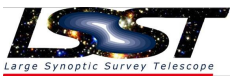
Big data in astronomical observations

William O'Mullane

Data management
LSST
Tucson, AZ USA

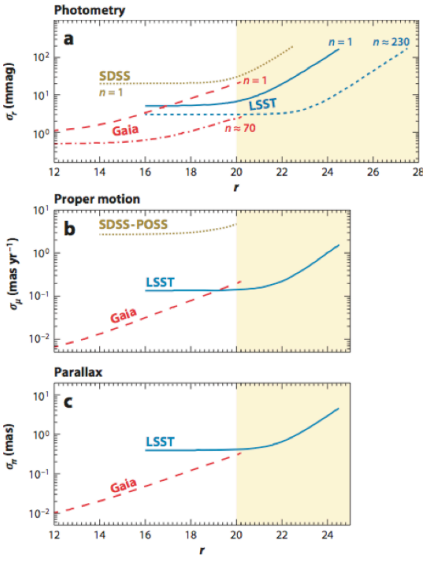
4rd July 2019
ISSI
Bern, Switzerland



A little about myself

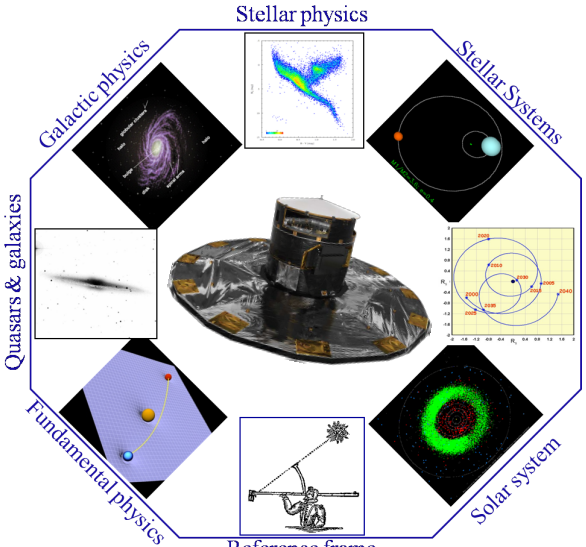
- 1985ish started with BASIC on a Commodore
- 1993 MSc BSc Computer Science, University College Cork, Ireland
- 1993 - 1997 Spacecraft Control Systems (C++), ESA ESOC Germany
- 1997 - 2001 Hipparcos, Integral, Planck, Gaia, Bepi-Sax (C,Java,Oracle, HTM, HEALPix), ESA ESTEC Netherlands
- 2001-2003 Commercial programming - some Astronomy (Java)
- 2003-2005 The Johns Hopkins, SDSS and National Virtual Observatory (C,C#,Java,Sqlserver)
- 2005-2014 Gaia Astrometric Solution and Science Operations (Java, Oracle, Intersystems Cache)
- 2012 PhD in Physics on Implementing the Gaia Astrometric Solution, Barcelona University
- 2014-2017 ESA Division head - all Science Ground Segments in Development
- 2017- LSST Data Management Project Manager (Python,C++), Deputy Project Manager for Software (control systems)

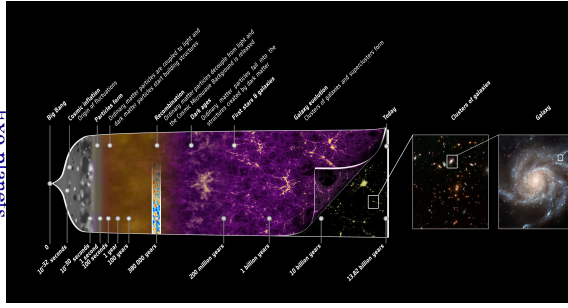
Gaia and LSST continuum



- Gaia: excellent astrometry (and photometry), but only to r > 20
- LSST: photometry to r < 27.5 and time resolved measurements to r < 24.5
- Complementary: photometric, proper motion and trigonometric parallax errors are similar around r=20

The Milky Way disk belongs to Gaia, and the halo to LSST (plus very faint and/or very red sources, such as white dwarfs and LT(Y) dwarfs). Zeljko Ivezić

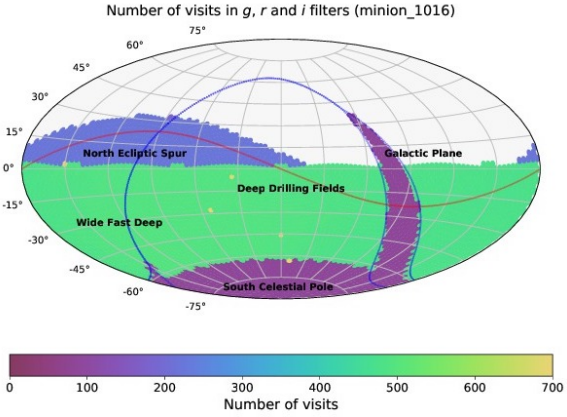

Stellar physics
Stellar Systems
Exo-planets
Solar system
Reference frame
Fundamental physics
Quasars & galaxies
Galactic physics


