Astronet 2021 - Computing Panel

Top Level Key recommendations

Key 1: *[People]* we recommend developing and investing in a professional software engineering / computational skills base in Astronomy. This has many implications and requirements, including career development with clear progression pathways in academia and improving the diversity of the workforce. Such careers have to be promoted and considered as an integral part of our science/research portfolio. Traditional metrics for academic performance are often inappropriate for measuring the impact and usefulness of computationally focussed outputs. New assessment criteria, for example based on industrial models, should be adopted by the astronomy community to give proper credit to essential contributions that technicians and software engineers provide.

Key 2: *[Data, Open, Archive, User-to-Data]* we recommend that missions and facilities plan an integrated approach for data products and software tools: their design, delivery, maintenance and development should be sufficiently planned for and resourced already at the onset and for the lifetime of the mission/facility. We recommend that initiatives be supported for the long term preservation and scientific use of data.

Key 3: *[Data, Archive, User-to-Data, Open, Green]* we recommend to adopt and further develop a "Tiered" approach for Data Infrastructure, for all types of data pertaining to astrophysics, including models, simulations and mocks, and where beneficial to connect with similar frameworks developed for other disciplines of science. A `Tiered' Distributed Analysis model maximises the transformation of science data to scientific results, organising and sharing expertise to benefit from economies of scale, with attention given to preventing the exclusion of individual communities, as well as monitoring and minimising the environmental cost.

Key 4 : *[Open, User-to-Data, People, Green]* we recommend embracing a fully collaborative, open and synergistic view when it comes to the astronomy-computing ecosystem, encompassing data, software, processing, analysing and modelling. The community should implement and support official mechanisms to monitor, acknowledge and reward behaviours that model this ambition.
Open science, data and software sharing, archives, cloud computing, platforms and service infrastructure represent various facets of an integrated view of computing in astronomy: this should be acknowledged and acted upon.

Key 5: *[Green]* We recommend that Astronet produces or commissions a biennial quantitative report to assess the carbon footprint of computing in Astronomy. The initial review should define clear measurable metrics against which progress can be evaluated. We further recommend that Astronet strongly encourages the use of efficient programming languages and computational architectures for intensive computations, the training of its scientists and developers in this regard, and strives to ensure that all computation performed is strictly required to achieve the desired science goals - all with the aim of minimising the environmental cost.

Key 6: we recommend that Astronet develop specific actions to coordinate cross-cutting activities. Transparent mechanisms should be defined and implemented to monitor and report on the progress pertaining to computing in Astronomy. This may involve the implementation of dedicated polls and studies, coordinated working groups reaching out to research groups and centres, communities, national and transnational entities, leading to an exposed and clear mandate for technology and solution watches.

The implementation of such recommendations requires an update of our research crediting and promoting system, as well as a more collaborative and ordered set of resources (e.g., funding agencies, grants, computing centres).

Keywords in Brackets refer to specific Sections of this document:

- *[Open]* = The paradigm change: Open Science, Data and Software sharing
- *[Data]* = Mission / Facility data processing
- *[Archive]* = Archives for data and software
- *[User-to-Data]* = The next Generation Tools: Cloud computing, Platforms, Collaborative frameworks
- *[Green]* =Green Data Infrastructures, Reducing the Carbon Footprint
- *[People]* = Training, Careers, People

1. General Introduction

The wide and diverse range of astrophysics research in this roadmap leads to cross-cutting requirements for data handling techniques, software, simulations and computing. Astrophysics research has already become a data-intensive endeavour, and the future large, complex and inter-dependent data sets that will be generated by observatories, space missions, mock experiments, theoretical model simulations and numerical simulations require new tools and approaches for doing science.

Scientific questions are driving the need for surveys that e.g., cover large regions of the sky, obtain deeper exposures or include multi-frequency/messenger coverage. The sheer data volumes from such experiments will increase at a high rate with a number of projects/facilities expected to have annual data production of the order of petabytes, and hundreds of petabytes for the extreme example of SKA. The archives of the future large data producing telescopes and missions will dwarf the existing ones in terms of volume and complexity.

We have also entered the era of multi-messenger astronomy. Going beyond the often-quoted advent of gravitational wave science or the synergy with astroparticle physics experiments, the trend is towards the need to combine data from observations across the electromagnetic spectrum and beyond. This is in addition to the transient object searches of time domain astronomy, leading to high flux event streams where rapid detection, classification and follow-up observations bring new challenges for computing and analysis.

Dedicated simulations, either conducted as numerical experiments or modelling efforts, also reach levels of complexity and volumes which call for an updated assessment of how we run them, process them, and share and distribute their outputs and associated data products. This is all the more true when auxiliary products and models (e.g., mock simulations, various realisations of the same datasets), themselves challenging to produce and store, must be associated with those primary simulation datasets.

We can expect users to interact with data from diverse sources. The need to jointly analyse datasets from a variety of instruments/facilities with different characteristics, as well as to connect the observational and simulation/modelling landscapes further motivates the development of flexible and transparent frameworks where the focus shifts from *data* towards the *applied expertise and tools*. This is sometimes reinforced by the paradigm shift pertaining to sharing information and the push for “Open Science”, often emphasised by national and European funding agencies and organisations. *The added value will come from the access to expertise, supported by media and tools that alleviate the difficulty of dealing with varied datasets.*

This motivates new ways to deal with data, using e.g., science analysis platforms that enable a "bringing computing to the data" paradigm. We further need to look at the technical resources, infrastructures, identify potential obstacles and challenges (e.g., the scaling and access to fast network connections). Most importantly, we need to scrutinise the existence and access to relevant expertise and the synergy with new profiles.

The technologies that will enable cross-cutting activities related to data and computing are evolving extremely rapidly. Astronomy, and indeed all scientific domains, will benefit greatly from increases in network speed, access to large volume data storage, fast computing as well as from the adoption and application of machine learning, informatics and the rise of data science as a discipline. Many fields face common challenges which at some level can be addressed by initiatives and developments such as the European Grid Initiative (EGI), and European Partnerships for the European Open Science Cloud (EOSC) and for High Performance Computing. There are also benefits that will come from widespread use of common tools and methodologies, such as collaborative development tools (e.g. Gitlab), sharable notebooks and workflows (e.g. via JupyterHubs), and building on ground-up developments such as `astropy`. From the individual researcher, up to the largest projects, the cross-cutting technologies that will influence the next decade will involve the wider context of computing and data infrastructures for science in general.

This document aims to provide a few principles and key points that are designed to support a pragmatic approach to address such challenges. It emphasises the need to actively monitor the ongoing and future technology and paradigm changes associated with computing in astronomy. It requires acting immediately and globally, with close coordination between missions, facilities, funding agencies, and key actors in the astronomical community. These challenges are not specific to computing in astronomy, are shared by other sciences, and must sometimes be viewed at the societal level. We thus need to coordinate such an upgrade of our services and infrastructures with other disciplines, building on the robust baseline and specifications we have deployed over the last decades, while making use of the quite unique vantage point and attractiveness of astronomy.

Finally, we should acknowledge that many of the proposals made here come with an apparent cost. However, investment in data brings added value to the wider astronomical community, and can help reduce waste in other areas, for instance, in removing costs associated with inefficient access to high quality science ready data products. As a community we need to provide a clear sense of direction and make choices, to support the investments in people, computing infrastructure etc, that we recommend. Given how vital the contribution of computing is to modern astronomy and astrophysics, it is a science-driven necessary cost and must be treated as a top priority item, in contrast with what has so often been done in the past. This is a prerequisite for high quality and shared science. We should also consider that implementing the coordination advocated in the present document would partly address the existing resource (software, hardware, possibly staff) fragmentation, which in turn would boost efficiency and create potential cost margins.

