Mon. Not. R. Astron. Soc. **000**, 1–18 (2014) Printed 5 November 2014 (MN LATEX style file v2.2)

SPACE WARPS: Crowd-sourcing the Discovery of Gravitational Lenses

Phil Marshall,[1,2⋆] Aprajita Verma,[1] Anupreeta More,[3] Amit Kapadia,[4] Michael Parrish,[4] Chris Snyder,[4] Julianne Wilcox, Elisabeth Baeten, Christine Macmillan, Claude Cornen, Chris Davis,[2] Surhud More,[3] Michael Baumer,[2] Chris Lintott,[1] Robert Simpson,[1] David Miller,[4] Arfon Smith,[4] Edward Paget,[4] Prasenjit Saha,[5] Rafael Kueng,[5] Edwin Simpson,[1] Kelly Borden,[4] Tom Collett, Thomas Jennings, Matthias Tecza,[1] Layne Wright and possibly others

[1] *Dept. of Physics, University of Oxford, Keble Road, Oxford, OX1 3RH, UK*
[2] *Kavli Institute for Particle Astrophysics and Cosmology, Stanford University, 452 Lomita Mall, Stanford, CA 94035, USA*
[3] *Kavli IPMU (WPI), University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa 277-8583, Japan*
[4] *Adler Planetarium, Chicago, IL, USA*
[5] *Department of Physics, University of Zurich, Switzerland*

to be submitted to MNRAS

ABSTRACT

SPACE WARPS is a web-based service that enables the discovery of strong gravitational lenses in wide-field imaging surveys by large numbers of people. Carefully produced color composite images are displayed to volunteers via a classification interface which records their estimates of the positions of candidate lensed features. Simulated lenses, and expert-classified non-lenses, are inserted into the image stream at random intervals; this training set is used to give the volunteers feedback on their performance, and to estimate a dynamically-updated probability for any given image to contain a lens. Low probability systems are retired from the site periodically, concentrating the sample towards a set of candidates; this "stage 1" set is then re-classified by the volunteers in a second refinement stage. Analyzing the classification of the training set, we predict that the first stage alone should yield a sample that is C% complete, while leading to the rejection of R% of the initial target sample. Having divided the 150 square degree CFHTLS imaging survey into 430000 overlapping 70 by 70 arcminute tiles and displayed them on the site, we were joined by 33000 volunteers who contributed X million image classifications over the course of N months. The sample was reduced to 3500 stage 1 candidates; these were then refined to yield a sample of 1400 candidates rankable by their stage 2 probability. We expect this sample to be X% complete and Y% pure at a threshold of 95% classification probability. We find that, on average, and given the assumptions we make in our analysis, we need 9 classifications per image during the first stage, X in the second. We estimate the mean information contributed per person to be X bits, over a session lasting, on average, N classifications per volunteer, and present the highly skewed distributions of these quantities. We comment on the scalability of the SPACE WARPS system to the wide field survey era, and its potential to operate beyond its design as a supervised classification system.

To Do (Phil): revisit abstract (#29).

Key words: gravitational lensing – methods: statistical – methods: citizen science

⋆ pjm@slac.stanford.edu

1 INTRODUCTION

Strong gravitational lensing – the formation of multiple, magnified images of background objects due to the deflection of light by massive foreground objects – is a very powerful astrophysical tool, enabling a wide range of science projects. The image separations and distortions provide information about the mass distribution in the lens (e.g. Auger et al. 2010b; Sonnenfeld et al. 2012, 2013), including on subgalactic scales (e.g. Dalal & Kochanek 2002; Vegetti et al. 2010; Hezaveh et al. 2013). Any strong lens can provide magnification of a factor of 10 or more, providing a deeper, higher resolution view of the distant universe through these "cosmic telescopes" (e.g. Stark et al. 2008; Newton et al. 2011). Lensed quasars enable cosmography via the time delays between the multiple images' lightcurves (e.g. Tewes et al. 2013; Suyu et al. 2013), and study of the accretion disk itself through the microlensing effect (e.g. Poindexter et al. 2008). All of these science projects would benefit from being able to draw from a larger sample of lenses.

In the last decade the numbers of detections of these rare cosmic alignments has increased by an order of magnitude, thanks to wide field surveys such as CLASS (Browne et al. 2003, e.g.), SDSS (e.g. Bolton et al. 2006; Auger et al. 2010a; Treu et al. 2011; Inada et al. 2012), CFHTLS (e.g. More et al. 2012; Gavazzi et al. 2014), Herschel (Negrello et al. 2014) and SPT (e.g. Vieira et al. 2013), among others. As the number of known lenses has increased, new types have been discovered, leading to entirely new investigations. Compound lenses (Gavazzi et al. 2008; Collett et al. 2012) and lensed supernovae (Quimby et al. 2014) are good examples of this.

Because they are rare, strong lenses are expensive to find. The most efficient searches to date have made use of relatively clean signals such as the presence of emission or absorption features at two distinct redshifts in the same optical spectrum (e.g. Bolton et al. 2004), or the strong "magnification bias" towards detecting strongly-lensed sources in the sub-mm waveband (e.g. Negrello et al. 2010). Such searches have to be efficient, because they require expensive high resolution imaging follow-up; consequently they have so far produced yields in the tens to hundreds. An alternative approach is to search images of sufficiently high resolution and color contrast, and confirm the systems as gravitational lenses by modeling the survey data themselves (Marshall et al. 2009). Several square degrees of HST images have been searched, yielding several tens of galaxy-scale lenses (e.g. Moustakas et al. 2007; Faure et al. 2008; Jackson 2008; More et al. 2012; Pawase et al. 2014). Similarly, searches of over a hundred square degrees of CFHT Legacy Survey ground-based imaging, also with sub-arcsecond image quality, have revealed a smaller number of wider image separation group-scale systems (e.g. Cabanac et al. 2007; More et al. 2012). Detecting galaxy-scale lenses from the ground is hard, but feasible albeit lower efficiency and requiring HST or spectroscopic follow-up to confirm the candidates as lenses (e.g. Gavazzi et al. 2014).

How can we scale these lens searches up to imaging surveys covering a hundred times the sky area, such as the almost-all sky surveys planned with LSST and Euclid, while reducing our dependence on expensive follow-up confirmation observations? There are two approaches to detecting lenses in imaging surveys. The first one is robotic: automated analysis of object catalogs and/or the survey images. The candidate samples produced by these methods have, to date, not been of high purity (see e.g. Marshall et al. 2009; More et al. 2012; Gavazzi et al. 2014), with visual inspection by teams of humans still required to narrow down the robotically-generated samples. In this approach, the image data may or may not be explicitly modelled by the robots as if it contained a gravitational lens, but the visual inspection can be thought of as a "mental modeling" step. Systems classified by an inspector to be good lens candidates are deemed as such because the features in the image can be explained by a model of what gravitational lenses do contained in the inspector's brain. The second approach simply cuts out the robot middleman: Faure et al. (2008); Jackson (2008) and Pawase et al. (2014) all performed entirely visual searches for lenses in HST imaging.

Visual image inspection seems, at present, unavoidable at some level when seraching for gravitational lenses. The technique has some drawbacks, however. First is that humans are only humans, and they make mistakes. The solution to this is to operate in teams, providing multiple classifications of the same images in order to catch errors and correct them. Second, and relatedly, is that humans get tired. With a well-designed classification interface, a human might be able to inspect images at a rate of one astronomical object per second (provided the majority are indeed uninteresting). At 10^4 massive galaxies, and 10 lenses, per square degree, visual lens searches in good quality imaging data are limited to a few square degrees per inspector per day. Scaling to thousands of square degrees therefore means either robotically reducing the number of targets for inspection, or increasing the number of inspectors, or both.

For example, a 10^4 square degree survey containing 10^8 photometrically-selected massive galaxies and 10^5 lenses could only be searched by 10 inspectors at a mean rate of 1 galaxy per second and 10 inspections per galaxy in about 14 years. Reducing the inspection time by a factor of 400 to two weeks would require a robot to reduce the target sample to 25 per square degree. However, at this point the required purity, 40%, would very likely require the average classification time per object to be more like 10 seconds per object. Hiring 10 inspectors to assess complex images full time full time for five months may not be the most cost-effective or reliable strategy. Alternatively, a team of 10^6 inspectors could, in principle, make the required 10^9 image classifications, 10^3 each, in a few weeks; robotically reducing the target list would lead to a proportional decrease in the required team size.

Systematic detection of rare astronomical objects by such "crowd-sourced" visual inspection has recently been achieved by the online citizen science project PlanetHunters (Schwamb et al. 2012). PlanetHunters was designed to enable the discovery of transiting exoplanets in data taken by the Kepler satellite; a community of N inspectors from the general public found, after each undergoing a small amount of training, N new exoplanet candidates by visual inspection of the Kepler lightcurves that were presented in a custom web-based classification interface. The older Galaxy Zoo morphological classification project (Lintott et al. 2008) has also enabled the discovery of rare objects, via its flexible inspection interface and discussion forum (Lintott et al. 2009).

Indeed, several of us (AV,EB,CC,TJ,CM,LW) were active in an informal Galaxy Zoo gravitational lens search, an experience which led to the present hypothesis that a systematic online visual lens search could be successful.

In this paper, we describe the Space Warps website, an online system that enables crowd-sourced gravitational lens detection by inviting volunteers to classify astronomical survey images as containing lens candidates or not. In a companion paper (More et al, in preparation, herafter Paper II) we will present the new gravitational lenses discovered in our first experimental lens search, and begin to investigate the differences between lens detections made in Space Warps and those made with automated techniques. Here though, we try to answer the following questions:

- How reliably can we find gravitational lenses using the Space Warps system? What is the completeness of the sample produced?
- How noisy is the system? What is the purity of the sample produced?
- How quickly can lenses be detected, and non-lenses be rejected? How many classifications, and so how many volunteers are needed per target?
- What can we learn about the scalability of the crowdsourcing approach?

Our basic method in this paper is to analyze the performance of the Space Warps system on a "training set" of simulated lenses and known non-lenses. This allows us to estimate completeness and purity with respect to gravitational lenses that have the same properties of the training set. In Paper IIwe carry out similar tests but with a sample of known (real) lenses *in situ*.

This paper is organised as follows. In Section 2 we introduce the Space Warps classification interface and the volunteers who make up the Space Warps collaboration, explain how we use training images, and describe our two stage candidate selection strategy. We then briefly introduce, in Section 3 the particular dataset used in the first experimental tests of the Space Warps system, and how we prepared the images prior to displaying them in the web interface. In Section 4 we describe our methodology for interpreting the classifications made by the volunteers, and then present the results of system performance tests made on the training images in Section 5. We discuss the implications of our results for future lens searches in Section 6 and draw conclusions in Section 7.

2 EXPERIMENT DESIGN

The basic steps of a visual search for gravitational lenses are: 1) prepare images, 2) display them to an inspector, 3) record the inspector's classification of each image (as, for example, containing a lens candidate or not) and 4) analyzing those classifications (and all others) in order to produce a final candidate list. We describe step 1 in Section 3 and step 4 in Section 4. In this section we take a volunteer's eye view and begin by describing the Space Warps classification interface, the crowd of volunteers, and the interactions between the two.

2.1 Classification Interface

A screenshot of the Space Warps classification interface (CI) is shown in Figure 1. The CI is the centrepiece of the Space Warps website, `http://spacewarps.org`; the web application is written in coffeescript, css and html and follows the general design of others written by the Zooniverse team.[1] The focus of the CI is a large display of the current pre-prepared PNG image of the "subject" being inspected. When the image is clicked on by the volunteer, a marker symbol appears where the pointer was. Several markers can be placed. The next image moves rapidly in from a queue formed at the right hand side of the screen when the "Finished marking" button is pressed. At the same time, the positions of the markers are written out to the classification database, in an entry that also stores the ID of the subject, the username (or IP address) of the volunteer, a timestamp and some other metadata.

