Mon. Not. R. Astron. Soc. **000**, 000–000 (0000) Printed 16 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*

16 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

This paper presents morphological classifications of nearly 50,000 galaxies imaged in the Cosmic And Near-infrared Deep Extragalactic Legacy Survey (CANDELS; Grogin et al. 2011; Koekemoer et al. 2011) measured by the Galaxy Zoo[1] project (Lintott et al. 2008). Over 95,000 volunteers have contributed over 2,000,000 detailed galaxy classifications to this effort. We combine, on average, 43 independent classifications of each galaxy to produce detailed, quantitative morphological descriptions of these distant galaxies along many physical axes of interest.

The shape and appearance of a galaxy trace the underlying physical processes that have formed it and continue to influence its evolution. For example, the signatures of past merger events (from $z \sim 2$ onwards; Martig et al. 2012) are visible even at $z = 0$ in the form of a galactic bulge; the strength of the bulge is tied to the strength of the merger, as indeed the lack of a bulge indicates a lack of significant mergers (e.g., Kormendy et al. 2010). Likewise, other morphological features are tied to disk instabilities and resonances (e.g., warps, bars, rings **??**), and orbital changes from the disruptive (mergers; e.g., Darg et al. 2010b,a; **?**; **?**) to the relatively subtle (e.g., bars, **??**). Furthermore, combination of morphological parameters with other measures, such as environment, color, mass and star formation histories (e.g. Bamford et al. 2009; **?**; **?**; **?**), can provide more insight than either alone.

Morphological measures have a long history in astronomy (e.g., Hubble 1926; **?**; **?**; **?**; **?**). The computerized era of astrophysics has brought with it a number of automated morphological classification techniques. Some use multiple parameters to characterise a galaxy's distribution of light (Sérsic 1968; **?**; **?**), while others adopt a non-parametric approach, each reducing a galaxy to one number (and often used in combination; e.g. **?????**). These lend themselves relatively well to large-scale processing of images from galaxy surveys (e.g. **????**) and provide a uniform quantitative set of measures. Modern machine learning techniques are also well-tested and applicable to large data sets (**??**).

However, no computer has yet exceeded the human brain's capacity for pattern detection and serendipitous discovery. Visual morphologies remain among the most nuanced and powerful measures of galaxy structure. Galaxy Zoo combines the strengths of both visual and computer-driven approaches, using the Internet to collect more independent and complete visual classifications than any group

⋆ This publication has been made possible by the participation of more than 95,000 volunteers in the Galaxy Zoo project. The contributions of the more than 40,000 of those who registered a username with Galaxy Zoo are individually acknowledged at http://authors.galaxyzoo.org/ .
† E-mail: brooke.simmons@astro.ox.ac.uk
[1] zoo4.galaxyzoo.org

of astronomers is realistically capable of and combining these classifications via tested and proven techniques.

Since 2007, Galaxy Zoo has been a unique resource of quantitative and statistically robust visual galaxy morphologies. Prior to Galaxy Zoo CANDELS, three Galaxy Zoo projects have collected morphologies for over $1,000,000$ galaxies using the largest surveys to date to $z \sim 1$. These projects have been and continue to be extremely scientifically productive, both for the project team (**??????**) and for the larger scientific community (**?????????**).

Here we present the Galaxy Zoo visual morphologies of $49,555$ galaxies imaged by the largest near-infrared *Hubble Space Telescope (HST)* survey to date, CANDELS, which images galaxies at rest-frame optical wavelengths to $z \approx 2.7$.

In Section 2 we describe the observational data and the preparation of CANDELS images for use in Galaxy Zoo. In Section 3 we detail the collection of morphological classifications and the method of weighting and combining independent classifications for each galaxy. Section 4 compares Galaxy Zoo classifications to other morphological measurements. In Section 5 we show an example result using the classifications, and in Section 6 we summarize. Throughout this paper we use the AB magnitude system, and where necessary we adopt a cosmology consistent with ΛCDM, with $H_0 = 70$ km s^{-1}Mpc^{-1}, $\Omega_{\rm m} = 0.3$ and $\Omega_\Lambda = 0.7$ (Bennett et al. 2013).

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, 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 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?

3 CLASSIFICATION DATA

3.1 Definition of Terms

Throughout this paper we follow Willett et al. (2013) and **?** in adopting the following terms to describe different parts of the Galaxy Zoo software and data:

- **User** or **Volunteer**. Those classifying galaxies within the Galaxy Zoo software[2] are essential to the success of the project. While it is true that the software is written so that a user could in principle be a machine, during this project we have not included machine classifications, and thus we use the terms "user" and "volunteer" interchangeably here.

