Draft version August 19, 2017
Preprint typeset using LaTeX style AASTeX6 v. 1.0

INTEGRATING HUMAN AND MACHINE INTELLIGENCE IN GALAXY MORPHOLOGY CLASSIFICATION TASKS

Melanie R. Beck[1], Claudia Scarlata[1], Lucy F. Fortson[1], Chris J. Lintott[2, 3], Melanie A. Galloway[1], Kyle W. Willett[1], B. D. Simmons[2,4,7], Hugh Dickinson[1], Karen L. Masters[5], Philip J. Marshall[6], and Darryl Wright[2]

[1]Minnesota Institute for Astrophysics, University of Minnesota, Minneapolis, MN 55455, USA; beck@astro.umn.edu

[2]Oxford Astrophysics, Denys Wilkinson Building, Keble Road, Oxford OX1 3RH, UK

[3]New College, Oxford OX1 3BN, UK

[4]Center for Astrophysics and Space Sciences, Department of Physics, University of California, San Diego, CA 92093, USA

[5]Institute of Cosmology and Gravitation, University of Portsmouth, Portsmouth, UK

[6]Kavli Institute for Particle Astrophysics and Cosmology, P.O. Box 20450, MS29, Stanford, CA 94309, U.S.A.

[7]Einstein Fellow

ABSTRACT

Quantifying galaxy morphology is a challenging yet scientifically rewarding task. As the scale of data continues to increase with upcoming surveys, traditional classification methods will struggle to handle the load. We present a solution through an integration of visual and automated classifications, preserving the best features of both human and machine. We demonstrate the effectiveness of such a system through a re-analysis of visual galaxy morphology classifications collected during the Galaxy Zoo 2 (GZ2) project. We reprocess the top-level question of the GZ2 decision tree with a Bayesian classification aggregation algorithm dubbed SWAP, originally developed for the Space Warps gravitational lens project. Through a simple binary classification scheme we increase the classification rate nearly 5-fold classifying 226,124 galaxies in 92 days of GZ2 project time while reproducing labels derived from GZ2 classification data with 95.7% accuracy.

We next combine this with a Random Forest machine learning algorithm that learns on a suite of non-parametric morphology indicators widely used for automated morphologies. We develop a decision engine that delegates tasks between human and machine and demonstrate that the combined system provides a factor of 11.4 increase in the classification rate, classifying 210,543 galaxies in just 32 days of GZ2 project time with 93.5% accuracy. As the Random Forest algorithm requires a minimal amount of computational cost, this result has important implications for galaxy morphology identification tasks in the era of *Euclid* and other large-scale surveys.

Keywords: galaxies: general — galaxies: morphology — methods: data analysis — methods: machine learning

1. INTRODUCTION

Astronomers have made use of visual galaxy morphologies to understand the dynamical structure of these systems for nearly ninety years (e.g., Hubble 1936; de Vaucouleurs 1959; Sandage 1961; van den Bergh 1976; Nair & Abraham 2010; Baillard et al. 2011). The division between early-type and late-type systems corresponds, for example, to a wide range of parameters from mass and luminosity, to environment, colour, and star formation history (e.g., Kormendy 1977; Dressler 1980; Strateva et al. 2001; Blanton et al. 2003; Kauffmann et al. 2003; Nakamura et al. 2003; Shen et al. 2003; Peng et al. 2010); while detailed observations of morphological features such as bars and bulges provide information about the history of their host systems (e.g., review by Kormendy & Kennicutt 2004; Elmegreen et al. 2008; Sheth et al. 2008; Masters et al. 2011; Simmons et al. 2014). Modern studies of morphology divide systems into broad classes (e.g., Conselice 2006; Lintott et al. 2008; Kartaltepe et al. 2015; Peth et al. 2016), but a wealth of information can be gained from identifying new and often rare classes, such as low redshift clumpy galaxies (e.g., Elmegreen et al. 2013), polar-ring galaxies (e.g., Whitmore et al. 1990), and the green peas (Cardamone et al. 2009).

While the Galaxy Zoo project has provided a solution that scales visual classification for current surveys by harnessing the combined power of thousands of volunteers (Lintott et al. 2008, 2011; Willett et al. 2013, 2017; Simmons et al. 2017), producing a prolific amount

of scientific output (e.g., Land et al. 2008; Bamford et al. 2009; Darg et al. 2010; Schawinski et al. 2014; Galloway et al. 2015; Smethurst et al. 2016); upcoming surveys such as *LSST* and *Euclid* will require a different approach, imaging more than a billion new galaxies (LSST Science Collaboration et al. 2009; Laureijs et al. 2011). If detailed morphologies can be extracted for just 0.1% of this imaging, we will have millions of images to contend with. A project of this magnitude would take more than sixty years to classify at Galaxy Zoo's current rate and configuration. Standard visual morphology methods will thus be unable to cope with the scale of data.

Another approach has been the automated extraction of morphologies with the development of parametric (Sersic 1968; Odewahn et al. 2002; Peng et al. 2002), and non-parametric (Abraham et al. 1994; Conselice 2003; Abraham et al. 2003; Lotz et al. 2004; Freeman et al. 2013) structural indicators. While these scale well to large samples (e.g., Simard et al. 2011; Griffith et al. 2012; Casteels et al. 2014; Holwerda et al. 2014; Meert et al. 2016), they often fail to capture detailed structure and can provide only statistical morphologies with large uncertainties (e.g., Abraham et al. 1996; Bershady et al. 2000).

Machine learning techniques are becoming increasingly popular for classification and image processing tasks. Another automated approach, these generally work by defining a set of features that describe the morphology in an N-dimensional space. The location in this morphology space defines a morphological type for each galaxy. Learning the morphology space can be achieved through algorithms such as Support Vector Machines (Huertas-Company et al. 2008) or Principal Component Analysis (Watanabe et al. 1985; Scarlata et al. 2007). Another approach is through deep learning, a machine learning technique that attempts to model high level abstractions. Algorithms like convolutional and artificial neural networks (CNNs, ANNs) have been used for galaxy morphology classification with impressive accuracy (Ball et al. 2004; Banerji et al. 2010; Dieleman et al. 2015; Huertas-Company et al. 2015). A drawback to all machine learning classification techniques is the need for standardized training data, with more complex algorithms requiring more data. Furthermore, these data must be consistent for each survey: differences in resolution and depth can be implicitly learned by the algorithm making their application to disparate surveys challenging.

In this work we present a system that preserves the best features of both visual and automatic classifications, developing for the first time a framework that brings both human and machine intelligence to the task of galaxy morphology to handle the scale and scope of next generation data. We demonstrate the effectiveness of such a system through a re-analysis of visual galaxy morphology classifications collected during the Galaxy Zoo 2 project, and combine these with a Random Forest machine learning algorithm that trains on a suite of non-parametric morphology indicators widely used for automated morphologies. (Added: The primary goal of this paper is to generalize how such a system would work in the context of upcoming surveys like LSST and Euclid.) (Replaced: ~~In this paper~~ replaced with: As a proof of concept,) we focus on the first question of the Galaxy Zoo decision tree (Added: , however we briefly discuss how such a method can be implemented for more complex tasks). We demonstrate that our (Replaced: ~~method~~ replaced with: current implementation) provides a factor of 11.4 increase in the rate of galaxy morphology classification while maintaining at least 93.5% classification accuracy as compared to Galaxy Zoo 2 published data. We first present an overview of our framework, which also serves as a blueprint for this paper.

2. GALAXY ZOO EXPRESS OVERVIEW

The Galaxy Zoo Express (GZX) framework combines human and machine to increase morphological classification efficiency, both in terms of the classification rate and required human effort. Figure 1 presents a schematic of GZX including section numbers as a shortcut for the reader. We note that transparent portions of the schematic represent areas of future work which we explore in Section 7. Any system combining human and machine classifications will have a set of generic features: a group of human classifiers, at least one machine classifier, and a decision engine which determines how these classifications should be combined.

In this work we demonstrate our system through a re-analysis of Galaxy Zoo 2 (GZ2) classifications. This allows us to create simulations of human classifiers (described in Section 3). These classifications are used most effectively when processed with SWAP, a Bayesian code described in Section 4, first developed for the Space Warps gravitational lens discovery project (Marshall et al. 2016). These subjects provide the machine's training sample.

In Section 5, we incorporate a machine classifier. We have developed a Random Forest algorithm that trains on measured morphology indicators such as Concentration, Asymmetry, Gini coefficient and M_{20}, well-suited for the top-level question of the GZ2 decision tree, discussed below. After a sufficient number of subjects have been classified by humans, the machine is trained and its performance assessed through cross-validation. This procedure is repeated nightly and the machine's performance increases with the size of the training sample, albeit with a performance limit. Once the machine reaches

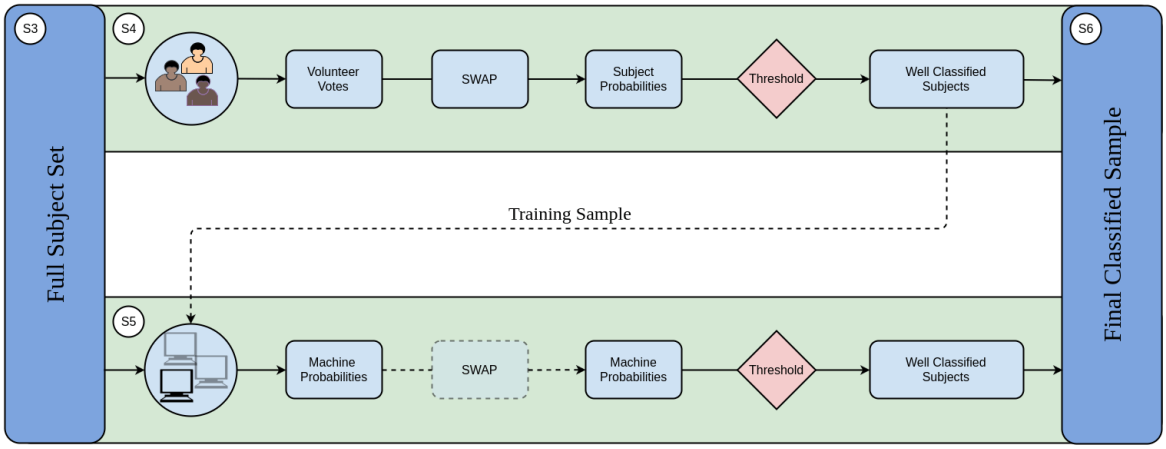

Figure 1. Schematic of our hybrid system. Humans provide classifications of galaxy images via a web interface. We simulate this with the Galaxy Zoo 2 classification data described in Section 3. Human classifications are processed with an algorithm described in Section 4. Subjects that pass a set of thresholds are considered human-retired (fully classified) and provide the training sample for the machine classifier as described in Section 5. The trained machine is applied to all subjects not yet retired. Those that pass an analogous set of machine-specific thresholds are considered machine-retired. The rest remain in the system to be classified by either human or machine. This procedure is repeated nightly. Our results are reported in Section 6.

an acceptable level of performance it is applied to the remaining galaxy sample.

Even with this simple description, one can see that the classification process will progress in three phases. First, the machine will not yet have reached an acceptable level of performance; only humans contribute to subject classification. Second, the machine's performance will improve; both humans and machine will be responsible for classification. Finally, machine performance will slow; remaining images will likely need to be classified by humans. These results are explored in Section 6. This blueprint allows even modest machine learning routines to make significant contributions alongside human classifiers and removes the need for ever-increasing performance in machine classification.

3. GALAXY ZOO 2 CLASSIFICATION DATA

Our simulations utilize original classifications made by volunteers during the GZ2 project. These data[1] are described in detail in Willett et al. (2013), though we provide a brief overview here. The GZ2 subject sample consists of 285,962 galaxies identified as the brightest 25% (r-band magnitude < 17) residing in the SDSS North Galactic Cap region from Data Release 7 and included subjects with both spectroscopic and photometric redshifts out to $z < 0.25$. Subjects were shown as colour composite images via a web-based interface[2] wherein volunteers answered a series of questions pertaining to the morphology of the subject. With the exception of the first question, subsequent queries were dependent on volunteer responses from the previous task creating a complex decision tree[3]. Using GZ2 nomenclature, a *classification* is the total amount of information about a subject obtained by completing all tasks in the decision tree. A subject is *retired* after it has achieved a sufficient number of classifications.

For our current analysis, we choose the first task in the tree: "Is the galaxy simply smooth and rounded, with no sign of a disk?" to which possible responses include "smooth", "features or disk", or "star or artifact". This choice serves two purposes: 1) this is one of only two questions in the GZ2 decision tree that is asked of every subject thus maximizing the amount of data we have to work with, and 2) our analysis assumes a binary task and this question is simple enough to cast as such. Specifically, we combine "star or artifact" responses with "features or disk" responses.

We assign each subject a descriptive label in order to validate our classification output with GZ2. GZ2 classifications are composed of volunteer vote fractions for each response to every task in the decision tree, denoted as $f_{\rm response}$. They are derived from the fraction of volunteers who voted for a particular response and are thus approximately continuous. A common technique is to place a threshold on these vote fractions to select samples with an emphasis on purity or completeness, depending on the science case. For our current analysis we choose a threshold of 0.5, that is, if $f_{\rm featured}+f_{\rm artifact} > f_{\rm smooth}$, the galaxy is labelled 'Featured', otherwise it is labelled 'Not'. We note that only 512 subjects in the

[1] data.galaxyzoo.org

[2] www.galaxyzoo.org

[3] A visualization of this decision tree can be found at https://data.galaxyzoo.org/gz_trees/gz_trees.html

GZ2 catalogue have a majority $f_{\rm artifact}$, contributing less than half a percent contamination when combining the "star or artifact" with "features or disk" responses.

