Mitigating AI Bias with Astronomical User-inspired Science

PI: Christopher J. Miller

1 Motivation

Bias in Artificial Intelligence (AI) refers to the systematic and repeatable errors in the outcomes produced by a predictive model due to the assumptions made in the machine learning process. Like any other domain, astronomical AI requires input data that can reflect biased human decisions. AI bias in astronomical research is primarily a concern for scientific endeavors. However, the nature of openness in astronomical data and software provides an opportunity to transfer research and development in Trustworthy AI to domains like employment, housing and healthcare, where discrimination and bias are not academic issues, but human ones. AI systems also often have a low level of (human) interpretability, which can lead to difficulties in identifying and mitigating bias compared to simpler analyses (e.g. linear systems, logistic regressions, etc). These are amplified when multiple survey datasets are combined, which is common in astronomy.

The goal of this proposal is to implement AI systems on astronomical data in order solve an interesting astrophysical question which requires the identification, quantification, and mitigation of bias. In doing so, we hope to pave the way for general approaches to trustworthy AI that can be used in other domains.

Much of modern astronomical research utilizes advanced statistics and machine learning. However, the future of astronomy requires artificial intelligence. Not only are the data voluminous (petabytes), but highly heterogeneous due to instrumental modes (e.g. imaging versus spectroscopy), observing conditions (e.g., mountaintop versus orbit), experimental designs/facilities (i.e., multi-partner international teams versus individual PIs), etc. At the same time, our universe is filled with an extraordinary range of astronomical objects with an equally wide range of absorption/emission mechanisms, many of which are also evolving in complex ways in a universe with a space-time that is itself changing.

These complex conditions create fundamental limitations to the physical insights humans alone can learn from the data. There are also realistic limits to the experiments we deploy and the data we collect. Recently and partly due to these complexities and limits, scientific advances in the physical sciences like astronomy, astrophysics and cosmology have become mostly incremental [38]. AI provides a promising way forward in terms of making larger leaps in our astrophysical knowledge and finding deeper connections between the objects and mechanisms which govern their emission.

AI aims to minimize the non-essential tasks astronomers do in order to to maximize time spent on things the machines cannot yet do. With this realization in mind, the application of AI techniques in astronomy is now growing at an exponential rate [47]. Inherent in the path towards true AI is need for astronomers to trust the outcome of the AI model. However, human astronomers are not perfect and imperfections in the models may stem from imperfections in the training data. As we highlight in this proposal, training set biases are hard to identify and can appear even after you think you have corrected for it.

We will study AI bias using an interesting and novel scientific use case which can only be solved with an AI system. Specifically, we aim to distinguish between competing models of galaxy evolution by mapping local/global morphological features to local/global spectral features. We will evolve a set of statistical tools to discover and quantify bias in the human-trained morphological features. These tools will then be embedded into the AI system in ways that will enable automated bias mitigation. The resultant models will be deconstructed, so that human astronomers can learn how and why they work and/or they fail. Finally, we will conduct observational follow-up on a small subset of galaxy features predicted by the AI system to assess trustworthiness.

2 Background Material

The astronomical sciences have long been a driver for advancing statistical theory and applications. Motivated by the precision measurements of stars and planets of Tycho Brahe, Kepler collapsed Copernicus's perfectly circular orbits into ellipses. This in turn gave rise to Newton's theory of gravity and the birth of mechanics. Precision astronomical measurements of Mercury's precessing orbit were explained by Einstein's theory of General Relativity. "[The statistical] requirements of astronomy (and geodesy) gave rise to the classical theory of errors and to the method of least squares" [46]. Laplace used his theory of probability to reason why hyperbolic orbits of comets are rare. In an early 19th century tour de force of data science, Gauss applied his theory of least squared errors to predict the sky location of a strange new moving object, later identified by astronomers von Zach and Olbers as the asteroid Ceres [12].

In 1994, Lahav [25] reminded the astronomy community that the fundamentals of artificial neural networks are the non-linear extensions of many traditional statistical machine learning (ML) techniques. As the data grew in quality and quantity, astronomers began to apply techniques like principal component analysis (for unsupervised learning), and Bayesian classifications. Astronomers were early adopters of KD-trees and Monte-Carlo Markov Chains to address speed issues [8, 24]. Nowadays, Bayseian and non-parametric statistical techniques dominate the astronomical and cosmological literature [47].

However, it was another 20 years for machine learning in astronomy to become common [50], as shown in Figure 1 of [47]. Unlike the historical and strong connection between astronomy and statistics, ML in astronomy is catching-up. This forms the basis for this proposal. Until recently, astronomers have typically used highly interpretable ML (IML), where the degree of human explainability is high (e.g., linear regression or decision trees). Our proposed methodologies will emphasize the use of low interpretability ML (LIML) which have experienced an acceleration of use in astronomy in the last 5 years.

2.1 AI: Artificial Intelligence

Driven by the success of the natural language tool ChatGPT, the term "AI" is ubiquitous in the press. ChatGPT is a "Generative Pretrained Transformer" model that is fine-tuned for human conversational purposes. It is a simulation of human conversation based on training. For this proposal, we define AI specifically as a system of data and processes which can achieve a result that would be difficult for humans to obtain through piece-wise analyses. It is more than Bayesian inference, classification or density estimation or LLMs.

True **Artificial Intelligence** (AI) allows computers and automated systems to execute functionality that traditionally necessitated human intelligence and decision-making capabilities. AI research focuses on comprehending the principles behind thinking and intelligent actions, and how these can be integrated into machines. Future (and true) AI systems will automatically perceive, learn, reason, communicate, and act. In their actions, they will show flexibility, resourcefulness, creativity, real-time responsiveness. Over time, true AI systems will use reflection and reasoning to demonstrate competence in their analyses, build trust with their human counterparts, and do so in way that uses natural language.

This proposal is designed to develop a framework for Trustworthy AI that can be applied to any discipline in the hopes of transferring AI knowledge from astronomy to other domains. We will do so using a user-inspired astronomical scientific challenge: can we learn an interpretable AI model which predicts spatio-spectral inferred properties inside galaxies (like stellar age and metallicity) from just the pixel imaging data? If so, we can create datasets that are an order of magnitude larger than what is currently available, and the machines can help us learn how the internal environment of a galaxy processes its stars.

AI is filled with acronyms. For this proposal, we will focus on terms required for our use-inspired research project. We succinctly define these with enough detail to understand how they will be used in the

proposed research. The characters in parentheses refer to their abbreviations in the rest of this proposal.

*** AI system:** A collection of applications that perform complex tasks that would have otherwise required significant human input. A "true" AI system is fully autonomous. For this proposal, an AI system is a complex scientific analysis which removes much, but all of the human intervention.

*** Loss Function (LF):** Loss functions are estimates of how good your model is in terms of predicting the expected outcome. For instance, Support Vector Machines (SVM) often employ a "Hinge" LF which penalizes the wrong predictions as well as correct predictions with low confidence. Other typical LFs include the mean squared error (ℓ^2), mean-absolute deviations, etc. AI-systems minimize the LF during training.

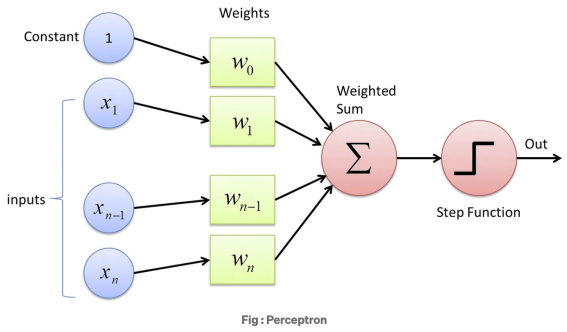

Figure 1: A single perceptron, where the data are weighted and summed and then turned into a number via an activation function, here the Step Function.

*** Perceptron:** The fundamental units of modern neural networks and used to separate two classes by a linear boundary. They have similarities to a probabilistic classification like Naive Bayes. The input data are multiplied by weights, summed, and turned into a number between zero and one. This latter step is the activation function. The weights are determined through minimizing the loss function.

*** Convolutional Neural Network (CNN):** Perceptrons are grouped into hidden layers where convolutions are performed. The connections between each perceptron correspond to the weight parameters of the network. The number of hidden layers, perceptrons, and connections, varies with the problem at hand. The activation functions adds non-linearity to the network. The goal of the CNN is to learn the data-specific kernels and the weights by minimizing the LF using "back propagation", using the gradient of the LF.

*** Residual Neural Network (ResNET):** Adds an additional connection that skips one or more layers of the network in the forward feed. If x/$F(x)$ are the input/output through a layer, the skipped layer has an output $H(x) = F(x) + x$ with a residual function $F(x) = H(x) - x$. This prevents "identify mapping" convergence, which can stop the learning process. The use of a residual also helps resolve issues where the gradients become too small.

*** Domain Adaption (DA):** DA method involves tuning models trained on one dataset to perform well on similar but independent datasets (e.g., a model trained on SDSS imaging applied to DES imaging). DA includes fine-tuning of the layers and weights using discrepancy metrics, parallel modeling to encourage domain invariant features (adversarial), and multi-source semi-supervised training.

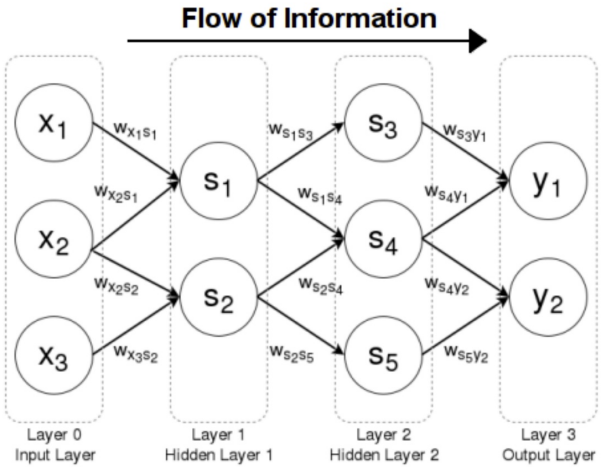

Figure 2: A CNN. The data x_i, the perceptrons, s_j, the weights, $W_{x_i s_j}$ and the output y_k.

*** Class Activation Mapping (CAM):** After running a CNN, we generate images which show the area of activation from a particular convolution layer. Gradient-weighted CAM (GradCAM) uses the gradient information (from the loss function) into the final layer. These maps help establish trust in model predictions by enabling increased interpretability.

We emphasize that in this proposal, the astronomer stays in-the-loop and plays two very specific roles: teaching and interpreting. Each of these is necessitated by the fact that we are trying to conduct science. Unlike some AI endeavors, *AI in the physical sciences requires foundational principals*. In other words, our goal is not to let AI re-discover or re-write the math, physics, and astrophysics found in our textbooks. Our goal is to use that textbook knowledge as the starting point for the next deeper level of learning. Therefore by choice, we will always be training the AI system in physically meaningful ways. Similarly, while a resultant model may have a low level of interpretability (LIML), the astronomer will still need to interpret it. A model that works to predict a complex process is great, unless that model lacks any connection to physical reality.

