History of changes

Page/Section	**Nature of change and reason (if applicable)**
Part A	
Risk table	Added risk related to Covid-19.
Deliverables table	Added more deliverables as requested. The following deliverables were added: D1.4, D1.6, D3.1, D3.3, D3.4, D3.5, D3.6, D3.7, D3.8, D3.9, D3.10, D3.11, D3.12, D3.13, D4.1, D4.2, D4.4, D5.1, D5.3, D5.5, D5.6, D5.7, D5.8, D5.9, D6.1, D6.2, D6.3, D6.4
Deliverables	Some titles of the original deliverables were slightly changed in the process of adding more deliverables,
Milestones	Timing for milestones M4 and M9 updated.
Part B	
Section 3.1	Tables 3.1a, 3.1b, 3.1c removed as requested.
Section 3.2	Tables 3.2a, 3.2b removed as requested.
Section 3.4	Table 3.4a removed as requested.
Section 2.2.1, page 24/25	Clarified in what sense the use of the `arXiv.org` repository satisfies the open access requirements.
Section 3.4, page 32	Clarified (second item on page) how many persons will be hired at each of ULEI, MPIA, INAF, UCAM.
Section 3.4, page 32/33	Clarified that for each of ULEI, MPIA, and UCAM, mandatory audit costs are included in 'other goods and services'. The budget justification table has been edited accordingly.
Gantt chart (figure 12)	The chart was updated with the correct project name and the updated milestone timings, and the deliverables were added.
Section 3.3	Clarified the role of external experts.
Section 4.1	Clarified the role of Price-Whelan in the NYU description.

Table of Contents

1 Excellence

1.1 Objectives

The objective of this proposal is to build a detailed selection function for the astrophysical data collected during the survey of the sky carried out by the European Space Agency's Gaia Mission. The 'selection function' is closely related to the concepts of data 'censoring' and 'truncation' (see chapter 8 in [20]), and is needed to account for missing data in statistical analyses. The Gaia survey selection function will be provided to the scientific community in the form of numerical tables and open source computer applications. This will enable scientists to straightforwardly apply the selection function to their analyses of the Gaia data. The outcome will be a dramatic boosting of the scientific exploitation of the Gaia mission data; the unlocking of science applications based on Gaia data that are not possible now; and an increase in scientific publications from this flagship European space mission. In detail the objectives are:

1. Develop a detailed mathematical formulation of a survey selection function, where the goal is to keep this as general as possible even if in this project the focus is on the Gaia survey. This will ensure that the expertise built up in this project is transferable to other astronomical surveys. The mathematical formulation should account for the layering of user imposed sample selections on top of the survey selection function and allow for a description of the selection function of combinations of surveys.

2. Research and develop a detailed description and modelling of the Gaia survey selection function and do the same for the combination of Gaia and other surveys. The results will set the requirements for the next two objectives.

3. Provide a detailed practical implementation of the Gaia selection function in the form of auxiliary data, which will be accessible through the ESA Science Data Centre, and open source tools, which will be made available through code hosting platforms.

4. Develop tools to incorporate the selection function in scientific analyses. These tools should allow for combining the Gaia survey selection function with user imposed sample restrictions and for combining Gaia with other surveys. The tools will be made available as open source code through code hosting websites.

5. Apply the selection function tools to example science cases. This will serve to demonstrate the benefits of carefully accounting for the selection function and at the same time provide worked examples to the community of prospective users of these tools. The application of the tools also serves as a means of rigorously testing them.

The long term curation of the results from this project will be ensured by the availability of the auxiliary data through the ESA Gaia archive and by the source code of the user-facing tools being embedded in or affiliated to the Astropy project.

1.1.1 Boosting the science exploitation of Gaia data

The European Space Agency's Gaia Mission [19] is the most successful of ESA's space astronomy missions as measured by the publication rate of papers using its data. Over 2500 papers appeared between April 2018 and February 2020 that use data from the second Gaia data release[1]. Among the many high profile science results published by European groups are the discovery of a spiral in the phase space structure of the Milky Way's disk ([2], indicative of a recent disturbance, most likely due to the Sagittarius dwarf galaxy), the first direct observational evidence of crystallization in the interiors of white dwarfs ([37], confirming a fifty year old prediction), and the uncovering of a major event 10 billion years into the Milky Way's past when a collision with a dwarf galaxy took place which contributed to the build-up of the Milky Way's stellar halo and thick disk [4, 5, 22].

But while there are numerous exciting results that are starting to transform many different fields of astrophysics, most of these have focused on 'low hanging fruit' in the Gaia data archive, and have barely scratched the surface of the information contained in the Gaia data. Specifically, almost all analyses to date have exploited the astounding Gaia data precision and accuracy on individual objects, but have side-stepped performing or exploiting **population studies**, i.e. the *statistical distribution of objects with certain properties*. For example, a result as elementary as a good determination of the radial profile of the Milky Way's disk stars — by simply counting them as a function of position — has yet to be carried out. This is because such studies inevitably require knowledge of the **selection function**, i.e. the probability that an object with certain observed properties at a given position on the sky would be catalogued by Gaia. This knowledge is indispensable for any population study and — with its billions-of-objects catalogue — population studies are the very core of the Gaia mission.

The Gaia selection function results from a highly non-trivial combination of the on-board source selection, telemetry losses, selections on data quality during processing, the scanning pattern of the mission, and selections on quality before the publication of the processed data in the Gaia archive. In practice the community has dealt with this complexity by resorting to drastic cuts on object properties and Gaia data quality flags to arrive at samples whose selection function is very simple (avoiding, for example, the well known problems associated with the estimation of distance as the inverse of the parallax [29]). Such analyses grossly under-exploit the information content of the Gaia data set, often by orders of magnitude.

Consequently, many high-profile science cases for which Gaia was built cannot be realized as long as the selection function of the Gaia survey remains unknown. It is the single goal of this proposal to remedy this shortcoming and dramatically boost the astrophysical utility and legacy value of the Gaia mission (which will remain the standard in fundamental astrophysical data for decades to come), by comprehensively characterizing the catalogue's selection function.

1.1.2 Context

ESA's Gaia mission [19] represents a European breakthrough in astrophysics, a cornerstone mission launched in 2013 which was aimed at producing the most accurate 3D map of the Milky Way to date. The resulting stereoscopic census of our Galaxy represents a giant leap in astrometric accuracy (reaching the 10–20 micro-arcsecond regime, which is equivalent to knowing the 3D positions of stars to 10 per cent precision over distances as large as 20 000 to 30 000 light-years) complemented by the only full sky homogeneous photometric survey with an image angular resolution comparable to that of the Hubble Space Telescope, as well as the largest spectroscopic survey ever undertaken.

The primary scientific aim of the mission is to map the structure of our Galaxy and unravel its formation history and subsequent evolution. Current cosmological models envisage the formation of large galaxies through the merging of smaller structures. Deciphering the assembly history of our Galaxy requires a detailed mapping of the structure, dynamics, chemical composition, and age distribution of its stellar populations. Ideally one would like to 'tag' individual stars to each of the progenitor building

[1] As indicated by the number of citations to the Gaia DR2 release paper

blocks of the Galaxy [17]. The Gaia mission is designed to provide the required fundamental data in the form of distances (through parallaxes), space velocities (through proper motions and radial velocities) and astrophysical characterisation (through multi-colour photometry) for massive numbers of stars throughout most of the Galaxy. It should be stressed however that Gaia is not simply a 'Milky Way mission' but is truly a multi-faceted *astrophysics mission* which will provide exciting scientific results covering many topics including: fundamental stellar physics across the Hertzsprung-Russell diagram, the characterisation of tens of millions of binary stars, unique samples of variable stars of nearly all types (including key cosmological distance calibrators), detection and orbital classification of thousands of exo-planetary systems, a comprehensive survey of objects ranging from huge numbers of minor bodies in our Solar System, through galaxies in the nearby Universe, to over a million distant quasars. Gaia will also provide a number of stringent tests of general relativity. Last but not least, a massive survey such as Gaia will uncover many surprises that the Universe still holds in store for us.

The data processing for the Gaia mission is carried out by a European consortium of institutes and funding agencies, the Data Processing and Analysis Consortium (DPAC). The task of the DPAC is to turn the raw telemetry from Gaia into data products ready for use by astronomers and scientists around the world. The data products are both the basic astrometric, photometric, and spectroscopic results as well as advanced data products, such as the astrophysical characterization of stars, derived from the former. Through an independent unit within the consortium the DPAC also ensures the quality of the published data through extensive technical and scientific validation. The publicly released data, which are disseminated through the ESA Science Data Center and its partners, are accompanied by extensive documentation which is also produced by DPAC. It should be stressed here that the DPAC holds no proprietary rights to the data which means that once the data products are validated and documented they are released publicly and world-wide without delay. In this sense Gaia represents a uniquely open astronomical survey concept.

From the early days of the Gaia mission it was foreseen to have incremental data releases based on increasing amounts of collected measurements. The various data products would be released in stages (where, for example, some of the advanced data products such as exoplanet catalogues can only be released after sufficient data has been collected and the various astrometric calibrations have reached the required levels of precision) and with increasing accuracy for each release. To date two data releases have taken place, Gaia DR1 in September 2016 and Gaia DR2 in April 2018. Both releases have been highly successful and have had major impacts on the fields of Galactic archaeology; Milky Way mass estimation; stellar streams, dwarf galaxies, globular clusters; open clusters; white dwarfs; the characterization of exoplanet host stars and determining the absolute sizes of protoplanetary disks, to name some prominent examples. The studies of individual stars and stellar systems benefit immensely from the now easily available high precision parallax data.

The next releases foreseen are Gaia EDR3 in the third quarter of 2020 and Gaia DR3 in the second half of 2021. The following release (Gaia DR4) will be based on all data collected during the nominal mission lifetime of Gaia and feature the full suite of data products, including the epoch data (individual measurements) for the astrometry, photometry, and spectroscopy[2]. These future Gaia data releases are expected to have an even larger impact on astronomy due to the much richer set of data products allowing for the combination of high accuracy astrometry and radial velocities with detailed astrophysical characterization of the sources in the Gaia catalogue (stellar parameters including chemical compositions and variability, parameters of multiple stars and exoplanets, medium resolution spectra around the Ca triplet region for millions of stars to magnitude 13, low resolution prism spectrophotometry for all sources in the Gaia catalogue).

Despite the major successes of Gaia DR1 and Gaia DR2 there is one serious weakness of the releases: the DPAC does not provide a detailed selection function for the Gaia survey. This aspect was never included in the funded DPAC efforts and is not planned to be added as DPAC work in the future. This implies that major Gaia science cases cannot be addressed by the astronomical community to the full

[2] See the Gaia data release scenario pages.

potential of the quality of the data. In section 1.1.3 we motivate why a detailed selection function is needed and how this will enhance the scientific exploitation of the data from this flagship European space mission.

Gaia instruments and survey strategy For the understanding of the work proposed here it is useful to briefly summarize the characteristics of the Gaia mission and how the measurements are collected and subsequently processed on ground. Much more detail can be found in the paper describing the Gaia mission [19]. The Gaia spacecraft and its instruments were designed to collect data that will allow the determination of highly accurate positions, parallaxes, and proper motions for > 1 billion sources brighter than magnitude 20.7 in its white-light photometric band G (covering the wavelength range 330–1050 nm). The astrometry is complemented by multi-colour photometry, measured for all sources observed by Gaia, and radial velocities which are collected for stars brighter than $G \approx 17$. Gaia carries two telescopes of which the collected light is combined onto a single focal plane servicing the three main instruments. The astrometric instrument collects source images in the G-band, where the fundamental inputs to the astrometric data processing consist of the precise times when the image centroids pass a fiducial line in the focal plane of the instrument. The photometric instrument is realized through two prisms dispersing the light entering the field of view of two dedicated sets of CCDs. The Blue Photometer (BP) operates over the wavelength range 330–680 nm, while the Red Photometer (RP) covers 640–1050 nm. The data collected by the photometric instrument consists of low resolution spectrophotometric measurements of the source spectral energy distributions. The spectroscopic instrument, also called the radial-velocity spectrometer (RVS), collects medium resolution ($R \sim 11\,700$) spectra over the wavelength range 845–872 nm, centred on the Calcium triplet region. The spectra are collected for all sources to $G \approx 17$ (16^{th} magnitude in the RVS filter band).

The Gaia sky survey strategy relies on the spacecraft slowly spinning around the axis perpendicular to the lines of sight of the two telescopes (which are separated by 106.5°). Every six hours the Gaia telescopes scan a great circle on the sky with an across scan field of view size of 0.7°. By slowly precessing the spin axis around the direction to the Sun the full sky can be surveyed multiple times with optimal sky coverage uniformity over the five years nominal mission lifetime.

The processing of the observations collected by Gaia is carried out by the DPAC through a series of complex pipelines which are part of a large iterative system (see section 7 in [19]).

Although the Gaia survey selection function is seemingly very simple, with only a magnitude limit imposed on the observations (see figure 9), each of the above components of the Gaia mission adds complexity to the survey selection function as illustrated in figure 10. A few examples are listed here:

- The data collection on the spacecraft is limited by the amount of sources that can be handled at any on time, meaning that for dense regions on the sky (above 10^6, $750\,000$, and $35\,000$ sources deg^{-2} for the astrometric, photometric, and spectroscopic instruments respectively) not all sources can be measured, leading to brighter effective survey limits in those regions. In addition the on-board storage capacity limitations combined with available ground station time lead to deletion of data on occasions when both Gaia telescopes are scanning along the Galactic plane for prolonged periods of time.

- Despite the optimal sky survey strategy realized through the revolving scanning technique, the number of times any given source is observed does vary with celestial position, in particular with the ecliptic latitude.

- Decisions are made along each of the various on-ground data processing steps on which data and/or sources to process, where periods of ‘bad inputs’ are excluded. This leads to gaps in the time coverage of measurements for a given source or to effectively brighter survey limits covering certain time periods (which translates to certain areas on the sky). Additionally, in early data releases the processing for certain data products may be limited to brighter sources or sources with specific astrophysical characteristics.

Each of the above examples complicates the Gaia survey selection function which in the end is very non-trivial to describe and handle. Nevertheless we must make an effort to provide the best possible description of the selection function as well as tools to incorporate it into scientific analyses as we motivate next.

1.1.3 Scientific motivation

1.1.3.1 The selection function: definition and indispensability

For any 'statistical' study using a catalogue — a study where it matters quantitatively how common or rare a certain type of object or circumstance is — the selection function of a catalogue is arguably as important as the catalogue itself. The catalogue *selection function*, $S_C(y)$, is a function of the observable (and thereby underlying physical) properties, y, of any astronomical object; and it specifies the probability that an object of properties y will enter the (Gaia) catalogue. A generic and central way of learning from any catalogue is to envision a set of physical or mathematical models $\lambda_0(z \mid \theta)$ — that predict the distribution of true quantities z, given a set of model parameters θ — and then ask what the 'data' (the catalogue entries) $\lambda_C(y \mid \theta)$ tell us about the relative plausibility of the model parameters. The relation between these quantities is generically

$$\lambda_C(y \mid \theta) = S_C(y) \int p(y \mid z) \lambda_0(z \mid \theta) \, \mathrm{d}z = S_C(y) \lambda_0(y \mid \theta) \, . \tag{1}$$

This immediately shows that no inference, is possible without knowledge of $\mathbf{S_C(y)}$. Most commonly the quantities y_s on which the selection function really depends are only a modest subset of all the catalogue observables, or all the quantities that are being modelled, y, such that the selection function separates, $S_C(y_s) \times S_0(y_\mathrm{other})$, with $S_0(y_\mathrm{other}) \approx 1$. For the Gaia catalogue selection function, the quantities y_s are (in presumably good approximation): the sky position (α, δ) of an object, its magnitude in the visible G band, and its $G_\mathrm{BP} - G_\mathrm{RP}$ colour from the spectrophotometric bands.[3] These four quantities are a small subset of the Gaia catalogue entries. So, why is it hard to determine a good approximation of the Gaia catalogue selection function, $S_C(y_s)$? This is because $S_C(y_s)$ is such a complex function of the y_s: the sky position enters through the source density ('crowding') and the scanning pattern, which sets the number of observation epochs and the distribution of scan angles; the scan angle in turn determines the (sky) direction in which the BP, RP and RVS spectra are dispersed, which in turn determines the probability that they are rendered unusable due to source-to-source contamination. The magnitudes determine not only the obvious aspect of the photon noise, but — in conjunction with the source density — also the decision of parsing the data into postage stamps for processing. This will be detailed in the methodology section of this proposal.

Given that $S_C(y_s)$ has not (yet) been determined for Gaia, why is it that many widely accepted results have already come from analysing the distribution of Gaia catalogue entries for large samples? This is for a number of reasons. First, some studies have essentially ignored (at their own peril) the selection function, thereby implicitly assuming $S_C(y_s) \approx 1$ for the regime of interest. The second and most frequent regime so far is that one can make additional 'sample cuts' by defining

$$S_\mathrm{cut}(y) = \begin{cases} 1 & \text{if } y \text{ within cuts} \\ 0 & \text{otherwise} \, . \end{cases}$$

If one then models only the restricted data $\lambda_\mathrm{cut}(y \mid \theta)$, Eq. (1) becomes simple, as one can then sensibly adopt

$$\lambda_\mathrm{cut}(y \mid \theta) \approx S_C(y) \; S_\mathrm{cut}(y) \, \lambda_0(y \mid \theta) \approx \begin{cases} \lambda_0(y \mid \theta) & \text{if } y \text{ within cuts, where } S_C(y_s) \approx 1 \\ 0 & \text{otherwise} \end{cases} \, .$$

[3] For objects with very high proper motions, e.g. in the Solar system, the selection function may also depend on $\vec{\mu}$; this interesting regime, affecting a very small fraction of the Gaia detections, is not the focus of this proposal.

A pertinent example would be a 'stellar halo study' $S_{\mathrm{cut}}(y)$ that limits their analysis to, say, blue-ish ($G_{\mathrm{BP}} - G_{\mathrm{RP}} < 1$) stars of $13.5 < G < 17.5$ at high Galactic latitudes, away from Globular clusters and the Magellanic Clouds. Then the assumption that the small sub-sample is nearly complete within the stated $S_{\mathrm{cut}}(y)$ is defensible.

The third regime is that one can analyze the distribution of quantities in $p(y_{\mathrm{other}})$, or $p(y_{\mathrm{other}} \mid y_s)$: the most obvious case is to analyze the 3D kinematics of stars in the Solar neighbourhood. The fourth regime is that the selection function $S_C(y_s)$ has been 'properly' determined and included in the modelling; however this has only happened to date for very small sub-regimes of the space spanned by y_s in the Gaia catalogue.

It is the comprehensive construction of this (approximate) selection function, $S_C(y_s)$, that is at the heart of this proposal. As we now sketch with a few examples, without knowing the selection function robustly and precisely it is impossible to extrapolate from *statements about populations in a catalogue* into *insights in fundamental physics and astrophysics.*

1.1.3.2 Highlight Gaia science that cannot be done without a good selection function

We now proceed to highlight a few science cases central to Gaia's scientific success that absolutely require a well-determined selection function to be feasible[4]. To illustrate the kinds of selection effects that can occur in the Gaia data we refer to figures 1–7. These figures serve also to reinforce the points made in the following text.

A) 3D-Mapping of the Milky Way

The overall 3D structure of Milky Way's stellar body: Creating an unprecedented, large-scale 3D map of our Galaxy is arguably the central seminal goal of the Gaia mission, at the heart of setting the fundamental parallax precision science requirements. Yet, to date, no large-scale map of the stellar density (most of which is in the Galactic disk) has emerged. This is because such a density map is not merely the combination of the (catalogued) stars distribution $\{y_s\}$ in position, magnitude and colour with parallactic distances. Linking $\{y_s\}$ to the spatial density of stars $\rho_*(X, Y, Z)$ requires modelling of the 3D dust distribution (there are large, coherent efforts underway to do this) and of $S_C(y_s)$. The selection function is so central here because the quantities to be modelled are exactly the quantities y_s on which the selection depends!

Unbiased distance estimates for all stars in samples of Gaia data are also of vital interest to the community for studies of the structure and dynamics of the Milky Way. Estimating the distances to stars in a Bayesian framework requires you to know both the Milky Way 3D density map and the selection function of the given sample. In its simplest form, our posterior on the distance to any star in the sample will be given by

$$\mathbb{P}(r|\varpi) = \frac{1}{Z}\mathbb{P}(\varpi|r, \sigma_\varpi)\mathbb{P}(r|l, b)\mathbb{P}(S|r, l, b)\,, \tag{2}$$

where $\mathbb{P}(r|l, b)$ refers to the 3D Milky Way map, expressed as the distance r along the line of sight specified by Galactic longitude and latitude (l, b), $\mathbb{P}(S|r, l, b)$ is the selection function and Z is a normalisation constant. The parallax and its uncertainty are ϖ and σ_ϖ, respectively. A detailed description of estimating unbiased distances from parallaxes is given in Luri et al. [29]

Mapping the 'Young' Milky Way with Cepheids: Gaia is also a time-domain survey, and variability-based identification of diagnostically precious classes of objects, such as classical Cepheids, are an integral part of its science deliverables [24]. In the Milky Way, Cepheids are the arguably ideal tracers to map the dynamics of the 'young' disk (stars with ages $\tau_{\mathrm{age}} < t_{\mathrm{dyn}} \approx 150$ Myrs), and to map the pattern of how the large-scale star-formation pattern proceeded across the Galactic disk in the last

[4] By 'feasible' in this context we mean: able to address a science question with a quantitative understanding of the dominant source of systematic uncertainty; and not falling short by more than an order-of-magnitude of the full pertinent information content of the Gaia catalogue.

