Draft version October 3, 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, Melanie Galloway, Kyle Willett, Hugh Dickenson
Minnesota Institute for Astrophysics, University of Minnesota, Minneapolis, MN 55454

Chris Lintott
Department of Physics, University of Oxford, Oxford OX1 3RH

Phil Marshall

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

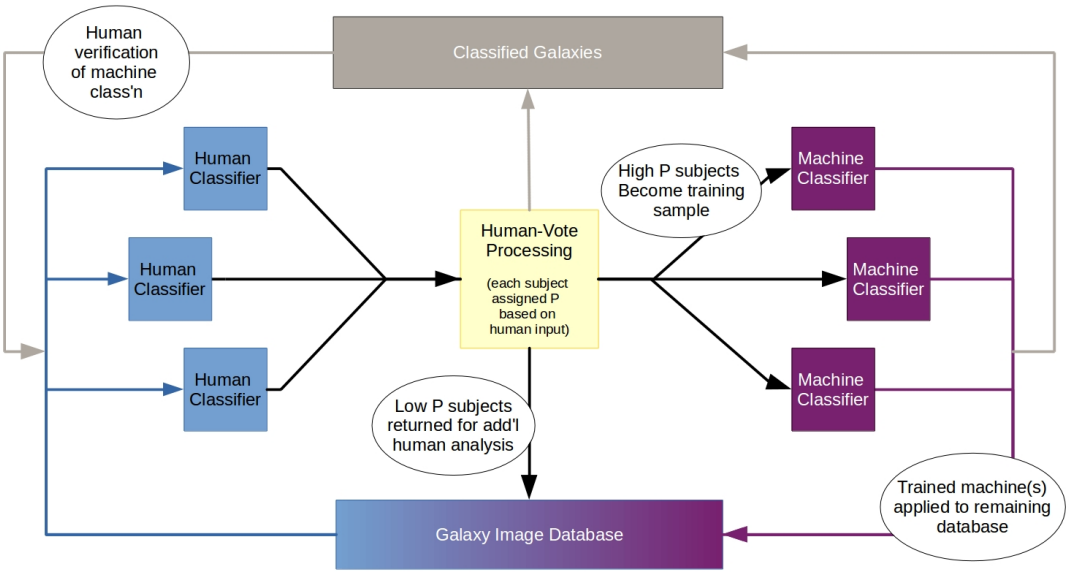

Figure 1. Schematic of our hybrid system. Human classifiers are shown images of galaxies via the Galaxy Zoo web interface. These classifications are recorded and processed according to section XXX. As a result of the processing, those subjects whose probabilties cross the classification thresholds are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

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 blah will be testing the method in various circumstances. Section blah will be results. Section blah will be Discussion/Conclusions. What sections do we want?

2. OVERVIEW OF THE METHOD?

Any system combining human and machine classifications will have a set of generic features which we must replicate.

First, a set of humans willing to classify data on request. We will simulate this using a database of classifications from the Galaxy Zoo project which we can draw on at will. These classifications are processed by a Bayesian code first developed for the Space Warps project (SWAP).

Secondly, we need a machine classifier; for this project, we have developed a random forest classifier using easily measured physical parameters such as CAS and Gini as input. See Section X for details.

Thirdly, we will need to make decisions about how the two sets of classifications are combined. After a batch of (human) classifications is processed, then the machine will be trained and its performance assessed against a validation sample. This process is repeated and the machine will grow in accuracy as the size of the training sample increases. Once the machine reaches some 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 classification will proceed in three phases. At first, the machine will not reach the acceptable level of performance and the only galaxies retired from classified are those for which human classifiers have reached consensus. Secondly, the machine will rapidly improve and both human and machine classifiers will be responsible for image retirement. Finally, improvement in the machine performance will slow, and the remaining images will need to be classified by humans. Working in this 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. In terms of GZ2, 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 the decision tree as shown in Fig **??**. In total, the data consist of over 14 million classifications from 83,943 individual volunteers.

Our first simulated run considers only the first task in the decision tree: 'Is the galaxy simply smooth and rounded, with no sign of a disk?', to which possible responses include 'smooth', 'feature or disk', and 'star or artifact'.

As we discuss in detail below, the software we use requires that every volunteer see a subset of subjects expertly identified by a member of the GZ team. We examine only those classifications made by one of the 30,894 volunteers that identified one or more of our gold standard sample. We note that these volunteers represent 36% of all users yet provided nearly 90% of the total Galaxy Zoo classification data.

