Draft version November 18, 2016
Preprint typeset using LaTeX style AASTeX6 v. 1.0

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

Melanie Beck, Claudia Scarlata, Lucy Fortson, Chris Lintott, Melanie Galloway, Kyle Willett, Brooke Simmons, Karen Masters, Hugh Dickinson, Phil Marshall, and Darryl Wright
Department of Physics, University of Oxford, Oxford OX1 3RH
and
Minnesota Institute for Astrophysics, University of Minnesota, Minneapolis, MN 55454

ABSTRACT

We implemented one of the first human-machine combos by running a kick ass simulation on previous citizen science data in conjunction with machine algorithms. And guess what? We can obtain at least an ORDER OF MAGNITUDE improvement in the efficiency of classification. So we got that going for us. Which is nice.

Keywords: editorials, notices — miscellaneous — catalogs — surveys

1. INTRODUCTION

Astronomers have made use of galaxy morphologies to understand the dynamical structure of these systems since at least the 1930s (Hubble ,Realm of the Nebulae). The division between early-type and late-type systems corresponds, for example, with a wide range of parameters from mass to environment and star formation history, and detailed observation of morphological features such as bars (Masters et al. 2010) and bulges (Simmons et al. 2013) provide information about the history of their host systems. Modern studies of morphology either seek to divide systems into broad classes (Lintott et al. 08, Karlatepe et al 14) or concentrate on identifying new and often rare classes, such as the green peas (Cardamone et al.) and beans (Schirmer et al.).

While attempts have been made to use proxies such as colour for morphology, there is no simple substitute for classifying by shape (Bamford et al. (red spirals), Schawinski et al (blue ellipticals). While Galaxy Zoo has, by recruiting a crowd, scaled human classification effort to cope with current surveys (Lintott et al, Willett et al., Simmons et al.), upcoming surveys including LSST and Euclid which will come online in the next five years will require a different approach.

The data promised from these massive surveys will provide a wealth of new information, including millions of never before seen galaxies. In order to understand the structure and evolution of these galaxies, methodologies must be developed to handle very large data sets. Standard visual morphology methods will be unable to cope with the size and current automated morphologies (see, for example, XXXXX) provide only statistical morphological structure with large uncertainties and impurities. Methods must be developed which can efficiently and accurately quantify the likelihood of given morphological features being present.

One solution is full automation. The standard approach is to define a set of features which describe the morphology in an N-dimensional space. Decision boundaries within this space define morphological types for each galaxy. Learning these decision boundaries can be achieved through standard machine learning techniques such as support vector machines (Huertas-Company 2008) or principal component analysis (Scarlata 2006). Another approach is through deep learning, a machine learning technique that attempts to model high level abstractions. Algorithms like convolutional neural networks have been used with impressive accuracy (Dieleman 2015, Huertas-Company 2016). However, a drawback to all automated classification techniques is the need for standardized training data, with more complex algorithms requiring more data. Furthermore, that data must be consistent for each survey as differences in resolution and depth are inherently learned by the algorithm making them difficult to apply to disparate surveys.

In this work we seek to develop a system which preserves the best features of both human and automatic classifications, developing for the first time a system which, at scale, brings both human and machine intelligence to the task of galaxy morphology. We demonstrate the effectiveness of such a system through a re-analysis of visual galaxy morphology classifications collected during the Galaxy Zoo 2 Project, and combine these classifications with a suite of non-parametric morphology indicators typically used for automated morphologies. Our method provides an order of magnitude improve-

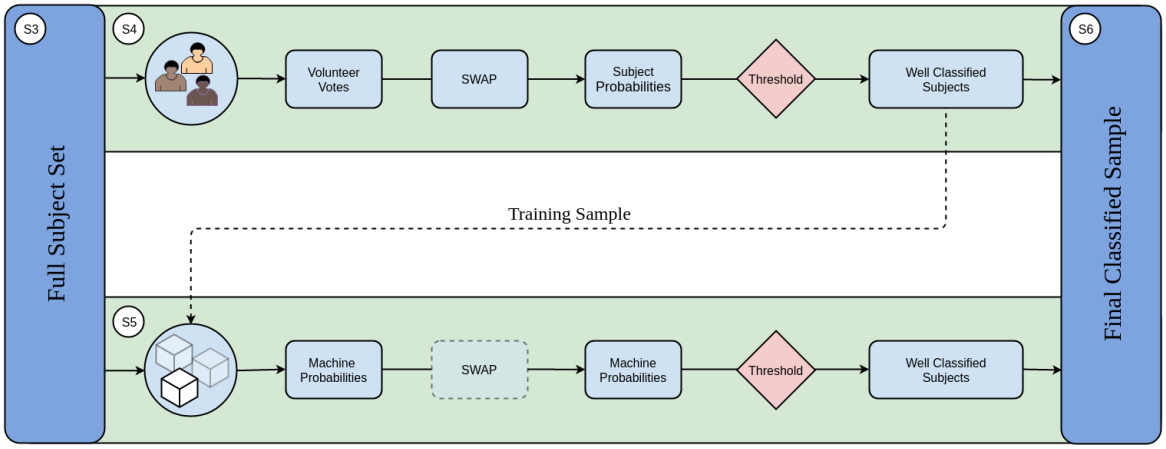

Figure 1. Schematic of our hybrid system. Human classifiers are shown images of galaxies via the Galaxy Zoo web interface. These classifications are recorded and processed according to section XXX. As a result of the processing, those subjects whose probabilties cross the classification thresholds are passed to the machine classifier as a training sample. The trained machine is then applied to the remaining subjects in the database (test sample). Those subjects which the machine classifies with high confidence are removed from the sample and considered fully classified. The rest remain in the database to be seen by human classifiers.

ment in the rate of galaxy morphology classifications and suggests that the challenge of future surveys can be met.

2. GALAXY ZOO EXPRESS OVERVIEW

In Galaxy Zoo Express (GZX) we combine humans and machines with the goal of increasing galaxy morphology classification efficiency, both in terms of the rate of classification over time and in the amount of human effort required. Figure 1 presents a schematic of GZX which includes section numbers as a shortcut for the savvy reader. We note that transparent portions of the schematic represent areas of future work which we explore in section 7.

Any system combining human and machine classifications will have a set of generic features: a group of human classifiers, at least one machine classifier, and a decision engine which determines how these classifications should be combined.

We draw from the Galaxy Zoo 2 (GZ2) classification database which allows us to create simulations of human classifiers (described in section 3). These classifications are used most effectively when processed with SWAP, a Bayesian code described in section 4, first developed for the Space Warps gravitational lens discovery project. These subjects become the basis for the machine's training sample.

In section 5 we incorporate a machine classifier; where, for this project, we have developed a random forest classifier which trains on easily measured physical parameters relevant to galaxy morphology such as Concentration, Asymmetry, Gini coefficient and M_{20} as input. After a batch of (human) classifications is processed, the machine is trained and its performance assessed against a validation sample. This procedure is repeated and the machine grows in accuracy as the size of the training sample increases (to a point). Once the machine reaches an acceptable level of performance it is run against the remaining galaxy sample. Images reliably classified by machine are not further classified by humans.

Even with this simple description, one can see that the classification process will progress in three phases. At first, the machine will not yet have reached the acceptable level of performance and the only subjects classified are those for which human classifiers have reached consensus. Secondly, the machine will rapidly improve and both humans and machine will be responsible for classification. Finally, improvement in the machine performance will slow and the remaining images will likely need to be classified by humans (if they can be classified at all). These results are explored in section 6. This blueprint allows even moderately successful machine learning routines to make significant contributions alongside human classifiers and removes the need for ever-increasing performance in machine classification.

3. GALAXY ZOO 2 CLASSIFICATION DATA

Our simulations utilize original classifications made by volunteers during the GZ2 project. These data are described in detail in Willett et al. (2013) though we provide a brief overview here. The GZ2 subject sample was designed to consist of the brightest 25% (r band magnitude < 17) of galaxies residing in the SDSS North Galactic Cap region from Data Release 7 and included subjects with both spectroscopic and photometric redshifts out to $z < 0.25$. In total, 285,962 subjects were

classified in the GZ2 Main Sample catalogs[1].

Subjects were shown as color composite images via a web-based interface wherein volunteers answered a series of questions pertaining to the morphology of the subject. With the exception of the first question, subsequent queries were dependent on volunteer responses from the previous task creating a complex decision tree. Using GZ2 nomenclature, a *classification* is defined as the total amount of information about a subject obtained by completing all tasks in the decision tree. A *task* represents a segment of the tree consisting of a *question* and possible *responses*. A subject is *retired* after it has achieved a sufficient amount of classification.

For the analysis in this paper we utilize the first task in the tree, 'Is the galaxy simply smooth and rounded, with no sign of a disk?' to which possible responses include 'smooth', 'features or disk', or 'star or artifact'. This choice serves two purposes: 1) this question is one of only two questions in the GZ2 decision tree that is asked of every subject thus maximizing the amount of data we can work with, and 2) our analysis will assume a binary task and this question is simple enough to mold into such a form.