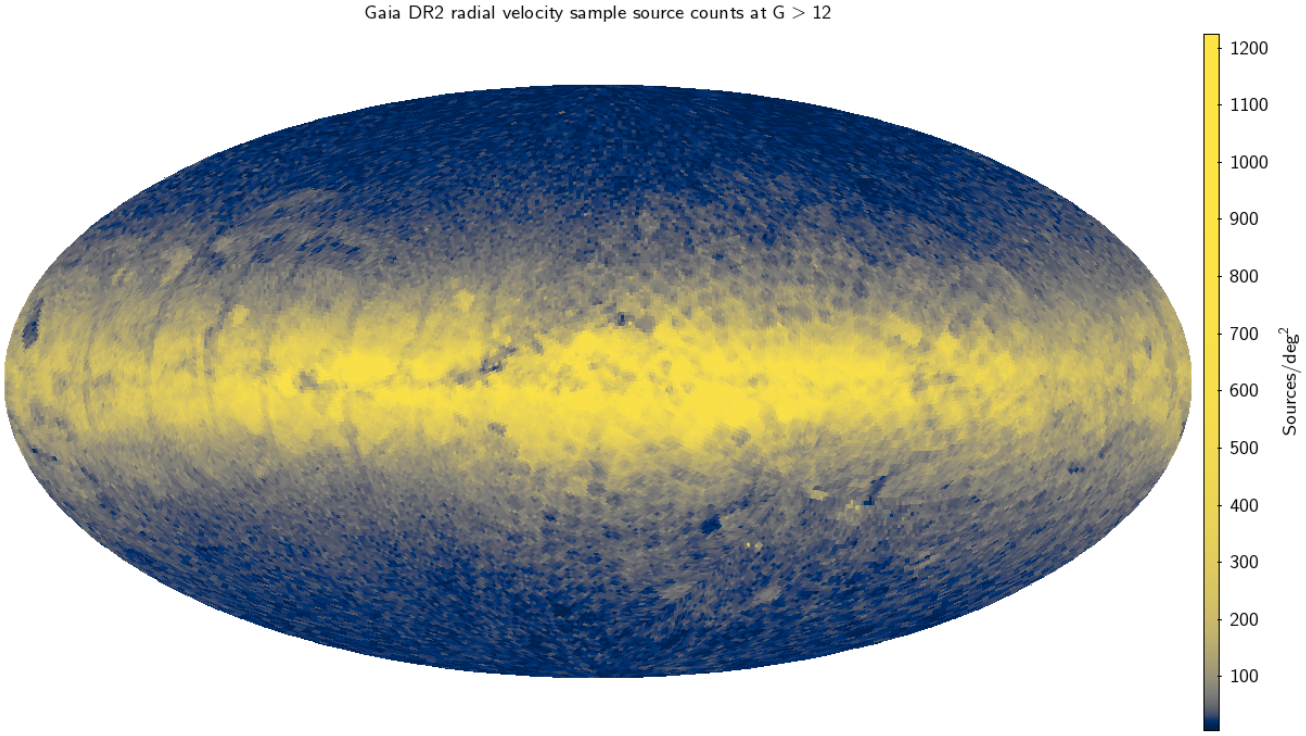

Figure 1: Source density distribution on the sky (number per square degree) of stars in the Gaia DR2 catalogue for which a radial velocity is listed and which are fainter than $G = 12$. The sky projection is in Galactic coordinates with the Milky Way bulge in the centre of the image and the longitude increasing to the left. Note the prominent patterns of squares in the southern hemisphere on the lower right hand side of the figure. These are imprints of the photographic plates on which the Guide Star Catalogue is based, which was one of the surveys used to construct the Initial Gaia Source List [36]. This list was used to bootstrap the creation of the Gaia source list (see section 1.3.2). Imprints from the spectroscopic component of the Sloan Digital Sky Survey can be seen as dark stripes crossing the Galactic plane on the left hand side. Although these imprints will gradually disappear in future Gaia data releases, this figure is a good illustration of the complications that can arise for selection functions of combined surveys.

$t_{\rm dyn}$ [35]. This is because Cepheids combine three qualities: they are luminous (and can be seen to great distances, even in the presence of quite severe dust extinction), their distance can be ascertained quite precisely through the period-luminosity relation (even in light of marginal parallax information), and their period, also precisely predicts their mass and hence age.

But of course, any such analysis revolves around ’what is the number density or Cepheids at distance X in direction Y at period/luminosity Z’; in short this requires both the selection function for Cepheids within Gaia; its modification when external surveys [such as ASAS-SN 27] are included; and, as most classical Galactic Cepheids are at very low latitude, it also requires combining the catalogue selection function $S_C(y_s)$ with a 3D dust map [eg. 21].

B) Milky Way Dynamics and Distribution of (Dark) Matter

Another seminal and central science case for the overall Gaia mission is the inference of the Galactic (dark) matter distribution, or equivalently the gravitational potential, from modelling the 6D phase-space distribution of stars. There are many sophisticated approaches to ‘dynamical modelling’, but the indispensable role of a well-understood selection function of the stars that serve as kinematic tracers can be illustrated with the arguably simplest form of dynamical modelling, the Jeans Equation. Consider

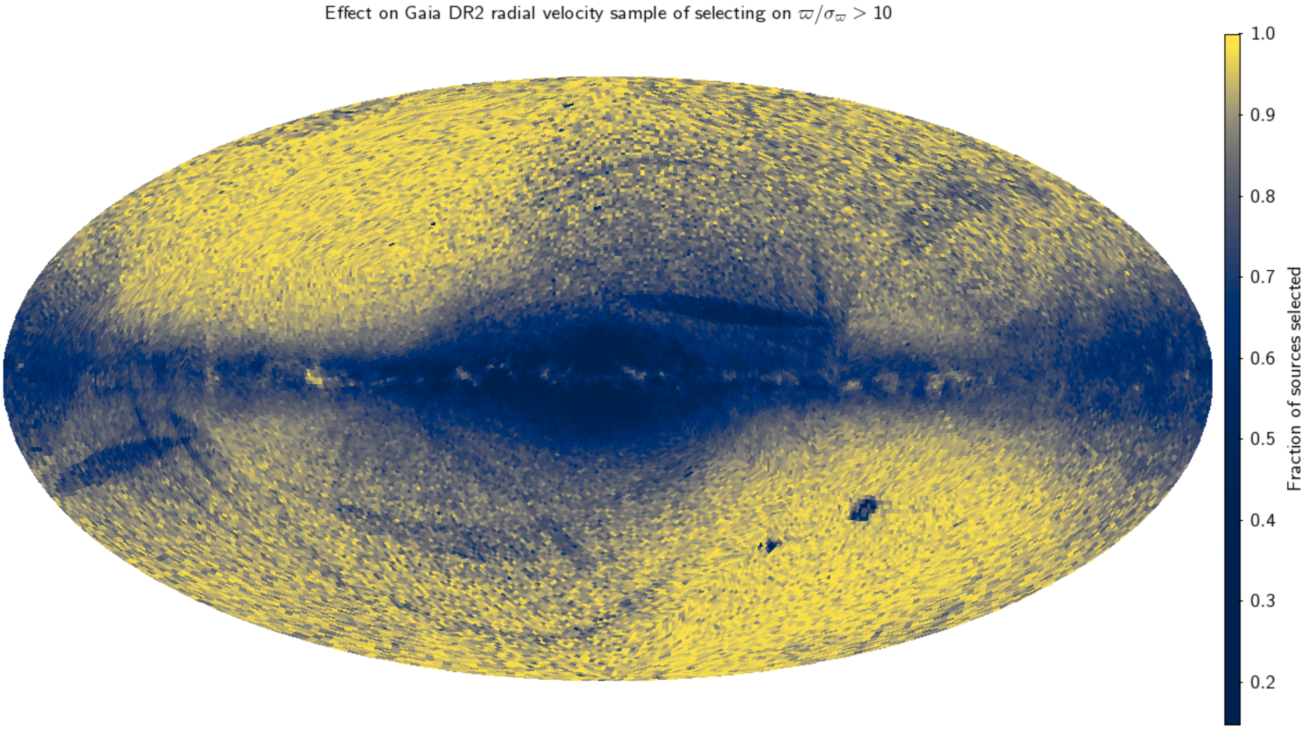
Effect on Gaia DR2 radial velocity sample of selecting on $\varpi/\sigma_\varpi > 10$
Fraction of sources selected
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2

Figure 2: The fraction of stars selected from the Gaia DR2 sample for which radial velocities are listed, when demanding that the parallax is larger than 10 times its uncertainty. Note how the ecliptic pole regions on the top left and bottom right are strongly favoured, whereas the fraction of selected sources drops dramatically along the Galactic plane. In addition there are oddly shaped regions with a lower fraction of selected sources. These reflect a combination of the scanning strategy followed by Gaia and filtering on data quality at various stages in the data processing.

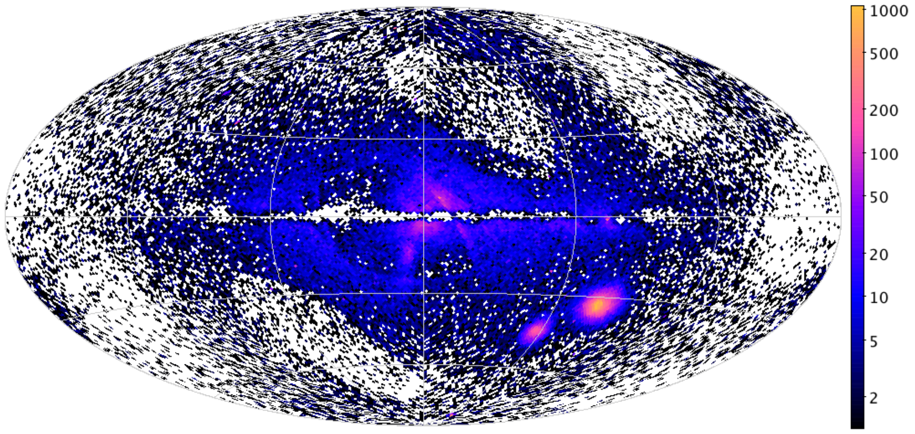

Figure 3: The sky distribution of confirmed RR Lyrae variables in the Gaia DR2 catalogue. Note the very strong selection effects on the sky due to the combination of the scanning strategy, the constraints on the minimum number of observations for classifying variable stars, and data quality filtering in the processing pipelines. For details see [23].

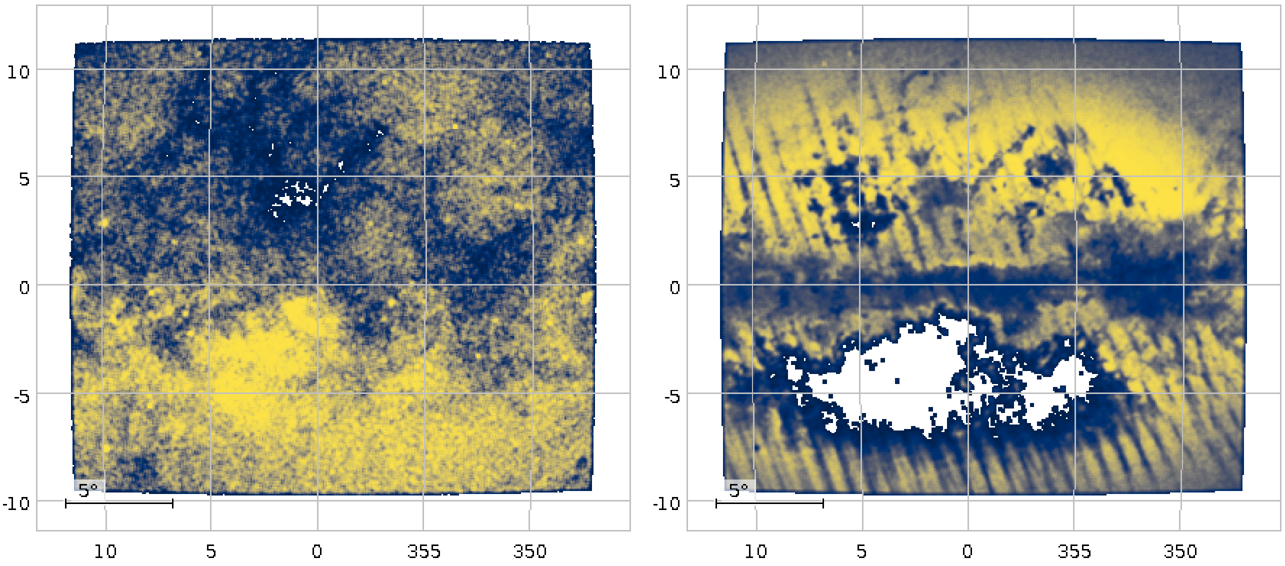

Figure 4: **Left** The distribution on the sky of sources brighter than $G = 13$ (300 thousand sources) in the direction of the Milky Way bulge region. No obvious selection effects are apparent, the darker regions being due to clouds of dust blocking the light from the stars behind them. The white pixels are empty of sources due to severe dust effects. **Right** The same for sources with $20.65 < G < 20.70$ (4 million sources). Note the strong striping pattern which again reflects the combination of the Gaia sky scanning strategy and the filtering on data quality during the data processing. The white areas are due to a combination of dust effects and the drop in catalogue completeness at these faint magnitudes.

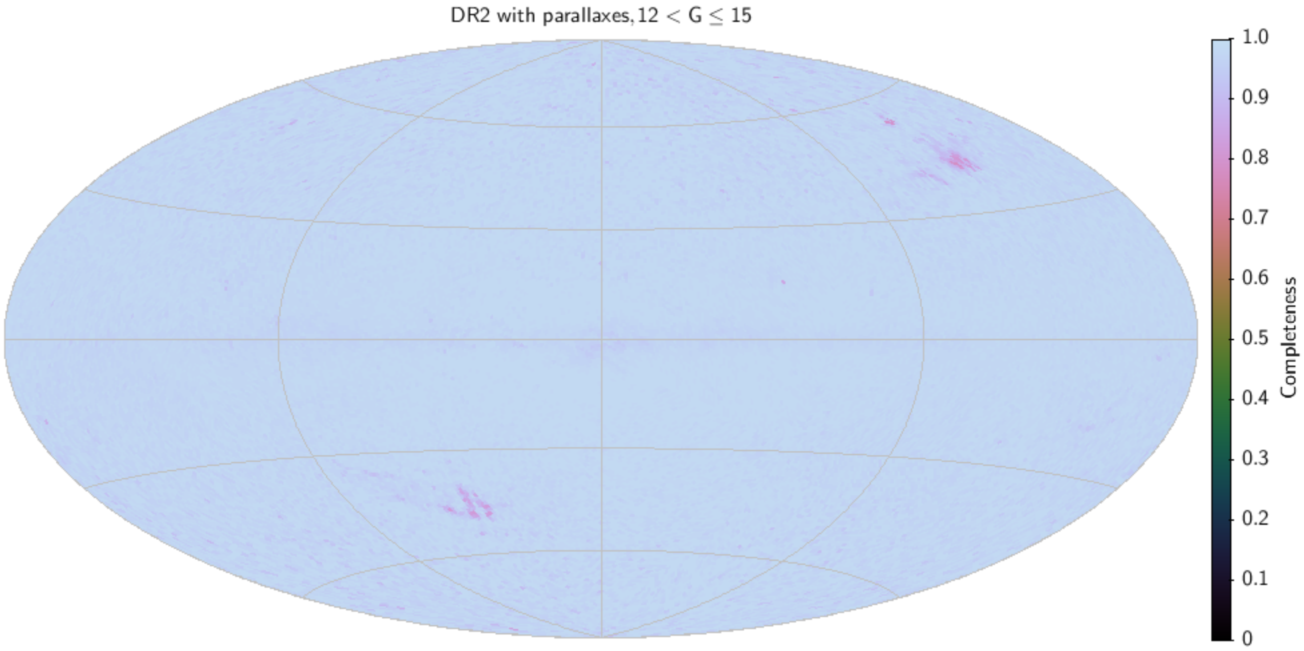

Figure 5: The completeness of the Gaia DR2 sample of sources at $12 < G \leq 15$ for which parallax and proper motion are listed in the catalogue. The completeness is high and generally uniform over this magnitude range. Compare to figure 6 which shows how the completeness is affected when combining this sample with the 2MASS catalogue.

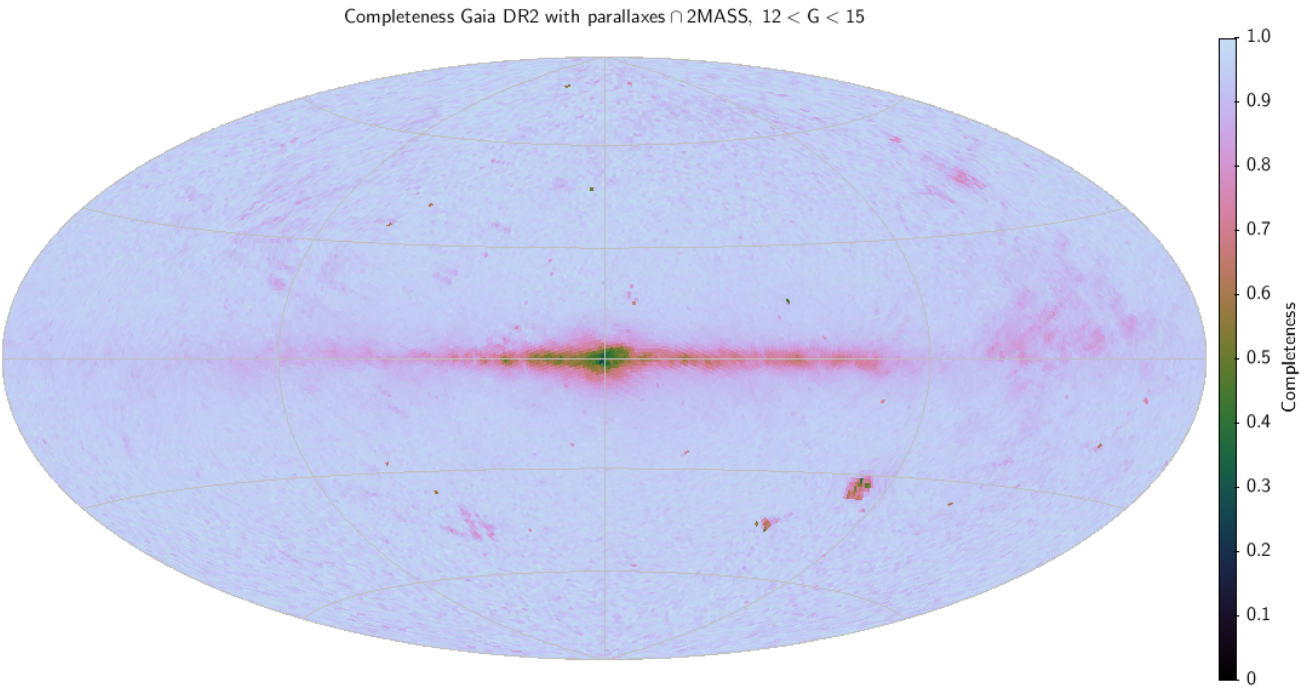

Figure 6: The completeness of the sample of sources at $12 < G \leq 15$ for which parallax and proper motion are listed in Gaia DR2 and high quality infrared photometry is listed in 2MASS. Note how combining these two catalogues leads to a drop in the completeness along the Galactic plane. This is due to crowding effects in the 2MASS catalogue. In addition, in regions away from the Galactic plane artefacts related to the 2MASS sky survey are seen. This figure is an illustration of the selection function complexities that can arise when combing data from multiple catalogues.

the disk dynamics in the vertical direction (the 'Oort limit'):

$$\frac{\partial[\rho_*(R,z)\overline{v_z^2}(R,z)]}{\partial z} + \frac{1}{R}\frac{\partial[R\rho_*\overline{v_z v_R}]}{\partial R} + \rho_*\frac{\Phi(R,z)}{\partial z} = 0 \text{ or } \frac{\partial[\rho_*(R,z)\overline{v_z^2(R,z)}]}{\partial z} + \rho_*\frac{\Phi(R,z)}{\partial z} = 0\,, \qquad (3)$$

where the second form is for separable motions. Dynamics involves not only motions (e.g. velocity dispersion), but also the tracer densities, and the spatial derivatives of the densities and motions. And the prediction of observables (the catalogue entries) from $\rho_*(R,z)$ is of course linearly dependent on the selection function, and on the spatial gradient of the selection function. **Therefore, Gaia data have taught us very little to date about dark matter near the Sun, or whether the dark matter in the Milky Way has a cusp or a core**.

This is also why all initial high-profile analyses of the motions in Gaia data [2, 18] were only *kinematic* not *dynamical* analysis. Traditionally, the 'rotation curve' has been the regime where the spatial tracer distribution matters least. But even the best Gaia rotation curve analyses [e.g. 13] are limited in measuring exactly where dark matter starts to dominate the mass distribution in our Galaxy by the 'asymmetric drift' – the analogous radial terms to Eq. (3) that include spatial tracer derivatives. In the work presented in [13] it is not only Gaia's selection function that needs to be known, but also that of the APOGEE survey; this present proposal also addresses the determination of survey-combined selection functions.

