Crowdsourcing quality control for Dark Energy Survey images

P. Melchior[a,b], E. Sheldon[c], A. Drlica-Wagner[d], E. S. Rykoff[e,f], T. M. C. Abbott[g], F. B. Abdalla[h,i], S. Allam[d], A. Benoit-Lévy[j,h,k], D. Brooks[h], E. Buckley-Geer[d], A. Carnero Rosell[l,m], M. Carrasco Kind[n,o], J. Carretero[p,q], M. Crocce[p], C. B. D'Andrea[r,s], L. N. da Costa[l,m], S. Desai[t,u], P. Doel[h], A. E. Evrard[v,w], D. A. Finley[d], B. Flaugher[d], J. Frieman[d,x], E. Gaztanaga[p], D. W. Gerdes[w], D. Gruen[y,z], R. A. Gruendl[n,o], K. Honscheid[a,b], D. J. James[g], M. Jarvis[aa], K. Kuehn[ab], T. S. Li[ac], M. A. G. Maia[l,m], M. March[aa], J. L. Marshall[ac], B. Nord[d], R. Ogando[l,m], A. A. Plazas[ad], A. K. Romer[ae], E. Sanchez[af], V. Scarpine[d], I. Sevilla-Noarbe[af,n], R. C. Smith[g], M. Soares-Santos[d], E. Suchyta[aa], M. E. C. Swanson[o], G. Tarle[w], V. Vikram[ag], A. R. Walker[g], W. Wester[d], Y. Zhang[w]

[a]Center for Cosmology and Astro-Particle Physics, The Ohio State University, Columbus, OH 43210, USA
[b]Department of Physics, The Ohio State University, Columbus, OH 43210, USA
[c]Brookhaven National Laboratory, Bldg 510, Upton, NY 11973, USA
[d]Fermi National Accelerator Laboratory, P. O. Box 500, Batavia, IL 60510, USA
[e]Kavli Institute for Particle Astrophysics & Cosmology, P. O. Box 2450, Stanford University, Stanford, CA 94305, USA
[f]SLAC National Accelerator Laboratory, Menlo Park, CA 94025, USA
[g]Cerro Tololo Inter-American Observatory, National Optical Astronomy Observatory, Casilla 603, La Serena, Chile
[h]Department of Physics & Astronomy, University College London, Gower Street, London, WC1E 6BT, UK
[i]Department of Physics and Electronics, Rhodes University, PO Box 94, Grahamstown, 6140, South Africa
[j]CNRS, UMR 7095, Institut d'Astrophysique de Paris, F-75014, Paris, France
[k]Sorbonne Universités, UPMC Univ Paris 06, UMR 7095, Institut d'Astrophysique de Paris, F-75014, Paris, France
[l]Laboratório Interinstitucional de e-Astronomia - LIneA, Rua Gal. José Cristino 77, Rio de Janeiro, RJ - 20921-400, Brazil
[m]Observatório Nacional, Rua Gal. José Cristino 77, Rio de Janeiro, RJ - 20921-400, Brazil
[n]Department of Astronomy, University of Illinois, 1002 W. Green Street, Urbana, IL 61801, USA
[o]National Center for Supercomputing Applications, 1205 West Clark St., Urbana, IL 61801, USA
[p]Institut de Ciències de l'Espai, IEEC-CSIC, Campus UAB, Carrer de Can Magrans, s/n, 08193 Bellaterra, Barcelona, Spain
[q]Institut de Física d'Altes Energies (IFAE), The Barcelona Institute of Science and Technology, Campus UAB, 08193 Bellaterra (Barcelona) Spain
[r]Institute of Cosmology & Gravitation, University of Portsmouth, Portsmouth, PO1 3FX, UK
[s]School of Physics and Astronomy, University of Southampton, Southampton, SO17 1BJ, UK
[t]Excellence Cluster Universe, Boltzmannstr. 2, 85748 Garching, Germany
[u]Faculty of Physics, Ludwig-Maximilians University, Scheinerstr. 1, 81679 Munich, Germany
[v]Department of Astronomy, University of Michigan, Ann Arbor, MI 48109, USA
[w]Department of Physics, University of Michigan, Ann Arbor, MI 48109, USA
[x]Kavli Institute for Cosmological Physics, University of Chicago, Chicago, IL 60637, USA
[y]Max Planck Institute for Extraterrestrial Physics, Giessenbachstrasse, 85748 Garching, Germany
[z]Universitäts-Sternwarte, Fakultät für Physik, Ludwig-Maximilians Universität München, Scheinerstr. 1, 81679 München, Germany
[aa]Department of Physics and Astronomy, University of Pennsylvania, Philadelphia, PA 19104, USA
[ab]Australian Astronomical Observatory, North Ryde, NSW 2113, Australia
[ac]George P. and Cynthia Woods Mitchell Institute for Fundamental Physics and Astronomy, and Department of Physics and Astronomy, Texas A&M University, College Station, TX 77843, USA
[ad]Jet Propulsion Laboratory, California Institute of Technology, 4800 Oak Grove Dr., Pasadena, CA 91109, USA
[ae]Department of Physics and Astronomy, Pevensey Building, University of Sussex, Brighton, BN1 9QH, UK
[af]Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas (CIEMAT), Madrid, Spain
[ag]Argonne National Laboratory, 9700 South Cass Avenue, Lemont, IL 60439, USA

Abstract

We have developed a crowdsourcing web application for image quality control employed by the Dark Energy Survey. Dubbed the "DES exposure checker", it renders science-grade images directly to a web browser and allows users to mark problematic features from a set of predefined classes. Users can also generate custom labels and thus help identify previously unknown problem classes. User reports are fed back to hardware and software experts to help mitigate and eliminate recognized issues. We report on the implementation of the application and our experience with its over 100 users, the majority of which are professional or prospective astronomers but not data management experts. We discuss aspects of user training and engagement, and demonstrate how problem reports have been pivotal to rapidly correct artifacts which would likely have been too subtle or infrequent to be recognized otherwise. We conclude with a number of important lessons learned, suggest possible improvements, and recommend this collective exploratory approach for future astronomical surveys or other extensive data sets with a sufficiently large user base. We also release open-source code of the web application and host an online demo version at http://des-exp-checker.pmelchior.net.

Keywords: surveys, Information systems: Crowdsourcing, Human-centered computing: Collaborative filtering

arXiv:1511.03391v2 [astro-ph.IM] 2 May 2016

1. Introduction

Large astronomical surveys produce vast amounts of data for increasingly demanding science applications. At the same time, the complexity of the instruments, operations, and the subsequent data analyses renders glitches and flaws inevitable, particularly during the early phases of experiments. Thus, mechanisms that facilitate the discovery and reporting of problems in the data, whether they originate from unexpected instrumental behavior or insufficient treatment in software, are important components of a data quality program. Due to the often unexpected nature of these features, algorithmic approaches to identify artifacts are generally infeasible and human inspection remains necessary. Human-generated reports can then be fed back to algorithm developers and hardware experts to mitigate and eliminate problems whenever possible. For current and upcoming surveys, the demands of carefully inspecting sizable portions of the data volume exceed the capabilities of individual, or a small team of, data management experts.