4. POST-PROCESSING OF HUMAN CLASSIFICATIONS

Galaxy Zoo decision trees require a large number of independent classifications for each subject where this value is typically set at forty individual volunteer classifications. Once a project reaches completion, GZ team scientists down-weight inconsistent and unreliable volunteers while the vast majority of volunteers are treated equally with no up-weighted volunteers. While this process reduces input from malicious users and 'bots' from contributing to the consensus, it doesn't reward consistent and correct volunteers. Furthermore, waiting until project completion doesn't allow for efficient utilization of super-users, those volunteers who are exceptional at classification tasks. [Do I need to cite something here?]

Instead, GZ:EXPRESS employs software adapted from the Space Warps Zoonivere project (Marshall et al. 2016) which searched for and successfully found 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. While full details can be found in Marshall et al. (2016), we briefly outline the method 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. The confusion matrix also encodes that volunteer's probabiliy of correctly identifying the absense of feature A (denoted as N) given that the subject does not exhibit feature A. 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 identifes a training subject, i.e., one which the scientest knows the correct label *a priori* while the volunteer does not. 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. Eventually most subjects cross a classification threshold which define wthether that subject has been confirmed or rejected for exhibiting feature A and the subject is considered to be retired. The software no longer records volunteer information for these subjects.

4.1. *Volunteer Training Sample*

Finally, another key feature of the original Space Warps project was the training of individual volunteers through the use of simulated lensed galaxies. Volunteers were shown 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 depending depending on their answer. In the next section we describe how we egineered the GZ2 data to mimic the Space Warps setup as closely as possible.

We found that the SWAP software does not perform well there are no 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 classified by expert eyes into various TTypes. Thoug helpful, this particular classification isn't quite apples to apples, as Nair was not being asked the same question that GZ2 volunteers were asked. Instead, we classified this 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' labels ('Featured' or 'Not') to the 3500 galaxies.

While 3500 galaxies is a sizeable undertaking for a handful of experts, it is not a large sample compared to the GZ2 data set. Thus, 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 leaset one of these 3500 galaxies. This reduces our raw data set from 16 million clicks to 14 million; from XXX unique volunteers to 33K.

We now have a retroactively designated training sample. When considering the raw data base, however, the classifications for these particular galaxies could have timestamps anywhere within the 14 month time span during which the original project ran. As previously stated, SWAP does not perform adequately unless the bulk of the training occurs at the beginning of a project's life. We therefore adjust the order of the classification timestamps such that annotations of training sample galaxies have timestamps well before all other GZ2 galaxies. Since it is implicitly assumed that a galaxy's classifications are independent and random (galaxy images are shown randomly to volunteers), the order of the classifications should have only as small effect, if any, on the results. When running a simulation, which pulls from the database according to timestamps, the training images will be the first to be processed through SWAP.

We have done our best to mimic the Space Warps project with the goal of producing meaningful results in a similar format. What we cannot reproduce at this time, however, is actual volunteer feedback. Space Warps gently guided their volunteers towards proper classification in real time by providing pop-up comments during the project. We obviously cannot reproduce this behavior after the fact though this difference should be kept in mind. We discuss this topic further is Section XXX Future Shits.

4.2. *SWAP Requirements*

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. Before the 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 classification thresholds and the prior probability of the subject. Specifically, we must choose

- $P_{S,0}$, the initial probability that a volunteer identifies a subject as being 'Smooth', $P_0(\text{“}S\text{”}|S)$
- $P_{N,0}$, the initial probability that a volunteer identifies a subject as being 'Not Smooth', $P_0(\text{“}N\text{”}|N)$
- p_0, the prior probability of a subject to be 'Smooth'.
- t_s, the threshold defining the minimum probability for a subject to be classified 'Smooth'
- t_n, the threshold defining the maximum probability for a subject to be classified 'Not Smooth'

We perform several simulations to explore SWAP performance compared to the original GZ2 project in terms of overall accuracy achieved and, perhaps more importantly, in terms of time efficiency. Thus, to evaluate SWAP performance we consider two basic metrics: the cumulative sum of classified subjects at a given point in GZ2 project time, c_{tot}, and the accuracy of those classifications as compared to the GZ2 labels, c_{acc}. (Willett et al. 2013) advise caution when using the GZ2 catalog

