HUBBLE’S TUNING FORK: A MACHINE LEARNING APPROACH

Brandon Bergerud1 and Ossian Mogensen2

1

Department of Physics and Astronomy, University of Iowa, Iowa City, IA 52242

2

Department of Computer Science, University of Iowa, Iowa City, IA 52242

# ABSTRACT

With the introduction of powerful telescopes such as the Hubble Space Telescope, vast quantities of high-fidelity imagery of remote galaxies have become available. Manual analysis of these images by experts has become infeasible, spawning citizen science projects such as Galaxy Zoo. However, the next generation of telescopes are expected to generate enormous volumes of data, going far beyond the capacity even of crowdsourced volunteers. In this study, we extend the work done on automatic galaxy image classification in the Galaxy Zoo 2 Kaggle challenge by training a convolutional neural network based on the Google Inception architecture on the Galaxy Zoo 2 dataset using transfer learning with the popular ImageNet dataset. The model was evaluated on the Galaxy Zoo 2 dataset and on expert-annotated datasets from the Sloan Digital Sky Survey**.** The results show that machine learning systems trained on the citizen-science annotated Galaxy Zoo dataset can successfully be used to classify other galaxy imagery datasets if a modestly sized manually labeled training set from the new data source can be acquired.

# INTRODUCTION

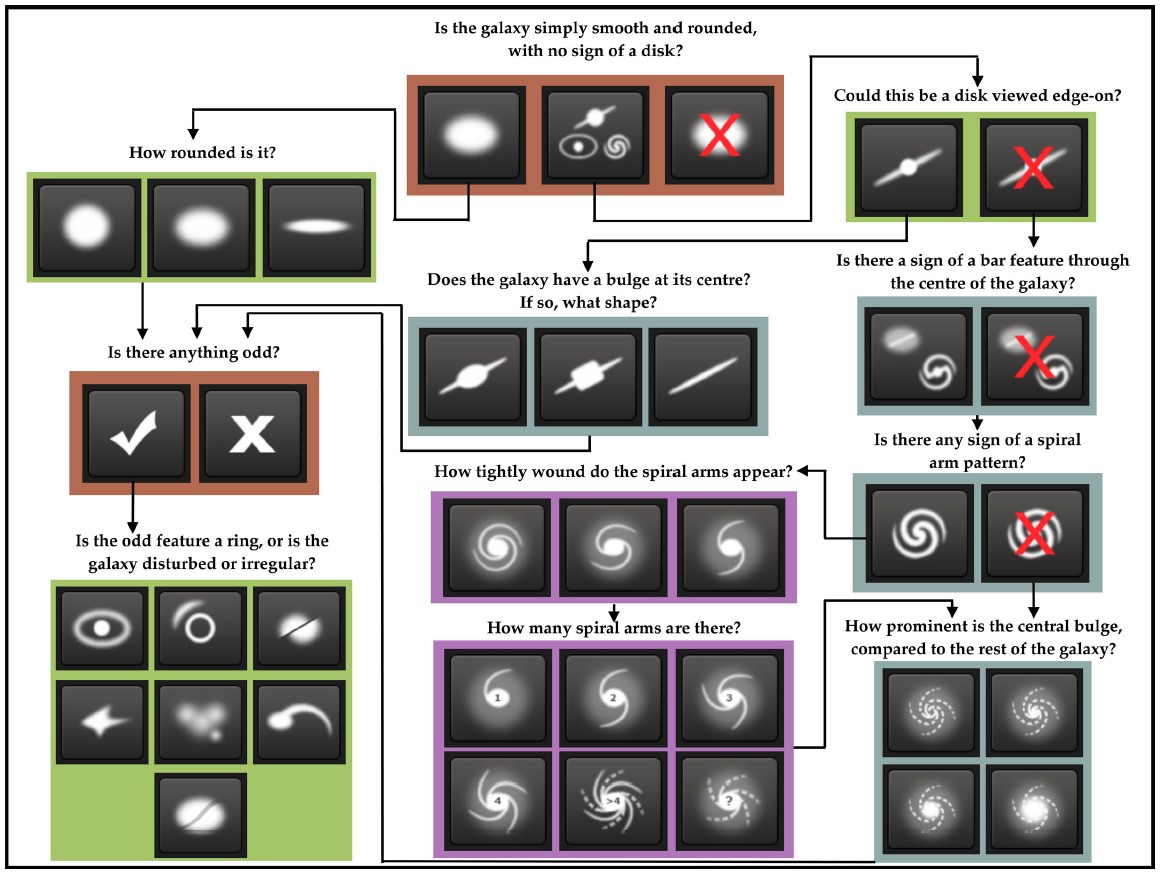
The size and scope of astronomy datasets have increased dramatically in recent years. The introduction of telescopes such as the Hubble Space Telescope (HST) and projects like the Sloan Digital Sky Survey (SDSS) have given astronomers access to imagery of millions of celestial objects. Traditional methods of data analysis, manually inspecting and classifying celestial objects, have become untenable in the face of this embarrassment of riches of data.

