Astro2020 Science White Paper

The Role of Machine Learning in the Next Decade of Cosmology

Thematic Areas: □ Planetary Systems □ Star and Planet Formation
□ Formation and Evolution of Compact Objects ☑ Cosmology and Fundamental Physics
□ Stars and Stellar Evolution □ Resolved Stellar Populations and their Environments
□ Galaxy Evolution □ Multi-Messenger Astronomy and Astrophysics

Principal Author:
Name: Michelle Ntampaka
Institution: Harvard Data Science Initiative
Center for Astrophysics | Harvard & Smithsonian
Email: michelle.ntampaka@cfa.harvard.edu
Phone: (617) 495-8752

Co-authors: Camille Avestruz[1,2], Steven Boada[3], João Caldeira[4], Jessi Cisewski-Kehe[5], Rosanne Di Stefano[6], Cora Dvorkin[7], August E. Evrard[8,9], Arya Farahi[10], Doug Finkbeiner[6], Shy Genel[11,12], Alyssa Goodman[6,13,14], Andy Goulding[15], Shirley Ho[10,11,15], Arthur Kosowsky[16], Paul La Plante[17], François Lanusse[18], Michelle Lochner[19,20], Rachel Mandelbaum[10], Daisuke Nagai[21], Jeffrey A. Newman[16], Brian Nord[1,4,22], J. E. G. Peek[23,24], Austin Peel[25], Barnabás Póczos[26], Markus Michael Rau[10], Aneta Siemiginowska[6], Dougal J. Sutherland[27], Hy Trac[10], Benjamin Wandelt[11,28,29]

Abstract: In recent years, machine learning (ML) methods have remarkably improved how cosmologists can interpret data. The next decade will bring new opportunities for data-driven cosmological discovery, but will also present new challenges for adopting ML methodologies and understanding the results. ML could transform our field, but this transformation will require the astronomy community to both foster and promote interdisciplinary research endeavors.

[1] Kavli Institute for Cosmological Physics, The University of Chicago, Chicago, IL 60637, USA

[2] Enrico Fermi Institute, The University of Chicago, Chicago, IL 60637, USA

[3] Department of Physics and Astronomy, Rutgers, the State University of New Jersey, Piscataway, NJ 08854, USA

[4] Fermi National Accelerator Laboratory, Batavia, IL 60510, USA

[5] Department of Statistics and Data Science, Yale University, New Haven, CT 06511, USA

[6] Center for Astrophysics | Harvard & Smithsonian, Cambridge, MA 02138, USA

[7] Department of Physics, Harvard University, Cambridge, MA 02138, USA

[8] Department of Physics and Michigan Center for Theoretical Physics, University of Michigan, Ann Arbor, MI 48109, USA

[9] Department of Astronomy, University of Michigan, Ann Arbor, MI 48109, USA

[10] McWilliams Center for Cosmology, Department of Physics, Carnegie Mellon University, Pittsburgh, PA 15213, USA

[11] Center for Computational Astrophysics, Flatiron Institute, New York, NY 10010, USA

[12] Columbia Astrophysics Laboratory, Columbia University, New York, NY 10027, USA

[13] Harvard Data Science Initiative, Harvard University, Cambridge, MA 02138, USA

[14] Radcliffe Institute for Advanced Study, Harvard University, Cambridge, MA 02138, USA

[15] Department of Astrophysical Sciences, Princeton University, Princeton, 08544, USA

[16] Department of Physics and Astronomy, University of Pittsburgh, Pittsburgh, PA 15260, USA

[17] Department of Physics and Astronomy, University of Pennsylvania, Philadelphia, PA 19104, USA

[18] Berkeley Center for Cosmological Physics, University of California, Berkeley, CA 94720, USA

[19] African Institute for Mathematical Sciences, Cape Town, 7945, South Africa

[20] South African Radio Astronomy Observatory, Cape Town, 7405, South Africa

[21] Department of Physics, Yale University, New Haven, CT 06520, USA

[22] Department of Astronomy and Astrophysics, University of Chicago, Chicago, IL 60637, USA