Gravitational lenses are rare: typically, most of the images will not contain a lens candidate, and these need to be quickly rejected by the inspector. The queue allows several images to be pre-loaded while the volunteer is classifying the current subject, and the rapid movement is designed to encourage volunteers to classify rapidly.

For the more interesting subjects, the CI offers two features that enable further investigation of the subjects. First is the "Quick Dashboard" (QD) a more advanced image viewer. This allows the viewer to compare three different contrast and color balance settings, to help bring out subtle features, and to zoom in on interesting regions of the image to assess small features. Markers can be placed in the Quick Dashboard just the same as in the main CI image viewer. The second is a link to that subject's page in the project discussion forum, `http://talk.spacewarps.org` (known as "Talk"). Here, volunteers can discuss the features they have seen either before they submit their classification, or after, if they "favorite" the subject. There is no "back" button: each volunteer may only classify a given subject once. However, the presence of an option to see what others think about any given subject before submitting your own classification means that the classifications may not be strictly independent; the advantage of this system is that volunteers can learn from others what constitutes a good lens candidate. In practice, we might expect this to be a relatively unimportant educational resource, given the explicit training we provide for the volunteers, which we describe in the next section.

2.2 Training

Gravitational lenses are unfamiliar objects to volunteers who are new to the site. New volunteers need to learn what lenses look like as quickly as possible, so that they can contribute informative classifications. They also need to learn what lenses do not look like, in order to reduce the false positive detection rate. There are three primary mechanisms in the Space Warps system for teaching the volunteers what to

[1] The Space Warps web application code is open source and can be accessed from `https://github.com/zooniverse/Lens-Zoo`

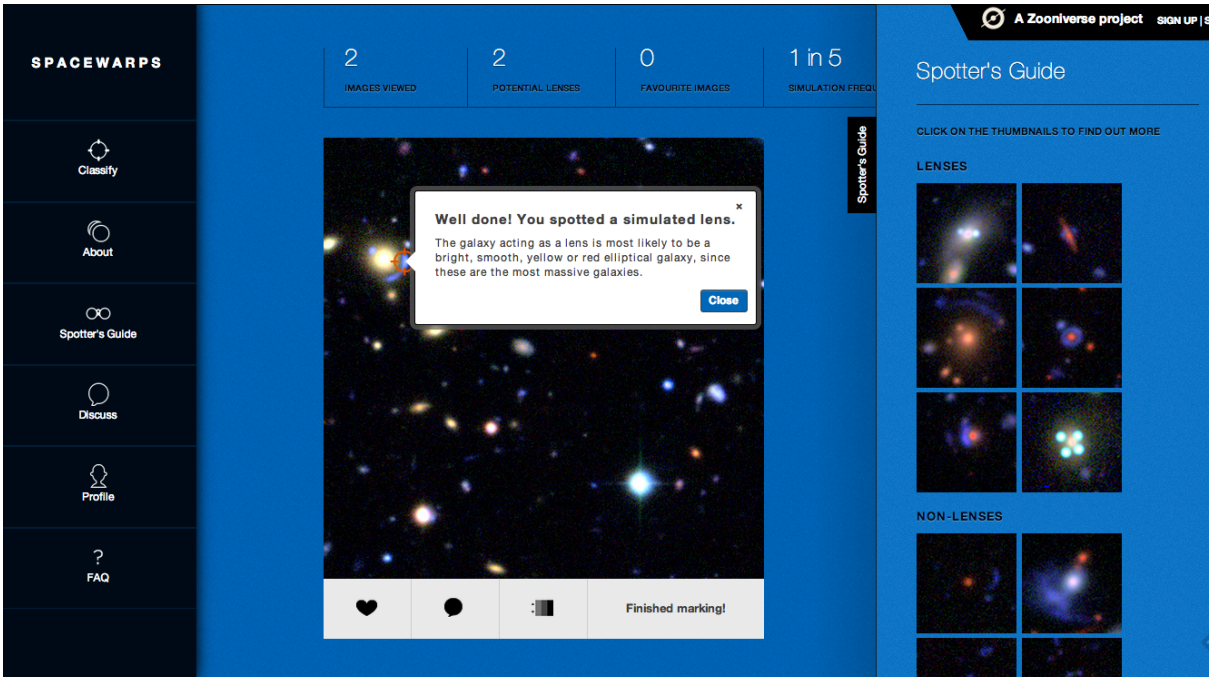

Figure 1. Screenshot of the Space Warps classification interface.

look for. These are, in the order in which they are encountered, an inline tutorial, instant feedback, and a "Spotter's Guide."

2.2.1 Inline Tutorial

New volunteers are welcomed to the site with a very short tutorial, in which the task is introduced, a typical image containing a simulated lens is displayed, and the marking procedure walked through, using pop-up message boxes. Subsequent images gradually introduce the more advanced features of the classification interface (the QD and Talk buttons), also using pop-up messages. The tutorial was purposely kept as short as possible so as to provide the minimal barrier to entry.

2.2.2 Training Subjects and Instant Feedback

The second image viewed after the initial tutorial image is already a survey image, in order to get the volunteers engaged in the real task as quickly as possible. Training continues beyond the first image tutorial through "training subjects" inserted randomly into the stream. These training subjects are either simulated lenses (known as "sims"), or survey images that were expert-classified (by AV, AM and PM) and found not to contain any lens candidates (these images are known as "duds"). The tutorial explains that the volunteers will be shown such training images. They are also informed that they will receive instant feedback about their performance after classifying (blind) any of these training subjects. Indeed, after a volunteer finishes marking a training subject and hits "Finished marking," a pop-up message is generated, containing either positive feedback for a successful classification (for example, "Well done! You spotted a simulated lens," as in Figure 1) or negative feedback for an unsuccessful one (for example, "There is no gravitational lens in this field!")

The initial frequency of the training images is set to be two in five; subjects are drawn randomly from the pool of training images with this frequency. The pool contains equal numbers of sims and duds, and the draw is made without replacement (for that volunteer). As the number of classifications made by a volunteer increases, this frequency is decreased, to $2/(5 \times 2^{(\mathrm{int}(N_c/20)+1)/2})$ (≈ 0.3 for the second 20 subjects, 0.2 for the third 20 subjects, and so on).

This training regime means that in the first 60 images viewed, each volunteer is shown (on average) 9 simulated gravitational lenses, and 9 empty fields. This is a much higher rate than the natural one: to try and avoid this leading to over-optimism among the inspectors (and a resulting high false positive rate), we display the current "Simulation Frequency" on the classification interface ("1 in 5" in Figure 1) and maintain the consistent theme in the feedback messages that lenses are rare.

In Figure **??** we show 12 example training images from the first Space Warps project (Section 3 below).

To Do (Phil, Aprajita): Make gallery of 6 stage 2 sims and 6 stage 2 duds, to give the reader a feel for what the sims look like, and also how the images were presented, stretch and scale-wise.

2.2.3 Spotter's Guide

The instant feedback provides real-time educational responses to the volunteers as they start classifying; as well as this dynamic system, Space Warps provides a static reference work for volunteers to consult when in doubt about how to perform the task. This "Spotter's Guide" is a set of webpages showing example lenses, both real and simulated,

and also some common false positives, drawn from the pool of survey images. The non-lenses were identified by three of us (AV, AM and PM) while inspecting a small set of survey images in order to define the "dud" training images. For easy reference, the lenses are divided by type (for example, "lensed galaxies," "lensed quasars" and "cluster lenses"), as are the false positives (for example, "Rings and Spirals," "Mergers," "Artifacts" and so on). The example images are accompanied by explanatory text. The Spotter's Guide is reached via a button on the left hand side, or the hyperlinked thumbnail images of the "Quick Reference" provided on the right hand side, of the classification interface.

Most of the text of the Spotter's Guide focuses on what lenses do or don't do; the website "Science page" contains a very brief introduction to how gravitational lenses work, which is fleshed out a little on the "FAQ" page. This also contains answers to frequently asked questions about the interface and the task set.

2.3 Staged Classification

We now describe briefly the two-stage strategy that was employed in the CFHTLS project, initial classification (involving the rejection of very large numbers of non-lenses), and refinement (to further narrow down the sample). The web application was reconfigured between the two stages, to assist in their functioning.

2.3.1 Stage 1: Initial Classification

The goal of a stage 1 classification is to achieve a high rejection rate, while maintaining high completeness. In this mode, therefore, the pre-loading of images was used to make the sliding in of new subjects happen quickly, to provide a sense of urgency: initial classification must be done fairly quickly for the search to be completed within a reasonable time period. We expect some trade-off between speed and accuracy, which we return to in the results section below. Completion of the search requires subjects to be "retired" over time, as a result of their being classified. We do this by analyzing the classifications on a daily basis, as described in Section 4 below. As subjects are retired, new ones are ingested into the web app for classification. This means that the discovery of lens candidates in stage 1 is truly a community effort: to detect a lens candidate, many non-lenses must first be rejected.

The stage 1 training set was chosen to be quite clear cut, in order to err on the side of inclusivity, and so ensure high completeness. For the training duds we selected several hundred images at random, and three of us (AV, AM, PM) inspected them and discarded anything that could be a lens candidate. The remainder were ingested into the site.

2.3.2 Stage 2: Refinement

At stage 2, the goal is to define a final sample that has both high completeness and high purity. To this end, a more demanding training set was defined, where the duds all contain an object identified by one of us (AV, AM and PM) as a potential false positive. Figure **??** shows some example images from the CFHTLS stage 2 training set.

To Do (Phil): Show some example sims and duds...

We also attempted to encourage discernment by changing the look and feel of the app, slowing down the arrival of new images, and switching the background color to bright orange to make it clear that a different task was being set. The frequency with which training images were shown was fixed at 1 in 3. Finally, the Spotter's Guide was upgraded to include more detailed discussion of various possible false positives.

To Do (Phil): write this section (#31)

3 DATA

We refer the reader to Paper IIfor the details of the particular set of imaging survey data used in this first Space Warps project. Here, we summarize very briefly the choices that were made, in order to provide the context for our general description and illustrations of the Space Warps system.

3.1 The CFHT Legacy Survey

The CFHT Legacy Survey (CFHTLS, **?**) covered XXX square degrees of approximately equatorial sky, in four patches distributed in right ascension, over the course of N years. With homegeneous image quality of, typically, X arcsec (in the i-band), and reaching limiting magnitudes of XX-XX in the *ugriz* filter set, this survey has yielded several dozen new gravitational lenses on both galaxy and group scales (). The quality of the data, combined with the presence of these comparison "known lens" samples, made this a natural choice against which to develop and test the Space Warps system. The CFHTLS is also well-representative of the data quality expected from several next-generation sky surveys, such as DES, KiDS, HSC and LSST.

In order to investigate the completeness of the previous, semi-automated, lens searches in the CFHTLS area, we designed a "blind search." We divided the mosaiced images into some 430,000 equal size, overlapping tiles, approximately 70 arcsec on a side. Each of these images is a "test image" (as opposed to a "training image"). In future, larger area projects we expect to implement the rather different strategy of producing image tiles centred on particular preselected "targets," to make for a more efficient (but less complete) survey. We do not expect the performance of citizen image inspectors to change significantly between these strategies: to first order, both strategies require the inspectors to learn what lenses look like, and then search the presented images for similar features.

