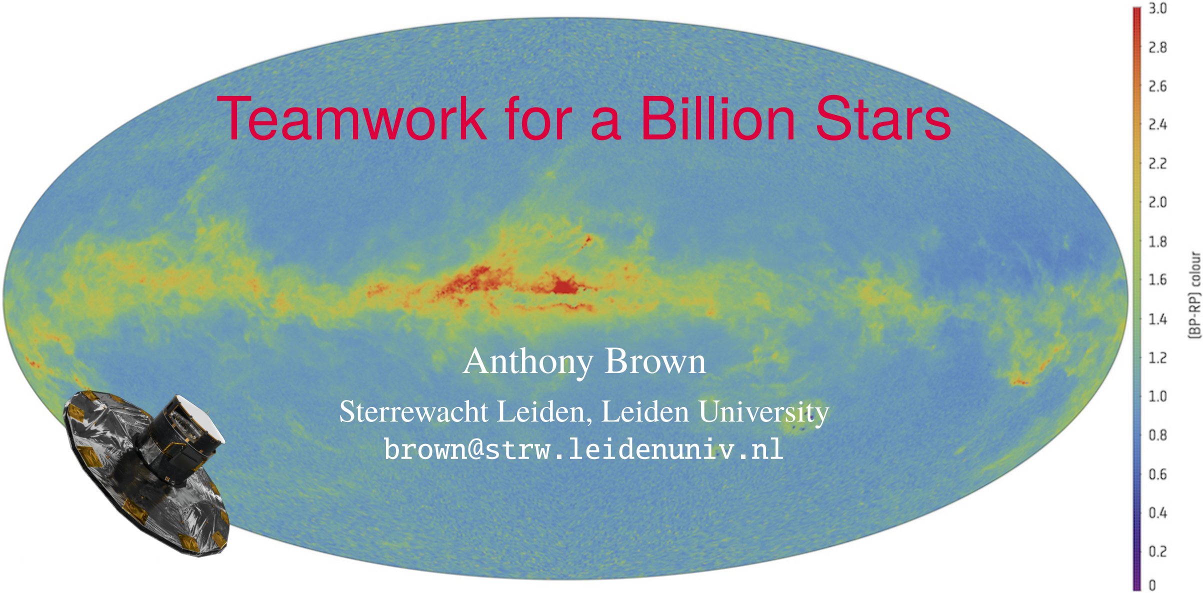
Teamwork for a Billion Stars

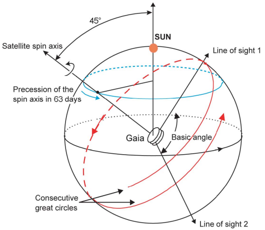


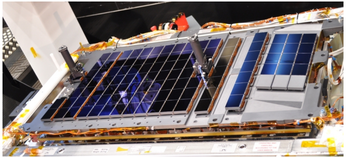
Anthony Brown

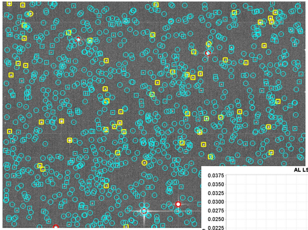
Sterrewacht Leiden, Leiden University

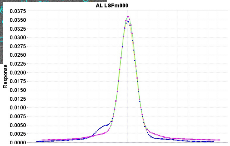
brown@strw.leidenuniv.nl

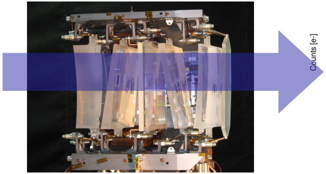
Gaia instruments and measurements

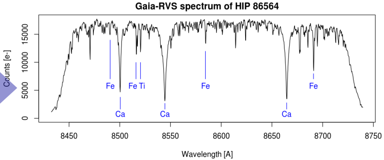


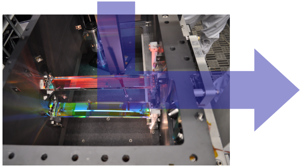


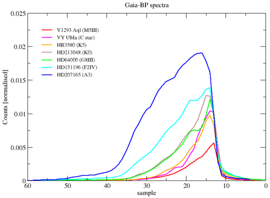


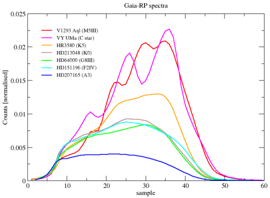












Find the source parameters

α, δ, ϖ, $\mu_{\alpha*}$, μ_δ, v_{rad}, orbit parameters multiple stars,
G, colours, T_{eff}, [Fe/H], $\log g$, A_0, solar system object orbits,
light curves, variable star classification, ...