[23] Space Telescope Science Institute, 3700 San Martin Dr, Baltimore, MD 21218, USA

[24] Department of Physics & Astronomy, Johns Hopkins University, Baltimore, MD 21218, USA

[25] AIM, CEA, CNRS, Université Paris-Saclay, Université Paris Diderot, Sorbonne Paris Cité, F-91191 Gif-sur-Yvette, France

[26] Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA 15213, USA

[27] Gatsby Computational Neuroscience Unit, University College London, London, UK

[28] Sorbonne Université, CNRS, UMR 7095, Institut d'Astrophysique de Paris (IAP), 75014 Paris, France

[29] Sorbonne Université, Institut Lagrange de Paris (ILP), 75014 Paris, France

The Role of Machine Learning in the Next Decade of Cosmology

Machine learning permeates our daily lives — performing tasks from identifying the people in a photograph to suggesting the next big purchase — but will it change the way we do research? The last decade has seen a remarkable rise in interdisciplinary machine learning (ML)-based astronomy research, offering enticing improvements in the ways we can interpret data. The next decade will see a continued rise in data-driven discovery as methods improve and data volumes grow, but realizing the full potential of ML — even in an era of unprecedented data volumes — presents challenges.

1 What Are Data Science and Machine Learning?

Data science is the study, application, and often the *art* of creating and using sophisticated algorithms and cutting-edge data analysis techniques to extract information from data. It includes the fields of statistics, machine learning, applied mathematics, and computer science. Data science is an inherently interdisciplinary endeavor that reaches across departmental lines to produce new and innovative ways to interpret simulated data and astronomical observations. When used well, data science provides methods for extracting more information from data sets than ever before, reducing bias and scatter, identifying interesting outliers, and inexpensively generating simulated data. When used *very* well, it guides our physical interpretation of observations and can lead to great discoveries.

While it's important to define what data science is, it's also important to define what it is *not*. Data science is not the study of how to store or disseminate data. While data storage and dissemination are important issues facing the astronomy community in the era of large surveys such as the Large Synoptic Survey Telescope (LSST), addressing these issues is not data science. Nor is data science equivalent to data analysis. When we refer to expanded uses of data science in cosmology, we do not include solidly established and easy-to-implement foundational tools such as chi-square analysis and linear regression. Data science also is not "big data," though the two often work side-by-side.

Data science encompasses a broad range of sophisticated data-analysis methodologies, and machine learning (ML) is just one tool in the toolbox. Machine learning research explores the development and application of algorithms that find patterns in data. In the context of astronomy, ML algorithms can be used to address a broad range of tasks including: describing complicated relationships, identifying data clusters and data outliers, reducing scatter by using complex or subtle signals, generating simulated data, classifying observations, addressing sparse data, and exploring data sets to understand the physical underpinning. Machine learning is gaining traction within the astronomy community, and compelling successful applications of ML indicate that it has the potential to be transformative in the upcoming decade.

2 Machine Learning Successes in Cosmology

Recent successes illustrate the potential for sophisticated machine learning data-analysis tools to make significant strides in cosmology. These successes include the following important results:

1. Galaxy clusters are sensitive to the underlying cosmological model, and low-scatter cluster mass proxies are one essential ingredient in using these objects to constrain parameters. ML has been shown to significantly reduce scatter in cluster mass estimates compared to more traditional methods (Ntampaka et al., 2015; Armitage et al., 2018; Ho et al., 2019).
2. Weak lensing maps can shed light on the fundamental nature of gravity and cosmic acceleration. ML has been used with such maps to discriminate between standard and modified gravity models that generate statistically similar observations (Peel et al., 2018). Non-Gaussianities in weak lensing maps can encode cosmological information, but these are hard to measure or parameterize. ML has been shown to tighten parameter constraints by a factor of five or more by harnessing these non-Gaussianities (Gupta et al., 2018; Ribli et al., 2018).
3. Next-generation cosmic microwave background (CMB) experiments will have increased sensitivity, enabling improved constraints on fundamental physics parameters. Achieving optimal constraints requires high signal-to-noise extraction of the projected gravitational potential from the CMB maps. ML techniques have been shown to provide competitive methods for this extraction, and are expected to excel in capturing hard-to-model non-Gaussian foreground and noise contributions (Caldeira et al., 2018).
4. N-body simulations are an effective approach to predicting structure formation of the universe, but are computationally expensive. ML has been used to predict structure formation of the universe, generating a full 3D N-body-like simulation with positions and velocities in $30ms$ (He et al., 2018). This method outperforms traditional fast analytical approximation and accurately extrapolates far beyond its training data.
5. Estimating cosmological parameters from the large-scale structure is traditionally done by calculating summary statistics of the observed large-scale structure traced by galaxies and then compared to the analytical theory. ML can be used to estimate cosmological parameters directly from the large-scale structure field and find more stringent constraints on the cosmological parameters (Ravanbakhsh et al., 2016).
6. Observations of the Epoch of Reionization can provide information about the earliest luminous sources. ML can classify the types of sources driving reionization (Hassan et al., 2018) and measure the duration of reionization to within 10%, given a semi-analytic model and a strong prior on the midpoint of reionization (La Plante & Ntampaka, 2018).
7. Topological data analysis (TDA) is an ML and statistical method for summarizing the shape of data. TDA has been useful for discriminating dark energy models on simulated data (Van de Weygaert et al., 2011), isolating structures of the cosmic web (Sousbie et al., 2011; Libeskind et al., 2018), and defining new types of structures in the cosmic web such as filament loops (Xu et al., 2018). TDA may also help constrain the sum of neutrino masses (Xu et al., 2018).
8. Supernova classification is a critical step in obtaining cosmological constraints from type Ia supernovae in photometric surveys such as LSST. ML has proven to be a powerful tool (e.g., Lochner et al., 2016) and has been successfully applied to the current largest public supernova dataset (Narayan et al., 2018). The public has become heavily involved in devel-

oping new classification techniques (The PLAsTiCC team et al., 2018; Malz et al., 2018).

9. Strong lensing probes cosmic structure along lines of sight. ML was the most effective method at correctly identifying strong lensing arcs in a recent data challenge, outperforming humans at this classification task (Lanusse et al., 2018). ML makes the analysis of strong lensing systems 10 million times faster than the state-of-the-art method (Hezaveh et al., 2017; Perreault Levasseur et al., 2017).

ML will not displace standard statistical reasoning for well-modeled phenomena. However, there are many cases where our current parametric models are inadequate to fully describe a physical system. These ML successes in cosmology imply that there is great potential for data-driven discovery, particularly as data sets grow and become more complex.

3 Challenges for the Next Decade

ML represents the next step in automation, driven by both rapidly increasing data volumes and the desire to prioritize human attention on tasks that require our insight and ingenuity. There is no doubt that ML techniques will become more powerful and widespread in the 2020s, transforming our ability to address previously intractable problems.

ML has demonstrated its potential to accelerate discovery in astrophysics, but challenges to more widespread adoption remain. New tools come with new failure modes, and ML poses the temptation to choose expediency over understanding. A common complaint about ML methods is that they are black boxes that cannot lead to physical understanding, but this need not be the case. Though it is difficult to understand the inner workings of a complex ML model, in many cases it is not impossible. There are ways to peer inside the box and to gain physical understanding from complex models. For example, Google's DeepDream project (Mordvintsev et al., 2015) originated as a way to visualize inputs that maximize activation in various layers of a neural network, and recent applications to cosmology indicate that the method can be used to gain physical understanding of astronomical systems (e.g., Ntampaka et al., 2018). Other recent developments to improve ML interpretability include saliency maps (Simonyan et al., 2013) that reveal which parts of an input most influence the output, and the deep k-nearest neighbors approach (Papernot & McDaniel, 2018) that shows which training examples have the most influence on a specific outcome. ML interpretability is an active area of research, and we expect further improvements in the quality and diversity of interpretation techniques in the next decade.