To force such a binary classification, we group 'star or artifact' responses with 'features or disk'. Additionally, we define 'true' labels for each GZ2 subject to which we can compare labels assigned by Galaxy Zoo Express (GZX). Specifically, we take the majority vote fraction as the label for that subject. If the majority resided under 'star or artifact' or 'feature or disk', it was labeled as 'Featured'; otherwise it was labeled 'Not'. The GZ2 catalog assigns every subject three types of volunteer vote fractions: raw, weighted, and debiased. Debiased vote fractions are calculated to correct morphological classifications for redshift bias, a task that GZX is not yet built to handle. The weighted vote fractions serve to downgrade malicious volunteers and bots, a task we perform as well. However, because the mechanism for determining malicious volunteers is entirely different between GZ2 and GZX, we use labels derived from the GZ2 raw vote fractions ($\mathrm{GZ2_{raw}}$) as the closest comparison. We note that only 512 subjects in the GZ2 catalog have a majority 'star or artifact' vote fraction, contributing less than half a percent contamination.

In total, the data consist of over 16 million classifications from 83,943 individual volunteers. As we discuss in Section 4, the algorithm we use requires that every volunteer see a subset of subjects that are expertly identified by a member of the science team. 30,984 volunteers identified one or more of our gold standard sample (see Section 4.1), thus we use only those classifications made by one of these "gold standard" volunteers. We note that these volunteers represent 36% of all users yet provide nearly 90% of the total Galaxy Zoo classification data, reducing the total number of classifications available for our simulated runs to approximately 14 million.

[1] https://data.galaxyzoo.org

4. EFFICIENCY THROUGH CLEVER HUMAN-VOTE PROCESSING

Galaxy Zoo 2 had a brute force subject retirement rule whereby a subject was considered retired when a target number of classifications had been reached. As the project required a large number of independent classifications for each subject, the retirement threshold was set rather high, typically at forty individual volunteer classifications. Once the project reached completion, inconsistent and unreliable volunteers were down-weighted as described in Willett et al. (2013). While this process reduces input from malicious users and 'bots', it doesn't reward the most consistent volunteers. Furthermore, waiting until project completion doesn't allow for efficient utilization of super-users, those volunteers who are exceptional at classification tasks.

As a first step towards increasing classification efficiency, we ran a sequence of simulations on GZ2 classifications employing an algorithm that more intelligently manages subject retirement. This software was adapted 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), the software predicted the probability that an image contained a gravitational lens given volunteers' classifications as well as their past experience after being shown a training sample consisting of simulated lensing events. We provide a brief overview here.

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

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

where $N_{\text{“}X\text{”}}$ is the number of classifications the volunteer labeled as type X, N_X is the number of subjects the volunteer has seen that were actually of type X, and d represents the history of the volunteer (all subjects they have seen). The software employs two prescriptions for

when the agent updates the volunteer's confusion matrix. In *Supervised* mode the probabilities are only updated after the volunteer identifies a training subject. In *Supervised and Unsupervised* mode, the agent updates the probabilities after every subject the volunteer identifies.

Each subject is assigned a prior probability that it exhibits feature A: $P(A) = p_0$. When a volunteer makes a classification C, Bayes' Theorem is used to derive how the subject's prior probability should be updated into a posterior:

$$P(A|C) = \frac{P(C|A)P(A)}{P(C|A)P(A) + P(C|N)P(N)} \quad (2)$$

where this value can then be calculated using the elements of the agent's confusion matrix. Marshall et al. (2016) show that perfect volunteers (i.e., those with $P(\text{“}A\text{”}|A) = 1.0$ and $P(\text{“}N\text{”}|N) = 1.0$ would calculate the posterior probability of the subject to be 1.0 which is not surprising (perfect classifiers are perfect!). However, they also show that *obtuse* classifiers (those with $P(\text{“}A\text{”}|A) = 0.0$ and $P(\text{“}N\text{”}|N) = 0.0$ also produce a posterior probability of 1.0; demonstrating that obtuse volunteers are just as helpful as perfect volunteers.

As the project progresses, each subject's prior probability is continually updated and is nudged to higher or lower probability depending on volunteer classifications. Probability thresholds can then be set such that subjects crossing these thresholds are highly likely (unlikely) to exhibit the feature of interest (or not). While most subjects will cross a classification threshold, some are indeterminate in that their posteriors simply bounce back and forth in probability space. Those that do cross a threshold are considered *retired*. The software no longer records volunteer information on these subjects.

4.1. *Volunteer Training Sample*

A key feature of the original Space Warps project was the training of individual volunteers through the use of simulated lensed galaxies. Volunteers were shown these simulated images interspersed with actual data with the simulated data shown predominately at the beginning of a volunteer's association with the project. After a volunteer submitted their classification, the system provided feedback in the form of a pop-up comment. In this section we describe how we engineer the GZ2 data to mimic the Space Warps setup as closely as possible, though we note that retroactively training volunteers in real time is obviously not possible.

We find that the SWAP algorithm does not perform well without the use of designated training images. Furthermore, SWAP performs optimally when these training images are introduced towards the beginning of a volunteer's contribution to the project. This allows each volunteer's confusion matrix to update sufficiently before intense classification of test images commences. To mimic this behavior we select a sample of ∼3500 SDSS galaxies which overlaps the Nair & Abraham (2010) catalog. This catalog contains ∼14K galaxies classified by eye into various T-Types, a numerical index of a galaxy's stage along the Hubble sequence. Though helpful, this particular classification isn't directly comparable to GZ2 as discussed in Willett et al. (2013). Therefore, we took the additional step of reclassifying a subsample on the Zooniverse platform. [2] These expert classifications were provided by approximately 15 members of the Galaxy Zoo science team ranging from advanced graduate students, post docs, and faculty. The question posed to our science team was identical to the original GZ2 question and at least five experts saw each galaxy. Once classifications were complete, votes were aggregated and a simple majority was used to provide an expert label to each of the 3500 subjects. However, not every volunteer in the GZ2 database classified at least one of these ad-hoc training images. Because we wish to recreate the conditions of the Space Warps project, we remove from our data all volunteers who don't classify at least one of these 3500 subjects. This reduces our raw data set from 16 million classifications to 14 million; from 90K unique volunteers to 30K.

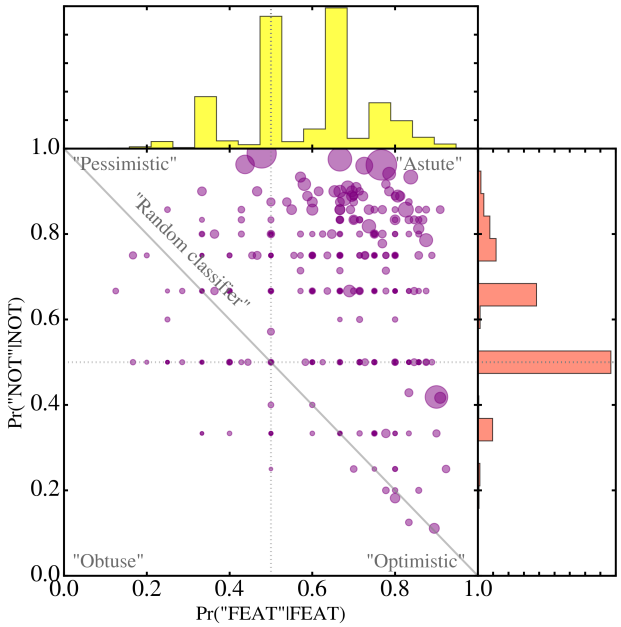

Figure 2. Volunteer probabilities for our fiducial SWAP run.

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

Finally, the classifications for these particular subjects could have timestamps anywhere within the 14-month time span during which the original project ran. We therefore adjust the order of the classification timestamps such that annotations of training sample subjects have timestamps well before all other GZ2 subjects. When running a simulation, which pulls from the database according to timestamps, the training image classifications are the first to be processed with SWAP. In each of the following simulated runs, the first four days consist of processing training image classifications only after which the remaining GZ2 subjects classifications are processed.

Figure 2 demonstrates the volunteer training we achieve with this scheme (figure adapted from Marshall et al. (2016)). This figures shows the confusion matrices for 1000 volunteers where the size of the circle is proportional to the number of training images each volunteer saw. The histograms represent the distribution of all volunteers for each probability. Nearly 25% of volunteers fall into the 'Astute' category indicating they are generally good at correctly identifying both 'Featured' and 'Not' (smooth) subjects. The spikes in the histograms are due to 48% of volunteers that have one of their probabilities at 0.5 even after training. These volunteers see only one expertly-classified training image (say, 'Featured'), thus their probability in the other ('Not') remains unchanged. Finally, 4% of volunteers have 0.5 for both probabilities after training. These are likely volunteers who see two training images of the same type where one is classified correctly but the other is classified incorrectly, essentially resetting their probability.

4.2. *Fiducial SWAP simulation*

To simulate a live project we run SWAP on a regular timestep which we set as $\Delta t = 1$ day. At each timestep, the software pulls from the database all volunteer classifications which have timestamps within that range. We cycle through three months of GZ2 classification data for each simulation we discuss below. However, before a simulation can be run, a number of parameters which control the behavior of SWAP must first be chosen. These include the initial confusion matrix assigned to each volunteer, the retirement thresholds, and the prior probability of the subject. Specifically, we must choose

- $P_{F,0}$, the initial probability that a volunteer identifies a subject as being 'Featured', $P_0(\text{“}F\text{”}|F)$
- $P_{N,0}$, the initial probability that a volunteer identifies a subject as being 'Not', $P_0(\text{“}N\text{”}|N)$
- p_0, the prior probability of a subject to be 'Featured'.
- t_F, the threshold defining the minimum probability for a subject to be retired as 'Featured'.
- t_N, the threshold defining the maximum probability for a subject to be retired as 'Not'.