Astronomers have successfully turned to citizen science projects such as Galaxy Zoo to leverage vast numbers of volunteers to help classify objects. The human visual system can, with little effort or training, provide image recognition capabilities that match or exceed the state of the art in computer image recognition.

With the dawn of a new generation of telescopes, astronomy is threatened to be deluged in a sea of data. The GAIA spacecraft will produce a 3D map of over 1 billion astronomical objects (Gaia Collaboration *et al.* 2016). The Thirty Meter Telescope (TMT) (Skidmore *et al.* 2015) and the 40-meter European Extremely Large Telecope (E-ELT) will view the visible universe at unprecedented depth. The Large Synoptic Survey Telescope (LSST) is estimated to generate 15 TB of data each night as it surveys the entire sky (Ivezic *et al.* 2009). Even these vast

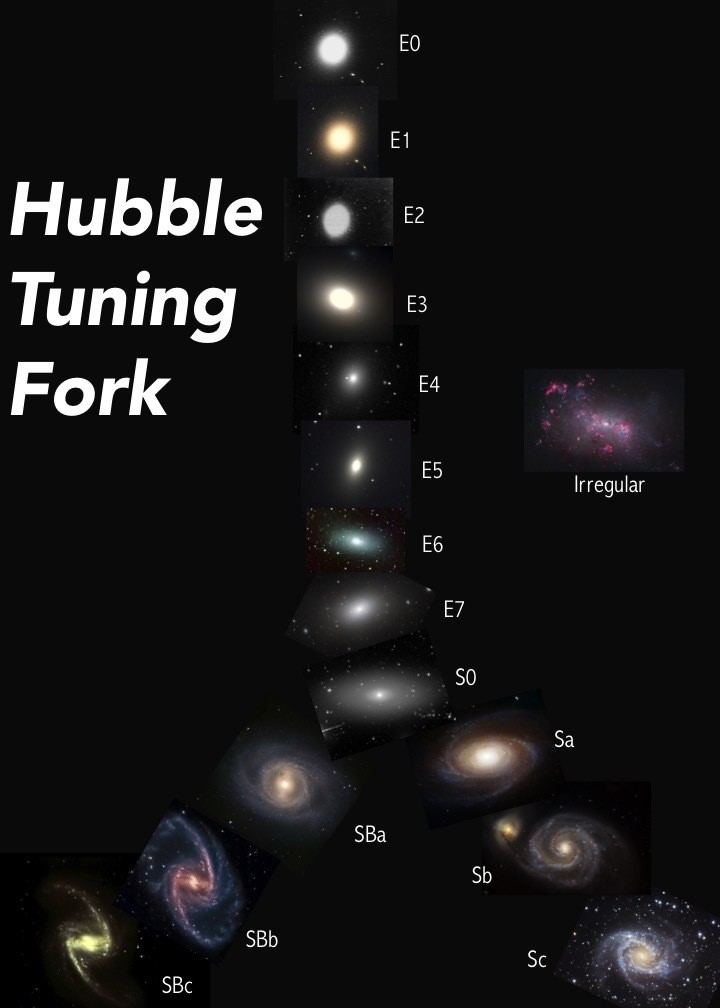
sums of data pale in comparison to the output expected from the monsuvian Square Kilometer Array (SKA). Such enormous amounts of data are beyond the ability of crowdsourcing to handle: they can only be handled by leveraging supercomputers, sophisticated algorithms, and machine learning.

