Astro2020 State of the Profession Consideration White Paper

Realizing the potential of astrostatistics and astroinformatics

September 27, 2019

Principal Author:

Name: *Gwendolyn Eadie*[4,5,6,15,17]

Email: eadieg@uw.edu

Co-authors:

Thomas Loredo[1,19], Ashish A. Mahabal[2,15,16,18], Aneta Siemiginowska[3,15], Eric Feigelson[7,15], Eric B. Ford[7,15], S.G. Djorgovski[2,20], Matthew Graham[2,15,16], Željko Ivezić[6,16], Kirk Borne[8,15], Jessi Cisewski-Kehe[9,15,17], J. E. G. Peek[10,11], Chad Schafer[12,19], Padma A. Yanamandra-Fisher[13,15], C. Alex Young[14,15]

[1]Cornell University, Cornell Center for Astrophysics and Planetary Science (CCAPS) & Department of Statistical Sciences, Ithaca, NY 14853, USA
[2]Division of Physics, Mathematics, & Astronomy, California Institute of Technology, Pasadena, CA 91125, USA
[3]Center for Astrophysics | Harvard & Smithsonian, Cambridge, MA 02138, USA
[4]eScience Institute, University of Washington, Seattle, WA 98195, USA
[5]DIRAC Institute, Department of Astronomy, University of Washington, Seattle, WA 98195, USA
[6]Department of Astronomy, University of Washington, Seattle, WA 98195, USA
[7]Penn State University, University Park, PA 16802, USA
[8]Booz Allen Hamilton, Annapolis Junction, MD, USA
[9]Department of Statistics & Data Science, Yale University, New Haven, CT 06511, USA
[10]Department of Physics & Astronomy, Johns Hopkins University, Baltimore, MD 21218, USA
[11]Space Telescope Science Institute, Baltimore, MD 21218, USA
[12]Department of Statistics & Data Science Carnegie Mellon University, Pittsburgh, PA, USA
[13]Founder, The PACA Project, Space Science Institute, Boulder, CO 80301, USA
[14]NASA Goddard Space Flight Center, Greenbelt, MD 20771 USA
[15]American Astronomical Society Working Group on Astroinformatics and Astrostatistics
[16]American Astronomical Society Working Group on Time-Domain Astronomy
[17]American Statistical Association Astrostatistics Interest Group
[18]International Astronomical Union Commission B3 on Astroinformatics & Astrostatistics
[19]LSST's Informatics and Statistics Science Collaboration (ISSC)
[20]International AstroInformatics Association

1 The growing impact of astrostatistics and astroinformatics

Astrostatistics and astroinformatics (A&A) comprise interdisciplinary research combining astronomy with one or more of the information sciences, including statistics, machine learning, data mining, computer science, information engineering, and related fields. For the Astro2010 decadal survey, nearly 100 astronomers and information scientists submitted two State of the Profession Position Papers (Borne *et al.*, 2009; Loredo *et al.*, 2009) highlighting the potential of the then-emerging areas of astrostatistics and astroinformatics to make transformative contributions to astronomy, if only support for research and education in those areas could be enhanced. In the decade since, the size and impact of A&A has grown dramatically, despite only modest changes in formal support of these areas.

In the time since Astro2010, the community of A&A researchers has grown tremendously in size. Scholarly societies and large astronomy projects have responded with the creation of several A&A groups, with a combined membership of several hundred astronomers and information scientists: LSST's Informatics and Statistics Science Collaboration (ISSC, 2009, 72 members), the International Astrostatistics Association[1] (IAA, 2012, 601 members), the American Astronomical Society Working Group in Astroinformatics and Astrostatistics[2], (WGAA; 2012, 116 members) the American Astronomical Society Working Group on Time Domain Astronomy[3] (2014), the American Statistical Association Astrostatistics Interest Group[4] (2014, 111 members), the IEEE Astrominer Task Force (2014), the International Astronomical Union Commission B3 on Astroinformatics & Astrostatistics[5] (2015, 239 members), and the International AstroInformatics Association[6] (2019, 182 members).

Various teams from within the A&A community have submitted multiple Science White Papers addressing recent and future A&A science impacts in various areas of astronomy, and APC White Papers addressing specific A&A considerations such as the needs of petascale A&A research, and education and collaboration support issues.

This White Paper is authored by leaders of the A&A groups listed above, and reflects broad A&A support considerations discussed across their memberships. It briefly **highlights the strong and growing impact of A&A, identifies key issues hampering the growth of this new field, and offers recommendations for improved support of both research and education in A&A.** This WP is not comprehensive; it does not address a number of astroinformatics issues, especially in the arenas of data systems, cyberinfrastructure, Virtual Observatory, etc..

