Mon. Not. R. Astron. Soc. **000**, 000–000 (0000) Printed 10 March 2015 (MN LATEX style file v2.2)

Galaxy Zoo: Detailed Morphological Classifications for 48,000 galaxies from CANDELS⋆

B. D. Simmons[1]†, and a *lot* of other people to be named later

[1] *Oxford Astrophysics, Denys Wilkinson Building, Keble Road, Oxford OX1 3RH, UK*

10 March 2015

ABSTRACT

To be rewritten, probably last.

Galaxies are sometimes really far away. The distant ones are pretty cool, because they tell us what the Universe was like back when it was just a kid, or maybe a teenager. You really have to look hard to see these galaxies, but once you do, what you do see tells you a whole lot. I mean, it's not exactly a WYSIWYG type of thing: there's a lot of work to figure out what the faint stuff you see really means. We did a bunch of work, and we think we did pretty well. Also, we compared to others who have done different kinds of work to try and answer some of the same questions. But we have a unique way of answering them, so here are those answers, and you can use them to answer other questions about the Universe.

Key words:
galaxies: general — galaxies: evolution — galaxies: morphology — galaxies: structure

1 INTRODUCTION

Outline:
Galaxy morphologies trace physics & dynamics and are therefore important.

They have a long history and there are many types, including those that use proxies [which are useful but not without problems] and those that reduce a galaxy and its billions of stars to a single number [same]. Visual morphologies are in many ways still ideal as they are able to provide incredibly detailed information about galaxies.

Visual morphologies have a scale problem without many eyes, and they're often not quantified. Galaxy Zoo to the rescue.

Galaxy Zoo started in 2007 and has provided robust, quantified visual classifications for more than 1,000,000 galaxies to $z \sim 1$.

Much use has been made of these, including [summarize and include non-GZ papers].

Here we present the visual morphologies of galaxies imaged by the *HST* treasury survey CANDELS.

Section summary and cosmology.

2 OBSERVATIONAL DATA

2.1 Images

The Cosmic Assembly Near-infrared Extragalactic Legacy Survey (CANDELS; Grogin et al. 2011; Koekemoer et al. 2011) is an *HST* Treasury programme combining optical and near-infrared imaging from the Advanced Camera for Surveys (ACS) and Wide Field Camera 3 (infrared channel; WFC3/IR) across five well-studied survey fields (GOODS-North and -South, Giavalisco et al. 2004; EGS, Davis et al. 2007; UDS, Lawrence et al. 2007, Cirasuolo et al. 2007; and COSMOS, Scoville et al. 2007) using a two-tiered "deep" and "wide" approach. Each of the wide fields (UDS, COSMOS, EGS and flanking fields to the GOODS-S and GOODS-N deep fields) are imaged over 2 orbits in WFC3/IR, split in a 2:1 ratio between filters F160W and F125W, respectively, with parallel exposures in F606W and F814W using ACS. Each of the deep fields (GOODS-S and GOODS-N) are imaged over at least 4 orbits each in both the F160W and F125W filters and 3 orbits in the F105W filter, with ACS exposures in F606W and F814W in parallel. These are reduced and combined to produce a single mosaic for each field in each band, with drizzled resolutions of 0.03″ and 0.06″ per pixel for ACS and WFC3/IR, respectively (a process described in detail by Koekemoer et al. 2011).

Here we use the CANDELS ACS and WFC3/IR images from within the first set of data to be classified within the Galaxy Zoo interface. Those data cover the COSMOS,

⋆ This publication has been made possible by the participation of more than COUNT volunteers in the Galaxy Zoo project. Their contributions are individually acknowledged at http://authors.galaxyzoo.org/ .
† E-mail: brooke.simmons@astro.ox.ac.uk

GOODS-South, and UDS fields. The 4th release of Galaxy Zoo included all detections with $H \leqslant 25.5$ from these 3 combined fields, comprising 49,555 unique images. These were shown to visitors to the website galaxyzoo.org[1] starting on the 10th of September, 2012.

The images shown to the public are colour composites of ACS I ($F814W$), WFC3 J ($F125W$), and WFC3 H ($F160W$) filters for the blue, green and red channels, respectively. The angular sizes of the images in different filters are matched, and the native point-spread functions (PSFs) are used. The images are combined with an asinh stretch (described in detail in **?**) with a non-linearity value of 3.0.