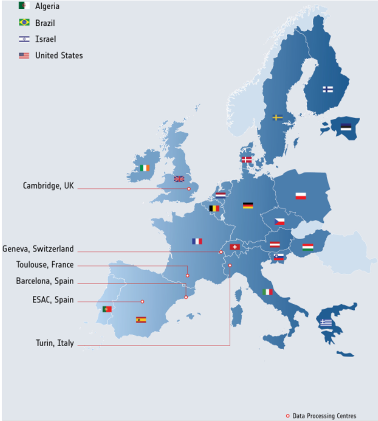
and instrument (calibration) parameters

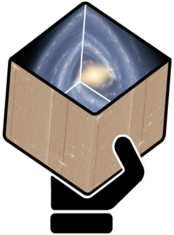
{Collection of parameters describing Gaia}

that best explain the Gaia observations.

Teamwork to deliver the promise of Gaia

- 10+ years of effort
- 550 scientists and engineers
- 160 institutes
- 24 countries and ESA
- Six data processing centres
- Funding: national space/funding agencies and ESA



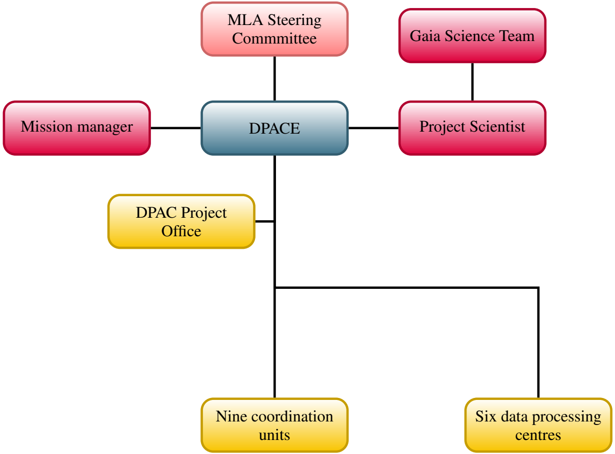


0 1 0 0 0 0 0 0 1 1 0 0 1 0 0 1 1 0 0 1 1 0 1 1 0 0 1 0 1 0

Data Processing & Analysis Consortium

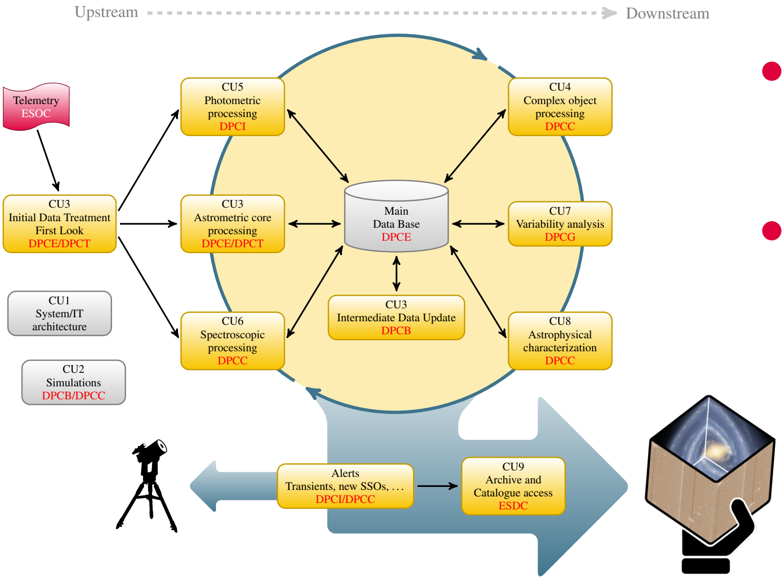
α δ ϖ $\mu_{\alpha*}$ μ_δ G ...

