Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY

MNRAS **476,** 5516–5534 (2018) doi:10.1093/mnras/sty503
Advance Access publication 2018 March 6

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] B. D. Simmons,[2,4]† Melanie A. Galloway,[1] Kyle W. Willett,[1] 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*
[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, Dennis Sciama Building, Burnaby Road, Portsmouth, PO1 3FX UK*
[6]*Kavli Institute for Particle Astrophysics and Cosmology, P.O. Box 20450, MS29, Stanford, CA 94309, USA*

Accepted 2018 February 21. Received 2018 February 4; in original form 2016 December 15

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 d of GZ2 project time while reproducing labels derived from GZ2 classification data with 95.7 per cent 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 at least a factor of 8 increase in the classification rate, classifying 210 803 galaxies in just 32 d of GZ2 project time with 93.1 per cent 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.

Key words: methods: data analysis – methods: statistical – galaxies: statistics – galaxies: structure.

1 INTRODUCTION

Astronomers have made use of visual galaxy morphologies to understand the dynamical structure of these systems for nearly 90 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. reviews by Kormendy & Kennicutt 2004; Elmegreen, Bournaud & Elmegreen 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

★ E-mail: melaniebeck81@gmail.com
† Einstein Fellow

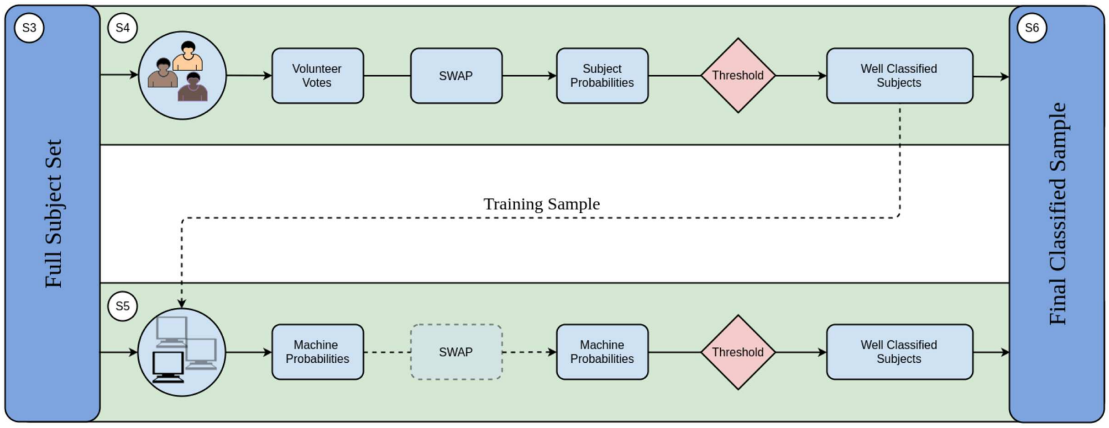

Figure 1. Schematic of our hybrid system. Humans provide classifications of galaxy images via a web interface. We simulate this with the Galaxy Zoo 2 classification data described in Section 2. Human classifications are processed with an algorithm described in Section 3. Subjects that pass a set of thresholds are considered human-retired (fully classified) and provide the training sample for the machine classifier as described in Section 4. 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 5.

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 per cent of this imaging, we will have millions of images to contend with. A project of this magnitude would take more than 60 yr 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; Abraham, van den Bergh & Nair 2003; Conselice 2003; Lotz, Primack & Madau 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, Jangren & Conselice 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, Willett & Dambre 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 (RF) machine learning algorithm that trains on a suite of non-parametric morphology indicators widely used for automated morphologies. 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. As a proof of concept, we focus on the first question of the Galaxy Zoo decision tree. We demonstrate that our current implementation provides at least a factor of 8 increase in the rate of galaxy morphology classification while maintaining at least 93.5 per cent 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.

1.1 Galaxy Zoo Express overview

The Galaxy Zoo Express (GZX) framework combines human and machine to increase morphological classification efficiency, in terms of both the classification rate and required human effort. Fig. 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 6. 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) crowd-sourced classifications as described in Section 2. We compute 'ground truth' labels for each galaxy in the GZ2 sample from the published GZ2 classification catalogue

(Section 2.1). The GZ2 data allow us to create simulations of human classifiers whose classifications are used most effectively when processed with SWAP, a Bayesian code first developed for the Space Warps gravitational lens discovery project (Marshall et al. 2016) and described in Section 3 provide the machine's training sample. SWAP aggregates the crowd-sourced classifications of galaxy images (hereafter, *subjects*) producing a final label for each subject (Section 3.3). We show that SWAP produces significant gains in classification efficiency as well as a reduction of human effort in Sections 3.4 and 3.5. In Section 3.6 we compare these labels to the 'ground truth' labels computed from GZ2's traditional crowd-sourced classification method. Subjects classified by SWAP then provide the machine's training sample.

In Section 4, we incorporate a machine classifier. We develop an RF 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. Section 4.4 discusses the decision engine we develop that delegates tasks between human classification and the RF. After a sufficient number of subjects have been classified by humans via SWAP, 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 an acceptable level of performance, it is applied to the remaining galaxy sample as explored in Section 4.5.

The results of our combined GZX system are provided in Section 5. 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. This result is detailed in Section 5.1. Furthermore, in Section 5.2, we find evidence that the RF may be capable of correctly identifying subjects that humans miss providing a complimentary approach to galaxy classification. 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. Discussion and conclusions are presented in Section 6.

2 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 per cent (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 disc?' to which possible responses include 'smooth', 'features or disc', or 'star or artefact'. This choice serves two purposes: (1) this is one of only two questions in the GZ2 decision tree that is asked about 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 artefact' responses with 'features or disc' responses.

2.1 'Ground truth' labels

We assign each subject a descriptive label in order to validate our classification output against that of GZ2. GZ2 classifications are composed of volunteer vote fractions for each response to every task in the decision tree, denoted as $f_{\rm response}$. The most basic of these is computed simply as $f_{\rm r} = n_{\rm r}/n_{\rm t}$, that is, the number of votes of response r divided by the total number of votes for task t. Vote fractions 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 GZ2 catalogue have a majority $f_{\rm artifact}$, contributing less than half a percent contamination when combining the 'star or artefact' with 'features or disc' 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 thus derive labels from the simple 'raw' vote fractions defined above, and designate the resulting labels as $\rm GZ2_{raw}$. In total, the data consist of over 14 million classifications from 83 943 individual volunteers.

The $\rm GZ2_{raw}$ labels we compute from GZ2 vote fractions are used solely to validate our classification method and are thus considered 'ground truth', though this is, of course, subjective. Furthermore, we envision our framework being applied to never-before-classified image sets for which 'ground truth' labels would not yet exist. Nevertheless, in Appendix A we show how different choices of our descriptive GZ2 labels change the perceived quality of our classification system and demonstrate that our method yields robust galaxy classifications.

3 EFFICIENCY THROUGH INTELLIGENT HUMAN-VOTE AGGREGATION

Galaxy Zoo 2 did not have a predictive retirement rule, rather each galaxy received a median of 44 independent classifications. Once

[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.

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 an overview here; interested readers are encouraged to refer to Marshall et al. (2016) for additional details.

3.1 The SWAP algorithm

SWAP evaluates the accuracy of individual classifiers based on their responses to subjects where the true classification is known, and applies those evaluations to the consensus classifications of subjects where the true classification is unknown in order to improve classification efficiency and reduce the classification effort required to complete a project. In order to achieve this, SWAP 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 to correctly identify feature A given that the subject exhibits feature A; and the probability to correctly identify 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 normalized 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; while 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.

3.2 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 engagement 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, 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 new expert labels for these galaxies that are consistent with the labels we defined for GZ2 classifications. These are provided by 15 professional astronomers, including members of the Galaxy Zoo science team, through the Zooniverse platform.[4] 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 data set 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.

3.3 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 '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 $\Delta t = 1$ d, during which SWAP processes all volunteer classifications with timestamps within that range. This is performed for three months worth of GZ2 classification data. Hereafter, we refer to this as *GZ2 project time* where 0 marks the first day of the original GZ2 project.

