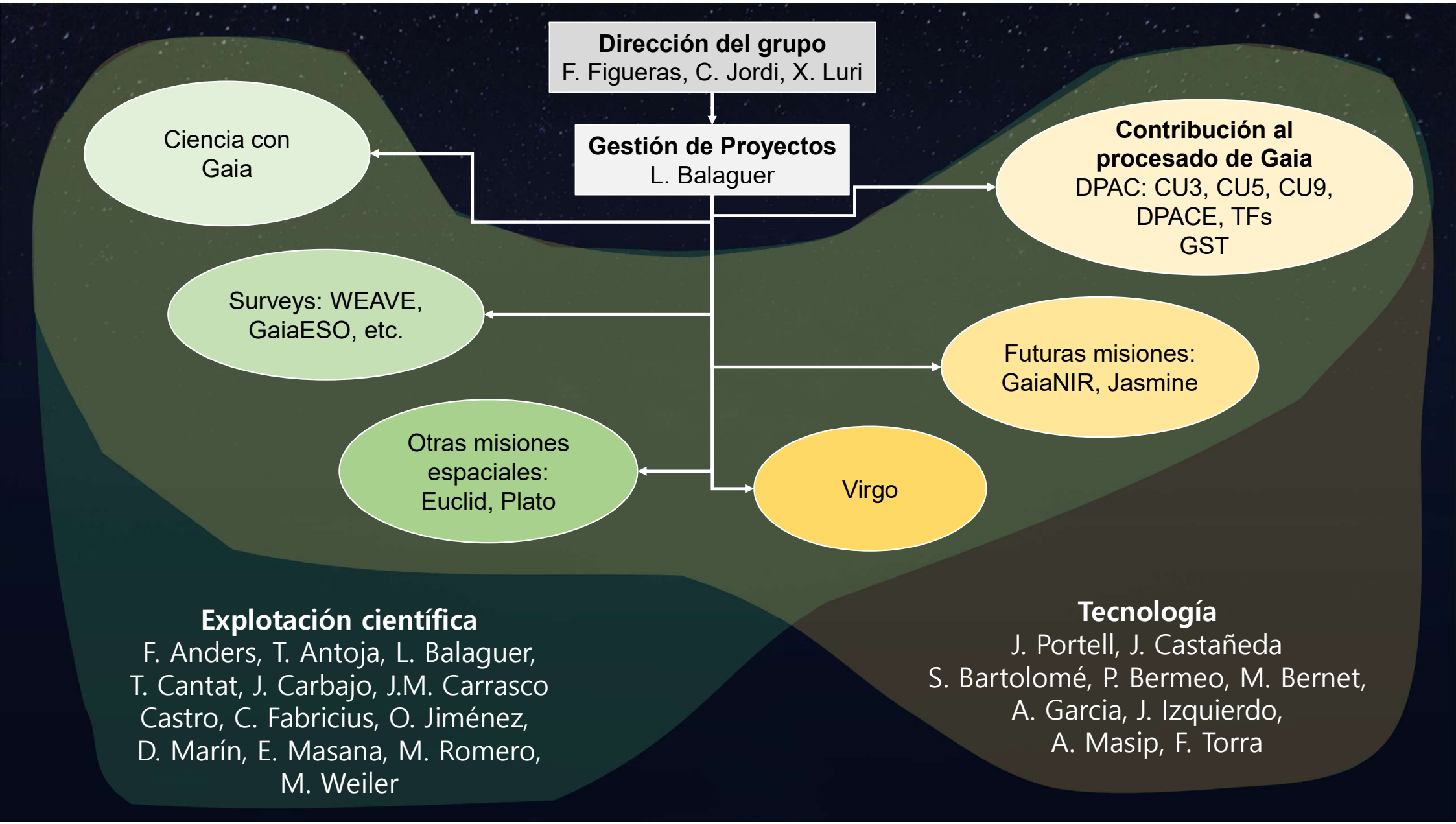
http://gaia.ub.edu

Equip Gaia a la UB

~27 membres
Científics/Enginyers

IPs: X. Luri, C. Jordi, F. Figueras

Equip multidisciplinari
Ciència i Tecnologia

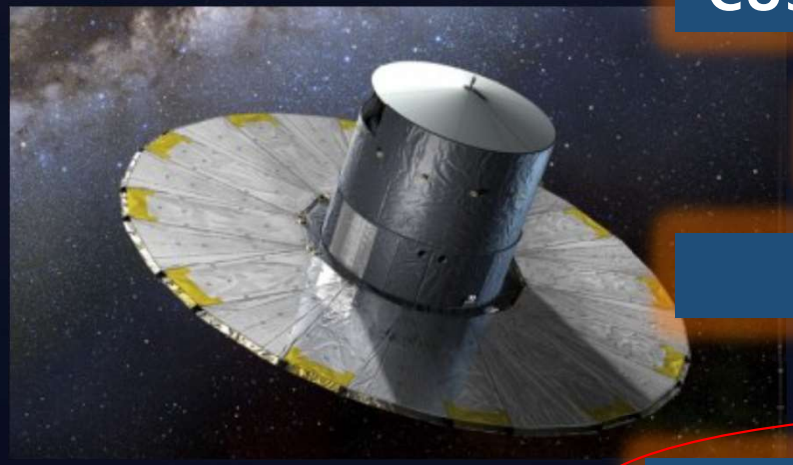

Dirección del grupo
F. Figueras, C. Jordi, X. Luri
Gestión de Proyectos
L. Balaguer
Ciencia con Gaia
Contribución al procesado de Gaia
DPAC: CU3, CU5, CU9, DPACE, TFs
GST
Surveys: WEAVE, GaiaESO, etc.
Futuras misiones: GaiaNIR, Jasmine
Otras misiones espaciales: Euclid, Plato
Virgo
Explotación científica
F. Anders, T. Antoja, L. Balaguer, T. Cantat, J. Carbajo, J.M. Carrasco Castro, C. Fabricius, O. Jiménez, D. Marín, E. Masana, M. Romero, M. Weiler
Tecnología
J. Portell, J. Castañeda
S. Bartolomé, P. Bermeo, M. Bernet, A. Garcia, J. Izquierdo, A. Masip, F. Torra

El projecte científic Gaia@UB

Història i evolució de la Galàxia
- història formació estel·lar
- funció de massa inicial
- disc i el seu ondulació
- barra central
- braços espirals
- acreció de galàxies
- poblacions estel·lars
- cúmuls estel·lars
- determinació de distàncies

Desenvolupament tecnològic
- compressió
- aparellament d'observacions
- mineria de dades
- intel·ligència artificial