2.2 Gold Standard Labeling Bias in Supervised Learning

Consider a human labelled dataset $\mathscr{D} = \{(\mathbf{x}_i, \tilde{y}_i)\}_{i=1}^{N}$ composed of pairs $(\mathbf{x}_i, \tilde{y}_i)$ of features $\mathbf{x}_i$ and human generated labels $\tilde{y}_i$. We assume these pairs to be sampled independently from a data distribution $p(\mathbf{x}, \tilde{y})$ defined over $\mathscr{X} \times \mathscr{Y}$. Given these initial (or source) data, we want to determine a learned function $f_{\mathbf{w}} : \mathscr{X} \to \mathscr{Y}$ that maps the input features to the labels, with parameters $\mathbf{w}$ to be fitted.

In real life, it may be very hard to obtain the actual real label y_i, which it is usually called the gold standard or ground truth (GT). An estimate of this gold standard may be obtained by using a set of labels created by one or many human annotators. In other words, we assume the existence of an unknown *ground truth* label $y_i \in \mathscr{Y}$ for each original label $\tilde{y}_i$. However, real data is never perfect, and labels given by the annotators can be systematically biased due to poor quality of the observed data they are labeling. For example, when labeling data by watching images, or video, human labels may be biased because of image resolution or video frames per seconds. This bias cannot be overcome by "re-training" the annotators, or by incorporating measurement error, which can in principal be very small. This is a significant challenge for fields which require training sets for which an absolute ground truth classification is impossible (e.g., astronomy).

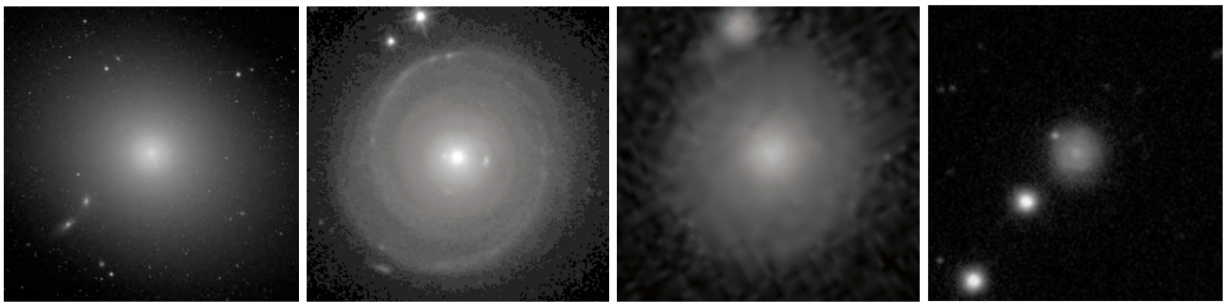

Figure 3: This is an example of ground truth labeling bias. Panel (a) is an elliptical galaxy, which we consider "featureless" for training. Panels (b-d) are the same face on spiral galaxy. Panel (b) is viewed from space via the Hubble Space Telescope (HST) and the spiral features are evident. Panel (c) is viewed from the ground via the Sloan Digital Sky Survey (SDSS). Panel (d) is the same as (c), except over a wider field-of-view. For the SDSS data, the lower quality results in a featureless morphology. The SDSS image was classified as a featureless elliptical galaxy as part of the Galaxy Zoo project by $> 95\%$ of the annotators. In other words, there is essentially no labelling error, even though the higher quality HST image shows the label to be incorrect.

Labeling bias is becoming increasingly important in the AI era, where it becomes difficult to create ground-truth classifications due to the shear size and complexity of the data. In astronomy, researchers have

gone from needing ground truth catalogs based on 1000s of galaxies to millions in less than 10 years. One solution has been to evolve from using professional astronomers to "citizen scientists" via crowd sourcing approaches. The ground truth catalogs then include only those classifications where the majority of the crowd on a single classification. Yet even when the entire crowd agrees, it can still be wrong. An example of this is shown in **Figure 3**. Of course, when one wants to apply a learning model using biased ground truth data, the biases carry through into the learning and classification process.

We emphasize that the labeling bias we are focusing on is not the same as measurement bias. For instance, [5] have reviewed statistical methods for detecting measurement bias in psychological and educational tests. They define this bias as a systematic inaccuracy of measurement. In our case, the measurement by the annotator is "correct", as the human eye cannot identify the spiral features due to limitations of the data. Only higher quality data would change the label.

The type of bias we address in this proposal was studied by the Galaxy Zoo (GZ) team. Bamford et al. [2] corrected the vote fractions obtained from the crowd sourcing system by assuming that the morphologies do not evolve with redshift (distance) within bins of fixed galaxy physical size and luminosity. Later, [53] adapted this technique to Galaxy Zoo 2 (GZ2), taking in consideration that it uses a decision tree rather than a single question (as in GZ), meaning that all tasks but the first one depend on responses to tasks higher in the decision tree. After that, [18] presented an improved method that addresses the questions on GZ2 decision tree with multiple responses (such as number of spiral arms), showing that the method from [53] does not always adjust the vote fractions correctly. Their method aims to make the vote distributions consistent at different redshifts rather than the mean vote fractions values as in [2] and [53]. However as we shall show in the next section, significant labeling bias remains in these calibrated morphologies.

2.2.1 Quantifying Ground Truth Labeling Bias

So how do we identify and quantify labeling bias? [2] quantified the bias in GZ labels by measuring the redshift distribution in bins of luminosity and physical size. These are *intrinsic parameters* of the galaxies and at the same time there is an *intrinsic label distribution*. As in [2], we define properties $\beta = \{\beta_1, \cdots, \beta_{n_\beta}\}$ on which we define $N_{\mathscr{B}}$ multi-dimensional bins $\mathscr{B}_q$. Given a set of K labels (e.g. $K = 2$ for spirals and ellipticals), in each bin $\mathscr{B}_q$, we calculate the *intrinsic class fraction* of objects with each label as $f_{k,q}$. For typical galaxy morphology data sets, we define $\beta_i = (R_i, M_i, z_i)$, where R_i is the physical radius (in kpc), M_i is the absolute magnitude, and z_i is the redshift for object i. In other words, given a fixed bin q in galaxy physical size, luminosity, and redshift, $f_{k=\text{spiral},q}$ defines the *intrinsic fraction* of spirals compared to the total number of galaxies in bin q.

Consider the set of *observed* properties of the objects. The fractions in each class of an unbiased labeled data set should not vary systematically as a function of an *observable parameter*, like image resolution. As in Figure 3, the high resolution galaxy will not be mislabeled as featureless. The low resolution version will. The same can be said for the angular size, which is exemplified by comparing a classification based on Figure 3 (c) versus (d). As in the observable parameters, we define the set of observed properties $\alpha = \{\alpha_j\}_{j=1}^{n_\alpha}$.

We can then build tests which look to identify dependencies in the labels with observable parameters. As in the binning of the observable parameters, we define the set of observed properties $\alpha = \{\alpha_j\}_{j=1}^{n_\alpha}$ and create single dimensional bins on each observed property. We do this within each of the intrinsic $\mathscr{B}_q$ bins $\mathscr{A}_{j,q}$. For typical galaxy morphological data sets, we define $\alpha_i = (r_i/\text{PSF}_i)$ where r_i is the angular size and PSF_i is the estimated size of the point spread function at the galaxy location in the same units as its angular size.

We define a set of ranges (l bins) for the observed property and calculate the *observed class fraction* $f_{j,l,q,k}$ as in equation 1, where $N_{\mathscr{A}_{j,q}}$ is the total number of objects with the observed property α_j in bin

$\mathscr{A}_{j,q}$. $\delta_{\hat{y}_i,k}$ is the Kronecker delta given an estimate of each galaxy i's classification $\hat{y}_i$ for class k. The right-hand sides sum over all galaxies which are simultaneously in the observed single property bin $\mathscr{A}_{j,q}$ and the intrinsic property multi-dimensional bin $\mathscr{B}_q$.

Equations 1-3: Defining the bias.

$$f_{l,j,q,k} = \frac{1}{N_{\mathscr{A}_{j,q}}} \sum_{\substack{i|\alpha_{i,j}\in\mathscr{A}_{j,q} \\ \beta_i\in\mathscr{B}_q}} \delta_{\hat{y}_i,k}, \quad (1)$$

$$\sigma^2_{j,k,q} = \frac{1}{N_{\mathscr{A}_{j,q}}} \sum_{l=1}^{N_{\mathscr{A}_{j,q}}} (f_{j,l,q,k} - f_{k,q})^2. \quad (2)$$

$$L = \sqrt{\frac{1}{KN_{\mathscr{B}}n_\alpha} \sum_{j,k,q} {\sigma_{j,k,q}}^2}, \quad (3)$$

We also calculate the intrinsic class fraction $f_{k,q}$ and the ℓ^2- Euclidean difference to the observed class fraction $f_{j,l,q,k}$ over all l. Equation 2 should be ~ 0 when there is no difference between the intrinsic and observed class fractions, i.e., *when the classifications are unbiased with respect to an observable.* We estimate $\hat{f}_{k,q}$ by using $f_{j,l,k,q}$ for the bin l in property j which is likely to have the least bias. For example, for $\alpha_j = r/\sigma_{\rm PSF}$, then we calculate $\hat{f}_{k,q}$ for the bin which includes the largest values of $r/\sigma_{\rm PSF}$, since it should contain the least biased classifications. It is also possible that $f_{k,q}$ could be predicted from theory. We extend this to all classes and intrinsic and observed properties and define the *classification bias* in equation 3, where K is the number of classes.

The above formalism for bias can be generalized. For instance, suppose one was studying the population trends of rabbits and hares in different environments (desert, forest, tundra, etc). The intrinsic properties could include fur color or length and genetics inform us on the underlying true $f_{k,q}$. The observed property would be the size of the animal in a digital image used by a human annotator to classify it as either a hare or a rabbit.

2.2.2 Identifying Ground Truth Labeling Bias

Ground truth labeling bias in galaxy morphologies was first examined by our team (Cabrera et al. 2018). For different sets of labels, Cabrera et al. calculated equation 3. To ensure a fair comparison between labeled datasets, the number of galaxies was kept fixed, as defined by the smallest dataset. We examined the expert labels from [14] (F07), as well as [35] (NA10), the crowd-sourced Galaxy Zoo Biased (GZB) sample, the Galaxy Zoo DeBiased (GZDB) sample [2], and the labels from a machine learning algorithm used by [21] (HC11). HC11 used the F07 classifications for training a Support Vector Machine to define the galaxy classes.