3.2 Image Presentation

The CFHTLS g, r and i-band images have the greatest depth and best image quality, and we chose to focus on this set (although the u and z-band images were also available for inspection in Talk). We made colour composite images

using publically-available code[2] following the prescription of **?**. Specifically, we first rescaled the pixel values of each channel image into flux units, and then applied an arcsinh stretch. The stretch parameters were chosen using a small random sample of images, to ensure that the background noise was just visible, and that the centres of bright, intermediate resdhift galaxies were not saturated. The color scales were chosen to maximize the contrast between faint extended objects. These parameters were then fixed during the production of all the tiles, in order to allow straightforward comparison between one image and another, and for intuition to be built up about the appearance of stars and galaxies across the survey. Alternative algorithms, such as adjusting the stretch and scale dynamically according to, for example, the root-mean-square pixel value in each image, can lead to better presentation of bright objects, but in doing so they tend to hide the faint features in those images: we needed to optimize the detectability of these faint features. Examples of CFHTLS training set images can be seen in Figure **??**.

4 CLASSIFICATION STAGES AND ANALYSIS

Having described the classification interface, the training images and the test images, we now outline our methodology for interpreting the interactions of the volunteers with the identification interface, and then describe how we applied this methodology in the two classification stages in the CFHTLS project in the Space Warps Analysis Pipeline (SWAP) code.[3]

Each classification made is logged in a database, storing subject IDs, (anonymous) volunteer IDs, a timestamp and the classification results. The *kind* of subject – whether it is a training subject (a simulated lens or a known non-lens) or a test subject (an unseen image drawn from the survey) – is also recorded. For all subjects, the positions of all Markers are recorded, in pixel coordinates. For training subjects, we also store the "classification" of the subject as a lens, or a non-lens, and also the type of object present in the image. These types are summarized in Table **??**. This classification is used to provide instant feedback, but is also the basic measurement used in a probabilistic classification of every subject based on all image views to date.

While the Space Warps web app is live, and classifications are being made, we perform an online analysis of the classifications, updating a probabilistic model of every (anonymous) volunteer's data, and also updating the lens probability of each subject (in both the training and test sets), on a daily basis. This gives us a dynamic estimate of the posterior probability for any given subject being a lens, given all classifications of it to date. Assigning thresholds in this lens probability then allows us to make good decisions about whether or not to retire a subject from the system, in order to focus attention on new images.

[2] The open source colour image composition code used in this work is available from `http://github.com/drphilmarshall/HumVI`

[3] The open source SWAP code is available from `https://github.com/drphilmarshall/SpaceWarps`

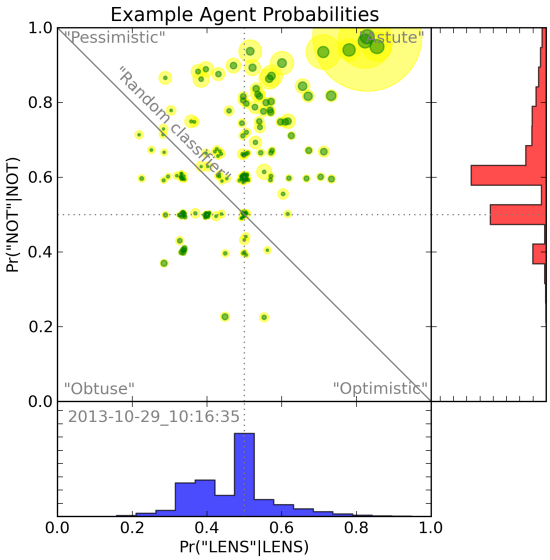

Figure 2. Typical Space Warps Agent confusion matrix elements. At a particular snapshot, 200 random Agents are shown distributed over the unit plane, with a tendency to move towards the "astute" region in the upper right hand quadrant as the Agents' volunteers view more images. Yellow point size is proprtional to the number of images classified; green point size shows Agent-perceived "skill."

The details of how the lens probabilities are calculated are given in Appendix A. In summary:

- Each volunteer is assigned a simple software agent, characterised by a confusion matrix. The two independent elements of this matrix are the probabilities, as estimated by the agent, that the volunteer is going to be 1) correct when they report that an image contains a lens when it really does contain a lens, Pr("LENS"|LENS, T), and 2) correct when they report that an image does not contain a lens when it really doesn't contain a lens, Pr("NOT"|LENS, T).
- Each agent updates its confusion matrix elements based on the number of times its volunteer has been right in each way while classifying subjects from the training set, accounting for noise early on due to small number statistics: T is the set of all training images seen to date.
- Each agent uses its confusion matrices to update, via Bayes' theorem, the probability of an image from the test set containing a lens, Pr(LENS|C, T), when that image is classified by its volunteer. (C is the set of all classifications made of this subject.)

Figure 2 shows a random selection of Agents' confusion matrix elements, as they were on a particular day towards the end of the CFHTLS project. Many volunteers classify only a small number of images, and so their Agents' confusion matrix elements remain close to their initial values of (0.5,0.5). As more images are classified (shown by the yellow point size), the Agents' matrix elements tend to move towards higher values, as the volunteers attain greater skill levels (green point sizes, see Section 5 below) and the Agent learns more about them. In this quadrant, the Agents perceive their volunteers to be "astute." This trend is more clearly seen in Figure 3, which shows the skill of the same

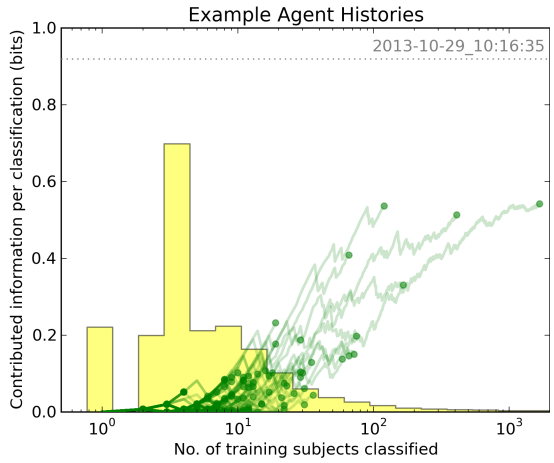

Figure 3. Typical SPACE WARPS Agent histories. The "skill" of the same 200 random Agents as in Figure 2 is plotted in green as a function of the number of subjects classified ("effort"). The yellow histogram in the background shows the distribution of effort.

sample of Agents as the number of images classified increases. The histogram shows the distribution of classification number: a long tail to very high "effort" can be seen.

In Section 5 below, we define several quantities based on the probabilities listed above that serve to quantify the performance of the crowd in terms of the information they provide via their classifications, and report on the performance of the system in returning a sample of lens candidates as a function of $\Pr(\mathrm{LENS}|C,T)$ threshold.

During stage 1 classification of the CFHTLS images, we assigned a prior probability for each image to contain a lens of 2×10^{-4}, based on a rough estimate of the number of expected lenses in the survey. We then assigned two values of the images' posterior probability, $\Pr(\mathrm{LENS}|C,T)$, to define "detection" and "rejection" thresholds. These were set to be 0.95 and (approximately symmetrically in the logarithm of probability), 10^{-7}. Subjects that attained probability of less than the rejection threshold were scheduled for retirement and subsequently ignored by the analysis code. Subjects crossing the detection threshold were not retired from the website, but instead left in the system so that more volunteers could see them. The progress of the subjects is illustrated in Figure 4. Subjects appear on this plot at the tip of the arrow, at zero classifications and prior probability; they then drift downwards as they are classified by the crowd, with each Agent applying the appropriate kick in probability based on what it hears its volunteer say. Encouragingly, sims (blue) tend to end up with high probability, while duds (red) pile up at low probability; test subjects (black) mostly drift to low probability, but some go the other way. The latter will help make up the candidate sample. As this plot shows, around 10 classifications are required for a subject to reach the retirement threshold.

The analysis code was run every night during the project, and subjects retired in batches after its completion. This introduced some inefficiency, because some classifications were accumulated in the time between them crossing the rejection threshold and the subject actually being retired from the website. As subjects were retired from the

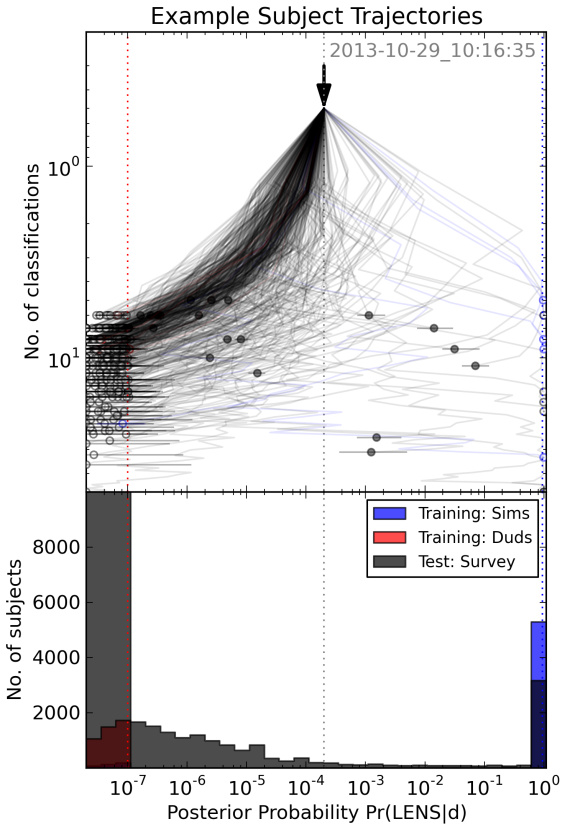

Figure 4. Typical SPACE WARPS stage 1 subject trajectories. Subjects drift downwards in the top panel as they are classified, while being nudged left and right by the Agents as they interpret the volunteers input. The dotted vertical lines show (left to right) the retirement threshold, prior probability, and the detection threshold.

site, more subjects were activated. In this way, the volunteers who down-voted images for not containing any lensed features enabled new images to be shown to other members of the community.

When all the subjects had either been retired, or classified around 10 times or more, the web app was paused and reconfigured for stage 2. The sample of subjects classified during stage 2 was selected to be all those that passed the detection threshold ($\Pr(\mathrm{LENS}|C,T)>0.95$) at stage 1. These were classified for one week, with no retirement but a maximum classification number of 50 each. The number of subjects at stage 2 was small enough that we did not need to retire any: instead, we simply collected classifications for a fixed period of time (about 4 weeks). Without the time pressure motivating the online analysis, we implemented an "offline" version of the analysis that performs a joint inference of all Agent confusion matrix elements and subject probabilities simultaneously, for comparison. We describe how this differs from the online analysis in the appendix.

5 RESULTS

In this section we present our findings about the performance of the Space Warps system, in terms of the information contributed by the crowd in Section 5.1, and the overall classifications of the training set that they made.

5.1 Crowd Properties

We define the following properties of the crowd, as characterised by their agents, and plot the distributions of their logarithms in Figure 5.