At the intersection of ML and cosmology lies a unique opportunity for the benefit of both fields. ML is likely to accelerate discovery in cosmology through multiple applications and modalities (classification, regression, reinforcement learning). Cosmology, in return, provides new tasks and challenges for ML researchers and well-understood data sets for testing ML methods. The challenges provided by cosmology open opportunities for breakthroughs in the fundamental understanding of ML.

More advanced analyses of cosmological data place stringent requirements on the interpretability of results. This represents a key hurdle for applying ML to cosmology: the assessment of uncertainty and the removal of bias. Integrating traditional statistical methods with modern ML models may provide a solution, but this will require cross-disciplinary collaboration among statisticians, ML researchers, and cosmologists. During the 2020s, it is plausible that we could train, characterize, and use ML with the same rigor that we bring to more conventional statistical analysis. This is a significant shift in how we approach our research, and supporting this shift will require the community's investment in education, interdisciplinary research endeavors, and the development and transfer of methodologies from the computer science community.

4 Opportunities in the 2020s and Beyond

The assertion that "astronomy is entering the era of big data" has become cliché. And yet, we cannot help but note that upcoming data sets, both big and small, will provide rich opportunities to use machine learning for teasing out complex correlations. Here, we provide a few examples of those opportunities.

Big Data Opportunities With LSST: The LSST survey (LSST Science Collaboration et al., 2009) will provide the optical astronomical community with an unprecedented data rate. It will cover nearly the entire visible southern sky roughly every three days for a decade, providing $\sim$1,000 exposures total at each position (split across 6 passbands). With 500 petabytes of images, and a database including tens of trillions of observations of tens of billions of objects (Ivezić et al., 2008), LSST's discovery potential will be enormous — but standard analysis methods will not enable the community to unlock the full potential of LSST. Its high source density, as well as the transient nature of some phenomena (e.g., asteroids and supernovae), will present a new set of challenges related to source identification and classification in this colossal dataset. For example, LSST expects to identify, and subsequently classify, 10 million rapid-response transient alerts on any given night. The continued development and future implementation of carefully designed ML algorithms at both the image processing (e.g., Goulding et al., 2018; Dai & Tong, 2018; Ackermann et al., 2018) and catalog (e.g., Narayan et al., 2018; Malz et al., 2018) levels have the potential for producing significant advances in our ability to efficiently extract scientifically useful information (e.g., classification, distance, morphology, and mass) from the LSST data. However, these ML methodologies will require further exploration to fully understand their feasibility and general applicability to LSST.

Big Data Opportunities in Radio Astronomy: The Hydrogen Epoch of Reionization Array (HERA), a radio interferometer seeking to provide the first detection of the 21 cm power spectrum from the Epoch of Reionization, is projected to produce 50 TB of data per night of observation when completed in late 2019. ML can identify radio frequency interference present in data from HERA faster and more reliably than traditional algorithms (Kerrigan et al., in prep.). ML is also well-poised to replace other key aspects of the data analysis and reduction pipeline, such as the calibration of antennae and automatic identification of malfunctioning equipment. Further into the decade, the Square Kilometre Array (SKA) will feature data volumes even larger than HERA, with ML representing a viable path for analyzing and reducing these data in real time.

Pipeline Optimization Opportunities: ML has the potential to have a dramatic impact on the efficiency of cosmological experiments. For example, real-bogus (Brink et al., 2013) is an ML system for determining whether a transient detected in photometric variation is a true variable object or simply an artifact. Similarly, the Dark Energy Camera Plane Survey (Schlafly et al., 2018) uses a simple deep neural network to find images that have nebulosity, and, thus, require a separate processing algorithm. Furthermore, ML has the potential to simplify and accelerate the building of statistical inference pipelines in the context of full forward models through Likelihood-

Free Inference (Alsing et al., 2018) and the automated extraction of informative features from data sets (Charnock et al., 2018). The cosmological explorations of the 2020s will require excellent quality control while simultaneously handling unprecedented volumes of data, and ML is a strong option for pipeline optimization.