In **Figure 4**, we show bias which can be fairly compared between datasets, but not between panels. The top row shows the bias for distance (or redshift) as the observable parameter and the physical radius R and absolute magnitude M as the intrinsic parameters. This is effectively how Galaxy Zoo quantified and debiased the crowd classifications [2]. The second row is when $\alpha = r/PSF$ is the observed parameter (as described above).

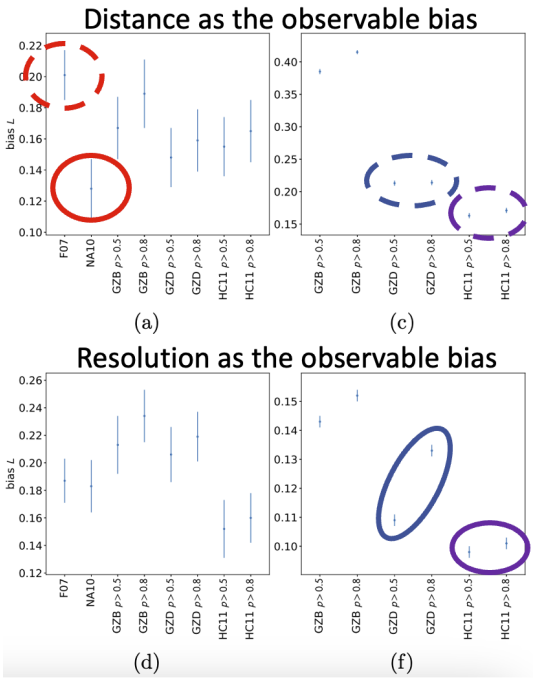

Figure 4: The red circles are expert labels, whose bias varies widely. Purple ellipses show the machine learned labels, usually least biased, even when trained on F07. The blue ellipses show the Galaxy Zoo redshift de-biased labels from [2].

We summarize **Figure 4**:

1. In the top left panel, we see that the F09 expert annotations show much more redshift bias than the NA10 expert annotations. However, the experts are equal when resolution is the biasing parameter.
2. In most of the comparisons, the HC11 machine learned labels are the least biased.
3. There is often an *increase* in the bias when the classifications are more certain ($p > 0.8$). This is evident in the lower right panel (i.e., when resolution is the biasing parameter).

Items (2) and (3) above were the most surprising results. Why do the HC11 labels, trained on the biased F07 labels, have the lowest bias? Why are high confidence labels more biased than low confidence ones?

HC11 used galaxy light concentrations, ellipticities, and colors to train their SVM. Perhaps the machine is using them in a way which is non-trivial to understand (i.e., it has low interpretability). In terms of confidence, we suggest that featureless galaxies are expected to be more confidently classified, even when wrong as a result of low image resolution. The obvious consequence is an increased bias. The most worrisome result is the bottom right panel, which shows that the de-biasing procedure applied in [2] made the GZD (p>0.8) labels much worse for high-confidence labels when resolution is the biasing parameter. Yet the high-confidence labels are the most likely to be used for scientific analysis.

2.2.3 Mitigating Labeling Bias in AI Systems

In order to mitigate the bias, we have to incorporate it into the loss function. We revise the nominal ℓ^2 LF using the likelihood of the data $\mathscr{D}$ given the fitted parameters $\mathbf{w}$ in equations (4-7). The first term of equation (6) is the likelihood function and models the bias in the data by assuming that the biased labels $\tilde{y}_i$ depend only on the ground truth labels y_i and the observed parameters α_i of each ith object. In practice, we model $p(\tilde{y}_i = \text{featureless}|y_i = \text{spiral}, \alpha_i)$ as $e^{-\alpha_i^2/(2\theta^2)}$, where θ is a parameter fitted at an earlier stage to minimize the bias in equation 3. If α is defined as the galaxy's size over the PSF, then $\lim_{\alpha\to 0} p(\tilde{y} = \text{featureless}|y = \text{spiral}, \alpha) = 1$ and $\lim_{\alpha\to\infty} p(\tilde{y} = \text{smooth}|y = \text{disk}, \alpha) = 0$.

Low resolution spiral galaxies will always be mislabelled as featureless galaxies by the annotators and high resolution spiral galaxies are never mislabelled. On the other hand, we assume that truly featureless galaxies are never mis-identified as spirals. Thus we have $p(\tilde{y} = \text{spiral}|y = \text{featureless}, \alpha) = 0$ and $p(\tilde{y} = \text{featureless}|y = \text{featureless}, \alpha) = 1$.

Equations 4-7: The likelihood used in the LF

$$p(\mathscr{D}|\mathbf{w}, \{\alpha_i\}_{i=1}^N) = \prod_{i=1}^N p(\tilde{y}_i|\mathbf{x}_i, \mathbf{w}, \alpha_i), \quad (4)$$

$$= \prod_{i=1}^N \sum_{y_i} p(\tilde{y}_i, y_i|\mathbf{x}_i, \mathbf{w}, \alpha_i), \quad (5)$$

$$= \prod_{i=1}^N \sum_{y_i} p(\tilde{y}_i|y_i, \alpha_i) p(y_i|\mathbf{x}_i, \mathbf{w}), \quad (6)$$

$$\mathscr{L} = -\log p(\mathscr{D}|\mathbf{w}, \{\alpha_i\}_{i=1}^N). \quad (7)$$

The second term of equation (6) corresponds to a prior probabilistic classification model used to infer the ground truth labels from the observed features $\mathbf{x}_i$ (e.g. 2D imaging data) and the model parameters $\mathbf{w}$. Since our initial work is simply a binary classification, we can write $p(y_i|\mathbf{x}_i, \mathbf{w})$ as $(a_i p_i + b_i(1-p_i))$, where $a_i = p_{0|1}^{(1-\tilde{y}_i)}(1-p_{0|1})^{\tilde{y}_i}$ and $b_i = p_{1|0}^{\tilde{y}_i}(1-p_{1|0})^{\tilde{y}_i}$. We require that the biased labels $\tilde{y}$ do not depend directly on the features $\mathbf{x}$, while the true labels do. When training a de-biased model, we add the negative of the log-likelihood of equation (6) into the normal MSE loss function.

In our most recent work, we implemented de-biasing into two AI-systems to learn labels [33]. The first technique uses 1D catalog data for each galaxy to enable a classification via logistic regression. While these catalog data were measured from the 2D imaging data, the pixels themselves are not directly used in the classification. The parameters used to make classifications in via the 1D catalog data are the Sérsic index [44],

the ellipticity, and the half-light radius as classification features. To train this model we follow [5] and maximize the log-likelihood from equation 7 using the Expectation-Maximization algorithm. The second classification model employs Deep Learning techniques via CNN and requires only the 2D galaxy imaging data.

Dataset / Method	Bias CV18
a) GZ2B (Willett et al. 2013)	0.3696 ± 0.0095
b) GZ2D (Hart et al. 2016)	0.3106 ± 0.0108
c) ResNet50 over GZ2B	0.3781 ± 0.0091
d) ResNet50 over GZ2D	0.3258 ± 0.0107
e) DCV14 over GZ2B	0.2994 ± 0.0102
f) DDB (ours) over GZ2B	**0.2866 ± 0.0112**

Table 1: (a,b) are biases L inherent to the GZ classifications. GZD is debiased statistically to remove redshift trends of the classes. (c,d) are classes based on training our AI system on (a,b) classes. (e,f) and de-biased classes after incorporating the likelihood in equation 3. (e) uses catalog data for training and prediction. (f) uses pixel data.

We used a ResNet50 [19] CNN and added an additional dense layer with 1024 neurons and ReLU activation. As input we used JPEG images and we trained our model by minimizing the negative of the log-likelihood (equation (7)), where features $\mathbf{x}$ corresponding to the 2D galaxy imaging data and parameters $\mathbf{w}$ to the weights of the ResNet50 model. In this exercise, we have updated the Galaxy Zoo classes (GZ2B) to be from [53]. These classes were de-biased (GZ2D) using a revised scheme which is still based on the redshift distribution [18]

In the above **Table 1**, we show the relative class biases using the same galaxies for all calculations (but different labels). We see that the revised statistical de-biasing for GZ2 scheme works better than the original (i.e, GZ2D $<$ GZ2B). Interestingly, we see that our ResNET50 CNN recovers the human classifications quite well and also the bias. In other words, using just the pixels, the machine is as biased as the human. However, when we include information about what could be biased through the loss in equation 7, the machine is able to do a better job (i.e., logistic regression (DVV14) and the CNN (DDB) in rows e,f).

The next step is to assess the trustworthiness in the AI model. We do so by using domain-specific knowledge: the Sérsic index [44]. From decades of research, we know that disks (ellipticals) have flat (steep) radial projected light profiles. We also know that the human labels for spiral disk galaxies will not be biased. As previously noted, this is simply because if spiral features are visible, the only class the human will choose is disk. Therefore, the Sérsic index for spiral galaxies is not contaminated by elliptical galaxies (although it is not complete, since misclassified ellipticals are not included).

In **Figure 5**, we show the frequency of Sérsic indices in the GZ2B sample which had their labels changed from featureless (ellipitcal) to disk (spiral). The expectation is that most will have a low-power index. Our de-biased AI model follows the same distribution as the GZ2B data, which as noted is uncontaminated by ellipticals. The GZ2D de-biased labels show a much higher frequency of ellipitcal-like steep profiles in these disk galaxies. As a separate test. we have identified numerous new examples like Figure 3, where the HST data clearly show spiral features in mislabeled ellipticals.

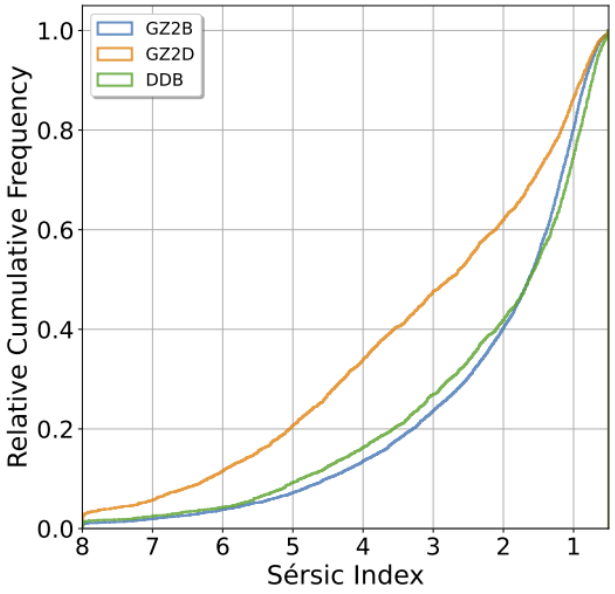

Figure 5: The power-law light profiles for galaxies with switched labels from featureless to spiral-like. Since spirals cannot be mis-classified as featureless, the de-biased frequency should match the original labels (GZ2DB). While our AI system works, the GZ2D de-biasing scheme is identifying galaxies with steep light profiles as disk-like spirals.