Fig. 2 (adapted from Fig. 4 of Marshall et al. 2016) demonstrates the volunteer assessment we achieve at the end of our simulation, and shows confusion matrices for 1000 randomly selected volunteers. The circle size is proportional to the number of gold standard subjects each volunteer classified. If we were to examine this figure immediately prior to the start of classifications, it would show all points as small circles stacked precisely at the centre of the figure since each volunteer is initially assigned a confusion matrix of (0.5, 0.5). As the simulation progresses, each volunteer's green circle is updated in both location and

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

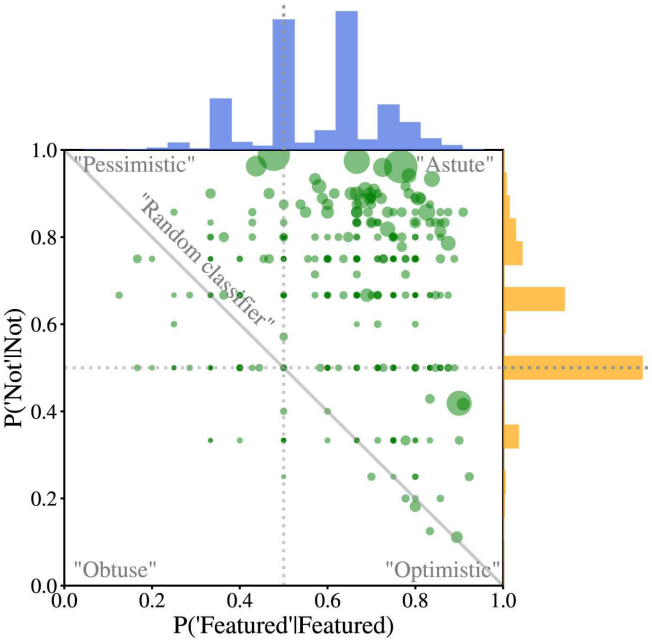

Figure 2. Confusion matrices for 1000 randomly selected GZ2 volunteers after fiducial SWAP assessment. Circle size is proportional to the number of gold standard subjects each volunteer classified. The histograms on top and right represent the distribution of each component of the confusion matrix for all volunteers. A quarter of GZ2 volunteers are 'Astute: they correctly identify both 'Featured' and 'Not' subjects more than 50 per cent 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.

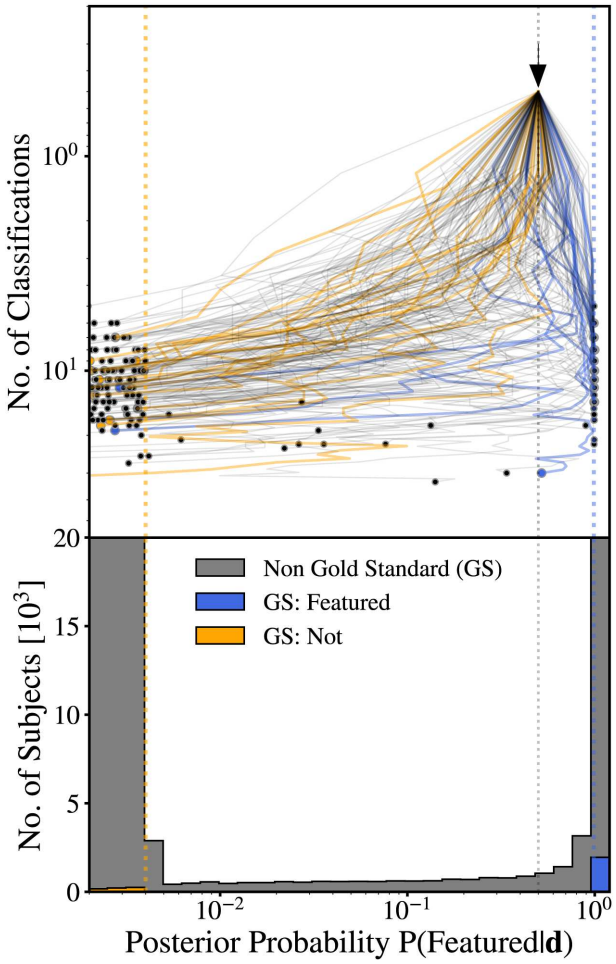

Figure 3. Posterior probabilities for GZ2 subjects. The top panel depicts the probability trajectories of 200 randomly selected GZ2 subjects. All subjects begin with a prior of 0.5 denoted by the arrow. Each subject's probability is nudged back and forth with each volunteer classification. From left 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.

size according to their assessment of gold standard subjects until arriving at the figure shown here. The histograms represent the distribution of each component of the confusion matrix for all volunteers. Nearly 25 per cent of volunteers are considered 'Astute' indicating they correctly identify both 'Featured' and 'Not' subjects more than 50 per cent of the time. Furthermore, as long as a volunteer's confusion matrix is different from a random classifier, 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 per cent 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.

Fig. 3 (adapted from fig. 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 blue and orange 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: of all subjects that SWAP 'sees', i.e. processes at least one classification, only eight per cent have not reached retirement by the end of our simulation.

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 GZ2 project time. We define a subject as GZ2-retired once it achieves at least 30 volunteer votes, encompassing 98.6 per cent of GZ2 subjects (this definition is quantified and its implications are explored in Section 3.4). 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. 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. Fig. 4 graphically ascribes semantic interpretations for the elements of this

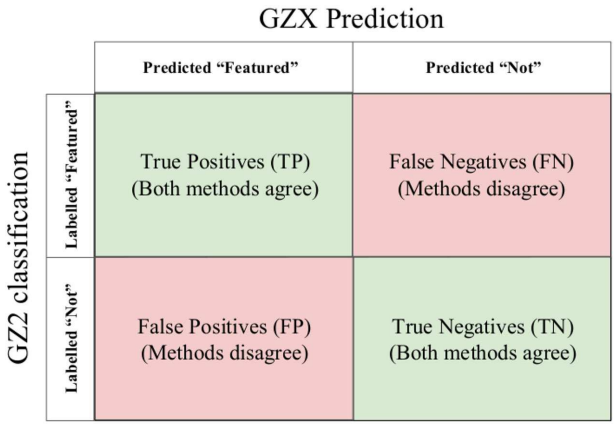

Figure 4. Confusion matrix for comparing our method to GZ2 which we consider to be 'ground truth' as discussed in Section 2.1. True positives (TP) and true negatives (TN) indicate that the predictions from our method agree with GZ2 for subjects labelled 'Featured' and 'Not', respectively. When the two classification methods disagree, the result is a sample of false negatives (FN) and false positives (FP). This allows us to easily compute quality metrics like accuracy, completeness, and purity with respect to GZ2 as shown in equations (3).

confusion matrix. From these we compute:

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

Thus a 100 per cent complete sample recovers *all* subjects labelled 'Featured' by GZ2, whereas a 100 per cent pure sample recovers *only* subjects labelled 'Featured' by GZ2. For example, by Day 20, SWAP retires 120K subjects with 96 per cent accuracy, 99.7 per cent completeness, and 92 per cent purity.

Fig. 5 and Table 1 detail the results of our fiducial SWAP simulation ('SWAP only') 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 per cent of the entire GZ2 sample in three months. Processing volunteer classifications through SWAP presents nearly a factor of 5 increase in classification efficiency. The top panel of Fig. 5 demonstrates the quality of those classifications as a function of time and establishes that our full SWAP-retired sample is 95.7 per cent accurate, 99 per cent complete, and 86.7 per cent pure. We discuss these small discrepancies in Section 3.6.

3.4 Intelligent subject retirement

That SWAP achieves a classification rate nearly 5 times faster than GZ2 comes with a caveat: we consider only the top-level question of the GZ2 decision tree. It can be argued that GZ2 did not need ∼40 votes per subject to achieve exquisite sampling for the top-level question but rather adequate sampling for the subqueries. It might therefore be the case that the top-level question could be accurately resolved with far fewer classifications. In order to put SWAP and GZ2 on equal footing, we determine the minimum number of votes, N, that the GZ2 project would need in order to replicate the original

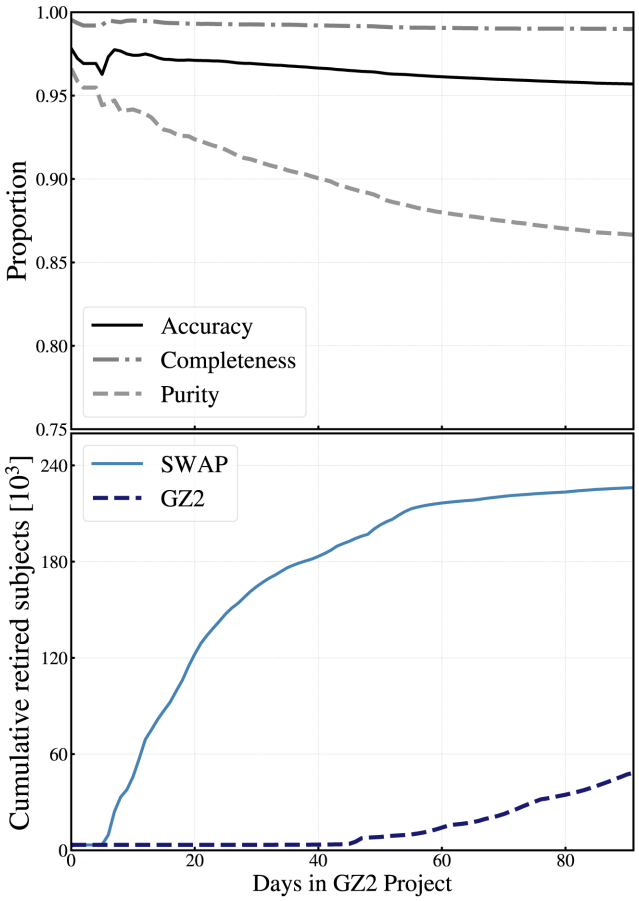

Figure 5. Fiducial SWAP simulation demonstrates a factor of 4.7 increase in the rate of subject retirement as a function of GZ2 project time (bottom panel, light blue) compared with the original GZ2 project (dashed dark blue). After 92 d, SWAP retires over 226K subjects, while GZ2 retires ∼48K. The top panel displays the quality metrics (greys). These are calculated by comparing labels predicted by SWAP to $GZ2_{raw}$ labels (Section 2) for the subject sample retired by that day of the simulation. Thus, on the final day, SWAP retires 226 124 subjects with 95.7 per cent accuracy, and with completeness and purity of 'Featured' subjects at 99 per cent and 86.7 per cent, 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 (see Section 3.4).