The GZ2 catalogue publishes three types of vote fractions for each subject: raw, weighted, and debiased. Debiased vote fractions are calculated to correct for redshift bias, a task that GZX does not perform. The weighted vote fractions account for inconsistent volunteers. The SWAP algorithm (described below) also has a mechanism to weight volunteer votes, however, the two methods are in stark contrast. For consistency, we derive labels from the raw vote fractions ($\rm GZ2_{raw}$) that have received no post-processing whatsoever. In total, the data consist of over 16 million classifications from 83,943 individual volunteers.

(Added: The labels we compute from GZ2 vote fractions are used solely to validate our classification method. We thus consider these labels as "ground truth" though, of course, this is subjective. Varying the threshold on which we define 'Featured' from 'Not', or choosing a different type of GZ2 vote fraction will yield slightly different results for the quality metrics we compute throughout this paper. However, we envision this method being applied to never-before-classified image sets and in such a case, no "ground truth" classification would yet exist. Rather than re-classifying GZ2 subjects, we instead endeavour to provide a general sense of the quality of our classifier in the context of the well-known quality of the GZ2 catalogue. In Appendix C we show how different choices of our descriptive GZ2 labels changes the perceived quality of our classification system and demonstrate that our method yields robust galaxy classifications.)

4. EFFICIENCY THROUGH INTELLIGENT HUMAN-VOTE AGGREGATION

Galaxy Zoo 2 had a brute-force subject retirement rule whereby each galaxy was to receive approximately forty independent classifications. Once the project reached completion, inconsistent volunteers were down-weighted (Willett et al. 2013), a process that does not make efficient use of those who are exceptionally skilled. To intelligently manage subject retirement and increase classification efficiency, we adapt an algorithm from the Zooniverse project Space Warps (Marshall et al. 2016), which searched for and discovered several gravitational lens candidates in the CFHT Legacy Survey (More et al. 2016). Dubbed SWAP (Space Warps Analysis Pipeline), this algorithm computed the probability that an image contained a gravitational lens given volunteers' classifications and experience after being shown a training sample consisting of simulated lensing events. We provide a brief overview here.

The algorithm assigns each volunteer an *agent* which interprets that volunteer's classifications. Each agent assigns a 2×2 confusion matrix to their volunteer which encodes that volunteer's probability of correctly identify feature A given that the subject exhibits feature A; and the probability of correctly identifying the absence of feature A (denoted N) given that the subject does not exhibit that feature. The agent updates these probabilities by estimating them as

$$P(\text{“}X\text{”}|X,\mathbf{d}) \approx \frac{\mathcal{N}_{\text{“}X\text{”}}}{\mathcal{N}_X} \tag{1}$$

where X is the true classification of the subject and "X" is the classification made by the volunteer upon viewing the subject. Thus $\mathcal{N}_{\text{“}X\text{”}}$ is the number of classifications the volunteer labelled as type X, $\mathcal{N}_X$ is the number of subjects the volunteer has seen that were actually of type X, and $\mathbf{d}$ represents the history of the volunteer, i.e., all subjects they have seen. Therefore the confusion matrix for a single volunteer goes as

$$\mathcal{M} = \begin{bmatrix} P(\text{“}A\text{”}|N,\mathbf{d}) & P(\text{“}A\text{”}|A,\mathbf{d}) \\ P(\text{“}N\text{”}|N,\mathbf{d}) & P(\text{“}N\text{”}|A,\mathbf{d}) \end{bmatrix} \tag{2}$$

where probabilities are normalised such that $P(\text{“}A\text{”}|A) = 1 - P(\text{“}N\text{”}|A)$.

Each subject is assigned a prior probability that it exhibits feature A: $P(A) = p_0$. When a volunteer makes a classification, Bayes' theorem is used to compute how that subject's prior probability should be updated into a posterior using elements of the agent's confusion matrix. As the project progresses, each subject's posterior probability is updated after every volunteer classification, nudged higher or lower depending on volunteer input. Upper and lower probability thresholds can be set such that when a subject's posterior crosses the upper threshold it is highly likely to exhibit feature A; if it crosses the lower threshold it is highly likely that feature A is absent. Subjects whose posteriors cross either of these thresholds are considered retired.

4.1. *Gold-standard sample*

A key feature of the original Space Warps project was the training of individual volunteers through the use of simulated images. These were interspersed with real imaging and were predominantly shown at the beginning of a volunteer's association with the project, allowing that volunteer's agent time to update before classifying real data. Volunteers were provided feedback in the form of a pop-up comment after classifying a training image. GZ2 did not train volunteers in such a way, presenting a challenge when applying SWAP to GZ2 classifications. Though we cannot retroactively train GZ2 volunteers,

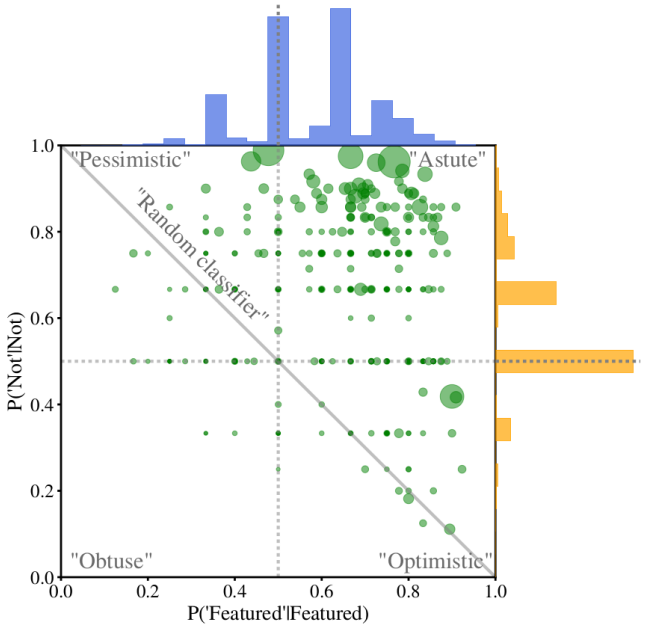

Figure 2. Confusion matrices for 1000 randomly selected GZ2 volunteers after fiducial SWAP assessment. Circle size is proportional to the number of gold standard subjects each volunteer classified. The histograms on top and right represent the distribution of each component of the confusion matrix for all volunteers. A quarter of GZ2 volunteers are "Astute"; they correctly identify both 'Featured' and 'Not' subjects more than 50% of the time. The peaks at 0.5 in both distributions are due primarily to volunteers who see only one training image: only half of their confusion matrix is updated.

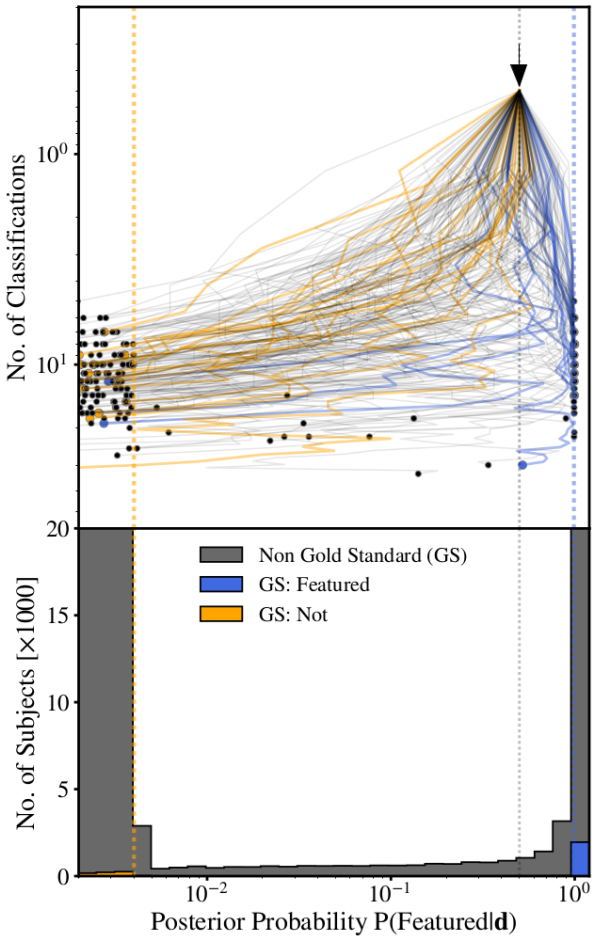

Figure 3. Posterior probabilities for GZ2 subjects. The top panel depicts the probability trajectories of 200 randomly selected GZ2 subjects. All subjects begin with a prior of 0.5 denoted by the arrow. Each subject's probability is nudged back and forth with each volunteer classification. From left to right the dotted vertical lines show the 'Not' threshold, prior probability, and 'Featured' threshold. Different colours denote different types of subjects. The bottom panel shows the distribution in probability for all GZ2 subjects by the end of our simulation, where the y axis is truncated to show detail.

we develop a gold standard sample and arrange the order of gold standard classifications in order to mimic the Space Warps system.

We create a gold standard sample by selecting 3496 SDSS galaxies representative of the relative abundance of T-Types, a numerical index of a galaxy's stage along the Hubble sequence, at $z \sim 0$ by considering galaxies that overlap with the Nair & Abraham (2010) catalogue, a collection of ∼14K galaxies classified by eye into T-Types. We generate expert labels for these galaxies that are consistent with the labels we defined for GZ2 classifications. These are obtained through the Zooniverse platform[4] from 15 professional astronomers, including members of the Galaxy Zoo science team. The question posed was identical to the original top-level GZ2 question and at least five experts classified each galaxy. Votes are aggregated and a simple majority provides an expert label for each subject. This ensures that our expert labels are defined in exactly the same manner as the labels we assign the rest of the GZ2 sample. Our final dataset consists of the GZ2 classifications made by those volunteers who classify at least one of these gold standard subjects. We thus retain for our simulation 12,686,170 classifications from 30,894 unique volunteers. When running SWAP, classifications of gold standard subjects are always processed first.

[4] The Project Builder template facility can be found at http://www.zooniverse.org/lab.

4.2. *Fiducial SWAP simulation*

Before we run a simulation, a number of SWAP parameters must be chosen: the initial confusion matrix for each volunteer's agent, ($P(\text{"}F\text{"}|F)$, $P(\text{"}N\text{"}|N)$); the subject prior probability, p_0; and the retirement thresholds, t_F and t_N. For our fiducial simulation, we initialize all confusion matrices at (0.5, 0.5), and set the subject prior probability, $p_0 = 0.5$. We set the 'Fea-

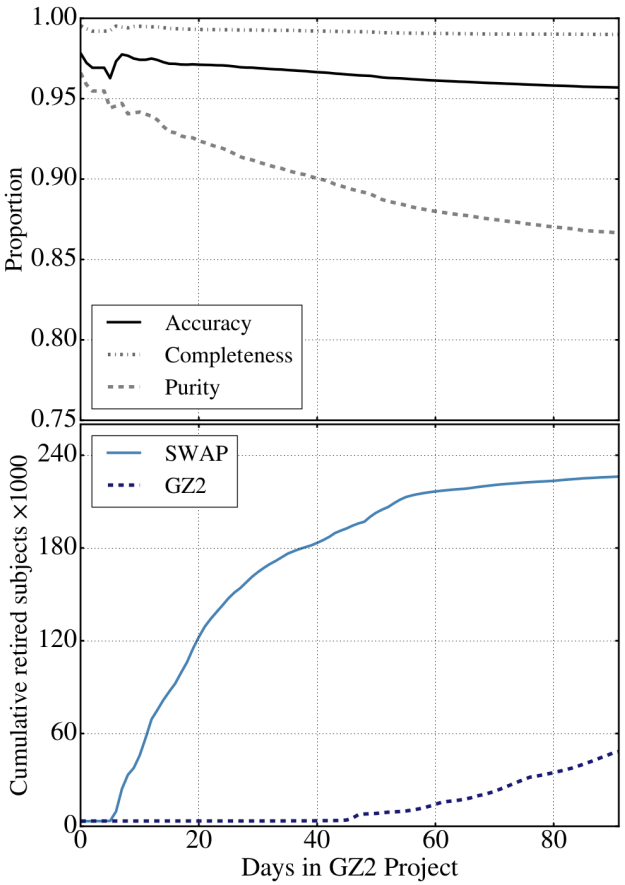

Figure 4. Fiducial SWAP simulation demonstrates a factor of 4-5 increase in the rate of subject retirement as a function of GZ2 project time (bottom panel, light blue) compared with the original GZ2 project (dashed dark blue). After 92 days, SWAP retires over 225K subjects, while GZ2 retires ∼48K. The top panel displays the quality metrics (greys). These are calculated by comparing SWAP-assigned labels to $GZ2_{raw}$ labels (Section 3) for the subject sample retired by that day of the simulation. Thus, on the final day, SWAP retires 226,124 subjects with 95.7% accuracy, and with completeness and purity of 'Featured' subjects at 99% and 86.7% respectively. The decrease in purity as a function of time is due, in part, to the fact that more difficult to classify subjects are retired later in the simulation.

tured' threshold, t_F, i.e., the minimum probability for a subject to be retired as 'Featured', to 0.99. Similarly, we set the 'Not' threshold, $t_N = 0.004$. In Appendix A we show that varying these parameters has only a small affect on the SWAP output. To simulate a live project, we run SWAP on a time step of $\Delta t = 1$ day, during which SWAP processes all volunteer classifications with timestamps within that range. This is performed for three months worth of GZ2 classification data.