Big Bang
Cosmic inflation
Origin of elements
Recombination
Dark ages
First stars & galaxies
Galaxy evolution
Today
Clusters of galaxies
Galaxy

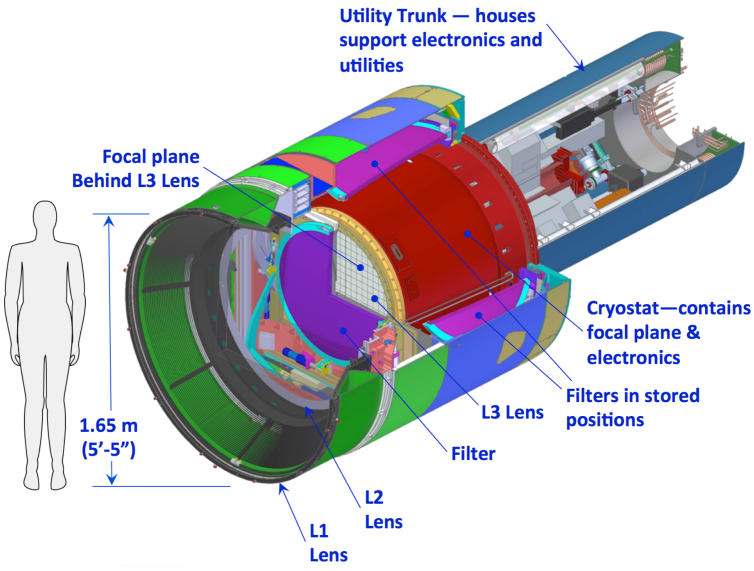
An optical/near-IR survey of half the sky in ugrizy bands to r 27.5 (36 nJy) based on 825 visits over a 10-year period: deep wide fast.

- 90% of time spent on uniform survey: every 3-4 nights, the whole observable sky scanned twice per night
- 100 PB of data: about a billion 16 Mpix images, enabling measurements for 40 billion objects!

see also http://www.lsst.org and Ivezic et al. (2008)-arXiv:0805.2366



10-year simulation of LSST survey: number of visits in u,g,r band (Aitoff projection of eq. coordinates)

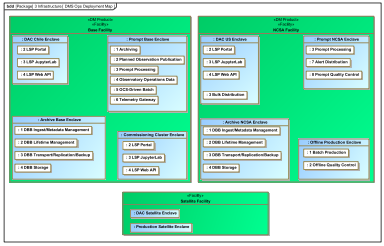


The largest astronomical camera:

- 2800 kg
- 3.2 Gpix

Site as imagined and in March 2019

DM build and deploy - already challenging



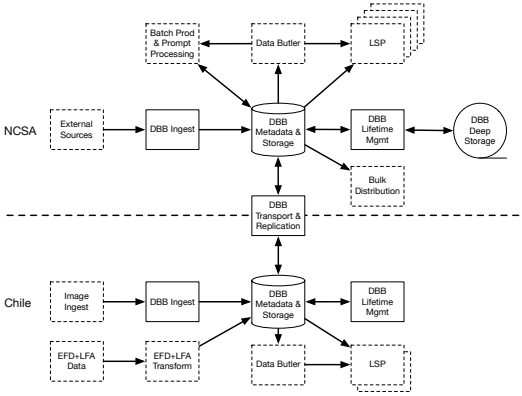
DM must build everything to get LSST products (see http://ls.st/dpdd) to the users.

- large data sets (20TB/night)
- complex analysis
- aiming for small systematics
- Science Alerts in under 2 minutes .. (aiming for 1 minute)

About $\frac{1}{2}$ million lines of code (C++/python) all open source on github

diagram K.T. Lim

Data Backbone

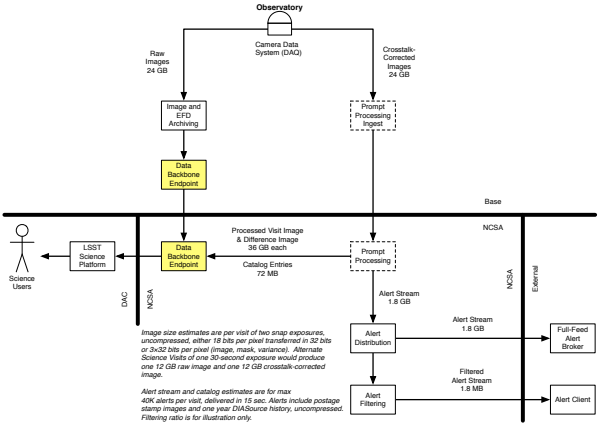


One small box on the previous slide was Data Backbone.
That hides several things

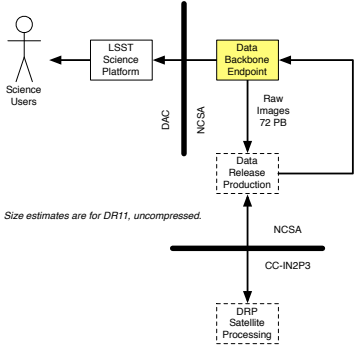
- Qserv - the LSST end user database
  - Custom Massive Parallel Processing (MPP)
  - allow queries on $\approx$ 20 Petabytes of tabular data
  - 4×10^{10} objects, 4×10^{13} sources (observations)
- All the networks : we now have fiber to the mountain and from La Serena to NCSA (two routes)