[2] All classifications discussed here were collected via web software.

• **Subject** or **Image**. Within the Zooniverse software, a subject is a unit of data to be classified. For other projects this may include light curves, groups of images, video or audio files. In Galaxy Zoo CANDELS, each subject consists of a single image, with the goal of classifying 1 galaxy per image. We therefore use the term "subject" interchangeably with "image" here. I'm not as comfortable with this one as I am with User/Volunteer. Maybe I should be stricter about Subjects.

• **Classification**. Galaxy Zoo CANDELS asks the user to complete several tasks to classify each subject. A classification is a unit of data that consists of 1 complete flow through the decision tree described in Section 3.2.

• **Task** and **Question**; **Response** and **Answer**. Each task in Galaxy Zoo CANDELS consists of a single question, with 2 or more possible responses, 1 of which the user selects as their answer in order to move on to the next task.

3.2 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.

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 appearance 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 Melvin et al. 2014) (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 3.3 below.

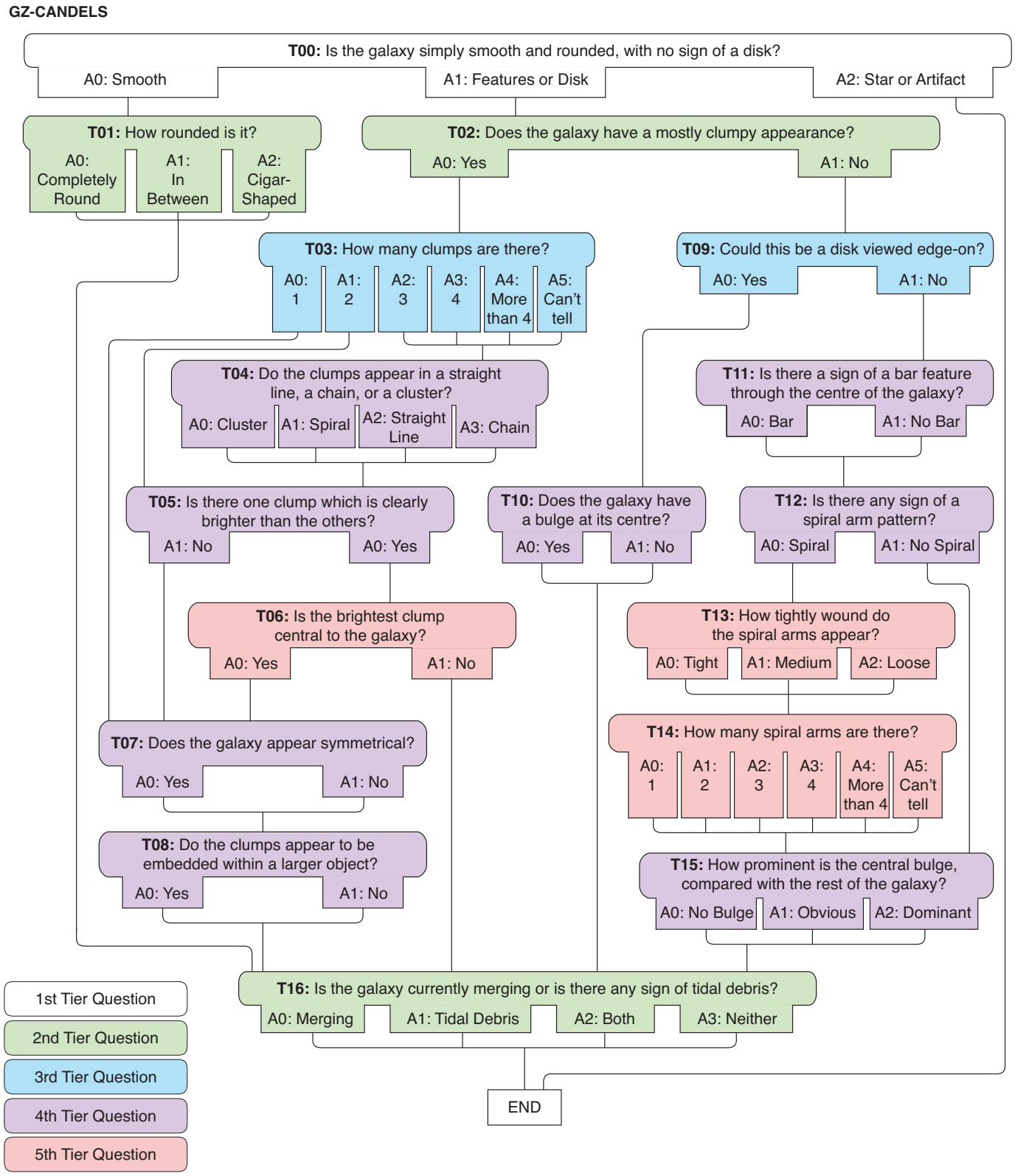

Figure 1. The Decision Tree for Galaxy Zoo: CANDELS in visual format. There are 16 tasks, with one question per task and up to 6 answers per question. Questions are coloured according to the minimum number of branches prior to that question. All users are asked the first question (task T00), and there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