C) Stellar Streams and 'empty' dark matter halos with Gaia

Dynamical analyses, where it is sensible to model $p(\vec{v} \mid \vec{x})$ rather than $p(\vec{v},\vec{x})$, depend less on detailed knowledge of the selection function, as $S_C(y)$ depends most strongly on $\vec{x}$, or on (α,δ,D), with D the distance. A beautiful recent example of this is Erkal et al. [15], which used Gaia proper motions (and

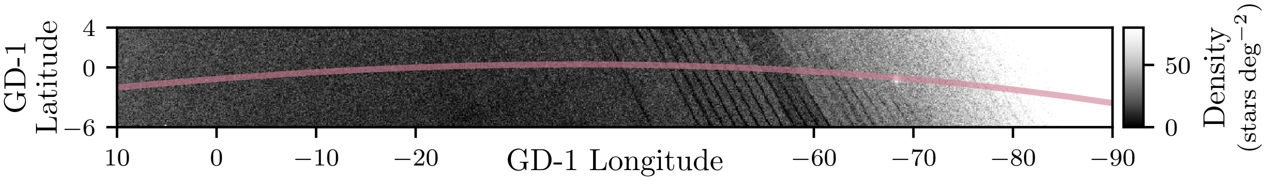

Figure 7: The incompleteness of the Gaia catalogue causes us to see stripes [25] in the stellar number density around the stream GD-1 (the track of this stream is shown in pink), which could be wrongly interpreted as gaps ripped by the passage of dark matter sub-halos.

other information) along the known 'Orphan stream' to constrain the mass of the Large Magellanic Cloud. However, one need not look further than the current holy grail of stellar stream science: the diagnosis of the impact of purported 'empty' (i.e. star-free) dark matter sub-halos in the Milky Way: a close passage by a stream of such dark matter halos would create a gap or caustic in the stream. In the analysis of Bonaca et al. [6] that provides the currently best evidence that such empty halos actually exist — a central and untested prediction of λCDM cosmogony — the detailed knowledge of the selection function is the limiting systematic. A powerful illustration of this effect is shown in Fig. 7, a modified version of a plot from Ibata et al. [25] who argued that the striped selection function of Gaia — arising from the scanning law — appears as gaps in the stellar density along the stream GD-1.

D) Stellar Physics: Binary Stars in Gaia

Among the many aspects of stellar physics where Gaia can be transformative, binary stars are an area of exceptional future importance [e.g. 11]. Already Gaia has offered breakthroughs in identifying wide binaries, as adjacent stars of the same proper motion. Yet, this science direction brings up — and is severely limited by — another aspect of the selection function: the probability of an object getting selected is not only a function of its own properties, y_s, but also of its 'neighbours'. This is illustrated in figure 8, which shows that the minimal distance $\Delta\theta$ at which a second (fainter) star has to be to also have a good proper motion measurement depends on the magnitude difference ΔG. In El-Badry et al. [14] this has been modelled well, but of course this modelling is only a small sub-aspect of the much broader selection function.

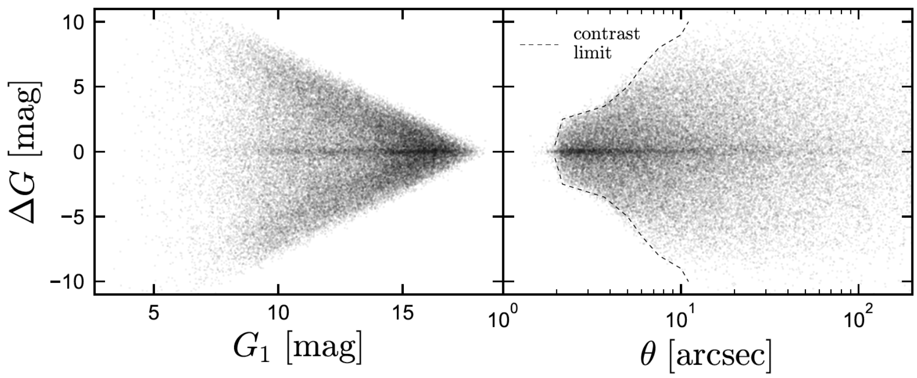

Figure 8: Illustration of the inherent complexities of the Gaia selection function ('contrast limit', right panel), when considering close pairs of sources, from El-Badry et al. [14], who discovered a population of twin-binaries (near-identical stars) to $\gtrsim$ 1000 au separations, of unknown formation mechanism. Unsurprisingly, the minimal angular separation at which stars of a certain magnitude difference ΔG both have good proper motions (so as to be recognized as comoving) depends strongly on ΔG.

Binaries are in this respect only the minimal case of either a star cluster or of crowded stellar fields in general. It is unsurprising that not a single publication has used Gaia DR2 to study the crowded regions of our Galaxy, such as the bulge (see figure 4) and in particular Baade's window, which are some of the most interesting regions of our Galaxy. In crowded fields, the existence of BP and RP 'magnitudes' in the Gaia catalogue, which come from re-integrated slitless spectra, has a very complex dependence on position and magnitude.

Without rigorous knowledge of $S_C(y_s)$ in the 'crowded source' regime, binary science, cluster science and the central Galaxy will remain inaccessible to Gaia studies that do justice to the data's information content.

There are added levels of complexity in determining and characterizing the selection function. Often, it is of interest — or at least tempting — to make use of an extensive set of quality flags as part of a sample selection. If such cuts only eliminate a tiny subset, for reasons unrelated to the object's physical properties, this is without consequence. But if far-reaching cuts are made in quality flags such as RUWE (renormalized unit weight error) or on data uncertainties (velocity errors, fractional parallax errors) a whole new level of complexities arises. While in general it may be advisable to base sample selection on observables, not their uncertainties, this issue cannot be ignored. This is an area where it may not be possible to devise (in the context of this proposal) a user-ready recipe for $S_C(y_s)$.

Taken together, far better knowledge of the Gaia catalogue selection function $S_C(y_s)$ is needed to

- **map our Galaxy in 3D**
- **model the dynamics of our Galaxy, constrain the distribution of dark matter and detect 'empty' dark matter halos**
- **unleash Gaia's potential for stellar astrophysics, even on aspects as elementary as estimating luminosity functions and frequencies of special objects, or learn about binary stars.**
- **start exploring what Gaia can teach us in 'crowded fields', clusters, and the inner galaxy.**

And while better knowledge of the Gaia catalogue selection function $S_C(y_s)$ is scientifically indispensable, it is not part of the DPAC deliverables.

In the subsequent sections we lay out how to provide this. We will focus on the position-magnitude aspects of the selection function, and on the selection functions that arise when combining Gaia data sets with external information, spectroscopic surveys or deeper imaging surveys.

1.2 Relation to the work programme

This proposal is in response to the call 'Space 2018–2020' and is specifically focused on the topic 'Scientific Data Exploitation' (SPACE-30-SCI-2020). How the work proposed here addresses the SPACE-30-SCI-2020 challenges is listed in the following table.

Call specific challenge/scope	How GaiaUnlimited addresses these	Project outcome/benefit	Work package(s)
Support the data exploitation of European missions and instruments, in conjunction, when relevant, with international missions.	GaiaUnlimited specifically aims to boost the scientific exploitation of the ESA Gaia mission data in combination with data from other international astronomical surveys.	Higher number and higher quality of publications based on data from a European space mission.	2, 3, 4, 5, 6

Call specific challenge/scope	**How GaiaUnlimited addresses these**	**Project outcome/benefit**	**Work package(s)**
Projects may rely on the data available through ESA Space Science Archives when possible or other means (e.g. instrumentation teams).	GaiaUnlimited relies on the Gaia data publicly available through the ESA archives and on other astronomical survey data publicly available through international data centres, or available to us through members of the survey teams participating in this project. In addition we have close relations with the Gaia Data Processing and Analysis Consortium experts (some GaiaUnlimited participants being members of DPAC).	Information necessary to the construction of the selection function. Access to survey (instrument) team expertise.	1, 3, 4, 5
Combination and correlation of this data with international scientific mission data, as well as with relevant data produced by ground-based infrastructures all over the world.	The cross-survey selection functions involve combining Gaia and other international surveys, such as Pan-STARRS and GALAH, which were produced with ground based telescopes in the US and Australia.	Improved insights into the Gaia selection function and new insights into the selection functions of the other surveys that boost the scientific exploitation of their data.	5, 6
The combined data shall further increase the scientific return and enable new research activities using existing data sets.	The availability of a selection function for Gaia and combinations of Gaia with other surveys will boost the quality of the scientific data exploitation of all surveys and also unlock new science applications which are not possible now.	Higher number and higher quality of publications based on data from a combination of a European space mission and ground based astronomical surveys.	2, 3, 4, 5, 6

Call specific challenge/scope	**How GaiaUnlimited addresses these**	**Project outcome/benefit**	**Work pack-age(s)**
These activities shall add scientific value through analysis of the data, leading to scientific publications and higher level data products, tools and methods.	An essential component of GaiaUnlimited is to apply the selection function tools to the scientific exploitation of the Gaia data in combination with other surveys. The specific objective of this project is to provide data products and tools to enable users to apply Gaia survey selection function in their scientific research.	Scientific publications and data products as well as open source tools that allow for the use of the selection functions.	4, 6
When possible, enhanced data products should be suitable for feeding back into the ESA archives.	The data products resulting from GaiaUnlimited efforts will be made available to the scientific community through the ESA archive.	Enhanced value of the ESA archive. Discoverable data products to which public and efficient access is provided. Long term curation of the data products.	1, 4
Resulting analyses should help preparing future European and international missions.	The expertise built up in constructing the Gaia selection function will be transferred to other surveys and missions through the data, tools, and documentation made publicly available.	Better scientific exploitation of data from future European space missions or ground based surveys. Improved design of future European and international astronomical surveys. Enhance the science exploitation of future European space missions and inform the design of such missions to ensure the resulting surveys have tractable selection functions.	2, 3, 4, 5
International cooperation is encouraged in particular with countries active in space exploration and space science.	The participants in this proposal represent an international collaboration with in particular the involvement of partners from the USA and Australia, both countries which are active in space exploration and space science.	Strengthen European and international collaboration. Enhanced links between Europe and international space powers.	1

Call specific challenge/scope	**How GaiaUnlimited addresses these**	**Project outcome/benefit**	**Work package(s)**
Involve post-graduate scientists, engineers and researchers.	We specifically target young post-doctoral scientists/researchers to work on the GaiaUnlimited project.	Work experience at world-class European astronomical institutes. Exposure to a big space mission project, both to ESA and the Gaia data processing consortium. Experience in advanced data analysis methods applied to 'big data' problems.	1, 3, 4, 5, 6
Promotion of gender balance.	We will adhere to the principles of the European Charter for Researchers and the Code of Conduct for the recruitment of the post-doctoral researchers, taking care to ensure equal opportunity and gender balance.	Improved gender balance among post-graduate scientists.	1
Proposals are also expected to add value to existing activities on European and international levels, and to enhance and broaden research partnerships.	Provide data products and tools to incorporate the Gaia survey selection function in scientific analyses. Do the same for a few combinations of Gaia and other surveys. Achieve this through a team of European and international partners.	Boosting the quality and quantity of scientific exploitation of European space mission and international survey data. Strengthening and deepening of the collaborations between the partners in GaiaUnlimited who represent both European and international research institutes.	1, 4, 5

1.3 Concept and methodology

1.3.1 Concept

The goal of this project is to progress from the concept of a selection function for the Gaia survey and combinations of Gaia with other surveys (for which the statistical and mathematical principles are known and simple implementations have been published and used, Technology Readiness Level 2/3) to a publicly available practical implementation of the selection functions (in the form of data products and open source computer applications, TRL 8). These can then be applied by researchers in their scientific exploitation of the Gaia data, which will unlock new astrophysical investigations as explained in section 1.1.3 above and increase the number of publications from Gaia as well as other surveys.

Our goals will be achieved through research and development (including application development) conducted at the GaiaUnlimited partner institutes. The bulk of the effort is budgeted for the European partners but essential expertise and effort will contributed by the partners in the US and Australia (see section 3.3 for the motivation to involve these non-EU partners). The realization of the Gaia selection function requires:

- a refined operational (mathematical) definition of the concept of a selection function as a framework for the overall effort (TRL 2);
- a detailed understanding of the Gaia sky scanning history and the various points at which measurements or objects are lost or discarded for one reason or another;
- bringing the above together in a specific realization of the selection function for Gaia (TRL 3);
- using the knowledge of the Gaia selection function to produce selections function for combinations of data from Gaia with other surveys (TRL 3);
- developing the necessary software tools and producing numerical data which together constitute a practical implementation of the above selection functions (TRL 4/5);
- testing the tools by applying them to real science cases and making prototype versions publicly available for user testing (TRL 7);
- making the software tools and numerical data publicly available to enable scientists to incorporate the selection function in their analyses of Gaia and other survey data (TRL 8).

The GaiaUnlimited effort must be guided by developing in parallel a number of concrete and diverse science applications in order to ensure that the requirements on the selection function (level of detail, best practical implementation) are properly understood and incorporated in the data products and tools delivered through the efforts proposed here. Section 1.3.2 describes in detail how the above elements are implemented through the work packages listed in Section 3.1.

Finally, we stress at this point that individual members of the astronomical community are already making progress toward constructing the Gaia selection function. We point in particular to the recent work by Boubert & Everall [8, 9] which includes data products and tools for the user[5]. However, these efforts, although extremely valuable, are somewhat ad-hoc and dependent on the voluntary work by individual astronomers who are naturally focused on solving the problems most pertinent to their own research. What is needed is a concerted effort to come to a generally applicable and usable Gaia selection function. Nevertheless the existing community efforts underline that GaiaUnlimited is timely and that the efforts we propose to embark on are indeed realistic.

1.3.1.1 Stakeholders

The stakeholders in this effort are:

- the astronomical scientific community, which will benefit from the enhanced capability to exploit the Gaia (and other survey) data, and has repeatedly expressed the need for a practically usable selection function;
- the DPAC will profit from the detailed understanding of the selection function during the quality control of the Gaia data products and gain insights into how to improve the data processing such that the selection function is kept as simple and tractable as possible;
- ESA will benefit from insights into future mission design that grow out of the detailed investigations of the selection function and from the strongly enhanced legacy value of the Gaia archive combined with the tools to properly account for selection effects;

[5] https://github.com/DouglasBoubert/selectionfunctions

- other sky survey projects which either base their target lists on the Gaia catalogue or use Gaia data as a reference for their calibrations will benefit from a detailed understanding of how their selection functions are affected by the use of Gaia data.

In return the expertise of these stakeholders will benefit this project:

- the scientific community will be engaged through publicly available prototypes of the GaiaUnlimited tools and community workshops at which user feedback and requirements can be captured;
- DPAC will play a natural role, through providing information and expertise, in understanding in detail all the selection effects entering the Gaia catalogue;
- the ESA expertise with data archiving and dissemination can be tapped for the efforts to make the selection function publicly available;
- the data from other sky surveys will be used in out research on the Gaia selection function.

1.3.1.2 Links to international research and innovation activities

The links to ESA and DPAC are a natural element in this proposal because of the role of the coordinator and the contact persons at the European partners who are members of the DPAC consortium (see section 3.3). Both the ESA scientific space programme and DPAC will benefit from the activities of GaiaUnlimited as explained above. Links to other international astronomical survey projects exist through the membership of some of the GaiaUnlimited team members in these surveys (e.g. WEAVE, 4MOST, Pans-Starss, GALAH). The links to the scientific community are established through the academic institutes where the work will take place and through participation in international workshops and conferences. Further links exist with the MW-Gaia COST action[6] in which several of the partners on this proposal participate. This will facilitate participation in workshops and conferences organized by MW-Gaia which are natural venues for the dissemination of the knowledge and tools built within GaiaUnlimited.

1.3.2 Methodology

The Gaia selection function in an ideal world would look like the probabilistic graphical model shown in figure 9. In this figure and the subsequent text the three main Gaia instruments as referred to as ‘AF’ for Astrometric Field (collecting the astrometric data), ‘XP’ which is short for the BP and RP photometric instruments, and RVS, the radial velocity spectrometer. The model in figure 9 shows that AF and XP observations of sources are collected whenever one of the Gaia telescopes scans over a source which has a magnitude G_source brighter than the magnitude limit G_lim. If a source is bright enough in G_RVS data are also collected by the RVS instrument. The pointing of the Gaia telescopes is determined by a ‘scan law’ which is fixed for the mission duration. The probability of a source entering the Gaia catalogue would thus be determined only by its true apparent magnitude G_source, the number of times the Gaia telescopes scanned across the source, and the uncertainty σ_G in the on-board estimate of the source brightness. The probability that a radial velocity is listed in addition to astrometry and photometry would be determined independently by the true apparent magnitude of the source in the RVS wavelength range, $G_\mathrm{RVS,source}$, in combination with the scan law and the uncertainty σ_{G_RVS} of the on-board estimate of $G_\mathrm{RVS,source}$. The selection function for any derived data products, such as the astrophysical parameters of the sources, would be driven entirely by the basic data (astrometry, photometry, radial velocity, XP spectra, RVS spectra, or any combination thereof) used as inputs.

In reality the Gaia survey selection function is vastly more complicated owing to the many stages in between source detection and final data product at which decisions are made on the spacecraft whether

[6] https://www.mw-gaia.org/

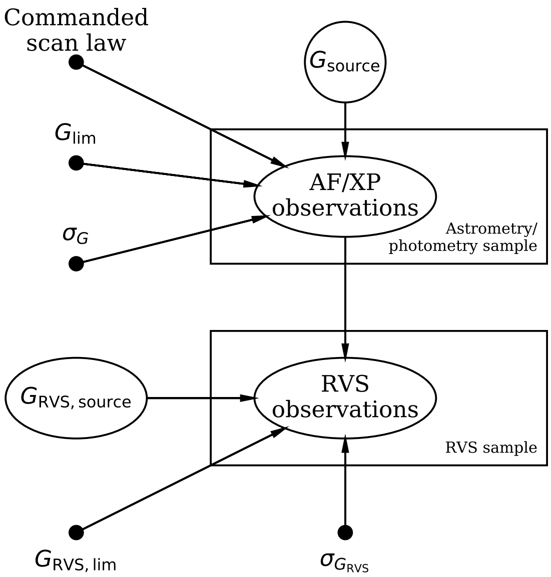

Figure 9: Idealized Gaia selection function. Here it is assumed that the spacecraft exactly follows the scan law as commanded from ground without any interruptions. The on-board decision on whether or not a source is observed depends only on the G and G_{RVS} magnitudes of the source, the corresponding fixed survey limits, and the uncertainties on the magnitude estimates. See the text for a more detailed explanation.

or not to make a certain observation, and in the on-ground data processing whether or not to include that observation. This is illustrated in figure 10.

- The actual sky-scanning strategy executed by Gaia will differ from the desired one due to imperfections in the attitude control and interruptions of observations due to outages such as station keeping manoeuvres or on-board activities that require interrupting the measurement process (e.g., decontamination through heating of the payload, refocusing of the telescopes, more details can be found in [19]).

- At the detection stage the brightness of the sources is estimated (with some uncertainty). This is indicated by G_{VPU} in the figure 10, where VPU stands for ‘Video Processing Unit’, essentially an on-board computer. In parallel an attempt is done to weed out spurious detections due to cosmic rays or PSF spikes, and sources are prioritized in order to deal with the limitation on the amount of sources that can be handled by the VPUs at any one time.

- Following confirmation that a detected source is real (through a second measurement of the source flux) astrometric and photometric data are collected, subject to further limitations in the amount of sources that can be handled (indicated with ‘VPU resource management’ in figure 10).

- RVS observations are collected only for sources bright enough in G_{RVS} which is estimated for bright sources through extrapolation from the on-board estimate G_{VPU} of G_{source} or through an estimate $G_{\mathrm{RVS,VPU}}$ from the flux in the RP spectrophotometric data. The latter introduces a source colour dependence.

- The data collected on-board is partly transmitted to the ground stations in real-time (during spacecraft contact periods) and is partly stored on-board for later transmission. The on-board storage limitations lead to data being deleted if it stays in the storage area too long (a consequence of limited availability of ground station time and the size of the data storage area). This

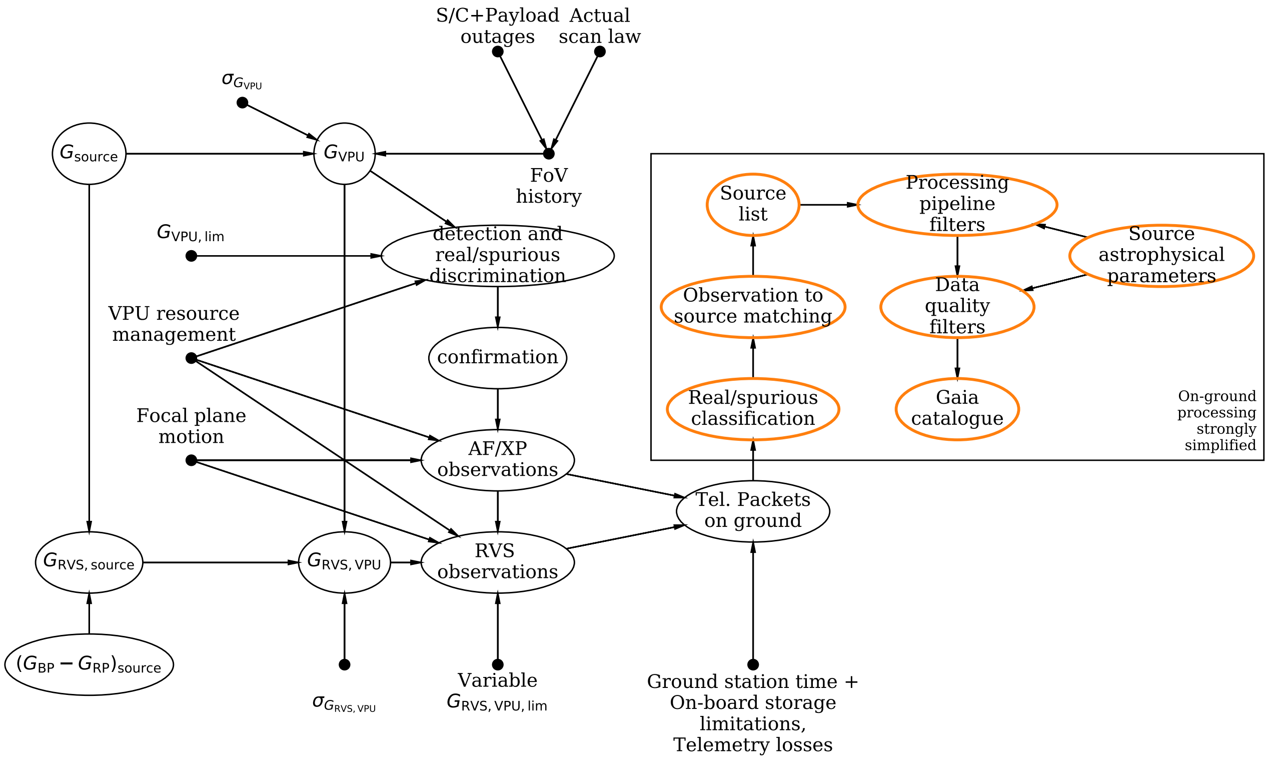
S/C+Payload outages
Actual scan law
$\sigma_{G_{\rm VPU}}$
$G_{\rm source}$
$G_{\rm VPU}$
FoV history
$G_{\rm VPU,\,lim}$
detection and real/spurious discrimination
VPU resource management
confirmation
Focal plane motion
AF/XP observations
$G_{\rm RVS,\,source}$
$G_{\rm RVS,\,VPU}$
RVS observations
Tel. Packets on ground
$(G_{\rm BP} - G_{\rm RP})_{\rm source}$
$\sigma_{G_{\rm RVS,\,VPU}}$
Variable $G_{\rm RVS,\,VPU,\,lim}$
Ground station time + On-board storage limitations, Telemetry losses
Source list
Processing pipeline filters
Source astrophysical parameters
Observation to source matching
Data quality filters
Real/spurious classification
Gaia catalogue
On-ground processing strongly simplified

Figure 10: Simplified diagram of the actual Gaia selection function. The left hand side of the diagram provides a fairly complete picture of all the ingredients on-board the spacecraft that contribute to the selection function before the raw data is in hand on the ground. The right hand side shows in strongly simplified form how the Gaia selection function is further affected by the decisions taken along the steps from the initial treatment of the raw data to the final published Gaia catalogue. See text for more detailed explanations.

happens primarily during so-called 'galactic plane scans', when both Gaia telescopes are pointing at locations in the Milky Way plane on the sky. The data deletion decision is taken on the basis of a detailed priority table stored on-board. Additional data losses may occur due to lost telemetry packets (a minor contribution affecting roughly one in a million packets).