Sources in the dataset vary greatly in size and surface brightness, and therefore a single set of values for channel scalings is not adequate to capture the variety of features across the images. We therefore use a variable scaling based on the magnitude and size of each target source. For each image the R, G, and B channels have a fixed ratio of [not sure; must get this from Jeyhan], and the multiplier can vary between A and B. Figure **??** shows examples of these colour composites across a wide range of source fluxes and sizes.

Each colour image is 424 pixels square. The angular size of the image varies, such that the colour image encompasses at least 3 times the 80% flux radius of the target source, with a minimum screen-to-WFC3 zoom ratio of 1:10 and a maximum ratio of 3:1. The Galaxy Zoo interface loads the normal colour images by default, and the user may choose to display an inverted colour image, but may not otherwise change the image scaling or size within the software while performing the classification.

2.2 Photometry

Brief description of photometric catalogs. Focus on IJH because that's all the images consider. Do mention all the extra stuff available, but note that the classifications themselves don't depend on them.

2.3 Redshifts

Some of them have speczs. Lots of them have photzs, including CANDELS, 3D-HST and several previous surveys.

Comparison of photozs and speczs; might be able to just reference the other papers.

What do we do about those without redshifts? We use them with caution?

2.4 Calibration and Simulated Images

Mention the duplicated images in GDS, which were put in both identically (this was by accident - so we got double each of these and can theoretically check the variance in classifications, but that's kind of a stupid justification and I'm not quite sure how to mention these without saying "oopsies") and at 2-epoch depth (though this has not actually happened yet so I may need to just leave it out). Also it'd be nice to mention the FERENGIfied images, but ... those aren't in yet either. Basically, the only reason to have this section at the moment is to say "oops".

[1] Archived at zoo4.galaxyzoo.org

3 CLASSIFICATION DATA

3.1 Decision Tree

The goal of Galaxy Zoo CANDELS is to provide detailed quantitative visual morphologies of galaxies observed by the deepest, most complete *HST* multi-wavelength legacy survey to date. There are many morphological features of interest, including both broad questions about a galaxy's overall appearance and more detailed questions about specific features.

We employ a tree-based structure for collecting information on these morphological features, a strategy that has been used successfully in both Galaxy Zoo 2 (refs here) and Galaxy Zoo: Hubble (and here). The decision tree, shown visually in Figure 1 and in text in Table 1, first asks the classifier to choose between the broad categories of "smooth and rounded", "features or disk", and "star or artifact". The next question either exits the classification (if the classifier has indicated the image is of a star or artifact) or asks for further details about the galaxy.

If the classifier has indicated in the first question that the galaxy has features or a disk, a series of follow-up questions are asked about features such as clumps, spiral patterns, bulge strength, and the presence of a bar. If the classifier has instead indicated the galaxy is mostly smooth and rounded, the next question asks them to rate the overall roundedness, a question roughly corresponding to an axis ratio measurement. Finally, when the classifier has finished answering all follow-up questions about either the "smooth" or "featured" galaxy, they are then asked whether the galaxy is undergoing a merger, has tidal tails, or has both, or neither.

The tree-based structure has a number of advantages. First, it collects substantially more information on each galaxy than a single question would, and captures a more detailed classification of higher-order structures while minimising the effort required on the part of the classifier by only asking for relevant inputs based on the answers provided to previous questions.

Second, it focuses the classifier on a single feature at a time, highlighting each feature. This resets the attention of the classifier with each new question and avoids the problems that may result when a person is presented with a large number of decision tasks at once, including a decrease in optimal decision-making (references for overchoice) and a reduced ability to recognise the unexpected (references for inattentional blindness).

Third, the tree-based structure is especially optimal for an interface which may collect classifications from users who have never before seen an image of a galaxy and may seek additional training. Within the interface, the classifier may optionally display training images in a "Help" section that shows different examples of the feature relevant to the current question. Asking single-topic questions in turn permits a full set of training images to be available throughout the classification without placing an unnecessary cognitive load on the classifier.

Other advantages we should mention?