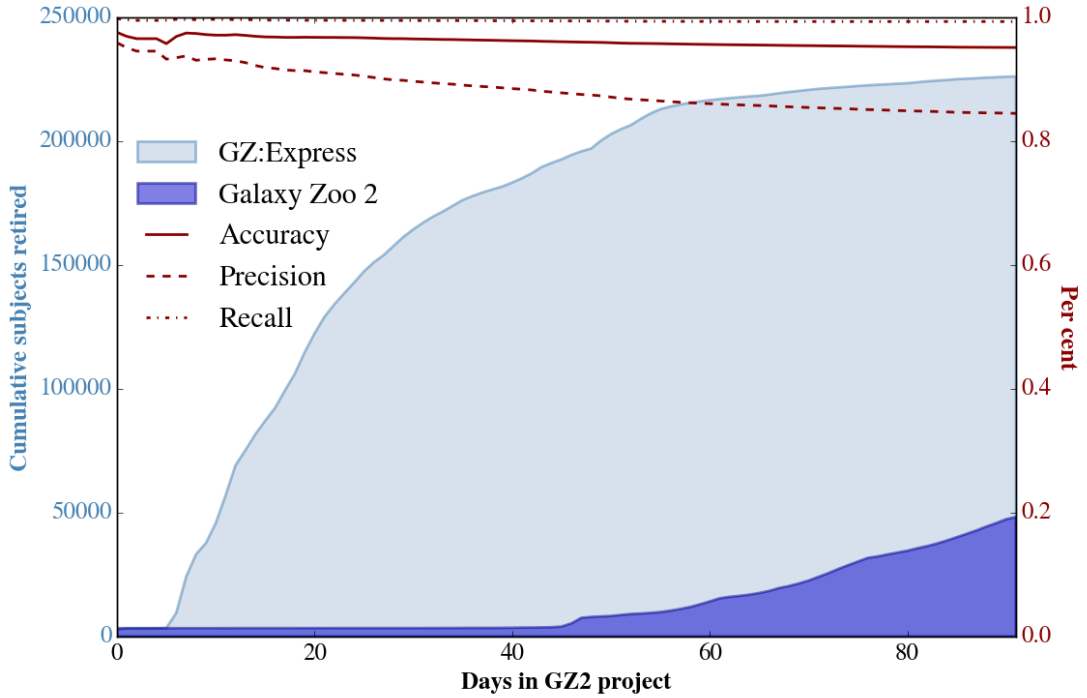

Figure 2. Simulation of top level GZ2 question reprocessed using the SWAP software only. In shaded grey are the actual number of volunteer votes. The blue shows the cumulative number of retired subjects according to the original GZ2 retirement scheme whereby a subject must achieve forty volunteer votes. The orange and yellow shading represents the cumulative number of retired subjects according to the SWAP retirement system, in that subject probabilities must cross an appropriate threshold for being labeled as having a feature or not.

to select subsamples of galaxies with a given morphology. They define various thresholds to aid the community in sample selection of clean or complete samples. In this case, we want to give a label to every object in the catalog. Every subject is given three different type of vote fractions: raw, weighted, and debiased. GZ2 debiased vote fractions are calculated to correct morphological classifications for the effects of redshift bias, a task that SWAP was not built to handle. GZ2 weighted vote fractions serve to downgrade malicious volunteers and/or bots, a task SWAP was intended to perform as well. However, because the mechanism for determining malicious volunteers is entirely different between the two schemes, we use GZ2 raw vote fractions as the closest apples to apples comparison.

Specifically, we take the majority raw vote fraction as the label for that galaxy. If the majority resided under 'star or artifact' or 'feature or disk', it was labeled as 'Featured'; otherwise it was labeled as 'Not'. We note that under this definition, only 512 subjects had a majority of 'star or artifact' and thus comprise an exceedingly small portion of the overall sample.

Figure 2 shows SWAP subject retirement as a function of GZ2 project time compared with the original GZ2 retirement scheme. GZ2 retirement was defined as a predetermined number of volunteer classifications. Galaxy Zoo projects typically require an average of 40 volunteer classifications for consensus. The blue shaded region represents the cumulative number of retired subjects as a function of GZ2 project time where we use a more lenient retirement definition: namely, if on that day of the GZ2 project, a galaxy had at least 30 classifications, it was considered retired. In yellow and orange are the cumulative number of subjects retired via the SWAP software where retirement is defined by a galaxy's probability crossing a retirement threshold. It is immediately obvious that by clever and adaptive processing of volunteer classifications speed and efficiency of subject retirement can be dramatically increased.

In figure **??** we evaluate the SWAP software by considering its accuracy, recall and precision (red, blue and green respectively) as compared to the GZ2 labels defined by raw vote fractions. This being a Smooth or Not run instead of a Featured or Not run, I'm not going to talk about the overall shape because it's going to change. BOO. These curves are, in part, a function of the parameters listed above and we now turn to a discussion of how these figures change when one or more of the SWAP parameters is adjusted.

