

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, AND PHIL MARSHALL

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

Galaxy morphology has been a facet of study for nearly 100 years. The visual structure of galaxies provides clues into the nature of galactic evolution and history. Galaxy morphology remains just as important today with new and rare galaxy types that probe various avenues of galactic evolution, from the discovery of the green peas to ... These diverse galaxy types are Several studies utilizing visual classifications have lead to advanced understanding of the evolution of galaxies: bar fractions, merger rates, etc.

Upcoming surveys like LSST and Euclid will be coming online in the next five years. 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 put in place to handle the vast nature of the data. Standard visual morphology methods will be unable to cope with the size and current automated morphologies provide only statistical morphological structure with large uncertainty and impurities. Methods must be developed which can efficiently and accurately quantify and qualify that nature of galaxy morphology.

In this work we present the first system which combines both human and machine intelligence to the task of galaxy morphology. We demonstrate this through a re-analysis of visual galaxy morphology classifications collected during the Galaxy Zoo 2 project. We combine these classifications with a suite of non-parametric morphology indicators typically used for automated morphologies. Our method provides an order of magnitude improvement in the rate of galaxy morphology classi-

cations.

Old Stuff vvv

The age of Big Data is upon us. The astrophysics community has been shifting focus, preparing for the ways in which both science and how we perform that science will change. Several stalwart astronomical data repositories have begun integrating analytical tools at the source, most notably CasJobs, because downloading data will become a thing of the past when these data are terabytes in size. Advances in data storage, telescope reduction pipelines, and data access are all required.

In addition to how we manage and compute on Big Data, we must also prepare for how we identify objects in the sky. We will have access to more sky, both in terms of area and depth, than ever before. Once those data have been reduced and stored, we must find efficient, clever ways of identifying transients, radio relics, gravitational lenses, galaxy morphologies, and more, all while allowing for an unprecedented amount of new serendipitous discoveries. If not, LSST, Euclid, WFIRST will overwhelm us.

In this paper we consider the particular problem of galaxy morphology. This challenge is actually several combined because it encapsulates the need to identify the mundane from the unique or rare and, ideally, requires an incredible amount of detail in order to withdraw useful science. Additionally, morphology is a great place to start because we can already begin to plan for the future by considering the data we currently have. The imaging techniques of future surveys will change mostly in resolution and depth; things we can account for. Another reason galaxy morphology is a useful example is that we can draw on vast, well-established citizen

science projects which have contributed to several past publications and have lead to serendipitous discovery on multiple occasions. There is no doubt that to spurn this resource would be a disservice to science.

Of course, galaxy morphology has drawbacks as well. The community firmly clings to a classification system developed nearly 100 years ago which works well only for galaxies in the local universe. However, the endeavor has merit as understanding the nature of galactic structure provides us with a means to disentangle galactic evolution.

In this paper we present a framework for combining human and machine intelligence for morphology classification tasks. This idea is not new in computer science circles as several studies have been done, many on Galaxy Zoo data [**Microsoft research guy, Chris: citations, plz.**] However, these studies are generally unknown to the astrophysics community and are, in some cases, impractical. The novelty of our approach lies in its practicality, simplicity, and in our ability to explore the repercussions of our system by simulating various outcomes on previously collected data.

We draw on visual classifications from the Galaxy Zoo project. We implement an algorithm first developed for the Space Warps Zooniverse project and demonstrate for the first time that this algorithm is well suited to galaxy morphology classification and significantly increases the rate of classification compared to Galaxy Zoo. We measure standard galaxy morphology indicators and use these as a feature vector for an implementation of a Random Forest machine learning algorithm. We develop a decision engine which dictates how human and machine classifications are combined and show that our approach achieves a full order of magnitude increase in the classification rate without sacrificing the quality of those classifications. We begin with an overview of the Galaxy Zoo Express framework.

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 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 both subjects with spectroscopic and photometric redshifts out to $z < 0.25$. In total, 285,962 subjects were classified in the GZ2 Main Sample catalogs¹. Of these, 243,500 have spectroscopic redshifts while 42,462 have only photometric redshifts.

Subjects were shown as color composite images via a web-based interface wherein volunteers answered a se-

¹ data.galaxyzoo.org

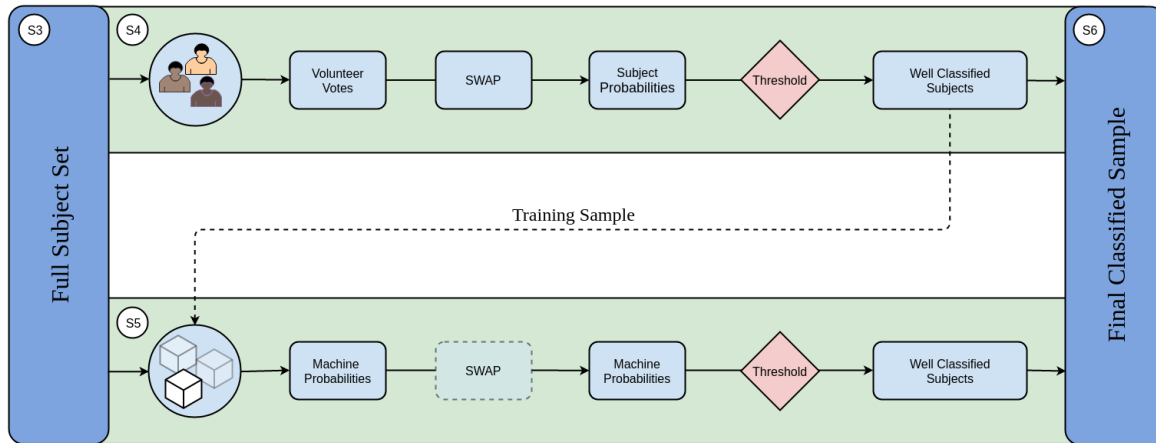


Figure 1. Schematic of our hybrid system. Human classifiers are shown images of galaxies via the Galaxy Zoo web interface. These classifications are recorded and processed according to section XXX. As a result of the processing, those subjects whose probabilities cross the classification thresholds are 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.

ries 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 a set of ‘true’ labels for each GZ2 subject to which we can compare labels assigned by Galaxy Zoo Express (GZX). 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 GZ2 raw vote fractions as the closest comparison. Specifi-

cally, we take the majority raw 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’. 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 software we use requires that every volunteer see a subset of subjects that are expertly identified by a member of the GZ team. We thus utilize only those classifications made by one of the 30,894 volunteers that identified one or more of our gold standard sample (see Section 4.1). 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 to approximately 14 million.

