rijksuniversiteit groningen

Rijksuniversiteit Groningen

Bachelor's Thesis

Fluffy or bright? Using citizen science to find low-surface brightness and ultra-diffuse galaxies in the Fornax cluster

Author:
Christiaan Seerden

Supervisor:
Prof. dr. R. F. Peletier
Second supervisor:
T. Saifollahi

July, 2021

Abstract

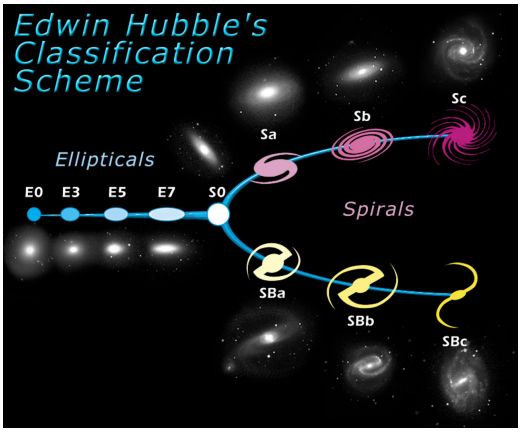
In this thesis, we analyze over 200,000 classifications contributed by over 3,000 individual volunteer citizen scientists, of over 6,000 images from the Fornax Deep Survey, in an attempt to discover low-surface brightness galaxies in this cluster not found previously, or to confirm those found in previously created catalogues. We find that the users correctly classify at least 30 (and up to 48 depending on strictness of classification criteria) of the 232 low-surface brightness galaxies from a forthcoming catalogue of Fornax low-surface brightness galaxies. By performing our own visual classification, we find an additional 63 objects classified by the users as so-called 'fluffy' galaxies, that were excluded from this forthcoming catalogue, but that from visual inspection do not show enough structure to be excluded from a catalogue of Fornax low-surface brightness galaxies. We find that objects with effective mean r-band surface brightness exceeding 25 mag/arcsec^2 are never identified by the volunteers as galaxies, even though a significant fraction of low-surface brightness galaxies are at least this faint, or fainter, placing a limit on the viability of using user-contributed classifications in identifying low-surface brightness galaxies like those in Fornax. We believe that the accuracy of users could be increased through various improvements to the tutorial provided to them before they make their first classification.

Contents

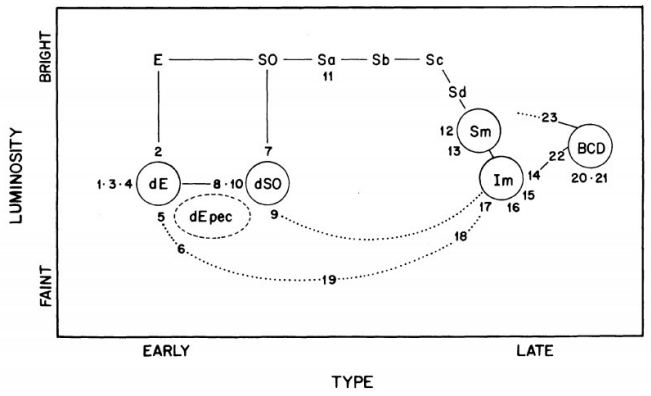
1 Introduction

Our universe contains some few hundred billion galaxies, each containing hundreds of millions to hundreds of trillions of stars. By virtue of them containing such huge quantities of *stuff*, galaxies are extremely interesting objects to study. Not all galaxies, however, are easy to find, which naturally makes them harder to study. Faint, extended galaxies, however, are still galaxies, and thus our human inquisitiveness leads us ever onward in our search of finding as many of these objects as we can. Much has been discovered about galaxies since we first aimed our sights at the cosmos. Research into their makeup, formation, and evolution are all active fields of study. In this thesis, we will aim to aid in the search for more of the small, faint galaxies that we call low-surface brightness galaxies, in the galactically nearby Fornax galaxy cluster. These galaxies are so faint that even our computer algorithms still have trouble identifying them sometimes.

1.1 Low Surface Brightness dwarfs and Ultra-Diffuse Galaxies



(a) Edwin Hubble's galaxy classification schema. Credit: ESA/Hubble



(b) Dwarf galaxy classification from Binggeli et al. (fig. 1, [15]). Giant galaxy classes are also shown along the 'bright' part of the vertical axis.

Figure 1

By number, dwarf galaxies are the dominant galaxy population in the Universe [18]. As their name suggests, we understand dwarf galaxies to be small ($R_{eff} \leq 1.5\,\mathrm{kpc}$), and like non-dwarf galaxies, we can readily morphologically categorize them into various subclasses (see figure 1a). Though generally a significant fraction (approximately 60%) of the galaxies in our Universe are spirals, for dwarfs it is quite the opposite: most dwarf galaxies are ellipticals, spheroidals and irregulars [16] (irregulars are essentially the dwarf version of spiral galaxies). The dwarf spiral 'void' is seen in figure 1b, where we see that there is no direct companion between Sa and Sb galaxies in the 'bright' part, and dwarf types in the 'faint' part in the image.

Current cosmological models tell us that smaller systems often form the building blocks of larger ones (e.g. [24, 5]). It is for reasons such as this that we may take a particular interest in the study of small galaxies like dwarfs.

In this thesis, we will concern ourselves with low-surface brightness (LSB) (dwarf) galaxies, and we will also consider the slightly larger ultra-diffuse galaxies (UDGs). The distinction between these two types (LSB dwarfs and UDGs) is based on effective radius: LSB dwarfs will extend up to 1.5 kpc, and for UDGs we consider $1.5\,\mathrm{kpc} < R_{eff} < 10\,\mathrm{kpc}$ [26]. Note that we consider the surface brightness regime of approximately $\mu_{e,r} \gtrsim 23$ mag arcsec^{-2}. [12] to be the 'low surface brightness' regime, where $\mu_{e,r}$ denotes the effective mean r-band surface brightness of an object.

Low surface brightness galaxies contain considerable amounts of dark matter in their halos, especially compared to giant galaxies. However, LSB galaxy formation is still poorly understood (see e.g. Martin et al. [11], or Bhattacharjee et al. [2]). Detection of these galaxies for further study, thus, is of vital importance. The more of these galaxies we find, the better our understanding of them can become.

1.1.1 *The Fornax Cluster*

In particular, this project concerns a (likely, see section 1.3) selection of low-surface brightness galaxies from the Fornax cluster. The Fornax cluster is the second-richest galaxy cluster (after Virgo) within 20 Mpc of the Galaxy with a virial mass of $7 \times 10^{13} M_{\odot}$, making it a great site for the study of galactic evolution and galactic dynamics [8]. In the past decade, the Fornax Deep Survey [13] has been created through observations with OmegaCAM on the VST. This survey goes 3 magnitudes deeper than the complete Fornax Cluster Catalogue from 1989, making it the deepest survey of this cluster to date. This improved depth allows for better study of exactly the objects we're looking for - the small, faint galaxies that might have previously escaped the view of then-state-of-the-art optics. The size and improvement of quality of this data compared to previous surveys allows us to study in more detail the properties and history of many objects in this cluster (see e.g. [14, 19]).

1.2 Citizen science and the Zooniverse

In citizen science projects, volunteers, who are generally untrained in the field before coming to a project, are asked to conduct science on platforms like Zooniverse or SciHub, often by means of visual identification (like in the project this thesis analyses), usually after following a short tutorial put together by the project managers that aims to guide users in the right direction. Classification projects may concern a wide variety of subjects, from animal species to astronomical objects. Involving a large group of volunteers in a project may either aid (or in cases possibly even replace entirely) work otherwise done by computers in cases where there is simply too much data to look at for a small group of experts, or in cases where an untrained eye can, with a slight nudge in the right direction in the form of a short tutorial, match up against the opinion of an expert. Volunteer contributions may also serve to identify objects that warrant a further manual inspection, which is also a way to reduce the size of the extremely large datasets and surveys that are produced nowadays.

In the past decade, significant attention has been drawn to the use of citizen science in the field of astronomy, on the web most notably with the Galaxy Zoo project [27], in which 10^5 participants made more than 4×10^7 morphological classifications of galaxies [10]. In a similar successor project called Galaxy Zoo Supernovae [17], volunteers were shown images of potential supernovae, and through a series of questions asked of each volunteer, a score would be assigned to the object in question based on how astrophysically interesting it was perceived to be (as e.g. a transient or supernova). Classifications and transcriptions gathered from citizen science projects are leading to the expansion of knowledge in a large variety of scientific disciplines (see for example the list of Zooniverse-related scientific publications at [28]).

1.3 Space Fluff

This thesis focuses on the analysis one particular Zooniverse project: Space Fluff, executed by Anna Lanteri with the Sundial international training network [20]. The following description is taken from the Space Fluff page on Zooniverse:

> The better our telescopes get, the fainter the galaxies we see with them can be. Recently a new population of faint objects has attracted the attention of the scientific community: we need your help to make sense of what we see!
>
> Since it is a challenging field due to the nature of the objects, we want to be prepared for the future big surveys. We have built this project with a medium sized dataset from the Fornax Deep Survey to classify these objects with citizen science and to study how this classification can compare with the traditional one. This will allow us to be ready when it is time to add a way bigger set from the KIDS survey of the same objects. Stay tuned for that! [21]

Space Fluff aims to find what the project describes as 'fluffy' galaxies in the Fornax cluster, referring to the UDGs and LSB (dwarf) galaxies described in the introduction to this work, by visually distinguishing between this type of galaxy and other galaxies and non-galaxy objects that appear in the images, but are for example actually far behind the Fornax cluster. It is hard for algorithms to distinguish between a high-surface brightness background galaxy, and a low-surface brightness cluster member (see also section 1.5), and hence Space Fluff was created to explore the effectiveness of the possibility of using human classification instead. Aside from the potential scientific benefit, Space Fluff also serves as a public outreach project.

The ultimate goal is to create a complete catalogue of LSB galaxies in the Fornax cluster, since naturally, if we are to study these objects as a group, we would first like to know where they are.

In the Space Fluff project, images of a total of 6362 possible LSB dwarf/UDG candidates were shown to a total of 3700 unique users (registered or unregistered), 2136 of them with accounts registered on the Zooniverse website, the rest of them as "not-logged-in" users. These users made a total of 233,375 classifications (from late October 2020 to mid-January 2021) across three so-called workflows: 'on-the-go', 'classify' and 'hardcore'. So as no to waste volunteers' time with objects that are easily seen not to be part of the Fornax cluster, or even not galaxies at all, the objects served in the project were run through the MTObjects source extraction algorithm [22] to provide a reasonable initial guess that the objects shown may in fact be LSB galaxies in the Fornax cluster.

In the main 'Classify!' workflow, users are shown an image of one of the objects, following a short initial training session (see the Appendix at the end of this work for a short overview of the Space Fluff training/field guide), and tasked with answering the following question: "Look at the very center of the image: do you see a single galaxy or a group of far away objects?". The user is invited to provide one of the following answers: 'Galaxy', 'Group of objects (Cluster)' or 'Something else/empty center'. If the user judges the center of the image to contain a galaxy, they will be asked whether they believe it to be 'fluffy' or 'bright' (see an excerpt of the Space Fluff field guide in the Appendix). This distinction forms the essence of a classification. An object may be a galaxy that is beyond the Fornax cluster that made it into the image simply by being in the line of sight, or it may be a low-surface brightness galaxy in the Fornax cluster. Using a physical argument resulting from data about previously confirmed cluster members (relations between color, magnitude, surface brightness and *concentration* (see section 1.6)), we can distinguish many background galaxies from cluster members. Objects that the users are guided to mark as 'bright', are supposed background galaxies, whereas 'fluffy' galaxies are more likely to be LSBs/UDGs in the Fornax cluster.

If a user thinks the see a group of objects in the center of the image, rather than a single galaxy (or even nothing), they won't be asked any further questions. The hardcore workflow goes deeper into the properties of the object in the image, about its color, or whether the user believes the object to have a nucleus or not. The workflow diagram in figure 27 (see the Appendix) is taken from the 'about' page of the Space Fluff project on Zooniverse, and describes the flow of questions presented to a user based on their choices throughout a classification.

The goal of this thesis is to analyze the answers provided by these volunteer users, and determine whether they can, with any accuracy, correctly identify LSB galaxies and UDGs purely based on a visual inspection of an image. In the process of our analysis, we will examine the behavior of the users in the project and attempt to find any correlation between the photometric properties of these objects and the resulting consensus among the votes of the volunteers on the identity of these 6362 objects.

1.4 Analyzing citizen science data

Aceves-Bueno et al. [1] draw a comparison between various methods used to analyze citizen science projects. The most common, and most intuitive, comparison method is percent agreement between a ground truth dataset (verified data, or analysis performed by experts) and the volunteer consensus. The main drawback of this method is a lack of framework that this presents for hypothesis testing. Another drawback is the fact that this methodology fails to account for agreement by chance. Other methods used in citizen science analysis are statistical methods like T tests, various correlations (Spearman's, Pearson's), chi-square tests and linear regressions. All of these methods require some sort of ground truth labeled into categories for complete comparison. Still, despite these drawbacks, it is intuitive to quantify a user consensus in terms of percentages (a majority is always a majority, so 50% agreement is in every case a good initial number at which to draw a line). Lintott et al. use percentage cutoffs of 80% and 95% to define 'clean' and 'superclean' samples of classifications in [10] in the first Galaxy Zoo project, however they refer to work by Darg et al. [3], who present a case where only 40 percent agreement is needed to correctly classify a sample of galaxy mergers, thereby leading to the conclusion that using a single pre-defined classification threshold is not always the best choice.

In our analysis of Space Fluff data, we will compare the classifications done by volunteers to a likely ground truth (LGT) catalogue (see section 1.6), containing a number of galaxies identified by Venhola

et al. (*in prep.*) as low-surface brightness galaxies in the Fornax cluster. However, the project does not contain a set of expert classifications on all of the objects, meaning a complete comparison between volunteers and experts cannot be done in this case. The most suitable method for analysis then becomes the simplest: percentage agreement. In order to determine the accuracy of volunteers in classifying Fornax LSB galaxies, we will be comparing the agreement among users for each task presented in the project, and comparing various levels of consensus to the likely ground truth. We will use 50% and 75% classification thresholds throughout this work, but will also come to find an argument for the usage of a 90% threshold in a specific situation later in this work.

A final question is that of user expertise. How does one weigh the level of skill of an individual user? Weighting each user's classifications can be done, and has been explored, e.g. by Lintott et al. and Darg et al. Lintott et al. find that, in their case of Galaxy Zoo, weighting users according to their agreement with the majority does not produce a significant change in classification outcome. A pitfall of upweighting a user who often agrees with the majority is the fact that this assumes that the majority is always accurate, which will always be main research question for a citizen science project, and thus cannot simply be assumed.

When discussing initial results from the Planet Hunters project [4], Fischer et al. note they aim to determine individual users' capibilities by inserting artificially generated light curves of planet transits and analyzing user performance for those curves. This is essentially a form of comparison against ground truth, as in this case data is generated for which the scientists know with certainty what they want the classification outcome to be.

For the bulk of our analysis, we will not perform weighting of individuals' classifications, but we will at a later stage investigate what changes in our results if we only consider classifications made by the most active subset of users, and also what happens if we discard the first few classifications made by each user. The latter follows from the assumption that users learn as they gain experience in a project.

1.5 Astronomical object detection: human vs. machine

In recent years, technology has drastically improved astronomical object detection, or "source extraction". Combined with the need for fast, large-scale detection engendered by the ever-increasing size of astronomical surveys, automated detection methods may outscale what is possible by humans. Space Fluff makes use of classification done by humans, instead of automated methods. How does the feasibility of the two different methodologies compare? In this section we will briefly go over various methods of algorithmic astronomical object detection, making note of strengths and weaknesses. Our analysis of Space Fluff in later sections will then reveal whether or not manual classification stacks up against algorithmic detection.

In what follows, we will very briefly compare a number of source extraction algorithms (SourceExtractor (also "SExtractor", or "SE"), NoiseChisel, ProFound and Max-Tree Objects (also "MTO" or "MTObjects")), following a comparison between these algorithms done by Haigh et al. [6]. Generally, source extraction methods all follow the same steps (see sec. 2.1 of [6]): (1) identify and measure background light, (2) threshold the image relative to the background, (3) locate sources exceeding the threshold level, (4) produce a catalogue of sources and their measured properties. Implementation details of these steps vary across algorithms of course.

Most relevant to our work is Max-Tree Objects (MTO). This is the algorithm that was used by Venhola et al. to extract the same set of over 6,000 objects that were presented to users in Space Fluff. MTO uses tree-based morphological operators: the leaves in a tree represent local maxima (the brightest pixels in some region), the nodes represent connecting areas of the image, and the root represents the entire image. Compare this to SExtractor or ProFound, where instead of trees, various levels of thresholds are used to distinguish objects from background pixels. SExtractor uses multiple backgrounds, spaced in exponential steps, and ProFound only uses one threshold level initially. Other programs use dendrograms: hierarchical representations of images, with nodes representing connections between local maxima, which is a more refined method than thresholding, where you end up with only a single unbroken object instead of a series of nodes. The complexity of implementation exceeds the level of this work, so we refer to [6] for a more complete description.

Comparing algorithms can be done for example by inserting a number of simulated objects into an

image, and analyzing how many of those objects are properly identified by the algorithm. Haigh et al. find that MTO sometimes finds diffuse regions in images and allocates them to the wrong object, meaning a manual classification must still be done to identify cases where this occurred. They also find that SExtractor and ProFound are incapable of detecting the outskirts of objects accurately. This may present a problem in the case of low-surface brightness galaxies, where objects are generally extremely faint to begin with, and thus often no bright areas exist in the images for these programs to identify as objects. According to Haigh et al., NoiseChisel and MTO were more accurate in finding these faint objects, but still experienced difficulty in identifying the edges of objects, and also in separating similar nested objects. Finally, Haigh et al. state that MTO consistently statistically outperformed other methods when testing simulated data.

Considering the performance of MTO on faint objects, we can state that MTO is a good choice for a source extraction algorithm for the low-surface brightness galaxies we consider in Space Fluff. The question remains, how does this algorithm stack up against the manual classification of the Space Fluff volunteers? This is the question we will explore throughout the remainder of this work.

1.6 Ground truth, selection cuts and object properties

If and when Space Fluff users reach a consensus on the identity of an object, how can we subsequently judge the accuracy of this consensus? We have no information on the level of experience of expertise of an individual user, nor of the group of volunteers as a whole. To analyse to any extent the accuracy of the user consensus, we will be comparing user classifications to a catalogue of 265 Fornax cluster UDGs and LSB dwarfs produced by Venhola et al. (*unpublished, in preparation*) resulting from work following [26, 25]. Throughout our analysis, we will in places refer to this catalogue in terms of "likely ground truth objects", "the (likely) ground truth catalogue", or "the LGT catalogue". Note that we emphasize *likely* ground truth when referring to this catalogue. This is because one of the aims of this work is to compare the performance of humans to that of machine algorithms, when it comes to classifying faint objects that may not have clearly defined edges in images. A possible result of our analysis may be that the human classifications are strictly better than that as performed by algorithms, in which case a catalogue resulting from human classifications would become the ground truth, instead.

The next subsection describes the procedure used by Venhola et al. to separate likely non-cluster member galaxies from the likely cluster members. We note that the catalogue produced by this procedure, the FDSDC [25], is a Fornax dwarf catalogue, and not necessarily a LSB galaxy/UDG catalogue. However, the selection procedure serves mainly to filter out background galaxies (galaxies not in the Fornax cluster at all), and thus is still relevant to our situation. A final note is that Venhola et al. only consider candidates with a semi-major axis of at least 2 arcsec for this dwarf catalogue.

1.6.1 *Selection cuts*

As mentioned above, Venhola et al. [25] perform so-called 'selection cuts' on Fornax dwarf candidates to filter out likely non-cluster members. By using a physical argument, objects can be excluded from cluster membership on the basis of color (the filters used for this are SDSS filters r', g', i', and the color selection cut uses the $g' - r'$ color), surface brightness and concentration. Concentration in this case is defined as follows:

$$C = 5 \times \log \frac{R_{80\%}}{R_{20\%}}$$

where $R_{80\%}$ and $R_{20\%}$ denote the radii enclosing that percentage of the galaxy's light.

Using known values of the above-mentioned color, surface brightness and concentation of spectroscopically confirmed Fornax cluster members, Venhola et al. filter out unlikely Fornax cluster member dwarf galaxies. The argument for being able to use these parametric selection cuts is based on three assumptions, derived by Venhola et al. from comparison to a sample of galaxies with known redshifts: (1) cluster galaxies become bluer with decreasing luminosity, (2) the surface brightness of cluster galaxies decreases with decreasing luminosity, (3) faint cluster galaxies are less concentrated (referring to the concentration parameter C) than background galaxies.

By excluding objects that do not adhere to these relations, the majority of background galaxies can be identified and thus excluded from cluster membership. Specifically, Venhola et al. perform three selection

cuts:

1. *Color cut*: Objects at least 0.15 magnitudes redder than the brightest spectroscopically confirmed galaxies in the cluster are filtered out. The objects Venhola et al. filter have $g' - r' > 0.95$ mag. The $g' - i'$ color is similar. Objects with $g' - i' > 1.35$ are also filtered out.

2. *Surface brightness cut*: Perform a linear fit on confirmed cluster galaxies in the magnitude-surface brightness space, and exclude candidate galaxies whose surface brightness is at least three standard deviations above the fit.

3. *Concentration cut*: Fit the magnitude-C relation for cluster galaxies with $r' < 16$ mag. Exclude objects that are 2 standard deviations above this fit, and have $C > 3.5$. For Space Fluff objects, the magnitude-C relation for $r' < 16$ mag is not relevant, since none of the Space Fluff objects reside in this r' magnitude regime.

Though these selection cuts already exclude a large number of cluster candidates, Venhola et al. proceed to perform a further selection based on visual classification, after which only 577 of the 1497 candidates that survive their selection cuts remain. This manual classification might be a matter of experience and expertise, and thus is hard to quantify for reproduction (which is in part the reason the Space Fluff project was created to begin with). In section 3.3, we will reproduce the selection cuts on Space Fluff objects in an effort to filter out a fraction of the likely background galaxies that might have been classified by users as fluffy galaxies. We will also perform a manual visual classification of objects that survive selection cuts, and are classified by users as fluffy galaxies, to compare our intuition against that of Venhola et al., and to possibly identify a number of galaxies that may or may not still be good candidates for catalogue inclusion.

1.6.2 *Object properties*

In addition to the 265 objects in the likely ground truth catalogue, an unfiltered catalogue (Venhola et al., *unpublished*) of all candidates considered for cluster membership in the creation of the likely ground truth catalogue was made available to us. This catalogue contains photometric properties and a few other parameters that we can use in our analysis of the Space Fluff project. These properties result from GALFIT models of the objects, applied to images from the FDS. Whenever we refer to any photometric properties, or parameters like surface brightness and concentration, we will mean the properties as provided to us through this catalogue. The surface brightness we use,$\mu_{e,r}$, is the mean effective surface brightness in the r'-band (with the dimension of magnitudes per square arcsecond).

Of the 6036 UDG candidates in this unfiltered catalogue, 5440 are present in the Space Fluff project as candidates for identification. The Space Fluff images are also selected from the FDS. The objects that are not listed in this unfiltered set of 6036 candidates are expected to be from images containing artifacts (which makes parameter extraction tricky), or objects otherwise not chosen for classification, presumably on the basis of them obviously not being cluster members in some way or another. Like for all objects that do have properties in the catalogue, we will analyse user classifications for these objects also.