We begin with a fiducial simulation in which we set $P_{F,0}$, $P_{N,0}$, and p_0 equal to 0.5. We let $t_F = 0.99$ and $t_N = 0.004$.

Because our ultimate goal is to increase the efficiency of galaxy classification, we use as a metric the cumulative number of retired subjects as a function of the original GZ2 project time for both the original GZ2 project and the SWAP output. GZ2 retirement was defined by the number of volunteer classifications, requiring ∼40 individual volunteers to reach classification consensus for each subject. We use a slightly more lenient definition and consider a subject GZ2-retired after it achieves 30 volunteer votes since not every subjec reached the 40-classification threshold as discussed in Willett et al. (2013). SWAP retirement is determined by a subject's posterior probability crossing either of the retirement thresholds defined above.

However, it is important not to prioritize efficiency at the expense of quality. Towards this end, we also consider the quality metrics of accuracy, purity and sample completeness as a function of GZ2 project time. These are defined in the standard way where accuracy is the number of correctly identified subjects divided by the total number retired; completeness is the number of correctly identified 'Featured' subjects divided by the number of actual 'Featured' retired; and purity is the number of correctly identified 'Featured' subjects divided by the number of subjects retired as 'Featured'.

Using as truth the labels we defined in section 3, we compute these metrics on the subject set retired *by that day of the GZ2 project.* For example, as shown in Figure 3, on the 20th day of the GZ2 project, SWAP has retired 120K subjects. Comparing those SWAP labels to their GZ2 labels, we find that the sample is 96% accurate, nearly 100% complete (that is, of all the subjects we retire up to that point, we successfully identify all that are 'Featured'), and 92% pure.

Figure 3 shows the results of our fiducial SWAP simulation compared to the original GZ2 project. The right hand axis shows the cumulative number of retired subjects as a function of GZ2 project time. After 90 days, GZ2 retires 50K subjects while SWAP retires more than 225K. In other words, we classify 80% of the entire GZ2 sample in three months. The original GZ2 project took approximately a year to complete. We thus achieve nearly an order of magnitude increase in classification time. One can also consider the amount of human effort necessary to perform these classification tasks. Our SWAP run required 2.3×10^6 volunteer

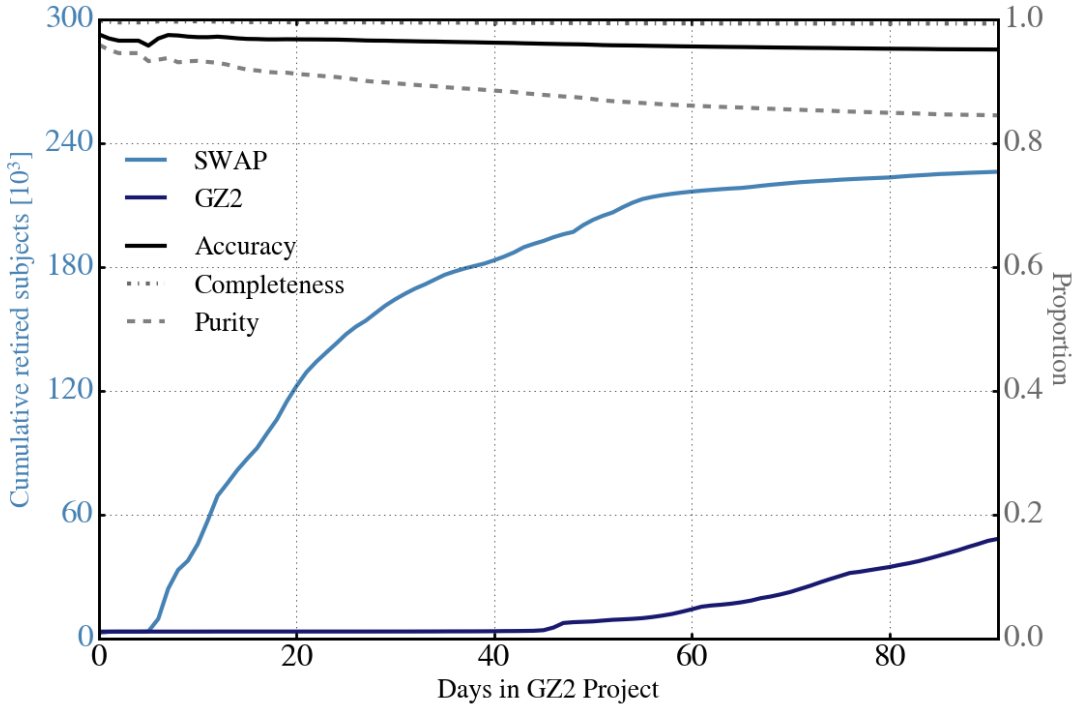

Figure 3. Fiducial SWAP simulation demonstrating the dramatic increase in the subject retirement rate as a function of GZ2 project time (in light blue) compared with the original GZ2 project (dark blue) corresponding to the left hand axis. Not only is efficiency increased by nearly an order of magnitude, we maintain high quality classifications as shown by high marks in accuracy, completeness and purity which correspond to the right hand axis. Specifically, these metrics are computed on the sample obtained *by that day in GZ2 project time*, e.g. on day 20, these metrics are computed on the 120K subjects which SWAP has classified by that time; their SWAP labels being compared to 'true' labels derived from published GZ2 data.

votes to retire these 225K subjects. GZ2 required nearly 10×10^6 for the exact same subject set. Again, this is nearly an order of magnitude reduction in the human effort required to classify this data set! This reduction in human effort can be seen directly in Figure 4 which shows the volunteer vote distributions achieved through SWAP (light blue) compared to GZ2 (dark blue) for the ∼225K retired subjects. GZ2, as expected, has a distribution that peaks around ∼45 unique volunteers classifying each subject with 99% of subjects having at least 25 classifications (Willett et al. 2013). SWAP, on the other hand, has a distribution which peaks around 10 classifications before retirement indicating that most subjects don't need as much human effort to reach a sufficient level of consensus. Some subjects are 'easy' to classify and can be retired in as few as 3 classifications, while subjects with less consensus will take more classifications, each one kicking the subject back and forth in probability space before it eventually crosses one of the retirement thresholds. This explains the tail towards higher classifications requirement.

It is obvious that by clever and adaptive processing of volunteer classifications, efficiency of subject classification and retirement can be increased by a factor of

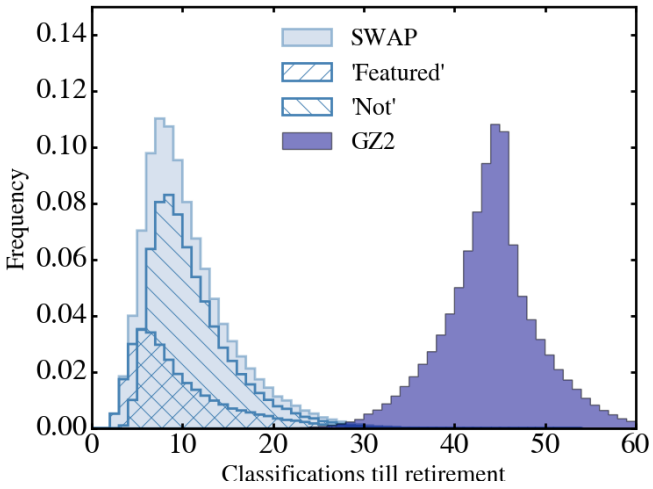

Figure 4. SWAP requires less human effort than GZ2 as evidenced by comparing the number of classifications until retirement for the ∼225K subjects retired by both SWAP and GZ2. GZ2 requires ∼45 classifications per subject before retirement. In contrast, most subjects only need approximately 10 classifications until retirement when processing volunteer votes with SWAP. Overall, SWAP can retire the same number of subjects but with an order of magnitude less human effort.

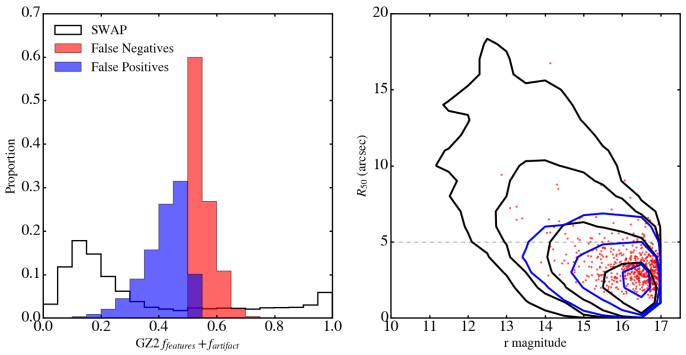

Figure 5. SWAP and GZ2 labels disagree approximately 5% of the time, depending on the choice of initial SWAP parameters. The disagreement stems, in large part, from 'true' labels derived from uncertain GZ2 vote fractions as shown in the left panel where the majority of SWAP's false negative and false positive subjects have GZ2 vote fractions which fall in the range 0.4 - 0.6. Furthermore, it is unsurprising that these same subjects are physically more difficult to classify being, on average, smaller and fainter than the sample as a whole as shown in the right panel.

four. The exact nature of subject retirement and associated quality metrics will be, in part, a function of initial SWAP parameters. We explore the SWAP parameter space in depth in Appendix B.