"Effort:" The number of test images, N_C, classified by a volunteer. In stage 1, the mean effort per agent was 263; in the shorter stage 2 it was 81.
"Experience:" The number of training images, N_T, classified by a volunteer. In stage 1, the mean experience per agent was 29; in stage 2 (where the training image fequency was set higher) it was 34.
"Skill:" The expectation value of the information gain, $\langle \Delta I \rangle_{0.5}$ should the next subject classified have lens probability 0.5 (Appendix A), in bits. Random classifiers have $\langle \Delta I \rangle_{0.5} = 0.0$, perfect classifiers have $\langle \Delta I \rangle_{0.5} = 1.0$. All agents start with $\langle \Delta I \rangle_{0.5} = 0.0$; this increases as training subjects are classified, and the agent's estimates of its confusion matrix elements improve. In stage 1, the mean skill per agent was 0.04 bits; in stage 2 it was 0.05.
"Contribution:" The integrated skill over a volunteer's test subject classification history, and represents the total contribution to the project that volunteer (see the appendix for more discussion of this quantity). In stage 1, the mean contribution per agent was 34.9 bits; in stage 2 it was 33.5.
"Information:" The total information ΔI generated by the agent during the volunteer's classification activity. This quantity depends on the value of each subject's lens probability when that subject was presented to the volunteer (Appendix A), and so there is an element of luck involved with this quantity: if you never see a high probability subject, it's hard to generate a large amount of information. You make your own luck by classifying more subjects.

The leftmost column of Figure 5 shows how the last four of these properties depends on the effort expended by the volunteers. We see that experience is strongly correlated with effort (as training images are presented throughout each stage, albeit at decreasing frequency), and that this is also true for skill. In the second row of Figure 5 we see that while skills of greater than 0.1 can be attained after just a few training images, most agents of such low experience have significantly lower skill. The volunteers in question only classify a few subjects before leaving. However, at high values of experience and effort, the skill is *always high*. There seem to be very few agents logging large numbers of classifications at low skill (although there are one or two exceptions): almost all high effort "super-users" have high skill. These two properties are reflected in both the contributions these volunteers make (third row) and the information they generate (fourth row).

The distributions for the stage 2 agents (orange) are qualitatively similar to those for the stage 1 agents (blue). Differences are: 1) the maximum effort possible at stage 2 is smaller, because fewer subjects were available to be classified, but 2) the mean effort expended at stage 2 was higher (perhaps because the subjects were higher probability, and more compelling); 3) the information generated per agent was higher at stage 2, because the subjects had higher probability.

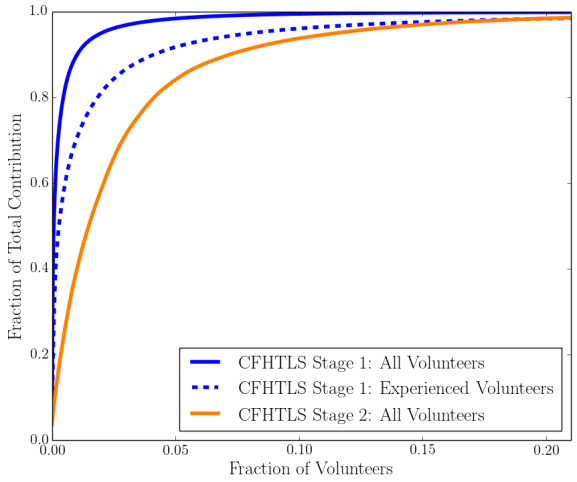

Figure 6. Narrow cumulative distributions of the contributions made by the agents: for example, 90% of the stage 2 contributions were made by the highest contributing 7% of the crowd. The stage 1 agents are shown in blue, the stage 2 agents in orange. "Experienced volunteers" classified 10 or more training subjects.

Figure 5 shows the Space Warps crowd to have quite broad distributions of logarithmic effort, skill, and contribution. To better quantify the contributions made by the volunteers, we show their cumulative distribution on a linear scale in Figure 6. This plot shows clearly the importance of the hardest-working, most active volunteers: at stage 1, 1.0% of the volunteers – 375 people – made 90% of the contribution. At stage 2, where it was not possible to make as many classifications before running out of subjects, 7.2% of the volunteers – 141 people – made 90% of the contribution.

However, it is not the case that only these small groups were capable of making this large contribution. The cumulative distribution of agent skill is shown in Figure 7: these distributions are significantly broader than the corresponding distributions of agent contribution in Figure 6. The most skilled 20% of agents possess only 79% of the skill at stage 1, and 77% at stage 2. The inexperienced volunteers also possess a significant fraction of the skill: the most skillful 20% of experienced volunteers (1824 people) possess just 43% of the total skill. The level of contribution made at Space Warps by experienced volunteers is largely a matter of choice (or perhaps, availability of time!).

5.2 Sample Properties

We now quantify the performance of the Space Warps system in terms of the recovery of the training set images. At stage 1, this set contains around 5712 simulated lenses, and 450 duds; at stage 2, we used 152 simulated lenses and 201 duds. Figure 8 shows receiver operating characteristic (ROC) curves for CFHTLS stage 1 and stage 2. These plots show the true positive rate (TPR), the number of sims correctly detected divided by the total number of sims in the

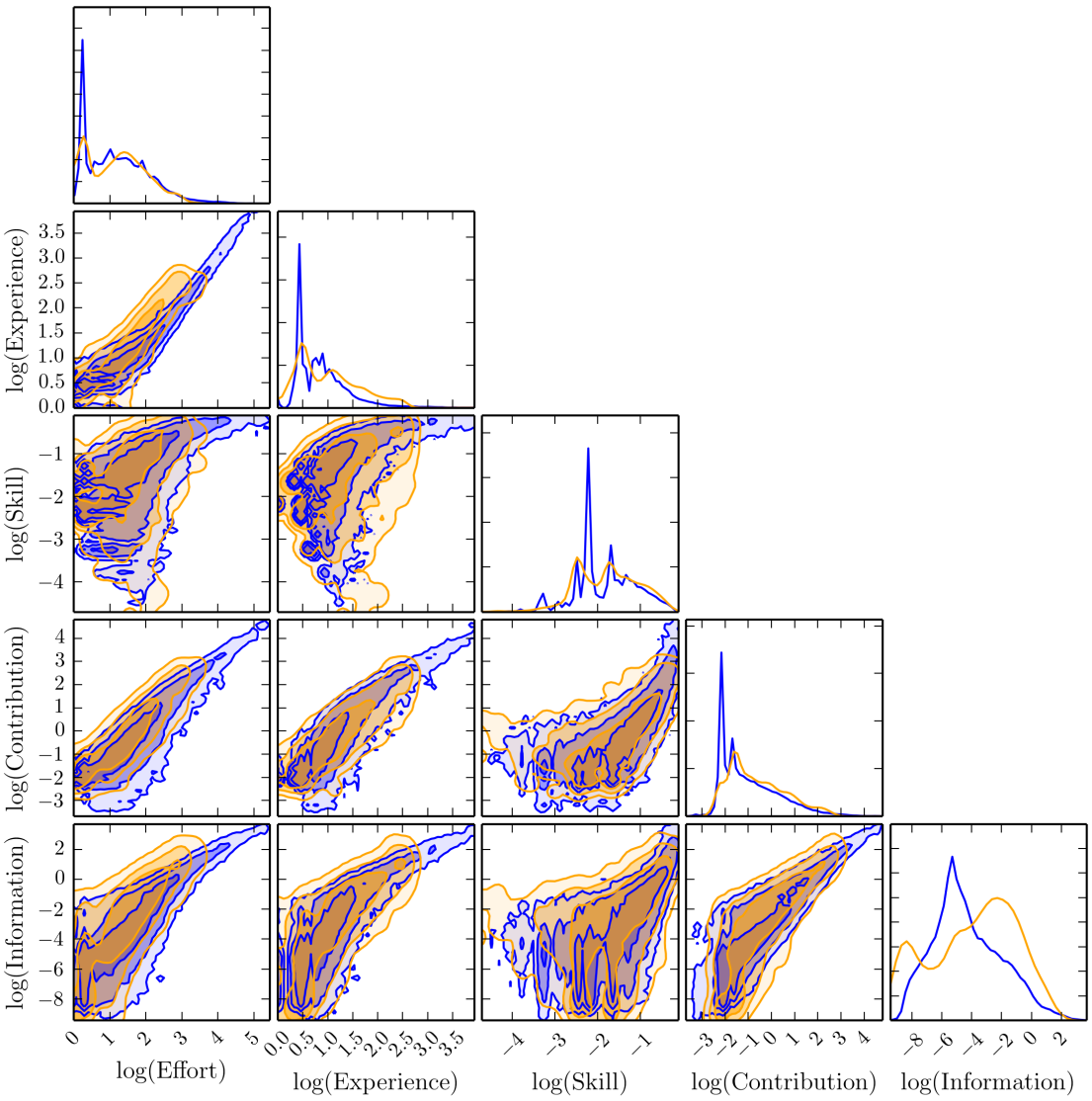

Figure 5. Key properties and contributions of the Space Warps crowd. Plotted are the 1-D and 2-D marginalized distributions for the logarithms of the properties of the agents described in the text. The stage 1 agents are shown in blue, the stage 2 agents in orange.

training set, and the false positive rate (FPR), the number of duds incorrectly detected divided by total number of duds in the training set, both for a given sample of detections defined by a particular probability threshold, which varies along the curves. In both stages, these curves show that true positive rates of around 90% were achieved, at very low false positive rates. For comparison we show the results of an analysis where the classifications of training images were ignored, and none of the agents allowed to learn: they were instead assigned initial values of their confusion matrices of 0.75 for each element, which then remained constant. This emulates a very simple unweighted voting scheme, where all classifications are treated equally. In this case, the TPR remains under 80% in stage 1 and 60% in stage 2, thus quantifying the benefit of including training images and allowing the agents to learn. The choice of initial confusion matrix is not very important: the same 0.75 initial values applied to normal, learning agents results in a slightly lower TPR than the default case, indicating that initializing the system less conservatively results in slightly worse performance.

The dot-dashed curves show the impact of the offline analysis. At stage 1 the results are very similar to the online version that was actually run (solid line). However, at stage 2 there is marginal evidence of there being greater benefit to doing the analysis offline. Over 85% TPR is achieved at zero FPR in the offline analysis , while if one is willing to accept a false positive rate of 5%, the true positive rate rises to over 95%, showing that some of the sims that were missed in the online analysis may be being recovered by doing the analysis offline. The same is true at stage 1, but to a lower degree. Assuming Poisson statistics for the fluctuations in the numbers of recovered lenses, the uncertainty

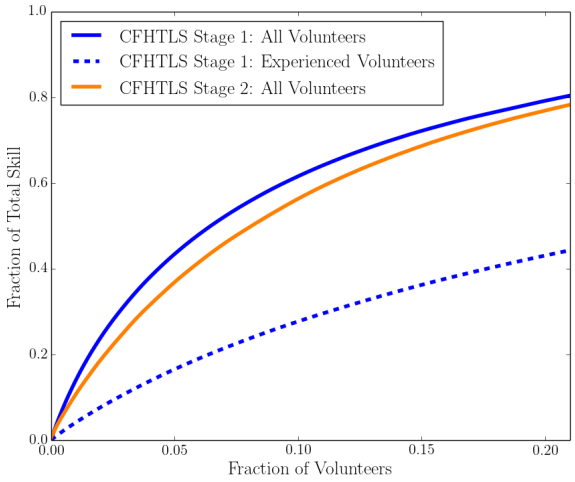

Figure 7. Broad cumulative distributions of agent skill: the most skilled 20% of the crowd only possess 79% of the skill at stage 1. The stage 1 agents are shown in blue, the stage 2 agents in orange. "Experienced volunteers" classified 10 or more training subjects.

in the measured stage 2 TPR values is around 8%, but the online and offline samples are highly correlated, such that the uncertainty on the difference between the ROC curves is somewhat less than this. Still, a larger validation set is needed to test these algorithmic choices more rigorously. At stage 1, high TPR can be measured to better than 1%.

