Draft version December 6, 2016
Preprint typeset using LaTeX style AASTeX6 v. 1.0

INTEGRATING HUMAN AND MACHINE INTELLIGENCE IN MORPHOLOGY CLASSIFICATION TASKS WITH GALAXY ZOO EXPRESS

Melanie Beck, Claudia Scarlata, Lucy F. Fortson, Chris J. Lintott, Melanie A. Galloway, Kyle W. Willett, B. D. Simmons, Karen L. Masters, Phil Marshall, Hugh Dickinson, and Darryl Wright
Minnesota Institute for Astrophysics, University of Minnesota, Minneapolis, MN 55454

Department of Physics, University of Oxford, Oxford, UK

Center for Astrophysics and Space Sciences, Department of Physics, University of CaliforniaSan Diego, San Diego, CA

ICG Portsmouth, UK

and

IPAC

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 to the scale problem through an integration of visual and machine classifications that preserves the best features of both human and automatic classifications. 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 machine algorithm that learns on a suite of non-parametric morphology indicators widely used for automated morphologies. Our method provides an order of magnitude improvement in the rate of galaxy morphology classifications while maintaining highly accurate, pure and complete samples, and requires a minimal amount of computational cost. This result has important implications for the development of galaxy morphology identification tasks in the era of Euclid and other large-scale surveys.

Keywords: galaxy morphology — classification — machine learning

1. INTRODUCTION

Astronomers have made use of galaxy morphologies to understand the dynamical structure of these systems since at least the 1930s (e.g., Hubble 1936). The division between early-type and late-type systems corresponds, for example, with a wide range of parameters from mass to environment and star formation history (e.g., Shen et al. 2003; Peng et al. 2010), and detailed observation of morphological features such as bars and bulges provide information about the history of their host systems (e.g., review by Kormendy & Kennicutt 2004; 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 the green peas (Cardamone et al. 2009) and beans (Schirmer et al. 2016).

While Galaxy Zoo has provided a solution that scales visual classification for current surveys (Willett et al. 2013, 2016; Simmons et al. 2016), upcoming surveys such as LSST and Euclid will require a different approach, one appropriate for tackling very large data sets. Standard visual morphology methods will be unable to contend with the scale of data, and current automated morphologies provide only statistical morphological structure with large uncertainties (e.g., Abraham et al. 1996; Bershady et al. 2000).

One solution is full automation, which can be achieved via multiple approaches. One approach is to define a set of features which describe the morphology in an N-dimensional space. Decision boundaries within this space define morphological types for each galaxy. Learning these decision boundaries can be achieved through machine learning techniques such as support vector machines (Huertas-Company et al. 2008) or principal com-

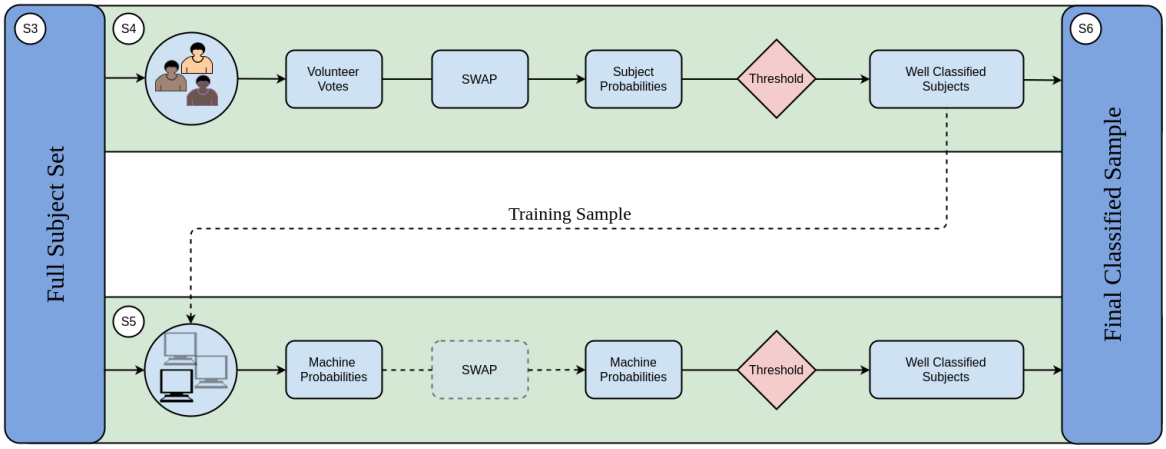

Figure 1. Schematic of our hybrid system. Humans provide classifications of galaxy images via a web interface. We simulate this with the Galaxy Zoo 2 classification dataset described in section 3. Human classifications are processed with an algorithm described in section 4. Subjects that pass a set of 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-retried. The rest remain in the system to be seen by either human or machine. This procedure is repeated on a nightly timestep. Our results are reported in section 6.

ponent analysis (Scarlata et al. 2007). Another approach is through deep learning, a machine learning technique that attempts to model high level abstractions. Algorithms like convolutional neural networks have been used with impressive accuracy (Dieleman et al. 2015; Huertas-Company et al. 2015). However, a drawback to all automated classification techniques is the need for standardized training data, with more complex algorithms requiring more data. Furthermore, that data must be consistent for each survey: differences in resolution and depth can be inherently learned by the algorithm making their application to disparate surveys challenging.

In this work we seek a system that preserves the best features of both human and automatic classifications, developing for the first time a system 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 classifications with a suite of non-parametric morphology indicators widely used for automated morphologies. Our method provides an order of magnitude improvement in the rate of galaxy morphology classifications while maintaining highly accurate, pure and complete samples.

2. GALAXY EXPRESS OVERVIEW

In Galaxy Express (GX) we combine 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 GX which includes section numbers as a shortcut for the savvy 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.

We draw from the Galaxy Zoo 2 (GZ2) classification database which 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 the machine classifier. We have developed a random forest algorithm that trains on measured morphology indicators such as Concentration, Asymmetry, Gini coefficient and M_{20}. After a sufficient number of subjects have been classified by humans, the machine is trained and its performance assessed against a validation sample. This procedure is repeated nightly and the machine's performance increases with size of the training sample, albeit with a performance limit. Once the machine reaches 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 hu-

mans. 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 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 $z < 0.25$[1].

