Limitations of majority agreement in crowdsourced image interpretation

*Carl F Salk[1,2], Tobias Sturn[1], Linda See[1], Steffen Fritz[1],

Transactions in GIS, 2016

(1) Ecosystems Services and Management Program, International Institute for Applied Systems Analysis, Schlossplatz 1, A-2361 Laxenburg, Austria

(2) Southern Swedish Forest Research Center, Swedish University of Agricultural Sciences, Box 52, S-23053 Alnarp, Sweden

*Corresponding author: salk@iiasa.ac.at; +43(0) 2236 807 293

Running Head: Limits of the crowd in VGI

Keywords: Crowdsourcing, Volunteered geographic information, Expert validation, Image interpretation, Task difficulty

Abstract

Crowdsourcing can efficiently complete tasks that are difficult to automate, but the quality of crowdsourced data is tricky to evaluate. Algorithms to grade volunteer work often assume that all tasks are similarly difficult, an assumption that is frequently false. We use a cropland identification game with over 2600 participants and 165,000 unique tasks to investigate how best to evaluate the difficulty of crowdsourced tasks and to what extent this is possible based on volunteer responses alone. Inter-volunteer agreement exceeded 90% for about 80% of the images and was negatively correlated with volunteer-expressed uncertainty about image classification. 343 relatively difficult images were independently classified as cropland, non-cropland or impossible by two experts. The experts disagreed weakly (one said impossible while the other rated as cropland or non-cropland) on 27% of the images, but disagreed strongly (cropland vs. non-cropland) on only 7%. Inter-volunteer disagreement increased significantly with inter-expert disagreement. While volunteers agreed with expert classifications for most images, over 20% would have been mis-categorized if only the volunteers' majority vote was used. We end with a series of recommendations to manage the challenges posed by heterogeneous tasks in crowdsourcing campaigns.

1 Introduction

Crowdsourcing is a powerful tool to perform tasks requiring human input that would be prohibitively expensive if paid for in a conventional way. Crowdsourcing has emerged from a business-oriented domain (Howe, 2006) in which different types of micro-tasks are outsourced to a willing labor force or interested volunteers but in recent years has been adopted for a broader array of data collection and processing tasks. Where the goal of data collection or processing is scientific, the involvement of citizens in research has been termed 'citizen science' (Bonney et al., 2009). Many citizen science projects maintain the traditional micro-task approach of crowdsourcing, ranging from relatively simple to highly-skilled tasks such as bird identification (Silvertown, 2009). However, involvement at higher levels, e.g. in hypothesis generation and research design, is a goal of many citizen science projects (Haklay, 2013). When crowdsourced data has a spatial aspect, it is often referred to as 'Volunteered Geographic Information' (VGI; Goodchild, 2007). VGI can be solicited in a variety of ways, particularly through mobile phones and social media. Regardless of the purpose of crowdsourcing, data quality is a fundamental issue that arises for inputs generated by non-specialists, whether the eventual goal is scientific analysis (Hunter et al., 2013) or conflation with authoritative data (Pourabdollah et al., 2013). Data quality assessment is of particular importance as it has implications for how volunteers are motivated, evaluated and rewarded (Oreg and Nov, 2008; Raddick et al., 2013).

Citizen science contributors may be accorded credit for their work in a variety of ways. On the most basic level, contributors take part for some sense of personal reward. A study of the highly successful Galaxy Zoo project identified at least 12 distinct personal motivations for taking part (Raddick et al, 2013). Volunteers may be awarded points proportioned in some way to their work, for instance for the total number of tasks completed, the number of tasks completed correctly, the accuracy of task completion, or some combination of these metrics (Ipeirotis et al., 2010; Wang et al., 2013), among

others. Contributor credit may be confined entirely to the 'game world' such that points accrued cannot be converted into cash or other goods. In-game recognition has proven to be a powerful motivator in many scientific discovery games (Mekler et al., 2013). Points earned by volunteers may also be converted to tangible rewards, for instance by paying a fixed amount for successful task completion, as is done using the Amazon Mechanical Turk platform (Buhrmester et al., 2011). Intermediate solutions are also possible, for instance entering top players into prize draws (See et al., 2014b) or awarding them co-authorship for their contributions to resulting manuscripts (Fritz et al., 2013; See et al., 2014a). The simplest reward structure for crowdsourcing campaigns is to award points uniformly for each task completed, or for each task completed successfully. Rewarding quantity (rather than quality) is a common, but slowly changing, feature of crowdsourcing efforts (Wang et al., 2013). Uniform rewards may be appropriate where task difficulty does not vary greatly, but what happens when this is not true? A more accurate evaluation of contributor quality, and a more nuanced awarding of credit, should take the difficulty of individual tasks into account.

Some authors suggest that it is possible to evaluate the correct answer to a task and volunteer quality using only volunteer-contributed data, or as one publication memorably put it, 'to grade a test without knowing the answers' (Bachrach et al., 2012). Dawid and Skene (1979) proposed what may have been the first algorithm attempting to do this, using an iterative maximum likelihood method to simultaneously estimate the correct response and the error rate of each rater. More recent work has built on this method, improving its efficiency, and perhaps extending its usefulness to somewhat noisier data, typically using Bayesian estimation (Wang et al., 2013; Bachrach et al., 2012; Whitehill et al., 2009; Welinder et al., 2010). Indeed, some successful crowdsourcing campaigns such as the 'ESP Game' for image labeling (von Ahn and Dabbish, 2004) and Galaxy Zoo (Lintott et al., 2008) have made little or no use of validated tasks. Because of the substantial literature on these expert-free methods, and their

appeal for crowdsourcing applications, we examine the consequences of assuming that the wisdom of the crowd is correct.

We address these questions in the context of a simple land-cover classification task. There is an extensive literature on the factors that make images difficult for humans to classify, particularly in the field of aerial photography and satellite imagery. For instance, studies have quantified how complex backgrounds slow down recognition of foreground objects (Lloyd and Hodgson, 2002) and the minimum resolution needed to identify disaster-caused damage to buildings (Battersby et al., 2012). The psychological underpinnings of these factors have also received substantial attention (e.g. Hoffman, 1990; Bianchetti, 2014). Our work takes the opposite approach. Rather than building a model of task difficulty from a basis of cognition and image composition, we ask what can be learned about task difficulty from the patterns of player responses themselves. This knowledge is particularly valuable in the context of crowdsourcing. Even if an image classification activity has a well-developed theory of what makes it difficult, applying this theory would require a separate evaluation of each image. For some activities, computers may be able to rate the difficulty of tasks, but in this case the tasks themselves are unlikely to require human interpretation, and expert pre-screening of images would defeat the purpose of crowdsourcing.

As a first step toward relating payment or reputational credit for crowdsourcers to the difficulty of the work they complete, we evaluate different metrics for assessing the difficulty of tasks using a land-cover classification example from the Cropland Capture game. We compare disagreement among volunteers and volunteer-reported uncertainty with expert-derived measures. The results show that while the different difficulty measures show positive relationships with one another, evaluations based on volunteers' data alone tend to underestimate the difficulty of tasks. Further, for a non-trivial fraction of tasks, the wisdom of the crowd was wrong, greatly complicating the assessment of other, non-

validated tasks. However, our findings are limited by the difficulty of implementing after-the-fact expert validations when both the number of tasks and number of volunteers is large. We end with recommendations on game design to avoid this problem and achieve more robust within-game evaluations of task difficulty.

