Draft version October 26, 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, 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

The age of Big Data is upon us. Has been upon us. The astrophysics comnunity is already shifting focus, preparing for the way in which our science will change and the way in which we perform our science will chagne. Look at the new CasJobs – This is the type of shit we need: where analytical tools are integrated at the source of the data repository. Downloading datasets is a thing of the past. you can't do Big Data science if you have to constantly move dem data around.

Another area we need to get ready for is how we label all that shit in the sky. We absolutely love labelling things and it's damn necessary too! And the more sky we see both in terms of area and depth is going to grow huge AF. We need to find efficient, clever ways of picking out transients, radio shits, gravitational lenses, galaxy morphology, ... make a really big list with things that are rare or common or time-domain-y. LSST, Euclid, WFIRST are going to swamp us.

In this paper we consider the particular problem of galaxy morphology. This challenge is actually several combined because it necessitates 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 of Today. The imaging techniques of future surveys will change mostly in resolution and depth; things we can account for.

Another great reason to use morphology as an example is that we can draw on vast, well-established citizen science projects which have contributed to several past publications and have lead to serenditious discovery on multiple occasions. There is no doubt that to spurn this resource would be a disservice to science!!!!

So then. Morphology it is. And don't think that morphology is just a waste of time either. While there is certainly always room for improvement in our classification system including the fact that our categories were made up 100 years ago and only work for the local universe... putting galaxies into categories helps us learn about the way dem galaxies be living their lives.

The idea of combing human and machine classifications IS NOT NEW. That shit's old AF and a big topic of study in computer science circles; circles we astronomers have never been invited to but of which we should still be aware. **Citations from Chris go here!** So this idea is not novel. What IS novel is one of the first practical applications and the ability to explore the repercussions of such a system by simulating various outcomes on previously collected data.

In this paper we consider visual classifications from both citizen scientists through the use of Galaxy Zoo data as well as expert visual classifications from various published catalogs as well as visual classifications from within our own team. We will combine these with various parameters which originally sought to automatically classifiy galaxy morphology. parameters like the Gini coefficient, M20, CAS, etc. We'll wrap this all up in a neat little package by throwing it all in the supervised machine learning algorithm black box which I'll actually explain. And out will pop some sweet classifications!

With all that said, start the paper! Section blah will be the components of the method. Section blah will be detail about post-processing visual classifications. Section blah will be about the machine algorithm. Section

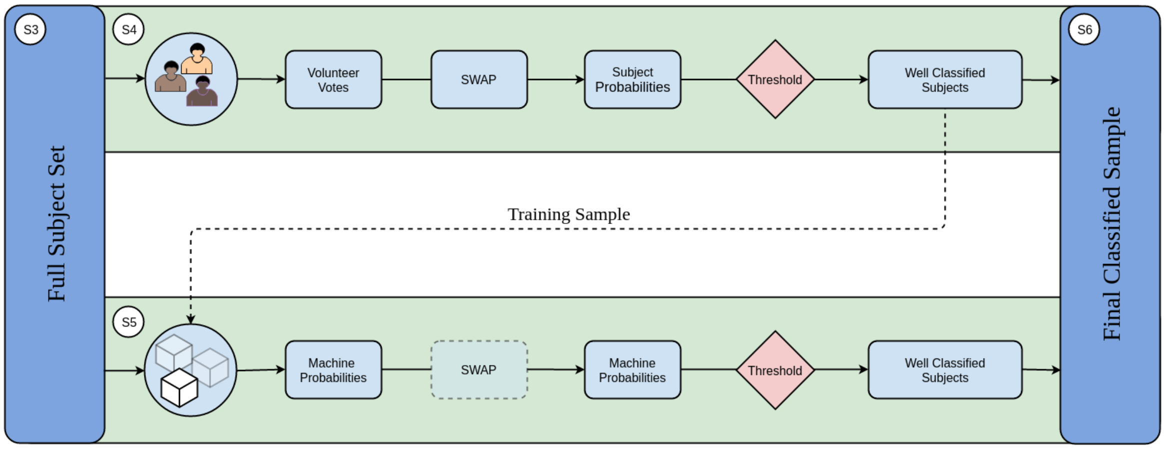

Figure 1. Schematic of our hybrid system. Human classifiers are shown images of galaxies via the Galaxy Zoo web interface. These classifications are recorded and processed according to section XXX. As a result of the processing, those subjects whose probabilties cross the classification thresholds are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

blah will be testing the method in various circumstances. Section blah will be results. Section blah will be Discussion/Conclusions.

2. GALAXY ZOO EXPRESS OVERVIEW

Any system combining human and machine classifications will have a set of generic features which we must replicate: a group of human classifiers, at least one machine classifier, and a decision engine which determines how these classifications should be combined. In Galaxy Zoo Express (GZX) we combine humans and machines in series such that the subjects humans classify provide the training sample a machine requires in order to learn. The goal of GZX is to increase the efficiency of galaxy morphology classification both in terms of the rate of classification over time, and in the amount of human effort required without sacrificing the quality of classification. We present a schematic of GZX in figure 1 and briefly outline our method here. This schematic includes significant section numbers and the savvy reader is encouraged to skip to their section of interest after reading section 3.

