Information Science and E-Infrastructure Challenges in Astronomy

Authors
Kai Polsterer (HITS), Harry Enke (AIP), Torsten Enßlin (MPA)
Contributors
Matthias Bartelmann (ITA), Benjamin Hoyle (LMU), Jochen Weller (LMU)

1. Executive summary

Astronomical research is driven by advances in observational techniques. The resulting data stream is currently growing rapidly in size and complexity, providing new opportunities but also challenges for astronomy. Data, software, and scientific publications should be exchanged openly. To maximize the benefit of future, present, and past datasets for the German astronomical community the following three challenges require attention in the next decade:

The information challenges: The process of how data is turned into knowledge is currently undergoing a dramatic change. Machine learning and Bayesian inference are on the rise, and the increased compute and storage capabilities make them feasible. Information theory provides the conceptual framework to understand and design optimal inference methods as well as data compression schemes required for handling the immense future data rates. The emerging information science has promises and challenges. The German astronomical community needs to participate in this interdisciplinary development to ensure its future innovativeness.

The infrastructure challenges: Data has to be curated: storage, indexing, annotation, sharing, visualization, processing, archiving, documenting requires an adapted and easily accessible digital infrastructure for Open Science. Derived data products, the involved software, and related publications have to be curated and more tightly coupled as well. The significant progress of the last decades in terms of a general scientific data infrastructure by service providers like computing centers and the astronomy-specific build-up of an interoperable software and data ecosystem by the Virtual Observatory alliances require a continuation of efforts and a clear communication of the community needs.

The societal challenges: The German astronomical community has to catch up on the information age. The academic curriculum should encompass elements from statistics (probabilistic inference, uncertainty quantification) and computer science (efficient programming, data handling). Training and consulting programs should facilitate the usage of the available e-science infrastructure like web services for data, software, and computational resources. Sharing of data, software, and infrastructure should be encouraged and supported to ensure a vibrant scientific community. Astronomers are on the forefront of Open Science and Open Access and it is in the best interest for the field to invest in these directions further.

2. The information challenges

2.1 Machine learning

Machine learning methods have entered a golden age in computer science, and advancements appear at an accelerated rate due to both improvements in algorithms and the gigantic increases in computing power. Astronomy has to adopt machine learning methods and to participate in their further development in order to cope with the unprecedented rise in astronomical data volumes, complexities, and heterogeneity. Machine learning can be regarded as a toolkit, and as such, all of its tools have their specific advantages and disadvantages that need to be studied,

characterized, understood, and applied. In particular, machine learning provides efficient tools to explore and characterize relationships between data in high-dimensional datasets. This allows machine learning to be applied to such problems as the construction of star- galaxy- and quasar catalogues, photometric redshift estimation, the clustering of galaxies by type, and the identification of rare strong lensing systems. Machine learning methods are unavoidable when attempting to handle and steer increasingly complex astronomical instrumentation and their data outputs. Many machine learning algorithms have strong links to the theory of networks and statistical non-equilibrium physics.

However, some of the machine learning methods contain many millions of tuneable parameters and are heuristic in nature. This means they often appear as black boxes to their user. An understanding of their reliability and predictive power is currently obtained on a case-by-case base for each astronomical application. How different machine learning methods represent knowledge also needs to be both better understood, and quantified for astrophysics purposes. Such an understanding will arise, and already does, from the investigation of machine learning in the light of network theory, non-equilibrium physics, and information theory.

2.2 Information theory

Information theory, is the language describing inference under uncertainty. Formally, it is the extension of the Boolean logic of *true* and *false* to the realms of uncertainty as captured by probability theory in its Bayesian interpretation. Despite the existence for more than sixty years of a formal proof for probability theory being the only consistent (real valued) extension of logic, i.e. the Cox theorem, the adoption of probabilistic or Bayesian methods has been slow in the past, partly due to their high computational demands. This is currently changing.