4. EFFICIENCY THROUGH CLEVER HUMAN-VOTE PROCESSING

Galaxy Zoo requires a large number of independent classifications for each subject and this value is typically set at forty individual volunteer classifications. Once a project reaches completion, Willett et al. (2013) down-weight inconsistent and unreliable volunteers. While this process reduces input from malicious users and ‘bots’, it doesn’t reward 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 instead employ software adapted from the Space Warps Zooniverse project (Marshall et al. 2016) which searched for and discovered several gravitational

lens candidates in the CFHT Lensing Survey (cite XXX). 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. 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

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

where $N_{“X”}$ 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.

In addition to agent probabilities, each subject begins with 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 agent should update the subject’s prior probability 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 $P(“A”|A) = 1.0$ and $P(“N”|N) = 1.0$) 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 $P(“A”|A) = 0.0$ and $P(“N”|N) = 0.0$) 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 to exhibit the feature of interest (or highly unlikely!). While most subjects will cross a classification threshold, not

all do. In particular, subjects for which volunteers are unsure will simply bounce back and forth in probability space indefinitely. 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 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 found that the SWAP software does not perform well without the use of designated training images. Furthermore, the software requires that these training images be introduced at the beginning of the project to allow volunteer confusion matrices 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 expertly classified into various TTypes. Though helpful, this particular classification isn’t directly comparable to GZ2. Instead, we reclassified the subsample amongst the Galaxy Zoo science team by building a small project on the Zooniverse platform. The question posed to our science team was identical to the original question posed to the volunteers. Approximately 15 members of the GZ science team contributed to these classifications and at least five experts saw each galaxy. Experts this case range from advanced graduate students, post docs, and several seasoned faculty members. Once classification was complete, the votes were aggregated and a simple majority was used to provide expert label to the 3500 subjects.

While 3500 subjects is a sizable sample, not every volunteer saw 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 never classify at least one of these 3500 subjects. This reduces our raw data set from 16 million clicks to 14 million; from 90K unique volunteers to 30K.

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 sub-

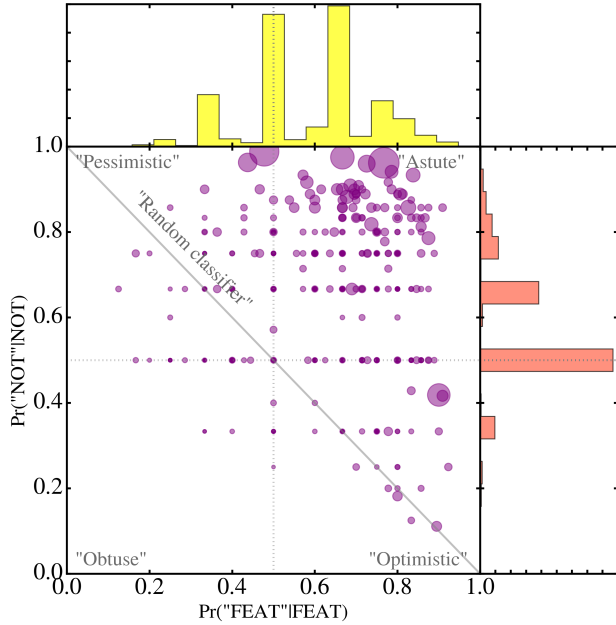


Figure 2. Volunteer probabilities for our fiducial SWAP run.

jects. Since it is implicitly assumed that a subject’s classifications are independent and random, the order of the classifications should have a small effect, if any, on the results. When running a simulation, which pulls from the database according to timestamps, the training image classifications are the first to be processed with SWAP.

Figure 2 demonstrates the volunteer training we achieve with this scheme (figure adapted from Marshall et al. (2016)). This figure shows the achieved confusion matrices for 1000 volunteers where the size of the circle is proportional to the number of training images each volunteer saw. Most of our volunteers fall into the ‘Astute’ category indicating they are exceptional at correctly identifying both ‘Featured’ and ‘Not’ (smooth) subjects. The full distribution of each probability for all volunteers can be seen in the histograms. The spikes at $p = 0.5$ for each distribution indicate that of those volunteers who saw only one subject (say, ‘Featured’), their probability in the other (‘Not’) remains unchanged.

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 cho-

sen. 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’, $P_0(“F”|F)$
- $P_{N,0}$, the initial probability that a volunteer identifies a subject as being ‘Not’, $P_0(“N”|N)$
- 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 consider 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 more lenient definition and consider a subject GZ2-retired after it achieves just 30 volunteer votes. SWAP retirement is determined by a subject’s posterior probability crossing either of the retirement thresholds defined above.

On the other hand, we don’t want 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

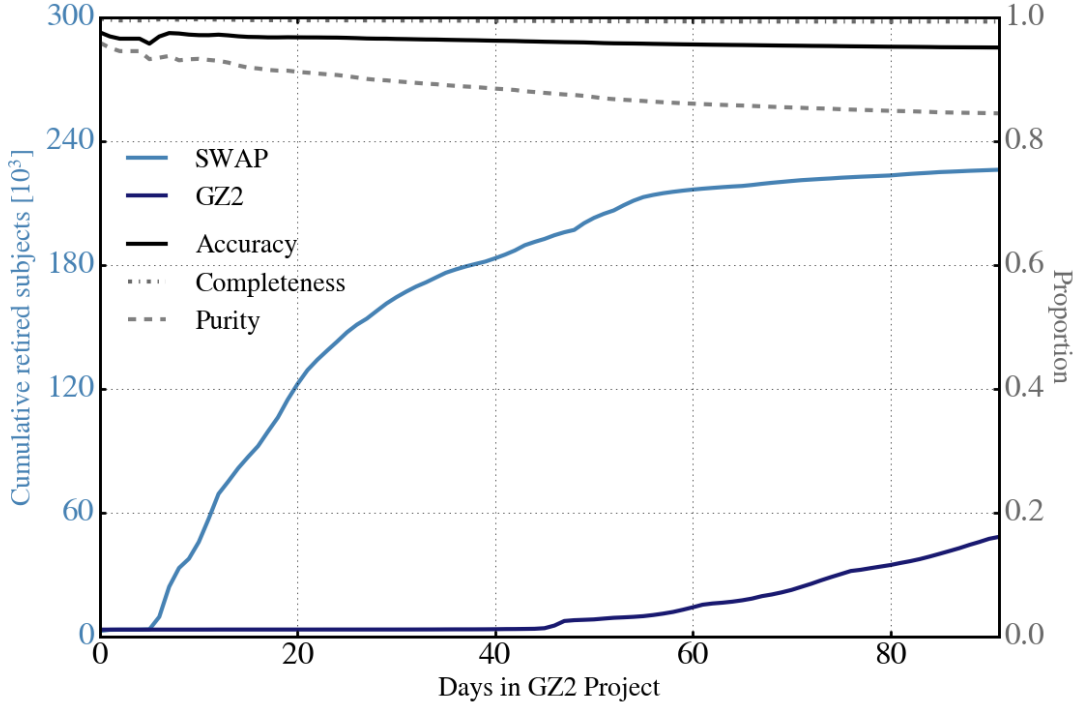


Figure 3. Fiducial SWAP simulation demonstrating the dramatic increase in the subject retirement rate as a function of GZ2 project time (in light blue) compared with the original GZ2 project (dark blue) corresponding to the left hand axis. Not only is efficiency increased by nearly an order of magnitude, 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.