Crowdsourcing has seen tremendous success in the past few years in many applications where a critical task cannot be performed by computers but where the amount of data to be gathered, processed, or analyzed exceeds the capabilities of even the most dedicated human. Examples can be found in non-profit, academic, commercial, or activist settings. In astronomy, one of the first implementations of crowdsourcing was to gather information about the 1833 Leonid meteor storm (Olmsted, 1834a,b; Littmann & Suomela, 2014). In recent years, widespread access to the internet has made such efforts easier to realize, allowing for larger crowds and quicker turnaround of results. The preeminent early adopter of this web-based mode of operations is the Galaxy Zoo project (Lintott et al., 2008), designed to visually classify the morphology of galaxies from the Sloan Digital Sky Survey (York et al., 2000). Galaxy Zoo has led to Zooniverse,[1] currently the largest online portal for citizen science projects. At the time of writing, Zooniverse findings have been published in over 80 articles across several disciplines of science.

We have built a web-based crowdsourcing application for image quality control for the Dark Energy Survey (DES; Dark Energy Survey Collaboration, 2005).[2] DES is a 5,000 deg^2 survey in five photometric bands (*grizY*) operating from the Blanco 4m telescope at the Cerro Tololo Inter-American Observatory (CTIO). Its 570 megapixel imager DECam (Flaugher et al., 2015) comprises 62 science CCDs (2048×4096 pixels) and 12 focus and guiding CCDs (2048×2048 pixels), covering a roughly hexagonal footprint of 3 deg^2. Each region of the survey footprint will be observed 10 times in each band over the course of five years. The total number of exposures is thus expected to be approximately 10^5, with a data volume in science images of order 100 TB. An overview of the data processing and management pipeline (DESDM) is given by Mohr et al. (2012) and Desai et al. (2012).

The application, dubbed the "DES exposure checker", is geared for a professional user base of several hundred scientists and seeks to identify flaws in the DES images that will be used for most science analyses. Problems discovered with this application can then be fixed in hardware or in subsequent data processing runs. Our approach ties in with other quality control efforts, which automatically analyze the latest exposures and flag cases of e.g. bad observing conditions (Honscheid et al., 2012; Diehl et al., 2014), and allow the inspection of final coadded data products, both images and object catalogs (Balbinot et al., 2012). Our concept profits from the experience of a prior ad hoc crowdsourcing effort in DES. During the so-called Science Verification phase in 2012, flaws in DES imaging have been identified by a small "eyeball squad", whose reports were relayed to DES operations and data management experts on a daily basis. While essential to improving the performance of the instrument during this early phase, the effort did not scale well with the increasing number of incoming images. In the remainder of this paper, we will show how to create a scalable solution to data quality control by providing an engaging user experience, while simultaneously maximizing the utility of the report collection. The reports resulting from the DES exposure checker have already been used to inform the modification of existing algorithms, the development of new algorithms, and a general improvement the quality of the DES data.

Concept

For many crowdsourcing applications, classification is the critical task performed by humans. Specifically, which class, of a predefined set, does a given test object belong to? To render this question accessible to as wide an audience as possible, the task is ideally broken up into yes-or-no questions. The approach we present here differs insofar as we ask our users not only to classify image flaws according to a given classification scheme, but also to extend the scheme if new classes of flaws arise. We can add this complexity because our audience consists of mostly experienced or prospective astronomers, who can contribute their prior knowledge to improve the classification scheme. This feedback system, catered to a professional—but not necessarily expert—user base, is the main novelty of our approach. Specifically, we want to bring together three different objectives in one application:

(1) Show DES participants how science-grade images really look.

(2) Enable users to discover and classify image flaws of known or unknown type.

(3) Aggregate problem reports to improve data and processing quality.

Objective (1) has an educational direction. To optimally exploit the data, participants should understand how the survey performs: what its capabilities and limitations really are. In addition, new participants, such as beginning students, should get a visual introduction to the current level of image quality. Objective (1) also provides a substantial source of motivation. During the early phases of an experiment, surprising features may be

*Corresponding author
Email address:* peter@pmelchior.net (P. Melchior)

present in the data. We therefore anticipate that participants are eager to satisfy their curiosity, and we seek to streamline the application such that they can efficiently inspect the images.

Practical limitations are at the heart of Objective (2). The flaws we seek to identify stem from instrumental malfunctions, sub-par observational conditions, and, foremost, an insufficient treatment of various artifacts during the stages of data processing. Because that constitutes an extensive and poorly defined list of potential problems, human inspection remains irreplaceable. The key to Objective (2) is to leverage the user's curiosity to perform this seemingly mundane task. Beyond curiosity, we expect—and our experience confirms—that users are genuinely motivated by a concern and a sense of responsibility for their respective scientific application. Consequently, different users care about different problems, and a diverse user base (according to scientific background, level of experience, etc.) is thus beneficial to identify different kinds of problems. The challenge of diversity lies in the difficulty in obtaining unequivocal classifications of known and especially previously unknown flaws.

Objective (3) enhances the overall utility of the application. With the collection of problem reports, we seek to provide feedback to experts in charge of data management and developers of data reduction and analysis pipelines. We are able to provide a list of training cases to assess the performance of new or improved processing algorithms. In particular, with a sufficiently large user base we can generate rapid feedback if unexpected and even rare problems exists in any new data release.

The remainder of this article describes how we seek to combine these objectives. Although interconnected, we separate the topics broadly according to the three aspects listed above. In Section 2 we describe the presentation of the image data and the main elements of the user interface; Section 3 deals with user interaction and engagement; and Section 4 summarizes our findings on how the application has been used and what improvements have been made to data processing in DES based on the problems our users identified. We conclude in Section 5.

2. Presentation

Like many citizen science projects, we adopted a web-based approach to reach as many users (affiliated with the survey) as possible and to allow them to inspect images wherever and whenever they want. We also sought to eliminate installation hassles, which means that all features needed to be readily available without the need for extra software.

The web application is built with a SQLite database[3] and PHP scripting[4] on the server side, with most of the functionality happening on the client side. The HTML page design derives from the Bootstrap framework[5], and essentially all interactive elements are written in JavaScript, utilizing the jQuery library.[6]

To allow maximum flexibility in handling the images, we decided not to convert them into static images (e.g., in PNG format), but to load FITS files directly into the browser. For displaying the images in our viewer, we employ the pure JavaScript implementation WebFITS[7] (Kapadia & Smith, 2013). This approach allows the users to alter properties of the visualization, such as image stretch, zoom, or the inclusion of overlay information, on the client side, without the need to reload additional data. Even when we deal with FITS files of only one of the 62 DECam CCDs, the full image comprises 4096×2048 pixels, which would exceed the size of virtually all commercial computer monitors and visually overwhelm the user. We therefore decided to compromise between providing the most authentic data unit, namely the entire unaltered image, and the most enjoyable user experience, which generally means reducing file sizes and load times. We decided to downsample the FITS image by a factor of 4 on the server side, reducing the image to 1024×512 pixels which, with compression, is about 450 kB in size. This choice yields images that conveniently fit into most browser windows and still provide enough information to spot even small flaws in the images. An example screenshot of the viewer page is shown in Figure 1.