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.

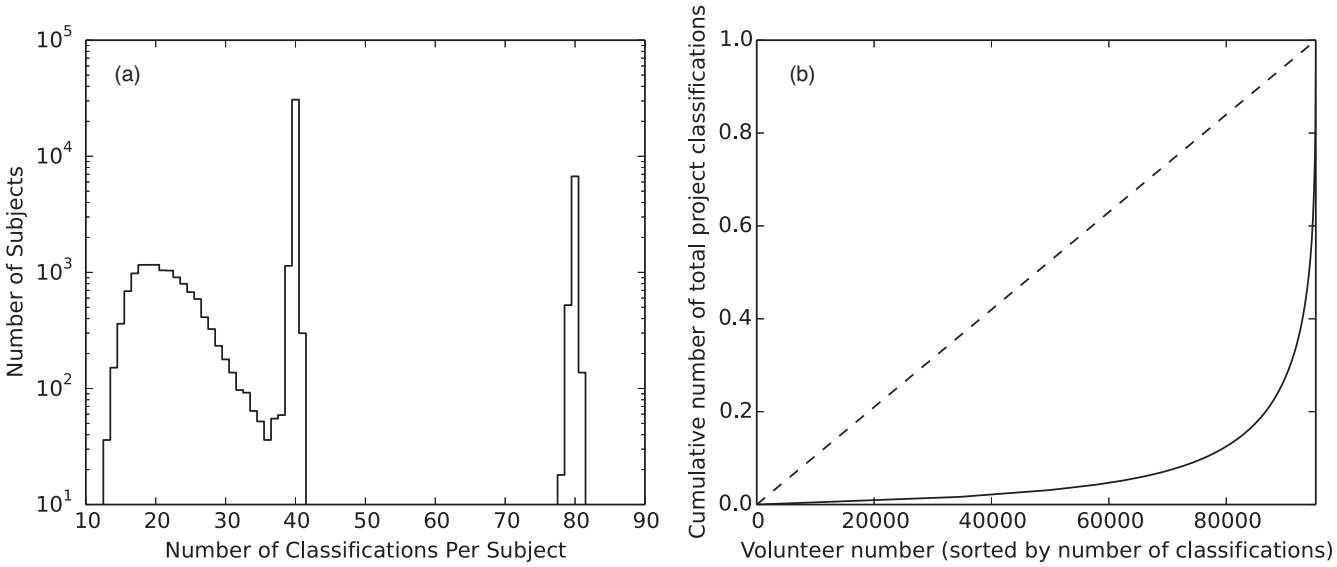

Figure 2. Basic information on classifications. *Left*: Distribution of number of classifications per subject in Galaxy Zoo CANDELS. The majority of images have 40 independent classifications each; a subset of 13,392 were retired early after being identified as too faint and low-surface-brightness for 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).

3.3 Raw classifications

The first classification of an image from CANDELS was registered on the Galaxy Zoo interface[3] 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 (Willett et al. 2013). 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.

[3] zoo4.galaxyzoo.org

This is typical of past Galaxy Zoo projects and Zooniverse[4] 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.

3.4 User Weighting

Multiple methods of user weighting have been successfully employed by many different Zooniverse projects (Lintott et al. 2008; Bamford et al. 2009; Lintott et al. 2011; **?**; **?**; **?**; **?**). 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

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

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

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

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 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 previous projects; Bamford et al. 2009; Willett et al. 2013). 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 $w \lesssim 0.2$), 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.