We dig into the neural net using GradCAM activation maps. In **Figure 6**, we show where the CNN focuses during the classification training. Recall that DDB is the ResNet50 with the de-biasing loss function. The others use a standard ℓ^2 loss. The top row is the average of the activation maps for spirals, the lower for featureless ellipticals.

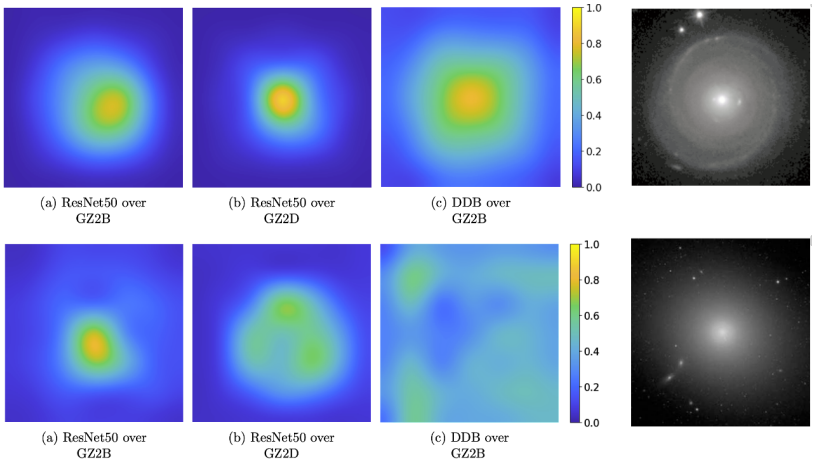

Figure 6: Activation maps which show where the CNN was focused when training on the classes. The top row is for spiral disk-like galaxies. The bottom for featureless ellipticals.

For spirals, the CNN always focuses on the core. This is surprising, given that the human eye is drawn to the spiral features in the outskirts. Maybe CNN is placing more emphasis on the radial light profile than the features themselves (e.g., as in Figure 5)? With the revised loss function, the machine shifts some of its attention to the outskirts.

For featureless ellipticals, there is very little difference between spirals and ellipticals for the sample with known biases (panel a). As we switch to training the ResNET50 on a sample with lower bias (panel b), we see the activation moves away from the core and into the outer regions. We then see a maximal difference in the biased versus un-biased data for the ResNet50 with the revised loss (panel c) where the core is ignored. For the galaxy in Figure 3c, the AI system weighed the core light over the featureless outskirts when deciding it is a disk-like spiral. On its own, it might have noticed a shallow light profile.

3 User-inspired Science

3.1 Summary of the AI Preliminary Work

At this point, our proposal has provided background on modern AI tools and terms used in morphological classifications in Section 2.1. We then followed with a definition and explanation of gold standard labeling bias in Section 2.2. From our prior work in [5], we defined a way to quantify this bias based on the observed and intrinsic properties of the data (Section 2.2.1). We then showed real examples where this bias exists, including in expertly labeled data as well as data that had been "de-biased" based on galaxy redshift distributions (Figure 4 in Section 2.2.2). We also showed that machine learning applications like SVM can somehow mitigate this bias, even when the supervised learning is trained on biased labels. In Section 2.2.3, we integrated bias mitigation CNN classification systems via a revised loss function. We showed that our AI system produces labels with lower bias then the original labels, the Galazy Zoo de-biased labels, and our logistic regression machine learned labels. We assessed truthfulness in Figure 5 by evaluating the light profiles for those galaxies which changed their labels from featureless (elliptical) to disk (spiral-like). Finally, we used activation maps to understand how the machines de-bias. We learned that for disk galaxies, the CNN focuses on the galaxy core with some attention towards the outskirts. For featureless morphologies, the CNN ignores the core and spends its reources in the galaxy outskirts.

We now define a science case which requires bias mitigation within CNN classifications. The **expected significance** is an order of magnitude increase in the amount of data available for the type of research described *without deploying a new survey*. We may even expand this science to include redshift evolution, which is currently impossible. Of course, to achieve our goals we will need to trust the AI classifications.

User-inspired Science Case:

There are competing models for galaxy formation. In dissipative collapse (top-down) models [6, 10, 26], lower gas densities from shallower potentials lead to less efficient star formation and self-enrichment. Measured from the high density core to the lower density outskirts, one should detect negative metallicity gradients and slightly positive stellar age gradients. If we instead consider merger driven evolution in a ΛCDM universe, present-day galaxies formed their central parts first, through either cold stream accretion, violent disk instabilities [9] or galaxy merging. Afterwords, there was a gradual buildup of a disk forming from the inside out [34]. In this model we expect both metallicities and stellar ages to show negative gradients [49].

Using intergal field unit spectroscopy (IFU), Delgado et al. [16] found strong negative age and weakly negative metallicity gradients out to $> 2-3\times$ the half-light radius with a steepness that depends on the morphological classification. This suggests inside out formation with the twist that stars inside galaxies can be "morphologically quenched". Chen et al. [7] found negative age and metallicity gradients, but with a strong dependence on a galaxies stellar mass and little dependence on morphology. Due to signal-to-noise constraints, they only probed out to ~ 1.5 HLR. Parikh et al. [37] found clear differences in the age and metallicity gradients between early type ellipticals (ETG) and late type spirals (LTG). The LTG show negative gradients while the ETG show flat profiles. There is clearly scientific debate in an active area of research. Have we really quantified these gradients? Do they depend on the galaxy's stellar mass or on its morphology? How do we make progress in light of the immense expense of astronomical survey observations which require spatio-spectral data?

Perhaps the radial gradient is the wrong measure? Detailed morphology and galaxy evolution are entwined. Bars can move gas inwards [41] which could driving and/or shutting down star formation [22, 45]. Galaxy bulges may quench star formation in regions well outside the core [4, 11, 30] and inside-out quenching [48]. Tidal features from past mergers most likely also play some role [20, 23]. Disentangling the complex interplay between morphology and galaxy evolution requires both the measurement of detailed morphology and also spatially resolved galaxy properties from spectroscopy. As emphasized in Fraser-McKelvie et al. [13], azimuthally averaging a galaxy's star formation history can result in a loss important structural information needed to understand its true formation history. The efforts described above use a few hundred to a few thousand galaxies with spectral-spatio data. When split by morphological type and mass, the binned samples are even smaller. However, getting deeper or more IFU data is an expensive proposition.

Our solution to these challenges is to train an AI system to predict spatio-spectral properties of the stellar light inside galaxies from their imaging data alone. Can we build an AI system that maps resolved stellar properties inside galaxies, (like mass, age, or metallicity), with their internal shape properties (like bars, arms, and bulges)? If so, we will be able to provide the community a way to make transformative progress in galaxy formation studies through an order of magnitude increase in the sample sizes used for galaxy evolution studies. Researchers can use the model to infer stellar mass, metallicity, and age on pixels which fall below S/N thresholds for spectra (e.g., in the outskirts). Astronomers can then probe to larger radii and build samples with homogeneous coverage. They can disentangle the features and the mechanisms which cloud our understanding of galaxy formation.

A critical component will be the identification, quantification, and mitigation of bias as discussed previously. Finally, we will attempt to verify the model and any bias using new observational data from the Integral Field Unit for Magellan (IFUM) [32].

3.2 Preliminary Results

Astronomers have previously "learned" models which map between photometric data and stellar mass [3, 51]. Typically, the stellar masses are measured by regressing broad band photometry or spectroscopy against single stellar population models. Sometimes dynamical masses are used [39]. These models have led to hundreds (if not thousands) of papers which undertook individualized science based on photometrically inferred galaxy stellar masses. Imagine the impact if one could build a model that covers a range from galaxy-sized mass scales ($10^{10} M_{\odot}$) all the way down to *galaxy feature-sized* mass scales ($10^{6} M_{\odot}$).

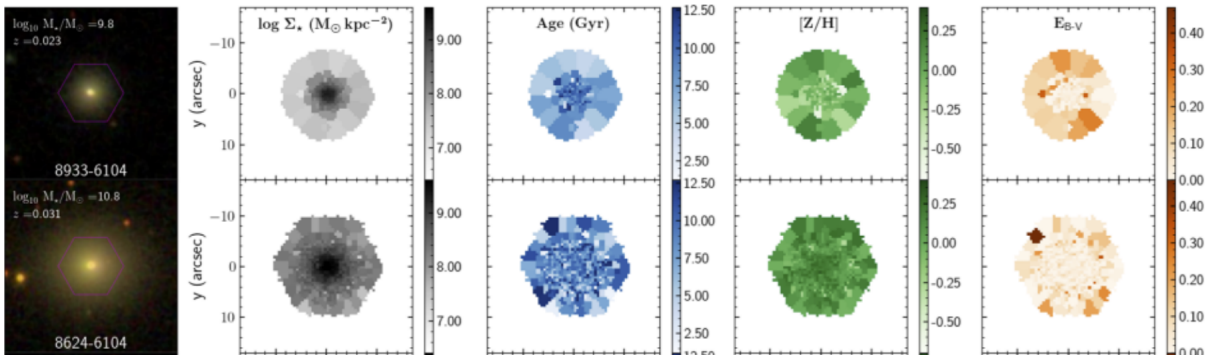

Figure 7: Left shows the multi-band SDSS photometry observed in 180s (3×1 minute drifts over each CCD/band in *gri*). The other panels show inferred astrophysical parameters from the spaxel spectroscopy using the MaNGA Integral Field Unit (IFU). Each galaxy's spectroscopic exposure required ∼8000-10000s.

For our preliminary effort, we will utilize the spatio-spectra data from the SDSS MaNGA survey. In **Figure 7**, we show an example of the data from MaNGA Value Added Catalog called *Firefly* (Neumann et al. 2022). The astrophysical properties are inferred in Voronoi Tesselation (VT) cells which have near constant signal-to-noise in the spectroscopic continua. The spectra in these cells are fit to stellar population models to infer properties like stellar mass, age, metallicity (Z) and internal extinction E(B-V).

We next built an AI system to map the VT pixel data from 5-band images to their spectroscopically inferred VT cell stellar masses. As truth input, we used the MaNGA VT cell stellar masses. The target data is the *ugriz* imaging data, with fluxes corrected for Galactic extinction. The redshift is treated as a free parameter in the training. We believe this is the first attempt to use an AI system to model the stellar distribution inside galaxies at kpc scales using just the 2D image data.