We first simulate human classifiers using a database of classifications (described in section 3) from the Galaxy Zoo 2 (GZ2) project which we can draw on at will. We increase the classification rate and decrease the required human effort by processing these classification using a Bayesian code first developed for the Space Warps project (SWAP) described in section 4. 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 will grow 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 galaxies retired from classification are those for which human classifiers have reached consensus. Secondly, the machine will rapidly improve and both humans and machine will be responsible for image retirement. 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 be used 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 resolved 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 catalogs (reference website?). 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 series of questions pertaining to the morphology of the subject. 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*. With the exception of the first task, subsequent tasks were dependent on volunteer responses from the previous task creating a complex decision tree.

For the analysis in this paper we select the first question 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'. 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. Additionally, we must also 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 GXZ 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. Specifically, 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'.

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) downweight 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

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

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 see 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:

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

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 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.

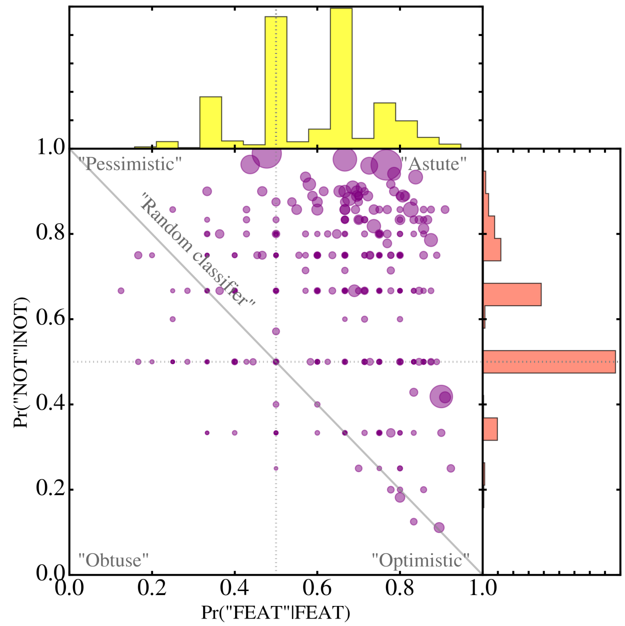

Figure 2. Volunteer probabilities for our fiducial SWAP run.

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 subjects. 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 figures 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 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 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 $\sim$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 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 $\sim$225K retired subjects. GZ2, as expected, has a distribution that peaks around $\sim$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 retire-

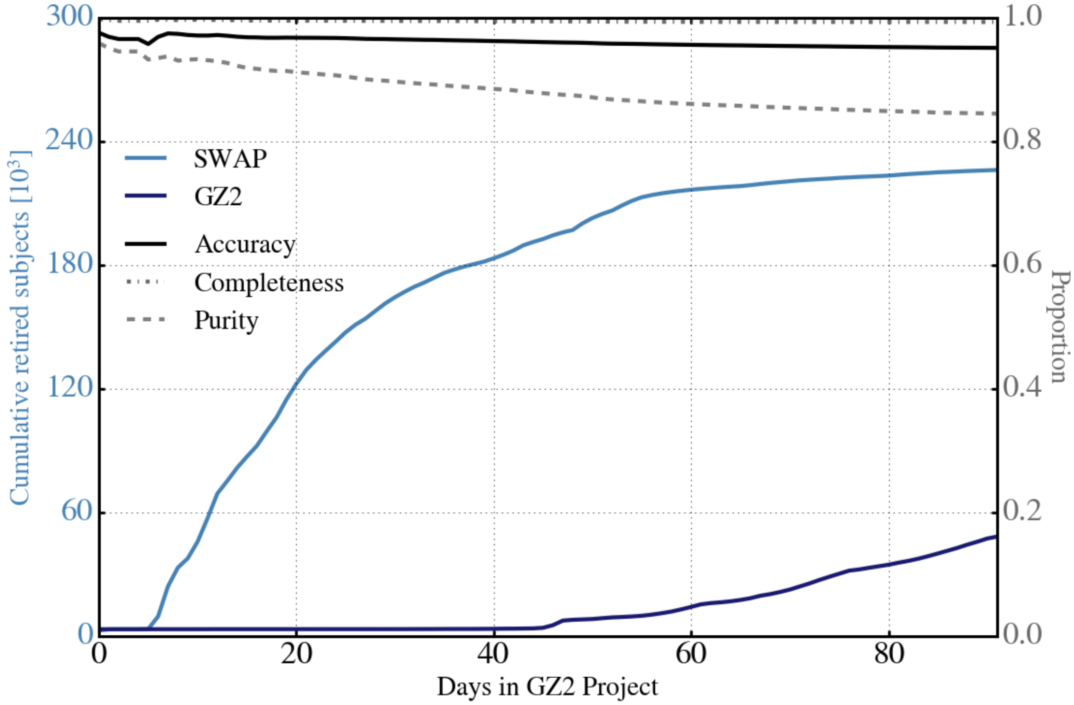

Figure 3. Fiducial SWAP simulation demonstrating the dramatic increase in the subject retirement rate as a function of GZ2 project time (in light blue) compared with the original GZ2 project (dark blue) corresponding to the left hand axis. Not only is efficiency increased by nearly an order of magnitude,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