The majority of the page is occupied by the image itself, with the elements of user interaction being confined to a single row above the image. At the most elementary level, only two buttons are needed. If users consider the image to be flawless, they click "Next" and are presented the next one. If they recognize flaws, they can select it from the dropdown menu labeled "Problem" (see left panel of Figure 2) and then click on the image to mark the location of the problem.

Additional information is provided on the right side, where we show the data release ("SVA1"), the unique exposure, CCD number, and the filter. Users can interact with the image by toggling on or off masked regions (shown as a blue overlay on the image). The masks, which are generated by DESDM to reject areas with e.g. bad columns or saturated stars, are stored in a second image extension in the FITS file and flattened to binary data format. In this process, any information as to why a pixel has been masked is lost, but the alternative, a multi-level mask selection or multi-color display of different masks, causes overly complicated visualizations while providing little actionable information.

Finally, users can toggle the image stretch between settings that emphasize bright or faint features. Both use the same arcsinh scaling (e.g. Lupton et al., 1999) but differ in the minimum intensity considered. The gray-scale value v of a pixel with intensity I is given by

$$v = 255 \, \frac{\mathrm{arcsinh}(I) - m}{M - m},\tag{1}$$

where $M = \mathrm{arcsinh}(\max(I))$ and $m = \mathrm{arcsinh}(\min(I))$ for the faint or $m = \mathrm{arcsinh}(1)$ for the bright setting. Since the images are background-subtracted, i.e. the mean intensity of the sky is set to zero, the latter choice sets the majority of non-object pixels to black, while the former places visual emphasis on low-level noise.
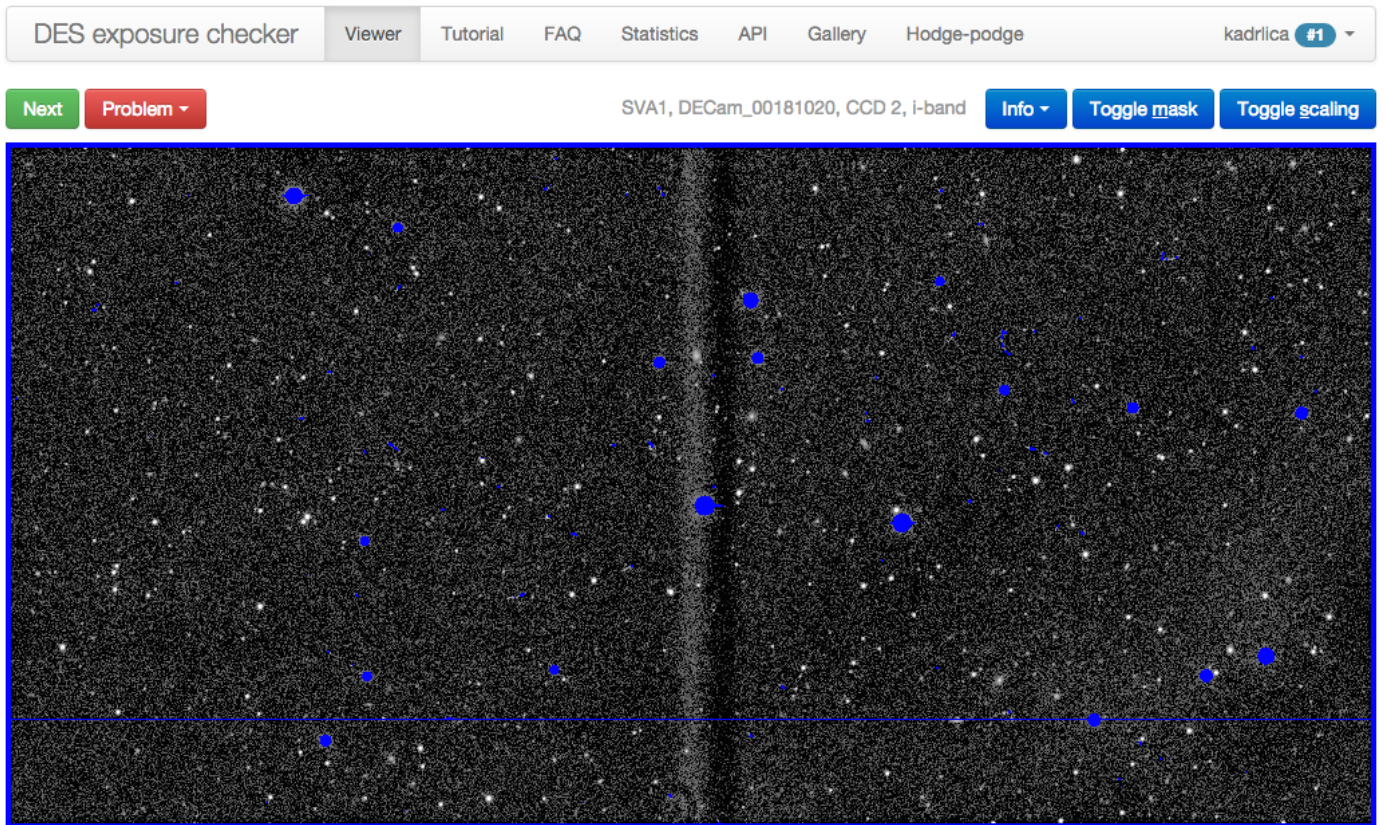
---

[3] http://sqlite.org
[4] http://php.net
[5] http://getbootstrap.com
[6] http://jquery.com

[7] https://github.com/kapadia/WebFITS

Figure 1: Screenshot of the DES exposure checker viewer page. It shows the fully reduced (i.e. science grade) FITS image of a single DECam CCD as delivered by an early version of the data processing pipeline. This image has been sky background subtracted and downsampled by a factor of four. Pixels colored in blue have been masked by the DESDM image processing pipeline. All elements of user interaction reside in the row above the image. The linear feature extending vertically accross the center of the image is the result of a shutter failure in the flat field for this exposure (Section 4.2.1).
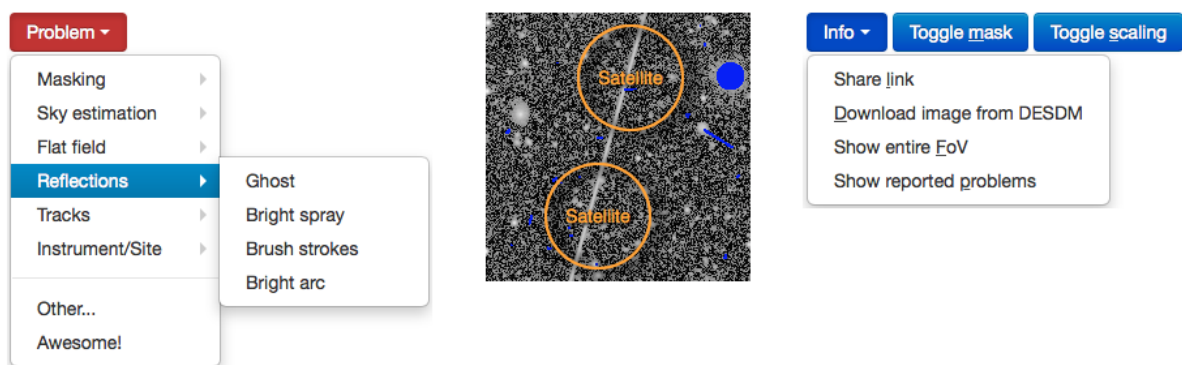
Figure 2: User interaction with the viewer page. *Left:* Problem dropdown selector with open sub-menu. *Center:* Cutout of a FITS image shown in the viewer with two problem markers (orange circles) indicating an earth-orbiting satellite trail (cf. Section 4.2.2). *Right:* Info dropdown selector.