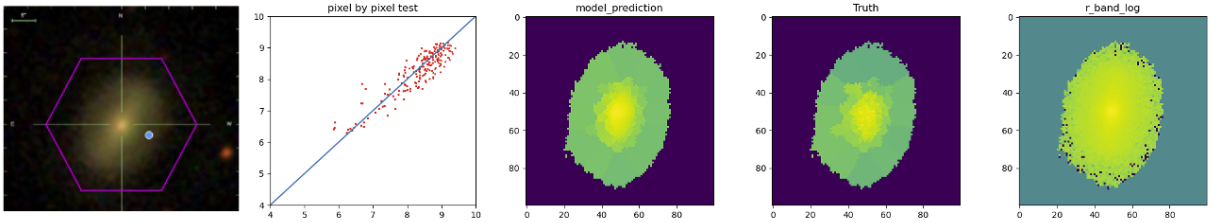

Figure 8: From left to right: The SDSS color composite of the MaNGA target; next is the log surface stellar mass ($M_{\odot}$) predicted purely from the photometric pixel fluxes (y-axis) versus the mass from the Firefly spectroscopic fitting. The scatter is ~ 0.25dex. We then show the model prediction from the CNN, which should resemble the Truth, which is taken directly from the Firefly data and used to train the model. These heatmaps are log stellar mass in VT cells. The model prediction was made using *only* the pixel-level photometry, and we show just the r-band in the right-most panel. We fed the CNN *ugriz* bands to train the model after applying a galactic extinction correction to the pixel fluxes.

In **Figure 8**, we show the results for a single galaxy. We find that the photometry alone can predict the stellar mass per VT pixel down to very small mass scales (e.g., $10^{6} M_{\odot}$).

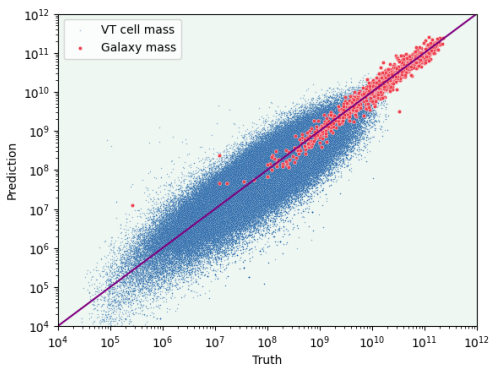

Figure 9: The VT cell (blue) and summed galaxy (red) stellar mass predictions versus truths for the MaNGA data. We train on 70% and we use 10% for tuning the CNN hyperparameters. We test on 20% of the data.

We have identified two key components to the success of this modeling. First, the locations of the VT cells. For instance, if we scramble the VT pixels, we lose most of the correlation. Second, the filter bands. If we just use the *r*-band image in the modeling, we again lose most of the correlation. So to succeed as in Figure 8, the model requires all of the 5-band *ugriz* pixel information *and* the locations, sizes, and shapes of the VT cells which contain the "true" stellar masses. The machine then learns the additional covariant dependence between pixels, i.e., that the stellar populations in spiral arms are different from spiral bulges.

In **Figure 9**, we show predictions of our convolution neural network trained on 2000 MaNGA galaxies. We used 300 galaxies to converge on the hyperparameters of the CNN (e.g., the number of layers). We then tested on the remaining 700 galaxies. The red is the integrated stellar masses per galaxy. The blue dots are the individual VT cells in the 700 galaxies. The scatter for the total stellar mass is less than 0.15dex and less than 0.35dex for the cells. This is an exciting result. Our AI system can infer stellar mass from imaging pixels at good precision for over 5 orders of magnitude in mass and at scales of kpc. It can recover integrated galaxy stellar masses for an additional 2 orders of magnitude at even better precision.

3.3 Planned Effort

Based on our preliminary efforts, we propose to conduct the science use case described earlier in this section. We need more detailed morphological classifications, as well as descriptive internal features. We will use the larger and higher quality GZ morphologies from GZ2 based on DECaLS and other surveys and expand to more than two classes. We will also use the Galaxy10 data and sub-classes [28]. For this effort, we will need to re-derive our loss function for multi-class probabilities. We will create AI system models that have had their biases mitigated and publish them as we study the radial metallicity and age gradients as a function of (localized and integrated) stellar mass and morphology. With the AI model, our sample sizes will be many times larger than any previous study. Lastly, we will use crowd-sourced feature masks from GZ-3D [31] to train models specific to the feature type.

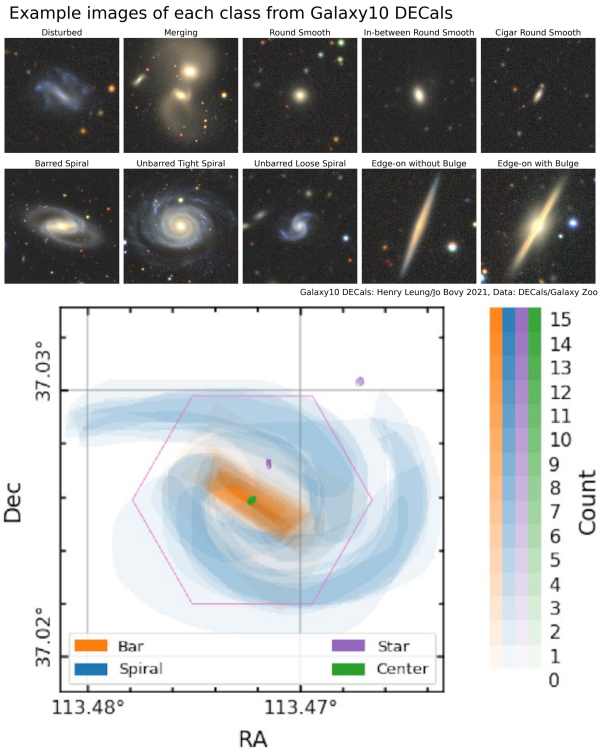

Figure 10: Top: The 10 classes of Galaxy10. Bottom: GZ-3D crowded annotated features in a spiral, which we can use to train better models specific to the feature-type.

In all cases will assess the gold standard labeling bias discussed in Section 2. We will investigate additional parameters which can cause bias (e.g., S/N, observed spectral resolution, VT cell size, etc). We will use our experience with GradCAM to interpret the models. Finally, we will apply for IFUM time on the 6.5m Magellan telescope (through the internal Michigan telescope allocation committee) to collect data in regions where the S/N is too low for SDSS. We will use the data to assess trustworthiness.

4 Broader Impacts

Machine learning is playing an increase role in decision making in the health fields. Consider Lebovitz et al. (2021) [27], who found high uncertainty of experts' know-what knowledge captured in ground truth labels used to train and validate ML models used in decision making in a hospital setting. Zhang et al. (2020) [55] recognized the increasing use of supervised learning methods for segmentation tasks in the medical image domain. They studied the reliability of individual annotators and the true segmentation label distributions, using two coupled CNNs. However, there are few examples of identifying and mitigating gold standard labeling bias.

The National Institutes of Standards and Technology recent published Special Publication 1270 which brings to light the challenges of AI bias [43]. We quote from the Executive Summary:

[H]uman and systemic institutional and societal factors are significant sources of AI bias as well, and are currently overlooked. Successfully meeting this challenge will require taking all forms of bias into account. This means expanding our perspective beyond the machine learning pipeline to recognize and investigate how this technology is both created within and impacts our society.

Our prior work on the subject of gold standard labeling bias is part of the bias that is being overlooked. As emphasized in the first section of this proposal, our goal is to implement AI systems on astronomical data in order solve an interesting astrophysical question which requires the identification, quantification, and mitigation of bias. In doing so, we hope to pave the way for general approaches to trustworthy AI that can be used in other domains. In this section, we explain how.

4.1 Collaboration with MIDAS

As described in detail in the Facilities, Equipment, and Other Resources, the PI is an affiliate member and will collaborate with the The Michigan Institute for Data Science (MIDAS), a reputed center of excellence in data science and AI. This proposal contains elements inside four of the five main "pillars" on which MIDAS stands. However the fifth pillar: "transforming health interventions through the adoption of cutting-edge analytics", is one where our algorithms and research described in Section 2 can be most transformative. The PI will collaborative with MIDAS to build out our techniques to identify, quantify, and mitigate bias with a focus on public health datasets.

4.2 Success from bias quantification, not just AUC/ROC

We will start by accessing data used in a recent Kaggle competition which may show gold standard bias: the RSNA Screening Mammography Breast Cancer Detection dataset. These data contain radiogram images with metadata which can be used as the intrinsic and observable parameters. Unlike trying to reach the highest rank in terms of accuracy, we will instead use these data to assess the gold standard labeling bias. The PI will work closely with staff and researchers in MIDAS to identify additional datasets which can be used for these bias studies. The PI will also work with MIDAS to disseminate our tools and research to appropriate audiences, including through seminars and colloquia, as well as through student poster presentations and possibly dedicated workshops.

5 Proposed Effort

- **Data wrangling**: We will use the latest Galaxy Zoo SDSS morphologies [54]. We will use the SDSS DR17 imaging and catalog data [1]. We will us DECaLS multi-band galaxy imaging and catalog data

using the Legacy Survey release 9 [42]. We will use the 10 classifcations from Galaxies10, which also includes DECaLS [52]. We will use the SDSS DR17 Firefly MaNGA IFU resolved stellar populations [36]. We will use the Galaxy Zoo 3D maps [31].

- **Algorithmic development:** Besides image resolution (i.e., observed size/observed PSF), we will incorporate apparent magnitude, sky surface brightness, and object crowding into the bias quantification via equation 3. We will move from a binary classification model to a multi-class model in the loss function via equation 7. As we utilize the spectroscopy, we will evaluate additional observable parameters such as spectral resolution, S/N, and VT cell sizes as possible sources of labeling bias. In this case, as opposed to human annotators, the labels would be from the SED fitting to the SEDs (i.e., with stellar mass or age or metallicity as a "label"). Our plan is to utilize binary classifications (young, old, metal rich, metal poor, etc) as the labels. Any bias we find would then be a result of a human implemented regression scheme.
- **Science implementation:** With the models in hand, our team will analyze the galaxies for metal and age gradients using the pixels. We will be able to measure radial gradients to much further distances. We will also conduct these same exercises using the predictions from the pixels for datasets with morphological features.
- **AI Trustworthiness:** Conduct analyses similar to Figure 5. The PI will lead an observing proposal to collect IFU data in galaxies with predictions made by the model. These data will serve as a baseline truth (like the HST image compared to the SDSS image in Figure 3.

