Making Better Use of the Crowd

Jennifer Wortman Vaughan
Microsoft Research, New York City
641 Avenue of the Americas, 7th Floor
New York, NY 10011
jenn@microsoft.com

First version: December 5, 2016
This version: April 20, 2017

Abstract

This tutorial provides a comprehensive overview of the landscape of crowdsourcing research, targeted at the machine learning community. We begin with a showcase of innovative uses of crowdsourcing, covering direct applications to machine learning systems, crowdsourcing for hybrid intelligence, and large scale studies of human behavior online. We then dig into recent research aimed at understanding who crowdworkers are, how they behave, and what this should teach us about best practices for interacting with the crowd. We discuss research on worker honesty, how to boost the quality of crowdwork using both well-designed monetary incentives and intrinsic motivation, and whether crowdworkers are really independent. Along the way, we lay out detailed tips and best practices that are rarely discussed in the literature yet crucial for ensuring that crowdsourcing-based research succeeds.

Keywords: crowdsourcing; hybrid intelligence systems; online experimentation; incentives; mechanical turk

1 Introduction

Over the last decade, crowdsourcing has been used to harness the power of human computation to solve tasks that are notoriously difficult to solve with computers alone, such as determining whether or not an image contains a tree, rating the relevance of a website, or verifying the phone number of a business.

The machine learning community was early to embrace crowdsourcing as a tool for quickly and inexpensively obtaining the vast quantities of labeled data needed to train machine learning systems. For example, in their highly influential paper, Snow et al. [64] used crowdworkers to annotate linguistic data for common natural language processing tasks such as word sense disambiguation and affect recognition. Similar ideas were applied to problems like annotating medical images [58] and discovering and labeling image attributes or features [55, 56, 81]. This simple idea—that crowds could be used to generate training data for machine learning algorithms—inspired a flurry of algorithmic work on how to best elicit and aggregate potentially noisy labels [18, 27, 31–33, 45, 63, 73, 79, 80].

In the majority of this work, it is assumed that once collected, the labeled data is handed off to a machine learning algorithm for use in training a model. This handoff is typically where the interaction with the crowd ends. The idea is that the learned model should be able to make autonomous predictions or actions. In other words, the crowd provides the data, but the ultimate goal is to eventually take humans out of the loop.

This might lead one to ask: *What other problems could the crowd solve?*

This tutorial begins with a showcase of innovative uses of crowdsourcing that go far beyond the collection of labeled data. These fall into three basic categories:

- **Direct applications to machine learning.** (Section 2) The crowd can be used to evaluate machine learning models [11], cluster data [21, 67], and debug the large and complex machine learning models used in fields like computer vision and speech recognition [49, 50, 54].
- **Hybrid intelligence systems.** (Section 3) These "human in the loop" AI systems leverage the complementary strengths of humans and machines in order to achieve more than either could achieve alone. While the study of hybrid intelligence systems is relatively new, there are already compelling examples that suggest their great potential for applications like real-time on-demand closed captioning of day-to-day conversations [38–41, 51], "communitysourced" conference planning [3, 14, 35], and crowd-powered writing and editing [5, 34, 36, 61, 69].
- **Large scale studies of human behavior online.** (Section 4) Crowdsourcing is gaining popularity among psychologists and social scientists who use platforms like Amazon Mechanical Turk to quickly and easily recruit large pools of subjects for survey-based research and behavioral experiments. Such experiments can benefit computer science too. With the rise of social computing, computer scientists can no longer ignore the effects of human behavior when reasoning about the performance of computer systems. Experiments allow us to better model things like how humans perceive security threats [70], understand numbers [4], and react to annoying advertisements [19], which leads to better designed algorithms and systems.

Viewed through another lens, we can think of these three categories of applications as illustrating the potential of crowdsourcing to influence machine learning, AI systems more broadly, and finally, all of computer

science (and even fields beyond computer science).

Section 5 of this tutorial is centered on one of the most obvious and important yet often overlooked aspects of crowdsourcing: *The crowd is made of people.* It contains a deep dive into recent research aimed at understanding who crowdworkers are, how they behave, and what this should teach us about best practices for interacting with the crowd.

Section 5.2 discusses the commonly held belief among machine learning researchers that crowdsourcing platforms are riddled with bad actors out to scam requesters. It includes the results of a research study that showed that crowdworkers on the whole are basically honest [66] and recent work suggesting that some crowdworkers misrepresent themselves when necessary to obtain work [10].

Sections 5.3 and 5.4 describe experiments that have explored how to boost the quality and quantity of crowdwork by appealing to both well-designed monetary incentives (such as performance-based payments [26, 28, 76, 77]) and intrinsic sources of motivation (such as piqued curiosity [42] or a sense of meaning [9, 59]).

Section 5.5 is a discussion of recent research—both qualitative [23] and quantitative [78]—that has opened up the black box of crowdsourcing to uncover that crowdworkers are not independent contractors, but rather a network with a rich communication structure.

Taken as a whole, this research has a lot to teach us about how to most effectively interact with the crowd. Throughout Section 5, best practices for engaging with crowdworkers are identified and called out. These best practices are rarely mentioned in the literature but make a huge difference in whether or not research studies succeed.

Crowdsourcing has the potential for major impact on the way we design, test, and evaluate machine learning and AI systems, but to unleash this potential we need more creative minds exploring novel ways to use it. This tutorial is intended to inspire the reader to find novel ways of using crowdsourcing in his or her own research, and to provide the reader with the resources needed to avoid common pitfalls.

A note on organization: This tutorial has been written in such a way that each section can be read and understood on its own; there are no major dependencies between sections. Readers who are primarily interested in tips and best practices for crowdsourcing can skip to Sections 5 and 6.

2 Direct Applications to Machine Learning

The next few sections provide a guided tour of a broad range of innovative applications of crowdsourcing that go far beyond the collection of data. We begin by describing several direct applications of crowdsourcing to machine learning. We use the term "crowdsourcing" very generally here to encompass both paid and volunteer crowdwork, done by experts or nonexperts, on any general or specialized crowdsourcing platform. Rather than committing to any specific definition, it should be interpreted in the broadest sense.

2.1 Crowdsourcing Labels and Features

While we will not focus too much attention on this area, it would be remiss to avoid mentioning the application that first got the machine learning community excited about crowdsourcing: generation of labeled data. In the basic setting, crowdworkers are presented with unlabeled data instances (such as websites or images) and are asked to supply labels (for instance, a binary label indicating whether or not the website contains profanity or a list of keywords describing the content of the image). Since the supplied labels can be noisy, the same instances may be presented to multiple crowdworkers and the workers' responses combined. There is a large body of research aimed at developing algorithmic approaches for the aggregation of these noisy labels from multiple workers [18, 27, 31–33, 45, 63, 73, 79, 80], a large portion of which builds on the expectation-maximization approach from the seminal paper of Dawid and Skene [15].

Crowdsourcing approaches have been applied to generate labeled data for natural language processing [64], computer vision [58], and many other fields. One of the behavioral studies discussed in Section 4 even used crowdsourced data labeling as a first step in a more complex experiment.

Crowdsourcing can also be used to identify and subsequently label diverse sets of salient features [81] such as attributes of images [55, 56]. The advantage of using a crowd over automated techniques is the ability to discover features that rely on knowledge and background experience unique to humans. For example, if a data set consists of images of celebrities, a human might use their background knowledge to define features such as "actor," "politician," or "married."

2.2 Crowd Evaluation of Learned Models

One application of crowdsourcing that has really taken off in some segments of the machine learning community is the use of crowdsourcing to evaluate learned models. This is especially useful for unsupervised models for which there is no objective notion of ground truth.

As an example, consider topic models. A topic model discovers thematic topics from a set of documents, for instance, New York Times articles from the past year. In this context, a *topic* is a distribution over words in a vocabulary. Every word in the vocabulary occurs in every topic, but with a different probability or weight. For example, a topic model might learn a food topic that places high weight on `cheese`, `kale`, and `bread`, or a politics topic that places high weight on `election`, `senate`, and `bill`.

Topic models are often used for data exploration and summarization. In order to be useful in these contexts, the learned model should be human-interpretable in the sense that the topics it discovers should make sense to people. However, human interpretability is hard to quantify, and as a result, topic models are often evaluated based on other criteria such as predictive power.

Chang et al. [11] proposed using crowdsourcing as a way to measure the interpretability of a set of topics. In particular, they designed a *word intrusion* task in which a crowdworker is presented with a randomly ordered list of the most common words from a topic along with one intruder word that has low weight for that topic but high weight for another topic. The worker is then asked to identify the intruder. If the topic is coherent, then picking out the intruder should be easy (think {`cheese`, `bread`, `steak`, `election`, `mushroom`, `kale`}). If not, the intruder would be harder to pick out. The average error that crowdworkers make on this task can thus be used as a proxy for how interpretable or coherent topics are. Chang et al.

found that previous measures of success like high log likelihood of held out data do not necessarily imply human interpretability.