The right hand side of figure 10 shows a strongly simplified representation of further decisions taken along the DPAC pipelines on whether or not to process certain sources or parts of their data.

- As a first step in the processing sources are classified as real or spurious, where the latter are typically caused by bright star PSF spikes or stray light features. Sources classified as spurious are 'blacklisted' and subsequently not considered for processing. This classification process is imperfect.
- The next step is to group observations of the same source together and produce the so-called 'source list'. This is a table linking a given source to all its observations. This process is redone from scratch for each data release and is subject to errors, such as observations being assigned to the wrong source, or the observations of one source being assigned to multiple different sources. Two examples of specific challenges for the creation of the source list are crowded regions, where the source density on the sky can lead to observation to source matching uncertainties, and variable stars, where a strongly varying apparent brightness can confuse the matching process.
- Once the source list is created all subsequent processing steps are taken, from generating the astrometry, photometry, and radial velocity data, to the astrophysical characterization of sources,

to the publication of a Gaia data release. At each step further decisions are taken on including or not sources or parts of their data. Often these decisions are explicitly or implicitly dependent on the source astrophysical parameters which will introduce further source brightness and colour dependencies into the selection function.

From the description above it is clear that the Gaia survey selection function is very complex and, depending on the particular science case being addressed with Gaia data, the selection function must be known in great detail. An example is the list of binaries to be produced on the basis of Gaia data. The list will consist of various solution types (varying from mere indications of binarity to complete orbit solutions) which will be derived only for sources for which there is sufficient evidence that they are not single stars. The resulting list will thus be incomplete in a way that intricately depends on the binary star parameters, with many binaries still hiding among the sources not examined in detail in the DPAC pipeline.

The objective of GaiaUnlimited is to derive a detailed description of the Gaia selection function which will be provided to and used by the scientific community. This work is divided into the tasks below which map to the work packages listed in section 3.1 (the management of GaiaUnlimited, WP1, is discussed in section 3.2).

Operational definition of the Gaia survey selection function (WP2) To guide the overall effort on the Gaia selection function it is important to precisely define what we mean by a selection function. We choose to take a probabilistic approach (Eq. 1), thus asking the question: what is the probability of source X appearing in sub-sample Y of the Gaia catalogue? The mathematical formulation should at the top level be applicable to any astronomical survey but detailed mathematical definitions tailored to the Gaia survey may be needed. The formulation should account for the multiple layers of selection described above and allow for incomplete information on the data collection process. The formulation should also make it possible to accommodate additional user imposed selections (such as studying only sources with certain astrophysical parameters or demanding some minimum value of the data used). A hierarchical probabilistic formulation is a natural choice but other approaches are not excluded. This task should be completed within the first 12 months of this project.

Specification of the Gaia selection function (WP3) With the above mathematical formulation as the framework, this task is concerned with researching and working out in detail the description and modelling of the Gaia survey selection function (where we anticipate that some aspects cannot be described in all detail and will require simplifications or modelling). All the necessary information will be collected to describe or model in detail the elements in figure 10, which will require close interactions with experts from the DPAC. An assessment will be made of the necessary level of detail to which the ingredients of the selection function must be known (in order to keep the scope of the effort realistic). The resulting description of the selection function will be hierarchical, starting from the basic selection steps and specializing to selection functions for specific subsets of the Gaia catalogue (such as, for example, the variable stars), and will allow for the inclusion of user imposed additional selection criteria. Part of this task will also be to collect data from other surveys which may aid in the understanding of the Gaia selection function. Examples are higher spatial resolution data from the Hubble Space Telescope to study selection effects in crowded regions, or photometric/spectroscopic data which can aid in the understanding of the selection effects on the distribution of source astrophysical parameters in Gaia samples.

Tools to incorporate selection functions in scientific analyses (WP4) To enable the scientific community to use the Gaia selection function, a set of data products and tools will be provided. The tools will implement in code the mathematical formulation and detailed description/modelling of the Gaia selection function. A means to layer user selections on top of the Gaia selection function will also be provided. The tools will be backed by the data products needed to use them such as numerical tables. These data products will be made available through, e.g., https://zenodo.org/, and in query-able form through the Gaia archive at ESA. The code underlying the tools will be open source and accessible through popular code-hosting facilities such as https://github.com/. Extensive documentation will also be provided in the form of papers in open access journals and detailed online documentation. A web-portal will be set up as entry point to the data, tools, and documentation.

An important component of this task is making prototype versions of the tools publicly available and to bring users together in community workshop to use and test the tools. The feedback will be used to improve the prototypes.

Selection functions for combinations of Gaia and other surveys (WP5) In many science applications the Gaia data will be combined with data from other surveys, often leading to the production of hybrid catalogues based on the input surveys [e.g., 1]. In such applications one needs to account for the combined selection function for the intersection of the surveys involved. In this task a generic method to combine survey selection functions will be researched and developed. The method will then be applied to generate selection functions for the combination of Gaia and a limited number of popular sky surveys. The tools and data products needed to combine selection functions will be made available through the channels mentioned above. The scientific community can use these to construct selection functions for other combinations of surveys.

Apply the selection function tools to selected science cases (WP6) We believe that the best way to guide the overall efforts in this proposal and to thoroughly test the implementation of the Gaia survey selection functions, is to apply these to a number of science cases. The science cases will be developed in parallel to the selection function, both as a means to provide early tests of — and feedback to — the tasks above, and as a means to keep the researchers involved in this effort motivated. The results of the science applications will be published in the open access peer reviewed literature. It is very important for their future career that the post-doctoral researchers carrying out the bulk of the work will also have a publication featuring a science application to add to their CVs.

1.4 Ambition

As pointed out in section 1.1 there is currently no detailed selection function available for the Gaia survey. It is not part of the DPAC tasks or data products, and the limited attempts that have been made in the community to characterize the selection function have led mostly to ad-hoc implementations for specific science cases. In fact the vast majority of works on the Gaia data so far ignore the selection effects altogether. More generic attempts at describing the Gaia selection function have been undertaken by Boubert & Everall[7], and by Rybizki, Drimmel & Carrasco[8]. However, these efforts only scratch the surface of the complexities of the Gaia selection function. The efforts proposed here will thus go well beyond the state of the art:

- For the first time a Gaia survey selection function will be provided at the level of detail and with the tools needed to take it into account in any science application. This will push the scientific exploitation of Gaia (and other survey) data beyond the current state of the art by unlocking Gaia science applications which are currently not possible (section 1.1.3.2).
- The efforts to describe and implement the Gaia selection function will lead to a clarification and better definition of the concept of selection functions and biases. This will benefit the development of selection functions for other surveys now and beyond the lifetime of this project.

[7] https://github.com/DouglasBoubert/selectionfunctions

[8] https://github.com/jan-rybizki/gdr2_completeness

- The availability of the selection function will improve all the scientific analyses of the Gaia data. This will lead to a much better scientific exploitation of the Gaia data in that selection biases can be properly accounted for, while the availability of open source tools and corresponding open access data will facilitate a much better reproducibility of results. Overall this will lead to more publications from this European space mission.
- The innovations in terms of the mathematical formulation of the selection function and a usable detailed practical implementation thereof can be transferred to other sky surveys, both past and future, enhancing the quality of the scientific exploitation and their legacy value.

Timing of this proposal We stress here that the time to research and implement a detailed selection function for the Gaia survey is *now*:

- To understand the selection function in detail, accounting for all the issues described above and in section 1.3.1, requires access to the Gaia spacecraft and data processing expertise within DPAC and ESA. It is thus essential to start this work while the DPAC is still active.
- Already now the availability of a selection function for the Gaia DR2 catalogue would represent a major boost to the science exploitation of those data.
- Which each upcoming Gaia data release the complexities of the selection function will increase thus making any effort postponed to the future much more costly in terms of human resources required.

2 Impact

2.1 Expected impacts

We expect the following impacts directly related to the work programme:

- The availability of a detailed Gaia selection function coupled with the data and tools to use it will **enhance the quality and reproducibility the scientific exploitation of data from a flagship European space mission**. Thus great value is added to the existing European efforts dedicated to the data processing for the Gaia mission as well as to the efforts to produce the publicly available Gaia data releases.
- A readily available Gaia selection function will enable science cases to be addressed that are currently hampered by the lack of a detailed description of Gaia survey selection biases. This will translate to **an increase in the number of scientific publications based on European space missions**.
- Although the participants on this proposal have all worked together in various constellations in the past on different topics, this is the first time that this group of experts, inspired and excited to invest time in the details of the Gaia selection function, will work closely together, thus enhancing an existing collaboration and broadening the expertise of the partners involved.
- The partner institutes will benefit from having Gaia and survey selection function experts in house, expertise which can then be turned into higher quality scientific exploitation of astronomical data and more publications (based on data from European space missions and surveys).

- **The European scientific community will gain a competitive advantage** in the exploitation of space mission data with respect to their peers worldwide through the expertise and the tools built within GaiaUnlimited. The expertise will be transferred to the community through the workshops that GaiaUnlimited will organize, and through presentations at relevant scientific conferences.
- The expertise on the Gaia selection function will be transferred to DPAC. This will facilitate turning the selection function into a standard DPAC data product for the community. In addition a detailed selection function allows for a more careful validation of the DPAC data products. For example, features in the population statistics of sub-samples of the Gaia data can then be correctly attributed to a selection bias or a problem in the data processing. The result will be better quality control before each Gaia data release.
- The expertise gained in this project can be transferred to other surveys. This is already partly planned in this proposal through work package 5 on combined selection functions. However we expect that other survey teams will seek out the expertise developed in this project and pick up the methods and tools we make available. This will enhance both European and international activities. **Examples of future European projects that stand to benefit are the 4MOST and WEAVE spectroscopic surveys and the Euclid and Plato space missions.** Examples of **international projects that will benefit are future surveys such as LSST and SDSS-V (USA) as well as existing surveys such as GALAH (Australia)** which rely heavily on complementary Gaia data.
- **The scientific community world-wide will benefit from the higher quality scientific analyses** that are possible with the public availability of the Gaia survey selection function.

On a broader societal level the impact will be through the expertise built up with understanding in detail how the interpretation of large amounts of collected data is affected by selection biases and how this limitation can be addressed through a proper description of the way the data was collected. This will be of benefit to any data science application and we expect that some of the researchers funded through this proposal might opt for a future career in European companies of which the business contains a large data science component.

Potential barriers/obstacles Other than the normal risks associated with scientific research we do not foresee any barriers to achieving the impacts above. The work proposed is not subject to regulation standards and presents no workforce issues. Although follow-up steps could be foreseen, e.g. further detailing the Gaia selection function beyond what is achieved in GaiaUnlimited, the absence of such steps will not present a problem. GaiaUnlimited is self-contained in that respect. The project implementation does of course have some risk factors as listed in the table of risks.

2.2 Measures to maximize impact

2.2.1 Dissemination and exploitation of results

The area in which we expect to have the highest impact is the scientific exploitation of results from the Gaia mission and other surveys. Correspondingly our dissemination efforts will be aimed at scientists using astronomical survey data and we will primarily use the standard academic channels:

- Presentation, description and documentation of our results through scientific publications in peer-reviewed journals. We will make use of open access journals and we will always make our publications available through https://arxiv.org. The latter repository is listed in the Registry of Open Access Repositories as http://roar.eprints.org/89/, and is the most popular repository for the fields of astronomy and astrophysics. We will deposit all our journal articles in this repository either upon submission or upon acceptance of the articles in question (the choice is left up to the authors). In this way the article is openly accessible (as PDF and in original source format, typically LaTeX) in parallel to any period during which the journal article is only open to subscribers (this is typically a one-year period for astronomical journals). Hence, although the journal articles are 'green open access', depositing them with `arXiv.org` in parallel effectively ensures 'gold open access'.

- Our results will be presented and promoted at scientific meetings where we aim to reach a diverse audience through the variety of science applications addressed as part of our efforts.
- We will organize three community workshops in which the Gaia selection function experts will offer training to interested scientists in the use of our data products and tools (deliverables D1.3, D1.5 D1.7). Our results from the science applications in work package 6 will serve as excellent worked examples.

This project will collect the data needed to understand the Gaia selection function, such as the details of the actual scan law used or of the spacecraft operation interruptions. These data will be part of the tools needed to use the Gaia selection function and thus will be stored in a form suitable for computer use or for storage in a database.

- All data generated/collected in this project will be made available publicly. This will be done through the ESA Gaia archives and any other astronomical data centres interested in hosting these data. In addition the data will be deposited with https://zenodo.org.
- The source code for the computer tools will be made available publicly through well known open access source code repositories such as http://github.com or through code hosting services available at the institutes participating in this proposal.
- A website will be setup (hosted by Leiden University) which will serve as entry point to the data, tools, and documentation generated in this project. To enhance visibility cross-links will be made with, for example, ESA websites related to Gaia.

The Gaia mission operations (including mission extensions) will end at the latest around end 2024 and the data processing will continue for a number years thereafter before the final (legacy) Gaia archive version can be released. This is beyond the lifetime of the project proposed here and we thus intend to transfer the expertise on the selection function to the DPAC aiming to make the selection function a data product provided by DPAC. This is most naturally provided together with the documentation through the ESA archive. Hence a long term curation of the corresponding data is foreseen as part of the Gaia legacy archive maintained by ESA. The Gaia archive is expected to remain the standard in astrometric and photometric data for decades to come.

The long term curation of the GaiaUnlimited tools and corresponding source code is best addressed by embedding it in an existing eco-system. The Astropy project[9] is a good candidate; as stated on the website, 'Astropy is a community effort to develop a common core package for Astronomy in Python and foster an ecosystem of inter-operable astronomy packages'. One of the GaiaUnlimited team (Price-Whelan) is a major contributor to Astropy [3] and with his contacts and expertise in hand we will work to ensure the GaiaUnlimited tools are embedded in Astropy or released as an affiliated package.

2.2.2 Communication activities

As mentioned above the community explicitly targeted by this proposal, professional astronomers, will be reached through the well known academic outreach channels: scientific publications, presentations at scientific conferences, and open access to the data and tools resulting from this project. In addition

[9] https://www.astropy.org/

we plan to engage the general public through media releases of the results from application of the Gaia selection function to specific science cases, highlighting how the work on the selection function enables much better interpretation of, for example, binary star and exoplanet population statistics. This will also be used as a hook to explain how a science like astronomy can enhance data analysis capabilities for society at large. After all, the concept of a ‘selection function’ is applicable to population modelling across all fields of research; spelling out a framework that is both intuitive and mathematically sound matters widely, far beyond Gaia or astrophysics.

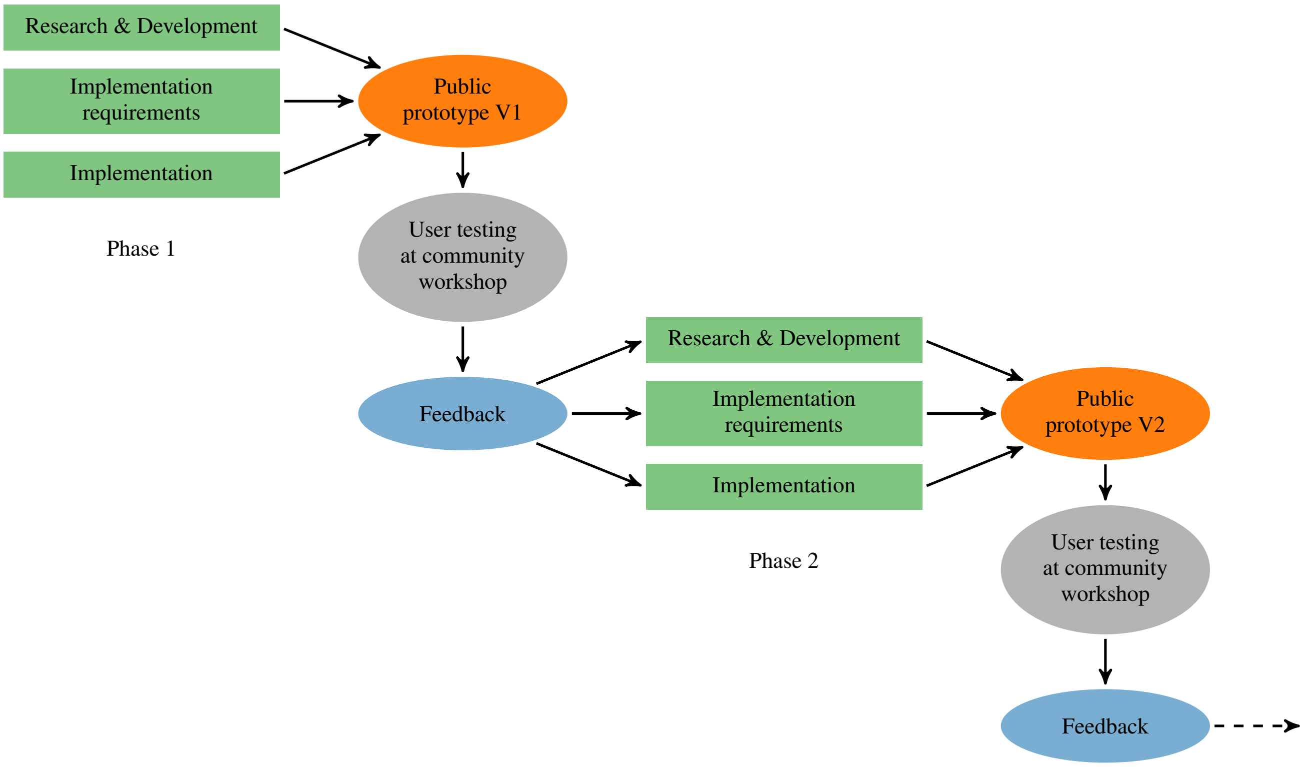

Figure 11: Iterative workflow for GaiaUnlimited.

3 Implementation

3.1 Work plan — Work packages, deliverables

3.1.1 Overall structure

The overall approach to the work plan is to have the research and development of the selection function progress in parallel with the implementation thereof. We made this choice because of the complexities of the selection function and its realization in the form of tools and auxiliary data, and because of the relatively short duration of the project. The work flow we foresee for GaiaUnlimited is shown in figure 11. It is an iterative workflow in which the findings in each of the work packages 3–6 will inform the others (see figure 13). For example the research on the selection function will lead to requirements on the implementation, while at the same time the conversion of research ideas into practical tools will lead to the questioning and revision of those same ideas.