3.5 Use of Classifications in Practice

The branched nature of the decision tree (Figure 1) means that selection of a sample of galaxies for a given morphological investigation may depend on a number of factors. For example, it is possible to choose a quantitative threshold for selection of a sample of galaxies with a given feature or combination of features corresponding to one's optimal trade-off between sample completeness and purity. One may also weight a population analysis by the vote fraction for a particular morphological feature (making the assumption that the probability of a galaxy having that feature, or the strength of the feature, is a function of the vote fraction). However, for all tasks below T00 in the tree, it is important

[4] zooniverse.org

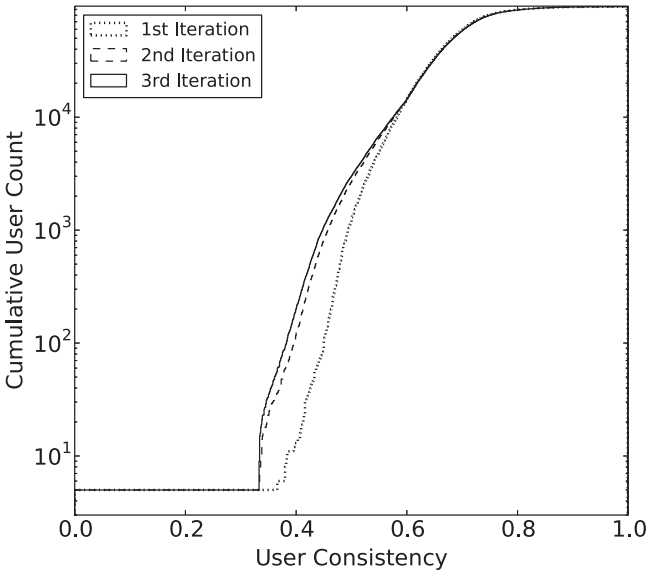

Figure 3. Distribution of user consistencies $\overline{\kappa}$ after 1 (dashed), 2 (dotted), and 3 (solid) iterations of the consistency-based weighting method (described in Section 3.4). Convergence of this method requires relatively few iterations: further iterations do not change significantly from the solid curve. Approximately 85 per cent of users have $\overline{\kappa} \geqslant 0.6$ and weights $w = 1$.

to consider the responses to the tasks above that question in this analysis.

For example, a study with the goal of examining spiral galaxies would ideally use a sample selected by considering the responses to task T12, "Is there any sign of a spiral arm pattern?" If a pure sample of galaxies with clear spiral arms is desired, a threshold may be selected at a high vote fraction for $f_{\rm spiral}$. If the threshold considers only this vote fraction, however, the final sample will likely be contaminated by galaxies where the spiral vote fraction is dominated by noise because only a small number of people reached that task (e.g., a warped edge-on disk).

In order to reach task T12, a user must give specific answers to the questions "Is the galaxy simply smooth and rounded, with no sign of a disk?" (T00), "Does the galaxy have a mostly clumpy appearance?" (T02), and "Could this be a disk viewed edge-on?" Each of these classifications should be considered in the context of the study's goals in order to select as pure a sample as possible whilst minimising contamination and bias.

If a moderately complete sample is desired, for example, the user could select thresholds for the selection such as $f_{\rm features} > 0.5$, $f_{\rm not\ clumpy} > 0.5$, $f_{\rm not\ edge-on} > 0.5$. Because most galaxies with these classifications will have received 80 classifications apiece (Section 3.3), these chained thresholds mean the minimum number of volunteers who will have answered the spiral question is $80 \times 0.5^3 = 10$. Higher thresholds will further restrict the minimum number of respondents to the deeper-branched question. If lower thresholds are desired, we recommend that the selection explicitly require a minimum number of respondents to each task.

There is no single set of thresholds that is ideal for all situations. However, in the data release accompanying this paper, we include "clean" selections of galaxies with different morphological features. These are detailed further below, but we additionally encourage users of this rich data set to experiment with different threshold/weight combinations in order to achieve their scientific goals.

3.6 Data release and "clean" samples

This paper includes the release of the raw and weighted classifications for each of the 49,555 subjects in the Galaxy Zoo CANDELS sample. In addition to each raw and weighted vote fraction for each task, we include the raw and weighted number of answers to each task, as well as for the whole galaxy overall. This combines for a total of 136 quantities for each subject, not including the subject ID or any other metadata. Clearly this is too much information to present a sample table in print here. However, the data is structured thusly for each task number NN with $i = 0$ to $n - 1$ responses:

`t[`NN`]_[quest_abbrev]_a[`i`]_[resp_abbrev]_frac` : the raw fraction of users who gave this response. `quest_abbrev` and `resp_abbrev` are abbreviated versions of the specific question and response, respectively.

`t[`NN`]_[quest_abbrev]_a[`i`]_[resp_abbrev]_weighted_frac` : the weighted fraction of users who gave this response.

`t[`NN`]_[quest_abbrev]_count` : the raw count of users who responded to this task.

