"It Would Be Super Helpful": Opportunities for Non-Expert Collection of Environmental Data

Alex Cabral
Harvard University
Cambridge, Massachusetts
acabral@seas.harvard.edu

ABSTRACT

There is increased recognition of the need for local climate data, particularly in underserved communities, which are also expected to be the most affected by climate change. Included in these communities are the American urban poor and minorities, who are more likely to suffer or die from extreme heat [23, 27], and many in the Global South, who are facing growing food shortages due to the increase of climate-related disasters [8]. Because of financial and geographic constraints, gathering local climate data from these communities can be challenging. To address these numerous issues, I propose leveraging my experience as a former software engineer and middle and high school educator to create citizen science and citizen sensing tools that will engage students, particularly in underserved communities, in local environmental data generation and analysis via active, project-based learning. An important feature of this work is the generation of relevant, useful data for both climate scientists and community leaders as they aim to better track and plan for the effects of climate change. For my final project, I designed and conducted interviews with climate and community experts and stakeholders to determine what local data would be useful and how it could be framed for experts to trust the data. I used grounded theory method techniques to analyze two interviews and found four emerging themes – reputation and trust, risks of data collection, quality over quantity, and the need for more data. These results, in conjunction with the findings from my future interviews, will help create a set of environmental data that non-experts can collect, will highlight how to make non-expert data trustworthy, and will guide the design of future technological interventions for environmental benefit.

CCS CONCEPTS

• **Human-centered computing** → **Human computer interaction (HCI)**; *Field studies*; • **Applied computing** → **Environmental sciences**; • **Information systems** → *Information retrieval.*

KEYWORDS

hci, citizen science, environmental data, data collection

COMPASS '20, June, 2020, Guayaquil, Ecuador

ACM ISBN 978-1-4503-9999-9/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

ACM Reference Format:
Alex Cabral. 2020. "It Would Be Super Helpful": Opportunities for Non-Expert Collection of Environmental Data. In *COMPASS '20: ACM SIGCAS Computing and Sustainable Societies, June, 2020, Guayaquil, Ecuador*. ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

There is increased recognition of the need for local environmental data, particularly to understand climate change on a local scale for the design and execution of appropriate adaptation technologies. However, gathering local climate data can be challenging due to geographical complexity, financial constraints, or lack of current climate data technologies. To address these issues, scholars are increasingly turning to citizen science and citizen sensing tools to get non-expert assistance in collecting data for low or no cost. These tools have had success for specific tasks, but there is also an opportunity to use them more generally as a means of data collection. In addition, there is an untapped opportunity to also use these participatory technologies to learn about social systems and their interactions with physical and natural systems.

Most citizen science and citizen sensing projects fail to find participants that are representative of the diverse communities in which the projects are deployed [11, 22]. Driven primarily by institutional researchers, such projects reflect, and may even amplify, the expectations, goals, and constraints of the researchers. Flipping this model so that communities design the tools would provide great benefits. Communities could address their specific interests and issues while engaging in direct action with expert assistance and feedback. Researchers could tackle questions of real-world impact that draw broader interest and participation. In addition, researchers could learn how different communities prioritize, view, and interact with the systems around them based on factors such as demographics and the physical environment. For these reasons, scholars are beginning to push for community or collaboratively designed citizen science and citizen sensing tools.

An important feature of generalized or community-driven citizen science and citizen sensing tools is the generation of relevant, useful data for research scientists and community leaders. In particular, this is essential to track and plan for the effects of climate change on a local scale. Prior research has looked into trust overall for data, machine learning, and artificial intelligence. However, there has not yet been a deep dive into the specifics of environmental data from a Computer Science or Human-Computer Interaction perspective. With the rise of low-cost technologies to collect environmental data and an increase in the general public interest of environmental issues, it is essential that Computer Scientists learn more about environmental data to ensure that environmental tools are not created or used without contributing in a positive way.

In this project, I designed a semi-structured interview to conduct with climate and community experts and stakeholders to determine what local data would be useful to collect and how it could be framed to be trustworthy. I then identified and recruited a set of stakeholders to reach out to. I also began the lengthy IRB process to get approval for the interviews so I can publish the results in the future. I conducted two interviews, one with an environmental scientist and one with a flood risk analyst. I recorded audio for the two interviews then transcribed the audio to text. Using the grounded theory method, I began analyzing the text from the two interviews to identify emerging themes.

Although this work is still in its infancy, I am already gaining useful insights from the two interviews I have conducted. Both interviewees suggested that non-experts could help collect data and that it would be useful for them to do so. Additional emergent themes include the connection between reputation and trust, potential dangers of data collection, that high quality data is better than a large amount of data, and the need for more data coverage in many locales.

Through these interviews I aim to answer a number of questions:

- What environmental data can non-experts reliably collect?
- What environmental data can non-industry sensors reliably collect?
- How can non-expert or non-industry sensor collected data be presented so that it is trusted?
- How can non-expert or non-industry sensor collected data be presented so that it is useful for experts?
- What are the risks of non-experts and non-industry sensors collecting data?

To develop meaningful answers to these questions, I will need to conduct additional interviews. I plan to conduct these over the next two months so that I can push to submit the results to a conference in early March. Additionally, getting results early will allow me to start working towards the larger goal of educational citizen science and citizen sensing tools.

