*Machine Learning in Astronomy: Possibilities and Pitfalls*
*Proceedings IAU Symposium No. 368, 2025*
*J. McIver, A. Mahabal & C. Fluke, eds.*
doi:10.1017/S1743921323001394

# How do we design data sets for Machine Learning astronomy?

**Renée Hložek**[1,2]

[1]Dunlap Institute for Astronomy & Astrophysics, 50 St. George Street, Toronto ON M5S 3H4

[2]David A. Dunlap Department for Astronomy & Astrophysics, 50 St. George Street, Toronto ON M5S 3H4
email: hlozek@dunlap.utoronto.ca

**Abstract.** Many problems in astronomy and physics lend themselves to solutions from machine learning methods for the detection and classification of astronomical signals, and model inference from those signals. The historic presentation of machine learning methods as 'black boxes' has generated push back from some in the the physics/astronomy communities regarding how useful they are to truly uncover the physical laws that govern our world. Skepticism about the applicability of new computational methods in scientific inference is not new; we highlight connections between the machine learning contexts and previous computational paradigm shifts in astronomy. Moreover, several advances in methodologies challenge the assumption that machine learning 'gives us answers that we can use but do not understand' to standing physics questions. We summarize some astronomical machine learning data challenges used in astronomy and how we can use challenges on different scales to test different parts/use cases of our analysis methods.

**Keywords.** machine learning, surveys, observations, statistics

## 1. Introduction

We are firmly in the artificial intelligence/machine-learning (AI/ML) assisted age. As the volume of data in astronomy increases to the 'petabyte age', astronomers are constantly looking for methods to process raw data to compressed science-ready products and discard data that are not useful for astronomical inference. Advances in the complexity and robustness of machine learning methods have been equally matched with an increase in their use in astronomy and physics contexts.

### 1.1. *Computational advances lead to both innovation and skepticism*

Machine learning is currently a widely used tool in astronomical research, with advances in algorithm development and methodology driving rapid breakthroughs in previously intractable problems. As breathtaking as this current paradigm is, it mirrors a similar 'revolution' in astronomy: that of simulating large gravitational systems in the late 1960s and 1970s. Advances in computing led to an explosion of the so-called 'N-body' simulations of gravitation. As shown in Figure 1, following their initial use, publications using the phrase 'N-body' increased dramatically following the similar increase in computational power. It is sometimes taken for granted that N-body simulations are an excellent way to study the dynamics of gravitational systems, however when they were first developed there was real skepticism of the validity of the methods and the divergence

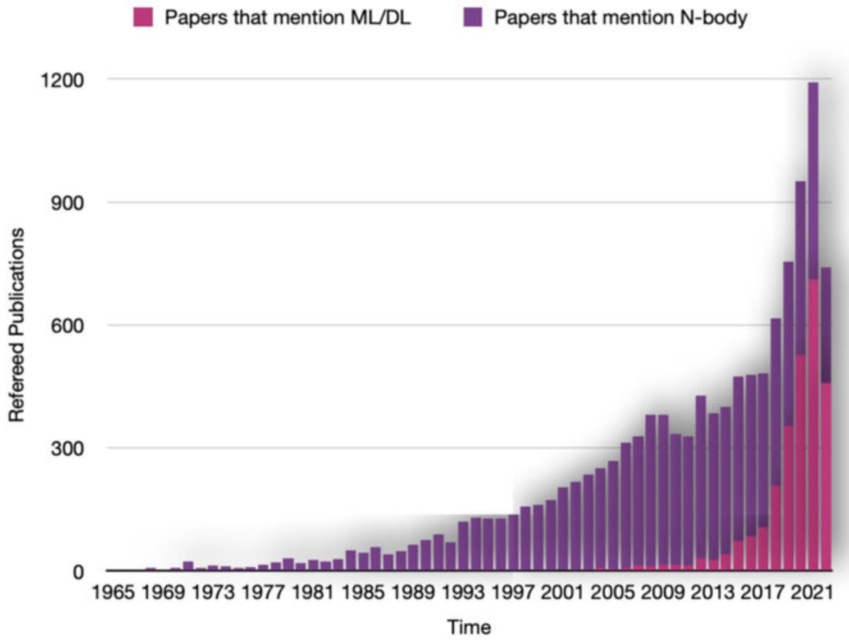



**Figure 1.** The number of refereed publications mentioning artificial intelligence/machine learning (pink) compared to those mentioning N-body methods (purple) as a function of time returned by the NASA Astrophysics Data System online at NASA ADS. The AI/ML revolution can be seen as the rapid growth in the number of publications as a function of time.

between groups following different prescriptions. Similarly, simulations diverged rapidly even with tiny changes in the initial conditions, as noted in early studies Miller (1964, 1971).

In a meeting to discuss these new methods, M. Lecar issued a word of caution and suggested the need for further comparative studies, given the one attempted for the 1968 meeting of researchers in dynamics:

*It seems that numerical experiments are becoming increasingly popular, and before things get out of hand, I feel that further tests of reproducibility are imperative. However, because of the difficulties in handling volumes of data from different computers, it would be worthwhile for each of us first to initiate statistical studies of our own; varying the accuracy and the microscopic initial conditions, and searching for stable quantities. These results should be sent to all the participants in this study, and perhaps by next summer, we can hold a second comparison study. Lecar (1968)*

As summarized in Quinlan & Tremaine (1992), this sensitivity of N-body simulations to small errors was not simply a numerical artifact that would dissolve with additional computational power, and so the community had to establish (and compare) best practices for how to regularize the integration methods in these systems. The community did not, however, abandon the N-body methods due to these potential pitfalls. Comparisons across groups on 'standard systems', clear descriptions of methods and exploration of the impacts of methodological assumptions has led to the continued success of this technique in astronomy for systems across a range of scales.

https://doi.org/10.1017/S1743921323001394 Published online by Cambridge University Press

## 1.2. *What is machine learning in astronomy used for?*

Machine learning methods in astronomy have been used since the mid-late 1990s (see Baron 2019, for an excellent comparison of different algorithms and their application in astronomy), with early methods focusing on spectral analysis (Connolly et al. 1995; Boroson & Green 1992) and principal component analysis slowly replaced with more recent use cases including deep learning, neural networks and diffusion methods (Cireşan et al. 2012; Fukushima 1980; Ho et al. 2020; Ćiprijanović et al. 2022). The use cases in astronomy can be summarized in roughly three categories, namely the *detection* of signals in a faint and often varying background; the *clustering/classification* of objects into various categories through either unsupervised (clustering) or supervised (classification) methods; and statistical *inference* of certain parameters or models allowed by a given data set.

### *1.2.1. Signal detection*

Astronomical signals can vary from the 'extremely bright/loud' bursts of radiation to faint hints of signal within a bright background. Examples of bright objects include Fast Radio Bursts (FRBs), which exhibit a characteristic 'sweep' in frequency-time space as the signal passes through the electrons along the line of sight. First detected in 2007 (Lorimer et al. 2007), fast radio bursts are exciting new objects in the sky that display brighten and fade rapidly in radio bands (ranging from milliseconds to a few seconds in duration). Some objects repeat this brightening/fading, while others appear to brighten only once (The CHIME/FRB Collaboration et al. 2023). Several surveys are optimized to find many of these new objects in the hopes of uncovering their origin and the physical mechanism behind the bursts (CHIME/FRB Collaboration et al. 2021; Connor et al. 2023; Vanderlinde et al. 2019; Lin et al. 2022; van Leeuwen et al. 2023; Megias Homar et al. 2023). The gravitational wave signatures of the in-fall of pairs of massive bodies like black holes and neutron stars also have a characteristic shape, but are a much fainter signal relative to background effects. First detected in 2015 (Abbott et al. 2016), these signals are powerful probes of the model of General Relativity and paired with electromagnetic counterparts, they offer a window into the energetics of these powerful events. Detecting the signals above the background and/or distinguishing the signal from an instrumental artifact is an active area of research (Alvarez-Lopez et al. 2023; Bini et al. 2023; George & Huerta 2018b,a; Razzano et al. 2023; Wang et al. 2023; Cabero et al. 2020). Assigning an astrophysical host to both fast radio bursts and gravitational wave sources is complicated by the uncertainty in spatial localization of the signal and the density of galaxies in the sky.