DPAC organization and structure



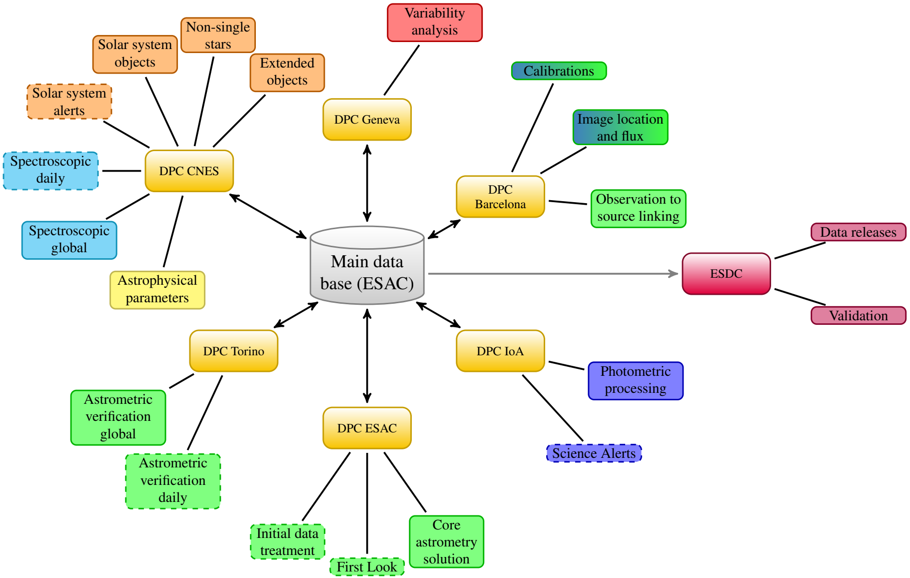
- Coordination Units (CUs)
  - organized around specific processing task
  - mix of scientists and IT specialists
  - design and development of scientific algorithms, validation/release of results
  - spread over many institutes
- Data Processing Centres (DPCs)
  - IT infrastructure (HW and SW)
  - integration and operation of CU codes
  - six physical locations
- Project Office
  - day-to-day coordination and monitoring; planning and scheduling
  - the 'glue' between the DPAC entities
- DPAC Executive
  - overall scientific coordination

DPAC structure and data flows

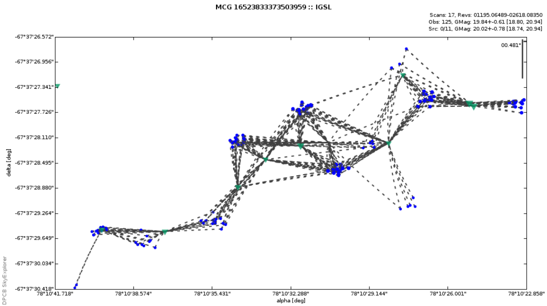


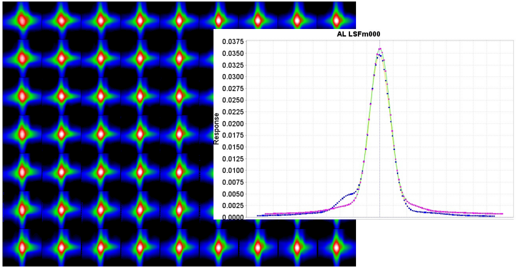
- Daily (real time) data processing
  - detailed payload health monitoring, alerts, initial calibrations
- Cyclic processing
  - achieve ultimate accuracy through iterating DPAC systems
  - increasing amounts of data treated each cycle
  - results to Gaia data releases

DPAC structure and data flows



Example: the path to eclipsing binaries





First processing steps (CU3/CU5/DPCB)

1. Linking observations to sources
2. Basic calibrations (PSF, background, bias, CTI)
3. Determine image locations and flux
   - account for source colour and CTI