At several points during the project the results will be combined into prototype implementations of the selection function tools. These will be tested both internally to the project and through the public availability of two prototype implementations. Specifically, we plan to organize two community workshops just after the public release of the prototypes to train members from the astronomical community in the use of the tools and to collect user feedback. The results from these tests will feed into the next round of research, development, and implementation. The scientific nature of much of the

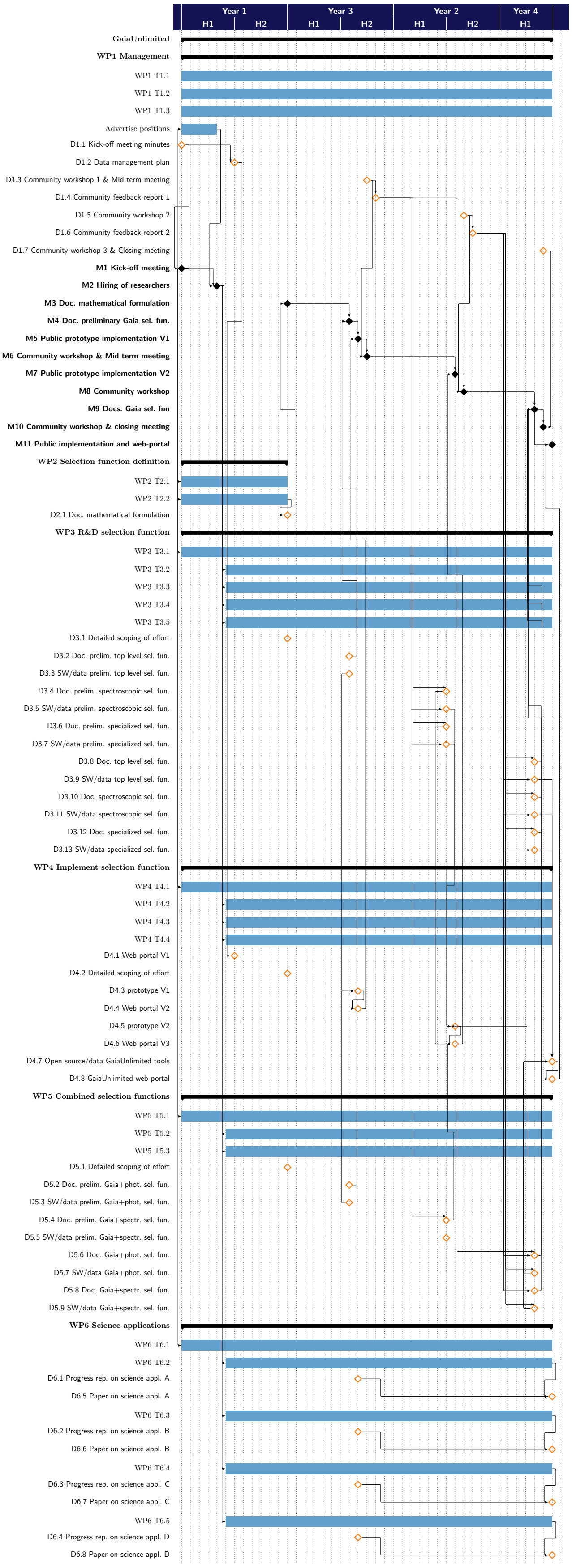

Figure 12: Gantt chart for the GaiaUnlimited project.

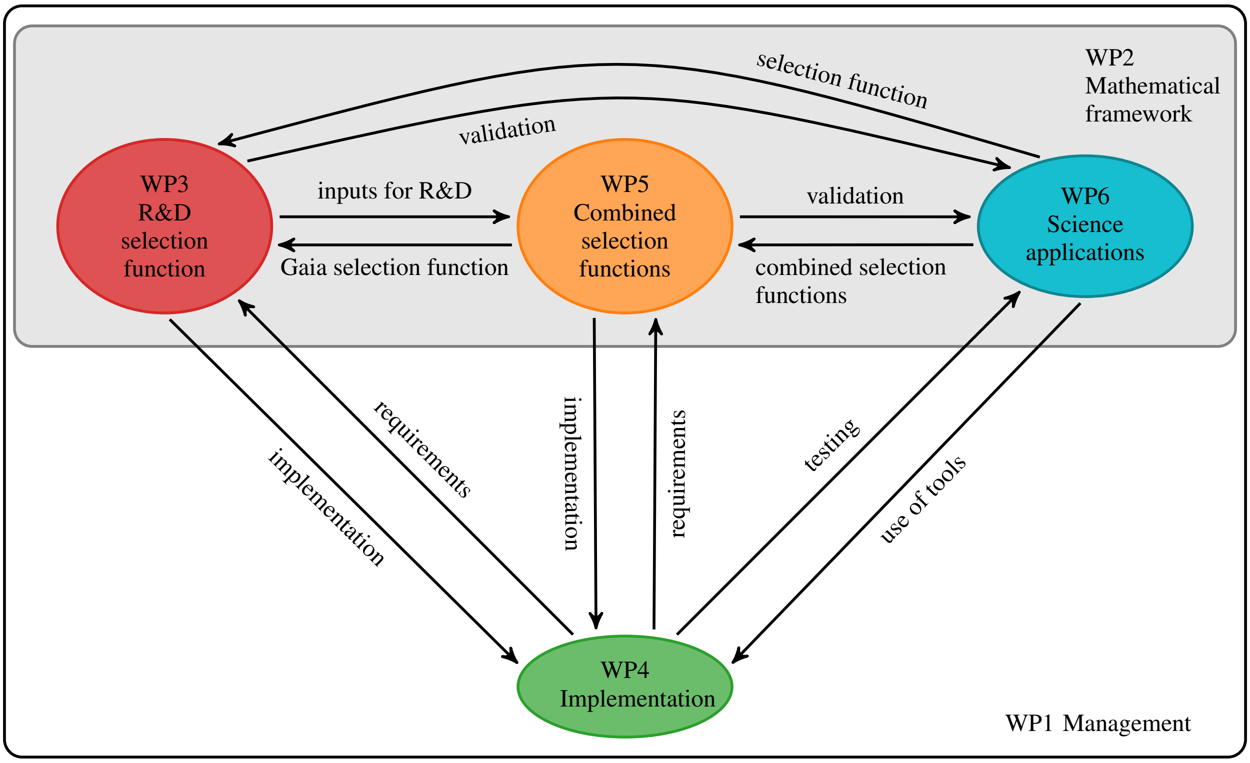

Figure 13: The dependencies between the work packages in GaiaUnlimited.

work conducted in GaiaUnlimited will very much benefit from an iterative approach. Close coordination between the work packages will of course be essential and this is foreseen in the management structure.

3.1.2 Timing of the work packages

The overall work plan is structured as follows:

- The first 6 months will be used to recruit the post-doctoral researchers to be funded from this proposal and to carry out the tasks in work package 2. Having a mathematical framework for the selection function in place early on in the project is important for guiding the research and the implementation phases.

- During the next three years of the project the efforts in work packages 3 to 6 will happen in parallel as illustrated in figure 11. The iterative approach outline above will be used in this phase.

- The selection function implementation prototypes to be developed in each phase will increase in their level of detail and complexity. We explicitly foresee two public prototypes (see figure 12 and WP4), which will be presented and tested at two community workshops. Further internal prototypes are also foreseen. Initially we may chose to implement a prototype selection function that depends only on sky position and apparent magnitudes (figure 9), without any consideration for observation interruptions on the spacecraft, or on-board or ground processing effects. It would then be natural to include spacecraft outages, as these events are well-documented. Depending on the level of complexity envisaged, Prototype 2 may include on board processing effects like magnitude estimates (given the input source catalogue) and telemetry effects. Prototype 2 could also include pipeline data processing interventions that occur on the ground, or how astrophysical environments would affect the observable properties (e.g., binarity, variability). The scope of each prototype will be iteratively guided by user testing, feedback, and the executive board.

3.1.3 Dependencies

Figure 13 shows the dependencies between the various work packages in GaiaUnlimited. The arrows indicate a 'depends on' relationship. For example WP5 depends on WP3 for the details of the Gaia selection function in order to produce combined selection functions. The reverse dependency shows that studying combined selection functions will also lead to constraints on the Gaia selection function. WP6 provides the science applications which depend on WPs 3 and 5 for the selection functions to apply and on WP4 for the tools to do so. In return WP6 provides crucial validation of the results from WPs 3 and 5 and extensive testing of the implementation provided by WP4. WP2 provides the framework for WPs 3/5/6 as indicated by the grey rectangle. The management work package (WP1) ensures the overall coordination of the GaiaUnlimited effort.

3.2 Management structure, milestones and procedures

3.2.1 Management

The GaiaUnlimited consortium being relatively small the management structure can be kept simple in order to make the coordination of the efforts efficient. This structure works well for the small group of highly motivated and self-driven researchers which constitute GaiaUnlimited. The management consists of three levels, reflecting the structuring of the GaiaUnlimited work packages.

Administrative management This first level concerns the administrative management of GaiaUnlimited, including financial management and reporting to the European Commission. It is carried out by the GaiaUnlimited coordinator (A. Brown) with the support of the Leiden Observatory institute management team.

Overall project management and coordination This second level concerns the global scientific and technical control and coordination of the project, ensuring that the tasks are properly carried out and remain on schedule, and that the objectives of GaiaUnlimited are fulfilled. It is carried out by the **GaiaUnlimited executive board**, formed by the managers of the work packages: A. Brown, H.-W. Rix, R. Drimmel, and V. Belokurov. In order to keep communications within the consortium as efficient as possible the main contacts at NYU and MONA (D.W. Hogg and A. Casey) will have a standing invitation to attend the executive board meetings (see below).

Work package management The third level of management concerns the main work packages. Each WP manager has the responsibility of supervising the execution of the core tasks assigned to their host institution, the work of the partners carrying out specific parts of the tasks attached to the work package, and report to the executive board accordingly.

We do not further specify the management procedures for each work package beyond the items given above, since the work in each of the work packages 2 to 6 will be integrated in already existing and well established teams at each host institution.

This management structure together with the procedures described below is sufficient for a relatively small team spread over only six institutes working together closely in a scientific environment. Much of the coordination and supervision will take place through day-to-day contacts at the institutes involved and between institutes via e-mail and ad-hoc telecons. Similar structures were applied successfully to other EU-funded projects (FP7 GENIUS Collaborative project, FP7 GREAT Marie-Curie Initial Training Network) in which the GaiaUnlimited coordinator was an active participant.

3.2.2 Innovation management

The aim of GaiaUnlimited is to develop and deliver a new product (the Gaia selection function and tools to apply it) to the market of astronomy researchers. The wish for a selection function for the Gaia survey has been expressed on numerous occasions, meaning that this product will readily find a large user base. In order to uncover opportunities the astronomical community will be engaged during the lifetime of GaiaUnlimited through the availability of public prototype implementations of the Gaia selection function. Two community workshops will be organized just after the release of the prototypes at which users can test the tools by applying them to their science cases. At the same time they will be asked for feedback in the form of missing features or desired improvements. Community feedback will also be obtained through presentation of GaiaUnlimited results at relevant scientific conferences.

3.2.3 Procedures and reporting

We list here the procedures that will apply to the management of GaiaUnlimited:

- The GaiaUnlimited coordinator will report to the European Commission following the rules applicable for research and innovation projects within the Horizon 2020 programme.
- GaiaUnlimited will hold at least three plenary meetings open to participation of all its members: the kick-off, mid-term, and closing meeting. The participation of the partner institute coordinators (or a representative) will be mandatory. Other plenary meetings may be held if needed.
- Three community workshops will be organized which will be held in parallel to the plenary meetings. These meetings will be the occasion to train the community in the use of GaiaUnlimited tools and to have the users test the tools through using them.
- The executive board will hold teleconferences at least once a month to track the status of the tasks. The minutes of these teleconferences will be made available to the members of the consortium.
- The executive board shall ensure the coordination of the GaiaUnlimited work with the wider activities on the Gaia mission and its data releases, in particular ensuring the coordination with ESA and DPAC (see section 2.2.1). This can be done efficiently through membership of several GaiaUnlimited team members in DPAC.

3.2.4 Consortium agreement

The internal workings of GaiaUnlimited and the roles and responsibilities of the partners will be formalized through a consortium agreement which will be drawn up and agreed at the start of the programme. It will define the decision taking procedures in GaiaUnlimited and the mechanisms for conflict resolution, which will be channelled through the Executive Board, as well as the management of intellectual property rights as described in section 2.2.1 and section 3.2.2 above.

3.2.5 Recruitment strategy

The recruitment policy for GaiaUnlimited will conform to the principles of the European Charter for Researchers and the Code of Conduct for their recruitment. It will take place in a globally coordinated way during the six-month setup phase described in section 3.1.2, placing an emphasis on individual excellence and capacity for team working, while taking care to ensure equal opportunity and gender balance. The recruitment strategy will be in line with institutional practices, where each of the partner institutes has equal opportunity and gender policies in place.

3.3 Consortium as a whole

The idea for this proposal grew out of discussions at several Gaia Sprints[10] in which all of the GaiaUnlimited members participated, Hogg, Casey, Price-Whelan, and Rix being members of the organizing team of the Sprints. In particular at the Gaia Sprint that took place in Santa Barbara (USA) in March 2019, two sessions were dedicated solely to discussing the Gaia selection function. It was concluded that only a significant and dedicated effort, specifically funded, could realistically lead to the construction of the Gaia selection function and the corresponding tools and data needed to use it.

The consortium represents a combination of Gaia/DPAC experts (Brown, Drimmel, Fouesneau) and experts that have given much thought to the principles of selections functions and worked on aspects of the Gaia selection function (Hogg, Rix, Drimmel, Price-Whelan, Casey, Fouesneau, Belokurov, Everall). All the participants have undertaken scientific analyses of the Gaia data that clearly highlight the need for a selection function [e.g., 13, 14, 32].

- The Gaia expertise in the consortium is contributed by long-standing DPAC members who are deeply involved in the data processing for the Gaia mission. They provide expert insight into the detailed ingredients of the Gaia selection function and, more importantly, have an extensive network of contacts with experts in DPAC who can provide more detailed information where needed. The presence of the Gaia experts is crucial to the guidance of the researchers to be funded through GaiaUnlimited. We note that the other participants have extensive experience as users of Gaia data and thus bring an important complementary perspective to GaiaUnlimited.

- The expertise on the mathematical formulation of the selection function and on selection functions in general is essential to the success of GaiaUnlimited. The expertise stems from the actual construction of selection functions and applying these to analyses of a number of large surveys [e.g., 10, 13, 14, 16].

- The requirements on a selection function and how it is provided in its practical form must be guided by the experience gained through advanced scientific analyses of data such as from the Gaia mission. The experience from successes and failures in past analyses will be very important in setting the priorities for the research and development in GaiaUnlimited. The participants of this proposal provide the required expertise.

- The partners from Australia and the USA contribute deep knowledge of selection function issues based on their past scientific work. Their perspective as experienced users of the Gaia data will be essential to ensure that the developments in GaiaUnlimited stay focused on satisfying the needs of the astronomical community in general, and not just the needs of DPAC experts.

As mentioned in section 1.3.1 community efforts to produce (parts of) the Gaia selection function are already underway, such as the works by Boubert & Everall [8, 9][11] and Rybizki, Drimmel & Carrasco[12]. Everall and Drimmel are participants in this proposal and we will work closely with D. Boubert (Oxford) and will thus benefit from his expertise built up with concrete research into the Gaia selection function, as well as from his expertise in creating user tools.

The participants in GaiaUnlimited have a long track record of effective collaboration as can be seen from the many joint publications. Through the organization of Gaia Sprints and other workshops[13] the participants have a proven track record of engaging the astronomical community, an important aspect of making sure that the tools and data delivered by GaiaUnlimited fulfil actual user needs.

The coordinator, A. Brown, is also a member of the management committee of the COST action CA18104 'Revealing the Milky Way with Gaia'[14] and is co-I on the recently submitted proposal for a

[10] http://gaia.lol

[11] https://github.com/DouglasBoubert/selectionfunctions

[12] https://github.com/jan-rybizki/gdr2_completeness

[13] For example the Gaia DR2 Exploration Lab https://www.cosmos.esa.int/web/gaia-dr2-exploration.

[14] http://www.mw-gaia.org/

Marie Skłodowska-Curie Innovative Training Network also called ‘Revealing the Milky Way with Gaia’. We foresee participation in the conferences organized through the MW-Gaia COST action as part of the dissemination activities and if the ITN proposal is successful we would offer training in the use of the Gaia selection function to the PhD students in that network. This could take place for example at one of the schools planned by the ITN, or through participation of the PhD students in our planned community workshops (deliverables D1.3, D1.5, D1.7).

Role of visiting experts As noted in the next section we have also reserved a budget to support visits by experts to the European partners in GaiaUnlimited. Examples are: Gaia Data Processing and Analysis Consortium experts with whom we need to work closely to understand aspects of the Gaia instruments, observation strategy, and data processing that are essential to the implementation of the selection function (WP3, WP4, WP5); experts on the photometric and spectroscopic surveys for which we want to implement combined selections functions with Gaia (WP5); experts on the scientific applications that we will pursue to guide the GaiaUnlimited work (WP6). The experts will contribute their knowledge and insights to GaiaUnlimited in the way customary for scientific collaborations. At this stage we do not know the names of the experts we will invite. We included below a rough estimate of the number of visits to expect in order to calculate a corresponding budget. It is important to keep flexibility with respect to visiting experts, so we can react to the way the work in GaiaUnlimited develops.

3.4 Resources to be committed

The person-month distribution for the GaiaUnlimited effort is shown in the summary of effort table. The workload is concentrated mostly in WPs 3–6, reflecting that in those work packages the bulk of the work will be carried out by the researchers to be hired through GaiaUnlimited funding. The larger number of person-months assigned to MPIA reflects that combining catalogues and generating the corresponding selection functions carries the significant overhead of handling and matching multiple large catalogues.

We summarize below how we arrived at the project financial budget.

- Salary costs are calculated as follows:
  - Full salary costs are included for all researchers to be hired, which amounts to a total of 156 person-months at the rates corresponding to the institute at which they will be employed.
  - At each of ULEI, MPIA, INAF, and UCAM one researcher at the post-doctoral level will be hired for 36 person-months. At MPIA an additional researcher will be hired for 12 person-months, at the MSc or post-doctoral level, to work on tasks T4.3 and T4.4.
  - Full salary costs are included for the participants from the staff at ULEI, INAF, and UCAM (Brown, Drimmel, Belokurov).
  - Salary costs for the staff from MPIA, MONA, and NYU are listed but treated as in kind contributions, *no EU funds are requested for these salary costs.* This explains the zero budget request for MONA and NYU.
- Travel costs are calculated as follows:
  - For the staff at ULEI, MPIA, INAF, UCAM we estimate 1 trip in Europe per half year (7 trips total), budgeted at 1250 Euro each, and one trip to the USA, and one to Australia, both budgeted at 2500 Euro.
  - For the researchers hired at ULEI, MPIA, INAF, UCAM we estimate 1 trip in Europe per half year (6 trips total, with 2 extra trips at MPIA reflecting the extra 12 person-months),

budgeted at 1250 Euro each, and one trip to the USA, and one to Australia, both budgeted at 2500 Euro.

- For ULEI € 13 500 is added for an extended (6 month) visit to the work with the partner at MONA (task T4.4).
- A budget for visitors is included for each institute which will allow for a visit by the NYU and MONA partners to each of the European institutes over the lifetime of the project (each trip at 2500 Euro). The travel costs will be paid by the European institutes, participating in the project, and will not be invoiced to NYU and MONA.
- Each European institute participating in the project was assigned an additional 2500 Euro to accommodate for visitors from within Europe. These visitors are foreseen to be experts in the research areas of GaiaUnlimited who will be advise on specific aspects of the project.

- Equipment costs are estimated at 5500 Euro each for ULEI, MPIA, INAF, and UCAM. This allows for the provision of high end laptops and PCs to the researchers which is a necessity for this data and compute intensive project. The following details on the accounting of equipment costs hold for each beneficiary:

  ULEI The equipment will be used only for the project for the entire project duration. The equipment is immediately depreciated one hundred per cent after deployment.

  MPIA The equipment will be used only for the project for the entire project duration. The depreciation costs are divided over three years. The total amount will be separated into three years, beginning from the month of deployment.

  INAF The equipment will be used only for the project for the entire project duration. INAF has a 36 month depreciation policy for equipment.

  UCAM The equipment will be used only for the project for the entire project duration. PCs and laptops are costed as consumables and or not depreciated.

- ULEI, MPIA, INAF, and UCAM all have budgets for other goods and services. These include publication costs, and audit costs where mandatory, as detailed below.
- At ULEI € 10 000 is reserved for support of the workshops and setting up a web portal. In addition € 10 000 is reserved for auditing costs.
- NOTE: As audit costs were not listed in the original proposal for MPIA and UCAM, the travel costs for each of these partners has been lowered by 2000€. This amount was then added to 'other goods and services' in order to keep the total budget fixed. Thus travel costs for MPIA will be € 34 250 and for UCAM € 31 750.

For completeness the tables below contain the justifications for the partners ULEI, MPIA, INAF, and UCAM.