As the idea of using crowdsourcing to evaluate topic models caught on, researchers also began thinking about how to incorporate feedback from crowds to improve the model over time. For example, Hu et al. [30] allow crowdworkers to suggest constraints capturing words that should be associated with the same topic but are not, and then use these constraints to improve the model.

Crowdsourcing has also been used, for instance, for evaluation of translations [7] and for relevance evaluation in information retrieval tasks [2].

2.3 Human Debugging of Machine Learning Models

In fields like computer vision, speech recognition, translation, and natural language processing, systems often consist of several discrete components linked together to perform a complex task. For example, consider the problem of semantic segmentation which involves partitioning an image into multiple semantically meaningful parts and labeling each part with a class. There are promising approaches to this problem that use conditional random fields (CRFs) or other machine learning models to integrate feedback from independent components that perform various scene understanding tasks like object detection, scene recognition, and segmentation.

If a system designer wants to improve performance, it is not always clear which component to focus attention on. To solve this problem, Parikh and Zitnick [54] proposed the clever idea of *human debugging*, in which humans are used to uncover bottlenecks in AI systems, and applied this idea to several problems in computer vision.[1]

Human debugging helps identify which component in a system is the "weakest link." The basic idea is simple. To quantify how much an improvement to a particular component would benefit the system as a whole, we could imagine replacing this component with something (close to) perfectly accurate and testing how much the system improves. Since for many vision and language tasks human performance is an upper bound on what we might expect from a machine, we could replace the component with a human instead.

Mottaghi et al. [49, 50] applied this idea in order to analyze the performance of a CRF that has been used in the computer vision community for scene understanding. They replaced each component with crowdworkers from the popular crowdsourcing platform Amazon Mechanical Turk and measured the change in performance of both the component in isolation and the system as a whole.

One of their most interesting findings was that humans are actually less accurate than machines at one particular task (classifying super-pixels) yet when human classifications were plugged into the CRF, the system performance improved. This suggests that making fewer mistakes classifying super-pixels is not enough. Rather it is more important that the classifier makes the right kind of mistakes. This kind of feedback helps designers know where to focus their effort.

Recently, Nushi et al. [52] took this idea one step further, allowing crowdworkers to propose targeted fixes to the machine components of a larger system and then evaluating the effect of various component fixes on the overall system performance.

[1] For more details on applications of crowdsourcing to computer vision, see the recent survey of Kovashka et al. [37].

2.4 Crowdsourcing Similarity Functions and Crowd Clustering

We next discuss several examples of recent work that showed how crowdsourcing can be applied to solve unsupervised learning tasks like estimating some notion of object similarity or clustering objects.

Similarity matrices, which assign similarity scores to pairs of objects, are useful for exploratory data analysis, data visualization, clustering, or classification using kernel-based approaches such as support vector machines. There are automated techniques for discovering similarities, but these can fail to uncover similarities that rely on specific semantic knowledge or experience that is unique to humans. Returning to the previous example of a celebrity image data set, a human might consider background knowledge about a celebrity's profession or home country when determining how similar two celebrities are.

Tamuz et al. [67] considered the problem of estimating a similarity matrix over all pairs of n objects from human judgments. They proposed an adaptive algorithm based on comparisons of triples which is able to learn a similarity matrix with a relatively small number of human judgments. Using their approach, they were able to answer questions like which necktie would be a good substitute for another, a task that would perhaps be difficult for a machine without specialized human knowledge.

Around the same time, Gomes et al. [21] suggested an approach to the problem of crowd-based clustering that involves presenting relatively small sets of objects to each member of the crowd and asking for a clustering of these objects. These partial clusterings are then used to train a Bayesian model with the goal of producing a full clustering over all objects.

For these applications, providing good instructions and clear guidance to workers is key. If the discovered features or clusters are intended to be used towards a certain goal, communicating this goal to the crowd can help them to identify the most salient aspects of the data on which to focus.

We will return to the idea of crowdsourced clustering later when we discuss hybrid intelligence systems for constrained optimization in Section 3.2.

3 Hybrid Intelligence Systems

A *hybrid intelligence system* is a "human in the loop" AI system made up of both human components and machine components. These systems are designed to leverage the complementary strengths of humans and machines with the hope of accomplishing more than would be possible using humans or machines alone. We'll next describe a couple of particularly compelling hybrid intelligence systems.

3.1 Hybrid Intelligence for Speech Recognition

We begin with a very creative hybrid intelligence system built by Lasecki, Bigham, and colleagues [38–41, 51]. This system addresses the problem of closed captioning. Closed captioning is something that can be done reasonably well using existing speech recognition techniques under ideal circumstances, for example, when the voice recording is high quality and the system has been trained on data from the particular speaker. It does not work as well on low quality audio, with novel speakers, with speakers with heavy accents, or with language that contains a lot of technical jargon. For these scenarios, the best results come from hiring

a professional stenographer, but high quality stenographers can charge as much as $200-$300 an hour and are not available on demand.

The question that these researchers asked is whether it would be possible to provide less expensive, real-time, on-demand closed captioning to users who need help understanding lectures, meetings, or other day-to-day conversations.

To achieve this, Lasecki et al. built a hybrid intelligence system that makes use of cutting edge techniques from natural language processing and crowdsourcing. Here is the basic idea. When a user would like to obtain closed captions, he starts recording. The audio is sent simultaneously to several crowdworkers. These workers aren't expected to be able to fully transcribe the speech, which is way too fast to transcribe on a normal keyboard. Instead, each worker transcribes sentence fragments. The system adjusts the speed and volume of the speech to focus each worker's attention on distinct overlapping components. The system then uses language models and AI techniques to combine the workers' text into one complete and coherent stream that is delivered back to the user with a delay of only a couple seconds.

If the crowdworkers are domain experts—for example, work study students generating closed captions for a technical math or science class—the system would be capable of capturing all of the jargon that is hard to get right using traditional automated closed captioning systems.

This system is truly able to take advantage of the complementary strengths of people and machines to achieve high quality results. It is also an example of a broader phenomenon of using crowdsourcing to compensate for poor AI in the short term. Maybe one day the AI will be good enough for personalized closed captioning with low quality audio, but using the crowd allows us to achieve this faster.

3.2 Hybrid Intelligence for Constrained Optimization

Cobi[2] [3, 14, 35] is a hybrid intelligence system for conference planning. Cobi is based on the idea of *communitysourcing*: it draws on the specialized expertise of people within a research community to plan out conference schedules.

The team that developed Cobi is part of the CHI community. CHI is a huge conference. In 2013, the year that Cobi was deployed, it accepted around 400 papers which were to be presented in 16 parallel tracks. Scheduling talks for this many papers is a huge feat. The organizers must minimize conflicts between sessions so that participants can see all of the talks they're interested in. Essentially they want to solve one big constrained optimization problem, but without direct access to the constraints—in this case, the overlap of interest in different papers and attendees' preferences more generally.

Cobi was designed to efficiently collect this information from the community. It includes individual components that elicit preferences, hard and soft constraints, and similarity judgments. Using this information, it provides an interface backed by optimization tools that chairs can use to fine tune the schedule to satisfy as many constraints as possible and produce something coherent.

As one example of how the crowd is used, authors are presented with lists of papers that are potentially related to theirs and asked for judgments about which papers belong in the same session and which papers should not be scheduled at the same time. Authors are motivated to provide this information because it is in

[2] http://projectcobi.com

their own best interest for their talk to land in a coherent session without any major timing conflicts. When Cobi was used at CHI, the authors of 87% of accepted submissions opted to provide feedback.

Where do the lists of potentially related papers come from? This is where the crowdsourced similarity judgments come in. In fact, the designers of Cobi treat this as a clustering problem and solve it using crowd clustering techniques like the ones we discussed in Section 2.4.

It is important to note that the conference chairs always maintain control, so in addition to the constraints coming from the community, they are able to use their own specialized knowledge to reason about tradeoffs between constraints and make sure the resulting schedule is sensible on the whole.

3.3 Hybrid Intelligence for Writing

The next application we'll consider is writing [61, 69]. Over the last few years, multiple hybrid intelligence systems have been introduced that aim to utilize crowd workers to speed up and improve the quality of writing. These include Soylent [5], Crowdforge [36], and Mechanical Novel [34], among others.

Before getting into the details of how a hybrid intelligence system for writing might work, let's step back and take a moment to think about the process of writing. In a series of recent papers, Teevan et al. [68, 69] have been advocating a technique they call "selfsourcing". Selfsourcing takes one of the primary principles of crowdsourcing—breaking a large task into many bitesize microtasks, each of which can be completed in isolation—and applies it to our every-day attempts to get work done. Teevan et al. [69] argue that writing is most effectively done as a three-step process:

1. **Collect content.** Often the most difficult part of writing is getting started. Most of us have had the experience of staring at a blank screen trying to figure out where to begin. The writing process is arguably much easier if we start off with a more manageable and less intimidating task, such as generating short bursts of content. This content can be on the level of individual ideas that we'd like to communicate, and we can generate these ideas over time whenever they come to us.
2. **Organize content.** Once these ideas have been generated and we're no longer staring at a blank screen, we can try to organize this content. This can involve first clustering the content by theme and then sorting the clusters in a logical order.
3. **Turn content into writing.** Next, we can begin the actual writing by turning each cluster of ideas into a coherent paragraph or section. Once this has been done, we find ourselves with a full draft, which we can edit, polish, and finalize as usual.