The Galaxy Zoo Kaggle challenge was a competition in 2013 to produce a machine learning model that could replicate the classifications of citizen science volunteers on a dataset of 70000 galaxy images captured by HST. The top models using Convolutional Neural Networks performed well in this challenge, but the challenge did not investigate how well models trained on the Galaxy Zoo dataset would generalize to other datasets. To generalize a model trained on the Galaxy Zoo dataset, a mapping system between the Galaxy Zoo 2 decision



**Figure 1**. The Galaxy Zoo 2 decision tree. Image from Willett *et al.* (2013).

2



**Figure 2**. Hubble’s tuning fork model. From

http://ay17-chusic.blogspot.com/2015/10/20-hubbletuning-fork.html

tree classification scheme and the Hubble Tuning Fork scheme (Figure 2) was developed. A machine learning system based on the Google’s Inception architecture for convolutional neural networks was trained on the Galaxy Zoo dataset to produce Tuning Fork classifications and tested against a 3rd party dataset of images classified manually by experts. The results from our system show that useful classifications can be achieved by applying a model trained on citizen-scientist labeled Galaxy Zoo 2 data on new datasets.

# RELATED WORK

In the astronomical community, the few automated galaxy classification systems have relied on more traditional methods, focusing on aggressive feature extraction algorithms making use of domain knowledge (such as WND-CHARM) to identify relationships among galaxies. These, however, have tended to focus on the more narrow classification of spirals and ellipticals, occasionally including edge-on spirals and irregular galaxies, and often work with much smaller datasets (see Dieleman *et al.* 2015 for a discussion). Kuminski and Shamir (2016) were rather unique when they made use of the “super clean” galaxies from the Galaxy Zoo 1 catalog (Lintott *et al.* 2008) to classify 3 000 000 galaxies into spirals and ellipticals.

While the top methods achieve ∼ 95% accuracy when separating ellipticals and spirals, they tend to perform much worse when the number of categories increases (de la Calleja and Fuentes 2004).

Gauthier *et al.* (2016), students in Prof. Ng’s machine learning class at Stanford, recently looked at several machine learning methods for classifying galaxies using the GZ2 dataset. While acknowledging the difficulty of directly classifying to the Hubble types, they sought to bridge the gap by modeling certain features, such as “roundness” and “diskiness”. They utilized the GZ2 decision tree to assign each galaxy to one of five categories: disc, spiral, elliptical, round, and other. In their preprocessing stage, images were cropped to reduce the file size, as well as reduce the number of nearby sources contaiminating the images. The galaxies were then rotated to align the principle axis, before proceeding with a background subtraction.

To further reduce the dimensionality of the problem, the authors applied principal component analysis (PCA), selecting the top 125 components to maintain *>* 99% of the variance. To classify the galaxies, they utilized a support vector machine (SVM) with a radial basis function (RBF) kernel, a decision tree, random forest, knearest neighbors, and an AdaBoost classifier, determing the classification accuracy using 10-fold cross validation. Overall, random forest produce the best results, achieving 67% accuracy. The poor success rate lead them to look into predicting probabilities (regression) rather than directly modeling the classes, similar to the Galaxy Zoo Kaggle challenge. They achieved better results in this regard, attaining ∼ 95% accuracy.

Overall, the biggest source of error was misclassifying spiral galaxies into the “other” category, which they attributed to the faintness (low signal-to-noise) of the spiral arms in many images. In addition, examining Figure 3 in their paper and comparing the original image with the 125 PC image, it appears that their method may hinder extracting the spiral arms by smoothing the disk, making classification more difficult. While this may be necessary for more traditional machine learning methods, deep learning can deal directly with the large feature space.