2 Methods

2.1 The Cropland Capture game

Cropland Capture is an online and mobile game in which volunteers are shown an image, either remotely-sensed or taken with a ground-based camera, and have to answer a yes/no question on whether any cropland is present in the image. The cutoff for a 'yes' response was any cropland at all, even if just a tiny fraction of the area contained cropland. If uncertain, volunteers had the option to respond 'maybe.' The game ran for about six months, from November 2013 to May 2014. As a motivation for their participation, the top three volunteers each week (as measured by their points earned) had their names entered into a raffle to take place at the end of the competition. Prizes awarded through the drawing included smartphones and tablets. Some volunteers were among the top three in many different weeks; this gave them increased chances to win a prize in the drawing. To encourage participation in the final phase of the competition, additional prizes were awarded during each of the last five weeks.

For the purpose of assigning points, the correctness of an answer was determined with only data supplied by the crowd, but with some forgiveness in cases where the volunteers did not show a clear consensus. If there was at least 80% agreement among volunteer ratings (excluding responses of 'maybe'), then only ratings agreeing with the majority were counted as correct. When this condition

was not met (i.e. the proportion of ratings of cropland was between 20% and 80%) both 'no' and 'yes' were treated as correct answers. The values of 20% and 80% were chosen to balance false positive and false negative responses. Had the cutoffs been closer to 0% and 100%, a few contrary ratings of an easy image could result in all responses receiving points. Had they been closer to 50%, more potentially correct ratings of difficult images would not be credited. Responses of 'maybe' were neither awarded points nor penalized. Later in this paper the idea of a 'correct' rating will be treated in more nuanced ways, but this method was chosen as a fast technique for run-time evaluation of responses.

Most image locations were taken from the Zhao *et al.* (2014) global validation set. From each point, pixels of varying scene size were taken. Although high-resolution imagery was preferred, in some cases Landsat imagery was used if no alternative could be found. Images ranged from 100 to 1000 m on a side, but filled the same area on the volunteer's screen regardless of the land area covered. The satellite images were supplemented with ground-based photographs taken from the Degree Confluence Project website (confluence.org). At some locations, multiple images with different spatial extents were used in the game.

An example volunteer's view of the game is seen in Figure 1. As stated above, the volunteer was asked to determine whether the image contained cropland or not, or if they were unsure. Cropland was defined as arable land supporting annual or perennial crops (FAO, 2013). This definition includes temporary agricultural crops (i.e. bare harvested land in winter qualifies), market and kitchen gardens, land fallow for less than five years and 'permanent' long-term crops such as orchards or coffee but excludes forest plantations for wood, rather than food or fodder. Grazing lands fall outside this definition, but land regularly mowed for hay is considered cropland. The volunteers had access to a gallery of annotated example images to help them understand these distinctions and that the cutoff for calling an image cropland was presence of any at all within the red box (see examples in Figure 2).

However, there was no formal training period to help encourage as many volunteers as possible to participate.

2.2 Expert validation

As a basis for several of the task difficulty measures presented here we undertook an expert validation exercise to provide information independent from the volunteer ratings. In the validation exercise 343 images were provided to two remote sensing specialists (LS and SF) who independently classified them according to the same scheme used in the game ('yes', 'no' or 'maybe'). Images for expert validation were not chosen randomly, but rather on the basis of several criteria. First, they were chosen for having been rated many times. Over 95% of these images have been rated over 30 times by volunteers. Second, they were stratified on the basis of falling into one or more of several difficulty categories. The categories included two groups of apparently easy images in which volunteers overwhelmingly agreed on a rating of either cropland or no cropland with very few 'maybe' responses, and several groups of potentially very difficult images. These difficult groups were (1) images with a high proportion of 'maybe' responses, (2) images receiving substantial numbers of both 'no' and 'yes' ratings, (3) images where top raters who usually agree with the majority classification did not do so, and (4) images where the top raters disagreed with one another. In practice it was possible for an image to fall into more than one of these categories. It is important to note that on average these expert-validated images are much more difficult than the broader population of images included in the game. This stratified sampling was necessary to achieve sufficient representation of the more difficult categories for a statistically robust comparison of their properties. Stratification also helped to differentiate the quality of work among individual raters (Salk et al., 2015).

The expert validators worked independently and did not agree on all images. Rather than viewing this as a problem, we used expert disagreements and a further validation exercise as the basis

for an ordinal difficulty ranking scheme. Based on the pairs of expert responses for each image, they were categorized into 'easy', 'moderate' and 'difficult' groups. An image was called 'easy' if the raters agreed on a 'cropland' or 'no cropland' rating. A 'moderate' classification was given in the case of weak disagreement between the experts, when one responded 'maybe' and the other responded with 'cropland' or 'no cropland.' Finally, a 'difficult' rating resulted if there was a strong disagreement (one expert said 'cropland' and the other 'no cropland') or if both raters responded 'maybe'.

All of the moderate and difficult images and some of the easy images were subjected to a secondary validation exercise in which the two experts worked together with another observer (CS) to come to a consensus on the assessment of an image. This process drew on both the original images and also looked at Google Earth in the area of the image for landscape context and potential higher-resolution views. However, the ultimate classification was based on what a viewer could reasonably infer from the image presented in the game; other information was used only for confirmation. If it appeared impossible to determine whether an image contained cropland without additional information, that image was put into an 'impossible' category. These images are examples of 'maybe' arguably being the correct rating. Taken together these validation exercises yielded seven categories: easy/moderate/hard cropland, easy/moderate/hard non-cropland, and impossible to determine. However, relatively few images fell in the moderate and hard bins, so those images were combined into a single category for analysis.

2.3 Data analysis

In addition to the ordinal ranking derived from the expert validations, several metrics were devised as potential measures of image difficulty. The most basic was the proportion of 'maybe' ratings returned for a particular image, in other words, the number of 'maybe' responses divided by the total number of responses, a value that can theoretically vary from 0 to 100%. The second was the level of

disagreement with the majority rating (inter-volunteer disagreement, excluding responses of 'maybe'), equal to the count of minority ratings divided by the total number of cropland and non-cropland ratings. This metric has a theoretical maximum of 50% in the case of an equal number of cropland and non-cropland responses. These calculations and all analyses described below were implemented in R version 3.0.3 (R Core Team, 2014).

Because of the differences among these metrics (continuous vs. categorical and different distributional characteristics), a variety of parametric and non-parametric statistics were used. To analyze the relationship between disagreement with the majority classification and rate of using the 'maybe' response we used major-axis (MA) regression (also known as type II regression). Unlike the more commonly used ordinary least squares (OLS) regression, MA regression admits error in both variables. Thus, its results do not change when the independent and dependent variables are switched. This is a useful feature when there is no hypothesized causal relationship among the variables, as is the case here. Both of these variables also had extremely non-normal distributions (Shapiro-Wilk test, $p<<.05$), particularly due to many values of zero. This situation was improved somewhat, but not entirely corrected, by log transformation (a small constant of .02 was added to both variables to prevent 0 being transformed to negative infinity).