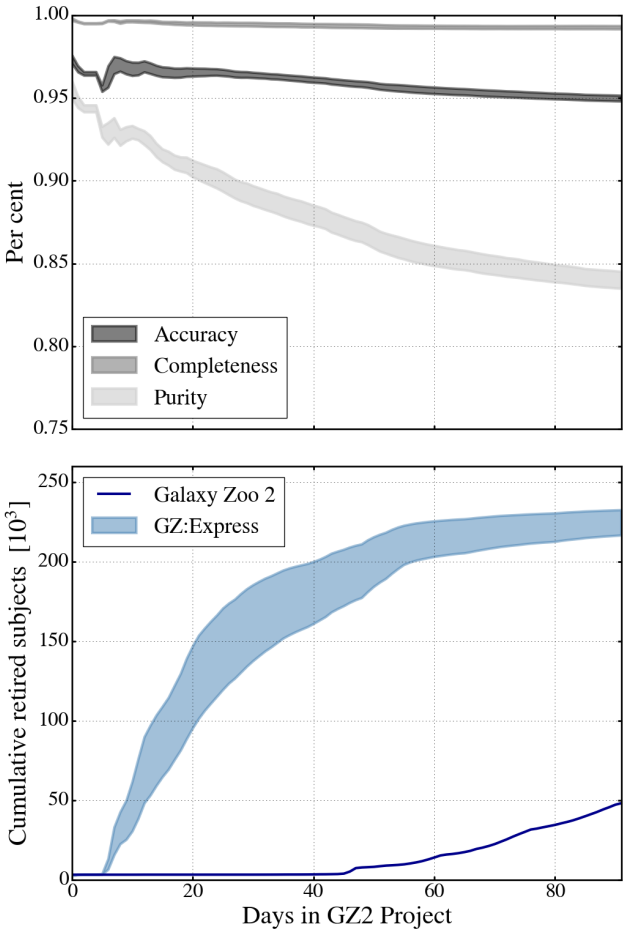

Figure 3. GZX/SWAP output as a function of GZ2 project days for a range of initial confusion matrix values.

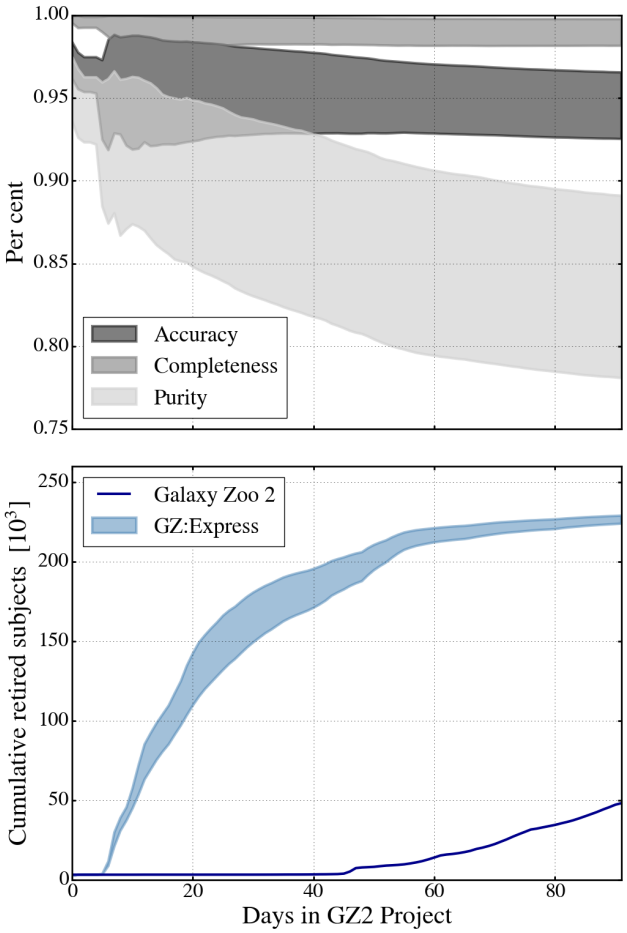

Figure 4. GZX/SWAP output as a function of GZ2 project days for a range of subject prior probabilities.

4.3. *SWAP Simulation Outcomes*

Initial agent confusion matrix. Space Warps agents assigned each volunteer $P(\text{“}A\text{”}|A), P(\text{“}N\text{”}|N) = (0.5, 0.5)$, assuming that humans started out no better than random classifiers. We explored a range of initial confusion matrix probabilities. We find that we are largely insensitive to the initial agent confusion matrix values. The majority of GZ2 volunteers achieve confusion matrices designating them as astute classifiers, regardless of their initial assigned values. The small variations observed in the SWAP output can be visualized in Figure 3. The bottom figure depicts the cumulative number of retired subjects as a function of the number of GZ2 project days where the light blue range shows the spread due to the initial $P_{S,0}$ and $P_{N,0}$ ranging from 0.4 to 0.6. Regardless of the initial input values, we achieve a total classification of $\sim$ 225K $\pm$3.5% subjects. The top figure explores various evaluation metrics as a function of the number of GZ2 project days including the overall classification accuracy, completeness, and purity of the classification. The spread is within a couple percentage points for any metric. Overall we maintain accuracy around 95%, as well as completeness of 99% while maintaining purity around 84%.