GZ2 outcome for the top-level classification task for a canonical 95 per cent of its sample.

We compute the raw vote fractions (f_{featured}, f_{smooth}, and f_{artifact}) for every subject in the GZ2 sample using only the first N classifications for $N \in [10, 15, 20, 25, 30, 35]$. From this, we compute descriptive labels as described in Section 2.1. Our SWAP simulation did not retire every subject in the GZ2 sample. We therefore select 100 random subsamples each consisting of 226 124 subjects, and compute the accuracy and the total number of GZ2 classifications necessary to retire each subsample. These results are shown in the bottom panel of Fig. 6 for each value of N along with the accuracy and total classifications for our SWAP simulation. We see that GZ2 needs at least 35 votes per subject in order to achieve consistent class labels 95 per cent of the time, a full 3.5 times more classifications than SWAP needs to achieve the same accuracy. Furthermore, this justifies our choice of defining a subject as GZ2-retired once it reaches at least 30 classifications.

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

	Days	Subjects retired	Human effort (classifications)	Accuracy (per cent)	Purity (per cent)	Completeness (per cent)
Galaxy Zoo 2	430	285 962	14 144 142	–	–	–
SWAP only	92	226 124	2 298 772	95.7	86.7	99.0
SWAP+RF	32	210 803	936 887	93.1	83.2	94.0

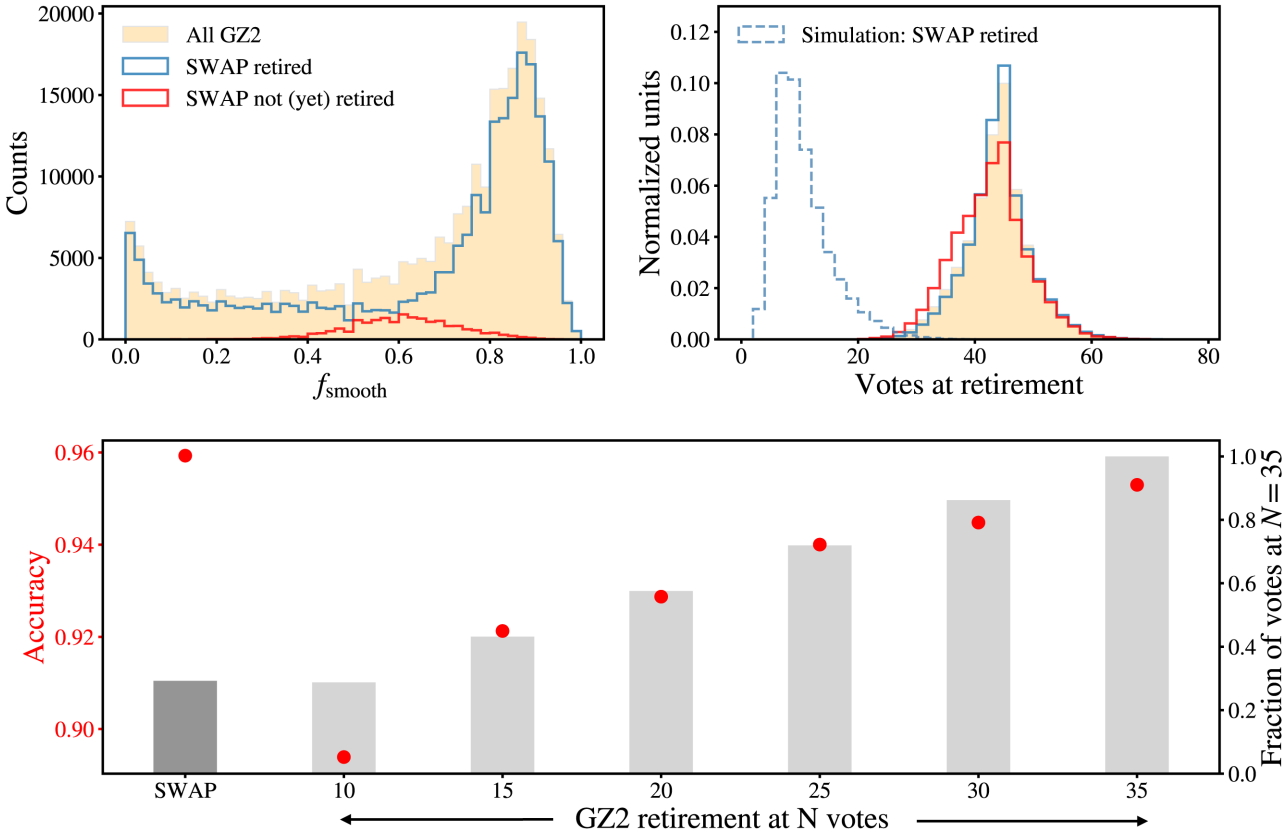

Figure 6. SWAP's intelligent retirement mechanism requires only 30 per cent of the classifications that GZ2 needs for the top-level question due to SWAP's ability to retire easier subjects quickly, while more difficult subjects remain in the system to accrue additional classifications. *Top panels:* The top left-hand panel shows f_{smooth} for the entire GZ2 sample (orange), the subjects retired by SWAP (blue), and subjects that SWAP has not yet retired by the end of our simulation (red). The latter distribution peaks at $f_{smooth} \sim 0.6$, which can intuitively be understood as the most difficult to classify subjects: those with $f_{smooth} \leq 0.5$ are easily identified as 'Featured', while those with $f_{smooth} \geq 0.8$ are more obviously 'Not'. The top right-hand panel provides additional evidence showing the number of votes at retirement for both the original GZ2 project (solid lines) and our SWAP simulation (dashed blue). The left-skew inherent in the red SWAP-not-yet-retired sample is due to difficult-to-classify subjects that received only 30–40 classifications during the GZ2 project. Even after processing all available classifications, SWAP cannot retire these subjects without additional volunteer input. *Bottom panel:* Here we compare SWAP to results of simulations of GZ2 run with a lower retirement limit in order to evaluate whether or not GZ2's considerable number of votes per subject are necessary solely to populate subqueries. Solid bars show the number of classifications required to retire the same number of galaxies as SWAP (dark grey) for different fixed retirement limits in GZ2 (light grey). The height of the bars is normalized to show the counts relative to the highest simulated GZ2 retirement limit we test ($N = 35$, right vertical axis). The accuracy of the classifications for these simulated GZ2 runs against the full GZ2 project is shown as red points (left vertical axis). If GZ2 retirement were set at a level ($N = 10$) that reproduces the total number of classifications logged by SWAP, the accuracy would be below 90 per cent (versus SWAP's 96 per cent). Instead, GZ2 requires, at minimum, 3.5 times as many votes to approach the same accuracy (95 per cent) as SWAP. Simulated GZ2 sessions were run 100 times, randomly selecting subsamples with the same number of galaxies as were retired during our fiducial SWAP simulation. Quantities shown are averages of these trials; statistical error bars are too small to be seen.

SWAP's performance can be explained through its retirement mechanism. GZ2 did not have a predictive retirement rule, rather the project was declared complete when the median classification count for the ensemble reached a value that was deemed to be sufficient for accurate characterization of the classification. In contrast, SWAP retires 'easier' subjects first while harder subjects remain in the system for longer (requiring many more votes to nudge that subject's posterior across a retirement threshold). Evidence for this can be seen in the top two panels of Fig. 6. The top left-hand panel shows the distribution of f_{smooth} for the entire GZ2 sample (orange), the SWAP-retired sample (blue), and the sample of subjects which SWAP has not yet retired, of which there are ∼19K at the end of our simulation. The SWAP-retired sample generally follows the same distribution as GZ2-full except for the noticeable dip around

$f_{\rm smooth}$=0.6. In contrast, the SWAP-not-yet-retired sample peaks at $f_{\rm smooth}$=0.6. These subjects can be interpreted as being the most difficult to classify which can be understood intuitively: galaxies with $f_{\rm smooth}\leq$0.5 are easily identified as having features, while galaxies with $f_{\rm smooth}\geq$0.8 are more obviously elliptical.

This is further corroborated in the top right-hand panel of Fig. 6, which shows the distribution of the number of classifications a subject had at the time of retirement. The solid lines show this distribution from the original GZ2 project for the same subsamples as the top left-hand panel. For comparison, the dashed line shows the number of classifications at retirement realized during our SWAP simulation. Again, we see that the SWAP-retired sample is representative of GZ2 as a whole. However, the distribution for the SWAP-not-yet-retired sample is skewed towards fewer total classifications.

To understand this, consider the following: GZ2 served subject images at random with the exception that, towards the end of the project, subjects with low numbers of classifications were shown at a higher rate (Willett et al. 2013). The median number of classifications was 44 with the full distribution shown in orange in the top right-hand panel of Fig. 6. Our SWAP simulation processes these classifications in the same order as the original project (with the exception that gold-standard subject classifications are processed first as described in Section 3.2). Because our simulations cycle through only 92 d of GZ2 data, there are three general scenarios for why a subject has not yet been retired through SWAP: (1) SWAP has seen only a few of the many classifications for a given subject and it is not yet enough to retire it, (2) SWAP has seen many of the classifications for a subject but that subject is difficult; if we ran the simulation longer to process the remaining GZ2 classifications, SWAP would eventually retire it, and (3) SWAP has seen most or all of the classifications for a subject but it is difficult and there are few or no remaining GZ2 classifications; without additional volunteer input, these subjects will never be retired by SWAP.