`t[`NN`]_[quest_abbrev]_weight` : the weighted count of users who responded to this task.

For example, the information available for task T00, which has 3 responses, is structured as:

```
t00_smooth_or_featured_a0_smooth_frac
t00_smooth_or_featured_a1_features_frac
t00_smooth_or_featured_a2_star_or_artifact_frac
t00_smooth_or_featured_a0_smooth_weighted_frac
t00_smooth_or_featured_a1_features_weighted_frac
t00_smooth_or_featured_a2_star_or_artifact_weighted_frac
t00_smooth_or_featured_count
t00_smooth_or_featured_weight
```

The sum of raw `_frac` fractions adds to 1.0, as does the sum of `_weighted_frac` fractions. Multiplying the `_frac` values (raw fractions) by the `_count` (raw user counts) will retrieve the number of people who gave a specific response; likewise with weighted answer counts from `_weighted_frac` and `_weight`. As the user weighting described in Section 3.4 assigns a weight of $w \leqslant 1$ to each classifier, the weighted vote count must be less than or equal to the raw vote count. While the raw vote counts and fractions are provided for completeness, we recommend that users of this data set use the weighted fractions and counts.

In addition to the vote fractions for each subject, we provide a set of flags for each subject that indicates its member or non-member status in a "clean" sample of galaxies of a specific type. We select separate clean samples of smooth, featured, clumpy, edge-on, and spiral galaxies. These samples contain exemplars of each galaxy type with minimal contamination of the sample, and are correspondingly highly incomplete. They are selected according to the following thresholds using weighted vote fractions:

Smooth - Task T00: $f_{\rm smooth} > 0.9$, $f_{\rm star\ or\ artifact} < 0.2$

Featured - Task T00: $f_{\rm features} > 0.9$, $f_{\rm star\ or\ artifact} < 0.2$

Clumpy - Task T00: $f_{\rm smooth} > 0.9$, $f_{\rm star\ or\ artifact} < 0.2$; Task T02: $f_{\rm clumpy} > 0.8$

Edge-on - Task T00: $f_{\rm smooth} > 0.9$, $f_{\rm star\ or\ artifact} < 0.2$; Task T02: $f_{\rm not\ clumpy} > 0.5$; Task T09: $f_{\rm edge-on} > 0.7$

Spiral - Task T00: $f_{\rm smooth} > 0.9$, $f_{\rm star\ or\ artifact} < 0.2$,; Task T02: $f_{\rm not\ clumpy} > 0.5$; Task T09: $f_{\rm not\ edge-on} > 0.7$; Task T12: $f_{\rm spiral} > 0.8$

We provide these flags for the convenience of the end user, but we additionally encourage those wishing to use Galaxy Zoo classifications to investigate whether a different set of thresholds would be optimal for their own science case.

4 COMPARISON TO OTHER VISUAL CLASSIFICATIONS

Most of the galaxies in the CANDELS data set have additional visual classifications available in the form of expert classifications from astronomers and students who are members of the CANDELS team. Analysis of the full set of classifications in that separate project is still underway; the first release of classifications from the GOODS-South field is presented by Kartaltepe et al. (2014), hereafter K14, who also detail the project design and objectives, including the classification interface. Consensus classifications from the UDS field are also available (Kartaltepe et al., in preparation). For each galaxy in all fields, between 3 and 7 (typically 3) members of the CANDELS team provided classifications.

The classification scheme described in K14 is substantially different to that presented here. Firstly, while that project collects detailed classifications about a number of possible structural features (with 37 different responses possible), they do not always align precisely with the questions asked in Galaxy Zoo CANDELS. For example, the Main Morphology Class of K14 requires the user to select at least one option from among "Disk", "Spheroid", and "Peculiar/Irregular" galaxy types, along with options for "Point Source/Compact" and "Unclassifiable". The last of these is not an option Galaxy Zoo provides, and the first two are not necessarily the same as task T00's responses of "Features or Disk". While Galaxy Zoo does ask about bulges, it does so after multiple branches of the decision tree, and therefore this is not easily comparable to a 1st-tier question.

In fact, *all* responses collected by the CANDELS team interface are 1st-tier questions: the user is presented with all 37 options at once. Additionally, colour composites are not used in that project. Images from each ACS and WFC3 filter are presented separately within the interface, with an option for the user to specify when classifications differ significantly between filters. Users may also view the segmentation map in the $F160W$ band, and in the Perl/DS9 version of the CANDELS team interface the user may adjust the stretch of the image as well. These options are not available to Galaxy Zoo volunteers. On the other hand, the Galaxy Zoo decision tree asks multiple questions designed to elucidate the configuration of clumps in a galaxy, whereas the CANDELS team interface does not.