Subjects were shown as color composite images via a web-based interface 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. 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 amount of classification.

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 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.

By combining the "star or artifact" vote fraction, $f_{artifact}$, with the "features or disk" vote fraction, $f_{features}$ we obtain a binary response. Here, a vote fraction is simply the fraction of volunteers who voted for a particular response. We define a label for each GZ2 subject as the majority vote fraction, that is, if $f_{features}+f_{artifact} > f_{smooth}$, the galaxy is labeled 'Featured', otherwise it is labeled 'Not'. We note that only 512 subjects in the GZ2 catalog have a majority $f_{artifact}$, contributing less than half a percent contamination.

The GZ2 catalog assigns every subject three types of volunteer vote fractions: 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 or malicious volunteers, a task we perform as well. However, because our mechanism is entirely different from GZ2, we derive labels from the raw vote fractions ($\mathrm{GZ2_{raw}}$). In total, the data consist of over 16 million classifications from 83,943 individual volunteers.

[1] data.galaxyzoo.org

4. EFFICIENCY THROUGH CLEVER HUMAN-VOTE PROCESSING

Galaxy Zoo 2 had a brute-force subject retirement rule whereby each galaxy was required to achieve a threshold of approximately forty independent classifications. Once the project reached completion, inconsistent and unreliable volunteers were down-weighted as described in Willett et al. (2013). While this process reduces input from any malicious users or 'bots', it doesn't 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 predicted 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 identifying feature 'A', given that the subject actually exhibits feature A; and the probability of correctly identifying the absence of feature A, given that the subject does not exhibit that feature. The agent updates these probabilities by estimating them as

$$P(\text{“}X\text{”}|X,d) \approx \frac{N_{\text{“}X\text{”}}}{N_X} \tag{1}$$

where $N_{\text{“}X\text{”}}$ is the number of classifications the volunteer labeled as type X, N_X is the number of subjects the volunteer has seen that were actually of type X, and d represents the history of the volunteer, i.e., all subjects they have seen.

Each subject is assigned a prior probability that it exhibits feature A: $P(A) = p_0$. When a volunteer makes a classification, C, Bayes' Theorem is used to derive 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 probability is continually updated, nudged higher or lower depending on volunteer input. Probability thresholds can be set such that subjects crossing a threshold are highly likely to exhibit the feature of interest or the absence thereof.

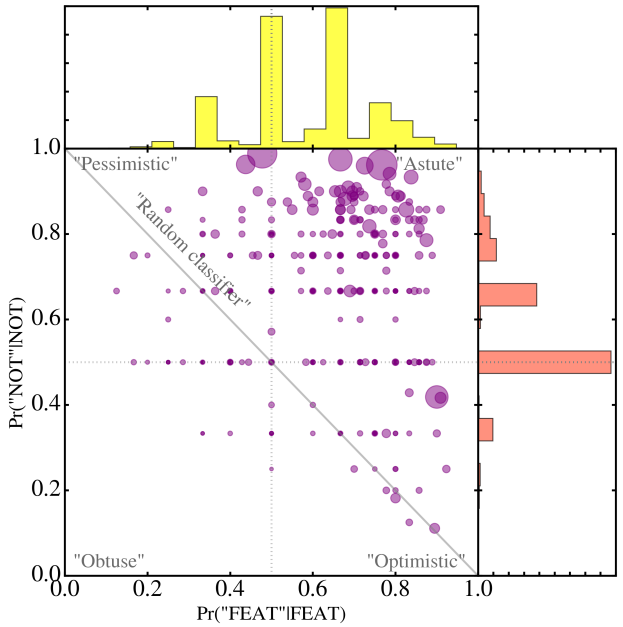

Figure 2. Confusion matrices for 1000 GZ2 volunteers where the size of the circle is proportional to the number of training images that volunteer classified. The histograms on top and right represent the distribution of each component of the confusion matrix for all volunteers. A quarter of GZ2 volunteers are "Astute"; they are adept at correctly identifying both 'Featured' and 'Not' subjects. The peaks at 0.5 in both distributions are an artifact of training: 48% of volunteers see only one training image; only half of their confusion matrix is updated.

These subjects are then considered retired.

4.1. *Volunteer Training 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. 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 which presents a challenge when applying SWAP to GZ2 classifications. We describe how we engineer the GZ2 data to mimic the Space Warps system as closely as possible.

We create a gold standard sample by selecting ∼3500 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) catalog, a collection of ∼14K galaxies classified by eye into T-Types. Expert classifications were obtained through the Zooniverse platform[2] from 15 professional astronomers, including members of the Galaxy Zoo science team. The question posed was identical to the original GZ2 question and at least five experts classified each galaxy. Votes are aggregated and a simple majority provides an expert label for each subject. 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. Classifications of gold standard subjects are always processed first.

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, the subject prior probability, and the retirement thresholds. 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 'Featured' 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 B 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 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 that 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 are generally good at correctly identifying both 'Featured' and 'Not' subjects. 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.

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 30 volunteer votes since not ev-