### *1.2.2. Clustering and classification*

Once astronomical signals have been detected, a separate goal is to separate the signals into different classes (or sub-classes) of objects. Supervised approaches use sets of real or simulated data with labels to train machine learning algorithms to recognize characteristic features in the data. These data typically contain changing flux in a given electromagnetic band (optical, radio, X-ray etc.) as a function of time. In the case of FRBs, various machine learning methods have been developed to classify these objects and distinguishing them apart from radio frequency interference (RFI) signals present in the data, including supervised (Connor & van Leeuwen 2018; Zhang et al. 2020; Agarwal et al. 2020) and unsupervised methods Zhu-Ge et al. (2023). In the case of the gravitational wave data, recent approaches separate the detected signals into different classes based on models of distances to the detected signals (T. C. Abbott et al. 2022).

https://doi.org/10.1017/S1743921323001394 Published online by Cambridge University Press

In both cases, very rapid follow-up of sources by complementary telescopes maximizes the scientific return of the observation, and so classifiers need to return results in a short period of time. Similar time constraints exist for the classification of events in gamma ray and X-ray telescopes, where detections of interest need to be flagged and communications sent out to other telescopes for 'target-of-opportunity' observations (Tohuvavohu et al. 2020; DeLaunay & Tohuvavohu 2022).

In optical astronomy, time-series classification and clustering has enjoyed a longer history of development. In this case detection itself is less of a challenge as the signals typically trigger a discovery 'alert' depending on a brightness threshold relative to the background noise. 'Forced' or archival photometric data are then produced from observations of the same location on the sky before the outburst/detection to determine the pre-outburst brightness of the host or environment.

#### *1.2.3. Inference*

While ML algorithms can provide predictive results for the detection or classification of new signals within noisy data, physicists are driven by the need to understand underlying processes that govern the behaviour of stars, galaxies, planets and the cosmos. In order to infer parameters of the model and uncertainties, very different machine learning architectures must be used. One method used to perform inference using ML is using Bayesian Neural Networks. In this formalism, weights and biases attached to the neural network are not specified as individual parameters but with probability distributions. The result of these distributions is that they can yield *post facto* distributions in the learned parameters of the network, providing model output that is more interpretable and transferable. While the approach of adding uncertainty to model networks was around from the early 1990s (Denker & LeCun 1990; Buntine & Weigend 1991; MacKay 1992), the application to modern machine learning gathered steam in recent years (Blundell et al. 2015) and is now commonly used in astrophysics to build models of galaxy formation and morphology (Walmsley et al. 2019; Dunn et al. 2023; Reza et al. 2022; Piras et al. 2023; Lucie-Smith et al. 2023), supernova modelling (Stein et al. 2022) and large-scale structure clustering (Sullivan et al. 2023; Modi et al. 2021), to name but a few examples.

## 2. Some Pitfalls/Critiques of Machine learning in Astronomy

In order to take full advantage of the machine assisted revolution in astronomy, we need to understand and address some of the critiques leveled at the use of machine learning in astronomy. We list some of the common pitfalls and describe attempts by the community to address them.

### 2.1. *Machine Learning provides 'black boxes' without any physics understanding*

The use of machine learning in astronomy/physics is often criticized for being a 'black box' that is not understood, or that parts of the model are not transparently discussed. As Rudin (2018) defines it, this black box could be because the model is a function that is too complicated (e.g. recursive) for any human to comprehend or that the function contains parts that are proprietary (something more common in medical, rather than astronomical applications). A key distinction that Rudin makes is between the *explainability* of the black box/algorithm, and the *interpretability* of the model. It is the latter that is of most interest in physics, as whatever model is used, we want it to illuminate wherever possible the underlying physical processes that the system is trying to model. Moreover, some astronomers see machine learning models as tools to use within the context of their analysis, and use off-the-shelf methods and tools to ensure speed and robustness of their

https://doi.org/10.1017/S1743921323001394 Published online by Cambridge University Press

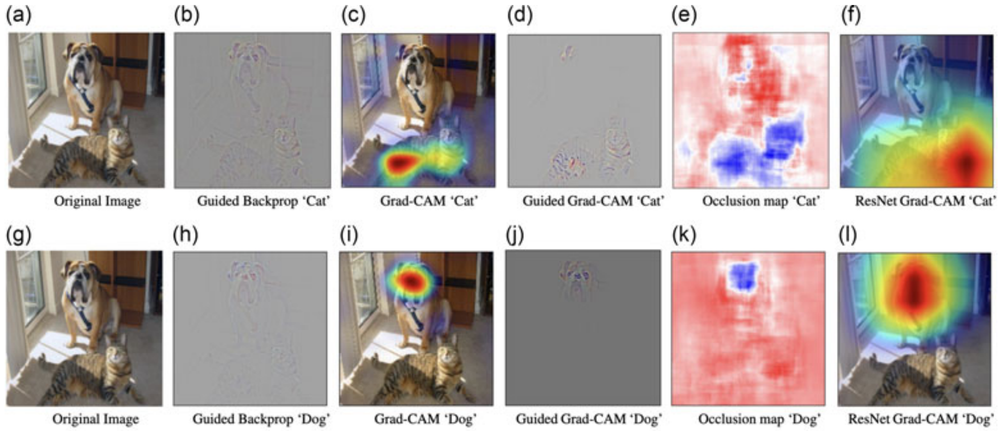



**Figure 2.** The Grad-CAM algorithm uses a class activation map weighted by the gradient information flowing into the final layer of a CNN to determine which parts of the data (in this case an image) are yielding the classification of 'cat' or 'dog'. Figure reproduced from (Selvaraju et al. 2016).

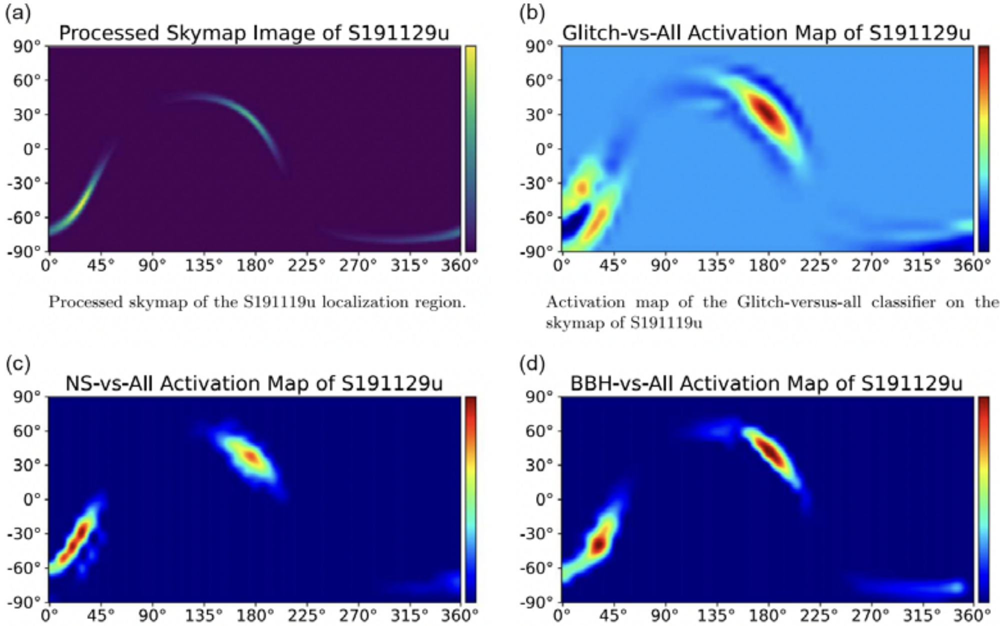
analysis pipelines. As such, less time is often spent on either justifying the approach used (or the selection of the network architecture) and explaining why it is useful in the specific astronomical context. This 'means-to-an-end' approach can perpetuate the myth that these tools lack value or that they do not allow for deeper understanding of the physics.

As outlined in Grojean et al. (2022), some methods have been introduced to develop *post facto* interpretability of trained ML models, including 'local' methods that trace every outcome in terms of all the input variables provided (e.g. LIME, Tulio Ribeiro et al. 2016), or more global methods analyse which variables are the most 'important' to the algorithm, or that produce the most variation of output through variables like the Gini coefficient (Giorgi & Gigliarano 2017).