It is this third category that skews the red distribution towards fewer GZ2 votes. These are difficult-to-classify subjects that have only 30–40 GZ2 classifications, all of which are processed by SWAP, but these subjects remain unretired. This is an indication that such subjects should have continued to accrue classifications in order to reach strong consensus.

We have demonstrated that SWAP retires subjects intelligently: quickly retiring easy-to-classify subjects while allowing those that are more difficult to collect additional classifications. SWAP thus requires only 30 per cent of the votes that GZ2 needs and retires nearly 5 times as many subjects during the three months of GZ2 project time that we include in our simulation.

3.5 Reducing human effort

SWAP's intelligent retirement mechanism is characterized, in large part, by the way SWAP estimates volunteer classification ability. This in turn allows for a dramatic reduction in the amount of human effort (votes) required. To see this more clearly, 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 per cent 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.

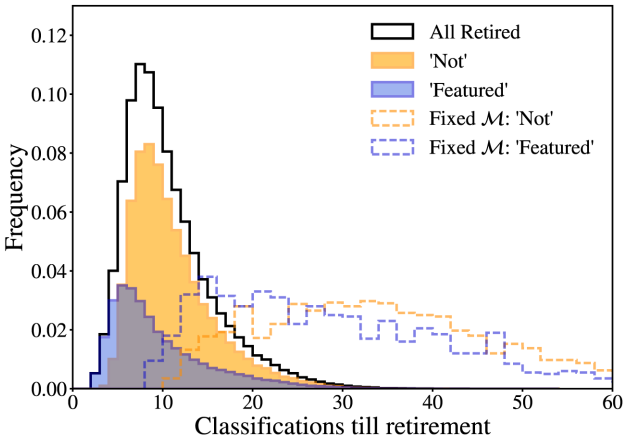

Figure 7. SWAP's volunteer-weighting mechanism provides a factor of 3 reduction in the human effort required to retire GZ2 subjects. The filled histograms show the number of volunteer classifications per subject achieved during our SWAP simulation broken down by class label, where the solid black line 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.

The results are presented in Fig. 7. The filled blue and orange histograms show the number of classifications per subject achieved from our SWAP simulation, where volunteer agent confusion matrices are those from Fig. 2. The dashed blue and orange distributions are the results from our toy model. When SWAP accounts for volunteer ability, 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 16–45 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 3.

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 3 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.

3.6 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 $\text{GZ2}_{\rm raw}$ label, SWAP disagrees about 5 per cent of the time. We thus examine the false positives (subjects SWAP labels as 'Featured' but $\text{GZ2}_{\rm raw}$ labels as 'Not') and false negatives (subjects SWAP labels as 'Not'but $\text{GZ2}_{\rm raw}$ labels as 'Featured'). We explore these subjects in redshift, magnitude, physical size, and concentration but 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

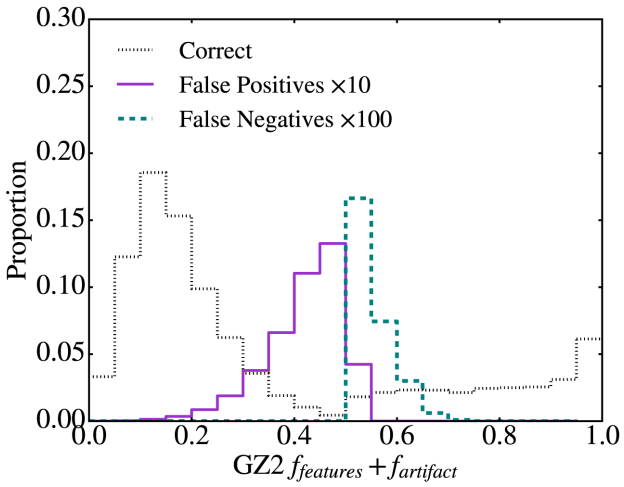

Figure 8. Distribution of GZ2 $f_{\rm featured}$+$f_{\rm artifact}$ vote fractions for subjects correctly identified by SWAP (dotted grey), along with those identified as false positives (solid purple) and false negatives (dashed teal). The false positives and false negatives are scaled by factors of 10 and 100, respectively for easier comparison. From Section 2, subjects with values >0.5 are defined as 'Featured', however, the teal distribution indicates that SWAP labels them as 'Not'. This is not necessarily a flaw of SWAP: 68.9 per cent of incorrectly identified subjects have $0.4 \leq f_{\rm featured}+f_{\rm artifact} \leq 0.6$, nearly the same range as a 68 per cent confidence interval around our choosen threshold. The overlap between the false positives and false negatives is due to subjects that are exactly 50-50; by default these are labelled 'Not'.

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.

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 per cent confidence interval on a subject with $f_{\rm smooth}$=0.5 is then (0.42, 0.57) assuming 40 classifications, each with a probability of 0.5. Fig. 8 shows the distribution of $f_{\rm featured}$+$f_{\rm artifact}$ 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 $0.4 < f_{\rm featured}+f_{\rm artifact} < 0.6$. 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 posterior probability crossing a retirement threshold, thus combining the best qualities of GZ2 and SWAP.

3.7 Summary

We demonstrate nearly a factor of 5 increase in the classification rate, a reduction of at least a factor of 3 in the human effort necessary to maintain that increased rate, all while maintaining 95 per cent 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 per cent (see Appendix B). Exploring those subjects wherein SWAP and GZ2 disagree, we conclude that the majority of this disagreement stems from the stochastic nature of $\rm GZ2_{raw}$ labels. We now turn our focus towards incorporating a machine classifier utilizing these SWAP-retired subjects as a training sample.

4 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 4.4.

4.1 Random Forests

We use an 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. RFs 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).

4.2 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 an 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.

4.3 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 per cent of galaxy pixels, as measured from SDSS DR12 *i*-band imaging (see Appendix C). 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, Jangren & Conselice 2000; Conselice, Bershady & Jangren 2000; Abraham, van den Bergh & Nair 2003; Conselice 2003; Lotz, Primack & Madau 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.

4.4 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 3, 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 per cent '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 an increasing sample size for fixed hyperparameters. Fig. 9 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 overfitting. However, as the training

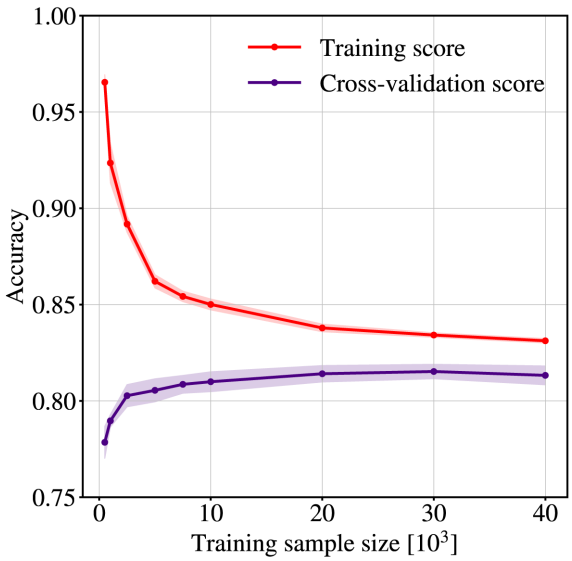

Figure 9. Learning curve for an RF with fixed hyperparameters. These curves show the mean accuracy computed during cross-validation and on the training sample, where the shaded regions denote the standard deviation. When the training sample size is small, the machine accurately identifies its own training sample but is unable to generalize to unseen data as evidenced by a low cross-validation score. This score increases with the size of the training sample but eventually plateaus indicating that larger training samples provide little in additional performance.

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 per cent 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.