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

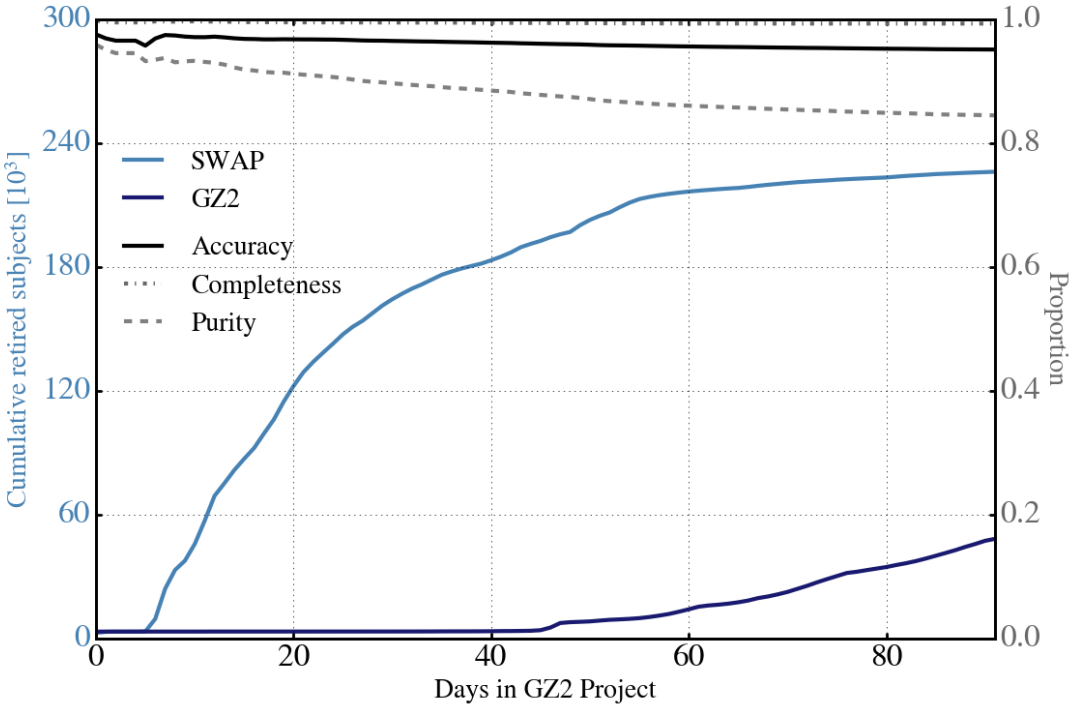

Figure 3. Fiducial SWAP simulation demonstrates a factor of 4-5 increase in the rate of subject retirement as a function of GZ2 project time (left axis, light blue) compared with the original GZ2 project (dark blue). The right axis corresponds to the quality metrics (greys). These are calculated using $GZ2_{raw}$ labels developed in section 3 as "truth" on the subject sample retired by that day of the simulation. Thus, on the final day, over 225K subjects are retired with accuracy of 95.7%, completeness of 99%, and purity of 86.7%. The decrease in purity as a function of time is partially due to the fact that more difficult to classify subjects are retired later in the simulation. Quality metrics can be influenced by careful selection of input SWAP parameters discussed in detail in Appendix B.

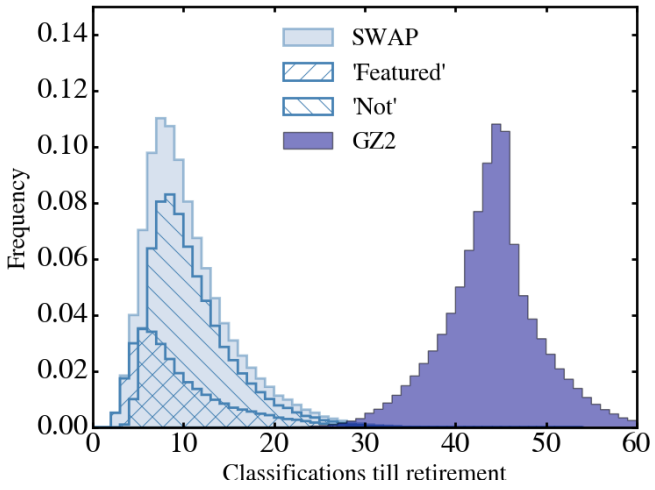

Figure 4. SWAP requires 4-5 times less human effort than GZ2 as evidenced by the distribution of the number of classifications a subject requires for retirement for the ∼225K subjects retired during our fiducial run. As expected, the GZ2 distribution peaks around 45 classifications per subject. In contrast, most subjects need only 9 classifications when processing with SWAP. Furthermore, 'easy' subjects can reach retirement in as few as 3-4 classifications.

ery subject reached the 40 volunteer threshold (Willett et al. 2013). 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. 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'. Thus, a complete sample has no false negatives whereas a pure sample has no false positives. We compute these metrics by comparing the SWAP-assigned labels of the cumulatively retired subject set to the $GZ2_{raw}$ labels for each day of the simulation. For example, by Day 20, SWAP retires 120K subjects with 96% accuracy, 99.7% completeness, and 92% purity.

Figure 3 and Table 1 detail the results of our fiducial SWAP simulation compared to the original GZ2

project. The left axis shows the cumulative number of retired subjects as a function of GZ2 project time. By the end of our simulation, GZ2 retires ∼50K subjects while SWAP retires 226,124 subjects. We thus classify 80% of the entire GZ2 sample in three months. The original GZ2 project took approximately one year to classify as many subjects, representing a factor of four increase in the classification rate. The right axis of Figure 3 corresponds to 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.

There is also a reduction in the human effort required to perform this classification task. Figure 4 shows the distribution of the number of volunteer classifications per subject achieved through SWAP (light blue) and GZ2 (dark blue) for the 226K subjects retired in our fiducial run. GZ2's distribution peaks at ∼45 indicating that, on average, 45 unique volunteers classify each subject. On the other hand, SWAP's distribution peaks around 9 classifications per subject. Furthermore, subjects that are 'easy' to classify (i.e., 'Featured') require even fewer classifications to reach strong consensus. More precisely, SWAP processes 2.3×10^6 volunteer classifications while GZ2 records $\sim 10^7$ for the same subject set. SWAP reduces human effort by more than a factor of four.

As we demonstrate in Appendix B, varying the initial SWAP parameters from the fiducial values does not substantially change the results presented here. The largest influence comes from choosing unrealistic subject prior probabilities which can mildly degrade the quality of the resulting classifications. More importantly, none of these effects significantly alters our human and machine integration.

4.3. *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 $\mathrm{GZ2_{raw}}$ label, SWAP disagrees about 5% of the time. We thus examine the false positives (subjects SWAP labels as 'Featured' but $\mathrm{GZ2_{raw}}$ labels as 'Not') and false negatives (subjects SWAP labels as 'Not' but $\mathrm{GZ2_{raw}}$ labels as 'Featured').