With these two rather simple operations, users can quickly investigate different features of the image, e.g. whether an area is properly masked or structure is present in the noise. These interactions are rendered swiftly in the browser because the data are already fully loaded from the FITS file. We intentionally restricted user interactions with the image to maintain focus on the task at hand: namely to identify clearly visible problems. We therefore do not allow panning or zooming of the image,

and only provide two image stretches and no color option.[8]

Once a problem has been selected from the dropdown menu, users can click on the image as often as they deem necessary to identify problematic features. If multiple problems are identified, they may be of the same or of different kind. We added two extensions to this scheme. To identify false positive problems, e.g. the application of a mask on an area that should not have been masked, we provide a "False" button to be pressed after the problem is selected. For problems of unknown type, the problem selector "Other…" brings up a text box where users are asked to concisely describe what they see. The same workflow applies to the "Awesome!" selector for the case when users find a celestial body that is remarkable in its own right.

To support tablet computers, the act of marking a flaw is achieved with a single click or tap on the image. We decided to use circle markers so that we can place a problem label in their interior.[9] The marker and label color is yellow, which is both visually striking and clearly distinguishable from both bright and faint areas in the gray-scale image. A closeup of a marked area is shown in the center panel of Figure 2.

Finally, users can request additional information about the image from the "Info" dropdown menu (see right panel of Figure 2). We provide an unique URL for each image, so that users can share interesting cases, as well as the URL where the original FITS file (prior to our compression and background subtraction steps) can be downloaded for detailed inspection outside of the application. Because some problematic features span several CCDs and are easier to recognize when seen on a larger scale, we provide a static PNG image of the entire field of view (see Figure 3) with the option of switching to any other CCD of the same exposure. By selecting "Show reported problems", the viewer will render all problem markers it has received for the loaded image, thus allowing users to inform themselves what others users have found during their inspection. We exploit this dynamic problem lookup for user training in the Section 3.2.

3. User interaction and engagement

The central design goal of the viewer page is to render the work convenient and to not frustrate the user:

- We sought to minimize the distance on screen and number of clicks to perform the most common tasks.
- For changes in the visualization (such as image stretch, mask or problem overlay etc), we implemented keyboard shortcuts.
- Whenever the application loads information from the server, we show a spinner image to indicate a short wait time.[10]

---

[8]In a prototype, we explored a full-featured adjustment of the image stretch, which may improve the recognition of subtle problems at the expense of inspection time. A colored scheme for the image pixels instead of gray-scale would have increased the dynamic range but interferes with the colored problem markers, the visibility of which we considered critical.

[9]We adhere to Jones (1989) and use the × symbol for false-positive reports, since × rarely, if ever, marks the right spot.

[10]While most queries to the server only require a few milli-seconds, retrieving a new FITS image can take up to one second. Without the spinner, the viewer would feel unresponsive during the load time, and without visual feedback users were found to click "Next" prematurely.
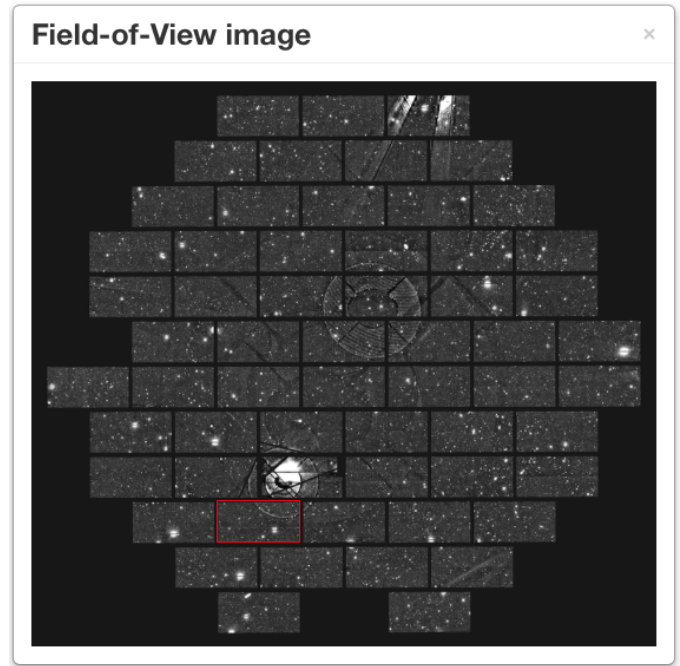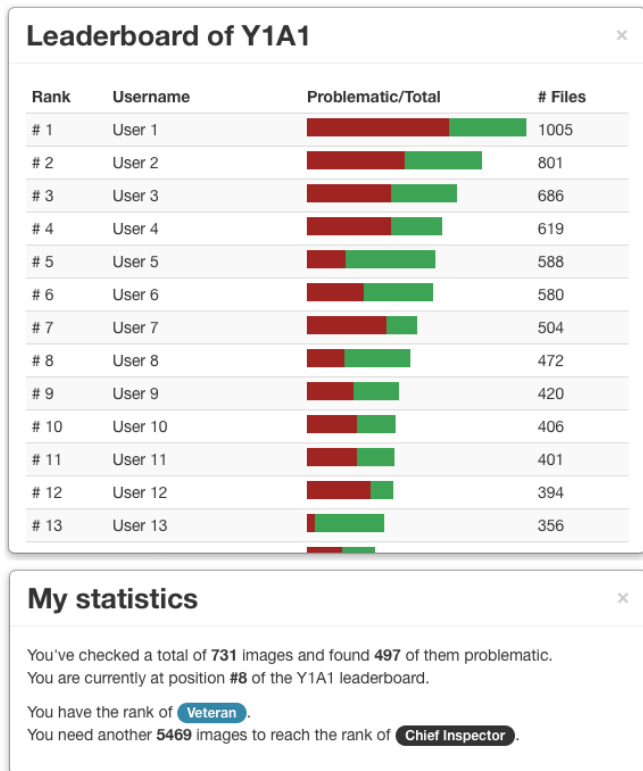

Figure 3: Field-of-View visualization. Because many image artifacts (like the so-called "ghosts" caused by reflected light from bright stars, see Section 4.2.5 for details) extend beyond a single CCD, it can be helpful to inspect the entire DECam array of a given exposure. The CCD image currently loaded in the viewer is highlighted by the red frame. Users can navigate to another CCD by clicking on its position in this image. Two CCDs are not shown; one failed early in the survey (Diehl et al., 2014), and the other experiences a time-variable gain.

