Elements of effective machine learning datasets in astronomy

Bernie Boscoe*
Department of Computer Science
Occidental College
Los Angeles, CA 90041
boscoe@oxy.edu

Tuan Do
Department of Physics and Astronomy
UCLA
Los Angeles, CA 90025
tdo@astro.ucla.edu

Evan Jones
Department of Physics and Astronomy
UCLA
Los Angeles, CA 90025
evan.jones@astro.ucla.edu

Yunqi Li
Department of Physics and Astronomy
UCLA
Los Angeles, CA 90025

Kevin Alfaro
Department of Physics and Astronomy
UCLA
Los Angeles, CA 90025

Christy Ma
Department of Physics and Astronomy
UCLA
Los Angeles, CA 90025

Abstract

In this work, we identify elements of effective machine learning datasets in astronomy and present suggestions for their design and creation. Machine learning has become an increasingly important tool for analyzing and understanding the large-scale flood of data in astronomy. To take advantage of these tools, datasets are required for training and testing. However, building machine learning datasets for astronomy can be challenging. Astronomical data is collected from instruments built to explore science questions in a traditional fashion rather than to conduct machine learning. Thus, it is often the case that raw data, or even downstream processed data is not in a form amenable to machine learning. We explore the construction of machine learning datasets and we ask: what elements define effective machine learning datasets? We define effective machine learning datasets in astronomy to be formed with well-defined data points, structure, and metadata. We discuss why these elements are important for astronomical applications and ways to put them in practice. We posit that these qualities not only make the data suitable for machine learning, they also help to foster usable, reusable, and replicable science practices.

1 Introduction

In recent years astronomy has seen a wide application of machine learning (ML) in numerous subfields, from exoplanets and stellar astrophysics to extragalactic and cosmology applications [5] [22]. In particular, the imminent start of the largest sky surveys ever conducted have motivated astronomers to adopt machine learning methods to filter, analyze, and extract information from these surveys. A large number of astronomy publications cite the Legacy Survey of Space and Time (LSST) and the Euclid space mission as the main drivers for implementing machine learning processes [9]

Machine Learning and the Physical Sciences workshop, NeurIPS 2022.

[17]. The sheer volume of data to be generated by those telescopes is far more than any other survey or mission to date [1] [24] [18] [12] [14].

Astronomy as a field has famously and painstakingly made numerous datasets freely available in the forms of mission archives and survey repositories, but there has been less work in general on defining and building common practices for creating machine learning datasets [6]. We claim traditional mission and survey catalogs are effective astronomy datasets: they comprise careful calibrations, calculations, meticulously detailed descriptions and imagery, typically in a format enabling SQL queries [21]. They are queryable, available, interpretable for humans and readable by astronomers' tools. To employ machine learning techniques, these datasets are commonly transformed for tools developed in industry like PyTorch and TensorFlow [13][15]. This laborious transformation process to an ML-ready dataset requires choices about what content, structure, and metadata to include. However, little work has been done on understanding this process and identifying what constitutes an effective machine learning dataset for astronomy [16].

In this paper, we outline three elements in effective machine learning datasets for astronomy and suggest adopting them in practice. In our view, effective datasets are useful, usable, and reusable for all researchers. We propose that effective ML datasets have three characteristics: well-defined data points, well-defined structure, and well-defined metadata. In the sections below, we explore these three characteristics through the lens of astronomy datasets for machine learning. We use the term well defined to describe datasets conceptually, structurally, and holistically. The datasets should possess the necessary characteristics for being effective, useful, usable, and reusable [23].

2 Well-defined data points

Astronomical data in an upstream state often consists of unstructured raw images and spectra from telescopes that require further processing into data products for use in scientific analysis. Additionally, further processing is required to transform these data products into structured data points (tabular, time series, etc.) or image formats for machine learning. Some areas of research do not use image data, while others use combinations of both image and measurement data. In this paper we provide best practices using both; but image data can be eliminated and the data flows would remain similar.

Well-defined data points are outcomes of processing raw data, and have clear boundaries delineating what information is and is not contained in the dataset. Some boundaries stem from choices made by the people previously or presently involved in the dataset creation process, while other boundaries come from structural limitations such as instrument configurations and the size and construction of the mirrors. These boundaries may be mutable or immutable, depending on the context; our focus is on documenting these decision-making processes as an integral part of ML datasets, as associated metadata.