At the turn of the century, SDSS — the first large-scale, public, digital sky survey — dramatically increased interest in statistics as well as machine learning and other computational sciences. Indeed, SDSS is cited as an early example of the so-called Fourth Paradigm of science — *data-intensive science*, colloquially called "big data science" (Hey *et al.*, 2009; Bell *et al.*, 2009). Many astronomical data sources embody the classic "three Vs" — ***volume, variety and velocity*** —

[1] http://iaa.mi.oa-brera.inaf.it/IAA/home.html
[2] https://aas.org/comms/working-group-astroinformatics-and-astrostatistics-wgaa
[3] https://aas.org/comms/working-group-time-domain-astronomy-wgtda
[4] https://community.amstat.org/astrostats/home
[5] https://www.iau.org/science/scientific_bodies/commissions/B3/
[6] http://astroinformatics.info/

distinguishing data-intensive science, particularly with recent wide-field surveys, optical/infrared integral field units, and radio interferometric instruments from earlier smaller and narrowly focused astronomy problems. Time domain survey astronomy has emerged as a major endeavor as wide-field telescopes, both large and small, are dedicated to repeated photometric measurements of celestial populations, producing particularly large datasets. Extracting sound science from complex data often requires advanced statistical and computational methods, even at relatively smaller volumes. Challenging A&A research problems arise across the full spectrum of dataset scales. To highlight broader data science challenges, researchers have added other "Vs" to the list of big-data Vs, most notably *veracity*, referring to the need to quantify uncertainty in data-based inferences, whether based on big datasets or modest ones.[7]

Recent *data challenges* provide excellent examples of the value of considering diverse methodologies for complex problems, and highlight the need to seek interdisciplinary collaboration. A decade ago, the Gravitational LEnsing Accuracy Testing (GREAT) weak lensing shear measurement competitions, GREAT08 and GREAT10 (including galaxy and star/PSF shape measurement challenges, Bridle *et al.* 2010; Kitching *et al.* 2012, 2013), were announced in *Annals of Applied Statistics*. They drew submissions from dozens of teams, many submitting results from multiple methods, with several teams comprising non-astronomers. The GREAT08 prize went to a pair of computer scientists; the GREAT10 galaxy prize went to a pair of astronomers new to weak lensing, with organizers describing it as *"a major success in its effort to generate new ideas and attract new people into the field."* The more recent 2018 Photometric LSST Astronomical Time-Series Classification Challenge (PLAsTiCC, Kessler *et al.* 2019) took advantage of newer crowd-sourcing tools (via the Kaggle platform), and attracted over 1000 submissions. Among the top five performing teams, only a single participant was an astronomer.

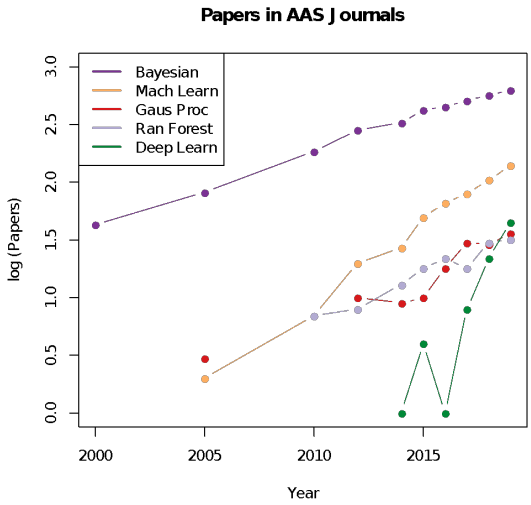

Figure 1: Research papers using the emerging methodology published in AAS Journals since 2000. The number of papers/yr is log10 scale. Methodology is marked in colors: Bayesian - purple, Machine Learning - yellow, Gaussian Processes - red, Random Forests - grey, Deep Learning - green.

The rise of advanced methodologies is recent and incredibly rapid with significant response by the research community. Basic bibliometric statistics, displayed in Fig. 1, show that the use of many modern approaches and methods — e.g., Bayesian statistics, machine learning, Gaussian processes, random forests, and deep learning — is growing exponentially. The *Appendix* describes selected themes of emerging, advanced A&A research, highlighting the breadth of methods and applications, and the rapid growth of interest in adapting state-of-the art data science methods to astronomy. During the 2013–19 period, the number of jobs emphasizing data analysis methodology offered to Ph.D. astronomers (both post-doc and faculty positions) approximately

[7] https://mapr.com/blog/top-10-big-data-challenges-serious-look-10-big-data-vs/

doubled[8]. The Astrostatistics Facebook group[9] has over 4000 members with new members joining daily. The community interest reflects the need for new methodology, and also for communication and training, as the standard training of astronomers lags behind.

The growth of interest in A&A research stands in contrast to serious deficiencies in support of the A&A education and research enterprise. The following section highlights three key gaps between needs and available resources for realizing the potential of A&A to meet current and emerging science challenges. The final section offers specific recommendations to close these gaps.

2 Key Issues: Education, funding, and quality control

Recent developments in methodology were not widely anticipated and have proceeded rapidly. This has resulted in unfamiliar challenges and imbalances for the educational, funding, and quality control structures of the field. In the following subsections, we describe in detail the challenges and imbalances in each of these three areas. From here, it is clear that actions are needed by different segments of the community — universities, observatories and institutes, funding agencies, and leadership organizations like the National Academy of Sciences — and we outline some recommendations in Section 3.