This work makes a number of contributions to the Human-Computer Interaction and broader Computer Science field. First, this work will result in a set of environmental data that non-experts can collect. This list will help guide future citizen science, citizen sensing, and other environmental technology projects. Additionally, this project will provide information on the factors that make environmental data useful and trustworthy for experts. These findings may also prove helpful in understanding what makes data useful and trustworthy in other domains.

This work can also help make a broader contribution to society by providing useful information to help address data justice. By highlighting the data that can be collected by non-experts, this work will guide designers and developers to create tools that can collect that data. These tools can then be deployed in areas that are currently underserved in terms of data representation.

2 BACKGROUND AND RELATED WORKS

This work draws from a number of research areas, combining different knowledge sources for both the problem formulation and project implementation.

2.1 Need for Local Data

Growing research in Climate Justice – the framing of climate change as an ethical, humanitarian issue rather than a purely environmental one – has shown that there are a number of injustices related to climate change, including the lack of local climate data in many communities [1, 10, 19, 23]. Additional scholars have pointed out the need for increased local environmental data for tasks such as rainwater measurement [1], regional climate prediction [24], and accommodation of topographical diversity [28]. Yet other scholars have described the difficulty in generating regional predictions from the existing global climate models [7, 21, 29, 30].

2.2 Citizen Science and Citizen Sensing

Citizen science and citizen sensing are increasingly used to gather local data for various tasks [2, 3, 12, 15], and there are a number of external studies that strongly promote the efficacy of these technologies [5, 25, 26]. Through these initiatives, researchers have made many essential findings, including changes in bird migration patterns and differences in temperature and air quality within a single city or community [2, 4]. These findings have clear implications for environmental and public health, and are also linked to social sciences via findings that factors such as temperature can affect crime rate in a city [9, 18].

2.3 Trust and Citizen Science

Some scholars have found that the data collected by citizen scientists may be poor-quality, misleading, or even malicious [6, 14]. This has resulted in distrust of citizen science data for many scientists, even when the quality is comparable to that collected by experts [14, 16, 17].

2.4 Grounded Theory Method

There are a number of different methods used to analyze qualitative results in research. In this work, I have chosen to use the grounded theory method [20]. In grounded theory method, all theories are built directly from the data rather than from prior thoughts or hypotheses. Theories are then iteratively revised and tested through successive rounds of data collection and analysis.

The main form of analysis in grounded theory method is thematic coding based on the emergent themes from the data. There are three phases of coding practice generally used in grounded theory method – open coding, axial coding, and selective coding.

2.4.1 Open Coding. In the open coding phase, the researcher begins by writing simple descriptive labels for different phrases or sentences. The codes generally refer to persons, objects, or concepts found in the data. Over time, certain codes will begin to repeatedly emerge, and the researcher tracks these codes. Many researchers use software tools to maintain a list of the codes and easily apply codes to additional portions of the transcript.

2.4.2 Axial Coding. In the axial coding phase, the researcher examines the recurring open codes to find relationships among them. These relationships lead to clusters that are then given a name. These named clusters are generally referred to as "categories", and

the main goal of this phase is to find powerful categories that describe the data. The abstract, conceptual categories are referred to as *axial codes*.

2.4.3 Selective Coding. After time, the researcher will determine that some axial codes are more important than others. The researcher will then focus on the more important codes and ignore the less important ones. This act of selecting certain codes is known as *selective coding*. In this phase, the researcher determines which topics to pursue based on the pervasiveness of different codes and categories. In addition, the researcher is expected to choose the codes based on the underlying themes and theories derived from the data.

3 METHODS

I completed a number of steps to complete this project and also prepare for the upcoming interviews I will have to conduct. Through these steps, I have begun to fine tune the process so that future interviews will run smoothly and quickly.

3.1 Stakeholder Identification

I began this project by identifying the various stakeholders and thus potential participants for the interview. The most important factor was that the participants work with environmental data in a professional capacity. With guidance from a professor at Harvard, I determined that this included urban planners, environmental scientists, and other government or community workers. Although this may not include the full list of jobs or roles that interact with environmental data, I felt that this represented most of the roles and also covered those that have the biggest impact with their work.

3.2 Interview Design

I then pivoted to design the interview that I was going to conduct. I chose to design a semi-structured interview, in which I had a set of questions to ask each participant but could add in additional questions as necessary based on the participant's responses.

The first interview question was "What do you do?". This question allowed me to get detailed information on what the participant does and how he or she uses environmental data. Additionally, this question acted as a last check to ensure that the participant did indeed work with environmental data. If I found at this time that the participant does not work with environmental data, I would end the study.

I then devised a list of questions to ask the participant about their most recent project. I chose to frame the questions in relation to a specific project rather than the general case because people have been shown to have better memory for specific events rather than general scenarios [13]. In addition, it is more likely that the participant's most recent project will be representative of the general case instead of a unique outlier. Furthermore, the participant will likely indicate if the project is especially different than the usual case.

For the participant's most recent project, I asked the following questions:

- What kind of data did you use in that project?
- How was the data collected?
- What made you trust the data?
- What data do you wish you had access to for that project?
- How might a non-expert help collect data for the project?
- What would worry or concern you about a non-expert collecting data for this project?
- How did you balance quality versus quantity for the data you used in the project?

From these questions I could gather useful information about the types of data that non-experts can help collect, how the data might be collected, and how to make the data trustworthy. In addition, by going through the details of a specific project, I could build a deep understanding of the workflow.