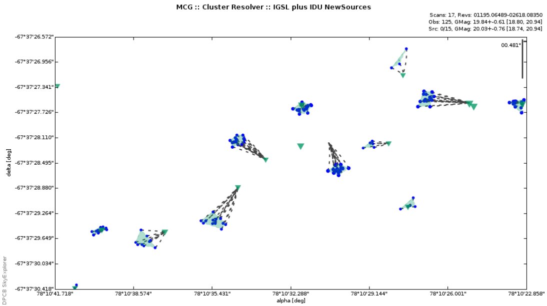
Inputs:

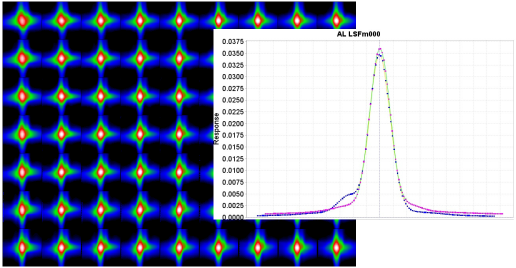
- Raw data (for a given fraction of mission length)
- Source list, photometry, spacecraft attitude, and astrometry from previous processing cycle

Outputs:

- Match table, calibrations, image location and flux

Example: the path to eclipsing binaries





First processing steps (CU3/CU5/DPCB)

1. Linking observations to sources
2. Basic calibrations (PSF, background, bias, CTI)
3. Determine image locations and flux
   - account for source colour and CTI

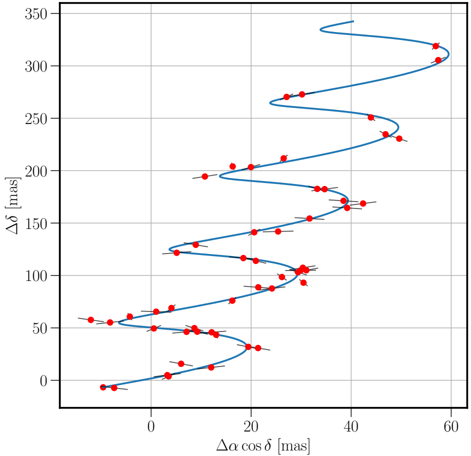
Inputs:

- Raw data (for a given fraction of mission length)
- Source list, photometry, spacecraft attitude, and astrometry from previous processing cycle

Outputs:

- Match table, calibrations, image location and flux

Example: the path to eclipsing binaries



Astrometric Global Iterative Solution (AGIS, CU3/DPCE)

1. Model spacecraft attitude
2. Model field-of-view to focal plane transformation (geometric calibration)
3. Account for relativistic effects of solar system bodies
4. Solve for source astrometric parameters

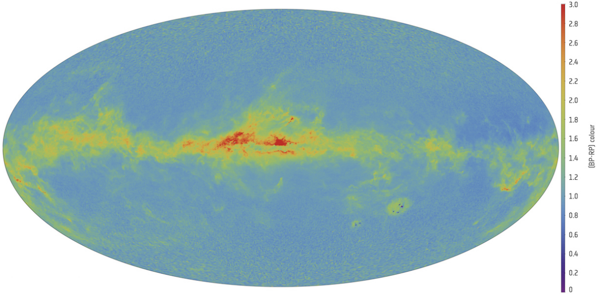
Inputs:

- Match table
- Image locations

Outputs:

- Epoch astrometry for all sources
- Spacecraft attitude as a function of time
- Geometric calibration as function of time

Example: the path to eclipsing binaries



Average $(G_{\rm BP} - G_{\rm RP})$ for sources at $G < 17$

<u>Photometric processing (CU5/DPCI)</u>

1. Establish internal photometric system
2. Solve for source photometry on this system
3. Tie to physical flux/wavelength scales through spectrophotometric standard stars

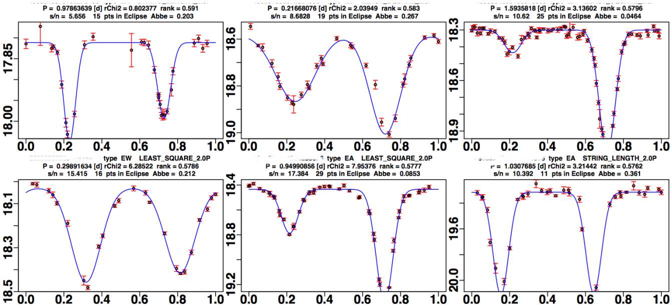
Inputs:

- Match table, image flux in G
- Source astrometry, spacecraft attitude, geometric calibration
- Raw BP/RP data and basic calibrations (electronic bias)

Outputs:

- Epoch photometry in G, $G_{\rm BP}$, $G_{\rm RP}$
- BP/RP spectrophotometry (Gaia DR3+)
- Pass-band, wavelength, flux calibrations

Example: the path to eclipsing binaries



<u>Variable source processing (CU7/DPCG)</u>

1. Analyse photometric time series
2. Classify and parameterize variable source types

Inputs:

- $G(t)$, $G_{\mathrm{BP}}(t)$, $G_{\mathrm{RP}}(t)$
- Source parallaxes

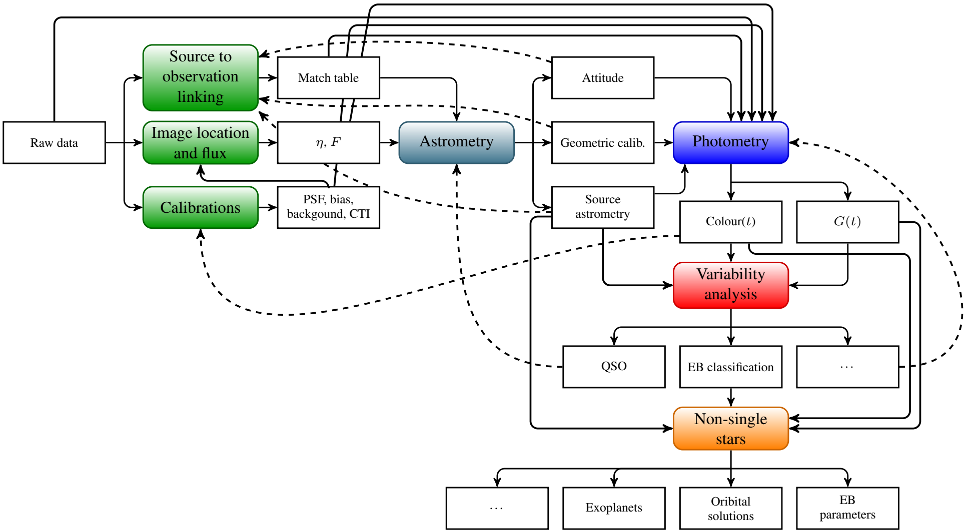
Outputs:

- Variable source classification
- Variable source parameters

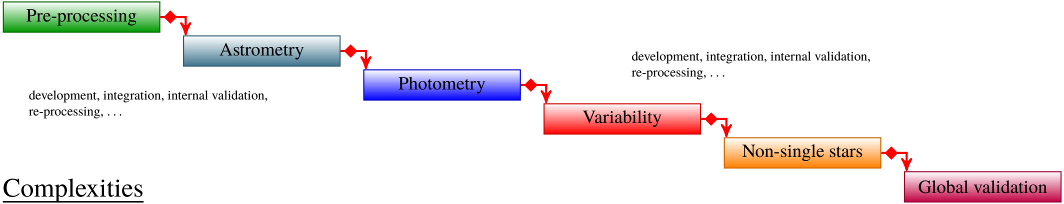
<u>Non-single star processing (CU4/DPCC)</u>

- Full modelling of non-single stars using inputs from upstream DPAC systems
  - in this example: determine eclipsing binary star parameters

Example: the path to eclipsing binaries



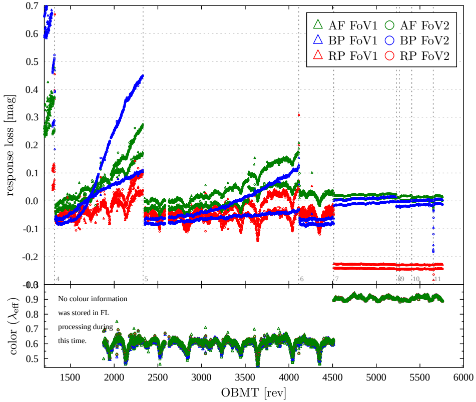
Example: the path to eclipsing binaries



Complexities

- Five DPAC systems involved in this example
  - operations spread over 5 different data processing centres
  - scientific teams spread over numerous institutes
- Each processing step takes months of development and operations
  - developments needed to adjust to realities of data
  - developments to cope with increasing precision and sensitivity to systematics
  - operations time increases with amount of data collected
  - data validation, transfer, and consolidation in between processing steps
  - global validation prior to release
- Many interfaces to coordinate and monitor ⟶ shared data model essential
- Feedback loops imply iterating between processing steps