In order to understand what their machine-learning algorithms are learning (or rather, what in the data they are responding to), methods like Gradient-weighted Class Activation Mapping (Grad-CAM, Selvaraju et al. 2016) to determine what parts of an image determine the classification/typing of the image. Grad-CAM is based on the Class Activation Mapping (CAM, Zhou et al. 2015) method which creates class activation maps by using the global average pooling performed in Convolutional Neural Networks (CNNs Fukushima 1980) to indicate parts of the image associated with the target category. Illustrated in Figure 2, the Grad-CAM algorithm modifies this by using the gradient information flowing into the final convolutional layer of the network as a measure of the importance of each neuron to the particular region of the image/data in determining the classification. This method is more robust to the underlying CNN and can provide useful insight into what features your machine learning method is actually learning. This has successfully been applied to the astronomical context in determining what the ML algorithms were learning when determining whether a gravitational wave signal was astrophysical or an instrumental glitch (T. C. Abbott et al. 2022, as shown in Figure 3).

Others use the approach of using regressive algorithms to 'directly infer' physical laws themselves. Schmidt & Lipson (2009) suggested that they could recover natural physical laws from observations alone by using machine learning methods, however further inspection by Hillar & Sommer (2012) suggested that their fitness functions within the algorithm incorporated Hamilton's equations of motions and Newton's second law

https://doi.org/10.1017/S1743921323001394 Published online by Cambridge University Press




Processed skymap of the S191119u localization region.

Activation map of the Glitch-versus-all classifier on the skymap of S191119u

Activation map of the NS-versus-all classifier on the skymap of S191129u.

Activation map of the BBH-versus-all classifier on the skymap of S191129u

**Figure 3.** Grad-CAM as applied to `GWSkyNet` classifications of binary black hole or neutron star pairs. Figure reproduced from (T. C. Abbott et al. 2022).

directly. Hillar & Sommer (2012) suggest ways to enforce a general fitness function that is not prone to this bias.

In the past decade, physics-informed machine learning models (Raissi et al. 2019) incorporate physical laws (for example those that govern the time-dependent dynamics of the system or those that control the energy budget) directly into the algorithm as prior information, rather than applying a generic machine-learning algorithm to the problem and assuming no prior information. Including these laws as rules into the learning algorithm itself can lead to more efficient and robust algorithm performance on complex problems in physics, and can reduce the dependence on training, or allow unsupervised training altogether. See (Karniadakis et al. 2021) for a review of these methods and their applicability to problems in physics and mathematics.

Applying these physics-aware or physics-based machine learning approaches to solve problems in physics and astronomy is a growing area of study and has particular applicability in the study of nonlinear dynamics and fluid mechanics (see e.g. Shukla et al. 2022; Dai et al. 2023; Rosofsky et al. 2022; Karpov et al. 2022; Iyer et al. 2022; Lucie-Smith et al. 2023; Ntampaka et al. 2021). A challenge suggested in Nord et al. (2019) is to facilitate the use of these physics-informed algorithms to solve some of the large and complex modeling required in the current age of large astronomical data.

### 2.2. *Machine learning does not provide error estimates*

Another criticism leveraged against machine learning methods historically has been a lack of meaningful error estimates from 'black box' models, however the past decade has seen the introduction of ML approaches specifically designed to provide uncertainty estimates on the parameters of interest within a model. These innovations were developed to enable machine learning *inference* of the model parameters (a few examples of

https://doi.org/10.1017/S1743921323001394 Published online by Cambridge University Press

current uses of inference in astronomy are included above). In fact, there is a move in the community to compare different approaches of uncertainty quantification (UQ) in physics and astronomy (see Caldeira & Nord 2020; Psaros et al. 2023; Ćiprijanović et al. 2022; Mohan et al. 2022; Ntampaka & Vikhlinin 2022, for some recent examples).

Machine-learning approaches and methods are being developed at a somewhat faster pace than the general astronomical community up-take. However, community standards continue to evolve towards these more interpretable approaches. Astronomy research planning efforts in e.g. the United States and Canada occur every decade and included these planning efforts are white papers and discussions on machine learning trends, resources and challenges – provided a natural platform for the discussions to move the field towards methods that can provide uncertainty estimates, are interpretable and physically motivated (e.g. Nord et al. 2019; Dvorkin et al. 2022; Siemiginowska et al. 2019).

### 2.3. *Machine learning compresses (and loses) information*

In order to process large data sets through machine learning algorithms, these data are often 'compressed' onto a smaller set of features/axes/vectors of information. While this can be criticized as 'losing information' in the system (e.g. by splitting up a video into many still frames, thereby potentially losing the temporal connection between successive frames), it is also important to address the problem of over-fitting in ML analyses. This occurs when the ML model starts fitting to noise in the data set, given enough model flexibility.

Deep learning methods were developed specifically to handle very large and complex data sets (e.g. 2D images and human speech) where there is a risk of under-fitting/not building a complete and predictive model. Changes to model complexity and network architecture can be made to ensure that the model does not have too much flexibility and freedom if there is concern of overfitting noise. Here we can learn from colleagues in optics and machine vision, who investigate when (and what) data to add into a system to obtain model that is optimally fit and that remains predictive (Boulahia et al. 2021). The challenge remains for the astronomy community to test and review any ML framework used for a specific analysis rather than dismissing the tools themselves. Some groups are already thinking at this abstract level and evaluating robustness, (Ćiprijanović et al. 2022), and examining how we can generalize the generating and simulating of data sets for such machine-learning contests (Lewis et al. 2022).

## 3. Leveraging data challenges in Astronomy machine learning

Astronomy has a rich history of simulating data sets in advance of telescope operations, and through regular advances in telescope development, has been in a 'data-rich' paradigm for many years. While the era of 'data-deluge' from large survey instruments like the Vera C. Rubin Observatory (Rubin Obs, LSST Science Collaboration et al. 2009; Bianco et al. 2022; LSST Dark Energy Science Collaboration 2012) and the Square Kilometer Array (SKA, Square Kilometre Array Cosmology Science Working Group et al. 2020) will soon be upon us, using current surveys (see e.g. van Roestel et al. 2021; T. M. C. Abbott et al. 2022) and simulations of the sky we will observe with these new facilities (LSST Dark Energy Science Collaboration (LSST DESC) et al. 2021; Levrier et al. 2009; Ramírez-Pérez et al. 2022). This ability to simulate, and to test pipelines on existing data sets with 'known' ground truth, and the community standard of transferring methods to different observing conditions and telescope configurations is a competitive advantage in astronomy, and one that distinguishes us from other data-rich fields where simulations are either extremely costly to obtain, complicated, or where data

https://doi.org/10.1017/S1743921323001394 Published online by Cambridge University Press

are highly individualized due to experimental conditions (An 2018; Klingner et al. 2022; Errington et al. 2021).

*What makes an effective astronomical machine-learning data challenge?* Boscoe et al. (2022) describe how effective astronomical data challenges need to have well-defined data points (including only what is relevant to the challenge, ensuring there are tags for data quality, including well-modeled uncertainties) and even make recommendations for the structural form of data (HDF5 rather than the common FITS format that is so common in astronomy. In addition to the data format, it is important to have well-structured and easy to access metadata, to enable supervised learning without the full domain knowledge that is sometimes mistakenly assumed of astronomers participating in these challenges. These recommendations for the well-defined data, data-structure and metadata can be encapsulated in the recommendation that the data challenge *goals are clearly defined* and articulated to participants, that *any required domain knowledge is communicated* and that the *performance metrics are well-established and communicated*. This can be challenging given the multi-variate needs of the astronomical community interested in a particular data challenge, however the preparatory work in defining goals, assumed knowledge and metrics can greatly change the number of potential participants and the success of the challenge. We highlight a few astronomical challenges below.

### 3.1. *Time-series classification challenges*

The sky has some objects in it (galaxies, groups and clusters of galaxies) that change slowly with time, other objects that fluctuate in brightness (e.g. energetic jets from the centre of galaxies, spinning neutron stars or pulsating stars nearing stellar death) and 'new' or transient objects that are a result of explosions, the collision of pairs of massive objects or other undefined phenomena. Classifying these transient objects into different categories is needed in order to analyze specific samples/types, and astronomers have been developing tools for rapid classification in tandem with the increase in the survey speed of new telescopes. Early surveys like the Catalina Real-Time Transient Sky Surveys (Drake et al. 2012) and the Palomar-Quest Digital Synoptic Sky Survey (Djorgovski et al. 2008) ushered in an era of wide-field transient science and presented lessons learned for processing of transient notifications or 'alerts' (see e.g. Graham et al. 2012). The Sloan Digital Sky Survey (Kessler et al. 2009; Sako et al. 2018) was focused on developing a sample of Type Ia supernovae (SNIa) for cosmological analysis and developed methods to classify the observed transients into different types to facilitate spectroscopic follow up of likely SNIa candidates.