Participant no. 1 (ULEI)	**Cost (EUR)**	**Justification**
Travel	47250	1 project meeting in Europe for 1 participant; 6 project meetings in Europe for 2 participants; 2 project meetings outside Europe for 2 participants; 1 visit from MONA for 1 participant; 1 visit from NYU for 1 participant; 2 visits to ULEI by 1 expert; 1 extended working visit to MONA for 1 participant.
Equipment	5500	Laptop and high-end PC.
Other goods and services	23000	Publication costs 3000€, auditing costs 10 000€, workshop and web portal costs 10 000€.
Total	75750	

Participant no. 2 (MPIA)	**Cost (EUR)**	**Justification**
Travel	34250	1 project meeting in Europe for 1 participant; 4 project meetings in Europe for 2 participants; 2 project meetings in Europe for 3 participants; 2 project meetings outside Europe for 2 participants; 1 visit from MONA for 1 participant; 1 visit from NYU for 1 participant; 2 visits to MPIA by 1 expert. Overall travel costs lowered by € 2000 as noted above.
Equipment	5500	Laptop and high-end PC.
Other goods and services	5000	Publication costs 3000€, audit costs 2000€.
Total	44750	

Participant no. 3 (INAF)	**Cost (EUR)**	**Justification**
Travel	33750	1 project meeting in Europe for 1 participant; 6 project meetings in Europe for 2 participants; 2 project meetings outside Europe for 2 participants; 1 visit from MONA for 1 participant; 1 visit from NYU for 1 participant; 2 visits to INAF by 1 expert.
Equipment	5500	Laptop and high-end PC.
Other goods and services	3000	Publication costs.
Total	42250	

Participant no. 4 (UCAM)	**Cost (EUR)**	**Justification**
Travel	31750	1 project meeting in Europe for 1 participant; 6 project meetings in Europe for 2 participants; 2 project meetings outside Europe for 2 participants; 1 visit from MONA for 1 participant; 1 visit from NYU for 1 participant; 2 visits to UCAM by 1 expert. Overall travel costs lowered by € 2000 as noted above.
Equipment	5500	Laptop and high-end PC.
Other goods and services	5000	Publication costs 3000€, audit costs 2000€.
Total	42250	

Bibliography

[1] Anders, F., Khalatyan, A., Chiappini, C., et al. 2019, A&A, 628, A94 23

[2] Antoja, T., Helmi, A., Romero-Gómez, M., et al. 2018, Nature, 561, 360 4, 12

[3] Astropy Collaboration, Price-Whelan, A. M., Sipőcz, B. M., et al. 2018, AJ, 156, 123 26, 45

[4] Belokurov, V., Erkal, D., Evans, N. W., Koposov, S. E., & Deason, A. J. 2018, MNRAS, 478, 611 4

[5] Belokurov, V., Sanders, J. L., Fattahi, A., et al. 2019, arXiv e-prints, arXiv:1909.04679 4

[6] Bonaca, A., Hogg, D. W., Price-Whelan, A. M., & Conroy, C. 2019, ApJ, 880, 38 13

[7] Bonaca, A., Hogg, D. W., Price-Whelan, A. M., & Conroy, C. 2019, ApJ, 880, 38 45

[8] Boubert, D. & Everall, A. 2020, submitted 18, 33

[9] Boubert, D., Everall, A., & Holl, B. 2020, submitted 18, 33

[10] Bovy, J., Rix, H.-W., Liu, C., et al. 2012, ApJ, 753, 148 33

[11] Breivik, K., Price-Whelan, A. M., D'Orazio, D. J., et al. 2019, arXiv e-prints, arXiv:1903.05094 13

[12] Clementini, G., Ripepi, V., Molinaro, R., et al. 2019, A&A, 622, A60 59

[13] Eilers, A.-C., Hogg, D. W., Rix, H.-W., & Ness, M. K. 2019, ApJ, 871, 120 12, 33

[14] El-Badry, K., Rix, H.-W., Tian, H., Duchêne, G., & Moe, M. 2019, MNRAS, 489, 5822 13, 33

[15] Erkal, D., Belokurov, V., Laporte, C. F. P., et al. 2019, MNRAS, 487, 2685 12

[16] Everall, A. & Das, P. 2020, MNRAS, 493, 2042 33

[17] Freeman, K. & Bland-Hawthorn, J. 2002, ARA&A, 40, 487 5

[18] Gaia Collaboration, Katz, D., Antoja, T., et al. 2018, A&A, 616, A11 12, 59

[19] Gaia Collaboration, Prusti, T., de Bruijne, J. H. J., et al. 2016, A&A, 595, A1 4, 6, 20

[20] Gelman, A., Carlin, J., Stern, H., et al. 2013, Bayesian Data Analysis, Third Edition, Texts in Statistical Science (Chapman & Hall/CRC) 3

[21] Green, G. M., Schlafly, E. F., Finkbeiner, D., et al. 2018, MNRAS, 478, 651 9

[22] Helmi, A., Babusiaux, C., Koppelman, H. H., et al. 2018, Nature, 563, 85 4

[23] Holl, B., Audard, M., Nienartowicz, K., et al. 2018, A&A, 618, A30 10

[24] Holl, B., Audard, M., Nienartowicz, K., et al. 2018, A&A, 618, A30 8

[25] Ibata, R., Thomas, G., Famaey, B., et al. 2020, arXiv e-prints, arXiv:2002.01488 13

[26] Inno, L., Matsunaga, N., Bono, G., et al. 2013, ApJ, 764, 84 59

[27] Kochanek, C. S., Shappee, B. J., Stanek, K. Z., et al. 2017, PASP, 129, 104502 9

[28] Kollmeier, J. A., Zasowski, G., Rix, H.-W., et al. 2017, arXiv e-prints, arXiv:1711.03234 60

[29] Luri, X., Brown, A. G. A., Sarro, L. M., et al. 2018, A&A, 616, A9 4, 8

[30] Poggio, E., Drimmel, R., Andrae, R., et al. 2020, Nature Astronomy, 2397 59

[31] Poggio, E., Drimmel, R., Lattanzi, M. G., et al. 2018, MNRAS, 481, L21 59

[32] Price-Whelan, A. M. & Bonaca, A. 2018, ApJ, 863, L20 33

[33] Price-Whelan, A. M. & Bonaca, A. 2018, ApJ, 863, L20 45

[34] Read, J. I. 2014, Journal of Physics G Nuclear Physics, 41, 063101 58

[35] Skowron, D. M., Skowron, J., Mróz, P., et al. 2019, Science, 365, 478 9

[36] Smart, R. L. & Nicastro, L. 2014, A&A, 570, A87 9, 41

[37] Tremblay, P.-E., Fontaine, G., Fusillo, N. P. G., et al. 2019, Nature, 565, 202 4

4 Members of the consortium

4.1 Participants (applicants)

4.1.1 Universiteit Leiden

General description Leiden Observatory is the astronomical institute of the Faculty of Science of Leiden University. Established in 1633, it is the oldest university observatory in operation today, with a very rich tradition. Leiden Observatory carries out world class research in the formation of structures in the universe and the origin and evolution of galaxies, the detection and characterization of exoplanets, and the formation of stars and planetary systems. The institute consists of about 35 faculty and adjunct faculty, 50 postdoctoral researchers, 50 MSc and 80 PhD students, and 30 support staff. Leiden Observatory offers an excellent educational programme at the Bachelor's and Master's levels and a renowned PhD programme. Within the Faculty of Science, the institute closely collaborates with the Leiden Institute of Physics, the Mathematical Institute and the Leiden Institute of Advanced Computer Science.

Key persons Dr. Anthony Brown (male) is a faculty member of Leiden Observatory. He has published over 75 papers in the peer-reviewed literature, and supervised 9 PhD students and several postdocs. He has been deeply involved with the Gaia mission since 1997, having contributed to the optimization of the mission, the preparations for the data processing, and the implementation of several algorithms in the DPAC data processing pipelines. He currently serves as the chair of the Gaia Data Processing and Analysis Consortium and is a member of the Gaia Science Team. The efforts described in this proposal will very much benefit from the extensive expertise of Brown on both the Gaia spacecraft and its instruments as well as the DPAC processing pipelines. The close connections to ESA and DPAC through Brown's role in the latter will facilitate the communications regarding details of the selection function, the integration of the results in the Gaia archive, and the transfer of expertise to DPAC.

Dr. Zuzanna Kostrzewa-Rutkowska (female) is a post-doctoral fellow at Leiden Observatory. She works with Brown on the crowded field photometric data processing for the Gaia mission. GaiaUnlimited can benefit from her expertise when it comes to developing the crowded field aspects of selection functions.

Dr. Elena Sellentin (female) is an assistant professor at Leiden Observatory and a world-class expert on advanced statistical analysis of astronomical data. Brown and the GaiaUnlimited team will benefit from her expertise in the mathematical and statistical development of the Gaia selection function.

Relevant publications

- Gaia Collaboration, **Brown**, et al., 2018, Gaia Data Release 2. Summary of the contents and survey properties, A&A 616, A1
- Snellen & **Brown**, 2018, The mass of the young planet Beta Pictoris b through the astrometric motion of its host star, Nature Astronomy 2, 883
- Zari, Hashemi, **Brown**, et al., 2018, 3D mapping of young stars in the solar neighbourhood with Gaia DR2, A&A 620, A172

- Luri, **Brown**, et al., 2018, Gaia Data Release 2. Using Gaia parallaxes, A&A 616, A9
- Helmi, Babusiaux, Koppelman, Massari, Veljanoski, **Brown**, 2018, The merger that led to the formation of the Milky Way's inner stellar halo and thick disk, Nature 563, 7729

Relevant previous projects Brown was the main contact point at ULEI for the following projects:

- FP6 ELSA Marie-Curie Research Training Network
- ESF GREAT Research Network Programme
- FP7 GREAT Marie-Curie Initial Training Network
- FP7 GENIUS Collaborative project (SPA.2013.2.1-01 Exploitation of space science and exploration data)
- BIG-SKY-EARTH Trans Domain COST Action

Significant infrastructure Leiden Observatory shares a building with the Lorentz Center, an international centre for workshops in the sciences which offers excellent facilities for hosting workshops, conferences, and schools. In particular, office space is available for bringing researchers together for extended periods.

4.1.2 Max-Planck-Institut für Astronomie (MPIA)

General description MPIA is a large and leading research institution, with three departments 'Galaxies', 'Star and Planet Formation', and (as of 1.3.2020) 'Atmosphere Physics of Exoplanets'. Its research approaches encompass observation, instrumentation development, and theoretical modelling. MPIA has played a key role in Gaia (having led the DPAC CU8, source astrophysical classification and parametrization, development), and Gaia data are used across the institute: MPIA researches have over 150 refereed publications involving Gaia or Gaia data, since Gaia DR1. MPIA has a long-term staff of 25 members, and about 60 post-docs and 60 PhD students.

Key persons Prof. Hans-Walter Rix (male) is director of the 'Galaxies and Cosmology' department at MPIA, and adjunct faculty at Heidelberg University. He has published over 500 peer-refereed papers (about 150 of them on the Milky Way or Gaia), and over the last 20 years has supervised 15 students, and 30 post-docs. He has held an ERC Advanced Grant on studying the dynamical structure of the Galactic disk, and is now the project scientist for the SDSS-V survey: the only survey to provide consistent all-sky, high-quality spectra of stars to complement Gaia. Rix is co-organizer of the Gaia Sprint workshops.

Dr. Morgan Fouesneau (male) is a postdoctoral fellow at MPIA. He is a member of DPAC and makes major contributions to the astrophysical classification and parametrization pipeline, both the processing algorithms and the scientific validation of the results. He is working on hybrid catalogues that combine Gaia with other surveys in order to produce more accurate source astrophysical parameters. His experience with hybrid catalogues will be essential to the efforts to produce combined selection functions. Fouesneau leads the classification pipeline for the 4MOST survey which will complement Gaia with optical spectra in the southern hemisphere and he was a co-organizer of the Gaia DR2 Exploration Lab.

Relevant publications

- Bovy & **Rix**, 2013, ApJ, 779, 115 (290 cit) A Direct Dynamical Measurement of the Milky Way's Disk Surface Density Profile, Disk Scale Length, and Dark Matter Profile
- **Rix** & Bovy, 2013, A&A Review, 21, 61 (165 cit.) Mapping and modelling the Galactic disk
- El-Badry & **Rix** , 2019, MNRAS, 489, 5822 (13 cit.) Gaia discovery of an equal-mass 'twin' binary population reaching 1000 + au separations
- Rezaei Kh, Bailer-Jones, Schlafly & **Fouesneau**, 2018, A&A, 616, id.A44. (cit. 4) Three-dimensional dust mapping in the Orion complex, combining Gaia-TGAS, 2MASS, and WISE.
- Poggio, et. al. (incl **Fouesneau**), 2020, 'Evidence of a dynamically evolving Galactic warp', Nature Astronomy (doi:10.1038/s41550-020-1017-3)

Relevant previous projects

- FP7-PEOPLE-2007-1-1-ITN
- ERC-2012-ADG-20120216
- FP7-INFRASTRUCTURES-2012-1

Significant infrastructure MPIA employs >100 PhD and PostDoc students: for them it offers a rich and structured research environment, a mentoring system and career development. MPIA has privileged access to MPG supercomputing resources, and has extensive in-house computing. MPIA can host (also extended) workshop, from days to weeks, and from 5 to 100 persons. MPIA has extensive access to facilities for Gaia follow-up observations.

4.1.3 INAF - Osservatorio Astrofisico di Torino

General description The Istituto Nazionale di Astrofisica (INAF) promotes, coordinates, and carries out, within the context of programs of the European Union and other International Entities, research activities in the fields of Astronomy, Radioastronomy, Space Astrophysics, and Cosmic Physics, in collaboration with Italian Universities as well as other private and public, national and international research organizations. In particular, INAF coordinates research activities carried out at national facilities (the Telescopio Nazionale Galileo, TNG, in the Canary Islands, and the Large Binocular Telescope, LBT, on Mt. Graham, Arizona) and at 18 research institutes and astronomical observatories on Italian soil. The proposed project will be undertaken at the Osservatorio Astrofisico di Torino (OATo), one of the INAF institutes.

Key persons Dr. Ronald Drimmel (male) is a research astronomer (staff) of INAF whose research interest is primarily the structure and dynamics of the Milky Way, especially the investigation of non-axisymmetric structures such as the spiral arms and the warp. While currently he works with Gaia data, he has analysed infrared and far-infrared all-sky data from the COBE satellite to derive a model of the distribution of the stars and interstellar dust in the Milky Way on large scales. Drimmel has been a key member of the Gaia Data Processing and Analysis Consortium (DPAC), serving as it's Deputy Chair from 2006 to 2012. Currently he is contributing to the construction of an all-sky map of the total Galactic extinction. Drimmel is also on the Gaia Payload Experts group, responsible for monitoring and evaluating the performance of the Gaia instrument.

Dr. Richard Smart (male) has been active in Gaia since it's conception. He was the provider of the Initial Gaia Source List [36] and is member of the Gaia Ultra Cool Dwarf and Gaia Ground Based Tracking work package and is leading the effort to produce the Gaia Catalogue of Nearby Stars that contains all stars to M9 within 100 pc of the Sun. Smart was the coordinator of the EU FP7 RISE program Interpretation and Parameterization of Extremely COOL objects (IPERCOOL, 247593), the Italian coordinator for the GENIUS program and has managed 5 postdocs at the Observatory.

Dr. Alessandro Spagna (male) is a senior research astronomer (staff) of INAF. His main scientific interests are in the field of ground-based and space astrometry, in the construction of wide field surveys for the study of the Galactic stellar populations and the chemo-dynamical evolution of the Milky Way. In particular his research activity focuses on the formation of the Galactic thick disc and stellar halo. He has been also involved in various projects for the construction of all-sky catalogues, such as Hipparcos, Guide Star Catalog II, and Gaia.

Dr. Eloisa Poggio (female) is a postdoctoral fellow at INAF and member of DPAC. Her main scientific interests include the structure and dynamics of the disk of the Milky Way, with particular focus on non-axisymmetric features such as the warp and spiral arms. She is currently developing a model for the Galaxy, aimed at estimating Galactic parameters via statistical bayesian inference including the survey selection function.

Relevant publications

- Poggio, **Drimmel**, et. al., 2020, Evidence of a dynamically evolving Galactic warp, Nature Astronomy (doi:10.1038/s41550-020-1017-3)

- Poggio, **Drimmel**, et. al., 2018, The Galactic warp revealed by Gaia DR2 kinematics, MNRAS Letters 481, L21

- Gaia Collaboration, Katz, Antoja, Romero-Gómez, **Drimmel**, et al., 2018 Gaia Data Release 2. Mapping the Milky Way disc kinematics A&A 616, A11

- **Drimmel** & Spergel, 2001, Three-dimensional Structure of the Milky Way Disk: The Distribution of Stars and Dust beyond 0.35 R_{solar} ApJ, 556, 181

- **Drimmel**, 2000, Evidence for a two-armed spiral in the Milky Way A&A, 358, L13

Relevant previous projects Drimmel was the main contact point at INAF for the following past projects:

- FP6 ELSA Marie-Curie Research Training Network

Drimmel also participated in the following EU funded projects:

- ESF GREAT Research Network Programme
- FP7 GREAT Marie-Curie Initial Training Network
- FP7 GENIUS Collaborative project (SPA.2013.2.1-01 Exploitation of space science and exploration data)

Significant infrastructure In the context of this proposal, the most relevant facilities that INAF participants have access to is the Marconi supercomputer at the CINECA, a national high-performance computing centre. Marconi is classified in Top500 list among the most powerful supercomputer: rank 12 in November 2016, and rank 19 in the November 2019 list. Also available is an Oracle Database Machine at Trieste Observatory.

4.1.4 Institute of Astronomy at the University of Cambridge (UCAM)

General description Institute of Astronomy (IoA) is one of the three Astronomy departments at the University of Cambridge, the other two being Cavendish Astrophysics Group and the Astrophysics Group at the Department of Applied Mathematics and Theoretical Physics. The IoA has 18 permanent staff members, some 100 postdoctoral fellows, approximately 50 graduate students and 25 support staff. The Institute also hosts the Cambridge Astronomical Survey Unit, the DPAC Coordination Unit 5 in charge of Photometric Processing and the Gaia-ESO survey. The Institute's research portfolio includes the full spectrum of the cutting edge Astrophysics topics, from the Early Universe to Exo-planets and instrument design and construction. The Institute played a uniquely active role in the scientific exploration of the Gaia DR1 and DR2 resulting in more than 200 publications by the IoA members (over the last 4 years only). The direct daily interface between the broad and active research body and CASU and CU5 makes UCAM an especially attractive environment for the proposed project.

Key persons Prof. Vasily Belokurov (male) is a Professor of Astronomy at the University of Cambridge. He is a survey scientist, whose expertise is in data mining, galactic dynamics and machine learning. The broad theme of his research is galaxy formation and evolution, with a particular focus on the Milky Way halo, low-mass galaxies and local constraints on Dark Matter. Belokurov has supervised more than 20 graduate students and worked closely with an appreciable number of postdocs. He has published more than 250 papers in refereed journals. It is worth mentioning separately the papers that present the analysis based on the data from the ESA's Gaia mission. Gaia Data Release 1 was delivered in 2016, while the Gaia DR2 arrived less than 2 year ago. So far, Belokurov has published >40 papers in refereed journals, relying on either Gaia DR1 or DR2 data.

Andrew Everall (male) is a PhD student at the Institute of Astronomy working with supervisors Prof. Vasily Belokurov and Prof. Wyn Evans. His core focus is on constraining the Milky Way Dark Matter distribution and specifically the local DM density. He has worked significantly on selection functions for multi fibre spectrographs and has already performed a substantial amount of work on selection functions for Gaia DR2 and subsets therein with Dr. Douglas Boubert.

Relevant publications

- Grady, **Belokurov** & Evans, 2020, MNRAS, 492, 3128, Age demographics of the Milky Way disc and bulge
- **Everall** & Das, 2020, MNRAS, 493, 2042, seestar: Selection functions for spectroscopic surveys of the Milky Way
- Erkal, **Belokurov**, et al., 2019, MNRAS, 487, 2685, The total mass of the Large Magellanic Cloud from its perturbation on the Orphan stream
- Necib, Lisanti, **Belokurov**, 2019, ApJ, 874, 3, Inferred Evidence for Dark Matter Kinematic Substructure with SDSS-Gaia
- **Belokurov** et al., 2018, MNRAS, 478, 611, Co-formation of the disc and the stellar halo

Relevant previous projects

- Starting Grant (StG), PE9, ERC-2012-StG_20111012
- Gaia-ESO survey. A large European survey which was specifically aimed at collecting spectroscopic data complementary to the Gaia mission.

Significant infrastructure

4.1.5 Monash University

General description Monash University is a world-class teaching and research institution located in Melbourne, Australia. The School of Physics & Astronomy at Monash University includes researchers spanning observational and theoretical astrophysics, with focussed groups on exoplanets, binary stars, stellar astrophysics, galactic structure, gravitational waves, and more. Statistical inference and data analysis are common to all of these research groups, making Monash University an excellent place to host researchers who are developing the Gaia selection function and applying it to key areas in astrophysics.

Key persons Dr. Andrew Casey (male) is a faculty member at Monash University and an Australian Research Council DECRA Fellow. He is currently six years post-PhD and has an extensive track record of high-quality peer-reviewed publications: >100 publications that have accrued 5,988 citations, yielding a h-index of 34 and a m-index of 3.4. Casey is a co-organiser of the Gaia Sprints workshops (2016–2019, inclusive), which have helped strengthen the astrophysics community who are working with Gaia data.

Relevant publications

- **Casey**, et al., 2019, A Data-driven Model of Nucleosynthesis with Chemical Tagging in a Lower-dimensional Latent Space, 2019, ApJ, 887, 73
- **Casey**, et al., 2019, Tidal Interactions between Binary Stars Can Drive Lithium Production in Low-mass Red Giants, 2019, ApJ, 880, 125
- Simpson, Martell, Da Costa, **Casey**, et al., 2019, The GALAH survey: co-orbiting stars and chemical tagging, MNRAS, 482, 5302
- Williams, Belokurov, **Casey**, Evans, 2017, On the run: mapping the escape speed across the Galaxy with SDSS, MNRAS, 468, 2359
- **Casey** et al., 2017, The RAVE-on Catalog of Stellar Atmospheric Parameters and Chemical Abundances for Chemo-dynamic Studies in the Gaia Era, ApJ, 840, 59

Relevant previous projects

- Gaia-ESO survey. A large European survey which was specifically aimed at collecting spectroscopic data complementary to the Gaia mission.

Significant infrastructure As of 2020 Monash University joins the Australian Research Council Centre of Excellence in Astrophysics in 3-Dimensions (ASTRO 3D), a \$30 M AUD (18.3 M Euro) initiative for astrophysics research in Australia. This initiative includes researchers working on the formation and evolution of the Milky Way galaxy, single and binary star evolution, and other key science goals for the Gaia mission. The Australian-led GALAH survey is one of the key projects of ASTRO 3D, with the goal to obtain high-resolution optical spectra for up to $\sim 10^6$ stars. Currently the GALAH/Gaia sub-sample represents the highest quality set of stars with precise astrometry and chemical abundance information. This provides a unique opportunity to study the chemo-dynamical history of the Milky Way in exquisite detail, yet the Gaia and GALAH samples are both subject to selection function effects. Monash University is the ideal place for researching binary population inference with the Gaia selection function, and the complement of the Gaia selection functions with other surveys (e.g. GALAH).

4.1.6 New York University

General description The Department of Physics at New York University hosts 36 tenure and tenure-track professors, two clinical faculty members, a few dozen research scientists and postdoctoral fellows, and over 80 graduate and 90 undergraduate students. The majority of the faculty are engaged in cutting-edge research in astrophysics and cosmology, particle and astro-particle experimental and theoretical physics, hard and soft condensed matter physics, biophysics, fluid dynamics, and applied mathematical physics.