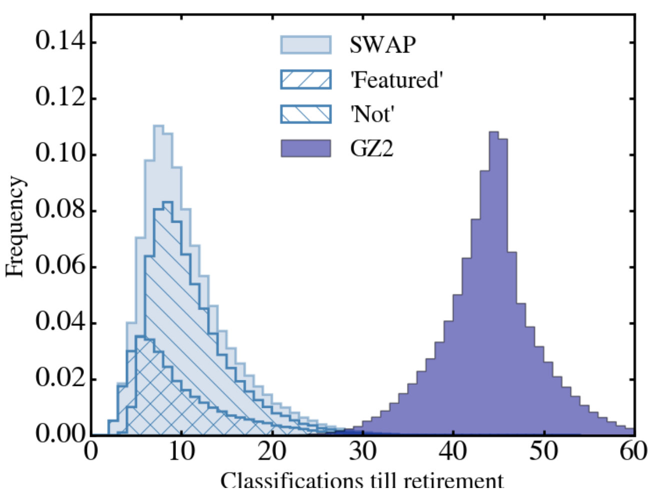

Figure 4. SWAP requires less human effort than GZ2 as evidenced by comparing the number of classifications until retirement for the ∼225K subjects retired by both SWAP and GZ2. GZ2 requires ∼45 classifications per subject before retirement. In contrast, most subjects only need approximately 10 classifications 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.

ment 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 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

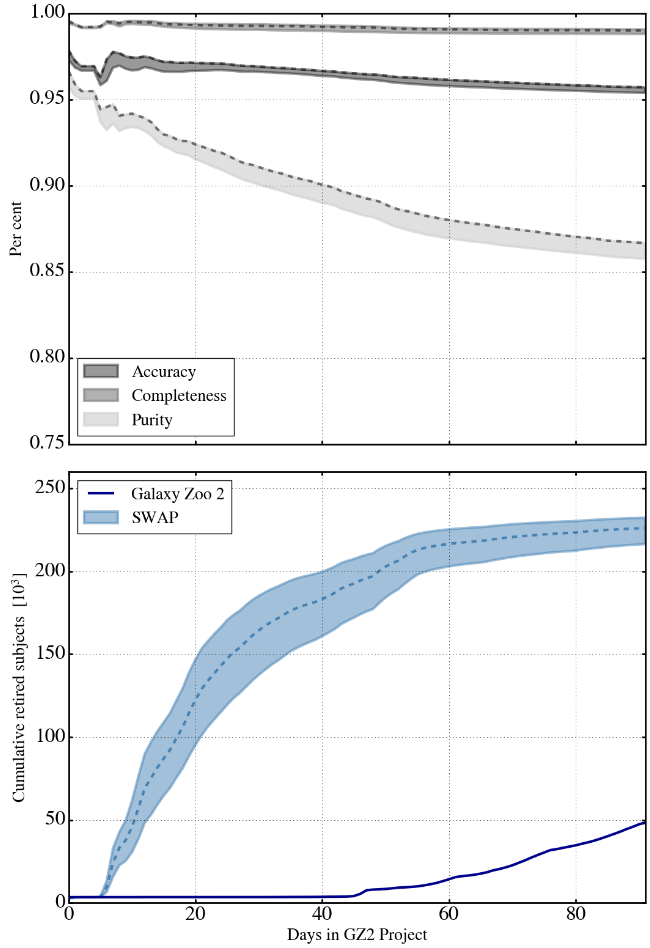

Figure 5. GZX/SWAP output as a function of GZ2 project days for a range of initial confusion matrix values.