Despite these significant differences, it is nevertheless helpful to compare the CANDELS team classifications to the Galaxy Zoo CANDELS classifications. Figure 4 shows the comparison of vote fractions in four categories: Featured, Merger or Interaction, Edge-On, and Barred. For all comparisons below we have compared the subset of sources in CANDELS above the surface brightness limit described in Section **??** which have visual classifications from both teams, which have *not* been deemed "unclassifiable" by the CANDELS team, and which have *not* been rejected as stars or artifacts by more than 50% of classifiers for either project.

4.1 Featured Galaxies

We seek to compare the overall classification of CANDELS galaxies as "smooth" or "featured" between Galaxy Zoo and the CANDELS team. However, the team interface described in K14 does not specifically ask about this distinction. It does ask about disks and spheroids; however, equating "smooth" to "spheroid" and "featured" to "disk" requires assumptions about galaxies at $z > 1$ that we would prefer to avoid.

We therefore compare the "Features or Disk" vote fraction for T00 in Galaxy Zoo CANDELS to a combination of vote fractions from the CANDELS team classifications. We choose a set of structures that are unambigously inconsistent with a smooth light distribution in a galaxy, namely spiral arms, clumps, and bar features.

Within the CANDELS team classification interface, a user may indicate the presence of a bar or spiral arms by selecting one response for each, $f_{\rm bar,CT}$ or $f_{\rm spiral,CT}$. The clumpy classification, however, is actually a rating of both "clumpiness" and "patchiness", in a 3×3 grid from 0 to 2 along each axis. According to K14, "Clumps are concentrated independent knots of light while patches are more diffuse structures." Both are distinct from a smooth light distribution, so we include both in the creation of a "featured" vote for the CANDELS team classifications.

We follow an approach similar to **?** in combining clumpy classifications within this matrix of possible responses. Each vote is weighted by the strength of features that it indicates, by assigning a weight of 0.25 for each level along each axis. For a clumpiness rating i and a patchiness rating j, the weight for that vote fraction is

$$w_{ij} = 0.25(i + j).$$

For example, the weight for $C_1P_2 = 0.75$. As the maximum value within the selection matrix is C_2P_2, the maximum weight is 1. The overall clumpy vote fraction for a given object is then

$$f_{\rm clumpy,CT} = \sum_i \sum_j w_{ij} f_{ij}.$$

We note that users may make multiple selections within the clumpiness/patchiness rating matrix, so the weighted, summed vote fraction $f_{\rm clumpy}$ can in principle exceed 1.

Figure 4a shows the summed "featured" vote fraction for the CANDELS team, $f_{\rm bar,CT} + f_{\rm spiral,CT} + f_{\rm clumpy,CT}$, versus the Galaxy Zoo vote fraction for the response "Features or Disk" to task T00, for the 13,195 galaxies that have been classified by both and that meet the surface-brightness and other criteria described just prior to this subsection. The figure shows the 2-D histogram via hexagonal shading, indicating that in both projects a high vote for features of any kind is relatively rare (most galaxies have $f_{\rm features} \sim 0$ and $f_{\rm features,CT} \sim 0$).