2.1 The education gap

Astronomers are well-trained in mathematics relating to physical processes in order to do astrophysics, but not in applied mathematics, statistics, and computer science relating to extraction of reliable information from complex, noisy datasets. They are typically conversant with computer programming and processing on a moderate scale, but many are not prepared for the world of Big Data with challenges in data storage, access, and efficient analysis on high performance multicore computers, and modern software development practices.

The problem arises in the curriculum of physical scientists: courses in modern statistics, applied mathematics, and computer science are not in the required curriculum. For computation, this deficiency has been recently documented among astronomers: a survey of ≈ 1100 astronomers found that 90% write software but only 8% received substantial training in software development (Momcheva and Tollerud, 2015). Informal on-the-job training is adequate for some purposes, but limits reproducibility, results in inefficient duplication, and can lead to mediocrity, or even unnecessary failure, for more challenging problems. The methodology needed for astronomy and astrophysics is so diverse that specialized coursework in the usage of statistical software environments and computer resources is essential for the astronomical research enterprise.

This deficit in the education of astronomers has been repeatedly noted: recent NASA[10] and National Academy of Sciences[11] reports emphasize the need for professional training beyond

[8] https://asaip.psu.edu/resources/jobs

[9] https://www.facebook.com/groups/astro.r/

[10] Big Data @ STScI: Enhancing STScI's Astronomical Data Science Capabilities over the Next Five Years (2016), http://archive.stsci.edu/reports/BigDataSDTReport_Final.pdf

[11] Optimizing the U.S. Ground-Based Optical and Infrared Astronomy System (2015), National Academy Press https://www.nap.edu/

long-standing formal education in the physical sciences. Some progress has been made. Textbooks on statistical methodology and data analysis (with computer codes) for astrophysics are available and taught in some universities (e.g., Feigelson and Babu, 2012; Ivezić *et al.*, 2014; Bailer-Jones, 2017; Hilbe *et al.*, 2017)[12]. Informal Summer Schools, Hack Days, and tutorials have proliferated. However, except perhaps for disorganized Facebook-based discussion forums, these resources are touching relatively few astronomers (perhaps 10%).

University curricula are not renovating fast enough to match the needs of methodological education for future space scientists, and professional development resources are insufficiently funded or organized to meet the needs of the research community.

2.2 The funding gap

Grants to universities and other institutions specifically designed to improve methodology for astronomical research are very scarce. NASA closed its only grant program in this area in 2011 Applied Information Science Research Program (AISRP). This program was critical, for example, to the development of the worldwide Virtual Observatory (Szalay, 2014), and funded many smaller-scale A&A efforts, including development of SAOImage-DS9 and work on new statistical and machine learning algorithms by individual investigators and multi-university collaborations (the NASA *Astrophysics Research, Analysis & Enabling Technology 2011 Review Panel* evaluated the AISRP program in more detail[13]). NSF has had short-lived interdisciplinary grant programs to promote mathematical developments for astronomy, and has supported astrostatistics at SAMSI programs. While astronomers do have access to agency-wide programs in cyberscience such as the Computational and Data-Enabled Science and Engineering (CDS&E) and various cyber infrastructure programs, the success rate of such proposals may be *de facto* limited by level of buy-in from NSF's Division of Astronomical Sciences.

Other scientific fields do not have these structural problems. Biostatistics is taught in most universities and has been heavily funded by NIH for decades with many large grant programs[14]. Statistics and informatics for Earth sciences have been well-funded by the NSF and NASA[15], coordinated by the inter-agency Big Earth Data Initiative, with results presented in several dozen poster sessions at annual AGU meetings.

[12] see https://asaip.psu.edu/resources/recent-books/methodology-books-for-astronomy

[13] See the July 2011 section of the NAC Astrophysics Subcommittee site.

[14] For illustration, the following NIH grant programs are available in cancer research (statfund.cancer.gov/funding): Big Data to Knowledge; Development of Informatics Technology; Informatics Technology for Cancer Research; Bridging the Gap between Cancer Mechanism and Population Science; Spatial Uncertainty: Data, Modeling, and Communication; Cancer Intervention and Surveillance Modeling Network; Short Courses on Mathematical, Statistical, and Computation Tools for Studying Biological Systems; NIDCR Grants for Data Analysis and Statistical Methodology applied to Genome-wide Data; NLM Express Analysis in Biomedical Informatics; Integrative Omics Data Analysis for Biomedical Informatics; New Computational Methods for Understanding the Functional Role of DNA Variants.