I then asked a set of general questions to gather additional information about the possibility of incorporating non-expert collected data. Those questions were:

- How often do you use data that you did not collect?
- How often do you make data that you do collect publicly available?
- How would you feel about using citizen-collected data in your work?

The answers to these questions provide insight to the feelings of shared data. In particular, they highlight the willingness of experts to use non-expert collected data and whether that willingness differs from other shared data sources.

For interviews conducted in person, I could also choose to ask to see one or more datasets, assuming they do not include sensitive information. This visualization may help in my understanding of the challenges and also provide insight into useful tools that can be created.

3.3 IRB Submission

After completing the interview design, I then wrote and submitted a protocol for IRB approval of my project. This is a necessary step to conduct human-subjects research and publish the results. To complete this process, I filled in a 30 page template, submitted a copy of my interview questions, wrote and submitted a consent form, and also wrote and submitted a copy of the recruitment email I would use to get participants. The recruitment email and consent form are both included in the appendix.

I am still waiting for IRB approval, but have passed the pre-review stage. I expect to have the full approval by early January.

3.4 Recruitment

I began my recruitment by directly contacting people I know who work with environmental data. I contacted them via email using a modified version of the recruitment email in Appendix A. The interview was designed to last between 30 and 60 minutes, but I told the participants that it would take 60 minutes to ensure that we did not run out of time. I also provided a $25 gift card for compensation.

After my IRB protocol is approved, I will begin to recruit participants whom I do not already know. To find participants, I will ask people I know to recommend others for the study. I will also go through department and governmental web pages to find faculty, staff, and researchers who seem to be a good fit for the study.

3.5 Conducting Interviews

After recruiting and scheduling time with participants, I conducted the interviews. I have currently only conducted two interviews, but the process will be the same for the additional interviews I will conduct in the future.

The interviews were scheduled to take 60 minutes. I began by reviewing the consent form with the participants to ensure that they agreed to the terms and to answer any questions or concerns they had. After receiving verbal confirmation the participant understood and agreed to the terms, I began the formal interview.

The first step was to begin the audio recording. I used Otter[1], a smartphone application that records audio and creates a textual transcription while recording. In addition, I made a separate recording on my laptop using Quicktime, in the event a backup is needed.

I then proceeded to go through the interview questions, as laid out in section 3.2. I followed the ordering of the questions as designed, unless something the participant said drove me to ask an ancillary or clarifying question.

After completing the interview questions, I gave participants the opportunity to add any last comments or thoughts. I then gave the participants their compensation, a $25 gift card.

The first interview was conducted over the phone and the second interview was conducted in person. I expect most interviews to happen in person, but am also prepared to conduct virtual interviews if necessary. For in person interviews, the participant receives a $25 Visa gift card. For virtual interviews, the participant receives a $25 gift card of their choosing. In an effort to promote sustainability, I chose not to give Amazon gift cards, which tend to be common for research compensation.

3.6 Data Analysis

To analyze the data, I first transcribed the audio files to text. The Otter app I recorded interviews with did this automatically, but the transcriptions were only about 70% accurate. I listened to the interview in full and read through the transcript, making corrections as necessary.

I then read through the corrected transcripts, pulling out interesting quotations and manually creating open codes (based on the grounded theory method) for noteworthy phrases. For this project I made a list of themes and phrases, but moving forward I will use a software application to assist in this process. I am considering NVivo[2] or Delve[3] for this task.

Moving forward, I will complete the steps of axial coding and selective coding, as described in the literature for grounded theory method. This will be easier to complete when I have additional interview data. The codes from the first round of interview will determine the second round of interviews when I try to disprove the theories grounded in the data from the first round.

4 RESULTS

I have successfully completed two interviews, one with a research scientist studying mercury and another with a flood risk analyst. I transcribed the audio recording of each interview to text, then examined the text to find emerging themes. In this section, I provide an overview of each interview and highlight the emerging themes that appeared in both.

4.1 Participant 1

Participant 1 (P1) is a Program Manager for a program affiliated with the National Flood Insurance Program. In this role, P1 works with actuaries and engineers to determine how to best set insurance prices for insurance policies covered by the National Flood Insurance Program. P1 coordinates with others to determine how to best use data to set the prices for flood insurance policies. Prices are set each year, so the general goal is to use the data to determine an individual's flood risk for the next calendar year.

P1 primarily relies on two data sets for this work – a data set collected by the federal government and another one provided by one or more private companies. The government data is based on hydraulics and hydrolics, and is thus commonly referred to as *H & H modeling*. This type of modeling is deterministic, so focuses on showing extreme events, such as the 100 year flood. Therefore, this data set alone is inadequate because floods do not occur on 100 year cycles, but are rather dynamic and can occur frequently.

To address this shortcoming, the team also uses a set of stochastic models such as catastrophe modeling. These models run simulations based on past events and possible outcomes through a number of algorithms numerous times, to show what the full range of possibilities is. Compared to H & H models, which aim to provide a house by house view, these stochastic models provide a view of a wider range.

The data for these models come from a variety of sources. The data for the catastrophe models is proprietary because the models are created by private, for-profit corporations. The models are constructed based on prior events and also take into account factors such as development changes and patterns, land use, and population. In addition, the catastrophe models run simulations in the future to try and predict today's flood scenarios.