2 Extracting and parsing the classification data

2.1 Space Fluff on Zooniverse

Figure 27 in the Appendix contains a flow diagram of the whole project as it was presented to users on the Zooniverse page of Space Fluff. For completeness and clarity, we will shortly state the makeup of every workflow in tables 1 and 2. Since these are the questions ultimately actually shown to users in the final version of the project, these tables are perhaps more useful than the workflow diagram, which also outlines Sundial's thoughts on the relevance and possible interpretation of answers in each task.

Task	Question
T0	Look at the very center of the image: do you see a single galaxy or a group of far away objects?
T1	Is the galaxy fluffy or is it bright?
T2	What color is the galaxy?
T3	Does the galaxy have a visible core?
T4	What shape is the galaxy?
T5	How would you describe the texture of the galaxy?
T9	Our bad! what do you see instead?

Table 1: Reference of task indices and corresponding questions. We will often refer to tasks by their identifier (e.g. 'T0' or 'task 0') throughout this work for the sake brevity.

Task	Unique answers
T0	Galaxy Group of objects (Cluster) Something else/empty center
T1	Fluffy Bright
T2	Impossible to say White/blue Red/yellow
T3	No/Unsure Yes, a bright point Yes, a bulge
T4	Distorted/disturbed Elliptical Round
T5	Smooth and fuzzy Smooth and dense Clumpy and/or featured
T9	Something else Looks like a small star Scattered light Nothing: background too bright or galaxy too faint

Table 2: Reference of task indices and corresponding unique answers. Users were only able to select from these predefined answers when classifying an object.

As mentioned previously, the first task presented to each user, which we will hereafter refer to as *T0* or *task 0*, is presented regardless of the workflow chosen by the user. In the "On-the-Go" workflow, T0 is the only question presented to a user. The second task, T1, is presented to users in either the "Classify!" or "Hardcore" workflow, provided that they answer T0 with 'Galaxy'. In the "Classify!" workflow, only T0 and T1 are present. In the Hardcore workflow, tasks 2 through 5 are also presented in this case. Finally, T9 is only presented in the Hardcore flow, and only after a user answers 'Something else/empty center' for T0.

2.2 Parsing Space Fluff classifications

2.2.1 *Initial extraction*

The classification phase has been successfully executed prior to this thesis, leading to some few hundred thousand total classifications of a total of over six thousand images. The analysis in this work builds, in part, off the data files and code generated by Anna Lanteri, available on GitHub [9] under an MIT license.

With the completion of the process of gathering volunteer classifications, it is time for data analysis. The output format of the project's data is a number of files with comma-separated values (CSV), in which every classification is given its own row, and each of the three workflows is distinctly labeled, so we may individually assess each workflow, or combine overlapping tasks (to reiterate: in Space Fluff, only task 0 is present in each workflow, and T1 is present in both "Classify!" and "Hardcore". Remaining questions are only present in "Hardcore").

To prepare these classifications for analysis, we must first extract the data into a usable format. The numerical analysis in this work was all conducted in the Python programming language, using the Pandas [23] and NumPy [7] frameworks for the processing and manipulation of the raw data into usable so-called *dataframes* (in the Pandas documentation, a dataframe is described as "Two-dimensional, size-mutable, potentially heterogeneous tabular data."[1]).

The initial data parsing process concerns the following matters:

1. Not all classifications were properly completed. Users may decide not to complete any task by simply leaving the web page. This results in a reduction of the total number of usable classifications. If, for example, in 'task 0' in the *Classify!* workflow, a user answers that they see a Galaxy, but then they don't follow up by providing an answer for whether they think the galaxy is 'fluffy' or 'bright' (by clicking off the webpage, for example), it will have been registered in the raw dataset, but we discard the classification for the purposes of this analysis. Note, however, that in the Hardcore workflow, due to the nature of some of the questions, users were allowed to opt out of completing a task regarding an object (excluding the initial task, which always had to be answered).

2. Some classifications were done in alpha and beta stages of the project. We discard these to exclude any classifications made potentially only to test the system.

3. Users may have been presented objects they had already seen, either in the same workflow, or in another workflow (a few hundred users did classifications in multiple workflows). We exclude any classifications of an object made by a single user after the first classification they made of that object. Of the 233,375 total classifications across the three workflows, 10,316 classifications, performed by 233 users, are filtered out due to this criterion.

In the CSV files generated, a single 'annotations' column per row contains the name of each task completed in that classification, along with the answer provided by the user for these tasks. For ease of analysis, we cast the answer given by the user for every task to a new dataframe column. This allows for easy grouping using the programming libraries we've chosen. Note that the (Python) programming code pertaining to the analysis of this data is available in the Appendix and on Github[2].

Figure 2 shows distributions of the number of remaining valid classifications, after the filtering described above. We also include the distribution after combining all three workflows. In the execution of the project, the order in which objects were served to the user was such that every object would receive a statistically significant amount of votes in each workflow (e.g. in "Classify!", this target (also called an object's 'retirement limit') was 15 classifications, and in "Hardcore" it was 10 classifications). Combining the three workflows is a good way to increase statistical significance, since we will in all cases end up with an equal or larger amount of classifications per object than if we were to consider only a single workflow.

2.2.2 *Combining workflows*

Something to note for later stages of our analysis is how we will combine results from each of the three different workflows in our final classification of each object: since the "On-the-Go workflow" only asks

[1] https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.html
[2] https://github.com/Seerden/SpaceFluff

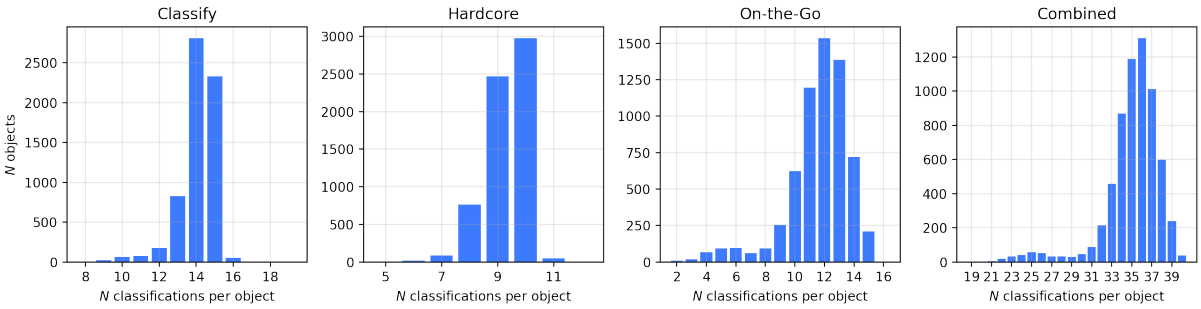


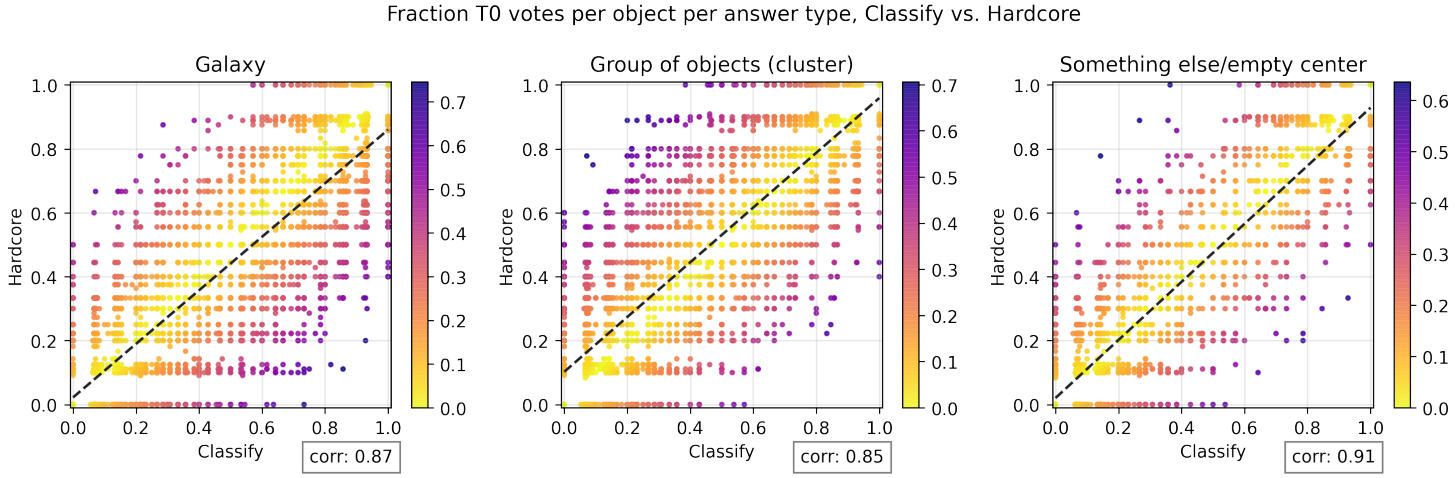
Figure 2: Votes per object for each of the three Space Fluff workflows, and the combined dataframe containing all workflows, after filtering. Note that values on each y-axis are not normalized.

users whether an object is a galaxy or not, its use is limited for our analysis. Later in our analysis, we will at times restrict ourselves to examining objects for which the user consensus for T0 is for example 'galaxy' or 'something else/empty center'. When examining these consensus, we will combine classifications from all three workflows to arrive at an answer, as there is no reason to distinguish between a 'galaxy' vote in the "On-the-Go" workflow from a 'galaxy' vote in the "Hardcore" workflow.

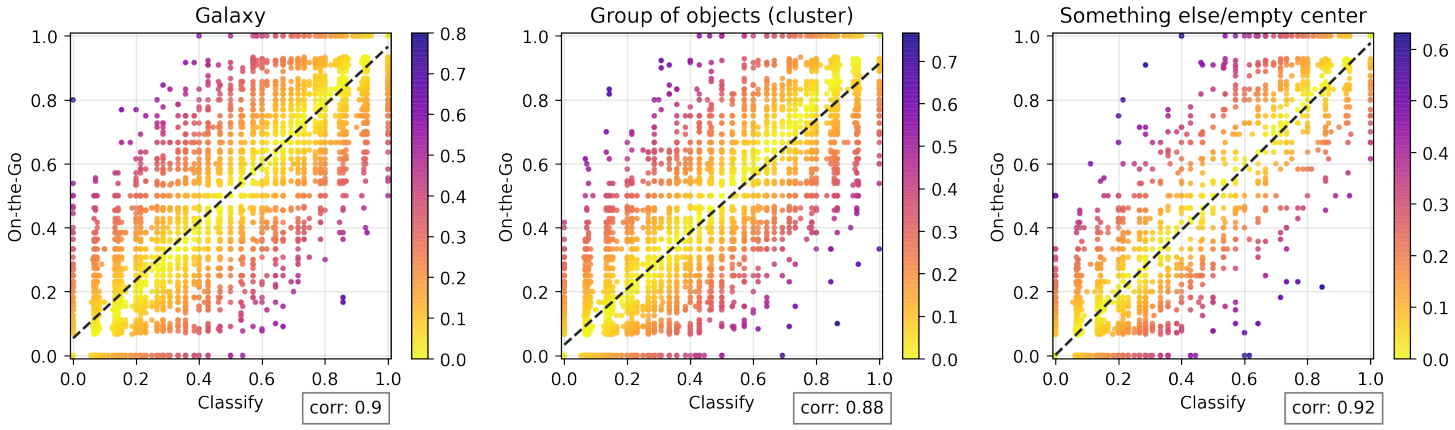
Figure 3 (see next page) shows a rudimentary comparison between the votes an object gets in task 0 in each workflow (recall that task 0 is the only task that exists in every workflow). We compare each workflow to each other workflow, and since there are only three possible answers for this task, we show the comparison for every unique answer type in a separate plot for completeness. The strong correlation between answers across workflows that we observe in this figure affirms the feasibility of simply combining all T0 answers for every object rather than, for example, employing some kind of weighting scheme. The correlation coefficient we use in this figure, and also at other points in this work, is the Pearson correlation coefficient, which is a measure of linear correlation between datasets. It can be computed by dividing the covariance of variables by the product of their standard deviations, so for two variables, X and Y, we would use the following;

$$\rho_{X,Y} = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y}$$

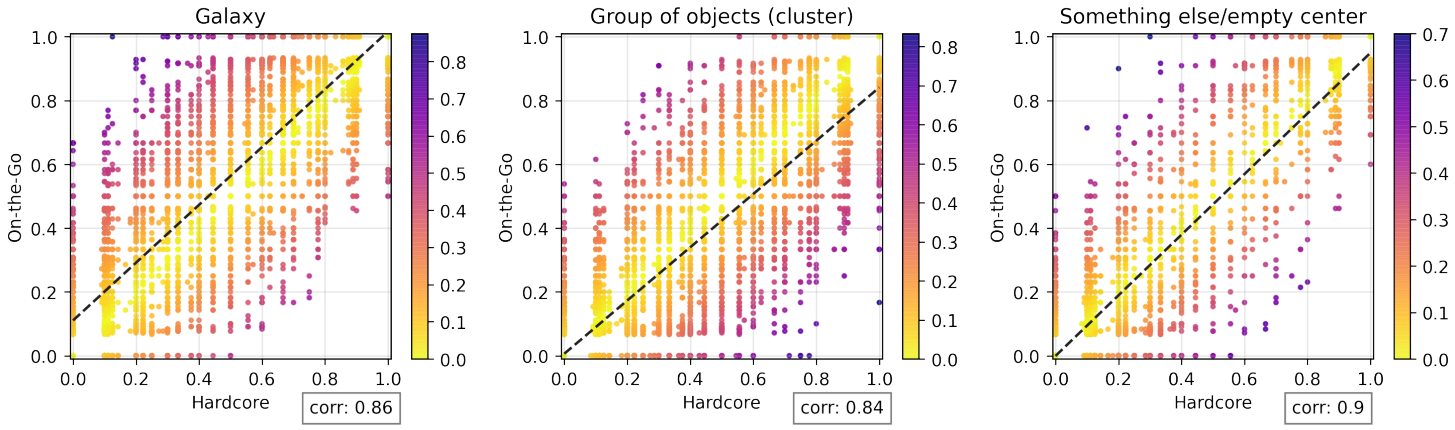
The correlation coefficient is in reality a matrix, but since the diagonal terms describe the relation between a variable and itself, the correlation coefficient for these will be 1. When we mention the correlation coefficient in this work, we describe the coefficient for two different variables. In the case of figure 3, it would be the correlation between the percentage of 'galaxy' votes an object receives in one workflow, and the percentage in another workflow.



(a) Classify vs. Hardcore



(b) Classify vs. On-the-Go



(c) Hardcore vs. On-the-Go

Figure 3: Comparison of T0 votes per object between each combination of workflows. The *corr* annotation denotes the Pearson correlation coefficient (see text). The color mapping indicates the absolute value of the difference between the x- and y-values of every point, and the dashed line on every plot indicates a simple linear fit.

2.2.3 *Object parameters and photometric properties*

As discussed in section 1.6, we have access to a catalogue/dataset containing a number of parameters for the majority of the objects in the Space Fluff project. The names and coordinates of the objects in this catalogue match those used in the Space Fluff project, so we can easily assign these photometric properties to the objects in our dataframes. As a sanity check, we also matched the coordinates of the objects between Space Fluff and the LGT catalogue dataset to ensure no anomalous data was present (e.g. an object with the same name in the two data sets having completely different coordinates).

At various points in our analysis, it is useful for visualisation purposes to include images of the objects we are discussing as they were presented to users in Space Fluff. This does not prove quantitative data, but gives us a clearer picture of the overall situation.

2.3 User engagement

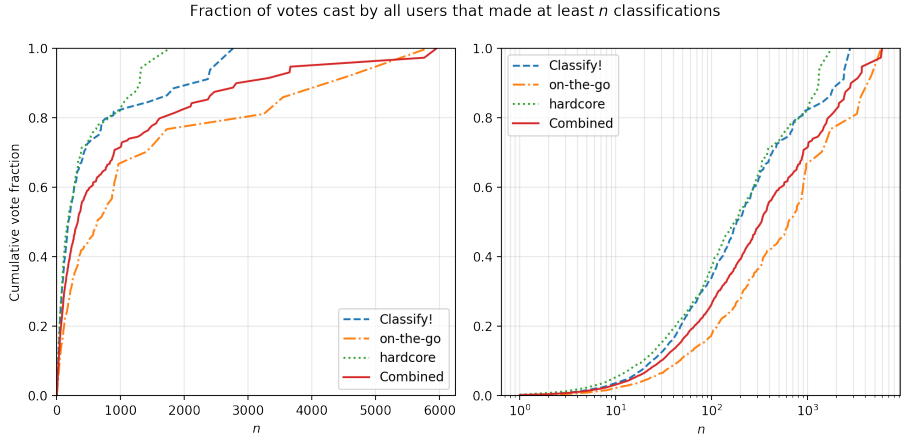


Figure 4: Cumulative vote fraction of each workflow as a function of users that made at least n classifications, in linear and in log scale. The y-axis denotes the fraction of total votes cast by all the users that performed less than or equal to n classifications.

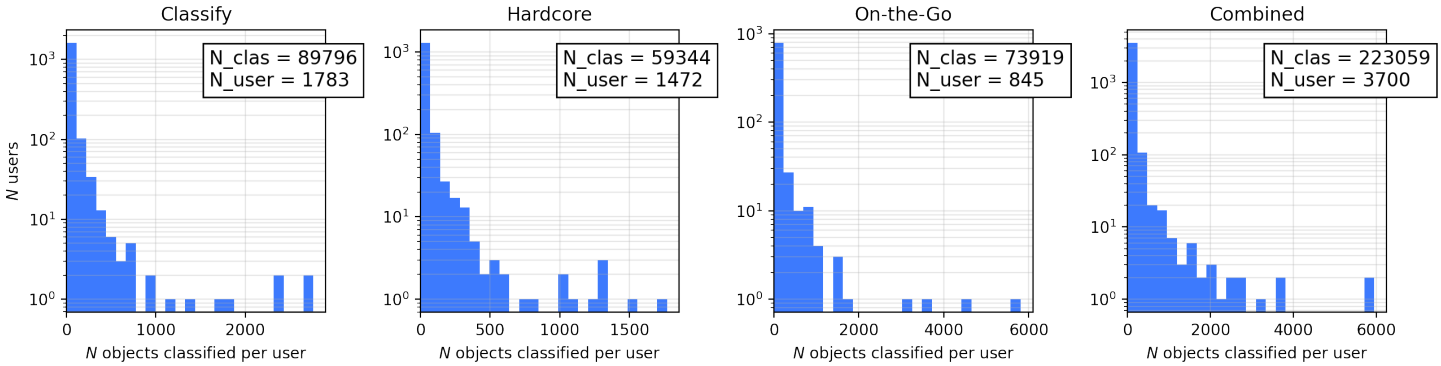


Figure 5: Distribution per-workflow of number of classifications made per user. The caption N_{clas} denotes the total number of classifications in the entire workflow, and N_{user} denotes the number of users in the workflow.

Figure 5 shows the distribution of number of classifications made per user, per workflow. Figure 4 indicates the distribution of total votes as a function of users included. Note the similarity between the "Classify" and "Hardcore" workflows when considering only the users that made 1000 classifications or fewer, and the divergence thereafter (the 'power-users', if you will, in the Hardcore category, made fewer classifications than the power-users in Classify). In "On-the-Go" we also see that a small fraction of users performed far more classifications than the rest. We also observe that users made more classifications, on

average, in "On-the-Go" than in the other two workflows. Table 3 describes a few basic statistics regarding the number of classifications made per user, per workflow. Additionally, we find that the distribution of time taken per classification is in line with the complexity of each workflow. The mean time per classification is 11, 20 and 41 seconds for "On-the-Go", "Classify!" and "Hardcore" respectively (after filtering out classifications that took more than an hour, as we assume the user in those cases left the page to do something else, and came back at a later date to finish that classification).

workflow	mean	median	standard deviation
Classify	50.36	14.00	162.68
Hardcore	40.32	9.00	121.89
On-the-Go	87.48	12.00	345.77
Combined	60.29	11.00	239.06

(a) Basic numbers on the amount of classifications done per user across workflows

workflow	mean (s)	median (s)
Classify	19.8	5.8
Hardcore	40.8	14.9
On-the-Go	11.2	3.0

(b) Basic information about average classification time (in seconds) for each workflow after filtering improbably long durations.

Table 3: Basic information on the average number of classifications per user in each workflow, and the average classification duration per workflow.

3 Analysing the classifications

This analysis can be partitioned as follows:

1. Compare the voting behavior of each individual user to that of all the users as a group.
2. Compare the classification behavior of the users as a group to the (photometric) parameters of the objects they classify.
3. Compare the users' behavior, combined with (photometric) parameters, to the likely ground truth catalogue.

3.1 User classification behavior among their peers

Since we cast every answer given by a user to a dataframe column, it becomes relatively straightforward to find a consensus for any given object, and any given task. Before we compare the volunteers' consensus to photometric properties, or to the likely ground truth catalogue (see 1.6), we will first examine how a consensus forms, by inspecting individual users' classification behavior.

Figure 6 describes what we will refer to as a user's 'precision' among their peers for three tasks. Because of the large variation in number of classifications per user, we plot the horizontal axis on both linear and logarithmic scale for clarity. To compute a user's precision for a task, f_{task}, we simply take the consensus vote for every object, for every task, and compare the answer the user submitted for that task to this consensus. If the user's answer coincides with the consensus (in this case, the 'consensus' answer is the answer that gets the most votes for that object, for that task), we say they 'match' the consensus. A user's precision, then, is simply the total amount of times they matched the consensus, divided by the total amount of times they submitted an answer for that task (for any object).

$$f_{\text{task}} = \frac{N_{\text{match}}}{N_{\text{match}} + N_{\text{no match}}}$$

Another way to compare the agreement between users in general is by evaluating Fleiss' kappa (κ). Fleiss' κ is a so-called inter-rater reliability statistic that takes a number of 'raters' (the users in Space Fluff), who categorize a number of 'subjects' (the Space Fluff objects) into a number of 'categories' (the unique answers given for each task). The computation goes as follows:

Let N be the total number of subjects. Let n be the number of ratings per subject, and k the number of categories into which ratings are assigned. Label the subjects with $i \in \{1, ..., N\}$, and the categories with $j \in \{1, ..., k\}$, then n_{ij} is the number of raters that assigned the ith subject to category j. Then, the proportion of assignments to the jth category is $p_j = \frac{1}{Nn} \sum_{i=1}^{N} N_{ij}$, with $\sum_{j=1}^{k} p_j = 1$. The extent to which users agree for the ith subject is

$$P_i = \frac{1}{n(n-1)} \sum_{j=1}^{k} n_{ij}(n_{ij} - 1)$$

The mean extent to which raters agree on subjects, then, is

$$\bar{P} = \frac{1}{N} \sum_{i=1}^{N} P_i$$

Finally, we can compute Fleiss κ as follows:

$$\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e},$$

where $\bar{P}_e = \sum_{j=1}^{k} p_j^2$

The interpretation of κ varies depending on the situation, though generally $\kappa \leq 0$ denotes agreement worse than chance, $0.01 \leq \kappa \leq 0.6$ denotes slight to moderate agreement (increasing κ means better agreement), and values between 0.61 and 1 denote almost perfect agreement. We will use this measure at a few places throughout this work when taking subsets, so that we have access to some measure with which to compare various subsets of classifications.