Structural boundaries describe 'fixed' data collected and stored from instruments, wavelengths, time durations, and regions of sky, to name a few. Boundaries resulting from human choice, such the selection of objects, are malleable; and can be reworked to create different datasets. Datasets are blended results of iterations of decisions. For example, a machine learning dataset might include galaxies over a particular region of the sky. The selection of galaxies is a choice while the region of the sky may be limited by the original survey parameters. Documenting details for why a particular dataset's attributes were chosen are helpful for others to understand and potentially reuse data. In this sense, the boundaries can define a limited ground truth. Well-defined data attributes formed from careful documentation and decision-making processes should be documented and consistent. This allows astronomers to have confidence in the use and reuse of the data and in understanding potential biases in the data.

When selecting and organizing data points, astronomers must make decisions about: 1) quantifying the quality of data points, 2) establishing criteria for included data points, 3) establishing outlier criteria, and 4) identifying and potentially removing missing or low-quality data.

Data quality measures are a way to signify how much trust the data curators have on the resulting data points. For example, some data points have higher uncertainties or have an increased chance of being artifacts due to limits of the instruments used to make the measurements. By including attributes such as data quality flags [2], data curators can help users filter data according to quality that might be necessary for different ML models. For example, some machine learning models are much more sensitive to noisy data than others.

Outliers are important to identify and potentially remove. Two main types of outliers are: (1) data points outside of the typical sample distribution and (2) data points that are erroneous measurements. In machine learning, training data that contains outliers can strongly bias the predictions of models. Outliers that are confirmed to be artifacts or errors that make them ‘mistakes’ should likely be discarded, as machine learning algorithms will ‘learn’ these errors and skew results. A different approach must be taken with outliers that are likely to have been measured correctly but have values that are very far away from the rest of the data distribution. This type of outlier should be kept in the data but considerations must be taken about how to deal with it. Identification of these outliers require characterizing the statistical distribution of the existing data points. Documenting these outliers will alert users to their potential effects during ML training. When dealing with outliers, one approach users can take is to bin them as a category, thereby reducing their effect or otherwise diminishing their power with other statistical techniques.

Missing data points are also important to consider. Missing data might be in the form of a single attribute of a data point, or, it might mean a large sector of data points is not available. This might happen because of instrumentation limitations. For example, some astronomical datasets may contain both imaging and spectroscopic observations of many objects. However, because spectroscopy is much more difficult to obtain than images, some objects may be missing spectroscopic data. One way to mitigate problematic missing data is to use simulated data [3]. In machine learning, interpolation and extrapolation techniques can mitigate these issues, with varying degrees of success. Sampling techniques such as these may shape a better representation of a possible ground truth for a particular study. Astronomers should interpret results sensitive to missing data as a way of handling known limitations.

3 Well-defined dataset structure

The form and structure of datasets are important for both machine learning and astronomy, but additional work must be taken to translate astronomical data points into ML-ready datasets. Structural considerations such as the data format, tabular shape, and image sizes and dimensions can have major impacts on the type of machine learning models able to be used.

In astronomy, data is often stored in the FITS file format, which is a broad data format that can store metadata, imaging, spectra, and tabular data [7][20]. While FITS files are highly compatible with, and designed for astronomical software, the files are not standardized enough to use directly in machine learning models. Machine learning tools expect inputs of certain data types, and are not flexible enough to handle aberrations in data structures.

While there are API libraries available such as AstroPy [4] to convert FITS files into machine learning amenable data formats, this transformation is not trivial, and can be problematic. PNG files are the most common image format used in mainstream machine learning, with lower resolution and smaller dimensional images comprising many datasets such as ImageNet or MNIST. PNG and JPG file formats use at the smallest 8 bits per channel, up to three channels (RGB). FITS image files, on the other hand, may contain many channels of various wavelengths at once, for example the COSMOS survey [19][11] contains thirty different bands of wavelengths for its image data.