diagram K.T. Lim

Data flow



Lots to do every night ..
Plus annually there is a data release

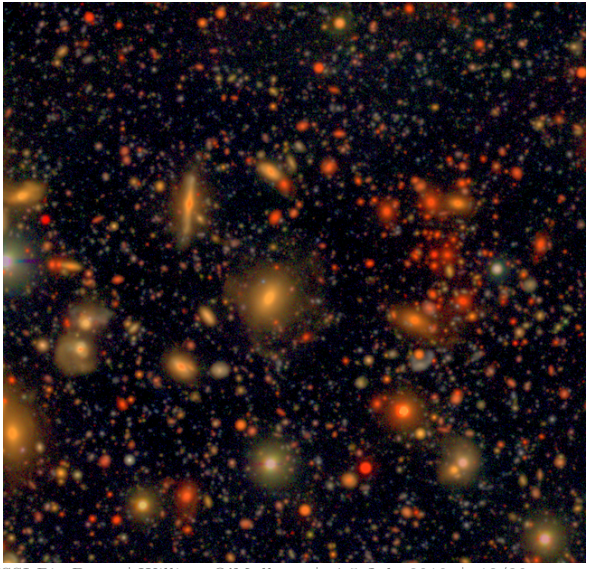


Images from K.T. Lim

SDSS image

Nice colors Lupton et al. (2004)
$\approx 3.5'$

Image Robert Lupton

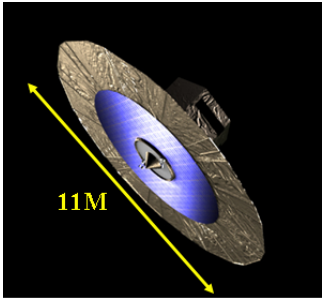


HSC image (COSMOS) from g,r(1.5 hrs) ,i(3 hrs) PSF matched co-add (≈ 27.5)
Challenges:

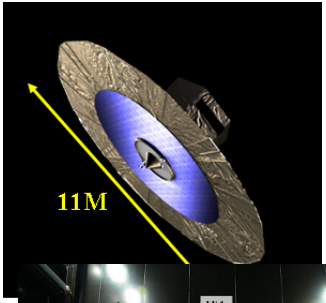
- Unknown statistical distributions, Truncated, censored and missing data, Unreliable quantities (e.g. unknown systematics and random errors)
- PSF - short exposure - atmosphere dominated ?
  - cosmic shear signal from weak lensing
- Photometry challenging - will Gaia help LSST ..
- Everything is blended!!