Bayesian methods demonstrate to be most efficient and accurate in many astrophysical areas ranging from planet detection, stellar and galaxy spectroscopy, categorization and characterization of objects, and many other areas of astronomical data interpretation. One of the past obstacles inhibiting general application of Bayesian methods was the lack of user-friendly software tools and the computational demands. The last decade has partly changed this, with the emergence of generic sampling methods to tackle complex inference problems which are more and more adopted by the astronomical community. This research continues to ramp up and it is essential that the German astronomical community actively participates in it.

Imaging, the conversion of raw instrument data into an image suitable for human inspection as well as for the analysis by machines, e.g. by the automatic construction of catalogs from image data, can be seen as a Bayesian procedure. The instrument data is usually too incomplete, noisy, and poorly calibrated to fully determine all degrees of freedom of an image. Additional information on the measurement process, as well as on the morphology, spectrum or other properties of the observed objects have to be added to turn the data into an usable image. In the past, this has mostly been done by ad-hoc procedures or by regularized inverse problem solvers. Both, ad-hoc methods and regularization can be interpreted as ways to inject prior information on the signal into the inference. With the current advance of Bayesian methods, the process of inclusion of prior information was made explicit, providing us with imaging methods tuned towards specific signals and purposes. This is setting the standard now, and developments of inference schemes need to take note of this standard. For example, the German community developed such Bayesian imaging software for radio-and gamma-ray astronomy and demonstrated it to be superior to traditional algorithms.

2.3 The next innovations

Astronomy with its needs for higher sensitivity, resolution, and fidelity has always been a cross-disciplinary technology driver. This is also true for the information science and infrastructure research. The German astronomical community would benefit over-proportionally from joining the long ongoing efforts in other disciplines, like artificial intelligence, and in strengthening the efforts arising from its own community, like the usage of field theoretical methods. Examples of such efforts can be found in experimental particle physics (filtering and compression of huge data streams), medical imaging (multidimensional datasets to be imaged using complex models), automatic image processing (industrial process control, brain science), remote sensing (meteorology, geology), and humanities (analysis and categorization of textual and non-textual data like music and images).

A vigorous participation of astronomy in this area could be achieved by establishing funding lines to initiate high risk research on astronomical information science, and to consolidate the results in case of success. Below, we list several high-potential innovative research directions of the German astronomical community from which other disciplines could benefit as well.

Artificial intelligence: With the current innovation speed in machine learning, the transition to a wide spread usage of artificial intelligence systems might not be far. The German astronomical community should actively participate in this research in order to join the forerunners of the technological application. As machine learning and artificial intelligence methods are not unproblematic in their application to personalized sociological data, astronomy is an ideal playground to investigate the potential, drawbacks, and possibly necessary regulations of such methods without risking direct harm to humans.

Field theory: Bayesian inference and the probabilistic interpretation of data is closely related to well-developed disciplines of physics. The close connection between statistical physics and information theory can be traced back at least to Shannon and Jaynes. Statistical field theories, for quantum or classical systems alike, are structurally very similar, if not identical, to mathematical representations of the problem of Bayesian inference from large data sets, as provided by the novel branch of information field theory developed within the German astronomical community. Similarly, close conceptual connections are revealed between the statistical physics of spin systems and the gradually emerging understanding of the deep or shallow networks used in artificial intelligence, automatic pattern recognition and other branches of machine learning. These relations are now being uncovered and addressed in joint, intense efforts between applied mathematics, statistical physics and computer science. It will be essential for astronomy to join this effort and possibly to recover its formerly leading role in data analysis.

Novel simulation schemes: Computer simulations are an essential data source for understanding astrophysical systems. Germany has a high profile in the area of astrophysical computer simulations. To maintain this leading role, the next set of intelligent simulation schemes should be developed in Germany. These will include, among others, data assimilation schemes that re-simulate our own Universe or estimate unknown parameters of the dynamics, uncertainty-aware simulations, information-aware schemes that refine computational grids to optimize their prognosis accuracy or to incorporate sub-grid information. The methodology for an informative comparison of our detailed simulations and precise observations needs to be developed.