Streamlining the viewer makes it less likely that users turn away from the application, but does not provide any motivation to use—or even continue to use—it in the first place.

3.1. Motivation

We pointed out in Section 1 that users involved in the survey have a genuine desire to understand the problems that may affect their science analyses and actively want to help improve the data quality. However, we sought to incorporate principles of gamification to provide additional motivation for users to keep using the application. Gamification results in a "blurring of the relationship between work and play" (e.g. Greenhill et al., 2014), and renders the task at hand into a competitive game that follows its own set of rules to define a winner. We employed a very simple metric to define success in our "game", namely the number of images users have inspected, irrespective of whether they found a problem or how many different problems they marked. The ranking of all users is shown as a leaderboard in the viewer (see top panel of Figure 4), and the top 15 users are shown on the start page of the website as an incentive to rank above the cut. To prevent users from tricking the system by simply submitting empty problem reports—hitting "Next" repeatedly without inspecting the images—we resort to

Leaderboard of Y1A1

Rank	Username	Problematic/Total	# Files
# 1	User 1		1005
# 2	User 2		801
# 3	User 3		686
# 4	User 4		619
# 5	User 5		588
# 6	User 6		580
# 7	User 7		504
# 8	User 8		472
# 9	User 9		420
# 10	User 10		406
# 11	User 11		401
# 12	User 12		394
# 13	User 13		356

My statistics

You've checked a total of **731** images and found **497** of them problematic.
You are currently at position **#8** of the Y1A1 leaderboard.

You have the rank of Veteran.
You need another **5469** images to reach the rank of Chief Inspector.

Figure 4: Elements of gamification. *Top:* Leaderboard with rank, username (anonymized for the purpose of this publication), a stacked bar chart indicating the total number and the number of flawless images (green), and the image count for the given data release ("Y1A1"). *Bottom:* User statistics listing the total number of images inspected (combining multiple data releases), the position on the release leaderboard, the current rank and, the number of images needed to attain the next-higher rank.

peer pressure. The leaderboard shows the fraction of images each user has considered flawless in green, triggering discussions among users when some of them appear "too soft".[11]

We also implement a reward system with virtual badges. Based on the total number of images inspected, the user progresses through a rank system, with higher ranks becoming increasingly hard to attain:

- "Rookie" for 10 images
- "Frequent Checker" for 62 images (corresponding to the number of DECam CCDs)
- "Top Performer" for 5×62 images
- "Veteran" for 10×62 images
- "Chief Inspector" for 100×62 images

---

[11]We do not provide any messaging system for users, but usernames are mostly recognizable and users know each other, so that email remained the preferred means of communication.
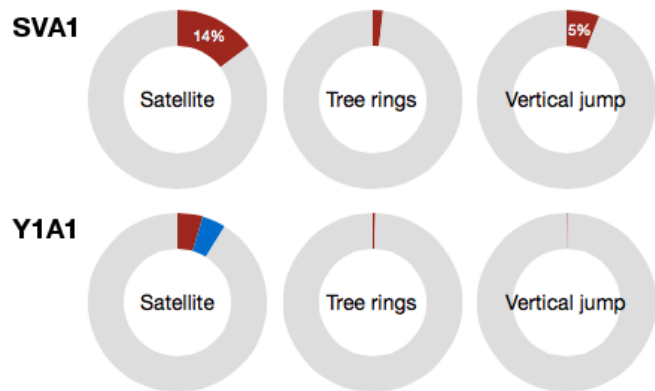
Figure 5: Aggregated statistics of problem reports, for the two first data releases of DES, SVA1 and Y1A1. Each donut chart shows the reported fraction of problematic images to exhibit this kind of flaw (red). False-positive fractions are colored in blue. See Section 3.2 for a discussion of the reliability of the abundance estimates.

Whenever a user crosses one of the thresholds, a window will appear and congratulate the user for being awarded the next-higher rank. We chose a very low threshold for the lowest award to provide early gratification for most users. Once awarded, the colored badge is shown next to the username in the top-right corner of Figure 1, highlighting the user's progress. Users also have access to their current statistics: the number of images they have inspected, their current rank, and the number of images they need to inspect for the next rank (see bottom panel of Figure 4).

We furthermore created a dynamic gallery comprised of images flagged as "Awesome" in the viewer. Listing the username and timestamp together with the user's description of the object not only yields some bragging rights for the discoverer, it also allows other users to share the enjoyment of remarkable images of astronomical objects (a fascination that influenced many users to pursue careers in physics or astronomy). While most objects in the gallery are well-known nearby spiral galaxies, in a few cases users suspected that they had discovered a so-far unknown object. When this was the case, we attempted to verify the nature of the object in question. In one instance, we confirmed that a user had spotted a comet, C/2010 R1, that had been discovered only a few years earlier.

Finally, we engaged users by sending email newsletters that announced updates, such as the availability of a new data set or improvements to the application, and also by giving out bonus points for the leaderboard position whenever a user inspected 10 images during the days from December 1 to 24, 2013. Both approaches led to at least twice as many user contributions compared to a regular day.

3.2. Consensus on problems

So far we have discussed the creation of an engaging application to interact with DES images, promoting Objective (1). To address Objective (2) and Objective (3) it is critical that

users be able to find flaws, recognize and classify them, and even identify previously unclassified flaws. The primary element of user training is an extensively vetted tutorial page, which describes common problem classes and the best current understanding of their origins.[12] Next to each description, we show a thumbnail of a characteristic example, which is linked to the viewer page, so that one can access the original FITS image and the problem markers reported by all users. Thus, the tutorial provides users with a visual reference of prototypical problem cases together with an understanding of their causes. The connection between cause and effect further satisfies the desire of survey participants "to understand the data".

In addition to the user tutorial, we implemented a statistic overview page, where problem reports were aggregated and listed for each problem class (a selection is shown in Figure 5). We show how often each problem occurs, and the fraction of false positive problem reports for each class.[13] The problem classes are linked to example images drawn at random from the images with corresponding problem reports. Beyond the prototypical cases shown in the tutorial, the random selection yields a dynamic impression of the variation between problem cases, or, more precisely, how they are perceived by users.

At this point we have to emphasize that these statistics do not necessarily provide a fair assessment of the overall data quality in DES. Many reports refer to insignificant flaws such as incompletely masked cosmic ray hits, which leave bright but only very short and thin streaks and therefore affect very small areas. Hardly any science application would be badly impacted by them, but we still count such an image as "reportedly problematic". Furthermore, the relative abundances of problem classes can become skewed towards the most obvious problems. For instance, large ghosts like those visible in Figure 3, often compel users to open the Field-of-View image and iterate through *all* affected CCDs, marking the same ghost on several images, whereas they normally only inspect one CCD image of a given exposure if there is no large problem visible. Also, with the overall data quality improving over time, users tend to shift their attention to subtler flaws, so that the aggregated statistics may not directly reflect the improvements to data quality that have actually been made.

