Elements of Effective Machine Learning Datasets in Astronomy

Bernie Boscoe [1], Tuan Do [2], Evan Jones [2], Yunqi Li [2], Kevin Alfaro [2] , Christy Ma [2]

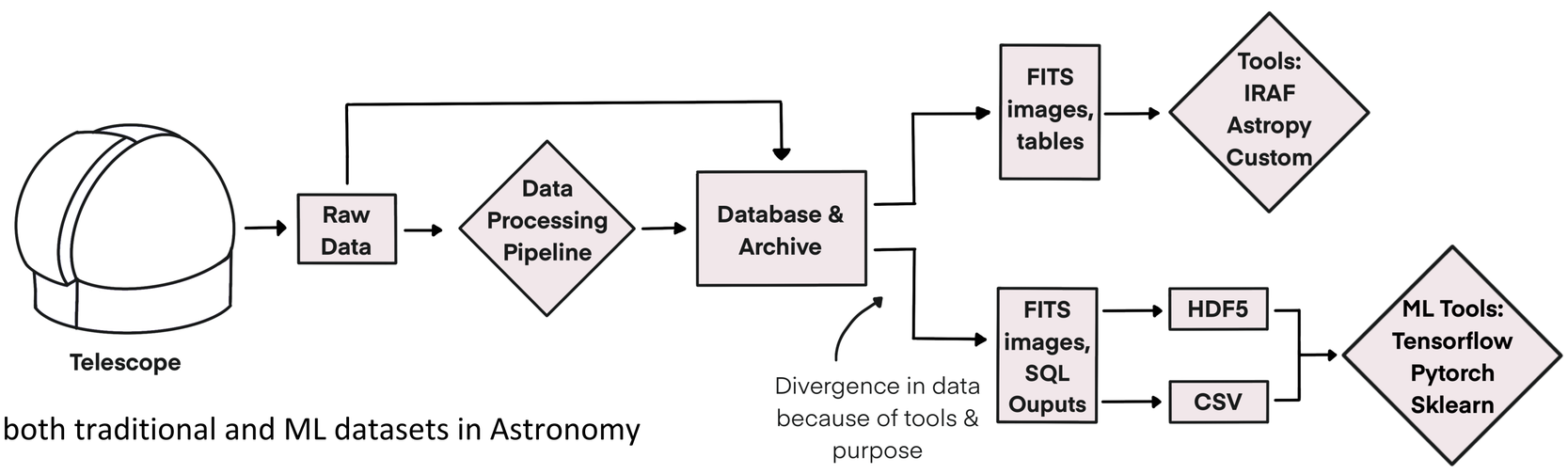
[1] Occidental College, Los Angeles [2] Physics and Astronomy Department, University of California, Los Angeles

Motivation

The construction of datasets for machine learning in astronomy can be challenging and labor-intensive. Astronomical data is collected from instruments built to explore science questions and resulting data forms are not (yet) amenable for machine learning. We ask: *what elements define effective machine learning datasets?* We define effective machine learning datasets to be formed with **well-defined data points**, **structure**, and **metadata.** We posit our suggestions will also foster usable, reusable, and replicable science.

Well-defined data points: transforming upstream data for ML

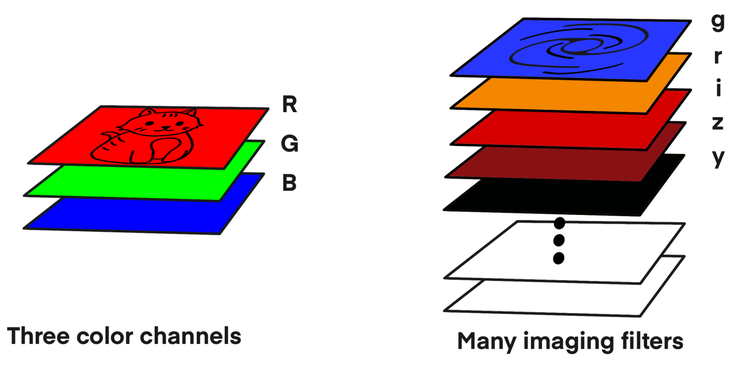
Data points evolve from decision iterations. Data points for ML have particular considerations including:
1) quantification of data point quality, 2) establishing criteria for included data points, 3) establishing outlier criteria, and 4) identifying and potentially removing missing or low-quality data.



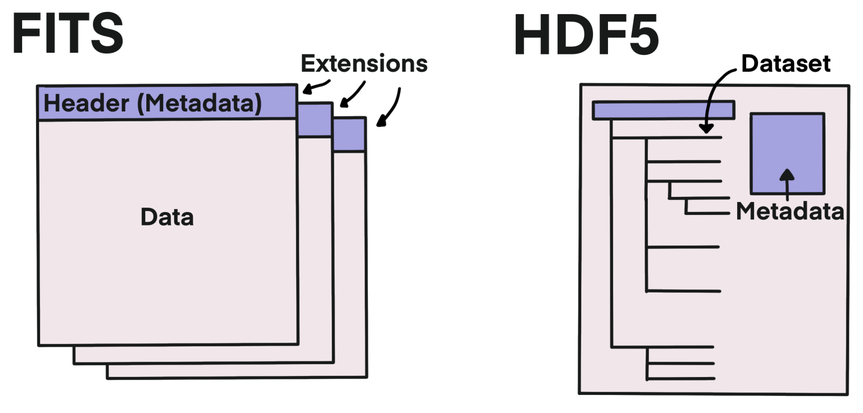
Data flow for both traditional and ML datasets in Astronomy

Well-defined dataset structure: preventing information loss

Machine learning tools like TensorFlow or PyTorch typically use downsized and downscaled image data for training in three or less channel PNG files; whereas astronomy data are observed at many different wavelengths. Preservation of precise measurements is important to construct appropriate ground truths. The FITS file format, while optimal for astronomy, does not lend itself to ML tool ingestion; we recommend the HDF5 file format as a suitable replacement downstream, and is able to preserve information such as multiple image channels, image detail, and precise measurements.



A comparison of channels: left, a color PNG file, right, an example of a galactic image with many layers



-- stable format
-- long history
-- compatible with most astronomical software
-- not easily used for machine learning

-- machine learning-friendly format
-- hierarchical data structure
-- metadata stored as objects
-- read into arrays for batching

Considerations such as data format, tabular shape, image sizes and dimensions can have major impacts on a dataset's efficacy for machine learning.

Well-defined metadata: context is key

Well-defined metadata includes 1) all contextual information relevant to data origins, 2) features and form of the dataset, and 3) motivations for the dataset with respect to the initial scientific goal. One example of information to preserve is the SQL queries used on the archive data from missions or surveys. Versioning schemas for datasets and associated metadata should be attached to publications and/or results.

References

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.

E. W. Greisen. "FITS: A Remarkable Achievement in Information Exchange". en. In: Information Handling in Astronomy - Historical Vistas. Ed. by André Heck. Astrophysics and SpaceScience Library 285. 00002. Springer Netherlands, Jan. 2002, pp. 71–87. ISBN: 978-1-4020-1178-8 978-0-306-48080-5306-48080-8_5.

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