Key persons Prof. David W. Hogg (male) is Professor of Physics and Data Science in the Center for Cosmology and Particle Physics in the Department of Physics. He is also Group Leader for the Astronomical Data Group in the Center for Computational Astrophysics of the Flatiron Institute, a staff member in the Centre for Data Science at New York University, and he has an affiliation with the Max-Planck-Institut für Astronomie. His main research interests are in observational cosmology, especially approaches that use galaxies to infer the physical properties of the Universe. He also works on the properties and kinematics of stars in the Galaxy, and the measurement and discovery of planets around other stars. In all areas, he is interested in developing the engineering systems that make these projects possible. He has published 250 peer-reviewed publications and has supervised or co-supervised 11 PhD students and numerous postdoctoral researchers. He is the chair of the organising committee for all Gaia Sprints held to date.

Dr. Adrian Price-Whelan (male) is a postdoctoral fellow at the Center for Computational Astrophysics at the Flatiron Institute. His main research centres on the structure and formation of the Milky Way. His analysis of the GD-1 stream in the Gaia DR2 data led to the discovery of gaps along the stream which could point to the presence of dark matter substructures in the halo of the Milky Way[7, 33]. This work was also one of the triggers of the discussions on the need for a Gaia selection function. Price-Whelan is a major contributor to Astropy [3], which is the eco-system in which the GaiaUnlimited tools are likely to be embedded. Price-Whelan is a co-organizer of the Gaia Sprint workshops.

Price-Whelan will contribute to GaiaUnlimited in an advisory capacity. He will contribute his insights on survey selection functions and how to implement these in code, and he will act as a liaison to the Astropy project (see section 2.2.1 and WP4).

Relevant publications

- Bonaca, **Hogg**, **Price-Whelan**, et al., 2019, The Spur and the Gap in GD-1: Dynamical Evidence for a Dark Substructure in the Milky Way Halo, ApJ, 880, 38
- Anderson, **Hogg**, et al., 2018, Improving Gaia Parallax Precision with a Data-driven Model of Stars, AJ, 156, 145
- Bonaca & **Hogg**, 2018, The Information Content in Cold Stellar Streams, ApJ, 867, 101
- Eilers, **Hogg**, Rix, & Ness, 2019, The Circular Velocity Curve of the Milky Way from 5 to 25 kpc, ApJ, 871, 120
- Rezaei Kh, Bailer-Jones, **Hogg**, et al., 2018, Detection of the Milky Way spiral arms in dust from 3D mapping, A&A, 618, 168

Relevant previous projects

- NSF Astronomy and Astrophysics Research Grant AST-1517237, New Probabilistic Methods for Observational Cosmology

- NSF Cyber-Enabled Discovery Type I Grant (IIS-1124794), A Unified Probabilistic Model of Astronomical Imaging
- NASA HST Archival Research Grant (AR-13250), Probabilistic Self-calibration of the WFC3 IR Channel
- NSF AST-0908357, Dynamical models from kinematic data: The Milky Way Disk and Halo

Significant infrastructure New York University and the Flatiron Institute are both located in New York City. Together these institutes host hundreds of visiting researchers every year and the Flatiron Institute has been host for multiple Gaia Sprints in the past, offering a unique and collaborative space that cultivates a lively atmosphere for research. Hogg has strong collaborations with astrophysicists and data scientists from all research institutions in the New York area.

4.2 Third parties involved in the project (third party resources)

No third parties are involved.

5 Ethics and Security

5.1 Ethics

GaiaUnlimited is a project devoted to research and development that will lead to data products and tools to apply selection functions to the analysis of data collected by a space astrophysics mission and other surveys. Hence, no ethics issues are expected to arise in the project. This is indicated in the form filled out online in the Funding & tender opportunities portal. Specific clarifications:

- Research and development in GaiaUnlimited is conducted on publicly available data from astronomical surveys. We will also make use of information from the Gaia Data Processing and Analysis Consortium (DPAC) to which the coordinator and several participants have access (as members of DPAC). For the participants who are not members of DPAC but need access to its non-public information, a standard non-disclosure agreement (NDA) already exists which the participants in question will sign. In particular we stress:
  - the research, development, and implementation of tools proposed here do not involve human participants, human tissue, human embryos/foetuses, cells, or animals.
  - no personal data is collected or processed for the purpose of the effort proposed.
- Two third countries are involved, Australia and the United States of America.
  - They will participate in the GaiaUnlimited research, development, and implementation of tools. As such they will be working with the publicly available data mentioned above and any source code written by these partners (as part of the GaiaUnlimited effort) will become part of the GaiaUnlimited tools that will be made publicly available.
  - No personal data will be imported from or exported to the non-EU partners.
  - No activities will be undertaken by the non-EU partners that are prohibited in the EU.
  - Any data exchange with the non-EU partners will only involve publicly available data to which the non-EU partners have access. In case DPAC information needs to be exchanged the NDA referred to above will also be applied to the non-EU partners.
- The purpose of GaiaUnlimited is to advance astronomical research by enabling a much more advanced use of Gaia and other survey data than is possible today. Hence the focus of the efforts proposed here is exclusively civilian.

5.2 Security

- Activities or results raising security issues: NO
- ‘EU-classified information’ as background or results: NO

Table 5.2a: List of work packages

Work package number	**Work package title**	**Lead participant no.**	**Lead participant name**	**Person-months**	**Start month**	**End month**
WP1	Management	1	ULEI	14	1	42
WP2	Definition of the selection function	2	MPIA	7	1	12
WP3	Research and develop the Gaia selection function	4	UCAM	41	1	42
WP4	Practical implementation and dissemination of the Gaia selection function	1	ULEI	45	1	42
WP5	Selection function for combinations of Gaia and other surveys	2	MPIA	42	1	42
WP6	Science applications	3	INAF	56	1	42
	TOTAL			205		

Work package description

Table 5.2b: Description of work packages

Work package number	**WP1**	**Lead beneficiary**		ULEI	
Work package title	**Management**				
Participant number	**1**	2	3	4	5	6
Short name	**ULEI**	MPIA	INAF	UCAM	MONA	NYU
Person-months	**8**	2	2	2	0	0
Start month	1		**End month**		42	

Objectives

This package provides the administrative management and overall scientific coordination of GaiaUnlimited following the structure and procedures described in section 3.2.

Description of work

This work package concerns the global administrative management and coordination tasks inside the consortium, including progress monitoring, financial management, intellectual property management, innovation management (section 3.2.2), project documentation, and communication and dissemination activities. It also includes the coordination with institutions and bodies relevant for the development of the Gaia archive like ESA, DPAC and MW-Gaia, as well as the representation of GaiaUnlimited in meetings or committees related to this coordination.

Task T1.1: Administrative management (M1-M42) *Leader:* ***ULEI****. Contributors: ULEI*
Overall coordination and administrative management as described above. The coordinator will be aided in his tasks by the Leiden Observatory project support staff.

6 ULEI person-months

Task T1.2: Overall project management and coordination (M1-M42) *Leader:* ***ULEI****.*
Contributors: MPIA, INAF, UCAM
Overall scientific and technical control and coordination of the project, ensuring that the tasks are properly carried out and remain on schedule, and that the objectives of GaiaUnlimited are fulfilled. This is done by the four main partners through the GaiaUnlimited Executive Board. As stated in section 3.2, the partners NYU and MONA will have a standing invitation to meetings of the Executive Board in order to keep communications efficient.

2 MPIA + 2 INAF + 2 UCAM person-months

Task T1.3: Communications and dissemination (M1-M42) *Leader:* ***ULEI****.*
Contributors: ULEI
Ensure that results from GaiaUnlimited are disseminated to the astronomical community through the appropriate academic channels (publications, conference presentations, community workshops). Engage this community by making available prototype implementations of the tools and provide channels for feedback. Organize three community workshops to provide training on the use of the selection function implementations and to collect feedback on the GaiaUnlimited tools.

2 ULEI person-months

Deliverables

The deliverables of this work package are the data management plan, the web-portal for access to GaiaUnlimited tools and data, and the organisation of the GaiaUnlimited plenary meetings (section 3.2.3) and community workshops.

D1.1 Kick-off meeting (plenary). The minutes of meeting will be delivered as a record that the meeting took place. **(M1)**

D1.2 Data management plan. **(M6)**

D1.3 Community workshop no. 1 and mid-term meeting (plenary). **(M21)**

D1.4 Report no. 1 summarizing community feedback on GaiaUnlimited tools. **(M22)**

D1.5 Community workshop no. 2. **(M32)**

D1.6 Report no. 2 summarizing community feedback on GaiaUnlimited tools. **(M33)**

D1.7 Community workshop no. 3 and closing meeting (plenary). **(M41)**

Work package number	**WP2**	**Lead beneficiary**	MPG			
Work package title	**Definition of the selection function**					
Participant number	1	**2**	3	4	5	6
Short name	ULEI	**MPIA**	INAF	UCAM	MONA	NYU
Person-months	1	**2**	1	1	1	1
Start month	1		**End month**	12		

Objectives

The objective of this work package is to research and implement a precise mathematical formulation of the concept of a survey selection function. The results will be written up as a scientific paper (to be published in the open access peer-reviewed literature) that will guide the rest of the work to be done within GaiaUnlimited. The mathematical formulation of the selection function will account for the following aspects:

- Applicability to arbitrary astronomical sky surveys.
- Selection functions for combinations of surveys can be described within the same formalism.
- Accommodate multiple selection layers (e.g., the selection function intrinsic to a survey combined with additional selections imposed by the user of the data).
- Unknowable aspects of the selection function should be included. For example, it is not guaranteed that all sources of data loss along the detection-telemetry-processing chain for Gaia can be traced, but this can be approximated through knowing which repeated measurements were included in the catalogue when expected.

Description of work

Task T2.1: Work package management (M1-M12) *Leader:* ***MPIA****. Contributors: MPIA*
Management and coordination of the specific work on defining the selection function and coordination of writing the paper corresponding to deliverable D2.1. This includes planning the work to be done, assigning sub-tasks and organizing the necessary meetings to discuss progress.

1 MPIA person-month

Task T2.2: R&D formulation selection function (M1-M12) *Leader:* ***MPIA****.*
Contributors: All other
Research and implement the mathematical formulation of survey selection functions.

1 MPIA + 1 ULEI + 1 INAF + 1 UCAM + 1 MONA + 1 NYU person-months

Role of partners

MPIA will lead Task T2.1.

ULEI, INAF, UCAM, NYU, MONA will contribute to the research and implementation of the survey selection function mathematical formulation and will contribute to writing the paper corresponding to deliverable D2.1.

Deliverables

D2.1 Document on the mathematical formulation of survey selection functions (to be submitted to open access peer-reviewed journal). **(M12)**

Work package number	**WP3**	**Lead beneficiary**		UCAM		
Work package title	**Research and develop the Gaia selection function**					
Participant number	1	2	3	**4**	5	6
Short name	ULEI	MPIA	INAF	**UCAM**	MONA	NYU
Person-months	2	6	11	**19**	2	1
Start month	1		**End month**		42	

Objectives

The objective of this work package is to research and develop in detail the description and modelling of the Gaia survey selection function. The framework for the efforts in this work package is provided by deliverable D2.1.

1. Develop the overall survey selection function, focusing on the Gaia astrometric, photometric, and spectroscopic surveys and combinations thereof. In addition a specialized selection function for specific subsets of the Gaia survey will be developed. To keep the scope of the work realistic we will focus on the case of binary stars and Cepheids (variable stars) in the Gaia catalogue. The development of further specialized Gaia selection functions is left to the community who can use the methods and tools developed in this work package.

2. Investigate to what extent detailed information is needed on the Gaia pointing history, its on-board detection algorithm, and the data losses introduced along the various steps from on-board measurement to the final data products. Can the selection function be reverse engineered from publicly available information?

3. Interface with DPAC to gain a deeper understanding of the many ingredients of the Gaia selection function and collect the information necessary to describe or model spacecraft or data processing pipeline decisions that affect the selection function.

Description of work

Task T3.1: Work package management (M1-M42) *Leader:* ***UCAM****. Contributors: UCAM*
Management and coordination of the research and development of the selection function and coordination of writing the paper corresponding to deliverables D3.8, D3.10, and D3.12. This includes planning the work to be done, assigning sub-tasks and organizing the necessary meetings to discuss progress.
1 UCAM person-month

Task T3.2: Collect information from DPAC (M7-M42) *Leader:* ***ULEI****. Contributors: ULEI*
Interface with DPAC to collect the necessary information to describe or model the events/decisions on board Gaia and the decisions in the data processing pipelines that affect the selection function.
2 ULEI person-months

Task T3.3: R&D top level Gaia selection functions (M7-M42) *Leader:* ***UCAM****. Contributors: INAF, NYU*
Research and development of the top-level Gaia selection function which describes the probability that a source enters (combinations of) the Gaia astrometric, photometric or radial velocity surveys.
10 UCAM + 5 INAF + 1 NYU person-months

Task T3.4: R&D spectroscopic Gaia selection functions (M7-M42) *Leader:* ***MPIA****. Contributors: UCAM, MONA*
Research and development of the Gaia spectroscopic selection function. This refers to the probability that a BP/RP and/or and RVS spectrum was measured for a source in the Gaia survey.
6 MPIA + 3 UCAM + 2 MONA person-months

Task T3.5: R&D specialized Gaia selection functions (M7-M42) *Leader:* ***INAF****. Contributors: INAF, UCAM*
Research and develop the methods needed to create selection functions for specific subsets of the Gaia survey, binary stars and Cepheid variables.
6 INAF + 5 UCAMperson-months

Role of partners

UCAM will lead Tasks T3.1 and T3.3.

ULEI will lead Task T3.2.

MPIA will lead Task T3.3.

INAF will lead Task T3.5

All partners will contribute to the research and development of the Gaia selection functions and to the writing of the resulting papers.

Deliverables

D3.1 Report with detailed scoping of the research and development effort. **(M12)**

D3.2 Documentation of preliminary top level Gaia selection functions. **(M19)**

D3.3 Prototype software modules and data for the preliminary top level Gaia selection functions implementation in Prototype V1. **(M19)**

D3.4 Documentation of preliminary spectroscopic Gaia selection functions. **(M30)**

D3.5 Prototype software modules and data for the preliminary spectroscopic Gaia selection function implementation in Prototype V2. **(M30)**

D3.6 Documentation of preliminary specialized Gaia selection functions. **(M30)**

D3.7 Prototype software modules and data for the preliminary specialized Gaia selection function implementation in Prototype V2. **(M30)**

D3.8 Documentation of the top level Gaia selection functions (also to be submitted to as paper to a peer-reviewed journal). **(M40)**

D3.9 Software modules and data for the final top level Gaia selection functions implementation. **(M40)**

D3.10 Documentation of the spectroscopic Gaia selection functions (also to be submitted to as paper to a peer-reviewed journal). **(M40)**

D3.11 Software modules and data for the final spectroscopic Gaia selection functions implementation. **(M40)**

D3.12 Documentation of the specialized Gaia selection functions (also to be submitted to as paper to a peer-reviewed journal). **(M40)**

D3.13 Software modules and data for the final specialized Gaia selection functions implementation. **(M40)**

Work package number	**WP4**	**Lead beneficiary**		ULEI		
Work package title	**Practical implementation and dissemination of the Gaia selection function**					
Participant number	**1**	2	3	4	5	6
Short name	**ULEI**	MPIA	INAF	UCAM	MONA	NYU
Person-months	**23**	12	6	2	1	1
Start month	1		**End month**		42	

Objectives

This work package provides the practical implementation of the Gaia selection functions in terms of data and associated open source computer applications. Dissemination of these results through the ESA Gaia archive and code hosting web-sites is also part of this work package. The detailed objectives are:

1. Implement the selection function as defined and developed in detail in work packages 2, 3 and 5 in the form of open source computer applications and associated numerical data (e.g., in the form of tables). This corresponds to deliverables D3.2–D3.13.
2. Implement a tool that allows for layering user defined selections on the Gaia archive data on top of the survey selection functions.
3. Provide implementations of a number of combined selection functions for intersections of Gaia and selected other surveys. This corresponds to deliverables D5.2–D5.9.
4. Ensure that the implementation is embedded in a well-established ecosystem. We are targeting Astropy to which one of the participants (Price-Whelan, NYU) is a major contributor.
5. Identify code hosting options and make the Gaia selection function source code available publicly.
6. Agree with ESA/DPAC on the hosting of the numerical data associated with the Gaia selection function in the ESA archive ecosystem.
7. Set up a website which will be the portal to the GaiaUnlimited results. It will provide information, updates during the project lifetime, links to the sites hosting the code and data, and ensure EU funding visibility.
8. Disseminate the results through public prototype versions of the GaiaUnlimited data and tools, and by providing training during the community workshops on the use of the tools. Present the GaiaUnlimited tools at relevant conferences and workshops.

Description of work

Task T4.1: Work package management (M1-M42) *Leader: **ULEI**. Contributors: ULEI*
Management and coordination of the implementation and dissemination of the Gaia survey selection function. This includes setting up the code hosting and interfacing with ESA/DPAC on hosting the numerical data for the selection function in the Gaia archive. The organization of the dissemination activities and creation of the GaiaUnlimited web-portal are also part of this task.

3 ULEI person-months

Task T4.2: Implementation (M7-M42) *Leader: **ULEI**. Contributors: NYU*
Implement the Gaia selection functions (both general and specialized to specific Gaia sub-samples) in terms of open source software tools and associated numerical data. Test the implementation through applications to science cases. Port the implementation to the relevant code hosting site and transfer the numerical data to the ESA Gaia archive. To ensure the long term curation it is important to embed the GaiaUnlimited tools in an existing stable ecosystem. Astropy is an excellent candidate and in this task we will also include the effort to integrate the GaiaUnlimited tools therein, relying on the expertise from NYU.

14 ULEI + 1 NYU person-months

Task T4.3: Implementation of tools to include user selections (M7-M42) *Leader:* ***MPIA****. Contributors: MPIA*
Develop and implement tools to chain together the Gaia survey selections function with additional user imposed selection on the Gaia archive data. The tools should produce an overall effective selection function for the user-selected sample.
9 MPIA person-months

Task T4.4: Implementation of tools for combined selection functions (M7-M42) *Leader:* ***INAF****. Contributors: MPIA, UCAM, ULEI, MONA*
Implement the selection functions for combinations of Gaia and other surveys based on the methodology developed in work package 5. This effort will focus on the combination of Gaia and the two surveys mentioned in task T5.3. In particular for the GALAH survey an extended visit by the ULEI researcher to MONA is foreseen in order to profit from the access to GALAH data and the spectroscopic survey expertise.
3 MPIA + 6 INAF + 2 UCAM + 6 ULEI + 1 MONA person-months

Role of partners

ULEI will lead Task T4.1.

ULEI will lead Task T4.2.

MPIA will lead Task T4.3.

INAF will lead Task T4.4.

All partners will contribute to the implementation and testing of the tools developed in this work package. In particular the science applications pursued by the partners will constitute a strong test of the Gaia selection function implementation.

Deliverables

D4.1 First version of web-portal for GaiaUnlimited. **(M6)**

D4.2 Report with detailed scoping of the development and implementation effort. **(M12)**

D4.3 Prototype V1: open source and open data implementation of the top level Gaia selection functions. **(M20)**

D4.4 Second version of web-portal for GaiaUnlimited, ready for testing by community. **(M20)**

D4.5 Prototype V2: open source and open data implementation of the overall and specialized Gaia selection functions and selection functions for combinations of Gaia and other surveys. **(M31)**

D4.6 Third version of web-portal for GaiaUnlimited, including feedback from community. **(M31)**

D4.7 Open source and open data implementation of all selection functions developed in GaiaUnlimited. **(M42)**

D4.8 Final version of the Web-portal providing access to the Gaia selection function data products and tools delivered by GaiaUnlimited. **(M42)**

Work package number	**WP5**	**Lead beneficiary**	MPIA		
Work package title	**Selection function for combinations of Gaia and other surveys**				
Participant number	1	**2**	3	4	5	6
Short name	ULEI	**MPIA**	INAF	UCAM	MONA	NYU
Person-months	2	**21**	9	7	2	1
Start month	1	**End month**	42		

Objectives

The objective of this work package is to research and develop methods to derive selection functions for the combination of Gaia and other large sky surveys. The framework for the efforts in this work package is provided by deliverable D2.1. Within the lifetime of the project it is not realistic to make combinations of Gaia and an arbitrary numbers of other surveys. Hence the concrete implementation will focus on two selected cases which are to be implemented in work package 4. The detailed objectives are

1. Research and develop a generic method for constructing selection functions for combinations of surveys.
2. Apply this method to a selected number of cases of the combination of Gaia with another survey. We will focus our efforts on one example each of a combination of Gaia with a photometric and a spectroscopic sky survey.

Description of work

Task T5.1: Work package management (M1-M42) *Leader:* ***MPIA****. Contributors: MPIA*
Management and coordination of the research and development of the combined selection function and coordination of writing the paper corresponding to deliverables D5.6 and D5.8. This includes planning the work to be done, assigning sub-tasks and organizing the necessary meetings to discuss progress.

2 MPIA person months

Task T5.2: Generic combination method. (M7-M42) *Leader:* ***MPIA****. Contributors: ULEI, INAF, UCAM*
Research and develop a generic method for constructing selection functions for combinations of surveys. Provide documentation on the methods, including additional practical guidance instructions for complex combination. The implementation of the combination method will be done within work package 4.