2. [OPEN] The paradigm change: Open Science, Data and Software sharing

Key Points

- Open-1: The Astrophysics community should be encouraged to engage with, and participate in, EOSC and other European or national Open Science initiatives, to benefit

from them, and to ensure that astronomy requirements and feedback are taken into account.
- Open-2: Promoting and adopting principles of FAIR, Open data and open-source software should become the default requirement for all of Astronomy within the next decade. This should commit the research community at large, as well as all the associated actors: research institutes, national and transnational organisations, missions and observatories. Doing so will be essential to fully exploit the capabilities of next-generation facilities, maximise the impact of scientists' research, and build up a virtuous and inclusive environment that promotes the development of tools.
- Open-3: Encourage community development of software, supported by long-term funding, and provide proper credit to such efforts (see Open-4).
- Open-4: The current framework for reward and recognition in astronomy research is outdated and requires radical changes to acknowledge the increasing contribution of software developers, data scientists and data stewards to the scientific yield of our facilities and instruments.
- Open-5: Given the potential costs associated with making Big data sets, simulations sets, and software Open and compliant with the FAIR principles, it is crucial that that community establishes whether complete access is useful in all cases. Where the choice is made not to make particular datasets FAIR, this should be done because of an untenably high cost-benefit ratio for the whole astronomy community and not to e.g., artificially increase the journal article output of those who do have full access.

Open Science and Open Data are central to the Strategic Research and Innovation Agenda of the European Open Science Cloud (EOSC), which is now entering its 10-year implementation phase following the creation of the EOSC Association. EOSC will "*enable a trusted, virtual, federated environment in Europe to store, share and re-use research data across borders and scientific disciplines*" based on the precept that all data artefacts should be Findable, Accessible, Interoperable and Reusable[1] (FAIR).

It is incumbent upon the astronomy and astrophysics communities within Europe to ensure that they continue working towards the adoption of FAIR principles by defining sets of common metadata standards and agreed procedures to enable widespread data sharing. Substantial progress has already been achieved with the astronomical Virtual Observatory (VO), which is a framework for FAIR astronomy data based on common internationally agreed interoperability standards. Since 2002, the International Virtual Observatory Alliance (IVOA) has advised and overseen development of the VO with support from numerous aligned national initiatives. The VO is now a mature framework that is used by astronomers around the World. It is embedded in major astronomy archives and data centres including ESA, ESO and CDS, and is integrated into a large number of tools, online services and software frameworks.

The usefulness and scientific yield of Open and FAIR astrophysical data is substantially reduced without the knowledge, expertise and appropriate software to process and analyse them. An outstanding challenge for the astronomy community in the coming decade is to build on the success of VO by defining common strategies for sharing software, its documentation and contextual metadata about its intended execution environment. One example of ongoing work to address this challenge is the ESCAPE Open Software and Services Repository (OSSR). The OSSR prototype aims to provide "*a sustainable open-access repository to share scientific software and services to the science community and enable open science*", but it is still to be seen whether such

[1] We emphasise that the "R" in FAIR stands for the principle of "reusability", not to be confused with "reproducibility".

activities will lead to a relevant milestone and concrete adoption and usage from the research community. Existing technologies and paradigms including unit testing, version control, continuous integration and deployment and containerisation can undoubtedly be applied in the context of astronomy, but best practises for how and when to use them remain to be established.

It is widely accepted that the construction and maintenance costs of modern astrophysical observatories make large-scale collaboration between scientists, institutes and national governments essential. A fact that seems less well recognised is that similarly collaborative activity will be required to effectively exploit datasets with the diversity, size and complexity that these new observatories will provide. Computing centres and coordinated calls for computing time are increasingly requesting that research programmes (publish and) share the outcome of the simulations they conduct. Those models and astrophysical simulations often need to be associated with specific datasets to enable their interpretation. This is a challenging task in itself, even prior to any consideration regarding a broader sharing effort. While the associated efforts and cost need to be acknowledged and addressed, effective data and software sharing demonstrably enhances the productivity of astronomers, expands the scale and scope of the projects they can undertake and improves the quality of the results they produce. Sharing of software and data, with appropriate credit, also increases the impact by allowing others to build on past effort and previous investments.

There is a general need to more strongly incentivise astronomers to embrace principles for data and software sharing. Achieving this will likely require us to reconsider our funding and recognition systems that predominantly measure success and productivity according to the numbers of refereed journal articles that researchers produce. This system naturally discourages sharing of data and software because astronomers will be reluctant to release data until they have extracted and published any scientific insights they contain. While we must acknowledge the relevance and usefulness of established proprietary times when it comes to acquiring data and motivating its timely analysis, it is important to realise that further disincentives may persist unless we improve mechanisms to recognise data and software as legitimate scientific products in their own right. We must establish channels to accept and promote articles that emphasise software or data that are not systematically attached to a scientific interpretation[2]. At the same time, the community must find ways to recognise and reward the substantial contribution made by technicians and software engineers whose work is *essential* but is not readily publishable in *any* form. Industrial models for performance assessment and career progression may be appropriate for this purpose and should be seriously considered.

If the incredible potential of forthcoming astrophysical observatories is to be realised, this instinct for proprietary software and closed datasets cannot survive the next decade. The astronomy community must establish new mechanisms that actively encourage collaboration and reward activities like *Open software development* and *Open data stewardship* as essential activities within astrophysical research. The technologies required to provide recognition and reward for these activities and the digital assets they produce already exist. Web-based utilities like Zenodo provide a citable Digital Object Identifier (DOI) that can be associated with datasets or software products. Journals dedicated for the publication and exposure of software are emerging (e.g., Journal of Open Source Software - JOSS - https://joss.theoj.org/). While numerous digital assets related to astronomy now have DOIs associated with them, the practice of citing these identifiers is yet to become widespread. Instead, citations for software or data are often attributed to articles

[2] Note the recent advent of dedicated journals, e.g., the Royal Astronomical Society "*Techniques and Instruments" see announcement at* RAS launches new multi-disciplinary journal)

describing science results they were used to derive. This is doubly damaging because it implicitly diminishes the perceived importance of the software or data product and denies a second citation to the relevant software developer or data steward.

We also need to clearly state that DOIs, by themselves, have little value if the associated content is not properly curated and/or reviewed. There is therefore a need to associate a process with sufficient capacity to acknowledge, deal with and review the incoming flow of shared data and software products.[3]

When the Open community-led software development paradigm is followed, the rewards can be substantial. A compelling example is provided by the `astropy` project, which has revolutionised the way that astrophysical analysis software is designed, developed, extended and delivered. It has undoubtedly enhanced the productivity of astronomy researchers around the World. The collaborative development effort exemplified by the `astropy` project , which includes experts and research groups from all corners of the community, as well as engaged and exposed (scientific and engineering) staff from missions or observatories, helps to alleviate the burden on individual astronomers who can help to develop, extend or improve small subsets of the overall framework. Moreover, the widespread usage of `astropy`, in conjunction with well developed methodologies for unit testing, helps to ensure robust software by allowing astronomers to find and report bugs that might have gone unnoticed in a closed-source project. While we need to fully acknowledge both the short and long term direct costs and commitments such efforts entail (i.e., in terms of personnel), the overall gain in reach and agility are huge benefits shared by the community. We further need again to monitor where the credit for such achievements go, and properly promote the developers and actors who made it possible.

It should be clearly acknowledged that this voluntary model for scientific software development also has drawbacks. Contributors to community-maintained software packages are predominantly early-career researchers who often find themselves without time to maintain their involvement when their careers shift or progress. The loss of experienced developers and their expertise can leave substantial sections of the code without a knowledgeable maintainer and may impact the ultimate longevity of a software package. These statements are not limited to the Astropy package - they apply to a whole ecosystem of widely used, freely contributed and scientifically useful software products. Notable examples include Numpy, Scipy, Matplotlib, Jupyter but there are many more. The community should consider new funding mechanisms that allow skilled developers to maintain long-term involvement with particular software packages and frameworks. Prolific users of scientific software should consider assigning members of their in-house pools of software engineers as contributors to community-developed software as an in-kind contribution.