4.5 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 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 of 'Featured'

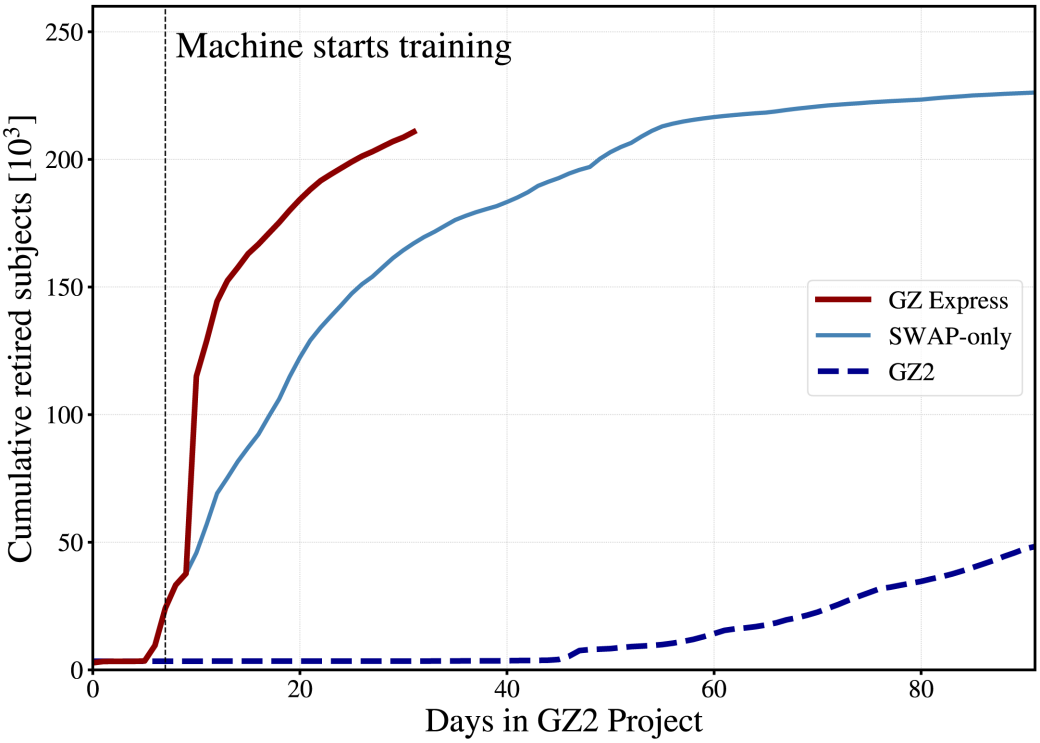

Figure 10. By incorporating a machine classifier, GZX (red) increases the classification rate by an order of magnitude compared to GZ2 (dashed dark blue) and outperforms the SWAP-only run (light blue), retiring more than 200K subjects in just 27 d of GZ2 project time. The dashed black line marks the 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 d having retired over 210K galaxies. See Table 1 for details.

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.

5 RESULTS

We perform a full GZX simulation incorporating our RF with the fiducial SWAP run discussed in Section 3.3. 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 d, having retired ∼210K subjects as detailed in Table 1.

We present these results in Fig. 10, 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, as reported in Table 1. Accuracy and purity remain within a few percent of the SWAP-only run at 93.1 per cent and 83.2 per cent, respectively. Instead we see a 5 per cent decline in the completeness. While the SWAP-only run identified 99 per cent of 'Featured' subjects, incorporation of the machine seems to miss a significant portion, thus dropping GZX completeness to 94.0 per cent. 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 d. Our GZX simulation processes a total of 936 887 visual classifications. As presented in Section 3.4, GZ2 requires at least 35 votes per subject to obtain galaxy classifications that are consistent 95 per cent of the time. At best, GZ2 could have retired 26 768 subjects with the classifications we process during our GZX run. This implies that we have increased the classification rate by at least a factor of 8, while requiring only 13 per cent as many human classifications. We next explore the composition of those classifications.

5.1 Who retires what, when?

In the top panel of Fig. 11, 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 3.3 for reference. Two things are immediately obvious. First, each component shoulders approximately half of the retirement burden with the machine and SWAP responsible for ∼98K and ∼112K 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

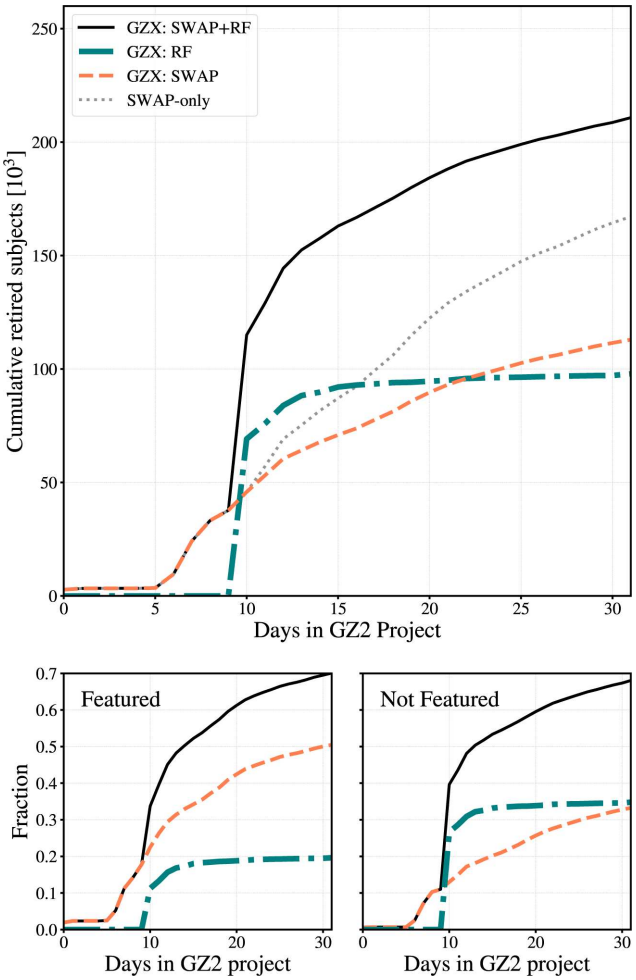

Figure 11. Contributions to subject retirement by both classifying agents of GZX: human (SWAP, orange) and machine (RF, teal). The top panel shows cumulative subject retirement for GZX as a whole (solid black), along with that attributed to the RF and SWAP. The dotted grey line shows the fiducial SWAP-only 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 73.7 per cent of the entire GZ2 sample in 32 d, retiring the same proportion of 'Featured' and 'Not' subjects as indicated by the black lines. However, humans retire 30 per cent more 'Featured' subjects than the machine, while both components retire a similar proportion of 'Not' subjects.

humans. However, the machine's performance plateaus quickly; the third phase is again dominated by human classifications.

In the bottom panels of Fig. 11, we consider the class composition of subjects retired by SWAP and the RF. The left (right)-hand 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 73.7 per cent 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 per cent of all 'Featured' subjects while the machine retires only 20 per cent. This divergence does not exist for 'Not' subjects where each component contributes 33–34 per cent.

What is the source of this discrepancy? Each night the machine trains on a sample composed consistently of 30–40 per cent 'Featured' subjects but does not retire a similar proportion, indicating that the 30 per cent 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.

5.2 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.

We isolate the 7060 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 7.2 per cent 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 sample is shown in Fig. 12.

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 3.6, 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 40 per cent 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 45 per cent of false positives have $f_{featured} + f_{artifact} \leq 0.35$, and this discrepancy is not as easily explained. In Fig. 12 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 that the subject is discy and worth further investigation.

Accordingly, this suggests that, in some cases, the morphology indicators we measure are sufficient for the machine to recognize 'Featured' galaxies regardless of the labels humans provide. Fig. 13 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 Fig. 14, 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

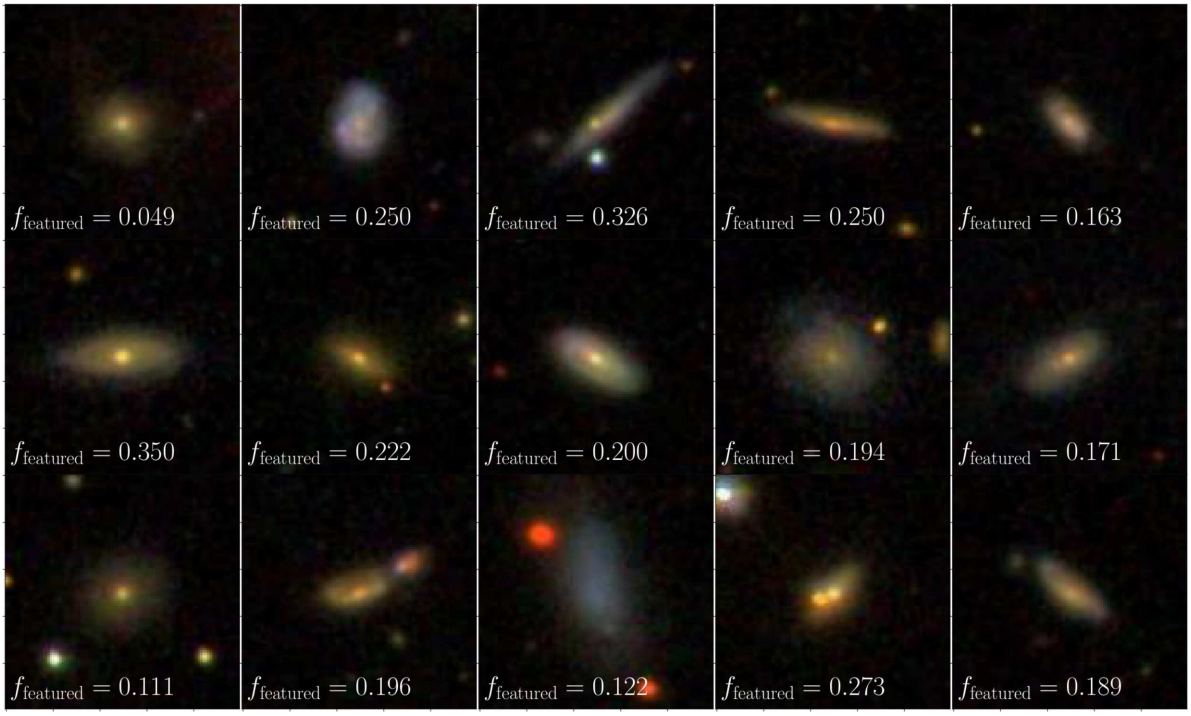

Figure 12. A random subsample of subjects identified as false positives: labelled by machine as 'Featured' but as 'Not' according to $GZ2_{raw}$. We display $f_{featured}$ in the lower left corner, that is, the fraction of volunteers who classified the subject as 'Featured'. Values are typically under 0.35 indicating that GZ2 volunteers strongly believed these to be 'smooth' ('Not'). Fortunately, the machine is able to identify these subjects as 'Featured' due to their measured morphology diagnostics.