[15] The following NSF grant programs are available in geosciences, in addition to agency-wide programs in mathematics and cyberscience: Collaboration in Mathematical Geosciences; EarthCube; Geoinformatics; Signals in the Soil; Advanced Digitization of Biodiversity Collections. NASA operates the: Earth Observing System Data and Information System; NASA Center for Climate Simulation

2.3 The quality gap

The culture of our research community and funding agencies is fully cognizant that major advances are driven by improvements in instrumentation, and that these instruments require software for operation and knowledge extraction. It is less well recognized that new instruments give rise to science questions so diverse and complex that traditional data analysis procedures are often inadequate. The knowledge and skills of the statistician, applied mathematician, and algorithmic computer scientist need to be incorporated into programs that currently emphasize engineering and physical science in order to fully achieve the scientific potential of instruments and telescopes and the data they provide. These issues might be divided into two stages: (1) data reduction through software pipelines (often developed within instrumentation groups, but ripe for methodological improvements from the broader community); and (2) science analysis that is performed by hundreds of scientists dispersed through U.S. universities and abroad. Both stages benefit from modern statistical and computational methods; in some cases, the science result is completely inaccessible without state-of-the-art methodology.

High standards for analysis methodology are not set consistently for publications, instrument analysis pipelines, science analysis software developed by national observatories or space mission science centers, or for software produced by extramural science programs. The result is uneven quality in data and science analysis products; the methods used in astronomical software systems for data processing and science analysis are often inappropriate and/or obsolete (e.g., see Protassov *et al.*, 2002; Tak *et al.*, 2018). Limited peer review resources often make these kinds of problems undetectable until after publication.

3 Strategic Plan

We have outlined an unusual situation for our profession: the historical unfamiliarity of research based on advanced cross-disciplinary methodology, and the rapidity of its growth, have led to imbalances that hinder research. Strides made in methodologies and computer science are often not incorporated into astronomical research because we lack adequate educational, funding, and quality control structures. In the last decade or so, *data science* has gained traction in both industry and academia; privately funded data science centers have appeared in industry and data science institutes have appeared in universities. While these centers and institutes have contributed to the development of new methodologies in A&A, they are often not fully utilized by astronomy departments at universities and do not provide enough focused support for astronomy. Thus, a serious organizational commitment from the astronomy community at many levels is needed.

The problems outlined in Section 2 can be substantially rectified if concerted effort is made by the funding agencies, national observatories and mission centers, universities, and scholarly societies. Large projects could not only fund software pipelines, but also cross-disciplinary study and oversight so the pipelines and associated science analysis software incorporate modern statistical and computational methods. National institutes could nurture internal teams devoted to methodology and hire consultants to advise large hardware and research groups. Universities can offer undergraduate and graduate courses in statistics, informatics, and the data sciences within

astronomy programs, assuring students interested in data-intensive research careers sufficient curricular flexibility to become appropriately trained.

Cross-disciplinary interest groups that have emerged in scholarly societies can be energized with funds to organize collaborative research efforts, conferences and workshops, and informal education tutorials.

The obvious result of the lack of investment and commitment by the American astronomical enterprise in astrostatistics and astroinformatics is the loss of astronomical results, particularly relating to Big Data science from LSST and its predecessor instruments. During the past decade, A&A has been established as an important research area in the astronomical community. The importance of this research should be also recognized by agencies and universities, and supported by appropriate changes in the funding and educational structures.

With these issues in mind, we offer the following recommendations. We estimate that the total new cost for implementing our specific research and training recommendations is a few million dollars annually, a small fraction of annual spending in astronomy. This small investment will have a disproportionately large impact on A&A and on astronomy as a whole. Our team does not have the resources and expertise to assess costs in detail; further, several recommendations involve adjusting the balance of various resources (monetary and otherwise) across multiple stakeholders. *We propose that the Astro2020 survey recommend that the AAS or another appropriate body establish a committee to review the support of A&A, using these recommendations as a starting point.* The committee should be provided sufficient resources and access to stakeholders to enable developing detailed and realistic recommendations for improved support of A&A.

3.1 Closing the education gap

3.1.1 Universities and National Observatories

- Universities should revise the curriculum in undergraduate physical science to require courses in applied statistics, mathematics, and computer science. At the graduate level, specialized courses in computational methods and usage of statistical methods should be incorporated into the astronomy and astrophysics curriculum. Specifically, students should learn how to use modern astronomy computing environments, and how to harness modern computing hardware efficiently.

- Universities and national observatories should develop information science courses for astronomers at the undergraduate and graduate levels.

- Universities and national observatories should financially support summer schools and cross-disciplinary workshops on advanced methods, both to train astronomical data science researchers and to integrate this emerging area into mainstream astronomy.

- Universities and national observatories should establish specialized permanent appointments for data science in astronomy, as routinely as they now do for observers/instrumentalists and theorists. Cross-appointment permanent positions (e.g., with statistics departments, computer science departments, etc.) should also be considered.

3.1.2 NSF and NASA

- NSF and NASA should establish mechanisms to support educators interested in developing, curating, improving, maintaining, and/or disseminating astroinformatics materials that accelerate and improve astroinformatics education in the community.
- NSF and NASA should survey their existing programs, at the agency and center levels (e.g. NASA centers), which support the A&A education of the community within these agencies. This includes programs such as the Frontier Development Lab from NASA Ames, NASA Goddard's astropy summer schools, and the astroinformatics working groups at Goddard (and other centers).