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 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 ~ 225 K 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 an order of magnitude. However, the exact nature of subject retirement and associated quality metrics will be, in part, a function of initial SWAP parameters. We now turn to a discussion of how classification and retirement change as a function of SWAP input parameters.

4.3. 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

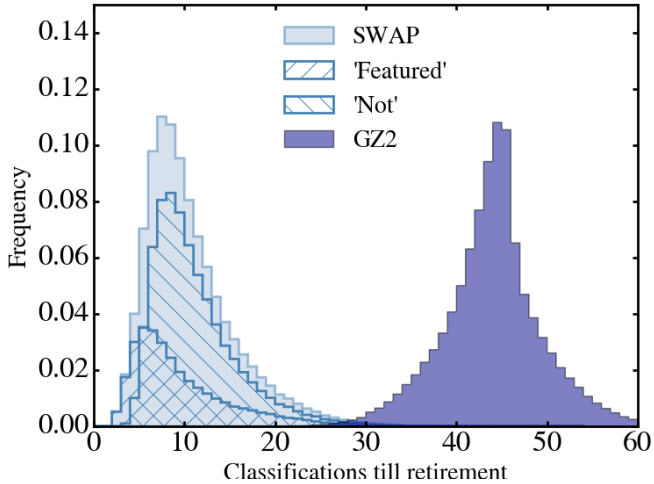


Figure 4. SWAP requires less human effort than GZ2 as evidenced by comparing the number of classifications until retirement for the $\sim 225\text{K}$ subjects retired by both SWAP and GZ2. GZ2 requires ~ 45 classifications per subject before retirement. In contrast, most subjects only need approximately 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.

everything else remaining constant. Results of these simulations compared to the fiducial run are shown in Figure 5. We find that we are largely insensitive to the initial agent confusion matrix probabilities both in terms of the overall number of retired subjects and in the quality of their SWAP labels.

Predictably, when $(P_{F,0}, P_{N,0})$ are low, we retire fewer subjects in the same time frame and more subjects when $(P_{F,0}, P_{N,0})$ are high. This is easy to understand since it takes longer for volunteers to become strong, astute classifiers when they are initially given values denoting them as obtuse. Even with this handicap, most volunteers become astute classifiers by the end of the simulation. Overall, we retire $\sim 225\text{K} \pm 3.5\%$ subjects as shown by the light blue spread in the bottom panel of Figure 5 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

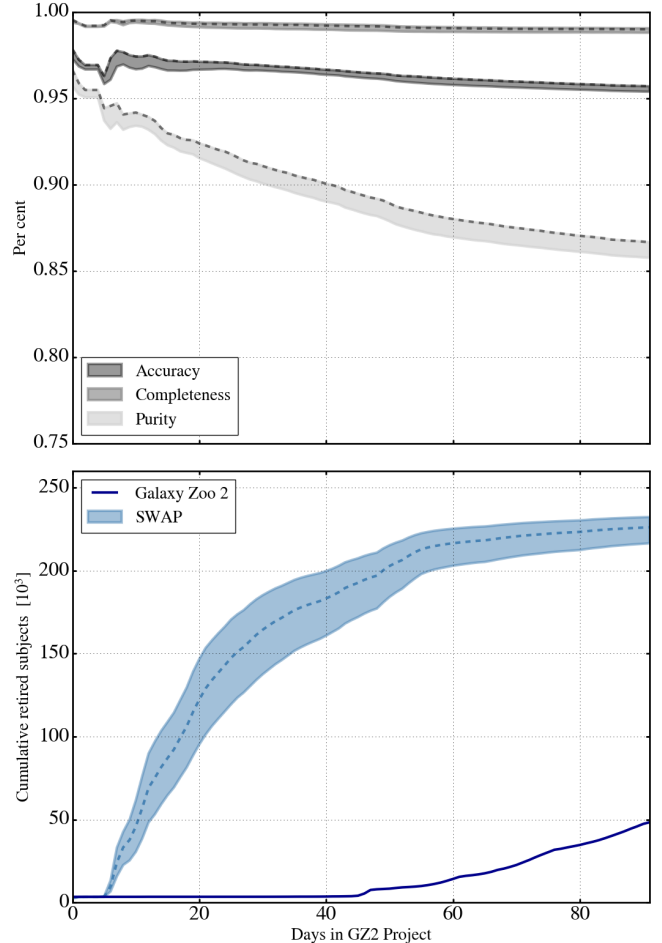


Figure 5. GZX/SWAP output as a function of GZ2 project days for a range of initial confusion matrix values.

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

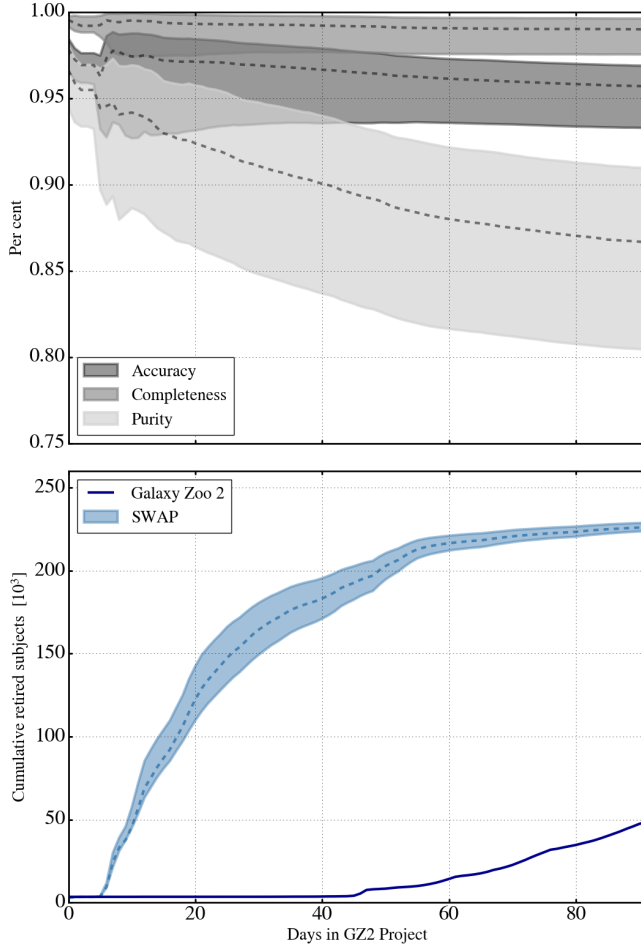


Figure 6. GZX/SWAP output as a function of GZ2 project days for a range of subject prior probabilities.