The image-specific rate of 'maybe' responses was analyzed as a function of the expert-assessed difficulty category. Typically, this analysis would be an ANOVA, however, the distribution of the 'maybe' rate within each category was clearly non-normal (Shapiro-Wilk test, $p<<.05$). As such, non-parametric tests for heteroscedasticity (i.e. differences among categories' variances) were required. We used the Brown-Forsythe test which confirmed differences among these groups' variances ($p<<.05$). Because both of these situations violate assumptions of a standard ANOVA, we used Welch's ANOVA, a non-parametric alternative that is robust to both non-normality and non-constant variance. An appropriate

post-hoc test in this situation would be a Games-Howell test but this appears not to be implemented in any R package. As a conservative alternative, we applied a Kolmogorov-Smirnov test (which is robust to non-normality and unequal variance) to each of the three pairs of difficulty categories and applied a Bonferroni correction to the resulting *p*-values by multiplying them by 3.

The image-specific rate of disagreement of the crowd with experts was also non-normally distributed within the easy and difficult image groups (Shapiro-Wilk test, $p<<.05$) and heteroscedastic (Brown-Forsythe test, $p=.006$). The difference between these two groups was assessed using the Kolmogorov-Smirnov test.

3 Results

The Cropland Capture game provided ratings on a total of 165,439 unique images delivered by 2,738 volunteers. Of these images, 50.3% were remotely-sensed with the balance being land-based photographs. Taken together, a total of 4,547,083 ratings were undertaken across all images and volunteers. This number includes a few images that were seen repeatedly by the same volunteer to test his or her consistency. Normally a volunteer saw about 2% of images more than once, although a few volunteers contributed more ratings than there were images, so necessarily had a substantial number of repeat views. Images received between one and 5,179 ratings each. The median number of ratings per image was 13, with most images receiving between 10 and 100 ratings (Figure 3). A few images underwent extremely large numbers of ratings due to unintentional assignment of many different image numbers to a single image, and were thus rated much more frequently than intended. These duplicate images were combined into a single identifier for analysis.

Expert-provided validations agreed perfectly 66.5% of the time. In other words, the two experts both chose ‘no,’ ‘maybe,’ or ‘yes’ for two-thirds of the images. The experts showed weak disagreement (one chose ‘maybe’ while the other chose ‘yes’ or ‘no’) 26.5% of the time. They exhibited strong

disagreement (one said 'yes' while the other said 'no') only 7.0% of the time (Figure 4). After the consensus building exercise, the ratings of 97.4% of the images on which the experts agreed were left unchanged if ratings of 'maybe' are considered to indicate that an image is impossible. For 42.8% of the images on which the experts showed weak disagreement the consensus exercise determined that the image was impossible to classify. The corresponding figure was 45.8% for images on which the experts showed strong disagreement (Figure 4).

Of the images in the Cropland Capture game, most showed low apparent difficulty levels. Among-volunteer disagreement exceeded 10% for only 20.8% of images (Figure 5A). The proportion of 'maybe' responses exceeded 10% for 4.0% of images (Figure 5B). Of the 343 expert-validated images, 258, or 75%, were considered to be possible to rate by a volunteer. Of these 258, most images (219, or 85%) were considered easy to rate correctly (Figure 5C). The remainders were either classified as moderately difficult (12) or difficult (27). Because of the small sample sizes for the last two groups, they were combined into a 'moderate to difficult' group for subsequent analyses. As images were not chosen randomly for this task (to ensure sufficient representation of potentially difficult images), these expert-validated images over-estimate the average difficulty of tasks in the Cropland Capture game.

Volunteer-expressed uncertainty ('maybe' rate) was significantly but weakly positively correlated with the rate of disagreement among volunteers for a particular image (major axis regression, $R^2 = .230$, $p=.001$, n=45,991; Figure 6). However, note that even with log transformations, this relationship is far from normally distributed and homoscedastic. In particular, there is an inflated number of zeros, especially for the 'maybe' rate, which is why the regression line is so low in Figure 6. Regardless of these distributional problems, it is important to note that all combinations of high and low disagreement rates and 'maybe' rates are seen among the images used in Cropland Capture (Figure 6).

For images in the expert-determined 'easy' category, the volunteer classification usually agreed with the expert validation, regardless of whether the correct classification was cropland or non-cropland (Figure 7A, 6B, 8B). For these images, just over 50% of images had an agreement rate with experts of ≥90%. For moderately difficult to hard images, volunteer majority votes appeared random (at least visually as the sample size was not large enough for robust statistics) regardless of whether the image was classified as cropland or non-cropland by experts (Figures 6C, 6D). For images determined to be impossible to evaluate, the crowd had a strong bias toward ratings of non-cropland (Wilcoxon rank sum test with hypothesized mean of .5, p=.0001; Figure 7E). In spite of this skewed rating of impossible images, there was no bias in either direction for easy images; for cropland images, 76.5% of volunteer ratings agreed with expert validations; for non-cropland images, 77.9% of volunteer ratings agreed with expert validations. These values were statistically indistinguishable (Kolmogorov-Smirnov test, p=1.00).

The rate of 'maybe' responses increased with expert-assessed difficulty of images (Figure 8). Images rated as easy had responses of 'maybe' on average 2.95% of the time. This rate increased to 6.76% for moderate to difficult images and 10.5% for impossible images. These values differ significantly from one another (Welch's ANOVA: $p < .0001$, $F = 40.816$, $df = 2$). A post-hoc test showed that the easy group differed strongly from the difficult and impossible groups (Kolmogorov-Smirnov test, $p<.0001$ after Bonferroni correction), but the difference between the difficult and impossible groups was weak (K-S test, $p= .0924$; p=.277 after Bonferroni correction).

The majority vote of the volunteers disagreed with the expert validations for 22% of images (grey bars in Figure 9A). The total proportion of volunteers disagreeing with the expert validation was greater for moderate to difficult images (49%; Figure 9C) than for easy images (23%; Figure 9B). The difference between these two groups of images was statistically strong (Kolmogorov-Smirnov test; p<.0001).

4 Discussion

This paper has highlighted some of the challenges of assessing the difficulty of image classification tasks based on volunteer responses alone. Although we have shown consistent positive relationships among three independent task difficulty metrics (volunteer-expressed uncertainty, inter-volunteer disagreement, and expert evaluation), we have also demonstrated that without careful implementation, all of these methods have their limits. In particular, the frequency of majority classifications disagreeing with expert validations leads to significant questions about some commonly recommended methods of inferring correct image classifications. The remainder of this paper discusses these issues and ends with recommendations for game design to make robust use of gamified data collection campaigns.