The government-owned H & H models are built on public data that the government already has. A lot of the information is gleaned from their own insurance policies, and they also incorporate information from different agencies that collect data on flood heights and other factors. A primary goal of these models is to develop *instance damage curves*. These curves aim to address the following scenario – if I have one foot of water in my house, how much damage, from an insurance payout standpoint, does that mean and how does it change as the water increases from one foot of water to two or more feet of water?

Because the two models and two sets of data contrast so greatly, P1 found benefit in using both of them. the government data is used to make flood maps, and these are useful from a deterministic standpoint to determine if someone is in the flood zone and thus has a 100% chance in any given year of being flooded. The probabilistic models, such as the catastrophe models, are used to determine in any given year what the risk is of having any harmful impacts from floods.

In regards to trust, P1 had an interesting view. P1 did not necessarily trust the data that came from the private companies for the catastrophe models, but instead recognized that the data met the

[1] https://otter.ai
[2] https://www.qsrinternational.com/nvivo
[3] https://delvetool.com/

needs of the job. Because of the shortcomings of the government data, P1 appreciated having an external, distinct data set to use. This was especially important because the government data is used to create flood maps, which are also being used to make decisions about how land use happens and where development can happen. P1 found this problematic because those kinds of decisions will last for 20 or 30 years, whereas the flood insurance rate maps are supposed to be a point in time of flood risk. Additionally, the flood maps are not updated often and can be very wrong. This was the case in Houston after Hurricane Harvey, when an extraordinary amount of people who were flooded were not considered at risk by the flood map.

P1 indicated that citizens could help collect data, especially for *pluvial floods*. Pluvial floods are those which are caused by excessive rainfall as opposed to storm surges, tidal waves, or rising sea level. Pluvial floods are difficult to measure because they are generally unpredictable and the height of the flood can differ widely within a single city or community. Because citizens experience pluvial floods firsthand, they may be able to help measure the height of the flood at different locations in the city. However, P1 also indicated that there may be an inherent danger to measuring flood levels. Floods can be higher than expected, causing a person to possibly drown. Thus this specific task may be better suited for citizen sensing rather than citizen science.

Another area where non-experts may help collect data is in detailed heights of water after a flood. P1 indicated that the *depth damage function*, which is used to predict flood risk, was not created from a reliable set of data. In many cases, the amount of water in any given town, and how much of a financial impact that would have in terms of damage, was often guessed. From house to house, there were averages and probabilities, but the underlying data was often not there. It is difficult for government officials and homeowners to get that data because one may have a general sense of how high the water is after a flood, but that data is not necessarily captured and fed into all of these models. If homeowners had the right tools and information to measure the water, this could help improve the models in the future.

A potential risk of non-experts collected flood data is that it may be dangerous. Because flood heights can be hard to determine, and there may be uncertainty of the underlying elevation and terrain, people measuring the flood water may underestimate the height and perhaps drown.

P1 did not use publicly available data besides the government data used to create flood maps, which is likely already public. In addition, P1 did not share data publicly for others to use.

Finally, P1 highlighted additional challenges with the government data used to create flood maps. The data was sometimes of poor quality because of coding errors, transitions between different systems, or different ways of collecting the data. An additional challenge is that the data for the catastrophe models could not assist with the quality issues because the level of analysis is quite different.

4.2 Participant 2

Participant 2 (P2) is a research scientist studying the biogeochemistry of mercury. Specifically, P2 studies the exchanges between the terrestrial biosphere – plants and soils – and the atmosphere. P2 is developing models to understand the controls on that exchange between the atmosphere and biosphere. Those models will then be integrated into larger global models so the cycle can be linked to anthropogenic emission sources. The goal is to identify where mercury will be stored over 10 year, 100 year, and even longer timescales.

The main driver of this work is that the existing model must be updated to accommodate new data that has been collected in the past decade by the US Geological Survey. The USGS conducted a large-scale soil survey, in which they dug thousands of holes to measure all of the metals in the soils. P2 is now using that data to help determine whether the current model correctly tracks this distribution of mercury accumulation.

P2 also collaborates with researchers who measure the direct exchange of mercury between the atmosphere and the forest canopy. Again the goal of this work is to determine whether the parameters in the model are accurately representing these intensively measured sites.

P2 trusts the data from the large soil study because it came from the USGS. P2 mentioned that whom the data comes from is extremely important. Additionally, the larger the effort is to collect the data, the more confident P2 is that somebody has put time into ensuring that the methods are robust. This is also tied to the reputation of who released the data – because the USGS is large and relatively well-funded, P2 trusts that someone ensured the quality of the data and its collection methods. P2 also mentioned reading about the pilot project and the results of that pilot project as an additional means to trust the data, even though the report does not indicate whether the data from the pilot are actually good or not.

P2 would like to have access to flux data, which measures how mercury concentrations scale with the air mass. This data can help identify mercury exchange between the forest and boundary layer above it. P2 mentioned that this data does exist for some places, but they are not collocated with the soil measurement locations. This makes it difficult to build a robust model because there is not a complete understanding of the mercury exchange system in a single environment. Currently P2 handles that by interpolating and filling in data from other locations, but there is a level of uncertainty generated from this process.

A lot of difficulties in P2's work arise from the cost of data collection. Atmospheric mercury sensors cost tens of thousands of dollars, and that does not include the cost of human labor to set up, maintain, and evaluate data from the sensors. Additionally, P2 mentioned that many projects cannot be driven to completion because the grant funding expires or runs out before the project is finished. P2 highlighted this as a shortcoming of existing data collection methods and an opportunity for non-experts to have a positive impact.