3.2 Closing the funding gap

3.2.1 Universities and National Observatories

- University astronomy departments and National Observatories should work with internal data science institutes and other departments (e.g. statistics, mathematics, computer science) to offer competitive, interdisciplinary postdoctoral fellowships in A&A.
- Universities should financially support multidisciplinary PhDs in astronomy (e.g., in A&A). This would also encourage graduates to enter astronomy programs even if they are interested in possibly pursuing a more general career in data science.

3.2.2 NSF and NASA

- NSF should provide interdisciplinary grant support for research related to A&A.
- NSF Division of Astronomical Sciences should pursue partnerships in support of medium and large projects that have a significant astronomical data science component.
- With community input, NASA should be urged to reorganize its support of data analysis and information science research. There should be a focus on financial support for both routine and advanced data analysis research that serves space-based astrophysics through development, adaptation, validation and application of modern A&A methods.
- Both NASA and NSF A&A research programs should implement explicitly multi-tiered support, with different categories of research of various duration and levels of funding. Long-term funding must be included, especially targeting young researchers.
- NASA and NSF should encourage reviewers of postdoctoral fellowship applications to recognize A&A, including both proposed A&A research and/or a track-record of high-quality A&A in previous research publications.
- NASA and NSF should instate a 3-year interdisciplinary fellowship program in astronomical data sciences. This would encourage young scientists to pursue A&A careers, and would bring recognition to these scientists and to the discipline.

- NASA and NSF should also support astronomical data science research targeting **infrastructure** (e.g., data management and computational resource management research, including development of astronomy-oriented parallel, grid, and cloud computing software environments, and maintaining the critical software tools). Such support should be separated from support from focused, *science-driven data science research*, either via separate programs, or via explicitly identified proposal categories within a single program.

- Similarly, NASA and NSF should support the development of significant public, open-source software that provides important science-enabling technology, much like the development of a new instrument. As with instruments, significant codebases need maintenance, and funding channels need to support major updates of widely-used codebases similar to how instrument maintenance is supported.

- Agencies should develop or adapt funding opportunities enabling support of A&A data challenges, like those mentioned in § 1. Data challenges (e.g. PLAsTiCC) draw in participants from communities outside of astronomy, and could lead to more interdisciplinary collaboration and higher quality research.

- We echo the recommendations made in the NASA Task Force on Big Data for SMD[16]:

  Recommendation: SMD should establish a new division that would focus on cross-cutting data science and computing projects and whose responsibilities would include establishing the Data Science Applications Program which will promote bringing modern data science methodologies into SMD's data analysis worlds including the science operations of SMD's missions.

  Recommendation: In staffing the Science Committee and the four thematic Science Advisory Committees, SMD should ensure that at least one appointment on each of these committees is reserved for an expert who is a routine user of high-performance computers (NASA's or others), is active in employing modern data science methodologies, and/or is deeply involved in the science operations of large, complex scientific data archives.

  Recommendation: NASA should make prioritized investments in computing and analysis hardware, workflow software and education and training to substantially accelerate modeling workflows. NASA should take the lead to make substantial increases in: ... software modernization; resources to develop new data analysis paradigms; education and training workshops, scientific conferences and journal special collections to effect a culture acceptance of the importance of workflow development and management; . . . lossy data compression and more advanced methods for signal detection.

3.3 Closing the quality gap

- Journals should maintain high standards for analysis methodology and algorithms. This may involve supporting a statistics/informatics editor (as is currently done by the AAS journals), and modifying review processes to ensure that papers with significant A&A

[16] https://science.nasa.gov/science-committee/subcommittees/big-data-task-force

content are examined by reviewers with expertise in both the relevant astrophysics and the relevant information science. Authors should be strongly encouraged (perhaps required, in some circumstances) to make computational results reproducible, e.g., by publishing software, repositories, and/or computational "notebooks" along with papers.

- Astronomy curricula should include training in best practices in development of software, e.g., by encouraging or requiring formal training in programming and development practices from faculty actively engaged in astroinformatics education, computer science departments, and/or from well-vetted training programs (e.g., Software Carpentry).
- Support interdisciplinary collaborations that seek funding to include substantive contributions from experts in specific algorithms and/or computational methods that could advance their research goals.
- Funding agencies should encourage production of open-source software; this development model improves code quality in broad scientific applications.

4 Appendix: Emerging data science themes in astronomy

Here we briefly survey a selection of important data science areas where recent developments in statistics and machine learning are beginning to make significant impacts in astronomy[17]. This survey is by no means exhaustive, nor are the highlighted applications meant to be endorsements of specific approaches. Rather, this survey is intended to display the broad scope of data science research in astronomy and its potential for producing qualitative advances in our ability to distill science from data. Few if any of the highlighted approaches are covered in the recent spate of books on statistics and machine learning in astronomy, emphasizing the need for interdisciplinary research and collaboration.