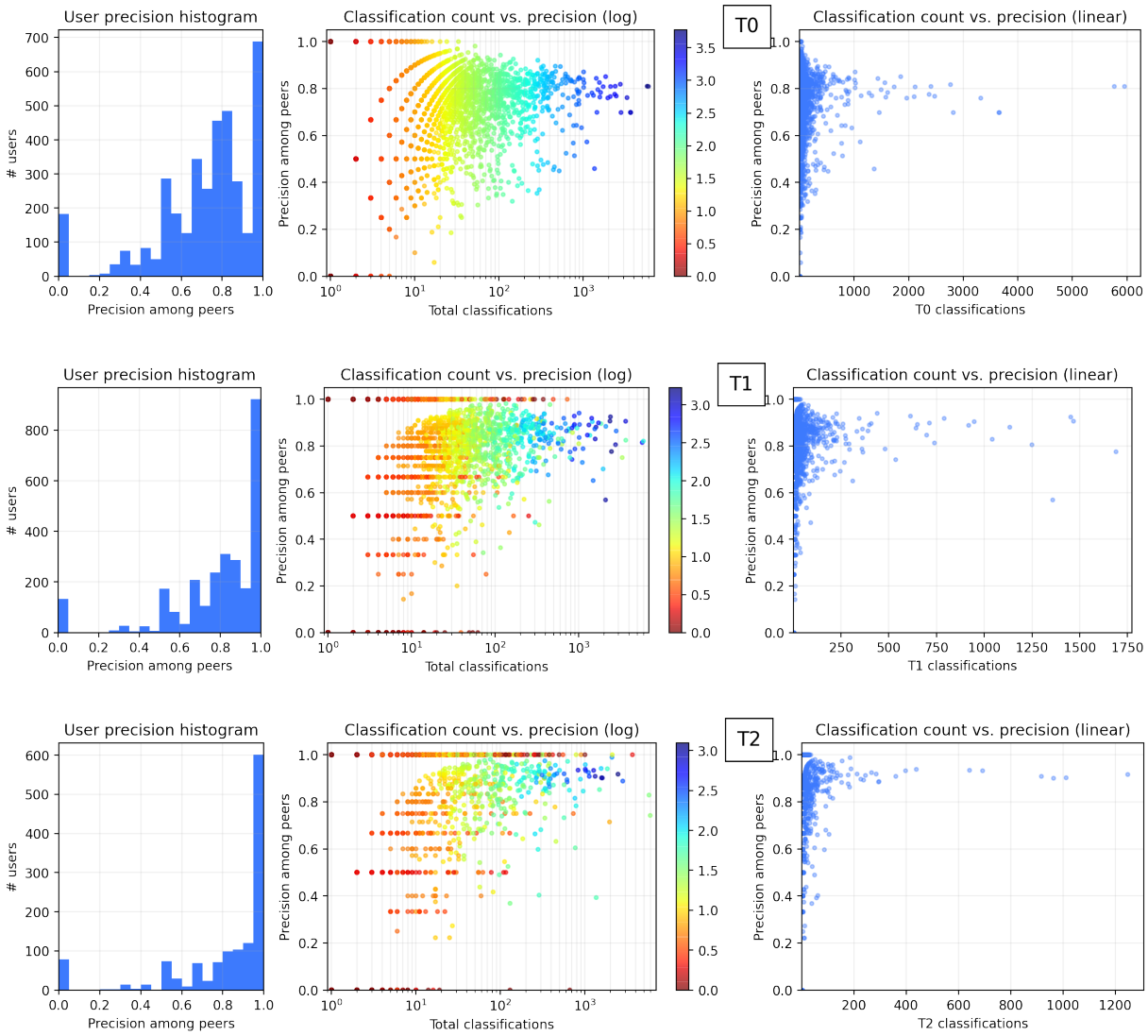


Figure 6: Precision of users among their peers for three tasks (T0, T1, T2). Each row of graphs describes one task. The leftmost plot is a histogram of user precision for that task. The central plot compares each user's average precision for that task to the *total* amount of classifications they made. The colormap indicates (on log base 10 scale) the amount of times they answered *the specific task*. The rightmost plot describes (in linear scale) the user's average precision for a task compared to the amount of times each user answered the specific task.

3.2 Comparing classifications to photometric properties

3.2.1 *Objects without photometric properties*

Of the 6362 objects in the project, 922 do not have properties associated with them in the unfiltered catalogue/dataset. None of these objects eventually make it into the likely ground truth catalogue. These objects are, however, in the Space Fluff dataset, so we must decide how to treat them. In many of these cases, we can verify by inspection of the objects' thumbnail images (see e.g. figure 7) that the reason the images do not have photometric properties assigned to them (i.e. the objects do not exist in the unfiltered catalogue) is because the objects, even though they were pre-selected by MTO for Space Fluff, are for example (1) not easily definable due to the presence of a nearby very bright object, (2) not definable due to image quality issues, like the presence of artifacts, or (3) appear in the image as a group of background objects scattered around the center, rather than a single object in the center.

Figure 8a shows the vote distribution for task 0 ("Look at the very center of the image: do you see

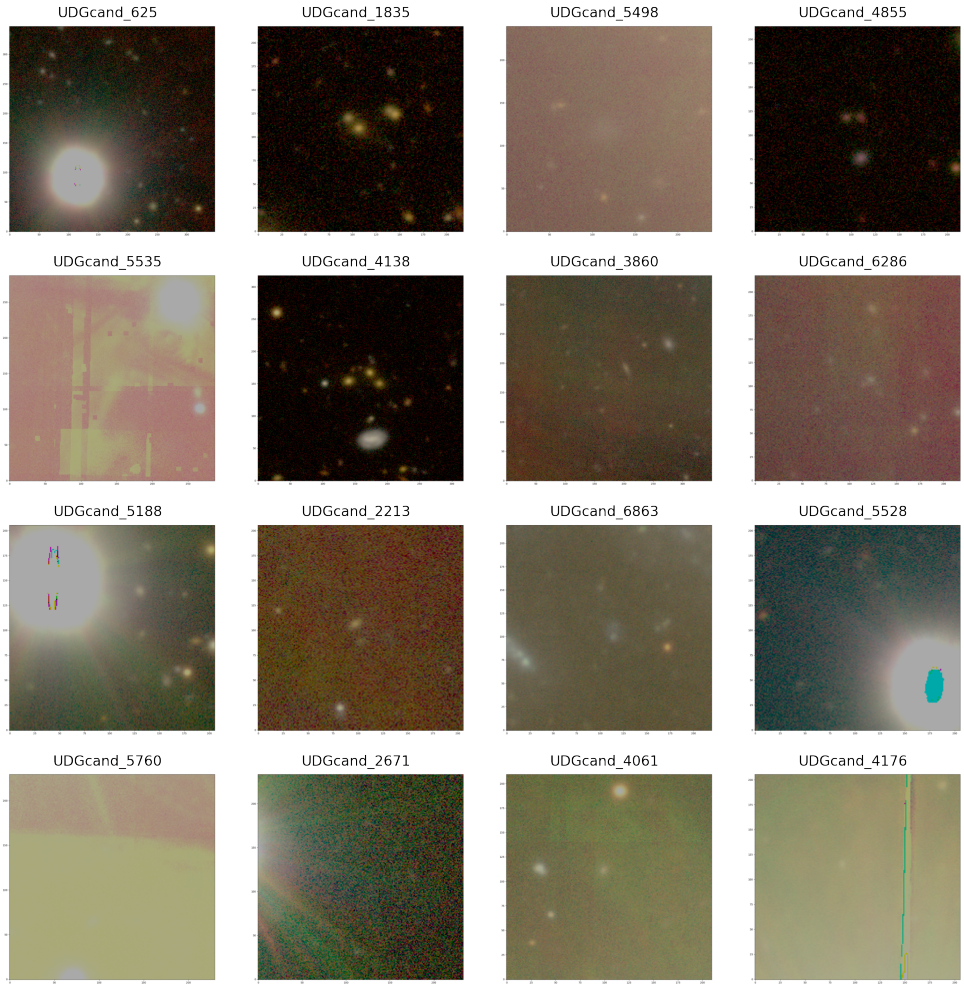
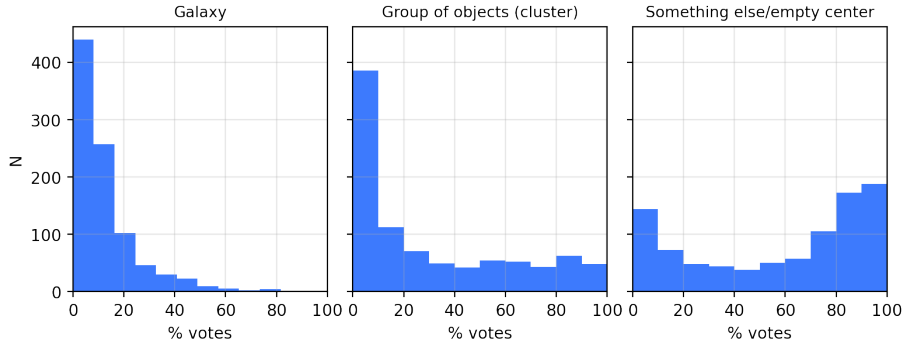
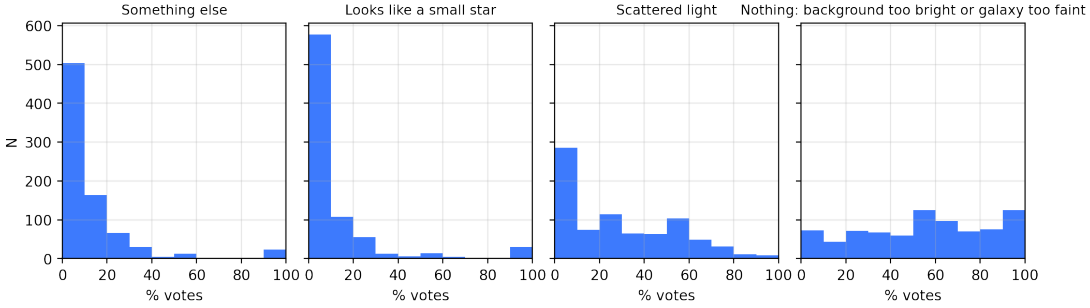


Figure 7: 4x4 image grid of a random selection of objects that do not have photometric properties associated with them in the unfiltered dataset by Venhola et. al (see section 1.6).

a single galaxy or a group of far away objects?") answers for the objects without photometric properties. Assuming these objects indeed do not have properties due to any mixture of the reasons outlined above, the low percentage of 'galaxy' votes for these objects indicates that the users are generally adept at finding these anomalous objects. Of the 922 objects without properties, only 19 are considered galaxies by the majority of the users that classified them. Due to the nature of the project, task 0 was the only one presented to users in case they classified an object as being a 'group of objects'. As mentioned in section 2.1, task 9 ("Our bad! what do you see instead?") was presented in the Hardcore workflow in case users voted 'something else/empty center' in task 0. We include the distribution of task 9 votes for objects without properties in figure 8b.



(a) Task 0 distributions



(b) Task 9 distributions

Figure 8: Vote distributions for objects without properties in the unfiltered catalogue (see section 1.6). For reference, Fleiss' κ computed for task 0 for objects without properties is 0.69, indicating moderate to strong agreement among users.

3.2.2 *Objects with photometric properties*

In this section, we will compare the users' classifications for the objects that have (photometric) properties in the unfiltered catalogue (see 1.6). We will present comparisons for the most relevant tasks in the project one-by-one. For reference, Fleiss' κ for task 0 for objects with properties is 0.66, indicating moderate to strong agreement among users in general, but slightly lower agreement on average than for the objects without properties.

Task 0 ("Look at the very center of the image: do you see a single galaxy or a group of far away objects?") The most important answer in this task is 'galaxy'. In figure 9, we compare the (photometric) properties of objects to the number of galaxy votes they receive. Interesting to note is the relation between magnitude and percentage votes for 'galaxy': objects with fainter r' magnitudes are less likely to be considered to be a galaxy by a majority. This relation can be observed in figure 30 (in the Appendix), where we plot the r' magnitude against the percentage of 'galaxy' votes each object receives, along with a linear fit. The Pearson correlation coefficient for this relation is -0.53, which indicates a moderate (but certainly not strong) correlation.

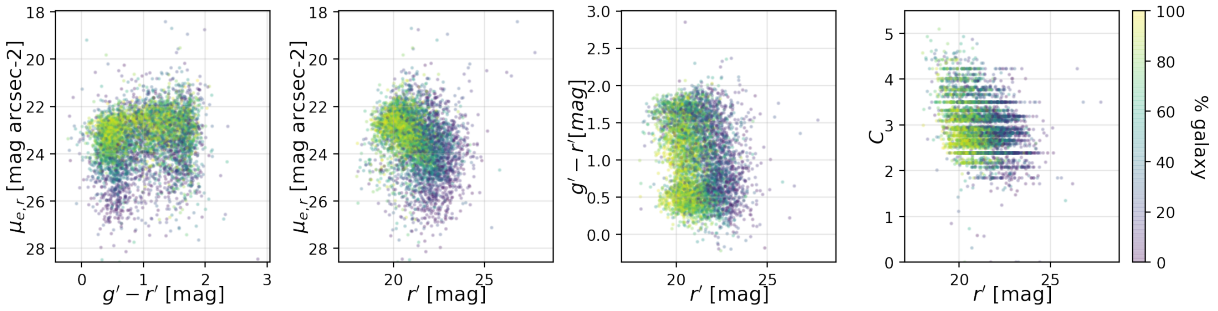
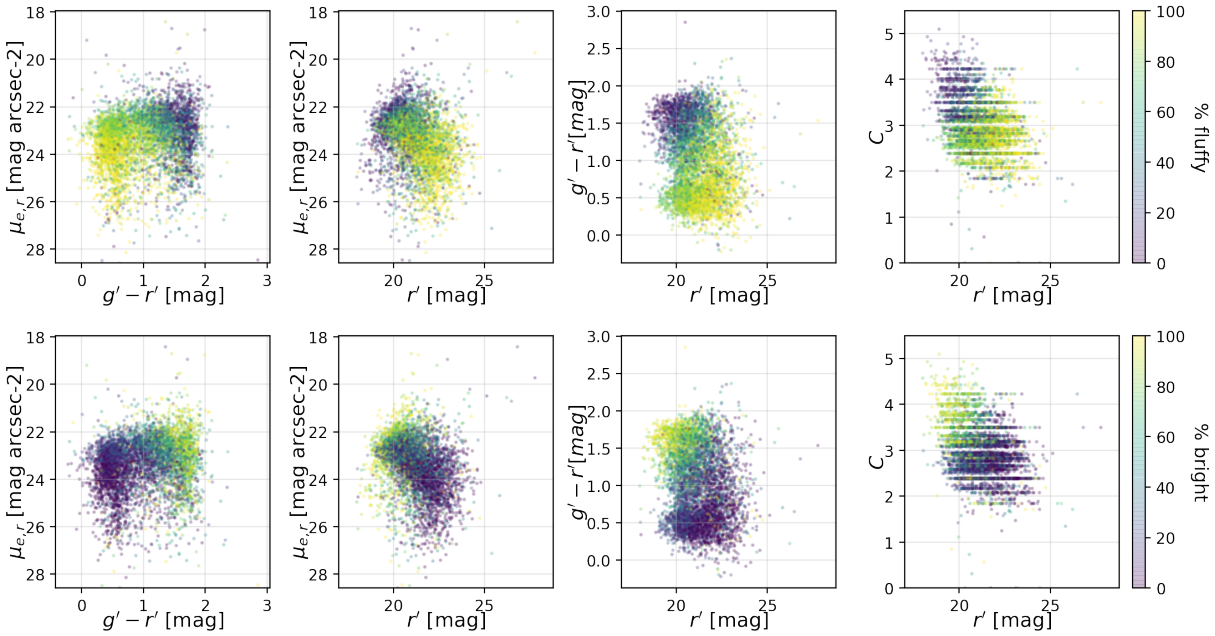


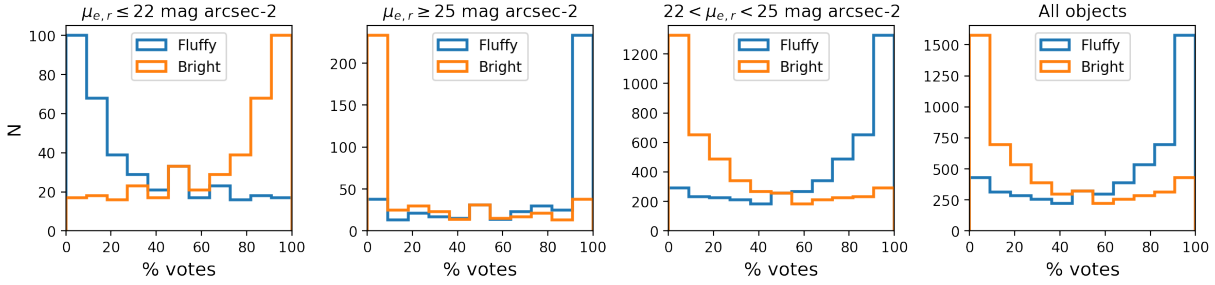
Figure 9: Comparison between Space Fluff objects in various parameter spaces. The color mapping indicates the percentage of users that classify an object as a galaxy.

Task 1 ("Is the galaxy fluffy or is it bright?") Figure 10 displays the objects in parameter space again, but now colored by the percentage of votes for each task 1 answer. In figure 10b we see that, generally, the majority of users classify an object as bright if its (mean effective r-band) surface brightness exceeds 22 mag/arcsec^2, and fluffy if the surface brightness is below this value. Note that we do not account for whether the majority of users thinks the object in question is even a galaxy in the first place. Task 1 is only shown to a user if they already classify the object as a galaxy. This means that objects only classified as galaxies by a minority of users are considered equal to objects with a majority 'galaxy' consensus in this plot. Subfigure (c) describes more clearly the correlation between 'fluffy' votes and color, and surface brightness. The Pearson correlation is much stronger between color and 'fluffy' votes, than is is between surface brightness and 'fluffy' votes.

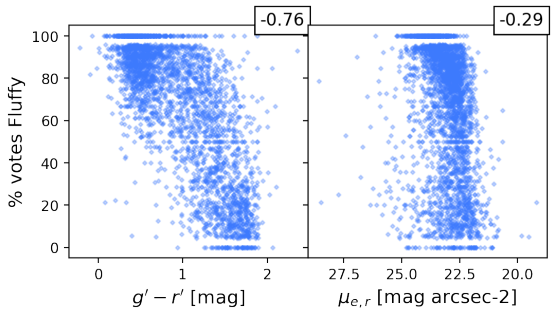
Referring to an excerpt of the Space Fluff field guide (see the Appendix), we note that a disproportionate number of objects indicated there as 'bright' also displays this color bias, which is most likely an unfortunate coincidence.



(a) Objects in parameter space, with the percentages of T1 votes for 'fluffy' and 'bright' as color bars.



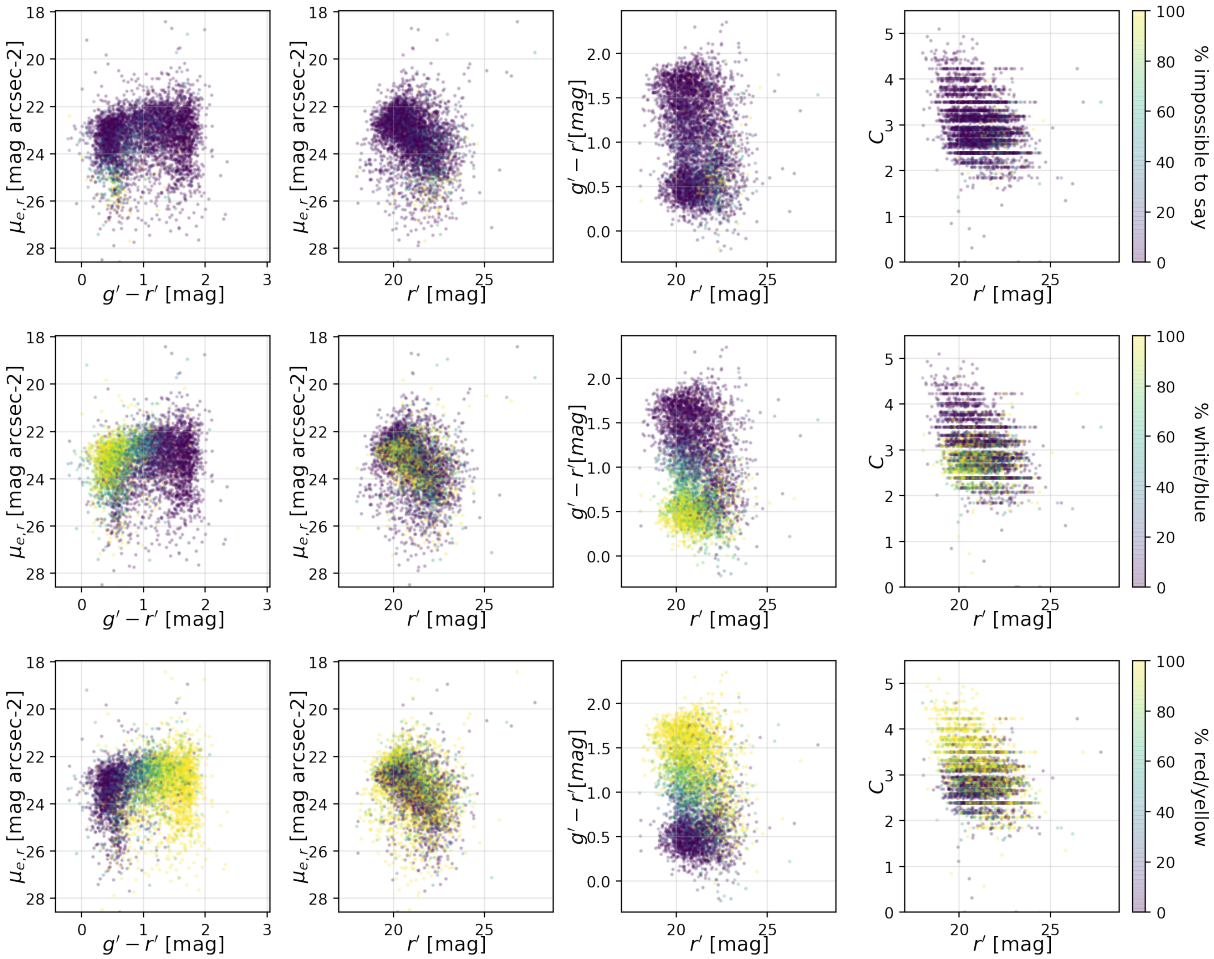
(b) Percentages of T1 votes for various subsets of objects of various surface brightnesses. Note that we do not distinguish here between an object that has, for example, 22 votes with all in favor of 'fluffy', and one that only has 2 votes with both for 'fluffy'. The vertical axis denotes the number of objects in each bin. Also note that the fact that the two options mirror each other perfectly is simply due to the fact that there are only two possible answers for T1.



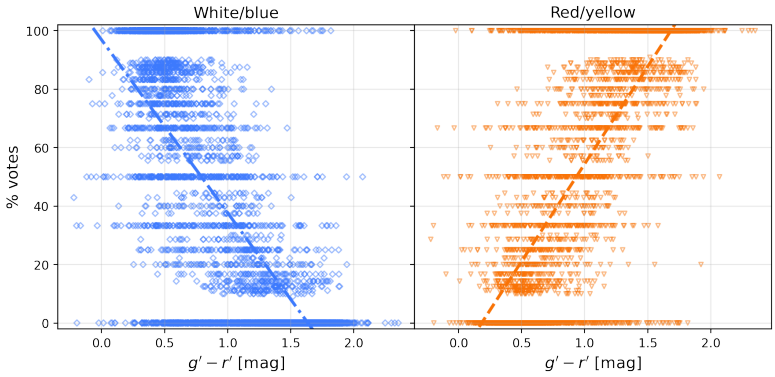
(c) Percentage of T1 votes for 'fluffy' compared to object color and surface brightness. The annotation in the top right of each subplot denotes the Pearson correlation coefficient.