All volunteer-derived difficulty metrics assessed in this study increased with expert-assessed difficulty. This suggests that inter-expert comparisons are an effective way to assess how hard a task is. Further, our results present indirect evidence that expert difficulty ratings are more robust than volunteer-derived ratings. This can be seen from the surprisingly infrequent use of the 'maybe' rating by participants, in spite of the risk of losing points for incorrect classifications. The very hardest images (those determined to be impossible to rate through the multi-step expert evaluation process) were only classified as 'maybe' 10.5% percent of the time. It may be possible to improve on this situation through altered game design, an issue we return to below.

For a non-trivial fraction of images, the majority classification disagreed with the expert classification. This is cause for serious concern and suggests that the guidance provided to Cropland Capture volunteers (majority-derived feedback as to whether a rating is correct) was insufficient. Additional steps are needed to improve the quality of responses, particularly on very difficult tasks. One possible solution is better training of volunteers. In Cropland Capture there was no formal training

period. Volunteers had to rely on their own knowledge and intuition, plus a small library of images with explanations (see examples in Figure 2). A second (and related) solution is to provide feedback as to whether a rating is correct on the basis of expert-validated images only. Because Cropland Capture awarded points based on majority classifications of images, it is likely (see below) that the crowd developed rating norms that were at odds with those used by expert validators and mapmakers. While these suggestions may reduce confusion by volunteers, they cannot guarantee that no image will have a majority classification at odds with its expert classification. This finding calls for a different approach to game design, one that has a central and ongoing role for expert-validated images, a task we turn to in the following section.

As discussed in the introduction, some schemes for evaluating the correct response to a task are based on volunteer data alone (e.g. Dawid and Skene,1979; Bachrach et al., 2012; Wang et al., 2013). Such algorithms may suffice when all tasks are relatively simple and most ratings disagreeing with the majority can be attributed to rater inattention. However, where the task pool contains many difficult items, as in Cropland Capture (Figure 9), these schemes will likely fail. It is difficult to see how any such procedure could have correctly classified a task where the crowd was confidently wrong, as in the right hand side of the distributions in Figure 9. In Bayesian models, strong priors about the correct classification of an image could theoretically solve this problem. However, if such information were available, one might reasonably ask why crowdsourcing is necessary; after all, the answer to the question is already known with a fairly high degree of confidence. In real crowdsourcing tasks, such information is usually lacking, so these solutions all rest on the assumption that the wisdom of the crowd is basically correct.

Volunteers used the 'maybe' response surprisingly rarely. This can be seen most starkly for images with no visible land, for instance due to clouds or corrupted files. Most such images had less

than 50% ‘maybe’ ratings, even though they are clearly uncategorizable. Why this happens warrants further exploration. We can think of three possible explanations. The first is that wrong answers were not penalized strongly enough. Answers that disagreed with the majority were given a -1 point penalty, which has the same magnitude of the +1 point reward for correct answers. However, since penalties were only given for responses that disagreed with a majority of 80% or more, even random guessing would result in positive point accumulation. Thus, volunteers might not have seen any benefit to choosing ‘maybe’, resulting in no points earned or lost, when a random guess would average a positive point gain. A second explanation is that volunteers simply found it easier to consider only two possibilities, and in the relatively rare cases where images were indisputably impossible, they simply chose a response out of habit. The final explanation is that this behavior was a result of the wording of the question “Is there any cropland in the red box?” (Figure 1). This question could be interpreted as referring to the image itself, rather than the landscape (hopefully) depicted in the image, as the game designers intended. Thus, the question effectively becomes ‘Is any cropland visible within the red box?” to which an answer of ‘no’ is defensible when the image does not depict land cover at all.

In reality, the possibilities outlined in the previous paragraph are not mutually exclusive. However, the best evidence we have is for the third possibility. While systematically searching all 165,000 images for blank/cloudy/shaded scenes was not practical, those that we did encounter by chance were typically given many ‘no’ responses, a moderate number of ‘maybe’ responses, and very few answers of ‘yes.’ This pattern suggests many volunteers made a particularly literal interpretation of the question. This evidence opposes the first possibility, since ‘no’ ratings outnumbered ‘yes’ by a big enough margin so that only ‘no’ responses would have been credited as correct. This process could have been iterative and self-reinforcing. As more volunteers rated these bad images as not containing cropland, and saw that they received a point for such a rating, they may have come to see this as a correct answer.

5 Lessons for game design

The best solution to the problems discussed above is to prevent them in the first place. In this section, we present some strategies to accommodate widely varying task difficulty. These recommendations aim to increase the efficiency of crowdsourced data acquisition and make more effective use of both experts' and volunteers' time. The list includes well-defined tasks, pre-validation of select tasks by experts, guidance for new volunteers, real-time performance monitoring, outlets for expression of uncertainty, and eliminating shortcuts that undermine the game's scientific purpose.

Clearly define the task and questions. This may seem obvious, but seemingly simple tasks and instructions can be subject to misinterpretation. We suspect that the Cropland Capture game suffered from the wording of its single question 'Is there cropland in the red box?' (Figure 1). Superficially, this seems like a simple and straightforward question. Indeed, it seems to have been effective for simple-to-evaluate images. However, for impossible images (e.g. obscured by clouds or blank due to failed downloads), many volunteers answered 'no' (Figure 7E). This is literally correct in that there was no cropland in the red box, but from the perspective of someone wanting to use this data to validate maps, it is a misleading response. Ratings of 'maybe' for these problematic images may have been more common if the question had been worded differently, for instance 'Is there any cropland in the area bounded by the red box?' Explicit examples and training would also help with this problem. This kind of difficulty is not always easy to foresee, so pre-testing before public release of a game may help identify 'unknown unknowns' if time and resources permit. Ideally, this step should also include estimating the number of tasks, volunteers, and tasks completed by each volunteer since they determine how many images and what difficulty of images should be validated by experts. However, these numbers depend on publicity, task difficulty and volunteer interest, so may not be easy to predict.

Develop an expert-validated image bank before opening the game to public participation. This step brings several benefits. The most basic is amassing a set of core tasks whose correct answers are clearly agreed upon by experts. These tasks should be chosen such that they represent the breadth of difficulty of tasks included in the crowdsourcing campaign, and some assessment of task difficulty should be done. Such an image bank provides several benefits. Assigning these validated tasks to all volunteers makes it possible to statistically assess volunteers' performance against one another and against experts in real time. This also ensures that all volunteers rate enough validated images so that they can be robustly compared with one another (Salk et al., 2015). Information on task difficulty can provide a more nuanced view on the skill level of individual volunteers. Developing an expert-validated image bank also gives the team a chance to learn about possible pitfalls in the game or image set. For instance, this process may help elucidate ambiguities in the task definition or determine specific types of images that volunteers may find difficult so that a set of examples can be provided to explain what the correct answer is in these situations.

Provide explicit guidance, especially when a new volunteer joins. Volunteers will quickly develop habits of how to perform tasks as they take part in citizen science campaigns. For the sake of data quality, it is essential that the volunteers develop a mental model that accurately reflects the scientific goals of the project. Explicit guidance to participants will help prevent problems described in this paper, including the pervasive response of 'no' for unclassifiable images (Figure 7E) and confusion between pasture and cropland (Figure 2). As new volunteers will be excited to begin tasks, the instruction period could take the same form as the game itself, with feedback following each task. These training tasks can be ordered for gradually increasing difficulty to help build volunteer confidence and avoid the frustration of immediately being faced with hard images.