4.3. *Disagreements between SWAP and GZ2*

Galaxy Zoo's strength comes from the consensus of dozens of volunteers voting on each subject. Processing votes with SWAP effectively reduces this consensus to its bare minimum. Though we typically recover the GZ2 label, SWAP disagrees about 5% of the time. In this section we examine the main effects driving this disagreement whereby we again turn to our fiducial simulation, isolating all false positives and false negatives.

We find that the majority of these disagreements are due to uncertainties in the GZ2 label. This is shown in the left panel of Figure 5 where we show the distribution of the `features_or_disk` + `star_or_artifact` GZ2 vote fractions for the entire sample retired by SWAP (solid black lines). Recall that we group these together as 'Featured' during our SWAP simulations and so group them here for a fair comparison. Furthermore, because we used a majority vote fraction to derive labels, any subject with $f_{features} + f_{artifact} > 0.5$ would be labeled 'Featured'. However the false negative (red) distribution shows that a portion of these still attained a 'Not' label through SWAP (with the opposite being true for the blue false positive sample). In fact, the majority of incorrectly labeled subjects have $0.4 \leq f_{features} + f_{artifact} \leq 0.6$, indicating that the GZ2 vote fractions are simply too uncertain to provide high quality labels.

Another effect contributing to the disagreement is related but more subtle and concerns the order in which volunteers cast their votes. Consider a subject where the first N volunteers classify it as X while the subsequent majority of volunteers classify the subject as Y. Here the power of GZ2's large consensus shines because this lopsided voting is averaged out. SWAP, however, will swiftly kick that subject's probability over a retirement threshold when encountering such a chain of classifications. The result is SWAP labels which disagree with GZ2 labels purely due to order statistics. Consider a subject with a $f_{features} + f_{artifact} = 0.2$, where 40 unique volunteers have classified this subject yielding a GZ2 label of 'Not'. What is the probability that this subject will obtain a label 'Featured' through SWAP? In other words, what is the probability that the first eight volunteers all voted 'Featured'? If we assume all volunteers have $P_{F,0} = 0.5$, it is trivial to compute $1/2^8 = 0.39\%$. Of course, this effect is more complicated to model in aggregate since volunteers have confusion matrices that vary with time and subjects can be retired with a varying amount of volunteer classifications.

This leads to another subtle effect that can cause SWAP labels to disagree with GZ2 labels. Volunteers generally do not have $(P_{F,0}, P_{N,0}) = (0.5, 0.5)$. We find that, occasionally, some subjects are retired with only two classifications because one of the volunteers had an exceptionally large confusion matrix. We estimate this effect at XXX.

To prevent these issues, one could a.) require each volunteer classify a certain number of subjects to increase their confusion matrix values or b.) require each subject reach a minimum number of classifications before allowing it's probability to cross a threshold. The difference lies between averaging over a volunteer's burn-in phase versus a subject's burn-in phase. The latter is preferable as the majority of Zooniverse volunteers contribute only a small amount of classifications to any given project. Requiring they achieve a minimum classification count before allowing their contribution to count would hamstring the effectiveness of citizen science projects.

4.4. *Summary*

We have demonstrated that, regardless of the initial configuration of the SWAP software, we achieve 4-5x increase in the efficiency of classification corresponding to nearly an order of magnitude decrease in required human effort. All of this can be obtained with accuracy over 95%, 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%. We've explored those subjects where the SWAP label and the GZ2 label disagree and have shown that the majority of the disagreement lies in the uncertainty of our 'true' labels with small contributions from idiosyncrasies

in SWAP. We now turn our focus towards incorporating a machine classifier utilizing these SWAP-retired subjects as a training sample.

5. EFFICIENCY THROUGH INCORPORATION OF MACHINE CLASSIFIERS

In this section we construct the full Galaxy Zoo Express by incorporating supervised learning, the machine learning task of inference from labeled training data. The training data consist of a set of training examples, and must include an input feature vector and a desired output label. Generally speaking, a supervised learning algorithm analyzes the training data and produces an inferred function that can then be mapped to new examples. An optimized algorithm will correctly determine class labels for unseen data. In general, most classification algorithms can handle prediction of several labels simultaneously. Work has been done to predict the entirety of GZ2 classification labels using deep learning (Dieleman et al. 2015) with great success. However, it is still simpler for a machine to predict fewer labels than it is to predict several dozen, [**citation?**], with the additional bonus that fewer class labels require less training data. By processing human classifications through SWAP we obtain a discrete, binary task for a machine to tackle. We briefly outline the technical details of our machine classifier before turning towards the decision engine we develop for GZX in section 5.5.

5.1. *Random Forests*

Because our task is simple, we choose a simple machine. In particular, we use a Random Forest (RF) algorithm, 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. As such, decision trees alone are prone to overfitting the training data thus precluding them from generalizing well to new data. Random Forests mitigate this effect by combining the output label from a multitude of decision trees. In particular we use the `RandomForestClassifier` from the Python module `scikit-learn` (Pedregosa et al. 2011).

5.2. *Cross-validation*

Of fundamental importance is the task of choosing an algorithm's hyperparameters, values which determine how the machine learns. In the case of a RF, one must choose the maximum depth of the tree, the minimum leaf size, the maximum number of leaf nodes, etc. The goal is to determine which values will optimize the machine's performance and thus cannot be choosen *a priori*. Idealy, one would train the machine with every combination of parameters and consider the resulting performance by testing the trained machine on a sample withheld from the training sample so as not to contaminate the results. Formally, we perform k-fold cross-validation whereby the training sample is split into k subsamples. One such subsample is withheld while the remaining data is used to train the machine. This is performed k times and the average performance value is recorded. The entire process is repeated for every combination of the specified hyperparameter space and values that optimize the output are choosen.

5.3. *Feature Representation and Pre-Processing*

Machine learning algorithms require a feature vector for each training example. This vector is composed of D individual numeric quantities associated with the subject which the machine will use to discern that subject from others in the training sample. To segregate 'Featured' from 'Not' our feature set draws on ZEST (Scarlata et al. 2007) and is composed of Concentration, Asymmetry, Gini, M_{20} and Source Extractor's ellipticity (See Appendix 9 for details concerning the measurement process). These non-parametric indicators have long been used to quantify galaxy morphology in an automated fashion **citations: Conselice? Peth? Huertas-company?**. Altogether, these features describe a five dimensional parameter space in which the machine attempts to distinguish between the two classes. As the RF algorithm is capabale of handling high-dimensional parameter spaces, in a future paper we will explore increasing our feature space to include parametric morphology indicators such as Sersic index and B/T ratio.

Another benefit of the RF algorithm is the flexibility with which it can accept input features. Most algorithms require that feature vectors be processed suth that all dimensions lie on the same scale. This is not necessary with an RF. The only preprocessing required in our case is the removal of morphological paramters which were not well-measured, i.e. catastrophic failures.

5.4. *Training and Validation Samples*

We are now ready to discuss the training sample in more detail. As we showed in the previous section, SWAP retires subjects far more rapidly than GZ2 by adaptively tracking volunteer skill and subject probabilities in a Bayesian framework. This provides us with a way of quickly generating large subject samples with accurate labels provided by human classifications that are dynamically generated as a function of GZ2 project time. For the following analysis we again build off of

our fiducial model where, according to Figure 3, SWAP retires XXX subjects within 10 days.

As discussed above, in addition to a training sample we also desire a validation sample to estimate the generalization (true) error of our trained machine. For this purpose we maximize the utility of our expertly classified sample. This sample thus provides training to our volunteers and verification for our machine.

5.5. *Decision Engine*

A number of decisions must be made before attempting to train the machine. Which SWAP subjects should be designated as the training sample? When should we attempt the first training session? How do we decide when the machine has successfully trained enough to be applied to unseen subjects? These are the core issues that we address in our machine learning decision engine.

Which subjects should provide the training sample? As mentioned above, SWAP yields a probability that a subject exhibits the feature of choice. A RF requires a distinct label so we use only those subjects which have crossed either of the retirement thresholds. However, subjects do not cross these thresholds with equal rates. At any given stage in the simulation, the ratio of retired 'Featured' to 'Not' is not guaranteed to be balanced, thus yielding an unbalanced training sample. However, as a first test, we allow the machine to learn on all high probability subjects.

When should we attempt the first training session? During the couple days of the simulation, SWAP retires a few hundred subjects. One could, in principle, train a machine with such a small sample, but the resulting predictions on the test sample will be exceptionally poor. Furthermore, the machine won't know that it's performing poorly. For example, if a training sample consists of 100 'Featured' subjects, the machine will subsequently predict that every member of the test sample is also 'Featured' with high probability. This is obviously wrong. In practice, a much larger training size is required for the machine to learn the true parameter space in which the feature vectors reside, but there is no hard rule for choosing this number. Because RF is a simple model, we initially require that the training sample consist of at least 10K subjects before attempting the first training session.