Figure 10

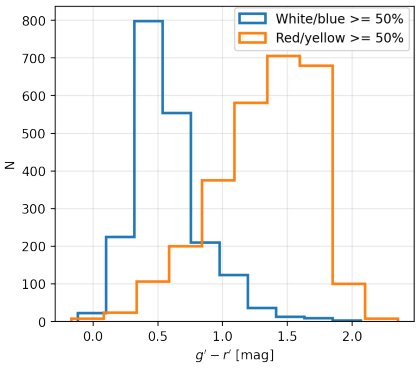
Task 2 ("What color is the galaxy?") From figure 11(b and c) it is evident that the color a user perceives is correlated moderately strongly to the astronomical color of the object. In subfigure (b): for white/blue, the Pearson correlation coefficient is $\sim$ -0.73, and for red/yellow it is $\sim$ 0.78. Note that this subfigure includes even objects with only one T2 vote. If we increase the threshold, the correlation will increase (e.g. when considering only objects with at least five T2 votes, the correlation coefficients become approx. -0.87 and 0.89, respectively). In the Appendix, we include a small image grid of some of the objects that have $g' - r' < 0.5$ (bluer objects), that get are voted by the majority as being red/yellow, and have at least five T2 votes (this excludes the objects that only got one T2 vote to begin with, in which case the color classification is not statistically significant).



(a) Comparison between Space Fluff objects in various parameter spaces. The colorbar indicates the percentage of users that vote for that option.



(b) Percentage of votes for each type of answer, plotted against the $g' - r'$ color of the objects. The points each indicate an individual object. The straight line in each subplot indicates a linear fit.



(c) Histogram of $g' - r'$ color for objects with at least 50% of their T2 votes for 'white/blue', and the same for 'red/yellow'.

Figure 11: Distributions in parameter space, and histograms, of objects and users' task 2 votes.

3.3 Reproducing selection cuts on Space Fluff data

To gain a better idea of the relation between users' classifications and the ground truth, we will in this section retrace the selection cut procedure done by Venhola et al. [25] for the objects classified as fluffy galaxies by the Space Fluff users, in order to filter out the bulk of the most likely non-cluster (i.e. background) galaxies. We do this keeping in mind the main objective of the Space Fluff project, which is to determine the viability of purely using manual classification as done by a large group of users. Any additional steps we need to take to refine the users' decisions reduces the strength of the idea that user classifications can be a pure indicator for a galaxy's Fornax cluster membership.

Figure 12a describes the selection cuts as performed by Venhola et al. in [25]. Figure 12b compares results from [25] to those in the what we call the likely ground truth catalogue (which is the result of Venhola et al., *in prep.*). In subfigure (a), the objects excluded by each selection cut are in the grey shaded region. In subfigure (b), the grey dashed line indicates the surface brightness cut.

To perform the selection cuts, we begin with the set of objects the users classify as fluffy galaxies (see also section 3.4.1), and perform the selection on those objects. The remaining objects (all artifacts, groups of objects, galaxies the users don't consider fluffy, etc.) in the Space Fluff dataset have essentially been excluded from selection by the users themselves already.

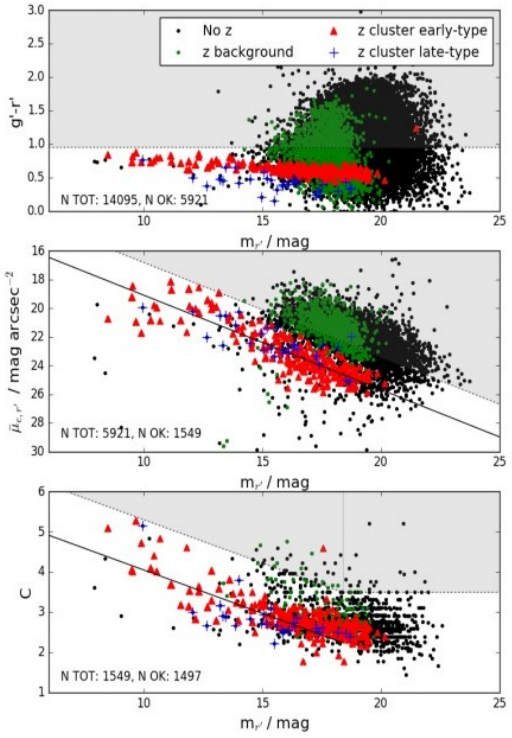


Fig. 16. Illustration of our main criteria for distinguishing the cluster and background galaxies from each other. The panels from top to bottom show how the g'-r' color (also g'-i' cut was used which looks very similar), the mean effective surface brightness $\bar{\mu}_{e,r'}$, and the concentration parameter C of the spectrally confirmed (Drinkwater et al. 2000) early- (red symbols) and late-type (blue symbols) cluster and background galaxies (green symbols), scale with the r'-band apparent magnitude ($m_{r'}$). The solid lines show the fits to the early-type Fornax cluster galaxies, and the dotted lines show the selection limits. The excluded areas are shaded with gray. The black dots correspond to objects with no spectra available. The numbers in each plot correspond to the total number of galaxies before the cut, and the number of galaxies that remain after the cut. The two lower panels show only the galaxies that have not been excluded in the previous steps.

(a) Description of the selection cuts, from Venhola et al. (2018) [figure 16, [25]].

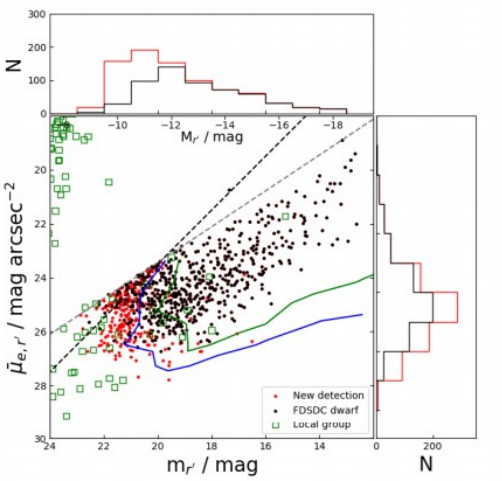


Fig. 7. Apparent r'-band magnitudes ($m_{r'}$) and mean effective surface brightnesses ($\bar{\mu}_{e,r'}$) of the FDSDC galaxies (*black symbols*) compared with the LSB extension galaxies described in this work (*red symbols*). The histograms on the x- and y-axes show the distributions of the magnitudes and surface brightnesses with the same colors as in the scatter plot. The blue and green contours show the 50% and 70% detection limits found using the artificial galaxies (Section 3.2) and the black and grey dashed lines show the $a > 2$ arcsec and surface brightness selection limits, respectively. For a comparison, we also show a sample of Local group dwarfs from Brodie et al. (2011).

(b) Comparison between FDSDC (the catalogue resulting from [25]) and the likely ground truth catalogue (Venhola et al., *unpublished*).

Figure 12: Selection cuts as performed by Venhola et al. Note that these two subfigures are from different works; the color schemes of data points between the two subfigures bear no relation.

Color cut Excluding any objects with color $g' - r' > 0.95$ mag removes 150 of the 1050 objects classified as fluffy galaxies. Also excluding those with $g' - i' > 1.35$ excludes an additional handful of objects.

Surface brightness cut Extracting and fitting to the same set of confirmed galaxies as done in [25] is slightly outside of the scope of this thesis, but as an approximation, we instead take the likely ground truth catalogue and approximate the linear fit from the shape of the catalogue objects in the magnitude-surface brightness space. In figure 13 we show our approximation to the linear regression that determines the surface brightness cut.

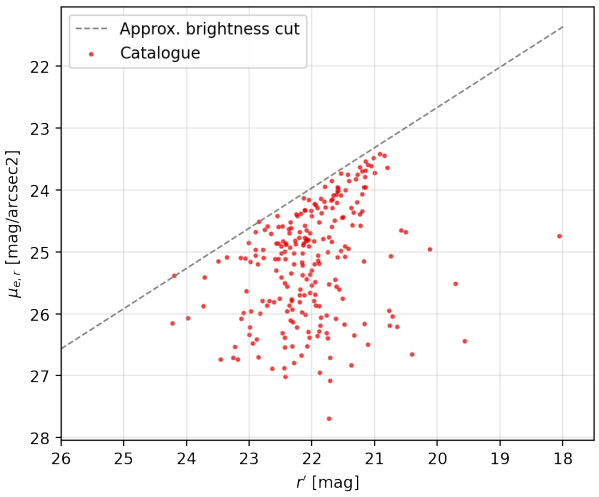


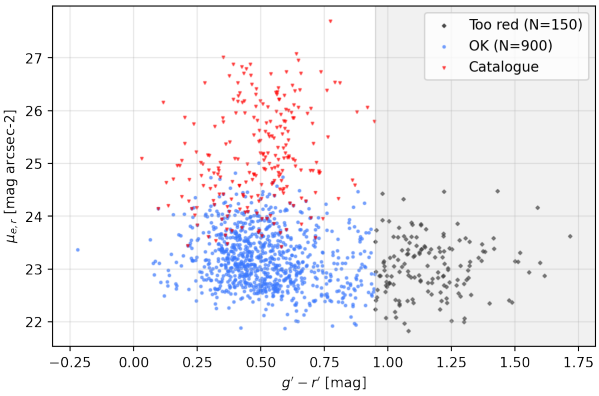
Figure 13: Likely ground truth objects in magnitude-surface brightness space, along with our visually derived estimate of the linear regression that determines the surface brightness cut. For the sake of reproduction: our approximation is of the form $\mu_{e,r} = 9.771 + r' \times 0.645$. Colors and axes are chosen to reflect those in figure 12b for the sake of comparison.

Concentration cut Performing the concentration cut by removing objects with a concentration parameter $C > 3.5$ excludes another 3 objects the users classified as fluffy galaxies.

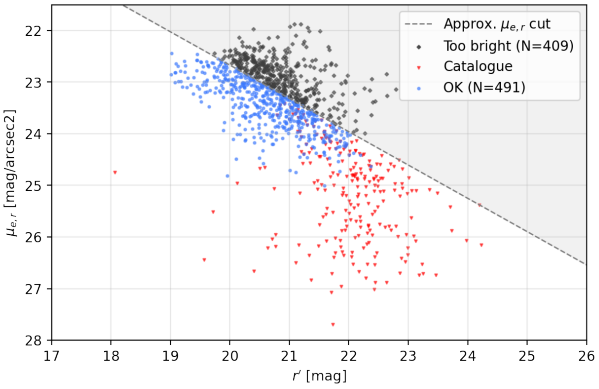
We visualize the selection cut process using our approximation to the surface brightness cut from Venhola et al. in figure 14. From the 1050 user-classified fluffy galaxies, 488 remain after these selection cuts. In table 4 we list the results of performing these selection cuts on various subsets of Space Fluff data, where we also include the results for the 90% threshold on fluffy votes (see section 3.4.2). We see that among the subsets, the highest survival rate occurs for the user-classified fluffy galaxies.

For comparison, we also ran 1000 simulations with randomly assigned T0 'galaxy' vote percentages, performing selection cuts on the subset of galaxies that get at least 50% and 75% votes for galaxy, and find that, on average, $22.0 \pm 0.6\%$ and $22.0 \pm 1.0\%$ of objects survive the selection cuts for the 50% and 75% thresholds, respectively (since the votes are randomly assigned, it makes sense that the percentage threshold does not make a difference in the mean survival rate). This is on the order of the overall percentage of objects with properties that survive the selection cuts, as expected.

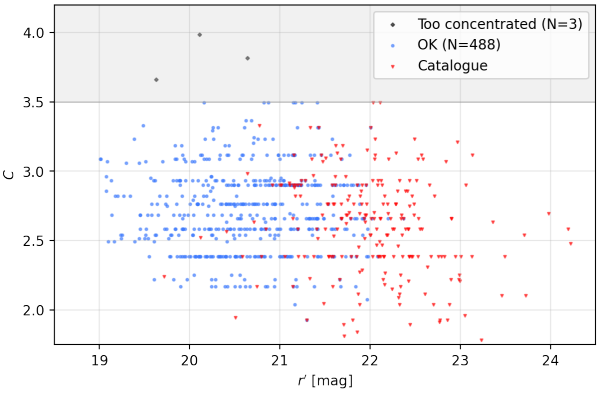
We reiterate that, since the selection cuts are set up so that the likely ground truth objects retain their alleged cluster membership, and the selection cuts in all cases decrease (or maintain) the number of objects in the Space Fluff dataset possibly in the Fornax cluster, this is a rather artifical statistic for the purposes of analyzing the accuracy of user classifications. However, for the purposes of picking a subset of objects from a large dataset for further inspection, this does significantly reduce the size of the dataset, with few disadvantages (a small percentage of actual cluster members may be labeled as background galaxies by the selection procedure, as outlined in [25]).



(a) Color cut



(b) Surface brightness cut



(c) Concentration cut

Figure 14: Performing selection cuts on fluffy galaxies as classified by the users. Points excluded due to each selection cut (colored in black each time) do not make it into the next subfigure. The objects marked "OK" are those that survive the cut, i.e. those that can still be considered for cluster membership after the selection step. We also show the likely ground truth objects (marked as 'Catalogue') for the sake of comparison. Note of course that there is some overlap between "OK" objects and likely ground truth.

3.4 Comparing user classifications to likely ground truth

After comparing the voting behavior of the users to photometry, we now turn to the other half of our analysis; in this section we will relate the user consensus to the likely ground truth catalogue in order to determine the usefulness of manual classification in the process of identifying UDGs.

subset of objects	N	N_{OK}	% OK (rounded to nearest integer)
All objects with properties	5440	1196	22%
User-classified galaxies (75% threshold)	1973	532	27%
Galaxies also classified as fluffy (75% threshold)	1050	473	45%
Galaxies also classified as fluffy (90% threshold)	548	317	57%

Table 4: Selection cuts performed on various subsets of Space Fluff data. The N column denotes the total number of objects in the subset, N_{OK} the number of objects that survive the selection cuts, and $\%OK$ the percentage of objects that survive the selection cuts.

3.4.1 *Fluffy galaxies according to the users*

Of the 6362 objects in the dataset, 1050 are classified as fluffy galaxies (with 75% classification thresholds). What objects do the users consider to be fluffy galaxies? In figure 15, we compare properties of objects in the overall dataset to the subset of objects that are considered galaxies (by at least 75% of users) and also fluffy (with two vote thresholds; the number 1050 we mention above is obtained using a threshold of 75%).

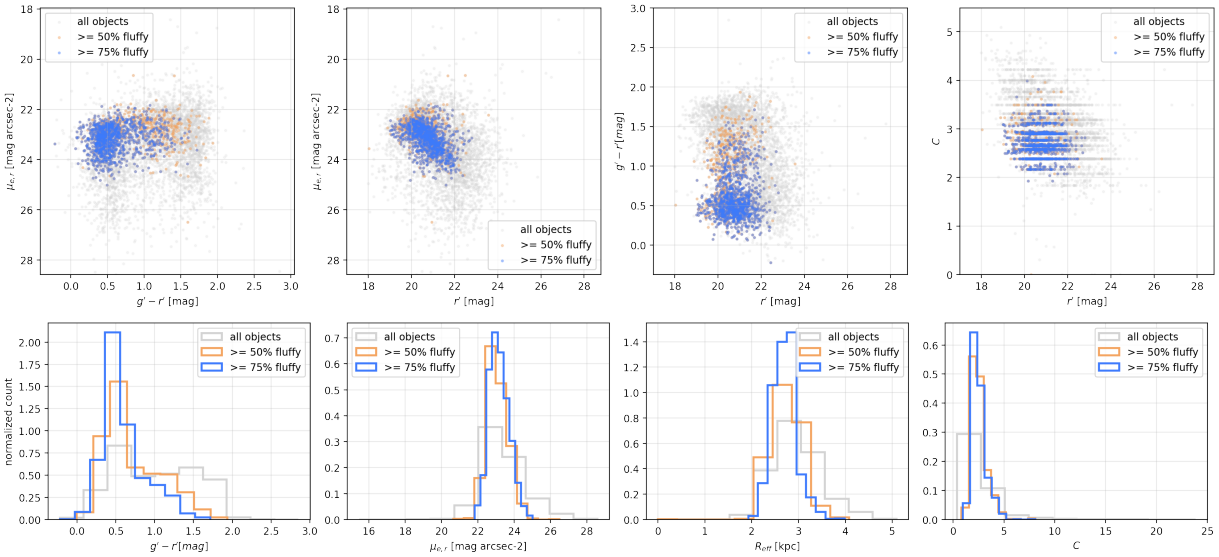


Figure 15: Comparison in parameter space between all objects in Space Fluff, and objects that are classified galaxies ($\geq$ 75% threshold) and fluffy (with thresholds of $\geq$ 50% and $\geq$ 75%).

We note a few things: (1) redder objects get classified as fluffy galaxies much less often than bluer objects. Approx. 48% of the objects in the overall dataset (that have properties) have a color of $g' - r' \geq 1.0$ mag, but for objects the users classify as fluffy galaxies (galaxy $\geq$ 75%, fluffy $\geq$ 50%), only approx. 24% of the objects satisfy this inequality, and even fewer if we bump the 'fluffy' threshold to $\geq$ 75% of votes: then, only 12% of objects have $g' - r' \geq 1.0$ mag. (2) Lower surface brightness objects are less likely to be considered fluffy galaxies. While 25% of objects with properties in the complete dataset have a surface brightness fainter than 24 mag arsec-2 (i.e. $\mu_{e,r} \geq 24$ mag arcsec-2), of the objects classified as fluffy, only 8% are this faint or fainter.

The above can be visualized in figure 16, where we display the fraction of objects remaining in the '$\geq$ 75% fluffy' subset and the overall dataset as we progressively increase surface brightness and color thresholds. We also see here that the number of objects considered 'fluffy' decreases more sharply as a function of surface brightness than the number of objects in the overall dataset does, for example.

3.4.2 *Likely ground truth catalogue objects*

Figure 17 displays the distribution of votes across all 230 objects in the likely ground truth catalogue that are also present in Space Fluff. There are another 35 objects in the catalogue that are not present in Space Fluff, and thus those naturally don't have any classifications. It is evident from the figure that only a few

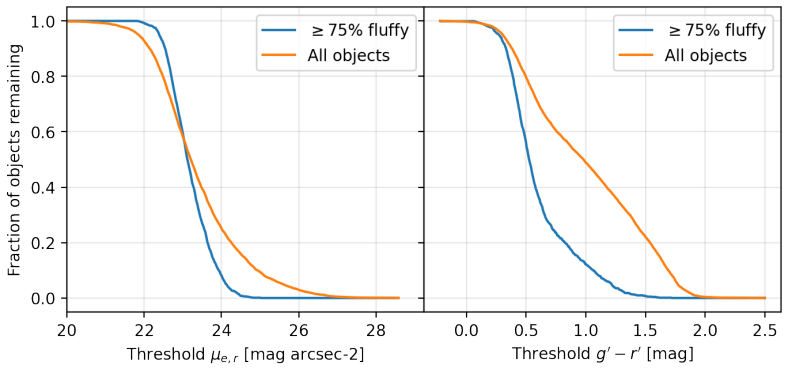


Figure 16: Cumulative fraction of objects remaining in two subsets of the data as we increase surface brightness and color thresholds, respectively. The vertical axis denotes the fraction of objects for which the quantity on the horizontal axis is at least equal to the value given on the horizontal axis (e.g. for the 'all objects' set, a fraction of ~ 0.5 (so approx. 48%) of objects have color $g' - r' \geq 1.0$ mag. Note for the blue line, we consider objects that are voted 'galaxy' by at least 75% of users, and also 'fluffy' for T1 by at least 75% of users.

of these objects are considered groups of objects by the users. In fact, only 19 of the 230 objects ($\sim 8.7\%$) get the majority ($\geq 50\%$) of their votes in favor of this option, which is quite a bit less than the same statistic for the complete dataset ($\sim 33.7\%$). Furthermore, we note that Fleiss' κ for task 0 for the LGT objects is 0.57, significantly lower than for the general set of objects with properties, indicating that users have a harder time forming a clear consensus on the identity of these objects.

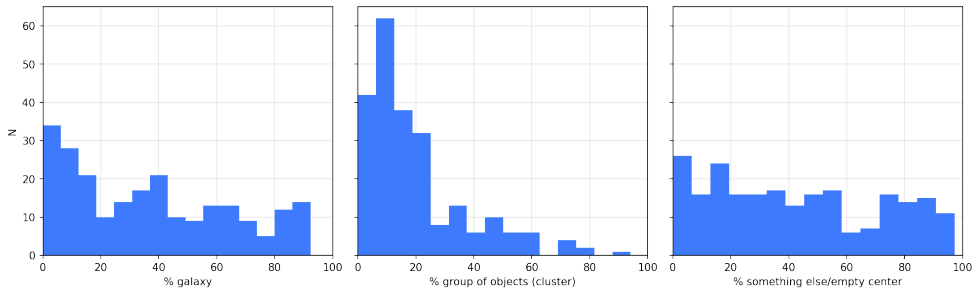


Figure 17: Vote distributions for task 0 of all likely ground truth catalogue objects that were presented to users for classification in Space Fluff. The vertical axis indicates the number of objects.

Figure 18 contains comparisons in parameter space and of parameter distributions between the overall dataset and the likely ground truth objects.

Ground truth objects classified as fluffy Of the 232 likely ground truth objects, only 75 ($\sim 32\%$) are voted to be galaxies by at least half the users that classified them. If we shift this threshold from 50% to 75%, only 30 of the 232 objects are classified as galaxies. Interestingly, the overwhelming majority of users that see these objects, classify them as fluffy galaxies rather than bright galaxies, which is exactly what we would hope for, because the 'fluffy' identifier is supposed to indicate an LSB galaxy by the definition of the project. These 30 objects each have at least 12 'task 1' (T1) votes, and each has at least 91.3% of its T1 votes in favor of 'fluffy'. This is a much larger fraction than the general subset of objects with a 'galaxy' consensus: for objects with at least 75% votes for 'galaxy', only $\sim 27.7\%$ have at least 90% of their T1 votes as 'fluffy'.

Only 8 of the 232 likely ground objects get fewer than 50% of their T1 votes for fluffy. Among these, only two have more than 3 T1 votes. It is clear, then, that the difficulty in identifying the likely ground truth objects is not that they do not appear fluffy, but that they are extremely difficult for the users to identify as galaxies to begin with. Figure 19 displays five of these objects for the sake of comparison.

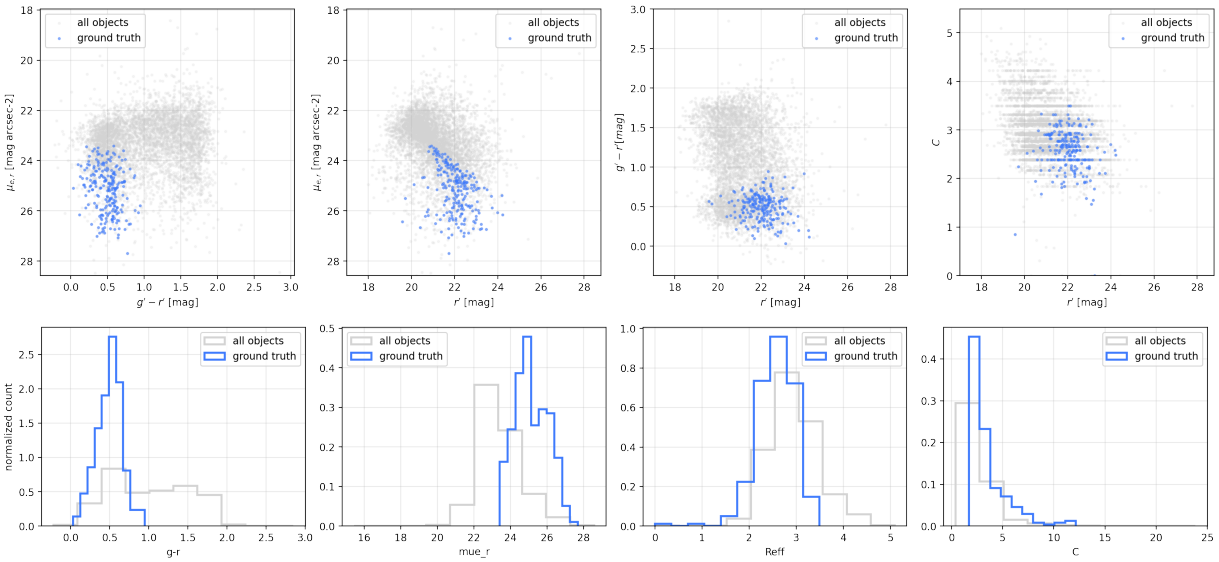


Figure 18: Top row: a comparison in parameter space (color, magnitude, surface brightness, effective radius and concentration) between the likely ground truth objects and the complete Space Fluff dataset. Bottom row: comparison of the distribution of various parameters between these same two categories. Note the labels on each subfigure, as the top and bottom figures don't necessarily correspond.