Monitor volunteer performance in real time. The overall performance of individual volunteers can be a useful proxy for the quality of their individual ratings (Allahkbash et al., 2013). However, performance can vary over time, so monitoring changes in volunteer performance and providing instant feedback are essential. This information could be incorporated into scoring mechanisms, for example by adjusting the number of points awarded per task as a function of volunteer performance (Wang et al., 2013). This process is strengthened by building it on a robust basis of expert-validated images to reduce the risk that an incorrect crowd-level consensus emerges.

Encourage expression of uncertainty. Participants should be strongly encouraged to make their best judgments about the correct classification of a task, but sometime tasks are impossible. For instance, some images in the Cropland Capture game were obscured by clouds or shadows. Based on this information, it was not possible to tell if there was cropland in the area. In these situations, volunteers need some way of expressing uncertainty. While Cropland Capture is innovative in that a 'maybe' response was available, as discussed above, this option was under-used, especially for effectively impossible tasks (Figure 7E). This situation could be fixed in several ways. First, wording tweaks may encourage the use of this option. Rather than admitting uncertainty (which some volunteers may view as negative, akin to admitting failure), the 'maybe' option could be named or framed in a way that volunteers don't view as reflecting badly on themselves. The option could be called 'impossible' or a 'bad image' to encourage volunteers to view this information as useful data, rather than a failure to answer a question. Points could even be awarded for 'maybe/impossible' responses under certain circumstances, for instance when the majority of raters clearly agree on that response or expert validation has determined an image to be impossible. Raters could also be encouraged to use the 'maybe' response by showing them example images (for instance with clouds or poor resolution) that experts determined to be impossible.

Beware of loopholes that volunteers can abuse. While we argue that the 'maybe' response is often a necessary game design feature, allowing volunteers to not answer a question can come with downsides. If a game is structured so that 'maybe' responses allow volunteers to skip a task and escaping the risk of an incorrect response, then volunteers may 'game the game' by avoiding difficult images and only completing the easiest and most unambiguous tasks. Because a basic goal of crowdsourcing is to elicit answers to questions that computers struggle with, such a pattern of play would defeat the purpose of applying human intellect to confusing and ambiguous problems. This problem could be overcome in several ways. For instance, limits could be placed on the number of 'maybe' responses allowed per hundred tasks completed. These should be based on estimates of what proportion of images are truly impossible to rate, information that could be gleaned from the expert validation exercise.

The above advice is intended for citizen science games with simple, discrete, tasks. Not all of it will apply to more complex or open-ended tasks of which an individual volunteer is expected to complete relatively few during their participation, such as the 'FoldIt' game (Bohannon, 2009). However, crowdsourcing design often benefits from breaking tasks into the simplest possible components (Bernstein et al., 2010). When the game tasks are simple, fast and of variable difficulty, our recommendations are particularly relevant.

Acknowledgements

This research was supported by a IIASA postdoctoral fellowship to Carl Salk and the ERC CrowdLand (617754) and SIGMA (603719) projects.

References

Bachrach Y, Minka T, Guiver J and Graepel T 2012 How to grade a test without knowing the answers – A Bayesian graphical model for adaptive crowdsourcing and aptitude testing. *Proceedings of the 29th International Conference on machine Learning*

Battersby S E, Hodgson M E and Wang J 2012 Spatial resolution imagery requirements for identifying structure damage in a hurricane disaster: A cognitive approach. *Photogrammetric Engineering and Remote Sensing*, 78:625-635.

Bernstein M S, Little G, Miller R C, Hartmann B, Ackerman M S, Karger D R, Crowell D and Panovich K 2010 Soylent: A word processor with a crowd inside. *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, 313-322

Bianchetti R A 2014 Looking Back to Inform the Future: The Role of Cognition in Forest Disturbance Characterization from Remote Sensing Imagery. PhD Dissertation, Pennsylvania State University. 140 pp.

Bohannon J 2009 Gamers Unravel the Secret Life of Protein. *Wired Magazine* 17

Bonney R, Cooper C B, Dickinson J, Kelling S, Phillips T, Rosenberg K V and Shirk J 2009 Citizen science: A developing tool for expanding science knowledge and scientific literacy. *BioScience* 59: 977–984

Buhrmester M, Kwang T and Gosling S D 2011 Amazon's Mechanical Turk A New Source of Inexpensive, Yet High-Quality, Data? *Perspectives on Psychological Scienc*e 6: 3–5

Dawid A P and Skene A M 1979 Maximum likelihood estimation of observer error-rates using the EM algorithm. *Applied Statistics* 28: 20-28

FAO 2013 FAO Glossary. WWW document, http://faostat.fao.org/site/375/default.aspx

Fritz S, See L, van der Velde M, Nalepa R A, Perger C, Schill C, McCallum I, Schepaschenko D, Kraxner F, Cai X, Zhang X, Ortner S, Hazarika R, Cipriani A, Di Bella C, Rabia A H, Garcia A, Vakolyuk M, Singha K, Beget M E, Erasmi S, Albrecht F, Shaw B and Obersteiner M 2013 Downgrading recent estimates of land available for biofuel production. *Environmental Science and Technology* 47: 1688–1694

Goodchild M F 2007 Citizens as sensors: the world of volunteered geography. *GeoJournal* 69: 211–221

Haklay M 2013 Citizen science and volunteered geographic information: Overview and typology of participation, in: Sui D, Elwood S and Goodchild M (Eds.) *Crowdsourcing Geographic Knowledge* Springer Netherlands: 105–122.Hoffman R R 1990 Remote perceiving: A step toward a unified science of remote sensing. *Geocarto International* 2:3-13.

Howe J 2006 The rise of crowdsouring. *Wired Magazine* 14

Hunter J, Alabri A and van Ingen C 2013 Assessing the quality and trustworthiness of citizen science data. *Concurrency and Computation: Practice and Experience* 25: 454–466

Ipeirotis P G, Provost F and Wang J 2010 Quality management on Amazon Mechanical Turk. *Proceedings of the Second Human Computation Workshop*

Lintott C J, Schawinski K, Solsar A, Land K, Bamford S, Thomas D, Raddick M J, Nichol R C, Szalay A, Andreescu D, Murray P and Vandenberg J 2008 Galaxy Zoo: morphologies derived from visual inspection of galaxies. *Monthly Notices of the Royal Astronomical Society* 389:1179-1189.

from the Sloan Digital Sky Survey

Lloyd R and Hodgson M E 2002 Visual search for land use objects in aerial photographs. *Cartography and Geographic Information Science* 29:3-15.