We further need to acknowledge the fact that FAIR principles represent a very high standard to reach. Despite the obvious advantages and rewards that come when data and software are widely shared, it can be technically challenging and costly on many fronts (e.g., staffing, expertise, funding). There may be cases when the significant cost of adopting FAIR principles is not sustainable or even desirable if the cost/benefit ratio is deemed to be too high. Notable examples of when this may be true include the petabyte scale raw data produced by next-generation facilities and the equally large data sets produced by current simulations. While these data must certainly be retained for as long as possible, making these genuinely Big datasets fully accessible

[3] We do not address the potential move towards Open Access (for publications), for which many regimes (and associated business models) have been proposed and debated: it is an important and complex topic, but outside of the scope of the present document.

implies sophisticated technological solutions for cataloguing and retrieval, as well as substantial networking resources to enable their transfer around the World. It is also unclear whether access to these raw data is actually useful to the vast majority of the research community. The computational resources required to process these data into science-ready products are not available to most astronomers, so it is very likely that providing access to smaller volumes of partially or completely reduced data is both more tractable and more desirable. Another plausible solution involves implementing access protocols via dedicated analysis services. Such services should implement software frameworks that deliver the required data products directly and deploy those frameworks via an associated service that provides a manageable interface to the Big data. Later in this document (see the [Archives] and [Data-to-User] Sections) we briefly discuss modes of analysis that can be applied when even the reduced data products are too large to be analysed locally.

3. [DATA] Mission / Facility Data Processing

Key Points

- Data-1: Data products and software tools should be recognised as an integral part of the design and development of the mission, instrument and/or facility, and the production of those adequately resourced as part of its development plan.
- Data-2: In that context, advanced data products and analysis tools should be considered as a key component at all phases of the development, operation and post operations lifecycles. Resources shall be available throughout the lifecycle to deliver the data products, with sufficient capability to ensure final product delivery (reflecting that final data sets can be large/complex).
- Data-3: in response to the increasing complexity of instruments, missions, facilities, and simulations being developed, and in order to best utilise expertise in data analysis, connecting the overall data processing development and operations with expert centres may be required. A tiered approach is highlighted where Tier 2 Data Processing Centres support the data processing needs of Tier 1 facilities, and provide a live and collaborative interface to Tier 3 instrument teams and the wider community, exploiting the use of standards in building these connections between facility-to-data-to-science.

The motivation for ground-based observational facilities and space missions[4] (hereafter ‘missions’ or ‘facilities’) to produce science-ready data products has grown very significantly over the last two decades. This has been driven by various factors, including the increasing complexity of modern datasets, the need to optimally imprint the know-how of instrument-builders and science experts into the delivered datasets, and the wish to maximise the scientific return and serve a broader community of scientists who should focus on scientific exploitation. With trends such as multi-messenger astronomy, and cross-fields fertilisation, the value of data is becoming more dependent on the availability of expertise and tools, rather than on its bare existence. In practice, “science-ready” data products and tools to address them now represent the key outputs of facilities and missions, and deserve further attention.

[4] We discuss astronomical observational facilities and missions to encompass major telescopes (space or ground based) and their associated instruments. We also include facilities that may be a telescope/instrument that is operated to carry out large scale surveys. For simplicity the words ‘mission’ or ‘facilities’ are used to refer to both space and ground based telescopes, instruments and large scale surveys. Current examples are e.g., Gaia, Euclid, PLATO, ESO-VLT, ESO-ELT, ESO-4MOST, SKA, CTA, and other astronomical infrastructures identified in the ESFRI roadmap. We note that large numerical simulations nowadays produce data of complexity comparable to observational data and thus can and should be treated as missions/facilities themselves.

These new facilities increasingly involve the generation of larger and more complex data sets, where there is often a requirement to incorporate a range of external data sets to achieve the key science aims. For instance, the ESA M2 Euclid mission, with key aims to probe cosmology through galaxy shear and clustering measurements, requires data not only from the Euclid spacecraft (broadband optical imaging and near infrared spectroscopy) but also multi-band photometry from ground-based imaging surveys. Likewise the final science data goals from the ESA M3 PLATO mission require not only the onboard high cadence photometry, but also high resolution spectroscopy for the high priority targets that will be sourced from a dedicated ground based observational programme. Facilities such as ESO's 4MOST rely on the availability of associated ground and space based image survey data in order to construct their input target catalogues.

In parallel, the success of a mission is increasingly being measured not only in terms of delivery of the new mission or facility on time and on cost, but also on the scientific productivity of the mission through its operational and post operational phases in terms of science outputs, which should be focused on scientific discoveries and should encompass a wide range of impact measures (societal or economic impacts, including e.g., innovation, education, outreach).

Further, the ambition of the science goals of new missions (only the strongest science cases are successful) and complexity of the resulting instrumentation, mean that it is unrealistic for any new mission to provide only raw data to the community. Over the last few decades the release of so called Level 1 products (broadly data with instrumental signatures removed) has been seen as a minimum key deliverable from any mission or facility (e.g. the public SDSS sky survey releases). More recently though, and especially for large survey missions, there has been an increasing expectation of the delivery of Level 2 data products along with key extracted astrophysical parameters (e.g. in the case of spectroscopic surveys, Level 1 spectra and Level 2 catalogues with source measurements of key chemical abundances (Galactic) or redshifts (extragalactic)). The generation of these layered data products in turn requires development of sophisticated processing chains, effective data management systems and significant hardware infrastructures.

The increasing trend towards delivery of advanced data products requires the corresponding development of software and tools which must be regarded as core elements in the deliverables for any new facility/instrument. The availability of both the data products and either the software and tools that generated the data products, or a fully documented description of the software is essential in ensuring the transparent, fair, direct and robust scientific exploitation of newly delivered data (see the [Open] Section). This software development must be resourced at a sufficient level to match the requirements on data delivery, and is a key factor in the total cost, covering development, operations, and post operations, of any new mission, facility or instrument.

In short, ensuring the timely delivery of these complex, quality assured science data products, delivered in well documented releases, is a demanding process that must be adequately resourced. Effective mission/facility science data releases is increasingly a requirement in ensuring that the science goals of the missions are met, that the facilities are scientifically productive, and that the wider astronomical community are able to exploit the mission and complementary science data from other missions/facilities for their science.

The Gaia model - Large projects (especially in the case of ESA missions) are able to devote sufficient resources to enable the construction of highly performant science pipelines to generate their science data products. For instance, the ESA Gaia mission planned from an early stage, a series of increasingly rich Data Releases through the lifetime of the project. The complexity and richness of the data products released by Gaia has increased with each new release. All Gaia data releases are fully quality assured and documented. The quality of these releases, and effective distribution to the community, through the ESA Gaia Archive and partner centres, underpins the huge range of scientific advances made by the community based on the Gaia data. The data analysis model adopted by Gaia, with the analysis chains being developed and operated by the Gaia Data Processing and Analysis Consortium (DPAC) is an example of a successful large astronomy data infrastructure, noting that the DPAC is constituted as a community funded organisation, with participation from ESA. The Gaia model reflects that the analysis of Gaia data is so complex that it could not be carried out in the wider community, rather a critical mass of expertise is concentrated in the DPAC, which then ensures a cost effective delivery of science ready data to the community which they are then able to scientifically exploit.

Towards the SKA - The new generations of radio telescopes (like LOFAR and SKA) produce prodigious amounts of very complex raw data which can not be permanently stored and provided to users at their original time and frequency resolution. Processing the data is challenging and ongoing development and experience often leads to valuable improvements in calibration and (image) data quality. SKA has adopted a tiered processing approach, with the Observatory delivering science quality data which has been sufficiently reduced in volume to be distributed to a global network of SKA regional centres (SRCs) that will take up further processing and analysis of the data by the scientific community. The SRCs will also host the (distributed) and globally accessible primary science archive - consisting of both the observatory data products and advanced data products created in the SRCs while the Observatory retains a (backup) copy of its observatory data products. The development and operation of the SRCs is also largely a community funded activity - outside the scope of the SKA Observatory.