Like P1, P2 felt that non-experts could help collect data successfully. However, P2 highlighted a unique concern for non-expert data. P2 mentioned that a single data reading in isolation may not be particularly useful. In the case of mercury, one needs to know about the rain, soil readings, and atmospheric readings to make sense of the mercury levels. P2 also felt that it would be challenging to figure out how to analyze the data, especially if it relied on physical samples as mercury readings do. Furthermore, there is the

risk of mercury contamination by trace elements that non-experts may not be able to stop. For example, many dental fillings have mercury in them, and although it is unlikely that these fillings will contaminate samples, it highlights how common, everyday items may have an impact on data collection.

P2 prefers a small amount of high quality data instead of a large amount of data sets. This is because the integration of each additional data set takes a substantial amount of time. In addition, the search cost to find additional data sets is quite high. P2 indicated that the current search process for a data set is to spend a moderate amount of time searching the web before resorting to an advisor or other professor for guidance. Generally, P2 incorporates the highest quality data set first, then incorporates additional data sets, only as necessary to cover the gaps or shortcomings of the prior ones. Each additional data set that is added is generally the highest quality of those that are not yet included in the model.

A shortcoming of the data that P2 uses is that most of the readings come from the developed world. This is particularly challenging in the developing world, because most of the resources are focused on mercury readings near gold mines, where mercury is known to be pervasive, and not in areas such as forests. P2 mentioned a recent study that suggested tree foliage having a large impact on mercury levels in the atmosphere. However, it is difficult to trust the results of that study because the data all comes from the northern hemisphere. In fact, P2 said there are only two atmospheric mercury sensors in the entire southern hemisphere.

P2 mentioned additional factors that aid in the data search and usage process. One factor is inter-comparison of data sets. When a new data set is released, there are standards used as a benchmark to measure how good the data set is. For example, all new models or data sets will use a specific type of Montana soil and the researchers will release the results using that soil type. This allows researchers to make comparisons with existing data sets they have based on the results of the model using that data set compared to others. Generally, the new data or model is trusted if the results fit within a certain range of what they currently find in other data sets and models.

Another factor is common formatting for data sets in repositories. Many environmental fields have large data repositories with numerous data sets. To submit to the common repository, a researcher must adhere to the common format. This helps for easier data sharing and widespread data usage.

4.3 Emerging Themes

From the initial analysis of the two interviews I've conducted, I have found a number of themes that seem to be pervasive. These themes are highlighted an described in this section.

4.3.1 Reputation and Trust. The most prominent factor that seems to affect trust for both P1 and P2 is the reputation of who is providing the data. For P2, this relates to the reputation of the researcher or institute that provides the data. For P1, this refers to the companies that provide the external data and the means they use to gather it. For both P1 and P2, even if they do not know the particulars of how the data is collected, they trust it if it is collected by certain individuals or institutes.

4.3.2 Risks of Data Collection. P1 and P2 both indicated potential risks in non-experts collecting data for their work. P1 identified physical risks, such as personal harm or drowning. P2 identified risks in data contamination, especially because non-experts may not know what objects have mercury in them.

4.3.3 Quality over Quantity. When asked about the preference between high quality and a large quantity of data, both P1 and P2 answered that they prefer to have high quality data. P1 would like a lot of high quality data, but P2 prefers a small number of high quality data sets. For P2, this is important because of the overhead in integrating additional data sets into the model.

4.3.4 Need more Data. Both P1 and P2 addressed the need for more data, particularly in specific areas. P1 described the need for pluvial flood data and for increased flood data in areas that have varying elevation. P2 addressed the lack of mercury reading stations in the southern hemisphere and the overall shortage of atmospheric mercury readings in areas that do not have a large amount of research.

4.4 Noteworthy Quotations

"It's not necessarily that I trusted the data. It was more that the type of data that they were providing met the needs that my client had." – P1

"We're starting at a pretty low bar, and the way that [we] map things because it's based on this deterministic way of looking at floods like it's already not doing a good job of communicating risks." – P1

"We ended up using both [government] data and this probabilistic modeling data, because we felt that the blend married the best of both worlds." – P1

"... if either of those things is off by six inches, it can have a pretty drastic impact on what your expected damages are, which will in turn impact what your flood insurance rates might be." – P1

"I think having a better sense of what those - it's both like how the floods happen, and where the water goes, and how it goes, and also the impacts of those flooding events would be interesting data to get." – P1

"I would say always high quality [data] and always a lot of it." – P1

"There are lots of questions that are not sexy enough to get research funding, but will still be useful to know over the long term." – P2

"I mean, science is networked and it's not a great answer but I think I just happen to trust the person who it came from" – P2

"I know it's because when they do this kind of work, they have resources allocated to making sure that the data are good data." – P2

"... to the point of citizen science that's probably a question of like how much ancillary data do you need to collect in order to make meaning out of the record that you've created." – P2

> "If you're looking at data, and you want to be skeptical about its quality, you would have to think about how could a 13 year old mess this up. And, depending on what's being collected there could be wildly different answers." – P2

5 DISCUSSION

I am finding value in this project just from the two interviews that I have conducted. Although I will need to conduct additional interviews to have meaningful, generalizable results, the emerging themes and initial findings from these interviews is already providing useful information for my long term thesis project.