Adopting the online stage 1 analysis, and the offline stage 2 analysis, we show in Figure 9 a plot of the more familiar (to astronomers) quantities completeness versus purity in the two stages. As in Figure 8, the detection threshold varies along the curves. Completeness is defined as the number of correctly detected sims divided by the total number of sims in the training set, while purity is the number of correct detections divided by the total number of detections.[4] If the training set samples from the same systems as the test set, then the completeness of the training set will be equivalent to the completeness of the test set. The purity depends on the proportion of sims to duds, and so the purity of the test set must be approximated by rescaling the training set to the expected proportion of lens systems to not-lens systems in the survey. First we compute the expected number of false positives by multiplying the FPR by the expected number of non-lenses in the survey. Then we multiply the TPR by the expected number of lenses in the survey, to get the expected number of true positives. The sum of the true positives and the false positives gives the expected sample size; dividing the expected number of true positives by this sample size gives the purity. Note that the completeness is invariant to this transformation.

The stage 1 curves are truncated by the retirement of subjects in this phase, which sets the minimum size of this sample. We see from the solid blue curve that over 90% completeness was reached, albeit in a sample with around 20% purity. To investigate the completeness to the three different types of lens in the training set, we repeat the same procedure, only now we consider only the detections of a certain kind of lens and of the non-lenses in the training set. We estimate the expected number of lenses and non-lens false positives by dividing the lens and dud sets into equal fractions.

To Do (Chris): That still could be worded better... # 66

The lensed quasar part of the training set yielded the highest completeness, suggesting that these were the easiest sims to spot. The lensed galaxies were recovered at understandably lowest completeness, although the wider image separation lensing clusters were not very different. At stage 2, where no retirement was carried out, it was possible to reach 100% purity: indeed, the knee of the curve is at just under 90% completeness. However, the purity decreases rapidly if higher completeness than this is sought. Interestingly, the lensing clusters can be seen to pose problems for the system, with its curve turning over at just 60% completeness.

The optimal sample in this simulated lens search experiment would have been constructed with a threshold value of $\Pr(\mathrm{LENS}|C,T) > 0.47$. At 100% purity and 89% completeness, it would have contained around 89 lens candidates.

6 DISCUSSION

What can we learn from the results of the previous section, for future projects? Potential improvements to the Space Warps system can be divided into three categories, performance, efficiency and capacity.

6.1 Improving performance: reducing incompleteness and impurity

We investigated the source of the incompleteness and impurity visible in Figure 9, by examining the stage 2 false negatives and false positives, and their behaviour as they are classified using the online analysis trajectory plots introduced in Section 4 (Figure 4). Figure 10 shows 2 example simulated lenses that were missed ($\Pr(\mathrm{LENS}|C,T) < 10^{-5}$) by the Space Warps system, and 2 example non-lenses that were incorrectly flagged as candidates ($\Pr(\mathrm{LENS}|C,T) > 0.95$) by the Space Warps system. While the final stage 2 analysis was done offline, these trajectory plots still serve to illustrate the approximate classification histories of the subjects.

In some cases the rejection of the false negatives is understandable: the lensed features are faint, or in some cases, somewhat unrealistic as lenses. However, in other cases a reasonably obvious lens was passed over. This mainly seems to be due to noise in the system: when only low-skill classifiers view a subject, all the updates to its posterior probability are small, and if none are very confident about the presence of lensing, the subject follows a random walk down its trajectory plot. This can be seen in the top left panel of the figure.

The false positives show similar behaviour, for example in the bottom lefthand panel of the figure. As well as subjects being "unlucky" in this way, there are two less common failure modes associated with mistakes made by higher skill users, illustrated in the righthand column of the figure. In the false negative subject shown in the top right-hand panel,

[4] The completeness is equivalent to the TPR and is also known elsewhere as the "recall." The purity is also known as the "precision."

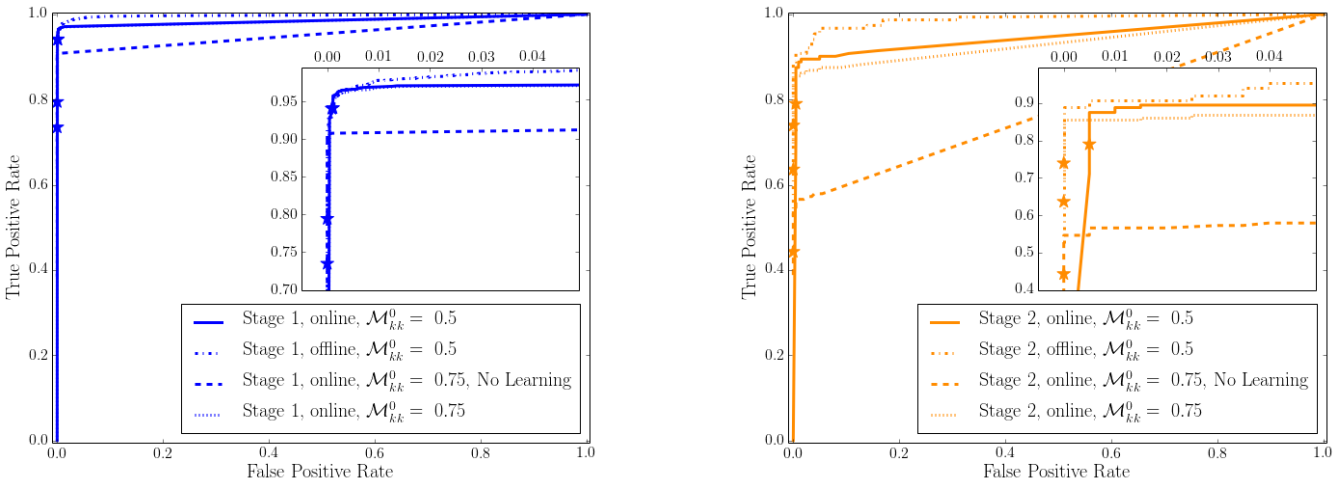

Figure 8. Receiver operating characteristic curves for the SPACE WARPS system, using the CFHTLS training set. Left: stage 1, right: stage 2.

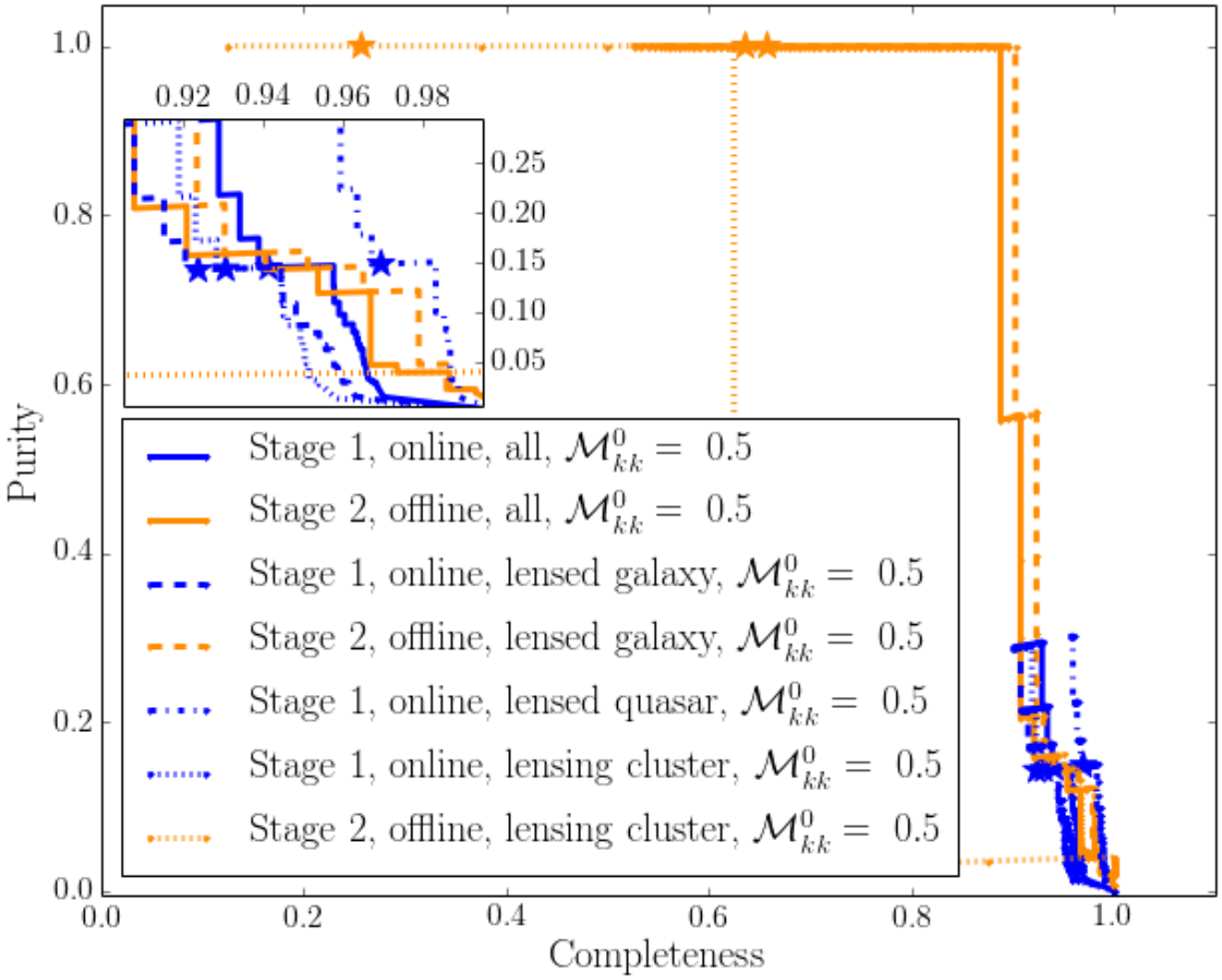

Figure 9. Completeness-estimated purity curves for the SPACE WARPS system, using the CFHTLS training set.

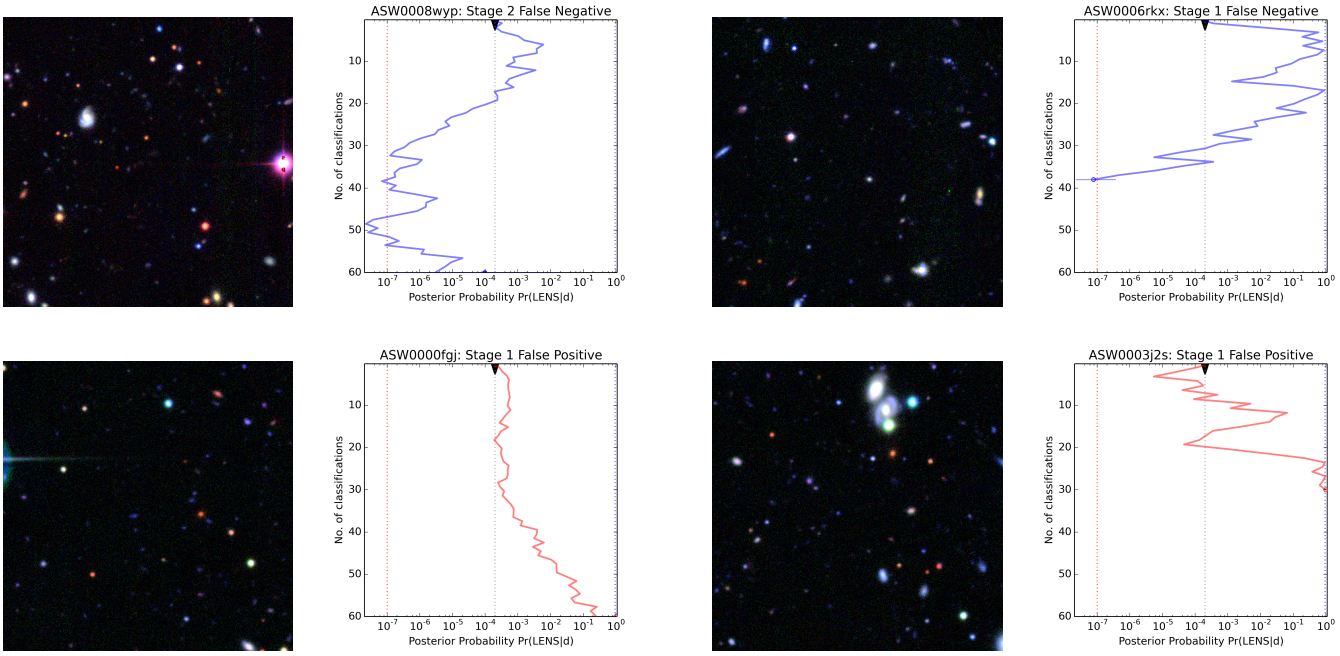

Figure 10. Illustrative examples of false negatives (blue) and false positives (red) from the classification of the Space Warps CFHTLS training set. The trajectory plot for each of the 4 subjects is shown to the right of its image. lefthand panels show the most common types of trajectories: short step random walks resulting from no high skill classifiers viewing the subject. The righthand panels show examples where higher skill classifiers have been involved, causing random walks with longer step lengths (top) and catastrophic mis-classifications (bottom right).