3. The infrastructure challenges

3.1 The data age

Astronomy has always been based on observations. New techniques in instrumentation and telescope design together with large-scale simulations have led to a dramatic increase in available data. Not only the volume, but also the complexity and variety of the data and its production rate are just some of the future challenges to deal with. A deep understanding of the astrophysical processes in astronomical objects can only emerge from a combined and coherent analysis of the full information available in the multi-dimensional data space. Going to multi-wavelength analysis and exploiting the time domain are just two examples. In the past, we could see a boost in publications and an extended use of observational data by providing those through publicly available archives. Survey astronomy is no longer just providing plain catalogs. Archives are not just used to reproduce scientific results. Publications in (refereed) journals available on-line are more than just an addition to the printed versions. Simulations provide insight into unobservable regimes of astronomical objects. Preserving the data legacy, but more importantly, enabling science exploration and exploitation of the current and future data flood require technological and sociological adaption. Astronomy is evolving from an instrumentation, observation, and theory driven research field into a data driven science. We need to catch up to a vigorous evolution taking place elsewhere. Four major aspects have been identified for the German astronomical e-infrastructure to be addressed to be able to internationally compete in the future: Technology, Infrastructure, Software, and Methodology.

Many of these aspects are generic for many disciplines of science and are addressed by ongoing general developments on e-infrastructure, driven by computer science, applied mathematics and computing centers. The German astronomical community needs to communicate its priorities, to inspire these developments and to liaise with them. Some of these aspects, however, are specific to the astronomical community and need to be tackled within that community.

3.2 Technology and Infrastructure

Data-driven science would strongly benefit from an infrastructure that supports all stages within the process of knowledge discovery. These start with the retrieval and storage of data and meta-data from observations, simulations, publications, and of higher level data products like data-releases, catalogs, and value added data. An ideal infrastructure would support processing, analysis, and inspection in a manageable way. This would enable reprocessing of data and therefore make results reproducible. Computationally expensive and data-access intense experiments could be seen as a different kind of observation. These could be performed at generally available data-science facilities, which grant access to the German astronomical community and other data intense scientific communities.

For a data driven astronomy, an e-infrastructure is needed that is capable of storing observational data in all its occurrences, including raw data, if applicable, intermediate data products, and protected data that might become public after a certain time.

In astronomy, all data dimensions like location, photon or particle energy, timing, and polarization are of certain relevance for the scientific interpretation. Not only the data needs to be available, the necessary meta-information keeping the data provenance must be accessible in an open electronic form. This is true for derived higher level

data products, like object catalogs and calibrated images, as well as for raw data from instruments. In particular the raw data are of long lasting value, as they contain the full information from which the derived data products can always be regenerated, if the necessary algorithms were documented and kept in accessible software repositories. The true value of raw data pays off whenever more sophisticated data analysis methods become available, when this data is re-analyzed in combination with other data sets, and whenever a historical record of some area of the sky is needed in order to help classifying a transient phenomenon.

Data products generated by simulations already play and will play an even more important role to understand physical processes. Hence, the same infrastructure that handles observational data should also handle simulation results. For both types of data, post-processed products, annotated information, scientific / technical publications, and software packages should be stored and linked to the data, respectively.

To properly support researchers, information on data versions or different releases must be preserved. In addition, the combination of catalogs from different sources would benefit from an infrastructure to jointly store and process catalogs as well as from interfaces and standards to uniformly access the primary data and catalogs. A thorough process of data curation is necessary, collecting and providing as much meta-data as possible. This includes technical data of the used instrumentation / parameters and setups of simulations, footprints of surveys, good statistical data describing the noise / likelihood functions / uncertainties if applicable. Essentially, this means that publications based on publicly available data, using publicly available software, also need to be publicly available. This implies that this type if Open Science is impossible without Open Access, and whatever links to it. A very important challenge for the future will be to provide accurate provenance data. This will enable scientists to validate the process the data was generated in. The international Virtual Observatory has already done much work towards making astronomical data digitally accessible and usable in unified formats. Such efforts need continuation.