So far we've been talking about a single author completing each of these steps in isolation, but this doesn't have to be the case. The same process could easily be completed by a set of collaborators who independently generate initial ideas, organize content, and turn clusters of content into writing.

Taking it one step further, some of these steps—organizing content, turning content into initial paragraphs of text, perhaps some of the editing that follows—do not need to be completed by someone who is an expert on the topic. It is easy to imagine that these could be completed by crowdworkers, traditional machine learning or AI approaches, or some combination of the two. It's not important that these steps are done perfectly,

because the author will still be involved in the final editing pass. This is the type of hybrid intelligence system that researchers envision [61, 68, 69].

3.4 Hybrid Intelligence for Information Aggregation

We next discuss combinatorial prediction markets [1, 25]. A prediction market is a financial market in which traders can buy or sell securities with payoffs that are linked to future events. For example, in an election market, traders might buy or sell a security that is worth \$1 if a particular candidate wins and nothing otherwise. If a trader believes that the probability of this candidate winning is p and wants to maximize her expected payoff, then she should be willing to buy this security at any price less than $\$p$, since with probability p she would get \$1. Similarly, she should be willing to sell at any price greater than $\$p$. For this reason, we can think of the current market price of this security as capturing traders' collective beliefs about how likely it is that the candidate will win the election.

How is a prediction market an example of hybrid intelligence? Well, consider a prediction market for a United States Presidential election. What if we wanted to allow traders to express more complicated information than just the probability that a particular candidate will win the general election? To achieve the best possible aggregation of information, we might like to allow traders to express more specialized information such as the probability that a Democrat wins in North Carolina or the probability that a Republican wins in at least one of Ohio and Pennsylvania.

There are several challenges that arise. First, if we allow traders to construct arbitrary securities, there will be a liquidity problem. If a trader has highly specialized information, it will be difficult for her to find a counterparty to trade with. This problem can be solved by implementing the market using an automated market maker, a centralized algorithmic agent that is always willing to buy and sell any security at some price which typically depends on the history of trade [25]. However, with large and complex security space, it is computationally hard for a market maker to maintain coherent prices and keep its loss bounded [12]. Using techniques from convex optimization, we can design market makers that generate coherent prices (and therefore coherent predictions) over large outcome spaces while maintaining bounded monetary loss for the market maker [1]. This gives us an algorithmic way to aggregate even the most specialized beliefs and information of a crowd.

3.5 Hybrid Intelligence Systems in Industry

So far we have focused on hybrid intelligence systems that have come out of the research community, but it's worth mentioning that human-in-the-loop systems are widely used in industry as well. To name just a few examples, Stitch Fix, which provides personalized style recommendations, trains machine learning algorithms to suggest items that a user might like and then sends the output of these algorithms to a human expert to curate and pare down.[3] Twitter employs contract workers to interpret search terms that suddenly spike in meaning, often due to recent mentions in the media or pop culture that an algorithm trained on stale data would miss.[4] PatternEx uses machine learning to identify suspicious activity that could indicate a security threat. This suspicious activity is then examined by a human, who determines whether there is a

[3] http://multithreaded.stitchfix.com/blog/2016/03/29/hcomp1/
[4] http://nyti.ms/2gzOo4K

real attack, and the human's feedback is used to improve the system.[5]

4 Large Scale Studies of Human Behavior Online

Let's zoom out one more time and consider how crowdsourcing can benefit the larger computer science community.

Over the last few years, psychologists and social scientists have increasingly turned to crowdsourcing platforms to recruit subjects to complete surveys and participate in behavioral experiments that traditionally would have been run in labs on campus. Crowdsourcing provides easy access to large and diverse pools of subjects and studies have shown that classic results from psychology and behavioral economics can be replicated by these crowdworkers [29, 53]. Crowdsourcing also allows faster iteration between the development of new theories and experimentation, allowing researchers to speed up their overall research process [47].

While it is not the focus of this tutorial, crowdsourcing allows researchers to push the boundaries of experiment design by recruiting larger pools of subjects and interacting with subjects over longer time horizons than would be possible in a traditional lab setting. (See, for example, the recent work of Mao et al. [46] who examined what happened when 94 subjects each played 400 ten-round games of Prisoners Dilemma over the course of a month, an order of magnitude more than prior work.)

At the same time, computer scientists have begun to take more interest in conducting behavioral experiments and survey research of their own. With the increasing prevalence of social computing and other systems that depend on large scale human participation, it is not a good idea for computer scientists to ignore human behavior [13]. By running experiments to develop better models of human behavior, we are able to design better algorithms and better social computing systems.

This section contains examples of ways in which crowdsourcing has been used to conduct different types of user studies and human subject experiments online.

4.1 Human Behavior and Security

We begin with a simple example of a survey-based study.

It is widely known that Internet users have a tendency to choose overly predictable passwords. However, there has been relatively little research aimed at understanding how well users understand the security or predictability of the passwords they choose.

Ur et al. [70] used Mechanical Turk to conduct a study on user perception of password security. They surveyed crowdworkers to find out, for example, how these workers perceived the security of different password generation strategies and what each worker's mental model of an attacker is.

As an example, in one component of their study, they showed each worker pairs of passwords and asked the worker which was more secure. To evaluate these assessments, they compared them with a gold standard obtained by looking at the number of guesses that it would take for a modern password-cracking program to

[5] http://tek.io/1Vy01KB

guess each password and calling one password more secure than another if the number of guesses required to crack it is more than an order of magnitude larger. They observed that many of their participants overestimated the benefit of replacing letters with numbers or symbols, incorrectly rating `p@ssw0rd` as more secure than `pAsswOrd`. They also underestimated the risk of including common keyboard patterns (e.g., `qwertyuiop`).

4.2 Human Behavior and the Communication of Numbers

The news is filled with numbers that are notoriously difficult to understand. Is a one hundred billion dollar cut to the United States federal budget large or small? It certainly sounds like a lot of money, but how does it compare with overall federal spending? To reason about the impact of such a budget cut, it may help to know that one hundred billion dollars is roughly 3% of the 2015 federal budget, or about a sixth of the amount that the United States spends annually on the military. It's also about 30% of the net worth of Beyoncé or about $5 for every person in New York State.

Barrio et al. [4] used crowdsourcing to test whether people's understanding of numbers could be improved through the use of such perspectives.

The first step was to generate useful perspectives. To do this, they extracted 64 quotes containing measurements from recent New York Times front page articles. They asked crowdworkers on Amazon Mechanical Turk to generate perspectives for these quotes using a specialized template. To determine which perspectives were the most useful, they had other crowdworkers rate their usefulness. Finally they extracted the most useful perspectives according to these ratings. Here are a few examples:

> *The Ohio National Guard brought 33,000 gallons of drinking water to the region, while volunteers handed out bottled water at distribution centers set up at local high schools.*
>
> *To put this into perspective, 33,000 gallons of water is about equal to the amount of water it takes to fill 2 average swimming pools.*

> *They also recommended safety programs for the nations gun owners; Americans own almost 300 million firearms.*
>
> *To put this into perspective, 300 million firearms is about 1 firearm for every person in the United States.*

Barrio et al. [4] used these top rated perspectives in a sequence of experiments designed to test whether perspectives increased three different proxies of numerical comprehension: recall, estimation, and error detection. For example, in the recall experiments, crowdworkers were randomly assigned to view news quotes either with or without an added perspective. After viewing the quotes, they were distracted with a quick game of Tetris, and then asked to recall the numbers from the quotes that they had viewed.

They authors found support for the benefits of perspectives across all experiments. As one example, 55% of participants who viewed the corresponding perspective were able to recall the number of firearms in the United States, while only 40% of those who viewed the quote without the perspective were able to do so.

This project is a user study, but it also hints at the possibility of building a hybrid intelligence system that could automatically extract numerical values from news articles and use the crowd to generate the most salient perspectives to display with each one in order to strengthen readers' comprehension.

4.3 Human Behavior and Online Advertising

The last example we showcase is a particularly creative project by Goldstein et al. [19, 20] who used crowdsourcing to try to answer a high-impact business question. Publishers on the Internet display banner ads on their websites to generate profit. Some of these ads are clearly more annoying than others. What Goldstein et al. asked is whether displaying these annoying ads is costing publishers money, and if so, whether it is possible to quantify how much money it is costing.

Their experiment involved two steps, both of which used crowdsourcing.

The goal of the first step was to identify some examples of good and bad ads. To do this, they presented workers on Mechanical Turk with ads and had the crowd rate how annoying each ad was. This is basically just a data labeling task and is a fairly standard use of Mechanical Turk. Aggregating these ratings across crowdworkers, they produced lists of the most annoying and least annoying banner ads.