Recognizing the challenges that this classification had presented, Kessler *et al.* developed the Supernova Photometric Classification Challenge (SNPhotCC, Kessler et al. 2010) which was a classification challenge based on simulated supernovae of three different types assuming the photometric properties of a survey like the (then upcoming) Dark Energy Survey. The challenge led to broad community participation (Kessler et al. 2010) and engaged new groups in optical classification. The challenge highlighted that the community particularly struggled with the increase of data (and increase in noise) with survey depth, and that the change in relative numbers of different populations when going from precursor data to larger volume data sets. As a result, when simulating a new survey for the upcoming Rubin Observatory, the Photometric LSST Astronomical Time-series Classification Challenge (PLAsTiCC, Kessler et al. 2019; Hložek et al. 2020; Malz et al. 2018) included many more types of transients, a large difference in the size of the training sample relative to the test sample (as shown in Figure 4), and a collection of rare objects that were explicitly excluded from any labelled training data. The three-month PLAsTiCC (hosted on a public data challenge website at Kaggle.com) had over 1000 participants from within and outside the astronomical community and has

https://doi.org/10.1017/S1743921323001394 Published online by Cambridge University Press

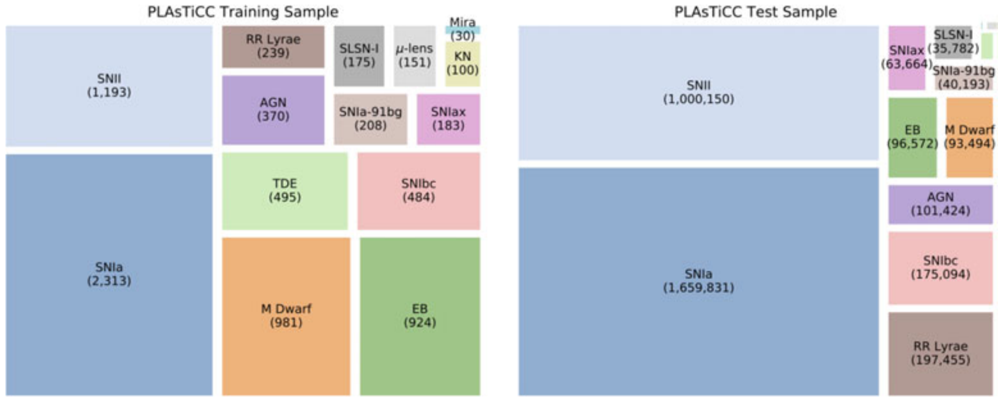



**Figure 4.** The $\simeq 8000$ objects in the training set and 3.5 million test objects in the PLAsTiCC survey. The size of the boxes are proportional to the relative numbers in each set and the absolute numbers are in parentheses. Figure reproduced from Hložek et al. (2020).

already led to 45 publications on methods and approaches to classification and anomaly detection.

This challenge included time-series data for the full LSST survey from Rubin, as processed flux measurements. The primary way that new transients will be communicated from the telescope, however, is through packages of 'alerts' released by the telescope to a specific set of alert brokers developed within the community. The next instalment in developing classification challenges was to provide the light curves in this 'alert' format, and to increase the complexity of the simulations to control for host environment (Lokken et al. 2023). The Extended LSST Astronomical Time-series Classification Challenge (ELAsTiCC, Narayan & ELAsTiCC Team 2023) was focused at these brokers, testing their ability to ingest and process (and classify) a stream of alerts similar to what will be generated from Rubin. The challenge is available online.

PLAsTiCC and ELAsTiCC are broad challenges including a range of different object types. Narrower, domain-specific classification challenges can include more model variety and employ more specific science-focused metrics for evaluating performance between methods. The LSST AGN classification challenge (Yu et al. 2022) provided mock observations based on templates from existing data and focused on assessing the ability of participants to select AGN and parameterize their light-curves, and the ability to determine the photometric redshift of the AGN candidate.

While not considered 'transient', finding planets around other stars requires similar techniques as the exoplanets often transit their host stars, or generate faint shifts to the motion of the host star. The Radial Velocity Challenge (Dumusque et al. 2017) simulated radial velocity signals (Dumusque 2016) from the 'wobble' that an exoplanet would generate for its host star. The challenge was focused on testing methods in advance of the Transiting Exoplanet Survey Satellite (TESS, Ricker et al. 2014).

The Laser Interferometer Space Antenna (LISA) Data challenges (Baghi 2022) are a series of challenges designed to test community methods for extracting gravitational wave signals from noisy data in preparation for the European Space Agency (ESA)- and NASA-funded LISA mission, with the current purpose of tackling mild source confusion within an idealized instrumental noise framework. The current challenge continues a long line of previous challenges developed by the Mock LISA Data Challenge task force (Babak et al. 2008) to prepare the community for LISA analysis.

Similarly, the Australian Square Kilometre Array Pathfinder (ASKAP) group designed a data challenge to test the fidelity of radio source finding (Hopkins et al. 2015) for

https://doi.org/10.1017/S1743921323001394 Published online by Cambridge University Press

the SKA pathfinder, and led to other challenges focused on specific telescopes like the PARKES radio telescope (Yong et al. 2022) and the SKA itself (Hartley et al. 2023), and the development of new methods and approaches (Vafaei Sadr et al. 2019; Riggi et al. 2019; Bonavera et al. (2021).

### 3.2. *Image analysis challenges*

While transient surveys focus on identifying specific types of objects that appear/disappear or vary in brightness on the sky, astronomers interested in the 'static sky' are often more concerned with measuring the shape of objects, classifying objects into different types (e.g. spiral or elliptical galaxies).

For example, the GRavitational lEnsing Accuracy Testing challenges (GREAT, Mandelbaum et al. 2014; Kitching et al. 2010) required participants to develop lensing shear estimators and compared performance on increasingly complex simulated test data. The LSST-DESC 3x2pt Tomography Optimization Challenge (Zuntz et al. 2021), was challenge with a more specialized purpose: to determine the optimal tomographic binning scheme for a photometric survey for measurements of the '3x2pt' (the three different types of two-point correlation function measurements used as the data observable that summarizes the clustering of galaxies, and the shapes of those galaxies as they are distorted by cosmic lensing.) This challenge, targeted at members of the LSST Dark Energy Science Collaboration (DESC) was a way to galvanize the DESC community to generating an optimal method of defining samples for the upcoming Rubin LSST survey.

The above lensing challenges focus on the subtle affects from weak gravitational lensing. 'The Strong Gravitational Lens Finding Challenge' (Metcalf et al. 2018) instead asked participants to identify 100,000 candidate strong lens systems from millions of noisy images. By comparing algorithm performance across all entries, the challenge presented also gained new insights e.g. for the need for multi-band imaging in order to correctly identify strong lens systems.

### 3.3. *The power of citizen science*

While many challenges in astronomy are reasonably technical and can require some amount of domain knowledge, large-scale challenges that take advantage of the energy of 'citizens' can generate large amounts of data and find novel objects. One of the classic examples of citizen science is the Galaxy Zoo project (Lintott et al. 2008), which started off asking citizens to make simple classifications between spiral and elliptical galaxies in SDSS images, but has grown to encompass citizen science challenges across disciplines (see the full suite at zooniverse.org) and led to the discovery in astronomy of novel objects like the 'green pea' galaxies, named after their greenish appearance and small size in the SDSS images. Further study of these new objects discovered by the citizen scientists revealed their role in the reionization of the universe (Cardamone et al. 2009). The classifications from citizen science volunteers can then be used to train and refine machine-learning methods for discovery and classification (Andersson et al. 2023; Peek & White 2021; Kruk & Merín 2023; Razzano et al. 2023; Jimenez et al. 2023).

## 4. Conclusions

Data challenges are an important step in preparing for the complexity of future data sets, and for developing new approaches to analyze current data. Astronomers have the benefit of extensive and well-labelled existing data sets, and the community practice of simulating new data with different intrinsic properties and observing conditions. When developing a data challenge for training/testing machine learning methods in astronomy, the first step should be reducing and controlling the complexity of the data presented, and deciding on the limited number of goals of the challenge.