Figure 2 (adapted from Figure 4 of Marshall et al. 2016) demonstrates the volunteer assessment we achieve, and shows confusion matrices for 1000 randomly selected volunteers. The circle size is proportional to the number of gold standard subjects each volunteer classified. The histograms represent the distribution of each component of the confusion matrix for all volunteers. Nearly 25% of volunteers are considered "Astute" indicating they correctly identify both 'Featured' and 'Not' subjects more than 50% of the time. Furthermore, as long as a volunteer's confusion matrix is different from random, they provide useful information to the project. The spikes at 0.5 in the histograms are due to volunteers who see only one gold standard subject (i.e., 'Featured'), leaving their probability in the other ('Not') unchanged. Additionally, 4% of volunteers have a confusion matrix of (0.5, 0.5) indicating these volunteers classified two gold standard subjects of the same type, one correctly and one incorrectly.

Figure 3 (adapted from Figure 5 of Marshall et al. 2016) demonstrates how subject posterior probabilities are updated with each classification. The arrow in the top panel denotes the prior probability, $p_0 = 0.5$. With each classification, that prior is updated into a posterior probability creating a trajectory through probability space for each subject. The blue and orange lines show the trajectories of a random sample of 'Featured' and 'Not' subjects from our gold standard sample, while the black lines show the trajectories of a random sample of GZ2 subjects that were not part of the gold standard sample. The similarly coloured vertical dashed lines correspond to the retirement thresholds, t_F and t_N. The lower panel shows the full distribution of GZ2 subject posteriors at the end of our simulation, where the y-axis has been truncated to show detail. An overwhelming majority of subjects cross one of these retirement thresholds.

Our goal is to increase the efficiency of galaxy classification. We therefore use as a metric the cumulative number of retired subjects as a function of the original GZ2 project time. We define a subject as GZ2-retired once it achieves at least 30 volunteer votes, encompassing 98.6% of GZ2 subjects. In contrast, a subject is considered SWAP-retired once its posterior probability crosses either of the retirement thresholds defined above.

However, it is important not to prioritize efficiency at the expense of quality. (Replaced: ~~We thus also consider the metrics of accuracy, purity and completeness as a function of GZ2 project time These are defined as follows: accuracy is the number of correctly identified subjects divided by the total number retired; completeness is the number of correctly identified 'Featured' subjects divided by the number of actual 'Featured' retired; and purity is the number of correctly identified 'Featured' subjects divided by the number of subjects retired as 'Featured'.~~ replaced with: Because we have a binary classification we can construct a confusion matrix from which we can compute the quality metrics of accuracy, completeness and purity as a

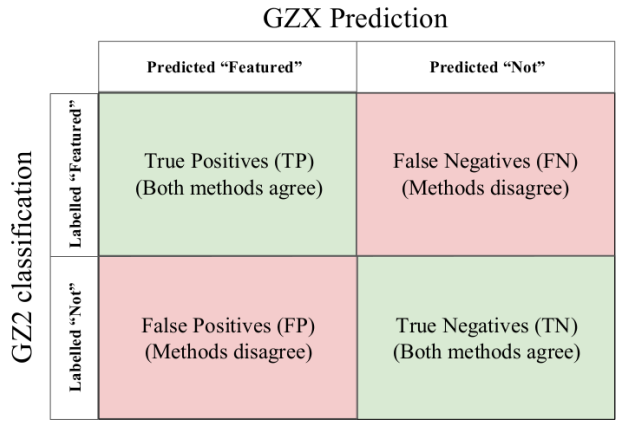

Figure 5. Confusion matrix for comparing GZ2 classifications to our method. True Positives and True Negatives indicate that both our method and GZ2 predict a given subject as being 'Featured' or 'Not', respectively. When our method predicts a subject as being 'Featured' ('Not') that GZ2 classified as 'Not' ('Featured'), the result is a False Negative (Positive). This allows us to easily compute quality metrics like accuracy, completeness, and purity with respect to GZ2 labels as shown in Equations 3.

function of GZ2 project time by comparing our predicted labels to the $GZ2_{raw}$ labels. Figure 5 graphically depicts the elements of this confusion matrix. From this we compute:)

$$\begin{aligned} \text{accuracy} &= \frac{TP+TN}{TP+FP+TN+FN} \\ \text{completeness} &= \frac{TP}{TP+FN} \\ \text{purity} &= \frac{TP}{TP+FP} \end{aligned} \tag{3}$$

(Replaced: ~~Thus, a complete sample has no false negatives whereas a pure sample has no false positives.~~ replaced with: Thus, a complete sample recovers *all* subjects labelled 'Featured' by GZ2, whereas a pure sample recovers *only* subjects labelled 'Featured' by GZ2.) For example, by Day 20, SWAP retires 120K subjects with 96% accuracy, 99.7% completeness, and 92% purity.

Figure 4 and Table 1 detail the results of our fiducial SWAP simulation compared to the original GZ2 project. The bottom panel shows the cumulative number of retired subjects as a function of GZ2 project time. By the end of our simulation, GZ2 (dashed dark blue) retires ~50K subjects while SWAP (solid light blue) retires 226,124 subjects. We thus classify 80% of the entire GZ2 sample in three months. Here we consider the number of classifications logged each day and not the length of time spent on a single classification. Under the assumption that collapsing the GZ2 decision tree to a single question would not have decreased the number of classifications collected each day during the GZ2 project, processing

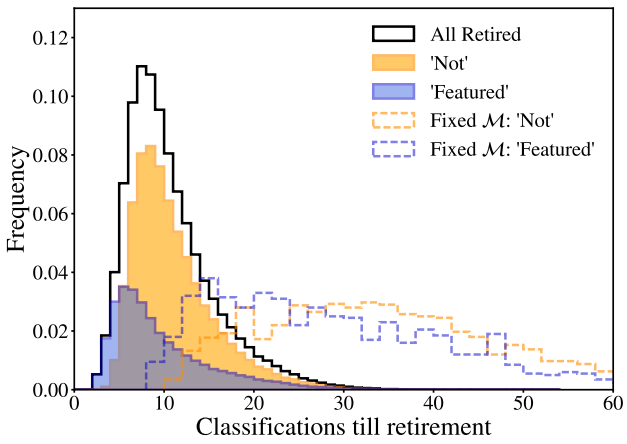

Figure 6. SWAP's volunteer-weighting mechanism provides a factor of three reduction in the human effort required to retire GZ2 subjects. The filled histograms show the number of volunteer classifications per subject achieved from our simulation of GZ2 data broken down by class label, where the solid black line is the total. The dashed histograms are results from our toy model in which we simulate volunteers with fixed confusion matrices, effectively disengaging SWAP's volunteer-weighting mechanism. These broad distributions require ~3 times more classifications per subject to reach the same retirement thresholds.

volunteer classifications through SWAP presents a striking increase in classification efficiency. The top panel of Figure 4 demonstrates the quality of those classifications as a function of time and establishes that our full SWAP-retired sample is 95.7% accurate, 99% complete, and 86.7% pure. We discuss these small discrepancies in Section 4.5.

(Added:

4.3. *What about the GZ2 decision tree?*

) Our results thus far have compared the increased efficiency achieved for the top-level question of the GZ2 decision tree. However, it can be argued that this comparison is unfair since GZ2 needed additional classifications in order to populate the lower levels of the decision tree. In order to put this comparison on more equal footing, we explore what would have happened in GZ2 had instead asked only the first question, presumably requiring fewer classifications overall to achieve the same results.

To do this we compute the raw vote fractions (f_{featured}, f_{smooth}, and f_{artifact}) for every subject in the GZ2 sample from the first N classifications for $N \in [15, 20, 25, 30, 35]$. From these vote fractions we then compute the binary descriptive label as described in Section 3. Comparing these new labels to those

4.4. *Reducing human effort*

The SWAP algorithm effectively weights volunteers according to how many gold standard subjects they correctly identify. This mechanism provides a reduction in the amount of human effort required to perform this classification task. To see this, we consider a toy model wherein we simulate volunteers with fixed confusion matrices. We simulate 1000 'Featured' subjects and 1000 'Not' subjects each with prior, $p_0 = 0.5$. We simulate 100 volunteer agents all with the same fixed confusion matrix of (0.63, 0.65), where these values are computed as the average $P(\text{“}F\text{”}|F)$ and $P(\text{“}N\text{”}|N)$ from our assessment of real volunteers, excluding the spikes at 0.5. We generate volunteer classifications based on this confusion matrix (i.e., volunteers will correctly identify 'Featured' subjects 63% of the time) and update the subject's posterior probability with each classification. We track how many classifications are required for each subject's posterior to cross either the 'Featured' or 'Not' retirement thresholds.

The results are presented in Figure 6. The filled blue and orange histograms show the number of volunteer classifications per subject achieved from our simulation of GZ2 data, where volunteer agent confusion matrices are those from Figure 2. The dashed blue and orange distributions are the results from our toy model. When SWAP accounts for volunteer skill, most subjects are retired with between 6 and 15 votes, with a median of 9 votes. In contrast, when every volunteer is given equal weighting, subjects require 16 to 45 votes with a median of 30 votes before crossing one of the retirement thresholds. Thus the volunteer weighting scheme embedded in SWAP can reduce the amount of human effort required to retire subjects by a factor of three

This reduction will be, in part, a function of the number of gold standard subjects each volunteer sees. Our gold standard sample was chosen to be representative of morphology rather than evenly distributed among GZ2 volunteers. We thus find that half of our volunteers classify only one or two gold standard subjects. That we achieve a factor of three reduction when only half of our volunteer pool has seen ≥ 2 gold standard subjects suggests that an additional reduction of human effort is possible with more extensive volunteer training.

4.5. *Disagreements between SWAP and GZ2*

Galaxy Zoo's strength comes from the consensus of dozens of volunteers voting on each subject. Processing votes with SWAP reduces the number of classifications to reach consensus. Though we typically recover the GZ2_{raw} label, SWAP disagrees about 5% of the time. We thus examine the false positives (subjects SWAP labels as 'Featured' but GZ2_{raw} labels as 'Not') and false negatives (subjects SWAP labels as 'Not' but GZ2_{raw} labels as 'Featured'). We explore these subjects in redshift, magnitude, physical size, and concentration. We find no correlation with any of these variables, suggesting that, at least for this galaxy sample, the reliability of morphology depends on factors that are not captured by these coarse measurements. This is perhaps unsurprising since GZ2 subjects were selected from the larger GZ1 sample to be the brightest, largest and nearest galaxies: precisely those subjects most accessible for visual classification.

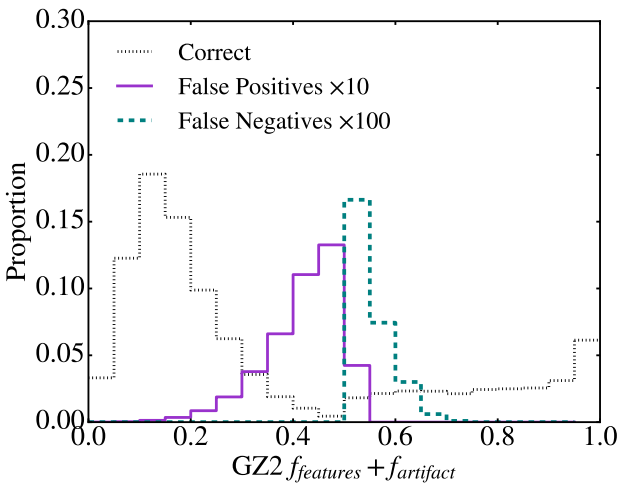

Figure 7. Distribution of GZ2 $f_{\text{featured}}+f_{\text{artifact}}$ vote fractions for subjects correctly identified by SWAP (dotted grey), along with those identified as false positives (solid purple), and false negatives (dashed teal). The false positives and false negatives are scaled by factors of 10 and 100 respectively for easier comparison. From Section 3, subjects with values > 0.5 are defined as 'Featured', however, the teal distribution indica tes that SWAP labels them as 'Not'. This is not a flaw of SWAP: 68.9% of incorrectly identified subjects have $0.4 \leq f_{\text{featured}}+f_{\text{artifact}} \leq 0.6$ suggesting that GZ2_{raw} labels are simply too uncertain. The overlap between the false positives and negatives is due to subjects that are exactly 50-50; by default these are labelled 'Not'.

Instead we consider the stochastic nature of GZ2 vote fractions, which can be estimated as binomial. Let success be a response of "smooth" and failure be any other response. The 68% confidence interval on a subject with $f_{\text{smooth}} = 0.5$ is then $(0.42, 0.57)$ assuming 40 classifications, each with a probability of 0.5. Figure 7 shows the distribution of $f_{\text{featured}}+f_{\text{artifact}}$ for the false positives (solid purple), and the false negatives (dashed teal) compared to the subjects where SWAP and GZ2 agree (dotted grey). Recall that if this value is greater than 0.5, the subject is labelled 'Featured'. The majority of disagreements between SWAP and GZ2 are for subjects that have $0.4 < f_{\text{featured}}+f_{\text{artifact}} < 0.6$. It is thus unsurprising that SWAP and GZ2 disagree most within the approximate confidence interval of our selected GZ2 threshold. We note that the distribution overlap between false positives and false negatives is due

to subjects that do not have a majority; these are labelled 'Not' by default.