Nonlinear dimensionality reduction. Many astronomical datasets "live" in a high-dimensional space. An observed spectral energy distribution comprising measurements in dozens, hundreds, or thousands of spectral bands may be considered to be a vector in a sample space with a dimension for each band. An image with a million spatial pixels may be considered as a vector in a million-dimensional sample space. Empirically, a collection of many cases of such data very often lies on or near a low-dimensional manifold in the full sample space. Discovering such a manifold can enable dramatic improvement in inference tasks such as classification or regression (characterizing correlations). When that manifold is a hyperplane, it may be found using techniques from linear algebra that are well-known to astronomers, *principal component analysis* (PCA) being the best-known example. But more often, the manifold will be a complex curve or surface, and discovering it requires tools for *nonlinear dimensionality reduction*. This has been a major research area in statistics and machine learning in the last decade, with several new techniques making significant impacts in diverse areas of astronomy. Several of the most successful approaches rely on analysis of the matrix of pairwise distances or similarities of the data (*spectral clustering* or *spectral connectivity analysis*). Examples include use of diffusion maps for supernova classification, locally-biased spectral graph analysis for describing SDSS

[17] see also the Science White Paper submitted to the Astro2020 in March 2019 (Siemiginowska *et al.*, 2019)

galaxy spectra, t-Distributed Stochastic Neighbor Embedding (t-SNE) for classification of supernovae and stellar spectra, and convolutional neural network based autoencoders for radio galaxy classification (Richards *et al.*, 2011; Lawlor *et al.*, 2016; Lochner *et al.*, 2016; Reis *et al.*, 2018; Ma *et al.*, 2019).

Sparsity. A similar but complementary kind of reduction in complexity can occur in the parameter space used to describe the "true" signals underlying observed data. Signals are often best described in a *transform* space, e.g., Fourier space for periodic time-domain signals, or wavelet or shapelet spaces for images. Empirically, many natural signals have very sparse representations in an appropriately selected transform space. E.g., although the Fourier transform of a light curve with N time samples has O(N) Fourier coefficients, periodic light curves can be described with many fewer than O(N) coefficients. Similarly, image compression exploits the observation that natural images with N pixels are well-described with many fewer than O(N) coefficients in, say, a discrete cosine transform (DCT) or wavelet basis. Information scientists are devising new models and algorithms that exploit knowledge of sparsity to improve signal recovery. A notable example is *compressed sensing*, a class of signal processing techniques that exploits sparsity in a transform space to enable recovery of complex signals even when the *data* are relatively sparse (e.g., not fully covering Fourier space). Example applications in astronomy include improving image recovery in radio interferometry, measuring cosmological image distortions due to weak lensing, and inverting solar flare differential emission measure (DEM) data. The method now appears in about 30 astronomy papers/yr. (Hastie *et al.*, 2015; Wiaux *et al.*, 2009; Leonard *et al.*, 2012; Carrillo *et al.*, 2014; Cheung *et al.*, 2015)

Deep learning. Many flexible model architectures in statistics and machine learning are built by composition of a large number of simple elements. Examples include basis expansions (e.g., linear combinations of Fourier or wavelet basis functions) and artificial neural nets (ANNs, linear combinations of simple but nonlinearly tunable "activation functions"). Early theoretical work on the approximation power of ANNs showed that the flexibility of such compositions could be greatly enhanced by layering: for models with fixed "width" (the number of linearly combined components), flexibility can be greatly enhanced by adding "depth" (using the outputs of components of one layer as inputs to a new layer of superposed components). Since the early 2000s, advances in training algorithms for deep models, combined with wide availability of massively parallel computing capability via GPUs and large training sets, have enabled deep learning (DL) algorithms to leapfrog competitors in many industrial applications (e.g., speech and image recognition and classification). A key aspect of these algorithms has been inclusion of dimension-reducing layers, e.g., via tunable convolution or subsampling operations. These enable DL models to discover rich *hierarchical feature representations* of data, e.g., describing an image in terms of edges with different orientations at the lowest level, groups of edges comprising more complex features at the next level, and so on. It has long been appreciated that feature selection is crucial to the performance of machine learning algorithms; DL models can partly automate feature selection. DL is being applied to diverse learning tasks in many areas of astronomy, e.g., classification of galaxy images and stellar spectra, image deblending, photometric redshift estimation, classification of variables and transients, and source discovery in multimessenger astronomy (Goodfellow *et al.*, 2016; Hoyle, 2016; Mahabal *et al.*, 2017; Pasquet *et al.*, 2019; Allen *et al.*, 2019; Boucaud *et al.*, 2019; Wu *et al.*, 2019; Muthukrishna *et al.*, 2019).