In the second step, they performed an experiment aimed at estimating how much (monetary) value or utility web users get from *not* being exposed to these annoying ads. They gave crowdworkers what looks like a normal crowdsourcing task, labeling email from the Enron email database as either spam or not spam. Next to each email, they showed either no ad, an ad that was determined to be good in step 1, or an ad that was determined to be annoying. They also varied how much they paid users for labeling each email. Workers could label as many emails as they wanted and would see a new ad each time, but always an ad of the same type. By looking at the number of emails that workers chose to classify in each treatment, Goldstein et al. were able to estimate how much more money it would be necessary to pay workers to get them to perform the same number of classification tasks when exposed to bad ads as opposed to good ads or no ads at all. From that, they were able to estimate how much publishers lose by showing bad ads.

Without getting into detail on the methods used to generate these numbers, there are a few key takeaways. First, displaying good ads doesn't hurt publishers too much, in the sense that workers only chose to classify slightly fewer emails with good ads compared to no ads at all.

However, displaying annoying ads hurts a lot. Goldstein et al. estimated that they would have to pay about $1 extra to generate 1000 views using a bad ad compared with a good ad or no ad, which is extremely expensive for banner ads. In other words, publishers are likely losing money by displaying annoying ads unless they are charging these advertisers significantly more.

5 The Human Side of Crowdsourcing

When incorporating crowdsourcing into their research, it is common for researchers to treat the nebulous "crowd" as a black box, ignoring the fact that this crowd is made up of real people with their own life experiences, preferences, and relationships. In this section, we provide an overview of research aimed at understanding who crowdworkers are and how they behave. We argue that this research has important

implications for the practice of crowdsourcing and distill these implications into concrete recommendations of best practices to follow.

Understanding the crowd can help anyone who wants to use crowdsourcing do so more effectively, whether the goal is to generate training data, evaluate machine learning models, debug machine learning systems, or learn about human behavior. It can help answer questions like how much to pay for a task and how to set up the right payment structure. It can teach us how and why to communicate effectively, how to attract more people to a task, how much to worry about spam, and how to avoid common pitfalls. As we'll see, it also has implications about the independence of crowdworkers' responses, which are especially critical to understand when using crowdsourcing to study human behavior.

Gaining a basic understanding of the crowd is also a necessary precondition for the goal of one day being able to formally reason about the performance of hybrid intelligence systems in the way that traditional computer science techniques allow us to reason about the performance of computer systems run on machine alone. Suppose that we would like to analyze the running time, scalability, or correctness of a computer system augmented with human components, or predict the impact of design decisions when there are humans in the loop. To do this, we need to have in mind a model of how the humans in the system are likely to behave. Are they mostly honest or will they cheat you if they can? Do they respond rationally to financial incentives? If we make unrealistic assumptions about the people who make up this system, then the system might not behave as we intended, so it's crucial to base our models on how people actually behave [13].

While many of the takeaways and lessons learned apply equally well to other crowdsourcing platforms, most of the research in this section looks specifically at workers on one particular platform, Amazon Mechanical Turk. For a brief discussion of alternative platforms, see Appendix A. Amazon Mechanical Turk is a platform for crowdsourcing *microtasks* that has become popular within the research community. On Mechanical Turk, task requesters post small-scale *human intelligence tasks* (referred to within the Turk community as "HITs") along with the amount of money that they are willing to pay. Workers can then browse the set of tasks available and choose the tasks they want to work on.

5.1 Crowdworker Demographics

Over the years there have been several studies published which examine the demographics of workers on Mechanical Turk. We mention only a few key statistics that help paint a picture of the worker pool. These come from MTurk Tracker,[6] a project aimed at tracking the demographics of Mechanical Turk over time by continually releasing tasks containing demographic surveys on Mechanical Turk to obtain up-to-date information about workers [16]. While there are some potential problems with this approach (for instance, not all workers on Mechanical Turk choose to do surveys, so this is perhaps a better reflection of the population of workers who do survey work), the results are more or less in line with other studies.

According to the MTurk Tracker data:

- About 70-80% of Mechanical Turk workers are from the United States, while about 10-20% are from India, but the breakdown of workers varies significantly throughout the day. The prevalence of workers from the U.S. and India makes sense because Mechanical Turk offers payment only in U.S. dollars,

[6] http://www.behind-the-enemy-lines.com/2015/04/demographics-of-mechanical-turk-now.html

Indian rupees, or Amazon credit.

- The breakdown between male and female workers is fairly close to even, though it varies a bit by country.
- For crowdworkers in the U.S., the (self-reported) median household income is in the range of $40K-$60K, which is in line with the median U.S. household income. The median for Indian workers is less than $15K, with many Indian workers reporting a household income of less than $10K per year.

There is evidence that other crowdsourcing platforms, such as CrowdFlower and Prolific Academic (see Appendix A), attract more European workers and lower income workers than Mechanical Turk [57].

Goodman and Paolacci [22] provide a nice overview of the similarities and differences between the population of Mechanical Turk workers and the US population as a whole, as well as the populations traditionally used for consumer studies.

It is important to note that the demographics of available workers vary widely based on the time of day and, to a lesser extent, day of the week. Through a large study of intertemporal demographic differences on Mechanical Turk, Casey et al. [8] found that, even restricting attention to workers from the U.S., the demographics of available workers change dramatically over the course of a day. For example, they found that workers who completed their task at night were more likely to be single than those who completed it in the morning, and more likely to be completing the task on a smartphone.

Takeaways and Related Best Practices:

- The population of Mechanical Turk workers is not a representative sample of the population at large, but can be used to build representative panels when needed.
- Requesters should consider the influence of timing on the population of workers who complete their tasks. The pool of workers available in the morning may be very different than the pool available in the evening. This is especially important when running experiments with multiple treatments. Ideally workers should be assigned to treatments randomly upon accepting the task so that timing has no effect on the assignment.

5.2 Dishonesty Among Workers

Papers on label aggregation often begin with a discussion of the prevalence of noise in crowdsourced labels. In particular, they often mention the common notion that crowdsourcing platforms are riddled with spammers who try to cheat the system to make money. Of course spammers and bots exist on crowdsourcing platforms, as they do all throughout the Internet, but some evidence suggests that they are not as widespread as one might think and that the majority of crowdworkers are trying to do good work.

Suri et al. [66] ran a study to test how honest the population of crowdworkers on Mechanical Turk are. Their study used a trick from the behavioral economics literature that allowed them to measure how trustworthy workers are on the whole even without being able to detect individual lies.

The idea is simple. Each worker was asked to roll a die (or simulate rolling a die on an external website) and report the value of her roll, a random number between 1 and 6. For completing this task, the worker

received a base payment of $0.25 plus a bonus of $0.25 times her roll. For example, if a worker reported rolling a 4, she would receive a total payment of $1.25. Total payments were therefore between $0.50 and $1.75. Workers knew that there was no way for the requester to verify the outcomes of their rolls.

If all workers followed the instructions and honestly reported their rolls, the mean of the rolls reported would be close to 3.5. What Suri et al. [66] observed was not so far off from this. The mean of the rolls reported by the 175 participants who took part in the study was 3.91. On the whole, workers had a tendency to overreport rolls of 5 and 6 and underreport rolls of 1 and 2, but this misreporting was far from universal, even in this extreme case in which workers had a risk-free way to directly benefit from being dishonest.

Suri et al. [66] also tested whether increasing the ability of the requester to detect cheating would lead to higher levels of honesty. They conducted a variant of the die rolling experiment in which each worker was asked to report the results of thirty independent die rolls. This time, each worker received a base payment of $0.25 plus the sum of their reported die rolls, for a total payment between $0.55 and $2.05.

This time around, the mean reported role of the 232 participants was 3.57—still statistically significantly higher than the expected mean of 3.5, but significantly closer too. Examining the individual behavior of the subjects, only 3 of the 232 were found to have an average report significantly greater than we'd expect if they were honest, and only 1 chose to maximally game the system by always reporting 6.

While these results suggest that workers are largely honest, recent studies have shown that some workers are willing to deceive requesters when this deception is necessary to gain access to additional work. In particular, recent contemporaneous studies by Chandler and Paolacci [10] and Wessling et al. [74] showed that a substantial number of workers misrepresent themselves in order to satisfy eligibility criteria of studies. For example, in one study of over two thousand workers, Chandler and Paolacci [10] asked workers to report whether they were the parent or guardian of a child with autism. In one treatment, workers were told this was part of a prescreening test for a subsequent study, while in another (the control) they were not. 7.8% of workers in the prescreening treatment reported being the parent or guardian of a child with autism, compared with 4.3% in the control, suggesting that approximately 3.5% of workers in the prescreening treatment provident a fraudulent response. While 3.5% of workers may seem small, note that due to the low number of workers who would truly qualify for the follow-up study, this prescreening step would lead to a high prevalence of impostors (45%) in a follow-up.

In another study, Chandler and Paolacci [10] found that when payments were sufficiently high, 16% of participants made a second attempt to pass a prescreening survey, identifying themselves as a different gender the second time around after initially being blocked from a study.