We find the majority of these disagreements are due to uncertainties in the $\mathrm{GZ2_{raw}}$ label. Figure 5 shows the distribution of $f_{features} + f_{artifact}$ for the false positives (blue), and the false negatives (red). Recall that if this value is greater than 0.5, the subject is labeled 'Featured'. The majority of incorrectly labeled subjects have $0.4 \leq f_{features} + f_{artifact} \leq 0.6$, indicating that the GZ2 raw vote fractions are simply too uncertain to provide high quality labels. We note that the distribution overlap is due to subjects that do not have a majority; these are labeled 'Featured' by default.

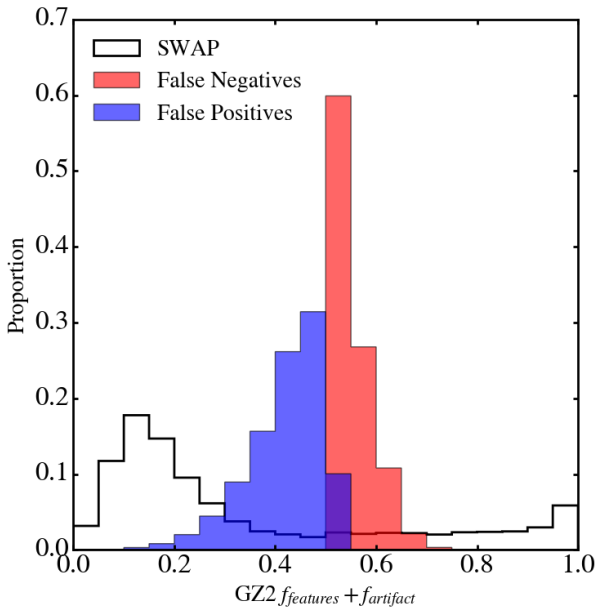

Figure 5. Distribution of GZ2 $f_{features} + f_{artifact}$ vote fractions for subjects identified as false positives (blue), false negatives (red) and the full sample retired by SWAP (solid black line). The proportions are independent for each of the three distributions. From section 3, subjects with values > 0.5 are defined as 'Featured', however, the red distribution indicates that SWAP labels them as 'Not'. This is not a flaw of SWAP: 68.9% of incorrectly identified subjects have $0.4 \leq f_{features} + f_{artifact} \leq 0.6$ suggesting that $\mathrm{GZ2_{raw}}$ labels are simply too uncertain.

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 some cases, these expert-level volunteers disagreed with the majority. These issues can be mitigated by requiring each subject reach a minimum number of classifications before allowing its probability to cross a threshold, thus combining the best qualities of GZ2 and SWAP.

4.4. *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 B). Exploring those subjects wherein SWAP

and GZ2 disagree, we conclude that the majority of this disagreement stems from uncertainty in $GZ2_{raw}$ 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 Express by incorporating supervised learning, the machine learning task of inference from labeled 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 analyzes the training data and produces a function that can be mapped to new examples. An 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 overfitting, precluding them from generalizing well to new data. Random Forests mitigate this effect by combining the output label from a multitude of decision trees. In particular we use the `RandomForestClassifier` from the Python module `scikit-learn` (Pedregosa et al. 2011).

5.2. *Cross-validation*

Of fundamental importance is the task of choosing an algorithm's hyperparameters, values which determine how the machine learns. In the case of a RF, these include the maximum depth of the tree, the minimum leaf size, the maximum number of leaf nodes, among others. 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 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 is 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 hyperparameter space and values that optimize the output are chosen.

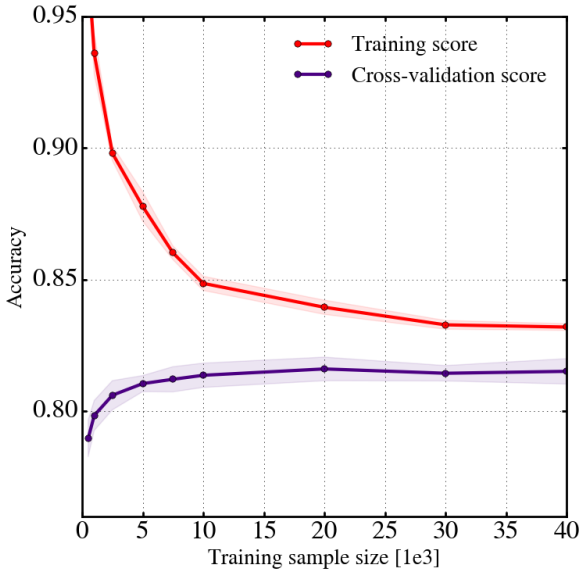

Figure 6. Learning curve for a Random Forest with fixed hyperparameters. The training score is the accuracy of the trained machine applied to its own training sample. The cross-validation score is the accuracy of the machine computed during the cross-validation process. When the training sample size 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 behavior plateaus indicating that larger training sample sizes provide little in additional performance.

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 M_{20}, the second-order moment of light for the brightest 20% of galaxy pixels (Appendix A details the measurement process). Coupled with SExtractor's measurement of ellipticity (Bertin & Arnouts 1996), we provide the machine with a five dimensional morphology parameter space. These non-parametric diagnostics have long been used to quantify galaxy morphology in an automated fashion (e.g., Abraham et al. 1996; Bershady 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 preprocessing 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 simulaton, 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 6 demonstrates such a curve wherein we plot the accuracy from both the k-fold cross-validation, and the trained machine applied to its own training sample. 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*

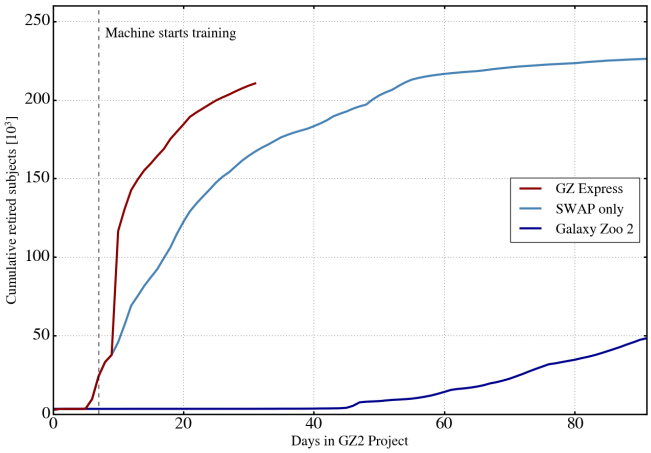

Figure 7. Galaxy Express reduces the classification rate by an order of magnitude after incorporating a machine classifier, retiring more than 200K subjects in just 27 days. The dashed line marks the first night the machine trains. After several additional nights of training, it is deemed optimized and 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.