References

Allen, G., Andreoni, I., Bachelet, E., Berriman, G. B., Bianco, F. B., Biswas, R., Carrasco Kind, M., Chard, K., Cho, M., Cowperthwaite, P. S., Etienne, Z. B., George, D., Gibbs, T., Graham, M., Gropp, W., Gupta, A., Haas, R., Huerta, E. A., Jennings, E., Katz, D. S., Khan, A., Kindratenko, V., Kramer, W. T. C., Liu, X., Mahabal, A., McHenry, K., Miller, J. M., Neubauer, M. S., Oberlin, S., Olivas, Jr, A. R., Rosofsky, S., Ruiz, M., Saxton, A., Schutz, B., Schwing, A., Seidel, E., Shapiro, S. L., Shen, H., Shen, Y., Sipőcz, B. M., Sun, L., Towns, J., Tsokaros, A., Wei, W., Wells, J., Williams, T. J., Xiong, J., and Zhao, Z. (2019). Deep Learning for Multi-Messenger Astrophysics: A Gateway for Discovery in the Big Data Era. *arXiv e-prints*.

Bailer-Jones, C. A. L. (2017). *Practical Bayesian Inference: A Primer for Physical Scientists*. Cambridge University Press.

Bell, G., Hey, T., and Szalay, A. (2009). Beyond the data deluge. *Science*, **323**(5919), 1297–1298.

Borne, K., Accomazzi, A., Bloom, J., Brunner, R., Burke, D., Butler, N., Chernoff, D. F., Connolly, B., Connolly, A., Connors, A., *et al.* (2009). Astroinformatics: A 21st Century Approach to Astronomy. In *astro2010: The Astronomy and Astrophysics Decadal Survey*, volume 2010, page P6.

Boucaud, A., Huertas-Company, M., Heneka, C., Ishida, E. E. O., Sedaghat, N., de Souza, R. S., Moews, B., Dole, H., Castellano, M., Merlin, E., Roscani, V., Tramacere, A., Killedar, M., and Trindade, A. M. M. (2019). Photometry of high-redshift blended galaxies using deep learning. *arXiv e-prints*.

Bridle, S., Balan, S. T., Bethge, M., Gentile, M., Harmeling, S., Heymans, C., Hirsch, M., Hosseini, R., Jarvis, M., Kirk, D., Kitching, T., Kuijken, K., Lewis, A., Paulin-Henriksson, S., Schölkopf, B., Velander, M., Voigt, L., Witherick, D., Amara, A., Bernstein, G., Courbin, F., Gill, M., Heavens, A., Mandelbaum, R., Massey, R., Moghaddam, B., Rassat, A., Réfrégier, A., Rhodes, J., Schrabback, T., Shawe-Taylor, J., Shmakova, M., van Waerbeke, L., and Wittman, D. (2010). Results of the GREAT08 Challenge: an image analysis competition for cosmological lensing. *MNRAS* , **405**, 2044–2061.

Carrillo, R. E., McEwen, J. D., and Wiaux, Y. (2014). *Purify: A new algorithmic framework for next-generation radio-interferometric imaging*. IEEE.

Cheung, M. C. M., Boerner, P., Schrijver, C. J., Testa, P., Chen, F., Peter, H., and Malanushenko, A. (2015). Thermal Diagnostics with the Atmospheric Imaging Assembly on board the Solar Dynamics Observatory: A Validated Method for Differential Emission Measure Inversions. *ApJ* , **807**, 143.

Feigelson, E. D. and Babu, G. J. (2012). *Frontmatter*, pages i–vi. Cambridge University Press.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press. http://www.deeplearningbook.org.

Hastie, T., Tibshirani, R., and Wainwright, M. (2015). *Statistical Learning with Sparsity*. New York: Chapman and Hall/CRC.

Hey, T., Tansley, S., and Tolle, K. (2009). *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Microsoft Research.

Hilbe, J. M., De Souza, R. S., and Ishida, E. E. (2017). *Bayesian models for astrophysical data: using R, JAGS, Python, and Stan*. Cambridge University Press.

Hoyle, B. (2016). Measuring photometric redshifts using galaxy images and Deep Neural Networks. *Astronomy and Computing*, **16**, 34–40.

Ivezić, Ž., Connolly, A., Vanderplas, J., and Gray, A. (2014). *Statistics, Data Mining and Machine Learning in Astronomy*. Princeton University Press.

Kessler, R., Narayan, G., Avelino, A., Bachelet, E., Biswas, R., Brown, P. J., Chernoff, D. F., Connolly, A. J., Dai, M., Daniel, S., Di Stefano, R., Drout, M. R., Galbany, L., González-Gaitán, S., Graham, M. L., Hložek, R., Ishida, E. E. O., Guillochon, J., Jha, S. W., Jones, D. O., Mandel, K. S., Muthukrishna, D., O'Grady, A., Peters, C. M., Pierel, J. R., Ponder, K. A., Prša, A., Rodney, S., and Villar, V. A. (2019). Models and Simulations for the Photometric LSST Astronomical Time Series Classification Challenge (PLAsTiCC). *arXiv e-prints*.

Kitching, T. D., Balan, S. T., Bridle, S., Cantale, N., Courbin, F., Eifler, T., Gentile, M., Gill, M. S. S., Harmeling, S., Heymans, C., Hirsch, M., Honscheid, K., Kacprzak, T., Kirkby, D., Margala, D., Massey, R. J., Melchior, P., Nurbaeva, G., Patton, K., Rhodes, J., Rowe, B. T. P., Taylor, A. N., Tewes, M., Viola, M., Witherick, D., Voigt, L., Young, J., and Zuntz, J. (2012). Image analysis for cosmology: results from the GREAT10 Galaxy Challenge. *MNRAS* , **423**, 3163–3208.