5.1 Personnel

- The PI will lead the project, conduct some of the research, manage the graduate student research assistant (GSRA), and coordinate with the unfunded collaborators. The PI will conduct the annual reporting effort. The PI is an experienced observer and will conduct the observational effort. The PI will spend 50% of his available research time dedicated to this project.
- The GSRA will conduct some of the research, as coordinated by the PI. The GSRA will be responsible for a first author manuscript, expected to be submitted to journal by December, 2026. The GSRA will be 100% of their available research time on this project.
- Dr. Regier (Statisics) and Dr. Cabrera-Vivas (Computer Science) are listed as unfunded collaborators on this proposal. They are each leaders in their fields and their leadership in advancing Bayesian techniques (Regier) and machine learning (Cabrera) will be helpful for the work described in this proposal. The PI has professional relationships with these researchers where they often discuss topics related to this proposal. These discussions may result in direct avenues of research which align with this proposal. Such avenues will be pursued (as necessary) through funded resources (e.g., the PI or the graduate student). As unfunded collaborators, any time that they spend on this project will be voluntary and not required for the successful completion of the effort described in the work plan.

5.2 Timeline

- Year 1: Data wrangling and extending the bias models to a larger number of classes. Apply for observing time on IFUM.
- Year 2: Publish the bias measurements and mitigation results for the new datasets (e.g., GZ2 and DECaLS). Conduct observations. Continue with algorithm development. Publish the MaNGA Firefly AI system model to predict spatio-spectro properties from image pixel data.
- Year 3: Finalize the IFUM data analysis. Publish the new CNN model which incorporates morphology and features.

The work identified in the Broader Impacts will be conducted concurrently.

6 Results from Prior NSF Support

NSF AST-1812739 *Constraining Cosmological Parameters with Galaxy Cluster Phase Spaces* 06/13/2018 - 06/30/2023, PI: C. Miller, $382,882

6.1 Intellectual Merit:

We aimed to study the escape-velocity/weak-lensing masses for 100 galaxy clusters. We identified a set of $\sim$60 clusters meeting our spectroscopic criteria and lensing criteria. Another 20 clusters were observed using the Michigan Magellan Fiber Spectrograph (M2FS). These data were reduced and preliminary phase-spaces were analyzed. We reached 80% of our original sample size goals. In the process, we built a python-based M2FS spectroscopic pipeline and verified it against SDSS and other literature spectra and with repeat M2FS observations. This pipeline was released on Dr. Anthony Kremin's *github* and there is a paper in preparation. Our team was the first to apply the Action-based AGAMA theoretical phase-space package to assess the accuracy and precision of observed phase-space edges. We also incorporated multiple N-body simulations to conduct an end-to-end analysis of the effects of halo mass, cosmology, and velocity anisotropy on the escape edge accuracy and precision. None of these were found to impact the edge measurement beyond a few percent. This is an important discovery negating the original suggestion by Diaferio and Geller in the late 1990s. We then made the first escape-edge cluster mass measurement on Abell 1063 using our escape edge inference techniques and showed that (a) for well-sampled phase-spaces, the statistical precision is competitive with weak lensing and (b) that the escape inferred mass matches very closely to weak lensing (as well as SZ masses). The outcomes from this effort put our team at state-of-the-art when it comes to cluster escape-edge measurement and dynamical mass inference. Papers submitted or published as part of this effort are cited in the References as [15, 17, 29, 40].

6.2 Broader Impacts

This project enabled all of the students involved to further their educational background and also gain meaningful employment. Two students earned their PhD under this proposal: Dr. Vitali Halenka, 2019 PhD in Physics: "In Quest of Devising Tools of Probing Cosmology and Gravity using Galaxy Clusters". Dr. Anthony Kremin, 2020 PhD in Physics. "Dynamical Constraints of Galaxy Clusters via Spectroscopic Observations. " In addition, Paige Vansickle and Ray Wang both conducted undergraduate research and senior theses for this NSF project as part of their BSci in Astronomy. Other BSci Astronomy undergraduate students who contributed were Mason Cleveland, Samuel Kim, and Nick Susemiehl.

Training and Employment inside Astronomy: Dr. Kremin took a post-doctoral research position at Lawrence Berkeley National Labs, working on the data team for Dark Energy Spectroscopic Instrument (DESI). Ray is attending graduate school for his PhD at Michigan State University. Nick went to the Georgia Institute of Technology to earn his Masters Degree in Analytics. He is now a Data Analysist at the NASA Exoplanet Archive at Caltech. P. Vansickle is a Planetarium Educator at the Cernan Earth and Space Center.

Training and Employment outside Astronomy: Dr. Halenka became a Data Scientist for SubscriberWise and is now the CEO of the company. Samuel Kim is a Data Science Analyst at Discover Financial Services. Mason Cleveland is a Programmer Analyst The MGH/HST Martinos Center for Biomedical Imaging

Outreach: The PI aimed to re-develop the Astronomy Merit Badge for BSA Scouting. An outline and plan was defined and discussed with local BSA Council Members where it was determined that the BSA Scouting calendar for Merit Badge updates fell outside the window of the funded period. The PI instead engaged on the badge effort by completing training for a Merit Badge Councilor (MBC) appointment within BSA. A MBC is a teacher, a mentor, and a helping hand in the progression of Scouting. The PI then taught the current Astronomy badge requirements to $\sim$40 youths over two years of summer camp at Cole Canoe Base. As part of that effort, I was able to engage youths with the promise of STEM career through Astronomy.

References

[1] Abdurro'uf, Katherine Accetta, Conny Aerts, Víctor Silva Aguirre, Romina Ahumada, Nikhil Ajgaonkar, N. Filiz Ak, Shadab Alam, Carlos Allende Prieto, Andrés Almeida, Friedrich Anders, Scott F. Anderson, Brett H. Andrews, Borja Anguiano, Erik Aquino-Ortíz, Alfonso Aragón-Salamanca, Maria Argudo-Fernández, Metin Ata, Marie Aubert, Vladimir Avila-Reese, Carles Badenes, Rodolfo H. Barbá, Kat Barger, Jorge K. Barrera-Ballesteros, Rachael L. Beaton, Timothy C. Beers, Francesco Belfiore, Chad F. Bender, Mariangela Bernardi, Matthew A. Bershady, Florian Beutler, Christian Moni Bidin, Jonathan C. Bird, Dmitry Bizyaev, Guillermo A. Blanc, Michael R. Blanton, Nicholas Fraser Boardman, Adam S. Bolton, Médéric Boquien, Jura Borissova, Jo Bovy, W. N. Brandt, Jordan Brown, Joel R. Brownstein, Marcella Brusa, Johannes Buchner, Kevin Bundy, Joseph N. Burchett, Martin Bureau, Adam Burgasser, Tuesday K. Cabang, Stephanie Campbell, Michele Cappellari, Joleen K. Carlberg, Fábio Carneiro Wanderley, Ricardo Carrera, Jennifer Cash, Yan-Ping Chen, Wei-Huai Chen, Brian Cherinka, Cristina Chiappini, Peter Doohyun Choi, S. Drew Chojnowski, Haeun Chung, Nicolas Clerc, Roger E. Cohen, Julia M. Comerford, Johan Comparat, Luiz da Costa, Kevin Covey, Jeffrey D. Crane, Irene Cruz-Gonzalez, Connor Culhane, Katia Cunha, Y. Sophia Dai, Guillermo Damke, Jeremy Darling, Jr. Davidson, James W., Roger Davies, Kyle Dawson, Nathan De Lee, Aleksandar M. Diamond-Stanic, Mariana Cano-Díaz, Helena Domínguez Sánchez, John Donor, Chris Duckworth, Tom Dwelly, Daniel J. Eisenstein, Yvonne P. Elsworth, Eric Emsellem, Mike Eracleous, Stephanie Escoffier, Xiaohui Fan, Emily Farr, Shuai Feng, José G. Fernández-Trincado, Diane Feuillet, Andreas Filipp, Sean P. Fillingham, Peter M. Frinchaboy, Sebastien Fromenteau, Lluís Galbany, Rafael A. García, D. A. García-Hernández, Junqiang Ge, Doug Geisler, Joseph Gelfand, Tobias Géron, Benjamin J. Gibson, Julian Goddy, Diego Godoy-Rivera, Kathleen Grabowski, Paul J. Green, Michael Greener, Catherine J. Grier, Emily Griffith, Hong Guo, Julien Guy, Massinissa Hadjara, Paul Harding, Sten Hasselquist, Christian R. Hayes, Fred Hearty, Jesús Hernández, Lewis Hill, David W. Hogg, Jon A. Holtzman, Danny Horta, Bau-Ching Hsieh, Chin-Hao Hsu, Yun-Hsin Hsu, Daniel Huber, Marc Huertas-Company, Brian Hutchinson, Ho Seong Hwang, Héctor J. Ibarra-Medel, Jacob Ider Chitham, Gabriele S. Ilha, Julie Imig, Will Jaekle, Tharindu Jayasinghe, Xihan Ji, Jennifer A. Johnson, Amy Jones, Henrik Jönsson, Ivan Katkov, Dr. Khalatyan, Arman, Karen Kinemuchi, Shobhit Kisku, Johan H. Knapen, Jean-Paul Kneib, Juna A. Kollmeier, Miranda Kong, Marina Kounkel, Kathryn Kreckel, Dhanesh Krishnarao, Ivan Lacerna, Richard R. Lane, Rachel Langgin, Ramon Lavender, David R. Law, Daniel Lazarz, Henry W. Leung, Ho-Hin Leung, Hannah M. Lewis, Cheng Li, Ran Li, Jianhui Lian, Fu-Heng Liang, Lihwai Lin, Yen-Ting Lin, Sicheng Lin, Chris Lintott, Dan Long, Penélope Longa-Peña, Carlos López-Cobá, Shengdong Lu, Britt F. Lundgren, Yuanze Luo, J. Ted Mackereth, Axel de la Macorra, Suvrath Mahadevan, Steven R. Majewski, Arturo Manchado, Travis Mandeville, Claudia Maraston, Berta Margalef-Bentabol, Thomas Masseron, Karen L. Masters, Savita Mathur, Richard M. McDermid, Myles Mckay, Andrea Merloni, Michael Merrifield, Szabolcs Meszaros, Andrea Miglio, Francesco Di Mille, Dante Minniti, Rebecca Minsley, Antonela Monachesi, Jeongin Moon, Benoit Mosser, John Mulchaey, Demitri Muna, Ricardo R. Muñoz, Adam D. Myers, Natalie Myers, Seshadri Nadathur, Preethi Nair, Kirpal Nandra, Justus Neumann, Jeffrey A. Newman, David L. Nidever, Farnik Nikakhtar, Christian Nitschelm, Julia E. O'Connell, Luis Garma-Oehmichen, Gabriel Luan Souza de Oliveira, Richard Olney, Daniel Oravetz, Mario Ortigoza-Urdaneta, Yeisson Osorio, Justin Otter, Zachary J. Pace, Nelson Padilla, Kaike Pan, Hsi-An Pan, Taniya Parikh, James Parker, Sebastien Peirani, Karla Peña Ramírez, Samantha Penny, Will J. Percival, Ismael Perez-Fournon, Marc Pinsonneault, Frédérick Poidevin, Vijith Jacob Poovelil, Adrian M. Price-Whelan, Anna Bárbara de Andrade Queiroz, M. Jordan Raddick, Amy Ray, Sandro Barboza Rembold, Nicole Riddle, Rogemar A. Riffel, Rogério Riffel, Hans-Walter Rix, Annie C. Robin, Aldo Rodríguez-Puebla, Alexandre Roman-Lopes, Carlos Román-Zúñiga, Benjamin Rose, Ashley J. Ross, Graziano Rossi,