Finally, we address the problem reports posted in the umbrella category "Other", chosen when users are uncertain about the type of problem. Similar to the gallery of "Awesome" objects, we list these cases together with the free-form description provided by the user on a page dubbed "Hodge-podge". Each problem report is linked to the corresponding FITS image in the viewer, and the page also shows how often the same description has been used on different images. The hodge-podge page is linked from the tutorial and constitutes a dynamic extension of the latter to currently unclassified problems. We ask users

Figure 6: Number of users within bins of either inspected images per user (blue) or total number of problems reported per user (green). The bins span one decade and are indicated by the vertical white lines, the horizontal position within each bin is given by its average submission number. Poissonian uncertainties are shown.

(as part of the tutorial and in email newsletters) to inspect the hodge-podge list to integrate these problems in their visual classification scheme and to adopt the existing description whenever they come across a fitting case.[14] Because the list of hodge-podge problems is ranked according to their occurrence, users are presented the most common—and hence, most important—problems first, similar to the concept of ranked comments on discussion websites such as http://stackexchange.com.

This feedback loop, with which we dynamically integrate and extend the list of problem examples and even classes to take advantage of what users have reported, is the core element that distinguishes our approach from most crowdsourcing applications. For instance, "Vertical jump" was a common "Other" report early on and was therefore upgraded to a dropdown option (cf. Section 4.2.1). We comment on the limitations of the current approach and how they can be remedied in Section 5.

4. Results

4.1. Usage

To date, 112 users have signed up on the DES exposure checker website, constituting a considerable fraction of the scientists and students in the DES Collaboration. For the first two data releases of DES, called SVA1 (with observations from 2012–2013) and Y1A1 (from 2013–2014), they have inspected a total of 21,147 images and submitted 39,213 reports.[15]

As indicated in top panel of Figure 4, only a few users are responsible for the majority of the submissions. Figure 6 shows

---

[12]We did not provide a facility to discuss the possible origin of artifacts as part of the tutorial, but rather relied on the DES collaboration wiki pages. These discussions were thus documented in a way that was familiar to survey participants and the conclusions were summarized in the tutorial.

[13]In this context, false positive refers to cases where the processing pipeline masked an area of an image without any visible artifact.

[14]The description field in the viewer page pre-loads existing descriptions for swift selection and to avoid description mismatch e.g. from typos.

[15]To facilitate inspection of the images and discussion within the collaboration, the viewer can be used without registration or the requirement to submit problem reports. The number of image views exceeds 250,000.

in more detail that about one half of the users inspected dozens or hundreds of images. When looking at the nearly flat distribution of number of problem reports per user (of which there can be several for each image), it becomes even more apparent that the most prolific users also attempted to be the most thorough.

To support studies of user reliability, we do not randomly present images to each user. Instead, we seek to maximize the number of images that have been inspected by at least three different users. Because a sizable fraction of the reports come from a small number of users, this cannot always be achieved, and the average number of distinct users per image is 1.58. Whenever multiple users inspected the same images, we found that they identified the same flaw(s) in 37% of the images, with the most common form of disagreement being caused by more subtle flaws identified by one user but not by others. This type of incompleteness is not a major concern for achieving Objective (3), because any generated problem report will be conditioned on at least one user having found the problem in question. The second most common report mismatch stems from actual disagreement regarding a flaw's classification, followed by the rare case, in which two users identified two distinct and mutually exclusive problems. We refer to Section 5 for a discussion about the advantages of a user reliability assessment.

4.2. Benefits for DES

Our main aim, besides informing survey participants, is to discover flaws in the data so that they can be corrected. Our effort is similar to the work of e.g. the "LSC glitch group", which visually monitored transients for LIGO and identified several instrumental causes for false alarms (Blackburn et al., 2008). To facilitate the information exchange we provide an API to retrieve aggregated and anonymized reports for individual problems. Data processing experts can thus obtain lists of images and locations of problematic features to test and improve the quality of various image reduction and data processing steps. These lists are especially useful for issues that occur rarely, where it would take a single user a significant amount of time to assemble a sufficiently large data set.

4.2.1. Shutter failure during flat field exposures

The first interesting finding based on problem reports is related to the exposure shown in Figure 1, specifically the vertical banding feature in the center of the image. This was one of the first artifacts identified by users of the exposure checker website, and was found to be present in a large number exposures from the SV period of DES. The problem reports provided critical information for an investigation of the origin of the flaw. It was ultimately traced back to a number of flat field images that were read out while the camera shutter was partially open, causing a strong gradient in horizontal direction. Because several flat field images are combined to suppress the variance of the correction and the resulting so-called "super-flat" is then applied to all images taken over a period of approximately two weeks, the flaw was present in about 5% of all images of the SVA1 data, making it relatively easy to find. However, the amount of information about the artifact and the speed with

which that information became available through the exposure checker API, enabled a rapid re-design of the super-flat production as well as a change to the data acquisition that prevented the problem from occuring in subsequent data releases. As shown in Figure 5, where this artifact is labeled "Vertical jump", the problem was prevalent in SVA1 but vanishes in Y1A1 after the aforementioned changes.

4.2.2. Orbital satellite trails

The DES exposures are 90 seconds in g, r, i, z-bands and 45 seconds in Y-band. During this time, earth-orbiting satellites can traverse a significant fraction of the DES focal plane leaving extended bright trails across multiple CCDs (a trail segment can be seen in the central panel of Figure 2). Unlike the shutter failures, this artifact was expected before the implementation of the exposure checker. However, an algorithm had not been developed for identifying and masking this contamination. Humans excel at perceiving linear features in noisy data (e.g. Quackenbush, 2004), and the exposure checker users quickly identified a large sample of trails. While not specifically designed for this purpose, the background subtraction and coarse binning we chose for the viewer proved beneficial for enhancing the contrast between satellite trails and the sky background. The automated algorithm developed for satellite trail detection took advantage of the same background subtraction and binning steps employed by the exposure checker, and then applied a Hough transform (Hough, 1959; Duda & Hart, 1972) for satellite trail detection and characterization. Due to the efficiency of the human eye in recognizing trails, the appropriate problem reports from the exposure checker was used as a truth sample for testing and optimizing the performance of the automated algorithm. Because we also ask our users to report false-positive satellite masks (cf. Figure 5), the parameters of the algorithm were adjusted such that it does not over-aggressively trigger for proper elongated features like large elliptical galaxies. To date, the automated algorithm achieves a true-positive rate of ~80% with a false-positive rate of ~2%.[16]

4.2.3. SExtractor background estimation

During investigation of the first data release, users noticed a systematic darkening of the sky background near the corners of many CCDs (see left panel of Figure 7). Initially labeled "dark corners" or "dark boundary", it was one of the most common problem reports from the hodge-podge list of that data release. Occurring in crowded stellar fields (e.g. close to the Large Magellanic Cloud) or when bright stars happen to fall close to the boundaries of CCDs, further investigation revealed that these artifacts stem from the sky background estimation performed by SExtractor (Bertin & Arnouts, 1996). Once identified, this issue was subsequently mitigated by the data processing team through adjustment of the SExtractor parameters.