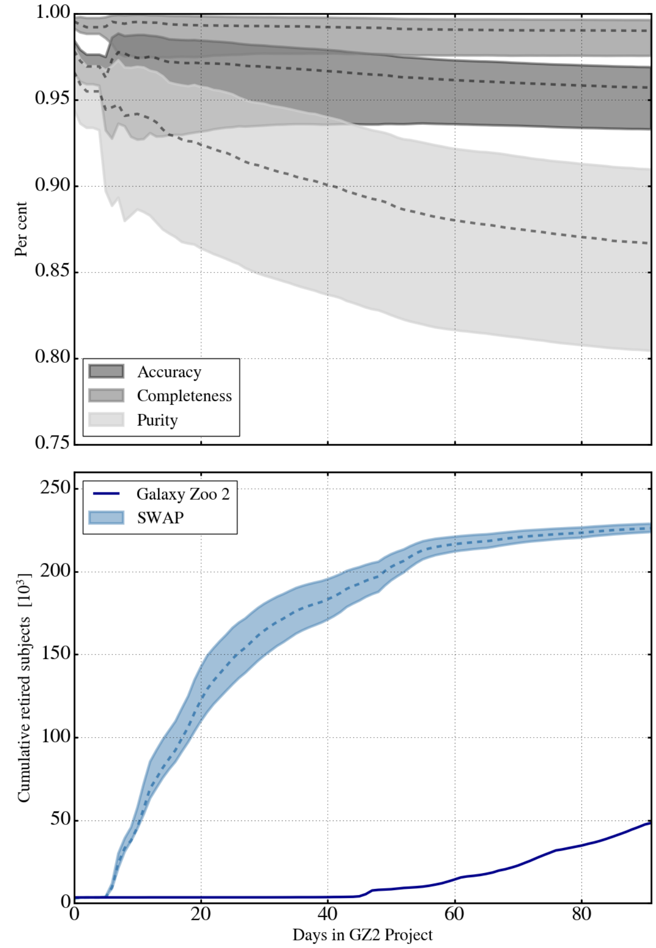

Figure 6. GZX/SWAP output as a function of GZ2 project days for a range of subject prior probabilities.

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 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 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 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 $\pm5\%$.

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.

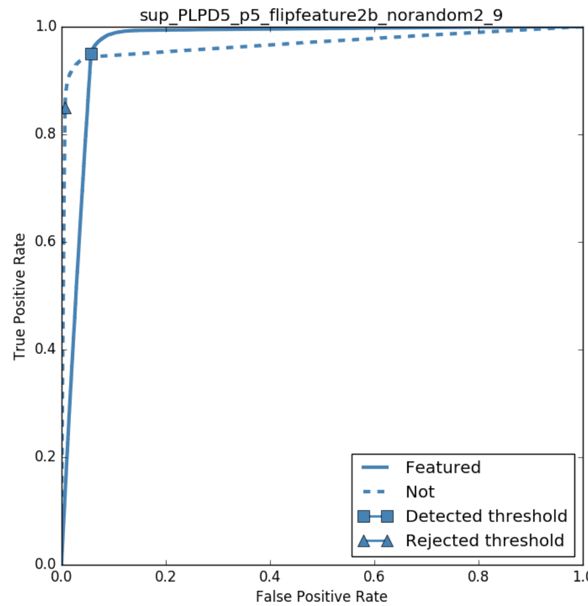

Figure 7. Retirement thresholds are adequate for our purposes.

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 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 uncer-

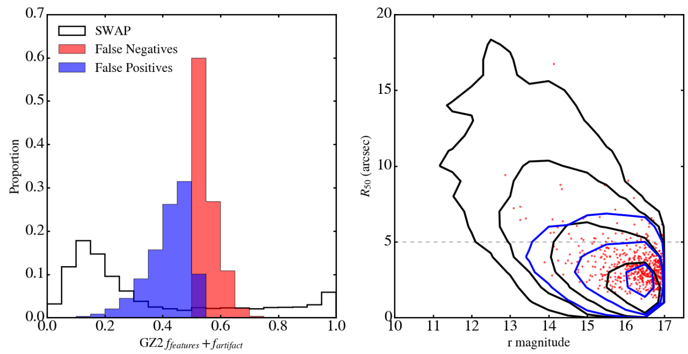

Figure 8. SWAP and GZ2 labels disagree approximately 5% of the time, depending on the choice of initial SWAP parameters. The disagreement stems, in large part, from 'true' labels derived from uncertain GZ2 vote fractions as shown in the left panel where the majority of SWAP's false negative and false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

tain. 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 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 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 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 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 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 process. 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 receive 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 has learned all it can for being the simple model that it is.

Once the machine has been fully trained, it is then applied to the test sample. In this case, the test sample is any subject which has either not reached retirement through SWAP, or is not part of the validation sample. Since the total number of subjects in GZ2 is XXX, the validation sample comprises XXX, the initial training sample is 10K, thus the first test sample contains XXX subjects. The test sample decreases as a function of project time in tandem with the increasing training sample.

5.6. *The Machine Shop / Feedback Loop*

A typical run which incorporates the machine begins with human classifications processed through SWAP for several days. During that time, the available training set builds up until it crosses the 10K threshold. At this point, the machine 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 properly 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.

At this point, another decision must be made. What constitutes a confident machine classification? Some models allow one to obtain a probability for each respective 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 which subjects the machine is most confident about (though we note it is not a true measure of confidence). Only subjects which receive a class prediction with $p_{machine} >= 0.9$ are considered retired and are removed from the system. Subjects which are not retired by the machine are subsequently fed back to human classifiers for further input during the next timestep.