Mekler E D, Brühlmann F, Opwis K and Tuch A N 2013 Do points, levels and leaderboards harm intrinsic motivation?: an empirical analysis of common gamification elements. *Proceedings of the First International Conference on Gameful Design, Research, and Applications* 66-73

Oreg S and Nov O 2008 Exploring motivations for contributing to open source initiatives: The roles of contribution context and personal values. *Computers in Human Behavior* 24: 2055–2073

Pourabdollah A, Morley J, Feldman S and Jackson M 2013 Towards an authoritative OpenStreetMap: conflating OSM and OS OpenData national maps' road network. *ISPRS International Journal of Geo-Information* 2: 704–728

Raddick M J, Bracey G, Gay P L, Lintott C J, Cardamone C, Murray P, Schawinski K, Szalay A S and Vandenberg J 2013 Galaxy Zoo: Motivations of citizen scientists. *Astronomy Education Review* 12: 010106

R Core Team 2014 R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Salk C F, Sturn T, See L, Fritz S and Perger C 2015 Assessing quality of volunteer crowdsourcing contributions: Lessons from the Cropland Capture game. *International Journal of Digital Earth* in press

See L, Schepaschenko D, Lesiv M, McCallum I, Fritz S, Comber A, Perger C, Schill C, Zhao Y, Maus V, Siraj M A, Albrecht F, Cipriani A, Vakolyuk M, Garcia A, Rabia A H, Singha K, Marcarini A A, Kattenborn T, Hazarika R, Schepaschenko M, van der Velde M, Kraxner F and Obersteiner M 2014a Building a hybrid land cover map with crowdsourcing and geographically weighted regression. *ISPRS Journal of Photogrammetry and Remote Sensing*, in press

See L, Sturn T, Perger C, Fritz S, McCallum I and Salk C 2014b Cropland Capture: A gaming approach to improve global land cover, in: 17th AGILE International Conference on Geographic Information Science. Castellon, Spain

Silvertown J 2009 A new dawn for citizen science. *Trends in Ecology and Evolution* 24: 467-471

von Ahn, L and Dabbish L 2004 Labeling images with a computer game. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* 319-326.

Wang J, Ipeirotis P G and Provost F 2013 Quality-based pricing for crowdsourced workers. *NYU Stern Research Working Paper* CBA 13-06

Welinder P, Branson S, Belongie S and Perona P 2010 The multidimensional wisdom of crowds. *Advances in Neural Information Processing Systems* 23

Whitehill J, Ruvolo P, Wu T, Bergsma J and Movellan J 2009 Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. *Advances in Neural Information Processing Systems* 22

Zhao Y, Gong P, Yu L, Hu L, Li X, Li C, Zhang H et al. 2014 Towards a common validation sample set for global land-cover mapping. *International Journal of Remote Sensing*, 35: 4795–4814

Figures

Figure 1. A screenshot of the Cropland Capture game as seen by volunteers.

Figure 2. Examples of image that were provided to volunteers in the Cropland Capture game. Panels A and B show scenes that include cropland which is outlined in red. Note that most of the area in panel B is not cropland. However, the correct response is still 'yes' because the question asks if there is *any* cropland present. Panels C and D are scenes with no cropland. Panel D is particularly difficult, hence the explanation included in the image. This image would have been considered impossible in the expert validation process.

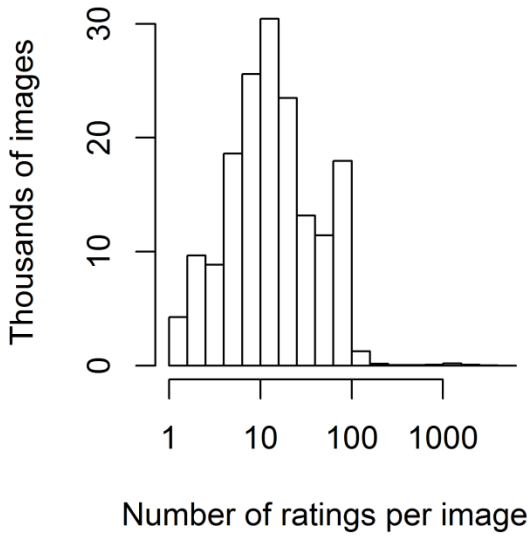

Figure 3. Histogram of number of ratings per image in the Cropland Capture game. Note the log scale on the x-axis.