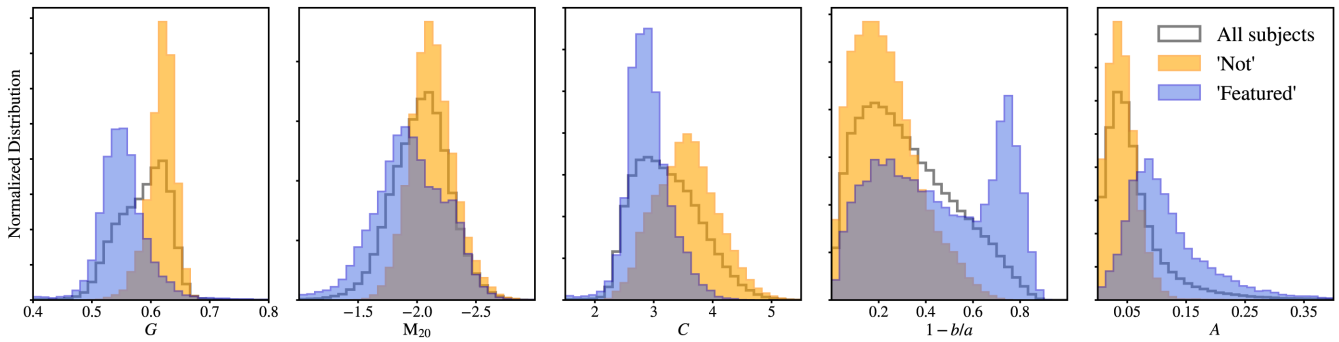

Figure 13. The RF is trained on a 5-dimensional morphology parameter space. We show the distribution of each morphology indicator for machine-retired 'Featured' (blue) and 'Not'(orange) subjects compared to the full GZ2 subject sample (black). The difference between 'Featured' and 'Not'subjects is in stark contrast for all distributions except, perhaps, M_{20}.

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 Fig. B2. 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.

6 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

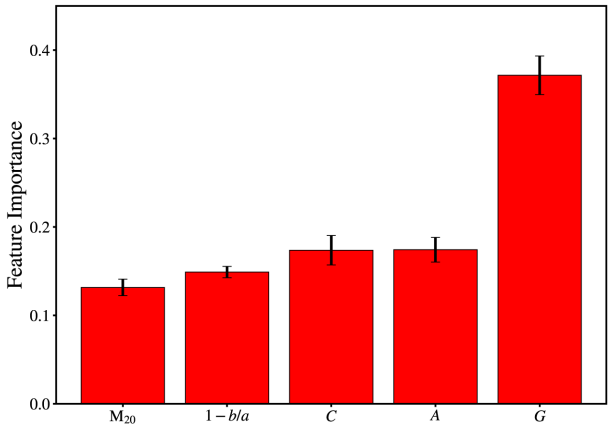

Figure 14. The RF's ranked feature importance averaged over all nights of training with black bars indicating the standard deviation. A larger value corresponds to higher importance. The machine computes feature importance according to how much each feature increases the purity of the resulting split averaged over all trees in the forest. The RF places great importance in the Gini coefficient though we note that it can underrepresent the importance of highly correlated features such as concentration.

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: multiband morphology diagnostics will be included in our future feature vector. However, the RF algorithm is not easily adapted to handle measurement errors or class labels with continuous distributions. A key feature of GZ2 vote fractions is their use in determining the strength of a morphological feature. Although both SWAP and our RF provide class predictions that are continuous, we apply thresholds to discretize the classification. To fully utilize the information provided, sophisticated algorithms should be considered such as deep convolutional neural networks (CNNs) 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 Fig. 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 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 yr to classify. [5] If the *Euclid* sample is on the high end, GZX as currently implemented could classify the brightest 20 per cent during the 6 yr of its observing mission. As currently implemented, we obtain accuracy around 95 per cent potentially leaving hundreds of thousands of galaxies with unreliable classifications. In a companion paper that seeks to identify supernovae, Wright et al. (2017) 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.

6.1 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 d of GZ2 project time to classify nearly 80 per cent of the entire galaxy sample.

Furthermore, we have implemented and tested an RF 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 per cent of GZ2 galaxies in just 32 d of GZ2 project time. This represents a factor of at least 8 increase in the classification rate as well as nearly 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 ∼94 per cent accuracy 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*.

ACKNOWLEDGEMENTS

We are grateful to the anonymous referees for helpful comments and suggestions which greatly improved this manuscript. 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 partial support from the US National Science Foundation (NSF) Grant AST-1413610. LF and DW also gratefully acknowledge partial support from NSF IIS 1619177. 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 issued by the Chandra X-ray Observatory

[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.

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. This publication uses data generated via the Zooniverse.org platform, development of which is funded by generous support, including a Global Impact Award from Google, and by a grant from the Alfred P. Sloan Foundation.

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

REFERENCES

Abraham R. G., Valdes F., Yee H. K. C., van den Bergh S., 1994, ApJ, 432, 75
Abraham R. G., Tanvir N. R., Santiago B. X., Ellis R. S., Glazebrook K., van den Bergh S., 1996, MNRAS, 279, L47
Abraham R. G., van den Bergh S., Nair P., 2003, ApJ, 588, 218
Astropy Collaboration et al., 2013, A&A, 558, A33
Baillard A. et al., 2011, A&A, 532, A74
Ball N. M., Loveday J., Fukugita M., Nakamura O., Okamura S., Brinkmann J., Brunner R. J.2004, MNRAS, 348, 1038
Bamford S. P. et al., 2009, MNRAS, 393, 1324
Banerji M. 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. et al., 2003, ApJ, 594, 186
Breiman L., 2001, Mach. Learn., 45, 5
Cardamone C. et al., 2009, MNRAS, 399, 1191
Casteels K. R. V. et al., 2014, MNRAS, 445, 1157
Conselice C. J., 2003, ApJS, 147, 1
Conselice C. J., 2006, MNRAS, 373, 1389
Conselice C. J., Bershady M. A., Jangren A., 2000, ApJ, 529, 886
Darg D. W. 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., Muñoz-Tun C., Dewberry J., Putko J., Teich Y., Popinchalk M., 2013, ApJ, 774, 86
Freeman P. E., Izbicki R., Lee A. B., Newman J. A., Conselice C. J., Koekemoer A. M., Lotz J. M., Mozena M.2013, MNRAS, 434, 282
Galloway M. A. et al., 2015, MNRAS, 448, 3442
Glasser G. J., 1962, J. Am. Stat. Assoc., 57, 648
Griffith R. L. et al., 2012, ApJS, 200, 9
Holwerda B. W. et al., 2014, ApJ, 781, 12
Hubble E. P., 1936, The Realm of the Nebulae. Yale Univ. Press
Huertas-Company M., Rouan D., Tasca L., Soucail G., Le Fèvre O., 2008, A&A, 478, 971
Huertas-Company M. et al., 2015, ApJS, 221, 8
Kartaltepe J. S. et al., 2015, ApJS, 221, 11
Kauffmann G. 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. et al., 2008, MNRAS, 388, 1686
Laureijs R. et al., 2011, preprint (arXiv:1110.3193)
Lintott C. J. et al., 2008, MNRAS, 389, 1179
Lintott C. et al., 2011, MNRAS, 410, 166
Lotz J. M., Primack J., Madau P., 2004, AJ, 128, 163
LSST Science Collaboration Abell et al., 2009, preprint (arXiv:0912.0201)
Marshall P. J. et al., 2016, MNRAS, 455, 1171
Masters K. L. et al., 2011, MNRAS, 411, 2026
Meert A., Vikram V., Bernardi M., 2016, MNRAS, 455, 2440
More A. et al., 2016, MNRAS, 455, 1191
Nair P. B., Abraham R. G., 2010, ApJS, 186, 427
Nakamura O., Fukugita M., Yasuda N., Loveday J., Brinkmann J., Schneider D. P., Shimasaku K., SubbaRao M.2003, AJ, 125, 1682
Odewahn S. C., Cohen S. H., Windhorst R. A., Philip N. S., 2002, ApJ, 568, 539
Pedregosa F. et al., 2011, J. Mach. Learn. Res., 12, 2825
Peng C. Y., Ho L. C., Impey C. D., Rix H.-W., 2002, AJ, 124, 266
Peng Y.-j. et al., 2010, ApJ, 721, 193
Peth M. A. et al., 2016, MNRAS, 458, 963
Sandage A., 1961, The Hubble Atlas of Galaxies. Carnegie Institute of Washington, Washington DC
Scarlata C. et al., 2007, ApJS, 172, 406
Schawinski K. et al., 2014, MNRAS, 440, 889
Sersic J. L., 1968, Atlas de Galaxias Australes. Universidad Nacional de Cordoba, Observatorio Astronomico
Shen S., Mo H. J., White S. D. M., Blanton M. R., Kauffmann G., Voges W., Brinkmann J., Csabai I., 2003, MNRAS, 343, 978
Sheth K. 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. et al., 2014, MNRAS, 445, 3466
Simmons B. D. et al., 2017, MNRAS, 464, 4420
Smethurst R. J. et al., 2016, MNRAS, 463, 2986
Snyder G. F. et al., 2015, MNRAS, 454, 1886
Strateva I. et al., 2001, AJ, 122, 1861
Taylor M. B., 2005, in Shopbell P., Britton M., Ebert R. eds., ASP Conf. Ser., Vol. 347, Astronomical Data Analysis Software and Systems XIV, p. 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., Steiman-Cameron T. Y., Sackett P. D., Olling R. P., 1990, AJ, 100, 1489
Willett K. W. et al., 2013, MNRAS, 435, 2835
Willett K. W. et al., 2017, MNRAS, 464, 4176
Wright D. E. et al., 2017, MNRAS, 472, 1315