These astronomical challenges provide great ways of galvanizing the astronomical community to develop new methods, and if well-curated the data set will remain useful well beyond the individual data challenge. It may be that data challenges need to be developed in stages, increasing the complexity of the data used in each stage, and relaxing additional assumptions about metadata or data complexity with each challenge. Rather than seeing this as a bug, it can be approached as a smooth way to understand the community response to data complexity.

A challenge that remains for the community is to include rewards for methods that are physically interpretable, and to develop standards for how to communicate and derive machine learning solutions to problems in astronomy. We are lucky as a community to continually have exciting new data on the horizon, and look forward to more machine-assisted solutions to understand the Universe.

## References

Abbott, B. P., et al. (2016). Tests of general relativity with GW150914. *Phys. Rev. Lett.*, *116*(22), 221101. ([Erratum: Phys.Rev.Lett. 121, 129902 (2018)]) doi: 10.1103/PhysRevLett.116.221101

Abbott, T. C., Buffaz, E., Vieira, N., Cabero, M., Haggard, D., Mahabal, A., & McIver, J. (2022, March). GWSkyNet-Multi: A Machine-learning Multiclass Classifier for LIGO-Virgo Public Alerts. *ApJ*, *927*(2), 232. doi: 10.3847/1538-4357/ac5019

Abbott, T. M. C., Aguena, M., Alarcon, A., Allam, S., Alves, O., Amon, A., ... DES Collaboration (2022, January). Dark Energy Survey Year 3 results: Cosmological constraints from galaxy clustering and weak lensing. *Phys. Rev. D*, *105*(2), 023520. doi: 10.1103/PhysRevD.105.023520

Agarwal, D., Aggarwal, K., Burke-Spolaor, S., Lorimer, D. R., & Garver-Daniels, N. (2020, September). FETCH: A deep-learning based classifier for fast transient classification. *MNRAS*, *497*(2), 1661–1674. doi: 10.1093/mnras/staa1856

Alvarez-Lopez, S., Liyanage, A., Ding, J., Ng, R., & McIver, J. (2023, April). GSpyNet-Tree: A signal-vs-glitch classifier for gravitational-wave event candidates. *arXiv e-prints*, arXiv:2304.09977. doi: 10.48550/arXiv.2304.09977

An, G. (2018, September). The crisis of reproducibility, the denominator problem and the scientific role of multi-scale modeling. *Bull Math Biol*, *80*(12), 3071–3080.

Andersson, A., Lintott, C., Fender, R., Bright, J., Carotenuto, F., Driessen, L., ... Whittle, I. (2023, April). Bursts from Space: MeerKAT - The first citizen science project dedicated to commensal radio transients. *arXiv e-prints*, arXiv:2304.14157. doi: 10.48550/arXiv.2304.14157

Babak, S., Baker, J. G., Benacquista, M. J., Cornish, N. J., Crowder, J., Cutler, C., ... Challenge-2 participants, t. (2008, June). Report on the second Mock LISA data challenge. *Classical and Quantum Gravity*, *25*(11), 114037. doi: 10.1088/0264-9381/25/11/114037

Baghi, Q. (2022, April). The LISA Data Challenges. *arXiv e-prints*, arXiv:2204.12142. doi: 10.48550/arXiv.2204.12142

Baron, D. (2019, April). Machine Learning in Astronomy: a practical overview. *arXiv e-prints*, arXiv:1904.07248. doi: 10.48550/arXiv.1904.07248

Bianco, F. B., Ivezić, Ž., Jones, R. L., Graham, M. L., Marshall, P., Saha, A., ... Willman, B. (2022, January). Optimization of the Observing Cadence for the Rubin Observatory Legacy Survey of Space and Time: A Pioneering Process of Community-focused Experimental Design. *ApJS*, *258*(1), 1. doi: 10.3847/1538-4365/ac3e72

Bini, S., Vedovato, G., Drago, M., Salemi, F., & Prodi, G. A. (2023, March). An autoencoder neural network integrated into gravitational-wave burst searches to improve the rejection of noise transients. *arXiv e-prints*, arXiv:2303.05986. doi: 10.48550/arXiv.2303.05986

Blundell, C., Cornebise, J., Kavukcuoglu, K., & Wierstra, D. (2015, May). Weight Uncertainty in Neural Networks. *arXiv e-prints*, arXiv:1505.05424. doi: 10.48550/arXiv.1505.05424

Bonavera, L., Suarez Gomez, S. L., González-Nuevo, J., Cueli, M. M., Santos, J. D., Sanchez, M. L., ... de Cos, F. J. (2021, April). Point source detection with fully convolutional networks. Performance in realistic microwave sky simulations. *A&A*, *648*, A50. doi: 10.1051/0004-6361/201937171

Boroson, T. A., & Green, R. F. (1992, May). The Emission-Line Properties of Low-Redshift Quasi-stellar Objects. *ApJS*, *80*, 109. doi: 10.1086/191661

Boscoe, B., Do, T., Jones, E., Li, Y., Alfaro, K., & Ma, C. (2022, November). Elements of effective machine learning datasets in astronomy. *arXiv e-prints*, arXiv:2211.14401. doi: 10.48550/arXiv.2211.14401

Boulahia, S. Y., Amamra, A., Madi, M. R., & Daikh, S. (2021, nov). Early, intermediate and late fusion strategies for robust deep learning-based multimodal action recognition. *Mach. Vision Appl.*, *32*(6). doi: 10.1007/s00138-021-01249-8

Buntine, W. L., & Weigend, A. S. (1991). Bayesian back-propagation. *Complex Syst.*, *5*.

Cabero, M., Mahabal, A., & McIver, J. (2020, November). GWSkyNet: A Real-time Classifier for Public Gravitational-wave Candidates. *ApJ*, *904*(1), L9. doi: 10.3847/2041-8213/abc5b5

Caldeira, J., & Nord, B. (2020, April). Deeply Uncertain: Comparing Methods of Uncertainty Quantification in Deep Learning Algorithms. *arXiv e-prints*, arXiv:2004.10710. doi: 10.48550/arXiv.2004.10710

Cardamone, C., Schawinski, K., Sarzi, M., Bamford, S. P., Bennert, N., Urry, C. M., ... Vandenberg, J. (2009, November). Galaxy Zoo Green Peas: discovery of a class of compact extremely star- forming galaxies. *MNRAS*, *399*, 1191–1205. doi: 10.1111/j.1365-2966.2009.15383.x

CHIME/FRB Collaboration, Amiri, M., Andersen, B. C., Bandura, K., Berger, S., Bhardwaj, M., ... Zwaniga, A. V. (2021, December). The First CHIME/FRB Fast Radio Burst Catalog. *ApJS*, *257*(2), 59. doi: 10.3847/1538-4365/ac33ab

Ćiprijanović, A., Kafkes, D., Snyder, G., Sánchez, F. J., Perdue, G. N., Pedro, K., ... Wild, S. M. (2022, September). DeepAdversaries: examining the robustness of deep learning models for galaxy morphology classification. *Machine Learning: Science and Technology*, *3*(3), 035007. doi: 10.1088/2632-2153/ac7f1a

Cireşan, D., Meier, U., & Schmidhuber, J. (2012, February). Multi-column Deep Neural Networks for Image Classification. *arXiv e-prints*, arXiv:1202.2745. doi: 10.48550/arXiv.1202.2745

Connolly, A. J., Szalay, A. S., Bershady, M. A., Kinney, A. L., & Calzetti, D. (1995, September). Spectral Classification of Galaxies: an Orthogonal Approach. *AJ*, *110*, 1071. doi: 10.1086/117587

Connor, L., Ravi, V., Catha, M., Chen, G., Faber, J. T., Lamb, J. W., ... Yadlapalli, N. (2023, February). Deep Synoptic Array science: Two fast radio burst sources in massive galaxy clusters. *arXiv e-prints*, arXiv:2302.14788. doi: 10.48550/arXiv.2302.14788

Connor, L., & van Leeuwen, J. (2018, December). Applying Deep Learning to Fast Radio Burst Classification. *AJ*, *156*(6), 256. doi: 10.3847/1538-3881/aae649

Dai, Z., Moews, B., Vilalta, R., & Dave, R. (2023, March). Physics-informed neural networks in the recreation of hydrodynamic simulations from dark matter. *arXiv e-prints*, arXiv:2303.14090. doi: 10.48550/arXiv.2303.14090