The Galaxy Zoo Kaggle challenge showed the power of convolutional neural networks (CNNs) when it comes to galaxy classification. Rather than relying on domain knowledge, the models had to learn to identify features on their own and were able to successfully reproduce the probabililty distributions of the citizen scientists. The processing pipeline for the top performing model (Dieleman *et al.* 2015) is schematically illustrated in Figure 3, which we shall examine next.

|  |
| --- |
| 3    **Figure 3**. Processing pipeline for the top model in the Galaxy Zoo Kaggle competition. From Dieleman *et al.* (2015). |

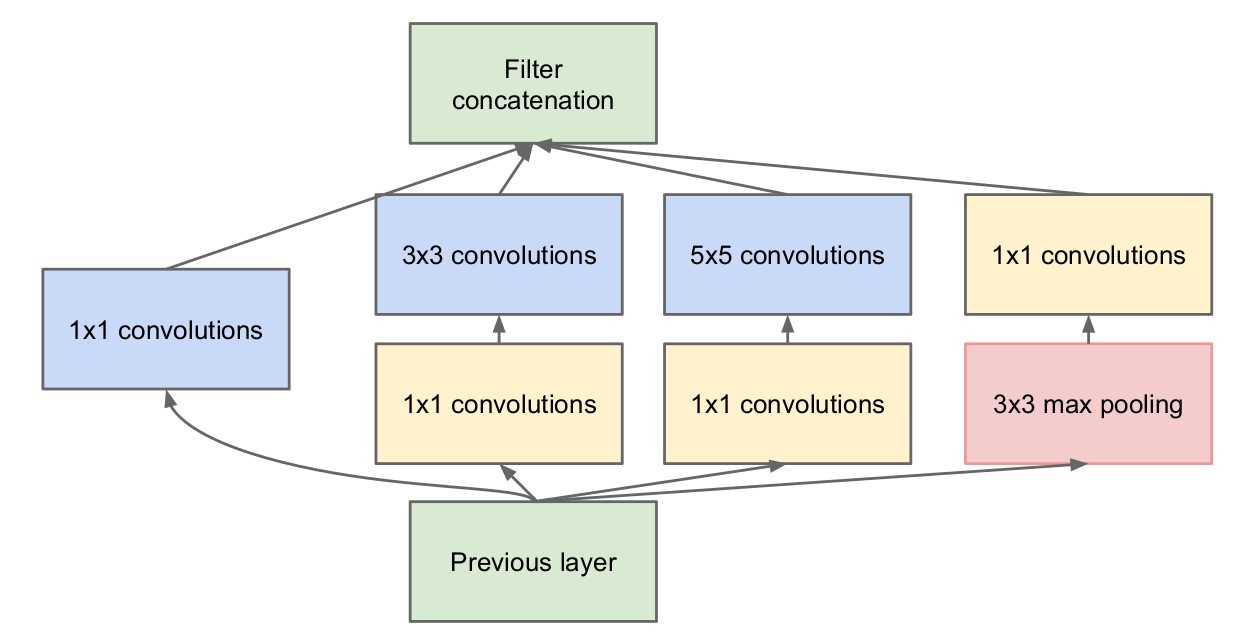
The winning algorithm was an ensemble method, averaging the results of many different CNNs. In the preprocessing stage, the image was cropped and rescaled several times down to 69 × 69 pixel images, which occasionally removed part of the galaxy. In some of their models, they used SExtractor to estimate the position and size of the galaxy, allowing them to center and rescale the galaxies to a standardized size. In addition, grayscaling was examined, although this lead to worse results. Due to the limited sample size, the number of images was increased by performing random pertubations, such as rotating, translating, scaling, and flipping, as well adjusting the color brightness on demand so that the model was never trained on the exact image twice. In addition, the number of images was increased by rotating, flipping and cropping each image into 16 different, but overlapping, 45 × 45 pixel images representing different viewpoints. Each of the 16 images were then passed together through the CNN, which performed several convolutions and poolings before concatenating the results and passing through a couple fully connected layers to output the final categorical probabilities. The probabilities were then averaged over 17 models.

Overall, the model did quite well, achieving ∼ 99% accuracy. It struggled most with the larger angular sized galaxies (more nearby), as well as those that were not radially symmetric.

# APPROACH

The existing systems from the Galaxy Zoo 2 Kaggle challenge do an excellent job of replicating the voting patterns of citizen science volunteers on the Galaxy Zoo 2 dataset. We developed an automated system trained on the large annotated Galaxy Zoo datasets to classify new imagery from other sources using the popular Hubble Tuning Fork scheme. While this can be done to some extent using the kaggle models, it requires cross-correlating expertly annoted images to find the optimal probability cutoffs to transform the probability distributions to Hubble types, adding an additional layer of complexity that the machine wasn’t required to learn. We developed a mapping between the two classification schemes to enable training of a general purpose galaxy classification model.