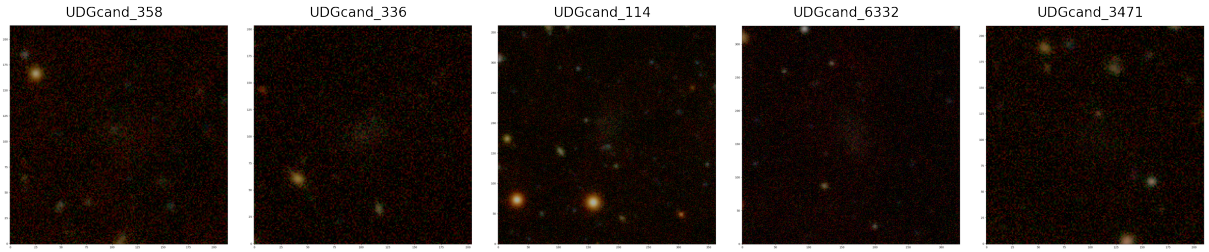
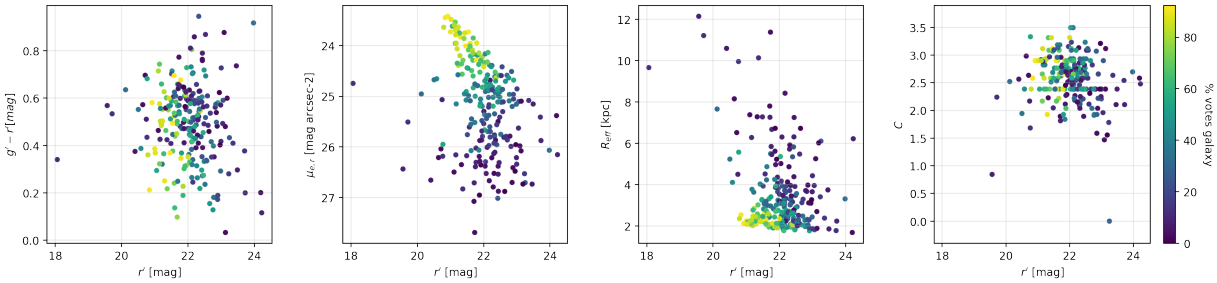


Figure 19: Five objects from the LGT catalogue with $\mu_{e,r} > 26$. Each of these was classified as a galaxy by fewer than 15% of users. Note that the galaxy is in the exact center of the image - do not be distracted by nearby interlopers!

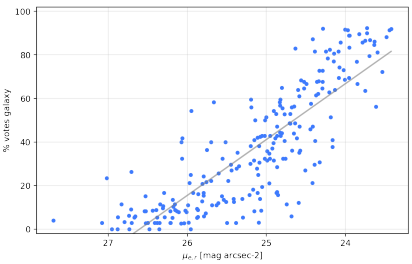
Ground truth objects not classified as fluffy Why do the other 202 likely ground truth objects not get classified as fluffy galaxies? In figure 20, we plot some of the same object parameters we've plotted before, but now we color-code them by the amount of T0 'galaxy' votes the objects receive. The second subfigure in figure 20a alludes to a relation between surface brightness and votes in favor of 'galaxy'. We quantify this relationship in figure 20b by including a simple linear fit, and by computing the Pearson correlation coefficient, which evaluates to a relatively strong correlation of 0.84 (or rather -0.84, but because higher surface brightness corresponds to a lower number, we invert the x-axis on the plot, which inverts the sign of the correlation). Figure 20c repeats (b), but for the entire dataset. The correlation is much weaker in this case ($\sim$ 0.34), but we do note from this subfigure that lower surface brightness objects generally receive fewer 'galaxy' votes than higher surface brightness ones.

Fluffy objects not in ground truth, and selection cuts Of the 1050 objects classified by the users as fluffy galaxies (see section 3.4.1), only 30 ($< 3\%$) are actually in the likely ground truth catalogue. In figure 20 we related the surface brightness of an object to the percentage of 'galaxy' votes it receives, and noted that for this intersection of ground truth objects and objects classified as 'galaxy', all 30 objects classified as galaxies are also classified as 'fluffy' galaxies.

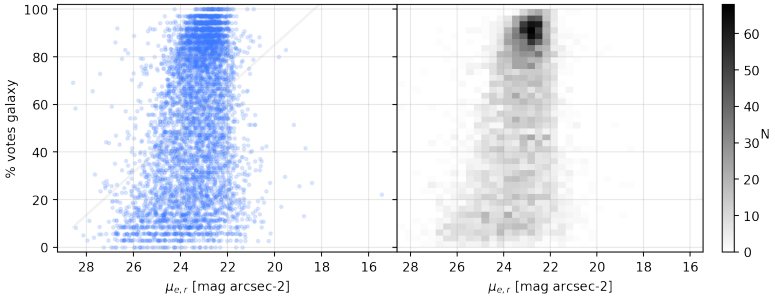
What about the hundreds of remaining objects classified as fluffy galaxies? Have the users identified new cluster members, or are there physical reasons for the discrepancy between catalogue and user consensus? One thing to note is that we have thus far only considered the likely ground truth catalogue for comparison. In its preparation, Venhola et al. disregard any objects already present in the Fornax Deep



(a) Objects in the likely ground truth catalogue plotted in parameter space, colored by percentage of T0 'galaxy' votes. Each point represents one object.



(b) Surface brightness plotted against percentage of T0 'galaxy' votes. The grey line represents a simple linear fit (of the form $y \approx 25.65 \times x + 682.49$). The points represent only objects in the likely ground truth catalogue



(c) Surface brightness plotted against percentage of T0 votes for 'galaxy'. The grey line represents a simple linear fit. The points represent all objects in Space Fluff for which a $\mu_{e,r}$ value is present in the dataset. On the left: scatter plot, each point represents one object. On the right: two-dimensional heatmap, where the number of points in each square is indicated by the color bar.

Figure 20

Survey Dwarf Catalogue (FDSDC), which is the catalogue that resulted from the work in [25]. These are Fornax cluster dwarf galaxies, so not necessarily LSB dwarfs or UDGs, however we could argue that if the users correctly identify one of these galaxies, they have still succeeded; a significant fraction of these dwarfs are, indeed, of low surface brightness.

Cross-matching the coordinates of the centers of each of the FDSDC galaxies to within 3 arcseconds of the center of a Space Fluff object (this procedure is also used in Venhola et al., *unpublished*), which at the distance to the Fornax cluster (approx. 19.95 Mpc) corresponds to a center-to-center distance of at most $\approx$ 280 pc; significantly smaller than what we generally expect a Fornax dwarf to be, yields a significant number of objects also classified by the users as fluffy galaxies. Setting the threshold for galaxy classification at 75% of votes or more, and 75% for fluffy classification, we find that 39 of the 124 FDSDC galaxies present in the Space Fluff project are classified by Space Fluff users as fluffy galaxies. We display the parameters of all the FDSDC objects we find in the Space Fluff dataset, alongside the objects from the likely ground truth catalogue, in figure 21. It is possible that comparison to other existing LSB dwarf/UDG catalogues reveal more intersections with objects classified as fluffy galaxies by Space Fluff users. This does, however, not change anything about the difficulty the users faced in identifying the lower surface brightness objects, however it does serve to show the usefulness of classification by humans.

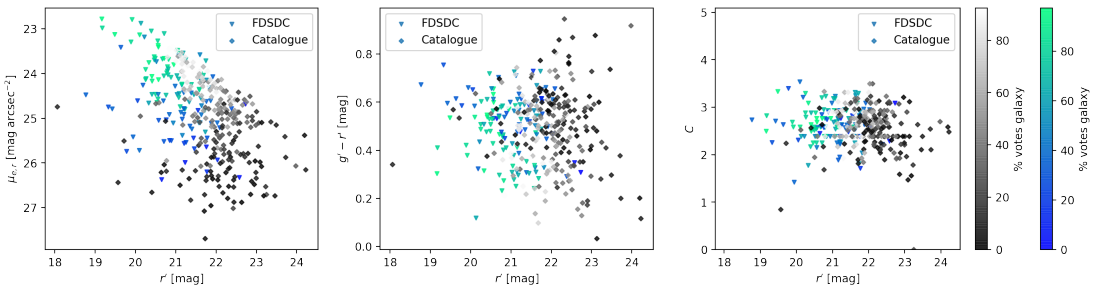


Figure 21: FDSDC and likely ground truth objects in parameter space.

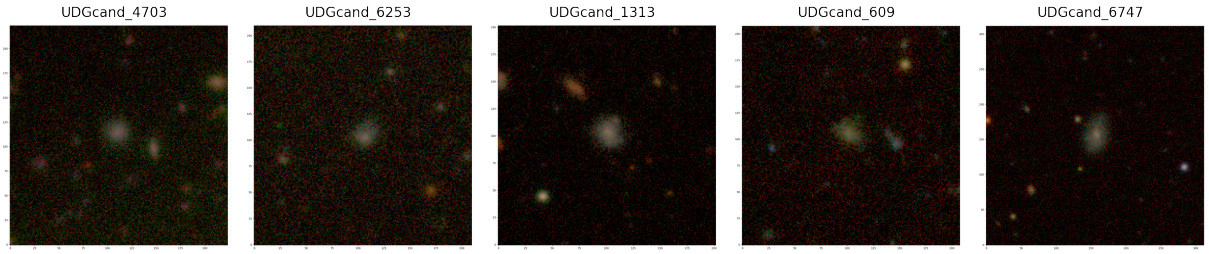
3.5 Manual classification on user-selected fluffy galaxies

As previously described, after performing parametric selection cuts, Venhola et al. go on to perform a manual selection of the remaining objects, labeling all the objects with clear structure like spirals and bulges, and excluding them from the catalogue, since those structures are not expected to exist in low-surface brightness galaxies in the Fornax cluster. This selection is hard to reproduce manually, however, to gain a better idea of the process, and also to identify any objects that may not have been selected for the likely ground truth catalogue by Venhola et al., but could conceivably be included based on visual inspection, we will attempt to reproduce this process.

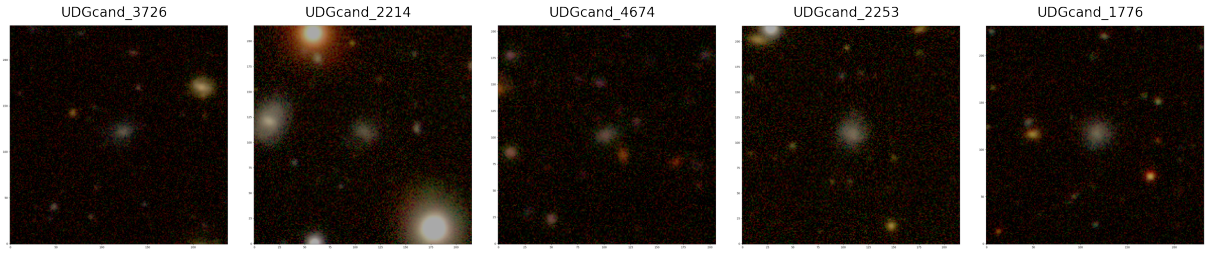
We start with the set of objects classified by users as fluffy galaxies (with a 75% classification threshold), that also survive our reproduction of the paramametric selection cuts, and that do *not* make it into the likely ground truth catalogue. We create 5×5 grids of thumbnail images as shown to the users in Space Fluff, and put them next to grids of the likely ground truth objects, and compare, thereby essentially walking through part of the process the Space Fluff users also experienced.

We do this in two passes. First, we choose any objects that do not appear too bright and show no obvious structure like spirals or strong bulges. After this initial loose filtering, we end up with 95 (of 445) objects that warrant a second inspection. Then, for the second inspection, we narrow the comparison slightly by creating new image grids containing only these images. We take the subset of LGT objects that are at least as bright as the faintest object classified as 'fluffy galaxy' by the users (refer back to previous sections, where we found that objects fainter than 25 mag/arcsec^2 are never classified as galaxies by the majority of users), and visually compare those to the objects that remain after our first pass. We find 22 objects that we believe to show clear enough structure to exclude from cluster membership candidacy. Note that because these objects are so faint, and generally show so little structure, it is quite hard to say with complete certainty that these objects do or do not show clear bulges, for example.

Doing the same thing for the objects that are classified by users as fluffy galaxies if we exclude the first 50 classifications per user, yields an additional 9 objects we believe could be included in the catalogue, bringing the total from 73 to 82. We cross-match the coordinates of these objects to the FDSDC and find that 19 of these 82 objects are actually included in that catalogue already, which leaves 63 objects that we recommend for further inspection. We list the identifiers (names) of all these objects and their thumbnail images in the Appendix. Venhola et al. might have recognized structure in these images, or they may show structure in images from other surveys, or perhaps they found another physical reason to exclude these objects from their final catalogue. We display a small set of these images in figure 22 (refer to the Appendix for thumbnails of all 63 objects), and also plot them in parameter space to compare them to the LGT catalogue in figure 23, and the total set of fluffy galaxies identified by users that survive the selection cuts. We then leave it as an exercise to the reader to make their own judgement: after having seen various examples of images throughout this work (and also in the Appendix, where we include a number of examples from the field guide), do these objects appear similar enough to the LGT catalogue objects to warrant reconsidering their inclusion/exclusion in the catalogue?

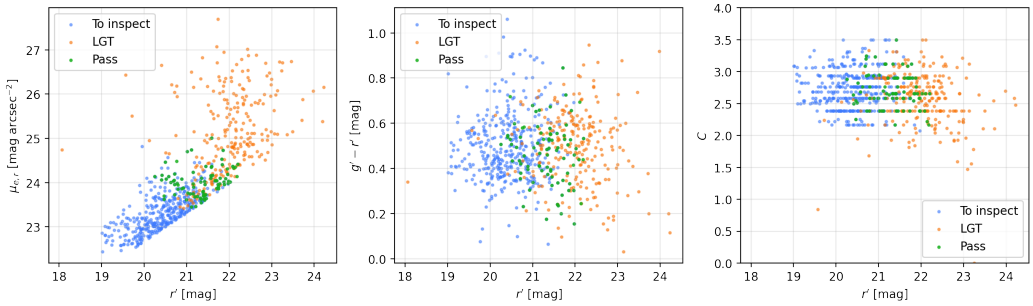


(a) Selection of possible cluster members after our two-pass visual classification.

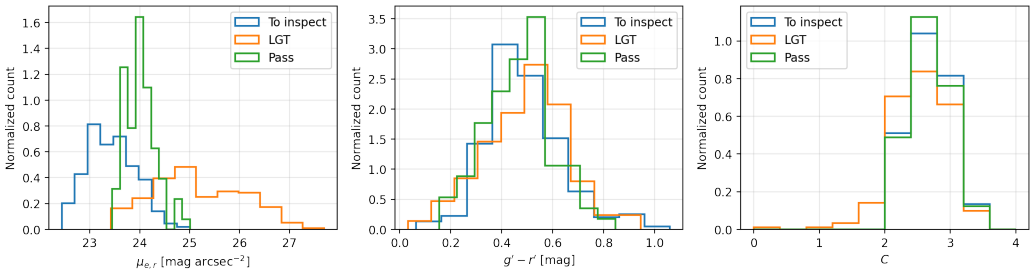


(b) Selection of LGT catalogue objects brighter than $\mu_{e,r} = 25$ mag/arcsec^2.

Figure 22: Comparison of a small selection of image thumbnails of LGT catalogue objects with $\mu_{e,r} \leq 25$ mag/arcsec^2, and objects classified as fluffy galaxies by the users that we believe show no clear enough structure to be excluded from the LGT catalogue. The top row displays a few objects we believe could be reconsidered for inclusion in the LGT catalogue, and the bottom row displays actual LGT catalogue members.



(a) Distributions in parameter space.



(b) Parameter histograms.

Figure 23: Comparison between LGT catalogue, 'fluffy galaxies' selected by the users, and the subset of objects from the 'fluffy galaxies' that we believe could be included in the catalogue based on our 2-pass visual inspection. The blue 'To inspect' points denote the fluffy galaxies as classified by the users, that don't make it into the LGT catalogue. The orange 'LGT' points are all the LGT catalogue objects, and the green 'Pass' objects are the 63 objects we believe might still be candidates for inclusion in the LGT catalogue.

3.6 Classification accuracy of experienced users

An important question in citizen science projects is how important each individual user's contributions should be. An expert or highly-skilled classifier is likely to provide more trustworthy results, however this requires prior knowledge of users and their level of expertise by some quantifiable measure. In Space Fluff, we have no access to any of this information about individual users. One thing we can expand on, however, results from figure 6, where we noted that users that make more classifications typically end up being more precise (among their peers, which does not necessarily correlate to accuracy of cluster member identification). What happens to the overall classification consensus if we account for this trend, by excluding the first few classifications a user makes? In doing this, we assume that, during the initial classifications made by a user, they are still exploring the project, and still have to process the types of images they are seeing to get a feeling for which objects they are likely to put into which category (galaxy/empty center, fluffy/bright, etc.). Depending on the number of classifications we exclude, some users' contributions may be entirely removed from consideration, as there are many users that made only a few classifications.

Figure 24 describes the total amount of classifications remaining per object for an increasing number of excluded classifications per user. We see that, eventually, the number of classifications remaining for some objects falls far enough that we cannot statistically significantly consider the votes cast for that object, which places a limit on how many votes we can realistically exclude. Excluding up to the first 50 classifications per user leaves most objects with over 20 total classifications, and no objects with fewer than 5. Extending the exclusion to the first 250 classifications per user will leave us with some objects receiving only 1 classification, which is not statistically significant.

Fleiss' κ for task 0 increases steadily with increasing levels of exclusion, from 0.67 for $n_{ex} = 0$, to 0.70 for $n_{ex} = 50$. Where, $n_{ex} = x$ denotes that for that subset, each user's first x votes have been discarded.

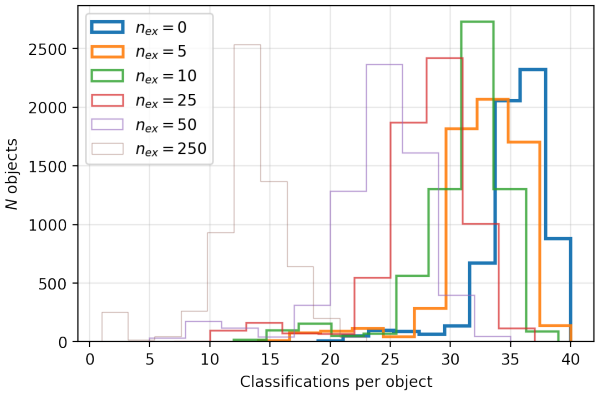


Figure 24: Total number of classifications remaining when excluding the first n_{ex} classifications made by each user.

The overall shape of the distributions for tasks 0 and 1 does not vary appreciably when excluding classifications (see figure 32 in the Appendix), however there are a number of objects for which the user consensus changes when excluding classifications. In figure 25, we compare the sets of objects remaining for exclusion of $n_{ex} = 0$ and $n_{ex} = 50$, and plot the task 0 'galaxy' votes of objects that receive at least 75% of their T0 votes for 'galaxy' in one of these sets, and not in the other. We find that the difference in 'galaxy' votes rarely exceeds 20%. We also note that there are more objects that gain a 'galaxy' consensus if we exclude each user's first 50 votes, than there are galaxies that fall below the 75% threshold if we exclude these 50 votes per user.

In figure 26, we show the percentage of votes received for 'galaxy' and 'fluffy' between the non-exclusive dataset and the $n_{ex} = 50$ dataset, considering now only objects that are classified as fluffy galaxies (75% thresholds for both fluffy and galaxy) in one of the sets, but not in the other. We note that the difference in votes for either of these answers only rarely exceeds 20%, however there is a significant number of objects that end up not being classified as fluffy galaxies in one set, but do end up being classified in the other.

How does the comparison to the likely ground truth (LGT) catalogue change if, instead of including every single classification, we exclude the first n classifications per user? Table 5 describes the number

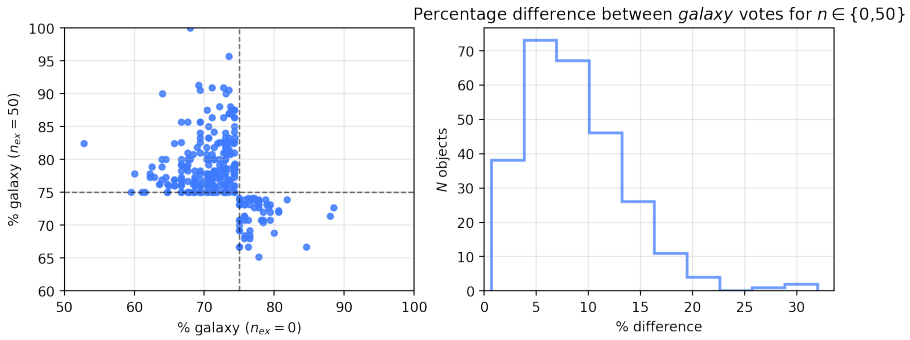


Figure 25: Percentage of 'galaxy' votes for objects that receive at least 75% of T0 votes for 'galaxy' in either the complete dataset, or the dataset in which the first 50 votes per user are filtered, but not in the other subset. The dashed lines at 75% on either axis denote the cutoff we consider as threshold for overall 'galaxy' consensus.

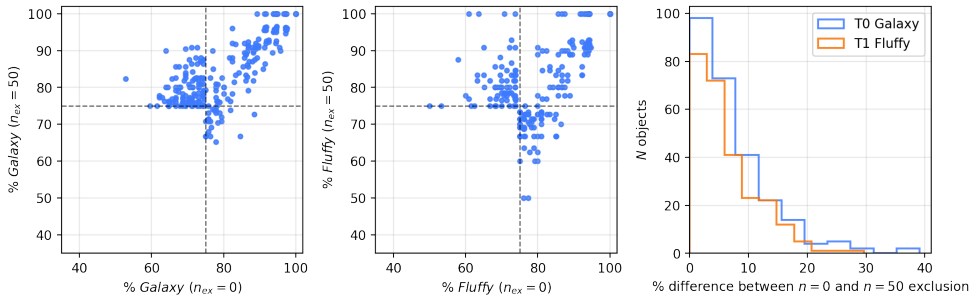


Figure 26: Percentage of votes for 'galaxy' and 'fluffy' for objects that are classified as fluffy galaxies in either the complete dataset, or in the set excluding each user's first 50 classifications.

of LGT objects classified as fluffy galaxies by the users as a function of the number of excluded leading classifications n. We find that increasing the threshold of exclusion yields more and more likely ground truth objects, indicating that users get better at identifying fluffy galaxies as they classify more and more objects. However, the users still come nowhere near identifying all 232 LGT objects. Lowering the threshold for 'galaxy' consensus yields more objects, however this might be a result of selection bias - objects that are considered by more users as empty images instead of galaxies are almost certainly not going to be high-surface brightness objects (as those appear brighter in the image). For the users that then do consider them galaxies, they are almost guaranteed to be considered fluffy instead of bright, which would lead to a 'correct' classification.

n_{ex}	N_{LGT} ($\geq$ 50% galaxy threshold)	N_{LGT} ($\geq$ 75% galaxy threshold)
0	73 (4.5%)	30 (2.9%)
5	74 (4.6%)	33 (3.1%)
25	74 (4.4%)	35 (3.1%)
50	76 (4.6%)	38 (3.4%)
250	97 (5.5%)	48 (4.0%)

Table 5: Number of leading votes per user excluded (n_{ex}) versus likely ground truth (LGT) catalogue objects 'retrieved' (N_{LGT}), i.e. classified by the users as fluffy galaxies (fluffiness threshold of 75%, galaxy threshold of either 50% or 75% for comparison). The percentages next to the number of LGT objects retrieved in each column denotes that number as a percentage of the total number of fluffy galaxies in that subset. Higher is better, and ideally this would be 100%, assuming the LGT catalogue is complete and ideal.

