Mon. Not. R. Astron. Soc. **000**, 1–11 (2014) Printed 1 June 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, Surhud More,[3] Michael Baumer,[2] Chris Davis,[2] Chris Lintott,[1] Robert Simpson,[1] David Miller,[4] Arfon Smith,[4] Edward Paget,[4] Prasenjit Saha,[5] Rafael Kueng,[5] 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 Institute IPMU, University of Tokyo, 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.

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

⋆ pjm@astro.stanford.edu

© 2011 RAS

## 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 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?

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. this section, we also investigate the properties of the crowd, to where the information is coming from. 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`. 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 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 pro-

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

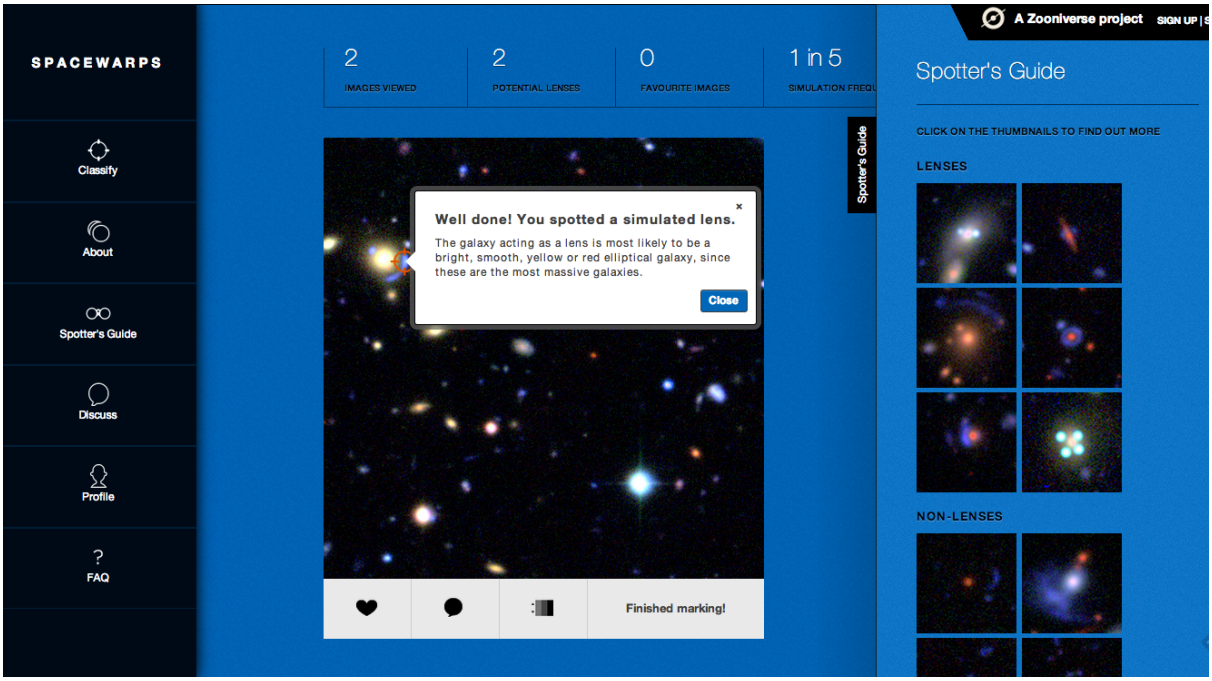


**Figure 1.** Screenshot of the Space Warps classification interface.

cedure 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 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!")

**Question from PJM: Does the feedback lead to volunteers leaving the site more quickly, or less quickly? Does negative feedback have more impact than positive?**

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 $2/(5 \times 2^{(\mathrm{int}(N_c/20)+1)/2})$ ($\approx 0.3$ 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.

### *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 Stage 1: Initial Classification

Interface fast due to pre-loading of images, and minimizing interaction. Trade-off between speed and accuracy. Decreasing training rate.

Quick dashboard provides simple ways to explore further: zoom, contrast controls.

Spotting lenses: Markers to be placed. Two reasons: first, to give good feedback. Second, to focus attention.

Non-lenses marked? Favourite button instead, enabling serendipitous discovery of other interesting things, separate from lenses.

Retirement of low probability systems. Concentrates sample, provides more "bacon" (while slightly skewing "sim frequency"). Note that this feature means that everyone contributes to detection of lenses: luck is made for the few that happen to see the new lenses, by the masses that did the rejection. Group effort.

Sims vs duds leads to inclusive search – click on anything you think etc...

### 2.4 Stage 2: Refinement

Goal: assess candidates, reject false positives by comparing with training set of non-obvious non-lenses. Produce a sample rankable by probability.