several high skill classifiers update the subject upwards in log probability by some way each time, but other, comparable skill classifiers mis-classify the system to lower probability. If we had retired subjects after crossing a high probability detection threshold, this subject would have been correctly identified as a lens candidate; instead, it was kept in the system and eventually scattered down to retirement. The trajectory looks like a random walk, but with bigger steps. The bottom righthand panel shows an example of a final, apparently rare, failure mode: we see some short-step random walk behaviour, followed by a mis-classification by a very high skill "expert" classifier after 20 classifications that promotes the subject to high probability.

There are a number of places where we can address these problems and improve system performance: adding flexibility to the classification interface, educating the volunteers, assigning subjects for classification, and interpreting the classification data.

We might wonder whether some of the mistakes made by reasonably high skill classifiers could have been corrected by those classifiers themselves, had they had access to a "go back" option. While clearly enabling error reduction, we might worry about such a mechanism enabling classifiers to have a guess, then go and read the Talk discussion, and then return to the classification interface to change their mind to agree with the forum. In fact, this is already possible on the current interface – and indeed, it is not obvious that such correlations would necessarily lead to poorer system performance. This could be tested by presenting a fraction of the volunteers with a version of the site that actively suggested that they take this approach, and then tracking the relative performance of "collaborative" classifications compared with independent ones. We leave this to further work.

Mistakes by both high and low skill classifiers could be reduced by improving the training in the system, which could be done in several ways. One is to make more training images available to those who want or need it. A basic level of training images are needed for SWAP to build up an accurate picture of each classifier's skill – but one could imagine volunteers *choosing* to see more training images (still at random) in their stream. We could also experiment with providing greater training rates early on for all volunteers; the risk associated with this is that retention rates may drop if too few "fresh" test images are shown early on. Another way of improving the training could be to provide more information about what lenses do. In this project, the Spotter's Guide was always available on the site, but as a passive background resource. We might consider providing more links to this guide in the feedback messages shown to the classifiers as they go – or perhaps extending these messages to themselves include more explanations and example images. We might also investigate a more dynamic Spotter's Guide: a set of manipulable model lenses illustrating all the possible image configurations that that those deflectors can make could help volunteers gain understanding of what lensed features can look like very quickly. Such a toy is under development.

However, most identification failures (around 60%) seem to be due to the random walk problem. Short, random direction steps arise from low skill classifiers, arising from classifier inexperience (Section 5.1). While improved training will help reduce this effect, it is here that targeted task assignment could also help improve system performance, by bringing higher skill volunteers in to the mix. We advertised

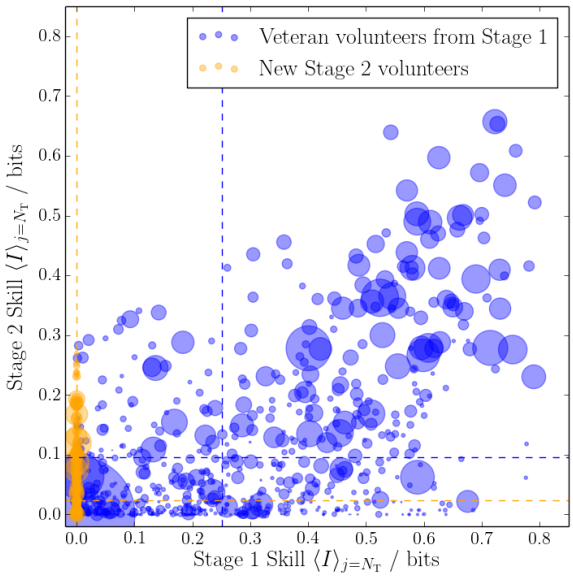

Figure 11. Stage 2 classifier skill, as a function of their stage 1 skill. Veterans from stage 1 are shown in blue, while new volunteers are shown in orange. Point size is proportional to the number of stage 2 classifications made, while dashed lines are drawn at the mean values for each sample. Restricting stage 2 to classifiers with high stage 1 skill could reduce the noise in the system.

stage 2 of the current project to all registered users; it was taken up by stage 1 veterans with a broad range of skill, and also picked up a significant number of new users who did not have enough time to aquire high skill (since stage 2 was quite short): of the 1964 volunteers who took part in the stage 2 classification round, only 774 were veterans from stage 1. Figure 11 shows how the stage 1 skill maps on to stage 2 skill and contribution.

This figure suggests that there may be some benefit to opening stage 2 on the basis of stage 1 agent skill, in order to reduce the noise in the system generated by new and low skill volunteers in this more difficult classification stage. A simple implementation of this would be to switch the web app to the stage 2 version when either a certain stage 1 skill (and/or effort) threshold was passed. Good strategy would probably be to switch back to stage 1 after some number of stage 2 classifications (to avoid losing the high skill users from stage 1 entirely), and also to make it very clear which stage the classifier was at (since stage 2 requires more careful attention). This is something we plan to experiment with in future.

Finally, there could be some performance gains to be made by improving the SWAP agent model, or its implementation. Low skill is typically a result of inexperience (Section 5.1), but it could be the agent that is inexperienced as well as the classifier or its agent. In a future paper we plan to investigate the use of the test images as well as the training images in accelerating the agent's learning (Davis et al, in preparation). We also plan to investigate the use of offline analysis at Stage 1, for the same reason (see also Section 6.2 below.) One might also consider looking at introducing more conditional dependences in the confusion matrix elements, to allow for some classifiers having greater skill in spotting one type of lens than another, or, more generally, as a function of lens property (such as color, brightness, and image separation). In the current model, all agents are considered to be completely independent, whereas in fact we might expect there to be significant clustering of the confusion matrix elements. A hierarchical model for the crowd, with hyper-parameters describing the distribution of confusion matrix elements across the population, may well accelerate the agent learning process by including the notion of one agent being likely to be similar to its neighbours. Finally, it is worth noting that the model of **?** explicitly avoids the assumption of the agent confusion matrix elements being constant in time, allowing the development of volunteer skill to be more accurately tracked – and that they did see some time-evolution in the supernova zoo classifiers' skill. Finding a way to incorporate such a learning model into SWAP, while retaining its statistically online character is an interesting challenge for future work.

6.2 Improving efficiency

Table 1 shows the total effort, contribution, skill and information generated in both stage 1 and 2 of the CFHTLS project, with the total numbers of agents and subjects for comparison. These numbers allow us to quantify the efficiency of the system.

The contribution per classification is defined in terms of a hypothetical subject with lens probability of 0.5; one bit of information is needed to update such a subject's lens probability to either zero or one. This means that a maximally complete classification stage would yield a total contribution (summed over all agents) equal to the number of subjects. The ratio of this hypothetical optimum to the actual total contribution is therefore a measure of the stage's inefficiency. We find our inefficiency (by dividing column 2 by column 3) to be 33% and 17% in stage 1 and 2 respectively. In stage 1, this inefficiency is due to the daily processing: we were not able to retire subjects fast enough, and so they, remained in the system, being over-classified. Indeed, only 3705745 classifications were needed to retire all the subjects: the ratio of this to the total number made is 34%. (The remaining 1% is due to not all subjects being classified to 1 or 0 probability.) At stage 2, we did not retire any subjects at all; the inefficiency in this case was by design, to give everyone a chance to appreciate what they had found together. (An unwanted side effect of this policy was noted in Figure 11 in the previous sub-section.)

It is clear that to increase the efficiency of the system we need to reduce the time lag between the classification being made and its outcome being analyzed. The optimal way to do this would be to have the web app itself analyze the classifications in a fully online system. This is under investigation for future projects. With the classification data being analyzed in real time, there may still be a place for a daily or weekly offline analysis: this could potentially reduce the false negative and false positive rates by "resurrecting" subjects that had been retired by the online system before the agents had time to learn enough about their classifier's high skill.

Table 1. Total crowd and subject sample properties from the CFHTLS project.

Stage	Subjects J	Contribution $\sum_k^K \langle \Delta I \rangle_{0.5}{}_k^{\rm total}$ (bits)	Agents K	Skill $\sum_k^K \langle \Delta I \rangle_{0.5}{}_k$ (bits)	Classifications $\sum_k^K N_{{\rm C},k}$	Candidates $N_{\rm det}$	Information $\sum_j^J \sum_k^K \Delta I_{j,k}$ (bits)
1	427064	1292016.3	36982	1471.9	10802125	3368	91122.6
2	3679	21895.8	1964	102.4	224745	90	1640.4

6.3 Increasing capacity

Finally, we comment on the size of the first Space Warps crowd, and how the system might be scaled up for future surveys. In Section 5.1 and Table 1, a rough picture emerged for the Space Warps crowd of it containing a few 10^4 volunteers, with a few 10^3 achieving considerable skill, and a few 10^2 having the time to make a significant contribution. Slightly more quantitively, we might note that the total skill of the crowd, computed by summing the skill of all the agents, is a measure of the effective crowd size, in the sense that a crowd of perfect classifiers would be of this size. By this measure,the stage 1 crowd was equivalent to a team of 1470 perfect classifiers, while the stage 2 crowd was equivalent to a team of 102 perfect classifiers. With this same crowd, we can see that surveys providing a few 10^4 subjects would be completed quickly, if the high contribution rate of the current crowd were to be repeated.

There are (at least) two ways in which we might increase the numbers of high-contribution volunteers for larger projects in future. The first is simply to increase the total crowd size, and hope that a similar fraction of volunteers make large contributions. Greater exposure of the website to the public through mass media woudl help. Another option is to advertise the project to new groups of volunteers by translating the website into other languages (something which is now supported by the Zooniverse). A multi-lingual user base would come with its own set of challenges, especially in terms of volunteers' continuing training and interactions on Talk.

The second way to scale up the number of high contribution classifiers is to increase the rate at which new volunteers become dedicated volunteers. Based on feedback from the wider Space Warps community, this could potentially be achieved through closer collaboration with the science team. It's also possible that the dynamic stage 2 system proposed in the previous section may act as an incentive to some volunteers to increase their contribution (while gaining skill in order to "reach" stage 2). Reducing the rate at which new volunteers lose interest could also play a role. Anecdotally, it seems fairly common for new volunteers to be wary of classifying at all, for fear of introducing errors. Better explanation of how their early classifications are analyzed could help assuage these fears: Section 5.1 shows that effectively down-weighting new users' classifications (by setting their agents' initial confusion matrix elements to those of a random classifier) leads to best performance, a result which shoudl be of some comfort to the nervous volunteer.