Are the LGT objects found by users across these subsets of classifications the same objects? Comparing the $n_{ex} = 0$ subset to the $n_{ex} = 50$ one, we find that of the 30 objects classified as fluffy galaxies when $n_{ex} = 0$, 29 are also found by the $n_{ex} = 50$ subset, meaning there is significant overlap, and we can consider the $n_{ex} = 50$ subset to be a better sample. For the $n_{ex} = 250$ set, again 29 are found that

were also found when $n_{ex} = 0$, meaning this subset is an even better selection. Thus, we find that users that spend more time in the project generally do correctly classify a larger subsample Fornax LSB galaxies.

As an additional comparison, we can also choose to include *all* classifications, but only those done by the so-called power users in the project. If we define a power-user to be someone who makes at least 250 classifications, then the results are similar to the results $n_{ex} = 250$, but the accuracy compared to the LGT catalogue is slightly lower. The users retrieve 41 LGT objects, which is 3.5% of all the fluffy galaxies they find (compare to table 5). This means these results lie between the $n_{ex} = 50$ and $n_{ex} = 250$ subsets, which makes sense, since in this case we only consider the users that made at least 250 classifications, but now we include their first 250 classifications. If we assume (by using the trends we found previously, like the results from figure 6) that users get better as they classify more and more objects, this naturally would also apply to power-users, and excluding their first classifications would also yield a better accuracy than including these first classifications.

4 Results and discussion

Space Fluff tasked users with identifying an object from the Fornax Deep Survey into one of three categories, based off a single image of the object: (1) galaxies, (2) groups of objects, (3) empty images/artifacts. Our main interest are the subset of galaxies we wish to have classified as 'fluffy', as we consider those to be LSB galaxies in the Fornax cluster. These objects generally have much less structure than, for example, the spiral galaxies studied in the immensely popular Galaxy Zoo spiral galaxies citizen science project. Because of this general lack of structure, combined with their low surface brightness, the assumption is that it will be harder for the untrained eye to identify these objects as galaxies. In this work, we analyzed the classification behavior of the Space Fluff volunteer users to find correlations between the user consensus and the (photometric) properties of the objects.

4.1 Correlation of classification behavior and photometric properties

Of the 922 objects in the project that do not have photometric properties assigned to them in the unfiltered catalogue of possible cluster candidates, produced by Venhola et al. (*in prep.*), none are generally considered by the users to be fluffy galaxies. Very few are even considered galaxies at all (only 24 of these objects are classified as galaxies by at least half the users that classified them, and none of them are classified as galaxies by 90% or more of the users).

For objects considered galaxies by the users, we will describe the correlations between photometry and user classifications, below.

4.1.1 *Magnitude*

We find a moderate correlation between an object's r' magnitude and the percentage of votes an object receives in task 0 for 'galaxy', with higher (fainter) magnitudes generally leading to a lower percentage of galaxy votes.

4.1.2 *Color*

The $g' - r'$ color of an object is strongly correlated to the color perceived by the users. An increase in $g' - r'$ leads to an increase in user votes for "Red/yellow", and a decrease in $g' - r'$ leads to an increase for "White/blue" votes.

Another strong correlation exists between $g' - r'$ color and consensus on whether a galaxy is fluffy or bright. We find a stronger correlation here than we do when comparing the actual mean effective surface brightness to fluffy/bright votes. We find that objects that are more yellow/red are typically considered bright, whereas white/blue objects are considered fluffy more often.

4.1.3 *Surface brightness*

Across the whole dataset, we do not find a strong correlation between an object's mean r-band effective surface brightness and the likelihood of users forming a consensus on whether or not an object is a fluffy galaxy. The faintest objects in the dataset are most of the time not considered galaxies by the majority of

users. Over 90% of the objects in the likely ground truth set have surface brightness $\mu_{e,r} \geq 24$ mag/arcsec^2, but for objects classified by users as fluffy galaxies, this is only 8%. Only 7 of the objects with $\mu_{e,r} > 25$ mag/arcsec^2 in the LGT catalogue are classified as galaxies by the majority of users, even though over half of the likely ground truth catalogue is this faint or fainter.

The fact that users do not consider the faintest objects to be galaxies may be due to image quality and resolution, which might make it hard to distinguish between a true empty image and a very faint image. It might also be an inherent limitation to what the users consider a galaxy. Placing a stronger emphasis in the training session the users undergo on the fact that these very faint objects are indeed very faint galaxies, instead of for example artifacts, might lead to a more accurate classification for these faintest objects.

4.2 Accuracy of cluster member classifications

We find that the majority of the likely ground truth catalogue objects do not reach a majority fluffy galaxy consensus. Over two-thirds of these objects are not even considered galaxies by the majority of the users that classified them.

232 likely ground truth objects were included in the Space Fluff project. Of these, only 30 are accurately identified as fluffy galaxies by the users in accordance with the thresholds we set for classification (at least 75% of votes for 'galaxy', and then at least 75% of votes for 'fluffy'). However, among the users that do classify these catalogue objects as galaxies, the consensus generally is that they are fluffy galaxies rather than bright ones. If we decrease the threshold for the acceptance of a galaxy consensus to 50% (a simple majority), the amount of catalogue targets classified as fluffy galaxies more than doubles, to 73. However, we must note that there is a large number of objects that the users classify as fluffy galaxies even though they are not in the likely ground truth catalogue. Depending on the classification threshold, approximately half of these can be ruled out from Fornax cluster membership on the account of being too red, too bright, or too concentrated. Manual classification done by Venhola et al. (in [25] and the unpublished paper that produces the likely ground truth catalogue we refer to throughout this work) rules out the remaining few hundred objects on the basis of morphology.

4.2.1 *Experienced users*

We find that as users gain more experience classifying objects in the project, they become more adept at correctly classifying likely ground truth catalogue objects as fluffy galaxies. Excluding up to the first 250 classifications per user (thereby also completely discarding many users' classifications in the case that those users made fewer than 250 total classifications) yields an increasing number of LGT objects classified as fluffy galaxies, however the total number of votes per object decreases simultaneously, leading to a decrease in statistical significance of these remaining classifications. It is hard to balance the benefit of the increased accuracy to the decreased significance, as we do not have enough information on the expertise of individual users to weigh their classifications based on any other metric.

4.3 Manual classification

We performed a manual classification of all the objects that are classified by users as fluffy galaxies, and that survive the parametric selection cuts, and find that, of these 450 or so objects, 63 look similar enough to likely ground truth objects in our opinion that they could not be excluded from cluster membership based on visual identification alone. This means that we do not believe clear spiral structure and bulges are present in these images. Whether Venhola et al. excluded these objects based on some other criterion that is outside of the scope of this thesis, or if they did judge them to show clear signs of spirals or bulges is not known to us. Therefore we present this set of objects as candidate Fornax cluster LSBs, with the caveat that a single visual classification should not suffice to conclusively include them, just that further inspection might be warranted.

4.4 Suggestions for similar future projects

There are a few factors that limit the depth of this analysis that could be mitigated in a similar project involving contributions from citizen scientists:

- The lack of knowledge about the users. Reconsidering the level of balance between the exposure of personally identifiable information (PII) and useful background information of a user's level of training or expertise would be extremely useful in weighting the classifications done by an individual user against those by other users.
- The lack of a sizeable expert (also called 'gold standard') classification dataset to compare the volunteer classifications to, limits the extent to which we can determine the accuracy of users. Instead, we compared our results to a single catalogue without individually labeled reasons for exclusion of objects. In our project, we were left with a number of objects classified by the users as fluffy galaxies that were subsequently ruled out from cluster membership on the basis of a manual morphological classification by Venhola et al. (*in prep.*). A distinctly labeled set of further criteria that determines whether or not an object that survives the selection cuts actually becomes included in the final catalogue, would not only be helpful in the analysis of a project like this after its execution, but could also be very useful as additional training information for any users willing to partake in classification.
- Other aspects of the training procedure might also be improved. A more diverse presentation of labeled objects might prevent the situation we encountered in our analysis, where users apparently judge the fluffiness or brightness of an object based on its color, rather than other aspects like surface brightness.
- The fact that only a single image was presented per object severely limits the decision process a user undergoes when classifying an object. The effect of this is likely to be more noticeable in a project like Space Fluff, where the objects of interest already have rather few features. Some features that guide the user in making a classification might be different in another filter, so including images from several (combinations of) filters and levels of contrast would help the user make a more involved decision.

5 Conclusions and summary

In this work we analyzed the classifications made by volunteer users of possible low-surface brightness galaxies in the Fornax cluster. Based off only a single image containing an object in the Fornax Deep Survey, extracted by the Max-Tree Objects algorithm, users were asked to decide whether that object was a galaxy, and subsequently whether that galaxy appeared to them as 'fluffy' or 'bright', where the fluffiness or brightness of an object supposedly distinguishes between a brighter galaxy simply appearing in the line-of-sight within the image, and a fluffy galaxy being a low-surface brightness galaxy within the Fornax cluster.

We compared the classifications made by the users as a group to a likely ground truth catalogue of 265 low-surface brightness galaxies, of which 232 were present in the Space Fluff project for users to attempt to identify, among some 6,000 other objects that were other galaxies not selected by Venhola et al. for their final catalogue on the basis of morphology and photometric parameter cuts. We compared results depending on various selection criteria, like the thresholds we use for classification (e.g. if half of the users classify something as a galaxy, we can consider it a galaxy, but we can also put this limit at a more stringent 75% in accordance with other citizen science projects), or an exclusion of each user's first few classifications, where we consider these initial classifications to be part of the user's learning process as they familiarize themselves with the project.

We find that experienced users are more adept at identifying the likely ground truth objects as fluffy galaxies, retrieving between 30 and 48 of the 232 likely ground truth objects as fluffy galaxies when taking into account each classification, or only those after each user's 50th classification respectively. We find that users experience difficulty recognizing the lowest surface brightness galaxies as galaxies, instead they believe many of these images to contain no galaxy at all. This is due to the extreme faintness of these galaxies, which indeed often makes them hard to spot in the image. Of all objects classified as fluffy by the users, approximately 69% are brighter than the brightest object in the likely ground truth catalogue.

Based on various selection cuts, determined by a physical argument derived from properties of spectrally confirmed Fornax cluster members, approximately half of the fluffy galaxies selected by users can be ruled out from cluster membership. Depending on the exact selection criteria we apply to the user classifications, this leaves a few hundred objects as possible cluster members, which are subsequently ruled out by Venhola et al. on the basis of morphology (presence of spirals, strong bulges). We performed our own visual classification of these few hundred galaxies and find 73 faint galaxies that we believe do not show strong enough structure to be excluded from the likely ground truth catalogue purely based on visual classification.

The lack of prior knowledge about individual users, and also the lack of an expert set of example classifications, makes it hard to categorize the remaining thousands of objects in the dataset, or to determine in a statistical framework the overall accuracy of classifications.

Any future project involving visual inspection of low-surface brightness galaxies is recommended to place a stronger focus on the initial training provided to users. In our project, we found that the training images provided appeared to show a bias towards yellow/red objects as being 'bright', which translated directly into the relation between the astronomical color of an object and how likely a user was to classify it as bright, where ideally we would instead want this classification to happen based on the surface brightness of the object. Providing users with more options for comparison, like images in various filters, or varying levels of contrast, or even a comparison tool where they can more easily directly compare the image they are looking at to other objects they have already seen, may help the user make a more involved decision in their classification process, and ultimately possibly lead to better classification accuracy.

Concluding, we find that users accurately classify the non-galaxy objects in the dataset, but they experience difficulty identifying the fainter galaxies, which unfortunately means that the vast majority of likely ground truth objects are not classified as galaxies, and thus we do not obtain a complete catalogue of Fornax low-surface brightness galaxies using this citizen science process. We leave a recommendation for further expert classification of a small subset of galaxies we believe might be included in a LSB catalogue, as they survive parametric selection cuts, are classified by users as fluffy galaxies, and appear to us as not showing any obvious morphological properties that would exclude them from cluster membership.

References

[1] E. Aceves-Bueno, A. S. Adeleye, M. Feraud, Y. Huang, M. Tao, Y. Yang, and S. E. Anderson. The accuracy of citizen science data: A quantitative review. *The Bulletin of the Ecological Society of America*, 98(4):278–290, 2017. doi: https://doi.org/10.1002/bes2.1336. URL https://esajournals.onlinelibrary.wiley.com/doi/abs/10.1002/bes2.1336.

[2] P. Bhattacharjee, P. Majumdar, M. Das, S. Das, P. S. Joarder, and S. Biswas. Multiwavelength analysis of low surface brightness galaxies to study possible dark matter signature. *Monthly Notices of the Royal Astronomical Society*, 501(3):4238–4254, Dec 2020. ISSN 1365-2966. doi: 10.1093/mnras/staa3877. URL http://dx.doi.org/10.1093/mnras/staa3877.

[3] D. W. Darg, S. Kaviraj, C. J. Lintott, K. Schawinski, M. Sarzi, S. Bamford, J. Silk, R. Proctor, D. Andreescu, P. Murray, R. C. Nichol, M. J. Raddick, A. Slosar, A. S. Szalay, D. Thomas, and J. Vandenberg. Galaxy Zoo: the fraction of merging galaxies in the SDSS and their morphologies. *Monthly Notices of the Royal Astronomical Society*, 401(2):1043–1056, 01 2010. ISSN 0035-8711. doi: 10.1111/j.1365-2966.2009.15686.x. URL https://doi.org/10.1111/j.1365-2966.2009.15686.x.

[4] D. A. Fischer, M. E. Schwamb, K. Schawinski, C. Lintott, J. Brewer, M. Giguere, S. Lynn, M. Parrish, T. Sartori, R. Simpson, and et al. Planet hunters: the first two planet candidates identified by the public using the kepler public archive data. *Monthly Notices of the Royal Astronomical Society*, 419(4):2900–2911, Nov 2011. ISSN 0035-8711. doi: 10.1111/j.1365-2966.2011.19932.x. URL http://dx.doi.org/10.1111/j.1365-2966.2011.19932.x.

[5] A. Frebel, E. N. Kirby, and J. D. Simon. Linking dwarf galaxies to halo building blocks with the most metal-poor star in sculptor. *Nature*, 464(7285):72–75, Mar 2010. ISSN 1476-4687. doi: 10.1038/nature08772. URL https://doi.org/10.1038/nature08772.

[6] Haigh, Caroline, Chamba, Nushkia, Venhola, Aku, Peletier, Reynier, Doorenbos, Lars, Watkins, Matthew, and Wilkinson, Michael H. F. Optimising and comparing source-extraction tools using objective segmentation quality criteria. *A&A*, 645:A107, 2021. doi: 10.1051/0004-6361/201936561. URL https://doi.org/10.1051/0004-6361/201936561.

[7] C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. F. del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, and T. E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, Sept. 2020. doi: 10.1038/s41586-020-2649-2. URL https://doi.org/10.1038/s41586-020-2649-2.

[8] Iodice, E., Spavone, M., Capaccioli, M., Peletier, R. F., van de Ven, G., Napolitano, N. R., Hilker, M., Mieske, S., Smith, R., Pasquali, A., Limatola, L., Grado, A., Venhola, A., Cantiello, M., Paolillo, M., Falcon-Barroso, J., D´Abrusco, R., and Schipani, P. The fornax deep survey with the vst - v. exploring the faintest regions of the bright early-type galaxies inside the virial radius. *A&A*, 623:A1, 2019. doi: 10.1051/0004-6361/201833741. URL https://doi.org/10.1051/0004-6361/201833741.

[9] A. Lanteri. Data analysis for spacefluff. URL https://github.com/hwiks/spacefluff. Accessed 2021-05-06.

[10] C. J. Lintott, K. Schawinski, A. Slosar, K. Land, S. Bamford, D. Thomas, M. J. Raddick, R. C. Nichol, A. Szalay, D. Andreescu, P. Murray, and J. Vandenberg. Galaxy Zoo: morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey*. *Monthly Notices of the Royal Astronomical Society*, 389(3):1179–1189, 09 2008. ISSN 0035-8711. doi: 10.1111/j.1365-2966.2008.13689.x. URL https://doi.org/10.1111/j.1365-2966.2008.13689.x.

[11] G. Martin, S. Kaviraj, C. Laigle, J. E. G. Devriendt, R. A. Jackson, S. Peirani, Y. Dubois, C. Pichon, and A. Slyz. The formation and evolution of low-surface-brightness galaxies. *Monthly Notices of the Royal Astronomical Society*, 485(1):796–818, Feb 2019. ISSN 1365-2966. doi: 10.1093/mnras/stz356. URL http://dx.doi.org/10.1093/mnras/stz356.

[12] S. McGaugh. Dwarf and Low Surface Brightness Galaxies. In R. Bender and R. L. Davies, editors, *New Light on Galaxy Evolution*, volume 171, page 97, Jan. 1996.

[13] R. Peletier and FDS Team. FDS - The Fornax Ultra-Deep Imaging Survey: Evolution of Dwarf Galaxies. In *VST in the Era of the Large Sky Surveys*, page 40, June 2018. doi: 10.5281/zenodo.1303950.

[14] M. A. Raj, E. Iodice, N. R. Napolitano, M. Hilker, M. Spavone, R. F. Peletier, H.-S. Su, J. Falcon-Barroso, G. van de Ven, M. Cantiello, D. Kleiner, A. Venhola, S. Mieske, M. Paolillo, M. Capaccioli, and P. Schipani. The fornax deep survey with vst x. the assembly history of the bright galaxies and intra-group light in the fornax a subgroup. *Astronomy & Astrophysics*, 640, 2020. ISSN 0004-6361. doi: 10.1051/0004-6361/202038043. URL https://doi.org/10.1051/0004-6361/202038043.

[15] A. Sandage and B. Binggeli. Studies of the Virgo cluster. III. A classification system and an illustrated Atlas of Virgo cluster dwarf galaxies. , 89:919–931, July 1984. doi: 10.1086/113588.

[16] J. M. Schombert, R. A. Pildis, J. A. Eder, and J. Oemler, Augustus. Dwarf Spirals. , 110:2067, Nov. 1995. doi: 10.1086/117669.

[17] A. M. Smith, S. Lynn, M. Sullivan, C. J. Lintott, P. E. Nugent, J. Botyanszki, M. Kasliwal, R. Quimby, S. P. Bamford, L. F. Fortson, K. Schawinski, I. Hook, S. Blake, P. Podsiadlowski, J. Jönsson, A. Gal-Yam, I. Arcavi, D. A. Howell, J. S. Bloom, J. Jacobsen, S. R. Kulkarni, N. M. Law, E. O. Ofek, and R. Walters. Galaxy zoo supernovae. *Monthly Notices of the Royal Astronomical Society*, 412(2):1309–1319, 2011. doi: https://doi.org/10.1111/j.1365-2966.2010.17994.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-2966.2010.17994.x.

[18] R. Smith, S. Phillipps, J. B. Jones, R. A. H. Morris, R. M. Smith, M. J. Drinkwater, and A. M. Karick. Infrared surface photometry of dwarf galaxies in fornax. *Monthly Notices of the Royal Astronomical Society*, 420(4):3412–3426, 2012. doi: https://doi.org/10.1111/j.1365-2966.2011.20266.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-2966.2011.20266.x.

[19] M. Spavone, E. Iodice, G. van de Ven, J. Falcon-Barroso, M. A. Raj, M. Hilker, R. P. Peletier, M. Capaccioli, S. Mieske, A. Venhola, N. R. Napolitano, M. Cantiello, M. Paolillo, and P. Schipani. The fornax deep survey with vst viii. connecting the accretion history with the cluster density. *Astronomy & Astrophysics*, 639, 2020. ISSN 0004-6361. doi: 10.1051/0004-6361/202038015. URL https://doi.org/10.1051/0004-6361/202038015.

[20] Sundial. Sundial - about. URL https://www.astro.rug.nl/~sundial/. Accessed 2021-05-31.

[21] Sundial. Space fluff, 2021. URL https://www.zooniverse.org/projects/sundial-itn/space-fluff. Accessed 2021-05-31.

[22] P. Teeninga, U. Moschini, S. C. Trager, and M. H. Wilkinson. Improved detection of faint extended astronomical objects through statistical attribute filtering. In *International Symposium on Mathematical Morphology and Its Applications to Signal and Image Processing*, pages 157–168. Springer, 2015.

[23] The pandas development team. pandas-dev/pandas: Pandas, Feb. 2020. URL https://doi.org/10.5281/zenodo.3509134.

[24] E. Tolstoy. Dwarf galaxies: Important clues to galaxy formation. *Astrophysics and Space Science*, 284 (2):579–588, Apr 2003. ISSN 1572-946X. doi: 10.1023/A:1024006006003. URL https://doi.org/10.1023/A:1024006006003.

[25] A. Venhola, R. Peletier, E. Laurikainen, H. Salo, E. Iodice, S. Mieske, M. Hilker, C. Wittmann, T. Lisker, M. Paolillo, and et al. The fornax deep survey with the vst. *Astronomy Astrophysics*, 620: A165, Dec 2018. ISSN 1432-0746. doi: 10.1051/0004-6361/201833933. URL http://dx.doi.org/10.1051/0004-6361/201833933.

[26] Venhola, Aku, Peletier, Reynier, Laurikainen, Eija, Salo, Heikki, Lisker, Thorsten, Iodice, Enrichetta, Capaccioli, Massimo, Kleijn, Gijs Verdoes, Valentijn, Edwin, Mieske, Steffen, Hilker, Michael, Wittmann, Carolin, van de Ven, Glenn, Grado, Aniello, Spavone, Marilena, Cantiello, Michele, Napolitano, Nicola, Paolillo, Maurizio, and Falcón-Barroso, Jesús. The fornax deep survey with vst - iii. low surface brightness dwarfs and ultra diffuse galaxies in the center of the fornax cluster. *A&A*, 608:A142, 2017. doi: 10.1051/0004-6361/201730696. URL https://doi.org/10.1051/0004-6361/201730696.

[27] Zooniverse. Galaxy zoo, 2021. URL https://www.zooniverse.org/projects/zookeeper/galaxy-zoo/. Accessed 2021-05-31.

[28] Zooniverse. Galaxy zoo: About, 2021. URL https://www.zooniverse.org/about/publications. Accessed 2021-05-31.

6 Appendix

6.1 Space Fluff user tutorial

The tutorial or 'field guide' itself is available on the Space Fluff page on Zooniverse[3], but we will include a few images below of objects that are intended (according to this field guide) to be classified a certain way, on the next few pages, in figures 28 and 29. Figure 27 describes the workflow of the project as described on the Space Fluff Zooniverse.