Our model differs slightly from the format of the Kaggle challenge. The Kaggle Galaxy Zoo challenge formulated the problem as a regression on the class probabilities, defined as the ratio of citizen science volunteers that gave a given galaxy a certain classification. To match the structure of our gold standard Tuning Fork scheme data, we instead treated this as a classification problem and selected only those galaxies whose vote fractions are within our chosen threshold for each Hubble 4



**Figure 4**. Inception block. The top image recognition CNNs in recent years use many inception blocks in their networks. From Szegedy *et al.* (2014).

type. This favors the more nearby galaxies, whose properties the top performing model in the Kaggle competition had a harder time predicting accurately. The system might also serve as a more interactive tool that could serve as a complement to the galaxy classification lab in *Stars, Galaxies, and the Universe*.

Based on prior work, the best approach to galaxy classification is a Deep Convolutional Neural Network. The top image recognition CNNs in recent years have used the inception model (Szegedy *et al.* 2014) as a building block in their networks (Figure 4). We follow this trend.

Since many of the images used in Galaxy Zoo 2 had poor consensus among the citizen scientists, we pruned the dataset to retain only high-confidence classifications for model training.

The Galaxy Zoo 1 dataset contains 11845 galaxies classified as spiral, elliptical, or merged. The Galaxy Zoo 2 data consists of 29407 images gathered from the Sloan Digital Sky Survey (SDSS), classified on the Hubble system by converting the Citizen Scientist ratings into Hubble types. To test the performance of our system, an independent expert-annotated dataset of 13507 galaxy images was obtained from the Sloan Digital Sky Survey (SDSS).

# Implementation

Our galaxy classification model was built on top of Google’s Tensorflow and the open-source deep learning library Keras. Keras has achieved popularity through ease of use and a wide selection of pre-built models. We use the Keras implementation of Google’s Inception V3 architecture, initialized to weights based on ImageNet pre-training. Training an Inception model from scratch is computationally intensive. By re-using the ImageNet weights provided by Google, we were able to considerably reduce training time.

Models were trained on the datasets in two stages. First, the final softmax layer was discarded, and a new softmax layer trained from scratch with all other network weights held constant from the ImageNet initialization. In the second step, the last three layers were retrained for 1000 epochs at a low learning rate (lr=0.001, momentum=0.9**)** using categorical cross-entropy loss, with all other weights held constant. The classes were weighted to mitigate the large class imbalances**.** The imagery is already centered and scaled to display the galaxy at a consistent size and location, but additional preprocessing was done to randomly rotate and flip the training images for each epoch.

Training and prediction was done on the University of Iowa ARGON high performance computing cluster. ARGON offers a wide range of compute configurations, including nodes equipped with nVidia workstation class GPU accelerators. For performance reasons, GPU acceleration proved necessary to train the inception model.

Three different models were trained on different datasets. One model was trained and tested on the Galaxy Zoo 1 dataset. Galaxy Zoo 1 is inadequate for Hubble classification purposes as the available class labels only distinguish whether a galaxy is a spiral, elliptical, or lenticular. However, there is prior work (de la Calleja and Fuentes 2004) showing 90% accuracy on classifying the Galaxy Zoo 1 dataset, which allows us to compare the performance of our system to earlier work in the field. In addition, a model was trained on the Galaxy Zoo 2 dataset (GZ2), and a model was created by re-training the GZ2-trained model on the expert-labeled dataset (GS2-EXP). The model built on GZ2 was tested on GZ2 and EXP and the model built on GZ2-EXP was tested on EXP. This allows us to assess the generalization performance of a classifier trained on the Galaxy Zoo 2 dataset when applied to a new dataset.