Takeaways and Related Best Practices:

- Most workers are honest most of the time. Most of them are trying to do a good job.
- Some workers will deceive requesters to gain access to work. If prescreening is necessary for a study, it should be done with care, ideally in advance as part of a separate stand-alone task. Workers should be prevented from attempting screens more than once.[7]
- Use care to avoid attacks. Even if most workers are honest, this is the Internet and scammers do exist.[8]

[7] For a lengthier discussion of best practices for screening participants, see Chandler and Paolacci [10].

[8] Despite attempts to carefully build a manipulation-proof system, the author of this paper and her collaborators once ended up

5.3 Monetary Incentives

We next turn our attention to the question of what motivates crowdworkers, and how both monetary incentives and intrinsic motivation can be used to improve the quantity and quality of crowdwork.

When researchers first started incorporating crowdsourcing into their work, a big part of the appeal was access to inexpensive data. Over time, the general viewpoint on this has shifted a bit. More emphasis has been placed on the ethical considerations of online labor markets and the importance of paying a fair wage. Many crowdworkers rely on the money they earn on platforms like Mechanical Turk to make ends meet. While crowdworkers are considered contractors and therefore are not covered by minimum wage laws, paying at least minimum wage is the decent thing to do. It is good for the requester too since it helps maintain better relationships with workers, a point we will return to later.

An effective and widely used method of setting payments is to follow this rule of thumb:[9] Ask your colleagues or students to complete your task or give your task to a small number of crowdworkers in order to calculate an estimate of how long it takes to complete. Use that estimate to make sure that workers receive the equivalent of the United States minimum wage (or higher).

Beyond the moral argument for paying well, it is natural to ask whether paying more improves the quality of crowdwork. There have been several studies examining the effects of payment on quality of work [6, 26, 44, 48, 59, 62, 75–77]. We describe here the results of one particular study aimed at determining the effect of *performance-based payments* [28]. We choose to focus on this study for two reasons. First, it is comprehensive, offering potential explanations for disparities in previous results. Second, it serves as a nice illustrative example of a behavioral experiment run on Mechanical Turk and allows us to touch on issues that arise, such as when randomization should take place.

Performance-based payments are payments that reward crowdworkers for higher quality work. Most commonly, a worker is offered some base payment (in the examples below, this will be $0.50, which was chosen using the rule of thumb above) just for completing a task with the opportunity to earn a bonus payment (say, an additional $1) for submitting work that the requester judges to be good. These types of bonuses are fairly common on platforms like Mechanical Turk.

Prior to the work of Ho et al. [28], the literature on the effects of payments in crowdsourcing markets was inconclusive and in some cases appeared contradictory. Several papers argued that paying higher amounts increases the quantity of crowdsourced work, but not the quality [6, 44, 48, 59]. (We will return to this shortly.) There was work showing that performance-based payments improve quality [26, 77] and other work showing that performance-based payments *do not* improve quality [62]. Finally there was work suggesting, perhaps surprisingly, that when performance-based payments are used, the quality of work does not depend on the size of bonus payments [76].

To make sense of and expand on these results, Ho et al. [28] set out to run a sequence of experiments with the goal of uncovering when, why, and on which tasks performance-based payments lead to higher quality. Their first experiment was a warm-up experiment to verify that performance-based payments can indeed lead to higher-quality crowdwork on some task. At the same time, they wanted to test the hypothesis that

paying a particularly devious worker more than six hundred times after this worker devised a complicated trick to exploit a bug in order to submit work more than once. Mason and Suri [47] mention a similar incident.

[9] The author first learned of this technique via personal communication with Sid Suri (2014) and has heard it from several other researchers since.

even when a requester is not explicitly offering a bonus for high quality work, there may be a kind of *implicit* performance-based payment effect on workers. In particular, on Mechanical Turk, work is not automatically accepted. While this doesn't often occur, a requester always has the opportunity to reject poor quality work and refuse payment. If a worker thinks her work is likely to be rejected if it is not high enough quality, then she may act as if payments are performance-based even if they're not.

For this experiment, Ho et al. [28] posted a proofreading task in which they showed workers a block of text in which they had inserted a total of 20 common English-language typos. The block of text was an image so workers couldn't simply copy it and run it through a spellcheck. They chose this task because it had a few useful properties. First, quality was easily measurable since they inserted the typos themselves. Second, they suspected that if workers exerted more effort it would lead to better results, which might make performance-based payments more effective.

They offered a base payment of $0.50 and a bonus of $1. They considered three bonus treatments: one in which there was no bonus and no mention of a bonus, one in which all workers automatically received the bonus just for accepting the task, and one in which workers received the bonus only if they found at least 75% of the typos found by other workers. This type of relative threshold was chosen because it is something that could be implemented in practice without needing to know the total number of typos that could be found. Additionally, the authors did not want to make it obvious to workers that they had inserted the typos into the text themselves and therefore knew how many there were. They chose an automatic bonus instead of a larger base payment so that they could post the task only once with a fixed base of $0.50 and randomly assign treatments after workers accepted the task.

To test the implicit performance-based pay hypothesis, the authors also considered two base treatments: one in which it was explicitly guaranteed that the work would be accepted as long as at least one typo was found, and a treatment in which no such guarantee was mentioned.

The results have a few takeaways. First, as one might perhaps expect, guaranteeing payment hurts. In other words, the implicit performance-based payment effect is real. Second, performance-based payments do indeed improve performance on this task. Finally, on this task, simply paying more (that is, giving a bonus independent of quality, just for accepting the task) also improves the quality of work. This is somewhat surprising at it contradicts what was observed in prior experiments. However, while it led to a similar improvement in quality, giving unconditional bonuses costs the requester a lot more than using performance-based pay since the bonuses are awarded to everyone.

The next few experiments that Ho et al. [28] ran were meant to test how sensitive this initial experiment was to the choice of task and to the particular parameters we chose (in this case, payments and thresholds). First, they tested whether the results are robust to different choices of threshold by varying what workers needed to do in order to receive the bonus payment. The control had no bonus. Additional treatments required workers to find either 25%, 75%, or all of the typos found by other workers to receive the bonus. Finally, they considered a treatment in which workers were asked to find at least 5 typos total to receive the bonus.

There are a couple of interesting things to note. First performance-based payments led to high quality work for a wide range of thresholds. This is good news since it means the quality improvement is not too sensitive. Quality was slightly worse when workers were asked to identify all of the typos found by other workers, perhaps because they were less confident that they would be able to achieve this goal and gave up. Interestingly, it seems that making the threshold for payment uncertain leads to a big improvement. Since 20 typos were inserted, 25% of the typos would be 5, but quality was much higher when workers were asked

to find 25%.

Experiments designed to test the effect of the bonus size showed that as long as the bonus offered was big enough, quality improved. Offering a very small bonus (in this case $0.05) actually led to a small apparent decrease in performance, though this decrease is not statistically significant.

These results are especially interesting because they may explain some of the disparities in prior work. The paper of Shaw et al. [62] that claimed that performance-based payments don't improve quality used bonus payments that were extremely small compared with the base, so perhaps they were just in the regime where the payments were too small to help. On the other hand, the paper of Yin et al. [76] that noted that bonus sizes don't matter only considered bonuses that were relatively large compared with the base, a regime in which Ho et al. [28] also saw no statistically significant differences in quality when we vary the bonus size.

Finally, Ho et al. [28] asked what types of tasks are amenable to improvement from performance-based payments, testing the theory that these payments work well on *effort-responsive* tasks for which putting in more effort leads to higher quality. To objectively measure which tasks were effort-responsive, they looked at the relationship between the time it took each worker to complete a task and that worker's quality to see whether quality improves with time, a proxy for effort. They found that tasks like proofreading and spotting differences in images were effort-responsive, while handwriting recognition and audio transcription were not. Additionally, their experiments revealed that performance-based payments led to improved quality on proofreading and spotting differences, but not the others. This suggests that whether a task is effort-responsive may indeed play a role in whether quality can be improved using performance-based pay.

Let us briefly return to the question of when simply paying a higher flat payment per task increases work quality. In very recent work, Ye et al. [75] found that higher payments can lead to higher quality work for effort-responsive tasks and presented evidence that this increase in quality is due to an increase in perceived fairness. Additionally, they showed that when crowdworkers believe they are being paid fairly, they work faster and feel more satisfied with their work.

Takeaways and Related Best Practices:

- To be fair to workers, aim to pay at least the U.S. minimum wage. To figure out how much to pay, pilot your task to get a sense of how long it takes to complete. Paying higher than minimum wage can improve your relationship with workers.
- Most evidence suggests that simply paying more for a task does not improve quality, though in some cases it does. New research suggests that this may be tied to whether or not a task is effort-responsive, as well as workers' perceived notion of fairness.
- Performance-based payments can improve quality for effort-responsive tasks. If you aren't sure if your task is effort-responsive, try running a pilot to check the relationship between the time that workers spend on the task and the quality of their work.
- To be effective, bonus payments should be large relative to the base payment. As long as your bonus payment is large enough to make the reward salient, the precise amount doesn't matter too much, nor does the precise quality threshold a worker must meet to receive the bonus.

5.4 Intrinsic Motivation

Although Mechanical Turk is a paid crowdsourcing system, there have been several studies examining the effect of intrinsic, non-monetary sources of motivation for crowdworkers who use the platform.