---

[16]Nearly all of the trails missed by the automated algorithm have a contrast of $< 3\sigma$ with respect to the sky background. These trails are clearly visible to the human eye, but are algorithmically difficult to distinguish from elongated elliptical galaxies and large-scale imperfections in the background model.
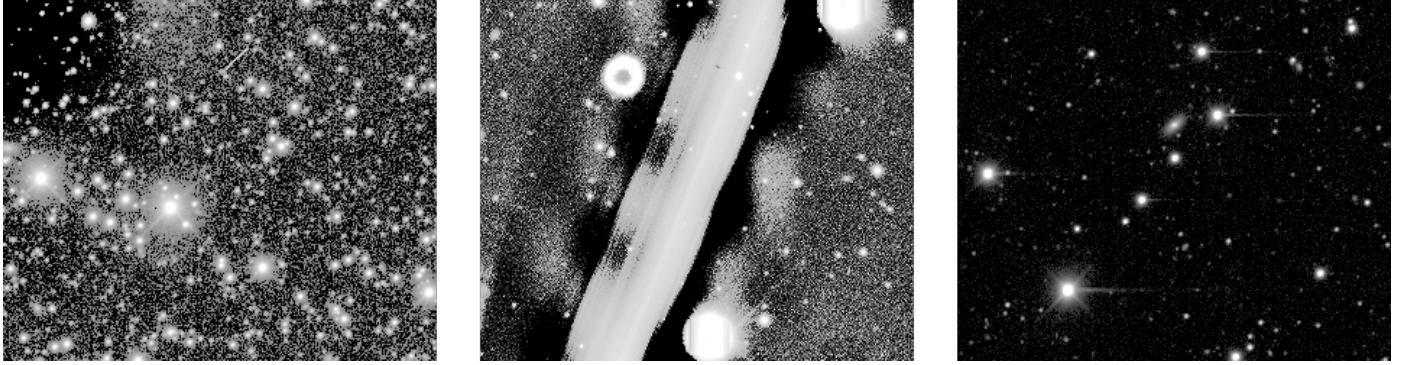
Figure 7: Example problem cases. *Left:* Sky background estimation was found to be problematic in dense stellar fields, particularly at the edges and corners of the CCDs (cf. Section 4.2.3). *Center:* The light of airplanes ruins large portions of affected exposures (the image shown covers almost an entire CCD; cf. Section 4.2.4). *Right:* A small number of CCDs shows stars with one-sided trails, the cause of which is currently under investigation (cf. Section 4.2.7).

4.2.4. Airplanes

Two commercial Chilean flight paths pass close to CTIO. As opposed to bright orbital satellites which impact at most 10% of the CCDs out of a full exposure, airplanes may render a large portion of the focal plane entirely unusable for scientific purposes (see central panel of Figure 7). While the occurrence of airplanes in the DES imaging is rare (several exposures per season), their tremendous impact makes it critical that affected exposures be identified and removed. Because airplanes can completely ruin a single-CCD image, the field-of-view visualization in the viewer page is particularly helpful to allow users to see their entire track across the DECam focal plane. Additionally, once an airplane is spotted in a single CCD, users can quickly navigate to other affected CCDs and tag them, too. Users identified four airplane tracks in the DES Science Verification data release and the affected CCDs were subsequently excluded from further analysis.

4.2.5. Spurious reflections

While every care was taken to avoid scattering surfaces during the construction of DECam, some scattered light artifacts are visible when observing in the vicinity of bright stars. They appear as large-scale diffuse features with characteristic morphologies (e.g. the shape of the telescope pupil) originally identified by the eyeball squad and easily visible in exposure checker images. Since scattered light artifacts often span multiple CCDs, this is another problem class that benefits from the field-of-view visualization: often the characteristic patterns of scattered light artifacts are only clearly recognizable when multiple CCDs can be examined (cf. Figure 3).

Feedback from the exposure checker provided a test sample to validate automated masking algorithms, which can predict the location of the most intense reflections from a detailed ray-tracing simulation of the optical path, and to identify other sources of scattered light. Whereas some scattered light features are unavoidable,[17] others, such as reflections off of the

walls of the filter changer, were mitigated by judicious application of anti-reflective paint once the cause was identified (Section 6.3 in Flaugher et al., 2015).

4.2.6. Cosmic Rays

Thousands of cosmic rays interact with the DECam CCDs during each DES exposure. Cosmic rays deposit excess charges into pixels leaving tracks that are much narrower than the point spread function of the image. During SVA1 processing the cosmic-ray masking algorithm was known to not perform optimally. This poor performance was independently confirmed by the results of the exposure checker. An improved masking algorithm was implemented for Y1A1 and the improved performance was validated by exposure checker users. We caution that while users are clearly able to identify catastrophic failures in cosmic-ray masking, the small sizes and large numbers of cosmic-ray tracks make it unlikely that the reports are complete. This renders it difficult to quantify the detailed performance of an algorithm based on the exposure checker output alone.

4.2.7. Star trails

An example of a relatively rare but visually recognizable feature is shown in the right panel of Figure 7. Dubbed "horizontal star trails", it has been identified by several users in 45 images spread over the SVA1 and Y1A1 releases. Our subsequent investigation revealed that all reports refer to images from only three CCDs, that these CCDs are far apart in the focal plane, and that neighboring CCDs of the same exposure are unaffected. The trails occur behind (in readout direction) moderately bright stars and are preferentially clustered on one half of the affected CCDs. A detailed investigation of this artifact is underway.

5. Lessons learned

Visual quality control is a feasible application of crowd-sourcing. We operate under the assumption that participants

---

[17] for example, the image "ghosting" from double reflections of light off of

the CCDs despite their anti-reflective coating

in scientific experiments, such as the optical imaging survey DES, are genuinely interested in data flaws relevant to their science cases. Indeed, we found—and have been told—that if the interaction with the application is streamlined and frustration-free, our users enjoyed exploring the data, either to inspect the quality of new releases or, for new participants, to get to know their current state. Many users suggested improvements to the applications (e.g. the field-of-view visualization) that extended its capabilities and made the exploration more efficient and illuminating. We also found that users can be further motivated by newsletters, gaming incentives, such as badges and leaderboards, and follow-up investigations of unexpected discoveries. These findings all point to an engaged user base, and our usage statistics support such an interpretation.

By now, users have submitted more than 39,000 reports. Many of the reported flaws were initially not expected, so we created the umbrella class "Other", for which user-defined labels are required. By rank-ordering those labels, we could identify the most common types of flaws and extend our classification scheme to currently 26 distinct classes. The tutorial, which summarizes the classes and the current knowledge of their causes, has become an often-visited visual catalog of DES data flaws, and has clear educational value for survey participants. We furthermore provide an API to query anonymized problem reports. This database has been instrumental to identify the origin of several problem classes and to generate specific test cases for algorithms that mitigate or entirely eliminate the impact on subsequent science analyses.