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 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 6, 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 7 which depicts the receiver operating characteristic (ROC) curve for our fiducial

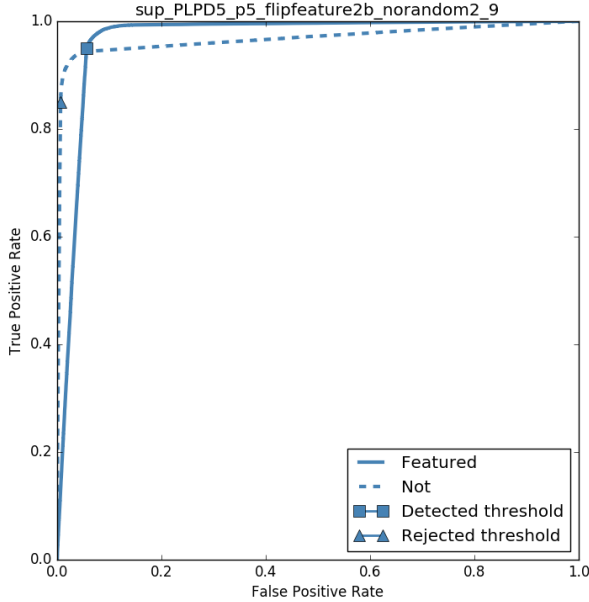


Figure 7. Retirement thresholds are adequate for our purposes.

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.

4.4. 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 8 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

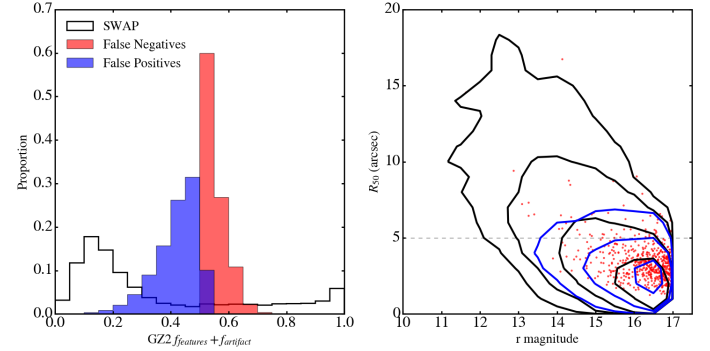


Figure 8. SWAP and GZ2 labels disagree approximately 5% of the time, depending on the choice of initial SWAP parameters. The disagreement stems, in large part, from ‘true’ labels derived from uncertain GZ2 vote fractions as shown in the left panel where the majority of SWAP’s false negative and false 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.

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.

4.5. Summary

We have demonstrated that, regardless of the initial configuration of the SWAP software, we achieve nearly an order of magnitude 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 sample accuracy over 95%, almost perfect completeness of subjects designated as ‘Featured’, and with a sample 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. Under this assumption, we now turn our focus towards incorporating machine classifiers utilizing these SWAP-retired subjects as training samples.

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 chosen *a priori*. Ideally, 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 chosen.

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 (Scarlati et al. 2007) and is composed of Concentration, Asymmetry, Gini, M_{20} and Source Extractor’s ellipticity (See Appendix A 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 capable 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 such 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 parameters 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 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-validation 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 classifying 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 $p_{\text{machine}} \geq 0.9$ 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 feed-

back 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 of the run 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. At this point, SWAP has now 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 ($p_{\text{machine}} \geq 0.9$ for nearly 71K subjects, thus 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 9 where GZX (red) is compared to our fiducial SWAP only run (light blue) and GZ2 subject retirement (dark blue).

The top panel of Figure 9 shows our usual quality metrics for both SWAP-only and GZX, again using the GZ2 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*

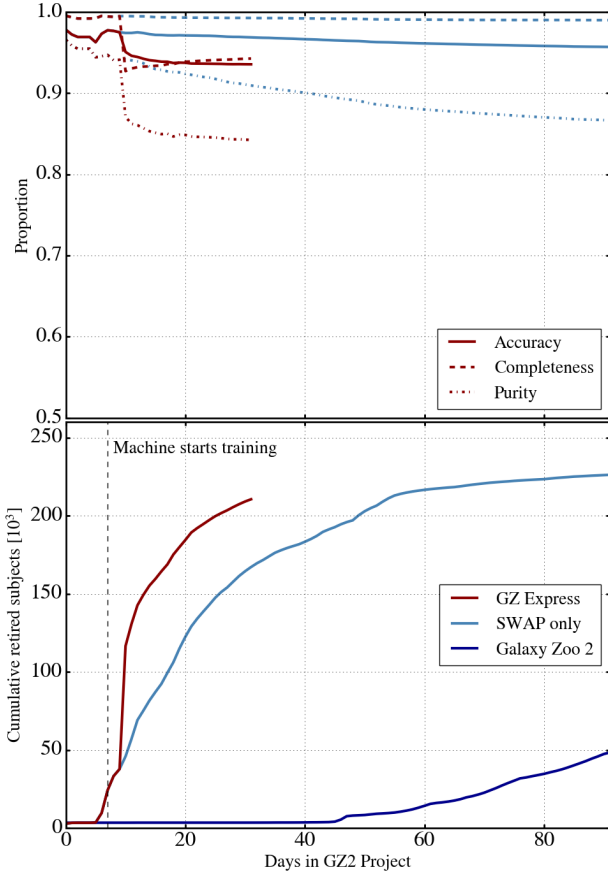


Figure 9. Incorporating the machine reduces the total time to classify over 200K subjects in the GZ2 sample to 23 days.