When has the machine trained enough? We assess our machine's learning status by first considering a learning curve, an illustration of a model's performance with increasing sample size for fixed model complexity. An example is shown in **Fig XXX** for a RF with fixed hyperparameters. The cross-validation score is the accuracy resulting from k-fold cross-validation. The training score is the resulting machine applied to the training sample. When the sample size is small, the

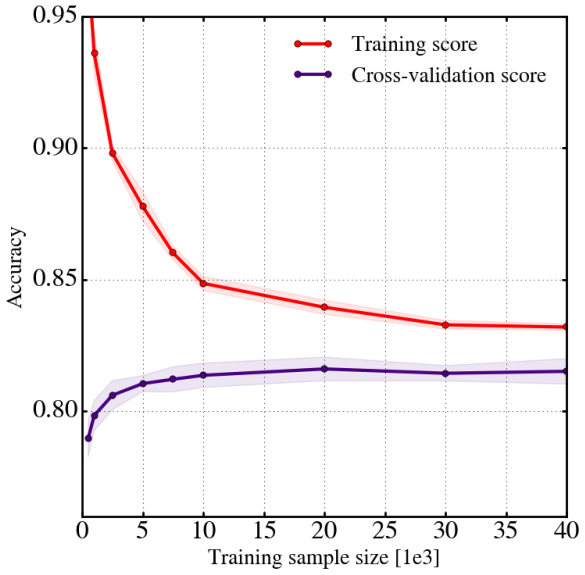

Figure 6. Learning Curve!

cross-validation score is low while the training score is high. This is a clear demonstration of a model overfitting the data. As the training sample size increases, the cross-validation score increases while the training score decreases. Eventually both plateau, regardless of how large the training sample grows. This demonstrates that, after a certain point, for a fixed complexity model, larger training sets yield little additional gain. That the training score reduces almost to the cross-validation score signifies that this particular model is not well suited to capturing the complexity of the data set. A more sophisticated model would, in turn, likely require a larger training sample.

We use this general feature of any machine learning process to guide our decision making. We cannot reproduce a true learning curve because the cross-validation procedure we perform can, in principle and in practice, yield a different set of machine hyperparameters that are most appropriate for the training sample it received that night. Instead, we look for the characteristic cross-validation score plateau. At each timestep, our software keeps record of the machine's training performance, including how well it scores on the training sample (to estimate overfitting), cross-validation score, and the best hyperparameters. When the machine's cross-validations score remains within 1% on three consecutive nights, we deem the machine's performance acceptable.

5.6. *The Machine Shop*

We can now describe a full GZX simulation. A typical run begins with human classifications processed through SWAP for several days. During that time, humans 'train' on the gold standard sample and then begin clas-

sifying the main sample. Once they retire at least 10K subjects, these are passed to the machine which trains for the first time. A suite of performance metrics are recorded by a machine *agent*, similar in construction to SWAP's *agents*. Each night, the machine agent determines whether or not the machine has sufficiently trained by assessing all previous nights of training, comparing the variation in performance metrics. Once the machine has passed the criterion laid out above, the agent introduces the machine to the test sample which consists of any subject that has not yet been seen by humans, has not yet reached retirement through SWAP, has not been previously classified by itself, and is not part of the gold standard sample.

What constitutes a machine classification? Once the trained machine is applied to the test sample it will provide predictions for every subject therein, but these predictions are not made with the same certainty. Most machine algorithms allow one to obtain a probability associated with each subject's predicted label. In the case of an RF, 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. We use this probability to assess subjects about which the machine is most confident (though we note it is not a true measure of confidence). Only subjects that receive a class prediction with $p_{machine} \geq 0.9$ are considered retired and hence removed from the system. The remaining subjects have the possibility of being classified by humans (or the machine again) during future timesteps.

This is the embodiment of our feedback loop. Those subjects on which the machine is least confident can be examined by human classifiers, potentially becoming part of the training sample during a future cycle. Ideally, this increased training sample will cover a larger portion of the parameter space in which the machine can learn. With additional subjects spanning all of the parameter space, the machine can quickly achieve its maximum performance.

In the future we will create a more sophisticated feedback mechanism whereby we can send particularly challenging subjects to super-users for identification (where these users can be identified by their high confusion matrix values as determined by SWAP, see section 4). Additionally we can perform periodic spot-checks to confirm the machine's performance by cross-checking with humans for label consistency.

6. RESULTS

We perform a full GZX simulation incorporating both SWAP and our RF machine using the fiducial SWAP run discussed in section 4.2. The machine attempts its first training on Day 8 with an initial training sample of nearly 20K subjects. The machine undergoes several additional nights of training, each time with a larger training sample. By Day 12, the machine's *agent* has assessed that the machine is suitably trained. By this point, SWAP has provided over 40K subjects as a training sample. The machine predicts class labels for the remaining 230K GZ2 subjects which are not already retired by SWAP or part of the expertly classified sample. Of those, the machine strongly predicts labels ($p_{machine} >= 0.9$ for nearly 71K subjects, immediately and dramatically increasing the overall sample of retired subjects. As the simulation progresses, retirement by both SWAP and the machine tapers off, though at different rates. We end the simulation after 32 days, having sufficiently classified well over 200K subjects. We present these results in the bottom panel of Figure 7 where GZX (red) is compared to our fiducial SWAP-only run (light blue) and GZ2 subject retirement (dark blue).

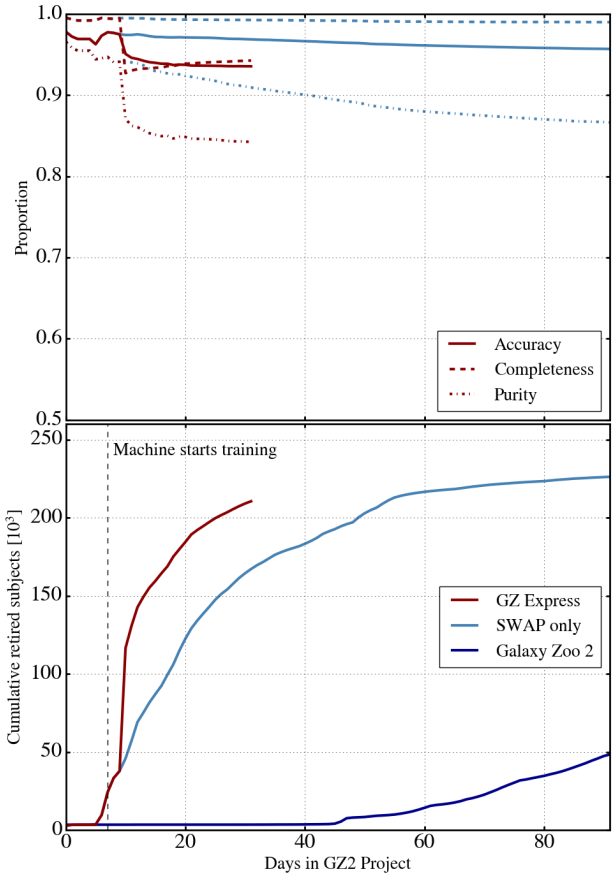

Figure 7. Incorporating the machine reduces the total time to classify over 200K subjects in the GZ2 sample to 23 days.

The top panel of Figure 7 shows our usual quality metrics for both SWAP-only and GZX, again using the $GZ2_{raw}$ labels discussed in section 3. Accuracy and completeness of the cumulative sample for the combined system each remain around 95% and purity remains around 85%. We note that incorporating the machine yields similar purity as when we allow SWAP to handle the entirety of the sample for a similar-sized subject set. Instead we seem to make a small sacrifice in the completeness of the 'Featured' sample which, in turn, affects our accuracy. Whereas SWAP alone identifies nearly all 'Featured' subjects it encounters, the machine appears to miss a significant portion of these thus dropping the completeness of that day's cumulative sample. We discuss this behavior below.

By dynamically generating a training sample through a more sophisticated analysis of human classifications coupled with a machine classifier trained through a simple feedback mechanism, we are able to retire over 200K GZ2 subjects in 27 days. SWAP alone retires as many subjects in 50 days while GZ2 requires XX *months* to retire as many and 14 months to retire the entire catalog of 285K subjects. GZX thus provides *a full order of magnitude decrease in classification time* over the traditional crowd-sourced approach. We next explore the composition of those classifications.

6.1. *Who retires what, when?*

In the top panel of Figure 8 we explore the relative contribution to subject retirement from both the machine (blue) and SWAP (red). The solid black line shows the full GZX retirement and the dashed line shows the SWAP-only fiducial run from section 4.2. Two things are immediately obvious. First, we see that, over the course of our simulation, the machine retires a total of $\sim$ 100K subjects while SWAP retires $\sim$ 110K subjects. Each shoulders approximately half of the classification burden. Secondly, the behavior in the rates of these retirements is clearly very different. Retirement through SWAP proceeds at a relatively constant rate. In constrast, the machine contributes a surge in retirement during its first couple days, quickly overtaking the human contribution, but plateaus thereafter. By Day 22, the human contribution through SWAP again becomes the primary mode of subject retirement.