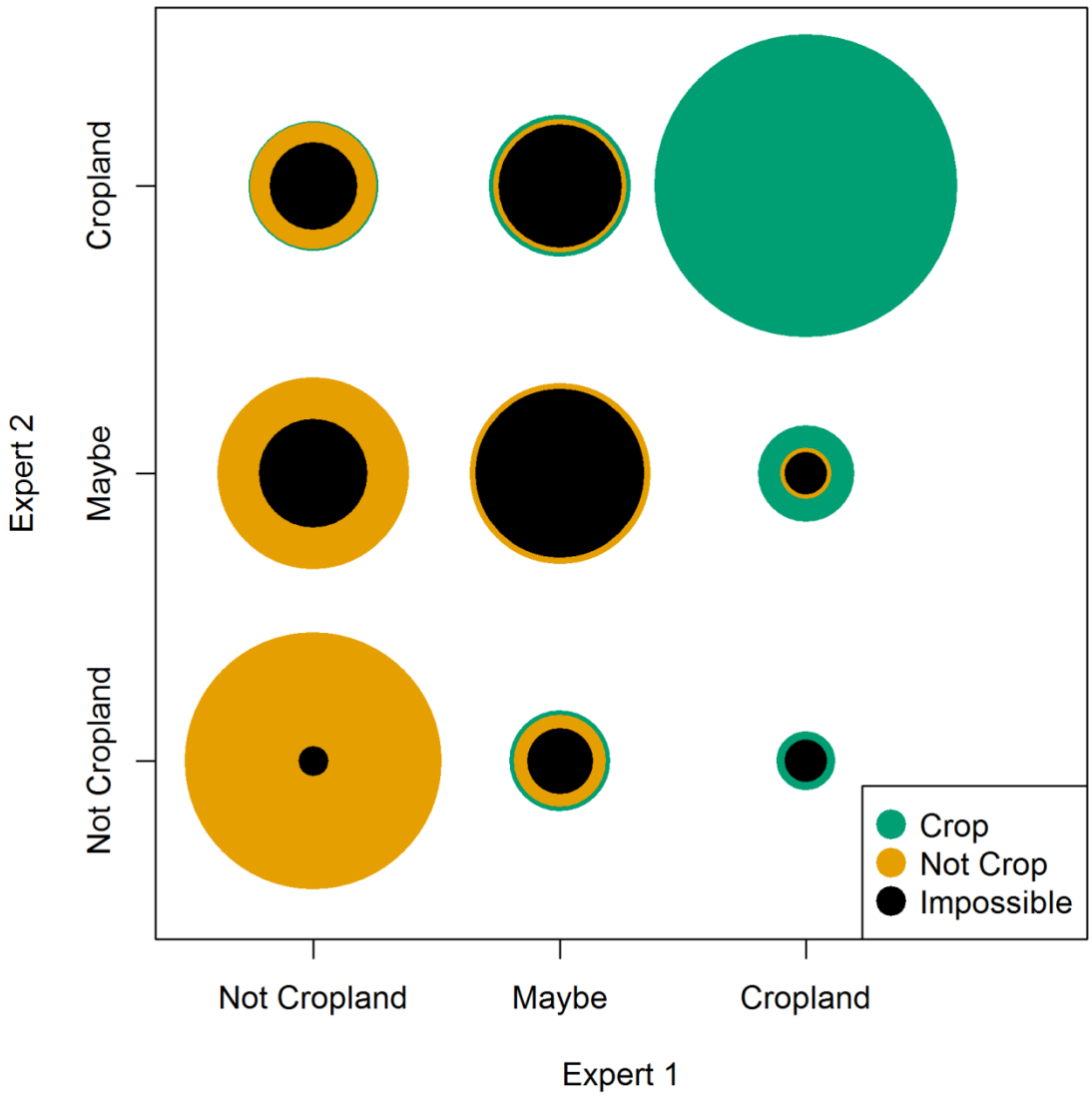

Figure 4. Agreement between the two experts' initial classifications and the final consensus classification in the expert validation exercise. The area covered by symbols is proportional to the number of images in each category. Each expert's ratings are on a separate axis, thus points on the diagonal indicate agreement between the two experts. Coloration indicates the consensus rating following the initial validation exercise. Green indicates cropland, orange means no cropland and black that the image was considered impossible to evaluate, for instance due to clouds or low resolution. Note that the images selected for expert validation in this exercise were more difficult than the average

image included in the Cropland Capture game; a more representative sample of the images would have likely resulted in much less disagreement between the two experts.

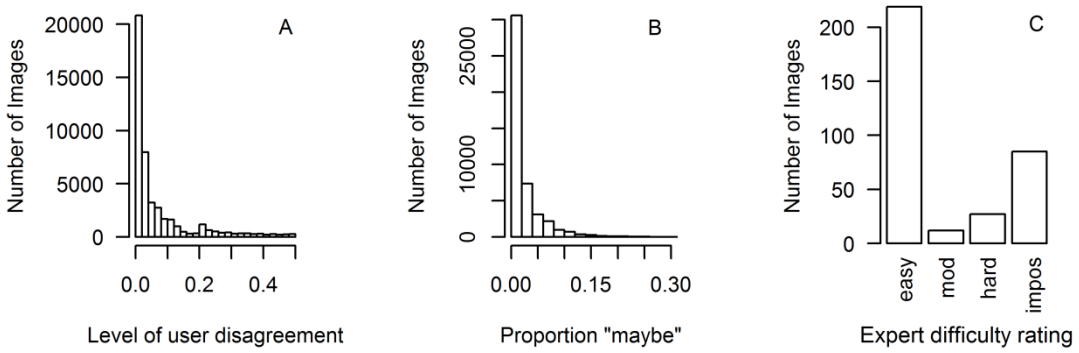

Figure 5. Three different measures of the difficulty of the image classification tasks in the Cropland Capture game. (A) Histogram of among-volunteer disagreement about the presence of cropland in all 45,991 images with >25 volunteer ratings from the Cropland Capture game. (B) Histogram of proportion of 'maybe' responses on the same set of images as in (A). (C) Bar plot of the expert-determined difficulty of 343 images; 'mod'= moderate difficulty, 'impos' = impossible to determine based on image presented. See main text for details on how these categories were determined.

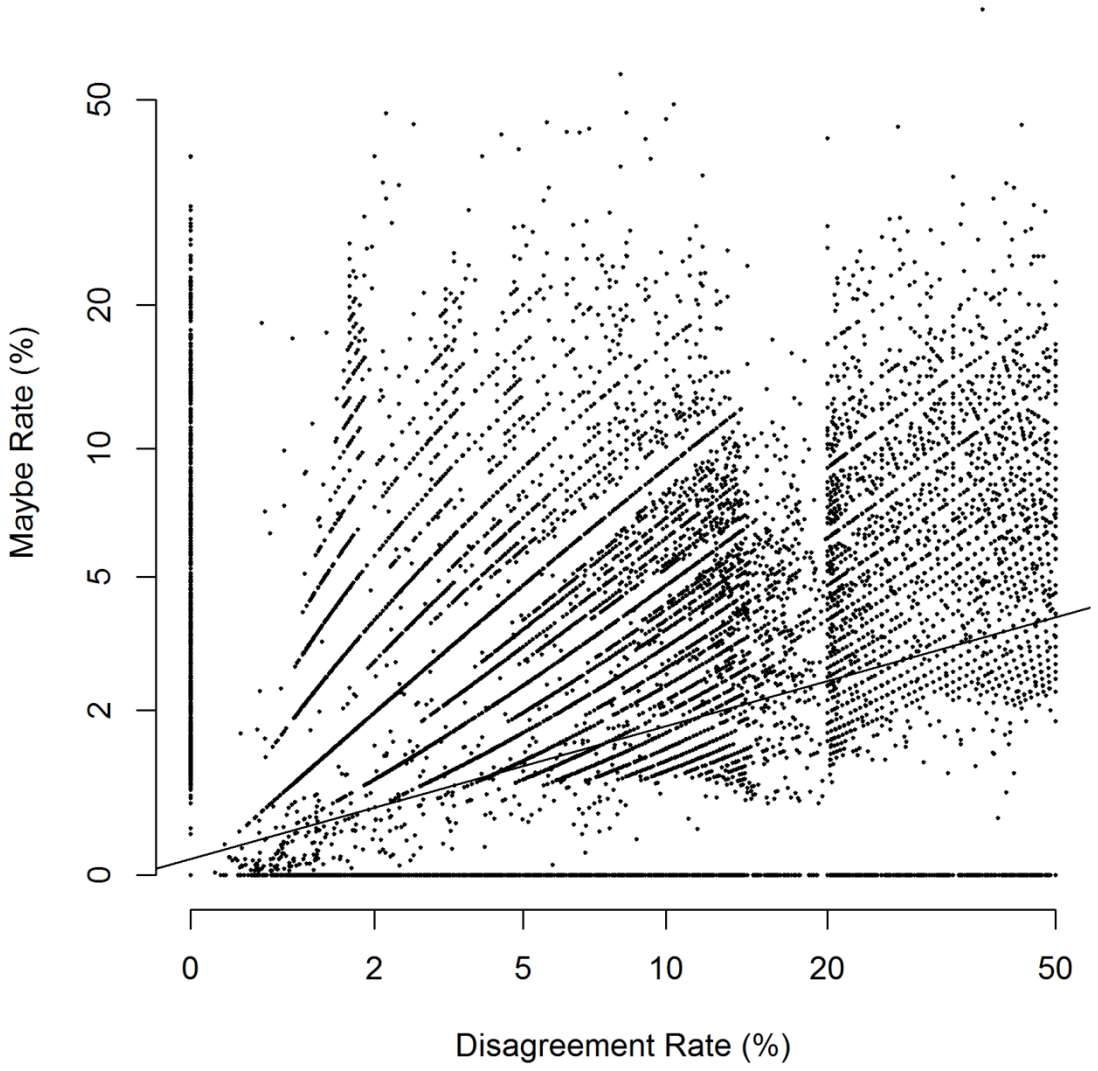

Figure 6. The relationship between the rate of disagreement with the majority vote on an image's classification and the proportion of 'maybe' ratings that the image received. Each point corresponds to a single image used in the game. The variables were log transformed (plus a small constant so that values of zero could be used in the analysis) to improve normality. While serious distributional problems remain after transformation, there was a positive relationship between these variables (see main text).