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 10 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 $\sim 100\text{K}$ subjects while SWAP retires $\sim 110\text{K}$ 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 from Days 5 through 25 before beginning to taper off. In contrast, the machine contributes a massive surge in retirement

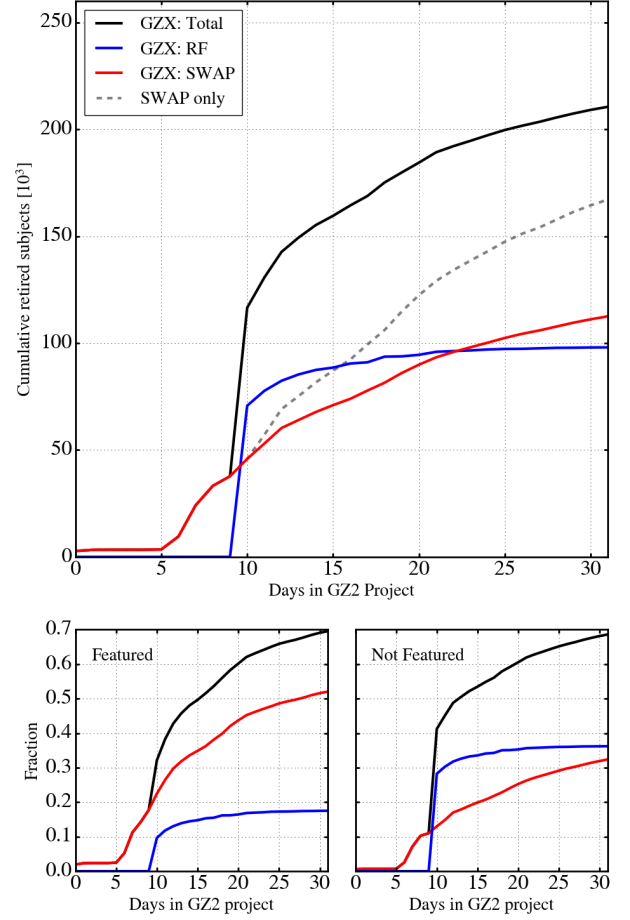


Figure 10. The relative contributions of the human and machine components to the number of retired subjects are nearly equal but display very different behaviors over the course of our simulation (top panel). The bottom panel shows the total fraction of subjects retired split by class label. Overall, 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.

during its first couple days quickly overtaking the human contribution but plateaus dramatically thereafter. By Day 22, the human contribution through SWAP again becomes the primary mode of subject retirement.

Thus we 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 becomes a workhorse and takes a monstrous share of the retirement burden though humans also contribute. The machine quickly reaches it’s peak performance, classifying all subjects it’s capable of recognizing and the third phase is thus left to humans who again dominate the the rate of subject retirement.

Finally, we consider the composition of the two classifying agents in the bottom panels of Figure 10. The left panel shows the retired fraction of all subjects identified

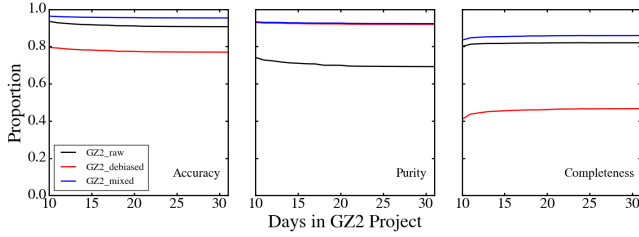


Figure 11. The performance of our RF algorithm varies depending on which set of GZ2 labels are provided as ‘truth’. Each panel shows the accuracy, purity, and completeness computed on the cumulatively retired subject set retired by that day in the simulation for three different sets of GZ2 labels: the raw vote fractions used throughout the analysis thus far, the debiased vote fractions, and a combination of the two described in the text.

as ‘Featured’ according to the GZ2 labels we assigned in section 3 and similarly for those subjects designated as ‘Not’ in the right panel. Of the cumulative subjects SWAP retires each night, 38% to 40% of them are consistently and correctly identified as ‘Featured’ culminating in the retirement of over 50% of the total number of ‘Featured’ subjects in the GZ2 sample. In contrast, of the cumulative subjects retired by the machine, only 14% of them are correctly identified as ‘Featured’ the first night the machine is run, increasing to only 20% by the end of the simulation. In total, the machine retires about 18% of the actual subjects labeled as ‘Featured’ in the GZ2 sample. This suggests that humans are optimized to identify obviously ‘Featured’ galaxies while the machine seems to identify the converse. That humans and machine might be optimized somewhat divergently is perhaps not unheard of when one considers that they are presented with different information and learn in entirely distinct ways. However, the machine trains on a sample consistently composed of 40% ‘Featured’ subjects provided by SWAP yet doesn’t identify as many when applied to the test sample each night does deserve deeper consideration. In the next section we explore the machine’s retirement to understand the core of its performance.

6.2. Machine-only performance

The analysis of the results thus far have assumed that comparison to the GZ2 labels computed in section 3 are appropriate as ‘truth’ for both SWAP’s output and the machine’s. This was obviously the best choice for SWAP’s output as our goal was to directly compare human consensus through two different aggregation algorithms. However, the machine does not learn in the same way humans learn and is not presented with the same information that humans are given. It is thus reasonable to presume that the machine and humans are biased in different ways. For example, it is already well

known that visual classifications are highly sensitive to a galaxy’s redshift and surface brightness. For this reason, Willett et al. (2013) compute corrections to the volunteer vote fractions.

However, we surmise that the machine is not biased in the same way. It receives a 5 dimensional morphology parameter space which is very different from the much higher dimensional space of a color jpeg that humans see. However, these parameters are specifically known to correlate with various galaxy morphology properties in a predictable way. For example, galaxies with very high ellipticity are predominantly edge-on and would thus fall under the ‘Featured’ category, while galaxies with high concentrations would suggest they belong to the ‘Not’ category. More importantly, the efficacy of these parameters is not heavily dependent on the redshift range of the GZ2 galaxy sample [citations]. Thus, these parameters provide a less-biased feature space which suggests that the GZ2 vote fractions we computed before might not be appropriate.

We can see evidence for this by considering how the quality metrics we compute for the machine’s performance vary when comparing the output to different GZ2 ‘truth’ labels. Figure 11 shows our standard quality metrics but now considering only the subjects which the machine retires. The black lines show the machine’s performance when using the GZ2 raw vote fractions as done previously. The red lines are the performance when using the GZ2 debiased vote fractions. This has the effect of dramatically increasing the purity of the sample (fewer ‘Not’ subjects contaminate the sample) but the completeness (how many ‘Featured’ we recover) drops significantly. This might seem to be evidence to the contrary but we note that the GZ2 debiasing technique affects a full third of the galaxies in the GZ2 catalog. Since the machine is not trained on these labels, it doesn’t recover the significantly increased fraction of ‘Featured’ galaxies. Finally, we construct a mix of these two labels in the following way: Under the assumption that the machine is identifying more ‘Featured’ subjects than it seems, we look at the group of false positives, those subjects which the machine identified as ‘Featured’ but which the GZ2 raw labels claim are ‘Not’. To this subset we apply the GZ2 debiased label and the results are shown in blue. All metrics improve dramatically and are, in fact, better than either individual set of labels.