As exemplified by ESA Gaia (see dedicated panel), large space-born missions often plan and devote sufficient resources to provide robust delivery and support for science data products. The picture and approach taken for new ground based facilities is more mixed. One the one hand, there are commendable and efficient engagements, including e.g., the European side of the ALMA observatory which has developed a network of expertise centres (the European ALMA Regional Centres, or European ARCs) connecting with the scientific community and “staffed (...) throughout the lifetime of a project, from proposal preparation to data analysis”. On the other hand, new instruments that are added to major telescopes often have development budgets that do not cover longer term operations, and hence the data processing systems delivered for these new instruments are limited to basic pipelines, which are assumed to be run by the eventual users of the facility. This leads to an uneven situation when it comes to the maturity and readiness of the delivered software and tools, depending on parameters associated with the scientific focus of the instrument-builders, available expertise during the design and development of the mission, and the achieved balance (or tension) between key efforts to deliver a given instrumentation. For instance, ESO facility instruments typically are delivered with basic instrument pipelines that can be run by the user to reduce their raw observational data. There are very significant efforts and associated resources engaged in processing reduced data for the users. However, it almost unavoidably often depends on specific choices, and science data products are not generally available to the general astronomical community except for a given set of programmes. There are obvious exceptions, where the instrument is foreseen to be used in a larger ‘survey’ operation mode, and for which data products to be released publicly are planned from the start. Even in such cases, the resources available for the data analysis system development and operations may be limited.

In the development of new facilities with more modest budgets, there are opportunities to benefit from shared/common approaches. Complementary paths for the implementation of software and analysis systems could be coordinated between the mission/facility/observatory and the community. These should be planned as early as possible, certainly in the initial phases of the facility development. There is an opportunity to benefit from expertise in thematic processing centres in the development of pipelines and data infrastructures that meet the needs of science requirements generated by the new instrument teams. Coordinating a guided effort where deliverables are developed, promoted, and implemented via an integrated communication between the builders, the coordinators (e.g., an observatory or mission) and the scientific community will help ensure an improved common standard of data product delivery for all new instrumentation. A tiered approach encompassing identified centres of expertise for experimental and simulated data analysis represents an efficient route towards the development of robust, mission-specific, tools, as well as a more agile set of packages and frameworks. Those expertise centres can interact with the science teams of the projects that they are working with to ensure the analysis systems meet the project's science requirements, and can directly connect with the scientific community via exposed two-way communication and collaborative channels.

The critical ingredient for an efficient and functional structure is that *this sense of direction is planned early, represents an integral part of the design and development of the facility, in close coordination with potential partners and with the scientific community*.

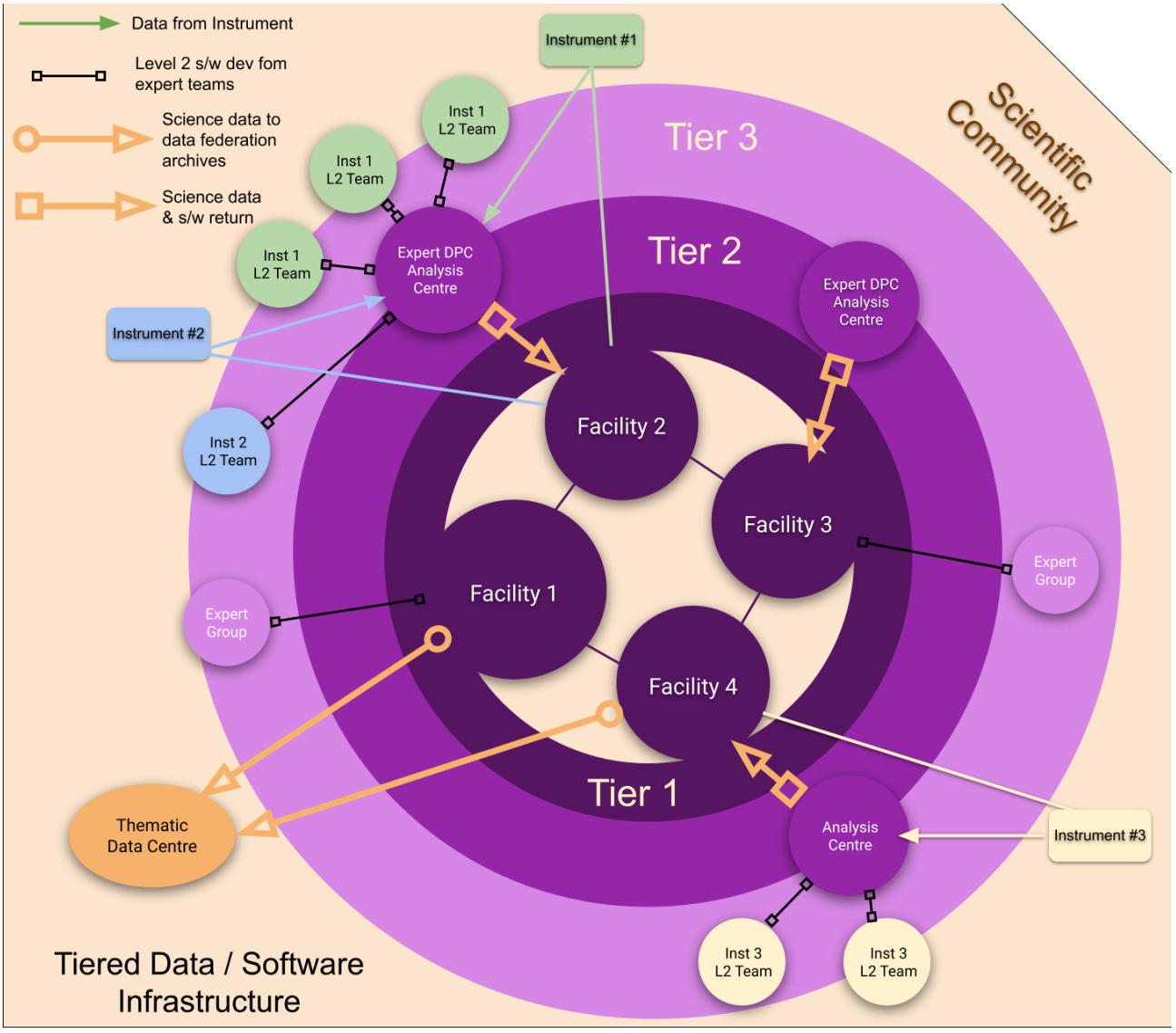

Figure 1: This diagram illustrates a tiered processing infrastructure model for the case of major astronomical facilities and survey instruments. Tier 1 facility operators such as ESO and SKA which are responsible for multiple data generating telescopes and their instrument suites, pass data from these to community provided Tier 2 expert data processing centres (DPC), which have critical expertise in specific techniques or wavelength domains (e.g. optical/near-IR, thermal-IR, radio, etc), including mocks and numerical simulations. Tier 2 centres connect directly with the scientific community and provide an interface to Tier 3 expert development teams and groups in the wider community who are responsible for the development of specialised processing chains that provide higher level value added science products. This is notwithstanding *the need for Tier 1 facilities to directly connect and be embedded within the scientific community at large*, providing an *open collaborative channel*. Such a layered and distributed approach will facilitate the inter-connection between (all level) data centres, facilities and the community.

4. [ARCHIVE] Archives for Data and Software

Key Points

- Archive-1: Pursue long-term investment in state-of-the-art archives for storage, exploration and exploitation of data associated with missions and facilities.
- Archive-2: Consider alternative storage solutions for data that require limited access.
- Archive-3: Keep data accessible and live where possible, facilitating and encouraging updates of methods/pipelines, metadata and associated usage examples.
- Archive-4: Consider direct access to on-the-fly user-tuned computation of data and models (e.g., IRFs), watching out for computational and environmental costs.
- Archive-5: Make development and preservation of software an integral part of astronomy portfolio (as it is for data), including efforts on the Open Science side (see Open-3), promoting associated achievements and careers (upgrading our crediting system, see Open-4), and appropriate long-term funding (including full-time tenure positions).
- Archive-6 Embed the existing and new data centres and archives, in a coordinated way, in the global inter-connected, interoperable data-and-computing infrastructure described above.