This is the embodiment of our feedback loop. Those subjects on which the machine is least confident are judged by human classifiers, potentially becoming part of the training sample during the next cycle. Ideally, this increased training sample now covers an additional portion of the parameter space that the machine was unfamiliar with or unable to learn. With additional subjects spanning all of the parameter space, the machine can quickly achieve its maximum performance.

6. RESULTS

How well does the overall human/machine system perform together?

When does the machine kick in? How quickly does it learn?

Efficency of classification increased by order of magnitude.

We perform a full run incorporating both SWAP and the machine using the fiducial SWAP run. The machine attempts its first training on Day 7 of the run. The machine attempts several additional nights of training, with a larger training sample each night. By Day 13, the machine *agent* has assessed that the machine is suitably prepared to analyze the test sample. At that point, SWAP has already retired 68K subjects (this is the machine's training sample). The machine predicts classes for the remaining subjects, approximately 200K. Of those, the machine strongly predicts classes for nearly 70K subjects, thus dramatically increasing the overall sample of retired subjects for the run. As the simulation progresses, retirement by both SWAP and the machine tapers off. We end the simulation after 23 days, having sufficiently classified over 200K subjects.

The bottom panel of Fig 9 compares Galaxy Zoo Express, SWAP only, and Galaxy Zoo 2 subject retirement. By dynamically generating a training sample through a more sophisticated analysis of human classifications, we are able to retire over 200K GZ2 subjects in 23 days. With SWAP alone, we retired this many subjects in 60 days. GZ2 took 9 months to retire as many and 14 months to classify the entire catalog of 295K subjects.

The top panel of Fig 9 shows our usual quality metrics for both SWAP-only and GZX. Even incorporating such a simplistic model as a RF allows us to dramatically increase our classification efficiency without sacrificing much in terms of quality. Accuracy for the combined system remains above 90%, while purity and completeness remain above 85%. Furthermore, we note that incorporating the machine yields similar over-all quality metrics as when we allow SWAP to handle the entirety of the sample. Both yield purity 85%. Instead we seem to make a small sacrifice in the completeness of the sample when incorporating the machine: whereas SWAP alone provides nearly perfect completeness, the machine cannot keep up with that.

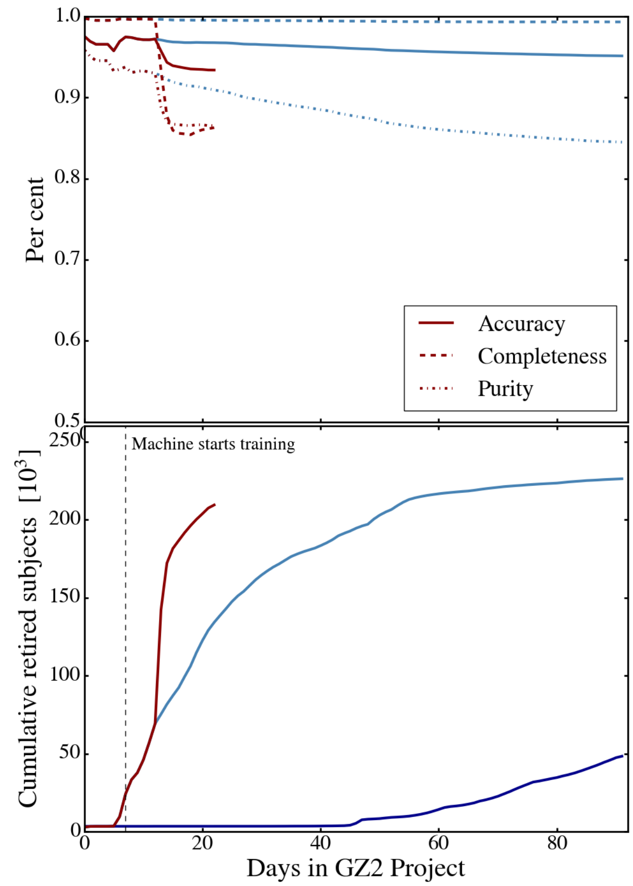

Figure 9. Incorporating the machine reduces the total time to classify over 200K subjects in the GZ2 sample to 23 days.