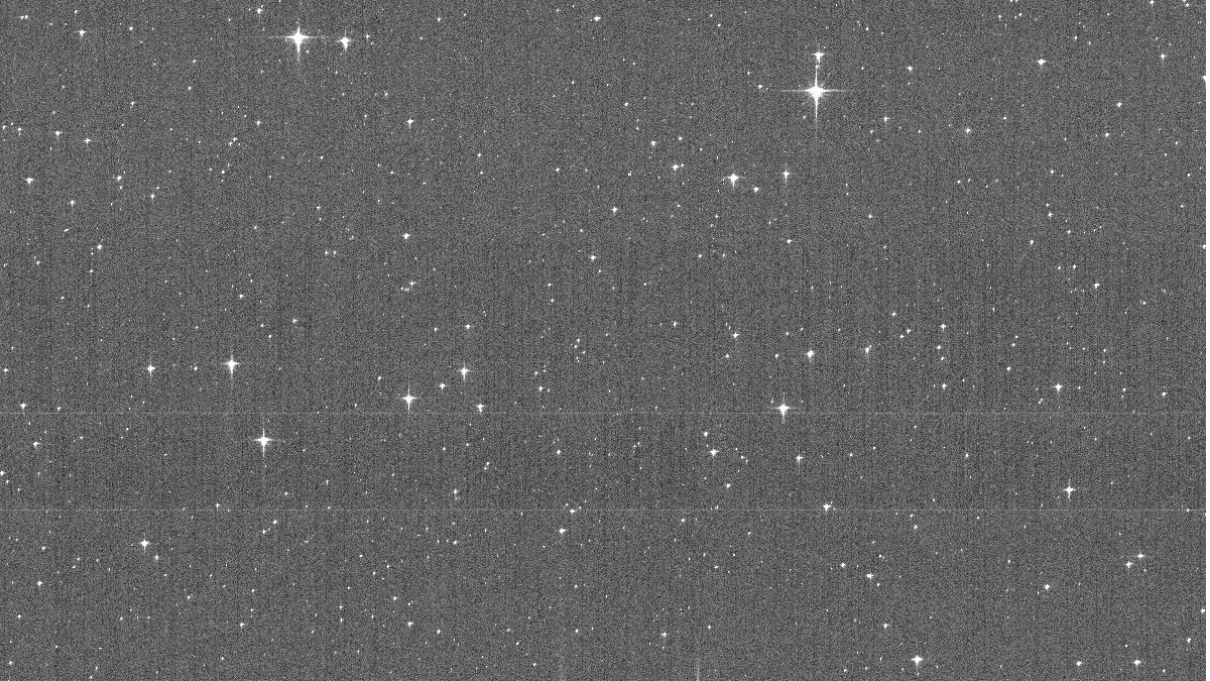
Satellite overview

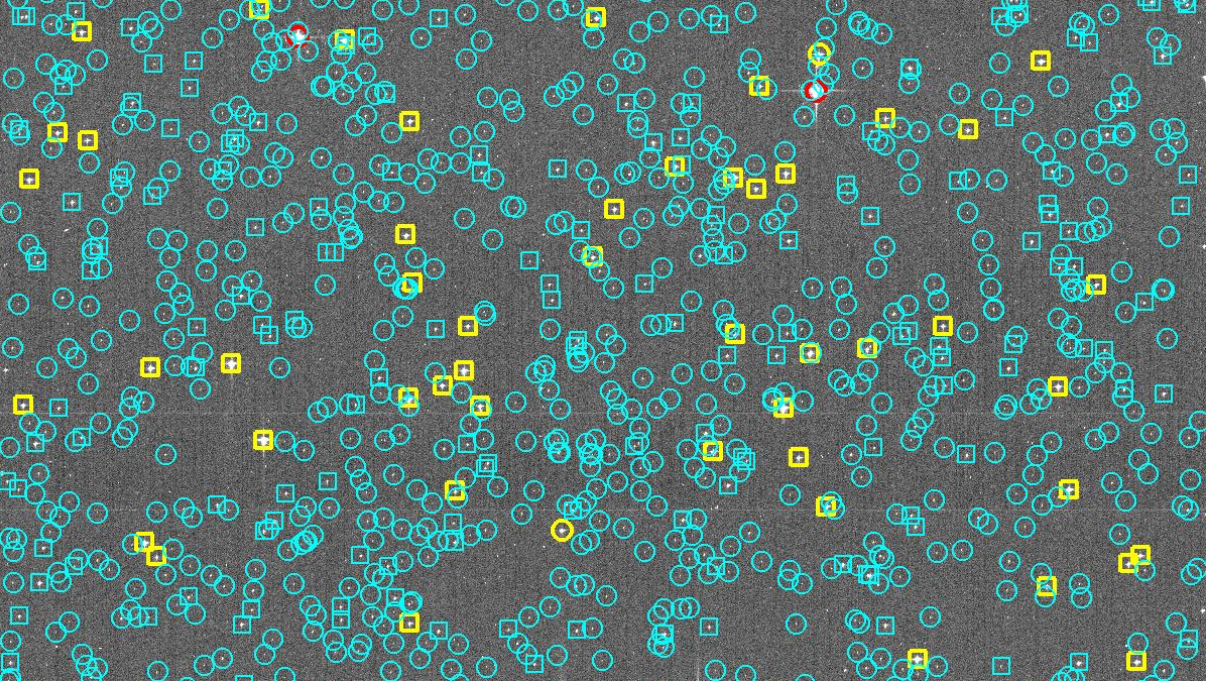
- Mission:
  - ESA Corner Stone 6
    - ESA provided the hardware and launch
      Mass: 2120 kg (payload 743 kg)
      Power: 1631 W (payload 815 W)
  - Launched December 19$^{\text{th}}$ 2013
  - Stereoscopic Census of Galaxy over 5 years
    - Extended 2 yrs - request for five
  - μarcsec Astrometry G < 20 (10^9 sources)
  - Radial Velocities G < 16
  - Photometry millimag G < 20



- Mission:
  - ESA Corner Stone 6
    - ESA provided the hardware and launch
      Mass: 2120 kg (payload 743 kg)
      Power: 1631 W (payload 815 W)
  - Launched December 19$^{\text{th}}$ 2013
  - Stereoscopic Census of Galaxy over 5 years
    - Extended 2 yrs - request for five
  - μarcsec Astrometry G < 20 (10^9 sources)
  - Radial Velocities G < 16
  - Photometry millimag G < 20
- Final catalogue $\approx$ 202?