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 prediction, 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 with $p_{machine} \geq 0.9$ 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 7 where subject retirement with GZX (red) is compared to our fiducial SWAP-only run (light blue) and GZ2 (dark blue). Using the $GZ2_{raw}$ labels as before, we compute our usual quality metrics on the full sample of GZX-retired subjects. 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 behavior 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. GZX thus provides an order of magnitude increase in the rate of classification 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 8 we explore the individual contributions to GZX subject retirement from the RF (blue) and SWAP (red). The solid black line shows the total GZX retirement (SWAP+RF), while the dashed 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, as we presumed. 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.

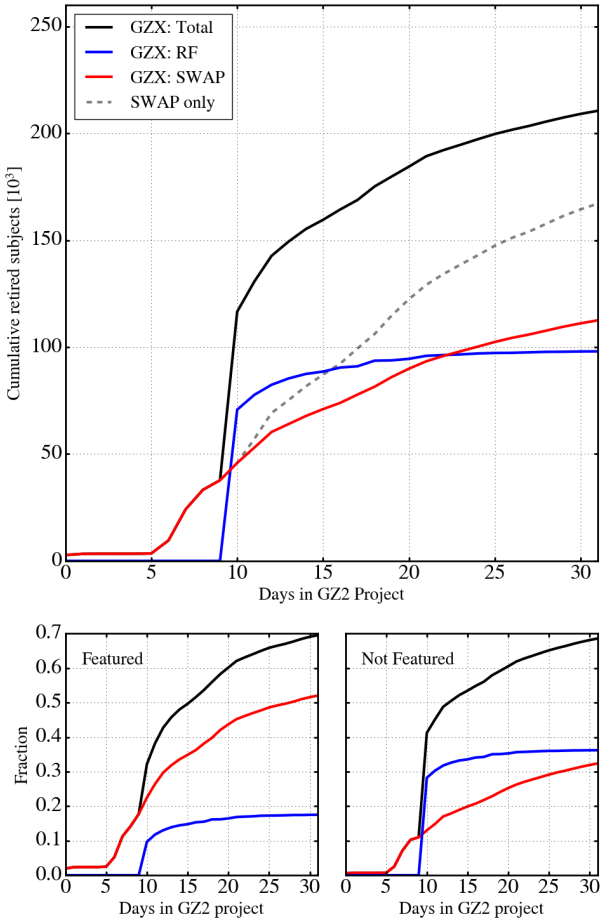

Figure 8. The top panel demonstrates that the relative contributions to subject retirement are nearly equal between the human (red) and machine (blue) components but display different behaviors over the course of the simulation with the dashed line showing the fiducial SWAP-only run for comparison. The bottom panels shows what fraction of GZ2 subjects are retired, separated by class label. Overall, GX retires 73.6% of the entire GZ2 sample in 32 days distributed evenly between 'Featured' and 'Not' 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.

In the bottom panels of Figure 8, 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' sub-

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

jects 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 artifact 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. We argue that the machine classifications are, in fact, valid and complimentary to human classifications.

We visually examine several hundred subjects that were deemed false positives, i.e., galaxies retired as 'Featured' by the machine which have 'Not' $GZ2_{raw}$ labels. Figure 9 shows a random sample of images we inspected. The value in the lower left corner is the raw GZ2 smooth vote fraction, f_{smooth}; the fraction of volunteers who classified that subject as 'Not'. The sample consists primarily of edge-on disks and disk galaxies with low surface brightness features.

That the machine can identify 'Featured' galaxies that humans classify as 'Not' has two contributing factors: 1) the first task of the GZ2 decision tree asks a very specific question that does not necessarily correlate with a split between early- and late-type galaxies, and 2) the machine learns on morphology diagnostics that are very different from visual inspection. Regarding the first point, we see that f_{smooth} is generally between 0.5 and 0.65 for this sample, indicating that volunteers have not reached a strong consensus. This behavior could be modified by providing actual training images and live feedback as performed in Marshall et al. (2016). The second point suggests that the morphology indicators we measure are sufficient for the machine to recognize 'Featured' galaxies regardless of the labels humans provide. It remains to be seen whether this generalizes to other machine algorithms. We note that most of these galaxies are labeled 'Featured' after GZ2's debiasing process.

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.

GX is perhaps one of the simplest ways to combine human and machine intelligence and its impressive performance motivates a higher level of sophistication. Towards this end, 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, 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 such as deep convolutional neural networks (CNN) or Latent Dirichlet allocation (LDA), an algorithm that is frequently used in document processing, should be considered. 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 with our current implementation we can cope with upcoming data volumes