Chandler and Kapelner [9] showed that workers are more active when tasks are framed as meaningful. They recruited workers on Mechanical Turk to label medical images. In one treatment, workers were told that they were labeling tumor cells and that the results of their work would be used to assist medical researchers. In the control, they were given no context for the task at all. In a third treatment, they were given no context and additionally told that the labels they generated would not be recorded; that is, all of their work would be discarded.

They found that when workers were told their work would benefit medical research, the quantity of work that they produced increased compared with the control, but their work was not significantly more accurate. On the other hand, when workers were told their work would be discarded, the quality of their work was worse than the control, but the quantity of work produced was similar.

Similar effects were observed by Rogstadius et al. [59] who compared the behavior of workers who were told they were performing work for a nonprofit organization "dedicated to saving lives by improving health throughout the world" with workers told they were working for a for-profit pharmaceutical manufacturer.

And of course, beyond paid crowdsourcing systems, the motivation to engage in meaningful work has been a major driver in the success of citizen science platforms and other volunteer-based crowdsourcing systems like the Zooniverse[10] and Science at Home[11].

Recently, Law et al. [42] examined the possibility of appealing to workers' curiosity as a source of intrinsic motivation. Their work was inspired by the *information gap theory* of curiosity, which suggests that when people are made aware that there is a gap in their knowledge, they actively seek out the information needed to fill in this gap. They suggested several "curiosity interventions" aimed at stoking workers' curiosity. While some interventions increased worker productivity, there is some subtlety in how to most effectively engage workers' curiosity.

Finally, gamification [17, 72] has also proved useful as a source of intrinsic motivation in both paid and unpaid crowdsourcing settings.

Takeaways and Related Best Practices:

- Crowdworkers produce more work when they know they are performing a meaningful task, but the quality of their work might not improve.
- Gamification and attempts to stoke curiosity can also increase worker productivity. Make tasks fun and interesting to workers!

[10] https://www.zooniverse.org
[11] https://www.scienceathome.org

5.5 Crowdworker Communication and Forum Usage

There is another assumption about the crowd that people often make without even realizing it: that crowdworkers are independent. When researchers post a task on Mechanical Turk, they expect the responses they receive from different workers to be uncorrelated, or even i.i.d.

Recent ethnographic studies have shown that this is not the case [23, 24]. Extensive interviews with crowdworkers have uncovered that it is common for workers to help each other with administrative overhead (especially in India, where even figuring out how to receive payment can be nontrivial), share information about good tasks and reputable (or irreputable) task requesters, and more generally recreate the social connections and social support that are missing from crowdwork. In other words, there is a hidden communication network behind websites like Mechanical Turk.

Yin et al. [78] attempted to quantify this hidden network in order to better understand the scale and structure of the network and how is it used, focusing on Mechanical Turk. This is challenging because this network is not something that is accessible from an API or easily scrapeable. A lot of the communication goes on offline, for example, through text messaging or even in person. Because of this, the authors needed to devise a way to map the network that would elicit as many "true" edges as possible and avoid eliciting edges that are not real (for example, by paying per edge reported). Additionally, they wanted to do this in a way that would preserve workers' privacy, and therefore could not simply ask workers to report other workers' Mechanical Turk IDs since Turk IDs are not anonymous [43].

To do this, Yin et al. [78] created a web app which they posted as a task on Mechanical Turk. When a worker accepted the task, he was first asked to create a nickname for himself. He then filled out a brief demographic survey and answered a couple of free-form questions about his experience on Mechanical Turk. These questions were carefully selected based on the results of a pilot study in which workers were asked what they were most interested in knowing about other workers on Mechanical Turk. The worker was then asked to pause and swap nicknames with other workers he knows who had already completed the task or might be interested in completing it. This process of swapping nicknames was how the network of communication was constructed. Workers were also able to return and add more nicknames later. When a worker added a connection to another worker, he was asked a few questions like how he usually communicates with this worker and what they communicate about.

Finally, workers were given a chance to explore the partially constructed worker network, viewing the network structure, basic information on all workers (including their answers to the questions from the pilot), and more extensive information about those workers with whom they had exchanged nicknames.

Over a period of several weeks, 10,354 workers completed the task. Based on previous estimates of the number of active workers on Mechanical Turk during any given period of time, this is roughly a census of the worker population during that period [65]. These workers reported a total of 5,268 connections.

Roughly 13% of workers were connected to at least one other worker. On average these workers had 7.6 connections, and the maximum degree of any worker was 321. The largest connected component contained 994 workers, or about 72% of connected workers.

While workers reported communicating in many different ways, the network was primarily enabled by the use of forums. 90% of all edges were between pairs of workers who communicate via forums, and 86% are between pairs who communicate exclusively through forums. These forums create visible subcommunities

in the network. The authors' analysis showed that these subcommunities differ in terms of topological structure, dynamics, and the content of communication, with some acting more as social communities and others more like broadcasting platforms.

Yin et al. [78] found that connected workers tended to find our task earlier. Additionally, connected workers were more likely to have been active on Mechanical Turk longer and more likely to have achieved Mechanical Turk's Master level qualification. They also had a higher approval rate on average. While this experiment was not sufficient to show any causal relationship between connectivity and success on Mechanical Turk, it is consistent with the possibility that being connected has informational advantages to workers.

Takeaways and Related Best Practices:

- Forum usage is widespread on Mechanical Turk. You can think of forums as the virtual equivalent of a "water cooler" for crowdworkers [23]. Workers go to these forums to share information about good and bad tasks and requesters.
- Engage with workers on forums. If you are a new requester, introduce yourself to the workers before you post your first task.[12] Even if you're not new to Turk, posting information about your tasks on the forums can be a good recruiting tool.
- Actively monitor forum discussion of your task. We know that workers discuss tasks on forums. For some tasks, this can be beneficial; workers might share tips to help others complete your tasks more efficiently or accurately. In other cases, however, this discussion can be a big problem. This is particularly true if you are running a behavioral experiment with several different treatments. In these cases, including a polite request to avoid talking about your task in the task instructions or as part of the exit interview can be extremely effective, especially if you explain why. However, it is inevitable that someone will mention your task on the forums anyway, in which case you want to catch this quickly and shut the conversation down.
- Be careful about assuming independence. The workers who complete your task are not an i.i.d. sample of all workers on Mechanical Turk. For many applications, this is not a problem. But when coming up with your research methodology, make sure that you aren't implicitly assuming independence in a way that matters.

6 Additional Best Practices

We conclude this paper by mentioning a few additional best practices that are crucial for effectively using Mechanical Turk for research, yet are rarely (if ever) mentioned in the literature. Some are covered in the guides of Mason and Suri [47] (focused on crowdsourcing for behavioral research) and Goodman and Paolacci [22] (focused on crowdsourcing for consumer research). Others are simply folklore in the crowdsourcing community.

[12] As an example, Yin et al. [78] posted a notification on the forum TurkerNation when they were ready to launch the preliminary trial run of their network experiment in order to recruit a first batch of workers who would be likely to know each other and therefore add links.

The first few tips have to do with cultivating a good relationship with crowdworkers and building a good reputation for yourself as a requester. There are several reasons why this is important. As discussed above, crowdworkers talk about good and bad requesters on forums. It is also common for crowdworkers to use tools that allow them to view a requester rating when viewing a task or be notified when a favorite requester posts a task. Maintaining a good reputation leads to higher interest in your tasks.

- Actively monitor your requester email account and respond to questions. In deciding when to launch your task, make sure you will be able to set aside enough time to communicate with workers as needed. This takes some advanced planning, but it is worth it.
- Approve work quickly. Workers are not paid until their work is approved. Approving work quickly goes a long way towards maintaining a good relationship with workers.
- Avoid rejecting work. Maintaining a high approval rate is very important to workers. Many requesters only allow workers with sufficiently high approval rates to complete their tasks. Rejecting work from a well-meaning worker can therefore harm that worker's chance of earning future income. In general, work should be rejected only in the most extreme circumstances.
- Be an ethical requester. As a start, review and follow the guidelines for academic requesters[13] that were posted as part of the Dynamo project [60].

We conclude with a few more tips to help crowdsourcing projects run smoothly.

- Pilot, pilot, pilot! No matter how carefully you think through the design of your task, your first implementation probably will not be perfect, especially if you're doing something novel. Run pilot studies on your project collaborators, on your colleagues or students who are not directly involved in your project, and eventually, on small batches of crowdworkers. (If running an experiment, make sure to exclude these workers from future iterations of the task.)
- Iterate as many times as needed. It can be time consuming, but it is much better to catch bugs early rather than discovering them after you've fully launched your task.
- Create clear instructions. Pilot your task to collect feedback to make sure that your instructions can be understood. In some cases, it can make sense to insert quiz questions in the instructions to test worker comprehension and correct any misunderstandings.
- Create an attractive and easy-to-use interface. This is crucial for both keeping workers engaged and reducing errors from misunderstandings. Use pilots to test your interface too.
- When appropriate, conduct exit surveys for workers who have completed your task. Find out whether the instructions were clear, how they approached the task, and if they ran into any potential bugs or other issues.