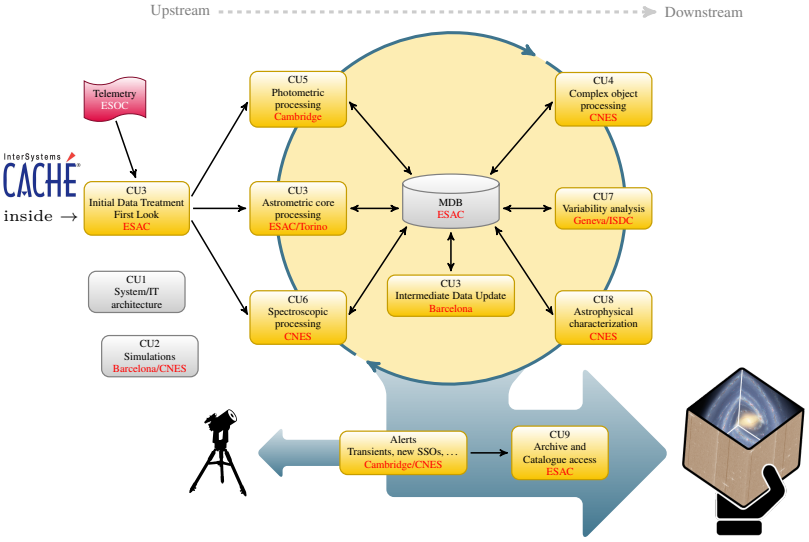


Data processing - challenging complex data

- Astrometric centroid of the CCD image to be determined to an accuracy of 1% of the pixel size!
  - There will be 10^{12} images $\approx$ 100TB downlink need to handle $\approx$ 1PB
  - At 1 millisecond each that is $\approx$ 30 years
- Reconstructed attitude is required to order 10 μarcsec
  - Path of light through instrument needed to nanometre level
  - System must be extremely stable
  - Must consider relativistic light bending from solar system objects.
- Attitude and Geometric calibration can only by done using Gaia's own observational data. (AGIS) (O'Mullane et al., 2011; Lindegren et al., 2012)
  - this requires a significant portion of the data to be processed iteratively

- Astrometric centroid of the CCD image to be determined to an accuracy of 1% of the pixel size!
  - There will be 10^{12} images $\approx$ 100TB downlink need to handle $\approx$ 1PB
  - At 1 millisecond each that is $\approx$ 30 years
- Reconstructed attitude is required to order 10 μarcsec
  - Path of light through instrument needed to nanometre level
  - System must be extremely stable
  - Must consider relativistic light bending from solar system objects.
- Attitude and Geometric calibration can only by done using Gaia's own observational data. (AGIS) (O'Mullane et al., 2011; Lindegren et al., 2012)
  - this requires a significant portion of the data to be processed iteratively