Kitching, T. D., Rowe, B., Gill, M., Heymans, C., Massey, R., Witherick, D., Courbin, F., Georgatzis, K., Gentile, M., Gruen, D., Kilbinger, M., Li, G. L., Mariglis, A. P., Meylan, G., Storkey, A., and Xin, B. (2013). Image Analysis for Cosmology: Results from the GREAT10 Star Challenge. *ApJS* , **205**, 12.

Lawlor, D., Budavári, T., and Mahoney, M. W. (2016). Mapping the Similarities of Spectra: Global and Locally-biased Approaches to SDSS Galaxies. *ApJ* , **833**(1), 26.

Leonard, A., Dupé, F.-X., and Starck, J.-L. (2012). A compressed sensing approach to 3D weak lensing. *A&A*, **539**, A85.

Lochner, M., McEwen, J. D., Peiris, H. V., Lahav, O., and Winter, M. K. (2016). Photometric Supernova Classification with Machine Learning. *ApJS* , **225**(2), 31.

Loredo, T. J., Accomazzi, A., Bloom, J., Borne, K., Brunner, R., Burke, D., Butler, N., Chernoff, D. F., Connolly, B., Connolly, A., *et al.* (2009). The Astronomical Information Sciences: A Keystone for 21st-Century Astronomy. In *astro2010: The Astronomy and Astrophysics Decadal Survey*, volume 2010, page P34.

Ma, Z., Xu, H., Zhu, J., Hu, D., Li, W., Shan, C., Zhu, Z., Gu, L., Li, J., and Liu, C. (2019). A Machine Learning Based Morphological Classification of 14,245 Radio AGNs Selected from the Best-Heckman Sample. *ApJS* , **240**(2), 34.

Mahabal, A., Sheth, K., Gieseke, F., Pai, A., Djorgovski, S. G., Drake, A. J., and Graham, M. J. (2017). Deep-learnt classification of light curves. In *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 2757–2764.

Momcheva, I. and Tollerud, E. (2015). Software Use in Astronomy: an Informal Survey. *arXiv e-prints*, page arXiv:1507.03989.

Muthukrishna, D., Narayan, G., Mandel, K. S., Biswas, R., and Hložek, R. (2019). RAPID: Early Classification of Explosive Transients using Deep Learning. *arXiv e-prints*.

Pasquet, J., Bertin, E., Treyer, M., Arnouts, S., and Fouchez, D. (2019). Photometric redshifts from SDSS images using a convolutional neural network. *A&A*, **621**, A26.

Protassov, R., van Dyk, D. A., Connors, A., Kashyap, V. L., and Siemiginowska, A. (2002). Statistics, Handle with Care: Detecting Multiple Model Components with the Likelihood Ratio Test. *ApJ* , **571**(1), 545–559.

Reis, I., Poznanski, D., Baron, D., Zasowski, G., and Shahaf, S. (2018). Detecting outliers and learning complex structures with large spectroscopic surveys - a case study with APOGEE stars. *MNRAS* , **476**(2), 2117–2136.

Richards, J. W., Starr, D. L., Butler, N. R., Bloom, J. S., Brewer, J. M., Crellin-Quick, A., Higgins, J., Kennedy, R., and Rischard, M. (2011). On Machine-learned Classification of Variable Stars with Sparse and Noisy Time-series Data. *ApJ* , **733**, 10.

Siemiginowska, A., Eadie, G., Czekala, I., Feigelson, E., Ford, E. B., Kashyap, V., Kuhn, M., Loredo, T., Ntampaka, M., Stevens, A., *et al.* (2019). The Next Decade of Astroinformatics and Astrostatistics. In *BAAS*, volume 51, page 355.

Szalay, A. S. (2014). From AISR to the Virtual Observatory. In *American Astronomical Society Meeting Abstracts #223*, volume 223 of *American Astronomical Society Meeting Abstracts*, page 203.02.

Tak, H., Ghosh, S. K., and Ellis, J. A. (2018). How proper are Bayesian models in the astronomical literature? *MNRAS* , **481**(1), 277–285.

Wiaux, Y., Jacques, L., Puy, G., Scaife, A. M. M., and Vandergheynst, P. (2009). Compressed sensing imaging techniques for radio interferometry. *MNRAS* , **395**, 1733–1742.

Wu, C., Wong, O. I., Rudnick, L., Shabala, S. S., Alger, M. J., Banfield, J. K., Ong, C. S., White, S. V., Garon, A. F., Norris, R. P., Andernach, H., Tate, J., Lukic, V., Tang, H., Schawinski, K., and Diakogiannis, F. I. (2019). Radio Galaxy Zoo: CLARAN - a deep learning classifier for radio morphologies. *MNRAS* , **482**, 1211–1230.