Kate H. R. Rubin, Mara Salvato, Sebástian F. Sánchez, José R. Sánchez-Gallego, Robyn Sanderson, Felipe Antonio Santana Rojas, Edgar Sarceno, Regina Sarmiento, Conor Sayres, Elizaveta Sazonova, Adam L. Schaefer, Ricardo Schiavon, David J. Schlegel, Donald P. Schneider, Mathias Schultheis, Axel Schwope, Aldo Serenelli, Javier Serna, Zhengyi Shao, Griffin Shapiro, Anubhav Sharma, Yue Shen, Matthew Shetrone, Yiping Shu, Joshua D. Simon, M. F. Skrutskie, Rebecca Smethurst, Verne Smith, Jennifer Sobeck, Taylor Spoo, Dani Sprague, David V. Stark, Keivan G. Stassun, Matthias Steinmetz, Dennis Stello, Alexander Stone-Martinez, Thaisa Storchi-Bergmann, Guy S. Stringfellow, Amelia Stutz, Yung-Chau Su, Manuchehr Taghizadeh-Popp, Michael S. Talbot, Jamie Tayar, Eduardo Telles, Johanna Teske, Ani Thakar, Christopher Theissen, Andrew Tkachenko, Daniel Thomas, Rita Tojeiro, Hector Hernandez Toledo, Nicholas W. Troup, Jonathan R. Trump, James Trussler, Jacqueline Turner, Sarah Tuttle, Eduardo Unda-Sanzana, José Antonio Vázquez-Mata, Marica Valentini, Octavio Valenzuela, Jaime Vargas-González, Mariana Vargas-Magaña, Pablo Vera Alfaro, Sandro Villanova, Fiorenzo Vincenzo, David Wake, Jack T. Warfield, Jessica Diane Washington, Benjamin Alan Weaver, Anne-Marie Weijmans, David H. Weinberg, Achim Weiss, Kyle B. Westfall, Vivienne Wild, Matthew C. Wilde, John C. Wilson, Robert F. Wilson, Mikayla Wilson, Julien Wolf, W. M. Wood-Vasey, Renbin Yan, Olga Zamora, Gail Zasowski, Kai Zhang, Cheng Zhao, Zheng Zheng, Zheng Zheng, and Kai Zhu. The Seventeenth Data Release of the Sloan Digital Sky Surveys: Complete Release of MaNGA, MaStar, and APOGEE-2 Data. *Astrophys. J. Suppl.* , 259(2):35, April 2022. doi: 10.3847/1538-4365/ac4414.

[2] Steven P. Bamford, Robert C. Nichol, Ivan K. Baldry, Kate Land, Chris J. Lintott, Kevin Schawinski, Anže Slosar, Alexander S. Szalay, Daniel Thomas, Mehri Torki, Dan Andreescu, Edward M. Edmondson, Christopher J. Miller, Phil Murray, M. Jordan Raddick, and Jan Vandenberg. Galaxy Zoo: the dependence of morphology and colour on environment*. *MNRAS* , 393(4):1324–1352, March 2009. doi: 10.1111/j.1365-2966.2008.14252.x.

[3] Eric F. Bell, Daniel H. McIntosh, Neal Katz, and Martin D. Weinberg. The Optical and Near-Infrared Properties of Galaxies. I. Luminosity and Stellar Mass Functions. *Astrophys. J. Suppl.* , 149(2):289–312, December 2003. doi: 10.1086/378847.

[4] Asa F. L. Bluck, J. Trevor Mendel, Sara L. Ellison, Jorge Moreno, Luc Simard, David R. Patton, and Else Starkenburg. Bulge mass is king: the dominant role of the bulge in determining the fraction of passive galaxies in the Sloan Digital Sky Survey. *MNRAS* , 441(1):599–629, June 2014. doi: 10.1093/mnras/stu594.

[5] Guillermo F. Cabrera, Chris J. Miller, and Jeff Schneider. Systematic labeling bias: De-biasing where everyone is wrong. In *Pattern Recognition (ICPR), 2014 22nd International Conference on*, page in press. IEEE, 2014.

[6] R. G. Carlberg. Dissipative formation of an elliptical galaxy. *Astrophys. J.* , 286:403–415, November 1984. doi: 10.1086/162615.

[7] Guangwen Chen, Hong-Xin Zhang, Xu Kong, Zesen Lin, Zhixiong Liang, Xinkai Chen, Zuyi Chen, and Zhiyuan Song. The Most Predictive Physical Properties for the Stellar Population Radial Profiles of Nearby Galaxies. *Astrophys. J.* , 895(2):146, June 2020. doi: 10.3847/1538-4357/ab8cc2.

[8] Nelson Christensen, Renate Meyer, Lloyd Knox, and Ben Luey. Bayesian methods for cosmological parameter estimation from cosmic microwave background measurements. *Classical and Quantum Gravity*, 18(14):2677–2688, July 2001. doi: 10.1088/0264-9381/18/14/306.

[9] A. Dekel, Y. Birnboim, G. Engel, J. Freundlich, T. Goerdt, M. Mumcuoglu, E. Neistein, C. Pichon, R. Teyssier, and E. Zinger. Cold streams in early massive hot haloes as the main mode of galaxy formation. *Nature* , 457(7228):451–454, January 2009. doi: 10.1038/nature07648.

[10] O. J. Eggen, D. Lynden-Bell, and A. R. Sandage. Evidence from the motions of old stars that the Galaxy collapsed. *Astrophys. J.* , 136:748, November 1962. doi: 10.1086/147433.

[11] Jerome J. Fang, S. M. Faber, David C. Koo, and Avishai Dekel. A Link between Star Formation Quenching and Inner Stellar Mass Density in Sloan Digital Sky Survey Central Galaxies. *Astrophys. J.* , 776(1):63, October 2013. doi: 10.1088/0004-637X/776/1/63.

[12] Eric G. Forbes. Gauss and the Discovery of Ceres. *Journal for the History of Astronomy*, 2:195, January 1971. doi: 10.1177/002182867100200305.

[13] Amelia Fraser-McKelvie, Michael Merrifield, Alfonso Aragón-Salamanca, Thomas Peterken, Karen Masters, Coleman Krawczyk, Brett Andrews, Johan H. Knapen, Sandor Kruk, Adam Schaefer, Rebecca Smethurst, Rogério Riffel, Joel Brownstein, and Niv Drory. SDSS-IV MaNGA: stellar population gradients within barred galaxies. *MNRAS* , 488(1):L6–L11, September 2019. doi: 10.1093/mnrasl/slz085.

[14] Masataka Fukugita, Osamu Nakamura, Sadanori Okamura, Naoki Yasuda, John C. Barentine, Jon Brinkmann, James E. Gunn, Mike Harvanek, Takashi Ichikawa, Robert H. Lupton, Donald P. Schneider, Michael A. Strauss, and Donald G. York. A Catalog of Morphologically Classified Galaxies from the Sloan Digital Sky Survey: North Equatorial Region. *Astron. J.* , 134(2):579–593, August 2007. doi: 10.1086/518962.

[15] Jesse B. Golden-Marx and Christopher J. Miller. The Impact of Environment on Late-time Evolution of the Stellar Mass-Halo Mass Relation. *Astrophys. J.* , 878(1):14, June 2019. doi: 10.3847/1538-4357/ab1d55.

[16] R. M. González Delgado, E. Pérez, R. Cid Fernandes, R. García-Benito, R. López Fernández, N. Vale Asari, C. Cortijo-Ferrero, A. L. de Amorim, E. A. D. Lacerda, S. F. Sánchez, M. D. Lehnert, and C. J. Walcher. Spatially-resolved star formation histories of CALIFA galaxies. Implications for galaxy formation. *Astron. & Astrophys.* , 607:A128, November 2017. doi: 10.1051/0004-6361/201730883.

[17] Vitali Halenka, Christopher J. Miller, and Paige Vansickle. Quantifying the Projected Suppression of Cluster Escape Velocity Profiles. *arXiv e-prints*, art. arXiv:2003.02733, March 2020. doi: 10.48550/arXiv.2003.02733.

[18] Ross E. Hart, Steven P. Bamford, Kyle W. Willett, Karen L. Masters, Carolin Cardamone, Chris J. Lintott, Robert J. Mackay, Robert C. Nichol, Christopher K. Rosslowe, Brooke D. Simmons, and Rebecca J. Smethurst. Galaxy Zoo: comparing the demographics of spiral arm number and a new method for correcting redshift bias. *Monthly Notices of the Royal Astronomical Society*, 461(4):3663–3682, 07 2016. ISSN 0035-8711. doi: 10.1093/mnras/stw1588. URL `https://doi.org/10.1093/mnras/stw1588`.

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[20] Philip F. Hopkins, Darren Croton, Kevin Bundy, Sadegh Khochfar, Frank van den Bosch, Rachel S. Somerville, Andrew Wetzel, Dusan Keres, Lars Hernquist, Kyle Stewart, Joshua D. Younger, Shy

Genel, and Chung-Pei Ma. Mergers in ΛCDM: Uncertainties in Theoretical Predictions and Interpretations of the Merger Rate. *Astrophys. J.* , 724(2):915–945, December 2010. doi: 10.1088/0004-637X/724/2/915.

[21] M. Huertas-Company, J. A. L. Aguerri, M. Bernardi, S. Mei, and J. Sánchez Almeida. Revisiting the Hubble sequence in the SDSS DR7 spectroscopic sample: a publicly available Bayesian automated classification. *Astron. & Astrophys.* , 525:A157, January 2011. doi: 10.1051/0004-6361/201015735.

[22] Shardha Jogee, Nick Scoville, and Jeffrey D. P. Kenney. The Central Region of Barred Galaxies: Molecular Environment, Starbursts, and Secular Evolution. *Astrophys. J.* , 630(2):837–863, September 2005. doi: 10.1086/432106.

[23] Sugata Kaviraj. The importance of minor-merger-driven star formation and black hole growth in disc galaxies. *MNRAS* , 440(4):2944–2952, June 2014. doi: 10.1093/mnras/stu338.

[24] Jeremy Martin Kubica, Joseph Masiero, Andrew Moore, Robert Jedicke, and Andrew J. Connolly. Variable kd-tree algorithms for efficient spatial pattern search. Technical Report CMU-RI-TR-05-43, Carnegie Mellon University, Pittsburgh, PA, September 2005.