APPENDIX A: EXPLORING THE QUALITY OF GALAXY ZOO: EXPRESS

In this section we consider the robustness of GZX by computing several sets of 'ground truth' labels from the GZ2 catalogue. Recall in Section 2 we defined a subject as 'Featured' if $f_{\rm featured} + f_{\rm artifact} \geq f_{\rm smooth}$, a threshold, t, of 0.5. Here we compute new descriptive labels by allowing that threshold to vary where $t \in [0.2, 0.3, 0.4, 0.5]$. Any subject with $f_{\rm featured} + f_{\rm artifact} \geq t$ is labelled 'Featured', otherwise it is labelled 'Not'. We recalculate the quality metrics of accuracy, purity, and completeness on the sample of galaxies retired during the full GZX simulation (SWAP+RF) for each threshold and each type of GZ2 vote fraction: raw, weighted, and debiased. The results are shown in Fig. A1. GZX classifications are quite robust, with accuracy fluctuating by only a few percent for $\rm GZ2_{raw}$ and $\rm GZ2_{weighted}$ labels computed using a threshold between 0.3 and 0.5. Instead we see a trade-off between purity and completeness. Decreasing the threshold results in more subjects labelled as 'Featured', which in turn increases sample purity while simultaneously decreasing completeness.

That the $\rm GZ2_{debiased}$ 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 a posteriori. It is also unsurprising that the $\rm GZ2_{raw}$ and $\rm GZ2_{weighted}$ classifications are in such tight agreement. $\rm GZ2_{weighted}$ 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.

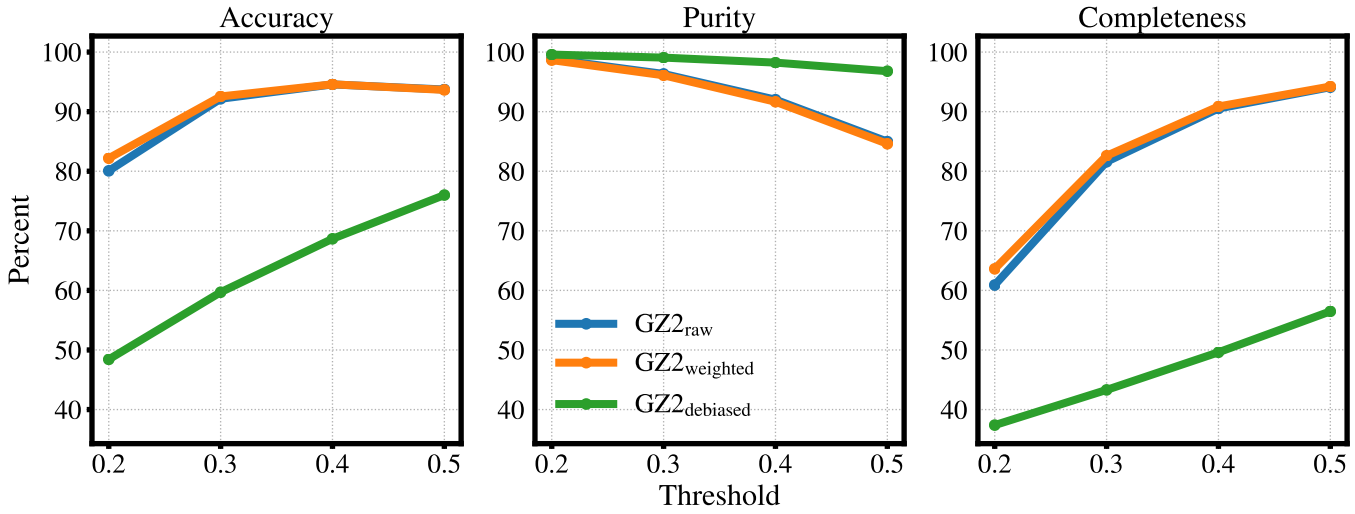

Figure A1. Quality metrics computed on the subjects retired during the GZX simulation for a range of thresholds and GZ2 vote fraction types.

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 show in Appendix B, changing the prior has little effect on the retirement rate but does result in considerable variability in the completeness and purity of the resulting classifications. The choice of whether to optimize SWAP to recover pure or complete samples is a decision for a given science team.

APPENDIX B: EXPLORING SWAP'S PARAMETER SPACE

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

B1 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-hand panel of Fig. B1. We find that SWAP is largely insensitive to the initial confusion matrix in terms of both the subject retirement rate and classification quality.

We retire $\sim$225K$\pm$3.5 per cent subjects as shown by the light blue-shaded region in the bottom left-hand panel of Fig. B1, 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. Regardless, most volunteers become astute classifiers by the end of the simulation. The top left-hand panel demonstrates our usual quality metrics as computed in Section 3.3. The dashed lines again denote the fiducial run. We maintain $\sim$95 per cent accuracy, 99 per cent completeness, and $\sim$84 per cent purity, and no metric changes by $>$ 2 per cent 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 per cent of retired subjects are the same among all simulations, and, of these, 99 per cent 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.

B2 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, disc, 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 affect 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-hand panels of Fig. B1.

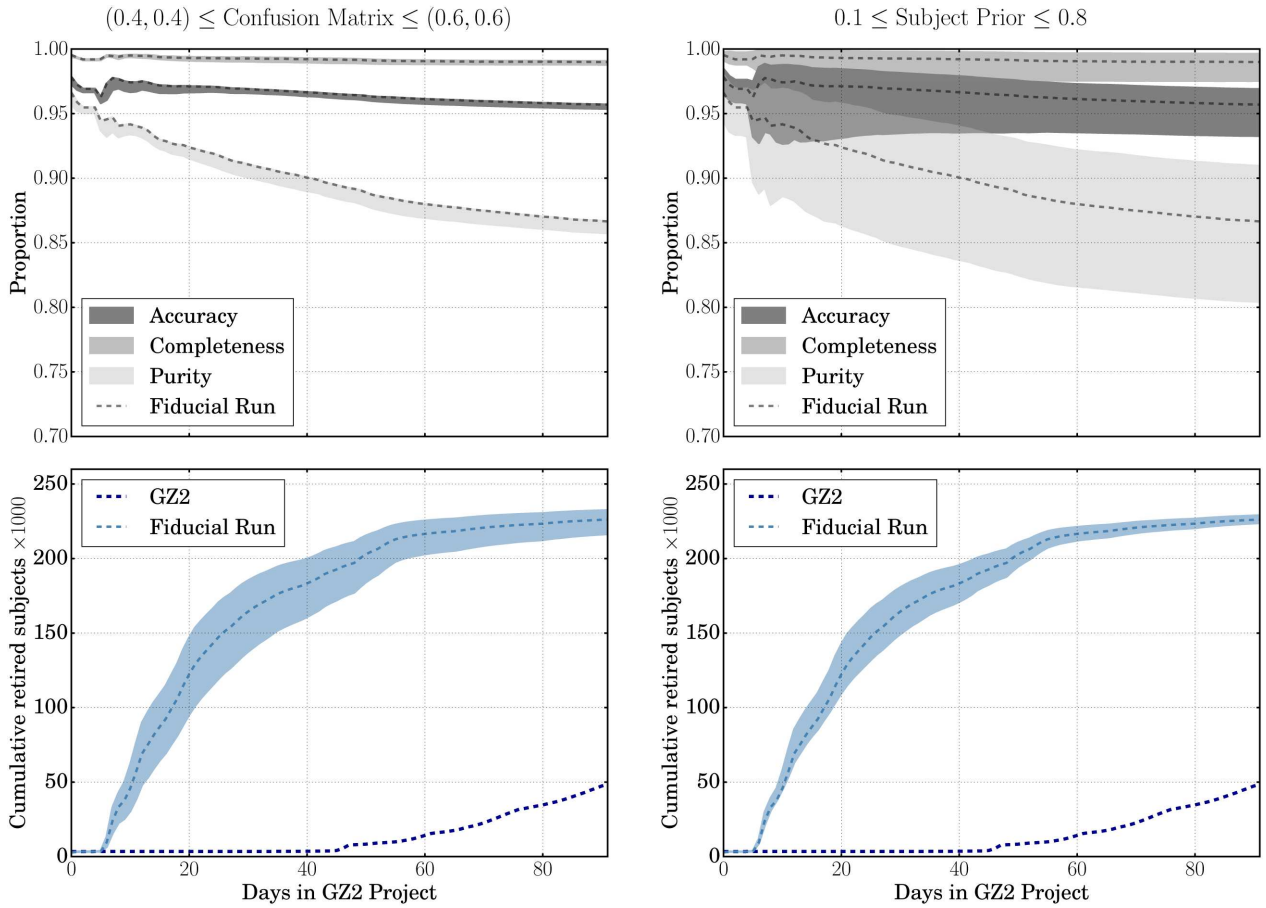

Figure B1. SWAP performance does not dramatically change even with a range of input parameters (shaded regions) as compared to the fiducial run of Section 3.3 (dashed lines). *Left.* The quality (top) and retirement rate (bottom) when the confusion matrix is initialized as (0.4, 0.4) and (0.6, 0.6), with all other input parameters remaining constant. *Right.* Same as the left-hand 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.