The German astronomical community has to deal with several challenges that can just be solved as a community effort. It has to develop strategies to professionally host huge amounts of data, build and operate storage facilities as well as archives. Remote visualization of the huge data-sets will play an important role as the current way of downloading and locally inspecting the data is failing. Thus, technologies and infrastructure installations are needed that support the required data processing, accessing, and transferring rates.

The German astronomical community therefore strongly encourages the build up of data science centers and strongly supports the already started efforts of institutional and national computing centers to develop into this direction. The German Astronomical Virtual Observatory community, coordinating with the international Virtual Observatory, is well prepared to participate in the process of making such centers interoperable and well usable for astronomers by providing the necessary standards for data, software, and service annotation.

3.3 Software and Methodology

Cutting edge data science requires an environment to develop, test, and exchange the software implementing its intelligent algorithms. Methodological advances are required to efficiently deal with the increasing complexity and sizes of present and future data sets. For time-domain astronomy completely new approaches have to be

developed. Heterogeneous sampling, heteroscedastic errors, missing and censored data require information theory based algorithms, since missing information has somehow be augmented. For some tasks, online and incremental learning approaches might be required. Dedicated pipelines for pre- and post-processing the data will be required, in addition to those which are integral part of an instrumentation project.

Software development leading to tools and libraries becomes as important as building instrumentation. In the past, discoveries were often enabled through new developments in observation techniques and instrumentation. Along with more powerful instruments producing a plethora of data, the techniques for data mining acquired a new role for the machinery of discoveries.

Scalability and availability become more important with the new demands of data access and data processing. Available software tools should be annotated and registered with meta-information denoting their hardware and infrastructure requirements. This would permit the distribution of such tools as a service in cloud based environments or data centers. Applications that do not require an intensive installation and configuration work will be adapted by a larger number of scientists. An open access to those tools, developed through publicly funded projects can be more than the pure availability of source code. Tools must be easily accessible and searchable via maintained registries, otherwise they will not be used within the community. Efforts to build up such registries like the Astronomical Software Code Library, should therefore be encouraged and supported.

Innovations usually are made by individuals and small groups of scientists. Ways to finance the customization of novel algorithms for the general user would permit to harvest the intellectual capital of the community. Coordination is needed for the development of standards which are mandatory for data-formats, data-access as well as protocols to enable software tools to exchange information and to interoperate. On an international level, the worldwide developments are standardized and coordinated through the international virtual observatory alliance (IVOA). This institution is also responsible for reporting back to the IAU. On a national level, the work of the German Astronomical Virtual Observatory (GAVO) has to be continued. National coordination of standardization and interoperability aspects are just some aspects that have to be continued. We need competence groups to support the community with training and consulting on how to make use of the available standards and tools. Most important is to have an independent instance to ensure that a longterm strategy is being pursued.

4. The societal challenges

The ongoing transition of astronomy towards a data and information science, which takes full advantage of its emerging e-infrastructure, requires sociological changes through education and training, a commitment to Open Science, and a modification in the way work is acknowledged and funded.

4.1 Education and training

The astronomical curriculum should reflect that astronomy is a data-intense science. A joint effort across astronomy, physics, computer science and possibly mathematics and statistics is needed to include advanced inference into the general physics curriculum. This would embrace the theory and practice of signal processing, data compression, pattern recognition and anomaly detection. To set up such a curriculum would require a coordination with other disciplines and initiatives, in particular the

ongoing establishment of data science curricula at German Universities. It would probably cover:

- Foundations and applications of Bayesian inference
- Advanced image processing and filtering methods
- Time efficient programming and optimization
- Usage of compute infrastructures and cloud services
- Basic tools of machine learning

Additionally, the astronomical curriculum should include more of the available and growing set of data mining tools.