The disadvantage of a tree-based classification structure concerns the dependencies introduced into the vote fractions by such a structure. A classifier cannot, for example, answer that the same galaxy has both a mostly smooth appear-

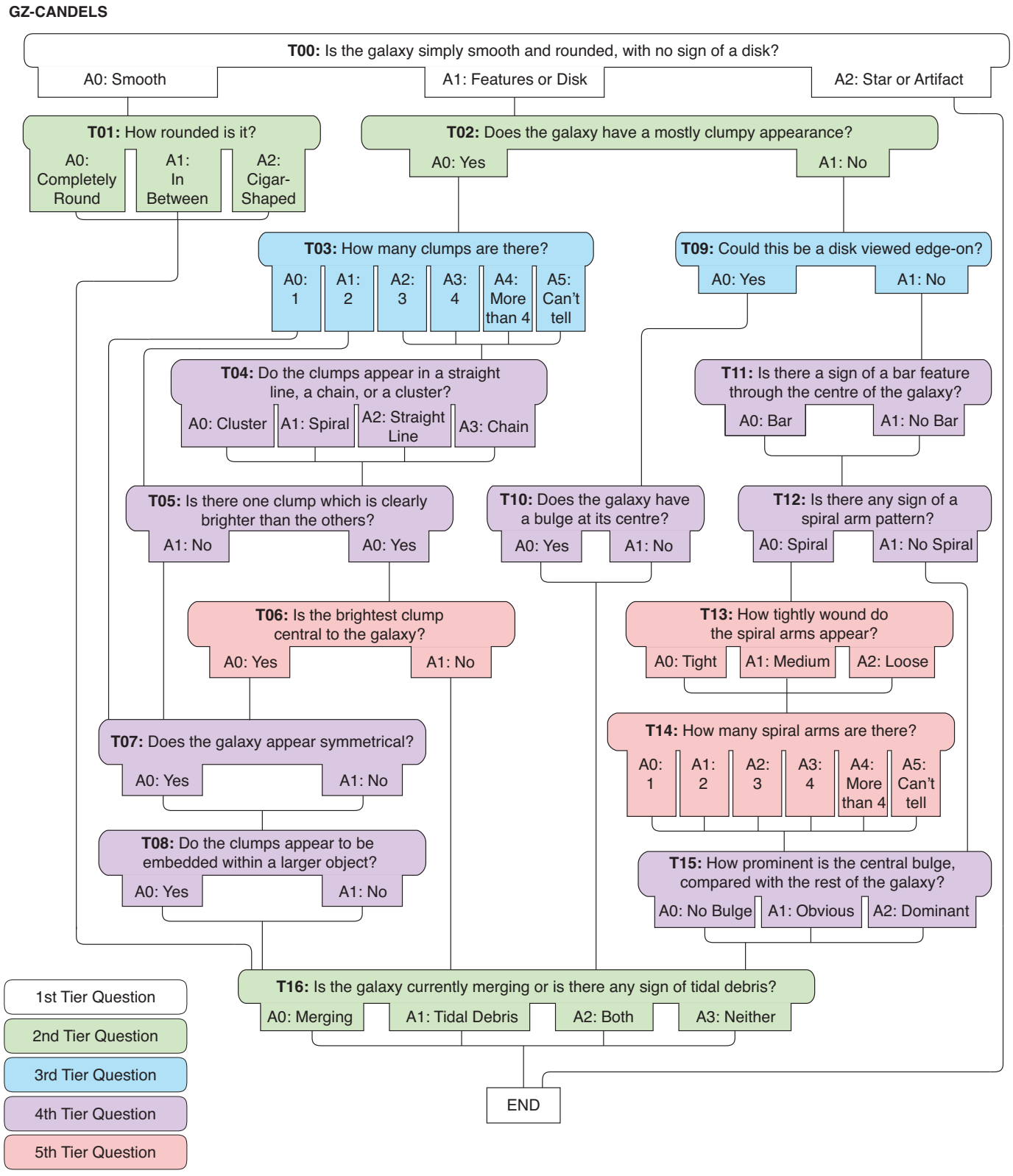

Figure 1. The Decision Tree for Galaxy Zoo: CANDELS in visual format. There are 16 tasks, with one question per task and up to 6 answers per question. Questions are coloured according to the minimum number of branches prior to that question. All users are asked the first question (task T00), and there are 4 subsequent levels of branching. The tree is also shown in text in Table 1.

ance and also has a spiral feature. This is in some ways an advantage, as it prevents contradictory and unphysical classifications, but it also means that an analysis of morphological vote fractions with the goal of examining spiral galaxies must account for the fact that whether a given classifier reached the spiral branch of the decision tree depends on their answer to the questions preceding it.