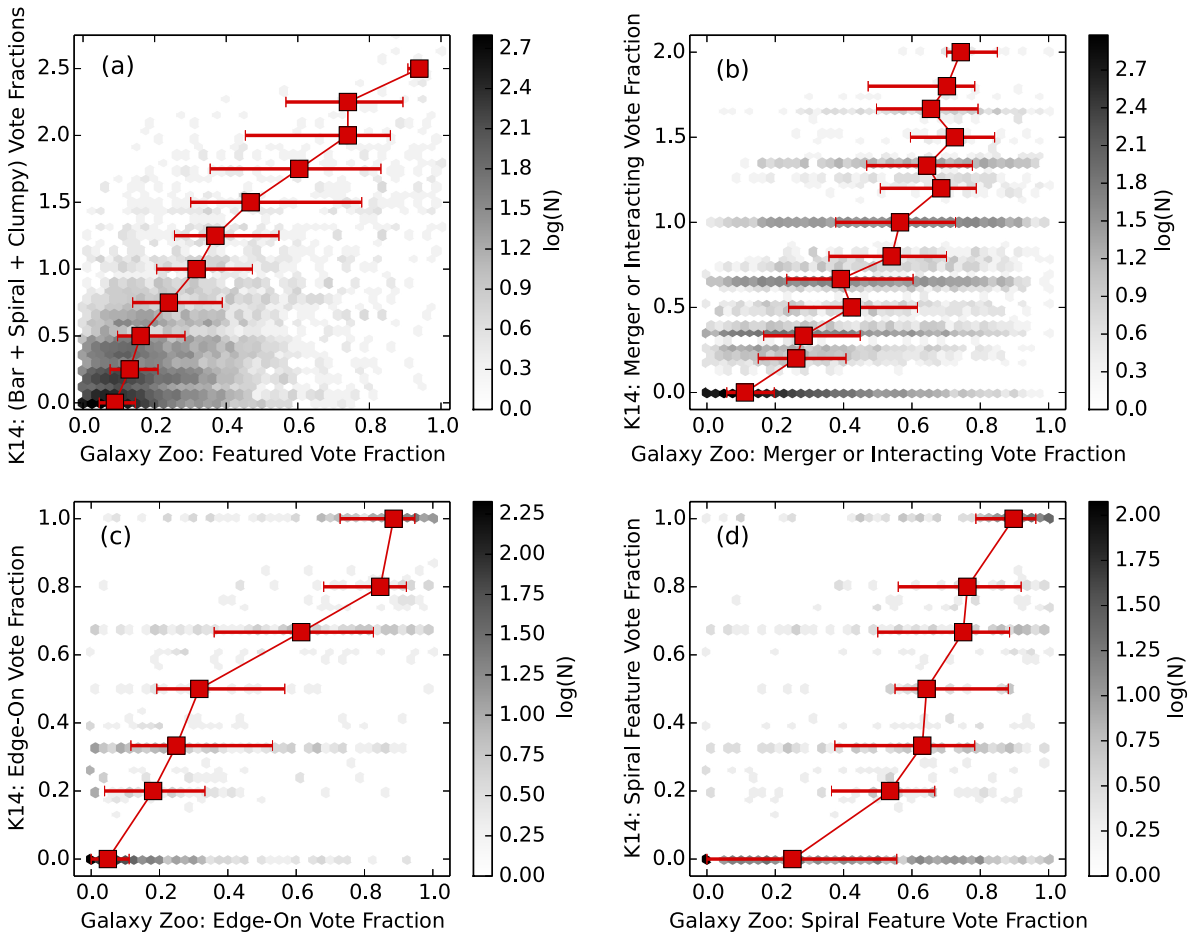

Figure 4. Comparison of Galaxy Zoo classifications with visual classifications from the CANDELS team (Kartaltepe et al. 2014, K14). The classification questions differ between the two projects, but we have selected 4 different classifications which are the most similar: (a) the sum of vote fractions in K14 for spiral, bar, and clumpy features, versus the Galaxy Zoo vote fraction for "Features or Disk" in task T00. Note that the sum of these vote fractions from K14 can add to > 1; (b) vote fractions for merger or interactions (task T16 in the Galaxy Zoo decision tree) for those subjects not identified as "Star or Artifact" in task T00; (c) vote fractions for the presence of an edge-on disk (task T09) for subjects that are neither artifacts nor predominantly smooth, nor dominated by clumps; and (d) vote fractions for the presence of spiral arms (task T12) for those subjects in panel (c) that are not edge-on. In all panels, the number of individual galaxies in a given hexagon in parameter space is shown by its shaded value. Red squares show the median Galaxy Zoo vote fraction and within a given K14 vote fraction bin; error bars show the interquartile region (from the 25th to 75th percentile). The different visual classification methods track each other well across many different kinds of structural classification.

The featured vote fractions track each other well, with a clear correlation shown by the red points indicating the median Galaxy Zoo vote fraction within bins of the CANDELS team vote fraction. There are virtually no galaxies for which the CANDELS team voted strongly for features being present but the Galaxy Zoo volunteers did not. It is also rare for the Galaxy Zoo volunteers to find a proportionally higher vote fraction for features than the CANDELS team, although the few examples seen in this parameter space may contain examples of distraction bias in a classification interface that presents dozens of choices simultaneously. This is clearly a small effect, however: on the whole the classifications agree very well with each other.