New job-profiles are expected to emerge in astronomy. Technology experts are required to implement and follow new trends and developments. Data-scientists are necessary to provide expertise with analyzing large and complexly structured data-sets as well as working on new methodologies and algorithms. Data-managers will be required to preserve, curate and organize the data that is produced by instruments and simulations as well as higher-level data products like catalogs. Software experts have to be part of the community to guarantee a certain level of software projects as well as the proper implementation of algorithms.

4.2 Open science

An important aspect of E-Science concerns new ways of publishing. Driven in part by pressure from funding agencies, Open Science becomes the leading paradigm. In parallel, scientific demands are growing for raw or derived data and at least rudimentary analysis software to be made available together with publications in order to increase the transparency of research and facilitate the reproducibility of scientific results. At the same time, animations and videos quickly become part of scientific publications as purely electronic publication formats spread.

The Public Knowledge Project[1] impressively demonstrates what developed infrastructure for community-based publication already can do. Particularly in view of quasi-monopolistic, commercial publishers, it will be important for astronomy like for all other scientific areas to seriously consider electronic publishing of papers, figures, data, software, animations, videos and possible other formats in its own responsibility to maximize their benefit for science and the public. Establishing an approach to new forms of publishing as a community within the next decade should be a strong point on the EScience agenda in astronomy.

4.3 Acknowledgements, Policies and Funding

Data curation, data publication, software development for astronomy, community work on standardization, maintaining of infrastructures, support of software tools are no longer menial work but necessary part of the scientific process in astronomy. This has to be acknowledged to acquire and keep excellent personnel in this highly competitive market. Furthermore, acknowledgement and financing for those that train and support the community in the usage of e-infrastructure and information science is needed. With consultants and support assistances available locally, comparably to the ACI-REF program of the NSF, the investments in a modern e-infrastructure will give a good return.
Acknowledgement of these activities may have different forms: for data curation and publication, for software development and publication the community needs to find suitable forms for review and reward. Also, there are no suitable metrics for these

1 https://pkp.sfu.ca/

contributions to the scientific process[2], and without push from the community, none will be considered. Correspondingly, the Research Funding Organisations (RFO), and evaluation procedures from Research Performing Organisations (RPO), e.g. MPG, Leibniz Association, HGF, rarely appreciate these contributions. This is not specific for astronomy, but the initiatives to change this should get stronger support.

Most of German RPO have recently updated and enhanced their policies for Open Access. The astronomy community has already facilities like arXiv available, but these do not count for scientific reputation building, which limits greatly their usefulness. The pre/postprint servers like arXiv need full support, also DataCite as a way to identify published data sets. We will need registries and repositories like ASCL for citing and publishing the software that was used in a publication. Like for scientific papers, the scientific data and the software needs improved procedures for validation and identification. Open Access is part of a transformation to take back full control of the results of the research by the science community. It should be vigorously supported by implementing Open Access in the institutes of the astronomy community.

5. Conclusion: Key infrastructure needed for researchers in Germany

5.1 Information science

Initiate a cross disciplinary effort on information science and applications in data-intensive areas of science with a strong participation from the astronomy community. Establish funding opportunities for innovative, risky research on machine learning as well as information theory and a DFG priority program for this field. Ways to fund the customization of prototypic methods would ensure the sustainability of the investments.

5.2 Information infrastructure

Consolidate exisiting EScience groups in astronomy and build these up into competence centers around astronomical data centers for the support of the community in the usage of the emerging e-infrastructure and developing tools for more efficient scientific work with the digital data and providing the environment for Open Science. It provides also a sustainable way for supporting the German Astrophysical Virtual Observatory (GAVO) and its contributions to the international Virtual Observatory.

5.3 Education and open science

Initiate and support cross disciplinary effort for a revision of the curriculum in the natural sciences to include the requirements of data intensive research.
Develop modules for the curriculum of astronomy which reflect the data intensive methods and tools.
Support Open Science initiatives on each level, especially in the astronomical institutions.

[2] See the statements on “Kerndatensatz-Forschung”, e.g. “Stellungnahme_leibniz-data.pdf”, Ausschussdrucksache 18(18)86 d, p. 2