Two other effects contribute to the disagreement between SWAP and GZ2. First, as the number of classifications used to retire a galaxy decreases, the likelihood of misclassification by random chance increases. Second, disagreement arises due to expert-level volunteers whose confusion matrices are close to 1.0. These volunteers are essentially more strongly weighted, allowing that subject's posterior to cross a retirement threshold in as few as two classifications. In rare cases, despite training, some expert-level volunteers get it wrong compared to the gold-standard labels. These issues can be mitigated by requiring each subject reach a minimum number of classifications in addition to its probability crossing a retirement threshold, thus combining the best qualities of GZ2 and SWAP.

4.6. *Summary*

We demonstrate a factor of four or more increase in classification efficiency while maintaining 95% accuracy, nearly perfect completeness of 'Featured' subjects, and with a purity that can be controlled by careful selection of input parameters to be better than 90% (see Appendix A). Exploring those subjects wherein SWAP and GZ2 disagree, we conclude that the majority of this disagreement stems from the stochastic nature of $\mathrm{GZ2_{raw}}$ labels. We now turn our focus towards incorporating a machine classifier utilizing these SWAP-retired subjects as a training sample.

5. EFFICIENCY THROUGH INCORPORATION OF MACHINE CLASSIFIERS

We construct the full Galaxy Zoo Express by incorporating supervised learning, the machine learning task of inference from labelled training data. The training data consist of a set of training examples, and must include an input feature vector and a desired output label. Generally speaking, a supervised learning algorithm analyses the training data and produces a function that can be mapped to new examples. A properly optimized algorithm will correctly determine class labels for unseen data. By processing human classifications through SWAP, we obtain a set of binary labels by which we can train a machine classifier. We briefly outline the technical details of our machine below, turning towards the decision engine we develop in Section 5.4.

5.1. *Random Forests*

We use a Random Forest (RF) algorithm (Breiman 2001), an ensemble classifier that operates by bootstrapping the training data and constructing a multitude of individual decision tree algorithms, one for each subsample. An individual decision tree works by deciding which of the input features best separates the classes. It does this by performing splits on the values of the input feature that minimize the classification error. These feature splits proceed recursively. Decision trees alone are prone to over-fitting, precluding them from generalizing well to new data. Random Forests mitigate this effect by combining the output labels from a multitude of decision trees. Specifically, we use the `RandomForestClassifier` from the Python module `scikit-learn` (Pedregosa et al. 2011).

5.2. *Grid Search and Cross-validation*

Of fundamental importance is the task of choosing an algorithm's hyperparameters, values which determine how the machine learns. For a RF, key quantities include the maximum depth of individual trees (`max_depth`), the number of trees in the forest (`n_estimators`), and the number of features to consider when looking for the best split (`max_features`). The goal is to determine which values will optimize the machine's performance and thus these values cannot be chosen *a priori*. We perform a grid search with k-fold cross-validation whereby the training sample is split into k subsamples. One subsample is withheld to estimate the machine's performance while the remaining data are used to train the machine. This is performed k times and the average performance value is recorded. The entire process is repeated for every combination of the hyperparameters in the grid space and values that optimize the output are chosen. In this work we let $k = 10$, however, we leave this as an adjustable input parameter. In the interest of computational speed, we set `n_estimators` $= 30$ and perform the grid search for `max_depth` over the range $[5, 16]$, and `max_features` over the range $[\sqrt{D}, D]$, where D is the number of features in the feature vector, described below.

5.3. *Feature Representation and Pre-Processing*

The feature vector on which the machine learns is composed of D individual numeric quantities associated with the subject that the machine uses to discern that subject from others in the training sample. To segregate 'Featured' from 'Not', we draw on ZEST (Scarlata et al. 2007) and compute concentration, asymmetry, Gini coefficient, and $\mathrm{M_{20}}$, the second-order moment of light for the brightest 20% of galaxy pixels as measured from SDSS DR12 i-band imaging (see Appendix B). Coupled with SExtractor's measurement of ellipticity (Bertin & Arnouts 1996), we provide the machine with a $D = 5$ dimensional morphology parameter space. These non-parametric diagnostics have long been used to distinguish between early- and late-type galaxies in an automated fashion (e.g., Abraham et al. 1996; Bershady

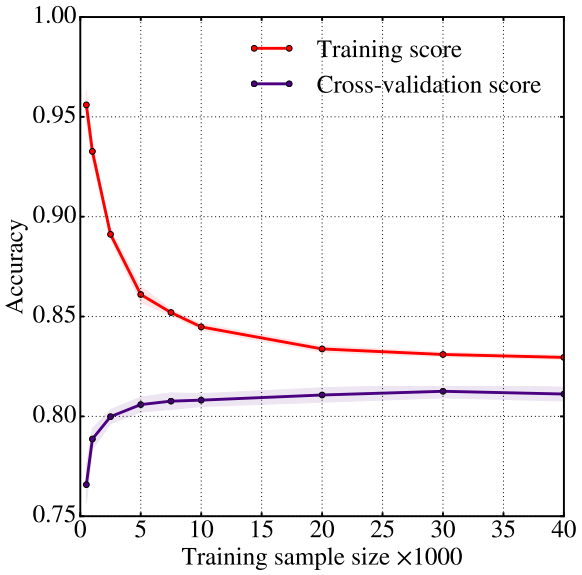

Figure 8. Learning curve for a Random Forest with fixed hyperparameters. These curves show the mean accuracy computed during cross-validation and on the training sample, where the shaded regions denote the standard deviation. When the training sample size is small, the machine accurately identifies its own training sample but is unable to generalize to unseen data as evidenced by a low cross-validation score. As the training sample size increases, the cross-validation score increases. This behaviour plateaus indicating that larger training samples provide little in additional performance.

et al. 2000; Conselice et al. 2000; Abraham et al. 2003; Conselice 2003; Lotz et al. 2004; Snyder et al. 2015). Because the RF algorithm handles a variety of input formats, the only pre-processing step we perform is the removal of poorly-measured morphological indicators, i.e. catastrophic failures.

5.4. *Decision Engine*

A number of decisions must be addressed before attempting to train the machine. In particular, which subjects should be designated as the training sample? When should the machine attempt its first training session? When has the machine's performance been optimized such that it will successfully generalize to unseen subjects? The field of machine learning provides few hard rules for answering these questions, only guidelines and best practices. Here we briefly discuss our approach for the development of our decision engine.

As discussed in detail in Section 4, SWAP yields a probability that a subject exhibits the feature of interest. While some machine algorithms can accept continuous input labels, the RF requires distinct classes. We thus use only those subjects which have crossed either of the retirement thresholds. Though we find that SWAP consistently retires 35-40% 'Featured' subjects on any given day of the simulation, a balanced ratio of 'Featured' to 'Not' isn't guaranteed. Highly unbalanced training samples should be resampled to correct the imbalance; however, as we exhibit only a mild lopsidedness, we allow the machine to train on all SWAP-retired subjects.

SWAP retires a few hundred subjects during the first days of the simulation. In principle, a machine can be trained with such a small sample, but will be unable to generalize to unseen data. We estimate a minimum number of training samples and the machine's ability to generalize by considering a learning curve, an illustration of a machine's performance with increasing sample size for fixed hyperparameters. Figure 8 demonstrates such a curve wherein we plot the accuracy from both the 10-fold cross-validation, and the trained machine applied to its own training sample for a random sample of GZ2 subjects required to be balanced between 'Featured' and 'Not'. We fix the RF's hyperparameters as follows: `max_depth` = 8, `n_estimators` = 30, and `max_features` = 2. When the sample size is small, the cross-validation score is low and the training score is high, a clear sign of over-fitting. However, as the training sample size increases, the cross-validation score increases and eventually plateaus, indicating that larger training sets will yield little additional gain.

We estimate this plateau begins when the training sample reaches 10,000 subjects and require SWAP retire at least this many before the machine attempts its first training. We estimate the machine has trained sufficiently if the cross-validation score fluctuates by less than 1% for three consecutive nights of training to ensure we have reached the plateau. This requires that we record the machine's training performance each night, including how well it scores on the training sample, the cross-validation score, and the best hyperparameters.

5.5. *The Machine Shop*

We can now describe a full GZX simulation, which begins with human classifications processed through SWAP for several days. Once at least 10K subjects have been retired, their feature vectors are passed to the machine for its inaugural training. A suite of performance metrics are recorded by a machine agent, similar in construction to SWAP's agents. This agent determines when the machine has trained sufficiently by assessing the variation in performance metrics for all previous nights of training. Once the machine has been optimized, the agent introduces it to the test sample consisting of any subject that has not yet reached retirement through SWAP and is not part of the gold standard sample.

Analogous to SWAP, we generate a retirement rule for machine-classified subjects. In addition to the class pre-

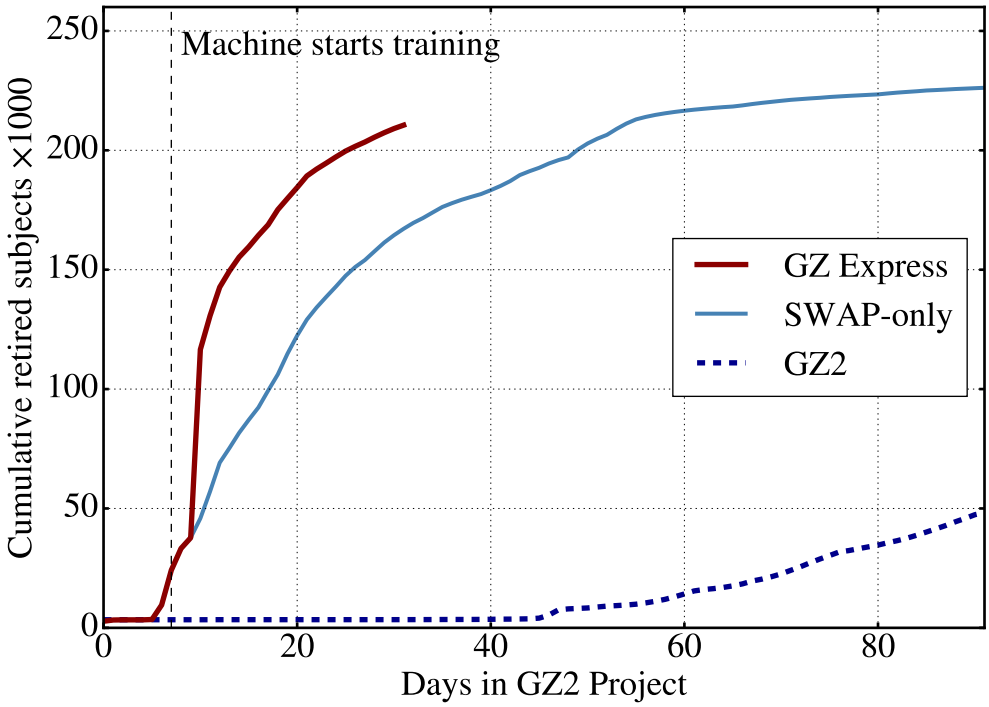

Figure 9. By incorporating a machine classifier, GZX (red) increases the classification rate by an order of magnitude compared to GZ2 (dashed dark blue) and out-performs the SWAP-only run (light blue), retiring more than 200K subjects in just 27 days of GZ2 project time. The dashed black line marks the first night the machine trains. After several additional nights of training, it is deemed optimized and allowed to retire subjects. Both humans and machine then contribute to retirement. We end the simulation after 32 days having retired over 210K galaxies. See Table 1 for details.

diction, the RF algorithm computes the probability for each subject to belong to each class. This probability is simply the average of the probabilities of each individual decision tree, where the probability of a single tree is determined as the fraction of subjects of class X on a leaf node. Only subjects that receive a class prediction of 'Featured'with $p_{\rm machine} \geq 0.9$ ($p_{\rm machine} \leq 0.1$ for 'Not') are considered retired. The remaining subjects have the possibility of being classified by humans or the machine on a future night of the simulation. This constitutes the core of our passive feedback mechanism. Subjects that are not retired by the machine can instead be retired by humans, thus providing the machine a more fully sampled morphology parameter space on future training sessions.

6. RESULTS

We perform a full GZX simulation incorporating our RF with the fiducial SWAP run discussed in Section 4.2. The machine attempts its first training on Day 8 with an initial training sample of ∼20K subjects. It undergoes several additional nights of training, each time with a larger training sample. By Day 12, SWAP has provided over 40K subjects for training and the machine's agent has deemed the machine optimized. The machine predicts class labels for the remaining 230K GZ2 subjects. Of those, the machine retires over 70K, dramatically increasing the subset of retired subjects. We end the simulation after 32 days, having retired ∼210K subjects as detailed in Table 1.

We present these results in Figure 9 where subject retirement with GZX (red) is compared to our fiducial SWAP-only run (light blue) and GZ2 (dashed dark blue). Using the $\rm GZ2_{raw}$ labels as before, we compute our usual quality metrics on the full sample of GZX-retired subjects; reported in Table 1. Accuracy and purity remain within a few percent of the SWAP-only run at 93.5% and 84.2% respectively. Instead we see a 5% decline in the completeness. While the SWAP-only run identified 99% of 'Featured' subjects, incorporation of the machine seems to miss a significant portion thus dropping GZX completeness to 94.3%. We discuss this behaviour below.