DeLaunay, J., & Tohuvavohu, A. (2022, December). Harvesting BAT-GUANO with NITRATES (Non-Imaging Transient Reconstruction and Temporal Search): Detecting and Localizing the Faintest Gamma-Ray Bursts with a Likelihood Framework. *ApJ*, *941*(2), 169. doi: 10.3847/1538-4357/ac9d38

Denker, J. S., & LeCun, Y. (1990). Transforming neural-net output levels to probability distributions. In *Advances in neural information processing systems 3* (p. 853–859). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Djorgovski, S. G., Baltay, C., Mahabal, A. A., Drake, A. J., Williams, R., Rabinowitz, D., ... Ellman, N. (2008, March). The Palomar-Quest digital synoptic sky survey. *Astronomische Nachrichten*, *329*(3), 263. doi: 10.1002/asna.200710948

Drake, A. J., Djorgovski, S. G., Mahabal, A., Prieto, J. L., Beshore, E., Graham, M. J., ... Williams, R. (2012, April). The Catalina Real-time Transient Survey. In E. Griffin, R. Hanisch, & R. Seaman (Eds.), *New horizons in time domain astronomy* (Vol. 285, p. 306–308). doi: 10.1017/S1743921312000889

Dumusque, X. (2016, Aug). Radial velocity fitting challenge. I. Simulating the data set including realistic stellar radial-velocity signals. *A&A*, *593*, A5. doi: 10.1051/0004-6361/201628672

Dumusque, X., Borsa, F., Damasso, M., Díaz, R. F., Gregory, P. C., Hara, N. C., ... Udry, S. (2017, February). Radial-velocity fitting challenge. II. First results of the analysis of the data set. *A&A*, *598*, A133. doi: 10.1051/0004-6361/201628671

Dunn, M., Ćiprijanović, A., Nord, B., & Mobasher, B. (2023, January). Galaxy Morphology Classification Using Bayesian Neural Networks for LSST. In *American astronomical society meeting abstracts* (Vol. 55, p. 105.13).

Dvorkin, C., Mishra-Sharma, S., Nord, B., Villar, V. A., Avestruz, C., Bechtol, K., ... Villaescusa-Navarro, F. (2022, March). Machine Learning and Cosmology. *arXiv e-prints*, arXiv:2203.08056. doi: 10.48550/arXiv.2203.08056

Errington, T. M., Denis, A., Perfito, N., Iorns, E., & Nosek, B. A. (2021, dec). Reproducibility in cancer biology: Challenges for assessing replicability in preclinical cancer biology. *eLife*, *10*, e67995. Retrieved from https://doi.org/10.7554/eLife.67995 doi: 10.7554/eLife.67995

Fukushima, K. (1980). Neocognitron: a self organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biol Cybern*, *36*(4), 193–202.

George, D., & Huerta, E. (2018a). Deep learning for real-time gravitational wave detection and parameter estimation: Results with advanced ligo data. *Physics Letters B*, *778*, 64–70. Retrieved from https://www.sciencedirect.com/science/article/pii/S0370269317310390 doi: https://doi.org/10.1016/j.physletb.2017.12.053

George, D., & Huerta, E. A. (2018b, Feb). Deep neural networks to enable real-time multimessenger astrophysics. *Phys. Rev. D*, *97*, 044039. Retrieved from https://link.aps.org/doi/10.1103/PhysRevD.97.044039 doi: 10.1103/PhysRevD.97.044039

Giorgi, G. M., & Gigliarano, C. (2017). The gini concentration index: A review of the inference literature. *Journal of Economic Surveys*, *31*(4), 1130–1148. doi: https://doi.org/10.1111/joes.12185

Graham, M. J., Djorgovski, S. G., Mahabal, A., Donalek, C., Drake, A., & Longo, G. (2012, August). Data challenges of time domain astronomy. *arXiv e-prints*, arXiv:1208.2480. doi: 10.48550/arXiv.1208.2480

Grojean, C., Paul, A., Qian, Z., & Strümke, I. (2022, March). Interpretable machine learning in Physics. *arXiv e-prints*, arXiv:2203.08021. doi: 10.48550/arXiv.2203.08021

Hartley, P., Bonaldi, A., Braun, R., Aditya, J. N. H. S., Aicardi, S., Alegre, L., ... Zuo, S. (2023, March). SKA Science Data Challenge 2: analysis and results. *arXiv e-prints*, arXiv:2303.07943. doi: 10.48550/arXiv.2303.07943

Hillar, C., & Sommer, F. (2012, October). Comment on the article "Distilling free-form natural laws from experimental data". *arXiv e-prints*, arXiv:1210.7273. doi: 10.48550/arXiv.1210.7273

Hložek, R., Ponder, K. A., Malz, A. I., Dai, M., Narayan, G., Ishida, E. E. O., ... Setzer, C. N. (2020, December). Results of the Photometric LSST Astronomical Time-series Classification Challenge (PLAsTiCC). *arXiv e-prints*, arXiv:2012.12392. doi: 10.48550/arXiv.2012.12392

Ho, J., Jain, A., & Abbeel, P. (2020, June). Denoising Diffusion Probabilistic Models. *arXiv e-prints*, arXiv:2006.11239. doi: 10.48550/arXiv.2006.11239

Hopkins, A. M., Whiting, M. T., Seymour, N., Chow, K. E., Norris, R. P., Bonavera, L., ... van der Horst, A. J. (2015, October). The ASKAP/EMU Source Finding Data Challenge. *Publ. Astron. Soc. Australia*, *32*, e037. doi: 10.1017/pasa.2015.37

Iyer, K. G., Speagle, J. S., Caplar, N., Forbes, J. C., Gawiser, E., Leja, J., & Tacchella, S. (2022, August). Stochastic Modelling of Star Formation Histories III. Constraints from Physically-Motivated Gaussian Processes. *arXiv e-prints*, arXiv:2208.05938. doi: 10.48550/arXiv.2208.05938

Jimenez, M., Alfaro, E. J., Torres Torres, M., & Triguero, I. (2023, February). CzSL: Learning from citizen science, experts and unlabelled data in astronomical image classification. *arXiv e-prints*, arXiv:2302.00366. doi: 10.48550/arXiv.2302.00366

Karniadakis, G. E., Kevrekidis, I. G., Lu, L., Perdikaris, P., Wang, S., & Yang, L. (2021, January). Physics-informed machine learning. *Nature Reviews Physics*, *3*(6), 422–440. doi: 10.1038/s42254-021-00314-5

Karpov, P. I., Huang, C., Sitdikov, I., Fryer, C. L., Woosley, S., & Pilania, G. (2022, November). Physics-informed Machine Learning for Modeling Turbulence in Supernovae. *ApJ*, *940*(1), 26. doi: 10.3847/1538-4357/ac88cc

Kessler, R., Bassett, B., Belov, P., Bhatnagar, V., Campbell, H., Conley, A., ... Varughese, M. (2010, December). Results from the Supernova Photometric Classification Challenge. *Publications of the Astronomical Society of the Pacific*, *122*, 1415. doi: 10.1086/657607

Kessler, R., Becker, A. C., Cinabro, D., Vanderplas, J., Frieman, J. A., Marriner, J., ... York, D. (2009, November). First-Year Sloan Digital Sky Survey-II Supernova Results: Hubble Diagram and Cosmological Parameters. *ApJS*, *185*(1), 32–84. doi: 10.1088/0067-0049/185/1/32

Kessler, R., Narayan, G., Avelino, A., Bachelet, E., Biswas, R., Brown, P. J., ... Transient and Variable Stars Science Collaboration (2019, Sep). Models and Simulations for the Photometric LSST Astronomical Time Series Classification Challenge (PLAsTiCC). *Publications of the Astronomical Society of the Pacific*, *131* (1003), 094501. doi: 10.1088/1538-3873/ab26f1

Kitching, T., Balan, S., Bernstein, G., Bethge, M., Bridle, S., Courbin, F., ... Voigt, L. (2010, September). Gravitational Lensing Accuracy Testing 2010 (GREAT10) Challenge Handbook. *arXiv e-prints*, arXiv:1009.0779. doi: 10.48550/arXiv.1009.0779

Klingner, C. M., Denker, M., Grün, S., Hanke, M., Oeltze-Jafra, S., Ohl, F. W., ... Ritter, P. (2022). Overcoming the reproducibility crisis - results of the first community survey of the german national research data infrastructure for neuroscience. *bioRxiv*. doi: 10.1101/2022.04.07.487439

Kruk, S., & Merín, B. (2023, April). Citizen Science with ESA Science Data - The Hubble Asteroid Hunter project. *Europhysics News*, *54*(2), 28–31. doi: 10.1051/epn/2023206

Lecar, M. (1968). Bulletin astronomique de l'observatoire de paris. In (Vol. 3). Centre National de la Research Scientifique.