# Results

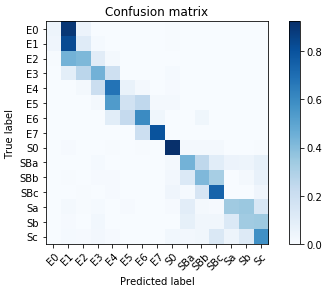
The system provides excellent performance on GZ1 and is able to generate meaningful predictions on all datasets. On the Galaxy Zoo 1 dataset, we report 95% classification accuracy, compared to 90% in previous literature (de la Calleja and Fuentes 2004). On Galaxy Zoo 2, we obtain 58% accuracy and a macro-averaged F1-measure of 0.496. Without retraining, the generalization performance on a new imagery dataset is disappointing at 23% with a macro-averaged F1-measure at 0.229, but after retraining on a small expert-labeled training set the performance improves to 46% accuracy and an F1-measure of 0.459. These results show that galaxy classification models trained on the Galaxy Zoo 2 dataset generalize well to other galaxy datasets if a small manually labeled dataset for retraining can be found.

|  |  |  |  |
| --- | --- | --- | --- |
| Train\Test dataset | GZ1 | GZ2 | EXP |
| GZ1 | 95% |  |  |
| GZ2 |  | 58% | 23% |
| GZ2-EXP |  |  | 46% |

**Table 1**. Accuracy of the models evaluated on the three data sets.

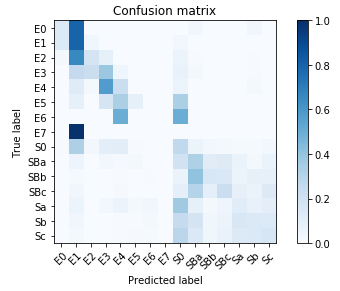
|  |  |  |
| --- | --- | --- |
| Train\Test dataset | GZ2 | EXP |
| GZ2 | .496 | .229 |
| GZ2-EXP |  | .459 |

**Table 2**. Macro-averaged F1 measure of the models evaluated on the Hubble classification scheme datasets.

****

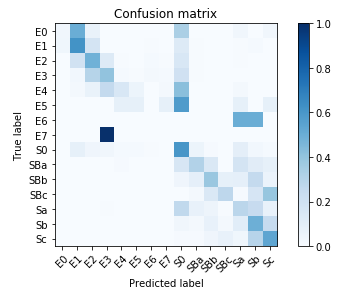
**Figure 5.** Normalized confusion matrix of GZ2-trained model evaluated on GZ2 test set.

Figure 5 shows the performance of the GZ2-trained model on the GZ2 test set. There is clear separation of the spirals, lenticulars (S0), and ellipticals, as we would expect from the results on GZ1. The ellipticals are tightly grouped around the diagonal; even if the classification is not correct, almost all galaxies are either in the correct class or an (almost correct) adjacent class. The spirals show a larger spread between classes, but most of the galaxies are still either in the correct class or an adjacent class.



**Figure 6.** Normalized confusion matrix of GZ2-trained model evaluated on EXP test set.

The GZ2 trained model when applied directly to the expert-labeled test set shows fairly poor performance, as can be seen in the confusion matrix (Figure 6). It does not consistently manage to distinguish spirals and ellipticals, much less discriminate different classes of spirals and ellipticals. It is possible that some of these differences reflect persistent differences in how experts and citizen scientists label galaxies, or perhaps a bias in our mapping from the Galaxy Zoo 2 decision tree to the Hubble Tuning Fork. However, when the top 3 inception modules are retrained on a training set selected from the expert-labeled data, the performance improves and is comparable to what we see on the original GZ2 dataset for most classes. A large number of both elliptical and spiral galaxies are incorrectly labeled as S0 galaxies, and the E6 and E7 galaxies are consistently mislabeled as Sa/Sb and E3, respectively. Despite these shortcomings, the system does show reasonable prediction accuracy for the other classes.



**Figure 7.** Normalized confusion matrix of GZ2-trained model evaluated on EXP test set.

This study shows that galaxy classification models trained on the Galaxy Zoo 2 dataset can be applied to previously-unseen galaxy imagery with good classification performance if a modestly sized manually-labeled training set from the new datasource is available.

# Future work

A challenge of the Galaxy Zoo datasets is the probabilistic class membership of the citizen scientist rated images. In this study we only retained high-confidence classifications, which potentially excludes important but ambiguous-looking corner cases. An improvement would be to acquire expert help to rate the discarded galaxies.