The need for state-of-the-art archives does not need much justification any more, considering the heavy and successful usage made by the astronomical community, thanks to the implementation of transparent, efficient and accessible archival channels. The proportion of archival data in the overall scientific production has significantly increased over the years. Archives that provide effective access to their science data products, are increasingly used for archival research, which can in turn lead to the acquisition of new observational or simulation data to further the initial investigations. While the implementation, curation and expansion of archives is a demanding task, astronomy is one among a few privileged sciences where extensive archive services have developed steadily and robustly over the years. Beyond the associated efforts and resources, their implementation benefited from the data standardisation and tools offered by the maturation of the Virtual Observatory.

In the context of data preservation, archives are no longer just a way to store and access raw data. They increasingly expose advanced products and implement interactive and visualisation tools, advanced search services to optimise the user experience and their scientific content and objectives have significantly evolved over the last decade. Moreover, those centralised archives are no longer the only interface by which scientists can harvest and connect their datasets. The concept of data preservation is broadening in connection with the increased need for sharing and

widespread distribution (see [Open] Section). All of these considerations imply that the preservation of data has become a multi-facetted challenge which calls for a coordinated, yet diverse set of approaches depending on the type and usage of data, i.e. raw, reduced, mock or simulated data. The following focuses on selected items within the broader picture, with interoperability, curation and documentation always being key requirements.

Raw data

In the coming decades, a few space and ground-based observatories will deliver raw and low-level data of unprecedented complexity and volume (at the petabyte/exabyte scale), while a good fraction of all raw data will have to be reprocessed multiple times. These datasets are expensive to acquire and potentially irreplaceable, so robust, backup (and possibly mirrored) data storage infrastructures are essential. Ideally, the HPC and HTC facilities that run associated reduction pipelines are co-located with the repositories for the data they process to avoid time-consuming network transfer.

It remains unclear whether universal access to these low-level data is feasible or even useful for all experiments and observatories. While this may sound like a provocative statement which could lead to restricted access to critical datasets, it is actually meant as a way to bring a fresh perspective on such data, already applied e.g., in the context of large-scale simulations, and to maximise the exploitation of existing scale-effective solutions. To provide usable open access to such large data volumes is technologically challenging and can be extremely costly, with corresponding requirements for network connectivity and bandwidth if processing close to the archive cannot be provided. Where possible and applicable, provision should be made for retrieval of data subsets, and technologies including data streaming should be considered to minimise local data storage requirements for the data users. For long term storage of low-level data with limited access, different technical solutions may be appropriate. If data are stored indefinitely then they must be curated, annotated with comprehensive metadata and potential hands-on usage tutorials to ensure that they remain interpretable.

Reduced Data

Open access to intermediate reduced data and post-analysis high-level data products is likely to be of a more direct utility for the astronomical community. Fortunately the volume of these data is (often but not always) expected to be much smaller than their low-level counterparts, and the technological challenges associated with publishing them are correspondingly reduced.

Online science analysis platforms are increasingly prevalent and such facilities plus other similar facilitating frameworks and services (including archives providing advanced services and tools) will likely become the most common means of accessing science-ready data. This is fortunate because the sheer volume of forthcoming datasets may make download and local storage of reduced data products impractical or unfeasible. Existing archives are also often not designed a priori to sustain an agile and flexible expansion towards new data products or frameworks. In principle high level data can be regenerated from their lower-level counterparts. However, the computational and environmental costs of doing so may be substantial, making redundant, mirrored storage cheaper and preferable in some cases. Data losses notwithstanding, revisions to data processing pipelines and improved instrument calibrations may necessitate regeneration of intermediate level data products and provision should be made for sufficient computational resources. To promote

straightforward reproducibility of scientific results, retention of used data pipelines and processes (on top of the raw data) using appropriate versioning plus metadata may sometimes be desirable. Those requirements also have clear ties with the development of dedicated services and new frameworks to bring software (and the user) to data (as opposed to bringing data to the user, see the [User-to-data] Section of the present document).

Software

Software development entails substantial investment of time and intellectual effort, making tools for reduction and analysis almost as valuable as the data they process. For the processing and reduction software stacks associated with large observatories, industry-standard software development and preservation practises should be adopted. In particular, effective use of version control, robust unit testing and comprehensive documentation are essential to ensure stable code that is correctly applied. However, it is clearly not the whole story, as version control does not ensure that software is preserved and exposed appropriately or serves its purpose.

The specific requirements of different science communities within astronomy result in the majority of high-level analysis tools being developed as required by the astronomers themselves. Unfortunately, for many individual researchers who develop software, there is currently little incentive to perform time-intensive tasks like writing documentation or unit tests. Software development activity receives little recognition in the astronomy community, which can impact individuals' career progression if it reduces their publication and citation indices. Failure to adopt good practises for software development, albeit for pragmatic reasons, leads to adverse effects including non-reproducibility of results if software evolves without proper versioning, misuse of tools producing spurious results if code is not documented and duplication of effort if code is lost, difficult to adopt, or never published. These issues can be mitigated by providing proper recognition for software development activities and training early career researchers in development best practises.

Another way to alleviate the software development burden on individual researchers is to encourage open software development principles and procedures (see the [Open] Section). Community-developed open-source frameworks like `astropy` or the `yt-project` provide excellent examples of how collaborative software development can promote reliability, useability and utility for high-level analysis tools. Frameworks like `astropy` aim to provide comprehensive documentation with usage examples, which reduces the chance that misapplied software will generate spurious scientific results.

In some cases, preservation of software source code may not be sufficient. The context in which software executes may also affect the results it generates. Containerisation frameworks (like Docker and Singularity) enable preservation and restoration of software within defined runtime environments to help achieve reproducibility for data analyses. Containerisation also allows software to be straightforwardly deployed on cloud computation infrastructure and container runtime engines have been developed that run on HPC infrastructure (e.g. NERSC Shifter). And as any piece of software, containers need maintenance efforts to be planned and executed when relevant.

Using metadata to associate software with the data that it is designed to process can also prevent its misuse. And again, detailed usage examples as well as dedicated tutorials should help preserve the expertise that is possibly the hardest part to safeguard on long timescales.

Preservation of software is thus closely inter-connected with efforts on the Open Science side (as it is in the case of data). This should definitely be recognised as an integral part of the scientific port-folio, and supported by robust and long-term funding, as opposed to solely individual-project-based and time-limited efforts.

Instrument Response functions

Several instruments including CTA and KM3Net will require large scale simulations to generate their instrument response functions (IRFs). These IRFs will likely evolve over the lifetime of the observatories and may change abruptly to reflect observing conditions or temporary reconfigurations of the instrument. Science analyses require IRFs that correspond to the epoch when the data were obtained, so it is important that all historical response functions are retained. If the simulated data that are used to generate the IRFs must be retained this could imply substantial storage requirements, increasing at a rate that depends on how regularly IRFs must be regenerated. In some cases it may be feasible to store only the reduced IRFs themselves and discard the raw simulation data, with corresponding storage cost savings. Emulators or trained deep learning models may provide a mechanism to compress simulation results without discarding them completely.

5. [USER-TO-DATA] The next Generation Tools: Cloud computing, Platforms, Collaborative frameworks

Key Points

- User-to-Data-1 Support researchers in addressing multi-source datasets, models and simulations, encourage the “bring the user to the data” cross-mission/facility approach, with expert centres, missions and observatories as engaged facilitators.
- User-to-Data-2 Encourage and coordinate the development of open source research-oriented platforms and analysis/collaborative frameworks close to data. Promote and support coordination of the fragmented computing infrastructures in the framework of the Tiered data-and-computing infrastructure described above.
- User-to-Data 3 Monitor and address dependence and constraints associated with private providers (potential “vendor lock-in”), using mirroring, diversification and easy re-deployment of cloud computing services (e.g., encouraging the development and usage of framework standards)
- User-to-Data-4 Imprint such developments in the long-term planning of infrastructures, with associated long-term investments coordinated with the missions and facilities.
- User-to-Data-5 Encourage cross-disciplinary efforts in the platform building processes, and promote and facilitate careers with professional software engineering and computing skills.