9 MPIA + 1 ULEI + 6 INAF + 4 UCAM person months

Task T5.3: Combined selection functions. (M7-M42) *Leader:* ***MPIA****. Contributors: All other*

First, applying the method developed in task T5.2 to an external validation of the Gaia selection function using the DECaLS survey. Second, construct the selection function for the combination of Gaia and another survey. The focus will be on one photometric sky survey that GaiaUnlimited team members know well, i.e. the Pan-Starss PS1 release, and on one spectroscopic survey. For the latter we choose GALAH as an example of a recently started survey which is already well underway and to which GaiaUnlimited partner MONA has access along with the necessary expertise. Note that this task also includes matching the source list of Gaia to that of the other surveys. This should be done carefully to keep the combined selection function tractable, thus a significant effort is implied.

10 MPIA + 1 ULEI + 3 INAF + 3 UCAM + 2 MONA + 1 NYU person months

Role of partners

MPIA will lead Tasks T5.1, T5.2, and T5.3.

All other partners will contribute to tasks T5.2 and T5.3.

Deliverables

D5.1 Report with detailed scoping of the research and development effort. **(M12)**

D5.2 Report documenting the preliminary version of the combination of Gaia and a large photometric survey. **(M19)**

D5.3 Prototype software modules and data for the preliminary implementation of the combination of Gaia and a large photometric survey in Prototype V1. **(M19)**

D5.4 Report documenting the preliminary version the combination of Gaia and a large spectroscopic survey. **(M30)**

D5.5 Prototype software modules and data for the preliminary implementation of the combination of Gaia and a large spectroscopic survey in Prototype V2. **(M30)**

D5.6 Report documenting the methods to construct combined selection functions for Gaia and photometric surveys, to be submitted to peer-reviewed journal. **(M40)**

D5.7 Software modules and data for the final implementation of the combination of Gaia and a large photometric survey. **(M40)**

D5.8 Report documenting the methods to construct combined selection functions for Gaia and spectroscopic surveys, to be submitted to peer-reviewed journal. **(M40)**

D5.9 Software modules and data for the final implementation of the combination of Gaia and a large spectroscopic survey. **(M40)**

Work package number	**WP6**	**Lead beneficiary**		INAF		
Work package title	**Science applications**					
Participant number	1	2	**3**	4	5	6
Short name	ULEI	MPIA	**INAF**	UCAM	MONA	NYU
Person-months	13	13	**15**	13	2	0

Start month	1	End month	42

Objectives

The objective of this work package is to apply the Gaia selection function to concrete science cases. This serves to test the Gaia selection function implementation, to provide the community with worked examples of how to use the Gaia selection function, and to motivate the researchers involved in GaiaUnlimited and ensure that they have a scientific publication to enhance their prospects for future jobs, should they choose to continue in academia.

Here we present four exciting science cases and list the aspects of the selection function implementation that would be tested in each case. The science cases may change in their details. However, the GaiaUnlimited board will ensure that the actual science cases will cover the testing needed for the implementation produced in WP4.

A. Binary stars in Gaia Binary stars have a complex dependency on the Gaia selection function that depends on the source observables (e.g. colour and apparent magnitude, see Fig.8) and the orbital properties (e.g. period, inclination, mass ratio). In this application we would apply the selection functions for Gaia astrometry and spectroscopy, as well as the selection function specialized to binaries, to understand what Gaia would, or would not, observe for a fictitious population of binary stars. Given the selection functions we would research how the Gaia observables can be used to detect and partially characterise binary star systems, and more importantly to understand which systems would not have been detected or characterized.

By applying the selection functions to example populations of binary stars (e.g. using COMPAS, https://compas.science/, which is used and developed by a large group at MONA) we will perform population inference to make appropriate comparisons between models and the Gaia catalogue. With unbiased characterisation of binary star populations, and comparisons between observations and models where the selection functions are appropriately applied, we expect to provide unbiased answers to many long-standing questions in stellar multiplicity research, including: how stellar multiplicity changes with metallicity, understanding the initial mass ratio distribution of stars, the timescales of tidal circularisation, to name but a few.

Aspects tested: overall Gaia selection function; spectroscopic Gaia selection function; specialized Gaia selection function: deliverables D3.2, D3.3, D3.4, D3.5, D3.6, D3.7, D4.3, D4.5.

B. The Oort Limit, and Dark Matter near the Sun The positions and motions of stars near the Sun (< 500 pc), can constrain the gravitational potential at the Solar radius, and — when compared to a direct census of the stellar mass — can inform us about dark matter near the Sun. This is an analysis that has been carried out with ground-based data since the pioneering work of Oort [e.g. 34]. The robustness of results has been hampered by the fact that the local dark matter density is the difference between the total one (measured by dynamics) and the stellar one, obtained by star counts (or colour-magnitude diagram analysis); Gaia can do far better on both aspects, leading to a much better 'difference' of the two terms.

The dynamical mass (to be measured to $< 10\%$) requires exceptionally good knowledge of the selection function, as star-tracer densities and their derivatives enter in the analysis (see Eq. 3). The census of stellar mass can happen via placing stars onto the colour-magnitude diagram (using parallaxes and precise colours), inferring each star's mass and adding them up. But in any magnitude limited sample, the mix of luminous to faint stars will change as a function of distance: both dust extinction and the selection function will need to be modelled precisely.

This foreseen paper will aim to provide a benchmark on how to incorporate the Gaia selection function into a dynamical analysis (and tell us how much dark matter there is near the Sun).

Aspects tested: overall Gaia selection function; cross-catalogue selection function; combination of Gaia selection function with (externally determined) 3D dust map: deliverables D3.2, D3.3, D5.4, D5.5, D4.3, D4.5.

C. Mapping non-axisymmetric structures of the Milky Way via Cepheids A particularly important population are classical Cepheids. With ages between 50 and 500 Myr, classical Cepheids are excellent tracers of the young stellar population in the Galactic disc and spiral arms. Their characteristic variation in luminosity makes them easily recognized even in the severe crowding or high extinction conditions typical in these regions of the Milky Way. Their pulsation characteristics (period, mean magnitudes/colours) are linked through tight period– luminosity and period-Wesenheit relations, which make classical Cepheids primary distance indicators for use at distances beyond which standard parallax methods break down [12, 26]. To accurately map the non-axysymmetric structures of the disk, it is of fundamental importance to characterize the selection function of the sample, to avoid being biased by observational effects, such as foreground extinction and the sample completeness.

Moreover, to understand the dynamical nature of the spiral arms and of the warp, mapping their geometry in configuration space is not sufficient. To understand why these structures exist and how they evolve, it is necessary to investigate their signature in velocity space. We are able to trace the bulk motions in the disc using the proper motions and line-of-sight velocities provided by Gaia for millions of stars, which has already provided indications of non-axisymmetric streaming motions in the giants [see Fig. 5 in 18], as well as the kinematic signature of the warp in the systematic vertical velocities of both young and old populations [31]. Furthermore, the warp precession has been measured for a sample of 12 million giants using a simple spatial and kinematic model of the Galaxy and taking into account a preliminary Gaia selection function and measurement errors [30].

As the Gaia survey of line-of-sight velocities from the most recent data release DR2 is limited to stars brighter than about 12.5 magnitude, the volume of the disc that can be fully mapped in velocity space is still somewhat limited. This volume will gradually increase with future Gaia data releases, but here again the Gaia data for classical Cepheids are a potential treasure. With relatively accurate distances spanning the Galactic disk, and Gaia proper motions, there is the potential to map the vertical kinematics of the entire Galactic disk. This would allow us to adopt and constrain more sophisticated models of the warp, or even detailed simulations of disk-satellite interactions, opening the door to the field of Galactic seismology. With additional radial velocity for these stars, either from Gaia or other surveys, we will also be able to probe in-plane non-axisymmetric streaming motions, and a fuller understanding of the Milky Ways spiral arms. Such data will potentially allow us to paint a more global picture of the dynamical state of the Milky Way. However, for a proper interpretation of the data into a dynamical model, the Gaia selection function for Cepheids must be understood. This is particularly important as our sample of Cepheids is relatively small, being a tiny fraction of the stellar disk. Using such a small sample as a kinematic tracer requires an accurate selection function of this population to avoid a biased picture.

Aspects tested: selection function of variability-classified objects in the Gaia catalogue; cross-catalogue selection function; combination of Gaia selection function with (externally determined) 3D dust map: deliverables D3.2, D3.3, D3.6, D3.7,D5.4, D5.5, D4.3, D4.5

D. Evolution of the Milky Way disk GALAH is a high-resolution spectroscopic survey of $\sim 10^6$ stars with the goal to measure up to 28 chemical elements for each star. The union of

Gaia and GALAH represents a unique data subset. GALAH has the largest suite of chemical elements measured for any large set of stars, and because most GALAH stars are bright and nearby, no other spectroscopic survey benefits more from Gaia's exquisite astrometric precision. Currently no other sample is better suited to test formation models for the Milky Way disk, or to trace it's detailed chemical evolution with time.

This combined data set also presents a suitable challenge, in that the selection function for both catalogues must be modelled. In theory GALAH has a selection function that depends primarily on the apparent magnitude ($12 < V < 14$ mag) and sky position ($10^\circ < |b| < 60^\circ$ and sources must be visible from the sourthern hemisphere). In practice, however, the true selection function is complicated by weather, instrument errors, fibre allocations (e.g., due to crowded fields, or if the stellar density was 1–1.5 times the fibre density then the extra sources may not be re-observed) and by other astrophysical properties (e.g., dust and distance both contribute to determining whether a star is observed). Compared to most ground-based spectroscopic surveys, the *intended* GALAH selection function is relatively simple, yet complicated by other astrophysical properties, and real world effects.

As part of this objective we will combine the Gaia and GALAH selection functions in order to understand the kinematic and chemical evolution of the nearby Milky Way disk. The process of combining selection functions from multiple surveys will serve as an example for all other astronomy surveys to properly model their own selection functions, and to combine it appropriately with the Gaia selection function to be developed in this program. With the Gaia/GALAH subset, and its defined selection function, we will set out to test galaxy formation models (in a qualitative way), and to constrain dynamical effects such as radial migration, blurring, and churning, in a quantitative way.

The modelling approach put in place here can be transferred to upcoming spectroscopic surveys of the Milky Way, such as SDSS-V [28], which will offer detailed element abundances across the entire sky, with a 20 times larger sample. And, the larger the sample, the smaller $1/\sqrt{N_\mathrm{sample}}$, and consequently the more crucial the knowledge of the selection function.

Aspects tested: overall Gaia selection function, selection function for combined surveys: deliverables D3.2, D3.3, D5.4, D5.5, D4.3, D4.5

Description of work

Task T6.1: Work package management (M1-M42) *Leader:* ***INAF****. Contributors: INAF*
Management and coordination of the efforts on the science applications of the Gaia selection function. This includes monitoring the progress of the scientific papers, organizing the necessary meetings to discuss the paper contents, and ensuring that the papers expose the use of the main elements of the tools developed in GaiaUnlimited.

2 INAF person-months

Task T6.2: Science application A. (M7-M42) *Leader:* ***ULEI****. Contributors: ULEI, MONA*
Carry out the research related to science application A and write a paper for a peer-reviewed journal.

13 ULEI + 1 MONA person-months

Task T6.3: Science application B. (M7-M42) *Leader:* ***MPIA****. Contributors: MPIA*
Carry out the research related to science application B and write a paper for a peer-reviewed journal.

13 MPIA person-months

Task T6.4: Science application C. (M7-M42) *Leader:* ***INAF****. Contributors: INAF*
Carry out the research related to science application C and write a paper for a peer-reviewed journal.
13 INAF person-months

Task T6.5: Science application D. (M7-M42) *Leader:* ***UCAM****. Contributors: UCAM*
Carry out the research related to science application D and write a paper for a peer-reviewed journal.
13 UCAM +1 MONA person-months

Role of partners

ULEI will lead Task T6.2.

MPIA will lead Task T6.3.

INAF will lead Task T6.4.

UCAM will lead Task T6.5.

All partners will contribute to each of the tasks T6.2–T6.5.

The post-doctoral researchers funded through GaiaUnlimited are expected to do the bulk of the scientific analysis for the tasks above, supported by local supervision.

Deliverables

D6.1 Progress report on science application A. **(M20)**

D6.2 Progress report on science application B. **(M20)**

D6.3 Progress report on science application C. **(M20)**

D6.4 Progress report on science application D. **(M20)**

D6.5 Paper on science application A submitted to peer reviewed journal. **(M42)**

D6.6 Paper on science application B submitted to peer reviewed journal. **(M42)**

D6.7 Paper on science application C submitted to peer reviewed journal. **(M42)**

D6.8 Paper on science application D submitted to peer reviewed journal. **(M42)**

Table 5.2c: Deliverable list

Deliverable number	Deliverable name	WP no.	Lead participant	Type	Dissemination Level	Delivery date (proj. month)
D1.1	Kick-off meeting (plenary). The minutes of meeting will be delivered as a record that the meeting took place.	WP1	ULEI	OTHER	CO	1
D1.2	Data management plan.	WP1	ULEI	R	PU	6

Continued on next page

D4.1	First version of web-portal for GaiaUnlimited.	WP4	ULEI	DEC	PU	6
D2.1	Document on the mathematical formulation of survey selection functions (to be submitted to open access peer-reviewed journal).	WP2	MPIA	R	PU	12
D3.1	Report with detailed scoping of the research and development effort.	WP3	UCAM	R	CO	12
D4.2	Report with detailed scoping of the development and implementation effort.	WP4	ULEI	R	CO	12
D5.1	Report with detailed scoping of the research and development effort.	WP5	UCAM	R	CO	12
D3.2	Documentation of preliminary top level Gaia selection functions.	WP3	UCAM	R	PU	19
D3.3	Prototype software modules and data for the preliminary top level Gaia selection functions implementation in Prototype V1.	WP3	UCAM	DEM	PU	19
D5.2	Report documenting the preliminary version of the combination of Gaia and a large photometric survey.	WP5	MPIA	R	PU	19
D5.3	Prototype software modules and data for the preliminary implementation of the combination of Gaia and a large photometric survey in Prototype V1.	WP5	MPIA	DEM	PU	19
D4.3	Prototype V1: open source and open data implementation of the top level Gaia selection functions.	WP4	ULEI	DEM	PU	20
D4.4	Second version of web-portal for GaiaUnlimited, ready for testing by community.	WP4	ULEI	DEC	PU	20
D6.1	Progress report on science application A.	WP6	ULEI	R	CO	20
D6.2	Progress report on science application B.	WP6	MPIA	R	CO	20
D6.3	Progress report on science application C.	WP6	INAF	R	CO	20

Continued on next page

D6.4	Progress report on science application D.	WP6	UCAM	R	CO	20
D1.3	Community workshop no. 1 and mid-term meeting (plenary).	WP1	ULEI	OTHER	PU	21
D1.4	Report no. 1 summarizing community feedback on GaiaUnlimited tools.	WP1	ULEI	R	PU	22
D3.4	Documentation of preliminary spectroscopic Gaia selection functions.	WP3	UCAM	R	PU	30
D3.5	Prototype software modules and data for the preliminary spectroscopic Gaia selection function implementation in Prototype V2.	WP3	UCAM	DEM	PU	30
D3.6	Documentation of preliminary specialized Gaia selection functions.	WP3	UCAM	R	PU	30
D3.7	Prototype software modules and data for the preliminary specialized Gaia selection function implementation in Prototype V2.	WP3	UCAM	DEM	PU	30
D5.4	Report documenting the preliminary version the combination of Gaia and a large spectroscopic survey.	WP5	MPIA	R	PU	30
D5.5	Prototype software modules and data for the preliminary implementation of the combination of Gaia and a large spectroscopic survey in Prototype V2.	WP5	MPIA	DEM	PU	30
D4.5	Prototype V2: open source and open data implementation of the overall and specialized Gaia selection functions and selection functions for combinations of Gaia and other surveys.	WP4	ULEI	DEM	PU	31
D4.6	Third version of web-portal for GaiaUnlimited, including feedback from community.	WP4	ULEI	DEC	PU	31
D1.5	Community workshop no. 2.	WP1	ULEI	OTHER	PU	32
D1.6	Report no. 2 summarizing community feedback on GaiaUnlimited tools.	WP1	ULEI	R	PU	33

Continued on next page

D3.8	Documentation of the top level Gaia selection functions (also to be submitted to as paper to a peer-reviewed journal).	WP3	UCAM	R	PU	40
D3.9	Software modules and data for the final top level Gaia selection functions implementation.	WP3	UCAM	OTHER	PU	40
D3.10	Documentation of the spectroscopic Gaia selection functions (also to be submitted to as paper to a peer-reviewed journal).	WP3	UCAM	R	PU	40
D3.11	Software modules and data for the final spectroscopic Gaia selection functions implementation.	WP3	UCAM	OTHER	PU	40
D3.12	Documentation of the specialized Gaia selection functions (also to be submitted to as paper to a peer-reviewed journal).	WP3	UCAM	R	PU	40
D3.13	Software modules and data for the final specialized Gaia selection functions implementation.	WP3	UCAM	OTHER	PU	40
D5.6	Report documenting the methods to construct combined selection functions for Gaia and photometric surveys, to be submitted to peer-reviewed journal.	WP5	MPIA	R	PU	40
D5.7	Software modules and data for the final implementation of the combination of Gaia and a large photometric survey.	WP5	MPIA	OTHER	PU	40
D5.8	Report documenting the methods to construct combined selection functions for Gaia and spectroscopic surveys, to be submitted to peer-reviewed journal.	WP5	MPIA	R	PU	40
D5.9	Software modules and data for the final implementation of the combination of Gaia and a large spectroscopic survey.	WP5	MPIA	OTHER	PU	40

Continued on next page

D1.7	Community workshop no. 3 and closing meeting (plenary).	WP1	ULEI	OTHER	PU	41
D4.7	Open source and open data implementation of all selection functions developed in GaiaUnlimited.	WP4	ULEI	DEC	PU	42
D4.8	Final version of the Web-portal providing access to the Gaia selection function data products and tools delivered by GaiaUnlimited.	WP4	ULEI	DEC	PU	42
D6.5	Paper on science application A submitted to peer reviewed journal.	WP6	ULEI	R	PU	42
D6.6	Paper on science application B submitted to peer reviewed journal.	WP6	MPIA	R	PU	42
D6.7	Paper on science application C submitted to peer reviewed journal.	WP6	INAF	R	PU	42
D6.8	Paper on science application D submitted to peer reviewed journal.	WP6	UCAM	R	PU	42

Table 5.2d: List of milestones

Milestone number	**Milestone name**	**Related WPs**	**Estimated date**	**Means of verification**
M1	Kick-off meeting	WP1	1	Organized by GaiaUnlimited executive board
M2	Hiring of researchers	WP1	4	Partners notify GaiaUnlimited executive board
M3	Document on mathematical formulation of selection function submitted to peer-reviewed journal	WP2	12	Document approved by GaiaUnlimited executive board
M4	Document describing the preliminary Gaia selection function.	WP3	19	Document approved by GaiaUnlimited executive board
M5	Public prototype V1 of selection function implementation	WP3, 4, 5	20	Software and data released and validated by user group
M6	Community workshop and mid-term meeting	WP1	21	Organized by GaiaUnlimited executive board
M7	Public prototype V2 of selection function implementation	WP4	31	Software and data released and validated by user group
M8	Community workshop	WP1	32	Organized by GaiaUnlimited executive board
M9	Papers and documentation describing the Gaia and combined selection functions ready to submit to peer-reviewed journals	WP3, 4, 5	40	Verified by GaiaUnlimited executive board
M10	Closing meeting and community workshop	WP1	41	Organized by GaiaUnlimited executive board
M11	Publicly available selection function implementation: open source tools, open access data, and a web-portal	WP1, 3, 4, 5	42	Software and data released and validated by user group

Table 5.2e: Critical risks for implementation

Description of Risk	**WPs involved**	**Proposed risk-mitigation measures**
Late recruitment of researchers. Medium	WP3–6	Six month preparatory phase added in schedule to accommodate early advertizing of open positions. Use professional network of GaiaUnlimited participants. Executive board will track recruitment.
Recruited researchers leaving early. Medium	WP3–6	Keep staff motivated, including through supporting career development. Coordinator and executive board will monitor the well-being of the personnel.
Unexpected complexity of tasks. High	WP3–6	Prioritize the GaiaUnlimited efforts and make timely decisions on dropping tasks deemed too complex.
Lack of information from DPAC. Low	WP3–5	Several GaiaUnlimited team members are members of DPAC. Coordinator can identify alternative routes within DPAC to obtaining information.
Coordination problems with ESA and DPAC leading to the selection function data products not being available through ESA archives. Low	WP4	Plan for alternative distribution channels. Coordinator to maintain close connections to ESA and DPAC.
Delay of Gaia data releases. Can lead to difficulties in coordinating with DPAC/ESA (due to their prioritizing working on the releases). Medium	WP3–6	Coordinator to monitor the situation and agree mitigation measures with ESA/DPAC. We stress that even in the extremely unlikely case that no further Gaia data releases were to happen, the GaiaUnlimited effort will still have a major impact on the exploitation of the existing Gaia DR2 data.
New lock-down measures and travel restriction across Europe and the world related to Covid-19 or a new pandemic. High	WP1–6	Coordinator and participants will monitor the situation and take timely measures to move planned physical meetings online. Note that regular telecons are already planned as outlined in section 3.2.3, which make the consortium more robust against new pandemics. If foreseen working visits cannot take place the work can be done remotely or the timing of the visit adjusted, taking into account the project schedule.

Table 5.2f: Summary of staff effort

Partic. no.	**Partic. short name**	**WP1**	**WP2**	**WP3**	**WP4**	**WP5**	**WP6**	**Total person months**
1	ULEI	**8**	1	2	**23**	2	13	49
2	MPIA	2	**2**	6	12	**21**	13	56
3	INAF	2	1	11	6	9	**15**	44
4	UCAM	2	1	**19**	2	7	13	44
5	MONA	0	1	2	1	2	2	8
6	NYU	0	1	1	1	1	0	4
Total		14	7	41	45	42	56	205