The independent expert-labeled dataset in this study originates from the Sloan Digital Sky Survey, which is also where the Galaxy Zoo data stems from. Future work should investigate whether the same generalization performance will hold on imagery from other observatories.

# REFERENCES

Gaia Collaboration, T. Prusti, J. H. J. de Bruijne, A. G. A.

Brown, A. Vallenari, C. Babusiaux, C. A. L. Bailer-Jones, U. Bastian, M. Biermann, D. W. Evans, and et al., [A&A](http://dx.doi.org/10.1051/0004-6361/201629272) [**595**, A1 (2016),](http://dx.doi.org/10.1051/0004-6361/201629272) [arXiv:1609.04153 [astro-ph.IM].](http://arxiv.org/abs/1609.04153)

W. Skidmore, TMT International Science Development Teams, and T. Science Advisory Committee, [Research in Astronomy and Astrophysics](http://dx.doi.org/%2010.1088/1674-4527/15/12/001) [**15**, 1945 (2015),](http://dx.doi.org/%2010.1088/1674-4527/15/12/001) [arXiv:1505.01195 [astro-ph.IM].](http://arxiv.org/abs/1505.01195)

Z. Ivezic, J. A. Tyson, T. Axelrod, D. Burke, C. F. Claver, K. H.

Cook, S. M. Kahn, R. H. Lupton, D. G. Monet, P. A. Pinto,

M. A. Strauss, C. W. Stubbs, L. Jones, A. Saha, R. Scranton,

C. Smith, and LSST Collaboration, in *American Astronomical Society Meeting Abstracts #213*, Bulletin of the American Astronomical Society, Vol. 41 (2009) p. 366.

K. W. Willett, C. J. Lintott, S. P. Bamford, K. L. Masters, B. D.

Simmons, K. R. V. Casteels, E. M. Edmondson, L. F. Fortson,

S. Kaviraj, W. C. Keel, T. Melvin, R. C. Nichol, M. J. Raddick, K. Schawinski, R. J. Simpson, R. A. Skibba, A. M. Smith, and

D. Thomas, [MNRAS](http://dx.doi.org/%2010.1093/mnras/stt1458) [**435**, 2835 (2013),](http://dx.doi.org/%2010.1093/mnras/stt1458) [arXiv:1308.3496.](http://arxiv.org/abs/1308.3496)

S. Dieleman, K. W. Willett, and J. Dambre, [MNRAS](http://dx.doi.org/10.1093/mnras/stv632) [**450**, 1441](http://dx.doi.org/10.1093/mnras/stv632)

[(2015),](http://dx.doi.org/10.1093/mnras/stv632) [arXiv:1503.07077 [astro-ph.IM].](http://arxiv.org/abs/1503.07077)

J. de la Calleja and O. Fuentes, [MNRAS](http://dx.doi.org/10.1111/j.1365-2966.2004.07442.x) [**349**, 87 (2004).](http://dx.doi.org/10.1111/j.1365-2966.2004.07442.x)

E. Kuminski and L. Shamir, [ApJS](http://dx.doi.org/10.3847/0067-0049/223/2/20) [**223**, 20 (2016),](http://dx.doi.org/10.3847/0067-0049/223/2/20) [arXiv:1602.06854.](http://arxiv.org/abs/1602.06854)

1. J. Lintott, K. Schawinski, A. Slosar, K. Land, S. Bamford,
2. Thomas, M. J. Raddick, R. C. Nichol, A. Szalay,D. Andreescu, P. Murray, and J. Vandenberg, [MNRAS](http://dx.doi.org/%2010.1111/j.1365-2966.2008.13689.x) [**389**, 1179 (2008),](http://dx.doi.org/%2010.1111/j.1365-2966.2008.13689.x) [arXiv:0804.4483.](http://arxiv.org/abs/0804.4483)

A. Gauthier, A. Jain, and E. Noordeh, [(2016).](http://cs229.stanford.edu/proj2016/report/GauthierJainNoordeh-GalaxyMorphology-report.pdf)

1. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov,
2. Erhan, V. Vanhoucke, and A. Rabinovich, ArXiv e-prints(2014), [arXiv:1409.4842 [cs.CV]](http://arxiv.org/abs/1409.4842)