Mon. Not. R. Astron. Soc. **000**, 1–5 (2013) Printed 10 February 2014 (MN LaTeX style file v2.2)

Space Warps: I. Crowd-sourcing the Discovery of Gravitational Lenses

Phil Marshall,[1,2*] Aprajita Verma,[1] Anupreeta More,[3] and others

[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*

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 flexible interface, which records their estimates of the positions of candidate lensed features. Simulated lenses, and expert-classified non-lenses, are inserted into the 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 daily, concentrating the sample towards a set of candidates; these are then re-classified by the volunteers in a second refinement stage. Analyzing the classification of the training set, we predict that the first stage should yield a sample with C% completeness and P% purity, 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 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 estimate the mean information contributed per person to be X bits, over a session lasting, on average, N classifications per volunteer – although the distributions of these quantities are highly skewed. We comment on the scalability of the Space Warps system, and its potential to operate beyond its design as a supervised classification system.

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

1 INTRODUCTION

Scientific motivation. Applications of lenses: group-scale arcs, galaxy-galaxy lenses, lensed quasars.

Problem of rarity. Imaging surveys. Problem of purity/false positives.

Review of progress to date. Methods in SL2S, SQLS. Contrast with SLACS.

Scaling to wide field era. Automated methods: problems. Need for good training sets. Need for quality control: always present.

Novel solution: crowd-sourcing. Brief review of similar problems. PlanetHunters. Space Warps as an experiment.

In this paper, we describe the Space Warps website, an online system that enables crowd-sourced detection of gravitational lenses. In a companion paper we will present the new gravitational lenses discovered in our first imaging survey dataset, 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?

* pjm@slac.stanford.edu

- 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 crowd-sourcing 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 a "Pseudo-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 allows us to track the speed with which the crowd learns about lenses, and also gives us a dynamic estimate of the posterior probability for any given subject being a lens, given all classifications of it. Assigning thresholds in this lens probability allows us to make good decisions about whether or not to accept a subject into the collection of candidates visible in Talk, and also whether or not to carry on classifying a subject at all.

In this paper we focus on the training data, investigating how the crowd's ability to identify gravitational lenses during the course of the project, and the completeness and purity of the lens candidate sample generated.

Describe SWAP here. (Move from appendix.)

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 DATA MODEL

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 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 t he 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: 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 by the television). To cope with this uncertainty, we assign a 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. We now describe how each agent works.

Each agent assumes that the probability of a volunteer recognising any given simulated lens as a lens is some number, $\Pr(\text{“LENS”}|\mathrm{LENS}, T)$, 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, $\Pr(\text{“NOT”}|\mathrm{NOT}, T)$, 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:

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

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

As an example, 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 are as helpful as *perfect* ones!

Indeed, the information content of each classification can be estimated by an agent before the next classification 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.

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:

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

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.

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 user increases. For simplicity, the discussion thus far assumed the case when the confusion matrix is known perfectly. Let us first discuss how to characterize the noise in the confusion matrix. ** This needs work **

For ease of notation, we will denote $\mathrm{Pr}(C_k|\mathrm{LENS}, T_k) \equiv p_L$ and $\mathrm{Pr}(C_k|\mathrm{NOT}, T_k) \equiv p_N$. In reality, there is a probability distribution for both p_L and p_N. Let p_0 be the prior probability of the subject being a lens. Then the posterior probability, p_0' of the subject being a lens after the classification C_k is

$$p_0' = \frac{p_L p_0}{[p_L p_0 + p_N(1-p_0)]}, \tag{A13}$$

The posterior probability distribution p_0' can be obtained by marginalizing over the probability distributions of p_L, p_N and the prior probability distribution p_0 such that,

$$P(p_0') = \int p_0' P(p_L) P(p_N) P(p_0) dp_L dp_N dp_0 \,. \tag{A14}$$

This marginalization is not analytically tractable. Therefore, we have implemented the following Monte-Carlo solution for this problem.

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