[3] https://www.zooniverse.org/projects/sundial-itn/space-fluff/classify

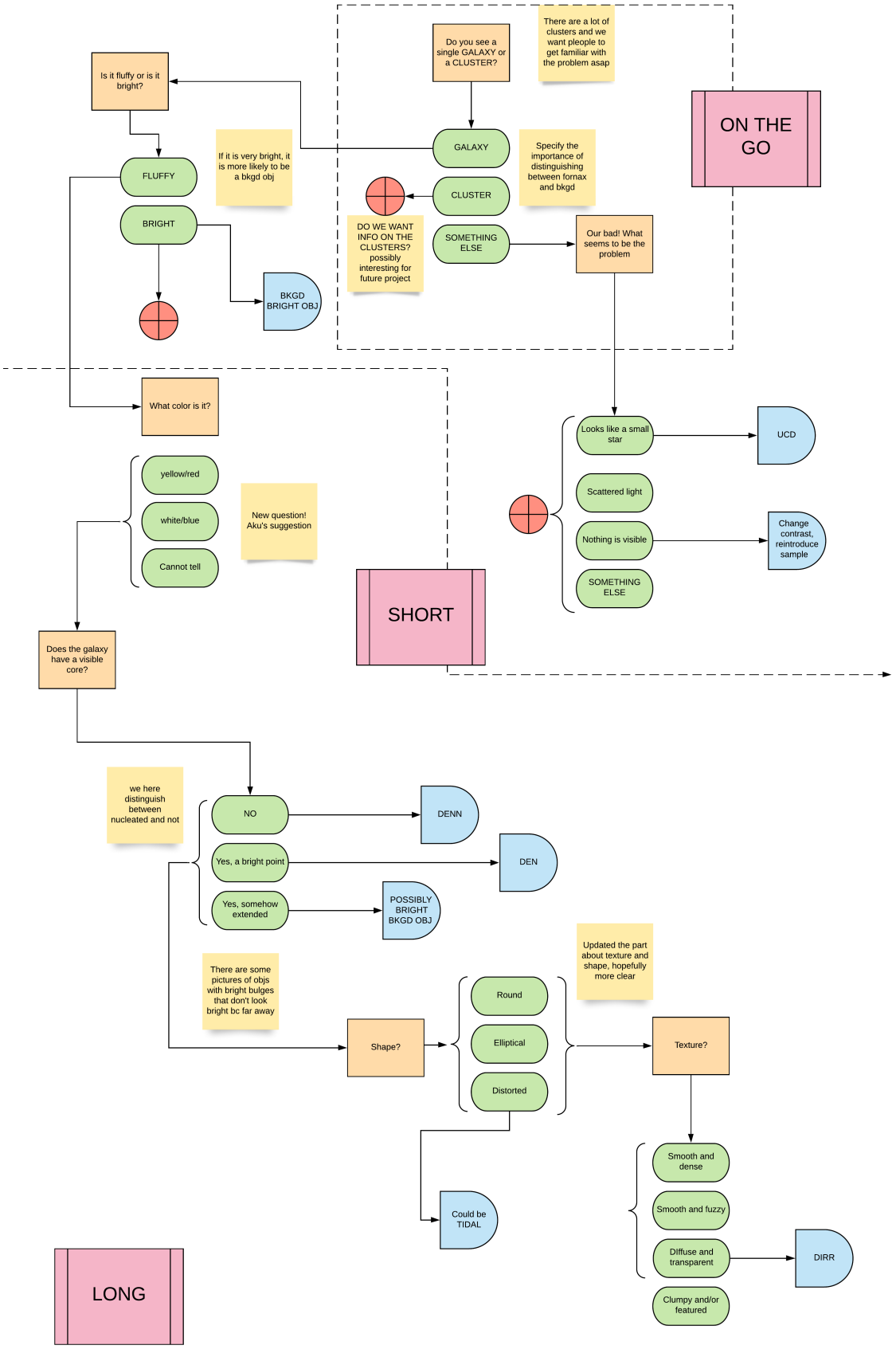


Figure 27: Space Fluff workflow diagram, from [21] (best viewed in color). Rounded green boxes denote the unique answer to each orange box, which indicates one of the questions presented to the users (also called 'tasks' hereafter).

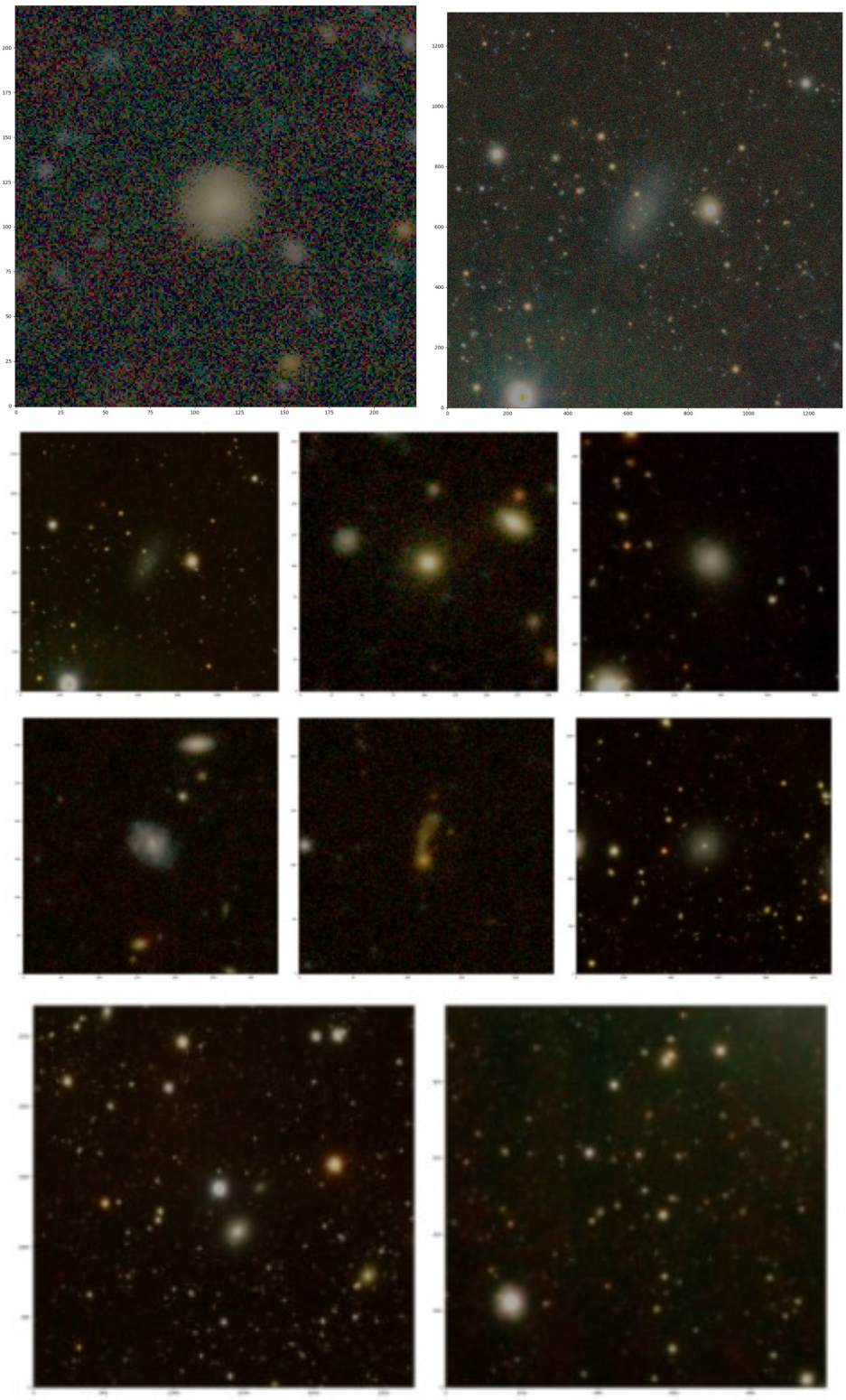


Figure 28: Examples from Space Fluff field guide of objects intended to be classified as galaxies

One of the trickiest questions to answer is: does the galaxy look 'fluffy' or 'bright'?

For 'fluffy' we mean a dim, faint,fuzzy looking object with a smooth light gradient. Something like this:

This are the kind of objects we are looking for in the first place! They can also look fainter, like this:

When we say 'bright' on the other hand, we mean an object that looks more like an orb of light. Here are some examples:

Sometimes however it is not so clear cut. Sometimes you get images looking like this:

In this cases, you can ask yourself: does it look like a bright object (see above), that looks faint due to the distance? As in, is the image grainier that usual, with less clear edges (like in the case of the last one)? Then you can still select 'bright'. Otherwise, if it looks like an example of 'fluffy' with an hyperactive core, select 'fluffy' anyway.

This question has the purpose to help us select those objects that are bright and in the background of Fornax, and exclude them from the longer workflow. We prefer to err on the side of caution and keep more than needed.

That said, your best guess is always the best answer!

Figure 29: Field guide section on classification of fluffy and bright galaxies, taken directly from the Space Fluff Zooniverse page [21].

6.2 Remaining plots for completeness

Figure 30 belongs to section 3.2.2 and displays the relation between task 0 'galaxy' votes an object receives and its r' magnitude.

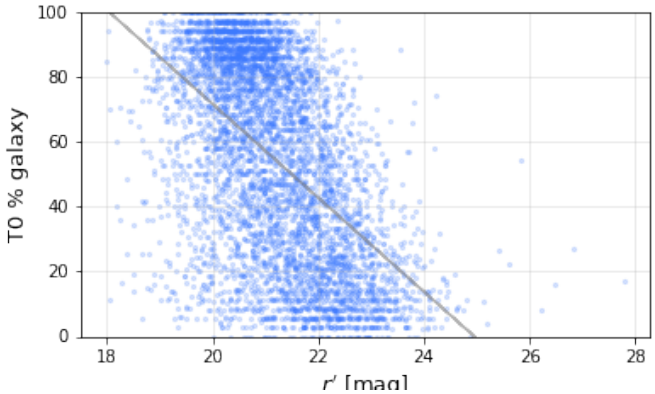


Figure 30: Comparison between r' magnitude and percent votes 'galaxy' per object. The grey line indicates a linear fit, mainly to guide the eye. The Pearson correlation coefficient for this relation is -0.53.

Figure 31 displays the few objects that have at least five T2 votes (which asks users about galaxy color), that have $g' - r' < 0.5$, and are voted by more than half their users as 'red/yellow'.

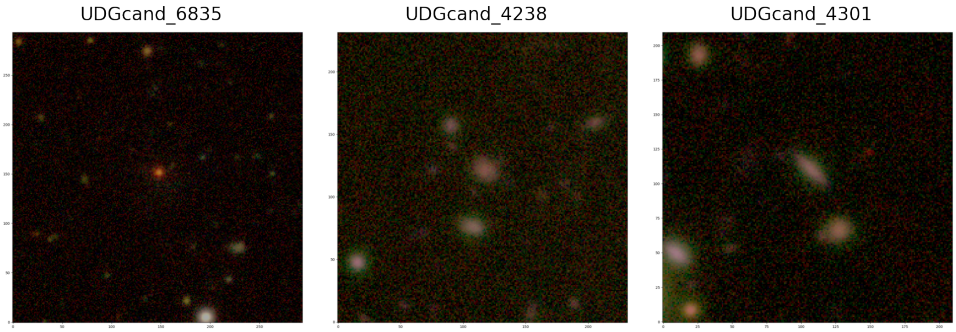
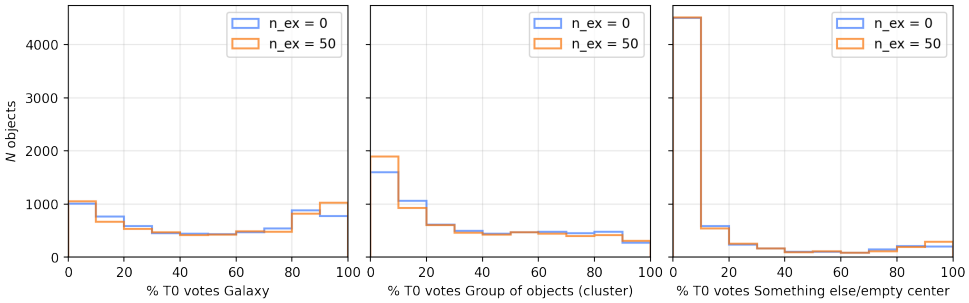
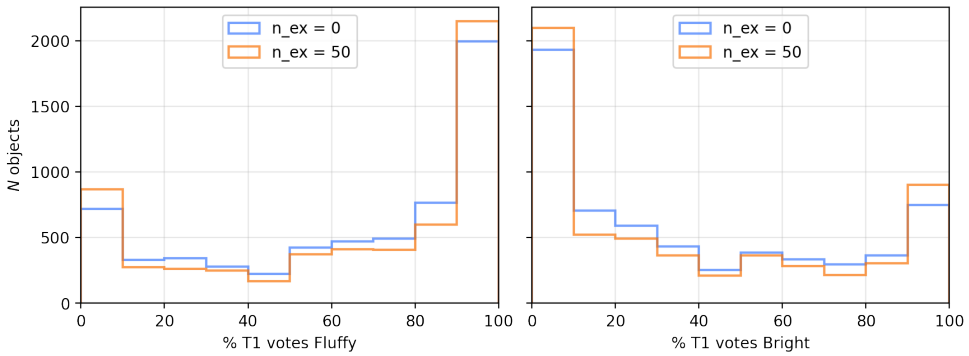


Figure 31: Objects with $g' - r' < 0.5$ that have at least have half their T2 votes in favor of 'red/yellow'.

Figure 32 belongs to section 3.6, it compares vote distribution for tasks 0 and 1 between the complete dataset and a filtered dataset (where we exclude each user's first 50 classifications).



(a) Task 0 answers



(b) Task 1 answers

Figure 32: Vote distributions for task 0 and 1 answers, comparing the set of all classifications ($n_{ex} = 0$) to the set we obtain if we exclude each user's first 50 classifications ($n_{ex} = 50$).

6.3 List of possible Fornax LSBs as resulting from our visual classification

Below is a list of object identifiers for the objects we believe look similar enough to likely ground truth objects that they cannot be ruled out from cluster membership based only on visual identification. The actual names of the objects in the Space Fluff project, then, are "UDGcand_n", with n being the identifier number from the list below. Figures 33, 34 and 35 on the next few pages display the thumbnail images of each of these objects. Figures 36 and 37 on the pages thereafter display thumbnail images of objects that are already in the LGT catalogue, for comparison. We state for emphasis that the set of images of objects in the LGT catalogue that we present here is a random sample, which also includes objects with surface brightnesses of ≥ 25 mag/arcsec^2. For a more direct comparison between objects we still recommend for possible inclusion in the catalogue, and those actually in the LGT catalogue of similar surface brightness, refer to figure 22.

141	3729	2077
106	2363	3185
136	1150	4341
163	6685	3279
282	1057	3333
226	4703	609
1436	3410	4557
4805	412	5810
1719	2497	4276
3195	3226	2105
1820	2182	1313
2810	6129	6058
432	4904	5994
5121	637	1837
1212	4331	3194
3736	2109	6253
1571	3253	1979
6747	6024	4870
7127	434	2339
5142	5242	7069
1002	4966	7051

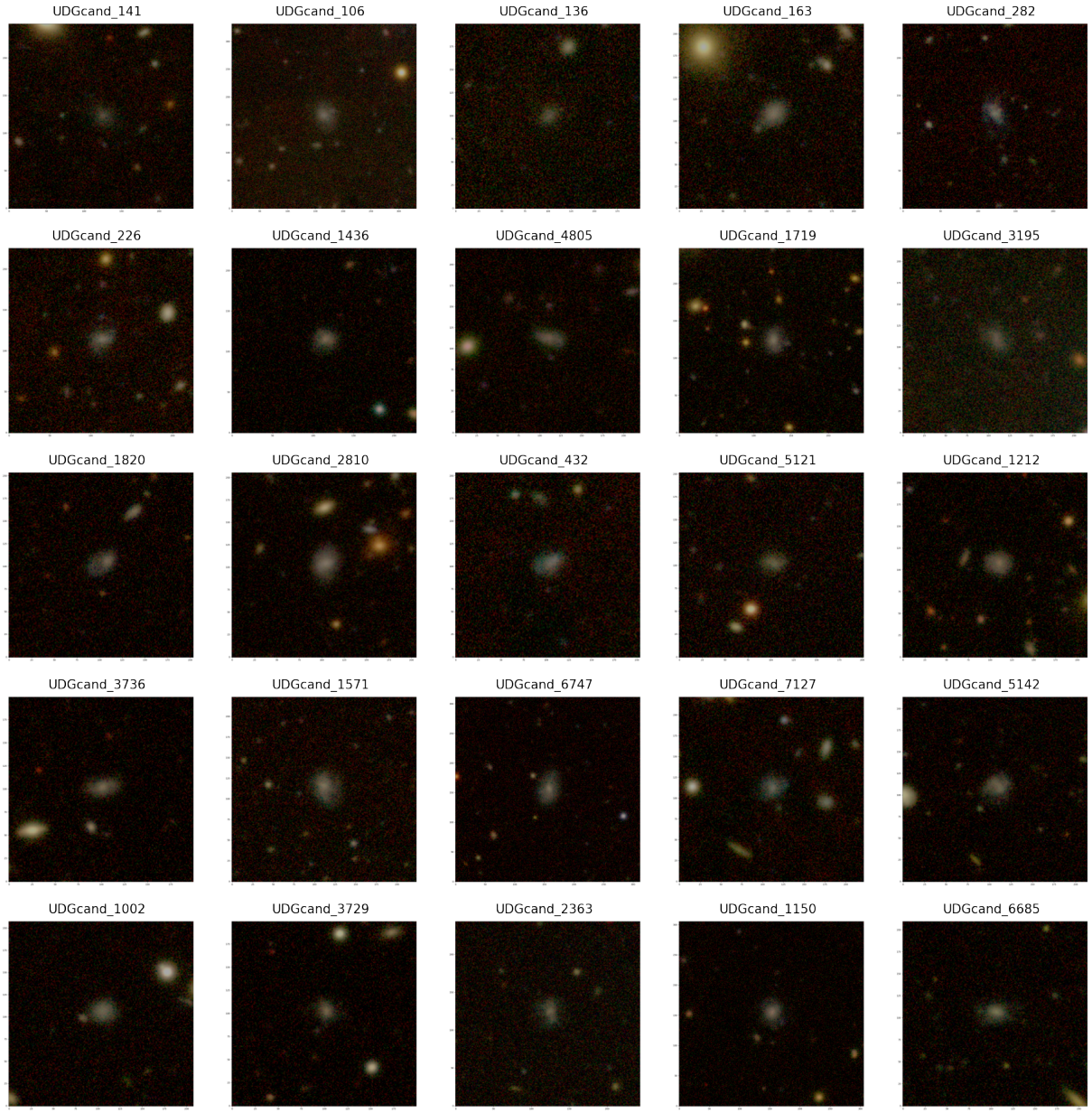


Figure 33: (1/3) Thumbnail images of objects we believe are faint and structure-less enough to possibly still be included in the LGT catalogue.

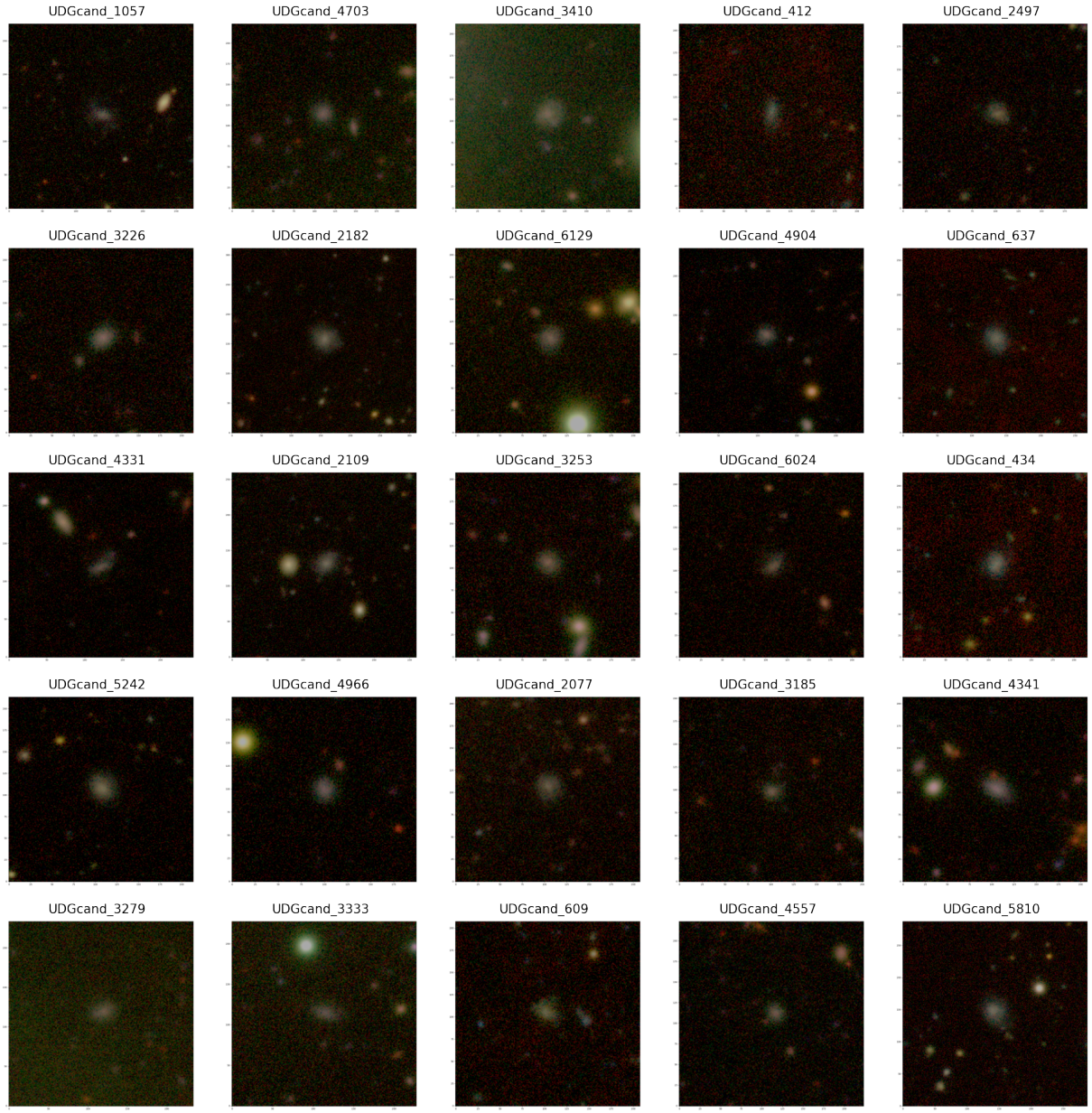


Figure 34: (2/3) Thumbnail images of objects we believe are faint and structure-less enough to possibly still be included in the LGT catalogue.