[25] O. Lahav and M. C. Storrie-Lombardi. *Neural networks in astronomy*. Pergamon Press, 1994.

[26] Richard B. Larson. Dynamical models for the formation and evolution of spherical galaxies. *MNRAS* , 166:585–616, March 1974. doi: 10.1093/mnras/166.3.585.

[27] Sarah Lebovitz, Natalia Levina, and Hila Lifshitz-Assaf. Is AI Ground Truth Really 'True'? The Dangers of Training and Evaluating AI Tools Based on Experts. *Management Information Systems Quarterly*, 45:1501–1525, 2021. doi: 10.25300/MISQ/2021/16564.

[28] Henry W. Leung and Jo Bovy. Deep learning of multi-element abundances from high-resolution spectroscopic data. *MNRAS* , 483(3):3255–3277, March 2019. doi: 10.1093/mnras/sty3217.

[29] Wentao Luo, Jiajun Zhang, Vitali Halenka, Xiaohu Yang, Surhud More, Christopher J. Miller, Lei Liu, and Feng Shi. Emergent Gravity Fails to Explain Color-dependent Galaxy-Galaxy Lensing Signal from SDSS DR7. *Astrophys. J.* , 914(2):96, June 2021. doi: 10.3847/1538-4357/abf4c2.

[30] Karen L. Masters, Robert C. Nichol, Ben Hoyle, Chris Lintott, Steven P. Bamford, Edward M. Edmondson, Lucy Fortson, William C. Keel, Kevin Schawinski, Arfon M. Smith, and Daniel Thomas. Galaxy Zoo: bars in disc galaxies. *MNRAS* , 411(3):2026–2034, March 2011. doi: 10.1111/j.1365-2966.2010.17834.x.

[31] Karen L. Masters, Coleman Krawczyk, Shoaib Shamsi, Alexander Todd, Daniel Finnegan, Matthew Bershady, Kevin Bundy, Brian Cherinka, Amelia Fraser-McKelvie, Dhanesh Krishnarao, Sandor Kruk, Richard R. Lane, David Law, Chris Lintott, Michael Merrifield, Brooke Simmons, Anne-Marie Weijmans, and Renbin Yan. Galaxy Zoo: 3D - crowdsourced bar, spiral, and foreground star masks for MaNGA target galaxies. *MNRAS* , 507(3):3923–3935, November 2021. doi: 10.1093/mnras/stab2282.

[32] Mario Mateo, John I. Bailey, Yingyi Song, Jeffrey Crane, Charlie Hull, Stephen Shectman, and Christoph Birk. IFUM: integral field units for Magellan. In Christopher J. Evans, Julia J. Bryant, and Kentaro Motohara, editors, *Ground-based and Airborne Instrumentation for Astronomy IX*, volume 12184 of *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, page 121845P, August 2022. doi: 10.1117/12.2629506.

[33] Esteban Medina-Rosales, Guillermo Cabrera-Vives, and Christopher J. Miller. Mitigating Bias in Deep Learning: Training Unbiased Models on Biased Data for the Morphological Classification of Galaxies. *arXiv e-prints*, art. arXiv:2308.11007, August 2023. doi: 10.48550/arXiv.2308.11007.

[34] H. J. Mo, Shude Mao, and Simon D. M. White. The formation of galactic discs. *MNRAS* , 295(2): 319–336, April 1998. doi: 10.1046/j.1365-8711.1998.01227.x.

[35] Preethi B. Nair and Roberto G. Abraham. A Catalog of Detailed Visual Morphological Classifications for 14,034 Galaxies in the Sloan Digital Sky Survey. *Astrophys. J. Suppl.* , 186(2):427–456, February 2010. doi: 10.1088/0067-0049/186/2/427.

[36] Justus Neumann, Daniel Thomas, Claudia Maraston, Lewis Hill, Lorenza Nanni, Oliver Wenman, Jianhui Lian, Johan Comparat, Violeta Gonzalez-Perez, Kyle B. Westfall, Renbin Yan, Yanping Chen, Guy S. Stringfellow, Matthew A. Bershady, Joel R. Brownstein, Niv Drory, and Donald P. Schneider. The MaNGA FIREFLY value added catalogue: resolved stellar populations of 10 010 nearby galaxies. *MNRAS* , 513(4):5988–6012, July 2022. doi: 10.1093/mnras/stac1260.

[37] Taniya Parikh, Daniel Thomas, Claudia Maraston, Kyle B. Westfall, Brett H. Andrews, Nicholas Fraser Boardman, Niv Drory, and Grecco Oyarzun. SDSS-IV MaNGA: radial gradients in stellar population properties of early-type and late-type galaxies. *MNRAS* , 502(4):5508–5527, April 2021. doi: 10.1093/mnras/stab449.

[38] Michael Park, Erin Leahey, and Russell J. Funk. Papers and patents are becoming less disruptive over time. *Nature* , 613(7942):138–144, January 2023. doi: 10.1038/s41586-022-05543-x.

[39] A. Rettura, P. Rosati, V. Strazzullo, M. Dickinson, R. A. E. Fosbury, B. Rocca-Volmerange, A. Cimatti, S. di Serego Alighieri, H. Kuntschner, B. Lanzoni, M. Nonino, P. Popesso, D. Stern, P. R. Eisenhardt, C. Lidman, and S. A. Stanford. Comparing dynamical and photometric-stellar masses of early-type galaxies at z $\sim$1. *Astron. & Astrophys.* , 458(3):717–726, November 2006. doi: 10.1051/0004-6361:20065273.

[40] Alex Rodriquez, Christopher J Miller, Vitali Halenka, and Anthony Kremin. Escape Velocity Mass of Abell 1063. *submitted to Astrophys. J.* , May 2023.

[41] K. Sakamoto, S. K. Okumura, S. Ishizuki, and N. Z. Scoville. Bar-driven Transport of Molecular Gas to Galactic Centers and Its Consequences. *Astrophys. J.* , 525(2):691–701, November 1999. doi: 10.1086/307910.

[42] D. Schlegel, A. Dey, D. Herrera, S. Juneau, M. Landriau, D. Lang, A. Meisner, J. Moustakas, A. Myers, E. Schlafly, F. Valdes, B. Weaver, M. Zhang, R. Zhou, and DESI Legacy Imaging Surveys Team. DESI Legacy Imaging Surveys Data Release 9. In *American Astronomical Society Meeting Abstracts*, volume 53 of *American Astronomical Society Meeting Abstracts*, page 235.03, January 2021.

[43] Reva Schwartz, Apostol Vassilev, Kristen K. Greene, Lori Perine, Andrew Burt, and Patrick Hall. Towards a standard for identifying and managing bias in artificial intelligence, 2022-03-15 04:03:00 2022.

[44] J. L. Sersic. *Atlas de galaxias australes*. Córdoba: Obs. Astronómico, 1968.

[45] Ravi K. Sheth and Giuseppe Tormen. On the environmental dependence of halo formation. *MNRAS* , 350(4):1385–1390, June 2004. doi: 10.1111/j.1365-2966.2004.07733.x.

[46] O. Sheynin. On the history of the principle of least squares. *Archive for History of Exact Sciences*, 46 (1):39–54, August 1993.

[47] Michael J. Smith and James E. Geach. Astronomia ex machina: a history, primer and outlook on neural networks in astronomy. *Royal Society Open Science*, 10(5):221454, May 2023. doi: 10.1098/rsos.221454.

[48] Ashley Spindler and David Wake. The differing relationships between size, mass, metallicity and core velocity dispersion of central and satellite galaxies. *MNRAS* , 468(1):333–345, June 2017. doi: 10.1093/mnras/stx427.

[49] Philip Taylor and Chiaki Kobayashi. The metallicity and elemental abundance gradients of simulated galaxies and their environmental dependence. *MNRAS* , 471(4):3856–3870, November 2017. doi: 10.1093/mnras/stx1860.

[50] J.T. Vanderplas, A.J. Connolly, Ž. Ivezić, and A. Gray. Introduction to astroml: Machine learning for astrophysics. In *Conference on Intelligent Data Understanding (CIDU)*, pages 47 –54, oct. 2012. doi: 10.1109/CIDU.2012.6382200.

[51] Jakob Walcher, Brent Groves, Tamás Budavári, and Daniel Dale. Fitting the integrated spectral energy distributions of galaxies. *Astrophys. Space Sci.* , 331:1–52, January 2011. doi: 10.1007/s10509-010-0458-z.

[52] Mike Walmsley, Chris Lintott, Tobias Géron, Sandor Kruk, Coleman Krawczyk, Kyle W. Willett, Steven Bamford, Lee S. Kelvin, Lucy Fortson, Yarin Gal, William Keel, Karen L. Masters, Vihang Mehta, Brooke D. Simmons, Rebecca Smethurst, Lewis Smith, Elisabeth M. Baeten, and Christine Macmillan. Galaxy Zoo DECaLS: Detailed visual morphology measurements from volunteers and deep learning for 314 000 galaxies. *MNRAS* , 509(3):3966–3988, January 2022. doi: 10.1093/mnras/stab2093.

[53] Kyle W. Willett, Chris J. Lintott, Steven P. Bamford, Karen L. Masters, Brooke D. Simmons, Kevin R. V. Casteels, Edward M. Edmondson, Lucy F. Fortson, Sugata Kaviraj, William C. Keel, Thomas Melvin, Robert C. Nichol, M. Jordan Raddick, Kevin Schawinski, Robert J. Simpson, Ramin A. Skibba, Arfon M. Smith, and Daniel Thomas. Galaxy Zoo 2: detailed morphological classifications for 304 122 galaxies from the Sloan Digital Sky Survey. *MNRAS* , 435(4):2835–2860, November 2013. doi: 10.1093/mnras/stt1458.

[54] Kyle W. Willett, Melanie A. Galloway, Steven P. Bamford, Chris J. Lintott, Karen L. Masters, Claudia Scarlata, B. D. Simmons, Melanie Beck, Carolin N. Cardamone, Edmond Cheung, Edward M. Edmondson, Lucy F. Fortson, Roger L. Griffith, Boris Häußler, Anna Han, Ross Hart, Thomas Melvin, Michael Parrish, Kevin Schawinski, R. J. Smethurst, and Arfon M. Smith. Galaxy Zoo: morphological classifications for 120 000 galaxies in HST legacy imaging. *MNRAS* , 464(4):4176–4203, February 2017. doi: 10.1093/mnras/stw2568.

[55] Le Zhang, Ryutaro Tanno, Mou-Cheng Xu, Chen Jin, Joseph Jacob, Olga Ciccarelli, Frederik Barkhof, and Daniel C. Alexander. Disentangling human error from the ground truth in segmentation of medical images. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.