Reducing image resolution, number of channels, and image dimensions can greatly affect the amount of information contained in each image. For example, astronomers use multiple units for brightness, but in transformations to PNG these values would be reduced to a scale of integers from 0-255, greatly reducing each measurement’s precision. Therefore, it is not recommended in many cases to transform astronomy images stored in FITS files to PNGs or JPGs. Also, FITS holds metadata about each image in its header section which also needs to be retained, and would be lost in a format transformation.

Instead of PNG files, a more flexible format that works well with machine learning algorithms like HDF5 [8] is a wiser choice. HDF5 provides a better compromise between preserving all the information in FITS and ease of use in ML models. HDF5 is a library and associated file format that can store most types of multi-dimensional array data. This means it is able to preserve all wavelength channels in an astronomical image. The HDF5 data format is capable of storing metadata, tabular data, and image data, and works well to enable data to be ingested into TensorFlow or PyTorch. One major downside to HDF5 for astronomers is that its flexibility of storing metadata and data varies widely in terms of structure and labeling. Therefore, the FITS structure astronomers are familiar with is absent in HDF5, and conversions between the two formats are not one-to-one mappings.

Because of this issue, datasets for machine learning in astronomy must retain critical information spanning two file formats. Best practices include providing code to produce the HDF5 file in addition to descriptions of the storage of metadata alongside the data itself. There is no elegant solution; perhaps in the future a more seamless system could be developed to ingest FITS images into machine learning, but image sizes and the number of channels could be too large for model training. Issues of computing power, training time, and file size are of serious importance when constructing ML datasets.

Restrictions imposed on data formats originate from tool design outside of astronomy, so the structure of machine learning datasets must be made to align the science goals with requirements from the tools. For instance, computer vision for recognizing galaxy morphologies may need to work for galaxies of many different scale sizes, but ML tools require images to be a specific dimension. In most cases, the images from telescopes are much larger than the sizes that are accepted by ML tools. Structuring astronomical data for current ML tools can limit the amount of information that was originally available, for example in mainstream ML tools, to decrease training time image dimensions and resolutions are routinely resized and downscaled. Astronomers building ML datasets should consider and articulate potential information loss from transforming the data.

4 Well-defined metadata

Well-defined metadata includes: 1) all contextual information relevant to the creation of the dataset, 2) the features and form of the dataset, 3) motivations for creating the dataset with respect to the initial scientific goal.

To document contextual information, an astronomer should document the creation of the dataset starting from how the original data was obtained from the archives or instruments to the dataset's final form. For example, users of most astronomy archives use SQL queries to extract subsets of data; queries used to form ML datasets should be preserved. Filtering and processing steps should also be documented and explained. Often, ML datasets in astronomy end up being different versions of similar data; specific versioning schemas should be enacted to ensure consistency in subsequent use of each dataset.

With respect to feature and form preservation, well-defined tabular metadata should include metadata for each column and should include details such as units, descriptions, and how features were obtained. For images from FITS files, relevant data from the FITS headers should be preserved outside of FITS, if possible, and made available with the final machine learning dataset. The file format version (if any) and tools that can read the dataset should be documented in an archival file format.

Metadata detailing the motivation for the creation of the dataset enables users to better understand ways it contextually could be used. Often, ML datasets for astronomy are created with specific science goals in mind (e.g. photometric redshifts [10]). Decisions about the boundaries of the initial dataset are easier to understand if the original goals are known. By documenting the goals and the requirements of the initial scientific investigations, users can situate the dataset with respect to their own science goals.

5 Conclusion

Datasets with well-defined data points, structure, and metadata invite reusability, which makes scientific studies more reliable and efficient. Because of the intensive labor that necessarily goes into creating machine learning-ready datasets, reusing these datasets makes enormous sense to researchers. Machine learning datasets possessing the elements we outlined in this work are a way forward to effectively use massive datasets that are unable to be manually explored. We recommend that astronomers create effective datasets with an eye toward machine learning-ready characteristics, for the benefit of all data-intensive science.

References

[1] Viviana Acquaviva. “Pushing the Technical Frontier: From Overwhelmingly Large Data Sets to Machine Learning”. In: *Proceedings of the International Astronomical Union* 15.S341 (Nov. 2019). arXiv:1901.05978 [astro-ph], pp. 88–98. ISSN: 1743-9213, 1743-9221. DOI: 10.1017/S1743921319003077. URL: http://arxiv.org/abs/1901.05978.