We see this more explicitly by examining these subjects’ distributions in the various morphology parameters. Figure 12 depicts the distributions for each of the five morphology parameters split in various ways. The top row shows the difference between those subjects the machine retires as ‘Featured’ and ‘Not’. There is a marked contrast between groups in Gini, Concen-

tration, and ellipticity. The middle row shows the distributions for those subjects which were determined to be false positives and false negatives according to their GZ2 raw labels. We see that the false positive distribution is nearly identical for every morphological parameter to the distribution of those correctly identified as ‘Featured’. Furthermore, if we instead plot those labeled as false positives according to their GZ2 debiased label, the sharp peak in the ellipticity distribution disappears entirely. Thus the machine truly is identifying more ‘Featured’ subjects than at first supposed. Finally, the bottom row shows the distributions of all machine-retired subjects with those which the machine did not retire, i.e., those where the machine predicted labels with $p_{\text{machine}} < 0.9$. The distribution of those not retired by the machine display properties of both ‘Featured’ and ‘Not’ as evidenced by a broader Gini and Concentration distributions. Though these two distributions seem to be more similar to the distribution of ‘Featured’ subjects, the lack of obvious peak in the ellipticity distribution could indicate that the machine needs at least 3 parameters to strongly predict a ‘Featured’ label.

Another desirable quality of the RF algorithm is its interpretability. The trees in the forest each split on values for every dimension of the feature vector where splits that reduce the prediction error are preferred. From this one can reconstruct which features are most important for determining class predictions. Unsurprisingly, our machine puts a great deal of weight on the Gini coefficient which is clearly evident in the decisive separation between subject subsets in the top two rows of Figure 12. The RF next places nearly equal weight on Concentration and ellipticity. **double check this and values**

We recognize that deep learning algorithms are becoming increasingly popular and incredibly powerful at accurate prediction. However, a core drawback to these algorithms is their black box quality. Not only is it difficult to interpret which features they identify, it has yet to be shown how or if these features directly relate to the physical properties of galaxies. Furthermore, while huertas-company, deilman (and others?) have demonstrated that these algorithms can be trained to very accurately predict just as humans would, we know that human classification is inherently biased (moreso for the general public, but even experts disagree!). By using a simple off-the-shelf random forest we may not achieve the same level of accuracy as a deep learning algorithm, however we argue this is due to the fact that this machine can be optimized to, in a sense, counteract human bias.

The motivating goal behind GZX is to provide a practical framework to increase the rate of galaxy morphology classification without sacrificing the quality of those classifications. We thus advocate that one should use

their machine of choice in place of a RF if so desired. This choice should be motivated by the science goals of the project.

7. IMPLICATIONS AND RECOMMENDATIONS

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 a simple decision engine which guides how these two agents interact, we achieve an order of magnitude increase in the classification rate. Furthermore, this is achieved while maintaining accurate identification of both ‘Featured’ galaxies those which are ‘Not’ to over 95% when considering the mixed GZ2 labels discussed in the previous section. Additionally, we recover nearly 95% of those galaxies dubbed ‘Featured’ with contamination at a modest 15%. In this section we discuss our road map for the future including immediate steps we can take to further strengthen GZX, implications and practicalities for citizen science, and implications for upcoming massive surveys such as LSST, Euclid and WFIRST.

7.1. Visions for the Near Future

GZX represents the simplest and perhaps most intuitive way to combine human and machine intelligence. That it performs as well as it does gives the authors goosebumps and motivates a higher level of sophistication. Here we discuss a handful of the most straightforward steps we will take to achieve that next level, implementation that we believe 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 entirely passive in that no decisions are made about which subjects provide that feedback. This was due to the obvious constraints of our simulations. However, with SWAP we can leverage the most skilled volunteers as a means of spot-checking that the machine is performing as expected. This could additionally provide a timely warning: if the machine consistently fails the spot-checking, a new algorithm could instead take its place.

A more sophisticated machine learning algorithm. We have demonstrated that dramatic results can be gained using a simple random forest algorithm. This machine has several desirable qualities. One can understand the machine’s performance directly by examining the distribution of output in each of the input features as we have done in section 6.2. Thus one can monitor that performance and make further decisions