Reconfigured website: more detailed SG, more detailed feedback. Orange background to make it obvious stage 2 is different. Slower image presentation. Higher, constant training rate.

## 3 DATA

Definitions: training subjects and test subjects. Sims and duds.

### 3.1 The CFHT Legacy Survey

Describe survey. Refs.

Why this one? Good IQ, deep, colorful, homogeneous. Precursor to Stage III and IV imaging surveys, DES, KIDS, LSST etc. Already searched by robots: enables comparison of techniques. Lenses not yet found by robots, detectable by humans?

Blind search strategy. Preparation of data: divide survey into overlapping tiles.

### 3.2 Image Presentation

Presentation of images. Uniform scales, to build intuition and avoid rescales due to bright objects. Arcsinh stretch, to bring out low SB features. Approximately optimized, how? Examples of images.

## 4 CLASSIFICATION ANALYSIS

In this section we outline our methodology for interpreting the interactions of the volunteers with the identification interface. 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.

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.

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, $\Pr(\text{"LENS"}|\text{LENS}, T)$, and 2) correct when they report that an image does not contain a lens when it really doesn't contain a lens, $\Pr(\text{"NOT"}|\text{LENS}, T)$.
- 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, $\Pr(\text{LENS}|C, T)$, when that image is classified by its volunteer. ($C$ is the set of all classifications made of this subject.)

In the next section we present results in terms of all three of these probabilities, as we investigate the performance of the Space Warps system.