Simplified processing overview



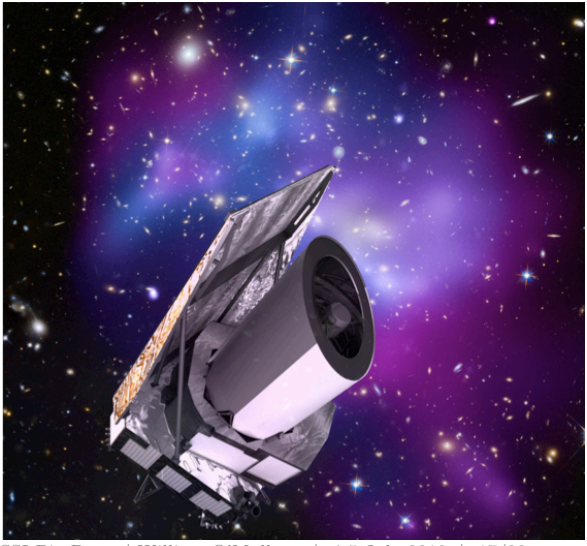
CU=Coordination Unit

Daily $50 - 100$GB compressed

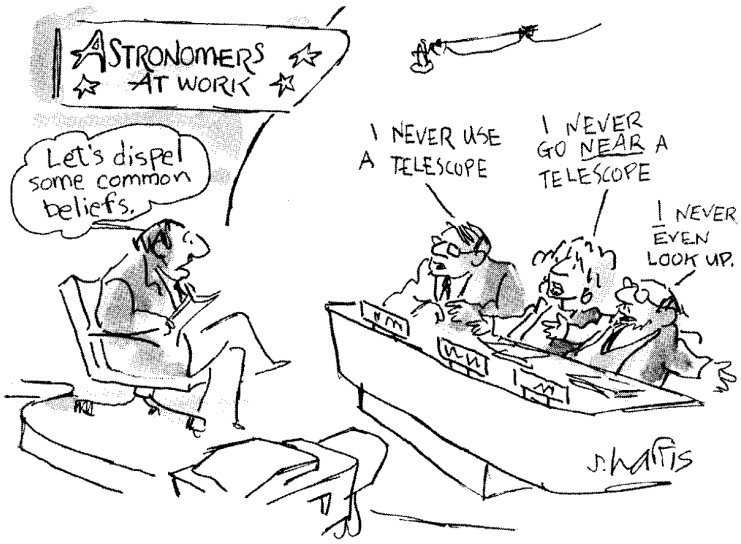
5yrs 1PB

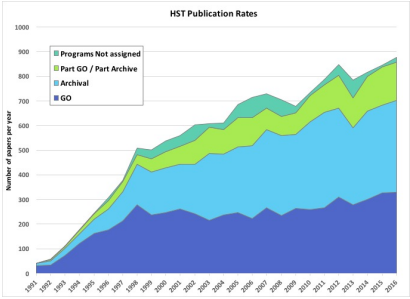
(see also O'Mullane et al. (2006), O'Mullane et al. (2009))

Next for ESA will be Euclid



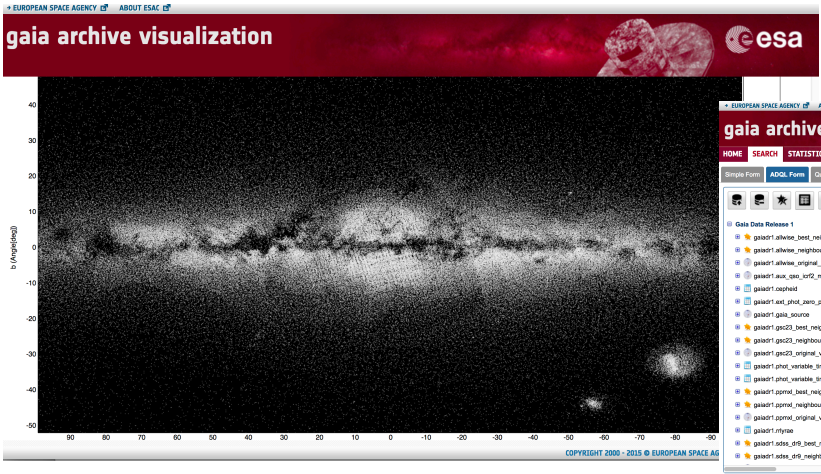
- Gaia had 8Mbps downlink - Euclid will have 55Mbps!
  - No planned outages in Data Acquisition
- A little like LSST in terms of goals
  - Very demanding core science requirements.
- Complex ground segment (and consortium)
  - Large legacy science community; both the data right holders (Consortium) and the wider community through public data releases.
- Estimated 30PB of date over 6 years
- Launch date 2022

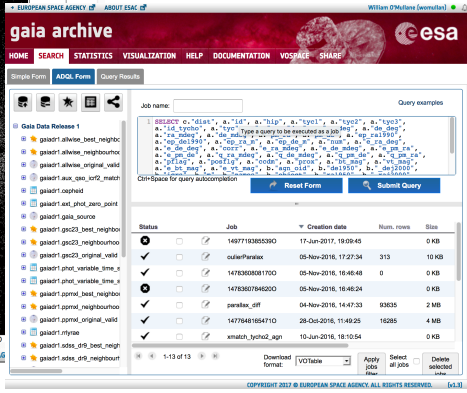




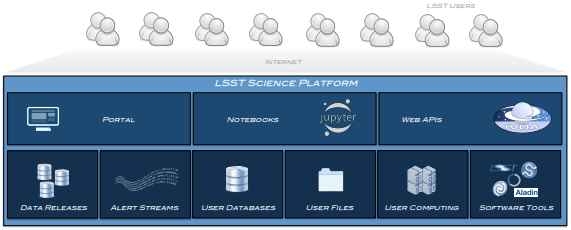
https://archive.stsci.edu/hst/bibliography/pubstat.html

...indicates archival research probably play an important role in the scientific success of XMM-Newton Ness et al. (2014)





All Gaia data is publicly accessible at https://gea.esac.esa.int/archive/



For DR2:

- Computing:2,400 cores ($\approx$ 18 TFLOPs)
- File storage: $\approx$ 4PB (VOSpace)
- Database storage: $\approx$ 3PB (MYDB)

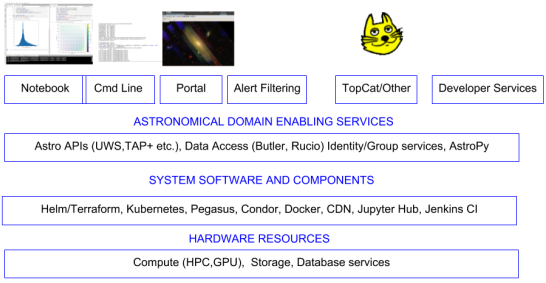
The Science Platform has three user facing aspects: the Portal (novice), the JupyterLab (intermediate), and the Web APIs (expert and remote tools).
Vision: LSE-319 — Design: LDM-542 — Test: DMTR-51

This is the sort of environment users now expect to have - it is no longer novel. We are finally Bringing code to the data - almost did with GAVIP Vagg et al. (2016)
Bruno will have more to say on this topic.