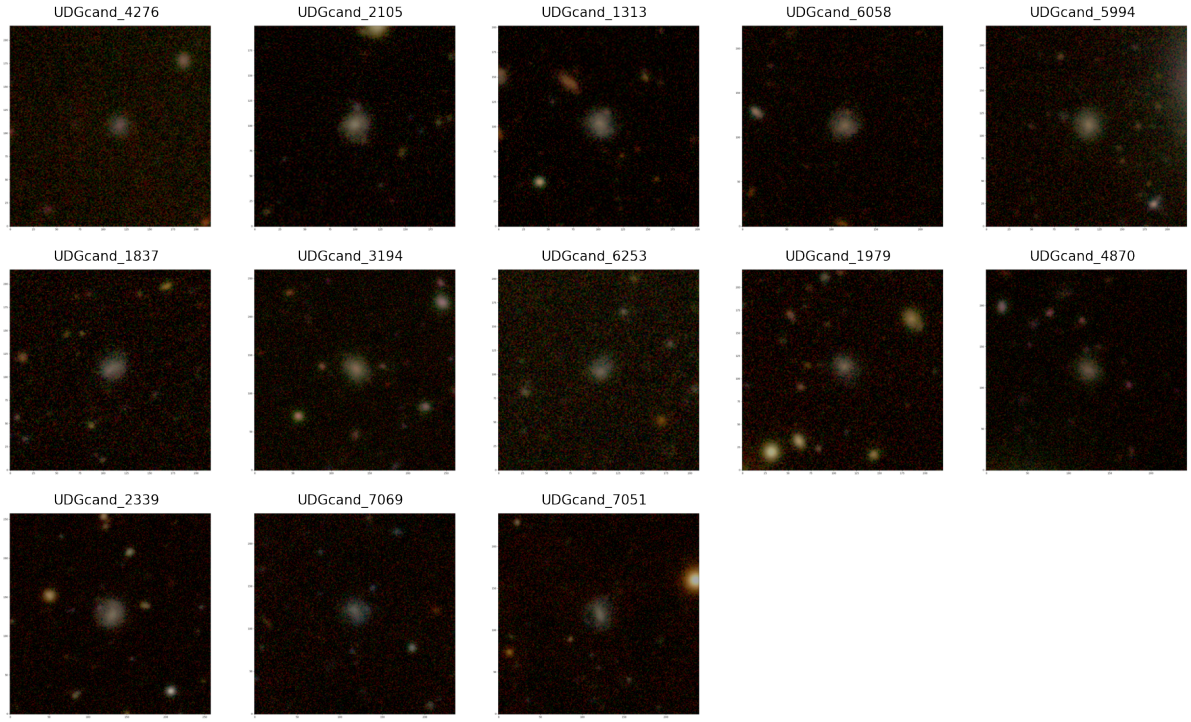


Figure 35: (3/3) Thumbnail images of objects we believe are faint and structure-less enough to possibly still be included in the LGT catalogue.

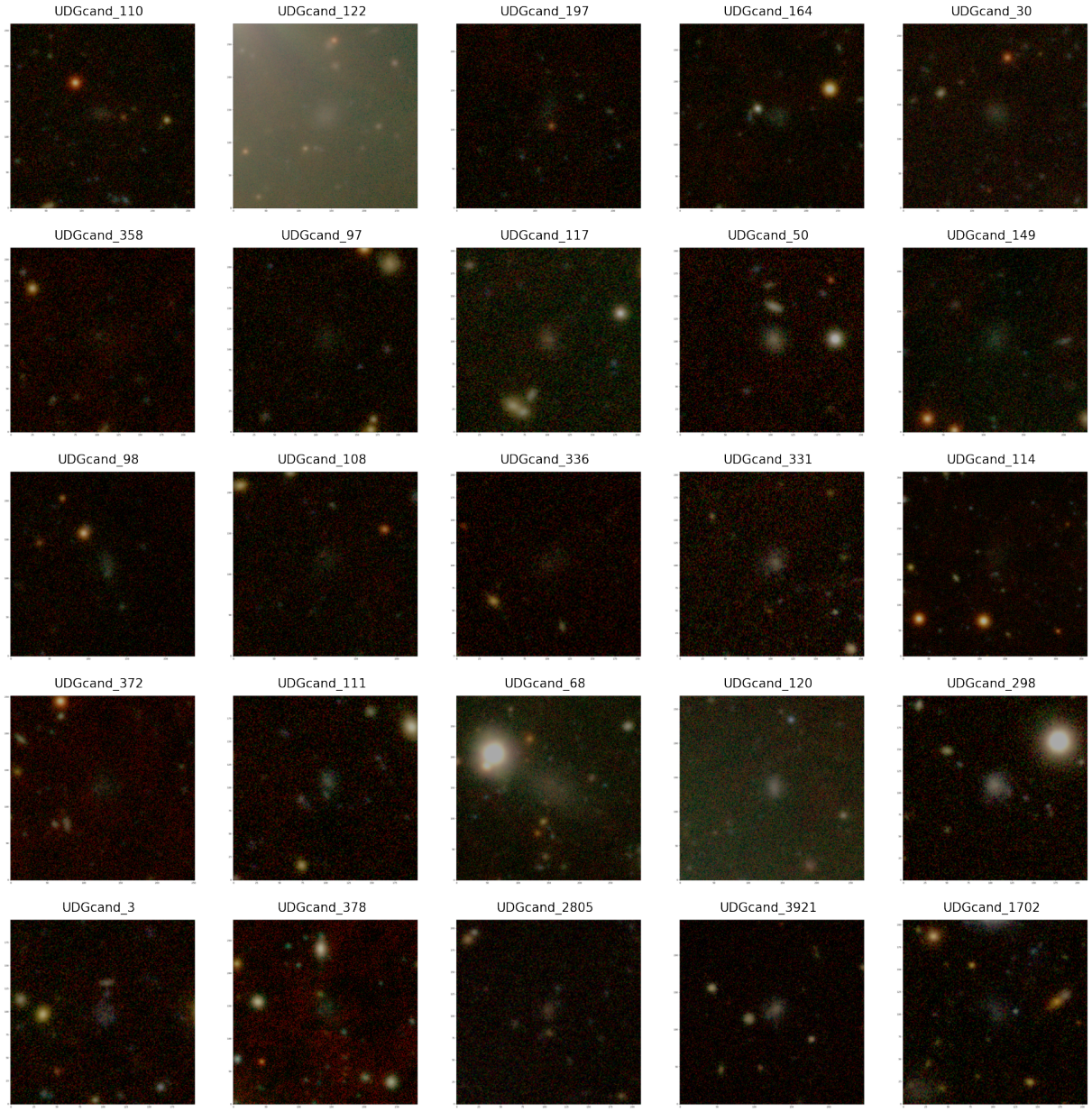


Figure 36: (1/2) Thumbnail images of objects in the LGT catalogue.

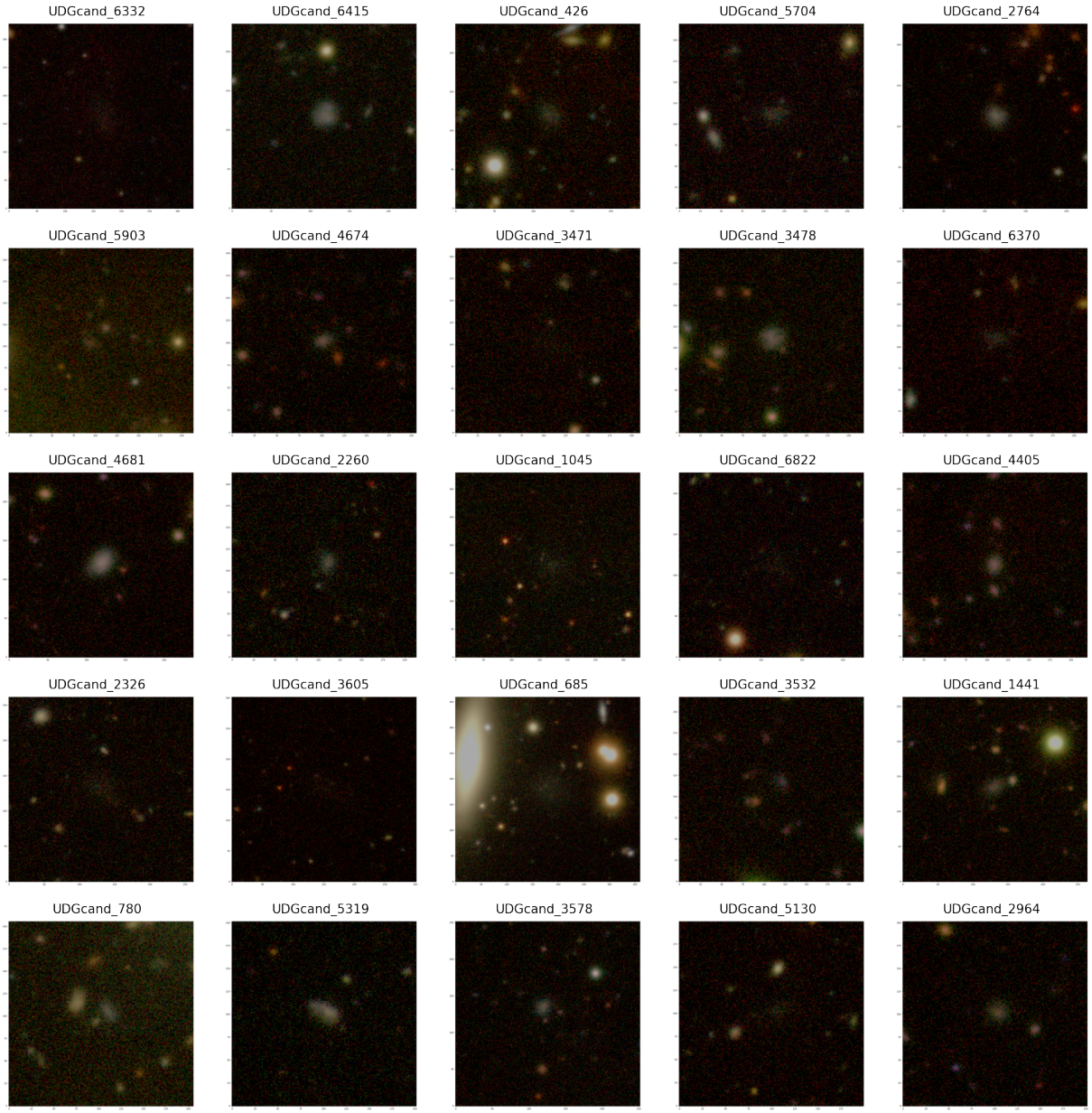


Figure 37: (2/2) Thumbnail images of objects in the LGT catalogue.

6.4 Code

A complete overview of the (Python) code written during the writing of this thesis is available on GitHub[4], however, this section contains a part of the code written for the creation and parsing of the dataframes themselves. Code that serves only to create the plots used in the figures throughout this thesis are left on GitHub, as they can be recreated easily once one obtains the parsed dataframes. The same goes for various types of queries that serve only to filter the dataframes. The code on GitHub is mostly in the form of Jupyter Notebooks, with most Notebooks containing enough information to be standalone to a certain extent.

6.4.1 *Dataframe creation*

```
1  import os
2  from .sf import (
3    getFilename,
4    extract_task_value,
5    parseTime,
6    get_power_users,
7    percentageVotesForAnswer,
8    extractTaskValue,
9    get_task_0_value_counts)
10 from .helpers import json_parser
11 from datetime import date
12 import pandas as pd
13 import numpy as np
14 import json
15
16 def make_df_classify(workflow, task_indices=[0,1]): # [0,1] are the indices from the classify
       workflow
17   """
18     Create a dataframe where each contains a single classification, from a Zooniverse .csv file.
19     @param {str} workflow: one of 'classify', 'onthego' and 'hardcore'
20     @param {List[Int]} task_indices: list of task indices present in the given workflow
21   """
22   converters = { column_name: json_parser for column_name in ['annotations', 'subject_data', '
       metadata'] }
23
24   cwd = os.path.dirname(os.path.abspath(__file__))
25   csv_filenames = {
26     'classify': 'classify-classifications',
27     'hardcore': 'classify-hardcore-edition-classifications',
28     'onthego': 'classify-on-the-go-classifications'
29   }
30   pathstring = '../SpaceFluff/zooniverse_exports/{}.csv'.format(csv_filenames[workflow])
31   loc = os.path.join(cwd, pathstring)
32   df = pd.read_csv(loc, delimiter=",", converters=converters)
33
34   df.insert(0, 'Filename', df['subject_data'].apply(getFilename))
35
36   tasks = ['T{}'.format(i) for i in task_indices]
37   for task in tasks:
38     df[task] = df['annotations'].apply(lambda x: extractTaskValue(x, task))
39
40   df = df[~df['T0'].isnull()] # if user didn't answer T0, the classification is void and can be
       removed safely
41
42   # filter out classifications from beta
43   df['created_at'] = parseTime(df['created_at'])
44   end_of_beta = pd.Timestamp(date(2020, 10, 20), tz='utc')
45   df = df[df['created_at'] > end_of_beta]
46
47   try:
48     df['isRetired'] = df['metadata'].apply(lambda x: x.get('subject_selection_state', {}).get('
       retired'))
49     df['alreadySeen'] = df['metadata'].apply(lambda x: x.get('subject_selection_state', {}).get(
       'already_seen'))
50
51     # filter alreadySeen or retired rows, and drop obsolete columns from the dataframe
       altogether
52     df = df.query('(isRetired == False) & (alreadySeen == False)')
53     df = df.drop(['isRetired', 'alreadySeen', 'gold_standard'], axis=1)
54   except:
55     pass
56
57   return df
58
59 def make_df_vote_threshold(df, vote_count_threshold):
60   users_and_votes = get_power_users(df, vote_count_threshold)
61   usernames = [user['username'] for user in users_and_votes]
```

[4]https://github.com/Seerden/SpaceFluff

```
62
63      df = df[df['user_name'].isin(usernames)]
64
65      return df
66
67  def make_df_tasks_with_props(df, candidate_names, object_info, onthego=False):
68      # create a temporary dataframe containing only classifications where 'task0' == 'Galaxy'
69      df_galaxy = df[df['T0'] == 'Galaxy']
70      galaxy_names = df_galaxy['Filename']
71
72      df_task0 = make_df_task0(df, candidate_names, onthego)
73
74      if not onthego:
75          groupby_name = df_galaxy[['Filename', 'T0', 'T1']].groupby(['Filename'])
76          galaxy_task1_values = []
77          for name in set(galaxy_names):
78              group = groupby_name.get_group(name) # get all classifications of this object from df
79
80              rowObj = {
81                  "name": name
82              }
83
84              for answer in ['Fluffy', 'Bright']: # add 'fluffy' and 'bright' columns
85                  rowObj['%_{}'.format(answer)] = round(list(group['T1']).count(answer)*100/group.shape
        [0], 1)
86
87              none_count = group[group['T1'].isnull()].shape[0] # also manually add 'None' row since
        None is parsed to NaN otherwise
88              rowObj['%_None'] = round(none_count*100/group.shape[0], 1)
89
90              galaxy_task1_values.append(rowObj) # append rowObj to list
91
92          df_task1 = pd.DataFrame(galaxy_task1_values)
93          df_tasks = df_task1.merge(df_task0, on='name', how='outer')
94      else:
95          df_tasks = df_task0
96      df_tasks_with_props = df_tasks.merge(object_info, how='outer', on='name') # merge properties
          onto dataframe
97      df_tasks_with_props = df_tasks_with_props[~df_tasks_with_props['#_votes'].isnull()] # filter
          out objects without actual votes
98
99      return df_tasks_with_props
100
101 def make_df_task0(df, candidate_names, onthego):
102     # group df by filename, so that each group contains only rows belonging to that object
103     gr = df[['Filename', 'T0']].groupby('Filename')
104
105     task0Values = [] # create empty list to push results to
106     for objectName in candidate_names:
107         # loop over every group created above to accumulate 'task 0' votes ('galaxy'/'group of
          objects'/'something else')
108         try:
109             task0_values = gr.get_group(objectName)['T0']
110             counts, votes = get_task_0_value_counts(task0_values)
111
112             countObj = {
113                 "name": objectName,
114                 "counts": counts,
115                 "#_votes": votes
116             }
117
118             task0Values.append(countObj)
119         except:
120             continue
121
122     df_task0 = pd.DataFrame(task0Values)
123
124     clusterstring = 'Group_of_objects_(Cluster)' if not onthego else 'Group_of_objects_(cluster)'
125     answer_types = ['Galaxy', clusterstring, 'Something_else/empty_center']
126
127     for ans_type in answer_types:
128         vote_percentage_column = df_task0['counts'].apply(
129             lambda x: percentageVotesForAnswer(x, ans_type))
130         df_task0['%_votes_{}'.format(ans_type)] = vote_percentage_column
131
132     # filter dataframe and only leave objects with more than 5 votes
133     # df_task0 = df_task0[df_task0['# votes'] > 5]
134
135     return df_task0
136
137 def make_df_vote_threshold(df, vote_count_threshold):
138     users_and_votes = get_power_users(df, vote_count_threshold)
139     usernames = [user['username'] for user in users_and_votes]
140
141     df = df[df['user_name'].isin(usernames)]
142
143     return df
```

Dataframe creation helpers

```
import pandas as pd
import numpy as np
import os
import json
from datetime import date

def getFilename(subject_data):
  """
  Given the subject_data field from a row of one of our SpaceFluff dataframes, extract the name of the object being classified
  by extracting the 'Filename'|'image'|'IMAGE' field".

  To be used with df[column].apply()

  @returns {string} filename of the object being classified, including the extension '_insp.png'
  """

  keys = list(subject_data.values())[0].keys()
  accessKey = (
    "Filename" if "Filename" in keys else "image" if "image" in keys else "IMAGE" if "IMAGE" in keys else None)

  if accessKey:
    return list(subject_data.values())[0][accessKey][:-9]
  else:
    print("No filename found!")

def getMetadataValue(metadata, field):
  '''
    @param metadata metadata column from a row in a SpaceFluff dataframe
    @param {string} field: 'retired' | 'already_seen'
    @returns {boolean} value of `field` within the row's metadata column
  '''
  return metadata['subject_selection_state'][field]

def parseTime(created_at):
  '''
  @param {df column} created_at: df['created_at'] column
  '''
  return pd.to_datetime(created_at, format="%Y-%m-%d %H:%M:%S %Z")

def getGroupSize(group):
  '''
    @param {pd.core.frame.DataFrame} pandas dataframe group
    @returns number of rows in group (corresponds to number of columns in case of parsed SpaceFluff dataframe)
  '''
  return group.shape[0]

def extract_task_value(task_index, row):
  try:
    return row[task_index]['value']
  except:
    return

def percentageVotesForAnswer(counts, answer):
  '''
  @param counts: a df column like {galaxy: 15, group of objects (cluster): 10, something else/empty center: 2}
  @paramanswer: one of the keys of `counts`
  '''

  totalVotes = sum(counts.values())

  if not answer in counts.keys():
    return 0

  votesForAnswer = counts[answer]

  return round(100*votesForAnswer/totalVotes, 1)

def extractTaskValue(annotations, task):
  '''
  @param {list} annotations: annotations column for a row in a SpaceFluff dataframe
  @param {string} task: one of 'Ti', where i \in 0,2,1,3,4,5,9
  @returns {string | None} value the user provided for the given task, or None
  '''

  filtered = list(filter(lambda x: x['task'] == task, annotations))
  if len(filtered) > 0:
    return filtered[0]['value']

def extract_retired_info(subject_data):
  '''
    @param subject_data: (dataframe 'subject_data' column)
```

```
    '''
    return list(subject_data.values())[0]["retired"]

def get_power_users(df, vote_count_threshold):
    """
    @param df: parsed dataframe where each row is a single classification
    @param {int} vote_count_threshold: return only users that made at least this many valid
        classifications
    """

    groupby_username = df[['user_name']].groupby(['user_name'])
    groupby_username_filtered = groupby_username.filter(lambda x: x.shape[0] >= vote_count_threshold)

    grouped = groupby_username_filtered.groupby(['user_name'])

    filtered_usernames_and_votes = []
    for username, vote_count in grouped:
      filtered_usernames_and_votes.append({
        "username": username,
        "votes": len(vote_count)
      })

    return filtered_usernames_and_votes

def get_task_0_value_counts(row):
    "Get␣task␣0␣value␣counts␣for␣one␣row␣of␣a␣group␣of␣classifications"
    row = list(row)

    # value_counts = {answer: 0 for answer in answer_types}
    value_counts = {}
    for vote in row:
      if value_counts.get(vote):
        value_counts[vote] += 1
      else:
        value_counts[vote] = 1

    return value_counts, len(row)

def as_array(lst):
    'Turn␣a␣Python␣list␣into␣a␣NumPy␣array'
    if type(lst) == np.ndarray:
      return lst
    return np.array(lst)

def get_running_vote_fraction(df):
    """
    Returns a list of
      (% votes by users that case <= n votes)/total votes
    as a function of n
    @param df: `df`-like dataframe, where each row corresponds to a single classification made by a
        single user
    """
    users_and_classification_counts = []

    for k, v in df.groupby('user_name').groups.items():
      users_and_classification_counts.append({
        'username': k,
        'classifications': len(v)
      })

    cls_per_user = [entry['classifications'] for entry in users_and_classification_counts]
    total_votes = sum(cls_per_user) # total number of votes made
    sorted_vote_counts = sorted(cls_per_user) # sorted list of number of classifications per user

    # create dictionary with keys: # votes per user, values: # users that cast that amount of votes
    countDict = {}
    for entry in sorted_vote_counts:
      countDict[entry] = countDict.get(entry, 0) + 1

    fractions = []
    for vote_count, occurrence_rate in countDict.items():
      fractions.append([vote_count, vote_count*occurrence_rate/total_votes, occurrence_rate])
    counts, fractions, users_included = as_array(fractions).T

    # create a running fraction of total votes cast in a single loop
    running_fraction = []
    for i, fr in enumerate(fractions):
      if i == 0:
        val = fr
      else:
        val = fr+running_fraction[i-1]
      running_fraction.append(val)

    return [
      users_and_classification_counts,
      cls_per_user,
      counts,
      running_fraction
```

```
]
```

Other helper functions

```
import json
import sys
import numpy as np
import pandas as pd

def json_parser(data):
  return json.loads(data)

def df_to_json(df, path):
  "Save a dataframe ‘df‘ as .json file in the specified (relative) location ‘path‘ as ’./path.json’"
  df.to_json('{}/{}.json'.format(sys.path[0], path))

def get_cols(df, cols):
  '''
    Extract values of the specified columns ‘cols‘ from a dataframe ‘df‘
    @param df: input dataframe
    @param cols: list of columns, e.g. ['Name', 'Date', 'Votes']
    @returns list of lists, where each list contains all values for that column,
    @example:
      name, date, votes = get_cols(df, ['Name', 'Date', 'Votes'])
  '''
  return df[cols].T.values

def get_column_names(s, df):
  'Retrieve from a dataframe ‘df‘ the list of column names that start with (sub)string ‘s‘'
  cols = df.columns.tolist()
  return list(filter(lambda x: x.startswith(s), cols))

# Fleiss' kappa computation
def get_P_i(row, answers):
  '''
    Compute the proportions for each answer for a single 'subject' (Space Fluff object)
    @param row: a row in ‘df_votes‘-like dataframe
    @param answers: list of unique answers for task for which we're computing Fleiss' kappa

  '''
  n = sum(row.values()) # number of votes for this category for this object
  if n < 2:
    return 1
  else:
    val_sum = 0
    for answer in answers:
      val = row.get(answer, 0)
      val_sum += val*val
    return val_sum/(n*(n-1))

def fleiss_kappa(df, df_votes, task):
  '''
    Compute Fleiss' kappa for a single task, for a df_votes-like dataframe.
    @param df: dataframe where each row corresponds to a single classification
    @param df_votes: dataframe derived from ‘df‘, where each row corresponds to a single object,
      its parameters, and the number of votes it got for each answer in each task
    @param task: one of 'T0', 'T1', etc.
  '''

  N = df.shape[0] # total number of votes

  answers = df[task].unique()

  p_js = np.array([df_votes['T0'].apply(lambda x: x.get(ans, 0)).values.tolist() for ans in answers])/N

  P_is = df_votes[task].apply(lambda x: get_P_i(x, answers))
  P_bar = sum(P_is)/df_votes.shape[0]
  P_bar_e = np.sum(p_js**2)

  kappa = (P_bar - P_bar_e)/(1 - P_bar_e)

  return kappa
```