Gaia



CU2 - Simuladors

CU3 - Astrometria

CU5 - Fotometria

CU9 - Arxiu

Big data

Mineria de dades

HPC/HTC

Cloud computing

Metodologia

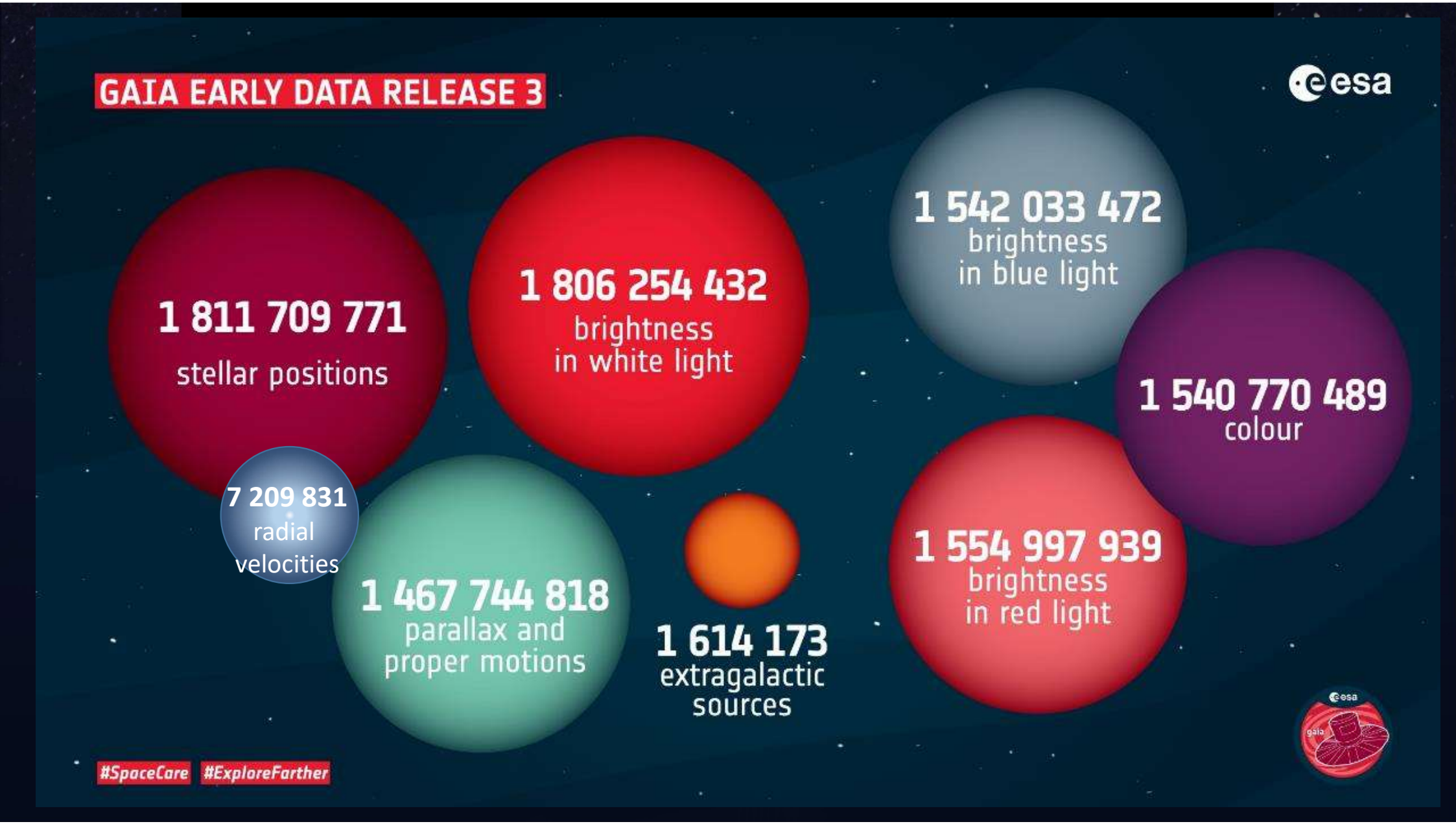
Línia de recerca

Computació avançada per a la missió Gaia

Necessitats de computació intensiva

Donat el volum d'informació que genera la missió Gaia el nostre equip ha hagut d'adquirir expertesa en diverses àrees:

- *High-Performance computing* i *High-Throughput Computing*
- Mineria de dades, Inteligència Artificial, Big Data
- Eines de computació: Spark-Hadoop, computació al núvol


GAIA EARLY DATA RELEASE 3
esa
1 811 709 771
stellar positions
1 806 254 432
brightness
in white light
1 542 033 472
brightness
in blue light
1 540 770 489
colour
7 209 831
radial
velocities
1 467 744 818
parallax and
proper motions
1 614 173
extragalactic
sources
1 554 997 939
brightness
in red light
#SpaceCare
#ExploreFarther
esa
gaia

MareNostrum (BSC-CNS)

- Projecte estratègic BSC-CNS
- Dedicat al processat de les dades de Gaia (IDU)
- Aplicacions de HPC
- Milions d'hores de CPU anuals

Clusters de rang mitjà (CSUC)

- Simulacions
- Desenvolupament de prototips de sistemes de processat (IDT)
- Explotació científica

Clusters i servidors propis (UB)

- Desenvolupament de programari
- Validació de dades
- Explotació científica
- Docència

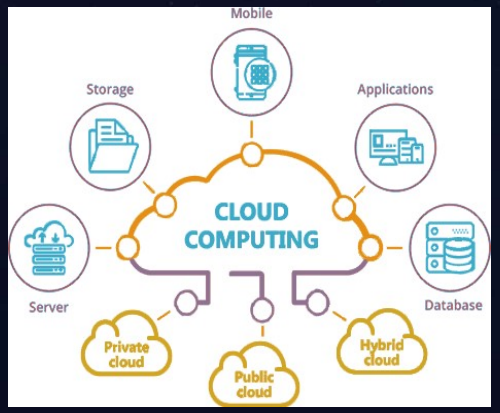
Eines per a l'explotació científica de Gaia