Levrier, F., Wilman, R. J., Obreschkow, D., Kloeckner, H. R., Heywood, I. H., & Rawlings, S. (2009, January). Mapping the SKA Simulated Skies with the S3-Tools. In *Wide field astronomy & technology for the square kilometre array* (p. 5). doi: 10.22323/1.132.0005

Lewis, A., Voetberg, M., Nord, B., Jones, C., Hložek, R., Ciprijanovic, A., & Perdue, G. N. (2022, July). DeepBench: A library for simulating benchmark datasets for scientific analysis. In *Machine learning for astrophysics* (p. 32).

Lin, H.-H., Lin, K.-y., Li, C.-T., Tseng, Y.-H., Jiang, H., Wang, J.-H., ... Zhu, H.-M. (2022, September). BURSTT: Bustling Universe Radio Survey Telescope in Taiwan. *PASP*, *134*(1039), 094106. doi: 10.1088/1538-3873/ac8f71

Lintott, C. J., Schawinski, K., Slosar, A., Land, K., Bamford, S., Thomas, D., ... Vandenberg, J. (2008, September). Galaxy Zoo: morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey. *MNRAS*, *389*(3), 1179–1189. doi: 10.1111/j.1365-2966.2008.13689.x

Lokken, M., Gagliano, A., Narayan, G., Hložek, R., Kessler, R., Crenshaw, J. F., ... LSST Dark Energy Science Collaboration (2023, April). The simulated catalogue of optical transients and correlated hosts (SCOTCH). *MNRAS*, *520*(2), 2887–2912. doi: 10.1093/mnras/stad302

Lorimer, D. R., Bailes, M., McLaughlin, M. A., Narkevic, D. J., & Crawford, F. (2007, November). A Bright Millisecond Radio Burst of Extragalactic Origin. *Science*, *318*(5851), 777. doi: 10.1126/science.1147532

LSST Dark Energy Science Collaboration. (2012, November). Large Synoptic Survey Telescope: Dark Energy Science Collaboration. *arXiv e-prints*, arXiv:1211.0310. doi: 10.48550/arXiv.1211.0310

LSST Dark Energy Science Collaboration (LSST DESC), Abolfathi, B., Alonso, D., Armstrong, R., Aubourg, É., Awan, H., ... Zuntz, J. (2021, March). The LSST DESC DC2 Simulated Sky Survey. *ApJS*, *253*(1), 31. doi: 10.3847/1538-4365/abd62c

LSST Science Collaboration, Abell, P. A., Allison, J., Anderson, S. F., Andrew, J. R., Angel, J. R. P., ... Zhan, H. (2009, December). LSST Science Book, Version 2.0. *arXiv e-prints*, arXiv:0912.0201. doi: 10.48550/arXiv.0912.0201

Lucie-Smith, L., Peiris, H. V., & Pontzen, A. (2023, May). Explaining dark matter halo density profiles with neural networks. *arXiv e-prints*, arXiv:2305.03077.

MacKay, D. J. C. (1992, 05). A Practical Bayesian Framework for Backpropagation Networks. *Neural Computation*, *4*(3), 448–472. doi: 10.1162/neco.1992.4.3.448

Malz, A., Hložek, R., Allam, J., Tarek, Bahmanyar, A., Biswas, R., Dai, M., ... Variable Stars Science Collaboration (2018, September). The Photometric LSST Astronomical Time-series Classification Challenge (PLAsTiCC): Selection of a performance metric for classification probabilities balancing diverse science goals. *ArXiv e-prints*, arXiv:1809.11145.

Mandelbaum, R., Rowe, B., Bosch, J., Chang, C., Courbin, F., Gill, M., ... Schrabback, T. (2014, May). The Third Gravitational Lensing Accuracy Testing (GREAT3) Challenge Handbook. *The Astrophysical Journal Supplement Series*, *212*, 5. doi: 10.1088/0067-0049/212/1/5

Megias Homar, G., Meyers, J. E., & Kahn, S. M. (2023, March). Prompt Detection of Fast Optical Bursts with the Vera C. Rubin Observatory. *arXiv e-prints*, arXiv:2303.02525. doi: 10.48550/arXiv.2303.02525

Metcalf, R. B., Meneghetti, M., Avestruz, C., Bellagamba, F., Bom, C. R., Bertin, E., ... Vernardos, G. (2018, February). The Strong Gravitational Lens Finding Challenge. *ArXiv e-prints*, arXiv:1802.03609.

Miller, R. H. (1964, July). Irreversibility in Small Stellar Dynamical Systems. *ApJ*, *140*, 250. doi: 10.1086/147911

Miller, R. H. (1971, January). Experimental studies of the numerical stability of the gravitational n-body problem. *Journal of Computational Physics*, *8*, 449–464. doi: 10.1016/0021-9991(71)90023-4

Modi, C., Lanusse, F., Seljak, U., Spergel, D. N., & Perreault-Levasseur, L. (2021, April). CosmicRIM: Reconstructing Early Universe by Combining Differentiable Simulations with Recurrent Inference Machines. *arXiv e-prints*, arXiv:2104.12864. doi: 10.48550/arXiv.2104.12864

Mohan, D., Scaife, A. M. M., Porter, F., Walmsley, M., & Bowles, M. (2022, April). Quantifying uncertainty in deep learning approaches to radio galaxy classification. *MNRAS*, *511*(3), 3722–3740. doi: 10.1093/mnras/stac223

Narayan, G., & ELAsTiCC Team. (2023, January). The Extended LSST Astronomical Time-series Classification Challenge (ELAsTiCC). In *American astronomical society meeting abstracts* (Vol. 55, p. 117.01).

Nord, B., Connolly, A. J., Kinney, J., Kubica, J., Narayan, G., Peek, J. E. G., ... Tollerud, E. J. (2019, September). Algorithms and Statistical Models for Scientific Discovery in the Petabyte Era. In *Bulletin of the american astronomical society* (Vol. 51, p. 224). doi: 10.48550/arXiv.1911.02479

Ntampaka, M., Ho, M., & Nord, B. (2021, November). Building Trustworthy Machine Learning Models for Astronomy. *arXiv e-prints*, arXiv:2111.14566. doi: 10.48550/arXiv.2111.14566

Ntampaka, M., & Vikhlinin, A. (2022, February). The Importance of Being Interpretable: Toward an Understandable Machine Learning Encoder for Galaxy Cluster Cosmology. *ApJ*, *926*(1), 45. doi: 10.3847/1538-4357/ac423e

Peek, J., & White, R. (2021, June). Search By Image: Citizen Science and Deep Learning for next-generation archives. In *American astronomical society meeting abstracts* (Vol. 53, p. 301.06).

Piras, D., Peiris, H. V., Pontzen, A., Lucie-Smith, L., Guo, N., & Nord, B. (2023, June). A robust estimator of mutual information for deep learning interpretability. *Machine Learning: Science and Technology*, *4*(2), 025006. doi: 10.1088/2632-2153/acc444

Psaros, A. F., Meng, X., Zou, Z., Guo, L., & Karniadakis, G. E. (2023, March). Uncertainty quantification in scientific machine learning: Methods, metrics, and comparisons. *Journal of Computational Physics*, *477*, 111902. doi: 10.1016/j.jcp.2022.111902

Quinlan, G. D., & Tremaine, S. (1992, December). On the reliability of gravitational N-body integrations. *MNRAS*, *259*(3), 505–518. doi: 10.1093/mnras/259.3.505

Raissi, M., Perdikaris, P., & Karniadakis, G. E. (2019, February). Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, *378*, 686–707. doi: 10.1016/j.jcp.2018.10.045

Ramírez-Pérez, C., Sanchez, J., Alonso, D., & Font-Ribera, A. (2022, May). CoLoRe: fast cosmological realisations over large volumes with multiple tracers. *J. Cosmology Astropart. Phys.*, *2022*(5), 002. doi: 10.1088/1475-7516/2022/05/002

Razzano, M., Di Renzo, F., Fidecaro, F., Hemming, G., & Katsanevas, S. (2023, March). GWitchHunters: Machine learning and citizen science to improve the performance of gravitational wave detector. *Nuclear Instruments and Methods in Physics Research A*, *1048*, 167959. doi: 10.1016/j.nima.2022.167959