DPAC challenges



Reality of the data

- Complexity, non-trivial uncertainty properties
- 10^{12}–10^{13} observations
  - any statistical outlier can and will occur
- Unanticipated features
  - Excess stray light
  - Basic angle variations
  - Attitude disturbances due to micro-clanks
  - Continued contamination and throughput loss evolution

Consequences

- More post-launch development time needed across all DPAC systems
- Increase in complexity of results validation

DPAC challenges

- Planning, scheduling
  - Staggered process of bringing the various DPAC systems fully online
  - Downstream systems in development while upstream systems started operations
  - Upstream systems in operation delivering data to downstream systems in development
- Software and IT infrastructure
  - development and testing timescales often underestimated
  - mix of astronomers and software engineers essential, but can also lead to communications problems
- Lots of time spent on understanding and preparing data
  - ordering — are all necessary auxiliary data available? — adapt to data model changes
- Data volume and data transfers
  - volume estimates for planning
  - transfers time scales between DPCs significant
- Data accounting
- In parallel to operations:
  - bugs, performance issues
  - reprocessing of data to fix errors
  - testing, rehearsals, deployment new/updated software

Coordination and communication

- Regular telecons between PO, CUx, and DPCy
  - keep scientists and data processing staff on same page
- Weekly PO-DPCs telecon
  - interfaces, schedule, raise and solve issues
- Monthly payload experts telecon
  - detailed Gaia payload health monitoring
- Participate in MOC-SOC planning meetings
- 1–2 times a year CU plenary meeting (with DPC representatives)
- Consortium meeting every $\sim$ 18 months (since 2015)
  - bring everyone up to speed, team building
- Monthly DPACE telecons, and bi-annual DPACE meeting
  - overall coordination
- Tools: Wiki, JIRA, Subversion

Lessons learned

- Big data ⟶ 'this should not happen' does not apply!
  - anticipate software development to deal with unexpected data features
- Project Office is essential
  - Need staff dedicated to coordinating consortium
  - DPAC scientist have many other tasks distracting from management/coordination
- Cyclic (iterative) operations imply post-launch development in parallel to operations
  - downstream systems need to adjust to changes in upstream data products
  - routine operations for some systems only achieved after several cycles
  - increasing precision in the data requires development effort to deal with more challenging calibration
  - substantial post-launch development effort must be included in funding profiles
- Deliver real data early on to all systems for testing
  - discover early on that real data is not perfect
  - early adaptation to data model changes
- Advanced data products essential to validation of core processing systems
- Be ready (resources!) to re-process stretches of data or re-start operational runs
- Shared data model essential to managing interfacing
  - your data model *will* change

Lessons learned

- Big data $\longrightarrow$ 'this should not happen' does not apply!
  - anticipate software development to deal with unexpected data features
- Project Office is essential
  - Need staff dedicated to coordinating consortium
  - DPAC scientist have many other tasks distracting from management/coordination
- Cyclic (iterative) operations imply post-launch development in parallel to operations
  - downstream systems need to adjust to changes in upstream data products
  - routine operations for some systems only achieved after several cycles
  - increasing precision in the data requires development effort to deal with more challenging calibration
  - substantial post-launch development effort must be included in funding profiles
- Deliver real data early on to all systems for testing
  - discover early on that real data is not perfect
  - early adaptation to data model changes
- Advanced data products essential to validation of core processing systems
- Be ready (resources!) to re-process stretches of data or re-start operational runs
- Shared data model essential to managing interfacing
  - your data model *will* change