Objectiu: proporcionar eines avançades de computació i anàlisi per a l'explotació científica de Gaia

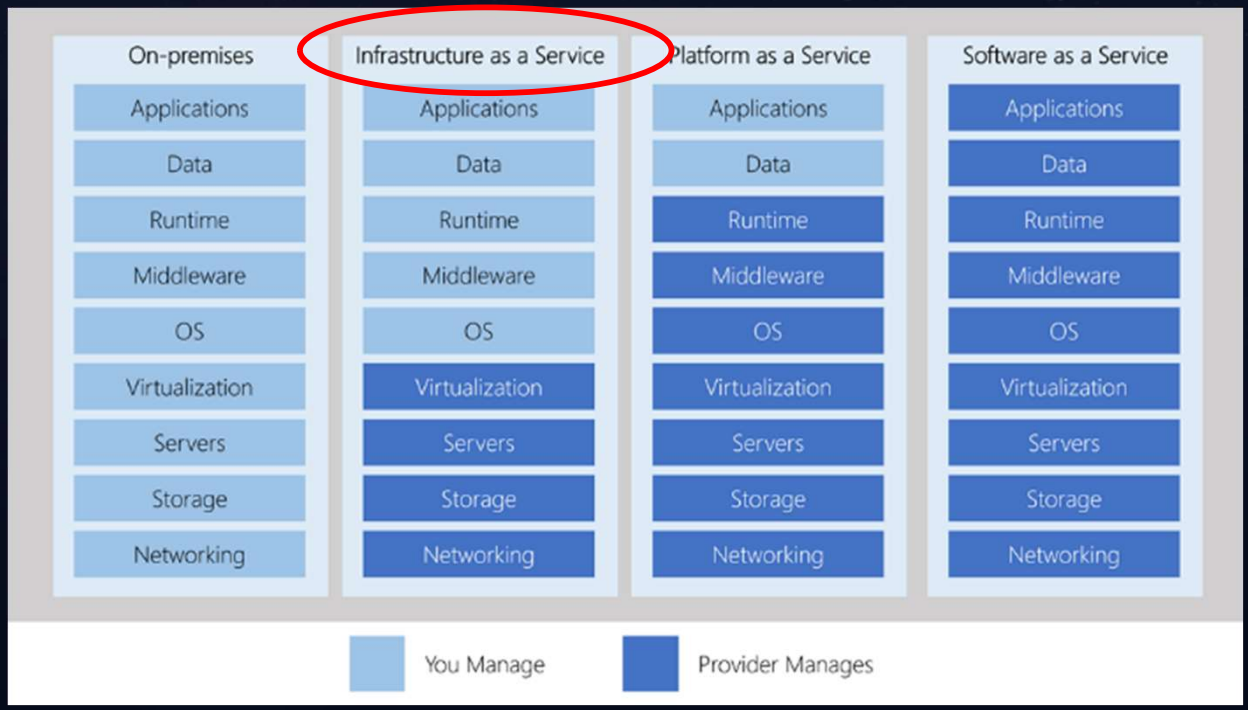
- Subproducte de les nostres responsabilitats en Gaia
- Basat en projectes europeus:
  - GENIUS
  - Spacious
  - OCRE
- Lligat a la "Red Española de Gaia", xarxa finançada pel programa "Redes de Excelencia" del ministeri, liderada per l'equip Gaia a la UB

Computació al núvol (I)

Al grup Gaia-UB vam començar a valorar l'ús de la computació al núvol després del projecte GENIUS, seguit de projectes H2020 per al seu desenvolupament en la explotació de Gaia.



No suposen estrictament una nova tecnologia, sinó més aviat una innovació en l'accés i forma d'ús dels recursos de computació



Tipus de serveis als núvol

Computació al núvol (II)

Objectiu: adquisició de *know-how* en l'ús de les grans núvols comercials (CCS)

- Aplicació a projectes propis: reducció de costos i flexibilitat
- Avaluació de la seva aplicació a l'explotació massiva de dades de Gaia com a futur a complement a l'arxiu: coordinació amb els prototips de ESA (Google cloud)
- Disseminació de l'expertesa en la comunitat Gaia (REG)

OCRE

Open Clouds for Research Environments

Accelerate cloud adoption in the European research community

Commodity-type commercial digital services necessary for interdisciplinary research activities.