7 CONCLUSIONS

To Do (Phil): write conclusions (#45).

Summary of system.

Crowd-sourced gravitational lens detection works, in terms of the classification of the training set as described here, in the following specific ways:

- Participation (crowd size, activity rate) enabled project completion
- Both stages (1 and 2) achieved the required rejection rates
- Integrated humanpower = X (stage 1) and y (stage 2), cf hours taken by small team of experts
- Nightly processing is inefficient: more classifications were made than was necessary during peak participation. Need kafka...
- Retirement rate. False negatives: which sims were missed?
- The optimal true positive rate (completeness) and false positive rate in the training set were estimated to be TPR% and FPR% at Stage 1, assuming a detection threshold of xxx.
- In the "refinement" stage 2, X% of the stage 1 candidates were rejected (with P less than threshold), and the remainder assigned lens "probabilities." Ranking subjects by their Stage 2 lens probability gives an ROC curve for the system with X properties. The optimal true positive rate (completeness) and false positive rate in the training set were estimated as TPR% and FPR%;
- The lens-finding crowd shows some interesting properties, with consequences for future scalability
- The information comes predominantly from volunteers with agents with P = ...
- The agents show a high mean information per classification, which increased/decreased with time; this does/doesn?t correlate with active crowd size, showing how the crowd changed over time...

Sum up, end.

ACKNOWLEDGEMENTS

We thank all XXXmembers of the Space Warps community for their contributions to the project so far. A complete list of collaborators is given at... In particular we would like to recognise the efforts of XXX, YYY etc in moderating the discussion.

We are also grateful to Brooke Simmons, David Hogg, XXX and YYY for many useful conversations about citizen science and gravitational lens detection. PJM was given

support by the Royal Society, in the form of a research fellowship, and by the U.S. Department of Energy under contract number DE-AC02-76SF00515. AV acknowledges support from the Leverhulme Trust in the form of a research fellowship. The work of AM and SM was supported by World Premier International Research Center Initiative (WPI Initiative), MEXT, Japan. The work of AM was also supported in part by National Science Foundation Grant No. PHYS-1066293 and the hospitality of the Aspen Center for Physics. PJM and ES thank the Institute of Astronomy and Astrophysics, Academia Sinica (ASIAA) and Taiwan's Ministry of Science and Technology (MOST) for their financial support of the workshop "Citizen Science in Astronomy" in March 2014, at which some parts of the SWAP analysis was developed.

The Space Warps project is open source. The web app was developed at https://github.com/Zooniverse/Lens-Zoo while the SWAP analysis software was developed at https://github.com/drphilmarshall/SpaceWarps.

APPENDIX A: PROBABILISTIC CLASSIFICATION ANALYSIS

Our aim is to enable the construction of a sample of good lens candidates. Since we aspire to making logical decisions, we define a "good candidate" as one which has a high posterior probability of being a lens, given the data: Pr(LENS|$\mathbf{d}$). Our problem is to approximate this probability. The data $\mathbf{d}$ in our case are the pixel values of a colour image. However, we can greatly compress these complex, noisy sets of data by asking each volunteer what they think about them. A complete classification in Space Warps consists of a set of Marker positions, or none at all. The null set encodes the statement from the volunteer that the image in question is "NOT" a lens, while the placement of any Markers indicates that the volunteer considers this image to contain a "LENS". We simplify the problem by only using the Marker positions to assess whether the volunteer correctly assigned the classification "LENS" or "NOT" after viewing (blindly) a member of the training set of subjects.

How should we model these compressed data? The circumstances of each classification are quite complex, as are the human classifiers in general: the volunteers learn more about the problem as they go, but also inevitably make occasional mistakes (perhaps because a lens is difficult to see, or they became distracted during the task). To cope with this uncertainty, we assign a simple software *agent* to partner each volunteer. The agent's task is to interpret their volunteer's classification data as best it can, using a model that makes a number of necessary approximations. These interpretations will then include uncertainty arising as a result of the volunteer's efforts and also the agent's approximations, but they will have two important redeeming features. First, the interpretations will be quantitative (where before they were qualititative), and thus will be useful in decision-making. Second, the agent will be able to predict, using its model, the probability of a test subject being a LENS, given both its volunteer's classification, and its volunteer's experience. In this appendix we describe how these agents work, and other aspects of the Space Warps analysis pipeline (SWAP).

A1 Agents and their Confusion Matrices

Each agent assumes that the probability of a volunteer recognising any given simulated lens as a lens is some number, Pr("LENS"|LENS, $\mathbf{d}^{\rm t}$), that depends only on what the volunteer is currently looking at, and all the previous training subjects they have seen (and not on what type of lens it is, how faint it is, what time it is, *etc.*). Likewise, it also assumes that the probability of a volunteer recognising any given dud image as a dud is some other number, Pr("NOT"|NOT, $\mathbf{d}^{\rm t}$), that also depends only on what the volunteer is currently looking at, and all the previous training subjects they have seen. These two probabilities define a 2 by 2 "confusion matrix," which the agent updates, every time a volunteer classifies a training subject, using the following very simple estimate:

$$\Pr(\text{“}X\text{”}|X, \mathbf{d}^{\rm t}) \approx \frac{N_{\text{“}X\text{”}}}{N_X}. \quad \text{(A1)}$$

Here, X stands for the true classification of the subject, *i.e.* either LENS or NOT, while "X'' is the corresponding classification made by the volunteer on viewing the subject. N_X is the number of lenses the volunteer has been shown, while $N_{\text{“}X\text{”}}$ is the number of times the volunteer got their classifications of this type of training subject right. $\mathbf{d}^{\rm t}$ stands for all $N_{\rm LENS} + N_{\rm NOT}$ training data that the agent has heard about to date.

The full confusion matrix of the $k^{\rm th}$ volunteer's agent is therefore:

$$\mathcal{M}^k = \begin{bmatrix} \Pr(\text{“LENS”}|\text{NOT}, \mathbf{d}^{\rm t}_k) & \Pr(\text{“LENS”}|\text{LENS}, \mathbf{d}^{\rm t}_k) \\ \Pr(\text{“NOT”}|\text{NOT}, \mathbf{d}^{\rm t}_k) & \Pr(\text{“NOT”}|\text{LENS}, \mathbf{d}^{\rm t}_k) \end{bmatrix},$$
$$= \begin{bmatrix} \mathcal{M}_{LN} & \mathcal{M}_{LL} \\ \mathcal{M}_{NN} & \mathcal{M}_{NL} \end{bmatrix}^k. \quad \text{(A2)}$$

Note that these probabilities are normalized, such that Pr("NOT"|NOT) = 1 − Pr("LENS"|NOT).

Now, when this volunteer views a test subject, it is this confusion matrix that will allow their agent to update the probability of that test subject being a LENS. Let us suppose that this subject has never been seen before: the agent assigns a prior probability that it is (or contains) a lens is

$$\Pr(\text{LENS}) = p_0 \quad \text{(A3)}$$

where we have to assign a value for p_0. In the CFHTLS, we might expect something like 100 lenses in 430,000 images, so $p_0 = 2 \times 10^{-4}$ is a reasonable estimate. The volunteer then makes a classification C_k (= "LENS" or "NOT"). We can apply Bayes' Theorem to derive how the agent should update this prior probability into a posterior one using this new information:

$$\Pr(\text{LENS}|C_k, \mathbf{d}^{\rm t}_k) = \quad \text{(A4)}$$
$$\frac{\Pr(C_k|\text{LENS}, \mathbf{d}^{\rm t}_k) \cdot \Pr(\text{LENS})}{[\Pr(C_k|\text{LENS}, \mathbf{d}^{\rm t}_k) \cdot \Pr(\text{LENS}) + \Pr(C_k|\text{NOT}, \mathbf{d}^{\rm t}_k) \cdot \Pr(\text{NOT})]},$$

which can be evaluated numerically using the elements of the confusion matrix.

A2 Examples

Suppose we have a volunteer who is always right about the true nature of a training subject. Their agent's confusion

matrix would be

$$\mathcal{M}^{\rm perfect} = \begin{bmatrix} 0.0 & 1.0 \\ 1.0 & 0.0 \end{bmatrix}. \quad \text{(A5)}$$

On being given a fresh subject that actually is a LENS, this hypothetical volunteer would submit $C =$ "LENS". Their agent would then calculate the posterior probability for the subject being a $LENS$ to be

$$\Pr(\text{LENS}|\text{"LENS"}, \mathbf{d}^{\rm t}_k) = \frac{1.0 \cdot p_0}{[1.0 \cdot p_0 + 0.0 \cdot (1-p_0)]} = 1.0, \quad \text{(A6)}$$

as we might expect for such a *perfect* classifier. Meanwhile, a hypothetical volunteer who (for some reason) wilfully always submits the wrong classification would have an agent with the column-swapped confusion matrix

$$\mathcal{M}^{\rm obtuse} = \begin{bmatrix} 1.0 & 0.0 \\ 0.0 & 1.0 \end{bmatrix}, \quad \text{(A7)}$$

and would submit $C =$ "NOT" for this subject. However, such a volunteer would nevertheless be submitting useful information, since given the above confusion matrix, their agent would calculate

$$\Pr(\text{LENS}|\text{"NOT"}, T_k) = \frac{1.0 \cdot p_0}{[1.0 \cdot p_0 + 0.0 \cdot (1-p_0)]} = 1.0. \quad \text{(A8)}$$

Obtuse classifiers turn out to be as helpful as *perfect* ones.

A3 Online SWAP: Updating the Subject Probabilities

Suppose the $k{+}1^{\rm th}$ volunteer now submits a classification, on the same subject just classified by the $k^{\rm th}$ volunteer. We can generalise Equation A4 by replacing the prior probability with the current posterior probability:

$$\Pr(\text{LENS}|C_{k+1}, \mathbf{d}^{\rm t}{}_{k+1}, \mathbf{d}) = \quad \text{(A9)}$$

$$\frac{1}{Z}\Pr(C_{k+1}|\text{LENS}, \mathbf{d}^{\rm t}{}_{k+1}) \cdot \Pr(\text{LENS}|\mathbf{d}) \quad \text{(A10)}$$

$$\text{where } Z = \Pr(C_{k+1}|\text{LENS}, \mathbf{d}^{\rm t}{}_{k+1}) \cdot \Pr(\text{LENS}|\mathbf{d}) + \Pr(C_{k+1}|\text{NOT}, \mathbf{d}^{\rm t}{}_{k+1}) \cdot \Pr(\text{NOT}|\mathbf{d}),$$

and $\mathbf{d} = \{C_k, \mathbf{d}^{\rm t}_k\}$ is the set of all previous classifications, and the set of training subjects seen by each of those volunteers. $\Pr(\text{LENS}|\mathbf{d})$ is the fundamental property of each test subject that we are trying to infer. We track $\Pr(\text{LENS}|\mathbf{d})$ as a function of time, and by comparing it to a lower or upper thresholds, make decisions about whether to retire the subject from the classification interface or promote it in Talk, respectively.