We again find that SWAP is consistent in terms of subject retirement which varies by only 1 per cent. However, as can be seen in the top panel, the variation in our quality metrics is more pronounced. First, 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 per cent of retired subjects are common to all simulations. Secondly, of those that are common, only 94 per cent 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 ±5 per cent.

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 data set. Indeed, the best performance is achieved when p_0= 0.35. This reflects the distribution of 'Featured' subjects as determined by $GZ2_{raw}$ 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.

B3 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-hand panel of Fig. B2, 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

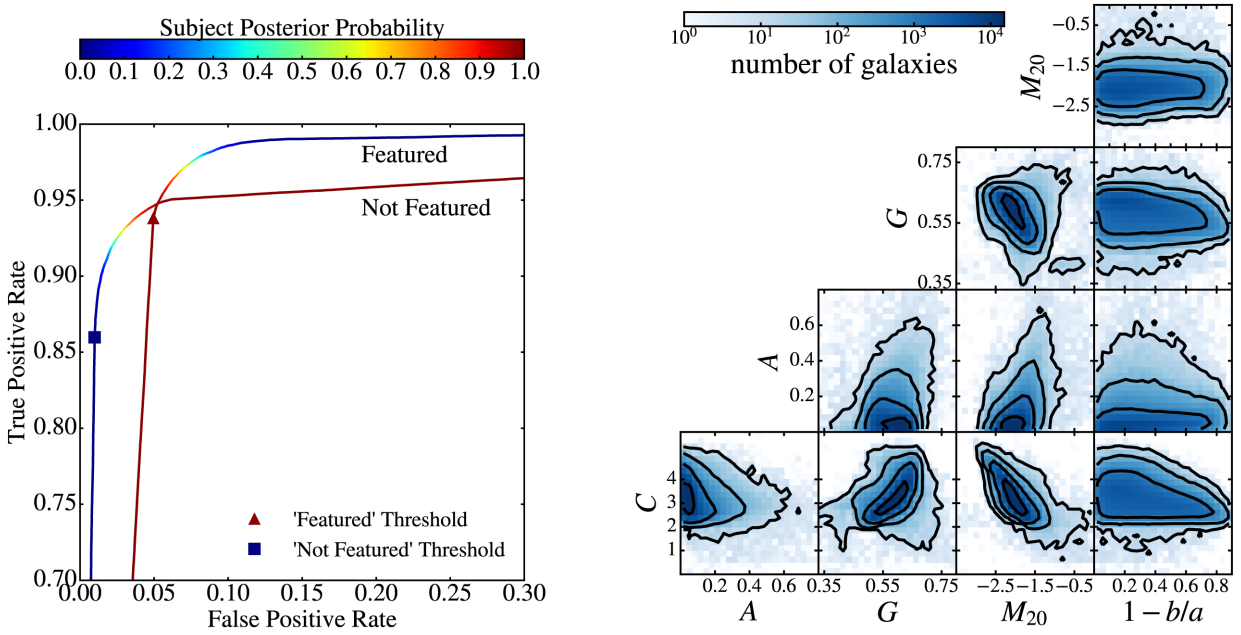

Figure B2. *Left.* Identifying 'Featured' subjects is independent of identifying 'Not' subjects. Both ROC curves use all subjects processed by SWAP where the score used to create the ROC curve is simply each subject's achieved posterior probability. The Featured curve demonstrates how well we identify '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'.

false-positive rate for a discriminatory threshold or score with a perfect classifier achieving 100 per cent 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.

APPENDIX C: MEASURING NONPARAMETRIC MORPHOLOGICAL DIAGNOSTICS ON SDSS STAMPS

In order to train our RF machine learning algorithm, we measure non-parametric morphology diagnostics for the GZ2 galaxy sample. SExtractor and their pixels replaced with values that mimic the background in that region. We obtain *i*-band imaging (with central wavelength 7480 Å) from SDSS Data Release 12 for 290 059 galaxies, representing 98.2 per cent of the GZ2 main galaxy sample. Postage stamps of each galaxy are cut from these fields where the dimensions of each cutout are 4×Petrosian radius as measured by the SDSS pipeline. Galaxies located within 4 Petrosian radii of the edge of a field were excluded as image mosaicking was not performed. This removed 7962 galaxies resulting in a final sample of 282 350 GZ2 galaxy postage stamps, or 95.6 per cent of the original sample.

These postage stamps undergo a cleaning process in order to remove the light from nearby sources so as not to contaminate the light profile of the galaxy of interest. Each stamp is processed through Source Extractor (SExtractor, version. 2.8.6; Bertin & Arnouts 1996). Two sets of parameters are used as it is not feasible to find a single set of parameters that properly identifies all 282K galaxies. The first is designed to identify bright sources, while the second is better optimized to detect fainter objects. SExtractor segmentation maps are used to identify the boundaries of each detected object in an image. By design, the galaxy of interest is located at the centre of the cutout. Extraneous sources are then identified from both the bright and faint segmentation maps and the pixels corresponding to these sources are replaced with a random value consistent with the background in that postage stamp.

We compute the following widely adopted nonparametric measurements of the galaxy light distribution on the cleaned postage stamps:

Concentration is computed as $C = 5\log(r_{80}/r_{20})$ (Bershady et al. 2000), where r_{80} and r_{20} are the radii containing 80 per cent and 20 per cent of the galaxy light, respectively. We define the total flux as that within 1.5 Petrosian radii, and the galaxy centre is that determined by the asymmetry minimization (described below; Lotz et al. 2004). Small values of this ratio tend to indicate discy 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 background noise is applied (as in e.g. Conselice et al. (2000); Lotz et al. (2004)), i.e.

$$A = \frac{\sum_{x,y} |I - I_{180}|}{2\sum |I|} - B_{180} \quad \text{(C1)}$$

Table C1. Summary of morphology measurements made on ∼282K galaxies from the GZ2 sample.

	Number	% Success	Notes
Full Galaxy Zoo 2 sample	295 305		
Postage stamps	282 350	95.6	per cent of full sample
Concentration	281 927	99.85	per cent of postage stamps
Asymmetry	282 334	99.99	per cent of postage stamps
Gini coefficient	282 323	99.99	per cent of postage stamps
M_{20}	282 194	99.94	per cent of postage stamps
Ellipticity $(1 - b/a)$	282 350	100.0	per cent of postage stamps
All morphologies successful	281 801	95.4	per cent of full sample

where I is the galaxy flux in each pixel (x, y), I_{180} is the image rotated by 180 deg about the galaxy's central pixel, and B_{180} is the average asymmetry of the background. A is summed over all pixels within one Petrosian radius of the galaxy's centre and then normalized by a corresponding measure in the original image. The centre is determined by minimizing A as described in Conselice et al. (2000).

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

$$G = \frac{1}{|\bar{X}|n(n-1)} \sum_{i}^{n} (2i - n - 1)|X_i| \tag{C2}$$

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 per cent of the galaxy flux. We compute it as

$$M_{tot} = \sum_{i}^{n} f_i[(x_i - x_c)^2 + (y_i - y_c)^2] \tag{C3}$$

$$M_{20} = \log_{10}\left(\frac{\sum_i M_i}{M_{tot}}\right), \text{while} \sum_i fi < 0.2 f_{tot} \tag{C4}$$

where f_i is the flux in pixel (x_i, y_i), and (x_c, y_c) is the galaxy's centre which is determined by minimizing the total moment, $M_{\rm tot}$, in a similar fashion as is done for the asymmetry. The galaxy pixels are then ranked by flux in descending order and M_i is summed over the brightest pixels until that sum equals 20 per cent of the total galaxy flux within one Petrosian radius, $f_{\rm tot}$, normalized by $M_{\rm tot}$. 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 successfully measure all morphological indicators for 281 801 SDSS galaxies. Some galaxies are lost at each stage of the measurement process due to various failures. For example, on rare occasions the minimization of the asymmetry centre fails to converge. The number of galaxies with successful measurements at each stage is listed in Table C1. The relation between these diagnostics for the full sample is shown in the right-hand panel of Fig. B2. The code developed to clean and compute these morphology indicators is open source and can be found at https://github.com/melaniebeck/measure_morphology.

This paper has been typeset from a TEX/LATEX file prepared by the author.