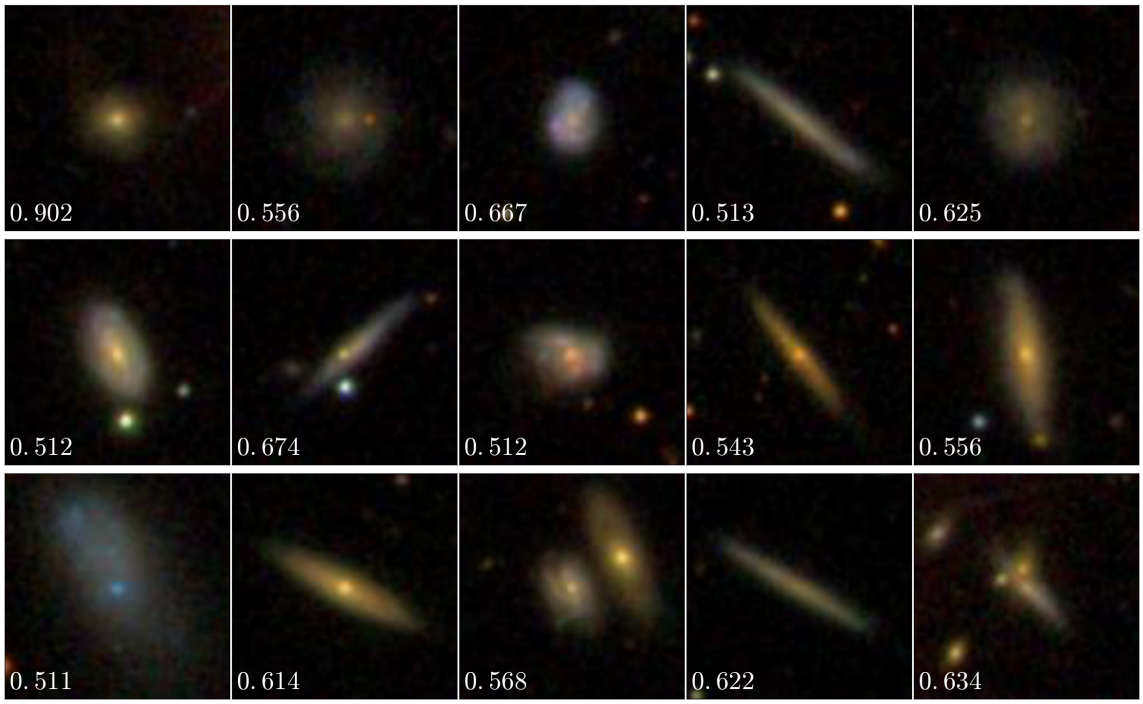

Figure 9. A subsample of galaxies labeled 'Not' according to $GZ2_{raw}$ labels but which the machine identifies as 'Featured'. We display f_{smooth} in the lower left corner, that is, the fraction of volunteers who classified the subject as 'smooth' ('Not'). Values are typically between 0.5 and 0.65 indicating that GZ2 does not reach a strong consensus. Fortunately, the machine is able to identify these subjects as 'Featured' due to their measured morphology diagnostics.

from large surveys. By some estimates, Euclid is expected to obtain measurable morphology with VIS for approximately $10^6 - 10^7$ galaxies. Traditional visual classification at the rate achieved today through Galaxy Zoo would require approximately 300 years to classify. GZX as currently implemented could classify the brightest 20% of the Euclid sample during the six years of its observing mission. We predict this is a lower limit as more sophisticated machine learning algorithms will undoubtedly be capable of further efficiency.

7.1. *Conclusions*

We test an innovative system 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 simple 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 rate of classification allowing us to retire over 70% of GZ2 galaxies in just 32 days of GZ2 project time, representing an order of magnitude increase in the classification rate 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 which humans miss, providing a complementary approach to visual classification. The gain in classification speed allows us to tackle the massive amounts of data soon to be forthcoming from large surveys like LSST, Euclid, and WFIRST.

MB thanks John Wallin, Steven Bamford, and Boris Häußler for discussions which helped elucidate things and stuff, 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, Karen Masters, Steven Bamford, Kevin Schawinski, and Rebecca Smethurst for providing expert classifications in addition to those provided by the authors.

MB, CS, LF, KW, and MG gratefully acknowledge support from the US National Science Foundation Grant

AST1413610. MB acknowledges additional support through New College and Oxford University's Balzan Fellowship as well as the University of Minnesota's Doctoral Dissertation Fellowship. Travel funding was supplied to MB, in part, by the University of Minnesota's 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 issued 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.

APPENDIX

A. MEASURING NONPARAMETRIC MORPHOLOGICAL DIAGNOSTICS ON SDSS STAMPS

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. Large values of this ratio tend to indicate disky galaxies, while smaller 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 background noise is applied (as in e.g. Conselice et al. (2000)), i.e.,

$$A = \frac{\sum_{x,y} |I - I_{180}|}{2\sum |I|} - B_{180} \tag{A1}$$

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

$$G = \frac{1}{|\bar{X}|n(n-1)} \sum_{i}^{n} (2i - n - 1)|X_i| \tag{A2}$$

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

$$M_{tot} = \sum_{i}^{n} f_i[(x_i - x_c)^2 + (y_i - y_c)^2] \tag{A3}$$

$$M_{20} = \log_{10}(\frac{\sum_i M_i}{M_{tot}}), \quad \text{while} \sum_i fi < 0.2 f_{tot} \tag{A4}$$

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-center knots of light.

Finally, we use the ellipticity, $\epsilon = 1 - b/a$, of the light distribution as measured by SExtractor which computes the semimajor axis a and semiminor 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 Figure A1. The code developed to clean and compute these morphology indicators is open source and can be found at `https://github.com/melaniebeck/measure_morphology`.

B. 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 with confusion matrix $(P_{F,0}, P_{N,0}) = (0.5, 0.5)$, which presumes that volunteers are no better than random classifiers. We perform two