We thus clearly see three epochs of subject retirement, as we presumed. In the first phase, humans are the only contributors to the total number of retired subjects. Once the machine has trained, it immediately contributes as much to subject retirement as humans. However, the machine's performance plateaus faster than human performance; the third phase of subject retirement is again dominated by human classifications.

Finally, we consider the composition of the two classifying agents in the bottom panels of Figure 8. The left (right) panel shows the retired fraction of all subjects identified as 'Featured' ('Not') as a function of project time. Over the course of the simulation, SWAP retires over 50% of all 'Featured' galaxies in the GZ2 sample while the machine retires only 18%. Furthermore, this divergence only exists for 'Featured' galaxies. In the lower right panel of Figure 8, we see that both humans and machine identify similar proportions (33-37%) of the total number of 'Not' galaxies.

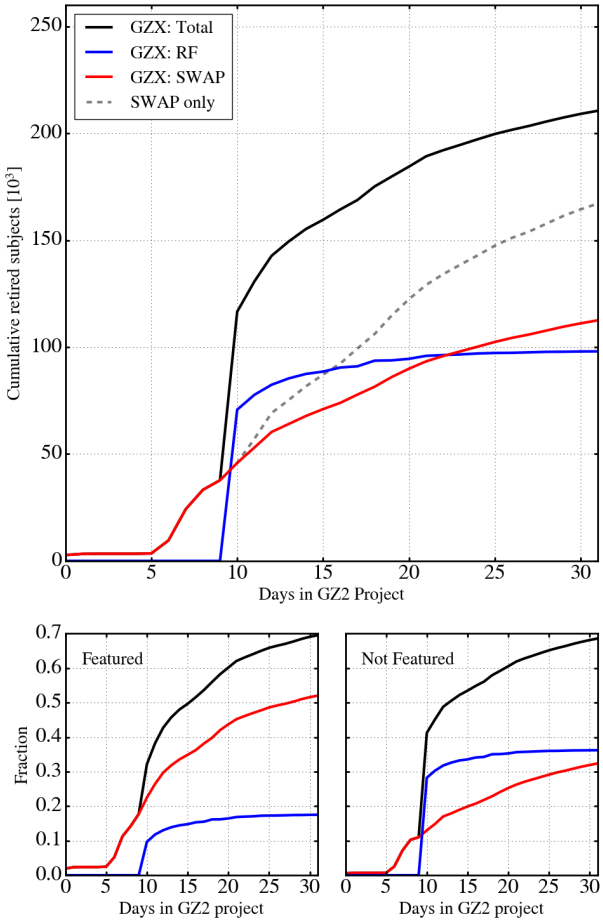

Figure 8. The relative contributions of the human and machine components to the number of retired subjects are nearly equal but display very different behaviors over the course of our simulation (top panel). The bottom panel shows the total fraction of subjects retired split by class label. Overall, GZX retires over 70% of the entire GZ2 sample in 32 days but the composition from each component is quite different. Humans retire more 'Featured' subjects than the machine and vice versa.

What is the source of this discrepancy? Why does the machine not identify with strong probability the remaining 30% of 'Featured' galaxies left unclassified by the end of our simulation? Are humans simply better at identifying featured galaxies with the machine optimized to identify the 'Not' subjects? Each night the machine trains, it receives a training sample that is con-

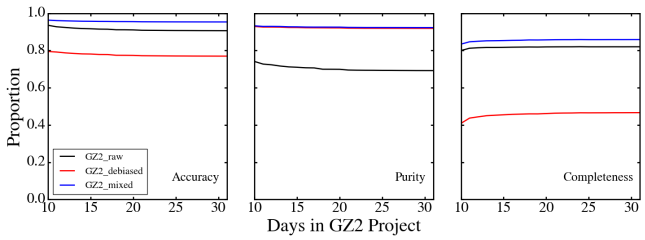

Figure 9. The performance of our RF algorithm varies depending on which set of GZ2 labels are provided as 'truth'. Each panel shows the accuracy, purity, and completeness computed on the cumulatively retired subject set retired by that day in the simulation for three different sets of GZ2 labels: the raw vote fractions used throughout the analysis thus far, the debiased vote fractions, and a combination of the two described in the text.

sistently composed of 30-40% 'Featured' galaxies, yet a similar proportion of 'Featured' galaxies in the test sample do not receive high $p_{machine}$. Is this an artifact of our particular choice in machine, in the human-machine mechanism we've implemented here, or something else entirely? In the next section we explore the machine's performance in the context of its training through human classifiers.

6.2. *Machine performance*

The analysis of the results thus far have assumed that $GZ2_{raw}$ labels are the best 'truth' for both the machine and SWAP's output. This was the best choice for SWAP since our goal was to directly compare human consensus through two different aggregation algorithms. However, the machine does not learn in the same way, nor is it presented with the same information that humans are given.

Figure 9 examines the quality of the classifications made by the machine during the course of the simulation. The black lines represent our standard quality metrics computed from $GZ2_{raw}$ labels. The red lines are the same metrics computed from $GZ2_{debiased}$ labels (themselves computed in an identical fashion to the $GZ2_{raw}$ labels described in section 3). It is immediately apparent that different definitions of 'truth' lead to significant variation in our quality metrics.

We find the machine is identifying a subsample of galaxies that are labeled 'Not' according to $GZ2_{raw}$ labels but are labeled 'Featured' according to $GZ2_{debiased}$. This can be seen in the middle panel which shows the purity of 'Featured' galaxies increases significantly when applying $GZ2_{debiased}$. This implies that subjects considered false positives (labeled as 'Not' by $GZ2_{raw}$) are truly 'Featured' galaxies. We solidify this idea by computing the blue lines in Figure 9 wherein we apply $GZ2_{debiased}$ labels to the sample of false positives and $GZ2_{raw}$ labels otherwise. The quality of the machine classifications improves in all categories.

This behavior indicates that the machine is sensitive to a different information space and is able to identify 'Featured' galaxies that humans originally classify as 'Not'; labels which were ultimately recovered through GZ2's debiasing process. We see this explicitly by examining the morphology distributions for subsamples of machine-retired galaxies as shown in Figure 10. The top row shows the difference between these distributions for the 'Featured' and 'Not' samples, while the middle row shows the false positive and false negative samples. We see that the morphology distributions for the false positive sample are nearly identical to the morphology distributions for the 'Featured' sample. This suggests that this combination of morphology indicators, specifically Gini, concentration, and ellipticity, are enough for the machine to recognize 'Featured' galaxies regardless of what labels humans provide.

Why then doesn't the machine identify *all* galaxies that are ultimately labeled as 'Featured' by $GZ2_{debiased}$ labels? We know it does not as evidenced by the paltry 42% $GZ2_{debiased}$ completeness displayed in the right-most panel of Figure 9. This is likely a consequence of the machine training on labels provided by SWAP which match $GZ2_{raw}$ labels to 95%. We would expect a higher level of completeness if we instead provided the machine with debiased labels.

The machine is retiring more 'Featured' galaxies than $GZ2_{raw}$ labels suggest but not all that the $GZ2_{debiased}$ labels suggest. We thus conclude that two effects contribute to the disparity in the retirement ratios of 'Featured' galaxies between humans and machines indicated in Figure 8.

1.) The machine learns on different information in a different way than humans, and

2.) The machine trains on labels which have not yet been debiased.

The first is largely the result of the particular machine and feature vector we have chosen, though it remains to be seen if this generalizes to other machine algorithms. The second is a consequence of the current state of our human+machine combination and advocates for a dynamical debiasing process to be performed after SWAP aggregation and before labels are passed to the machine. This would likely provide the best sample purity and completeness, however, such work is beyond the scope of this paper.

7. LOOKING FORWARD

We've demonstrated the first practical framework for combining human and machine intelligence in galaxy morphology classifications tasks. By reprocessing the original Galaxy Zoo 2 classifications with SWAP, incorporating a supervised machine learning algorithm, and

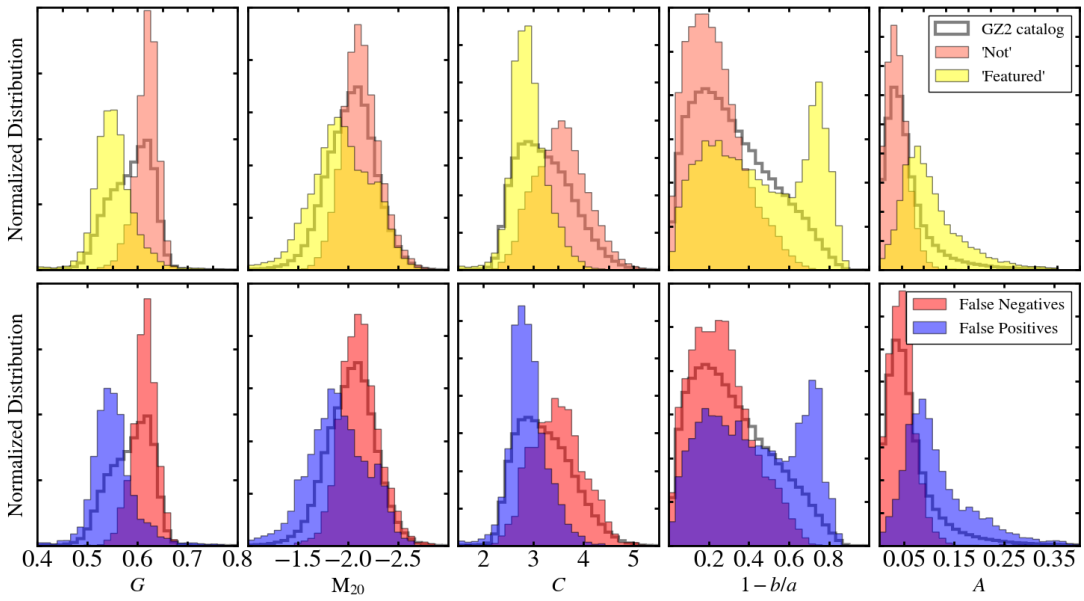

Figure 10. What the machine can and can't classify.

implementing a simple decision engine which guides how these two agents interact, we achieve an order of magnitude increase in the classification rate. Our results have implications for the future of citizen science in general and Galaxy Zoo in particular, however we focus here on a brief discussion of our next steps and potential applications to large upcoming surveys.

