Mon. Not. R. Astron. Soc. **000**, 1–4 (2013) Printed 17 October 2013 (MN LATEX style file v2.2)

SPACE WARPS: I. Crowd-sourcing the Discovery of Gravitational Lenses

Phil Marshall,[1⋆] Aprajita Verma,[1] Anupreeta More,[2] Amit Kapadia,[3]
Kelly Borden,[3] Chris Lintott,[1] David Miller,[3] Robert Simpson,[1] Arfon Smith,[3]
Elisabeth Baeten, Claude Cornen, Cecile Faure, Thomas Jennings,
Jean-Paul Kneib[4], Rafael Kueng,[5] Stuart Lowe, Christine Macmillan,
Surhud More,[2] Prasenjit Saha,[5] Matthias Tecza[1], Julianne Wilcox, Layne Wright

[1] *Dept. of Physics, University of Oxford, Keble Road, Oxford, OX1 3RH, UK*
[2] *Kavli Institute IPMU, University of Tokyo, Japan*
[3] *Adler Planetarium, Chicago, IL, USA*
[4] *EPFL, Lausanne, Switzerland*
[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 arbitrarily 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 as well as estimate a dynamically-updated probability for any given image containing a lens. High probability systems are filtered into the Zooniverse TALK discussion system; this smaller sample is further classified and analysed by volunteers into a final set. We analyze the classification of the training set itself, and find that 1) <volunteers learn>... 2) <the filter works>...

Key words: gravitational lensing – methods: statistical

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. Other papers in this series will present new gravitational lenses discovered in our first imaging survey dataset, and investigate the differences between lens detections made in SPACE WARPS and those made with automated techniques. Here, we try to answer the following questons:

- 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 accurately can gravitational lenses be located on the sky during the identification process?
- How quickly can lenses be detected, and non-lenses be

⋆ dr.phil.marshall@gmail.com

rejected? How many volunteers are needed per target, and how quickly can they mobilise?

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 via scientific discussion and further classification (voting).

2.1 Tutorial Material

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

Inline tutorial. Merge into stream. Instant feedback, positive and negative. Anecdotal support for this.

2.2 The Classification Interface

Image presentation. Consistency in size and appearance.

Interface fast due to pre-loading of images, and minimizing interaction.

Spotting lenses: Markers to be placed. Two reasons: first, to give good feedback. Second, to improve quality of discussion later.

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

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

2.3 Collaborative Filtering with "Talk"

Feeding in collections, following online analysis. Taking subjects to Talk individually, too.

Mental modeling. Further inspection on Dashboard. Voting.

Final sample generation.

3 DATA

Training data and test data. Refer to Paper II for CFHTLS details.

Presentation of images. Arcsinh stretch, calibration. Approximately optimized.

4 IDENTIFICATION 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.

Where to describe probabilistic classifier stuff? Here or in an appendix? Not sure.

5 IDENTIFICATION RESULTS

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

5.1 Learned Behavior

5.2 Sample completeness and purity

Vary the acceptance and rejection thresholds, look at completeness and purity as a function of this. How much time can be saved by retiring subjects into candidate list or rejection list?

6 DISCUSSION

Challenges for future.

7 CONCLUSIONS

We draw the following conclusions:

- Ability to detect lenses
- What we can say about the next steps

Summarize, 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(LENS|**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("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("NOT"|NOT) = 1 − Pr("LENS"|NOT).

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

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

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

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

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

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

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

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 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

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

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.

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.

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 $\Pr(C_k|\text{LENS}, T_k) \equiv p_L$ and $\Pr(C_k|\text{NOT}, T_k) \equiv p_N$. 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

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

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,

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

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

REFERENCES

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