One of the great benefits of this project is that it taught me a lot about environmental data, including its collection, usage, challenges, and analysis. As a Computer Scientist, it is sometimes difficult to understand others' challenges with data, and this project helped me to do so. For example, I had not considered the difficulty in working with data sets that have wildly different formats. This is a feature I will certainly need to include in my long term thesis project to ensure that researchers can easily integrate data from various users and places.

The biggest takeaway I have now is the issue of trust. Both P1 and P2 indicated that reputation is strongly tied to trust. This is a hurdle I will have to tackle to make data collected by middle and high schoolers trustworthy. P2 indicated that there are two ways of looking at this problem – people may say how in the world could you trust a 13 year old to collect data, or people may say how in the world could a 13 year old mess this up? I think the second viewpoint will be useful in my work. Perhaps by highlighting the tasks that data collectors completed and by making the tasks as foolproof as possible, that will help address the issue of trust.

Another main takeaway from these findings is that there are a number of underserved areas in regards to data collection. Reaching these areas is a primary objective of mine. I will begin by focusing on local and national areas before reaching out to international ones. I expect this work to be extremely difficult, but hope that it will address the calls for increased regional data and data justice.

6 LIMITATIONS

There are a number of limitations to this work. Of course, the most obvious is that I currently only have two participants. However, even when I do have a larger number of participants, because it is focused on specific roles, the findings may not be widely generalizable outside of those roles. It is also not clear that the findings will be general for other kinds of data or for locations outside of the United States.

Another limitation is that environmental science and environmental data span a number of different fields and subfields. It is unlikely that I will have representative participation for all of these different fields. Thus, I will have to determine how broadly to apply my findings to the totality of environmental data. It may be possible to find related literature or existing tools to collect different types of data rather than having to conduct an interview about all the different types.

7 FUTURE WORK

The most pressing future work for this project is to conduct additional interviews. It is difficult to make broad generalizations with only the two interviews that I have conducted thus far. I would like to conduct at least another four interviews for the initial round of analysis. After identifying theories based on grounded theory method, I would then like to conduct at least an additional six interviews in an attempt to disprove those theories. Ideally, I would have closer to 20 or 30 interviews rather than just 12.

Additional future work is to extend this series of interviews beyond environmental data users. Because my ultimate goal is to develop educational citizen science and citizen sensing tools that can be used for non-specific tasks, I will also interview teachers to see how open they are to using these technologies in the classroom.

Furthermore, I plan to use the findings of this study in my future work developing educational tools. I will focus on the types of data that seem simplest and most reasonable for non-experts to collect and include features that have been deemed important in making data trustworthy.

I ultimately hope to submit these findings to a conference. I am targeting the ACM COMPASS (Computing and Sustainable Societies) conference, but because the deadline is in early March, I may have to instead plan for CSCW (Computer Supported Cooperative Work) or CHI (Human Factors in Computing Systems).

8 CONCLUSION

With the rapidly changing environment around the globe, there is a need for additional local data to track these changes. Non-experts have the opportunity to assist in this effort through citizen science and citizen sensing technologies. In this work, I strive to build a list of environmental data that non-experts can collect and to determine the factors that will make the data useful and trustworthy for experts. I designed a comprehensive interview and conducted two interviews to gather data. I began analysis on the interview transcripts using techniques from the grounded theory method. From this analysis, I found four emerging themes – reputation and trust, risk of data collection, quality over quantity, and the need for more data. I will continue to conduct interviews and find emerging themes for at least 10 additional participants.

This work makes a number of contributions for Computer Science and the broader society. It will result in the creation of a set of environmental data that non-experts can collect to help with research and community planning. It will also highlight the factors that make environment data trustworthy. These findings will help guide the design of future citizen science and citizen sensing tools. They can also inspire other technologies that use or collect environmental data.

REFERENCES

[1] Nassir S. Al-Amri and Ali M. Subyani. 2020. *Analysis of Rainfall, Missing Data, Frequency and PMP in Al-Madinah Area, Western Saudi Arabia.* Springer International Publishing, Cham, 235–248. https://doi.org/10.1007/978-3-030-21874-4_9

[2] Corey T. Callaghan and Dale E. Gawlik. 2015. Efficacy of eBird data as an aid in conservation planning and monitoring. *Journal of Field Ornithology* 86, 4 (2015), 298–304. https://doi.org/10.1111/jofo.12121 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/jofo.12121

[3] An-Jung Cheng, Yan-Ying Chen, Yen-Ta Huang, Winston H. Hsu, and Hong-Yuan Mark Liao. 2011. Personalized Travel Recommendation by Mining People Attributes from Community-contributed Photos. In *Proceedings of the 19th ACM*

International Conference on Multimedia (MM '11). ACM, New York, NY, USA, 83–92. https://doi.org/10.1145/2072298.2072311

[4] Jane E Clougherty, Iyad Kheirbek, Holger M Eisl, Zev Ross, Grant Pezeshki, John E Gorczynski, Sarah Johnson, Steven Markowitz, Daniel Kass, and Thomas Matte. 2013. Intra-urban spatial variability in wintertime street-level concentrations of multiple combustion-related air pollutants: the New York City Community Air Survey (NYCCAS). *Journal of Exposure Science and Environmental Epidemiology* 23, 3 (2013), 232.