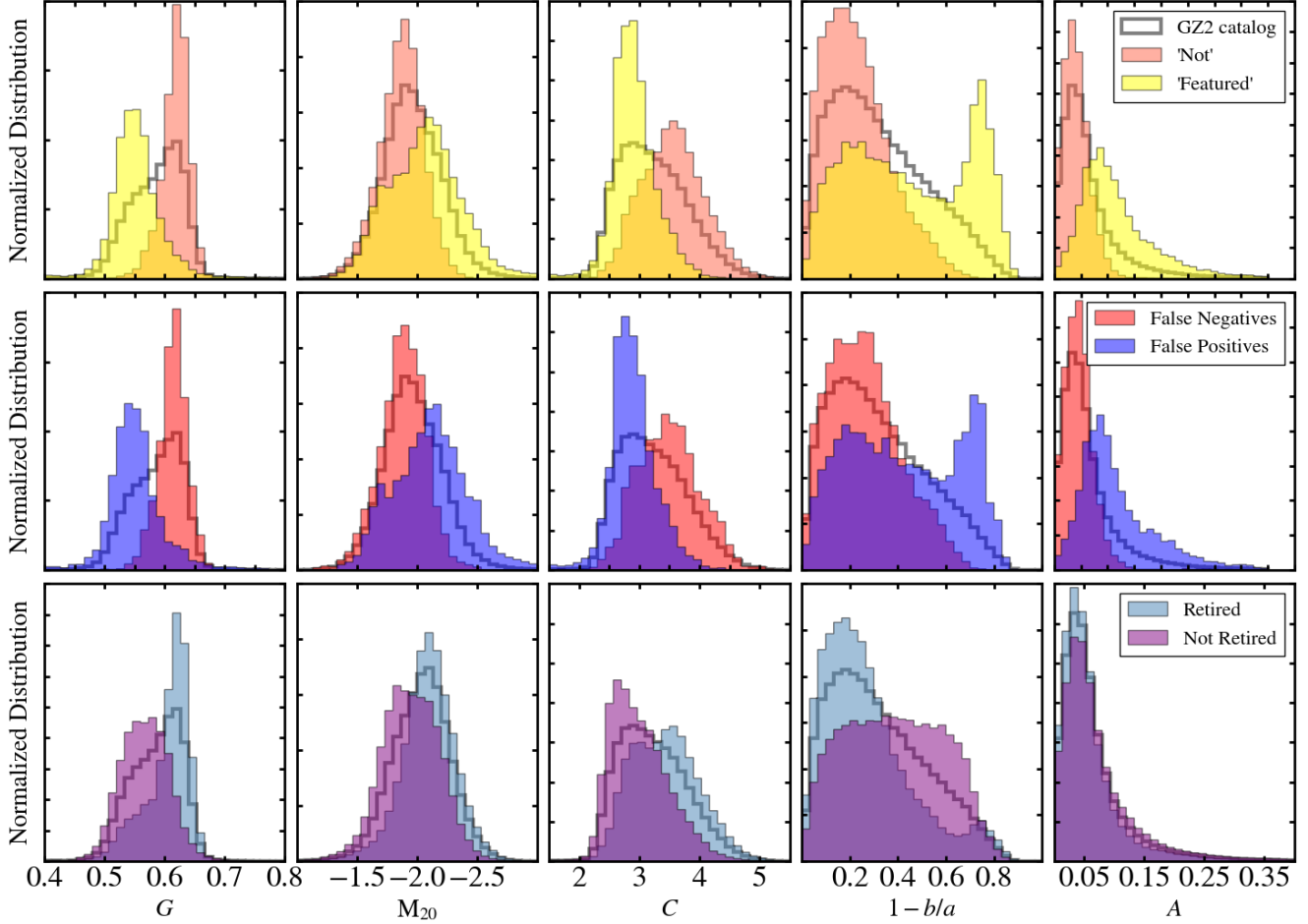


Figure 12. What the machine can and can't classify.

based on it (does the machine need additional training? should a different machine be used? etc.). Secondly, this algorithm can inherently output probabilities associated with each prediction and identify which features were most important to that prediction. Finally, this machine is very fast to train. Even with a considerable cross-validation, the RF can train on 75K subjects in under an hour with an 2.9GHz 8-core laptop.

That being said, a RF is still a very simplistic machine. It is highly sensitive to outliers in the feature vector space. It cannot accept measurement error on features. And it can only accept distinct class labels (though it can certainly handle more than 2 as we have done here!). One way to better utilize all available information would be to implement an algorithm that could accept a continuous distribution as a label since this is precisely what SWAP outputs for every subject. Furthermore, providing a probability distribution of possible output labels would also be wise as there are galaxies that exhibit both 'Featured' and 'Not' properties (S0 galaxies, anyone?). Galaxy morphology is only distinct at the far ends of the spectrum with a great many subjects wad-

ing in the middle, exhibiting a mixed set of features. Luckily, algorithms exist that have these capabilities including deep convolutional neural networks (CNN) such as Google's TensorFlow or Python's Theano, and Latent Dirichlet allocation (LDA), an algorithm that is frequently used in document processing. It is already known that CNNs can accurately predict galaxy morphology labels but application of LDA to this type of task has never been attempted, to our knowledge.

A larger feature vector for each subject. Our choice of feature vector stems from work on ZEST (Scar-lata et al. 2007) who used the same set of features in a Principal Component Analysis to classify 120K galaxies in the COSMOS field. Though G , M_{20} , C , A , e have been well studied for decades and known to correlate with physical properties of galaxies. There are additional morphological parameters that correlate with galaxy properties which a machine algorithm could find more informative than the features we provide here. In particular, the MID statistics of ? have been shown to correlate well for higher redshift galaxies and are less sensitive to noise. Our current features coupled with

the MID statistics are all non-parametric indicators and could easily be implemented as basic measurements in upcoming survey reduction pipelines. Additionally, information could be gained by the inclusion of parametric features such as Sersic index and bulge to total ratio, though these are more time consuming to measure in bulk.

We note that the morphology indicators we have discussed so far are geared towards distinguishing between early- and late-type galaxies. However, our framework is applicable to far more than just ‘Featured’ or ‘Not’. A different question would require a set of features that appropriately identify the morphology of interest. It is well known in machine-learning circles that choosing an appropriate feature vector is more of an art than a science, thus these choices should be tested on a subsample prior to implementation in GZX.

A larger confusion matrix for SWAP. Our simulations considered a binary classification task as this is how SWAP was originally built. It was created with the goal of identifying rare events to which it was only necessary to ask if the image in question contained a gravitational lens or not. We have demonstrated that SWAP can be used for pure classification and for a task which is perhaps the most common question in the local Universe. And though much can be gained from simple binary tasks, there will always be occasion for a more complex classification. Specifically, any morphology task concerned with the strength of a feature (i.e., How pronounced is the galactic bulge, if one exists?) will be difficult to force into a binary box.

The core of SWAP revolves around the confusion matrix that an agent assigns to each volunteer who classifies. Implementing SWAP3D, extending the confusion matrix to three dimensions should be relatively straightforward. Higher dimensions than this will quickly grow in complexity without much in additional gain. We note that most questions in the Galaxy Zoo decision tree currently have or 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 on the same training sample simultaneously and their predictions aggregated with SWAP. We have already implemented an agent which is assigned to our machine classifier. This agent tracks the machine’s performance over time and could, in principle, be used in exactly the same way as agents assigned to human classifiers.

There is much work in the machine learning community concerning ensemble algorithms [citations?] and, in fact, a Random Forest is such an algorithm. These types of algorithms, and RF specifically, are created to reduce variance (sensitivity to small fluctuations in the training set) without increasing bias (error induced by erroneous assumptions in the learning algorithm, i.e., missing relevant relations between features). It remains to be tested whether SWAP aggregation of a group of machines algorithms will yield similar fruit.

7.2. Visions for Citizen Science

The work presented here has provided us with unprecedented insight into the Galaxy Zoo project and, by extension, citizen science projects. Here we present a few of our thoughts and suggestions for maximizing the wisdom of the crowd.

Don’t waste volunteers’ time. A core tenant of the Citizen Science Alliance (?)/ citizen science is not wasting human effort. If a project can be accomplished by a small group of experts, by an automated task, or a combination of both, citizen scientists should not be tapped. SWAP alone presents a means to reduce the required effort from citizen scientists by a factor of 4-5. After incorporating a machine, we reduce human effort by XXX. Human eyes thus remain crucial to classification tasks as training samples don’t generate themselves, but the burden can and should be reduced.

Towards that end, the Zooniverse has plans and is currently developing several of the techniques presented here into their online platform. Though we note that our framework doesn’t necessarily require Galaxy Zoo or the Zooniverse, it is the largest citizen science web platform in the world. Once SWAP has been implemented on their back end, they are an obvious choice for handling at least the crowd-sourcing aspect of classification tasks. [Chris: feel free to adjust/add/subtract from this section as you see fit!]

Engagement = More votes = faster classification. Nearly every Zooniverse project has a surge of volunteer effort within the first few days of a project’s inception. The results we present here are due in large part to this phenomenon. If one looks at the rate of incoming volunteer classifications for the original Galaxy Zoo 2 project, it peaks on day 2-3 at over 400K! Within one week of GZ2’s inception, XXX classifications were made by XXX volunteers. After the initial excitement dies down, most projects receive a greatly reduced but steady stream of classifications. Therefore, take advantage of this initial peak in classifications by generating excitement with emphasis on volunteer engagement. This leads to a greater the surge in volunteer classifications at the outset of a project, and allows SWAP and the machine to tackle substantial data sets.