6.1. *Identifying the Point of No Classification*

We now turn to a discussion of what subjects can and can't be classified and why. We argue that our method provides a quick way to determine which subjects will inevitably be suitable for either visual or machine clas-

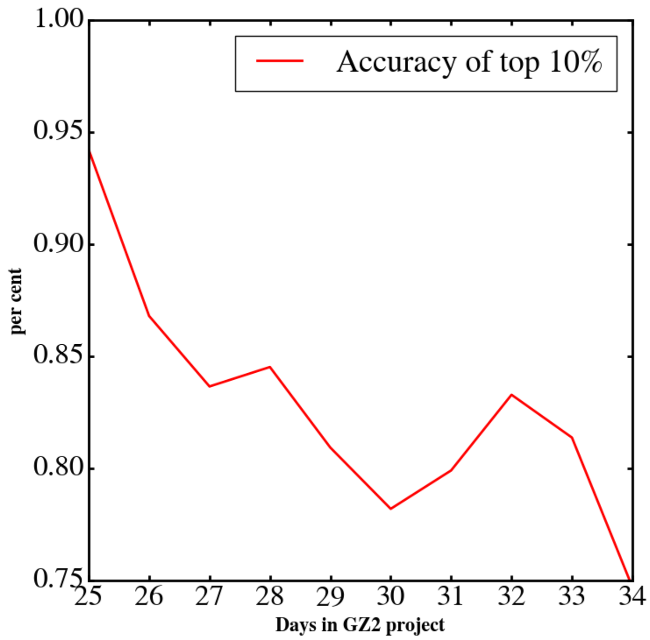

Figure 10. Incorporating the machine reduces the total time to classify over 200K subjects in the GZ2 sample to 23 days.

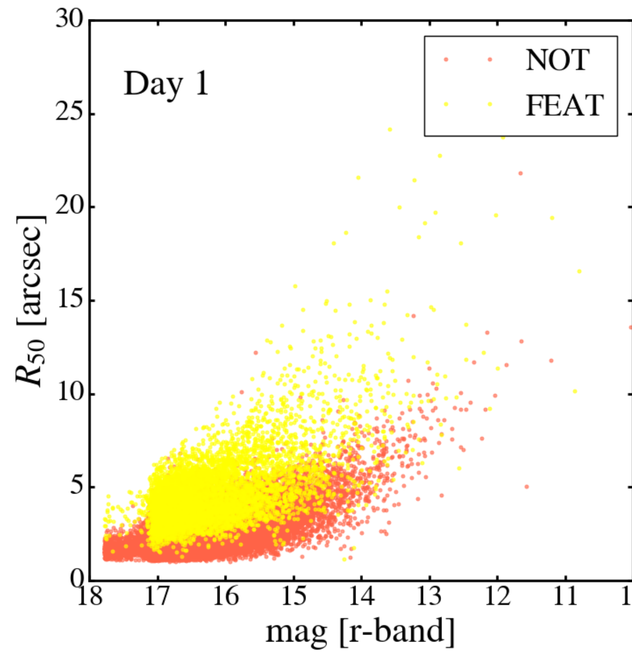

Figure 11. You can't classify what isn't there... Lack of information leads to sub-par classifications by the machine.**Make this a GIF in the online manuscipt?**

sification and which subjects simply provide too little information for robust classification. This knowledge, coupled with the fast classification rate will enable large scale surveys to efficiently and effectively tackle morphological classification tasks.

Why is the machine so bad? If we're getting completeness for the machine 70% it's not better than just assuming the the majority-class.... I need to dig into this a bit more.

What does the Machine get wrong? Why? In Fig 10 we show the accuracy of the machine classifier for the 10 days it predicts on the test sample. We see that the first application yields exceptionally good accuracy considering the machine of choice. However, that accuracy drops down to 75% by the end of the simulation. We examine in the size-magnitude plane those subjects retired by machine as shown in Fig A1. We see see the first classification by the machine includes a healthy mix of both 'Featured' and 'Not' subjects over a broad range of magnitude and sizes. These subjects are "easy to classify" as they are large and bright. However, by the end of the run the machine (combined with SWAP) have classifed the easiest subjects; those with the most information stored in their images. On the last day we peform the simulation, the machine is only classifying small, faint subjects; all the larger and brighter subjects have been retired. Small, faint subjects are notoriously difficult to classify. As such, when comparing the resulting prediction from the machine to the original label provided by GZ2, it is less likely that machine matches that prediction. Indeed, it is just as likely that the original GZ2 classifications are inherently wrong themselves. **Show a random sample of original jpg images?** This trend can even be seen in the SWAP-only output. Fig **??** shows the accuracy of the SWAP output slowly decreases over the life of the project coupled with the strong tapering of classifications are both sure signs that the easiest to classify galaxies have been retired leaving more difficult, if even classifiable, subjects in the queue.

We also examine those subjects which are not classified at all, either in the SWAP-only run or the combine run. I am sure they will all suck but need to throw up some distributions REDSHIFT, MAGNITUDE, SIZE. Indeed, the original GZ2classifications were never intended to be used directly as a catalog. The authors strongly urge cuts down the decision tree as as well as in magnitude and redshift. Even the power of human consensus cannot classify information which is not provided by an image. **The third way we increase efficiency is to effectively identify those subjects for which additional human or machine intervention is not practical and will not yield appropriate classifications.**

7. IMPLICATIONS FOR FUTURE SURVEYS / VISIONS FOR THE FUTURE

We've now identified several ways to suss out those subjects which require additional intervention. If SWAP can't classify it, then potentially these subjects should be diverted to experts. If a machine can't classify it, then those subjects can be relegated back to humans. Thus we have a cute little chain of command!