A4 Information Gain per Classification, Agent "Skill" and "Contribution"

With an agent's confusion matrix in hand we can compute the *information* generated in any given classification. This will depend on the confusion matrix elements (Equation A2) but also on the probability of the subject being classified containing a lens. The quantity of interest is the relative entropy, or Kullback-Leiber divergence, between the prior and posterior probabilities for the possible truths T given the submitted classification C:

$$\begin{aligned}\Delta I =& \sum_T \Pr(T|C) \log_2 \frac{\Pr(T|C)}{\Pr(T)} \\ =& \Pr(\text{LENS}|C) \log_2 \frac{\Pr(C|\text{LENS})}{\Pr(C)} \\ &+ \Pr(\text{NOT}|C) \log_2 \frac{\Pr(C|\text{NOT})}{\Pr(C)}, \end{aligned} \quad \text{(A11)}$$

where, as above, C can take the values "LENS" or "NOT". Substituting for the posterior probabilities using Equation A4 we get an expression that just depends on the elements of the confusion matrix $\mathcal{M}$ and the pre-classification subject lens probability $\Pr(\text{LENS}) = p$:

$$\begin{aligned}\Delta I =& p\frac{\mathcal{M}_{CL}}{p_c} \log_2 \frac{\mathcal{M}_{CL}}{p_c} \\ &+ (1-p)\frac{\mathcal{M}_{CN}}{p_c} \log_2 \frac{\mathcal{M}_{CN}}{p_c}, \end{aligned} \quad \text{(A12)}$$

where the common denominator $p_c = p\mathcal{M}_{CL} + (1-p)\mathcal{M}_{CN}$. This expression has many interesting features. If p is either zero or one, $\Delta I(C) = 0$ regardless of the value of C or the values of the confusion matrix elements: if we know the subject's status with certainty, additional classifications supply no new information. If we set p to be the prior probability, Equation A12 tells us how much information is generated by classifying it all the way to $p = 1$ (which a perfect classifier, with $\mathcal{M}_{LL} = \mathcal{M}_{NN} = 1$, can do in a single classification). For a prior probability of 2×10^{-4}, 12.3 bits are generated in such a "detection." Conversely, only 0.0003 bits are generated during the rejection of a subject with the same prior: we are already fairly sure that each subject does not contain a lens! Imperfect classifiers (with $\mathcal{M}_{LL}$ and $\mathcal{M}_{NN}$ both less than 1) generate less than these maximum amounts of information each classification; the only classifiers that generate zero information are those that have $\mathcal{M}_{LL} = 1 - \mathcal{M}_{NN}$ (or equivalently, $\mathcal{M}_{CL} = \mathcal{M}_{CN}$ for all values of C). We might label such classifiers as "random", since they are as likely to classify a subject as a "LENS" no matter the true content of that subject.

Equation A12 suggests a useful information theoretical definition of the classifier skill perceived by the agent. At a fixed value of p, we can take the expectation value of the information gain ΔI over the possible classifications that could be made:

$$\begin{aligned}\langle \Delta I \rangle =& \sum_C \sum_T \Pr(T|C)\Pr(C) \log_2 \frac{\Pr(T|C)}{\Pr(T)} \\ =& -\sum_T \Pr(T) \log_2 \Pr(T) \\ &+ \sum_C \Pr(C) \sum_T \Pr(T|C) \log_2 \Pr(T|C) \\ =& p\left[\mathcal{S}(\mathcal{M}_{LL}) + \mathcal{S}(1-\mathcal{M}_{LL})\right] \\ &+ (1-p)\left[\mathcal{S}(\mathcal{M}_{NN}) + \mathcal{S}(1-\mathcal{M}_{NN})\right] \\ &- \mathcal{S}\left[p\mathcal{M}_{LL} + (1-p)(1-\mathcal{M}_{NN})\right] \\ &- \mathcal{S}\left[p(1-\mathcal{M}_{LL}) + (1-p)\mathcal{M}_{NN}\right] \end{aligned} \quad \text{(A13)}$$

where $\mathcal{S}(x) = x \log_2 x$. If we choose to evaluate $\langle \Delta I \rangle$ at $p = 0.5$, the result has some useful properties. While random classifiers presented with $p = 0.5$ subjects have $\langle \Delta I \rangle_{0.5} = 0.0$ as expected, perfect classifiers appear to the agents to

have $\langle \Delta I \rangle_{0.5} = 1.0$. This suggests that $\langle \Delta I \rangle_{0.5}$, the amount of information we expect to gain when a classifier is presented with a 50-50 subject, is a reasonable quantification of *normalised skill*. A consequence of this choice is that the integrated skill (over all agents' histories) should come out to be approximately equal to the number of subjects in the survey, when the search is "complete" (and all subjects are fully classified). Therefore, a particular agent's integrated skill is a reasonable measure of that classifier's *contribution* to the lens search.

We conservatively initialize both elements of each agent's confusion matrix to be $\mathcal{M}^0_{LL} = \mathcal{M}^0_{NN} = 0.5$, that of a maximally ambivalent random classifier, so that all agents start with zero skill. While this makes no allowance for volunteers that actually do have previous experience of what gravitational lenses look like, we might expect it to help mitigate against false positives. Anyone who classifies more than one image (by progressing beyond the tutorial) makes a non-zero information contribution to the project.

The total information generated during the CFHTLS project is shown in Table 1. Interpreting these numbers is not easy, but we might do the following. Dividing this by the amount of information it takes to classify a Space Warps subject all the way to the detection threshold (lens probability 0.95), and then multiplying by the survey inefficiency gives us a very rough estimate for the effective number of detections corresponding to the crowd's contribution: these are 2830 and 25 bits for stages 1 and 2 respectively. These figures are close to the numbers of detections given in column 7 of the table.

A5 Uncertainty in the Agent Confusion Matrices

Finally, the confusion matrix obtained from the application of Equation A1 has some inherent noise which reduces as the number of training subjects classified by the agent's volunteer increases. For simplicity, the discussion has thus far assumed the case when the confusion matrix is known perfectly; in practice, we allow for uncertainty in the agent confusion matrices by averaging over a small number of samples drawn from Binomial distributions characterised by the matrix elements $\Pr(C_k|\mathrm{LENS}, \mathbf{d}^{\mathrm{t}}_k)$ and $\Pr(C_k|\mathrm{NOT}, \mathbf{d}^{\mathrm{t}}_k)$. The associated standard deviation in the estimated subject probability provides an error bar for this quantity.

A6 Offline SWAP

The probabilistic model described above does not need to be implemented as an online inference. Indeed, it might be more appropriate to perform the inference of all Agent confusion matrix elements and Subject probabilities simultaneously, so that the early classifications are not effectively downweighted as a result of the Agent's ignorance. It might also be that this ignorance builds in some conservatism to the system, reducing the noise due the early classifications if they are unreliable. In the joint analysis, the basic assumption that is built into the Agents, that their volunteers have innate and unchanging talent for lens spotting parameterised by two constant confusion matrix elements which simply need to be inferred given the data, is implemented in full. The effect is that of applying the time-averaged confusion matrices, rather than one that evolves as the Agents (and in the real world, the volunteers) learn.

The mathematics of the offline inference are presented elsewhere (in preparation). Here we briefly note that we maximize the joint posterior probability distribution for all the model parameters (some 66,000 confusion matrix elements and 430,000 subject probabilities) with a simple expectation-maximisation algorithm. This procedure takes approximately the same CPU time as the stage 2 online analysis, because no matrix inversions are required in the algorithm. The algorithm scales well and is actually faster than the online analysis with the larger stage 1 dataset. The expectation-maximisation algorithm is robust to initial starting parameters in, e.g., initial Agent confusion matrix elements and Subject probabilities. The plots in Section 5.2 show the difference in performance between the online and offline analyses.

REFERENCES

Auger, M. W., Treu, T., Bolton, A. S., Gavazzi, R., Koopmans, L. V. E., Marshall, P. J., Moustakas, L. A., & Burles, S. 2010a, ApJ, 724, 511

Auger, M. W., Treu, T., Gavazzi, R., Bolton, A. S., Koopmans, L. V. E., & Marshall, P. J. 2010b, ApJL, 721, L163

Bolton, A. S., Burles, S., Koopmans, L. V. E., Treu, T., & Moustakas, L. A. 2006, ApJ, 638, 703

Bolton, A. S., Burles, S., Schlegel, D. J., Eisenstein, D. J., & Brinkmann, J. 2004, AJ, 127, 1860

Browne, I. W. A., et al. 2003, MNRAS, 341, 13

Cabanac, R. A., et al. 2007, A&A, 461, 813

Collett, T. E., Auger, M. W., Belokurov, V., Marshall, P. J., & Hall, A. C. 2012, MNRAS, 424, 2864

Dalal, N., & Kochanek, C. S. 2002, ApJ, 572, 25

Faure, C., et al. 2008, ApJS, 176, 19

Gavazzi, R., Marshall, P. J., Treu, T., & Sonnenfeld, A. 2014, ApJ, 785, 144

Gavazzi, R., Treu, T., Koopmans, L. V. E., Bolton, A. S., Moustakas, L. A., Burles, S., & Marshall, P. J. 2008, ApJ, 677, 1046

Hezaveh, Y., Dalal, N., Holder, G., Kuhlen, M., Marrone, D., Murray, N., & Vieira, J. 2013, ApJ, 767, 9

Inada, N., et al. 2012, AJ, 143, 119

Jackson, N. 2008, MNRAS, 389, 1311

Lintott, C. J., et al. 2008, MNRAS, 389, 1179

—. 2009, MNRAS, 399, 129

Marshall, P. J., Hogg, D. W., Moustakas, L. A., Fassnacht, C. D., Bradač, M., Schrabback, T., & Blandford, R. D. 2009, ApJ, 694, 924

More, A., Cabanac, R., More, S., Alard, C., Limousin, M., Kneib, J.-P., Gavazzi, R., & Motta, V. 2012, ApJ, 749, 38

Moustakas, L. A., et al. 2007, ApJL, 660, L31

Negrello, M., et al. 2010, Science, 330, 800

—. 2014, MNRAS, 440, 1999

Newton, E. R., Marshall, P. J., Treu, T., Auger, M. W., Gavazzi, R., Bolton, A. S., Koopmans, L. V. E., & Moustakas, L. A. 2011, ApJ, 734, 104

Pawase, R. S., Courbin, F., Faure, C., Kokotanekova, R., & Meylan, G. 2014, MNRAS, 439, 3392

Poindexter, S., Morgan, N., & Kochanek, C. S. 2008, ApJ, 673, 34

Quimby, R. M., et al. 2014, ArXiv e-prints
Schwamb, M. E., et al. 2012, ApJ, 754, 129
Sonnenfeld, A., Treu, T., Gavazzi, R., Marshall, P. J., Auger, M. W., Suyu, S. H., Koopmans, L. V. E., & Bolton, A. S. 2012, ApJ, 752, 163
Sonnenfeld, A., Treu, T., Gavazzi, R., Suyu, S. H., Marshall, P. J., Auger, M. W., & Nipoti, C. 2013, ApJ, 777, 98
Stark, D. P., Swinbank, A. M., Ellis, R. S., Dye, S., Smail, I. R., & Richard, J. 2008, Nature, 455, 775
Suyu, S. H., et al. 2013, ApJ, 766, 70
Tewes, M., et al. 2013, A&A, 556, A22
Treu, T., Dutton, A. A., Auger, M. W., Marshall, P. J., Bolton, A. S., Brewer, B. J., Koo, D. C., & Koopmans, L. V. E. 2011, MNRAS, 417, 1601
Vegetti, S., Koopmans, L. V. E., Bolton, A., Treu, T., & Gavazzi, R. 2010, MNRAS, 408, 1969
Vieira, J. D., et al. 2013, Nature, 495, 344

This paper has been typeset from a TEX/ LATEX file prepared by the author.