By dynamically generating a training sample through a more sophisticated analysis of human classifications coupled with a machine classifier, we retire more than 200K GZ2 subjects in just 27 days. Visual classification through SWAP alone retires as many in 50 days, while GZ2 requires a full year. Though our analysis considers only the top-level task of GZ2's decision tree, GZX

suggests a tantalizing potential to increase the classification rate by an order of magnitude over the traditional crowd-sourced approach. We next explore the composition of those classifications.

6.1. *Who retires what, when?*

In the top panel of Figure 10 we explore the individual contributions to GZX subject retirement from the RF (dash-dotted teal) and SWAP (dashed orange). The solid black line shows the total GZX retirement (SWAP+RF), while the dotted grey line depicts the fiducial SWAP-only run from Section 4.2 for reference. Two things are immediately obvious. First, each component shoulders approximately half of the retirement burden with the machine and SWAP responsible for ∼100K and ∼110K subjects respectively. Secondly, the rate of retirement exhibited by the two components is in stark contrast. SWAP retires at a relatively constant rate while the machine retires dramatically at the beginning of its application, quickly surpassing the human contribution, and plateaus thereafter. We thus clearly see three epochs of subject retirement. In the first phase, humans are the only contributors to subject retirement. Once the machine is optimized, it immediately contributes more to retirement than humans. However, the machine's performance plateaus quickly; the third phase is again dominated by human classifications.

In the bottom panels of Figure 10, we consider the class composition of subjects retired by SWAP and the RF. The left (right) panel shows the retired fraction of GZ2 subjects identified as 'Featured' ('Not') according to their $GZ2_{raw}$ labels as a function of GZ2 project time. Overall, GZX retires 73.6% of the GZ2 subject sample and this is evenly distributed between 'Featured' and 'Not' subjects as indicated by the solid black lines in both panels. However, SWAP retires more than 50% of all 'Featured' subjects while the machine retires only 18%. This divergence does not exist for 'Not' subjects where each component contributes 33-37%.

What is the source of this discrepancy? Each night the machine trains on a sample composed consistently of 30-40% 'Featured' subjects but does not retire a similar proportion, indicating that the 30% of non-retired 'Featured' subjects do not receive high $p_{machine}$. In the following section we explore whether this is an artefact of our choice in machine or in the human-machine combination implemented here.

6.2. *Machine performance*

Throughout our analysis we have defined 'Featured' and 'Not' subjects by their $GZ2_{raw}$ labels as this was the most compatible choice for comparison with SWAP output. However, the machine does not learn in the same way, nor is it presented with the same information. Machine and human classifications each provide valuable and complementary information for identifying 'Featured' galaxies.

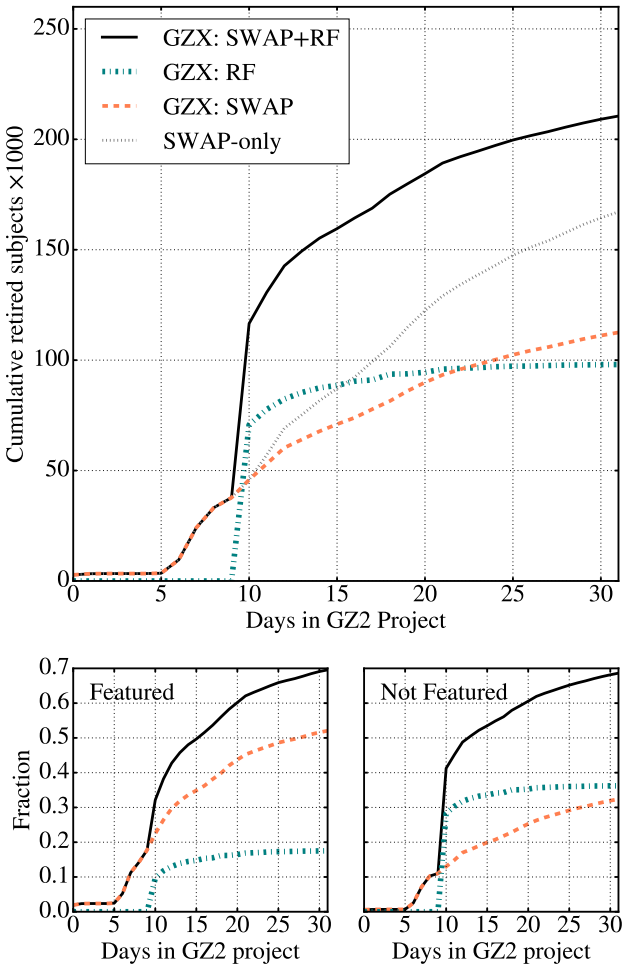

Figure 10. Contributions to subject retirement by both classifying agents of GZX: human (SWAP) and machine. The top panel shows cumulative subject retirement for GZX as a whole (solid black), along with that attributed to the RF (dash-dotted teal), and SWAP (dashed orange). The dotted grey line shows the fiducial SWAP-only run for comparison. Retirement totals for humans and machine are nearly equal over the course of the simulation but display different behaviours: SWAP's retirement rate is almost constant while the RF contributes substantially after its initial application and then plateaus. The bottom panels show what fraction of GZ2 subjects are retired, separated by class label. Overall, GZX retires 73.6% of the entire GZ2 sample in 32 days, retiring the same proportion of 'Featured' and 'Not' subjects as indicated by the black lines. However, humans retire 30% more 'Featured' subjects than the machine, while both components retire a similar proportion of 'Not' subjects.

We isolate the 6127 subjects that were deemed false positives, i.e., galaxies retired by the machine as 'Featured' that have 'Not' $GZ2_{raw}$ labels, a sample that comprises only 6.25% of all subjects the machine retires. We visually examine several hundred and assess that, to the expert eye, a majority are, in fact, 'Featured'. A random

Table 1. Summary of key quantities for GZ2 and our various simulations. All quality metrics are calculated using $GZ2_{raw}$ labels.

Simulation Summary

	Days	Subjects Retired	Human Effort (classifications)	Accuracy (%)	Purity (%)	Completeness (%)
Galaxy Zoo 2	430	285962	16,340,298	–	–	–
SWAP only	92	226124	2,298,772	95.7	86.7	99.0
SWAP+RF	32	210543	932,017	93.5	84.2	94.3

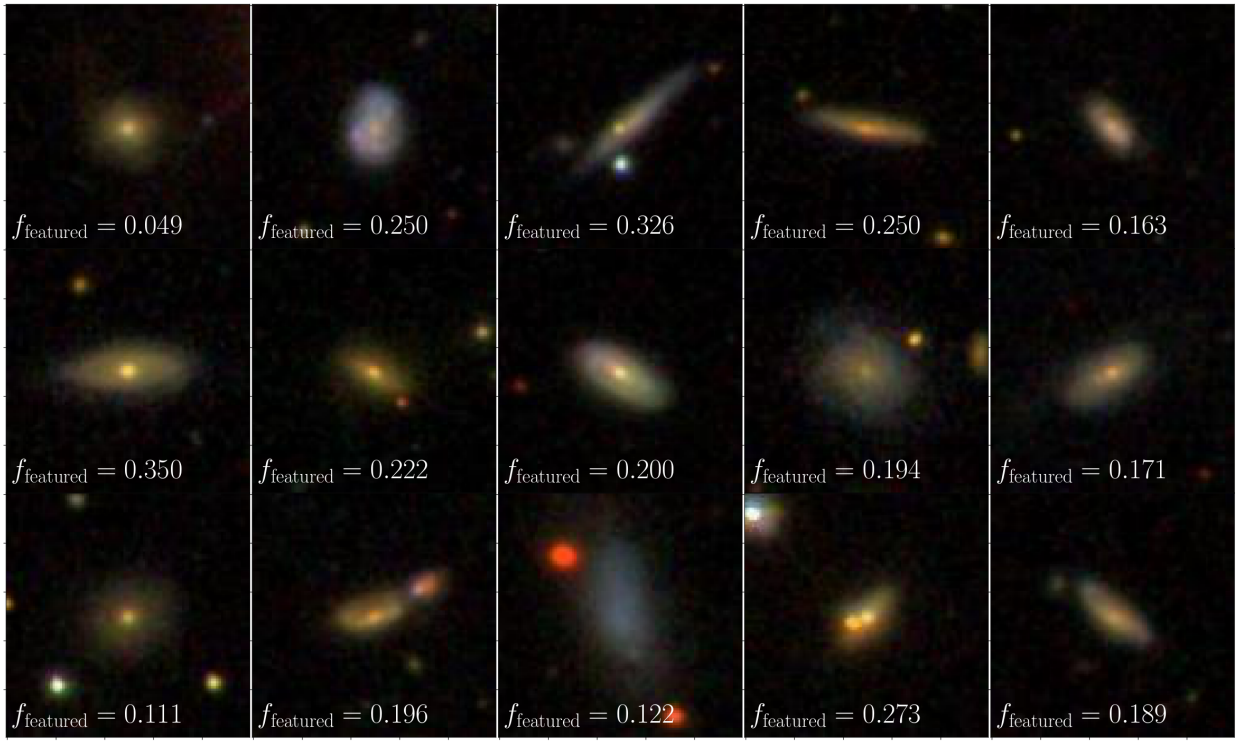

Figure 11. A random subsample of subjects identified as false positives: labelled by machine as 'Featured', but as 'Not' according to $GZ2_{raw}$. We display f_{smooth} in the lower left corner, that is, the fraction of volunteers who classified the subject as 'smooth' ('Not'). Values are typically between 0.5 and 0.65 indicating that GZ2 volunteers did not reach a strong consensus. Fortunately, the machine is able to identify these subjects as 'Featured' due to their measured morphology diagnostics.

sample is shown in Figure 11.

That the machine strongly identifies these galaxies as 'Featured' ($p_{machine} \geq 0.9$) where humans instead classify them as 'Not' ($f_{featured} < 0.5$) has several contributing factors: 1) as discussed in Section 4.5, the threshold we chose carries with it a confidence interval such that subjects with $0.4 < f_{featured}+f_{artifact} < 0.6$ are most likely to receive disagreeing labels from other classifying agents, 2) the first task of the GZ2 decision tree asks a question that does not necessarily correlate with a split between early- and late-type galaxies, and 3) the machine learns on morphology diagnostics that are very different from visual inspection.

We find that 41.4% of these false positives have $0.4 \leq f_{featured}+f_{artifact} < 0.5$ indicating that the disagreement between humans and machine is likely due to the labels we assign at our given threshold. However, we also find that 43.5% of false positives have $f_{featured}+f_{artifact} \leq 0.35$, and this discrepancy is not as easily explained. In Figure 11 we examine a random sample of false positives in this regime where, for clarity, we display only the $f_{featured}$ value in the lower left corner. The majority of these subjects are discs lacking features such as spiral arms or strong bars. Whether this is the reason the majority of volunteers classify these objects as "smooth" is beyond the scope of this paper, however, this behaviour might be modified by providing actual training images and live feedback as performed in Marshall et al. (2016). We suggest that, at least for this particular question, if either human or machine identifies a subject as 'Featured', it is likely the subject is discy and worth further investigation.

Accordingly, this suggests that, in some cases, the morphology indicators we measure are sufficient for the

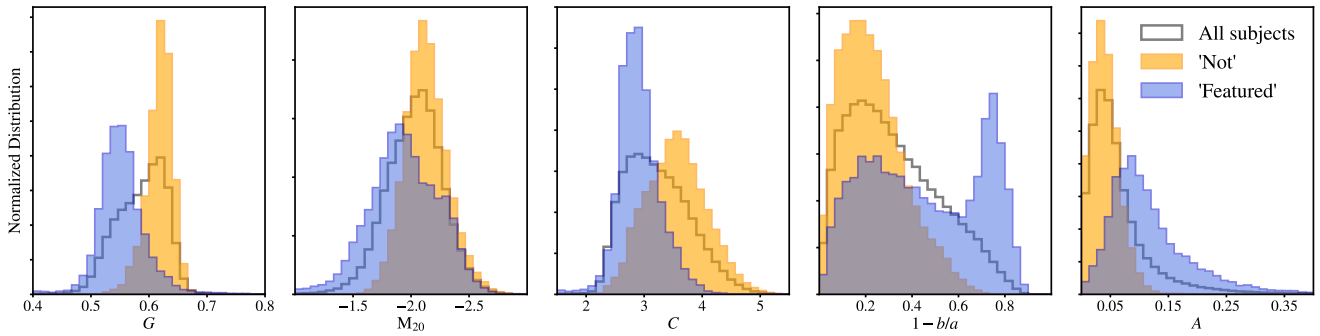

Figure 12. The RF is trained on a 5-dimensional morphology parameter space. We show the distribution of each morphology indicator for machine-retired 'Featured' (blue) and 'Not' (orange) subjects compared to the full GZ2 subject sample (black). The difference between 'Featured' and 'Not' subjects is in stark contrast for all distributions except, perhaps, M_{20}.