Accounting for dependencies of questions in deeper branches of the decision tree on higher-level questions is, however, a manageable task which has been undertaken successfully in many previous studies of specific galaxy structural features (for specific examples, see [citation bomb here]).

The Galaxy Zoo CANDELS decision tree is shown in visual form in Figure 1 and in text form in Table 1. We note that this tree is most similar to the tree used in the Galaxy Zoo: Hubble project (shown in **?**) (note to self to check that the full tree is actually shown there), which also has an additional branch identifying clumpy galaxies and focusing on the detailed structure of galaxy clumps. There are small differences, however: for example, Task 10, the question about a bulge in an edge-on disk, is a Yes/No question here, whereas in previous iterations of the decision tree this question also asked whether the bulge shape was rounded or boxy. Additionally, the final question in the tree (Task 16) is substantially different from previous versions and is here only concerned with galaxy mergers and tidal features.

After the classification of each image is finished, the classifier is asked "Would you like to discuss this object?" If the classifier selects "No", a new image is shown for classification. If the classifier selects yes, a new window opens with a discussion page focused on the image they have just classified. Within this part of the Galaxy Zoo software, called Talk, users may ask questions and make comments on specific images, or engage in more general discussions. Users may also "tag" images and discussions using a format identical to Twitter's hashtag system. Some of these tags were used in the pre-analysis of Galaxy Zoo CANDELS data, on which more details are given in Section **??** below.

3.2 Raw classifications

The first classification of an image from CANDELS was registered on the Galaxy Zoo interface[2] on the 10th of September 2012. The final classification considered here, in the first phase of Galaxy Zoo CANDELS, was registered on the 30th of November 2013. Between these times, the site collected 2,149,206 classifications of 52,076 CANDELS subjects from 41,552 registered volunteers and 53,714 web browser sessions where the user did not log in. For all analysis presented here we have assumed that each unregistered browser session contains classifications from a single, unique volunteer.

Subjects within a given Galaxy Zoo sample are chosen randomly for classification, so that the number of independent classifications per galaxy builds up uniformly through the full sample. Once a pre-set classification limit has been reached, the subject is retired from the active classification pool. The initial goal for Galaxy Zoo CANDELS was to obtain at least 40 independent classifications for each galaxy.

This uniform retirement limit was modified twice during the project. In the first instance, a pre-analysis of the dataset performed when the average number of classifications per galaxy had reached approximately 20 revealed 11,837 subjects where further classification was unlikely to provide any additional information. These subjects were identified with the help of a set of subjects tagged in the Galaxy Zoo Talk software as "#toofainttoclassify" and "#FHB" (which stands for "Faint Hubble Blob"). Tags in Galaxy Zoo Talk are generally highly incomplete; thus the 204 tagged subjects were used as tracers during a further examination of all subjects in magnitude-surface brightness parameter space. The selection, made from initial photometry, was deliberately conservative, retiring only those subjects where it was clear that the classification vote fractions had converged at all tiers of the classification tree. During this analysis, an additional 1,555 subjects were identified as highly likely to be stars or artifacts and were also retired.

The second modification of the retirement limit was implemented 1 year after the project start. At this time, the retirement limit was raised to 80 classifications for all galaxies where at least 20% of volunteers had answered "Features or Disk" to the first question (task T00 in Figure 1 and Table 1). This is a higher retirement limit than in previous Galaxy Zoo projects, and it is justified by the increased complexity of the question tree compared to, e.g., Galaxy Zoo 2 (**?**). The Galaxy Zoo CANDELS question tree has an additional branch level, and the number of volunteers answering a question is typically reduced at each branch point. Thus, 40 classifications at the first question may not be enough to ensure convergence in, for example, task 14, "How many spiral arms are there?", a 5th-tier question with 6 possible answers. The increased retirement limit affected 7,402 subjects.

Figure 2a shows the distribution of total classification counts within the sample. The majority of subjects received 40 classifications, but the distribution is asymmetric: there are peaks at $\sim$ 20, 40, and 80 classifications, consistent with the description above. The Lorenz curve of classifications by volunteers is shown in Figure 2b. The curve is highly skewed from the 1 : 1 line that would be seen if all volunteers contributed the same number of classifications; the top 9% volunteers contributed 80% of total classifications. The Gini coefficient for classfications, i.e., the fractional difference in area under the Lorenz curve versus the dashed line, is 0.86. This is typical of past Galaxy Zoo projects and Zooniverse[3] citizen research projects in general (could cite VOLCROWE CiSE paper here).