Subject prior probability, p_0. The prior probability assigned to each subject is determined by 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.) within a dataset. In our case, we desire to simply find galaxies with any feature at all, however, this is dependent on mass and redshift, among other characteristics. The original GZ2 sample was selected primarily on magnitude, then 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. As such, designating a single prior is not clear-cut. We thus explore how various p_0 affect the SWAP outcome.

The bottom panel of figure 4 shows that the overall number of retired subjects changes very little regardless

of p_0. In fact, p_0 has a smaller effect than $P_{S,0}$ and $P_{N,0}$ retiring $\sim$ 226K $\pm$1% of subjects over the course of the simulated project time. Instead, p_0 can change what label a subject is given which, in turn, has an effect on evaluation metrics. More subjects are classified as 'Not' when p_0 is small since their prior is already closer to the rejection threshold; these same subjects are classifed as 'Featured' when p_0 is larger. This explains the larger spread of the evaluation metrics in the top panel of figure 4. Overall, the average values of accuracy, completeness, and purity remain similar over the course of the project compared to figure 3. Instead, the spread can be as large as 5%, in the case of purity. If the desired science is highly contingent on a pure sample, it falls to the user to choose a prior wisely in order to maximize this output.

How should one go about this task? We find that the prior is, in essence, a reflection of the dataset at hand. The raw GZ2 classification data are skewed towards labeling subjects as 'Not', rather than 'Featured'. Overall, only 35% of volunteer votes for any subject are **not** 'Smooth'. Knowing this *a priori*, we can set the value of $p_0 = 0.35$. Everything else remaining constant, we find that this does, indeed, maximize SWAP output.

Thus one could develop a prescription for determining the best p_0 for a given dataset by initializing batches of subjects with various (though intelligently chosen) p_0 to discover, in real time, which value precisely maximizes output and performance.

Retirement thresholds. The Space Warps project set their retirement thresholds equidistant in logspace. Their prior was significantly small to begin with as lenses are expected to be very rare. However, 'Smooth' galaxies are almost as equally likely to exist as 'Featured' galaxies at low redshift (for appropriate choices of mass or luminosity). Thus care must be taken when setting t_s and t_n for retirement. These parameters are directly responsible for the label assigned to a given subject. When these thresholds are low, subjects more easiliy attain the appropriate probability to cross that threshold. This can increase speed of classification, however it also greatly affects accuracy. If the one volunteer disagreed with the previous volunteer on the nature of the subject, her vote could reverse the subject's probability away from the threshold, thus prolonging it's classification phase.

Regardless of the parameters with which one begins (within reason) the number of retired subjects grows rapidly. GZ2 requires nearly 50 days of volunteer input before a significant number of subjects can be retired (10K) whereas processing that input through SWAP retires 175K - 225K subjects, depending on the initial conditions. Perhaps of greater merit is the drastically reduced workload on the volunteers. Retiring subjects through traditional means requires nearly 2.5×10^6 volunteer votes to retire 60K subjects while SWAP requires only 4×10^5, **an order of magnitude reduction in human effort!** We now turn towards our implementation of a supervised machine to further classification efficiency.

5. MACHINE CLASSIFIER

Supervised learning is the machine learning task of inference from labeled training data. The training data consist of a set of training examples, including an input (feature) vector and a desired output (or 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?**] Fortunately, by handling individual features and processing human classifications through SWAP, we arrive with a discrete, binary task for a machine to tackle. However, in the future we will explore more sophisticated algorithms which are optimized to handle a input continuum since the actual output of SWAP is a smoothly ranging probability for each subject.

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, M20 and 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: 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 so as to all be on the same scale or between 0 and 1. 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. As we showed in the previous section, SWAP retires subjects far more rapidly than GZ2 by adaptively tracking volunteer skill and subject probabilities. This provides us with a way of quickly generating considerably large subject samples with accurate labels provided by human classifications. These retired subjects are the basis of the machine's training sample. That training sample is dynamically generated as a function of project time. Specifically, within 10 days, SWAP has accumulated 20K subjects which can form the basis for machine training.