The computing landscape is changing both from the view point of astronomers performing large-scale simulations and those analysing data from large-scale observational infrastructures. In the case of simulations, the current codes can scale up and take advantage of new computing paradigms, such as accelerator platforms, and HPC facilities can provide more CPU/GPU flops. As the size of the data from these simulations correspondingly grows, the analysis, transfer, and long-term storage of the data is becoming a severe computational bottleneck. Increasing amounts and resolution of the observational data, on the other hand, did not before pose a real HPC challenge, but nowadays it increasingly does: to manipulate the increasingly large data sets,

memory and computing requirements bring this work to the realm of HPC computing. The challenges, goals, and the required infrastructure for both communities approach each other - the big data analysis and HPC computing borders are getting increasingly blurred. The current trend is computation becoming cheaper, but bandwidths of memory and data transfer are not growing at the same speed. Hence, any type of computation/data analysis tasks will become limited by these bandwidths, when the data sizes increase.

The implications of these trends are plentiful and important to be considered by the astronomical community itself, but also by the funding and mission-based agencies. Firstly, data at Petascale and beyond is too large to be "moved" around, hence users should be brought close to the data as opposed to the data being distributed (via downloads) to the users ("user-to-data" -policy). The increasing size of the data has the consequence that, even if it is hosted in an open access data storage facility, it would not be "open", as the capacity for downloading and utilising the data at local sites would be difficult if not impossible. It will also violate the principles of diversity, as some researchers from rich and therefore better-equipped institutes would have increased chances to exploit the data in comparison to researchers from poorer institutes. The policies for data access and the infrastructures built should better reflect the demands arising from large-scale data.

Instead of merely providing data access, the efforts should therefore concentrate on providing data analysis tools and resources on the site, where the data is stored. This necessarily means that in addition to the access, there should be HPC resources and additional software stack on site, which we refer to as "research-oriented platforms", so that any user could have access to the data, have basic data analysis toolbox available to operate on the data on site, and furthermore to be able to contribute to the development of the tools on the software stack. The structure of the platform can be envisioned to follow the tiered structure discussed before: the HPC resource level would be the lowest, invisible, layer to the community user, the data, software, and cloud services level the layers above, whereto the user would have access depending on the engagement level desired.

For similar reasons, addressing multi-source datasets together with models (and simulations) synergistically requires dedicated services and development frameworks with an emphasis on both robustness and adaptability. A global need to share analysis tools and allow further developments close to the data is emerging. Such frameworks have been and are being developed and implemented in various scientific disciplines, including Astronomy. The ESA Datalabs promote the idea of "bringing the questions to the data", while ESCAPE ESFRI Science Analysis Platform (ESAP) is meant as a "flexible science platform for the analysis of open access data available through EOSC". Pangeo (https://pangeo.io) is another example of a more global collaborative ecosystem with the Earth science community. Implementing and promoting cloud computing and collaborative research platforms may first help alleviate obstacles associated with downloads Most importantly, it will be key to allow efficient and innovative science derived from multi-source, multi-messenger, big-scale, and complex datasets. They are a promising thread to follow when it comes to environmental costs (see [Green] Section). Last but not least, we need to admit that we cannot predict the future: completely new attractive and relevant technologies, protocols, or ways to conduct our research may emerge. It is critical that we keep a live update on the requirements encompassed by our activities, associated with a pro-active technology watch.

Many challenges lie ahead if we wish to establish such a new (and modern) way of doing research. To be successful, it requires a broad coordination of the main actors in the community, including the research institutes and organisations, missions and observatories, to engage into and connect to that scheme. It also requires even more dedicated efforts to develop codes as open-source

community efforts, and efficient collection, cataloguing, and sharing of the developed algorithms and software. While platforms do exist, including a few dedicated to Astronomy (e.g. the Astrophysics Source Code Library, or ASCL), software is often deposited across various repositories that are either not well curated or exposed. Committing to these activities would enable them to focus on exposing and sharing their unique expertise at the service of science (Smith et al. 2020), to benefit from serving the Open science objective and FAIR principles. While open source frameworks should be encouraged, there may exist attractive technologies or services provided by the private sector: in such cases, the community at large needs to monitor, assess and address the potential risks associated with the dependence to such products (e.g., “vendor lock-in” in cloud computing). This calls for a diversification and easiness of deployment of cloud computing and platforms (possibly via established standards), as well as adaptable schemes for down-scaled datasets.

Some data transfers in between different sites are inevitably still required (and desirable). For this purpose, new tools exploiting, e.g. streaming software solutions, or better utilisation and further development of compression schemes, would be desirable. Obviously such work, constituting the design and implementation of complex and non-astronomical software, cannot always be done by the astronomical community alone. The same constraint applies to the build-up and maintenance of the research-oriented platforms; the thinking of building and funding only hardware is no longer meeting up the requirements of infrastructures resulting in/dealing with massive amounts of data. For such infrastructures to be useful, considerable human resources should be funded alongside the hardware. This is not the current trend in the infrastructure funding schemes, and poses a major challenge for astronomical infrastructures in the future. Special care should finally be taken to ensure that all parties have sufficient and fast network connections, as to prevent the exclusion of less privileged regions and institutions.

The HPC paradigms are in constant change. For some time, the astrophysical community has been preparing for the paradigm change from CPU computing to GPU and hybrid platforms, currently taking place in HPC facilities all over the world. Such paradigm shifts are not at all trivial for the adaptation of the community’s codes; this problem is only to become more pronounced in the Exascale and quantum computers. Reacting to these changes has taken up and will take up a lot of resources and attention from the principal function of the researchers in our community: producing high-quality science. AI is bringing forward great potential for data analysis tasks, and the HPC centres are also conforming their hardware to adapt to the increasing need of such tasks. The constant change will force us to soon tackle other challenging paradigms, again requiring a lot of resources (and dedicated skills) to develop novel algorithms, and perhaps even going back to old paradigms such as field-programmable gate arrays.

In reaction to the increasing prevalence of petabyte and exabyte-scale datasets, the astronomy community has rapidly adopted new paradigms for Big Data analysis. Prominent among these are Deep Learning techniques (colloquially referred to as AI) that can be used to discover complex patterns and relationships within and between high-dimensional datasets. The rapid development of Deep Learning techniques during the last decade was primarily driven by commercial companies and facilitated by very large, labelled datasets and novel computational architectures - most notably GPUs. In astronomy, large collections of reliably labelled data are less common which has made training *supervised* Deep Learning models more challenging, notwithstanding some notable successes. Innovative approaches like the *Galaxy Zoo* citizen science project have been used to collect labels for millions of astronomical images and more recently, astronomers have begun to explore methods that can be used to compensate for limited or absent data labels.

Domain-adaptation techniques have allowed highly realistic simulated data (with known ground-truth labels) to be used to train supervised Deep Learning models before fine-tuning with a small number of human-labelled real data. With increasingly large, high quality datasets, unsupervised and self-supervised techniques become increasingly feasible, and allow complex data representations to to be learned from *unlabeled* data. Deep Learning techniques will certainly be required to efficiently analyse data from next-generation surveys. The community needs to promote the skills and provide the computational resources (like GPUs and TPUs) that these techniques require.

The requirement is not only to produce code that can compile, run and scale up in the novel HPC platforms - the society has the demand of energy efficiency, and low carbon footprint (see the [Green] Section). Hence, we need to pay much more attention to producing optimal solutions of the data analysis and simulation tools. Plenty of know-how in ICT is similarly required in building the research-oriented platforms. Therefore, for both of these tasks, cross-disciplinary efforts in between astro- and ICT sectors would be strongly beneficial, and careers with professional software engineering / computing skills should be promoted and facilitated (see the [People] Section).

6. [GREEN] Green Data Infrastructures, Reducing the Carbon Footprint

Key Points

- Green-1 Quantifying, monitoring and exposing the carbon footprint of computing in astronomy should become a default for research institutes, organisations, computing centres (as part of a more global assessment of the impact of our research activities). We should implement a regular reporting exercise, coordinate the gathering of such data for the community, aim at clear objectives and specific recommendations.
- Green-2 The community at large should aim at minimising its energy consumption altogether while encouraging energy suppliers to invest in power generation capacity using carbon neutral sources.
- Green-3 Code optimisation should become a key both for the sake of efficiency and to minimise the environmental cost. Efforts should focus on the training of scientists and associated developers, promoting careers around code optimisation and adaptation.