Apply all these performance metrics to the datasets expected from Euclid, LSST, etc. Estimate reduction in classification time.

7.1. *Short-term Visions*

Immediate steps we can take to make an even stronger system.

Classifications are not debiased. It is a known issue that any visual classification will be biased by several effects, especially redshift and size. Lack of depth and resolution at higher redshift tend to produce classifications which lack observable features. Galaxy Zoo 2 is able to account for redshift bias by adjusting the vote fractions produced by the volunteers. This step is done after the fact by considering for every Galaxy Zoo catalog Willett et al. (2013), **?**, **?**. Throughout this paper, we compare predicted classes to a majority raw GZ2 vote fractions. These vote fractions are not corrected for redshift bias. As states above, we chose these vote fractions because SWAP processes the original votes and does not have functionality to debias these votes. Thus, this step must still be done after subjects are retired or the project completes.

Where does the gold standard sample come from? Of utmost importance to GZX is the establishment of a gold standard sample. We utilized this sample both to provide "training" for volunteers and to provide validation for the machine. Applications of GZX to other surveys and projects will require PIs to develop appropriate gold standard samples of their own data. How should these samples be developed? How should the labels be applied? Luckily, the Zooniverse Project Builder is an open source, free to use, web based service which allows teams to build and classify their own data. How large should the sample be? Unfortunately, it is outside the scope of this paper to produce a larger gold standard sample for testing purposes. However, we have unequivocally demonstrated that with a sample of 3500 subjects, we are able to achieve adequate results. **It would be soooo cool to run SWAP again with fewer gold standard subjects; or play around with how the classifications on gold standard subjects are distributed throughout the data. :(**

We didn't actually train volunteers. The Space Warps project provided feedback to volunteers after their classification of simulated data thus providing practical training for their volunteers. We stress that, although we have strived to mimic this process, we are reprocessing data and thus could not provide feedback in real time. The extent of this effect outside the scope of this paper, however, results of an application of SWAP to a live project are explored in Wright2016, in prep **??**. **Notes: Fewer users trained. Fewer training images. Less front-loading (how far apart can the training images be staggered and still produce good results?)**

Low redshift regime – what can we do for higher z data? GZ2 is a low redshift sample of galaxies, those to which the Hubble tuning fork could be applied. At higher redshift, these class labels break down. Similarly, standard methods of quantifying morphological structure also suffer as M20 and Gini coefficient are susceptible to noise (I think, cite XXX). Additionally, as the shapes of galaxies are significantly different at higher redshift, different metrics of morphology would be more appropriate. We could easily extend our machine to perform on high redshift samples by incorporating the MID statistics **??**. Another future test will be to examine how GZX performs on Galaxy Zoo: Hubble and Galaxy Zoo: CANDELS datasets.

Naive machinery – RF is too simplistic Fig **??** shows the pseudo-learning curve for fiducial GZX simulation. We note that it is not a true learning curve because the RF model is not fixed for each night of training, though the range of hyperparameters found does not vary drastically. **My Prediction:** As in the example shown in Fig **??**, we see that the cross-validation score increases as a function of training sample size (which is a proxy for project time) while the training score decreases.

That these two curves meet and plateau is a strong indication that this particular model is simply unable to adequately reflect the data. A more sophisticated model should instead of used, however, this is beyond the scope of this paper.

In principle, any desired machine could be used including deep neural networks such as that provided by Google's TensorFlow or Python's Theano. Though the standard complications remain concerning the interpretation of CNNs, crowd-sourcing with SWAP provides a convenient and efficient way to quickly accumulate the large training samples that CNNs require.

7.2. *My Vision for Galaxy Zoo / Citizen Science*

Every iteration since the original Galaxy Zoo project has adopted a decision tree which yield dozens of individual class labels from several tasks asked of their users. This was adopted for several reasons (I am assuming): 1. Maximizing the information that could be gathered for each subject, 2. Minimizing the effort volunteers must spend on each classification. To "not waste time", most tasks in the decision tree are directly dependent on the preceding task and only ask volunteers questions that are sensible for the subject at hand, i.e., not asking whether a subject has spiral arms when the volunteer classified it as 'Smooth'. The downside to this structure is the complications in both analysis of the volunteer votes for the creation of GZ catalogs and the ability and extent to which one can use a GZ catalog. True statistical samples cannot be created because not all subjects are asked the same question. Users of the GZ catalogs can mine for purity but cannot hope to achieve completeness. Selecting samples for scientific analysis requires a slew of cuts on every task in the GZ decision tree preceding the question of interest.