As discussed above, in addition to a training sample we also need 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 included in the training sample? What should be the minimum number of training subjects before a first attempt? When should the trained machine be applied to our test sample, i.e. subjects in the dataset which have not yet been seen?

Which subjects should provide the training sample? As mentioned above, SWAP yields a probability that a subject exhibits the feature of choice. With a different choice of machine, one could incorporate every subject with the whole spectrum of probabilities as input labels. However, an RF requires a distinct label and thus we use only those subjects which have crossed either the detection or rejection thresholds and are thus most likely to be 'Featured' or 'Not'. However, subjects do not cross these thresholds with equal likelihood. At any given stage in the simulation, the balance of 'Featured' to 'Not' is not guaranteed to be even. This would then provide an unbalanced training sample which could hamper the learning process. However, as a first test, we allow the machine to learn on all high probability subjects.

What should be the minimum number of training subjects before a first attempt at training? On the first few days in the simulation, SWAP retires a few hundred subjects. Though one can, in principle, train a machine with such a small sample, the resulting predictions on the test sample will be exceptionally poor. Furthermore, the machine won't know that is performing poorly. For example, if a training sample of 100 'Featured' subjects is provided, the machine will 'learn' and 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. 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 should the trained machine be applied to the test sample? Again, there is no magic size that a training sample should be before attempting classifica-

tion of the test sample. We instead attempt to assess the machine's learning before applying it to the test sample. We do this by considering the machine's learning curve, which illustrates a model's performance with increasing sample size for fixed model complexity. An example is shown in **Fig XXX** for an 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, the accuracy of the method will plateau and larger training sets yield little 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 help guide our decision making process. We technically cannot reproduce a true learning curve because the cross-validation procedure can, in principle and in practice, yield a different set of machine hyperparameters that are most appropriate for the training sample delivered to it that "night". Instead, we look for the characteristic cross-validation score plateau. At each timestep, a record of the machine's training is kept including its performance on the current training sample, cross-validation score, and the best hyperparameters. When the machine's cross-validations score remains within a 1% range on three concurrent runs, we deem the machine is as learned as it will be for the simple model we have supplied. At this point the trained machine is then applied to the test sample.

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 the SWAP processing, 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 and separately.

what are sensible criteria for using the machine? If we change these criteria, how does performance change?

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.

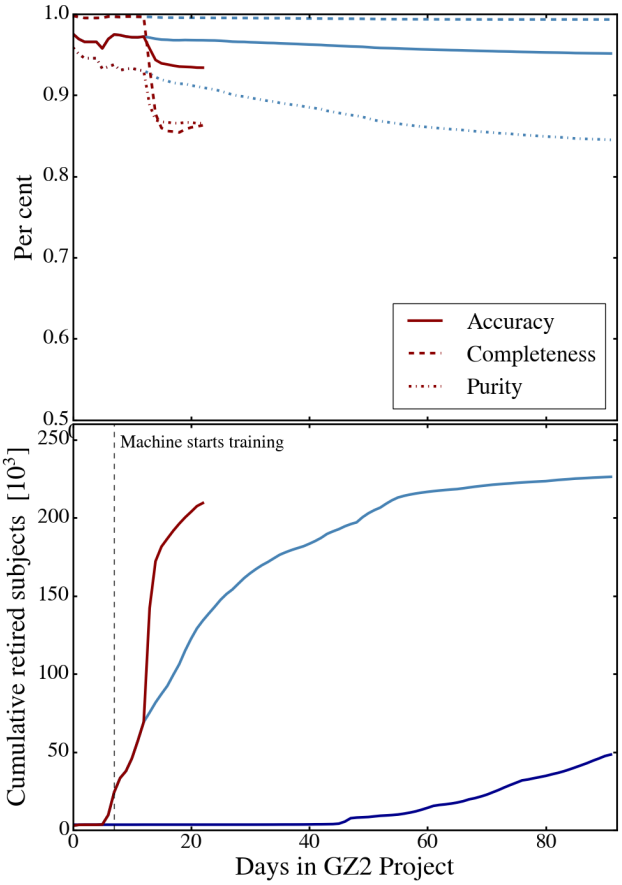

Figure 5. Incorporating the machine reduces the total time to classify over 200K subjects in the GZ2 sample to 23 days.