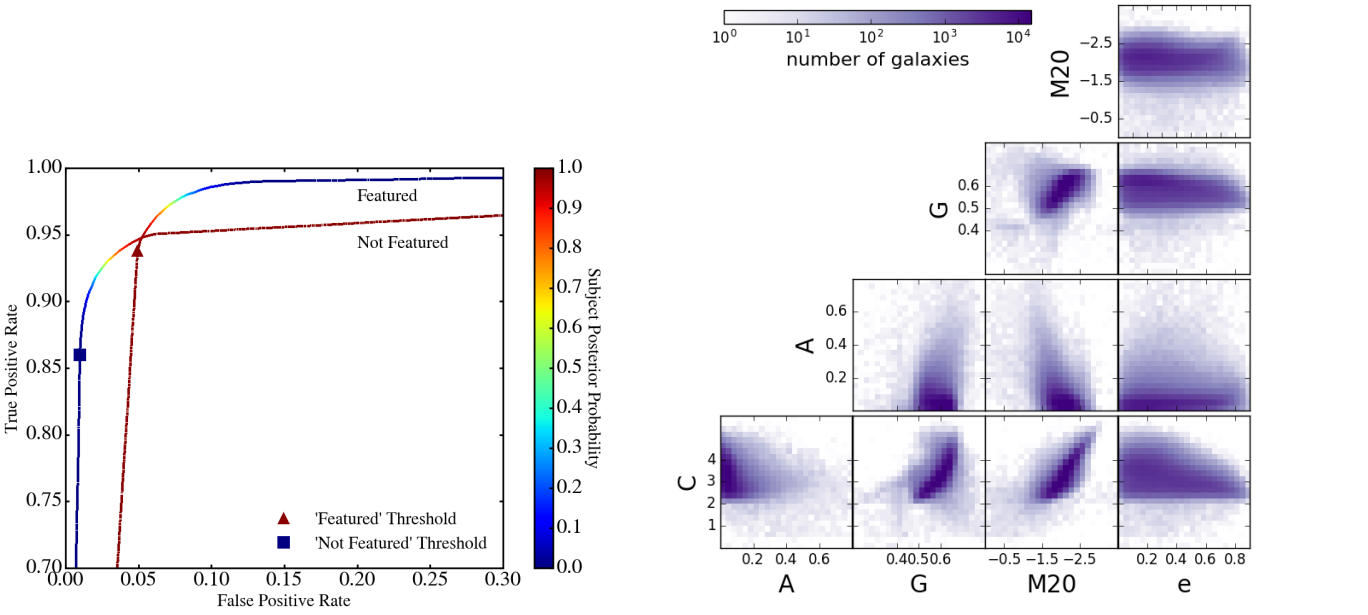

Figure A1. *Left.* Identifying 'Featured' subjects is independent of identifying 'Not' subjects. Both ROC curves use all subjects processed by SWAP where the score used to create the ROC curve is simply each subject's achieved posterior probability. The Featured curve demonstrates how well we identify '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 with scores that lie closest to the upper left corner. Our 'Featured' threshold is nearly as optimumal as possible though our 'Not' threshold could have been slightly better. *Right.* Relation between morphology diagnostics measured for more than 280K SDSS galaxies classified during the GZ2 project.

simulations wherein we allow $(P_{F,0},\, P_{N,0}) = (0.4, 0.4)$, slightly obtuse volunteers, and $(P_{F,0},\, P_{N,0}) = (0.6, 0.6)$, slightly astute volunteers with everything else remaining constant. Results of these simulations compared to the fiducial run are shown in Figure B2. We find that we are largely insensitive to the initial agent confusion matrix probabilities both in terms of the overall number of retired subjects and in the quality of their SWAP labels.

Predictably, when $(P_{F,0},\, P_{N,0})$ are low, we retire fewer subjects in the same time frame and more subjects when $(P_{F,0},\, P_{N,0})$ are high. This is easy to understand since it takes longer for volunteers to become strong, astute classifiers when they are initially given values denoting them as obtuse. Even with this handicap, most volunteers become astute classifiers by the end of the simulation. Overall, we retire $\sim$ 225K $\pm$3.5% subjects as shown by the light blue spread in the bottom panel of Figure B2 where the dashed blue line denotes the fiducial run.

The top panel depicts the same quality metrics computed before where the dashed lines again denote the fiducial run. The spread is within a couple per cent for any metric. Overall we maintain accuracy around 95%, as well as completeness of 99% while maintaining purity around 84%. This spread can be due to three different effects: 1) classifying a different subset of subjects, 2) retiring subjects in a different order, and 3) subjects acquiring a different SWAP label in different simulations.

We find that SWAP is exceptionally consistent. Of all the subjects retired in these runs, we find that over 99% of them are the same subjects between simulations. Of those consistent between runs, we find that SWAP gives the same label for more than 99% of the subjects. What changes between runs is the order in which subjects are classified. In the low $(P_{F,0},\, P_{N,0})$ run, subjects take longer to classify compared to the fiducial run (i.e., they retire on a later date in GZ2 project time). Subjects in the high $(P_{F,0},\, P_{N,0})$ run retire earlier in GZ2 project time. This can cause a variation in accuracy, completeness or purity because these values are calculated on a day to day basis; if we're working with a slightly different make-up of subjects on a given day, we can expect to compute different values for these metrics. These effects each contribute less than one per cent variation and thus we see a high level of consistency between these simulations.

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 to simply find galaxies that