## 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.

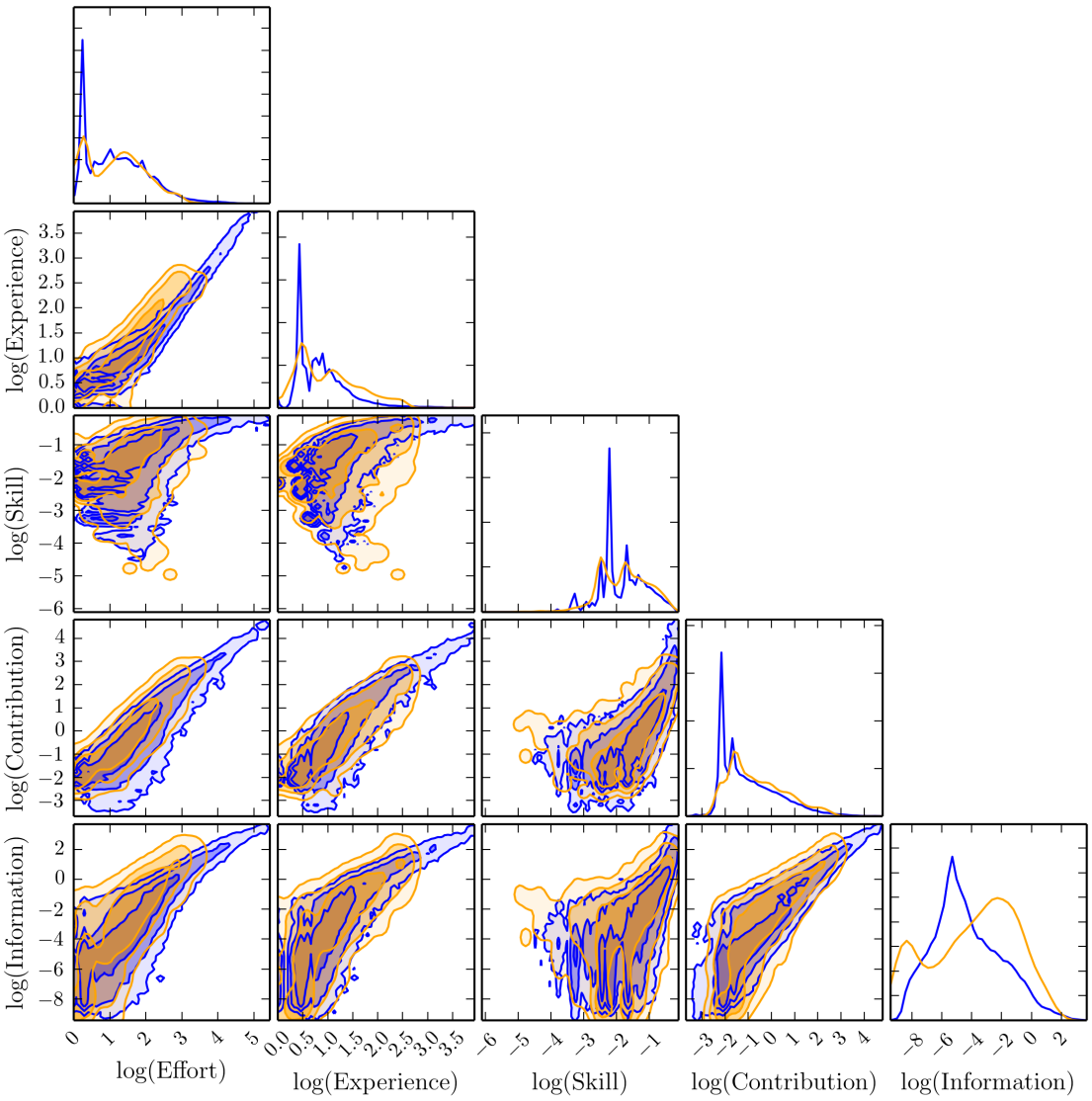


**Figure 2.** Key properties and contributions of the Space Warps crowd. Plotted are the 1-D and 2-D marginalized distributions for the logarithms of the properties of the agents described in the text. The stage 1 agents are shown in blue, the stage 2 agents in orange.

### 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 2.

**"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, $\langle I \rangle$ should the next subject classified have lens probability 0.5 Appendix A, in bits. Random classifiers have $\langle I \rangle = 0.0$, perfect classifiers have $\langle I \rangle = 1.0$. All agents start with $\langle I \rangle = 0.0$; 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 34.3 bits; in stage 2 it was 33.5.

**"Information:"** The total information $\delta I$ 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 sub-

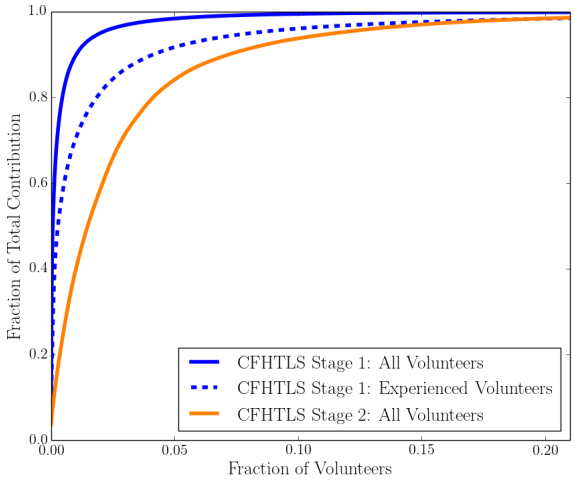


**Figure 3.** Narrow cumulative distributions of the contributions made by the agents: for example, 90% of the stage 2 contributions were made by the highest contributing 7% of the crowd. The stage 1 agents are shown in blue, the stage 2 agents in orange. "Experienced volunteers" classified 10 or more training subjects.

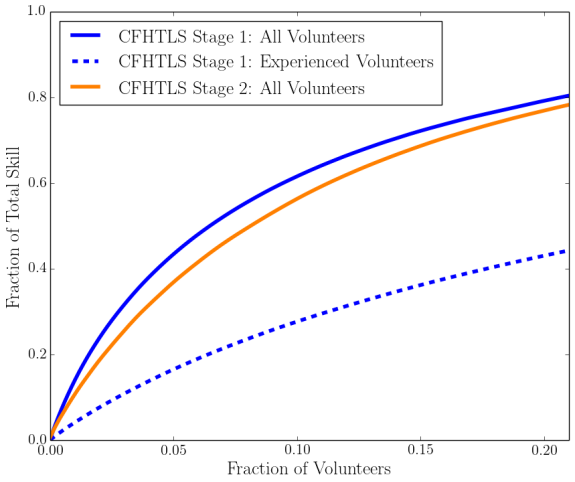


**Figure 4.** Broad cumulative distributions of agent skill: the most skilled 20% of the crowd only possess 79% of the skill at stage 1. The stage 1 agents are shown in blue, the stage 2 agents in orange. "Experienced volunteers" classified 10 or more training subjects.

ject, it's hard to generate a large amount of information. You make your own luck by classifying more subjects.

The leftmost column of Figure 2 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 2 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). Most 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 classifed, 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 2 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 3. 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.