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 5 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 5 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

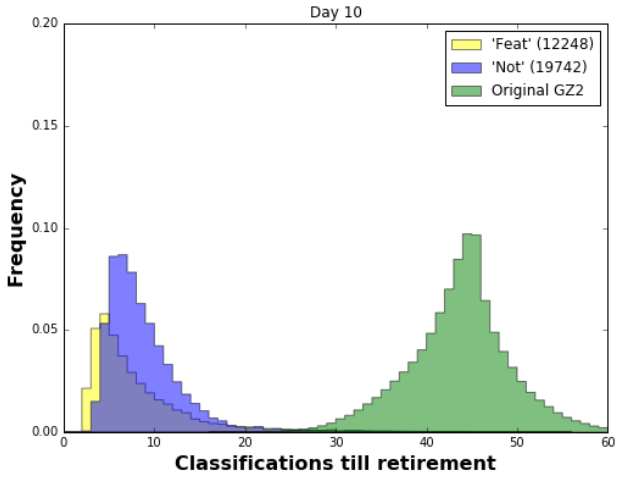

Figure 6. SWAP vote distributions are totally sick. **Make this a gif in the online journal version! Cuz it's super cool and stuff.**

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. **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.**

7. DISCUSSION

7.1. *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 classification 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.

What does SWAP get wrong? Why? Galaxy Zoo's strength comes, in part, from the consensus of dozens of volunteers voting on each subject. Processing votes with SWAP effectively reduces the consensus to its bare mininum. Though we typically are able to recover the original GZ2 class label, that is not alway the case. Of those subjects that fail, a small portion of them are due to unlikely ordering of volunteer votes. We found %?? of subjects in which the first N volunteers classified

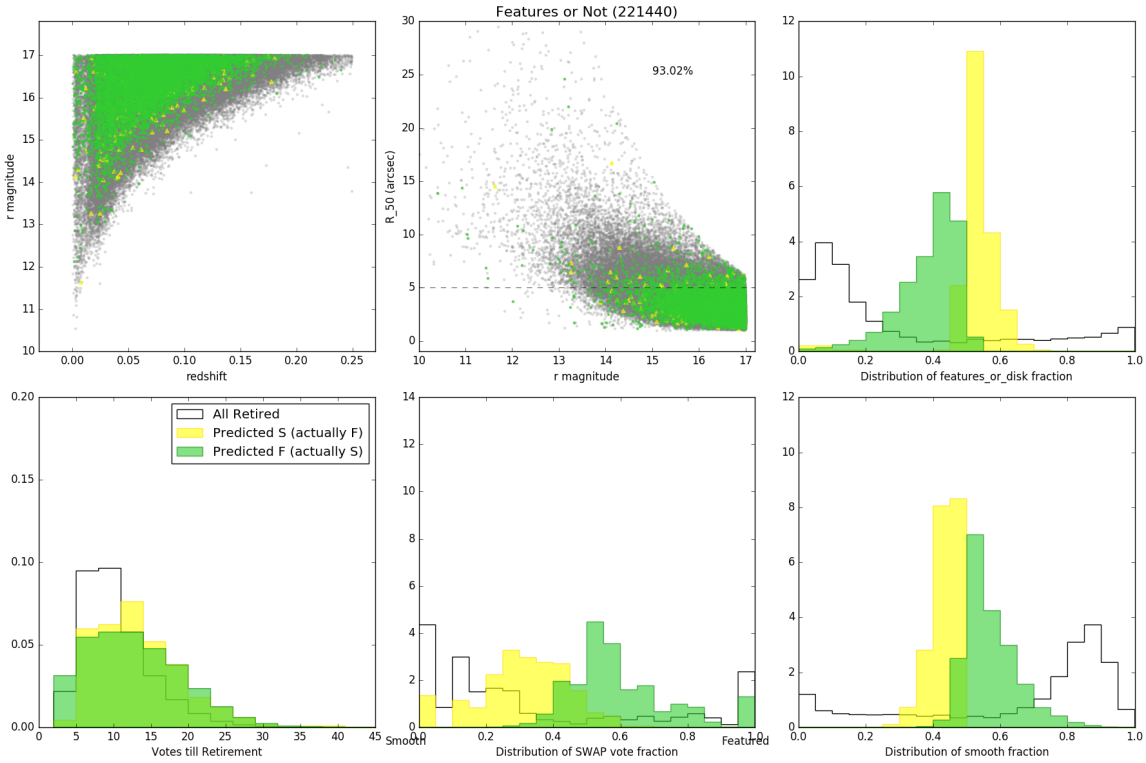

Figure 7. Why and when does SWAP fail?

them as X, yet the next 2*N classified them as Y. Thus, the majority class label given to the subject in GZ2 will be Y, but SWAP will yield a class label of X because it processes only a fraction of the votes that GZ2 does. This is one of the ways SWAP increases efficiency but at a small price. We can see this in Fig 6

To prevent this, one could a.) require each user classify a certain number of subjects to smooth out their burn-in phase thus yielding a better-trained visual classifier, b.) require each subject reach a bare minimum of classifications before allowing it's probability to reach a threshold. The difference between waiting out a user'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 assessing their data would hamstring the effectiveness of the citizen science ethos.