See also Bauer et al. (2019)
Soon, if not already, Data will be looking for astronomers not vice versa. Looking at Hubble archive 7K datasets have no publications.

- Proprietary data may have had its day .. if we want people to look at data we need to remove barriers.
- Networks and infrastructure are improving but more needs to be done
- Who looks after all the data ?
  - For space science in Europe ESAC preserves data - it is relatively small
  - IPAC, HEASARC and STScI pretty much deal with NASA data - still not one location.
  - Who looks after all the ground based data ?
  - There is no long term preservation plan for LSST or other big telescopes in the USA (Alex wil probably have more to say on this)

We all constantly redesign and rebuild wheels, this will become too expensive as data volume grows ..



- We should agree a component based cyber infrastructure model and work on improving specific components to plug in - right now we are all building TAP, Designing Databases , deploying Jupyter ...
- Data models like CAOM from IVOA are going to be essential going forward so we can inter-operate on data

Filesystems are End Of Life - object stores should not be confused with the object databases many of us struggled with in the 90s and 00s .. Google and Amazon do not run filesystems they run object stores.

- LSST allows 9 months for a Data Release Processing Cycle - probably about 6 months actual processing.
- The original requirement for Gaia astrometric solution was three months.
- Traditional batch systems and shared nothing architecture may not always work
  - Gaia Astrometric Solution required temporal spatial access and had global matrices
  - There are processes for LSST which will have similar problems e.g. Forward Global Calibration Model (FGCM) https://github.com/lsst/fgcmcal (Burke et al., 2018)
- preparing and staging for tasks can take ever longer as our processing becomes more sophisticated
- And then there is machine learning!
  - Automated discovery - need it great !
  - Reproducibility, understandability ..

Software is no longer an afterthought

- In ESA (personal opinion) Gaia was a game changer - the Software was seen as critical - I had to fight to keep it off the Launch Critical Items List. We still had a meager 10% or so of the budget.
- LSST recognized early on that the processing was as important as the telescope - DM is one of 4 Subsystem with about 20% of the project budget
- We need to consider long term software support and the software eco system
  - Open source and publicly scrutinized algorithms
  - Agree community cyber infrastructure model (previous slide)
  - Though I made my career in astronomy/computing .. its not easy .. and I really had to do the management thing
- Education and outreach - How do we do better?
  - Does everyone need to be a data scientist/programmer ?
  - Educate astronomers/managers on how to open source
  - How do we foster inclusion ?

Its all team work on big projects

The End

Acronym	Description
AGIS	Astrometric Global Iterative Solution
Archive	The repository for documents required by the NSF to be kept. These include documents related to design and development, construction, integration, test, and operations of the LSST observatory system. The archive is maintained using the enterprise content management system DocuShare, which is accessible through a link on the project website www.project.lsst.org.
CAOM	Common Astronomical Observation Model
CCD	Charge-Coupled Device
CI	Continuous Integration
CU	Coordination Unit
Camera	The LSST subsystem responsible for the 3.2-gigapixel LSST camera, which will take more than 800 panoramic images of the sky every night. SLAC leads a consortium of Department of Energy laboratories to design and build the camera sensors, optics, electronics, cryostat, filters and filter exchange mechanism, and camera control system.
DM	Data Management
DMTR	DM Test Report
Data Backbone	The software that provides for data registration, retrieval, storage, transport, replication, and provenance capabilities that are compatible with the Data Butler. It allows data products to move between Facilities, Enclaves, and DACs by managing caches of files at each endpoint, including persistence to long-term archival storage (e.g. tape).
Data Management	The LSST Subsystem responsible for the Data Management System (DMS), which will capture, store, catalog, and serve the LSST dataset to the scientific community and public. The DM team is responsible for the DMS architecture, applications, middleware, infrastructure, algorithms, and Observatory Network Design. DM is a distributed team working at LSST and partner institutions, with the DM Subsystem Manager located at LSST headquarters in Tucson.

Acronyms II

Data Release	The approximately annual reprocessing of all LSST data, and the installation of the resulting data products in the LSST Data Access Centers, which marks the start of the two-year proprietary period.
ESA	European Space Agency
ESAC	European Space Astronomy Centre
ESOC	European Space Operations Centre
ESTEC	European Space Technology Engineering Centre
FGCM	Forward Global Calibration Model
GAVIP	Gaia Added Value Interface Platform
GB	Gigabyte
HEALPix	Hierarchical Equal-Area iso-Latitude Pixelisation
HEASARC	NASA's Archive of Data on Energetic Phenomena
HSC	Hyper Suprime-Cam
HTM	Hierarchical Triangular Mesh
IPAC	No longer an acronym; science and data center at Caltech
IR	Infra Red
ISSI	International Space Science Institute
IVOA	International Virtual-Observatory Alliance
LDM	LSST Data Management (Document Handle)
LSE	LSST Systems Engineering (Document Handle)
LSST	Large Synoptic Survey Telescope
MPP	Massively Parallel Process
NASA	National Aeronautics and Space Administration
NCSA	National Center for Supercomputing Applications
Operations	The 10-year period following construction and commissioning during which the LSST Observatory conducts its survey
PB	PetaByte