To alleviate these concerns, our vision for the future of Galaxy Zoo classifications would consist of several, simple, binary questions. Each question would be tracked by a separate version of SWAP and, if appropriate, a machine (or several). The questions themselves would need to be redesigned though the spirit could remain unchanged. An answer of 'Featured' = Yes would automatically provide an answer of 'No' to every subsequent question down the 'Smooth' path of the decision tree. If a volunteer answers "Could this be an edge on disk?" with "Yes", a series of 'No's would be entered for the questions which would have followed, had the volunteer answered 'No'. In this way, volunteers will still only see questions which they believe are pertinent to the subject but the various SWAP agent(s) assigned to this user will interpret and extrapolate their answers such that ALL questions in the "tree" are answered thus providing statistical reliability, ease of classification analysis, and straightforward data products.

Additional modification of the existing SWAP software will be necessary to achieve these goals. First, SWAP3D will need to be developed in order to handle questions to which a binary option doesn't make sense, i.e. "How many spiral arms...". It should be relatively trivial to extend SWAP's confusion matrices into the third dimension. Secondly, architecture will need to be put in place to allow agents assigned to a volunteer in one task to communicate the anti-answer to agents in another task. Alternatively, the architecture could be redistributed such that a single agent is still assigned to a volunteer but participates in several analysis chains independently.

Each thread or task can also be assigned different machine learning algorithms. It is important to keep in mind that different machines will achieve various performance levels depending on what they train on. Our particular machine would not be able to answer the question "Is there a bar or not?" because G, M20, CAS, etc. are not suited for detecting bars. Providing the machine with information pertinent to the subject will be crucial. This is an area where deep learning techniques could lend additional benefits.

7.3. *Long-term Visions – Data Deluge*

How well will our system be able to tackle the challenges of LSST, Euclid, etc?

Incorporating mutliple machines

Incorporating more clever machines

Different strategies for different types of data – Darryl's paper

8. CONCLUSIONS

We outline and test a novel, new system for efficiently classifying galaxy morphologies, a task that will surly prove difficult with the onset of such surveys as LSST, Euclid, and WFIRST. Our system incorporates the native ability of the human mind to identify the abstract and novel with machine learning algorithms who provide speed and brute force. These, coupled with a more sophisticated analysis of volunteer votes through SWAP and a decision engine which delegates tasks between human and machine combine to create a classification system which is more than an order of magnitude faster than anything presented before. Efficiency is achieved through the dynamic generation in real time of a training sample by human classifiers which allows the machine to quickly understand the parameter space.

Classification efficiency is achieved in three ways:

1. sophisticated analysis of human visual classifications; reducing the amount of time humans would be required to contribute.

2. Sharing the burden of classification with a workhorse machine learning algorithm, in this case, a simple Random Forest. This quick and easy algorithm is fast to train, requiring little in the way of computation time, yet accounts for more than 50% of the overall subject retirement count.

3. Rapid identification of subjects for which classifications, visual or machine, are unrealistic.

– We demonstrate that even a simplistic and naive machine can still provide a significant boost to classification efficiency without dramatically deteriorating classification quality. Furthermore, simple machines are quick and easy to train.

– This doesn't have to be done on the Zooniverse platform. The code is (will be) publicly available and free to use or incorporate as the user sees fit. It can be used with small groups of experts or with your favorite citizen scientists. However, work has already begun by Zooniverse developers to incorporate these techniques into the backend of their existing, stable platform. They have a well-defined base of volunteers, significant funding and nearly ten years of expertise in citizen science.

– Take advantage of the initial peak of classifications characteristic of every Zooniverse project. The larger the excitement generated for the project, the greater the surge in volunteer classifications at the outset, the more quickly SWAP and machine can tackle substantial data sets.

9. ACKNOWLEDGEMENTS

MB thanks John Wallin, Stephen Bamford, Boris H. for discusions 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 ...

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

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 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.

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

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

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

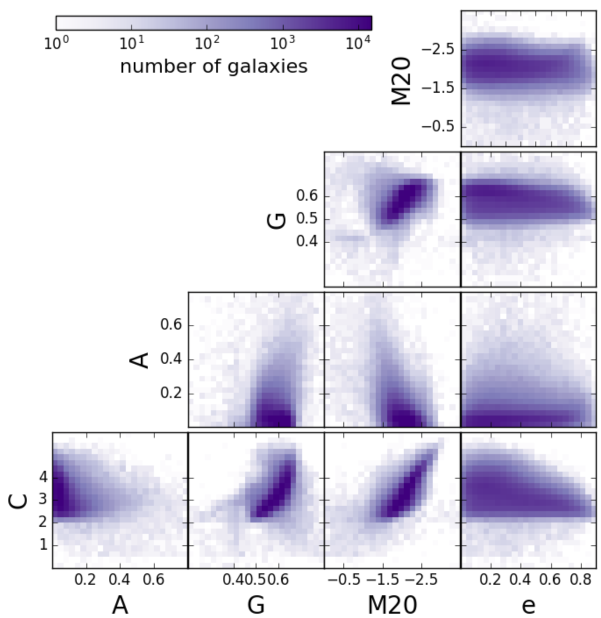

Figure A1. That's a lot of parameters!