Low Signal Opportunities With *eROSITA*: While the most obvious targets of opportunity are observations with unprecedented data volumes, fully harnessing our smaller data sets provides rich opportunities for applying ML methodology as well. For example, the upcoming *eROSITA* mission is estimated to find more than 90,000 galaxy clusters with masses $> 10^{13.7}h^{-1}\,M_{\odot}$ (Pillepich et al., 2012). While the mission will detect clusters out to $z \sim 2$, a significant fraction of these cluster observations will be in the low-photon regime of 100 photons or fewer (Pillepich et al., 2018). Fully utilizing this sample will require developing techniques that provide low scatter mass proxies in the low signal regime, and ML is one viable option for this.

Archival Data Opportunities: The potential for ML to help make great scientific strides is not limited to these upcoming data sets. Research based on *Hubble* archival data, for example, outnumbers those on new observations (Villard, 2011), showing that there is a vast, untapped potential even in the community's archival data.

The astronomy community already makes a significant investment in state-of-the-art instrumentation, software development to support data reduction (e.g., Astropy Collaboration et al., 2013; Craig et al., 2015; Hazelton et al., 2017; Mommert, 2017; Price-Whelan et al., 2018), and data management (e.g., Sands, 2015; Darch et al., 2017). There remains a strong need for this community to invest in the interdisciplinary development and application of cutting-edge ML techniques to interpret our rich and complex data, and help propel us into the next decade.

We encourage the astronomy community to invest in education and interdisciplinary research efforts that will transfer knowledge and methods from the ML research community to our field. We also encourage efforts to build communities of practice for ML-based studies, especially those that profitably join simulated and observational survey data. In the 2020s and beyond, these communities could cluster around discipline-focused hubs, or science gateways[1], offering researchers access to open-source software, relevant data products, and common analysis workflows.

5 Summary

Recent applications of machine learning techniques to cosmological questions have made remarkable improvements in the way we interpret our data, and these compelling successes imply that ML has the potential to be transformative to our field. This transformation will require the astronomy community to cultivate and support research endeavors that cross traditional discipline boundaries, but the payoff has the potential to be steep. Machine learning will give cosmologists access to the data analysis methods that we need to fully utilize our rich data sets and make great scientific leaps forward over the next decade.

[1] https://sciencegateways.org/

References

Ackermann, S., Schawinski, K., Zhang, C., Weigel, A. K., & Turp, M. D. 2018, MNRAS , 479, 415

Alsing, J., Wandelt, B., & Feeney, S. 2018, MNRAS , 477, 2874

Armitage, T. J., Kay, S. T., & Barnes, D. J. 2018, ArXiv e-prints, arXiv:1810.08430

Astropy Collaboration, Robitaille, T. P., Tollerud, E. J., et al. 2013, A&A, 558, A33

Brink, H., Richards, J. W., Poznanski, D., et al. 2013, MNRAS , 435, 1047

Caldeira, J., Wu, W. L. K., Nord, B., et al. 2018, arXiv e-prints, arXiv:1810.01483

Charnock, T., Lavaux, G., & Wandelt, B. D. 2018, Phys. Rev. D, 97, 083004

Craig, M. W., Crawford, S. M., Deil, C., et al. 2015, ccdproc: CCD data reduction software, ascl:1510.007

Dai, J.-M., & Tong, J. 2018, arXiv e-prints, arXiv:1807.10406

Darch, P. T., Sands, A. E., Borgman, C., Golshan, M. S., & Traweek, S. 2017, in American Astronomical Society Meeting Abstracts, Vol. 229, American Astronomical Society Meeting Abstracts #229, 128.01

Goulding, A. D., Greene, J. E., Bezanson, R., et al. 2018, Publications of the Astronomical Society of Japan, 70, S37

Gupta, A., Matilla, J. M. Z., Hsu, D., & Haiman, Z. 2018, Phys. Rev. D, 97, 103515

Hassan, S., Liu, A., Kohn, S., & La Plante, P. 2018, ArXiv e-prints, arXiv:1807.03317