PSF	Point Spread Function
Project Manager	The person responsible for exercising leadership and oversight over the entire LSST project; he or she controls schedule, budget, and all contingency funds
Qserv	Proprietary Database built by SLAC for LSST
SDSS	Sloan Digital Sky Survey
Science Platform	A set of integrated web applications and services deployed at the LSST Data Access Centers (DACs) through which the scientific community will access, visualize, and perform next-to-the-data analysis of the LSST data products.
Subsystem	A set of elements comprising a system within the larger LSST system that is responsible for a key technical deliverable of the project.
TAP	Table Access Protocol
TB	TeraByte
XMM	X-ray Multi-mirror Mission (ESA; officially known as XMM-Newton)
arcmin	arcminute minute of arc (unit of angle)
arcsec	arcsecond second of arc (unit of angle)
astrometry	In astronomy, the sub-discipline of astrometry concerns precision measurement of positions (at a reference epoch), and real and apparent motions of astrophysical objects. Real motion means 3-D motions of the object with respect to an inertial reference frame; apparent motions are an artifact of the motion of the Earth. Astrometry per se is sometimes confused with the act of determining a World Coordinate System (WCS), which is a functional characterization of the mapping from pixels in an image or spectrum to world coordinate such as (RA, Dec) or wavelength.
calibration	The process of translating signals produced by a measuring instrument such as a telescope and camera into physical units such as flux, which are used for scientific analysis. Calibration removes most of the contributions to the signal from environmental and instrumental factors, such that only the astronomical component remains.
camera	An imaging device mounted at a telescope focal plane, composed of optics, a shutter, a set of filters, and one or more sensors arranged in a focal plane array.

metric	A measurable quantity which may be tracked. A metric has a name, description, unit, references, and tags (which are used for grouping). A metric is a scalar by definition. See also: aggregate metric, model metric, point metric.

References

Bauer, A.E., Bellm, E.C., Bolton, A.S., et al., 2019, arXiv e-prints, arXiv:1905.05116 (arXiv:1905.05116), ADS Link

[DMTR-51], Bosch, J., Chiang, H.F., Gower, M., et al., 2017, LDM-503-02 (HSC Reprocessing) Test Report, DMTR-51, URL https://ls.st/DMTR-51

Burke, D.L., Rykoff, E.S., Allam, S., et al., 2018, AJ, 155, 41 (arXiv:1706.01542), doi:10.3847/1538-3881/aa9f22, ADS Link

[LDM-542], Dubois-Felsmann, G., Lim, K.T., Wu, X., et al., 2017, LSST Science Platform Design, LDM-542, URL https://ls.st/LDM-542

Ivezic, Z., et al., 2008, ArXiv e-prints (arXiv:0805.2366), ADS Link

[LSE-319], Jurić, M., Ciardi, D., Dubois-Felsmann, G., 2017, LSST Science Platform Vision Document, LSE-319, URL https://ls.st/LSE-319

Lindegren, L., Lammers, U., Hobbs, D., et al., 2012, A&A, 538, A78 (arXiv:1112.4139), doi:10.1051/0004-6361/201117905, ADS Link

Lupton, R., Blanton, M.R., Fekete, G., et al., 2004, PASP, 116, 133 (arXiv:astro-ph/0312483), doi:10.1086/382245, ADS Link

Ness, J.U., Parmar, A.N., Valencic, L.A., et al., 2014, Astronomische Nachrichten, 335, 210 (arXiv:1311.5751), doi:10.1002/asna.201312001, ADS Link

O'Mullane, W., Lammers, U., Bailer-Jones, C., et al., 2006, ArXiv Astrophysics e-prints (arXiv:astro-ph/0611885), ADS Link

O'Mullane, W., Hernández, J., Hoar, J., Lammers, U., 2009, In: D. A. Bohlender, D. Durand, & P. Dowler (ed.) Astronomical Data Analysis Software and Systems XVIII, vol. 411 of Astronomical Society of the Pacific Conference Series, 470, ADS Link

O'Mullane, W., Lammers, U., Lindegren, L., Hernandez, J., Hobbs, D., 2011, Experimental Astronomy, 31, 215 (arXiv:1108.2206), doi:10.1007/s10686-011-9248-z, ADS Link

Vagg, D., O'Callaghan, D., O'Hógáin, F., et al., 2016, In: Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, vol. 9913 of SPIE, 99131V (arXiv:1605.09287), doi:10.1117/12.2233619, ADS Link