Reza, M., Zhang, Y., Nord, B., Poh, J., Ciprijanovic, A., & Strigari, L. (2022, July). Estimating Cosmological Constraints from Galaxy Cluster Abundance using Simulation-Based Inference. In *Machine learning for astrophysics* (p. 20). doi: 10.48550/arXiv.2208.00134

Ricker, G. R., Winn, J. N., Vanderspek, R., Latham, D. W., Bakos, G. Á., Bean, J. L., ... Villasenor, J. (2014, Aug). Transiting Exoplanet Survey Satellite (TESS). In *Space telescopes and instrumentation 2014: Optical, infrared, and millimeter wave* (Vol. 9143, p. 914320). doi: 10.1117/12.2063489

Riggi, S., Vitello, F., Becciani, U., Buemi, C., Bufano, F., Calanducci, A., ... Umana, G. (2019, October). Cuc(aesar) source finder: Recent developments and testing. *Publ. Astron. Soc. Australia*, *36*, e037. doi: 10.1017/pasa.2019.29

Rosofsky, S. G., Al Majed, H., & Huerta, E. A. (2022, March). Applications of physics informed neural operators. *arXiv e-prints*, arXiv:2203.12634. doi: 10.48550/arXiv.2203.12634

Rudin, C. (2018, November). Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *arXiv e-prints*, arXiv:1811.10154. doi: 10.48550/arXiv.1811.10154

Sako, M., Bassett, B., Becker, A. C., Brown, P. J., Campbell, H., Wolf, R., ... Zheng, C. (2018, June). The Data Release of the Sloan Digital Sky Survey-II Supernova Survey. *PASP*, *130*(988), 064002. doi: 10.1088/1538-3873/aab4e0

Schmidt, M., & Lipson, H. (2009). Distilling free-form natural laws from experimental data. *Science*, *324*(5923), 81–85. doi: 10.1126/science.1165893

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2016, October). Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. *arXiv e-prints*, arXiv:1610.02391. doi: 10.48550/arXiv.1610.02391

Shukla, K., Xu, M., Trask, N., & Karniadakis, G. E. (2022, May). Scalable algorithms for physics-informed neural and graph networks. *arXiv e-prints*, arXiv:2205.08332. doi: 10.48550/arXiv.2205.08332

Siemiginowska, A., Eadie, G., Czekala, I., Feigelson, E., Ford, E. B., Kashyap, V., ... Young, C. A. (2019, May). The Next Decade of Astroinformatics and Astrostatistics. *BAAS*, *51*(3), 355. doi: 10.48550/arXiv.1903.06796

Square Kilometre Array Cosmology Science Working Group, Bacon, D. J., Battye, R. A., Bull, P., Camera, S., Ferreira, P. G., ... Zuntz, J. (2020, March). Cosmology with Phase 1 of the Square Kilometre Array Red Book 2018: Technical specifications and performance forecasts. *Publ. Astron. Soc. Australia*, *37*, e007. doi: 10.1017/pasa.2019.51

Stein, G., Seljak, U., Böhm, V., Aldering, G., Antilogus, P., Aragon, C., ... Nearby Supernova Factory (2022, August). A Probabilistic Autoencoder for Type Ia Supernova Spectral Time Series. *ApJ*, *935*(1), 5. doi: 10.3847/1538-4357/ac7c08

Sullivan, J. M., Prijon, T., & Seljak, U. (2023, March). Learning to Concentrate: Multi-tracer Forecasts on Local Primordial Non-Gaussianity with Machine-Learned Bias. *arXiv e-prints*, arXiv:2303.08901. doi: 10.48550/arXiv.2303.08901

The CHIME/FRB Collaboration, Andersen, B. C., Bandura, K., Bhardwaj, M., Boyle, P. J., ... Zwaniga, A. (2023, January). CHIME/FRB Discovery of 25 Repeating Fast Radio Burst Sources. *arXiv e-prints*, arXiv:2301.08762. doi: 10.48550/arXiv.2301.08762

Tohuvavohu, A., Kennea, J. A., DeLaunay, J., Palmer, D. M., Cenko, S. B., & Barthelmy, S. (2020, September). Gamma-Ray Urgent Archiver for Novel Opportunities (GUANO): Swift/BAT Event Data Dumps on Demand to Enable Sensitive Subthreshold GRB Searches. *ApJ*, *900*(1), 35. doi: 10.3847/1538-4357/aba94f

Tulio Ribeiro, M., Singh, S., & Guestrin, C. (2016, February). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *arXiv e-prints*, arXiv:1602.04938. doi: 10.48550/arXiv.1602.04938

Vafaei Sadr, A., Vos, E. E., Bassett, B. A., Hosenie, Z., Oozeer, N., & Lochner, M. (2019, April). DEEPSOURCE: point source detection using deep learning. *MNRAS*, *484*(2), 2793–2806. doi: 10.1093/mnras/stz131

van Leeuwen, J., Kooistra, E., Oostrum, L., Connor, L., Hargreaves, J. E., Maan, Y., ... Ziemke, J. (2023, April). The Apertif Radio Transient System (ARTS): Design, commissioning, data release, and detection of the first five fast radio bursts. *A&A*, *672*, A117. doi: 10.1051/0004-6361/202244107

van Roestel, J., Duev, D. A., Mahabal, A. A., Coughlin, M. W., Mróz, P., Burdge, K., ... Kulkarni, S. R. (2021, June). The ZTF Source Classification Project. I. Methods and Infrastructure. *AJ*, *161*(6), 267. doi: 10.3847/1538-3881/abe853

Vanderlinde, K., Liu, A., Gaensler, B., Bond, D., Hinshaw, G., Ng, C., ... Kaspi, V. (2019, October). The Canadian Hydrogen Observatory and Radio-transient Detector (CHORD). In *Canadian long range plan for astronomy and astrophysics white papers* (Vol. 2020, p. 28). doi: 10.5281/zenodo.3765414

Walmsley, M., Smith, L., Lintott, C., Gal, Y., Bamford, S., Dickinson, H., ... Wright, D. (2019, 10). Galaxy Zoo: probabilistic morphology through Bayesian CNNs and active learning. *Monthly Notices of the Royal Astronomical Society*, *491*(2), 1554–1574. Retrieved from https://doi.org/10.1093/mnras/stz2816 doi: 10.1093/mnras/stz2816

Wang, Y.-T., Liu, H.-Y., & Piao, Y.-S. (2023, February). Self-supervised learning for gravitational wave signal identification. *arXiv e-prints*, arXiv:2302.00295. doi: 10.48550/arXiv.2302.00295

Yong, S. Y., Hobbs, G., Huynh, M. T., Rolland, V., Petersson, L., Norris, R. P., ... Zic, A. (2022, November). SPARKESX: Single-dish PARKES data sets for finding the uneXpected - a data challenge. *MNRAS*, *516*(4), 5832–5848. doi: 10.1093/mnras/stac2558

Yu, W., Richards, G., Buat, V., Brandt, W. N., Banerji, M., Ni, Q., ... Yang, J. (2022, July). *Lsstc agn data challenge 2021.* Zenodo. Retrieved from https://doi.org/10.5281/zenodo.6878414 doi: 10.5281/zenodo.6878414

Zhang, C., Wang, C., Hobbs, G., Russell, C. J., Li, D., Zhang, S. B., ... Ren, Z. Y. (2020, October). Applying saliency-map analysis in searches for pulsars and fast radio bursts. *A&A*, *642*, A26. doi: 10.1051/0004-6361/201937234

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (2015, December). Learning Deep Features for Discriminative Localization. *arXiv e-prints*, arXiv:1512.04150. doi: 10.48550/arXiv.1512.04150

Zhu-Ge, J.-M., Luo, J.-W., & Zhang, B. (2023, February). Machine learning classification of CHIME fast radio bursts - II. Unsupervised methods. *MNRAS*, *519*(2), 1823–1836. doi: 10.1093/mnras/stac3599

Zuntz, J., Lanusse, F., Malz, A. I., Wright, A. H., Slosar, A., Abolfathi, B., ... LSST Dark Energy Science Collaboration (2021, October). The LSST-DESC 3x2pt Tomography Optimization Challenge. *The Open Journal of Astrophysics*, *4*(1), 13. doi: 10.21105/astro.2108.13418