The values in Figure 2 are raw classification counts; while raw classification counts and vote fractions are certainly useful, we additionally apply a user weighting scheme to classifications to produce a cleaner set of vote fractions for each subject. The user weighting is described in further detail below.

3.3 User Weighting

Multiple methods of user weighting have been successfully employed by many different Zooniverse projects (**???????**).

[2] zoo4.galaxyzoo.org

[3] zooniverse.org

Task	Question	Responses	Next
T00	*Is the galaxy simply smooth and rounded, with no sign of a disk?*	smooth	01
		features or disk	02
		star or artifact	**end**
T01	*How rounded is it?*	completely round	16
		in between	16
		cigar-shaped	16
T02	*Does the galaxy have a mostly clumpy appearance?*	yes	03
		no	09
T03	*How many clumps are there?*	1	07
		2	05
		3	04
		4	04
		more than four	04
		can't tell	04
T04	*Do the clumps appear in a straight line, a chain or a cluster?*	cluster	05
		spiral	05
		straight line	05
		chain	05
T05	*Is there one clump which is clearly brighter than the others?*	yes	06
		no	07
T06	*Is the brightest clump central to the galaxy?*	yes	07
		no	16
T07	*Does the galaxy appear symmetrical?*	yes	08
		no	08
T08	*Do the clumps appear to be embedded within a larger object?*	yes	16
		no	16
T09	*Could this be a disk viewed edge-on?*	yes	10
		no	11
T10	*Does the galaxy have a bulge at its centre?*	yes	16
		no	16
T11	*Is there a sign of a bar feature through the centre of the galaxy?*	bar	12
		no bar	12
T12	*Is there any sign of a spiral arm pattern?*	spiral	13
		no spiral	15
T13	*How tightly wound do the spiral arms appear?*	tight	14
		medium	14
		loose	14
T14	*How many spiral arms are there?*	1	15
		2	15
		3	15
		4	15
		more than four	15
		can't tell	15
T15	*How prominent is the central bulge, compared with the rest of the galaxy?*	no bulge	16
		just noticeable	16
		obvious	16
		dominant	16
T16	*Is the galaxy currently merging or is there any sign of tidal debris?*	merging	**end**
		tidal debris	**end**
		both	**end**
		neither	**end**

Table 1. [I'd like to make this a 2-column type table, split after T08, but I don't really have the energy...] The Galaxy Zoo CANDELS decision tree, comprising 16 tasks and 51 responses. Each task is comprised of a single question and up to 6 possible responses. The first question is Task 00, and a classification is completed by responding to all subsequent questions until the end of the tree is reached. The 'Next' column indicates the subsequent task the classifier is directed to upon choosing a specific response. Although a classifier will flow through the tree from top to bottom, there is no path through the tree that includes all tasks.