It is high time that the astronomical community accepts that our activities, which form part of the everyday business of our scientific research, have a significant impact on our environment and contribute to the climate emergency that is the result of increased levels of greenhouse gases. Making the changes that form a meaningful contribution to addressing the issue come at a cost, both financially and in adaptations to the way we work, interact and collaborate. One side-effect of the Covid-19 pandemic has been the dramatic increase in working from home and reduction in travel, which have forced us to explore new ways of collaboration and have shown that at least some changes are possible. But we aren’t there yet.

The factors that contribute to the carbon footprint of our profession are diverse and have different relative weights. Those include (super)computing, travel, flights & commuting to work, the building and operation of our observatories, offices and facilities and should be addressed in their respective contexts (Burtscher et al. 2021).

Supercomputing as well as the distribution, archiving and retrieval associated with simulations and increasingly large data-sets from our observatories and space instruments play an ever increasing

role in modern astronomy. It forms a significant component of our energy consumption. Whether these activities also contribute to the carbon footprint depends on the way in which power is generated. Traditional carbon-fuel is still widely used, and therefore the move to renewable and sustainable energy sources would contribute to lowering astronomy's carbon footprint. The largest impact will be achieved if our activities lead directly to new or additional investments in power generation capacity using carbon neutral sources (e.g. wind, solar, tidal, geothermal). All institutes, universities and observatories are encouraged to continue their transition to non-fossil fuels and wherever possible ensure this also applies to the (super)computing centres, cloud providers, data centres as well as network providers that we work with. We should be mindful of the fact that the move towards a carbon-neutral future can have a negative impact on the fairness of access to computing resources. The move towards reuse of heat generated by computing and hot or warm water cooling also deserves to be pursued.

The astronomical community has, for many years, benefitted from the dependable and predictable increase in affordable and readily available computing (e.g. Moore's Law). This has coincided with a growing dependence on interpreted languages and little incentive to invest in optimisation and acceleration of codes. While many will benefit from the ease of using interpreted languages such as Python, the community needs to also maintain the skills in compiled languages.
We should keep in mind that compiled languages are still in use in the libraries and routines that are called from interpreted languages. There are many optimisers and transpilers available now that can be used to increase efficiency (e.g. numba, cupy) and users are encouraged to continue to contribute to valuable community resources such as `astropy` and `numpy`, while the emergence of new languages may ease the transition towards more efficient interpreted languages.

As new hardware solutions emerge, codes require demanding refactoring which is often challenging to deploy due to the lack of dedicated expertise, resources or a lack of awareness. Scientists should be trained in the general optimisation of codes both for the sake of efficiency and the gain in environmental cost. Such optimisations should also be incentivised by crediting such efforts as part of the research port-folio, and careers in general. The implementation will require this expertise to be deployed at the service of or within the scientific collaborations, which itself needs robust and long-term funding and acknowledgements from academia and funding agencies.

Providing specific recommendations goes admittedly well beyond the scope of the present report. We must acknowledge that we are still lacking the basic tools, processes and information to quantify the present state of affairs, establish concrete objectives, and report on our progress (Henderson et al. 2020, as an illustration). The above-mentioned focus on the global environmental cost of computing in Astronomy desperately needs to be reflected in **concrete**, **quantified** and **controllable** actions, within a more global framework addressing the impact of our research activities. This in turn requires proactive monitoring and regular reporting exercises fed by all intervening actors in a guided and transparent process.

7. [People] Training, Careers, People

Key Points

- People-1 Establish and promote metrics to encourage data scientists and data stewards, HPC and software experts to pursue careers within Astronomy, and exploit the attractiveness of astronomy to recruit and promote talented staff.

- People-2 Encourage multi-disciplinary collaborative programmes fostering long-term human resources pertaining to software, coding, ICT and HPC, favour mobility, value and pro-actively increase the diversity of the workforce.
- People-3 Promote the (computing/software/ICT) existing talent of those who are actually educated in an astronomical context, nurturing the combined motivation for and understanding of the field.
- People-4 Fully integrate professional software and computing engineering career paths within our research scheme/system and promote them as part of the “astronomy” world.

While software (and computing in general) is increasingly seen as a critical ingredient of e.g., new astronomical facilities, and more generally of scientific endeavours, the skills and work of the associated developers (engineers, scientists) are too often hidden behind project or facility fences, and are neither actively promoted nor even exposed. The engagement of missions towards Open Science (see [Open] Section) may partly help resolving this unbalanced situation, but there is still a long way to go. More alarming is the fact that such a situation may be prolonged due to our established perspective towards computing and software, and becomes unsustainable in the face of the novel ways we are conducting scientific work. Given the increasing numerical challenges the astronomy community are facing, there is a dire need for professional and coordinated support in terms of computer scientists, software and hardware engineers, data stewards that combine both ICT and astronomy/astrophysics expertise.

In particular, multidisciplinary university studies and diplomas at the crossroad between astronomy and informatics should be fostered, including the definition of key specialisations for future astronomical computing. Careers for these new profiles, from HPC experts to data and software scientists, in astronomy, should be properly defined, made attractive, promoted and rewarded with dedicated career paths and metrics. They should be integrated at various levels of the tiered approach, within the research institutes, the data centres, computing facilities, missions and observatories and be part of multi-disciplinary teams with complementary skills. Such expertise should be active in community forums, and exposed in relevant committees, panels, jurys and executive authorities.

One effective path is to identify and promote the internal talent, already present within the astronomy research and education ecosystems, that has computing expertise. This approach may ease the challenges of securing long-term funding and career development when such experts can grow within a fixed ecosystem and their work is thus fully acknowledged as a part of the collaboration entity. While such a path may look attractive, it certainly cannot fill in the required expertise for the development, implementation and maintenance of professional computer-related efforts. Another complementary and possibly central approach is to connect with external computing experts, or create active trans-disciplines bridges. While this requires a dedicated effort to share the relevant astrophysics-related research objectives, it does favour mobility and nurture the diversity of perspectives, a key catalyst for creativity and scientific progress. More generally speaking, there is a critical need for increasing the diversity of the workforce, better reflecting society in its richness. Building up the next generation of computing experts staff in Astronomy (and science) is a unique opportunity to fully embrace such a guiding principle.

Coordinated efforts should also be encouraged and supported to specifically address computing challenges in astronomy. This includes the development of multidisciplinary collaborative programmes and the mutualisation of ICT human resources on common European-level projects. For instance, to adapt existing and widely used codes to new architectures in the upcoming

exascale era so that it is efficient (see [Green] Section of the present document) can benefit the whole community and avoid duplicating the efforts in a sub-optimal way.

Long-term investments at all levels of astro-computing are thus highly desirable. Significant funding should be invested for computing-related FTEs in astronomy, not only for local and punctual support on specific projects but also at the national and European level, within the facilities, the research institutes and academic world. Those careers should be pro-actively sponsored, recognised and rewarded as part of the astronomy port-folio, benefiting and committing to the appropriate exposure (e.g., conferences, panels, papers).

8. References

- ★ Borgman, C. L., & Wofford, M. F. (2021), From Data Processes to Data Products: Knowledge Infrastructures in Astronomy · Harvard Data Science Review
- ★ Burtscher, L., Dalgleish, H., Barret, D., et al., 2021, Nature Astronomy 5, 857
- ★ Chang et al. (astro2020 white paper): [1903.04590] Cyberinfrastructure Requirements to Enhance Multi-messenger Astrophysics
- ★ Desai, V., et al. (2019). A Science Platform Network to Facilitate Astrophysics in the 2020s. *Bulletin of the AAS*, *51*(7). Retrieved from https://baas.aas.org/pub/2020n7i14
- ★ Henderson, P., et al. (2020) Towards the systematic reporting of the energy and carbon footprints of machine learning. Journal of Machine Learning Research 21, 1, https://jmlr.org/papers/v21/20-312.html
- ★ Patterson, D. et al., (2021), Carbon emissions and large neural network training. Preprint at https://arxiv.org/abs/2104.10350
- ★ Smith, A. et al. [1907.06320] Astro2020 APC White Paper: Astronomy should be in the clouds
- ★ Smith A. et al. [1907.06981] Astro2020 APC White Paper: Elevating the Role of Software as a Product of the Research Enterprise
- ★ Wilkinson, M., Dumontier, M., Aalbersberg, I. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 3, 160018 (2016). The FAIR Guiding Principles for scientific data management and stewardship | Scientific Data

Acknowledgements

The Panel would like to warmly thank all contributions and feedback received during the course of the writing process. More specifically, we would like to acknowledge input from: Dominique Aubert, Frédéric Bournaud, Wolfram Freudling, Olivier Hainaut, Bob Mann, Marco Molinaro, Ralf Palsa, Felix Stoehr, Jean-Pierre Villote, John Swinbank.