[13] http://wiki.wearedynamo.org/index.php/Guidelines_for_Academic_Requesters

Acknowledgments

This paper is based on notes that were originally prepared to accompany the authors' tutorial at NIPS 2016. Thanks to all of the people—far too many to name—who sent pointers and suggestions of research to include. Thanks to Dan Goldstein, Chien-Ju Ho, Jake Hofman, Andrew Mao, Roozbeh Mottaghi, Sid Suri, Jaime Teevan, Ming Yin, Haoqi Zhang, and all of their collaborators for discussing their research and allowing the author to borrow material from their slides for the tutorial presentation at NIPS. Huge thanks to Chien-Ju Ho, Andrew Mao, Joelle Pineau, Sid Suri, Hanna Wallach, and especially Ming Yin for extended discussions and valuable feedback on prior versions of these notes. And thanks one more time to Sid, for passing down the many unwritten best practices of crowdsourcing over the course of several collaborations.

References

[1] Jacob Abernethy, Yiling Chen, and Jennifer Wortman Vaughan. Efficient market making via convex optimization, and a connection to online learning. *ACM Transactions on Economics and Computation*, 1(2):Article 12, 2013.

[2] Omar Alonso. Implementing crowdsourcing-based relevance experimentation: An industrial perspective. *Information Retrieval*, 16(2):101–120, 2013.

[3] Paul André, Haoqi Zhang, Juho Kim, Lydia B. Chilton, Steven P. Dow, and Robert C. Miller. Community clustering: Leveraging an academic crowd to form coherent conference sessions. In *HCOMP*, 2013.

[4] Pablo J. Barrio, Daniel G. Goldstein, and Jake M. Hofman. Improving comprehension of numbers in the news. In *CHI*, 2016.

[5] Michael Bernstein, Greg Little, Rob Miller, Bjoern Hartmann, Mark Ackerman, David Karger, David Crowell, and Katrina Panovich. Soylent: A word processor with a crowd inside. In *UIST*, 2010.

[6] Michael Buhrmester, Tracy Kwang, and Samuel D. Gosling. Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, 2011.

[7] Chris Callison-Burch. Fast, cheap, and creative: Evaluating translation quality using Amazon's Mechanical Turk. In *EMNLP*, 2009.

[8] Logan Casey, Jesse Chandler, Adam Seth Levine, Andrew Proctor, and Dara Z. Strolovitch. Intertemporal differences among MTurk worker demographics. Working paper on PsyArXiv, 2017.

[9] Dana Chandler and Adam Kapelner. Breaking monotony with meaning: Motivation in crowdsourcing markets. *Journal of Economic Behavior and Organization*, 90:123–133, 2013.

[10] Jesse J. Chandler and Gabriele Paolacci. Lie for a dime: When most prescreening responses are honest but most study participants are imposters. *Social Psychological and Personality Science*, 2017 (To appear).

[11] Jonathan Chang, Jordan Boyd-Graber, Chong Wang, Sean Gerrish, and David M. Blei. Reading tea leaves: How humans interpret topic models. In *NIPS*, 2009.

[12] Yiling Chen, Lance Fortnow, Nicolas Lambert, David Pennock, and Jennifer Wortman Vaughan. Complexity of combinatorial market makers. In *ACM EC*, 2008.

[13] Yiling Chen, Arpita Ghosh, Michael Kearns, Tim Roughgarden, and Jennifer Wortman Vaughan. Mathematical foundations of social computing. *Communications of the ACM*, 59(12):102–108, December 2016.

[14] Lydia Chilton, Juho Kim, Paul André, Felicia Cordeiro, James Landay, Dan Weld, Steven P. Dow, Robert C. Miller, and Haoqi Zhang. Frenzy: Collaborative data organization for creating conference sessions. In *CHI*, 2014.

[15] Philip Dawid and Allan Skene. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, 28(1):20–28, 1979.

[16] Djellel Eddine Difallah, Michele Catasta, Gianluca Demartini, Panagiotis G. Ipeirotis, and Philippe Cudré-Mauroux. The dynamics of micro-task crowdsourcing: The case of Amazon MTurk. In *WWW*, 2015.

[17] Oluwaseyi Feyisetan, Elena Simperl, Max Van Kleek, and Nigel Shadbolt. Improving paid microtasks through gamification and adaptive furtherance incentives. In *WWW*, 2015.

[18] Arpita Ghosh, Satyen Kale, and Preston McAfee. Who moderates the moderators? Crowdsourcing abuse detection in user-generated content. In *ACM EC*, 2011.

[19] Daniel G. Goldstein, R. Preston McAfee, and Siddharth Suri. The cost of annoying ads. In *WWW*, 2013.

[20] Daniel G. Goldstein, Siddharth Suri, R. Preston McAfee, Matthew Ekstrand-Abueg, and Fernando Diaz. The economic and cognitive costs of annoying display advertisements. *Journal of Marketing Research*, 51(6):742–752, 2014.

[21] Ryan Gomes, Peter Welinder, Andreas Krause, and Pietro Perona. Crowdclustering. In *NIPS*, 2011.

[22] Joseph K. Goodman and Gabriele Paolacci. Crowdsourcing consumer research. *Journal of Consumer Research*, 2017 (To appear).

[23] Mary L. Gray, Siddharth Suri, Syed Shoaib Ali, and Deepti Kulkarni. The crowd is a collaborative network. In *CSCW*, 2016.

[24] Neha Gupta, David Martin, Benjamin V. Hanrahan, and Jacki O'Neil. Turk-life in India. In *The Intnernational Conference on Supporting Groupwork*, 2014.

[25] R. Hanson. Combinatorial information market design. *Information Systems Frontiers*, (1):105–119, 2003.

[26] Christopher G. Harris. You're hired! An examination of crowdsourcing incentive models in human resource tasks. In *WSDM 2011 Workshop on Crwdsourcing for Search and Data Mining*, 2011.

[27] Chien-Ju Ho, Shahin Jabbari, and Jennifer Wortman Vaughan. Adaptive task assignment for crowdsourced classification. In *ICML*, 2013.

[28] Chien-Ju Ho, Aleksandrs Slivkins, Siddharth Suri, and Jennifer Wortman Vaughan. Incentivizing high quality crowdwork. In *WWW*, 2015.

[29] John J. Horton, David Rand, and Richard Zeckhauser. The online laboratory: Conducting experiments in a real labor market. *Experimental Economics*, 14(3):399–425, 2011.

[30] Yuening Hu, Jordan Boyd-Graber, Brianna Satinoff, and Alison Smith. Interactive topic modeling. *Machine Learning*, 95:423–469, 2014.

[31] David Karger, Sewoong Oh, and Devavrat Shah. Iterative learning for reliable crowdsourcing systems. In *NIPS*, 2011.

[32] David Karger, Sewoong Oh, and Devavrat Shah. Budget-optimal task allocation for reliable crowdsourcing systems. *Operations Research*, 62:1–24, 2014.

[33] Ashish Khetan and Sewoong Oh. Achieving budget-optimality with adaptive schemes in crowdsourcing. In *NIPS*, 2016.

[34] Joy Kim, Sarah Sterman, Allegra Argent Beal Cohen, and Michael S. Bernstein. Mechanical novel: Crowdsourcing complex work through reflection and revision. In *CSCW*, 2017.

[35] Juho Kim, Haoqi Zhang, Paul André, Lydia B. Chilton, Wendy Mackay, Michel Beaudouin-Lafon, Robert C. Miller, and Steven P. Dow. Cobi: A community-informed conference scheduling tool. In *UIST*, 2013.

[36] Aniket Kittur, Boris Smus, Susheel Khamkar, and Robert E. Kraut. Crowdforge: Crowdsourcing complex work. In *UIST*, 2011.

[37] Adriana Kovashka, Olga Russakovsky, Li Fei-Fei, and Kristen Grauman. Crowdsourcing in computer vision. *Foundations and Trends in Computer Graphics and Vision*, 2016 (To appear).

[38] Raja S. Kushalnagar, Walter S. Lasecki, and Jeffrey P. Bigham. A readability evaluation of real-time crowd captions in the classroom. In *ASSETS*, 2012.

[39] Walter S. Lasecki and Jeffrey P. Bigham. Online quality control for real-time crowd captioning. In *ASSETS*, 2012.

[40] Walter S. Lasecki, Christopher D. Miller, Adam Sadilek, Andrew Abumoussa, Donato Borrello, Raja Kushalnagar, and Jeffrey P. Bigham. Real-time captioning by groups of non-experts. In *UIST*, 2012.

[41] Walter S. Lasecki, Christopher D. Miller, and Jeffrey P. Bigham. Warping time for more effective real-time crowdsourcing. In *CHI*, 2013.

[42] Edith Law, Ming Yin, Joslin Goh, Kevin Chen, Michael Terry, and Krzysztof Z. Gajos. Curiosity killed the cat, but makes crowdwork better. In *CHI*, 2016.

[43] Matthew Lease, Jessica Hullman, Jeffrey P. Bigham, Michael S. Bernstein, Juho Kim, Walter S. Lasecki, Saeideh Bakhshi, Tanushree Mitra, and Robert C. Miller. Mechanical Turk is not anonymous. In *Social Science Research Network (SSRN) Online*, 2013.