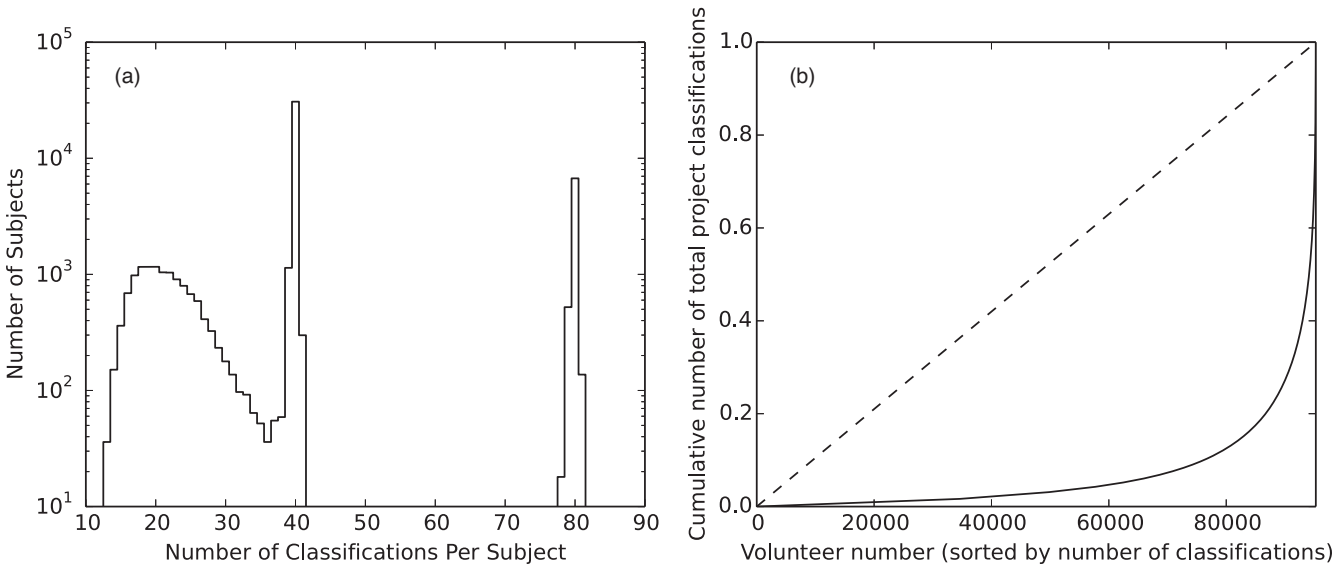

Figure 2. Basic information on classifications. *Left*: Distribution of number of classifications per subject in Galaxy Zoo CANDELS. The majority of images have 40 independent classifications each; a subset of 13,392 were retired early after being identified as too faint and low-surface-brightness for additional classifications to be useful (11,837) or as stars or artifacts (1,555). Subsequently, 7,402 subjects where at least 20% of classifiers registered a vote for "Features or Disk" in the first task were re-activated with a retirement limit of 80 classifications, in order to ensure a complete sampling of the deepest branches of the question tree. *Right*: Cumulative distribution of classifications by volunteers, where the volunteers are sorted in order of least to most classifications contributed (Lorenz curve for classifiers). If every volunteer had contributed the same number of classifications, the Lorenz curve would be equal to the dashed curve. The top 9% of users contributed 80% of the classifications (Gini coefficient = 0.86).

In general, the optimal choice of user weighting depends on the amount of information available per subject and the goal of the project. In Galaxy Zoo CANDELS the goal is to converge to a classification for each galaxy whilst still allowing for unexpected discoveries, and there is ample information from classifiers but little information on the "ground truth", i.e., we do not know what the true intrinsic classification is for even a modest fraction of the sample.

For these reasons, we adopt an iterative consensus-based weighting method, following previous Galaxy Zoo projects. This weighting scheme effectively identifies the small proportion of classifiers whose contributions are routinely errant compared to other classfiiers (or consistent with random inputs) and downweights their contributions, while preserving the inputs from the vast majority of users.

Weights for each user are computed based on a mean consistency factor, $\overline{\kappa}$, which is the average of consistencies for each of that user's classifications. For a given classification i composed of a series of completed tasks t answered about a specific subject, we compare the user's answer to each task with the aggregated classifications of other users of the same subject. Each task has a_t answers from all users, each of which is assigned to one of $N_{r,t}$ possible responses to the task. We define the vote fraction for a particular response r as $f_r \equiv a_r/a_t$, where a_r is the number of positive answers for that response (i.e., the number of classifiers who selected that response out of all possible responses to the task).

For each task that was completed by the classifier in classification i, the consistency index κ_r for each response r to that task t is

$$\kappa_r = \begin{cases} f_r & \text{if the classifier's answer corresponds to this response,} \\ (1 - f_r) & \text{if the answer does not correspond.} \end{cases} \quad (1)$$

The consistency for that task, κ_t, is the average of these indices over all possible responses. For example, if a classifier responded "Star or Artifact" to Task T00 for a particular subject, and the overall vote fractions on that task for that subject are ("Smooth", "Features or Disk", "Star or Artifact") = $(0.1, 0.6, 0.3)$, then the user's consistency for Task T00 for this classification is

$$\kappa_t = [(1 - 0.1) + (1 - 0.6) + 0.3]\,/3 = 0.5\overline{3}.$$

In the above example, the user's answer to Task T00 leads to the end of the workflow (Table 1), so this κ_t is also equal to the user's consistency for the overall classification, κ_i. More generally, the classification consistency is the answer-weighted average of the task consistencies:

$$\kappa_i = \frac{\sum_t \kappa_t a_t}{\sum_t a_t}, \quad (2)$$

where each sum is over the number of tasks the user completed during the classification.