Another way that SWAP fails to match with the GZ2 label is through the classification of subjects which are "on the edge". These subjects receive between 40-60% vote fractions as assessed by GZ2. When GZ2 is unsure about a subject, SWAP isn't sure either and the SWAP class label will instead depend on the skill and confusion matrix of the volunteers who contribute to that subject. For example, a subject which GZ2 labels as 'Smooth' with a smooth vote fraction of 0.6 can end up classified by SWAP as being 'Featured'. This is most likely due to a.) the original GZ2 label is inherently uncertain and b.) the particular volunteers who classified this subjects have exceptionally high confusion matrices allowing them to more forcefully 'nudge' subjects over a threshold.

What does the Machine get wrong? Why? In Fig 8 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 9. 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.

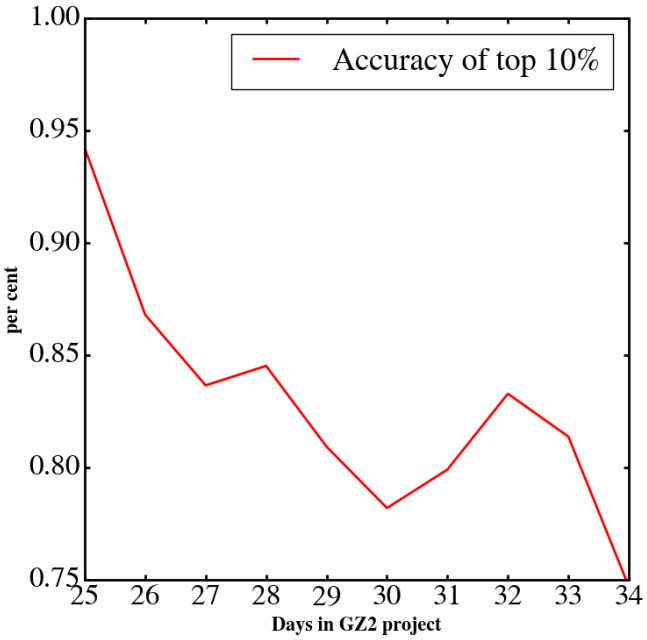

Figure 8. Incorporating the machine reduces the total time to classify over 200K subjects in the GZ2 sample to 23 days.

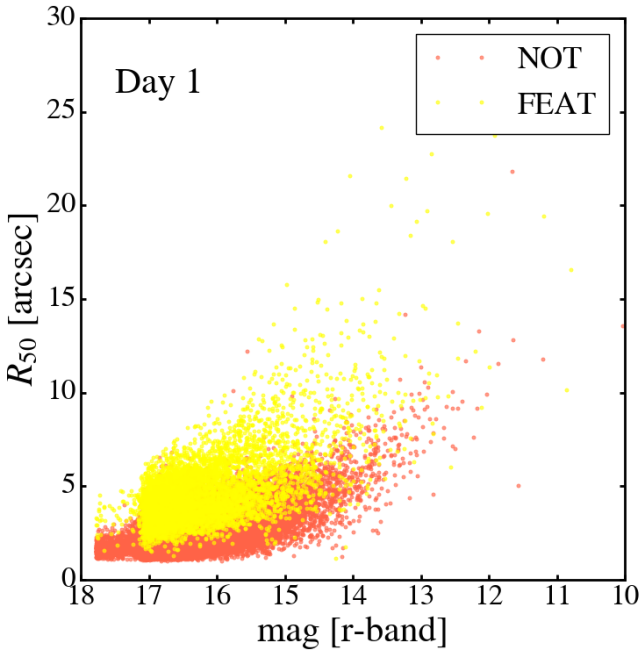

Figure 9. You can't classify what isn't there... Lack of information leads to sub-par classifications by the machine.**Make this a GIF in the online manuscipt?**

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.**

7.2. *Limitations of GZX*

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.

7.3. *My Vision for GZExpress/Galaxy Classifications*

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.

8. IMPLICATIONS FOR FUTURE SURVEYS

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.

9. CONCLUSIONS

recap all the things.

– 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.

10. ACKNOWLEDGEMENTS

NSF Grant

Thesis Research Travel Grant / University of Minnesota

Balzan Fellowship / Oxford Uni

APPENDIX

A. MEASURING MORPHOLOGICAL PARAMETERS ON SDSS CUTOUTS

So we did a LOT of work to measure all that shit.

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