Lessons learned

- Big data $\longrightarrow$ 'this should not happen' does not apply!
  - anticipate software development to deal with unexpected data features
- Project Office is essential
  - Need staff dedicated to coordinating consortium
  - DPAC scientist have many other tasks distracting from management/coordination
- Cyclic (iterative) operations imply post-launch development in parallel to operations
  - downstream systems need to adjust to changes in upstream data products
  - routine operations for some systems only achieved after several cycles
  - increasing precision in the data requires development effort to deal with more challenging calibration
  - substantial post-launch development effort must be included in funding profiles
- Deliver real data early on to all systems for testing
  - discover early on that real data is not perfect
  - early adaptation to data model changes
- Advanced data products essential to validation of core processing systems
- Be ready (resources!) to re-process stretches of data or re-start operational runs
- Shared data model essential to managing interfacing
  - your data model *will* change

Lessons learned

- Big data $\longrightarrow$ 'this should not happen' does not apply!
  - anticipate software development to deal with unexpected data features
- Project Office is essential
  - Need staff dedicated to coordinating consortium
  - DPAC scientist have many other tasks distracting from management/coordination
- Cyclic (iterative) operations imply post-launch development in parallel to operations
  - downstream systems need to adjust to changes in upstream data products
  - routine operations for some systems only achieved after several cycles
  - increasing precision in the data requires development effort to deal with more challenging calibration
  - substantial post-launch development effort must be included in funding profiles
- Deliver real data early on to all systems for testing
  - discover early on that real data is not perfect
  - early adaptation to data model changes
- Advanced data products essential to validation of core processing systems
- Be ready (resources!) to re-process stretches of data or re-start operational runs
- Shared data model essential to managing interfacing
  - your data model *will* change

Lessons learned

- Big data $\longrightarrow$ 'this should not happen' does not apply!
  - anticipate software development to deal with unexpected data features
- Project Office is essential
  - Need staff dedicated to coordinating consortium
  - DPAC scientist have many other tasks distracting from management/coordination
- Cyclic (iterative) operations imply post-launch development in parallel to operations
  - downstream systems need to adjust to changes in upstream data products
  - routine operations for some systems only achieved after several cycles
  - increasing precision in the data requires development effort to deal with more challenging calibration
  - substantial post-launch development effort must be included in funding profiles
- Deliver real data early on to all systems for testing
  - discover early on that real data is not perfect
  - early adaptation to data model changes
- Advanced data products essential to validation of core processing systems
- Be ready (resources!) to re-process stretches of data or re-start operational runs
- Shared data model essential to managing interfacing
  - your data model *will* change

Lessons learned

- Big data $\longrightarrow$ ‘this should not happen’ does not apply!
  - anticipate software development to deal with unexpected data features
- Project Office is essential
  - Need staff dedicated to coordinating consortium
  - DPAC scientist have many other tasks distracting from management/coordination
- Cyclic (iterative) operations imply post-launch development in parallel to operations
  - downstream systems need to adjust to changes in upstream data products
  - routine operations for some systems only achieved after several cycles
  - increasing precision in the data requires development effort to deal with more challenging calibration
  - substantial post-launch development effort must be included in funding profiles
- Deliver real data early on to all systems for testing
  - discover early on that real data is not perfect
  - early adaptation to data model changes
- Advanced data products essential to validation of core processing systems
- Be ready (resources!) to re-process stretches of data or re-start operational runs
- Shared data model essential to managing interfacing
  - your data model *will* change

Lessons learned

- Big data ⟶ 'this should not happen' does not apply!
  - anticipate software development to deal with unexpected data features
- Project Office is essential
  - Need staff dedicated to coordinating consortium
  - DPAC scientist have many other tasks distracting from management/coordination
- Cyclic (iterative) operations imply post-launch development in parallel to operations
  - downstream systems need to adjust to changes in upstream data products
  - routine operations for some systems only achieved after several cycles
  - increasing precision in the data requires development effort to deal with more challenging calibration
  - substantial post-launch development effort must be included in funding profiles
- Deliver real data early on to all systems for testing
  - discover early on that real data is not perfect
  - early adaptation to data model changes
- Advanced data products essential to validation of core processing systems
- Be ready (resources!) to re-process stretches of data or re-start operational runs
- Shared data model essential to managing interfacing
  - your data model *will* change


Leiden 2015


IEEC
Sitges 2017
Gaia
DPAC