4.2 Merging or Interacting Galaxies

The last task (T16) in the Galaxy Zoo decision tree asks whether the user sees evidence of a merger, or of tidal interaction, or both, or neither. The CANDELS team Interaction Class asks the user to decide whether the galaxy is a merger, or whether there is interaction within the segmentation map, or outside of it, with an additional option for a non-interacting companion. There is also a separate flag within the CANDELS team classification to indicate whether a galaxy has tidal arms. Because these selections between projects are similar but not exactly the same, we choose to compare the sum of all signs of interaction of any kind within both projects. Specifically, we consider the sum of vote fractions within the CANDELS team classifications for "Merger", "Interaction within Segmap", "Interaction beyond Segmap", and "Tidal Arms", while for Galaxy Zoo we consider the sum of vote fractions for "Merger", "Tidal Interaction", or "Both". Given this selection, the maximum value for the combined CANDELS team vote is 2, whereas for Galaxy Zoo the maximum vote fraction is 1.

Figure 4b compares these fractions for each galaxy in

the same way as Figure 4a, with darker shaded bins representing a higher number of galaxies within that bin, and with red squares indicating the median Galaxy Zoo vote fraction within a given range of vote fractions for the CANDELS team. The striations seen in the figure reflect the finite number of possible vote fractions within the CANDELS team votes; this structure was not seen in Figure 4a due to the weighted combination of clumpy vote fractions.

The majority of images received 3 independent classifications each by the CANDELS team, and thus the majority of vote fractions fall within fractions of thirds. We therefore choose asymmetric bins within which to calculate median Galaxy Zoo vote fractions, with the aim of minimising the spread in source counts per bin: bins falling in increments of thirds include only sources with those exact CANDELS team vote fractions, and each bin between those values covers every other value. We adopt this binning strategy in Figures 4c and d as well.

Although Galaxy Zoo and the CANDELS team measure different aspects of mergers differently, in combination the merger/interaction vote fractions clearly correlate. As in the comparison between overall featured fractions, there are more examples where the Galaxy Zoo vote fraction is notably higher than the CANDELS team vote fraction than vice-versa. Examination of galaxies where Galaxy Zoo $f_{\text{merger or interaction}} > 0.5$ and CANDELS team $f_{\text{merger or interaction,CT}} = 0$ indicates some cases where a merger or tidal feature is clearly present, but others where it is less obvious whether a nearby companion is interacting.

Indeed, among this sample the CANDELS team vote fraction for "Non-Interacting Companion" is considerably higher on average than for the overall sample. This option is not explicitly available to Galaxy Zoo volunteers, although even moderately experienced users, particularly those who participate in discussions within the community Talk software, will in general select "Neither" if they decide the companion is not interacting. This explains why the number of galaxies showing this mismatch is much smaller (less than 2% of the sample) than the overall number of galaxies which CANDELS team classifications mark as having a non-interacting companion. Future analyses of mergers and interacting galaxies may find a combination of Galaxy Zoo and CANDELS team classifications useful for eliminating the effects of distraction bias and distinguishing between interacting and non-interacting companions.

4.3 Edge-On Galaxies

The branched nature of the Galaxy Zoo decision tree (Figure 1) means that, in general, a smaller number of people answer the question "Could this be a disk viewed edge-on?" than complete a classification

4.4 Spiral Galaxies

5 VISUAL CLASSIFICATIONS AND SERSIC INDICES

Distribution of Sersic indices for different galaxy types - I'm thinking of this as a result rather than a comparison, as they measure different things.

e.g. we can use this to talk about how many smooth galaxies are actually disks *assuming* n=1 is a disk

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

Bamford S. P. et al., 2009, MNRAS, 393, 1324
Bennett C. L. et al., 2013, ApJS, 208, 20
Cirasuolo M. et al., 2007, MNRAS, 380, 585
Darg D. W. et al., 2010a, MNRAS, 401, 1552
Darg D. W. et al., 2010b, MNRAS, 401, 1043
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
Hubble E. P., 1926, ApJ, 64, 321
Kartaltepe J. S. et al., 2014, ArXiv e-prints, 1401.2455
Koekemoer A. M. et al., 2011, ApJS, 197, 36
Kormendy J., Drory N., Bender R., Cornell M. E., 2010, ApJ, 723, 54
Lawrence A. et al., 2007, MNRAS, 379, 1599
Lintott C. et al., 2011, MNRAS, 410, 166
Lintott C. J. et al., 2008, MNRAS, 389, 1179
Martig M., Bournaud F., Croton D. J., Dekel A., Teyssier R., 2012, ApJ, 756, 26
Melvin T. et al., 2014, MNRAS
Scoville N. et al., 2007, ApJS, 172, 1
Sérsic J. L., 1968, Atlas de galaxias australes. Cordoba, Argentina: Observatorio Astronomico, 1968
Willett K. W. et al., 2013, MNRAS, 435, 2835