Our findings are applicable to upcoming astronomical surveys and other extensive but error-prone data sets with a sufficiently large user base. We see clear benefits of the crowd-sourcing approach to rapidly improve data quality in the early phases of an experiment, and to educate participants about its real-world limitations. In the light of growing data volumes and rates of forthcoming surveys such as LSST,[18] we emphasize that a complete inspection of all images is not necessary to reveal various kinds of problems present in the data. When faced with truly overwhelming data volumes, we recommend pre-selecting images with potential problems for human inspection to keep the size of an effective user base within reasonable limits and to provide enough problems, so that the users remain motivated by their discoveries.

The application code currently operates with images in the astronomical FITS format but can be modified to work with any two-dimensional data representation. A demo version of the application as employed by DES is hosted online[19], and we publicly release the source code.[20]

Potential improvements

- While we recommend that users read the tutorial, there was no formal training process. To establish a more informed and confident user base, an interactive introduction could be made mandatory, wherein each new user is presented with several thoroughly vetted problem cases of increasing difficulty. While such a requirement could be perceived as onerous labor and might reduce the number of active users from the beginning, it would likely improve the classification accuracy. Following Marshall et al. (2016), we thus recommend a short and engaging training session for new users to even out differences in professional experience.

- The number of recognized problem classes has by now grown to 26. To help users during classification, prototypical images from the tutorial could be made available on the viewer page as a swift visual index of the classification, for instance in the form of a collapsible side bar.

- To maximize the utility of the problem reports with respect to Objective (3), i.e. to help inform the improvement of data processing algorithms, we need problem classifications with small type-I error, whereas type-II errors are irrelevant.[21] In this context, it is important to note that most mismatches in problem reports stem from type-II errors, whereas cases where users identify the same flaw but classify it differently (thereby causing at least one type-I error) are less frequent. Nonetheless, type-I errors could be reduced if we knew how often they occur for any given user, which would allow us to weigh reports according to the probability $p(c\,|\,u)$ of any user u correctly classifying the problem of interest c. Lintott et al. (2008) suggested a scheme, in which the agreement with the majority classification for objects with multiple submissions from different users determines the user weights. Because it is difficult to produce sufficient redundancy with a smaller user base, we prefer an alternative proposed by Marshall et al. (2016): The necessary information can be gathered by occasionally inserting training images into the normal stream presented in the viewer and tracking how often users correctly identify flaws. We note that verifying whether a user has found a pre-determined problem is a non-trivial task for problem classes with large spatial extent (like ghosts or reflections): In contrast to simpler distance-based success criteria, complex polygonal shapes would be required for such training images to determine if the markers were placed within respective target areas.

- The random insertion of training images would also help to re-calibrate the overall problem statistics, i.e. the total number $N(c)$ of problem reports for class c as shown in Figure 5. One can imagine a case where a given user is rather impervious to a problem class, biasing its reported abundance low, or a case of a thorough user who inspects neighboring CCDs when they might be affected by the same problem, causing the abundance to be biased high. By determining how often users report a particular

---

problem when it has been presented to them in a hidden test, $N_t(c\,|\,u)$, one can estimate a corrected abundance as $\sum_u N(c\,|\,u)/N_t(c\,|\,u)$.[22]

- The hodge-podge list of unknown problem classes currently ranks new cases by their total occurrence. Hence, type-I errors pose the same kind of problem as for the pre-determined classification scheme, namely if one user mistakenly starts to identify some aspect of the image as a new class of problems when it is a known class or no problem at all. The ranking could be improved if the problem reports are weighed with $\sum_c p(c\,|\,u)$, i.e. the probability of the submitting user u correctly classifying *all* existing problems.

- An alternative mechanism to rank the hodge-podge list, potentially in addition to simple or weighted occurrence, is user voting. By giving each user a vote to either agree or disagree with the existence of the proposed problem class, one could reduce the impact of poorly informed or even rogue users who continue to assign a particular description to a problem that is either already known or non-existent.[23]

- It could be beneficial to discuss if an image is problematic or what kind of flaw can be observed directly within the viewer, e.g. in a discussion thread below the image, instead of on a separate wiki page. For instance, in case of uncertain classifications from the hodge-podge list, experienced users could provide their comments and help clarify whether the flaw is of a new kind. The downside of such a solution is that these discussions would be spread out over many distinct images, whereas a wiki page provides a central location for a coordinated discussion. Such a downside could be overcome through the implementation of a discussion forum where individual images could be discussed and two-way references could be formed between images and forum topics.

- The gallery of "Awesome" objects comprises many of the most extended or otherwise remarkable objects of the survey. An automated script that queries our API and creates multi-color images of all objects on the awesome list would visibly recognize the effort of the users and provide a valuable resource for outreach purposes.

We encourage any interested party to contact us with further suggestions of improvements through the issue tracker system of the code repository.

---

[22]Note that for the former case $N_t(c\,|\,u) = p(c\,|\,u)$, whereas the latter case requires that reports of the same class for images inspected by the same user directly after the test images need to be combined.

[23]Due to the small user base of professional collaborators who mostly know each other, we never experienced intentionally malicious behavior.

References

References

Balbinot E., Santiago B., Girardi L., da Costa L. N., Maia M. A. G., Pellegrini P. S. S., Makler M., 2012, in Astronomical Society of the Pacific Conference Series Vol. 461, p. 287

Bertin E., Arnouts S., 1996, Astronomy and Astrophysics Supplement Series, 117, 393

Blackburn L. et al., 2008, Classical and Quantum Gravity, 25, 184004

Dark Energy Survey Collaboration 2005, arXiv:astro-ph/0510346

Desai S. et al., 2012, ApJ, 757, 83

Diehl H. T. et al., 2014, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series Vol. 9149, p. 91490V

Duda R. O., Hart P. E., 1972, Commun. ACM, 15, 11

Flaugher B. et al., 2015, AJ, 150, 150

Greenhill A., Holmes K., Lintott C., Simmons B., Masters K., Cox J., Graham G., 2014, in GAME-ON 2014 15th International Conference on Intelligent Games and Simulation, pp 15–24

Honscheid K. et al., 2012, in Software and Cyberinfrastructure for Astronomy II, pp 845112–845112–15

Hough P. V. C., 1959, in Proceedings, 2nd International Conference on High-Energy Accelerators and Instrumentation, HEACC 1959 Vol. C590914, pp 554–558

Jones I., 1989, The Last Crusade

Kapadia A., Smith A., 2013, in Astronomical Society of the Pacific Conference Series Vol. 475, p. 299

Lintott C. J. et al., 2008, Monthly Notices of the Royal Astronomical Society, 389, 1179

Littmann M., Suomela T., 2014, Endeavour, 38, 130

Lupton R. H., Gunn J. E., Szalay A. S., 1999, AJ, 118, 1406

Marshall P. J. et al., 2016, MNRAS, 455, 1171

Mohr J. J. et al., 2012, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series Vol. 8451, 84510D

Olmsted D., 1834a, American Journal of Science and Arts, 25, 363

Olmsted D., 1834b, American Journal of Science and Arts, 26, 132

Quackenbush L. J., 2004, Photogrammetric Engineering & Remote Sensing, 70, 1383

York D. G. et al., 2000, AJ, 120, 1579