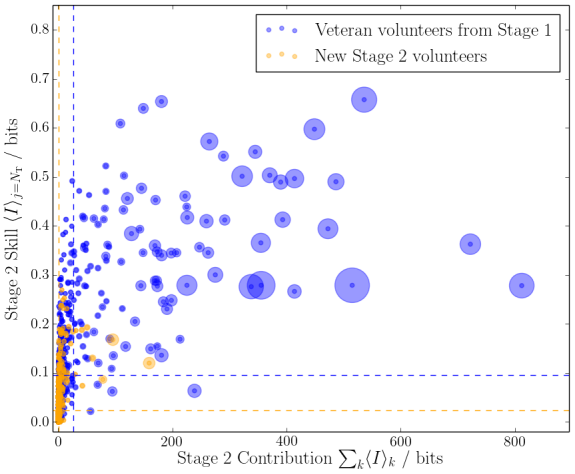


**Figure 5.** Stage 2 agent properties, separating veterans from stage 1 from new volunteers. Point size represents total information generated, dashed lines are drawn at the mean values for each sample.

However, it is not the case that only these small groups were capable of making this large contribution: 9097 volunteers were "experienced" in that they had all classified at least 10 training images; the 375 highest contributing volunteers make up just 4% of this experienced volunteer pool. The cumulative distribution of agent skill is shown in Figure 4: these distributions are significantly broader than the corresponding distributions of agent contribution in Figure 3. 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 (1820 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!).

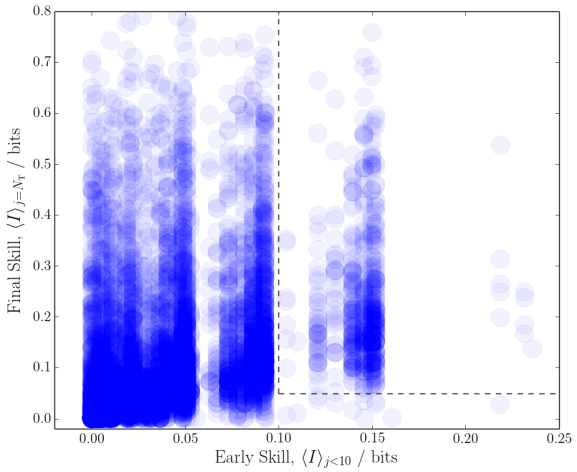


**Figure 6.** Predictability of final skill from early skill (after 10 training classifications) in stage 1 agents. The dashed lines enclose 649 agents with early skill greater than 0.1 who retain final skill of greater than 0.05.

We now look at the transition between stage 1 and stage 2 in more detail. Of the 1964 volunteers who took part in the stage 2 classification round, only 774 were veterans from stage 1. However, these volunteers showed significantly higher skill levels than the new users, and also made more classifications (and hence made greater contributions. This is illustrated in Figure 5. In this figure, point size is proportional to information generated: the higher skill, higher effort stage 1 veterans generated most of the information.

Having seen the importance of the high skill volunteers at both stage 1 and stage 2, we might ask, can we predict agent skill after, say, the volunteer has experienced 10 training subjects? In Figure 6 we plot this "early skill" against final skill, for all experienced stage 1 volunteers. The majority of the crowd shows little correlation, but at the high early skill end, some predictability appears. The 649 experienced stage 1 volunteers who had early skill greater than 0.1 went on to attain a mean final skill of 0.22, with 97% remaining at skill 0.05 or higher. This suggests that it might be worth tracking volunteers' skill as a project progresses, in order to encourage those showing an aptitude for the task to take part in the more difficult activities, such as stage 2 refinement.

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 hypthetical 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. Only 3705704 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!

The total skill of the crowd, computed by summing the skill of all the agents, is, similarly, a measure of the effective crowd size: a crowd of perfect classifiers would be this size. The stage 1 crowd was equivalent to a team of 1469.9 perfect classifiers; the stage 2 crowd was equivalent to a team of 102.4 perfect classifiers.

The total information generated during the survey is harder to interpret, because it's exactly the high probability candidates that we leave in the system that give rise to most of the information generated. What we can do is divide the total information generated by the amount of information it takes to classify a Space Warps subject all the way to the detection threshold (lens probability 0.95), divide by this number, and then multiply by the survey inefficiency to get a very rough estimate for the effective number of detections corresponding to the crowd's contribution. With a prior probability of $2 \times 10^{-4}$, it takes 11.6 bits to make a detection; the effective number of detections contributed by the stage 1 and stage 2 volunteers is then 2670 and 24 respectively. These figures are close to the numbers of detections given in column 7 of Table 1. The uncertainty in the interpretation of the information generated provides further justification for our focus on the expected information gain as a measure of volunteer contribution.

### 5.2 Sample completeness and purity

Rejection rate. Completeness and purity at P ¿ retirement, P ¿ 95%, and as function of probability P. Compare stage 1 and stage 2.

Summarize performance at some fiducial threshold: eg P = 95%.

## 6 DISCUSSION

Challenges for future.

## 7 CONCLUSIONS

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...