Hazelton, B. J., Jacobs, D. C., Pober, J. C., & Beardsley, A. P. 2017, The Journal of Open Source Software, 2, 140

He, S., Li, Y., Feng, Y., et al. 2018, arXiv e-prints, arXiv:1811.06533

Hezaveh, Y. D., Perreault Levasseur, L., & Marshall, P. J. 2017, Nature , 548, 555

Ho, M., Rau, M. M., Ntampaka, M., et al. 2019, arXiv e-prints, arXiv:1902.05950

Ivezić, Ž., Kahn, S. M., Tyson, J. A., et al. 2008, arXiv e-prints, arXiv:0805.2366

Kerrigan, J., La Plante, P., Kohn, S., et al. in prep.

La Plante, P., & Ntampaka, M. 2018, ArXiv e-prints, arXiv:1810.08211

Lanusse, F., Ma, Q., Li, N., et al. 2018, MNRAS , 473, 3895

Libeskind, N. I., Van De Weygaert, R., Cautun, M., et al. 2018, Monthly Notices of the Royal Astronomical Society, 473, 1195

Lochner, M., McEwen, J. D., Peiris, H. V., Lahav, O., & Winter, M. K. 2016, The Astrophysical Journal Supplement Series, 225, 31

LSST Science Collaboration, Abell, P. A., Allison, J., et al. 2009, ArXiv e-prints, arXiv:0912.0201

Malz, A., Hložek, R., Allam, T., et al. 2018, Arxiv:1809.11145v1

Mommert, M. 2017, PHOTOMETRYPIPELINE: Automated photometry pipeline, ascl:1703.004

Mordvintsev, A., Olah, C., & Tyka, M. 2015, in Google AI Blog

Narayan, G., Zaidi, T., Soraisam, M. D., et al. 2018, The Astrophysical Journal Supplement Series, 236, 9

Ntampaka, M., Trac, H., Sutherland, D. J., et al. 2015, ApJ , 803, 50

Ntampaka, M., ZuHone, J., Eisenstein, D., et al. 2018, arXiv e-prints, arXiv:1810.07703

Papernot, N., & McDaniel, P. 2018, ArXiv e-prints, arXiv:1803.04765

Peel, A., Lalande, F., Starck, J.-L., et al. 2018, ArXiv e-prints, arXiv:1810.11030

Perreault Levasseur, L., Hezaveh, Y. D., & Wechsler, R. H. 2017, ApJ , 850, L7

Pillepich, A., Porciani, C., & Reiprich, T. H. 2012, MNRAS , 422, 44

Pillepich, A., Reiprich, T. H., Porciani, C., Borm, K., & Merloni, A. 2018, MNRAS , 481, 613

Price-Whelan, A. M., Sipőcz, B. M., Günther, H. M., et al. 2018, AJ , 156, 123

Ravanbakhsh, S., Lanusse, F., Mandelbaum, R., Schneider, J., & Poczos, B. 2016, arXiv e-prints, arXiv:1609.05796

Ribli, D., Ármin Pataki, B., & Csabai, I. 2018, ArXiv e-prints, arXiv:1806.05995

Sands, A. E. 2015, in American Astronomical Society Meeting Abstracts, Vol. 225, American Astronomical Society Meeting Abstracts #225, 422.04

Schlafly, E. F., Green, G. M., Lang, D., et al. 2018, The Astrophysical Journal Supplement Series, 234, 39

Simonyan, K., Vedaldi, A., & Zisserman, A. 2013, ArXiv e-prints, arXiv:1312.6034

Sousbie, T., Pichon, C., & Kawahara, H. 2011, Monthly Notices of the Royal Astronomical Society, 414, 384

The PLAsTiCC team, Allam, T., Bahmanyar, A., et al. 2018, Arxiv:1810.00001v1

Van de Weygaert, R., Vegter, G., Edelsbrunner, H., et al. 2011, in Transactions on Computational Science XIV, Springer-Verlag, 60–101

Villard, R. 2011

Xu, X., Cisewski-Kehe, J., Green, S. B., & Nagai, D. 2018, arXiv preprint arXiv:1811.08450