[5] Seth Cooper, Firas Khatib, Adrien Treuille, Janos Barbero, Jeehyung Lee, Michael Beenen, Andrew Leaver-Fay, David Baker, Zoran Popović, and Foldit players. 2010. Predicting protein structures with a multiplayer online game. *Nature* 466 (2010), 756–760. https://doi.org/10.1038/nature09304

[6] Alycia W. Crall, Gregory J. Newman, Thomas J. Stohlgren, Kirstin A. Holfelder, Jim Graham, and Donald M. Waller. 2011. Assessing citizen science data quality: an invasive species case study. *Conservation Letters* 4, 6 (2011), 433–442. https://doi.org/10.1111/j.1755-263X.2011.00196.x arXiv:https://conbio.onlinelibrary.wiley.com/doi/pdf/10.1111/j.1755-263X.2011.00196.x

[7] Kerry Emanuel. 2018. *What We Know about Climate Change.* MIT Press.

[8] FAO, IFAD, UNICEF, WFP, and WHO. 2018. The State of Food Security and Nutrition in the World 2018. Building climate resilience for food security and nutrition.

[9] Simon Field. 1992. The Effect of Temperature on Crime. *The British Journal of Criminology* 32, 3 (07 1992), 340–351. https://doi.org/10.1093/oxfordjournals.bjc.a048222 arXiv:http://oup.prod.sis.lan/bjc/article-pdf/32/3/340/909182/32-3-340.pdf

[10] James Garvey. 2008. *The Ethics of Climate Change: Right and Wrong in a Warming World.* Continuum.

[11] Daniel Gooch, Annika Wolff, Gerd Kortuem, and Rebecca Brown. 2015. Reimagining the Role of Citizens in Smart City Projects. In *Adjunct Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2015 ACM International Symposium on Wearable Computers (UbiComp/ISWC'15 Adjunct)*. ACM, New York, NY, USA, 1587–1594. https://doi.org/10.1145/2800835.2801622

[12] Mordechai Haklay and Patrick Weber. 2008. OpenStreetMap: User-Generated Street Maps. *IEEE Pervasive Computing* 7, 4 (Oct 2008), 12–18. https://doi.org/10.1109/MPRV.2008.80

[13] Alisha C Holland, Donna Rose Addis, and Elizabeth A Kensinger. 2011. The neural correlates of specific versus general autobiographical memory construction and elaboration. *Neuropsychologia* 49, 12 (2011), 3164–3177.

[14] Jane Hunter, Abdulmonem Alabri, and Catharine van Ingen. 2013. Assessing the quality and trustworthiness of citizen science data. *Concurrency and Computation: Practice and Experience* 25, 4 (2013), 454–466. https://doi.org/10.1002/cpe.2923 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/cpe.2923

[15] Qijun Jiang, Frank Kresin, Arnold K Bregt, Lammert Kooistra, Emma Pareschi, Edith Van Putten, Hester Volten, and Joost Wesseling. 2016. Citizen sensing for improved urban environmental monitoring. *Journal of Sensors* 2016 (2016).

[16] Margaret Kosmala, Andrea Wiggins, Alexandra Swanson, and Brooke Simmons. 2016. Assessing data quality in citizen science. *Frontiers in Ecology and the Environment* 14, 10 (2016), 551–560. https://doi.org/10.1002/fee.1436 arXiv:https://esajournals.onlinelibrary.wiley.com/doi/pdf/10.1002/fee.1436

[17] Edith Law, Krzysztof Z. Gajos, Andrea Wiggins, Mary L. Gray, and Alex Williams. 2017. Crowdsourcing As a Tool for Research: Implications of Uncertainty. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '17)*. ACM, New York, NY, USA, 1544–1561. https://doi.org/10.1145/2998181.2998197

[18] Dennis Mares. 2013. Climate change and crime: monthly temperature and precipitation anomalies and crime rates in St. Louis, MO 1990–2009. *Crime, Law and Social Change* 59, 2 (01 Mar 2013), 185–208. https://doi.org/10.1007/s10611-013-9411-8

[19] Randall V. Martin, Michael Brauer, Aaron van Donkelaar, Gavin Shaddick, Urvashi Narain, and Sagnik Dey. 2019. No one knows which city has the highest concentration of fine particulate matter. *Atmospheric Environment: X* 3 (2019), 100040. https://doi.org/10.1016/j.aeaoa.2019.100040

[20] Michael J Muller and Sandra Kogan. 2010. Grounded theory method in HCI and CSCW. *Cambridge: IBM Center for Social Software* (2010), 1–46.

[21] Jonathan T. Overpeck, Gerald A. Meehl, Sandrine Bony, and David R. Easterling. 2011. Climate Data Challenges in the 21st Century. *Science* 331, 6018 (11 2 2011), 700–702. https://doi.org/10.1126/science.1197869

[22] Rajul E Pandya. 2012. A framework for engaging diverse communities in citizen science in the US. *Frontiers in Ecology and the Environment* 10, 6 (2012), 314–317. https://doi.org/10.1890/120007 arXiv:https://esajournals.onlinelibrary.wiley.com/doi/pdf/10.1890/120007

[23] Mary Robinson. 2018. *Climate Justice: Hope, Resilience, and the Fight for a Sustainable Future.* Bloomsbury Publishing.