[2] Hiroaki Aihara et al. “Second Data Release of the Hyper Suprime-Cam Subaru Strategic Program”. In: *Publications of the Astronomical Society of Japan* 71.6 (Dec. 2019). arXiv:1905.12221 [astro-ph], p. 114. ISSN: 0004-6264, 2053-051X. DOI: 10.1093/pasj/psz103. URL: http://arxiv.org/abs/1905.12221.

[3] A. Ćiprijanović et al. “DeepMerge: Classifying high-redshift merging galaxies with deep neural networks”. en. In: *Astronomy and Computing, Volume 32, article id. 100390.* 32 (July 2020), p. 100390. ISSN: 2213-1337. DOI: 10.1016/j.ascom.2020.100390.

[4] The Astropy Collaboration et al. “The Astropy Project: Sustaining and Growing a Community-oriented Open-source Project and the Latest Major Release (v5.0) of the Core Package”. en. In: (June 2022). DOI: 10.3847/1538-4357/ac7c74. URL: https://arxiv.org/abs/2206.14220v1.

[5] Sander Dieleman, Kyle W. Willett, and Joni Dambre. “Rotation-invariant convolutional neural networks for galaxy morphology prediction”. In: *Monthly Notices of the Royal Astronomical Society* 450.2 (June 2015), pp. 1441–1459. ISSN: 0035-8711. DOI: 10.1093/mnras/stv632. URL: https://doi.org/10.1093/mnras/stv632.

[6] Timnit Gebru et al. “Datasheets for Datasets”. In: *arXiv:1803.09010 [cs]* (Mar. 2018). arXiv: 1803.09010. URL: http://arxiv.org/abs/1803.09010.

[7] E. W. Greisen. “FITS: A Remarkable Achievement in Information Exchange”. en. In: *Information Handling in Astronomy - Historical Vistas*. Ed. by André Heck. Astrophysics and Space Science Library 285. 00002. Springer Netherlands, Jan. 2002, pp. 71–87. ISBN: 978-1-4020-1178-8 978-0-306-48080-5. URL: http://link.springer.com/chapter/10.1007/0-306-48080-8_5.

[8] *HDF5, Hierarchical Data Format, Version 5*. eng. web page. May 2022. URL: https://www.loc.gov/preservation/digital/formats/fdd/fdd000229.shtml.

[9] Željko Ivezić et al. *LSST: from Science Drivers to Reference Design and Anticipated Data Products*. arXiv:0805.2366 [astro-ph]. May 2018. DOI: 10.3847/1538-4357/ab042c. URL: http://arxiv.org/abs/0805.2366.

[10] Evan Jones et al. *Photometric Redshifts for Cosmology: Improving Accuracy and Uncertainty Estimates Using Bayesian Neural Networks*. Number: arXiv:2202.07121 arXiv:2202.07121 [astro-ph]. Feb. 2022. DOI: 10.48550/arXiv.2202.07121. URL: http://arxiv.org/abs/2202.07121.

[11] C. Laigle et al. “THE COSMOS2015 CATALOG: EXPLORING THE 1 &lt$\mathsemicolon$ $\less$i$\greater$z$\less$/i$\greater$ &lt$\mathsemicolon$ 6 UNIVERSE WITH HALF A MILLION GALAXIES”. en. In: *The Astrophysical Journal Supplement Series* 224.2 (June 2016). Publisher: American Astronomical Society, p. 24. ISSN: 0067-0049. DOI: 10.3847/0067-0049/224/2/24. URL: https://doi.org/10.3847/0067-0049/224/2/24.

[12] Rachel Mandelbaum. “Weak Lensing for Precision Cosmology”. In: *Annual Review of Astronomy and Astrophysics* 56.1 (2018). _eprint: https://doi.org/10.1146/annurev-astro-081817-051928, pp. 393–433. DOI: 10.1146/annurev-astro-081817-051928. URL: https://doi.org/10.1146/annurev-astro-081817-051928.