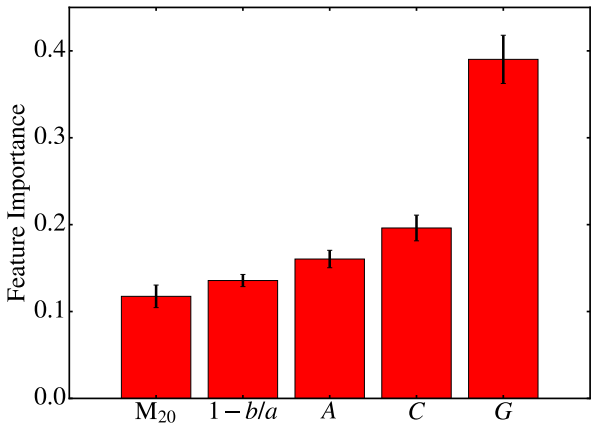

Figure 13. The RF's ranked feature importance averaged over all nights of training with black bars indicating the standard deviation. A larger value corresponds to higher importance. The machine computes feature importance according to how much each feature increases the purity of the resulting split averaged over all trees in the forest. The RF places great importance in the Gini coefficient though we note that it can under-represent the importance of highly correlated features such as concentration.

machine to recognize 'Featured' galaxies regardless of the labels humans provide. Figure 12 shows the distribution of each morphology indicator for all subjects the machine retires as 'Featured' (blue) and 'Not' (orange) compared to the full GZ2 subject set. The difference between 'Featured' and 'Not' is stark in all but the M_{20} distribution. This can be seen explicitly in Figure 13 in which we show the RF's ranked feature importances, where large values indicate higher importance. Feature importance is computed as how much each feature decreases the impurity of a split in a tree. The impurity decrease from each feature is then averaged over all trees and ranked. We show the feature importance averaged over all nights of training with black bars indicating the standard deviation. The machine finds the Gini coefficient most important for class prediction, placing little emphasis on M_{20}. It is well known that the Gini coefficient is more sensitive to noise than other diagnostics, however, we point out that when a machine is faced with two or more correlated features any of them can be used as the predictor. Once chosen, the importance of the others is reduced. This explains why Concentration is ranked much lower than Gini even though they are strongly correlated as seen in Figure A2. That the machine relies heavily on these two morphology diagnostics is unsurprising as concentration has long been an automated predictor between early- and late-type galaxies (Abraham et al. 1994, 1996; Shen et al. 2003).

The complementary nature of human and machine classification can best be utilized by a feedback mechanism in which a portion of machine-retired subjects are reviewed by humans. Subjects that display excessive disagreement should be verified by an expert (or expert-user). In the same way that humans increase the machine's training sample over time, subjects that the machine properly identifies can become part of the humans' training sample.

7. LOOKING FORWARD

We have demonstrated the first practical framework for combining human and machine intelligence in galaxy morphology classification tasks. While we focus below on a brief discussion of our next steps and potential applications to large upcoming surveys, we note that our results have implications for the future of citizen science and Galaxy Zoo in particular.

GZX is perhaps one of the simplest ways to combine human and machine intelligence and its impressive performance motivates a higher level of sophistication. A first step will be an implementation of SWAP that can handle a complex decision tree. In addition, we envision multiple forms of active feedback in addition to our passive feedback mechanism. SWAP allows us to leverage the most skilled volunteers to review galaxies difficult for either human or machine to classify. Additionally,

machine-retired subjects should contribute to the training sample for humans in an analogous fashion to what we have already implemented.

Secondly, our RF can be improved by providing it information equal to what humans receive: multi-band morphology diagnostics will be included in our future feature vector. However, the Random Forest algorithm is not easily adapted to handle measurement errors or class labels with continuous distributions. To fully utilize the information provided by SWAP, sophisticated algorithms should be considered such as deep convolutional neural networks (CNN) or Latent Dirichlet allocation (LDA), an algorithm that is frequently used in document processing. Furthermore, there is no reason to limit to a single machine. As hinted at in Figure 1, several machines could train simultaneously, their predictions aggregated through SWAP, creating an on-the-fly machine ensemble.

With the above upgrades implemented, we expect performance of both the classification rate and quality to further increase. However, even our current implementation can cope with upcoming data volumes from large surveys. By some estimates, *Euclid* is expected to obtain measurable morphology with its visual instrument (VIS) for approximately $10^6 - 10^7$ galaxies (Laureijs et al. 2011). Visual classification at the rate achieved with Galaxy Zoo today would require 12–120 years to classify.[5] If the *Euclid* sample is on the high end, GZX as currently implemented could classify the brightest 20% during the six years of its observing mission. As currently implemented, we obtain accuracy around 95% potentially leaving hundreds of thousands of galaxies with unreliable classifications. In a companion paper that seeks to identify supernovae, Wright et al. (submitted) demonstrate a dramatic increase in accuracy through an entirely different human-machine combination whereby the scores from human and machine are averaged together with the combined score yielding the most reliable classification. Again, a combination of both approaches will allow us to take full advantage of legacy output from large scale surveys.

7.1. *Conclusions*

In this paper we design and test Galaxy Zoo Express, an innovative system[6] for the efficient classification of galaxy morphology tasks that integrates the native ability of the human mind to identify the abstract and novel with machine learning algorithms that provide speed and brute force. We demonstrate for the first time that the SWAP algorithm, originally developed to identify rare gravitational lenses in the Space Warps project, is robust for use in galaxy morphology classification. We show that by implementing SWAP on GZ2 classification data we can increase the rate of classification by a factor of 4-5, requiring only 90 days of GZ2 project time to classify nearly 80% of the entire galaxy sample.

Furthermore, we have implemented and tested a Random Forest algorithm and developed a decision engine that delegates tasks between human and machine. We show that even this simple machine is capable of providing significant gains in the classification rate when combined with human classifiers: GZX retires over 70% of GZ2 galaxies in just 32 days of GZ2 project time. This represents a factor of 11.4 increase in the classification rate as well as an order of magnitude reduction in human effort compared to the original GZ2 project. This is achieved without sacrificing the quality of classifications as we maintain accuracy well above 90% throughout our simulations. Additionally, we have shown that training on a 5-dimensional parameter space of traditional non-parametric morphology indicators allows the machine to identify subjects that humans miss, providing a complementary approach to visual classification. The gain in classification speed allows us to tackle the massive amount of data promised from large surveys like *LSST* and *Euclid*.

MB thanks Steven Bamford and Boris Häußler for insightful discussions on citizen science and Galaxy Zoo; and John Wallin and Marc Huertas-Company for several enlightening conversations on machine learning and classification. We are grateful to Elisabeth Baeten, Micaela Bagley, Karlen Shahinyan, Vihang Mehta, Steven Bamford, Kevin Schawinski, and Rebecca Smethurst for providing expert classifications in addition to those provided by the authors. PJM acknowledges Aprajita Verma and Anupreeta More for their ongoing collaboration on the Space Warps project.

MB, CS, LF, KW, and MG gratefully acknowledge support from the US National Science Foundation Grant AST-1413610. MB acknowledges additional support through New College and Oxford University's Balzan Fellowship as well as the University of Minnesota Doctoral Dissertation Fellowship. Travel funding was supplied to MB, in part, by the University of Minnesota Thesis Research Travel Grant. CJL recognizes support from a grant from the Science & Technology Facilities Council (ST/N003179/1). BDS acknowledges support from Balliol College, Oxford, and the National Aeronautics and Space Administration (NASA) through Einstein Postdoctoral Fellowship Award Number PF5-160143 is-

[5] We note that the classification rate of GZ2 was 4 times higher than GZ's current steady rate.

[6] Our code can be found at https://github.com/melaniebeck/GZExpress

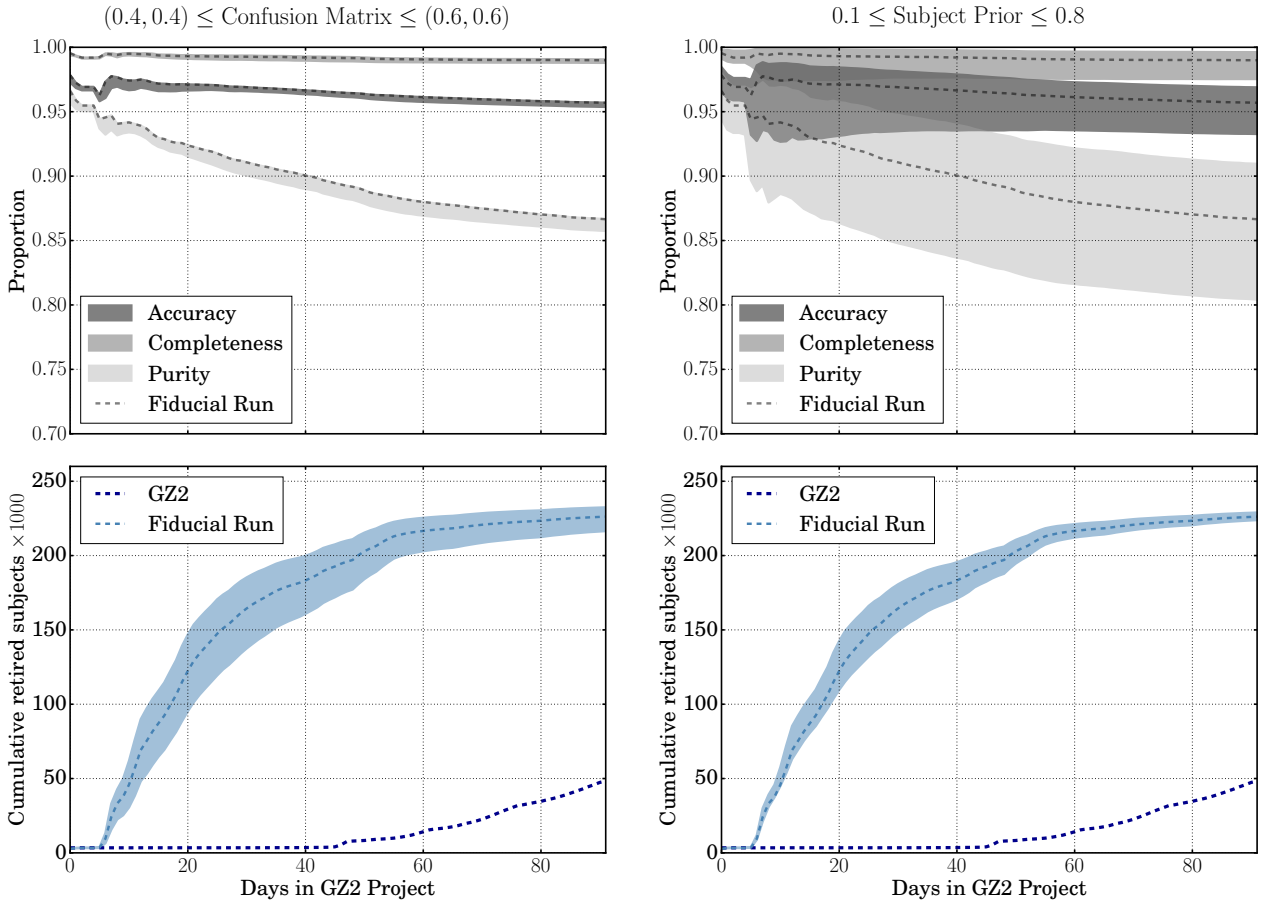

Figure A1. SWAP performance does not dramatically change even with a range of input parameters (shaded regions) as compared to the fiducial run of Section 4.2 (dashed lines). *Left.* The quality (top) and retirement rate (bottom) when the confusion matrix is initialized as (0.4, 0.4) and (0.6, 0.6), with all other input parameters remaining constant. *Right.* Same as the left panel but allowing the subject prior probability, $p_0 = 0.2, 0.35$ and 0.8. Changing the confusion matrix has little impact on the quality of the labels but varies the total number of subjects retired. In contrast, changing the subject prior is more likely to affect the classification quality rather than the total number of subjects retired.

sued by the Chandra X-ray Observatory Center, which is operated by the Smithsonian Astrophysical Observatory for and on behalf of NASA under contract NAS8-03060. The work of PJM is supported by the U.S. Department of Energy under contract number DE-AC02-76SF00515.

Software: scikit-learn (Pedregosa et al. 2011), Astropy (Astropy Collaboration et al. 2013), TOPCAT (Taylor 2005)

APPENDIX

A. EXPLORING SWAP'S PARAMETER SPACE

In this Appendix we explore the SWAP parameter space and assess the effects on subject retirement.

Initial agent confusion matrix. In our fiducial simulation each volunteer was assigned an agent whose confusion matrix was initialized at $(0.5, 0.5)$, which presumes that volunteers are no better than random classifiers. We perform two simulations wherein we initialize agent confusion matrices as $(0.4, 0.4)$, slightly obtuse volunteers; and $(0.6, 0.6)$, slightly astute volunteers, with everything else remaining constant. Results of these simulations compared to the fiducial run are shown in the left panel of Figure A1. We find that SWAP is largely insensitive to the initial confusion matrix both in terms of the subject retirement rate and classification quality.

We retire ∼225K±3.5% subjects as shown by the light blue shaded region in the bottom left panel of Figure A1, where the dashed blue line denotes the fiducial run. Predictably, when the confusion matrix probabilities are low, we retire fewer subjects than when these probabilities are high for a given period of time. This is easy to understand since it takes longer for volunteers to become astute classifiers when they are initially given values denoting them as obtuse.