OCRE: projecte Gaia UB

Open Clouds for Research Environments

100k€ en recursos de computació al núvol per a un projecte integrat d'explotació de dades de Gaia

We aim to use the unprecedented accuracy of the Early Data Release 3 (EDR3) of the European Space Agency mission Gaia published in December 2020, to investigate, for the first time in a holistic way, the Milky Way (MW) star formation and its interaction with its satellite galaxies.

Després de diversos retards s'iniciarà el novembre de 2021

We aim to use

- the Gaia EDR3 catalogue (1.8 billion stars)
- a Galaxy Model (BGM Fast)
- Bayesian techniques
- Machine Learning
- Data Mining tools
- an N-body simulation of interacting galaxies

to develop three complementary scientific analysis:

- to evaluate if the star formation history in the last 6-7 Gyrs is driven by the interaction of the Milky Way with the Sagittarius satellite galaxy;
- to perform an orbital analysis of the interaction of the Magellanic clouds with an unprecedented spatial and temporal resolution;
- to search for substructures in the Milky Way that are predicted by the cosmological models and reveal crucial information on its collision history and its Dark Matter content;

Our project will consume mainly IaaS and PaaS cloud services:

1. A Spark cluster (e.g. Azure HDinsight or Google Dataproc) for preliminary calculations and asingle big run for 8 days using 1,424 CPUs, complemented by 3TB in a Data-Lake for 10 months
2. A Linux Virtual Machine (VM) with 32 vCPUs for 3 months
3. A Linux VM up to 128 vCPUs or larger, depending the response capability of ADQL queries to the Gaia archive for approximately 100,000 hours of CPU;
4. A Machine Learning Service to use mainly the Scikit-Learn library and shared Jupyter Notebooks
5. The necessary data transactions and storage for Gaia data, catalogues, simulations and products of scientific analysis
6. A reserve a budget for VM redundancy, unexpected cloud costs or further calculations due to unexpected interesting results

Our project will highlight the key benefits of the Commercial Cloud Services (CCS) by demonstrating the efficiency of the scientific use of different cloud infrastructures.

- We will demonstrate the **agility** of the CCS by quickly deploying the data mining infrastructure needed for searching substructures in the Milky Way, in order to lead the publication of the analysis after EDR3.

- We will show the **scalability** of CCS by running the BGM FASt Galaxy Model in a Spark cluster (e.g. Dataproc or HDInsight).

- In parallel, the N-body code (ART) will run continuously in a big virtual machine during more than 70 days, illustrating the **high availability** (e.g. using redundancy in Azure Availability Zones) and the fault tolerance of CCS.

Our project as a whole will illustrate the flexibility of CCS and also its cost-effectiveness , showing how we can afford different types of computational infrastructures for a given period of time (1 to 10 months) paying only for the consumed resources.

Consells per usar el núvol

- Familiaritzeu-vos amb els serveis generals que ofereixen els proveïdors del núvol (màquines virtuals; bases de dades, Iaas, Paas, Saas)
- Comproveu quins serveis generals s'adapten a les necessitats computacionals del projecte
- Reuniu-vos amb alguns dels proveïdors. Poden ajudar a afinar les necessitats del projecte
- Definiu clarament els serveis que necessita el vostre projecte i quina potència computacional.
- Utilitzeu les calculadores de preus disponibles per estimar els pressupostos
- Deixeu-vos aconsellar per l'equip de la UB per gestionar correctament la contractació

Experiència personal

- El núvol és una eina més: habitualment complementa altres recursos, no els substitueix
- La dificultat administrativa per a la seva contractació ha sigut un obstacle significatiu per a la seva adopció, però aquest problema s'està resolent (licitacions GEANT, Rediris, CSUC, experiència UB)
- L'adopció del núvol requereix un canvi d'hàbits i mentalitat, sobretot en planificació i control de la despesa
- Aporta molta flexibilitat: disponibilitat a demanda, escalabilitat
- Hi ha molta documentació disponible i proves gratuïtes (Azure@UB)
- Corba d'aprenentatge raonable; no requereix més suport d'experts que altres recursos computacionals avançats (segons com en requereix menys)


Moltes gràcies