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

| Stage | Subjects $J$ | Contribution $\sum_k^K \langle I \rangle_k^{\rm total}$ (bits) | Agents $K$ | Skill $\sum_k^K \langle I \rangle_k$ (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 | 1291997.4 | 36929 | 1469.9 | 10768759 | 3367 | 91110.7 |
| 2 | 3679 | 21895.8 | 1964 | 102.4 | 224745 | 91 | 1641.9 |

• 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. The Space Warps project is open source, and 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|**d**). Our problem is to approximate this probability. The data **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.

### 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, $T$), 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, $T$), 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:

$$\Pr(\text{“}X\text{”}|X,T) \approx \frac{N_{\text{“}X\text{”}}}{N_X}. \quad \text{(A1)}$$

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. $T$ 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:

$$\mathcal{M}^k = \begin{bmatrix} \Pr(\text{“LENS”}|\text{NOT},T_k) & \Pr(\text{“LENS”}|\text{LENS},T_k) \\ \Pr(\text{“NOT”}|\text{NOT},T_k) & \Pr(\text{“NOT”}|\text{LENS},T_k) \end{bmatrix}. \quad \text{(A2)}$$

Note that these probabilities are normalized, such that $\Pr(\text{“NOT”}|\text{NOT}) = 1 - \Pr(\text{“LENS”}|\text{NOT})$.

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 $p_0 = 2\times10^{-4}$ 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:

$$\Pr(\text{LENS}|C_k,T_k) = \quad \text{(A4)}$$
$$\frac{\Pr(C_k|\text{LENS},T_k)\cdot\Pr(\text{LENS})}{[\Pr(C_k|\text{LENS},T_k)\cdot\Pr(\text{LENS}) + \Pr(C_k|\text{NOT},T_k)\cdot\Pr(\text{NOT})]},$$

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

$$\Pr(\text{LENS}|\text{“LENS”},T_k) = \frac{1.0\cdot p_0}{[1.0\cdot p_0 + 0.0\cdot(1-p_0)]} = 1.0, \quad \text{(A6)}$$

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 Information Contribution

The information likely to be contributed by each agent for a given subject can be estimated before the next classification of that subject is made, just from its confusion matrix. The Shannon entropy generated by a classifier upon performing a classification is

$$\langle S_k\rangle = -P_{\rm right}\cdot\log_2 P_{\rm right} - P_{\rm wrong}\cdot\log_2 P_{\rm wrong}\,, \quad \text{(A9)}$$

where $P_{\rm right}$ and $P_{\rm wrong}$ are the averages of the diagonal and the off-diagonal elements of the confusion matrix, respectively, and $\langle S_k\rangle$ is measured in “bits.” These averages represent the probability of a classifier to get a classification right or wrong, respectively. We define the information contributed by a classifier as

$$I_k = 1 - S_k\,. \quad \text{(A10)}$$

Equation A10 gives the required result, that both the hypothetical *perfect* and *obtuse* classifiers contribute 1 bit of information each, per classification. Classifiers whose agent's confusion matrix is such that $P_{\rm right} = P_{\rm wrong} = 0.5$, contribute zero bits of information. Such users identify a lens correctly with the same probability as they misclassify a dud image to contain a lens, and thus their classification is of no value.

We conservatively initialise all the elements of the agents' confusion matrices to be 0.5, that of a random classifier. This makes no allowance for volunteers that actually do have previous experience of what gravitational lenses look like, but should help prevent large numbers of false positives being assigned high probability. Plotting $\langle I_k\rangle$ as a function of time will, to some extent, illustrate the learning process undergone by the $k^{\rm th}$ volunteer-agent partnership.

### A4 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:

$$\Pr(\text{LENS}|C_{k+1},T_{k+1},\mathbf{d}) = \quad \text{(A11)}$$
$$\frac{1}{Z}\Pr(C_{k+1}|\text{LENS},T_{k+1})\cdot\Pr(\text{LENS}|\mathbf{d}) \quad \text{(A12)}$$
$$\text{where } Z = \Pr(C_{k+1}|\text{LENS},T_{k+1})\cdot\Pr(\text{LENS}|\mathbf{d}) + \Pr(C_{k+1}|\text{NOT},T_{k+1})\cdot\Pr(\text{NOT}|\mathbf{d}),$$

and $\mathbf{d} = \{C_k, T_k\}$ 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.

### A5 Uncertainty in the Agent Confusion Matrices

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 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 $\Pr(C_k|\mathrm{LENS}, T_k)$ and $\Pr(C_k|\mathrm{NOT}, T_k)$. The associated standard deviation in the estimated subject probability provides an error bar for this quantity.

## 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.