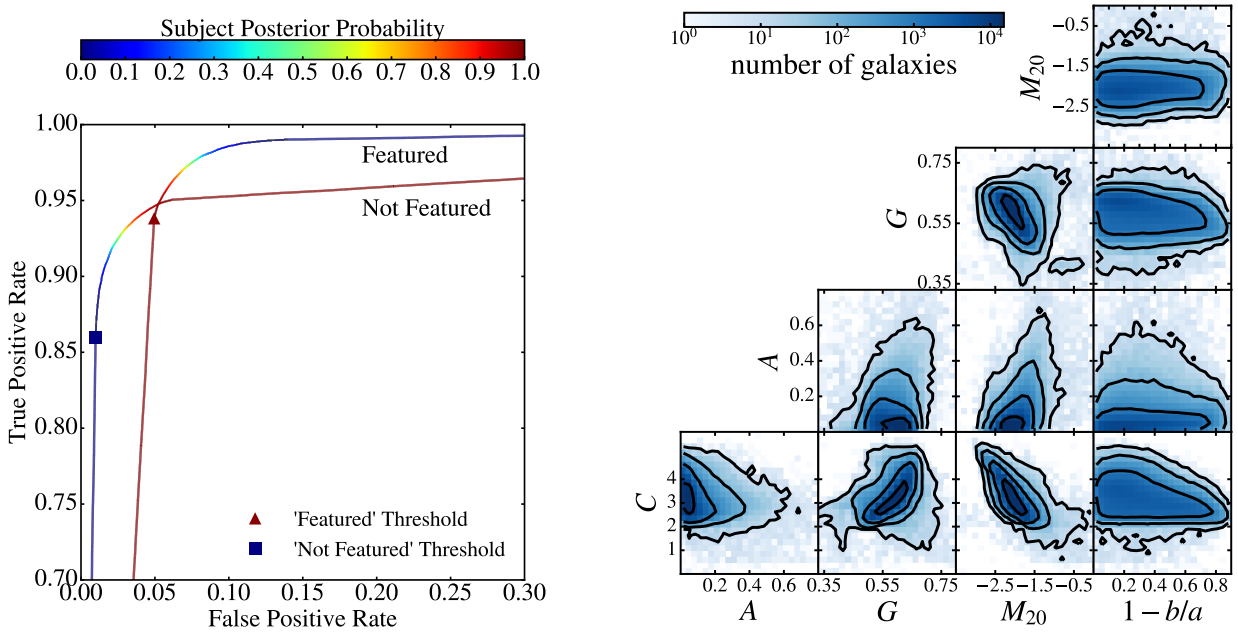

Figure A2. *Left.* Identifying 'Featured' subjects is independent of identifying 'Not' subjects. Both ROC curves use all subjects processed by SWAP where the score used to create the ROC curve is simply each subject's achieved posterior probability. The Featured curve demonstrates how well we identify 'Featured' subjects with a threshold of 0.99, while the Not Featured curve demonstrates how well we identify 'Not' subjects with a threshold of 0.004. Typically, best performance is achieved by the score associated with the upper-left-most part of the curve. Our 'Featured' threshold is nearly optimal, while our 'Not' threshold could be improved since the blue square is not as close to the upper left hand corner as other possible values of the subject posterior. *Right.* Relation between measured morphology diagnostics for more than 280K SDSS galaxies. Most of these galaxies are processed through SWAP, receiving a posterior probability that estimates how likely each is to be 'Featured' or 'Not'.

Regardless, most volunteers become astute classifiers by the end of the simulation. The top left panel demonstrates our usual quality metrics as computed in Section 4.2. The dashed lines again denote the fiducial run. We maintain ∼95% accuracy, 99% completeness, and ∼84% purity; and no metric changes by $> 2\%$ regardless of initial confusion matrix values.

This spread is due to three effects: 1) subjects can receive an alternate SWAP label in different simulations, 2) subjects can be retired in a different order, and 3) the set of retired subjects is not guaranteed to be common to all runs. We find SWAP to be highly consistent: more than 99% of retired subjects are the same among all simulations, and, of these, 99% receive the same label. Instead we find that the order in which subjects are retired changes between runs. When the confusion matrix is low, subjects take longer to classify compared to the fiducial run (i.e., they retire on a later date in GZ2 project time). Likewise, subjects retire sooner when the confusion matrix is high. This can cause quality metrics to vary since they are calculated on a day to day basis. These effects each contribute less than one per cent variation and thus we see a high level of consistency between simulations.

Of interest, perhaps, is that the quality metrics for these simulations are not symmetric about the fiducial run. However, in the Bayesian framework of SWAP, an agent with confusion matrix (0.4, 0.4) contributes as much information as an agent with confusion matrix (0.6, 0.6). The quality metrics computed are thus within a per cent of each other. In either case, we find that initializing agents at (0.5, 0.5) provides optimal performance for the 'training' we simulate with our current approach. Further assessment would require a live project with real-time training and feedback.

Subject prior probability, p_0. The prior probability assigned to each subject is an educated guess of the frequency of that characteristic in the scope of the data at hand. For galaxy morphologies, this number should be an estimate of the probability of observing a desired feature (bar, disk, ring, etc.). In our case, we desire simply to find galaxies that are 'Featured'; however, this is dependent on mass, redshift, physical size, etc. The original GZ2 sample was selected primarily on magnitude and redshift. As there was no cut on galaxy size (with the exception that each galaxy be larger than the SDSS PSF), the sample includes a large range of masses and sizes. Designating a single prior is not clear-cut; we thus explore how various p_0 values effect the SWAP outcome.

We run simulations allowing p_0 to take values 0.2, 0.35, and 0.8 and compare these to the fiducial run, with everything else remaining constant. The results are shown in the right panels of Figure A1. We again find that SWAP is consistent in terms of subject retirement which varies by only 1%. However, as can be seen in the top panel, the variation in our quality metrics is more pronounced. Firstly, though we retire nearly the same number of subjects over the course of each simulation, they are less consistent than our previous runs. That is, only 95% of retired subjects are common to all simulations. Secondly, of those that are common, only 94% receive the same label from SWAP indicating that changing the prior is more likely to produce a different label for a given subject than changing the initial agent confusion matrix. Finally, there is also a larger spread for the day on which a subject is retired as compared to the fiducial run. These trends all contribute to a broader spread in accuracy, completeness, and purity as a function of project time. We stress, however, that although more substantial than the previous comparison, these variations are all within $\pm 5\%$.

We can understand these variations more intuitively by considering the following. Recall that our retirement thresholds, t_F and t_N, have not changed in these simulations. When p_0 is small, the subject's probability is already closer to t_N in probability space, and thus more subjects are classified as 'Not' compared to the fiducial run. Similarly, when p_0 is large, some of these same subjects can instead be classified as 'Featured' because p_0 is already closer to t_F. Obviously, both outcomes cannot be correct. We find that the simulation with $p_0 = 0.8$ performs the worst of any run; this is a direct reflection of the fact that this prior is not suitable for this question or this dataset. Indeed, the best performance is achieved when $p_0 = 0.35$. This reflects the distribution of 'Featured' subjects as determined by $\mathrm{GZ2_{raw}}$ labels and is more characteristic of the expected proportion of 'Featured' galaxies in the local universe. As a value far from the correct value can have a significant impact on the classification quality, it is important to choose a prior wisely.

Retirement thresholds, t_F and t_N. Retirement thresholds are directly related to the time that a subject will spend in SWAP before retirement. If we lower t_F (and/or raise t_N), more subjects will be retired compared to the fiducial run as each subject will have a smaller swath of probability space in which to fluctuate before crossing one of these thresholds. On the other hand, if we raise t_F (and/or lower t_N), it will take longer for subjects to cross one of these thresholds. This also increases the likelihood of some subjects never crossing either threshold, instead oscillating indefinitely through probability space.

What thresholds should one choose? To answer this question, we consider the left panel of Figure A2, which depicts the receiver operating characteristic (ROC) curve for our fiducial simulation, an illustration of performance as a function of a threshold for a binary classifier. ROC curves display the true positive rate against the false positive rate for a discriminatory threshold or score with a perfect classifier achieving 100% true positives and no false positives. The value of the threshold optimal for predicting class labels would be that which allows the ROC curve to reach the upper-left-most point in the diagram. We have two thresholds to consider and thus we plot the curve twice: once under the assumption that "true positives" denote correctly identified 'Featured' subjects; and again under the assumption that "true positives" instead denote correctly identified 'Not' subjects. In both cases, the colour of the line corresponds to the subject posterior probability. We mark the location of $t_F = 0.99$ and $t_N = 0.004$ from our fiducial run with a red triangle and blue square respectively. We see that t_F is nearly optimal but t_N could be improved upon.

B. MEASURING NONPARAMETRIC MORPHOLOGICAL DIAGNOSTICS ON SDSS STAMPS

In order to train our Random Forest machine learning algorithm, we measure non-parametric morphology diagnostics for the GZ2 galaxy sample.

We obtain i-band imaging from SDSS Data Release 12. Postage stamps are made from the SDSS fields for each galaxy with dimensions of 3 Petrosian radii. Galaxies located within 3 Petrosian radii of the edge of a field were excluded. Postage stamps undergo a cleaning process whereby nearby sources are identified with SExtractor (ver. 2.8.6; Bertin & Arnouts 1996) and their pixels replaced with values that mimic the background in that region. We compute the following widely adopted nonparametric measurements of the galaxy light distribution on the cleaned postage stamps:

Concentration is computed as $C = 5\log(r_{80}/r_{20})$ where r_{80} and r_{20} are the radii containing 80% and 20% of the galaxy light respectively. Small values of this ratio tend to indicate disky galaxies, while larger values correlate with early-type ellipticals.

Asymmetry quantifies the degree of rotational symmetry in the galaxy light distribution (not necessarily the physical shape of the galaxy as this parameter is not highly sensitive to low surface brightness features). A correction for

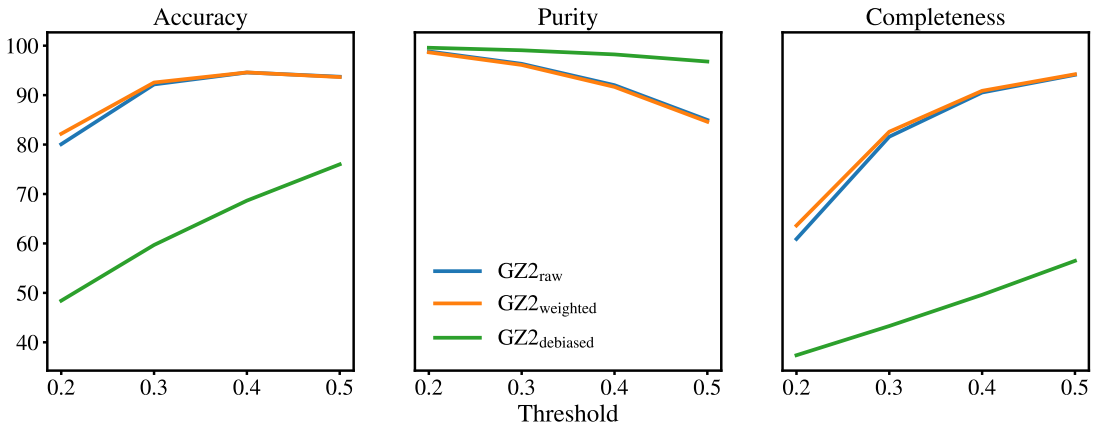

Figure C3. Quality metrics computed on the subjects retired during the GZX simulation for a range of thresholds and GZ2 vote fraction types.

background noise is applied (as in e.g. Conselice et al. (2000)), i.e.,

$$A = \frac{\sum_{x,y} |I - I_{180}|}{2 \sum |I|} - B_{180} \tag{B1}$$

where I is the galaxy flux in each pixel (x, y), I_{180} is the image rotated by 180 degrees about the galaxy's central pixel, and B_{180} is the average asymmetry of the background.

The Gini coefficient, G, (Glasser 1962; Abraham et al. 2003) describes how uniformly distributed a galaxy's flux is. If G is 0, the flux is distributed homogeneously among all galaxy pixels; if G is 1, the light is contained within a single pixel. This term correlates with C, however, G does not require that the flux be in the central region of the galaxy. We follow Lotz et al. (2004) by first ordering the pixels by increasing flux value, and then computing

$$G = \frac{1}{|\bar{X}|n(n-1)} \sum_{i}^{n} (2i - n - 1)|X_i| \tag{B2}$$

where n is the number of pixels assigned to the galaxy, and $\bar{X}$ is the mean pixel value.

M_{20} (Lotz et al. 2004) is the second order moment of the brightest 20% of the galaxy flux. We compute it as

$$M_{tot} = \sum_{i}^{n} f_i[(x_i - x_c)^2 + (y_i - y_c)^2] \tag{B3}$$

$$M_{20} = \log_{10}(\frac{\sum_i M_i}{M_{tot}}), \quad \text{while} \sum_i fi < 0.2 f_{tot} \tag{B4}$$

where M_{tot}, the total moment, is computed first and f_{tot} is the total flux. For centrally concentrated objects, M_{20} correlates with C but is also sensitive to bright off-centre knots of light.

Finally, we use the ellipticity, $\epsilon = 1 - b/a$, of the light distribution as measured by SExtractor which computes the semi-major axis a and semi-minor axis b from the second-order moments of the galaxy light.

In total, we measure morphological indicators for 282,350 SDSS galaxies. The relations between these diagnostics for the full sample is shown in the right panel of Figure A2. The code developed to clean and compute these morphology indicators is open source and can be found at https://github.com/melaniebeck/measure_morphology.