Following this calculation for the entire classification database, each user's average consistency is calculated as

$$\overline{\kappa} = \frac{1}{N_i} \sum_i \kappa_i. \quad (3)$$

Averaging over a user's individual consistency values for all classifications effectively downweights those contributions from users whose classifications regularly diverge from

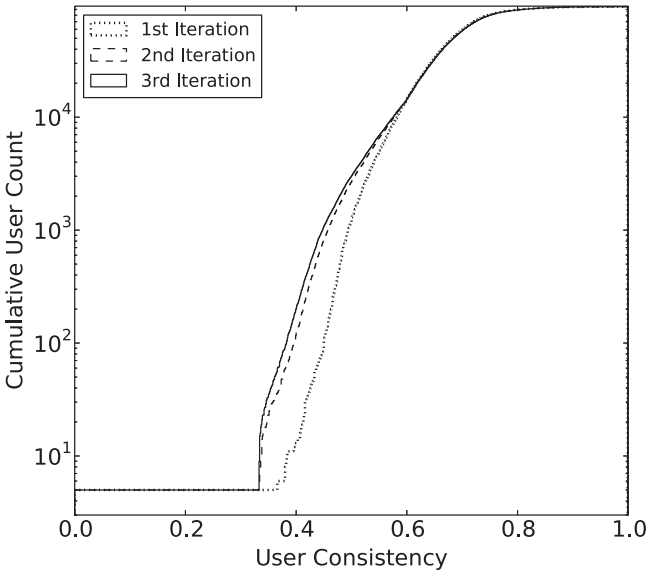

Figure 3. Distribution of user consistencies $\overline{\kappa}$ after 1 (dashed), 2 (dotted), and 3 (solid) iterations of the consistency-based weighting method (described in Section 3.3). Convergence of this method requires relatively few iterations: further iterations do not change significantly from the solid curve. Approximately 85 per cent of users have $\overline{\kappa} \geqslant 0.6$ and weights $w = 1$.

the consensus whilst preserving the diversity of classifications from volunteers who are *on average* consistent with each other. It also allows for the classifications of skilled volunteers to remain highly weighted even on difficult subjects where the individual consensus is skewed (e.g., if an image is very noisy or if a nearby artifact is distracting to less experienced volunteers).

The user weight is then calculated as

$$w = \min\left(1.0, (\overline{\kappa}/0.6)^{8.5}\right), \quad (4)$$

a formulation that preserves a uniform weighting for any classifier with $\overline{\kappa} \geqslant 0.6$ and downweights those with a lower consistency rating.

The weighted consensus classifications are then calculated for each subject by summing the weighted votes for each task and response, and reporting the vote fractions f for each. As the user weights are calculated via comparison with the consensus, which leads to a new consensus, this method can be iterated until the user weights converge to a stable value.

In practice, the number of iterations required to reach this goal is low (e.g., 3 or less; **??**). In Figure 3 we show the distribution of user consistencies after 1, 2 and 3 iterations of the above method. Approximately 3 per cent of users have consistency $\overline{\kappa} < 0.5$ (corresponding to a weight $w \approx 0.2$), whereas 85 percent of users have an end weight of $w = 1$. The vast majority of Galaxy Zoo volunteers contribute highly valuable information to the project.

4 COMPARISON TO OTHER CLASSIFICATIONS

Kartaltepe et al. 2014

Distribution of Sersic indices for different galaxy types

5 SOME KIND OF INITIAL RESULT

Clumpy galaxies as a function of redshift? Not too difficult to show this.

6 SUMMARY

Galaxies! We have galaxies!

ACKNOWLEDGMENTS

The development of Galaxy Zoo was supported in part by the Alfred P. Sloan Foundation. Galaxy Zoo was supported by The Leverhulme Trust.

This work is based on observations taken by the CANDELS Multi-Cycle Treasury Program with the NASA/ESA HST, which is operated by the Association of Universities for Research in Astronomy, Inc., under NASA contract NAS5-26555.

REFERENCES

Cirasuolo M. et al., 2007, MNRAS, 380, 585
Davis M. et al., 2007, ApJ, 660, L1
Giavalisco M. et al., 2004, ApJ, 600, L93
Grogin N. A. et al., 2011, ApJS, 197, 35
Koekemoer A. M. et al., 2011, ApJS, 197, 36
Lawrence A. et al., 2007, MNRAS, 379, 1599
Scoville N. et al., 2007, ApJS, 172, 1