Train your volunteers. The Space Warps Zooniverse project was one of the first citizen science projects to demonstrate that humans could be trained in a similar way as machine algorithms, and that this effect could be substantial and quantified. Although we have strived to mimic this process, we obviously could not provide feedback in real time. We have, however compared the classifications of the top 5 most skilled GZ2 volunteers (as quantified by SWAP) to our expert classifications for the sample of 3500 galaxies we described in section 4.1. We find that even the most skilled volunteers are still biased towards classifying galaxies as ‘Not’. We surmise this is due, in part, to the fact that GZ2 did not provide a training program which precluded volunteers from discerning S0-Sa galaxies, a task at which experts are more sensitive.

Applications of GZX to other projects will require PIs to develop appropriate gold standard samples for their data. We suggest that this expert sample could be developed using the Zooniverse Project Builder, an open source, free to use, web based service which allows science teams to build and classify their own data. How large should the sample be? Unfortunately, this question is outside the scope of this paper. However, our sister publication (??), uses SWAP in a live project classifying XXX and discusses the

Decision Trees versus Volunteer Experience Every iteration of Galaxy Zoo, with the exception of the the original project, has adopted a decision tree yielding dozens of possible class labels from several tasks. The downside to this structure is the complications arising in both aggregation of volunteer votes and the ease of which one can use the resulting GZ catalogs. True statistical samples cannot be created because the same questions are not asked of every subject. GZ catalogs can be used to identify pure samples at the expense of completeness, where these samples are generated through a series of cuts on every task in the decision tree preceding the question of interest.

Our vision for the future of Galaxy Zoo and other projects which utilize decision trees would have an instance of SWAP for each task in the tree and, if appropriate, one or several machine classifiers. The volunteer would still see a sensible subsequent question depending on her response but SWAP would appropriate a negative response for every subsequent task down the path she *doesn’t* take. For example, if she answered ‘Featured’ to the first question, SWAP would automatically record a negative response to every subsequent question down the ‘Smooth’ path of the decision tree. This would allow every subject in the sample to receive an answer for every question in the decision tree thus providing statistical reliability, easier vote aggregation analysis, and straightforward data products.

This would require additional SWAP architecture such that a single agent assigned to a volunteer would have the additional directive to participate in several instances of SWAP at once, where that volunteer is assigned as many confusion matrices as questions in the tree. These suggestions are motivated, in part, due to observed volunteer experience. Separating the decision tree into several distinct tasks has been shown to fall short in compelling volunteers to carry on the project [Chris: citations? comments?]. It seems volunteers prefer more integrated tasks. This is important because it relates to engagement which in turn can encourage additional volunteers and classifications as discussed above.

7.3. Dealing with the Data Deluge

The work in this paper has been motivated by the need for increased speed in the face of several large surveys to come online in the next 5 years. These projects promise to deliver more raw data per night than we have accumulated in the past decade. We 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 increase the rate of subject retirement for galaxy morphology tasks. But is this enough? Some predictions, Euclid is expected to obtain measurable morphology with VIS for $\sim 10^8$ galaxies, more than has ever been classified before. [Claudia: help here?] Galaxy Zoo is currently able to classify roughly 400 images per day **verify!**. At that rate, it would take over 3000 years to classify the entire Euclid sample.

Similarly, LSST promises to deliver XXX galaxies with measurable morphology each year. [Chris help. I know nothing of LSST.]

With GZX, we conceivably reduce classification time by an order of magnitude which puts us on the timescale of dozens to hundreds of years. Still a long time to wait for suitable catalogs. However, we are within arms reach of an adequate solution. The slack will have to be made up by machines with the exploration and implementation of more sophisticated machine algorithms with more robust feature vectors as previously discussed.

The Euclid collaboration is already considering which morphology indicators will be built into their reduction pipeline for galaxy morphology studies and we suggest that LSST do likewise [Are they already?]. Not only are these suitable for galaxy studies themselves, they provide the framework and feedback by which we can judge a machine classifier’s performance.

Is the performance enough? We’ve demonstrated that GZX as implemented here can provide classifications with accuracy around 95%. But 95% of a billion still leaves us 50 million galaxies with unreliable clas-

sifications, a daunting projection. Deep learning can provide a solution at the expense of interpretability if volunteers can be trained to produce classifications that are less biased. In addition, ensembles of machine classifiers aggregated through SWAP could provide another promising avenue. A third solution is presented in our sister publication ?? whereby accuracy is dramatically increased through an entirely different human-machine combination. In that paper we explore the effects of combining the human classification with the machine's directly and find that an average of the two 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, Stephen Bamford, Boris H. for discussions which helped illucidate things and stuff. Additionally, many thanks to Hugh Dickinson and Darryl Wright for careful reading of the manuscript. This work is funded by NSF Grant XXXXX. Additional funding was awarded to MB under the University of Minnesota's Thesis Research Travel Grant as well as New College and Oxford University's support through a Balzan Junior Visiting Research Fellowship. Partial funding was also supplied through the University of Minnesota's Doctoral Dissertation Fellowship awarded to MB.

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 ...

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

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 disk galaxies, while smaller values correlate with early-type ellipticals.

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

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

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 distributed 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

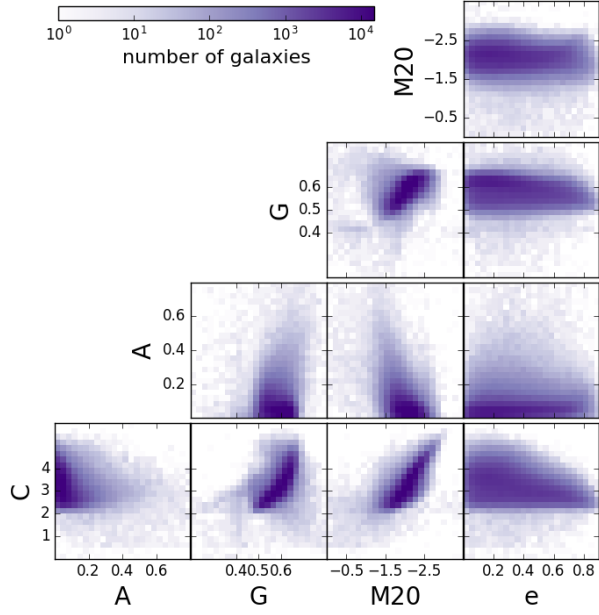


Figure A1. That's a lot of parameters!

region of the galaxy. We calculate G by first ordering the pixels by increasing flux value, and then computing

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

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.

$$M_{tot} = \sum_i^n f_i [(x_i - x_c)^2 + (y_i - y_c)^2] \quad (\text{A4})$$

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

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
- 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