7.1. *From Simple to Sophisticated*

GZX represents perhaps one of the simplest ways to combine human and machine intelligence. Its impressive performance motivates a higher level of sophistication. Here we discuss a handful of the most straightforward steps that will advance GZX to the next level, implementation that will not only further improve the rate of classification but also the overall quality of those classifications.

A more sophisticated feedback mechanism. As we mentioned in section 5.6, our feedback mechanisms is passive. However, SWAP allows us to leverage the most skilled volunteers to field those galaxies which are the most difficult for either human or machine classify.

A more sophisticated machine learning algorithm. The random forest algorithm is not easily adapted to handle measurement errors or class labels with continuous distributions. To fully utilize the information provided by SWAP, algorithms which can handle continuous distributions for the input and output subject labels should be explored. These include deep convolutional neural networks (CNN) or Latent Dirichlet allocation (LDA), an algorithm that is frequently used in document processing.

A larger confusion matrix for SWAP. The core of SWAP revolves around the confusion matrix that an agent assigns to each volunteer. Extending the confusion matrix to three dimensions should be relatively straightforward and would allow for more detailed morphology tasks. We note that most questions in the Galaxy Zoo decision tree could be easily modified to consist of three responses.

On-the-fly machine ensembles. Less straightforward than the previous tasks on our to-do list, this is nevertheless the next logical extension of GZX. SWAP allows aggregation of the crowd to produce meaningful subject classification. We see no reason why this same procedure could not be applied to an ensemble of machine classifiers as hinted at in Figure 1. In this scheme, several disparate machine algorithms could train simultaneously on the same training sample with their predictions aggregated through SWAP.

7.2. *Dealing with the Data Deluge*

The work in this paper is motivated by the need for increased speed in the face of several large surveys coming online in the next 5 years. These projects promise to deliver more raw data per night than we have accumulated in the past decade. Though these efforts are concentrated on probing dark energy, we can maximize the science output of these surveys. We thus briefly discuss how GZX might handle the data deluge.

Is an order of magnitude enough? We've demonstrated that we can achieve an order of magnitude in-

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

Is the performance enough? GZX as implemented here can provide accuracy around 95%. That would leave us with hundreds of thousands of galaxies with unreliable classifications. Such a dataset will provie a wealth of information for a slew of science cases, however, to extract the most from these datasets, additional gains in performance may be required. One solution is to incorporate deep learning algorithms which can provide higher quality classifications at the expense of interpretability. Ensembles of machine classifiers aggregated through SWAP could provide another promising avenue. A third solution is presented in our sister publication (Wright et al. 2016, submitted) whereby accuracy is dramatically increased through an entirely different human-machine combination. In that paper we explore the effects of averaging the human and machine classification scores and find that the combined score yields the most reliable classification.

8. CONCLUSIONS

We outline and test Galaxy Zoo Express, a novel system for the efficient classification of galaxy morphology tasks. Our system incorporates the native ability of the human mind to identify the abstract and novel with machine learning algorithms which 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 Galaxy Zoo 2 classification data we can increase the rate of classification by a factor of 4-5, requiring only 90 days of project time to sufficiently classify nearly 80% of the entire Galaxy Zoo 2 galaxy sample.

Furthermore, we have implemented and tested a simple Random Forest algorithm and developed a decision engine that delegates tasks between human and machine. We show that even a simple machine is capable of providing significant gains in the rate of classification allowing us to retire over 70% of GZ2 galaxies in just 32 days of GZ2 project time. This is an order of magnitude decrease in the time required compared to the original Galaxy Zoo 2 project. This is achieved without sacrificing the quality of classifications as we maintain accuracy well above 90% throughout our simulations. Additionally, this machine, though simplistic, can provide valuable insight to the classification process because it learns on a 5-dimensional parameter space consisting of a variety of traditional non-parametric morphology identifiers.

The gain in classification speed allow us to come within reach of being able to handle the massive amounts of data that upcoming large-scale surveys like LSST, Euclid, and WFIRST promise to provide. With a few modest upgrades to our GZX framework we are confident this method will be able to handle the gargantuan task of morphology classification of the billions of galaxies soon to be cataloged.

9. ACKNOWLEDGEMENTS

MB thanks John Wallin, Steven Bamford, and Boris Häußler for discusions which helped illucidate things and stuff. This work is funded by NSF Grant XXXXX. Additional support was provided to MB through New College and Oxford University's Balzan Fellowship as well as the University of Minnesota's Doctoral Dissertation Fellowship. Travel funding was further supplied to MB under the University of Minnesota's Thesis Research Travel Grant.

APPENDIX

A. MEASURING MORPHOLOGICAL PARAMETERS ON SDSS CUTOUTS

We measure all these things on cleaned postage stamps of SDSS i-band imaging. Images were obtained from DR12. Concetration measures the ...

$$C = 5\log(r_{80}/r_{20}) \tag{A1}$$

where r_{80} and r_{20} are the radii containing 80% and 20% of the galaxy light respectively. Large values of this ratio tend to indicate disky galaxies, while smaller values correlate with early-type ellipticals.

Asymmetry quantifies the degree of rotational symmetry in the galaxy light distribution (not necessarily the physical

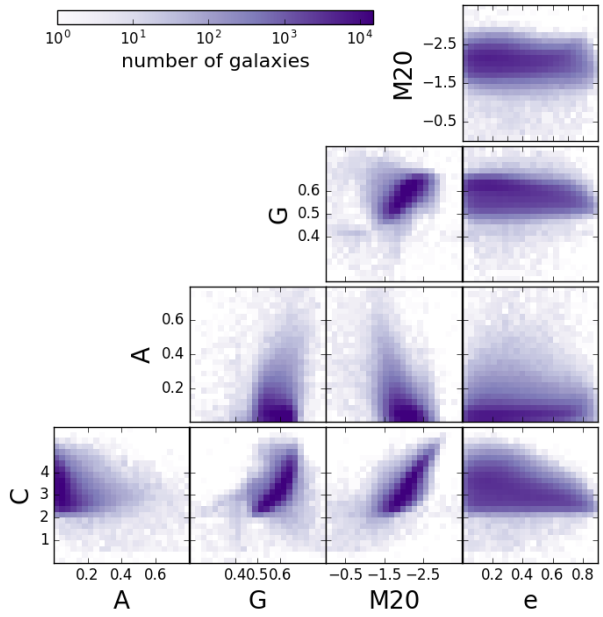

Figure A1. That's a lot of parameters!

shape of the galaxy as this parameter is not highly sensitive to low surface brightness features).

$$A = \frac{\sum_{x,y} |I - I_{180}|}{2 \sum |I|} - B_{180} \tag{A2}$$

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

The Gini coefficient, G, describes how uniformly distributed a galaxy's flux is. If G is 0, the flux is distrubuted homogeneously among all galaxy pixels.; while if G is 1, all of the light is contained within a single pixel. This term correlates with C, however, unlike concentration, G does not require that the flux be concentrated within the central region of the galaxy. We calculate G by first ordering the pixels by increasing flux value, and then computing

$$G = \frac{1}{|\bar{X}|n(n-1)} \sum_{i}^{n} (2i - n - 1)|X_i| \tag{A3}$$

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

M_{20} is the second order moment of the brightest 20% of the galaxy flux.

$$M_{tot} = \sum_{i}^{n} f_i[(x_i - x_c)^2 + (y_i - y_c)^2] \tag{A4}$$

$$M_{20} = \log_{10}(\frac{\sum_i M_i}{M_{tot}}), \quad \text{while} \sum_i fi < 0.2 f_{tot} \tag{A5}$$

B. EXPLORING SWAP'S PARAMETER SPACE

Initial agent confusion matrix. In our fiducial simulation each volunteer was assigned an agent with confusion matrix $(P_{F,0}, P_{N,0}) = (0.5, 0.5)$, which presumes that volunteers are no better than random classifiers. We perform two simulations wherein we allow $(P_{F,0}, P_{N,0}) = (0.4, 0.4)$, slightly obtuse volunteers, and $(P_{F,0}, P_{N,0}) = (0.6, 0.6)$, slightly astute volunteers with everything else remaining constant. Results of these simulations compared to the fiducial run are shown in Figure **??**. We find that we are largely insensitive to the initial agent confusion matrix probabilities both in terms of the overall number of retired subjects and in the quality of their SWAP labels.