[24] Quirin Schiermeier. 2010. The Real Holes in Climate Science. *Nature* 463 (2010), 284–287. https://doi.org/10.1038/463284a

[25] Robert Simpson, Kevin R. Page, and David De Roure. 2014. Zooniverse: Observing the World's Largest Citizen Science Platform. In *Proceedings of the 23rd International Conference on World Wide Web (WWW '14 Companion)*. ACM, New York, NY, USA, 1049–1054. https://doi.org/10.1145/2567948.2579215

[26] Brian L. Sullivan, Christopher L. Wood, Marshall J. Iliff, Rick E. Bonney, Daniel Fink, and Steve Kelling. 2009. eBird: A citizen-based bird observation network in the biological sciences. *Biological Conservation* 142, 10 (2009), 2282 – 2292. https://doi.org/10.1016/j.biocon.2009.05.006

[27] Christopher K. Uejio, Olga V. Wilhelmi, Jay S. Golden, David M. Mills, Sam P. Gulino, and Jason P. Samenow. 2011. Intra-urban societal vulnerability to extreme heat: The role of heat exposure and the built environment, socioeconomics, and neighborhood stability. *Health & Place* 17, 2 (2011), 498 – 507. https://doi.org/10.1016/j.healthplace.2010.12.005 Geographies of Care.

[28] Tongli Wang, Andreas Hamann, David L. Spittlehouse, and Trevor Q. Murdock. 2012. ClimateWNA—High-Resolution Spatial Climate Data for Western North America. *Journal of Applied Meteorology and Climatology* 51, 1 (2012), 16–29. https://doi.org/10.1175/JAMC-D-11-043.1 arXiv:https://doi.org/10.1175/JAMC-D-11-043.1

[29] Yuqing Wang, L. Ruby Leung, John L. McGregor, Dong-Kyou Lee, Wei-Chyung Wang, Yihui Ding, and Fujio Kimura. 2004. Regional Climate Modeling: Progress, Challenges, and Prospects. *Journal of the Meteorological Society of Japan* 82, 6 (2004), 1599–1628. https://doi.org/10.2151/jmsj.82.1599

[30] Jonathan M. Winter, Brian Beckage, Gabriela Bucini, Radley M. Horton, and Patrick J. Clemins. 2016. Development and Evaluation of High-Resolution Climate Simulations over the Mountainous Northeastern United States. *Journal of Hydrometeorology* 17, 3 (2016), 881–896. https://doi.org/10.1175/JHM-D-15-0052.1 arXiv:https://doi.org/10.1175/JHM-D-15-0052.1

A RECRUITMENT EMAIL

Dear _____,

My name is Alex Cabral and I am a PhD student at Harvard University running a study to understand how citizens can collect climate-related data that is useful, relevant, and trustworthy. I am writing to you because [I know that you work with this sort of data your website indicates that you work with this sort of data I was referred to contact you by _____]. I believe that with your background in [climate science urban planning city design] you would be able to provide helpful insight into my study.

I am looking to interview people who work with climate data to get a better understanding of what data they use, how they use it, how citizens may be able to help collect it, and how to make citizen-collected data useful and trustworthy. The interview will take about 60 minutes and can be conducted in person or via Skype. You will be compensated with a $25 gift certificate for your time, and I will be happy to collaborate with you in the future if you are interested in the work.

Please let me know if you would like to participate in this study and when you are available for an interview. Additionally, if you have any colleagues or contacts you believe would be helpful for this study, please let me know.

Thank you,
Alex Cabral

B CONSENT FORM

What is the purpose of this research? We want to learn what climate-related data is useful to scientists and city planners, and how we can make citizen-collected data useful for those stakeholders. This work will help us build tools for students to collect local climate-related data to help scientists and city planners better understand and plan for climate change.

What can I expect if I take part in this research? You will be asked a set of questions about the data you use for your work. Your participation in this study is completely voluntary, and you may

refuse to participate or withdraw from the study without penalty or loss of benefits to which you may otherwise be entitled. There are no risks anticipated in taking part in this study and you are free to leave at any time without penalty or loss of benefits to which you are otherwise entitled.

Your responses will be kept anonymous. We do not collect any information that could be used to directly establish your identity. Audio files will be deleted after they are transcribed. If you share any sensitive information, it will be deleted from the audio transcripts. After completing the interview, you will receive a $25 gift certificate for your time.

If you have questions or concerns about this research, please contact Alex Cabral, Maxwell Dworkin 242, 33 Oxford St, Cambridge, MA 02138, acabral@g.harvard.edu.

Whom to contact about your rights in this research, for questions, concerns, suggestions, or complaints that are not being addressed by the researcher, or research-related harm: Committee on the Use of Human Subjects in Research at Harvard University, 1414 Massachusetts Avenue, Second Floor, Cambridge, MA 02138.Phone: 617-496-2847 (CUHS). Email: cuhs@fas.harvard.edu The study will take about 60 minutes to complete.

What should I know about a research study?

- Whether or not you take part is up to you.
- Your participation is completely voluntary.
- You can choose not to take part.
- You can agree to take part and later change your mind.
- Your decision will not be held against you.
- Your refusal to participate will not result in any consequences or any loss of benefits that you are otherwise entitled to receive.
- You can ask all the questions you want before you decide.

Who can I talk to? If you have questions, concerns, or complaints, or think the research has hurt you, talk to the research team at acabral@g.harvard.edu.