APPENDIX I - Past Astronet recommendations

The past Astronet exercises (2010 / 2014) led to a series of recommendations which paved the way for and supported a number of key evolutions and implementations in Astronomy, as illustrated below (grouped by topics).

ASTRONET-1: NETWORKING AND PREPARING DECISIONS ON MAJOR INFRASTRUCTURES

Virtual Observatory

1. Provision of a public VO-compliant archive should be an integral part of the planning for any new facility. We recommend that data centres provide science-ready data.
2. Providers of astronomical tools should make them VO-compliant so they can easily talk to other VO tools and can be accessed within the VO environment.
3. The infrastructure established with EC support will need to be sustained by the national funding agencies to allow continuity of the VO.
4. The development of the VO should be coordinated with evolution of the generic e-infrastructure, and that evolution should reflect the domain-specific needs of astronomy.
5. To prepare for the challenges posed by large surveys, multi-wavelength astronomy and the VO, modelling codes need to be made modular.
6. Substantial investments are required in software that simulates mock data with the observational biases inherent in current and future facilities. Publication of such software in VO-compliant form should become an integral part of the construction of any instrument.

Simulations

7. Given the growing importance of sophisticated simulations for the future of astronomy, funding of theory must not fall far behind that provided for observational facilities.

Codes

8. Increasingly astronomy will depend on codes that are too complex to be written from scratch by students and postdocs, and astrophysicists throughout Europe must have access to state-of-the-art standard codes. These codes should be regarded as essential infrastructure on a par with major observational instruments.

ASL

9. A laboratory without walls called the Astrophysical Software Laboratory should be established to coordinate and fund software development and support, user training, and to set standards. Training and development funding would make it possible for codes to remain at the cutting edge of the field for extended periods. Development funding would also ensure that supported codes conformed to modular standards; the ASL would be the catalyst that enabled the community to establish these standards.
10 Code authors supported by the ASL should be committed to the open-source model.
11 The ASL would have an important role in nurturing the next generation of theorists and codes, both by funding postdoctoral positions within a programme of pan-European networks, and by supporting the development of innovative codes.
12 The ASL committee will select a few highly competitive astrophysics projects each year to send proposals to the European pan-science top-tier computers; this will ensure a significant share of CPU hours at the petascale level for astronomy.
13 The human resources required for the ASL are estimated at 50 FTE/yr. This number includes scientists who are already funded at the national levels, plus a core of researchers (estimated at about 20 FTE/yr) to be funded at European level, and who will be responsible for the ASL's activities and organisation. The ASL should be financed by the national agencies: a specified percentage of each agency budget should be reserved for it.

HPC, Grids and data links

14 Astronomy should continue to benefit from HPC all-science centres, and share the efforts to develop and increase continuously their performances in order to be at the forefront of the international competition.

15 The development of the top-tier HPC centres should not slow down that of the lower tiers: the whole pyramid of computers at different scales, national and local, is absolutely necessary to satisfy all computing needs.
16 Astronomy must exploit the grid infrastructure more widely, and contribute to the expansion of the capabilities of its middleware, in particular for data processing.
17 Data links within Europe and to the outside world need to be kept abreast of advances in technology. The VO is likely to require a different network architecture from that put in place for LHC science.
18 The possibility of using billions of otherwise idle processors for scientific calculations is now real, and could revolutionise data modelling. Astronomy should lead the way in this area, either by exploiting its popular appeal to get CPU owners to donate spare CPU cycles, or by initiating a classical market in such cycles. The ASL could possibly coordinate this activity, which could have a significant commercial spin-off.

ASTRONET-2: IMPLEMENTING THE ROADMAP AND PREPARING THE FUTURE

ASL
1: the ASTRONET Board needs to determine the status of the Astrophysical Software Laboratory in the near-future

Investment in data facilities
2: there is a need for continued investment in dedicated data facilities across Europe to keep pace with the data increase.

APPENDIX II - Relevant roadmaps

Astronet Reporting (2010/2013/2014) = https://www.astronet-eu.org/forums/roadmap-community-consultation
ESFRI Roadmap 2021 = https://www.esfri.eu/esfri-roadmap-2021
ESFRI Roadmap 2018: http://roadmap2018.esfri.eu/media/1066/esfri-roadmap-2018.pdf

EIROforum IT working group, . (2013, November 8). e-Infrastructure for the 21st century.. http://doi.org/10.5281/zenodo.7592

EOSC Strategic Research Innovation Agenda - https://www.eosc.eu/sites/default/files/EOSC-SRIA-V1.0_15Feb2021.pdf

EOSC Partnership - https://ec.europa.eu/info/sites/default/files/research_and_innovation/funding/documents/ec_rtd_he-partnership-open-science-cloud-eosc.pdf

Canadian Long-Range Plan: https://casca.ca/wp-content/uploads/2021/04/20UOT001_CASCA_LRP_EN_vFA2.0.pdf
US Decadal Review paper: A Science Platform Network to Facilitate Astrophysics in the 2020s: https://baas.aas.org/pub/2020n7i146/release/1
Lincei (2009): http://archives.esf.org/index.php?eID=tx_nawsecuredl&u=0&g=0&t=1620143683&hash=e878d8cc91f80d3229d9fa93342077b181ef7de4&file=/fileadmin/be_user/research_areas/PESC/Documents/FLOOKS/FL-Lincei-FINAL%20VERSION.pdf

INSU/FR (2019):
https://www.insu.cnrs.fr/sites/institut_insu/files/news/2021-04/Prospective_INSU_AA_2019.pdf

APPENDIX III - Existing infrastructures and coordinating efforts

- AERAP - Africa-Europe Science and Innovation Platform
- DPAC - Data Processing and Analysis Consortium - Gaia - Cosmos
- EOSC - European Open Science Cloud
- EOSC Hub
- ERIC - European Research Infrastructure Consortium | European Commission
- ESCAPE | The European Science Cluster of Astronomy & Particle physics ESFRI research infrastructures
- ESFRI - The European Strategy Forum on Research Infrastructures
- EuroHPC - European High Performance Computer Joint Undertaking
- EuroVO
- IRIS – Digital research infrastructure for STFC science
- Pangeo — Pangeo documentation

APPENDIX IV - Examples of science platforms and collaborative frameworks

- ESCAPE ESFRI Science Analysis Platform
- SciServer – Collaborative data-driven science
- STEP - Scientific Toolbox Exploitation Platform
- ESA Datalabs
- CosmoHub
- Galactica - The COAST group simulation DataBase

Glossary

- CPU: Central Processing Unit
- DPC: Data Processing Centre
- FAIR: Findability, Accessibility, Interoperability, and Reusability
- GPU: Graphics Processing Unit
- HPC: High Performance Computing
- HTC: High Throughput Computing
- ICT: Information and Communication Technology
- TPU: Tensor Processing Unit
- VO: Virtual Observatory

Panel

- Mark Allen (CDS, Strasbourg University- FR)
- Sandrine Codis (AIM, Paris-Saclay University - FR)
- Hugh Dickinson (Open University - UK)
- Eric Emsellem (Chair) (ESO, Garching - DE / CRAL, Université de Lyon - FR)

- Michiel van Haarlem (ASTRON - NL)
- Maarit Käpylä (Aalto University - FI)
- Agnieszka Pollo (Co-Chair) (NCBJ & UJ - PL)
- Nicholas Walton (IoA, University of Cambridge - UK)