Predictably, when $(P_{F,0}, P_{N,0})$ are low, we retire fewer subjects in the same time frame and more subjects when $(P_{F,0}, P_{N,0})$ are high. This is easy to understand since it takes longer for volunteers to become strong, astute classifiers

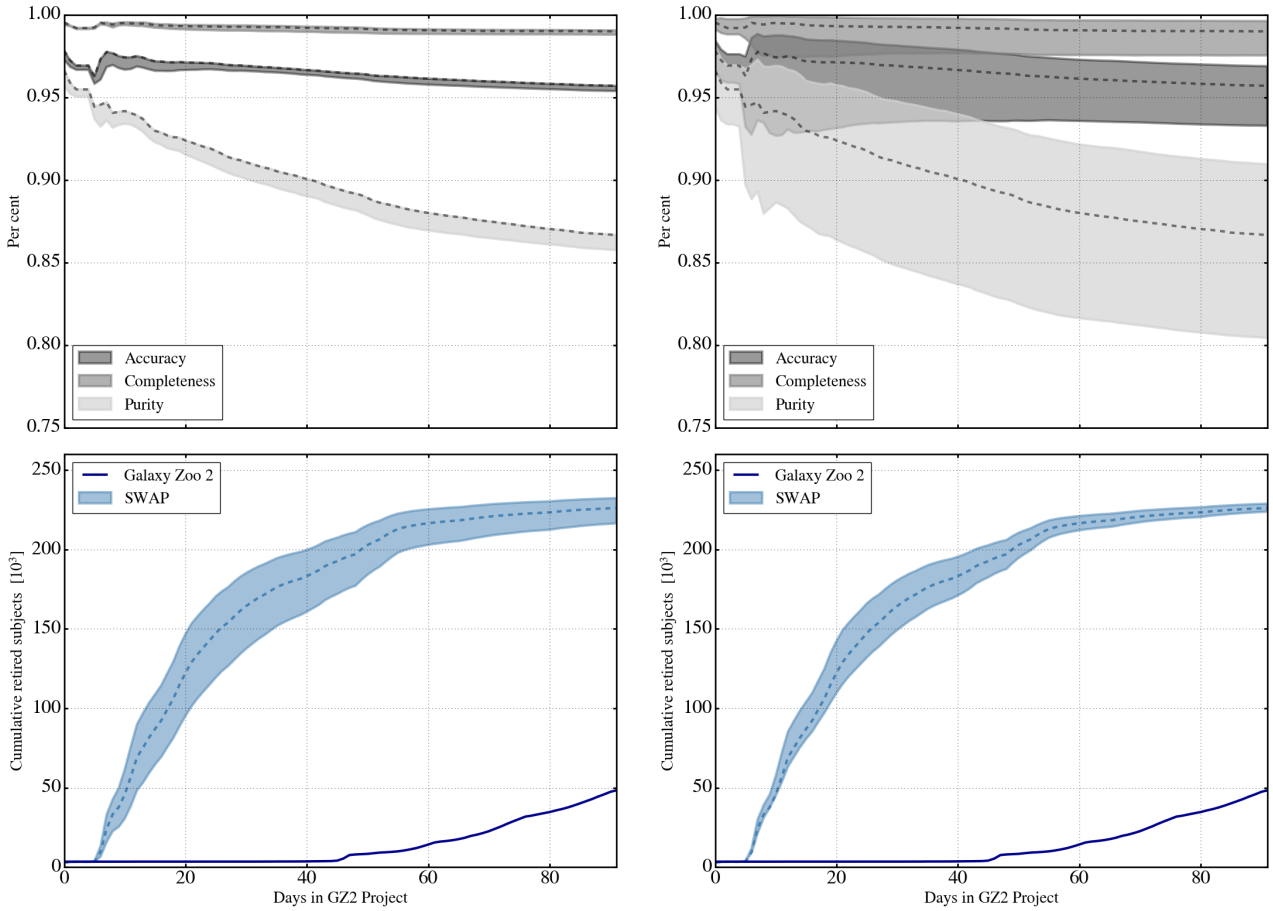

Figure B2. GZX/SWAP output as a function of GZ2 project days for a range of subject prior probabilities.

when they are initially given values denoting them as obtuse. Even with this handicap, most volunteers become astute classifiers by the end of the simulation. Overall, we retire $\sim$ 225K $\pm$3.5% subjects as shown by the light blue spread in the bottom panel of Figure **??** where the dashed blue line denotes the fiducial run.

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

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

Subject prior probability, p_0. The prior probability assigned to each subject is an educated guess of the frequency of that characteristic in the scope of the data at hand. For galaxy morphologies, this number should be an estimate of the probability of observing a desired feature (bar, disk, ring, etc.). In our case, we desire to simply find galaxies that are 'Featured', however, this is dependent on mass, redshift, physical size, etc. The original GZ2 sample was selected primarily on magnitude and redshift. As there was no cut on the galaxy size (with the exception that each galaxy be

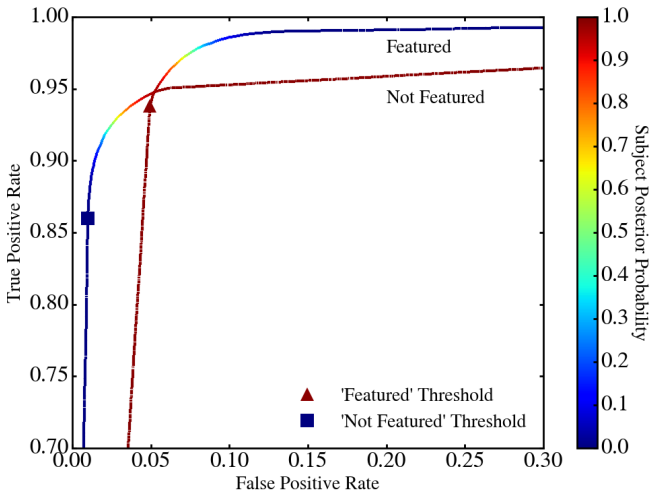

Figure B3. Identifying 'Featured' subjects is independent of identifying 'Not' subjects. Both ROC curves use all subjects processed by SWAP where the score used to create the ROC curve is simply each subject's achieved posterior probability. The Featured curve demonstrates how well we identify 'Featured' subjects with a threshold of 0.99, while the Not Featured curve demonstrates how well we identify 'Not' subjects with a threshold of 0.004. Typically, best performance is achieved with scores that lie closest to the upper left corner. Our 'Featured' threshold is nearly as optimumal as possible though our 'Not' threshold could have been slightly better.

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

We run several simulations where p_0 is allowed to take values 0.2, 0.35, and 0.8 and compare these to the fiducial run where $p_0 = 0.5$, everything else remaining constant. The results are shown in Figure B2, where again we find that SWAP is consistent in terms of the total number of subjects retired during the simulation which varies by only 1%. However, as can be seen in the top panel, the variation in our quality metrics is more pronounced and deserves some discussion.

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

We can get a handle on these variations more intuitively by considering the following. Recall that our retirement thresholds, t_F and t_N, have not changed in these simulations. Thus when p_0 is small, the subject probability is already closer to t_N, and more subjects are classified as 'Not' compared to the fiducial run. Similarly, when p_0 is large, some of these same subjects can instead be classified as 'Featured' because the prior probability is already closer to t_F. Obviously, both outcomes cannot be correct and we find that the simulation with $p_0 = 0.8$ performs the worst of any run which is a direct reflection of the fact that this prior is not suitable for this question nor for this dataset. For the mass, size, and redshift range of subjects in GZ2, we would not expect that 80% of them are 'Featured'. Indeed, the best performance is achieved when $p_0 = 0.35$. This reflects the actual distribution of 'Featured' subjects in the GZ2 sample as well as being similar to the expected proportion of 'Featured' galaxies in the local universe, depending on your definition **(cite studies of distribution of early and late type gals in local universe?).** Thus, the take-away here is to choose your prior wisely since a value far from the correct value can have a significant impact on the quality of your classifications.

Retirement thresholds. Retirement thresholds are directly related to the time that a subject will spend in SWAP before retiring. If we lower t_F (and/or raise t_N), more subjects will be retired compared to the fiducial run as each subject will have a smaller swath of probability space in which to bounce back and forth before crossing one of these thresholds. On the other hand, if we raise t_F (and/or lower t_N), it will take longer for subjects to cross one of these

thresholds. Additionally, this will also increase the likelihood of some subjects never crossing either threshold as there are always some which are nudged back and forth indefinitely through probability space.

What thresholds should one choose? To answer this question, we consider Figure B3 which depicts the receiver operating characteristic (ROC) curve for our fiducial simulation. The solid line shows the curve when considering the 'Featured' threshold while the dotted line corresponds to 'Not'. The square and triangle represent our thresholds, $(t_F, t_N) = (0.99, 0.004)$, on that curve at the end of the simulation. We see that t_F is nearly optimal but t_N could be improved upon.

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

REFERENCES

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

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

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

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

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

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

Willett, K. W., Lintott, C. J., Bamford, S. P., et al. 2013, MNRAS, 435, 2835