C. EXPLORING THE QUALITY OF GALAXY ZOO: EXPRESS

(Added: In this section we consider the robustness of GZX by computing several sets of "ground truth" labels from the GZ2 catalogue. We recalculate the quality metrics of accuracy, purity, and completeness on the final sample of retired galaxies from the full GZX simulation (SWAP+RF) for a range of thresholds and for each type of GZ2 vote fractions. Specifically, we define labels such that any subject with $f_{\rm featured} + f_{\rm artifact} \geq t$ is labelled 'Featured', otherwise it is labelled 'Not'. This is performed for a range of threshold values, $t = [0.2, 0.3, 0.4, 0.5]$, and using each type of GZ2 vote fraction: raw, weighted, and debiased.)

(Added: In Figure C3 we show how the perceived quality of GZX changes as a function of threshold and GZ2 vote fraction type. GZX classifications are quite robust, with accuracy fluctuating by only a couple percent for GZ2 labels computed using a threshold between 0.3 and 0.5. Though accuracy remains relatively flat for this range of thresholds, we see the trade off between purity and completeness. Decreasing the threshold on which labels are computed (i.e., more subjects are given a 'Featured' label) results in higher purity but lower completeness.)

(Added: That the $GZ2_{debiased}$ labels perform poorly is not surprising. These vote fractions are computed after considerable post-processing of the raw volunteer votes in order to remove the effects of redshift and surface brightness. As with any set of visual classifications, these biases must be accounted for and this is traditionally done retrospectively. It is also unsurprising that the $GZ2_{raw}$ and $GZ2_{weighted}$ classifications are in such tight agreement. $GZ2_{weighted}$ vote fractions are computed by down-weighting inconsistent volunteers of which there are relatively few. These two sets of vote fractions are thus very similar.)

(Added: When applying GZX to future imaging programs, there will be no "ground truth" labels for comparison. In some sense, these thresholds can be interpreted as a prior for the SWAP p_0 value, the initial probability for a subject to be 'Featured'. As we showed in Appendix **??** changing the prior has little affect on the retirement rate but does result in considerable variability in the completeness and purity of the resulting classifications. When applying GZX to a future sample, science teams will need to determine between prioritizing between these which will be dependent on the particular science case. Additionally,)

REFERENCES

Abraham, R. G., Tanvir, N. R., Santiago, B. X., et al. 1996, MNRAS, 279, L47

Abraham, R. G., Valdes, F., Yee, H. K. C., & van den Bergh, S. 1994, ApJ, 432, 75

Abraham, R. G., van den Bergh, S., & Nair, P. 2003, ApJ, 588, 218

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

Baillard, A., Bertin, E., de Lapparent, V., et al. 2011, A&A, 532, A74

Ball, N. M., Loveday, J., Fukugita, M., et al. 2004, MNRAS, 348, 1038

Bamford, S. P., Nichol, R. C., Baldry, I. K., et al. 2009, MNRAS, 393, 1324

Banerji, M., Lahav, O., Lintott, C. J., et al. 2010, MNRAS, 406, 342

Bershady, M. A., Jangren, A., & Conselice, C. J. 2000, AJ, 119, 2645

Bertin, E., & Arnouts, S. 1996, A&AS, 117, 393

Blanton, M. R., Hogg, D. W., Bahcall, N. A., et al. 2003, ApJ, 594, 186

Breiman, L. 2001, Machine Learning, 45, 5

Cardamone, C., Schawinski, K., Sarzi, M., et al. 2009, MNRAS, 399, 1191

Casteels, K. R. V., Conselice, C. J., Bamford, S. P., et al. 2014, MNRAS, 445, 1157

Conselice, C. J. 2003, ApJS, 147, 1

—. 2006, MNRAS, 373, 1389

Conselice, C. J., Bershady, M. A., & Jangren, A. 2000, ApJ, 529, 886

Darg, D. W., Kaviraj, S., Lintott, C. J., et al. 2010, MNRAS, 401, 1552

de Vaucouleurs, G. 1959, Handbuch der Physik, 53, 275

Dieleman, S., Willett, K. W., & Dambre, J. 2015, MNRAS, 450, 1441

Dressler, A. 1980, ApJ, 236, 351

Elmegreen, B. G., Bournaud, F., & Elmegreen, D. M. 2008, ApJ, 688, 67

Elmegreen, B. G., Elmegreen, D. M., Sánchez Almeida, J., et al. 2013, ApJ, 774, 86

Freeman, P. E., Izbicki, R., Lee, A. B., et al. 2013, MNRAS, 434, 282

Galloway, M. A., Willett, K. W., Fortson, L. F., et al. 2015, MNRAS, 448, 3442

Glasser, G. J. 1962, Journal of the American Statistical Association, 57, 648

Griffith, R. L., Cooper, M. C., Newman, J. A., et al. 2012, ApJS, 200, 9

Holwerda, B. W., Muñoz-Mateos, J.-C., Comerón, S., et al. 2014, ApJ, 781, 12

Hubble, E. P. 1936, The Realm of the Nebulae (Yale University Press)

Huertas-Company, M., Rouan, D., Tasca, L., Soucail, G., & Le Fèvre, O. 2008, A&A, 478, 971

Huertas-Company, M., Gravet, R., Cabrera-Vives, G., et al. 2015, ApJS, 221, 8

Kartaltepe, J. S., Mozena, M., Kocevski, D., et al. 2015, ApJS, 221, 11

Kauffmann, G., Heckman, T. M., White, S. D. M., et al. 2003, MNRAS, 341, 54

Kormendy, J. 1977, ApJ, 217, 406

Kormendy, J., & Kennicutt, Jr., R. C. 2004, ARA&A, 42, 603

Land, K., Slosar, A., Lintott, C., et al. 2008, MNRAS, 388, 1686

Laureijs, R., Amiaux, J., Arduini, S., et al. 2011, ArXiv e-prints, arXiv:1110.3193

Lintott, C., Schawinski, K., Bamford, S., et al. 2011, MNRAS, 410, 166

Lintott, C. J., Schawinski, K., Slosar, A., et al. 2008, MNRAS, 389, 1179

Lotz, J. M., Primack, J., & Madau, P. 2004, AJ, 128, 163

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

Marshall, P. J., Verma, A., More, A., et al. 2016, MNRAS, 455, 1171

Masters, K. L., Nichol, R. C., Hoyle, B., et al. 2011, MNRAS, 411, 2026

Meert, A., Vikram, V., & Bernardi, M. 2016, MNRAS, 455, 2440
More, A., Verma, A., Marshall, P. J., et al. 2016, MNRAS, 455, 1191
Nair, P. B., & Abraham, R. G. 2010, ApJS, 186, 427
Nakamura, O., Fukugita, M., Yasuda, N., et al. 2003, AJ, 125, 1682
Odewahn, S. C., Cohen, S. H., Windhorst, R. A., & Philip, N. S. 2002, ApJ, 568, 539
Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2011, Journal of Machine Learning Research, 12, 2825
Peng, C. Y., Ho, L. C., Impey, C. D., & Rix, H.-W. 2002, AJ, 124, 266
Peng, Y.-j., Lilly, S. J., Kovač, K., et al. 2010, ApJ, 721, 193
Peth, M. A., Lotz, J. M., Freeman, P. E., et al. 2016, MNRAS, 458, 963
Sandage, A. 1961, The Hubble atlas of galaxies
Scarlata, C., Carollo, C. M., Lilly, S., et al. 2007, ApJS, 172, 406
Schawinski, K., Urry, C. M., Simmons, B. D., et al. 2014, MNRAS, 440, 889
Sersic, J. L. 1968, Atlas de galaxias australes
Shen, S., Mo, H. J., White, S. D. M., et al. 2003, MNRAS, 343, 978
Sheth, K., Elmegreen, D. M., Elmegreen, B. G., et al. 2008, ApJ, 675, 1141
Simard, L., Mendel, J. T., Patton, D. R., Ellison, S. L., & McConnachie, A. W. 2011, ApJS, 196, 11
Simmons, B. D., Melvin, T., Lintott, C., et al. 2014, MNRAS, 445, 3466
Simmons, B. D., Lintott, C., Willett, K. W., et al. 2017, MNRAS, 464, 4420
Smethurst, R. J., Lintott, C. J., Simmons, B. D., et al. 2016, MNRAS, 463, 2986
Snyder, G. F., Torrey, P., Lotz, J. M., et al. 2015, MNRAS, 454, 1886
Strateva, I., Ivezić, Ž., Knapp, G. R., et al. 2001, AJ, 122, 1861
Taylor, M. B. 2005, in Astronomical Society of the Pacific Conference Series, Vol. 347, Astronomical Data Analysis Software and Systems XIV, ed. P. Shopbell, M. Britton, & R. Ebert, 29
van den Bergh, S. 1976, ApJ, 206, 883
Watanabe, M., Kodaira, K., & Okamura, S. 1985, ApJ, 292, 72
Whitmore, B. C., Lucas, R. A., McElroy, D. B., et al. 1990, AJ, 100, 1489
Willett, K. W., Lintott, C. J., Bamford, S. P., et al. 2013, MNRAS, 435, 2835
Willett, K. W., Galloway, M. A., Bamford, S. P., et al. 2017, MNRAS, 464, 4176

List of Changes

Added: The primary goal of this paper is to generalize how such a system would work in the context of upcoming surveys like LSST and Euclid., on page 2.

Replaced: ~~In this paper~~ replaced with: As a proof of concept,, on page 2.

Added: , however we briefly discuss how such a method can be implemented for more complex tasks, on page 2.

Replaced: ~~method~~ replaced with: current implementation, on page 2.

Added: The labels we compute from GZ2 vote fractions are used solely to validate our classification method. We thus consider these labels as "ground truth" though, of course, this is subjective. Varying the threshold on which we define 'Featured' from 'Not', or choosing a different type of GZ2 vote fraction will yield slightly different results for the quality metrics we compute throughout this paper. However, we envision this method being applied to never-before-classified image sets and in such a case, no "ground truth" classification would yet exist. Rather than re-classifying GZ2 subjects, we instead endeavour to provide a general sense of the quality of our classifier in the context of the well-known quality of the GZ2 catalogue. In Appendix C we show how different choices of our descriptive GZ2 labels changes the perceived quality of our classification system and demonstrate that our method yields robust galaxy classifications., on page 4.

Replaced: ~~We thus also consider the metrics of accuracy, purity and completeness as a function of GZ2 project time. These are defined as follows: accuracy is the number of correctly identified subjects divided by the total number retired; completeness is the number of correctly identified 'Featured' subjects divided by the number of actual 'Featured' retired; and purity is the number of correctly identified 'Featured' subjects divided by the number of subjects retired as 'Featured'.~~ replaced with: Because we have a binary classification we can construct a confusion matrix from which we can compute the quality metrics of accuracy, completeness and purity as a function of GZ2 project time by comparing our predicted labels to the $GZ2_{raw}$ labels. Figure 5 graphically depicts the elements of this confusion matrix. From this we compute: , on page 6.

Replaced: ~~Thus, a complete sample has no false negatives whereas a pure sample has no false positives.~~ replaced with: Thus, a complete sample recovers *all* subjects labelled 'Featured' by GZ2, whereas a pure sample recovers *only* subjects labelled 'Featured' by GZ2., on page 7.

Added:

C.1. *What about the GZ2 decision tree?*

, on page 7.

Added: In this section we consider the robustness of GZX by computing several sets of "ground truth" labels from the GZ2 catalogue. We recalculate the quality metrics of accuracy, purity, and completeness on the final sample of retired galaxies from the full GZX simulation (SWAP+RF) for a range of thresholds and for each type of GZ2 vote fractions. Specifically, we define labels such that any subject with $f_{\rm featured}+f_{\rm artifact}\geq t$ is labelled 'Featured', otherwise it is labelled 'Not'.

This is performed for a range of threshold values, $t = [0.2, 0.3, 0.4, 0.5]$, and using each type of GZ2 vote fraction: raw, weighted, and debiased. , on page 19.

Added: In Figure C3 we show how the perceived quality of GZX changes as a function of threshold and GZ2 vote fraction type. GZX classifications are quite robust, with accuracy fluctuating by only a couple percent for GZ2 labels computed using a threshold between 0.3 and 0.5. Though accuracy remains relatively flat for this range of thresholds, we see the trade off between purity and completeness. Decreasing the threshold on which labels are computed (i.e., more subjects are given a 'Featured' label) results in higher purity but lower completeness. , on page 19.

Added: That the $\mathrm{GZ2}_{\mathrm{debiased}}$ labels perform poorly is not surprising. These vote fractions are computed after considerable post-processing of the raw volunteer votes in order to remove the effects of redshift and surface brightness. As with any set of visual classifications, these biases must be accounted for and this is traditionally done retrospectively. It is also unsurprising that the $\mathrm{GZ2}_{\mathrm{raw}}$ and $\mathrm{GZ2}_{\mathrm{weighted}}$ classifications are in such tight agreement. $\mathrm{GZ2}_{\mathrm{weighted}}$ vote fractions are computed by down-weighting inconsistent volunteers of which there are relatively few. These two sets of vote fractions are thus very similar. , on page 20.

Added: When applying GZX to future imaging programs, there will be no "ground truth" labels for comparison. In some sense, these thresholds can be interpreted as a prior for the SWAP p_0 value, the initial probability for a subject to be 'Featured'. As we showed in Appendix **??** changing the prior has little affect on the retirement rate but does result in considerable variability in the completeness and purity of the resulting classifications. When applying GZX to a future sample, science teams will need to determine between prioritizing between these which will be dependent on the particular science case. Additionally, , on page 20.