[13] Martín Abadi et al. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. Software available from tensorflow.org. 2015. URL: https://www.tensorflow.org/.

[14] Jeffrey A. Newman and Daniel Gruen. “Photometric Redshifts for Next-Generation Surveys”. In: *Annual Review of Astronomy and Astrophysics* 60.1 (2022). _eprint: https://doi.org/10.1146/annurev-astro-032122-014611, pp. 363–414. DOI: 10.1146/annurev-astro-032122-014611. URL: https://doi.org/10.1146/annurev-astro-032122-014611.

[15] Adam Paszke et al. “PyTorch: An Imperative Style, High-Performance Deep Learning Library”. In: *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2019, pp. 8024–8035. URL: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

[16] J. Peek and R. White. "Search By Image: Citizen Science and Deep Learning for next-generation archives". In: 53 (June 2021). Conference Name: American Astronomical Society Meeting Abstracts ADS Bibcode: 2021AAS...23830106P, p. 301.06. URL: https://ui.adsabs.harvard.edu/abs/2021AAS...23830106P.

[17] Giuseppe D. Racca et al. "The Euclid mission design". In: arXiv:1610.05508 [astro-ph]. July 2016, 99040O. DOI: 10.1117/12.2230762. URL: http://arxiv.org/abs/1610.05508.

[18] J Sánchez et al. "The LSST DESC data challenge 1: generation and analysis of synthetic images for next-generation surveys". In: *Monthly Notices of the Royal Astronomical Society* 497.1 (Sept. 2020), pp. 210–228. ISSN: 0035-8711. DOI: 10.1093/mnras/staa1957. URL: https://doi.org/10.1093/mnras/staa1957.

[19] N. Scoville et al. "The Cosmic Evolution Survey (COSMOS): Overview". In: *The Astrophysical Journal Supplement Series* 172 (Sept. 2007). ADS Bibcode: 2007ApJS..172....1S, pp. 1–8. ISSN: 0067-0049. DOI: 10.1086/516585. URL: https://ui.adsabs.harvard.edu/abs/2007ApJS..172....1S.

[20] Michael Scroggins and Bernadette M. Boscoe. "Once FITS, Always FITS? Astronomical Infrastructure in Transition". In: *IEEE Annals of the History of Computing* 42.2 (Apr. 2020), pp. 42–54. ISSN: 1058-6180, 1934-1547. DOI: 10.1109/MAHC.2020.2986745. URL: http://arxiv.org/abs/1809.09224.

[21] Alexander S. Szalay et al. "Designing and mining multi-terabyte astronomy archives: the Sloan Digital Sky Survey". en. In: ACM Press, 2000, pp. 451–462. ISBN: 978-1-58113-217-5. DOI: 10.1145/342009.335439. URL: http://portal.acm.org/citation.cfm?doid=342009.335439.

[22] Mike Walmsley et al. "Galaxy Zoo DECaLS: Detailed visual morphology measurements from volunteers and deep learning for 314000 galaxies". In: *Monthly Notices of the Royal Astronomical Society* 509.3 (Jan. 2021), pp. 3966–3988. ISSN: 0035-8711. DOI: 10.1093/mnras/stab2093. URL: https://doi.org/10.1093/mnras/stab2093.

[23] Anne-Marie Weijmans et al. "Streamlining the Sloan Digital Sky Survey Public Data Releases: Changes Made and Lessons Learned". en. In: (), p. 4.

[24] John F. Wu and J. E. G. Peek. "Predicting galaxy spectra from images with hybrid convolutional neural networks". In: *arXiv:2009.12318 [astro-ph]* (Nov. 2020). arXiv: 2009.12318. URL: http://arxiv.org/abs/2009.12318.

Broader Impacts

We believe that our paper may have a positive impact in the field of astronomy, because we suggest best practices that can be incorporated into workflows to improve research practices surrounding the creation, curation, and preservation of datasets for machine learning in astronomy. We feel that potential ethical issues stemming from our work are minimal.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [N/A]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [N/A]

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

   (a) If your work uses existing assets, did you cite the creators? [N/A]

   (b) Did you mention the license of the assets? [N/A]

   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]

   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]

   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...

   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]