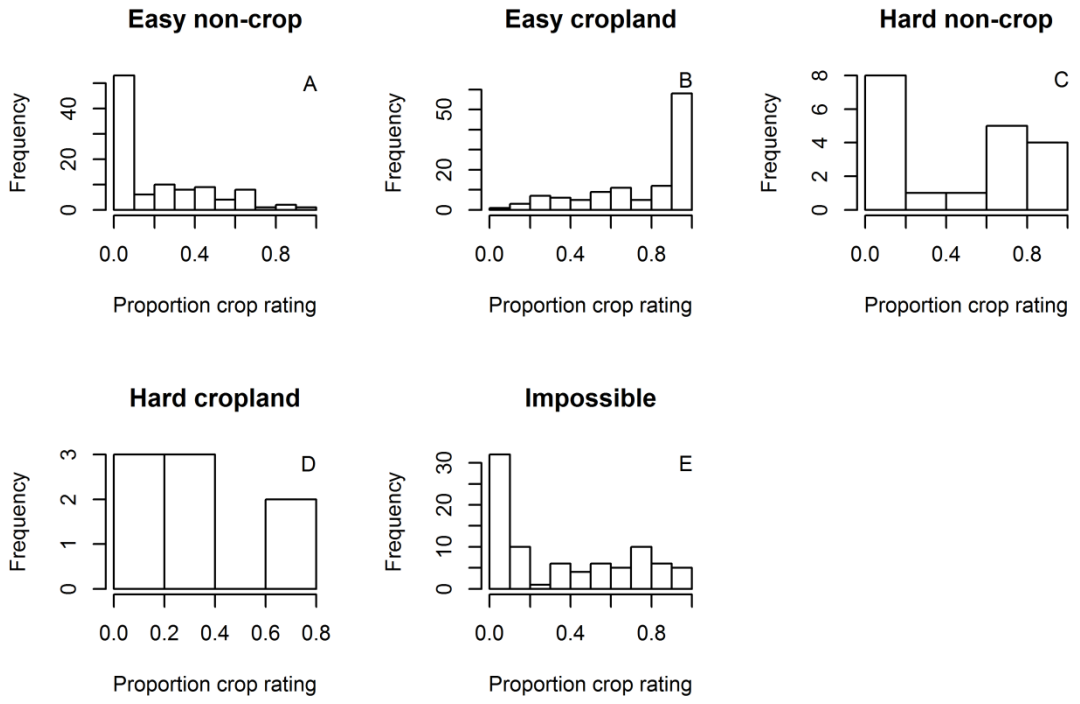

Figure 7. Distribution of volunteer classification of images as a function of expert-determined classification (cropland, no cropland, or impossible to say based on image). A proportion of 0 is equivalent to all non-maybe responses rating an image as not containing cropland; a value of 1 corresponds to all non-maybe responses claiming cropland. Note that the proportion of images in the easy, hard and impossible categories is not representative of the 165,000 images in the Cropland Capture game; these images are much more difficult than average (see main text).

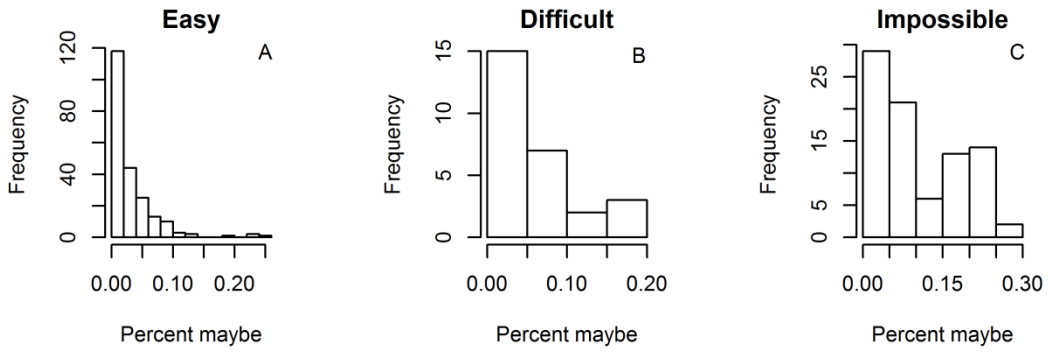

Figure 8. The rate of 'maybe' responses to images in the cropland capture game for images with expert-determined task difficulties of easy (A), moderate to difficult (B) and impossible (C). The means of these groups differ significantly from one another (see main text), with a consistent pattern of greater mean 'maybe' rating with increasing expert-rated difficulty. Note that the proportion of images in the 'easy,' 'hard' and 'impossible' categories is not representative of the 165,000 images in the Cropland Capture game; these images are much more difficult than average (see main text).

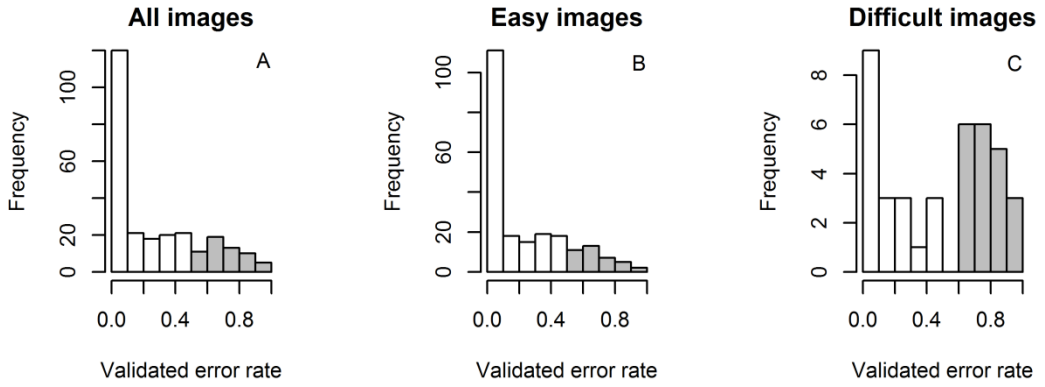

Figure 9. The rate of volunteer agreement with expert validations for different image difficulties. A) All images that experts determined a careful volunteer could reasonably be expected to rate correctly. B) The subset of the images in (A) that experts determined could easily be rated correctly. C) The subset of images in (A) that experts determined were moderately difficult or difficult, but not impossible, for volunteers to assess correctly. The unfilled bars indicate circumstances in which the majority classification was in agreement with the expert validation. The filled bars correspond to images for which the majority vote disagreed with expert classification. The mean validated error rate for the difficult images was significantly greater than for the easy images (see statistics in main text). The images included in the expert validation exercise were more difficult than the average image in the Cropland Capture game.