[44] Leib Litman, Jonathan Robinson, and Cheskie Rosenzweig. The relationship between motivation, monetary compensation, and data quality among US- and India-based workers on Mechanical Turk. *Behavioral Research Methods*, 47(2):519–528, 2014.

[45] Qiang Liu, Jian Peng, and Alexander Ihler. Variational inference for crowdsourcing. In *NIPS*, 2012.

[46] Andrew Mao, Lili Dworkin, Siddharth Suri, and Duncan J. Watts. Resilient cooperators stabilize long-run cooperation in the finitely repeated prisoners dilemma. *Nature Communications*, 2016 (To appear).

[47] Winter Mason and Siddharth Suri. Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods*, 44(1):1–23, 2012.

[48] Winter Mason and Duncan J. Watts. Financial incentives and the "performance of crowds". In *HCOMP*, 2009.

[49] Roozbeh Mottaghi, Sanja Fidler, Jian Yao, Raquel Urtasun, and Devi Parikh. Analyzing semantic segmentation using hybrid human-machine CRFs. In *CVPR*, 2013.

[50] Roozbeh Mottaghi, Sanja Fidler, Alan Yuille, Raquel Urtasun, and Devi Parikh. Human-machine CRFs for identifying bottlenecks in scene understanding. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2016.

[51] Iftekhar Naim, Daniel Gildea, Walter Lasecki, and Jeffrey P. Bigham. Text alignment for real-time crowd captioning. In *NAACL*, 2013.

[52] Besmira Nushi, Ece Kamar, Donald Kossmann, and Eric Horvitz. On human intellect and machine failures: Troubleshooting integrative machine learning systems. In *AAAI*, 2017.

[53] Gabriele Paolacci, Jesse Chandler, and Panagiotis G. Ipeirotis. Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making*, 5:411–419, 2010.

[54] Devi Parikh and C. Lawrence Zitnick. Human-debugging of machines. In *Second NIPS Workshop on Computational Social Science and the Wisdom of Crowds*, 2011.

[55] Genevieve Patterson and James Hays. Sun attribute database: Discovering, annotating, and recognizing scene attributes. In *CVPR*, 2012.

[56] Genevieve Patterson, Chen Xu, Hang Su, and James Hays. The sun attribute database: Beyond categories for deeper scene understanding. *International Journal of Computer Vision*, 108(1–2):59–81, 2014.

[57] Eyal Peera, Laura Brandimarteb, Sonam Samatc, and Alessandro Acquistic. Beyond the Turk: An empirical comparison of alternative platforms for crowdsourcing online behavioral research. 70:153–163, 2017.

[58] Vikas C. Raykar, Shipeng Yu, Linda H. Zhao, Gerardo Valadez, Charles Florin, Luca Bogoni, and Linda Moy. Learning from crowds. *Journal of Machine Learning Research*, 11:1297–1322, 2010.

[59] Jakob Rogstadius, Vassilis Kostakos, Aniket Kittur, Boris Smus, Jim Laredo, and Maja Vukovic. An assessment of intrinsic and extrinsic motivation on task performance in crowdsourcing markets. In *ICWSM*, 2011.

[60] Niloufar Salehi, Lilly Irani, Michael Bernstein, Ali Alkhatib, Eva Ogbe, Kristy Milland, and Clickhappier. We are dynamo: Overcoming stalling and friction in collective action for crowd workers. In *CHI*, 2015.

[61] Niloufar Salehi, Jaime Teevan, Shamsi Iqbal, and Ece Kamar. Communicating context to the crowd for complex writing tasks. In *CSCW*, 2017.

[62] Aaron D. Shaw, John J. Horton, and Daniel L. Chen. Designing incentives for inexpert human raters. In *CSCW*, 2011.

[63] Victor Sheng, Foster Provost, and Panagiotis Ipeirotis. Get another label? Improving data quality using multiple, noisy labelers. In *KDD*, 2008.

[64] Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. Cheap and fast—but is it good?: Evaluating non-expert annotations for natural language tasks. In *EMNLP*, 2008.

[65] Neil Stewart, Christoph Ungemach, Adam J. L. Harris, Daniel M. Bartels, Ben R. Newell, Gabriele Paolacci, and Jesse Chandler. The average laboratory samples a population of 7,300 Amazon Mechanical Turk workers. *Judgment and Decision Making*, September 2015.

[66] Siddharth Suri, Daniel Goldstein, and Winter Mason. Honesty in an online labor market. In *HCOMP*, 2011.

[67] Omer Tamuz, Ce Liu, Serge Belongie, Ohad Shamir, and Adam Kalai. Adaptively learning the crowd kernel. In *ICML*, 2011.

[68] Jaime Teevan, Daniel Libeling, and Walter Lasecki. Selfsourcing personal tasks. In *CHI*, 2014.

[69] Jaime Teevan, Shamsi Iqbal, and Curtis von Veh. Supporting collaborative writing with microtasks. In *CHI*, 2016.

[70] Blase Ur, Jonathan Bees, Sean M. Segreti, Lujo Bauer, and Lorrie Faith Cranor Nicolas Christin. Do users' perceptions of password security match reality? In *CHI*, 2016.

[71] Donna Vakharia and Matthew Lease. Beyond Mechanical Turk: An analysis of paid crowd work platforms. In *Proceedings of the iConference*, 2015.

[72] Luis von Ahn and Laura Dabbish. General techniques for designing games with a purpose. *Communications of the ACM*, 51(8):58–67, August 2008.

[73] Peter Welinder, Steve Branson, Serge Belongie, and Perona Pietro. The multidimensional wisdom of crowds. In *NIPS*, 2010.

[74] Kathryn Sharpe Wessling, Joel Huber, and Oded Netzer. Character misrepresentation by Amazon Turk workers: Assessment and solutions. *Journal of Consumer Research*, 2017 (To appear).

[75] Teng Ye, Sangseok You, and Lionel P. Robert Jr. When does more money work? Examining the role of perceived fairness in pay on the performance quality of crowdworkers. In *ICWSM*, 2017.

[76] Ming Yin, Yiling Chen, and Yu-An Sun. The effects of performance-contingent financial incentives in online labor markets. In *AAAI*, 2013.

[77] Ming Yin, Yiling Chen, and Yu-An Sun. Monetary interventions in crowdsourcing task switching. In *HCOMP*, 2014.

[78] Ming Yin, Mary L. Gray, Siddharth Suri, and Jennifer Wortman Vaughan. The communication network within the crowd. In *WWW*, 2016.

[79] Yuchen Zhang, Xi Chen, Dengyong Zhou, and Michael I. Jordan. Spectral methods meet EM: A provably optimal algorithm for crowdsourcing. *Journal of Machine Learning Research*, 17(102):1–44, 2016.

[80] Dengyong Zhou, Sumit Basu, Yi Mao, and John Platt. Learning from the wisdom of crowds by minimax entropy. In *NIPS*, 2012.

[81] James Zou, Kamalika Chaudhuri, and Adam Tauman Kalai. Crowdsourcing feature discovery via adaptively chosen comparisons. In *HCOMP*, 2015.

A Alternative Crowdsourcing Platforms

Most of the research described in this paper was performed on Amazon Mechanical Turk. Mechanical Turk is a widely recognized crowdsourcing platform used broadly in the research community, but it is not the right choice for everyone. In particular, it can be difficult to use from outside of the United States. Luckily, many alternatives are available. For example:

- CrowdFlower[14] is a crowdsourcing platform widely used in both industry and research. CrowdFlower offers specialized enterprise solutions for businesses with artificial intelligence and data science needs including search relevance evaluation, sentiment analysis, and data classification.
- ClickWorker[15] is a German crowdsourcing platform that attracts European workers. It provides support for specialized tasks such as translation, web research, and web content generation. It also provides tools for mobile crowdsourcing.
- Prolific Academic[16] is a UK-based crowdsourcing platform focused on connecting researchers with participants for their studies.
- Upwork[17] is an online freelancer marketplace focused not on microtasks but rather on larger scale jobs such as writing an article or designing a website.

For a thorough comparison of these and other alternative platforms, we direct the reader to Vakharia and Lease [71] and Peera et al. [57]. Vakharia and Lease [71] provide a thorough qualitative content analysis of seven platforms: ClickWorker, CrowdComputing Systems (now WorkFusion), CloudFactory, CrowdFlower,

[14] https://www.crowdflower.com
[15] https://www.clickworker.com
[16] https://www.prolific.ac
[17] https://www.upwork.com

CrowdSource, MobileWorks (now LeadGenius), and oDesk (now Upwork), examining factors like infrastructure and tools, support for fraud protection, and quality of work. Peera et al. [57] provide a detailed experimental comparison of three platforms: Mechanical Turk, CrowdFlower, and Prolific Academic. They compare dropout rates, response rates, workers' performance on attention-check questions, workers' reliability, workers' familiarity with common psychology studies (which would signal an overused population of subjects), and the ability to replicate classic psychology studies on each platform.