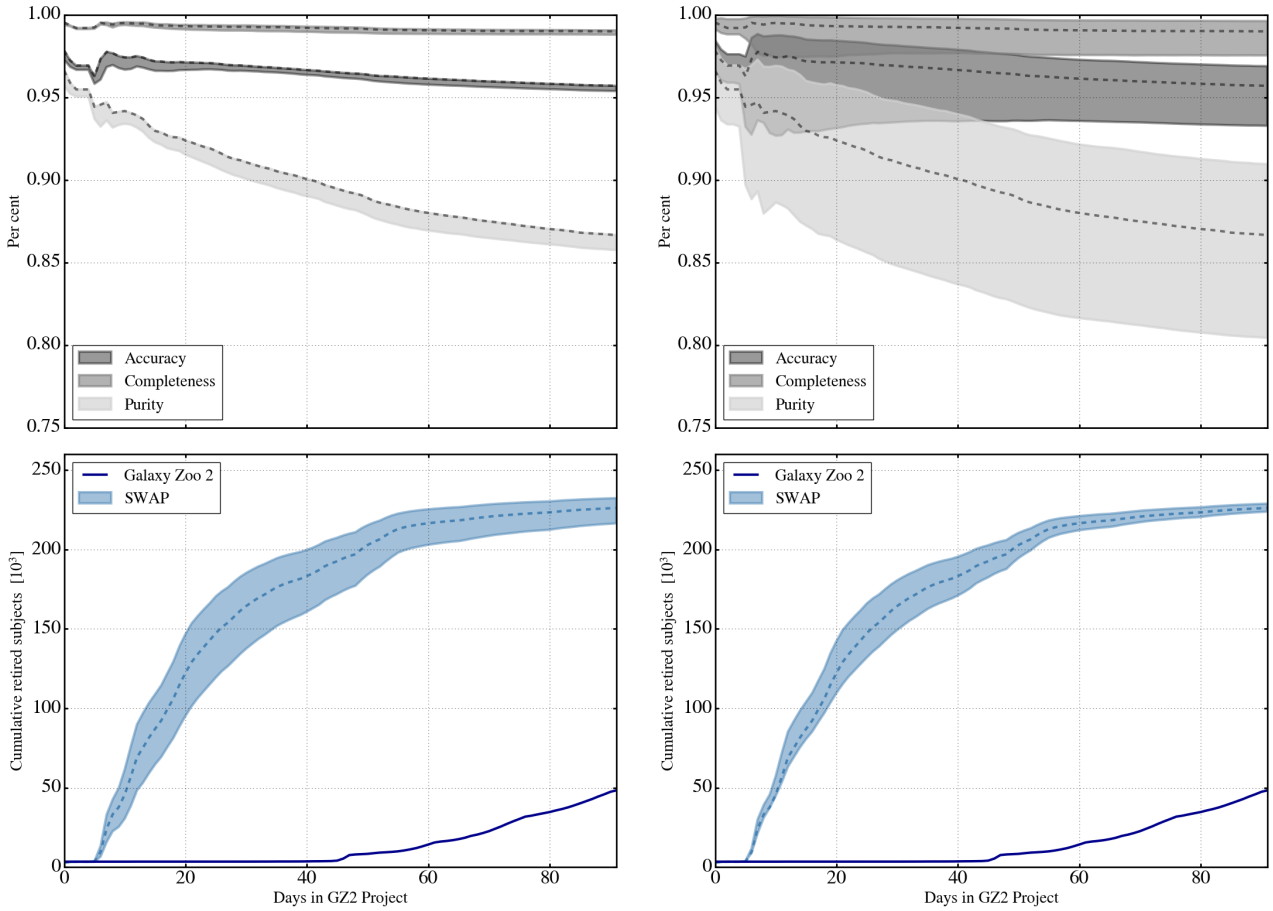

Figure B2. SWAP output does not dramatically change for a range of SWAP parameters. *Left.* The quality (top) and retirement rate (bottom) when different values of agent confusion matrix are assigned. *Right.* Same as the left panel but for various values of the subject prior probability. Changing the confusion matrix has little impact on the quality of the labels but varies the total number of subjects retired during the run. In contrast, changing the subject prior is more likelyt o affect the quality of the retirement labels rather than the total number retired.

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 the galaxy size (with the exception that each galaxy be larger than the SDSS PSF), the sample includes a large range of galaxy masses and sizes. Thus, designating a single prior is not clear-cut. We thus explore how various p_0 affect the SWAP outcome.

We run several simulations where p_0 is allowed to take values 0.2, 0.35, and 0.8 and compare these to the fiducial run where $p_0 = 0.5$, everything else remaining constant. The results are shown in Figure **??**, where again we find that SWAP is consistent in terms of the total number of subjects retired during the simulation which varies by only 1%. However, as can be seen in the top panel, the variation in our quality metrics is more pronounced and deserves some discussion.

Firstly, though we are retiring nearly the same number of subjects over the course of each simulation, they are less consistent than our previous simulations. That is, only 95% of the subjects are common to all runs. Secondly, of those that are common, only 94% receive the same label from SWAP. 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 in the day on which a subject is retired when compared to the fiducial run, being nearly equally likely to retire 'late' or 'early' regardless of p_0. These trends all contribute to a broader spread in accuracy, completeness, and purity as a function of project time. We stress, however, that though more substantial than the previous comparison, these variations are all within $\pm 5\%$.

We can get a handle on these variations more intuitively by considering the following. Recall that our retirement thresholds, t_F and t_N, have not changed in these simulations. Thus when p_0 is small, the subject probability is already

closer to t_N, and 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 the prior probability is already closer to t_F. Obviously, both outcomes cannot be correct and we find that the simulation with $p_0 = 0.8$ performs the worst of any run which is a direct reflection of the fact that this prior is not suitable for this question nor for this dataset. For the mass, size, and redshift range of subjects in GZ2, we would not expect that 80% of them are 'Featured'. Indeed, the best performance is achieved when p_0= 0.35. This reflects the actual distribution of 'Featured' subjects in the GZ2 sample as well as being similar to the expected proportion of 'Featured' galaxies in the local universe, depending on your definition. Thus, the take-away here is to choose your prior wisely since a value far from the correct value can have a significant impact on the quality of your classifications.

Retirement thresholds. Retirement thresholds are directly related to the time that a subject will spend in SWAP before retiring. 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 bounce back and forth 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. Additionally, this will also increase the likelihood of some subjects never crossing either threshold as there are always some which are nudged back and forth indefinitely through probability space.

What thresholds should one choose? To answer this question, we consider Figure A1 which depicts the receiver operating characteristic (ROC) curve for our fiducial simulation. The solid line shows the curve when considering the 'Featured' threshold while the dotted line corresponds to 'Not'. The square and triangle represent our thresholds, $(t_F, t_N) = (0.99, 0.004)$, on that curve at the end of the simulation. We see that t_F is nearly optimal but t_N could be improved upon.

Throughout this discussion we have computed quality metrics under the assumption that the 'true' labels provided by GZ2 are accurate. This is unlikely to be the case for every subject as Willett et al. (2013) explicitly caution against using the majority volunteer vote fraction as label since some of these are highly uncertain. We now turn to a brief discussion of those subjects where SWAP and GZ2 do not agree.

REFERENCES

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

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

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

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

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

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

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

—. 2006, MNRAS, 373, 1389

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

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

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

Hubble, E. P. 1936

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

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

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

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

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

More, A., Verma, A., Marshall, P. J., et al. 2016, MNRAS, 455, 1191

Nair, P. B., & Abraham, R. G. 2010, ApJS, 186, 427

Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2011, Journal of Machine Learning Research, 12, 2825

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

Scarlata, C., Carollo, C. M., Lilly, S., et al. 2007, ApJS, 172, 406

Schirmer, M., Malhotra, S., Levenson, N. A., et al. 2016, MNRAS, 463, 1554

Shen, S., Mo, H. J., White, S. D. M., et al. 2003, MNRAS, 343, 978

Simmons, B. D., Melvin, T., Lintott, C., et al. 2014, MNRAS, 445, 3466

Simmons, B. D., Lintott, C., Willett, K. W., et al. 2016, ArXiv e-prints, arXiv:1610.03070

Snyder, G. F., Torrey, P., Lotz, J. M., et al. 2015, MNRAS, 454, 1886

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. 2016, ArXiv e-prints, arXiv:1610.03068