Mon. Not. R. Astron. Soc. **000**, 1–6 (2014) Printed 7 May 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,[3] Kelly Borden,[3] Chris Lintott,[1] David Miller,[3] Edward Paget,[3] Michael Parrish,[3] Robert Simpson,[1] Arfon Smith,[3] Chris Snyder,[3] Elisabeth Baeten, Claude Cornen, Cecile Faure, Thomas Jennings, Rafael Kueng,[4] Christine Macmillan, Surhud More,[2] Prasenjit Saha,[4] Matthias Tecza,[1] Julianne Wilcox, Layne Wright

[1] *Dept. of Physics, University of Oxford, Keble Road, Oxford, OX1 3RH, UK*
[1] *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] *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

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

⋆ pjm@slac.stanford.edu

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. **??**), including on sub-galactic scales (e.g. **???**). 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. **??**). Lensed quasars enable cosmography via the time delays between the multiple images' lightcurves (e.g. **??**), and study of the accretion disk itself through the microlensing effect (e.g. **?**). 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 (**?**, e.g.), SDSS (e.g. **????**), CFHTLS (e.g. **??**), Herschel (**?**) and SPT (e.g. **?**), among others. As the number of known lenses has increased, new types have been discovered, leading to entirely new investigations. Compound lenses (**??**) and lensed supernovae (**?**) 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. **?**), or the strong "magnification bias" towards detecting strongly-lensed sources in the sub-mm waveband (e.g. **?**). 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 (**?**). Several square degrees of HST images have been searched, yielding several tens of galaxy-scale lenses (e.g. **????**). 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. **??**). 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. **?**).

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. **???**), 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: **??** and **?** 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 (**?**). PlanetHunters was designed to enable the discovery of transiting exoplanets in data taken by the Kepler satellite; a community of N inspectors 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 (**?**) has also enabled the discovery of rare objects, via its flexible inspection interface and discussion forum. 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. 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?

In Section 2 we introduce the Space Warps system, describing and explaining its various features. We then briefly...

2 EXPERIMENT DESIGN

Unfamiliar objects: need to learn what lenses look like, fast. Rare objects: need to be able to reject rapidly, and get through sample. Confusion with non-lenses: further filtering after advanced training, and scientific discussion in Talk.

Intro to Classification Interface. Basic description of site.

2.1 Training

Learning what lenses do: Spotter's Guide. Learning how lenses work (science page, FAQ).

More on what lenses do: inline tutorial and feedback. Merge into stream. Instant feedback, positive and negative. Anecdotal support for this.

Training requires lenses to be more common than is realistic. How to manage expectations, avoid high false positive rate? "Lenses are rare" messaging; simulation frequency marker.

2.2 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...

2.3 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, $\Pr(\mathrm{LENS}|C,T)$, 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

Understanding crowd, so we can help them learn faster. Understanding images given the crowd, so we can find lenses.

5.1 Crowd Properties

Enthusiasm: histogram of classifications, stage 1 vs stage 2. Information contributed. Correlation with number of classifications.

PL and PD as measures of skill. not quite talent, due to possibility of learning - but agents assume talent. Performance of crowd re PD and PL. Correlations with N classifications.

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

We draw the following conclusions:

• Crowd-sourced gravitational lens detection works, as shown on sims and duds:

• Participation (crowd size, activity rate) enabled project completion

• Both stages (1 and 2) achieved the required rejection rates

• Integrated humanpower = X, cf hours taken by small team of experts

• Nightly processing is inefficient: more classifications were made than was necessary during peak participation. Need kafka...

• Completeness and purity were estimated as C% and P%, from sim and dud recovery/miss rates. Which sims were missed? False negatives

• 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: $\Pr(\mathrm{LENS}|\mathbf{d})$. 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.

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, $\Pr(\text{“LENS”}|\text{LENS}, T)$, 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, $\Pr(\text{“NOT”}|\text{NOT}, T)$, 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:

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

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:

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

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

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

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:

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

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

$$\mathcal{M}^{\rm perfect} = \begin{bmatrix} 0.0 & 1.0 \\ 1.0 & 0.0 \end{bmatrix}. \tag{A5}$$

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

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

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

$$\mathcal{M}^{\rm obtuse} = \begin{bmatrix} 1.0 & 0.0 \\ 0.0 & 1.0 \end{bmatrix}, \tag{A7}$$

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

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

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

$$\langle S_k \rangle = -P_{\rm right} \cdot \log_2 P_{\rm right} - P_{\rm wrong} \cdot \log_2 P_{\rm wrong}\,, \tag{A9}$$

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

$$I_k = 1 - S_k\,. \tag{A10}$$

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:

$$\Pr(\mathrm{LENS}|C_{k+1}, T_{k+1}, \mathbf{d}) = \tag{A11}$$

$$\frac{1}{Z}\Pr(C_{k+1}|\mathrm{LENS}, T_{k+1}) \cdot \Pr(\mathrm{LENS}|\mathbf{d}) \tag{A12}$$

$$\begin{aligned} \text{where } \; Z =& \Pr(C_{k+1}|\mathrm{LENS}, T_{k+1}) \cdot \Pr(\mathrm{LENS}|\mathbf{d}) \\ &+ \Pr(C_{k+1}|\mathrm{NOT}, T_{k+1}) \cdot \Pr(\mathrm{NOT}|\mathbf{d}), \end{aligned}$$

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. $\Pr(\mathrm{LENS}|\mathbf{d})$ is the fundamental property of each test subject that we are trying to infer. We track $\Pr(\mathrm{LENS}|\mathbf{d})$ 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

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